Τ.Ε.Ι. ΠΑΤΡΑΣ
ΣΧΟΛΗ ΔΙΟΙΚΗΣΗΣ ΚΑΙ ΟΙΚΟΝΟΜΙΑΣ
ΤΜΗΜΑ ΛΟΓΙΣΤΙΚΗΣ

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΜΕ Θ Ε Μ Α :

«ΕΙΔΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΜΕ ΕΜΦΑΣΗ ΣΤΗΝ ΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΕΦΑΡΜΟΓΗ ΤΗΣ»

ΕΠΙΜΕΛΕΙΑ: ΒΑΝΑ ΒΑΣΙΛΙΚΗ
ΓΚΟΛΟΜΑΖΟΥ ΜΑΓΔΑΛΙΝΗ
ΞΕΠΑΠΑΔΕΑ ΣΟΥΛΤΑΝΑ

ΚΑΘΗΓΗΤΡΙΑ : ΚΑΡΥΩΤΗ ΒΑΣΙΛΙΚΗ

ΠΑΤΡΑ
ΙΟΥΛΙΟΣ 2009

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ	Σελ.4
ΚΕΦΑΛΑΙΟ 1ο	7-33
ΣΥΣΧΕΤΙΣΗ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ	
1.1 ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ ΣΤΑΤΙΣΤΙΚΗΣ –ΜΕΣΗ ΤΙΜΗ ΚΑΙ ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ	7
1.2 ΣΥΣΧΕΤΙΣΗ	11
1.3 ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ	17
1.4 ΣΥΝΑΡΤΗΣΙΑΚΗ ΕΞΑΡΤΗΣΗ	27
1.5 ΣΤΟΧΑΣΤΙΚΗ ΕΞΑΡΤΗΣΗ	28
1.6 ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΕΙΔΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ	30
ΚΕΦΑΛΑΙΟ 2ο	34-53
ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ	
2.1 ΓΕΝΙΚΑ	34
2.2 ΜΟΝΤΕΛΑ ΠΙΘΑΝΟΤΗΤΑΣ ΚΑΙ ΣΥΝΘΗΚΕΣ ΙΣΧΥΟΣ ΤΟΥΣ	34
2.3 ΟΡΙΣΜΟΣ ΣΦΑΛΜΑΤΟΣ	36
2.4 ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ	38
2.5 ΔΕΙΚΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ	44
2.6 ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΙ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ	47
2.7 ΓΡΑΦΙΚΗ ΠΑΡΑΣΤΑΣΗ ΤΩΝ ΥΠΟΛΟΙΠΩΝ	49
2.7.1 ΕΛΕΓΧΟΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΤΗΣ ΕΞΙΣΩΣΗΣ	50
2.7.2 ΕΛΕΓΧΟΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΤΟΥ ρ (συντελεστής συσχέτισης)	52
2.8 ΓΡΑΦΙΚΗ ΠΑΡΑΣΤΑΣΗ ΤΩΝ ΥΠΟΛΟΙΠΩΝ	52
ΚΕΦΑΛΑΙΟ 3ο	54-72
ΔΙΑΦΟΡΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΜΟΡΦΕΣ	
3.1 ΓΕΝΙΚΑ	54
3.2 ΕΠΙΛΟΓΗ ΚΑΜΠΥΛΗΣ ΚΑΤΑΛΛΗΛΗΣ ΜΟΡΦΗΣ	54

3.3 ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ ΚΑΙ ΒΕΛΤΙΣΤΗ ΠΑΡΑΒΟΛΗ	59
3.4 ΕΚΘΕΤΙΚΗ ΕΞΑΡΤΗΣΗ	61
3.5 ΥΠΕΡΒΟΛΙΚΗ ΕΞΑΡΤΗΣΗ	66
3.6 ΠΟΛΛΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ	70
3.7 ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ ΜΗ ΓΡΑΜΜΙΚΟΥ ΥΠΟΔΕΙΓΜΑΤΟΣ	70
ΚΕΦΑΛΑΙΟ 4^{ο}	73-80
ΕΦΑΡΜΟΓΗ ΤΗΣ ΜΕΘΟΔΟΥ ΠΑΛΙΝΔΡΟΜΗΣΗΣ	
4.1 ΤΕΛΙΚΗ ΕΦΑΡΜΟΓΗ	73
ΚΕΦΑΛΑΙΟ 5^{ο}	81-82
ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ	
ΒΙΒΛΙΟΓΡΑΦΙΑ	83

ΕΙΣΑΓΩΓΗ

Υπάρχει μια σειρά ερωτημάτων στα οποία ο ειδικευμένος επιστήμονας πρέπει να είναι ικανός να απαντήσει ώστε να βοηθήσει ουσιαστικά μετόχους και επενδυτές, όπως για παράδειγμα το προβλεπόμενο κέρδος από την επένδυση σε μετοχές ή σε ακίνητα. Απαραίτητη επίσης, είναι η συμβολή του στην επιλογή μεταξύ δύο ή περισσότερων στρατηγικών ανάπτυξης μιας εταιρίας. Ας υποθέσουμε, για παράδειγμα ότι μια ασφαλιστική εταιρία θέλει να αποφασίσει αν θα διαθέσει επιπλέον κεφάλαια για την πρόσληψη νέων πωλητών ή θα δώσει αντίστοιχο ποσό για να αυξήσει τη διαφημιστική της δαπάνη (Χαλικιάς, 2001). Η εμπεριστατωμένη απάντηση σε ένα τέτοιο ερώτημα δίνεται μετά από μετρήσεις και εφαρμογή κατάλληλων μαθηματικών προσεγγίσεων από ειδικούς επιστήμονες.

Η απάντηση στα παραπάνω ερωτήματα συνήθως εξαρτάται από διάφορους παράγοντες ορισμένους από τους οποίους δεν μπορεί να γνωρίζει επακριβώς. Παρόλα αυτά η εκτίμηση και πρόβλεψη παραμέτρων όπως το κέρδος μιας επένδυσης επιτυγχάνεται χρησιμοποιώντας στατιστικές μεθόδους. Επιπλέον υπάρχουν σημαντικά οικονομικά μεγέθη, όπως ο *πληθωρισμός*[1], που επηρεάζουν τη γενικότερη οικονομική στρατηγική ενός κράτους, κατά συνέπεια κρίνεται απαραίτητη η πρόβλεψή τους. Είναι προφανής η ισχυρή εξάρτηση του πληθωρισμού π.χ. από το ΑΕΠ (Ακαθάριστο Εθνικό προϊόν)[2] ή το ύψος του μέσου ημερομισθίου αλλά και από παράγοντες ανεξάρτητους της στενής οικονομικής πολιτικής όπως οι διεθνείς τιμές του πετρελαίου. Τα περισσότερα οικονομικά μεγέθη δηλαδή παρουσιάζουν πολύπλοκη εξάρτηση από ένα πλήθος παραμέτρων. Όμως, η μελέτη της εξάρτησης που εμφανίζουν δύο μεταβλητές, αφενός αποτελεί το απλούστερο πλαίσιο στο οποίο θα στηριχθεί η μελέτη πολυπλοκότερων εξαρτήσεων αφετέρου, πολλές φορές οδηγεί σε χρήσιμα και

[1] Ορίζεται ως η γενική τάση αύξησης της τιμής των αγαθών και υπηρεσιών, πρακτικά δηλαδή ως μείωση της αγοραστικής δύναμης των πολιτών (Γεωργακόπουλος κ.α., 1989)
[2] Ως το σύνολο της παραγωγής υλικών και μη υλικών (υπηρεσιών) αγαθών (Γεωργακόπουλος, ό.π.)

αξιόπιστα συμπεράσματα. Αυτό οφείλεται στο γεγονός ότι ορισμένα μεγέθη εξαρτώνται ισχυρά από έναν παράγοντα, όπως για παράδειγμα η αγοραστική δύναμη των μισθωτών από τον πληθωρισμό. Άρα είναι αναγκαίος ο προσδιορισμός μιας στατιστικής σχέσης μεταξύ τους προκειμένου οι υπεύθυνοι για την άσκηση οικονομική πολιτικής να λάβουν τα απαραίτητα μέτρα βάση πάντα των εκτιμηθέντων παραμέτρων.

Ο αριθμός των παραμέτρων που εισάγουμε κατά την μελέτη ενός φαινομένου εξαρτάται από την δομή και το είδος του προβλήματος. Η εισαγωγή ενός υπερβολικά μεγάλου αριθμού μεταβλητών δεν βοηθά πάντα στην καλύτερη περιγραφή ιδιαίτερα αν δεν επιδρούν όλες το ίδιο σημαντικά. Π.χ. ο ρυθμός ανάπτυξης της οικονομίας μιας χώρας εξαρτάται σε μεγάλο βαθμό από τη βιομηχανική δραστηριότητα της και τις διεθνείς τιμές του πετρελαίου. Αν όμως εισαγάγουμε και άλλες μεταβλητές (όπως ο ρυθμός ανάπτυξης ασιατικών χωρών για παράδειγμα) ακόμα και αν βρεθεί σημαντική συσχέτιση, το συμπέρασμα είναι λάθος και το φαινόμενο δεν περιγράφεται καλύτερα εφόσον δεν υπάρχει βάσιμη αιτία που να εξηγεί το αποτέλεσμα. Συνεπώς η σχέση αιτίου – αιτιατού είναι θεμελιώδης προκειμένου να ορισθεί και να μελετηθεί σωστά ένα πρόβλημα με σκοπό την εξαγωγή χρήσιμων και ουσιαστικών συμπερασμάτων.

Η συμβολή της Στατιστικής στην περιγραφή, επεξεργασία και πρόβλεψη – εκτίμηση των οικονομικών φαινομένων είναι καθοριστική. Επιτρέπει στον ερευνητή να διερευνήσει πιθανές σχέσεις μεταξύ μεταβλητών, ενώ παράλληλα προσφέρει τα απαραίτητα μαθηματικά εργαλεία για την πλήρη και επιστημονική διερεύνηση των σχέσεων αυτών. Η βασική θεωρία διατυπώνεται για τη μελέτη της εξάρτησης μιας μεταβλητής αποκλειστικά από μια άλλη και στην συνέχεια επεκτείνεται στην εξάρτηση από περισσότερες.

Στην παρούσα εργασία θα μελετηθεί η πιθανή εξάρτηση μεταξύ δύο μεταβλητών και η μαθηματική διατύπωσή της, εφόσον υφίσταται. Επίσης, θα

παρουσιασθούν οι μέθοδοι της Στατιστικής που όχι μόνο απαντούν στο ερώτημα αν υπάρχει ή όχι εξάρτηση αλλά επιπλέον ερευνούν τα όρια αξιοπιστίας των μαθηματικών σχέσεων που προκύπτουν μετά από την επεξεργασία τιμών δύο μεταβλητών που προκύπτουν από δειγματοληψία.

Στο πρώτο κεφάλαιο, θα αναλύσουμε τις απαραίτητες βασικές έννοιες περιγραφής της κατανομής μιας μεταβλητής που προέκυψε από δειγματοληψία, τις κυριότερες μαθηματικές σχέσεις εξάρτησης μεταξύ δύο μεταβλητών και τους τρόπους με τους οποίους ελέγχουμε την εξάρτηση μιας μεταβλητής από μια άλλη.

Στο δεύτερο κεφάλαιο, θα μελετήσουμε την απλούστερη περίπτωση συσχέτισης δύο μεταβλητών *(απλή γραμμική παλινδρόμηση)* και θα αναπτύξουμε τη μαθηματική τεχνική *(μέθοδος ελαχίστων τετραγώνων)* που απαιτείται για την επεξεργασία των τιμών των μεταβλητών προκειμένου να προσδιοριστεί η ακριβής σχέση μεταξύ τους. Πρέπει, εδώ να τονίσουμε ότι η υπολογιζόμενη σχέση αποτελεί μια προσέγγιση της πραγματικής σχέσης που ισχύει μεταξύ των μεταβλητών ενός πληθυσμού. Το μαθηματικό μοντέλο, όμως μας δίνει την δυνατότητα να ελέγξουμε και το ποσοστό ακρίβειας της προσδιοριζόμενης σχέσης.

Στο τρίτο κεφάλαιο, θα χρησιμοποιήσουμε τη μαθηματική θεωρία που αναπτύξαμε για την απλή γραμμική παλινδρόμηση προκειμένου να μελετήσουμε πιο πολύπλοκες εξαρτήσεις μεταξύ των μεταβλητών, οι οποίες όμως με κατάλληλες μετατροπές αντιμετωπίζονται με την ίδια μέθοδο.

Τέλος, στο τέταρτο κεφάλαιο, θα εφαρμόσουμε τη μέθοδο σε ένα πραγματικό πρόβλημα, τον υπολογισμό σχέσης που να δίνει την αξία ενός διαμερίσματος σε σχέση με τα τετραγωνικά εμβαδού του. Για το συγκεκριμένο πρόβλημα αναζητήσαμε και καταγράψαμε αγγελίας πώλησης διαμερισμάτων σε γνωστή ιστοσελίδα εφημερίδας.

ΚΕΦΑΛΑΙΟ 1ο

ΣΥΣΧΕΤΙΣΗ ΚΑΙ ΠΑΛΙΝΔΡΟΜΗΣΗ

1. 1 ΕΙΣΑΓΩΓΙΚΕΣ ΕΝΝΟΙΕΣ ΣΤΑΤΙΣΤΙΚΗΣ –ΜΕΣΗ ΤΙΜΗ ΚΑΙ ΤΥΠΙΚΗ ΑΠΟΚΛΙΣΗ

Στο κεφάλαιο αυτό θα δώσουμε τις *βασικές έννοιες της στατιστικής* που είναι απαραίτητες για τη μελέτη οικονομικών φαινομένων και θα περιγράψουμε την *παλινδρόμηση (regression)* μια από τις σημαντικότερες μεθόδους επεξεργασίας δεδομένων. Το πλεονέκτημα της παλινδρόμησης μέσω της εκτίμησης των παραμέτρων από το δείγμα των παρατηρήσεων είναι ότι μας δίνει τη δυνατότητα π.χ να προβλέψουμε τη *χρονική εξέλιξη οικονομικών παραμέτρων (forecasting of time series data)*, όταν μελετάμε την εξάρτηση ενός οικονομικού μεγέθους από τον παράγοντα του χρόνου να υπολογίσουμε διαστήματα εμπιστοσύνης η να ελέγξουμε την *ορθότητα υποθέσεων (hypothesis testing)* που έχουμε διατυπώσει για οικονομικά φαινόμενα, όπως για παράδειγμα την απόδοση μετοχών σε ένα χρονικό διάστημα μια δεκαετίας ή την εξάρτηση του δείκτη του χρηματιστηρίου από την τιμή του πετρελαίου. Η παλινδρόμηση εφαρμόζεται μόνο σε μεταβλητές που συσχετίζονται δηλαδή στις μεταβλητές που η μεταβολή της μιας έχει ως αποτέλεσμα την αντίστοιχη μεταβολή της άλλης.

Είναι σημαντικό όμως να τονίσουμε ότι οι εκτιμήσεις και οι προβλέψεις που προκύπτουν εφαρμόζοντας την τεχνική της παλινδρόμησης δεν είναι απόλυτες και πρέπει να καθορίζονται πλήρως, κάθε φορά οι προϋποθέσεις κάτω από τις οποίες ισχύουν (Sykes, Wikipedia, 2008). Σε αντίθετη περίπτωση, μπορούμε να οδηγηθούμε σε εσφαλμένα συμπεράσματα για το υπό μελέτη φαινόμενο. Επιπλέον, πάντα αποτελούν μοντελοποίηση της πραγματικότητας

και όχι ακριβή περιγραφή της κατά συνέπεια υπόκεινται σε σφάλματα. Το σημαντικό πλεονέκτημα των μαθηματικών τεχνικών είναι η δυνατότητα υπολογισμού και αξιολόγησης αυτών των σφαλμάτων. Πρακτικά, το σφάλμα δίνει ένα μέτρο του ποσοστού εξάρτησης της μεταβλητής που μελετάμε από παράγοντες που δε λαμβάνονται υπόψη. Για παράδειγμα, αν μελετήσουμε την τιμή πώλησης ενός διαμερίσματος σε συνάρτηση με το εμβαδόν του είναι προφανές ότι θα προκύψει εξάρτηση. Στην συνέχεια μπορούμε να υπολογίσουμε το σφάλμα των υπολογισμών μας. Π.χ. σφάλμα 35 % στην περίπτωση αυτή σημαίνει ότι η τιμή πώλησης εξαρτάται κατά 65% από το εμβαδόν και κατά 35% από άλλους παράγοντες όπως π.χ. η περιοχή ή ο όροφος που βρίσκεται, η παλαιότητα, κτλ.

Επιπλέον, είναι συνήθως πολύ δύσκολο και ασύμφορο να μετρήσουμε τις τιμές των μεταβλητών σε ολόκληρο τον πληθυσμό που μελετάμε, αν παραδείγματος χάρη εξετάζουμε τη σχέση μεταξύ μόρφωσης και μισθού των εργαζομένων (Sykes ο.π.). Για αυτό, οι μετρήσεις γίνονται σε ένα τυχαίο δείγμα, το οποίο προσδιορίζεται εφαρμόζοντας *τεχνικές δειγματοληψίας*[3] ώστε να είναι αξιόπιστο και τα συμπεράσματα που θα προκύψουν από αυτό να ισχύουν για ολόκληρο τον πληθυσμό. Όμως, είναι λογικό η δειγματοληψία να περιέχει ένα ποσοστό αβεβαιότητας.

Πριν προχωρήσουμε όμως στην παρουσίαση των τεχνικών για τη μελέτη της εξάρτησης δύο μεταβλητών είναι απαραίτητο να υπενθυμίσουμε δύο σημαντικά στατιστικά μεγέθη που αναφέρονται στη συμπεριφορά μιας μεταβλητής και είναι απαραίτητα για τη μαθηματική θεμελίωση των σχέσεων που διέπουν την εξάρτηση μιας μεταβλητής από μία ή περισσότερες άλλες. Αυτά είναι η *μέση τιμή* και η *διακύμανση* μιας μεταβλητής.

Έστω τυχαία μεταβλητή X με τιμές x_i με i=1, 2…..n, όπου n το μέγεθος του δείγματος ενός πληθυσμού που αποτελείται από N άτομα, το οποίο

[3] Δειγματοληψία είναι η διαδικασία με την οποία από ένα πληθυσμό επιλέγεται τυχαίο και αντιπροσωπευτικό δείγμα για την πραγματοποίηση μετρήσεων

προέκυψε μετά από συστηματική και αξιόπιστη δειγματοληψία. Ως μέση τιμή της μεταβλητής Χ ορίζεται το μέγεθος (Κιόχος, 1990)

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1.1)$$

όπου $\bar{x}$ η μέση τιμή της μεταβλητής Χ

x_i οι τιμές της μεταβλητής Χ που προέκυψαν με τη δειγματοληψία

n το μέγεθος του δείγματος

Από τον ορισμό της μέσης τιμής είναι φανερό ότι ο υπολογισμός της, ουσιαστικά, προκύπτει από μετρήσεις σε ένα μέρος του πληθυσμού και όχι σε ολόκληρο τον πληθυσμό. Αν μπορούσαμε να πάρουμε μετρήσεις από το σύνολο του πληθυσμού πιθανότατα θα προέκυπτε μια τιμή διαφορετική από αυτή που υπολογίζεται από την σχέση (1.1). Το ίδιο και αν χρησιμοποιούσαμε ένα διαφορετικό δείγμα. Οι δύο τιμές δε θα ήταν όμως σημαντικά διαφορετικές. Κατά συνέπεια ο υπολογισμός της μέσης τιμής εμπεριέχει ένα σφάλμα, που έχει ως αποτέλεσμα να μην μπορούμε να υπολογίσουμε ακριβώς την τιμή ενός μεγέθους. Αυτό που είναι, όμως, δυνατόν και πολύ χρήσιμο, να κάνουμε είναι να βρούμε μια εκτιμήτρια της πραγματικής τιμής του πληθυσμού η οποία για να μας δίνει έναν προσεγγιστικό υπολογισμό της πραγματικής τιμής του πληθυσμού θα πρέπει να πληροί κάποιες ιδιότητες όπως π.χ αμεροληψία, αποτελεσματικότητα κ.α. Σύμφωνα με τα παραπάνω γίνεται, για παράδειγμα, κατανοητό γιατί τα προγνωστικά ποσοστά των κομμάτων στις εκλογές διαφέρουν ανάλογα με την εταιρία δημοσκοπήσεων. Όμως τις περισσότερες φορές συμπίπτουν με αυτά που προκύπτουν από τις εκλογές, με μικρές αποκλίσεις.

Από το παραπάνω παράδειγμα γίνεται φανερό ότι η μέση τιμή ενός μεγέθους που αναφέρεται σε ένα πληθυσμό δεν είναι η πραγματική τιμή της μεταβλητής για το συγκεκριμένο πληθυσμό. Είναι όμως αντιπροσωπευτική και χρήσιμη στην εξαγωγή συμπερασμάτων. Επιπλέον, η Στατιστική δίνει τη δυνατότητα προσέγγισης του σφάλματος εισάγοντας τα μεγέθη *διακύμανση (ή διασπορά ή μεταβλητότητα)* και *τυπική απόκλιση*.

Η διακύμανση ορίζεται με την σχέση

$$\boldsymbol{Var}(\boldsymbol{X}) = \boldsymbol{E}(\boldsymbol{X} - m)^2 \quad (1.2)$$ (διακύμανση μέσου πλυθησμού)

$\mathrm{Var}(\bar{x}) = \sigma^2/n$ (διακύμανση δειγματικού μέσου)

με Vαr(X) την διακύμανση x_i και $\bar{x}$ όπως ορίσθηκαν παραπάνω.

Η τετραγωνική ρίζα της διακύμανσης ονομάζεται τυπική απόκλιση. Τα μεγέθη αυτά είναι ιδιαίτερα σημαντικά αφού μας δείχνουν την διασπορά της τυχαίας μεταβλητής X γύρω από τον αριθμητικό της μέσο.

$$s_x = \sqrt{\boldsymbol{Var}(\boldsymbol{X})} \quad (1.3)$$

Είναι φανερό από τον ορισμό ότι τόσο η διακύμανση όσο και η τυπική απόκλιση είναι θετικοί αριθμοί.

