ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ
ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΤΗ ΔΙΟΙΚΗΣΗ ΕΠΙΧΕΙΡΗΣΕΩΝ – ΔΙΟΙΚΗΣΗ ΟΛΙΚΗΣ ΠΟΙΟΤΗΤΑΣ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΣΕ ΔΙΑΔΙΚΑΣΙΑ ΠΑΡΑΓΩΓΗΣ

ΜΕΛΕΤΗ ΠΕΡΙΠΤΩΣΗΣ ΒΙΟΜΗΧΑΝΙΑΣ ΠΑΡΑΓΩΓΗΣ ΤΣΙΜΕΝΤΟΥ

ΜΑΓΚΑΝΑΡΗ ΕΙΡΗΝΗ
ΧΗΜΙΚΟΣ ΜΗΧΑΝΙΚΟΣ Ε.Μ.Π.

ΠΕΙΡΑΙΑΣ, ΦΕΒΡΟΥΑΡΙΟΣ 2006

ΕΥΧΑΡΙΣΤΙΕΣ

Η παρούσα Διπλωματική Εργασία εκπονήθηκε στα πλαίσια της επίτευξης των σπουδών μου στο Μεταπτυχιακό Πρόγραμμα στη Διοίκηση Επιχειρήσεων - Διοίκηση Ολικής Ποιότητας του Πανεπιστημίου Πειραιώς.

Με αφορμή την επιτυχή διεκπεραίωσή της, θα ήθελα να εκφράσω τις θερμές μου ευχαριστίες στον επιβλέποντα Αναπληρωτή Καθηγητή Μιχάλη Σφακιανάκη, για τη συνεργασία του και την υποστήριξη της προσπάθειάς μου. Επίσης, τον ευχαριστώ ιδιαίτερα για την εμπιστοσύνη που έδειξε στο πρόσωπό μου, με την πρόταση που μου έκανε για συνεργασία, αλλά και για την ευκαιρία που μου έδωσε να ασχοληθώ με πραγματικά δεδομένα, που αφορούν μια μεγάλη βιομηχανία παραγωγής τσιμέντου.

Θα ήθελα, επίσης, να ευχαριστήσω ιδιαίτερα τον κ. Α. Κατσιάμπουλα, που είναι Προϊστάμενος Ποιοτικού Ελέγχου σε ένα από τα εργοστάσια της συγκεκριμένης τσιμεντοβιομηχανίας. Η συνεργασία του για τη συλλογή των δεδομένων, αλλά και η βοήθειά του σε θέματα επιστήμης και τεχνολογίας του τσιμέντου, ήταν ιδιαίτερα καθοριστικές για την επεξεργασία που πραγματοποιήθηκε στην παρούσα εργασία.

Τέλος, ευχαριστώ πολύ τον κ. Ν. Λίτινα, Υπεύθυνο Διαχείρισης Ποιότητας της τσιμεντοβιομηχανίας, ο οποίος έκανε μερικές πολύ ορθές παρατηρήσεις, που με βοήθησαν πολύ στην ερμηνεία των αποτελεσμάτων της πραγματοποιούμενης ανάλυσης.

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Η ανάγκη για Στατιστική Ανάλυση

Στη βιομηχανία σήμερα δεν υπάρχει έλλειψη της "πληροφορίας". Ανεξάρτητα από το εάν μια διεργασία είναι μικρή, σαφής ή περίπλοκη, σε όλες υπάρχουν άφθονα όργανα μέτρησης. Τα όργανα αυτά μετράνε διάφορες παραμέτρους σχετικά με την παραγωγική διαδικασία, όπως τη θερμοκρασία, τη συγκέντρωση διαφόρων αντιδραστηρίων, το ρυθμό κατανάλωσης, την πίεση. Μερικές από αυτές τις παρατηρήσεις είναι διαθέσιμες σε τακτικά διαστήματα, για παράδειγμα κάθε πέντε λεπτά ή κάθε μισή ώρα, ενώ άλλες είναι υπό διαρκή παρατήρηση. Από την άλλη, μερικές παρατηρήσεις μπορεί να μην είναι εύκολο να μετρηθούν, γι' αυτό και να ελέγχονται σε αραιότερα διαστήματα. Δείγματα από το τελικό προϊόν μπορεί να λαμβάνονται σε τακτά διαστήματα και μετά από ανάλυση δίνουν μετρήσεις για διάφορες παραμέτρους που είναι σημαντικές για τον κατασκευαστή, όπως η επί τοις εκατό απόδοση, η τελική αντοχή του προϊόντος, το χρώμα. Σε πολλά εργοστάσια υπάρχει μεγάλη συσσώρευση δεδομένων παρόμοιων με αυτά που αναφέρθηκαν παραπάνω, που είτε συλλέγονται στα πλαίσια ενός ορθολογικού ελέγχου, είτε χωρίς κάποιο συγκεκριμένο λόγο, αλλά απλώς ως συνήθεια.

Ωστόσο, από τα δεδομένα μιας χημικής βιομηχανίας -αλλά και γενικότερα- μπορούν να εξαχθούν σημαντικά συμπεράσματα. Μια βασική στατιστική ανάλυση που μπορεί να εφαρμοστεί έχει ως σκοπό την εξαγωγή κάποιων κύριων χαρακτηριστικών των εξισώσεων και συσχετίσεων που κρύβονται μέσα στα δεδομένα. Αυτή η στατιστική ανάλυση είναι η παλινδρόμηση (απλή και πολλαπλή).

Σε οποιοδήποτε σύστημα υπάρχουν ποσοτικές μεταβλητές που αλλάζουν, εμφανίζεται το ενδιαφέρον να εξεταστεί η επίδραση που μερικές μεταβλητές ασκούν (ή φαίνεται να ασκούν) σε κάποιες άλλες. Μπορεί όντως να υπάρχει μια απλή λειτουργική σχέση μεταξύ δύο μεταβλητών. Ωστόσο, σε πολλές φυσικές διεργασίες αυτό αποτελεί την εξαίρεση και όχι τον κανόνα. Κι αυτό, επειδή συχνά υπάρχει μια λειτουργική σχέση, η οποία είναι πολύ περίπλοκη για να μπορέσει να περιγραφεί με απλούς όρους. Για παράδειγμα, σε μια τέτοια περίπτωση προσπαθούμε να προσεγγίσουμε αυτή τη σχέση χρησιμοποιώντας μια απλή μαθηματική συνάρτηση, π.χ. την πολυωνυμική, η οποία περιλαμβάνει τις κατάλληλες μεταβλητές και προσεγγίζει την πραγματική συνάρτηση για κάποια συγκεκριμένα διαστήματα τιμών των περιλαμβανομένων μεταβλητών.

Ακόμα και αν δεν υπάρχει κάποια λογική φυσική σχέση μεταξύ των μεταβλητών, μπορούμε να τις συσχετίσουμε με κάποια μορφή μαθηματικής εξίσωσης. Μπορεί η εξίσωση να μην έχει κάποια φυσική σημασία, ωστόσο πιθανόν να είναι πολύτιμη για την πρόβλεψη των τιμών κάποιων μεταβλητών σε σχέση με κάποιες άλλες, ίσως κάτω από κάποιους συγκεκριμένους περιορισμούς και προϋποθέσεις.

<u>Σκοπός Παρούσας Εργασίας</u>

Στην παρούσα εργασία μελετάται η περίπτωση μιας μεγάλης βιομηχανίας παραγωγής τσιμέντου της χώρας μας. Η εταιρεία ενδιαφέρεται για την ποιότητα του τελικού προϊόντος (τσιμέντο) και συγκεκριμένα για τις τελικές αντοχές σε θλίψη που αναπτύσσει το τσιμέντο στις 28 ημέρες. Γι' αυτό, σκοπός αυτής της εργασίας είναι η αρχική επεξεργασία μερικών βασικών χημικών και φυσικοχημικών παραμέτρων που παίζουν σημαντικό ρόλο στην παραγωγική διαδικασία, και η εύρεση πιθανών συσχετίσεων μεταξύ αυτών και των αντοχών του τσιμέντου στις 2 και στις 28 ημέρες. Απώτερος **στόχος** είναι η εύρεση ενός μοντέλου πολλαπλής παλινδρόμησης μεταξύ των αντοχών στις 28 ημέρες και των ανεξάρτητων μεταβλητών, που είναι κάποιες χημικές και φυσικοχημικές παράμετροι. Επίσης, αναζητείται ένα μοντέλο πολλαπλής παλινδρόμησης παρόμοιο με το παραπάνω, με τη διαφορά ότι η εξαρτημένη μεταβλητή είναι οι πρώιμες αντοχές του τσιμέντου στις 2 ημέρες. Τέλος, εφαρμόζονται μοντέλα απλής παλινδρόμησης μεταξύ των αντοχών στις 28 ημέρες και στις 2 ημέρες, καθώς και μεταξύ των αντοχών στις 28 ημέρες και στις 7 ημέρες.

Τα στοιχεία και δεδομένα που χρησιμοποιήθηκαν αφορούν δύο βασικούς τύπους τσιμέντου, το CEM II 42,5 και το OPC. Οι δύο αυτοί τύποι διαφέρουν μεταξύ τους ως προς το ποσοστό περιεχόμενου κλίνκερ και γύψου, την ποζολάνη και τον ασβεστόλιθο. Τα τελευταία δύο συστατικά υπάρχουν μόνο στον τύπο CEM II 42,5. Επίσης, οι δύο τύποι διαφέρουν ως προς τα βελτιωτικά, τα οποία υπάρχουν μόνο στο OPC.

Ο κάθε τύπος τσιμέντου παράγεται κυρίως από δύο διαφορετικούς μύλους. Ο κάθε μύλος, όμως, παρουσιάζει διαφορετική συμπεριφορά, καθώς μπορεί να έχει διαφορετική τεχνολογία και τεχνικά χαρακτηριστικά, άλλες συνθήκες μεταχείρισης, διαφορετικό χρόνο ζωής και απαιτήσεις συντήρησης. Για όλους αυτούς τους λόγους, οι συνθήκες παραγωγής μπορεί να διαφέρουν από μύλο σε μύλο. Γι' αυτό, κρίνεται σκόπιμο η στατιστική ανάλυση που πραγματοποιείται στην παρούσα εργασία να γίνει ξεχωριστά τόσο για κάθε τύπο τσιμέντου, όσο και για κάθε μύλο παραγωγής του ίδιου τύπου τσιμέντου. Αυτό, διότι στις στατιστικές αναλύσεις είναι αναγκαίο οι συνθήκες κάτω από τις οποίες γίνεται η δειγματοληψία (δεδομένα) να είναι όσο το δυνατόν πιο σταθερές.

Η εύρεση ενός μοντέλου παλινδρόμησης αποτελεί στην ουσία την εύρεση μιας μαθηματικής εξίσωσης, που συνδέει την εξαρτημένη μεταβλητή με κάποιες άλλες (ή άλλη) ανεξάρτητη. Η εξίσωση αυτή μπορεί να χρησιμοποιηθεί για **πρόβλεψη** της εξαρτημένης μεταβλητής (αντοχές τσιμέντου), πριν ακόμα παραχθεί το τσιμέντο. Αυτό είναι πολύ σημαντικό για κάθε βιομηχανία, γιατί δίνει τη δυνατότητα ελέγχου της παραγωγής, και συνεπώς επιτρέπει την πραγματοποίηση κάποιων επεμβάσεων για τη διατήρηση της ποιότητας του τελικού προϊόντος.

Ένα σημαντικό συμπέρασμα που προέκυψε στην παρούσα μελέτη είναι ότι το τσιμέντο αποτελεί ένα πολύ πολύπλοκο προϊόν, που εξαρτάται από πολλές παραμέτρους. Στα μοντέλα παλινδρόμησης για τις αντοχές στις 28 ημέρες (ή στις 2 ημέρες) φάνηκε ότι οι περιεκτικότητες σε κάποια βασικά οξείδια και μερικά άλλα φυσικοχημικά χαρακτηριστικά εξηγούν ένα μικρό ποσοστό της μεταβλητότητας των τελικών αντοχών. Αυτό, σε συνδυασμό με την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων, είναι ενδεικτικό για τη μη συμπερίληψη στο μοντέλο κάποιων βασικών

μεταβλητών. Οι μεταβλητές αυτές πολύ πιθανόν να είναι κάποιες λειτουργικές παράμετροι στις διάφορες φάσεις παραγωγής του τσιμέντου, όπως επισημαίνεται και από την επιστήμη και τεχνολογία του τσιμέντου. Επειδή, όμως, η μέτρηση των παραμέτρων αυτών είναι ιδιαίτερα δύσκολη, δεν υπάρχουν διαθέσιμα δεδομένα και μετρήσεις των παραμέτρων αυτών από την τσιμεντοβιομηχανία που μελετάται.

Περιεχόμενα Παρούσας Εργασίας

Η δομή της μελέτης είναι η ακόλουθη:

Κεφάλαιο 1: Στο κεφάλαιο αυτό αναφέρονται κάποια βασικά ιστορικά στοιχεία για το τσιμέντο, καθώς και η βασική ορολογία αυτού. Επίσης, παρουσιάζονται τα βασικά στάδια παραγωγής τσιμέντου, οι βασικοί τύποι και κατηγορίες αυτού και η επίδραση διαφόρων παραμέτρων της διαδικασίας παραγωγής στις τελικές ιδιότητες (αντοχές) του τσιμέντου. Το κεφάλαιο κλείνει με την παράθεση κάποιων βασικών στοιχείων για το προφίλ της προς μελέτη εταιρείας.

Κεφάλαιο 2: Το κεφάλαιο αυτό θέτει τις βασικές έννοιες της στατιστικής, οι οποίες είναι απαραίτητες για την κατανόηση της μετέπειτα ανάλυσης. Έτσι, γίνεται μια εισαγωγή για τις βασικές στατιστικές μεθόδους που υπάρχουν και τους τρόπους παρουσίασης των δεδομένων, ενώ ακολουθεί η παρουσίαση των βασικών χαρακτηριστικών των αριθμητικών δεδομένων (κεντρική τάση, διασπορά, σχήμα). Τέλος, γίνεται αναφορά στην κανονική κατανομή και τις βασικές ιδιότητες αυτής, και τίθενται οι βάσεις για την κατανόηση των ελέγχων υποθέσεων, που διέπουν τις περισσότερες στατιστικές αναλύσεις.

Κεφάλαιο 3: Το κεφάλαιο αυτό ασχολείται με την απλή γραμμική παλινδρόμηση και συσχέτιση μεταξύ διαφόρων μεταβλητών. Παρουσιάζεται ο συντελεστής συσχέτισης και τα διαγράμματα διασποράς, ενώ υποδεικνύεται το υπόδειγμα του μοντέλου απλής γραμμικής παλινδρόμησης και ο τρόπος υπολογισμού αυτού, με τη μέθοδο των ελαχίστων τετραγώνων. Επίσης, προσδιορίζεται ο συντελεστής προσδιορισμού και ο έλεγχος στατιστικής σημαντικότητας τόσο αυτού, όσο και του συντελεστή παλινδρόμησης.

Κεφάλαιο 4: Στο κεφάλαιο αυτό παρουσιάζεται το υπόδειγμα πολλαπλής γραμμικής παλινδρόμησης και η εκτίμηση αυτού, ενώ αναλύεται ο συντελεστής πολλαπλού προσδιορισμού και οι διάφοροι έλεγχοι στατιστικής σημαντικότητας. Τέλος, παρουσιάζονται τα διάφορα προβλήματα των μοντέλων παλινδρόμησης, σχετικά με την επιλογή των κατάλληλων μεταβλητών, την ύπαρξη πολυσυγγραμικότητας και την εξέταση των καταλοίπων.

Κεφάλαιο 5: Στο κεφάλαιο αυτό πραγματοποιείται το πρώτο βήμα στατιστικής ανάλυσης των δεδομένων που χρησιμοποιούνται. Συγκεκριμένα, εξετάζεται η μέτρηση της κεντρικής τάσης, η μέτρηση της διασποράς και ο έλεγχος του σχήματος των μεταβλητών που μελετώνται. Με βάση τα παραπάνω μέτρα, καθώς και με διάφορα διαγράμματα, ελέγχεται η ύπαρξη κανονικότητας, τα διαστήματα εμπιστοσύνης για τις τιμές κάθε μεταβλητής και η ύπαρξη εκτρόπων παρατηρήσεων.

Κεφάλαιο 6: Το κεφάλαιο αυτό αποτελεί το βασικό κεφάλαιο στατιστικής επεξεργασίας και αποτελεσμάτων. Εδώ βρίσκονται οι συσχετίσεις μεταξύ όλων των μεταβλητών και καθορίζονται τα μοντέλα πολλαπλής και απλής παλινδρόμησης για κάθε τύπο τσιμέντου και κάθε μύλο παραγωγής. Επίσης, ελέγχεται η ισχύς όλων των προϋποθέσεων της παλινδρόμησης, για να εκτιμηθεί εάν τα μοντέλα είναι έγκυρα ή αν πρέπει να διατηρήσουμε κάποια επιφυλακτικότητα ως προς αυτά.

Κεφάλαιο 7: Στο κεφάλαιο αυτό παρατίθενται τα βασικά συμπεράσματα και παρατηρήσεις που προκύπτουν από την επεξεργασία των δεδομένων και την ανάλυση παλινδρόμησης, όπως για παράδειγμα για το ποιες μεταβλητές φαίνονται ως οι σημαντικότερες σε κάθε τύπο τσιμέντου. Εδώ προκύπτει και το συμπέρασμα ότι οι ανεξάρτητες μεταβλητές που έχουν επιλεγεί δεν αρκούν για την ερμηνεία των αντοχών του τσιμέντου στις 2 και στις 28 ημέρες. Τέλος, αναφέρονται μερικές προτάσεις για περαιτέρω έρευνα.

ΚΑΤΑΛΟΓΟΣ ΠΙΝΑΚΩΝ

ΚΑΤΑΛΟΓΟΣ ΣΧΗΜΑΤΩΝ

ΚΑΤΑΛΟΓΟΣ ΔΙΑΓΡΑΜΜΑΤΩΝ

CEM II 42,5 – MT1

ΜΕΡΟΣ Α΄ - ΘΕΩΡΗΤΙΚΟ ΜΕΡΟΣ

ΚΕΦΑΛΑΙΟ 1: ΒΑΣΙΚΕΣ ΓΝΩΣΕΙΣ ΣΤΗΝ ΠΑΡΑΓΩΓΗ ΤΣΙΜΕΝΤΟΥ

1.1. Ιστορικά Στοιχεία για το Τσιμέντο

Από τότε που ο άνθρωπος αποφάσισε να αφήσει τα φυσικά σημεία διαβίωσής του είχε το βασικό μέλημα να βρει και να χρησιμοποιήσει κατά το δυνατόν ανθεκτικότερα υλικά, προκειμένου να κατασκευάσει τις κατοικίες που θα έμενε αυτός και η οικογένειά του. Από την αρχή εντόπισε τις προσπάθειές του στο να ανακαλύψει τα κατάλληλα συνδετικά υλικά που θα συνέδεαν τα δομικά στοιχεία, κυρίως λίθους, που χρησιμοποιούσε για την κατοικία του. Ο όρος τσιμέντο (cement) φαίνεται ότι πρωτοχρησιμοποιήθηκε κατά τους Ρωμαϊκούς χρόνους και τις αρχές του Μεσαίωνα. Με τον όρο αυτό χαρακτηρίζονταν υλικά με συνδετικές ιδιότητες και κυρίως κονιάματα ή μίγματα ασβέστου, ποζολάνης, νερού κλπ, που χρησιμοποιούσαν οι τότε κατασκευαστές για να συνδέσουν τους λίθους στις κατασκευές τους. Η κονία που αποκτούσε συνδετικές ιδιότητες κάτω από την επίδραση του νερού ονομαζόταν υδραυλικό τσιμέντο (hydraulic cement). Μερικές φορές στο παρελθόν, χρησιμοποιούνταν οι όροι ***pozzolana cement*** ή ***Santorin cement*** σε ένδειξη της φύσης του δείγματος. Το πρώτο, δηλαδή, αναφερόταν σε μίγμα ποζολάνης και ασβέστου και το δεύτερο ειδικότερα σε ποζολάνη από τη Σαντορίνη (θηραϊκή γη) και άσβεστο.

Στο δεύτερο ήμισυ του 18^{ου} αιώνα υπήρχε έντονη δραστηριότητα στην Ευρώπη για την ανακάλυψη ενός ανόργανου συνδετικού υλικού που θα μπορούσε να πήξει και να σκληρυνθεί με το νερό. Οι πρώτες ανακαλύψεις αφορούσαν κυρίως την άσβεστο. Παρατηρήθηκε ότι κατά το ψήσιμο (ασβεστοποίηση) του καθαρού ασβεστόλιθου προέκυπτε τελικά ένα υλικό, που με την προσθήκη νερού μετασχηματιζόταν σε μια παχύρρευστη μάζα που είχε συνδετικές ιδιότητες. Η άσβεστος αυτή ήταν γνωστή ως "παχιά άσβεστος". Όταν αυξάνονταν οι ποσότητες των προσμίξεων, που ήταν κυρίως αργιλικές ή πυριτικές, τότε το προϊόν της ασβεστοποίησης δεν ενυδατωνόταν εύκολα και η εκλυόμενη θερμότητα ήταν πολύ μικρή. Το 1758 ο Smeaton διαπίστωσε ότι οι άσβεστοι που περιείχαν μεγάλες ποσότητες (20-25%) αργιλικών υλικών, είχαν την ιδιότητα να σκληρύνονται κάτω από το νερό, χαρακτηριστικό που δεν είχε παρατηρηθεί με τις καθαρότερες ασβέστους. Παρατηρήθηκε, μάλιστα, ότι το φαινόμενο αυτό ήταν πιο έντονο όταν χρησιμοποιούταν μια ποζολάνη από μια περιοχή (Pozooli) κοντά στη Ρώμη. Οι καθαρές άσβεστοι παρατηρήθηκε ότι σκληραίνονταν στον αέρα, με την επίδραση του διοξειδίου του άνθρακα, για να ξανασχηματίσουν ανθρακικό ασβέστιο. Οι άσβεστοι που δεν ήταν καθαρές αντιδρούσαν μέσα στο νερό με τις πυριτικές και αργιλικές ενώσεις που περιείχαν ως προσμίξεις, για να σχηματίσουν ένυδρες ασβεσταργιλικές και ασβεστοπυριτικές ενώσεις, που είχαν ισχυρές συνδετικές ικανότητες και δεν χρειάζονταν διοξείδιο του άνθρακα για να σκληρυνθούν. Έτσι, η μεν καθαρή άσβεστος δεν μπορούσε να χρησιμοποιηθεί ως τσιμέντο, ενώ η μη καθαρή μπορούσε.

Για να διαφοροποιήσει ο Smeaton την καθαρή από τη μη καθαρή άσβεστο, χαρακτήρισε την καθαρή ως μη υδραυλική, ενώ τη μη καθαρή ως υδραυλική. Κατά

την έψηση των υδραυλικών ασβέστων, μερικά τεμάχια του ορυκτού που χρησιμοποιούταν ως πρώτη ύλη δέχονταν πολύ υψηλές θερμοκρασίες, με αποτέλεσμα να συντήκονται. Τα τεμάχια αυτά θεωρούνταν άχρηστα και απορρίπτονταν. Παρατηρήθηκε, όμως, ότι το προϊόν της άλεσης αυτών των τεμαχίων ήταν ένα πολύ καλό συνδετικό υλικό που οι Γάλλοι χρησιμοποίησαν με το όνομα "Grappier cement". Αποτέλεσμα των ερευνών του Smeaton ήταν να γίνει γνωστό, ότι εάν προστεθούν στον ασβεστόλιθο, πριν την έψηση, επιπρόσθετες ποσότητες αργιλικών υλικών, οι υδραυλικές ιδιότητες βελτιώνονται. Έτσι επιτεύχθηκε η παρασκευή των "Ceminently hydraulic limes".

Το 1800 ο Parker ανακάλυψε το "Roman cement", με θέρμανση σε θερμοκρασία υαλοποίησης αργίλων που περιείχαν μικρούς κρυστάλλους ασβεστολιθικής ύλης. Το 1822 κατοχύρωσε με δίπλωμα ευρεσιτεχνίας το "British cement", ενώ το 1824 ο Άγγλος J. Aspdin ανακάλυψε ότι ήταν δυνατό να παρασκευασθεί το τσιμέντο με πολύ καλύτερες ιδιότητες από τις υδραυλικές ασβέστους, εάν το μίγμα ασβέστου και αργίλου θερμαινόταν μέχρι επίτηξης. Το προϊόν της επίτηξης κατοχύρωσε με δίπλωμα ευρεσιτεχνίας με το όνομα Portland cement, για διαφοροποίηση από τις διάφορες ασβέστους. Το 1845 ο I.C. Jonson υπέδειξε ακριβέστερες αναλογίες και καταλληλότερες θερμοκρασίες για τις πρώτες ύλες και την έψηση, αντίστοιχα. Είναι ενδιαφέρον να σημειωθεί ότι ο Jonson εργάστηκε στην τσιμεντοβιομηχανία από το 1827 σε ηλικία 16 ετών. Πέθανε σε ηλικία 100 ετών το 1911 και θεωρείται ο πρώτος παρασκευαστής στην ιστορία του τσιμέντου.

Το 1850 ιδρύεται στη Γαλλία το πρώτο εργοστάσιο τσιμέντου, το 1855 στη Γερμανία, το 1875 στην Αμερική και το 1902 στην Ελλάδα. Το 1859 μετρήθηκαν οι πρώτες αντοχές του τσιμέντου από τον J. Grant, ενώ το 1887 ο H. Le Chatelier ανέπτυξε τις πρώτες θεωρίες για την ενυδάτωση του τσιμέντου. Το 1895 ο W. Michaelis εισήγαγε τη δοκιμή της σταθερότητας όγκου, το 1904 έγιναν οι πρώτες προδιαγραφές για το τσιμέντο από την Αμερικανική Ένωση Πολιτικών Μηχανικών και το 1924 ο R. Bogue προσεγγίζει με ικανοποιητική ακρίβεια την ορυκτολογική σύσταση του τσιμέντου.

Αντίστοιχες ημερομηνίες σταθμοί αναφέρονται για τις τεχνολογικές εξελίξεις κατά την παραγωγική διαδικασία του τσιμέντου (μύλοι, συστήματα έψησης κλπ), εξελίξεις που συνεχίζονται μέχρι σήμερα. Βασική ώθηση στη μαζική παραγωγή του τσιμέντου αποτέλεσε η εισαγωγή του περιστροφικού φούρνου το 1877. Στην Ελλάδα οι φούρνοι του τύπου αυτού εισήχθησαν το 1912. [9]

1.2. Ορολογία Τσιμέντου

Η βασική ορολογία του τσιμέντου που βοηθάει στην κατανόηση ορισμένων εννοιών της παρούσας εργασίας είναι η εξής:

Τσιμέντο **(γενικά)** είναι μία λεπτόκοκκη σκόνη με υδραυλικές ιδιότητες. Αποτελείται από οξείδια του ασβεστίου, πυριτίου, αργιλίου και σιδήρου, που είναι ενωμένα μεταξύ τους και αποτελούν το 90% του βάρους του. Το υπόλοιπο μέρος είναι γύψος και μικρές ποσότητες αλάτων μαγνησίου, καλίου, νατρίου και άλλων στοιχείων. Όταν αναμιγνύεται με νερό έχει την ιδιότητα να πήζει και να σκληραίνει, είτε στον αέρα, είτε κάτω από το νερό.

Τσιμέντο Portland είναι το προϊόν που προκύπτει μετά από έψηση σε θερμοκρασία κλινκεροποίησης (1380-1420^{o}C) ενός πλήρως ομογενοποιημένου μίγματος, που αποτελείται από περίπου 75% ασβεστολιθικά υλικά και 25% αργιλοπυριτικά υλικά, και στη συνέχεια συνάλεση του προκύπτοντος προϊόντος (κλίνκερ) με την κατάλληλη ποσότητα γύψου.

Υδραυλικές ιδιότητες είναι οι ιδιότητες που έχουν ορισμένα υλικά, όπως π.χ. το τσιμέντο, να σχηματίζουν κάτω από την επίδραση νερού σταθερές ένυδρες ενώσεις που είναι ελάχιστα υδατοδιαλυτές και έχουν μεγάλη συνάφεια μεταξύ των και με τα αδρανή. Οι ενώσεις αυτές με την πάροδο του χρόνου αυξάνουν τη συνοχή των πολτών και των κονιαμάτων που προέρχονται από αυτές, με αποτέλεσμα την ανάπτυξη αντοχών. Τα υλικά αυτά λέγονται υδραυλικές κονίες, σε αντίθεση με τις αερικές κονίες (π.χ. ο ασβέστης) που αναπτύσσουν τις αντοχές τους εκτιθέμενες στον αέρα.

Κύρια συστατικά (main constituents) είναι ειδικά επιλεγμένα ανόργανα υλικά, που προστίθενται σε κάποια φάση της παραγωγικής διαδικασίας στο τσιμέντο, σε αναλογία που ξεπερνά το 5% κατά βάρος της συνολικής ποσότητας κυρίων και δευτερευόντων συστατικών. Ενδεικτικά αναφέρεται ότι τα κύρια συστατικά του τσιμέντου είναι το κλίνκερ τσιμέντου Portland (αποτελείται κυρίως από οξείδια του ασβεστίου, πυριτίου, αργιλίου, σιδήρου), η σκωρία υψικαμίνων, τα ποζολανικά υλικά, οι ιπτάμενες τέφρες, το burnt shale, ο ασβεστόλιθος και το silica fume.

Δευτερεύοντα συστατικά (minor additional constituents) είναι ειδικά επιλεγμένα ανόργανα υλικά που προστίθενται σε κάποια φάση της παραγωγικής διαδικασίας στο τσιμέντο, σε αναλογία που δεν ξεπερνά το 5% κατά βάρος της συνολικής ποσότητας κύριων και δευτερευόντων συστατικών.

Πρόσθετα (additives) νοούνται συστατικά που προστίθενται σε μικρά ποσοστά (συνήθως μικρότερα από 1%) και τα οποία έχουν σκοπό να βελτιώσουν είτε την παραγωγική διαδικασία είτε τις ιδιότητες του τσιμέντου.

Ποζολάνη είναι ένα πυριτικό ή αργιλικό υλικό, το οποίο αν και μόνο του δεν έχει υδραυλικές ιδιότητες, όταν αλεσθεί και παρουσία νερού αντιδρά με την υδράσβεστο που προκύπτει από τις αντιδράσεις ενυδάτωσης των κύριων συστατικών του τσιμέντου σε συνήθη θερμοκρασία και σχηματίζει ενώσεις που έχουν υδραυλικές ιδιότητες.

Κλινκεροποίηση: περιλαμβάνει το σύνολο των αντιδράσεων που συμβαίνουν στις περιστροφικές καμίνους σε θερμοκρασίες μεγαλύτερες των 1150^{o}C τόσο στην υγρή φάση όσο και κατά την οριακή επίτηξη των κόκκων που οδηγούν τελικά στην παραγωγή του κλίνκερ.

Πήξη (setting) είναι το φαινόμενο κατά το οποίο ο τσιμεντοπολτός παύει να έχει πλαστικές ιδιότητες.

Σκλήρυνση (hardening) είναι η ανάπτυξη των αντοχών που λαμβάνει χώρα μετά την πήξη.

Γύψος είναι το θειικό ασβέστιο που προστίθεται στα άλλα συστατικά του τσιμέντου κατά την τελική άλεση του κλίνκερ, με σκοπό να ρυθμίσει την πήξη του τσιμέντου. Το ακριβές ποσοστό προσθήκης συνήθως κυμαίνεται μεταξύ 4 και 5% του βάρους του κλίνκερ.

1.3. Η Διαδικασία Παραγωγής του Τσιμέντου

Τα κύρια στάδια της διαδικασίας παραγωγής τσιμέντου είναι η προετοιμασία του μίγματος των πρώτων υλών, η έψηση του μίγματος πρώτων υλών και η άλεση του τσιμέντου. Η κατηγοριοποίηση αυτή διακρίνει την παραγωγική διαδικασία στις διεργασίες που προηγούνται της κυρίως διεργασίας μετασχηματισμού των πρώτων υλών σε προϊόν, που είναι η έψηση, και στις διεργασίες που έπονται και που προσδίδουν στο προϊόν επιμέρους ιδιότητες. Όσο και αν τα μηχανήματα με τα οποία γίνεται η έψηση (περιστροφικές κάμινοι) αποτελούν την καρδιά της τσιμεντοβιομηχανίας καθορίζοντας και τη δυναμικότητά της, και τα μηχανήματα των δύο άλλων σταδίων είναι εξίσου σημαντικά για την ορθολογική λειτουργία της βιομηχανικής μονάδας. Πιο αναλυτικά, τα στάδια παραγωγής του τσιμέντου είναι τα εξής:

1ο Στάδιο-Εξόρυξη Πρώτων Υλών: Οι πρώτες ύλες εξορύσσονται με τη χρήση ισχυρών εκσκαπτικών μηχανημάτων ή με τη χρήση εκρηκτικών υλών.

2ο Στάδιο-Θραύση Πρώτων Υλών: Τα υλικά θραύονται σε μεγάλους θραυστήρες σε τεμάχια, συνήθως μικρότερα των 30 χιλιοστών.

3ο Στάδιο-Αποθήκευση και Προομοιογένεια Πρώτων Υλών: Οι θραυσμένες πρώτες ύλες αποθηκεύονται (με σύγχρονη ανάμιξη) χωριστά κατά κατηγορία και από εκεί οδεύουν προς τους μύλους συνάλεσης σε αυστηρά καθορισμένη και συνεχώς ελεγχόμενη δοσολογία.

4ο Στάδιο-Ξήρανση και Άλεση Πρώτων Υλών: Οι μύλοι είναι μεταλλικοί κύλινδροι, με ισχυρή εσωτερική μεταλλική θωράκιση και περιέχουν πολλούς τόνους από σφαιρικά χαλύβδινα αλεστικά σώματα. Κατά την περιστροφική κίνηση των μύλων οι σφαίρες κονιοποιούν τις προθραυσμένες πρώτες ύλες σε κόκκους μέσης διαμέτρου. Το προϊόν αυτό ονομάζεται φαρίνα.

5ο Στάδιο-Ομογενοποίηση και Αποθήκευση Φαρίνας: Η φαρίνα οδηγείται στα ειδικά σιλό όπου συντελείται η ομογενοποίηση.

6ο Στάδιο-Έψηση: Μετά την ομογενοποίηση η φαρίνα περνάει από ένα σύστημα κυκλώνων που ονομάζεται προθερμαντής και υφίσταται μια προοδευτική θερμική κατεργασία μέχρι 900^{o}C. Στη συνέχεια οι περιστροφικοί κλίβανοι αναλαμβάνουν την έψηση. Οι περιστροφικοί κλίβανοι είναι μεταλλικοί κύλινδροι μήκους 50-150 μέτρων και διαμέτρου 3-5 μέτρων, με εσωτερική επένδυση από πυρότουβλα. Η περιστροφική κίνηση του κλίβανου και η κλίση του εξωθούν τη φαρίνα προς την έξοδο. Στην πορεία της συναντάει θερμοκρασίες που φθάνουν τους 1400^{o}C. Μέσα στον κλίβανο χάρη στις φυσικοχημικές της ιδιότητες, η φαρίνα μετατρέπεται σε ένα κοκκώδες προϊόν που λέγεται κλίνκερ.

7ο Στάδιο-Άλεση Τσιμέντου: Το κλίνκερ αποτελεί το βασικό συστατικό του τσιμέντου και από την ποιότητά του εξαρτάται στο μέγιστο βαθμό η ποιότητά του. Το τσιμέντο ως τελικό προϊόν είναι μια πολύ λεπτή σκόνη και για τη δημιουργία του απαιτείται συνάλεση κλίνκερ, γύψου και ορισμένων φυσικών ή τεχνητών υλικών, που προσδίδουν στο τσιμέντο ωφέλιμες ιδιότητες. Τέτοιες ύλες είναι οι ποζολάνες. Οι μύλοι τσιμέντου μοιάζουν με τους μύλους φαρίνας. Οι δοσολογίες των υλικών συνάλεσης είναι αυστηρά καθορισμένες και συνεχώς ελεγχόμενες. Οι διάφοροι τύποι τσιμέντων και το επίπεδο αντοχών τους, που αποτελεί και το σημαντικότερο χαρακτηριστικό τους, διαμορφώνονται από τη χημική σύσταση του κλίνκερ, το βαθμό άλεσης του τσιμέντου και την παρουσία ή όχι των διαφόρων πρόσθετων.

8ο Στάδιο-Σιλό Τσιμέντου: Το τσιμέντο αποθηκεύεται σε σιλό, που αποτελούν χώρους αποθήκευσης μέσης χρονικής διάρκειας.

9ο Στάδιο-Κατανάλωση: Το τσιμέντο διατίθεται στην κατανάλωση χύμα ή σε σάκους. Οι μεγαλύτερες ποσότητες διατίθενται χύμα με ειδικά σιλοφόρα αυτοκίνητα ή πλοία. [1]

1.4. Προτυποποίηση των Κοινών Τσιμέντων

1.4.1. Το Ευρωπαϊκό Πρότυπο prEN 197-1

Τα τελευταία χρόνια και στα πλαίσια της έκδοσης κοινών Ευρωπαϊκών Κανονισμών για όλες τις χώρες της CEN (Committee Europeenne de Normalisation) στην οποία μετέχει και η Ελλάδα, έχουν διαμορφωθεί σειρές προτύπων που αφορούν το τσιμέντο και το σκυρόδεμα. Στο σχετικό πρότυπο prEN 197-1 που αναφέρεται στην ενοποίηση των επιμέρους τύπων τσιμέντου που παράγονται στις διάφορες χώρες της Ευρώπης προβλέπονται οι πέντε τύποι και οι πολλές υποδιαιρέσεις που παρουσιάζονται στον Πίνακα 1.1. Σημειώνεται ότι ως τσιμέντα Portland ορίζονται αυτά των δύο πρώτων τύπων Ι και ΙΙ, όπου το κλίνκερ συμμετέχει σε ποσοστό μεγαλύτερο του 65%. [22]

<u>Πίνακας 1.1</u>: Τύποι Τσιμέντου σύμφωνα με το Ευρωπαϊκό Πρότυπο prEN 197-1

Τύπος	Ονομασία	Κύρια Συστατικά									Δευτ. Συστ.
		Κλίνκερ	Σκωρία	Silica fume	Ποζολάνες		Ιπτ. Τέφρες		Burnt shale	Ασβεστόλιθος	
		K	S	D	P	Q	V	W	T	L	
ΤΣΙΜΕΝΤΑ PORTLAND											
CEM I	I	95-100									0-5
ΣΥΝΘΕΤΑ ΤΣΙΜΕΝΤΑ PORTLAND											
CEM II	II/A-S	80-94	6-20								0-5
	II/B-S	65-79	21-35								
	II/A-D	90-94		6-10							0-5
	II/A-P	80-90			6-20						
	II/B-P	65-79			21-35						0-5
	II/A-Q	80-94									
	II/B-Q	65-79									
	II/A-V	80-94					6-20				
	II/B-V	65-79					21-35				0-5
	II/A-W	80-94						6-20			
	II/B-W	68-79						21-35			
	II/A-T	80-94							6-20		0-5
	II/B-T	65-79							21-35		
	II/A-L	80-94								6-20	0-5
	II/B-L	65-79								21-35	
	II/A-M	80-94	6-20								0-5
	II/B-M	65-79	21-35								
ΣΚΩΡΙΟΤΣΙΜΕΝΤΑ											
CEM III	III/A	35-64	36-65								
	III/B	20-34	66-80								0-5
	III/C	5-19	81-95								
ΠΟΖΟΛΑΝΙΚΑ ΤΣΙΜΕΝΤΑ											
CEM IV	IV/A	65-89		11-35							0-5
	IV/B	45-64		36-55							
ΣΥΝΘΕΤΑ ΤΣΙΜΕΝΤΑ											
CEM V	V/A	40-64	18-30		16-30						0-5
	V/B	20-39	31-50		31-50						

Από την 1η Απριλίου 2001, με απόφαση της Ευρωπαϊκής Επιτροπής, οι ποιότητες τσιμέντου που κυκλοφορούν στα κράτη-μέλη της Ευρωπαϊκής Ένωσης μπορούν να είναι πιστοποιημένες, σύμφωνα με τα νέα ευρωπαϊκά πρότυπα και να φέρουν τη διακριτική σήμανση CE. Από την 1η Απριλίου 2002 η σήμανση έγινε υποχρεωτική.

Ο Ελληνικός Οργανισμός Τυποποίησης (ΕΛΟΤ) και ειδικότερα η Επιτροπή "Τσιμέντο και Δομικοί Άσβεστοι" έχουν εκδώσει τα αντίστοιχα ελληνικά πρότυπα ήδη από τον Οκτώβριο του 2000, που είναι τα ίδια με αυτά του προτύπου prEN 197-1.

Το τσιμέντο -όπως και τα άλλα δομικά υλικά- πρέπει να πληροί συγκεκριμένες προδιαγραφές και να ανταποκρίνεται στα νέα ευρωπαϊκά πρότυπα, όσον αφορά τις ιδιότητες και τη σταθερότητα της παραγωγής του, για να μπορεί να διακινείται στο εξής ελεύθερα στις χώρες της Ευρωπαϊκής Ένωσης. Η σήμανση CE επάνω στο σακί και στα συνοδευτικά έγγραφα θα επιβεβαιώνει ότι το τσιμέντο εναρμονίζεται με τις απαιτήσεις των ευρωπαϊκών προτύπων και τις οδηγίες υγιεινής και ασφάλειας που αφορούν στα δομικά υλικά.

Πριν από τη σύνταξη των ευρωπαϊκών προτύπων, προηγήθηκε η κωδικοποίηση όλων των ποιοτήτων τσιμέντων κοινής αποδοχής και ευρείας χρήσεως που παράγονται στις

χώρες-μέλη της Ευρωπαϊκής Ένωσης και παράλληλα δημιουργήθηκε κοινή ορολογία, που θα διευκολύνει τη συνεννόηση μεταξύ μελετητών, χρηστών και κατασκευαστών. Σημειώνεται ότι η πιστοποίηση παρέχεται με βάση αυστηρότατα κριτήρια αξιολόγησης, κατόπιν εξωτερικής δειγματοληψίας και έλεγχο των παραγωγικών διεργασιών από ανεξάρτητο φορέα πιστοποίησης, διαπιστευμένο και κοινοποιημένο στην Ευρωπαϊκή Ένωση (ΕΛΟΤ).

1.4.2. Τύποι και Κατηγορίες Τσιμέντων

Κάθε χώρα παρασκευάζει τσιμέντο χρησιμοποιώντας τις πηγές πρώτων υλών που διαθέτει. Ανάλογα με τις υπάρχουσες πρώτες ύλες, παράγονται διάφοροι τύποι τσιμέντου (καθαρό ή αμιγές, με ποζολάνη, με ιπτάμενη τέφρα, σκωρία υψικαμίνου, πυριτική παιπάλη, ασβεστόλιθο κλπ), ποιότητες που δεν κυκλοφορούν σε όλα τα κράτη-μέλη της Ενωμένης Ευρώπης.

Τα νέα ευρωπαϊκά πρότυπα προδιαγράφουν το είδος και το ποσοστό των συστατικών που χρησιμοποιούνται στην παραγωγή 27 τύπων προϊόντων τσιμέντου και τα κατατάσσουν σε έξι κατηγορίες αντοχών, ανάλογα με την αντοχή τους σε θλίψη κονιάματος πρότυπης σύνθεσης και τρόπου παρασκευής.

Οι κανονικές αντοχές του τσιμέντου (standard strength) είναι οι θλιπτικές που προσδιορίζονται σύμφωνα με το prEN 196-1 στις 28 ημέρες και θα πρέπει να συμμορφώνονται με κάποιες συγκεκριμένες απαιτήσεις. Έχουν θεσπιστεί τρεις κατηγορίες αντοχών 32,5 N/mm^2, 42,5 N/mm^2 και 52,5 N/mm^2. Οι πρώιμες αντοχές μετρώνται στις 2 ημέρες, εκτός από την κατηγορία 32,5, όπου μετρώνται στις 7 ημέρες. Για κάθε κατηγορία προβλέπονται δύο τάξεις πρώιμων αντοχών, εκ των οποίων η πρώτη αναφέρεται στις κανονικές πρώιμες αντοχές και η άλλη, που συμβολίζεται με το γράμμα R, στις μεγάλες πρώιμες αντοχές ή, διαφορετικά, αντιστοιχεί σε τσιμέντα ταχείας ανάπτυξης αντοχών.

1.5. Επίδραση της Διαδικασίας Παραγωγής στις Ιδιότητες του Τσιμέντου

Οι ιδιότητες του τσιμέντου επηρεάζονται άμεσα από τη σύσταση και τη λεπτότητα του τσιμέντου. Σε μεγάλο ποσοστό, όμως, οι ιδιότητες αυτές και κυρίως οι πιο σημαντικές, που είναι η πήξη και η σκλήρυνση, είναι συνάρτηση και της διαδικασίας παραγωγής.

Για μια δεδομένη σύσταση και λεπτότητα τσιμέντου, η δραστικότητα των συστατικών του επηρεάζει το σχηματισμό της εσωτερικής δομής των προϊόντων της ενυδάτωσης του τσιμέντου. Η δραστικότητα αυτή είναι άμεσα συνυφασμένη με τις ιδιαιτερότητες της παραγωγικής διαδικασίας.

Το σύνολο των παραμέτρων που χαρακτηρίζουν την έψηση έχει τη μεγαλύτερη σπουδαιότητα για τη δραστικότητα των συστατικών του τσιμέντου. Πιο συγκεκριμένα, αυτή εξαρτάται από τη θερμοκρασία και τη διάρκεια της παραμονής του υλικού στη ζώνη κλινκεροποίησης, καθώς επίσης και από το ρυθμό ψύξης. Εκτός από το κλίνκερ, και τα άλλα υλικά (σκωρίες, ιπτάμενες τέφρες) που χρησιμοποιούνται ως κύρια συστατικά σε διάφορους τύπους τσιμέντου επηρεάζονται

σε μεγάλο βαθμό από τις θερμικές συνθήκες κάτω από τις οποίες παράχθηκαν ως υποπροϊόντα.

Επίσης, η ατμόσφαιρα που επικρατεί μέσα στην κάμινο κατά την παραγωγική διαδικασία έχει επίδραση στη δραστικότητα των συστατικών του τσιμέντου. Για παράδειγμα, κατά την έψηση και την ψύξη λαμβάνουν χώρα αντιδράσεις, στις οποίες το οξυγόνο και οι ατμοί των πτητικών ενώσεων που παράγονται έχουν ιδιαίτερη σημασία για το χαρακτηρισμό της ποιότητας του τσιμέντου που παράγεται από αυτές. Παράλληλα υπαγορεύουν μια σειρά από επεμβάσεις στη διαδικασία παραγωγής που κρίνονται απαραίτητες.

Η άλεση του τσιμέντου και ιδιαίτερα οι συνθήκες λειτουργίας του μύλου και του διαχωριστή έχουν καθοριστική επίδραση στη λεπτότητα του τσιμέντου, και μέσω αυτής, στις ιδιότητές του. Ακόμα, το περιβάλλον του μύλου, και συγκεκριμένα η υγρασία που υπάρχει στον αέρα ή στη γύψο, επιφέρει ορισμένα προβλήματα, όπως είναι η αρχική ενυδάτωση των συστατικών, τα οποία είναι μικρότερης σημασίας. Οι διαφοροποιήσεις στις λεπτότητες των επιμέρους συστατικών κατά τη συνάλεσή τους, που οφείλονται στις διαφορετικές αλεστικότητές τους, επηρεάζονται επίσης από την ακολουθούμενη τεχνολογία άλεσης και μπορούν να αρθούν με κατάλληλες επεμβάσεις.

Τέλος, το περιβάλλον του χώρου αποθήκευσης του κλίνκερ ή του τσιμέντου έχει αξιοσημείωτη επίδραση στις δύο ιδιότητες του τσιμέντου (πήξη και σκλήρυνση). [9]

Από τα παραπάνω φαίνεται ότι κάθε μία από τις κύριες φάσεις της παραγωγικής διαδικασίας του τσιμέντου έχει μικρότερη ή μεγαλύτερη επίδραση στην τελική του ποιότητα. Τα κύρια σημεία που παίζουν σημαντικό ρόλο στις τελικές ιδιότητες του τσιμέντου είναι τα εξής:

- Η σύσταση του τσιμέντου στα 4 βασικά οξείδια.
- Η λεπτότητα του τσιμέντου.
- Η σύσταση σε αλκάλια της τροφοδοσίας της καμίνου.
- Οι συνθήκες κλινκεροποίησης.
- Οι συνθήκες ψύξης.
- Οι συνθήκες άλεσης.
- Οι συνθήκες αποθήκευσης του τσιμέντου.

1.6. Προφίλ της Προς Μελέτη Εταιρείας

1.6.1. Αντικείμενο Εργασιών

Η Εταιρεία ιδρύθηκε το 1902 και εισήχθη στο Χ.Α.Α. δέκα χρόνια μετά, στις 22 Φεβρουαρίου 1912. Σήμερα έχει εδραιώσει την παρουσία της διεθνώς, καθώς διαθέτει και εκμεταλλεύεται παραγωγικές μονάδες σε πολλές χώρες του κόσμου. Ο κυρίαρχος στόχος της εταιρίας είναι η καθιέρωσή της ως Πολυεθνική Εταιρία, η οποία υπολογίζεται ως ανεξάρτητη δύναμη στην παγκόσμια αγορά των δομικών υλικών και συνδυάζει την επιχειρηματική ικανότητα και ανταγωνιστικότητα με σεβασμό για τον άνθρωπο, την κοινωνία και το περιβάλλον.

Πιο συγκεκριμένα ο συγκεκριμένος Όμιλος έχει στην κατοχή του:

- 11 μονάδες παραγωγής τσιμέντου, εκ των οποίων 4 στην Ελλάδα, 2 στις ΗΠΑ, 3 στα Βαλκάνια και 2 στη Μέση Ανατολή, συνολικής ετήσιας παραγωγικής δυναμικότητας 15 εκατομμυρίων τόνων
- 7 κέντρα διανομής τσιμέντου εκ των οποίων 2 βρίσκονται στις ΗΠΑ, 2 στην Αίγυπτο, και από 1 στην Ιταλία, Αγγλία και Γαλλία
- Ακόμη διατηρεί 67 μονάδες έτοιμου σκυροδέματος
- 10 λατομεία και 3 ορυχεία (19 εκ. τόνους)
- 1 μονάδα παραγωγής κονιαμάτων (INTERMIX)
- 1 μονάδα παραγωγής επιτραπέζιας πορσελάνης

Η ετήσια δυναμικότητα παραγωγής τσιμέντου της εταιρείας στην Ελλάδα ανέρχεται σε 6 εκατ. τόνους, κατέχοντας περίπου το 40% της αγοράς. Επιπλέον, η ετήσια παραγωγική δυναμικότητα τσιμέντου του Ομίλου στο εξωτερικό ανέρχεται σε άλλους 8 εκατ. τόνους.

Κατά την δεκαετία του '90, κυρίως, ο όμιλος καθετοποίησε σημαντικά τις δραστηριότητές του, επενδύοντας στους κλάδους ετοίμου σκυροδέματος, αδρανών υλικών και άλλων συναφών υλικών (κονιάματα, τσιμεντόλιθοι).

1.6.2. Στρατηγική

Η στρατηγική του Ομίλου συνοψίζεται ως εξής:

- Συνέχιση της διεθνούς επέκτασης στον κλάδο τσιμέντου με σκοπό την ενδυνάμωση της περιφερειακής παρουσίας στις 4 περιοχές που ήδη διατηρεί παραγωγική δραστηριότητα (Ελλάδα, ΗΠΑ, Βαλκάνια, Μέση Ανατολή).
- Περαιτέρω καθετοποίηση δραστηριοτήτων όπου οι συνθήκες αγοράς και ανταγωνισμού την ευνοούν.
- Συνεχής βελτίωση κόστους και παραγωγικότητας.
- Βέλτιστη αξιοποίηση του ανθρώπινου δυναμικού του και προσαρμογή του στην διεθνοποιημένη μορφή του.

1.6.3. Ανταγωνισμός – Επιχειρηματικοί Κίνδυνοι

Ο κλάδος παραγωγής τσιμέντου παρουσιάζει μια έντονη τάση συγκέντρωσης τα τελευταία χρόνια. Έτσι, σήμερα, το 43% της παγκόσμιας παραγωγής (πλην Κίνας) βρίσκεται στην κατοχή των πέντε μεγαλυτέρων ομίλων του κλάδου (Lafarge, Holcim, Cemex, Heidelberger, Italcementi). Η προς μελέτη εταιρεία βρίσκεται στην 15η θέση και με τις εξαγορές της παίρνει μέρος κατά μικρότερο ποσοστό στη συγκέντρωση του κλάδου.

Ο κλάδος τσιμέντου χαρακτηρίζεται από:

- Δυσκολία εισόδου νέων παικτών λόγω: α) δυσκολίας εύρεσης κατάλληλων πρώτων υλών, β) υψηλής απαίτησης κεφαλαίων και, γ) εξειδίκευσης παικτών.

- Τοπικό χαρακτήρα: δεν υπάρχουν παγκόσμιες οικονομίες κλίμακα ενώ το μεταφορικό κόστος είναι πολύ υψηλό (συγκρινόμενο με την αξία του προϊόντος).
- Έλλειψη ικανού ανταγωνιστικού υλικού (σε κόστος και ανθεκτικότητα).
- Υψηλό λειτουργικό cash flow.

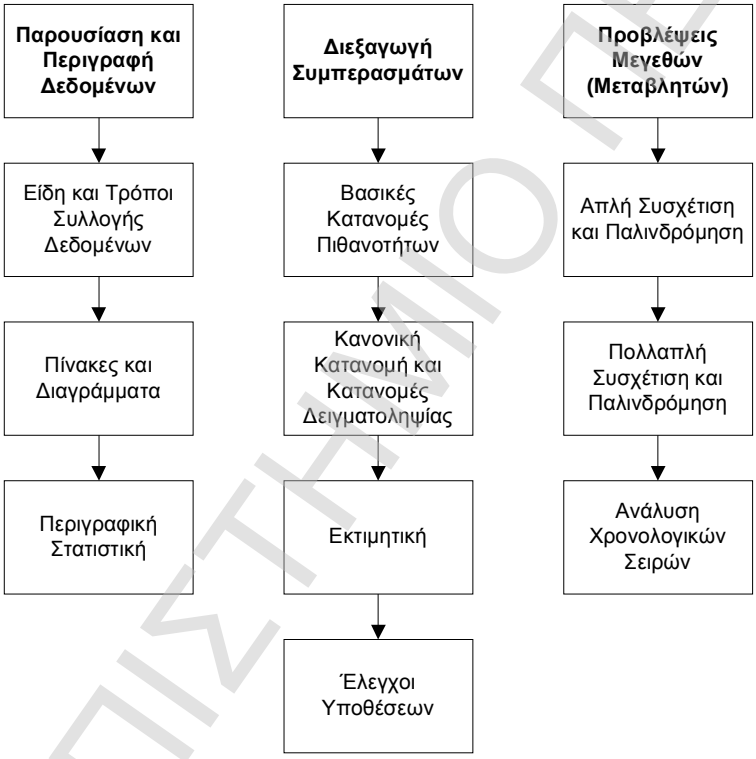
Όσον αφορά στους κινδύνους που ενδεχομένως να αντιμετωπίσει ο Όμιλος τα επόμενα χρόνια, αυτοί δεν διαφέρουν από αυτούς του παρελθόντος. Λόγω της φύσης του κύριου προϊόντος (commodity), οι τιμές πώλησης (και επομένως και η κερδοφορία) επηρεάζονται σημαντικά από την κατεύθυνση της ζήτησης αλλά και της προσφοράς. Η ζήτηση επηρεάζεται άμεσα από την πορεία των οικονομιών αλλά και από τα προγράμματα Δημοσίων Έργων. Είναι γεγονός ότι, στην περίοδο ύφεσης που περνούν οι βασικές οικονομίες, η ενδεχόμενη πτώση της ιδιωτικής οικοδομικής δραστηριότητας μπορεί να αντισταθμιστεί από τις αυξημένες Δημόσιες επενδύσεις σε έργα υποδομής (κυρίως στην Ελλάδα και τις ΗΠΑ). Από την πλευρά της προσφοράς (παραγωγής) τσιμέντου τα δεδομένα δεν μεταβάλλονται σημαντικά, καθώς χρειάζονται περίπου δύο χρόνια για να κτιστεί μια νέα γραμμή παραγωγής και υπάρχει αρκετός χρόνος για να προετοιμαστεί κανείς. Κίνδυνος όμως προκύπτει όταν επιδεινώνεται γρήγορα η ισορροπία προσφοράς και ζήτησης σε μια δεδομένη οικονομία ή περιοχή. Τότε "περισσεύουν" σημαντικές ποσότητες τσιμέντου στην τοπική αγορά που αναζητούν διέξοδο στις διεθνείς αγορές. Μια τέτοια συμπεριφορά μπορεί να έχει αλυσιδωτές αντιδράσεις σε πολύ μακρινές αγορές. Η κρίση της Ασίας αποτελεί παράδειγμα του κοντινού παρελθόντος και μια παρόμοια κατάσταση σε κάποιο άλλο μέρος του κόσμου δεν μπορεί να αποκλειστεί στο μέλλον. [1]

ΚΕΦΑΛΑΙΟ 2: ΒΑΣΙΚΕΣ ΕΝΝΟΙΕΣ ΤΗΣ ΣΤΑΤΙΣΤΙΚΗΣ

Σκοπός του κεφαλαίου είναι να παρουσιάσει κάποιες βασικές έννοιες της στατιστικής, πάνω στις οποίες βασίζεται όλη η ανάλυση και η επεξεργασία των αποτελεσμάτων στο Μέρος Β΄ της παρούσας εργασίας.

2.1. Βασικές Στατιστικές Μέθοδοι

Οι βασικές μέθοδοι στατιστικής ανάλυσης (ή αλλιώς ανάλυση δεδομένων) που χρησιμοποιούνται ευρύτατα σε διάφορους τομείς στο χώρο των επιχειρήσεων παρουσιάζονται στο Σχήμα 2.1.



Σχήμα 2.1: Βασικές Στατιστικές Μέθοδοι

Πιο συγκεκριμένα,

- **Περιγραφική Στατιστική**: Με τον όρο αυτόν περιγράφουμε τις μεθόδους που ασχολούνται με τη συλλογή, παρουσίαση και χαρακτηρισμό (ταξινόμηση) των δεδομένων ανάλογα με το είδος των χαρακτηριστικών που περιγράφουν. Το είδος των δεδομένων που αναλύουμε καθορίζει και την παράμετρο που θα χρησιμοποιήσουμε για να περιγράψουμε την τάση των δεδομένων. Για παράδειγμα, εάν το χαρακτηριστικό είναι ποσοτικό, αυτό που ενδιαφέρει τον αναλυτή είναι η μέση τιμή του χαρακτηριστικού, ενώ για ποιοτικά χαρακτηριστικά αυτό που μπορεί να τον ενδιαφέρει είναι η εύρεση κάποιων ποσοστών αυτού.

- **Επαγωγική Στατιστική**: Έτσι ορίζονται οι μέθοδοι που μας βοηθούν να εκτιμήσουμε τα χαρακτηριστικά ενός πληθυσμού, με βάση τα αποτελέσματα που προκύπτουν από τις παρατηρήσεις ενός δείγματος.

Σε έναν πληθυσμό το συγκεκριμένο χαρακτηριστικό που θέλουμε να αναλύσουμε ονομάζεται **παράμετρος του πληθυσμού** (population parameter). Εάν από ένα δείγμα του πληθυσμού αυτού ελέγξουμε πάλι το συγκεκριμένο χαρακτηριστικό, η τιμή που αυτό θα έχει αποτελεί μία **εκτίμηση** (statistic ή estimation) του άγνωστου χαρακτηριστικού. Η επαγωγική στατιστική μας επιτρέπει, χρησιμοποιώντας την εκτίμηση του δείγματος, να διεξάγουμε συμπεράσματα για τον πληθυσμό. Έτσι, με την κατάλληλη μέθοδο, μπορούμε να συμπεράνουμε σε ποιο διάστημα τιμών αναμένεται να διαμορφωθεί το χαρακτηριστικό που αναζητάμε.

Οι τεχνικές των προβλέψεων αποτελούν, επίσης, μία χρήσιμη κατηγορία στατιστικών μεθόδων. Συνήθως οι μέθοδοι αυτές συνδυάζονται με τις μεθόδους της περιγραφικής και επαγωγικής στατιστικής. [51]

2.2. Τρόποι Παρουσίασης Στατιστικών Δεδομένων

Καταρχάς σημειώνεται ότι τα δεδομένα είναι οι αριθμητικές πληροφορίες που συλλέγουμε και στη συνέχεια επεξεργαζόμαστε για να πάρουμε μια απόφαση. Τα δεδομένα ποικίλουν ανάλογα με το είδος του χαρακτηριστικού του οποίου αποτελούν την αριθμητική έκφραση. Υπάρχουν δύο μεγάλες κατηγορίες δεδομένων: τα ποσοτικά και τα ποιοτικά. Επειδή τα ποσοτικά και ποιοτικά χαρακτηριστικά σε ένα δείγμα που μελετάμε διαφέρουν (μεταβάλλονται) από περίπτωση σε περίπτωση, τα ονομάζουμε μεταβλητές. Μάλιστα, στο βαθμό που οι τιμές τους είναι τυχαίες (δηλαδή δεν τις γνωρίζουμε πριν εκτελέσουμε τη μέτρηση) καλούνται τυχαίες μεταβλητές.

Τέλος, σημειώνεται ότι οι ποσοτικές μεταβλητές διακρίνονται σε ασυνεχείς (διακριτές) και σε συνεχείς. Οι πρώτες αναφέρονται στα χαρακτηριστικά που εκφράζονται μόνο με ακέραιους αριθμούς, π.χ. αριθμός μελών οικογένειας, αριθμός επισκέψεων στον κινηματογράφο, αριθμός περιοδικών που αγοράζουν τα μέλη μιας οικογένειας κλπ. Αντίθετα, οι μεταβλητές που παίρνουν τιμές σε όλο το εύρος των τιμών ονομάζονται συνεχείς. Παράδειγμα, το βάρος (κιλά, γραμμάρια, δέκατα γραμμαρίων), το μήκος (μέτρα, εκατοστά, χιλιοστά), ο χρόνος (ώρες, λεπτά, δευτερόλεπτα) κλπ.

Η περιγραφή, οργάνωση και παρουσίαση των αριθμητικών δεδομένων γίνονται συνήθως με πίνακες και διαγράμματα.

2.2.1. Πίνακες

Η χρήση πινάκων ευνοεί κυρίως την παρουσίαση πολλών δεδομένων, χωρίς να κουράζει ιδιαίτερα τον αναγνώστη και ταυτόχρονα πληροφορώντας αυτόν για τη συμπεριφορά κάποιου συγκεκριμένου χαρακτηριστικού. Ένας τρόπος που επιτυγχάνεται το παραπάνω είναι με την παρουσίαση των δεδομένων σε αύξουσα

τάξη μεγέθους, δηλαδή από τη χαμηλότερη έως την υψηλότερη τιμή που παίρνει το χαρακτηριστικό.

Μερικές φορές, όταν υπάρχουν άπειρα δεδομένα, είναι προτιμότερο τα χαρακτηριστικά που μελετώνται να συγκεντρώνονται σε διαστήματα τιμών, προκειμένου η εικόνα της κατανομής τους να είναι πιο συνοπτική. Αυτού του είδους η ταξινόμηση ονομάζεται **κατανομή συχνοτήτων** του προς μελέτη χαρακτηριστικού. Δηλαδή, δείχνει πώς κατανέμονται τα χαρακτηριστικά σε διαστήματα τιμών. Έτσι, ο καινούριος περιληπτικός πίνακας που θα προκύψει θα δείχνει τη συχνότητα (πλήθος) των χαρακτηριστικών που ανήκουν στα αντίστοιχα διαστήματα τιμών.

Για τη σωστή κατασκευή ενός πίνακα κατανομής συχνοτήτων, πρέπει να προσέξουμε τα εξής τρία σημεία:

1. Την επιλογή του κατάλληλου αριθμού διαστημάτων, τα οποία ονομάζονται διαστήματα τάξης.

2. Το σωστό πλάτος (εύρος) των διαστημάτων τάξης.

3. Τον καθορισμό των ορίων των διαστημάτων τάξης, έτσι ώστε και να καλύπτονται όλες οι τιμές, αλλά και να μην υπάρχουν επικαλύψεις.

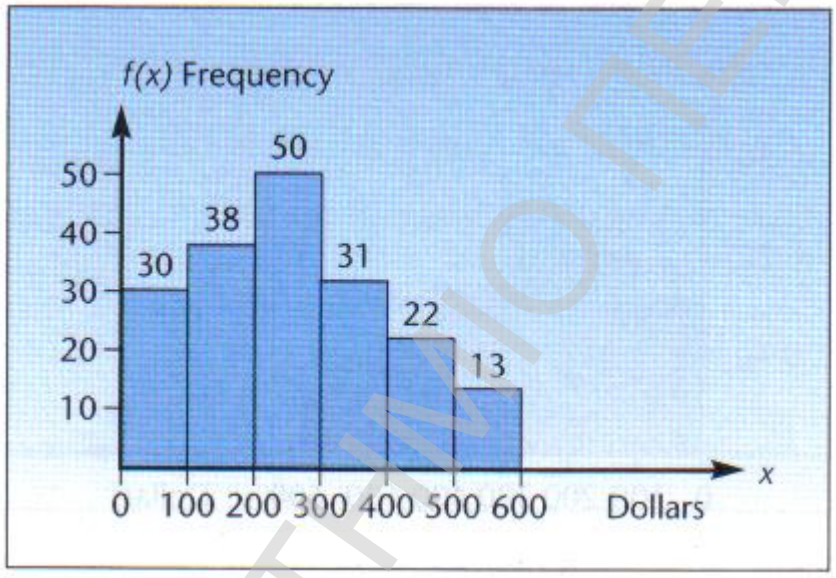
Πιο αναλυτικά, οι βασικές αρχές είναι οι εξής: Ο αριθμός των διαστημάτων τάξης να μην είναι ούτε μικρός ούτε μεγάλος. Σκοπός του πίνακα κατανομής συχνοτήτων είναι να ικανοποιεί τα κριτήρια της πληροφόρησης αλλά και της περίληψης. Να μην κουράζει, αλλά και να πληροφορεί.

Οι πληροφορίες που παίρνουμε από έναν πίνακα κατανομής συχνοτήτων είναι οι εξής: Πρώτα από όλα, βλέπουμε συνοπτικά στα διαστήματα τάξης τη συγκέντρωση (ή κατανομή) των τιμών του συγκεκριμένου χαρακτηριστικού. Οπότε, μπορούμε να βγάλουμε κάποια συμπεράσματα, όπως για παράδειγμα ποιο διάστημα εμφανίζει τη μεγαλύτερη συχνότητα. Επίσης, μπορούμε να υπολογίσουμε και τις σχετικές συχνότητες των διαστημάτων τάξης, δηλαδή τις απόλυτες συχνότητες των παρατηρήσεων εντός μιας τάξης, ως ποσοστό % του συνόλου των παρατηρήσεων για όλες τις τάξεις. Το πλεονέκτημα των σχετικών συχνοτήτων είναι ότι είναι προτυποποιημένες, δηλαδή έχουν άθροισμα ίσο με τη μονάδα. Τέλος, μπορούμε να υπολογίσουμε και τις αθροιστικές συχνότητες, οι οποίες προκύπτουν από τη διαδοχική άθροιση των απλών συχνοτήτων, και η πληροφορία που παρέχουν είναι πόσες παρατηρήσεις έχουμε σωρευτικά μέχρι μία τιμή. Εκφράζοντας τις αθροιστικές συχνότητες ως ποσοστά % του συνόλου των παρατηρήσεων, προκύπτουν οι σχετικές αθροιστικές συχνότητες.

Οι πίνακες κατανομών συχνοτήτων χρησιμοποιούνται συνήθως για την παρουσίαση ποσοτικών δεδομένων. Μπορούν, όμως, να χρησιμοποιηθούν και στις περιπτώσεις εκείνες που θέλουμε να δείξουμε την κατανομή των παρατηρήσεων ως προς ένα ποιοτικό χαρακτηριστικό.

2.2.2. Διαγράμματα

Τα πιο διαδεδομένα διαγράμματα παρουσίασης δεδομένων είναι τα **ιστογράμματα**. Ιστόγραμμα είναι ένα διάγραμμα με κάθετες στήλες (ή ιστούς) που έχουν ως βάση τα διαστήματα τάξης και ύψος ανάλογο με τον αριθμό (συχνότητα) των παρατηρήσεων που ανήκουν στα διαστήματα. Το Σχήμα 2.2 δείχνει το ιστόγραμμα της κατανομής των συχνοτήτων των δολαρίων που ξοδεύτηκαν από 184 πελάτες σε ένα συγκεκριμένο κατάστημα. Στον οριζόντιο άξονα απεικονίζεται η τιμή της μεταβλητής (ποσό σε δολάρια) και στον κάθετο άξονα οι συχνότητες (αριθμός πελατών). Η κλίμακα μέτρησης του οριζόντιου άξονα ξεκινάει από την τιμή του κάτω ορίου του πρώτου διαστήματος (μικρότερες τιμές δεν ενδιαφέρουν). Ο κάθετος, όμως, άξονας αρχίζει από την τιμή μηδέν, διότι το ύψος κάθε στήλης αντιστοιχεί στον αριθμό των παρατηρήσεων (συχνότητα) που ανήκουν στα διαστήματα τάξης.



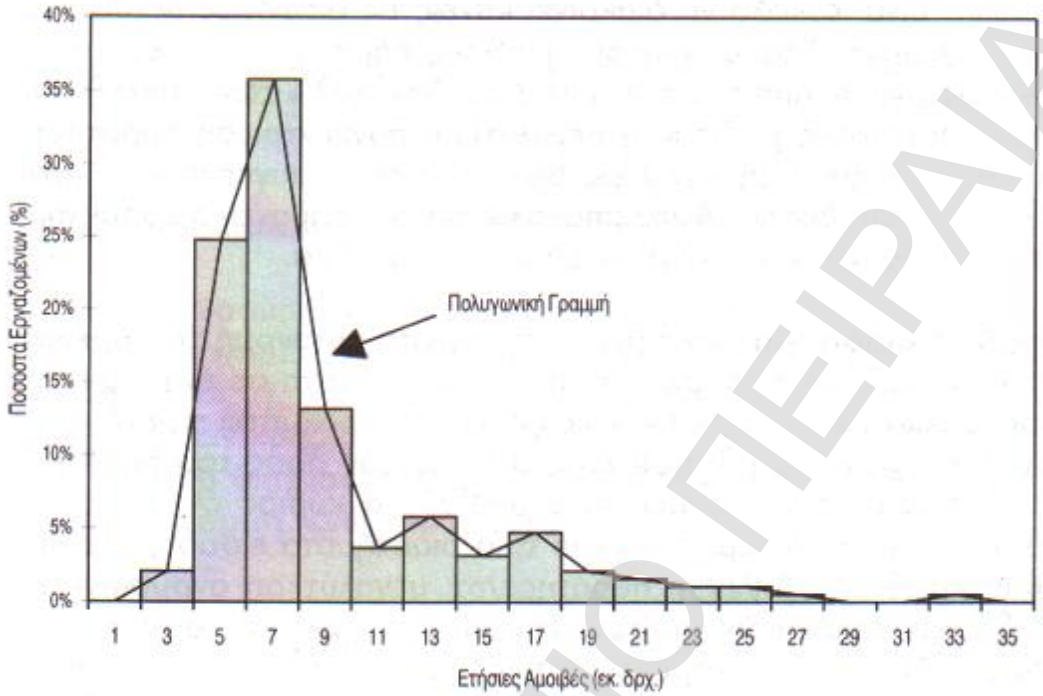
Σχήμα 2.2: Ιστόγραμμα Κατανομής Συχνοτήτων

Αντί των απόλυτων συχνοτήτων, μπορούμε να απεικονίσουμε στον κάθετο άξονα τις σχετικές συχνότητες, δηλαδή τα ποσοστά το συνόλου των παρατηρήσεων που ανήκουν στα διαστήματα τάξης. Γενικά, η χρήση των σχετικών συχνοτήτων είναι προτιμότερη των απλών συχνοτήτων, κυρίως στις περιπτώσεις που συγκρίνουμε διαφορετικές κατανομές με διαφορετικό συνολικό αριθμό παρατηρήσεων.

Ένας εναλλακτικός τρόπος απεικόνισης της κατανομής συχνοτήτων (απόλυτων ή σχετικών) είναι η **πολυγωνική γραμμή** που συνδέει τα "κεντρικά σημεία" των ιστών, όπως φαίνεται στο Σχήμα 2.3. Η φιλοσοφία αυτού του διαγράμματος είναι ότι τα "κεντρικά" σημεία αντιπροσωπεύουν τις συχνότητες, ενώ το σχήμα της πολυγωνικής γραμμής δίνει μια εικόνα του είδους της κατανομής.

Έτσι, τα σημεία της πολυγωνικής γραμμής αντιστοιχούν στους κεντρικούς όρους των διαστημάτων τάξης. Πιο συγκεκριμένα, σε κάθε διάστημα τάξης υπάρχει μια κεντρική τιμή που βρίσκεται στο μέσον του διαστήματος και ισούται με τον απλό μέσο όρο των άνω και κάτω ορίων. Επομένως, όταν κατασκευάζουμε την πολυγωνική γραμμή ένας εναλλακτικός τρόπος είναι να απεικονίζουμε στον οριζόντιο άξονα τις τιμές των κεντρικών όρων, αντί των ορίων των διαστημάτων. Το

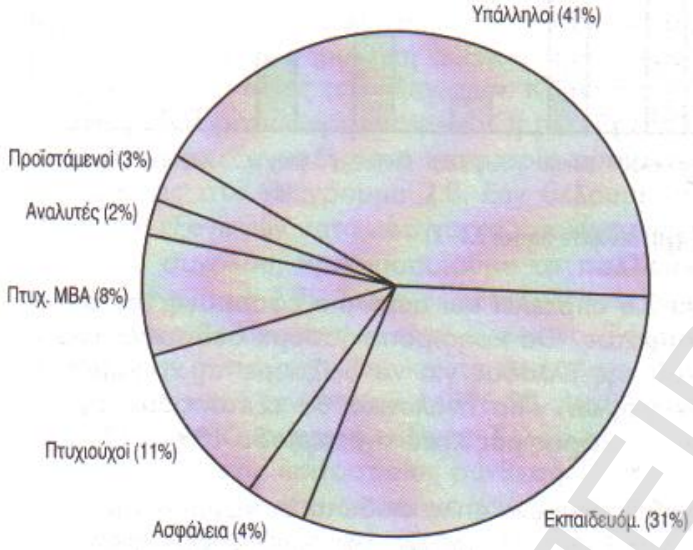
Σχήμα 2.3 δίνει την ποσοστιαία κατανομή συχνοτήτων των ετήσιων αμοιβών των εργαζομένων σε μια επιχείρηση και την πολυγωνική γραμμή.



<u>Σχήμα 2.3</u>: Ποσοστιαία Κατανομή Συχνοτήτων και Πολυγωνική Γραμμή

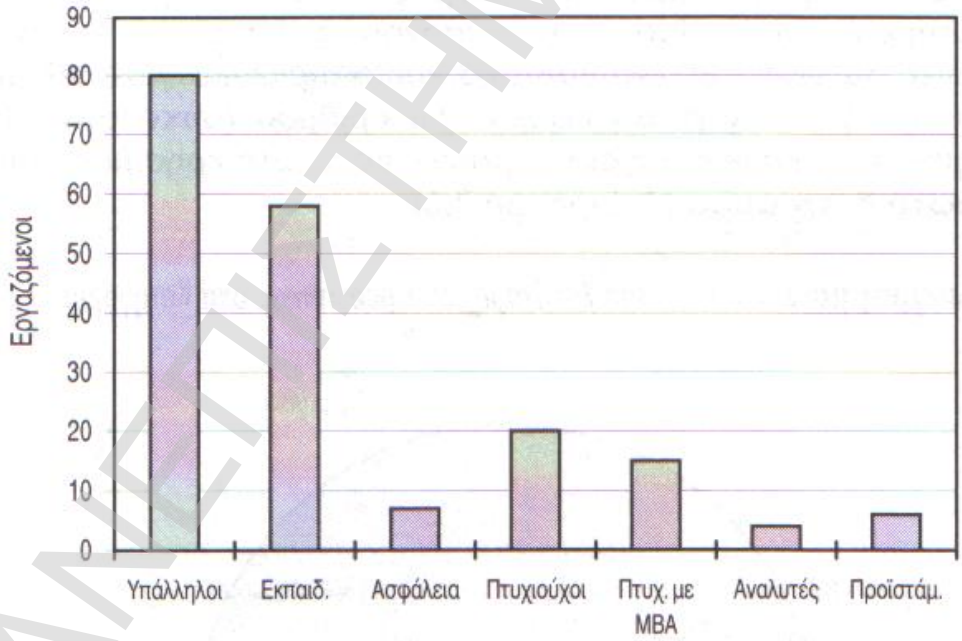
Επίσης, ένα άλλο διάγραμμα που αφορά τις κατανομές συχνοτήτων είναι η αθροιστική πολυγωνική γραμμή, η οποία απεικονίζει τις αθροιστικές συχνότητες. Εάν χρησιμοποιήσουμε τις σχετικές αθροιστικές συχνότητες, τότε το διάγραμμα ονομάζεται ποσοστιαία αθροιστική πολυγωνική γραμμή.

Ένα άλλο διάγραμμα που χρησιμοποιείται είναι το **κυκλικό διάγραμμα** (pie chart). Αυτό χρησιμοποιείται όταν θέλουμε να απεικονίσουμε την κατανομή των παρατηρήσεων σε κατηγορίες σύμφωνα με ένα ποιοτικό χαρακτηριστικό. Το Σχήμα 2.4 απεικονίζει την ποσοστιαία σύνθεση του προσωπικού που εργάζεται σε μια επιχείρηση, κατά κατηγορία εργαζομένων.



Σχήμα 2.4: Κυκλικό Διάγραμμα (pie chart)

Εάν θέλουμε να δώσουμε μια συγκεκριμένη εικόνα του αριθμού (συχνότητα) των εργαζομένων σε κάθε κατηγορία προσωπικού, τότε χρησιμοποιούμε το **ακιδωτό διάγραμμα** (Σχήμα 2.5). Τα ακιδωτά διαγράμματα ή διαγράμματα στηλών (bar charts) αποτελούν τον πιο συνηθισμένο τρόπο απεικόνισης οικονομικών δεδομένων.



Σχήμα 2.5: Ακιδωτό Διάγραμμα (bar chart)

Τα διαγράμματα αποτελούν τα βασικά και πιο χρήσιμα εργαλεία για την ανάλυση των αριθμητικών δεδομένων. Αποκαλύπτουν με μία ματιά τη συμπεριφορά των δεδομένων και δίνουν ενδείξεις για το είδος της ανάλυσης που πρέπει να ακολουθήσουμε. [10, 29]

2.3. Βασικά Χαρακτηριστικά Αριθμητικών Δεδομένων

Στην προηγούμενη ενότητα είδαμε πώς παρουσιάζονται τα δεδομένα με τη βοήθεια των πινάκων και των διαγραμμάτων. Όμως, πώς αξιοποιούνται αυτές οι πληροφορίες; Αν και η παρουσίαση των δεδομένων αποτελεί βασικό στοιχείο της περιγραφικής στατιστικής, δεν αρκεί για να αποκαλύψει όλη την εικόνα που περιέχουν τα δεδομένα. Η πλήρης ανάλυση των δεδομένων δεν βασίζεται μόνο στην παρουσίαση και παρατήρηση του τι προσπαθούν τα δεδομένα να αποκαλύψουν, αλλά περιλαμβάνει υπολογισμούς και εκτιμήσεις των βασικών χαρακτηριστικών, η ανάλυση των οποίων θα οδηγήσει και στην πλήρη κατανόηση της εικόνας που αποκαλύπτουν. Έτσι, σκοπός μας είναι να δούμε με ποιες μεθόδους θα επιτύχουμε τη συνοπτική περιγραφή των δεδομένων και στη συνέχεια την ερμηνεία τους.

Τρεις είναι οι βασικές ιδιότητες που χαρακτηρίζουν ένα σύνολο αριθμητικών δεδομένων:

- Η κεντρική τάση
- Η διασπορά ή μεταβλητότητα
- Το σχήμα

Η ανάλυση των δεδομένων περιλαμβάνει, μεταξύ άλλων, και μια σειρά μετρήσεων που θα περιγράφουν τις παραπάνω ιδιότητες: κεντρική τάση, διασπορά, και σχήμα της κατανομής. Εάν αυτές οι περιγραφικές μετρήσεις προκύπτουν από δεδομένα ενός δείγματος ονομάζονται **εκτιμήσεις** ή **στατιστικές** (estimations ή statistics). Ενώ, αν υπολογίζονται για όλον τον πληθυσμό, τότε ονομάζονται **παράμετροι**.

2.3.1. Μέτρηση της Κεντρικής Τάσης

Στις περισσότερες περιπτώσεις ένα σύνολο δεδομένων παρουσιάζει τάση συγκέντρωσης των τιμών του γύρω από μία κεντρική τιμή. Έτσι, για κάθε συγκεκριμένο σύνολο δεδομένων, είναι δυνατόν να επιλέξουμε κάποια τυπική τιμή ή μέσο που θα περιγράφει τη συμπεριφορά των τιμών. Δηλαδή, προσπαθούμε να βρούμε τον εκπρόσωπο των τιμών που θα τις αντιπροσωπεύει όποτε θα αναφερόμαστε σε αυτές.

Τρεις είναι οι συνηθέστεροι τρόποι μέτρησης της κεντρικής τάσης μιας ομάδας αριθμητικών δεδομένων: ο αριθμητικός μέσος, η διάμεσος και το σημείο μέγιστης συχνότητας (ή τύπος).

