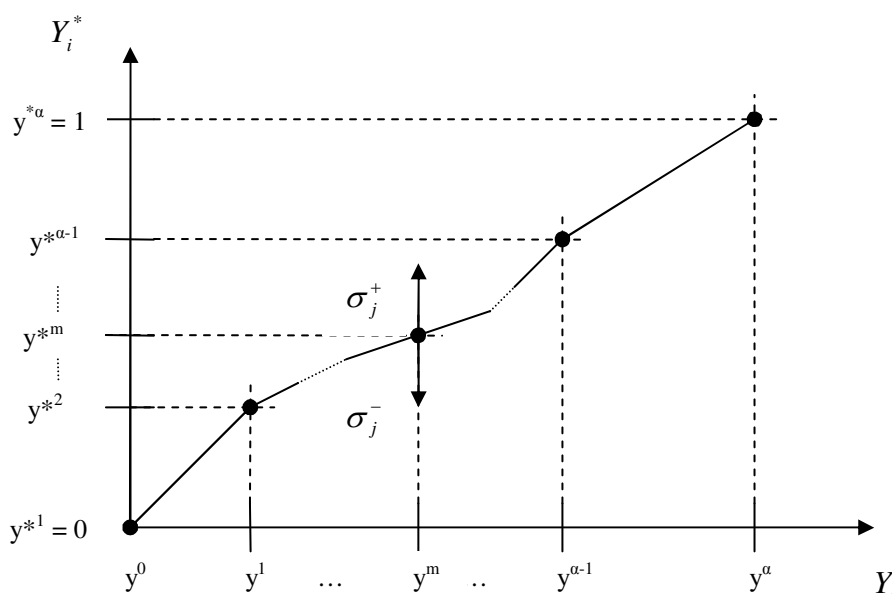




ΠΟΛΥΤΕΧΝΕΙΟ ΚΡΗΤΗΣ
ΤΜΗΜΑ ΜΗΧΑΝΙΚΩΝ ΠΑΡΑΓΩΓΗΣ ΚΑΙ ΔΙΟΙΚΗΣΗΣ
Μεταπτυχιακός Κύκλος Σπουδών
Τομέας Οργάνωσης και Διοίκησης
Πολυτεχνειούπολη, Κουνουπιδιανά, 73100 Χανιά

***“Βηματική ποιοτική παλινδρόμηση:
Ανάπτυξη και διερεύνηση εναλλακτικών μοντέλων”***



Διατριβή που υπεβλήθη για τη μερική ικανοποίηση των απαιτήσεων για την
απόκτηση Μεταπτυχιακού Διπλώματος Ειδίκευσης
υπό

Βασιλείου Α. Σταμπολίδη

Χανιά, 2010

© Copyright υπό Βασίλειο Α. Σταμπολίδη, Χανιά 2010

Η διατριβή του Σταμπολίδη Βασιλείου εγκρίνεται

Γρηγορούδης Ευάγγελος

Επίκουρος Καθηγητής τμήματος Μ.Π.Δ., Επιβλέπων _____

Δούμπος Μιχαήλ

Επίκουρος Καθηγητής τμήματος Μ.Π.Δ. _____

Μαρινάκης Ιωάννης

Λέκτορας τμήματος Μ.Π.Δ. _____

ΠΕΡΙΕΧΟΜΕΝΑ

ΠΕΡΙΕΧΟΜΕΝΑ	- 3 -
ΕΥΧΑΡΙΣΤΙΕΣ	- 6 -
ΚΕΦΑΛΑΙΟ 1 – ΕΙΣΑΓΩΓΗ.....	- 11 -
1.1 ΠΕΡΙΓΡΑΦΗ ΘΕΜΑΤΟΛΟΓΙΑΣ.....	- 11 -
1.2 ΣΚΟΠΟΣ ΤΗΣ ΕΡΓΑΣΙΑΣ.....	- 13 -
1.3 ΔΟΜΗ ΤΗΣ ΕΡΓΑΣΙΑΣ.....	- 14 -
ΚΕΦΑΛΑΙΟ 2 – ΣΤΑΤΙΣΤΙΚΗ ΒΗΜΑΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.....	- 15 -
2.1 ΕΠΙΛΟΓΗ ΤΟΥ ΚΑΛΥΤΕΡΟΥ ΜΟΝΤΕΛΟΥ ΠΑΛΙΝΔΡΟΜΗΣΗΣ	- 15 -
2.2 ΕΠΙΛΟΓΗ ΤΟΥ «ΚΑΛΥΤΕΡΟΥ» ΥΠΟΣΥΝΟΛΟΥ ΣΕ ΠΡΟΒΛΗΜΑΤΑ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....	- 16 -
2.3 ΕΠΙΛΟΓΗ ΤΟΥ ΚΑΛΥΤΕΡΟΥ ΥΠΟΣΥΝΟΛΟΥ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΧΡΗΣΗ ΜΕΘΟΔΩΝ ΒΗΜΑΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....	- 19 -
2.3.1 ΒΗΜΑΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ (STEPWISE REGRESSION)	- 19 -
2.3.2 ΠΡΟΣ ΤΑ ΕΜΠΡΟΣ ΕΠΙΛΟΓΗ (FORWARD SELECTION).....	- 21 -
2.3.3 ΠΡΟΣ ΤΑ ΠΙΣΩ ΑΠΑΛΟΙΦΗ (BACKWARD ELIMINATION).....	- 22 -
2.4 ΠΑΡΑΛΛΑΓΕΣ ΤΗΣ ΜΕΘΟΔΟΥ ΒΗΜΑΤΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....	- 23 -
2.5 ΓΕΝΙΚΗ ΚΡΙΤΙΚΗ ΣΤΗ ΒΗΜΑΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.....	- 24 -
2.6 ΚΡΙΤΗΡΙΑ ΑΞΙΟΛΟΓΗΣΗΣ ΤΩΝ ΜΟΝΤΕΛΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....	- 25 -
2.6.1 ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ (COEFFICIENT OF MULTIPLE DETERMINATION).....	- 25 -
2.6.2 ΠΡΟΣΑΡΜΟΣΜΕΝΟΣ ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ	- 26 -
2.6.3 ΜΕΣΟ ΑΘΡΟΙΣΜΑ ΤΩΝ ΤΕΤΡΑΓΩΝΩΝ ΤΩΝ ΥΠΟΛΟΙΠΩΝ	- 27 -
2.6.4 ΔΕΙΚΤΗΣ ΠΡΟΣΑΡΜΟΓΗΣ.....	- 27 -
2.6.5 ΚΡΙΤΗΡΙΟ ΤΟΥ MALLOW	- 28 -
2.7 ΚΡΙΤΗΡΙΑ ΓΙΑ ΤΗΝ ΕΙΣΑΓΩΓΗ – ΕΞΑΓΩΓΗ ΜΕΤΑΒΛΗΤΩΝ	- 29 -
2.7.1 ΣΤΑΤΙΣΤΙΚΗ F.....	- 29 -
2.7.2 Η ΤΙΜΗ P	- 30 -
2.8 ΚΡΙΤΙΚΗ ΣΤΗΝ ΕΠΙΛΟΓΗ ΜΟΝΤΕΛΟΥ ΠΑΛΙΝΔΡΟΜΗΣΗΣ	- 30 -
ΚΕΦΑΛΑΙΟ 3 – ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ	- 33 -
3.1 ΕΝΑΛΛΑΚΤΙΚΕΣ ΜΕΘΟΔΟΛΟΓΙΕΣ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ.....	- 33 -
3.1.1 ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΧΡΗΣΗ ΜΑΘΗΜΑΤΙΚΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ	- 33 -
3.1.2 ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΧΡΗΣΗ ΑΚΕΡΑΙΟΥ ΜΑΘΗΜΑΤΙΚΟΥ ΠΡΟΓΡΑΜΜΑΤΙΣΜΟΥ	- 35 -
3.1.3 ΒΗΜΑΤΙΚΗ ΠΑΡΑΛΛΑΓΗ ΧΡΗΣΙΜΟΠΟΙΗΣΗΣ ΤΟΥ AIC	- 37 -
3.1.4 ΠΡΟΤΕΙΝΟΜΕΝΗ ΠΑΡΑΛΛΑΓΗ ΤΗΣ ΠΡΟΣ ΤΑ ΠΙΣΩ ΒΗΜΑΤΙΚΗΣ ΑΠΑΛΟΙΦΗΣ	- 38 -
3.1.5 ΜΕΘΟΔΟΛΟΓΙΕΣ ΕΠΙΛΟΓΗΣ ΜΕΤΑΒΛΗΤΩΝ ΜΕ ΧΡΗΣΗ ΤΟΥ ΑΘΡΟΙΣΜΑΤΟΣ ΑΠΟΛΥΤΩΝ ΣΦΑΛΜΑΤΩΝ	- 39 -
3.2 ΣΥΓΚΡΙΣΗ ΕΝΑΛΛΑΚΤΙΚΩΝ ΜΕΘΟΔΩΝ ΓΙΑ ΕΠΙΛΟΓΗ ΤΟΥ ΚΑΛΥΤΕΡΟΥ ΥΠΟΣΥΝΟΛΟΥ.....	- 39 -
ΚΕΦΑΛΑΙΟ 4 – ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΛΟΓΙΑ.....	- 41 -
4.1 ΠΟΙΟΤΙΚΗ ΑΝΑΛΥΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ.....	- 41 -
4.2 ΕΝΑΛΛΑΚΤΙΚΑ ΚΡΙΤΗΡΙΑ ΒΕΛΤΙΣΤΟΠΟΙΗΣΗΣ	- 47 -
4.2.1 ΤΕΤΡΑΓΩΝΙΚΟ ΆΘΡΟΙΣΜΑ ΑΠΟΛΥΤΩΝ ΣΦΑΛΜΑΤΩΝ.....	- 47 -
4.2.2 ΣΥΝΤΕΛΕΣΤΗΣ ΠΡΟΣΔΙΟΡΙΣΜΟΥ R^2	- 47 -
4.3 ΔΙΑΔΙΚΑΣΙΑ ΕΙΣΑΓΩΓΗΣ – ΕΞΑΓΩΓΗΣ ΜΕΤΑΒΛΗΤΩΝ ΣΤΟ ΜΟΝΤΕΛΟ.....	- 52 -

4.3.1 ΠΡΟΣ ΤΑ ΕΜΠΡΟΣ ΒΗΜΑΤΙΚΗ ΠΟΙΟΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ	- 53 -
4.3.1.1 Άθροισμα απλών απόλυτων σφαλμάτων.....	- 53 -
4.3.1.2 Άθροισμα τετραγωνικών σφαλμάτων	- 54 -
4.3.1.3 Συντελεστής προσδιορισμού	- 56 -
4.3.2 ΠΡΟΣ ΤΑ ΠΙΣΩ ΒΗΜΑΤΙΚΗ ΠΟΙΟΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ.....	- 56 -
4.3.2.1 Άθροισμα απλών απόλυτων σφαλμάτων.....	- 57 -
4.3.2.2 Άθροισμα τετραγωνικών σφαλμάτων.....	- 59 -
4.3.2.3 Συντελεστής προσδιορισμού	- 59 -
ΚΕΦΑΛΑΙΟ 5 – ΠΡΑΚΤΙΚΕΣ ΕΦΑΡΜΟΓΕΣ.....	- 63 -
5.1 ΤΑ ΔΕΔΟΜΕΝΑ ΔΙΕΡΕΥΝΗΣΗΣ	- 63 -
5.1.1 ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ 1: ΤΑ ΔΕΔΟΜΕΝΑ WATER USAGE OF PRODUCTION PLANT.....	- 64 -
5.1.1.1 Άθροισμα Απόλυτων Σφαλμάτων.....	- 65 -
5.1.1.1.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 65 -
5.1.1.1.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 67 -
5.1.1.2 Άθροισμα Τετραγωνικών Σφαλμάτων.....	- 67 -
5.1.1.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 68 -
5.1.1.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 69 -
5.1.1.3 Συντελεστής προσδιορισμού	- 70 -
5.1.1.3.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 70 -
5.1.1.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα	- 71 -
5.1.2 ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ 2: ΤΑ ΔΕΔΟΜΕΝΑ SOLAR THERMAL ENERGY	- 74 -
5.1.2.1 Άθροισμα Απόλυτων Σφαλμάτων.....	- 75 -
5.1.2.1.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 77 -
5.1.2.1.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 78 -
5.1.2.2 Άθροισμα Τετραγωνικών Σφαλμάτων.....	- 78 -
5.1.2.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 80 -
5.1.2.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 80 -
5.1.2.3 Συντελεστής προσδιορισμού	- 81 -
5.1.2.3.1 Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση	- 81 -
5.1.2.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα	- 83 -
5.2 ΤΑ ΔΕΔΟΜΕΝΑ ΕΛΕΓΧΟΥ	- 86 -
5.2.1 ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ 3: ΤΑ ΔΕΔΟΜΕΝΑ BELLE AYR LIQUEFACTION RUNS	- 87 -
5.2.1.1 Άθροισμα Απόλυτων Σφαλμάτων.....	- 89 -
5.2.1.1.1 Προς τα εμπρός βηματική παλινδρόμηση.....	- 90 -
5.2.1.1.2 Προς τα πίσω βηματική παλινδρόμηση	- 91 -
5.2.1.2 Άθροισμα Τετραγωνικών Σφαλμάτων.....	- 92 -
5.2.1.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 93 -
5.2.1.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 94 -
5.2.1.3 Συντελεστής προσδιορισμού	- 95 -
5.2.1.3.1 Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση	- 95 -
5.2.1.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα	- 96 -
5.2.2 ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ 4: ΤΑ ΔΕΔΟΜΕΝΑ BODY FAT OF FEMALES	- 100 -
5.2.2.1 Άθροισμα Απόλυτων Σφαλμάτων.....	- 101 -
5.2.2.1.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 102 -
5.2.2.1.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 102 -
5.2.2.2 Άθροισμα Τετραγωνικών Σφαλμάτων.....	- 103 -
5.2.2.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 103 -
5.2.2.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 104 -
5.2.2.3 Συντελεστής προσδιορισμού	- 104 -
5.2.2.3.1 Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση	- 105 -
5.2.2.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα	- 106 -
5.2.3 ΠΡΑΚΤΙΚΗ ΕΦΑΡΜΟΓΗ 5: ΤΑ ΔΕΔΟΜΕΝΑ HALD	- 107 -
5.2.3.1 Άθροισμα Απόλυτων Σφαλμάτων.....	- 109 -
5.2.3.1.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 109 -
5.2.3.1.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 110 -
5.2.3.2 Άθροισμα Τετραγωνικών Σφαλμάτων.....	- 110 -
5.2.3.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση.....	- 111 -
5.2.3.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση	- 112 -
5.2.3.3 Συντελεστής προσδιορισμού	- 112 -
5.2.3.3.1 Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση	- 113 -

5.2.3.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα	114 -
ΚΕΦΑΛΑΙΟ 6 – ΣΥΜΠΕΡΑΣΜΑΤΑ	117 -
6.1 ΣΥΜΠΕΡΑΣΜΑΤΑ	117 -
6.2 ΜΕΛΛΟΝΤΙΚΕΣ ΠΡΟΕΚΤΑΣΕΙΣ.....	126 -
ΒΙΒΛΙΟΓΡΑΦΙΑ.....	129 -
ΠΑΡΑΡΤΗΜΑ.....	131 -
ΑΝΑΛΥΤΙΚΑ ΑΠΟΤΕΛΕΣΜΑΤΑ ΕΦΑΡΜΟΓΗΣ ΣΤΑ ΣΥΝΟΛΑ ΔΕΔΟΜΕΝΩΝ ΕΛΕΓΧΟΥ	131 -

Ευχαριστίες

Είναι γενικά αποδεκτό ότι η *επιστημονική έρευνα* αποτελεί ένα από τα σημαντικότερα εργαλεία ανάπτυξης και προόδου μιας οργανωμένης κοινωνίας. Είναι το εργαλείο το οποίο, βάσει ορισμένων κανόνων, μπορεί να οδηγήσει στη διεύρυνση των γνώσεων μας για τον κόσμο που μας περιβάλλει. Η χαρά και η ικανοποίηση από την απόκτηση καινούριων γνώσεων, συγκρίνεται με αυτή η οποία προέρχεται από τη αξιοποίηση των γνώσεων αυτών σε πρακτικό επίπεδο, τονώνοντας το αίσθημα της δημιουργικότητας. Αισθάνομαι πολύ τυχερός για το λόγο ότι βρέθηκα κι εγώ στη θέση να παρατηρήσω, από όσο κοντά επιτρέπει η ενασχόλησή μου με τις ακαδημαϊκές υποχρεώσεις ενός μεταπτυχιακού κύκλου σπουδών, τον επιστημονικό τρόπο με τον οποίο καταβάλλεται προσπάθεια για έρευνα.

Περισσότερο τυχερός αισθάνομαι διότι κατά την ενασχόλησή μου αυτή, βρήκα σημαντικότερη συμπαράσταση από τον Δρ. Ευάγγελο Γρηγορούδη, Επίκουρο Καθηγητή του τμήματος Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης, ο οποίος και επέβλεψε τη διπλωματική μου εργασία για την απόκτηση του Μεταπτυχιακού μου Διπλώματος. Κάτω από την υπομονετική καθοδήγησή του, κατάφερα να συγκεντρώσω το προϊόν της έρευνάς μου στις σελίδες αυτής της εργασίας. Θα ήθελα, επίσης, να ευχαριστήσω και τα υπόλοιπα μέλη της εξεταστικής επιτροπής της εργασίας μου, κκ. Δρ. Μιχαήλ Δούμπο, Επίκουρο Καθηγητή του τμήματος Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης και Δρ. Ιωάννη Μαρινάκη, Λέκτορα του τμήματος Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης, επειδή δέχτηκαν να μετάσχουν σε αυτή.

Είναι γεγονός ότι για να αποδώσει καρπούς μια επιστημονική έρευνα, χρειάζεται πολλή προσωπική εργασία, υπομονή και χρόνο. Χρειάζεται, ωστόσο, και αρκετή συμπαράσταση. Οι γονείς μου ήταν αυτοί που κυρίως συνέβαλαν υλικά και ηθικά για την ολοκλήρωση των ακαδημαϊκών μου υποχρεώσεων. Τους ευχαριστώ.

Τέλος, θα ήθελα να ευχαριστήσω όλους εκείνους οι οποίοι με τον τρόπο τους, ακόμα και εάν δεν το γνωρίζουν, συνέβαλαν και αυτοί στο να μπορέσω να ολοκληρώσω τους ακαδημαϊκούς μου στόχους.

Σύντομο Βιογραφικό Σημείωμα

Ο Σταμπολίδης Α. Βασίλειος γεννήθηκε στο Φρέχεν της Γερμανίας, το 1979. Τη βασική του εκπαίδευση την έλαβε στο Κιλκίς, όπου και αποφοίτησε από το 1^ο Γενικό Λύκειο Κιλκίς. Το 1999 εισήχθη στο Τμήμα Μηχανικών Παραγωγής και Διοίκησης του Πολυτεχνείου Κρήτης, στα Χανιά. Τον Μάρτιο του 2005 έλαβε το πτυχίο του Διπλωματούχου Μηχανικού Παραγωγής και Διοίκησης, με βαθμό 7,87 και στη συνέχεια έγινε αποδεκτή η αίτησή του για εγγραφή στο Μεταπτυχιακό Πρόγραμμα Ειδίκευσης του ίδιου τμήματος, στον τομέα Οργάνωσης και Διοίκησης. Η παρούσα εργασία, αποτελεί μέρος των απαιτήσεων για την ολοκλήρωση των σπουδών του για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης.

Περίληψη

Ο σκοπός των μεθόδων παλινδρόμησης είναι, σε γενικές γραμμές, να αποσαφηνιστεί, μέσω της εφαρμογής τους, η σχέση μεταξύ μιας ή περισσότερων ανεξάρτητων μεταβλητών και μιας εξαρτημένης μεταβλητής. Αυτό γίνεται μέσω της επιλογής του καλύτερου μοντέλου παλινδρόμησης, το οποίο περιγράφει περισσότερο ικανοποιητικά την παραπάνω ζητούμενη σχέση για τα συγκεκριμένα δεδομένα. Τα διαθέσιμα δεδομένα μπορεί να εκτείνονται από μερικές δεκάδες μέχρι και μερικές χιλιάδες. Το γεγονός αυτό καθιστά επιτακτική την ανάπτυξη τεχνικών, οι οποίες θα είναι σε θέση να «εντοπίζουν» εκείνες τις ανεξάρτητες μεταβλητές οι οποίες θα συνεισφέρουν περισσότερο στην εξηγηματική ή προβλεπτική ικανότητα του μοντέλου παλινδρόμησης, χωρίς να είναι απαραίτητος ο εξαντλητικός έλεγχος όλων των διαθέσιμων δεδομένων. Υπάρχει ήδη πληθώρα ερευνών στις οποίες έχουν χρησιμοποιηθεί γενικές μέθοδοι απλής και πολλαπλής γραμμικής παλινδρόμησης και έχουν προταθεί αρκετές τεχνικές για την επιλογή του καλύτερου υποσυνόλου ανεξάρτητων μεταβλητών που δομούν το τελικό μοντέλο.

Στα πλαίσια αυτής της εργασίας, επιχειρείται η ανάπτυξη και η διερεύνηση εναλλακτικών μοντέλων ποιοτικής ανάλυσης παλινδρόμησης, η πρακτικότητα και η λειτουργικότητα των οποίων εξετάζεται με την εφαρμογή τους σε πραγματικά δεδομένα. Κύρια επιδίωξη της εργασίας είναι να ενσωματωθούν και να προσαρμοστούν, σε πρακτικό επίπεδο, οι ήδη υπάρχουσες τάσεις και ιδέες στο χώρο της ποιοτικής ανάλυσης παλινδρόμησης. Επιπλέον στόχος της εργασίας, είναι να αναπτυχθεί μια μεθοδολογία για την επιλογή του καλύτερου υποσυνόλου ανεξάρτητων μεταβλητών, οι οποίες θα συμμετάσχουν στο τελικό μοντέλο παλινδρόμησης. Η προσπάθεια αυτή παρουσιάζει σχετικό ενδιαφέρον, καθώς, στα πλαίσια χρησιμοποίησης της ποιοτικής ανάλυσης παλινδρόμησης, η ερευνητική δραστηριότητα, μέχρι στιγμής, για την ανάπτυξη αντίστοιχων μεθοδολογιών, ουσιαστικά δεν έχει να επιδείξει κάποια αποτελέσματα.

Η ποιοτική ανάλυση παλινδρόμησης έχει χρησιμοποιηθεί σε πολλά προβλήματα πολυκριτήριας ανάλυσης. Βασική αρχή της συγκεκριμένης προσέγγισης, αποτελεί η υπόθεση ότι η «αξία» μιας εξαρτημένης μεταβλητής, Y , μπορεί να γραφεί ως ο σταθμισμένος μέσος των «αξιών» των ανεξάρτητων μεταβλητών, X_i . Η διαφορά με την απλή γραμμική παλινδρόμηση έγκειται στο γεγονός του διαφορετικού τρόπου υπολογισμού των συναρτήσεων αξιών, τόσο της εξαρτημένης μεταβλητής, όσο και των ανεξάρτητων μεταβλητών. Στόχος είναι να αναπτυχθεί ένα μη γραμμικό μοντέλο πρόβλεψης, με το οποίο θα επιτυγχάνεται η εκτίμηση των προαναφερθέντων συναρτήσεων αξιών, οι οποίες είναι γραμμικές μόνο κατά τμήματα. Το υπόδειγμα της ποιοτικής ανάλυσης παλινδρόμησης υποθέτει ότι η συνεισφορά ενός παράγοντα X_i στην εξαρτημένη μεταβλητή Y , δεν είναι αναλογική, αλλά εξαρτάται από την ονομαστική τιμή της μεταβλητής. Για την επιλογή του καλύτερου υποσυνόλου των ανεξάρτητων μεταβλητών, η ποιοτική ανάλυση παλινδρόμησης, στηριζόμενη στο γενικότερο πλαίσιο των αρχών του γραμμικού προγραμματισμού (προγραμματισμός στόχων), επιδιώκει την εκτίμηση των συναρτήσεων «αξίας» που αποδίδεται στην εξαρτημένη μεταβλητή και στις επιλεγόμενες ανεξάρτητες μεταβλητές, έτσι ώστε να υπάρχει η μεγαλύτερη δυνατή συνέπεια με την πραγματική τιμή της εξαρτημένης μεταβλητής. Με τον τρόπο αυτό, μπορούν να εξεταστούν αρκετά κριτήρια βελτιστοποίησης για την ανάπτυξη της τελικής προτεινόμενης μεθοδολογίας. Τέλος, για τον έλεγχο της ακρίβειας των αποτελεσμάτων της προτεινόμενης μεθοδολογίας,

αυτά αντιπαραβάλλονται με τα αντίστοιχα της κλασσικής βηματικής ανάλυσης παλινδρόμησης.

ΚΕΦΑΛΑΙΟ 1

ΕΙΣΑΓΩΓΗ

1.1 Περιγραφή Θεματολογίας

Στο κεφάλαιο αυτό, θα επιχειρηθεί μια προκαταρκτική εισαγωγή στο θέμα το οποίο διαπραγματεύεται η παρούσα εργασία, θα γίνει μια γενική παρουσίαση των επιμέρους αντικειμένων τα οποία εξετάζονται στη συνέχεια, ενώ, παράλληλα, καθορίζονται και οι στόχοι που τίθενται στα πλαίσια αυτής της εξέτασης, αναδεικνύοντας ταυτόχρονα την αναγκαιότητα της τελευταίας. Τέλος, καθορίζεται η δομή της παρούσας εργασίας.

Σε πολλές σύγχρονες επιστημονικές και επιχειρηματικές εφαρμογές, είναι επιβεβλημένη η ανάγκη για την επεξεργασία και διαχείριση σύνθετων και τεράστιων, πολλές φορές, όγκων δεδομένων, η ανάλυση των οποίων είναι ζωτικής σημασίας για την εξαγωγή κρίσιμων συμπερασμάτων. Η ανάπτυξη της πληροφορικής επιστήμης και η τεχνολογική πρόοδος η οποία επιτρέπει τη δημιουργία υπολογιστικών μονάδων με ολοένα και αυξανόμενη υπολογιστική και αποθηκευτική ισχύ, στέκονται αρωγοί στην προσπάθεια ανάπτυξης μεθόδων για τη διαχείριση τέτοιων δεδομένων. Ταυτόχρονα, διεξάγεται και ερευνητική προσπάθεια για την ανάπτυξη θεωρητικών μεθοδολογιών, οι οποίες θα είναι σε θέση, εφαρμοζόμενες, να αποδώσουν τα αναμενόμενα αποτελέσματα με βάση την αναφερόμενη διαχείριση δεδομένων. Παρόλα αυτά, η ανάγκη για τη διαχείριση και ανάλυση πολύ μεγάλων όγκων δεδομένων, στα πλαίσια σύνθετων προβλημάτων, δεν μπορεί να ικανοποιηθεί επαρκώς από τις μέχρι τώρα θεωρητικές και τεχνολογικές ανακαλύψεις. Επομένως, το ερευνητικό ενδιαφέρον στα πεδία αυτά, εμφανίζεται έντονο.

Ερευνητικό ενδιαφέρον παρουσιάζει και η περίπτωση διαχείρισης και επεξεργασίας δεδομένων μέσω της ανάλυσης παλινδρόμησης. Στα πλαίσια αυτής της ανάλυσης, αναζητούνται οι συσχετισμοί μεταξύ ενός ή πολλών χαρακτηριστικών (εξαρτημένη(ες) μεταβλητή(ές)) και μερικών άλλων επιμέρους χαρακτηριστικών (ανεξάρτητες μεταβλητές), με βάση τους οποίους, θα συγκροτηθεί ένα μοντέλο παλινδρόμησης με σκοπό την ανάλυση, την ερμηνεία ή την πρόβλεψη δεδομένων. Το μοντέλο παλινδρόμησης, συνίσταται από μια εξίσωση, γραμμική ή μη, η οποία συνδέει τα επιμέρους αναφερόμενα χαρακτηριστικά. Είναι ευνόητο ότι το επιδιωκόμενο μοντέλο παλινδρόμησης, είναι επιθυμητό να επιτυγχάνει τη μεγαλύτερη δυνατή αντιστοιχία μεταξύ των εκτιμώμενων δεδομένων, όπως προκύπτουν από αυτό, και των πραγματικών διαθέσιμων δεδομένων. Η μέχρι τώρα εμπειρία στο θέμα αυτό, δείχνει ότι κάθε σύνολο δεδομένων παρουσιάζει διαφορετική «συμπεριφορά», ανάλογα με τις ιδιότητες που έχει (ποιότητα δεδομένων, π.χ. υψηλές ή χαμηλές συσχετίσεις μεταξύ των μεταβλητών). Από την άλλη, σε αρκετές περιπτώσεις είναι δυνατόν να είναι διαθέσιμος ένας πολύ μεγάλος αριθμός χαρακτηριστικών, με ανάλογο πλήθος δεδομένων για αυτά. Καθίσταται φανερό ότι η

διαχείριση τέτοιων δεδομένων είναι χρονοβόρα και, αρκετά συχνά, επίπονη. Με βάση τη διαπίστωση αυτή, έχουν αναπτυχθεί αρκετές θεωρητικές τεχνικές (με ανάλογες πρακτικές εφαρμογές), οι οποίες στοχεύουν στη δόμηση ενός μοντέλου παλινδρόμησης, χωρίς να είναι απαραίτητος ο εξαντλητικός έλεγχος όλων των διαθέσιμων χαρακτηριστικών, εξασφαλίζοντας ταυτόχρονα, όσο είναι δυνατόν, και μεγάλη αξιοπιστία αποτελεσμάτων. Το πρόβλημα της επιλογής των πλέον κατάλληλων μεταβλητών, χωρίς τον εξαντλητικό έλεγχο όλων των διαθέσιμων δεδομένων είναι γνωστό και ως πρόβλημα επιλογής του καλύτερου υποσυνόλου μεταβλητών (best subset selection problem).

Η παρούσα εργασία, διαπραγματεύεται το θέμα της ανάπτυξης μοντέλων παλινδρόμησης, διερευνώντας την δυνατότητα πρακτικής εφαρμογής μιας εναλλακτικής μεθοδολογίας για το σκοπό αυτό. Πιο συγκεκριμένα, στόχος είναι να διερευνηθεί εάν είναι δυνατόν οι αρχές της ποιοτικής ανάλυσης παλινδρόμησης να μπορούν να χρησιμοποιηθούν με τέτοιο τρόπο ώστε να προταθεί ένα μοντέλο παλινδρόμησης για την περιγραφή ενός οποιουδήποτε συνόλου δεδομένων, χωρίς να είναι απαραίτητος ο εξαντλητικός έλεγχος όλων των διαθέσιμων δεδομένων. Για να γίνει κάτι τέτοιο, υιοθετήθηκε ένας «βηματικός» τρόπος ελέγχου για το ποια μεταβλητή είναι σε θέση να συμμετέχει στο αναζητούμενο μοντέλο παλινδρόμησης (ή, αντίστοιχα, να απορριφθεί από αυτό). Αξίζει να σημειωθεί εδώ ότι ήδη υπάρχουν αρκετές μεθοδολογίες οι οποίες προσεγγίζουν με βηματικό τρόπο την ανάπτυξη ενός μοντέλου παλινδρόμησης, έτσι ώστε να μην είναι απαραίτητος ο εξαντλητικός έλεγχος όλων των διαθέσιμων δεδομένων. Τέτοιου είδους μεθοδολογίες, είναι και αυτές που ανήκουν στην οικογένεια της κλασσικής βηματικής ανάλυσης παλινδρόμησης (stepwise regression analysis). Πιο αναλυτικά, υπάρχουν και χρησιμοποιούνται σχετικά συχνά σε διάφορες εφαρμογές, η *προς τα εμπρός επιλογή* (forward selection), η *προς τα πίσω απαλοιφή* (backward elimination) και η *βηματική παλινδρόμηση* (stepwise regression). Η προς τα εμπρός επιλογή, ξεκινά από τη μη συμμετοχή καμμιάς ανεξάρτητης μεταβλητής στο μοντέλο παλινδρόμησης και βήμα-βήμα εξετάζεται η συμμετοχή ή όχι κάποιας από τις διαθέσιμες μεταβλητές σε αυτό. η εισαγωγή αυτή στηρίζεται, σε κάθε βήμα (εκτός του αρχικού), στο γεγονός ότι είναι γνωστό ότι ήδη κάποιες συγκεκριμένες μεταβλητές συμμετέχουν σε αυτό και η συμμετοχή μιας άλλης πραγματοποιείται μέσω στατιστικών υποθέσεων, οι οποίες επιβεβαιώνονται ή όχι με βάση ένα προκαθορισμένο κατώφλι αποδοχής (π.χ. χρησιμοποίηση της στατιστικής F). Στην περίπτωση της προς τα πίσω απαλοιφής, η διαδικασία ξεκινά με την υπόθεση ότι όλες οι διαθέσιμες μεταβλητές συμμετέχουν στο μοντέλο παλινδρόμησης, οπότε σε κάθε επόμενο βήμα, εξετάζεται, πάλι με στατιστικές υποθέσεις και ένα προκαθορισμένο κατώφλι απόρριψης, η εξαγωγή μιας μεταβλητής από το μοντέλο. Τέλος, η βηματική παλινδρόμηση, αποτελεί έναν συνδυασμό των δύο προαναφερθέντων μεθοδολογιών και σε κάθε βήμα της εξετάζεται τόσο η συμμετοχή μιας μεταβλητής εκτός μοντέλου, όσο και η απόρριψη μιας άλλης, η οποία ήδη συμμετέχει σε αυτό. αξίζει να σημειωθεί, ότι με τον τρόπο αυτό, μια μεταβλητή η οποία κρίθηκε στατιστικά ικανή να συμμετέχει στο μοντέλο σε κάποιο βήμα αυτή, είναι δυνατόν σε επόμενο βήμα να απορριφθεί. Η διαδικασία στην περίπτωση αυτή ξεκινά από τη μη συμμετοχή καμμιάς μεταβλητής στο μοντέλο παλινδρόμησης και συνεχίζεται με βάση τη διαδικασία που περιγράφηκε, υιοθετώντας ένα κατώφλι αποδοχής και ένα κατώφλι απόρριψης για τον έλεγχο των στατιστικών υποθέσεων. Σε κάθε μία από τις παραπάνω μεθοδολογίες, το προτεινόμενο αποτέλεσμα είναι ένα σύνολο από προτεινόμενες μεταβλητές (οι οποίες

λαμβάνονται από το σύνολο των διαθέσιμων δεδομένων) και είναι αυτές οι οποίες τελικά θα απαρτίσουν το ζητούμενο μοντέλο παλινδρόμησης.

Η αξιοποίηση της ποιοτικής ανάλυσης παλινδρόμησης για την επιλογή του καλύτερου υποσυνόλου μεταβλητών, εξετάζεται στα πρότυπα των διαδικασιών που περιγράφηκαν προηγουμένως. Στα επόμενα κεφάλαια, περιγράφεται ο τρόπος με τον οποίο εξετάζεται τόσο η θεωρητική όσο και η πρακτική προσέγγιση του θέματος, χρησιμοποιώντας τις αρχές της ποιοτικής ανάλυσης παλινδρόμησης. Ας σημειωθεί εδώ ότι η ποιοτική ανάλυση παλινδρόμησης στηρίζεται στο γεγονός ότι η «συνεισφορά» μιας εξεταζόμενης ανεξάρτητης μεταβλητής στο τελικό μοντέλο παλινδρόμησης, εξαρτάται από την υπολογιζόμενη «αξία» της, και όχι από την ονομαστική τιμή αυτής (δηλαδή, δεν είναι αναλογική αυτή η συνεισφορά). Το γενικό πρότυπο που ακολουθείται, είναι η ανάπτυξη και υλοποίηση των κατάλληλων, σε κάθε βήμα, γραμμικών προγραμμάτων, με βάση τα διαθέσιμα δεδομένα, ώστε να προσδιοριστεί το καλύτερο υποσύνολο μεταβλητών. Όπως γίνεται φανερό και στα κεφάλαια που ακολουθούν, τα γραμμικά προγράμματα τα οποία αναπτύσσονται, είναι γραμμικά, μη γραμμικά και μικτού ακέραιου προγραμματισμού, ανάλογα με το κριτήριο βελτιστοποίησης που υιοθετείται κάθε φορά. Τα αποτελέσματα και τα επιμέρους συμπεράσματα από τη διαδικασία αυτή, περιγράφονται στα κεφάλαια που ακολουθούν.

1.2 Σκοπός της Εργασίας

Η παρούσα εργασία διαπραγματεύεται την αναζήτηση, την ανάδειξη και την εν δυνάμει πρακτική εφαρμογή των θεωρητικών δυνατοτήτων οι οποίες απορρέουν από το υπόβαθρο θεμελίωσης του βασικού μοντέλου της *βηματικής ποιοτικής παλινδρόμησης*. Με βάση το παραπάνω σκεπτικό, επιχειρείται η ανάπτυξη και η διερεύνηση εναλλακτικών μοντέλων ποιοτικής ανάλυσης παλινδρόμησης πολλών μεταβλητών, με στόχο την εξακρίβωση και αξιολόγηση της δυνατότητας που αυτά παρέχουν αναφορικά με την χρησιμοποίησή τους για την ανάπτυξη και εμπέδωση αξιόπιστων μοντέλων παλινδρόμησης, τα οποία θα μπορούν να χρησιμοποιηθούν για σκοπούς όπως η πρόβλεψη δεδομένων. Η σημαντικότητα της πρόβλεψης δεδομένων στα πλαίσια διοικητικών και διαχειριστικών που διέπουν τις σύγχρονες επιχειρήσεις και οργανισμούς, είναι στις μέρες μας αναγνωρισμένη, παρόλο που μπορεί κανείς να ισχυριστεί ότι ο βαθμός υιοθέτησης στρατηγικών για την πρόβλεψη δεδομένων από τις τελευταίες ίσως να μην είναι ανάλογος της σημαντικότητας που αποδίδεται στην πρόβλεψη δεδομένων.

Πιο συγκεκριμένα, σε ό,τι αφορά το σκοπό της εργασίας, βασική επιδίωξη αποτελεί η ανάπτυξη μιας μεθοδολογίας ενός μοντέλου βηματικής παλινδρόμησης, η οποία θα ενσωματώνει τα πλεονεκτήματα από τις σύγχρονες τάσεις και ιδέες που επικρατούν στο χώρο της ποιοτικής ανάλυσης παλινδρόμησης, σε συνδυασμό με τη χρήση του μαθηματικού προγραμματισμού. Η μεθοδολογία αυτή αποσκοπεί στο να προταθεί, μέσω αυτής, μια αξιόπιστη διαδικασία επιλογής του *καλύτερου υποσυνόλου μεταβλητών* από ένα σύνολο διαθέσιμων παρατηρήσεων, οι οποίες τελικά θα συμμετάσχουν στο υπό ανάπτυξη μοντέλο παλινδρόμησης. Το μοντέλο αυτό, μπορεί στη συνέχεια να χρησιμοποιηθεί για διάφορους επιμέρους σκοπούς.

Από τα παραπάνω, καθίσταται φανερό ότι η παρούσα εργασία με τους επιδιωκόμενους στόχους της, επιχειρεί να συνεισφέρει στην έρευνα σχετικά με την ανάδειξη μαθηματικών εργαλείων για την επιλογή του καλύτερου υποσυνόλου μεταβλητών με βάση το οποίο θα αναπτυχθεί ένα μοντέλο πολλαπλής παλινδρόμησης, μιας εξαρτημένης μεταβλητής. Η χρησιμότητα της επιλογής ενός υποσυνόλου μεταβλητών, γίνεται φανερή εάν αναλογιστεί κανείς το μεγάλο πλήθος των μεταβλητών που φαίνεται να επηρεάζουν μια υπό μελέτη (εξαρτημένη) μεταβλητή στα πρακτικά ζητήματα που απασχολούν ερευνητές και επιστήμονες σε ερευνητικό, επιστημονικό και βιομηχανικό επίπεδο. Από την άλλη, η έρευνα πάνω ζήτημα καθορισμού του καλύτερου υποσυνόλου μεταβλητών στα πλαίσια της πολλαπλής ανάλυσης παλινδρόμησης, δεν έχει να επιδείξει ιδιαίτερα αποτελέσματα, με συνέπεια το ερευνητικό ενδιαφέρον για το εν λόγω θέμα να παραμένει έντονο.

1.3 Δομή της Εργασίας

Η παρούσα εργασία, αποτελείται από έξι κεφάλαια. Η δομή της και το περιεχόμενο κάθε κεφαλαίου, φαίνεται στη συνέχεια:

Στο πρώτο κεφάλαιο επιχειρείται μια γενική εισαγωγή στο θέμα με το οποίο πραγματεύεται η εργασία και τίθενται οι στόχοι για τους οποίους πραγματοποιείται. Παράλληλα, ορίζεται η σκοπιμότητα των επιδιωκόμενων στόχων.

Στο δεύτερο κεφάλαιο, γίνεται μια απόπειρα παρουσίασης της στατιστικής βηματικής παλινδρόμησης, όπως αυτή περιγράφεται στη σχετική διαθέσιμη βιβλιογραφία.

Στο τρίτο κεφάλαιο, παρουσιάζονται μερικές ενδεικτικές προσπάθειες χρησιμοποίησης εναλλακτικών μοντέλων βηματικής παλινδρόμησης σε διάφορα πρακτικά ή θεωρητικά ζητήματα.

Στο τέταρτο κεφάλαιο γίνεται η αναλυτική περιγραφή της ανάπτυξης της μεθοδολογίας η οποία προτείνεται μέσα από την παρούσα εργασία. Πιο συγκεκριμένα, περιγράφεται το αναπτυχθέν μαθηματικό μοντέλο για την μεθοδολογία αυτή και παράλληλα παρουσιάζονται τα εναλλακτικά κριτήρια βελτιστοποίησης που χρησιμοποιήθηκαν στα πλαίσια ανάπτυξής της.

Στο πέμπτο κεφάλαιο επιχειρείται η εξακρίβωση της αξιοπιστίας του μαθηματικού μοντέλου που αναπτύχθηκε στο τέταρτο κεφάλαιο. Αυτό πραγματοποιείται με την εφαρμογή του σε πραγματικά δεδομένα που υπάρχουν διαθέσιμα στο διαδίκτυο. Στα πλαίσια της παρούσας εργασίας, τα διαθέσιμα δεδομένα χωρίζονται σε δεδομένα διερεύνησης και σε δεδομένα εφαρμογής. Τα αποτελέσματα αυτά συγκρίνονται με τα αντίστοιχα της στατιστικής βηματικής παλινδρόμησης.

Τέλος, στο έκτο κεφάλαιο, συγκεντρώνονται όλα τα συμπεράσματα τα οποία προκύπτουν από υλοποιείται στα προηγούμενα κεφάλαια, ενώ προτείνονται και μερικές ενδεικτικές μελλοντικές επεκτάσεις οι οποίες μπορούν να βρουν εφαρμογή στο προτεινόμενο μοντέλο.

ΚΕΦΑΛΑΙΟ

2

ΣΤΑΤΙΣΤΙΚΗ ΒΗΜΑΤΙΚΗ ΠΑΛΙΝΔΡΟΜΗΣΗ

Η στατιστική βηματική παλινδρόμηση είναι μια κατηγορία μεθόδων ανάπτυξης μοντέλων παλινδρόμησης, κατά τις οποίες το ζητούμενο μοντέλο αναπτύσσεται βηματικά. Σε κάθε βήμα μιας τέτοιας μεθόδου, εξετάζεται, βάσει κατάλληλων κριτηρίων, ποιες από τις διαθέσιμες μεταβλητές απόφασης θα συμμετάσχουν τελικά στο μοντέλο. Στη συνέχεια, θα δοθεί πρώτα μια γενική εικόνα για την επιλογή του καλύτερου μοντέλου παλινδρόμησης, θα περιγραφεί το πρόβλημα επιλογής του καλύτερου υποσυνόλου μεταβλητών από τις συνολικά διαθέσιμες για την ανάπτυξη ενός μοντέλου παλινδρόμησης και, τέλος, θα περιγραφούν οι πιο δημοφιλείς διαθέσιμες μέθοδοι βηματικής παλινδρόμησης, μαζί με τα κριτήρια που χρησιμοποιούνται για την αξιολόγηση των μοντέλων που αναπτύχθηκαν.

2.1 Επιλογή του καλύτερου μοντέλου παλινδρόμησης

Υπάρχουν αρκετά πεδία στα οποία μπορεί να εφαρμοστούν τα μοντέλα παλινδρόμησης. Οι τέσσερις γενικές κατηγορίες εφαρμογής των μοντέλων παλινδρόμησης, είναι οι εξής:

- *Περιγραφή δεδομένων*, για την περιγραφή μιας δεδομένη διαδικασίας ή τη μοντελοποίηση ενός πολύπλοκου συστήματος.
- *Πρόβλεψη και εκτίμηση*, για την πρόβλεψη μελλοντικών παρατηρήσεων ή την εκτίμηση της μέσης τιμής για την απόκριση.
- *Εκτίμηση παραμέτρων*, όπου έμφαση πρέπει να δοθεί στο θόρυβο, ο οποίος προέρχεται από τη διαγραφή μεταβλητών από το μοντέλο, και στις διακυμάνσεις των εκτιμώμενων συντελεστών.
- *Έλεγχος*, όπου απαιτείται ακριβής εκτίμηση των παραμέτρων (και άρα τα τυπικά σφάλματα των συντελεστών της παλινδρόμησης πρέπει να είναι μικρά).

Είναι φανερό η σημασία που είναι δυνατόν να αποδώσει κανείς στις πρακτικές εφαρμογές της ανάλυσης παλινδρόμησης. Με βάση τη διαπίστωση αυτή, μπορεί να δικαιολογηθεί το ενδιαφέρον για την ανάπτυξη μεθοδολογιών οι οποίες δεν είναι απαραίτητο να εξετάζονται όλα τα διαθέσιμα δεδομένα, καθώς ο υπολογιστικός φόρτος και χρόνος που απαιτούνται σε αυτή την περίπτωση, μπορεί να είναι σημαντικά μικρότεροι. Στη συνέχεια περιγράφεται το πρόβλημα επιλογής του καλύτερου υποσυνόλου μεταβλητών.

Έστω ότι απαιτείται ο υπολογισμός μιας γραμμικής συνάρτησης παλινδρόμησης για μια συγκεκριμένη εξαρτημένη μεταβλητή (απόκριση) Y , με βάση τις ανεξάρτητες μεταβλητές ή μεταβλητές πρόβλεψης (predictor variables) $X_1, X_2, \dots$,

X_k . Έστω επίσης ότι $Z_1, Z_2, \dots, Z_r$ είναι οι συναρτήσεις μιας ή περισσότερων ανεξάρτητων μεταβλητών X , οι οποίες αναπαριστούν ολόκληρο το σύνολο των μεταβλητών από τις οποίες θα επιλεγεί η συνάρτηση και ότι το σύνολο αυτό περιέχει οποιεσδήποτε συναρτήσεις, όπως τετράγωνα, γινόμενο ανυσμάτων, λογαρίθμους κτλ. για την επιλογή της συνάρτησης, χρησιμοποιούνται δύο, κυρίως, κριτήρια, τα οποία είναι και ανταγωνιστικά μεταξύ τους. Αυτά είναι:

- Εάν το ζητούμενο είναι η συνάρτηση που θα επιλεγεί να χρησιμοποιηθεί για προβλεπτικούς σκοπούς, τότε θα πρέπει να το μοντέλο να περιλαμβάνει όσο το δυνατόν περισσότερες συναρτήσεις Z , έτσι ώστε να καθοριστούν αξιόπιστες προσαρμοσμένες τιμές, βασιζόμενες στην πληροφορία που περιέχουν οι Z .
- Για μεγάλο αριθμό συναρτήσεων Z , το κόστος παρακολούθησης και απολαβής πληροφορίας από αυτές είναι μεγάλο, οπότε θα πρέπει κατά την επιλογή της συνάρτησης παλινδρόμησης αυτή να περιλαμβάνει όσο το δυνατόν λιγότερες συναρτήσεις Z . Με τον τρόπο αυτό μειώνεται και η διακύμανση για την εκτίμηση $\hat{Y}$ της Y .

Η διαδικασία συμβιβασμού μεταξύ αυτών των δύο κριτηρίων για την επιλογή της συνάρτησης παλινδρόμησης, συνήθως ονομάζεται επιλογή της καλύτερης συνάρτησης παλινδρόμησης. Για το σκοπό αυτό, δεν υπάρχει κάποια στατιστική μέθοδος. Αν το μέγεθος της πραγματικής τυχαίας διακύμανσης, σ^2 , των παρατηρήσεων είναι γνωστό, τότε η επιλογή θα ήταν πιο εύκολη. Πρακτικά, κάτι τέτοιο δε συμβαίνει συχνά, οπότε είναι απαραίτητη μια υποκειμενική κρίση από τον αναλυτή για την επιλογή.

Γενικά, έχουν διατυπωθεί διάφορες σκέψεις και προβληματισμοί για το αν ένα μοντέλο παλινδρόμησης είναι προτιμότερο να διαθέτει μικρό αριθμό μεταβλητών. Οι απόψεις αυτές, μπορούν, με βάση το περιεχόμενό τους, να χωριστούν σε δύο κατηγορίες: στατιστικού και μη στατιστικού περιεχομένου. Η κατηγορία των μη στατιστικών απόψεων περιλαμβάνει ισχυρισμούς που έχουν να κάνουν με παράγοντες οικονομικούς, κοινωνικούς, οικονομίας σκέψης, πρακτικότητας, αισθητικούς και υπολογιστικούς. Τέτοιοι ισχυρισμοί είτε μπορούν να εκφραστούν με «επίσημους» όρους και να ιδωθούν υπό το πρίσμα των σκοπών των μεθόδων παλινδρόμησης είτε να χρησιμοποιηθούν εκ των υστέρων, όταν συγκριθούν τα μοντέλα που προκύπτουν από διαφορετικές μεθόδους παλινδρόμησης. Η κατηγορία των στατιστικών απόψεων, κυρίως περιφέρεται γύρω από δύο θέματα: την ύπαρξη μεγάλου μεγέθους πινάκων συνδιακύμανσης για τους συντελεστές b και την φτωχή προβλεπτική ικανότητα νέων δεδομένων, όταν χρησιμοποιούνται υπεράριθμες μεταβλητές (Draper and Smith, 1981).

2.2 Επιλογή του «καλύτερου» υποσυνόλου σε προβλήματα παλινδρόμησης

Το πρόβλημα επιλογής του καλύτερου υποσυνόλου συνίσταται στην μοντελοποίηση μιας σχέσης μεταξύ μιας εξαρτημένης μεταβλητής Y και ενός υποσυνόλου από ένα δεδομένο σύνολο μεταβλητών $X_1, X_2, \dots, X_n$, χωρίς να είναι γνωστό το πιο υποσύνολο θα πρέπει να χρησιμοποιηθεί. Μια τέτοια διαδικασία είναι ιδιαίτερα χρήσιμη όταν ο αριθμός των μεταβλητών που είναι διαθέσιμες (n) είναι

πολύ μεγάλος και αν το σύνολο μεταβλητών $X_1, X_2, \dots, X_n$ θεωρείται ότι περιλαμβάνει περιττές ή άσχετες μεταβλητές. Γενικά, το πρόβλημα επιλογής του καλύτερου υποσυνόλου, μπορεί να θεωρηθεί ότι αποτελεί μια ειδική περίπτωση του προβλήματος επιλογής μοντέλου (για τη σχέση μεταξύ X_i και Y). Η επιλογή υποσυνόλων είναι περισσότερο συνδεδεμένη με το πλαίσιο της γραμμικής ανάλυσης παλινδρόμησης (George, 2000).

Για την επιλογή της καλύτερης συνάρτησης παλινδρόμησης, έχουν προταθεί διάφορες μέθοδοι, οι οποίες απαντώνται και σε πρακτικά προβλήματα. Οι μέθοδοι αυτοί είναι πιθανόν να μην καταλήξουν στο ίδιο αποτέλεσμα, ακόμα και αν εφαρμοσθούν στο ίδιο πρόβλημα. Η επιλογή αυτή λαμβάνει χώρα σε ένα «ιδανικό» περιβάλλον, υποθέτοντας ότι οι σωστές συναρτήσεις που περιέχονται στα σύνολα Z είναι γνωστά και ότι δεν υφίστανται ακραίες παρατηρήσεις στο δείγμα, γεγονός το οποίο δεν παρατηρείται συχνά στην πράξη. Μερικές από αυτές τις μεθόδους είναι και οι:

1. Όλες οι πιθανές συναρτήσεις παλινδρόμησης, με χρήση τριών κριτηρίων: δείκτες R^2 , s^2 και C_p του Mallows.
2. Το καλύτερο υποσύνολο συναρτήσεων παλινδρόμησης.
3. Η προς τα πίσω απαλοιφή (backward elimination).
4. Η βηματική (stepwise) παλινδρόμηση.
5. Η μέθοδος Ridge Regression.
6. Η μέθοδος PRESS.
7. Η παλινδρόμηση σε κύριες συνιστώσες (principal components regression).
8. Η παλινδρόμηση λανθάνουσας ρίζας (latent root regression).
9. Η παλινδρόμηση κατά φάση (stagewise regression).

Πολλές από τις μεθόδους αυτές, δεν επιλέγουν το απόλυτα καλύτερο μοντέλο παλινδρόμησης, αλλά ένα αποδεκτό μοντέλο. Με καμία από αυτές δεν μπορεί να εξασφαλιστεί ότι θα επιλεγεί το καλύτερο μοντέλο για ένα δεδομένο σύνολο μεταβλητών. Η εμπειρία, η επαγγελματική υποκειμενική κρίση του αναλυτή και άλλες υποκειμενικές παραδοχές συμβάλλουν στην τελική επιλογή του μοντέλου παλινδρόμησης. Για τη βελτίωση των αποτελεσμάτων επιλογής του καλύτερου μοντέλου, έχουν χρησιμοποιηθεί και μερικές παραλλαγές των μεθόδων 1 – 4. Γενικά, η διαδικασία επιλογής με βάση τις παραπάνω μεθόδους, πραγματοποιείται επαναληπτικά, επιλέγοντας πρώτα τη στρατηγική επιλογής των μεταβλητών για είσοδο στο μοντέλο και έπειτα λαμβάνει χώρα ο έλεγχος του μοντέλου που προκύπτει για πιθανές ακραίες τιμές των παρατηρήσεων και συσχετίσεις μεταξύ τους.

Με τη διαγραφή μεταβλητών από το μοντέλο, μπορεί να επιτευχθεί καλύτερη ακρίβεια του τελευταίου, ακόμα και αν μερικές από αυτές δεν είναι αμελητέας σημαντικότητας. Κάτι τέτοιο μπορεί να αληθεύει και στην περίπτωση της διακύμανσης της προβλεπόμενης τιμής για την απόκριση του μοντέλου. Με τη διαγραφή μεταβλητών, όμως, μπορεί να εισάγεται θόρυβος για την εκτίμηση των συντελεστών των μεταβλητών. Αν οι μεταβλητές που διαγράφονται έχουν μικρό αντίκτυπο για τον υπολογισμό της απόκρισης, τότε το μέσο τετραγωνικό σφάλμα (MSE) των εκτιμήσεων με θόρυβο θα είναι μικρότερο από τη διακύμανση των εκτιμήσεων χωρίς θόρυβο. Δηλαδή το επίπεδο του θορύβου που εισάγεται στο μοντέλο είναι μικρότερο από τη μείωση που παρατηρείται στη διακύμανση.

Πέραν των παραπάνω μεθόδων, έχουν προταθεί και άλλες. Μια σύντομη περιγραφή τέτοιων μεθόδων, φαίνεται στη συνέχεια. Στα πλαίσια της βηματικής παλινδρόμησης, εξέχουσα θέση κατέχει η τεχνική της λογιστικής παλινδρόμησης (logistic regression), όπου η εξαρτημένη μεταβλητή παίρνει τιμές 0 και 1. Οι υπολογισμοί είναι σχετικά χρονοβόροι, λόγω της επαναληπτικής διαδικασίας που ακολουθείται.

Στην περίπτωση της διαδοχικής αντικατάστασης μεταβλητών στο μοντέλο (sequential replacement algorithms), η βασική ιδέα είναι να αντικαταστήσουμε μια ήδη υπάρχουσα στο μοντέλο μεταβλητή, έτσι ώστε να οδηγηθούμε σε μικρότερο άθροισμα τετραγώνων των υπολοίπων. Για παράδειγμα, εάν ένα μοντέλο διαθέτει ήδη τις μεταβλητές $AB\Gamma\Delta$, ξεκινούμε δοκιμάζοντας να αντικαταστήσουμε την μεταβλητή A με κάποια άλλη εκτός του μοντέλου. Εάν προκύψει μικρότερο RSS , εκτελούμε την αντικατάσταση. Το ίδιο γίνεται για τις μεταβλητές B , Γ , και Δ . Η διαδικασία σταματά όταν δεν είναι δυνατόν να προκύψει μικρότερο RSS με αντικατάσταση κάποιας από τις μεταβλητές. Το πρόβλημα εδώ είναι ότι θα πρέπει να γίνει εξαντλητικός έλεγχος όλων των διαθέσιμων μεταβλητών και συνδυασμών. Το πρόβλημα αυτό λύνεται με τη χρήση διάφορων τεχνικών, όπως για παράδειγμα η τεχνική branch-and-bound.

Σε προβλήματα με πολλές μεταβλητές (εκατοντάδες π.χ.), η αντικατάσταση δύο μεταβλητών ταυτόχρονα φαίνεται να οδηγεί στην εύρεση καλύτερων υποσυνόλων με καλύτερη πιθανότητα. Ένας αλγόριθμος ο οποίος χρησιμοποιήθηκε με μεγάλη επιτυχία για το σκοπό αυτό, όπου τα υποσύνολα που εξετάζονται διαθέτουν κατά το μέγιστο p μεταβλητές, είναι και ο παρακάτω (Miller, 2002):

1. Δημιουργία τυχαίων υποσυνόλων, αποτελούμενων από περίπου $p/2$ μεταβλητές.
2. Στο κάθε υποσύνολο, λαμβάνονται υπόψη όλοι οι δυνατοί συνδυασμοί αντικατάστασης ζευγών μεταβλητών. Οι συνδυασμοί αυτοί περιλαμβάνουν την αντικατάσταση και ενός μόνο ζεύγους μεταβλητών ή την ταυτόχρονη αντικατάσταση περισσότερων ζευγών.
3. Εύρεση της καλύτερης αντικατάστασης. Εάν το RSS μειώνεται, λαμβάνει χώρα η αντικατάσταση και γίνεται επαναφορά στο βήμα 2.
4. Εάν το υποσύνολο των μεταβλητών που εξετάζεται δεν έχει το μέγιστο μέγεθος p , τότε εισάγεται σε αυτό οποιαδήποτε μεταβλητή και γίνεται επαναφορά στο βήμα 2.
5. Επανάληψη της διαδικασίας από το βήμα 1, έως ότου ικανοποιηθεί η απαίτηση για p μεταβλητές.

Οι τεχνικές branch-and-bound συνίστανται στη διαίρεση και ομαδοποίηση όλων των δυνατών υποσυνόλων από p μεταβλητές από ένα σύνολο n μεταβλητών, έτσι ώστε, κάθε «κλάδος» να περιέχει πχ μια συγκεκριμένη μεταβλητή. Μετά μπορούν να δημιουργηθούν υποομάδες («υποκλάδοι»), οι οποίοι να απαρτίζονται από υποσύνολα του αρχικού κλάδου. Στόχος μπορεί να είναι πχ η εύρεση ενός υποσυνόλου με 5 μεταβλητές το οποίο να δίδει το μικρότερο RSS . Η τεχνική αυτή είναι ιδιαίτερη χρήσιμη όταν υπάρχουν μεταβλητές οι οποίες πρέπει οπωσδήποτε να συμμετέχουν στα καλύτερα υποσύνολα (Miller, 2002).

Άλλες τεχνικές που έχουν προταθεί, είναι της ομαδοποίησης μεταβλητών (grouping variables), η παλινδρόμηση ridge (ridge regression), η nonnegative garrote

και η lasso (least absolute shrinkage and selection operator). Περισσότερες λεπτομέρειες για τις τεχνικές αυτές, μπορούν να βρεθούν στο (Miller, 2002).

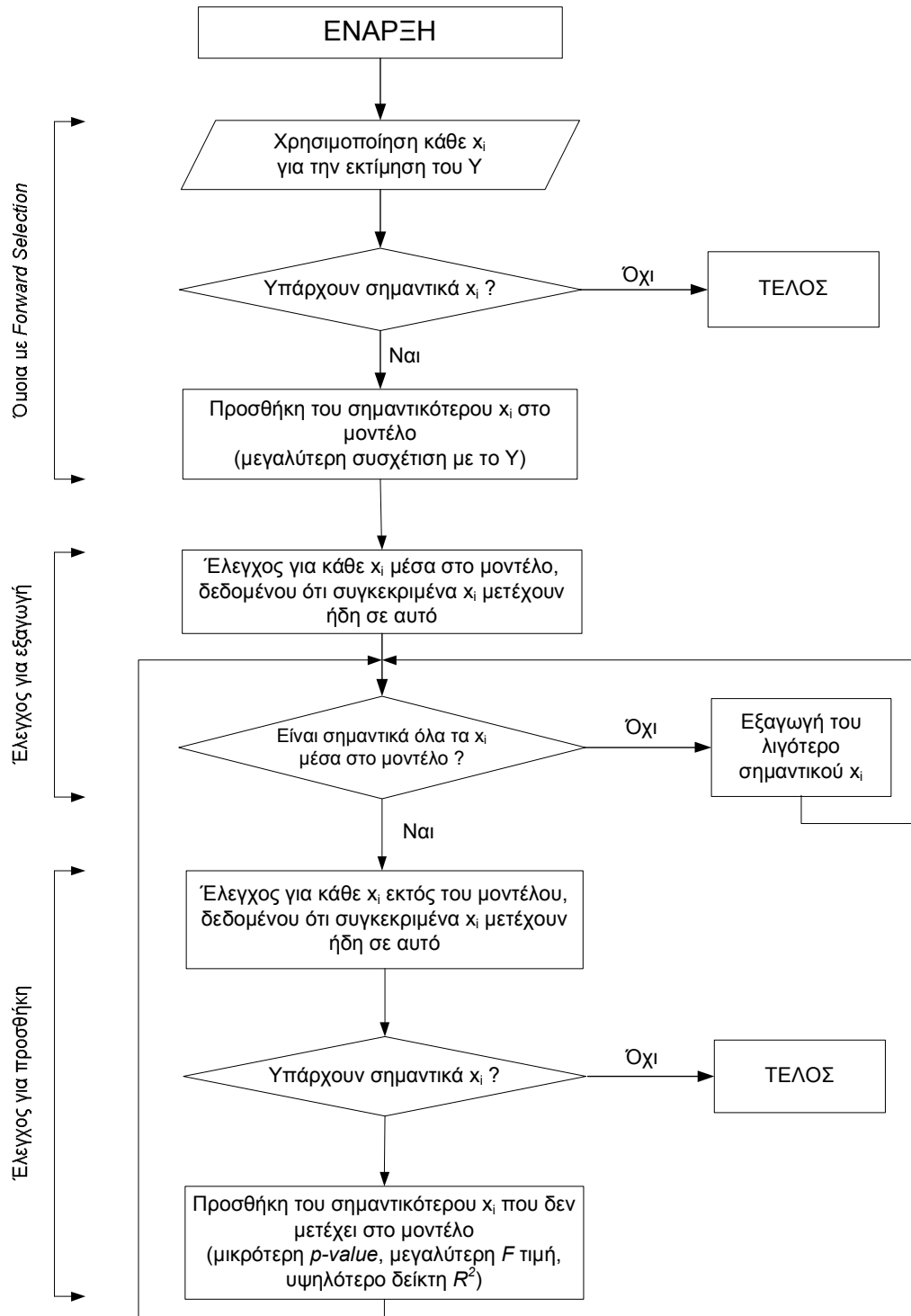
2.3 Επιλογή του καλύτερου υποσυνόλου μεταβλητών με χρήση μεθόδων βηματικής παλινδρόμησης

Στα πλαίσια των μεθόδων αυτών, επιχειρείται διαδοχικά η εισαγωγή ή διαγραφή μεταβλητών από την συνάρτησης παλινδρόμησης, μέχρις ότου αυτή να είναι ικανοποιητική. Το ποιες μεταβλητές και με ποια σειρά θα εξεταστούν για εισαγωγή στη συνάρτηση, εξαρτάται από το μέτρο του συντελεστή μερικής συσχέτισης (partial correlation coefficient) σε σχέση με την εξαρτημένη μεταβλητή Y . Ο συντελεστής αυτός χρησιμοποιείται ως μέτρο για τη σημαντικότητα των μεταβλητών οι οποίες πρόκειται να εξεταστούν για εισαγωγή στο μοντέλο. Γενικά, οι μέθοδοι αυτοί μπορούν να ενταχθούν σε τρεις ευρύτερες κατηγορίες τέτοιων μεθόδων: 1) προς-τα-εμπρός επιλογή, 2) προς-τα-πίσω απαλοιφή και 3) βηματική παλινδρόμηση, η οποία είναι συνδυασμός των δύο προηγούμενων κατηγοριών.