Πρακτικά, υποθέτοντας ότι η τυχαία μεταβλητή X του πληθυσμού ακολουθεί την κανονική κατανομή η οποία ως γνωστό είναι συμμετρική, οι τιμές του μεγέθους X σε ολόκληρο τον πληθυσμό κυμαίνονται μεταξύ x_μ - 3 σ_x και x_μ + 3 σ_x όπου x_μ είναι ο αριθμητικός μέσος του δείγματος και σ_x η τυπική απόκλιση του πληθυσμού η αλλιώς η τετραγωνική ρίζα της διακύμανσης . Αν επιλέξουμε μια τυχαία τιμή του πληθυσμού ανεξάρτητα αν είναι τιμή του

δείγματος η όχι, η πιθανότητα να ανήκει στο παραπάνω διάστημα είναι 99,74%. Παρατηρούμε συνεπώς, ότι από μετρήσεις τυχαίου δείγματος μπορέσαμε να προβλέψουμε με μικρό σφάλμα τις τιμές για σχεδόν ολόκληρο τον πληθυσμό.

Οι έννοιες της μέσης τιμής, διακύμανσης και τυπικής απόκλισης θα χρησιμοποιηθούν παρακάτω για τη μελέτη της σχέσης εξάρτησης μεταξύ δύο μεταβλητών.

1.2 ΣΥΣΧΕΤΙΣΗ

Όπως αναλύσαμε παραπάνω μας ενδιαφέρει να ελέγξουμε αν δύο (ή περισσότερες) μεταβλητές που αναφέρονται σε έναν πληθυσμό συνδέονται μεταξύ τους με κάποια σχέση και στη συνέχεια να διατυπώσουμε τη σχέση αυτή σε μαθηματική μορφή, δηλαδή με μια εξίσωση που να συνδέει τις δύο μεταβλητές. Το πρώτο βήμα, ο έλεγχος ύπαρξης στατιστικής σχέσης μεταξύ των δύο μεταβλητών ονομάζεται *συσχέτιση*, ενώ *παλινδρόμηση* είναι η μέθοδος που χρησιμοποιούμε προκειμένου να προσδιορίσουμε τη μαθηματική σχέση που τις συνδέει. Στη συνέχεια μπορούμε να χρησιμοποιήσουμε τη σχέση αυτή για να κάνουμε προβλέψεις για την τιμή της μιας, δηλαδή να εκτιμήσουμε τη μία μεταβλητή όταν γνωρίζουμε την τιμή της άλλης. Αναπόφευκτα, στην εκτίμηση που θα κάνουμε θα υπάρχει κάποιο σφάλμα το οποίο προσπαθούμε να ελαχιστοποιήσουμε.

Έστω λοιπόν ότι έχουμε n ζεύγη τιμών δύο μεταβλητών X, Y που προέκυψαν μετά από δειγματοληψία ενός πληθυσμού N ατόμων :

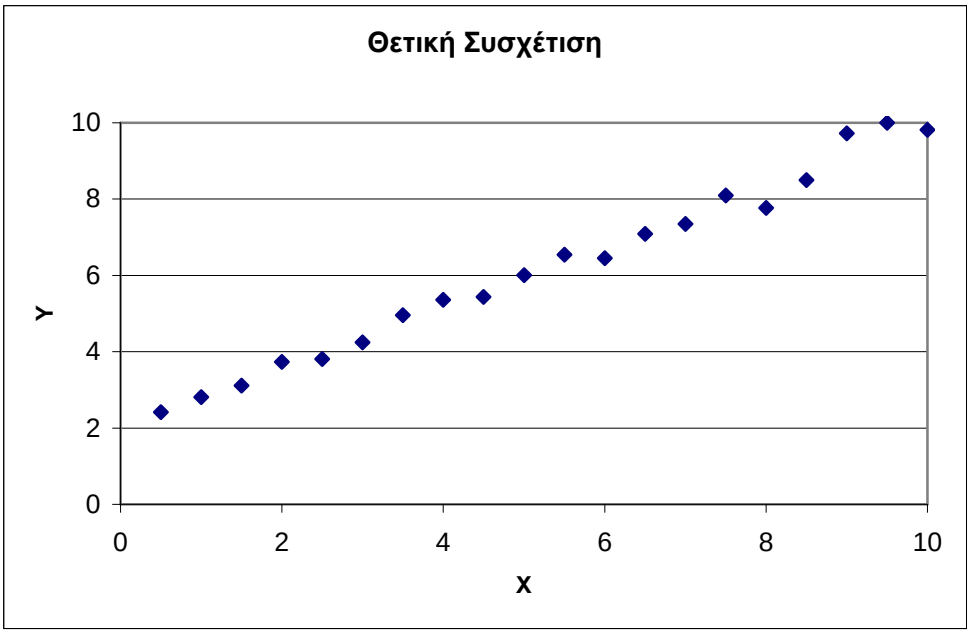
$$(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$$

Στο σημείο αυτό πρέπει να επισημάνουμε ότι υπάρχει διαφορά στον τρόπο προσέγγισης της παλινδρόμησης ανάλογα με την μεταβλητή που θεωρούμε ως δεδομένη και τη μεταβλητή που θέλουμε να εκτιμήσουμε. Την

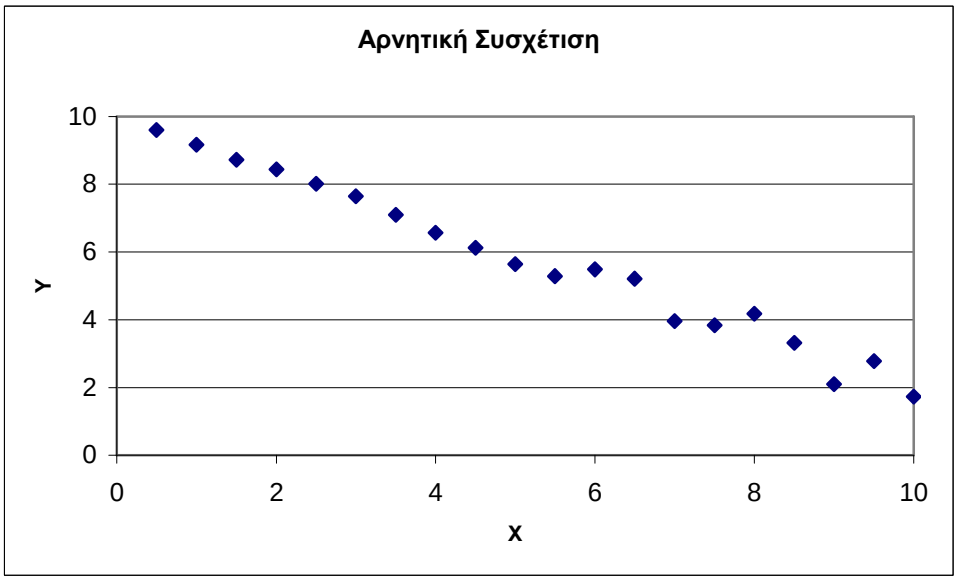
πρώτη την καλούμε *ανεξάρτητη μεταβλητή* και τη δεύτερη *εξαρτημένη (ή ερμηνευτική) μεταβλητή*. Έστω λοιπόν, ότι θέλουμε να εξετάσουμε τον τρόπο που οι τιμές της εξαρτημένης μεταβλητής Υ διαφοροποιούνται καθώς οι τιμές της ανεξάρτητης μεταβλητής Χ αλλάζουν. Ποια μεταβλητή είναι η εξαρτημένη και ποια η ανεξάρτητη καθορίζεται από την σχέση αιτίου – αποτελέσματος. Η ίδια μεταβλητή ανάλογα με το πρόβλημα μπορεί να είναι εξαρτημένη ή ανεξάρτητη. Η αύξηση του πληθωρισμού έχει ως αποτέλεσμα τη μείωση της αγοραστικής δύναμης των καταναλωτών εφόσον με το ίδιο ποσό χρημάτων αγοράζουν λιγότερα αγαθά. Αντίθετα η αύξηση των τιμών του πετρελαίου οδηγεί σε αύξηση του πληθωρισμού. Παρατηρούμε λοιπόν ότι η μεταβλητή 'πληθωρισμός' ανάλογα με το πρόβλημα είναι εξαρτημένη ή ανεξάρτητη.

Κατ' αρχήν πρέπει να εξετάσουμε αν οι δύο μεταβλητές συσχετίζονται ή όχι. Ένας απλός εποπτικός τρόπος για να ελεγχθεί η πιθανή συσχέτιση των δύο μεταβλητών είναι η τοποθέτηση των ζευγών (x_i, y_i) όπου $i=1, 2, \ldots, N$ σε ένα καρτεσιανό σύστημα αξόνων. Κάθε ζεύγος αντιστοιχεί σε ένα σημείο του επιπέδου με τετμημένη x_i και τεταγμένη y_i. Με τον τρόπο αυτό σχηματίζεται ένα σύνολο σημείων που ονομάζεται *νέφος σημείων* ή *διάγραμμα διασποράς*. Γραμμή που να ενώνει όλα τα σημεία του διαγράμματος διασποράς δεν είναι δυνατόν να βρεθεί. Είναι πιθανό όμως τα σημεία να συγκεντρώνονται σε συγκεκριμένες περιοχές του διαγράμματος ακολουθώντας μια νοητή καμπύλη που θυμίζει γνωστές μαθηματικές συναρτήσεις. Η συσχέτιση μπορεί να είναι *θετική* γραμμική όταν αύξηση της τιμής της Χ οδηγεί σε ανάλογη αύξηση και της Υ (Σχήμα 1) ή *αρνητική γραμμική* όταν η αύξηση της Χ έχει ως αποτέλεσμα αντίστοιχα ανάλογη μείωση της Υ (Σχήμα 2). Στις δυο αυτές περιπτώσεις η μαθηματική σχέση μεταξύ των δύο μεταβλητών είναι ευθεία γραμμή. Για παράδειγμα, μπορούμε να υποθέσουμε θετική γραμμική συσχέτιση παρουσιάζουν οι πωλήσεις ενός προϊόντος σε σχέση με την διαφημιστική δαπάνη (Χαλικιάς, 2001), ενώ αρνητική γραμμική οι δαπάνες θέρμανσης σε σχέση με τον αριθμό των παιδιών μιας οικογένειας (Κιόχος, 1990).

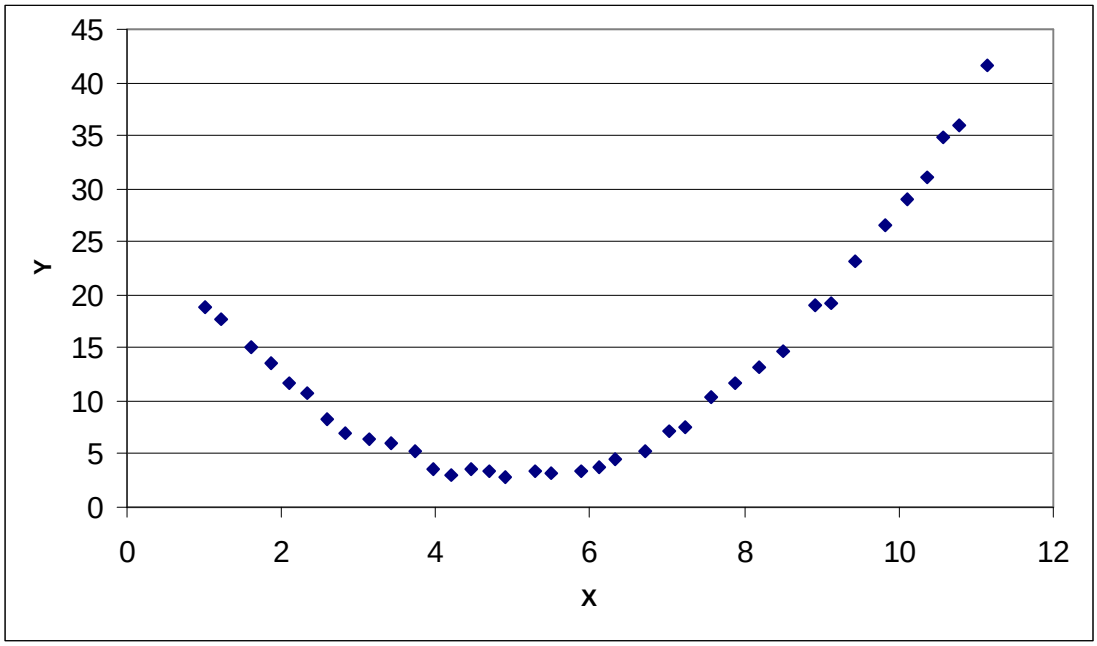
Προφανώς η γραμμική δεν είναι η μόνη πιθανή μαθηματική εξάρτηση μεταξύ δύο μεταβλητών. Άλλες πιθανές είναι η *παραβολική* (Σχήμα 3), η *εκθετική αύξουσα* (Σχήμα 4) ή *φθίνουσα* (Σχήμα 5) και η *υπερβολική* (Σχήμα 6). Τέλος, αν τα σημεία είναι διάσπαρτα τότε οι μεταβλητές δεν συσχετίζονται και λέμε ότι είναι *ασυσχέτιστες* (Σχήμα 7).



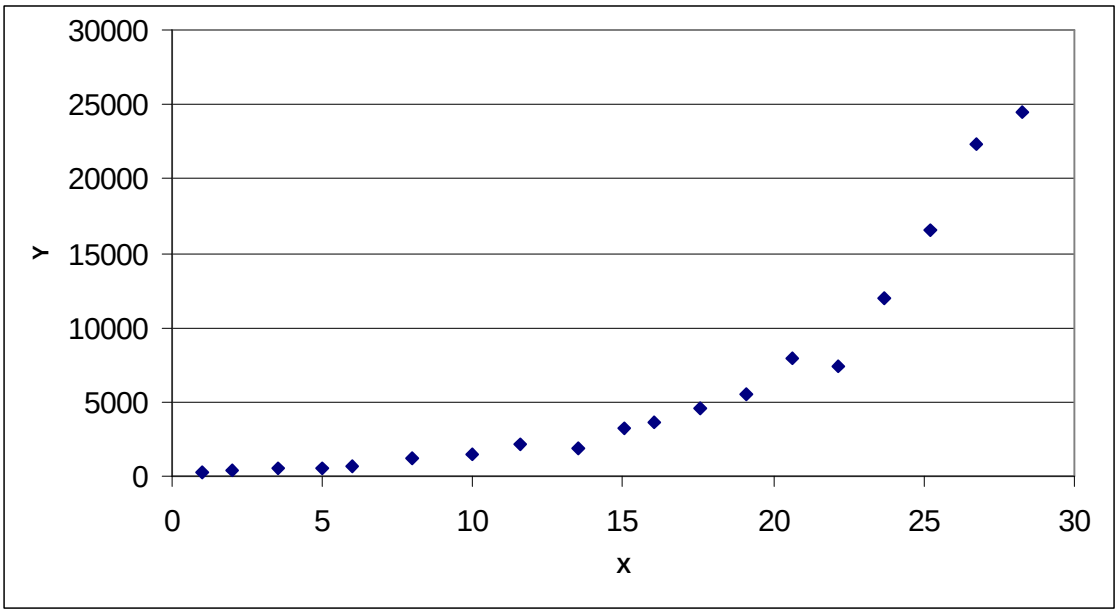
Σχήμα 1. Διάγραμμα διασποράς δύο μεταβλητών X και Y που εμφανίζουν θετική γραμμική συσχέτιση



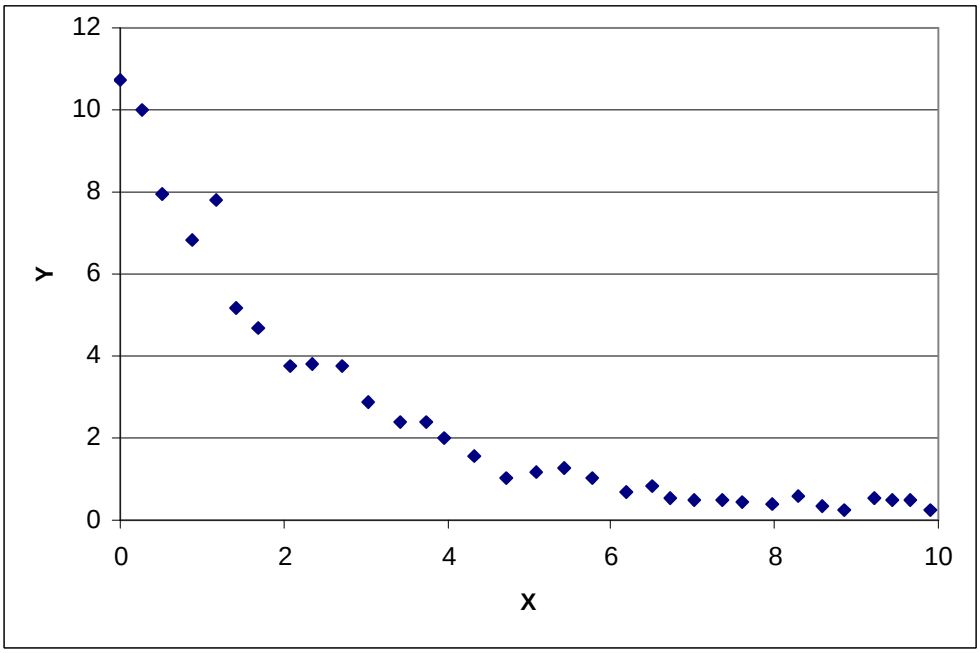
Σχήμα 2. Διάγραμμα διασποράς δύο μεταβλητών X και Y που εμφανίζουν αρνητική γραμμική συσχέτιση



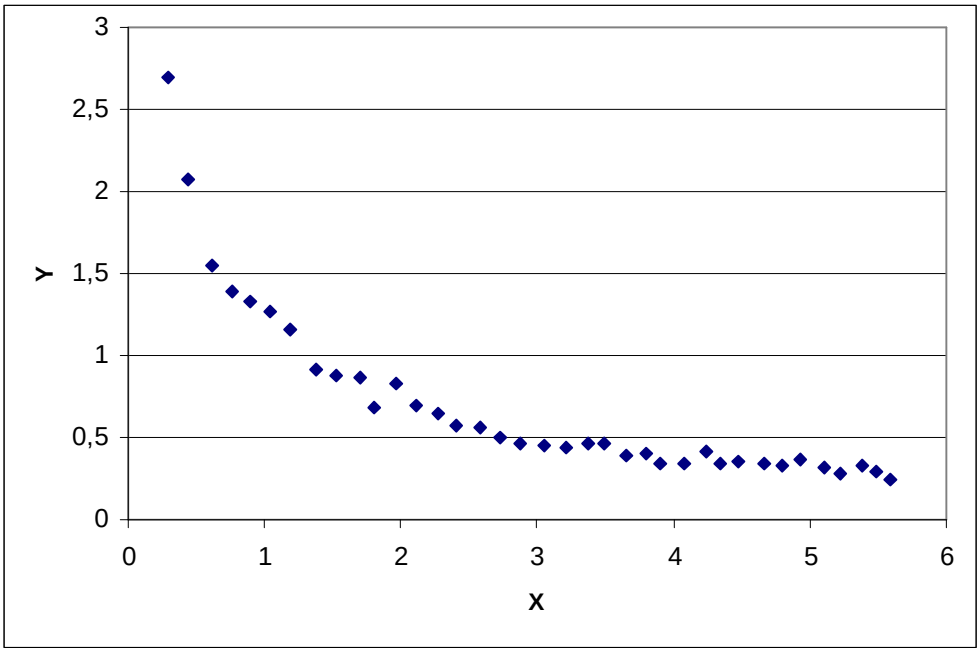
Σχήμα 3. Διάγραμμα διασποράς δύο μεταβλητών X και Y που εμφανίζουν παραβολική συσχέτιση



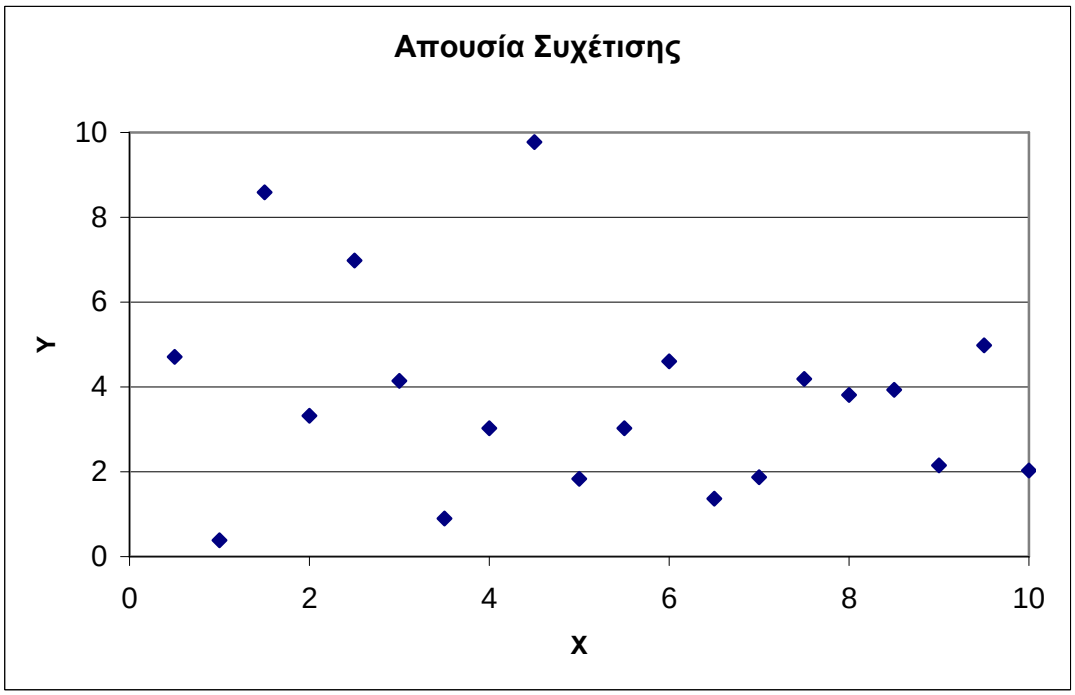
Σχήμα 4. Διάγραμμα διασποράς δύο μεταβλητών X και Y που εμφανίζουν αύξουσα εκθετική συσχέτιση



Σχήμα 5. Διάγραμμα διασποράς δύο μεταβλητών X και Y που εμφανίζουν φθίνουσα εκθετική συσχέτιση



Σχήμα 6. Διάγραμμα διασποράς δύο μεταβλητών X και Y που εμφανίζουν υπερβολική συσχέτιση



Σχήμα 7. Διάγραμμα διασποράς δύο ασυσχέτιστων μεταβλητών X και Y

1.3 ΣΥΝΤΕΛΕΣΤΗΣ ΣΥΣΧΕΤΙΣΗΣ

Το διάγραμμα διασποράς αποτελεί μια εποπτική εικόνα που μας οδηγεί να υποπτευθούμε και κατά συνέπεια να εξετάσουμε πιθανή συσχέτιση μεταξύ δύο μεταβλητών. Για το μαθητικό έλεγχο της συσχέτισης των μεταβλητών, όμως, εισάγουμε την έννοια του *δειγματικού γραμμικού συντελεστή συσχέτισης (correlation coefficient)* ο οποίος ορίζεται με την σχέση (Κιόχος, 1990)

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{(x_i - \bar{x})^2 (y_i - \bar{y})^2}} \qquad (1.4)$$

όπου r = συντελεστής συσχέτισης

x_i = τιμές της ανεξάρτητης μεταβλητής X

$\bar{x}$ = μέση τιμή της ανεξάρτητης μεταβλητής

y_i = τιμές της εξαρτημένης μεταβλητής

$\bar{y}$ = μέση τιμή της εξαρτημένης μεταβλητής

Ο r υπολογίζεται και αναφέρεται στις τιμές του δείγματος και αποτελεί εκτίμηση του πραγματικού συντελεστή γραμμικής συσχέτισης ρ ολόκληρου του πληθυσμού.