- ***Αριθμητικός Μέσος***

Ο αριθμητικός μέσος (ή μέσος) είναι ο συνηθέστερος τρόπος μέτρησης της κεντρικής τάσης. Υπολογίζεται από το άθροισμα όλων των τιμών διαιρούμενο με το πλήθος των παρατηρήσεων. Εάν συμβολίσουμε με X το χαρακτηριστικό που μετράμε (μεταβλητή) και με n το πλήθος των παρατηρήσεων, τότε οι τιμές συμβολίζονται με $X_1, X_2, \ldots, X_n$. Ο αριθμητικός μέσος, που συμβολίζεται με $\overline{X}$, υπολογίζεται με τον εξής τύπο:

$$\overline{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + ... + X_n}{n}$$

Ο μέσος που προκύπτει από τον παραπάνω τύπο ονομάζεται απλός ή αστάθμητος αριθμητικός μέσος (arithmetic mean), διότι προκύπτει από το απλό άθροισμα των τιμών της *X*. Από την άλλη, ο σταθμικός αριθμητικός μέσος υπολογίζεται ως εξής:

$$\overline{X} = \frac{\overline{X}_1 n_1 + \overline{X}_2 n_2 + ... + \overline{X}_k n_k}{n_1 + n_2 + ... + n_k}$$

ή

$$\overline{X} = \overline{X}_1 (n_1 / n) + \overline{X}_2 (n_2 / n) + ... + \overline{X}_k (n_k / n)$$

όπου

$$n = n_1 + n_2 + \ldots + n_k$$

Οι συντελεστές στάθμισης n_i/n ονομάζονται σχετικοί ή ποσοστιαίοι συντελεστές στάθμισης. Το άθροισμα των σχετικών συντελεστών στάθμισης ισούται με τη μονάδα. Ένας γενικός τύπος του σταθμικού μέσου εάν X_1, X_2, …, X_k είναι μια σειρά μετρήσεων με συντελεστές στάθμισης $w_1 + w_2 + \ldots + w_k$, αντίστοιχα, είναι ο ακόλουθος:

$$\overline{X} = \sum_{i=1}^{k} X_i w_i$$

όπου

$$\sum_{i=1}^{k} w_i = 1$$

Ο αριθμητικός μέσος είναι ο πιο χρήσιμος τρόπος μέτρησης της τάσης ενός δείγματος αριθμητικών δεδομένων. Βασικό του πλεονέκτημα είναι ότι η εκτίμησή του βασίζεται σε όλες τις τιμές του δείγματος. Για το λόγο αυτό, καλείται και επαρκής εκτιμητής της κεντρικής τάσης. Έχει όμως και μειονεκτήματα, καθώς παρασύρεται από τις ακραίες τιμές και πολλές φορές οδηγεί σε λανθασμένα συμπεράσματα. Έτσι, για κατανομές που υπάρχουν ακραίες τιμές ο αριθμητικός μέσος δεν αποτελεί αξιόπιστο τρόπο μέτρησης της τάσης. Σε αυτές τις περιπτώσεις είναι προτιμότερο να χρησιμοποιήσουμε τη διάμεσο, η οποία περιγράφεται στη συνέχεια.

- ***Διάμεσος***

Η διάμεσος (median) είναι η μεσαία τιμή μιας ομάδας τιμών ιεραρχημένων σε αύξουσα τάξη μεγέθους. Εάν δεν υπάρχουν δεσμοί (οι τιμές δεν συμπίπτουν μεταξύ τους), τότε οι μισές παρατηρήσεις είναι μικρότερες της διαμέσου και οι άλλες μισές μεγαλύτερες. Έτσι, η διάμεσος δείχνει την τιμή που χωρίζει τις παρατηρήσεις σε δύο ίσες υπο-ομάδες. Ο υπολογισμός της τιμής της είναι εύκολος και το μόνο που προϋποθέτει είναι ότι οι τιμές βρίσκονται σε αύξουσα τάξη μεγέθους.

Εάν ο αριθμός παρατηρήσεων είναι περιττός αριθμός, τότε η διάμεσος είναι η κεντρική τιμή. Δηλαδή, η (n+1)/2 παρατήρηση, εφόσον οι n τιμές του δείγματος

τεθούν σε αύξουσα τάξη μεγέθους. Για ζυγό αριθμό παρατηρήσεων η διάμεσος ισούται με τον απλό αριθμητικό μέσο των δύο κεντρικών τιμών. Δηλαδή των n/2 και (n/2)+1 παρατηρήσεων. Το πλεονέκτημα της διαμέσου είναι ότι δεν επηρεάζεται από τις ακραίες τιμές.

- ***Σημείο Μέγιστης Συχνότητας ή Κορυφή***

Το σημείο μέγιστης συχνότητας ή επικρατούσα τιμή ή κορυφή (mode) είναι η τιμή με τη μεγαλύτερη συχνότητα σε ένα σύνολο μετρήσεων. Εάν τα στοιχεία είναι αταξινόμητα, τότε η επικρατούσα τιμή είναι εκείνη που εμφανίζεται συχνότερα. Στην περίπτωση που οι μεταβλητές είναι ασυνεχείς (ακέραιος αριθμός), ο υπολογισμός του τύπου είναι εύκολος. Όμως, στην περίπτωση των συνεχών μεταβλητών, και για αταξινόμητα δεδομένα, ο προσδιορισμός του τύπου είναι αδύνατος, διότι είναι πολύ πιθανό όλες οι τιμές να διαφέρουν μεταξύ τους. Σε αυτές τις περιπτώσεις, πρώτα κατασκευάζουμε την κατανομή συχνοτήτων και στη συνέχεια εκτιμούμε το τύπο (επικρατούσα τιμή). Η τιμή του εντοπίζεται στο διάστημα με τη μεγαλύτερη συχνότητα και προσεγγίζεται διαγραμματικά από το ιστόγραμμα.

2.3.2. Μέτρηση της Διασποράς

Η δεύτερη σημαντική ιδιότητα που χαρακτηρίζει ένα σύνολο αριθμητικών δεδομένων είναι η διασπορά ή μεταβλητότητα. Η διασπορά είναι το μέγεθος της ανομοιογένειας μεταξύ των τιμών, δηλαδή πόσο διαφέρουν μεταξύ τους ή πόσο διεσπαρμένες είναι οι τιμές.

Πέντε είναι οι συνηθέστεροι τρόποι μέτρησης της διασποράς: το εύρος, η τεταρτημοριακή απόκλιση, η διακύμανση, η τυπική απόκλιση και ο συντελεστής μεταβλητότητας.

- ***Εύρος***

Το εύρος (range) είναι η διαφορά μεταξύ της μεγαλύτερης (X_{max}) και της μικρότερης τιμής (X_{min}) των δεδομένων. Τα πλεονεκτήματα του εύρους είναι ο εύκολος υπολογισμός του και ότι είναι κατανοητή η ερμηνεία του. Βασικό, όμως, μειονέκτημα είναι η επιρροή του από τις ακραίες τιμές (μέγιστες ή ελάχιστες). Γι' αυτούς τους λόγους, χρησιμοποιείται κατά κανόνα σε στατιστικές που αφορούν ποιοτικούς ελέγχους στη βιομηχανική παραγωγή. Σε αυτές τις περιπτώσεις, οι τεχνικές προδιαγραφές θέτουν όρια ανοχής (πλάτος, μήκος, βάρος, αντίσταση, κλπ), όπου η διαφορά μεταξύ της μέγιστης και ελάχιστης τιμής που παρατηρείται σε μαζική παραγωγή δεν πρέπει να υπερβαίνει μια προκαθορισμένη τιμή.

- ***Τεταρτημοριακή Απόκλιση***

Το μειονέκτημα του εύρους αντιμετωπίζεται με την τεταρτημοριακή απόκλιση ή ενδοτεταρτημοριακό εύρος (quartile deviation ή interquartile range, αντίστοιχα). Ως τεταρτημόρια ορίζονται οι τιμές που χωρίζουν το σύνολο των παρατηρήσεων σε τέσσερα ίσα (από πλευράς παρατηρήσεων) μέρη. Έτσι, μέχρι την τιμή του πρώτου

τεταρτημορίου (Q_1) βρίσκεται το 25% των παρατηρήσεων. Από το πρώτο τεταρτημόριο μέχρι το δεύτερο (που συμπίπτει με τη διάμεσο) έχουμε το επόμενο 25% των παρατηρήσεων, από το Q_2 έως το Q_3 έχουμε το επόμενο 25% των τιμών και τέλος από το Q_3 και μετά έχουμε το τελευταίο 25% των παρατηρήσεων. Ο τύπος υπολογισμού των τεταρτημορίων είναι *(n+1)P/100*, όπου *P* είναι η τιμή κάτω από την οποία βρίσκεται το *P%* των παρατηρήσεων και *n* είναι ο αριθμός των παρατηρήσεων.

Το ενδοτεταρτημοριακό εύρος υπολογίζεται από τη διαφορά μεταξύ του πρώτου και του τρίτου τεταρτημορίου, δείχνει δηλαδή το εύρος των τιμών που συγκεντρώνεται το μεσαίο (κεντρικό) 50% των παρατηρήσεων. Τα τεταρτημόρια έχουν το πλεονέκτημα ότι δεν επηρεάζονται από τις ακραίες τιμές, και επομένως, το ενδοτεταρτημοριακό εύρος δείχνει το εύρος των κεντρικών τιμών του δείγματος, χωρίς να λαμβάνει υπόψη τις ακραίες τιμές.

- ***Διακύμανση***

Τόσο το εύρος όσο και η τεταρτημοριακή απόκλιση που μετρούν τη συνολική και μεσαία διασπορά, αντίστοιχα, δεν λαμβάνουν υπόψη τη συμπεριφορά των υπολοίπων τιμών του δείγματος. Έτσι, χρειαζόμαστε έναν τρόπο μέτρησης της διασποράς που να βασίζεται σε όλες τις τιμές των δεδομένων και στον τρόπο που κατανέμονται. Ο πιο συνηθισμένος τρόπος μέτρησης της διασποράς, που βασίζεται σε όλες τις παρατηρήσεις και ταυτόχρονα μετράει τη συγκέντρωση των τιμών γύρω από τον αριθμητικό μέσο, είναι η διακύμανση (variance) και συμβολίζεται με s^2.

Η διακύμανση είναι ο μέσος όρος των τετραγωνικών αποκλίσεων των τιμών από τον αριθμητικό μέσο. Δηλαδή, για ένα δείγμα n τιμών, X_1, X_2,…, X_n, η διακύμανση ισούται με:

$$s^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}$$

- ***Τυπική Απόκλιση***

Η τυπική απόκλιση (standard deviation) είναι η τετραγωνική ρίζα της διακύμανσης και συμβολίζεται με s. Επομένως, ισχύει ότι:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1}}$$

Η τυπική απόκλιση εκφράζεται στις ίδιες μονάδες μέτρησης με την αρχική μεταβλητή, σε αντίθεση με τη διακύμανση όπου οι μονάδες υψώνονται στο τετράγωνο.

Είναι προφανές ότι όσο μικρότερη είναι η τιμή της διακύμανσης και, ως εκ τούτου, και της τυπικής απόκλισης, τόσο μικρότερη είναι η διασπορά των τιμών. Επίσης,

τόσο η διακύμανση όσο και η τυπική απόκλιση έχουν πάντα θετικές τιμές, διότι ο υπολογισμός τους βασίζεται σε άθροισμα τετραγώνων. Η μόνη περίπτωση να έχουν τιμή μηδέν είναι στην ακραία περίπτωση που όλες οι τιμές είναι ίσες μεταξύ τους, δηλαδή δεν υπάρχει καμία διασπορά μεταξύ των τιμών.

Συνεπώς, η διακύμανση και η τυπική απόκλιση μετρούν τη μέση διασπορά των τιμών γύρω από τον αριθμητικό μέσο. Ο αριθμητικός μέσος είναι το μέσο σημείο των τιμών ενός δείγματος, με αποτέλεσμα οι αποκλίσεις των τιμών, άλλες θετικές και άλλες αρνητικές, να αλληλοαναιρούνται και το άθροισμά τους να ισούται με μηδέν, δηλαδή:

$$\sum_{i=1}^{n}(X_i - \overline{X}) = 0$$

Η τυπική απόκλιση είναι η σπουδαιότερη εκτίμηση της διασποράς των τιμών. Όχι μόνο εκφράζει τη διασπορά τους γύρω από τον αριθμητικό μέσο, αλλά πληροφορεί και για τον τρόπο συγκέντρωσης των τιμών γύρω από το μέσο. Χωρίς την τυπική απόκλιση δεν θα είχαν αναπτυχθεί οι μέθοδοι της επαγωγικής στατιστικής.

- ***<u>Συντελεστής Μεταβλητότητας</u>***

Σε αντίθεση με τους παραπάνω τρόπους μέτρησης της διασποράς που εκφράζονται στις μονάδες μέτρησης του συγκεκριμένου χαρακτηριστικού, ο συντελεστής μεταβλητότητας (coefficient of variation) αποτελεί σχετική μέτρηση της διασποράς και ισούται με:

$$CV = \left(\frac{s}{\overline{X}}\right)$$

και εκφράζεται ως ποσοστό επί τοις εκατό (%). Έτσι, ο λόγος της τυπικής απόκλισης s προς τον αριθμητικό μέσο $\overline{X}$ εκφράζει τη διασπορά γύρω από το μέσο ως ποσοστό του αριθμητικού μέσου. Επίσης, ο συντελεστής μεταβλητότητας, ως σχετική μέτρηση της διασποράς, είναι ιδιαίτερα χρήσιμος όταν συγκρίνουμε τη μεταβλητότητα δύο ή περισσοτέρων ομάδων δεδομένων που εκφράζονται σε διαφορετικές μονάδες μέτρησης.

2.3.3. <u>Έλεγχος του Σχήματος</u>

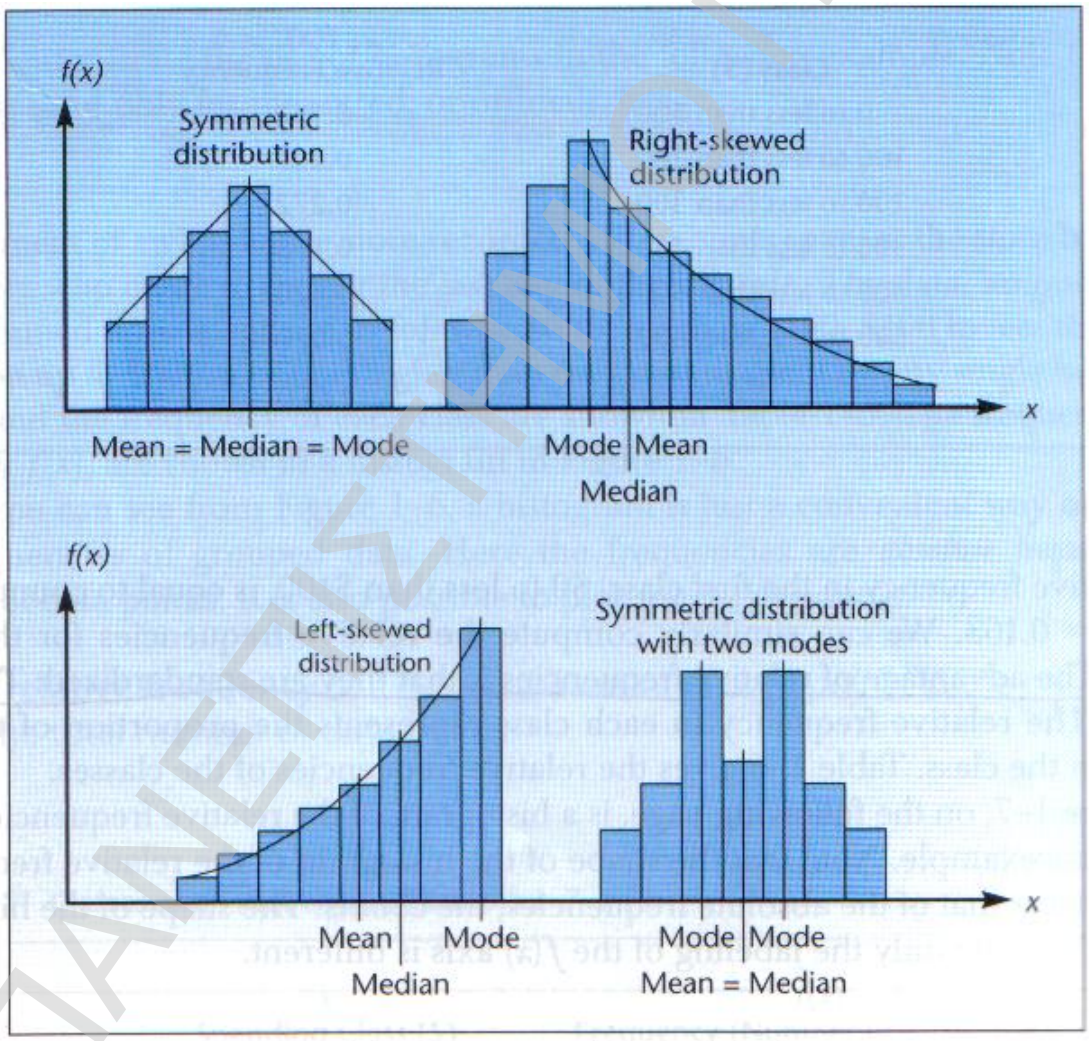
Η τρίτη βασική ιδιότητα ενός συνόλου αριθμητικών δεδομένων είναι το σχήμα που ακολουθεί η κατανομή τους, δηλαδή ο τρόπος που κατανέμονται τα δεδομένα. Οι δύο βασικές παράμετροι που καθορίζουν το σχήμα μιας κατανομής είναι η ασυμμετρία και η κύρτωση.

- ***<u>Ασυμμετρία</u>***

Μία κατανομή δεδομένων είναι συμμετρική ή ασυμμετρική (skew). Εάν τα δεδομένα κατανέμονται συμμετρικά, τότε όσο απομακρυνόμαστε από τη μέση τιμή (είτε προς

τα πάνω είτε προς τα κάτω) συναντάμε περίπου τον ίδιο αριθμό παρατηρήσεων. Για παράδειγμα, το ύψος των ανθρώπων κατανέμεται συμμετρικά, που σημαίνει ότι όσοι είναι οι ψηλοί τόσοι είναι περίπου και οι κοντοί.

Για να περιγράψουμε το σχήμα της κατανομής και να προσδιορίσουμε το είδος της συμμετρίας, αρκεί να συγκρίνουμε τον αριθμητικό μέσο με τη διάμεσο. Εάν ο μέσος είναι περίπου ίσος με τη διάμεσο, τότε θεωρούμε ότι τα δεδομένα κατανέμονται κανονικά, δηλαδή έχουμε μηδενική συμμετρία. Αντίθετα, εάν ο μέσος είναι μεγαλύτερος της διαμέσου, τότε τα δεδομένα χαρακτηρίζονται ως ασυμμετρικά, και η ασυμμετρία (skewness) ονομάζεται θετική. Και στην περίπτωση που ο μέσος είναι μικρότερος της διαμέσου η ασυμμετρία ονομάζεται αρνητική. Γενικότερα, όταν η κατανομή "απλώνεται" περισσότερο προς τα δεξιά από ό,τι στα αριστερά, τότε λέμε ότι υπάρχει θετική ασυμμετρία (ή ασυμμετρία προς τα δεξιά), ενώ όταν "απλώνεται" περισσότερο προς τα αριστερά, τότε υπάρχει αρνητική ασυμμετρία (ή ασυμμετρία προς τα αριστερά). Στο Σχήμα 2.6 συγκρίνονται διάφορες κατανομές ως προς την ασυμμετρία τους.



Σχήμα 2.6: Ασυμμετρία των Κατανομών

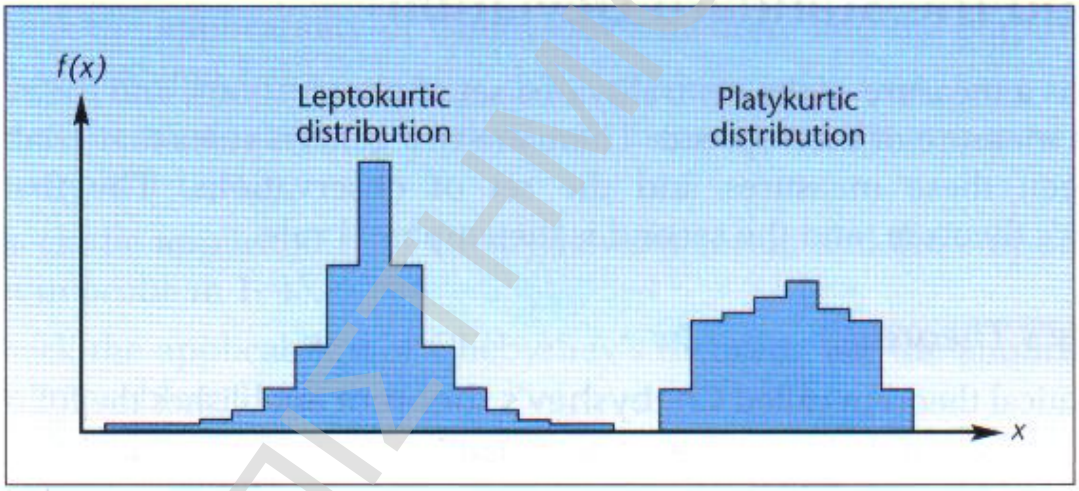
- ***Κύρτωση***

Ακόμα και αν δύο κατανομές έχουν τον ίδιο μέσο, τυπική απόκλιση και ασυμμετρία, μπορούν και πάλι να εμφανίζουν μεγάλη διαφορά στο σχήμα τους. Για το λόγο αυτό, πρέπει να ελέγχουμε την κύρτωση (kurtosis). Η κύρτωση εκφράζει το σχήμα της κορυφής της κατανομής. Όσο μεγαλύτερη είναι η κύρτωση, τόσο μεγαλύτερη είναι η κορυφή της κατανομής.

Η κύρτωση μπορεί να υπολογιστεί είτε ως απόλυτη είτε ως σχετική τιμή. Η απόλυτη κύρτωση έχει πάντα θετική τιμή. Η απόλυτη κύρτωση μιας κανονικής κατανομής είναι 3. Με βάση την τιμή αυτή υπολογίζεται η σχετική κύρτωση:

$$\textit{Σχετική Κύρτωση} = \textit{Απόλυτη Κύρτωση} - 3$$

Η σχετική κύρτωση μπορεί να είναι αρνητική. Στην ουσία, η σχετική κύρτωση είναι αυτή που χρησιμοποιείται. Μια αρνητική τιμή αυτής φανερώνει ότι η κατανομή είναι πιο επίπεδη από την κανονική κατανομή, και ονομάζεται πλατύκυρτη. Η θετική κύρτωση δηλώνει μια κατανομή με μεγαλύτερη κορυφή από την κανονική κατανομή, η οποία ονομάζεται λεπτόκυρτη. Οι δύο αυτές περιπτώσεις αναπαρίστανται στο Σχήμα 2.7. [7, 10, 11, 25]



Σχήμα 2.7: Κύρτωση των Κατανομών

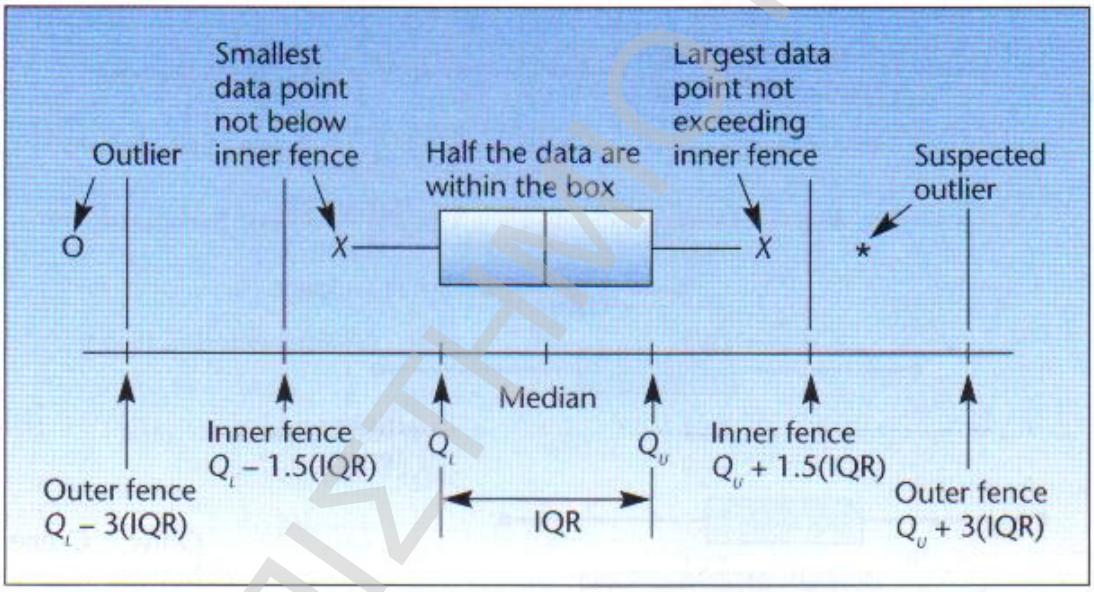
2.4. Διάγραμμα Πλαισίου-Απολήξεων (Box-and-Whisker Plot)

Τα διαγράμματα πλαισίου-απολήξεων (ή θηκογράμματα) αποτελούν μια στατιστική και γραφική τεχνική (Exploratory Data Analysis-EDA) που παρέχει τρόπους παρατήρησης των δεδομένων για να εντοπιστούν συσχετίσεις και τάσεις, και έκτροπες παρατηρήσεις. Επίσης, αποτελεί ένα τρόπο απόκτησης μιας γενικής εικόνας του συνόλου των δεδομένων, γρήγορα και εύκολα.

Το θηκόγραμμα αποτελεί ένα σύνολο πέντε βασικών μετρήσεων, οι οποίες είναι οι εξής:

1. Η διάμεσος των δεδομένων
2. Το κατώτερο τεταρτημόριο
3. Το ανώτερο τεταρτημόριο
4. Η μικρότερη παρατήρηση
5. Η μέγιστη παρατήρηση

Στο διάγραμμα αυτό, οι κάθετες πλευρές του πλαισίου (***hinges***) είναι στην ουσία το κατώτερο (Q_1) και το ανώτερο τεταρτημόριο (Q_3). Η κάθετη γραμμή εντός του πλαισίου είναι η διάμεσος. Οι οριζόντιες γραμμές καλούνται απολήξεις (*whiskers*) και ενώνουν το ανώτερο τεταρτημόριο με τη μεγαλύτερη παρατήρηση, και το κατώτερο τεταρτημόριο με τη μικρότερη παρατήρηση, εφόσον η μέγιστη και η ελάχιστη παρατήρηση βρίσκονται σε μια απόσταση μικρότερη από 1,5 φορές το ενδοτεταρτημοριακό εύρος από το αντίστοιχο τεταρτημόριο. Εάν κάποια παρατήρηση βρίσκεται μακρύτερα από την απόσταση αυτή θεωρείται ως *πιθανή* έκτροπη παρατήρηση, ενώ αν βρίσκεται σε απόσταση μεγαλύτερη από 3 φορές το ενδοτεταρτημοριακό εύρος θεωρείται ως έκτροπη. Χαρακτηριστικό είναι το Σχήμα 2.8.



Σχήμα 2.8: Διάγραμμα Πλαισίου-Απολήξεων

Το διάγραμμα πλαισίου-απολήξεων χρησιμεύει για τους ακόλουθους σκοπούς:

- Αναγνωρίζει τη θέση ενός σετ δεδομένων μέσω της διαμέσου.
- Αναγνωρίζει τη διασπορά των δεδομένων από το μήκος του πλαισίου (Q_3-Q_1) και το μήκος των οριζόντιων γραμμών (απολήξεις).
- Αναγνωρίζει πιθανή ασυμμετρία της κατανομής του σετ δεδομένων. Εάν το κομμάτι του πλαισίου στα δεξιά από τη διάμεσο είναι μεγαλύτερο από το κομμάτι αριστερά της διαμέσου ή εάν η δεξιά οριζόντια γραμμή είναι μακρύτερη από την αριστερή, τότε τα δεδομένα παρουσιάζουν ασυμμετρία προς τα δεξιά (και αντίστροφα). Εάν το πλαίσιο και οι απολήξεις είναι συμμετρικές, τα δεδομένα είναι κανονικά κατανεμημένα, με μηδενική ασυμμετρία.
- Αναγνωρίζει πιθανές ή πραγματικές έκτροπες παρατηρήσεις.

- Συγκρίνει δύο ή περισσότερα σετ δεδομένων, σχηματίζοντας ένα θηκόγραμμα για κάθε σετ δεδομένων και παραθέτοντας αυτά στην ίδια κλίμακα. [11, 47]

2.5. Η Κανονική Κατανομή

Η κανονική κατανομή (normal distribution) αναφέρεται σε συνεχείς μεταβλητές και αποτελεί την πιο σημαντική κατανομή πιθανοτήτων της στατιστικής για τρεις κυρίως λόγους:

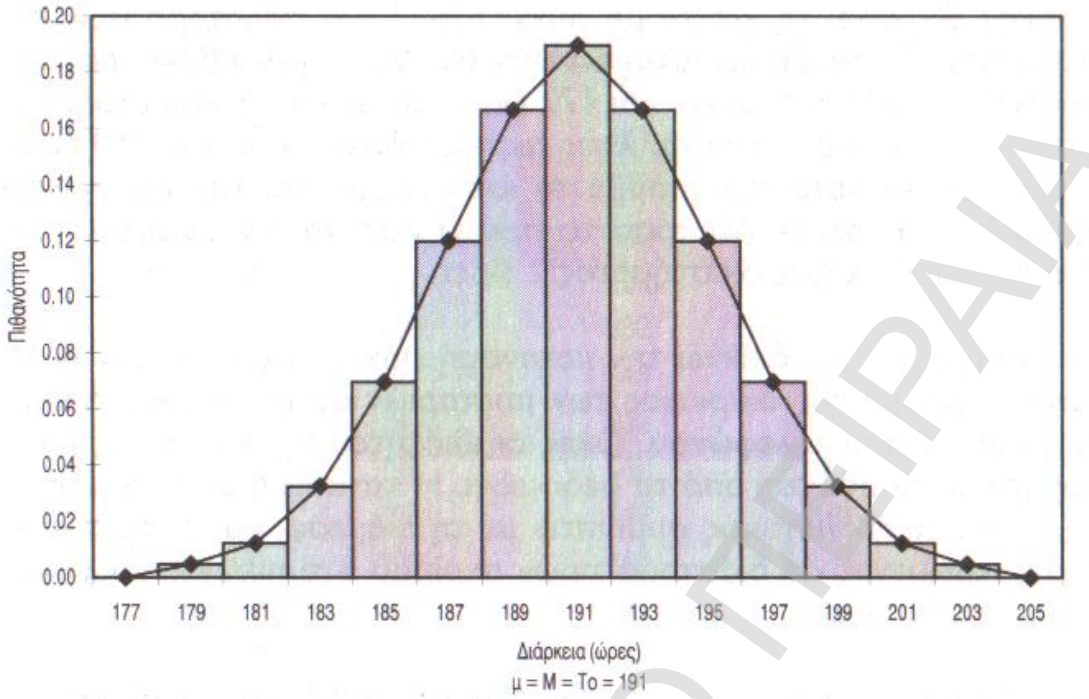
1. Τα περισσότερα συνεχή φαινόμενα ακολουθούν είτε με μεγάλη ακρίβεια είτε με μεγάλη προσέγγιση την κανονική κατανομή.

2. Με την κανονική κατανομή μπορούμε να προσεγγίσουμε πολλές ασυνεχείς κατανομές πιθανοτήτων.

3. Η κανονική κατανομή αποτελεί τη βάση της επαγωγικής στατιστικής.

Τα βασικά χαρακτηριστικά της κανονικής κατανομής είναι ότι ακολουθεί το σχήμα της "καμπάνας" και γι' αυτό είναι συμμετρική. Αποτέλεσμα αυτού είναι ο μέσος να είναι ίσος με τη διάμεσο και την επικρατούσα τιμή (σημείο μέγιστης συχνότητας). Δηλαδή, όλες οι τιμές των παραμέτρων της κεντρικής τάσης ή θέσης συμπίπτουν.

Στην πράξη οι μεταβλητές που παρατηρούμε, με βάση ένα δείγμα παρατηρήσεων, δεν συμπίπτουν απόλυτα με την κανονική κατανομή, ακόμα κι αν ο πληθυσμός από τον οποίο προέρχονται ακολουθεί τον κανονικό νόμο. Αυτό είναι λογικό, αφού η κανονική κατανομή είναι μια συνεχής κατανομή και απαιτούνται πάρα πολλές παρατηρήσεις για να δώσει το κατάλληλο σχήμα της καμπάνας. Στην ουσία, για ένα χαρακτηριστικό που ακολουθεί την κανονική κατανομή περιμένουμε οι παρατηρήσεις του δείγματος να δώσουν ένα συμμετρικό ιστόγραμμα, και το εύρος των παρατηρήσεων του δείγματος να βρίσκεται περίπου μεταξύ 3 τυπικών αποκλίσεων γύρω από τον αριθμητικό μέσο, δηλαδή το συνολικό εύρος των τιμών να είναι περίπου 6 τυπικές αποκλίσεις. Μία χαρακτηριστική κανονική κατανομή φαίνεται στο Σχήμα 2.9.



Σχήμα 2.9: Η Κανονική Κατανομή

Στις συνεχείς μεταβλητές δεν υπάρχουν απλές πιθανότητες, αλλά αυτό που γνωρίζουμε ή προσπαθούμε να προσεγγίσουμε είναι η μαθηματική συνάρτηση της κατανομής f(X). Η συνάρτηση μιας συνεχούς μεταβλητής ονομάζεται **συνάρτηση πυκνότητας πιθανότητας** (probability density function). Δηλαδή, είναι η συνάρτηση που εκφράζει την πυκνότητα (συχνότητα) εμφάνισης των παρατηρήσεων στα διάφορα διαστήματα τιμών. Η μαθηματική έκφραση της συνάρτησης πυκνότητας πιθανότητας της κανονικής κατανομής είναι η εξής:

$$f(X) = \frac{1}{s\sqrt{2p}} e^{-\frac{1}{2}(\frac{X-m}{s})^2}, \quad -\infty < X < +\infty$$

όπου

e = η βάση των νεπέρειων λογαρίθμων (περίπου 2,71828)
π = η γνωστή μαθηματική σταθερά (περίπου 3,14159)
μ = ο μέσος του πληθυσμού
σ = η τυπική απόκλιση του πληθυσμού
X = μια τιμή της συνεχούς τυχαίας μεταβλητής στο διάστημα $-\infty$ έως $+\infty$

Από την παραπάνω σχέση προκύπτει ότι οι πιθανότητες της τυχαίας μεταβλητής X εξαρτώνται μόνο από τις δύο παραμέτρους του πληθυσμού μ και σ. Κάθε φορά που ορίζουμε ένα συγκεκριμένο συνδυασμό μ και σ έχουμε και μια διαφορετική κανονική κατανομή. Αυτό σημαίνει ότι υπάρχουν άπειρες κανονικές κατανομές, αφού άπειρα είναι και τα χαρακτηριστικά (μεταβλητές) που ακολουθούν τον κανονικό νόμο. Έτσι, προέκυψε η λύση να τυποποιήσουμε τα δεδομένα και να χρησιμοποιήσουμε μια μόνο κατανομή, την **τυποποιημένη κανονική κατανομή**.

Η τυποποίηση των δεδομένων βασίζεται στην απόκλισή τους από το μέσο σε όρους της σ, σύμφωνα με τον τύπο:

$$Z = \frac{X - m}{s}$$

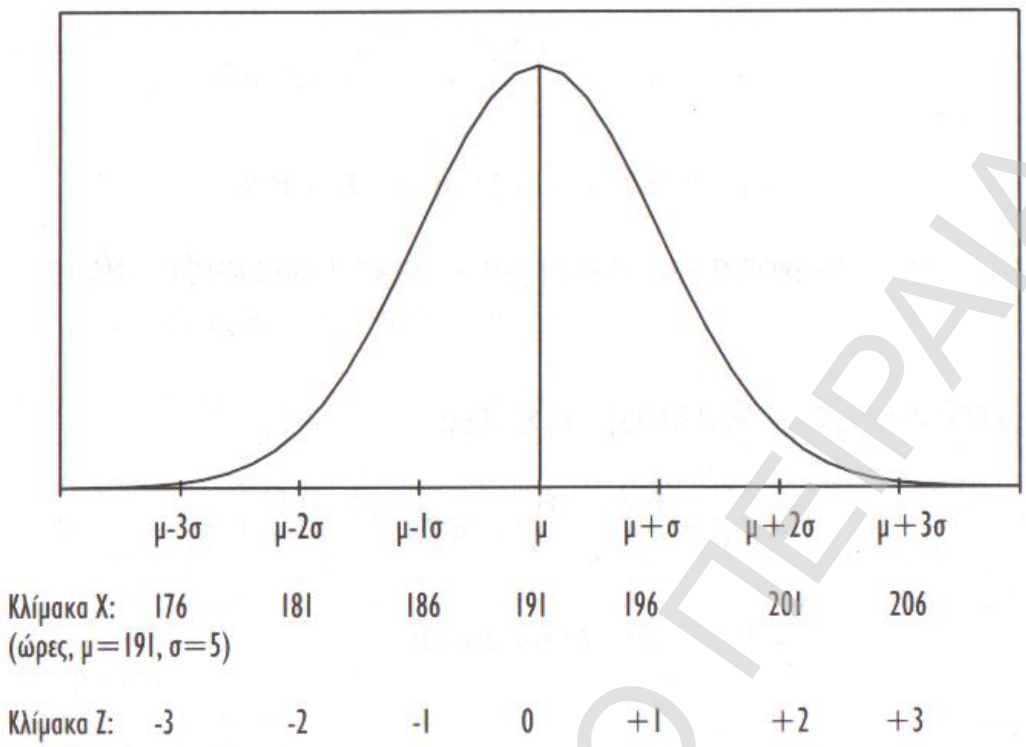
Η παραπάνω σχέση δημιουργεί μια νέα μεταβλητή, τη *Z*, που ονομάζεται τυποποιημένη κανονική μεταβλητή. Η *Z* κατανέμεται και αυτή κανονικά και έχει τις εξής ιδιότητες:

- Είναι ανεξάρτητη από τη μονάδα μέτρησης της μεταβλητής *X*, αφού δεν εκφράζεται σε καμία μονάδα μέτρησης.
- Έχει μέσο ίσο με μηδέν.
- Έχει διακύμανση ίση με τη μονάδα, που σημαίνει $\sigma_Z = 1$.

Η συνάρτηση πυκνότητας πιθανότητας της τυποποιημένης μεταβλητής *Z* είναι:

$$f(Z) = \frac{1}{\sqrt{2p}} e^{-\frac{1}{2}z^2}$$

και ονομάζεται τυποποιημένη κανονική κατανομή (standardized normal distribution). Με τον τύπο $Z = \frac{X - m}{s}$ μπορούμε να μετατρέψουμε τα δεδομένα μιας μεταβλητής που κατανέμεται κανονικά σε τυποποιημένη μορφή, και να υπολογίσουμε τις πιθανότητες χρησιμοποιώντας τους πίνακες της τυποποιημένης κανονικής κατανομής. Το Σχήμα 2.10 δείχνει την αντιστοιχία μεταξύ της μεταβλητής *X* που αναφέρθηκε στο παράδειγμα του Σχήματος 2.9 και της τυποποιημένης μεταβλητής *Z*.



Σχήμα 2.10: Η Τυποποιημένη Κανονική Κατανομή

Στο παραπάνω σχήμα βλέπουμε ότι για κάθε τιμή της μεταβλητής *X* αντιστοιχεί μια τυποποιημένη τιμή της *Z*. Έτσι, αν θέλουμε να υπολογίσουμε την πιθανότητα *P(X<181)*, είναι η ίδια με την πιθανότητα *P(Z<-2)*. Για τον υπολογισμό των πιθανοτήτων της κατανομής της *Z* χρησιμοποιούμε τους πίνακες της τυποποιημένης κανονικής κατανομής που υπάρχουν σε όλα τα βιβλία στατιστικής.

2.6. Έλεγχος Υποθέσεων

Στη στατιστική χρησιμοποιείται ευρύτατα ο έλεγχος υποθέσεων. Η υπόθεση είναι κάτι το οποίο δεν έχει ακόμα αποδειχθεί ότι είναι σωστό. Ο έλεγχος υποθέσεων είναι η διαδικασία που καθορίζει εάν μια υπόθεση είναι αληθής ή όχι. Τις περισσότερες φορές, η υπόθεση ελέγχεται ως προς τον αριθμητικό μέσο.

2.6.1. Η Μηδενική Υπόθεση (null hypothesis)

Το πρώτο βήμα στον έλεγχο υποθέσεων είναι ο καθορισμός αυτού μέσω της μηδενικής υπόθεσης. Η μηδενική υπόθεση είναι ένας ισχυρισμός για την τιμή κάποιας παραμέτρου ενός πληθυσμού. Είναι ένας ισχυρισμός ότι αυτό που δηλώνουμε ισχύει, μέχρι να έχουμε αρκετές στατιστικές αποδείξεις που οδηγούν στο αντίθετο συμπέρασμα.

Για παράδειγμα, μια μηδενική υπόθεση μπορεί να ισχυρίζεται ότι ο μέσος ενός συγκεκριμένου πληθυσμού έχει την τιμή 100. Εκτός κι αν αποκτήσουμε αρκετές

αποδείξεις ότι δεν είναι 100, θα δεχόμαστε πάντα ότι είναι 100. Αυτή η μηδενική υπόθεση διατυπώνεται ως εξής:

$$\mathbf{H_o: \mu = 100}$$

όπου το σύμβολο H_o υποδηλώνει τη μηδενική υπόθεση.

Η εναλλακτική υπόθεση είναι η άρνηση (αναίρεση) της μηδενικής υπόθεσης. Δηλαδή, για το παραπάνω παράδειγμα της H_o: μ = 100, η εναλλακτική υπόθεση δηλώνεται ως εξής:

$$\mathbf{H_1: \mu \neq 100}$$

Επειδή η μηδενική και η εναλλακτική υπόθεση ισχυρίζονται ακριβώς αντίθετα πράγματα, μόνο η μία από αυτές μπορεί να είναι αληθής. Η απόρριψη της μίας ταυτόχρονα σημαίνει την αποδοχή της άλλης.

Υποθέσεις μπορούν να γίνουν και με άλλες παραμέτρους, όπως την αναλογία (πιθανότητα) του πληθυσμού ή τη διασπορά αυτού. Επίσης, οι υποθέσεις μπορούν να διατυπωθούν όχι μόνο με ισότητες, αλλά και με ανισότητες. Πρέπει, όμως, να τονιστεί ότι το σύμβολο της ισότητας εμφανίζεται πάντα στη μηδενική υπόθεση.

Αν και η ιδέα της μηδενικής υπόθεσης είναι απλή, ο καθορισμός αυτής για μια συγκεκριμένη περίπτωση μπορεί να είναι δύσκολος. Είναι σημαντικό η μηδενική υπόθεση να διατυπωθεί σωστά και σαφώς, αλλιώς το τεστ δεν θα έχει νόημα. Ένας τρόπος για να διατυπώσουμε σωστά τη μηδενική υπόθεση είναι να έχουμε στο μυαλό μας ότι εάν η μηδενική υπόθεση είναι αληθής, τότε δεν χρειάζεται να γίνει καμία διορθωτική ενέργεια. Δηλαδή, η μηδενική υπόθεση πρέπει να διατυπώνεται με τον πλέον *καλοπροαίρετο* τρόπο, που ελαχιστοποιεί τις οποιεσδήποτε αμφιβολίες και ενέργειες που θα ακολουθήσουν.

2.6.2. Συγκέντρωση Αποδείξεων

Αναφέρθηκε προηγουμένως ότι θεωρούμε πάντα αληθή τη μηδενική υπόθεση, μέχρις αποδείξεως του αντιθέτου, αφού έχουν συγκεντρωθεί επαρκείς αποδείξεις. Οι αποδείξεις αυτές σίγουρα μπορούν να θεωρηθούν επαρκείς, όταν ελέγξουμε *όλον* τον πληθυσμό και υπολογίσουμε την ακριβής τιμή της παραμέτρου που εξετάζουμε. Με τον τρόπο αυτό θα είμαστε 100% σίγουροι για το αν ισχύει ή όχι η μηδενική υπόθεση. Αυτό βέβαια στην πράξη είναι πολύ δύσκολο ή και αδύνατο να γίνει, ενώ δεν συμφέρει οικονομικά. Γι' αυτό, οι αποδείξεις αυτές μπορούν να συγκεντρωθούν μόνο από ένα τυχαίο δείγμα του πληθυσμού.

Ένας σημαντικός περιορισμός στο να βγάζουμε συμπεράσματα από ένα δείγμα του πληθυσμού είναι ότι δεν μπορούμε να είμαστε 100% σίγουροι. Το κατά πόσο σίγουροι θα είμαστε εξαρτάται από το μέγεθος του δείγματος -θα πρέπει να είναι αρκετά μεγάλο ώστε να μας δώσει ικανοποιητικά διαστήματα εμπιστοσύνης και αρκετά μικρό ώστε να μειωθεί το κόστος- και από παραμέτρους όπως η διασπορά του πληθυσμού.

2.6.3. Σφάλματα Τύπου I και II

Στους στατιστικούς ελέγχους υποθέσεων, δεν μπορούμε να είμαστε απόλυτα σίγουροι για το αν η μηδενική υπόθεση αληθεύει ή όχι. Για το λόγο αυτό, υπάρχουν δύο τύποι σφαλμάτων που είναι πιθανό να εμφανιστούν σε αυτούς τους ελέγχους: το σφάλμα τύπου I εμφανίζεται όταν απορρίπτουμε μία μηδενική υπόθεση που αληθεύει, ενώ το σφάλμα τύπου II όταν αποδεχόμαστε μία λανθασμένη μηδενική υπόθεση (Σχήμα 2.11).

	H_o Αληθής	**H_o Λανθασμένη**
Αποδοχή H_o	Κανένα σφάλμα	Σφάλμα Τύπου II
Απόρριψη H_o	Σφάλμα Τύπου I	Κανένα σφάλμα

Σχήμα 2.11: Σφάλματα Τύπου I και II

Εάν πάντα δεχόμαστε τη μηδενική υπόθεση ως σωστή, τότε δεν θα εμφανιστεί ποτέ η πιθανότητα να απορρίψουμε μια σωστή μηδενική υπόθεση. Άρα, δεν θα διαπραχθεί ποτέ το σφάλμα τύπου I. Αυτό, όμως, αυτόματα σημαίνει ότι εάν δεχόμαστε πάντα τη μηδενική υπόθεση, τότε θα δεχθούμε σίγουρα και μία λανθασμένη υπόθεση. Δηλαδή, διαπράττουμε σφάλμα τύπου II. Γι' αυτό, πρέπει να βρούμε μια άριστη λύση που θα εξισορροπεί τις συνέπειες που προκύπτουν από κάθε τύπο σφάλματος, με κύριο κριτήριο το κόστος των συνεπειών αυτών.

2.6.4. P-value

Έστω ότι η μηδενική και η εναλλακτική υπόθεση είναι οι ακόλουθες:

$$H_o: \mu \geq 1.000$$
$$H_1: \mu < 1.000$$

Ένα τυχαίο δείγμα μεγέθους 30 δίνει ένα μέσο με τιμή 999. Καθώς ο μέσος του δείγματος είναι μικρότερος από 1.000, οι αποδείξεις φαίνεται ότι ευνοούν την απόρριψη της μηδενικής υπόθεσης. Όμως, δεν είμαστε σίγουροι για το αν πρέπει να απορριφθεί η H_o, καθώς η διαφορά του μέσου από το επιθυμητό όριο είναι ελάχιστη. Το ζήτημα που προκύπτει είναι κατά πόσο η H_o είναι αξιόπιστη (αληθεύει), ακόμα και όταν υπάρχουν αποδείξεις που δεν την ευνοούν. Επειδή η πιθανότητα η H_o να αληθεύει είναι δύσκολο να υπολογιστεί μαθηματικά, θέτουμε το ερώτημα:

Όταν ο πραγματικός μέσος είναι $\mu=1.000$, και για μέγεθος δείγματος ίσο με 30, ποια είναι η πιθανότητα να προκύψει ένας μέσος του δείγματος που είναι μικρότερος ή ίσος με την τιμή 999;

Η παραπάνω πιθανότητα ονομάζεται ***p-value*** και παίρνει τιμές από 0 έως 1. Εάν είναι 0, τότε σημαίνει ότι η H_o είναι σίγουρα λάθος, ενώ αν είναι 1 η H_o είναι σίγουρα σωστή. Μία τυχαία τιμή ***p-value*** ίση με 30% σημαίνει ότι υπάρχει περίπου 30% πιθανότητα ότι η H_o αληθεύει παρά τις αντίθετες ενδείξεις, ή ότι υπάρχει περίπου 70% πιθανότητα η H_o να είναι όντως λανθασμένη, όπως δείχνουν οι αποδείξεις. Δηλαδή, το ***p-value*** ορίζεται ως *η πιθανότητα να προκύψουν αποδείξεις από ένα δείγμα μεγέθους **n** που δεν ευνοούν τον ισχυρισμό της H_o, όταν αυτή (H_o) είναι όντως αληθής*.

2.6.5. Επίπεδο Σημαντικότητας

Η πιο κοινή τακτική στους στατιστικούς ελέγχους υποθέσεων είναι ο καθορισμός ενός επιπέδου εμπιστοσύνης, α, και η απόρριψη της H_o όταν το ***p-value*** είναι κάτω από αυτό. Δηλαδή, όταν το ***p-value*** είναι μικρότερο από α, απορρίπτουμε την H_o.

Οι συνήθεις τιμές του α είναι 10%, 5% και 1%. Για παράδειγμα, εάν α=5%, τότε όποτε το ***p-value*** είναι μικρότερο από 5% θα απορρίπτουμε την H_o. Ωστόσο, εάν το ***p-value*** είναι 6% δεν απορρίπτουμε την H_o, χωρίς αυτό να σημαίνει ότι η H_o είναι αληθής. Για το λόγο αυτό, λέμε ότι δεν μπορούμε να απορρίψουμε την H_o για α=5%, και όχι ότι δεχόμαστε την H_o για α=5%.

Το επίπεδο σημαντικότητας α είναι η μέγιστη πιθανότητα του σφάλματος τύπου Ι που εμείς θέτουμε, καθώς είναι το μέγιστο ***p-value*** στο οποίο απορρίπτουμε την H_o. Με άλλα λόγια, θέτοντας α=5% σημαίνει ότι δεχόμαστε μέχρι 5% πιθανότητα να διαπράξουμε σφάλμα τύπου Ι. Επίσης, το α έμμεσα καθορίζει και την πιθανότητα του σφάλματος τύπου ΙΙ. Εάν δεχθούμε ότι α=0, τότε μειώνεται μεν η πιθανότητα του σφάλματος τύπου Ι, αλλά αυξάνεται η πιθανότητα σφάλματος τύπου ΙΙ και παίρνει την τιμή 1. Για να μειώσουμε την πιθανότητα του σφάλματος τύπου ΙΙ πρέπει να αυξήσουμε το α, σε τέτοιο βαθμό που να μην είναι αρκετά μεγάλο ώστε να αυξάνεται σημαντικά η πιθανότητα σφάλματος τύπου Ι. [10, 11]

ΚΕΦΑΛΑΙΟ 3: ΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ ΚΑΙ ΣΥΣΧΕΤΙΣΗ

Στις περιπτώσεις που το ενδιαφέρον επικεντρώνεται στην εξέταση της συμπεριφοράς ενός χαρακτηριστικού, δηλαδή μιας τυχαίας μεταβλητής, τότε χρησιμοποιούμε τις γνωστές στατιστικές μεθόδους της εκτιμητικής και της στατιστικής επαγωγής: εκτίμηση παραμέτρων, εκτίμηση διαστήματος εμπιστοσύνης, έλεγχος υποθέσεων κλπ. Υπάρχουν, όμως, πολλές περιπτώσεις που οι αποφάσεις βασίζονται στη σχέση που πιθανόν να υπάρχει μεταξύ δύο ή περισσοτέρων μεταβλητών. Εάν αυτή η σχέση μπορεί να περιγραφεί επιτυχώς, τότε τα οφέλη είναι σημαντικά, καθώς μπορεί να γίνει πρόβλεψη και εκτίμηση της μιας από τις δυο μεταβλητές, εάν η άλλη είναι γνωστή.

Παρακάτω αναλύονται η ανάλυση της απλής γραμμικής συσχέτισης και παλινδρόμησης. Η ανάλυση της συσχέτισης χρησιμοποιείται στο πρώτο στάδιο της ανάλυσης, προκειμένου να διαπιστωθεί εάν υπάρχει στατιστική σχέση μεταξύ δύο μεταβλητών. Εάν η ανάλυση οδηγήσει στο συμπέρασμα ότι οι δύο μεταβλητές συσχετίζονται, τότε προχωρούμε στο δεύτερο στάδιο της ανάλυσης παλινδρόμησης, όπου θα περιγράψουμε αυτή τη σχέση.

3.1. Συντελεστής Συσχέτισης και Διαγράμματα Διασποράς

Η συσχέτιση μεταξύ δύο τυχαίων μεταβλητών είναι ένα μέτρο της γραμμικής σχέσης που υπάρχει μεταξύ αυτών. Ουσιαστικά, αποτελεί μια ένδειξη για το πόσο καλά οι δύο μεταβλητές κινούνται μαζί σε ευθεία γραμμή. Η συσχέτιση μεταξύ των *X* και *Y* είναι η ίδια με τη συσχέτιση μεταξύ των *Y* και *X*. Συνεπώς, ως συσχέτιση μεταξύ δύο τυχαίων μεταβλητών *X* και *Y* ορίζεται το μέτρο του βαθμού της γραμμικής συσχέτισης μεταξύ των δύο μεταβλητών. Πρέπει να τονιστεί, ωστόσο, ότι στην ανάλυση συσχέτισης υποθέτουμε ότι και οι δύο μεταβλητές *X* και *Y* ακολουθούν την κανονική κατανομή, με μέσους m_X και m_Y και τυπικές αποκλίσεις s_X και s_Y, αντίστοιχα.

Η ποσοτική μέτρηση της έντασης της γραμμικής σχέσης μεταξύ δύο μεταβλητών δύο δειγμάτων ονομάζεται συντελεστής συσχέτισης, r (correlation coefficient). Η εκτίμηση του συντελεστή συσχέτισης προκύπτει ως εξής:

$$r = \frac{\Sigma(X - \overline{X})(Y - \overline{Y})}{\sqrt{\Sigma(X - \overline{X})^2 \Sigma(Y - \overline{Y})^2}}$$

όπου: r = εκτίμηση του απλού συντελεστή συσχέτισης του πληθυσμού
n = μέγεθος δείγματος (ζεύγη τιμών)
X = τιμές της ανεξάρτητης μεταβλητής
$\overline{X}$ = μέσος αριθμητικός της X
Y = τιμές της εξαρτημένης μεταβλητής
$\overline{Y}$ = μέσος αριθμητικός της Y

Ο παραπάνω συντελεστής συσχέτισης ονομάζεται επίσης και **συντελεστής συσχέτισης κατά Pearson**.

Οι πιθανές τιμές του r και η ερμηνεία αυτών δίνονται παρακάτω. Το πρόσημο του r δείχνει το είδος (την κατεύθυνση) της σχέσης.

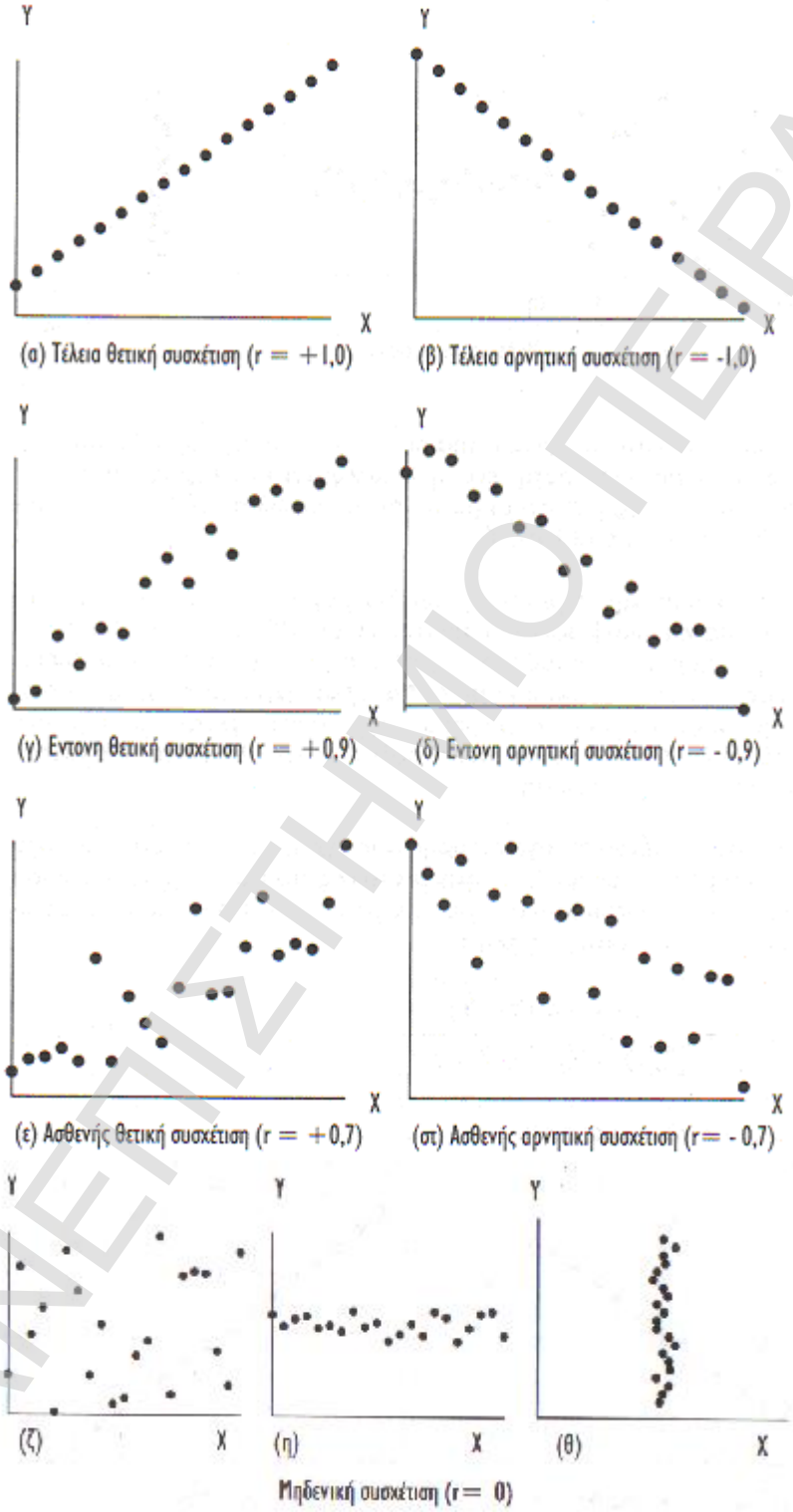
1. Όταν **r=0**, τότε δεν υπάρχει συσχέτιση. Δηλαδή, δεν υπάρχει γραμμική σχέση μεταξύ των δύο τυχαίων μεταβλητών.
2. Όταν **r=1**, υπάρχει τέλεια θετική γραμμική σχέση μεταξύ των δύο μεταβλητών. Αυτό σημαίνει, ότι όποτε η μία από τις δύο μεταβλητές *X* και *Y* αυξάνεται (ή μειώνεται), τότε και η άλλη μεταβλητή αυξάνεται (ή μειώνεται).
3. Όταν **r=-1**, υπάρχει τέλεια αρνητική γραμμική σχέση μεταξύ των δύο μεταβλητών. Αυτό σημαίνει, ότι όποτε η μία από τις δύο μεταβλητές *X* και *Y* αυξάνεται (ή μειώνεται), τότε η άλλη μεταβλητή μειώνεται (ή αυξάνεται).
4. Όταν η τιμή του **r** είναι **μεταξύ [0, 1] ή [-1, 0]**, τότε αυτό δηλώνει τη σχετική δύναμη της γραμμικής σχέσης μεταξύ των μεταβλητών. Για παράδειγμα, αν r=0,90 τότε υπάρχει μία σχετικά ισχυρή θετική σχέση μεταξύ των δύο μεταβλητών. Η τιμή r=-0,70 δηλώνει μια ασθενέστερη αρνητική γραμμική σχέση, ενώ η τιμή r=30 δηλώνει μια σχετικά ασθενή θετική γραμμική σχέση μεταξύ των *X* και *Y*.

Ένα άλλο χαρακτηριστικό του συντελεστή συσχέτισης είναι ότι δεν εκφράζεται σε καμία μονάδα μέτρησης, αλλά είναι καθαρός αριθμός, καθώς οι μονάδες μέτρησης του αριθμητή είναι ίδιες με τις μονάδες μέτρησης του παρονομαστή. Αυτό έχει το πλεονέκτημα ότι μπορούμε να συγκρίνουμε τους συντελεστές συσχέτισης για διαφορετικά ζεύγη μεταβλητών.

Ο πιο απλός τρόπος για να διαπιστώσουμε αν υπάρχει συσχέτιση μεταξύ δύο μεταβλητών είναι η κατασκευή του **διαγράμματος διασποράς**. Το διάγραμμα αυτό βασίζεται σε δύο ορθογώνιους άξονες που αντιστοιχούν στις δύο μεταβλητές: της εξαρτημένης μεταβλητής *Y* και της ανεξάρτητης *X*. Για κάθε ζεύγος τιμών (X_i, Y_i) αντιστοιχεί και ένα σημείο που συμβολίζεται με μια κουκκίδα, σταυρό ή κάποιο άλλο σύμβολο. Διευκρινίζεται ότι εξαρτημένη μεταβλητή *Y* είναι η μεταβλητή της οποίας τις μεταβολές θέλουμε να εξηγήσουμε. Ενώ, ανεξάρτητη *X* είναι η μεταβλητή που πιστεύουμε ότι επιδρά στην *Y*, προκαλεί τις μεταβολές της *Y* και επομένως χρησιμοποιείται για να εξηγήσουμε τη μεταβλητότητα που παρουσιάζει η *Y*.

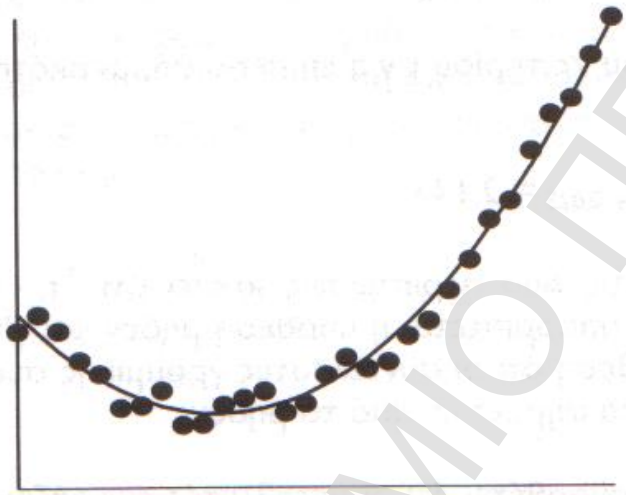
Όσο πιο συγκεντρωμένα είναι τα σημεία γύρω από μία ευθεία γραμμή, τόσο πιο δυνατή είναι η σχέση μεταξύ των δύο μεταβλητών. Χαρακτηριστικά διαγράμματα διασποράς φαίνονται στο Σχήμα 3.1. Όπως φαίνεται στο σχήμα αυτό, εάν τα σημεία βρίσκονται σε μια ευθεία γραμμή έχουμε τέλεια συσχέτιση (περιπτώσεις α και β), ενώ στις περιπτώσεις ζ, η και θ οι μεταβλητές *X* και *Y* δεν συσχετίζονται, δηλαδή δεν υπάρχει γραμμική σχέση εξάρτησης και η μία μεταβάλλεται ανεξάρτητα από την άλλη. Αυτό προκύπτει από τη συμπεριφορά των σημείων που είναι διάσπαρτα σε όλες τις περιοχές που καλύπτουν οι δύο άξονες, χωρίς να παρουσιάζουν καμία συστηματικότητα. Μια άλλη περίπτωση ανεξαρτησίας είναι το διάγραμμα διασποράς (η). Στην περίπτωση αυτή, ενώ η *X* καλύπτει ένα εύρος τιμών, η *Y* παραμένει σταθερή. Αυτό σημαίνει ότι η μεταβλητή *X* δεν έχει καμία επίδραση στη μεταβλητή *Y*. Τέλος, έχουμε την περίπτωση του διαγράμματος διασποράς (θ). Εδώ συμβαίνει

ακριβώς το αντίθετο. Η μεταβλητή *Y* μεταβάλλεται, ενώ η *X* παραμένει σταθερή. Δηλαδή, η μεταβλητή *Y* μεταβάλλεται ανεξάρτητα από την *X*.



<u>Σχήμα 3.1</u>: Διαγράμματα Διασποράς

Η ερμηνεία της τιμής του συντελεστή συσχέτισης χρειάζεται ιδιαίτερη προσοχή. Ο συντελεστής συσχέτισης εκφράζει την ένταση της σχέσης μεταξύ των μεταβλητών *X* και *Y*, μόνο όταν υπάρχει **γραμμική** συσχέτιση. Μια χαμηλή τιμή του r δεν σημαίνει πάντα ότι η σχέση είναι ασθενής. Οι μεταβλητές ενδέχεται να συσχετίζονται έντονα, αλλά η σχέση να είναι καμπυλόγραμμη, όπως δείχνει το Σχήμα 3.2. Η σχέση εξάρτησης είναι έντονη, ενώ ο συντελεστής συσχέτισης είναι μόλις 0,75. Αυτός είναι ένας ακόμη λόγος που αποδεικνύει τη χρησιμότητα του διαγράμματος διασποράς. Τέτοιες συσχετίσεις μπορούν να αναγνωριστούν έγκαιρα με την απλή απεικόνιση των σημείων.



Σχήμα 3.2: Διάγραμμα Διασποράς Καμπυλόγραμμης Σχέσης (r = 0,75)

Ένα άλλο σημείο που έχει ιδιαίτερη σημασία όταν ερμηνεύεται η τιμή του συντελεστή συσχέτισης, είναι η σχέση αιτίου-αποτελέσματος (causality) μεταξύ των μεταβλητών. Μια υψηλή τιμή του r δεν σημαίνει ότι υπάρχει σχέση αιτίας και αποτελέσματος μεταξύ των μεταβλητών. Επίσης, δεν σημαίνει ότι οι μεταβλητές σχετίζονται με άμεσο και ουσιώδη τρόπο. Για παράδειγμα, εάν βρούμε ότι δύο μεταβλητές αυξάνονται "ταυτόχρονα", αυτό μπορεί απλώς να είναι αποτέλεσμα της αύξησης μιας άλλης τρίτης μεταβλητής, και όχι της συσχέτισης μεταξύ των δύο συγκεκριμένων μεταβλητών. Γι' αυτό, πρέπει να ελεγχθεί η ύπαρξη άλλων μεταβλητών που επηρεάζουν και τις δύο μεταβλητές που μελετώνται [10, 11].

3.2. Απλή Γραμμική Παλινδρόμηση

3.2.1. Υπόδειγμα και Προϋποθέσεις Γραμμικής Παλινδρόμησης

Από την ανάλυση συσχέτισης που αναφέρθηκε στην προηγούμενη ενότητα μπορούμε να ελέγξουμε αν υπάρχει γραμμική σχέση μεταξύ δύο τυχαίων μεταβλητών. Για να μπορέσουμε να εκτιμήσουμε την εξίσωση που περιγράφει τη σχέση μεταξύ της εξαρτημένης μεταβλητής *Y* και της ανεξάρτητης μεταβλητής *X*, χρησιμοποιούμε μια συγκεκριμένη στατιστική μέθοδο: την ανάλυση παλινδρόμησης. Επειδή υπάρχουν μόνο δύο μεταβλητές, την καλούμε ανάλυση απλής παλινδρόμησης. Επιπλέον, επειδή η σχέση μεταξύ της εξαρτημένης και ανεξάρτητης μεταβλητής είναι γραμμική, η

μέθοδος ονομάζεται **απλή γραμμική παλινδρόμηση**. Ο σκοπός της απλής γραμμικής παλινδρόμησης είναι να περιγράψει τη σχέση μεταξύ των *X* και *Y* με ένα υπόδειγμα (μοντέλο), που έχει την εξής μορφή:

$$\boldsymbol{Y_i = \beta_o + \beta_1 X_i + \varepsilon_i}$$

όπου: Y_i = η τιμή της εξαρτημένης μεταβλητής
X_i = η τιμή της ανεξάρτητης μεταβλητής
β_o = το σημείο τομής του άξονα της *Y* από τη γραμμή παλινδρόμησης
β_1 = η κλίση της γραμμής παλινδρόμησης
ε_i = σφάλμα ή κατάλοιπο, δηλαδή η διαφορά μεταξύ της πραγματικής τιμής της *Y* και της τιμής της πρόβλεψης που προκύπτει από το υπόδειγμα

Το υπόδειγμα της απλής γραμμικής παλινδρόμησης βασίζεται στις ακόλουθες **βασικές προϋποθέσεις**:

1. Η σχέση μεταξύ των *X* και *Y* είναι γραμμική.
2. Οι τιμές της ανεξάρτητης μεταβλητής *X* θεωρούνται σταθερές (όχι τυχαία μεταβλητή). Οποιοσδήποτε τυχαίος παράγοντας επηρεάζει τις τιμές της *Y* θεωρείται ότι προέρχεται από το σφάλμα ε.
3. Τα σφάλματα ε ακολουθούν κανονική κατανομή, με μέσο μηδέν και σταθερή διακύμανση σ^2. Τα σφάλματα δεν συσχετίζονται μεταξύ τους σε διαδοχικές παρατηρήσεις.

Ωστόσο, ισχύει η ακόλουθη σχέση:

$$Y_i = \underbrace{b_o + b_1 X_i}_{\substack{\text{μη τυχαίος}\\ \text{παράγοντας}}} + \underbrace{e_i}_{\substack{\text{τυχαίος}\\ \text{παράγοντας}}}$$

Συνεπώς, δύο άμεσες συνέπειες των παραπάνω υποθέσεων είναι ότι κάθε παρατήρηση Y_i έχει μέση τιμή

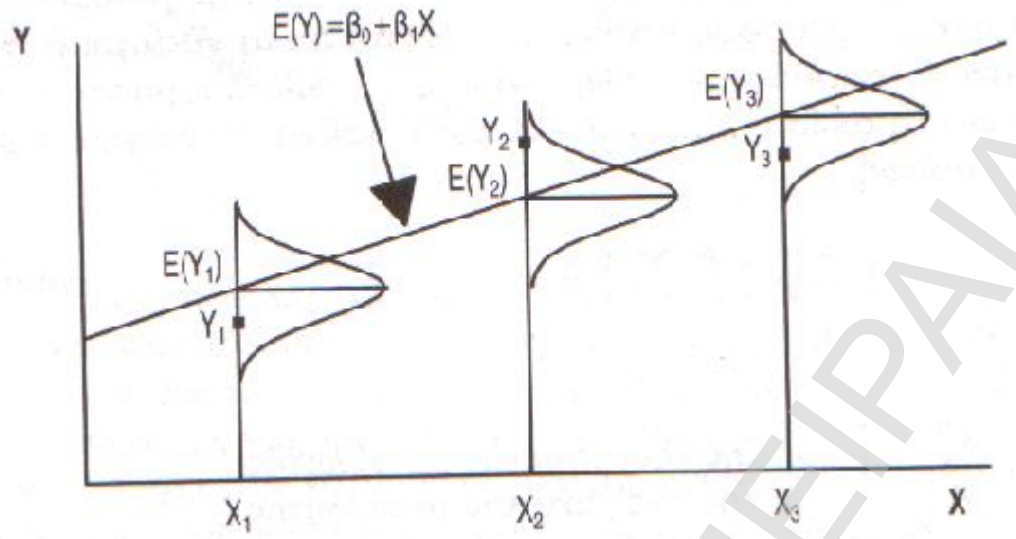
$$E(Y_i) = \mu_{Y|X} = \beta_o + \beta_1 X_i,$$

και διακύμανση

$$V(X_i) = \sigma^2,$$

και οι Y_i, Y_j είναι ασυσχέτιστες για κάθε ζεύγος i, j με $i \neq j$.

Το Σχήμα 3.3 απεικονίζει τις παραπάνω υποθέσεις.



<u>Σχήμα 3.3</u>: Διαγραμματική Απεικόνιση των Υποθέσεων της Γραμμικής Παλινδρόμησης

Η ευθεία γραμμή της παλινδρόμησης συνδέει τους μέσους της μεταβλητής *Y* που αντιστοιχούν στις τιμές της *X*. Όπως συμβαίνει με κάθε ευθεία γραμμή, έτσι και η γραμμή της παλινδρόμησης προσδιορίζεται από δύο παραμέτρους: τις β_o και β_1, που ονομάζονται **συντελεστές παλινδρόμησης**. Η β_o δίνει το σημείο που τέμνει η γραμμή παλινδρόμησης τον άξονα της *Y* και ονομάζεται τεταγμένη στην αρχή μηδέν. Είναι η προβλεπόμενη από το υπόδειγμα τιμή της *Y* για *X*=0. Ο συντελεστής β_1 δίνει την κλίση της γραμμής παλινδρόμησης. Είναι η μέση μεταβολή της εξαρτημένης μεταβλητής *Y* που αντιστοιχεί σε μεταβολή της *X* κατά μία μονάδα.

3.2.2. <u>Εκτίμηση της Εξίσωσης Παλινδρόμησης: Μέθοδος Ελαχίστων Τετραγώνων</u>

Σκοπός μας είναι να εκτιμήσουμε τις παραμέτρους του μοντέλου της παλινδρόμησης, δηλαδή τους συντελεστές β_o και β_1, κατά τέτοιο τρόπο, ώστε η ευθεία γραμμή που θα προκύψει να περιγράφει κατά τον καλύτερο δυνατό τρόπο τη σχέση μεταξύ των μεταβλητών *X* και *Y*. Η γραμμή της παλινδρόμησης πρέπει να περνάει κοντά από τα σημεία που αντιστοιχούν στα ζεύγη των παρατηρήσεων (X_i,Y_i), έτσι ώστε να ελαχιστοποιούνται τα σφάλματα της πρόβλεψης.

Έχει επικρατήσει στη διεθνή βιβλιογραφία να συμβολίζουμε με μικρούς ελληνικούς χαρακτήρες τις τιμές των παραμέτρων του πληθυσμού και με λατινικούς χαρακτήρες τις εκτιμήσεις τους από τα δεδομένα του δείγματος. Έτσι, ισχύει ότι:

$$b_o \xrightarrow[\text{εκτιμήσεις}]{} \beta_o$$

$$b_1 \longrightarrow \beta_1$$

Μόλις οι εκτιμήσεις αυτές γίνουν γνωστές, θα είμαστε σε θέση να προβλέπουμε τις τιμές της *Y* με την εξίσωση παλινδρόμησης:

$$\hat{Y} = b_o + b_1 X$$

Δηλαδή, η $\hat{Y}$ είναι η εκτίμηση της $E(Y) = \beta_o + \beta_1 X$. Έτσι, οι αποκλίσεις μεταξύ των πραγματικών τιμών της Y και των τιμών $\hat{Y}$, που συμβολίζονται με e, ισούνται με:

$$e_i = Y_i - \hat{Y}_i$$

ή

$$e_i = Y_i - (b_o + b_1 X_i), \quad \text{για i} = 1,\ldots,\text{n}$$

Επομένως, αναζητούμε εκείνες τις τιμές των b_o και bi που θα ελαχιστοποιούν τις αποκλίσεις (κατάλοιπα ή σφάλματα) e_i. Επειδή τα σφάλματα έχουν και θετικό και αρνητικό πρόσημο, προσπαθούμε να ελαχιστοποιήσουμε τα τετράγωνά τους, και μάλιστα το άθροισμά τους. Γι' αυτό και η μέθοδος ονομάζεται **μέθοδος ελαχίστων τετραγώνων**. Το άθροισμα των τετραγώνων των αποκλίσεων για τα n ζεύγη των παρατηρήσεων ισούται με:

$$\sum_{i=1}^{n} e_i^{\,2} = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n} [Y_i - (b_o + b_1 X_i)]^2$$

Παραγωγίζοντας την παραπάνω σχέση ως προς b_o και b_1 και εξισώνοντας τις δύο παραγώγους με το μηδέν, προκύπτουν δύο εξισώσεις με δύο αγνώστους (normal equations):

$$\sum_{i=1}^{n} Y_i = n b_o + b_1 \sum_{i=1}^{n} X_i$$

$$\sum_{i=1}^{n} Y_i X_i = b_o \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^{\,2}$$

Η λύση του συστήματος των παραπάνω εξισώσεων δίνει τα εξής αποτελέσματα:

- $b_1 = \dfrac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2}$
- $b_o = \overline{Y} - b_1 \overline{X}$

Οι εκτιμητές b_o και b_1 των ελαχίστων τετραγώνων έχουν τις εξής ιδιότητες:

1. Το άθροισμα των καταλοίπων e_i γύρω από τη γραμμή παλινδρόμησης ισούται με το μηδέν.
2. Η γραμμή παλινδρόμησης περνάει από το σημείο $(\overline{X}, \overline{Y})$ που αντιστοιχεί στους μέσους των μεταβλητών X και Y.
3. Οι συντελεστές των ελαχίστων τετραγώνων b_o και b_1 είναι αμερόληπτες εκτιμήσεις των συντελεστών παλινδρόμησης του πληθυσμού β_o και β_1, αντίστοιχα. Επομένως:

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

4. Επίσης, οι συντελεστές b_o και b_1 είναι αποτελεσματικές εκτιμήσεις των β_o και β_1, δηλαδή έχουν το μικρότερο τυπικό σφάλμα εκτίμησης.

3.2.3. Συντελεστής Προσδιορισμού

Το πιο ουσιαστικό ερώτημα που θα πρέπει να απαντηθεί πριν χρησιμοποιηθεί η εξίσωση παλινδρόμησης είναι ποια είναι η προβλεπτική ικανότητα της εξίσωσης και τι ποσοστό των μεταβολών της εξαρτημένης μεταβλητής *Y* οφείλεται στις επιδράσεις της *X*. Η απάντηση στα ερωτήματα αυτά θα καθορίσει και πόσο μπορούμε να εμπιστευτούμε τα αποτελέσματα μιας πρόβλεψης με βάση την εξίσωση παλινδρόμησης.

Η ανάλυση που ακολουθεί βασίζεται στα κατάλοιπα (ή σφάλματα) της εξίσωσης παλινδρόμησης, δηλαδή τα e_i. Όσο μεγαλύτερη είναι η επίδραση της *X* επί της *Y*, τόσο μικρότερα είναι τα κατάλοιπα, και αντιστρόφως. Αρχικά, πρέπει να υπολογιστεί η συνολική διασπορά γύρω από τη γραμμή παλινδρόμησης, δηλαδή το άθροισμα των τετραγώνων των αποκλίσεων των πραγματικών τιμών της *Y* από τις αντίστοιχες τιμές $\hat{Y}$ του υποδείγματος της παλινδρόμησης, επομένως, πρέπει να υπολογιστεί το **άθροισμα των τετραγώνων των σφαλμάτων** (sum of squared errors) και συμβολίζεται με SSE:

$$SSE = \sum e^2 = \sum (Y - \hat{Y})^2$$

Όπως αναφέρθηκε και νωρίτερα, το μοντέλο της παλινδρόμησης στον πληθυσμό ορίζεται από την εξίσωση $Y = E(Y) + \varepsilon = \beta_o + \beta_1 X + \varepsilon$. Ωστόσο, για την πλήρη περιγραφή του μοντέλου, εκτός από τους συντελεστές παλινδρόμησης β_o και β_1, πρέπει να γνωρίζουμε και τη διακύμανση του σφάλματος ε, δηλαδή το σ^2_ε. Η διακύμανση του σφάλματος είναι η παράμετρος που καθορίζει την ένταση της εξάρτησης της *Y* από την *X*. Η εκτίμησή της θα βασιστεί στο άθροισμα των τετραγώνων γύρω από τη γραμμή παλινδρόμησης, δηλαδή το SSE. Συμβολίζοντας την εκτίμηση του σ^2_ε με s^2_ε, προκύπτει ότι:

$$s^2_\varepsilon = \frac{\sum (Y - \hat{Y})^2}{n-2} = \frac{SSE}{n-2} = MSE$$

όπου MSE = Mean Square Error, και
n-2 = οι βαθμοί ελευθερίας.

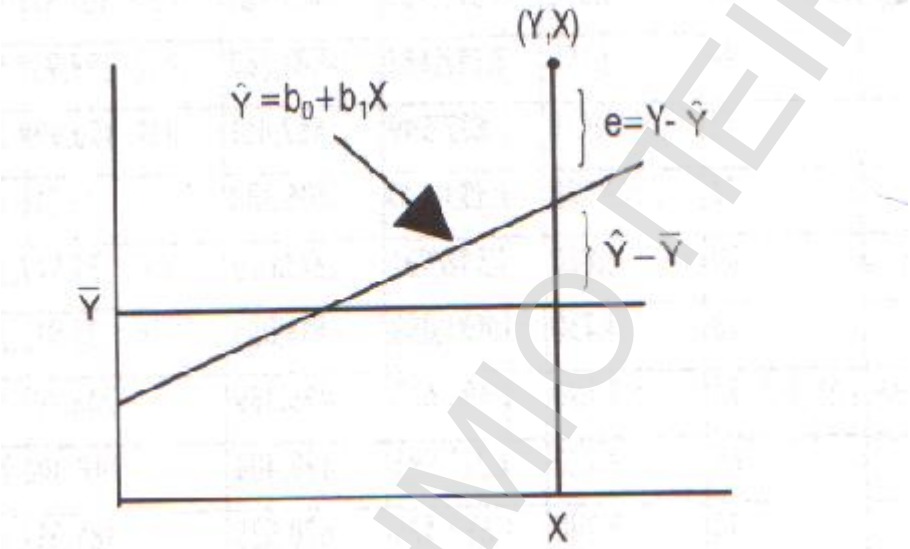
Χάνονται δύο βαθμοί ελευθερίας, διότι η εκτίμηση του s^2_ε βασίζεται στην εκτίμηση δύο παραμέτρων: των b_o και b_1.