2.3.1 Βηματική παλινδρόμηση (Stepwise Regression)

Ο αλγόριθμος της βηματικής παλινδρόμησης παρουσιάστηκε από τον Efroymson το 1960. Αποτελεί ένα συνδυασμό των μεθόδων της προς-τα-εμπρός επιλογής και της προς-τα-πίσω απαλοιφής. Η διαδικασία που ακολουθείται εδώ, είναι η εξής: αρχικά επιλέγεται το πιο συσχετισμένο με την εξαρτημένη μεταβλητή Y σύνολο Z , έστω το Z_1 , και υπολογίζεται η πρώτη τάξεως γραμμική συνάρτηση παλινδρόμησης, $\hat{Y} = f(Z_1)$. Στη συνέχεια ελέγχεται αν η μεταβλητή αυτή είναι σημαντική. Αν δεν είναι, η διαδικασία σταματά και ως καλύτερο μοντέλο επιλέγεται το $Y = \bar{Y}$, αλλιώς, η διαδικασία συνεχίζεται με την επόμενη προς εισαγωγή μεταβλητή. Για το σκοπό αυτό ελέγχονται όλοι οι συντελεστές μερικής συσχέτισης για όλες τις μεταβλητές (έστω Z_j , $j \neq 1$) που δε μετέχουν, ως αυτό το στάδιο, στη συνάρτηση, σε σχέση με την Y . Αυτό σημαίνει ότι οι Y και Z_j έχουν προσαρμοστεί, σε σχέση με την Z_1 , ως προς τις σχέσεις γραμμικότητάς τους, και η συσχέτιση υπολογίζεται για όλα τα $j \neq 1$. Από μαθηματικής άποψης, αυτό είναι ισοδύναμο με τον υπολογισμό των συσχετίσεων μεταξύ 1) των υπολοίπων από τη συνάρτηση παλινδρόμησης $\hat{Y} = f(Z_1)$ και 2) των υπολοίπων από κάθε από τις j συναρτήσεις παλινδρόμησης $\hat{Z}_j = f_j(Z_1)$, οι οποίες, στην πραγματικότητα, δεν έχουν υπολογιστεί. Έτσι, επιλέγεται το Z_j με το μεγαλύτερο συντελεστή μερικής συσχέτισης, έστω το Z_2 , και υπολογίζεται μια νέα συνάρτηση παλινδρόμησης, της μορφής $\hat{Y} = f(Z_1, Z_2)$. Η συνάρτηση αυτή εξετάζεται ως προς τη σημαντικότητά της και σημειώνεται η τυχόν βελτίωση στην τιμή R^2 (τετράγωνο του συντελεστή πολλαπλής συσχέτισης). Επίσης, εξετάζεται και οι τιμές των μερικών F και για τις δύο μεταβλητές που υπάρχουν πλέον στο μοντέλο. Η μικρότερη από αυτές τις τιμές συγκρίνεται με ένα κατάλληλο προεπιλεγμένο επίπεδο (ποσοστό) για τις τιμές F και ανάλογα με τη σύγκριση αυτή, η αντίστοιχη μεταβλητή περιλαμβάνεται στο μοντέλο ή απορρίπτεται. Ας σημειωθεί ότι αυτή η διαδικασία σύγκρισης, επαναλαμβάνεται σε κάθε στάδιο της μεθόδου και μπορεί να οδηγήσει σε απόρριψη μιας μεταβλητής η οποία σε προηγούμενο στάδιο φαινόταν κατάλληλη για εισαγωγή, εξαιτίας των συσχετίσεων που παρουσιάζουν οι

υπόλοιπες εισαχθείσες μεταβλητές στο μοντέλο. Η διαδικασία αυτή παρέχει μια κρίση για τη συνεισφορά της λιγότερο «πολύτιμης» μεταβλητής στο μοντέλο της παλινδρόμησης σε κάθε στάδιο της μεθόδου, ανεξάρτητα με τον αν αυτή εισήχθη στο μοντέλο στην τελευταία δοκιμή ή όχι. Η μεταβλητή αυτή μπορεί ανάλογα με το αποτέλεσμα της σύγκρισης να παραμείνει ή να απορριφθεί, ενώ η κατάλληλη συνάρτηση παλινδρόμησης υπολογίζεται για τις μεταβλητές που παραμένουν στο τέλος κάθε σταδίου της μεθόδου.



Σχήμα 2.1: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου βηματικής παλινδρόμησης.

Στη συνέχεια, για την καλύτερη μεταβλητή η οποία δεν έχει εισαχθεί έως αυτό το στάδιο στο μοντέλο (π.χ. αυτή που παρουσιάζει το μεγαλύτερο συντελεστή μερικής συσχέτισης με την Y , δοθέντων των μεταβλητών οι οποίες ήδη μετέχουν στο μοντέλο), ελέγχεται για το αν πληροί την προϋπόθεση εισαγωγής στο μοντέλο (σύγκριση με το επιλεγμένο επίπεδο της τιμής F). Αν πληροί την προϋπόθεση, εισάγεται στο μοντέλο και επαναλαμβάνεται ο έλεγχος των μερικών F τιμών για όλες τις μεταβλητές που συμμετέχουν, αν όχι, απορρίπτεται. Σταδιακά, η όλη διαδικασία σταματά όταν στο παρόν στάδιο δεν υπάρχουν άλλες μεταβλητές για απόρριψη και η επόμενη καλύτερη μεταβλητή για εισαγωγή δεν μπορεί να διατηρήσει τη θέση της μέσα στο μοντέλο. Συνήθως, κατά τη διαδικασία εισαγωγής μιας μεταβλητής στο μοντέλο, καταγράφεται και η αλλαγή που επιφέρει στην τιμή R^2 (Draper and Smith, 1981).

Εάν ένα μοντέλο περιέχει p μεταβλητές και παρουσιάζει άθροισμα τετραγώνων υπολοίπων RSS_p , η ποσότητα με την οποία συγκρίνεται η στατιστική F_{in} για είσοδο στο μοντέλο μιας μεταβλητής, φαίνεται παρακάτω (πρέπει $F_{in} < R$):

$$R = \frac{RSS_p - RSS_{p+1}}{RSS_{p+1} / (n - p - 1)}$$

όπου n το σύνολο των μεταβλητών, p οι ήδη υπάρχουσες στο μοντέλο μεταβλητές και RSS_{p+1} το άθροισμα τετραγώνων των υπολοίπων με την εισαγωγή της νέας μεταβλητής (μικρότερο από το RSS_p). Για την εξαγωγή μεταβλητής, η στατιστική F_{out} συγκρίνεται με την ποσότητα (πρέπει $F_{out} > R$):

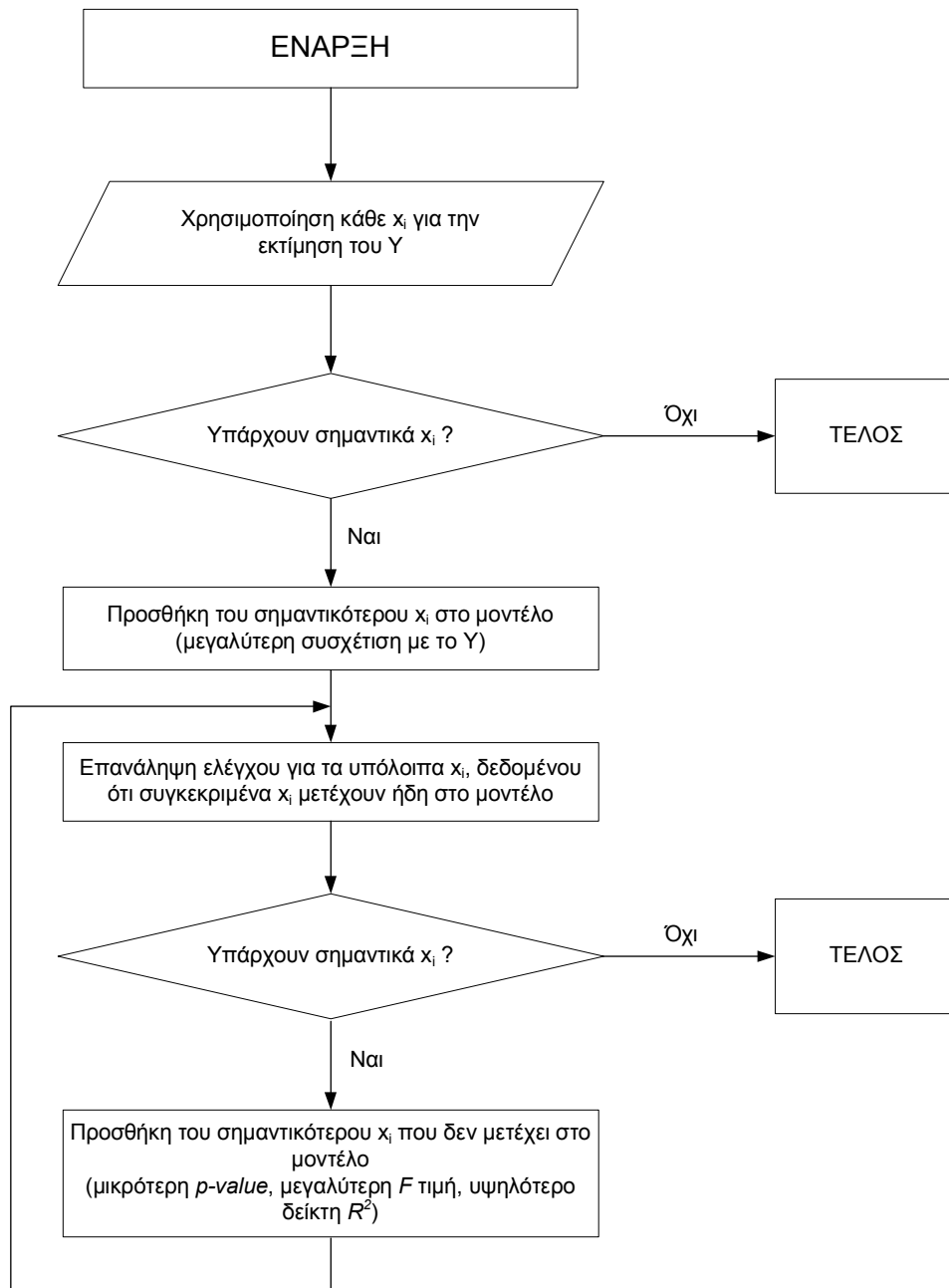
$$R = \frac{RSS_{p-1} - RSS_p}{RSS_p / (n - p)}$$

όπου RSS_{p-1} είναι το άθροισμα τετραγώνων των υπολοίπων που προκύπτει από την εξαγωγή κάποιας μεταβλητής (μικρότερο από το RSS_p). Η διαδικασία σταματά, όταν δεν μπορούν να μουν ή να εξαχθούν μεταβλητές στο μοντέλο.

2.3.2 Προς τα εμπρός επιλογή (Forward Selection)

Στη διαδικασία αυτή, γίνεται η υπόθεση ότι στο μοντέλο δεν υπάρχει καμμία μεταβλητή, εκτός ίσως από τον σταθερό όρο, εάν υιοθετηθεί η χρήση του. Στη συνέχεια γίνεται προσπάθεια κάποιες μεταβλητές να μουν διαδοχικά (μία-μία) στο μοντέλο. Αρχικά, στο μοντέλο εξετάζεται για εισαγωγή η μεταβλητή αυτή η οποία εμφανίζει την μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή. Αυτή είναι και η μεταβλητή που θα εμφανίσει τη μεγαλύτερη τιμή για την παράμετρο (στατιστική) F (F -statistic). Εάν η τιμή αυτή ξεπερνά μια προκαθορισμένη τιμή για εισαγωγή στο μοντέλο, έστω F_{in} , τότε εισάγεται σε αυτό. Με τον τρόπο αυτό, ελέγχεται η σημαντικότητα της εξεταζόμενης για είσοδο μεταβλητής. Εν συνεχεία, εξετάζεται η επόμενη μεταβλητή για είσοδο στο μοντέλο. Αυτή θα είναι αυτή που διαθέτει την αμέσως μεγαλύτερη συσχέτιση με την εξαρτημένη μεταβλητή. Ας σημειωθεί ότι οι επιμέρους αυτές συσχετίσεις των μεταβλητών ονομάζονται και μερικές συσχετίσεις (partial correlations). Πάλι εξετάζεται η σημαντικότητα της προς εισαγωγή μεταβλητής και αν η τιμή για την παράμετρο F αυτής ξεπερνά την F_{in} , δεδομένου ότι

ήδη συμμετέχει στο μοντέλο κάποια γνωστή μεταβλητή, τότε εισάγεται στο μοντέλο. Η διαδικασία συνεχίζεται με παρόμοιο τρόπο και τερματίζεται όταν είτε καμιά τιμή για την παράμετρο F δεν ξεπερνά την F_{in} , είτε όταν εισαχθεί στο μοντέλο και η τελευταία από τις διαθέσιμες μεταβλητές.

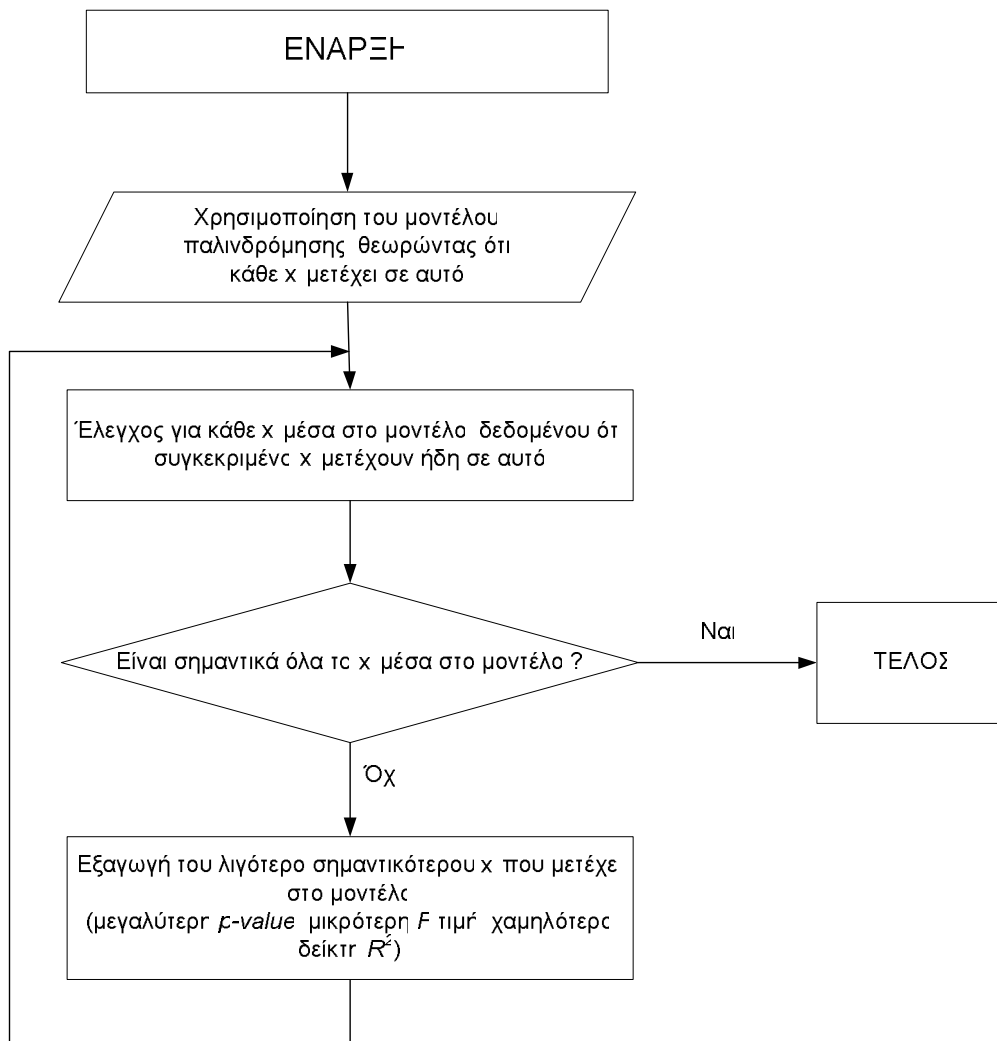


Σχήμα 2.2: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου προς τα εμπρός επιλογής.

2.3.3 Προς τα πίσω απαλοιφή (Backward Elimination)

Αυτή η διαδικασία λειτουργεί με την αντίθετη λογική της προς τα εμπρός επιλογής. Έτσι, αρχικά στο μοντέλο υπάρχουν όλες οι k , έστω, διαθέσιμες μεταβλητές. Υπολογίζονται στη συνέχεια οι τιμές της παραμέτρου F για κάθε μεταβλητή, σαν να επρόκειτο να ήταν η τελευταία για εισαγωγή στο μοντέλο. Η μικρότερη από αυτές τις

τιμές, συγκρίνεται με μια προκαθορισμένη τιμή για την F η οποία οριοθετεί την τιμή για την έξοδο μιας μεταβλητής από το μοντέλο (έστω F_{out}). Εάν η μικρότερη από τις τιμές για την F που υπολογίστηκαν είναι μικρότερη και από την τιμή F_{out} , τότε η αντίστοιχη μεταβλητή εξάγεται από το μοντέλο. Έπειτα, υπολογίζονται οι τιμές για την F για τις υπόλοιπες μεταβλητές στο μοντέλο, δεδομένου ότι μια γνωστή μεταβλητή εξήλθε του μοντέλου. Πάλι γίνεται η σύγκριση με την τιμή F_{out} και η διαδικασία συνεχίζεται με παρόμοιο τρόπο, έως ότου καμμία από τις τιμές της F για κάθε μεταβλητή μέσα στο μοντέλο δεν είναι μικρότερη από την F_{out} (Miller, 2000).



Σχήμα 2.3: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου προς τα πίσω απαλοιφής.

2.4 Παραλλαγές της μεθόδου βηματικής παλινδρόμησης

Στο σημείο αυτό θα περιγραφούν δύο παραλλαγές της μεθόδου βηματικής παλινδρόμησης. Η πρώτη αφορά την εφαρμογή της μεθόδου με δεδομένα επίπεδα αποδοχής και απόρριψης. Όταν η διαδικασία επιλογής σταματά, γίνεται ο καθορισμός των μεταβλητών στο τελικό επιλεγθέν μοντέλο. Με τη χρησιμοποίηση των μεταβλητών αυτών, έστω q στον αριθμό, σχηματίζονται όλα τα πιθανά σύνολα των q μεταβλητών από τις R αρχικές μεταβλητές και επιλέγεται το καλύτερο σύνολο.

Σχετικά με αυτή την παραλλαγή, έχουν παρατεθεί και ανάλογα σχόλια για την εφαρμογή της. Πιο συγκεκριμένα, κατά την εφαρμογή της παραλλαγής αυτής, μπορεί να παρουσιαστεί μια κατάσταση κατά την οποία να βρεθούν, όχι μία, αλλά δύο μεταβλητές υποψήφιες για εισαγωγή στο μοντέλο με παραπλήσιους συντελεστές μερικής συσχέτισης. Τότε θα μπορούσε κανείς να ισχυριστεί ότι δεν υπάρχει ικανή πληροφορία μέσα στα δεδομένα ώστε να πραγματοποιηθεί μια ξεκάθαρη επιλογή, καθιστώντας απαραίτητες άλλες *a priori* παραδοχές και την υποκειμενική κρίση του αναλυτή για την επιλογή του τελικού μοντέλου. Η παραπάνω παραλλαγή, μπορεί να αποτύχει στην περίπτωση κατά την οποία ένα μεγάλο σύνολο από μεταβλητές θα μπορούσε να είναι καλύτερο, αλλά δεν εξετάζεται ποτέ από τον αλγόριθμο της βηματικής παλινδρόμησης. Τέλος, σε ό,τι αφορά τον υπολογιστικό φόρτο της παραλλαγής αυτής, είναι μεγαλύτερος σε σχέση με τον αρχικό αλγόριθμο.

Η δεύτερη παραλλαγή σχετίζεται με τη χρησιμοποίηση λιγότερο περιοριστικών επιπέδων αποδοχής και απόρριψης στα πλαίσια της βηματικής παλινδρόμησης. Αυτό μπορεί να οδηγήσει στην αποδοχή περισσότερων ανεξάρτητων μεταβλητών, οι οποίες με τον αρχικό αλγόριθμο δε θα επιλέγονταν, εξετάζοντας έτσι τη συνεισφορά τους στο τελικό μοντέλο. Είναι πιθανό η διαδικασία να καταλήξει και σε διαφορετικό μοντέλο από αυτό της αρχικής μεθόδου.

Όπως και στην προηγούμενη περίπτωση και εδώ έχουν αναπτυχθεί διάφορες κριτικές σχετικά με την εφαρμογή της συγκεκριμένης παραλλαγής. Αυτή έχει αποδειχθεί χρήσιμη σε συγκεκριμένες περιπτώσεις. Πιο ειδικά, προσφέρει ένα εναλλακτικό μοντέλο με χαρακτηριστικά παρόμοια με αυτά της αρχικής μεθόδου, όταν το πρόβλημα που εξετάζεται είναι δύσκολο να επιλυθεί εξαιτίας των υψηλών συσχετίσεων μεταξύ των ανεξάρτητων μεταβλητών.

2.5 Γενική κριτική στη βηματική παλινδρόμηση

Σε ό,τι αφορά τις μεθόδους που ανήκουν στην οικογένεια της βηματικής παλινδρόμησης, έχουν αναπτυχθεί διάφορες κριτικές. Το κοινό στοιχείο των κριτικών αυτών εντοπίζεται στον ισχυρισμό ότι καμμία από τις μεθόδους αυτές δε μπορεί να εξασφαλίσει την ανάπτυξη του καλύτερου, κάθε φορά, μοντέλου παλινδρόμησης, ανεξαρτήτως μεγέθους. Ένα ακόμα σχόλιο σχετικά με τις μεθόδους αυτές, είναι ότι καταλήγουν πάντα σε μια συνάρτηση παλινδρόμησης, η οποία μπορεί να θεωρηθεί από μη πεπειραμένους αναλυτές ως βέλτιστη, ενώ ίσως δεν είναι. Σε αυτό το πρόβλημα συμβάλλει και το γεγονός ότι είναι πιθανόν να μην υπάρχει μόνο ένα πιθανό βέλτιστο υποσύνολο παραμέτρων, αλλά πολλά για την μοντελοποίηση ενός προβλήματος.

Κατά τη διαδικασία υλοποίησης των μεθόδων αυτών, η σειρά με την οποία οι μεταβλητές εισέρχονται στην συνάρτηση παλινδρόμησης, δεν πρέπει να συγχέεται με τη σημαντικότητα που έχουν αυτές, διότι μπορεί κάποια μεταβλητή που εισήλθε σε κάποιο στάδιο, σε κάποιο επόμενο να απορριφθεί. Τέλος, η εφαρμογή των μεθόδων αυτών, ακόμα και στο ίδιο πρόβλημα, δεν είναι απαραίτητο να οδηγήσει στο ίδιο τελικό μοντέλο. Αυτό οφείλεται στις συσχετίσεις μεταξύ των μεταβλητών που επιλέγονται για εισαγωγή στο μοντέλο, οι οποίες επηρεάζουν τη σειρά με την οποία εξετάζονται οι τελευταίες. Για όλους αυτούς τους λόγους, είναι απαραίτητο οι μέθοδοι της βηματικής παλινδρόμησης να χρησιμοποιούνται με προσοχή.

Ο Berk (1978) αναφέρει ότι ακόμα και μεταξύ τους οι μέθοδοι της βηματικής παλινδρόμησης είναι δυνατόν να μη συμφωνούν. Αυτό μπορεί να εμφανιστεί ιδιαίτερα για ακραίες περιπτώσεις δεδομένων (τα οποία δηλαδή, δεν απαντώνται συχνά στην πραγματικότητα). Ειδικότερα, γίνεται ο ισχυρισμός ότι η προς τα εμπρός επιλογή και η προς τα πίσω απαλοιφή, συμφωνούν μεταξύ τους στη διαδοχή των επιμέρους υποσυνόλων μέχρι την ανάδειξη του «καλύτερου» υποσυνόλου, μόνο στην περίπτωση που υπάρχει μια αυθαίρετα μεγάλη ανακολουθία με τα αποτελέσματα που δίνει η εξαντλητική εξέταση όλων των δυνατών υποσυνόλων. Ακόμα, μέσω συγκεκριμένων παραδειγμάτων, δεικνύεται ότι η προς τα εμπρός επιλογή είναι καλύτερη της προς τα πίσω απαλοιφής, όταν πρόκειται για μικρά υποσύνολα.

Ένα άλλο θέμα που προκύπτει με τη χρησιμοποίηση τέτοιων μεθόδων, είναι το εάν μπορεί να είναι κάποιος σίγουρος ότι οι μεταβλητές που δεν επιλέχθηκαν τελικά για το μοντέλο, μπορούν να προσφέρουν επιπλέον πληροφορία για την εξαρτημένη μεταβλητή. Δηλαδή, να είναι κάποιος σίγουρος ότι το υποσύνολο που επιλέχθηκε είναι πράγματι το καλύτερο. Σε γενικές γραμμές, δεν υπάρχει κάποιος τρόπος διαβεβαίωσης για κάτι τέτοιο. Ένας τρόπος με τον οποίο προτάθηκε να γίνεται έλεγχος του παραπάνω ζητήματος, είναι ο ορισμός τεχνητών μεταβλητών στο σύνολο των μεταβλητών, ώστε όταν εισαχθεί μια τέτοια, να σταματήσει η διαδικασία. Αυτό όμως, μπορεί να σημαίνει ότι στο μοντέλο είναι δυνατόν να εισαχθούν και μεταβλητές χωρίς συνεισφορά και ακόμα ότι ο ορισμός των τεχνητών μεταβλητών μπορεί να οδηγήσει σε αύξηση του υπολογιστικού φόρτου (Miller, 2000).

Τα πλεονεκτήματα, από την άλλη πλευρά, τα οποία παρουσιάζουν οι μέθοδοι αυτοί, είναι ότι είναι γρήγορες, εύκολα υλοποιήσιμες σε ηλεκτρονικό υπολογιστή, ενώ υπάρχουν ήδη αρκετά έτοιμα λογισμικά για τη χρησιμοποίηση τέτοιων μεθόδων.

2.6 Κριτήρια Αξιολόγησης των Μοντέλων Παλινδρόμησης

2.6.1 Συντελεστής προσδιορισμού (coefficient of multiple determination)

Ένα γενικώς αποδεκτό μέτρο για την ορθότητα και την καταλληλότητα ενός επιλεγμένου μοντέλου παλινδρόμησης, είναι ο συντελεστής προσδιορισμού, R^2 . Αν το δείγμα που είναι διαθέσιμο αποτελείται από p όρους, δηλαδή $p - 1$ ανεξάρτητες μεταβλητές και ένα συντελεστή β_0 , τότε ο συντελεστής προσδιορισμού, R_p^2 , υπολογίζεται από τη Σχέση (2.1):

$$R_p^2 = \frac{SS_R(p)}{S_{yy}} = 1 - \frac{SS_E(p)}{S_{yy}} \quad (2.1)$$

όπου $SS_R(p)$ και $SS_E(p)$ είναι το άθροισμα των τετραγώνων της παλινδρόμησης και το άθροισμα των τετραγώνων των υπολοίπων αντίστοιχα. Ο συντελεστής συσχέτισης R^2 ορίζεται και ως εξής (Drapper and Smith, 1981):

$$R^2 = (\text{SS λόγω της παλινδρόμησης}) / (\text{Συνολικό SS, διορθωμένο για το μέσο } \bar{Y})$$

όπου:

$$SS \text{ λόγω παλινδρόμησης} = SS_{reg} = \frac{\left[\sum_i X_i Y_i - \left(\sum_i X_i \right) \left(\sum_i Y_i \right) / n \right]^2}{\left[\sum_i X_i^2 - \left(\sum_i X_i \right)^2 / n \right]}$$

$$SS \text{ διορθωμένο στο } \bar{Y} = SS_{YY} = \sum_i Y_i^2 - \left(\sum_i Y_i \right)^2 / n$$

Με βάση τα παραπάνω, ο συντελεστής συσχέτισης R^2 υπολογίζεται και από τη Σχέση (2.2):

$$R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = 1 - \frac{\sum_i (\hat{Y}_i - Y_i)^2}{\sum_i (Y_i - \bar{Y})^2} \quad (2.2)$$

Ο συντελεστής αυτός είναι ένα μέτρο του ποσοστού της συνολικής διακύμανσης σχετικά με το $\bar{Y}$ που εξηγείται με βάση την παλινδρόμηση. Ουσιαστικά, ο συντελεστής R είναι η συσχέτιση μεταξύ του Y και της εκτίμησης για το Y , $\hat{Y}$. Δηλαδή ισχύει $r_{Y\hat{Y}} = R$.

Η χρησιμοποίηση αυτού του κριτηρίου από τους αναλυτές πραγματοποιείται για να βρεθεί ο αριθμός των μεταβλητών που θα εισαχθούν στο μοντέλο, μέχρι του σημείου να βρεθεί η μεταβλητή εκείνη η οποία δεν συνεισφέρει σημαντικά στη βελτίωση του R^2 . Το σημείο αυτό καθορίζεται υποκειμενικά από τον αναλυτή, δεδομένου ότι το R^2 παρουσιάζει ξεκάθαρη βελτίωση μέχρι ένα σημείο εισαγωγής μεταβλητών στο μοντέλο και από εκεί και έπειτα σταθεροποιείται στη μέγιστη τιμή που μπορεί να λάβει. Αυτό σημαίνει ότι δεν υπολογίζεται η βέλτιστη τιμή που μπορεί να λάβει το R^2 , αλλά καθορίζεται ένα «ικανοποιητικό» επίπεδο αυτού. Μοντέλα τα οποία παρουσιάζουν μεγάλες τιμές για το R^2 προτιμούνται έναντι αυτών με χαμηλές τιμές.

2.6.2 Προσαρμοσμένος συντελεστής προσδιορισμού

Για να αποφευχθούν οι δυσκολίες που παρουσιάζει η ερμηνεία των τιμών για το R^2 , οι αναλυτές καθόρισαν μία άλλη στατιστική παράμετρο σε ό,τι αφορά το μέτρο καταλληλότητας ενός μοντέλου παλινδρόμησης. Αυτή η παράμετρος είναι ο προσαρμοσμένος συντελεστής προσδιορισμού, $\bar{R}^2$. Για ένα μοντέλο παλινδρόμησης με p όρους, ο προσαρμοσμένος συντελεστής προσδιορισμού, $\bar{R}_p^2$, υπολογίζεται από τη Σχέση (2.3):

$$\bar{R}_p^2 = 1 - \left(\frac{n-1}{n-p} \right) \cdot (1 - R_p^2) \quad (2.3)$$

όπου n το σύνολο των παρατηρήσεων για τις ανεξάρτητες μεταβλητές.

Το $\bar{R}_p^2$ δεν είναι απαραίτητο να αυξάνεται με την εισαγωγή κάποιων μεταβλητών στο μοντέλο. Αντίθετα, αποδεικνύεται ότι εάν s μεταβλητές εισαχθούν σε ένα μοντέλο με αρχικό $\bar{R}^2$ ίσο με $\bar{R}_p^2$, τότε το $\bar{R}_{p+s}^2$ θα ξεπεράσει σε τιμή το $\bar{R}_p^2$, εάν και μόνο εάν οι συντελεστές μερικής συσχέτισης των εισαχθέντων s μεταβλητών ξεπερνούν τη μονάδα. Τα μοντέλα παλινδρόμησης τα οποία παρουσιάζουν το μεγαλύτερο $\bar{R}_p^2$ θεωρούνται ως καλύτερα.

2.6.3 Μέσο άθροισμα των τετραγώνων των υπολοίπων

Το μέσο άθροισμα των τετραγώνων των υπολοίπων, ως κριτήριο επιλογής p μεταβλητών για εισαγωγή σε ένα μοντέλο παλινδρόμησης, μπορεί να υπολογιστεί από τη Σχέση (2.4):

$$MS_E(p) = \frac{SS_E(p)}{n-p} \quad (2.4)$$

Η συμπεριφορά που παρουσιάζει το $MS_E(p)$ καθώς αυξάνεται το p , παρουσιάζει κάποια ιδιαιτερότητα. Αρχικά, με την αύξηση του p το $MS_E(p)$ μειώνεται και μετά από ένα σημείο σταθεροποιείται, ενώ μπορεί να μπορεί να παρατηρηθεί και μια τάση μικρής βελτίωσης από εκεί και πέρα. Η συμπεριφορά αυτή παρατηρείται καθώς το $SS_E(p)$ διαρκώς μειώνεται με την αύξηση του p . Η βηματική βελτίωση του $MS_E(p)$ μπορεί να παρατηρηθεί όταν μια μεταβλητή που εισέρχεται στο μοντέλο δεν είναι ικανή να αντισταθμιστεί για την απώλεια ενός βαθμού ελευθερίας, όπως φαίνεται στον παρονομαστή της Σχέσης 2.4. Με άλλα λόγια, η εισαγωγή μιας μεταβλητής σε ένα μοντέλο παλινδρόμησης με p όρους μπορεί να προκαλέσει αύξηση του $MS_E(p)$ σε σχέση με το $MS_E(p+1)$ εάν η αύξηση στο άθροισμα των τετραγώνων των υπολοίπων είναι μικρότερο από το $MS_E(p)$. Τα μοντέλα παλινδρόμησης τα οποία παρουσιάζουν το μικρότερο $MS_E(p)$, θεωρούνται τα καλύτερα, ενώ μπορεί να αποδειχθεί ότι η ελαχιστοποίηση του κριτηρίου $MS_E(p)$ σε σχέση με τη μεγιστοποίηση του κριτηρίου $\bar{R}_p^2$ είναι ισοδύναμη.

2.6.4 Δείκτης Προσαρμογής

Ένα από τα πιο συχνά χρησιμοποιούμενα κριτήρια για την αξιολόγηση των αποτελεσμάτων μιας διαδικασίας παλινδρόμησης, είναι ο δείκτης προσαρμογής. Αυτός στηρίζεται στον υπολογισμό των υπολοίπων της εν λόγω διαδικασίας. Η Σχέση (2.5), με την οποία υπολογίζεται ο δείκτης προσαρμογής, φαίνεται στη συνέχεια:

$$s_f = \sqrt{\frac{\sum_{i=1}^n (Y_i - \hat{Y})^2}{n - p}} \quad (2.5)$$

όπου n είναι ο αριθμός των διαθέσιμων παρατηρήσεων, p είναι ο αριθμός των μεταβλητών που συμμετέχουν στην εξίσωση παλινδρόμησης, Y_i είναι η παρατηρούμενη τιμή της Y και $\hat{Y}_i$ η εκτίμηση για την Y (Mandel, 1989).

2.6.5 Κριτήριο του Mallow

Σχετικά με την επιλογή μεταβλητών σε ένα μοντέλο παλινδρόμησης, ο Mallow πρότεινε ένα κριτήριο, το οποίο σχετίζεται με το μέσο τετραγωνικό σφάλμα μιας εξαρτημένης μεταβλητής. Το κριτήριο αυτό εκφράζεται μαθηματικά από τη Σχέση (2.6):

$$E[\hat{y}_i - E(y_i)]^2 = [E(y_i) - E(\hat{y}_i)]^2 + V(\hat{y}_i) \quad (2.6)$$

όπου $E(y_i)$ είναι η προσδοκώμενη τιμή της απόκρισης από το πραγματικό μοντέλο παλινδρόμησης και $E(\hat{y}_i)$ είναι η προσδοκώμενη τιμή από το μοντέλο το οποίο περιέχει ένα υποσύνολο p όρων. Συνεπώς, η διαφορά $E(y_i) - E(\hat{y}_i)$ είναι ο θόρυβος που περιέχεται στο i -στό στοιχείο, άρα ο όρος $[E(y_i) - E(\hat{y}_i)]^2$ είναι το τετράγωνο του θορύβου για το μέσο τετραγωνικό σφάλμα, ενώ το $V(\hat{y}_i)$ είναι η διακύμανση για το τελευταίο.

Έστω ότι το συνολικό τετράγωνο του θορύβου για μια συνάρτηση παλινδρόμησης με p όρους δίνεται από την Σχέση (2.7):

$$SS_B(p) = \sum_{i=1}^n [E(y_i) - E(\hat{y}_i)]^2 \quad (2.7)$$

Τότε το κανονικοποιημένο συνολικό τετραγωνικό σφάλμα μπορεί να ορισθεί ως ακολούθως (Σχέση (2.8)):

$$\Gamma_p = \frac{1}{\sigma^2} \left\{ \sum_{i=1}^n [E(y_i) - E(\hat{y}_i)]^2 + \sum_{i=1}^n V(\hat{y}_i) \right\} = \frac{SS_B(p)}{\sigma^2} + \frac{1}{\sigma^2} \sum_{i=1}^n V(\hat{y}_i) \quad (2.8)$$

Μπορεί να αποδειχθεί ότι $\sum_{i=1}^n V(\hat{y}_i) = p\sigma^2$, ενώ η προσδοκώμενη τιμή για το άθροισμα των τετραγώνων των υπολοίπων είναι ίση με $E[SS_E(p)] = SS_B(p) + (n - p) \cdot \sigma^2$. Τότε, αν το $\hat{\sigma}^2$ είναι μια καλή εκτίμηση για το σ^2 , μια εκτίμηση για το Γ_p , έστω C_p , μπορεί να προκύψει από τη Σχέση (2.9):

$$C_p = \frac{SS_E(p)}{\hat{\sigma}^2} - n + 2p \quad (2.9)$$

Σχετικά με τον τρόπο συμπεριφοράς του C_p ανάλογα με το μέγεθος του p , για συναρτήσεις παλινδρόμησης με μικρό θόρυβο αναμένεται να έχουν τιμές του C_p οι οποίες θα είναι κοντά στις τιμές που περιγράφονται από την ευθεία του $C_p = p$. Γενικά, όμως, μικρές τιμές για το C_p θεωρούνται επιθυμητές, ενώ ίσως είναι θεμιτή η αποδοχή κάποιου επιπέδου θορύβου στη συνάρτηση παλινδρόμησης, έτσι ώστε να μειωθεί το μέσο σφάλμα της εκτίμησης. Για καλά αποτελέσματα στη λήψη τιμών για το C_p , μια καλή εκτίμηση του σ^2 είναι απαραίτητη.

2.7 Κριτήρια για την εισαγωγή – εξαγωγή μεταβλητών

2.7.1 Στατιστική F

Η τιμή αυτή υπολογίζεται ως μέτρο σύγκρισης (λόγος) δύο παραγόντων. Οι παράγοντες αυτοί, αποτελούν τις διακυμάνσεις δύο δειγμάτων τα οποία συγκρίνονται. Με τον τρόπο αυτό μπορεί να ελεγχθεί η μηδενική υπόθεση: α) ότι οι διακυμάνσεις των δειγμάτων που συγκρίνονται προέρχονται από δείγματα με κανονική κατανομή, με κοινή διακύμανση, β) ότι οι μέσοι όροι των δύο δειγμάτων είναι ίσοι και γ) ότι δεν υπάρχει καμία σχέση μεταξύ της εξαρτημένης μεταβλητής και κάποιων από τις ανεξάρτητες μεταβλητές.

Στην περίπτωση της παλινδρόμησης, ο έλεγχος με βάση την στατιστική F , πραγματοποιείται συγκρίνοντας το μέσο τετραγωνικό άθροισμα των σφαλμάτων λόγω της παλινδρόμησης (ποσοστό διακύμανσης που εξηγείται με βάση την παλινδρόμηση), MS_R , προς το τετραγωνικό άθροισμα των υπολοίπων, MS_E . Αξίζει εδώ να σημειωθεί ότι με την έκφραση «μέσο τετραγωνικό άθροισμα», νοείται το αντίστοιχο άθροισμα, διαιρεμένο με τους βαθμούς ελευθερίας που αντιστοιχούν σε αυτό (ανάλογα με τις ανεξάρτητες μεταβλητές που μετέχουν στο μοντέλο). Ο Πίνακας 2.1 δείχνει τα αθροίσματα αυτά σε συνάρτηση με τους βαθμούς ελευθερίας που τους αντιστοιχούν (Montgomery and Peck, 1992).

Πίνακας 2.1: Ανάλυση διακύμανσης για τον έλεγχο σημαντικότητας στην παλινδρόμηση.

Πηγή Διακύμανσης λόγω:	Τετραγωνικό Άθροισμα	Βαθμοί Ελευθερίας	Μέσο Τετραγωνικό Άθροισμα	Στατιστική Μηδενικής Υπόθεσης, F_0
Παλινδρόμησης	SS_R	k	MS_R	MS_R / MS_E
Υπολοίπων	SS_E	$n - k - 1$	MS_E	
Συνολικά	S_{yy}	$n - 1$		

n : Συνολικές διαθέσιμες μεταβλητές

k : Μεταβλητές που συμμετέχουν στο μοντέλο

Ο έλεγχος σημαντικότητας στην πολλαπλή παλινδρόμηση, έγκειται στο να διαπιστωθεί ότι μια ανεξάρτητη μεταβλητή είναι συσχετισμένη με την εξαρτημένη. Επομένως, η σημαντικότητα μιας ανεξάρτητης μεταβλητής, μπορεί να αποδειχθεί απορρίπτοντας την μηδενική υπόθεση ότι υπάρχει ανεξάρτητη μεταβλητή με μηδενικό συντελεστή για το μοντέλο. Η απόρριψη ή η αποδοχή αυτής της υπόθεσης,

πραγματοποιείται με την σύγκριση της F_0 με την αντίστοιχη τιμή για την στατιστική F , η οποία μπορεί να βρεθεί σε ανάλογους στατιστικούς πίνακες. Η μηδενική υπόθεση απορρίπτεται όταν η τιμή της F_0 ξεπερνά την αντίστοιχη τιμή της F , όπως εντοπίζεται στους στατιστικούς πίνακες.

2.7.2 Η τιμή p

Ένας άλλος τρόπος στατιστικού ελέγχου της σημαντικότητας των ανεξάρτητων μεταβλητών, είναι η χρησιμοποίηση της στατιστικής p (p -value). Συνήθως, ορίζεται ένα κατώφλι για την τιμή αυτή (πολύ συχνά λαμβάνει την τιμή 0,05). Στην περίπτωση αυτή, δεχόμαστε ότι υπάρχει πιθανότητα 5% ότι τα αποτελέσματα τα οποία παρατηρούμε, έχουν τυχαία κατανομή σε σχέση με αυτά που διαθέτουμε (π.χ. εισαγωγή μιας συγκεκριμένης μεταβλητής στο μοντέλο), άρα μπορούμε να ισχυριστούμε ότι υπάρχει πιθανότητα 95% τα παρατηρούμενα αποτελέσματα έχουν κάποια σχέση με τα διαθέσιμα. Συνεπώς, κατά 95% μπορούμε να δεχτούμε ότι η εξεταζόμενη μεταβλητή είναι σημαντική για το μοντέλο (και άρα για εισαγωγή σε αυτό). Ας σημειωθεί ότι όσο μικρότερες τιμές για την στατιστική p λαμβάνουμε, τόσο περισσότερες είναι οι ενδείξεις ότι δεν ισχύει η μηδενική υπόθεση (στην περίπτωση αυτή, ότι η εξεταζόμενη μεταβλητή δεν είναι σημαντική για το μοντέλο).

2.8 Κριτική στην Επιλογή Μοντέλου Παλινδρόμησης

Στην παράγραφο αυτή, εξετάστηκαν μια σειρά από κριτήρια επιλογής των μεταβλητών οι οποίες θα μορφοποιήσουν το τελικό μοντέλο παλινδρόμησης. Αυτό που γίνεται φανερό, είναι ότι η τελική επιλογή δεν είναι τόσο ξεκάθαρη. Γενικά, είναι δυνατόν να εντοπιστούν κάποια σημαντικά στοιχεία κατά τη διαδικασία επιλογής των μεταβλητών και της ανάπτυξης του μοντέλου, στα οποία οι αναλυτές θα πρέπει να δίνουν κάποια έμφαση. Αυτά τα σημεία, είναι τα ακόλουθα:

1. Είναι η τελική συνάρτηση λογική; Δηλαδή, έχουν κάποιο νόημα οι επιλεγμένες μεταβλητές στο μοντέλο σχετικά με το εξεταζόμενο πρόβλημα;
2. Είναι το αναπτυχθέν μοντέλο χρήσιμο για τον σκοπό για τον οποίο αναπτύχθηκε; Για παράδειγμα, για ένα μοντέλο παλινδρόμησης για πρόβλεψη, το οποίο όμως περιέχει μια μεταβλητή η οποία δεν μπορεί να παρατηρηθεί κατά τη στιγμή που απαιτείται η πρόβλεψη, δεν είναι χρήσιμο.
3. Είναι οι συντελεστές παλινδρόμησης στο μοντέλο λογικοί; Δηλαδή, θα πρέπει να εξετάζεται το μέγεθος και το πρόσημο των συντελεστών ως προς τη ρεαλιστικότητά τους, αλλά και ως προς το τυπικό σφάλμα τους, το οποίο θα πρέπει να είναι μικρό.
4. Είναι οι διαγνωστικοί έλεγχοι για την καταλληλότητα και την επάρκεια του μοντέλου ικανοποιητικοί; Για παράδειγμα, θα πρέπει να ελέγχεται η ύπαρξη ακραίων τιμών στο μοντέλο.

Από τις ερωτήσεις αυτές φαίνεται ότι αν αυτές ληφθούν υπόψη σοβαρά και οι απαντήσεις εφαρμοστούν αυστηρά, τότε δεν πρόκειται να υπάρξει ένα τελικό ικανοποιητικό μοντέλο. Η υποκειμενική κρίση και η εμπειρία των αναλυτών παίζει σημαντικό ρόλο για το συγκεκριμένο σκοπό που αναπτύσσεται το μοντέλο. Επίσης,

ακόμα κι αν ένα μοντέλο περάσει με επιτυχία τους διαγνωστικούς ελέγχους, δεν υπάρχει καμμία εγγύηση ότι αυτό θα δώσει καλά αποτελέσματα εάν εφαρμοσθεί σε ένα καινούριο πρόβλημα. Επιπλέον, στο (Grechanovsky, 1987) δίνεται μια περιεκτική επισκόπηση των βηματικών μεθοδολογιών παλινδρόμησης, ενώ αναδεικνύονται και τα πιο συχνά προβλήματα τους.

ΚΕΦΑΛΑΙΟ

3

ΒΙΒΛΙΟΓΡΑΦΙΚΗ ΕΠΙΣΚΟΠΗΣΗ

Για το πρόβλημα επιλογής μεταβλητών σε ένα μοντέλο παλινδρόμησης, έχουν προταθεί κατά το παρελθόν αρκετές μεθοδολογίες, οι οποίες αξιοποιούν επίσης διαφορετικά εναλλακτικά κριτήρια και χαρακτηριστικά των διαθέσιμων δεδομένων για την τελική επιλογή. Σύμφωνα με τις μεθοδολογίες αυτές, δεν είναι απαραίτητο να χρησιμοποιηθεί το σύνολο των διαθέσιμων μεταβλητών στο υπό ανάπτυξη μοντέλο. Παρόλα αυτά, με βάση την αντιμετώπιση αυτή, απορρέουν και ανάλογες κριτικές αναφορικά με το κατά πόσο αντικειμενική και αξιόπιστη είναι η επιλογή των μεταβλητών, ειδικά όταν δεν εξετάζονται όλες οι διαθέσιμες μεταβλητές. Αυτός ο αντίλογος εγείρεται από τους ερευνητές που ασκούν την ανάλογη κριτική, καθώς δεν είναι απόλυτα εξακριβωμένο εάν ένας διαφορετικός συνδυασμός μεταβλητών από αυτόν που τελικά επιλέχθηκε μπορεί ή όχι να δώσει καλύτερα αποτελέσματα. Στη συνέχεια, θα παρουσιασθούν μερικές ενδεικτικές μελέτες, κατά τις οποίες χρησιμοποιήθηκαν εναλλακτικά μοντέλα βηματικής παλινδρόμησης για την ανάπτυξη ενός μοντέλου παλινδρόμησης.

Στη συνέχεια παρουσιάζονται διάφορες ερευνητικές προσπάθειες που αφορούν το πρόβλημα επιλογής μεταβλητών, όπως εντοπίζονται στη διεθνή βιβλιογραφία. Στην παρουσίαση αυτή, μπορούν να εντοπισθούν προσπάθειες σχετικά με α) το πρόβλημα επιλογής μεταβλητών προτείνοντας εναλλακτικές μεθόδους β) χρησιμοποίηση ή/και σύγκριση βηματικών μεθοδολογιών για την επιλογή μεταβλητών και γ) εξέταση της επίδοσης εναλλακτικών κριτηρίων για την επιλογή μεταβλητών.

3.1 Εναλλακτικές Μεθοδολογίες Επιλογής Μεταβλητών

3.1.1 Επιλογή Μεταβλητών με Χρήση μαθηματικού Προγραμματισμού

Μια προσέγγιση του προβλήματος επιλογής μεταβλητών στη γραμμική πολλαπλή παλινδρόμηση με βάση το μαθηματικό προγραμματισμό, επιχειρείται στο (Eksioğlu et al, 2005). Στο άρθρο αυτό, οι συγγραφείς υιοθετούν ένα από τα μη παραδοσιακά κριτήρια για την επιλογή του καλύτερου υποσυνόλου μεταβλητών, βασιζόμενοι στη «διαίσθηση» ότι για ένα καλό μοντέλο παλινδρόμησης, οι συσχετίσεις των ανεξάρτητων μεταβλητών με την εξαρτημένη πρέπει να είναι υψηλές, ενώ μεταξύ τους θα πρέπει να είναι χαμηλές.

Το κριτήριο που υιοθετείται, αφορά τη μεγιστοποίηση του αθροίσματος των συσχετίσεων των ανεξάρτητων μεταβλητών με την ανεξάρτητη μεταβλητή Y ($\sum_{i \in M} |\rho_{0i}|$), ενώ ταυτόχρονα ελαχιστοποιείται το άθροισμα των συσχετίσεων των ανεξάρτητων μεταβλητών μεταξύ τους ($\sum_{i,j \in M, i < j} |\rho_{ij}|$), θεωρώντας ότι επιλέγεται ένα υποσύνολο M ανεξάρτητων μεταβλητών από τις συνολικά N διαθέσιμες. Ως κριτήριο επιλέγεται ένας γραμμικός συνδυασμός των δύο αυτών αθροισμάτων, επιλέγοντας ένα βάρος $0 < a < 1$, το οποίο δείχνει και τη σημαντικότητα κάθε όρου συσχέτισης. Στο σημείο αυτό ορίζονται και οι μεταβλητές απόφασης, οι οποίες είναι δυαδικές (binary), ως εξής:

$$u_i = 1 \text{ εάν } i \in M, 0 \text{ αλλιώς } \forall i \in N$$

$$v_{ij} = 1 \text{ εάν } i, j \in M (u_i = v_{ij} = 1), 0 \text{ αλλιώς } \forall i, j \in N, \text{ με } i < j$$

Τότε, το γ.π. το οποίο πρέπει να επιλυθεί, λαμβάνει την ακόλουθη μορφή:

$$\max z_a = a \sum_{i \in N} |\rho_{0i}| u_i - (1-a) \sum_{i,j \in N, i < j} |\rho_{ij}| v_{ij}$$

υ.π.

$$u_i \in \{0,1\} \quad \forall i \in N,$$

$$v_{ij} \in \{0,1\} \quad \forall i, j \in N, \text{ με } i < j$$

Για να απλοποιηθεί το παραπάνω γ.π., εισάγονται νέοι περιορισμοί, οι οποίοι απαλείφουν τους περιορισμούς δυαδικότητας για τις v_{ij} . Το νέο μοντέλο είναι τώρα:

$$\max z_a = a \sum_{i \in N} |\rho_{0i}| u_i - (1-a) \sum_{i,j \in N, i < j} |\rho_{ij}| v_{ij}$$

υ.π.

$$u_i \in \{0,1\} \quad \forall i \in N,$$

$$v_{ij} \geq 0 \quad \forall i, j \in N, \text{ με } i < j$$

$$v_{ij} \geq u_i + u_j - 1 \quad \forall i, j \in N, \text{ με } i < j$$

$$v_{ij} \leq u_i \quad \forall i, j \in N, \text{ με } i < j$$

$$v_{ij} \leq u_j \quad \forall i, j \in N, \text{ με } i < j$$

Το παραπάνω πρόβλημα μπορεί να επιλυθεί παραμετρικά. Η δυνατότητα επιλογής του βάρους a από τον αναλυτή, το καθιστά «ευέλικτο». Εάν επιλυθεί για όλες τις τιμές του a (μεταξύ 0 και 1), λαμβάνεται ένα σύνολο εφικτών λύσεων. Για $a=0$, λαμβάνεται το σύνολο των λύσεων για τα οποία τα επιμέρους υποσύνολα αποτελούνται από μια μόνο ανεξάρτητη μεταβλητή. Από αυτά, όποιο διαθέτει την πιο υψηλή συσχέτιση με την Y , επιλέγεται ως καλύτερο. Από την άλλη, όταν $a=1$, λαμβάνεται σαν λύση το υποσύνολο το οποίο περιλαμβάνει όλες τις διαθέσιμες ανεξάρτητες μεταβλητές. Ένα υποσύνολο το οποίο περιλαμβάνει μία μόνο μεταβλητή, ονομάζεται «αρχικό» (SS^0), ενώ αυτό που περιέχει όλες τις μεταβλητές ονομάζεται τελικό (SS^F). Με χρήση αυτών των υποσυνόλων, τα υπόλοιπα μπορούν να ευρεθούν

επαναληπτικά. Τώρα, εάν SS^q και SS^r είναι δύο υποσύνολα που βρέθηκαν με τον παραπάνω τρόπο και a_l είναι η τιμή του βάρους a που τα καθιστά ισοδύναμα, βάσει της z_a , τότε:

$$z_a(SS^q) = z_a(SS^r) \Rightarrow a_l \theta^q - (1 - a_l) \phi^q = a_l \theta^r - (1 - a_l) \phi^r$$

όπου

$$\theta^q = \sum_{i \in N} |\rho_{0i}| \mu_i$$

$$\phi^q = \sum_{i, j \in N, i < j} |\rho_{ij}| v_{ij}$$

$$\theta^r = \sum_{i \in N} |\rho_{0i}| \mu_i$$

$$\phi^r = \sum_{i, j \in N, i < j} |\rho_{ij}| v_{ij}$$

Για την τιμή αυτή, a_l , επιλύεται το τελευταίο γ.π., δίνοντας τιμή για την αντικειμενική π.χ. $z_a(SS^l)$. Εάν $z_a(SS^l) > z_a(SS^q)$, η νέα λύση προστίθεται στο σύνολο των καλύτερων εφικτών λύσεων και η διαδικασία συνεχίζεται μέχρι να μη είναι δυνατόν να βρεθεί άλλο τέτοιο υποσύνολο. Τελικά, προκύπτει ένα σύνολο S από τέτοιες εφικτές λύσεις, από τις οποίες ο αποφασίζων μπορεί να επιλέξει ποιο επιθυμεί.

Οι συγγραφείς επίσης, προτείνουν και δύο ευρετικούς αλγορίθμους, προσαρμοσμένους στο προηγούμενο γ.π., έτσι ώστε να μειωθεί ο χρόνος εύρεσης των λύσεων για μεγάλα σύνολα δεδομένων. Οι δοκιμές που έγιναν σε 12 σύνολα δεδομένων, δεν αποσαφηνίζουν ποια μέθοδος είναι η καλύτερη (οι προτεινόμενες σε σχέση με τις κλασσικές βηματικές), ενώ σημειώνεται ότι το κύριο πλεονέκτημα των προτεινόμενων μεθόδων είναι ότι προτείνουν ένα σύνολο από καλύτερες λύσεις για να επιλέξει ο αποφασίζων.

3.1.2 Επιλογή Μεταβλητών με Χρήση Ακέραιου Μαθηματικού Προγραμματισμού

Σε μια άλλη εργασία οι Konno και Yamamoto (2009), προτείνουν μια μεθοδολογία επιλογής μεταβλητών με χρήση ακέραιου μαθηματικού προγραμματισμού, θέτοντας ως κριτήριο βελτιστοποίησης το άθροισμα των απόλυτων σφαλμάτων μεταξύ των πραγματικών και των εκτιμώμενων τιμών για την εξαρτημένη μεταβλητή Y . Θεωρώντας το ακόλουθο γραμμικό μοντέλο παλινδρόμησης, $Y = a_0 + a_1 X_1 + \dots + a_k X_k + \varepsilon$, μορφοποιούν το παρακάτω γ.π.:

$$\min \sum_{t=1}^T \left| y_t - \left(a_0 + \sum_{j=1}^k a_j x_{jt} \right) \right|$$

υ.π.

$$\sum_{j=1}^k z_j = s$$

$$\underline{a}_j z_j \leq a_j \leq \bar{a}_j z_j, \quad j = 1, 2, \dots, k$$

$$z_j \in \{0, 1\}, \quad j = 1, 2, \dots, k$$

όπου $\underline{a}_j$ και $\bar{a}_j$ η ανώτερη και η κατώτερη αντίστοιχα τιμή που μπορεί να λάβει η a_j , ενώ s είναι το πλήθος των μεταβλητών που επιθυμούμε να μην έχουν μηδενικούς συντελεστές. Το ανωτέρω γ.π., μπορεί να λάβει την παρακάτω μορφή:

$$\min \sum_{t=1}^T (u_t + v_t)$$

υ.π.

$$u_t - v_t = y_t - \left(a_0 + \sum_{j=1}^k a_j x_{jt} \right), \quad t = 1, 2, \dots, T$$

$$u_t \geq 0, \quad v_t \geq 0, \quad t = 1, 2, \dots, T$$

$$\sum_{j=1}^k z_j = s$$

$$\underline{a}_j z_j \leq a_j \leq \bar{a}_j z_j, \quad j = 1, 2, \dots, k$$

$$z_j \in \{0, 1\}, \quad j = 1, 2, \dots, k$$

Εάν πρόκειται για όχι πολύ μεγάλες τιμές των k , s , και T , το γ.π. αυτό μπορεί να επιλυθεί σε προγράμματα όπως το CPLEX 10.1. Οι συγγραφείς εδώ, προτείνουν και έναν αλγόριθμο δύο βημάτων για την επίλυση του αρχικού γ.π., με αντικειμενική συνάρτηση το άθροισμα των τετραγώνων των υπολοίπων. Αρχικά, η πρόταση είναι να επιλυθεί το ακόλουθο γ.π., χρησιμοποιώντας μια θετική ακέραιη τιμή $r \leq k - s$:

$$\min \sum_{t=1}^T \left| y_t - \left(a_0 + \sum_{j=1}^k a_j x_{jt} \right) \right|$$

υ.π.

$$\sum_{j=1}^k z_j = s + r$$

$$\underline{a}_j z_j \leq a_j \leq \bar{a}_j z_j, \quad j = 1, 2, \dots, k$$

$$z_j \in \{0, 1\}, \quad j = 1, 2, \dots, k$$

Εδώ γίνεται η υπόθεση ότι $z_j^* = 0$ για κάθε $j > s + r$. Ακόμα, γίνεται η θεώρηση ότι $\hat{a}_j = 0$ για σχεδόν όλα τα $j > s + r$, όταν το r είναι αρκετά μεγάλο. Τότε

η επίλυση του παρακάτω γ.π. μπορεί να δώσει μια βέλτιστη λύση για την ελαχιστοποίηση του αθροίσματος των τετραγώνων των υπολοίπων.

$$\min \sum_{t=1}^T \left(y_t - \left(a_0 + \sum_{j=1}^k a_j x_{jt} \right) \right)^2$$

υ.π.

$$\sum_{j=1}^{s+r} z_j = s$$

$$\underline{a}_j z_j \leq a_j \leq \bar{a}_j z_j, \quad j = 1, 2, \dots, k$$

$$z_j \in \{0, 1\}, \quad j = 1, 2, \dots, (s + r)$$

Δοκιμάζοντας τα ευρήματα αυτά σε 5 σύνολα δεδομένων, τα οποία παρήχθησαν τυχαία, έδειξε ότι υπάρχουν συνδυασμοί για διάφορα k και s , οι οποίοι δεν μπορούσαν να επιλυθούν μέσα σε 1.500 CPU sec, στην περίπτωση που η εξαρτημένη μεταβλητή είναι συσχετισμένη με τις ανεξάρτητες μεταβλητές.