Από τον ορισμό προκύπτει ότι ο συντελεστής γραμμικής συσχέτισης είναι καθαρός αριθμός. Οι τιμές του r κυμαίνονται μεταξύ του –1 (τέλεια αρνητική συσχέτιση) μέχρι 1 (τέλεια θετική συσχέτιση). Συνήθως παραδεχόμαστε για τις τιμές του r και την συσχέτιση των μεταβλητών:

$|r| = 0$ γραμμικά ασυσχέτιστα

$|r| \leq 0,3$ δεν έχουμε συσχέτιση

$0,3 \leq |r| \leq 0,6$ έχουμε ασθενή συσχέτιση

$0,5 \leq |r| \leq 0,7$ έχουμε μέση συσχέτιση

$0,7 \leq |r| \leq 0,8$ έχουμε ισχυρή συσχέτιση

$|r| \geq 0,8$ έχουμε πολύ ισχυρή συσχέτιση

$|r| = 1$ έχουμε τέλεια συσχέτιση

Ο συντελεστής γραμμικής συσχέτισης απαντά μόνο στο ερώτημα της πιθανής εξάρτησης μεταξύ των μεταβλητών, οπότε δεν μπορεί να χρησιμοποιηθεί για εκτίμηση της εξαρτημένης μεταβλητής και τη μελέτη της χρονικής εξέλιξης ενός φαινομένου. Επιπλέον περιορίζεται στην εύρεση σχέσης μόνο για μεταβλητές που παρουσιάζουν γραμμική συσχέτιση δηλαδή για όσες το διάγραμμα διασποράς δίνει ευθεία γραμμή. Δύο μεταβλητές είναι δυνατόν να παρουσιάζουν ισχυρή συσχέτιση που όμως να είναι καμπυλόγραμμης μορφής γεγονός που θα δώσει ψευδή αποτελέσματα του συντελεστή συσχέτισης. Επιπλέον, πάντα πρέπει να ελέγχουμε την σχέση αιτίου-αιτιατού στα φαινόμενα που περιγράφουμε.

Αν για παράδειγμα θεωρήσουμε ως μεταβλητές Χ και Υ τους αριθμούς αγορά αυτοκινήτων και κατανάλωσης πίτσας θα διαπιστώσουμε ισχυρή συσχέτιση. Το είδος αυτό ονομάζεται *νόθος συσχέτιση* επειδή ουσιαστικά κρύβει την εξάρτηση και των δύο μεταβλητών από μια τρίτη, την αλλαγή στον τρόπο διαβίωσης. Γίνεται, επομένως για άλλη μια φορά φανερός ο προσδιορισμός αιτίου – αποτελέσματος για την σωστή και χρήσιμη περιγραφή μιας οικονομικής μεταβλητής.

Για τη μελέτη της εξάρτησης μιας μεταβλητής από μια άλλη χρησιμοποιείται και το στατιστικό μέγεθος *συνδιακύμανση* (Κιόχος, 1990). Ορίζεται ως εξής:

$$Cov(X,Y) = \frac{1}{n}\sum_{i}^{n}(x_i - \bar{x})(y_i - \bar{y}) \qquad (1.5)$$

Συνήθως για τον αριθμητικό υπολογισμό της συνδιακύμανσης χρησιμοποιείται η παρακάτω σχέση καθώς απαιτεί λιγότερους αριθμητικούς υπολογισμούς.

$$Cov(X,Y) = \frac{1}{n}\sum_{i}^{n}x_i y_i - \overline{xy} \qquad (1.6)$$

Η σχέση (1.6) προκύπτει από την (1.5) με λίγες απλές πράξεις όπως φαίνεται παρακάτω

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

$$\frac{1}{n}\sum_{i}^{n}x_i y_i - \frac{1}{n}\bar{y}\sum_{i}^{n}x_i - \frac{1}{n}\bar{x}\sum_{i}^{n}y_i + \frac{1}{n}\sum_{i}^{n}\overline{xy}$$

$$= \frac{1}{n}\sum_{i}^{n}x_i y_i - 2\overline{xy} + \overline{xy} = \frac{1}{n}\sum_{i}^{n}x_i y_i - \overline{xy}$$

Επίσης να σημειώσουμε ότι αν η συνδιακύμανση είναι θετικός αριθμός οι μεταβλητές μεταβάλλονται ομόρροπα (αύξηση της μίας οδηγεί σε αύξηση της άλλης), ενώ όταν είναι αρνητικός μεταβάλλονται αντίρροπα.

Η συνδιακύμανση, σε αντίθεση με το συντελεστή συσχέτισης δεν είναι καθαρός αριθμός. Συνεπώς η τιμή της εξαρτάται από τις μονάδες μέτρησης των δύο μεταβλητών. Το γεγονός αυτό καθιστά το συντελεστή συσχέτισης καλύτερο μέτρο προσδιορισμού της ακριβούς γραμμικής σχέσης μεταξύ δύο μεταβλητών. Επίσης, για τον ίδιο λόγο είναι καταλληλότερος για τον προσδιορισμό της ανεξάρτητης μεταβλητής, που επηρεάζει σε ισχυρότερο βαθμό την εξαρτημένη αν η τελευταία δεν εξαρτάται από μία μόνο μεταβλητή (Χαλικιάς, 2001).

Παραδείγματος χάρη, ο προσδιορισμός του συντελεστή συσχέτισης διαφημιστικής δαπάνης - έσοδα πωλήσεων μιας ασφαλιστικής εταιρίας και του αντίστοιχου για αριθμό πωλητών- έσοδα πωλήσεων είναι απαραίτητος προκειμένου να αποφασιστεί ποια από τις δύο ανεξάρτητες μεταβλητές πρέπει να μεταβληθεί για την καλύτερη ανάπτυξη της επιχείρησης.

Ο συντελεστής συσχέτισης και η συνδιακύμανση συνδέονται με την σχέση

$$r = \frac{Cov(X,Y)}{s_x s_y} \qquad (1.7)$$

με σ_x και σ_y οι τυπικές αποκλίσεις των μεταβλητών X,Y όπως αυτές ορίσθηκαν στην προηγούμενη παράγραφο. Η τελευταία σχέση χρησιμοποιείται συχνά για τον υπολογισμό του συντελεστή συσχέτισης.

ΠΑΡΑΔΕΙΓΜΑ 1

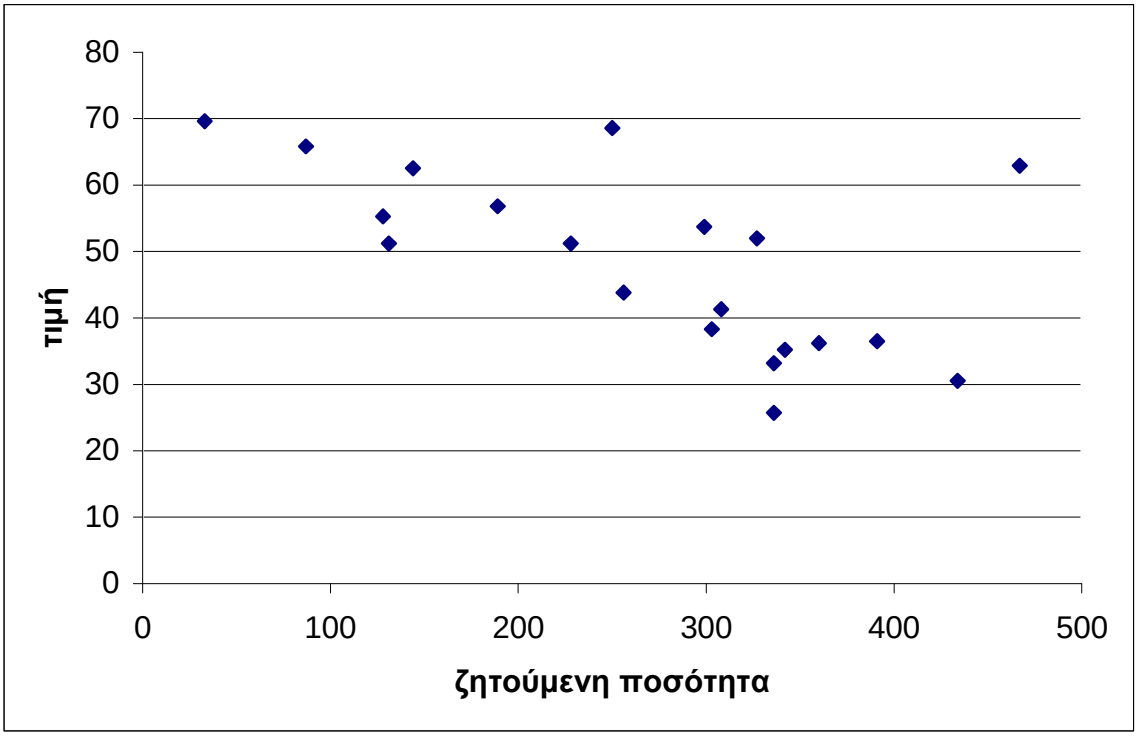
Μια εταιρία προστασίας του καταναλωτή ερευνά τη ζήτηση και την αντίστοιχη τιμή που χρεώνεται για μια χρηματοοικονομική υπηρεσία που

προσφέρεται από μια τράπεζα. Από την έρευνα προκύπτουν τα ακόλουθα δεδομένα.

Πίνακας 1.3 Δεδομένα Δειγματοληψίας

i	ζητούμενη ποσότητα	τιμή
1	33	69,6
2	87	65,8
3	128	55,3
4	131	51,2
5	189	56,8
6	228	51,2
7	256	43,8
8	308	41,3
9	299	53,7
10	327	52
11	360	36,2
12	336	33,2
13	303	38,3
14	342	35,2
15	434	30,5
16	391	36,5
17	336	25,7
18	467	62,9
19	250	68,6
20	144	62,5
Άθροισμα	5349	970,3
Μέση Τιμή	267,45	48,5

Θεωρώντας ως ανεξάρτητη μεταβλητή τη ζητούμενη ποσότητα και εξαρτημένη την τιμή το διάγραμμα διασποράς που προκύπτει για τα παραπάνω δεδομένα είναι:



Σχήμα 10. Διάγραμμα διασποράς μεταξύ ζητούμενης ποσότητας και τιμής

Από την μορφή του διαγράμματος διασποράς πιθανολογούμε αρνητική γραμμική συσχέτιση μεταξύ των δύο μεταβλητών. Για να υπολογίσουμε τον συντελεστή συσχέτισης και να ελέγξουμε την υπόθεση μας πρέπει να κατασκευάσουμε πίνακα αντίστοιχο με τον πίνακα 1.2 για τα δεδομένα δειγματοληψίας.

Πίνακας 1.4 Στοιχεία για τον υπολογισμό του συντελεστή συσχέτισης μεταξύ ζητούμενης ποσότητας και τιμής

i	ζητούμενη ποσότητα	τιμή	$(x_i - \bar{x})^2$	$(y_i - \bar{y})^2$	$(x_i - \bar{x})(y_i - \bar{y})$

1	33	69,6	54966,8	445,21	-4946,895
2	87	65,8	32562,2	299,29	-3121,785
3	128	55,3	19446,3	46,24	-948,26
4	131	51,2	18618,6	7,29	-368,415
5	189	56,8	6154,403	68,89	-651,135
6	228	51,2	1556,303	7,29	-106,515
7	256	43,8	131,1025	22,09	53,815
8	308	41,3	1644,303	51,84	-291,96
9	299	53,7	995,4025	27,04	164,06
10	327	52	3546,203	12,25	208,425
11	360	36,2	8565,503	151,29	-1138,365
12	336	33,2	4699,103	234,09	-1048,815
13	303	38,3	1263,803	104,04	-362,61
14	342	35,2	5557,703	176,89	-991,515
15	434	30,5	27738,9	324	-2997,9
16	391	36,5	15264,6	144	-1482,6
17	336	25,7	4699,103	519,84	-1562,94
18	467	62,9	39820,2	207,36	-2873,52
19	250	68,6	304,5025	404,01	-350,745
20	144	62,5	15239,9	196	-1728,3
Άθροισμα	5349	970,3	262775	3448,95	-18798,9
Μέση Τιμή	267,45	48,5	13138,75	172,4475	1269,929

Οπότε ο συντελεστής συσχέτισης ισούται

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

$$= \frac{-18798,9}{\sqrt{262775 \cdot 3448,95}} = 0,624 = 62,4\%$$

Κατά συνέπεια υπάρχει μέση συσχέτιση μεταξύ της ζήτησης της τραπεζικής υπηρεσίας και της τιμής που αυτή χρεώνεται.

Ο συντελεστής γραμμικής συσχέτισης αποτελεί ένα πρώτο βήμα στη μελέτη της σχέσης μεταξύ δύο μεταβλητών. Το επόμενο βήμα είναι ο μαθηματικός προσδιορισμός της καμπύλης που προκύπτει από το διάγραμμα διασποράς. Για το σκοπό αυτό στις επόμενες παραγράφους θα αναπτύξουμε τις δύο πιθανές μορφές εξάρτησης δύο μεταβλητών, την *συναρτησιακή* και την *στοχαστική*.

ΠΑΡΑΔΕΙΓΜΑ 2

Στον παρακάτω πίνακα δίνονται η ηλικία και οι τιμές της συστολικής αρτηριακής πίεσης δέκα γυναικών. Η πρώτη στήλη αφορά τον δείκτη i που ουσιαστικά είναι ένας αύξοντος αριθμός από 1 έως 12, αφού 12 είναι το πλήθος των δεδομένων μας (n = 12). Η δεύτερη στήλη αφορά τις τιμές ηλικίας των γυναικών, τις οποίες θεωρούμε ως την ανεξάρτητη μεταβλητή X. Η τρίτη στήλη αφορά τις τιμές της αρτηριακής πίεσης που θεωρούμε ως την εξαρτημένη μεταβλητή Y.

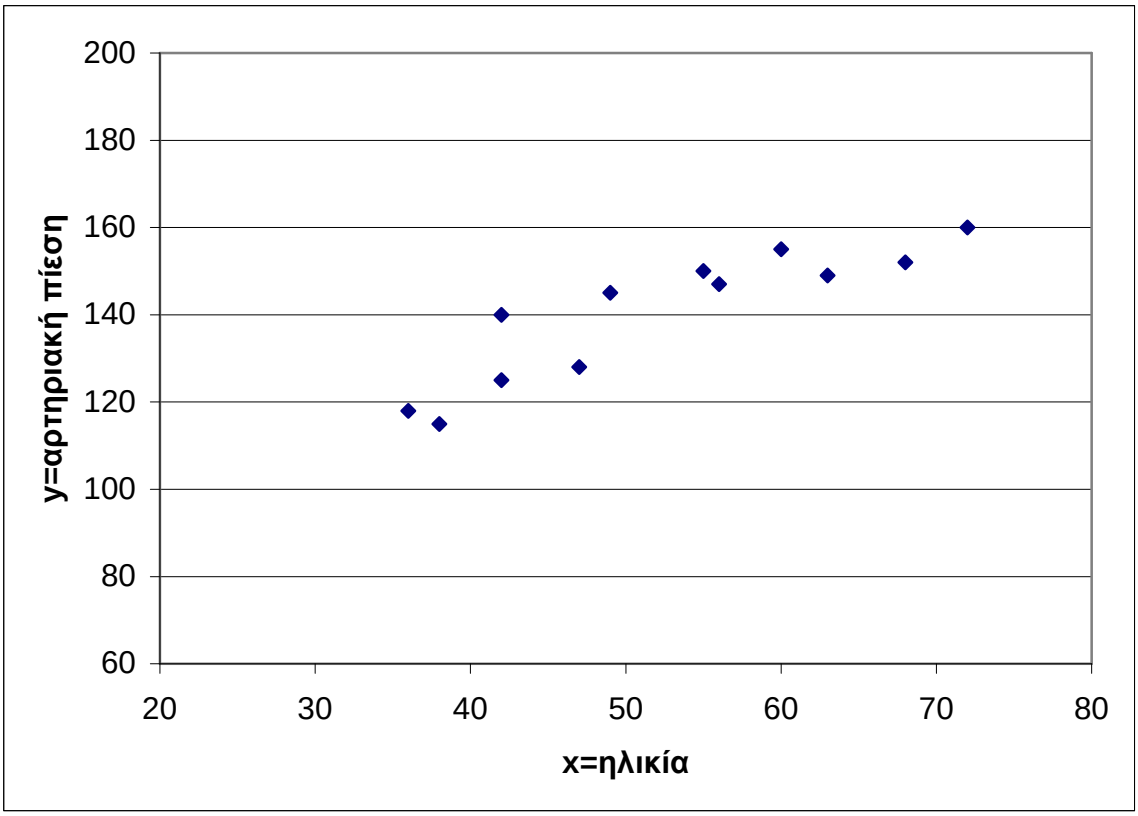
Θα υπολογίσουμε τον συντελεστή συσχέτισης, προκειμένου να ελέγξουμε την πιθανή συσχέτιση μεταξύ ηλικίας και αρτηριακής πίεσης.

Πίνακας 1.1 Δεδομένα Δειγματοληψίας

i	ηλικία xi	αρ.πίεση yi (mmHg)
1	56	147
2	42	125
3	72	160
4	36	118
5	63	149
6	47	128
7	55	150
8	49	145
9	38	115
10	42	140

11	68	152
12	60	155

Το διάγραμμα διασποράς για τα παραπάνω δεδομένα είναι:



Σχήμα 9. Διάγραμμα διασποράς μεταξύ ηλικίας και αρτηριακής πίεσης

Προκειμένου να διερευνήσουμε τη θετική γραμμική συσχέτιση που υποδηλώνει το διάγραμμα πρέπει να υπολογίσουμε το συντελεστή συσχέτισης. Για το σκοπό αυτό κατασκευάζουμε τον πίνακα 1.2, που περιέχει επιπλέον όλα τα δεδομένα που χρειάζονται για τον υπολογισμό του συντελεστή συσχέτισης βάσει της σχέσης (1.4). Για τον υπολογισμό των τιμών στα κελιά των στηλών αυτών χρησιμοποιούνται εκφράσεις των τιμών των μεταβλητών X και Y της

ίδιας γραμμής καθώς και οι μέσες τιμές των X και Y, $\bar{x}$ και $\bar{y}$, αντίστοιχα. Οι μέσες τιμές κάθε στήλης υπολογίζονται στις δύο τελευταίες γραμμές κάθε στήλης.

Πίνακας 1.2 Στοιχεία για τον υπολογισμό του συντελεστή συσχέτισης μεταξύ ηλικίας και αρτηριακής πίεσης.

i	x_i	y_i	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	56	147	13,44	44,44	24,44
2	42	125	106,78	235,11	158,44
3	72	160	386,78	386,78	386,78
4	36	118	266,78	498,78	364,78
5	63	149	113,78	75,11	92,44
6	47	128	28,44	152,11	65,78
7	55	150	7,11	93,44	25,78
8	49	145	11,11	21,78	-15,56
9	38	115	205,44	641,78	363,11
10	42	140	106,78	0,11	3,44
11	68	152	245,44	136,11	182,78
12	60	155	58,78	215,11	112,44
Άθροισμα	628	1684	1550,67	2500,67	1764,67
Μέση Τιμή	52,3333	140,3333	129,2222	208,3889	147,0556

Οπότε ο συντελεστής συσχέτισης είναι :

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

$$= \frac{147,06}{\sqrt{129,22 \cdot 208,39}} = 0,88 = 88\%$$

Κατά συνέπεια υπάρχει έντονη θετική γραμμική συσχέτιση, δηλαδή όσο αυξάνει η ηλικία αυξάνει και η αρτηριακή πίεση. Επιπλέον, υπάρχει λογική σχέση μεταξύ των δύο μεγεθών καθώς οι διαδικασίες γήρανσης του οργανισμού επηρεάζουν όλα τα όργανα κατά συνέπεια και την καρδιά.

1.4 ΣΥΝΑΡΤΗΣΙΑΚΗ ΕΞΑΡΤΗΣΗ

Συναρτησιακή εξάρτηση έχουμε όταν γνωρίζοντας την τιμή της ανεξάρτητης μεταβλητής Χ μπορούμε να εκτιμήσουμε ακριβώς την τιμή της εξαρτημένης μεταβλητής Υ. Στη περίπτωση αυτή όλα τα εμφανιζόμενα ζεύγη τιμών (x_i, y_i) επαληθεύουν μια συνάρτηση με τύπο:

$$y = f(x)$$

Δηλαδή σε κάθε τιμή x_i της μεταβλητής Χ αντιστοιχεί μία μόνο τιμή y_i της Υ χωρίς σφάλμα, η οποία προσδιορίζεται ακριβώς από την μαθηματική σχέση. Τα σημεία (x_i, y_i) βρίσκονται πάνω στη γραφική παράσταση της y = f(x).

