Μεταπτυχιακή εργασία

ΚΡΙΤΗΡΙΑ ΕΠΙΛΟΓΗΣ ΒΕΛΤΙΣΤΩΝ ΜΟΝΤΕΛΩΝ ΠΟΛΥΜΕΤΑΒΛΗΤΗΣ ΓΡΑΜΜΙΚΗΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ

Παπαδογιάννης Σπυρίδων

Κατεύθυνση:
Επιχειρησιακά Μαθηματικά

Πανεπιστήμιο Κρήτης
Ιούνιος 2007

Στα πλαίσια του διατμηματικού μεταπτυχιακού προγράμματος «Τα Μαθηματικά και οι Εφαρμογές τους» του τμήματος Μαθηματικών και του τμήματος Εφαρμοσμένων Μαθηματικών του Πανεπιστημίου Κρήτης

Επιβλέπων καθηγητής: Πουλίκος Πραστάκος
Τριμελής επιτροπή: κ.Πραστάκος, κ.Κρητικού, κ.Λουλάκης

Ευχαριστώ τον κ.Καμαριανάκη για τη βοήθεια και την καθοδήγησή του στη συγγραφή της εργασίας

ΠΕΡΙΕΧΟΜΕΝΑ

ΕΙΣΑΓΩΓΗ

Η παλινδρόμηση χρησιμοποιείται για να μελετηθούν οι σχέσεις μεταξύ μετρήσιμων μεταβλητών. Κατά τη γραμμική παλινδρόμηση, οι σχέσεις αυτές περιγράφονται μέσω ευθειών γραμμών ή γενικότερα μέσω γραμμικών εξισώσεων. Η αναζήτηση γραμμικών μοντέλων με τη βοήθεια πραγματικών δεδομένων αποτελεί αντικείμενο μελέτης και έρευνας για τις περισσότερες των επιστημών, όπως οι κοινωνικές επιστήμες, η φυσική, η μηχανική, η βιολογία, η ιατρική, η οικονομία, η επιχειρησιακή έρευνα και η τεχνολογία. Οι βασικότεροι λόγοι για τους οποίους προσαρμόζουμε τα δεδομένα που διαθέτουμε σε ένα κατάλληλο γραμμικό μοντέλο είναι η περιγραφή και η εκτίμηση των σχέσεων μεταξύ των μεταβλητών καθώς και η πρόβλεψη μελλοντικών τιμών για μια συγκεκριμένη μεταβλητή που μας ενδιαφέρει.

Η διάκριση μεταξύ των μεταβλητών γίνεται ως εξής: Μία από τις μεταβλητές επιλέγεται ως **εξαρτημένη** (απόκριση ή *response*),ενώ οι υπόλοιπες αποτελούν τις **ανεξάρτητες** (επεξηγηματικές ή *predictors*). Συλλέγοντας δεδομένα που αφορούν τις ανεξάρτητες μεταβλητές, προσπαθούμε να εκφράσουμε την εξαρτημένη μεταβλητή ως μια συνάρτηση, κατά προτίμηση γραμμική, των ανεξαρτήτων. Ένα μοντέλο που χαρακτηρίζει αυτή τη συνάρτηση, καθορίζει τους συντελεστές των επεξηγηματικών μεταβλητών αλλά και τη συμπεριφορά της εξαρτημένης μεταβλητής για δοσμένες τιμές των επεξηγηματικών.

Το ενδιαφέρον επικεντρώνεται στην εύρεση του βέλτιστου (γραμμικού) μοντέλου, αυτού δηλαδή που θα προσδιορίζει με το βέλτιστο δυνατό τρόπο την εξίσωση μεταξύ εξαρτημένης και ανεξαρτήτων μεταβλητών. Δύο κριτήρια που έρχονται σε αντίθεση μεταξύ τους αλλά προσδιορίζουν την επιλογή του βέλτιστου μοντέλου είναι τα εξής:

- Για να είναι το μοντέλο χρήσιμο ως προς την πρόβλεψη τιμών για την εξαρτημένη μεταβλητή πρέπει να περιλαμβάνει όσο το δυνατόν περισσότερες επεξηγηματικές μεταβλητές ή συναρτήσεις αυτών, ώστε να προκύπτουν αξιόπιστα αποτελέσματα.

- Λόγω του κόστους που προϋποθέτει η απόκτηση των πληροφοριών για ένα μεγάλο πλήθος δεδομένων, κατά συνέπεια και ο έλεγχος αυτών, πρέπει το μοντέλο να περιλαμβάνει όσο το δυνατό λιγότερες επεξηγηματικές μεταβλητές ή συναρτήσεις αυτών.

Ο συνδυασμός αυτών των δύο κριτηρίων καταλήγει στην επιλογή του βέλτιστου μοντέλου. Το ζητούμενο λοιπόν είναι η κατάλληλη επιλογή των μεταβλητών που θα συμπεριληφθούν στο μοντέλο. Το πρόβλημα της εύρεσης του βέλτιστου υποσυνόλου των μεταβλητών απασχολεί εδώ και πολύ καιρό τους εφαρμοσμένους στατιστικούς και ειδικότερα τα τελευταία χρόνια χάρη στη χρήση υπολογιστών μεγάλης ταχύτητας. Σε αρκετά άρθρα έχουν αναλυθεί διάφορες πτυχές του προβλήματος αλλά φαίνεται ότι ένας μέσος χρήστης δεν έχει σημαντικό κέρδος. Η έλλειψη δυνατότητας της επίλυσης του προβλήματος μπορεί να οφείλεται στο γεγονός ότι δεν έχει μοναδική λύση αλλά διαφορετικά μοντέλα μπορεί

να είναι εξίσου ικανοποιητικά. Πρέπει λοιπόν να δωθούν συγκεκριμένες απαντήσεις και οδηγίες στους εφαρμοσμένους στατιστικούς που επιχειρούν να ασχοληθούν με αυτό το ζήτημα.

Το πρόβλημα της επιλογής ενός υποσυνόλου των επεξηγηματικών μεταβλητών απαιτεί να ικανοποιούνται οι εξής προϋποθέσεις:

• ο ερευνητής να διαθέτει δεδομένα από έναν αρκετά μεγάλο αριθμό μεταβλητών, μεταξύ των οποίων βρίσκονται όλες οι απαραίτητες για το πρόβλημα μεταβλητές και κατάλληλες συναρτήσεις αυτών, καθώς και άλλες μεταβλητές που ενδέχεται να επηρεάζουν το πρόβλημα.

• ο ερευνητής να διαθέτει αξιόπιστα δεδομένα στα οποία να μπορεί να βασίσει τα τελικά του συμπεράσματα.

Στην πράξη, η έλλειψη ικανοποίησης αυτών των προϋποθέσεων μπορεί να κάνει μια λεπτομερή ανάλυση επιλογής υποσυνόλου να οδηγήσει σε λανθασμένα συμπεράσματα, ακόμα και αν έχουμε εργαστεί με λεπτομέρεια για να επιλέξουμε ένα υποσύνολο.

Δεν είναι καθόλου εύκολο να εξασφαλίσουμε ότι διαθέτουμε όλες τις σημαντικές για το πρόβλημά μας μεταβλητές. Η ανάλυση των υπολοίπων μπορεί να εμφανίσει διαφορετικές συναρτησιακές μορφές που θα έπρεπε να ληφθούν υπ'οψη, ακόμα και να προτείνουν μεταβλητές που δεν είχαν αρχικά συμπεριληφθεί. Τέτοιου τύπου φαινόμενα εντοπίζονται και αντιμετωπίζονται με λεπτομέρη εξέταση από τον ερευνητή.

Για να ελέγξουμε την αξιοπιστία των δεδομένων, τα γραφήματα των υπολοίπων μπορούν ξανά να προτείνουν μετασχηματισμούς ή να εμφανίσουν αναξιόπιστα ή ελλιπή δεδομένα ακόμα και ακραίες τιμές(*outliers*). Ένα σοβαρό πρόβλημα που ενδέχεται να προκύψει είναι η πολυσυγγραμμικότητα (*multicollinearity*) μεταξύ των ανεξάρτητων μεταβλητών. Η πολυσυγγραμικότητα προκύπτει όταν κάποιες μεταβλητές είναι στενά συσχετισμένες μεταξύ τους με αποτέλεσμα οι αντίστοιχες σε αυτές στήλες του πίνακα των δεδομένων να είναι γραμμικώς εξαρτημένες ή να έχουν ισχυρή εξάρτηση μεταξύ τους με αποτέλεσμα ο πίνακας αυτός να είναι ιδιάζων δηλαδή η ορίζουσά του να είναι πολύ κοντά ή ίση με το 0. Αυτό έχει ως άποτέλεσμα οι εκτιμητές των συντελεστών παλινδρόμησης να έχουν μεγάλη διασπορά, αφού αυτή έξαρτάται από τον αντίστροφο αυτού του πίνακα (όπως θα δούμε στο κεφάλαιο 1), γεγονός που σημαίνει μεγάλη απόκλιση αυτών από τις πραγματικές τους τιμές. Επιπλέον, η εξίσωση πρόβλεψης που προκύπτει μπορεί να είναι αρκετά αναξιόπιστη, ειδικά όταν χρησιμοποιείται έξω από την κλίμακα των αρχικών δεδομένων.

Η πολυσυγγραμμικότητα και το φαινόμενο των ελλιπών δεδομένων είναι δύο προβλήματα που πρέπει να αντιμετωπιστούν παράλληλα. Η αστάθεια της μεθόδου των ελαχίστων τετραγώνων και η εμφάνιση ιδιαζόντων πινάκων σημαίνει ότι τα γραφήματα των υπολοίπων μπορεί να μην εμφανίσουν ελλιπή δεδομένα ή λανθασμένες ενδείξεις. Η ανάγκη για διαδικασίες που να είναι περισσότερο ευσταθείς απέναντι σε τέτοια φαινόμενα είναι φανερή και θα μελετηθούν σε επόμενα κεφάλαια.

Το πρόβλημα της επιλογής της κατάλληλης εξίσωσης βασισμένης σε ένα υποσύνολο του αρχικού συνόλου των μεταβλητών εμπεριέχει τρεις βασικούς άξονες:

(i) την υπολογιστική μέθοδο που θα χρησιμοποιηθεί για να παρέχει τις πληροφορίες για την ανάλυση,

(*ii*) το κριτήριο που θα καθορίσει την επιλογή του κατάλληλου υποσυνόλου των μεταβλητών που θα συμπεριληφθούν στο μοντέλο και

(*iii*) την εκτίμηση των παραμέτρων της τελικής εξίσωσης.

Στο Κεφάλαιο 1, παρουσιάζεται η βασική μέθοδος εκτίμησης των παραμέτρων ενός γραμμικού μοντέλου πρόβλεψης, δηλαδή η μέθοδος ελαχίστων τετραγώνων. Έπειτα, περιγράφεται η ανάλυση της διασποράς του μοντέλου παλινδρόμησης (Πίνακας $ANOVA$) και ο συντελεστής προσδιορισμού R^2, ο οποίος αποτελεί ένα δείκτη καλής εφαρμογής του μοντέλου. Στη συνέχεια, δίνεται η έννοια τς μεροληψίας στην εκτίμηση των παραμέτρων του μοντέλου. Εξετάζονται επίσης κάποιες βασικές μέθοδοι υπολογισμού του βέλτιστου υποσυνόλου των επεξηγηματικών μεταβλητών που θα χρησιμοποιηθούν στην εξίσωση παλινδρόμησης, καθώς επίσης και τα βασικότερα κριτήρια επιλογής του υποσυνόλου αυτού. Τέλος, αναλύονται τρεις διαδικασίες μεροληπτικής εκτίμησης, οι οποίες καθιστούν το γραμμικό μοντέλο πιο ευσταθές σε σχέση με αυτό που προκύπτει όταν χρησιμοποιούμε τη μέθοδο ελαχίστων τετραγώνων.

Στο κεφάλαιο 2, περιγράφονται ορισμένες προτάσεις σχετικές με την επιλογή του βέλτιστου υποσυνόλου των επεξηγηματικών μεταβλητών. Οι *R.R.Hocking* και *R.N.Leslie* (1967) υπολογίζουν μόνο ορισμένα καταλλήλως επιλεγμένα υποσύνολα από τις μεταβλητές, με σκοπό να βρεθεί το βέλτιστο υποσύνολο με τον ελάχιστο δυνατό υπολογιστικό φόρτο. Οι *J.W.Gorman* και *R.J.Toman* (1966) χρησιμοποιούν ως κριτήριο επιλογής του βέλτιστου υποσυνόλου το στατιστικό C_p του *Mallows*, οπότε η τελική εξίσωση περιορίζεται μεταξύ λίγων υποσυνόλων. Οι $T.A.Bancroft$ και *W.J.Kennedy* (1971) βασίζονται σε διαδοχικούς ελέγχους σημαντικότητας και χρησιμοποιούν τις διαδικασίες "*Forward Selection*" και "*Sequential Deletion*" για την επιλογή των κατάλληλων μεταβλητών. Οι *H.J.Larson* και $T.A.Bancroft$ (1963) περιγράφουν και συγκρίνουν δύο διαδικασίες επιλογής των μεταβλητών που θα συμπεριληφθούν στο τελικό μοντέλο με βάση την ταξινόμησή τους σε μια σειρά σημαντικότητας. Οι *H.J.Larson* και $T.A.Bancroft$ (1963) προτείνουν την επιλογή μεταξύ του πλήρους μοντέλου (αυτού δηλαδή που περιέχει όλες τις μεταβλητές) και ενός περιορισμένου μοντέλου, που περιλαμβάνει μόνο κάποιες από τις μεταβλητές, οι οποίες καθορίζονται από τον ερευνητή. Έπειτα, υπολογίζεται η μέση τιμή και το μέσο τετραγωνικό σφάλμα του εκτιμητή της εξαρτημένης μεταβλητής. Τέλος, ο $T.A.Bancroft$ (1944) ασχολείται με τον έλεγχο της σημαντικότητας του συντελεστή μιας μεταβλητής, ώστε να τη διατηρήσουμε στο μοντέλο ή να την απαλείψουμε από αυτό.

Στο κεφάλαιο 3, περιγράφεται με τη βοήθεια του άρθρου του *R.Tibshirani* (1996) η μέθοδος *Lasso*, σύμφωνα με την οποία επιλύεται το πρόβλημα της ελαχιστοποίησης των τετραγώνων των σφαλμάτων του μοντέλου παλινδρόμησης υπό τον περιορισμό το άθροισμα των απολύτων τιμών των συντελεστών των επεξηγηματικών μεταβλητών να είναι μικρότερο ή ίσο από μία ρυθμιζόμενη παράμετρο. Γίνονται συγκρίσεις με άλλες μεθόδους, όπως η *subset selection* και η *ridge regression*, αλλά και η ίδια η μέθοδος των ελαχίστων τετραγώνων. Στη συνέχεια, παρουσιάζεται το δυϊκό πρόβλημα σύμφωνα με το άρθρο των $Osborne, M.R.$, $Presnell, B.$ και $Turlach, B.A.$ (2000) και αναλύεται η σχέση που έχουν οι λύσεις της μεθόδου *Lasso* με αυτές του δυϊκού προβλήματος.

Το κεφάλαιο 4 συγκεντρώνει το μεγαλύτερο ενδιαφέρον, αφού περιγράφει μια πολύ πρόσφατη μέθοδο του $B.Efron$ και των $T.Hastie, I.Johnstone$ και $R.Tishribani$ (2004),

η οποία ονομάζεται "Παλινδρόμηση Ελάχιστης Γωνίας" (*Least Angle Regression*) ή *LARS*. Η μέθοδος αυτή συνδυάζει με υπολογιστικά απλούστερο τρόπο τις *Lasso* και *Forward Stagewise* μεθόδους.

Στο κεφάλαιο 5, τέλος, εφαρμόζεται η μέθοδος *LARS*, αλλά και οι μέθοδοι *Lasso* και *Forward Stagewise* μέθοδοι πάνω σε πραγματικά δεδομένα με τη βοήθεια του στατιστικού πακέτου *R*. Προτείνονται βέλτιστα γραμμικά μοντέλα σύμφωνα με δύο βασικά κριτήρια επιλογής και, παράλληλα, αναλύονται τα αποτελέσματα και συγκρίνονται οι μέθοδοι.

Κεφάλαιο 1

ΒΑΣΙΚΟ ΠΡΟΒΛΗΜΑ - ΚΥΡΙΕΣ ΜΕΘΟΔΟΙ ΕΠΙΛΥΣΗΣ

Στο παρόν κεφάλαιο, περιγράφονται συνοπτικά κάποια βασικά χαρακτηριστικά της γραμμικής παλινδρόμησης, όπως επίσης και οι κύριες μέθοδοι επίλυσης του προβλήματος της επιλογής των επεξηγηματικών μεταβλητών και της εκτίμησης των παραμέτρων. Αρχικά, αναφέρεται η μέθοδος των ελαχίστων τετραγώνων, ο πίνακας ανάλυσης της διασποράς του μοντέλου παλινδρόμησης ($ANOVA$) και ο συντελεστής προσδιορισμού R^2 ($Weisberg2005$). Έπειτα, παρουσιάζεται η έννοια της μεροληψίας στους εκτιμητές του μοντέλου ($Draper, Smith1981$), ορισμένες βασικές υπολογιστικές μέθοδοι, τα κυριότερα κριτήρια επιλογής του κατάλληλου υποσυνόλου των μεταβλητών και, τέλος, αναλύονται τρεις βασικές μεροληπτικές μέθοδοι. ($Hocking1976$).

Σε πολλές περιπτώσεις μία γραμμική σχέση είναι πολύτιμη ως προς την περιγραφή της εξάρτησης μιας μεταβλητής από μία ή περισσότερες άλλες μεταβλητές. Υποθέτουμε ότι διαθέτουμε n παρατηρήσεις από ένα σύνολο p επεξηγηματικών μεταβλητών $X_1, X_2, ..., X_p$ που εισάγουμε (Πρέπει $n \geq p+1$). Διαθέτουμε επίσης μια εξαρτημένη μεταβλητή Y, τέτοια ώστε η j-οστή συνιστώσα της, Y_j, με $j = 1, 2, ...n$, να καθορίζεται από τη σχέση:

$$Y_j = \beta_0 + \sum_{i=1}^{p} \beta_i X_{ij} + e_j \tag{1.1}$$

όπου το X_{ij} εκφράζει την τιμή της X_i μεταβλητής για την j-οστή παρατήρηση, ενώ τα β_0 και β_i, $i = 1, ..., p$, εκφράζουν το σταθερό όρο του μοντέλου και τους συντελεστές των μεταβλητών X_i, $i = 1, ..., p$ αντίστοιχα. Τα υπόλοιπα (σφάλματα) e_j υποθέτουμε ότι είναι ανεξάρτητα κατανεμημένα (συνήθως ακολουθούν την κανονική κατανομή) με μέση τιμή 0 και άγνωστη διασπορά σ^2. (Οι τιμές εισόδου X_{ij} είναι συνήθως καθορισμένες τιμές, αλλά σε πολλές περιπτώσεις είναι προτιμότερο να τις θεωρούμε τυχαίες μεταβλητές και να θεωρήσουμε μια κοινή κατανομή των Y και $X_1, X_2, ..., X_p$, για παράδειγμα πολυμεταβλητή κανονική). Υποθέτουμε επιπλέον ότι οι μεταβλητές $X_1, X_2, ..., X_p$ περιέχουν όλες τις απαραίτητες για το πρόβλημά μας μεταβλητές παρόλο που ενδέχεται να συμπεριλαμβάνονται σε αυτές και ορισμένες ασήμαντες μεταβλητές.

Το μοντέλο (1.1) συχνά εκφράζεται με τη μορφή πινάκων ως εξής:

$$Y = X\beta + e, \tag{1.2}$$

με $Ee = 0$ και $Var(e) = E(ee') = \sigma^2 I$.

Εδώ, Y είναι το n-διάστατο διάνυσμα των παρατηρούμενων τιμών της εξαρτημένης μεταβλητής. Ο πίνακας X, που ονομάζεται "*Πίνακας Σχεδιασμού* (*Design Matrix*)", έχει διάσταση $n \times (p+1)$ και τάξη $p+1$, ενώ η πρώτη του στήλη αντιστοιχεί στο σταθερό όρο β_0 και τα στοιχεία της είναι μονάδες. Το β είναι το $(p+1)$-διάστατο διάνυσμα των αγνώστων συντελεστών παλινδρόμησης, όπου συμπεριλαμβάνεται και το β_0. Σε μερικές περιπτώσεις, είναι ευκολότερο να υποθέσουμε ότι οι τιμές των επεξηγηματικών μεταβλητών αλλά και της εξαρτημένης μεταβλητής έχουν εκφραστεί ως αποκλίσεις από τις δειγματικές μέσες τιμές για κάθε μεταβλητή. Σε άλλες περιπτώσεις υποθέτουμε επιπλέον ότι έχουν άθροισμα τετραγώνων ίσο με 1 για κάθε μεταβλητή. Σε τέτοιες περιπτώσεις, χρησιμοποιούμε τη σχέση (1.2) για να εκφράσουμε το μοντέλο, αλλά τονίζουμε ότι ο $n \times p$ πίνακας είναι αυτή τη φορά ο 'προσαρμοσμένος' (*adjusted*) πίνακας σχεδιασμού ή ο 'κανονικοποιημένος' (*standardized*) πίνακας σχεδιασμού.

1.1. Μέθοδος Ελαχίστων Τετραγώνων

Το ζητούμενο είναι να λυθεί το πρόβλημά μας, δηλαδή να προσδιοριστεί το διάνυσμα β στο μοντέλο (1.2) έτσι ώστε το σφάλμα e να είναι ελάχιστο. Η συνήθης μέθοδος για να γίνει αυτό είναι αυτή των ελαχίστων τετραγώνων. Η μέθοδος αυτή απαιτεί το άθροισμα τετραγώνων των σφαλμάτων, το οποίο θα συμβολίζουμε στο εξής με RSS (*Residual Sum of Squares*) να γίνεται ελάχιστο. Τα εκτιμώμενα β_i που θα προκύψουν, συμβολίζονται με $\hat{\beta}_i$ και το μοντέλο παίρνει τη μορφή:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_p X_p \tag{1.3}$$

όπου $\hat{Y}$ είναι η εκτίμηση της πραγματικής τιμής του Y για δοσμένα $X_1, X_2, ..., X_p$.

Σκοπός μας λοιπόν είναι να ελαχιστοποιήσουμε τα αθροίσματα:

$$RSS = \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} (Y_j - \beta_0 - \sum_{i=1}^{p} \beta_i X_{ij})^2 = (Y - X\beta)'(Y - X\beta) \tag{1.4}$$

Θα επιλέξουμε τους εκτιμητές $\hat{\beta}_i$ αυτούς που θα δώσουν την ελάχιστη τιμή για το RSS. Με τη βοήθεια του κριτηρίου της πρώτης και της δεύτερης παραγώγου, προκύπτει η παρακάτω σχέση.

$$X'X\beta = X'Y \tag{1.5}$$

Αυτή η σχέση παριστάνει ένα σύστημα $p+1$ γραμμικών εξισώσεων ως προς τις παραμέτρους $\beta_0, \beta_1, ..., \beta_p$ και λέγονται **κανονικές εξισώσεις** (*normal equations*). Αν ο πίνακας $X'X$ είναι αντιστρέψιμος,τότε το σύστημα (1.5) έχει μοναδική λύση την:

$$\hat{\beta} = (X'X)^{-1}X'Y \tag{1.6}$$

Αν ο $X'X$ είναι μη-αντιστρέψιμος,τότε η (1.5) έχει άπειρες λύσεις της μορφής:

$$\hat{\beta} = (X'X)^{-}X'Y + (I - (X'X)^{-}X'X)w \tag{1.7}$$

όπου $(X'X)^{-}$ είναι ο γενικευμένος αντίστροφος του $X'X$ και w ένα αυθαίρετο διάνυσμα.

Στη συνέχεια, για να εκτιμήσουμε τη λύση (1.6), υπολογίζουμε τη μέση τιμή και τη διασπορά του $\hat{\beta}$. Έχουμε λοιπόν:

$$E\hat{\beta} = E[(X'X)^{-1}X']E(Y) = (X'X)^{-1}X'X\beta = \beta \tag{1.8}$$

Επίσης,

$$\begin{aligned} Var(\hat{\beta}) \quad &= E(\hat{\beta} - E\hat{\beta})(\hat{\beta} - E\hat{\beta})' \\ &= E(\hat{\beta} - \beta)(\hat{\beta} - \beta)' = E[(X'X)^{-1}X'e][(X'X)^{-1}X'e]' \\ &= (X'X)^{-1}X'[E(ee')]X(X'X)^{-1} = (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned} \tag{1.9}$$

αφού $\hat{\beta} - \beta = (X'X)^{-1}X'Y - \beta = (X'X)^{-1}X'(X\beta + e) - \beta = (X'X)^{-1}X'e$.

Αν συμβολίσουμε με C τον πίνακα $(X'X)^{-1}$ και c_{ij} τα στοιχεία του,τότε η διασπορά του εκτιμητή $\hat{\beta}_i$ της παραμέτρου β_i θα είναι:

$$Var(\hat{\beta}_i) = \sigma^2 c_{ii} \tag{1.10}$$

με $i = 0, 1, ..., p$.

Η τυπική απόκλιση του $\hat{\beta}_i$ θα είναι λοιπόν:

$$\sigma(\hat{\beta}_i) = \sigma\sqrt{c_{ii}} \tag{1.11}$$

με $i = 0, 1, ..., p$.

Επίσης, η συνδιασπορά των $\hat{\beta}_i$ και $\hat{\beta}_j$ θα είναι:

$$cov(\hat{\beta}_i, \hat{\beta}_j) = \sigma^2 c_{ij} \tag{1.12}$$

για $i \neq j$ και $i, j = 0, 1, ..., p$.

Η παράμετρος σ^2 στις σχέσεις (1.10)-(1.12) είναι άγνωστη, οπότε πρέπει να την εκτιμήσουμε. Αφού το σ^2 είναι ουσιαστικά το μέσο τετραγωνικό "μέγεθος" των e_i, ο εκτιμητής του, $\hat{\sigma}^2$, θα υπολογίζεται από τη μέση τιμή των e_i^2. Υπό την προϋπόθεση ότι τα e_i είναι ασυσχέτιστες τυχαίες μεταβλητές με μέση τιμή 0 και κοινή διασπορά σ^2, μπορούμε να πάρουμε ως εκτιμητή του σ^2 το πηλίκο του RSS με τους **βαθμούς ελευθερίας** (df) του, όπου

df = πλήθος παρατηρήσεων - πλήθος παραμέτρων στο μοντέλο $= dim\Re^n - dim\Theta$
όπου Θ ο παραμετρικός χώρος.

Για την πολλαπλή παλινδρόμηση, $df = n-p-1$, οπότε ο εκτιμητής του σ^2 δίνεται από τη σχέση:

$$\hat{\sigma}^2 = \frac{R\hat{S}S}{n-p-1} \tag{1.13}$$

όπου $R\hat{S}S = ||Y - X\hat{\beta}||^2$

Η ποσότητα αυτή καλείται *μέσο τετραγωνικό σφάλμα* (*residual mean square* ή MRS). Γενικά, κάθε άθροισμα τετραγώνων διαιρούμενο με τους βαθμούς ελευθερίας που αντιστοιχούν σε αυτό καλείται *μέσο τετράγωνο*.

Αν στις υποθέσεις μας έχουμε επιπλέον ότι τα e_i ακολουθούν μια κανονική κατανομή, τότε το MRS θα είναι μια τυχαία μεταβλητή που ακολουθεί την κατανομή χ^2 με $n-p-1$ βαθμούς ελευθερίας ή συμβολικά:

$$(n-p-1)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p-1} \tag{1.14}$$

Αν αντικαταστήσουμε το $\hat{\sigma}^2$ στη θέση του σ^2 στην (1.9), βρίσκουμε την εκτιμώμενη διασπορά του $\hat{\beta}$:

$$\widehat{Var}(\hat{\beta}) = \hat{\sigma}^2(X'X)^{-1}. \tag{1.15}$$

Τονίζουμε ότι αν υπάρχει γραμμική εξάρτηση μεταξύ ορισμένων επεξηγηματικών μεταβλητών, παρατηρείται το φαινόμενο της *πολυσυγγραμμικότητας*. Ο πίνακας $X'X$ τότε είναι μη-αντιστρέψιμος, οπότε η τελευταία σχέση δεν μπορεί να χρησιμοποιηθεί. Επίσης, αν υπάρχει ισχυρή εξάρτηση μεταξύ των μεταβλητών, η ορίζουσα του πίνακα $X'X$ είναι πολύ μικρή, οπότε τα στοιχεία του πίνακα $(X'X)^{-1}$ παίρνουν πολύ μεγάλες τιμές, επομένως η διασπορά των συντελεστών παλινδρόμησης θα είναι πολύ μεγάλη. Τέτοιου είδους προβλήματα αντιμετωπίζονται με ευσταθείς διαδικασίες που θα αναλυθούν αργότερα.

1.2. Ανάλυση Διασποράς Μοντέλου Παλινδρόμησης (Πίνακας $ANOVA$)

Για την πολλαπλή παλινδρόμηση, η ανάλυση της διασποράς του μοντέλου είναι μια τεχνική που χρησιμοποιείται για να συγκρίνουμε μοντέλα που περιλαμβάνουν διαφορετικά σύνολα μεταβλητών. Ως βασικό παράδειγμα εδώ, το πλήρες μοντέλο

$$Y = X\beta + e \tag{1.16}$$

συγκρίνεται με το μοντέλο που δεν περιλαμβάνει καμία από τις μεταβλητές $X_1, X_2, ..., X_p$, δηλαδή το

$$Y = \beta_0\mathbf{1} + e, \tag{1.17}$$

όπου $\mathbf{1}$ είναι το $n \times 1$ διάνυσμα με στοιχεία μονάδες.