Η διασπορά της μεταβλητής *Y* (και γενικότερα) ορίζεται από το άθροισμα των τετραγώνων των αποκλίσεων των τιμών από το μέσο όρο τους, δηλαδή $\sum (Y - \bar{Y})^2$. Αυτό το άθροισμα ονομάζεται **συνολικό άθροισμα τετραγώνων** (total sum of squares) και συμβολίζεται με TSS. Συνεπώς:

$$TSS = \sum (Y - \bar{Y})^2$$

Το SSE αντιπροσωπεύει το μέρος της συνολικής μεταβλητότητας της *Y* που δεν εξηγείται από την εξίσωση παλινδρόμησης. Το υπόλοιπο, δηλαδή TSS-SSE, αποτελεί το μέρος της διασποράς της TSS που οφείλεται στις επιδράσεις της *X*. Με άλλα λόγια, η συνολική μεταβλητότητα της *Y* χωρίζεται σε δύο μέρη (συνιστώσες): στην "εξηγημένη" από την εξίσωση παλινδρόμησης, και στην "ανεξήγητη", δηλαδή εκείνη που οφείλεται στην επίδραση όλων των άλλων παραγόντων, εκτός της *X*. Έτσι, όπως απεικονίζεται και στο Σχήμα 3.4, η απόκλιση $(Y_i - \overline{Y})$ διακρίνεται σε:

$$(Y_i - \overline{Y}) = (Y_i - \hat{Y}_i) + (\hat{Y}_i - \overline{Y})$$



<u>Σχήμα 3.4</u>: Συνιστώσες της Διασποράς της *Y*

Υψώνοντας στο τετράγωνο τα δύο μέρη της παραπάνω σχέσης και αθροίζοντας για όλες τις τιμές της Y_i, i=1,…,n, προκύπτει η σχέση:

$$\sum (Y - \overline{Y})^2 = \sum (Y - \hat{Y})^2 + \sum (\hat{Y} - \overline{Y})^2$$

$$\left(\text{ισχύει ότι } 2 \cdot \sum (Y - \hat{Y})(\hat{Y} - \overline{Y}) = 0\right)$$

Το άθροισμα $\sum (\hat{Y} - \overline{Y})^2$ αποτελεί το μέρος της διασποράς της *Y* που οφείλεται στις επιδράσεις της *X*. Για το λόγο αυτόν, εξηγείται από την εξίσωση παλινδρόμησης, ονομάζεται **άθροισμα των τετραγώνων των τιμών της παλινδρόμησης** (sum of squared regression) και συμβολίζεται με SSR. Δηλαδή:

$$SSR = \sum (\hat{Y} - \overline{Y})^2$$

Επομένως, ισχύει η σχέση:

$$TSS = SSR + SSE$$

Το ποσοστό της συνολικής μεταβλητότητας της *Y* που εξηγείται από την εξίσωση παλινδρόμησης, δηλαδή οφείλεται στις επιδράσεις της *X*, ονομάζεται **συντελεστής προσδιορισμού** και συμβολίζεται με R^2:

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

ή

$$R^2 = 1 - SSE / TSS$$
$$= 1 - \sum(Y - \hat{Y})^2 / \sum(Y - \bar{Y})^2$$
$$= 1 - \frac{\sum Y^2 - b_o \sum Y - b_1 \sum YX}{\sum Y^2 - (\sum Y)^2 / n}$$

Από τα παραπάνω προκύπτει ότι ο συντελεστής προσδιορισμού R^2 παίρνει μόνο θετικές τιμές στο διάστημα [0,1]. Εφόσον δεν παίρνει αρνητικές τιμές, δεν γνωρίζουμε εάν η ευθεία κλίνει προς τα επάνω ή προς τα κάτω.

Επίσης, υπάρχει και εναλλακτικός τρόπος υπολογισμού του συντελεστή προσδιορισμού. Ισχύει ότι ο συντελεστής προσδιορισμού R^2 ισούται με το τετράγωνο του συντελεστή συσχέτισης r, δηλαδή:

$$R^2 = r^2$$

Η σχέση αυτή ισχύει μόνο για την απλή γραμμική παλινδρόμηση, όπου έχουμε μία ανεξάρτητη μεταβλητή. Σχετικά με την ερμηνεία και την αξιολόγηση του συντελεστή συσχέτισης r, αυτός πρέπει να είναι τουλάχιστον 0,7 (ή -0,7) για να χαρακτηριστεί η συσχέτιση έντονη. Και αυτό, διότι $R^2 = (0,7)^2 = 0,49 \approx 50\%$, που σημαίνει ότι πρέπει να ελέγχουμε τουλάχιστον το 50% της διασποράς της *Y* για να χαρακτηρίσουμε τη σχέση μεταξύ των μεταβλητών *X* και *Y* έντονη. Ένας συντελεστής συσχέτισης ίσος με 0,6 αποκαλύπτει μια ασθενή σχέση εξάρτησης , διότι $R^2 = (0,6)^2 = 0,36$, που σημαίνει ότι η *X* επηρεάζει λίγο περισσότερο από το ένα τρίτο των μεταβολών της *Y*.

3.2.4. Έλεγχος της Στατιστικής Σημαντικότητας του Συντελεστή Προσδιορισμού

Ο έλεγχος της στατιστικής σημαντικότητας του συντελεστή προσδιορισμού R^2 βασίζεται στη σχέση *TSS=SSR+SSE*, που δίνει τις δύο συνιστώσες της συνολικής μεταβλητότητας της *Y*. Ουσιαστικά, αυτό που πρέπει να ελεγχθεί είναι εάν το ποσοστό των μεταβολών της *Y* που οφείλεται στις επιδράσεις της *X* είναι διάφορο του μηδενός. Έτσι, η μηδενική υπόθεση (H_o) και η εναλλακτική αυτής (H_1) διατυπώνονται ως εξής:

- **Μηδενική Υπόθεση, H_o**: Η εξίσωση παλινδρόμησης δεν εξηγεί καθόλου τις μεταβολές της *Y* (το ποσοστό της εξηγημένης διασποράς της *Y* είναι μηδέν).
- **Εναλλακτική Υπόθεση, H_1**: Η εξίσωση παλινδρόμησης εξηγεί ένα μέρος των μεταβολών της *Y* (το ποσοστό της εξηγημένης διασποράς της *Y* είναι μεγαλύτερο του μηδενός).

Άρα, πρέπει να συγκριθούν οι δύο συνιστώσες της *TSS*, η εξηγημένη *SSR* και η ανεξήγητη *SSE*. Εάν η πρώτη είναι σημαντικά μεγαλύτερη της δεύτερης, σημαίνει ότι η επίδραση της εξίσωσης παλινδρόμησης είναι σημαντική. Στην αντίθετη περίπτωση που η ανεξήγητη *SSE* είναι σημαντικά μεγαλύτερη από την εξηγημένη *SSR*, το ποσοστό της *TSS* 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 (degrees of freedom-d.f.). Οι λόγοι που θα προκύψουν ονομάζονται **μέσα τετράγωνα** (mean squares-MS) και ο έλεγχος μεταξύ τους βασίζεται στην κατανομή F (ή κατανομή Snedecor).

Εάν η τιμή $F_{1,n-2}$ είναι μεγαλύτερη της κρίσιμης τιμής $F_{(1,n-2),\alpha}$ (όπου α = επίπεδο σημαντικότητας), τότε απορρίπτεται η μηδενική υπόθεση, και αντιστρόφως. Ο λόγος *SSE/n-2* ονομάζεται και **μέσο τετραγωνικό σφάλμα** (mean square error) και συμβολίζεται με *MSE*. Έτσι, η στατιστική F με βαθμούς ελευθερίας 1 και n-2 ισούται με:

$$F_{(1,n-2)} = \frac{\sum(\hat{Y} - \bar{Y})^2 / 1}{\sum(Y - \hat{Y})^2 / (n-2)} = [SSR/1] / [SSE/(n-2)] = SSR / MSE$$

3.2.5. Έλεγχος Στατιστικής Σημαντικότητας του Συντελεστή Παλινδρόμησης b_i

Από τα παραπάνω προκύπτει ότι εάν η ανεξάρτητη μεταβλητή *X* ασκεί στατιστικά σημαντική επίδραση στην εξαρτημένη μεταβλητή *Y*, τότε ο συντελεστής παλινδρόμησης του πληθυσμού β_1 θα είναι διάφορος του μηδενός. Επιπλέον, όπως συμβαίνει με όλες τις παραμέτρους που η εκτίμησή τους βασίζεται σε δείγμα παρατηρήσεων, έτσι και ο συντελεστής παλινδρόμησης b_1 υπόκειται στα σφάλματα της δειγματοληψίας. Αυτό σημαίνει ότι πρέπει να γνωρίζουμε όχι μόνο εάν ο β_1 είναι διάφορος του μηδενός, αλλά και σε ποιο διάστημα εμπιστοσύνης βρίσκεται η τιμή του συντελεστή παλινδρόμησης του πληθυσμού.

Δεν πρέπει να ξεχνάμε, ωστόσο, ότι ο συντελεστής παλινδρόμησης β_1 είναι εκείνος που "επωμίζεται" όλη την ευθύνη της περιγραφής της σχέσης εξάρτησης της *Y* από την *X*. Όλη η επιτυχία της ανάλυσης παλινδρόμησης εξαρτάται από την επιτυχή εκτίμηση του συντελεστή β_1. Με τον όρο επιτυχή στη στατιστική ορολογία, εννοούμε την εκτίμηση που είναι αμερόληπτη (ιδιότητα που εξασφαλίζεται από τη μέθοδο των ελαχίστων τετραγώνων) και αποτελεσματική, δηλαδή να έχει μικρό δειγματοληπτικό σφάλμα, και ως εκ τούτου, μεγάλη πιθανότητα να βρίσκεται κοντά στην πραγματική τιμή του πληθυσμού. Επομένως, τόσο ο έλεγχος της σημαντικότητας του β_1, όσο και η εκτίμηση του διαστήματος εμπιστοσύνης, είναι απαραίτητες ενέργειες πριν χρησιμοποιήσουμε την εξίσωση παλινδρόμησης, για παράδειγμα για προβλέψεις που θα αποτελέσουν τη βασική πληροφόρηση σε μια διαδικασία λήψης αποφάσεων.

Το τυπικό σφάλμα της κατανομής δειγματοληψίας του συντελεστή b_1 συμβολίζεται με σ_{b1} και δίνεται από τη σχέση:

$$s_{b_1} = \frac{s_e}{\sqrt{\sum (X - \overline{X})^2}}$$

όπου $\boldsymbol{\sigma_\varepsilon}$ είναι το **τυπικό σφάλμα εκτίμησης της εξίσωσης παλινδρόμησης**, δηλαδή η τετραγωνική ρίζα του σ^2_ε. Όμως, το σ_ε δεν είναι γνωστό, κι επομένως πρέπει να χρησιμοποιήσουμε την εκτίμησή του από τα δεδομένα του δείγματος, το s_e. Έτσι, η εκτίμηση του τυπικού σφάλματος σ_{b1} συμβολίζεται με s_{b1} και ισούται με:

$$\boldsymbol{s_{b1}} = \frac{s_e}{\sqrt{\sum (X - \overline{X})^2}}$$

Το $\boldsymbol{s_{b1}}$ ονομάζεται και τυπικό σφάλμα εκτίμησης του συντελεστή παλινδρόμησης β_1. Ο έλεγχος της σημαντικότητας του β_1 και η εκτίμηση του διαστήματος εμπιστοσύνης βασίζονται στο s_{b1}.

Η μηδενική και η εναλλακτική υπόθεση διατυπώνονται ως εξής:

- **Μηδενική Υπόθεση,** $\boldsymbol{H_o: \beta_1 = 0}$
- **Εναλλακτική Υπόθεση,** $\boldsymbol{H_1: \beta_1 \neq 0}$

Ο έλεγχος γίνεται με το κριτήριο t και n-2 βαθμούς ελευθερίας, δηλαδή:

$$t_{n-2} = \frac{b_1}{s_{b_1}}$$

Εάν η τιμή $|t_{n-2}|$είναι μεγαλύτερη της κρίσιμης τιμής $|t_{n-2,a/2}|$, απορρίπτεται η μηδενική υπόθεση, και αντίστροφα. Σημειώνεται ότι η παραπάνω σχέση κανονικά είναι: $t_{n-2} = \frac{b_1 - b_{1o}}{s_{b_1}}$, όπου β_{1o} είναι η τιμή του συντελεστή β_1 κάτω από τη μηδενική υπόθεση. Η συνηθέστερη, όμως, τιμή που μας ενδιαφέρει όταν κάνουμε τη μηδενική υπόθεση, είναι αυτή για $\beta_{1o}=0$.

Για την απλή γραμμική παλινδρόμηση, όλοι οι έλεγχοι που γίνονται οδηγούν στο ίδιο συμπέρασμα. Δηλαδή, εάν ο συντελεστής συσχέτισης είναι στατιστικά σημαντικός, σημαντικός θα είναι και ο συντελεστής προσδιορισμού και ο συντελεστής παλινδρόμησης. Εφόσον υπάρχει μία μόνο ανεξάρτητη μεταβλητή, η επίδρασή της θα δώσει την ίδια εικόνα μέσα από όλες τις παραμέτρους (r, b_1 και R^2). Συνεπώς, οι τρεις αυτοί έλεγχοι είναι μεταξύ τους ισοδύναμοι και δίνουν το ίδιο ακριβώς αποτέλεσμα, όσον αφορά την πιθανότητα της τιμής των κριτηρίων. Ωστόσο, ο έλεγχος του b_1 είναι αυτός που έχει επικρατήσει, διότι αυτός είναι που δίνει όλη την πληροφορία που ενσωματώνεται στην εξίσωση παλινδρόμησης. [10, 17, 40]

3.3. Άλλα Είδη Συσχέτισης

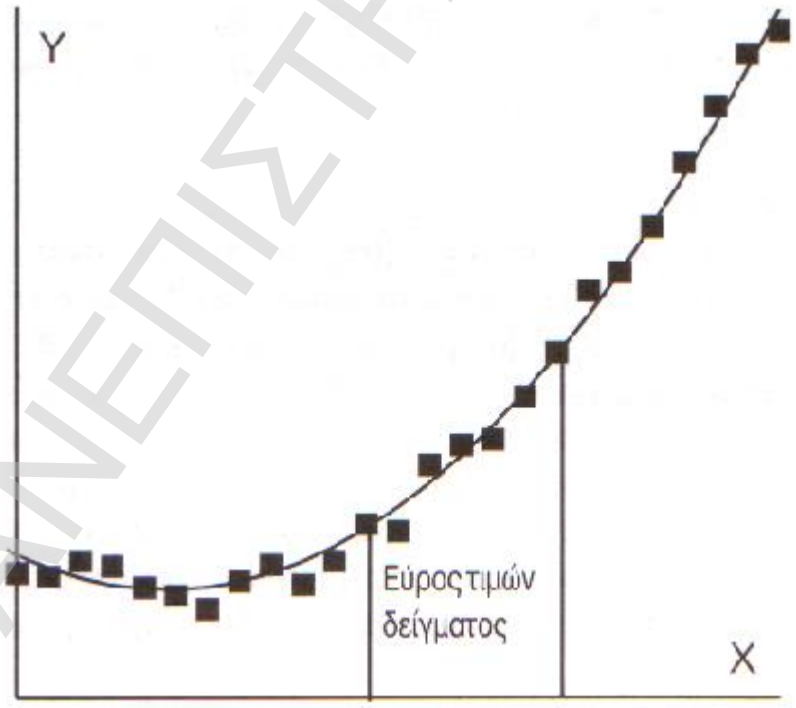
3.3.1. Καμπυλόγραμμη Συσχέτιση

Όπως είδαμε στην αρχή του κεφαλαίου, η σχέση μεταξύ δύο μεταβλητών δεν είναι πάντοτε γραμμική. Ωστόσο, η γραμμική συσχέτιση είναι αυτή με τη μεγαλύτερη εφαρμογή. Ακόμα και σε περιπτώσεις που η πραγματική σχέση μεταξύ των μεταβλητών είναι καμπυλόγραμμη, εάν το δείγμα καλύπτει σχετικά περιορισμένο εύρος τιμών (ή μικρή χρονική περίοδο), οι τιμές του δείγματος δεν επαρκούν για να αποκαλύψουν τη μη-γραμμική σχέση. Υπάρχουν, όμως, αρκετές περιπτώσεις που τα δεδομένα αποκαλύπτουν μη-γραμμική σχέση μεταξύ των μεταβλητών.

Αν και υπάρχουν αρκετές μαθηματικές εκφράσεις μη-γραμμικών σχέσεων μεταξύ δύο μεταβλητών, μία συνήθης μορφή που έχει καθιερωθεί κυρίως στο χώρο των οικονομικών αναλυτών είναι η ακόλουθη:

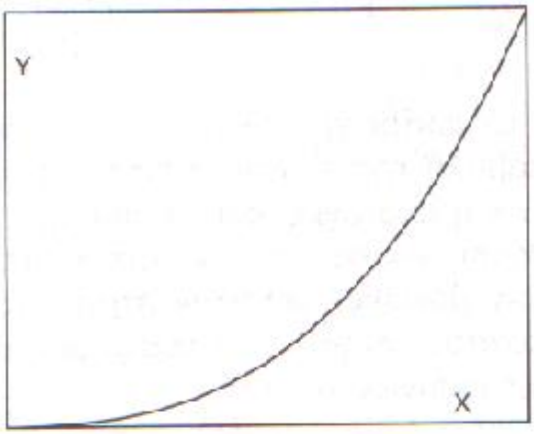
$$Y = b_o \cdot X^{b_1} \cdot e^u$$

όπου: Y = η τιμή της εξαρτημένης μεταβλητής
X = η τιμή της ανεξάρτητης μεταβλητής
β_o = μία σταθερά
β_1 = η τιμή του συντελεστή παλινδρόμησης
e = η βάση των νεπερείων λογαρίθμων
u = σφάλμα ή κατάλοιπο (η διαφορά μεταξύ της πραγματικής τιμής της $ln(Y)$ και της τιμής της πρόβλεψης $ln(\hat{Y})$ που προκύπτει από το υπόδειγμα), που κατανέμεται κανονικά με μέσο το μηδέν και διακύμανση σ^2_u.

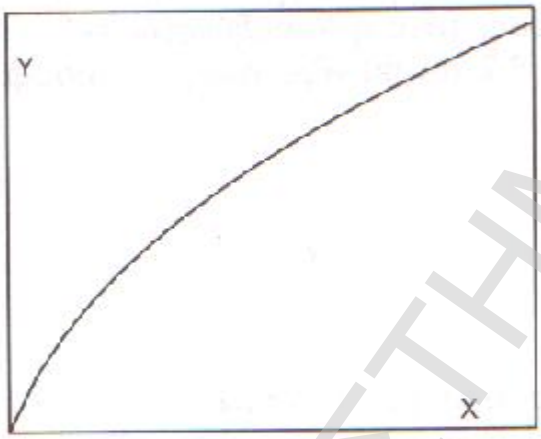


Σχήμα 3.5: Καμπυλόγραμμη Σχέση

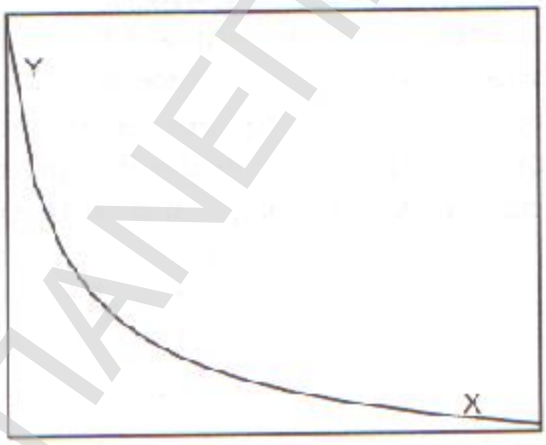
Η μορφή της καμπύλης που αντιστοιχεί στην παραπάνω εξίσωση εξαρτάται από το συντελεστή β_1. Για παράδειγμα, εάν ο συντελεστής είναι μεγαλύτερος της μονάδας, τότε η Y αυξάνεται εκθετικά σε σχέση με τη X. Το Σχήμα 3.6 απεικονίζει όλες τις πιθανές μορφές.



Οι θετικές μεταβολές της Y συνεχώς αυξάνονται.
Ο συντελεστής β_1 ικανοποιεί τη σχέση: $\beta_1 > 1$.



Οι θετικές μεταβολές της Y συνεχώς μειώνονται.
Ο συντελεστής β_1 ικανοποιεί τη σχέση: $0 < \beta_1 < 1$.



Οι αρνητικές μεταβολές της Y συνεχώς μειώνονται.
Ο συντελεστής β_1 ικανοποιεί τη σχέση: $\beta_1 < 0$.

<u>Σχήμα 3.6</u>: Μορφές της Καμπυλόγραμμης Σχέσης $Y=\beta_o X^{\beta_1}$ Ανάλογα με την Τιμή του Συντελεστή β_1

Το μοντέλο $Y = b_o \cdot X^{b_1} \cdot e^u$ δεν μπορεί να εκτιμηθεί με την κλασική μέθοδο των ελαχίστων τετραγώνων. Αυτή η αδυναμία οφείλεται στο γεγονός ότι το παραπάνω υπόδειγμα είναι μη-γραμμικό ως προς τις παραμέτρους β_o και β_1, που σημαίνει ότι η παραγώγιση του αθροίσματος των τετραγώνων των αποκλίσεων ως προς τα β_o και β_1 οδηγεί σε μη-γραμμικό σύστημα εξισώσεων. Η λύση μη-γραμμικών συστημάτων είναι αρκετά πολύπλοκη διαδικασία. Η εκτίμηση του υποδείγματος γίνεται με έμμεσο τρόπο, λογαριθμίζοντας και τα δύο μέλη της εξίσωσης, δηλαδή:

$$ln(Y) = ln(b_o \cdot X^{b_1} \cdot e^u)$$

ή

$$ln(Y) = ln(\beta_o) + \beta_1 ln(X) + u$$

Η παραπάνω σχέση αποτελεί ένα συνηθισμένο γραμμικό μοντέλο, όμοιο με εκείνο που περιγράφηκε στην προηγούμενη ενότητα. Η μόνη διαφορά είναι ότι αντί των αρχικών μεταβλητών X και Y χρησιμοποιούνται οι νεπέρειοι λογάριθμοι αυτών. Όλες οι υποθέσεις για την εφαρμογή της μεθόδου των ελαχίστων τετραγώνων ισχύουν και εδώ, αφού το σφάλμα u κατανέμεται κανονικά με μέσο το μηδέν και σταθερή διακύμανση ίση με σ^2_u. [10]

Έτσι, η προς εκτίμηση εξίσωση είναι:

$$\hat{Y} = b_o X^{b_1}$$

ή

$$ln(\hat{Y}) = ln(b_o) + b_1 ln(X)$$

Παρόμοια, προκύπτει ότι:

- $b_1 = \dfrac{n\sum \ln(X)\ln(Y) - \sum \ln(X)\sum \ln(Y)}{n[\sum \ln(X)^2] - [\sum \ln(X)]^2}$
- $\ln(b_o) = \dfrac{\sum \ln(Y) - b_1 \sum \ln(X)}{n}$

ΚΕΦΑΛΑΙΟ 4: ΠΟΛΛΑΠΛΗ ΓΡΑΜΜΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

3.4. Υπόδειγμα Πολλαπλής Γραμμικής Παλινδρόμησης

Στο προηγούμενο κεφάλαιο περιγράφηκε η μέθοδος της απλής γραμμικής συσχέτισης και παλινδρόμησης, η οποία περιορίζεται στην ανάλυση της σχέσης μεταξύ δύο μόνο μεταβλητών. Σε αυτό το κεφάλαιο περιγράφεται η πολλαπλή παλινδρόμηση, που αποτελεί την επέκταση της απλής παλινδρόμησης για περισσότερες από δύο μεταβλητές.

Ο σκοπός της πολλαπλής παλινδρόμησης είναι να περιγράψει τη σχέση μεταξύ της εξαρτημένης μεταβλητής Y και των k ανεξάρτητων μεταβλητών $X_1, X_2, \ldots, X_k$. Το μοντέλο υποθέτει ότι η Y είναι γραμμική συνάρτηση των k ανεξάρτητων μεταβλητών και έχει την εξής μορφή:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_\kappa + \varepsilon$$

όπου:

Y = η τιμή της εξαρτημένης μεταβλητής
$X_1, X_2, \ldots, X_k$ = οι τιμές των ανεξάρτητων μεταβλητών
β_o = μια σταθερά
$\beta_1, \beta_2, \ldots, \beta_k$ = οι συντελεστές παλινδρόμησης που περιγράφουν την επίδραση των ανεξάρτητων μεταβλητών
ε = σφάλμα ή κατάλοιπο, δηλαδή η διαφορά μεταξύ της πραγματικής τιμής της Y και της τιμής της πρόβλεψης που προκύπτει από το υπόδειγμα

Οι **βασικές προϋποθέσεις** του μοντέλου πολλαπλής παλινδρόμησης είναι οι ακόλουθες:

1. Τα σφάλματα ε_i είναι ανεξάρτητα μεταξύ τους και κατανέμονται κανονικά.
2. Οι αναμενόμενες τιμές (μέσοι) των σφαλμάτων ε_i είναι μηδέν.
3. Τα σφάλματα ε_i έχουν την ίδια διακύμανση σ^2_ε για όλους τους συνδυασμούς των τιμών των ανεξάρτητων μεταβλητών.
4. Στα πλαίσια της ανάλυσης παλινδρόμησης, οι μεταβλητές X_j θεωρούνται σταθερές ποσότητες, ενώ στην ανάλυση συσχέτισης είναι τυχαίες μεταβλητές. Σε κάθε περίπτωση, οι μεταβλητές X_j είναι *ανεξάρτητες από το σφάλμα ε*.

Επίσης, στην πολλαπλή παλινδρόμηση καλό είναι οι ανεξάρτητες μεταβλητές να είναι ανεξάρτητες μεταξύ τους, έτσι ώστε να μην δημιουργείται το πρόβλημα της πολυσυγγραμικότητας, που θα αναλυθεί σε επόμενη ενότητα.

Όπως έχει αναφερθεί, η εξαρτημένη μεταβλητή είναι συνάρτηση k ανεξάρτητων μεταβλητών. Συνεπώς, η μεταβλητότητα της Y δεν εξηγείται μόνο από τις μεταβολές μιας ανεξάρτητης μεταβλητής, όπως συμβαίνει στην απλή παλινδρόμηση, αλλά προστίθενται και άλλες ανεξάρτητες μεταβλητές σε μια προσπάθεια να ερμηνευτεί μεγαλύτερο μέρος της μεταβλητότητας της Y. Με άλλα λόγια, επεκτείνεται το

υπόδειγμα της απλής παλινδρόμησης με την προσθήκη και άλλων ανεξάρτητων μεταβλητών.

Επομένως, 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,…, X_k, και την τυχαία (κατάλοιπο) συνιστώσα ε, που ενσωματώνει όλους τους άλλους (εκτός των X_1, X_2,…, X_k) παράγοντες που επηρεάζουν τη διαμόρφωση της τιμής της *Y*. Δηλαδή:

$$Y = \beta_o + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_\kappa + \varepsilon =$$
$$= \underbrace{E(Y)}_{\text{μη τυχαίος παράγοντας}} + \underbrace{e}_{\text{τυχαίος παράγοντας}}$$

3.5. Εκτίμηση της Εξίσωσης της Πολλαπλής Γραμμικής Παλινδρόμησης

Σκοπός μας είναι η εκτίμηση των παραμέτρων του μοντέλου της πολλαπλής παλινδρόμησης, δηλαδή των συντελεστών β_o, β_1, …, β_k. Οι εκτιμήσεις από τα δεδομένα του δείγματος των συντελεστών πολλαπλής παλινδρόμησης του πληθυσμού (β_o, β_1, …, β_κ) συμβολίζονται με b_o, b_1, …, b_κ, αντίστοιχα. Έτσι, η εξίσωση που θα προκύψει από την εκτίμηση των **συντελεστών πολλαπλής παλινδρόμησης** είναι η:

$$\hat{Y} = b_o + b_1 X_1 + b_2 X_2 + \ldots + b_k X_k$$

Η παραπάνω εξίσωση σχηματίζει ένα υπερεπίπεδο στο χώρο των k+1 διαστάσεων, που δεν μπορεί να απεικονιστεί στο χαρτί των δύο διαστάσεων. Μόνο στην περίπτωση των τριών μεταβλητών (δύο ανεξάρτητες μεταβλητές) αντιστοιχεί στην συγκεκριμένη εξίσωση ένα επίπεδο στο χώρο των τριών διαστάσεων. Οι συντελεστές b_o, b_1, …, b_κ δείχνουν τη **μερική επίδραση** που ασκούν οι ανεξάρτητες μεταβλητές στην *Y*. Για παράδειγμα, ο συντελεστής b_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_k) παραμείνουν σταθερές. Επομένως, ο συντελεστής b_2 μετράει τη μερική επίδραση της ανεξάρτητης μεταβλητής X_2. Για το λόγο αυτόν, οι συντελεστές b_o, b_1, …, b_k ονομάζονται και συντελεστές μερικής παλινδρόμησης.

Η παραπάνω εξίσωση θα προκύψει από τη μέθοδο των ελαχίστων τετραγώνων, όπως και στην περίπτωση της απλής γραμμικής παλινδρόμησης. Δηλαδή, θα αναζητήσουμε εκείνες τις τιμές των b_o, b_1, …, b_k 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. Για την απλούστερη περίπτωση όπου υπάρχουν δύο ανεξάρτητες μεταβλητές, το υπόδειγμα που θα εκτιμήσουμε είναι το εξής:

$$\hat{Y} = b_o + b_1 X_1 + b_2 X_2$$

Η $\hat{Y}$ είναι η εκτίμηση της *E(Y)*, ενώ οι αποκλίσεις μεταξύ των πραγματικών τιμών της *Y* και των τιμών $\hat{Y}$ συμβολίζονται με e, δηλαδή:

$$e_i = Y_i - \hat{Y}_i$$

ή

$$e_i = Y_i - (b_o + b_1X_{1i} + b_2X_{2i}), \quad \text{για i = 1, …, n}$$

όπου n είναι το μέγεθος του δείγματος.

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

$$\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \sum_{i=1}^{n}[Y_i - (b_o + b_1X_{1i} + b_2X_{2i})]^2$$

ενώ οι κανονικές εξισώσεις που προκύπτουν είναι οι εξής:

$$\sum Y = nb_o + b_1\sum X_1 + b_2\sum X_2$$

$$\sum YX_1 = b_o\sum X_1 + b_1\sum X_1^{\,2} + b_2\sum X_1X_2$$

$$\sum YX_2 = b_o\sum X_2 + b_1\sum X_1X_2 + b_2\sum X_2^{\,2}$$

Η λύση του συστήματος των εξισώσεων δίνει τις τιμές b_1 και b_2. Ο συντελεστής b_o προκύπτει εύκολα από την πρώτη εξίσωση του συστήματος εξισώσεων, με απλή αντικατάσταση των τιμών των b_1 και b_2. 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.

3.6. Συντελεστής Πολλαπλού Προσδιορισμού

Στην απλή γραμμική παλινδρόμηση περιγράψαμε το συντελεστή προσδιορισμού, που μετρά το ποσοστό της μεταβλητότητας της *Y* που οφείλεται στις επιδράσεις της ανεξάρτητης μεταβλητής *X*. Στην πολλαπλή παλινδρόμηση χρησιμοποιείται, επίσης, ο ανάλογος συντελεστής, για να μετρηθεί το ποσοστό της μεταβλητότητας της *Y* που οφείλεται στις επιδράσεις **όλων μαζί των ανεξάρτητων μεταβλητών** $X_1, X_2,\ldots, X_k$. Επειδή στο υπόδειγμα της πολλαπλής παλινδρόμησης περιλαμβάνονται περισσότερες από μια ανεξάρτητες μεταβλητές, ο συντελεστής πολλαπλού προσδιορισμού μετράει τη συνολική επίδραση που δέχεται η *Y* από τις $X_1, X_2,\ldots, X_k$. Ο συντελεστής πολλαπλού προσδιορισμού ισούται με:

$$R^2 = SSR / TSS = 1 - SSE / TSS$$

όπου:

$$SSR = \sum (\hat{Y} - \bar{Y})^2$$
$$TSS = \sum (Y - \bar{Y})^2$$
$$SSE = \sum e^2 = \sum (Y - \hat{Y})^2$$

Από τον ορισμό τόσο του απλού, όσο και του πολλαπλού συντελεστή προσδιορισμού, προκύπτει ότι η προσθήκη μιας νέας ανεξάρτητης μεταβλητής στο υπόδειγμα θα οδηγήσει σε μείωση της ανερμήνευτης συνιστώσας (αποκλίσεις μεταξύ Y και $\hat{Y}$) και επομένως σε αύξηση της τιμής του συντελεστή R^2. Όμως, κάθε νέα ανεξάρτητη μεταβλητή "στοιχίζει" και ένα βαθμό ελευθερίας. Οι βαθμοί ελευθερίας ισούνται με n-k-1, όπου k 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" αύξηση της τιμής του R^2 που δεν θα έχει καμία αξία, όταν μάλιστα ο αριθμός των ανεξάρτητων μεταβλητών (k) είναι υψηλός σε σχέση με το μέγεθος του δείγματος (n).

Το πρόβλημα αυτό αντιμετωπίζεται με το "**διορθωμένο**" (adjusted) **συντελεστή πολλαπλού προσδιορισμού**, που λαμβάνει υπόψη την απώλεια των βαθμών ελευθερίας. Ο διορθωμένος συντελεστής $\bar{R}^2$ ισούται με:

$$\bar{R}^2 = 1 - \frac{SSE/[n-(k+1)]}{SST/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-(k+1)}$$

Ο διορθωμένος συντελεστής πολλαπλού προσδιορισμού δεν αυξάνεται πάντα, όταν νέες ανεξάρτητες μεταβλητές εισέρχονται στην εξίσωση παλινδρόμησης. Ωστόσο, εάν αυτός αυξηθεί σε μία τέτοια περίπτωση, τότε σίγουρα πρέπει η συγκεκριμένη νέα μεταβλητή να περιληφθεί στο μοντέλο παλινδρόμησης. Επίσης, ο $\bar{R}^2$ έχει σημαντικό ρόλο στις περιπτώσεις εκείνες που ο αριθμός των ανεξάρτητων μεταβλητών είναι αρκετά μεγάλος σε σχέση με το μέγεθος του δείγματος. Για μεγάλο (σε σχέση με το k) μέγεθος δείγματος, ο $\bar{R}^2$ διαφέρει ελάχιστα από τον R^2.

Ωστόσο, η συμμετοχή κάθε ανεξάρτητης μεταβλητής στην ερμηνεία των μεταβολών της Y μετριέται με έναν άλλο συντελεστή, το συντελεστή μερικού προσδιορισμού, που μετράει την επίδραση κάθε ανεξάρτητης μεταβλητής, εάν πρώτα αφαιρεθεί η επίδραση των υπόλοιπων ανεξάρτητων μεταβλητών.

3.7. Έλεγχοι Στατιστικής Σημαντικότητας

Οι έλεγχοι στατιστικής σημαντικότητας στην ανάλυση της πολλαπλής παλινδρόμησης έχουν σκοπό πρώτα να ελέγξουν αν η εξίσωση της παλινδρόμησης, στο σύνολό της, εξηγεί ένα σημαντικό μέρος των μεταβολών της εξαρτημένης μεταβλητής Y, και στη συνέχεια, εφόσον υπάρχουν περισσότερες από μία

ανεξάρτητες μεταβλητές, να ελέγξουμε τη σημαντικότητα των συντελεστών παλινδρόμησης ξεχωριστά. Έτσι, θα διαπιστώσουμε ποιες μεταβλητές ασκούν σημαντική επίδραση στην *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, ελέγχουμε αυτό που μετρά ο R^2, εάν το ποσοστό των μεταβολών της *Y* που οφείλεται στις επιδράσεις των ανεξάρτητων μεταβλητών είναι διάφορο του μηδενός. Έτσι, η μηδενική υπόθεση (H_o) και η εναλλακτική αυτής (H_1) διατυπώνονται ως εξής:

- **Μηδενική Υπόθεση, H_o**: Η εξίσωση παλινδρόμησης δεν εξηγεί καθόλου τις μεταβολές της *Y* (το ποσοστό της εξηγημένης διασποράς της *Y* είναι μηδέν) και επομένως $\boldsymbol{\beta_1 = \beta_2 = \ldots = \beta_k = 0}$.

- **Εναλλακτική Υπόθεση, H_1**: Η εξίσωση παλινδρόμησης εξηγεί ένα μέρος των μεταβολών της *Y* (το ποσοστό της εξηγημένης διασποράς της *Y* είναι μεγαλύτερο του μηδενός) και τουλάχιστον ένας συντελεστής $\boldsymbol{\beta_i \neq 0}$.

Επομένως, θα συγκρίνουμε τις δύο συνιστώσες της *TSS*, την εξηγημένη *SSR* και την ανεξήγητη *SSE*. Εάν η πρώτη είναι σημαντικά μεγαλύτερη της δεύτερης, σημαίνει ότι η επίδραση της εξίσωσης παλινδρόμησης είναι σημαντική. Στην αντίθετη περίπτωση που η ανεξήγητη *SSE* είναι σημαντικά μεγαλύτερη από την εξηγημένη *SSR*, το ποσοστό της *TSS* 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 (degrees of freedom-d.f.). Οι λόγοι που θα προκύψουν ονομάζονται μέσα τετράγωνα (mean squares-MS) και ο έλεγχος μεταξύ τους βασίζεται στην κατανομή F (ή κατανομή Snedecor).

Εάν η τιμή $F_{k,n-k-1}$ είναι μεγαλύτερη της κρίσιμης τιμής $F_{(k,n-k-1),\alpha}$ (όπου α = επίπεδο σημαντικότητας), τότε απορρίπτεται η μηδενική υπόθεση, και αντιστρόφως. Ο λόγος *SSE/(n-k-1)* ονομάζεται και **μέσο τετραγωνικό σφάλμα** (mean square error) και συμβολίζεται με *MSE*. Έτσι, η στατιστική F με βαθμούς ελευθερίας k και n-k-1 ισούται με:

$$F_{(k,n-k-1)} = \frac{\sum (\hat{Y} - \bar{Y})^2 / k}{\sum (Y - \hat{Y})^2 / (n-k-1)} = [SSR/k] / [SSE/(n-k-1)] = (SSR/k) / MSE$$

Εάν ο πρώτος 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 μεταβλητής *Y*, το επόμενο βήμα είναι να ελέγξουμε τη σημαντικότητα των συντελεστών μερικής παλινδρόμησης b_i, i=1,2,…,k. Όπως συμβαίνει με όλες τις παραμέτρους που η εκτίμησή τους βασίζεται σε δείγμα παρατηρήσεων, έτσι και οι συντελεστής παλινδρόμησης b_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 είναι εκείνοι που έχουν όλη την ευθύνη της περιγραφής της σχέσης εξάρτησης της *Y* από τις X_j.

Το τυπικό σφάλμα της κατανομής δειγματοληψίας του συντελεστή b_i συμβολίζεται με σ_{bi} και έτσι η μηδενική και η εναλλακτική υπόθεση διατυπώνονται ως εξής:

- **Μηδενική Υπόθεση, H_o: $\beta_i = 0$**, 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.

Ο έλεγχος γίνεται με το κριτήριο t και n-k-1 βαθμούς ελευθερίας, δηλαδή:

$$t_{n-k-1} = \frac{b_i}{s_{b_{i1}}}$$

Εάν η τιμή $|t_{n-k-1}|$είναι μεγαλύτερη της κρίσιμης τιμής $|t_{n-k-1,a/2}|$, απορρίπτεται η μηδενική υπόθεση, και αντίστροφα.

3.8. Συντελεστές Μερικού Προσδιορισμού

Το δυσκολότερο ίσως ερώτημα που η ανάλυση της πολλαπλής παλινδρόμησης πρέπει να απαντήσει είναι ποια είναι η συνεισφορά κάθε ανεξάρτητης μεταβλητής στην εξήγηση των μεταβολών της εξαρτημένης μεταβλητής *Y*. Η δυσκολία στην απάντηση αυτού του ερωτήματος προέρχεται από το γεγονός ότι και οι ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους. Έτσι, είναι δύσκολο να εκτιμήσουμε το ποσοστό συμμετοχής κάθε ανεξάρτητης μεταβλητής στην ερμηνεία των μεταβολών της *Y*. Η αλληλεπίδραση μεταξύ των ανεξάρτητων μεταβλητών εμποδίζει να μετρήσουμε τη συνεισφορά κάθε μεταβλητής, αφού η είσοδός τους στο υπόδειγμα δεν επηρεάζει μόνο τη μεταβλητή *Y*, αλλά και τις λοιπές ανεξάρτητες μεταβλητές *X*.

Η ιδανική περίπτωση είναι η εξής:

- Η μεταβλητή X_1 εξηγεί το κ % των μεταβολών της *Y*, σύμφωνα με το υπόδειγμα $\hat{Y} = b_o + b_1 X_1$
- Η μεταβλητή X_2 εξηγεί το π % των μεταβολών της *Y*, σύμφωνα με το υπόδειγμα $\hat{Y} = b_o' + b_1' X_1$
- Οι μεταβλητές X_1 και X_2 (και οι δύο μαζί) εξηγούν το (κ+π) % των μεταβολών της *Y*, σύμφωνα με το υπόδειγμα $\hat{Y} = b_o + b_1 X_1 + b_2 X_2$

Τα παραπάνω ισχύουν **μόνο** στην περίπτωση που ο συντελεστής συσχέτισης μεταξύ των μεταβλητών X_1 και X_2 είναι μηδέν, δηλαδή οι X_1 και X_2 δεν συσχετίζονται γραμμικά μεταξύ τους. Η σχέση που ισχύει είναι:

$$R^2_{Y,(X_1,X_2)} = R^2_{Y,X_1} + R^2_{Y,X_2}$$

Ωστόσο, σε πάρα πολλές περιπτώσεις τα πράγματα δεν είναι τόσο απλά, και οι ανεξάρτητες μεταβλητές συσχετίζονται σε κάποιο βαθμό μεταξύ τους, με αποτέλεσμα να μην ισχύει η παραπάνω σχέση. Στις περιπτώσεις αυτές, αυτό που μπορεί να μετρηθεί είναι το εξής: Η Y εκφράζεται ως γραμμική συνάρτηση της X_1 που ερμηνεύει ένα ποσοστό της συνολικής διασποράς της Y, ενώ το υπόλοιπο μένει ανερμήνευτο. Εάν προστεθεί στο υπόδειγμα και μία νέα μεταβλητή X_2, τι ποσοστό της ανερμήνευτης διασποράς μπορεί να ερμηνεύσει η νέα μεταβλητή X_2; Αυτό το ποσοστό της ανερμήνευτης διασποράς της Y, που δεν εξηγεί η επίδραση της X_1 και που θα ερμηνεύσει η νέα μεταβλητή X_2, ονομάζεται **συντελεστής μερικού προσδιορισμού** της X_2. Αυτός ο συντελεστής είναι το μόνο που μπορούμε να μετρήσουμε. Δηλαδή, τι ποσοστό της ανεξήγητης (από τις επιδράσεις της X_1) διασποράς της Y μπορεί να ερμηνεύσει η νέα ανεξάρτητη μεταβλητή X_2. Όμως, επειδή η X_2 συνήθως συσχετίζεται τόσο με την Y όσο και με την X_1, για να μετρήσουμε την πραγματική συνεισφορά της στο ανερμήνευτο από την X_1 μέρος της διασποράς της Y, πρέπει πρώτα να αφαιρέσουμε την επίδραση της X_1 και από την Y (για να μείνει το ανερμήνευτο μέρος) και από την X_2, για να προκύψει η καθαρή επίδραση της X_2, χωρίς την παρέμβαση της X_1.

Ο συντελεστής μερικού προσδιορισμού της μεταβλητής X_2 συμβολίζεται με $R^2_{Y,X_2/X_1}$ και σημαίνει ότι μετράμε την επίδραση της X_2 στην Y, αφού πρώτα αφαιρέσουμε τις επιδράσεις της X_1 στην Y και στην X_2. Η διαδικασία εκτίμησης του συντελεστή αυτού είναι η ακόλουθη:

1. Εκτιμούμε την εξίσωση παλινδρόμησης $\hat{Y} = b_o + b_1 X_1$ και υπολογίζουμε τα κατάλοιπα $Y - \hat{Y}$.
2. Εκτιμούμε την εξίσωση παλινδρόμησης $\hat{X}_2 = c_o + c_1 X_1$ και υπολογίζουμε τα κατάλοιπα $X_2 - \hat{X}_2$.
3. Ο συντελεστής μερικού προσδιορισμού $R^2_{Y,X_2/X_1}$ είναι ο συντελεστής προσδιορισμού μεταξύ των μεταβλητών $(Y - \hat{Y})$ και $(X_2 - \hat{X}_2)$.

Με ανάλογο τρόπο μπορεί να εκτιμηθεί και ο συντελεστής μερικού προσδιορισμού της ανεξάρτητης μεταβλητής X_1, $R^2_{Y,X_1/X_2}$.

Από τα παραπάνω προκύπτει ότι ο συντελεστής μερικού προσδιορισμού έχει και μια άλλη χρησιμότητα. Επιτρέπει στον αναλυτή να εκτιμήσει από πριν πόσο θα μειωθεί το ανερμήνευτο μέρος των μεταβολών της Y, εάν προστεθεί στο υπόδειγμα παλινδρόμησης μια νέα μεταβλητή. Έτσι, μπορούμε να επιλέξουμε, μεταξύ μιας ομάδας πιθανών ανεξαρτήτων μεταβλητών, εκείνες που θα συνεισφέρουν περισσότερο στη μείωση της ανερμήνευτης μεταβλητότητας της εξαρτημένης

μεταβλητής *Y*. Αυτός ο τρόπος επιλογής των ανεξάρτητων μεταβλητών ονομάζεται **πολλαπλή παλινδρόμηση με διαδοχική επιλογή των ανεξάρτητων μεταβλητών** (stepwise regression) και θα αναλυθεί περαιτέρω στην επόμενη ενότητα του κεφαλαίου. Επειδή οι υπολογισμοί που απαιτούνται για τη διαδοχική επιλογή των μεταβλητών *X* είναι πολύπλοκοι, η μέθοδος εφαρμόζεται μόνο με τη βοήθεια ειδικών στατιστικών προγραμμάτων. [10, 11]

3.9. Έλεγχος και Προβλήματα του Μοντέλου Παλινδρόμησης

3.9.1. Μέθοδοι Επιλογής των Κατάλληλων Μεταβλητών

Πολλές φορές γνωρίζουμε εκ των προτέρων ποιο μοντέλο παλινδρόμησης είναι σωστό, δηλαδή ποιες ανεξάρτητες μεταβλητές *X* πρέπει να συμπεριληφθούν στο μοντέλο, και ποια είναι η σχέση τους (γραμμική ή όχι) με την εξαρτημένη μεταβλητή *Y*. Ωστόσο, πολύ συχνά η θεωρία ή και οι προηγούμενες έρευνες δίνουν μόνο μια σαφή κατεύθυνση για το ποιο μοντέλο παλινδρόμησης είναι το πλέον κατάλληλο. Γι' αυτό, απαιτείται εντατικός έλεγχος των δεδομένων.

Καθώς στο υπόδειγμα παλινδρόμησης περιλαμβάνονται περισσότερες μεταβλητές *X*, προκύπτουν οι παρακάτω αλλαγές:

- Αυξάνεται η πρόβλεψη. Ο συντελεστής πολλαπλού προσδιορισμού R^2 αυξάνεται (όχι όμως και ο διορθωμένος συντελεστής), και η τυπική απόκλιση των καταλοίπων s_e μειώνεται. Όμως, είναι ουσιαστική αυτή η βελτίωση;
- Οι συντελεστές παλινδρόμησης περιγράφουν πώς οι επιπρόσθετες μεταβλητές επηρεάζουν το $\hat{Y}$. Είναι αυτοί οι συντελεστές στατιστικά σημαντικά διάφοροι του μηδενός και αρκετά μεγάλοι, ώστε να δηλώνουν σημαντική βελτίωση του μοντέλου;
- Οι μη σημαντικοί συντελεστές "συρρικνώνονται". Όμως, οι προστιθέμενες μεταβλητές μπορούν να αλλάξουν σημαντικά τα συμπεράσματά μας σχετικά με τις επιδράσεις των άλλων ανεξάρτητων μεταβλητών;

Εάν οι απαντήσεις στα παραπάνω ερωτήματα είναι θετικές, τότε οι επιπρόσθετες μεταβλητές καλό είναι να παραμένουν στο μοντέλο. Οι αρνητικές απαντήσεις υποδηλώνουν ότι οι συγκεκριμένες μεταβλητές συνεισφέρουν πολύ λίγο στο υπόδειγμα παλινδρόμησης, και γι' αυτό θα πρέπει να παραλειφθούν από αυτό.

Ένας βασικός στόχος που πρέπει να ισχύει στην επιλογή των μεταβλητών είναι στο τελικό μοντέλο να υπάρχει μία ισορροπία μεταξύ της απλότητας και της καλής προσαρμογής των δεδομένων στο μοντέλο (fit). Στην αγγλική ορολογία αυτό καλείται *parsimony*. Ο διορθωμένος συντελεστής R^2 είναι ένα πολύ καλό παράδειγμα ύπαρξης της προαναφερθείσας ισορροπίας, καθώς συνδυάζει ένα μέτρο προσαρμογής (R^2) και ένα μέτρο της διαφοράς στην πολυπλοκότητα μεταξύ των δεδομένων (μέγεθος δείγματος, n) και του μοντέλου (αριθμός εκτιμώμενων παραμέτρων, k). Η διαφορά n-k είναι στην ουσία οι βαθμοί ελευθερίας των καταλοίπων.

Στην επιλογή των κατάλληλων μεταβλητών *X* για την ανάλυση παλινδρόμησης εμφανίζονται δύο πιθανοί κίνδυνοι:

1. *Η περίληψη μιας άσχετης μεταβλητής*. Μία μεταβλητή X_k θεωρείται άσχετη, όταν η πραγματική τιμή του συντελεστή β_k είναι μηδέν, ή αρκετά μικρή ώστε να μην θεωρείται στατιστικά σημαντική. Εάν στο μοντέλο συμπεριληφθούν τέτοιες άσχετες μεταβλητές, τότε το μοντέλο έχει γίνει πιο πολύπλοκο, χωρίς να υπάρχει λόγος.

2. *Η παράβλεψη μιας σχετικής μεταβλητής*. Μία μεταβλητή X_k θεωρείται σχετική, όταν η πραγματική τιμή του συντελεστή β_k δεν είναι μηδέν, αλλά αρκετά μεγάλη ώστε να θεωρείται στατιστικά σημαντική, και όταν η μεταβλητή αυτή συσχετίζεται με άλλες ανεξάρτητες μεταβλητές *X*. Εάν στο μοντέλο παραλειφθούν τέτοιες μεταβλητές, τότε το μοντέλο δεν είναι αξιόπιστο, γιατί δεν αναφέρεται στην πραγματικότητα.

Η επιλογή των κατάλληλων ανεξάρτητων μεταβλητών είναι στην ουσία μια αντιστάθμιση πολλών παραγόντων και κινδύνων. Τα δεδομένα του δείγματος μπορούν να μας παραπλανήσουν για το ποιες μεταβλητές είναι σημαντικές. Ακόμα και αν $\beta_k \neq 0$, ο συντελεστής b_k μπορεί να είναι στατιστικά *μη* σημαντικός. Και αντιστρόφως, ακόμα και αν β_k=0, ο b_k μπορεί να είναι διάφορος του μηδενός. Ωστόσο, το ποια ρίσκα είναι τα χειρότερα εξαρτάται από την ισχύ των σχέσεων και το σκοπό της κάθε έρευνας. [26]

Οι μέθοδοι επιλογής μεταβλητών που χρησιμοποιούνται στην πράξη αναλύονται στη συνέχεια.

- **<u>Όλες οι πιθανές παλινδρομήσεις</u>**: Η μέθοδος αυτή συνιστά τον έλεγχο όλων των πιθανών παλινδρομήσεων για k ανεξάρτητες μεταβλητές, και την επιλογή του καλύτερου μοντέλου. Αν θεωρήσουμε ότι κάθε ένα από τα μοντέλα που θεωρούμε έχει και έναν σταθερό συντελεστή β_o, τότε υπάρχουν 2^k πιθανά μοντέλα. Αυτό, διότι καθεμιά από τις k ανεξάρτητες μεταβλητές μπορεί είτε να περιληφθεί στο μοντέλο, είτε να μην περιληφθεί, που σημαίνει ότι υπάρχουν δύο πιθανότητες για κάθε μεταβλητή, δηλαδή 2^k πιθανότητες για ένα μοντέλο που αποτελείται από k μεταβλητές. Για παράδειγμα, εάν υπάρχουν τέσσερις πιθανές μεταβλητές, τότε υπάρχουν 2^4=16 πιθανά μοντέλα: τέσσερα με μια μόνο μεταβλητή, έξι μοντέλα με δύο μεταβλητές, τέσσερα με τρεις μεταβλητές, ένα μοντέλο με όλες τις μεταβλητές και ένα μοντέλο χωρίς καμία μεταβλητή (μόνο με τον όρο β_o). Όπως φαίνεται, ο αριθμός των πιθανών μοντέλων παλινδρόμησης αυξάνεται πολύ γρήγορα, όσο αυξάνεται ο αριθμός των ανεξάρτητων μεταβλητών που περιλαμβάνονται στο μοντέλο.

  Τα διαφορετικά μοντέλα αξιολογούνται σύμφωνα με μερικά κριτήρια για την απόδοση του μοντέλου. Υπάρχουν αρκετά πιθανά κριτήρια: Μπορούμε να επιλέξουμε το μοντέλο με την υψηλότερη τιμή του *διορθωμένου* συντελεστή πολλαπλού προσδιορισμού ή το μοντέλο με το χαμηλότερο τυπικό σφάλμα καταλοίπων (s^2 ή MSE). Μπορούμε επίσης να βρούμε το μοντέλο με τον υψηλότερο *διορθωμένο* συντελεστή R^2 για έναν συγκεκριμένο αριθμό μεταβλητών, και έπειτα να εκτιμήσουμε την αύξηση του R^2 κάθε φορά που προστίθεται επιπλέον μια μεταβλητή, έτσι ώστε να δούμε αν η αύξηση στο R^2 είναι αξιόλογη, για να προσθέσουμε τη νέα μεταβλητή. Ένα άλλο κριτήριο που υπάρχει σε πιο εξειδικευμένα βιβλία είναι αυτό του στατιστικού μέτρου C_p του *Mallow*.

Η παραπάνω διαδικασία με τον έλεγχο όλων των πιθανών μοντέλων παλινδρόμησης είναι αρκετά κουραστική. Οι επόμενες τρεις μέθοδοι είναι όλες βηματικές διαδικασίες για τη δημιουργία του καλύτερου μοντέλου.

- **Επιλογή προς τα εμπρός (forward selection)**: Η επιλογή προς τα εμπρός ξεκινάει με ένα μοντέλο χωρίς μεταβλητές. Η μέθοδος στη συνέχεια θεωρεί όλα τα k μοντέλα, με μια ανεξάρτητη μεταβλητή το καθένα, και επιλέγει το μοντέλο με το πιο σημαντικό *F statistic*, κάνοντας την υπόθεση ότι τουλάχιστον ένα τέτοιο μοντέλο έχει ένα *F statistic* με *p-value* μικρότερο από μία προκαθορισμένη τιμή.

  Σημειώνεται ότι το μερικό *F statistic* ορίζεται ως εξής:

$$F_{(r,n-k-1)} = \frac{(SSE_R - SSE_F)/r}{MSE_F}$$

  όπου SSE_R είναι το άθροισμα των τετραγώνων των σφαλμάτων του μειωμένου μοντέλου, SSE_F είναι το άθροισμα των τετραγώνων των σφαλμάτων του μοντέλου με όλες τις μεταβλητές, $MSE_F=SSE_F/(n-k-1)$ είναι το μέσο τετραγωνικό σφάλμα του κανονικού μοντέλου, *k* είναι ο αριθμός των ανεξάρτητων μεταβλητών συνολικά στο κανονικό μοντέλο και *r* είναι ο αριθμός των μεταβλητών που απορρίπτονται από το κανονικό μοντέλο, σχηματίζοντας έτσι το μειωμένο μοντέλο.

  Στη συνέχεια, ελέγχονται οι μεταβλητές που έχουν παραμείνει εκτός μοντέλου και τα μερικά *F statistics* αυτών, και προστίθεται στο μοντέλο η μεταβλητή με τη μεγαλύτερη τιμή *F*, πάλι υποθέτοντας ότι τουλάχιστον μία μεταβλητή υπάρχει που ικανοποιεί το απαιτούμενο επίπεδο σημαντικότητας. Η διαδικασία αυτή συνεχίζεται έως ότου να μην υπάρχει καμία μεταβλητή εκτός μοντέλου, η οποία έχει μερικό *F statistic* που ικανοποιεί το επίπεδο σημαντικότητας που απαιτείται για να εισέλθει στο μοντέλο.

- **Προς τα πίσω απαλοιφή (backward elimination)**: Η μέθοδος αυτή λειτουργεί με έναν τρόπο αντίστροφο στη forward επιλογή. Το μοντέλο ξεκινάει με όλες τις k μεταβλητές. Στη συνέχεια υπολογίζονται τα μερικά *F statistics* για κάθε μεταβλητή, θεωρώντας καθεμιά από αυτές σαν την τελευταία μεταβλητή που θα εισέλθει στην εξίσωση παλινδρόμησης, δηλαδή εκτιμούμε κάθε μεταβλητή αναφορικά με τη συνεισφορά της σε ένα μοντέλο, το οποίο ήδη περιέχει όλες τις άλλες μεταβλητές. Εάν βρεθεί ότι το επίπεδο σημαντικότητας του μερικού *F statistic* μιας μεταβλητής δεν πληροί κάποια προκαθορισμένη τιμή (το *p-value* είναι μεγαλύτερο από ένα προκαθορισμένο *p-value*), τότε η μεταβλητή απαλείφεται από το μοντέλο. Όλα τα *F statistics* υπολογίζονται εκ νέου για το καινούριο, μειωμένο μοντέλο, και οι μεταβλητές που απομένουν ελέγχονται για το αν ικανοποιούν το προκαθορισμένο standard. Εάν βρεθεί ότι μια μεταβλητή έχει *p-value* μεγαλύτερο από αυτό που απαιτείται, η μεταβλητή απομακρύνεται από την εξίσωση. Η διαδικασία ακολουθεί μέχρι όλες οι μεταβλητές που παραμένουν στην εξίσωση παλινδρόμησης να είναι σημαντικές, ως προς τα μερικά *F statistics* τους.

- **Stepwise παλινδρόμηση**: Η μέθοδος αυτή είναι αυτή που χρησιμοποιείται πιο συχνά από όλες, και απαιτεί τη χρήση ηλεκτρονικού υπολογιστή. Η διαδικασία αυτή είναι στην ουσία ένα μίγμα από τις δύο παραπάνω βηματικές μεθόδους, την προς τα εμπρός επιλογή και την προς τα πίσω απαλοιφή. Στην προς τα εμπρός

επιλογή, όσες μεταβλητές εισέλθουν στην εξίσωση, παραμένουν μόνιμα εκεί. Η μέθοδος αυτή δεν επιτρέπει την επανεκτίμηση της σημαντικότητας μιας μεταβλητής, αφού αυτή μπει στο μοντέλο. Όπως θα δούμε στην επόμενη ενότητα, μπορεί να υπάρξουν προβλήματα πολυσυγγραμικότητας, που καθιστούν μια μεταβλητή περιττή σε ένα μοντέλο, εάν εισέλθουν σε αυτό άλλες μεταβλητές με περίπου την ίδια πληροφορία. Αυτό αποτελεί και την αδυναμία της τεχνικής της μεθόδου της προς τα εμπρός επιλογής. Παρομοίως, στη μέθοδο της προς τα πίσω απαλοιφής, μόλις μια μεταβλητή απορριφθεί από το μοντέλο, παραμένει μόνιμα έξω από αυτό. Καθώς, όμως, είναι πιθανό μια μεταβλητή να μην φαίνεται σημαντική λόγω της πολυσυγγραμικότητας και να απορριφθεί από το μοντέλο, ενώ στην πραγματικότητα έχει προβλεπτική ισχύ όταν κάποιες άλλες μεταβλητές απαλειφθούν, τότε καταλαβαίνουμε ότι και στην προς τα πίσω μέθοδο υπάρχουν περιορισμοί και αδυναμίες.

Η stepwise regression είναι ένας συνδυασμός των προς τα εμπρός και προς τα πίσω επιλογών και επανεκτιμά τη σημαντικότητα κάθε μεταβλητής σε *κάθε* στάδιο. Αυτό ελαχιστοποιεί την πιθανότητα να παραλειφθούν από το μοντέλο σημαντικές μεταβλητές, ή να περιληφθούν σε αυτό άλλες μη σημαντικές. Η διαδικασία δουλεύει ως εξής: ο αλγόριθμος ξεκινάει, όπως και στην προς τα εμπρός επιλογή, με την εύρεση του πιο σημαντικού μοντέλου με μια μεταβλητή. Έπειτα, ελέγχονται οι μεταβλητές εκτός μοντέλου μέσω των μερικών *F tests* και η πιο σημαντική μεταβλητή προστίθεται στο μοντέλο, εφόσον ικανοποιεί το επίπεδο σημαντικότητας εισόδου. Στο σημείο αυτό, η διαδικασία διαφοροποιείται από την προς τα εμπρός μέθοδο και εφαρμόζεται η προς τα πίσω απαλοιφή. Η αρχική μεταβλητή στο μοντέλο *επανεκτιμάται* για να ελεγχθεί εάν καλύπτει ακόμα τις προκαθορισμένες προδιαγραφές, ώστε να παραμείνει στο μοντέλο και μετά την προσθήκη της νέας μεταβλητής. Εάν βρεθεί ότι δεν τις καλύπτει, τότε απορρίπτεται από το μοντέλο. Έπειτα, οι μεταβλητές που είναι ακόμα εκτός μοντέλου ελέγχονται για το αν μπορούν να εισέλθουν στο μοντέλο, και η πιο σημαντική από αυτές, αν βρεθεί κάποια, προστίθεται. Όλες οι μεταβλητές μέσα στο μοντέλο ελέγχονται ξανά ως προς τη σημαντικότητά τους, μετά την είσοδο κάθε νέας μεταβλητής. Η διαδικασία συνεχίζει μέχρι να μην υπάρχουν μεταβλητές εκτός μοντέλου, οι οποίες θα έπρεπε να προστεθούν σε αυτό, και μεταβλητές μέσα στο μοντέλο, οι οποίες δεν θα έπρεπε να συμπεριλαμβάνονται. [11]

3.9.2. Πολυσυγγραμικότητα

Για την πολυσυγγραμικότητα έχουν γίνει ήδη κάποιες αναφορές σε προηγούμενες ενότητες. Η ύπαρξη πολυσυγγραμικότητας συναντάται πολύ συχνά στην πολλαπλή παλινδρόμηση, καθώς, εκτός από τις συσχετίσεις μεταξύ της μεταβλητής Y και των X_i, υπάρχουν συσχετίσεις και μεταξύ των ανεξαρτήτων μεταβλητών X_i. Ιδανικά, οι μεταβλητές X_i σε μια εξίσωση παλινδρόμησης είναι ασυσχέτιστες μεταξύ τους, και κάθε μεταβλητή περιέχει μια μοναδική πληροφορία για την Y, η οποία δεν περιέχεται σε καμία άλλη μεταβλητή X_i. Όταν αυτή η ιδανική κατάσταση συμβαίνει στην πράξη, τότε δεν υπάρχει πολυσυγγραμικότητα. Το άλλο άκρο είναι να υπάρχει *τέλεια συγγραμικότητα*. Για παράδειγμα, ας υποθέσουμε ότι έχουμε ένα μοντέλο παλινδρόμησης της Y με δυο ανεξάρτητες μεταβλητές X_1 και X_2. Τέλεια συγγραμικότητα υπάρχει όταν η μια μεταβλητή X μπορεί να εκφραστεί ακριβώς με

όρους της άλλης ανεξάρτητης μεταβλητής, για όλες τις παρατηρήσεις στο σύνολο των δεδομένων.

Δηλαδή:

Οι μεταβλητές X_1 και X_2 έχουν τέλεια συγγραμικότητα, εάν

$$\boldsymbol{X_1 = \alpha + bX_2}$$

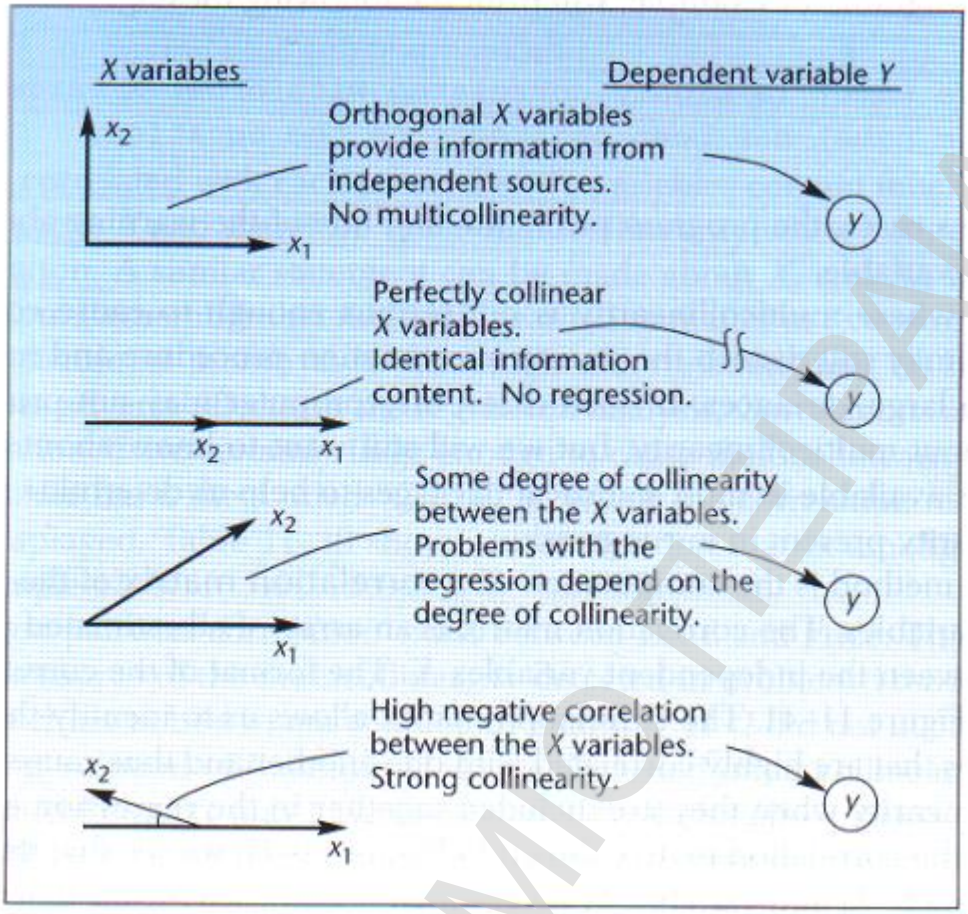
όπου α, b είναι πραγματικοί αριθμοί.

Από την παραπάνω εξίσωση φαίνεται ότι οι δύο μεταβλητές βρίσκονται σε ευθεία γραμμή, και ότι η μία προσδιορίζει τέλεια την άλλη. Στην περίπτωση αυτή, δεν υπάρχει καινούρια πληροφορία που θα μπορούσε να αποκτήσει η Y με την προσθήκη της X_2 σε ένα μοντέλο παλινδρόμησης που ήδη περιέχει την X_1 (ή και αντίστροφα).

Στην πράξη, οι περισσότερες περιπτώσεις είναι ανάμεσα στις δύο παραπάνω ακραίες καταστάσεις. Συχνά, υπάρχει κάποιος βαθμός συγγραμικότητας μεταξύ αρκετών από τις ανεξάρτητες μεταβλητές σε ένα μοντέλο παλινδρόμησης. Ένα μέτρο της συγγραμικότητας μεταξύ δύο μεταβλητών X_i είναι η *συσχέτιση* μεταξύ τους. Εάν και μια προϋπόθεση της παλινδρόμησης είναι ότι οι μεταβλητές X_i είναι σταθερές και μη τυχαίες μεταβλητές, εδώ χαλαρώνουμε λίγο αυτόν τον περιορισμό και μετράμε τη συσχέτιση μεταξύ των ανεξαρτήτων μεταβλητών (αυτό σημαίνει ότι είναι τυχαίες μεταβλητές). Όταν δύο μεταβλητές X_i βρεθούν να έχουν υψηλή συσχέτιση μεταξύ τους, τότε αναμένονται τα δυσμενή αποτελέσματα της πολυσυγγραμικότητας στη διαδικασία υπολογισμού του μοντέλου παλινδρόμησης.

Στην περίπτωση της τέλειας συγγραμικότητας, ο αλγόριθμος της παλινδρόμησης καταρρέει τελείως. Ακόμα και αν βρίσκαμε τις εκτιμήσεις των συντελεστών παλινδρόμησης, η διακύμανσή τους θα έτεινε στο άπειρο. Εάν ο βαθμός συγγραμικότητας δεν είναι πολύ υψηλός, περιμένουμε η διακύμανση των εκτιμήσεων της παλινδρόμησης (και τα τυπικά σφάλματα) να είναι μεγάλη. Όταν οι συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών της παλινδρόμησης είναι μικρού βαθμού, οι επιδράσεις της πολυσυγγραμικότητας μπορεί να μην είναι πολύ σοβαρές. Σε περιπτώσεις, όμως, ισχυρών συσχετίσεων, το πρόβλημα αυτό μπορεί να επηρεάσει αρκετά την παλινδρόμηση, έτσι ώστε να πρέπει να ληφθούν διορθωτικά μέτρα.

Εάν προσπαθήσουμε να φανταστούμε μια μεταβλητή και την πληροφορία που αυτή περιέχει σαν μια κατεύθυνση στο χώρο, τότε δύο ασυσχέτιστες μεταβλητές μπορούν να αναπαρασταθούν ως ορθογώνιες κατευθύνσεις στο χώρο, δηλαδή να σχηματίζουν γωνία 90^o μεταξύ τους. Οι τέλεια συσχετισμένες μεταβλητές έχουν κατευθύνσεις που σχηματίζουν γωνία 0^o ή 180^o μεταξύ τους, ανάλογα με το αν έχουν τέλεια θετική ή αρνητική συσχέτιση, αντίστοιχα. Μεταβλητές που είναι μερικώς συσχετισμένες σχηματίζουν κατευθύνσεις με γωνία μεταξύ των 0^o και 90^o για θετική συσχέτιση, ή μεταξύ των 90^o και 180^o για αρνητική συσχέτιση. Όσο πιο κοντά βρίσκεται η γωνία στις 0^o ή στις 180^o, τόσο μεγαλύτερη είναι η συγγραμικότητα. Το Σχήμα 4.1 απεικονίζει τις διάφορες κατευθύνσεις των μεταβλητών στο χώρο.



Σχήμα 4.1: Αναπαράσταση Συγγραμικότητας ως Σχέση μεταξύ δύο Διευθύνσεων στο Χώρο

ø ***Αιτίες Εμφάνισης Πολυσυγγραμικότητας***

Υπάρχουν αρκετοί λόγοι που μπορεί να προκαλέσουν πολυσυγγραμικότητα. Μια μέθοδος συλλογής δεδομένων μπορεί να προκαλέσει πολυσυγγραμικότητα, εάν χωρίς να το προσέξουμε συλλέγουμε δεδομένα με σχετικές τιμές σε διάφορες μεταβλητές. Σε αυτές τις περιπτώσεις, η βελτίωση της μεθόδου δειγματοληψίας μπορεί να λύσει το πρόβλημα. Σε άλλες περιπτώσεις, οι μεταβλητές μπορεί εκ φύσεως να συσχετίζονται μεταξύ τους, και οι προσαρμογές της δειγματοληψίας να μην έχουν κανένα θετικό αποτέλεσμα. Τότε, μία εκ των συσχετισμένων μεταβλητών πρέπει να απορριφθεί από το μοντέλο, ώστε να αποφευχθεί το πρόβλημα της συγγραμικότητας.

Στις βιομηχανικές διεργασίες, μερικές φορές υπάρχουν φυσικοί περιορισμοί στα δεδομένα. Για παράδειγμα, εάν εφαρμόσουμε ένα μοντέλο παλινδρόμησης μεταξύ της χημικής απόδοσης ενός Y σε σχέση με τη συγκέντρωση δύο στοιχείων X_1 και X_2 και η συνολική ποσότητα των υλικών στη διεργασία είναι σταθερή, τότε καθώς η συγκέντρωση του ενός χημικού αυξάνεται, η ποσότητα του άλλου πρέπει να μειωθεί. Στην περίπτωση αυτή, οι μεταβλητές X_1 και X_2 είναι αρνητικά συσχετισμένες, συνεπώς εμφανίζεται η πολυσυγγραμικότητα.

Άλλος ένας λόγος εμφάνισης συγγραμικότητας είναι η συμπερίληψη στο μοντέλο μεταβλητών X_i υψωμένες σε κάποια δύναμη (μεγαλύτερη από 2). Η συμπερίληψη

μιας μεταβλητής X^2 σε ένα μοντέλο παλινδρόμησης, όπου υπάρχει ήδη η μεταβλητή X, μπορεί να προκαλέσει συγγραμικότητα, αν τα δεδομένα περιορίζονται σε ένα στενό εύρος τιμών.

ø *Εντοπισμός της Ύπαρξης Πολυσυγγραμικότητας*

Πολλά στατιστικά προγράμματα στον ηλεκτρονικό υπολογιστή έχουν ενσωματωμένες προειδοποιήσεις σχετικά με την ύπαρξη πολυσυγγραμικότητας υψηλού βαθμού, που δημιουργούν σοβαρά προβλήματα στο μοντέλο παλινδρόμησης. Ωστόσο, ακόμα και σε πιο "ελαφριές" περιπτώσεις εμφάνισης του προβλήματος, μπορεί τα αποτελέσματα να μην είναι αντιπροσωπευτικά και να εμφανίζονται εκτιμητές με μεγάλες διακυμάνσεις. Αυτό σημαίνει ότι μερικοί συντελεστές παλινδρόμησης φαίνονται στατιστικά μη σημαντικοί, ενώ στην πραγματικότητα είναι. Στην περίπτωση αυτήν, ο ηλεκτρονικός υπολογιστής μπορεί να μην μας προειδοποιήσει, εμείς όμως θα πρέπει να ελέγξουμε για να μάθουμε. Υπάρχουν κυρίως δύο μέθοδοι για τον έλεγχο ύπαρξης πολυσυγγραμικότητας.

Η πρώτη μέθοδος είναι ο υπολογισμός του πίνακα συσχετίσεων των ανεξάρτητων μεταβλητών της παλινδρόμησης. Ο πίνακας συσχετίσεων μας επιτρέπει να εντοπίζουμε τις ανεξάρτητες μεταβλητές που συσχετίζονται ισχυρά μεταξύ τους και επομένως προκαλούν το πρόβλημα της πολυσυγγραμικότητας, όταν περιλαμβάνονται μαζί σε μια εξίσωση παλινδρόμησης. Όταν δύο μεταβλητές συσχετίζονται, αυτό σημαίνει ότι περιέχουν και οι δύο περίπου την ίδια πληροφορία για τη μεταβλητή Y.

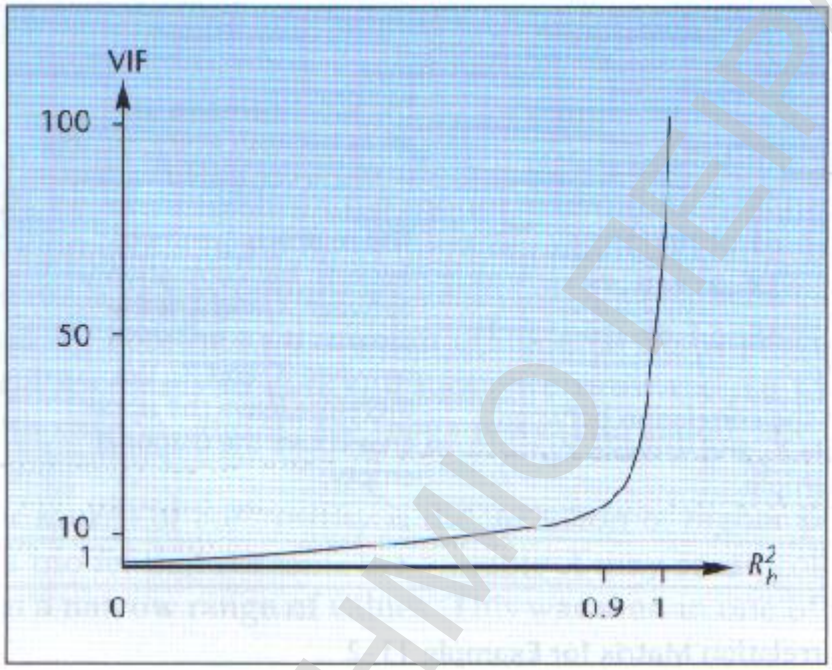
Ωστόσο, σε πολλές περιπτώσεις που υπάρχουν συσχετίσεις μεταξύ περισσοτέρων από δύο ζευγών, η πολυσυγγραμικότητα μπορεί να προκαλείται από το ζεύγος με τη μικρότερη συσχέτιση. Επίσης, μπορεί και να προκαλείται από πιο περίπλοκες συσχετίσεις των δεδομένων, σε σχέση με τις συσχετίσεις ανά ζεύγος. Γι' αυτό, υπάρχει και η δεύτερη μέθοδος εντοπισμού της πολυσυγγραμικότητας: ο πληθωριστικός παράγοντας διασποράς (Variance Inflation Factor – VIF).

Ο βαθμός πολυσυγγραμικότητας που οφείλεται σε μια μεταβλητή X_h, όταν στο μοντέλο υπάρχουν ήδη οι μεταβλητές $X_1,\ldots, X_k$, είναι μια συνάρτηση της πολλαπλής συσχέτισης μεταξύ της X_h και των άλλων μεταβλητών $X_1,\ldots, X_k$. Αν υποθέσουμε ότι εφαρμόζουμε ένα μοντέλο παλινδρόμησης, όχι της Y, αλλά της X_h σε σχέση με τις μεταβλητές $X_1,\ldots, X_k$, τότε προκύπτει ένας συντελεστής προσδιορισμού R_h^2. Ο συντελεστής αυτός αποτελεί ένα μέτρο της πολυσυγγραμικότητας που ασκείται από τη μεταβλητή X_h. Θυμίζουμε ότι ένα από τα προβλήματα της πολυσυγγραμικότητας είναι η σημαντική αύξηση των τιμών της διασποράς των εκτιμήσεων των συντελεστών παλινδρόμησης. Για να μετρήσουμε το φαινόμενο αυτό, χρησιμοποιούμε τον παράγοντα *VIF* της μεταβλητής X_h:

$$\boldsymbol{VIF(X_h) = 1/ (1-R_h^2)}$$

όπου R_h^2 είναι ο R^2 που προκύπτει από την παλινδρόμηση της μεταβλητής X_h, ως εξαρτημένης μεταβλητής, σε σχέση με τις άλλες X μεταβλητές που χρησιμοποιήθηκαν στην αρχική εξίσωση πρόβλεψης της Y.

Ο *VIF* της μεταβλητής X_h είναι ίσος με το λόγο της διασποράς της εκτίμησης του συντελεστή b_h της αρχικής παλινδρόμησης, ως προς τη διασπορά της εκτίμηση b_h, εάν η X_h δεν είναι συγγραμική με καμία άλλη εξαρτημένη μεταβλητή. Ο *VIF* είναι ο πληθωριστικός παράγοντας διασποράς των εκτιμήσεων, σε σύγκριση με το ποια θα ήταν η διασπορά, εάν η X_h δεν ήταν συγγραμική με καμία από τις άλλες X μεταβλητές στην παλινδρόμηση. Η χαρακτηριστική γραφική παράσταση της σχέσης μεταξύ του R_h^2 και του *VIF* φαίνεται στο παρακάτω σχήμα.



Σχήμα 4.2: Σχέση μεταξύ R_h^2 και *VIF*

Όπως φαίνεται από το παραπάνω σχήμα, όταν το R_h^2 αυξάνεται από 0,9 σε 1, το *VIF* αυξάνεται δραματικά. Στην ουσία, για R_h^2=1, το *VIF* προσεγγίζει το άπειρο. Ωστόσο, ακόμα και για τιμές R_h^2 μικρότερες από 0,9, το *VIF* είναι αρκετά μεγάλο. Μία τιμή *VIF*=6, για παράδειγμα, σημαίνει ότι η διασπορά του b_h είναι 6 φορές μεγαλύτερη από αυτή που θα έπρεπε να είναι, εάν δεν υπήρχε πολυσυγγραμικότητα.

Μερικές ενδείξεις πολυσυγγραμικότητας, τις οποίες πρέπει να ελέγχουμε είναι οι ακόλουθες:

1. Οι τιμές των διασπορών και των τυπικών σφαλμάτων των εκτιμήσεων των συντελεστών παλινδρόμησης b_i είναι διογκωμένες.
2. Το μέγεθος των εκτιμήσεων των συντελεστών παλινδρόμησης b_i μπορεί να διαφέρουν από αυτό που περιμένουμε.
3. Τα πρόσημα των b_i μπορεί να είναι αντίθετα από τα αναμενόμενα.
4. Η προσθήκη ή απομάκρυνση μεταβλητών προκαλεί μεγάλες αλλαγές στους b_i ή στα πρόσημά τους.
5. Η απομάκρυνση ενός σημείου από τα δεδομένα προκαλεί μεγάλες αλλαγές στους b_i ή στα πρόσημά τους.
6. Σε μερικές περιπτώσεις, τα *F-tests* προκύπτουν να είναι σημαντικά, ενώ κανένα από τα *t-tests* δεν βγαίνει σημαντικό.

Από τις επιδράσεις της πολυσυγγραμικότητας που προαναφέρθηκαν συμπεραίνουμε ότι οι εκτιμήσεις b_i δεν θεωρούνται αξιόπιστες. Η πιο σοβαρή επίδραση είναι η διόγκωση της τιμής της διασποράς, που κάνει μερικές μεταβλητές να φαίνονται μη σημαντικές.