Συναρτησιακές εξαρτήσεις συναντάμε κυρίως στις θετικές επιστήμες, π.χ. η σχέση μεταξύ της περιμέτρου Ρ και διαμέτρου D ενός κύκλου (P = π D). Παραδείγματα συναρτησιακών εξαρτήσεων που εμφανίζονται στις οικονομικές επιστήμες είναι:

Κέρδος = Εισπράξεις – Κόστος

Συνολικό Κόστος = Σταθερό κόστος + (Μεταβλητό κόστος

x αριθμό μονάδων που παρήχθησαν)

Οι παραπάνω σχέσεις είναι απόρροια κάποιου *προσδιοριστικού μοντέλου* (deterministic model) που περιγράφει πλήρως την αλληλεξάρτηση των

μεταβλητών (Κέρδος, Εισπράξεις, Κόστος κλπ). Η μελέτη τέτοιων σχέσεων είναι γνωστή από τα μαθηματικά.

1.5 ΣΤΟΧΑΣΤΙΚΗ ΕΞΑΡΤΗΣΗ

Στη πράξη είναι πολύ σπάνιο να έχουμε προσδιοριστικά μοντέλα που να περιγράφουν την εξάρτηση μιας μεταβλητής από μία άλλη. Για παράδειγμα, η τιμή πώλησης ενός διαμερίσματος εξαρτάται από πάρα πολλούς παράγοντες (μεταβλητές) όπως το μέγεθος, το έτος κατασκευής, η θέση του, η ποιότητα κατασκευής και πολλούς άλλους. Αν γνωρίζαμε κάποιους από αυτούς τους παράγοντες δδεν θα ήμασταν σε θέση να εκτιμήσουμε με ακρίβεια την τιμή πώλησης του διαμερίσματος. Αλλά ακόμα και αν γνωρίζαμε όλους αυτούς τους παράγοντες πάλι δεν θα ήμασταν σε θέση να την εκτιμήσουμε γιατί υπάρχουν και άλλοι παράγοντες που δεν γνωρίζουμε ή που δεν μπορούμε να μετρήσουμε. Οι παράγοντες αυτοί ενσωματώνονται στο σφάλμα ε_i. Στις περιπτώσεις αυτές λέμε ότι υπάρχει στοχαστική εξάρτηση μεταξύ της εξαρτημένης μεταβλητής και των ανεξάρτητων μεταβλητών.

Στις δραστηριότητες της καθημερινής ζωής υπάρχει ένα στοιχείο *τυχαιότητας* (randomness) για τον συνυπολογισμό του οποίου πρέπει να χρησιμοποιήσουμε ένα *στοχαστικό μοντέλο* (probabilistic model). Στα στοχαστικά μοντέλα χρησιμοποιούμε και ένα τυχαίο όρο ε που αναφέρεται στις μεταβλητές, μετρήσιμες και μη μετρήσιμες που δεν είναι μέρος του μοντέλου. Σε ένα γραμμικό μοντέλο, για παράδειγμα, η τιμή της εξαρτημένης μεταβλητής Υ θα σχετίζεται με αυτήν της ανεξάρτητης μεταβλητής Χ, σύμφωνα με την σχέση

$$y = \alpha + \beta x + \varepsilon \qquad (1.8)$$

Στο παράδειγμα που αναφέραμε παραπάνω θα μπορούσε κανείς να υποθέσει μια απλή γραμμική σχέση μεταξύ της τιμής πώλησης ενός διαμερίσματος και του μεγέθους του σε τετραγωνικά μέτρα. Αρχικά, υποθέτουμε ένα προσδιοριστικό μοντέλο συσχέτισης:

(τιμή πώλησης) = α (μέγεθος) + β

Είναι προφανές ότι δυο διαμερίσματα ίδιου μεγέθους δεν θα έχουν απαραίτητα την ίδια τιμή, και πρέπει να προσθέσουμε έναν ακόμα τυχαίο όρο ε στη παραπάνω σχέση μετατρέποντας το μοντέλο σε στοχαστικό.

(τιμή πώλησης) = α (μέγεθος) + β + ε

Ο όρος ε αντιπροσωπεύει όλους τους παράγοντες που δεν λάβαμε υπόψη στο μοντέλο μας, αλλά ακόμη και τους παράγοντες εκείνους που δεν μπορούσαμε να λάβουμε υπόψη. Στο σημείο αυτό πρέπει να σημειώσουμε ότι θεωρούμε πως το μοντέλο περιγράφει τη συμπεριφορά όλων των διαμερισμάτων που διατίθενται προς πώληση. Η τιμή πώλησης ενός υποσυνόλου διαμερισμάτων με συγκεκριμένο μέγεθος (π.χ. 100 τ.μ.) είναι μια τυχαία μεταβλητή της οποίας την τιμή δεν μπορούμε να προσδιορίσουμε με ακρίβεια γιατί υπάρχει ο τυχαίος όρος ε.

Στη γενική περίπτωση, στο στοχαστικό μοντέλο θεωρούμε ότι οι τιμές της εξαρτημένης μεταβλητής Y συνδέονται με αυτές της μεταβλητής X, μέσω της σχέσης:

$$y_i = f(x_i; \alpha, \beta) + \varepsilon_i \qquad (1.9)$$

Δηλαδή, οι τιμές της μεταβλητής Y για την δεδομένη τιμή x_i αποτελούνται από ένα σταθερό και ένα τυχαίο όρο.

Η παλινδρόμηση είναι η μαθηματική μέθοδος προσδιορισμού των όρων αυτών, ώστε να μπορούμε για οποιαδήποτε τιμή της ανεξάρτητης μεταβλητής να υπολογίσουμε την τιμή της εξαρτημένης.

1.6 ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΕΙΔΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Το είδος του μοντέλου που θα εισάγουμε για τον υπολογισμό των παραμέτρων της σχέσης εξαρτημένης – ανεξάρτητης μεταβλητής καθορίζεται από το διάγραμμα διασποράς. Αν από το διάγραμμα διασποράς φαίνεται ότι οι δυο μεταβλητές συσχετίζονται, έτσι ώστε να υπάρχει μια καμπύλη η οποία να μπορεί να χρησιμοποιηθεί για την πρόβλεψη της τιμής της μεταβλητής Y, όταν γνωρίζουμε την τιμή της X τότε το επόμενο βήμα είναι να προσδιοριστεί η μαθηματική εξίσωση που περιγράφει αυτήν την καμπύλη. Η εξίσωση αυτή πρέπει να είναι της μορφής

$$y = f(x)$$

έτσι ώστε γνωρίζοντας την τιμή της X να μπορούμε να υπολογίσουμε την τιμή της Y.

Αξίζει να σημειωθεί ότι η παραπάνω σχέση δηλώνει μια *συναρτησιακή εξάρτηση* της μεταβλητής Y από την X σε αντιδιαστολή με την *στοχαστική ή στατιστική εξάρτηση* την οποία ακολουθούν τα αρχικά δεδομένα. Η μετατροπή αυτή από στοχαστική σε συναρτησιακή εξάρτηση έχει ως αποτέλεσμα να υπάρχει σε κάθε ζεύγος δεδομένων (x_i, y_i) ένα σφάλμα ε_i μεταξύ της τιμής y_i και της τιμής $f(x_i)$,

$$\varepsilon_i = y_i - f(x_i) \qquad (1.10)$$

Στόχος της μεθόδου της παλινδρόμησης είναι η εύρεση της συνάρτησης f(x) που ελαχιστοποιεί τα σφάλματα ε_i. Κατ’ αυτό τον τρόπο το αναμενόμενο σφάλμα όταν προσπαθούμε να προσεγγίσουμε την τιμή y από την f(x) ελαχιστοποιείται.

Οι υποθέσεις που κάνουμε για τα σφάλματα είναι οι εξής:

- $E(\varepsilon_\iota) = 0$,
- $Var(\varepsilon_\iota) = \sigma^2_\varepsilon$ (διακύμανση σταθερή άρα ομοσκεδαστικότητα)
- $cov(\varepsilon_\iota, \varepsilon_j) = 0$ (τα σφάλματα είναι ασυσχέτιστα)
- Τα σφάλματα ακολουθούν την κανονική κατανομή. Η υπόθεση της κανονικότητας είναι εξαιρετικά σημαντική αφού αποτελεί την βάση για την διενέργεια των εκάστοτε στατιστικών τεχνικών

Έτσι, οι τιμές της Y για μια δεδομένη τιμή του X, $Y|x_i = x$ ακολουθούν μια κατανομή με μέση τιμή

$$E(Y|x_i = x) = f(x; \alpha, \beta) + E(\varepsilon_i|x_i = x) = f(x; \alpha, \beta)$$

Στη μέθοδο της παλινδρόμησης καλούμαστε να εκτιμήσουμε τις παραμέτρους της συνάρτησης f(x), α, β, ώστε στο τέλος να καταλήξουμε σε μια εκτίμηση για την τιμή της Y, για δεδομένη τιμή της μεταβλητής X, ίση με x. Για το λόγο αυτό λαμβάνουμε ένα δείγμα n παρατηρήσεων από τον πληθυσμό (x_1, y_1), (x_2, y_2), …, (x_n, y_n) και βάσει αυτού εκτιμούμε τις τιμές $\hat{a}$, $\hat{b}$ για τις παραμέτρους α, β ώστε να καταλήξουμε σε μια εκτίμηση $\hat{y}$ για την $E(Y|x_i = x)$.

$$\hat{y} = f(x; \hat{a}, \hat{b})$$

Ο προσδιορισμός της συνάρτησης f(x) δεν είναι εύκολος, ιδιαίτερα στην περίπτωση που δεν γνωρίζουμε τη σχέση εξάρτησης των δύο μεταβλητών, πράγμα που συμβαίνει στις περισσότερες περιπτώσεις. Είναι βέβαια πιθανό τη γενική μορφή της f(x) να την γνωρίζουμε από τη θεωρία που περιγράφει τη συσχέτιση των μεταβλητών X και Y αλλά συνήθως απαιτείται να την εικάσουμε από την μορφή που έχουν τα δεδομένα (x_i, y_i). Στην περίπτωση αυτή προσπαθούμε από τη μορφή του νέφους σημείων να ‘υποπτευθούμε’ την πιθανή εξάρτηση και στην συνέχεια υπάρχουν στατιστικές μέθοδοι που μας επιτρέπουν να ελέγξουμε την ορθότητα της υπόθεσης μας. Επιπλέον, η υπόθεση της γραμμικότητας μπορεί να συνιστά μια απλή μαθηματική μορφή της σχέσης

μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής, σε περίπτωση όμως που υποθέταμε άλλες μορφές εξάρτησης οι μαθηματικοί υπολογισμοί θα ήταν ιδιαίτερα περίπλοκοι ενώ και κατ επέκταση η εξαγωγή στατιστικών συμπερασμάτων ποιο δύσκολη.

Άρα, για την εύρεση της συνάρτησης f(x) πρέπει κατ' αρχήν να βρούμε την γενική μορφή. Για παράδειγμα, η f(x) μπορεί να περιγράφει

- ευθεία γραμμή, $f(x) = \alpha x + \beta$,
- πολυώνυμο n βαθμού, $f(x) = \alpha_n x^n + \alpha_{n-1} x^{n-1} + \ldots + \alpha_1 x + \alpha_0$,
- εκθετική συνάρτηση (αύξουσα ή φθίνουσα), $f(x) = A e^{k x}$
- υπερβολική, $f(x) = A\frac{1}{x}$

Τα είδη παλινδρόμησης μπορούν να συνοψιστούν σε απλή γραμμική παλινδρόμηση, πολλαπλή γραμμική παλινδρόμηση, παλινδρόμηση με την χρήση ψευδομεταβλητών ως ανεξάρτητες μεταβλητές, παλινδρόμηση όπου οι ψευδομεταβλητές είναι η εξαρτημένη μεταβλητή (μοντέλο logit), παλινδρόμηση όπου η ανεξάρτητη μεταβλητή είναι ο χρόνος, παλινδρομήσεις όπου μετασχηματίζουμε τα αρχικά υποδείγματα σε περιπτώσεις που καταργούμε κάποιες αρχικές υποθέσεις π.χ ετεροσκεδαστικότητα, παλινδρόμηση με περιορισμούς, μη γραμμική παλινδρόμηση όπως π.χ η εκθετική συνάρτηση που περιγράψαμε παραπάνω κ.τ.λ .

Η μαθηματικά απλούστερη μορφή της $f(x, d', b')$ είναι η απλή γραμμική παλινδρόμηση $Y=\alpha+\beta X$. Η ευθεία που προκύπτει μετά τον προσδιορισμό των α, β ονομάζεται *ευθεία παλινδρόμησης ή απλή γραμμική παλινδρόμηση*.

Η *παραβολική παλινδρόμηση* προκύπτει για πολυώνυμο βαθμού δύο.

Η μέθοδος των ελαχίστων τετραγώνων μας δίνει την δυνατότητα να υπολογίσουμε την μαθηματική σχέση σύνδεσης των δύο μεταβλητών αλλά και το σφάλμα λόγω της μετατροπής από στοχαστική σε συναρτησιακή εξάρτηση.

Πρακτικά, η διαδικασία αυτή είναι ο υπολογισμός των παραμέτρων της γενικής μορφής της f(x) την οποία έχουμε εικάσει από το νέφος σημείων. Στο επόμενο κεφάλαιο θα μελετήσουμε αναλυτικά μεθόδους προσδιορισμού για την περίπτωση της ευθείας γραμμής όχι μόνο επειδή είναι η μαθηματικά απλούστερη αλλά και επειδή με κατάλληλες μετατροπές των μεταβλητών X και Y περισσότερο πολύπλοκες μορφές της f(x) ανάγονται σε ευθεία γραμμή.

Συνοψίζοντας, η διαδικασία παλινδρόμησης είναι η εύρεση ενός μαθηματικού μοντέλου ικανού να περιγράφει ικανοποιητικά την πραγματικότητα. Είναι προφανές ότι το μοντέλο είναι αδύνατο να είναι απολύτως ακριβές. Όμως είναι απολύτως ικανό να χρησιμοποιηθεί για να προβλέψει τη χρονική εξέλιξη ή να ελέγξει την πιθανή συσχέτιση μεταξύ οικονομικών μεγεθών με ένα μικρό και προσδιορίσιμο σφάλμα.

ΚΕΦΑΛΑΙΟ 2ο
ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

2.1 ΓΕΝΙΚΑ

Στο προηγούμενο κεφάλαιο παρουσιάσαμε το διάγραμμα διασποράς και τον συντελεστή συσχέτισης που αποτελούν σημαντικά εργαλεία για τη μελέτη της εξάρτησης μεταξύ δύο μεταβλητών. Το επόμενο βήμα είναι ο προσδιορισμός της μαθηματικής σχέσης μεταξύ των μεταβλητών ώστε να μπορεί να υπολογιστεί η εξαρτημένη Υ για κάθε πιθανή τιμή της ανεξάρτητης Χ. Η απλούστερη εξάρτηση είναι η γραμμική.

Οπότε, στο κεφάλαιο αυτό, αρχικά θα διατυπώσουμε τις χαρακτηριστικότερες μορφές των συναρτήσεων για *γραμμική παλινδρόμηση* καθώς και τις προϋποθέσεις κάτω από τις οποίες το μοντέλο ισχύει. Στην συνέχεια θα αναπτύξουμε τις μαθηματικές σχέσεις για τον προσδιορισμό των παραμέτρων στην περίπτωση της *απλής γραμμικής παλινδρόμησης*.

2.2 ΜΟΝΤΕΛΑ ΠΙΘΑΝΟΤΗΤΑΣ ΚΑΙ ΣΥΝΘΗΚΕΣ ΙΣΧΥΟΣ ΤΟΥΣ

Η παραμετρική εξάρτηση των δύο μεταβλητών Χ και Υ μπορεί να προκύψει βάσει της θεωρίας ή από τον τρόπο που κατανέμονται τα δεδομένα μας σε ένα διάγραμμα διασποράς. Με την εισαγωγή των παραμέτρων προκύπτει ένα *μοντέλο πιθανότητας* (Πανάρετος, 2001) με το οποίο προσπαθούμε να περιγράψουμε την πραγματική σχέση μεταξύ των μεταβλητών. Είναι προφανές ότι η πραγματική εξάρτηση είναι περισσότερη πολύπλοκη και η πολυπλοκότητα αυτή εκφράζεται με την εισαγωγή του τυχαίου σφάλματος. Πρακτικά,

ελαχιστοποιώντας το σφάλμα προσδιορίζουμε την παραμετρική μορφή της f(x), ώστε στην συνέχεια να μπορέσουμε να χειριστούμε ευκολότερα τις μεταβλητές και να κάνουμε προβλέψεις.

Στην απλούστερη μαθηματική διατύπωση, της *απλής γραμμικής παλινδρόμησης*, η f(x) εξαρτάται από δύο αριθμούς τους α και β μέσω της παρακάτω απλής σχέσης:

$$f(x; \alpha, \beta) = \alpha x + \beta \qquad (2.1)$$

Πιο πολύπλοκες γραμμικές προσεγγίσεις δίνονται π.χ. από συναρτήσεις της μορφής (Πανάρετος, 2001)

$$f(x; \alpha_i, \beta) = \alpha_1 x_1 + \alpha_2 x_2 + \ldots.. + \alpha_k x_k + \beta + \varepsilon \qquad (2.2)$$

$$f(x; \alpha_i, \beta) = \alpha_1 x + \alpha_2 x^2 + \ldots.. + \alpha_k x^k + \beta + \varepsilon \qquad (2.3)$$

Η πρώτη περίπτωση αντιστοιχεί στην *πολλαπλή γραμμική* και η δεύτερη στην *πολλαπλή πολυωνυμική παλινδρόμηση*.

Στην μέθοδο της παλινδρόμησης, τις παραμέτρους αυτές δεν τις γνωρίζουμε αλλά προσπαθούμε να τις εκτιμήσουμε. Η μέθοδος των ελαχίστων τετραγώνων χρησιμοποιείται ακριβώς για την εκτίμηση $\hat{a}$ και $\hat{b}$ των παραμέτρων α και β. Μετά από αυτό μπορούμε να εκτιμήσουμε την αναμενόμενη τιμή $\hat{y}_o$ της Y, για μια τιμή x_o της X βάσει της σχέσης:

$$\hat{y}_0 = \hat{a}x_0 + \hat{b} \qquad (2.4)$$

Τονίζουμε, ότι η πραγματική τιμή y_o που λαμβάνει η Υ μπορεί να είναι πολύ διαφορετική από την $\hat{y}_o$, ανάλογα με το μέγεθος του σφάλματος ε. Στην περίπτωση της απλής γραμμικής παλινδρόμησης η θεώρηση ότι η μέση τιμή των σφαλμάτων είναι ίση με το μηδέν οδηγεί σε καλή εκτίμηση της μέσης τιμής της Υ.

Πρακτικά, η εφαρμογή της μεθόδου ελαχίστων τετραγώνων μετατρέπει το αρχικό στοχαστικό μοντέλο σε ντετερμινιστικό, με την εισαγωγή καταλλήλων συνθηκών. Εξυπακούεται ότι η μετατροπή αυτή δεν είναι ανεξάρτητη σφαλμάτων. Η μέθοδος όμως δίνει επιπλέον τη δυνατότητα υπολογισμού του σφάλματος.

2.3 ΟΡΙΣΜΟΣ ΣΦΑΛΜΑΤΟΣ

Όπως είδαμε στην προηγούμενη παράγραφο, υπάρχει μια διαφορά μεταξύ της εκτιμούμενης τιμής (στην οποία θα καταλήξουμε) και της πραγματικής τιμής της μεταβλητής Υ. Αυτό ισχύει, φυσικά ακόμα και για τις τιμές y_i των ζευγών των παρατηρήσεων (x_i, y_i). Δηλαδή για $X = x_i$ θα εκτιμούμε την τιμή της Υ ως $\hat{y}_i$ με την ακόλουθη σχέση που ονομάζεται *ευθεία παλινδρόμησης*:

$$\hat{y}_i = \hat{a}x_i + \hat{b} \qquad (2.5)$$

η οποία όμως είναι, εν γένει, διαφορετική από την πραγματική τιμή y_i που έχουμε καταγράψει για την Υ. Συμβολίζουμε με e_i την διαφορά των δύο τιμών

$$\hat{e}_i = y_i - \hat{y}_I \qquad (2.6)$$

ή

$$\hat{e}_i = y_i - \hat{a}x_i - \hat{b} \qquad (2.7)$$

Η διαφορά $e^{\wedge}$ εκτιμώμενο ονομάζεται *σφάλμα ή απόκλιση* της μετρούμενης παρατήρησης από την θεωρητική τιμή, που υπολογίζεται με την σχέση (2.5)

Είναι προφανές ότι μια καλή επιλογή των εκτιμητών $\hat{a}$ και $\hat{b}$ για τις παραμέτρους α και β, είναι αυτή που θα ελαχιστοποιήσει τα σφάλματα e_i ώστε η εκτιμούμενη τιμή $\hat{y}_i$ να προσαρμοστεί όσο το δυνατόν καλύτερα στις υφιστάμενες παρατηρήσεις. Οι τιμές των e_i μπορεί να είναι είτε θετικές είτε αρνητικές όμως και στις δύο περιπτώσεις υπάρχει απόκλιση οπότε θέλουμε η επιλογή των $\hat{a}$ και $\hat{b}$ να ελαχιστοποιεί την απόλυτη τιμή του e_i. Στην μέθοδο *ελαχίστων τετραγώνων* προσπαθούμε να εξάγουμε τα $\hat{a}$ και $\hat{b}$ για τα οποία ελαχιστοποιείται το άθροισμα των τετραγώνων Σe_i^2 (εξ ου και το όνομα της μεθόδου). Με τον τρόπο αυτό θα βρούμε μια εξίσωση η οποία θα προσαρμόζεται κατά τον καλύτερο δυνατό τρόπο στις παρατηρήσεις

Το άθροισμα των τετραγώνων των σφαλμάτων ορίζεται ως

$$\Sigma(e_i^2) = \sum_{i=1}^{n} (y_i - \hat{a}x_i - \hat{b})^2 \qquad (2.8)$$

Την έκφραση αυτή θα την χρησιμοποιήσουμε στην επόμενη παράγραφο για την εύρεση των $\hat{a}$ και $\hat{b}$.

2.4 ΕΦΑΡΜΟΓΗ ΜΕΘΟΔΟΥ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ

Όπως βλέπουμε από την προηγούμενη σχέση το άθροισμα των τετραγώνων των αποκλίσεων είναι *συνάρτηση των εκτιμητών* $\hat{a}$, $\hat{b}$ για τις παραμέτρους α και β που θεωρήσαμε στο αρχικό μας μοντέλο (Κιόχος 1994, Χαλικιάς 2001).

$$\Sigma e_i^2 = \varphi(\hat{a}, \hat{b}) \qquad (2.9)$$

Με τον σύμβολο φ παριστάνουμε την συνάρτηση την μορφή της οποίας θα υπολογίσουμε θέτοντας κατάλληλες συνθήκες.