Μπορούμε εύκολα να αποδείξουμε ότι για το μοντέλο (1.17) ισχύει $\hat{\beta}_0 = \overline{Y}$ και το RSS είναι ίσο με $\sum_{i=1}^{n}(Y_i - \hat{\beta}_0)^2 = \sum_{i=1}^{n}(Y_i - \overline{Y})^2$, ποσότητα που τη συμβολίζουμε με SYY, δηλαδή:

$$SYY = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 \tag{1.18}$$

Για το μοντέλο (1.16) ο εκτιμητής του β δίνεται από τη σχέση (1.6) και το RSS από τη σχέση (1.4). Προφανώς, ισχύει ότι $RSS < SYY$ και η διαφορά τους:

$$SSR = SYY - RSS \tag{1.19}$$

λέγεται *άθροισμα τετραγώνων λόγω της παλινδρόμησης* (*Sum of Squares due to Regression*) και αποτελεί το άθροισμα των τετραγώνων των τιμών του Y που εξηγείται από το μεγαλύτερο μοντέλο και δεν εξηγείται από το μικρότερο. Οι βαθμοί ελευθερίας που σχετίζονται με το SSR είναι ίσοι με τους βαθμούς ελευθερίας του SYY $(n-1)$ μείον τους β.ε. του RSS $(n-p-1)$, δηλαδή $n-1-(n-p-1) = p$ β.ε.

Από την (1.4) το RSS για $\beta = \hat{\beta}$ γράφεται:

$$\begin{aligned} RSS \quad &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X\hat{\beta} \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'X(X'X)^{-1}X'Y \\ &= Y'Y - 2\hat{\beta}'X'Y + \hat{\beta}'X'Y \\ &= Y'Y - \hat{\beta}'X'Y \end{aligned} \tag{1.20}$$

Η σχέση (1.18) μπορεί να γραφτεί ως εξής:

$$SYY = Y'Y - \frac{1}{n}(Y'\mathbf{1})^2, \tag{1.21}$$

Επίσης, ισχύει ότι:

$$\sum_{i=1}^{n}(Y_i - \overline{Y})^2 = Y'Y - \frac{1}{n}(Y'\mathbf{1})^2 \tag{1.22}$$

Έπειτα, λόγω της σχέσης (1.19) ισχύει:

$$SSR = Y'Y - \frac{1}{n}(Y'\mathbf{1})^2 - Y'Y + \hat{\beta}'X'Y = \hat{\beta}'X'Y - \frac{1}{n}(Y'\mathbf{1})^2 \tag{1.23}$$

Αυτά τα αποτελέσματα συνοψίζονται στον παρακάτω πίνακα της ανάλυσης της διασποράς του μοντέλου παλινδρόμησης, γνωστός και ως Πίνακας $ANOVA$ (*ANalysis Of VAriance*).

Πηγή	Αθροίσματα Τετραγώνων (SS)	Βαθμοί Ελευθερίας (df)	Μέσα Τετράγωνα (MS)
Παλινδρόμηση στα $X_1, X_2, ..., X_p$	$SSR = \hat{\beta}'X'Y - \frac{1}{n}(Y'1)^2$	p	$MS_R = \frac{SSR}{p}$
Υπόλοιπο (σφάλματα)	$RSS = Y'Y - \hat{\beta}'X'Y$	$n-p-1$	$MS_E = \frac{RSS}{n-p-1} = \hat{\sigma}^2$
Σύνολο	$SYY = Y'Y - \frac{1}{n}(Y'1)^2$	$n-1$	

Μπορούμε να προσδιορίσουμε τη σημαντικότητα της παλινδρόμησης συγκρίνοντας την τιμή του λόγου $F = \frac{MS_R}{MS_E}$ με την κρίσιμη τιμή $F_{p,n-p-1,\alpha}$ σε στάθμη σημαντικότητας α. Πιο συγκεκριμένα αν $\beta^* = (\beta_1, \beta_2, ..., \beta_p)'$, ελέγχουμε τη μηδενική υπόθεση

$$H_0 : \beta^* = 0,$$

έναντι της εναλλακτικής

$$H_1 : \beta^* \neq 0,$$

Αν $F > F_{p,n-p-1,\alpha}$, μπορούμε να συμπεράνουμε ότι η χρήση των μεταβλητών X_1, X_2, ..., X_p παρέχει ένα σημαντικά καλύτερο μοντέλο από αυτό που δεν περιέχει τις μεταβλητές αυτές, οπότε απορρίπτουμε την υπόθεση H_0 σε στάθμη σημαντικότητας α και χρησιμοποιούμε το μοντέλο (1.16). Αν $F < F_{p,n-p-1,\alpha}$ δεν έχουμε λόγο να απορρίψουμε την H_0, οπότε χρησιμοποιούμε το μοντέλο (1.17). Η αποδοχή της H_0 σημαίνει ότι από τα δεδομένα του πειράματος δεν προκύπτουν επαρκείς ενδείξεις για την ύπαρξη γραμμικής σχέσης μεταξύ των μεταβλητών Y και $X_1, X_2, ...X_p$. Αυτό δεν αποκλείει την ύπαρξη κάποιας άλλης σχέσης, π.χ. πολυωνυμικής, εκθετικής κλπ, μεταξύ των ίδιων μεταβλητών. Ο λόγος F θα ακολουθεί κατανομή $Fisher : F_{p,n-p-1}$ αν τα σφάλματα ακολουθούν κανονική κατανομή $N(0, \sigma^2)$ και επιπλέον η H_0 είναι αληθής.

1.3. Συντελεστής Προσδιορισμού R^2 ($Coefficient\ Of\ Determination$)

Αν διαιρέσουμε τη σχέση (1.19) με το SYY προκύπτει ότι:

$$\frac{SSR}{SYY} = 1 - \frac{RSS}{SYY} \tag{1.24}$$

Το αριστερό μέλος της (1.24) εκφράζει το ποσοστό της μεταβλητότητας του Y που εξηγείται από την παλινδρόμηση με τις μεταβλητές $X_1, X_2, ...X_p$ σε σχέση με τη συνολική μεταβλητότητα του Y. Το δεξί μέλος ισούται με 1 μείον την εναπομείνουσα ανεξήγητη μεταβλητότητα. Ως εκ τούτου, ορίζουμε ως **συντελεστή προσδιορισμού πολλαπλής παλινδρόμησης** R^2 την ποσότητα:

$$R^2 = \frac{SSR}{SYY} = 1 - \frac{RSS}{SYY} \tag{1.25}$$

Το R^2 υπολογίζεται εύκολα με τη βοήθεια του πίνακα $ANOVA$ και εκφράζει το πόσο ισχυρή είναι η σχέση μεταξύ του Y και των $X_1, X_2, ...X_p$ σύμφωνα με τα δεδομένα που

διαθέτουμε. Το γεγονός ότι $0 \leq R^2 \leq 1$ σημαίνει ότι όσο πιο κοντά στο 1 είναι το R^2, τόσο πιο ισχυρή είναι αυτή η σχέση, δηλαδή τόσο καλύτερη είναι η προσαρμογή στα δεδομένα μας, ενώ όσο πλησιάζει το 0 τόσο χειρότερη είναι η προσαρμογή αυτή.

Ένας δείκτης καλής εφαρμογής που συνδέεται με το R^2, είναι ο *προσαρμοσμένος συντελεστής προσδιορισμού πολλαπλής παλινδρόμησης* R_a^2 (*adjusted coefficient of determination*) που δίνεται από τη σχέση:

$$R_a^2 = 1 - \frac{RSS/(n-p-1)}{SYY/(n-1)} = 1 - (1-R^2)\frac{n-1}{n-p-1} \tag{1.26}$$

Τονίζουμε ότι το R^2 δεν είναι ο κατάλληλος δείκτης για να συγκρίνουμε ένα μοντέλο με q μεταβλητές με ένα μοντέλο με $p < q$ μεταβλητές, επειδή το R^2 αυξάνεται πάντα όταν μια νέα επεξηγηματική μεταβλητή εισέλθει στο μοντέλο και δεν μπορούμε να βγάλουμε ένα σωστό συμπέρασμα για τη σημαντικότητα της μεταβλητής αυτής. Σε αντίθεση με αυτό, το R_a^2 αποτελεί ένα "ποινικοποιημένο μέτρο καλής εφαρμογής" (*penalized measure of goodness of fit*), αφού υπάρχει πιθανότητα αν προσθέσουμε μία μεταβλητή στο μοντέλο να μειωθεί το R_a^2, οπότε να συμπεράνουμε ότι αυτή η μεταβλητή δεν είναι απαραίτητη για το μοντέλο. Αν όμως αυξηθεί το R_a^2, τότε η μεταβλητή αυτή είναι σημαντική για το μοντέλο και πρέπει να την συμπεριλάβουμε σε αυτό. Από τη σχέση (1.26), παρατηρούμε ότι το R_a^2 μπορεί να πάρει και αρνητικές τιμές, καθώς επίσης ότι είναι πάντα μικρότερο του R^2 αφού η "ποινή" (*penalty*) $\frac{n-1}{n-p-1}$ είναι ένας αριθμός μεγαλύτερος του 1.

Επίσης, ο συντελεστής προσδιορισμού R^2 μπορεί να χρησιμοποιηθεί και για τον έλεγχο της υπόθεσης $H_0 : \beta^* = 0$ αντί του F. Πράγματι, εύκολα αποδεικνύεται ότι:

$$F = \frac{n-p-1}{p}\frac{R^2}{1-R^2} \tag{1.27}$$

ή αλλιώς:

$$R^2 = \frac{pF/(n-p-1)}{1+pF/(n-p-1)} \tag{1.28}$$

1.4. Μεροληψία (*Bias*) Στους Εκτιμητές

Γνωρίζουμε ότι μια τυχαία μεταβλητή T είναι *αμερόληπτος εκτιμητής* της παραμέτρου θ όταν ισχύει $E(T) = \theta$. Όπως αποδείξαμε νωρίτερα, (σχέση 1.8), μελετώντας το μοντέλο $Y = X\beta + e$, με $EY = X\beta$ και $\hat{\beta} = (X'X)^{-1}X'Y$, ισχύει ότι $E\hat{\beta} = \beta$, δηλαδή ο εκτιμητής ελαχίστων τετραγώνων $\hat{\beta}$ είναι αμερόληπτος εκτιμητής του β. Τονίζουμε ότι το γεγονός αυτό ισχύει μόνο αν το μοντέλο που κατασκευάζουμε είναι το σωστό, δηλαδή για τα δεδομένα που διαθέτουμε ισχύει η σχέση $EY = X\beta$. Αν δεν είναι το σωστό, τότε οι εκτιμητές είναι μεροληπτικοί, δηλαδή $E(\hat{\beta}) \neq \beta$. Η τιμή της μεροληψίας:

$$bias = E(\hat{\beta}) - \beta \tag{1.29}$$

εξαρτάται, όπως θα δείξουμε, όχι μόνο από το μοντέλο που έχουμε προσαρμόσει, αλλά επίσης και από τις τιμές των επεξηγηματικών μεταβλητών που χρησιμοποιούνται στους υπολογισμούς. Όταν ένα προσχεδιασμένο πείραμα χρησιμοποιείται, η μεροληψία εξαρτάται τόσο από το σχέδιο του πειράματος, όσο και από το μοντέλο.

Έστω ότι προσαρμόζουμε το μοντέλο:

$$E(Y) = X_A \beta_A \tag{1.30}$$

όπου X_A είναι ο πίνακας σχεδιασμού που κατασκευάστηκε από τις παρατηρήσεις πάνω στο σύνολο μεταβλητών Α.

Η μέθοδος ελαχίστων τετραγώνων δίνει τον εκτιμητή:

$$\hat{\beta_A} = (X'_A X_A)^{-1} X'_A Y \tag{1.31}$$

Αν το μοντέλο αυτό είναι σωστό, τότε έχουμε ότι:

$$E(\hat{\beta_A}) = (X'_A X_A)^{-1} X'_A E(Y) = (X'_A X_A)^{-1} X'_A X_A \beta_A = \beta_A \tag{1.32}$$

Άρα, ο $\hat{\beta_A}$ είναι αμερόληπτος εκτιμητής του β_A.

Υποθέτουμε τώρα ότι προσαρμόζουμε ξανά το μοντέλο (1.30) έτσι ώστε ο $\hat{\beta_A}$ να εξακολουθεί να είναι το διάνυσμα των εκτιμώμενων συντελεστών. Έστω τώρα ότι αυτή τη φορά η πραγματική σχέση με την Y δεν είναι η (1.30) αλλά η:

$$E(Y) = X_A \beta_A + X_B \beta_B \tag{1.33}$$

Εδώ ο πίνακας σχεδιασμού X_B αφορά το σύνολο μεταβλητών Β.

Υπάρχουν, δηλαδή, οι όροι $X_B \beta_B$ που δεν συμπεριλάβαμε στη διαδικασία εκτίμησης που κάναμε. Προκύπτει τώρα ότι:

$$\begin{aligned} E(\hat{\beta_A}) &= (X'_A X_A)^{-1} X'_A E(Y) \\ &= (X'_A X_A)^{-1} X'_A (X_A \beta_A + X_B \beta_B) \\ &= (X'_A X_A)^{-1} X'_A X_A \beta_A + (X'_A X_A)^{-1} X'_A X_B \beta_B \\ &= \beta_A + S\beta_B \end{aligned} \tag{1.34}$$

όπου ο πίνακας:

$$S = (X'_A X_A)^{-1} X'_A X_B \tag{1.35}$$

καλείται *πίνακας μεροληψίας* (*bias matrix*). Παρατηρούμε ότι οι όροι $S\beta_B$ που χαρακτηρίζουν τη μεροληψία δεν εξαρτώνται μόνο από το προσαρμοσμένο μοντέλο και τα πραγματικά μοντέλα αλλά επίσης και από την επιλογή των πινάκων X_A και X_B. Επομένως, μπορούμε να επιλέξουμε τα σύνολα Α και Β έτσι ώστε να πετύχουμε την ελάχιστη μεροληψία.

1.5. Υπολογιστικές Μέθοδοι

Το βασικό πρόβλημά μας είναι να προσδιορίσουμε τη σχέση μεταξύ των μεταβλητών X_i, $i = 0, 1, ..., t$, που εισάγουμε και της εξαρτημένης μεταβλητής Y, όπως επίσης και το συνδυασμό των μεταβλητών X_i που περιγράφει με βέλτιστο τρόπο την Y. Ένας αντικειμενικός σκοπός της ανάλυσης αυτής είναι η επιλογή του υποσυνόλου των μεταβλητών που θα χρησιμοποιήσουμε στην τελική εξίσωση.

Για να καταλήξουμε σε μια τέτοια επιλογή, χρειαζόμαστε μοντέλα με διάφορους συνδυασμούς των μεταβλητών εισόδου. Αν το πλήθος τους, t, είναι μικρό, μπορούμε να υπολογίσουμε όλους τους 2^t συνδυασμούς, αλλά για μεγάλα t, αυτό είναι τελείως ασύμφορο.

Παρακάτω περιγράφουμε τις κυριότερες υπολογιστικές διαδικασίες κατά τις οποίες υπολογίζονται ορισμένοι ή ακόμα και όλοι οι δυνατοί συνδυασμοί των υποσυνόλων των επεξηγηματικών μεταβλητών.

1.5.1. Όλες οι πιθανές παλινδρομήσεις (*All Possible Regressions*)

Αν το t δεν είναι πολύ μεγάλο, μπορούμε να υπολογίσουμε όλα τα πιθανά μοντέλα που μπορούν να κατασκευαστούν από συνδυασμούς των επεξηγηματικών μεταβλητών, δηλαδή τα t μοντέλα στα οποία μόνο μία από τις μεταβλητές περιλαμβάνεται στο μοντέλο, τα $\begin{pmatrix} t \\ 2 \end{pmatrix}$ μοντέλα στα οποία δύο μόνο από τις μεταβλητές συμπεριλαμβάνονται κ.ο.κ. έως και το μοντέλο που περιλαμβάνει όλες τις μεταβλητές. Υπάρχουν στη βιβλιογραφία αποτελεσματικοί αλγόριθμοι που παρέχουν γρήγορο υπολογισμό όλων των πιθανών μοντέλων.

Η βασική ιδέα είναι να υπολογίσουμε τα 2^t υποσύνολα με τέτοιο τρόπο ώστε τα διαδοχικά υποσύνολα που υπολογίζουμε να διαφέρουν μεταξύ τους κατά μία μεταβλητή. Μπορούμε επίσης να αποφύγουμε να υπολογίσουμε περιττές αρχικά ποσότητες όπως τους συντελεστές παλινδρόμησης ή τον πίνακα $(X'X)^{-1}$. Όποιον αλγόριθμο πάντως κι αν ακολουθήσουμε πρέπει να έχουμε υπ'οψη μας τις απαιτήσεις αποθήκευσης, το πλήθος των υπολογισμών, το χρόνο που θα χρειαστεί ο υπολογιστής, την ακρίβεια και την ποσότητα των δοσμένων πληροφοριών.

1.5.2. Πολυβηματικές Μέθοδοι (*Stepwise Methods*)

Εξαιτίας του υπολογιστικού κόστους της προηγούμενης μεθόδου, έχουν προταθεί διάφορες μέθοδοι για τον υπολογισμό ενός μικρού μόνο πλήθους των υποσυνόλων είτε προσθέτοντας είτε απαλείφοντας μεταβλητές, μία κάθε φορά, σύμφωνα με ένα προκαθορισμένο κριτήριο. Τέτοιες διαδικασίες, που συχνά αναφέρονται ως *Πολυβηματικές Μέθοδοι* (*Stepwise Methods*), χωρίζονται σε δύο βασικές κατηγορίες: τη *Forward Selection* (*FS*) και την *Backward Elimination* (*BE*).

Μια σύντομη περιγραφή των δύο αυτών μεθόδων ακολουθεί παρακάτω.

α) *Forward Selection*

Κατά τη μέθοδο αυτή, ξεκινάμε με το μοντέλο που δεν περιέχει καμία μεταβλητή και στη συνέχεια προσθέτουμε μια μεταβλητή σε κάθε βήμα είτε μέχρις ότου όλες οι μεταβλητές εισέλθουν στην εξίσωση είτε μέχρι να ικανοποιηθεί ένα κριτήριο τερματισμού. Η μεταβλητή που προορίζεται να εισέλθει στην εξίσωση σε ένα βήμα είναι αυτή που θα δώσει το μεγαλύτερο λόγο F και επιπλέον αυτός ο λόγος θα είναι μεγαλύτερος από μία καθορισμένη τιμή. Αυτό σημαίνει ότι η μεταβλητή i προστίθεται στην εξίσωση η οποία περιλαμβάνει ήδη έστω p επεξηγηματικές μεταβλητές εφόσον ισχύει:

$$F_i = \max_i(\frac{RSS_p - RSS_{p+i}}{\hat{\sigma}^2_{p+i}}) > F_{in} \tag{1.36}$$

Εδώ, ο δείκτης $p+i$ αναφέρεται στο μοντέλο που περιλαμβάνει τις ήδη υπάρχουσες p μεταβλητές αλλά και τη μεταβλητή i που εισέρχεται σε αυτό, ενώ ο δείκτης p δηλώνει το μοντέλο με τις p μεταβλητές. Η φύση των υπολογισμών είναι τέτοια ώστε αν το F_{in} είναι αρκετά μεγάλο, οι υπολογισμοί θα τερματιστούν πριν συμπεριληφθούν όλες οι μεταβλητές στο μοντέλο. Συνήθως θεωρούμε ότι:

$$F_{in} = F_{1,n-p-1,\alpha} \tag{1.37}$$

για στάθμη σημαντικότητας α.

β) *Backward Elimination*

Εδώ, αρχίζοντας με την εξίσωση που περιέχει όλες τις μεταβλητές, απαλείφουμε μία σε κάθε βήμα. Έτσι, η μεταβλητή με το μικρότερο λόγο F απαλείφεται αν ο λόγος αυτός δεν ξεπερνά μια καθορισμένη τιμή. Δηλαδή, η μεταβλητή i διαγράφεται από την εξίσωση με p επεξηγηματικές μεταβλητές, εφόσον ισχύει:

$$F_i = \min_i(\frac{RSS_{p-i} - RSS_p}{\hat{\sigma}^2_p}) < F_{out} \tag{1.38}$$

Εδώ, το RSS_{p-i} δηλώνει το άθροισμα των τετραγώνων των σφαλμάτων που υπολογίζεται όταν η μεταβλητή i διαγράφεται από την ήδη υπάρχουσα εξίσωση p μεταβλητών. Είναι φανερό ότι αν το F_{out} είναι αρκετά μικρό, δε θα απαλειφθούν όλες οι μεταβλητές από το μοντέλο. Μπορούμε, αντιστοίχως, να θεωρήσουμε ότι:

$$F_{out} = F_{\alpha,1,n-p} \tag{1.39}$$

για στάθμη σημαντικότητας α.

γ) Άλλες αναφορές

Διάφοροι συγγραφείς προτείνουν συγκεκριμένες στάθμες σημαντικότητας α για κάθε μέθοδο, ενώ άλλοι προτείνουν διαφορετικά κριτήρια τερματισμού. Οι *Kennedy* και *Bancroft* (1971) αναπτύσσουν εκφράσεις για τη μεροληψία και το μέσο τετραγωνικό σφάλμα πρόβλεψης για τις FS και BE κάτω από περιοριστικές συνθήκες.

Έχουν προταθεί διάφοροι συνδυασμοί των δύο αυτών μεθόδων, ένας από τους οποίους συμβολίζεται με ES. Η μέθοδος ES είναι παρόμοια με την FS με τη διαφορά ότι σε κάθε βήμα ενδέχεται να διαγραφεί μια μεταβλητή, όπως συμβαίνει στην BE.

Έχουν αναφερθεί αρκετά μειονεκτήματα για τις πολυβηματικές διαδικασίες. Ένα από αυτά είναι ότι καμία από τις μεθόδους FS, BE, ES δεν εξασφαλίζει ότι το βέλτιστο υποσύνολο από ένα συγκεκριμένο πλήθος μεταβλητών εμφανίζεται. Επίσης, το γεγονός ότι προτείνεται μια σειρά σημαντικότητας για τις μεταβλητές ενδέχεται να είναι παραπλανητικό, αφού για παράδειγμα υπάρχει πιθανότητα η πρώτη μεταβλητή που εισέρχεται στο μοντέλο (στην FS μέθοδο) να μην είναι απαραίτητη όταν εισέλθουν κάποιες άλλες μεταβλητές στο μοντέλο ή μπορεί η πρώτη μεταβλητή που διαγράφεται στην BE μέθοδο να είναι η πρώτη που εισέρχεται στην FS μέθοδο. Οι *Gorman* και *Toman* (1966) παρατηρούν ότι είναι σπάνιο να υπάρχει ένα μόνο βέλτιστο υποσύνολο, αλλά ότι συνήθως υπάρχουν διάφορα εξίσου καλά υποσύνολα. Η συγκεκριμένη παρατήρηση θα μελετηθεί εκτενέστερα στο κεφάλαιο 2.

1.5.3. Βέλτιστα Υποσύνολα (*Optimal Subsets*)

Υπάρχει μια στοιχειώδης αλλά ουσιαστική αρχή στα προβλήματα ελαχιστοποίησης υπό περιορισμούς, η οποία αναφέρει ότι αν στο πρόβλημα προστεθούν περισσότεροι περιορισμοί, η βέλτιστη τιμή της αντικειμενικής συνάρτησης θα είναι ίση ή μεγαλύτερη από αυτή που υπολογίζεται στο αρχικό πρόβλημα. Για να εξετάσουμε με ποιό τρόπο αυτή η αρχή εφαρμόζεται στο πρόβλημα επιλογής υποσυνόλων, παρατηρούμε ότι το πρόβλημα μπορεί να θεωρηθεί ως ελαχιστοποίηση του αθροίσματος τετραγώνων των σφαλμάτων για το πλήρες μοντέλο, υπό τον περιορισμό ότι ορισμένοι συντελεστές είναι 0. Δηλαδή,

$$\begin{cases} minimize\ Q(\beta) \\ \\ \tau.\omega.\ \beta_i = 0,\ i \in I \end{cases} \tag{1.40}$$

Εδώ το $Q(\beta) = (Y - X\beta)'(Y - X\beta)$ είναι το άθροισμα τετραγώνων των σφαλμάτων και I είναι ένα συγκεκριμένο σύνολο δεικτών. Η αρχή της βελτιστοποίησης δείχνει ότι αν R_1 και R_2 είναι δύο σύνολα δεικτών και Q_1 και Q_2 τα αντίστοιχα αθροίσματα τετραγώνων των σφαλμάτων, τότε, αν το R_1 είναι υποσύνολο του R_2, προκύπτει ότι $Q_1 \leq Q_2$.

Έχουν προταθεί διάφοροι τρόποι αναζήτησης των βέλτιστων υποσυνόλων. Οι *Hocking* και *Leslie* (1967) ανέπτυξαν αλγορίθμους που εξασφαλίζουν ότι το βέλτιστο υποσύνολο για κάθε πλήθος μεταβλητών βρίσκεται υπολογίζοντας μόνο ένα ποσοστό των 2^t υποσυνόλων και όχι όλα τα υποσύνολα. Οι *LaMotte* και *Hocking* (1970) δημιούργησαν το πρόγραμμα *SELECT* κατά το οποίο το βέλτιστο υποσύνολο βρίσκεται υπολογίζοντας μόνο ένα μικρό ποσοστό από όλα τα δυνατά υποσύνολα. Επίσης, μπορούμε εκτός από το βέλτιστο υποσύνολο κάθε μεγέθους p, να λάβουμε υπ'οψη κάποια "σχεδόν βέλτιστα" υποσύνολα. Τέλος, έχει παρατηρηθεί ότι το πλήθος των υποσυνόλων που πρέπει να υπολογιστούν για να προσδιοριστούν τα βέλτιστα υποσύνολα εξαρτάται από τα δεδομένα που διαθέτουμε αλλά και από το πλήθος των μεταβλητών.

1.6. Κριτήρια Επιλογής των Μεταβλητών

1.6.1. Χρήσεις παλινδρόμησης

Το κριτήριο που χρησιμοποιείται για να αποφασίσουμε ποιό υποσύνολο από τις επεξηγηματικές μεταβλητές θα επιλέξουμε πρέπει να σχετίζεται με το πώς σκοπεύουμε να χρησιμοποιήσουμε το μοντέλο παλινδρόμησης. Πιθανές χρήσεις της εξίσωσης παλινδρόμησης είναι οι εξής:

- Απλή περιγραφή
- Πρόβλεψη και εκτίμηση
- Παρεκβολή
- Εκτίμηση των παραμέτρων
- Έλεγχος
- Κατασκευή μοντέλου

Αν ο στόχος μας είναι να αποκτήσουμε μια καλή *περιγραφή* της εξαρτημένης μεταβλητής και το κριτήριο για να προσαρμόσουμε τα δεδομένα είναι η ελαχιστοποίηση του RSS, τότε αναζητούμε εξισώσεις με όσο το δυνατόν μικρότερα αθροίσματα τετραγώνων των σφαλμάτων. Με αυτή την έννοια, η καλύτερη λύση είναι να κρατήσουμε όλες τις μεταβλητές αλλά σε μερικές περιπτώσεις δεν υπάρχει ουσιαστική διαφορά αν κάποιες μεταβλητές απαλειφθούν. Οι πιο πολλοί χρήστες προτιμούν το R^2 ως ένα ισοδύναμο μέτρο το οποίο αφού κυμαίνεται μεταξύ του 0 και του 1, είναι εύκολο να εξηγηθεί.

Η διάκριση μεταξύ της *πρόβλεψης* μιας μέλλουσας τιμής της εξαρτημένης μεταβλητής και της *εκτίμησης* μιας μέσης τιμής αυτής για συγκεκριμένα δεδομένα σχολιάζεται σε πολλά άρθρα σχετικά με την παλινδρόμηση. Στην πρώτη περίπτωση, στόχος μας είναι να χρησιμοποιήσουμε το μοντέλο παλινδρόμησης για να προβλέψουμε την τιμή της εξαρτημένης μεταβλητής του προβλήματός μας με βάση κάποιες νέες τιμές που εισάγουμε για τις επεξηγηματικές μεταβλητές. Στη δεύτερη περίπτωση, δεν εισάγουμε νέες τιμές, αλλά εκτιμούμε την εξαρτημένη μεταβλητή με βάση τα δεδομένα που χρησιμοποιήσαμε για να κατασκευάσουμε το μοντέλο παλινδρόμησης.

Ο κίνδυνος της *παρεκβολής* (*extrapolation*) έξω από την κλίμακα των δεδομένων που χρησιμοποιούμε για να υπολογίσουμε τους εκτιμητές είναι εμφανής αφού το μοντέλο μπορεί να μην είναι πλέον εφαρμόσιμο. Όμως, ακόμα κι αν το μοντέλο είναι κατάλληλο για τιμές των επεξηγηματικών μεταβλητών μέσα στην κλίμακα αυτή, ενδέχεται να είναι εφαρμόσιμο αλλά αναξιόπιστο έξω από την κλίμακα αυτή, λόγω των ασταθών, λόγω πολυσυγγραμμικότητας, συντελεστών παλινδρόμησης.