ø ***Λύσεις στο Πρόβλημα της Πολυσυγγραμικότητας***

1. Μία από τις καλύτερες λύσεις στο πρόβλημα της πολυσυγγραμικότητας είναι η απόρριψη κάποιων εκ των συσχετισμένων μεταβλητών από το μοντέλο παλινδρόμησης. Για παράδειγμα, έστω ότι έχουμε μια παλινδρόμηση της *Y* ως προς τις X_1, X_2, X_3 και X_4 και ότι η X_1 συσχετίζεται ισχυρά με τη X_4. Στην περίπτωση αυτή, μεγάλο μέρος της πληροφορίας για την Y που περιέχεται στη X_1 περιέχεται και στη X_4. Εάν απορριφθεί η μία εκ των δύο από το μοντέλο, τότε θα λυθεί το πρόβλημα της πολυσυγγραμικότητας και θα χαθεί κάποια μικρή πληροφορία για την *Y*. Συγκρίνοντας το R^2 με το διορθωμένο R^2 των μοντέλων παλινδρόμησης με και χωρίς μία εκ των μεταβλητών, μπορούμε να αποφασίσουμε ποια από τις δύο ανεξάρτητες μεταβλητές να απορρίψουμε από το μοντέλο. Επιθυμητό είναι να διατηρηθεί ένα υψηλό R^2, γι' αυτό και απορρίπτουμε μία μεταβλητή, αν το R^2 δεν μειώνεται πολύ μετά την απομάκρυνση της μεταβλητής αυτής από το μοντέλο. Όταν το διορθωμένο R^2 αυξάνεται με την απομάκρυνση μιας μεταβλητής, τότε σίγουρα πρέπει να απορρίψουμε τη μεταβλητή αυτή. Ωστόσο, μερικές φορές από την απομάκρυνση κάποιας συγκεκριμένης μεταβλητής μπορεί να εκδηλώσουμε κάποια μεροληψία, το οποίο δεν είναι επιθυμητό. Σε τέτοιες περιπτώσεις, καλό είναι "ζυγίζουμε" τις συνέπειες από τη μεροληψία που προκύπτει με τη διαγραφή μιας μεταβλητής, έναντι της αύξησης της διασποράς της εκτίμησης του συντελεστή παλινδρόμησης, όταν η μεταβλητή περιλαμβάνεται στο μοντέλο.

2. Όταν η πολυσυγγραμικότητα οφείλεται στις μεθόδους δειγματοληψίας που, εκ της φύσεώς τους, τείνουν προς την επιλογή στοιχείων με παρόμοιες τιμές μερικών από τις ανεξάρτητες μεταβλητές, τότε μια αλλαγή στο σχέδιο δειγματοληψίας, ώστε να συμπεριλάβει στοιχεία πέρα από το εύρος της πολυσυγγραμικότητας, μπορεί να μειώσουμε την έκταση του προβλήματος.

3. Μία άλλη μέθοδος που μπορεί να μειώσει ή και να εξαλείψει το φαινόμενο της πολυσυγγραμικότητας είναι ο μετασχηματισμός μερικών από τις μεταβλητές. Ο καλύτερος τρόπος για να γίνει αυτό είναι να σχηματιστούν νέοι συνδυασμοί των μεταβλητών *X* (*logX, 1/X*), οι οποίοι δεν συσχετίζονται μεταξύ τους, και στη συνέχεια να εφαρμοστεί το μοντέλο παλινδρόμησης με τους νέους μετασχηματισμούς και όχι τις αρχικές μεταβλητές. Με τον τρόπο αυτόν, η πληροφορία της αρχικής μεταβλητής διατηρείται, ενώ η πολυσυγγραμικότητα απαλείφεται. Ένας άλλος τρόπος μετασχηματισμού των μεταβλητών είναι το κεντράρισμα των δεδομένων, όπου οι μέσοι αφαιρούνται από τις μεταβλητές, σχηματίζοντας έτσι νέες μεταβλητές που εισέρχονται στο μοντέλο παλινδρόμησης.

4. Η πολυσυγγραμικότητα μπορεί να εξαλειφθεί με την εφαρμογή μιας διαδικασίας διαφορετικής από αυτή των ελαχίστων τετραγώνων, η οποία

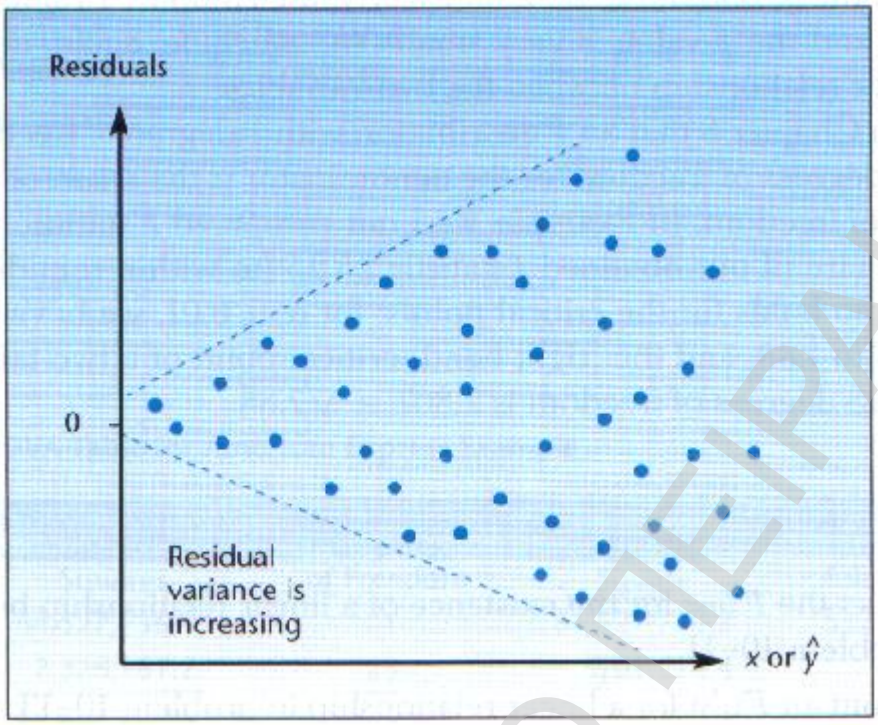
καλείται ραχοειδής παλινδρόμηση (***ridge regression***). Οι εκτιμήσεις των συντελεστών παλινδρόμησης εμφανίζουν κάποια μεροληψία, όμως μερικές φορές είναι προτιμότερο να ανεχτούμε λίγη αμεροληψία στις εκτιμήσεις αυτές, προκειμένου να μειωθούν οι υψηλές τιμές των διακυμάνσεων που προκύπτουν από την πολυσυγγραμικότητα. [11, 36, 42, 43]

3.9.3. Εξέταση των Καταλοίπων

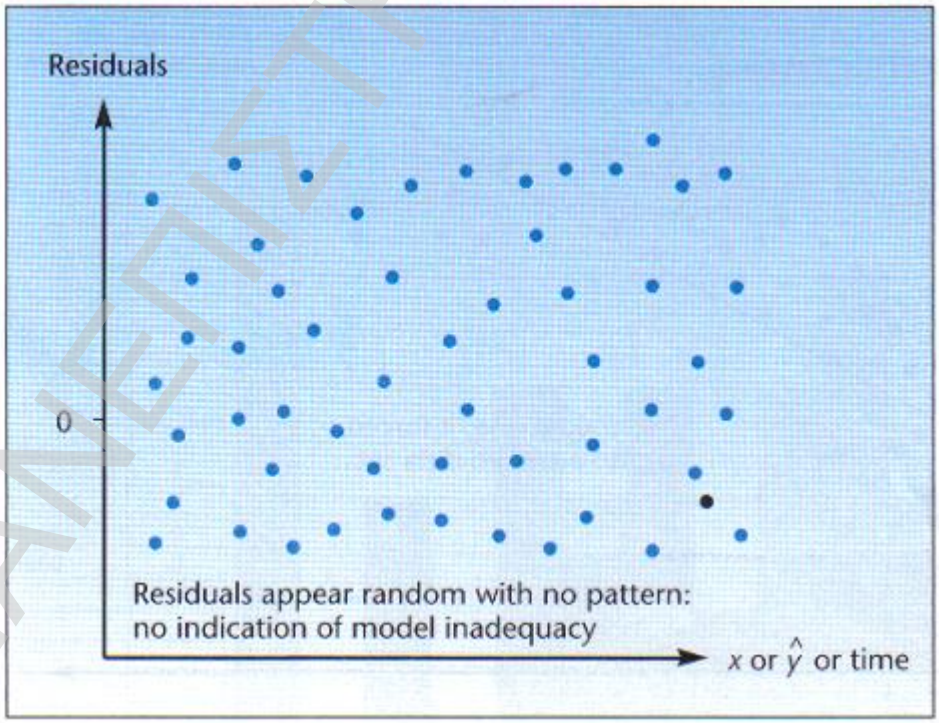
Όπως έχουμε δει και σε προηγούμενη ενότητα, τα κατάλοιπα ορίζονται ως οι n διαφορές $e_i = Y_i - \hat{Y}_i$, i = 1, 2, …, n, όπου Y_i είναι μία παρατήρηση και $\hat{Y}_i$ η τιμή που αντιστοιχεί στην αντίστοιχη παρατήρηση και προκύπτει από την εξίσωση παλινδρόμησης. Ουσιαστικά, τα κατάλοιπα είναι η διαφορά μεταξύ αυτού που πραγματικά παρατηρείται και αυτού που προβλέπεται από την παλινδρόμηση, δηλαδή η ποσότητα που η εξίσωση παλινδρόμησης δεν μπορεί να εξηγήσει. Όσον αφορά την εφαρμογή ενός μοντέλου παλινδρόμησης, υπάρχουν κάποιες προϋποθέσεις που πρέπει να ακολουθούν τα κατάλοιπα: να είναι ανεξάρτητα μεταξύ τους και με τις άλλες μεταβλητές, να έχουν μέσο μηδέν, σταθερή διακύμανση και να ακολουθούν κανονική κατανομή. Η τελευταία προϋπόθεση απαιτείται για τη διεξαγωγή των ***F-tests***. Επομένως, εάν το προσαρμοσμένο μοντέλο είναι σωστό, τα κατάλοιπα πρέπει να τείνουν να ικανοποιούν τις παραπάνω προϋποθέσεις, ή τουλάχιστον να μην οδηγούν σε κατάρριψη αυτών. Στη συνέχεια ακολουθούν διάφοροι τρόποι για τον έλεγχο των παραπάνω υποθέσεων.

ø *Έλεγχος Σταθερότητας της Διακύμανσης*

Μια γραφική παράσταση των σφαλμάτων με τις ανεξάρτητες μεταβλητές X ή τις προβλεπόμενες τιμές $\hat{Y}$ μπορεί να αποδείξει εάν η διασπορά των σφαλμάτων παραμένει σταθερή ή όχι. Η διασπορά των καταλοίπων φαίνεται από το εύρος του διαγράμματος διασποράς των καταλοίπων, καθώς η μεταβλητή X αυξάνεται. Εάν αυτό το εύρος αυξάνεται ή μειώνεται καθώς το X αυξάνεται, τότε δεν ισχύει η προϋπόθεση της σταθερής διακύμανσης. Το πρόβλημα αυτό καλείται *ετεροσκεδαστικότητα*. Όταν εμφανίζεται αυτό το πρόβλημα, δεν μπορούμε να χρησιμοποιήσουμε τη μέθοδο ελαχίστων τετραγώνων για την εκτίμηση της παλινδρόμησης και πρέπει να εφαρμόσουμε μία πιο πολύπλοκη μέθοδο, που ονομάζεται γενικευμένη (ή σταθμισμένη) μέθοδος ελαχίστων τετραγώνων. Το Σχήμα 4.3 δείχνει πώς είναι ένα διάγραμμα καταλοίπων όταν υπάρχει ετεροσκεδαστικότητα, ενώ το Σχήμα 4.4 όταν δεν εμφανίζεται το πρόβλημα.



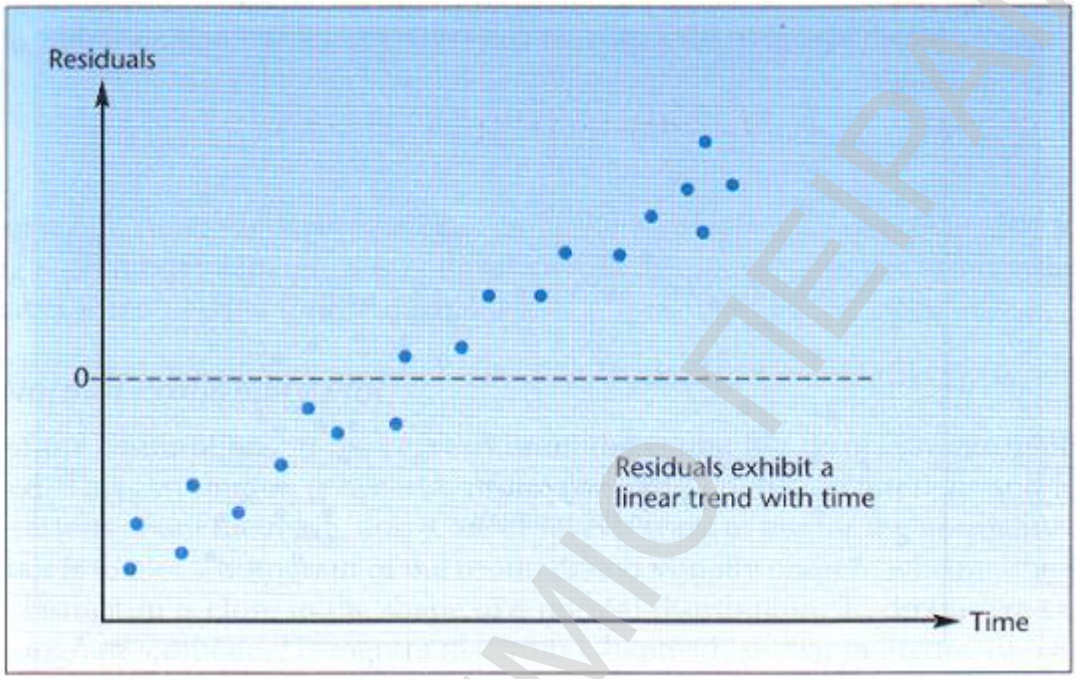
<u>Σχήμα 4.3</u>: Διάγραμμα Καταλοίπων Με Ετεροσκεδαστικότητα



<u>Σχήμα 4.4</u>: Διάγραμμα Καταλοίπων Χωρίς Ετεροσκεδαστικότητα

ø ***Έλεγχος για Μεταβλητές που Λείπουν***

Το Σχήμα 4.4 δείχνει επίσης πώς πρέπει να είναι τα κατάλοιπα όταν σχεδιάζονται ως προς το χρόνο (ή τη σειρά με την οποία συλλέγονται τα δεδομένα). Δεν πρέπει να εμφανίζεται κάποια τάση στα κατάλοιπα ως προς το χρόνο. Το Σχήμα 4.5 δείχνει μια γραμμική τάση των καταλοίπων ως προς το χρόνο.



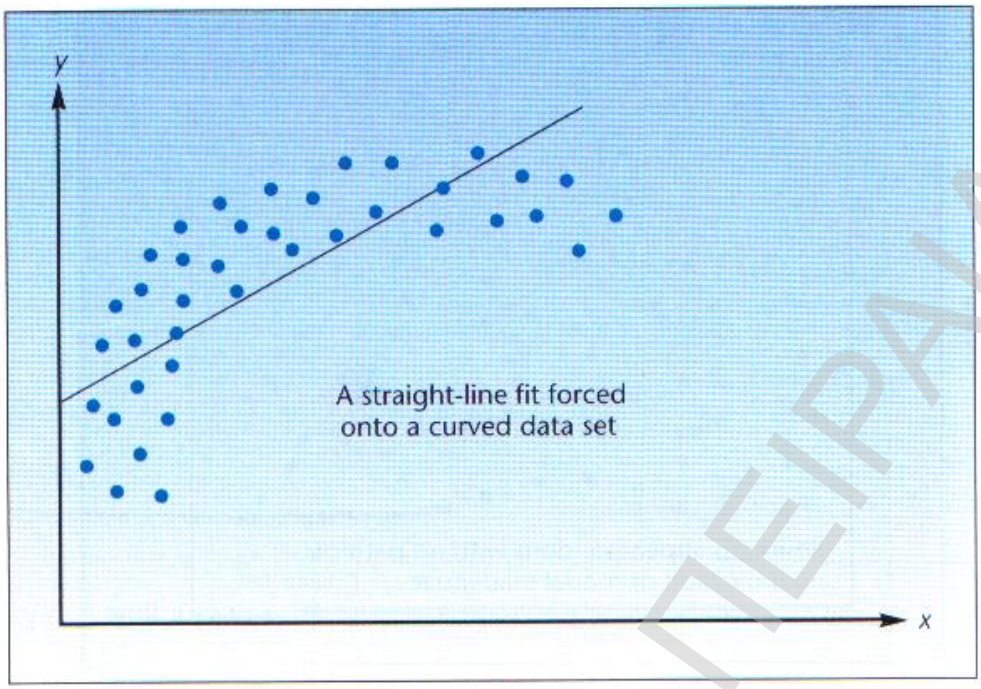
Σχήμα 4.5: Διάγραμμα Καταλοίπων που Εμφανίζουν Τάση με το Χρόνο

Εάν τα κατάλοιπα εμφανίζουν κάποιο σχέδιο (pattern) ως προς το χρόνο, τότε καλό είναι ο χρόνος να συμπεριληφθεί στο μοντέλο ως ανεξάρτητη μεταβλητή. Το ίδιο ισχύει και για οποιαδήποτε άλλη μεταβλητή: αν υπάρχει κάποια τάση ή σχέδιο στο διάγραμμα καταλοίπων ως προς μια μεταβλητή, η μεταβλητή πρέπει να συμπεριληφθεί στο μοντέλο παλινδρόμησης.

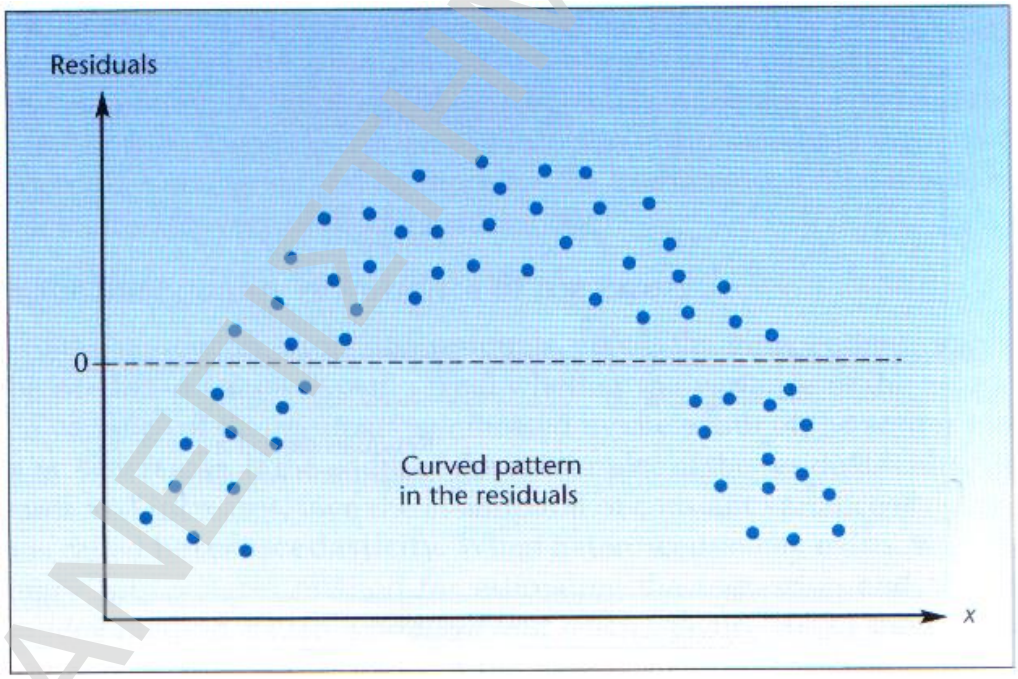
ø ***Εντοπισμός Καμπυλόγραμμης Σχέσης μεταξύ Y και X***

Εάν η σχέση μεταξύ της Y και X είναι καμπυλόγραμμη, τότε πιέζοντας να περάσουμε από τα δεδομένα μία ευθεία γραμμή δεν θα έχει καλά αποτελέσματα ως προς την εφαρμογή του μοντέλου. Αυτό φαίνεται στο Σχήμα 4.6. Στην περίπτωση αυτή, τα κατάλοιπα είναι αρχικά μεγάλα και με αρνητική τιμή, μετά μειώνονται, παίρνουν θετική τιμή, και μετά ξαναπαίρνουν αρνητική τιμή. Τα κατάλοιπα δεν είναι τυχαία και ανεξάρτητα, αλλά επιδεικνύουν καμπυλότητα. Το pattern αυτό φαίνεται στα δύο ακόλουθα σχήματα. Η κατάσταση αυτή μπορεί να βελτιωθεί με την προσθήκη της μεταβλητής X^2 στο μοντέλο.

Υπάρχει και στατιστική μέθοδος που προσδιορίζει την έλλειψη προσαρμογής του μοντέλου στα δεδομένα (lack of fit), η οποία σπάει το άθροισμα τετραγώνων των σφαλμάτων σε άθροισμα τετραγώνων που οφείλεται σε καθαρό σφάλμα και σε άθροισμα τετραγώνων εξαιτίας έλλειψης προσαρμογής.



Σχήμα 4.6: Αποτέλεσμα Προσαρμογής Ευθείας Γραμμής σε Καμπυλόγραμμα Δεδομένα

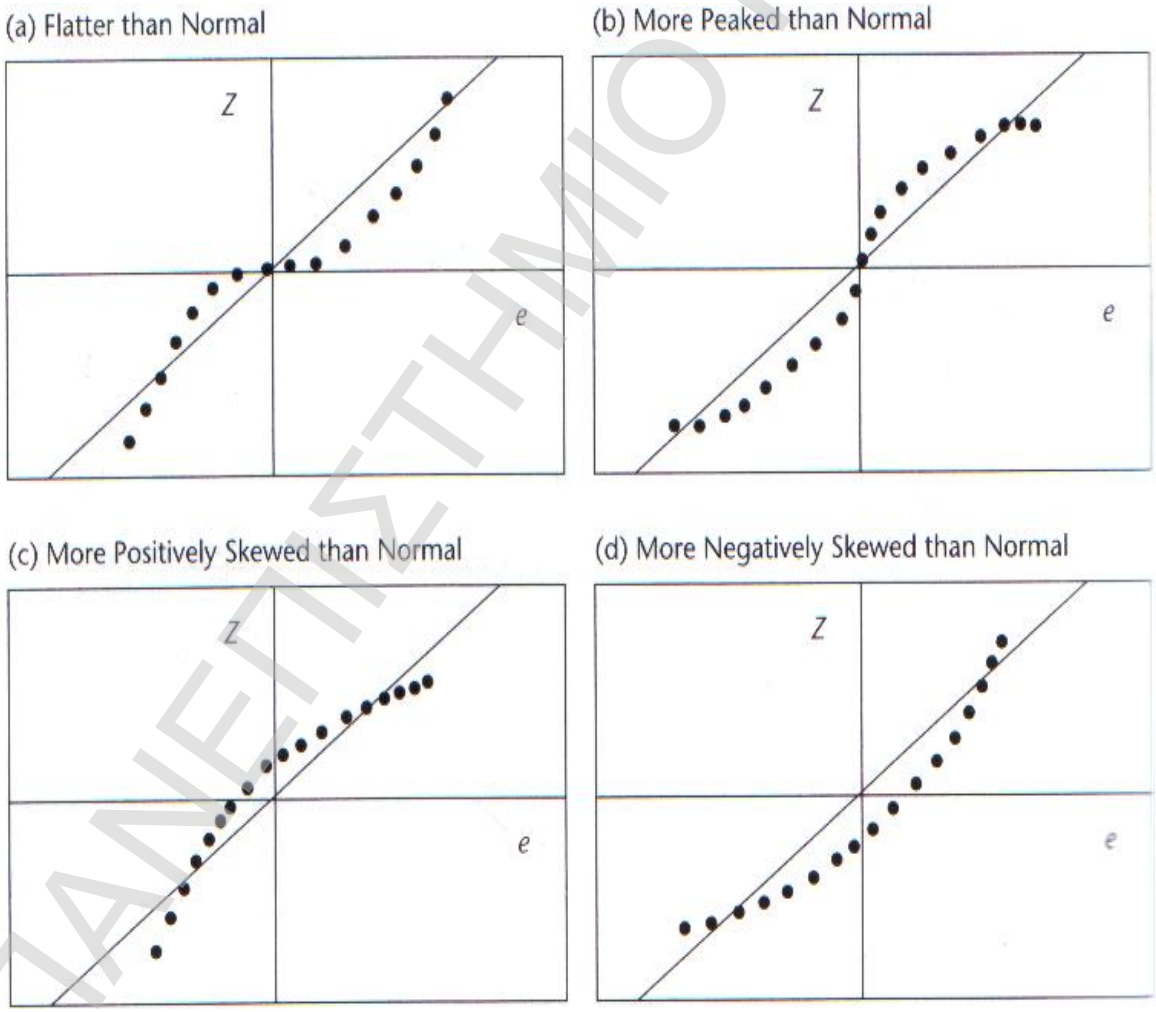


Σχήμα 4.7: Τάση των Καταλοίπων όταν Ευθεία Γραμμή Προσαρμόζεται σε Καμπυλόγραμμα Δεδομένα

ø ***Διάγραμμα Ελέγχου Κανονικότητας***

Η προϋπόθεση ύπαρξης κανονικότητας στα κατάλοιπα είναι απαραίτητο να ικανοποιείται, έτσι ώστε να υπολογίζονται τα διαστήματα πρόβλεψης και για να ισχύουν τα tests των μηδενικών υποθέσεων της παλινδρόμησης. Ένας τρόπος για να ελεγχθεί η κανονικότητα των καταλοίπων είναι να σχεδιάσουμε το ιστόγραμμα των καταλοίπων και να παρατηρήσουμε με το μάτι αν το σχήμα του ιστογράμματος είναι παρόμοιο με το σχήμα της κανονικής κατανομής (καμπάνα).

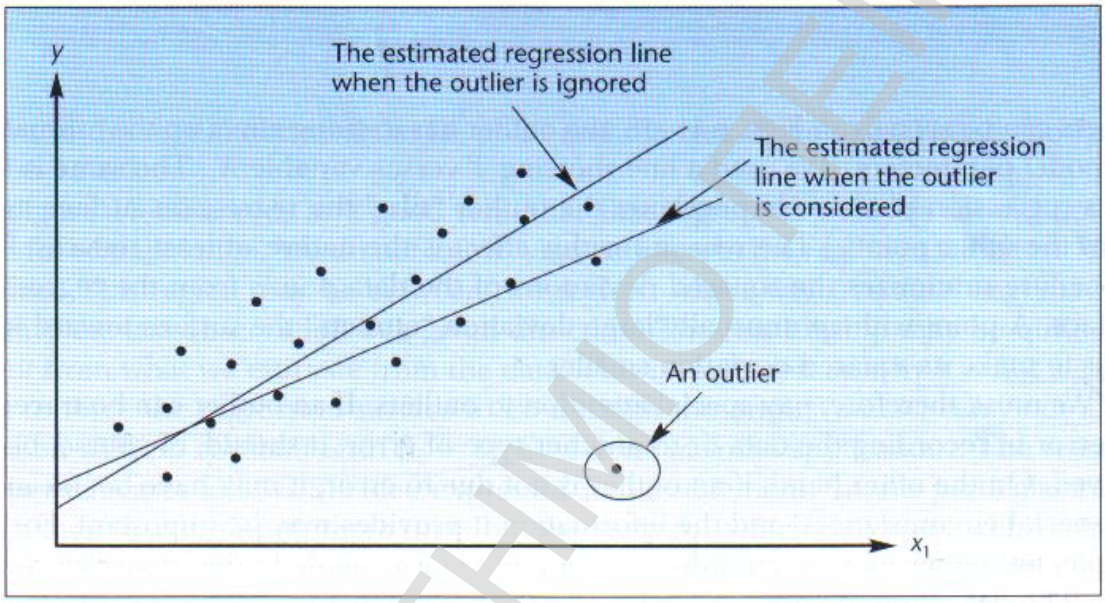
Μία καλύτερη μέθοδος για τον έλεγχο της κανονικότητας των καταλοίπων είναι η χρήση ενός διαγράμματος κανονικής κατανομής (normal probability plot). Στο διάγραμμα αυτό, οι τιμές των καταλοίπων απεικονίζονται στον οριζόντιο άξονα και οι αντίστοιχες τιμές *z* της κανονικής κατανομής στον κατακόρυφο άξονα. Εάν τα σφάλματα είναι κανονικά κατανεμημένα, πρέπει να ευθυγραμμίζονται πάνω στην ευθεία γραμμή που εμφανίζεται στο Σχήμα 4.8. Στο βαθμό που απέχουν από αυτή την ευθεία γραμμή, αντιστοιχεί και ο βαθμός απόκλισής τους από την κανονικότητα.



Σχήμα 4.8: Μορφές Μη Κανονικών Κατανομών σε Διαγράμματα Ελέγχου Κανονικότητας

ø *Έκτροπες Παρατηρήσεις και Παρατηρήσεις με Ισχυρή Επιρροή*

Ως έκτροπη παρατήρηση θεωρείται μία ακραία παρατήρηση. Είναι ένα σημείο που βρίσκεται μακριά από το υπόλοιπο σετ δεδομένων. Εξαιτίας αυτού, τα έκτροπα μπορεί να ασκούν μεγαλύτερη επίδραση στις εκτιμήσεις των παραμέτρων της παλινδρόμησης που προκύπτουν με τη μέθοδο ελαχίστων τετραγώνων από ό,τι άλλες παρατηρήσεις. Ο λόγος που συμβαίνει αυτό φαίνεται διαγραμματικά στο Σχήμα 4.9. Στο διάγραμμα αυτό φαίνεται η εκτιμώμενη από τα ελάχιστα τετράγωνα γραμμή παλινδρόμησης όταν δεν συμπεριλαμβάνεται η έκτροπη παρατήρηση, και η γραμμή παλινδρόμησης που συμπεριλαμβάνει την έκτροπη παρατήρηση.

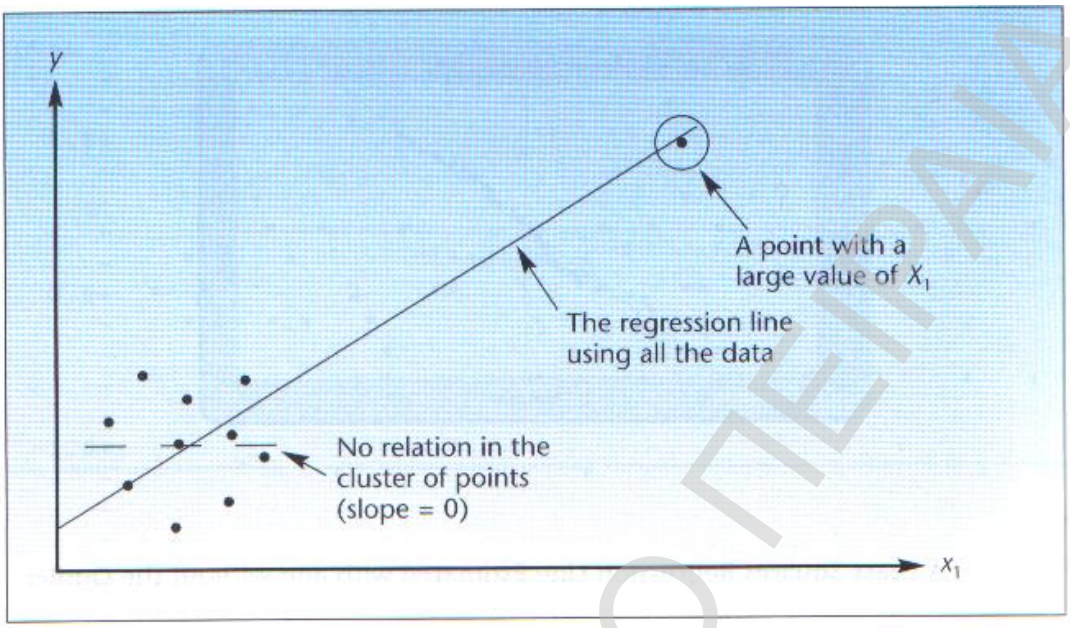


Σχήμα 4.9: Γραμμή Παλινδρόμησης με τη Μέθοδο Ελαχίστων Τετραγώνων Με και Χωρίς την Έκτροπη Παρατήρηση

Η έκτροπη παρατήρηση έχει σημαντική επίδραση στην εκτίμηση των παραμέτρων του μοντέλου. Ο λόγος που συμβαίνει αυτό είναι λόγω της φύσης των ελαχίστων τετραγώνων. Η διαδικασία ελαχιστοποιεί τις τυπικές αποκλίσεις των δεδομένων από την επιφάνεια παλινδρόμησης. Ένα σημείο με ασυνήθιστα μεγάλη τυπική απόκλιση ελκύει την επιφάνεια παλινδρόμησης προς τη μεριά του, έτσι ώστε να μικρύνει την τυπική του απόκλιση. Γι' αυτό, πρέπει να δίνουμε μεγάλη προσοχή στην ύπαρξη εκτρόπων. Εάν ένα έκτροπο βρεθεί ότι οφείλεται σε κάποιο σφάλμα κατά την καταχώρηση των δεδομένων ή σε κάποιο άλλο τύπο σφάλματος, τότε πρέπει να απομακρυνθεί. Από την άλλη, όμως, αν το έκτροπο δεν οφείλεται σε κάποιο σφάλμα, μπορεί αυτό να έχει προκληθεί κάτω από ειδικές συνθήκες, οπότε η πληροφορία που παρέχει μπορεί να είναι σημαντική. Σε τέτοιες περιπτώσεις καλό είναι να μην απορρίψουμε τις παρατηρήσεις αυτές, αλλά να χρησιμοποιήσουμε κάποια μέθοδο εναλλακτική αυτής των ελαχίστων τετραγώνων.

Ένα σημείο των δεδομένων που βρίσκεται μακριά από τα υπόλοιπα σημεία σε κάποια κατεύθυνση X_i ονομάζεται παρατήρηση με ισχυρή επιρροή, όταν επηρεάζει σημαντικά την προσαρμογή του μοντέλου παλινδρόμησης. Χαρακτηριστικά αυτό φαίνεται στο Σχήμα 4.10, όπου μαζί με το σημείο αυτό η κλίση της ευθείας είναι

διάφορη του μηδενός, ενώ χωρίς αυτό το σημείο η κλίση είναι μηδενική, δηλαδή δεν υπάρχει μοντέλο παλινδρόμησης.



Σχήμα 4.10: Επίδραση μιας Παρατήρησης Μακριά από το Σύνολο των Δεδομένων

ø *Αυτοσυσχέτιση Καταλοίπων*

Η αυτοσυσχέτιση είναι η συσχέτιση που υπάρχει μεταξύ των τιμών μιας μεταβλητής με τις τιμές της ίδιας μεταβλητής που έχουν χρονική καθυστέρηση μιας ή περισσοτέρων περιόδων. Για παράδειγμα, όταν παρατηρείται αυτοσυσχέτιση μεταξύ των καταλοίπων με χρονική υστέρηση 1 (μία περίοδος) υπάρχει συσχέτιση μεταξύ των τιμών e_i και e_{i-1}. Η συσχέτιση μεταξύ των σφαλμάτων του πληθυσμού ε_i και ε_{i-1} με χρονική υστέρηση 1 συμβολίζεται με ρ_1, με υστέρηση 2 με ρ_2 κ.ο.κ. Αντίστοιχα, η αυτοσυσχέτιση των σφαλμάτων του δείγματος συμβολίζεται με r_1, r_2 κλπ.

Η προϋπόθεση ότι τα σφάλματα της παλινδρόμησης δεν συσχετίζονται μεταξύ τους σημαίνει ότι αυτά δεν συσχετίζονται σε οποιαδήποτε χρονική υστέρηση (*lag*). Ισχύει, δηλαδή, ότι $\rho_1 = \rho_2 = \rho_3 = \ldots = 0$. Ένα στατιστικό τεστ αναπτύχθηκε το 1951 από τους Durbin και Watson για να ελεγχθεί αν παραβιάζεται αυτή η υπόθεση. Το ***Durbin-Watson test*** ελέγχει την ύπαρξη μόνο για την ύπαρξη πρώτης-τάξεως (lag 1) αυτοσυσχέτισης με τις ακόλουθες υποθέσεις:

$$H_o: \rho_1 = 0$$
$$H_1: \rho_1 \neq 0$$

Ενώ το *Durbin-Watson statistic* είναι το εξής:

$$d = \frac{\sum_{i=2}^{n}(e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2}$$

Χρησιμοποιώντας ένα συγκεκριμένο επίπεδο εμπιστοσύνης α (0,05 ή 0,01) από στατιστικούς πίνακες, μπορούμε να διεξάγουμε είτε ένα τεστ για $\rho_1 < 0$ είτε για $\rho_1 > 0$. Το τεστ έχει δύο κρίσιμα σημεία για τον έλεγχο ύπαρξης θετικής αυτοσυσχέτισης. Όταν το ***test statistic d*** είναι αριστερά από το μικρότερο κρίσιμο σημείο d_L, συμπεραίνουμε ότι υπάρχει ένδειξη θετικής αυτοσυσχέτισης των σφαλμάτων με lag 1. Όταν το d βρίσκεται μεταξύ του d_L και του μεγαλύτερου κρίσιμου σημείου d_U, δεν μπορούμε να καταλήξουμε σε κάποιο συμπέρασμα. Όταν το d είναι μεγαλύτερο από το d_U, συμπεραίνουμε ότι δεν υπάρχει ένδειξη αυτοσυσχέτισης τάξεως 1. Παρομοίως για την περίπτωση αρνητικής αυτοσυσχέτισης, αν το d είναι μεγαλύτερο από $4 - d_L$, υπάρχει ένδειξη αρνητικής αυτοσυσχέτισης τάξεως 1. Αν το d είναι μεταξύ $4 - d_U$ και $4 - d_L$, το τεστ δεν βγάζει συμπέρασμα, και αν το d είναι μικρότερο από $4 - d_U$, δεν υπάρχει ένδειξη αρνητικής αυτοσυσχέτισης. [11, 21, 23, 38]

ΜΕΡΟΣ Β΄ - ΑΠΟΤΕΛΕΣΜΑΤΑ

ΚΕΦΑΛΑΙΟ 5: ΣΤΑΤΙΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

5.1. Μεθοδολογία Στατιστικής Ανάλυσης Δεδομένων

Στο σημείο αυτό ξεκινάει η επεξεργασία των δεδομένων που χρησιμοποιήθηκαν στην παρούσα μελέτη. Τα δεδομένα προέρχονται από μετρήσεις δυόμισι ετών (Ιανουάριος 2003 μέχρι Αύγουστο 2005) στις περιεκτικότητες κάποιων χημικών ενώσεων, σε μερικές βασικές φυσικοχημικές ιδιότητες του τσιμέντου και στις αντοχές του τσιμέντου κατά τη δεύτερη, έβδομη και εικοστή όγδοη ημέρα. Οι μετρήσεις αυτές γίνονται για δύο τύπους τσιμέντου. Επειδή, όμως, ο κάθε τύπος τσιμέντου παράγεται σε δύο διαφορετικούς μύλους, τα δεδομένα κατηγοριοποιούνται περαιτέρω και για κάθε μύλο παραγωγής.

Η στατιστική ανάλυση των δεδομένων αποτελεί στην ουσία το πρώτο βήμα για οποιαδήποτε μελέτη στη στατιστική. Χρειάζεται να μελετηθεί σε βάθος η φύση των δεδομένων όλων των μεταβλητών (ανεξάρτητων και εξαρτημένων) που θα χρησιμοποιηθούν στο μοντέλο παλινδρόμησης. Με τον τρόπο αυτόν μπορούν να εντοπιστούν τα βασικά χαρακτηριστικά των μεταβλητών και να βρεθούν πιθανά προβλήματα ή καταπατήσεις των βασικών προϋποθέσεων που απαιτούνται για τη μετέπειτα ανάλυση.

Πιο συγκεκριμένα, η στατιστική ανάλυση των δεδομένων εξετάζει τη μέτρηση της κεντρικής τάσης (μέσος όρος), τη μέτρηση της διασποράς (διακύμανση, τυπική απόκλιση, εύρος) και τη μέτρηση του σχήματος (ασυμμετρία, κύρτωση). Επίσης, με βάση τα παραπάνω μέτρα και διάφορα διαγράμματα ελέγχονται η ύπαρξη κανονικότητας (έλεγχος προτυποποιημένης ασυμμετρίας και κύρτωσης ή έλεγχος μέσω ιστογράμματος), τα διαστήματα εμπιστοσύνης για τις τιμές κάθε μεταβλητής και η ύπαρξη εκτρόπων παρατηρήσεων, οι οποίες πρέπει να ελεγχθούν και, αν οφείλονται σε συγκεκριμένο αίτιο πέρα από την κανονική ροή των διαδικασιών μέτρησης ή παραγωγής, να απορριφθούν.

Πριν την ανάλυση για κάθε τύπο τσιμέντου OPC ή CEM II 42,5 και για κάθε μύλο παραγωγής αυτών, παρατίθενται στον Πίνακα 5.1 όλες οι συντομογραφίες και συμβολισμοί που χρησιμοποιούνται.

Πίνακας 5.1: Συμβολισμοί Μεταβλητών που Αναλύονται

Συμβολισμός	Επεξήγηση
OPC	Απλό Τσιμέντο Portland (Ordinary Portland Cement)
CEM II 42,5	Σύνθετο Τσιμέντο Portland
SiO_2	Οξείδιο του Πυριτίου (% κατά βάρος περιεκτικότητα)
Al_2O_3	Οξείδιο του Αργιλίου (% κατά βάρος περιεκτικότητα)
Blaine	Λεπτότητα (cm^2/gr ειδική επιφάνεια)
IR	Αδιάλυτο Υπόλειμμα (Insoluble Residue) (%)
LOI	Απώλεια Πύρωσης (Loss Of Ignition) (%)
Clk	Κλίνκερ (% περιεκτικότητα)
Gyp	Γύψος (% περιεκτικότητα)
Est2	Αντοχή σε θλίψη κατά τη δεύτερη ημέρα (Nt/mm^2)
Est7	Αντοχή σε θλίψη κατά την έβδομη ημέρα (Nt/mm^2)
Est28	Αντοχή σε θλίψη κατά την εικοστή όγδοη ημέρα (Nt/mm^2)
MT_i	Μύλος Παραγωγής Τσιμέντου

5.2. Σύνθετο Τσιμέντο Portland CEM II 42,5 – Μύλος Παραγωγής 1

5.2.1. Μεταβλητή SiO_2

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

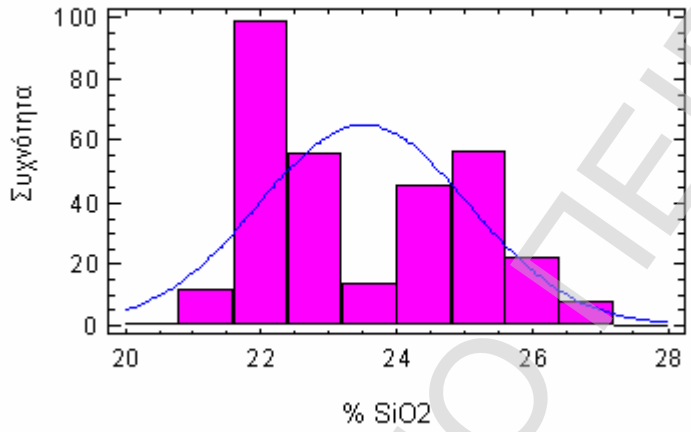
Από την επεξεργασία με το λογισμικό στατιστικού περιεχομένου Statgraphics Plus 5.1 προκύπτουν οι εξής τιμές των βασικών στατιστικών χαρακτηριστικών και τα διαστήματα εμπιστοσύνης:

Στατιστικά Μέτρα για SiO_2
Αριθμός Παρατηρήσεων = 315
Μέσος Όρος = 23,5141
Διασπορά = 2,3734
Τυπική Απόκλιση = 1,54059
Τυπικό Σφάλμα = 0,0868022
Ελάχιστη Τιμή = 20,41
Μέγιστη Τιμή = 26,88
Εύρος = 6,47
Προτυποποιημένη Ασυμμετρία = 2,42985
Προτυποποιημένη Κύρτωση = -4,69557

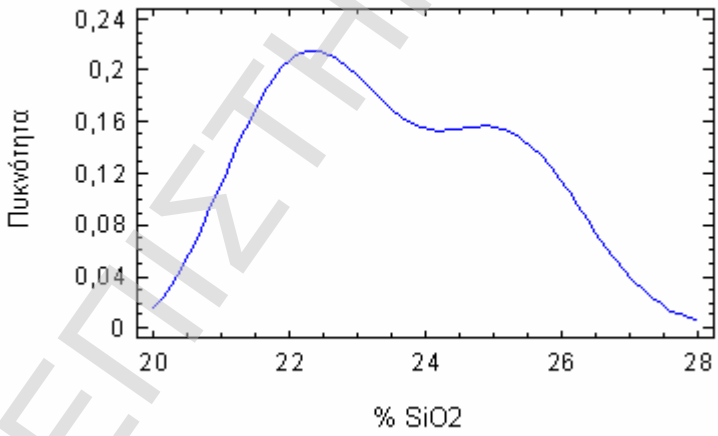
Υπάρχουν 315 σημεία στο σετ δεδομένων για τη διάρκεια των δυόμισι ετών. Ο μέσος όρος των παρατηρήσεων είναι περίπου 23,51, με μέγιστη τιμή 26,88 και ελάχιστη 20,41. Η διασπορά είναι 2,3734, ενώ η τυπική απόκλιση είναι 1,54059. Σημαντικές παράμετροι που πρέπει να ελέγχονται είναι η προτυποποιημένη ασυμμετρία και κύρτωση. Οι τιμές των παραμέτρων αυτών είναι ενδεικτικές για την ύπαρξη κανονικότητας. Εάν οι τιμές της προτυποποιημένης ασυμμετρίας και κύρτωσης είναι μεταξύ του διαστήματος [-2, 2], τότε η μεταβλητή μπορεί να θεωρηθεί ότι ακολουθεί κανονική κατανομή. Στη συγκεκριμένη περίπτωση, η μεταβλητή SiO_2 δεν είναι κανονικά κατανεμημένη.

Έλεγχος Κανονικότητας με Διαγράμματα

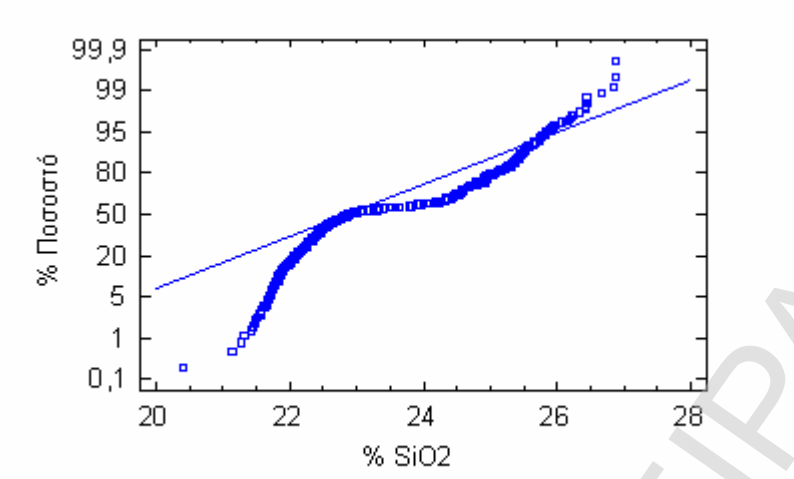
Από τον πρώτο τρόπο που αναφέρθηκε αμέσως παραπάνω, οι τιμές της προτυποποιημένης ασυμμετρίας και κύρτωσης δεν είναι μεταξύ του διαστήματος [-2,2], συνεπώς η μεταβλητή δεν ακολουθεί κανονική κατανομή. Αυτό φαίνεται και από το ιστόγραμμα, το διάγραμμα ίχνους της πυκνότητας, το διάγραμμα ελέγχου κανονικότητας, αλλά και από το θηκόγραμμα, το οποίο μπορεί να χρησιμοποιηθεί και για την εύρεση εκτρόπων παρατηρήσεων.



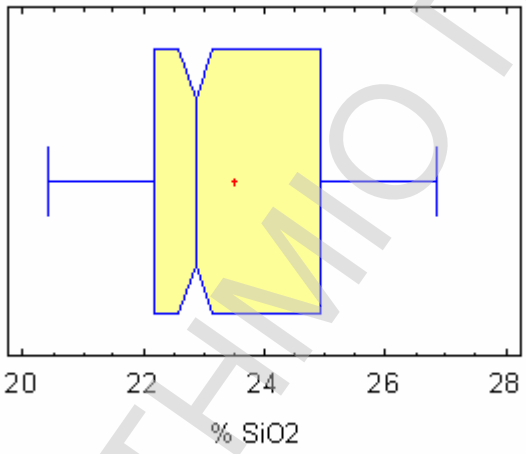
Διάγραμμα 5.1: Ιστόγραμμα της μεταβλητής SiO_2



Διάγραμμα 5.2: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής SiO_2



Διάγραμμα 5.3: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής SiO_2

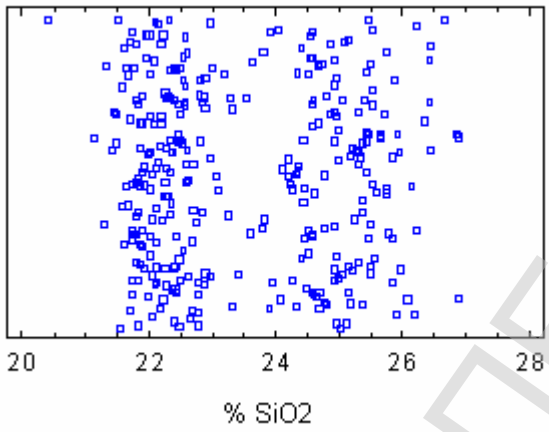


Διάγραμμα 5.4: Θηκόγραμμα της μεταβλητής SiO_2

Από το ιστόγραμμα βλέπουμε ότι η μεταβλητή απέχει πολύ από το κλασικό σχήμα "καμπάνας" της κανονικής κατανομής, το οποίο φαίνεται και από το διάγραμμα του διάγραμμα ίχνους της πυκνότητας. Επίσης, από το Διάγραμμα Ελέγχου Κανονικότητας κανονικά θα έπρεπε η μεταβλητή να "πέφτει" πάνω στην ευθεία γραμμή, το οποίο όμως δεν συμβαίνει. Στο θηκόγραμμα φαίνεται ότι ο μέσος όρος (κόκκινη κουκίδα) και η διάμεσος (μπλε κάθετη γραμμή) είναι αρκετά άνισα, οπότε η SiO_2 έχει πρόβλημα κανονικότητας. Από το τελευταίο διάγραμμα φαίνεται επίσης ότι δεν υπάρχει κάποια έκτροπη παρατήρηση.

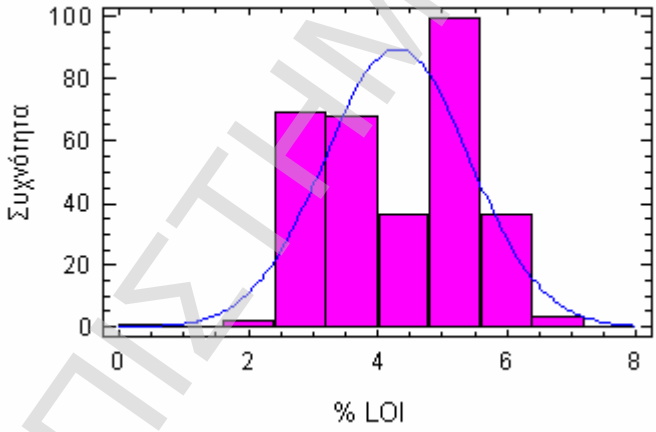
Ä Ωστόσο, ένα **σημαντικό πόρισμα** που προκύπτει από το ιστόγραμμα και το διάγραμμα ίχνους της πυκνότητας είναι ότι υπάρχουν **δύο μέσοι**. Δηλαδή, φαίνεται ότι κάτι έχει αλλάξει στην παραγωγική διαδικασία, ώστε να τεθούν νέοι στόχοι για την τιμή της συγκέντρωσης της μεταβλητής SiO_2. Όντως, από εκτενή έλεγχο των δεδομένων και με εξακρίβωση από τον υπεύθυνο της τσιμεντοβιομηχανίας, **η παραγωγή του τσιμέντου CEM II 42,5 άλλαξε τον Ιούλιο του 2004**. Συνεπώς, κρίθηκε σκόπιμο η στατιστική ανάλυση για αυτόν τον τύπο τσιμέντου να διεκπεραιωθεί για τα δεδομένα από τον Ιούλιο 2004 και μετά, ώστε να ισχύουν οι

ίδιες συνθήκες παραγωγής που εξασφαλίζουν πιο έγκυρα αποτελέσματα για την ανάλυση παλινδρόμησης που ακολουθεί στο επόμενο κεφάλαιο. Χαρακτηριστικό είναι και το διάγραμμα διασποράς της μεταβλητής SiO_2, όπου φαίνεται ότι τα δεδομένα συγκεντρώνονται σε δύο διαφορετικά, ανεξάρτητα πεδία τιμών.



Διάγραμμα 5.5: Διάγραμμα Διασποράς της μεταβλητής SiO_2

Η αλλαγή αυτή έγινε αισθητή και από το ιστόγραμμα της μεταβλητής LOI, όπως φαίνεται στο παρακάτω διάγραμμα:



Διάγραμμα 5.6: Ιστόγραμμα της μεταβλητής LOI

Ø *Συνεπώς, η στατιστική ανάλυση των δεδομένων για το* ***νέο*** *διάστημα ημερομηνιών είναι η ακόλουθη:*

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

Στατιστικά Μέτρα για SiO_2
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 25,1719
Διασπορά = 0,470435
Τυπική Απόκλιση = 0,685882
Τυπικό Σφάλμα = 0,0611033
Ελάχιστη Τιμή = 23,29

Μέγιστη Τιμή = 26,88
Εύρος = 3,59
Προτυποποιημένη Ασυμμετρία = 1,4065
Προτυποποιημένη Κύρτωση = -0,16684

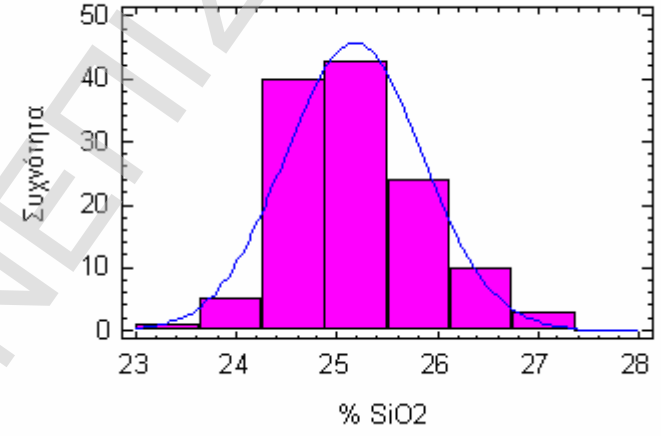
Διαστήματα Εμπιστοσύνης για SiO_2
95,0% διάστημα εμπιστοσύνης για μέσο όρο: 25,1719 +/- 0,120931 [25,051;25,2928]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [0,610372;0,782895]

Το πλήθος των δεδομένων μειώθηκε στις 126 παρατηρήσεις. Ο μέσος όρος είναι 25,17 (από 23,51) με μέγιστη και ελάχιστη τιμή, αντίστοιχα, 26,88 και 23,29. Η τυπική απόκλιση μειώθηκε αρκετά, από 1,54059 σε 0,685882, ενώ το εύρος (Range) σχεδόν υποδιπλασιάστηκε. Επίσης, η ασυμμετρία και η κύρτωση βρίσκονται εντός του διαστήματος [-2,2], δηλαδή φαίνεται ότι υπάρχει κανονικότητα. Οπότε, βλέπουμε ότι ήταν ορθή η επιλογή να μειωθεί ο αριθμός των δεδομένων στο χρονικό διάστημα όπου οι συνθήκες παρατήρησης και καταγραφής παρέμειναν σχετικά σταθερές.

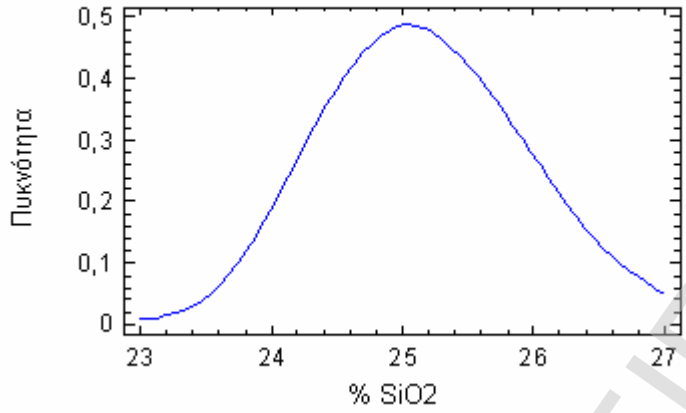
Τα διαστήματα εμπιστοσύνης δηλώνουν ότι, σε επαναλαμβανόμενη δειγματοληψία, τα διαστήματα αυτά θα περιλαμβάνουν την πραγματική τιμή του μέσου όρου ή της τυπικής απόκλισης του πληθυσμού από τον οποίο προέρχονται τα δεδομένα, κατά το 95% των συνολικών φορών. Με άλλα λόγια, ο μέσος της μεταβλητής SiO_2 βρίσκεται μεταξύ των τιμών 25,021 και 25,2928, ενώ η τυπική απόκλιση μεταξύ των τιμών 0,610372 και 0,782895.

Έλεγχος Κανονικότητας με Διαγράμματα

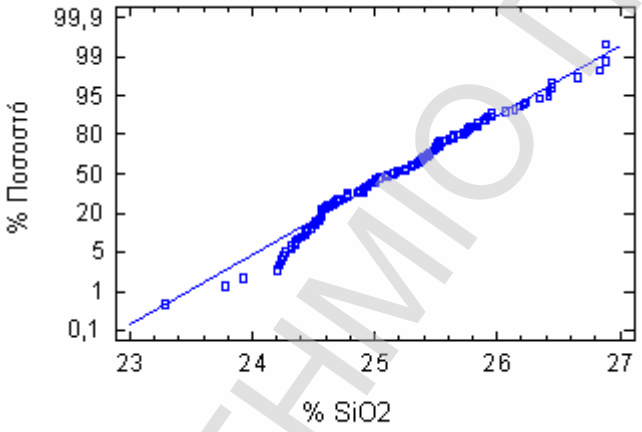
Αναμένουμε τα διαγράμματα που ακολουθούν να μας δείξουν την ύπαρξη κανονικότητας, αφού το πρώτο κριτήριο με την ασυμμετρία και την κύρτωση οδηγεί στο συμπέρασμα αυτό.



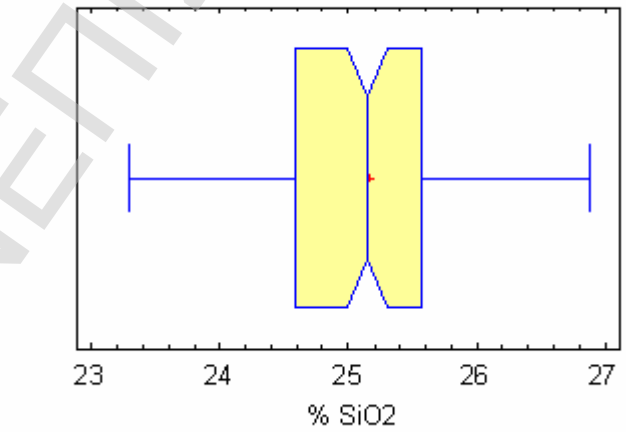
Διάγραμμα 5.7: Ιστόγραμμα της νέας μεταβλητής SiO_2



Διάγραμμα 5.8: Διάγραμμα Ίχνους της Πυκνότητας της νέας μεταβλητής SiO_2



Διάγραμμα 5.9: Διάγραμμα Ελέγχου Κανονικότητας της νέας μεταβλητής SiO_2



Διάγραμμα 5.10: Θηκόγραμμα της νέας μεταβλητής SiO_2

Από τα παραπάνω διαγράμματα βλέπουμε ότι η "νέα" μεταβλητή SiO_2 προσεγγίζει πολύ καλά την κανονική κατανομή.

5.2.2. Μεταβλητή Al_2O_3

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

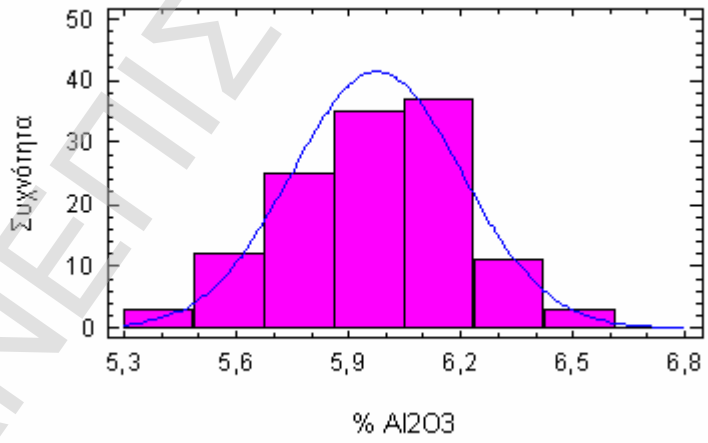
Το πλήθος του δείγματος είναι 126. Παρακάτω δίνονται οι τιμές για το μέσο όρο, τη διασπορά και την τυπική απόκλιση, το τυπικό σφάλμα, τη μέγιστη και ελάχιστη τιμή, την ασυμμετρία και κύρτωση και τα διαστήματα εμπιστοσύνης.

Στατιστικά Μέτρα για Al_2O_3
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 5,97571
Διασπορά = 0,0515847
Τυπική Απόκλιση = 0,227123
Τυπικό Σφάλμα = 0,0202337
Ελάχιστη Τιμή = 5,37
Μέγιστη Τιμή = 6,45
Εύρος = 1,08
Προτυποποιημένη Ασυμμετρία = -1,57524
Προτυποποιημένη Κύρτωση = -0,428845

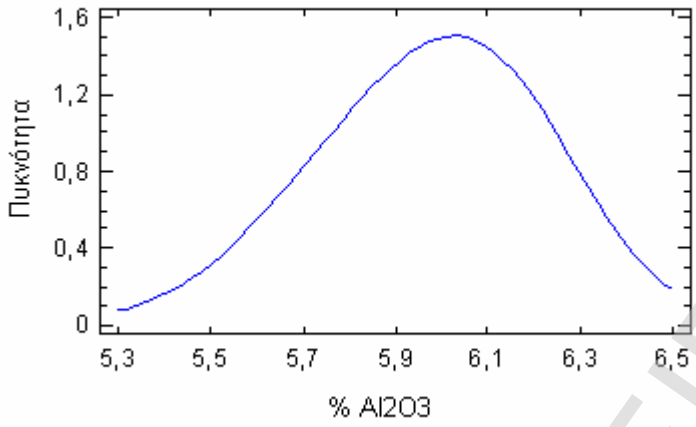
Διαστήματα Εμπιστοσύνης για Al_2O_3
95,0% διάστημα εμπιστοσύνης για μέσο όρο: 5,97571 +/- 0,0400451 [5,93567;6,01576]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [0,202118;0,259247]

Έλεγχος Κανονικότητας με Διαγράμματα

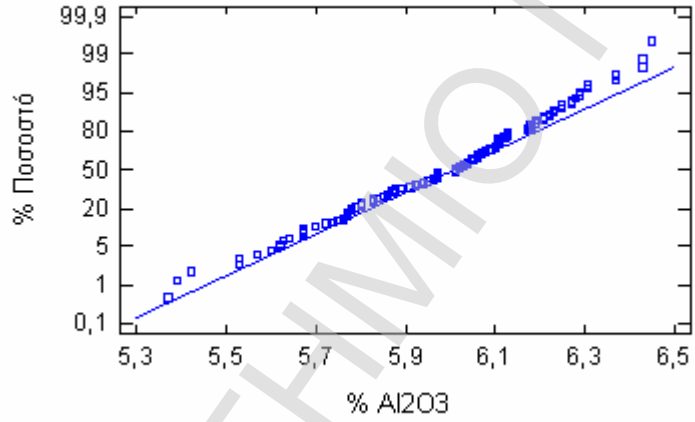
Η ασυμμετρία και κύρτωση από τον παραπάνω πίνακα φαίνεται ότι είναι εντός του διαστήματος [-2, 2], άρα η μεταβλητή Al_2O_3 ακολουθεί κανονική κατανομή. Το ίδιο προκύπτει και από τα παρακάτω διαγράμματα, αν και εμφανίζεται ελαφριά κύρτωση προς τα αριστερά.



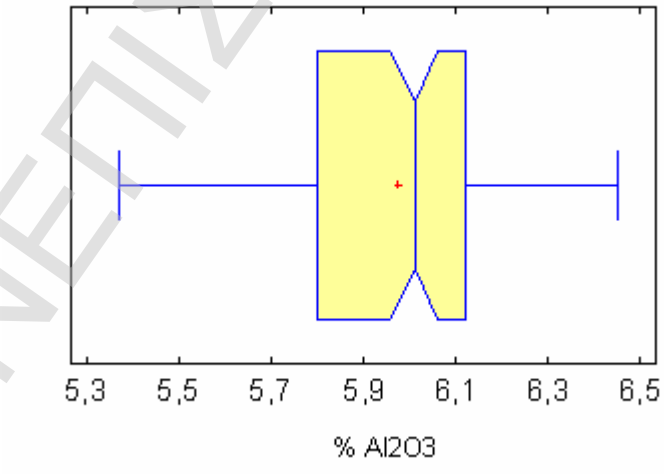
Διάγραμμα 5.11: Ιστόγραμμα της μεταβλητής Al_2O_3



<u>Διάγραμμα 5.12</u>: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής Al_2O_3



<u>Διάγραμμα 5.13</u>: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής Al_2O_3



<u>Διάγραμμα 5.14</u>: Θηκόγραμμα της μεταβλητής Al_2O_3

5.2.3. Μεταβλητή Blaine

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

Οι βασικές στατιστικές μετρήσεις για τη μεταβλητή αυτή είναι οι ακόλουθες:

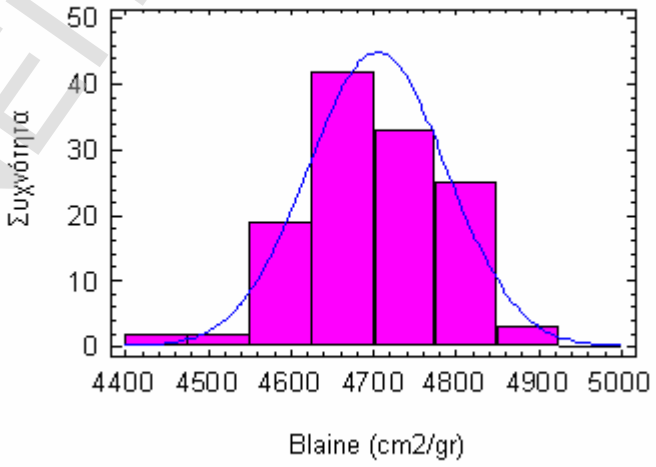
Στατιστικά Μέτρα για Blaine
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 4704,37
Διασπορά = 7061,59
Τυπική Απόκλιση = 84,0333
Τυπικό Σφάλμα = 7,48628
Ελάχιστη Τιμή = 4450,0
Μέγιστη Τιμή = 4900,0
Εύρος = 450,0
Προτυποποιημένη Ασυμμετρία = -1,20873
Προτυποποιημένη Κύρτωση = 0,466206

Διαστήματα Εμπιστοσύνης για Blaine
95,0% διάστημα εμπιστοσύνης για μέσο όρο: 4704,37 +/- 14,8163 [4689,55;4719,18]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [74,7818;95,9191]

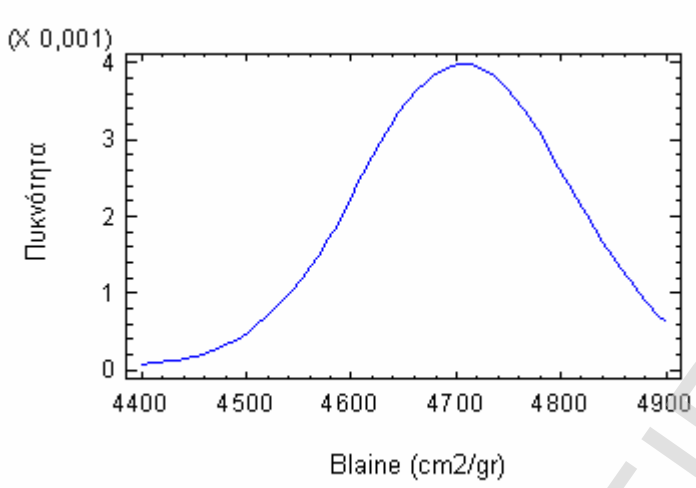
Αυτό που μας ενδιαφέρει από τις παραπάνω μετρήσεις είναι κυρίως η ασυμμετρία και η κύρτωση. Οι τιμές τους είναι αντίστοιχα -1,20873 και 0,466206, δηλαδή εντός του διαστήματος [-2, 2]. Συνεπώς, η μεταβλητή Blaine είναι κανονικά κατανεμημένη.

Έλεγχος Κανονικότητας με Διαγράμματα

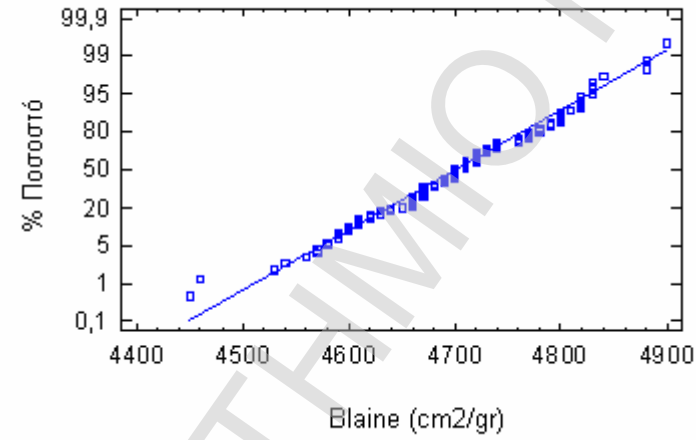
Τα τέσσερα παρακάτω διαγράμματα επιβεβαιώνουν την υπόθεση ύπαρξης κανονικότητας. Τονίζεται ότι το θηκόγραμμα είναι επίσης ενδεικτικό των μέτρων κεντρικής τάσης και θέσης (μέσος όρος, διάμεσος, διασποράς), καθώς επίσης και της ύπαρξης έκτροπων παρατηρήσεων. Στη συγκεκριμένη περίπτωση, βλέπουμε ότι υπάρχουν κάποιες μακρινές παρατηρήσεις που απέχουν πάνω από 1,5 φορά το ενδοτεταρτημοριακό εύρος, δεν θεωρούνται όμως έκτροπες παρατηρήσεις και γι' αυτό δεν απορρίπτονται.



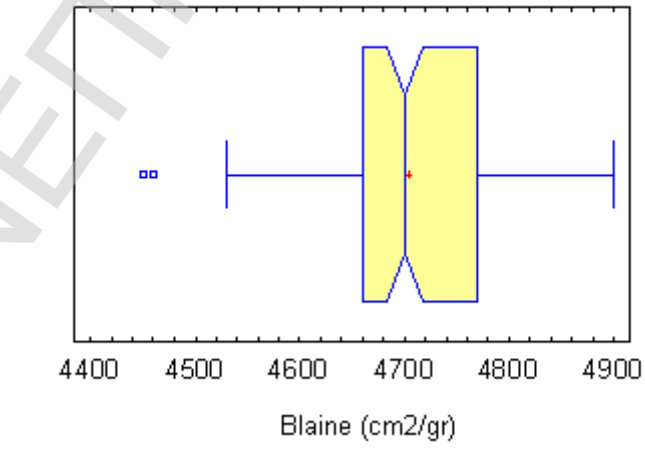
Διάγραμμα 5.15: Ιστόγραμμα της μεταβλητής Blaine



Διάγραμμα 5.16: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής Blaine



Διάγραμμα 5.17: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής Blaine



Διάγραμμα 5.18: Θηκόγραμμα της μεταβλητής Blaine

5.2.4. Μεταβλητή IR

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

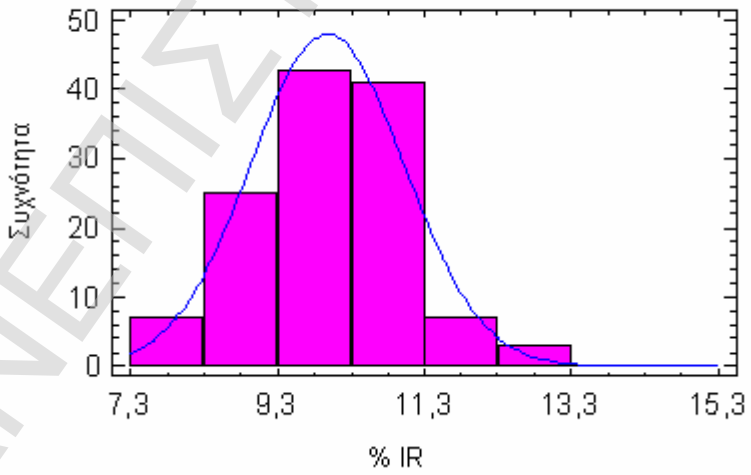
Στατιστικά Μέτρα για IR
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 9,97992
Διασπορά = 1,09463
Τυπική Απόκλιση = 1,04625
Τυπικό Σφάλμα = 0,093207
Ελάχιστη Τιμή = 7,65
Μέγιστη Τιμή = 13,3
Εύρος = 5,65
Προτυποποιημένη Ασυμμετρία = 1,11461
Προτυποποιημένη Κύρτωση = 0,716744

Διαστήματα Εμπιστοσύνης για IR
95,0% διάστημα εμπιστοσύνης για μέσο όρο: 9,97992 +/- 0,184469 [9,79545;10,1644]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [0,931062;1,19423]

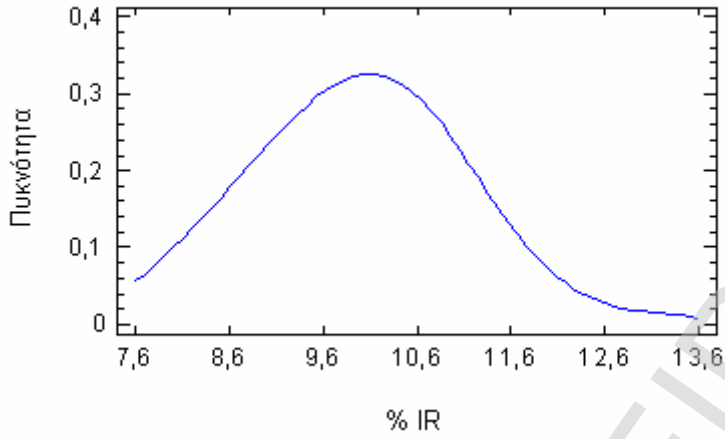
Και σε αυτή την περίπτωση, η μεταβλητή είναι κανονικά κατανεμημένη, με μέσο όρο περίπου 9,98 και τυπική απόκλιση 1,04625.

Έλεγχος Κανονικότητας με Διαγράμματα

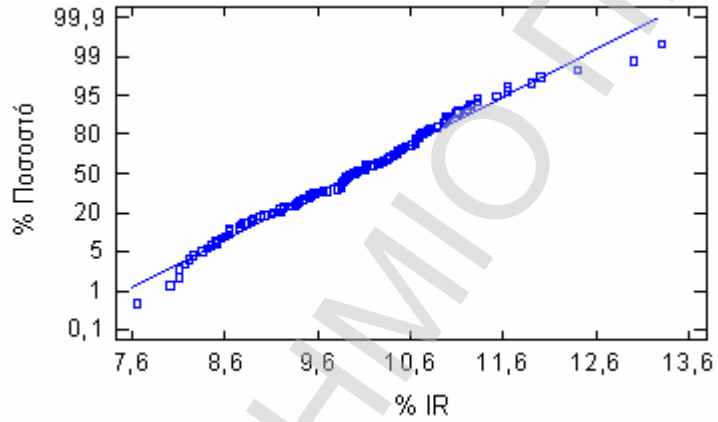
Τα διαγράμματα επιβεβαιώνουν την ύπαρξη κανονικότητας. Στο θηκόγραμμα διακρίνεται μια ελαφριά ασυμμετρία προς τα δεξιά, καθώς δύο τιμές ξεφεύγουν πέρα από 3 φορές την απόσταση του ενδοτεταρτημοριακού εύρους.



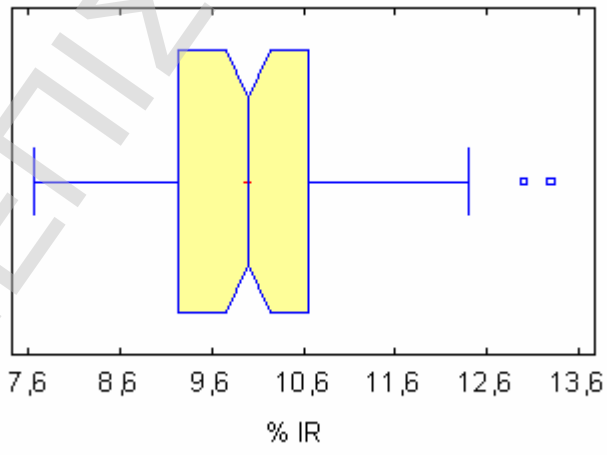
Διάγραμμα 5.19: Ιστόγραμμα της μεταβλητής IR



<u>Διάγραμμα 5.20</u>: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής IR



<u>Διάγραμμα 5.21</u>: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής IR



<u>Διάγραμμα 5.22</u>: Θηκόγραμμα της μεταβλητής IR

5.2.5. Μεταβλητή LOI

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

Στατιστικά Μέτρα για LOI
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 3,20976
Διασπορά = 0,0725671
Τυπική Απόκλιση = 0,269383
Τυπικό Σφάλμα = 0,0239985
Ελάχιστη Τιμή = 2,5
Μέγιστη Τιμή = 4,11
Εύρος = 1,61
Προτυποποιημένη Ασυμμετρία = 2,04221
Προτυποποιημένη Κύρτωση = 2,54233

Διαστήματα Εμπιστοσύνης για LOI
95,0% διάγραμμα εμπιστοσύνης για μέσο όρο: 3,20976 +/- 0,0474962 [3,16227;3,25726]
95,0% διάγραμμα εμπιστοσύνης για τυπική απόκλιση: [0,239726;0,307485]

Στην περίπτωση της μεταβλητής αυτής, η ασυμμετρία και κύρτωση **δεν** είναι εντός του διαστήματος [-2, 2], συνεπώς η LOI δεν ακολουθεί κανονική κατανομή. Αυτό έχει ως αποτέλεσμα να μην είναι έγκυρα τα διαστήματα εμπιστοσύνης που παρουσιάζονται, επειδή τα *F-* και *t-tests*, βάσει των οποίων υπολογίζονται, προϋποθέτουν την ύπαρξη κανονικότητας. Ωστόσο, επειδή ο αριθμός των παρατηρήσεων είναι κατά πολύ μεγαλύτερος από το νούμερο 30, η στατιστική θεωρεί σχετικά έγκυρα τα τεστ αυτά.

Ένας τρόπος που είναι γενικά αποδεκτός στη στατιστική και προτείνεται και από τα στατιστικά προγράμματα σχετικά με την ύπαρξη κανονικότητας, είναι ο *μετασχηματισμός των μεταβλητών* με τη χρήση κάποιων πολύ διαδεδομένων συναρτήσεων, όπως είναι η λογαριθμική (*logX*), η αντίστροφη (*1/X*) και η τετραγωνική ρίζα (*SQRT(X)*). Στη συγκεκριμένη περίπτωση της μεταβλητής LOI ο μετασχηματισμός σε *log(LOI)* δίνει τα ακόλουθα αποτελέσματα:

Στατιστικά Μέτρα για LOG(LOI)
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 1,16274
Διασπορά = 0,00694217
Τυπική Απόκλιση = 0,0833197
Τυπικό Σφάλμα = 0,00742271
Ελάχιστη Τιμή = 0,916291
Μέγιστη Τιμή = 1,41342
Εύρος = 0,497132
Προτυποποιημένη Ασυμμετρία = 0,454092
Προτυποποιημένη Κύρτωση = 1,86048

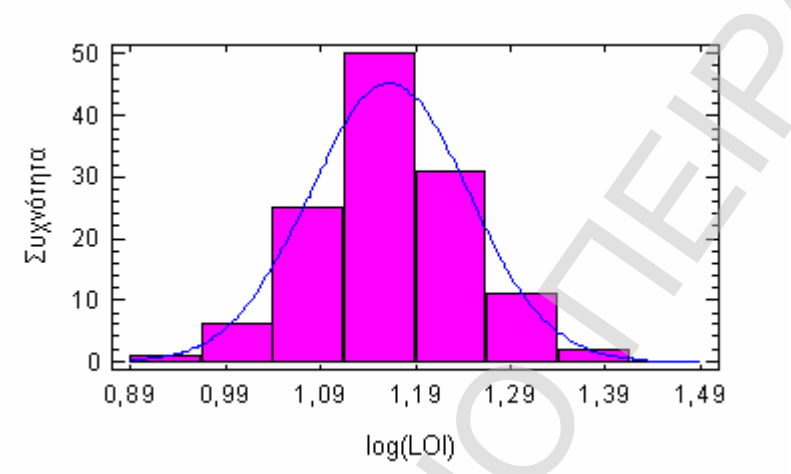
Διαστήματα Εμπιστοσύνης για LOG(LOI)
95,0% διάγραμμα εμπιστοσύνης για μέσο όρο: 1,16274 +/- 0,0146905 [1,14805;1,17743]
95,0% διάγραμμα εμπιστοσύνης για τυπική απόκλιση: [0,0741468;0,0951046]

Παρατηρούμε ότι η μεταβλητή *log(LOI)* είναι κανονικά κατανεμημένη. Ο μετασχηματισμός αυτός θα μπορούσε να χρησιμοποιηθεί στην ανάλυση παλινδρόμησης που ακολουθεί στο επόμενο κεφάλαιο, αν και δεν είναι απαραίτητο.