Μπορούμε να βρούμε πότε ελαχιστοποιείται η συνάρτηση αυτή χρησιμοποιώντας μεθόδους ανάλυσης πολλών μεταβλητών. Συγκεκριμένα, για να εμφανίζει ελάχιστο μια συνάρτηση πολλών μεταβλητών απαραίτητη συνθήκη είναι:

$$\frac{\partial f(\hat{a}, \hat{b})}{\partial \hat{a}} = 0 \qquad (2.10)$$

και

$$\frac{\partial f(\hat{a}, \hat{b})}{\partial \hat{b}} = 0 \qquad (2.11)$$

όπου το σύμβολο ∂ χρησιμοποιείται για να δηλώσει την μερική παράγωγο της συνάρτησης ως μια μεταβλητή της, δηλαδή την παράγωγο της συνάρτησης ως προς την μια μεταβλητή θεωρώντας ότι η άλλη είναι σταθερή. Έτσι

$$\frac{\partial f(a,b)}{\partial a} = \frac{\partial E(e_i^{\,2})}{\partial a} = \frac{\partial}{\partial a}\sum_{i=1}^{n}\frac{(y_i - ax_i - b)^2}{n} =$$

$$= \sum_{i=1}^{n}\left[\frac{\partial}{\partial a}\frac{(y_i - a\,x_i - b)^2}{n}\right] = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial a}(y_i - a\,x_i - b)^2\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[2(y_i - a\,x_i - b)\frac{\partial}{\partial a}(y_i - a\,x_i - b)\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n}\left[2(y_i - a\,x_i - b)(0 - x_i - 0)\right] = \frac{-2}{n}\sum_{i=1}^{n}\left[(y_i - a\,x_i - b)\,x_i\right] =$$

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

$$= \frac{-2}{n}\left[\sum_{i=1}^{n}(x_i y_i) - a\sum_{i=1}^{n}x_i^{\,2} - b\sum_{i=1}^{n}x_i\right] \qquad (2.12)$$

ομοίως

$$\frac{\partial f(a,b)}{\partial b} = \frac{\partial E(e_i^{\,2})}{\partial b} = \frac{\partial}{\partial b}\sum_{i=1}^{n}\frac{(y_i - a\,x_i - b)^2}{n} = \sum_{i=1}^{n}\left[\frac{\partial}{\partial b}\frac{(y_i - a\,x_i - b)^2}{n}\right] =$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[\frac{\partial}{\partial b}(y_i - a\,x_i - b)^2\right] = \frac{1}{n}\sum_{i=1}^{n}\left[2(y_i - a\,x_i - b)\frac{\partial}{\partial b}(y_i - a\,x_i - b)\right] =$$

$$\frac{1}{n}\sum_{i=1}^{n}\left[2(y_i - a\,x_i - b)(0 - 0 - 1)\right] = \frac{-2}{n}\sum_{i=1}^{n}\left[y_i - a\,x_i - b\right] =$$

$$= \frac{-2}{n}\left[\sum_{i=1}^{n}y_i - a\sum_{i=1}^{n}x_i - nb\right] \qquad (2\text{-}13)$$

Οπότε καταλήγουμε στο σύστημα εξισώσεων:

$$\left.\begin{aligned}\frac{\partial f(a,b)}{\partial a}=0\\ \frac{\partial f(a,b)}{\partial b}=0\end{aligned}\right\}\Rightarrow \begin{aligned}\sum_{i=1}^{n} x_i^{\,2}\, a+\sum_{i=1}^{n} x_i\, b=\sum_{i=1}^{n}(x_i y_i)\\ \sum_{i=1}^{n} x_i\, a+n\, b=\sum_{i=1}^{n} y_i\end{aligned} \qquad (2.14)$$

Το οποίο είναι ένα σύστημα εξισώσεων 2 x 2 με αγνώστους τα a και b.

Υπάρχουν πολλοί τρόποι για να λυθεί το σύστημα αυτό, εμείς θα χρησιμοποιήσουμε την μέθοδο Crammer (ή μέθοδο των οριζουσών). Οι ορίζουσες του συστήματος είναι

$$D=\begin{vmatrix}\sum_{i=1}^{n} x_i^{\,2} & \sum_{i=1}^{n} x_i\\ \sum_{i=1}^{n} x_i & n\end{vmatrix}=n\sum_{i=1}^{n} x_i^{\,2}-\left(\sum_{i=1}^{n} x_i\right)^2 \qquad (2.15)$$

$$D_a=\begin{vmatrix}\sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} x_i\\ \sum_{i=1}^{n} y_i & n\end{vmatrix}=n\sum_{i=1}^{n} x_i y_i-\sum_{i=1}^{n} x_i\sum_{i=1}^{n} y_i \qquad (2.16)$$

$$D_b=\begin{vmatrix}\sum_{i=1}^{n} x_i^{\,2} & \sum_{i=1}^{n} x_i y_i\\ \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i\end{vmatrix}=\sum_{i=1}^{n} x_i^{\,2}\sum_{i=1}^{n} y_i-\sum_{i=1}^{n} x_i\sum_{i=1}^{n} x_i y_i \qquad (2.17)$$

Οπότε η λύση του συστήματος είναι:

$$\hat{a} = \frac{D_a}{D} = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (2.18)$$

$$\hat{b} = \frac{D_b}{D} = \frac{\sum x_i^{\,2} \sum y_i - \sum x_i \sum x_i y_i}{n\sum x_i^{\,2} - \left(\sum x_i\right)^2} \qquad (2.19)$$

Η λύση αυτή ελαχιστοποιεί την συνάρτηση $\varphi(\hat{a},\hat{b})$ για την μέση τιμή των τετραγώνων των αποκλίσεων γιατί οι δεύτεροι παράγωγοι ως προς $\hat{a}$ και $\hat{b}$ είναι θετικές.

Για τον υπολογισμό των τιμών $\hat{a}$ και $\hat{b}$ εργαζόμαστε όπως και στην περίπτωση του συντελεστή συσχέτισης στο Κεφάλαιο 1.

Η παράμετρος $\hat{a}$ ονομάζεται *γωνιακός συντελεστής* ή *συντελεστής παλινδρόμησης* και καθορίζει την μεταβολή της εξαρτημένης μεταβλητής y από την ανεξάρτητη x. Στην πραγματικότητα η τιμή του $\hat{a}$ συμπίπτει με την μεταβολή του y όταν το x μεταβάλλεται κατά μία μονάδα όπως φαίνεται από τους παρακάτω υπολογισμούς για $x_2 = x_1 + 1$

$$\hat{y}_2 - \hat{y}_1 = (\hat{a}\,x_2 + \hat{b}) - (\hat{a}\,x_1 + \hat{b}) = \hat{a}\,(x_2 - x_1) = \hat{a}$$

Η παράμετρος $\hat{b}$ αντιστοιχεί στην εξαρτημένη μεταβλητή όταν η ανεξάρτητη είναι ίση με το μηδέν και ουσιαστικά είναι το σημείο τομής της ευθείας με τον άξονα Υ. Άρα, είναι δείκτης της εξάρτησης από τους υπόλοιπους παράγοντες, πλην της μεταβλητής x .

Παράδειγμα

Θα προσδιορίσουμε την ευθεία παλινδρόμησης και για τα δεδομένα του πίνακα 1.3 κατασκευάζοντας τον πίνακα 2.2

I	ζητούμενη ποσότητα	Τιμή (€)	$x_i y_i$	y_i^2
1	33	69,6	2296,8	1089
2	87	65,8	5724,6	7569
3	128	55,3	7078,4	16384
4	131	51,2	6707,2	17161
5	189	56,8	10735,2	35721
6	228	51,2	11673,6	51984
7	256	43,8	11212,8	65536
8	308	41,3	12720,4	94864
9	299	53,7	16056,3	89401
10	327	52	17004	106929
11	360	36,2	13032	129600
12	336	33,2	11155,2	112896
13	303	38,3	11604,9	91809
14	342	35,2	12038,4	116964
15	434	30,5	13237	188356
16	391	36,5	14271,5	152881
17	336	25,7	8635,2	112896
18	467	62,9	29374,3	218089
19	250	68,6	17150	62500
20	144	62,5	9000	20736
Άθροισμα	5349	970,3	240707,8	1693365

Από τα δεδομένα του πίνακα προκύπτουν οι τιμές των παραμέτρων

$$\sum_{i=1}^{20} x_i = 5349$$

$$\sum_{i=1}^{20} y_i = 970{,}3$$

$$\sum_{i=1}^{20} x_i y_i = 240707,8$$

$$\sum_{i=1}^{20} x_i^{\,2} = 1693365$$

Οπότε οι παράμετροι $\hat{a}$ και $\hat{b}$ είναι

$$\hat{a} = \frac{20 \cdot 240707,8 - 5349 \cdot 970,3}{20 \cdot 1693365 - 5349^2}$$

$$= \frac{4814156 - 5190135}{33867300 - 28611800} = -\frac{375979}{5255500} = -0,7$$

Περιμέναμε το αρνητικό πρόσημο από την μορφή του διαγράμματος διασποράς. Η τιμή αυτή δείχνει πόσο θα μεταβληθεί η αξία της τραπεζικής συναλλαγής αν η ζήτηση αυξηθεί κατά μια μονάδα και δείχνει την κλίση της ευθείας. Με ανάλογες πράξεις

$$\hat{b} = 67,65$$

Η τιμή του $\hat{b}$ μας δίνει την αξία της συναλλαγής για μηδενική (δηλαδή ελάχιστη ζήτηση).

Όπως έχουμε τονίσει η ευθεία που προκύπτει μετά τον υπολογισμό των παραμέτρων $\hat{a}$, $\hat{b}$ αποτελούν προσέγγιση της πραγματικής που καθορίζει την εξάρτηση των δύο μεταβλητών. Συνεπώς, πρέπει να γίνει έλεγχος της αξιοπιστίας των αποτελεσμάτων και προσδιορισθούν τα διαστήματα στα οποία οι εκτιμήσεις της τιμής Υ είναι αληθείς.

2.5 ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ

Ένα ερώτημα που προκύπτει μετά την κατασκευή της εξίσωσης παλινδρόμησης είναι κατά πόσο η εκτίμηση που δίνει για την εξαρτημένη μεταβλητή Y είναι άξια εμπιστοσύνης. Χρειαζόμαστε δηλαδή ένα δείκτη που να μας δείχνει κατά πόσο οι τιμές της Y εξαρτώνται από την ανεξάρτητη μεταβλητή X που θεωρήσαμε ή από τους υπόλοιπους παράγοντες που δεν λάβαμε υπόψη στο μοντέλο μας. Είναι προφανές ότι όσο περισσότερο εξαρτάται η τιμή της Y από άλλους παράγοντες τόσο μεγαλύτερη θα είναι η *διακύμανση* του σφάλματος ε στο θεωρούμενο στοχαστικό μοντέλο.

Ωστόσο, η διακύμανση του σφάλματος δεν μπορεί συνήθως να παρατηρηθεί καθώς δεν εξετάζουμε όλο τον πληθυσμό αλλά ένα περιορισμένο δείγμα αυτού. Έτσι είμαστε υποχρεωμένοι αντί για την διακύμανση του σφάλματος για ολόκληρο τον πληθυσμό (την οποία είναι αδύνατο να υπολογίσουμε), να χρησιμοποιήσουμε μια εκτίμηση αυτής, s_e βάσει των αποκλίσεων e που παρατηρούμε στο δείγμα μας μεταξύ της εκτιμούμενης και της παρατηρούμενης τιμή στα δεδομένα μας. Έτσι η εκτίμηση που κάνουμε είναι

$$s_e = \sqrt{\frac{\sum_{i=1}^{n} e^{\wedge 2}}{n-2}} \qquad (2.20)$$

με $$e^{\wedge}_i = y_i - \hat{y}_i \qquad (2.21)$$

Στην σχέση αυτή διαιρούμε με n-2 και όχι με το πλήθος n του δείγματος αφού εκτιμώντας τις παραμέτρους $\hat{a}$ και $\hat{b}$ έχουμε προσδιορίσει δύο από παράγοντες που επηρεάζουν την εξαρτημένη μεταβλητή. Το $\hat{e}_i$ ονομάζεται *δειγματικό υπόλοιπο (sample residual)* και είναι η διαφορά μεταξύ της τιμής της εξαρτημένης μεταβλητής που προκύπτει από δειγματοληψία και αυτής που υπολογίζεται από την ευθεία παλινδρόμησης για την ίδια ανεξάρτητη μεταβλητή.

Το άθροισμα *SSE (sum of squared errors)*

$$SSE = \sum_{i=1}^{n} e^2 \qquad (2.22)$$

ονομάζεται *άθροισμα των τετραγώνων των σφαλμάτων* και εκφράζει το κομμάτι της διακύμανσης της μεταβλητής Y που δεν εξηγείται από την διακύμανση της X.

Συγκεκριμένα η διακύμανση της μεταβλητής Y, Var(Y) είναι

$$\text{Var(Y)} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y})^2 \quad = \frac{1}{n} SST \qquad (2.23)$$

Όπου το *SST (total sum of squares)* καλείται *συνολικό άθροισμα τετραγώνων* και αναλύεται σε άθροισμα δύο όρων (με $\hat{y}_i$ παριστάνεται η τιμή της μεταβλητής Y όπως αυτή υπολογίζεται από την ευθεία ελαχίστων τετραγώνων)

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

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

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

$$= \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 + \sum(\hat{y}_i - \bar{y})^2 + 0 =$$

$$= \sum_{i=1}^{n}(e_i)^2 + \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = \text{SSE} + \text{SSR}$$

όπου ***SSR (sum of squared regression)*** ονομάζεται *άθροισμα των τετραγώνων της παλινδρόμησης*

$$\text{SSR} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \qquad (2.24)$$

Βλέπουμε δηλαδή ότι η διακύμανση του Y οφείλεται σε δύο παράγοντες, στην παλινδρόμηση εξαιτίας του υπολογισμού των εκτιμητών $\hat{a}$ και $\hat{b}$, δηλαδή της σχέσης της ανεξάρτητης μεταβλητής και στους υπόλοιπους παράγοντες που εκφράζονται με τον SSE. Ο δείκτης προσδιορισμού R^2 δείχνει κατά πόσον η διακύμανση της εξαρτημένης μεταβλητής Y οφείλεται στην ανεξάρτητη μεταβλητή X ή σε άλλους παράγοντες και ορίζεται ως

$$R^2=SSR/\ SST=\frac{\sum_{i=1}^{n}(\hat{y}_i-\bar{y})^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2} \qquad (2.25)$$

Ο συντελεστής R^2 είναι πάντα θετικός και παίρνει τιμές μεταξύ 0 και 1. Όσο πιο κοντά στην μονάδα είναι η τιμή του, τόσο καλύτερα προσεγγίζει η ευθεία ελαχίστων τετραγώνων που υπολογίσαμε την πραγματική εξάρτηση μεταξύ εξαρτημένης και ανεξάρτητης μεταβλητής. Προφανώς ο υπολογισμός του συντελεστή R^2 στηρίζεται στα δεδομένα δειγματοληψίας. Όμως, η τιμή του ισούται με το τετράγωνο του κανονικού συντελεστή συσχέτισης λόγω ελαχιστοποίησης των τετραγωνικών σφαλμάτων, δηλαδή ισχύει:

$$R^2=\rho^2 \qquad (2.26)$$

Και ισχύει στην περίπτωση της απλής γραμμικής παλινδρόμησης

2.6 ΑΝΑΛΥΣΗ ΔΙΑΚΥΜΑΝΣΗΣ ΚΑΙ ΔΙΑΣΤΗΜΑΤΑ ΕΜΠΙΣΤΟΣΥΝΗΣ

Η ευθεία ελαχίστων τετραγώνων $\hat{y}_i=\hat{a}x_i+\hat{b}$ μας δίνει την δυνατότητα πρόβλεψης της εξαρτημένης μεταβλητής από τιμές της ανεξάρτητης. Τίθεται όμως το ερώτημα πόσο αξιόπιστη είναι η τιμή $\hat{y}_i$ που υπολογίζουμε καθώς ο προσδιορισμός των συντελεστών a,b υπόκειται σε στατιστικά σφάλματα. Άλλωστε, όπως έχουμε σημειώσει συχνά στην εργασία αυτή οι υπολογισμοί στηρίζονται σε μετρήσεις δειγματοληψίας και όχι από το σύνολο του πληθυσμού. Συνεπώς είναι λογικό να παρουσιάζονται αποκλίσεις από τις πραγματικές τιμές που αφορούν ολόκληρο τον πληθυσμό. Η *ανάλυση διακύμανσης* είναι η μεθοδολογία με την οποία προσδιορίζονται οι πηγές των

αποκλίσεων από τις πραγματικές τιμές και ο βαθμός σημαντικότητας τους (Πανάρετος 2001).

Παρά τις αποκλίσεις, όμως, είναι δυνατόν να καθορίσουμε την ακρίβεια της τιμής που υπολογίσαμε χρησιμοποιώντας την έννοια του τυπικού σφάλματος και την κατανομή t, οι τιμές της οποίας δίνονται στο παράρτημα.

Το τυπικό σφάλμα της $\hat{y}$ δίνεται από την σχέση

$$s_y = \sqrt{\frac{SSE}{n-2}} \sqrt{\frac{1}{n} + \frac{\sum_{i=1}^{n} \left(x_i - \bar{x}\right)}{\sum_{i=1}^{n} x_i^{\,2} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}} \qquad (2.27)$$

Το σημαντικό στοιχείο είναι ότι από το τυπικό σφάλμα μπορούμε να υπολογίσουμε το διάστημα τιμών στο οποίο ανήκει η πραγματική τιμή της μεταβλητής Υ με πιθανότητα 1-α, που προκύπτει από την t-κατανομή. Αυτό είναι

$$y = \hat{y} \pm t_{a/2} s_y \qquad (2.28)$$

με $t_{a/2}$ την τιμή που προκύπτει από τους πίνακες της t κατανομής ανάλογα με την ακρίβεια που θέλουμε να υπολογίσουμε την μεταβλητή. Αν για παράδειγμα επιλέξουμε α=0,05 (5%) η πραγματική τιμή έχει πιθανότητα 95% να ανήκει στο διάστημα που επιλέξαμε.

Ομοίως το διάστημα εμπιστοσύνης για το β είναι ($\hat{b}$ – t n-2, a/2 s_b, $\hat{b}$ + t n-2, a /2 s_b)

Όπου $s\hat{b} = \hat{s}_\varepsilon / \sqrt{\Sigma(X-\bar{X})^2}$ επίπεδο σημαντικότητας

Και για το α ($\hat{a}$ – t n-2, a s_α, $\hat{a}$ + t n-2, a s_α)

Όπου s $\hat{a} = \hat{s}_\varepsilon \sqrt{\Sigma(X-\bar{X})^2} / \sqrt{n\Sigma(X-\bar{X})^2}$

2.7 ΕΛΕΓΧΟΣ ΥΠΟΘΕΣΕΩΝ

Αντί πολλές φορές να χρησιμοποιήσουμε διαστήματα εμπιστοσύνης για να βρούμε το εύρος (range) που μπορεί να κινούνται οι παράμετροι του πληθυσμού με κάποιο επίπεδο σημαντικότητας μπορούμε να πάρουμε αυθαίρετα μια τιμή και να αποφανθούμε αν η τιμή αυτή μπορεί να γίνει δεκτή ως υποψήφια τιμή της παραμέτρου του πληθυσμού η να απορριφθεί εξ αρχής. Η διαδικασία με την οποία γίνεται αυτό ονομάζεται έλεγχος υποθέσεων. Αυτό που μας ενδιαφέρει συνήθως είναι η στατιστική σημαντικότητα των συντελεστών διενεργούμε δηλ ελέγχους του τύπου π.χ για την κλίση της γραμμής παλινδρόμησης

H0 : α= 0 H1 : α ≠ 0

Για την διενέργεια του εν λόγω ελέγχου χρησιμοποιούμε την στατιστική $\hat{a} - 0 / \hat{s}a$. Ως γνωστό ο συντελεστής α ακολουθεί την κανονική κατανομή με μέσο α και διακύμανση var($\hat{a}$). Άρα η στατιστική $\hat{a}$ - Ε(α) / $\hat{s}a$ θα ακολουθεί την τυποποιημένη κατανομή student. Να σημειώσουμε πως όταν ελέγχουμε την στατιστική σημαντικότητα του α όπου Ε(α) βάζουμε την τιμή μηδέν. Με την χρήση των γνωστών πινάκων μπορούμε να καταλήξουμε σε συμπεράσματα για τον παραπάνω έλεγχο

Αν η παραπάνω στατιστική βρίσκεται μεταξύ των κρίσιμων τιμών

-ta/2, ν-2 , ta/2, ν-2 τότε δεχόμαστε την υπόθεση ότι ο συντελεστής α που είναι η κλίση της ευθείας είναι στατιστικά μη σημαντικός. Σε διαφορετική περίπτωση δεχόμαστε την εναλλακτική υπόθεση ότι ο συντελεστής είναι διάφορος του μηδενός δηλ. στατιστικά σημαντικός.

Με την ίδια λογική μπορούμε να κάνουμε έλεγχο στατιστικής σημαντικότητας για την σταθερά της εξίσωσης β ή για το $\hat{Y}$ Σε κάθε περίπτωση για να εξάγουμε την στατιστική t υπολογίζουμε από την μέθοδο ελαχίστων τετραγώνων τις δειγματικές παραμέτρους, στην συνέχεια αφαιρούμε την τιμή 0 και διαιρούμε όλο το παραπάνω με την δειγματική τυπική απόκλιση της παραμέτρου. Αφού υπολογίσουμε την στατιστική t ελέγχουμε αν αυτή βρίσκεται στην περιοχή απόρριψης η περιοχή αποδοχής η οποία περιοχή καθορίζεται από τις κρίσιμες τιμές. Να σημειώσουμε πως αν η διακύμανση των σφαλμάτων του πληθυσμού ήταν γνωστή τότε θα χρησιμοποιούσαμε την τυποποιημένη κανονική κατανομή z.

2.7.1 ΕΛΕΓΧΟΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΤΗΣ ΕΞΙΣΩΣΗΣ

Ένα βασικό πρόβλημα στατιστική επαγωγής που μας απασχολεί συνήθως στην παλινδρόμηση είναι αν υπάρχει πράγματι γραμμική σχέση μεταξύ των μεταβλητών X και Y. Μέχρι τώρα είδαμε ότι μπορούμε να το αντιμετωπίσουμε ελέγχοντας την υπόθεση H_0 : β=0 έναντι της H_1 : $\beta \neq 0$.Μπορούμε να ελέγξουμε όμως την υπόθεση αυτή κάνοντας χρήση της ανάλυση της διακύμανσης. Η στατιστική $(\hat{b} - b)\sqrt{\Sigma(ci - \overline{c})^2} / se$ ακολουθεί την τυποποιημένη κατανομή. Το τετράγωνό της θα ακολουθεί την τροποποιημένη κατανομή χ^2 με ν =1 βαθμούς ελευθερίας.

Επιπρόσθετα η στατιστική (δεν θα το αποδείξουμε για λόγους συντομίας) $((\boldsymbol{n}-2)\boldsymbol{S}_{\boldsymbol{e}}^{2}/s_{e}^{2}))/\boldsymbol{n}-2$ ακολουθεί την τροποποιημένη κατανομή χ^2 με ν-2 βαθμούς ελευθερίας. Άρα το πηλίκο των 2 παραπάνω θα ακολουθεί την κατανομή F με βαθμούς ελευθερίας ν=1, κ=ν-2. Αν υποθέσουμε ότι β=0 τότε ο αριθμητής κάτι πάλι που δεν θα το αποδείξουμε ισούται με $\Sigma(\hat{Y}_i-\bar{Y})^2$ ενώ ο παρονομαστής με $\Sigma\hat{e}_i^{\,2}/n-2$ Άρα η στατιστική $\Sigma(\hat{Y}_i-\bar{Y})^2/\ \Sigma\hat{e}_i^{\,2}/n-2$ θα ακολουθεί την στατιστική F1, ν-2, α. Η παραπάνω στατιστική μπορεί να γραφεί και ως $(R^2 / 1- R^2)$ n-2

Συνεπώς για να κάνουμε τον έλεγχο της στατιστικής σημαντικότητας του β δηλ. το έλεγχο της στατιστικής σημαντικότητας της εξίσωσης βρίσκουμε την στατιστική F η οποία όπως αναφέραμε μπορεί να υπολογιστεί με 2 τρόπους και την συγκρίνουμε με την κρίσιμη τιμή F1 , n-2 , 0,05. Αν η στατιστική είναι μικρότερη της κρίσιμης τιμής δεχόμαστε την H_0 ότι η εξίσωση δεν είναι στατιστικά σημαντική. Σε διαφορετική περίπτωση δεχόμαστε την H_1 ότι υφίσταται γραμμική σχέση μεταξύ της Y και της X. Τέλος να σημειώσουμε ότι στην περίπτωση της πολλαπλής παλινδρόμησης όπου έχουμε πολλές ανεξάρτητες μεταβλητές αν θέλουμε να ελέγξουμε την στατιστική σημαντικότητα της εξίσωσης θα χρησιμοποιήσουμε την στατιστική F και σε καμία περίπτωση δεν θα κάνουμε έλεγχο για την στατιστική σημαντικότητα κάθε συντελεστή της ερμηνευτικής μεταβλητής ξεχωριστά.

2.7.2 ΕΛΕΓΧΟΣ ΣΤΑΤΙΣΤΙΚΗΣ ΣΗΜΑΝΤΙΚΟΤΗΤΑΣ ΤΟΥ ρ (συντελεστής συσχέτισης)

Πολλές φορές θέλουμε να κάνουμε τον έλεγχο του αν ο συντελεστής συσχέτισης ρ μεταξύ της εξαρτημένης μεταβλητής Υ και της εξαρτημένης μεταβλητής Χ είναι στατιστικά σημαντικός. Αποφεύγοντας τις εκτενείς αναφορές για το πώς εξάγουμε την αντίστοιχη στατιστική απλώς αναφέρουμε ότι σε περίπτωση που θέλουμε να κάνουμε τον έλεγχο $H_0 : \rho = 0$ έναντι της εναλλακτικής $H_1 : \rho \neq 0$ χρησιμοποιούμε την στατιστική $r\sqrt{n-2} / \sqrt{1-r^2}$ Η παραπάνω στατιστική ακολουθεί την κατανομή student με ν-2 βαθμούς ελευθερίας. Αν η παραπάνω στατιστική είναι μικρότερη από την κρίσιμη τιμή t n-2 , 0,05 τότε δεχόμαστε την $H_0 : \rho = 0$

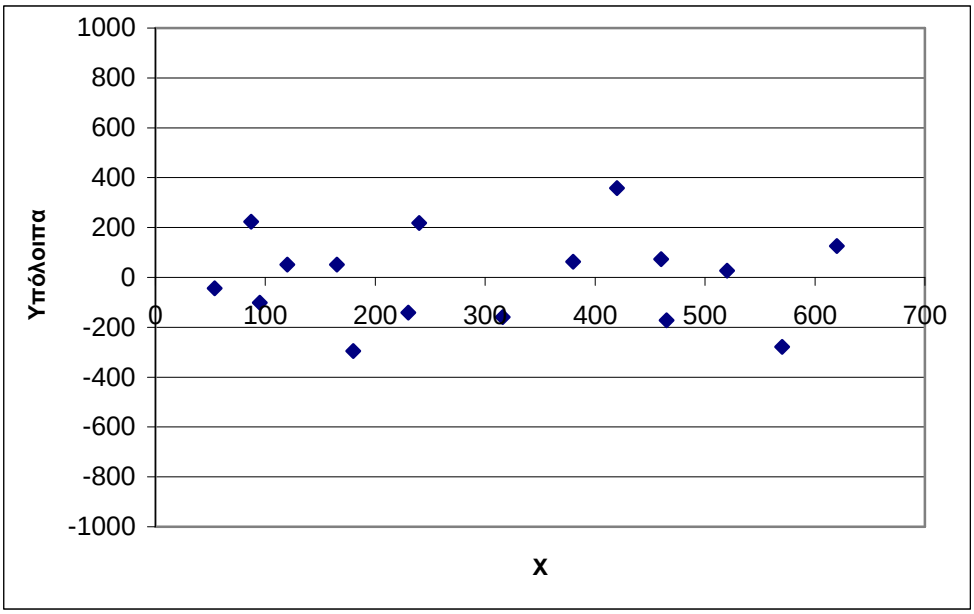
2.8 ΓΡΑΦΙΚΗ ΠΑΡΑΣΤΑΣΗ ΤΩΝ ΥΠΟΛΟΙΠΩΝ

Ένας δεύτερος τρόπος έλεγχου της αξιοπιστίας της ευθείας που προσδιορίζεται με την μέθοδο ελαχίστων τετραγώνων είναι με τα *δειγματικά υπόλοιπα (sample residuals).*

Η γραφική παράσταση των υπολοίπων σε συνάρτηση με τις τιμές της ανεξάρτητης μεταβλητής αποτελεί έναν άμεσο οπτικό δείκτη της αξιοπιστίας της προσέγγισης μέσω της ευθείας ελαχίστων τετραγώνων καθώς και της ακρίβειας των υποθέσεων του μοντέλου.

Για να ισχύει π.χ. η υπόθεση της γραμμικότητας του μοντέλου η γραφική παράσταση των υπολοίπων θα πρέπει να αντιστοιχεί περίπου σε ευθεία γραμμή.

Η πιο σημαντική όμως υπόθεση είναι αυτή της ανεξαρτησίας των μεταβλητών X και Y. Για να ισχύει πρέπει τα υπόλοιπα να κατανέμονται τυχαία και συμμετρικά γύρω από την ευθεία που αντιστοιχεί σε υπόλοιπο ίσο με μηδέν, όπως φαίνεται στο παρακάτω διάγραμμα.



ΚΕΦΑΛΑΙΟ 3ο
ΔΙΑΦΟΡΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΜΟΡΦΕΣ

3.1 ΓΕΝΙΚΑ

Η περίπτωση της απλής γραμμικής παλινδρόμησης, που αναλύθηκε στα προηγούμενα κεφάλαια, μπορεί να επεκταθεί και στην περίπτωση μη γραμμικής συσχέτισης μεταξύ μεταβλητών, για παράδειγμα όταν οι δύο μεταβλητές παρουσιάζουν παραβολική εξάρτηση. Τη μορφή της καλύτερης καμπύλης που προσαρμόζεται στις μετρήσεις μας την 'υποπτευόμαστε' από το διάγραμμα διασποράς.

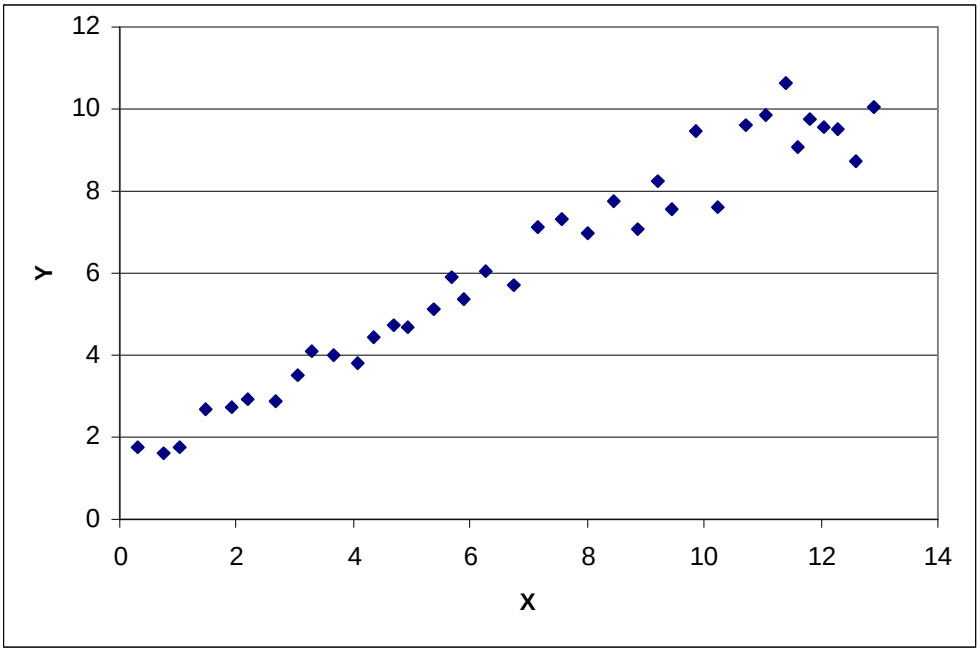
Επιπλέον, είναι δυνατόν εφαρμόζοντας κατάλληλους μετασχηματισμούς των μεταβλητών X και Y να μετατρέψουμε την σχέση εξάρτησης σε γραμμική και στην συνέχεια να εφαρμόσουμε σχέσεις αντίστοιχες με τις (2.10), (2.11) του δευτέρου κεφαλαίου.

Στο κεφάλαιο αυτό θα παρουσιάσουμε τον τρόπο υπολογισμού της καλύτερης καμπύλης σε περιπτώσεις που το διάγραμμα διασποράς δίνει εξάρτηση διαφορετική από τη γραμμική. Πριν όμως προχωρήσουμε σε αυτό οφείλουμε να προτείνουμε τεχνικές ικανές να μας οδηγήσουν στην εκλογή του κατάλληλου μοντέλου για το υπό μελέτη πρόβλημα.

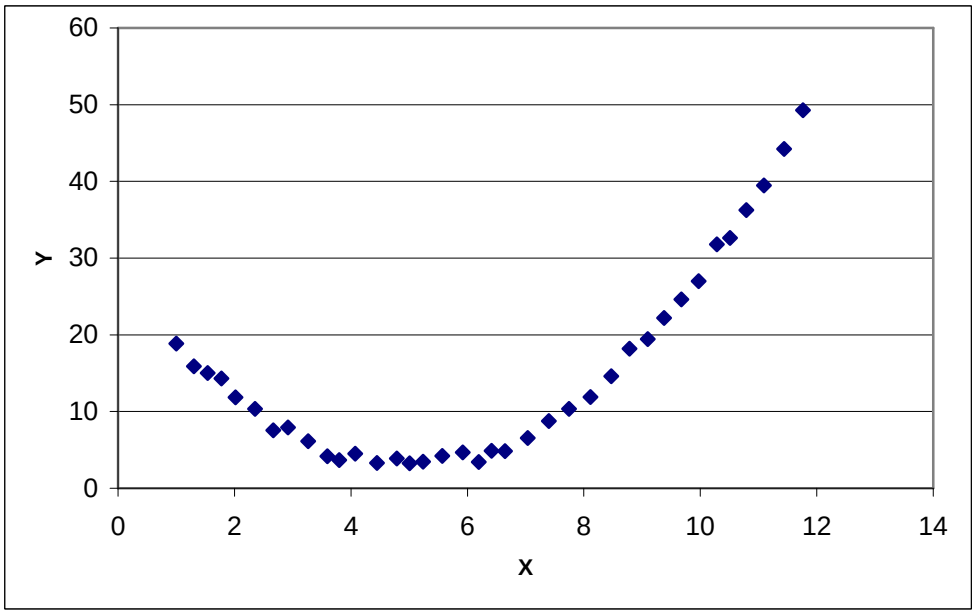
3.2 ΕΠΙΛΟΓΗ ΚΑΜΠΥΛΗΣ ΚΑΤΑΛΛΗΛΗΣ ΜΟΡΦΗΣ

Ας υποθέσουμε ότι έχουμε τοποθετήσει τις τιμές της ανεξάρτητης X και εξαρτημένης μεταβλητής Y, που προέκυψαν μετά από δειγματοληψία, σε

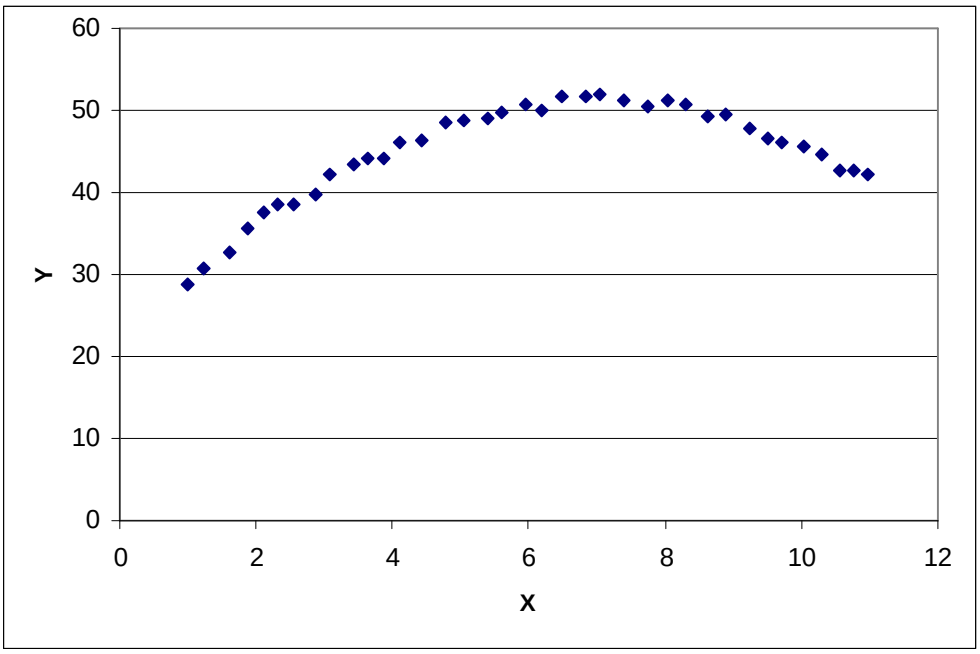
διάγραμμα διασποράς. Στο παρακάτω σχήμα (Κιόχος, 1990) δίνονται πιθανές μορφές του.



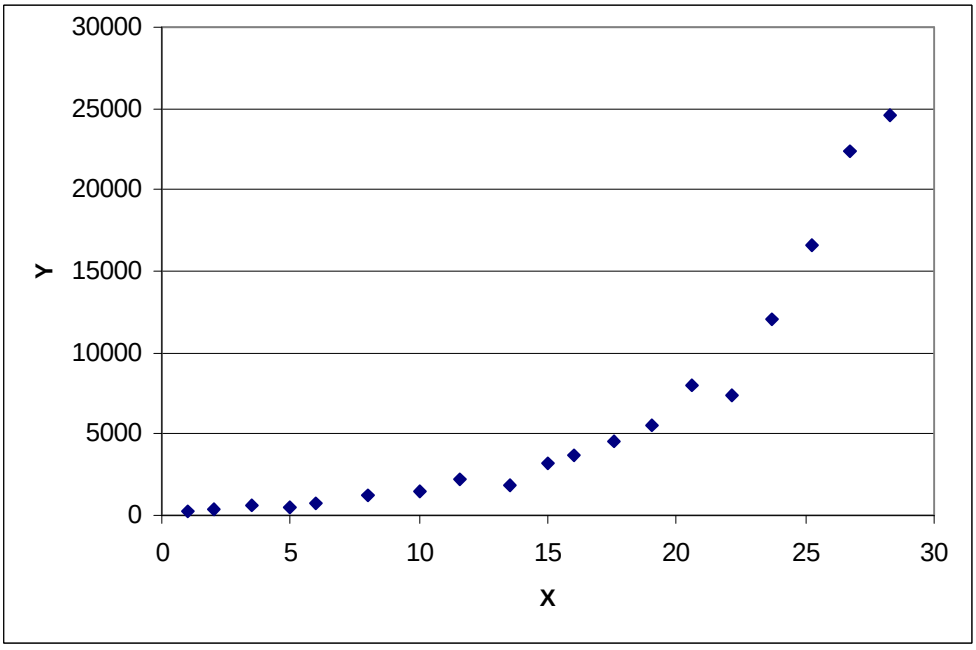
Γραμμική εξάρτηση (y = a x + b)



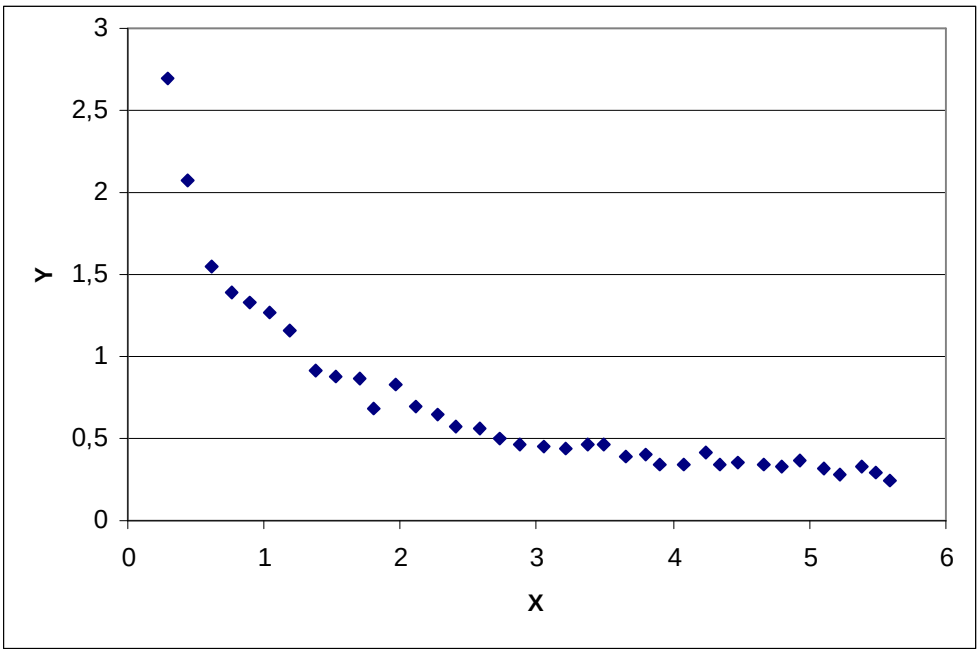
Παραβολική εξάρτηση ($y = a x^2 + bx + c, a > 0$)



Παραβολική εξάρτηση ($y = a\,x^2 + bx + c$, $a < 0$)



Εκθετική εξάρτηση ($y = a\,b^x$)



Υπερβολική εξάρτηση (y = 1/(a+bx))

Από τη γνώση της θεωρίας συναρτήσεων καταλαβαίνουμε ότι πιθανότατα: το πρώτο διάγραμμα δίνει γραμμική συσχέτιση μεταξύ των δύο μεταβλητών, το δεύτερο και το τρίτο *παραβολική* (με θετική και αρνητική συσχέτιση αντίστοιχα), το τέταρτο *υπερβολική* και το πέμπτο *εκθετική*.

Υπάρχουν όμως και άλλα κριτήρια τα οποία μπορούμε να εφαρμόζουμε για να προσδιορίσουμε την σχέση ανεξάρτητης-εξαρτημένης μεταβλητής. Η μεθοδολογία στηρίζεται στον υπολογισμό των διαφορών κάθε τιμής x_i, y_i των μεταβλητών X, Y από την προηγούμενη και την επόμενη (Κιόχος, 1990)

$$\Delta x_i = x_{i-1} - x_i \qquad (3.1)$$

$$\Delta y_i = y_{i-1} - y_i \qquad (3.2)$$

Η ανεξάρτητη μεταβλητή Χ ακολουθεί πάντα περίπου αριθμητική πρόοδο[4]. Η εξαρτημένη μεταβλητή όμως ακολουθεί διάφορες κατανομές. Ανάλογα με την μαθηματική σχέση των τιμών της Υ κατά συνέπεια και των διαφορών Δy προκύπτει και διαφορετική πιθανότερη καμπύλη για τις δειγματοληπτικές μετρήσεις. Ειδικότερα:

Α) όταν η μεταβλητή Υ παρουσιάζει περίπου αριθμητική πρόοδο ή ισοδύναμα οι διαφορές Δy είναι περίπου σταθερές τότε η καταλληλότερη καμπύλη είναι ευθεία και χρησιμοποιείται η μέθοδος ελαχίστων τετραγώνων για τον προσδιορισμό της .