Αν η *εκτίμηση των παραμέτρων* είναι ο στόχος μας, τότε πρέπει να εξετάσουμε τη μεροληψία που προκύπτει από την απαλοιφή των μεταβλητών αλλά και από την εκτιμώμενη διασπορά. Αν ο πίνακας X είναι ιδιάζων, προτείνονται μεροληπτικές διαδικασίες, οι οποίες δίνουν πιο ευσταθείς εκτιμητές, αλλά και μπορούν να οδηγήσουν σε μια εξίσωση πρόβλεψης που θα είναι πιο αποτελεσματική στην περίπτωση της παρεκβολής.

Το ζήτημα του *ελέγχου* προκύπτει όταν στόχος μας είναι να ελέγξουμε που κυμαίνονται οι τιμές της εξαρτημένης μεταβλητής όταν μεταβάλουμε την κλίμακα των δεδομένων που εισάγουμε. Σε αυτή την περίπτωση, χρειαζόμαστε ακριβείς εκτιμητές για τους συντελεστές παλινδρόμησης.

Σε πολλές έρευνες, στόχος μας είναι να *αναπτύξουμε ένα μοντέλο* για την εξαρτημένη μεταβλητή ως συνάρτηση των παρατηρήσεων των επεξηγηματικών μεταβλητών. Σε μια τέτοια κατάσταση, είναι απαραίτητο να επιλέξουμε μια κατάλληλη μέθοδο για τον προσδιορισμό των μεταβλητών που θα χρησιμοποιήσουμε αλλά και των σχέσεων μεταξύ τους.

1.6.2. Συναρτήσεις-Κριτήρια

Έχοντας υπ'οψη την παραπάνω λίστα των χρήσεων της παλινδρόμησης, έχουν προταθεί ορισμένα κριτήρια για να αποφασίσουμε ποιό υποσύνολο θα επιλέξουμε. Αυτά τα κριτήρια αφορούν τη συμπεριφορά διαφόρων συναρτήσεων ως προς τις μεταβλητές που συμπεριλαμβάνονται μέσα στο υποσύνολο. Πολλές από αυτές τις συναρτήσεις-κριτήρια είναι απλές συναρτήσεις του αθροίσματος τετραγώνων των σφαλμάτων για την εξίσωση με p επεξηγηματικές μεταβλητές που συμβολίζεται με RSS_p. Τα πιο συνηθισμένα κριτήρια είναι:

1. Το μέσο τετράγωνο των υπολοίπων:

$$RMS_p = \frac{RSS_p}{n-p}$$

2. Ο συντελεστής προσδιορισμού πολλαπλής παλινδρόμησης:

$$R_p^2 = 1 - \frac{RSS_p}{SYY}$$

3. Ο προσαρμοσμένος συντελεστής προσδιορισμού πολλαπλής παλινδρόμησης:

$$\overline{R}_p^2 = 1 - (1 - R_p^2)\frac{n-1}{n-p-1}$$

4. Η μέση διασπορά πρόβλεψης:

$$J_p = (n+p)RMS_p/n$$

5. Το συνολικό τετραγωνικό σφάλμα ή στατιστικό C_p του *Mallows* (1964):

$$C_p = RSS_p/\hat{\sigma}^2 + 2p - n$$

6. Ο μέσος όρος του μέσου τετραγωνικού σφάλματος:

$$S_p = RMS_p/(n-p-1)$$

7. Το κανονικοποιημένο άθροισμα τετραγώνων των σφαλμάτων:

$$RSS_p^* = e_p' D_p^{-1} e_p,$$

όπου $e_p = Y - \hat{Y}_p$ και $D_p = diag(I - X_p(X_p'X_p)^{-1}X_p')$

8. Το άθροισμα τετραγώνων της πρόβλεψης:

$$PRESS = e_p' D_p^{-2} e_p$$

Κατά τη διαδικασία $PRESS$ που χρησιμοποιεί το παραπάνω κριτήριο, γίνεται ο εξής έλεγχος για την καταλληλότητα του μοντέλου που έχουμε κατασκευάσει έχοντας στη διάθεσή μας n παρατηρήσεις. Αγνοούμε μία από τις n παρατηρήσεις και κατασκευάζουμε το μοντέλο σύμφωνα με τις εναπομείνουσες $n-1$ παρατηρήσεις. Έπειτα, σύμφωνα με αυτό το μοντέλο, υπολογίζουμε την τιμή της εξαρτημένης μεταβλητής από τις τιμές των επεξηγηματικών μεταβλητών για την παρατήρηση που αγνοήσαμε και βλέπουμε πόσο συμπίπτει με την πραγματική της τιμή που ήδη ξέρουμε. Επαναλαμβάνουμε τη διαδικασία αυτή n φορές αγνοώντας κάθε φορά μία διαφορετική παρατήρηση. Τέλος, υπολογίζουμε το συνολικό σφάλμα της πρόβλεψης, που δίνεται από τον παραπάνω τύπο.

1.7. Μεροληπτική Εκτίμηση

Οι μέχρι τώρα υπολογιστικές μέθοδοι που χρησιμοποιήσαμε για την κατασκευή μοντέλων βασίζονταν στη μέθοδο των ελαχίστων τετραγώνων. Εφόσον δηλαδή αποφασίσουμε ποιές μεταβλητές θα συμπεριληφθούν στο μοντέλο, χρησιμοποιούμε τη μέθοδο των ελαχίστων τετραγώνων για να εκτιμήσουμε τις παραμέτρους.

Όμως, υπάρχουν αρκετά άρθρα που προτείνουν εναλλακτικούς τρόπους οι οποίοι, αν και παράγουν μεροληπτικούς εκτιμητές, είναι προτιμότεροι λόγω ευσταθείας και λόγω του ότι αντιμετωπίζουν προβλήματα συγγραμικότητας. Εφόσον θέλουμε να χρησιμοποιήσουμε τελικά την εξίσωση παλινδρόμησης, η μεροληψία στους εκτιμητές μπορεί να μην αποτελέσει πρόβλημα. Πράγματι, όταν δε χρησιμοποιούμε το πλήρες μοντέλο, αλλά ένα περιορισμένο, οι εκτιμητές είναι γενικά μεροληπτικοί. Το ζήτημα είναι αν οι εκτιμητές ή οι μεταβλητές που προκύπτουν τελικά είναι καταλληλότεροι ή όχι.

Στην παράγραφο αυτή, αναλύονται τρεις διαδικασίες μεροληπτικής εκτίμησης:

- Η *Stein shrinkage*
- Η *Ridge regression* και
- Η *Principal component regression* και κάποιες παραλλαγές της.

Υποθέτουμε ότι όλες οι μεταβλητές, εξαρτημένες και ανεξάρτητες, έχουν κανονικοποιηθεί. Επομένως, τα στοιχεία των πινάκων $X'X$ και $X'Y$ ισούνται με τις αντίστοιχες συσχετίσεις των μεταβλητών. Σημειώνουμε ότι το σύμβολο $\hat{\beta}$ θα αναφέρεται πάντα στον

εκτιμητή ελαχίστων τετραγώνων ενώ οι υπόλοιποι εκτιμητές θα ξεχωρίζουν από κατάλληλους δείκτες.

1.7.1. *Stein Shrinkage*

Υποθέτουμε ότι η εξαρτημένη και οι ανεξάρτητες μεταβλητές ακολουθούν κανονική κατανομή με γνωστές μέσες τιμές. Ο *Stein* (1960) θεωρεί τη συνάρτηση απώλειας:

$$L = \frac{(\tilde{\beta} - \beta)'\Gamma(\tilde{\beta} - \beta)}{\sigma^2} \tag{1.41}$$

όπου Γ είναι ο πίνακας συνδιασποράς του διανύσματος των τιμών των μεταβλητών που εισάγουμε, σ^2 είναι η διασπορά των υπολοίπων για τη δεσμευμένη κατανομή της $y|x$ και $\tilde{\beta}$ ένας οποιοσδήποτε εκτιμητής του β. Αν θεωρήσουμε ότι $\tilde{\beta} = \hat{\beta}$, όπου $\hat{\beta}$ εδώ είναι είτε ο εκτιμητής ελαχίστων τετραγώνων είτε ο εκτιμητής μέγιστης πιθανοφάνειας, τότε ο κίνδυνος (αναμενόμενη απώλεια) ισούται με τη σταθερά $t/(n-t-1)$, όπου t το πλήθος των επεξηγηματικών μεταβλητών. Ο *Stein* έδειξε ότι ο εκτιμητής:

$$\tilde{\beta} = (1 - \frac{b(1-R^2)}{a(1-R^2)+R^2})\hat{\beta}, \tag{1.42}$$

για κατάλληλες επιλογές των a και b, δίνει μικρότερο κίνδυνο από τον $\hat{\beta}$. Συγκεκριμένα, παρατήρησε ότι ο κίνδυνος για $a = 0$ είναι μικρότερος από τη μέγιστη πιθανοφάνεια αν $b = (t-2)/(n-t+2)$ και πρότεινε τον εκτιμητή $\tilde{\beta}_s = c\hat{\beta}$, όπου

$$c = \max[(1 - \frac{t-2}{n-t+2}\frac{1-R^2}{R^2}), 0] \tag{1.43}$$

Γεωμετρικά, ο *Stein* προτείνει να συρρικνώσουμε τον εκτιμητή ελαχίστων τετραγώνων προς το 0. Με άλλα λόγια, ο $\tilde{\beta}_s$ είναι η λύση του προβλήματος:

$$\begin{cases} minimize_{\beta}(\beta - \hat{\beta})'(\beta - \hat{\beta}) \\ \\ \tau.\omega.\ \beta'\beta \leq d^2 \end{cases} \tag{1.44}$$

όπου η ακτίνα d καθορίζεται από τη σταθερά c.

1.7.2. *Ridge Regression*

Οι *Hoerl* και *Kennard* (1970) πρότειναν το μεροληπτικό εκτιμητή *ridge* για προβλήματα όπου τα διανύσματα των τιμών των επεξηγηματικών μεταβλητών δεν είναι ορθογώνια. Συγκεκριμένα, μελέτησαν τον εκτιμητή:

$$\beta(k) = (X'X + kI)^{-1}X'Y \tag{1.45}$$

όπου η σταθερά k καθορίζεται από το "ίχνος *ridge*" (*ridge trace*), δηλαδή τα γραφήματα του $\beta(k)$ ως προς το k.

Αυτό που οδήγησε στην ιδέα της *ridge regression* ήταν το γεγονός ότι μοντέλα για τα οποία ο πίνακας $X'X$ έχει μη-ομοιόμορφη δομή ιδιοτιμών μπορεί να οδηγήσει σε εκτιμητές ελαχίστων τετραγώνων που είναι μακριά από το πραγματικό τους σημείο. Για να το δούμε αυτό, έστω L_1 η ευκλείδια απόσταση του εκτιμητή ελαχίστων τετραγώνων, $\hat{\beta}$, από το πραγματικό του σημείο β. Τότε αν λ_i, $i = 1, 2, ..., t$ είναι οι ιδιοτιμές του $X'X$, αποδεικνύεται ότι:

$$E[L_1^2] = \sigma^2 \sum_{i=1}^{t} \frac{1}{\lambda_i} \tag{1.46}$$

Αν ένα ή περισσότερα από τα λ_i είναι πολύ μικρά, είναι φανερό ότι το $E[L_1^2]$ θα είναι πολύ μεγάλο άρα ο εκτιμητής $\hat{\beta}$ θα έχει απομακρυνθεί πολύ από την πραγματική του τιμή β.

Ο εκτιμητής *ridge* είναι παρόμοιος με τον εκτιμητή *Stein* με την έννοια ότι προκαλείται ξανά συρρίκνωση στον εκτιμητή ελαχίστων τετραγώνων προς το 0, αλλά σε αυτή την περίπτωση, η συρρίκνωση έχει γίνει σε συνάρτηση με τον πίνακα $X'X$. Πιο συγκεκριμένα, ο εκτιμητής *ridge*, ο οποίος μπορεί να γραφτεί αλλιώς ως εξής:

$$\tilde{\beta}_R = (I + k(X'X)^{-1})^{-1}\hat{\beta} \tag{1.47}$$

είναι η λύση του προβλήματος:

$$\begin{cases} minimize_\beta \ (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ \\ \tau.\omega. \ \beta'\beta \leq d^2 \end{cases} \tag{1.48}$$

όπου η ακτίνα d εξαρτάται από το k.

Πολλοί συγγραφείς έχουν μελετήσει την επιλογή της σταθεράς k στην (1.47). Οι *Hoerl* και *Kennard* (1970) πρότειναν για τη σταθερά k την τιμή η οποία δίνει τον εκτιμητή με την ελάχιστη μέση απόσταση του $\hat{\beta}$ από τον β. Πρότειναν να ερευνήσει κανείς το "*ridge trace*" για να εκτιμήσει το k. Άλλοι συγγραφείς πρότειναν εναλλακτικούς τρόπους για την εκτίμηση του k. Οι *Marquardt* και *Snee* (1973) πρότειναν να χρησιμοποιηθεί η τιμή του k για την οποία η μέγιστη VIF (*Variance Inflation Factor*) είναι μεταξύ του 1 και του 10 και μάλιστα πιο κοντά στο 1. Η VIF, που σχετίζεται με κάθε συντελεστή ξεχωριστά, εκφράζει πόσο μεγαλώνει η διασπορά αυτού του συντελεστή λόγω των συσχετίσεων μεταξύ των μεταβλητών. Συγκεκριμένα, τα VIF είναι τα διαγώνια στοιχεία του πίνακα:

$$Var(\tilde{\beta}_R)/\sigma^2 = (X'X + kI)^{-1}X'X(X'X + kI)^{-1} \tag{1.49}$$

Ο *Mallows* (1973) επέκτεινε την ιδέα των $C_p - plots$ σε $C_k - plots$ τα οποία μπορούν να χρησιμοποιηθούν για να καθοριστεί η τιμή του k. Συγκεκριμένα, πρότεινε το γράφημα του C_k ως προς το V_k, όπου

$$C_k = \frac{RSS_k}{\hat{\sigma}^2} - n + 2 + 2tr(XL),$$

$$V_k = 1 + tr(X'XLL')$$

και

$$L = (X'X + kI)^{-1}X'$$

Εδώ, το RSS_k είναι το άθροισμα τετραγώνων των σφαλμάτων σε συνάρτηση με το k. Προτείνεται να διαλέξουμε το k που ελαχιστοποιεί το C_k.

Άλλη αναφορά είναι αυτή του *Farebrother* (1975) ο οποίος προτείνει την τιμή $k = \hat{\sigma}^2/\hat{\beta}'\hat{\beta}$. Επίσης, παρατηρούμε ότι αν $X'X = I$, τότε η επιλογή του k που ελαχιστοποιεί το $E[L_1^2]$ είναι η $k = t\sigma^2/\beta'\beta$. Ο *Hoerl* (1975) πρότεινε την επιλογή $k = t\hat{\sigma}^2/\hat{\beta}'\hat{\beta}$ ως γενικό κανόνα. Οι έρευνές τους έδειξαν ότι ο εκτιμητής που προκύπτει δίνει εκτιμητές συντελεστών με μικρότερο μέσο τετραγωνικό σφάλμα από αυτό που δίνει ο εκτιμητής ελαχίστων τετραγώνων. Οι *Hoerl* και *Kennard* (1975) πρότειναν την αναδρομική διαδικασία όπου $k_{i+1} = t\hat{\sigma}^2/\beta_i'\beta_i$ με $\beta_i = \tilde{\beta}_R(k_i)$. Άλλοι συγγραφείς εξέτασαν περιπτώσεις όπου οι επεξηγηματικές μεταβλητές είναι 2 ή 3 και απέδειξαν ότι σε αυτές τις περιπτώσεις ο εκτιμητής *ridge* ενδέχεται να είναι χειρότερος από τον εκτιμητή ελαχίστων τετραγώνων. Μια φυσική επέκταση του εκτιμητή *ridge* είναι να θεωρήσουμε το διαγώνιο πίνακα K αντί του kI.

Ως ένα τελικό σχόλιο πάνω στη *ridge regression*, μπορεί να αποδειχθεί ότι ο εκτιμητής *ridge* είναι ισοδύναμος με αυτόν των ελαχίστων τετραγώνων όπου τα δεδομένα έχουν αυξηθεί κατά ένα φανταστικό παράγοντα τέτοιο ώστε η εξαρτημένη μεταβλητή είναι 0 και ένας διαγώνιος πίνακας έχει προστεθεί στον $X'X$. Προτείνεται λοιπόν να συλλεχθούν επιπλέον δεδομένα που θα βελτιώσουν τη σταθερότητα του πίνακα $X'X$.

1.7.3. *Principal Component Regression*

Υπάρχουν πολλές αιτίες πολυσυγγραμικότητας σε μεταβλητές παλινδρόμησης. Η ύπαρξη μικρών ιδιοτιμών για τον πίνακα $X'X$ είναι προειδοποίηση της παρουσίας ενός ή περισσοτέρων τέτοιων προβλημάτων. Είναι φανερό ότι αν υπάρχουν s μηδενικές ιδιοτιμές, το πλήθος των μεταβλητών που εισάγουμε μπορεί να μειωθεί κατά s. Αν υπάρχουν ιδιοτιμές που είναι κοντά στο 0, η κατάσταση δεν είναι ξεκάθαρη. Ενδέχεται να εκφράζουν πραγματικές γραμμικές εξαρτήσεις και η διαφορά τους από το 0 μπορεί να οφείλεται σε υπολογιστικά λάθη. Μπορεί όμως να είναι μη μηδενικές, αλλά να μαρτυρούν ισχυρές εξαρτήσεις. Η διαδικασία που θα ακολουθήσουμε σε αυτή την περίπτωση δεν είναι φανερή. Η *ridge regression* αγνοεί τη φύση της εξάρτησης και τροποποιεί τα δεδομένα ώστε να δώσει μεροληπτικούς εκτιμητές. Πολλοί συγγραφείς προτείνουν το μετασχηματισμό σε χώρο ορθογώνιων ανεξαρτήτων μεταβλητών που αντιστοιχούν σε μικρές ιδιοτιμές. Εξετάζουμε παρακάτω αυτή τη διαδικασία.

Έστω Λ ο διαγώνιος πίνακας των ιδιοτιμών λ_i του πίνακα $X'X$ και T ο ορθογώνιος πίνακας των ιδιοδιανυσμάτων του. Δηλαδή ισχύει η σχέση:

$$T'X'XT = \Lambda \tag{1.50}$$

και

$$T'T = I \tag{1.51}$$

Αν θέσουμε $Z = XT$ και $\beta = T\gamma$, το μοντέλο (1.2) γίνεται:

$$Y = Z\gamma + e \tag{1.52}$$

Ο εκτιμητής ελαχίστων τετραγώνων αυτού του μοντέλου καθορίζεται τότε από τη λύση των κανονικών εξισώσεων (1.6), οι οποίες γράφονται ως εξής:

$$\Lambda\hat{\gamma} = Z'Y \tag{1.53}$$

ή:

$$\hat{\beta} = T\hat{\gamma} \tag{1.54}$$

Τώρα, αν s από τις ιδιοτιμές αυτές είναι 0, προκύπτει ότι οι αντίστοιχες στήλες του Z είναι μηδενικές, άρα αυτές οι μεταβλητές απαλείφονται από το μοντέλο (1.52). Η εξίσωση (1.53) τότε έχει διάσταση $g = t - s$. Γράφουμε λοιπόν τους πίνακες Λ και T, καθώς και την παράμετρο γ ως εξής:

$$T = (T_g, T_s),$$

$$\gamma' = (\gamma'_g, \gamma'_s)$$

και

$$\Lambda = diag(\Lambda_g, \Lambda_s),$$

Οπότε η εξίσωση (1.53) γράφεται ως εξής:

$$\Lambda_g\hat{\gamma}_g = T'_g X'Y$$

και η (1.54) γράφεται:

$$\beta_g^+ = T_g\hat{\gamma}_g \tag{1.55}$$

Η μέθοδος αυτή έγινε γνωστή ως "*principal component regression*" και είναι βασισμένη στην παραπάνω ανάλυση. Δηλαδή, αν ο X είναι βαθμού t, αλλά s από τις ιδιοτιμές είναι πολύ μικρές ή ίσες με 0, τις θέτουμε όλες ίσες με 0 και υπολογίζουμε τον εκτιμητή του β από τη σχέση (1.55).

Είναι ενδιαφέρον να συγκρίνουμε αυτή τη διαδικασία με τη *Stein shrinkage* και τη *ridge regression*. Συγκεκριμένα, παρατηρούμε ότι η (1.55) είναι η λύση του προβλήματος:

$$\begin{cases} minimize_\beta\ (\beta - \hat{\beta})'X'X(\beta - \hat{\beta}) \\ \\ \tau.\omega.\ T'_s\beta = 0 \end{cases} \tag{1.56}$$

Μια εναλλακτική έκφραση για την (1.55) είναι η:

$$\beta_g^+ = (I - T_s T_s')\hat{\beta} \tag{1.57}$$

που θυμίζει τη σχέση (1.47) για τον εκτιμητή *ridge* αλλά είναι διαφορετική.

Αν πάρουμε την περίπτωση όπου $t = 2$ και $g = s = 1$, παρατηρούμε ότι:

$$X'X = \begin{pmatrix} 1 & \delta \\ \delta & 1 \end{pmatrix}$$

Οι ιδιοτιμές του πίνακα $X'X$ είναι οι $\lambda_1 = 1 + \delta$ και $\lambda_2 = 1 - \delta$, ενώ τα ιδιοδιανύσματα δίνονται από τον πίνακα:

$$T = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

Επομένως, αν το δ, για παράδειγμα, είναι κοντά στο 1, τότε το λ_2 είναι κοντά στο 0. Άρα, ο γραμμικός περιορισμός της (1.56) γίνεται $\beta_1 = \beta_2$.

Μια παραλλαγή της *principal component regression* που ο *Webster* (1974) πρότεινε καλείται "*latent root regression*". Για να συνδέσουμε την πρότασή του με την παρούσα ανάπτυξη, θέτουμε $A = [Y|X]$ ο οποίος θα είναι ο $n \times (t + 1)$ πίνακας των κανονικοποιημένων μεταβλητών. Έστω ότι ο Θ δηλώνει το διαγώνιο πίνακα των ιδιοτιμών, θ_i, του $A'A$ και ο H δηλώνει τον αντίστοιχο ορθογώνιο πίνακα των ιδιοδιανυσμάτων, h_i. Δηλαδή ισχύουν οι σχέσεις:

$$H'A'AH = \Theta \tag{1.58}$$

και

$$H'H = I \tag{1.59}$$

Επιπλέον, έστω ότι η πρώτη γραμμή του H συμβολίζεται με h' και τα υπόλοιπα στοιχεία του με H_t. Τότε, ο πίνακας Η γράφεται ως εξής:

$$H = \begin{pmatrix} h' \\ H_t \end{pmatrix} \tag{1.60}$$

Μπορούμε τώρα εύκολα να δείξουμε ότι ο εκτιμητής ελαχίστων τετραγώνων του β δίνεται από τη σχέση:

$$\hat{\beta} = -H_t \upsilon \tag{1.61}$$

όπου υ είναι η λύση του προβλήματος:

$$\begin{cases} minimize\ \upsilon'\Theta\upsilon \\ \\ \tau.\omega.\ h'\upsilon = 1 \end{cases} \tag{1.62}$$

Για να το διαπιστώσουμε αυτό, θέτουμε $\alpha = Hv$ και παρατηρούμε ότι η λύση είναι η $\hat{\alpha}' = (1, -\hat{\beta})$.

Ο *Webster* (1974) παρατήρησε ότι αν $\theta_i = 0$ για κάποιο i, τότε υπάρχει μια ακριβής σχέση μεταξύ εξαρτημένης και επεξηγηματικών μεταβλητών. Αν, επιπλέον, το αντίστοιχο στοιχείο του h' είναι ίσο με 0, τότε υπάρχει γραμμική εξάρτηση μεταξύ των επεξηγηματικών μεταβλητών. Βασισμένος σε αυτή την παρατήρηση, συμπέρανε ότι αν αυτές οι δύο ποσότητες δεν είναι ίσες με 0 αλλά είναι μικρές, το πρόβλημα είναι ασταθές. Κατά συνέπεια, πρότεινε αυτές οι ποσότητες να θεωρηθούνται ίσες με 0. Αυτό σημαίνει ότι τα αντίστοιχα στοιχεία του v είναι ίσα με 0, ή ισοδύναμα, το v είναι η λύση του προβλήματος:

$$\begin{cases} minimize\ v'\Theta v \\ \\ \tau.\omega.\ h'v = 1 \\ \\ (0\ I_s)v = 0 \end{cases} \tag{1.63}$$

Η (1.63) φανερώνει ότι έχουμε θέσει τα τελευταία s στοιχεία του v ίσα με 0.

Κεφάλαιο 2

ΕΠΙΛΟΓΗ ΜΕΤΑΒΛΗΤΩΝ - ΚΑΤΑΣΚΕΥΗ ΜΟΝΤΕΛΟΥ

Σε αυτό το κεφάλαιο αναλύονται συγκεκριμένες προτάσεις για την αντιμετώπιση του προβλήματος της επιλογής ενός βέλτιστου υποσυνόλου των επεξηγηματικών μεταβλητών που θα συμπεριληφθούν στο γραμμικό μοντέλο που θέλουμε να κατασκευάσουμε. Οι *R.R.Hocking* και *R.N.Leslie* (1967) υπολογίζουν τη μείωση του αθροίσματος των τετραγώνων λόγω παλινδρόμησης (SSR) κάθε φορά που απαλείφουμε μια μεταβλητή από το μοντέλο. Με αυτόν τον τρόπο, δημιουργείται μια σειρά σημαντικότητας των μεταβλητών και σύμφωνα με αυτή εισέρχονται οι μεταβλητές στο μοντέλο. Οι *J.W.Gorman* και *R.J.Toman* (1966) χρησιμοποιούν πολυβηματική παλινδρόμηση για την επιλογή των μεταβλητών με βάση το C_p κριτήριο. Οι *W.J.Kennedy* και *T.A.Bancroft* (1971) εκτελούν επαναλαμβανόμενους ελέγχους σημαντικότητας των μεταβλητών με βάση δύο διαδικασίες. Επίσης, υπολογίζουν τη μεροληψία και το μέσο τετραγωνικό σφάλμα της εκτιμώμενης τιμής της εξαρτημένης μεταβλητής για τις δύο αυτές διαδικασίες. Έπειτα, αναλύονται δύο άρθρα των *H.J.Larson* και *T.A.Bancroft* (1963). Στο πρώτο άρθρο, επιλέγουν το κατάλληλο υποσύνολο υπολογίζοντας τη μεροληψία και το μέσο τετραγωνικό σφάλμα του εκτιμητή της εξαρτημένης μεταβλητής με βάση δύο μεθόδους. Και στο δεύτερο άρθρο υπολογίζουν τις ποσότητες αυτές αλλά εδώ επιλέγουν μεταξύ του πλήρους μοντέλου και ενός περιορισμένου που έχει προκαθοριστεί από τον ερευνητή. Στο τελευταίο άρθρο που περιγράφεται, ο *T.A.Bancroft* (1944) χρησιμοποιεί προκαταρτικούς ελέγχους σημαντικότητας για τους συντελεστές παλινδρόμησης ώστε να επιλέξει το κατάλληλο μοντέλο.

2.1. Επιλογή του βέλτιστου υποσυνόλου των μεταβλητών με βάση την ταξινόμησή τους ως προς τη μείωση του αθροίσματος των τετραγώνων λόγω παλινδρόμησης όταν μια μεταβλητή απαλειφθεί από το μοντέλο

Οι *R.R.Hocking* και *R.N.Leslie* (1967) επισημαίνουν ότι κατά την επιλογή του βέλτιστου υποσυνόλου ή υποσυνόλων των επεξηγηματικών μεταβλητών για μια γραμμική παλινδρόμηση δύο είναι οι βασικές προϋποθέσεις. Πρώτον, πρέπει να καθοριστεί το κριτήριο επιλογής μεταξύ δύο υποσυνόλων. Δεύτερον, πρέπει να μειωθεί το υπολογιστικό κόστος. Οι παραπάνω συγγραφείς, λοιπόν, χρησιμοποιώντας ως βασικό κριτήριο το στατιστικό C_p,

αναπτύσσουν μια διαδικασία που υποδεικνύει καλές παλινδρομήσεις με το ελάχιστο των υπολογισμών.

Το πρόβλημα της γραμμικής παλινδρόμησης αφορά την εκτίμηση των συντελεστών του γραμμικού μοντέλου:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + e \tag{2.1}$$

από ένα σύνολο n παρατηρήσεων, όπου το σφάλμα e υποθέτουμε ότι έχει συνιστώσες e_i, $i = 1, ..., k$ οι οποίες ακολουθούν κατανομή $N(0, \sigma^2)$. Ως γνωστόν, από τη λύση των κανονικών εξισώσεων:

$$X'X\beta = X'Y \tag{2.2}$$

που προέκυψαν από την ελαχιστοποίηση του αθροίσματος τετραγώνων των σφαλμάτων, εκτιμούμε τους συντελεστές $\hat{\beta}_i$, $i = 1, ..., k$. Αν όλες οι πιθανές παλινδρομήσεις (2^k) υπολογιστούν, τότε από το γράφημα των μέσων τετραγώνων των σφαλμάτων e_i ως προς το πλήθος των μεταβλητών, μπορούμε να επιλέξουμε τη βέλτιστη παλινδρόμηση.