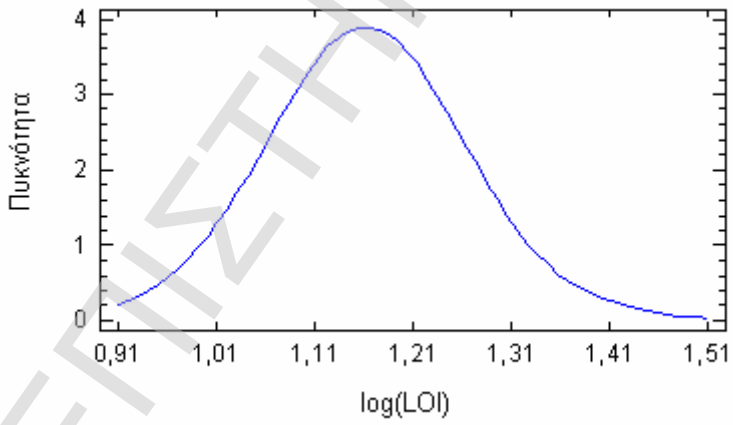
Συνήθως, αυτό που απαιτείται στην παλινδρόμηση είναι η ύπαρξη κανονικότητας στην εξαρτημένη μεταβλητή *Y*, συνεπώς και στα σφάλματα *ε*.

Έλεγχος Κανονικότητας με Διαγράμματα

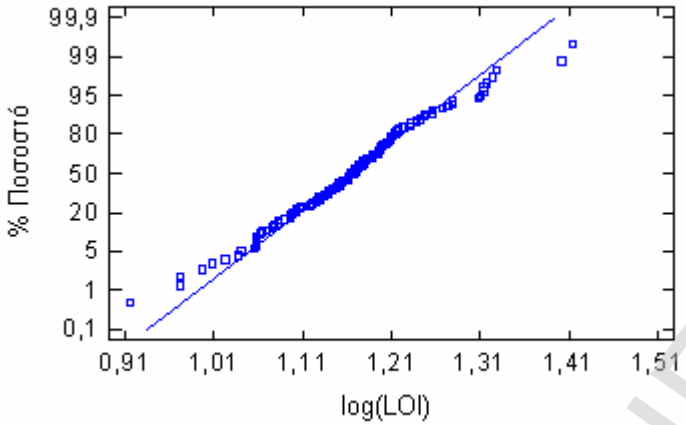
Στην περίπτωση αυτή, θα εξετάσουμε τα διαγράμματα για τη μεταβλητή *log(LOI)* και όχι τη LOI. Η ύπαρξη της κανονικότητας επαληθεύεται.



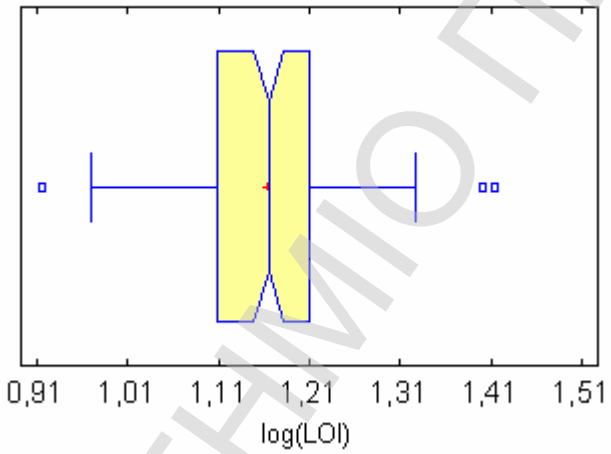
Διάγραμμα 5.23: Ιστόγραμμα της μεταβλητής log(LOI)



Διάγραμμα 5.24: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής log(LOI)



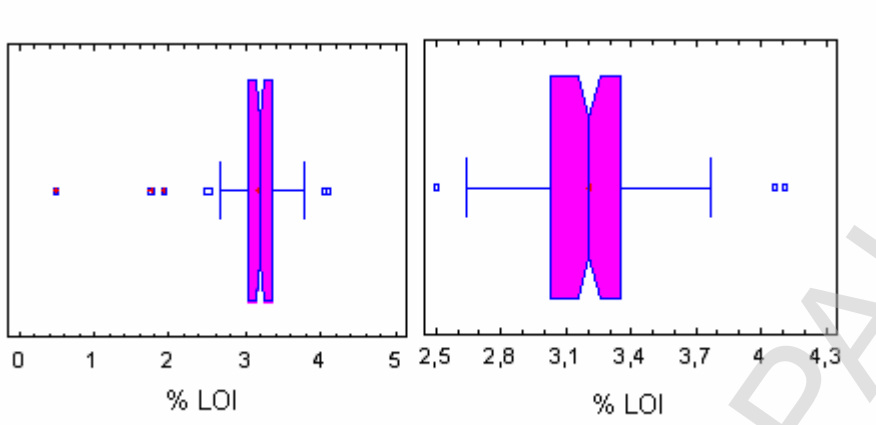
Διάγραμμα 5.25: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής log(LOI)



Διάγραμμα 5.26: Θηκόγραμμα της μεταβλητής log(LOI)

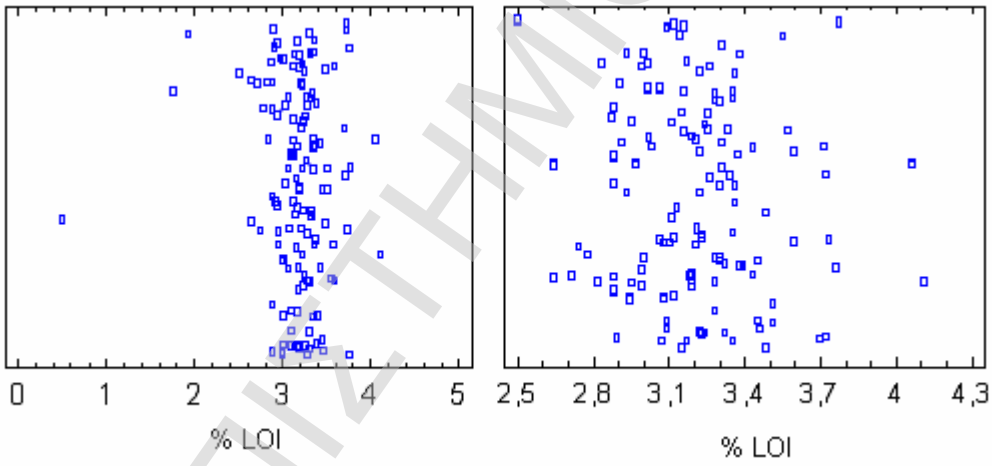
Έλεγχος για Έκτροπες Παρατηρήσεις

Πρέπει να σημειωθεί ότι οι παραπάνω παρατηρήσεις έχουν κατά κάποιο τρόπο "καθαριστεί" από έκτροπες παρατηρήσεις. Οι έκτροπες παρατηρήσεις μπορούν να εντοπιστούν τόσο από την απλή παρατήρηση των δεδομένων, όσο και από το θηκόγραμμα. Στο παρακάτω διάγραμμα φαίνεται το θηκόγραμμα πριν και μετά την απομάκρυνση εκτρόπων παρατηρήσεων, πάντα με την έγκριση των υπεύθυνων της χημικής βιομηχανίας. Οι έκτροπες παρατηρήσεις είναι αυτές που απέχουν πάνω από 3 φορές το ενδοτεταρτημοριακό εύρος και συμβολίζονται με τα μικρά τετραγωνάκια με τον κόκκινο σταυρό στη μέση.



<u>Διάγραμμα 5.27</u>: Θηκογράμματα της μεταβλητής LOI Με και Χωρίς Έκτροπες Παρατηρήσεις

Όπως φαίνεται στο παραπάνω διάγραμμα, αρχικά τρεις τιμές της μεταβλητής LOI είναι ασυνήθιστα μικρές σε σχέση με τις άλλες. Αυτό πιθανότατα οφείλεται σε λάθος μέτρηση, ή σε ύπαρξη ειδικών συνθηκών του περιβάλλοντος της παραγωγικής διαδικασίας, οι οποίες δεν αντιπροσωπεύουν τα επιθυμητά επίπεδα και προδιαγραφές που θέτει η εταιρεία. Για το λόγο αυτό απομακρύνονται. Τα διαγράμματα διασποράς με και χωρίς τα έκτροπα είναι επίσης ενδεικτικά της επιρροής που αυτά ασκούν.



<u>Διάγραμμα 5.28</u>: Διαγράμματα Διασποράς της μεταβλητής LOI Με και Χωρίς Έκτροπες Παρατηρήσεις

5.2.6. <u>Μεταβλητή Clk</u>

<u>Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος</u>

Οι βασικές στατιστικές παράμετροι της μεταβλητής Clk φαίνονται συγκεντρωτικά στον παρακάτω πίνακα.

<u>Στατιστικά Μέτρα για Clk</u>
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 80,7595
Διασπορά = 0,462589
Τυπική Απόκλιση = 0,680139

Τυπικό Σφάλμα = 0,0605916
Ελάχιστη Τιμή = 78,7
Μέγιστη Τιμή = 82,9
Εύρος = 4,2
Προτυποποιημένη Ασυμμετρία = 0,55208
Προτυποποιημένη Κύρτωση = 3,90546

Διαστήματα Εμπιστοσύνης για Clk
95,0% διάστημα εμπιστοσύνης για μέσο όρο: 80,7595 +/- 0,119918 [80,6396;80,8794]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [0,60526;0,776338]

Όπως φαίνεται από τις τιμές ασυμμετρίας και κύρτωσης, η μεταβλητή Clk δεν ακολουθεί κανονική κατανομή. Στην περίπτωση αυτή, γίνεται έλεγχος για το αν οι μεταβλητές *log(Clk)*, *1/Clk* ή *SQRT(Clk)* ακολουθούν κανονική κατανομή με τις τιμές ασυμμετρίας και κύρτωσης. Παρόλα αυτά, επειδή οι τιμές των παραμέτρων αυτών δεν είναι μεταξύ του διαστήματος [-2,2], συμπεραίνουμε ότι οι παραπάνω μεταβλητές δεν είναι κανονικά κατανεμημένες.

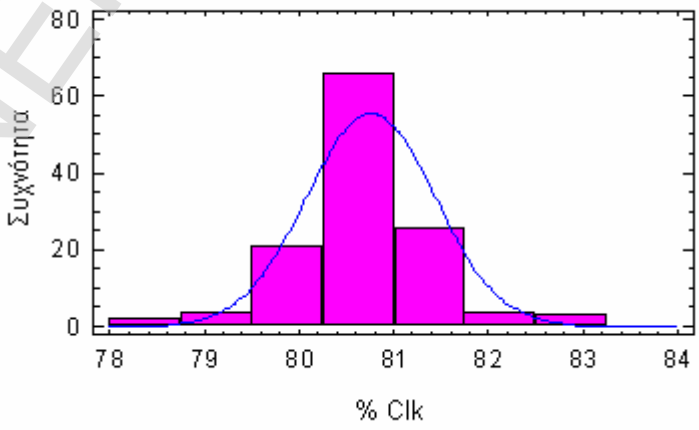
- **Στατιστικά Μέτρα για LOG(Clk)**
  Προτυποποιημένη Ασυμμετρία = 0,343613
  Προτυποποιημένη Κύρτωση = 3,87512

- **Στατιστικά Μέτρα για 1/Clk**
  Προτυποποιημένη Ασυμμετρία = -0,135299
  Προτυποποιημένη Κύρτωση = 3,85864

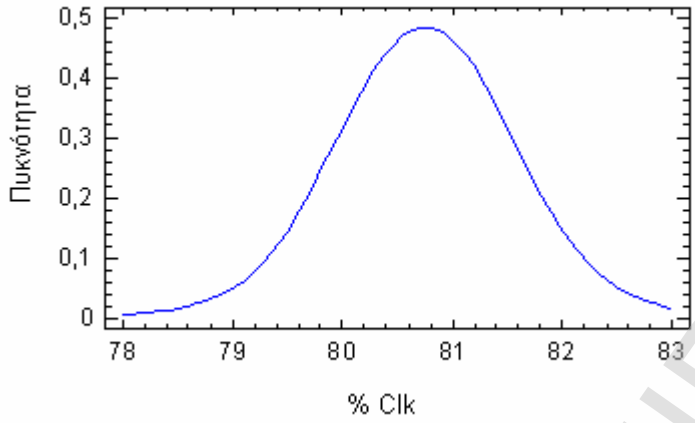
- **Στατιστικά Μέτρα για SQRT(Clk)**
  Προτυποποιημένη Ασυμμετρία = 0,447825
  Προτυποποιημένη Κύρτωση = 3,88855

Έλεγχος Κανονικότητας με Διαγράμματα

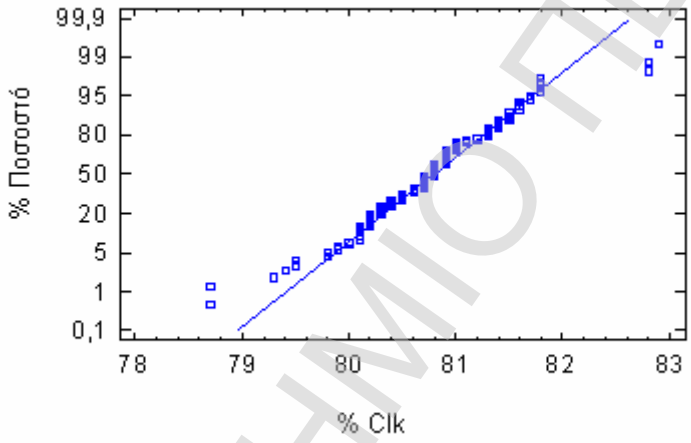
Τα παρακάτω διαγράμματα δείχνουν ότι υπάρχουν αποκλίσεις από την κανονικότητα. Το σχήμα καμπάνας είναι μεν εμφανές, όμως φαίνεται ότι υπάρχει πρόβλημα στην κύρτωση (λεπτόκυρτη κατανομή). Επίσης, στο διάγραμμα ελέγχου της κανονικότητας τα δεδομένα δεν ευθυγραμμίζονται τελείως πάνω στην ευθεία γραμμή, αλλά ξεφεύγουν αρκετά στα άκρα και, εκτός αυτού, έχουν πολύ συγκεντρωμένες τιμές σε κάποια σημεία.



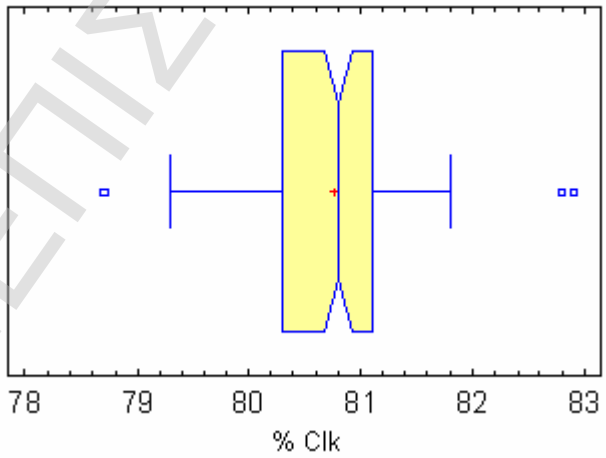
Διάγραμμα 5.29: Ιστόγραμμα της μεταβλητής Clk



Διάγραμμα 5.30: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής Clk



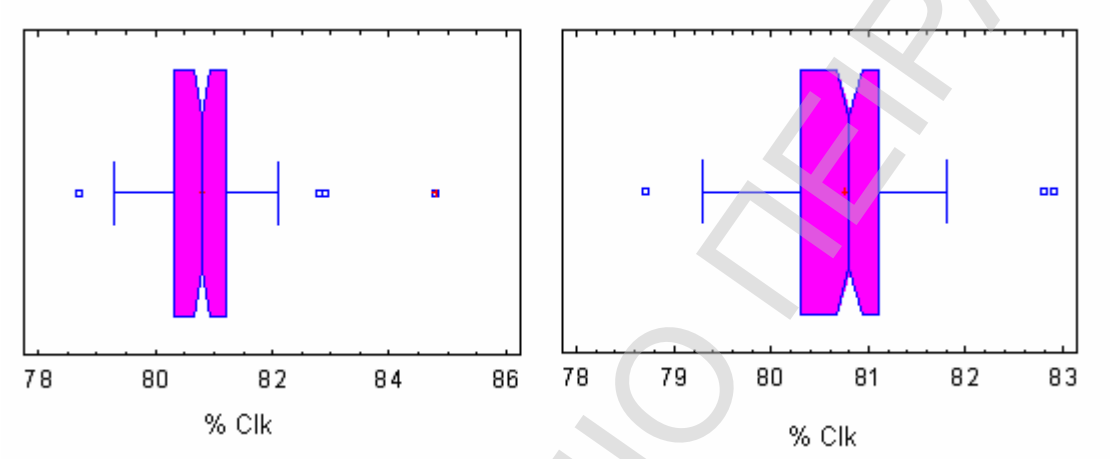
Διάγραμμα 5.31: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής Clk



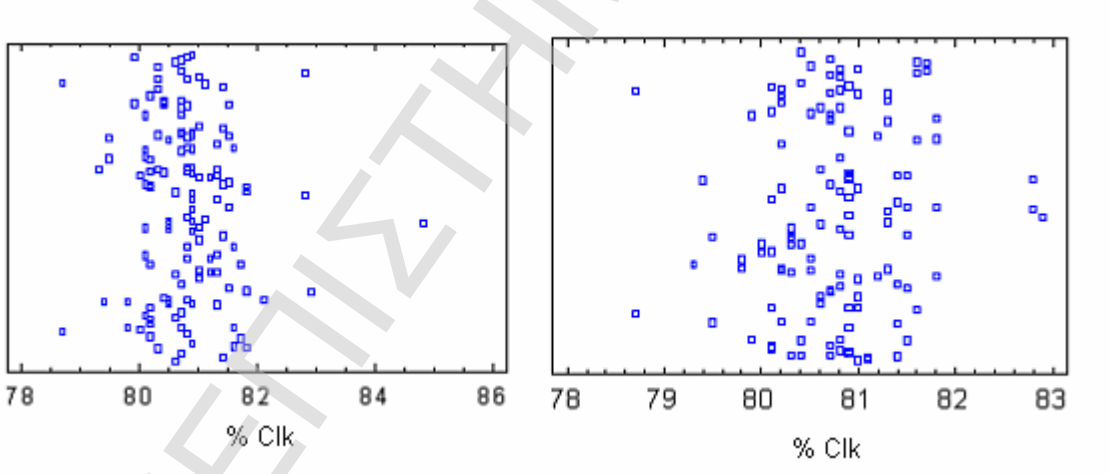
Διάγραμμα 5.32: Θηκόγραμμα της μεταβλητής Clk

Έλεγχος για Έκτροπες Παρατηρήσεις

Και στην περίπτωση της μεταβλητής αυτής, όπως και στην περίπτωση της μεταβλητής LOI, παρατηρήθηκαν κάποιες έκτροπες παρατηρήσεις, οι οποίες και απομακρύνθηκαν. Στα παρακάτω διαγράμματα φαίνεται το θηκόγραμμα και το διάγραμμα διασποράς, πριν και μετά την απομάκρυνση εκτρόπων παρατηρήσεων. Συγκεκριμένα, παρατηρήθηκε μία έκτροπη παρατήρηση, η οποία αποτελεί και αρκετά μακρινό σημείο, καθώς απέχει πάνω από 3 φορές την απόσταση του ενδοτεταρτημοριακού εύρους.



Διάγραμμα 5.33: Θηκογράμματα της μεταβλητής Clk Με και Χωρίς την Έκτροπη Παρατήρηση



Διάγραμμα 5.34: Διαγράμματα Διασποράς της μεταβλητής Clk Με και Χωρίς την Έκτροπη Παρατήρηση

5.2.7. Μεταβλητή Gyp

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

Στατιστικά Μέτρα για Gyp
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 5,44365
Διασπορά = 0,0957594
Τυπική Απόκλιση = 0,30945

Τυπικό Σφάλμα = 0,027568
Ελάχιστη Τιμή = 4,1
Μέγιστη Τιμή = 6,0
Εύρος = 1,9
Προτυποποιημένη Ασυμμετρία = -5,0297
Προτυποποιημένη Κύρτωση = 6,30796

Διαστήματα Εμπιστοσύνης για Gyp
95,0% διάστημα εμπιστοσύνης για μέσο όρο: 5,44365 +/- 0,0545606 [5,38909;5,49821]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [0,275382;0,353219]

Η μεταβλητή αυτή απέχει πολύ από την κανονικότητα. Οι μετασχηματισμοί της μεταβλητής δεν αρκούν για να διορθώσουν την κατάσταση.

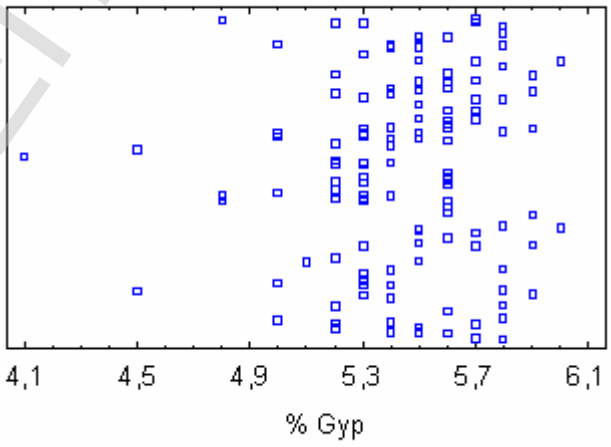
- **Στατιστικά Μέτρα για LOG(Gyp)**
  Προτυποποιημένη Ασυμμετρία = -6,57335
  Προτυποποιημένη Κύρτωση = 9,79476

- **Στατιστικά Μέτρα για 1/Gyp**
  Προτυποποιημένη Ασυμμετρία = 8,35585
  Προτυποποιημένη Κύρτωση = 14,5992

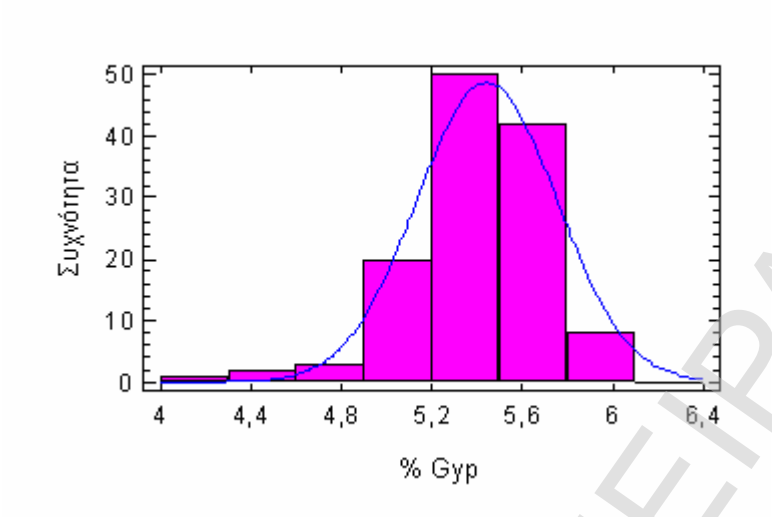
- **Στατιστικά Μέτρα για SQRT(Gyp)**
  Προτυποποιημένη Ασυμμετρία = -5,77342
  Προτυποποιημένη Κύρτωση = 7,90684

Έλεγχος Κανονικότητας με Διαγράμματα

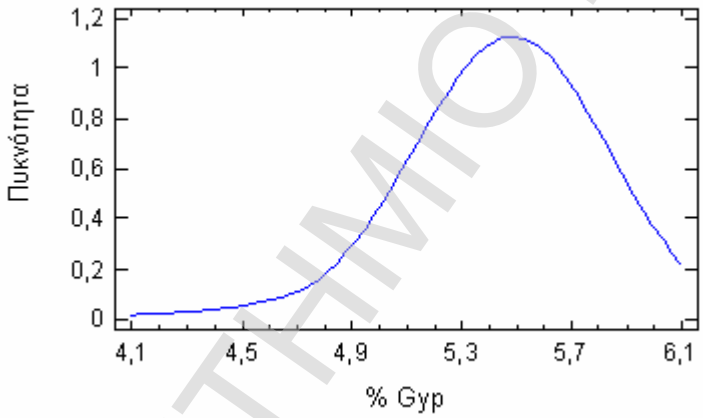
Τα παρακάτω διαγράμματα δείχνουν καθαρά τη μεγάλη απόκλιση από την κανονικότητα. Γενικά, παρατηρείται ότι οι τιμές των δεδομένων της μεταβλητής Gyp είναι πολύ συγκεντρωμένες γύρω από ένα σταθερό σημείο. Στη συγκεκριμένη περίπτωση αυτό συμβαίνει γιατί η Gyp είναι μια ελεγχόμενη μεταβλητή. Όντως, από τη χημεία του τσιμέντου γνωρίζουμε ότι η γύψος προστίθεται στο τελικό μίγμα της δημιουργίας τσιμέντου μαζί με το κλίνκερ, αφού είναι ήδη γνωστή η συγκέντρωση κλίνκερ, η οποία επηρεάζει και την απόφασή μας για την επιθυμητή συγκέντρωση γύψου. Αυτό φαίνεται κυρίως από το διάγραμμα διασποράς, αλλά και από το διάγραμμα ελέγχου της κανονικότητας που ακολουθούν.



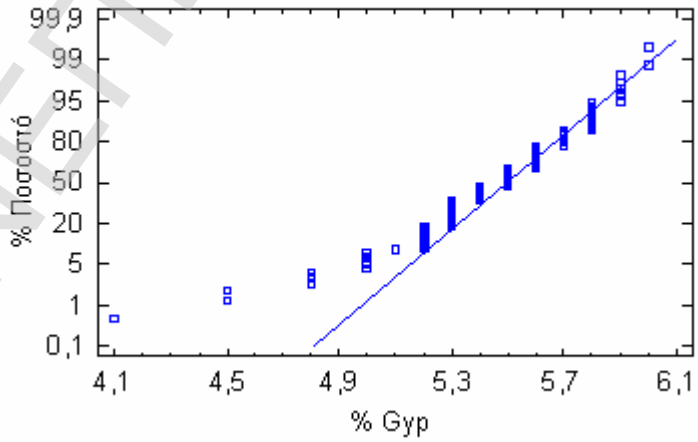
Διάγραμμα 5.35: Διάγραμμα Διασποράς της μεταβλητής Gyp



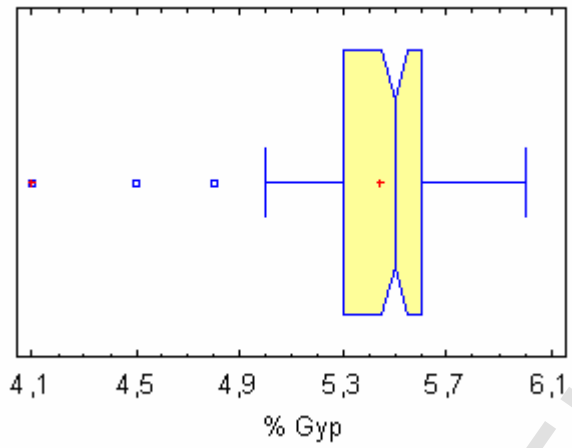
Διάγραμμα 5.36: Ιστόγραμμα της μεταβλητής Gyp



Διάγραμμα 5.37: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής Gyp



Διάγραμμα 5.38: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής Gyp

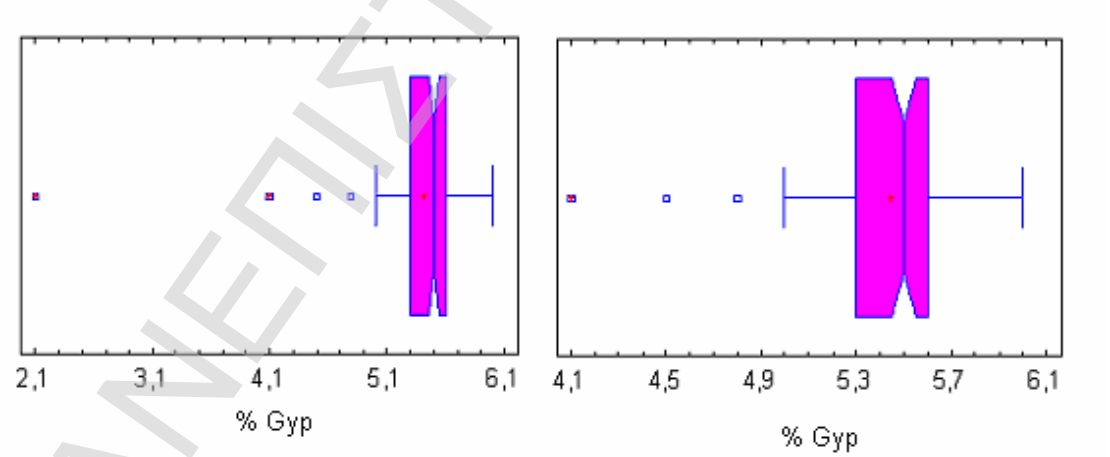


Διάγραμμα 5.39: Θηκόγραμμα της μεταβλητής Gyp

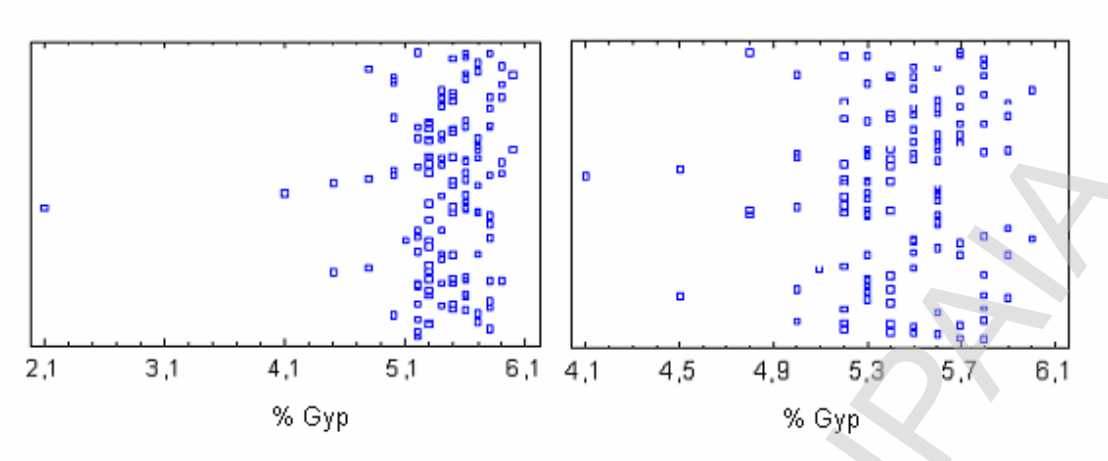
Έλεγχος για Έκτροπες Παρατηρήσεις

Από τις αρχικές παρατηρήσεις έχει αφαιρεθεί μία πολύ μακρινή παρατήρηση (τιμή Gyp περίπου 2,1), η οποία κατά πάσα πιθανότητα οφείλεται σε κάποιο λάθος μέτρησης. Υπάρχει και μια άλλη μακρινή παρατήρηση (τιμή Gyp περίπου 4,1) η οποία δεν απομακρύνεται, γιατί μετά από έλεγχο βρέθηκε ότι στη Gyp μετά από κάθε απομάκρυνση μιας παρατήρησης πάντα προκύπτει και μία άλλη μακρινή παρατήρηση (τετραγωνάκι με σταυρό στη μέση). Για το λόγο αυτό απομακρύνθηκε μόνο η παρατήρηση που είναι κατά πολύ διαφορετική από *όλες* τις υπόλοιπες.

Ακολουθούν τα θηκογράμματα και τα διαγράμματα διασποράς για τις περιπτώσεις όπου αρχικά υπάρχει η έκτροπη παρατήρηση και στη συνέχεια απομακρύνεται.



Διάγραμμα 5.40: Θηκογράμματα της μεταβλητής Gyp Με και Χωρίς την Έκτροπη Παρατήρηση



Διάγραμμα 5.41: Διαγράμματα Διασποράς της μεταβλητής Gyp Με και Χωρίς την Έκτροπη Παρατήρηση

5.2.8. Μεταβλητή Est2

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

Στατιστικά Μέτρα για Est2
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 23,7341
Διασπορά = 3,37059
Τυπική Απόκλιση = 1,83592
Τυπικό Σφάλμα = 0,163556
Ελάχιστη Τιμή = 19,7
Μέγιστη Τιμή = 28,4
Εύρος = 8,7
Προτυποποιημένη Ασυμμετρία = 2,53494
Προτυποποιημένη Κύρτωση = 0,320835

Διαστήματα Εμπιστοσύνης για Est2
95,0% διάστημα εμπιστοσύνης για: 23,7341 +/- 0,323699 [23,4104;24,0578]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [1,63379;2,09559]

Από τις τιμές της ασυμμετρίας παρατηρούμε ότι δεν ανήκουν στο διάστημα [-2,2], άρα υπάρχει απόκλιση από την κανονική κατανομή, που τείνει να ακυρώσει κάθε έλεγχο που σχετίζεται με την τυπική απόκλιση. Επειδή, όμως, η μεταβλητή Est2 είναι **εξαρτημένη** μεταβλητή, το οποίο σημαίνει ότι στην ανάλυση παλινδρόμησης η κανονικότητα αυτής είναι αναγκαία προϋπόθεση, γίνεται μια προσπάθεια μετασχηματισμού της. Αρχικά μετασχηματίζεται σε λογαριθμική συνάρτηση, η οποία είναι και η πιο "ήπια" μορφή μετασχηματισμού, σε σχέση με την τετραγωνική ρίζα και την αντίστροφη συνάρτηση. Τα αποτελέσματα για τη μεταβλητή ***log(Est2)*** είναι τα ακόλουθα:

Στατιστικά Μέτρα για LOG(Est2)
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 3,164
Διασπορά = 0,00581764
Τυπική Απόκλιση = 0,0762735

Τυπικό Σφάλμα = 0,00679498
Ελάχιστη Τιμή = 2,98062
Μέγιστη Τιμή = 3,34639
Εύρος = 0,365771
Προτυποποιημένη Ασυμμετρία = 1,57442
Προτυποποιημένη Κύρτωση = 0,0214929

Η ασυμμετρία τώρα βρίσκεται εντός του επιθυμητού διαστήματος, συνεπώς φαίνεται ότι ισχύει η προϋπόθεση της ύπαρξης κανονικότητας. Ωστόσο, επειδή η μεταβλητή αυτή μας ενδιαφέρει πολύ σχετικά με το εάν είναι κανονικά κατανεμημένη, ελέγχουμε για την ύπαρξη κανονικότητας και με μερικά στατιστικά tests, όπως με την κατανομή X^2 (Chi-Square) για την καλή προσαρμογή των δεδομένων και το *Shapiro-Wilks statistic*.

Έλεγχος Κανονικότητας με Διάφορα Tests

Τα τεστ που γίνονται για τον έλεγχο της ύπαρξης κανονικότητας της μεταβλητής log(Est2) είναι το τεστ για καλή προσαρμογή των δεδομένων στην κατανομή X^2 (*Chi-Square goodness-of-fit statistic*), το *Shapiro-Wilks statistic* και τα *Z scores* για ασυμμετρία και κύρτωση, τα οποία στην ουσία συμπίπτουν με το κριτήριο που χρησιμοποιείται γενικά στην εργασία αυτή. Η μηδενική υπόθεση ισχυρίζεται ότι η μεταβλητή ακολουθεί την κανονική κατανομή, συνεπώς επιθυμητό είναι το *p-value* που προκύπτει να είναι μεγαλύτερο από 10%, έτσι ώστε να μην μπορούμε να απορρίψουμε τη μηδενική υπόθεση σε 90% επίπεδο εμπιστοσύνης. Τα αποτελέσματα των tests είναι τα εξής:

Τεστ Κανονικότητας για LOG(Est2)

X^2 τεστ για έλεγχο καλής προσαρμογής = 43,2063
P-Value = 0,00654555

Shapiro-Wilks W τεστ = 0,966834
P-Value = 0,04237

Z τεστ για ασυμμετρία = 1,1313
P-Value = 0,257928

Z τεσυ για κύρτωση = 0,194762
P-Value = 0,845574

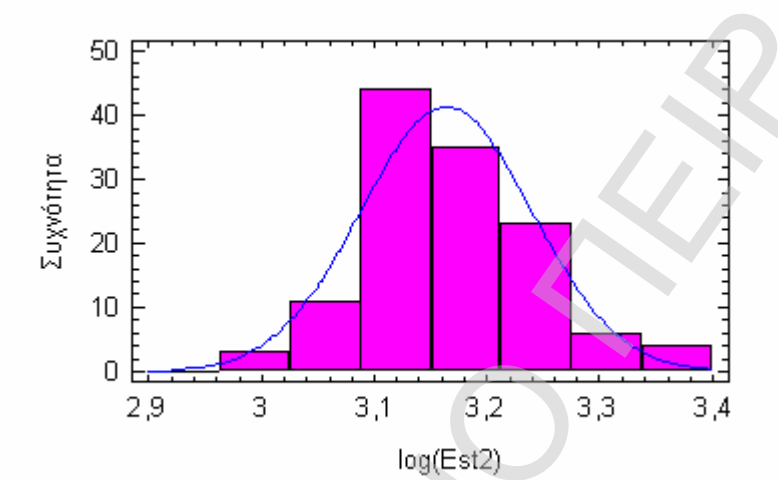
Τα *p-values* των δύο τελευταίων τεστ για την ασυμμετρία και κύρτωση είναι αρκετά μεγαλύτερα από 10%, συνεπώς δείχνουν ότι δεν μπορούμε να απορρίψουμε τη μηδενική υπόθεση ότι η *log(Est2)* κατανέμεται κανονικά. Ωστόσο, στα δύο άλλα τεστ (X^2 και Shapiro-Wilks) το *p-value* είναι μικρότερο από 5%, άρα κατά 95% μπορούμε να απορρίψουμε τη μηδενική υπόθεση. Συνεπώς, διατηρούμε μια επιφύλαξη για την ύπαρξη κανονικότητας στη συγκεκριμένη μεταβλητή, και ελέγχουμε παρακάτω οπτικά την κατανομή αυτής.

Έλεγχος Κανονικότητας με Διαγράμματα

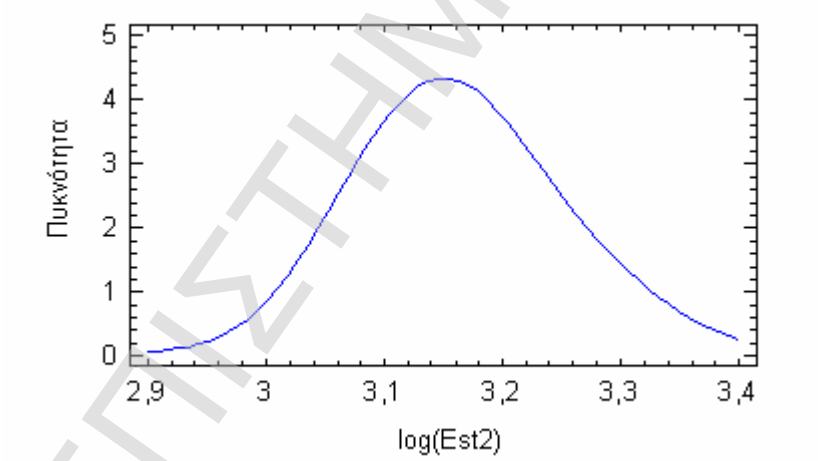
Από το ιστόγραμμα και το διάγραμμα ίχνους της πυκνότητας που ακολουθούν, φαίνεται ότι με βάση το σχήμα της κατανομής (ασυμμετρία και κύρτωση) δεν

μπορούμε να απορρίψουμε την υπόθεση ότι η μεταβλητή *log(Est2)* ακολουθεί κανονική κατανομή.

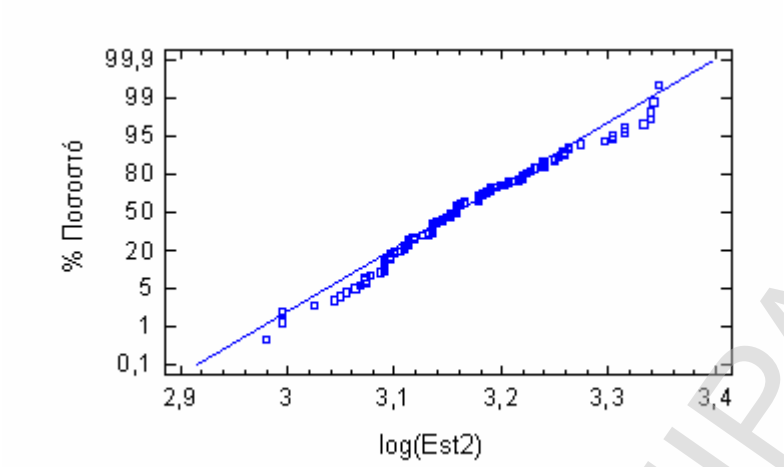
Από το διάγραμμα ελέγχου κανονικότητας τα σημεία προσεγγίζουν την ευθεία γραμμή, αν και υπάρχουν κάποιες αποκλίσεις από αυτήν. Τέλος, το θηκόγραμμα δείχνει μια ισορροπημένη κατανομή των δεδομένων (κεντραρισμένη και χωρίς έκτροπες παρατηρήσεις), ενώ η διάμεσος και ο μέσος όρος δεν έχουν πολύ μεγάλη διαφορά στις τιμές τους.



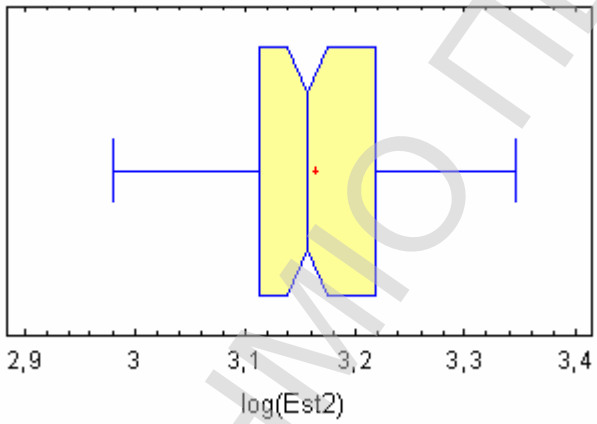
Διάγραμμα 5.42: Ιστόγραμμα της μεταβλητής log(Est2)



Διάγραμμα 5.43: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής log(Est2)



Διάγραμμα 5.44: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής log(Est2)



Διάγραμμα 5.45: Θηκόγραμμα της μεταβλητής log(Est2)

5.2.9. Μεταβλητή Est7

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

Στατιστικά Μέτρα για Est7
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 37,95
Διασπορά = 5,0438
Τυπική Απόκλιση = 2,24584
Τυπικό Σφάλμα = 0,200075
Ελάχιστη Τιμή = 33,2
Μέγιστη Τιμή = 43,7
Εύρος = 10,5
Προτυποποιημένη Ασυμμετρία = 1,86732
Προτυποποιημένη Κύρτωση = 0,0169914

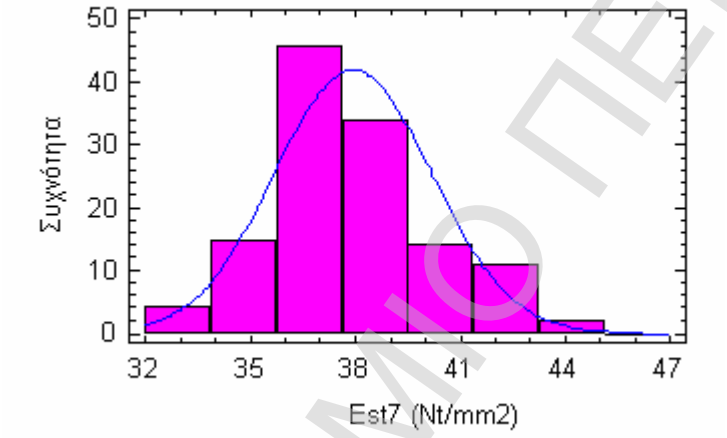
Διαστήματα Εμπιστοσύνης για Est7
95,0% διάστημα εμπιστοσύνης για μέσο όρο: 37,95 +/- 0,395975 [37,554;38,346]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [1,99859;2,5635]

Από τις τιμές της ασυμμετρίας και κύρτωσης παρατηρούμε ότι αυτές ανήκουν στο διάστημα [-2,2], άρα η μεταβλητή Est7 ακολουθεί την κανονική κατανομή.

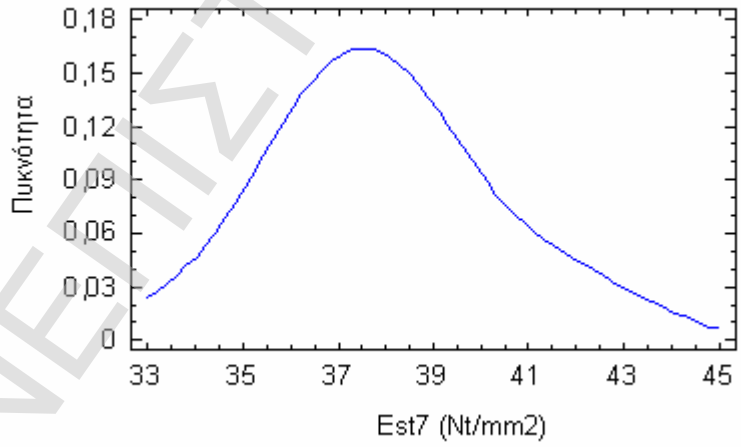
Έλεγχος Κανονικότητας με Διαγράμματα

Από το ιστόγραμμα και το διάγραμμα ίχνους της πυκνότητας που ακολουθούν φαίνεται, ότι με βάση το σχήμα της κατανομής (ασυμμετρία και κύρτωση) δεν μπορούμε να απορρίψουμε την υπόθεση ότι η μεταβλητή Est7 ακολουθεί κανονική κατανομή.

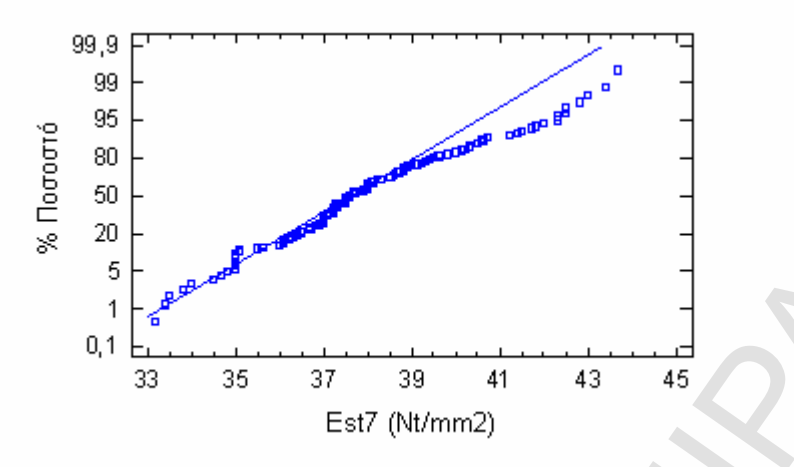
Από το διάγραμμα ελέγχου κανονικότητας φαίνεται ότι τα σημεία προσεγγίζουν την ευθεία γραμμή, αν και προς το τέλος υπάρχουν κάποιες αποκλίσεις από αυτήν. Τέλος, το θηκόγραμμα δείχνει μια σχετικά ισορροπημένη κατανομή των δεδομένων (αν και υπάρχουν τρεις μακρινές παρατηρήσεις), ενώ η διάμεσος και ο μέσος έχουν μια μικρή διαφορά στις τιμές τους.



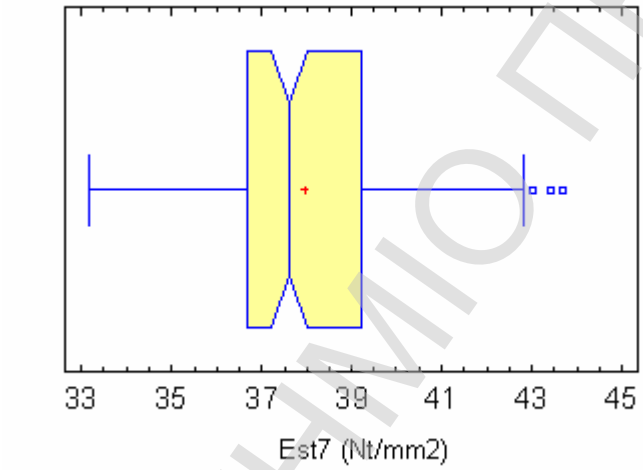
Διάγραμμα 5.46: Ιστόγραμμα της μεταβλητής Est7



Διάγραμμα 5.47: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής Est7



Διάγραμμα 5.48: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής Est7



Διάγραμμα 5.49: Θηκόγραμμα της μεταβλητής Est7

5.2.10. Μεταβλητή Est28

Μέτρα Κεντρικής Τάσης, Διασποράς και Σχήματος

Τα βασικά μέτρα κεντρικής τάσης, διασποράς και σχήματος είναι τα εξής:

Στατιστικά Μέτρα για Est28
Αριθμός Παρατηρήσεων = 126
Μέσος Όρος = 52,3325
Διασπορά = 4,96253
Τυπική Απόκλιση = 2,22767
Τυπικό Σφάλμα = 0,198457
Ελάχιστη Τιμή = 47,0
Μέγιστη Τιμή = 57,0
Εύρος = 10,0
Προτυποποιημένη Ασυμμετρία = 0,114449
Προτυποποιημένη Κύρτωση = -0,941302

Διαστήματα Εμπιστοσύνης για Est28
95,0% διάστημα εμπιστοσύνης για μέσο όρο: 52,3325 +/- 0,392772 [51,9398;52,7253]
95,0% διάστημα εμπιστοσύνης για τυπική απόκλιση: [1,98242;2,54276]

Η μεταβλητή Est28 ακολουθεί κανονική κατανομή, με βάση τον έλεγχο της ασυμμετρίας και κύρτωσης. Συνεπώς, δεν απαιτείται κανένας μετασχηματισμός, αφού όλοι οι έλεγχοι που εκτελούνται είναι έγκυροι. Η ύπαρξη κανονικότητας στην περίπτωση αυτή επιβεβαιώνεται και από τα τεστ X^2 και *Shapiro-Wilks*.

Έλεγχος Κανονικότητας με Διάφορα Tests

Τα αποτελέσματα και τα *p-values* των τεστ που πραγματοποιούνται είναι τα ακόλουθα:

Τεστ Κανονικότητας για Est28

X^2 τεστ για έλεγχο καλής προσαρμογής = 30,4127
P-Value = 0,137917

Shapiro-Wilks W τεστ = 0,972899
P-Value = 0,150683

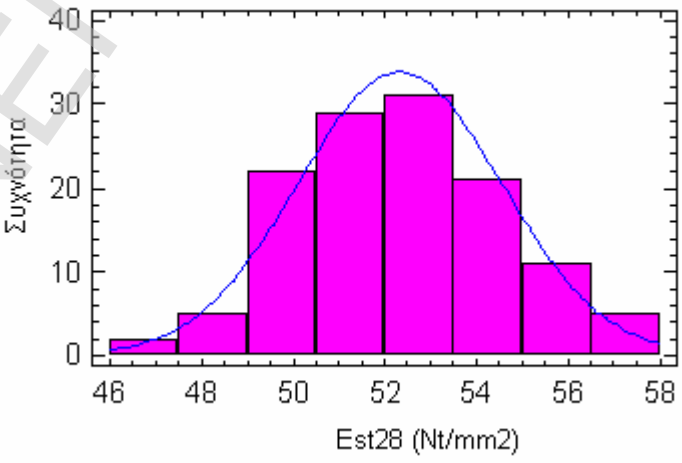
Z score για ασυμμετρία = 0,0841508
P-Value = 0,932931

Z score για κύρτωση = -1,04695
P-Value = 0,295122

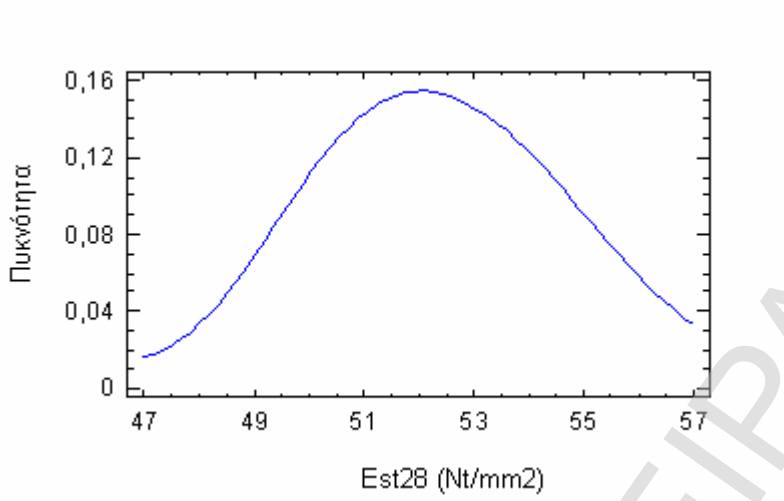
Και στα 4 παραπάνω τεστ το *p-value* είναι μεγαλύτερο από 10%, συνεπώς δεν μπορούμε να απορρίψουμε την υπόθεση ότι η *Est28* κατανέμεται κανονικά, σε επίπεδο εμπιστοσύνης 90% ή περισσότερο.

Έλεγχος Κανονικότητας με Διαγράμματα

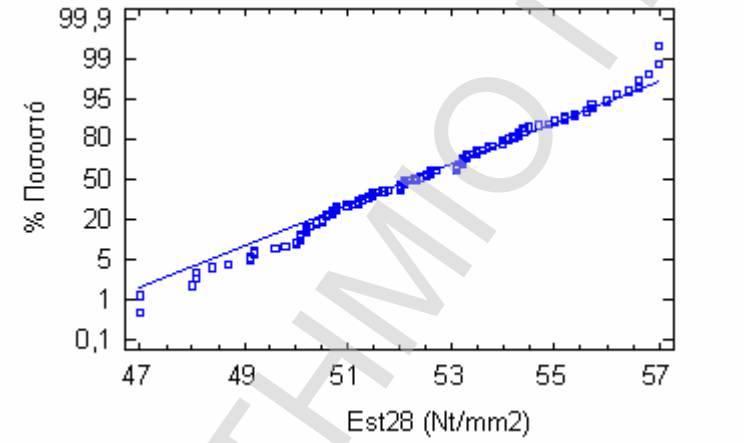
Η μεταβλητή προσεγγίζει σε ικανοποιητικό βαθμό την κανονική κατανομή. Αυτό φαίνεται οπτικά και από τα ακόλουθα διαγράμματα, όπου διακρίνεται το σχήμα καμπάνας, ενώ στο διάγραμμα ελέγχου κανονικότητας υπάρχουν πολύ λίγες αποκλίσεις των σημείων από την ευθεία γραμμή.



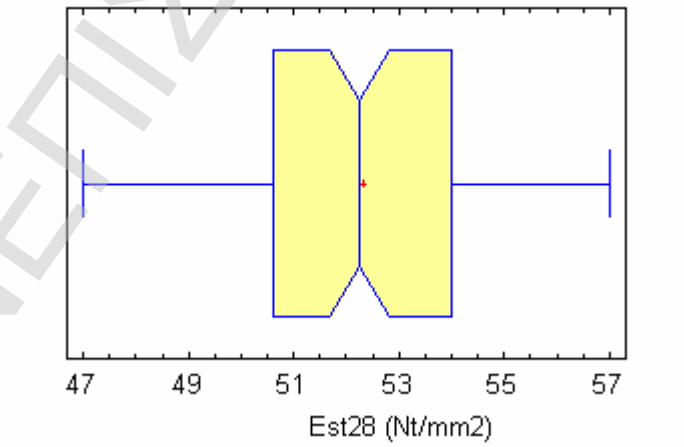
Διάγραμμα 5.50: Ιστόγραμμα της μεταβλητής Est28



Διάγραμμα 5.51: Διάγραμμα Ίχνους της Πυκνότητας της μεταβλητής Est28



Διάγραμμα 5.52: Διάγραμμα Ελέγχου Κανονικότητας της μεταβλητής Est28



Διάγραμμα 5.53: Θηκόγραμμα της μεταβλητής Est28

Τα παραπάνω αποτελέσματα συγκεντρωτικά φαίνονται στον Πίνακα 5.1.

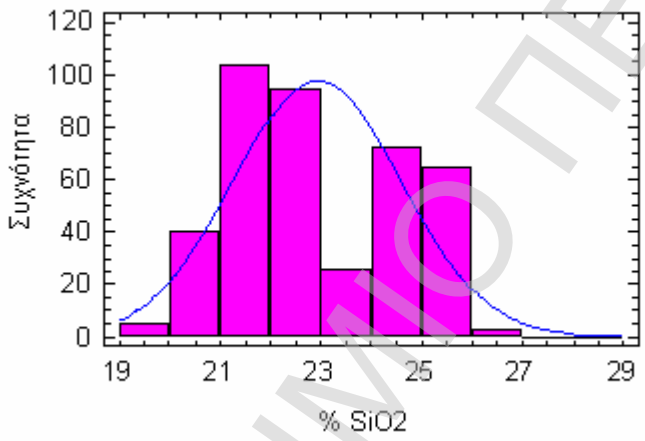
<u>Πίνακας 5.1</u>: Αποτελέσματα Στατιστικής Ανάλυσης για CEM II 42,5 – MT1

	SiO_2	Al_2O_3	*Blaine*	*IR*	*log(LOI)*	*Clk*
Μέσος Όρος	25,1719	5,97571	4704,37	9,97992	1,16274	80,7595
Διακύμανση	0,470435	0,0515847	7061,59	1,09463	0,00694217	0,462589
Τυπ. Απόκλιση	0,685882	0,227123	84,0333	1,04625	0,0833197	0,680139
Τυπ. Σφάλμα	0,061103	0,0202337	7,48628	0,09320	0,00742271	0,060591
Ελάχιστη Τιμή	23,29	5,37	4450,0	7,65	0,916291	78,7
Μέγιστη Τιμή	26,88	6,45	4900,0	13,3	1,41342	82,9
Προτυπ. Ασυμμετρία	1,4065	-1,57524	-1,20873	1,11461	0,454092	0,55208
Προτυπ. Κύρτωση	-0,16684	-0,428845	0,466206	0,71674	1,86048	3,90546
Ύπαρξη Κανονικότητας	**ΝΑΙ**	**ΝΑΙ**	**ΝΑΙ**	**ΝΑΙ**	**ΝΑΙ**	**ΟΧΙ**
	Gyp	*Est2*	*log(Est2)*	*Est7*	*Est28*	
Μέσος Όρος	5,44365	23,7341	3,164	37,95	52,3325	
Διακύμανση	0,095759	3,37059	0,005817	5,0438	4,96253	
Τυπ. Απόκλιση	0,30945	1,83592	0,0762735	2,24584	2,22767	
Τυπ. Σφάλμα	0,027568	0,163556	0,00679498	0,200075	0,198457	
Ελάχιστη Τιμή	4,1	19,7	2,98062	33,2	47,0	
Μέγιστη Τιμή	6,0	28,4	3,34639	43,7	57,0	
Προτυπ. Ασυμμετρία	-5,0297	2,53494	1,57442	1,86732	0,114449	
Προτυπ. Κύρτωση	6,30796	0,320835	0,0214929	0,0169914	-0,94130	
Ύπαρξη Κανονικότητας	**ΟΧΙ**	**ΟΧΙ**	**ΝΑΙ**	**ΝΑΙ**	**ΝΑΙ**	

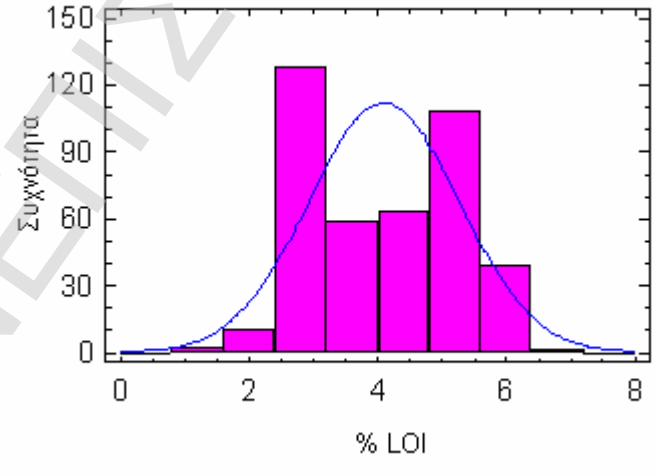
5.3. Σύνθετο Τσιμέντο Portland CEM II 42,5 – Μύλος Παραγωγής 4

Στην ενότητα αυτή ακολουθεί ακριβώς η ίδια διαδικασία στατιστικής ανάλυσης που εφαρμόστηκε και παραπάνω στο Μύλο 1. Δεν κρίνεται σκόπιμο να περιγραφεί αναλυτικά η διαδικασία, γι' αυτό οι σημαντικότερες στατιστικές παράμετροι για καθεμία από τις εννέα μεταβλητές είναι συγκεντρωμένες στον Πίνακα 5.2.

Ωστόσο, σημειώνεται ότι και σε αυτή την περίπτωση γίνεται αισθητή η διαφοροποίηση των τιμών των δεδομένων από τον Ιούλιο του 2004 και μετά, όταν άλλαξε η παραγωγή σε αυτόν τον τύπο τσιμέντου. Παρακάτω παρατίθεται το ιστόγραμμα για τη μεταβλητή SiO_2 και τη LOI, σε ένδειξη της αλλαγής που πραγματοποιήθηκε.



Διάγραμμα 5.54: Ιστόγραμμα της μεταβλητής SiO_2



Διάγραμμα 5.55: Ιστόγραμμα της μεταβλητής LOI

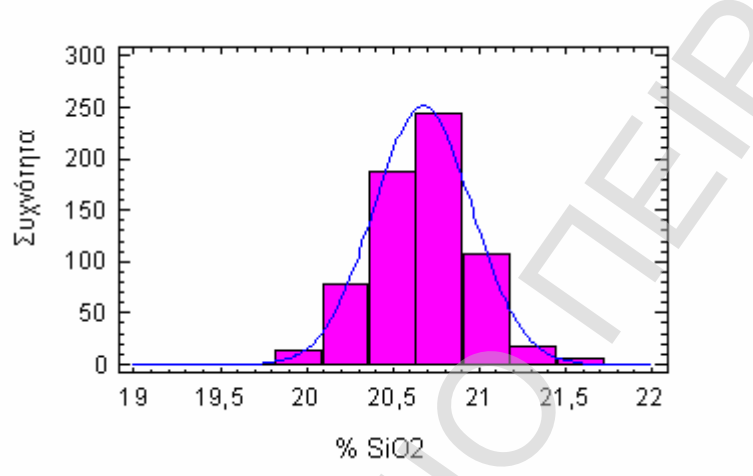
Πίνακας 5.2: Αποτελέσματα Στατιστικής Ανάλυσης για CEM II 42,5 – MT4

	SiO_2	Al_2O_3	*Blaine*	*IR*	*LOI*	*Clk*
Μέσος Όρος	24,9114	5,35099	4186,61	8,82744	2,96893	81,4554
Διακύμανση	0,373407	0,030475	5810,92	0,94475	0,07865	0,503658
Τυπ. Απόκλιση	0,611071	0,174573	76,2294	0,97198	0,280463	0,709689
Τυπ. Σφάλμα	0,05555	0,01587	6,92995	0,08836	0,02549	0,064517
Ελάχιστη Τιμή	22,74	4,95	3960,0	5,65	2,17	79,3
Μέγιστη Τιμή	26,9	5,79	4360,0	12,01	3,61	83,4
Προτυπ. Ασυμμετρία	-1,62774	-1,591	-1,50071	-2,23947	-1,45247	-1,30849
Προτυπ. Κύρτωση	3,815	-0,892642	1,45665	3,91279	0,045329	2,38938
Ύπαρξη Κανονικότητας	**ΟΧΙ**	**ΝΑΙ**	**ΝΑΙ**	**ΟΧΙ**	**ΝΑΙ**	**ΟΧΙ**
	Gyp	***Est2***	***Est7***	***Est28***		
Μέσος Όρος	5,21901	24,0587	38,9719	53,4545		
Διακύμανση	0,106552	1,79628	1,97454	3,07617		
Τυπ. Απόκλιση	0,326424	1,34025	1,40518	1,7539		
Τυπ. Σφάλμα	0,0296749	0,121841	0,127744	0,159445		
Ελάχιστη Τιμή	4,0	21,1	35,7	48,7		
Μέγιστη Τιμή	6,1	27,1	42,4	57,6		
Προτυπ. Ασυμμετρία	2,06229	0,76898	0,413715	-1,72604		
Προτυπ. Κύρτωση	3,21403	-0,671615	-0,207154	0,616994		
Ύπαρξη Κανονικότητας	**ΟΧΙ**	**ΝΑΙ**	**ΝΑΙ**	**ΝΑΙ**		

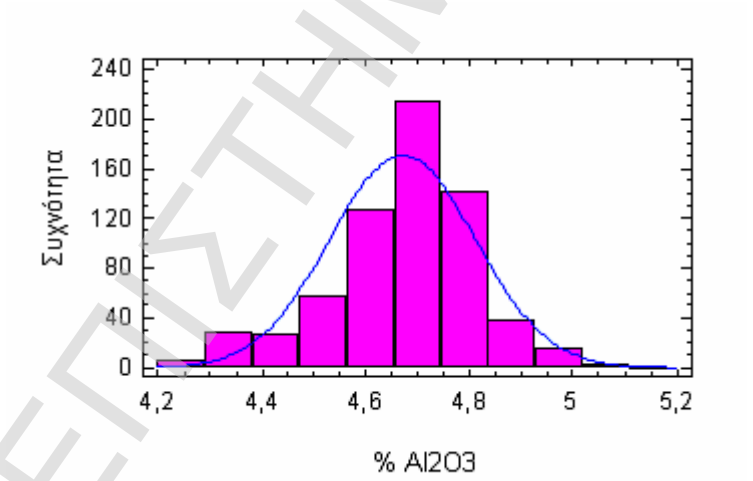
Από τις παραπάνω μεταβλητές οι SiO_2, IR, Clk και Gyp δεν παρουσιάζουν κανονικότητα. Επιχειρήθηκαν κάποιοι μετασχηματισμοί αυτών, αλλά το αποτέλεσμα παρέμεινε το ίδιο. Ωστόσο, οι μεταβλητές Est2 και Est28 είναι κανονικά κατανεμημένες και αυτό είναι το σημαντικό, καθώς οι μεταβλητές αυτές αποτελούν τις εξαρτημένες μεταβλητές στην ανάλυση παλινδρόμησης που ακολουθεί στο επόμενο κεφάλαιο.

5.4. Ordinary Portland Cement (OPC) – Μύλος Παραγωγής 3

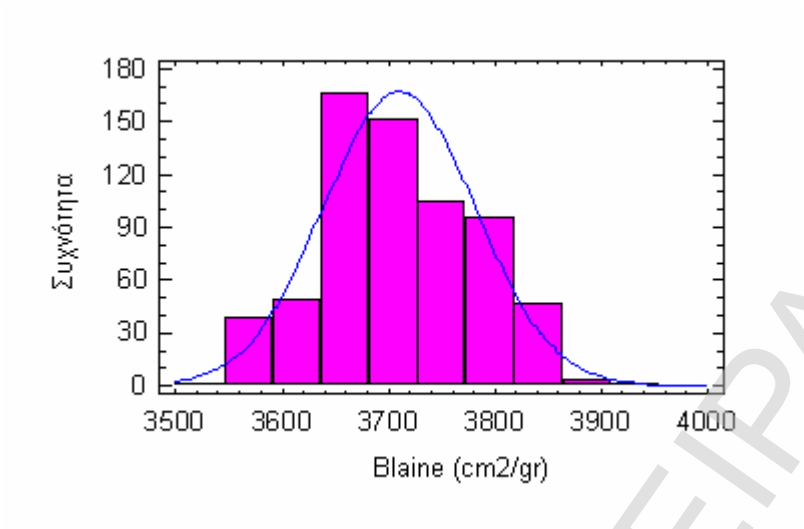
Στην περίπτωση αυτού του τύπου τσιμέντου, η επεξεργασία των δεδομένων και τα ιστογράμματα των μεταβλητών δεν δείχνουν ότι οι μεταβλητές έχουν υποστεί σημαντικές αλλαγές, που συνήθως οφείλονται στις συνθήκες παραγωγής. Για το λόγο αυτόν, τα δεδομένα που χρησιμοποιούνται είναι από τις αρχές του 2003 έως τις αρχές Αυγούστου 2005. Ενδεικτικά, παρακάτω φαίνονται τα ιστογράμματα για καθεμία από τις εννέα μεταβλητές.



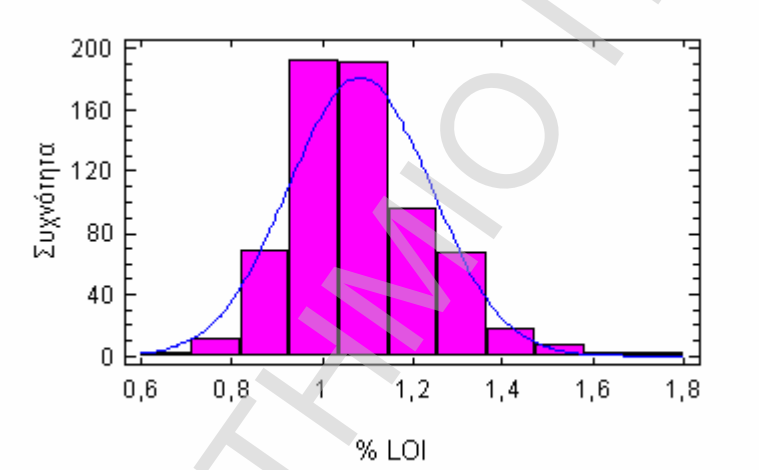
Διάγραμμα 5.56: Ιστόγραμμα της μεταβλητής SiO_2



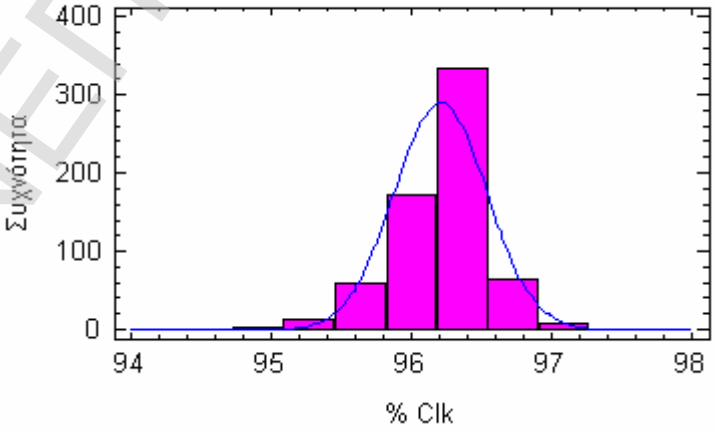
Διάγραμμα 5.57: Ιστόγραμμα της μεταβλητής Al_2O_3



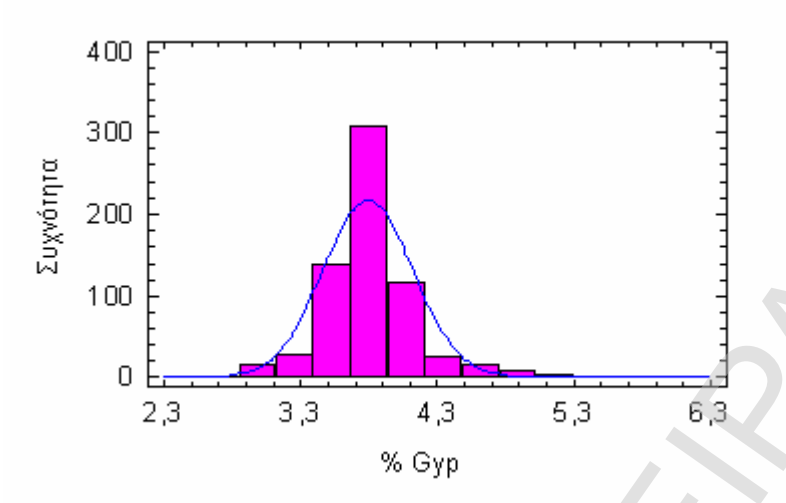
<u>Διάγραμμα 5.58</u>: Ιστόγραμμα της μεταβλητής Blaine



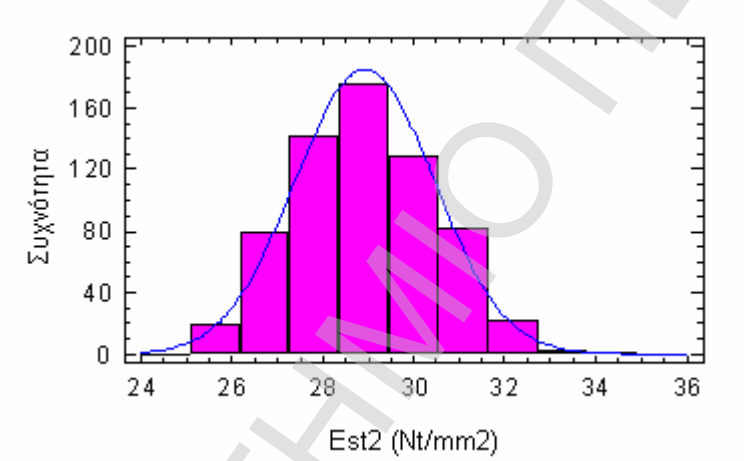
<u>Διάγραμμα 5.59</u>: Ιστόγραμμα της μεταβλητής LOI



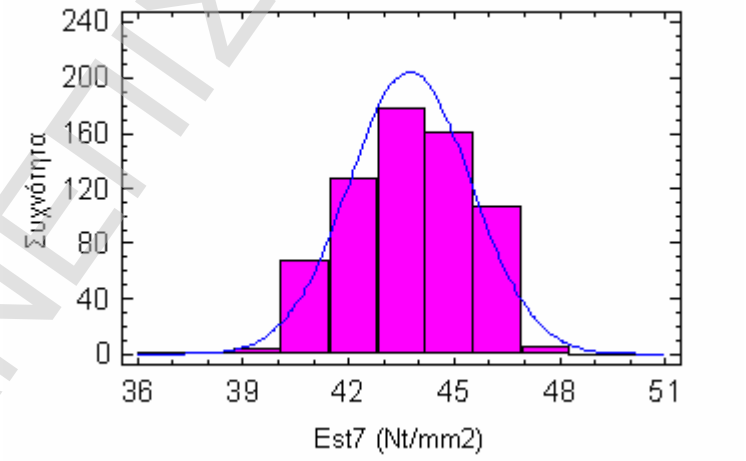
<u>Διάγραμμα 5.60</u>: Ιστόγραμμα της μεταβλητής Clk



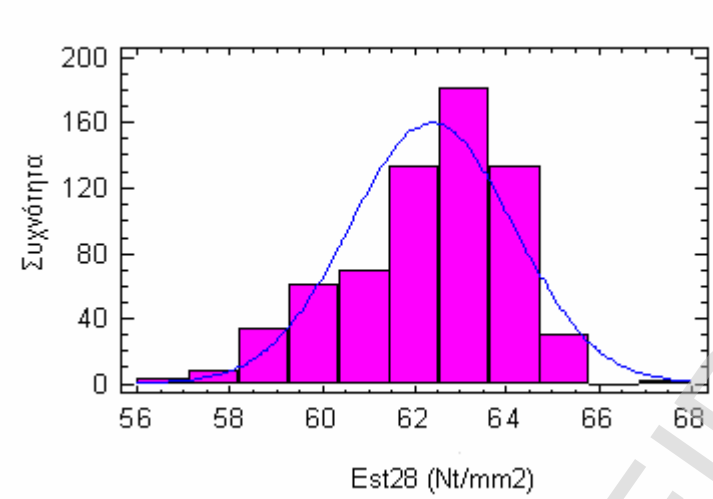
<u>Διάγραμμα 5.61</u>: Ιστόγραμμα της μεταβλητής Gyp



<u>Διάγραμμα 5.62</u>: Ιστόγραμμα της μεταβλητής Est2



<u>Διάγραμμα 5.63</u>: Ιστόγραμμα της μεταβλητής Est7



Διάγραμμα 5.64: Ιστόγραμμα της μεταβλητής Est28

Οι περισσότερες από τις παραπάνω μεταβλητές δεν ακολουθούν κανονική κατανομή, ακόμα και μετά από προσπάθεια μετασχηματισμού αυτών. Εξαίρεση αποτελεί μόνο η μεταβλητή Est2. Αυτό μπορεί να οφείλεται στους εξής λόγους:

- Μερικές μεταβλητές διατηρούνται μέσα σε αυστηρά επίπεδα τιμών, συνεπώς τα δεδομένα συγκεντρώνονται γύρω από κάποια τιμή και έχουν πολύ μικρή διασπορά. Τέτοιο παράδειγμα αποτελούν κυρίως οι μεταβλητές Clk, Gyp και LOI.

- Επίσης, μπορεί η έλλειψη κανονικότητας να οφείλεται στο γεγονός ότι οι μετρήσεις δεν γίνονται σε τακτά χρονικά διαστήματα (κάθε μέρα στη συγκεκριμένη περίπτωση), αλλά μπορεί να περάσουν και δύο εβδομάδες μεταξύ δύο μετρήσεων. Πιθανώς, αυτό συμβαίνει διότι κάθε μύλος δεν παράγει μόνο ένα είδος τσιμέντου, αλλά περισσότερα. Συνεπώς, οι μετρήσεις ενός είδους τσιμέντου σε ένα μύλο παρουσιάζουν κάποια χρονικά κενά.

- Τέλος, ακριβώς επειδή συμβαίνει αυτή η εναλλαγή στην παραγωγή, πολλές λανθασμένες μετρήσεις μπορούν να γίνουν κατά την έναρξη λειτουργίας του μύλου κάθε φορά, πιθανότατα επειδή έχουν μείνει κάποια υπολείμματα προηγούμενου τύπου τσιμέντου ή επειδή δεν έχουν σταθεροποιηθεί κάποιες συνθήκες.

Στον Πίνακα 5.3 φαίνονται οι τιμές όλων των βασικών στατιστικών παραμέτρων που εξετάζονται.

Πίνακας 5.3: Αποτελέσματα Στατιστικής Ανάλυσης για OPC – MT3

	SiO_2	Al_2O_3	Blaine	LOI	Clk
Μέσος Όρος	20,6761	4,67173	3710,2	1,08391	96,2065
Διακύμανση	0,0812464	0,0196016	5092,4	0,0249106	0,108131
Τυπ. Απόκλιση	0,285038	0,140006	71,3611	0,157831	0,328832
Τυπ. Σφάλμα	0,0111119	0,0054579	2,78194	0,0061528	0,0128192
Ελάχιστη Τιμή	19,83	4,25	3530,0	0,67	94,9
Μέγιστη Τιμή	21,63	5,03	3910,0	1,72	97,5
Προτυπ. Ασυμμετρία	-0,103991	-7,16985	0,831192	7,41531	-4,72395
Προτυπ. Κύρτωση	3,03349	2,97216	-2,93543	5,12283	10,7728
Ύπαρξη Κανονικότητας	**ΟΧΙ**	**ΟΧΙ**	**ΟΧΙ**	**ΟΧΙ**	**ΟΧΙ**
	Gyp	***Est2***	***Est7***	***Est28***	
Μέσος Όρος	3,79347	28,9288	43,7488	62,3821	
Διακύμανση	0,108131	2,35058	3,01519	3,1511	
Τυπ. Απόκλιση	0,328832	1,53316	1,73643	1,77514	
Τυπ. Σφάλμα	0,0128192	0,0599972	0,0680039	0,0694664	
Ελάχιστη Τιμή	2,5	25,1	37,0	56,7	
Μέγιστη Τιμή	5,1	33,9	47,8	67,2	
Προτυπ. Ασυμμετρία	4,72395	1,59464	-2,21478	-6,9872	
Προτυπ. Κύρτωση	10,7728	-1,98057	-1,9414	0,11466	
Ύπαρξη Κανονικότητας	**ΟΧΙ**	**ΝΑΙ**	**ΟΧΙ**	**ΟΧΙ**	

5.5. Ordinary Portland Cement (OPC) – Μύλος Παραγωγής 4

Στο Μύλο Παραγωγής 4 του OPC τα δεδομένα που χρησιμοποιούνται είναι από τις αρχές 2003 μέχρι τις 7/8/2005. Τα δεδομένα έχουν μετρηθεί σε τακτικά διαστήματα, ωστόσο όχι σε καθημερινή βάση (όπως και σε όλες τις άλλες περιπτώσεις). Λίγες μεταβλητές φαίνεται να είναι κανονικά κατανεμημένες. Οι μεταβλητές σχετικά με τις τελικές αντοχές του τσιμέντου στις 2 και στις 28 ημέρες (οι οποίες παρουσιάζουν και το μεγαλύτερο ενδιαφέρον) δεν ακολουθούν κανονική κατανομή. Η μεταβλητή Est7 είναι κανονικά κατανεμημένη, όταν μετασχηματιστεί σε log(Est7).

Στον Πίνακα 5.4 φαίνονται οι σημαντικότερες παράμετροι των μεταβλητών.