3.1.3 Βηματική Παραλλαγή Χρησιμοποίησης του AIC

Στο (Yamashita et al., 2007), παρουσιάζεται μια παραλλαγή χρησιμοποίησης του AIC (Akaike's Information Criterion) στα πλαίσια μιας προτεινόμενης βηματικής μεθοδολογίας. Η παραλλαγή αυτή τεκμηριώνεται σε μαθηματικό θεωρητικό επίπεδο και συγκρίνεται, με παρόμοιο τρόπο, με άλλα γνωστά κριτήρια που χρησιμοποιούνται σε βηματικές μεθοδολογίες (Partial F , Partial Correlation, Semi-Partial Correlation). Στην περίπτωση πολλαπλής γραμμικής παλινδρόμησης με μία εξαρτημένη μεταβλητή, n διαθέσιμες ανεξάρτητες μεταβλητές και k μεταβλητές να συνιστούν ένα υποσύνολο ανεξάρτητων μεταβλητών, το AIC είναι (για κανονική κατανομή των σφαλμάτων RSS – τετραγωνικό άθροισμα των υπολοίπων):

$$AIC = n \log \left(\frac{RSS}{n} \right) + 2k$$

Οι συγγραφείς αναλύοντας τα βήματα κάθε εξεταζόμενης μεθοδολογίας και προσαρμόζοντας κατάλληλα το παραπάνω κριτήριο σε κάθε μία από τις τρεις φάσεις για κάθε βήμα της διαδικασίας που προτείνουν (i) αρχικό βήμα για επιλογή μεταβλητής, ii) βήμα για επιλογή μεταβλητής και iii) βήμα για εξαγωγή μεταβλητής), καταλήγουν στο συμπέρασμα ότι η παραλλαγή που προτείνουν παρουσιάζει τα ίδια αποτελέσματα που θα λαμβάνονταν εάν χρησιμοποιείτο η μέθοδος της Partial F (χρησιμοποιείται στις κλασσικές βηματικές μεθόδους). Με βάση τη διαπίστωση αυτή, οι συγγραφείς ισχυρίζονται ότι η παραλλαγή που προτείνουν προσφέρει περισσότερα πλεονεκτήματα έναντι των υπολοίπων που εξετάζονται, καθώς είναι εύκολο να χρησιμοποιηθεί σε πιο γενικευμένα γραμμικά μοντέλα, μη γραμμικά μοντέλα και σε δεδομένα τα οποία δεν είναι απαραίτητο να ακολουθούν την κανονική κατανομή.

3.1.4 Προτεινόμενη Παραλλαγή της Προς τα Πίσω Βηματικής Απαλοιφής

Στο (Fuchs and Maria, 2007) παρουσιάζεται μια προσπάθεια παραλλαγής και επέκτασης της μεθόδου της προς τα πίσω απαλοιφής. Οι συγγραφείς ξεκινούν από το απλό βασικό μοντέλο παλινδρόμησης ($Y = X\beta + e$), για το οποίο, στην περίπτωση της μεθόδου των ελαχίστων τετραγώνων, προβαίνουν σε κάποιες στατιστικές παραδοχές:

- οι διαθέσιμες παρατηρήσεις περιέχουν θόρυβο, ο οποίος ακολουθεί την κατανομή Gauss.
- $\hat{\beta} \sim N\left(\beta^*, \sigma^2 (X^T X)^{-1}\right)$ με β^* τις πραγματικές τιμές των συντελεστών β
- $\hat{\sigma}^2 = \frac{RSS}{n-p}$ είναι μια εκτίμηση χωρίς θόρυβο για την σ^2
- $\frac{RSS}{\sigma^2} = \frac{(n-p)\hat{\sigma}^2}{\sigma^2} \sim \chi_{n-p}^2$
- $\hat{\beta}, \hat{\sigma}^2$ είναι ανεξάρτητα

Οι ερευνητές ορίζουν έναν πίνακα $Z = X + N$, ώστε ο N να περιέχει τα σφάλματα (θόρυβο) των παρατηρήσεων του X . Εάν οι X και Y είναι ανεξάρτητοι και $X \sim N(0,1)$, $Y \sim \chi_n^2$, τότε ο $Z = \frac{X}{\sqrt{Y/n}}$ θα ακολουθεί την κατανομή Student, με n

βαθμούς ελευθερίας. Μπορεί να αποδειχθεί, κατά τους ίδιους, ότι την κατανομή Student με $n-p$ βαθμούς ελευθερίας, ακολουθεί και η τυχαία μεταβλητή:

$$S_j = \frac{(\hat{\beta}_j - \beta_j^*) / \left(\sigma^2 (X^T X)^{-1}_{j,j} \right)^{1/2}}{(\hat{\sigma}^2 / \sigma^2)^{1/2}} = \frac{\hat{\beta}_j - \beta_j^*}{\left(\hat{\sigma}^2 \left((X^T X)^{-1}_{j,j} \right) \right)^{1/2}}$$

Η ιδέα εδώ είναι να κανονικοποιηθεί κατάλληλα κάθε εκτίμηση για τα β_j και αυτές να συγκριθούν με ένα κατώφλι, το οποίο θα ανταποκρίνεται σε ένα συγκεκριμένο διάστημα εμπιστοσύνης, υποθέτοντας ότι η πραγματική τιμή β_j^* είναι μηδέν. Έτσι, ξεκινώντας από το μοντέλο το οποίο περιέχει όλες τις μεταβλητές, σε κάθε βήμα ελέγχεται ο ακόλουθος κανόνας:

$$S_j = \frac{\hat{\beta}_j}{\left(\hat{\sigma}^2 \left((X^T X)^{-1}_{j,j} \right) \right)^{1/2}} \leq t_{n-p}(1-a/2)$$

Εάν ισχύει, τότε εξάγεται η μεταβλητή X_j από τον πίνακα X και η διαδικασία συνεχίζεται έως ότου δεν προκύψει μεταβλητή για την οποία να ισχύει ο παραπάνω κανόνας. Σε μια πιο γενικευμένη εκδοχή, οι συγγραφείς θεωρούν ότι μερικές μεταβλητές έχουν θόρυβο και μερικές άλλες όχι. Προσεγγίζοντας θεωρητικά το θέμα αυτό, προτείνουν ένα αλγόριθμο, ο οποίος και πάλι στηρίζεται σε έναν αντίστοιχο κανόνα εξαγωγής μεταβλητών από το μοντέλο.

3.1.5 Μεθοδολογίες Επιλογής Μεταβλητών με Χρήση του Αθροίσματος Απόλυτων Σφαλμάτων

Μια γενική επισκόπηση των μεθοδολογιών για την επιλογή του καλύτερου υποσυνόλου, μπορεί να βρεθεί στο (Andre et al., 2003). Στο άρθρο αυτό, οι συγγραφείς αντιπαραβάλλουν το κριτήριο του ελάχιστου αθροίσματος των απόλυτων σφαλμάτων (MSAE) με αυτό των ελαχίστων τετραγώνων και καταλήγουν ότι το πρώτο είναι μια πολύ καλή εναλλακτική έναντι του δεύτερου. Επίσης, απαριθμούν όλες τις μεθοδολογίες επιλογής μεταβλητών οι οποίες χρησιμοποιούν ή δυνητικά μπορούν να αξιοποιήσουν το κριτήριο MSAE. Αυτές διαχωρίζονται σε όσες εξετάζουν εξαντλητικά το σύνολο των δυνατών συνδυασμών παλινδρόμησης και σε όσες επιλέγουν μέρος μόνο των διαθέσιμων μεταβλητών. Για την τελευταία κατηγορία, εξετάζονται οι τρεις κλασσικές βηματικές, υιοθετώντας, ωστόσο, ως κριτήριο μια στατιστική που ενσωματώνει το άθροισμα των απόλυτων σφαλμάτων, υπό την παραδοχή ότι τα απόλυτα σφάλματα, ακολουθούν την κατανομή Laplace. Τότε, αποδεικνύεται, κατά τους ίδιους συγγραφείς, ότι η «συνεισφορά» μιας μεταβλητής x_i σε ένα μοντέλο παλινδρόμησης, μπορεί να υπολογισθεί από την έκφραση:

$$Q_{MSAE} = (SAE(i) - SAE) / (\hat{\tau} / 2)$$

όπου $\hat{\tau} = SAE/n$ όταν τα απόλυτα σφάλματα ακολουθούν την κατανομή Laplace, $SAE(i)$ είναι το άθροισμα απόλυτων σφαλμάτων του μοντέλου όταν αφαιρεθεί η μεταβλητή x_i από αυτό, ενώ SAE είναι το άθροισμα απόλυτων σφαλμάτων του μοντέλου πριν αφαιρεθεί η τελευταία. Έτσι, στα πλαίσια των βηματικών μεθόδων, η συνεισφορά των μεταβλητών, είτε πρόκειται για εισαγωγή είτε για εξαγωγή από το μοντέλο, εξετάζεται με βάση το Q_{MSAE} .

Στο αρχικό βήμα της προς τα εμπρός επιλογής, όποια μεταβλητή παρουσιάζει το μεγαλύτερο Q_{MSAE} , εξετάζεται σε ένα επίπεδο σημαντικότητας α σε σχέση με το μοντέλο $\hat{y}_i = \text{median}(y)$ και εάν είναι σημαντική, εισέρχεται στο μοντέλο και η διαδικασία συνεχίζεται με παρόμοιο τρόπο. Για την προς τα πίσω βηματική απαλοιφή, υπολογίζεται το Q_{MSAE} για κάθε μεταβλητή και αυτό εξετάζεται σε ένα επίπεδο σημαντικότητας α . Εάν προκύψει μια μεταβλητή μη σημαντική με αυτό τον τρόπο (εξετάζεται η μεταβλητή με το μικρότερο Q_{MSAE}), εξάγεται από το μοντέλο. Στη βηματική διαδικασία, ακολουθείται ένας συνδυασμός των παραπάνω.

3.2 Σύγκριση Εναλλακτικών Μεθόδων για Επιλογή του Καλύτερου Υποσυνόλου

Οι Xu και Zhang (2001), χρησιμοποίησαν τις τρεις πιο γνωστές μεθόδους βηματικής παλινδρόμησης (προς τα εμπρός επιλογή, προς τα πίσω απαλοιφή, βηματική επιλογή), ώστε να μελετηθούν και να συγκριθούν τα αποτελέσματα που δίδουν αυτές με άλλες μεθόδους επιλογής μεταβλητών. Οι υπόλοιπες μέθοδοι ήταν οι γενετικοί αλγόριθμοι, η leaps-and-bounds Regression και η orthogonal descriptors. Οι παραπάνω μέθοδοι χρησιμοποιήθηκαν για να μελετηθεί η τοξικότητα 35 νιτροβενζινών (nitrobenzenes), παρατηρώντας την επίδραση της συμμετοχής πέντε

επιμέρους συστατικών σε αυτές. Ωστόσο, επειδή η στερεομετρία των μορίων (η θέση στο χώρο - τοπολογία) μπορεί να έχει και διαφορετική επίδραση, χρησιμοποιήθηκαν και 22 επιπλέον μεταβλητές. Η εξέταση που πραγματοποιήθηκε, αφορά τη συμμετοχή 1 μέχρι 7 μεταβλητών στο τελικό μοντέλο, καθώς υιοθετήθηκε ο λόγος $N/m \geq 5$, όπου N είναι το σύνολο των ανεξάρτητων μεταβλητών και m οι επιλεγόμενες ανεξάρτητες μεταβλητές. Τα συμπεράσματα που απορρέουν από τη σύγκριση των αποτελεσμάτων, είναι ότι η μέθοδος των orthogonal descriptors είναι καλύτερη από τις υπόλοιπες, καθώς δίνει υψηλότερη τιμή για τον δείκτη συσχέτισης R^2 σε κάθε περίπτωση. Μια πιθανή εξήγηση που δίνουν οι συγγραφείς για το γεγονός αυτό, είναι ότι οι πληροφορίες οι οποίες περιέχονταν σε μερικές μεταβλητές, ήταν «συμπαγείς». Άλλη μια παρατήρηση είναι ότι σε ό,τι αφορά τις μεθόδους βηματικής παλινδρόμησης, οι μεταβλητές που φαίνονταν ως οι καλύτερες σε προηγούμενα στάδια, είναι δυνατό να μη συμμετέχουν στο τελικό μοντέλο, κάτι το οποίο δε φαίνεται να ισχύει στην περίπτωση της μεθόδου orthogonal descriptors. Έτσι, παρόλο που οι τρεις μέθοδοι της οικογένειας της βηματικής παλινδρόμησης είναι δυνατόν να δώσουν παραπλήσια αποτελέσματα, δεν υπάρχει τρόπος απόλυτης διαβεβαίωσης για την βελτιστότητα της τελικής επιλογής.

Σε άλλες ερευνητικές προσπάθειες, εντοπίζεται η απόπειρα εισαγωγής ενός νέου, πιο γενικευμένου κριτηρίου για την επιλογή μεταβλητών, το οποίο αντιπαραβάλλεται με το γνωστό κριτήριο C_p του Mallows (Kashid and Kulkarni, 2002), ενώ στο (Salt et al., 2007) παρουσιάζεται μια προσπάθεια βελτιωμένης εκτίμησης της στατιστικής F σχετικά με την επιλογή μεταβλητών σε ένα μοντέλο παλινδρόμησης.

ΚΕΦΑΛΑΙΟ

4

ΠΡΟΤΕΙΝΟΜΕΝΗ ΜΕΘΟΔΟΛΟΓΙΑ

Το κεφάλαιο αυτό περιγράφει τον τρόπο με τον οποίο αναπτύχθηκε το προταθέν μαθηματικό μοντέλο και παρουσιάζει τη μαθηματική δομή του. Εξετάζονται ακόμα τα επιμέρους κριτήρια που λήφθηκαν υπόψη κατά τη διάρκεια ανάπτυξης του μοντέλου, ενώ δίνεται και μια αναλυτική περιγραφή της διαδικασίας που υιοθετήθηκε για την εισαγωγή και εξαγωγή των υπό εξέταση διαθέσιμων μεταβλητών σε ένα μοντέλο, με βάση αυτά τα κριτήρια. Αρχικά, παρουσιάζονται τα βασικά στοιχεία που πλαισιώνουν την κεντρική ιδέα που διέπει την ποιοτική ανάλυση παλινδρόμησης.

4.1 Ποιοτική Ανάλυση Παλινδρόμησης

Στόχος της ποιοτικής ανάλυσης παλινδρόμησης, είναι να αναπτυχθεί ένα μη γραμμικό μοντέλο πρόβλεψης, σύμφωνα με το οποίο θα πραγματοποιηθεί η εκτίμηση των συναρτήσεων αξιών, τόσο της εξαρτημένης μεταβλητής, όσο και των ανεξάρτητων μεταβλητών που είναι διαθέσιμες. Η κεντρική ιδέα μιας τέτοιας μοντελοποίησης, με βάση συναρτήσεις αξιών, είναι ότι σε μια τέτοια προσέγγιση, η «αξία» της εξαρτημένης, για παράδειγμα, μεταβλητής, εξαρτάται από την αξία της μεταβολής των μη εξαρτημένων μεταβλητών που φαίνεται να την επηρεάζουν (μετέχουν στο μοντέλο). Με άλλα λόγια, ένα τέτοιο γεγονός ερμηνεύεται από την παρατήρηση ότι η συνεισφορά μιας μη εξαρτημένης μεταβλητής στην αξία της εξαρτημένης μεταβλητής, δεν είναι σταθερή και αναλογική, αλλά εξαρτάται από την ονομαστική τιμή της μεταβολής της πρώτης. Αξίζει να σημειωθεί, ωστόσο, ότι οι επιμέρους συναρτήσεις αξιών, είναι γραμμικές μόνο κατά τμήματα.

Για τον υπολογισμό των επιμέρους συναρτήσεων αξιών, λαμβάνεται υπόψη και η σημαντικότητα της κάθε εξεταζόμενης μεταβλητής στο μοντέλο. Η σημαντικότητα αυτή για κάθε μεταβλητή, υπολογίζεται με την εφαρμογή του μοντέλου, μέσω της απόδοσης βαρών (σχετικούς δείκτες σημαντικότητας) σε κάθε μεταβλητή. Τα βάρη αυτά, δεν είναι τίποτε άλλο παρά η μέγιστη συνεισφορά καθεμιάς μεταβλητής στην αξία της εξαρτημένης μεταβλητής και δίνονται σε κανονικοποιημένη μορφή, κάτι που επιτρέπει την άμεση σύγκριση μεταξύ της επιμέρους σημαντικότητας κάθε μεταβλητής.

Η αποτύπωση της εξίσωσης παλινδρόμησης σε ένα τέτοιο μοντέλο, με βάση τα παραπάνω, έχει ως εξής (Γρηγορούδης και Σίσκος, 2005):

$$Y^* = \sum_{i=1}^n b_i \cdot X_i^* \text{ δεδομένου ότι } \sum_{i=1}^n b_i = 1 \quad (4.1)$$

όπου οι συναρτήσεις αξιών Y^* και X_i^* είναι κανονικοποιημένες στο διάστημα $[0,1]$, ενώ με b_i συμβολίζεται ο συντελεστής βάρους της μεταβλητής i . Το γεγονός ότι οι παραπάνω συναρτήσεις αξιών είναι κανονικοποιημένες στο διάστημα $[0,1]$, σημαίνει ότι η μικρότερη (χειρότερη) και η μεγαλύτερη (καλύτερη) τιμή που μπορούν να λάβουν, είναι αντίστοιχα το 0 και το 1. Με τον τρόπο αυτό, οι περιορισμοί κανονικοποίησης μπορούν να γραφούν ως (Γρηγορούδης και Σίσκος, 2005):

$$\begin{cases} y^{*1} = 0, y^{*a} = 1 \\ x^{*1} = 0, x^{*a_i} = 1 \end{cases} \text{ για } i = 1, 2, \dots, n \quad (4.2)$$

Ο Πίνακας 4.1 θα βοηθήσει στην επεξήγηση των συμβολισμών που χρησιμοποιούνται στο κεφάλαιο αυτό.

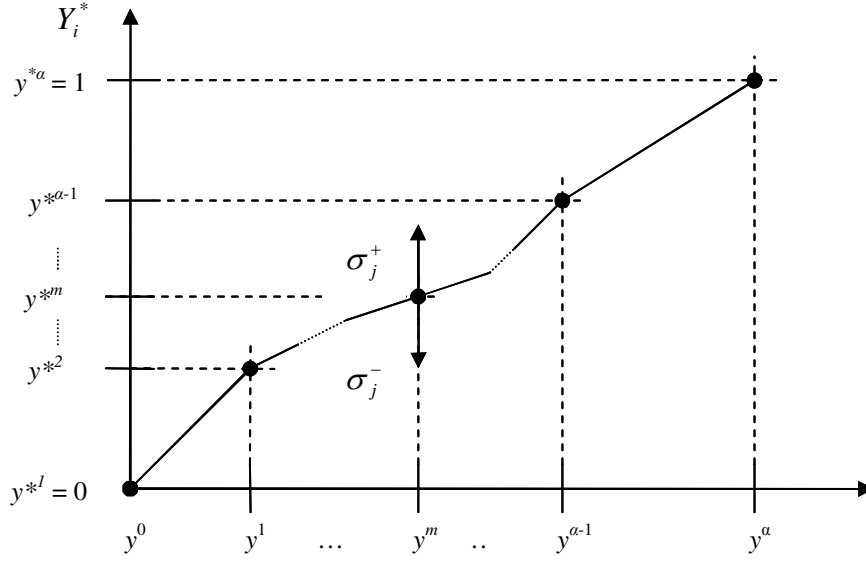
Πίνακας 4.1: Συμβολισμοί των μεταβλητών του μοντέλου και η σημασία τους.

Συμβολισμός	Σημασία
Y	μεταβολή της εξαρτημένης μεταβλητής
a	αριθμός επιπέδων διαχωρισμού της κλίμακας του Y
y^m	το m επίπεδο του Y ($m = 1, 2, \dots, a$)
n	αριθμός μεταβλητών
X_i	μεταβολή της i -στής μεταβλητής ($i = 1, 2, \dots, n$)
a_i	αριθμός επιπέδων διαχωρισμού της κλίμακας της i -στής μεταβλητής
x_i^k	το k επίπεδο της i -στής μεταβλητής ($k = 1, 2, \dots, a_i$)
Y^*	συνάρτηση αξιών του Y
y^{*m}	αξία του y^m επιπέδου μεταβολής του Y
X_i^*	συνάρτηση αξιών του X
x_i^{*k}	αξία του x_i^k επιπέδου μεταβολής του X_i
j	πλήθος διαθέσιμων παρατηρήσεων για την Y

Για την γραφική απεικόνιση των συναρτήσεων αυτών, χρησιμοποιείται ένα σύστημα αξόνων, στον κατακόρυφο άξονα του οποίου απεικονίζεται η αξία της εξεταζόμενης μεταβλητής, ενώ στον οριζόντιο άξονα απεικονίζεται η πραγματική τιμή της παραμέτρου βάση της οποίας εξετάζεται η απεικονιζόμενη μεταβλητή (π.χ. χρόνος, ποσότητα πωλήσεων, ποσότητα συστατικών κτλ.). Η κλίμακα που υιοθετείται στον οριζόντιο άξονα χωρίζεται σε επίπεδα (διαστήματα) διαχωρισμού, τα οποία αντιστοιχούν σε ανάλογα επίπεδα για τον κατακόρυφο άξονα.

Βασική επιδίωξη για ένα μοντέλο ποιοτικής παλινδρόμησης, είναι η εκτίμηση των συναρτήσεων αξιών Y^* και X_i^* με τέτοιο τρόπο, ώστε να υπάρξει η μεγαλύτερη δυνατή συμφωνία ανάμεσα στην πραγματική και την εκτιμώμενη τιμή για την εξαρτημένη μεταβλητή Y . Για το σκοπό αυτό, εισάγεται μια διπλή μεταβλητή σφάλματος στο μοντέλο και η εξίσωση παλινδρόμησης, όπως φαίνεται στη Σχέση (4.1), λαμβάνει την μορφή της Σχέσης (4.3) (Γρηγορούδης και Σίσκος, 2005). Το

Σχήμα 4.1 δείχνει ένα παράδειγμα μιας τέτοιας γραφικής απεικόνισης, όπου φαίνονται και τα σφάλματα απόκλισης.



Σχήμα 4.1: Γραφική απεικόνιση κλιμάκων και των σφαλμάτων απόκλισης.

$$\hat{Y}^* = \sum_{i=1}^n b_i \cdot X_i^* - \sigma^+ + \sigma^- \quad (4.3)$$

όπου $\hat{Y}^*$ είναι η εκτίμηση της συνάρτησης αξιών Y^* , σ^- και σ^+ είναι τα σφάλματα υποεκτίμησης και υπερεκτίμησης αντίστοιχα.

Όπως φαίνεται στο τυχαίο παράδειγμα απεικόνισης στο Σχήμα 4.1, οι συναρτήσεις αξιών είναι κατά τμήματα γραμμικές. Επειδή, όμως, υπάρχει περίπτωση μερικά από τα τμήματα αυτά να είναι παράλληλα προς τον οριζόντιο άξονα, τίθενται στο μοντέλο και οι ανάλογοι περιορισμοί μονοτονίας των συναρτήσεων αξιών, ώστε να εξασφαλιστεί ότι αυτές θα είναι γνήσια μονότονες. Αυτό πραγματοποιείται με την διαδικασία που περιγράφεται στη συνέχεια.

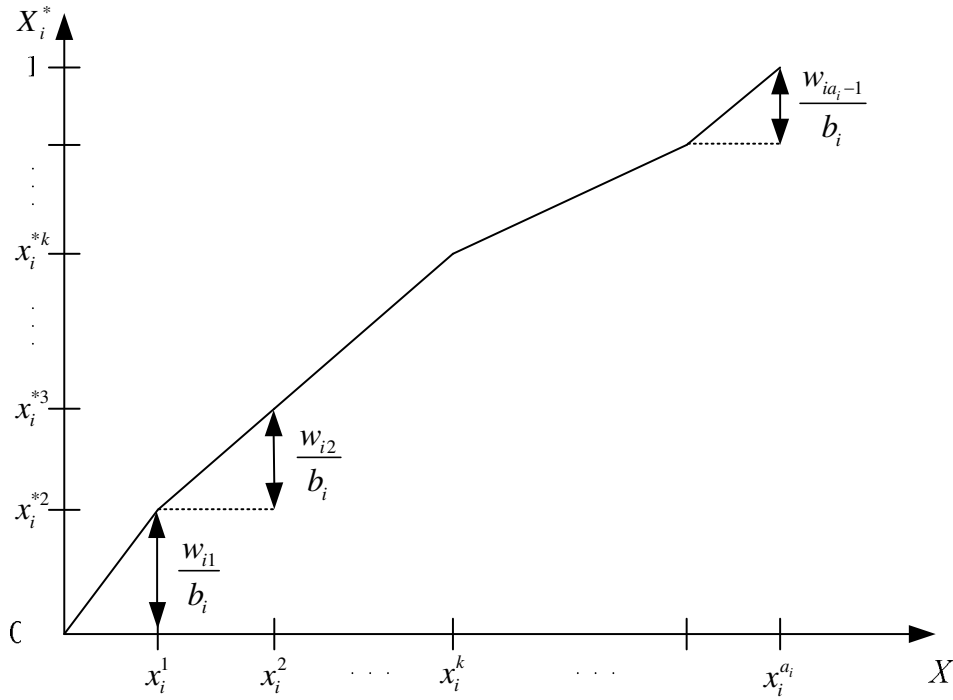
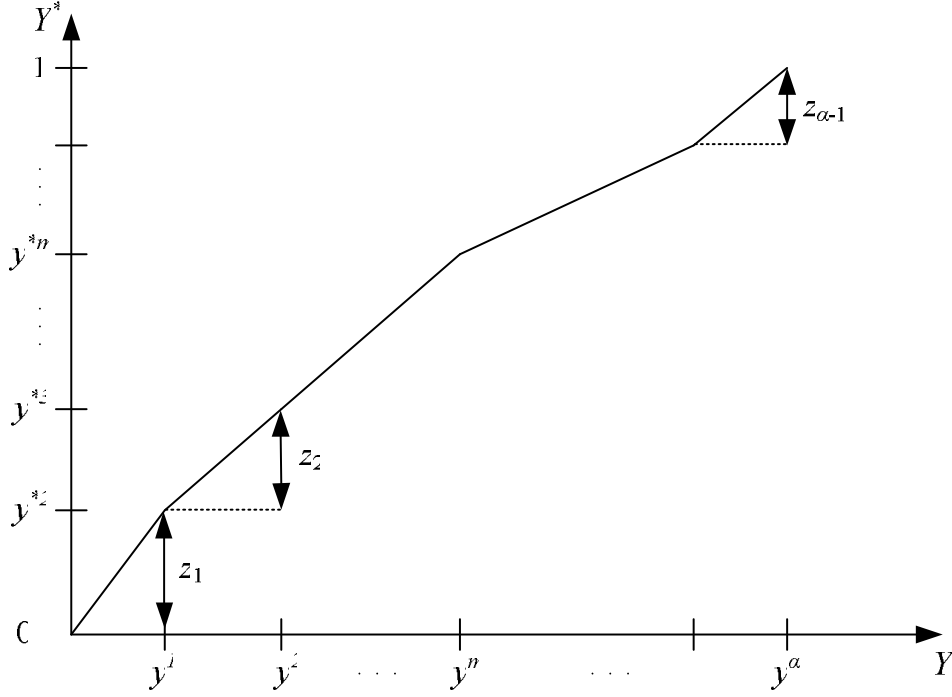
Αρχικά, ορίζονται οι ακόλουθες μεταβλητές (Γρηγορούδης και Σίσκος, 2005):

$$\begin{cases} z_m = y^{*m+1} - y^{*m} & \text{για } m = 1, 2, \dots, a-1 \\ w_{ik} = b_i x^{*k+1} - b_i x^{*k} & \text{για } k = 1, 2, \dots, a_i-1 \text{ και } i = 1, 2, \dots, n \end{cases} \quad (4.4)$$

Οι μεταβλητές z_m και w_{ik} εκφράζουν τα διαδοχικά βήματα αύξησης των συναρτήσεων αξιών Y^* και X_i^* αντίστοιχα. Εποπτική απεικόνιση των μεταβλητών αυτών, φαίνεται στο Σχήμα 4.2.

Στη συνέχεια εισάγονται τα κατώφλια προτίμησης για Y^* και X_i^* , έτσι ώστε να εξασφαλίζεται το ελάχιστο βήμα αύξησης των συναρτήσεων αυτών να μην είναι μηδενικό. Τα κατώφλια αυτά συμβολίζονται με γ και γ_i αντίστοιχα, ενώ οι σχέσεις

(4.4) τελικά λαμβάνουν τη μορφή των Σχέσεων (4.5) (Γρηγορούδης και Σίσκος, 2005).



Σχήμα 4.2: Γραφική απεικόνιση των μεταβλητών μετασχηματισμού z_m και w_{ik} .

$$\begin{cases} z_m = y^{*m+1} - y^{*m} \geq \gamma & \text{για } m = 1, 2, \dots, a-1 \\ w_{ik} = b_i x_i^{*k+1} - b_i x_i^{*k} \geq \gamma_i & \text{για } k = 1, 2, \dots, a_i-1 \text{ και } i = 1, 2, \dots, n \\ \gamma, \gamma_i > 0 \end{cases} \quad (4.5)$$

Συνεχίζοντας το μετασχηματισμό των παραπάνω σχέσεων, έχουμε:

$$\begin{cases} z_m \geq \gamma \\ w_{ik} \geq \gamma_i \end{cases} \Leftrightarrow \begin{cases} z_m - \gamma \geq 0 \\ w_{ik} - \gamma_i \geq 0 \end{cases} \Leftrightarrow \begin{cases} z'_m \geq 0 \\ w'_{ik} \geq 0 \end{cases} \quad (4.6)$$

Είναι φανερό ότι οι μεταβλητές Z'_m και W'_{ik} έχουν οριστεί όπως φαίνεται στη συνέχεια (Γρηγορούδης και Σίσκος, 2005):

$$\begin{cases} z_m = z'_m + \gamma & \text{για } m = 1, 2, \dots, a-1 \\ w_{ik} = w'_{ik} + \gamma_i & \text{για } k = 1, 2, \dots, a_i - 1 \text{ και } i = 1, 2, \dots, n \end{cases} \quad (4.7)$$

Χρησιμοποιώντας όλες τις παραπάνω σχέσεις, προκύπτει η ακόλουθη μορφή για το υπό συζήτηση μη γραμμικό μοντέλο πρόβλεψης. Ζητούμενο είναι, όπως αναφέρθηκε, η όσο το δυνατόν σύγκλιση μεταξύ της πραγματικής και εκτιμώμενης τιμής της εξαρτημένης μεταβλητής, οπότε το πρόβλημα ανάγεται σε πρόβλημα ελαχιστοποίησης του αθροίσματος των μεταβλητών σφάλματος που εισήχθησαν στη Σχέση (4.3) (Γρηγορούδης και Σίσκος, 2005):

$$\begin{aligned} \min F &= \sum_j (\sigma_j^- + \sigma_j^+) \\ \text{υ.}\pi. & \\ \sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - \sum_{m=1}^{a-1} c_j^m z'_m + \sigma_j^- - \sigma_j^+ &= \gamma \sum_{m=1}^{a-1} c_j^m - \sum_{i=1}^n \gamma_i \sum_{k=1}^{a_i-1} c_{ij}^k \\ \sum_{m=1}^{a-1} z'_m &= 1 - \gamma(a-1) \\ \sum_{i=1}^n \sum_{k=1}^{a_i-1} w'_{ik} &= 1 - \sum_{i=1}^n \gamma_i(a_i - 1) \\ z'_m, w'_m, \gamma, \gamma_i, \sigma_j^-, \sigma_j^+ &\geq 0 \quad \text{για κάθε } m, i, j, k \end{aligned} \quad (4.8)$$

όπου οι συντελεστές c_j^m και c_{ij}^k υπολογίζονται με χρήση γραμμικής παρεμβολής, καθώς αυτό επιτρέπεται λόγω της γραμμικότητας των επιμέρους αντίστοιχων τμημάτων των συναρτήσεων αξιών Y^* και X_i^* αντίστοιχα. Αναλυτικά, η χρήση της γραμμικής παρεμβολής για τον υπολογισμό των c_j^m και c_{ij}^k πραγματοποιείται μέσω των Σχέσεων (4.9) και (4.10) (Γρηγορούδης και Σίσκος, 2005):

$$c_j^m = \begin{cases} 0 & \alpha \nu Y_j \leq y^m \\ \frac{Y_j - y^m}{y^{m+1} - y^m} & \alpha \nu y^m \leq Y_j \leq y^{m+1}, \text{ για } m = 1, 2, \dots, a-1 \text{ και } \forall j \\ 1 & \alpha \nu Y_j \geq y^{m+1} \end{cases} \quad (4.9)$$

$$c_{ij}^k = \begin{cases} 0 & \alpha\nu X_{ij} \leq x_i^k \\ \frac{X_{ij} - x_i^k}{x_i^{k+1} - x_i^k} & \alpha\nu x_i^k \leq X_{ij} \leq x_i^{k+1}, \text{ για } i = 1, 2, \dots, n, k = 1, 2, \dots, a_i - 1 \text{ και } \forall j \\ 1 & \alpha\nu X_{ij} \geq x_i^{k+1} \end{cases} \quad (4.10)$$

Οι αρχικές μεταβλητές του προβλήματος υπολογίζονται με βάση τη βέλτιστη λύση του γ.π. της Σχέσης (4.8), καθώς αποδεικνύεται ότι ισχύουν οι παρακάτω σχέσεις (Γρηγορούδης και Σίσκος, 2005):

$$\begin{aligned} y^{*m} &= \sum_{t=1}^{a-1} z_t & \text{για } m = 2, 3, \dots, a \\ b_i &= \sum_{t=1}^{a_i-1} w_{it} & \text{για } i = 1, 2, \dots, n \\ x_i^{*k} &= \frac{\sum_{t=1}^{k-1} w_{it}}{\sum_{t=1}^{a_i-1} w_{it}} & \text{για } i = 1, 2, \dots, n \text{ και } k = 2, 3, \dots, a_i \end{aligned} \quad (4.11)$$

Αξίζει στο σημείο αυτό, να επισημανθεί ότι οι συντελεστές c_{ij}^k , όπως υπολογίζονται από τη Σχέση (4.10), ανταποκρίνονται στην περίπτωση κατά την οποία η εκάστοτε εξεταζόμενη μεταβλητή απόφασης (ανεξάρτητη μεταβλητή), έχει «θετική συνεισφορά» στην διαδικασία μορφοποίησης ενός μοντέλου συνάρτησης αξιών για την ερμηνεία και προβλεψη της εξαρτημένης μεταβλητής. Αυτό σημαίνει ότι μια τέτοια ανεξάρτητη μεταβλητή, εμφανίζει θετική συσχέτιση με την εξαρτημένη μεταβλητή. Εάν παρατηρηθεί ότι μια ανεξάρτητη μεταβλητή δεν παρουσιάζει τέτοια χαρακτηριστικά, η «συνεισφορά» της αντιμετωπίζεται ως «αρνητική», δηλαδή η συνάρτηση αξιών της λαμβάνει τέτοια φορά, ώστε ανάλογα με τις διαθέσιμες παρατηρήσεις να λαμβάνεται υπόψη και η αντίστοιχη αξία («συνεισφορά») κατά τον συνολικό υπολογισμό της αξίας μιας παρατήρησης με βάση τις αξίες των αντίστοιχων διαθέσιμων παρατηρήσεων για τις εξεταζόμενες ανεξάρτητες μεταβλητές. Για να γίνει πιο κατανοητή η παραπάνω περιγραφή, θα αναφερθούν και τα παρακάτω. Με τον τρόπο που έχουν ορισθεί οι συντελεστές c_j^m της Σχέσης (4.9), η συνάρτηση αξιών της εξαρτημένης μεταβλητής Y , λαμβάνει αύξουσα μορφή (μεγαλύτερες τιμές των παρατηρήσεων για την Y αντιστοιχούν σε μεγαλύτερες τιμές στον άξονα των αξιών). Το ίδιο ισχύει και για τους συντελεστές c_{ij}^k της Σχέσης (4.10). Εάν μια ανεξάρτητη μεταβλητή εμφανίζει αρνητική συσχέτιση με την Y , αυτό σημαίνει ότι μεγαλύτερες τιμές αυτής, θα πρέπει να αντιστοιχούν σε μικρότερες τιμές στον άξονα αξιών της συνάρτησης αξιών της (έτσι ορίζεται η «αρνητική συνεισφορά»). Το γεγονός αυτό λήφθηκε υπόψη σε όλες τις περιπτώσεις εξέτασης δεδομένων που χρησιμοποιούνται στο επόμενο κεφάλαιο για την αξιολόγηση των προτεινόμενων μεθοδολογιών που περιγράφονται στη συνέχεια.

4.2 Εναλλακτικά Κριτήρια Βελτιστοποίησης

Στην προηγούμενη παράγραφο, ορίστηκε το γ.π. (4.8), θέτοντας ως αντικειμενική συνάρτηση το άθροισμα των απολύτων σφαλμάτων τα οποία παρατηρούνται στην εκτίμηση για την αξία της εξαρτημένης μεταβλητής Y . Στα πλαίσια αυτής της εργασίας, υιοθετούνται και άλλα επιμέρους κριτήρια βελτιστοποίησης, με στόχο να εντοπιστεί ποιο από αυτά δίνει τα καλύτερα αποτελέσματα όσον αφορά τη συμφωνία ανάμεσα στις πραγματικές και εκτιμώμενες τιμές για την μεταβλητή Y .

4.2.1 Τετραγωνικό Άθροισμα Απόλυτων Σφαλμάτων

Ως πρώτο εναλλακτικό κριτήριο βελτιστοποίησης, έναντι του απλού αθροίσματος των απόλυτων σφαλμάτων, επιλέχθηκε το άθροισμα των τετραγώνων των τελευταίων. Στην περίπτωση αυτή, δεν παρουσιάζονται κάποιες αλλαγές στη δομή του γ.π. που ορίστηκε, απλώς αλλάζει η αντικειμενική συνάρτηση και έχουμε το γ.π. της Σχέσης (4.12).

$$\begin{aligned} \min F &= \sum_j (\sigma_j^-)^2 + (\sigma_j^+)^2 \\ \text{υ.π.} \\ \sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - \sum_{m=1}^{a-1} c_j^m z_m' + \sigma_j^- - \sigma_j^+ &= \gamma \sum_{m=1}^{a-1} c_j^m - \sum_{i=1}^n \gamma_i \sum_{k=1}^{a_i-1} c_{ij}^k \\ \sum_{m=1}^{a-1} z_m' &= 1 - \gamma(a-1) \\ \sum_{i=1}^n \sum_{k=1}^{a_i-1} w_{ik}' &= 1 - \sum_{i=1}^n \gamma_i(a_i-1) \\ z_m', w_m', \gamma, \gamma_i, \sigma_j^-, \sigma_j^+ &\geq 0 \quad \text{για κάθε } m, i, j, k \end{aligned} \tag{4.12}$$

4.2.2 Συντελεστής προσδιορισμού R^2

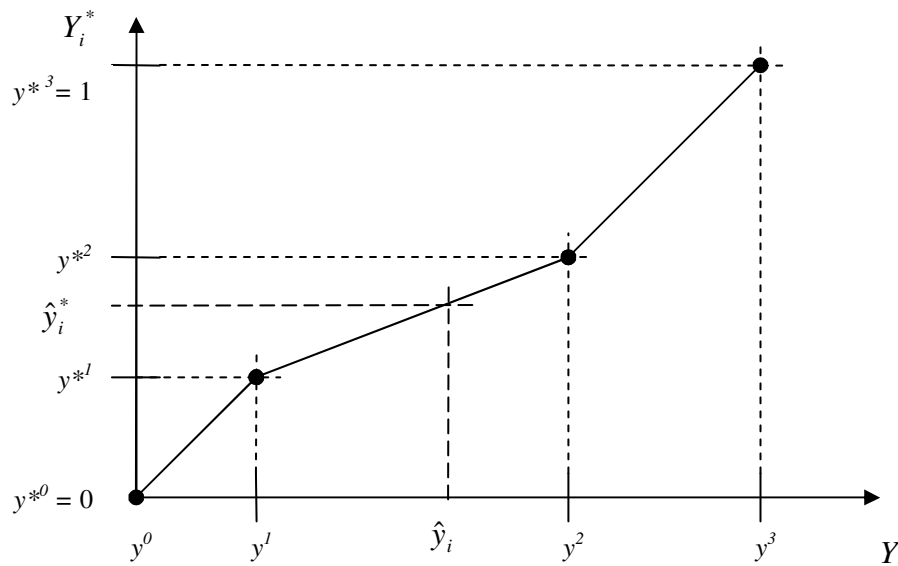
Ο δείκτης αυτός παρουσιάστηκε στο Κεφάλαιο 2 της παρούσης. Αφού ο ίδιος αυτός δείκτης αποτελεί έναν αξιόπιστο, κατά πολλούς ερευνητές, δείκτη αξιολόγησης των αποτελεσμάτων ενός μοντέλου παλινδρόμησης, επιλέχθηκε να είναι ένα ακόμη εναλλακτικό κριτήριο βελτιστοποίησης, με το σκεπτικό ότι όσο καλύτερες τιμές λαμβάνει αυτός ο δείκτης (όσο το δυνατόν πλησιέστερα ή ίσος στη μονάδα), τόσο καλύτερα θα πρέπει να είναι και τα αποτελέσματα που δίνει το μοντέλο στο οποίο αντιστοιχεί.

Για να εισαχθεί ως κριτήριο βελτιστοποίησης (αντικειμενική συνάρτηση) ο συντελεστής προσδιορισμού, θα πρέπει να ληφθούν υπόψη ορισμένες τροποποιήσεις οι οποίες επιβάλλεται να πραγματοποιηθούν στο βασικό μοντέλο (4.8). Ήδη από τον

ορισμό του δείκτη αυτού ($R^2 = \frac{\sum_i (\hat{Y}_i - \bar{Y})^2}{\sum_i (Y_i - \bar{Y})^2} = 1 - \frac{\sum_i (\hat{Y}_i - Y_i)^2}{\sum_i (Y_i - \bar{Y})^2}$, βλ. και Κεφάλαιο 2),

φαίνεται ότι, εκτός των σταθερών (και γνωστών) τιμών, στην αντικειμενική συνάρτηση εισάγεται απευθείας και η ζητούμενη εκτίμηση για την μεταβλητή Y , $\hat{Y}_i$. Αυτό σημαίνει ότι θα πρέπει να προβλεφθούν όλες οι δυνατές τιμές τις οποίες μπορεί να λάβει αυτή η εκτίμηση, με βάση τον τρόπο με τον οποίο ορίστηκε οι συνάρτηση αξιών Y^* , και να ενσωματωθούν στη δομή του νέου γ.π., εν ίδι επιπρόσθετων περιορισμών.

Έχει ήδη παρουσιασθεί ο τρόπος με τον οποίον ορίστηκαν οι συναρτήσεις αξιών Y^* και X_i^* . Είδαμε ότι αυτές είναι κατά τμήματα γραμμικές. Σε ένα από τα τμήματα αυτά, τα οποία οριοθετούνται από τα επιλεγόμενα διαστήματα διαχωρισμού, είναι δυνατόν να ανήκει και η υπολογιζόμενη εκτίμηση για την Y . Ωστόσο, επειδή δεν είναι δυνατόν να γνωρίζουμε εκ των προτέρων σε ποιο από τα διαστήματα αυτά θα ανήκει η εκτίμηση αυτή, η προσέγγιση που ακολουθείται είναι ότι κάθε ένα από τα διαστήματα αυτά μπορεί να περιέχει την εκτίμηση. Έτσι, θα πρέπει να προβλεφθούν τόσες περιπτώσεις πιθανών τιμών για την εκτίμηση, όσες ακριβώς προκύπτουν από τον ορισμό των διαστημάτων διαχωρισμού (βλ. Σχήμα 4.3). Ο επιλεγθείς αριθμός των διαστημάτων διαχωρισμού καθορίζει και τον αριθμό πιθανών τιμών που μπορεί να λάβει η εκτίμηση για την Y . Προκειμένου για κλίμακα βαθμού m , η τιμή της εκτίμησης για την Y , μπορεί να ανήκει σε κάθε ένα από τα $m-1$ διαστήματα διαχωρισμού, καθώς και στα δύο ακραία διαστήματα που ορίζονται από την ελάχιστη (και τις μικρότερες από αυτήν) και τη μέγιστη (και τις μεγαλύτερες από αυτήν) τιμή της κλίμακας για τις πραγματικές τιμές της Y .



Σχήμα 4.3: Εποπτικό παράδειγμα απεικόνισης της τιμής της εκτίμησης, $\hat{y}_i$, για την Y .

Όταν η υπολογισθείσα τιμή της εκτίμησης για την Y ανήκει σε ένα από τα $m-1$ διαστήματα διαχωρισμού, αυτή υπολογίζεται με γραμμική παρεμβολή σε αυτά, με βάση τους μαθηματικούς τύπους της Σχέσης (4.9), ενώ λαμβάνει την ελάχιστη τιμή

της κλίμακας εάν ανήκει στο διάστημα που ορίζεται από την ελάχιστη τιμή της κλίμακας και τις μικρότερες από αυτήν τιμές και την μεγαλύτερη τιμή της κλίμακας εάν ανήκει στο διάστημα που ορίζεται από την μεγαλύτερη τιμή της κλίμακας και τις μεγαλύτερες από αυτήν τιμές. Όλα τα παραπάνω, προκειμένου για την εκτίμηση της i -στής παρατήρησης για την Y , δίνονται αναλυτικά από τις Σχέσεις (4.13).

$$\hat{y}_i = \begin{cases} y^0 & \text{εάν } \sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - 0 \leq 0 \\ y^0 + \frac{\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - 0}{y^{*1} - y^{*0}} \cdot (y^1 - y^0) & \text{εάν } y^{*0} \leq \sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} \leq y^{*1} \\ y^1 + \frac{\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - y^{*1}}{y^{*2} - y^{*1}} \cdot (y^2 - y^1) & \text{εάν } y^{*1} \leq \sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} \leq y^{*2} \\ y^2 + \frac{\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - y^{*2}}{y^{*3} - y^{*2}} \cdot (y^3 - y^2) & \text{εάν } y^{*2} \leq \sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} \leq y^{*3} \\ \vdots & \\ y^{a-1} & \text{εάν } \sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - 1 \geq 0 \end{cases} \quad (4.13)$$

Οι πιθανές τιμές που μπορεί να λάβει η εκτίμηση για την Y , φαίνονται στις παραπάνω περιπτώσεις. Ωστόσο, όπως αναφέρθηκε και πριν, δεν είναι δυνατόν να είναι γνωστό το διάστημα στο οποίο θα ανήκει αυτή η εκτίμηση. Θα πρέπει, επομένως, να ληφθεί πρόνοια ώστε το μοντέλο να είναι σε θέση να διακρίνει το διάστημα διαχωρισμού, όπως ορίστηκαν στα προηγούμενα, στο οποίο ανήκει η υπολογισθείσα εκτίμηση για την Y . Για τον λόγο αυτό, η τιμή της εκτίμησης αυτής, μπορεί να ιδωθεί ως ένα άθροισμα των τιμών που περιγράφονται στις περιπτώσεις της Σχέσης (4.13), θέτοντας επιπλέον περιορισμούς στο γ.π., σύμφωνα με τους οποίους θα διακρίνεται το διάστημα διαχωρισμού στο οποίο ανήκει η εκτίμηση για την Y , «ενεργοποιώντας» μόνο τον κατάλληλο όρο του αθροίσματος που αναφέρθηκε πριν. Για να γίνει κατανοητή όλη η διαδικασία αυτή, θα οριστεί το παρακάτω άθροισμα για την εκτίμηση της τιμής για την i -στή παρατήρηση της Y :

$$\hat{y}_i = d_{1i} \cdot y_i^0 + d_{2i} \cdot \left(y_i^0 + \frac{\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - 0}{y_i^{*1} - y_i^{*0}} \cdot (y^1 - y^0) \right) + \dots + d_{a-i} \cdot y^{a-1} \quad (4.14)$$

Κάθε όρος του προηγούμενου αθροίσματος, περιλαμβάνει μία από τις περιπτώσεις τιμών που προβλέπονται στη Σχέση (4.13). Στο σημείο αυτό ορίζονται οι μεταβλητές d_{pi} , τόσες όσες ακριβώς είναι οι περιπτώσεις των τιμών που μπορεί να λάβει η εκτίμηση για την Y (όπως φαίνεται στην Σχέση (4.13)), έτσι ώστε να υπάρχει μια για κάθε όρο του αθροίσματος (4.14). Οι μεταβλητές αυτές ορίζονται να είναι δυαδικές,

λαμβάνοντας την τιμή 0 ή 1, ώστε τελικά να «ενεργοποιηθεί» μόνον ο κατάλληλος όρος του παραπάνω αθροίσματος. Πρόκειται, ουσιαστικά, για ψευδομεταβλητές, οι οποίες προορίζονται μονάχα για τον εντοπισμό του διαστήματος διαχωρισμού όπου ανήκει η υπολογιζόμενη εκτίμηση για την Y . Ο τρόπος με τον οποίο μπορεί να επιτευχθεί κάτι τέτοιο, είναι μέσω της εισαγωγής κατάλληλων περιορισμών αναφορικά με τις ψευδομεταβλητές d_{pi} και περιγράφεται στη συνέχεια.

Αφού οι ψευδομεταβλητές d_{pi} ορίστηκαν για να μετάσχουν στο άθροισμα, είναι λογικό να καθιστούν «ενεργό» τον όρο εκείνο για τον οποίο ικανοποιούνται οι συνθήκες που φαίνονται στη Σχέση (4.13). Ακόμα, επειδή αυτές ορίστηκαν να λαμβάνουν τις τιμές 0 και 1, για να «ενεργοποιηθεί» ο κατάλληλος όρος όταν ικανοποιούνται οι αντίστοιχες συνθήκες, θα πρέπει να ισχύουν και οι περιορισμοί της Σχέσης (4.15) για κάθε ψευδομεταβλητή d_{pi} , όπως οι τελευταίες ορίστηκαν πιο πάνω:

$$\begin{aligned}
 (2d_{1i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - 0 \right) &\leq 0 \\
 (2d_{2i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - 0 \right) &\geq 0 \\
 (2d_{2i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - y^{*1} \right) &\leq 0 \\
 (2d_{3i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - y^{*1} \right) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - y^{*2} \right) &\leq 0 \\
 &\vdots \\
 (2d_{a-li} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - y^{*m-2} \right) &\geq 0 \\
 (2d_{a-li} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w'_{ik} - 1 \right) &\leq 0
 \end{aligned} \tag{4.15}$$

Τέλος, για να εξασφαλισθεί ότι μία μόνο ψευδομεταβλητή θα λάβει την τιμή 1 κατά την επίλυση του προβλήματος, τίθεται και ένας ακόμη περιορισμός (Σχέση 4.16).

$$d_{1i} + d_{2i} + d_{3i} + \dots + d_{a-li} = 1 \tag{4.16}$$

Σύμφωνα με όλα τα παραπάνω, το νέο γ.π., με αντικειμενική συνάρτηση των συντελεστή πολλαπλής παλινδρόμησης R^2 , λαμβάνει την δομή της Σχέσης (4.17). Αξίζει να σημειωθεί ότι ανάλογα με τον τρόπο που ορίζονται οι κλίμακες της συνάρτησης αξιών, μπορούν να απαλειφθούν μερικοί από τους πιο πάνω περιορισμούς. Για παράδειγμα, εάν η οριζόντια κλίμακα των πραγματικών τιμών της Y ορισθεί έτσι ώστε για ακραίες τιμές να έχει την ελάχιστη και τη μικρότερη τιμή των διαθέσιμων παρατηρήσεων, τότε ο πρώτος, ο δεύτερος, ο τρίτος και ο τελευταίος περιορισμός της Σχέσης (4.15), μπορούν να παραλειφθούν από το γ.π.

$$\begin{aligned}
 \max F &= \frac{\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2}{\sum_{i=1}^n (y_i - \bar{Y})^2} = 1 - \frac{\sum_{i=1}^n (\hat{y}_i - y_i)^2}{\sum_{i=1}^n (y_i - \bar{Y})^2} \\
 \text{υ.π.} \\
 \sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - \sum_{m=1}^{a-1} c_j^m z_m' + \sigma_j^- - \sigma_j^+ &= \gamma \sum_{m=1}^{a-1} c_j^m - \sum_{i=1}^n \gamma_i \sum_{k=1}^{a_i-1} c_{ij}^k \\
 \sum_{m=1}^{a-1} z_m' &= 1 - \gamma(a-1) \\
 \sum_{i=1}^n \sum_{k=1}^{a_i-1} w_{ik}' &= 1 - \sum_{i=1}^n \gamma_i(a_i-1) \\
 \left(d_{1i} \cdot y_i^0 + d_{2i} \cdot \left(y_i^0 + \frac{\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - 0}{y_i^{*1} - y_i^{*0}} \cdot (y^1 - y^0) \right) + \dots + d_{a-i} \cdot y_i^{a-1} \right) - \hat{y}_i &= 0 \\
 (2d_{1i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - 0 \right) &\leq 0 \\
 (2d_{2i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - 0 \right) &\geq 0 \\
 (2d_{2i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - y^{*1} \right) &\leq 0 \\
 (2d_{3i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - y^{*1} \right) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - y^{*2} \right) &\leq 0 \\
 \vdots \\
 (2d_{a-i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - y^{*m-2} \right) &\geq 0 \\
 (2d_{a-i} - 1) \cdot \left(\sum_{i=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - 1 \right) &\leq 0 \\
 \sum_{p=1}^{a-1} \sum_{i=1}^n d_{pi} &= 1 \\
 z_m', w_m', \gamma, \gamma_i, \sigma_j^-, \sigma_j^+ &\geq 0 \quad \text{για κάθε } m, i, j, k \\
 d_{pi} &\text{ binary}
 \end{aligned} \tag{4.17}$$

Στην περίπτωση αυτού του γ.π., χρειάζεται να ληφθεί ειδική μέριμνα για τα απόλυτα σφάλματα που έχουν ορισθεί σε αυτό. Και αυτό διότι μπορεί να ναι μεν να ικανοποιούνται όλοι οι σχετικοί περιορισμοί, τα σφάλματα αυτά, ωστόσο, είναι δυνατόν να λάβουν απεριόριστα μεγάλες τιμές. Ο περιορισμός των απόλυτων σφαλμάτων, μπορεί να λάβει χώρα με τρεις τρόπους, θέτοντας τους αντίστοιχους περιορισμούς στο σώμα του ανωτέρω γ.π.

Ο πρώτος τρόπος συνίσταται στο να χρησιμοποιηθούν δύο μεταβλητές για τα απόλυτα σφάλματα, μια για το σφάλμα υποεκτίμησης (σ^-) και μια για το σφάλμα υπερεκτίμησης (σ^+), όπως φαίνεται στο γ.π. (4.17). Εδώ τίθενται οι περιορισμοί της Σχέσης (4.18):

$$\begin{aligned}\sigma_i^- \cdot \sigma_i^+ &= 0 \\ \sigma_i^-, \sigma_i^+ &\leq 1\end{aligned}\tag{4.18}$$

Ο πρώτος περιορισμός τίθεται ώστε να εξασφαλιστεί ότι ένα τουλάχιστον σφάλμα θα λάβει την τιμή 0, ενώ ο δεύτερος τίθεται για να περιοριστεί το εύρος τιμών (σε συνδυασμό με τον περιορισμό $\sigma^+, \sigma^- \geq 0$) που μπορούν να λάβουν οι μεταβλητές σφάλματος.

Ο δεύτερος τρόπος είναι και πάλι να χρησιμοποιηθούν δύο μη αρνητικές μεταβλητές για τα απόλυτα σφάλματα, αλλά αυτή τη φορά να μην περιοριστεί το εύρος τιμών αυτών, παρά μόνον το άθροισμά τους. Μια καλή σκέψη για το άνω όριο αυτού του αθροίσματος, είναι μια τιμή σχετικά κοντά στο αποτέλεσμα που έδωσε το γ.π. (4.12), όπως φαίνεται στη Σχέση (4.19):

$$\sum_i (\sigma_i^+ + \sigma_i^-) \leq (1 + \varepsilon) \cdot F^* \tag{4.19}$$

όπου ε είναι ένας μικρός θετικός αριθμός και F^* είναι η βέλτιστη τιμή της αντικειμενικής συνάρτησης F για το γ.π. (4.12).

Ο τρίτος τρόπος είναι να χρησιμοποιηθεί μια μόνο μεταβλητή σφάλματος, γεγονός που σημαίνει ότι αυτή μπορεί να λάβει τόσο θετικές, όσο και αρνητικές τιμές (επιπλέον της τιμής 0, φυσικά). Στην περίπτωση αυτή, αρκεί να περιορισθεί το εύρος τιμών της μεταβλητής αυτής, π.χ. όπως παρακάτω:

$$-1 \leq \sigma_i \leq 1 \tag{4.20}$$

4.3 Διαδικασία εισαγωγής – εξαγωγής μεταβλητών στο μοντέλο

Στην ενότητα αυτή θα περιγραφεί η διαδικασία με την οποία επιλέγεται ή εξέρχεται μια μεταβλητή από το υπό ανάπτυξη μοντέλο παλινδρόμησης. Η εισαγωγή ή εξαγωγή μιας μεταβλητής, εξαρτάται από την τιμή του χρησιμοποιούμενου κριτηρίου βελτιστοποίησης, όπως αυτά ορίστηκαν στις προηγούμενες παραγράφους. Η διαδικασία χαρακτηρίζεται από ένα σταδιακό (βηματικό) τρόπο για την λήψη απόφασης σχετικά με την εισαγωγή ή εξαγωγή μιας μεταβλητής.

Δύο τρόποι θα εξεταστούν στη συνέχεια, σύμφωνα με τους οποίους μπορεί να υλοποιηθεί η βηματική αυτή διαδικασία δόμησης του μοντέλου παλινδρόμησης. Ο πρώτος τρόπος είναι να θεωρηθεί ότι στο πρώτο βήμα καμμία μεταβλητή δε συμμετέχει στο μοντέλο και ότι οι μεταβλητές εισάγονται βηματικά στο μοντέλο (εάν συντρέχει λόγος βάσει του εξεταζόμενου κριτηρίου), ενώ ο δεύτερος είναι να υποτεθεί ότι όλες οι διαθέσιμες μεταβλητές αρχικά μετέχουν στο μοντέλο και, αν

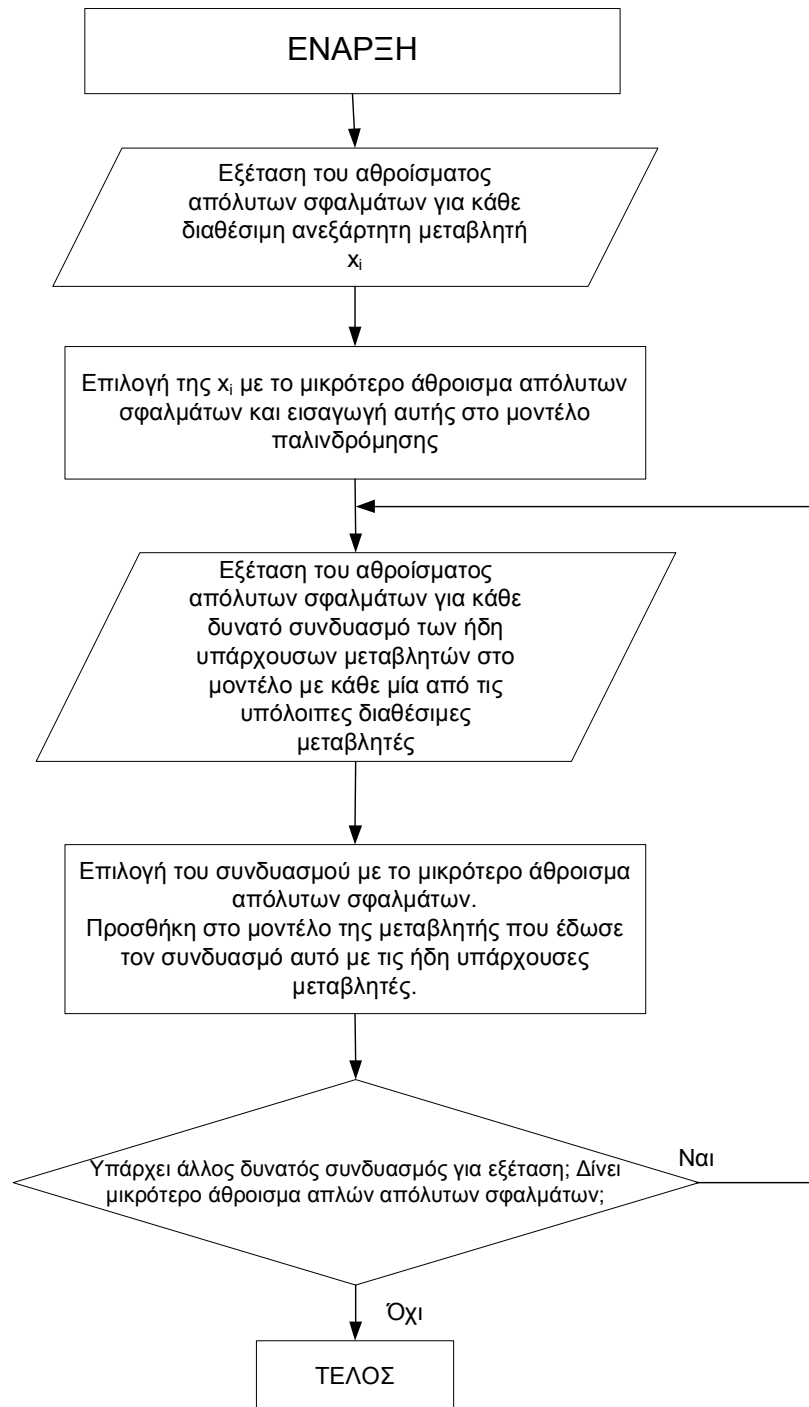
χρειαστεί, εξέρχονται βηματικά του μοντέλου. Στη συνέχεια δίνεται αναλυτική περιγραφή της διαδικασίας που ακολουθείται σε κάθε θεώρηση που αναφέρθηκε, ανάλογα με το εκάστοτε χρησιμοποιούμενο κριτήριο βελτιστοποίησης.

4.3.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Αρχικά, όπως αναφέρθηκε, καμμία μεταβλητή δε συμμετέχει στο μοντέλο. Το πρώτο βήμα αφορά στην επιλογή της πρώτης μεταβλητής στο μοντέλο. Αυτό θα πραγματοποιηθεί με την εξέταση της τιμής του κριτηρίου βελτιστοποίησης που χρησιμοποιείται στα γ.π. (4.8), (4.12) και (4.17). Στη συνέχεια εξετάζεται, με βάση τα ίδια κριτήρια, εάν μπορεί να εισαχθεί και άλλη μεταβλητή στο μοντέλο. Ο τρόπος με τον οποίο γίνεται αυτό, περιγράφεται στη συνέχεια, για κάθε κριτήριο βελτιστοποίησης ξεχωριστά.