Χρησιμοποιούμε το στατιστικό C_p του $Mallows$ (1964) που είχαμε αναφέρει στο προηγούμενο κεφάλαιο:

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} - (n - 2p) \tag{2.3}$$

όπου p είναι το πλήθος των μεταβλητών στο μοντέλο, RSS_p είναι το άθροισμα των τετραγώνων των σφαλμάτων στο μοντέλο που περιέχει p μεταβλητές και $\hat{\sigma}^2$ είναι ένας κατάλληλος εκτιμητής του σ^2. Ο $Mallows$ (1964) έδειξε ότι παλινδρομήσεις με μικρή μεροληψία έχουν C_p που είναι σχεδόν ίσο με p, επομένως αναζητούμε αυτές τις παλινδρομήσεις. Το τελευταίο συμπέρασμα θα αναλυθεί στην παράγραφο 2.2.

Επειδή το πλήθος των υπολογισμών αυξάνεται εκθετικά με το k, υπάρχει ανάγκη να μειωθεί το υπολογιστικό κόστος όταν το k είναι μεγάλο. Προφανώς, από τις $\begin{pmatrix} k \\ p \end{pmatrix}$ παλινδρομήσεις μεγέθους p, μόνο λίγες θεωρούνται καλές και θα είναι αυτές με το ελάχιστο C_p. Παρακάτω περιγράφεται η διαδικασία κατά την οποία μόνο ένα ποσοστό αυτών των υποσυνόλων χρειάζεται να υπολογιστεί ώστε να καθοριστούν οι ζητούμενες παλινδρομήσεις.

Είναι χρήσιμο να αναφερθούμε στην μείωση του αθροίσματος των τετραγώνων λόγω παλινδρόμησης (SSR όπως το συμβολίσαμε στο κεφ.1), όταν απαλείφουμε $r = k - p$ μεταβλητές από το μοντέλο. Το σύνολο των r μεταβλητών για το οποίο αυτή η μείωση γίνεται ελάχιστη καθορίζει τις p μεταβλητές που θα παραμείνουν στο μοντέλο. Το άθροισμα τετραγώνων των σφαλμάτων (RSS) για το μοντέλο που περιέχει αυτές τις p μεταβλητές θα είναι τότε ελάχιστο. Αν συμβολίσουμε με Red_p αυτή τη μείωση, το στατιστικό C_p μπορεί να γραφτεί:

$$C_p = \frac{Red_p}{\hat{\sigma}^2} + (2p - k)$$

Αν η $i-$οστή μεταβλητή απαλειφθεί, η μείωση δίνεται από το γινόμενο $\hat{\sigma}^2 t_i^2$, όπου:

$$t_i^2 = \frac{\hat{\beta}_i^{\,2}}{\hat{\sigma}^2_{\hat{\beta}_i}}$$

είναι το τετράγωνο της $t-$στατιστικής που συνδέεται με το β_i. Αν συμβολίσουμε τη μείωση λόγω απαλοιφής της $i-$οστής μεταβλητής με $\theta_i, i = 1, ..., k$, τότε ισχύει η σχέση:

$$\theta_i = \hat{\sigma}^2 t_i^2$$

Το πρώτο βήμα της διαδικασίας είναι να υπολογίσουμε το πλήρες μοντέλο λύνοντας τις κανονικές εξισώσεις (2.2) και έπειτα τα θ_i. Χωρίς περιορισμό της γενικότητας, υποθέτουμε ότι οι μεταβλητές είναι τέτοιες ώστε να ισχύει:

$$\theta_1 \leq \theta_2 \leq \ldots \leq \theta_k$$

Οπότε, το υποσύνολο μεγέθους $k-1$ που έχει ελάχιστο RSS είναι αυτό που προκύπτει όταν διαγράψουμε την πρώτη μεταβλητή, αφού αυτή δίνει το μικρότερο θ_i.

Πριν συνεχίσουμε, ας αναφέρουμε μια σημαντική ιδιότητα των τετραγωνικών μορφών: "Αν η μείωση του SSR λόγω της απαλοιφής οποιουδήποτε υποσυνόλου μεταβλητών, για το οποίο ο μέγιστος δείκτης είναι i, δεν είναι μεγαλύτερη από θ_{i+1}, τότε κανένα υποσύνολο που περιλαμβάνει μεταβλητές με δείκτες μεγαλύτερους του i δε θα δώσει μικρότερη μείωση". Η μέθοδος απαιτεί το πολύ $p+1$ στάδια για κάθε τιμή του $p = 1, ..., k-2$ και έχει ως εξής:

Στο πρώτο στάδιο, υπολογίζουμε τη μείωση του SSR λόγω της απαλοιφής των μεταβλητών $X_1, X_2, ..., X_r$ για $r = k - p$. Αν αυτή η μείωση δεν ξεπερνάει το θ_{r+1}, τότε η διαδικασία τερματίζεται και το μοντέλο που αποτελείται από τις p μεταβλητές $X_{r+1}, ..., X_k$ είναι το βέλτιστο μοντέλο μεγέθους p.

Αν η μείωση του SSR στο πρώτο στάδιο ξεπερνάει το θ_{r+1}, προχωράμε στο δεύτερο στάδιο όπου συμπεριλαμβάνουμε τη μεταβλητή X_{r+1} μεταξύ των υποψηφίων για απαλοιφή. Υπολογίζονται τότε οι $\begin{pmatrix} r \\ 1 \end{pmatrix}$ μειώσεις λόγω απαλοιφής οποιουδήποτε υποσυνόλου r μεταβλητών που επιλέγεται από τις πρώτες $r+1$ και το οποίο περιλαμβάνει την X_{r+1}. Αν η μικρότερη από τις $1 + \begin{pmatrix} r \\ 1 \end{pmatrix}$ μειώσεις δεν ξεπερνά τη θ_{r+2}, η διαδικασία τερματίζεται και το μοντέλο που αντιστοιχεί στη μικρότερη μείωση είναι το βέλτιστο.

Διαφορετικά, προχωράμε στο τρίτο στάδιο όπου οι μειώσεις υπολογίζονται για όλα τα υποσύνολα μεγέθους r, που επιλέγονται από τις πρώτες $r+2$ μεταβλητές και περιλαμβάνουν τη μεταβλητή X_{r+2}. Αυτές οι μειώσεις είναι συνολικά $\begin{pmatrix} r+1 \\ 2 \end{pmatrix}$. Η ελάχιστη από τις $1 + \begin{pmatrix} r \\ 1 \end{pmatrix} + \begin{pmatrix} r+1 \\ 2 \end{pmatrix}$ μειώσεις των τριών πρώτων σταδίων συγκρίνεται με τη θ_{r+3} και η διαδικασία τερματίζεται ή συνεχίζεται στο επόμενο στάδιο ανάλογα με το αποτέλεσμα της σύγκρισης.

Γενικά, σε κάθε στάδιο, έστω το $q-$οστό, υπολογίζονται συνολικά $\begin{pmatrix} r+q-2 \\ q-1 \end{pmatrix}$ μειώσεις και ελέγχουμε αν το βέλτιστο υποσύνολο έχει εντοπιστεί. Στο $q-$οστό στάδιο, ο

μεγαλύτερος δείκτης σε κάθε μεταβλητή είναι ο $r+q-1$ και άρα η αναζήτηση τερματίζεται αν η ελάχιστη από τις $\sum_{j=1}^{q} \begin{pmatrix} r+j-2 \\ j-1 \end{pmatrix}$ μειώσεις που υπολογίστηκαν στα πρώτα q στάδια δεν ξεπερνά την θ_{r+q} οπότε και το αντίστοιχο μοντέλο είναι το βέλτιστο. Αν όχι, προχωράμε στο $(q+1)-$οστό στάδιο, όπου εξετάζουμε τα υποσύνολα μεγέθους r που περιλαμβάνουν τη μεταβλητή $r+q$.

Έχει παρατηρηθεί ότι σπάνια υπολογίζονται όλες οι $\sum_{j=1}^{p+1} \begin{pmatrix} r+j-2 \\ j-1 \end{pmatrix} = \begin{pmatrix} k \\ p \end{pmatrix}$ παλινδρομήσεις.

2.2. Επιλογή μεταβλητών με τη βοήθεια πολυβηματικής παλινδρόμησης και με κριτήριο επιλογής το C_p

Οι *J.W.Gorman* και *R.J.Toman* (1966) επιχειρούν και αυτοί να μειώσουν το υπολογιστικό κόστος κατά τη διαδικασία αναζήτησης βέλτιστου μοντέλου υπολογίζοντας ένα μέρος μόνο των υποσυνόλων των μεταβλητών και χρησιμοποιώντας ως κριτήριο επιλογής το στατιστικό C_p.

Η εξίσωση (2.1) παίρνει τη μορφή:

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_1 + \hat{\beta}_2 X_2 + \ldots + \hat{\beta}_k X_k \tag{2.4}$$

όπου οι συντελεστές $\hat{\beta}_i, i = 0, 1, ..., k$, έχουν υπολογιστεί από τη μέθοδο των ελαχίστων τετραγώνων. Απαραίτητη είναι η επιλογή των σημαντικών μεταβλητών.

Πριν επιλεχθούν οι μεταβλητές, πρέπει να εξεταστούν τα δεδομένα προσεκτικά για στατιστικές ιδιομορφίες όπως η σειριακή συσχέτιση των σφαλμάτων (*serial correlation*), απόκλιση των σφαλμάτων από την κανονική κατανομή, ακραίες τιμές (*outliers*), καθώς και συναρτησιακές δυσκολίες όπως λάθος επιλογή των X_i στην πλήρη εξίσωση. Οι στατιστικές ατέλειες συχνά εντοπίζονται με προσεκτική εξέταση των σφαλμάτων μετά την προσαρμογή όλων των μεταβλητών στην εξίσωση.

Από τις 2^k εξισώσεις που μπορούμε να υπολογίσουμε, πρέπει να επιλεχθούν ορισμένες εξίσου καλές εξισώσεις για να περιγράψουν τα δεδομένα. Η επιλογή αυτή εξαρτάται από το αν χρειαζόμαστε έναν τύπο για την εκτίμηση των τιμών της εξαρτημένης μεταβλητής ή αν θέλουμε να εκτιμήσουμε την επίδραση κάθε επεξηγηματικής μεταβλητής στην εξαρτημένη μεταβλητή.

Ένας τρόπος για να εξετάσουμε εδώ τις 2^k εξισώσεις είναι η πολυβηματική παλινδρόμηση (*stepwise regression*) που περιγράφηκε στο προηγούμενο κεφάλαιο, αλλά με διαφορετικό κριτήριο τερματισμού. Δύο είναι οι βασικές παραλλαγές: η *forward selection* και η *backward elimination*. Στη *forward selection*, οι επεξηγηματικές μεταβλητές εισέρχονται μία κάθε φορά. Σε κάθε βήμα, η μεταβλητή, έστω η i-οστή, που δίνει τη μεγαλύτερη μείωση στο άθροισμα των τετραγώνων των σφαλμάτων και η ποσότητα $(\hat{\beta}_i/\hat{\sigma}_{\hat{\beta}_i})^2$ ξεπερνά μια προεπιλεγμένη τιμή μπαίνει στην εξίσωση. Κατά τη *backward elimination*, ξεκινάμε με την πλήρη εξίσωση και απαλείφουμε τις λιγότερο σημαντικές μεταβλητές μία κάθε φορά με παρόμοια κριτήρια.

Εφόσον ο υπολογισμός των 2^k εξισώσεων είναι οικονομικά ασύμφορος πρέπει να βρεθεί

Term	Complete Equation				Matrix of Simple Correlation Coefficients, r_{ij}					
	b_i	$\hat{\sigma}_{b_i}$	t_i	R_i^2		X_1	X_2	X_3	X_4	y
Constant	62.40									
X_1	1.55	0.74	2.08	0.97	X_1	1.0				
X_2	0.51	0.72	0.70	0.99	X_2	0.23	1.0			
X_3	0.10	0.75	0.14	0.98	X_3	−0.82	−0.14	1.0		
X_4	−0.14	0.71	0.20	0.99	X_4	−0.24	−0.97	0.03	1.0	
					y	0.73	0.82	−0.53	−0.82	1.0
F-Test	111.4									
RSS	47.9									
R_y^2	0.98									
$\hat{\sigma}^2$	5.98									

Πίνακας 2.1: Στατιστικά Αποτελέσματα [*J.W.Gorman, R.J.Toman* (1966) *Selection of variables for fitting equations to data, Technometrics Vol.*8, *No*1, *p.*36]

ένας τρόπος να απομονώσουμε τις καλύτερες εξισώσεις. Ο *C.Mallows* (1964) πρότεινε μια πολύ χρήσιμη διαδικασία κατά την οποία συγκρίνονται διάφορες εξισώσεις παλινδρόμησης.

Το συνολικό τετραγωνικό σφάλμα (μεροληπτικό και τυχαίο) για n παρατηρήσεις όταν χρησιμοποιούμε μια εξίσωση ισούται με:

$$\sum_{i=1}^{N}(y_i - \hat{y}_i)^2 + \sum_{i=1}^{N} Var(\hat{Y}_{pi})$$

όπου y_i η πραγματική τιμή της εξαρτημένης μεταβλητής,

$\hat{y}_i$ η εκτιμώμενη τιμή της y_i,

$(y_i - \hat{y}_i)$ η μεροληψία στην $i-$οστή παρατήρηση και

N το πλήθος παρατηρήσεων.

Το πρώτο άθροισμα εκφράζει το μεροληπτικό τετραγωνικό σφάλμα ενώ το δεύτερο εκφράζει το τυχαίο τετραγωνικό σφάλμα. Αν θέσουμε $SSB = \sum_{i=1}^{N}(y_i - \hat{y}_i)^2$ και συμβολίσουμε με Γ_p το κανονικοποιημένο συνολικό τετραγωνικό σφάλμα:

$$\Gamma_p = \frac{SSB}{\sigma^2} + \frac{1}{\sigma^2}\sum_{i=1}^{N} Var(\hat{Y}_{pi}) \tag{2.5}$$

Εφόσον ισχύει ότι: $\sum_{i=1}^{n} Var(\hat{Y}_{pi}) = p\sigma^2$, η (2.5) γίνεται:

$$\Gamma_p = \frac{SSB}{\sigma^2} + p \tag{2.6}$$

Στη συνέχεια, το άθροισμα τετραγώνων των σφαλμάτων, RSS_p, από μια εξίσωση με p μεταβλητές έχει αναμενόμενη τιμή:

$$E(RSS_p) = SSB + (N - p)\sigma^2$$

C_p's for All Equations		
Variables in Equation	p	C_p
None	1	443.2
X_1	2	202.7
X_2	2	142.6
X_1X_2	3	2.68
X_3	2	315.3
X_1X_3	3	198.2
X_2X_3	3	62.5
$X_1X_2X_3$	4	3.04
X_4	2	138.8
X_1X_4	3	5.51
X_2X_4	3	138.3
$X_1X_2X_4$	4	3.03
X_3X_4	3	22.4
$X_1X_3X_4$	4	3.49
$X_2X_3X_4$	4	7.34
$X_1X_2X_3X_4$	5	5.0

Πίνακας 2.2: Οι τιμές των C_p για κάθε υποσύνολο μεταβλητών [*J.W.Gorman, R.J.Toman (1966) Selection of variables for fitting equations to data, Technometrics Vol.8, No1, p.36*]

Αλλιώς:

$$SSB = E(RSS_p) - (N-p)\sigma^2$$

Οπότε η (2.6) γίνεται:

$$\Gamma_p = \frac{E(RSS_p)}{\sigma^2} - (N-p) + p = \frac{E(RSS_p)}{\sigma^2} - (N-2p)$$

Αν $\hat{\sigma}^2$ είναι ένας εκτιμητής του σ^2, το στατιστικό C_p, το οποίο δίνεται από τη σχέση:

$$C_p = \frac{RSS_p}{\hat{\sigma^2}} - (N-2p)$$

είναι ένας εκτιμητής του Γ_p.

Παρατηρούμε ότι αν $SSB \simeq 0$, τότε $E(RSS_p) \simeq (N-p)\sigma^2$, δηλαδή το $RSS_p = (N-p)\hat{\sigma_p}^2$, όπου $\hat{\sigma_p}^2$ ένας εκτιμητής της διασποράς σ_p^2 για μια εξίσωση p μεταβλητών, μπορεί να θεωρηθεί εκτιμητής του $(N-p)\sigma^2$, άρα:

$$C_p = \frac{(N-p)\hat{\sigma_p}^2}{\hat{\sigma^2}} - (N-2p)$$

Αλλά αν θεωρήσουμε ότι $\hat{\sigma_p}^2 \simeq \hat{\sigma}^2$, ισχύει:

$$C_p \simeq (N-p) - (N-2p) = p \tag{2.7}$$

Καταλήγουμε λοιπόν στο συμπέρασμα ότι αν τα C_p παρασταθούν γραφικά ως προς p, τότε τα C_p που αντιστοιχούν σε εξισώσεις με μικρή μεροληψία βρίσκονται κοντά στην ευθεία $C_p = p$, ενώ τα C_p των εξισώσεων με μεγάλη μεροληψία βρίσκονται πάνω από τη γραμμή αυτή. Υπάρχει περίπτωση ένα σημείο με μεγαλύτερο C_p από κάποιο άλλο να βρίσκεται πιο κοντά στην ευθεία $C_p = p$ από το άλλο, επειδή το πλήθος των μεταβλητών είναι μεγαλύτερο. Άρα, προσθέτοντας μεταβλητές που μειώνουν τη μεροληψία, μπορεί να αυξήσουν τη συνολική διασπορά. Μερικές φορές, για να πάρουμε μια πιο απλή εξίσωση, απαλείφουμε μερικούς όρους και δεχόμαστε μεγαλύτερη μεροληψία για να έχουμε μικρότερη διασπορά.

Παράδειγμα:

Ας δούμε ένα παράδειγμα με 4 μεταβλητές (Πίνακες 2.1 και 2.2). Τα αποτελέσματα δείχνουν μεγάλες γραμμικές εξαρτήσεις μεταξύ των μεταβλητών. Αυτό φαίνεται από τις τιμές των R_i^2 και οφείλεται στις μεγάλες συσχετίσεις των X_1 και X_3, αλλά και των X_2 και X_4. Το γράφημα των C_p ως προς το p υποδεικνύει 4 καλές εξισώσεις: (1,2,3), (1,2,4), (1,3,4) και (1,2). Δεν είναι γνωστό ποιά από αυτές είναι η βέλτιστη αλλά αν θέλουμε να απλουστεύσουμε την εξίσωση, παίρνουμε την (1,2).

Μπορούμε λοιπόν μέσω του C_p-στατιστικού να βρούμε κατάλληλες εξισώσεις τις οποίες εξετάζουμε αναλυτικά, λαμβάνοντας υπ'οψη το σκοπό για τον οποίο χρησιμοποιούμε την τελική εξίσωση αλλά και το ποιές εξισώσεις είναι λογικές και πρακτικά εύχρηστες.

2.3. Επιλογή μεταβλητών μέσω παλινδρόμησης βασισμένης σε επαναλαμβανόμενους ελέγχους σημαντικότητας και υπολογισμούς της μεροληψίας και του μέσου τετραγωνικού σφάλματος της εκτιμώμενης τιμής της εξαρτημένης μεταβλητής

Οι *W.J.Kennedy* και *T.A.Bancroft* (1971) ασχολούνται με την κατάλληλη επιλογή του υποσυνόλου των μεταβλητών που θα χρησιμοποιηθούν σε ένα μοντέλο πρόβλεψης με βάση διαδοχικούς ελέγχους σημαντικότητας. Πιο συγκεκριμένα, μελετούν δύο διαφορετικές διαδικασίες επιλογής μεταβλητών, που καλούνται "*Forward Selection*" και "*Sequential Deletion*". Υποθέτουμε ότι ο ερευνητής διαθέτει την κατάλληλη γνώση ώστε να διαλέξει r από τις k επεξηγηματικές μεταβλητές τις οποίες θεωρεί ότι πρέπει οπωσδήποτε να συμπεριλάβει στο μοντέλο. Επιπλέον, γνωρίζει τη σειρά σημαντικότητας των υπολοίπων $k - r$ μεταβλητών και τις διατάσσει από την x_{r+1} που θεωρεί την πιο σημαντική μεταβλητή από αυτές έως την x_k που θεωρεί τη λιγότερο σημαντική. Επίσης, διαθέτει n μετρήσεις για τις μεταβλητές $y, x_1, ...x_k$ και θέλει να προβλέψει την τιμή του y για διάφορες τιμές των $x_1, x_2, ..., x_k$ που θα εισάγει στο μέλλον.

Κατά τη διαδικασία *sequential deletion*, ξεκινάμε με το πλήρες μοντέλο και ελέγχουμε την υπόθεση αν ο συντελεστής της x_k είναι 0. Αν δεχτούμε την υπόθεση, απαλείφουμε τη x_k από το μοντέλο και ελέγχουμε αν ο συντελεστής της x_{k-1} είναι 0. Αν δεχτούμε την υπόθεση, απαλείφουμε τη x_{k-1} και ελέγχουμε αν ο συντελεστής της x_{k-2} είναι 0 κ.ο.κ. Συνεχίζουμε μέχρις ότου να απορρίψουμε την υπόθεση ότι ένας συντελεστής είναι 0 ή μέχρι να φτάσουμε στον έλεγχο του συντελεστή της x_r, οπότε διατηρούμε στην εξίσωση

τη x_r και όλες τις υπόλοιπες μεταβλητές που δεν έχουν ακόμα ελεγχθεί.

Κατά τη διαδικασία *forward selection*, ξεκινάμε με το μοντέλο που περιλαμβάνει τις r σημαντικές μεταβλητές και ελέγχουμε την υπόθεση αν ο συντελεστής της x_{r+1} είναι 0. Αν απορρίψουμε την υπόθεση, προσθέτουμε την x_{r+1} στο μοντέλο με τις $x_1, x_2, ..., x_r$ και ελέγχουμε αν ο συντελεστής της x_{r+2} είναι 0. Αν απορρίψουμε την υπόθεση προσθέτουμε τη x_{r+2} στην εξίσωση και ελέγχουμε το συντελεστή της x_{r+3} κ.ο.κ. Συνεχίζουμε μέχρι να δεχτούμε την υπόθεση ότι ένας συντελεστής είναι 0, οπότε δεν προσθέτουμε ούτε τη μεταβλητή που αντιστοιχεί σε αυτόν ούτε τις υπόλοιπες που δεν έχουμε ακόμα ελέγξει.

Προκειμένου να εξετάσουμε τώρα τις επιπτώσεις της χρήσης κάθε διαδικασίας, θα υπολογίσουμε σε κάθε περίπτωση τη μεροληψία αλλά και το μέσο τετραγωνικό σφάλμα της $\hat{y}$, όπου $\hat{y}$ η εκτιμώμενη τιμή της πραγματικής τιμής y της εξαρτημένης μεταβλητής.

2.3.1. Η sequential deletion διαδικασία

Υποθέτουμε ότι το μοντέλο $Y = X\beta + e$ παράγει τα δεδομένα μας, όπου Y είναι το $n \times 1$ διάνυσμα των τιμών της εξαρτημένης μεταβλητής για τις n παρατηρήσεις, X είναι ο $n \times (k+1)$ πίνακας των τιμών των επεξηγηματικών μεταβλητών και e είναι το $n \times 1$ διάνυσμα των σφαλμάτων. Υποθέτουμε επίσης ότι $E(ee') = \sigma^2 I$ και ότι ο πίνακας X είναι κανονικοποιημένος ώστε $X'X = I$. Εδώ, το y_i θα συμβολίζει την τιμή της εξαρτημένης μεταβλητής όταν i επεξηγηματικές μεταβλητές περιέχονται στο μοντέλο. Το A_i θα συμβολίζει το γεγονός ότι απαλείφουμε i μεταβλητές από το μοντέλο.

Το κριτήριο για τον έλεγχο της υπόθεσης $H_0 : \beta_i = 0$ για κάποιο i είναι το b_i^2/υ, όπου b_i ο εκτιμητής του συντελεστή β_i της μεταβλητής x_i, και $\upsilon = RSS$. Απορρίπτουμε την υπόθεση H_0, αν $b_i^2/\upsilon \geq \delta = F_{1,n-k-1,1-a}$, ενώ τη δεχόμαστε διαφορετικά. Τότε ισχύει ότι:

$$E(\hat{y}) = E(y_k|A_0)P(A_0) + E(y_{k-1}|A_1)P(A_1) + ... + E(y_r|A_{k-r})P(A_{k-r}) \tag{2.8}$$

και επειδή τα b_i είναι ανεξάρτητα έχουμε για $i = r+1, ..., k$ ότι:

$$E(y_i|A_{k-i})P(A_{k-i}) = [\beta_0 + \beta_1 X_1 + ... + \beta_{i-1}X_{i-1} + X_i E(b_i|A_{k-i})]P(A_{k-i})$$

και

$$E(y_r|A_{k-r}) = (\beta_0 + \sum_{j=1}^{r} \beta_j X_j)P(A_{k-r})$$

Έστω $\lambda_i = \beta_i^2/2\sigma^2$, $i = r, .., k$, και $F_s(z|\lambda)$ η αθροιστική συνάρτηση κατανομής της χ_s^2 με παράμετρο κεντρικότητας λ και s β.ε. Για $0 \leq i < k - r$ η πιθανότητα $P(A_i)$ είναι:

$$C_{k-i} = P(A_i) = \int_0^\infty [\prod_{j=k-i+1}^{k} F_1(\gamma y|\lambda_j)][1 - F_1(\gamma y|\lambda_{k-1})]g(y)dy$$

όπου $\gamma = \delta/m$, με $m = n - k - 1$, $g(y)$ η σ.π.π. της χ^2_m και $\prod_{j=k+1}^{k} F_1(x|\lambda_j) \equiv 1$

Έστω $r(b_{k-i}, b_{k-i+1}, ..., b_k, \upsilon)$ η κοινή πυκνότητα των τ.μ. $b_{k-i}, b_{k-i+1}, ...b_k$ και υ. Τότε έχουμε:

$$P(A_i) = \int_B ... \int r(b_{k-i}, ..., b_k, \upsilon) d\upsilon \prod_{j=k-i}^{k} db_j \tag{2.9}$$

όπου B είναι η περιοχή που ορίζεται από τα y για τα οποία ισχύουν οι ανισότητες: $b_k^2 < \delta\upsilon, ..., b_{k-i+1}^2 < \delta\upsilon$ και $b_{k-i}^2 \geq \delta\upsilon$.

Αν παραγωγίσουμε το C_{k-i} ως προς β_{k-i}, έχουμε:

$$\frac{\partial C_{k-i}}{\partial \beta_{k-i}} = \frac{\beta_{k-i}}{\sigma^2}[H(A_i) - P(A_i)] \tag{2.10}$$

όπου $H(A_i) = \int_0^\infty [\prod_{j=k-i+1}^{k} F_1(\gamma y|\lambda_j)][1 - F_3(\gamma y|\lambda_{k-i})]g(y)dy$.

Παραγωγίζοντας τη (2.9) ως προς β_{k-i} έχουμε:

$$\frac{\partial P(A_i)}{\partial \beta_{k-i}} = \frac{1}{\sigma^2} E(b_{k-i}|A_i)P(A_i) - \frac{\beta_{k-i}}{\sigma^2} P(A_i) \tag{2.11}$$

Εξισώνοντας τις (2.10) και (2.11), έχουμε:

$$E(b_{k-i}|A_i)P(A_i) = \beta_{k-i} H(A_i), 0 \leq i < k - r$$

Από την (2.8) βρίσκουμε το $E(\hat{y})$. Αν θέσουμε $\sum_{s=0}^{-1} P(A_s) \equiv 0$,τότε η μεροληψία του $\hat{y}$ εκφράζεται από τη σχέση:

$$bias(\hat{y}) = \sum_{t=r+1}^{k} \beta_t X_t [\sum_{s=0}^{k-t-1} P(A_s) + H(A_{k-t}) - 1] \tag{2.12}$$

Για να υπολογίσουμε τώρα το μέσο τετραγωνικό σφάλμα του $\hat{y}$, πρέπει να υπολογίσουμε αρχικά το $E(\hat{y})^2$. Οπότε, έχουμε:

$$E(\hat{y})^2 = E(y_k^2|A_0)P(A_0) + E(y_{k-1}^2|A_1)P(A_1) + ... + E(y_r^2|A_{k-r})P(A_{k-r}) \tag{2.13}$$

Επίσης, για $r < i \leq k$, έχουμε:

$$E(y_i^2|A_{k-i})P(A_{k-i}) = [(\beta_0 + \sum_{j=1}^{i-1} \beta_j x_j)^2 + \sigma^2(1/n + \sum_{j=1}^{i-1} x_j^2)$$

$$+2(\beta_0 + \sum_{j=1}^{i-1} \beta_j x_j) x_i E(b_i|A_{k-i}) + x_i^2 E(b_i^2|A_{k-i})]P(A_{k-i})$$

και

$$E(y_r^2|A_{k-r})P(A_{k-r}) = [(\beta_0 + \sum_{j=1}^{r}\beta_j x_j)^2 + \sigma^2(1/n + \sum_{j=1}^{r} x_j^2)]P(A_{k-r})$$

Η αναμενόμενη τιμή του y_i^2 εξαρτάται από αυτήν του b_i^2. Για να υπολογίσουμε το $E(b_i^2|A_{k-i})P(A_{k-i})$, παραγωγίζουμε δύο φορές την (2.9) ως προς το β_{k-i} και έχουμε:

$$\frac{\partial^2 P(A_i)}{\partial \beta_{k-i}^2} = \frac{1}{\sigma^4}[E(b_{k-i}^2|A_i) - 2\beta_{k-i}E(b_{k-i}|A_i) + \beta_{k-i}^2 - \sigma^2]P(A_i) \tag{2.14}$$

Ομοίως, παραγωγίζοντας δύο φορές το C_{k-i}, έχουμε:

$$\frac{\partial^2 C_{k-i}}{\partial \beta_{k-i}^2} = \frac{\beta_{k-i}^2}{\sigma^4}[P(A_i) - 2H(A_i) + T(A_i)] + \frac{1}{\sigma^2}[H(A_i) - P(A_i)] \tag{2.15}$$

όπου $T(A_i) = \int_0^\infty [\prod_{j=k-i+1}^{k} F_1(\gamma y|\lambda_j)][1 - F_5(\gamma y|\lambda_{k-i})]g(y)dy$.