Πίνακας 5.4: Αποτελέσματα Στατιστικής Ανάλυσης για OPC – MT4

	SiO_2	***Al_2O_3***	***Blaine***	***LOI***	***Clk***
Μέσος Όρος	20,7076	4,62805	3729,41	1,23027	96,2346
Διακύμανση	0,086164	0,0191078	5130,72	0,0265813	0,18694
Τυπ. Απόκλιση	0,293537	0,138231	71,6291	0,163038	0,432365
Τυπ. Σφάλμα	0,0139621	0,0065749	3,40705	0,0077549	0,0205655
Ελάχιστη Τιμή	19,78	4,23	3480,0	0,84	93,1
Μέγιστη Τιμή	21,65	4,97	3940,0	1,99	97,9
Προτυπ. Ασυμμετρία	1,69876	-6,45705	-1,67234	7,82274	-11,4225
Προτυπ. Κύρτωση	1,90624	2,03585	0,416617	9,40652	37,6729
Ύπαρξη Κανονικότητας	**ΝΑΙ**	**ΟΧΙ**	**ΝΑΙ**	**ΟΧΙ**	**ΟΧΙ**
	Gyp	***Est2***	***Est7***	***log(Est7)***	***Est28***
Μέσος Όρος	3,75928	27,4876	42,2557	3,74297	61,3568
Διακύμανση	0,16505	1,52726	2,77304	0,00153826	3,84201
Τυπ. Απόκλιση	0,406263	1,23582	1,66524	0,0392206	1,9601
Τυπ. Σφάλμα	0,019324	0,0587821	0,0792076	0,00186553	0,0932326
Ελάχιστη Τιμή	2,1	25,0	38,2	3,64284	55,9
Μέγιστη Τιμή	6,0	31,1	47,8	3,86703	66,7
Προτυπ. Ασυμμετρία	5,20115	3,58502	2,87203	1,98634	-3,6743
Προτυπ. Κύρτωση	17,1986	-0,532504	-0,395746	-0,883918	-1,878
Ύπαρξη Κανονικότητας	**ΟΧΙ**	**ΟΧΙ**	**ΟΧΙ**	**ΝΑΙ**	**ΟΧΙ**

6. ΚΕΦΑΛΑΙΟ 6: ΑΠΟΤΕΛΕΣΜΑΤΑ ΠΟΛΛΑΠΛΗΣ ΚΑΙ ΑΠΛΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

6.1. Μεθοδολογία Πολλαπλής και Απλής Παλινδρόμησης

Στην παρούσα μελέτη εξετάζονται όλες οι μεταβλητές που αναλύθηκαν στο προηγούμενο κεφάλαιο, **εκτός** από τις μεταβλητές **Clk** και **Gyp**. Οι μεταβλητές αυτές ελέγχονται σε αρκετά μεγάλο βαθμό, έτσι ώστε οι τιμές τους να είναι συγκεντρωμένες γύρω από κάποια τιμή. Συνεπώς, αφού η μεταβλητότητά τους είναι περιορισμένη και ελεγχόμενη, δεν συμπεριλαμβάνονται ως ανεξάρτητες μεταβλητές στα μοντέλα παλινδρόμησης. Από τις άλλες μεταβλητές, οι περισσότερες είναι οι ανεξάρτητες μεταβλητές (όλες οι περιεκτικότητες σε χημικές ουσίες και κάποια φυσικοχημικά χαρακτηριστικά), ενώ οι υπόλοιπες δύο, δηλαδή οι τελικές αντοχές του τσιμέντου κατά τη δεύτερη και την εικοστή όγδοη ημέρα, αποτελούν τις εξαρτημένες μεταβλητές. Οι τελικές αντοχές του τσιμέντου κατά την έβδομη ημέρα είναι η ανεξάρτητη μεταβλητή σε ένα μοντέλο απλής παλινδρόμησης.

Στόχο της μελέτης αποτελεί αρχικά η συσχέτιση όλων των μεταβλητών μεταξύ τους, και στη συνέχεια η εύρεση μιας εξίσωσης πολλαπλής παλινδρόμησης μεταξύ κάθε μιας από τις ανεξάρτητες μεταβλητές και όλων των υπόλοιπων ανεξαρτήτων. Επίσης, βρίσκεται η πλέον κατάλληλη εξίσωση συσχέτισης (απλή παλινδρόμηση) μεταξύ των τελικών αντοχών στις 28 ημέρες και των τελικών αντοχών στις 2 και στις 7 ημέρες. Η παραπάνω διαδικασία εφαρμόζεται για τους δύο τύπους τσιμέντου, για κάθε μύλο παραγωγής ξεχωριστά.

Όλες οι μεταβλητές που θα εξεταστούν είναι ήδη γνωστές και καθορισμένες. Από το προηγούμενο κεφάλαιο έχει γίνει η αρχική στατιστική ανάλυση των δεδομένων και έχει βρεθεί ποιες μεταβλητές είναι κανονικά κατανεμημένες και ποιες όχι. Στη συνέχεια της ανάλυσης ακολουθούν **διαδοχικά** τα εξής **βήματα**:

1. Έλεγχος συσχετίσεων μεταξύ των ζευγών όλων των μεταβλητών, εξαρτημένων και ανεξαρτήτων. Εάν οι μεταβλητές ακολουθούν κανονική κατανομή, επιλέγουμε τον πίνακα συσχετίσεων κατά *Pearson*. Εάν δεν υπάρχει κανονικότητα, τότε ελέγχουμε τον πίνακα *Spearman Rank Correlations*, ο οποίος είναι λιγότερο ευαίσθητος στις μακρινές παρατηρήσεις (outliers). Από τον πίνακα των συσχετίσεων ελέγχουμε δύο πράγματα: Πρώτον, βρίσκουμε η εξαρτημένη μεταβλητή αν και με ποιες μεταβλητές συσχετίζεται, συνεπώς υπάρχει πιθανότητα αυτές να συμπεριλαμβάνονται στο μοντέλο παλινδρόμησης (πολλαπλή και απλή). Αυτό βέβαια δεν είναι υποχρεωτικό να συμβεί, καθώς το γεγονός ότι υπάρχει συσχέτιση μεταξύ κάποιων μεταβλητών δεν σημαίνει απαραίτητα ότι υπάρχει και ερμηνευτικότητα ή αιτιότητα (causality), δηλαδή ότι η μία μεταβλητή επηρεάζει την άλλη. Δεύτερον, ελέγχουμε τις συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών. Σε περίπτωση που υπάρχουν τέτοιες συσχετίσεις, έχουμε πρόβλημα πολυσυγγραμικότητας και λαμβάνουμε μέτρα για την επίλυση του προβλήματος (βήμα 2).

2. Δημιουργούμε τα δύο μοντέλα πολλαπλής παλινδρόμησης των Est28 και Est2, περιλαμβάνοντας όλες τις ανεξάρτητες μεταβλητές. Εάν από τον πίνακα συσχετίσεων στο βήμα 1 προκύψει ότι μερικές από τις ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους, τότε εφαρμόζουμε επιπρόσθετα και τη μέθοδο βηματικής επιλογής μεταβλητών, με τη μέθοδο της προς τα εμπρός επιλογής (***forward selection***) που παρέχει το στατιστικό πρόγραμμα. Στο τελευταίο μοντέλο, ελέγχουμε πάλι αν υπάρχει συσχέτιση μεταξύ των μεταβλητών που έχουν προκριθεί. Αν αυτό συμβαίνει, τότε εφαρμόζουμε μία πιο προχωρημένη μέθοδο παλινδρόμησης, τη ραχοειδή παλινδρόμηση (***ridge regression***). Η μέθοδος αυτή χρησιμοποιείται στις περιπτώσεις ύπαρξης πολυσυγγραμικότητας.

3. Δημιουργούμε τα δύο μοντέλα απλής παλινδρόμησης μεταξύ της εξαρτημένης Est28 και της ανεξάρτητης Est2 και Est7. Στην περίπτωση αυτή, επιλέγουμε από ένα πλήθος μοντέλων απλής παλινδρόμησης το πλέον κατάλληλο για τα δεδομένα μας, με κριτήριο την υψηλότερη τιμή του συντελεστή προσδιορισμού R^2.

4. Στο τελευταίο βήμα, το οποίο είναι πολύ σημαντικό, γίνεται ο έλεγχος για την ισχύ των προϋποθέσεων της παλινδρόμησης. Έλεγχος γίνεται στην πολλαπλή παλινδρόμηση με τη μέθοδο της προς τα εμπρός επιλογής (καθώς αυτή είναι η πιο έγκυρη μέθοδος), και στις δύο περιπτώσεις απλής παλινδρόμησης. Πέρα από τον έλεγχο για ύπαρξη πολυσυγγραμικότητας, που εντοπίζεται και επιλύεται στο βήμα 2, ελέγχουμε για την ύπαρξη κανονικότητας στα κατάλοιπα, που στην ουσία ισοδυναμεί με την ύπαρξη κανονικότητας στην εξαρτημένη μεταβλητή. Επίσης, μέσα από τα διαγράμματα των καταλοίπων ως προς την εξαρτημένη και τις ανεξάρτητες μεταβλητές ελέγχουμε αν το μέσο σφάλμα είναι μηδέν και αν η διασπορά είναι σταθερή. Τέλος, από τα διαγράμματα των καταλοίπων ως προς τον αριθμό σειράς, αλλά κυρίως από το *test Durbin-Watson*, ελέγχουμε για την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

6.2. Σύνθετο Τσιμέντο Portland CEM II 42,5 – Μύλος Παραγωγής 1

Τα δεδομένα που χρησιμοποιούνται για την ανάλυση που ακολουθεί βρίσκονται στο χρονικό διάστημα 7/2004 έως 8/2005. Δηλαδή, χρησιμοποιούνται τα δεδομένα μετά την ημερομηνία αλλαγής της παραγωγής στο τσιμέντο CEM II 42,5.

6.2.1. Έλεγχος Συσχετίσεων μεταξύ των Μεταβλητών

Στο σημείο αυτό ελέγχεται ο πίνακας συσχετίσεων κατά Spearman, επειδή η μέθοδος αυτή ενδείκνυται για περιπτώσεις μη ύπαρξης κανονικότητας, και -όπως είδαμε στο προηγούμενο κεφάλαιο- υπάρχει κάποια επιφύλαξη για την ύπαρξη κανονικότητας στη μεταβλητή log(Est2). Τα αποτελέσματα φαίνονται στον Πίνακα 6.1.

<u>Πίνακας 6.1</u>: Πίνακας Συσχετίσεων των Μεταβλητών κατά Spearman

	SiO_2	***Al_2O_3***	***Blaine***	***IR***	***log(LOI)***	***log(Est2)***	***Est7***	***Est28***
SiO_2		0,0779 (126) 0,3839	**-0,2239** (126) **0,0123**	**0,3279** (126) **0,0002**	**0,2986** (126) **0,0008**	-0,1729 (126) 0,0532	-0,1176 (126) 0,1888	0,0361 (126) 0,6864
Al_2O_3	0,0779 (126) 0,3839		0,0062 (126) 0,9444	0,1055 (126) 0,2380	**-0,1821** (126) **0,0417**	**-0,2405** (126) **0,0072**	**-0,1908** (126) **0,0329**	-0,0884 (126) 0,3228
Blaine	**-0,2239** (126) **0,0123**	0,0062 (126) 0,9444		0,0269 (126) 0,7633	-0,0328 (126) 0,7140	**0,3471** (126) **0,0001**	**0,3844** (126) **0,0000**	**0,3181** (126) **0,0004**
IR	**0,3279** (126) **0,0002**	0,1055 (126) 0,2380	0,0269 (126) 0,7633		0,0492 (126) 0,5826	-0,0555 (126) 0,5353	0,0038 (126) 0,9661	-0,0549 (126) 0,5393
log(LOI)	**0,2986** (126) **0,0008**	**-0,1821** (126) **0,0417**	-0,0328 (126) 0,7140	0,0492 (126) 0,5826		**-0,2617** (126) **0,0034**	-0,1508 (126) 0,0917	**-0,1770** (126) **0,0479**
log(Est2)	-0,1729 (126) 0,0532	**-0,2405** (126) **0,0072**	**0,3471** (126) **0,0001**	-0,0555 (126) 0,5353	**-0,2617** (126) **0,0034**		**0,8467** (126) **0,0000**	**0,7030** (126) **0,0000**
Est7	-0,1176 (126) 0,1888	**-0,1908** (126) **0,0329**	**0,3844** (126) **0,0000**	0,0038 (126) 0,9661	-0,1508 (126) 0,0917	**0,8467** (126) **0,0000**		**0,7123** (126) **0,0000**
Est28	0,0361 (126) 0,6864	-0,0884 (126) 0,3228	**0,3181** (126) **0,0004**	-0,0549 (126) 0,5393	**-0,1770** (126) **0,0479**	**0,7030** (126) **0,0000**	**0,7123** (126) **0,0000**	

Ο παραπάνω πίνακας δείχνει τις συσχετίσεις μεταξύ όλων των μεταβλητών. Στην πρώτη γραμμή αναγράφονται οι συντελεστές συσχέτισης. Οι συντελεστές αυτοί μπορούν να πάρουν τιμές από -1 έως +1 και μετράνε το βαθμό συσχέτισης μεταξύ των μεταβλητών. Στη δεύτερη σειρά γράφεται ο αριθμός των ζευγών των τιμών των δεδομένων που χρησιμοποιούνται για τον υπολογισμό των συντελεστών. Ο τρίτος αριθμός σε κάθε κελί του πίνακα είναι το *p-value* που εξετάζει τη στατιστική σημαντικότητα των εκτιμώμενων συντελεστών συσχέτισης. *P-values* μικρότερες από 0,05 φανερώνουν στατιστικά σημαντικές μη-μηδενικές συσχετίσεις, σε 95 % επίπεδο εμπιστοσύνης. Τα ζεύγη των μεταβλητών που έχουν *p-values* κάτω από 0,05 επισημαίνονται με έντονο χρώμα στον παραπάνω πίνακα, και είναι τα ακόλουθα:

SiO2 και **Blaine**
SiO2 και **IR**
SiO2 και **LOG(LOI)**
Al2O3 και **LOG(LOI)**
Al2O3 και **LOG(Est2)**
Al2O3 και **Est7**
Blaine και **LOG(Est2)**
Blaine και **Est7**
Blaine και **Est28**
LOG(LOI) και **LOG(Est2)**
LOG(LOI) και **Est28**
LOG(Est2) και **Est7**
LOG(Est2) και **Est28**
Est7 και **Est28**

Πολλές ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους, συνεπώς υπάρχει πρόβλημα πολυσυγγραμικότητας. Γι' αυτό, στο μοντέλο παλινδρόμησης θα εφαρμοστεί και η μέθοδος της προς τα εμπρός επιλογής.

Συμπεραίνουμε, επίσης, ότι η εξαρτημένη μεταβλητή Est28 συσχετίζεται με τις μεταβλητές Blaine, log(LOI), log(Est2) και Est7. Άρα, στην πολλαπλή παλινδρόμηση είναι πιθανόν ότι οι μεταβλητές Blaine και log(LOI) θα είναι παρούσες, ενώ μπορούμε να εφαρμόσουμε ένα μοντέλο απλής παλινδρόμησης με τη μεταβλητή log(Est2) και την Est7.

Όσον αφορά την εξαρτημένη μεταβλητή log(Est2), παρατηρούμε ότι συσχετίζεται με τις μεταβλητές Al_2O_3, Blaine και log(LOI), οι οποίες είναι πιθανόν να προκριθούν στο μοντέλο παλινδρόμησης.

6.2.2. Πολλαπλή Παλινδρόμηση

- ***Εξαρτημένη Μεταβλητή Est28 – Όλες οι Μεταβλητές***

Με την επιλογή όλων των μεταβλητών, το μοντέλο παλινδρόμησης που προκύπτει είναι το εξής:

Est28 = 0,13235 + 0,917126*SiO2 - 1,79699*Al2O3 + 0,0109734*Blaine - 0,293312*IR - 7,6051*log(LOI)

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.2.

Πίνακας 6.2: Πολλαπλή Παλινδρόμηση για Est28-Όλες οι Μεταβλητές, CEM II 42,5, MT1

Εξαρτημένη μεταβλητή: Est28

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	0,13235	13,6958	0,0096635	0,9923
SiO2	0,917126	0,30469	3,01003	0,0032
Al2O3	-1,79699	0,808933	-2,22143	0,0282
Blaine	0,0109734	0,00217523	5,04469	0,0000
IR	-0,293312	0,181034	-1,6202	0,1078
LOG(LOI)	-7,6051	2,35705	-3,22654	0,0016

Ανάλυση Διακύμανσης

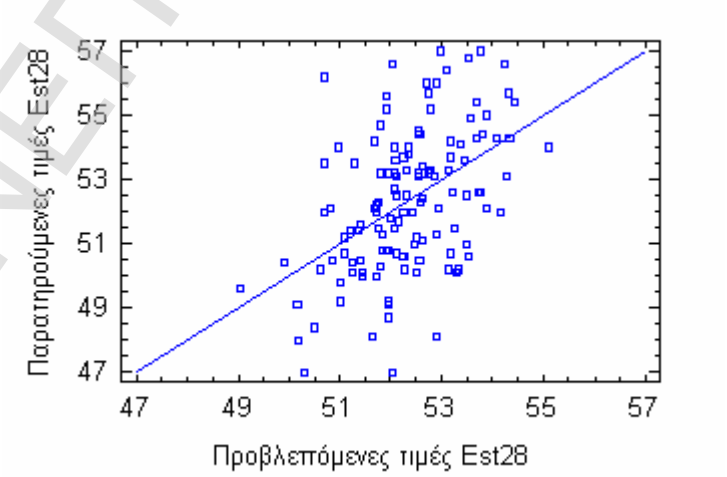
	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	144,449	5	28,8899	7,29	0,0000
Κατάλοιπα	475,867	120	3,96556		
Σύνολο	620,317	125			

R-2 = 23,2864 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 20,09 τοις εκατό
Τυπικό σφάλμα = 1,99137
Μέσο απόλυτο σφάλμα = 1,54666
Durbin-Watson statistic = 1,51351 (P=0,0029)

Από την ανάλυση διακύμανσης (Analysis of Variance) προκύπτει ότι το *p-value* είναι μικρότερο από 0,01, επομένως υπάρχει μια στατιστικά σημαντική σχέση ανάμεσα στις μεταβλητές, σε 99% επίπεδο εμπιστοσύνης.

Ο συντελεστής προσδιορισμού R^2 δείχνει ότι το προσαρμοσμένο μοντέλο εξηγεί 23,2864% της μεταβλητότητας στη μεταβλητή Est28. Ο διορθωμένος συντελεστής R^2, που είναι πιο κατάλληλος για σύγκριση μοντέλων πολλαπλής παλινδρόμησης, είναι 20,09%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 1,99137. Η τιμή αυτή μπορεί να χρησιμοποιηθεί για τη δημιουργία διαστημάτων πρόβλεψης για νέες παρατηρήσεις. Το μέσο απόλυτο σφάλμα (MAE) είναι 1,54666 και αποτελεί τη μέση τιμή των καταλοίπων. Το *Durbin-Watson (DW) statistic* ελέγχει τα κατάλοιπα για το αν υπάρχει αυτοσυσχέτιση. Επειδή το *p-value* είναι μικρότερο από 0,05, υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής Est28 σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές. Όσο πιο κοντά τα δεδομένα "πέφτουν" στη διαγώνια γραμμή, τόσο καλύτερο είναι το μοντέλο στην πρόβλεψη των παρατηρούμενων δεδομένων. Στη συγκεκριμένη περίπτωση οι παρατηρούμενες τιμές αποκλίνουν αρκετά από την ευθεία γραμμή, άρα το μοντέλο δεν θεωρείται ιδιαίτερα καλό. Αυτό, άλλωστε, υποδηλώνεται και από τη σχετικά μικρή τιμή του συντελεστή R^2.



<u>Διάγραμμα 6.1</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est28-Όλες οι Μεταβλητές, CEM II 42,5, MT1

- ***<u>Εξαρτημένη Μεταβλητή Est28 –Προς τα Εμπρός Επιλογή</u>***

Με την επιλογή της μεθόδου προς τα εμπρός επιλογής το μοντέλο παλινδρόμησης που προκύπτει είναι το εξής:

Est28 = 8,02767 + 0,00941782*Blaine

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.3.

<u>Πίνακας 6.3</u>: Πολλαπλή Παλινδρόμηση για Est28-Προς τα Εμπρός Επιλογή, CEM II 42,5, MT1

Εξαρτημένη μεταβλητή: Est28

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	8,02767	10,4704	0,76670	0,4447
Blaine	0,00941782	0,00222532	4,23212	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	78,2913	1	78,2913	17,91	0,0000
Κατάλοιπα	542,025	124	4,37117		
Σύνολο	620,317	125			

R-2 = 12,6212 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 11,9165 τοις εκατό
Τυπικό Σφάλμα = 2,09073
Μέσο απόλυτο σφάλμα = 1,71213
Durbin-Watson statistic = 1,39104 (P=0,0003)

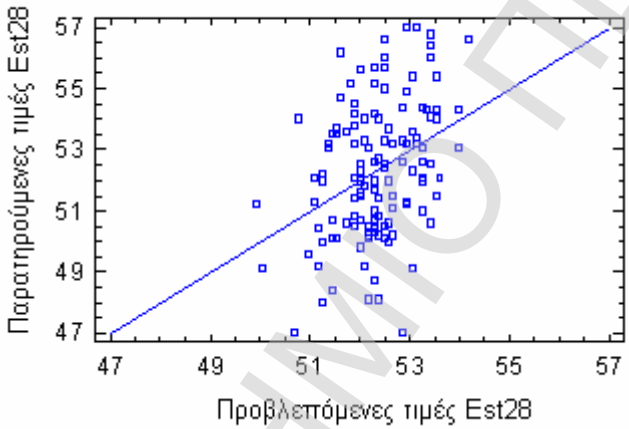
Η μεταβλητή Blaine είναι η μοναδική που προκρίνεται στο μοντέλο παλινδρόμησης. Από τον παραπάνω πίνακα των συσχετίσεων είδαμε ότι υπάρχει συσχέτιση μεταξύ της Est28 και Blaine, το οποίο ήταν και αναμενόμενο. Από την ανάλυση διακύμανσης (Analysis of Variance) προκύπτει ότι το ***p-value*** είναι μικρότερο από 0,01, επομένως υπάρχει μια στατιστικά σημαντική σχέση ανάμεσα στις μεταβλητές, σε 99% επίπεδο εμπιστοσύνης.

Ο συντελεστής προσδιορισμού R^2 δείχνει ότι το προσαρμοσμένο μοντέλο εξηγεί 12,6212% της μεταβλητότητας στη μεταβλητή Est28. Ο διορθωμένος συντελεστής R^2, που είναι πιο κατάλληλος για σύγκριση μοντέλων πολλαπλής παλινδρόμησης, είναι 11,9165%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 2,09073. Η τιμή αυτή μπορεί να χρησιμοποιηθεί για τη δημιουργία

διαστημάτων πρόβλεψης για νέες παρατηρήσεις. Το μέσο απόλυτο σφάλμα είναι 1,71213 και αποτελεί τη μέση τιμή των καταλοίπων. Το ***Durbin-Watson (DW) statistic*** ελέγχει τα κατάλοιπα για το αν υπάρχει αυτοσυσχέτιση. Επειδή το ***p-value*** είναι μικρότερο από 0,05, υπάρχει ένδειξη αυτοσυσχέτισης.

Επειδή η υψηλότερη τιμή *p*-value των ανεξάρτητων μεταβλητών, που είναι 0,0000 και ανήκει στη μεταβλητή Blaine, είναι μικρότερη από 0,01, συμπεραίνουμε ότι η μεταβλητή είναι σημαντική σε 99% επίπεδο εμπιστοσύνης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής Est28 σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές. Σε σχέση με το αντίστοιχο διάγραμμα της πολλαπλής παλινδρόμησης που περιλαμβάνει όλες τις μεταβλητές εμφανίζεται χειρότερο. Δηλαδή, η ευθεία γραμμή δεν προσαρμόζεται ιδιαίτερα καλά σε *όλα* τα δεδομένα.



<u>Διάγραμμα 6.2</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est28-Προς τα Εμπρός Επιλογή, CEM II 42,5, MT1

- ***<u>Εξαρτημένη Μεταβλητή log(Est2) –Όλες οι Μεταβλητές</u>***

Το μοντέλο παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής log(Est2) και των ανεξάρτητων μεταβλητών είναι το εξής:

log(Est2) = 2,09804 + 0,0125107*SiO2 - 0,102093*Al2O3 + 0,000373577*Blaine - 0,0032725*IR - 0,312761*LOG(LOI)

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων φαίνονται στον Πίνακα 6.4.

<u>Πίνακας 6.4</u>: Πολλαπλή Παλινδρόμηση για log(Est2)-Όλες οι Μεταβλητές, CEM II 42,5, MT1

Εξαρτημένη μεταβλητή: LOG(Est2)

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	2,09804	0,44876	4,6752	0,0000
SiO2	0,0125107	0,00998351	1,25314	0,2126
Al2O3	-0,102093	0,0265056	-3,85177	0,0002
Blaine	0,000373577	0,0000712738	5,24144	0,0000
IR	-0,0032725	0,00593179	-0,551688	0,5822
LOG(LOI)	-0,312761	0,0772313	-4,04967	0,0001

Ανάλυση Διακύμανσης

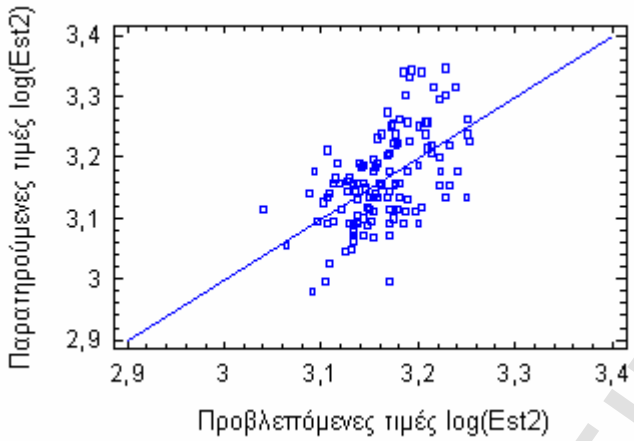
	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	0,216305	5	0,043261	10,16	0,0000
Κατάλοιπα	0,5109	120	0,0042575		
Σύνολο	0,727205	125			

R-2 = 29,7447 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 26,8174τοις εκατό
Τυπικό Σφάλμα = 0,0652495
Μέσο απόλυτο σφάλμα = 0,0524657
Durbin-Watson statistic = 1,12447 (P=0,0000)

Από την ανάλυση διακύμανσης προκύπτει ότι το *p-value* είναι μικρότερο από 0,01, επομένως υπάρχει μια στατιστικά σημαντική σχέση ανάμεσα στις μεταβλητές, σε 99% επίπεδο εμπιστοσύνης.

Ο συντελεστής προσδιορισμού R^2 εξηγεί 29,7447% της μεταβλητότητας στη μεταβλητή Est28. Ο διορθωμένος συντελεστής R^2 είναι 26,8174%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 0,0652495. Το μέσο απόλυτο σφάλμα είναι 0,0524657 και αποτελεί τη μέση τιμή των καταλοίπων. Το *Durbin-Watson statistic* έχει *p-value* μικρότερη από 0,05, άρα υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής log(Est2) σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές.



<u>Διάγραμμα 6.3</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής log(Est2)-Όλες οι Μεταβλητές, CEM II 42,5, MT1

- ***<u>Εξαρτημένη Μεταβλητή log(Est2) – Προς τα Εμπρός Επιλογή</u>***

Το μοντέλο παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής log(Est2) και των ανεξάρτητων μεταβλητών που προκύπτει από τη βηματική μέθοδο επιλογής των μεταβλητών είναι το εξής:

log(Est2) = 2,39995 - 0,0970642*Al2O3 + 0,000353906*Blaine – 0,275916*log(LOI)

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων φαίνονται στον Πίνακα 6.5.

<u>Πίνακας 6.5</u>: Πολλαπλή Παλινδρόμηση για log(Est2)-Προς τα Εμπρός Επιλογή, CEM II 42,5, MT1

Εξαρτημένη μεταβλητή: LOG(Est2)

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	2,39995	0,371871	6,4537	0,0000
Al2O3	-0,0970642	0,0260902	-3,72033	0,0003
Blaine	0,000353906	0,0000693917	5,10012	0,0000
LOG(LOI)	-0,275916	0,0710623	-3,88273	0,0002

Ανάλυση Διακύμανσης

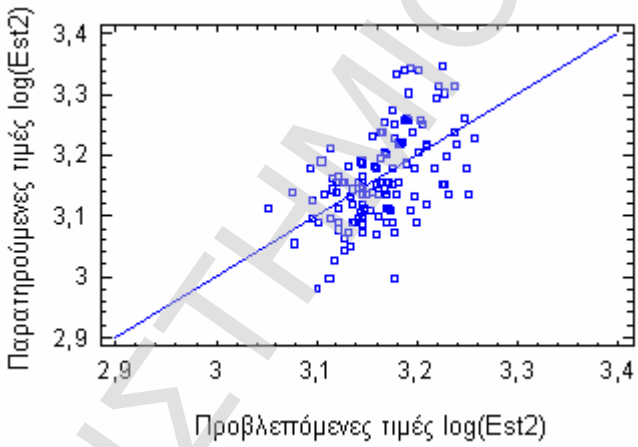
	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	0,209535	3	0,0698449	16,46	0,0000
Κατάλοιπα	0,51767	122	0,0042432		
Σύνολο	0,727205	125			

R-2 = 28,8137 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 27,0632 τοις εκατό
Τυπικό σφάλμα = 0,0651398
Μέσο απόλυτο σφάλμα = 0,0524958
Durbin-Watson statistic = 1,10136 (P=0,0000)

Από την ανάλυση διακύμανσης προκύπτει ότι το *p-value* είναι μικρότερο από 0,01, επομένως υπάρχει μια στατιστικά σημαντική σχέση ανάμεσα στις μεταβλητές, σε 99% επίπεδο εμπιστοσύνης.

Ο συντελεστής προσδιορισμού R^2 εξηγεί 28,8137% της μεταβλητότητας στη μεταβλητή Est28. Ο διορθωμένος συντελεστής R^2 είναι 27,0632%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 0,0651398. Το μέσο απόλυτο σφάλμα είναι 0,0524958 και αποτελεί τη μέση τιμή των καταλοίπων. Το *Durbin-Watson statistic* έχει *p-value* μικρότερη από 0,05, άρα υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής log(Est2) σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές.



<u>Διάγραμμα 6.4</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής log(Est2)-Προς τα Εμπρός Επιλογή, CEM II 42,5, MT1

Στο μοντέλο προκρίνονται τρεις ανεξάρτητες μεταβλητές. Από τον πίνακα συσχετίσεων προέκυψε ότι, όντως, αυτές οι τρεις μεταβλητές συσχετίζονται με την log(Est2). Κάτι άλλο, όμως, που πρέπει να ελεγχθεί είναι αν αυτές οι τρεις ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους. Φάνηκε από την προηγούμενη ανάλυση συσχέτισης ότι οι μεταβλητές Al_2O_3 και log(LOI) πιθανότατα συσχετίζονται. Ωστόσο, η συσχέτιση μεταξύ τους ελέγχεται και μέσω των εκτιμητριών των συντελεστών παλινδρόμησης, που παρέχει το πρόγραμμα μέσα στις επιλογές της πολλαπλής παλινδρόμησης. Τα αποτελέσματα παρατίθενται στον Πίνακα 6.6.

Πίνακας 6.6: Πίνακας Συσχετίσεων των Εκτιμητριών των Συντελεστών Παλινδρόμησης

	CONSTANT	Al2O3	Blaine	LOG(LOI)
CONSTANT	1,0000	-0,4235	-0,8613	-0,2978
Al2O3	-0,4235	1,0000	-0,0402	0,1778
Blaine	-0,8613	-0,0402	1,0000	0,0012
LOG(LOI)	-0,2978	0,1778	0,0012	1,0000

Όπως προκύπτει από τον παραπάνω πίνακα, δεν υπάρχουν συσχετίσεις με απόλυτες τιμές μεγαλύτερες από 0,5 (εξαιρουμένων των σταθερών όρων-constant, οι οποίοι δεν παίζουν ρόλο). Για το λόγο αυτό, δεν κρίνεται απαραίτητη η εφαρμογή της μεθόδου ραχοειδούς παλινδρόμησης, αφού η μέθοδος αυτή είναι σχεδιασμένη να παρέχει εκτιμήσεις των συντελεστών παλινδρόμησης, όταν οι ανεξάρτητες μεταβλητές συσχετίζονται *ισχυρά* μεταξύ τους.

6.2.3. Απλή Παλινδρόμηση

Από τον πίνακα των συσχετίσεων (Πίνακας 6.1) προέκυψε ότι η μεταβλητή **Est28** συσχετίζεται ισχυρά με τις μεταβλητές **log(Est2)** και **Est7**. Για το λόγο αυτόν, αλλά και επειδή η πρόβλεψη των τελικών αντοχών του τσιμέντου στις 28 μέρες όταν είναι γνωστές οι αντοχές του στις 2 ημέρες και στις 7 ημέρες παίζει σημαντικό ρόλο στους ανθρώπους της τσιμεντοβιομηχανίας, επιχειρείται η δημιουργία δύο μοντέλων απλής παλινδρόμησης μεταξύ των παραπάνω ζευγών μεταβλητών.

- ***Απλή Παλινδρόμηση μεταξύ Est28 και log(Est2)***

Για να αποφασίσουμε ποια μορφή απλής παλινδρόμησης θα επιλέξουμε, εξετάζουμε διάφορα εναλλακτικά μοντέλα που προτείνει το Statgraphics, με κριτήριο ποιο από αυτά έχει τη μεγαλύτερη τιμή σε R^2. Τα εναλλακτικά μοντέλα απεικονίζονται στον Πίνακα 6.7.

Πίνακας 6.7: Σύγκριση Εναλλακτικών Μοντέλων Απλής Παλινδρόμησης

Μοντέλο	Συσχέτιση	R-2
Αντίστροφο ως προς X	-0,7596	57,70%
Λογαριθμικό ως προς X	0,7591	57,62%
Καμπύλη S	-0,7589	57,59%
Τετραγωνικής ρίζας του X	0,7587	57,56%
Γραμμικό	0,7583	57,50%
Πολλαπλασιαστικό	0,7580	57,45%
Διπλής αντιστροφής	0,7577	57,41%
Τετραγωνικής ρίζας του Y	0,7577	57,41%
Εκθετικό	0,7569	57,29%
Αντίστροφο ως προς Y	-0,7551	57,02%
Λογιστικό	<no fit>	
Log probit	<no fit>	

Το αντίστροφο ως προς τη μεταβλητή *X* μοντέλο φαίνεται ότι έχει τη μεγαλύτερη ερμηνευτικότητα. Η μορφή του μοντέλου αυτού είναι: *Y=a+b/X*. Συνεπώς, επιλέγεται αυτό για την ανάλυση που ακολουθεί. Στον Πίνακα 6.8 αναγράφονται τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου, ενώ η εξίσωση της απλής παλινδρόμησης είναι η ακόλουθη:

Est28 = 122,989 - 223,43/log(Est2)

<u>Πίνακας 6.8</u>: Απλή Παλινδρόμηση μεταξύ Est28 και log(Est2), CEM II 42,5, MT1

Αντίστροφο ως προς X μοντέλο: Y = a + b/X

Εξαρτημένη μεταβλητή: Est28
Ανεξάρτητη μεταβλητή: LOG(Est2)

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
Σταθερός όρος	122,989	5,43383	22,634	0,0000
Κλίση	-223,43	17,1779	-13,0068	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	357,952	1	357,952	169,18	0,0000
Κατάλοιπα	262,365	124	2,11584		
Σύνολο	620,317	125			

Συντελεστής συσχέτισης = -0,759636
R-2 = 57,7047 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 57,3636 τοις εκατό
Τυπικό σφάλμα = 1,45459
Μέσο απόλυτο σφάλμα = 1,1737
Durbin-Watson statistic = 1,30714 (P=0,0000)

Το *p-value* στον πίνακα της ANOVA είναι μικρότερο από 0,01, συνεπώς υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Est28 και της log(Est2) στο 99% επίπεδο εμπιστοσύνης.

Το R^2 είναι 57,7047%, που δηλώνει ότι το μοντέλο που προσαρμόζεται στα δεδομένα εξηγεί κατά αυτό το ποσοστό τη μεταβλητότητα του Est28. Ο συντελεστής συσχέτισης είναι -0,759636, που δείχνει μια σχετικά ισχυρή σχέση μεταξύ των μεταβλητών. Το τυπικό σφάλμα της εκτίμησης είναι 1,45159 και ισούται με την τυπική απόκλιση των καταλοίπων. Από το *Durbin-Watson (DW) statistic* προκύπτει ότι το *p-value* είναι μικρότερο από 0,05, το οποίο φανερώνει την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

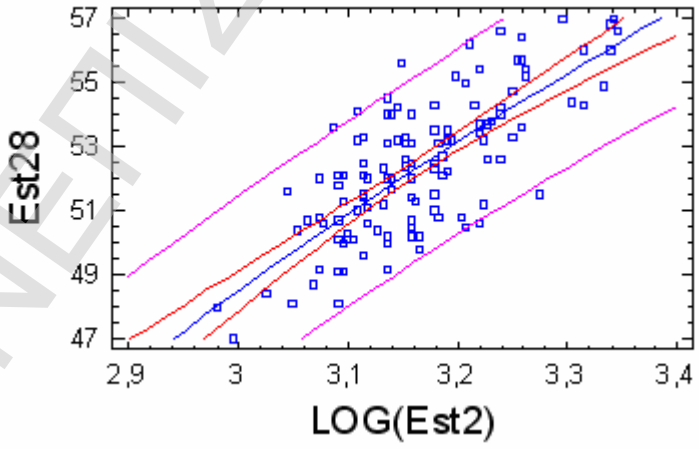
Για να ελέγξουμε αν το προτεινόμενο μοντέλο επαρκεί για να περιγράψει τις τιμές των δεδομένων, εκτελούμε ένα τεστ για έλεγχο της έλλειψης προσαρμογής αυτού, το οποίο καλείται ***Lack-of-Fit test***. Τα αποτελέσματα του τεστ αυτού φαίνονται στον Πίνακα 6.9.

Πίνακας 6.9: Έλεγχος Έλλειψης Προσαρμογής στην Απλή Παλινδρόμηση μεταξύ Est28 και log(Est2), CEM II 42,5, MT1

Ανάλυση Διακύμανσης με Lack-of-Fit τεστ

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	**P-Value**
Μοντέλο	357,952	1	357,952	169,18	0,0000
Κατάλοιπα	262,365	124	2,11584		
Lack-of-Fit	127,905	53	2,41331	1,27	**0,1692**
Καθαρό σφάλμα	134,459	71	1,89379		
Σύνολο	620,317	125			

Το τεστ για τον έλεγχο έλλειψης προσαρμογής είναι σχεδιασμένο για να καθορίζει εάν το επιλεγμένο μοντέλο επαρκεί για να περιγράψει τα παρατηρούμενα δεδομένα, ή εάν απαιτείται ένα πιο πολύπλοκο μοντέλο. Το τεστ πραγματοποιείται συγκρίνοντας τη μεταβλητότητα των καταλοίπων στο προτεινόμενο μοντέλο με τη μεταβλητότητα μεταξύ των παρατηρήσεων της ανεξάρτητης μεταβλητής *X*. Καθώς το ***p-value*** για το lack-of-fit τεστ στον πίνακα της ANOVA είναι μεγαλύτερο από 0,10, το μοντέλο φαίνεται να επαρκεί για τα παρατηρούμενα δεδομένα. Το διάγραμμα της σχέσης μεταξύ των μεταβλητών φαίνεται στο παρακάτω Διάγραμμα 6.5. Τα εσωτερικά όρια (κόκκινες γραμμές) δηλώνουν τα 95% διαστήματα εμπιστοσύνης για το μέσο Est28 των παρατηρήσεων, για δεδομένες τιμές log(Est2). Τα εξωτερικά όρια (ροζ γραμμές) δηλώνουν τα 95% διαστήματα *πρόβλεψης* για τις νέες παρατηρήσεις.



Διάγραμμα 6.5: Διάγραμμα Απλής Παλινδρόμησης των μεταβλητών Est28 και log(Est2), CEM II 42,5, MT1

- ***Απλή Παλινδρόμηση μεταξύ Est28 και Est7***

Το μοντέλο παλινδρόμησης που εμφανίζεται ως το καλύτερο για εφαρμογή στη συγκεκριμένη περίπτωση είναι το γραμμικό, όπως φαίνεται και στον Πίνακα 6.10.

Πίνακας 6.10: Σύγκριση Εναλλακτικών Μοντέλων Απλής Παλινδρόμησης

Μοντέλο	Συσχέτιση	R-2
Γραμμικό	0,7583	57,50%
Τετραγωνικής ρίζας του X	0,7580	57,45%
Λογαριθμικό ως προς X	0,7574	57,36%
Τετραγωνικής ρίζας του Y	0,7569	57,29%
Αντίστροφο ως προς X	-0,7554	57,06%
Εκθετικό	0,7554	57,06%
Πολλαπλασιαστικό	0,7552	57,03%
Καμπύλη S	-0,7539	56,83%
Αντίστροφο ως προς Y	-0,7520	56,55%
Διπλής αντιστροφής	0,7519	56,54%
Λογιστικό	<no fit>	
Log probit	<no fit>	

Στον Πίνακα 6.11 αναγράφονται τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου, ενώ η εξίσωση της απλής παλινδρόμησης είναι η ακόλουθη:

Est28 = 23,7877 + 0,752171*Est7

Πίνακας 6.11: Απλή Παλινδρόμηση μεταξύ Est28 και Est7, CEM II 42,5, MT1

Γραμμικό μοντέλο: Y = a + b*X

Εξαρτημένη μεταβλητή: Est28
Ανεξάρτητη μεταβλητή: Est7

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
Σταθερός όρος	23,7877	2,20754	10,7757	0,0000
Κλίση	0,752171	0,0580688	12,9531	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	356,698	1	356,698	167,78	0,0000
Κατάλοιπα	263,618	124	2,12595		
Σύνολο	620,317	125			

Συντελεστής συσχέτισης = 0,758305
R-2 = 57,5026 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 57,1599 τοις εκατό
Τυπικό σφάλμα = 1,45807
Μέσο απόλυτο σφάλμα = 1,17707
Durbin-Watson statistic = 1,11715 (P=0,0000)

Το *p-value* στον πίνακα της ANOVA είναι μικρότερο από 0,01, συνεπώς υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Est28 και της Est7 στο 99% επίπεδο εμπιστοσύνης.

Το R^2 είναι 57,5026%, που δηλώνει ότι το μοντέλο που προσαρμόζεται στα δεδομένα εξηγεί κατά αυτό το ποσοστό τη μεταβλητότητα του Est28. Ο συντελεστής συσχέτισης είναι 0,758305, που δείχνει μια σχετικά ισχυρή σχέση μεταξύ των μεταβλητών. Το τυπικό σφάλμα της εκτίμησης είναι 1,45807 και ισούται με την τυπική απόκλιση των καταλοίπων. Από το *Durbin-Watson (DW) statistic* προκύπτει ότι το *p-value* είναι μικρότερο από 0,05, το οποίο φανερώνει την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

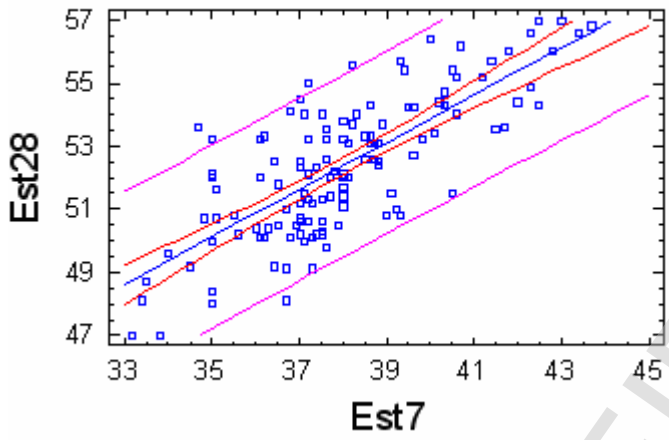
Για να ελέγξουμε αν το προτεινόμενο μοντέλο επαρκεί για να περιγράψει τις τιμές των δεδομένων, εκτελούμε το *Lack-of-Fit test*. Τα αποτελέσματα του τεστ αυτού φαίνονται στον Πίνακα 6.12.

<u>Πίνακας 6.12</u>: Έλεγχος Έλλειψης Προσαρμογής στην Απλή Παλινδρόμηση μεταξύ Est28 και Est7, CEM II 42,5, MT1

ΑνάλυσηΔιακύμανσης με Lack-of-Fit τεστ

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	**P-Value**
Μοντέλο	356,698	1	356,698	167,78	0,0000
Κατάλοιπα	263,618	124	2,12595		
Lack-of-Fit	116,721	65	1,79571	0,72	**0,9005**
Καθ.σφάλμα	146,897	59	2,48978		
Σύνολο	620,317	125			

Καθώς το *p-value* για το lack-of-fit test στον πίνακα της ANOVA είναι μεγαλύτερο από 0,10, το μοντέλο επαρκεί για να περιγράψει τα δεδομένα. Το διάγραμμα της σχέσης μεταξύ των μεταβλητών φαίνεται στο παρακάτω Διάγραμμα 6.6.



<u>Διάγραμμα 6.6</u>: Διάγραμμα Απλής Παλινδρόμησης των μεταβλητών Est28 και Est7, CEM II 42,5, MT1

6.2.4. <u>Έλεγχος των Προϋποθέσεων της Παλινδρόμησης</u>

Ο έλεγχος των προϋποθέσεων της πολλαπλής (με την προς τα εμπρός επιλογή) και των δύο απλών παλινδρομήσεων γίνεται κυρίως ως προς τα κατάλοιπα, και συγκεκριμένα για το αν αυτά είναι κανονικά κατανεμημένα, αν έχουν μέση τιμή ίση με μηδέν, αν έχουν σταθερή διασπορά και αν αυτοσυσχετίζονται. Ο έλεγχος για ύπαρξη πολυσυγγραμικότητας και οι διορθωτικές πράξεις έχουν μελετηθεί σε προηγούμενη ενότητα, κατά την ανάλυση της πολλαπλής παλινδρόμησης.

- Ο έλεγχος σχετικά με το αν τα κατάλοιπα ακολουθούν **κανονική κατανομή** γίνεται μέσω της εξαρτημένης μεταβλητής *Y*. Αυτό, διότι γνωρίζουμε από τη θεωρία της παλινδρόμησης ότι οι εξαρτημένες μεταβλητές *X* θεωρούνται *σταθερές* ποσότητες. Συνεπώς, οποιοσδήποτε τυχαίος παράγοντας επηρεάζει την *Y* οφείλεται αποκλειστικά στα κατάλοιπα. Από τη στατιστική ανάλυση που προηγήθηκε στο αμέσως προηγούμενο κεφάλαιο, προέκυψε ότι οι μεταβλητές Est28 και log(Est2) ακολουθούν κανονική κατανομή με βάση τον έλεγχο ασυμμετρίας και κύρτωσης, ενώ η μεταβλητή log(Est2) δεν φαίνεται να ακολουθεί κανονική κατανομή με βάση τα X^2 και *Shapiro-Wilks statistics*.

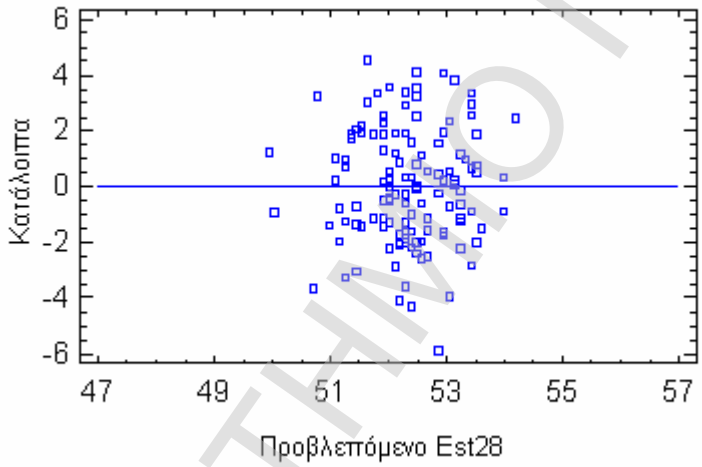
- Όσον αφορά τη **μέση τιμή** των καταλοίπων, που πρέπει να είναι μηδενική, αυτό μπορεί αν ελεγχθεί με δύο τρόπους: Πρώτον, από την τιμή του MAE (μέσο απόλυτο σφάλμα) που δίνεται από την ανάλυση διακύμανσης που συνοδεύει τα αποτελέσματα παλινδρόμησης. Δεύτερον, από τα διαγράμματα των καταλοίπων ως προς τις προβλεπόμενες τιμές *Y*. Από το τελευταίο διάγραμμα μπορεί να γίνει και έλεγχος για το αν η **διασπορά** των καταλοίπων παραμένει σταθερή, ή αν εμφανίζεται πρόβλημα ετεροσκεδαστικότητας. Τα αποτελέσματα για καθεμία από τις δύο πολλαπλές παλινδρομήσεις, αλλά και για την απλή παλινδρόμηση είναι τα εξής:

§ ***Πολλαπλή Παλινδρόμηση Est28 – Προς τα Εμπρός Επιλογή***

Το MAE ισούται με 1,71213, το οποίο είναι διάφορο του μηδενός. Από το διάγραμμα, όμως, των καταλοίπων ως προς τις προβλεπόμενες τιμές Est28 δεν φαίνεται να υπάρχει σημαντική απόκλιση από την υπόθεση ότι το μέσο σφάλμα έχει μηδενική τιμή. Αυτό φαίνεται από το γεγονός ότι οι τιμές είναι εξίσου διεσπαρμένες, τόσο πάνω από την γραμμή στον οριζόντιο άξονα, όσο και κάτω από αυτήν.

Επίσης, σχετικά με τη διασπορά, φαίνεται από το ίδιο διάγραμμα ότι αυτή φαίνεται να έχει παντού το ίδιο εύρος, καθώς προχωράμε δεξιά προς τον οριζόντιο άξονα.

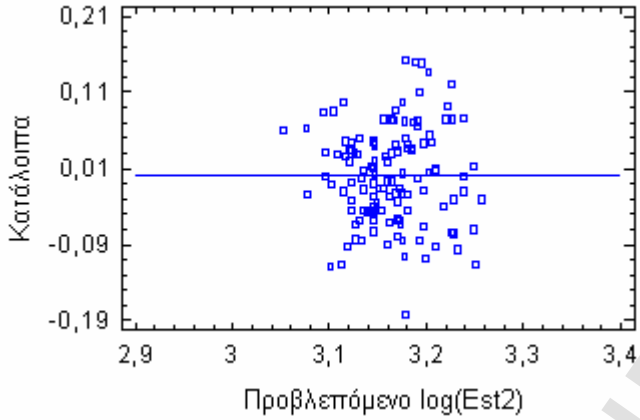
Το Διάγραμμα 6.7 απεικονίζει το διάγραμμα καταλοίπων ως προς τις προβλεπόμενες τιμές *Y*, για τον έλεγχο των προϋποθέσεων μηδενικής μέσης τιμής των σφαλμάτων και σταθερής διασποράς αυτών.



Διάγραμμα 6.7: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Πολλαπλή Παλινδρόμηση, CEM II 42,5, MT1

§ ***Πολλαπλή Παλινδρόμηση log(Est2) – Προς τα Εμπρός Επιλογή***

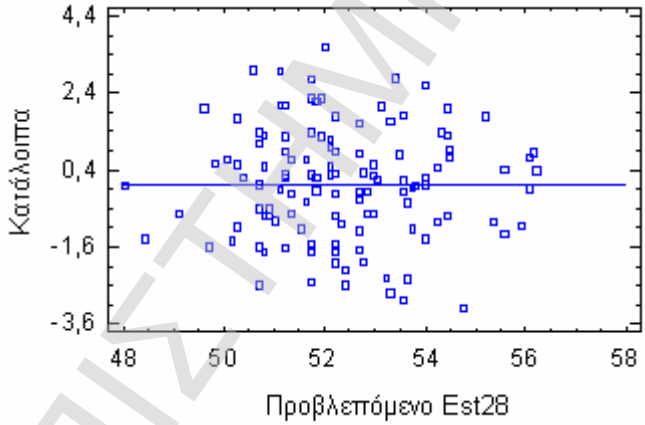
Το MAE ισούται με 0,0524958, το οποίο προσεγγίζει αρκετά την τιμή μηδέν. Από το διάγραμμα των καταλοίπων ως προς τις προβλεπόμενες τιμές log(Est2) φαίνεται επίσης ότι τα κατάλοιπα είναι ομοιόμορφα κατανεμημένα πάνω και κάτω από τη γραμμή του οριζόντιου άξονα. Η διασπορά φαίνεται επίσης να είναι ομοιόμορφη, χωρίς να σχηματίζει κάποιο ιδιαίτερο σχέδιο (pattern).



Διάγραμμα 6.8: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές log(Est2), Πολλαπλή Παλινδρόμηση, CEM II 42,5, MT1

§ ***Απλή Παλινδρόμηση Est28 και log(Est2)***

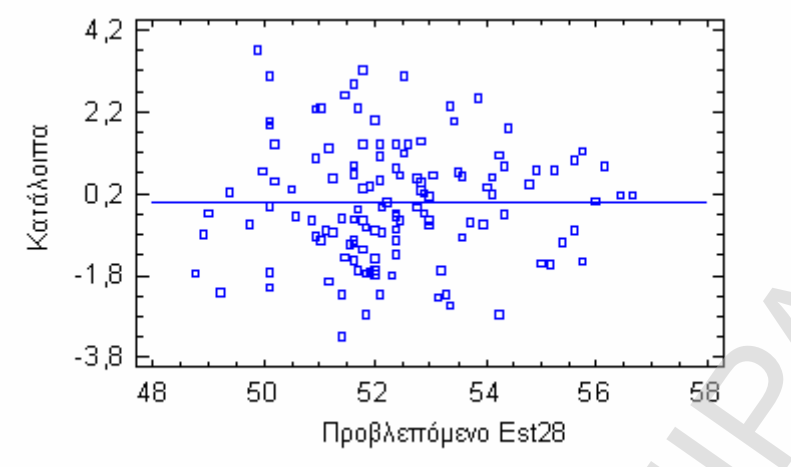
Σε αυτή την απλή παλινδρόμηση, το MAE είναι 1,1737, το οποίο δεν ισούται με την τιμή μηδέν. Από το διάγραμμα των καταλοίπων δεν φαίνεται κάποιο ιδιαίτερο πρόβλημα για την καταπάτηση αυτής της προϋπόθεσης. Η διασπορά φαίνεται να είναι ομοιόμορφα κατανεμημένη.



Διάγραμμα 6.9: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Απλή Παλινδρόμηση Est28 και log(Est2), CEM II 42,5, MT1

§ ***Απλή Παλινδρόμηση Est28 καιEst7***

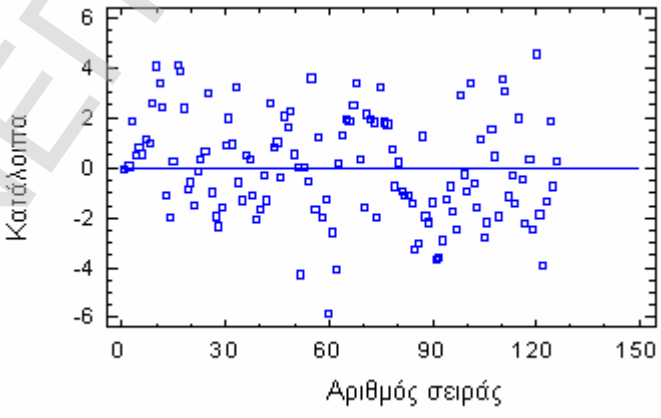
Στην απλή παλινδρόμηση, το MAE είναι 1,17707, το οποίο δεν ισούται με την τιμή μηδέν. Από το διάγραμμα των καταλοίπων δεν φαίνεται κάποιο ιδιαίτερο πρόβλημα, ενώ η διασπορά φαίνεται να είναι ομοιόμορφα κατανεμημένη.



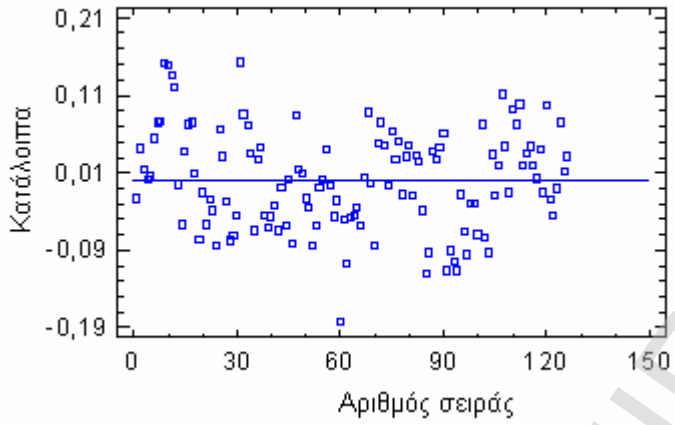
Διάγραμμα 6.10: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Απλή Παλινδρόμηση Est28 και Est7, CEM II 42,5, MT1

- Ο έλεγχος για την ύπαρξη **αυτοσυσχέτισης** μεταξύ των καταλοίπων γίνεται κατά κύριο λόγο με το *Durbin-Watson statistic*. Αν το *p-value* από το test αυτό είναι μικρότερο από 0,05, τότε υπάρχει σημαντική ένδειξη αυτοσυσχέτισης με lag 1. Ένας άλλος τρόπος ελέγχου ύπαρξης αυτοσυσχέτισης είναι το διάγραμμα των καταλοίπων ως προς τον αριθμό σειράς (row number). Βέβαια, ο τελευταίος έλεγχος γίνεται οπτικά, γι' αυτό και δεν θεωρείται αρκετά έγκυρος όσο ο πρώτος τρόπος ελέγχου. Τα *p-values* των Durbin-Watson tests και για τις τέσσερις περιπτώσεις παλινδρόμησης (δύο πολλαπλές και δύο απλές παλινδρομήσεις) είναι μικρότερα από 0,05. Άρα, υπάρχει αυτοσυσχέτιση μεταξύ των καταλοίπων. Αυτό φαίνεται και από τα διαγράμματα, τα οποία παρουσιάζονται παρακάτω.

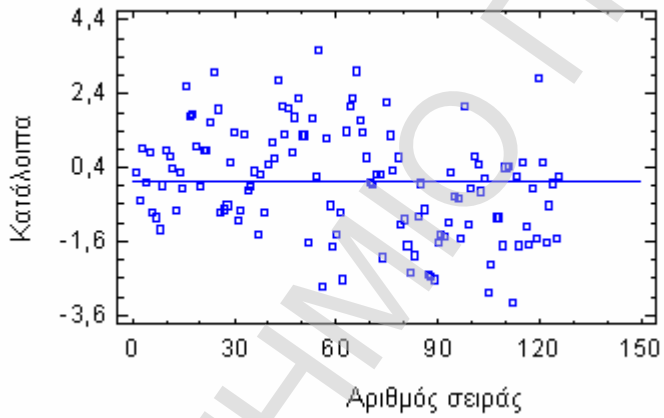
  Τα Διαγράμματα 6.11 και 6.12 παριστούν τα διαγράμματα καταλοίπων ως προς τον αριθμό σειράς της πολλαπλής παλινδρόμησης με εξαρτημένη μεταβλητή την Est28 και τη log(Est2), αντίστοιχα, και τα Διαγράμματα 6.13 και 6.14 το ίδιο διάγραμμα, για την περίπτωση της απλής παλινδρόμησης μεταξύ των μεταβλητών Est28 και logEst2 ή Est7.



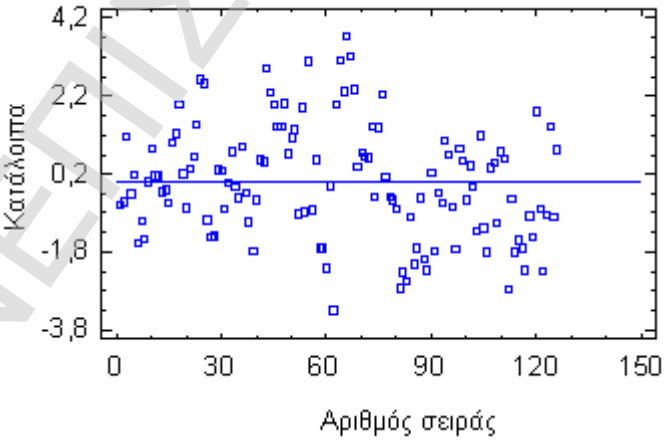
Διάγραμμα 6.11: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Πολλαπλή Παλινδρόμηση της Est28, CEM II 42,5, MT1



Διάγραμμα 6.12: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Πολλαπλή Παλινδρόμηση της log(Est2), CEM II 42,5, MT1



Διάγραμμα 6.13: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Απλή Παλινδρόμηση μεταξύ των Est28 και log(Est2), CEM II 42,5, MT1



Διάγραμμα 6.14: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Απλή Παλινδρόμηση μεταξύ των Est28 και Est7, CEM II 42,5, MT1

6.3. Σύνθετο Τσιμέντο Portland CEM II 42,5 – Μύλος Παραγωγής 4

Στην περίπτωση αυτή, ακολουθεί διαδικασία παρόμοια με αυτή που πραγματοποιήθηκε για την περίπτωση παραγωγής του ίδιου τύπου τσιμέντου στο μύλο 1. Τα αποτελέσματα είναι τα ακόλουθα:

6.3.1. Έλεγχος Συσχετίσεων μεταξύ των Μεταβλητών

Παρακάτω παρουσιάζεται ο πίνακας συσχετίσεων κατά Spearman, αφού δεν είναι όλες οι μεταβλητές κανονικά κατανεμημένες. Συγκεκριμένα, οι μεταβλητές SiO_2 και IR δεν ακολουθούν την κανονική κατανομή.

Πίνακας 6.13: Πίνακας Συσχετίσεων των Μεταβλητών κατά Spearman

	SiO_2	***Al_2O_3***	***Blaine***	***IR***	***LOI***	***Est2***	***Est7***	***Est28***
SiO_2		**0,8538** (121) **0,0000**	**0,3760** (121) **0,0000**	**0,7451** (121) **0,0000**	-0,1175 (121) 0,1981	0,0566 (121) 0,5350	0,0945 (121) 0,3008	0,1288 (121) 0,1584
Al_2O_3	**0,8538** (121) **0,0000**		**0,4254** (121) **0,0000**	**0,5880** (121) **0,0000**	-0,1284 (121) 0,1596	0,1338 (121) 0,1429	0,1631 (121) 0,0740	**0,2649** (121) **0,0037**
Blaine	**0,3760** (121) **0,0000**	**0,4254** (121) **0,0000**		**0,2334** (121) **0,0106**	0,0982 (121) 0,2819	0,1510 (121) 0,0980	0,1526 (121) 0,0946	0,1673 (121) 0,0669
IR	**0,7451** (121) **0,0000**	**0,5880** (121) **0,0000**	**0,2334** (121) **0,0106**		**0,2632** (121) **0,0039**	-0,0315 (121) 0,7299	0,0208 (121) 0,8196	0,0075 (121) 0,9344
LOI	-0,1175 (121) 0,1981	-0,1284 (121) 0,1596	0,0982 (121) 0,2819	**0,2632** (121) **0,0039**		-0,0859 (121) 0,3468	-0,0456 (121) 0,6171	-0,0812 (121) 0,3736
Est2	0,0566 (121) 0,5350	0,1338 (121) 0,1429	0,1510 (121) 0,0980	-0,0315 (121) 0,7299	-0,0859 (121) 0,3468		**0,5838** (121) **0,0000**	**0,4598** (121) **0,0000**
Est7	0,0945 (121) 0,3008	0,1631 (121) 0,0740	0,1526 (121) 0,0946	0,0208 (121) 0,8196	-0,0456 (121) 0,6171	**0,5838** (121) **0,0000**		**0,5929** (121) **0,0000**
Est28	0,1288 (121) 0,1584	**0,2649** (121) **0,0037**	0,1673 (121) 0,0669	0,0075 (121) 0,9344	-0,0812 (121) 0,3736	**0,4598** (121) **0,0000**	**0,5929** (121) **0,0000**	

Μεταβλητές με *p-values* μικρότερες από 0,05 φανερώνουν στατιστικά σημαντικές μη-μηδενικές συσχετίσεις, σε 95 % επίπεδο εμπιστοσύνης. Τα ζεύγη των μεταβλητών που έχουν *p-values* κάτω από 0,05 είναι τα ακόλουθα:

SiO2 και **Al2O3**
SiO2 και **Blaine**
SiO2 και **IR**
Al2O3 και **Blaine**
Al2O3 και **IR**
Al2O3 και **Est28**
Blaine και **IR**
IR και **LOI**

Est2 και **Est7**
Est2 και **Est28**
Est7 και **Est28**

Πολλές ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους, συνεπώς υπάρχει πρόβλημα πολυσυγγραμικότητας. Γι' αυτό, στο μοντέλο παλινδρόμησης θα εφαρμοστεί και η μέθοδος της προς τα εμπρός επιλογής.

Η εξαρτημένη μεταβλητή Est28 φαίνεται από τον παραπάνω πίνακα ότι συσχετίζεται με τις μεταβλητές Al_2O_3, Est2 και Est7. Άρα, στην πολλαπλή παλινδρόμηση περιμένουμε ότι πιθανότατα η μεταβλητή Al_2O_3 θα είναι παρούσα, ενώ μπορούμε να εφαρμόσουμε ένα μοντέλο απλής παλινδρόμησης με τη μεταβλητή Est2 και ένα άλλο με την Est7.

Όσον αφορά την εξαρτημένη μεταβλητή Est2, παρατηρούμε ότι δεν συσχετίζεται με καμία ανεξάρτητη μεταβλητή. Συνεπώς, κατά πάσα πιθανότητα, δεν θα προκύψει κάποιο μοντέλο παλινδρόμησης με τη μέθοδο της προς τα εμπρός επιλογής.

6.3.2. Πολλαπλή Παλινδρόμηση

- ***Εξαρτημένη Μεταβλητή Est28 - Όλες οι Μεταβλητές***

Το μοντέλο παλινδρόμησης που προκύπτει είναι το εξής:

Est28 = 48,9894 - 1,54905*SiO2 + 6,27521*Al2O3 + 0,00321422*Blaine - 0,026708*IR - 1,26153*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.14.

Πίνακας 6.14: Πολλαπλή Παλινδρόμηση για Est28-Όλες οι Μεταβλητές, CEM II 42,5, MT4

Εξαρτημένη μεταβλητή: Est28

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	48,9894	11,0913	4,41691	0,0000
SiO2	-1,54905	0,626298	-2,47334	0,0148
Al2O3	6,27521	1,63024	3,84926	0,0002
Blaine	0,00321422	0,00225612	1,42467	0,1570
IR	-0,026708	0,272009	-0,0981879	0,9220
LOI	-1,26153	0,64572	-1,95368	0,0532

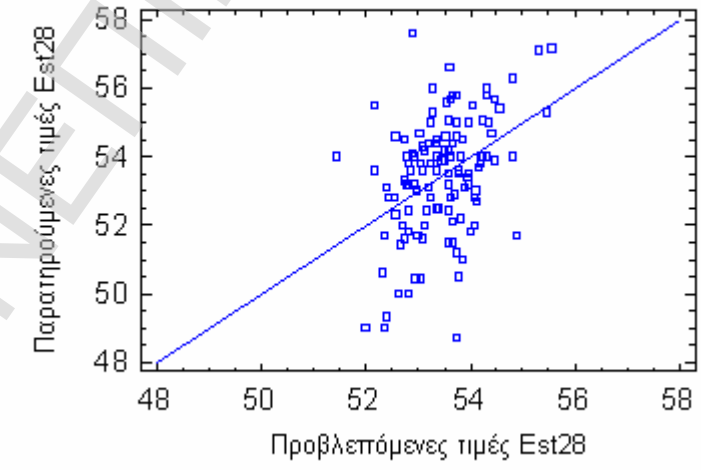
Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	62,3353	5	12,4671	3,40	0,0025
Κατάλοιπα	306,805	115	2,66787		
Σύνολο	369,14	120			

R-2 = 16,8866 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 13,273 τοις εκατό
Τυπικό σφάλμα = 1,63336
Μέσο απόλυτο σφάλμα = 1,26916
Durbin-Watson statistic = 1,47875 (P=0,0019)

Ο συντελεστής προσδιορισμού R^2 δείχνει ότι το προσαρμοσμένο μοντέλο εξηγεί 16,8866% της μεταβλητότητας στη μεταβλητή Est28. Ο διορθωμένος συντελεστής R^2, που είναι πιο κατάλληλος για σύγκριση μοντέλων πολλαπλής παλινδρόμησης, είναι 13,273%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 1,63336. Το μέσο απόλυτο σφάλμα είναι 1,26916 και αποτελεί τη μέση τιμή των καταλοίπων. Το Durbin-Watson statistic ελέγχει τα κατάλοιπα για το αν υπάρχει αυτοσυσχέτιση. Επειδή το *p-value* είναι μικρότερο από 0,05, υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής Est28 σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές. Όσο πιο κοντά τα δεδομένα "πέφτουν" στη διαγώνια γραμμή, τόσο καλύτερο είναι το μοντέλο στην πρόβλεψη των παρατηρούμενων δεδομένων. Στη συγκεκριμένη περίπτωση οι παρατηρούμενες τιμές αποκλίνουν πολύ από την ευθεία γραμμή, άρα το μοντέλο δεν θεωρείται ιδιαίτερα καλό.



<u>Διάγραμμα 6.15</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est28-Όλες οι Μεταβλητές, CEM II 42,5, MT4

- ***Εξαρτημένη Μεταβλητή Est28 – Προς τα Εμπρός Επιλογή***

Το μοντέλο παλινδρόμησης που προκύπτει είναι το εξής:

Est28 = 58,3727 - 1,48953*SiO2 + 6,63408*Al2O3 - 1,1151*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.15.

Πίνακας 6.15: Πολλαπλή Παλινδρόμηση για Est28-Προς τα Εμπρός Επιλογή, CEM II 42,5, MT4

Εξαρτημένη μεταβλητή: Est28

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	58,3727	6,60915	8,8321	0,0000
SiO2	-1,48953	0,455797	-3,26797	0,0014
Al2O3	6,63408	1,5879	4,17789	0,0001
LOI	-1,1151	0,536566	-2,07821	0,0399

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	56,7419	3	18,914	7,08	0,0002
Κατάλοιπα	312,398	117	2,67007		
Σύνολο	369,14	120			

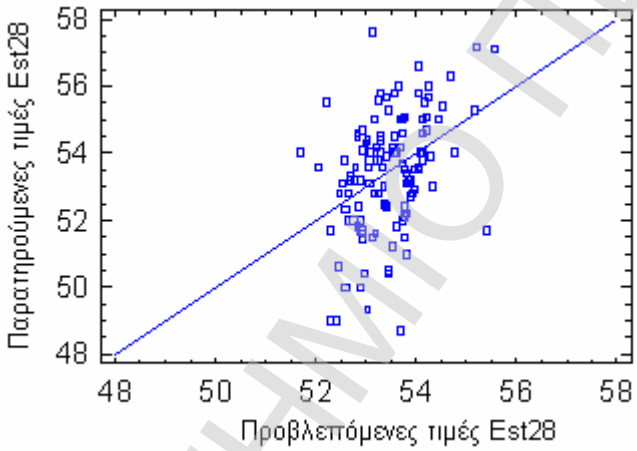
R-2 = 15,3714 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 13,2014 τοις εκατό
Τυπικό σφάλμα = 1,63403
Μέσο απόλυτο σφάλμα = 1,28002
Durbin-Watson statistic = 1,45919 (P=0,0013)

Από τις μεταβλητές που προκρίνονται στο μοντέλο μόνο η Al_2O_3 συσχετίζεται με την Est28. Οι SiO_2 και LOI πιθανότατα δεν θα έπρεπε να συμπεριλαμβάνονται. Αυτό μπορεί να οφείλεται στο ότι υπάρχει αυτοσυσχέτιση των καταλοίπων, καθώς το *p-value* στο *Durbin-Watson test* έχει τιμή μικρότερη του 0,05. Όταν εμφανίζεται αυτοσυσχέτιση στα κατάλοιπα, τα t statistics ορισμένων μεταβλητών εμφανίζονται να είναι πολύ μεγάλα, με αποτέλεσμα οι εκτιμώμενοι συντελεστές των μεταβλητών αυτών να φαίνονται στατιστικά σημαντικοί, ενώ στην πραγματικότητα δεν είναι. Επίσης, μία άλλη εξήγηση για το γεγονός ότι μπορεί να μην συμπεριλαμβάνονται στο μοντέλο οι μεταβλητές που πραγματικά επηρεάζουν την Est28, είναι η ύπαρξη πολυσυγγραμικότητας. Η πολυσυγγραμικότητα προκαλεί ακριβώς τα αντίθετα

αποτελέσματα από την αυτοσυσχέτιση, δηλαδή χαμηλές τιμές στα t tests, έτσι ώστε μερικές μεταβλητές που είναι στατιστικά σημαντικές να φαίνεται ότι δεν είναι.

Ο συντελεστής προσδιορισμού R^2 δείχνει ότι το προσαρμοσμένο μοντέλο εξηγεί 15,3714% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 είναι 13,2014%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 1,63403. Το μέσο απόλυτο σφάλμα είναι 1,28002 και αποτελεί τη μέση τιμή των καταλοίπων. Το ***Durbin-Watson (DW) statistic*** ελέγχει τα κατάλοιπα για το αν υπάρχει αυτοσυσχέτιση. Επειδή το *p-value* είναι μικρότερο από 0,05, υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής Est28 σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές. Στη συγκεκριμένη περίπτωση οι παρατηρούμενες τιμές αποκλίνουν πολύ από την ευθεία γραμμή, άρα το μοντέλο δεν θεωρείται ιδιαίτερα καλό.



<u>Διάγραμμα 6.16</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est28-Προς τα Εμπρός Επιλογή, CEM II 42,5, MT4

Επειδή οι μεταβλητές συσχετίζονται μεταξύ τους, με βάση τον προηγούμενο πίνακα συσχετίσεων, ελέγχουμε και τη συσχέτιση μεταξύ των εκτιμητριών των συντελεστών παλινδρόμησης:

<u>Πίνακας 6.16</u>: Πίνακας Συσχετίσεων των Εκτιμητριών των Συντελεστών Παλινδρόμησης

	CONSTANT	SiO2	Al2O3	LOI
CONSTANT	1,0000	-0,6630	0,1760	-0,3594
SiO2	-0,6630	1,0000	-0,8421	0,1149
Al2O3	0,1760	-0,8421	1,0000	-0,0615
LOI	-0,3594	0,1149	-0,0615	1,0000

Όπως φαίνεται από τον παραπάνω πίνακα, η συσχέτιση μεταξύ των μεταβλητών SiO_2 και Al_2O_3 είναι πολύ ισχυρή (μεγαλύτερη από 0,5). Για το λόγο αυτό, πρέπει να εφαρμοστεί μια πιο προχωρημένη μέθοδος παλινδρόμησης που επιλύει τα προβλήματα πολυσυγγραμικότητας, η οποία καλείται *ραχοειδής παλινδρόμηση*.

- ***Ραχοειδής Παλινδρόμηση***

Η ραχοειδής παλινδρόμηση είναι ειδικά σχεδιασμένη για να παρέχει εκτιμήσεις των συντελεστών παλινδρόμησης, όταν οι ανεξάρτητες μεταβλητές συσχετίζονται ισχυρά. Επιτρέποντας ένα μικρό ποσό μεροληψίας, η ακρίβεια των εκτιμητριών μπορούν να βελτιωθούν σημαντικά. Στη μέθοδο αυτή πρέπει να επιλέξουμε μία παράμετρο που θα εκτελεστεί η παλινδρόμηση, η οποία ονομάζεται *παράμετρος ραχοειδούς παλινδρόμησης*. Οι πληθωριστικοί παράγοντες διασποράς (Variance Inflation Factors-VIFs) μετρούν πόσο η διασπορά των εκτιμώμενων συντελεστών είναι "φουσκωμένη", σε σχέση με την περίπτωση όπου όλες οι ανεξάρτητες μεταβλητές είναι ασυσχέτιστες.

Καθώς η *παράμετρος ραχοειδούς παλινδρόμησης* μεγαλώνει από την τιμή 0,0, οι VIFs συχνά μειώνονται δραματικά στην αρχή, ενώ στη συνέχεια σχεδόν σταθεροποιούνται. Μία καλή τιμή για την *παράμετρο ραχοειδούς παλινδρόμησης* είναι η μικρότερη τιμή, μετά από την οποία οι VIFs αλλάζουν με πολύ αργό ρυθμό. Αυτό είναι μεν υποκειμενικό, ωστόσο υπάρχουν δύο τρόποι εντοπισμού της κατάλληλης τιμής της παραμέτρου αυτής. Ο πρώτος τρόπος αφορά στον έλεγχο των τιμών των προτυποποιημένων συντελεστών παλινδρόμησης. Η τιμή της παραμέτρου, μετά την οποία οι παραπάνω συντελεστές αλλάζουν τιμές με πολύ αργό ρυθμό, είναι η πλέον κατάλληλη προς χρήση. Ενδεικτικά παρατίθεται ο επόμενος πίνακας.

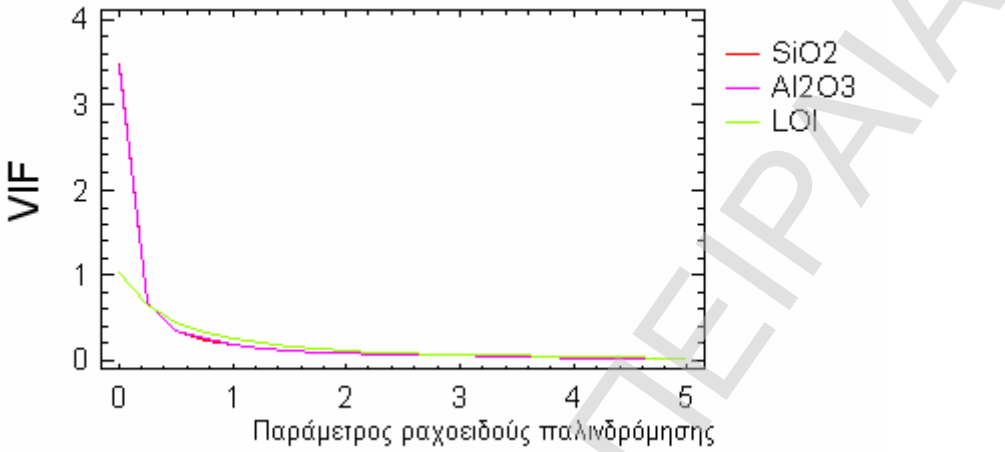
Πίνακας 6.17: Πίνακας Συντελεστών Παλινδρόμησης ως προς τις τιμές της Παραμέτρου Ραχοειδούς Παλινδρόμησης

Προτυποποιημένοι Συντελεστές Παλινδρόμησης

Ridge Parameter	**SiO2**	**Al2O3**	**LOI**
0,0	-0,518964	0,660317	-0,178313
0,25	-0,160599	0,289416	-0,12857
0,5	-0,0801057	0,197023	-0,104905
0,75	-0,0465913	0,153334	-0,0893168
1,0	-0,0290842	0,127168	-0,0779872
1,25	-0,0187594	0,109427	-0,0693031
1,5	-0,012189	0,0964527	-0,0624063
1,75	-0,00778377	0,0864697	-0,0567842
2,0	**-0,0047164**	**0,0785054**	**-0,0521071**
2,25	-0,00251915	0,0719773	-0,0481518
2,5	-0,000910668	0,0665128	-0,0447613
2,75	0,000286661	0,0618613	-0,0418215
3,0	0,00118939	0,0578473	-0,0392472
3,25	0,00187648	0,0543435	-0,0369739
3,5	0,00240289	0,0512555	-0,0349513
3,75	0,00280774	0,0485112	-0,0331398
4,0	0,00311942	0,0460546	-0,0315078
4,25	0,0033589	0,0438417	-0,0300299
4,5	0,00354188	0,0418369	-0,028685
4,75	0,0036803	0,0400117	-0,0274559
5,0	0,0037833	0,0383425	-0,0263282

Από τον παραπάνω πίνακα παρατηρούμε ότι μια καλή τιμή για την παράμετρο ραχοειδούς παλινδρόμησης είναι η τιμή 2,0. Αυτό μπορεί και να επιβεβαιωθεί από το

διάγραμμα των VIFs ως προς την παράμετρο. Το διάγραμμα αυτό αποτελεί τον δεύτερο τρόπο καθορισμού της τιμής της παραμέτρου. Μετά την τιμή 2,0, οι τιμές των VIFs σταθεροποιούνται αρκετά. Το διάγραμμα αυτό απεικονίζεται αμέσως παρακάτω, στο Διάγραμμα 6.17.



<u>Διάγραμμα 6.17</u>: Διάγραμμα Πληθωριστικών Παραγόντων Διασποράς ως προς την Παράμετρο Ραχοειδούς Παλινδρόμησης, Πολλαπλή Παλινδρόμηση ESt28, CEM II 42,5, MT4

Με βάση την τιμή 2,0 που επιλέχθηκε για την παράμετρο, το μοντέλο που προκύπτει από την παλινδρόμηση είναι το εξής:

Est28 = 50,5387 - 0,0135371*SiO2 + 0,788729*Al2O3 - 0,325856*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.18.

<u>Πίνακας 6.18</u>: Ραχοειδής Παλινδρόμηση για Est28, CEM II 42,5, MT4

Παράμετρος ραχοειδούς παλινδρόμησης = 2,0

Παράμετρος	Εκτίμηση	Variance Inflation Factor
CONSTANT	50,5387	
SiO2	-0,0135371	0,0789197
Al2O3	0,788729	0,0793429
LOI	-0,325856	0,110609

R-2 = 2,65589 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 0,159884 τοις εκατό
Τυπικό σφάλμα = 1,7364
Μέσο απόλυτο σφάλμα = 1,34392
Durbin-Watson statistic = 1,25748 (P=0,0000)

Τα βασικά συμπεράσματα που προκύπτουν από το παραπάνω μοντέλο είναι ότι ο συντελεστής R^2 είναι αρκετά μικρότερος από ό,τι προηγουμένως, και αυτό είναι λογικό, καθώς η Est28 συσχετίζεται μόνο με μία από τις ανεξάρτητες μεταβλητές (Al_2O_3). Επίσης, το πρόβλημα αυτοσυσχέτισης των καταλοίπων είναι ακόμα παρόν, καθώς το *p-value* στο *Durbin-Watson test* είναι μικρότερο από 0,05.

- ***Εξαρτημένη Μεταβλητή Est2 – Όλες οι Μεταβλητές***

Το μοντέλο παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής Est2 και των ανεξάρτητων μεταβλητών είναι το εξής:

Est2 = 19,2662 - 0,711632*SiO2 + 2,02924*Al2O3 + 0,00322085*Blaine + 0,046854*IR - 0,753215*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων φαίνονται στον Πίνακα 6.19.