4.3.1.1 Άθροισμα απλών απόλυτων σφαλμάτων

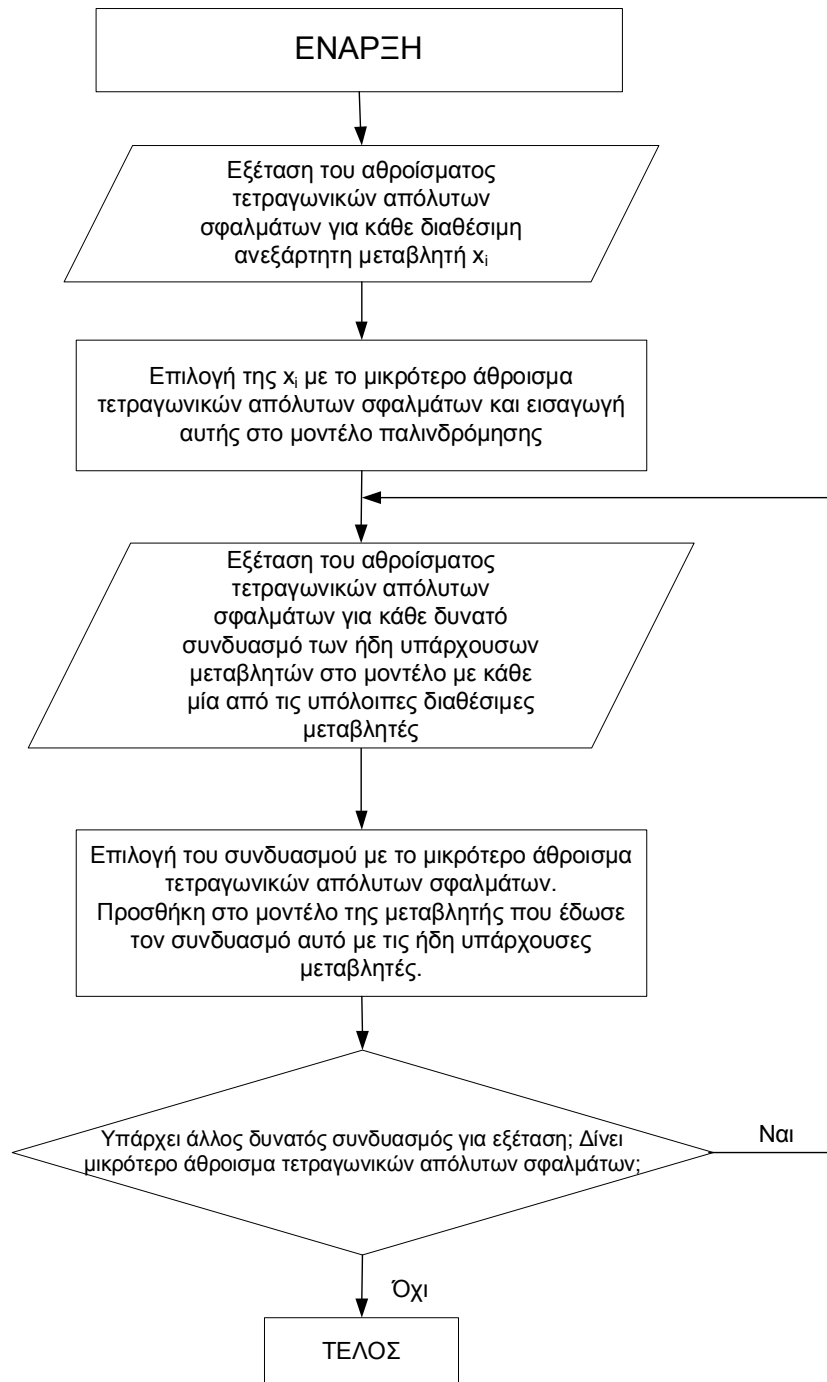
Στην περίπτωση αυτή, στο πρώτο βήμα υλοποιείται το γ.π. (4.8) ξεχωριστά για κάθε διαθέσιμη ανεξάρτητη μεταβλητή. Η μεταβλητή που έχει να επιδείξει το μικρότερο άθροισμα απόλυτων σφαλμάτων, επιλέγεται ως η πρώτη για εισαγωγή στο μοντέλο. Στη συνέχεια, εξετάζονται όλοι οι δυνατοί συνδυασμοί που δίνει η εισαχθείσα μεταβλητή με όλες τις υπόλοιπες διαθέσιμες. Για κάθε ένα τέτοιο συνδυασμό, υλοποιείται και πάλι το γ.π. (4.8), προσαρμοσμένο κατάλληλα για να εξεταστούν οι συνδυασμοί που αναφέρθηκαν. Ο συνδυασμός που θα δώσει το μικρότερο άθροισμα απόλυτων σφαλμάτων, επιλέγεται ως ο καλύτερος για το δεύτερο βήμα της διαδικασίας, κάτι που σημαίνει ότι εκτός από την πρώτη μεταβλητή που επιλέχθηκε, θα εισαχθεί στο μοντέλο και η άλλη μεταβλητή που υποδεικνύει ο επιλεγείς συνδυασμός. Το επόμενο βήμα είναι να εξεταστεί ο συνδυασμός του τελευταίου επιλεγέντα συνδυασμού μεταβλητών, με όλες τις υπόλοιπες διαθέσιμες ανεξάρτητες μεταβλητές. Και πάλι υλοποιείται το γ.π. (4.8), με τρόπο ώστε να ανταποκρίνεται στα χαρακτηριστικά του συνδυασμού που εξετάζεται. Ο συνδυασμός με το μικρότερο άθροισμα απλών απόλυτων σφαλμάτων, επιλέγεται ως ο επόμενος καλύτερος, άρα εισάγεται στο μοντέλο η μεταβλητή που συνιστά τον συνδυασμό αυτό μαζί με τις προηγούμενες που ήδη υπάρχουν στο μοντέλο. Η διαδικασία αυτή συνεχίζεται μέχρι να εξεταστούν όλοι οι δυνατοί συνδυασμοί που υποδεικνύονται με τον τρόπο ως οι καλύτεροι και τερματίζεται όταν δεν υπάρχει άλλος δυνατός συνδυασμός για να εξεταστεί ή δεν δώσει μικρότερο άθροισμα απλών απόλυτων σφαλμάτων από τον προηγούμενο. Τελικά, το προτεινόμενο μοντέλο παλινδρόμησης, αποτελείται από τις μεταβλητές που συνιστούν τον συνδυασμό με το μικρότερο άθροισμα απόλυτων σφαλμάτων, βάσει των αποτελεσμάτων που δίνει η υλοποίηση του γ.π. (4.8) σε κάθε βήμα της διαδικασίας. Το Σχήμα 4.4, δείχνει το διάγραμμα ροής αυτής της διαδικασίας.



Σχήμα 4.4: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου προς τα εμπρός βηματικής ποιοτικής παλινδρόμησης, για κριτήριο βελτιστοποίησης το άθροισμα των απλών απόλυτων σφαλμάτων.

4.3.1.2 Άθροισμα τετραγωνικών σφαλμάτων

Η διαδικασία που ακολουθείται με βάση το κριτήριο αυτό, δε διαφέρει σημαντικά από αυτήν της παραγράφου 4.3.1.1. Απλώς, σε κάθε βήμα υλοποιείται το γ.π. (4.12), το οποίο διαφέρει από το αντίστοιχο της Σχέση (4.12) μόνον στο ότι αλλάζει το κριτήριο βελτιστοποίησης.



Σχήμα 4.5: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου προς τα εμπρός βηματικής ποιοτικής παλινδρόμησης, για κριτήριο βελτιστοποίησης το άθροισμα των τετραγωνικών σφαλμάτων.

Η διαδικασία έχει ως εξής: αρχικά υλοποιείται το γ.π. (4.12) ξεχωριστά για κάθε διαθέσιμη ανεξάρτητη μεταβλητή. Η μεταβλητή που έχει να επιδείξει το μικρότερο άθροισμα τετραγωνικών σφαλμάτων, επιλέγεται ως η πρώτη για εισαγωγή στο μοντέλο. Στη συνέχεια, εξετάζονται όλοι οι δυνατοί συνδυασμοί που δίνει η εισαχθείσα μεταβλητή με όλες τις υπόλοιπες διαθέσιμες. Για κάθε ένα τέτοιο συνδυασμό, υλοποιείται και πάλι το γ.π. (4.12), προσαρμοσμένο κατάλληλα για να εξεταστούν οι συνδυασμοί που αναφέρθηκαν. Ο συνδυασμός που θα δώσει το μικρότερο άθροισμα τετραγωνικών σφαλμάτων, επιλέγεται ως ο καλύτερος για το δεύτερο βήμα της διαδικασίας, κάτι που σημαίνει ότι εκτός από την πρώτη μεταβλητή

που επιλέχθηκε, θα εισαχθεί στο μοντέλο και η άλλη μεταβλητή που υποδεικνύει ο επιλεγείς συνδυασμός. Η διαδικασία συνεχίζεται υλοποιώντας το γ.π. (4.12), με τρόπο ώστε να ανταποκρίνεται στα χαρακτηριστικά του κάθε συνδυασμού που εξετάζεται. Ο συνδυασμός με το μικρότερο άθροισμα τετραγωνικών σφαλμάτων, επιλέγεται ως ο επόμενος καλύτερος, άρα εισάγεται στο μοντέλο η μεταβλητή που συνιστά τον συνδυασμό αυτό μαζί με τις προηγούμενες που ήδη υπάρχουν στο μοντέλο. Όταν εξεταστούν όλοι οι δυνατοί συνδυασμοί που υποδεικνύονται με τον τρόπο ως οι καλύτεροι ή όταν δεν προκύψει συνδυασμός που να δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων από τον προηγούμενο, η διαδικασία τερματίζεται. Τελικά, το προτεινόμενο μοντέλο παλινδρόμησης, αποτελείται από τις μεταβλητές που συνιστούν τον συνδυασμό με το μικρότερο άθροισμα τετραγωνικών σφαλμάτων, βάσει των αποτελεσμάτων που δίνει η υλοποίηση του γ.π. (4.12) σε κάθε βήμα της διαδικασίας. Το Σχήμα 4.5, δείχνει το διάγραμμα ροής αυτής της διαδικασίας.

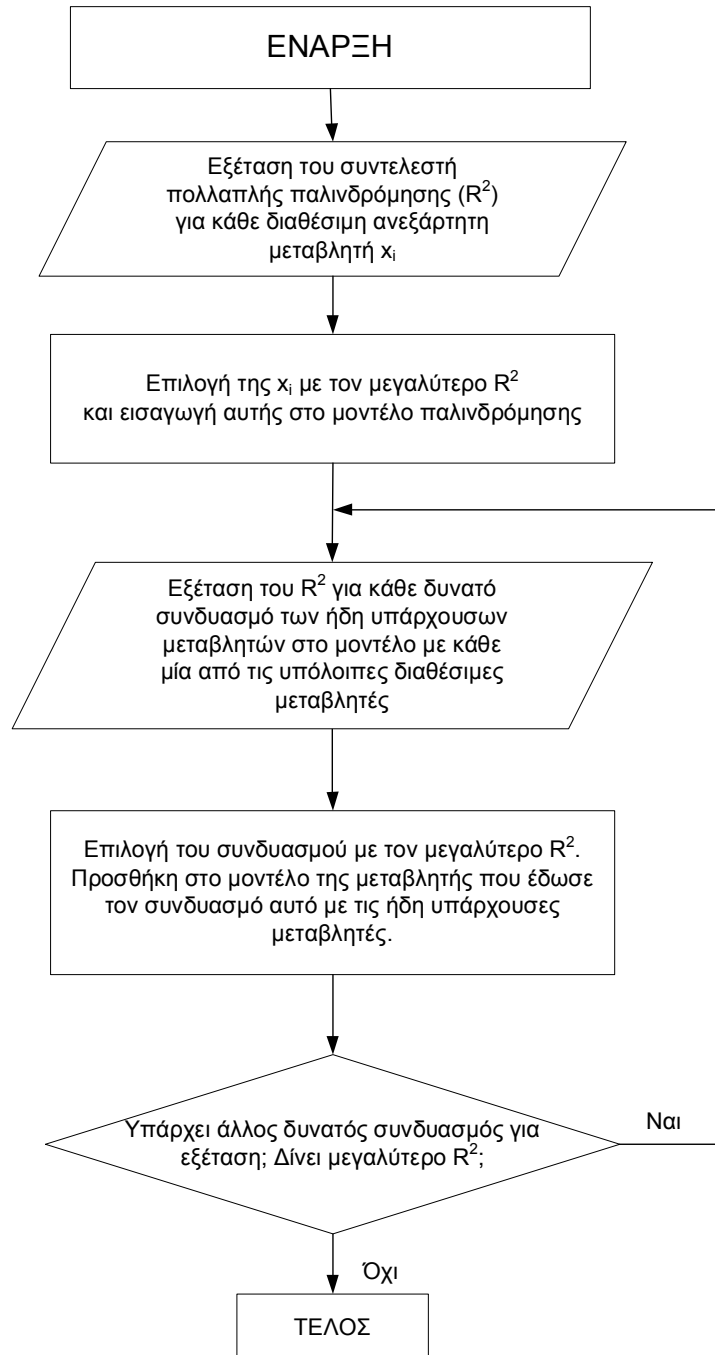
4.3.1.3 Συντελεστής προσδιορισμού

Με βάση αυτό το κριτήριο βελτιστοποίησης, η βηματική διαδικασία που ακολουθείται χαρακτηρίζεται από την υλοποίηση του γ.π. (4.17) σε κάθε βήμα της. Η διαφοροποίηση εδώ είναι ότι το κριτήριο βελτιστοποίησης μεγιστοποιείται σε κάθε εφαρμογή του γ.π. (4.17), οπότε ως καλύτερος συνδυασμός επιλέγεται αυτός ο οποίος δίνει τη μεγαλύτερη τιμή για το συγκεκριμένο κριτήριο. Η διαδικασία παρουσιάζει μεγάλη ομοιότητα στη διαδοχή των διεργασιών σε σχέση με τις προηγούμενες δύο μεθοδολογίες. Πιο αναλυτικά, ακολουθούνται τα εξής βήματα: στο πρώτο βήμα, υλοποιείται το γ.π. (4.17) για κάθε μία από τις διαθέσιμες ανεξάρτητες μεταβλητές. Η μεταβλητή που θα δώσει τον μεγαλύτερο συντελεστή πολλαπλής παλινδρόμησης, επιλέγεται ως η πρώτη που θα εισαχθεί στο μοντέλο. Ακολούθως, εξετάζονται όλοι οι δυνατοί συνδυασμοί αυτής της μεταβλητής, με όλες τις υπόλοιπες διαθέσιμες και για κάθε τέτοιο συνδυασμό, υλοποιείται το γ.π. (4.17). Ο συνδυασμός με το μεγαλύτερο συντελεστή πολλαπλής παλινδρόμησης, επιλέγεται ως ο καλύτερος για το συγκεκριμένο βήμα, οπότε η μεταβλητή που έδωσε μαζί την ήδη υπάρχουσα στο μοντέλο τον εν λόγω συνδυασμό, επιλέγεται για εισαγωγή στο μοντέλο παλινδρόμησης. Η ίδια διαδικασία ακολουθείται σε όλα τα επόμενα βήματα, έως ότου δεν βρεθεί συνδυασμός που να δίνει μεγαλύτερο R^2 από τον προηγούμενο ή εάν δεν υπάρχει άλλος δυνατός συνδυασμός μεταβλητών για να εξεταστεί. Ως καλύτερος συνδυασμός μεταβλητών αναδεικνύεται αυτός που θα προκύψει με το πέρας της παραπάνω διαδικασίας. Ένα διάγραμμα ροής των διεργασιών κατά την διαδικασία που περιγράφηκε πριν, φαίνεται στο Σχήμα 4.6.

4.3.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Η διαφορά αυτής της προσέγγισης με την προηγούμενη της παραγράφου 4.3.1, είναι ότι αρχικά γίνεται η υπόθεση της συμμετοχής όλων των διαθέσιμων ανεξάρτητων μεταβλητών στο μοντέλο. Σε κάθε βήμα αυτής της προσέγγισης, και πάλι υλοποιούνται τα γ.π. (4.8), (4.12) και (4.17), ανάλογα με τη μέθοδο που ακολουθείται. που χρησιμοποιείται. Στη συνέχεια εξετάζεται, με βάση τα αντίστοιχα κριτήρια βελτιστοποίησης, ποια μεταβλητή μπορεί να εξαχθεί από το μοντέλο. Ο

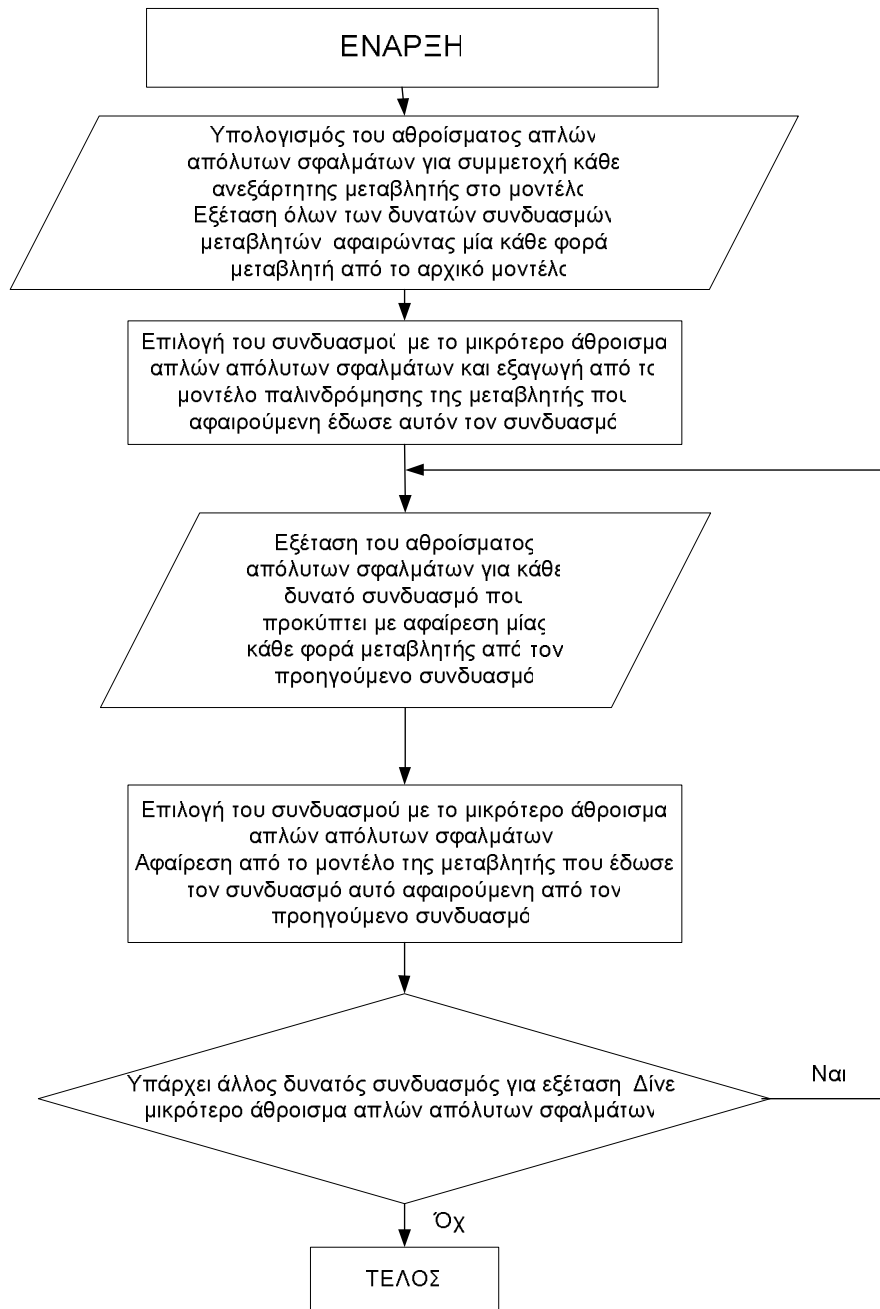
τρόπος με τον οποίο γίνεται αυτό, περιγράφεται στη συνέχεια, για κάθε κριτήριο βελτιστοποίησης ξεχωριστά.



Σχήμα 4.6: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου προς τα εμπρός βηματικής ποιοτικής παλινδρόμησης, για κριτήριο βελτιστοποίησης τον συντελεστή πολλαπλής παλινδρόμησης.

4.3.2.1 Άθροισμα απλών απόλυτων σφαλμάτων

Επειδή γίνεται η θεώρηση ότι όλες οι διαθέσιμες ανεξάρτητες μεταβλητές συμμετέχουν στο μοντέλο, υπολογίζεται το άθροισμα των απλών απόλυτων σφαλμάτων με βάση το γ.π. (4.8), για τον συγκεκριμένο συνδυασμό μεταβλητών.



Σχήμα 4.7: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου προς τα πίσω βηματικής ποιοτικής παλινδρόμησης, για κριτήριο βελτιστοποίησης το άθροισμα των απλών απόλυτων σφαλμάτων.

Το επόμενο βήμα είναι να εξεταστεί εάν μπορεί να εξαχθεί κάποια μεταβλητή από το μοντέλο, έτσι ώστε να οδηγηθούμε σε συνδυασμό μεταβλητών που να δίνουν μικρότερο άθροισμα απόλυτων σφαλμάτων από τον προηγούμενο. Για τον λόγο αυτό, υλοποιείται το γ.π. (4.8) για κάθε συνδυασμό μεταβλητών που υπολείπεται κατά μία μεταβλητή από τον αρχικό συνδυασμό. Εάν βρεθεί συνδυασμός μεταβλητών που να δίνει μικρότερο άθροισμα απλών απόλυτων σφαλμάτων από το αρχικό, επιλέγεται ως ο καλύτερος συνδυασμός για το βήμα αυτό και η μεταβλητή που αφαιρούμενη δίνει τον συνδυασμό αυτό, εξέρχεται από το μοντέλο. Στη συνέχεια, εξετάζεται εάν μπορεί να εξαχθεί, με παρόμοιο τρόπο, και άλλη μεταβλητή από το μοντέλο. Με βάση τον τελευταίο συνδυασμό μεταβλητών, εξετάζεται κάθε επόμενος δυνατός συνδυασμός,

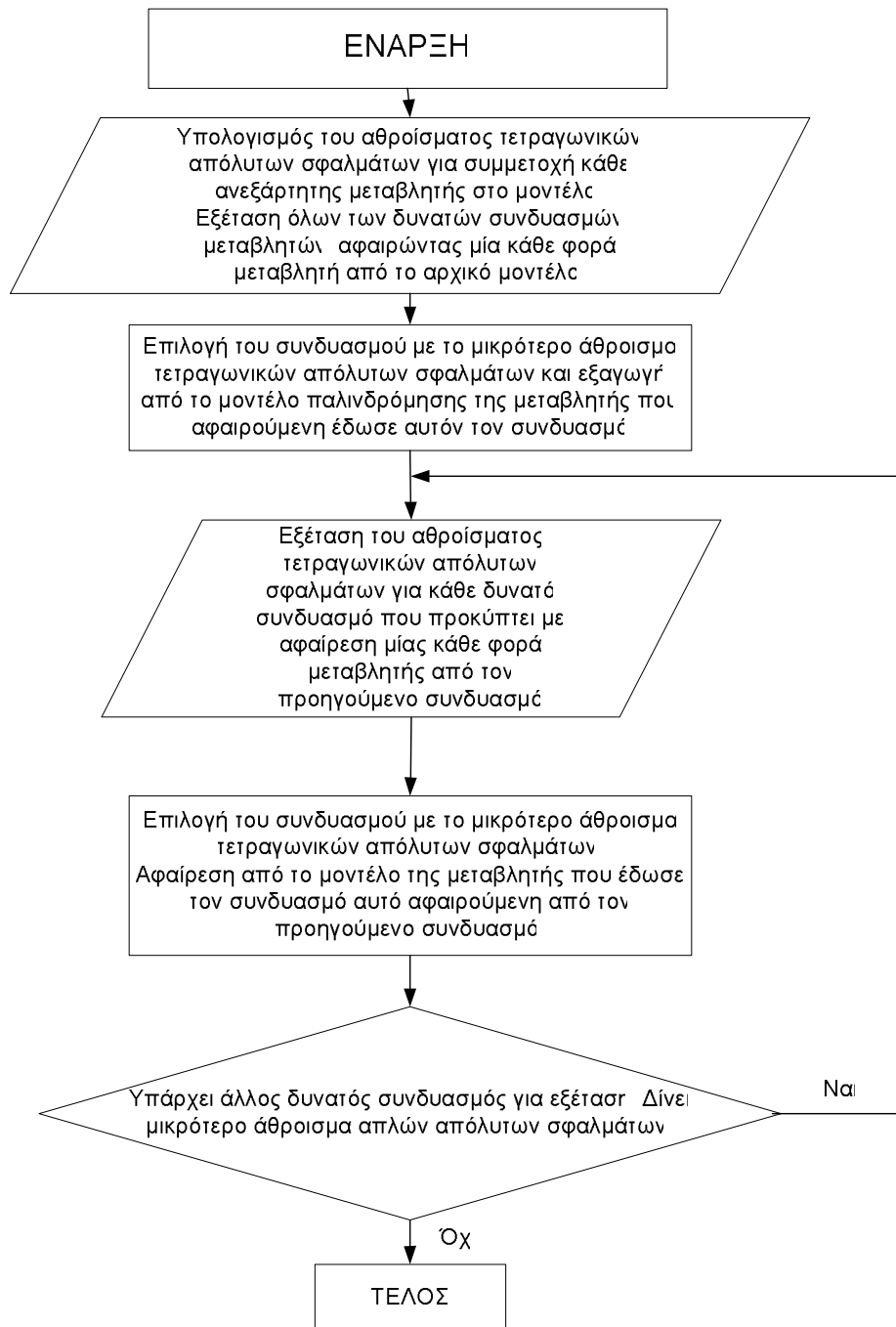
αφαιρώντας μία κάθε φορά μεταβλητή από αυτόν. Για τους συνδυασμούς αυτούς υλοποιείται το γ.π. (4.8) και εάν υπάρξει συνδυασμός που να δίνει μικρότερο άθροισμα απλών απόλυτων σφαλμάτων από τον προηγούμενο, επιλέγεται ως ο καλύτερος για το συγκεκριμένο βήμα και εξέρχεται από το μοντέλο η μεταβλητή που αφαιρούμενη από τον προηγούμενο συνδυασμό έδωσε τον συγκεκριμένο συνδυασμό. Η διαδικασία συνεχίζεται με παρόμοιο τρόπο, μέχρι να μην ευρεθεί συνδυασμός που να δίνει μικρότερο άθροισμα απλών απόλυτων σφαλμάτων από τον προηγούμενο ή μέχρι να μην υπάρχει άλλος δυνατός συνδυασμός για να εξεταστεί. Στο Σχήμα 4.7 δίνεται εποπτικά η ακολουθία των διεργασιών που ακολουθούνται στην προσέγγιση αυτή.

4.3.2.2 Άθροισμα τετραγωνικών σφαλμάτων

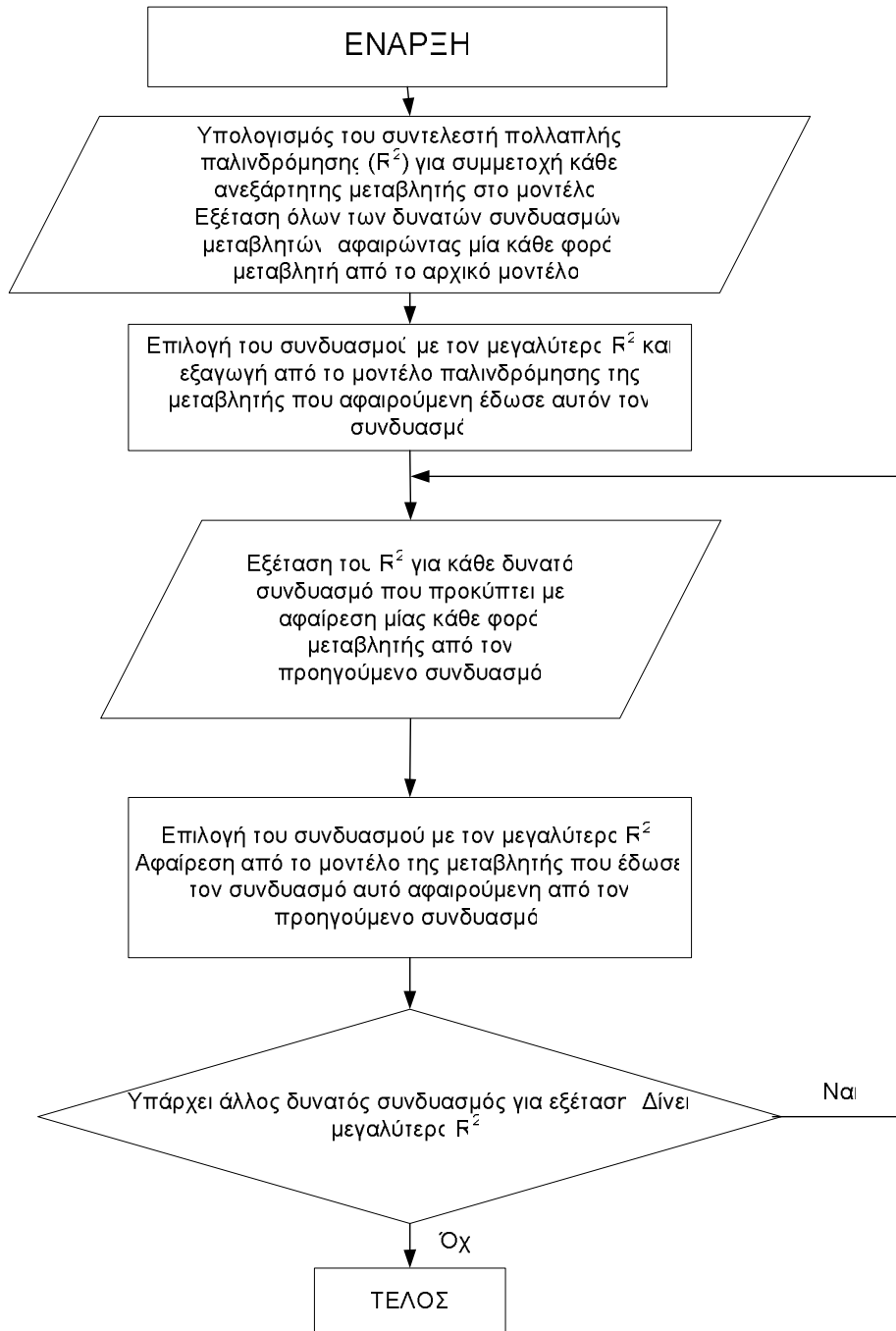
Στην περίπτωση αυτή, πάλι εξετάζεται εάν μπορεί να εξαχθεί κάποια μεταβλητή από το μοντέλο, υλοποιώντας σε κάθε βήμα υλοποιείται το γ.π. (4.12). Αφού αρχικά μετέχουν όλες οι διαθέσιμες ανεξάρτητες μεταβλητές στο μοντέλο, υπολογίζεται το άθροισμα των τετραγωνικών σφαλμάτων με χρήση του γ.π. (4.12). Έπειτα, πάλι εξετάζεται κάθε δυνατός συνδυασμός μεταβλητών, αφαιρώντας μία κάθε φορά μεταβλητή από το ήδη υπάρχον μοντέλο. Εάν προκύψει συνδυασμός που να δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων, επιλέγεται ως ο καλύτερος για το βήμα αυτό και η μεταβλητή που δίνει αφαιρούμενη τον συγκεκριμένο συνδυασμό, εξέρχεται του μοντέλου. Η διαδικασία συνεχίζει με παρόμοιο τρόπο, έως ότου δεν προκύψει συνδυασμός που να δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων από τον προηγούμενο ή όταν δεν μπορεί να υπάρξει άλλος δυνατός συνδυασμός για εξέταση. Το τελικό προτεινόμενο μοντέλο παλινδρόμησης, αποτελείται από τον συνδυασμό μεταβλητών που εξετάστηκε στο τελευταίο βήμα της διαδικασίας. Ένα διάγραμμα ροής για τις ακολουθούμενες διεργασίες αυτής της περίπτωσης, φαίνεται στο Σχήμα 4.8

4.3.2.3 Συντελεστής προσδιορισμού

Αρχικά υπολογίζεται η τιμή του R^2 με τη χρήση του γ.π. (4.17), υποθέτοντας ότι όλες οι διαθέσιμες ανεξάρτητες μεταβλητές, μετέχουν στο μοντέλο. Στη συνέχεια, αφαιρείται από τον αρχικό συνδυασμό μεταβλητών μία κάθε φορά μεταβλητή, ώστε να σχηματιστούν όλοι οι δυνατοί συνδυασμοί μεταβλητών που μπορεί να προκύψουν με τον τρόπο αυτό. Για κάθε έναν από τους συνδυασμούς αυτούς, υλοποιείται το γ.π. (4.17), ώστε να βρεθεί, εάν υπάρχει, ένας συνδυασμός μεταβλητών που να δίνει μεγαλύτερο R^2 . Εάν βρεθεί, ο συνδυασμός αυτός επιλέγεται ως ο καλύτερος για το συγκεκριμένο βήμα, ενώ η μεταβλητή που αφαιρείται για να προκύψει ο συνδυασμός αυτός, εξάγεται από το μοντέλο. Έπειτα, επαναλαμβάνεται η ίδια διαδικασία, ώστε να ευρεθεί, εάν υπάρχει, συνδυασμός μεταβλητών που να δίνει μεγαλύτερη τιμή για το R^2 από την αντίστοιχη του προηγούμενου βήματος. Εάν προκύψει, επιλέγεται ως ο καλύτερος για αυτό το βήμα πλέον και η μεταβλητή που αφαιρείται για να προκύψει ο εν λόγω συνδυασμός, εξάγεται από το μοντέλο. Η διαδικασία αυτή εφαρμόζεται σε κάθε βήμα, μέχρι να μην προκύψει συνδυασμός μεταβλητών που να δίνει μεγαλύτερο R^2 από τον συνδυασμό του προηγούμενου βήματος ή μέχρι να μην είναι δυνατόν να προκύψει άλλος δυνατός συνδυασμός για εξέταση. Το διάγραμμα ροής που αποτυπώνει αυτές τις διεργασίες, φαίνεται στο Σχήμα 4.9.



Σχήμα 4.8: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου προς τα πίσω βηματικής ποιοτικής παλινδρόμησης, για κριτήριο βελτιστοποίησης το άθροισμα των τετραγωνικών σφαλμάτων.



Σχήμα 4.9: Διάγραμμα ροής διαδικασιών στα πλαίσια της μεθόδου προς τα πίσω βηματικής ποιοτικής παλινδρόμησης, για κριτήριο βελτιστοποίησης τον συντελεστή πολλαπλής παλινδρόμησης.

ΚΕΦΑΛΑΙΟ

5

ΠΡΑΚΤΙΚΕΣ ΕΦΑΡΜΟΓΕΣ

Το κεφάλαιο αυτό περιγράφει όλες τις προσπάθειες για την πρακτική εφαρμογή του αναπτυχθέντος μοντέλου (Κεφάλαιο 4). Οι προσπάθειες αυτές έγιναν σε δεδομένα που αντλήθηκαν από βάσεις δεδομένων στο διαδίκτυο ή από βιβλία της διεθνούς βιβλιογραφίας, τα οποία ασχολούνται με το θέμα της ανάπτυξης μοντέλων παλινδρόμησης. Στη συνέχεια περιγράφονται και παρουσιάζονται αυτά τα δεδομένα, παρουσιάζονται τα αποτελέσματα που δίνει το αναπτυχθέν μοντέλο και ταυτόχρονα γίνονται κάποιες συγκρίσεις και σχόλια σχετικά με τα αποτελέσματα που προκύπτουν από άλλες μεθοδολογίες εφαρμοζόμενες στα ίδια δεδομένα. Οι μεθοδολογίες αυτές είναι οι τρεις πιο γνωστές μεθοδολογίες βηματικής παλινδρόμησης: προς τα εμπρός επιλογή, προς τα πίσω απαλοιφή, βηματική παλινδρόμηση. Για την σύγκριση με τις μεθοδολογίες αυτές, χρησιμοποιήθηκε το στατιστικό πακέτο SPSS v.12, εγκατεστημένο σε περιβάλλον Windows XP. Πρέπει να αναφερθεί στο σημείο αυτό ότι τα διαθέσιμα δεδομένα χωρίζονται σε δύο κατηγορίες: σε δεδομένα διερεύνησης και σε δεδομένα εφαρμογής. Η πρώτη κατηγορία χρησιμεύει για να γίνει αναλυτική διερεύνηση των συνδυασμών μεταβλητών που μπορούν να προκύψουν από τις διαθέσιμες μεταβλητές και να μελετηθεί βήμα προς βήμα η υπό πρόταση μεθοδολογία, ενώ η δεύτερη κατηγορία δεδομένων χρησιμοποιείται για να εφαρμοσθούν τα συμπεράσματα που συνάγονται από την εφαρμογή στην πρώτη. Ας σημειωθεί ότι στην μελέτη κάθε συνόλου δεδομένων, εκτιμάται και η τιμή μερικών επιπρόσθετων στατιστικών στοιχείων (δείκτες αξιολόγησης) με βάση τη διερεύνηση που υλοποιείται. Αυτά είναι ο συντελεστής προσδιορισμού R^2 (βλ. Κεφάλαιο 2), το άθροισμα των απόλυτων σφαλμάτων αξιών ($Sum(|\epsilon|)$), το άθροισμα των τετραγωνικών σφαλμάτων αξιών ($Sum(\epsilon^2)$), το άθροισμα των απόλυτων σφαλμάτων εκτίμησης (πρωτογενή σφάλματα, $Sum(|\epsilon|)$), το άθροισμα των τετραγωνικών σφαλμάτων εκτίμησης (πρωτογενή σφάλματα, $Sum(\epsilon^2)$) και το μέσο απόλυτο ποσοστιαίο σφάλμα ($MAPE$), τα οποία συνοψίζονται στον Πίνακα 5.1. Αξίζει να επισημανθεί ότι τα σφάλματα αξιών συμβολίζονται με σ , ενώ τα πρωτογενή σφάλματα συμβολίζονται με το λατινικό γράμμα ϵ . Ακόμα, με y_i συμβολίζεται η πραγματική διαθέσιμη τιμή για την i -στη παρατήρηση της Y , ενώ με $\hat{y}_i$ συμβολίζεται η αντίστοιχη υπολογιζόμενη εκτίμηση για αυτήν. Ο συμβολισμός αυτός ακολουθείται και στο υπόλοιπο μέρος του κεφαλαίου αυτού.

5.1 Τα δεδομένα διερεύνησης

Στη συνέχεια παρουσιάζονται τα σύνολα δεδομένων τα οποία χρησιμοποιήθηκαν ως δεδομένα διερεύνησης στα πλαίσια της παρούσας εργασίας. Για τα δεδομένα αυτά, εξετάστηκαν εξαντλητικά όλοι οι δυνατοί συνδυασμοί των

διαθέσιμων ανεξάρτητων μεταβλητών για κάθε σύνολο δεδομένων, με σκοπό να ιδωθεί βήμα προς βήμα η εφαρμογή της προτεινόμενης μεθοδολογίας. Τα αποτελέσματα που προκύπτουν από την προσπάθεια αυτή, παρουσιάζονται στη συνέχεια.

Πίνακας 5.1:Επιμέρους δείκτες αξιολόγησης για την ανάλυση των αποτελεσμάτων.

Δείκτης Αξιολόγησης	Σχέση Υπολογισμού
Sum(σ)	$\sum_{i=1}^n \left(\left \sum_{j=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - \sum_{m=1}^{a-1} c_j^m z_m' \right \right)$
Sum(σ ²)	$\sum_{i=1}^n \left(\sum_{j=1}^n \sum_{k=1}^{a_i-1} c_{ij}^k w_{ik}' - \sum_{m=1}^{a-1} c_j^m z_m' \right)^2$
Sum(e)	$\sum_{i=1}^n (y_i - \hat{y}_i)$
Sum(e ²)	$\sum_{i=1}^n (y_i - \hat{y}_i)^2$
MAPE	$\frac{1}{n} \cdot \sum_{i=1}^n \left(\frac{ y_i - \hat{y}_i }{y_i} \right)$

5.1.1 Πρακτική Εφαρμογή 1: Τα Δεδομένα Water Usage of Production Plant

Ο Πίνακας 5.2 περιέχει τα δεδομένα για την κατανάλωση νερού σε μια εγκατάσταση παραγωγικής διαδικασίας.

Πίνακας 5.2: Τα δεδομένα Water Usage of Production Plant.

a/a	x ₁	x ₂	x ₃	x ₄	y
1	58.8	7107	21	129	3067
2	65.2	6373	22	141	2828
3	70.9	6796	22	153	2891
4	77.4	9208	20	166	2994
5	79.3	14792	25	193	3082
6	81	14564	23	189	3898
7	71.9	11964	20	175	3502
8	63.9	13526	23	186	3060
9	54.5	12656	20	190	3211
10	39.5	14119	20	187	3286
11	44.5	16691	22	195	3542
12	43.6	14571	19	206	3125
13	56	13619	22	198	3022
14	64.7	14575	22	192	2922
15	73	14556	21	191	3950
16	78.9	18573	21	200	4488
17	79.4	15618	22	200	3295

Τα δεδομένα αντλήθηκαν από βάση δεδομένων η οποία διατηρείται στο διαδίκτυο και η ηλεκτρονική διεύθυνση στην οποία μπορούν να βρεθούν είναι: <http://www.statsci.org/data/general/water.html>. Ο σκοπός για τον οποίο συλλέχθηκαν τα δεδομένα που περιγράφονται στη συνέχεια, είναι για να εξεταστεί ο τρόπος με τον οποίο είναι δυνατόν να μειωθεί το κόστος λειτουργίας της εγκατάστασης, ελέγχοντας την κατανάλωση νερού, βάσει τεσσάρων παραγόντων που φαίνεται να επιδρούν σε αυτή. Οι παράγοντες οι οποίοι επιλέχθηκαν και συγκεντρώθηκαν, είναι οι: η μέση μηνιαία θερμοκρασία (x_1 , σε $^{\circ}F$), ποσότητα παραγωγής (x_2 , σε pounds – 1 pound = 0,45359 kg), αριθμός ημερών που λειτουργεί η εγκατάσταση μέσα σε ένα μήνα (x_3 , σε ημέρες), αριθμός ατόμων που εργάζονται στην εγκατάσταση μέσα σε ένα μήνα (x_4 , σε αριθμό ατόμων), ενώ η εξαρτημένη μεταβλητή είναι η μηνιαία κατανάλωση νερού (y , σε γαλόνια – 1 γαλόνι = 3,78541 λίτρα). Οι συσχετίσεις μεταξύ των μεταβλητών αυτών, φαίνονται στον Πίνακα 5.3.

Πίνακας 5.3: Συσχετίσεις μεταξύ των μεταβλητών των δεδομένων.

	x_1	x_2	x_3	x_4	y
x_1	1.000				
x_2	-0.024	1.000			
x_3	0.438	0.106	1.000		
x_4	-0.082	0.918	0.032	1.000	
y	0.286	0.631	-0.089	0.413	1.000

Εφαρμόζοντας τις τρεις βηματικές μεθόδους στα δεδομένα αυτά, λαμβάνονται τα αποτελέσματα του Πίνακα 5.4.

5.1.1.1 Άθροισμα Απόλυτων Σφαλμάτων

Εφαρμόζοντας την προτεινόμενη μεθοδολογία του τέταρτου κεφαλαίου της παρούσης εργασίας και υλοποιώντας το γ.π. (4.8) για κάθε συνδυασμό των διαθέσιμων στο σύνολο δεδομένων μεταβλητών, λαμβάνονται τα αποτελέσματα του Πίνακα 5.5. Στον πίνακα αυτόν, σκιασμένη φαίνεται η στήλη η οποία περιέχει τα αποτελέσματα για την αντικειμενική συνάρτηση του αντίστοιχου για την περίπτωση αυτή γ.π. (εδώ το γ.π. (4.8)).

5.1.1.1.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Εφαρμόζοντας την μεθοδολογία που περιγράφηκε στην παράγραφο 4.3.1.1, στο πρώτο βήμα αυτής, υλοποιείται το γ.π. (4.8) για κάθε μία από τις x_1 , x_2 , x_3 και x_4 . Τα αποτελέσματα φαίνονται στον Πίνακα 5.5. Με βάση τα αποτελέσματα αυτά, προκύπτει ότι η x_2 δίνει το μικρότερο άθροισμα απόλυτων σφαλμάτων (1,873), οπότε εισάγεται στο μοντέλο. Με δεδομένο ότι η x_2 βρίσκεται ήδη στο μοντέλο, εξετάζονται όλοι οι συνδυασμοί αυτής με όλες τις υπόλοιπες διαθέσιμες μεταβλητές, δηλαδή οι συνδυασμοί x_1 - x_2 , x_2 - x_3 και x_2 - x_4 . Τα αποτελέσματα αυτής της διερεύνησης, φαίνονται επίσης στον Πίνακα 5.5 και προκύπτει ότι ο συνδυασμός x_2 - x_3 δίνει το μικρότερο άθροισμα απόλυτων σφαλμάτων, οπότε επιλέγεται ως ο καλύτερος για το βήμα αυτό. μάλιστα, το άθροισμα αυτό είναι μικρότερο από αυτό του προηγούμενου βήματος (1,562), οπότε η x_3 επιλέγεται για είσοδο στο μοντέλο.

Πίνακας 5.4: Αποτελέσματα για την εφαρμογή της βηματικής παλινδρόμησης, με τη χρήση του πακέτου SPSS v.12, στα δεδομένα Water Usage of Production Plant.

Συνδυασμός	Βηματική Παλινδρόμηση			Εμπρός Επιλογή			Πίσω Απαλοιφή		
	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης
x1-x2-x3-x4	0.987	0.985	406.333	0.987	0.985	406.333	0.986	0.984	418.118
x1-x2-x3	0.986	0.984	418.118	0.986	0.984	418.118	0.986	0.984	418.118
x1-x2-x4	0.987	0.985	406.333	0.987	0.985	406.333	0.987	0.985	406.333
x1-x3-x4	0.987	0.985	406.333	0.987	0.985	406.333	0.987	0.985	406.333
x2-x3-x4	0.982	0.981	459.436	0.982	0.981	459.436	0.986	0.984	418.118
x1-x2	0.986	0.984	426.247	0.986	0.984	426.247	0.986	0.984	426.247
x1-x3	0.977	0.976	517.452	0.977	0.976	517.452	0.977	0.976	517.452
x1-x4	0.987	0.985	406.333	0.987	0.985	406.333	0.987	0.985	406.333
x2-x3	0.986	0.984	418.118	0.986	0.984	418.118	0.986	0.984	418.118
x2-x4	0.982	0.981	459.436	0.982	0.981	459.436	0.982	0.981	459.436
x3-x4	0.982	0.981	459.436	0.982	0.981	459.436	0.982	0.981	459.436
x1	0.959	0.956	696.535	0.959	0.956	696.535	0.959	0.956	696.535
x2	0.959	0.957	693.503	0.959	0.957	693.503	0.959	0.957	693.503
x3	0.977	0.976	517.452	0.977	0.976	517.452	0.977	0.976	517.452
x4	0.982	0.981	459.436	0.982	0.981	459.436	0.982	0.981	459.436

Πίνακας 5.5: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Water Usage of Production Plant.

Συνδυασμός	R^2	Sum(e)	Sum(σ^2)	Sum(e)	Sum(e^2)	MAPE
x1-x2-x3-x4	0.752	1.354	0.326	2219.426	815414.720	0.037
x1-x2-x3	0.750	1.351	0.256	2207.559	820776.737	0.037
x1-x2-x4	0.577	1.569	0.478	3328.110	1713683.054	0.061
x1-x3-x4	0.550	1.969	1.042	3285.906	1511067.235	0.056
x2-x3-x4	0.600	1.565	0.384	2643.432	1471990.358	0.043
x1-x2	0.615	1.578	0.404	3413.403	1446863.411	0.062
x1-x3	0.363	2.485	0.676	4028.918	2419314.487	0.071
x1-x4	0.369	2.107	1.364	4725.103	2595289.564	0.083
x2-x3	0.600	1.562	0.383	2637.249	1470239.249	0.042
x2-x4	0.489	1.877	0.755	4276.195	1847751.274	0.074
x3-x4	0.158	2.070	1.063	5874.031	5338260.255	0.103
x1	0.169	3.808	2.240	7516.176	5217940.463	0.135
x2	0.475	1.873	0.761	4292.912	2028193.374	0.073
x3	0.008	4.044	2.971	6782.771	6160363.330	0.112
x4	0.089	2.107	1.370	6408.970	4634826.679	0.112

Επειδή μπορούν να προκύψουν και άλλοι δυνατοί συνδυασμοί για εξέταση, η διαδικασία προχωρά και εξετάζονται οι συνδυασμοί μεταβλητών που περιέχουν τις x2 και x3, δηλαδή οι x1-x2-x3 και x2-x3-x4, υλοποιώντας το γ.π. (4.8). Από αυτούς, μικρότερο άθροισμα απόλυτων σφαλμάτων, και μάλιστα μικρότερο από το αντίστοιχο του προηγούμενου βήματος, δίνει ο x1-x2-x3 (1,351), οπότε επιλέγεται ως ο καλύτερος για το βήμα αυτό. Η μεταβλητή x1 επιλέγεται για είσοδο στο μοντέλο. Επειδή υπάρχει και άλλος ένας συνδυασμός για εξέταση, ο οποίος να περιλαμβάνει ήδη τις x1, x2 και x3, εξετάζεται και αυτός, δηλαδή ο συνδυασμός x1-x2-x3-x4, ο

οποίος δίνει μεγαλύτερο άθροισμα απόλυτων σφαλμάτων από τους συνδυασμούς του προηγούμενου βήματος, δεν επιλέγεται ως καλύτερος και άρα δεν υπάρχει μεταβλητή για είσοδο στο μοντέλο παλινδρόμησης. Η διαδικασία σταματά εδώ, αναδεικνύοντας ως καλύτερο υποσύνολο μεταβλητών για επιλογή τον συνδυασμό $x_1-x_2-x_3$, με τα χαρακτηριστικά που φαίνονται στον Πίνακα 5.5. Συνοπτικά τα βήματα που ακολουθήθηκαν, φαίνονται στον Πίνακα 5.6.

Πίνακας 5.6: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα απλών απόλυτων σφαλμάτων, στα δεδομένα Water Usage of Production Plant.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	x_1, x_2, x_3, x_4	x_2	1.873	x_2
2	$x_1-x_2, x_2-x_3, x_2-x_4$	x_2-x_3	1.562	x_3
3	$x_1-x_2-x_3, x_2-x_3-x_4$	$x_1-x_2-x_3$	1.351	x_1
4	$x_1-x_2-x_3-x_4$	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: $x_1-x_2-x_3$				

5.1.1.1.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Χρησιμοποιώντας τα αποτελέσματα του Πίνακα 5.5 και εφαρμόζοντας τη μεθοδολογία που περιγράφηκε στην παράγραφο 4.3.1.2, μπορούν να ληφθούν τα παρακάτω αποτελέσματα. Θεωρείται αρχικά ότι μετέχουν όλες οι διαθέσιμες ανεξάρτητες μεταβλητές στο μοντέλο (και οι τέσσερις), για τις οποίες εξετάζεται το άθροισμα των απλών απόλυτων σφαλμάτων. Το άθροισμα αυτό συγκρίνεται με το αντίστοιχο που προκύπτει από τη χρήση του γ.π. (4.8) για όλους τους συνδυασμούς που προκύπτουν, εάν αφαιρεθεί μία κάθε φορά μεταβλητή από τον αρχικό. Η σύγκριση αυτή δείχνει ότι ο συνδυασμός μεταβλητών $x_1-x_2-x_3$ δίνει άθροισμα απόλυτων σφαλμάτων μικρότερο από αυτό του αρχικού συνδυασμού, οπότε επιλέγεται ως ο καλύτερος για το βήμα αυτό. Επειδή ο συνδυασμός αυτός προκύπτει αφαιρώντας τη x_4 από τον προηγούμενο συνδυασμό, η τελευταία εξέρχεται από το μοντέλο. Στο επόμενο βήμα προκύπτει ότι κανένας από τους εξεταζόμενους δυνατούς συνδυασμούς δε δίνει μικρότερο άθροισμα απόλυτων σφαλμάτων από τον αντίστοιχο του προηγούμενου βήματος, οπότε η διαδικασία σταματά εδώ, αναδεικνύοντας τον συνδυασμό $x_1-x_2-x_3$ ως τον καλύτερο προτεινόμενο για τη δόμηση του μοντέλου παλινδρόμησης. Τα επιμέρους στατιστικά χαρακτηριστικά που προκύπτουν για την επιλογή αυτού του συνδυασμού, φαίνονται στον Πίνακα 5.5, ενώ ο Πίνακας 5.7 συνοψίζει τα βήματα που ακολουθήθηκαν στην διαδικασία που περιγράφηκε πριν.

5.1.1.2 Άθροισμα Τετραγωνικών Σφαλμάτων

Στην περίπτωση αυτή, θα ακολουθηθεί παρόμοια διαδικασία με πριν, με τη διαφορά ότι τώρα, κριτήριο βελτιστοποίησης θεωρείται το άθροισμα των τετραγωνικών σφαλμάτων (τα αποτελέσματα για κάθε συνδυασμό φαίνονται στην τέταρτη στήλη του Πίνακα 5.8).

Πίνακας 5.7: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα απλών απόλυτων σφαλμάτων, στα δεδομένα Water Usage of Production Plant.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3, x1-x2-x4, x1-x3-x4, x2-x3-x4	x1-x2-x3	1.351	x4
2	x1-x2, x1-x3, x2-x3	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x3				

Πίνακας 5.8: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.12), στα δεδομένα Water Usage of Production Plant.

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sum(ϵ)	Sum(ϵ^2)	MAPE
x1-x2-x3-x4	0.721	1.674	0.237	3093.184	905556.275	0.054
x1-x2-x3	0.722	1.672	0.237	3088.636	902722.823	0.054
x1-x2-x4	0.621	1.798	0.338	3639.302	1295965.946	0.065
x1-x3-x4	0.414	2.340	0.383	4826.935	2008377.653	0.084
x2-x3-x4	0.555	1.801	0.327	3383.502	1482773.924	0.057
x1-x2	0.622	1.797	0.338	3637.605	1295585.672	0.065
x1-x3	0.335	2.626	0.640	4516.028	2470691.434	0.079
x1-x4	0.225	2.815	0.640	6083.105	3150226.377	0.107
x2-x3	0.556	1.798	0.327	3379.552	1480041.570	0.057
x2-x4	0.549	2.105	0.450	4088.472	1561454.655	0.072
x3-x4	0.149	2.594	0.612	5829.122	4232005.376	0.100
x1	0.161	4.525	1.555	7615.136	5554980.881	0.133
x2	0.551	2.118	0.450	4074.187	1554321.534	0.071
x3	0.000	4.386	1.573	7984.170	6311011.535	0.136
x4	0.060	2.722	0.867	6833.228	5301055.007	0.118

5.1.1.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Θεωρώντας ότι αρχικά καμμία μεταβλητή δε μετέχει στο μοντέλο, εξετάζονται όλες οι διαθέσιμες ανεξάρτητες μεταβλητές, εφαρμόζοντας το γ.π. (4.12) για κάθε μία από αυτές, ως προς το άθροισμα των τετραγωνικών σφαλμάτων. Η εξέταση αυτή δείχνει ότι το μικρότερο άθροισμα των τετραγωνικών σφαλμάτων το παρουσιάζει η x2, οπότε εισάγεται στο μοντέλο. Στο επόμενο βήμα εξετάζονται οι συνδυασμοί οι οποίοι συμπεριλαμβάνουν τη μεταβλητή αυτή και εξετάζονται ως προς το άθροισμα των τετραγωνικών σφαλμάτων που παρουσιάζουν. Προκύπτει ότι στο βήμα αυτό καλύτερος συνδυασμός είναι ο x1-x2, ο οποίος παρουσιάζει και μικρότερο άθροισμα των τετραγωνικών σφαλμάτων από τον συνδυασμό του προηγούμενου βήματος, οπότε εισάγεται στο μοντέλο η μεταβλητή x1. Στη συνέχεια, σύμφωνα με τα όσα προτείνει η μεθοδολογία, εξετάζονται οι πιθανοί συνδυασμοί οι οποίοι περιέχουν τις μεταβλητές x1 και x2. Από αυτούς, γίνεται φανερό ότι ο συνδυασμός x1-x2-x3

είναι καλύτερος από τον αντίστοιχο του προηγούμενου βήματος, οπότε γίνεται η εισαγωγή της x_3 στο μοντέλο. Στο επόμενο βήμα εξετάζεται ο συνδυασμός x_1 - x_2 - x_3 - x_4 (ο οποίος περιέχει τις και τρεις μεταβλητές του προηγούμενου βήματος), ωστόσο δεν προκύπτει να δίνει μικρότερο άθροισμα των τετραγωνικών σφαλμάτων από τον συνδυασμό του προηγούμενου βήματος, με αποτέλεσμα η διαδικασία να σταματά εδώ. Καλύτερο υποσύνολο μεταβλητών προτείνεται να είναι το x_1 - x_2 - x_3 με τα χαρακτηριστικά που φαίνονται στον Πίνακα 5.8. Τα βήματα της παραπάνω διαδικασίας, αποτυπώνονται στον Πίνακα 5.9.

Πίνακας 5.9: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα των τετραγωνικών σφαλμάτων, στα δεδομένα Water Usage of Production Plant.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	x_1, x_2, x_3, x_4	x_2	0.450	x_2
2	x_1 - x_2 , x_2 - x_3 , x_2 - x_4	x_2 - x_3	0.338	x_3
3	x_1 - x_2 - x_3 , x_2 - x_3 - x_4	x_1 - x_2 - x_3	0.237	x_1
4	x_1 - x_2 - x_3 - x_4	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_1 - x_2 - x_3				

5.1.1.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Στην παράγραφο αυτή θα εφαρμοσθεί η διαδικασία που περιγράφεται στην παράγραφο 4.3.2.2 στα δεδομένα του Πίνακα 5.8. Αρχικά, θεωρείται ότι μετέχουν όλες οι διαθέσιμες μεταβλητές στο μοντέλο και υπολογίζεται το άθροισμα τετραγωνικών σφαλμάτων για το συνδυασμό αυτό. Έπειτα, αφαιρείται μία κάθε φορά μεταβλητή από τον αρχικό συνδυασμό και οι συνδυασμοί που προκύπτουν, εξετάζονται ως προς το άθροισμα τετραγωνικών σφαλμάτων που παρουσιάζουν, εφαρμόζοντας το γ.π. (4.12) σε κάθε έναν από αυτούς. Τα αθροίσματα αυτά συγκρίνονται με το προηγούμενο και προκύπτει ότι ο συνδυασμός x_1 - x_2 - x_3 παρουσιάζει μικρότερο άθροισμα τετραγωνικών σφαλμάτων από τον συνδυασμό του προηγούμενου βήματος. Συνεπώς, επιλέγεται ως ο καλύτερος για το βήμα αυτό και η μεταβλητή x_4 εξέρχεται από το μοντέλο. Στη συνέχεια εξετάζονται όλοι οι δυνατοί συνδυασμοί που προκύπτουν από την αφαίρεση μίας κάθε φορά μεταβλητής από τον συνδυασμό x_1 - x_2 - x_3 (του προηγούμενου βήματος) και εξετάζονται πάλι ως προς το άθροισμα τετραγωνικών σφαλμάτων. Η εξέταση των συνδυασμών αυτών δείχνει ότι δεν υπάρχει συνδυασμός με άθροισμα τετραγωνικών σφαλμάτων μικρότερο από τον αντίστοιχο του προηγούμενου βήματος, οπότε η διαδικασία σταματά εδώ, με προτεινόμενο καλύτερο υποσύνολο το x_1 - x_2 - x_3 . Τα χαρακτηριστικά του υποσυνόλου αυτού φαίνονται στον Πίνακα 5.8, ενώ η διαδικασία που ακολουθήθηκε παρουσιάζεται στον Πίνακα 5.10.

Πίνακας 5.10: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα των τετραγωνικών σφαλμάτων, στα δεδομένα Water Usage of Production Plant.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Αθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3, x1-x2-x4, x1-x3-x4, x2-x3-x4	x1-x2-x3	0.237	x4
2	x1-x2, x1-x3, x2-x3	-	-	-

Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x3

5.1.1.3 Συντελεστής προσδιορισμού

Στην περίπτωση αυτή, αλλάζει το κριτήριο βελτιστοποίησης του χρησιμοποιούμενου γ.π. Το γ.π. το οποίο χρησιμοποιείται, είναι αυτό της Σχέσης 4.17. Αξίζει, ωστόσο, να σημειωθεί ότι παρατηρώντας την μαθηματική έκφραση της αντικειμενικής συνάρτησης σε αυτό το γ.π., ο παρονομαστής είναι ένας σταθερός αριθμός ο οποίος προκύπτει μετά από κατάλληλες αριθμητικές πράξεις, οπότε το πρόβλημα ανάγεται στη μεγιστοποίηση του αριθμητή της πρώτης έκφρασης $(\sum_{i=1}^n (\hat{y}_i - \bar{Y})^2)$ ή στην ελαχιστοποίηση του αριθμητή της δεύτερης έκφρασης $(\sum_{i=1}^n (\hat{y}_i - y_i)^2)$. Στην συγκεκριμένη περίπτωση, επιλέχθηκε η ελαχιστοποίηση της δεύτερης έκφρασης, γεγονός το οποίο εκτιμάται ότι συμβάλει στη μείωση της πολυπλοκότητας των υπολογισμών και, κατά συνέπεια, του υπολογιστικού φόρτου. Ας επισημανθεί ότι για την επίλυση του εν λόγω μικτού μη γραμμικού προβλήματος, χρησιμοποιήθηκε το υπολογιστικό πακέτο LINGO 8.0, σε περιβάλλον Windows XP.

Τα δεδομένα που χρησιμοποιούνται στην περίπτωση αυτή, φαίνονται στον Πίνακα 5.2, ενώ τα αποτελέσματα των διερευνήσεων για κάθε εξεταζόμενο συνδυασμό με τη χρήση του γ.π. 4.17. φαίνονται στον Πίνακα 5.11.

5.1.1.3.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Στην περίπτωση αυτή, ακολουθείται η διαδικασία της παραγράφου 4.3.2.3. Με βάση τα αποτελέσματα του Πίνακα 5.11, υπάρχει καλύτερος συνδυασμός, με μία μεταβλητή λιγότερη από όσες του αρχικού, καθώς δίνει μικρότερη τιμή για την αντικειμενική συνάρτηση στο υπό εξέταση πρόβλημα ελαχιστοποίησης (και κατά συνέπεια, μεγαλύτερη τιμή για τον συντελεστή πολλαπλής συσχέτισης). Ο συνδυασμός αυτός είναι ο x1-x2-x3 και στο επόμενο βήμα εξετάζονται οι συνδυασμοί οι οποίοι προκύπτουν από αυτόν με την αφαίρεση μίας κάθε φορά μεταβλητής. Η εξέταση των συνδυασμών αυτών με την ανάλογη υλοποίηση του γ.π. (4.17), υποδεικνύει ότι δεν υπάρχει κάποιος συνδυασμός ο οποίος να δίνει καλύτερο αποτέλεσμα από αυτόν του προηγούμενου βήματος. Έτσι, η διαδικασία σταματά στο

σημείο αυτό και καλύτερος συνδυασμός μεταβλητών αναδεικνύεται, με βάση τη διαδικασία αυτή, ο x1-x2-x3. Τα βήματα τα οποία ακολουθήθηκαν στην περίπτωση αυτή, φαίνονται στον Πίνακα 5.12.

Πίνακας 5.11: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.17), στα δεδομένα Water Usage of Production Plant.