Εξισώνοντας τις (2.14) και (2.15) έχουμε:

$$E(b_{k-i}^2|A_i)P(A_i) = \beta_{k-i}^2 T(A_i) + \sigma^2 H(A_i)$$

Αν αντικαταστήσουμε στην (2.13) τις σχέσεις που βρήκαμε, βρίσκουμε την έκφραση για την $E(\hat{y})^2$ οπότε τελικά το *μέσο τετραγωνικό σφάλμα* του $\hat{y}$ είναι το:

$$\begin{aligned}
MSE(\hat{y}) &= E(\hat{y})^2 - E^2(\hat{y}) + [bias(\hat{y})]^2 \\
&= \sum_{j=r+1}^{k}[(\sum_{i=r+1}^{j-1}\beta_i x_i)^2 P(A_{k-j}) - 2(\sum_{i=r+1}^{k}\beta_i x_i)\beta_j x_j \sum_{j=0}^{k-j-1} P(A_s) \\
&+ \sigma^2(1/n + \sum_{i=1}^{j-1} x_i^2)P(A_{k-j}) + (\sigma^2 x_j^2 - 2\beta_j x_j \sum_{i=1}^{k}\beta_i x_i)H(A_{k-j}) \\
&+ \beta_j^2 x_j^2 T(A_{k-j})] + \sigma^2(1/n + \sum_{i=1}^{r} x_i^2)P(A_{k-r}) + (\sum_{i=r+1}^{k}\beta_i x_i)^2
\end{aligned}$$

όπου $\sum_{i=k}^{k-1}\beta_i x_i \equiv 0$.

Αν $\delta = 0$, δηλαδή αν απορρίπτουμε πάντα την H_0, τότε ισχύει ότι $MSE(\hat{y}) = \sigma^2(1/n + \sum_{i=1}^{k} x_i^2)$. Όταν $\delta \to \infty$, δηλαδή αν χρησιμοποιούμε πάντα μόνο τις r πρώτες μεταβλητές στο μοντέλο, τότε ισχύει ότι $MSE(\hat{y}) = \sigma^2(1/n + \sum_{i=1}^{k} x_i^2) + (\sum_{i=r+1}^{k}\beta_i x_i)^2$.

2.3.2. Η forward selection διαδικασία

Αν υποθέσουμε τώρα ότι το A_i συμβολίζει το γεγονός να εισέλθουν i μεταβλητές στην εξίσωση, υπολογίζουμε με όμοιο τρόπο τη μεροληψία και το μέσο τετραγωνικό σφάλμα της $\hat{y}$. Προκύπτουν οι τύποι:

$$bias(\hat{y}) = \sum_{i=1}^{k-r} \sum_{j=r+1}^{r+i} \beta_j x_j H_j(A_i) - \sum_{i=r+1}^{k} \beta_i x_i$$

και

$$\begin{aligned} MSE(\hat{y}) &= \sigma^2(1/n + \sum_{i=1}^{r} x_i^2) + \sum_{j=1}^{k-r}\{ \sum_{m+r+1}^{r+j} x_m^2[\beta_m^2 T_m(A_j) + \sigma^2 H_m(A_j)] \\ &+ 2\sum_{i=r+1}^{j} \sum_{t=r+2, i<t}^{j} \beta_i \beta_t x_i x_t S_{it}(A_j)\} - 2(\sum_{i=r+1}^{k} \beta_i x_i) \sum_{i=1}^{k-r} \sum_{j=r+1}^{r+i} \beta_j x_j H_j(A_i) \\ &+ (\sum_{i=r+1}^{k} \beta_i x_i)^2 \end{aligned}$$

όπου

$$S_{ij}(A_t) = \int_0^{\infty} \prod_{s=r+1, s\neq i,j}^{r+t} [1 - F_1(\gamma y|\lambda_s)][1 - F_3(\gamma y|\lambda_i)][1 - F_3(\gamma y|\lambda_j)] F_1(\gamma y|\lambda_{r+t+1}) g(y) dy$$

2.4. Σύγκριση δύο διαδικασιών επιλογής μεταβλητών ως προς τη μεροληψία και το μέσο τετραγωνικό σφάλμα του εκτιμητή της εξαρτημένης μεταβλητής

Οι *H.J.Larson* και *T.A.Bancroft* (1963) ασχολούνται με την επιλογή των μεταβλητών που θα χρησιμοποιηθούν για την κατασκευή ενός γραμμικού μοντέλου. Χρησιμοποιούνται δύο μέθοδοι επιλογής μεταβλητών που βασίζονται στη συσχέτιση μεταξύ των μεταβλητών.

Πολλές φορές στην παλινδρόμηση, αντιμετωπίζουμε την ειδική κατηγορία των "ελλιπώς καθορισμένων μοντέλων" με την έννοια ότι το πλήθος των επεξηγηματικών μεταβλητών που θα συμπεριληφθούν στο τελικό γραμμικό μοντέλο πρέπει να καθοριστούν από κάποιο κανόνα επιλογής βασισμένο στα δεδομένα της έρευνας. Όταν αναφερόμαστε σε "πλήρως καθορισμένα μοντέλα" εννοούμε σχεδιασμένα πειράματα όπου μία μόνο σωστή ανάλυση μπορεί να υπάρξει και οι έλεγχοι σημαντικότητας είναι πλήρως καθορισμένοι προτού τα πειραματικά αποτελέσματα να είναι διαθέσιμα. Στο άρθρο των *H.J.Larson* και *T.A.Bancroft* (1963) εξετάζονται προβλήματα που αφορούν ελλιπώς καθορισμένα μοντέλα και περιλαμβάνουν τη χρήση προκαταρτικών ελέγχων σημαντικότητας.

Πολλοί διαφορετικοί αντικειμενικοί κανόνες και μέθοδοι διαδικασιών έχουν προταθεί για τον καθορισμό, σε τέτοιες περιπτώσεις, του πλήθους των επεξηγηματικών μεταβλητών που θα συμπεριληφθούν στο τελικό γραμμικό μοντέλο. Οι *H.J.Larson* και *T.A.Bancroft* (1963) προτείνουν δύο κανόνες, την *sequential deletion* και την *forward selection* που εξετάστηκαν στο προηγούμενο άρθρο των *W.J.Kennedy* και *T.A.Bancroft* (1971). Εδώ αναφέρονται ως ¨Διαδικασία Α' και ¨Διαδικασία Β' αντιστοίχως. Και οι δύο διαδικασίες προϋποθέτουν ότι ο ερευνητής έχει επιπλέον γνώση της σειράς σημαντικότητας των

επεξηγηματικών μεταβλητών. Αυτό επιτυγχάνεται είτε από θεωρητική εξέταση των μεταβλητών είτε από προηγούμενη εμπειρία με παρόμοια δεδομένα. Αν αυτοί οι τρόποι δεν είναι εφικτοί,τότε ο ερευνητής μπορεί να κάνει μια προκαταρτική μελέτη με ανεξάρτητα δεδομένα ή να χρησιμοποιήσει ένα δείγμα από τα διαθέσιμα δεδομένα για να αποφασίσει τη σειρά σημαντικότητας των επεξηγηματικών μεταβλητών. Εναλλακτικά, μια τέτοια σειρά μπορεί να κατασκευαστεί αν θεωρήσουμε τη μεταβλητή (έστω τη x_1) με τη μεγαλύτερη συσχέτιση με την y ως την πιο σημαντική, τη μεταβλητή (έστω τη x_2) με τη μεγαλύτερη συσχέτιση με τις y και x_1 ως τη δεύτερη πιο σημαντική κ.ο.κ.

Σκοπός της μελέτης αυτής είναι να εξεταστούν οι συνέπειες των δύο διαδικασιών ως προς τη μεροληψία και το μέσο τετραγωνικό σφάλμα του εκτιμητή της y. Η συνάρτηση μεροληψίας και στις δύο περιπτώσεις προκύπτει να είναι τελικά μια γραμμική συνάρτηση των "άμφισβητούμενων" μεταβλητών x_i, $(i = r+1, r+2, ..., k)$. Έχουν κατασκευαστεί ειδικοί πίνακες με τις τιμές των παραμέτρων β_i/σ για άμεσο υπολογισμό των δύο μεροληπτικών συναρτήσεων και των αντιστοίχων μέσων τετραγωνικών σφαλμάτων για κάθε τιμή των x_i, στην περίπτωση όπου δύο είναι οι αμφισβητούμενες μεταβλητές. Για παραπάνω από δύο τέτοιες μεταβλητές, η κατασκευή ειδικών πινάκων είναι περίπλοκη.

Διαδικασία A

Υποθέτουμε ότι το μοντέλο $y = X\beta + e$ παράγει τα δεδομένα μας, όπου y είναι το $n \times 1$ διάνυσμα των παρατηρούμενων τιμών της εξαρτημένης μεταβλητής, X είναι ο $n \times k$ πίνακας των τιμών των επεξηγηματικών μεταβλητών και e είναι το $n \times 1$ διάνυσμα των συνιστωσών του σφάλματος e. Υποθέτουμε επίσης ότι το e ακολουθεί κανονική κατανομή με $E(e) = 0$ και $E(ee') = \sigma^2 I$ και ότι οι επεξηγηματικές μεταβλητές είναι κανονικοποιημένες έτσι ώστε $X'X = I$.

Αν η διασπορά σ^2 είναι γνωστή, το κριτήριο ελέγχου της υπόθεσης $\beta_i = 0$ είναι b_i^2/σ^2, $i = r+1, ..., k$, όπου το b_i είναι ο εκτιμητής του β_i. Η υπόθεση αυτή απορρίπτεται αν $b_i^2/\sigma^2 \geq \lambda$ (το $100\alpha\%$ σημείο της χ_1^2 κατανομής) ενώ διαφορετικά αυτή γίνεται δεκτή. Τότε, η αναμενόμενη τιμή του $\hat{y}$ είναι:

$$E(\hat{y}) = E(y_k|A_0)P(A_0) + E(y_{k-1}|A_1)P(A_1) + \ldots + E(y_r|A_{k-r})P(A_{k-r})$$

Επίσης,

$$E(y_i|A_{k-i})P(A_{k-i}) = [\beta_0 + \beta_1 x_1 + ... + \beta_{i-1}x_{i-1} + x_i E(b_i|A_{k-i})]P(A_{k-i})$$

για $i = r+1, r+2, ..., k$ και

$$E(y_r|A_{k-r})P(A_{k-r}) = [\beta_0 + \beta_1 x_1 + ... + \beta_r x_r]P(A_{k-r})$$

αφού οι εκτιμητές b_i είναι ανά δύο ανεξάρτητοι. Έπειτα, ορίζουμε:

$$F_i = 2\exp\{-\frac{1}{2}(\lambda + \beta_i^2/\sigma^2)\}\sinh(\beta_i\sqrt{\lambda}/\sigma)$$

και

$$H_i = \{\int_{-\infty}^{-\sqrt{\lambda}-\beta_i/\sigma} + \int_{\sqrt{\lambda}-\beta_i/\sigma}^{+\infty}\}(1/\sqrt{2\pi})\exp(-z^2/2)dz$$

με $i = r+1, r+2, ..., k$. Παρατηρούμε ότι τα F_i και H_i είναι συναρτήσεις μόνο των λ και β_i/σ. Από τις υποθέσεις, έχουμε ότι τα b_i ακολουθούν κανονική κατανομή με μέση τιμή β_i και διασπορά σ^2, $i = 1, 2, ..., k$. Επομένως:

$$E(b_k|A_0)P(A_0) = \frac{\sigma}{\sqrt{2\pi}}F_k + \beta_k H_k$$

και

$$E(b_{k-i}|A_i)P(A_i) = \prod_{j=k-i+1}^{k}[1-H_j][\frac{\sigma}{\sqrt{2\pi}}F_{k-i} + \beta_{k-i}H_{k-i}]$$

με $i = 1, 2, ..., k-r-1$.

Δεχόμαστε ότι ισχύει $\sum_{j=0}^{k-r} P(A_j) \equiv 1$ και $\sum_{j=0}^{i} P(A_{k-r-j}) \equiv \prod_{m=r+i+1}^{k}[1-H_m]$, οπότε μπορούμε να γράψουμε ότι:

$$E(\hat{y}) = \beta_0 + \sum_{i=1}^{k}\beta_i x_i + \sum_{i=r+1}^{k} x_i \prod_{j=i+1}^{k}[1-H_j][\frac{\sigma}{\sqrt{2\pi}}F_i - \beta_i(1-H_i)]$$

όπου $\prod_{j=r+1}^{k}(1-H_j) \equiv 1$.

Τελικά, εφόσον ξέρουμε ότι:

$$bias \equiv E(\hat{y}) - [\beta_0 + \sum_{i=1}^{k}\beta_i x_i]$$

θα ισχύουν οι σχέσεις:

$$bias = \sum_{i=r+1}^{k} x_i \prod_{j=i+1}^{k}(1-H_j)\{(\sigma/\sqrt{2\pi})F_i - \beta_i(1-H_i)\}$$

και

$$bias/\sigma = \sum_{i=r+1}^{k} x_i \prod_{j=i+1}^{k}(1-H_j)\{(1/\sqrt{2\pi})F_i - (\beta_i/\sigma)(1-H_i)\} \qquad (2.16)$$

Παρατηρούμε ότι τα F_i και H_i είναι συναρτήσεις μόνο του λ και του β_i/σ. Ο τύπος (2.16) δείχνει ότι η ποσότητα $bias/\sigma$ είναι συνάρτηση μόνο των συντελεστών β_i/σ και της στάθμης σημαντικότητας α των προκαταρτικών ελέγχων των υποθέσεων. Οπότε, με τη χρήση των ειδικών πινάκων, υπολογίζουμε άμεσα τη μεροληψία του $\hat{y}$. Εξάλλου, εφόσον οι εκτιμητές b_i των β_i ακολουθούν κατανομή $N(\beta_i, \sigma^2)$, οι τιμές των $bias/\sigma$ θα κυμαίνονται από -3 έως 3. Για αυτόν το λόγο, οι πίνακες χρησιμοποιούν τιμές στο διάστημα [-3,3].

Στην περίπτωση που η διασπορά σ^2 είναι άγνωστη, η διαδικασία για τον υπολογισμό της μεροληψίας του $\hat{y}$ διαφέρει μόνο στα κριτήρια ελέγχου των υποθέσεων. Για να ελέγξουμε αν $\beta_i = 0$ συγκρίνουμε το b_i^2/υ με το $\lambda = F_{1,n-k-1,\alpha}$, όπου $\upsilon = RSS$. Το μέσο τετραγωνικό σφάλμα $MSE(\hat{y})$, όταν η διασπορά σ^2 είναι γνωστή, υπολογίζεται με παρόμοιο τρόπο και δίνεται από τον εξής τύπο:

$$\begin{aligned}\frac{MSE(\hat{y})}{\sigma^2} &= (\frac{1}{n}+\sum_{i=1}^{k}x_i^2)+2\sum_{i=r+1}^{k}x_i\{\frac{\beta_i}{\sigma}-\frac{F_i}{\sqrt{2\pi}(1-H_i)}\}(\sum_{j=1}^{k}\frac{\beta_j}{\sigma}x_j)\prod_{m=i}^{k}[1-H_m] \\ &+ \sum_{i=r+1}^{k}x_i^2\{(\sqrt{\frac{\lambda}{2\pi}}G_i+\frac{1}{\sqrt{2\pi}}\frac{\beta_i}{\sigma}F_i-(\frac{\beta_i^2}{\sigma^2}+1)(1-H_i)\}\prod_{m=i+1}^{k}(1-H_m)\end{aligned}$$

όπου $G_i = 2\exp\{-\frac{1}{2}(\lambda+\beta_i^2/\sigma^2)\}\cosh(\beta_i\sqrt{\lambda}/\sigma), i = r+1, ..., k$.

2.4.2. Διαδικασία B

Λαμβάνοντας υπ'οψη τους ίδιους συμβολισμούς και τις ίδιες υποθέσεις, υπολογίζουμε τη μεροληψία $bias$ του $\hat{y}$ και το $bias/\sigma$ για τη διαδικασία B. Όταν η διασπορά σ^2 είναι γνωστή, προκύπτει ο εξής τύπος:

$$\frac{bias}{\sigma} = \sum_{i=r+1}^{k}x_i[\frac{1}{\sqrt{2\pi}}F_i\prod_{j=r+1}^{i-1}H_j-\frac{\beta_i}{\sigma}(1-\prod_{j=r+1}^{i}H_j)] \tag{2.17}$$

Παρατηρούμε ότι και σε αυτή την περίπτωση, το $bias/\sigma$ εξαρτάται μόνο από τους λόγους β_i/σ και το α μέσω των συναρτήσεων F_i και H_i. Επομένως, μπορούμε ξανά να χρησιμοποιήσουμε ειδικούς πίνακες με τις διάφορες τιμές των β_i/σ, όταν δύο είναι οι αμφισβητούμενες μεταβλητές. Στην περίπτωση που η διασπορά σ^2 είναι άγνωστη, την εκτιμάμε πάλι χρησιμοποιώντας το υ και κάνοντας παρόμοιες υποθέσεις.

Το μέσο τετραγωνικό σφάλμα $MSE(\hat{y})$ κατά τη διαδικασία B υπολογίζεται με παρόμοιο τρόπο και χρησιμοποιώντας τους ίδιους συμβολισμούς. Προκύπτει έτσι ο παρακάτω τύπος:

$$\begin{aligned}\frac{MSE(\hat{y})}{\sigma^2} &= (\frac{1}{n}+\sum_{i=1}^{r}x_i^2)-(\sum_{i=r+1}^{k}\frac{\beta_i}{\sigma}x_i)^2 \\ &- 2(\sum_{m=r+1}^{k}\frac{\beta_m}{\sigma}x_m)\sum_{i=r+1}^{k}x_i\{\prod_{j=r+1}^{i}H_j[\frac{\beta_i}{\sigma}+\frac{F_i}{\sqrt{2\pi}H_i}]-\frac{\beta_i}{\sigma}\} \\ &+ 2\sum_{i=r+1}^{k-1}\sum_{j=r+2,i<j}^{k}x_ix_j\prod_{m=r+1}^{j}H_m[\frac{\beta_i}{\sigma}+\frac{F_i}{\sqrt{2\pi}H_i}][\frac{\beta_j}{\sigma}+\frac{F_j}{\sqrt{2\pi}H_j}] \\ &+ \sum_{i=r+1}^{k}x_i^2\prod_{j=r+1}^{i-1}H_j[\sqrt{\frac{\lambda}{2\pi}}G_i+\frac{1}{\sqrt{2\pi}}\frac{\beta_i}{\sigma}F_i+(1+\frac{\beta_i^2}{\sigma^2})H_i]\end{aligned}$$

2.4.3. Μη-ορθογώνια περίπτωση

Όλοι οι παραπάνω υπολογισμοί έγιναν με την προϋπόθεση ότι οι επεξηγηματικές μεταβλητές $x_1, x_2, ..., x_k$ είναι ορθογώνιες. Είναι εύκολο να αποδειχθεί ότι οι συναρτήσεις του *bias* και $MSE(\hat{y})$ και στις δύο διαδικασίες είναι ανεξάρτητες της υπόθεσης αυτής. Σε μια δοσμένη μη-ορθογώνια περίπτωση, μια παραμετρικοποίηση του μοντέλου που περιλαμβάνει γραμμικούς συνδυασμούς των $x_1, ..., x_k$ και γραμμικούς συνδυασμούς των $\beta_1, \beta_2, ..., \beta_k$ αρκεί για να μεταβούμε μέσω ενός μετασχηματισμού σε ένα νέο σύνολο μεταβλητών $(x'_1, x'_2, ..., x'_k)$ και ένα νέο σύνολο παραμέτρων $(\beta'_1, \beta'_2, ..., \beta'_k)$ τέτοια ώστε η υπόθεση της ορθογωνιότητας να ικανοποιείται. Τότε, για οποιεσδήποτε τιμές των x_i, έστω $x_{10}, x_{20}, ..., x_{k0}$ ισχύει μέσω του μετασχηματισμού ότι $\sum x_{i0}\beta_i \equiv \sum x'_{i0}\beta'_i$ και $\sum x_{i0}b_i \equiv \sum x'_{i0}b'_i$, $i = 1, 2, ..., k$, όπου x'_{i0}, β'_i και b'_i είναι οι αντίστοιχες μετασχηματισμένες τιμές. Προφανώς, αφού τα *bias* και $MSE(\hat{y})$ εξαρτώνται από τα x_i, β_i και b_i μόνο μέσω τέτοιων αθροισμάτων γινομένων, αυτές οι συναρτήσεις παραμένουν αμετάβλητες μετά από αυτή την παραμετροποίηση.

2.4.4. Σύγκριση των δύο διαδικασιών

Ας κάνουμε τώρα κάποιες παρατηρήσεις και συγκρίσεις πάνω στις συναρτήσεις των $bias/\sigma$ και MSE των δύο διαδικασιών. Συμβολίζουμε με δ_A και δ_B την ποσότητα $bias/\sigma$ για τη διαδικασία Α και τη διαδικασία Β αντίστοιχα. Μέσα από παραδείγματα, μπορεί να διαπιστωθεί ότι όλοι οι συντελεστές και των δύο συναρτήσεων δ_A και δ_B αυξάνονται καθώς το α μειώνεται. Επίσης, αν $|\frac{\beta_i}{\sigma}| \leq 2$, οι συντελεστές των δ_A και δ_B αυξάνονται ενώ αν $|\frac{\beta_i}{\sigma}| > 2$, αυτοί οι συντελεστές μειώνονται.

Παρατηρούμε επίσης τα εξής: όταν η πραγματική τιμή των $\frac{\beta_i}{\sigma}(i = r+1, ..., k)$ είναι ίση με 0, τότε τα δ_A και δ_B είναι και αυτά ίσα με 0, άρα ο εκτιμητής $\hat{y}$ είναι αμερόληπτος και στις δύο περιπτώσεις. Επίσης, αν $\lambda = 0$, δηλαδή αν απορρίπτουμε πάντα την H_0 και άρα χρησιμοποιούμε όλες τις μεταβλητές, τα δ_A και δ_B είναι ξανά ίσα με 0. Αν $\lambda \to \infty$, δηλαδή αν δεχόμαστε πάντα την H_0 και άρα μόνο οι πρώτες r μεταβλητές συμπεριλαμβάνονται στο μοντέλο, τότε τα δ_A και δ_B τείνουν στο άθροισμα $-\sum_{i=r+1}^{k}(\beta_i/\sigma)x_i$. Αν συγκρίνουμε τις δύο διαδικασίες ως προς το $bias/\sigma$, βρίσκουμε ότι γενικά η διαδικασία Α είναι καλύτερη από τη Β.

Αν τα γ_A^2 και γ_B^2 συμβολίζουν το $MSE(\hat{y})$ για τις διαδικασίες Α και Β αντίστοιχα, παρατηρούμε τα εξής: Αν $\lambda = 0$, τα γ_A^2 και γ_B^2 ισούνται και τα δύο με $(\frac{1}{n} + \sum_{i=1}^{k} x_i^2)$. Αν $\lambda \to \infty$, θα είναι ίσα με $(\frac{1}{n} + \sum_{i=1}^{k} x_i^2) + [\sum_{i=r+1}^{k}(\beta_i/\sigma)x_i]^2$. Μέσα από κάποια παραδείγματα, φαίνεται ότι το γ_A^2 είναι ελαφρώς μεγαλύτερο του γ_B^2, αλλά γενικώς η διαφορά μεταξύ αυτών των δύο, μπορεί να θεωρηθεί αμελητέα όταν το $\sum_{i=1}^{r} x_i^2$ είναι αρκετά μεγάλο.

2.5. Επιλογή μεταξύ πλήρους και περιορισμένου μοντέλου και υπολογισμός μεροληψίας και μέσου τετραγωνικού σφάλματος για την αναμενόμενη

τιμή της y

Οι $H.J.Larson$ και $T.A.Bancroft$ (1963) ασχολούνται και αυτή τη φορά με τα ελλιπώς καθορισμένα μοντέλα και την επιλογή των μεταβλητών που θα συμπεριληφθούν στην τελική εξίσωση, αλλά με διαφορετικό τρόπο. Το συγκεκριμένο άρθρο αφορά τις συνέπειες της διάσπασης των επεξηγηματικών μεταβλητών από τον πειραματιστή, σε δύο κατηγορίες: αυτή με τις μεταβλητές (m το πλήθος) για τις οποίες είναι σίγουρος ότι είναι απαραίτητες για το μοντέλο πρόβλεψης και αυτή με τις αμφισβητούμενες μεταβλητές ($k-m$ το πλήθος) τις οποίες θα συμπεριλάβει στο τελικό μοντέλο αν απορρίψει την κοινή υπόθεση ότι όλοι οι αμφισβητούμενοι συντελεστές είναι 0.

Πιο αναλυτικά, υποθέτουμε ότι το πραγματικό μοντέλο είναι το εξής:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + e \tag{2.18}$$

όπου οι συνιστώσες του σφάλματος e είναι ανεξάρτητες και κανονικά κατανεμημένες με μέση τιμή 0 και άγνωστη διασπορά σ^2. Οι παρατηρήσεις είναι n το πλήθος και τα x_i είναι μετασχηματισμένα έτσι ώστε να είναι ορθογώνια μεταξύ τους με μέση τιμή 0 και αθροίσμα τετραγώνων ίσο με 1.

2.5.1. Κανόνας επιλογής μοντέλου

Ο κανόνας της διαδικασίας που χρησιμοποιείται είναι ο εξής: Παίρνουμε ένα δείγμα n παρατηρήσεων πάνω στις $x_1, x_2, ..., x_k$ και y και χρησιμοποιούμε την κλασσική μέθοδο των ελαχίστων τετραγώνων για να υπολογίσουμε τα $b_0, b_1, ..., b_k$, δηλαδή τους εκτιμητές των $\beta_0, \beta_1, ..., \beta_k$ αντίστοιχα. Το άθροισμα τετραγώνων λόγω της παλινδρόμησης στις $x_1, x_2, ..., x_k$ χωρίζεται σε δύο μέρη: το πρώτο είναι λόγω της παλινδρόμησης στα $x_1, x_2, ..., x_m$ (απαραίτητες) και το δεύτερο λόγω της παλινδρόμησης στις $x_{m+1}, x_{m+2}, ..., x_k$ (αμφισβητούμενες). Λόγω των υποθέσεων για τα x_i, τα αθροίσματα τετραγώνων είναι ίσα με $b_1^2 + b_2^2 + ... + b_m^2$ και $b_{m+1}^2 + b_{m+2}^2 + ... + b_k^2$ αντίστοιχα. Έστω ότι το υ συμβολίζει το δειγματικό μέσο τετραγωνικό σφάλμα, δηλαδή το άθροισμα τετραγώνων των σφαλμάτων RSS διαιρούμενο με τους αντίστοιχους β.ε. που είναι ίσοι με $k - m$.

Ελέγχουμε τώρα την υπόθεση $H_0 : \beta_{m+1} = \beta_{m+2} = ... = \beta_k = 0$, χρησιμοποιώντας ως κριτήριο ελέγχου το:

$$F_0 = \frac{b_{m+1}^2 + b_{m+2}^2 + \ldots + b_k^2}{(k-m)\upsilon} \tag{2.19}$$

το οποίο αποτελεί μια μεταβλητή που ακολουθεί κατανομή F με $k-m$ και $n-k-1$ β.ε. Αν $F_0 \geq \lambda$ (όπου $\lambda = F_{k-m,n-k-1,\alpha}$), απορρίπτουμε την H_0 και χρησιμοποιούμε στο μοντέλο όλες τις μεταβλητές. Αν $F_0 < \lambda$, δεχόμαστε την H_0 και χρησιμοποιούμε μόνο τις πρώτες m μεταβλητές στο μοντέλο. Άρα, η εκτιμώμενη τιμή της y, η $\hat{y}$, θα συμβολίζεται είτε με y_k (ο δείκτης δηλώνει το πλήθος των επεξηγηματικών μεταβλητών στο μοντέλο) είτε με y_m, ανάλογα με το αποτέλεσμα του ελέγχου της H_0. Το πρόβλημα που πρέπει να λυθεί είναι να καθοριστεί η αναμενόμενη τιμή και η μεροληψία του $\hat{y}$ και το μέσο τετραγωνικό σφάλμα της $\hat{y}$.

2.5.2. Υπολογισμός μεροληψίας

Υπολογίζουμε αρχικά το $E(\hat{y})$. Έστω ότι το γεγονός A_0 δηλώνει αποδοχή της H_0, ενώ το A_1 δηλώνει απόρριψη της H_0. Τότε, ισχύει:

$$E(\hat{y}) = E(y_m|A_0)P(A_0) + E(y_k|A_1)P(A_1) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + \sum_{i=m+1}^{k} x_i E(b_i|A_1)P(A_1) \tag{2.20}$$

Οπότε, αρκεί να υπολογίσουμε τα $E(b_i|A_1)P(A_1)$, για $i = m+1, ..., k$.