Πίνακας 6.19: Πολλαπλή Παλινδρόμηση για Est2-Όλες οι Μεταβλητές, CEM II 42,5, MT4

Εξαρτημένη μεταβλητή: Est2

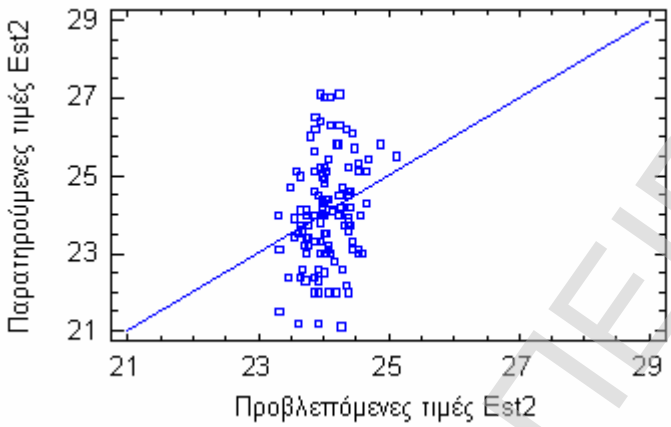
Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	19,2662	9,00761	2,13888	0,0346
SiO2	-0,711632	0,508635	-1,3991	0,1645
Al2O3	2,02924	1,32396	1,5327	0,1281
Blaine	0,00322085	0,00183226	1,75785	0,0814
IR	0,046854	0,220906	0,212099	0,8324
LOI	-0,753215	0,524408	-1,43632	0,1536

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	13,199	5	2,6398	1,50	0,1952
Κατάλοιπα	202,354	115	1,7596		
Σύνολο	215,553	120			

R-2 = 6,12331 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 2,04172 τοις εκατό
Τυπικό σφάλμα = 1,3265
Μέσο απόλυτο σφάλμα = 1,01186
Durbin-Watson statistic = 1,28198 (P=0,0000)

Ο συντελεστής προσδιορισμού R^2 εξηγεί 6,12331% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος R^2 είναι αρκετά χαμηλότερος και ισούται με 2,04172%. Το τυπικό σφάλμα είναι 1,3265, ενώ και σε αυτή την περίπτωση ο συντελεστής *Durbin-Watson* δεν είναι επιθυμητός, συνεπώς έχουμε πρόβλημα αυτοσυσχέτισης καταλοίπων.



Διάγραμμα 6.18: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est2-Όλες οι Μεταβλητές, CEM II 42,5, MT4

- ***Εξαρτημένη Μεταβλητή Est2 – Προς τα Εμπρός Επιλογή***

Το μοντέλο παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής Est2 και των ανεξάρτητων μεταβλητών που προκύπτει από τη βηματική μέθοδο επιλογής των μεταβλητών είναι το εξής:

Est2 = 24,0587

Όπως φαίνεται από το παραπάνω μοντέλο, υπάρχει μόνο σταθερός όρος μέσα σε αυτό και καμία ανεξάρτητη μεταβλητή. Αυτό το περιμέναμε, καθώς η Est2 δεν συσχετιζόταν με καμία από τις ανεξάρτητες μεταβλητές στον πίνακα συσχετίσεων. Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων φαίνονται στον Πίνακα 6.20.

Πίνακας 6.20: Πολλαπλή Παλινδρόμηση για Est2-Προς τα Εμπρός Επιλογή, CEM II 42,5, MT4

Εξαρτημένη μεταβλητή: Est2

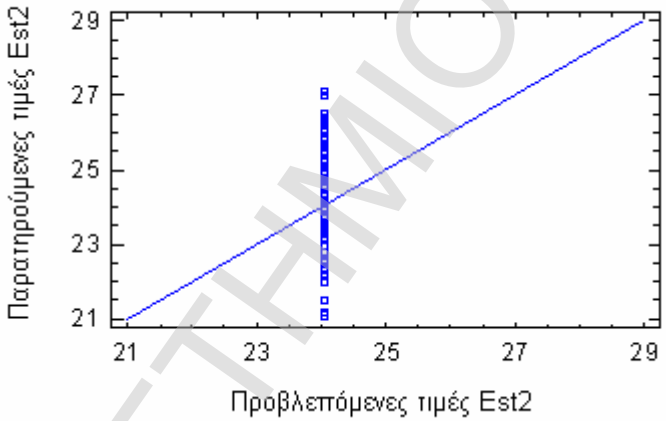
Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	24,0587	0,121841	197,459	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	0,0	0			
Κατάλοιπα	215,553	120	1,79628		
Σύνολο	215,553	120			

R-2 = 0,0 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 0,0 τοις εκατό
Τυπικό σφάλμα = 1,34025
Μέσο απόλυτο σφάλμα = 1,06042
Durbin-Watson statistic = 1,27681 (P=0,0000)

Ο συντελεστής προσδιορισμού R^2 εξηγεί 0,0% της μεταβλητότητας στη μεταβλητή Est28. Δηλαδή, στην ουσία δεν υπάρχει μοντέλο παλινδρόμησης. Αυτό φαίνεται και από το παρακάτω διάγραμμα, όπου οι τιμές της *Y* είναι σταθερές.



Διάγραμμα 6.19: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est2-Προς τα Εμπρός Επιλογή, CEM II 42,5, MT4

6.3.3. Απλή Παλινδρόμηση

Οι μεταβλητές **Est28, Est2** και **Est7** από τον πίνακα συσχετίσεων φάνηκε ότι συσχετίζονται μέτρια. Για το λόγο αυτόν, επιχειρείται η δημιουργία ενός μοντέλου απλής παλινδρόμησης μεταξύ των μεταβλητών αυτών.

- ***Απλή Παλινδρόμηση μεταξύ Est28 και Est2***

Για να αποφασίσουμε ποια μορφή απλής παλινδρόμησης θα επιλέξουμε, εξετάζουμε διάφορα εναλλακτικά μοντέλα ως προς το ποιο από αυτά έχει τη μεγαλύτερη τιμή σε R^2. Τα εναλλακτικά μοντέλα απεικονίζονται στον Πίνακα 6.21.

Πίνακας 6.21: Σύγκριση Εναλλακτικών Μοντέλων Απλής Παλινδρόμησης

Μοντέλο	Συσχέτιση	R-2
Τετραγωνικής ρίζας τουX	0,4714	22,22%
Λογαριθμικό ως προς X	0,4714	22,22%
Γραμμικό	0,4713	22,21%
Τετραγωνικής ρίζας του Y	0,4708	22,17%
Αντίστροφο ως προς X	-0,4707	22,15%
Πολλαπλασιαστικό	0,4705	22,14%
Εκθετικό	0,4703	22,12%
Καμπύλη S	-0,4700	22,09%
Αντίστροφο ως προς Y	-0,4691	22,00%
Διπλής αντιστροφής	0,4690	21,99%
Λογιστικό	<no fit>	
Log probit	<no fit>	

Το μοντέλο της τετραγωνικής ρίζας της μεταβλητής *X* φαίνεται ότι έχει το μεγαλύτερο R^2. Η μορφή του μοντέλου αυτού είναι: *Y=a+b*SQRT(X)*. Συνεπώς, επιλέγεται αυτό για την ανάλυση που ακολουθεί. Στον Πίνακα 6.22 αναγράφονται τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου, ενώ η εξίσωση της απλής παλινδρόμησης είναι η ακόλουθη:

Est28 = 23,748 + 6,05875*sqrt(Est2)

Πίνακας 6.22: Απλή Παλινδρόμηση μεταξύ Est28 και Est2, CEM II 42,5, MT4

Μοντέλο Τετραγωνικής Ρίζας του X: Y = a + b*sqrt(X)

Εξαρτημένη μεταβλητή: Est28
Ανεξάρτητη μεταβλητή: Est2

Παράμετρος	Εκτίμηση	Τυπ. Σφάλμα	T Statistic	P-Value
Σταθερός όρος	23,748	5,09663	4,65955	0,0000
Κλίση	6,05875	1,03908	5,8309	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Μέσο Άθροισμα Β.ε. Τετραγώνων	F-Ratio	P-Value
Μοντέλο	82,0299	1	82,0299	34,00	0,0000
Κατάλοιπα	287,11	119	2,41269		
Σύνολο	369,14	120			

Συντελεστής συσχέτισης = 0,471401
R-2 = 22,2219 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 21,5683 τοις εκατό
Τυπικό σφάλμα = 1,55328
Μέσο απόλυτο σφάλμα = 1,22735
Durbin-Watson statistic = 1,29365 (P=0,0000)

Το *p-value* στον πίνακα της ANOVA είναι μικρότερο από 0,01, συνεπώς υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Est28 και της Est2 στο 99% επίπεδο εμπιστοσύνης.

Το R^2 είναι 22,2219%, που δηλώνει ότι το μοντέλο που προσαρμόζεται στα δεδομένα εξηγεί κατά αυτό το ποσοστό τη μεταβλητότητα του Est28. Ο συντελεστής συσχέτισης είναι 0,471401, που δείχνει μια σχετικά ασθενή σχέση μεταξύ των μεταβλητών. Το τυπικό σφάλμα της εκτίμησης είναι 1,55328 και ισούται με την τυπική απόκλιση των καταλοίπων. Από το *Durbin-Watson (DW) statistic* προκύπτει ότι το *p-value* είναι μικρότερο από 0,05, το οποίο φανερώνει την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

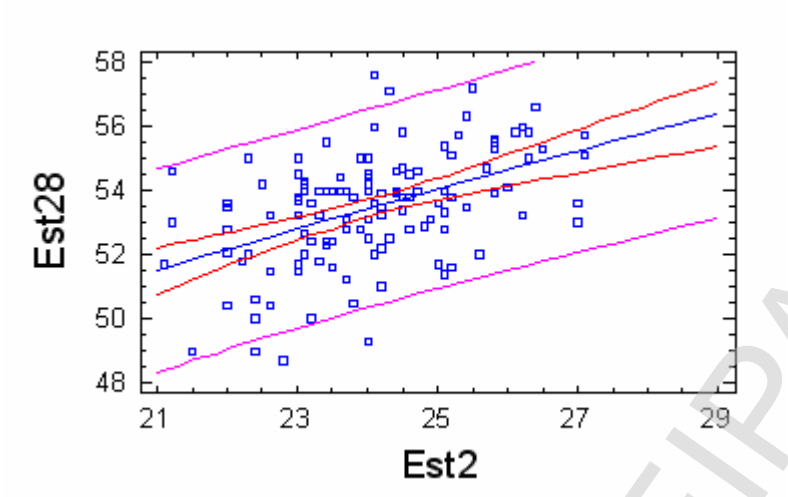
Για να ελέγξουμε αν το προτεινόμενο μοντέλο επαρκεί για να περιγράψει τις τιμές των δεδομένων, εκτελούμε το *Lack-of-Fit test*. Τα αποτελέσματα του test αυτού φαίνονται στον Πίνακα 6.23.

Πίνακας 6.23: Έλεγχος Έλλειψης Προσαρμογής στην Απλή Παλινδρόμηση μεταξύ Est28 και Est2, CEM II 42,5, MT4

Ανάλυση Διακύμανσης με Lack-of-Fit τεστ

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	**P-Value**
Μοντέλο	82,0299	1	82,0299	34,00	0,0000
Κατάλοιπα	287,11	119	2,41269		
Lack-of-Fit	131,742	45	2,92759	1,39	**0,1013**
Καθ.σφάλμα	155,369	74	2,09957		
Σύνολο	369,14	120			

Καθώς το *p-value* για το lack-of-fit test στον πίνακα της ANOVA είναι μεγαλύτερο από 0,10, το μοντέλο φαίνεται να επαρκεί (οριακά) για τα παρατηρούμενα δεδομένα. Το διάγραμμα της σχέσης μεταξύ των μεταβλητών φαίνεται στο παρακάτω Διάγραμμα 6.20. Τα εσωτερικά όρια (κόκκινες γραμμές) δηλώνουν τα 95% διαστήματα εμπιστοσύνης για το μέσο Est28 των παρατηρήσεων για τις συγκεκριμένες τιμές Est2 από τα δεδομένα. Τα εξωτερικά όρια (ροζ γραμμές) δηλώνουν τα 95% διαστήματα *πρόβλεψης* για τις νέες παρατηρήσεις.



<u>Διάγραμμα 6.20</u>: Διάγραμμα Απλής Παλινδρόμησης των μεταβλητών Est28 και Est2, CEM II 42,5, MT4

- ***<u>Απλή Παλινδρόμηση μεταξύ Est28 και Est7</u>***

Για να αποφασίσουμε ποια μορφή απλής παλινδρόμησης θα επιλέξουμε, εξετάζουμε διάφορα εναλλακτικά μοντέλα ως προς το ποιο από αυτά έχει τη μεγαλύτερη τιμή σε R^2. Τα εναλλακτικά μοντέλα απεικονίζονται στον Πίνακα 6.24.

<u>Πίνακας 6.24</u>: Σύγκριση Εναλλακτικών Μοντέλων Απλής Παλινδρόμησης

Μοντέλο	Συσχέτιση	R-2
Γραμμικό	0,6136	37,65%
Τετραγωνικής ρίζας του X	0,6134	37,63%
Λογαριθμικό ως προς X	0,6132	37,60%
Τετραγωνικής ρίζας του Y	0,6131	37,59%
Εκθετικό	0,6126	37,52%
Αντίστροφο ως προς X	-0,6124	37,50%
Πολλαπλασιαστικό	0,6124	37,50%
Καμπύλη S	-0,6118	37,43%
Αντίστροφο ως προς Y	-0,6112	37,35%
Διπλής αντιστροφής	0,6108	37,31%
Λογιστικό	<no fit>	
Log probit	<no fit>	

Το γραμμικό μοντέλο φαίνεται ότι έχει το μεγαλύτερο R^2. Η μορφή του μοντέλου αυτού είναι: $Y=a+b* X$. Συνεπώς, επιλέγεται αυτό για την ανάλυση που ακολουθεί. Στον Πίνακα 6.25 αναγράφονται τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου, ενώ η εξίσωση της απλής παλινδρόμησης είναι η ακόλουθη:

Est28 = 23,6077 + 0,765856*Est7

Πίνακας 6.25: Απλή Παλινδρόμηση μεταξύ Est28 και Est7, CEM II 42,5, MT4

Γραμμικό Μοντέλο: Y = a + b* X

Εξαρτημένη μεταβλητή: Est28
Ανεξάρτητη μεταβλητή: Est7

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
Σταθερός όρος	23,6077	3,52333	6,70039	0,0000
Κλίση	0,765856	0,0903486	8,47668	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	138,976	1	138,976	71,85	0,0000
Κατάλοιπα	230,164	119	1,93415		
Σύνολο	369,14	120			

Συντελεστής συσχέτισης = 0,613586
R-2 = 37,6487 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 37,1248 τοις εκατό
Τυπικό σφάλμα = 1,39074
Μέσο απόλυτο σφάλμα = 1,10171
Durbin-Watson statistic = 1,47972 (P=0,0019)

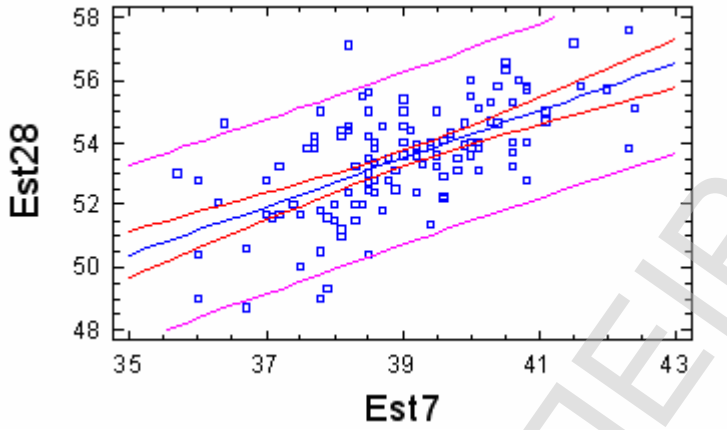
Το *p-value* στον πίνακα της ANOVA είναι μικρότερο από 0,01, συνεπώς υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Est28 και της Est7 στο 99% επίπεδο εμπιστοσύνης. Το R^2 είναι 37,6487%, ενώ ο συντελεστής συσχέτισης είναι 0,613586, που δείχνει μια μέτρια σχέση μεταξύ των μεταβλητών. Το τυπικό σφάλμα της εκτίμησης είναι 1,39074. Από το *Durbin-Watson statistic* προκύπτει ότι το *p-value* είναι μικρότερο από 0,05, το οποίο φανερώνει την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων. Για να ελέγξουμε αν το προτεινόμενο μοντέλο επαρκεί για να περιγράψει τις τιμές των δεδομένων, εκτελούμε το *Lack-of-Fit test*. Τα αποτελέσματα του test αυτού φαίνονται στον Πίνακα 6.26.

Πίνακας 6.26: Έλεγχος Έλλειψης Προσαρμογής στην Απλή Παλινδρόμηση μεταξύ Est28 και Est7, CEM II 42,5, MT4

Ανάλυση Διακύμανσης με Lack-of-Fit τεστ

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	**P-Value**
Μοντέλο	138,976	1	138,976	71,85	0,0000
Κατάλοιπα	230,164	119	1,93415		
Lack-of-Fit	105,492	46	2,29331	1,34	**0,1286**
Καθ. σφάλμα	124,671	73	1,70783		
Σύνολο	369,14	120			

Καθώς το *p-value* για το lack-of-fit test στον πίνακα της ANOVA είναι μεγαλύτερο από 0,10, το μοντέλο φαίνεται να επαρκεί για τα παρατηρούμενα δεδομένα. Το διάγραμμα της σχέσης μεταξύ των μεταβλητών φαίνεται στο παρακάτω Διάγραμμα 6.21.



Διάγραμμα 6.21: Διάγραμμα Απλής Παλινδρόμησης των μεταβλητών Est28 και Est7, CEM II 42,5, MT4

6.3.4. Έλεγχος των Προϋποθέσεων της Παλινδρόμησης

Ο έλεγχος των προϋποθέσεων της πολλαπλής και της απλής παλινδρόμησης ως προς τα κατάλοιπα γίνεται με τον ακόλουθο τρόπο:

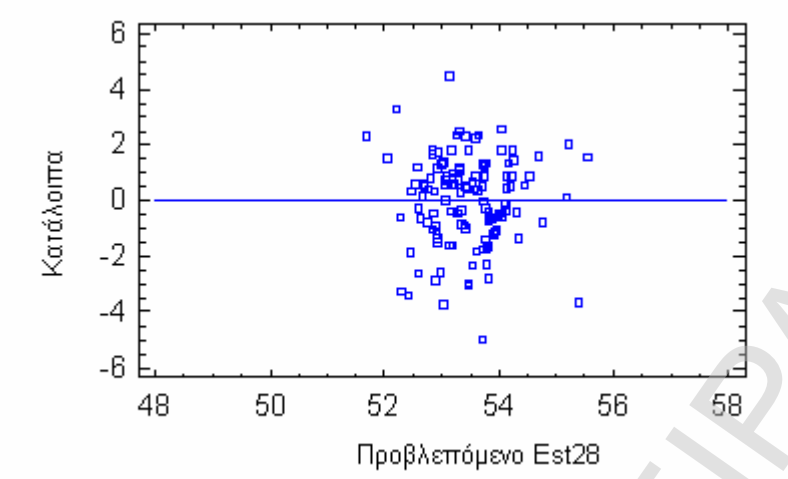
- Σχετικά με την υπόθεση για **κανονική κατανομή** των καταλοίπων και, συνεπώς, της εξαρτημένης μεταβλητής *Y*, γνωρίζουμε από τη στατιστική ανάλυση που προηγήθηκε στο αμέσως προηγούμενο κεφάλαιο ότι οι μεταβλητές Est28 και Est2 ακολουθούν κανονική κατανομή.

- Όσον αφορά τη **μέση τιμή** και τη **διασπορά** των καταλοίπων, τα αποτελέσματα του ελέγχου είναι τα εξής:

  - ***Πολλαπλή Παλινδρόμηση Est28 – Προς τα Εμπρός Επιλογή***

    Το μέσο απόλυτο σφάλμα (MAE) ισούται με 1,28002, το οποίο είναι διάφορο του μηδενός. Από το διάγραμμα, όμως, των καταλοίπων ως προς τις προβλεπόμενες τιμές Est28 δεν φαίνεται να υπάρχει σημαντική απόκλιση από την υπόθεση ότι το μέσο σφάλμα έχει μηδενική τιμή. Σχετικά με τη διασπορά, φαίνεται από το ίδιο διάγραμμα ότι έχει παντού το ίδιο εύρος, καθώς προχωράμε δεξιά προς τον οριζόντιο άξονα.

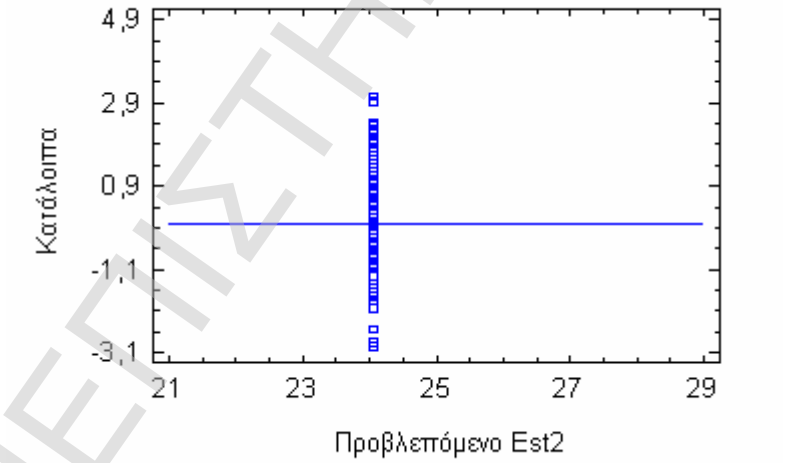
    Το Διάγραμμα 6.22 απεικονίζει το διάγραμμα καταλοίπων ως προς τις προβλεπόμενες τιμές *Y*, για τον έλεγχο των προϋποθέσεων μηδενικής μέσης τιμής των σφαλμάτων και σταθερής διασποράς αυτών.



Διάγραμμα 6.22: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Πολλαπλή Παλινδρόμηση, CEM II 42,5, MT4

§ *Πολλαπλή Παλινδρόμηση Est2 – Προς τα Εμπρός Επιλογή*

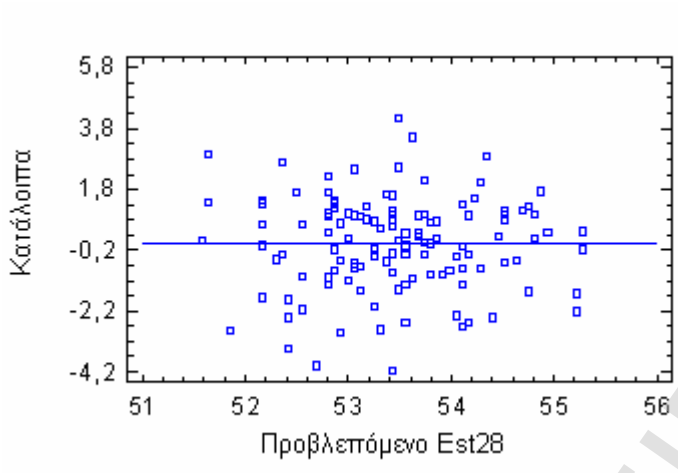
Το MAE ισούται με 1,06042, το οποίο διαφέρει από την τιμή μηδέν. Από το διάγραμμα των καταλοίπων ως προς τις προβλεπόμενες τιμές Est2 φαίνεται ότι τα κατάλοιπα είναι ομοιόμορφα κατανεμημένα πάνω και κάτω από τη γραμμή του οριζόντιου άξονα, δηλαδή ότι η μέση τιμή αυτών προσεγγίζει το μηδέν. Όσον αφορά τη διασπορά, αυτή παραμένει σταθερή, καθώς οι προβλεπόμενες τιμές του Est2 παραμένουν σταθερές.



Διάγραμμα 6.23: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est2, Πολλαπλή Παλινδρόμηση, CEM II 42,5, MT4

§ *Απλή Παλινδρόμηση Est28 και Est2*

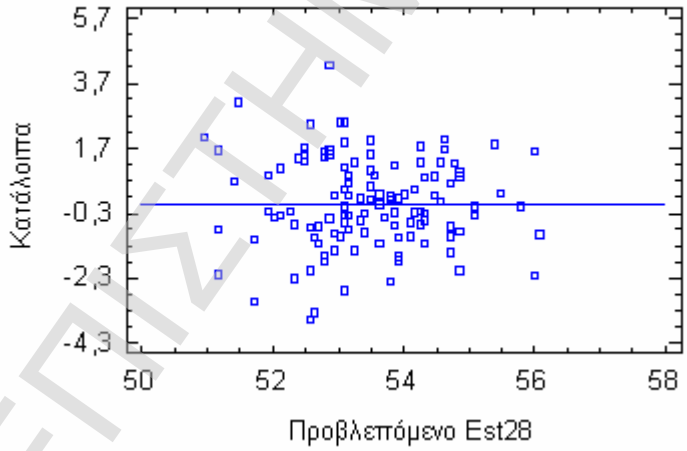
Στην απλή παλινδρόμηση, το MAE είναι 1,22735, το οποίο δεν ισούται με την τιμή μηδέν. Από το διάγραμμα των καταλοίπων δεν φαίνεται κάποιο ιδιαίτερο πρόβλημα για την καταπάτηση αυτής της προϋπόθεσης. Η διασπορά φαίνεται να είναι ομοιόμορφα κατανεμημένη.



Διάγραμμα 6.24: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Απλή Παλινδρόμηση Est28 και Est2, CEM II 42,5, MT4

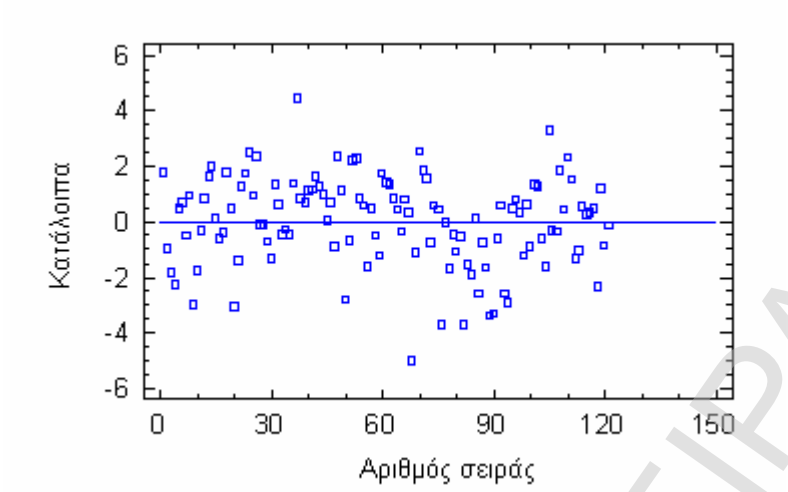
§ ***Απλή Παλινδρόμηση Est28 και Est7***

Στην απλή παλινδρόμηση, το MAE ισούται με 1,10171, το οποίο δεν ισούται με την τιμή μηδέν. Από το διάγραμμα των καταλοίπων δεν φαίνεται κάποιο ιδιαίτερο πρόβλημα για την καταπάτηση αυτής της προϋπόθεσης. Η διασπορά δεν φαίνεται να είναι ομοιόμορφα κατανεμημένη, οπότε υπάρχει πρόβλημα ετεροσκεδαστικότητας.

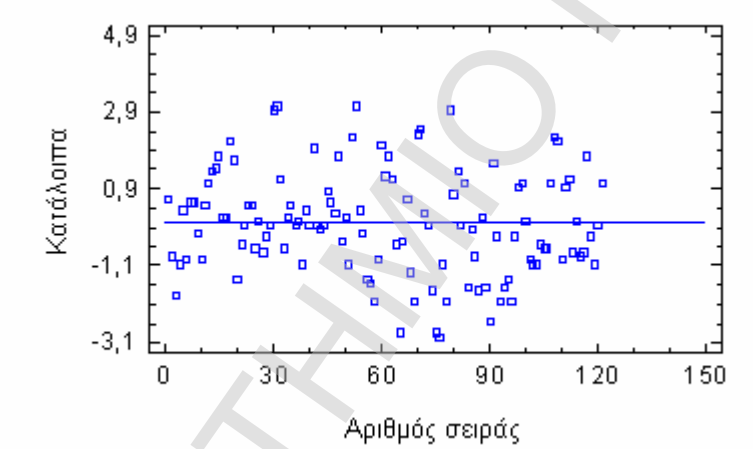


Διάγραμμα 6.25: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Απλή Παλινδρόμηση Est28 και Est7, CEM II 42,5, MT4

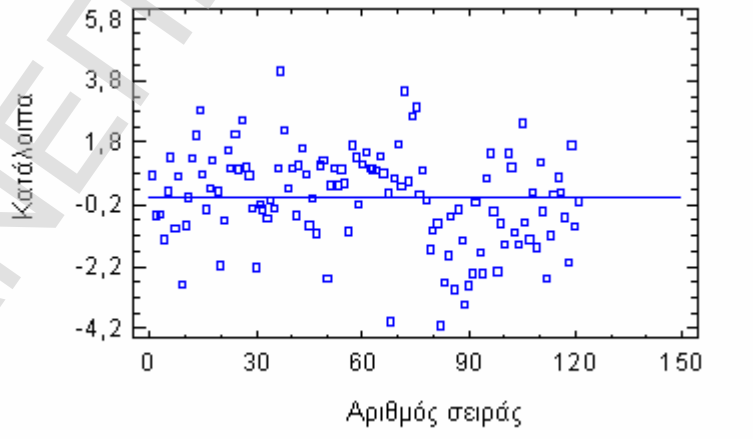
ü Σχετικά με την ύπαρξη **αυτοσυσχέτισης** μεταξύ των καταλοίπων, προκύπτει ότι τα *p-values* των Durbin-Watson tests και για τις τέσσερις περιπτώσεις παλινδρόμησης είναι μικρότερα από 0,05. Άρα, υπάρχει αυτοσυσχέτιση μεταξύ των καταλοίπων. Αυτό γίνεται αντιληπτό και από τα διαγράμματα, τα οποία παρουσιάζονται παρακάτω.



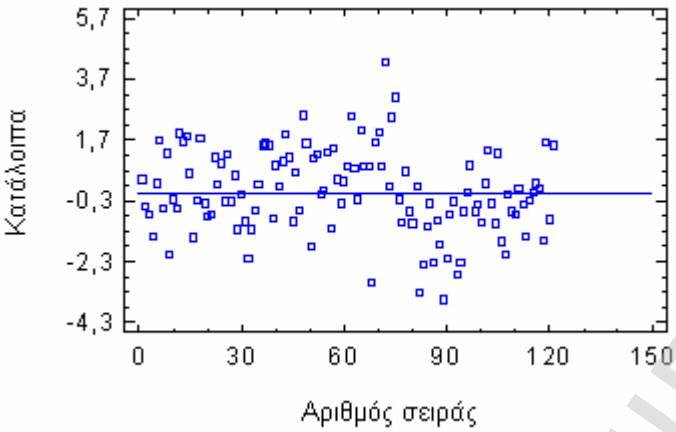
Διάγραμμα 6.26: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Πολλαπλή Παλινδρόμηση της Est28, CEM II 42,5, MT4



Διάγραμμα 6.27: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Πολλαπλή Παλινδρόμηση της Est2, CEM II 42,5, MT4



Διάγραμμα 6.28: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Απλή Παλινδρόμηση μεταξύ των Est28 και Est2, CEM II 42,5, MT4



Διάγραμμα 6.29: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Απλή Παλινδρόμηση μεταξύ των Est28 και Est7, CEM II 42,5, MT4

6.4. Απλό Τσιμέντο Portland OPC – Μύλος Παραγωγής 3

6.4.1. Έλεγχος Συσχετίσεων μεταξύ των Μεταβλητών

Παρακάτω παρουσιάζεται ο πίνακας συσχετίσεων κατά Spearman, αφού οι περισσότερες μεταβλητές δεν είναι κανονικά κατανεμημένες.

Πίνακας 6.27: Πίνακας Συσχετίσεων των Μεταβλητών κατά Spearman

	SiO_2	***Al_2O_3***	***Blaine***	***LOI***	***Est2***	***Est7***	***Est28***
SiO_2		-0,0673 (653) 0,0855	-0,0209 (653) 0,5937	**-0,1290** (653) **0,0010**	**-0,4470** (653) **0,0000**	**-0,3427** (653) **0,0000**	0,0549 (653) 0,1607
Al_2O_3	-0,0673 (653) 0,0855		**-0,4028** (653) **0,0000**	-0,0577 (653) 0,1405	0,0278 (653) 0,4770	**0,1004** (653) **0,0104**	**0,2755** (653) **0,0000**
Blaine	-0,0209 (653) 0,5937	**-0,4028** (653) **0,0000**		0,0728 (653) 0,0631	-0,0005 (653) 0,9900	**-0,1303** (653) **0,0009**	**-0,3342** (653) **0,0000**
LOI	**-0,1290** (653) **0,0010**	-0,0577 (653) 0,1405	0,0728 (653) 0,0631		-0,0486 (653) 0,2146	**-0,1260** (653) **0,0013**	**-0,2081** (653) **0,0000**
Est2	**-0,4470** (653) **0,0000**	0,0278 (653) 0,4770	-0,0005 (653) 0,9900	-0,0486 (653) 0,2146		**0,7939** (653) **0,0000**	**0,2435** (653) **0,0000**
Est7	**-0,3427** (653) **0,0000**	**0,1004** (653) **0,0104**	**-0,1303** (653) **0,0009**	**-0,1260** (653) **0,0013**	**0,7939** (653) **0,0000**		**0,5295** (653) **0,0000**
Est28	0,0549 (653) 0,1607	**0,2755** (653) **0,0000**	**-0,3342** (653) **0,0000**	**-0,2081** (653) **0,0000**	**0,2435** (653) **0,0000**	**0,5295** (653) **0,0000**	

Μεταβλητές με *p-values* μικρότερες από 0,05 φανερώνουν στατιστικά σημαντικές μη-μηδενικές συσχετίσεις, σε 95 % επίπεδο εμπιστοσύνης. Τα ζεύγη των μεταβλητών που έχουν *p-values* κάτω από 0,05 είναι τα ακόλουθα:

SiO2 και **LOI**
SiO2 και **Est2**
SiO2 και **Est7**
Al2O3 και **Blaine**
Al2O3 και **Est7**
Al2O3 και **Est28**
Blaine και **Est7**
Blaine και **Est28**
LOI και **Est7**
LOI και **Est28**
Est2 και **Est7**
Est2 και **Est28**
Est7 και **Est28**

Η εξαρτημένη μεταβλητή Est28 φαίνεται από τον παραπάνω πίνακα ότι συσχετίζεται με αρκετές μεταβλητές, όπως τις Al_2O_3, Blaine, LOI, Est2 και Est7. Άρα, στην πολλαπλή παλινδρόμηση περιμένουμε ότι κάποιες από τις παραπάνω μεταβλητές θα είναι παρούσες, ενώ μπορούμε να εφαρμόσουμε ένα μοντέλο απλής παλινδρόμησης με τη μεταβλητή Est2 και την Est7. Όσον αφορά την εξαρτημένη μεταβλητή Est2, παρατηρούμε ότι συσχετίζεται μόνο με την ανεξάρτητη μεταβλητή SiO_2.

6.4.2. Πολλαπλή Παλινδρόμηση

- ***Εξαρτημένη Μεταβλητή Est28 – Όλες οι Μεταβλητές***

Το μοντέλο παλινδρόμησης που προκύπτει είναι το εξής:

Est28 = 67,6419 + 0,149859*SiO2 + 2,71224*Al2O3 - 0,00513764*Blaine - 1,81537*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.28.

Πίνακας 6.28: Πολλαπλή Παλινδρόμηση για Est28-Όλες οι Μεταβλητές, OPC, MT3

Εξαρτημένη μεταβλητή: Est28

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	67,6419	7,16997	9, 43406	0,0000
SiO2	0,149859	0,225616	0,664223	0,5065
Al2O3	2,71224	0,5066	5,35381	0,0000
Blaine	-0,00513764	0,000996338	-5,15652	0,0000
LOI	-1,81537	0,409871	-4,42913	0,0000

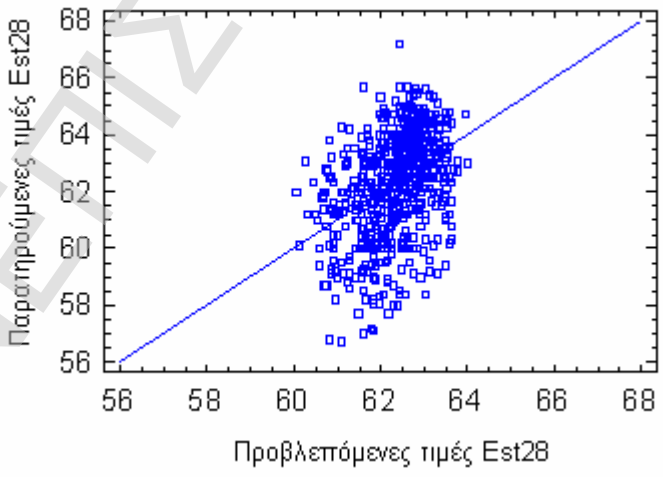
Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	340,244	4	85,0611	32,13	0,0000
Κατάλοιπα	1713,02	647	2,64764		
Σύνολο	2053,27	651			

R-2 = 16,5709 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 16,0551 τοις εκατό
Τυπικό σφάλμα = 1,62716
Μέσο απόλυτο σφάλμα = 1,28224
Durbin-Watson statistic = 1,06605 (P=0,0000)

Ο συντελεστής προσδιορισμού R^2 δείχνει ότι το προσαρμοσμένο μοντέλο εξηγεί 16,5709% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 είναι 16,0551%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 1,62716. Το μέσο απόλυτο σφάλμα (MAE) είναι 1,28224 και αποτελεί τη μέση τιμή των καταλοίπων. Το *Durbin-Watson (DW) statistic* ελέγχει τα κατάλοιπα για το αν υπάρχει αυτοσυσχέτιση. Επειδή το *p-value* είναι μικρότερο από 0,05, υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής Est28 σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές. Όσο πιο κοντά τα δεδομένα "πέφτουν" στη διαγώνια γραμμή, τόσο καλύτερο είναι το μοντέλο στην πρόβλεψη των παρατηρούμενων δεδομένων. Στη συγκεκριμένη περίπτωση οι παρατηρούμενες τιμές αποκλίνουν αρκετά από την ευθεία γραμμή, άρα το μοντέλο δεν θεωρείται ιδιαίτερα καλό.



<u>Διάγραμμα 6.30</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est28-Όλες οι Μεταβλητές, OPC, MT3

- ***Εξαρτημένη Μεταβλητή Est28 – Προς τα Εμπρός Επιλογή***

Το μοντέλο παλινδρόμησης που προκύπτει είναι το εξής:

Est28 = 70,9219 + 2,69899*Al2O3 - 0,00515807*Blaine - 1,8558*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.29.

Πίνακας 6.29: Πολλαπλή Παλινδρόμηση για Est28-Προς τα Εμπρός Επιλογή, OPC, MT3

Εξαρτημένη μεταβλητή: Est28

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	70,9219	5,19622	13,6487	0,0000
Al2O3	2,69899	0,505989	5,33409	0,0000
Blaine	-0,00515807	0,000995434	-5,18174	0,0000
LOI	-1,8558	0,405152	-4,58051	0,0000

Ανάλυση Διακύμανσης

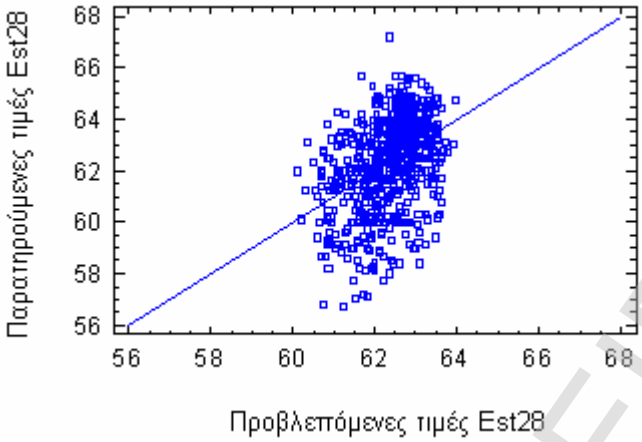
	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	339,076	3	113,025	42,73	0,0000
Κατάλοιπα	1714,19	648	2,64536		
Σύνολο	2053,27	651			

R-2 = 16,514 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 16,1275 τοις εκατό
Τυπικό σφάλμα = 1,62646
Μέσο απόλυτο σφάλμα = 1,28276
Durbin-Watson statistic = 1,0648 (P=0,0000)

Από τις μεταβλητές που προκρίνονται στο μοντέλο όλες συσχετίζονται με την Est28. Ο συντελεστής προσδιορισμού R^2 δείχνει ότι το προσαρμοσμένο μοντέλο εξηγεί 16,514% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 είναι 16,1275%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 1,62646. Το μέσο απόλυτο σφάλμα (MAE) είναι 1,28276 και αποτελεί τη μέση τιμή των καταλοίπων. Το *Durbin-Watson (DW) statistic* ελέγχει τα κατάλοιπα για το αν υπάρχει αυτοσυσχέτιση. Επειδή το *p-value* είναι μικρότερο από 0,05, υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής Est28 σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές. Στη συγκεκριμένη περίπτωση οι

παρατηρούμενες τιμές αποκλίνουν αρκετά από την ευθεία γραμμή, άρα το μοντέλο δεν θεωρείται ιδιαίτερα καλό.



Διάγραμμα 6.31: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est28-Προς τα Εμπρός Επιλογή, OPC, MT3

Στη συνέχεια ελέγχεται η συσχέτιση μεταξύ των εκτιμητριών των συντελεστών παλινδρόμησης:

Πίνακας 6.30: Πίνακας Συσχετίσεων των Εκτιμητριών των Συντελεστών Παλινδρόμησης

	CONSTANT	Al2O3	Blaine	LOI
CONSTANT	1,0000	-0,1074	-0,1938	-0,1809
Al2O3	-0,1074	1,0000	0,4397	0,0279
Blaine	-0,1938	0,4397	1,0000	-0,0545
LOI	-0,1809	0,0279	-0,0545	1,0000
Clk	-0,9632	-0,1022	-0,0513	0,1680

Όπως φαίνεται από τον παραπάνω πίνακα, η συσχέτιση μεταξύ των μεταβλητών δεν είναι σε κανένα ζεύγος πολύ ισχυρή, αφού έχει τιμή μικρότερη από 0,5. Για το λόγο αυτό, δεν εφαρμόζεται *ραχοειδής παλινδρόμηση*, αφού η μέθοδος αυτή προϋποθέτει ισχυρές συσχετίσεις μεταξύ των μεταβλητών.

- ***Εξαρτημένη Μεταβλητή Est2 – Όλες οι Μεταβλητές***

Το μοντέλο παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής Est2 και των ανεξάρτητων μεταβλητών είναι το εξής:

Est2 = 78,5401 - 2,52516*SiO2 + 0,316477*Al2O3 + 0,000665548*Blaine - 1,24825*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων φαίνονται στον Πίνακα 6.31.

<u>Πίνακας 6.31</u>: Πολλαπλή Παλινδρόμηση για Est2-Όλες οι Μεταβλητές, OPC, MT3

Εξαρτημένη μεταβλητή: Est2

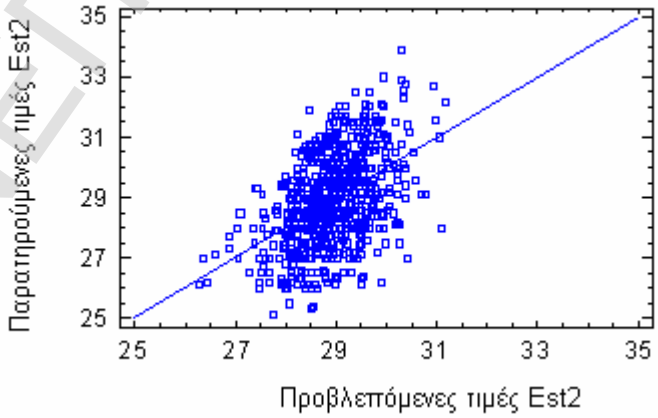
Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	78,5401	5,9704	13,1549	0,0000
SiO2	-2,52516	0,187869	-13,4411	0,0000
Al2O3	0,316477	0,421843	0,750224	0,4531
Blaine	0,000665548	0,000829646	0,802208	0,4224
LOI	-1,24825	0,341298	-3,65737	0,0003

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	340,909	4	85,2271	46,42	0,0000
Κατάλοιπα	1187,78	647	1,83582		
Σύνολο	1528,69	651			

R-2 = 22,3007 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 21,8204 τοις εκατό
Τυπικό σφάλμα = 1,35493
Μέσο απόλυτο σφάλμα = 1,10337
Durbin-Watson statistic = 1,07476 (P=0,0000)

Ο συντελεστής προσδιορισμού R^2 εξηγεί το 22,3007% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 εξηγεί το 21,8204%. Στο παρακάτω διάγραμμα φαίνεται η σχέση μεταξύ των παρατηρούμενων τιμών της Est2 και των τιμών που προβλέπει το μοντέλο.



<u>Διάγραμμα 6.32</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est2-Όλες οι Μεταβλητές, OPC, MT3

- ***Εξαρτημένη Μεταβλητή Est2 – Προς τα Εμπρός Επιλογή***

Το μοντέλο παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής Est2 και των ανεξάρτητων μεταβλητών που προκύπτει από τη βηματική μέθοδο επιλογής των μεταβλητών είναι το εξής:

Est2 = 82,634 - 2,53221*SiO2 - 1,2486*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων φαίνονται στον Πίνακα 6.32.

Πίνακας 6.32: Πολλαπλή Παλινδρόμηση για Est2-Προς τα Εμπρός Επιλογή, OPC, MT3

Εξαρτημένη μεταβλητή: Est2

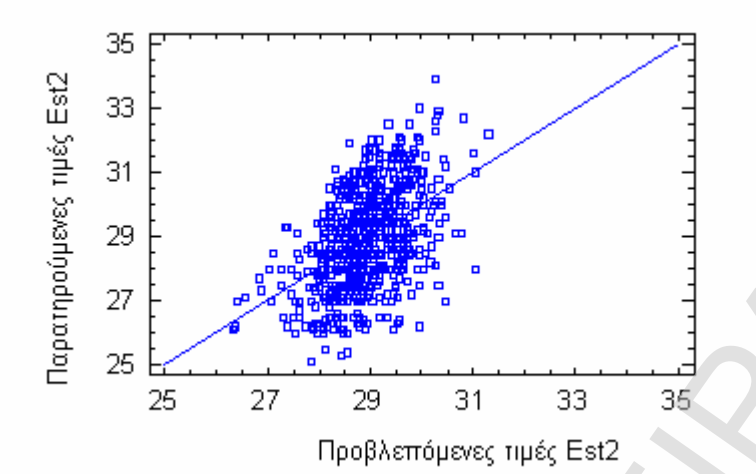
Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	82,634	3,94913	20,9246	0,0000
SiO2	-2,53221	0,187534	-13,5027	0,0000
LOI	-1,2486	0,339665	-3,67597	0,0002

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	339,366	2	169,683	92,59	0,0000
Κατάλοιπα	1189,32	649	1,83254		
Σύνολο	1528,69	651			

R-2 = 22,1998 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 21,96 τοις εκατό
Τυπικό σφάλμα = 1,35372
Μέσο απόλυτο σφάλμα = 1,10305
Durbin-Watson statistic = 1,08158 (P=0,0000)

Από τον παραπάνω πίνακα συσχετίσεων προέκυψε ότι η Est2 συσχετίζεται με τη SiO_2, αλλά όχι με τη LOI. Αυτό πιθανότατα να οφείλεται στην ύπαρξη αυτοσυσχέτισης στα κατάλοιπα, αφού το *p-value* του *Durbin-Watson test* είναι μικρότερο από 0,05. Ο συντελεστής προσδιορισμού R^2 εξηγεί το 22,1998% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 εξηγεί το 21,96%. Στο παρακάτω διάγραμμα φαίνεται η σχέση μεταξύ των παρατηρούμενων τιμών της Est2 και των τιμών που προβλέπει το μοντέλο.



Διάγραμμα 6.33: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est2-Προς τα Εμπρός Επιλογή, OPC, MT3

Ο παρακάτω πίνακας δείχνει τη συσχέτιση μεταξύ των εκτιμητριών των συντελεστών παλινδρόμησης:

Πίνακας 6.33: Πίνακας Συσχετίσεων των Εκτιμητριών των Συντελεστών Παλινδρόμησης

	CONSTANT	SiO2	LOI
CONSTANT	1,0000	-0,9957	-0,2387
SiO2	-0,9957	1,0000	0,1483
LOI	-0,2387	0,1483	1,0000

Από τον παραπάνω πίνακα φαίνεται ότι δεν υπάρχουν σημαντικές συσχετίσεις μεταξύ των εκτιμητριών των συντελεστών παλινδρόμησης.

6.4.3. Απλή Παλινδρόμηση

- ***Απλή Παλινδρόμηση μεταξύ Est28 και Est2***

Για να αποφασίσουμε ποια μορφή απλής παλινδρόμησης θα επιλέξουμε, εξετάζουμε διάφορα εναλλακτικά μοντέλα ως προς το ποιο από αυτά έχει τη μεγαλύτερη τιμή σε R^2. Τα εναλλακτικά μοντέλα απεικονίζονται στον Πίνακα 6.34.

Πίνακας 6.34: Σύγκριση Εναλλακτικών Μοντέλων Απλής Παλινδρόμησης

Μοντέλο	Συσχέτιση	R-2
Αντίστροφο ως προς Y	-0,2417	5,84%
Εκθετικό	0,2415	5,83%
Τετραγωνικής ρίζας του Y	0,2414	5,83%
Γραμμικό	0,2412	5,82%
Τετραγωνικής ρίζας του X	0,2395	5,74%
Πολλαπλασιαστικό	0,2380	5,67%
Λογαριθμικό ως προς X	0,2378	5,65%
Διπλής αντοστροφής	0,2344	5,49%
Καμπύλη S	-0,2342	5,48%
Αντίστροφο ως προς X	-0,2339	5,47%
Λογιστικό	<no fit>	
Log probit	<no fit>	

Το αντίστροφο μοντέλο ως προς τη μεταβλητή *Y* φαίνεται ότι έχει το μεγαλύτερο R^2. Η μορφή του μοντέλου αυτού είναι: *Y=1/(a+b* X)*. Συνεπώς, επιλέγεται αυτό για την ανάλυση που ακολουθεί. Στον Πίνακα 6.35 αναγράφονται τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου, ενώ η εξίσωση της απλής παλινδρόμησης είναι η ακόλουθη:

Est28 = 1/(0,0181698 - 0,0000735029*Est2)

Πίνακας 6.35: Απλή Παλινδρόμηση μεταξύ Est28 και Est2, OPC, MT3

Αντίστροφο ως προς Y Μοντέλο: Y = 1/(a + b*X)

Εξαρτημένη μεταβλητή: Est28
Ανεξάρτητη μεταβλητή: Est2

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
Σταθερός όρος	0,0181698	0,00033508	54,2254	0,0000
Κλίση	-0,0000735029	0,0000115667	-6,35469	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	0,00000828002	1	10,00000828002	40,38	0,0000
Κατάλοιπα	0,000133482	651	2,05042E-7		
Σύνολο	0,00014176	652			

Συντελεστής συσχέτισης = -0,240806
R-2 = 5,79874 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 5,69615 τοις εκατό
Τυπικό σφάλμα = 0,000453125
Μέσο απόλυτο σφάλμα = 0,000359082
Durbin-Watson statistic = 0,795859 (P=0,0000)

Το *p-value* στον πίνακα της ANOVA είναι μικρότερο από 0,01, συνεπώς υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Est28 και της Est2 στο 99% επίπεδο εμπιστοσύνης.

Το R^2 είναι 5,79874%, που δηλώνει ότι το μοντέλο που προσαρμόζεται στα δεδομένα εξηγεί κατά αυτό το ποσοστό τη μεταβλητότητα του Est28. Ο συντελεστής συσχέτισης είναι -0,240806, που δείχνει μια ασθενή σχέση μεταξύ των μεταβλητών. Το τυπικό σφάλμα της εκτίμησης είναι 0,000453125 και ισούται με την τυπική απόκλιση των καταλοίπων. Από το *Durbin-Watson (DW) statistic* προκύπτει ότι το *p-value* είναι μικρότερο από 0,05, το οποίο φανερώνει την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

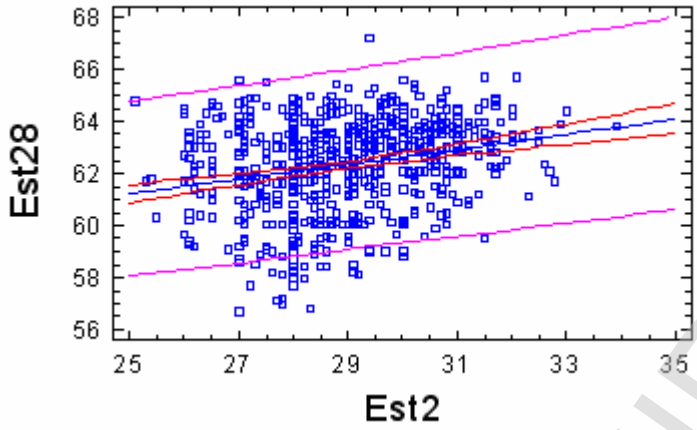
Για να ελέγξουμε αν το προτεινόμενο μοντέλο επαρκεί για να περιγράψει τις τιμές των δεδομένων, εκτελούμε το *Lack-of-Fit test*. Τα αποτελέσματα του test αυτού φαίνονται στον Πίνακα 6.36.

<u>Πίνακας 6.36</u>: Έλεγχος Έλλειψης Προσαρμογής στην Απλή Παλινδρόμηση μεταξύ Est28 και Est2, OPC, MT3

Ανάλυση Διακύμανσης με Lack-of-Fit τεστ

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	**P-Value**
Μοντέλο	0,00000828002	1	10,00000828002	40,38	0,0000
Κατάλοιπα	0,000133482	651	2,05042E-7		
Lack-of-Fit	0,0000172116	72	2,39051E-7	1,19	**0,1461**
Καθ.σφάλμα	0,000116271	579	2,00813E-7		
Σύνολο	0,000141762	652			

Καθώς το *p-value* για το lack-of-fit test στον πίνακα της ANOVA είναι μεγαλύτερο από 0,10, το μοντέλο φαίνεται να επαρκεί για τα παρατηρούμενα δεδομένα. Το διάγραμμα της σχέσης μεταξύ των μεταβλητών φαίνεται στο παρακάτω Διάγραμμα 6.34.



Διάγραμμα 6.34: Διάγραμμα Απλής Παλινδρόμησης των μεταβλητών Est28 και Est2, OPC, MT3

- ***Απλή Παλινδρόμηση μεταξύ Est28 και Est7***

Για να αποφασίσουμε ποια μορφή απλής παλινδρόμησης θα επιλέξουμε, εξετάζουμε διάφορα εναλλακτικά μοντέλα ως προς το ποιο από αυτά έχει τη μεγαλύτερη τιμή σε R^2. Τα εναλλακτικά μοντέλα απεικονίζονται στον Πίνακα 6.37.

Πίνακας 6.37: Σύγκριση Εναλλακτικών Μοντέλων Απλής Παλινδρόμησης

Μοντέλο	Συσχέτιση	R-2
Γραμμικό	0,5054	25,54%
Τετραγωνικής ρίζας του Y	0,5048	25,49%
Εκθετικό	0,5042	25,43%
Τετραγωνικής ρίζας του X	0,5035	25,35%
Αντίστροφο ως προς Y	-0,5029	25,29%
Λογαριθμικό ως προς X	0,5015	25,15%
Πολλαπλασιαστικό	0,5004	25,04%
Αντίστροφο ως προς X	-0,4970	24,70%
Καμπύλη S	-0,4960	24,60%
Διπλής αντιστροφής	0,4948	24,48%
Λογιστικό	<no fit>	
Log probit	<no fit>	

Το γραμμικό μοντέλο φαίνεται ότι έχει το μεγαλύτερο R^2. Συνεπώς, επιλέγεται αυτό για την ανάλυση που ακολουθεί. Στον Πίνακα 6.38 αναγράφονται τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου, ενώ η εξίσωση της απλής παλινδρόμησης είναι η ακόλουθη:

Est28 = 39,7667 + 0,516898*Est7

<u>Πίνακας 6.38</u>: Απλή Παλινδρόμηση μεταξύ Est28 και Est7, OPC, MT3

Γραμμικό Μοντέλο: Y = a + b*X

Εξαρτημένη μεταβλητή: Est28
Ανεξάρτητη μεταβλητή: Est7

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
Σταθερός όρος	39,7667	1,51558	26,2385	0,0000
Κλίση	0,516898	0,0346157	14,9325	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	524,453	1	524,453	222,98	0,0000
Κατάλοιπα	1528,82	650	2,35202		
Σύνολο	2053,27	651			

Συντελεστής συσχέτισης = 0,505394
R-2 = 25,5423 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 25,4278 τοις εκατό
Τυπικό σφάλμα = 1,53363
Μέσο απόλυτο σφάλμα = 1,19526
Durbin-Watson statistic = 0,814245 (P=0,0000)

Το *p-value* στον πίνακας της ANOVA είναι μικρότερο από 0,01, συνεπώς υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Est28 και της Est7 στο 99% επίπεδο εμπιστοσύνης.

Το R^2 είναι 25,5423%, που δηλώνει ότι το μοντέλο που προσαρμόζεται στα δεδομένα εξηγεί κατά αυτό το ποσοστό τη μεταβλητότητα του Est28. Ο συντελεστής συσχέτισης είναι 0,505394, που δείχνει μια σχετικά ασθενή σχέση μεταξύ των μεταβλητών. Το τυπικό σφάλμα της εκτίμησης είναι 1,53363 και ισούται με την τυπική απόκλιση των καταλοίπων. Από το *Durbin-Watson (DW) statistic* προκύπτει ότι το *p-value* είναι μικρότερο από 0,05, το οποίο φανερώνει την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

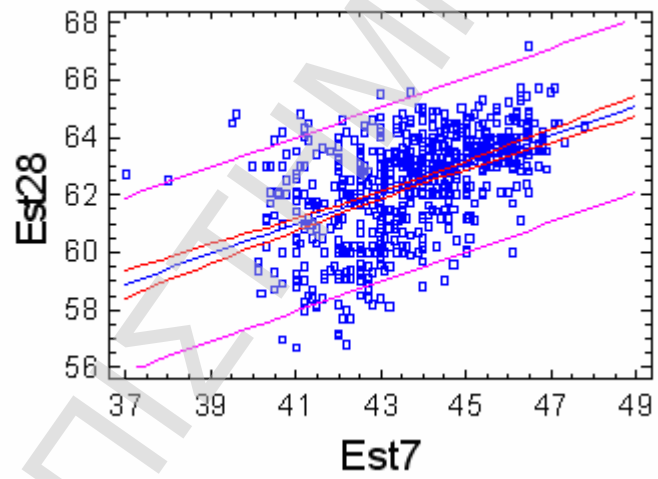
Για να ελέγξουμε αν το προτεινόμενο μοντέλο επαρκεί για να περιγράψει τις τιμές των δεδομένων, εκτελούμε το *Lack-of-Fit test*. Τα αποτελέσματα του test αυτού φαίνονται στον Πίνακα 6.39.

Πίνακας 6.39: Έλεγχος Έλλειψης Προσαρμογής στην Απλή Παλινδρόμηση μεταξύ Est28 και Est7, OPC, MT3

Ανάλυση Διακύμανσης με Lack-of-Fit τεστ

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	**P-Value**
Μοντέλο	524,453	1	524,453	222,98	0,0000
Κατάλοιπα	1528,82	650	2,35202		
Lack-of-Fit	273,452	76	3,59805	1,65	**0,0009**
Καθ.σφάλμα	1255,36	574	2,18705		
Σύνολο	2053,27	651			

Καθώς το *p-value* για το lack-of-fit test στον πίνακα της ANOVA είναι μικρότερο από 0,01, το μοντέλο φαίνεται να *μην* επαρκεί για τα παρατηρούμενα δεδομένα, για 99% επίπεδο εμπιστοσύνης. Το διάγραμμα της σχέσης μεταξύ των μεταβλητών φαίνεται στο παρακάτω Διάγραμμα 6.35. Τα εσωτερικά όρια (κόκκινες γραμμές) δηλώνουν τα 95% διαστήματα εμπιστοσύνης για το μέσο Est28 των παρατηρήσεων για τις συγκεκριμένες τιμές Est7 από τα δεδομένα. Τα εξωτερικά όρια (ροζ γραμμές) δηλώνουν τα 95% διαστήματα *πρόβλεψης* για τις νέες παρατηρήσεις.



Διάγραμμα 6.35: Διάγραμμα Απλής Παλινδρόμησης των μεταβλητών Est28 και Est7, OPC, MT3

6.4.4. Έλεγχος των Προϋποθέσεων της Παλινδρόμησης

Ο έλεγχος των προϋποθέσεων της πολλαπλής και της απλής παλινδρόμησης ως προς τα κατάλοιπα γίνεται με τον ακόλουθο τρόπο:

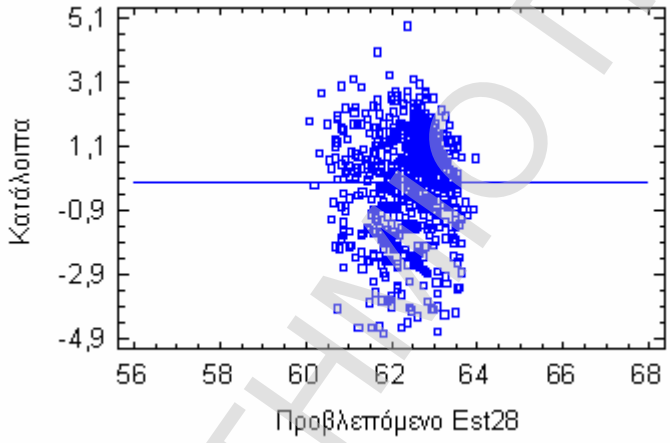
- Σχετικά με την υπόθεση για **κανονική κατανομή** των καταλοίπων και, συνεπώς, της εξαρτημένης μεταβλητής *Y*, γνωρίζουμε από τη στατιστική ανάλυση που προηγήθηκε στο αμέσως προηγούμενο κεφάλαιο ότι η μεταβλητή Est28 δεν ακολουθεί κανονική κατανομή, ενώ η Est2 ακολουθεί.

ü Όσον αφορά τη **μέση τιμή** και τη **διασπορά** των καταλοίπων, τα αποτελέσματα του ελέγχου είναι τα εξής:

§ ***Πολλαπλή Παλινδρόμηση Est28 – Προς τα Εμπρός Επιλογή***

Το MAE ισούται με 1,28276, το οποίο είναι διάφορο του μηδενός. Αυτό φαίνεται και από το διάγραμμα των καταλοίπων ως προς τις προβλεπόμενες τιμές Est28, όπου υπάρχουν περισσότερα κατάλοιπα κάτω από τη γραμμή του οριζόντιου άξονα από ό,τι πάνω από αυτήν. Σχετικά με τη διασπορά, φαίνεται από το ίδιο διάγραμμα ότι έχει παντού το ίδιο εύρος, καθώς προχωράμε δεξιά προς τον οριζόντιο άξονα.

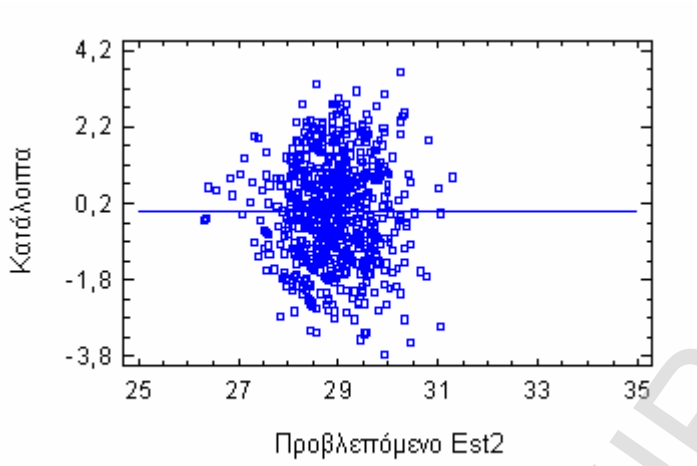
Το Διάγραμμα 6.36 απεικονίζει το διάγραμμα καταλοίπων ως προς τις προβλεπόμενες τιμές *Y*, για τον έλεγχο των προϋποθέσεων μηδενικής μέσης τιμής των σφαλμάτων και σταθερής διασποράς αυτών.



Διάγραμμα 6.36: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Πολλαπλή Παλινδρόμηση, OPC, MT3

§ ***Πολλαπλή Παλινδρόμηση Est2 – Προς τα Εμπρός Επιλογή***

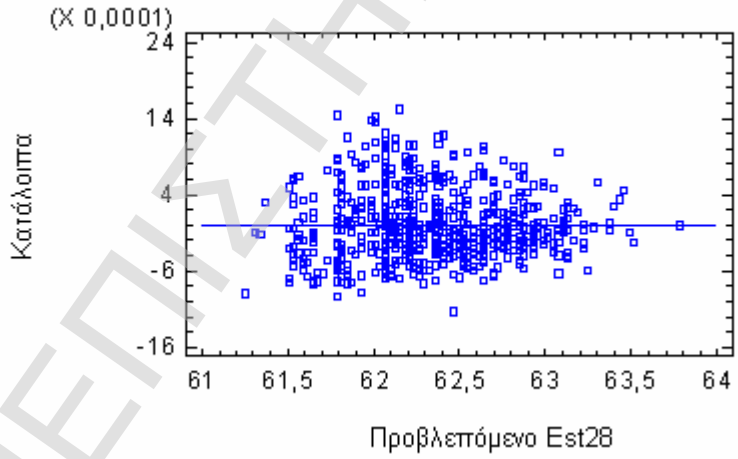
Το MAE ισούται με 1,10305, το οποίο διαφέρει από την τιμή μηδέν. Από το διάγραμμα των καταλοίπων ως προς τις προβλεπόμενες τιμές Est2 φαίνεται ότι τα κατάλοιπα είναι ομοιόμορφα κατανεμημένα πάνω και κάτω από τη γραμμή του οριζόντιου άξονα, δηλαδή ότι η μέση τιμή αυτών προσεγγίζει το μηδέν. Όσον αφορά τη διασπορά, αυτή παραμένει σταθερή για όλες τις προβλεπόμενες τιμές του Est2.



Διάγραμμα 6.37: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est2, Πολλαπλή Παλινδρόμηση, OPC, MT3

§ ***Απλή Παλινδρόμηση Est28 και Est2***

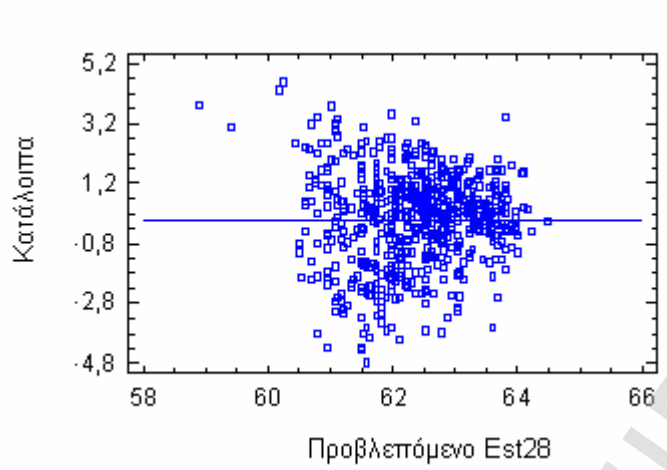
Στην απλή παλινδρόμηση, το MAE ισούται με 0,000359082, το οποίο προσεγγίζει πάρα πολύ την τιμή μηδέν. Στο διάγραμμα, όμως, των καταλοίπων ως προς τις προβλεπόμενες τιμές Est28, δεν φαίνεται τα κατάλοιπα να είναι ομοιόμορφα κατανεμημένα γύρω από τον οριζόντιο άξονα. Επίσης, η διασπορά δεν φαίνεται να παραμένει σταθερή, το οποίο πιθανότατα να υποδηλώνει πρόβλημα ετεροσκεδαστικότητας.



Διάγραμμα 6.38: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Απλή Παλινδρόμηση Est28 και Est2, OPC, MT3

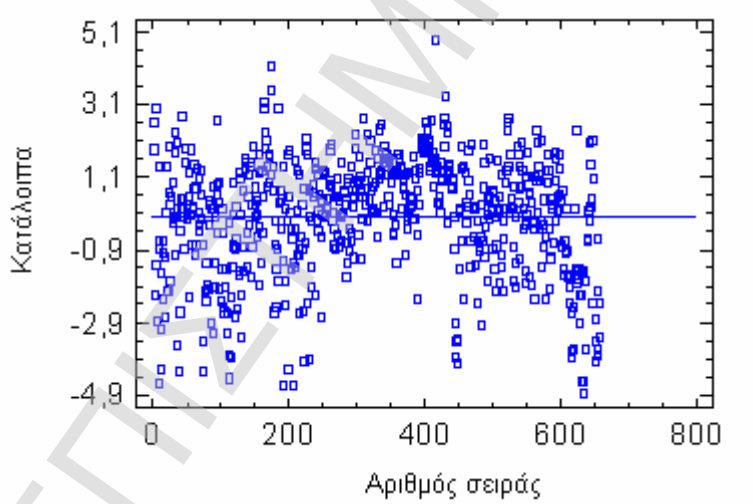
§ ***Απλή Παλινδρόμηση Est28 και Est7***

Στην περίπτωση αυτή, το MAE ισούται με 1,19526, το οποίο δεν προσεγγίζει πάρα πολύ την τιμή μηδέν. Στο διάγραμμα, όμως, των καταλοίπων ως προς τις προβλεπόμενες τιμές Est28, φαίνεται ότι τα κατάλοιπα είναι ομοιόμορφα κατανεμημένα γύρω από τον οριζόντιο άξονα. Επίσης, η διασπορά φαίνεται να παραμένει σταθερή.

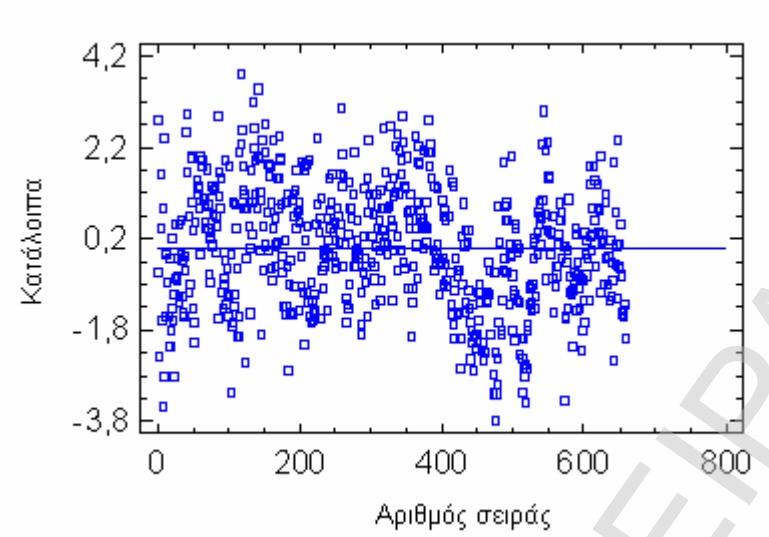


Διάγραμμα 6.39: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Απλή Παλινδρόμηση Est28 και Est7, OPC, MT3

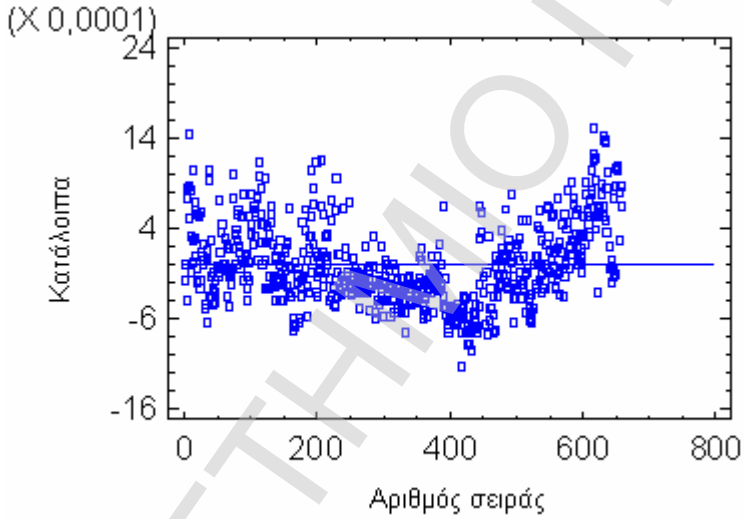
- Σχετικά με την ύπαρξη **αυτοσυσχέτισης** μεταξύ των καταλοίπων, προκύπτει ότι τα *p-values* των Durbin-Watson tests και για τις τρεις περιπτώσεις παλινδρόμησης είναι μικρότερα από 0,05. Άρα, υπάρχει αυτοσυσχέτιση μεταξύ των καταλοίπων. Αυτό γίνεται αντιληπτό και από τα διαγράμματα, τα οποία παρουσιάζονται παρακάτω.



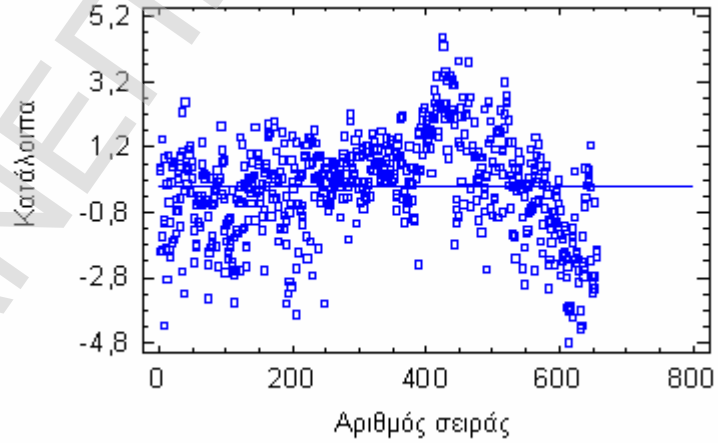
Διάγραμμα 6.40: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Πολλαπλή Παλινδρόμηση της Est28, OPC, MT3



Διάγραμμα 6.41: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Πολλαπλή Παλινδρόμηση της Est2, OPC, MT3



Διάγραμμα 6.42: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Απλή Παλινδρόμηση μεταξύ των Est28 και Est2, OPC, MT3



Διάγραμμα 6.43: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Απλή Παλινδρόμηση μεταξύ των Est28 και Est7, OPC, MT3

6.5. Απλό Τσιμέντο Portland OPC – Μύλος Παραγωγής 4

6.5.1. Έλεγχος Συσχετίσεων μεταξύ των Μεταβλητών

Παρακάτω παρουσιάζεται ο πίνακας συσχετίσεων κατά Spearman, αφού οι περισσότερες μεταβλητές δεν είναι κανονικά κατανεμημένες.

Πίνακας 6.40: Πίνακας Συσχετίσεων των Μεταβλητών κατά Spearman

	SiO_2	Al_2O_3	Blaine	LOI	Est2	log(Est7)	Est28
SiO_2		**-0,1041** (442) **0,0288**	**0,2362** (442) **0,0000**	**0,1311** (442) **0,0059**	**-0,4664** (442) **0,0000**	**-0,5449** (442) **0,0000**	**-0,1710** (442) **0,0003**
Al_2O_3	**-0,1041** (442) **0,0288**		**-0,3579** (442) **0,0000**	**-0,1920** (442) **0,0001**	**0,0944** (442) **0,0475**	**0,1736** (442) **0,0003**	**0,3948** (442) **0,0000**
Blaine	**0,2362** (442) **0,0000**	**-0,3579** (442) **0,0000**		**0,2037** (442) **0,0000**	-0,0066 (442) 0,8901	**-0,1058** (442) **0,0263**	**-0,2596** (442) **0,0000**
LOI	**0,1311** (442) **0,0059**	**-0,1920** (442) **0,0001**	**0,2037** (442) **0,0000**		**-0,1756** (442) **0,0002**	**-0,1605** (442) **0,0008**	**-0,1470** (442) **0,0020**
Est2	**-0,4664** (442) **0,0000**	**0,0944** (442) **0,0475**	-0,0066 (442) 0,8901	**-0,1756** (442) **0,0002**		**0,8217** (442) **0,0000**	**0,3169** (442) **0,0000**
log(Est7)	**-0,5449** (442) **0,0000**	**0,1736** (442) **0,0003**	**-0,1058** (442) **0,0263**	**-0,1605** (442) **0,0008**	**0,8217** (442) **0,0000**		**0,4466** (442) **0,0000**
Est28	**-0,1710** (442) **0,0003**	**0,3948** (442) **0,0000**	**-0,2596** (442) **0,0000**	**-0,1470** (442) **0,0020**	**0,3169** (442) **0,0000**	**0,4466** (442) **0,0000**	

Μεταβλητές με *p-values* μικρότερες από 0,05 φανερώνουν στατιστικά σημαντικές μη-μηδενικές συσχετίσεις, σε 95 % επίπεδο εμπιστοσύνης. Τα ζεύγη των μεταβλητών που έχουν *p-values* κάτω από 0,05 είναι τα ακόλουθα:

SiO2 και **Al2O3**
SiO2 και **Blaine**
SiO2 και **LOI**
SiO2 και **Est2**
SiO2 και **LOG(Est7)**
SiO2 και **Est28**
Al2O3 και **Blaine**
Al2O3 και **LOI**
Al2O3 και **Est2**
Al2O3 και **LOG(Est7)**
Al2O3 και **Est28**
Blaine και **LOI**
Blaine και **LOG(Est7)**
Blaine και **Est28**
LOI και **Est2**
LOI και **LOG(Est7)**

LOI και **Est28**
Est2 και **LOG(Est7)**
Est2 και **Est28**
LOG(Est7) και **Est28**

Πολλές ανεξάρτητες μεταβλητές συσχετίζονται μεταξύ τους, συνεπώς υπάρχει πρόβλημα πολυσυγγραμικότητας. Γι' αυτό, στο μοντέλο παλινδρόμησης θα εφαρμοστεί και η μέθοδος της προς τα εμπρός επιλογής.

Η εξαρτημένη μεταβλητή Est28 φαίνεται από τον παραπάνω πίνακα ότι συσχετίζεται με αρκετές μεταβλητές, όπως τις SiO_2, Al2O3, Blaine, LOI, Est2 και log(Est7). Άρα, στην πολλαπλή παλινδρόμηση περιμένουμε ότι κάποιες από τις παραπάνω μεταβλητές θα είναι παρούσες, ενώ μπορούμε να εφαρμόσουμε ένα μοντέλο απλής παλινδρόμησης με τη μεταβλητή Est2 και τη log(Est7). Όσον αφορά την εξαρτημένη μεταβλητή Est2, παρατηρούμε ότι συσχετίζεται με τις μεταβλητές SiO_2, Al_2O_3 και LOI.

6.5.2. Πολλαπλή Παλινδρόμηση

- ***Εξαρτημένη Μεταβλητή Est28 – Όλες οι Μεταβλητές***

Το μοντέλο παλινδρόμησης που προκύπτει είναι το εξής:

Est28 = 49,7472 - 0,435043*SiO2 + 5,94733*Al2O3 - 0,00164097*Blaine - 0,639201*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.41.

Πίνακας 6.41: Πολλαπλή Παλινδρόμηση για Est28-Όλες οι Μεταβλητές, OPC, MT4

Εξαρτημένη μεταβλητή: Est28

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	49,7472	8,19822	6,06805	0,0000
SiO2	-0,435043	0,287813	-1,51155	0,1314
Al2O3	5,94733	0,648991	9,16395	0,0000
Blaine	-0,00164097	0,0012729	-1,28916	0,1980
LOI	-0,639201	0,52201	-1,2245	0,2214

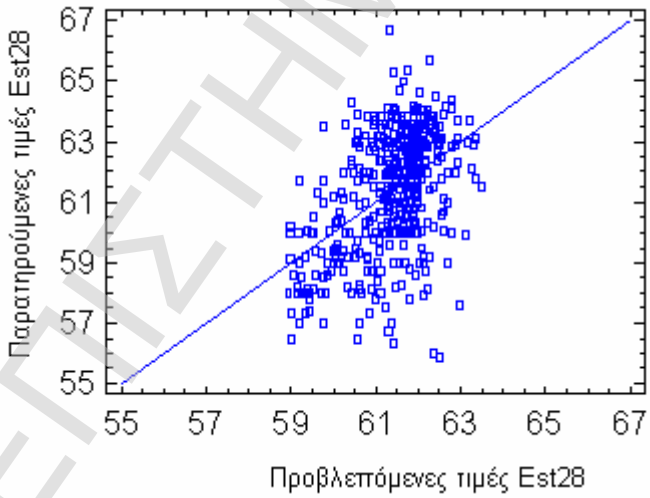
Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	375,865	4	93,9662	31,14	0,0000
Κατάλοιπα	1318,46	437	3,01707		
Σύνολο	1694,32	441			

R-2 = 22,1838 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 21,4715 τοις εκατό
Τυπικό σφάλμα = 1,73697
Μέσο απόλυτο σφάλμα = 1,37146
Durbin-Watson statistic = 1,32599 (P=0,0000)

Ο συντελεστής προσδιορισμού R^2 δείχνει ότι το προσαρμοσμένο μοντέλο εξηγεί 22,1838% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 είναι 21,4715%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 1,73697. Το μέσο απόλυτο σφάλμα (MAE) είναι 1,37146 και αποτελεί τη μέση τιμή των καταλοίπων. Το *Durbin-Watson (DW) statistic* ελέγχει τα κατάλοιπα για το αν υπάρχει αυτοσυσχέτιση. Επειδή το *p-value* είναι μικρότερο από 0,05, υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής Est28 σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές.



<u>Διάγραμμα 6.44</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est28-Όλες οι Μεταβλητές, OPC, MT4

- ***<u>Εξαρτημένη Μεταβλητή Est28 – Προς τα Εμπρός Επιλογή</u>***

Το μοντέλο παλινδρόμησης που προκύπτει είναι το εξής:

Est28 = 31,3154 + 6,49115*Al2O3

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων είναι αυτά που φαίνονται στον Πίνακα 6.42.