Εξεταζόμενος Συνδυασμός	Τιμή Αντικειμενικής Συνάρτησης $(\sum_{i=1}^n (\hat{y}_i - y_i)^2)$	$\sum_{i=1}^n (y_i - \bar{Y})^2$	Υπολογιζόμενος R^2
x1-x2-x3-x4	571912	3192631.529	82.09%
x1-x2-x3	566759.9		82.25%
x1-x2-x4	1007990		68.43%
x1-x3-x4	750779		76.48%
x2-x3-x4	1292100		59.53%
x1-x2	992556		68.91%
x1-x3	2314120		27.52%
x1-x4	2089820		35,54%
x2-x3	1288840		59.63%
x2-x4	1390330		56.45%

Πίνακας 5.12: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για τον συντελεστή πολλαπλής παλινδρόμησης, στα δεδομένα Water Usage of Production Plant.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	R^2 Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3, x1-x2-x4, x1-x3-x4, x2-x3-x4	x1-x2-x3	82.18%	x4
2	x1-x2, x1-x3, x2-x3	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x3				

5.1.1.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα

Στα προηγούμενα εξετάστηκε η «επίδοση» των προτεινόμενων μεθοδολογιών, στα δεδομένα Water Usage of Production Plant. Τα δεδομένα αυτά διαθέτουν 4 ανεξάρτητες μεταβλητές και μία εξαρτημένη μεταβλητή. για το μοντέλο παλινδρόμησης, σε σχέση με τις κλασσικές μεθόδους βηματικής παλινδρόμησης. Ο Πίνακας 5.13 συγκεντρώνει τα αποτελέσματα των συγκρινόμενων μεθοδολογιών.

Πίνακας 5.13: Αποτελέσματα των χρησιμοποιούμενων προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Water Usage of Production Plant.

Εξεταζόμενος Συνδυασμός	Βηματική Παλινδρόμηση		Βηματική Ποιοτική Παλινδρόμηση			
	Εμπρός Επιλογή	Πίσω Απαλοιφή	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
	R^2	R^2	$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4	0.987	0.986	1.354	0.752	0.237	0.721
x1-x2-x3	0.986	0.986	1.351	0.75	0.237	0.722
x1-x2-x4	0.987	0.987	1.569	0.577	0.338	0.621
x1-x3-x4	0.987	0.987	1.969	0.55	0.383	0.414
x2-x3-x4	0.982	0.986	1.565	0.6	0.327	0.555
x1-x2	0.986	0.986	1.578	0.615	0.338	0.622
x1-x3	0.977	0.977	2.485	0.363	0.64	0.335
x1-x4	0.987	0.987	2.107	0.369	0.64	0.225
x2-x3	0.986	0.986	1.562	0.6	0.327	0.556
x2-x4	0.982	0.982	1.877	0.489	0.45	0.549
x3-x4	0.982	0.982	2.07	0.158	0.612	0.149
x1	0.959	0.959	3.808	0.169	1.555	0.161
x2	0.959	0.959	1.873	0.475	0.45	0.551
x3	0.977	0.977	4.044	0.008	1.573	0
x4	0.982	0.982	2.107	0.089	0.867	0.06

Στον ίδιο πίνακα, παρατηρούνται μεγάλες αποκλίσεις μεταξύ του R^2 ο οποίος προκύπτει για τους αντίστοιχους εξεταζόμενους συνδυασμούς με βάση τη χρησιμοποίηση της κλασσικής μεθόδου βηματικής παλινδρόμησης σε σύγκριση με την προτεινόμενη μεθοδολογία βηματικής ποιοτικής παλινδρόμησης. Οι αποκλίσεις στα αποτελέσματα αυτά αμβλύνονται, εάν συγκριθούν μεταξύ τους τα αποτελέσματα που δίνει η χρησιμοποίηση των δύο διαφορετικών κριτηρίων (άθροισμα απλών και τετραγωνικών σφαλμάτων) στην ίδια (προτεινόμενη) μεθοδολογία. Προχωρώντας παραπέρα το συλλογισμό αυτό, παρατηρείται ότι για τα δύο αυτά κριτήρια, τα αποτελέσματα που λαμβάνονται για κάθε συνδυασμό για το ένα κριτήριο, είναι, κατά κανόνα, αντίστοιχα με αυτά που λαμβάνονται για το άλλο κριτήριο (συνδυασμοί που δίδουν μικρότερο άθροισμα σφαλμάτων, εμφανίζουν μεγαλύτερο συντελεστή πολλαπλής παλινδρόμησης). Ακολουθώντας, εξάλλου, τα βήματα τα οποία προτείνονται στις παραγράφους 4.3.1.1 και 4.3.1.2, προκύπτει ότι, με βάση τα δύο αυτά κριτήρια, καταλήγουμε στο ίδιο τελικό προτεινόμενο καλύτερο υποσύνολο μεταβλητών.

Χρησιμοποιώντας το κριτήριο του συντελεστή πολλαπλής παλινδρόμησης στα πλαίσια της προτεινόμενης βηματικής ποιοτικής παλινδρόμησης, μπορούν να γίνουν οι παρακάτω παρατηρήσεις στα δεδομένα του Πίνακα 5.14, ο οποίος συνοψίζει τα αποτελέσματα για τα τρία κριτήρια που εξετάζονται στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής ποιοτικής παλινδρόμησης.

Πίνακας 5.14: Αποτελέσματα των χρησιμοποιούμενων τριών κριτηρίων, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Water Usage of Production Plant.

Βηματική Ποιοτική Παλινδρόμηση					
Εξεταζόμενος Συνδυασμός	R^2	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
		$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4	82.09%	1.354	0.752	0.237	0.721
x1-x2-x3	82.18%	1.351	0.75	0.237	0.722
x1-x2-x4	68.43%	1.569	0.577	0.338	0.621
x1-x3-x4	76.48%	1.969	0.55	0.383	0.414
x2-x3-x4	59.53%	1.565	0.6	0.327	0.555
x1-x2	68.91%	1.578	0.615	0.338	0.622
x1-x3	27.52%	2.485	0.363	0.64	0.335
x1-x4	34.54%	2.107	0.369	0.64	0.225
x2-x3	59.63%	1.562	0.6	0.327	0.556
x2-x4	56.45%	1.877	0.489	0.45	0.549

Αυτό που μπορεί να ισχυριστεί κανείς, είναι ότι τα αποτελέσματα για τη χρησιμοποίηση του συντελεστή πολλαπλής παλινδρόμησης ως κριτήριο βελτιστοποίησης στα πλαίσια της προτεινόμενης μεθοδολογίας, οδηγεί, σε γενικές γραμμές, σε καλύτερα ή παραπλήσια αποτελέσματα για κάθε συνδυασμό από τα αντίστοιχα κριτήρια του αθροίσματος των απόλυτων σφαλμάτων. Ακόμα, όμως, μια σημαντική παρατήρηση είναι ότι ακολουθώντας τις διαδικασίες που περιγράφονται στις ενότητες 4.3.1 και 4.3.2, όλες οι μεθοδολογίες και τα τρία κριτήρια υποδεικνύουν το ίδιο υποσύνολο ως το καλύτερο για συμμετοχή στο ζητούμενο μοντέλο παλινδρόμησης.

Τέλος, παρατηρώντας τους επιπλέον δείκτες που εξετάστηκαν κατά τη διάρκεια των διερευνήσεων, είναι δυνατόν να εξαχθούν τα παρακάτω συμπεράσματα, παρατηρώντας τις τιμές του Πίνακα 5.15. Η γενική παρατήρηση η οποία μπορεί να γίνει, είναι ότι τα αποτελέσματα για τους επιμέρους δείκτες, είναι αντίστοιχα για κάθε συνδυασμό ανάλογα με το εξεταζόμενο κριτήριο. Δηλαδή, όσο βελτιώνεται ή χειροτερεύει η τιμή ενός δείκτη για κάθε συνδυασμό μεταβλητών, αντίστοιχα μεταβάλλεται και η τιμή των υπόλοιπων δεικτών για κάθε κριτήριο. Αυτό θα μπορούσε να αποτελέσει αφορμή για την εξέταση αυτών των δεικτών ως αξιόπιστου κριτηρίου στα πλαίσια της προτεινόμενης μεθοδολογίας. Μια ακόμη παρατήρηση, ωστόσο, η οποία μπορεί να γίνει εδώ, είναι ότι το άθροισμα των απλών σφαλμάτων πρωτογενών ($\sum |e|$) για κάθε συνδυασμό εμφανίζεται να είναι χειρότερο από την περίπτωση που το άθροισμα των τετραγωνικών σφαλμάτων χρησιμοποιείται ως κριτήριο (αντικειμενική συνάρτηση) και, αντίστροφα, το άθροισμα των τετραγωνικών πρωτογενών σφαλμάτων ($\sum (e)^2$) για κάθε συνδυασμό εμφανίζεται χειρότερο στην περίπτωση που χρησιμοποιείται ως αντικειμενική συνάρτηση το άθροισμα απλών απόλυτων σφαλμάτων. Τέλος, όλοι οι υπόλοιποι δείκτες, εκτός του αθροίσματος απλών/τετραγωνικών σφαλμάτων, φαίνεται να έχουν καλύτερες επιδόσεις, για το συγκεκριμένο σύνολο δεδομένων, στην περίπτωση του αθροίσματος απλών απόλυτων σφαλμάτων.

Πίνακας 5.15: Αποτελέσματα των επιμέρους χρησιμοποιούμενων δεικτών, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Water Usage of Production Plant.

Εξεταζόμενος Συνδυασμός	Απλά Απόλυτα Σφάλματα				Τετραγωνικά Σφάλματα			
	R^2	Sum(lel)	Sum(e ²)	MAPE	R^2	Sum(lel)	Sum(e ²)	MAPE
x1-x2-x3-x4	0.752	2219.426	815414.72	0.037	0.721	3093.184	905556.275	0.054
x1-x2-x3	0.75	2207.559	820776.737	0.037	0.722	3088.636	902722.823	0.054
x1-x2-x4	0.577	3328.11	1713683.054	0.061	0.621	3639.302	1295965.95	0.065
x1-x3-x4	0.55	3285.906	1511067.235	0.056	0.414	4826.935	2008377.65	0.084
x2-x3-x4	0.6	2643.432	1471990.358	0.043	0.555	3383.502	1482773.92	0.057
x1-x2	0.615	3413.403	1446863.411	0.062	0.622	3637.605	1295585.67	0.065
x1-x3	0.363	4028.918	2419314.487	0.071	0.335	4516.028	2470691.43	0.079
x1-x4	0.369	4725.103	2595289.564	0.083	0.225	6083.105	3150226.38	0.107
x2-x3	0.6	2637.249	1470239.249	0.042	0.556	3379.552	1480041.57	0.057
x2-x4	0.489	4276.195	1847751.274	0.074	0.549	4088.472	1561454.66	0.072
x3-x4	0.158	5874.031	5338260.255	0.103	0.149	5829.122	4232005.38	0.1
x1	0.169	7516.176	5217940.463	0.135	0.161	7615.136	5554980.88	0.133
x2	0.475	4292.912	2028193.374	0.073	0.551	4074.187	1554321.53	0.071
x3	0.008	6782.771	6160363.33	0.112	0	7984.17	6311011.54	0.136
x4	0.089	6408.97	4634826.679	0.112	0.06	6833.228	5301055.01	0.118

5.1.2 Πρακτική Εφαρμογή 2: Τα Δεδομένα Solar Thermal Energy

Τα δεδομένα αυτά παρουσιάζονται στο Παράρτημα Β του βιβλίου των Montgomery και Peck (1992) και αφορούν τη συνολική ροή θερμότητας (y , σε kwatts) στην επιφάνεια ενός ηλιακού συλλέκτη, ως συνάρτηση πέντε επιμέρους παραγόντων. Οι παράγοντες αυτοί είναι: η ηλιακή ακτινοβολία (x_1 , σε watts/m²), η εστιακή απόσταση του συλλέκτη στην ανατολική διεύθυνση (x_2 , σε ίντσες), η εστιακή απόσταση του συλλέκτη στην νότια διεύθυνση (x_3 , σε ίντσες), η εστιακή απόσταση του συλλέκτη στη βόρεια διεύθυνση (x_4 , σε ίντσες) και η ώρα της ημέρας (x_5). Οι συσχετίσεις μεταξύ όλων των μεταβλητών, φαίνονται στον Πίνακα 5.16, ενώ οι τιμές αυτών των δεδομένων, φαίνονται στον Πίνακα 5.17.

Πίνακας 5.16: Συσχετίσεις μεταξύ των μεταβλητών των δεδομένων.

	x1	x2	x3	x4	x5	y
x1	1.000					
x2	-0.203	1.000				
x3	-0.025	-0.329	1.000			
x4	-0.584	-0.117	0.287	1.000		
x5	-0.494	-0.065	0.697	0.685	1.000	
y	0.623	0.102	0.112	-0.849	-0.351	1.000

Πίνακας 5.17: Τα δεδομένα Solar Thermal Energy (Montgomery and Peck, 1992).

a/a	y	x ₁	x ₂	x ₃	x ₄	x ₅
1	271.8	783.35	33.53	40.55	16.66	13.2
2	264	748.45	36.5	36.19	16.46	14.11
3	238.8	684.45	34.66	37.31	17.66	15.68
4	230.7	827.8	33.13	32.52	17.5	10.53
5	251.6	860.45	35.75	33.71	16.4	11
6	257.9	875.15	34.46	34.14	16.28	11.31
7	263.9	909.45	34.6	34.85	16.06	11.96
8	266.5	905.55	35.38	35.89	15.93	12.58
9	229.1	756	35.85	33.53	16.6	10.66
10	239.3	769.35	35.68	33.79	16.41	10.85
11	258	793.5	35.35	34.72	16.17	11.41
12	257.6	801.65	35.04	35.22	15.92	11.91
13	267.3	819.65	34.07	36.5	16.04	12.85
14	267	808.55	32.2	37.6	16.19	13.58
15	259.6	774.95	34.32	37.89	16.62	14.21
16	240.4	711.85	31.08	37.71	17.37	15.56
17	227.2	694.85	35.73	37	18.12	15.83
18	196	638.1	34.11	36.76	18.53	16.41
19	278.7	774.55	34.79	34.62	15.54	13.1
20	272.3	757.9	35.77	35.4	15.7	13.63
21	267.4	753.35	36.44	35.96	16.45	14.51
22	254.5	704.7	37.82	36.26	17.62	15.38
23	224.7	666.8	35.07	36.34	18.12	16.1
24	181.5	568.55	35.26	35.9	19.05	16.73
25	227.5	563.1	35.56	31.84	16.51	10.58
26	253.6	704.05	35.73	33.16	16.02	11.28
27	263	709.6	36.46	33.83	15.89	11.91
28	265.8	726.9	36.26	34.89	15.83	12.65
29	263.8	697.15	37.2	36.27	16.71	14.06

Για τα δεδομένα αυτά και για κάθε δυνατό συνδυασμό μεταβλητών ξεχωριστά, εφαρμόστηκαν οι τρεις κλασσικές βηματικές μέθοδοι. Τα αποτελέσματα από την εφαρμογή αυτών των μεθόδων που αναφέρθηκαν, συνοψίζονται στον Πίνακα 5.18.

5.1.2.1 Αθροισμα Απόλυτων Σφαλμάτων

Η εξέταση όλων των δυνατών συνδυασμών των ανεξάρτητων μεταβλητών ως προς το άθροισμα των απλών απόλυτων σφαλμάτων, με εφαρμογή του γ.π. (4.8) σε κάθε έναν από αυτούς, έδωσε τα αποτελέσματα του Πίνακα 5.19. Με βάση τα περιεχόμενα του πίνακα αυτού, θα περιγραφούν στη συνέχεια τα βήματα που

ακολουθούνται στις προτεινόμενες μεθοδολογίες βηματικής ποιοτικής παλινδρόμησης.

Πίνακας 5.18: Αποτελέσματα για την εφαρμογή της βηματικής παλινδρόμησης, με τη χρήση του πακέτου SPSS v.12, στα δεδομένα Solar Thermal Energy.

Συνδυασμός	Βηματική Παλινδρόμηση			Εμπρός Επιλογή			Πίσω Απαλοιφή		
	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης
x1-x2-x3-x4-x5	0.999	0.998	9.874	0.999	0.998	9.874	0.999	0.998	9.874
x1-x2-x3-x4	0.999	0.998	9.874	0.999	0.998	9.874	0.999	0.998	9.874
x1-x2-x3-x5	0.995	0.995	17.433	0.995	0.995	17.433	0.996	0.995	16.868
x1-x2-x4-x5	0.997	0.997	13.946	0.997	0.997	13.946	0.997	0.997	13.946
x1-x3-x4-x5	0.996	0.996	16.681	0.996	0.996	16.681	0.996	0.996	16.681
x2-x3-x4-x5	0.998	0.998	11.443	0.998	0.998	11.443	0.998	0.998	11.443
x1-x2-x3	0.995	0.995	17.433	0.995	0.995	17.433	0.995	0.995	17.433
x1-x2-x4	0.997	0.996	15.657	0.997	0.996	15.657	0.997	0.996	15.657
x1-x2-x5	0.995	0.995	17.433	0.995	0.995	17.433	0.995	0.995	17.433
x1-x3-x4	0.996	0.996	16.681	0.996	0.996	16.681	0.996	0.996	16.681
x1-x3-x5	0.996	0.995	17.541	0.996	0.995	17.541	0.996	0.995	17.541
x1-x4-x5	0.994	0.993	20.351	0.994	0.993	20.351	0.994	0.993	20.351
x2-x3-x4	0.998	0.997	13.025	0.998	0.997	13.025	0.998	0.997	13.025
x2-x3-x5	0.995	0.995	17.971	0.995	0.995	17.971	0.995	0.995	17.971
x2-x4-x5	0.993	0.992	22.488	0.993	0.992	22.488	0.993	0.992	22.488
x3-x4-x5	0.995	0.995	17.825	0.995	0.995	17.825	0.995	0.995	17.825
x1-x2	0.995	0.995	17.433	0.995	0.995	17.433	0.995	0.995	17.433
x1-x3	0.995	0.995	18.291	0.995	0.995	18.291	0.995	0.995	18.291
x1-x4	0.993	0.993	21.097	0.993	0.993	21.097	0.993	0.993	21.097
x1-x5	0.994	0.993	20.351	0.994	0.993	20.351	0.994	0.993	20.351
x2-x3	0.992	0.992	23.040	0.992	0.992	23.040	0.992	0.992	23.040
x2-x4	0.993	0.992	22.488	0.993	0.992	22.488	0.993	0.992	22.488
x2-x5	0.991	0.991	24.054	0.991	0.991	24.054	0.991	0.991	24.054
x3-x4	0.992	0.992	22.768	0.992	0.992	22.768	0.992	0.992	22.768
x3-x5	0.995	0.994	18.851	0.995	0.994	18.851	0.995	0.994	18.851
x4-x5	0.981	0.980	35.117	0.981	0.980	35.117	0.981	0.980	35.117
x1	0.992	0.992	22.764	0.992	0.992	22.764	0.992	0.992	22.764
x2	0.991	0.991	24.054	0.991	0.991	24.054	0.991	0.991	24.054
x3	0.990	0.990	25.035	0.990	0.990	25.035	0.990	0.990	25.035
x4	0.981	0.980	35.117	0.981	0.980	35.117	0.981	0.980	35.117
x5	0.963	0.961	49.238	0.963	0.961	49.238	0.963	0.961	49.238

Πίνακας 5.19: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Solar Thermal Energy.

Συνδυασμός	R^2	Sum(e)	Sum(σ^2)	Sum(e)	Sum(e^2)	MAPE
x1-x2-x3-x4-x5	0.891	1.258	0.163	132.462	1687.465	0.019
x1-x2-x3-x4	0.892	1.254	0.162	132.262	1681.557	0.019
x1-x2-x3-x5	0.808	1.787	0.364	196.578	2846.848	0.028
x1-x2-x4-x5	0.739	1.478	0.257	238.918	4008.825	0.034
x1-x3-x4-x5	0.886	1.279	0.160	137.600	1677.277	0.020
x2-x3-x4-x5	0.885	1.573	0.197	152.781	1997.432	0.023
x1-x2-x3	0.634	2.188	0.755	263.800	5814.614	0.038
x1-x2-x4	0.760	1.466	0.256	225.025	3646.155	0.032
x1-x2-x5	0.670	1.862	0.325	286.236	5451.192	0.041
x1-x3-x4	0.887	1.276	0.160	137.464	1667.465	0.020
x1-x3-x5	0.717	1.855	0.369	271.450	4940.639	0.038
x1-x4-x5	0.725	1.480	0.262	261.253	4538.000	0.037
x2-x3-x4	0.880	1.623	0.203	157.157	2076.676	0.024
x2-x3-x5	0.850	1.999	0.335	184.045	2767.507	0.028
x2-x4-x5	0.458	2.300	0.426	325.873	11663.823	0.045
x3-x4-x5	0.884	1.570	0.196	153.005	2001.816	0.023
x1-x2	0.561	2.190	0.753	329.621	7240.255	0.046
x1-x3	0.599	2.207	0.798	312.256	6548.625	0.044
x1-x4	0.764	1.468	0.261	238.611	3743.374	0.034
x1-x5	0.539	1.865	0.377	379.682	8885.221	0.053
x2-x3	0.020	3.589	1.043	453.846	18841.746	0.070
x2-x4	0.734	2.290	0.419	254.383	4700.115	0.036
x2-x5	0.373	3.175	0.970	521.139	13694.536	0.073
x3-x4	0.879	1.620	0.202	157.040	2079.896	0.024
x3-x5	0.802	2.034	0.413	208.317	3266.073	0.031
x4-x5	0.591	2.326	0.642	362.847	8119.976	0.051
x1	0.528	2.207	0.801	393.039	8542.914	0.056
x2	0.015	5.425	3.464	674.346	25714.871	0.097
x3	0.051	4.125	2.034	510.388	20924.678	0.079
x4	0.668	2.314	0.643	310.236	6416.695	0.044
x5	0.385	3.289	1.166	530.136	16201.784	0.075

5.1.2.1.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Η πρώτη ανεξάρτητη μεταβλητή που θα εισαχθεί στο μοντέλο παλινδρόμησης, είναι η x_4 , καθώς παρουσιάζει το μικρότερο άθροισμα απλών απόλυτων σφαλμάτων. Στη συνέχεια εξετάζονται όλοι οι δυνατοί συνδυασμοί που περιέχουν την x_4 ως προς το άθροισμα απόλυτων σφαλμάτων, εφαρμόζοντας σε κάθε έναν από αυτούς το γ.π. (4.8). Η εξέταση αυτή δείχνει ότι ο συνδυασμός x_3 - x_4 εμφανίζει μικρότερο άθροισμα απόλυτων σφαλμάτων από το αντίστοιχο του προηγούμενου βήματος, οπότε επιλέγεται ως ο καλύτερος για το βήμα αυτό και η μεταβλητή x_3 εισάγεται στο

μοντέλο. Στο επόμενο βήμα πρέπει να εξεταστούν οι δυνατοί συνδυασμοί που περιέχουν τις x_3 - x_4 και να εξεταστούν ως προς το άθροισμα απόλυτων σφαλμάτων, με εφαρμογή του γ.π. (4.8) σε αυτούς. Προκύπτει με τον τρόπο αυτό ότι η x_1 πρέπει να εισαχθεί στο μοντέλο, καθώς ο συνδυασμός x_1 - x_3 - x_4 δίνει μικρότερο άθροισμα απόλυτων σφαλμάτων από το αντίστοιχο του προηγούμενου βήματος. Ακολούθως, εξετάζονται με παρόμοιο τρόπο οι συνδυασμοί που περιέχουν τις x_1 - x_3 - x_4 και προκύπτει ότι καλύτερος είναι αυτός που περιέχει τις x_1 - x_2 - x_3 - x_4 . Η διαδικασία σταματά εδώ, καθώς δεν υπάρχει άλλος συνδυασμός που να περιέχει αυτές τις μεταβλητές και να δίνει μικρότερο άθροισμα απλών απόλυτων σφαλμάτων, με βάση το γ.π. (4.8), οπότε ως καλύτερο υποσύνολο προτείνεται το x_1 - x_2 - x_3 - x_4 . Τα παραπάνω βήματα, συνοψίζονται στον Πίνακα 5.20.

Πίνακας 5.20: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα απλών απόλυτων σφαλμάτων, στα δεδομένα Solar Thermal Energy.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	x_1, x_2, x_3, x_4, x_5	x_4	2.314	x_4
2	x_1 - x_4 , x_2 - x_4 , x_3 - x_4 , x_4 - x_5	x_3 - x_4	1.620	x_3
3	x_1 - x_3 - x_4 , x_2 - x_3 - x_4 , x_3 - x_4 - x_5	x_1 - x_3 - x_4	1.276	x_1
4	x_1 - x_2 - x_3 - x_4 , x_1 - x_3 - x_4 - x_5	x_1 - x_2 - x_3 - x_4	1.254	x_2
5	x_1 - x_2 - x_3 - x_4 - x_5	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_1 - x_2 - x_3 - x_4				

5.1.2.1.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Με βάση τα αποτελέσματα του Πίνακα 5.19, στο πρώτο βήμα της διαδικασίας θα πρέπει να εξαχθεί από το μοντέλο η μεταβλητή x_5 , καθώς αφαιρούμενη δίνει το συνδυασμό (x_1 - x_2 - x_3 - x_4) με το μικρότερο άθροισμα απόλυτων σφαλμάτων από όλους τους υπόλοιπους εξεταζόμενους και, ταυτόχρονα, το άθροισμα αυτό είναι μικρότερο από αυτό που δίνει ο αρχικός συνδυασμός (με όλες τις ανεξάρτητες μεταβλητές να μετέχουν στο μοντέλο). Η διαδικασία τερματίζεται εδώ, καθώς δεν προκύπτει άλλος συνδυασμός στη συνέχεια που να δίνει μικρότερο άθροισμα απόλυτων σφαλμάτων. Όλα αυτά φαίνονται στον Πίνακα 5.21.

5.1.2.2 Άθροισμα Τετραγωνικών Σφαλμάτων

Στην παράγραφο αυτή εξετάζονται τα αποτελέσματα που δίνουν οι προτεινόμενες μεθοδολογίες βηματικής ποιοτικής παλινδρόμησης, με βάση τα αποτελέσματα του Πίνακα 5.22, όπου και συνοψίζονται τα αποτελέσματα για την εφαρμογή του γ.π. (4.12) σε όλους τους πιθανούς συνδυασμούς μεταβλητών που φαίνονται σε αυτόν.

Πίνακας 5.21: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα απλών απόλυτων σφαλμάτων, στα δεδομένα Solar Thermal Energy.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3-x4, x1-x2-x3-x5, x1-x2-x4-x5, x1-x3-x4-x5, x2-x3-x4-x5	x1-x2-x3-x4	1.254	x5
2	x1-x2-x3, x1-x2-x4, x1-x3-x4, x2-x3-x4	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x3-x4				

Πίνακας 5.22: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.12), στα δεδομένα Solar Thermal Energy.

Συνδυασμός	R^2	Sum(ϵ)	Sum(σ^2)	Sum(ϵ)	Sum(ϵ^2)	MAPE
x1-x2-x3-x4-x5	0.892	1.468	0.127	164.086	1597.691	0.023
x1-x2-x3-x4	0.893	1.466	0.126	163.459	1583.057	0.023
x1-x2-x3-x5	0.861	1.877	0.226	197.482	2060.225	0.028
x1-x2-x4-x5	0.822	1.685	0.169	201.827	2717.992	0.029
x1-x3-x4-x5	0.886	1.466	0.128	164.940	1683.195	0.024
x2-x3-x4-x5	0.874	1.788	0.170	175.754	1882.329	0.026
x1-x2-x3	0.574	2.582	0.433	347.350	6603.808	0.050
x1-x2-x4	0.829	1.647	0.168	197.797	2601.118	0.028
x1-x2-x5	0.693	2.074	0.283	288.352	4819.453	0.041
x1-x3-x4	0.887	1.464	0.127	164.348	1669.845	0.023
x1-x3-x5	0.801	2.047	0.247	239.429	2957.087	0.034
x1-x4-x5	0.783	1.809	0.197	247.386	3434.038	0.035
x2-x3-x4	0.872	1.794	0.170	177.654	1904.043	0.026
x2-x3-x5	0.841	2.157	0.275	212.465	2419.484	0.031
x2-x4-x5	0.754	2.411	0.323	250.329	4095.702	0.036
x3-x4-x5	0.866	1.765	0.170	177.356	1987.451	0.026
x1-x2	0.554	2.752	0.480	363.796	7232.639	0.052
x1-x3	0.452	2.796	0.575	415.722	9655.634	0.058
x1-x4	0.790	1.800	0.196	242.977	3315.672	0.034
x1-x5	0.619	2.144	0.357	348.006	7147.894	0.049
x2-x3	0.010	3.694	0.995	522.054	20148.705	0.079
x2-x4	0.758	2.403	0.321	248.571	4023.589	0.035
x2-x5	0.575	3.293	0.612	363.585	7394.499	0.052
x3-x4	0.863	1.782	0.170	181.353	2031.377	0.027
x3-x5	0.794	2.295	0.307	236.227	3042.625	0.034
x4-x5	0.723	2.610	0.393	276.589	4947.767	0.039
x1	0.428	2.893	0.682	460.658	10771.332	0.063
x2	0.018	5.504	1.867	653.031	22994.981	0.094
x3	0.039	4.741	1.374	604.150	25462.796	0.092
x4	0.727	2.602	0.390	274.500	4877.180	0.039
x5	0.517	3.613	1.050	563.151	16317.818	0.079

5.1.2.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Η εξέταση κάθε μιας ανεξάρτητης μεταβλητής ως προς το άθροισμα των τετραγωνικών σφαλμάτων, δείχνει ότι η μεταβλητή x_4 θα εισέλθει πρώτη στο μοντέλο παλινδρόμησης. Με την εφαρμογή του γ.π. (4.12) στους δυνατούς συνδυασμούς που περιέχουν την x_4 , προκύπτει ότι ο συνδυασμός x_3 - x_4 είναι ο καλύτερος για αυτό το βήμα, οπότε εισέρχεται στο μοντέλο η x_3 . Ακολούθως, με την σύγκριση των συνδυασμών που περιέχουν τις x_3 - x_4 , ως προς το άθροισμα των τετραγωνικών σφαλμάτων που παρουσιάζουν, φαίνεται ότι προτιμάται ο συνδυασμός x_1 - x_3 - x_4 . Συνεπώς, στο μοντέλο εισάγεται η x_1 . Με παρόμοιο τρόπο, από τους συνδυασμούς που περιέχουν τις x_1 - x_3 - x_4 , επιλέγεται ο x_1 - x_2 - x_3 - x_4 , οπότε εισέρχεται στο μοντέλο η x_2 . Η διαδικασία τερματίζεται στο σημείο αυτό, καθώς δεν υπάρχει συνδυασμός που να περιέχει τις x_1 - x_2 - x_3 - x_4 και ταυτόχρονα να δίνει άθροισμα τετραγωνικών σφαλμάτων μικρότερο από αυτό του προηγούμενου βήματος. Τα παραπάνω βήματα, παρουσιάζονται στον Πίνακα 5.23.

Πίνακας 5.23: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα των τετραγωνικών σφαλμάτων, στα δεδομένα Solar Thermal Energy.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	x_1, x_2, x_3, x_4, x_5	x_4	0.390	x_4
2	x_1 - x_4 , x_2 - x_4 , x_3 - x_4 , x_4 - x_5	x_3 - x_4	0.170	x_3
3	x_1 - x_3 - x_4 , x_2 - x_3 - x_4 , x_3 - x_4 - x_5	x_1 - x_3 - x_4	0.127	x_1
4	x_1 - x_2 - x_3 - x_4 , x_1 - x_3 - x_4 - x_5	x_1 - x_2 - x_3 - x_4	0.126	x_2
5	x_1 - x_2 - x_3 - x_4 - x_5	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_1 - x_2 - x_3 - x_4				

5.1.2.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Λαμβάνοντας υπόψη τα δεδομένα του Πίνακα 5.22, η μεταβλητή x_5 πρέπει να εξέλθει από το μοντέλο παλινδρόμησης, καθώς με τον τρόπο αυτό προκύπτει ο συνδυασμός x_1 - x_2 - x_3 - x_4 , ο οποίος δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων τόσο σε σχέση με τον αρχικό συνδυασμό, όσο και σε σχέση με τους υπόλοιπους συνδυασμούς που εξετάζονται στο βήμα αυτό.

Η εξέταση των συνδυασμών του επόμενου βήματος, δείχνει ότι δεν υπάρχει άλλος συνδυασμός που να δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων, οπότε η διαδικασία σταματά εδώ. Ο Πίνακας 5.24 συνοψίζει την όλη αυτή διαδικασία.

Πίνακας 5.24: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα τετραγωνικών σφαλμάτων, στα δεδομένα Solar Thermal Energy.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3-x4, x1-x2-x3-x5, x1-x2-x4-x5, x1-x3-x4-x5, x2-x3-x4-x5	x1-x2-x3-x4	0.126	x5
2	x1-x2-x3, x1-x2-x4, x1-x3-x4, x2-x3-x4	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x3-x4				

5.1.2.3 Συντελεστής προσδιορισμού

Κριτήριο για την επιλογή του καλύτερου συνδυασμού στην περίπτωση αυτή, αποτελεί η αντικειμενική συνάρτηση του γ.π. (4.17). Και εδώ επιλέγεται να ελαχιστοποιηθεί η έκφραση $\sum_{i=1}^n (y_i - \hat{y}_i)^2$, η οποία αντιστοιχεί στο άθροισμα των τετραγώνων των πρωτογενών σφαλμάτων e_i . Η ελαχιστοποίηση της έκφρασης αυτής, αντιστοιχεί στη μεγιστοποίηση του R^2 , όπως προβλέπει το γ.π. (4.17). Η εξέταση των συνδυασμών για το εν λόγω σύνολο δεδομένων, με βάση το γ.π. (4.17), δίνει τα αποτελέσματα του Πίνακα 5.25.

5.1.2.3.1 Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση

Τα περιεχόμενα του Πίνακα 5.25, φανερώνουν ότι υπάρχει συνδυασμός, με μια μεταβλητή λιγότερη από του αρχικού (στον οποίο όλες οι μεταβλητές συμμετέχουν), ο οποίος δίνει τιμή της εξεταζόμενης αντικειμενικής συνάρτησης μικρότερη από την αντίστοιχη του αρχικού συνδυασμού. Κάτι τέτοιο σημαίνει ότι ο συνδυασμός αυτός δίνει και καλύτερο R^2 από τον αρχικό. Ο συνδυασμός αυτός προκύπτει με αφαίρεση της μεταβλητής x5 από το αρχικό μοντέλο. Εφαρμόζεται στη συνέχεια το γ.π (4.17) για όλους τους δυνατούς συνδυασμούς οι οποίοι προκύπτουν αφαιρώντας ακόμη μία μεταβλητή, προκειμένου να εξεταστεί εάν μπορεί να προκύψει συνδυασμός με καλύτερο R^2 από αυτό του προηγούμενου βήματος. Η εξέταση αυτή δείχνει ότι δεν προκύπτει συνδυασμός με καλύτερη τιμή για την υπό εξέταση αντικειμενική συνάρτηση, οπότε η διαδικασία σταματά εδώ, με καλύτερο υποσύνολο μεταβλητών να υποδεικνύεται το x1-x2-x3-x4. Τα βήματα τα οποία ακολουθήθηκαν για τη διαδικασία αυτή, συγκεντρώνονται στον Πίνακα 5.26.

Πίνακας 5.25: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.17), στα δεδομένα Solar Thermal Energy.

Εξεταζόμενος Συνδυασμός	Τιμή Αντικειμενικής Συνάρτησης $(\sum_{i=1}^n (\hat{y}_i - y_i)^2)$	$\sum_{i=1}^n (y_i - \bar{Y})^2$	Υπολογιζόμενος R^2
x1-x2-x3-x4-x5	1318.63	14718.51	91.041%
x1-x2-x3-x4	1119.62		92.393%
x1-x2-x3-x5	2295.57		84.404%
x1-x2-x4-x5	2339.3		84.106%
x1-x3-x4-x5	1436.45		90.241%
x2-x3-x4-x5	1844.52		87.468%
x1-x2-x3	5300.46		63.988%
x1-x2-x4	2318.49		84.248%
x1-x2-x5	3966.84		73.049%
x1-x3-x4	1376.07		90.651%
x1-x3-x5	2716.94		81.541%
x1-x4-x5	2867.97		80.515%
x2-x3-x4	1777.71		87.922%
x2-x3-x5	1870.11		87.294%
x2-x4-x5	3169.48		78.466%
x3-x4-x5	1985.83		86.508%
x1-x2	6013.95		59.140%
x1-x3	5604.56		61.922%
x1-x4	3112.73		78.852%
x1-x5	4893.56		66.752%
x2-x3	13426.02		8.781%
x2-x4	3211.65		78.180%
x2-x5	7040.7		52.164%
x3-x4	1876.37		87.252%
x3-x5	3041.47		79.336%
x4-x5	4535.8		69.183%

Πίνακας 5.26: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για τον συντελεστή πολλαπλής παλινδρόμησης, στα δεδομένα Solar Thermal Energy.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	R^2 Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3-x4, x1-x2-x3-x5, x1-x2-x4-x5, x1-x3-x4-x5, x2-x3-x4-x5	x1-x2-x3-x4	91.187%	x5
2	x1-x2-x3, x1-x2-x4, x1-x3-x4, x2-x3-x4	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x3-x4				

5.1.2.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα

Στην παράγραφο αυτή, θα πραγματοποιηθεί η σύγκριση των αποτελεσμάτων που δίδουν οι προτεινόμενες μεθοδολογίες βηματικής ποιοτικής παλινδρόμησης. Ο Πίνακας 5.27 συγκεντρώνει τα αποτελέσματα των συγκρινόμενων μεθοδολογιών.

Πίνακας 5.27: Αποτελέσματα των χρησιμοποιούμενων προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Solar Thermal Energy.

Εξεταζόμενος Συνδυασμός	Βηματική Παλινδρόμηση		Βηματική Ποιοτική Παλινδρόμηση			
	Εμπρός Επιλογή	Πίσω Απαλοιφή	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
	R^2	R^2	$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4-x5	0.999	0.999	1.258	0.891	0.127	0.892
x1-x2-x3-x4	0.999	0.999	1.254	0.892	0.126	0.893
x1-x2-x3-x5	0.995	0.996	1.787	0.808	0.226	0.861
x1-x2-x4-x5	0.997	0.997	1.478	0.739	0.169	0.822
x1-x3-x4-x5	0.996	0.996	1.279	0.886	0.128	0.886
x2-x3-x4-x5	0.998	0.998	1.573	0.885	0.170	0.874
x1-x2-x3	0.995	0.995	2.188	0.634	0.433	0.574
x1-x2-x4	0.997	0.997	1.466	0.760	0.168	0.829
x1-x2-x5	0.995	0.995	1.862	0.670	0.283	0.693
x1-x3-x4	0.996	0.996	1.276	0.887	0.127	0.887
x1-x3-x5	0.996	0.996	1.855	0.717	0.247	0.801
x1-x4-x5	0.994	0.994	1.480	0.725	0.197	0.783
x2-x3-x4	0.998	0.998	1.623	0.880	0.170	0.872
x2-x3-x5	0.995	0.995	1.999	0.850	0.275	0.841
x2-x4-x5	0.993	0.993	2.300	0.458	0.323	0.754
x3-x4-x5	0.995	0.995	1.570	0.884	0.170	0.866
x1-x2	0.995	0.995	2.190	0.561	0.480	0.554
x1-x3	0.995	0.995	2.207	0.599	0.575	0.452
x1-x4	0.993	0.993	1.468	0.764	0.196	0.790
x1-x5	0.994	0.994	1.865	0.539	0.357	0.619
x2-x3	0.992	0.992	3.589	0.020	0.995	0.010
x2-x4	0.993	0.993	2.290	0.734	0.321	0.758
x2-x5	0.991	0.991	3.175	0.373	0.612	0.575
x3-x4	0.992	0.992	1.620	0.879	0.170	0.863
x3-x5	0.995	0.995	2.034	0.802	0.307	0.794
x4-x5	0.981	0.981	2.326	0.591	0.393	0.723
x1	0.992	0.992	2.207	0.528	0.682	0.428
x2	0.991	0.991	5.425	0.015	1.867	0.018
x3	0.990	0.990	4.125	0.051	1.374	0.039
x4	0.981	0.981	2.314	0.668	0.390	0.727
x5	0.963	0.963	3.289	0.385	1.050	0.517

Σε γενικές γραμμές, μπορεί να διατυπωθεί ο ισχυρισμός ότι ισχύουν οι παρατηρήσεις που έγιναν για τα αποτελέσματα στο προηγούμενο σετ δεδομένων. Ότι δηλαδή για κάθε εξεταζόμενο συνδυασμό υπάρχουν μεγάλες αποκλίσεις μεταξύ του R^2 των κλασσικών μεθόδων βηματικής παλινδρόμησης σε σχέση με τα αντίστοιχα για την προτεινόμενη μεθοδολογία βηματικής ποιοτικής παλινδρόμησης. Ακόμα, οι αποκλίσεις στα αποτελέσματα αυτά αμβλύνονται, εάν συγκριθούν μεταξύ τους τα αποτελέσματα που δίνει η χρησιμοποίηση των δύο διαφορετικών κριτηρίων (άθροισμα απλών και τετραγωνικών σφαλμάτων) στην ίδια (προτεινόμενη) μεθοδολογία. Παράλληλα, ισχύει και εδώ η παρατήρηση ότι συνδυασμοί που δίδουν μικρότερο άθροισμα σφαλμάτων, εμφανίζουν μεγαλύτερο συντελεστή πολλαπλής παλινδρόμησης. Δεν ισχύει, ωστόσο, πάντα η αρχή της αναλογικότητας, δηλαδή δεν μπορεί να ισχυριστεί κανείς ότι μεταξύ δύο συγκεκριμένων συνδυασμών αυτός που παρουσιάζει το μικρότερο άθροισμα απλών ή τετραγωνικών σφαλμάτων θα παρουσιάζει και τον μεγαλύτερο συντελεστή πολλαπλής παλινδρόμησης. Εφαρμόζοντας τις διαδικασίες που περιγράφονται στις παραγράφους 4.3.1.1 και 4.3.1.2, προκύπτει ότι, με βάση τα δύο αυτά κριτήρια, καταλήγουμε στο ίδιο τελικό προτεινόμενο καλύτερο υποσύνολο μεταβλητών. Όλα δηλαδή τα κριτήρια για την προτεινόμενη μεθοδολογία, καταλήγουν στο τελικό υποσύνολο μεταβλητών.

Σε ότι αφορά τις παρατηρήσεις οι οποίες μπορούν να γίνουν συγκρίνοντας τα αποτελέσματα για όλα τα προτεινόμενα κριτήρια, αυτές μπορούν να προκύψουν εξετάζοντας τα περιεχόμενα του Πίνακα 5.28. Ο τελευταίος πίνακας συνοψίζει τα αποτελέσματα για τα τρία κριτήρια που εξετάζονται στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής ποιοτικής παλινδρόμησης.

Αυτό που μπορεί να ισχυριστεί κανείς, είναι ότι τα αποτελέσματα για τη χρησιμοποίηση του συντελεστή πολλαπλής παλινδρόμησης ως κριτήριο βελτιστοποίησης στα πλαίσια της προτεινόμενης μεθοδολογίας, οδηγεί, σε γενικές γραμμές, σε καλύτερα ή παραπλήσια αποτελέσματα για κάθε συνδυασμό από τα αντίστοιχα κριτήρια του αθροίσματος των απόλυτων σφαλμάτων. Ακόμα, όμως, μια σημαντική παρατήρηση είναι ότι ακολουθώντας τις διαδικασίες που περιγράφονται στις ενότητες 4.3.1 και 4.3.2, και τα τρία κριτήρια υποδεικνύουν το ίδιο υποσύνολο ως το καλύτερο για συμμετοχή στο ζητούμενο μοντέλο παλινδρόμησης.

Στο σημείο αυτό θα εξεταστούν τα αποτελέσματα που προκύπτουν για τις επιδόσεις των επιμέρους δεικτών που υιοθετήθηκαν κατά τη διάρκεια των διερευνήσεων. Ο Πίνακας 5.28 συνοψίζει τα αποτελέσματα για τους δείκτες αυτούς για κάθε κριτήριο και κάθε συνδυασμό.

Και σε αυτή την περίπτωση, φαίνεται να ισχύει η παρατήρηση που έγινε για το σύνολο δεδομένων που εξετάστηκε στην προηγούμενη ενότητα. Δηλαδή, και εδώ παρουσιάζονται μερικές αποκλίσεις μεταξύ των τιμών των δεικτών για κάθε συνδυασμό ανάλογα με το κριτήριο βελτιστοποίησης. Μόνο που για το συγκεκριμένο σύνολο δεδομένων, φαίνεται ότι, σε γενικές γραμμές, έχουμε καλύτερα αποτελέσματα για τον R^2 και τους υπόλοιπους δείκτες στην περίπτωση ελαχιστοποίησης του αθροίσματος των τετραγωνικών σφαλμάτων. Ακόμα, μπορεί να γίνει η παρατήρηση ότι με βάση την τιμή της εκάστοτε χρησιμοποιούμενης αντικειμενικής συνάρτησης, όλα τα επιμέρους κριτήρια αξιολόγησης που χρησιμοποιούνται, παρουσιάζουν τιμές τέτοιες, ώστε να είναι σε αντιστοιχία με την τιμή της αντικειμενικής συνάρτησης. Δηλαδή, μικρές τιμές του αθροίσματος για τα

απλά και τα τετραγωνικά σφάλματα αξιών, οδηγούν αντίστοιχα σε μικρές τιμές για το άθροισμα των απλών και τετραγωνικών πρωτογενών σφαλμάτων, ενώ οδηγούν σε μεγάλες τιμές για τον R^2 .

Πίνακας 5.28: Αποτελέσματα των χρησιμοποιούμενων τριών κριτηρίων, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Solar Thermal Energy.

Βηματική Ποιοτική Παλινδρόμηση					
Εξεταζόμενος Συνδυασμός	R^2	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
		$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4-x5	91.041%	1.258	0.891	0.127	0.892
x1-x2-x3-x4	92.393%	1.254	0.892	0.126	0.893
x1-x2-x3-x5	84.404%	1.787	0.808	0.226	0.861
x1-x2-x4-x5	84.106%	1.478	0.739	0.169	0.822
x1-x3-x4-x5	90.241%	1.279	0.886	0.128	0.886
x2-x3-x4-x5	87.468%	1.573	0.885	0.170	0.874
x1-x2-x3	63.988%	2.188	0.634	0.433	0.574
x1-x2-x4	84.248%	1.466	0.760	0.168	0.829
x1-x2-x5	73.049%	1.862	0.670	0.283	0.693
x1-x3-x4	90.651%	1.276	0.887	0.127	0.887
x1-x3-x5	81.541%	1.855	0.717	0.247	0.801
x1-x4-x5	80.515%	1.480	0.725	0.197	0.783
x2-x3-x4	87.922%	1.623	0.880	0.170	0.872
x2-x3-x5	87.294%	1.999	0.850	0.275	0.841
x2-x4-x5	78.466%	2.300	0.458	0.323	0.754
x3-x4-x5	86.508%	1.570	0.884	0.170	0.866
x1-x2	59.140%	2.190	0.561	0.480	0.554
x1-x3	61.922%	2.207	0.599	0.575	0.452
x1-x4	78.852%	1.468	0.764	0.196	0.790
x1-x5	66.752%	1.865	0.539	0.357	0.619
x2-x3	8.781%	3.589	0.020	0.995	0.010
x2-x4	78.180%	2.290	0.734	0.321	0.758
x2-x5	52.164%	3.175	0.373	0.612	0.575
x3-x4	87.252%	1.620	0.879	0.170	0.863
x3-x5	79.336%	2.034	0.802	0.307	0.794
x4-x5	69.183%	2.326	0.591	0.393	0.723
x1	51.338%	2.207	0.528	0.682	0.428
x2	-18.425%	5.425	0.015	1.867	0.018
x3	-25.876%	4.125	0.051	1.374	0.039
x4	69.374%	2.314	0.668	0.390	0.727
x5	24.725%	3.289	0.385	1.050	0.517

Πίνακας 5.29: Αποτελέσματα των επιμέρους χρησιμοποιούμενων δεικτών, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Solar Thermal Energy.

Εξεταζόμενος Συνδυασμός	Απλά Απόλυτα Σφάλματα				Τετραγωνικά Σφάλματα			
	R^2	Sum(l _{el})	Sum(e ²)	MAPE	R^2	Sum(l _{el})	Sum(e ²)	MAPE
x1-x2-x3-x4-x5	0.891	132.462	1687.465	0.019	0.892	164.086	1597.691	0.023
x1-x2-x3-x4	0.892	132.262	1681.557	0.019	0.893	163.459	1583.057	0.023
x1-x2-x3-x5	0.808	196.578	2846.848	0.028	0.861	197.482	2060.225	0.028
x1-x2-x4-x5	0.739	238.918	4008.825	0.034	0.822	201.827	2717.992	0.029
x1-x3-x4-x5	0.886	137.600	1677.277	0.020	0.886	164.940	1683.195	0.024
x2-x3-x4-x5	0.885	152.781	1997.432	0.023	0.874	175.754	1882.329	0.026
x1-x2-x3	0.634	263.800	5814.614	0.038	0.574	347.350	6603.808	0.050
x1-x2-x4	0.760	225.025	3646.155	0.032	0.829	197.797	2601.118	0.028
x1-x2-x5	0.670	286.236	5451.192	0.041	0.693	288.352	4819.453	0.041
x1-x3-x4	0.887	137.464	1667.465	0.020	0.887	164.348	1669.845	0.023
x1-x3-x5	0.717	271.450	4940.639	0.038	0.801	239.429	2957.087	0.034
x1-x4-x5	0.725	261.253	4538.000	0.037	0.783	247.386	3434.038	0.035
x2-x3-x4	0.880	157.157	2076.676	0.024	0.872	177.654	1904.043	0.026
x2-x3-x5	0.850	184.045	2767.507	0.028	0.841	212.465	2419.484	0.031
x2-x4-x5	0.458	325.873	11663.823	0.045	0.754	250.329	4095.702	0.036
x3-x4-x5	0.884	153.005	2001.816	0.023	0.866	177.356	1987.451	0.026
x1-x2	0.561	329.621	7240.255	0.046	0.554	363.796	7232.639	0.052
x1-x3	0.599	312.256	6548.625	0.044	0.452	415.722	9655.634	0.058
x1-x4	0.764	238.611	3743.374	0.034	0.790	242.977	3315.672	0.034
x1-x5	0.539	379.682	8885.221	0.053	0.619	348.006	7147.894	0.049
x2-x3	0.020	453.846	18841.746	0.070	0.010	522.054	20148.705	0.079
x2-x4	0.734	254.383	4700.115	0.036	0.758	248.571	4023.589	0.035
x2-x5	0.373	521.139	13694.536	0.073	0.575	363.585	7394.499	0.052
x3-x4	0.879	157.040	2079.896	0.024	0.863	181.353	2031.377	0.027
x3-x5	0.802	208.317	3266.073	0.031	0.794	236.227	3042.625	0.034
x4-x5	0.591	362.847	8119.976	0.051	0.723	276.589	4947.767	0.039
x1	0.528	393.039	8542.914	0.056	0.428	460.658	10771.332	0.063
x2	0.015	674.346	25714.871	0.097	0.018	653.031	22994.981	0.094
x3	0.051	510.388	20924.678	0.079	0.039	604.150	25462.796	0.092
x4	0.668	310.236	6416.695	0.044	0.727	274.500	4877.180	0.039
x5	0.385	530.136	16201.784	0.075	0.517	563.151	16317.818	0.079

5.2 Τα Δεδομένα Ελέγχου

Στην περίπτωση των δεδομένων διερεύνησης, εξετάστηκαν εξαντλητικά όλοι οι δυνατοί συνδυασμοί των διαθέσιμων για κάθε σύνολο δεδομένων ανεξάρτητων μεταβλητών για να ακολουθηθούν τα βήματα των προτεινόμενων προσεγγίσεων της ποιοτικής βηματικής παλινδρόμησης (βλ. ενότητα 5.1). Στα δεδομένα ελέγχου που θα

εξεταστούν στη συνέχεια, η διερεύνηση θα πραγματοποιηθεί όχι εξαντλητικά, αλλά με βάση τα αποτελέσματα σε κάθε βήμα εφαρμογής των προτεινόμενων προσεγγίσεων που εξετάζονται. Τα δεδομένα ελέγχου αποτελούνται από τρία σύνολα δεδομένων, τα οποία παρουσιάζονται στη συνέχεια, και για κάθε ένα από αυτά περιγράφεται η διαδικασία εφαρμογής των προτεινόμενων προσεγγίσεων.

5.2.1 Πρακτική Εφαρμογή 3: Τα Δεδομένα Belle Ayr Liquefaction Runs

Το συγκεκριμένο σύνολο δεδομένων, μπορεί να βρεθεί στο (Montgomery and Peck, 1992) και περιλαμβάνει συνολικά οκτώ μεταβλητές (μαζί με την εξαρτημένη μεταβλητή). Η εξαρτημένη μεταβλητή y , η οποία στην προκειμένη περίπτωση αφορά την εκπομπή ποσότητας CO_2 κατά την διεργασία υγροποίησης ενός χημικού μίγματος τριών συστατικών, εκφράζεται συναρτήσει επτά μεταβλητών. Οι μεταβλητές αυτές είναι: απαιτούμενος χρόνος (x_1 , σε min), θερμοκρασία (x_2 , σε $^{\circ}\text{C}$), ποσοστό διάλυσης (x_3), κέρδος σε καύσιμο πετρέλαιο (x_4 , σε $\text{gR}/100 \text{ gR}$ μίγματος), συνολική ποσότητα άνθρακα (x_5), συνολικός διαλύτης (x_6), κατανάλωση υδρογόνου (x_7). Τα δεδομένα φαίνονται στον Πίνακα 5.30, ενώ οι συσχετίσεις μεταξύ όλων των μεταβλητών των δεδομένων αυτών, συγκεντρώνονται στον Πίνακα 5.31.

Πίνακας 5.30: Τα δεδομένα Belle Ayr Liquefaction Runs (Montgomery and Peck, 1992).

a/a	x_1	x_2	x_3	x_4	x_5	x_6	x_7	y
1	5.1	400	51.37	4.24	1484.83	2227.25	2.06	36.98
2	26.4	400	72.33	30.87	289.94	434.9	1.33	13.74
3	23.8	400	71.44	33.01	320.79	481.19	0.97	10.08
4	46.4	400	79.15	44.61	164.76	247.14	0.62	8.53
5	7	450	80.47	33.84	1097.26	1645.89	0.22	36.42
6	12.6	450	89.9	41.26	605.06	907.59	0.76	26.59
7	18.9	450	91.48	41.88	405.37	608.05	1.71	19.07
8	30.2	450	98.6	70.79	253.7	380.55	3.93	5.96
9	53.8	450	98.05	66.82	142.27	213.4	1.97	15.52
10	5.6	400	55.69	8.92	1362.24	2043.36	5.08	56.61
11	15.1	400	66.29	17.98	507.65	761.48	0.6	26.72
12	20.3	400	58.94	17.79	377.6	566.4	0.9	20.8
13	48.4	400	74.74	33.94	158.05	237.08	0.63	6.99
14	5.8	425	63.71	11.95	130.66	1961.49	2.04	45.93
15	11.2	425	67.14	14.73	682.59	1023.89	1.57	13.09
16	27.9	425	77.65	34.49	274.2	411.3	2.38	15.79
17	5.1	450	67.22	14.48	1496.51	2244.77	0.32	21.6
18	11.7	450	81.48	29.69	652.43	978.64	0.44	35.19
19	16.7	450	83.88	26.33	458.42	687.62	8.82	26.14
20	24.8	450	89.38	37.98	312.25	468.38	0.02	8.6
21	24.9	450	79.77	25.66	307.08	460.62	1.72	11.63
22	39.5	450	87.93	22.36	193.61	290.42	1.88	9.59
23	29	450	79.5	31.52	155.96	233.95	1.43	4.42
24	5.5	460	72.73	17.86	1392.08	2088.12	1.35	38.89
25	11.5	450	77.88	25.2	663.09	994.63	1.61	11.19
26	5.2	470	75.5	8.66	1464.11	2196.17	4.78	75.62
27	10.6	470	83.15	22.39	720.07	1080.11	5.88	36.03

Πίνακας 5.31: Συσχετίσεις μεταξύ των μεταβλητών των δεδομένων.

	X ₁	X ₂	X ₃	X ₄	X ₅	X ₆	X ₇	y
X ₁	1.000							
X ₂	-0.192	1.000						
X ₃	0.489	0.630	1.000					
X ₄	0.673	0.177	0.814	1.000				
X ₅	-0.736	0.166	-0.463	-0.577	1.000			
X ₆	-0.817	0.127	-0.559	-0.665	0.885	1.000		
X ₇	-0.178	0.255	0.073	-0.114	0.144	0.141	1.000	
y	-0.673	0.159	-0.385	-0.558	0.695	0.804	0.383	1.000

Πίνακας 5.32: Αποτελέσματα για την εφαρμογή της βηματικής παλινδρόμησης, με τη χρήση του πακέτου SPSS v.12, στα δεδομένα Belle Ayr Liquefaction Runs.

a/a	Βηματική Μέθοδος			
	Προς τα Εμπρός Επιλογή		Προς τα Πίσω Απαλοιφή	
	Συνδυασμός	R ²	Συνδυασμός	R ²
1	x6	0.870	x1-x2-x3-x4-x5-x6-x7	0.905
2	x6-x7	0.905	x1-x2-x3-x4-x6-x7	0.905
3	x1-x6-x7	0.905	x1-x2-x3-x5-x6-x7	0.905
4	x2-x6-x7	0.905	x1-x2-x4-x5-x6-x7	0.905
5	x3-x6-x7	0.905	x1-x3-x4-x5-x6-x7	0.905
6	x4-x6-x7	0.905	x2-x3-x4-x5-x6-x7	0.905
7	x5-x6-x7	0.905	x1-x2-x3-x6-x7	0.905
8	x1-x2-x6-x7	0.905	x1-x2-x4-x6-x7	0.905
9	x1-x3-x6-x7	0.905	x1-x2-x5-x6-x7	0.905
10	x1-x4-x6-x7	0.905	x1-x3-x4-x6-x7	0.905
11	x1-x5-x6-x7	0.905	x1-x3-x5-x6-x7	0.905
12	x2-x3-x6-x7	0.905	x1-x4-x5-x6-x7	0.905
13	x2-x4-x6-x7	0.905	x2-x3-x4-x6-x7	0.905
14	x2-x5-x6-x7	0.905	x2-x3-x5-x6-x7	0.905
15	x3-x4-x6-x7	0.905	x2-x4-x5-x6-x7	0.905
16	x3-x5-x6-x7	0.905	x3-x4-x5-x6-x7	0.905
17	x4-x5-x6-x7	0.905	x1-x2-x6-x7	0.905
18	x1-x2-x3-x6-x7	0.905	x1-x3-x6-x7	0.905
19	x1-x2-x4-x6-x7	0.905	x1-x4-x6-x7	0.905
20	x1-x2-x5-x6-x7	0.905	x1-x5-x6-x7	0.905
21	x1-x3-x4-x6-x7	0.905	x2-x3-x6-x7	0.905
22	x1-x3-x5-x6-x7	0.905	x2-x4-x6-x7	0.905
23	x1-x4-x5-x6-x7	0.905	x2-x5-x6-x7	0.905
24	x2-x3-x4-x6-x7	0.905	x3-x4-x6-x7	0.905
25	x2-x3-x5-x6-x7	0.905	x3-x5-x6-x7	0.905
26	x2-x4-x5-x6-x7	0.905	x4-x5-x6-x7	0.905
27	x3-x4-x5-x6-x7	0.905	x1-x6-x7	0.905
28	x1-x2-x3-x4-x6-x7	0.905	x2-x6-x7	0.905
29	x1-x2-x3-x5-x6-x7	0.905	x3-x6-x7	0.905
30	x1-x2-x4-x5-x6-x7	0.905	x4-x6-x7	0.905
31	x1-x3-x4-x5-x6-x7	0.905	x5-x6-x7	0.905
32	x2-x3-x4-x5-x6-x7	0.905	x6-x7	0.905
33	x1-x2-x3-x4-x5-x6-x7	0.905	x6	0.870

Στα δεδομένα του Πίνακα 5.30, εφαρμόστηκαν οι τρεις κλασσικές βηματικές μεθοδολογίες. Η εφαρμογή των μεθόδων αυτών, μέσω του πακέτου SPSS v.12, έδωσε

τα αποτελέσματα του Πίνακα Α.1, του Παραρτήματος. Στο σημείο αυτό παρατίθενται στον Πίνακα 5.32 τα βήματα τα οποία οδηγούν στο «καλύτερο» υποσύνολο μεταβλητών, βάση των αποτελεσμάτων που έδωσαν οι κλασσικές βηματικές μέθοδοι. Αυτό που παρατηρείται, είναι ότι οι κλασσικές βηματικές μεθοδολογίες, καταλήγουν από πολύ νωρίς στο μεγαλύτερο R^2 για το συγκεκριμένο σύνολο δεδομένων. Ακόμα, φαίνεται ότι από το δεύτερο βήμα των μεθοδολογιών αυτών, δεν έχει νόημα ούτε η προσθήκη, αλλά ούτε και η αφαίρεση κάποιας μεταβλητής, καθώς δεν υπάρχει κάποια αλλαγή στον R^2 , με αποτέλεσμα, εάν η διαδικασία σταματήσει στο δεύτερο βήμα των δύο εξεταζόμενων μεθόδων, αυτές καταλήγουν σε διαφορετικό «καλύτερο» υποσύνολο μεταβλητών.

5.2.1.1 Αθροισμα Απόλυτων Σφαλμάτων

Εδώ θα εξεταστούν οι προτεινόμενες προσεγγίσεις, με βάση τα αποτελέσματα που δίνει η εφαρμογή του γ.π. (4.8) σε κάθε βήμα τους, μέχρι η διαδικασία να καταλήξει στο προτεινόμενο ως «καλύτερο» υποσύνολο μεταβλητών. Αρχικά εξετάζεται η προς τα εμπρός βηματική ποιοτική παλινδρόμηση (βλ. Πίνακα 5.33).

Πίνακας 5.33: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Belle Ayr Liquefaction Runs.

α/α	Εξεταζόμενη Μέθοδος			
	Προς τα Εμπρός Ποιοτική Παλινδρόμηση		Προς τα Πίσω Ποιοτική Παλινδρόμηση	
	Συνδυασμός	Sum(sl)	Συνδυασμός	Sum(sl)
1	x6	3.86132	x1-x2-x3-x4-x5-x6-x7	3.11873
2	x1-x6	3.86277	x1-x2-x3-x4-x5-x6	3.19531
3	x2-x6	3.64066	x1-x2-x3-x4-x5-x7	3.27795
4	x3-x6	3.61351	x1-x2-x3-x4-x6-x7	3.11732
5	x4-x6	3.32231	x1-x2-x3-x5-x6-x7	3.15584
6	x5-x6	3.86405	x1-x2-x4-x5-x6-x7	3.12898
7	x6-x7	3.32279	x1-x3-x4-x5-x6-x7	3.19929
8	x1-x4-x6	3.31521	x2-x3-x4-x5-x6-x7	3.13906
9	x2-x4-x6	3.21847	x1-x2-x3-x4-x6	3.19403
10	x3-x4-x6	3.32666	x1-x2-x3-x4-x7	3.27661
11	x4-x5-x6	3.29840	x1-x2-x3-x6-x7	3.15389
12	x4-x6-x7	3.25530	x1-x2-x4-x6-x7	3.12696
13	x1-x2-x4-x6	3.19488	x1-x3-x4-x6-x7	3.22818
14	x2-x3-x4-x6	3.21757	x2-x3-x4-x5-x6	3.21824
15	x2-x4-x5-x6	3.21944	x2-x3-x4-x6-x7	3.13825
16	x2-x4-x6-x7	3.14836		
17	x1-x2-x4-x6-x7	3.12696		
18	x2-x3-x4-x6-x7	3.13825		
19	x2-x4-x5-x6-x7	3.14987		
20	x1-x2-x3-x4-x6-x7	3.11732		
21	x1-x2-x4-x5-x6-x7	3.12898		
22	x1-x2-x3-x4-x5-x6-x7	3.11873		
Προτεινόμενο καλύτερο υποσύνολο		x1-x2-x3-x4-x6-x7	x1-x2-x3-x4-x6-x7	

Ας σημειωθεί ότι στον Πίνακα 5.33, για την προς τα εμπρός διαδικασία, με έντονους χαρακτήρες σημειώνονται οι μεταβλητές που εισέρχονται στο μοντέλο, ενώ για την προς τα πίσω διαδικασία, σημειώνονται με τον ίδιο τρόπο οι αντίστοιχες μεταβλητές εξέρχονται του μοντέλου παλινδρόμησης. Με πλάγιους και κόκκινους χαρακτήρες, σημειώνονται οι καλύτεροι συνδυασμοί κάθε βήματος, για την αντίστοιχη εξεταζόμενη μέθοδο. Όπως φαίνεται και στον τελευταίο πίνακα, οι δύο διαδικασίες καταλήγουν στο ίδιο προτεινόμενο υποσύνολο μεταβλητών. Η περιγραφή των διαδικασιών, φαίνεται στη συνέχεια.