Β) ορίζουμε ως δεύτερη μεταβολή της εξαρτημένης μεταβλητής το μέγεθος

$$\Delta^2 y_i = \Delta y_i - \Delta y_{i-1} \qquad (3.3)$$

με Δy την διαφορά όπως ορίζεται από την σχέση (4.2). Όταν οι δεύτερες διαφορές $\Delta^2 y$ της μεταβλητής Υ είναι σταθερές ή περίπου σταθερές η καλύτερη καμπύλη παραβολή και ο τρόπος υπολογισμού της θα παρουσιασθεί στην παράγραφο 4.3

Γ) όταν οι τιμές της y_i σχηματίζουν περίπου γεωμετρική πρόοδο[5] η εξάρτηση είναι εκθετική και θα αναλυθεί στην παράγραφο 4.4

Δ) και τέλος, όταν ο λόγος $\frac{x_i}{y_i}$ ακολουθεί περίπου αριθμητική πρόοδο, η εξάρτηση είναι υπερβολική και θα παρουσιασθεί στη παράγραφο 4.5.

Η επιλογή της κατάλληλης μορφής καμπύλης είναι σημαντική διότι, όταν την προσδιορίσουμε μπορούμε να προχωρήσουμε σε εκτίμηση των τιμών της

[4] Μια σειρά αριθμών ακολουθεί αριθμητική πρόοδο όταν καθένας προκύπτει από τον προηγούμενο με πρόσθεση σταθερού αριθμού

[5] μια σειρά αριθμών ακολουθεί γεωμετρική πρόοδο όταν κάθε αριθμός προκύπτει από τον προηγούμενο μετά από πολλαπλασιασμό με σταθερό αριθμό.

εξαρτημένης μεταβλητής για τιμές της ανεξάρτητης που δεν ανήκουν στο δείγμα. Άρα, οφείλουμε να είμαστε προσεκτικοί στην χρήση των παραπάνω εργαλείων ώστε να επιτύχουμε την καλύτερη προσέγγιση του προβλήματος μέσω των μετρήσεων δειγματοληψίας.

3.3 ΜΕΘΟΔΟΣ ΕΛΑΧΙΣΤΩΝ ΤΕΤΡΑΓΩΝΩΝ ΚΑΙ ΒΕΛΤΙΣΤΗ ΠΑΡΑΒΟΛΗ

Όπως και στην περίπτωση της γραμμικής συσχέτισης ο προσδιορισμός της βέλτιστης παραβολής προκύπτει από την ελαχιστοποίηση του σφάλματος. Η γενική μορφή της σχέσης εξάρτησης των μεταβλητών X και Y για το δεύτερο σχήμα είναι γνωστή και ίση με :

$$y_i = a + bx_i + cx_i^2 \qquad (3.4)$$

Η εφαρμογή της μεθόδου ελαχίστων τετραγώνων καθιστά δυνατό τον προσδιορισμό των παραμέτρων a, b, c της παραπάνω καμπύλης. Τονίζουμε για άλλη μια φορά ότι η καμπύλη είναι προσέγγιση της πραγματικής που ισχύει για ολόκληρο τον πληθυσμό στον οποίο μελετάμε το συγκεκριμένο πρόβλημα.

Ο προσδιορισμός των παραμέτρων a,b,c προκύπτει από την λύση του συστήματος (Κιόχος, 1990, Spiegel, 1977)

$$\sum_{i=1}^{n} y_i = na + b\sum_{i=1}^{n} x_i + c\sum_{i=1}^{n} x_i^2 \qquad (3.5)$$

$$\sum_{i=1}^{n} x_i y_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^2 + c\sum_{i=1}^{n} x_1^3 \qquad (3.6)$$

$$\sum_{i=1}^{n} x_i^2 y_i = a\sum_{i=1}^{n} x_i^2 + b\sum_{i=1}^{n} x_i^3 + c\sum_{i=1}^{n} x_i^4 \qquad (3.7)$$

Οι παράμετροι $d, b, kai\ c$ είναι:

$$d = \frac{\begin{vmatrix} \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} y_i^2 \\ \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} y_i^3 \\ \sum_{i=1}^{n} x_i y_i^2 & \sum_{i=1}^{n} y_i^3 & \sum_{i=1}^{n} y_i^4 \end{vmatrix}}{D} \qquad (3.8)$$

$$b = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} x_i & \sum_{i=1}^{n} y_i^2 \\ \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i y_i & \sum_{i=1}^{n} y_i^3 \\ \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} x_i y_i^2 & \sum_{i=1}^{n} y_i^4 \end{vmatrix}}{D} \qquad (3.9)$$

$$c = \frac{\begin{vmatrix} n & \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} x_i \\ \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} x_i y_i \\ \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} y_i^3 & \sum_{i=1}^{n} x_i y_i^2 \end{vmatrix}}{D} \qquad (3.10)$$

με D

$$D = \begin{vmatrix} n & \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} y_i^2 \\ \sum_{i=1}^{n} y_i & \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} y_i^3 \\ \sum_{i=1}^{n} y_i^2 & \sum_{i=1}^{n} y_i^3 & \sum_{i=1}^{n} y_i^4 \end{vmatrix} \qquad (3.11)$$

Η παραβολή που προκύπτει μετά τον υπολογισμό των παραμέτρων είναι αυτή που πλησιάζει περισσότερο στο μεγαλύτερο ποσοστό των μετρήσεων όπως απεικονίζονται στο διάγραμμα διασποράς.

3.4 ΕΚΘΕΤΙΚΗ ΕΞΑΡΤΗΣΗ

Αν η εξάρτηση των δύο μεταβλητών είναι εκθετικής μορφής δηλαδή

$$y_i = ab^{x_i} \qquad (3.8)$$

εφαρμόζοντας λογαριθμική μετατροπή και στα δύο μέλη έχουμε, μετά από λίγες πράξεις εφαρμόζοντας γνωστές ιδιότητες των λογαρίθμων:

$$\log(a \cdot b^x) =$$

$$\log a + \log b^x \qquad (3.9)$$

$$\log y_i = \log a + x_i \log b$$

(Οι ιδιότητες που χρησιμοποιήσαμε είναι

1. $\log(a \cdot b) = \log a + \log b$
2. $\log a^k = k \log a$)

Από την παραπάνω σχέση παρατηρούμε ότι ο λογάριθμος της εξαρτημένης μεταβλητής παρουσιάζει γραμμική εξάρτηση με την ανεξάρτητη μεταβλητή.

Σύμφωνα με την θεωρία ελαχίστων τετραγώνων όπως αναλύθηκε στο Κεφάλαιο 2 οι εξισώσεις προσδιορισμού των παραμέτρων a,b είναι

$$\sum_{i=1}^{n} \log y_i = n \log a + \log b \sum_{i=1}^{n} x_i \qquad (3.10)$$

$$\sum_{i=1}^{n} x_i \log y_i = \log a \sum_{i=1}^{n} x_i + \log b \sum_{i=1}^{n} x_i^2 \qquad (3.11)$$

Εργαζόμενοι όπως στο κεφάλαιο 2 βρίσκουμε τις ακόλουθες λύσεις για τα a, b,

$$\log a = \frac{\sum_{i=1}^{n} x_i^{\,2} \sum_{i=1}^{n} \log y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i \log y}{n \sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i \right)^2} \qquad (3.12)$$

$$\log b = \frac{n\sum_{i=1}^{n} x_i \log y - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} \log y_i}{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2} \qquad (3.13)$$

Τυπικό παράδειγμα εκθετικής εξάρτησης είναι τα κέρδη μιας επένδυσης με τη χρονική διάρκεια αυτής.

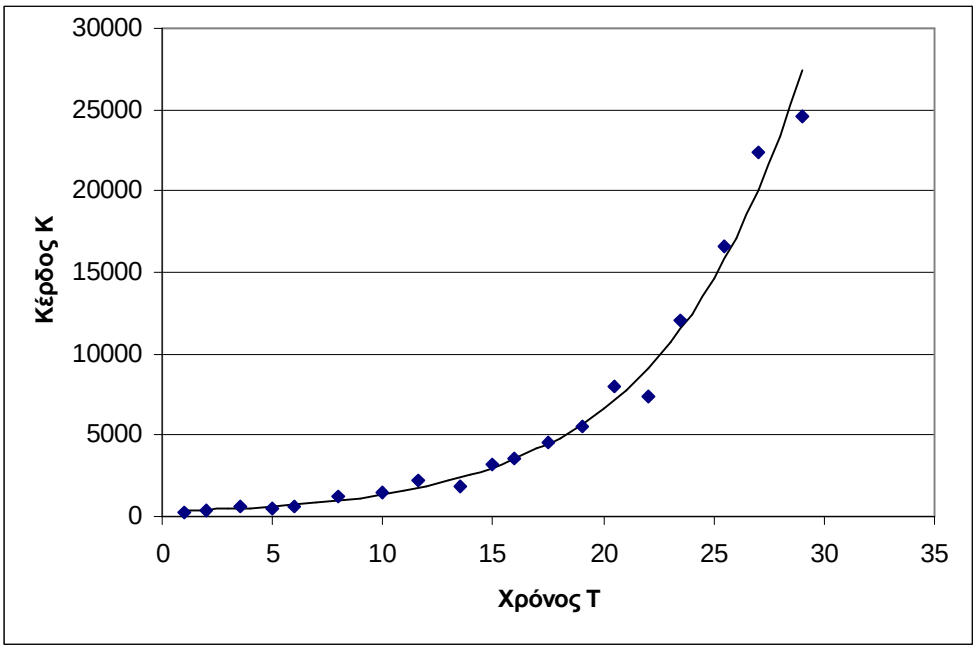
Παράδειγμα

Στον πίνακα 3.1 δίνονται τα κέρδη Κ αρχικής επένδυσης 1000 ευρώ σε συνάρτηση με το χρόνο Τ.

Πίνακας 3.1 Δεδομένα για τον υπολογισμό της εκθετικής εξάρτησης μεταξύ χρόνου επένδυσης και κερδών

i	Έτη t_i	Κέρδος k_i	log (k_i)	t_i log k_i	t_i^2
1	1	263	2,419956	2,419956	1
2	2	390	2,591065	5,182129	4
3	3,5	603	2,780317	9,731111	12,25
4	5	537	2,729974	13,64987	25
5	6	669	2,825426	16,95256	36
6	8	1208	3,082067	24,65654	64
7	10	1466	3,166134	31,66134	100
8	13,5	1905	3,279895	44,27858	182,25
9	15	3180	3,502427	52,53641	225
10	11,6	2200	3,342423	38,7721	134,56
11	16	3623	3,559068	56,94509	256
12	17,5	4575	3,660391	64,05684	306,25
13	19	5495	3,739968	71,05939	361
14	20,5	7944	3,900039	79,9508	420,25
15	22	7423	3,870579	85,15275	484
16	23,5	12022	4,079977	95,87945	552,25
17	25,5	16555	4,218929	107,5827	650,25
18	27	22380	4,34986	117,4462	729
19	29	24570	4,390405	127,3217	841
άθροισμα	275,6	117008	65,4889	1045,236	5384,06

Το διάγραμμα διασποράς και η βέλτιστη καμπύλη δίνονται στο παρακάτω σχήμα



Για τον προσδιορισμό της βέλτιστης καμπύλης χρησιμοποιήθηκαν οι υπολογισμοί με βάση τα δεδομένα του πίνακα.

$$\sum_{i=1} t_i = 275,6$$

$$\sum_{i=1}^{19} \log k_i = 65,4889$$

$$\sum_{i=1}^{19} t_i \log k_i = 1045,236$$

$$\sum_{i=1}^{19} t^2{}_i = 5384,06$$

n = 19

Άρα

$$\log a = = \frac{5384,06 \cdot 65,4889 - 275,6 \cdot 1045,236}{19 \cdot 5384,06 - (275,6)^2} = 2,45$$

$$\log b = = \frac{19 \cdot 1045,236 - 275,6 \cdot 65,4889}{19 \cdot 5384,06 - (275,6)^2} = 0,06874$$

Απ’ όπου βρίσκουμε

a = 281,8

b = 1,171

και η εξίσωση παλινδρόμησης είναι

$$y = 281{,}8 \cdot 1{,}171^{x}$$

Οι τιμές χρησιμοποιήθηκαν για την κατασκευή του ακόλουθου διαγράμματος διασποράς το οποίο και δείχνει την εκθετική εξάρτηση

3.5 ΥΠΕΡΒΟΛΙΚΗ ΕΞΑΡΤΗΣΗ

Οι μεταβλητές X, Y παρουσιάζουν υπερβολική συσχέτιση όταν συνδέονται με μία σχέση της μορφής

$$\frac{1}{Y} = a + bX \qquad (3.12)$$

Ας θεωρήσουμε, λοιπόν το μαθηματικό μοντέλο το οποίο προσεγγίζει κατά βέλτιστο βαθμό τις μετρήσεις δειγματοληψίας. Έχει όμοια μαθηματική μορφή με την εξίσωση (3.12) αλλά διαφορετική τιμή για τις παραμέτρους α,β. Οι παράμετροι a ,b του στατιστικής καμπύλης προκύπτουν θέτοντας

$$z_i = \frac{1}{y_i}$$

και επιλύνοντας το παρακάτω σύστημα:

$$\sum_{i=1}^{n} z_i = na + b\sum_{i=1}^{n} x_i \qquad (3.13)$$

$$\sum_{i=1}^{n} x_i z_i = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} x_i^{\,2} \qquad (3.14)$$

Παράδειγμα (Κιόχος 1990)

Δίνονται τα δεδομένα του παρακάτω πίνακα. Να υπολογισθούν οι συντελεστές a,b της βέλτιστης υπερβολικής καμπύλης.

	x_i	y_i	$z_i=1/y_i$	$z_i x_i$	x_i^2
	0	½	2	0	0
	1	1/5	5	5	1
	2	1/8	8	16	4
	4	1/14	14	56	16
σύνολο	7		29	77	21

Με τον μετασχηματισμό $z_i = \frac{1}{y_i}$ η καμπύλη $\frac{1}{y_i} = a + bx_i$ μετατρέπεται στην $z_i = a + bx_i$. Από την μέθοδο ελαχίστων προκύπτει το παρακάτω σύστημα υπολογισμού των συντελεστών a,b:

$$\sum_{i=1}^{5} z_i = na + b\sum_{i=1}^{n} x_i$$

$$\sum_{i=1}^{5} x_i z_i = a\sum_{i=1}^{5} x_i + b\sum_{i=1}^{5} x_i^{\,2}$$

$\Rightarrow$

$$\begin{aligned} 29 &= 4a + 7b \\ 77 &= 7a + 21b \end{aligned} \quad \Rightarrow$$

a=2 και b=3

Οπότε η βέλτιστη υπερβολή είναι η

$$y = \frac{1}{2 + 3x}$$

3.6 ΠΟΛΛΑΠΛΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Μέχρι τώρα ασχοληθήκαμε με την εξάρτηση μιας μεταβλητής από μία και μόνο μια άλλη. Στις περισσότερες περιπτώσεις όμως ένα οικονομικό μέγεθος εξαρτάται από περισσότερες από μια παραμέτρους. Άρα η εξαρτημένη μεταβλητή είναι συνάρτηση περισσότερων ανεξάρτητων. Σε αυτή την περίπτωση, απαιτείται να προσδιορισθεί και ο βαθμός επίδρασης καθεμιάς από τις ανεξάρτητες στην εξαρτημένη μεταβλητή. Ουσιαστικά, η επίδραση γίνεται φανερή από τον συντελεστή της κάθε μεταβλητής στη μαθηματική σχέση που

προσδιορίζουμε. Η θεμελιώδης αρχή της ελαχιστοποίησης των τετραγωνικών σφαλμάτων ισχύει και στην περίπτωση αυτή.

Έστω λοιπόν μια μεταβλητή Υ που εξαρτάται από n ανεξάρτητες μεταβλητές x_1, x_2, …..x_n. Η σχέση που συνδέει την εξαρτημένη μεταβλητή με περισσότερες από μία ανεξάρτητες ονομάζεται *πολλαπλή παλινδρόμηση*. Η συσχέτιση μεταξύ δύο μεταβλητών γίνεται φανερή με την κατασκευή του διαγράμματος διασποράς, αφού οι δύο μεταβλητές μπορούν να παρουσιασθούν στις δύο διαστάσεις του χαρτιού. Αν οι δύο μεταβλητές συσχετίζονταν ήταν δυνατό να εκτιμηθεί η καμπύλη συσχετίσεως τους εφαρμόζοντας τη μέθοδο ελαχίστων τετραγώνων. Στην περίπτωση της πολλαπλής παλινδρόμησης προκύπτει μια επιφάνεια ή υπερεπιφάνεια. Το αντίστοιχο διάγραμμα διασποράς πρέπει να απεικονισθεί σε χώρο αντίστοιχο διαστάσεων. Κάτι τέτοιο είναι δυνατό μόνο στην περίπτωση εξάρτησης από δύο μεταβλητές οπότε και έχουμε αναπαράσταση στον τρισδιάστατο χώρο.

Ας δούμε, λοιπόν, την απλούστερη περίπτωση όπου η εξαρτημένη μεταβλητή είναι γραμμική συνάρτηση δύο άλλων, όπως για παράδειγμα οι δαπάνες μια οικογένειας ανάλογα με το μέγεθος και το ετήσιο εισόδημα (Μπένος,1997) . Δηλ.

$$Y=\alpha+\beta x_1+\gamma x_2 +\varepsilon \qquad (3.15)$$

Από τη μέθοδο ελαχίστων τετραγώνων γνωρίζουμε ότι το τετραγωνικό σφάλμα πρέπει να ελαχιστοποιείται οπότε προκύπτουν οι εξισώσεις των παραμέτρων a, b, c της επιφάνειας που προσεγγίζει καλύτερα τις δειγματοληπτικές μετρήσεις. Η αύξηση του αριθμού των παραμέτρων που πρέπει να υπολογισθούν οδηγεί σε αντίστοιχη αύξηση και του αριθμού των εξισώσεων.

Οι εξισώσεις είναι :

$$\sum_{i=1}^{n} y_i - a - bx_{1i} - cx_{2i} = 0$$

$$\sum_{i=1}^{n} (y_i - a - bx_{1i} - cx_{2i})(-x_{1i}) = 0$$

$$\sum (y_i - a - bx_{1i} - cx_{2i})(-x_{2i}) = 0$$

Η λύση των εξισώσεων επιτυγχάνεται κυρίως με την χρήση υπολογιστή εξαιτίας της πολυπλοκότητας τους οπότε παρουσιάζεται αυξημένη πιθανότητα αριθμητικού λάθους.

3.7 ΜΕΤΑΣΧΗΜΑΤΙΣΜΟΣ ΜΗ ΓΡΑΜΜΙΚΟΥ ΥΠΟΔΕΙΓΜΑΤΟΣ

Ας υποθέσουμε μια συνάρτηση παραγωγής της μορφής Gobb – Douglas $Q = AL^{b1}K^{b2}u$ όπου υποθέτουμε ότι E(u) = 1

Λογαριθμίζοντας το παραπάνω υπόδειγμα και χρησιμοποιώντας τις ιδιότητες των λογαρίθμων λαμβάνουμε

$\log \boldsymbol{y} = \log \boldsymbol{A} + \boldsymbol{b}1 \log \boldsymbol{L} + \boldsymbol{b}2 \log \boldsymbol{K} + \log \boldsymbol{u}$ και θέτοντας logQ=Y , logA=b ,logL=X1, logK=X2, logu= e λαμβάνουμε την $Y= b+ b_1X1+b_2X2+ e$ (1)

Στην παραπάνω εξίσωση χρησιμοποιούμε την μέθοδο ελαχίστων τετραγώνων για να λάβουμε τις εκτιμήτριες. Θα πρέπει επίσης E(ε) =0

Δηλ E(logu)= $\log \sqrt[n]{\boldsymbol{u}_2} \boldsymbol{u}_2 ... \boldsymbol{u}_n = 0$

Το πρόβλημα είναι ότι ο λογάριθμος του γεωμετρικού μέσου E(ε) είναι πάντα μικρότερος του λογαρίθμου του αριθμητικού μέσου.

Κάνοντας τις πράξεις με βάση το παραπάνω παράδειγμα βρίσκουμε ότι ο λογάριθμος του αριθμητικού μέσου ισούται με το μηδέν. Άρα το E(ε) (λογάριθμος του γεωμετρικού μέσου) θα είναι μικρότερο του μηδενός συνεπώς δεν θα ισχύει μια βασική υπόθεση του κλασικού υποδείγματος. Για να αντιμετωπίσουμε αυτό το πρόβλημα μετασχηματίζουμε εκ νέου την (1) προσθαφαιρώντας την ποσότητα E(ε). Σε αυτή την περίπτωση η εξίσωση θα γίνει

$Y= b +E(\varepsilon) +b_1X1+b_2X2+ e –E(\varepsilon)$

Θέτωντας $b +E(\varepsilon) = b^*$ και $e –E(\varepsilon) =\varepsilon^*$ η εξίσωση μετατρέπεται σε

$Y= b^* +b_1X1+b_2X2+ \varepsilon^*$ όπου $E(\varepsilon^*) = 0$ και έτσι εκτιμάμε με OLS

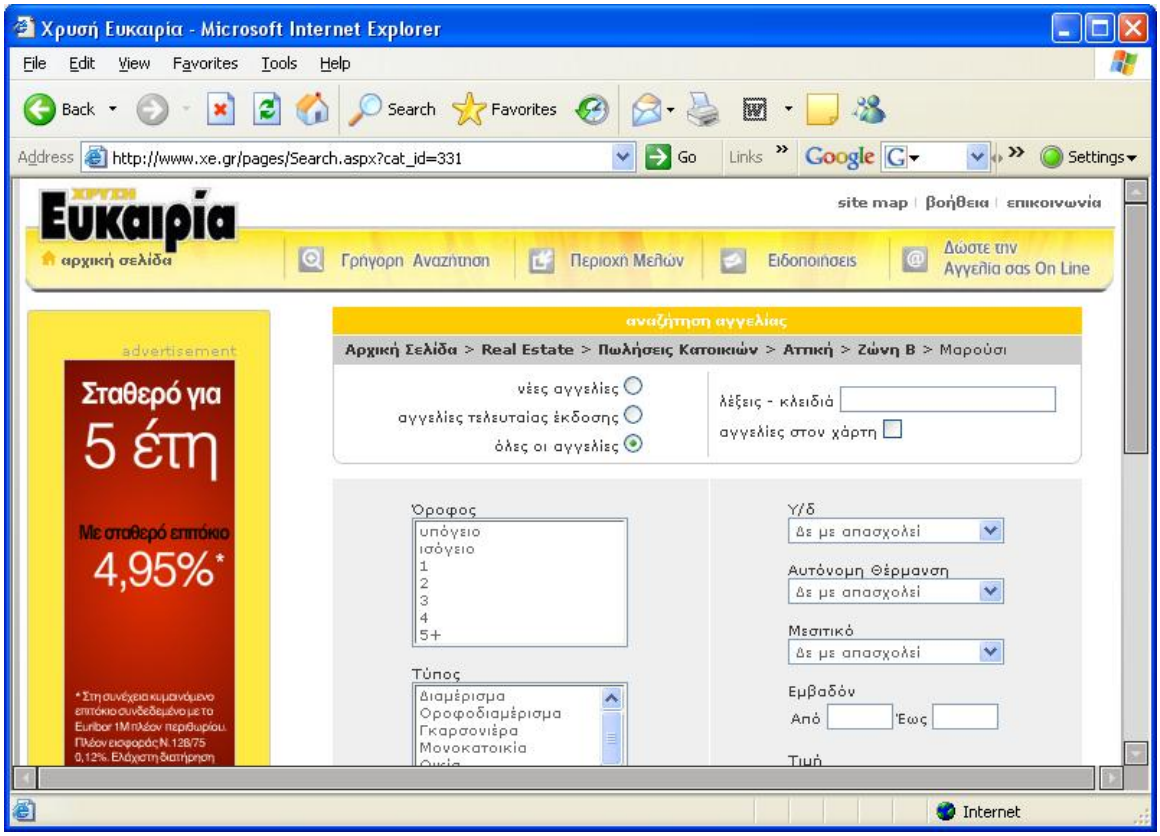
ΚΕΦΑΛΑΙΟ 4ο
ΕΦΑΡΜΟΓΗ ΤΗΣ
ΜΕΘΟΔΟΥ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

4.1 ΤΕΛΙΚΗ ΕΦΑΡΜΟΓΗ

Στο κεφάλαιο αυτό θα εφαρμόσουμε τη μέθοδο της παλινδρόμησης για να βρούμε τον τρόπο συσχέτισης της τιμής πώλησης μιας κατοικίας από το μέγεθός της σε τετραγωνικά μέτρα.