Τα b_i, για $i = m+1, ..., k$, ακολουθούν κανονική κατανομή με μέση τιμή β_i και διασπορά σ^2, ενώ το υ, που είναι ο εκτιμητής του σ^2, ακολουθεί την κατανομή της $\sigma^2\chi^2/(n-k-1)$ με $n-k-1$ β.ε. Επιπλέον, τα $b_{m+1}, ..., b_k, \upsilon$ είναι λόγω των υποθέσεων ανεξάρτητα κατανεμημένα με κοινή συνάρτηση πυκνότητας πιθανότητας ίση με:

$$f(b_{m+1}, ..., b_k) = K \exp\{-\sum_{i=m+1}^{k} \frac{(b_i - \beta_i)^2}{2\sigma^2} - \frac{(n-k-1)\upsilon}{2\sigma^2}\} \upsilon^{\frac{n-k-3}{2}} \tag{2.21}$$

όπου $K = \frac{1}{\Gamma(\frac{n-k-1}{2})} (\frac{n-k-1}{2\sigma^2})^{\frac{n-k-1}{2}} (\frac{1}{2\pi\sigma^2})^{\frac{k-m}{2}}$. Η περιοχή A_1 ορίζεται από τη σχέση $F_0 \geq \lambda$, ή αλλιώς από τη σχέση:

$$\upsilon \leq \frac{b_{m+1}^2 + b_{m+2}^2 + \ldots + b_k^2}{(k-m)\lambda}$$

Τότε, η:

$$P(A_1) = \int\int \ldots \int_{A_1} f(b_{m+1}, ..., b_k, \upsilon) \prod_{i=m+1}^{k} db_i d\upsilon = g(\theta) \tag{2.22}$$

είναι η πιθανότητα το $(k-m)\lambda/(n-k-1)$ να είναι μικρότερο από $(k-m)F^*/(n-k-1)$ όπου F^* είναι μια τυχαία μεταβλητή που ακολουθεί μη-κεντρική κατανομή F με $k-m$ και $n-k-1$ β.ε. και παράμετρο κεντρικότητας

$$\theta = \frac{1}{2} \sum_{i=m+1}^{k} \beta_i^2/\sigma^2 \tag{2.23}$$

Από (2.21) και (2.22) προκύπτει ότι:

$$\begin{aligned} \frac{\partial P(A_1)}{\partial \beta_i} &= \int\int \ldots \int_{A_1} \frac{b_i - \beta_i}{\sigma^2} f(b_{m+1}, ..., b_k, \upsilon) \prod_{i=m+1}^{k} db_i d\upsilon \\ &= \frac{E(b_i - \beta_i|A_i)P(A_1)}{\sigma^2} \end{aligned} \tag{2.24}$$

με $i = m+1, ..., k$.

Από (2.22) και (2.23), έχουμε ότι:

$$\frac{\partial P(A_1)}{\partial \beta_i} = g'(\theta)\frac{\partial \theta}{\partial \beta_i} = g'(\theta)\frac{\beta_i}{\sigma^2} \tag{2.25}$$

με $i = m+1, ..., k$.

Εξισώνοντας τις (2.24) και (2.25) έχουμε:

$$E(b_i|A_i)P(A_i) = \beta_i[g(\theta) + g'(\theta)] \tag{2.26}$$

με ι=μ+1,...,κ.

Οπότε από (2.20) και (2.26), έχω ότι:

$$E(\hat{y}) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + [g(\theta) + g'(\theta)] \sum_{i=m+1}^{k} \beta_i x_i \tag{2.27}$$

Αν τώρα πάρουμε τη μερική παράγωγο της $P(A_1)$ ως προς το θ αυτή τη φορά, έχουμε ότι:

$$g'(\theta) = -g(\theta) + P[\frac{k-m+2}{n-k-1} F^*_{k-m+2,n-k-1}(\theta) > \frac{k-m}{n-k-1}\lambda]$$

Αν θέσουμε ίση την τελευταία πιθανότητα με $h(\theta)$, έχουμε ότι $h(\theta) = g'(\theta) + g(\theta)$. Άρα, η (2.27) γίνεται:

$$E(\hat{y}) = \beta_0 + \sum_{i=1}^{m} \beta_i x_i + h(\theta) \sum_{i=m+1}^{k} \beta_i x_i \tag{2.28}$$

Οπότε, το *bias* του $\hat{y}$ είναι:

$$bias = \beta_0 + \sum_{i=1}^{k} \beta_i x_i - E(\hat{y}) = [1 - h(\theta)] \sum_{i=m+1}^{k} \beta_i x_i$$

και

$$\delta = \frac{bias}{\sigma} = [1 - h(\theta)] \sum_{i=m+1}^{k} \frac{\beta_i}{\sigma} x_i \tag{2.29}$$

Παρατηρούμε ότι αν $\lambda = 0$, δηλαδή αν χρησιμοποιούμε πάντα το πλήρες μοντέλο, τότε $\delta = 0$. Αν τώρα $\lambda \to \infty$, δηλαδή αν χρησιμοποιούμε το περιορισμένο μοντέλο, τότε $\delta \to \sum_{i=m+1}^{k} \beta_i x_i/\sigma$. Επίσης, αν κάποια από τα $\beta_{m+1}, ..., \beta_k$ γίνονται πολύ μεγάλα, σημαίνει ότι απορρίπτουμε πάντα την υπόθεση ότι είναι ίσα με 0 άρα και χρησιμοποιούμε το πλήρες μοντέλο, οπότε ισχύει ότι $\delta \to 0$.

2.5.3. Υπολογισμός μέσου τετραγωνικού σφάλματος

Για να υπολογίσουμε το $MSE(\hat{y})$ χρειαζόμαστε το $E(\hat{y})^2$. Ισχύει τότε:

$$\begin{aligned} E(\hat{y})^2 &= E(y_m^2|A_0)P(A_0) + E(y_k^2|A_1)P(A_1) \\ &= \sigma^2(\frac{1}{n} + \sum_{i=1}^{m} x_i^2) + (\beta_0 + \sum_{i=1}^{m} \beta_i x_i)^2 + 2(\beta_0 + \sum_{i=1}^{m} \beta_i x_i) \sum_{i=m+1}^{k} x_i E(b_i|A_1)P(A_1) \\ &+ 2 \sum_{i=m+2}^{k} \sum_{j=m+1, i>j}^{k-1} x_i x_j E(\beta_i \beta_j|A_1)P(A_1) + \sum_{i=m+1}^{k} x_i^2 E(b_i^2|A_1)P(A_1) \end{aligned} \tag{2.30}$$

Παραγωγίζοντας την $P(A_1)$ δύο φορές ως προς β_i έχουμε:

$$\begin{aligned} \frac{\partial^2 P(A_1)}{\partial \beta_i^2} &= \int\int \dots \int_{A_1} [-\frac{1}{\sigma^2} + \frac{(b_i - \beta_i)^2}{\sigma^4}] f(b_{m+1}, ..., b_k, \upsilon) \prod_{i=m+1}^{k} db_i d\upsilon \\ &= [-\frac{1}{\sigma^2} + \frac{E[b_i^2 - 2\beta_i b_i + \beta_i^2|A_1]}{\sigma^4}] P(A_1) \end{aligned} \tag{2.31}$$

Αν παραγωγίσουμε τώρα τη $g(\theta)$ δύο φορές ως προς β_i έχουμε:

$$\frac{\partial^2 g(\theta)}{\partial \beta_i^2} = g''(\theta)\frac{\beta_i^2}{\sigma^4} + \frac{g'(\theta)}{\sigma^2} \tag{2.32}$$

Εξισώνοντας τις (2.31) και (2.32) έχουμε:

$$E(b_i^2|A_1)P(A_1) = \sigma^2 h(\theta) + \beta_i^2[g''(\theta) + 2h(\theta) - g(\theta)] \tag{2.33}$$

με $i = m+1, ..., k$.

Επίσης ισχύει ότι:

$$\begin{aligned} \frac{\partial^2 P(A_1)}{\partial \beta_i \partial \beta_j} &= \int\int \dots \int_{A_1} [\frac{(b_i - \beta_i)(b_j - \beta_j)}{\sigma^4}] f(b_{m+1}, ..., b_k, \upsilon) \prod_{i=m+1}^{k} db_i d\upsilon \\ &= \frac{1}{\sigma^4} E[b_i b_j - b_i \beta_j - b_j \beta_i + \beta_i \beta_j|A_1] P(A_1) \end{aligned}$$

και

$$\frac{\partial^2 g(\theta)}{\partial \beta_i \partial \beta_j} = g''(\theta)\frac{\beta_i \beta_j}{\sigma^4}, i \neq j$$

Οι δύο τελευταίες σχέσεις, λοιπόν, μας δίνουν:

$$E(b_i b_j|A_1)P(A_1) = \beta_i \beta_j[g''(\theta) + 2h(\theta) - g(\theta)] \tag{2.34}$$

$i, j = m+1, ..., k, i \neq j$.

Συνδυάζοντας τις (2.26), (2.30), (2.33) και (2.34) λοιπόν έχουμε:

$$\begin{aligned} E(\hat{y})^2 &= \sigma^2(\frac{1}{n} + \sum_{i=1}^{m} x_i^2 + h(\theta) \sum_{i=m+1}^{k} x_i^2) + (\beta_0 + \sum_{i=1}^{m} \beta_i x_i)^2 \\ &+ 2(\beta_0 + \sum_{i=1}^{m} \beta_i x_i) \sum_{i=m+1}^{k} \beta_i x_i h(\theta) + [g''(\theta) + 2h(\theta) - g(\theta)](\sum_{i=m+1}^{k} \beta_i x_i)^2 \end{aligned}$$

Παίρνοντας τη δεύτερη παράγωγο του $g(\theta)$ ως προς θ, ισχύει ότι:

$$g''(\theta) = g(\theta) - 2h(\theta) + P[\frac{k-m+4}{n-k-1} F^*_{k-m+4,n-k-1}(\theta) > \frac{k-m}{n-k-1}\lambda]$$

Θέτοντας με $r(\theta)$ την τελευταία πιθανότητα, προκύπτει τελικά ότι:

$$\gamma^2 = \frac{MSE(\hat{y})}{\sigma^2} = \frac{1}{n} + \sum_{i=1}^{m} x_i^2 + h(\theta) \sum_{i=m+1}^{k} x_i^2 + [r(\theta) - 2h(\theta) + 1](\sum_{i=m+1}^{k} \frac{\beta_i}{\sigma} x_i)^2 \quad (2.35)$$

Μπορούμε να κάνουμε παρόμοιες παρατηρήσεις και για το γ^2. Επίσης, παρατηρούμε ότι τα δ και γ^2 είναι σχετικά απλές συναρτήσεις των β_i/σ και λ. Υπάρχουν και για αυτά ειδικοί πίνακες για τον υπολογισμό τους.

2.6. Επιλογή κατάλληλου μοντέλου μέσω της χρήσης προκαταρτικών ελέγχων σημαντικότητας των συντελεστών παλινδρόμησης

Σε εφαρμογές στατιστικής θεωρίας, υπάρχει συχνά αβεβαιότητα σχετικά με το αν προσδιορίσαμε κατάλληλα κάποιους παράγοντες, όπως την κατανομή που ακολουθούν τα δεδομένα, τη διασπορά του μοντέλου παλινδρόμησης, τους συντελεστές των μεταβλητών κλπ. Σε τέτοιες περιπτώσεις, είναι απαραίτητη η χρήση προκαταρτικών ελέγχων σημαντικότητας για να εξετάσουμε την ορθότητα των εκτιμήσεών μας.

Ο *T.A.Bancroft* (1944) εξετάζει τον *έλεγχο ενός συντελεστή μιας παλινδρόμησης*. Μετά από μία παλινδρόμηση μπορεί να είμαστε αβέβαιοι για το αν είναι κατάλληλο να διατηρήσουμε στο μοντέλο μία συγκεκριμένη επεξηγηματική μεταβλητή. Έστω ότι έχουμε προσαρμόσει τα δεδομένα μας στο μοντέλο $y = b_1 x_1 + b_2 x_2$ και θέλουμε να επιλέξουμε ανάμεσα σε αυτό και στο περιορισμένο μοντέλο $y = b'_1 x_1$. Έστω ότι το πραγματικό μοντέλο εκφράζεται από τη σχέση $y = \beta_1 x_1 + \beta_2 x_2$. Σε αυτή την περίπτωση ένας κανόνας για να αποφασίσουμε να κρατήσουμε ή όχι τη μεταβλητή x_2 στο μοντέλο είναι να ελέγξουμε το λόγο $F = \theta^2/\gamma^2$, όπου θ^2 είναι η μείωση του αθροίσματος τετραγώνων λόγω της απαλοιφής της μεταβλητής x_2 από το μοντέλο $y = b_1 x_1 + b_2 x_2$ και γ^2 είναι το μέσο τετράγωνο των σφαλμάτων. Αν το F είναι μικρότερο από μια προκαθορισμένη τιμή, τότε απαλείφουμε τη x_2 και χρησιμοποιούμε το b'_1 ως εκτιμητή του β_1. Αν το F είναι μεγαλύτερο από αυτή την τιμή, τότε διατηρούμε στην εξίσωση τη x_2 και χρησιμοποιούμε το b_1 ως εκτιμητή του β_1.

Αναλυτικότερα, έχουμε το μοντέλο:

$$y = \beta_1 x_1 + \beta_2 x_2 + e \tag{2.36}$$

Υποθέτουμε ότι οι μεταβλητές x_1 και x_2 έχουν διασπορές ίσες με 1 και συντελεστή συσχέτισης ρ, οπότε ισχύουν:

$$S(x_1^2) = n-1, S(x_2^2) = n-1, S(x_1 x_2) = \rho(n-1)$$

όπου n είναι το πλήθος των παρατηρήσεων και $S(x_1^2)$ είναι το άθροισμα των παρατηρούμενων τιμών x_1^2 στο δείγμα, με παρόμοιες έννοιες για τα $S(x_2^2)$ και $S(x_1 x_2)$.

Στη συνέχεια, εκτελούμε τον εξής ορθογώνιο μετασχηματισμό:

$$\begin{cases} \xi_1 = x_1 \\ \\ \xi_2 = x_2 - \rho x_1 \end{cases}$$

Τότε, η (2.36) γίνεται:

$$y = \beta_1 \xi_1 + \beta_2(\xi_2 + \rho \xi_1) + e$$

Προκύπτει, λοιπόν, ότι:

$$S(\xi_1^2) = n-1, S(\xi_2^2) = (n-1)(1-\rho^2), S(\xi_1 \xi_2) = 0$$

Επομένως,

$$S(y\xi_1) = \beta_1(n-1) + \beta_2 \rho(n-1) + S(x_1 e)$$

και

$$S(y\xi_2) = \beta_2(n-1)(1-\rho^2) + S[(x_2 - \rho x_1)e]$$

Αν συμβολίσουμε με B_1 τον εκτιμητή του συντελεστή του ξ_1 και με B_2 τον εκτιμητή του συντελεστή του ξ_2, έχουμε ότι:

$$B_1 S(\xi_1^2) = S(\xi_1 y), B_2 S(\xi_2^2) = S(\xi_2 y)$$

Η μείωση τότε στο συνολικό άθροισμα τετραγώνων λόγω της παλινδρόμησης στη x_1 όταν αγνοήσουμε τη x_2 είναι:

$$B_1 S(y\xi_1) = \frac{[S(\xi_1 y)]^2}{S(\xi_1^2)} = \frac{[(\beta_1 + \beta_2 \rho)(n-1) + S(x_1 e)]^2}{n-1}$$

ενώ η μείωση στο συνολικό άθροισμα τετραγώνων λόγω της παλινδρόμησης στη x_2 όταν στο μοντέλο είναι ήδη η x_1 είναι:

$$B_2 S(y\xi_2) = \frac{[S(\xi_2 y)]^2}{S(\xi_2^2)} = \frac{[\beta_2(n-1)(1-\rho^2) + S(x_2 - \rho x_1)e]^2}{(n-1)(1-\rho^2)}$$

Επομένως, η μείωση στο συνολικό άθροισμα τετραγώνων λόγω της παλινδρόμησης στις x_1 και x_2 είναι το άθροισμα των δύο παραπάνω ποσοτήτων, οι οποίες είναι ανεξάρτητα κατανεμημένες.

Έστω ότι b'_1 είναι ο εκτιμώμενος συντελεστής της x_1 στο μοντέλο παλινδρόμησης όπου η x_2 έχει απαλειφθεί. Οπότε, ισχύει:

$$b'_1 = B_1 = \frac{S(\xi_1 y)}{S(\xi_1^2)} = \frac{(\beta_1 + \beta_2\rho)(n-1) + S(x_1 e)}{n-1} \tag{2.37}$$

Επομένως:

$$E(b'_1) = \beta_1 + \beta_2\rho \tag{2.38}$$

αφού $E[S(x_1 e)] = 0$.

Έστω ότι b_2 είναι ο εκτιμώμενος συντελεστής του x_2 όταν στο μοντέλο παλινδρόμησης συμπεριλαμβάνονται οι μεταβλητές x_1 και x_2. Τότε:

$$b_2 = B_2 = \frac{S(\xi_2 y)}{S(\xi_2^2)} = \frac{(n-1)\beta_2(1-\rho^2) + S[(x_2 - \rho x_1)e]}{(n-1)(1-\rho^2)}$$

Η διασπορά του b_2 είναι:

$$Var(b_2) = \frac{S(\xi_2^2)}{[S(\xi_2^2)]^2} = \frac{1}{S(\xi_2^2)} = \frac{1}{(n-1)(1-\rho^2)}$$

Οι κανονικές εξισώσεις για το μοντέλο $y = b_1 x_1 + b_2 x_2$ γράφονται ως εξής:

$$\begin{cases} b_1 S(x_1^2) + b_2 S(x_1 x_2) = S(x_1 y) \\ \\ b_1 S(x_1 x_2) + b_2 S(x_2^2) = S(x_2 y) \end{cases}$$

Επίσης, για τον b'_1 ισχύει ότι:

$$b'_1 = \frac{S(x_1 y)}{S(x_1^2)} = b_1 + b_2 \frac{S(x_1 x_2)}{S(x_1^2)}$$

Επομένως,

$$b'_1 = b_1 + b_2\rho$$

ή

$$b_1 = b'_1 - b_2\rho \tag{2.39}$$

Επομένως, από τις (2.38) και (2.39) προκύπτει ότι:

$$E(b_1) = \beta_1 + \beta_2\rho - \rho E(b_2)$$

όπου αν $\rho = 0$, παρατηρούμε ότι το b_1 είναι αμερόληπτος εκτιμητής του β_1.

Για να επιλέξουμε τώρα ανάμεσα στα μοντέλα $y = b_1x_1+b_2x_2$ και $y' = b'_1x_1$, αρκεί να υπολογίσουμε την τιμή του λόγου $F = \theta^2/\gamma^2$, όπου $\theta^2 = B_2S(y\xi_2) = \frac{[\beta_2(n-1)(1-\rho^2)+S(x_2-\rho x_1)e]^2}{(n-1)(1-\rho^2)}$ και $\gamma^2 = S(y-Y)^2 = B_1S(y\xi_1)+B_2S(y\xi_2) = \frac{[(\beta_1+\beta_2\rho)(n-1)+S(x_1e)]^2}{n-1} + \frac{[\beta_2(n-1)(1-\rho^2)+S(x_2-\rho x_1)e]^2}{(n-1)(1-\rho^2)}$. Έπειτα, συγκρίνουμε το F με μια προκαθορισμένη τιμή, έστω το $\lambda = F_{1,n-3}$. Αν το $F < \lambda$, τότε παραλείπουμε τη x_2 και χρησιμοποιούμε το b'_1 ως εκτιμητή του β_1. Αν το $F > \lambda$, τότε διατηρούμε τη x_2 στην εξίσωση και χρησιμοποιούμε το b_1 ως εκτιμητή του β_1. Ο εκτιμητής του β_1 που προκύπτει τελικά καλείται b^*.

Θα υπολογίσουμε τώρα τη μεροληψία του b^*. Στην περίπτωση όπου $F < \lambda$, δηλαδή στο μοντέλο υπάρχει μόνο η x_1, θα ισχύει η σχέση (2.38). Αν $F \geq \lambda$, δηλαδή στο μοντέλο περιέχονται και οι δύο μεταβλητές, εκτελούμε τους απαραίτητους υπολογισμούς και καταλήγουμε στον τύπο:

$$E(b_2) = \frac{\beta_2}{P(u \leq \frac{1}{\lambda c})} \sum_{i=0}^{\infty} \frac{\alpha^i e^{-\alpha}}{i!} I_{x_0}(\frac{n-3}{2}, \frac{3}{2}+i),$$

όπου $u = \frac{\gamma^2}{b_2^2}$, $c = \frac{1}{(n-1)(1-\rho^2)}$, $\alpha = \frac{\beta_2^2}{2c}$ και $x_0 = \frac{1}{\frac{\lambda}{n-3}+1}$.

Αποδείξαμε ότι ισχύει:

$$E(b_1) = \beta_1 + \beta_2\rho - \rho E(b_2)$$

Επομένως, βρίσκουμε τη μεροληψία του b^* που είναι:

$$bias = \rho\beta_2[1 - \sum_{i=0}^{\infty} \frac{\alpha^i e^{-\alpha}}{i!} I_{x_0}(\frac{n-3}{2}, \frac{3}{2}+i)]$$

Παρατηρούμε ότι αν $\lambda = 0$, τότε $E(b_2) = \beta_2$ και $bias = 0$, ενώ όταν $\lambda \to \infty$, τότε $bias \to \rho\beta_2$. Επίσης, κάνουμε τις εξής παρατηρήσεις:

- Δεν υπάρχει μεροληψία στην εκτίμηση του β_1 αν $\rho = 0$ ή $\beta_2 = 0$
- Ο συντελεστής του β_2 στον τύπο είναι απολύτως μικρότερος ή ίσος του 1
- Το πρόσημο του $bias$ εξαρτάται από τα πρόσημα των ρ και β_2. Είναι θετικό αν $\rho\beta_2 > 0$, ενώ είναι αρνητικό όταν $\rho\beta_2 < 0$
- Η μεροληψία στην εκτίμηση του β_1 είναι ανεξάρτητη του β_1.

Το πρόβλημα που αναπτύχθηκε αποτελεί μια ειδική περίπτωση του γενικότερου προβλήματος της χρήσης του ελέγχου σημαντικότητας ως κριτήριο επιλογής του πλήθους των επεξηγηματικών μεταβλητών που θα χρησιμοποιηθούν στο μοντέλο:

$$y = b_1x_1 + b_2x_2 + \ldots + b_kx_k.$$

Κεφάλαιο 3

ΣΥΡΡΙΚΝΩΣΗ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΚΑΙ Η ΜΕΘΟΔΟΣ *LASSO*

Στο κεφάλαιο αυτό παρουσιάζεται η μέθοδος *Lasso*, η οποία χρησιμοποιείται για την εκτίμηση των συντελεστών των επεξηγηματικών μεταβλητών σε γραμμικά μοντέλα, σύμφωνα με το σχετικό άρθρο του *R.Tibshirani* (1996). Πρόκειται για το πρόβλημα της ελαχιστοποίησης των τετραγώνων των σφαλμάτων αλλά υπό τον περιορισμό το άθροισμα των απολύτων τιμών των εκτιμώμενων συντελεστών των επεξηγηματικών μεταβλητών να μην ξεπερνάει την τιμή μιας ρυθμιζόμενης παραμέτρου. Αυτό καθιστά κάποιους από τους συντελεστές ίσους με το 0, επομένως το μοντέλο γίνεται ευκολότερα ερμηνεύσιμο και η διασπορά του μειώνεται. Στη συνέχεια, περιγράφεται το δυϊκό πρόβλημα σύμφωνα με το σχετικό άρθρο των *M.R.Osborne*, *B.Presnell*, και *B.A.Turlach* (2000) και συγκρίνονται οι λύσεις των δύο αυτών προβλημάτων.

Θεωρούμε τα δεδομένα $(x^i, y_i), i = 1, 2, ..., N$, όπου το $x^i = (x_{i1}, ..., x_{ip})'$ είναι το διάνυσμα των τιμών των επεξηγηματικών μεταβλητών $X_1, X_2, ..., X_p$ για την i-οστή παρατήρηση ενώ το y_i συμβολίζει την τιμή της εξαρτημένης μεταβλητής Y για την i-οστή παρατήρηση. Οι εκτιμητές ελαχίστων τετραγώνων (*Ordinary Least Squares* ή OLS) λαμβάνονται ελαχιστοποιώντας το άθροισμα τετραγώνων των σφαλμάτων (RSS). Όμως, ο πειραματιστής είναι συχνά ανικανοποίητος από τη χρήση των OLS εκτιμητών. Δύο είναι οι βασικοί λόγοι: Ο πρώτος αφορά την *ακρίβεια της πρόβλεψης*, αφού οι OLS εκτιμητές δίνουν συχνά μικρή μεροληψία για την εκτιμώμενη τιμή της Y αλλά μεγάλη διασπορά για τους συντελεστές παλινδρόμησης. Η ακρίβεια της πρόβλεψης μπορεί να βελτιωθεί συρρικνώνοντας κάποιους συντελεστές ή θέτοντάς τους ίσους με 0. Με αυτόν τον τρόπο, λαμβάνουμε περισσότερη μεροληψία αλλά μειώνουμε τη διασπορά και επομένως βελτιώνουμε τη συνολική ακρίβεια. Ο δεύτερος λόγος αφορά την *ερμηνεία* του τελικού μοντέλου. Έχοντας μεγάλο αριθμό επεξηγηματικών μεταβλητών, θέλουμε συχνά να καθορίσουμε ένα υποσύνολο αυτών, ώστε να μπορούμε να ερμηνεύσουμε ευκολότερα το μοντέλο.

Οι δύο βασικές τεχνικές βελτίωσης των OLS εκτιμητών, δηλαδή η *Subset selection* και η *ridge regression*, έχουν και οι δύο μειονεκτήματα. Η *Subset selection* παρέχει μοντέλα που μπορούν να ερμηνευτούν αλλά ενδέχεται να είναι εξαιρετικά μεταβαλλόμενα αφού κατά την εφαρμογή της μεθόδου διατηρούμε ή απαλείφουμε μεταβλητές από το μοντέλο, κατά συνέπεια μικρές αλλαγές στα δεδομένα δίνουν συχνά πολύ διαφορετικά μοντέλα

μειώνοντας την ακρίβεια της πρόβλεψης. Η *ridge regression* είναι μια διαδικασία κατά την οποία συρρικνώνονται οι συντελεστές και άρα είναι πιο σταθερή. Όμως, δε μηδενίζεται κανένας συντελεστής, οπότε παίρνουμε μοντέλα που δύσκολα ερμηνεύονται.

Σε αυτό το κεφάλαιο, προτείνεται μια νέα τεχνική, που ονομάζεται *LASSO* (*Least Absolute Shrinkage and Selection Operator*) ή *Ελάχιστη Απόλυτη Συρρίκνωση και Τελεστής Επιλογής*. Αυτή η τεχνική συρρικνώνει κάποιους συντελεστές ενώ θέτει τους υπόλοιπους ίσους με 0, οπότε προσπαθεί να διατηρήσει τα καλά χαρακτηριστικά της *subset selection* και της *ridge regression*.

3.1. Περιγραφή της Μεθόδου *Lasso*

Υποθέτουμε ότι είτε οι παρατηρήσεις είναι ανεξάρτητες είτε ότι τα y_i είναι ανεξάρτητα με δοσμένα τα x_{ij}. Υποθέτουμε επίσης ότι τα x_{ij} είναι κανονικοποιημένα έτσι ώστε $\sum_i x_{ij}/N = 0$ και $\sum_i x_{ij}^2/N = 1$.

Αν το διάνυσμα $\hat{\beta}$ των εκτιμώμενων συντελεστών είναι το $(\hat{\beta}_1, ..., \hat{\beta}_p)'$, τότε ο *Lasso* εκτιμητής $(\hat{\alpha}, \hat{\beta})$ ορίζεται ως εξής:

$$Y = a + X\beta + e$$

ή

$$\begin{cases} (\hat{\alpha}, \hat{\beta}) = \arg\min\{\sum_{i=1}^{N}(y_i - \alpha - \sum_j \beta_j x_{ij})^2\} \\ \\ \tau.\omega.\ \sum_j |\beta_j| \leq t \end{cases} \tag{3.1}$$

Εδώ το $t \geq 0$ είναι μια ρυθμιζόμενη *παράμετρος*. Για κάθε t, η λύση για το α είναι η $\hat{\alpha} = \bar{y}$. Μπορούμε, χωρίς περιορισμό της γενικότητας, να θεωρήσουμε ότι $\bar{y} = 0$ και άρα να παραλείψουμε το α. Η παράμετρος $t \geq 0$ προσδιορίζει το μέγεθος της συρρίκνωσης των εκτιμητών. Έστω $\hat{\beta}_j^0, j = 1, ..., p$ οι εκτιμητές ελαχίστων τετραγώνων και $t_0 = \sum_{j=1}^{p} |\hat{\beta}_j^0|$. Για $t < t_0$, θα προκληθεί συρρίκνωση των λύσεων προς το 0 και μερικοί συντελεστές μπορεί να γίνουν ίσοι με 0. Για παράδειγμα, αν $t = t_0/2$, το αποτέλεσμα θα είναι παρόμοιο με αυτό της εύρεσης του καλύτερου υποσυνόλου μεγέθους $p/2$.