5.2.1.1.1 Προς τα εμπρός βηματική παλινδρόμηση

Αρχικά γίνεται η θεώρηση ότι καμμία μεταβλητή δε μετέχει στο μοντέλο παλινδρόμησης. Έτσι, εξετάζεται κάθε διαθέσιμη ανεξάρτητη μεταβλητή των δεδομένων, με εφαρμογή του γ.π. (4.8), για να διαπιστωθεί ποια δίνει το μικρότερο άθροισμα απλών απόλυτων σφαλμάτων. Από τον έλεγχο αυτό, προκύπτει ότι η x_6 είναι η μεταβλητή που θα εισαχθεί πρώτη στο μοντέλο. Στη συνέχεια, εξετάζονται όλοι οι δυνατοί συνδυασμοί των διαθέσιμων μεταβλητών, οι οποίοι περιέχουν την x_6 . Σε κάθε έναν από τους συνδυασμούς αυτούς, εφαρμόζεται πάλι το γ.π. (4.8) και προκύπτει ότι καλύτερος συνδυασμός μεταβλητών για το βήμα αυτό είναι ο x_4 - x_6 , αφού παρουσιάζει μικρότερο άθροισμα απόλυτων σφαλμάτων από τον συνδυασμό του προηγούμενου βήματος, οπότε εισάγεται στο μοντέλο η x_4 .

Πίνακας 5.34: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα των απλών απόλυτων σφαλμάτων, στα δεδομένα Belle Ayr Liquefaction Runs.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	$x_1, x_2, x_3, x_4, x_5, x_6, x_7$	x_6	3.861	x_6
2	x_1 - x_6 , x_2 - x_6 , x_3 - x_6 , x_4 - x_6 , x_5 - x_6 , x_6 - x_7	x_4 - x_6	3.322	x_4
3	x_1 - x_4 - x_6 , x_2 - x_4 - x_6 , x_3 - x_4 - x_6 , x_4 - x_5 - x_6 , x_4 - x_6 - x_7	x_2 - x_4 - x_6	3.218	x_2
4	x_1 - x_2 - x_4 - x_6 , x_2 - x_3 - x_4 - x_6 , x_2 - x_4 - x_5 - x_6 , x_2 - x_4 - x_6 - x_7	x_2 - x_4 - x_6 - x_7	3.148	x_7
5	x_1 - x_2 - x_4 - x_6 - x_7 , x_2 - x_3 - x_4 - x_6 - x_7 , x_2 - x_4 - x_5 - x_6 - x_7	x_1 - x_2 - x_4 - x_6 - x_7	3.127	x_1
6	x_1 - x_2 - x_3 - x_4 - x_6 - x_7 , x_1 - x_2 - x_4 - x_5 - x_6 - x_7	x_1 - x_2 - x_3 - x_4 - x_6 - x_7	3.117	x_3
7	x_1 - x_2 - x_3 - x_4 - x_5 - x_6 - x_7	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_1 - x_2 - x_3 - x_4 - x_6 - x_7				

Σύμφωνα με τα όσα προτείνονται στην προσέγγιση αυτή, στο επόμενο βήμα πρέπει να εξεταστούν, με παρόμοιο τρόπο, όλοι οι δυνατοί συνδυασμοί των ανεξάρτητων μεταβλητών, οι οποίοι περιέχουν και τις x_4 - x_6 . Η εξέταση αυτή δείχνει ότι ο συνδυασμός x_2 - x_4 - x_6 παρουσιάζει μικρότερο άθροισμα απόλυτων σφαλμάτων από αυτόν του προηγούμενου βήματος, οπότε επιλέγεται ως ο καλύτερος για το βήμα αυτό και η μεταβλητή x_2 εισέρχεται στο μοντέλο. Με όμοιο τρόπο, εξετάζονται οι συνδυασμοί που αυτή τη φορά περιέχουν και τις x_2 - x_4 - x_6 . Καλύτερος συνδυασμός αναδεικνύεται στο βήμα αυτό ο x_2 - x_4 - x_6 - x_7 , οπότε εισάγεται και η x_7 στο μοντέλο παλινδρόμησης. Ακολούθως εξετάζονται οι συνδυασμοί που να περιέχουν τις x_2 - x_4 - x_6 - x_7 . Στο βήμα αυτό, καλύτερος συνδυασμός αναδεικνύεται ο x_1 - x_2 - x_4 - x_6 - x_7 , η x_1 εισάγεται στο μοντέλο. Εξετάζοντας παρομοίως τους συνδυασμούς που περιέχουν τις x_1 - x_2 - x_4 - x_6 - x_7 , προκύπτει ότι στο μοντέλο πρέπει να εισαχθεί η x_3 , καθώς ο συνδυασμός x_1 - x_2 - x_3 - x_4 - x_6 - x_7 δίνει μικρότερο άθροισμα απόλυτων σφαλμάτων από τον συνδυασμό του προηγούμενου βήματος. Απομένει στη συνέχεια να εξεταστεί ο συνδυασμός που περιέχει όλες τις ανεξάρτητες μεταβλητές, αλλά ο έλεγχος αυτός δείχνει ότι δεν προκύπτει άθροισμα απόλυτων σφαλμάτων μικρότερο από το αντίστοιχο στο προηγούμενο βήμα. Έτσι, η διαδικασία σταματά εδώ και καλύτερο υποσύνολο μεταβλητών για το μοντέλο παλινδρόμησης αναδεικνύεται το x_1 - x_2 - x_4 - x_6 - x_7 . Όλη η παραπάνω διαδικασία, παρουσιάζεται στα βήματα του Πίνακα 5.34.

5.2.1.1.2 Προς τα πίσω βηματική παλινδρόμηση

Στην παράγραφο αυτή, θα ακολουθηθεί η αντίστροφη από αυτήν της προηγούμενης παραγράφου διαδικασία. Θεωρώντας ότι όλες οι ανεξάρτητες μεταβλητές μετέχουν στο μοντέλο, υπολογίζεται, βάση του γ.π.(4.8), το άθροισμα των απολύτων σφαλμάτων του συνδυασμού αυτού, το οποίο προκύπτει ίσο με 3.119.

Πίνακας 5.35: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα απλών απόλυτων σφαλμάτων, στα δεδομένα Belle Ayr Liquefaction Runs.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x_1 - x_2 - x_3 - x_4 - x_5 - x_6 x_1 - x_2 - x_3 - x_4 - x_5 - x_7 x_1 - x_2 - x_3 - x_4 - x_6 - x_7 x_1 - x_2 - x_3 - x_5 - x_6 - x_7 x_1 - x_2 - x_4 - x_5 - x_6 - x_7 x_1 - x_3 - x_4 - x_5 - x_6 - x_7 x_2 - x_3 - x_4 - x_5 - x_6 - x_7	x_1 - x_2 - x_3 - x_4 - x_6 - x_7	3.117	x_5
2	x_1 - x_2 - x_3 - x_4 - x_6 , x_1 - x_2 - x_4 - x_6 - x_7 , x_1 - x_3 - x_4 - x_6 - x_7 , x_1 - x_2 - x_3 - x_6 - x_7 , x_1 - x_2 - x_3 - x_4 - x_6 , x_2 - x_3 - x_4 - x_6 - x_7	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_1 - x_2 - x_3 - x_4 - x_6 - x_7				

Έπειτα εξετάζονται με παρόμοιο τρόπο όλοι οι δυνατοί συνδυασμοί που προκύπτουν με αφαίρεση μίας κάθε φορά μεταβλητής που μετέχει στον συνδυασμό του προηγούμενου βήματος. Διαπιστώνεται έτσι ότι ο συνδυασμός $x_1-x_2-x_3-x_4-x_6-x_7$ δίνει μικρότερο άθροισμα απόλυτων σφαλμάτων από τον συνδυασμό στο προηγούμενο βήμα, οπότε επιλέγεται ως ο καλύτερος για το βήμα αυτό και η μεταβλητή x_5 εξέρχεται από το μοντέλο. Ακολουθώντας, σχηματίζονται οι δυνατοί συνδυασμοί που προκύπτουν με αφαίρεση μίας κάθε φορά μεταβλητής από τον τελευταίο καλύτερο συνδυασμό. Με εφαρμογή του γ.π. (4.8) σε κάθε έναν από αυτούς, διαπιστώνεται ότι κανένας δεν παρουσιάζει μικρότερο άθροισμα απλών απόλυτων σφαλμάτων από τον προηγούμενο καλύτερο συνδυασμό. Επομένως, η διαδικασία τερματίζεται εδώ, με καλύτερο υποσύνολο μεταβλητών το $x_1-x_2-x_3-x_4-x_6-x_7$. Τα βήματα που περιγράφηκαν πριν, παρουσιάζονται εποπτικά στον Πίνακα 5.35.

5.2.1.2 Άθροισμα Τετραγωνικών Σφαλμάτων

Εδώ εφαρμόζονται οι προτεινόμενες προσεγγίσεις, υλοποιώντας το γ.π. (4.12) σε κάθε βήμα τους. Τα αποτελέσματα των διερευνήσεων αυτών, φαίνονται στον Πίνακα 5.36.

Πίνακας 5.36: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.12), στα δεδομένα Belle Ayr Liquefaction Runs.

α/α	Εξεταζόμενη Μέθοδος			
	Προς τα Εμπρός Ποιοτική Παλινδρόμηση		Προς τα Πίσω Ποιοτική Παλινδρόμηση	
	Συνδυασμός	Sum(σ^2)	Συνδυασμός	Sum(σ^2)
1	x_6	1.10052	$x_1-x_2-x_3-x_4-x_5-x_6-x_7$	0.75478
2	x_1-x_6	1.06922	$x_1-x_2-x_3-x_4-x_5-x_6$	0.79353
3	x_2-x_6	0.93211	$x_1-x_2-x_3-x_4-x_5-x_7$	0.79239
4	x_3-x_6	0.97482	$x_1-x_2-x_3-x_4-x_6-x_7$	0.75461
5	x_4-x_6	0.92539	$x_1-x_2-x_3-x_5-x_6-x_7$	0.76618
6	x_5-x_6	1.08096	$x_1-x_2-x_4-x_5-x_6-x_7$	0.75647
7	x_6-x_7	0.85310	$x_1-x_3-x_4-x_5-x_6-x_7$	0.79480
8	$x_1-x_4-x_6$	0.90383	$x_2-x_3-x_4-x_5-x_6-x_7$	0.77565
9	$x_2-x_4-x_6$	0.82054	$x_1-x_2-x_3-x_4-x_6$	0.79314
10	$x_3-x_4-x_6$	0.92300	$x_1-x_2-x_3-x_4-x_7$	0.85229
11	$x_4-x_5-x_6$	0.91807	$x_1-x_2-x_3-x_6-x_7$	0.76614
12	$x_4-x_6-x_7$	0.82193	$x_1-x_2-x_4-x_6-x_7$	0.75635
13	$x_1-x_2-x_4-x_6$	0.79936	$x_1-x_3-x_4-x_6-x_7$	0.79797
14	$x_2-x_3-x_4-x_6$	0.81412	$x_2-x_3-x_4-x_6-x_7$	0.77556
15	$x_2-x_4-x_5-x_6$	0.82078		
16	$x_2-x_4-x_6-x_7$	0.77814		
17	$x_1-x_2-x_4-x_6-x_7$	0.75635		
18	$x_2-x_3-x_4-x_6-x_7$	0.77556		
19	$x_2-x_4-x_5-x_6-x_7$	0.77821		
20	$x_1-x_2-x_3-x_4-x_6-x_7$	0.75461		
21	$x_1-x_2-x_4-x_5-x_6-x_7$	0.75647		
22	$x_1-x_2-x_3-x_4-x_5-x_6-x_7$	0.75478		
Προτεινόμενο καλύτερο υποσύνολο		$x_1-x_2-x_3-x_4-x_6-x_7$	$x_1-x_2-x_3-x_4-x_6-x_7$	

Αυτό που εύκολα γίνεται αντιληπτό, είναι ότι οι δύο εξεταζόμενες προσεγγίσεις, καταλήγουν στο ίδιο τελικό προτεινόμενο υποσύνολο μεταβλητών. Παρατηρώντας τους Πίνακες 5.35 και 5.36, διαπιστώνεται ότι σε κάθε βήμα κάθε προσέγγισης, εξετάζονται ακριβώς οι ίδιοι συνδυασμοί, κάτι το οποίο σημαίνει ότι η εξέταση του άθροισματος των απλών και των τετραγωνικών σφαλμάτων αξιών, οδηγεί ακριβώς στα ίδια αποτελέσματα, σε ό,τι αφορά τα βήματα των εξεταζόμενων διαδικασιών. Στη συνέχεια εξετάζονται αναλυτικά οι δύο προσεγγίσεις. Αρχή γίνεται με την προς τα εμπρός βηματική παλινδρόμησης.

5.2.1.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Στο πρώτο βήμα εξετάζονται όλες οι ανεξάρτητες μεταβλητές ως προς το άθροισμα τετραγωνικών σφαλμάτων. Προκύπτει ότι η x_6 παρουσιάζει το μικρότερο τέτοιο άθροισμα (1.10052), οπότε επιλέγεται για είσοδο στο μοντέλο παλινδρόμησης. Στο βήμα που ακολουθεί, εξετάζονται όλοι οι δυνατοί συνδυασμοί που προκύπτουν, θεωρώντας ότι η x_6 θα μετέχει σε αυτούς. Με εφαρμογή του γ.π. (4.12) σε κάθε συνδυασμό αυτού του βήματος, διαπιστώνεται ότι ο συνδυασμός x_6 - x_7 παρουσιάζει το μικρότερο άθροισμα τετραγωνικών σφαλμάτων. Έτσι, στο μοντέλο εισάγεται η x_7 και η διαδικασία συνεχίζεται εξετάζοντας τους συνδυασμούς που περιέχουν τις x_6 - x_7 . Η εξέταση αυτή δείχνει ότι καλύτερος συνδυασμός αναδεικνύεται, για το βήμα αυτό, ο x_2 - x_6 - x_7 , οπότε εισέρχεται στο μοντέλο η x_2 .

Πίνακας 5.37: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα των τετραγωνικών σφαλμάτων, στα δεδομένα Belle Ayr Liquefaction Runs.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	$x_1, x_2, x_3, x_4, x_5, x_6, x_7$	x_6	1.101	x_6
2	x_1 - x_6 , x_2 - x_6 , x_3 - x_6 , x_4 - x_6 , x_5 - x_6 , x_6 - x_7	x_6 - x_7	0.853	x_7
3	x_1 - x_6 - x_7 , x_2 - x_6 - x_7 , x_3 - x_6 - x_7 , x_4 - x_6 - x_7 , x_5 - x_6 - x_7	x_2 - x_6 - x_7	0.815	x_2
4	x_1 - x_2 - x_6 - x_7 , x_2 - x_3 - x_6 - x_7 , x_2 - x_5 - x_6 - x_7 , x_2 - x_4 - x_6 - x_7	x_2 - x_4 - x_6 - x_7	0.778	x_4
5	x_1 - x_2 - x_4 - x_6 - x_7 , x_2 - x_3 - x_4 - x_6 - x_7 , x_2 - x_4 - x_5 - x_6 - x_7	x_1 - x_2 - x_4 - x_6 - x_7	0.756	x_1
6	x_1 - x_2 - x_3 - x_4 - x_6 - x_7 , x_1 - x_2 - x_4 - x_5 - x_6 - x_7	x_1 - x_2 - x_3 - x_4 - x_6 - x_7	0.755	x_3
7	x_1 - x_2 - x_3 - x_4 - x_5 - x_6 - x_7	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_1 - x_2 - x_3 - x_4 - x_6 - x_7				

Στο επόμενο βήμα εξετάζονται οι δυνατοί συνδυασμοί που περιέχουν τις μεταβλητές αυτές και προκύπτει ότι μικρότερο άθροισμα τετραγωνικών σφαλμάτων, σε σχέση με το αντίστοιχο του προηγούμενου βήματος, δίνει ο συνδυασμός $x_2-x_4-x_6-x_7$. Συνεπώς, στο μοντέλο εισάγεται η x_4 . Με παρόμοιο τρόπο εξετάζονται οι συνδυασμοί που περιέχουν τις $x_2-x_4-x_6-x_7$. Προκύπτει ότι καλύτερος συνδυασμός για το βήμα αυτό είναι ο $x_1-x_2-x_4-x_6-x_7$, οπότε εισάγεται στο μοντέλο η x_1 . Η διαδικασία συνεχίζεται εξετάζοντας τους συνδυασμούς που περιέχουν τις μεταβλητές του καλύτερου συνδυασμού για το τελευταίο εξετασθέν βήμα και προκύπτει ότι νέος καλύτερος συνδυασμός αναδεικνύεται ο $x_1-x_2-x_3-x_4-x_6-x_7$. Στο μοντέλο εισάγεται η x_3 , ενώ η διαδικασία σταματά στο σημείο αυτό, καθώς ο συνδυασμός που απομένει να εξεταστεί (ο οποίος περιέχει όλες τις διαθέσιμες ανεξάρτητες μεταβλητές), δε δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων από τον καλύτερο συνδυασμό του προηγούμενου βήματος. Επομένως, καλύτερο προτεινόμενο υποσύνολο μεταβλητών αναδεικνύεται το $x_1-x_2-x_3-x_4-x_6-x_7$. Η όλη διαδικασία φαίνεται στα βήματα που περιέχει ο Πίνακας 5.37.

5.2.1.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Ο συνδυασμός ο οποίος περιέχει όλες τις διαθέσιμες ανεξάρτητες μεταβλητές, δίνει άθροισμα τετραγωνικών σφαλμάτων, υλοποιώντας το γ.π. (4.12), ίσο με 0.75478. Αφαιρώντας από τον συνδυασμό αυτόν μία κάθε φορά μεταβλητή, προκύπτουν επτά συνδυασμοί, οι οποίοι εξετάζονται και αυτοί ως το άθροισμα των τετραγωνικών σφαλμάτων που παρουσιάζουν με βάση το γ.π. (4.12). Προκύπτει, όμως, ότι ο συνδυασμός $x_1-x_2-x_3-x_4-x_6-x_7$ δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων σε σχέση με τον αρχικό συνδυασμό, οπότε η x_5 πρέπει να εξέρθει του μοντέλου παλινδρόμησης.

Πίνακας 5.38: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα τετραγωνικών σφαλμάτων, στα δεδομένα Belle Ayr Liquefaction Runs.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος από το Μοντέλο Παλινδρόμησης
1	$x_1-x_2-x_3-x_4-x_5-x_6$ $x_1-x_2-x_3-x_4-x_5-x_7$ $x_1-x_2-x_3-x_4-x_6-x_7$ $x_1-x_2-x_3-x_5-x_6-x_7$ $x_1-x_2-x_4-x_5-x_6-x_7$ $x_1-x_3-x_4-x_5-x_6-x_7$ $x_2-x_3-x_4-x_5-x_6-x_7$	$x_1-x_2-x_3-x_4-x_6-x_7$	0.755	x_5
2	$x_1-x_2-x_3-x_4-x_6$, $x_1-x_2-x_4-x_6-x_7$, $x_1-x_3-x_4-x_6-x_7$, $x_1-x_2-x_3-x_6-x_7$, $x_1-x_2-x_3-x_4-x_6$, $x_2-x_3-x_4-x_6-x_7$	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: $x_1-x_2-x_3-x_4-x_6-x_7$				

Από τον τελευταίο συνδυασμό που επιλέχθηκε ως ο καλύτερος, αφαιρείται πάλι μία κάθε φορά μεταβλητή και οι νέοι συνδυασμοί εξετάζονται με βάση την αντικειμενική συνάρτηση του γ.π. (4.12). Διαπιστώνεται στο σημείο αυτό ότι κανένας από τους συνδυασμούς αυτούς δε δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων από τον προηγούμενο, οπότε η διαδικασία σταματά εδώ. Καλύτερο προτεινόμενο υποσύνολο μεταβλητών αναδεικνύεται με τον τρόπο αυτό ο συνδυασμός $x_1-x_2-x_3-x_4-x_6-x_7$. Η διαδικασία αυτή, φαίνεται στον Πίνακα 5.38.

5.2.1.3 Συντελεστής προσδιορισμού

Στο σημείο αυτό θα αναζητηθεί το καλύτερο υποσύνολο μεταβλητών για τα δεδομένα του Πίνακα 5.30, με την υλοποίηση του γ.π. (4.17) για κάθε εξεταζόμενο συνδυασμό ο οποίος προκύπτει κατά τη διαδικασία διερεύνησης.

5.2.1.3.1 Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση

Αρχικά, γίνεται η θεώρηση ότι στο μοντέλο παλινδρόμησης μετέχουν όλες οι διαθέσιμες μεταβλητές. Για τον συνδυασμό αυτό των μεταβλητών, εφαρμόζεται το γ.π. (4.17) και προκύπτει η τιμή για την αντικειμενική που εξετάζεται, η οποία αντιστοιχεί σε ένα συγκεκριμένο συντελεστή πολλαπλής παλινδρόμησης (στην προκειμένη περίπτωση, προκύπτει $R^2 = 47.821\%$). Στη συνέχεια, από τον αρχικό συνδυασμό αφαιρείται μία κάθε φορά μεταβλητή και για κάθε έναν από τους συνδυασμούς που προκύπτουν με τον τρόπο αυτό, υλοποιείται το γ.π. (4.17). Τα αποτελέσματα συγκρίνονται με το αντίστοιχο του αρχικού συνδυασμού.

Πίνακας 5.39: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.17), στα δεδομένα Belle Ayr Liquefaction Runs.

Εξεταζόμενος Συνδυασμός	Τιμή Αντικειμενικής Συνάρτησης $(\sum_{i=1}^n (\hat{y}_i - y_i)^2)$	$\sum_{i=1}^n (y_i - \bar{Y})^2$	Υπολογιζόμενος R^2
$x_1-x_2-x_3-x_4-x_5-x_6-x_7$	3983.98	7635.20	47.821%
$x_1-x_2-x_3-x_4-x_5-x_6$	2007.65		73.705%
$x_1-x_2-x_3-x_4-x_5-x_7$	2664.56		65.102%
$x_1-x_2-x_3-x_4-x_6-x_7$	2631.02		65.541%
$x_1-x_2-x_3-x_5-x_6-x_7$	2597.15		65.985%
$x_1-x_2-x_4-x_5-x_6-x_7$	2541.07		66.719%
$x_1-x_3-x_4-x_5-x_6-x_7$	3463.13		54.643%
$x_2-x_3-x_4-x_5-x_6-x_7$	2291.77		69.984%
$x_1-x_2-x_3-x_4-x_5$	3183.4		58.306%
$x_1-x_2-x_3-x_4-x_6$	2582.86		66.172%
$x_1-x_2-x_3-x_5-x_6$	2505.86		67.180%
$x_1-x_2-x_4-x_5-x_6$	2583.34		66.165%
$x_1-x_3-x_4-x_5-x_6$	3387.84		55.629%
$x_2-x_3-x_4-x_5-x_6$	2681.33		64.882%

Η εξέταση αυτή δείχνει ότι υπάρχει συνδυασμός με καλύτερη τιμή για την αντικειμενική συνάρτηση που εξετάζεται από την αντίστοιχη για τον αρχικό συνδυασμό. Ο συνδυασμός αυτός είναι ο $x_1-x_2-x_3-x_4-x_5-x_6$, οπότε εξάγεται η μεταβλητή x_7 από το μοντέλο. Από τον συνδυασμό $x_1-x_2-x_3-x_4-x_5-x_6$, αφαιρείται μία κάθε φορά μεταβλητή, έτσι ώστε στο βήμα αυτό να εξεταστούν όλοι οι συνδυασμοί οι οποίοι προκύπτουν με τον τρόπο αυτό με χρήση του γ.π. (4.17). Τα αποτελέσματα συγκρίνονται με το αντίστοιχο για τον συνδυασμό από τον οποίο προέκυψαν όλοι οι προηγούμενοι και ελέγχεται εάν υπάρχει κάποιος ο οποίος να δίνει καλύτερο αποτέλεσμα για την υπό εξέταση αντικειμενική. Η εξέταση αυτή, δείχνει ότι δεν προκύπτει τέτοιος συνδυασμός στο βήμα αυτό, οπότε η διαδικασία σταματά εδώ και καλύτερο υποσύνολο μεταβλητών υποδεικνύεται το $x_1-x_2-x_3-x_4-x_5-x_6$. Τα αποτελέσματα των διερευνήσεων αυτών δίνονται στον Πίνακα 5.39, ενώ τα βήματα αυτής της διαδικασίας, φαίνονται στον Πίνακα 5.40.

Πίνακας 5.40: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για τον συντελεστή πολλαπλής παλινδρόμησης, στα δεδομένα Belle Ayr Liquefaction Runs.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	R^2 Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3-x4-x5-x6	x1-x2-x3-x4-x5-x6	73.705%	x7
	x1-x2-x3-x4-x5-x7			
	x1-x2-x3-x4-x6-x7			
	x1-x2-x3-x5-x6-x7			
	x1-x2-x4-x5-x6-x7			
	x1-x3-x4-x5-x6-x7			
	x2-x3-x4-x5-x6-x7			
2	x1-x2-x3-x4-x5, x1-x2-x3-x4-x6, x1-x2-x3-x5-x6, x1-x2-x4-x5-x6, x1-x3-x4-x5-x6, x2-x3-x4-x5-x6	-	-	-
	Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x3-x4-x5-x6			

5.2.1.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα

Στην παράγραφο αυτή, θα πραγματοποιηθεί η σύγκριση των αποτελεσμάτων που δίδουν οι προτεινόμενες μεθοδολογίες βηματικής ποιοτικής παλινδρόμησης. Ο Πίνακας 5.41 συγκεντρώνει τα αποτελέσματα των συγκρινόμενων μεθοδολογιών.

Σε γενικές γραμμές, μπορεί να διατυπωθεί ο ισχυρισμός ότι ισχύουν οι παρατηρήσεις που έγιναν για τα αποτελέσματα στο προηγούμενο σετ δεδομένων. Ότι δηλαδή για κάθε εξεταζόμενο συνδυασμό υπάρχουν μεγάλες αποκλίσεις μεταξύ του R^2 των κλασσικών μεθόδων βηματικής παλινδρόμησης σε σχέση με τα αντίστοιχα για την προτεινόμενη μεθοδολογία βηματικής ποιοτικής παλινδρόμησης.

Πίνακας 5.41: Αποτελέσματα των χρησιμοποιούμενων προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Belle Ayr Liquefaction Runs.

Εξεταζόμενος Συνδυασμός	Βηματική Παλινδρόμηση		Βηματική Ποιοτική Παλινδρόμηση			
	Εμπρός Επιλογή	Πίσω Απαλοιφή	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
	R^2	R^2	$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4-x5-x6-x7	0.905	0.905	3.11873	0.49701	0.75478	0.66509
x1-x2-x3-x4-x5-x6	0.87	0.87	3.19531	0.59934	0.79353	0.67042
x1-x2-x3-x4-x5-x7	0.847	0.847	3.27795	0.59298	0.79239	0.60822
x1-x2-x3-x4-x6-x7	0.905	0.905	3.11732	0.50445	0.75461	0.66625
x1-x2-x3-x5-x6-x7	0.905	0.905	3.15584	0.54482	0.76618	0.65142
x1-x2-x4-x5-x6-x7	0.905	0.905	3.12898	0.56621	0.75647	0.67338
x1-x3-x4-x5-x6-x7	0.905	0.905	3.19929	0.48426	0.79480	0.63573
x2-x3-x4-x5-x6-x7	0.905	0.905	3.13906	0.49163	0.77565	0.64642
x1-x2-x3-x4-x6	0.87	0.87	3.19403	0.60163	0.79314	0.67108
x1-x2-x3-x4-x7	0.834	0.834	3.27661	0.55450	0.85229	0.64918
x1-x2-x3-x6-x7	0.905	0.905	3.15389	0.54905	0.76614	0.65682
x1-x2-x4-x6-x7	0.905	0.905	3.12696	0.56753	0.75635	0.67506
x1-x3-x4-x6-x7	0.905	0.905	3.22818	0.52063	0.79797	0.65079
x2-x3-x4-x6-x7	0.905	0.905	3.13825	0.49836	0.77556	0.64671
x1-x2-x4-x6	0.813	0.813	3.19488	0.64205	0.79936	0.67543
x2-x3-x4-x6	0.826	0.844	3.21757	0.60059	0.81412	0.66763
x2-x4-x5-x6	0.87	0.87	3.21944	0.62158	0.82078	0.69969
x2-x4-x6-x7	0.834	0.834	3.14836	0.52701	0.77814	0.60414
x1-x4-x6	0.87	0.87	3.31521	0.52997	0.90383	0.63789
x2-x4-x6	0.87	0.87	3.21847	0.60670	0.82054	0.66705
x3-x4-x6	0.87	0.87	3.32666	0.51399	0.92300	0.60558
x4-x5-x6	0.87	0.87	3.29840	0.51808	0.91807	0.60968
x4-x6-x7	0.905	0.905	3.25530	0.49860	0.82193	0.64313
x1-x6	0.87	0.87	3.86277	0.64800	1.06922	0.64546
x2-x6	0.87	0.87	3.64066	0.67531	0.93211	0.68585
x3-x6	0.87	0.87	3.61351	0.49606	0.97482	0.56099
x3-x7	0.667	0.667	5.30425	0.38882	1.82410	0.31407
x4-x6	0.87	0.87	3.32231	0.51484	0.92539	0.62536
x5-x6	0.87	0.87	3.86405	0.54623	1.08096	0.60918
x6-x7	0.905	0.905	3.32279	0.52256	0.85310	0.63434
x6	0.87	0.87	3.86132	0.64046	1.10052	0.64006

Ακόμα, οι αποκλίσεις στα αποτελέσματα αυτά αμβλύνονται, εάν συγκριθούν μεταξύ τους τα αποτελέσματα που δίνει η χρησιμοποίηση των δύο διαφορετικών κριτηρίων (άθροισμα απλών και τετραγωνικών σφαλμάτων) στην ίδια (προτεινόμενη) μεθοδολογία. Παράλληλα, ισχύει και εδώ η παρατήρηση ότι συνδυασμοί που δίδουν μικρότερο άθροισμα σφαλμάτων, εμφανίζουν μεγαλύτερο συντελεστή πολλαπλής παλινδρόμησης. Δεν ισχύει, ωστόσο, πάντα η αρχή της αναλογικότητας, δηλαδή δεν

μπορεί να ισχυριστεί κανείς ότι μεταξύ δύο συγκεκριμένων συνδυασμών αυτός που παρουσιάζει το μικρότερο άθροισμα απλών ή τετραγωνικών σφαλμάτων θα παρουσιάζει και τον μεγαλύτερο συντελεστή πολλαπλής παλινδρόμησης. Εφαρμόζοντας τις διαδικασίες που περιγράφονται στις παραγράφους 4.3.1.1 και 4.3.1.2, προκύπτει ότι, με βάση τα δύο αυτά κριτήρια, καταλήγουμε στο ίδιο τελικό προτεινόμενο καλύτερο υποσύνολο μεταβλητών. Όλα δηλαδή τα κριτήρια για την προτεινόμενη μεθοδολογία, καταλήγουν στο τελικό υποσύνολο μεταβλητών.

Σε ότι αφορά τις παρατηρήσεις οι οποίες μπορούν να γίνουν συγκρίνοντας τα αποτελέσματα για όλα τα προτεινόμενα κριτήρια, αυτές μπορούν να προκύψουν εξετάζοντας τα περιεχόμενα του Πίνακα 5.42. Ο τελευταίος πίνακας συνοψίζει τα αποτελέσματα για τα τρία κριτήρια που εξετάζονται στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής ποιοτικής παλινδρόμησης.

Πίνακας 5.42: Αποτελέσματα των χρησιμοποιούμενων τριών κριτηρίων, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Belle Ay Liquefaction Runs.

Βηματική Ποιοτική Παλινδρόμηση					
Εξεταζόμενος Συνδυασμός	R^2	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
		$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4-x5-x6-x7	47.821%	3.11873	0.49701	0.75478	0.66509
x1-x2-x3-x4-x5-x6	73.705%	3.19531	0.59934	0.79353	0.67042
x1-x2-x3-x4-x5-x7	65.102%	3.27795	0.59298	0.79239	0.60822
x1-x2-x3-x4-x6-x7	65.541%	3.11732	0.50445	0.75461	0.66625
x1-x2-x3-x5-x6-x7	65.985%	3.15584	0.54482	0.76618	0.65142
x1-x2-x4-x5-x6-x7	66.719%	3.12898	0.56621	0.75647	0.67338
x1-x3-x4-x5-x6-x7	54.643%	3.19929	0.48426	0.79480	0.63573
x2-x3-x4-x5-x6-x7	69.984%	3.13906	0.49163	0.77565	0.64642
x1-x2-x3-x4-x5	58.306%	3.48694	0.69936	0.83875	0.60455
x1-x2-x3-x4-x6	66.172%	3.19403	0.60163	0.79314	0.67108
x1-x2-x3-x5-x6	67.180%	3.33701	0.62766	0.80717	0.64279
x1-x2-x4-x5-x6	66.165%	3.19618	0.63641	0.79970	0.67482
x1-x3-x4-x5-x6	55.629%	3.27333	0.51132	0.89221	0.59874

Αυτό που μπορεί να ισχυριστεί κανείς, είναι ότι τα αποτελέσματα για τη χρησιμοποίηση του συντελεστή πολλαπλής παλινδρόμησης ως κριτήριο βελτιστοποίησης στα πλαίσια της προτεινόμενης μεθοδολογίας, οδηγεί, σε γενικές γραμμές, σε καλύτερα ή παραπλήσια αποτελέσματα για κάθε συνδυασμό από τα αντίστοιχα κριτήρια του αθροίσματος των απόλυτων σφαλμάτων. Ακόμα, όμως, μια σημαντική παρατήρηση είναι ότι ακολουθώντας τις διαδικασίες που περιγράφονται στις ενότητες 4.3.1 και 4.3.2, και τα τρία κριτήρια υποδεικνύουν το ίδιο υποσύνολο ως το καλύτερο για συμμετοχή στο ζητούμενο μοντέλο παλινδρόμησης. Στο σημείο αυτό, θα εξεταστούν τα αποτελέσματα που προκύπτουν για τις επιδόσεις των επιμέρους δεικτών που υιοθετήθηκαν κατά τη διάρκεια των διερευνήσεων.

Πίνακας 5.43: Αποτελέσματα των επιμέρους χρησιμοποιούμενων δεικτών, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Belle Ayr Liquefaction Runs.

Εξεταζόμενος Συνδυασμός	Απλά Απόλυτα Σφάλματα				Τετραγωνικά Σφάλματα			
	R^2	Sum(e)	Sum(e^2)	MAPE	R^2	Sum(e)	Sum(e^2)	MAPE
x1-x2-x3-x4-x5-x6-x7	0.49701	283.547	7183.544	0.558	0.66509	244.253	5203.534	0.502
x1-x2-x3-x4-x5-x6	0.59934	305.195	7799.146	0.605	0.67042	248.879	4990.179	0.493
x1-x2-x3-x4-x5-x7	0.59298	291.345	8036.468	0.609	0.60822	270.680	6294.832	0.547
x1-x2-x3-x4-x6-x7	0.50445	279.757	6980.166	0.550	0.66625	244.536	5188.698	0.502
x1-x2-x3-x5-x6-x7	0.54482	247.453	6362.269	0.516	0.65142	250.148	5636.951	0.520
x1-x2-x4-x5-x6-x7	0.56621	257.787	5899.511	0.500	0.67338	243.960	5339.324	0.502
x1-x3-x4-x5-x6-x7	0.48426	302.542	8523.173	0.592	0.63573	263.869	5758.424	0.522
x2-x3-x4-x5-x6-x7	0.49163	281.872	7281.066	0.557	0.64642	249.683	5476.791	0.515
x1-x2-x3-x4-x6	0.60163	301.419	7663.703	0.601	0.67108	248.653	4978.021	0.493
x1-x2-x3-x4-x7	0.55450	264.573	6523.691	0.557	0.64918	254.182	5418.485	0.525
x1-x2-x3-x6-x7	0.54905	244.682	6226.701	0.511	0.65682	247.554	5551.869	0.513
x1-x2-x4-x6-x7	0.56753	256.057	5825.108	0.496	0.67506	244.417	5323.704	0.502
x1-x3-x4-x6-x7	0.52063	296.812	7889.982	0.575	0.65079	262.812	5438.573	0.512
x2-x3-x4-x6-x7	0.49836	278.473	7097.344	0.550	0.64671	249.897	5474.716	0.515
x1-x2-x4-x6	0.64205	284.923	6850.973	0.563	0.67543	255.761	5227.331	0.505
x2-x3-x4-x6	0.60059	291.714	7694.900	0.587	0.65304	261.899	5339.656	0.520
x2-x4-x5-x6	0.62158	290.607	7380.124	0.578	0.66668	255.181	5415.734	0.510
x2-x4-x6-x7	0.52701	272.498	6658.904	0.533	0.65751	249.898	5627.713	0.516
x1-x4-x6	0.52997	420.228	14054.130	0.770	0.63789	274.172	6416.292	0.530
x2-x4-x6	0.60670	300.047	7357.166	0.590	0.66705	254.518	5403.661	0.509
x3-x4-x6	0.51399	373.566	11849.280	0.705	0.60558	277.846	6591.621	0.530
x4-x5-x6	0.51808	399.125	13925.690	0.756	0.60968	273.693	6808.596	0.543
x4-x6-x7	0.49860	316.837	8653.529	0.613	0.64313	269.569	5715.993	0.533
x1-x6	0.64800	283.050	7770.043	0.513	0.64546	285.180	7714.661	0.518
x2-x6	0.67531	259.760	5477.921	0.553	0.68585	236.658	4763.720	0.480
x3-x6	0.49606	448.038	17158.640	0.829	0.56099	309.591	7693.557	0.575
x4-x6	0.51484	351.365	10140.280	0.655	0.62536	275.546	6626.059	0.538
x5-x6	0.54623	314.466	8941.487	0.572	0.60918	280.814	7788.450	0.531
x6-x7	0.52256	292.113	8034.785	0.577	0.63434	279.465	5885.867	0.539
x6	0.64046	278.856	7542.508	0.520	0.64006	287.032	7807.246	0.522

Ο Πίνακας 5.43 συνοψίζει τα αποτελέσματα για τους δείκτες αυτούς, για κάθε κριτήριο και κάθε συνδυασμό. Για το συγκεκριμένο σύνολο δεδομένων, παρουσιάζονται μερικές σημαντικές αποκλίσεις μεταξύ των τιμών των δεικτών για τους περισσότερους εξεταζόμενους συνδυασμούς, ανάλογα με το κριτήριο βελτιστοποίησης. Αυτό που φαίνεται να ισχύει, είναι ότι οι επιμέρους δείκτες αποδίδουν καλύτερα στην περίπτωση ελαχιστοποίησης του αθροίσματος των τετραγωνικών σφαλμάτων. Αξιοσημείωτο, πάντως, είναι το γεγονός ότι κατά την εξέταση των επιμέρους συνδυασμών στα βήματα που ακολουθήθηκαν, υπάρχουν συνδυασμοί οι οποίοι υπερτερούν σε ορισμένους δείκτες αξιολόγησης έναντι των αντίστοιχων του υποδεικνυόμενου ως «καλύτερου» συνδυασμού. Ένα τέτοιο γεγονός ίσως υποδεικνύει ότι οι δείκτες αυτοί είναι ανταγωνιστικοί μεταξύ τους.

5.2.2 Πρακτική Εφαρμογή 4: Τα Δεδομένα Body Fat of Females

Αυτά τα δεδομένα είναι διαθέσιμα στο διαδίκτυο και βρίσκονται στη βάση δεδομένων που διατηρείται στην ακόλουθη ηλεκτρονική διεύθυνση: <http://www.sci.usq.edu.au/staff/dunn/Datasets/applications/health/bodyfat.html>.

Πίνακας 5.44: Τα δεδομένα Body Fat of Females.

a/a	y	x ₁	x ₂	x ₃
1	11.9	19.5	43.1	29.1
2	22.8	24.7	49.8	28.2
3	18.7	30.7	51.9	37
4	20.1	29.8	54.3	31.1
5	12.9	19.1	42.2	30.9
6	21.7	25.6	53.9	23.7
7	27.1	31.4	58.5	27.6
8	25.4	27.9	52.1	30.6
9	21.3	22.1	49.9	23.2
10	19.3	25.5	53.5	24.8
11	25.4	31.1	56.6	30
12	27.2	30.4	56.7	28.3
13	11.7	18.7	46.5	23
14	17.8	19.7	44.2	28.6
15	12.8	14.6	42.7	21.3
16	23.9	29.5	54.4	30.1
17	22.6	27.7	55.3	25.7
18	25.4	30.2	58.6	24.6
19	14.8	22.7	48.2	27.1
20	21.1	25.2	51	27.5

Τα δεδομένα που φαίνονται στον Πίνακα 5.44 αφορούν την πρόβλεψη της ποσότητας λίπους στο γυναικείο σώμα, με βάση τις διαστάσεις τριών ανθρωπομετρικών χαρακτηριστικών, για 20 υγιείς γυναίκες ηλικίας 20-34 ετών. Τα χαρακτηριστικά αυτά είναι: διάσταση τρικέφαλων (x₁, σε άγνωστο μετρικό σύστημα), περιφέρεια μηρών (x₂, σε άγνωστο μετρικό σύστημα), περιφέρεια αγκώνα χεριού (x₃, σε άγνωστο μετρικό σύστημα). Αξίζει να σημειωθεί ότι οι μετρήσεις λήφθηκαν με τη χρήση ακριβής μεθόδου και εξοπλισμού, οπότε θα ήταν χρήσιμο εάν προέκυπτε ένα αξιόπιστο μοντέλο παλινδρόμησης για την πρόβλεψη λίπους στο σώμα με βάση τα τρία εξεταζόμενα

χαρακτηριστικά, αφού οι μετρήσεις για τα υπόλοιπα χαρακτηριστικά μπορούν να ληφθούν εύκολα και ανέξοδα. Οι συσχετίσεις όλων των μεταβλητών μεταξύ τους, παρουσιάζονται στον Πίνακα 5.45.

Πίνακας 5.45: Συσχετίσεις μεταξύ των μεταβλητών των δεδομένων.

	x_1	x_2	x_3	y
x_1	1.000			
x_2	0.924	1.000		
x_3	0.458	0.085	1.000	
y	0.843	0.878	0.142	1.000

Στον Πίνακα 5.46 συγκεντρώνονται τα αποτελέσματα των διερευνήσεων για κάθε συνδυασμό μεταβλητών των διαθέσιμων δεδομένων.

Πίνακας 5.46: Αποτελέσματα για την εφαρμογή της βηματικής παλινδρόμησης, με τη χρήση του πακέτου SPSS v.12, στα δεδομένα Body Fat of Females.

Συνδυασμός	Βηματική Παλινδρόμηση			Εμπρός Επιλογή			Πίσω Απαλοιφή		
	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης
$x_1-x_2-x_3$	0.986	0.985	2.584	0.986	0.985	2.584	0.986	0.985	2.584
x_1-x_2	0.983	0.982	2.760	0.983	0.982	2.760	0.983	0.982	2.760
x_1-x_3	0.986	0.985	2.584	0.986	0.985	2.584	0.986	0.985	2.584
x_2-x_3	0.974	0.973	3.429	0.974	0.973	3.429	0.974	0.973	3.429
x_1	0.983	0.982	2.760	0.983	0.982	2.760	0.983	0.982	2.760
x_2	0.974	0.973	3.429	0.974	0.973	3.429	0.974	0.973	3.429
x_3	0.936	0.932	5.408	0.936	0.932	5.408	0.936	0.932	5.408

Πίνακας 5.47: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Body Fat of Females.

α/α	Εξεταζόμενη Μέθοδος			
	Προς τα Εμπρός Ποιοτική Παλινδρόμηση		Προς τα Πίσω Ποιοτική Παλινδρόμηση	
	Συνδυασμός	Sum(sl)	Συνδυασμός	Sum(sl)
1	x_1	2.45523	$x_1-x_2-x_3$	1.92439
2	x_2	2.04633	x_1-x_2	2.02053
3	x_3	4.21008	x_1-x_3	2.46032
4	x_1-x_2	2.02053	x_2-x_3	1.92304
5	x_2-x_3	1.92304	x_2	2.04633
6	$x_1-x_2-x_3$	1.92439	x_3	4.21008
Προτεινόμενο καλύτερο υποσύνολο		x_2-x_3	x_2-x_3	

5.2.2.1 Άθροισμα Απόλυτων Σφαλμάτων

Στην ενότητα αυτή θα εξεταστούν τα αποτελέσματα που δίνουν οι προτεινόμενες μεθοδολογίες, με βάση αρχικά το γ.π. (4.8), στα εξεταζόμενα δεδομένα. Τα αποτελέσματα των διερευνήσεων αυτών, φαίνονται στον Πίνακα 5.47. Αναλυτικά αποτελέσματα για όλους τους συνδυασμούς μεταβλητών αυτού του συνόλου

δεδομένων, στην περίπτωση εξέτασης του γ.π. (4.8), μπορούν να βρεθούν στον Πίνακα Α.4 του Παραρτήματος.

5.2.2.1.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Στην αρχή εξετάζονται οι τρεις διαθέσιμες ανεξάρτητες μεταβλητές ως προς το άθροισμα απόλυτων σφαλμάτων που παρουσιάζουν. Από αυτές, μικρότερο τέτοιο άθροισμα παρουσιάζει η x_2 , η οποία και εισέρχεται στο μοντέλο. Εξετάζοντας τους συνδυασμούς με δύο μεταβλητές που προκύπτουν συμπεριλαμβάνοντας τη x_2 σε αυτούς, διαπιστώνεται ότι καλύτερος συνδυασμός για το βήμα αυτό είναι ο x_2 - x_3 , καθώς παρουσιάζει μικρότερο άθροισμα απόλυτων σφαλμάτων από τον συνδυασμό του προηγούμενου βήματος και, άρα, η x_3 εισέρχεται στο μοντέλο. Στη συνέχεια, εξετάζεται ο συνδυασμός που περιέχει και τις τρεις ανεξάρτητες μεταβλητές, αλλά διαπιστώνεται ότι αυτός δεν έχει μικρότερο άθροισμα απόλυτων σφαλμάτων από τον προηγούμενο καλύτερο συνδυασμό. Έτσι, η διαδικασία σταματά εδώ και καλύτερο υποσύνολο μεταβλητών για το μοντέλο παλινδρόμησης, υποδεικνύεται το x_2 - x_3 . Αξίζει, ωστόσο, στο σημείο αυτό να γίνει η παρατήρηση ότι με βάση το κριτήριο του αθροίσματος των απλών σφαλμάτων, επιλέγεται ο συνδυασμός x_2 - x_3 , αλλά καλύτερο συντελεστή πολλαπλής παλινδρόμησης, παρουσιάζει ο συνδυασμός x_1 - x_2 , όπως μπορεί να διαπιστωθεί στον Πίνακα Α.4 του Παραρτήματος. Ο Πίνακας 5.48 συνοψίζει την διαδικασία αυτή.

Πίνακας 5.48: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα των απλών απόλυτων σφαλμάτων, στα δεδομένα Body Fat of Females.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	x_1, x_2, x_3	x_2	2.046	x_2
2	x_1 - x_2, x_2 - x_3	x_2 - x_3	1.923	x_3
3	x_1 - x_2 - x_3	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_2 - x_3				

5.2.2.1.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Η διαδικασία που ακολουθείται εδώ, είναι η αντίστροφη της προηγούμενης παραγράφου. Αρχικά, όλες οι ανεξάρτητες μεταβλητές μετέχουν στο μοντέλο παλινδρόμησης και παρουσιάζουν άθροισμα απόλυτων σφαλμάτων ίσο με 1.924, όπως προκύπτει με υλοποίηση του γ.π. (4.8). Αφαιρώντας μία κάθε φορά μεταβλητή από τον αρχικό συνδυασμό, προκύπτουν τρεις νέοι συνδυασμοί, οι οποίοι εξετάζονται ως προς το άθροισμα απόλυτων σφαλμάτων. Ο συνδυασμός x_2 - x_3 εμφανίζει μικρότερο άθροισμα απόλυτων σφαλμάτων από τον αρχικό συνδυασμό, οπότε επιλέγεται ως ο καλύτερος για το βήμα αυτό και η μεταβλητή x_1 εξέρχεται του μοντέλου. Οι συνδυασμοί που εξετάζονται στη συνέχεια, είναι δύο και προκύπτει ότι κανείς τους δε δίνει καλύτερο αποτέλεσμα για το άθροισμα απόλυτων σφαλμάτων. Συνεπώς, η διαδικασία τερματίζεται εδώ και ως καλύτερο υποσύνολο μεταβλητών

προτείνεται το x_2 - x_3 . Σημειώνεται και στην περίπτωση αυτή, ότι, με βάση τα στοιχεία του Πίνακα Α.4 του Παραρτήματος, καλύτερο συντελεστή πολλαπλής παλινδρόμησης φαίνεται να εμφανίζει ο συνδυασμός x_1 - x_2 . Στον Πίνακα 5.49 αποτυπώνονται τα βήματα της παραπάνω διαδικασίας.

Πίνακας 5.49: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα των απλών απόλυτων σφαλμάτων, στα δεδομένα Body Fat of Females.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x_1 - x_2 , x_1 - x_3 , x_2 - x_3	x_2 - x_3	1.923	x_1
2	x_2, x_3	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_2 - x_3				

5.2.2.2 Άθροισμα Τετραγωνικών Σφαλμάτων

Οι προτεινόμενες μεθοδολογίες θα εξεταστούν εδώ βάσει του κριτηρίου του αθροίσματος των τετραγωνικών σφαλμάτων, με τη χρήση του γ.π. (4.12), ώστε να εντοπιστεί το «καλύτερο» υποσύνολο μεταβλητών. Ο Πίνακας 5.50 συγκεντρώνει όλα τα αποτελέσματα αυτής της εξέτασης, ενώ αναλυτικά αποτελέσματα για όλους τους συνδυασμούς μεταβλητών, στην περίπτωση αυτή, μπορούν να βρεθούν στον Πίνακα Α.5 του Παραρτήματος.

Πίνακας 5.50: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.12), στα δεδομένα Body Fat of Females.

α/α	Εξεταζόμενη Μέθοδος			
	Προς τα Εμπρός Ποιοτική Παλινδρόμηση		Προς τα Πίσω Ποιοτική Παλινδρόμηση	
	Συνδυασμός	Sum(σ^2)	Συνδυασμός	Sum(σ^2)
1	x_1	0.532849	x_1 - x_2 - x_3	0.344421
2	x_2	0.498789	x_1 - x_2	0.379035
3	x_3	1.732601	x_1 - x_3	0.53333
4	x_1 - x_2	0.379035	x_2 - x_3	0.359602
5	x_2 - x_3	0.359602		
6	x_1 - x_2 - x_3	0.344421		
Προτεινόμενο καλύτερο υποσύνολο		x_1 - x_2 - x_3	x_1 - x_2 - x_3	

5.2.2.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Η εφαρμογή του γ.π. (4.12) στις τρεις ανεξάρτητες μεταβλητές των διαθέσιμων δεδομένων, έδειξε ότι μικρότερο άθροισμα τετραγωνικών σφαλμάτων, έδωσε η x_2 , οπότε αυτή θα εισέλθει πρώτη στο μοντέλο παλινδρόμησης. Οι συνδυασμοί που περιέχουν την x_2 , εξετάζονται με παρόμοιο τρόπο και διαπιστώνεται ότι καλύτερος συνδυασμός για το βήμα αυτό, είναι ο x_2 - x_3 , οπότε η x_3 εισάγεται στο μοντέλο. Στο επόμενο βήμα εξετάζεται ο εναπομείνας συνδυασμός x_1 - x_2 - x_3 , ο

οποίος παρουσιάζει μικρότερο άθροισμα απόλυτων σφαλμάτων από τον προηγούμενο καλύτερο συνδυασμό και η x_1 εισάγεται στο μοντέλο. Έτσι, η διαδικασία λαμβάνει τέλος στο σημείο αυτό, αφού δεν υπάρχουν άλλοι δυνατοί συνδυασμοί να εξεταστούν, και καλύτερο υποσύνολο μεταβλητών αναδεικνύεται το x_1 - x_2 - x_3 . Αξιοσημείωτο είναι το γεγονός ότι, όπως και στην περίπτωση του αθροίσματος απλών απόλυτων σφαλμάτων, έτσι και στην περίπτωση του αθροίσματος των τετραγωνικών σφαλμάτων, άλλο υποσύνολο υποδεικνύεται ως το καλύτερο, αλλά με βάση τον συντελεστή πολλαπλής παλινδρόμησης ο οποίος προκύπτει κατά την εξέταση, υποδεικνύεται καλύτερο υποσύνολο άλλος συνδυασμός μεταβλητών. Πιο συγκεκριμένα, εδώ, με βάση το άθροισμα των τετραγωνικών σφαλμάτων, καλύτερο υποσύνολο υποδεικνύεται το x_1 - x_2 - x_3 , αλλά καλύτερο συντελεστή πολλαπλής παλινδρόμησης εμφανίζει το υποσύνολο x_1 - x_2 (βλ. Πίνακα Α.5 του Παραρτήματος). Στον Πίνακα 5.51 φαίνονται τα βήματα τα οποία ακολουθήθηκαν στα προηγούμενα.

Πίνακας 5.51: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα των τετραγωνικών σφαλμάτων, στα δεδομένα Body Fat of Females.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	x_1, x_2, x_3	x_2	0.499	x_2
2	x_1 - x_2 , x_2 - x_3	x_2 - x_3	0.36	x_3
3	x_1 - x_2 - x_3	x_1 - x_2 - x_3	0.344	x_1
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x_1 - x_2 - x_3				

5.2.2.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Ακολουθώντας τα βήματα τα οποία περιγράφονται στην παράγραφο 4.3.2.2, προκύπτει ότι ο αρχικός συνδυασμός x_1 - x_2 - x_3 παρουσιάζει άθροισμα τετραγωνικών σφαλμάτων ίσο με 0.344 και ότι κανένας από τους συνδυασμούς που εξετάζονται στο επόμενο βήμα, αφαιρώντας μία κάθε φορά μεταβλητή από αυτόν, δεν παρουσιάζει μικρότερο άθροισμα τετραγωνικών σφαλμάτων. Έτσι, η διαδικασία σταματά εδώ και καλύτερο προτεινόμενο υποσύνολο μεταβλητών για συμμετοχή στο μοντέλο παλινδρόμησης, με βάση την προσέγγιση αυτή, αναδεικνύεται το x_1 - x_2 - x_3 . Παρόλα αυτά, ισχύει και εδώ η παρατήρηση ότι καλύτερο υποσύνολο αναδεικνύεται το x_1 - x_2 - x_3 με βάση το άθροισμα των τετραγωνικών σφαλμάτων, ενώ καλύτερο συντελεστή πολλαπλής παλινδρόμησης εμφανίζει το x_1 - x_2 (βλ. Πίνακα Α.5 του Παραρτήματος). Ο Πίνακας 5.52 συνοψίζει όλα τα παραπάνω.

5.2.2.3 Συντελεστής προσδιορισμού

Ακολουθώντας τη διαδικασία που περιγράφεται στην παράγραφο 4.3.2.3, υπολογίζεται ο συντελεστής προσδιορισμού για τη συμμετοχή όλων των διαθέσιμων μεταβλητών που εξετάζονται στην περίπτωση αυτή. Ο Πίνακας 5.53 συνοψίζει τα αποτελέσματα των διερευνήσεων για τους συνδυασμούς που εξετάστηκαν με βάση το γ.π. (4.17).

Πίνακας 5.52: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα των τετραγωνικών σφαλμάτων, στα δεδομένα Body Fat of Females.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2, x1-x3, x2-x3	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1- x2-x3				

Πίνακας 5.53: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.17), στα δεδομένα Body Fat of Females.

Εξεταζόμενος Συνδυασμός	Τιμή Αντικειμενικής Συνάρτησης $\sum_{i=1}^n (\hat{y}_i - y_i)^2$	$\sum_{i=1}^n (y_i - \bar{Y})^2$	Υπολογιζόμενος R^2
x1-x2-x3	104.201	495,39	78.966%
x1-x2	129.699		73.819%
x1-x3	154.348		68.843%
x2-x3	205		58.618%

5.2.2.3.1 Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση

Προκύπτει εδώ ότι η τιμή για τον R^2 , για τον αρχικό συνδυασμό, είναι ίση με 78.966%. Στη συνέχεια εξετάζονται όλοι οι συνδυασμοί που προκύπτουν με αφαίρεση μίας κάθε φορά μεταβλητής από τον αρχικό συνδυασμό, υλοποιώντας για κάθε έναν από αυτούς το γ.π. (4.17). Τα αποτελέσματα αυτών των διερευνήσεων, δεικνύουν ότι δεν υπάρχει συνδυασμός ο οποίος να διαθέτει συντελεστή πολλαπλής παλινδρόμησης καλύτερο από αυτόν του προηγούμενου βήματος, οπότε η διαδικασία σταματά στο σημείο αυτό. Καλύτερο υποσύνολο υποδεικνύεται, με τον τρόπο αυτό, το x1-x2-x3. Τα βήματα που περιγράφηκαν προηγουμένως, συνοψίζονται στον Πίνακα 5.54.

Πίνακας 5.54: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για τον συντελεστή πολλαπλής παλινδρόμησης, στα δεδομένα Body Fat of Females.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	R^2 Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2, x1-x3, x2-x3	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x3				

5.2.2.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα

Στην παράγραφο αυτή συγκεντρώνονται όλα τα αποτελέσματα των διερευνήσεων που πραγματοποιήθηκαν για όλα τα κριτήρια στα υπό εξέταση δεδομένα. Ο Πίνακας 5.55 συγκεντρώνει τα αποτελέσματα των συγκρινόμενων μεθοδολογιών.

Πίνακας 5.55: Αποτελέσματα των χρησιμοποιούμενων προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Body Fat of Females.

Εξεταζόμενος Συνδυασμός	Βηματική Παλινδρόμηση		Βηματική Ποιοτική Παλινδρόμηση			
	Εμπρός Επιλογή	Πίσω Απαλοιφή	Απλά Απόλυτα Σφάλματα		Τετραγωνικά	
	R^2	R^2	$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3	0.986	0.986	1.92439	0.685	0.344421	0.680
x1-x2	0.983	0.983	2.02053	0.714	0.379035	0.734
x1-x3	0.986	0.986	2.46032	0.657	0.53333	0.625
x2-x3	0.974	0.974	1.92304	0.684	0.359602	0.675
x1	0.983	0.983	2.45523	0.697	0.532849	0.626
x2	0.974	0.974	2.04633	0.677	0.498789	0.669
x3	0.936	0.936	4.21008	0.049	1.732601	0.061

Και σε αυτή την περίπτωση, παρατηρούνται μεγάλες αποκλίσεις μεταξύ του R^2 ο οποίος προκύπτει για τους αντίστοιχους εξεταζόμενους συνδυασμούς με βάση τη χρησιμοποίηση της κλασσικής μεθόδου βηματικής παλινδρόμησης σε σύγκριση με την προτεινόμενη μεθοδολογία βηματικής ποιοτικής παλινδρόμησης. Οι αποκλίσεις στα αποτελέσματα αυτά αμβλύνονται, εάν συγκριθούν μεταξύ τους τα αποτελέσματα που δίνει η χρησιμοποίηση των δύο διαφορετικών κριτηρίων (άθροισμα απλών και τετραγωνικών σφαλμάτων) στην ίδια (προτεινόμενη) μεθοδολογία. Μάλιστα, αυτό το οποίο παρατηρείται για τα δύο αυτά κριτήρια, είναι ότι τα αποτελέσματα που λαμβάνονται για κάθε συνδυασμό και για κάθε κριτήριο, είναι, παραπλήσια μεταξύ τους. Επιπλέον, από τα δεδομένα του Πίνακα 5.55, δεν φαίνεται σαφώς εάν τα αποτελέσματα ενός κριτηρίου υπερτερούν έναντι των αντίστοιχων για το άλλο κριτήριο.

Στο σημείο αυτό θα εξεταστούν τα αποτελέσματα που λαμβάνονται για τις τρεις προσεγγίσεις της προτεινόμενης μεθοδολογίας. Εξετάζονται, δηλαδή, οι επιδόσεις των επιμέρους κριτηρίων που επιλέγονται ως κριτήρια βελτιστοποίησης. Ο Πίνακας 5.56 συνοψίζει τα αποτελέσματα για τα τρία αυτά κριτήρια.

Η παρατήρηση εδώ είναι ότι δεν είναι εύκολα δυνατόν κανείς να συμπεράνει ποιο από τα τρία κριτήρια δίνει τα καλύτερα αποτελέσματα για το υπό εξέταση σύνολο δεδομένων. Ωστόσο, φαίνεται ότι το κριτήριο του αθροίσματος των απλών απόλυτων σφαλμάτων δίνει καλύτερα αποτελέσματα από τα αντίστοιχα για το άθροισμα των τετραγωνικών σφαλμάτων, ενώ το κριτήριο του συντελεστή πολλαπλής παλινδρόμησης κατά περίπτωση δίνει καλύτερα, χειρότερα ή παραπλήσια αποτελέσματα. Τέλος, στον Πίνακα 5.57 συνοψίζονται τα αποτελέσματα που

εξάγονται για τους επιπλέον δείκτες που εξετάστηκαν κατά τη διάρκεια των διερευνήσεων.

Πίνακας 5.56: Αποτελέσματα των χρησιμοποιούμενων τριών κριτηρίων, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Body Fat of Females.

Βηματική Ποιοτική Παλινδρόμηση					
Εξεταζόμενος Συνδυασμός	R^2	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
		$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3	79.200%	1.92439	0.685	0.344421	0.680
x1-x2	73.819%	2.02053	0.714	0.379035	0.734
x1-x3	68.843%	2.46032	0.657	0.53333	0.625
x2-x3	58.618%	1.92304	0.684	0.359602	0.675
x1	70.134%	2.45523	0.697	0.532849	0.626
x2	59.523%	2.04633	0.677	0.498789	0.669
x3	-5.845%	4.21008	0.049	1.732601	0.061

Πίνακας 5.57: Αποτελέσματα των επιμέρους χρησιμοποιούμενων δεικτών, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Body Fat of Females.