Τα στατιστικά δεδομένα τα αντλήσαμε από την ηλεκτρονική έκδοση της εφημερίδας «Χρυσή Ευκαιρία», Στην ιστοσελίδα της συγκεκριμένης εφημερίδας (www.xe.gr) είναι δυνατή η αναζήτηση αγγελιών πώλησης κατοικιών σε διάφορα μέρη της Ελλάδας. Εμείς αναζητήσαμε αγγελίες από τον Δήμο Αμαρουσίου Αττικής προκειμένου να περιορίσουμε κάπως το μέγεθος του δείγματος. Από τις αγγελίες που προέκυψαν ξεχωρίσαμε αυτές (που είναι και οι περισσότερες) στις οποίες αναφέρονται τα τετραγωνικά μέτρα της κατοικίας καθώς και η τιμή πώλησης της κατοικίας. Καταβλήθηκε προσπάθεια ώστε να αποφευχθούν τυχόν επαναλήψεις τις ίδιας αγγελίας. Δεν έγινε καμία διάκριση ως προς τον τύπο της κατοικίας που αφορούσε η αγγελία, π.χ. αν επρόκειτο για μονοκατοικία ή διαμέρισμα ή μεζονέτα, αν περιλαμβάνονται βοηθητικοί χώροι, όπως αποθήκη ή πάρκιν κλπ.

Προέκυψαν έτσι, 25 αγγελίες που αφορούν πωλήσεις κατοικιών στο Μαρούσι Αττικής. Το εμβαδόν αυτών κυμαίνεται από 28 έως 201 τ.μ. και η τιμή πώλησης από 60.000 έως 400 χιλιάδες ευρώ.



Εικόνα 4.1. Η ιστοσελίδα αναζήτησης αγγελιών της ηλεκτρονικής έκδοσης της εφημερίδας «Χρυσή Ευκαιρία».

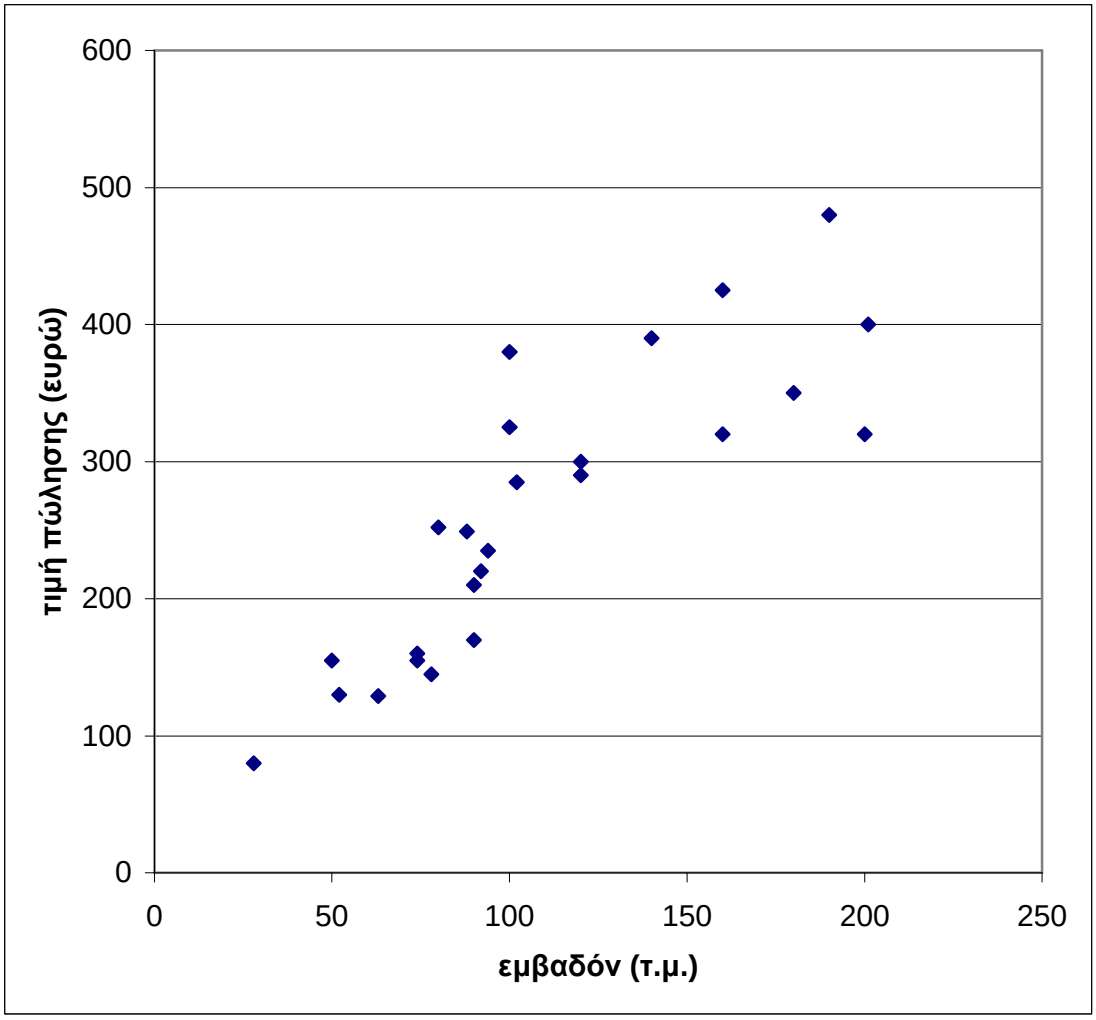
Ως ανεξάρτητη μεταβλητή θεωρούμε το εμβαδόν και ως εξαρτημένη την τιμή πώλησης.

Τα δεδομένα δειγματοληψίας δίνονται στον παρακάτω πίνακα:

Πίνακας 4.1 Δεδομένα Δειγματοληψίας

i	επιφάνεια (τ.μ.)	τιμή πώλησης (χ.ευρώ)
1	28	80
2	50	155
3	52	130
4	63	129
5	74	155
6	74	160
7	78	145
8	80	252
9	88	249
10	90	170
11	90	210
12	92	220
13	94	235
14	100	325
15	100	380
16	102	285
17	120	290
18	120	300
19	140	390
20	160	320
21	160	425
22	190	480
23	180	350
24	200	320
25	201	400

Το διάγραμμα διασποράς είναι



Για να υπολογίσουμε τον συντελεστή συσχέτισης κατασκευάζουμε τον ακόλουθο πίνακα και χρησιμοποιούμε την σχέση (1.4)

Πίνακας 4.2 Στοιχεία για τον υπολογισμό του συντελεστή συσχέτισης μεταξύ εμβαδού και τιμής πώλησης διαμερίσματος.

ι	επιφάνεια (τ.μ.)	τιμή πώλησης (χ.ευρώ)	$(x_i-\bar{x})^2$	$(y_i-\bar{y})^2$	$(x_i-\bar{x})(y_i-\bar{y})$
1	28	80	11664	61504	26784
2	50	155	2500	24025	7750
3	52	130	2704	16900	6760
4	63	129	3969	16641	8127
5	74	155	5476	24025	11470
6	74	160	5476	25600	11840
7	78	145	6084	21025	11310
8	80	252	6400	63504	20160
9	88	249	7744	62001	21912
10	90	170	8100	28900	15300
11	90	210	8100	44100	18900
12	92	220	8464	48400	20240
13	94	235	8836	55225	22090
14	100	325	10000	105625	32500
15	100	380	10000	144400	38000
16	102	285	10404	81225	29070
17	120	290	14400	84100	34800
18	120	300	14400	90000	36000
19	140	390	19600	152100	54600
20	160	320	25600	102400	51200
21	160	425	25600	180625	68000
22	190	480	36100	230400	91200
23	180	350	32400	122500	63000
24	200	320	40000	102400	64000
25	201	400	40401	160000	80400
άθροισμα	2726	6555	364422	2047625	845413
μέση τιμή	136	328			

Οπότε ο συντελεστής r είναι

$$r = \frac{845413}{\sqrt{364422 \cdot 2047625}} = 0{,}97 = 97\%$$

Άρα έχουμε πολύ ισχυρή συσχέτιση

Στον πίνακα 4.3 είναι οι απαραίτητοι υπολογισμοί για τον προσδιορισμό των παραμέτρων $\hat{a}\ kai\ \hat{b}$ της ευθείας παλινδρόμησης.

Πίνακας 4.3 Δεδομένα για τον υπολογισμό της ευθείας παλινδρόμησης μεταξύ εμβαδού και τιμής πώλησης διαμερίσματος

ι	επιφάνεια (X) (τ.μ.)	τιμή πώλησης (χ.ευρώ) (Y)	$x_i y_i$	x_i^2
1	28	80	2240	784
2	50	155	7750	2500
3	52	130	6760	2704
4	63	129	8127	3969
5	74	155	11470	5476
6	74	160	11840	5476
7	78	145	11310	6084
8	80	252	20160	6400
9	88	249	21912	7744
10	90	170	15300	8100
11	90	210	18900	8100
12	92	220	20240	8464
13	94	235	22090	8836
14	100	325	32500	10000
15	100	380	38000	10000
16	102	285	29070	10404
17	120	290	34800	14400
18	120	300	36000	14400
19	140	390	54600	19600
20	160	320	51200	25600
21	160	425	68000	25600
22	190	480	91200	36100
23	180	350	63000	32400
24	200	320	64000	40000
25	201	400	80400	40401
άθροισμα	2726	6555	820869	353542
μέση τιμή	136	328		

Οπότε προκύπτει $\hat{a} = 2{,}1\ \ kai\ \ \hat{b} = 56{,}8$. Άρα η ευθεία παλινδρόμησης είναι

y=2,1x+56,8 που σημαίνει ότι αν αυξηθεί η επιφάνεια κατά 1 τ.μ τότε η τιμή θα αυξηθεί κατά 2,1 €.

Χρησιμοποιώντας την τελευταία σχέση για ένα διαμέρισμα 100τ.μ. αναμένεται να πληρώσουμε 236,8 δηλαδή περίπου 240 χιλιάδες ευρώ.

Τέλος για τον υπολογισμό του δείκτη R^2 έχουμε

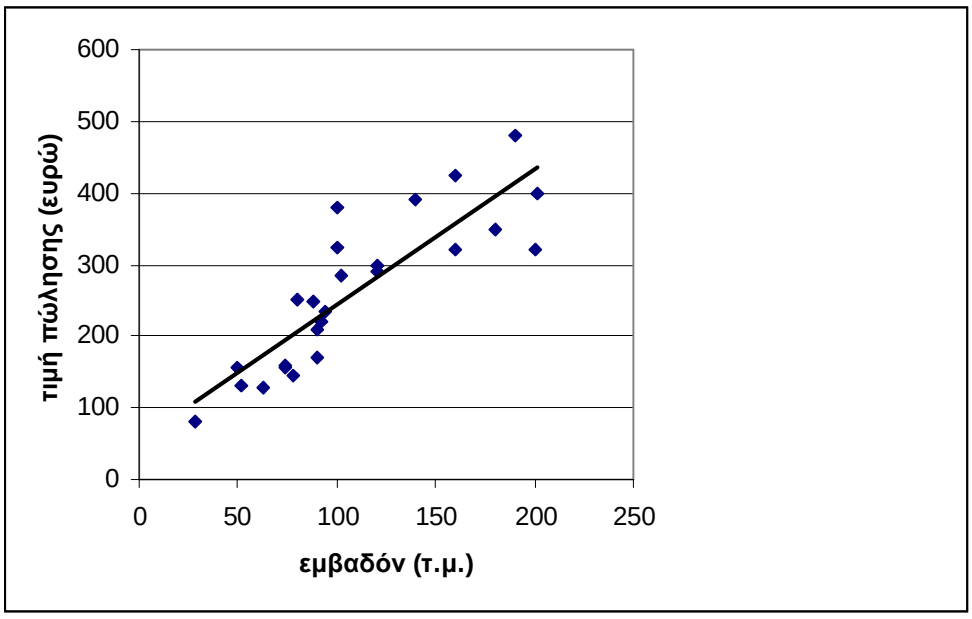
Πίνακας 4.4. Δεδομένα για του συντελεστή R^2 μεταξύ εμβαδού και τιμής πώλησης διαμερίσματος

i	επιφάνεια (τ.μ.)	τιμή πώλησης (χ.ευρώ)	$\hat{y}_i$	$(\hat{y}_i - \bar{y})^2$	$(y_i - \bar{y})^2$
1	28	80	107,2	48752,64	61504
2	50	155	146,8	21550,24	24025
3	52	130	150,4	22620,16	16900
4	63	129	170,2	28968,04	16641
5	74	155	190	36100	24025
6	74	160	190	36100	25600
7	78	145	197,2	38887,84	21025
8	80	252	200,8	40320,64	63504
9	88	249	215,2	46311,04	62001
10	90	170	218,8	47873,44	28900
11	90	210	218,8	47873,44	44100
12	92	220	222,4	49461,76	48400
13	94	235	226	51076	55225
14	100	325	236,8	56074,24	105625
15	100	380	236,8	56074,24	144400
16	102	285	240,4	57792,16	81225
17	120	290	272,8	74419,84	84100
18	120	300	272,8	74419,84	90000
19	140	390	308,8	95357,44	152100
20	160	320	344,8	118887	102400
21	160	425	344,8	118887	180625
22	190	480	398,8	159041,4	230400
23	180	350	380,8	145008,6	122500
24	200	320	416,8	173722,2	102400
25	201	400	418,6	175226	160000
άθροισμα	2726	6555	6326,8	1820805	2047625

μέση τιμή	136	328			

και ο συντελεστής προκύπτει R^2=0,88 δηλαδή 88%. Άρα η τιμή του διαμερίσματος εξαρτάται κατά 88% από το εμβαδόν του.

Η ευθεία παλινδρόμησης δίνεται στο ακόλουθο διάγραμμα



Πρόβλεψη μεμονωμένης τιμής Y

Έστω ότι να βρούμε διάστημα εμπιστοσύνης για το ποια θα είναι κατά μέσο όρο η τιμή για επιφάνεια 50 τ.μ

Το διάστημα εμπιστοσύνης είναι

$$\hat{Y}_0 \pm t, n-2, 0,05 s.e \sqrt{1/n + (X_0 - \bar{X})^2 / \Sigma(X_i - \bar{X})^2}$$

161 ± 2,39*59*0,31.

Τελικά για X=50 Δ.Ε= (117 204)

ΚΕΦΑΛΑΙΟ 5ο
ΓΕΝΙΚΑ ΣΥΜΠΕΡΑΣΜΑΤΑ

Στην εργασία αυτή εξετάσαμε τον τρόπο εξάρτησης μιας μεταβλητής από μια άλλη, κυρίως σε ότι αφορά την γραμμική εξάρτηση είτε θετική είτε αρνητική. Ιδιαίτερα χρήσιμα εργαλεία, για τον σκοπό αυτό αποτελούν το διάγραμμα διασποράς, ο συντελεστής συσχέτισης, ο προσδιορισμός της ευθείας παλινδρόμησης και ο δείκτης προσδιορισμού.

Το διάγραμμα διασποράς αποτελεί την οπτική απεικόνιση των μεταβλητών σε ένα σύστημα αξόνων. Η μορφή μας καθοδηγεί την απόφαση μας για το ποια μαθηματική συνάρτηση θα διαλέξουμε ως πιθανή για να περιγράφει την εξάρτηση των μεταβλητών μας.

Ο συντελεστής συσχέτισης και η ευθεία παλινδρόμησης περιορίζονται σε γραμμικές εξαρτήσεις (και σε όσες μπορούν με κατάλληλους μετασχηματισμούς να μετατραπούν σε γραμμικές όπως π.χ. η εκθετική). Ενώ ο συντελεστής συσχέτισης μας δείχνει το αν η μεταβλητές Y και X κινούνται η όχι προς την ίδια κατεύθυνση, η εξίσωση παλινρόμησης μας δείχνει πέρα από τον βαθμό συσχέτισης μεταξύ Y και X και την μαθηματική σχέση που συνδέει τα X με τα Y.

Μέσω της διαδικασίας ελαχίστων τετραγώνων είμαστε σε θέση όχι μόνο να πραγματοποιήσουμε σημειακή εκτίμηση των παραμέτρων του πληθυσμού αλλά και να βρούμε διαστήματα εμπιστοσύνης ή να διεξάγουμε έλεγχο υποθέσεων για τις παραμέτρους του πληθυσμού έστω και με κάποιο σφάλμα. Είναι φανερό ότι οι εκτιμήτριες που θα λαμβάνουμε κάθε φορά θα είναι διαφορετικές ανάλογα με το δείγμα, συνεπώς σε κάθε δειγματοληψία θα είναι διαφορετικές οι εξισώσεις παλινδρόμησης. Λόγω της τυχαιότητας του σφάλμα τόσο οι

συντελεστές όσο και η εξαρτημένη μεταβλητή είναι τυχαίες μεταβλητές ενώ υποθέτουμε ότι το Χ παραμένει σταθερό από δείγμα σε δείγμα είναι δηλ. μη τυχαία μεταβλητή

Όπως αναφέραμε στην εργασία η υπόθεση περί κανονικής κατανομής του σφάλματος διευκολύνει αφάνταστα τη χρήση των συγκεκριμένων στατιστικών τεχνικών. Η υπόθεση αυτή όπως και πολλές άλλες που αναφέρουμε μπορεί όμως να μην ισχύουν. Σε αυτή την περίπτωση για κάθε υπόθεση που αναιρούμε εισάγουμε ποιο περίπλοκες στατιστικές τεχνικές έτσι ώστε να απεικονιστεί η πραγματικότητα με τον καλύτερο δυνατό τρόπο και να μεγιστοποιηθεί η ακρίβεια των αποτελεσμάτων . Η περιπλοκότητα όμως του οικονομικού περιβάλλοντος είναι όμως τόσο μεγάλη που ακόμα και η χρήση συνθέτων στατιστικών τεχνικών μπορεί να μην επαρκεί.

Τέλος, εξετάσαμε και τον συντελεστή προσδιορισμού ο οποίος μας δείχνει το ποσοστό εξάρτησης της εξαρτημένης μεταβλητής από την ανεξάρτητη η αλλιώς κατά πόσο η εξαρτημένη μεταβλητή ερμηνεύεται από το χ και κατά πόσο από τα σφάλματα. Ένα υψηλό R^2 (measure of fitness) μας δείχνει ότι η προσαρμογή της εκτιμώμενης ευθείας στις παρατηρήσεις είναι ικανοποιητική. Παρόλα αυτά μια υψηλή τιμή δεν είναι ικανοποιητική αφού μπορεί να οφείλεται στην εισαγωγή ανεξάρτητων μεταβλητών οι οποίες όμως μπορεί να μην έχουν οικονομική σημασία. Επιπρόσθετα όταν εισάγουμε νέες ερμηνευτικές μεταβλητές αυξάνεται η διακύμανση των εκτιμητριών γεγονός που θα οδηγήσει στην λήψη μεγαλύτερων Δ.Ε για τις παραμέτρους του πληθυσμού κάτι που είναι αρνητικό.

Παρόλο τους περιορισμούς της ανάλυσης μας οι οποίοι απορρέουν από την διατύπωση ισχυρών υποθέσεων δεν πάει να αποτελεί κάτι που φαίνεται και από τα παραδείγματα ένα ικανοποιητικό σημείο εκκίνησης για την διερεύνηση της σχέσης διαφόρων οικονομικών μεταβλητών αλλά και για την διατύπωση κάποιων στοιχειωδών συμπερασμάτων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Κιόχος, Π., Στατιστική , Εκδόσεις Σταμούλη 1990
2. Μπένος, Β. Περιγραφική Στατιστική, Εκδόσεις Σταμούλης, 1997
3. Πανάρετος Ι, Γραμμικά μοντέλα με έμφαση στις εφαρμογές, Εκδόσεις Πανεπιστημίου Αθηνών 2001
4. Σταυρινός Βασίλης, Οικονομετρία
5. Spiegel, M. Πιθανότητες και Στατιστική, ΕΣΠΙ, Αθήνα 1977
6. Strang, G., Linear Algebra and its Applications, Tomson Books, 2006
7. Sykes, A., An introduction to regression analysis, wikipedia 2008, http://en.wikipedia.com
8. Χαλικιάς Γ. Ιωάννης, Στατιστική μέθοδος ανάλυσης για επιχειρηματικές αποφάσεις, εκδόσεις Ροσυλή 2001