Ο *Breiman* (1993) πρότεινε τη μέθοδο *garotte* η οποία ελαχιστοποιεί το άθροισμα:

$$\begin{cases} \sum_{i=1}^{N}(y_i - \alpha - \sum_j c_j \hat{\beta}_j^0 x_{ij})^2 \\ \\ \tau.\omega.\ c_j \geq 0, \sum_j c_j \leq t \end{cases} \tag{3.2}$$

Η *garotte* ξεκινάει με τους *OLS* εκτιμητές και τους συρρικνώνει με τη βοήθεια μη αρνητικών συντελεστών των οποίων το άθροισμα έχει άνω φράγμα το t. Ο *Breiman* (1993) έδειξε ότι η *garotte* δίνει μικρότερο σφάλμα πρόβλεψης από τη *subset selection* αλλά περίπου το ίδιο με τη *ridge regression* εκτός εάν το πραγματικό μοντέλο έχει πολλούς μικρούς μη μηδενικούς συντελεστές.

Ένα μειονέκτημα της *garotte* είναι ότι η λύση της εξαρτάται από το πρόσημο και την τιμή των OLS εκτιμητών. Αν δηλαδή οι OLS εκτιμητές δε συμπεριφέρονται καλά, το ίδιο θα ισχύει και για τους *garotte* εκτιμητές. Αντίθετα, η *Lasso* αποφεύγει τη χρήση των OLS εκτιμητών.

Υποθέτουμε τώρα ότι X είναι ο $n \times p$ πίνακας σχεδιασμού και ότι είναι ορθοκανονικός, δηλαδή $X'X = I$. Μπορεί εύκολα να αποδειχθεί ότι οι λύσεις της εξίσωσης (3.1) είναι:

$$\hat{\beta}_j = sign(\hat{\beta}_j^0)(|\hat{\beta}_j^0| - \gamma)^+ \tag{3.3}$$

όπου το γ καθορίζεται από τη συνθήκη $\sum_{j=1}^{p} |\hat{\beta}_j| = t$.

Στην περίπτωση του ορθοκανονικού πίνακα σχεδιασμού, η επιλογή του βέλτιστου υποσυνόλου μεγέθους k σημαίνει ότι πρέπει να επιλέξουμε τους k απολύτως μεγαλύτερους συντελεστές και να θέσουμε τους υπόλοιπους ίσους με 0. Για κάποιο λ, αυτό είναι ισοδύναμο με το να θέσουμε $\hat{\beta}_j = \hat{\beta}_j^0$ αν $|\hat{\beta}_j^0| > \lambda$, ή $\hat{\beta}_j = 0$ διαφορετικά. Η *ridge regression* τότε ελαχιστοποιεί το:

$$\sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij})^2 + \lambda \sum_j \beta_j^2$$

όπου λ ο *Lagrange* πολλαπλασιαστής. Ισοδύναμα, ελαχιστοποιεί το:

$$\begin{cases} \sum_{i=1}^{N} (y_i - \sum_j \beta_j x_{ij})^2 \\ \\ \tau.\omega. \sum_j \beta_j^2 \leq t \end{cases} \tag{3.4}$$

Οι λύσεις της *ridge regression* είναι:

$$\frac{1}{1+\gamma} \hat{\beta}_j^0$$

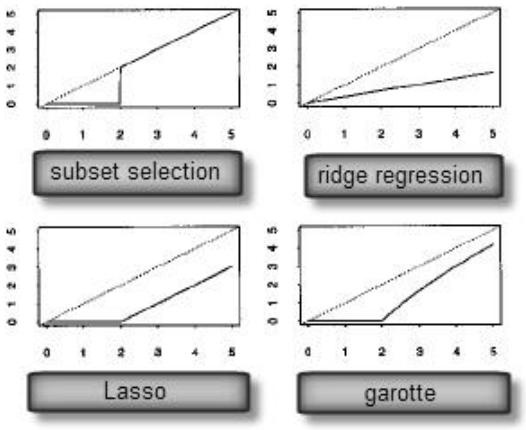
όπου το γ εξαρτάται από το λ ή το t.

Προκύπτει επίσης ότι οι *garotte* εκτιμητές είναι:

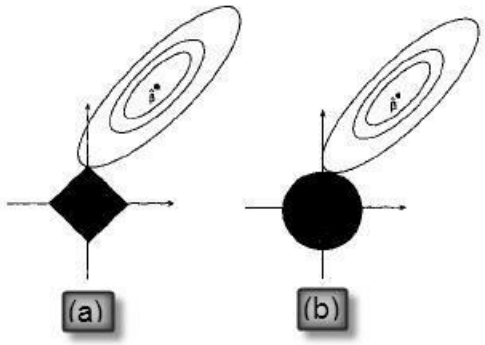
$$(1 - \frac{\gamma}{\hat{\beta}_j^{0\,2}})^+ \hat{\beta}_j^0$$

Το Σχήμα 3.1 δείχνει τη μορφή αυτών των συναρτήσεων, δηλαδή το πώς μεταβάλλονται οι εκτιμώμενοι συντελεστές παλινδρόμησης σε κάθε περίπτωση. Η διακεκομένη γραμμή απεικονίζει τη μεταβολή του εκτιμητή ελαχίστων τετραγώνων. Η *ridge regression* μειώνει τους συντελεστές κατά ένα σταθερό παράγοντα, ενώ η *Lasso* τους ελαττώνει κατά ένα σταθερό συντελεστή θέτοντας κάποιους ίσους με 0. Η *garotte* μοιάζει πολύ με τη *Lasso* αλλά δίνει μικρότερη συρρίκνωση για μεγαλύτερους συντελεστές.

Είναι φανερό από το Σχήμα 3.1 ότι η μέθοδος *Lasso* δίνει συχνά συντελεστές ίσους με 0. Για να εξετάσουμε γιατί συμβαίνει αυτό και στη μη-ορθοκανονική περίπτωση και γιατί δε συμβαίνει στη *ridge regression* η οποία χρησιμοποιεί τον περιορισμό $\sum \beta_j^2 \leq t$ αντί του $\sum |\beta_j| \leq t$, το Σχήμα 3.2 θα μας βοηθήσει (εδώ $p = 2$).



Σχήμα 3.1: Μορφή της Συρρίκνωσης του Συντελεστή [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267-288$]



Σχήμα 3.2: Εικόνα Εκτίμησης για τη (a)*Lasso* και για τη (b)*Ridge Regression* [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267-288$]

Το άθροισμα $\sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2$ ισούται με την τετραγωνική μορφή:

$$(\beta - \hat{\beta}^0)' X' X (\beta - \hat{\beta}^0)$$

σύν μια σταθερά. Οι ελλιπτικές τροχιές της συνάρτησης αυτής φαίνονται στο Σχήμα $3.2(a)$. Το κέντρο τους είναι οι OLS εκτιμητές, ενώ ο ρόμβος είναι η φραγμένη περιοχή που καθορίζεται από τον περιορισμό της μεθόδου. Η λύση της *Lasso* είναι το πρώτο σημείο πάνω στο οποίο οι τροχιές ακουμπούν το ρόμβο, και αυτό συμβαίνει μερικές φορές σε γωνία, η οποία θα αντιστοιχεί σε μηδενικό συντελεστή. Στο Σχήμα $3.2(b)$, παριστάνεται η *ridge regression*, όπου δεν υπάρχουν γωνίες οπότε δύσκολα επιτυγχάνεται μηδενικός συντελεστής.

Ένα ενδιαφέρον ερώτημα προκύπτει από το Σχήμα. Είναι δυνατόν τα πρόσημα των *Lasso* εκτιμητών να είναι διαφορετικά από αυτά των OLS εκτιμητών; Παρατηρούμε ότι

εφόσον οι μεταβλητές είναι κανονικοποιημένες, οι άξονες των ελλείψεων για $p = 2$ έχουν διαφορά 45° από τους άξονες των συντελεστών. Μπορεί να αποδειχθεί ότι οι τροχιές θα ακουμπούν το ρόμβο στο τεταρτημόριο που περιέχει τους OLS εκτιμητές. Όμως, για $p > 2$ και όταν υπάρχει έστω και μία μέτρια συσχέτιση στα δεδομένα, αυτό παύει να ισχύει.

Παρόλο που η *garotte* διατηρεί τα πρόσημα των OLS εκτιμητών, η *Lasso* μπορεί να τα αλλάξει. Ακόμα και σε περιπτώσεις όπου οι *Lasso* εκτιμητές έχουν το ίδιο πρόσημο με τους *garotte* εκτιμητές, η παρουσία των OLS εκτιμητών στον υπολογισμό των *garotte* εκτιμητών μπορεί να τους κάνει να συμπεριφέρονται διαφορετικά. Το άθροισμα $\sum c_j \hat{\beta}_j^0 x_{ij}$ με περιορισμό $\sum c_j \leq t$ μπορεί να γραφτεί ως $\sum \beta_j x_{ij}$ με περιορισμό $\sum \beta_j / \hat{\beta}_j^0 \leq t$. Αν για παράδειγμα ισχύει $p = 2$ και $\hat{\beta}_1^0 > \hat{\beta}_2^0 > 0$, αυτό σημαίνει γεωμετρικά ότι επιμηκύνουμε τον οριζόντιο άξονα του ρόμβου στο σχήμα 3.2(α). Άρα, μεγαλύτερες τιμές για το β_1 και μικρότερες τιμές για το β_2 προτιμώνται από την *garotte*.

Υποθέτουμε τώρα ότι $p = 2$ και ότι χωρίς περιορισμό της γενικότητας οι εκτιμητές $\hat{\beta}_1^0, \hat{\beta}_2^0$ είναι θετικοί. Τότε, από τη σχέση (3.3), προκύπτει ότι οι *Lasso* εκτιμητές είναι οι:

$$\hat{\beta}_j = (\hat{\beta}_j^0 - \gamma)^+$$

όπου το γ είναι τέτοιο ώστε $\hat{\beta}_1 + \hat{\beta}_2 = t$. Ο τύπος ισχύει για $t \leq \hat{\beta}_1^0 + \hat{\beta}_2^0$ ακόμα κι αν οι επεξηγηματικές μεταβλητές είναι συσχετισμένες μεταξύ τους. Λύνοντας ως προς γ έχουμε:

$$\begin{cases} \hat{\beta}_1 = (\frac{t}{2} + \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2})^+ \\ \\ \hat{\beta}_2 = (\frac{t}{2} - \frac{\hat{\beta}_1^0 - \hat{\beta}_2^0}{2})^+ \end{cases} \tag{3.5}$$

Μπορεί να αποδειχθεί ότι σε αντίθεση με τη *Lasso*, η μορφή της συρρίκνωσης στη *ridge regression* εξαρτάται από τη συσχέτιση των μεταβλητών.

Αν θέλουμε τώρα να υπολογίσουμε το τυπικό σφάλμα του *Lasso* εκτιμητή, παρατηρούμε ότι ο εκτιμητής αυτός είναι μια μη γραμμική και μη παραγωγίσιμη συνάρτηση των τιμών της y ακόμα και για συγκεκριμένο t. Οπότε, είναι δύσκολο να υπολογίσουμε έναν ακριβή εκτιμητή του τυπικού σφάλματός του. Όμως, η σταθεροποίηση του t είναι ανάλογη της επιλογής ενός βέλτιστου υποσυνόλου, οπότε μπορούμε να χρησιμοποιήσουμε το τυπικό σφάλμα του εκτιμητή ελαχίστων τετραγώνων για αυτό το υποσύνολο.

Μπορούμε να προσεγγίσουμε τον *Lasso* εκτιμητή γράφοντας το $\sum |\beta_j|$ ως $\sum \beta_j^2 / |\beta_j|$. Τότε, αν $\tilde{\beta}$ είναι ο *Lasso* εκτιμητής, μπορούμε να προσεγγίσουμε τη λύση με τη βοήθεια της *ridge regression* της μορφής $\beta^* = (X'X + \lambda W^-)^{-1} X'y$, όπου W ο διαγώνιος πίνακας με διαγώνια στοιχεία $|\tilde{\beta}_j|$, $j = 1, 2, ..., p$, W^- ο γενικευμένος αντίστροφός του W και λ τέτοιο ώστε $\sum |\beta_j^*| = t$. Ο πίνακας συνδιασποράς των εκτιμητών μπορεί τότε να προσεγγιστεί από τον πίνακα:

$$(X'X + \lambda W^-)^{-1} X'X (X'X + \lambda W^-)^{-1} \hat{\sigma^2} \tag{3.6}$$

όπου $\hat{\sigma^2}$ είναι ένας εκτιμητής του σφάλματος της διασποράς.

3.2. Σφάλμα πρόβλεψης και εκτίμηση του t

Σε αυτή την παράγραφο, θα περιγράψουμε τρεις μεθόδους για εκτίμηση της *Lasso* παραμέτρου t της μεθόδου *Lasso*:

- $cross - validation$
- γενικευμένη $cross - validation$
- έναν αναλυτικό αμερόληπτο εκτιμητή του κινδύνου

Διαθέτουμε τις παρατηρήσεις (Y, X), όπου Y είναι το $N + 1$ διάνυσμα των τιμών της εξαρτημένης μεταβλητής και X είναι ο $N \times p$ πίνακας των παρατηρούμενων τιμών για τις επεξηγηματικές μεταβλητές. Υποθέτουμε ότι συνδέονται μέσω της σχέσης:

$$Y = h(X) + e \tag{3.7}$$

όπου $E(e) = 0$, $var(e_i) = \sigma^2$, $i = 1, 2, ..., p$, και $h(X)$ μια συνάρτηση του πίνακα X.

Το μέσο τετραγωνικό σφάλμα ενός εκτιμητή $\hat{h}(X)$ του $h(X)$ ορίζεται ως εξής:

$$ME = E[\hat{h}(X) - h(X)]^2 \tag{3.8}$$

Ένα παρόμοιο με το ME μέτρο είναι το *σφάλμα πρόβλεψης του* $\hat{h}(X)$ και δίνεται από τον τύπο:

$$PE = E[Y - \hat{h}(X)]^2 = ME + \sigma^2 \tag{3.9}$$

Αν εκτιμήσουμε το σφάλμα πρόβλεψης για τη μέθοδο *Lasso* μέσω της $cross-validation$, η *Lasso* θα δίνεται σε συνάρτηση με την κανονικοποιημένη παράμετρο $s = t/\sum \hat{\beta}_j^0$ και εκτιμάμε το σφάλμα πρόβλεψης για τιμές του s μεταξύ 0 και 1. Επιλέγουμε την τιμή $\hat{s}$ που θα δώσει τη μικρότερη PE. Μπορούμε να γράψουμε τα αποτελέσματα σε συνάρτηση με το ME αντί του PE. Για γραμμικά μοντέλα όπου $h(X) = X\beta$, ισχύει:

$$ME = (\hat{\beta} - \beta)'V(\hat{\beta} - \beta)$$

όπου V ο πίνακας συνδιασποράς του X.

Μια *δεύτερη μέθοδος* για την εκτίμηση του t μπορεί να προκύψει από γραμμική προσέγγιση του t ως εξής. Γράφοντας τον περιορισμό $\sum |\beta_j| \leq t$ ως $\sum \beta_j^2/|\beta_j| \leq t$, αυτός είναι ισοδύναμος με το να προσθέσουμε ένα *Lagrangian* όρο $\lambda \sum \beta_j^2/|\beta_j|$ στο άθροισμα τετραγώνων των σφαλμάτων, όπου το λ θα εξαρτάται από το t. Άρα, η λύση $\tilde{\beta}$ αποτελεί τελικά το $ridge$ εκτιμητή:

$$\tilde{\beta} = (X'X + \lambda W^-)^{-1}X'y$$

όπως είχε αναφερθεί στην προηγούμενη παράγραφο.

Επομένως, το πλήθος των σημαντικών παραμέτρων στην περιορισμένη μορφή $\tilde{\beta}$ μπορεί να προσεγγιστεί από τη σχέση:

$$p(t) = tr\{X(X'X + \lambda W^{-})^{-1}X'\}$$

Έστω $RSS(t)$ το άθροισμα τετραγώνων των σφαλμάτων όταν υπάρχει ο περιορισμός t. Υπολογίζουμε τότε την παρακάτω ποσότητα, η οποία έχει τη μορφή μιας *γενικευμένης* $cross-validation$:

$$GCV(t) = \frac{1}{N}\frac{RSS(t)}{[1-p(t)/N]^2}$$

Η *τρίτη μέθοδος* στηρίζεται στον αμερόληπτο εκτιμητή του *Stein*. Έστω z ένα τυχαίο διάνυσμα που ακολουθεί την κατανομή $N(\mu, I)$. Αν $\hat{\mu}$ ένας εκτιμητής του μ και $\hat{\mu} = z + g(z)$, όπου $g : \Re^p \to \Re^p$ μια σχεδόν παντού παραγωγίσιμη συνάρτηση, τότε ο *Stein* (1981) έδειξε ότι:

$$E_\mu\|\hat{\mu} - \mu\|^2 = p + E_\mu(\|g(z)\|^2 + 2\sum_{i=1}^{p} dg_i/dz_i)$$

Μπορούμε να εφαρμόσουμε αυτό το αποτέλεσμα στον *Lasso* εκτιμητή (3.3). Αν δηλώσουμε το εκτιμώμενο τυπικό σφάλμα του $\hat{\beta}_j^0$ με $\hat{\tau} = \hat{\sigma}/\sqrt{N}$, όπου $\hat{\sigma}^2 = \sum(y_i - \hat{y}_i)^2/(N-p)$, τότε οι τυπικές αποκλίσεις $\hat{\beta}_j^0/\hat{\tau}$ είναι ελάχιστα εξαρτημένες μεταξύ τους και από την παραπάνω σχέση προκύπτει ο τύπος:

$$R(\hat{\beta}(\gamma)) \approx \hat{\tau}^2\{p - 2\hat{j} + \sum_{j=1}^{p}\max(|\hat{\beta}_j^0/\hat{\tau}|, \gamma)^2\},$$

όπου το $\hat{j}$ είναι τέτοιο ώστε $|\hat{\beta}_j^0/\hat{\tau}| < \gamma$. Η ποσότητα $R(\hat{\beta}(\gamma))$ είναι ένας σχεδόν *αμερόληπτος εκτιμητής του κινδύνου*, δηλαδή του μέσου τετραγωνικού σφάλματος $E[\hat{\beta}(\gamma) - \beta]^2$, όπου $\hat{\beta}_j(\gamma) = sign(\hat{\beta}_j^0)(|\hat{\beta}_j^0/\hat{\tau}| - \gamma)^+$, $j = 1, 2, ..., p$. Οπότε, μπορούμε να πάρουμε το γ εκείνο που ελαχιστοποιεί τον $R[\hat{\beta}(\gamma)]$:

$$\hat{\gamma} = \arg\min_{\gamma\geq 0}\{R[\hat{\beta}(\gamma)]\}$$

Έτσι προκύπτει ο εκτιμητής της *Lasso* παραμέτρου t:

$$\hat{t} = \sum_{j=1}^{p}(|\hat{\beta}_j^0| - \hat{\gamma})^+$$

Παραδείγματα μας δείχνουν ότι αυτή η μέθοδος δίνει καλούς εκτιμητές για το t. Έστω $X'X = V$, $Z = XV^{-1/2}$ και $\theta = \beta V^{-1/2}$. Εφόσον οι στήλες του X έχουν κανονικοποιηθεί, η περιοχή $\sum|\theta_j| \leq t$ διαφέρει από τη $\sum|\beta_j| \leq t$ αλλά δίνουν προβολές ίσου μέτρου. Άρα, η βέλτιστη τιμή του $\hat{t}$ είναι περίπου η ίδια σε κάθε περίσταση. Πάντως, όσον αφορά

το υπολογιστικό κόστος, η μέθοδος του *Stein* υπερέχει της $cross-validation$ εκτίμησης του t.

3.3. Αλγόριθμος υπολογισμού των λύσεων της Μεθόδου *Lasso*

Έστω $t \geq 0$. Το πρόβλημα (3.1) μπορεί να θεωρηθεί ως ένα πρόβλημα ελαχίστων τετραγώνων με 2^p ανισοτικούς περιορισμούς, οι οποίοι αντιστοιχούν στα 2^p διαφορετικά πιθανά πρόσημα των $\beta_j, j = 1, ..., p$. Περιγράφεται παρακάτω η διαδικασία που μας δίνει τη λύση για το γραμμικό πρόβλημα ελαχίστων τετραγώνων υπό ένα γενικό γραμμικό ανισοτικό περιορισμό $G\beta \leq h$. Εδώ G είναι ένας $m \times p$ πίνακας που αντιστοιχεί σε m γραμμικούς ανισοτικούς περιορισμούς στο διάνυσμα β μήκους p. Για το πρόβλημά μας, όμως, το $m = 2^p$ μπορεί να είναι πολύ μεγάλο, οπότε η άμεση εφαρμογή της διαδικασίας αυτής απαιτεί πολύ μεγάλο πλήθος υπολογισμών.

Το πρόβλημα, όμως, μπορεί να λυθεί παρουσιάζοντας τους ανισοτικούς περιορισμούς διαδοχικά ως εξής: Έστω $g(\beta) = \sum_{i=1}^{N}(y_i - \sum_j \beta_j x_{ij})^2$ και δ_i, $i = 1, 2, ..., 2^p$, τα διανύσματα της μορφής $(\pm 1, \pm 1, ..., \pm 1)$ μήκους p. Τότε η συνθήκη $\sum |\beta_j| \leq t$ είναι ισοδύναμη της $\delta_i'\beta \leq t$ για κάθε i. Για δοσμένο β, έστω $E = \{i : \delta_i'\beta = t\}$ και $S = \{i : \delta_i'\beta < t\}$. Ας συμβολίσουμε με G_E τον πίνακα του οποίου οι γραμμές είναι ίσες με τα δ_i για τα οποία ισχύει ότι $i \epsilon E$. Έστω $\mathbf{1}$ το μοναδιαίο διάνυσμα πλήθους ίσου με το πλήθος των γραμμών του G_E.

Ο αλγόριθμος ξεκινά με $E = \{i_0\}$, όπου $\delta_{i_0} = sign(\hat{\beta})$, όπου $\hat{\beta}$ ο OLS εκτιμητής. Δίνει τότε τη λύση για το πρόβλημα ελαχίστων τετραγώνων με τον περιορισμό ότι $\delta_{i_0}'\beta \leq t$ και έπειτα ελέγχει αν $\sum |\beta_j| \leq t$. Αν ισχύει, τότε ο υπολογισμός έχει τελειώσει. Αν όχι, το i για το οποίο παραβιάζεται ο περιορισμός προστίθεται στο E και η διαδικασία συνεχίζεται μέχρις ότου $\sum |\beta_j| \leq t$.

Ο αλγόριθμος λοιπόν έχει ως εξής:

1. Ξεκίνα με $E = i_0$ όπου $\delta_{i_0} = sign(\hat{\beta^0})$ και $\hat{\beta^0}$ ο OLS εκτιμητής.

2. Βρες το $\hat{\beta}$ που ελαχιστοποιεί το $g(\beta)$ έτσι ώστε $G_E\beta \leq t\mathbf{1}$

3. Όταν ισχύει ότι $\sum |\hat{\beta}_j| > t$,

4. Πρόσθεσε το i στο σύνολο E όπου $\delta_i = sign(\hat{\beta})$. Βρες το $\hat{\beta}$ που ελαχιστοποιεί το $g(\beta)$ ώστε να ισχύει ότι $G_E\beta \leq t\mathbf{1}$.

Η διαδικασία πρέπει να συγκλίνει σε ένα πεπερασμένο πλήθος βημάτων αφού μόνο ένα στοιχείο προστίθεται στο E σε κάθε βήμα και υπάρχουν συνολικά 2^p στοιχεία. Το τελικό αποτέλεσμα είναι η λύση του αρχικού προβλήματος. Μια παραλλαγή της διαδικασίας απαλείφει τα στοιχεία του E (βήμα 4) για τα οποία ο ισοτικός περιορισμός δεν ικανοποιείται. Αυτή είναι πιο αποτελεσματική διαδικασία αλλά δεν είναι φανερό αν συγκλίνει ή όχι. Το γεγονός ότι ο αλγόριθμος πρέπει να σταματήσει μετά από το πολύ 2^p επαναλήψεις επιβαρύνει τους υπολογισμούς αν το p είναι μεγάλο. Στην πράξη όμως, έχει βρεθεί ότι ο

μέσος όρος των επαναλήψεων που απαιτούνται είναι της τάξης του $0.5p - 0.75p$, άρα ο αλγόριθμος είναι αρκετά ικανοποιητικός.

Ένας τελείως διαφορετικός αλγόριθμος έχει προταθεί από τον *David Gay*. Γράφουμε κάθε β_j ως τη διαφορά $\beta_j^+ - \beta_j^-$, όπου β_j^+ και β_j^- είναι δύο μη αρνητικοί αριθμοί. Τότε λύνουμε το πρόβλημα ελαχίστων τετραγώνων με τους περιορισμούς $\beta_j^+ \geq 0$, $\beta_j^- \geq 0$ και $\sum_j \beta_j^+ + \sum_j \beta_j^- \leq t$. Με αυτόν τον τρόπο μετατρέπουμε το αρχικό πρόβλημα με p μεταβλητές και 2^p περιορισμούς σε ένα πρόβλημα με περισσότερες μεταβλητές $(2p)$ αλλά αρκετά λιγότερους περιορισμούς $(2p+1)$. Μπορεί να δειχθεί ότι αυτό το πρόβλημα έχει την ίδια λύση με το αρχικό και είναι ελαφρώς πιο γρήγορο.

3.4. Παραδείγματα - Συγκρίσεις Μεθόδων

Στα παρακάτω παραδείγματα, συγκρίνουμε τον εκτιμητή ελαχίστων τετραγώνων με τους εκτιμητές της *Lasso*, της *garotte*, της *Subset selection* και της *ridge regression*.

Method	*Median mean-squared error*	*Average no. of 0 coefficients*	*Average ŝ*
Least squares	2.79 (0.12)	0.0	—
Lasso (cross-validation)	2.43 (0.14)	3.3	0.63 (0.01)
Lasso (Stein)	2.07 (0.10)	2.6	0.69 (0.02)
Lasso (generalized cross-validation)	1.93 (0.09)	2.4	0.73 (0.01)
Garotte	2.29 (0.16)	3.9	—
Best subset selection	2.44 (0.16)	4.8	—
Ridge regression	3.21 (0.12)	0.0	—

Πίνακας 3.1: Αποτελέσματα Παραδείγματος 1 [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267-288$]

Παράδειγμα 1:
Εδώ έχουμε προσαρμόσει 50 σύνολα δεδομένων 20 παρατηρήσεων από το μοντέλο:

$$y = \beta' x + \sigma\varepsilon$$

όπου $\beta = (3, 1.5, 0, 0, 2, 0, 0, 0)'$ και το σφάλμα ακολουθεί την τυπική κανονική κατανομή. Η συσχέτιση μεταξύ x_i και x_j είναι $\rho^{|i-j|}$ με $\rho = 0.5$. Θέτουμε $\sigma = 3$ και παίρνουμε σφάλμα της τάξεως του 5.7. Ο πίνακας 3.1 δείχνει τα μέσα τετραγωνικά σφάλματα 200 προσομοιώσεων από το μοντέλο αυτό. Η *Lasso* συμπεριφέρεται καλύτερα και ακολουθεί η *garotte* και η *ridge regression*. Η εκτίμηση της *Lasso* παραμέτρου μέσω της γενικευμένης $cross-validation$ μεθόδου συμπεριφέρεται καλύτερα. Η *subset selection* βρίσκει το σωστό σχεδόν πλήθος των μηδενικών συντελεστών αλλά πάσχει από μεγάλη μεταβλητότητα.

Ο πίνακας 3.2 δείχνει τα πέντε μοντέλα που επέλεγε πιο συχνά η *Lasso*: παρόλο που το σωστό μοντέλο (1,2,5) επιλέγεται μόνο 2,5% των περιπτώσεων, το 95,5% των μοντέλων περιέχει τις (1,2,5). Τα μοντέλα που επιλέγει πιο συχνά η *subset regression* φαίνονται

Model	*Proportion*
1245678	0.055
123456	0.050
1258	0.045
1245	0.045
13 others	
125 (and 5 others)	0.025

Πίνακας 3.2: Τα μοντέλα που επέλεγε πιο συχνά η *Lasso* (γενικευμένη *cross−validation*) στο Παράδειγμα 1 [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267-288$]

Model	Proportion
125	0.240
15	0.200
1	0.095
1257	0.040

Πίνακας 3.3: Τα μοντέλα που επέλεγε η *subset selection* στο Παράδειγμα 1 [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267-288$]

στον πίνακα 3.3. Το σωστό μοντέλο επιλέγεται πιο συχνά (24% των περιπτώσεων) αλλά μόνο το 53,5% των μοντέλων περιέχει τις (1,2,5).

Παράδειγμα 2:

Το παράδειγμα αυτό είναι το ίδιο με το πρώτο με τη διαφορά ότι $\beta_j = 0,85\ \forall j$. Το σ είναι και πάλι ίσο με 3, ενώ το σφάλμα είναι της τάξεως του 1,8. Ο πίνακας 3.4 δείχνει ότι η *ridge regression* συμπεριφέρεται καλύτερα και ακολουθεί η *Lasso*.