Εξεταζόμενος Συνδυασμός	Απλά Απόλυτα Σφάλματα				Τετραγωνικά Σφάλματα			
	R^2	Sum(e)	Sum(e ²)	MAPE	R^2	Sum(e)	Sum(e ²)	MAPE
x1-x2-x3	0.75	2207.559	820776.737	0.037	0.722	3088.636	902722.823	0.054
x1-x2	0.615	3413.403	1446863.411	0.062	0.622	3637.605	1295585.67	0.065
x1-x3	0.363	4028.918	2419314.487	0.071	0.335	4516.028	2470691.43	0.079
x2-x3	0.6	2637.249	1470239.249	0.042	0.556	3379.552	1480041.57	0.057
x1	0.169	7516.176	5217940.463	0.135	0.161	7615.136	5554980.88	0.133
x2	0.475	4292.912	2028193.374	0.073	0.551	4074.187	1554321.53	0.071
x3	0.008	6782.771	6160363.33	0.112	0	7984.17	6311011.54	0.136

Ως συμπέρασμα μπορεί να προκύψει ότι όταν βελτιώνεται/χειροτερεύει η τιμή ενός κριτηρίου από συνδυασμό σε συνδυασμό, παρατηρείται και μια σχετική βελτίωση/χειροτέρευση της επίδοσης των υπόλοιπων κριτηρίων. Τέλος, διαφαίνεται και στην περίπτωση αυτή, ότι το κριτήριο του αθροίσματος των απλών σφαλμάτων, αποδίδει καλύτερα από αυτό του αθροίσματος των τετραγωνικών σφαλμάτων.

5.2.3 Πρακτική Εφαρμογή 5: Τα Δεδομένα Hald

Τα δεδομένα αυτά, τα οποία παρουσιάστηκαν ήδη από το 1952, αποτελούν «κλασσικό» παράδειγμα εφαρμογής σε πολλές ερευνητικές εργασίες αλλά και βιβλία που ασχολούνται με το θέμα επιλογής μεταβλητών. Αφορούν την ποσότητα θερμότητας (σε θερμίδες) που ενυπάρχει σε κάθε γραμμάριο τσιμέντου (y), ως συνάρτηση της ποσότητας τεσσάρων συστατικών στοιχείων που περιλαμβάνονται στο μίγμα τσιμέντου. Τα στοιχεία αυτά είναι: tricalcium aluminate (x₁), tricalcium silicate (x₂), tetracalcium alumino ferrite (x₃), dicalcium silicate (x₄). Τα διαθέσιμα δεδομένα, φαίνονται στον Πίνακα 5.58. Στόχος είναι να προσδιοριστεί ποια από τα συστατικά αυτά, μπορούν να συνεισφέρουν στη δόμηση ενός όσο το δυνατόν

αξιόπιστου μοντέλου παλινδρόμησης, το οποίο θα προβλέπει την ποσότητα θερμότητας ανά γραμμάριο τσιμέντου, με βάση τις ποσότητες συμμετοχής του εκάστοτε συστατικού στο μίγμα που συνιστά το τσιμέντο. Οι συσχετίσεις μεταξύ όλων των μεταβλητών του Πίνακα 5.58, φαίνονται στον Πίνακα 5.59.

Πίνακας 5.58: Τα δεδομένα Hald (Montgomery and Peck, 1992).

a/a	y (cal/gr)	x ₁ (gr)	x ₂ (gr)	x ₃ (gr)	x ₄ (gr)
1	78.5	7	26	6	60
2	74.3	1	29	15	52
3	104.3	11	56	8	20
4	87.6	11	31	8	47
5	95.9	7	52	6	33
6	109.2	11	55	9	22
7	102.7	3	71	17	6
8	72.5	1	31	22	44
9	93.1	2	54	18	22
10	115.9	21	47	4	26
11	83.8	1	40	23	34
12	113.3	11	66	9	12
13	109.4	10	68	8	12

Πίνακας 5.59: Συσχετίσεις μεταξύ των μεταβλητών των δεδομένων.

	x1	x2	x3	x4	y
x1	1				
x2	0.229	1			
x3	-0.824	-0.139	1		
x4	-0.245	-0.973	0.03	1	
y	0.731	0.816	-0.535	0.821	1

Πίνακας 5.60: Αποτελέσματα για την εφαρμογή της βηματικής παλινδρόμησης, με τη χρήση του πακέτου SPSS v.12, στα δεδομένα Hald.

Συνδυασμός	Βηματική Παλινδρόμηση			Εμπρός Επιλογή			Πίσω Απαλοιφή		
	R ²	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R ²	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R ²	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης
x1-x2-x3-x4	1.000	0.999	2.418	1.000	0.999	2.418	1.000	0.999	2.418
x1-x2-x3	0.993	0.991	9.268	0.993	0.991	9.268	0.993	0.991	9.268
x1-x2-x4	0.998	0.998	4.299	0.998	0.998	4.299	0.998	0.998	4.299
x1-x3-x4	0.972	0.966	17.673	0.972	0.966	17.673	0.972	0.966	17.673
x2-x3-x4	0.991	0.989	10.041	0.991	0.989	10.041	0.993	0.991	9.336
x1-x2	0.974	0.970	16.843	0.974	0.970	16.843	0.974	0.970	16.843
x1-x3	0.972	0.966	17.673	0.972	0.966	17.673	0.972	0.966	17.673
x1-x4	0.854	0.827	40.101	0.854	0.827	40.101	0.854	0.827	40.101
x2-x4	0.991	0.989	10.041	0.991	0.989	10.041	0.991	0.989	10.041
x1	0.730	0.707	52.225	0.730	0.707	52.225	0.730	0.707	52.225
x2	0.961	0.958	19.738	0.961	0.958	19.738	0.961	0.958	19.738
x3	0.704	0.679	54.648	0.704	0.679	54.648	0.704	0.679	54.648
x4	0.661	0.633	58.447	0.661	0.633	58.447	0.661	0.633	58.447

Τα αποτελέσματα αφορούν την εφαρμογή αυτών των τριών γνωστών βηματικών μεθόδων στα δεδομένα του Πίνακα 5.58 και συγκεντρώνονται στον Πίνακα 5.60. Σε αυτόν τον πίνακα διακρίνονται και τα διάφορα κριτήρια αξιολόγησης του μοντέλου παλινδρόμησης που προκύπτει. Αναλυτικά, όλα τα αντίστοιχα αποτελέσματα για όλους τους συνδυασμούς αυτού του συνόλου δεδομένων, μπορούν να βρεθούν στον Πίνακα Α.6 του Παραρτήματος.

5.2.3.1 Άθροισμα Απόλυτων Σφαλμάτων

Η διαδικασία που ακολουθούνται στα πλαίσια αυτού του κριτηρίου, περιγράφονται στις παραγράφους 4.3.1.1 και 4.3.2.1. Τα αποτελέσματα από την εφαρμογή τους στα δεδομένα Hald, δίνονται στη συνέχεια, στον Πίνακα 5.61.

Πίνακας 5.61: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Hald.

α/α	Εξεταζόμενη Μέθοδος			
	Προς τα Εμπρός Ποιοτική Παλινδρόμηση		Προς τα Πίσω Ποιοτική Παλινδρόμηση	
	Συνδυασμός	Sum(sl)	Συνδυασμός	Sum(sl)
1	x1	1.48504	x1-x2-x3-x4	0.53775
2	x1-x2	0.59751	x1-x2-x3	0.60232
3	x1-x3	1.48845	x1-x2-x4	0.53409
4	x1-x4	0.64308	x1-x3-x4	0.646069
5	x1-x2-x3	0.60232	x2-x3-x4	0.85442
6	x1-x2-x4	0.53409	x1-x2	0.59751
7	x1-x2-x3-x4	0.53775	x1-x4	0.64308
8			x2-x4	1.30724
Προτεινόμενο καλύτερο υποσύνολο		x1-x2-x4	x1-x2-x4	

5.2.3.1.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Υλοποιώντας το γ.π. (4.8) για κάθε μία από τις τέσσερις ανεξάρτητες μεταβλητές των δεδομένων, διαπιστώνεται ότι η x1 δίνει το μικρότερο άθροισμα απλών απόλυτων σφαλμάτων και γίνεται η πρώτη μεταβλητή που θα εισαχθεί στο μοντέλο παλινδρόμησης (βλ. Πίνακα 5.61). Στο επόμενο βήμα εξετάζονται οι συνδυασμοί που περιέχουν την x1 και προκύπτει ότι ο συνδυασμός x1-x2 παρουσιάζει μικρότερο άθροισμα απόλυτων σφαλμάτων σε σχέση με τον προηγούμενο καλύτερο συνδυασμό, οπότε η x2 επιλέγεται να εισαχθεί στο μοντέλο. Στη συνέχεια εξετάζονται οι συνδυασμοί που περιλαμβάνουν τις x1-x2. Από αυτούς, ο συνδυασμός x1-x2-x4 παρουσιάζει ακόμη μικρότερο άθροισμα απόλυτων σφαλμάτων από τον προηγούμενο καλύτερο συνδυασμό, οπότε γίνεται ο νέος καλύτερος συνδυασμός και η x4 εισάγεται στο μοντέλο. Η διαδικασία τερματίζεται εδώ, καθώς ο επόμενος μοναδικός συνδυασμός για εξέταση, δε δίνει μικρότερο άθροισμα απόλυτων σφαλμάτων από αυτό του προηγούμενου βήματος. Καλύτερο προτεινόμενο υποσύνολο μεταβλητών, με βάση αυτή τη προσέγγιση, προκύπτει να είναι το x1-x2-x4, ενώ τα βήματα που ακολουθήθηκαν, φαίνονται στον Πίνακα 5.62.

Πίνακας 5.62: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα απλών απόλυτων σφαλμάτων, στα δεδομένα Hald.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	x1, x2, x3, x4	x1	1.485	x1
2	x1-x2, x1-x3, x1-x4	x1-x2	0.598	x2
3	x1-x2-x3, x1-x2-x4	x1-x2-x4	0.534	x4
4	x1-x2-x3-x4	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x4				

5.2.3.1.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Ο συνδυασμός ο οποίος περιλαμβάνει όλες τις διαθέσιμες ανεξάρτητες μεταβλητές, δίνει άθροισμα απόλυτων σφαλμάτων ίσο με 0.538. Στους συνδυασμούς που προκύπτουν στο επόμενο βήμα, εφαρμόζεται το γ.π. (4.8) για κάθε έναν ξεχωριστά και διαπιστώνεται ότι ο συνδυασμός x1-x2-x4 δίνει μικρότερο άθροισμα απλών απόλυτων σφαλμάτων από τον αρχικό. Άρα, η μεταβλητή x3 εξέρχεται από το μοντέλο και η διαδικασία συνεχίζεται με την εξέταση των επόμενων συνδυασμών. Η εξέταση αυτή δείχνει ότι δεν υπάρχει άλλος συνδυασμός ο οποίος να διαθέτει μικρότερο άθροισμα απόλυτων σφαλμάτων, οπότε σταματάει η διαδικασία εδώ και προτείνεται ο συνδυασμός x1-x2-x4 ως το καλύτερο υποσύνολο για συμμετοχή στο μοντέλο παλινδρόμησης. Όλα αυτά, συνοψίζονται στον Πίνακα 5.63.

Πίνακας 5.63: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα των απλών απόλυτων σφαλμάτων, στα δεδομένα Hald.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Απόλυτων Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3, x1-x2-x4, x1-x3-x4, x2-x3-x4	x1-x2-x4	0.534	x3
2	x1-x2, x1-x4, x2-x4	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1- x2-x4				

5.2.3.2 Άθροισμα Τετραγωνικών Σφαλμάτων

Και σε αυτή την περίπτωση, εξετάζονται τα αποτελέσματα που δίνουν οι προτεινόμενες μεθοδολογίες, με βάση αυτή τη φορά το γ.π. (4.12). Τα αποτελέσματα που προέκυψαν, παρουσιάζονται στον Πίνακα 5.64.

Πίνακας 5.64: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.12), στα δεδομένα Hald.

a/a	Εξεταζόμενη Μέθοδος			
	Προς τα Εμπρός Ποιοτική Παλινδρόμηση		Προς τα Πίσω Ποιοτική Παλινδρόμηση	
	Συνδυασμός	Sum(σ^2)	Συνδυασμός	Sum(σ^2)
1	x1	0.338462	x1-x2-x3-x4	0.045055
2	x1-x2	0.052167	x1-x2-x3	0.052701
3	x1-x3	0.339633	x1-x2-x4	0.044747
4	x1-x4	0.076316	x1-x3-x4	0.076546
5	x1-x2-x3	0.052701	x2-x3-x4	0.103232
6	x1-x2-x4	0.044747	x1-x2	0.052167
7	x1-x2-x3-x4	0.045055	x1-x4	0.076316
8			x2-x4	0.328136
Προτεινόμενο καλύτερο υποσύνολο		x1-x2-x4	x1-x2-x4	

5.2.3.2.1 Προς τα εμπρός βηματική ποιοτική παλινδρόμηση

Στο πρώτο βήμα αυτής της διαδικασίας, επιλέγεται για εισαγωγή στο μοντέλο παλινδρόμησης η μεταβλητή x1, καθώς παρουσιάζει το μικρότερο άθροισμα τετραγωνικών σφαλμάτων, βάσει του γ.π. (4.12). Ακολούθως, εξετάζονται με παρόμοιο τρόπο οι πιθανοί συνδυασμοί που σχηματίζονται περιέχοντας την x1 σε αυτούς. Το συμπέρασμα στο βήμα αυτό είναι ότι ο συνδυασμός x1-x2 δίνει μικρότερο άθροισμα τετραγωνικών σφαλμάτων από τον προηγούμενο καλύτερο συνδυασμό, οπότε η x2 εισάγεται στο μοντέλο. Τώρα πρέπει να εξεταστούν οι συνδυασμοί που περιέχουν τις x1-x2, ως προς το ίδιο κριτήριο. Η εξέταση αυτή δείχνει ότι υπάρχει συνδυασμός με μικρότερη τιμή στο εξεταζόμενο κριτήριο, σε σχέση με τον συνδυασμό του προηγούμενου βήματος, και αυτός είναι ο x1-x2-x4. Συνεπώς, η x4 επιλέγεται για εισαγωγή στο μοντέλο στο βήμα αυτό. Επόμενος συνδυασμός που απομένει για εξέταση, είναι αυτός που περιλαμβάνει όλες τις διαθέσιμες ανεξάρτητες μεταβλητές.

Πίνακας 5.65: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα εμπρός βηματική ποιοτική παλινδρόμηση, για το άθροισμα τετραγωνικών σφαλμάτων, στα δεδομένα Hald.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Είσοδος Μεταβλητής στο Μοντέλο Παλινδρόμησης
1	x1, x2, x3, x4	x1	0.338	x1
2	x1-x2, x1-x3, x1-x4	x1-x2	0.052	x2
3	x1-x2-x3, x1-x2-x4	x1-x2-x4	0.045	x4
4	x1-x2-x3-x4	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x4				

Η εφαρμογή, ωστόσο, του γ.π. (4.12) σε αυτόν, δείχνει ότι το άθροισμα των τετραγωνικών σφαλμάτων που παρουσιάζει, δεν είναι μικρότερο από το αντίστοιχο του προηγούμενου καλύτερου συνδυασμού. Άρα, η διαδικασία σταματά εδώ και το καλύτερο υποσύνολο μεταβλητών που προκύπτει με βάση αυτήν, είναι το $x_1-x_2-x_4$. Στον Πίνακα 5.65 παρουσιάζονται τα βήματα που ακολουθήθηκαν στην προσέγγιση αυτή.

5.2.3.2.2 Προς τα πίσω βηματική ποιοτική παλινδρόμηση

Ακολουθώντας αντίστροφη λογική από αυτή της προηγούμενης παραγράφου, εξετάζεται, με χρήση του γ.π. (4.12) το άθροισμα τετραγωνικών σφαλμάτων του συνδυασμού ο οποίος περιέχει όλες τις διαθέσιμες ανεξάρτητες μεταβλητές. Το άθροισμα αυτό προκύπτει ίσο με 0.045 και, στη συνέχεια, εξετάζονται όλοι οι δυνατοί συνδυασμοί που προκύπτουν με αφαίρεση μίας κάθε φορά μεταβλητής από τον αρχικό συνδυασμό. Ο συνδυασμός $x_1-x_2-x_4$ φαίνεται ότι έχει μικρότερο άθροισμα τετραγωνικών σφαλμάτων από τον προηγούμενο συνδυασμό, οπότε επιλέγεται ως ο καλύτερος για το βήμα αυτό. Η μεταβλητή x_3 εξέρχεται του μοντέλου και με βάση τον τελευταίο συνδυασμό, σχηματίζονται οι συνδυασμοί που θα εξεταστούν, με χρήση του γ.π. (4.12), στο επόμενο βήμα. Από αυτούς προκύπτει ότι κανένας δε διαθέτει μικρότερο άθροισμα τετραγωνικών σφαλμάτων, με αποτέλεσμα η διαδικασία να τερματιστεί στο σημείο αυτό και καλύτερο υποσύνολο μεταβλητών να υποδειχθεί το $x_1-x_2-x_4$. Στον Πίνακα 5.66 φαίνονται τα βήματα που ακολουθήθηκαν.

Πίνακας 5.66: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για το άθροισμα των τετραγωνικών σφαλμάτων, στα δεδομένα Hald.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	Άθροισμα Τετραγωνικών Σφαλμάτων Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	$x_1-x_2-x_3, x_1-x_2-x_4, x_1-x_3-x_4, x_2-x_3-x_4$	$x_1-x_2-x_4$	0.0447	x_3
2	$x_1-x_2, x_1-x_4, x_2-x_4$	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: $x_1-x_2-x_4$				

5.2.3.3 Συντελεστής προσδιορισμού

Όπως και στις προηγούμενες περιπτώσεις συνόλων δεδομένων που εξετάστηκαν, και εδώ επιλέχθηκε να ελαχιστοποιηθεί η έκφραση $\sum_{i=1}^n (\hat{y}_i - y_i)^2$, η οποία αντιστοιχεί στο άθροισμα των τετραγώνων των διαφορών των εκτιμώμενων τιμών για τις παρατηρήσεις της εξαρτημένης μεταβλητής Y από τις πραγματικές παρατηρούμενες τιμές αυτής. Η εξέταση των συνδυασμών μεταβλητών για το εν λόγω σύνολο δεδομένων, με βάση το γ.π. (4.17), δίνει τα αποτελέσματα του Πίνακα 5.67.

Πίνακας 5.67: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.17), στα δεδομένα Hald.

Εξεταζόμενος Συνδυασμός	Τιμή Αντικειμενικής Συνάρτησης $\sum_{i=1}^n (\hat{y}_i - y_i)^2$	$\sum_{i=1}^n (y_i - \bar{Y})^2$	Υπολογιζόμενος R^2
x1-x2-x3-x4	58.46	2715.76	97.847%
x1-x2-x3	99.87		96.323%
x1-x2-x4	58.03		97.863%
x1-x3-x4	81.41		97.002%
x2-x3-x4	125.62		95.374%
x1-x2	96.85		96.434%
x1-x4	80.98		97.018%
x2-x4	827.61		69.526%

5.2.3.3.1 Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση

Με βάση τον Πίνακα 5.67, προκύπτει ότι υπάρχει συνδυασμός ο οποίος να περιέχει μία μεταβλητή λιγότερη από τον αρχικό (ο οποίος περιέχει όλες τις διαθέσιμες μεταβλητές) και δίνει καλύτερη τιμή για τον συντελεστή πολλαπλής παλινδρόμησης σύμφωνα με το γ.π. (4.17). Συνεπώς, ο συνδυασμός αυτός, εν προκειμένω ο x1-x2-x4, είναι ο καλύτερος για το βήμα αυτό και η μεταβλητή που αφαιρείται από τον αρχικό συνδυασμό για να προκύψει αυτός (η x3), εξάγεται από το μοντέλο παλινδρόμησης. Στη συνέχεια αφαιρείται μία κάθε φορά μεταβλητή από τον καινούριο καλύτερο συνδυασμό και όλοι οι συνδυασμοί που προκύπτουν, εξετάζονται με βάση το γ.π. (4.17). Τα αποτελέσματα των συνδυασμών αυτών συγκρίνονται με το αποτέλεσμα του προηγούμενου βήματος, ώστε να εξεταστεί εάν υπάρχει νέος καλύτερος συνδυασμός. Στη συγκεκριμένη περίπτωση, δεν προκύπτει νέος καλύτερος συνδυασμός, οπότε η διαδικασία σταματά εδώ και καλύτερος συνδυασμός αναδεικνύεται ο x1-x2-x4. Τα βήματα που ακολουθήθηκαν, φαίνονται στον Πίνακα 5.68.

Πίνακας 5.68: Αποτελέσματα της προτεινόμενης μεθοδολογίας για την προς τα πίσω βηματική ποιοτική παλινδρόμηση, για τον συντελεστή πολλαπλής παλινδρόμησης, στα δεδομένα Hald.

Βήμα	Εξεταζόμενοι Συνδυασμοί	Καλύτερος Συνδυασμός	R^2 Καλύτερου Συνδυασμού	Έξοδος Μεταβλητής από το Μοντέλο Παλινδρόμησης
1	x1-x2-x3, x1-x2-x4, x1-x3-x4, x2-x3-x4	x1-x2-x4	97.86%	x3
2	x1-x2, x1-x3, x2-x3	-	-	-
Προτεινόμενο Καλύτερο Υποσύνολο Μεταβλητών: x1-x2-x4				

5.2.3.4 Σύγκριση Αποτελεσμάτων - Συμπεράσματα

Στην παράγραφο αυτή, θα συνοψισθούν τα κυριότερα συμπεράσματα τα οποία μπορούν να προκύψουν από την προηγούμενη ανάλυση στα δεδομένα Hald. Μια πρώτη σύγκριση των αποτελεσμάτων των εξεταζόμενων μεθοδολογιών, φαίνεται στον Πίνακα 5.69, ενώ στον Πίνακα A.10 (βλ. Παράρτημα), φαίνεται μια αναλυτική σύγκριση των αποτελεσμάτων για όλους τους δυνατούς συνδυασμούς μεταβλητών.

Πίνακας 5.69: Αποτελέσματα των χρησιμοποιούμενων προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Hald.

Εξεταζόμενος Συνδυασμός	Βηματική Παλινδρόμηση		Βηματική Ποιοτική Παλινδρόμηση			
	Εμπρός Επιλογή	Πίσω Απαλοιφή	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
	R^2	R^2	$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4	0.99957	0.99957	0.53775	0.92389	0.045055	0.971903
x1-x2-x3	0.99291	0.99291	0.60232	0.94421	0.052701	0.963971
x1-x2-x4	0.99847	0.99847	0.53409	0.93251	0.044747	0.972212
x1-x3-x4	0.97163	0.97163	0.646069	0.96981	0.076546	0.962824
x2-x3-x4	0.99084	0.99280	0.85442	0.88072	0.103232	0.944592
x1-x2	0.97423	0.97423	0.59751	0.93935	0.052167	0.964651
x1-x3	0.97163	0.97163	1.48845	0.49647	0.339633	0.487448
x1-x4	0.85391	0.85391	0.64308	0.96999	0.076316	0.962982
x2-x4	0.99084	0.99084	1.30724	0.70979	0.328136	0.670633
x1	0.72971	0.72971	1.48504	0.59224	0.338462	0.488046
x2	0.96139	0.96139	1.50082	0.57885	0.360673	0.64175
x3	0.70405	0.70405	2.22113	0.23307	0.723887	0.330544
x4	0.66147	0.66147	1.65020	0.68657	0.380352	0.649438

Όπως και για τα προηγούμενα σύνολα δεδομένων που εξετάστηκαν, και στην περίπτωση αυτή παρατηρούνται, κατά περίπτωση, μεγάλες αποκλίσεις μεταξύ του R^2 ο οποίος προκύπτει για τους αντίστοιχους εξεταζόμενους συνδυασμούς με βάση τη χρησιμοποίηση της κλασσικής μεθόδου βηματικής παλινδρόμησης σε σύγκριση με την προτεινόμενη μεθοδολογία βηματικής ποιοτικής παλινδρόμησης. Οι αποκλίσεις στα αποτελέσματα αυτά αμβλύνονται, εάν συγκριθούν μεταξύ τους τα αποτελέσματα που δίνει η χρησιμοποίηση των δύο διαφορετικών κριτηρίων (άθροισμα απλών και τετραγωνικών σφαλμάτων) στην ίδια (προτεινόμενη) μεθοδολογία. Μια ακόμη παρατήρηση η οποία μπορεί να γίνει, είναι ότι, για το συγκεκριμένο σύνολο δεδομένων, το κριτήριο του αθροίσματος των τετραγωνικών σφαλμάτων, φαίνεται να δίνει καλύτερα αποτελέσματα για τον συντελεστή πολλαπλής παλινδρόμησης για την συντριπτική πλειοψηφία των εξεταζόμενων συνδυασμών. Ακολουθώντας, εξάλλου, τα βήματα τα οποία προτείνονται στις παραγράφους 4.3.1.1 και 4.3.1.2, προκύπτει ότι, με βάση τα δύο αυτά κριτήρια, καταλήγουμε στο ίδιο τελικό προτεινόμενο καλύτερο υποσύνολο μεταβλητών.

Ο Πίνακας 5.70, συνοψίζει τα αποτελέσματα για τα τρία κριτήρια που εξετάζονται στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής ποιοτικής παλινδρόμησης.

Πίνακας 5.70: Αποτελέσματα των χρησιμοποιούμενων τριών κριτηρίων, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Hald.

Εξεταζόμενος Συνδυασμός	Βηματική Ποιοτική Παλινδρόμηση				
	R^2	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
		$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4	97.847%	0.53775	0.92389	0.04505	0.971903
x1-x2-x3	96.323%	0.60232	0.94421	0.05270	0.963971
x1-x2-x4	97.863%	0.53409	0.93251	0.04474	0.972212
x1-x3-x4	97.002%	0.64607	0.96981	0.07654	0.962824
x2-x3-x4	95.374%	0.85442	0.88072	0.10323	0.944592
x1-x2	96.434%	0.59751	0.93935	0.05216	0.964651
x1-x3	64.907%	1.48845	0.49647	0.33963	0.487448
x1-x4	97.018%	0.64308	0.96999	0.07631	0.962982
x2-x4	69.526%	1.30724	0.70979	0.32813	0.670633
x1	65.308%	1.48504	0.59224	0.33846	0.488046
x2	65.250%	1.50082	0.57885	0.36067	0.64175
x3	16.588%	2.22113	0.23307	0.72388	0.330544
x4	79.936%	1.65020	0.68657	0.38035	0.649438

Αυτό το οποίο παρατηρείται, είναι ότι η χρησιμοποίηση του συντελεστή πολλαπλής παλινδρόμησης ως κριτήριο βελτιστοποίησης στα πλαίσια της προτεινόμενης μεθοδολογίας, οδηγεί, σε γενικές γραμμές, σε καλύτερα ή παραπλήσια αποτελέσματα για κάθε συνδυασμό από τα αντίστοιχα κριτήρια του αθροίσματος των απόλυτων σφαλμάτων. Ακόμα, όμως, μια σημαντική παρατήρηση είναι ότι ακολουθώντας τις διαδικασίες που περιγράφονται στις ενότητες 4.3.1 και 4.3.2, όλες οι μεθοδολογίες για τα τρία κριτήρια υποδεικνύουν το ίδιο υποσύνολο ως το καλύτερο για συμμετοχή στο ζητούμενο μοντέλο παλινδρόμησης.

Τέλος, παρατηρώντας τους επιπλέον δείκτες που εξετάστηκαν κατά τη διάρκεια των διερευνήσεων, είναι δυνατόν να εξαχθούν κάποια επιπλέον συμπεράσματα, παρατηρώντας τις τιμές του Πίνακα 5.71. Ένα πρώτο τέτοιο συμπέρασμα είναι ότι το κριτήριο του αθροίσματος των τετραγωνικών σφαλμάτων δίνει σχετικά καλύτερα αποτελέσματα για τον δείκτη του συντελεστή πολλαπλής παλινδρόμησης σε σύγκριση με τα αντίστοιχα για το κριτήριο του αθροίσματος των απλών απόλυτων σφαλμάτων. Μια άλλη παρατήρηση για το συγκεκριμένο σύνολο δεδομένων, είναι ότι ο δείκτης του μέσου ποσοστιαίου απόλυτου σφάλματος (MAPE), δε προκύπτει κατά αναλογία με τον υπολογιζόμενο δείκτη R^2 . Υπάρχουν, δηλαδή, περιπτώσεις συνδυασμών, κατά τις οποίες ο R^2 προκύπτει καλύτερος για το κριτήριο των τετραγωνικών σφαλμάτων, αλλά το MAPE εμφανίζεται χειρότερο ή παραπλήσιο με το αντίστοιχο για το κριτήριο των απλών απόλυτων σφαλμάτων. Τέλος, στην πλειοψηφία των περιπτώσεων που εξετάστηκαν, οι υπόλοιποι δείκτες φαίνεται να είναι σε συμφωνία με τον δείκτη R^2 (όσο καλύτερος εμφανίζεται αυτός σε κάποια

περίπτωση εξεταζόμενου συνδυασμού, τόσο καλύτεροι εμφανίζονται και οι εν λόγω δείκτες).

Πίνακας 5.71: Αποτελέσματα των επιμέρους χρησιμοποιούμενων δεικτών, στα πλαίσια των προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Hald.

Εξεταζόμενος Συνδυασμός	Απλά Απόλυτα Σφάλματα				Τετραγωνικά Σφάλματα			
	R^2	Sum(e)	Sum(e^2)	MAPE	R^2	Sum(e)	Sum(e^2)	MAPE
x1-x2-x3-x4	0.92389	39.928	299.205	0.030	0.971903	27.546	95.256	0.022
x1-x2-x3	0.94421	35.892	175.202	0.028	0.963971	31.680	114.198	0.025
x1-x2-x4	0.93251	37.117	286.137	0.028	0.972212	27.358	94.307	0.022
x1-x3-x4	0.96981	21.254	98.373	0.018	0.962824	34.007	124.225	0.028
x2-x3-x4	0.88072	40.433	390.951	0.032	0.944592	43.044	204.883	0.035
x1-x2	0.93935	38.318	217.013	0.030	0.964651	31.396	112.490	0.025
x1-x3	0.49647	99.240	1908.902	0.087	0.487448	108.906	1802.537	0.094
x1-x4	0.96999	21.158	97.882	0.018	0.962982	33.954	123.654	0.028
x2-x3	0.83473	48.224	465.102	0.038	0.896476	49.620	327.185	0.040
x2-x4	0.70979	76.423	866.786	0.061	0.670633	92.076	949.614	0.073
x3-x4	0.89399	42.510	381.520	0.035	0.938268	47.813	223.741	0.040
x1	0.59224	86.110	1357.349	0.075	0.488046	108.835	1800.152	0.093
x2	0.57885	90.040	1369.861	0.069	0.64175	96.919	1039.598	0.076
x3	0.23307	152.643	2692.614	0.129	0.330544	135.731	2449.374	0.118
x4	0.68657	87.988	887.591	0.072	0.649438	98.176	980.324	0.079

ΚΕΦΑΛΑΙΟ

6

6.1 Συμπεράσματα

Στο κεφάλαιο αυτό θα συνοψισθούν τα κυριότερα συμπεράσματα τα οποία μπορούν να προκύψουν με βάση την ανάλυση η οποία προηγήθηκε σε όλα τα προηγούμενα κεφάλαια. Στο τέλος του κεφαλαίου, θα παρατεθούν μερικές σκέψεις-προτάσεις για τις προεκτάσεις ή/και βελτιώσεις οι οποίες μπορούν να λάβουν χώρα στο μέλλον σχετικά με όσα έχουν προταθεί στην παρούσα εργασία.

Στα προηγούμενα κεφάλαια, έγινε μια προσπάθεια να αναπτυχθεί το θεωρητικό υπόβαθρο μιας νέας προτεινόμενης μεθοδολογίας. Αυτή η μεθοδολογία ανήκει στην ευρύτερη κατηγορία μεθοδολογιών που ως στόχο έχουν την ανάδειξη και υπόδειξη (επιλογή) του καλύτερου υποσυνόλου ανεξάρτητων μεταβλητών για συμμετοχή σε ένα μοντέλο πολλαπλής παλινδρόμησης, με βάση τα διαθέσιμα δεδομένα που εξετάζονται. Πιο ειδικά, εξετάζεται εάν η συγκεκριμένη μεθοδολογία είναι σε θέση να εκμεταλλευτεί τις δυνατότητες της ποιοτικής παλινδρόμησης, ενσωματώνοντας τες σε μια ενιαία βηματική διαδικασία επιλογής ανεξάρτητων μεταβλητών, οι οποίες θα συγκροτήσουν το ζητούμενο μοντέλο παλινδρόμησης.

Το κύριο χαρακτηριστικό της προτεινόμενης μεθοδολογίας, είναι ότι ανάλογα με το πλήθος των διαθέσιμων ανεξάρτητων μεταβλητών για τα δεδομένα που εξετάζονται, αναπτύσσονται σε κάθε βήμα και τα ανάλογα γραμμικά προγράμματα, ώστε να εξεταστεί η «συνεισφορά» ή μη των εξεταζόμενων ανεξάρτητων μεταβλητών. Ένα άλλο χαρακτηριστικό της μεθοδολογίας αυτής είναι ότι, επειδή ακριβώς στηρίζεται στις αρχές της ποιοτικής παλινδρόμησης, η «συνεισφορά» μιας εξεταζόμενης ανεξάρτητης μεταβλητής υπολογίζεται με βάση την «αξία» της σε μια διαθέσιμη παρατήρηση, λόγω της ονομαστικής της τιμής, η οποία θα συνυπολογιστεί στον συνολικό προσδιορισμό της «αξίας» της εξαρτημένης μεταβλητής για την παρατήρηση αυτή.

Στα πλαίσια ανάπτυξης της υπό πρότασης μεθοδολογίας, δοκιμάστηκε η χρησιμοποίηση επιμέρους κριτηρίων βελτιστοποίησης ως αντικειμενικές συναρτήσεις στα υλοποιούμενα γ.π. Το γεγονός αυτό δίνει τη δυνατότητα επιμέρους συγκρίσεων, για το ποιο από αυτά κριτήρια αποδίδει καλύτερα με βάση τα όσα προτείνονται στα προηγούμενα. Στην παρούσα εργασία, χρησιμοποιήθηκαν τρία εναλλακτικά τέτοια κριτήρια: το άθροισμα των απλών απόλυτων σφαλμάτων, το άθροισμα των τετραγωνικών σφαλμάτων και ο συντελεστής προσδιορισμού. Ας σημειωθεί ότι, όπως ορίστηκαν τα εξεταζόμενα γ.π., τα παραπάνω απόλυτα σφάλματα αναφέρονται στην απόκλιση της τελικής «αξίας» της εξαρτημένης μεταβλητής από την «αξία» που προκύπτει συνυπολογίζοντας τις επιμέρους «αξίες» των ανεξάρτητων μεταβλητών που συμμετέχουν στο υλοποιούμενο γ.π. Παρόλα αυτά, χρησιμοποιήθηκαν και

επιμέρους δείκτες για την αξιολόγηση των αποτελεσμάτων που έδωσε η εφαρμογή της προτεινόμενης μεθοδολογίας.

Συνολικά, η προτεινόμενη μεθοδολογία εφαρμόστηκε σε πέντε διαφορετικά σύνολα δεδομένων. Για όλα τα εξεταζόμενα σύνολα δεδομένων, χρησιμοποιήθηκε μια τετραβάθμια κλίμακα διαχωρισμού για τον υπολογισμό και την κατασκευή των επιμέρους συναρτήσεων αξιών των διαθέσιμων μεταβλητών. Ο Πίνακας 6.1 συνοψίζει τα χαρακτηριστικά αυτών των συνόλων δεδομένων, ενώ ο Πίνακας 6.2 συγκεντρώνει τις επιμέρους κατανομές των διαθέσιμων, για κάθε σύνολο δεδομένων, μεταβλητών.

Πίνακας 6.1: Τα χαρακτηριστικά των χρησιμοποιηθέντων συνόλων δεδομένων.

Τίτλος Συνόλου Δεδομένων	Διαθέσιμες Παρατηρήσεις	Πλήθος Ανεξάρτητων Μεταβλητών
Water Usage of Production Plant	17	4
Solar Thermal Energy	29	5
Belle Ayr Liquefaction Runs	27	7
Body Fat of Females	20	3
Hald Cement	13	4

Πίνακας 6.2: Κατανομή των διαθέσιμων παρατηρήσεων κάθε συνόλου δεδομένων στην χρησιμοποιούμενη τετραβάθμια κλίμακα διαχωρισμού των δεδομένων.

Τίτλος Συνόλου Δεδομένων	Διάστημα Διαχωρισμού	Πλήθος Παρατηρήσεων των Διαθέσιμων Μεταβλητών σε Κάθε Διάστημα Διαχωρισμού							
		Y	x1	x2	x3	x4	x5	x6	x7
Water Usage of Production Plant	1 ^ο	5	6	5	5	5	-	-	-
	2 ^ο	6	5	6	9	6	-	-	-
	3 ^ο	6	6	6	3	6	-	-	-
Solar Thermal Energy	1 ^ο	10	9	10	10	10	10	-	-
	2 ^ο	9	10	10	10	9	10	-	-
	3 ^ο	10	10	9	9	10	9	-	-
Belle Ayr Liquefaction Runs	1 ^ο	9	8	11	9	9	9	9	9
	2 ^ο	9	9	13	9	9	9	9	9
	3 ^ο	9	10	3	9	9	9	9	9
Body Fat of Females	1 ^ο	6	7	6	6	-	-	-	-
	2 ^ο	7	6	7	7	-	-	-	-
	3 ^ο	7	7	7	7	-	-	-	-
Hald Cement	1 ^ο	5	4	5	3	4	-	-	-
	2 ^ο	4	4	4	5	5	-	-	-
	3 ^ο	4	5	4	5	4	-	-	-

Ο Πίνακας 6.3, συγκεντρώνει τα αποτελέσματα για όλους τους επιμέρους δείκτες αξιολόγησης που υιοθετήθηκαν κατά τη διεξαγωγή της διερεύνησης με βάση τις προτεινόμενες βηματικές μεθοδολογίες, για όλα τα σύνολα δεδομένων και για τον «καλύτερο» συνδυασμό μεταβλητών που αναδεικνύεται για το καθένα από αυτά. Το «καλύτερο» υποσύνολο μεταβλητών, σημειώνεται στον πίνακα αυτόν και για τις δύο διαδικασίες που ακολουθούνται για τις προτεινόμενες μεθοδολογίες, δηλαδή για την *Προς τα Εμπρός Βηματική Ποιοτική Παλινδρόμηση (ΕΒΠΠ)* και την *Προς τα Πίσω Βηματική Ποιοτική Παλινδρόμηση (ΠΒΠΠ)*.

Πίνακας 6.3: Συγκεντρωτικά αποτελέσματα για τα χρησιμοποιούμενα κριτήρια αξιολόγησης, σε όλα τα σύνολα δεδομένων και για το καλύτερο υποσύνολο μεταβλητών που υποδεικνύουν οι προτεινόμενες μεθοδολογίες για αυτά.

	Εξεταζόμενο Κριτήριο Αξιολόγησης	Εξεταζόμενο Σύνολο Δεδομένων				
		<i>Water Usage of Production Plant</i>	<i>Solar Thermal Energy</i>	<i>Belle Ayr Liquefaction Runs</i>	<i>Body Fat of Females</i>	<i>Hald Cement</i>
Άθροισμα Απλών Απλότων Σφαλμάτων Αξιών	Sum(σ)	1.351	1.254	3.117	1.923	0.534
	Sum(σ ²)	0.256	0.162	0.864	0.389	0.069
	Sum(el)	2207.559	132.262	279.757	38.777	37.117
	Sum(e ²)	820776.74	1681.557	6980.166	161.333	286.137
	MAPE	0.037	0.019	0.550	0.107	0.028
	R ²	0.750	0.892	0.504	0.684	0.933
	ΕΒΠΠ	x1-x2-x3	x1-x2-x3-x4	x1-x2-x3-x4-x6-x7	x2-x3	x1-x2-x4
	ΠΒΠΠ	x1-x2-x3	x1-x2-x3-x4	x1-x2-x3-x4-x6-x7	x2-x3	x1-x2-x4
Άθροισμα Τετραγωνικών Σφαλμάτων Αξιών	Sum(σ)	1.672	1.466	3.304	2.087	0.589
	Sum(σ ²)	0.237	0.126	0.755	0.344	0.045
	Sum(el)	3088.636	163.459	244.536	46.589	27.358
	Sum(e ²)	902722.82	1583.057	5188.698	169.948	94.307
	MAPE	0.054	0.023	0.502	0.124	0.022
	R ²	0.722	0.893	0.666	0.680	0.972
	ΕΒΠΠ	x1-x2-x3	x1-x2-x3-x4	x1-x2-x3-x4-x6-x7	x1-x2-x3	x1-x2-x4
	ΠΒΠΠ	x1-x2-x3	x1-x2-x3-x4	x1-x2-x3-x4-x6-x7	x1-x2-x3	x1-x2-x4
R ²	Sum(σ)	1.554	1.505	3.515	2.252	0.587
	Sum(σ ²)	0.340	0.173	1.446	0.950	0.075
	Sum(el)	2116.904	126.493	190.405	38.766	16.809
	Sum(e ²)	566760.063	1119.621	2007.645	104.201	58.033
	MAPE	0.036	0.018	0.356	0.108	0.014
	R ²	0.823	0.924	0.749	0.790	0.979
	ΠΒΠΠ	x1-x2-x3	x1-x2-x3-x4	x1-x2-x3-x4-x5-x6	x1-x2-x3	x1-x2-x4

Για το πρώτο σύνολο δεδομένων που εξετάζεται (Water Usage of Production Plant), η εφαρμογή των γ.π. (4.8), (4.12) και (4.17) δίνει τέτοια αποτελέσματα, ώστε εάν εφαρμοσθούν οι διαδικασίες που αναφέρονται στην ενότητα 4.3 της παρούσης, να προτείνεται το ίδιο υποσύνολο μεταβλητών ως το καλύτερο για συμμετοχή στο μοντέλο παλινδρόμησης. Αξίζει να σημειωθεί ότι όλες οι προτεινόμενες μεθοδολογίες, υποδεικνύουν τον ίδιο συνδυασμό ως τον «καλύτερο» για αυτό το σύνολο δεδομένων. Ωστόσο, όταν χρησιμοποιείται ως αντικειμενικό κριτήριο το άθροισμα των απόλυτων σφαλμάτων αξιών, αυτό πράγματι ελαχιστοποιείται στην περίπτωση των απλών απόλυτων σφαλμάτων, δηλαδή οι υπόλοιπες μεθοδολογίες παρουσιάζουν μεγαλύτερο άθροισμα απόλυτων σφαλμάτων αξιών για τον ίδιο συνδυασμό μεταβλητών. Κάτι αντίστοιχο συμβαίνει για τα κριτήρια που χρησιμοποιούνται κάθε φορά ως αντικειμενικά. Για το συγκεκριμένο σύνολο δεδομένων, οι μεθοδολογίες που χρησιμοποιούν το άθροισμα των σφαλμάτων αξιών (απλά ή τετραγωνικά), δίνουν ως αποτέλεσμα παραπλήσιο R^2 μεταξύ τους για το ίδιο υποσύνολο μεταβλητών (75% και 72.2% αντίστοιχα), ενώ υπάρχει σημαντική διαφοροποίηση σε ό,τι αφορά το άθροισμα των πρωτογενών σφαλμάτων (Sum(lcl)). Ένα τέτοιο γεγονός, ίσως υποδηλώνει μια ανταγωνιστική φύση μεταξύ των δύο αυτών ειδών σφαλμάτων. Όταν χρησιμοποιείται ο R^2 ως αντικειμενικό κριτήριο, ο τελευταίος παρουσιάζεται σαφώς βελτιωμένος (82.3%) έναντι των δύο άλλων μεθοδολογιών, για το ίδιο υποσύνολο μεταβλητών, όπως και το άθροισμα των πρωτογενών σφαλμάτων, όπως αναμενόταν, αλλά, ταυτόχρονα, παρουσιάζει ένα άθροισμα απλών απόλυτων σφαλμάτων αξιών μεταξύ των αντίστοιχων τιμών για τις δύο υπόλοιπες μεθοδολογίες. Τέλος, αξίζει να σημειωθεί ότι το MAPE υπολογίζεται παραπλήσιο όταν χρησιμοποιούνται ως αντικειμενικά κριτήρια το άθροισμα των απλών απόλυτων σφαλμάτων αξιών και ο R^2 (3.7% και 3.6% αντίστοιχα), ενώ στην περίπτωση του αθροίσματος των τετραγωνικών σφαλμάτων αξιών, εμφανίζεται λίγο χειρότερο (5.4%).

Στην περίπτωση του δεύτερου εξεταζόμενου συνόλου δεδομένων (Solar Thermal Energy), πάλι υποδεικνύεται ο ίδιος συνδυασμός μεταβλητών ως ο «καλύτερος» για όλες τις χρησιμοποιούμενες προτεινόμενες μεθοδολογίες. Όπως φαίνεται στον Πίνακα 6.3, το MAPE εμφανίζεται παραπλήσιο για όλες τις μεθοδολογίες (1.9%, 2.3% και 1.8%). Σημαντική διαφοροποίηση εμφανίζεται εδώ στις περιπτώσεις μεταξύ των αντικειμενικών κριτηρίων του αθροίσματος απλών και τετραγωνικών σφαλμάτων αξιών. Στην πρώτη περίπτωση (απλά σφάλματα αξιών), εμφανίζεται άθροισμα ίσο με 1.254, ενώ στη δεύτερη (τετραγωνικά σφάλματα αξιών) ίσο με 1.466. Ακόμα, το άθροισμα των πρωτογενών σφαλμάτων εμφανίζεται καλύτερο στην περίπτωση του κριτηρίου του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών σε σύγκριση με την περίπτωση του κριτηρίου των τετραγωνικών σφαλμάτων (132.262 έναντι 163.459 αντίστοιχα), ενώ όταν χρησιμοποιείται ως αντικειμενικό κριτήριο ο R^2 , το άθροισμα των πρωτογενών σφαλμάτων εμφανίζεται καλύτερο από αυτό για το κριτήριο των τετραγωνικών σφαλμάτων αξιών (126.493 και 163.459 αντίστοιχα). Αξίζει να σημειωθεί ότι τα αντίστροφα ακριβώς ισχύουν στην περίπτωση εξέτασης του πρώτου συνόλου δεδομένων, Water Usage of Production Plant. Ένα ακόμα γεγονός, το οποίο χρίζει προσοχής, είναι ότι ο R^2 παρουσιάζει την καλύτερη τιμή του στην περίπτωση εξέτασης του συντελεστή προσδιορισμού (92,4%, τιμή εμφανώς καλύτερη από τις αντίστοιχες για τα απλά απόλυτα και τετραγωνικά σφάλματα, δηλαδή 89,2% και 89,3%). Τέλος, το άθροισμα των πρωτογενών σφαλμάτων εμφανίζεται ελαφρώς χειρότερο στην περίπτωση των απλών απόλυτων σφαλμάτων, σε σύγκριση με το κριτήριο του R^2 .

Στο τρίτο σύνολο δεδομένων που εξετάζεται (Belle Ayr Liquefaction Runs), η εξέταση των αντικειμενικών κριτηρίων του αθροίσματος των απλών και τετραγωνικών σφαλμάτων αξιών, οδηγεί στην υπόδειξη του ίδιου υποσυνόλου μεταβλητών ως «καλύτερο» ($x_1-x_2-x_3-x_4-x_6-x_7$). Στην περίπτωση του R^2 , όμως, η εξέταση καταλήγει σε διαφορετικό υποσύνολο ($x_1-x_2-x_3-x_4-x_5-x_6$). Σε ό,τι αφορά τις δύο πρώτες μεθοδολογίες, το άθροισμα των απλών απόλυτων σφαλμάτων αξιών ελαχιστοποιείται για την περίπτωση εξέτασης του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών και, αντίστοιχα, το άθροισμα των τετραγωνικών σφαλμάτων αξιών ελαχιστοποιείται για την περίπτωση εξέτασης του αθροίσματος των τετραγωνικών σφαλμάτων αξιών. Το άθροισμα των πρωτογενών σφαλμάτων εμφανίζεται καλύτερο στην περίπτωση εξέτασης του αθροίσματος των τετραγωνικών σφαλμάτων αξιών και σε αυτή εμφανίζεται επίσης καλύτερος ο R^2 σε σχέση με το άθροισμα των απλών απόλυτων σφαλμάτων αξιών (66.6% έναντι 50.4%), ενώ και το MAPE εμφανίζεται καλύτερο στην περίπτωση αυτή (50.2% έναντι 55%), αλλά είναι αρκετά υψηλό και στις δύο περιπτώσεις. Όταν εξετάζεται η περίπτωση του R^2 ως αντικειμενικό κριτήριο, το άθροισμα των απλών απόλυτων σφαλμάτων αξιών εμφανίζεται χειρότερο από ό,τι στις δύο άλλες περιπτώσεις αλλά το άθροισμα των πρωτογενών σφαλμάτων εμφανίζεται καλύτερο στην περίπτωση εξέτασης του R^2 . Με βάση το γεγονός αυτό ίσως μπορεί να εξηγηθεί το γεγονός ότι ο τελευταίος εμφανίζεται σαφώς βελτιωμένος έναντι των αντίστοιχων περιπτώσεων στις υπόλοιπες εξεταζόμενες μεθοδολογίες (74.9% έναντι 50.4% στην περίπτωση του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών και 66.6% στην περίπτωση του αθροίσματος των τετραγωνικών σφαλμάτων αξιών). Τέλος, και το MAPE υπολογίζεται να είναι σαφώς καλύτερο σε σχέση με τις δύο άλλες μεθοδολογίες (35.6% έναντι 55% στην περίπτωση του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών και 50.2% στην περίπτωση του αθροίσματος των τετραγωνικών σφαλμάτων αξιών), παρουσιάζοντας, ωστόσο, και αυτό σχετικά υψηλή τιμή.

Το επόμενο σύνολο δεδομένων που εξετάστηκε, είναι το Body Fat of Females. Εδώ παρατηρείται το γεγονός να υποδεικνύεται διαφορετικό υποσύνολο μεταβλητών ως το καλύτερο, με βάση τα κριτήρια του αθροίσματος απλών (x_2-x_3) και τετραγωνικών ($x_1-x_2-x_3$) απόλυτων σφαλμάτων, ενώ χρησιμοποιώντας το κριτήριο του R^2 προτείνεται το ίδιο υποσύνολο με την περίπτωση των τετραγωνικών σφαλμάτων ($x_1-x_2-x_3$). Τα χαρακτηριστικά για το καλύτερο υποσύνολο βάση των διαφορετικών κριτηρίων που εξετάζονται, φαίνονται στον Πίνακα 6.3. Αυτό που μπορεί να διαπιστωθεί εδώ, είναι ότι για τον προτεινόμενο ως «καλύτερο» συνδυασμό μεταβλητών, το άθροισμα των απλών απόλυτων σφαλμάτων αξιών ελαχιστοποιείται για την περίπτωση εξέτασης του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών και, αντίστοιχα, το άθροισμα των τετραγωνικών σφαλμάτων αξιών ελαχιστοποιείται για την περίπτωση εξέτασης του αθροίσματος των τετραγωνικών σφαλμάτων αξιών. Το άθροισμα των πρωτογενών σφαλμάτων εμφανίζεται καλύτερο στην περίπτωση εξέτασης του R^2 , όπως, θεωρητικά, αναμενόταν, ενώ παραπλήσια τιμή παρουσιάζει για την περίπτωση εξέτασης των τετραγωνικών σφαλμάτων αξιών (38.777 και 38.766 αντίστοιχα). Το MAPE, ωστόσο, φαίνεται να είναι καλύτερο στην περίπτωση εξέτασης του R^2 , ενώ για την περίπτωση εξέτασης των απλών απόλυτων σφαλμάτων, η αντίστοιχη τιμή είναι παραπλήσια (0.108 και 0.107 αντίστοιχα). Τέλος, οι μεθοδολογίες εξέτασης του αθροίσματος απλών και τετραγωνικών σφαλμάτων, δίνουν παραπλήσιο αποτέλεσμα για την τιμή του R^2 (0.684 και 0.680 αντίστοιχα), ενώ ο τελευταίος εμφανίζεται σαφώς βελτιωμένος, στην περίπτωση που χρησιμοποιείται ως αντικειμενικό κριτήριο εξέτασης (0.790).

Τελευταίο σύνολο δεδομένων που εξετάστηκε, είναι το Hald Cement. Σε αυτό το σύνολο δεδομένων, όλες οι εξεταζόμενες προτεινόμενες μεθοδολογίες καταλήγουν στο ίδιο καλύτερο υποσύνολο μεταβλητών (x_1 - x_2 - x_4). Σε ό,τι αφορά τα επιμέρους χαρακτηριστικά, αυτά φαίνονται στον Πίνακα 6.3. Εύκολα διαπιστώνεται ότι το άθροισμα των απλών απόλυτων σφαλμάτων αξιών ελαχιστοποιείται για την περίπτωση εξέτασης του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών και, αντίστοιχα, το άθροισμα των τετραγωνικών σφαλμάτων αξιών ελαχιστοποιείται για την περίπτωση εξέτασης του αθροίσματος των τετραγωνικών σφαλμάτων αξιών. Ειδικότερα, αυτό είναι το σύνολο δεδομένων για το οποίο παρουσιάζεται το μικρότερο άθροισμα απόλυτων σφαλμάτων αξιών σε σύγκριση με όλα τα υπόλοιπα σύνολα δεδομένων που εξετάστηκαν. Το άθροισμα των πρωτογενών σφαλμάτων ελαχιστοποιείται για την περίπτωση εξέτασης του R^2 (58.033 έναντι 286.137 για την περίπτωση εξέτασης του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών και 94.307 για την περίπτωση εξέτασης του αθροίσματος των τετραγωνικών σφαλμάτων αξιών). Ο R^2 εμφανίζει τη μεγαλύτερη τιμή του όταν χρησιμοποιείται ως αντικειμενικό κριτήριο εξέτασης (0.979 έναντι 0.933 για την περίπτωση εξέτασης του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών και 0.972 για την περίπτωση εξέτασης του αθροίσματος των τετραγωνικών σφαλμάτων αξιών). Επιπλέον, η τιμή αυτή για τον R^2 , είναι η μεγαλύτερη σε σύγκριση με όλα τα υπόλοιπα σύνολα δεδομένων που εξετάστηκαν στην αντίστοιχη περίπτωση. Τέλος, το MAPE ελαχιστοποιείται στην περίπτωση εξέτασης του R^2 ως αντικειμενικό κριτήριο (0.014 έναντι 0.028 για την περίπτωση εξέτασης του αθροίσματος των απλών απόλυτων σφαλμάτων αξιών και 0.022 για την περίπτωση εξέτασης του αθροίσματος των τετραγωνικών σφαλμάτων αξιών), ενώ η τιμή αυτή είναι η μικρότερη σε σχέση με τις αντίστοιχες περιπτώσεις για όλα τα υπόλοιπα σύνολα δεδομένων που εξετάστηκαν.

Στο σημείο αυτό, μπορούν να γίνουν μερικές επιμέρους γενικές παρατηρήσεις. Το κύριο χαρακτηριστικό των δεδομένων για το Hald Cement, είναι ότι οι ανεξάρτητες μεταβλητές που υπάρχουν σε αυτό, παρουσιάζουν πολύ υψηλή συσχέτιση με την εξαρτημένη μεταβλητή, όπως φαίνεται στον Πίνακα 6.4. Αυτό το γεγονός ίσως εξηγεί τις «καλές» τιμές για τα επιμέρους κριτήρια αξιολόγησης. Ακόμα, το μικρό μέσο ποσοστιαίο απόλυτο σφάλμα (πρωτογενή σφάλματα) το οποίο παρατηρείται στο συγκεκριμένο σύνολο δεδομένων, ίσως δείχνει ότι οι τιμές αυτές δεν εξαρτώνται τόσο πολύ από τον αριθμό των ανεξάρτητων μεταβλητών, αλλά περισσότερο από τα ποιοτικά χαρακτηριστικά των δεδομένων, όπως είναι οι συσχετίσεις μεταξύ τους. Επιπρόσθετα, η εφαρμογή των τριών κλασσικών βηματικών μεθοδολογιών σε αυτά τα δεδομένα, έδειξε ότι για κάθε συνδυασμό ανεξάρτητων μεταβλητών, το αποτέλεσμα ήταν ένας πολύ υψηλός δείκτης R^2 (βλ. Πίνακα 5.60).

Πίνακας 6.4: Συσχετίσεις μεταξύ των μεταβλητών των δεδομένων Hald Cement.

	x_1	x_2	x_3	x_4	y
x_1	1				
x_2	0.229	1			
x_3	-0.824	-0.139	1		
x_4	-0.245	-0.973	0.03	1	
y	0.731	0.816	-0.535	0.821	1

Από την άλλη, φαίνεται ότι στην περίπτωση εξέτασης του συνόλου δεδομένων Belle Ayr Liquefaction Runs λαμβάνονται σχετικά χαμηλές επιδόσεις στα κριτήρια αξιολόγησης. Το σύνολο αυτό περιλαμβάνει τον μεγαλύτερο αριθμό

διαθέσιμων ανεξάρτητων μεταβλητών και τον δεύτερο μεγαλύτερο αριθμό διαθέσιμων παρατηρήσεων. Ταυτόχρονα, διαθέτει και το χειρότερο MAPE (ο δείκτης αυτός είναι υψηλότερος και για τα υπόλοιπα αντικειμενικά κριτήρια που εξετάζονται για το συγκεκριμένο σύνολο δεδομένων), κάτι που σημαίνει ότι τα ποιοτικά χαρακτηριστικά των δεδομένων που εξετάζονται στην περίπτωση αυτή, οδηγούν στα παραπάνω αποτελέσματα. Οι συσχετίσεις μεταξύ των δεδομένων του Belle Ayr Liquefaction Runs δεν εμφανίζουν ούτε ακραία μεγάλες, ούτε ακραία μικρές διαφορές μεταξύ τους. Ωστόσο, παρατηρώντας τον Πίνακα 5.30, φαίνεται ότι για τη μεταβλητή x_2 , τα δεδομένα για αυτήν είναι τέτοια ώστε να επαναλαμβάνονται σε πολλές από τις διαθέσιμες παρατηρήσεις. Ειδικότερα, το γεγονός αυτό οδηγεί στη μη ισομερή, κατά το δυνατόν, κατανομή των δεδομένων αυτών, στην τετραβάθμια κλίμακα διαχωρισμού που χρησιμοποιήθηκε στα πλαίσια της ποιοτικής βηματικής παλινδρόμησης. Ίσως το γεγονός αυτό επιδρά αρνητικά στη διακριτική ευχέρεια της προτεινόμενης μεθοδολογίας, καθώς στα υπόλοιπα σύνολα δεδομένων, δεν παρατηρείται τέτοιο φαινόμενο, εκτός από το σύνολο δεδομένων Water Usage of Production Plant, στο οποίο, με βάση τον Πίνακα 6.2, παρατηρείται επίσης μια σχετικά ανισομερή κατανομή των δεδομένων στα διαστήματα διαχωρισμού στην περίπτωση της μεταβλητής x_3 . Εκτός από την προφανή ανισομερή κατανομή των διαθέσιμων δεδομένων στις περιπτώσεις που αναφέρθηκαν, το άλλο κοινό το οποίο είναι δυνατόν να παρατηρήσει κανείς, είναι ότι οι μεταβλητές για τις οποίες παρατηρείται το φαινόμενο αυτό, παρουσιάζουν και μια πολύ μικρή σχετικά συσχέτιση με την εξαρτημένη μεταβλητή ($\rho_{Yx3} = -0.089$ για το σύνολο δεδομένων Water Usage of Production Plant και $\rho_{Yx2} = 0.159$ για το σύνολο δεδομένων Belle Ayr Liquefaction Runs). Από την άλλη, στην περίπτωση του συνόλου δεδομένων Hald, τα διαθέσιμα δεδομένα για την x_3 παρουσιάζουν και αυτά μια μικρή, σχετικά, ανισομερή κατανομή στα διαστήματα διαχωρισμού, αλλά ισχύει $\rho_{Yx2} = 0.535$. Το γεγονός αυτό ίσως υποδηλώνει ότι η ανισομερή κατανομή των διαθέσιμων δεδομένων στα διαστήματα διαχωρισμού της χρησιμοποιούμενης κλίμακας, οδηγεί σε μείωση της ακρίβειας της διακριτικής ευχέρειας της μεθοδολογίας, ειδικά όταν η συσχέτιση της μεταβλητής, για την οποία παρατηρείται η ανισομέρεια αυτή της κατανομής, με την εξαρτημένη μεταβλητή, είναι σχετικά χαμηλή. Παρόλα αυτά, στην περίπτωση εξέτασης του συνόλου δεδομένων Body Fat of Females, η κατανομή των δεδομένων στα διαστήματα διαχωρισμού είναι ισομερής, αλλά οι εξεταζόμενες μεθοδολογίες, δεν καταλήγουν στο ίδιο προτεινόμενο υποσύνολο μεταβλητών. Στο συγκεκριμένο σύνολο δεδομένων, είναι διαθέσιμες τρεις ανεξάρτητες μεταβλητές, εκ των οποίων οι δύο έχουν υψηλή συσχέτιση με την εξαρτημένη μεταβλητή ($\rho_{Yx1} = 0.843$ και $\rho_{Yx2} = 0.878$), ενώ η τρίτη διαθέσιμη ανεξάρτητη μεταβλητή, παρουσιάζει σχετικά χαμηλή συσχέτιση με την εξαρτημένη μεταβλητή ($\rho_{Yx3} = 0.142$). Τα αποτελέσματα ίσως να ήταν διαφορετικά, εάν και η τρίτη ανεξάρτητη μεταβλητή παρουσίαζε υψηλότερη συσχέτιση με την εξαρτημένη μεταβλητή ή εάν υπήρχαν διαθέσιμες περισσότερες ανεξάρτητες μεταβλητές με ενδιάμεσες τιμές συσχέτισης με την εξαρτημένη μεταβλητή.