<u>Πίνακας 6.42</u>: Πολλαπλή Παλινδρόμηση για Est28-Προς τα Εμπρός Επιλογή, OPC, MT4

Εξαρτημένη μεταβλητή: Est28

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	31,3154	2,78276	11,2534	0,0000
Al2O3	6,49115	0,601013	10,8003	0,0000

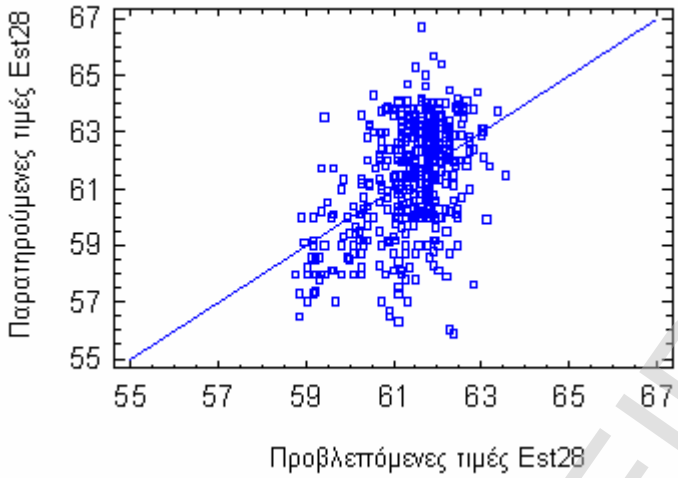
Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	355,052	1	355,052	116,65	0,0000
Κατάλοιπα	1339,27	440	3,0438		
Σύνολο	1694,32	441			

R-2 = 20,9554 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 20,7757 τοις εκατό
Τυπικό σφάλμα = 1,74465
Μέσο απόλυτο σφάλμα = 1,38825
Durbin-Watson statistic = 1,31498 (P=0,0000)

Η μεταβλητή η Al_2O_3 που προκρίνεται στο μοντέλο συσχετίζεται με την Est28, όπως φαίνεται στον αρχικό πίνακα συσχετίσεων. Ο συντελεστής προσδιορισμού R^2 δείχνει ότι το προσαρμοσμένο μοντέλο εξηγεί 20,9554% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 είναι 20,7757%. Το τυπικό σφάλμα της εκτίμησης δείχνει ότι η τυπική απόκλιση των καταλοίπων είναι 1,74465. Το μέσο απόλυτο σφάλμα (MAE) είναι 1,38825 και αποτελεί τη μέση τιμή των καταλοίπων. Το Durbin-Watson statistic ελέγχει τα κατάλοιπα για το αν υπάρχει αυτοσυσχέτιση. Επειδή το *p-value* είναι μικρότερο από 0,05, υπάρχει ένδειξη αυτοσυσχέτισης.

Το παρακάτω διάγραμμα δείχνει τις παρατηρούμενες τιμές της μεταβλητής Est28 σε σχέση με τις προβλεπόμενες από το μοντέλο τιμές. Όσο πιο κοντά τα δεδομένα "πέφτουν" στη διαγώνια γραμμή, τόσο καλύτερο είναι το μοντέλο στην πρόβλεψη των παρατηρούμενων δεδομένων. Στη συγκεκριμένη περίπτωση οι παρατηρούμενες τιμές αποκλίνουν αρκετά από την ευθεία γραμμή, άρα το μοντέλο δεν θεωρείται ιδιαίτερα καλό.



<u>Διάγραμμα 6.45</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est28-Προς τα Εμπρός Επιλογή, OPC, MT4

- ***<u>Εξαρτημένη Μεταβλητή Est2 –Όλες οι Μεταβλητές</u>***

Το μοντέλο παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής Est2 και των ανεξάρτητων μεταβλητών είναι το εξής:

Est2 = 53,3072 - 1,95586*SiO2 + 0,975766*Al2O3 + 0,00304356*Blaine - 0,96334*LOI

Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων φαίνονται στον Πίνακα 6.43.

<u>Πίνακας 6.43</u>: Πολλαπλή Παλινδρόμηση για Est2-Όλες οι Μεταβλητές, OPC, MT4

Εξαρτημένη μεταβλητή: Est2

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	53,3072	5,11353	10,4247	0,0000
SiO2	-1,95586	0,179519	-10,895	0,0000
Al2O3	0,975766	0,4048	2,41049	0,0163
Blaine	0,00304356	0,000793955	3,83341	0,0001
LOI	-0,96334	0,325597	-2,95869	0,0033

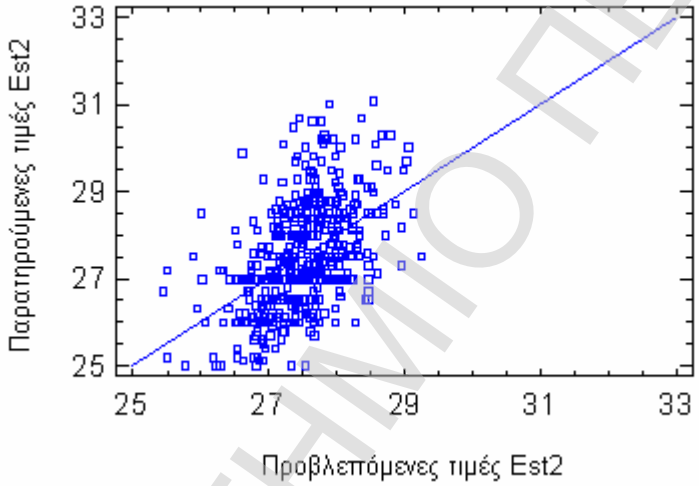
Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	160,578	4	40,1444	34,20	0,0000
Κατάλοιπα	512,944	437	1,17378		
Σύνολο	673,522	441			

R-2 = 23,8415 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 23,1444 τοις εκατό
Τυπικό σφάλμα = 1,08341
Μέσο απόλυτο σφάλμα = 0,868805
Durbin-Watson statistic = 1,18183 (P=0,0000)

Ο συντελεστής προσδιορισμού R^2 εξηγεί το 23,8415% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 εξηγεί το 23,1444%. Το *p-value* στο *Durbin-Watson test* είναι μικρότερο από 0,05, άρα υπάρχει αυτοσυσχέτιση μεταξύ των καταλοίπων.

Στο παρακάτω διάγραμμα φαίνεται η σχέση μεταξύ των παρατηρούμενων τιμών της Est2 και των τιμών που προβλέπει το μοντέλο.



<u>Διάγραμμα 6.46</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est2-Όλες οι Μεταβλητές, OPC, MT4

- ***<u>Εξαρτημένη Μεταβλητή Est2 – Προς τα Εμπρός Επιλογή</u>***

Το μοντέλο παλινδρόμησης μεταξύ της εξαρτημένης μεταβλητής Est2 και των ανεξάρτητων μεταβλητών που προκύπτει από τη βηματική μέθοδο επιλογής (forward selection) των μεταβλητών είναι το εξής:

Est2 = 53,3072 - 1,95586*SiO2 + 0,975766*Al2O3 + 0,00304356*Blaine - 0,96334*LOI

Το παραπάνω μοντέλο επιλέγει κάποιες μεταβλητές με τη μέθοδο της προς τα εμπρός επιλογής. Παρατηρούμε, ωστόσο, ότι όλες οι μεταβλητές προκρίνονται, δηλαδή το μοντέλο είναι όμοιο με το αμέσως προηγούμενο. Τα χαρακτηριστικά του μοντέλου και οι τιμές των βασικών παραμέτρων φαίνονται στον Πίνακα 6.44.

<u>Πίνακας 6.44</u>: Πολλαπλή Παλινδρόμηση για Est2-Προς τα Εμπρός Επιλογή, OPC, MT4

Εξαρτημένη μεταβλητή: Est2

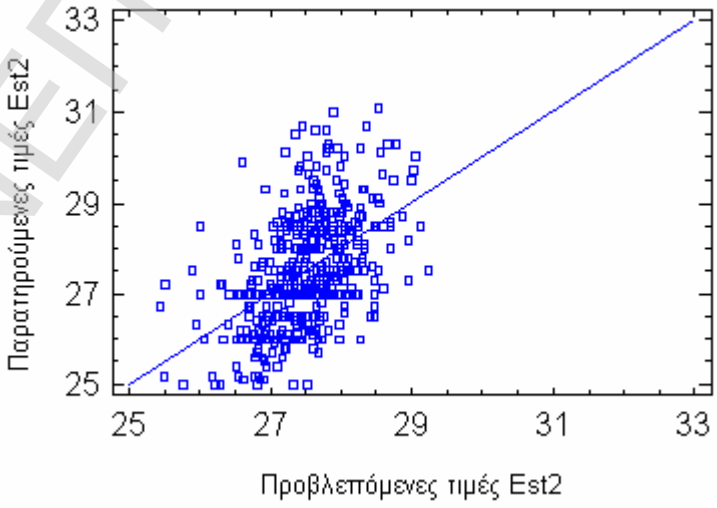
Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
CONSTANT	53,3072	5,11353	10,4247	0,0000
SiO2	-1,95586	0,179519	-10,895	0,0000
Al2O3	0,975766	0,4048	2,41049	0,0163
Blaine	0,00304356	0,000793955	3,83341	0,0001
LOI	-0,96334	0,325597	-2,95869	0,0033

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	160,578	4	40,1444	34,20	0,0000
Κατάλοιπα	512,944	437	1,17378		
Σύνολο	673,522	441			

R-2 = 23,8415 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 23,1444 τοις εκατό
Τυπικό σφάλμα = 1,08341
Μέσο απόλυτο σφάλμα = 0,868805
Durbin-Watson statistic = 1,18183 (P=0,0000)

Από τον παραπάνω πίνακα συσχετίσεων προέκυψε ότι η Est2 συσχετίζεται με όλες τις μεταβλητές που προκρίνονται στο παραπάνω μοντέλο, εκτός από τη Blaine. Ο συντελεστής προσδιορισμού R^2 εξηγεί το 23,8415% της μεταβλητότητας στη μεταβλητή Est28, ενώ ο διορθωμένος συντελεστής R^2 εξηγεί το 23,1444%. Στο παρακάτω διάγραμμα φαίνεται η σχέση μεταξύ των παρατηρούμενων τιμών της Est2 και των τιμών που προβλέπει το μοντέλο. Το *p-value* στο *Durbin-Watson test* είναι μικρότερο από 0,05, άρα υπάρχει αυτοσυσχέτιση μεταξύ των καταλοίπων.



<u>Διάγραμμα 6.47</u>: Διάγραμμα Πολλαπλής Παλινδρόμησης της μεταβλητής Est2-Προς τα Εμπρός Επιλογή, OPC, MT4

Από τον παρακάτω πίνακα των συσχετίσεων μεταξύ των εκτιμητριών των συντελεστών παλινδρόμησης προκύπτει ότι δεν υπάρχουν σημαντικές συσχετίσεις, που θα μπορούσαν να οδηγήσουν στην εφαρμογή της *ραχοειδούς παλινδρόμησης*.

Πίνακας 6.45: Πίνακας Συσχετίσεων των Εκτιμητριών των Συντελεστών Παλινδρόμησης

	CONSTANT	SiO2	Al2O3	Blaine	LOI
CONSTANT	1,0000	-0,6304	-0,5841	-0,5585	-0,0538
SiO2	-0,6304	1,0000	0,0213	-0,1807	0,0038
Al2O3	-0,5841	0,0213	1,0000	0,3309	0,1359
Blaine	-0,5585	-0,1807	0,3309	1,0000	-0,1331
LOI	-0,0538	0,0038	0,1359	-0,1331	1,0000

6.5.3. Απλή Παλινδρόμηση

- ***Απλή Παλινδρόμηση μεταξύ Est28 και Est2***

Οι μεταβλητές **Est28** και **Est2** από τον πίνακα συσχετίσεων φάνηκε ότι συσχετίζονται ασθενώς. Παρόλα αυτά, επιχειρείται η δημιουργία ενός μοντέλου απλής παλινδρόμησης μεταξύ των δύο μεταβλητών.

Για να αποφασίσουμε ποια μορφή απλής παλινδρόμησης θα επιλέξουμε, εξετάζουμε διάφορα εναλλακτικά μοντέλα ως προς το ποιο από αυτά έχει τη μεγαλύτερη τιμή σε R^2. Τα εναλλακτικά μοντέλα απεικονίζονται στον Πίνακα 6.46.

Πίνακας 6.46: Σύγκριση Εναλλακτικών Μοντέλων Απλής Παλινδρόμησης

Μοντέλο	Συσχέτιση	R-2
Γραμμικό	0,3143	9,88%
Τετραγωνικής ρίζας του X	0,3137	9,84%
Τετραγωνικής ρίζας του Y	0,3134	9,82%
Λογαριθμικό ως προς X	0,3130	9,80%
Εκθετικό	0,3125	9,76%
Αντίστροφο ως προς X	-0,3114	9,69%
Πολλαπλασιαστικό	0,3112	9,68%
Αντίστροφο ως προς Y	-0,3105	9,64%
Καμπύλη S	-0,3096	9,58%
Διπλής αντιστροφής	0,3077	9,47%
Λογιστικό	<no fit>	
Log probit	<no fit>	

Το γραμμικό μοντέλο έχει το μεγαλύτερο R^2, συνεπώς επιλέγεται αυτό για την ανάλυση που ακολουθεί. Στον Πίνακα 6.47 αναγράφονται τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου, ενώ η εξίσωση της απλής παλινδρόμησης είναι η ακόλουθη:

Est28 = 47,6529 + 0,498547*Est2

<u>Πίνακας 6.47</u>: Απλή Παλινδρόμηση μεταξύ Est28 και Est2, OPC, MT4

Γραμμικό Μοντέλο: Y = a + b* X

Εξαρτημένη μεταβλητή: Est28
Ανεξάρτητη μεταβλητή: Est2

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
Σταθερός όρος	47,6529	1,97506	24,1274	0,0000
Κλίση	0,498547	0,0717805	6,94545	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	167,403	1	167,403	48,24	0,0000
Κατάλοιπα	1526,92	440	3,47028		
Σύνολο	1694,32	441			

Συντελεστής συσχέτισης = 0,314329
R-2 = 9,88024 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 9,67542 τοις εκατό
Τυπικό σφάλμα = 1,86287
Μέσο απόλυτο σφάλμα = 1,53724
Durbin-Watson statistic = 1,03203 (P=0,0000)

Το *p-value* στον πίνακα της ANOVA είναι μικρότερο από 0,01, συνεπώς υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Est28 και της Est2 στο 99% επίπεδο εμπιστοσύνης.

Το R^2 είναι 9,88024%, που δηλώνει ότι το μοντέλο που προσαρμόζεται στα δεδομένα εξηγεί κατά αυτό το ποσοστό τη μεταβλητότητα του Est28. Ο συντελεστής συσχέτισης είναι 0,314329, που δείχνει μια ασθενή σχέση μεταξύ των μεταβλητών. Το τυπικό σφάλμα της εκτίμησης είναι 1,86287 και ισούται με την τυπική απόκλιση των καταλοίπων. Από το *Durbin-Watson (DW) statistic* προκύπτει ότι το *p-value* είναι μικρότερο από 0,05, το οποίο φανερώνει την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

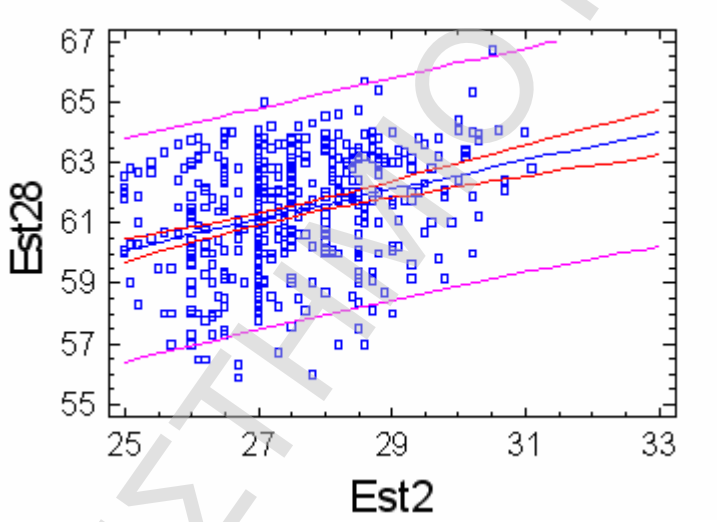
Για να ελέγξουμε αν το προτεινόμενο μοντέλο επαρκεί για να περιγράψει τις τιμές των δεδομένων, εκτελούμε το *Lack-of-Fit test*. Τα αποτελέσματα του test αυτού φαίνονται στον Πίνακα 6.48.

Πίνακας 6.48: Έλεγχος Έλλειψης Προσαρμογής στην Απλή Παλινδρόμηση μεταξύ Est28 και Est2, OPC, MT4

Ανάλυση Διακύμανσης με Lack-of-Fit τεστ

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	**P-Value**
Μοντέλο	167,403	1	167,403	48,24	0,0000
Κατάλοιπα	1526,92	440	3,47028		
Lack-of-Fit	210,106	54	3,89085	1,14	**0,2416**
Καθ.σφάλμα	1316,82	386	3,41144		
Σύνολο	1694,32	441			

Καθώς το *p-value* για το lack-of-fit test στον πίνακα της ANOVA είναι μεγαλύτερο από 0,10, το μοντέλο φαίνεται να επαρκεί για τα παρατηρούμενα δεδομένα. Το διάγραμμα της σχέσης μεταξύ των μεταβλητών φαίνεται στο παρακάτω Διάγραμμα 6.48.



Διάγραμμα 6.48: Διάγραμμα Απλής Παλινδρόμησης των μεταβλητών Est28 και Est2, OPC, MT4

- ***Απλή Παλινδρόμηση μεταξύ Est28 και log(Est7)***

Οι μεταβλητές **Est28** και **log(Est7)** από τον πίνακα συσχετίσεων φάνηκε ότι συσχετίζονται σε κάποιο βαθμό, γι' αυτό επιχειρείται η δημιουργία ενός μοντέλου απλής παλινδρόμησης μεταξύ των δύο μεταβλητών.

Για να αποφασίσουμε ποια μορφή απλής παλινδρόμησης θα επιλέξουμε, εξετάζουμε διάφορα εναλλακτικά μοντέλα ως προς το ποιο από αυτά έχει τη μεγαλύτερη τιμή σε R^2. Τα εναλλακτικά μοντέλα απεικονίζονται στον Πίνακα 6.49.

Πίνακας 6.49: Σύγκριση Εναλλακτικών Μοντέλων Απλής Παλινδρόμησης

Μοντέλο	Συσχέτιση	R-2
Γραμμικό	0,4387	19,24%
Τετραγωνικής ρίζας του X	0,4383	19,21%
Λογαριθμικό ως προς X	0,4380	19,18%
Αντίστροφο ως προς X	-0,4373	19,12%
Τετραγωνικής ρίζας του Y	0,4371	19,11%
Εκθετικό	0,4356	18,98%
Πολλαπλασιαστικό	0,4350	18,92%
Καμπύλη S	-0,4343	18,86%
Αντίστροφο ως προς Y	-0,4324	18,70%
Διπλής αντιστροφής	0,4312	18,59%
Λογιστικό	<no fit>	
Log probit	<no fit>	

Το γραμμικό μοντέλο έχει το μεγαλύτερο R^2 και επιλέγεται αυτό για την ανάλυση που ακολουθεί. Στον Πίνακα 6.50 αναγράφονται τα βασικά χαρακτηριστικά και αποτελέσματα του μοντέλου, ενώ η εξίσωση της απλής παλινδρόμησης είναι η ακόλουθη:

Est28 = - 20,6971 + 21,9221*log(Est7)

Πίνακας 6.50: Απλή Παλινδρόμηση μεταξύ Est28 και log(Est7), OPC, MT4

Γραμμικό Μοντέλο: Y = a + b* X

Εξαρτημένη μεταβλητή: Est28
Ανεξάρτητη μεταβλητή: log(Est7)

Παράμετρος	Εκτίμηση	Τυπικό Σφάλμα	T Statistic	P-Value
Σταθερός όρος	-20,6971	8,01443	-2,58248	0,0101
Κλίση	21,9221	2,14108	10,2388	0,0000

Ανάλυση Διακύμανσης

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	P-Value
Μοντέλο	326,012	1	326,012	104,83	0,0000
Κατάλοιπα	1368,31	440	3,1098		
Σύνολο	1694,32	441			

Συντελεστής συσχέτισης = 0,43865
R-2 = 19,2414 τοις εκατό
R-2 (προσαρμοσμένο στους β.ε.) = 19,0579 τοις εκατό
Τυπικό σφάλμα = 1,76346
Μέσο απόλυτο σφάλμα = 1,43126
Durbin-Watson statistic = 1,04092 (P=0,0000)

Το *p-value* στον πίνακα της ANOVA είναι μικρότερο από 0,01, συνεπώς υπάρχει μια στατιστικά σημαντική σχέση μεταξύ της Est28 και της log(Est7) στο 99% επίπεδο εμπιστοσύνης.

Το R^2 είναι 19,2414%, που δηλώνει ότι το μοντέλο που προσαρμόζεται στα δεδομένα εξηγεί κατά αυτό το ποσοστό τη μεταβλητότητα του Est28. Ο συντελεστής συσχέτισης είναι 0,43865, που δείχνει μια ασθενή σχέση μεταξύ των μεταβλητών. Το τυπικό σφάλμα της εκτίμησης είναι 1,76346 και ισούται με την τυπική απόκλιση των καταλοίπων. Από το *Durbin-Watson (DW) statistic* προκύπτει ότι το *p-value* είναι μικρότερο από 0,05, το οποίο φανερώνει την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

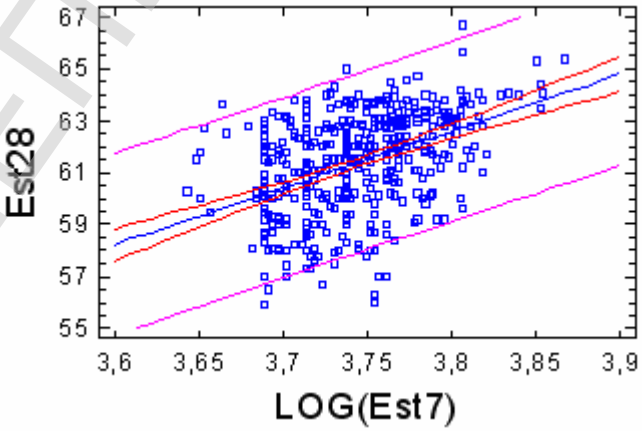
Για να ελέγξουμε αν το προτεινόμενο μοντέλο επαρκεί για να περιγράψει τις τιμές των δεδομένων, εκτελούμε το *Lack-of-Fit test*. Τα αποτελέσματα του test αυτού φαίνονται στον Πίνακα 6.51.

<u>Πίνακας 6.51</u>: Έλεγχος Έλλειψης Προσαρμογής στην Απλή Παλινδρόμηση μεταξύ Est28 και log(Est7), OPC, MT4

Ανάλυση Διακύμανσης με Lack-of-Fit τεστ

	Άθροισμα Τετραγώνων	Β.ε.	Μέσο Άθροισμα Τετραγώνων	F-Ratio	**P-Value**
Μοντέλο	326,012	1	326,012	104,83	0,0000
Κατάλοιπα	1368,31	440	3,1098		
Lack-of-Fit	241,295	72	3,35131	1,09	**0,2948**
Καθ.σφάλμα	1127,02	368	3,06255		
Σύνολο	1694,32	441			

Καθώς το *p-value* για το lack-of-fit test στον πίνακα της ANOVA είναι μεγαλύτερο από 0,10, το μοντέλο φαίνεται να επαρκεί για τα παρατηρούμενα δεδομένα. Το διάγραμμα της σχέσης μεταξύ των μεταβλητών φαίνεται στο παρακάτω Διάγραμμα 6.49.



<u>Διάγραμμα 6.49</u>: Διάγραμμα Απλής Παλινδρόμησης των μεταβλητών Est28 και log(Est7),OPC, MT4

6.5.4. Έλεγχος των Προϋποθέσεων της Παλινδρόμησης

Ο έλεγχος των προϋποθέσεων της πολλαπλής και της απλής παλινδρόμησης ως προς τα κατάλοιπα γίνεται με τον ακόλουθο τρόπο:

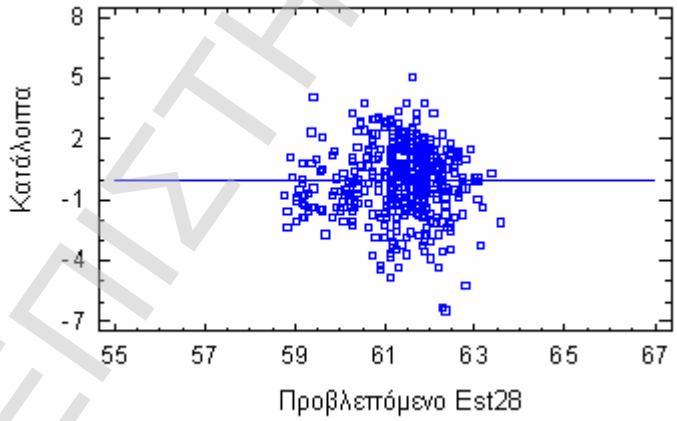
- Σχετικά με την υπόθεση για **κανονική κατανομή** των καταλοίπων και, συνεπώς, της εξαρτημένης μεταβλητής *Y*, γνωρίζουμε από τη στατιστική ανάλυση που προηγήθηκε στο αμέσως προηγούμενο κεφάλαιο ότι καμία από τις μεταβλητές Est28 και Est2 δεν ακολουθεί κανονική κατανομή.

- Όσον αφορά τη **μέση τιμή** και τη **διασπορά** των καταλοίπων, τα αποτελέσματα του ελέγχου είναι τα εξής:

  - ***Πολλαπλή Παλινδρόμηση Est28 – Προς τα Εμπρός Επιλογή***

    Το μέσο απόλυτο σφάλμα ισούται με 1,38825, το οποίο είναι διάφορο του μηδενός. Αυτό φαίνεται και από το διάγραμμα των καταλοίπων ως προς τις προβλεπόμενες τιμές Est28, όπου φαίνεται να υπάρχουν περισσότερα κατάλοιπα κάτω από τη γραμμή του οριζόντιου άξονα από ό,τι πάνω από αυτήν. Σχετικά με τη διασπορά, φαίνεται από το ίδιο διάγραμμα ότι δεν παραμένει σταθερή, άρα πιθανώς να υπάρχει πρόβλημα ετεροσκεδαστικότητας.

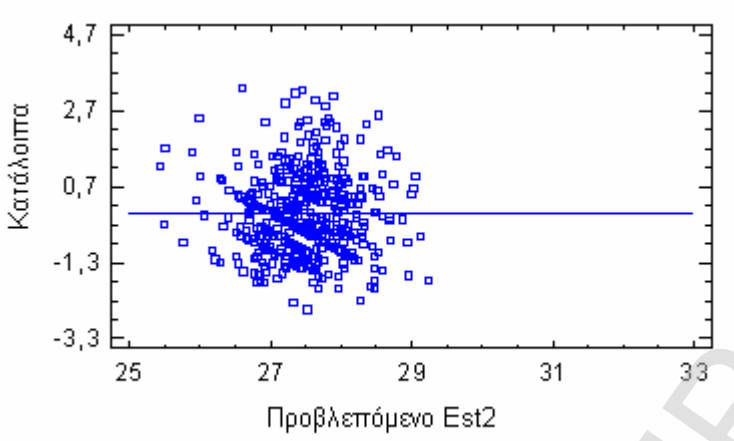
    Το Διάγραμμα 6.50 απεικονίζει το διάγραμμα καταλοίπων ως προς τις προβλεπόμενες τιμές *Y*, για τον έλεγχο των προϋποθέσεων μηδενικής μέσης τιμής των σφαλμάτων και σταθερής διασποράς αυτών.



Διάγραμμα 6.50: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Πολλαπλή Παλινδρόμηση, OPC, MT4

  - ***Πολλαπλή Παλινδρόμηση Est2 – Προς τα Εμπρός Επιλογή***

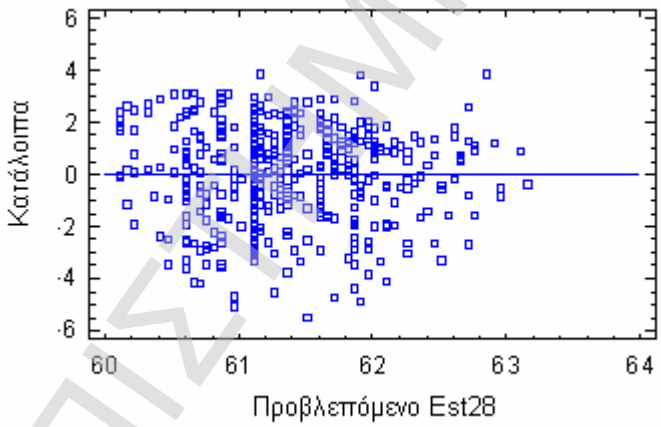
    Το MAE ισούται με 0,868805, το οποίο διαφέρει λίγο από την τιμή μηδέν. Από το διάγραμμα των καταλοίπων ως προς τις προβλεπόμενες τιμές Est2 δεν φαίνεται ιδιαίτερα ότι τα κατάλοιπα δεν είναι ομοιόμορφα κατανεμημένα πάνω και κάτω από τη γραμμή του οριζόντιου άξονα. Όσον αφορά τη διασπορά, αυτή φαίνεται να παραμένει σχετικά σταθερή για όλες τις προβλεπόμενες τιμές του Est2.



Διάγραμμα 6.51: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est2, Πολλαπλή Παλινδρόμηση, OPC, MT4

§ *Απλή Παλινδρόμηση Est28 και Est2*

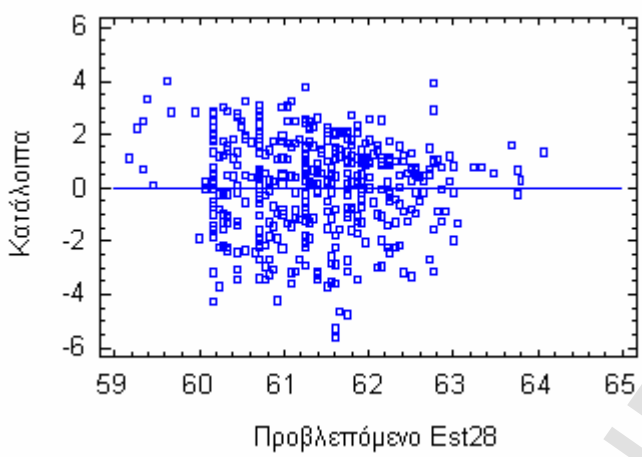
Στην απλή παλινδρόμηση, το MAE ισούται με 1,53724, το οποίο διαφέρει από την τιμή μηδέν. Αυτό φαίνεται και στο Διάγραμμα 6.52. Η διασπορά δεν φαίνεται να παραμένει σταθερή, καθώς μειώνεται όσο προχωράμε προς τον οριζόντιο άξονα.



Διάγραμμα 6.52: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Απλή Παλινδρόμηση Est28 και Est2, OPC, MT4

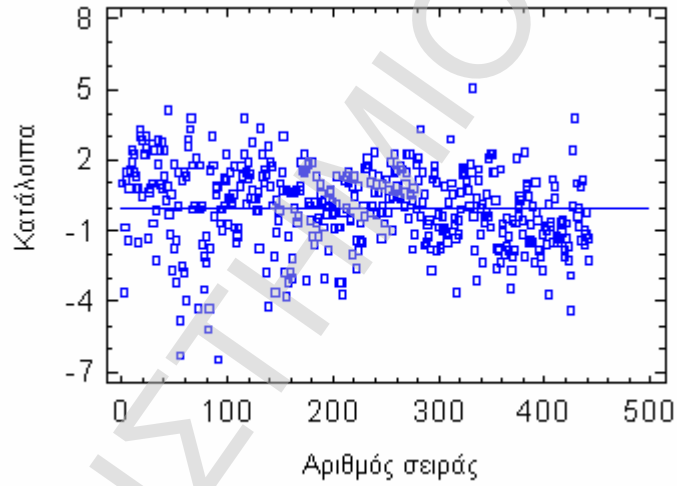
§ *Απλή Παλινδρόμηση Est28 και log(Est7)*

Σε αυτή την απλή παλινδρόμηση, το MAE ισούται με 1,43126, το οποίο διαφέρει από την τιμή μηδέν. Στο διάγραμμα των καταλοίπων ως προς τις προβλεπόμενες τιμές Est28 φαίνεται μια ελαφριά απόκλιση της μέσης τιμής των σφαλμάτων από την τιμή μηδέν. Η διασπορά δεν φαίνεται να παραμένει σταθερή, καθώς μειώνεται όσο προχωράμε προς τον οριζόντιο άξονα.

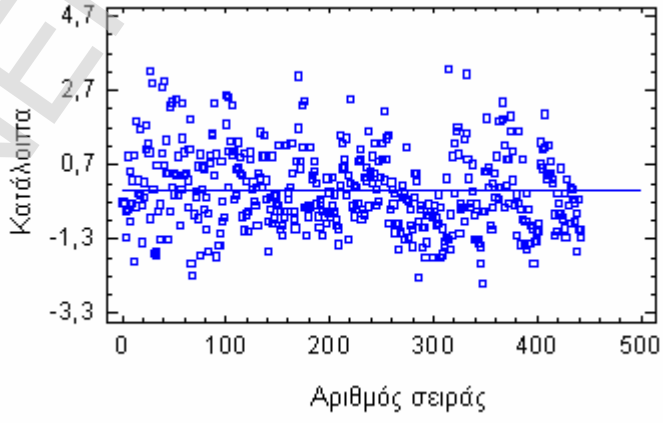


Διάγραμμα 6.53: Διάγραμμα Καταλοίπων ως προς τις Προβλεπόμενες τιμές Est28, Απλή Παλινδρόμηση Est28 και log(Est7), OPC, MT4

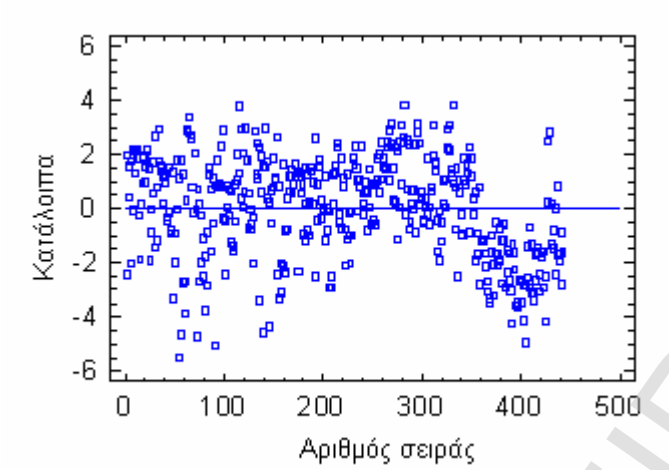
- Σχετικά με την ύπαρξη **αυτοσυσχέτισης** μεταξύ των καταλοίπων, προκύπτει ότι τα p-values και για τις τρεις περιπτώσεις παλινδρόμησης είναι μικρότερα από 0,05. Άρα, υπάρχει αυτοσυσχέτιση μεταξύ των καταλοίπων. Αυτό γίνεται αντιληπτό και από τα διαγράμματα, τα οποία παρουσιάζονται παρακάτω.



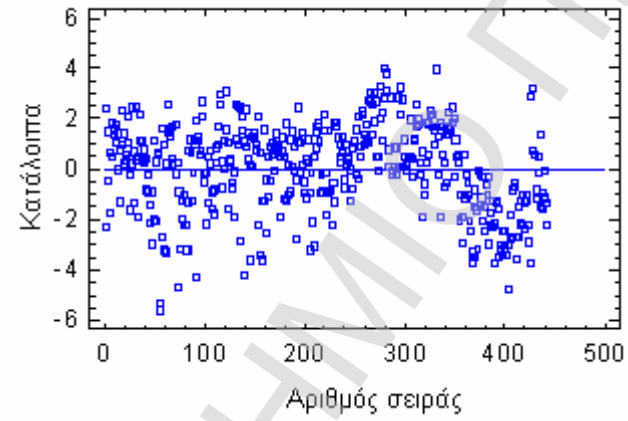
Διάγραμμα 6.54: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Πολλαπλή Παλινδρόμηση της Est28, OPC, MT4



Διάγραμμα 6.55: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Πολλαπλή Παλινδρόμηση της Est2, OPC, MT4



Διάγραμμα 6.56: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Απλή Παλινδρόμηση μεταξύ των Est28 και Est2, OPC, MT4



Διάγραμμα 6.57: Διάγραμμα Καταλοίπων ως προς τον Αριθμό Σειράς για την Απλή Παλινδρόμηση μεταξύ των Est28 και log(Est7), OPC, MT4

6.6. Σύνοψη Αποτελεσμάτων

Από την προηγούμενη ανάλυση παλινδρόμησης προέκυψαν κάποια μοντέλα πολλαπλής παλινδρόμησης και κάποια απλής παλινδρόμησης, για κάθε τύπο τσιμέντου και για κάθε μύλο παραγωγής ξεχωριστά. Οι μεταβλητές που συμπεριλαμβάνονται σε κάθε μοντέλο πολλαπλής παλινδρόμησης, καθώς και οι τιμές των συντελεστών παλινδρόμησης αυτών, φαίνονται στον Πίνακα 6.52. Στον ίδιο πίνακα φαίνονται επίσης και κάποια σημαντικά στατιστικά μέτρα, όπως ο διορθωμένος συντελεστής προσδιορισμού (R^2 προσαρμοσμένος στους βαθμούς ελευθερίας), το τυπικό σφάλμα της εκτίμησης και το μέσο απόλυτο σφάλμα (Mean Absolute Error). Επιθυμητό είναι το R^2 να είναι όσο το δυνατόν μεγαλύτερο, ενώ το τυπικό σφάλμα και το μέσο απόλυτο σφάλμα όσο το δυνατόν μικρότερα και κοντά στο μηδέν.

<u>Πίνακας 6.52</u>: Πολλαπλή Παλινδρόμηση των Εξαρτημένων Μεταβλητών Est28 και Est2-Όλες οι Μεταβλητές και Προς τα Εμπρός Επιλογή

	CEMII 42,5 - MT1				CEMII 42,5 - MT4			
	Est28		log(Est2)		Est28		Est2	
Συντελ.	Όλες οι μεταβλητές	Προς τα Εμπρός επιλογή	Όλες οι μεταβλητές	Προς τα Εμπρός επιλογή	Όλες οι μεταβλητές	Προς τα Εμπρός επιλογή	Όλες οι μεταβλητές	Προς τα Εμπρός επιλογή
Σταθερό	0,13235	8,02767	2,09804	2,39995	48,9894	58,3727	19,2662	24,0587
SiO2	0,917126		0,0125107		-1,54905	-1,48953	-0,711632	
Al2O3	-1,79699		-0,102093	-0,0970642	6,27521	6,63408	2,02924	
Blaine	0,0109734	0,00941782	0,000373577	0,000353906	0,00321422		0,00322085	
IR	-0,293312		-0,0032725		-0,026708		0,046854	
LOI					-1,26153	-1,1151	-0,753215	
log(LOI)	-7,6051		-0,312761	-0,275916				
R^2 (β.ε.)	20,09	11,9165	26,8174	27,0632	13,273	13,2014	2,04172	0
Τυπικό Σφάλμα	1,99137	2,09073	0,0652495	0,0651398	1,63336	1,63403	1,3265	1,34025
Μέσο Απόλυτο Σφάλμα	1,54666	1,71213	0,0524657	0,0524958	1,26916	1,28002	1,01186	1,06042

	OPC - MT3				OPC - MT4			
	Est28		Est2		Est28		Est2	
Συντελ.	Όλες οι μεταβλητές	Προς τα Εμπρός επιλογή	Όλες οι μεταβλητές	Προς τα Εμπρός επιλογή	Όλες οι μεταβλητές	Προς τα Εμπρός επιλογή	Όλες οι μεταβλητές	Προς τα Εμπρός επιλογή
Σταθερό	67,6419	70,9219	78,5401	82,634	49,7472	31,3154	53,3072	53,3072
SiO2	0,149859		-2,52516	-2,53221	-0,435043		-1,95586	-1,95586
Al2O3	2,71224	2,69899	0,316477		5,94733	6,49115	0,975766	0,975766
Blaine	-0,00513764	-0,00515807	0,000665548		-0,00164097		0,00304356	0,00304356
LOI	-1,81537	-1,8558	-1,24825	-1,2486	-0,639201		-0,96334	-0,96334
R^2 (β.ε.)	16,0551	16,1275	21,8204	21,96	21,4715	20,7757	23,1444	23,1444
Τυπικό Σφάλμα	1,62716	1,62646	1,35493	1,35372	1,73697	1,74465	1,08341	1,08341
Μέσο Απόλυτο Σφάλμα	1,28224	1,28276	1,10337	1,10305	1,37146	1,38825	0,868805	0,868805

Στον παρακάτω Πίνακα 6.53 φαίνονται οι τιμές των συντελεστών a και b των μοντέλων απλής παλινδρόμησης μεταξύ των Est28 και Est2 ή Est7, καθώς και κάποια άλλα βασικά στατιστικά μέτρα (R^2, τυπικό σφάλμα, συντελεστής συσχέτισης), που βοηθούν στην αξιολόγηση του μοντέλου.

Πίνακας 6.53: Απλή Παλινδρόμηση μεταξύ Est28-Est2 και Est28-Est7

	CEMII 42,5 - MT1		CEMII 42,5 - MT4	
	Est28-log(Est2)	Est28-Est7	Est28-Est2	Est28-Est7
ΜΟΝΤΕΛΟ	**Y=a+b/X**	**Y=a+b*X**	**Y=a+b*sqrt(X)**	**Y=a+b*X**
Σταθερά a	122,989	23,7877	23,748	23,6077
Συντελεστής b	-223,43	0,752171	6,05875	0,765856
Συντελεστής Συσχέτισης	-0,759636	0,758305	0,471401	0,613586
R^2	57,7047	57,5026	22,2219	37,6487
Τυπικό Σφάλμα	1,45459	1,45807	1,55328	1,39074
Μέσο Απόλυτο Σφάλμα	1,1737	1,17707	1,22735	1,10171
	OPC – MT3		**OPC - MT4**	
	Est28-Est2	Est28-Est7	Est28-Est2	Est28-log(Est7)
ΜΟΝΤΕΛΟ	**Y=1/(a+b*X)**	**Y=a+b*X**	**Y=a+b*X**	**Y=a+b*X**
Σταθερά a	0,0181698	39,7667	47,6529	-20,6971
Συντελεστής b	-0,0000735029	0,516898	0,498547	21,9221
Συντελεστής Συσχέτισης	-0,240806	0,505394	0,314329	0,43865
R^2	5,79874	25,5423	9,88024	19,2414
Τυπικό Σφάλμα	0,000453125	1,53363	1,86287	1,76346
Μέσο Απόλυτο Σφάλμα	0,000359082	1,19526	1,53724	1,43126

ΚΕΦΑΛΑΙΟ 7: ΣΥΜΠΕΡΑΣΜΑΤΑ ΚΑΙ ΠΡΟΤΑΣΕΙΣ ΓΙΑ ΠΕΡΑΙΤΕΡΩ ΕΡΕΥΝΑ

7.1. Σύνοψη της Μεθοδολογίας της Εργασίας

Στο δεύτερο μέρος των αποτελεσμάτων της παρούσας εργασίας έγινε αρχικά μία στατιστική ανάλυση των βασικών χαρακτηριστικών των μεταβλητών που μελετώνται. Εξετάστηκε ο μέσος όρος, η διασπορά, η τυπική απόκλιση, η μέγιστη και η ελάχιστη τιμή και οι τιμές της προτυποποιημένης ασυμμετρίας και κύρτωσης. Ιδιαίτερη σημασία έχουν οι δύο τελευταίες παράμετροι, διότι αποτελούν το βασικό κριτήριο για τον έλεγχο ύπαρξης κανονικότητας. Συγκεκριμένα, αν οι τιμές αυτές είναι μεταξύ των τιμών [-2,2], τότε η μεταβλητή κατανέμεται κανονικά. Έγιναν, ωστόσο, και κάποιοι οπτικοί έλεγχοι για την κανονικότητα μέσω κάποιων διαγραμμάτων, όπως το ιστόγραμμα, το διάγραμμα ίχνους της πυκνότητας, το διάγραμμα ελέγχου της κανονικότητας και το θηκόγραμμα. Ειδικά για τις μεταβλητές Est2 και Est28 έγινε και έλεγχος ύπαρξης κανονικότητας μέσω του X^2 *τεστ* και του ***Shapiro-Wilks*** *τεστ*. Τα τεστ αυτά σε μερικές περιπτώσεις δείχνουν διαφορετικά αποτελέσματα από τον έλεγχο που γίνεται με το διάστημα τιμών ασυμμετρίας και κύρτωσης. Σε αυτές τις περιπτώσεις, δεν απορρίπτουμε την ύπαρξη κανονικότητας, αλλά διατηρούμε μια επιφυλακτικότητα για την ισχύ της υπόθεσης αυτής.

Ο έλεγχος της κανονικότητας αποτελεί καθοριστικό σημείο για μια μελέτη, καθώς η έλλειψη αυτής ακυρώνει την ισχύ των στατιστικών τεστ που λαμβάνουν υπόψη την τυπική απόκλιση. Έτσι, για παράδειγμα, τα διαστήματα εμπιστοσύνης χάνουν κατά κάποιο βαθμό την εγκυρότητά τους, όταν η μεταβλητή δεν είναι κανονικά κατανεμημένη. Ιδιαίτερη σημαντικότητα έχει η έλλειψη κανονικότητας στις μεταβλητές Est2 και Est28, επειδή αυτές αποτελούν τις εξαρτημένες μεταβλητές της πολλαπλής και απλής παλινδρόμησης. Η ύπαρξη κανονικότητας στις εξαρτημένες μεταβλητές και στα κατάλοιπα σε ένα μοντέλο παλινδρόμησης αποτελεί μία από τις βασικές προϋποθέσεις μιας τέτοιας ανάλυσης.

Σε ορισμένες από τις μεταβλητές που δεν ακολουθούσαν κανονική κατανομή, έγιναν κάποιοι μετασχηματισμοί σε συναρτήσεις που είναι λιγότερο ευαίσθητες στις μακρινές και έκτροπες παρατηρήσεις. Οι πιο συνηθισμένες τέτοιες συναρτήσεις είναι η λογαριθμική (log$X$), η τετραγωνική ρίζα (SQRTX) και η αντίστροφη (1/X). Με τους μετασχηματισμούς αυτούς διατηρείται η αρχική πληροφορία που περιέχει κάθε μεταβλητή, ενώ ταυτόχρονα αποκτούν εγκυρότητα τα συνήθη στατιστικά τεστ. Για παράδειγμα, μετασχηματισμοί στη συγκεκριμένη εργασία έγιναν στην περίπτωση του τσιμέντου CEM II 42,5 στο MT1, όπου χρησιμοποιήθηκε η λογαριθμική συνάρτηση για τις μεταβλητές LOI και Est2. Και σε άλλες περιπτώσεις η λογαριθμική συνάρτηση φάνηκε να "συγκεντρώνει" περισσότερο όλες τις παρατηρήσεις μεταξύ τους, αλλά όσο και αν πέτυχε μια καλύτερη προσέγγιση στην κανονική κατανομή, ο έλεγχος των τιμών της ασυμμετρίας και κύρτωσης δεν επαλήθευσαν το απαιτούμενο κριτήριο (τιμές μεταξύ [-2,2]). Για το λόγο αυτό, οι μεταβλητές δεν μετασχηματίστηκαν, αλλά διατηρήθηκαν στην αρχική τους μορφή.

Τα δεδομένα που χρησιμοποιήθηκαν για την ανάλυση ήταν αρχικά για το χρονικό διάστημα από τις αρχές του 2003 έως τον Αύγουστο του 2005, και για τους δύο τύπους τσιμέντου. Βρέθηκε, όμως, για το CEM II 42,5 (και στους δύο μύλους) ότι υπάρχουν δύο διαφορετικοί πληθυσμοί μέσα στο δοθέν σετ δεδομένων. Αυτό έγινε αντιληπτό από το διάγραμμα διασποράς, όπου τα δεδομένα ήταν κατανεμημένα ανομοιόμορφα, σχηματίζοντας δύο ανεξάρτητες ομάδες δεδομένων στο καρτεσιανό επίπεδο. Επίσης, από το ιστόγραμμα φάνηκε ότι υπάρχουν δύο κορυφές, δηλαδή στην ουσία δύο μέσοι διαφορετικών πληθυσμών. Οι παραπάνω διαφοροποιήσεις στα διαγράμματα φάνηκαν κυρίως στην επεξεργασία των μεταβλητών SiO_2 και LOI. Μετά από προσεκτική μελέτη των τιμών των δεδομένων στο αρχείο Excel, φάνηκε ότι οι τιμές των δύο παραπάνω μεταβλητών άλλαξαν αισθητά από την 1/7/2004. Εξακριβώθηκε ότι την ημερομηνία αυτή η χημική βιομηχανία άλλαξε παραγωγή στο συγκεκριμένο τύπο τσιμέντου, συνεπώς χρησιμοποιήθηκαν τα δεδομένα από 1/7/2004 και μετά. Μία τέτοια αλλαγή στην παραγωγή δεν παρατηρήθηκε για το OPC, γι' αυτό και τα δεδομένα που χρησιμοποιήθηκαν εκεί είναι για όλη την περίοδο των 2,5 ετών.

Μία άλλη διαδικασία που πραγματοποιήθηκε στην παραπάνω ανάλυση είναι ο εντοπισμός και η απομάκρυνση πιθανών εκτρόπων παρατηρήσεων. Οι έκτροπες παρατηρήσεις είναι πολύ πιθανό ως και βέβαιο ότι θα υπάρχουν σε ένα μεγάλο σετ δεδομένων, καθώς πολλά απρόοπτα πράγματα μπορούν να συμβούν κατά την παραγωγική διαδικασία. Οι συνθήκες αλλάζουν από κάποιες επιθυμητές ή ανεπιθύμητες αιτίες, και αυτό μπορεί να αλλάξει τις τιμές των μετρήσεων. Για παράδειγμα, μπορεί τα μηχανήματα ή τα όργανα μέτρησης να πάθουν κάποια βλάβη κάποια στιγμή, και συνεπώς να δίνουν λάθος μετρήσεις μέχρι να διορθωθούν. Στη συγκεκριμένη ανάλυση, μερικές τιμές δεδομένων που απείχαν πάνω από τρεις φορές την απόσταση του ενδοτεταρτημοριακού εύρους απομακρύνθηκαν από τα δεδομένα, με την έγκριση των ανθρώπων της τσιμεντοβιομηχανίας. Μετά την απομάκρυνσή τους φάνηκαν αμέσως κάποιες σημαντικές αλλαγές, όπως η ύπαρξη κανονικότητας και η αλλαγή του μέσου και της διασποράς.

Στο τσιμέντο CEM II 42,5 υπάρχει κανονικότητα σχεδόν σε όλες τις μεταβλητές, και στους δύο μύλους παραγωγής. Οι μεταβλητές, όμως, Clk και Gyp απέχουν πολύ από την κανονικότητα. Οι μεταβλητές αυτές φαίνεται ότι ελέγχονται καλά, ώστε να παίρνουν κάποιες συγκεκριμένες τιμές. Στον τύπο τσιμέντου OPC σχεδόν όλες οι μεταβλητές *δεν* ακολουθούν κανονική κατανομή. Ένας πιθανός λόγος που συμβαίνει αυτό είναι ότι, λόγω του πολύ μεγάλου διαστήματος των 2,5 ετών που συγκεντρώθηκαν τα δεδομένα, οι τιμές εμφανίζονται συνολικά πολύ συγκεντρωμένες, καθώς δεν διακρίνονται κάποιες τυχαίες διακυμάνσεις που δημιουργούν την κανονική κατανομή σε μεμονωμένα χρονικά διαστήματα. Επίσης, γενικά και στους δύο τύπους τσιμέντου, η λήψη δεδομένων κατά άτακτα χρονικά διαστήματα και η εναλλαγή στη χρήση των μύλων παραγωγής μπορεί να οφείλεται σε ένα βαθμό για τη *μη* ύπαρξη κανονικότητας. Τέλος, κατά ομολογία των υπευθύνων της βιομηχανίας, το τσιμέντο OPC έχει λιγότερα κύρια συστατικά από το CEM II 42,5 -το οποίο έχει και ποζολάνες-, οπότε μπορεί να ελεγχθεί καλύτερα.

Στη συνέχεια της επεξεργασίας των δεδομένων αναπτύχθηκαν κάποια μοντέλα παλινδρόμησης. Συγκεκριμένα, εφαρμόστηκε πολλαπλή παλινδρόμηση για τους δύο τύπους τσιμέντου και για κάθε μύλο παραγωγής αυτών, λαμβάνοντας σε κάθε περίπτωση ως εξαρτημένη μεταβλητή τις τελικές αντοχές του τσιμέντου στις 28 ημέρες και τις πρόωρες αντοχές του τσιμέντου στις 2 ημέρες. Για κάθε μία από τις

εξαρτημένες μεταβλητές εφαρμόστηκαν δύο μέθοδοι παλινδρόμησης: η μέθοδος που περιλαμβάνει όλες τις ανεξάρτητες μεταβλητές, και η μέθοδος της προς τα εμπρός επιλογής, κατά την οποία επιλέγονται να συμπεριληφθούν στο μοντέλο μόνο οι μεταβλητές που είναι στατιστικά σημαντικές. Επίσης, υπολογίστηκαν και κάποια μοντέλα απλής παλινδρόμησης, για όλες τις περιπτώσεις τσιμέντων και μύλων παραγωγής που αναφέρθηκαν παραπάνω. Στα μοντέλα αυτά, επιχειρείται η εύρεση ενός μοντέλου πρόβλεψης των τελικών αντοχών του τσιμέντου στις 28 ημέρες (εξαρτημένη μεταβλητή), όταν είναι γνωστές οι αντοχές του τσιμέντου στις 2 ή στις 7 ημέρες (ανεξάρτητες μεταβλητές).

Στην πολλαπλή παλινδρόμηση δεν περιλαμβάνονται οι μεταβλητές **Clk** και **Gyp**. Οι μεταβλητές αυτές ελέγχονται σε αρκετά μεγάλο βαθμό, έτσι ώστε οι τιμές τους να είναι συγκεντρωμένες γύρω από κάποια τιμή. Συνεπώς, αφού η μεταβλητότητά τους είναι περιορισμένη και ελεγχόμενη, δεν λαμβάνονται υπόψη ως ανεξάρτητες μεταβλητές.

Αρχικά στην ανάλυση παλινδρόμησης έγινε έλεγχος συσχετίσεων μεταξύ των ζευγών όλων των μεταβλητών, εξαρτημένων και ανεξαρτήτων. Από τον πίνακα των συσχετίσεων ελέγχθηκε πρώτα η εξαρτημένη μεταβλητή, αν και με ποιες μεταβλητές συσχετίζεται, έτσι ώστε να προκύψει μια πρώτη κατεύθυνση για το ποιες μεταβλητές πιθανότατα θα συμπεριληφθούν στο μοντέλο παλινδρόμησης. Αυτό βέβαια δεν συνέβη σε όλες τις περιπτώσεις, καθώς το γεγονός ότι υπάρχει συσχέτιση μεταξύ κάποιων μεταβλητών δεν σημαίνει απαραίτητα ότι υπάρχει και ερμηνευτικότητα ή αιτιότητα (causality), δηλαδή ότι η μία μεταβλητή επηρεάζει την άλλη. Κατά δεύτερον, ελέγχθηκαν οι συσχετίσεις μεταξύ των ανεξάρτητων μεταβλητών. Σε περίπτωση που υπήρξαν τέτοιες συσχετίσεις, δηλαδή σε όλες τις περιπτώσεις, εφαρμόστηκε η μέθοδος της προς τα εμπρός επιλογής.

Στο επόμενο βήμα της ανάλυσης παλινδρόμησης δημιουργήθηκαν τα δύο μοντέλα πολλαπλής παλινδρόμησης των Est28 και Est2, αρχικά περιλαμβάνοντας όλες τις ανεξάρτητες μεταβλητές και έπειτα εφαρμόζοντας τη μέθοδο της προς τα εμπρός επιλογής. Στο μοντέλο της τελευταίας μεθόδου, ελέγχθηκε πάλι η ύπαρξη συσχέτισης μεταξύ των μεταβλητών που προκρίθηκαν, έτσι ώστε αν αυτό συμβαίνει, να εφαρμοστεί η ραχοειδής παλινδρόμηση. Αυτό συνέβη μόνο στην περίπτωση του μοντέλου πολλαπλής παλινδρόμησης της εξαρτημένης μεταβλητής Est28, στον τύπο CEM II 42,5 στο MT4. Στη συνέχεια, δημιουργήθηκαν τα δύο μοντέλα απλής παλινδρόμησης μεταξύ της εξαρτημένης Est28 και της ανεξάρτητης Est2 και Est7. Στην περίπτωση αυτή, επιλέχθηκε από ένα πλήθος μοντέλων απλής παλινδρόμησης το πλέον κατάλληλο για τα δεδομένα μας, με κριτήριο την υψηλότερη τιμή του συντελεστή προσδιορισμού R^2.

Στο τελευταίο βήμα της ανάλυσης παλινδρόμησης έγινε ο έλεγχος για την ισχύ των προϋποθέσεων της παλινδρόμησης. Έλεγχος έγινε μόνο στην πολλαπλή παλινδρόμηση με τη μέθοδο της προς τα εμπρός επιλογής (καθώς αυτή θεωρείται η πιο έγκυρη μέθοδος), και στις δύο περιπτώσεις απλής παλινδρόμησης. Επίσης, ελέγχθηκε η ύπαρξη κανονικότητας στα κατάλοιπα, που στην ουσία ισοδυναμεί με την ύπαρξη κανονικότητας στην εξαρτημένη μεταβλητή. Τέλος, μέσα από τα διαγράμματα των καταλοίπων ως προς την εξαρτημένη και τις ανεξάρτητες μεταβλητές ελέγχθηκε αν το μέσο σφάλμα είναι μηδέν και αν η διασπορά είναι σταθερή, ενώ, από τα διαγράμματα των καταλοίπων ως προς τον αριθμό σειράς, και

κυρίως από το τεστ ***Durbin-Watson***, ελέγχθηκε η ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων.

Από τα αποτελέσματα της πολλαπλής και της απλής παλινδρόμησης προκύπτουν τα πλέον σημαντικά συμπεράσματα της παρούσας μελέτης, και γι' αυτό αναλύονται διεξοδικότερα στην παρακάτω ενότητα.

7.2. Κύρια Συμπεράσματα της Εργασίας

Ø Από την ανάλυση παλινδρόμησης προέκυψαν κάποια μοντέλα πολλαπλής παλινδρόμησης και κάποια απλής παλινδρόμησης, για κάθε τύπο τσιμέντου και για κάθε μύλο παραγωγής ξεχωριστά. Από τα μοντέλα αυτά μπορούμε να συμπεράνουμε ποιες μεταβλητές είναι οι πιο σημαντικές για κάθε τύπο τσιμέντου και κάθε μύλο παραγωγής, ελέγχοντας τις μεταβλητές που προκρίνονται με τη μέθοδο της προς τα εμπρός επιλογής. Έτσι, για τον τύπο τσιμέντου **CEM II 42,5** στο **MT1** η μεταβλητή Blaine φαίνεται να είναι η πλέον σημαντική για τις τελικές αντοχές του τσιμέντου στις 28 ημέρες. Από την άποψη της χημείας του τσιμέντου, αυτό είναι λογικό, καθώς η λεπτότητα (Blaine) του τσιμέντου επηρεάζει σε μεγάλο βαθμό την τελική αντοχή αυτού σε θλίψη. Όσον αφορά τις πρώιμες αντοχές των 2 ημερών, σημαντικές φαίνονται οι μεταβλητές Al_2O_3, Blaine και log(LOI). Στο **MT4** του ίδιου τύπου τσιμέντου, για τις αντοχές Est28 σημαντικές φαίνονται οι μεταβλητές SiO_2, Al_2O_3 και LOI, ενώ για τις αντοχές Est2 καμία μεταβλητή δεν φαίνεται αρκετά σημαντική. Αυτό δεν είναι απόλυτα λογικό από τη χημεία του τσιμέντου, ωστόσο υπάρχουν κάποιοι λόγοι εξήγησης του φαινομένου. Πρώτα από όλα, ο μύλος παραγωγής 4 έχει διαφορετική τεχνολογία από το μύλο 1, γι' αυτό μπορεί άλλες μεταβλητές (όπως κάποια λειτουργικά χαρακτηριστικά) να παίζουν πιο σημαντικό ρόλο. Δεύτερον, εξαιτίας της ύπαρξης αυτοσυσχέτισης μεταξύ των καταλοίπων, πολλές μεταβλητές που δεν είναι στατιστικά σημαντικές φαίνεται ότι είναι, αφήνοντας εκτός μοντέλου άλλες σημαντικές μεταβλητές (π.χ. Blaine).

Στον τύπο τσιμέντου **OPC** στο **MT3** οι αντοχές Est28 φαίνεται να επηρεάζονται ιδιαίτερα από τις μεταβλητές Al_2O_3, Blaine και LOI, ενώ οι αντοχές Est2 από τις SiO_2 και LOI. Στο **MT4** οι αντοχές Est28 επηρεάζονται από τη μεταβλητή Al_2O_3, ενώ οι αντοχές Est2 από όλες τις μεταβλητές. Και σε αυτό τον τύπο τσιμέντου η λεπτότητα (Blaine) δεν προκρίνεται σε όλες τις περιπτώσεις, το οποίο ερμηνεύεται πιθανότατα από την αυτοσυσχέτιση των καταλοίπων.

Κάτι άλλο που παρατηρούμε είναι ότι οι αντοχές Est28 δεν εμφανίζουν την ίδια συμπεριφορά με τις αντοχές Est2. Όσο και αν αυτό φαίνεται παράξενο, καθώς πρόκειται για το ίδιο δείγμα τσιμέντου σε διαφορετικό χρονικό διάστημα, δείχνει πόσο πολύπλοκο είναι το τσιμέντο σχετικά με την ανάπτυξη αντοχών, καθώς επηρεάζεται από πάρα πολλούς παράγοντες (ορυκτολογική σύσταση, λειτουργικά χαρακτηριστικά σε όλες τις φάσεις της παραγωγικής διαδικασίας). Για παράδειγμα, από μερικές μελέτες έχει βρεθεί ότι τα αλκάλια της τροφοδοσίας της καμίνου αυξάνουν τις αρχικές, αλλά μειώνουν τις τελικές αντοχές του τσιμέντου.

Από τις τιμές του συντελεστή προσδιορισμού R^2 συμπεραίνουμε, ότι η μεταβλητότητα των ανεξάρτητων μεταβλητών σε **όλα** τα μοντέλα πολλαπλής

παλινδρόμησης δεν εξηγεί σε ικανοποιητικό βαθμό τη μεταβλητότητα της εξαρτημένης μεταβλητής (Est2 και Est28). Αυτό δείχνει την περίπλοκη φύση του τσιμέντου, ταυτόχρονα, όμως, μπορεί να είναι ενδεικτικό για το γεγονός ότι λείπουν από το μοντέλο κάποιες βασικές μεταβλητές. Ήδη, γνωρίζουμε από την τεχνολογία του τσιμέντου πως υπάρχουν πολλές λειτουργικές παράμετροι που επηρεάζουν τις τελικές αντοχές του τσιμέντου, όπως για παράδειγμα η θερμοκρασία κλινκεροποίησης ή ο ρυθμός ψύξης του κλίνκερ. Δυστυχώς, όμως, οι παράμετροι αυτές είναι πολύ δύσκολο να μετρηθούν, γι' αυτό και δεν ήταν διαθέσιμες από τους υπεύθυνους της τσιμεντοβιομηχανίας.

Το πρόβλημα της ύπαρξης αυτοσυσχέτισης μεταξύ των καταλοίπων συνήθως εμφανίζεται στις χρονοσειρές, και υποδηλώνει την ανάγκη συμπερίληψης του χρόνου ως ανεξάρτητη μεταβλητή, π.χ. όταν υπάρχει εποχικότητα σε κάποια προϊόντα. Επειδή το φαινόμενο αυτό δεν συμβαίνει στην παραγωγή του τσιμέντου, ένα συμπέρασμα που προκύπτει είναι ότι δεν έχουν συμπεριληφθεί κάποιες σημαντικές μεταβλητές, όπως αυτές που αναφέρθηκαν στην προηγούμενη παράγραφο. Επίσης, ένας άλλος πιθανός λόγος -που αναφέρεται περισσότερο στην περίπτωση της απλής παλινδρόμησης- είναι η μη ύπαρξη επαρκούς κανονικότητας. Πολλές μεταβλητές, ακόμα και οι εξαρτημένες, δεν ακολουθούν σε όλες τις περιπτώσεις την κανονική κατανομή. Ακόμα κι αν η ασυμμετρία και η κύρτωση αυτών είναι εντός των ορίων [-2,2], τα ιστογράμματα και τα θηκογράμματα δεν δείχνουν το ίδιο. Γενικά, αυτό που ισχύει στη Στατιστική είναι το εξής: *το γεγονός ότι δεν απορρίπτω τη μηδενική υπόθεση για ύπαρξη κανονικότητας, δεν σημαίνει ότι την αποδέχομαι*.

Ø Σχετικά με τα μοντέλα της απλής παλινδρόμησης, παρατηρούμε ότι στην απλή παλινδρόμηση μεταξύ Est28 και Est2 εφαρμόζονται κάθε φορά διάφορα μοντέλα ως τα πλέον κατάλληλα (γραμμικό, αντίστροφο ως προς *Y*, αντίστροφο ως προς *X*, τετραγωνική ρίζα του *X*). Αντίθετα, στην περίπτωση της απλής παλινδρόμησης μεταξύ Est28 και Est7 το μοντέλο που εφαρμόζεται σε όλες τις περιπτώσεις είναι το γραμμικό. Αυτό δείχνει ότι κατά την έβδομη ημέρα η συμπεριφορά του τσιμέντου ως προς την ανάπτυξη των αντοχών του έχει σταθεροποιηθεί σε κάποιο βαθμό, ενώ μέχρι τότε υπάρχουν αρκετοί παράγοντες που επηρεάζουν τη συμπεριφορά αυτή. Ενδεικτικό είναι και το γεγονός ότι το R^2 είναι αρκετά μεγαλύτερο στις απλές παλινδρομήσεις του Est28 με το Est7, παρά με το Est2 (εξαίρεση αποτελεί μόνο η περίπτωση του CEM II 42,5, MT1).

Ικανοποιητικό μοντέλο μπορεί να θεωρηθεί μόνο αυτό στην περίπτωση του τύπου τσιμέντου CEM II 42,5, MT1, τόσο για την παλινδρόμηση μεταξύ Est28 και Est2 όσο και μεταξύ Est28 και Est7. Το R^2 είναι περίπου 57%, το οποίο μπορεί να θεωρηθεί σχετικά ικανοποιητικό. Σχετικά με την ύπαρξη αυτοσυσχέτισης μεταξύ των καταλοίπων, αυτή οφείλεται πιθανότατα στην έλλειψη επαρκούς κανονικότητας.

7.3. Προτάσεις για Περαιτέρω Έρευνα

Γενικά, αξίζει να σημειωθεί ότι η Στατιστική αποτελεί ένα σημαντικό εργαλείο για κάθε χημική βιομηχανία (και όχι μόνο). Μέσω διαφόρων απλών και πιο σύνθετων αναλύσεων μπορούν να συναχθούν διάφορα συμπεράσματα, από τη συμπεριφορά των μεταβλητών που μελετώνται. Μια απλή συσχέτιση μεταξύ δύο μεταβλητών

μπορεί να οδηγήσει στη δημιουργία ενός μοντέλου πρόβλεψης, το οποίο μπορεί να αποβεί καθοριστικό για τον έλεγχο της ποιότητας των τελικών προϊόντων. Ένα μοντέλο πολλαπλής παλινδρόμησης μπορεί να εξυπηρετήσει σε ακόμα μεγαλύτερο βαθμό τις ανάγκες για πρόβλεψη. Ωστόσο, για τη συγκεκριμένη μελέτη περίπτωσης, καλό είναι να επιλεχθούν και άλλες τιμές δεδομένων σε άλλες χρονικές περιόδους, οι οποίες θα είναι πιο τακτικές. Αυτό, διότι οι μεταβλητές δεν φαίνεται να ακολουθούν ικανοποιητικά την κανονική κατανομή, ίσως επειδή η δειγματοληψία δεν γίνεται συστηματικά σε τακτικά χρονικά διαστήματα, αλλά μπορεί να περάσουν και δύο εβδομάδες μεταξύ δύο διαδοχικών μετρήσεων. Το γεγονός ότι μερικές μεταβλητές πιθανότατα να μην ακολουθούν την κανονική κατανομή επειδή ελέγχονται σε μεγάλο βαθμό είναι επιθυμητό, από την άλλη, όμως, η έλλειψη επαρκούς κανονικότητας αμφισβητεί την εγκυρότητα πολλών στατιστικών τεστ.

Επίσης, κάτι άλλο που προτείνεται να εφαρμοστεί από τη συγκεκριμένη βιομηχανία είναι η συγκέντρωση δεδομένων και από μερικές λειτουργικές παραμέτρους, όπως οι συνθήκες κλινκεροποίησης και ψύξης, οι οποίες επηρεάζουν τις τελικές αντοχές του τσιμέντου. Από την παρούσα μελέτη φάνηκε ότι οι υπάρχουσες μεταβλητές, αν και είναι γνωστό ότι παίζουν σημαντικό ρόλο στις αντοχές του τσιμέντου, δεν εξηγούν σε ικανοποιητικό βαθμό τη μεταβλητότητα των αντοχών σε θλίψη. Συνεπώς, υπάρχουν και άλλες παράμετροι που πρέπει να ληφθούν υπόψη. Επίσης, η αυτοσυσχέτιση των καταλοίπων πιθανότατα να δείχνει τη μη συμπερίληψη κάποιας ή κάποιων σημαντικών μεταβλητών.

Τέλος, για πιο έγκυρα και αξιόπιστα αποτελέσματα, μπορούν να εφαρμοστούν συστήματα προσομοίωσης ή και μερικά ιδιαίτερα αναπτυγμένα προγράμματα στον υπολογιστή, όπως τα τεχνητά νευρωνικά δίκτυα. Ως προσομοίωση ορίζεται η τεχνική χρήσης αντιπροσωπευτικών ή τεχνητών δεδομένων για την αναπαράσταση σε ένα μοντέλο των διαφόρων καταστάσεων που μπορούν να συμβούν στην πραγματική λειτουργία του συστήματος. Η διαδικασία της προσομοίωσης περιλαμβάνει τόσο την κατασκευή του μοντέλου του υπό μελέτη συστήματος, όσο και την πειραματική χρήση του μοντέλου για τη διερεύνηση ενός συγκεκριμένου προβλήματος. Έτσι, μπορεί κανείς να φανταστεί την προσομοίωση σαν μια εφαρμοσμένη μέθοδο που έχει σαν σκοπό να επιτύχει την περιγραφή της συμπεριφοράς του συστήματος, τη διατύπωση θεωριών ή υποθέσεων που ανταποκρίνονται στην παρατηρούμενη συμπεριφορά και την πρόβλεψη μελλοντικής συμπεριφοράς, όπως ποια θα είναι τα αποτελέσματα των αλλαγών στο σύστημα ή στον τρόπο λειτουργίας του. Για παράδειγμα, η προς μελέτη τσιμεντοβιομηχανία μπορεί να δημιουργήσει ένα μοντέλο προσομοίωσης του συστήματος παραγωγής της, και να μπορέσει να προβλέψει ποια θα είναι η τελική ποιότητα του προϊόντος (αντοχές σε θλίψη), εάν γνωρίζει κάποιες άλλες παραμέτρους κατά τα προηγούμενα στάδια της παραγωγικής διαδικασίας. [6]

Τα τεχνητά νευρωνικά δίκτυα είναι προγράμματα για υπολογιστές, που προσομοιώνουν τη βιολογική οργάνωση και τη λειτουργία των βιολογικών νευρώνων. Βασικό τους πλεονέκτημα είναι η ευπλαστότητα, όπως συμβαίνει με τα εγκεφαλικά μας κύτταρα. Έτσι, τα τεχνητά νευρωνικά δίκτυα δεν χρειάζεται να επαναπρογραμματιστούν αν αλλάξει το περιβάλλον. Επιπλέον, μπορούν να «μαθαίνουν» από μόνα τους αυτό που πρέπει να υπολογίσουν, χάρη σε ειδικά προγράμματα που σταδιακά διορθώνουν τα λάθη τους, καθώς μεταβάλλεται η κατάσταση. Επίσης, η μεγάλη χρησιμότητα των νευρωνικών δικτύων έγκειται στην ικανότητά τους να επεξεργάζονται και να αναπαριστούν τόσο γραμμικές όσο και

μη-γραμμικές σχέσεις. Αυτό εξυπηρετεί τα πολύπλοκα συστήματα εξισώσεων, στα οποία ανήκει και η περίπτωση της ανάπτυξης των τελικών αντοχών του τσιμέντου. Για το λόγο αυτό, τα τεχνητά νευρωνικά δίκτυα μπορούν να χρησιμοποιηθούν αντί της κλασικής μεθόδου της ανάλυσης παλινδρόμησης για την πρόβλεψη των τελικών αντοχών του τσιμέντου. [13, 39]

ΒΙΒΛΙΟΓΡΑΦΙΑ

ΕΛΛΗΝΙΚΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

1. Ιστοσελίδα και Αρχεία της προς μελέτη Εταιρείας
2. **Κονδύλης Ε.**, *"Στατιστικές Τεχνικές Διοίκησης Επιχειρήσεων"*, Εκδόσεις Interbooks
3. **Οικονόμου Γ., Γεωργίου Α.**, *"Ποσοτική Ανάλυση για τη Λήψη Διοικητικών Αποφάσεων"*, Τόμοι Α, Β, Εκδόσεις Ε. Μπένου
4. **Παπαϊωάννου Τ., Λουκάς Σ.**, *"Εισαγωγή στη Στατιστική"*, Εκδόσεις Σταμούλη
5. **Παρισάκης Γ., Κασελούρη Β., Τσίμας Σ., Φτίκος Χ.**, *"Χημεία και Τεχνολογία Τσιμέντου"*, Έκδοση Εθνικού Μετσόβιου Πολυτεχνείου, Αθήνα 1992
6. **Σοφοτάσιος Δ.Π., Σπυράκης Π.Γ., Τριανταφύλλου Β.Δ., Χατζηλυγερούδης Ι.Κ.**, *"Προγραμματισμός και Έλεγχος Παραγωγής"*, Εκδόσεις Gutenberg, Αθήνα 2002
7. **Σφακιανάκης Μ.**, *"Πρακτική Πληροφορική και Εφαρμογές"*, Εκδόσεις Πατάκη, Αθήνα 2000
8. **Σφακιανάκης Μ.**, *"Προσομοίωση και Εφαρμογές"*, Εκδόσεις Πατάκη
9. **Τσίμας Σ., Τσιβιλής Σ.**, *"Επιστήμη και Τεχνολογία Τσιμέντου"*, Έκδοση Εθνικού Μετσόβιου Πολυτεχνείου, Αθήνα 2001
10. **Χαλικιάς Ι.Γ.**, *"Στατιστική-Μέθοδοι Ανάλυσης για Επιχειρηματικές Αποφάσεις"*, Εκδόσεις Rosili, Αθήνα 2001

ΞΕΝΟΓΛΩΣΣΗ ΒΙΒΛΙΟΓΡΑΦΙΑ

11. **Aczel A.D., Sounderpandian J.**, *"Complete Business Statistics"*, Fifth Edition, McGraw Hill, 2002
12. **Akkurt Sedat, Tayfur Gokmen, Can Sever,** *"Fuzzy logic model for the prediction of cement compressive strength"***, Cement and Concrete Research** Vol. 34, 2004, pp. 1429-1433
13. **Baykasoglu Adil, Dereli Turkay, Tanis Serkan,** *"Prediction of cement strength using soft computing techniques"*, **Cement and Concrete Research** Vol. 34, 2004, pp. 2083-2090
14. **Bazant Zdenek P., Becq-Giraudon Emilie**, *"Statistical prediction of fracture parameters of concrete and implications for choice of testing standard"*, **Cement and Concrete Research** Vol. 32, 2002, pp. 529-556
15. **Bentz D.P., Haecker C.J., Fenga X.P. , Stutzman P.E.**, *"Prediction of cement physical properties by virtual testing"*, VDZ CONGRESS 2002, Technical Field 1: Influence of process technology on the manufacturing of market-oriented cements
16. **Berenson M.L., Levine D.M., Krehbiel T.C.**, *"Basic Business Statistics"*, Eighth Edition, Prentice Hall, 2002
17. **Cook D., Weisberg S.**, *"An Introduction to Regression Graphics"*, John Wiley and Sons, 1994

18. **Dengiz Berna, Bektas Tolga, Ultanir A. Eren,** *"Simulation optimization based DSS application: A diamond tool production line in industry"*, **Simulation Modelling: Practice and Theory** 2005
19. **Devore J., Peck R.**, *"Statistics-The Exploration and Analysis of Data"*, Fourth Edition, Duxbury Thomson Learning, 2001
20. **Dingstad Gunvor Irene, Westad Frank, Næs Tormod**, *"Three case studies illustrating the properties of ordinary and partial least squares regression in different mixture models"*, **Chemometrics and Intelligent Laboratory Systems** Vol. 71, 2004, pp. 33-45
21. **Draper N.R., Smith H.**, *"Applied Regression Analysis"*, John Wiley & Sons, New York, 1991
22. **European Committee for Standardization**, *prEn 197-1*, Draft September 1996
23. **Faraway J.**, *"Linear Models with R"*, Chapman and Hall/CRC, 2005
24. **Goh T.N.**, *"Operating frameworks for statistical quality engineering"*, **International Journal of Quality & Reliability Management** Vol. 17, No.2, 2000, pp. 180-188
25. **Gulezian R.C.**, *"Elements of Business Statistics"*, Saunders Company, 1979
26. **Hamilton L.C.**, *"Regression with Graphics-A Second Course in Applied Statistics"*, Duxbury Press, California, 1992
27. **Haus Frederique, Boissel Olivier, Junter Guy-Alain**, *"Multiple regression modelling of mineral base oil biodegradability based on their physical properties and overall chemical composition"*, **Chemosphere** Vol. 50, 2003, pp. 939-948
28. **Hebden Julia**, *"Statistics for Economists"*, Philip Allan Publishers, 1981
29. **Hoel P.G., Jessen R.J.**, *"Basic Statistics for Business and Economics"*, Second Edition, John Wiley and Sons, 1977
30. **Holst Lars, Bolmsjo Gunnar**, *"Simulation integration in manufacturing system development: a study of Japanese industry"*, **Industrial Management & Data Systems** Vol. 101, No. 7, 2001, pp. 339-356
31. **Hwang Kwangryul, Noguchi Takafumi, Tomosawa Fuminiro**, *"Prediction model of compressive strength development of fly-ash concrete"*, **Cement and Concrete Research** Vol. 34, 2004, pp. 2269-2276
32. **Karmel D.H., Polasek M.**, *"Applied Statistics for Economists"*, Fourth Edition, Pitman Publishing, 1977
33. **Levin R.I., Rubin D.S.**, *"Statistics for Management"*, Seventh Edition, Prentice Hall, 1998
34. **Makrymichalos M., Antony J., Antony F., Kumar M.**, *"Statistical thinking and its role for industrial engineers and managers in the 21st century"*, **Managerial Auditing Journal** Vol. 20, No. 4, 2005, pp. 354-363
35. **Mason R.D.**, *"Statistical Techniques in Business and Economics"*, Fourth Edition, Richard D. Irwin, 1978
36. **Montgomery D.C., Peck E.A., Vining G.G.**, *"Introduction to Linear Regression Analysis"*, Third Edition, John Wiley and Sons, 2001
37. **Negro C., Alonso A., Blanco A., Tijero J.**, *"Breaking load and bending strength prediction in manufacture of fibre cement composites using artificial neural networks and a flocculation sensor"*, **Composites: Part A: applied science and manufacturing** Vol. 36, 2005, pp. 1617–1626
38. **Neter J., Wasserman W.**, *"Applied Linear Statistical Models"*, Richard D. Irwin, 1974

39. **Raaymakers W.H.M., Weijters A.J.M.M.**, *"Makespan estimation in batch process industries: A comparison between regression analysis and neural networks"*, **European Journal of Operational Research** Vol. 145, 2003, pp.14-30
40. **Rice J.A.**, *"Mathematical Statistics and Data Analysis"*, Second Edition, Duxbury Press, 1995
41. **Richardson S.B., "***Statistical Analysis"*, Second Edition, The Ronald Press Company, New York 1964
42. **Salvatore D.**, *"Managerial Economics in a Global Economy"*, Fifth Edition, THOMSON-South-Western, 2004
43. **Seber G.A.F., Lee A.J.**, *"Linear Regression Analysis"*, Second Edition, John Wiley and Sons, 2003
44. **Sincich T.**, *"Business Statistics by Example"*, Fifth Edition, Prentice Hall, 1996
45. **Tango Siqueira**, *"An extrapolation method for compressive strength prediction of hydraulic cement products"*, **Cement and Concrete Research** Vol. 28, No. 7, 1998, pp. 969-983
46. **Tsivilis S., Parissakis G.**, *"A mathematical model for the prediction of cement strength"*, **Cement and Concrete Research** Vol.25, No.1, 1995, pp.9-14
47. **Tukey John W.**, *"Exploratory Data Analysis"*, Addison-Wesley Publishing Company, 1977
48. **website of ABB**, *"Performance management for the cement industry"*, http://library.abb.com/GLOBAL/SCOT/SCOT244.NSF/VerityDisplay/DD0A88CB10F5C72EC12570490034774B/$File/Cement_Knowledge_Manager_3BHS128037_REVA_lr.pdf
49. **Weerahandi S.**, *"Exact Statistical Methods for Data Analysis"*, Springer, 1995
50. **Williams E.J.**, *"Regression Analysis"*, John Wiley and Sons, 1959
51. **Wonnacott T.H., Wonnacott R.J.**, *"Introductory Statistics for Business and Economics"*, Second Edition, John Wiley and Sons, 1977
52. **Ye G., Van Breugel K., Fraaij A.L.A.**, *"Experimental study and numerical simulation on the formation of microstructure in cementitious materials at early age"*, **Cement and Concrete Research** Vol. 33, 2003, pp.233-239
53. **Yi Seong-Tae, Moon Young-Ho, Kim Jin-Keun**, *"Long-term strength prediction of concrete with curing temperature"*, **Cement and Concrete Research** Vol. 35, 2005, pp. 1961 – 1969
54. **Zelic J., Rusic D., Krstulovic R.**, *"A mathematical model for prediction of compressive strength in cement–silica fume blends"*, **Cement and Concrete Research** Vol. 34, 2004, pp. 2319-2328