Μετά την εξέταση των συνόλων δεδομένων και τα αποτελέσματα τα οποία παρουσιάστηκαν, φαίνεται ότι κανένα κριτήριο από αυτά που εξετάστηκαν δε δίνει τα καλύτερα αποτελέσματα για όλους τους δείκτες, σε όλα τα σύνολα δεδομένων. Επιπλέον, υπάρχουν και περιπτώσεις κατά τις οποίες οι προτεινόμενες μεθοδολογίες δεν καταλήγουν στο ίδιο καλύτερο υποσύνολο μεταβλητών. Αυτό, όμως, το οποίο αξίζει να επισημανθεί, είναι ότι, όπως παρατηρήθηκε στην ανάλυση του πέμπτου κεφαλαίου της παρούσης, μπορεί να υποδειχθεί ένα υποσύνολο ως το καλύτερο με

βάση ένα από τα κριτήρια που εξετάζονται, αλλά εξετάζοντας τους επιμέρους δείκτες που παρουσιάστηκαν, να προτείνεται ένα διαφορετικό υποσύνολο μεταβλητών. Αυτό, για παράδειγμα, συμβαίνει στην περίπτωση του συνόλου δεδομένων Body Fat of Females. Πιο αναλυτικά, για το συγκεκριμένο σύνολο δεδομένων, εάν ληφθεί ως κριτήριο βελτιστοποίησης το άθροισμα των απλών απόλυτων σφαλμάτων (γ.π.(4.8)), λαμβάνονται τα αποτελέσματα του Πίνακα 6.5. Σε αυτόν τον πίνακα, φαίνεται ότι από τους συνδυασμούς που εξετάζονται, καλύτερο αποτέλεσμα για το εν λόγω κριτήριο, εμφανίζει ο συνδυασμός x2-x3. Παρόλα αυτά, εξετάζοντας τους επιμέρους δείκτες του Πίνακα 6.5, παρατηρούμε ότι υπάρχουν συνδυασμοί με καλύτερες επιδόσεις σε αυτούς. Για παράδειγμα, ο συνδυασμός x1-x2 παρουσιάζει καλύτερο R^2 και ταυτόχρονα διαθέτει και μικρότερο MAPE από τον προτεινόμενο καλύτερο συνδυασμό, ενώ ο συνδυασμός x1-x2-x3 παρουσιάζει καλύτερο R^2 , αλλά δεν υπερτερεί σε κανέναν άλλο δείκτη από τον προτεινόμενο καλύτερο συνδυασμό. Επομένως, εγείρεται ένα ζήτημα σχετικά με το ποιο κριτήριο μπορεί να θεωρηθεί ότι δίδει τα πιο αξιόπιστα αποτελέσματα.

Πίνακας 6.5: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Body Fat of Females.

Συνδυασμός	R^2	Sum(e)	Sum(σ^2)	Sum(e)	Sum(e^2)	MAPE
x1-x2-x3	0.68484	1.92439	0.38842	38.78044	161.04685	0.10695
x1-x2	0.71398	2.02053	0.74836	40.87315	155.30146	0.104544
x1-x3	0.65657	2.46032	0.97105	45.90043	180.19121	0.127054
x2-x3	0.68433	1.92304	0.38850	38.77718	161.33294	0.106949
x1	0.69705	2.45523	0.97305	43.56803	160.59881	0.122291
x2	0.67690	2.04633	0.81767	45.72661	161.98678	0.134024
x3	0.04882	4.21008	2.57254	92.61833	673.07976	0.26614

Η παραπάνω διαπίστωση δίνει την αφορμή για ένα επιπλέον ερώτημα σχετικά με την επιλογή του «καλύτερου» κριτηρίου. Και αυτό είναι εάν τα κριτήρια που έχουν χρησιμοποιηθεί στις προτεινόμενες μεθοδολογίες, είναι ανταγωνιστικά μεταξύ τους. Για παράδειγμα, ένα ερώτημα είναι εάν δίνοντας έμφαση στο κριτήριο του R^2 το άθροισμα των απλών ή τετραγωνικών σφαλμάτων εκτινάσσεται σε υπερβολικά μεγάλα επίπεδα, κάτι το οποίο, ίσως, καθιστά την προσπάθεια πρόβλεψης, μέσω ενός τέτοιου μοντέλου παλινδρόμησης, μη αξιόπιστη. Αντίστροφα, επιχειρώντας έναν όσο το δυνατόν μεγαλύτερο περιορισμό του αθροίσματος των απλών/τετραγωνικών σφαλμάτων, μήπως οδηγούμαστε σε χειρότερες τιμές για τον R^2 ; Συνεπώς, το ζήτημα που τίθεται, είναι μήπως θα ήταν χρήσιμη μια έρευνα για τη «συμπεριφορά» αυτών των κριτηρίων. Μια τέτοια έρευνα, πιθανώς να οδηγούσε σε ενδείξεις για το εάν τα εν λόγω κριτήρια μπορούν να χρησιμοποιηθούν στα πλαίσια ενός γ.π. ποιοτικής παλινδρόμησης με περισσότερες της μίας αντικειμενικές συναρτήσεις, οι οποίες μπορεί να είναι και ανταγωνιστικές μεταξύ τους (επιλογή με πολλαπλά κριτήρια).

Στο σημείο αυτό αξίζει να επισημανθεί ότι στις περιπτώσεις των συνδυασμών μεταβλητών που εξετάστηκαν, σε κάθε σύνολο δεδομένων, με βάση το άθροισμα των απόλυτων σφαλμάτων (απλών/τετραγωνικών), χρησιμοποιήθηκαν τα γ.π. (4.8) και (4.12), τα οποία αναπτύχθηκαν σε περιβάλλον Excel 2003. Για την επίλυσή τους, χρησιμοποιήθηκε ο ενσωματωμένος επίλυτής (Solver) του συγκεκριμένου λογιστικού πακέτου. Ο χρόνος επίλυσης για κάθε εξεταζόμενο συνδυασμό, σε κάθε περίπτωση συνόλου δεδομένων, κατά μέσο όρο δεν ξεπερνούσε τα 60 δευτερόλεπτα. Ο χρόνος αυτός, μπορούσε να είναι ακόμη λιγότερος, αναλόγως των αρχικών συνθηκών για τις

μεταβλητές απόφασης του προβλήματος. Στην περίπτωση που η εξέταση πραγματοποιήθηκε για το γ.π. (4.17), χρησιμοποιήθηκε το λογισμικό Lingo 8.0. Ο χρόνος επίλυσης σε αυτές τις περιπτώσεις, αυξήθηκε δραματικά, καθώς το πρόβλημα ανάχθηκε στην επίλυση επιμέρους μικτών ακέραιων μη γραμμικών γ.π. Στην πλειονοψηφία των συνδυασμών που εξετάστηκαν σε όλες τις περιπτώσεις των συνόλων δεδομένων που χρησιμοποιήθηκαν, παρατηρήθηκε ότι η επίλυση ήταν ικανή να εντοπίσει μια τοπική βέλτιστη λύση, αλλά ο χρόνος επίλυσης μέχρι και εκείνη τη στιγμή, μπορούσε να ξεπερνά ακόμη και τις 48 ώρες.

Με βάση τις παραπάνω διαπιστώσεις, φαίνεται ότι η προτεινόμενη βηματική μεθοδολογία ποιοτικής παλινδρόμησης, είναι σε θέση να οδηγήσει σε χρήσιμα συμπεράσματα, σε ό,τι αφορά την επιλογή του «καλύτερου» υποσυνόλου μεταβλητών για συμμετοχή σε ένα μοντέλο παλινδρόμησης. Πιο συγκεκριμένα, η διαμόρφωση και επίλυση του κατάλληλου γ.π. για κάθε βήμα εξέτασης ενός συνόλου δεδομένων, φαίνεται να αξιοποιεί τις ιδιαιτερότητες και τις διαφοροποιήσεις των κάθε φορά χρησιμοποιούμενων δεδομένων. Το γεγονός αυτό οδηγεί σε διαφορετικά διακριτά αποτελέσματα κάθε φορά, τα οποία είναι άμεσα συγκρίσιμα και μπορούν να αποτελέσουν έναν, επίσης συγκεκριμένο, «οδηγό» για την επιλογή ή μη συγκεκριμένων εξεταζόμενων μεταβλητών. Επιπλέον, από την προηγούμενη ανάλυση στο πέμπτο κεφάλαιο, γίνεται φανερό ότι η όλη ακολουθούμενη διαδικασία, καταλήγει σε κάποιο αποτέλεσμα, χωρίς να είναι απαραίτητος ο εξαντλητικός έλεγχος όλων των δυνατών συνδυασμών μεταβλητών που είναι διαθέσιμοι για ένα σύνολο δεδομένων.

Ωστόσο, από την ανάλυση που πραγματοποιήθηκε στην ενότητα 5.1 (δεδομένα διερεύνησης), είναι δυνατόν να παρατηρηθούν μεγάλες αποκλίσεις στις τιμές που λαμβάνουν τα κριτήρια αξιολόγησης που υιοθετήθηκαν για την εξέταση των επιμέρους συνδυασμών μεταβλητών. Πιο συγκεκριμένα, αυτό που παρατηρείται, είναι μεγάλες, σχετικές, αποκλίσεις στις τιμές των κριτηρίων αξιολόγησης ανάμεσα σε συνδυασμούς μεταβλητών οι οποίοι περιέχουν μία μόνο μεταβλητή και σε αυτούς που περιέχουν περισσότερες των δύο μεταβλητών. Οι αντίστοιχες αποκλίσεις, όμως, μεταξύ των συνδυασμών που περιέχουν τον ίδιο αριθμό μεταβλητών, δεν είναι τόσο μεγάλες. Ειδικότερα, όσο αυξάνει ο αριθμός των μεταβλητών που συμμετέχουν σε έναν εξεταζόμενο συνδυασμό, οι αποκλίσεις αυτές είναι συγκρίσιμες μεταξύ τους, επειδή πιθανόν να αξιοποιείται μεγαλύτερο της «πληροφορίας» (ιδιαιτερότητες στις συσχετίσεις με την εξαρτημένη μεταβλητή και στην κατανομή των διαθέσιμων δεδομένων στην κλίμακα διαχωρισμού) που περιέχεται στα υπό εξέταση δεδομένα. Αξίζει να σημειωθεί ότι οι παραπάνω διαπιστώσεις, ισχύουν και στην περίπτωση της εξέτασης των δεδομένων ελέγχου (ενότητα 5.2). Τίθεται, επομένως, ένα ερώτημα, σχετικά με τη χρησιμότητα των δύο προσεγγίσεων (προς τα εμπρός και προς τα πίσω διαδικασία): είναι δυνατόν να προκύψει κάποιο συμπέρασμα αναφορικά για το ποια προσέγγιση είναι προτιμότερο να χρησιμοποιηθεί; Εάν χρησιμοποιηθεί η προς τα πίσω διαδικασία, τότε είναι πιθανότερο να αξιοποιείται περισσότερη «πληροφορία», όπως αναφέρθηκε πριν, από το σύνολο δεδομένων και αφαιρώντας «υποψήφιας» ανεξάρτητες μεταβλητές η διαδικασία αυτή ίσως φτάσει σε ένα προτεινόμενο υποσύνολο με την εξέταση λίγων συνδυασμών (εάν τελικά είναι θέμα ξιοποίησης «πληροφορίας»). Από την άλλη, χρησιμοποιώντας την προς τα εμπρός διαδικασία, αυτή θα συνεχιστεί μέχρι να αξιοποιηθεί η απαιτούμενη «πληροφορία». Στην πορεία αυτή, ίσως είναι απαραίτητη η εξέταση περισσότερων συνδυασμών από ό,τι στην προς τα πίσω διαδικασία, κάτι που σημαίνει ότι απαιτείται περισσότερος

υπολογιστικός χρόνος, αλλά η εξέταση περισσότερων συνδυασμών, μπορεί να συνδεθεί με περισσότερη «αξιοπιστία» στα εξαγόμενα αποτελέσματα. Η εμπειρία από την εξέταση τόσο των δεδομένων διερεύνησης (ενότητα 5.1, 2 σύνολα δεδομένων), όσο και των δεδομένων ελέγχου (ενότητα 5.2, 3 σύνολα δεδομένων), δείχνει ότι, πράγματι, η προς τα πίσω διαδικασία φθάνει σε κάποιο αποτέλεσμα σε λιγότερα βήματα από ό,τι η προς τα εμπρός διαδικασία, για το ίδιο σύνολο δεδομένων.

6.2 Μελλοντικές Προεκτάσεις

Στην ενότητα αυτή θα προταθούν μερικές μελλοντικές προεκτάσεις τις οποίες μπορεί να λάβει η προτεινόμενη μεθοδολογία, με βάση τις παρατηρήσεις που έγιναν στην προηγούμενη ανάλυση. Στα πλαίσια εξέτασης όλων των συνόλων δεδομένων που χρησιμοποιήθηκαν στο πέμπτο κεφάλαιο, υιοθετήθηκαν τρία κριτήρια αξιολόγησης ως αντικειμενικά (ως αντικειμενική συνάρτηση στα υλοποιούμενα γ.π.). εξετάστηκαν, ωστόσο, και μερικά επιμέρους κριτήρια αξιολόγησης για τον έλεγχο των αποτελεσμάτων που προέκυψαν. Από την ανάλυση η οποία προηγήθηκε, προκύπτει ότι όταν χρησιμοποιείται ως αντικειμενικό κριτήριο ο R^2 , αυτός υπολογίζεται να είναι μεγαλύτερος από ό,τι προκύπτει για τη χρήση του αθροίσματος των απλών και τετραγωνικών σφαλμάτων αξιών, για το ίδιο σύνολο δεδομένων. Κάτι μπορεί να υποδηλώνει την ανταγωνιστική φύση των κριτηρίων αυτών και, κατ' επέκταση την ανταγωνιστική φύση μεταξύ των σφαλμάτων αξιών και των πρωτογενών σφαλμάτων. Εάν χρησιμοποιηθεί το άθροισμα των σφαλμάτων αξιών ως αντικειμενικό (για να ελαχιστοποιηθεί), το άθροισμα των πρωτογενών σφαλμάτων θα αυξηθεί. Αντίθετα, εάν χρησιμοποιηθεί το άθροισμα των πρωτογενών σφαλμάτων ως αντικειμενικό (για να ελαχιστοποιηθεί), το άθροισμα των σφαλμάτων αξιών θα αυξηθεί. Συνεπώς, τίθεται το ερώτημα, ποιο από τα δύο είδη σφαλμάτων είναι προτιμότερο να χρησιμοποιηθεί ως αντικειμενικό; Δεν είναι εύκολο να δοθεί μια σαφής απάντηση στο ερώτημα αυτό, οπότε αυτό ανάγεται στο ερώτημα «έως ποιο επίπεδο είναι ανεκτή η αύξηση του αθροίσματος του ενός είδους σφαλμάτων, ώστε να επιτευχθεί ελαχιστοποίηση του αθροίσματος του άλλου είδους σφαλμάτων». Η προσέγγιση του περιορισμού του αθροίσματος σφαλμάτων αξιών χρησιμοποιήθηκε στην περίπτωση χρήσης του R^2 ως αντικειμενικού κριτηρίου, περιορίζοντας το εύρος του αθροίσματος των απόλυτων σφαλμάτων αξιών στο 10-20% του βέλτιστου αθροίσματος που υπολογίστηκε με χρήση του τελευταίου ως αντικειμενικό κριτήριο, για κάθε εξεταζόμενο συνδυασμό μεταβλητών. Επομένως, στο μέλλον ίσως είναι πρακτικό να διερευνηθεί εάν υπάρχει ένα «βέλτιστο» επίπεδο «περιορισμού» του αθροίσματος του ενός είδους σφαλμάτων έναντι του άλλου είδους. Μια άλλη προσέγγιση η οποία μπορεί να ακολουθηθεί, είναι αυτή της χρήσης πολύ-αντικειμενικών γ.π., τα οποία θα ελαχιστοποιούν ταυτόχρονα το άθροισμα και των δύο ειδών σφαλμάτων που αναφέρθηκαν, λαμβάνοντας υπόψη ότι η πολυπλοκότητα των γ.π. που θα πρέπει να αναπτυχθούν, ίσως αυξηθεί κατά μεγάλο βαθμό.

Ενδιαφέρον θα παρουσίαζε και η εξέταση του MAPE ως αντικειμενικό κριτήριο. Η υιοθέτησή του ως κριτήριο αξιολόγησης στα σύνολα δεδομένων που εξετάστηκαν, οδήγησε σε χρήσιμα συμπεράσματα, κάτι που σημαίνει ότι η αξιοποίησή του ως αντικειμενικό κριτήριο μπορεί να οδηγήσει σε ανάλογα συμπεράσματα. Θα μπορούσε κάποιος να ισχυριστεί, ωστόσο, ότι η ελαχιστοποίηση του MAPE είναι παρόμοια περίπτωση με αυτή της ελαχιστοποίησης του αθροίσματος των πρωτογενών σφαλμάτων. Μόνο που στην περίπτωση του MAPE, λαμβάνεται

υπόψη ο αριθμός των μεταβλητών που μετέχουν στον υπό εξέταση συνδυασμό. Τα γ.π. τα οποία θα πρέπει να υλοποιηθούν σε αυτή την περίπτωση, είναι μη γραμμικά, όπως στην περίπτωση του αθροίσματος των τετραγωνικών σφαλμάτων αξιών που εξετάστηκαν στην προηγούμενη ανάλυση.

Τα σύνολα δεδομένων που εξετάστηκαν στο πέμπτο κεφάλαιο, χαρακτηρίζονται από σχετικά μικρό αριθμό διαθέσιμων παρατηρήσεων και ανεξάρτητων μεταβλητών. Αυτό έγινε για να διαπιστωθεί η λειτουργικότητα ή μη της προτεινόμενης μεθοδολογίας. Τα πραγματικά προβλήματα με τα οποία έρχονται αντιμέτωποι διάφοροι οργανισμοί και επιχειρήσεις, είναι, στις περισσότερες περιπτώσεις, πολύ μεγαλύτερα. Άρα, θα ήταν πρακτικό να εξεταστεί στο μέλλον η «συμπεριφορά» της προτεινόμενης μεθοδολογίας σε κάποια αντιπροσωπευτικά προβλήματα αυτού του είδους. Βέβαια, θα πρέπει να ληφθεί υπόψη το γεγονός ότι ο χρόνος επίλυση των μικτών ακέραιων μη γραμμικών γ.π. που αναπτύχθηκαν στα πλαίσια εξέτασης των πέντε συνόλων δεδομένων που εξετάστηκαν, ήταν χρονοβόρα, λόγω της πολυπλοκότητάς τους. Αυτό μπορεί να σημαίνει ότι αυξάνοντας τον αριθμό των δεδομένων που εξετάζονται, ίσως αυξηθεί και ο χρόνος επίλυσης του προβλήματος. Επομένως, θα ήταν χρήσιμο να εξεταστεί ο τρόπος με τον οποίο είναι (εάν είναι) δυνατόν να ελαττωθεί ο απαιτούμενος χρόνος επίλυσης, με τη χρήση των κατάλληλων τεχνικών επίλυσης. Στον τομέα αυτό, θα μπορούσε να εξεταστεί η χρησιμοποίηση προηγμένων τεχνικών επίλυσης προβλημάτων βελτιστοποίησης μικτού ακέραιου μη γραμμικού προγραμματισμού, όπως είναι για παράδειγμα τα *νευρωνικά δίκτυα*.

Η εξέταση η οποία πραγματοποιήθηκε για όλα τα σύνολα δεδομένων, έγινε υλοποιώντας τα κατάλληλα γ.π. σε κάθε εξεταζόμενο βήμα. Θα ήταν, ωστόσο, πρακτικής σημασίας η μελλοντική εξέταση ενσωμάτωσης της διαδικασίας αυτής σε ένα εξειδικευμένο λογισμικό, ώστε η τελευταία να είναι προσιτή και σε μη εξοικειωμένους χρήστες.

Βιβλιογραφία

- André D. S. C., Narula C.S., Elian N. S. and Tavares A. R. (2003), An overview of the variable selection methods for the minimum sum of absolute errors regression, *Statistics in Medicine*, 22, pp. 2101-2111.
- Berk, K. N. (1978), Comparing subset regression procedures, *Technometrics*, 20, pp. 1-6.
- Draper N.R and Smith H. (1981), *Applied regression analysis*, John Wiley and Sons, New York.
- Eksioglou B., Demirer R. and Capar I. (2005), Subset selection in multiple linear regression: a new mathematical approach, *Computers & Industrial Engineering*, 49, pp. 155-167.
- Fuchs J-J. and Maria S. (2007), A new approach to variable selection using the TLS approach, *IEEE Transactions on Signal Processing*, 55 (1), pp. 10-19.
- George, E.I. (2000), The variable selection problem, *Journal of the American Statistical Association*, 72 (358), pp. 341-354.
- Grechanovsky E. (1987), Stepwise Regression procedures: overview, problems, results, and suggestions, *Annals of New York Academy of Sciences*, 491, pp. 197-232.
- Kashid D.N. and Kulkarni S.R. (2002), A more general criterion for subset selection in multiple linear regression, *Communication in Statistics – Theory and Methods*, 31 (5), pp. 795-811.
- Konno H. and Yamamoto R. (2009), Choosing the best of variables in regression analysis using integer programming, *J. Glob. Optim.*, 44, pp. 273-282.
- Mandel, J. (1989), Some thoughts on variable-selection in multiple regression, *Journal of Quality Technology*, 21 (1), pp. 2-6.
- Miller, A. (2002), *Subset Selection in regression*, 2nd Edition, Chapman and Hall/CRC, Florida.
- Miller, A. J. (2000), Another look at subset selection using linear least square, *Communication in Statistics – Theory and Methods*, 29 (9-10), pp. 2005-2017.
- Montgomery D.C. and Peck E.A. (1992), *Introduction to linear regression analysis*, , John Wiley and Sons inc, New York.
- Salt W.D., Subhash A., Crichton R, and Livingstone J.D. (2007), An improved approximation to the estimation of the critical F values in best subset regression, 43, pp. 143-149.
- Γρηγορούδης Ε. και Σίσκος Ι. (2005), Ποιότητα υπηρεσιών και μέτρηση ικανοποίησης του πελάτη – το σύστημα MUSA, 2^η Έκδοση, Εκδόσεις Νέων Τεχνολογιών, Αθήνα.
- Xu Lu, Zhang Wen-Jun (2001), Comparison of different methods for variable selection, *Analytica Chimica Acta*, 446, pp. 477-483.

Yamashita T, Yamashita K. and Kamimura R. (2007), A stepwise AIC method for variable selection in linear regression, Communications in Statistics – Theory and Methods, 36 (13), pp. 2395-2403.

ΠΑΡΑΡΤΗΜΑ

Αναλυτικά αποτελέσματα εφαρμογής στα σύνολα δεδομένων ελέγχου

Πίνακας Α.1: Αποτελέσματα για την εφαρμογή της κλασσικής βηματικής παλινδρόμησης, με τη χρήση του πακέτου SPSS v.12, στα δεδομένα Belle Ayr Liquefaction Runs.

Συνδυασμός	Βηματική Παλινδρόμηση			Εμπρός Επιλογή			Πίσω Απαλοιφή		
	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης
x1-x2-x3-x4-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2-x3-x4-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x2-x3-x4-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x1-x2-x3-x4-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2-x3-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2-x4-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x3-x4-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x2-x3-x4-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2-x3-x4-x5	0.826	0.813	12.553	0.826	0.813	12.553	0.844	0.825	12.133
x1-x2-x3-x4-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x2-x3-x4-x7	0.834	0.813	12.533	0.834	0.813	12.533	0.834	0.813	12.533
x1-x2-x3-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x2-x3-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x1-x2-x3-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2-x4-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x2-x4-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x1-x2-x4-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x3-x4-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x3-x4-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x1-x3-x4-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x3-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x4-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x2-x3-x4-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x2-x3-x4-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x2-x3-x4-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x2-x3-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x2-x4-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x3-x4-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2-x3-x4	0.813	0.798	13.024	0.813	0.798	13.024	0.813	0.798	13.024
x1-x2-x3-x5	0.826	0.813	12.553	0.826	0.813	12.553	0.844	0.825	12.133
x1-x2-x3-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x2-x3-x7	0.834	0.813	12.533	0.834	0.813	12.533	0.834	0.813	12.533
x1-x2-x4-x5	0.826	0.813	12.553	0.826	0.813	12.553	0.844	0.825	12.133
x1-x2-x4-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x2-x4-x7	0.834	0.813	12.533	0.834	0.813	12.533	0.834	0.813	12.533
x1-x2-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661

Πίνακας Α.1 (συνέχεια)

Συνδυασμός	Βηματική Παλινδρόμηση			Εμπρός Επιλογή			Πίσω Απαλοιφή		
	R^2	$\bar{R}^2$	Τυπικό	R^2	$\bar{R}^2$	Τυπικό	R^2	$\bar{R}^2$	Τυπικό
			Σφάλμα Εκτίμησης			Σφάλμα Εκτίμησης			Σφάλμα Εκτίμησης
x1-x2-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x1-x2-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x3-x4-x5	0.843	0.823	12.200	0.843	0.823	12.200	0.843	0.823	12.200
x1-x3-x4-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x3-x4-x7	0.809	0.785	13.451	0.809	0.785	13.451	0.809	0.785	13.451
x1-x3-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x3-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x1-x3-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x4-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x4-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x1-x4-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x2-x3-x4-x5	0.826	0.813	12.553	0.826	0.813	12.553	0.843	0.823	12.200
x2-x3-x4-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x2-x3-x4-x7	0.812	0.789	13.318	0.812	0.789	13.318	0.816	0.794	13.175
x2-x3-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x2-x3-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x2-x3-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x2-x4-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x2-x4-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x2-x4-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x2-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x3-x4-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x3-x4-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x3-x4-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x3-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x4-x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2-x3	0.813	0.798	13.024	0.813	0.798	13.024	0.813	0.798	13.024
x1-x2-x4	0.813	0.798	13.024	0.813	0.798	13.024	0.813	0.798	13.024
x1-x2-x5	0.826	0.813	12.553	0.826	0.813	12.553	0.844	0.825	12.133
x1-x2-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x2-x7	0.834	0.813	12.533	0.834	0.813	12.533	0.834	0.813	12.533
x1-x3-x4	0.809	0.785	13.451	0.809	0.785	13.451	0.809	0.785	13.451
x1-x3-x5	0.822	0.808	12.719	0.822	0.808	12.719	0.822	0.808	12.719
x1-x3-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x3-x7	0.790	0.764	14.084	0.790	0.764	14.084	0.790	0.764	14.084
x1-x4-x5	0.794	0.786	13.405	0.794	0.786	13.405	0.794	0.786	13.405
x1-x4-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x4-x7	0.546	0.528	19.919	0.546	0.528	19.919	0.546	0.528	19.919
x1-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x1-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x2-x3-x4	0.786	0.769	13.926	0.786	0.769	13.926	0.786	0.769	13.926
x2-x3-x5	0.826	0.813	12.553	0.826	0.813	12.553	0.826	0.813	12.553
x2-x3-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x2-x3-x7	0.816	0.794	13.175	0.816	0.794	13.175	0.816	0.794	13.175

Πίνακας Α.1 (συνέχεια)

Συνδυασμός	Βηματική Παλινδρόμηση			Εμπρός Επιλογή			Πίσω Απαλοιφή		
	R^2	$\bar{R}^2$	Τυπικό	R^2	$\bar{R}^2$	Τυπικό	R^2	$\bar{R}^2$	Τυπικό
			Σφάλμα Εκτίμησης			Σφάλμα Εκτίμησης			Σφάλμα Εκτίμησης
x2-x4-x5	0.826	0.813	12.553	0.826	0.813	12.553	0.826	0.813	12.553
x2-x4-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x2-x4-x7	0.812	0.789	13.318	0.812	0.789	13.318	0.812	0.789	13.318
x2-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x2-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x2-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x3-x4-x5	0.843	0.823	12.200	0.843	0.823	12.200	0.843	0.823	12.200
x3-x4-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x3-x4-x7	0.791	0.764	14.073	0.791	0.764	14.073	0.791	0.764	14.073
x3-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x3-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x3-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x4-x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x4-x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x4-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x5-x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1-x2	0.813	0.798	13.024	0.813	0.798	13.024	0.813	0.798	13.024
x1-x3	0.763	0.744	14.683	0.763	0.744	14.683	0.763	0.744	14.683
x1-x4	0.313	0.287	24.486	0.313	0.287	24.486	0.313	0.287	24.486
x1-x5	0.794	0.786	13.405	0.794	0.786	13.405	0.794	0.786	13.405
x1-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x1-x7	0.201	0.170	26.418	0.546	0.528	19.919	0.546	0.528	19.919
x2-x3	0.778	0.760	14.193	0.778	0.760	14.193	0.778	0.760	14.193
x2-x4	0.786	0.769	13.926	0.786	0.769	13.926	0.786	0.769	13.926
x2-x5	0.826	0.813	12.553	0.826	0.813	12.553	0.826	0.813	12.553
x2-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x2-x7	0.714	0.691	16.107	0.714	0.691	16.107	0.714	0.691	16.107
x3-x4	0.765	0.747	14.595	0.765	0.747	14.595	0.765	0.747	14.595
x3-x5	0.822	0.808	12.719	0.822	0.808	12.719	0.822	0.808	12.719
x3-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x3-x7	0.667	0.640	17.399	0.667	0.640	17.399	0.667	0.640	17.399
x4-x5	0.794	0.786	13.405	0.794	0.786	13.405	0.794	0.786	13.405
x4-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x4-x7	0.546	0.528	19.919	0.546	0.528	19.919	0.546	0.528	19.919
x5-x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x5-x7	0.847	0.834	11.802	0.847	0.834	11.802	0.847	0.834	11.802
x6-x7	0.905	0.898	9.276	0.905	0.898	9.276	0.905	0.898	9.276
x1	0.201	0.170	26.418	0.201	0.170	26.418	0.201	0.170	26.418
x2	0.670	0.657	16.973	0.670	0.657	16.973	0.670	0.657	16.973
x3	0.596	0.580	18.788	0.596	0.580	18.788	0.596	0.580	18.788
x4	0.313	0.287	24.486	0.313	0.287	24.486	0.313	0.287	24.486
x5	0.794	0.786	13.405	0.794	0.786	13.405	0.794	0.786	13.405
x6	0.870	0.865	10.661	0.870	0.865	10.661	0.870	0.865	10.661
x7	0.546	0.528	19.919	0.546	0.528	19.919	0.546	0.528	19.919

Πίνακας Α.2: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Belle Ayr Liquefaction Runs.

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x3-x4-x5-x6-x7	0.49701	3.11873	0.86791	283.547	7183.544	0.557613
x1-x2-x3-x4-x5-x6	0.59934	3.19531	1.00563	305.1954	7799.146	0.605128
x1-x2-x3-x4-x5-x7	0.59298	3.27795	0.91834	291.3447	8036.468	0.609466
x1-x2-x3-x4-x6-x7	0.50445	3.11732	0.86362	279.7574	6980.166	0.549938
x1-x2-x3-x5-x6-x7	0.54482	3.15584	0.88566	247.4534	6362.269	0.515743
x1-x2-x4-x5-x6-x7	0.56621	3.12898	0.83021	257.7866	5899.511	0.49986
x1-x3-x4-x5-x6-x7	0.48426	3.19929	0.87624	302.5423	8523.173	0.592474
x2-x3-x4-x5-x6-x7	0.49163	3.13906	0.87450	281.8723	7281.066	0.557245
x1-x2-x3-x4-x5	0.69936	3.48694	1.28233	194.501	3750.733	0.403258
x1-x2-x3-x4-x6	0.60163	3.19403	0.99659	301.4193	7663.703	0.600615
x1-x2-x3-x4-x7	0.55450	3.27661	1.03219	264.573	6523.691	0.557422
x1-x2-x3-x5-x6	0.62766	3.33701	0.84568	253.5837	5981.971	0.522613
x1-x2-x3-x5-x7	0.58825	3.30924	1.01165	240.3604	5993.434	0.528498
x1-x2-x3-x6-x7	0.54905	3.15389	0.88478	244.6823	6226.701	0.510863
x1-x2-x4-x5-x6	0.63641	3.19618	1.02212	290.3812	7090.844	0.572603
x1-x2-x4-x5-x7	0.56279	3.34082	1.06618	271.8243	6632.082	0.542476
x1-x2-x4-x6-x7	0.56753	3.12696	0.82974	256.0569	5825.108	0.496267
x1-x2-x5-x6-x7	0.59747	3.17084	0.85643	242.5506	5401.673	0.476857
x1-x3-x4-x5-x6	0.51132	3.27333	1.18388	413.6323	13894.56	0.76541
x1-x3-x4-x5-x7	0.43226	3.36337	1.13860	341.8341	12079.45	0.690628
x1-x3-x4-x6-x7	0.52063	3.22818	0.86610	296.8123	7889.982	0.574842
x1-x3-x5-x6-x7	0.50991	3.23547	0.91684	279.1332	7882.249	0.566555
x1-x4-x5-x6-x7	0.45489	3.17048	0.91050	302.5719	8590.448	0.600668
x2-x3-x4-x5-x6	0.59008	3.21824	1.00615	302.8989	7933.402	0.606134
x2-x3-x4-x5-x7	0.40207	3.57558	1.07272	296.0357	8030.121	0.574608
x2-x3-x4-x6-x7	0.49836	3.13825	0.87141	278.4731	7097.344	0.550413
x2-x3-x5-x6-x7	0.53746	3.17692	0.89555	252.9477	6470.131	0.522499
x2-x4-x5-x6-x7	0.52184	3.14987	0.87387	275.7882	6825.092	0.539834
x3-x4-x5-x6-x7	0.46150	3.22379	0.94170	318.4	9338.872	0.625328
x1-x2-x3-x4	0.60922	3.34258	1.10495	282.7228	7195.498	0.586912
x1-x2-x3-x5	0.65308	3.46880	0.92617	252.8597	5666.634	0.533325
x1-x2-x3-x6	0.62726	3.33349	0.84620	253.9137	5993.945	0.523146

Πίνακας Α.2: (συνέχεια)

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x3-x7	0.58355	3.30752	1.02315	239.9909	5972.772	0.526556
x1-x2-x4-x5	0.67173	3.41596	1.14956	241.9436	5090.604	0.505252
x1-x2-x4-x6	0.64205	3.19488	1.01107	284.9234	6850.973	0.563365
x1-x2-x4-x7	0.57263	3.33983	1.04225	255	5687.713	0.513013
x1-x2-x5-x6	0.69226	3.55032	1.46286	246.6498	5048.74	0.520902
x1-x2-x5-x7	0.56799	3.34091	1.07509	268.854	6472.667	0.536877
x1-x2-x6-x7	0.59851	3.16964	0.85553	241.1712	5340.486	0.474254
x1-x3-x4-x5	0.50026	3.44755	1.27731	415.601	15195.67	0.796147
x1-x3-x4-x6	0.50943	3.31589	1.26907	430.373	14323.38	0.779732
x1-x3-x4-x7	0.50123	3.41546	1.11878	300.2031	8285.75	0.605766
x1-x3-x5-x6	0.48473	3.49092	1.20195	424.495	15868.52	0.801537
x1-x3-x5-x7	0.49865	3.38898	1.10140	293.3844	8385.174	0.602652
x1-x3-x6-x7	0.51563	3.26812	0.91970	288.2512	8192.086	0.575231
x1-x4-x5-x6	0.53289	3.27346	1.18458	404.0185	13645.93	0.757134
x1-x4-x5-x7	0.52432	3.39645	1.05960	286.3418	7470.621	0.579278
x1-x4-x6-x7	0.52394	3.22732	0.86679	295.5084	7835.219	0.573411
x1-x5-x6-x7	0.54003	3.24695	0.91087	276.4106	7267.774	0.54877
x2-x3-x4-x5	0.45206	3.64977	1.19432	326.9732	9121.29	0.625574
x2-x3-x4-x6	0.60059	3.21757	0.99852	291.7137	7694.9	0.587465
x2-x3-x4-x7	0.52101	4.20439	1.43638	258.7111	4976.864	0.543084
x2-x3-x5-x6	0.59726	3.38708	0.86949	267.1437	6255.515	0.54522
x2-x3-x5-x7	0.37298	3.61443	1.10939	292.9293	7890.38	0.556342
x2-x3-x6-x7	0.54124	3.17539	0.89484	250.8131	6348.479	0.518232
x2-x4-x5-x6	0.62158	3.21944	1.02556	290.6065	7380.124	0.578066
x2-x4-x5-x7	0.43499	3.64184	1.09535	284.2012	6876.766	0.519748
x2-x4-x6-x7	0.52701	3.14836	0.86966	272.4981	6658.904	0.533355
x2-x5-x6-x7	0.56666	3.20192	0.88197	254.6258	5963.317	0.500429
x3-x4-x5-x6	0.49656	3.29811	1.18290	409.0299	14152.21	0.764829
x3-x4-x5-x7	0.32656	3.67822	1.15925	367.264	13548.86	0.708007
x3-x4-x6-x7	0.49444	3.25624	0.93016	318.1283	8712.951	0.614812
x3-x5-x6-x7	0.48593	3.27006	0.93820	287.7404	8469.567	0.590667
x4-x5-x6-x7	0.47542	3.22399	0.94477	314.1689	9316.718	0.623689
x1-x2-x3	0.65363	3.46633	0.92635	253.0956	5666.727	0.533533

Πίνακας Α.2: (συνέχεια)

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x4	0.67275	3.41371	1.14994	242.3787	5090.729	0.505502
x1-x2-x5	0.69046	3.55070	1.46019	247.8435	5066.995	0.531405
x1-x2-x6	0.69246	3.547	1.46246	247.1515	5051.286	0.522491
x1-x2-x7	0.58214	3.34282	1.07539	250.9501	5557.221	0.50543
x1-x3-x4	0.52251	3.53324	1.46102	417.3681	14073.74	0.780083
x1-x3-x5	0.49708	3.67875	1.21900	424.0146	16069.91	0.818825
x1-x3-x6	0.50429	3.56241	1.20976	444.8471	16732.24	0.819218
x1-x3-x7	0.50667	3.43374	1.13257	297.3508	8335.52	0.604007
x1-x4-x5	0.52887	3.47308	1.42155	393.0715	13352.55	0.757827
x1-x4-x6	0.52997	3.31521	1.26971	420.2281	14054.13	0.770432
x1-x4-x7	0.52358	3.43502	1.17936	298.1026	7838.992	0.595631
x1-x5-x6	0.59374	3.86437	1.91811	358.539	11069.94	0.666942
x1-x5-x7	0.53449	3.40670	1.08832	293.3812	8062.713	0.582842
x1-x6-x7	0.54675	3.27973	0.91060	284.6787	7505.332	0.555146
x2-x3-x4	0.31540	4.44210	1.46178	363.7924	8671.844	0.761668
x2-x3-x5	0.53128	3.93303	1.14447	266.6146	6295.977	0.555463
x2-x3-x6	0.59707	3.38407	0.87007	267.434	6268.27	0.545677
x2-x3-x7	0.46829	4.54047	2.21176	263.4683	4799.004	0.488145
x2-x4-x5	0.44591	3.70940	1.32377	350.4913	10469.79	0.658148
x2-x4-x6	0.60670	3.21847	1.01702	300.0466	7357.166	0.589753
x2-x4-x7	0.52109	4.20354	1.44760	256.8005	5067.529	0.535676
x2-x5-x6	0.67595	3.64325	1.57878	258.4659	5463.511	0.545779
x2-x5-x7	0.47102	3.72630	1.30307	258.8117	5615.708	0.44892
x2-x6-x7	0.56794	3.20087	0.88123	253.2355	5897.039	0.497851
x3-x4-x5	0.37461	3.77694	1.41710	441.8965	16158.5	0.801931
x3-x4-x6	0.51399	3.32666	1.55478	373.5663	11849.28	0.705216
x3-x4-x7	0.53084	4.43187	1.73653	278.1913	5969.591	0.540483
x3-x5-x6	0.47635	3.53791	1.20575	427.6689	16290.99	0.811537
x3-x5-x7	0.31893	3.73840	1.16728	377.0046	13449.43	0.712806
x3-x6-x7	0.48925	3.30677	0.94510	294.4723	8592.348	0.596889
x4-x5-x6	0.51808	3.29840	1.18380	399.1246	13925.69	0.756438
x4-x5-x7	0.35170	3.70684	1.21786	344.898	11504.29	0.65786
x4-x6-x7	0.49860	3.25530	0.93027	316.8372	8653.529	0.613044

Πίνακας Α.2: (συνέχεια)

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x5-x6-x7	0.51549	3.28865	0.93528	284.9029	7820.824	0.570704
x1-x2	0.68970	3.54828	1.45845	248.6257	5080.119	0.536988
x1-x3	0.50822	3.81117	1.64531	440.5096	16536.56	0.842585
x1-x4	0.54478	3.53339	1.46270	406.9338	13810.81	0.770819
x1-x5	0.55083	4.03801	1.70573	400.387	13663.57	0.788001
x1-x6	0.64800	3.86277	1.93618	283.0497	7770.043	0.512606
x1-x7	0.53921	3.45350	1.12406	294.3099	7632.801	0.585551
x2-x3	0.45808	4.65350	2.13505	271.9036	5053.241	0.483016
x2-x4	0.31101	4.43919	1.45869	375.2499	8835.941	0.781795
x2-x5	0.54576	4.14387	2.73017	250.5973	4634.921	0.488322
x2-x6	0.67531	3.64066	1.58027	259.7598	5477.921	0.552556
x2-x7	0.20599	4.86909	3.55017	375.2263	9108.656	0.808503
x3-x4	0.17175	4.71610	2.78604	319.1562	8857.707	0.470558
x3-x5	0.35951	4.15912	1.43247	468.4482	18054.54	0.862464
x3-x6	0.49606	3.61351	1.21436	448.0383	17158.64	0.829219
x3-x7	0.38882	5.30425	1.95828	281.7801	5934.763	0.456826
x4-x5	0.39495	3.77802	1.42835	424.1509	15252.37	0.776988
x4-x6	0.51484	3.32231	1.57932	351.3646	10140.28	0.655112
x4-x7	0.53129	4.43013	1.73953	279.321	6028.038	0.54367
x5-x6	0.54623	3.86405	1.93667	314.4658	8941.487	0.57222
x5-x7	0.39015	3.87396	1.28070	332.3755	10323.91	0.614341
x6-x7	0.52256	3.32279	0.93869	292.1134	8034.785	0.576819
x1	0.57340	4.09589	1.58867	397.7369	13531.3	0.774347
x2	0.18339	4.86666	3.56251	385.1081	9910.051	0.918524
x3	0.15077	6.06654	3.44098	614.8883	23402.78	1.378574
x4	0.48728	4.71240	2.79502	266.902	4767.2	0.534851
x5	0.47529	4.52862	2.75144	299.9551	8842.064	0.541265
x6	0.64046	3.86132	1.94357	278.8556	7542.508	0.519583
x7	0.12158	6.26476	3.79208	380.3215	9212.977	0.730252

Πίνακας Α.3: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.12), στα δεδομένα Belle Ayr Liquefaction Runs.

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x3-x4-x5-x6-x7	0.66509	3.31006	0.75478	244.2531	5203.534	0.502114
x1-x2-x3-x4-x5-x6	0.67042	3.42821	0.79353	248.879	4990.179	0.49334
x1-x2-x3-x4-x5-x7	0.60822	3.50259	0.79239	270.6803	6294.832	0.54654
x1-x2-x3-x4-x6-x7	0.66625	3.30364	0.75461	244.5364	5188.698	0.501648
x1-x2-x3-x5-x6-x7	0.65142	3.32187	0.76618	250.1483	5636.951	0.519721
x1-x2-x4-x5-x6-x7	0.67338	3.29904	0.75647	243.9601	5339.324	0.502126
x1-x3-x4-x5-x6-x7	0.63573	3.47306	0.79480	263.8689	5758.424	0.521597
x2-x3-x4-x5-x6-x7	0.64642	3.34845	0.77565	249.683	5476.791	0.515339
x1-x2-x3-x4-x5	0.60455	3.62993	0.83875	278.5977	5951.473	0.544023
x1-x2-x3-x4-x6	0.67108	3.42454	0.79314	248.6534	4978.021	0.492657
x1-x2-x3-x4-x7	0.64918	3.50476	0.85229	254.1823	5418.485	0.525235
x1-x2-x3-x5-x6	0.64279	3.55492	0.80717	266.6912	5556.645	0.533975
x1-x2-x3-x5-x7	0.59447	3.56406	0.81475	273.7271	6593.239	0.5583
x1-x2-x3-x6-x7	0.65682	3.31267	0.76614	247.554	5551.869	0.513311
x1-x2-x4-x5-x6	0.67482	3.36845	0.79970	255.717	5233.952	0.505585
x1-x2-x4-x5-x7	0.62487	3.47815	0.79640	267.8254	6195.094	0.539818
x1-x2-x4-x6-x7	0.67506	3.28983	0.75635	244.4168	5323.704	0.501596
x1-x2-x5-x6-x7	0.67401	3.38705	0.78906	230.8214	5155.189	0.481385
x1-x3-x4-x5-x6	0.59874	3.73490	0.89221	274.152	6534.792	0.522171
x1-x3-x4-x5-x7	0.57625	3.62723	0.82236	285.4702	7159.101	0.560741
x1-x3-x4-x6-x7	0.65079	3.49576	0.79797	262.8119	5438.573	0.512085
x1-x3-x5-x6-x7	0.62674	3.48790	0.80986	262.893	6065.845	0.527616
x1-x4-x5-x6-x7	0.61176	3.45993	0.78263	259.4536	5705.248	0.528023
x2-x3-x4-x5-x6	0.65240	3.50168	0.81439	262.0506	5355.252	0.521036
x2-x3-x4-x5-x7	0.42652	3.95402	0.95267	301.6959	7616.133	0.590924
x2-x3-x4-x6-x7	0.64671	3.34691	0.77556	249.8973	5474.716	0.515366
x2-x3-x5-x6-x7	0.64447	3.36633	0.78671	247.0726	5655.849	0.519123
x2-x4-x5-x6-x7	0.65711	3.33305	0.77821	249.6831	5628.924	0.515745
x3-x4-x5-x6-x7	0.61064	3.50801	0.81771	269.6046	6193.032	0.540101
x1-x2-x3-x4	0.65518	3.58271	0.87505	255.4836	5132.489	0.517052
x1-x2-x3-x5	0.57668	3.70505	0.86454	284.0191	6380.094	0.565444
x1-x2-x3-x6	0.64330	3.55279	0.80671	266.7845	5549.283	0.533874

Πίνακας Α.3: (συνέχεια)

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x3-x7	0.64624	3.56298	0.87642	255.8761	5541.347	0.537412
x1-x2-x4-x5	0.62459	3.51835	0.84567	276.0867	6127.003	0.538557
x1-x2-x4-x6	0.67543	3.36612	0.79936	255.7608	5227.331	0.505351
x1-x2-x4-x7	0.66763	3.47475	0.85845	251.9691	5482.738	0.521607
x1-x2-x5-x6	0.69969	3.68868	0.90043	233.5381	4566.693	0.473884
x1-x2-x5-x7	0.60414	3.64593	0.85008	265.9823	6280.604	0.532582
x1-x2-x6-x7	0.67825	3.36463	0.78956	232.6736	5113.972	0.481341
x1-x3-x4-x5	0.55131	3.84083	0.91948	301.3554	7429.772	0.558029
x1-x3-x4-x6	0.61220	3.75322	0.89952	276.7801	6394.991	0.518077
x1-x3-x4-x7	0.64612	3.75168	0.90965	259.7071	5258.498	0.521643
x1-x3-x5-x6	0.56157	3.87106	0.93649	295.9927	7503.361	0.558229
x1-x3-x5-x7	0.55886	3.72141	0.84560	294.4641	7798.713	0.584123
x1-x3-x6-x7	0.64259	3.51564	0.81400	265.5935	5732.453	0.521206
x1-x4-x5-x6	0.62209	3.71480	0.89598	272.4129	6595.651	0.534647
x1-x4-x5-x7	0.60050	3.61408	0.82610	286.2421	6854.468	0.559721
x1-x4-x6-x7	0.66785	3.47519	0.79927	264.7334	5349.518	0.515119
x1-x5-x6-x7	0.64785	3.50788	0.82262	265.9011	5732.016	0.518616
x2-x3-x4-x5	0.44116	4.05991	0.99840	288.7844	7124.951	0.567194
x2-x3-x4-x6	0.65304	3.49928	0.81412	261.8987	5339.656	0.52048
x2-x3-x4-x7	0.41800	4.60995	1.27378	317.2541	5968.227	0.680811
x2-x3-x5-x6	0.62962	3.61715	0.82670	265.8763	5585.001	0.537487
x2-x3-x5-x7	0.40544	3.94063	0.98713	293.4256	7928.248	0.592811
x2-x3-x6-x7	0.64479	3.36535	0.78664	247.4013	5653.261	0.519312
x2-x4-x5-x6	0.66668	3.42995	0.82078	255.1813	5415.734	0.510367
x2-x4-x5-x7	0.46636	3.87175	0.97991	294.5549	7457.501	0.584065
x2-x4-x6-x7	0.65751	3.33073	0.77814	249.8975	5627.713	0.515762
x2-x5-x6-x7	0.65145	3.41823	0.81424	238.5138	5491.909	0.499467
x3-x4-x5-x6	0.59241	3.77609	0.91621	276.0226	6773.318	0.536403
x3-x4-x5-x7	0.39391	4.10811	0.98518	350.162	9860.767	0.652622
x3-x4-x6-x7	0.62607	3.52809	0.82052	270.2605	5849.632	0.532919
x3-x5-x6-x7	0.60230	3.52611	0.83333	269.2714	6596.151	0.548265
x4-x5-x6-x7	0.63111	3.47363	0.81969	269.83	5987.332	0.540365
x1-x2-x3	0.64693	3.66797	0.89890	259.0548	5411.134	0.533835

Πίνακας Α.3: (συνέχεια)

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x4	0.66744	3.51620	0.88532	258.0637	5351.029	0.522706
x1-x2-x5	0.66590	3.85061	0.94261	249.1877	5028.989	0.511643
x1-x2-x6	0.70039	3.684	0.90005	233.7137	4561.837	0.473797
x1-x2-x7	0.68052	3.55476	0.90830	231.084	4994.259	0.49411
x1-x3-x4	0.59340	3.98997	1.00367	285.6599	6598.772	0.542281
x1-x3-x5	0.52316	4.02188	0.97528	318.7702	8211.599	0.590433
x1-x3-x6	0.57356	3.91131	0.94874	300.8033	7426.614	0.553153
x1-x3-x7	0.64240	3.86904	0.94184	258.7518	5355.604	0.531075
x1-x4-x5	0.57487	3.82430	0.92412	299.8058	7508.296	0.571295
x1-x4-x6	0.63789	3.74500	0.90383	274.1719	6416.292	0.529586
x1-x4-x7	0.66689	3.74661	0.91237	260.9407	5141.377	0.523936
x1-x5-x6	0.61996	4.12063	1.04993	280.9755	7999.498	0.521209
x1-x5-x7	0.56757	3.79791	0.86971	297.9789	7418.001	0.570593
x1-x6-x7	0.66231	3.52920	0.82760	270.6009	5474.283	0.514569
x2-x3-x4	0.41555	4.62889	1.27496	318.4195	6022.309	0.682362
x2-x3-x5	0.39483	4.164	1.08911	298.5532	7518.16	0.591273
x2-x3-x6	0.62972	3.61590	0.82634	266.0813	5582.142	0.537634
x2-x3-x7	0.34497	4.87280	1.40269	308.2376	6019.147	0.750628
x2-x4-x5	0.48886	4.04851	1.01308	281.3085	6769.883	0.559709
x2-x4-x6	0.66705	3.42745	0.82054	254.5176	5403.661	0.508923
x2-x4-x7	0.44343	4.53631	1.28758	315.5084	6002.464	0.696552
x2-x5-x6	0.68521	3.73572	0.93239	236.7293	4773.623	0.480256
x2-x5-x7	0.45031	3.98608	1.17345	309.9499	8121.668	0.603433
x2-x6-x7	0.65497	3.40638	0.81450	240.2293	5452.469	0.499566
x3-x4-x5	0.44322	4.29282	1.09246	324.6425	8580.427	0.6136
x3-x4-x6	0.60558	3.79403	0.92300	277.8456	6591.621	0.529527
x3-x4-x7	0.44733	4.88783	1.45051	324.6914	6284.768	0.69427
x3-x5-x6	0.54630	3.94827	0.96370	306.7044	7816.245	0.58209
x3-x5-x7	0.39530	4.08886	1.03680	344.3448	10041.39	0.653213
x3-x6-x7	0.62072	3.55162	0.83721	270.4838	6126.311	0.538833
x4-x5-x6	0.60968	3.75693	0.91807	273.6931	6808.596	0.543236
x4-x5-x7	0.43339	4.06370	1.00526	329.6851	9053.814	0.632461
x4-x6-x7	0.64313	3.50901	0.82193	269.5687	5715.993	0.533171

Πίνακας Α.3: (συνέχεια)

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x5-x6-x7	0.62002	3.55202	0.84914	276.3313	6148.82	0.544258
x1-x2	0.68927	3.80363	0.95934	240.994	4718.133	0.508676
x1-x3	0.57237	4.16049	1.05488	293.7418	7076.205	0.558123
x1-x4	0.62782	3.97511	1.01326	281.427	6588.766	0.555254
x1-x5	0.58525	4.24700	1.07729	303.8256	8496.885	0.559197
x1-x6	0.64546	4.17450	1.06922	285.1795	7714.661	0.518312
x1-x7	0.67825	3.82328	0.96248	254.3498	4780.728	0.511778
x2-x3	0.39875	4.99698	1.41347	288.0788	5369.042	0.735982
x2-x4	0.48211	4.56156	1.28838	302.1661	5606.12	0.683728
x2-x5	0.47759	4.52505	1.46406	313.5324	7420.954	0.58656
x2-x6	0.68585	3.73334	0.93211	236.6579	4763.72	0.479904
x2-x7	0.13615	5.74225	2.40559	441.3933	11042.45	1.100876
x3-x4	0.42739	5.21019	1.52815	338.2441	7360.091	0.721952
x3-x5	0.35737	4.58259	1.24705	371.8681	10405.16	0.675864
x3-x6	0.56099	3.98876	0.97482	309.5913	7693.557	0.574703
x3-x7	0.31407	5.75742	1.82410	360.1697	7253.594	0.886497
x4-x5	0.44409	4.29264	1.09222	324.2727	8565.568	0.613217
x4-x6	0.62536	3.78113	0.92539	275.5455	6626.059	0.538066
x4-x7	0.44777	4.88682	1.45005	324.4704	6280.698	0.694354
x5-x6	0.60918	4.18643	1.08096	280.8142	7788.45	0.530699
x5-x7	0.41803	4.16461	1.20326	360.3642	10164.19	0.663135
x6-x7	0.63434	3.56727	0.85310	279.4654	5885.867	0.538943
x1	0.64479	4.35297	1.15861	304.5534	8490.859	0.560087
x2	0.14566	6.40893	3.15402	454.6374	11550.43	1.207852
x3	0.10250	6.52519	2.23504	520.6244	13408.81	1.242855
x4	0.42800	5.20841	1.52723	337.9798	7355.014	0.722138
x5	0.47520	4.88284	1.76718	351.8325	10608.05	0.616239
x6	0.64006	4.21711	1.10052	287.0322	7807.246	0.522097
x7	0.10038	6.59093	2.99498	450.8747	10406.15	1.137339

Πίνακας Α.4: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Body Fat of Females.

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x3	0.68484	1.92439	0.38842	38.78044	161.04685	0.10695
x1-x2	0.71398	2.02053	0.74836	40.87315	155.30146	0.104544
x1-x3	0.65657	2.46032	0.97105	45.90043	180.19121	0.127054
x2-x3	0.68433	1.92304	0.38850	38.77718	161.33294	0.106949
x1	0.69705	2.45523	0.97305	43.56803	160.59881	0.122291
x2	0.67690	2.04633	0.81767	45.72661	161.98678	0.134024
x3	0.04882	4.21008	2.57254	92.61833	673.07976	0.26614

Πίνακας Α.5: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.12), στα δεδομένα Body Fat of Females.

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x3	0.679726	2.087357	0.344421	46.58897	169.9479	0.124399
x1-x2	0.734315	2.260203	0.379035	42.34495	135.7187	0.114457
x1-x3	0.625411	2.649073	0.53333	51.79706	213.3225	0.13826
x2-x3	0.674614	2.059648	0.359602	43.74278	169.8593	0.117449
x1	0.625966	2.648753	0.532849	51.7994	213.0806	0.138217
x2	0.669348	2.553869	0.498789	48.08001	172.2272	0.129822
x3	0.06128	4.584689	1.732601	90.11363	650.8328	0.261493

Πίνακας Α.6: Αποτελέσματα για την εφαρμογή της βηματικής παλινδρόμησης, με τη χρήση του πακέτου SPSS v.12, στα δεδομένα Hald.

Συνδυασμός	Βηματική Παλινδρόμηση			Εμπρός Επιλογή			Πίσω Απαλοιφή		
	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης	R^2	$\bar{R}^2$	Τυπικό Σφάλμα Εκτίμησης
x1-x2-x3-x4	1.000	0.999	2.418	1.000	0.999	2.418	1.000	0.999	2.418
x1-x2-x3	0.993	0.991	9.268	0.993	0.991	9.268	0.993	0.991	9.268
x1-x2-x4	0.998	0.998	4.299	0.998	0.998	4.299	0.998	0.998	4.299
x1-x3-x4	0.972	0.966	17.673	0.972	0.966	17.673	0.972	0.966	17.673
x2-x3-x4	0.991	0.989	10.041	0.991	0.989	10.041	0.993	0.991	9.336
x1-x2	0.974	0.970	16.843	0.974	0.970	16.843	0.974	0.970	16.843
x1-x3	0.972	0.966	17.673	0.972	0.966	17.673	0.972	0.966	17.673
x1-x4	0.854	0.827	40.101	0.854	0.827	40.101	0.854	0.827	40.101
x2-x3	0.961	0.958	19.738	0.961	0.958	19.738	0.961	0.958	19.738
x2-x4	0.991	0.989	10.041	0.991	0.989	10.041	0.991	0.989	10.041
x3-x4	0.704	0.679	54.648	0.704	0.679	54.648	0.765	0.723	50.829
x1	0.730	0.707	52.225	0.730	0.707	52.225	0.730	0.707	52.225
x2	0.961	0.958	19.738	0.961	0.958	19.738	0.961	0.958	19.738
x3	0.704	0.679	54.648	0.704	0.679	54.648	0.704	0.679	54.648
x4	0.661	0.633	58.447	0.661	0.633	58.447	0.661	0.633	58.447

Πίνακας Α.7: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.8), στα δεδομένα Hald.

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x3-x4	0.92389	0.53775	0.06922	39.928	299.205	0.030
x1-x2-x3	0.94421	0.60232	0.08280	35.892	175.202	0.028
x1-x2-x4	0.93251	0.53409	0.06929	37.117	286.137	0.028
x1-x3-x4	0.96981	0.646069	0.0967403	21.254	98.373	0.018
x2-x3-x4	0.88072	0.85442	0.14634	40.433	390.951	0.032
x1-x2	0.93935	0.59751	0.08251	38.318	217.013	0.030
x1-x3	0.49647	1.48845	0.91577	99.240	1908.902	0.087
x1-x4	0.96999	0.64308	0.09622	21.158	97.882	0.018
x2-x3	0.83473	0.96835	0.16170	48.224	465.102	0.038
x2-x4	0.70979	1.30724	0.62827	76.423	866.786	0.061
x3-x4	0.89399	0.95546	0.18288	42.510	381.520	0.035
x1	0.59224	1.48504	0.91826	86.110	1357.349	0.075
x2	0.57885	1.50082	0.57114	90.040	1369.861	0.069
x3	0.23307	2.22113	1.11215	152.643	2692.614	0.129
x4	0.68657	1.65020	0.71044	87.988	887.591	0.072

Πίνακας Α.8: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.12), στα δεδομένα Hald.

Συνδυασμός	R^2	Sum(σ)	Sum(σ^2)	Sumlel	Sum(e^2)	MAPE
x1-x2-x3-x4	0.971903	0.592471	0.045055	27.546	95.256	0.022
x1-x2-x3	0.963971	0.665133	0.052701	31.680	114.198	0.025
x1-x2-x4	0.972212	0.588761	0.044747	27.358	94.307	0.022
x1-x3-x4	0.962824	0.799954	0.076546	34.007	124.225	0.028
x2-x3-x4	0.944592	0.952499	0.103232	43.044	204.883	0.035
x1-x2	0.964651	0.659922	0.052167	31.396	112.490	0.025
x1-x3	0.487448	1.631638	0.339633	108.906	1802.537	0.094
x1-x4	0.962982	0.798468	0.076316	33.954	123.654	0.028
x2-x3	0.896476	1.062517	0.139738	49.620	327.185	0.040
x2-x4	0.670633	1.635362	0.328136	92.076	949.614	0.073
x3-x4	0.938268	1.137644	0.132609	47.813	223.741	0.040
x1	0.488046	1.628046	0.338462	108.835	1800.152	0.093
x2	0.64175	1.79277	0.360673	96.919	1039.598	0.076
x3	0.330544	2.506472	0.723887	135.731	2449.374	0.118
x4	0.649438	1.853286	0.380352	98.176	980.324	0.079

Πίνακας Α.9: Αποτελέσματα για την εφαρμογή του προτεινόμενου γ.π. (4.17), στα δεδομένα Hald.

Εξεταζόμενος Συνδυασμός	Τιμή	
	Αντικειμενικής Συνάρτησης $(\sum_{i=1}^n (\hat{y}_i - y_i)^2)$	Υπολογιζόμενος R^2 $\sum_{i=1}^n (y_i - \bar{Y})^2$
x1-x2-x3-x4	58.46	97.847%
x1-x2-x3	99.87	96.323%
x1-x2-x4	58.03	97.863%
x1-x3-x4	81.41	97.002%
x2-x3-x4	125.62	95.374%
x1-x2	96.85	96.434%
x1-x3	953.05	64.907%
x1-x4	80.98	97.018%
x2-x3	264.91	90.245%
x2-x4	827.61	69.526%
x3-x4	136.19	94.985%
x1	942.16	65.308%
x2	943.73	65.250%
x3	2265.28	16.588%
x4	544.89	79.936%

Πίνακας Α.10: Αποτελέσματα των χρησιμοποιούμενων προτεινόμενων μεθοδολογιών βηματικής παλινδρόμησης, στα δεδομένα Hald.

Εξεταζόμενος Συνδυασμός	Βηματική Παλινδρόμηση		Βηματική Ποιοτική Παλινδρόμηση			
	Εμπρός Επιλογή	Πίσω Απαλοιφή	Απλά Απόλυτα Σφάλματα		Τετραγωνικά Σφάλματα	
	R^2	R^2	$\sum_i (\sigma_i^+ + \sigma_i^-)$	R^2	$\sum_i ((\sigma_i^+)^2 + (\sigma_i^-)^2)$	R^2
x1-x2-x3-x4	0.99957	0.99957	0.53775	0.92389	0.045055	0.971903
x1-x2-x3	0.99291	0.99291	0.60232	0.94421	0.052701	0.963971
x1-x2-x4	0.99847	0.99847	0.53409	0.93251	0.044747	0.972212
x1-x3-x4	0.97163	0.97163	0.646069	0.96981	0.076546	0.962824
x2-x3-x4	0.99084	0.99280	0.85442	0.88072	0.103232	0.944592
x1-x2	0.97423	0.97423	0.59751	0.93935	0.052167	0.964651
x1-x3	0.97163	0.97163	1.48845	0.49647	0.339633	0.487448
x1-x4	0.85391	0.85391	0.64308	0.96999	0.076316	0.962982
x2-x3	0.96139	0.96139	0.96835	0.83473	0.139738	0.896476
x2-x4	0.99084	0.99084	1.30724	0.70979	0.328136	0.670633
x3-x4	0.70405	0.76530	0.95546	0.89399	0.132609	0.938268
x1	0.72971	0.72971	1.48504	0.59224	0.338462	0.488046
x2	0.96139	0.96139	1.50082	0.57885	0.360673	0.64175
x3	0.70405	0.70405	2.22113	0.23307	0.723887	0.330544
x4	0.66147	0.66147	1.65020	0.68657	0.380352	0.649438