Method	*Median mean-squared error*	*Average no. of 0 coefficients*	*Average ŝ*
Least squares	6.50 (0.64)	0.0	—
Lasso (cross-validation)	5.30 (0.45)	3.0	0.50 (0.03)
Lasso (Stein)	5.85 (0.36)	2.7	0.55 (0.03)
Lasso (generalized cross-validation)	4.87 (0.35)	2.3	0.69 (0.23)
Garotte	7.40 (0.48)	4.3	—
Subset selection	9.05 (0.78)	5.2	—
Ridge regression	2.30 (0.22)	0.0	—

Πίνακας 3.4: Αποτελέσματα του Παραδείγματος 2 [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267-288$]

Παράδειγμα 3:

Εδώ το μοντέλο είναι το ίδιο με αυτό του παραδείγματος 1 με τη διαφορά ότι $\beta = (5,0,0,0,0,0,0,0)$ και $\sigma = 2$, ενώ το σφάλμα είναι της τάξεως του 7. Τα αποτελέσματα στον πίνακα 3.5 δείχνουν ότι η *garotte* και η *subset regression* συμπεριφέρονται καλύτερα από τις υπόλοιπες, και τις ακολουθεί η *Lasso*. Η *ridge regression* δεν είναι κατάλληλη εδώ και έχει μεγαλύτερο μέσο τετραγωνικό σφάλμα από την εκτίμηση με ελάχιστα τετράγωνα.

Method	*Median mean-squared error*	*Average no. of 0 coefficients*	*Average ŝ*
Least squares	2.89 (0.04)	0.0	—
Lasso (cross-validation)	0.89 (0.01)	3.0	0.50 (0.03)
Lasso (Stein)	1.26 (0.02)	2.6	0.70 (0.01)
Lasso (generalized cross-validation)	1.02 (0.02)	3.9	0.63 (0.04)
Garotte	0.52 (0.01)	5.5	—
Subset selection	0.64 (0.02)	6.3	—
Ridge regression	3.53 (0.05)	0.0	—

Πίνακας 3.5: Αποτελέσματα του Παραδείγματος 3 [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267-288$]

Παράδειγμα 4:

Εδώ προσαρμόζουμε 50 σύνολα δεδομένων με το καθένα να έχει 100 παρατηρήσεις από 40 μεταβλητές. Εδώ η *subset regression* δεν είναι εφαρμόσιμη αφού $p > 30$. Ορίζουμε τις επεξηγηματικές μεταβλητές $x_{ij} = z_{ij} + z_i$, όπου z_{ij} και z_i είναι ανεξάρτητες τυχαίες μεταβλητές και ακολουθούν την τυπική κανονική κατανομή. Υποθέτουμε ότι υπάρχει συσχέτιση ίση με 0,5 ανάμεσα σε κάθε ζευγάρι επεξηγηματικών μεταβλητών. Ο συντελεστής είναι το διάνυσμα $\beta = (0, 0, ..., 0, 2, 2, ..., 2, 0, 0, ..., 0, 2, 2, ..., 2)$, όπου υπάρχουν 10 ίσα στοιχεία σε κάθε *block*. Ορίζουμε $y = \beta' x + 15\varepsilon$ όπου το ε ακολουθεί τυπική κανονική κατανομή και έχουμε σφάλμα της τάξεως του 9. Τα αποτελέσματα στον πίνακα 3.6 δείχνουν ότι η *ridge regression* συμπεριφέρεται καλύτερα, με τη *Lasso* να ακολουθεί.

Method	*Median mean-squared error*	*Average no. of 0 coefficients*	*Average ŝ*
Least squares	137.3 (7.3)	0.0	—
Lasso (Stein)	80.2 (4.9)	14.4	0.55 (0.02)
Lasso (generalized cross-validation)	64.9 (2.3)	13.6	0.60 (0.88)
Garotte	94.8 (3.2)	22.9	—
Ridge regression	57.4 (1.4)	0.0	—

Πίνακας 3.6: Αποτελέσματα του Παραδείγματος 4 [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267 - 288$]

Η μέση τιμή των *Lasso* συντελεστών σε καθένα από τα 4 *blocks* των 10 είναι 0,5 (0,06), 0,92 (0,07), 1,56 (0,08) και 2,33 (0,09) αντίστοιχα. Παρόλο που η *Lasso* δίνει μόνο 14,4 μηδενικούς συντελεστές κατά μέσο όρο, η μέση τιμή του $\hat{s}$ (0,55) είναι πολύ κοντά στην πραγματική αναλογία, δηλαδή 0,5.

3.5. Συμπεράσματα πάνω στη μέθοδο *Lasso*

Στις πρώτες 4 παραγράφους, προτάθηκε μια πρόσφατη μέθοδος για συρρίκνωση και επιλογή μεταβλητών για προβλήματα παλινδρόμησης. Η *Lasso* δεν εστιάζει σε υποσύνολα αλλά ορίζει μια συνεχή διαδικασία συρρίκνωσης που μπορεί να δώσει συντελεστές που είναι ακριβώς ίσοι με 0. Μέσα από παραδείγματα, διαπιστώσαμε ότι η *Lasso* αποδεικνύεται μερικές φορές καλύτερη ακόμα και από την *subset selection* ή την *ridge regression*. Μελετήθηκε το κέρδος της κάθε μεθόδου σε τρεις διαφορετικές περιπτώσεις:

1. *Μικρό πλήθος σημαντικών μεταβλητών*: Εδώ η *subset selection* είναι η καλύτερη, η *Lasso* όχι τόσο καλή, ενώ η *ridge regression* δεν ενδείκνυται.

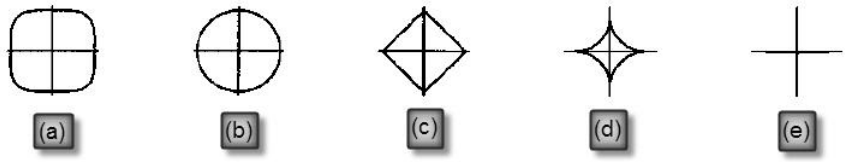
2. *Μέτριο πλήθος μέτρια σημαντικών μεταβλητών*: Εδώ η *Lasso* συμπεριφέρεται καλύτερα, με τη *ridge regression* να ακολουθεί, ενώ τελευταία έρχεται η *subset selection*.

3. *Μεγάλο πλήθος ασήμαντων μεταβλητών*: Η *ridge regression* κρίνεται η καλύτερη, ακολουθεί η *Lasso*, ενώ τελευταία έρχεται ξανά η *subset selection*.

Η μέθοδος *garotte* του *Breiman* (1993) συμπεριφέρεται λίγο καλύτερα από τη *Lasso* στην πρώτη περίπτωση και λίγο χειρότερα από αυτήν στα επόμενα δύο. Τα συμπεράσματα αυτά αναφέρονται στην ακρίβεια της πρόβλεψης. Οι *subset selection*, *Lasso* και *garotte* υπερτερούν της *ridge regression* στην κατασκευή μοντέλων που θέλουμε να ερμηνεύονται εύκολα.

Υπάρχουν στη βιβλιογραφία πολλές παραλλαγές αυτών των μεθόδων που προσπαθούν να τις βελτιώσουν. Οι *Frank* και *Friedman* (1993) προτείνουν μια γενίκευση της *ridge regression* και της *subset selection* κατά την οποία προσθέτουμε έναν επιπλέον όρο της μορφής $\lambda \sum_j |\beta_j|^q$ στο άθροισμα τετραγώνων των σφαλμάτων. Αυτό είναι ισοδύναμο με τον περιορισμό της μορφής $\sum_j |\beta_j|^q \leq t$ και ονομάζεται "γέφυρα". Η *Lasso* αντιστοιχεί σε $q = 1$. Προτείνουν ότι η κοινή εκτίμηση των β_j και q ενδέχεται να είναι μια αποτελεσματική στρατηγική αλλά δεν έχουν αναλυθεί τέτοια συμπεράσματα.

Το Σχήμα 3.3 δείχνει την κατάσταση στις δύο διαστάσεις. Η *subset selection* αντιστοιχεί στο $q \to 0$. Η τιμή $q = 1$ έχει το πλεονέκτημα να είναι πιο κοντά στην *subset selection* παρά στη *ridge regression*, όπου $q = 2$, και επίσης είναι η μικρότερη τιμή του q που δίνει μια κυρτή περιοχή.



Σχήμα 3.3: Τροχιές της σταθερής τιμής του $\sum_j |\beta_j|^q$
$(a)q = 4$, $(b)q = 2$, $(c)q = 1$, $(d)q = 0.5$, $(e)q = 0.1$ [*R.Tibshirani* (1996), *Regression Shrinkage and Selection via the Lasso*, *J.R.Statist.Soc.B*, $58, No.1, pp.267 - 288$]

Τα ενθαρρυντικά αποτελέσματα που αναφέρθηκαν εδώ δείχνουν ότι περιορισμοί με απόλυτες τιμές αποδεικνύονται χρήσιμοι για μια μεγάλη ποικιλία από στατιστικά προβλήματα εκτίμησης. Περαιτέρω έρευνα χρειάζεται για να εξετάσουμε αυτά τα ενδεχόμενα.

3.6. Το δυϊκό πρόβλημα της μεθόδου *Lasso*

Ο εκτιμητής της μεθόδου *Lasso* υπολογίζεται, όπως είδαμε, από το πρόβλημα (3.1). Αν θεωρήσουμε ότι $\bar{y} = 0$ και παραλείψουμε το α (αφού $\hat{\alpha} = \bar{y}$), τότε το πρόβλημα βελτιστοποίησης για την εύρεση του εκτιμητή της *Lasso* μπορεί να γραφτεί ως εξής:

$$\begin{cases} minimize_{\beta_1,\ldots,\beta_p} \frac{1}{2} \sum_{i=1}^{N} (y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \\ \tau.\omega. \sum_{j=1}^{p} |\beta_j| \leq t \end{cases} \tag{3.10}$$

Το πρόβλημα αυτό είναι δύσκολο να λυθεί και ο αλγόριθμος του *Tibshirani* (1996) που περιγράφηκε στην παράγραφο 3.3 δεν είναι αποτελεσματικός όταν το p είναι μεγάλο,

ενώ δεν μπορεί καθόλου να χρησιμοποιηθεί όταν $p > N$. Οι $M.R.Osborne, B.Presnell,$ και $B.A.Turlach$ (2000) εξετάζουν το δυϊκό πρόβλημα βελτιστοποίησης και παρέχουν έναν αλγόριθμο για τον υπολογισμό του εκτιμητή της μεθόδου $Lasso$, ο οποίος μπορεί να εφαρμοστεί και στην περίπτωση όπου $p > N$.

Ένα ισοδύναμο πρόβλημα με το (3.10) μπορεί να δωθεί ως εξής:

$$minimize_{\beta_1,...,\beta_p}\{\frac{1}{2}\sum_{i=1}^{N}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\} \tag{3.11}$$

όπου ο $Lagrangian$ πολλαπλασιαστής λ είναι ένας μη αρνητικός αριθμός.

Έστω $y = (y_1, ..., y_N)'$ το διάνυσμα των παρατηρήσεων για την εξαρτημένη μεταβλητή, $X = (x_1, ..., x_p)'$ ο $N \times p$ πίνακας των παρατηρήσεων για τις επεξηγηματικές μεταβλητές, όπου $x_j = (x_{1j}, ..., x_{Nj})'$ είναι η $j-$οστή του στήλη και έστω $A = X'X$. Υποθέτουμε ότι ο X έχει μέγιστη τάξη. Έστω $N(X) \subset \Re^p$ ο μηδενοχώρος του πίνακα X και β^0 ο OLS εκτιμητής. Προφανώς, αν $p \leq N$, τότε $N(X) = \{0\}$ και η λύση $\beta^0 = A^{-1}X'y$ είναι μοναδική. Αν όμως $p > N$, ο $N(X)$ έχει διάσταση $p - N$, ο β^0 δεν είναι μοναδικός και η σχέση $X(\beta^0 + \eta) = y$ ισχύει για κάθε $\eta \in N(X)$. Σε κάθε περίπτωση, πάντως, ορίζουμε:

$$t_0 = \min_{\eta \in N(X)} ||\beta^0 + \eta||_1$$

όπου $||\beta||_1 = \sum_{i=1}^{p}|\beta_i|$ είναι η L^1-νόρμα στον $\Re^p$. Υποθέτουμε ότι $t < t_0$, αφού αν $t \geq t_0$ ο εκτιμητής $Lasso$ είναι ισοδύναμος με τον OLS εκτιμητή.

Το πρόβλημα (3.10) μπορεί να γραφτεί ως εξής:

$$\begin{cases} minimize_\beta f(\beta) \\ \\ \tau.\omega.\ g(\beta) \geq 0 \end{cases} \tag{3.12}$$

όπου

$$f(\beta) = \frac{1}{2}(y - X\beta)'(y - X\beta) = \frac{1}{2}r'r$$

και

$$g(\beta) = t - \sum_{i=1}^{p}|\beta_i|$$

Αφού η f είναι συνεχής και η περιοχή που παίρνει τιμές το β είναι συμπαγής, το πρόβλημα (3.12) έχει σίγουρα λύση. Επιπλέον, αφού $t < t_0$, κάθε λύση β^* της (3.12) θα βρίσκεται στο σύνορο, δηλαδή θα ισχύει: $||\beta^*||_1 = t$. Εφόσον η g είναι κοίλη, η περιοχή που ορίζεται από την ανισότητα $g(\beta) \geq 0$ είναι κυρτή και αφού και η f είναι κυρτή, είναι φανερό ότι το σύνολο των λύσεων της (3.12) θα είναι κυρτό. Επίσης, αν $p \leq N$, τότε η f είναι αυστηρά κυρτή, επομένως η λύση είναι μοναδική. Τα συμπεράσματα αυτά συνοψίζονται στο παρακάτω θεώρημα:

Θεώρημα 1 *(Ύπαρξη και μοναδικότητα)*
Αν $t < t_0$, ισχύουν τα εξής:

α)Αν $p \leq N$, τότε το (3.12) έχει μοναδική λύση β^* και $||\beta^*|| = t$.

β)Αν $p > N$, τότε το (3.12) έχει λύση β^* και η ισότητα $||\beta^*||_1 = t$ ισχύει για κάθε λύση. Αν β_1^* και β_2^* είναι δύο λύσεις, τότε ο κυρτός συνδυασμός $\rho\beta_1^* + (1-\rho)\beta_2^*$ είναι επίσης λύση όταν $0 \leq \rho \leq 1$.□

Αν θεωρήσουμε το (3.12) ως ένα πρόβλημα κυρτού προγραμματισμού, η *Lagrangian* θα είναι:

$$L(\beta, \lambda) = f(\beta) - \lambda g(\beta) \tag{3.13}$$

Αν ορίσουμε

$$L^*(\beta) = \sup_{\lambda \geq 0} L(\beta, \lambda) \tag{3.14}$$

τότε θα ισχύει:

$$L^*(\beta) = \{ \begin{array}{l} f(\beta), \text{ αν } g(\beta) \geq 0 \\ +\infty, \text{ αν } g(\beta) < 0 \end{array}$$

Επομένως, η ελαχιστοποίηση του $L^*(\beta)$ είναι ισοδύναμη με την επίλυση του προβλήματος (3.12). Το πρόβλημα:

$$minimize_\beta L^*(\beta) \tag{3.15}$$

λέγεται *αρχικό πρόβλημα* και η $f(\beta)$ *αρχική αντικειμενική συνάρτηση.*

Για $\lambda \geq 0$, η *δυϊκή αντικειμενική συνάρτηση* ορίζεται ως εξής:

$$L_*(\lambda) = \inf_\beta L(\beta, \lambda) \tag{3.16}$$

και το *δυϊκό πρόβλημα* θα είναι το:

$$maximize_{\lambda \geq 0} L_*(\lambda) \tag{3.17}$$

Αν $\lambda \geq 0$, η $L(\beta, \lambda)$ είναι μια κυρτή συνάρτηση ως προς β και το $L(\beta, \lambda) \to +\infty$ καθώς το $||\beta||_1 \to +\infty$. Άρα, η $L(\beta, \lambda)$ έχει τουλάχιστον ένα ελάχιστο ως προς β και ισχύει ότι το $\bar{\beta}$ ελαχιστοποιεί την $L(\beta, \lambda)$ αν και μόνο αν το $p-$διάστατο μηδενικό διάνυσμα $\mathbf{0}$ είναι στοιχείο της μερικής παραγώγου $\partial_\beta L(\beta, \lambda)$ (*Osborne*, 1985). Η μερική παράγωγος δίνεται από τη σχέση:

$$\partial_\beta L(\beta, \lambda) = -X'r + \lambda v$$

όπου οι συνιστώσες του διανύσματος $v = (v_1, ..., v_p)$ είναι της μορφής $v_i = 1$ αν $\beta_i > 0$, $v_i = -1$ αν $\beta_i < 0$ και $v_i \in [-1, 1]$ αν $\beta_i = 0$. Οπότε, αν το $\bar{\beta}$ ελαχιστοποιεί το $L(\beta, \lambda)$, ισχύει ότι:

$$0 = -X'\bar{r} + \lambda v_1 \tag{3.18}$$

για κάποιο v_1 της παραπάνω μορφής, ενώ $\bar{r} = y - X\bar{\beta}$.

Για κάθε διάνυσμα v, ισχύει ότι $v'\bar{\beta} = ||\bar{\beta}||_1$, οπότε η σχέση (3.18) μας δίνει $\lambda = \bar{r}'X\bar{\beta}/||\bar{\beta}||_1$. Αν $\bar{\beta} \neq 0$, τότε ισχύει ότι $||v||_\infty = 1$, οπότε προκύπτει ότι $\lambda = ||X'\bar{r}||_\infty$. Χρησιμοποιώντας τις δύο αυτές εκφράσεις για το λ, έχουμε ότι:

$$\begin{aligned}
L_*(\lambda) = L(\bar{\beta}, \lambda) &= \frac{1}{2}\bar{r}'r - \frac{\bar{r}'X\bar{\beta}}{||\bar{\beta}||_1}(t - ||\bar{\beta}||_1) \\
&= \frac{1}{2}\bar{r}'r - \frac{\bar{r}'X\bar{\beta}}{||\bar{\beta}||_1}t + \bar{r}'X\bar{\beta} \\
&= \frac{1}{2}\bar{r}'(\bar{r} + X\bar{\beta}) + \frac{1}{2}\bar{r}'(X\bar{\beta}) - \frac{\bar{r}'X\bar{\beta}}{||\bar{\beta}||_1}t \\
&= \frac{1}{2}\bar{r}'(y + X\bar{\beta}) - \frac{\bar{r}'X\bar{\beta}}{||\bar{\beta}||_1}t \\
&= \frac{1}{2}(y - X\bar{\beta})'(y + X\bar{\beta}) - \frac{\bar{r}'X\bar{\beta}}{||\bar{\beta}||_1}t \\
&= \frac{1}{2}y'y - \frac{1}{2}\bar{\beta}'A\bar{\beta} - \frac{\bar{r}'X\bar{\beta}}{||\bar{\beta}||_1}t \\
&= \frac{1}{2}y'y - \frac{1}{2}\bar{\beta}'A\bar{\beta} - ||X'\bar{r}||_\infty t
\end{aligned}$$

Πολλές φορές, για να λύσουμε προβλήματα βελτιστοποίησης, χρειαζόμαστε τη σχέση μεταξύ αρχικού και δυϊκού προβλήματος. Τα δύο παρακάτω θεωρήματα αφορούν τη σχέση μεταξύ του αρχικού (3.12) και του δυϊκού (3.17) προβλήματος. Το πρώτο προκύπτει απ' ευθείας από τους ορισμούς των L^* και L_*:

Θεώρημα 2 *(Ασθενής δυϊκότητα)*
Αν β^* είναι μια λύση του (3.10) και $\bar{\lambda}$ είναι μια λύση του δυϊκού προβλήματος (3.17), τότε ισχύει:

$$L_*(\bar{\lambda}) \leq L^*(\beta^*)\square$$

Με τη βοήθεια του *Osborne* (1985) αποδεικνύεται το επόμενο θεώρημα. Ορίζουμε τη συνάρτηση $v(z) = \inf_{\beta \in \{\beta: g(\beta) \geq z\}} f(\beta)$ και χρησιμοποιώντας την κυρτότητά της, το λήμμα 1.6.2 και το θεώρημα 1.6.2 του *Osborne* (1985) καταλήγουμε στο παρακάτω συμπέρασμα.

Θεώρημα 3 *(Ισχυρή δυϊκότητα)*
Αν β^* είναι μια λύση του (3.10) και λ^* είναι ο *Lagrangian* πολλαπλασιαστής που αντιστοιχεί στο β^*, τότε το λ^* είναι μια λύση του δυϊκού προβλήματος (3.17) και ισχύει:

$$L_*(\lambda^*) = L(\beta^*, \lambda^*)\square$$

Στη συνέχεια, θα ασχοληθούμε με την περίπτωση όπου $p > N$. Αν β^* είναι μια λύση του (3.10), ορίζουμε ως $V(\beta^*)$ το σύνολο όλων των διανυσμάτων $\boldsymbol{\varepsilon}$ των οποίων οι συνιστώσες είναι της μορφής: $e_i = 1$ αν $\beta_i^* > 0$, $e_i = -1$ αν $\beta_i^* < 0$ και $e_i = 1$ ή $e_i = -1$

αν $\beta_i^* = 0$. Αν το β^* έχει $l < p$ συνιστώσες ίσες με 0, τότε το σύνολο $V(\beta^*) = \{\boldsymbol{\varepsilon}_1, ..., \boldsymbol{\varepsilon}_k\}$ περιέχει $k = 2^l$ διανύσματα. Έστω E ο $k \times p$ πίνακας, του οποίου η $i-$οστή γραμμή είναι η e_i'. Παρατηρούμε ότι $||\boldsymbol{\varepsilon}||_1 = p$ και $\beta^{*\prime}\boldsymbol{\varepsilon} = t < t_0$, για κάθε $\boldsymbol{\varepsilon} \in V(\beta^*)$. Τότε, ισχύει η σχέση:

$$(\beta^* - \frac{t}{p}\boldsymbol{\varepsilon})'e = 0, \forall \boldsymbol{\varepsilon} \in V(\beta^*)$$

Υποθέτουμε τώρα ότι $p > N$ και ότι το $\beta^\dagger$ είναι επίσης μια λύση του (3.10). Έστω η η διαφορά τους, δηλαδή $\eta = \beta^\dagger - \beta^*$. Από το θεώρημα 1, έχουμε ότι και ο κυρτός συνδυασμός τους $\rho\beta^\dagger + (1-\rho)\beta^* = \rho(\beta^\dagger - \beta^*) + \beta^* = \beta^* + \rho\eta$ είναι επίσης λύση για $0 \leq \rho \leq 1$ και μάλιστα $||\beta^* + \rho\eta||_1 = t$. Όμως, το $||y - X\beta||_2$ είναι σταθερό για κάθε λύση, οπότε θα ισχύει ότι $\eta \in N(X)$. Επίσης, από τη συνθήκη $||\beta^* + \rho\eta||_1 = t$ προκύπτει ότι $\eta'\boldsymbol{\varepsilon} \leq 0$ για κάθε $\boldsymbol{\varepsilon} \in V(\beta^*)$. Για να το αποδείξουμε αυτό, παρατηρούμε αρχικά ότι $||b||_1 \geq b'w$ για κάθε $b, w \in \Re^p$ με $||w||_\infty \leq 1$. Οπότε, αν $\eta'\boldsymbol{\varepsilon} > 0$ για κάποιο $\boldsymbol{\varepsilon} \in V(\beta^*)$, τότε $||\beta^* + \rho\eta||_1 \geq (\beta^* + \rho\eta)'\boldsymbol{\varepsilon} = t + \rho\eta'\boldsymbol{\varepsilon} > t$, άτοπο.

Αν τώρα συμβολίσουμε με $C(\beta^*)$ τον κυρτό κώνο που ορίζεται ως εξής:

$$C(\beta^*) = \{x \in \Re^p : x'\boldsymbol{\varepsilon} \leq 0, \forall \boldsymbol{\varepsilon} \in V(\beta^*)\}$$

τότε τα παραπάνω συνοψίζονται στο παρακάτω θεώρημα.

Θεώρημα 4

Το β^* είναι η μοναδική λύση του προβλήματος (3.10) αν και μόνο αν $C(\beta^*) \cap N(X) = \{\mathbf{0}\}$, όπου $\mathbf{0}$ το $p-$διάστατο μηδενικό διάνυσμα.□

Από το θεώρημα αυτό προκύπτει ότι αν $p > N$ και το β^* δεν είναι μοναδική λύση, τότε θα υπάρχει ένα μη μηδενικό διάνυσμα $\eta \in C(\beta^*) \cap N(X)$, δηλαδή ισχύει ότι $\eta \in N(X)$ και $\eta'\boldsymbol{\varepsilon} \leq 0$ για κάθε $\boldsymbol{\varepsilon} \in V(\beta^*)$. Αν κινηθούμε προς την κατεύθυνση όπου $\eta'\boldsymbol{\varepsilon} = 0$ για κάθε $\boldsymbol{\varepsilon} \in V(\beta^*)$., τότε θα φτάσουμε σε μια λύση η οποία θα έχει τουλάχιστον μία λιγότερη μη αρνητική συνιστώσα απ'οτι το β^*. Αν όμως το η είναι τέτοιο ώστε $\eta \in N(X)$ και $\eta'\boldsymbol{\varepsilon} > 0$ για τουλάχιστον ένα $\boldsymbol{\varepsilon} \in V(\beta^*)$, τότε το πλήθος των μη αρνητικών συνιστωσών της λύσης β^* αυξάνεται.

Για να αποδείξουμε τώρα το παρακάτω θεώρημα που αφορά το μέγιστο δυνατό πλήθος των μη-μηδενικών συνιστωσών μιας λύσης, παραθέτουμε το παρακάτω λήμμα.

Λήμμα 1

Η τάξη του πίνακα E είναι ίση με $l + 1$, δηλαδή το $V(\beta^*)$ περιέχει $l + 1$ γραμμικώς ανεξάρτητα διανύσματα. Επιπλέον, καμία γραμμή του E δεν είναι θετικός γραμμικός συνδυασμός των υπολοίπων γραμμών, ούτε το μηδενικό διάνυσμα είναι θετικός γραμμικός συνδυασμός των γραμμών του E.

Απόδειξη:

Έστω $I = \{i_1, ..., i_l\}$, όπου $\beta_{i_j}^* = 0$, για $j = 1, ..., l$. Για $k = 1, ..., l$, θεωρούμε τα διανύσματα $\boldsymbol{\varepsilon}_k \in V(\beta^*)$, τα οποία έχουν στοιχεία $e_{i_k,k} = -1$ και $e_{i_j,k} = 1$ για

$j = 1, ..., k-1, k+1, ..., l$ και έστω ότι το $\bar{e}_{l+1} \in V(\beta^*)$ έχει στοιχεία $e_{i_j,k} = 1$ για $j = 1, ..., l$. Είναι φανερό ότι αυτά τα $l+1$ διανύσματα είναι γραμμικώς ανεξάρτητα και ότι κάθε άλλο διάνυσμα που ανήκει στο $V(\beta^*)$ μπορεί να γραφτεί ως γραμμικός συνδυασμός αυτών των διανυσμάτων. Αυτό αποδεικνύει ότι ο Ε έχει τάξη $l+1$. Υποθέτουμε τώρα ότι υπάρχουν θετικοί αριθμοί λ_j και δείκτες k_j, με $j = 1, ..., q$, τέτοιοι ώστε το $\sum_{j=1}^{q} \lambda_j \boldsymbol{\varepsilon}_{k_j}$ είναι ίσο είτε με το $\boldsymbol{\varepsilon}_k$ για κάποιο k που δεν είναι κάποιο από τα $k_1, ..., k_q$, είτε με το μηδενικό διάνυσμα. Αφού $||\beta^*||_1 = t$, θα υπάρχει ένα τουλάχιστον i_0 με $\beta^*_{i_0} \neq 0$ και οι i_0-οστές συνιστώσες όλων των $\boldsymbol{\varepsilon} \in V(\beta^*)$ είναι είτε όλες ίσες με 1 είτε όλες ίσες με -1. Άρα, το $\sum_{j=1}^{q} \lambda_j$ είναι ίσο με 1 αν $\sum_{j=1}^{q} \lambda_j \boldsymbol{\varepsilon}_{k_j} = \boldsymbol{\varepsilon}_k$ ή 0 αν $\sum_{j=1}^{q} \lambda_j \boldsymbol{\varepsilon}_{k_j} = 0$. Η δεύτερη περίπτωση καταλήγει σε άτοπο. Για την πρώτη περίπτωση, αφού όλες οι συνιστώσες των $\boldsymbol{\varepsilon}_k$ και $\boldsymbol{\varepsilon}_{k_j}$, $j = 1, ..., q$, έχουν απόλυτη τιμή ίση με 1, είναι φανερό ότι πρέπει να ισχύει $e_{k_j,i} = e_{ki}$ για κάθε $j = 1, ..., q$ και $i = 1, ..., p$. Αυτό είναι άτοπο, αφού τα $\boldsymbol{\varepsilon}$ που ανήκουν στο $V(\beta^*)$ πρέπει να είναι όλα διαφορετικά μεταξύ τους. □

Θεώρημα 5

Αν $p > N$ και το β^* είναι κανονική λύση του (3.10), δηλαδή ισχύει $N(E) \cap N(X) = \{0\}$, τότε το β^* έχει το πολύ N μη μηδενικές συνιστώσες.

Απόδειξη:

Έστω β^* μια τέτοια λύση με l μηδενικές και $p-l$ μη μηδενικές συνιστώσες. Τότε από το Λήμμα 1, ο $N(E)$ είναι ένας $(p-l-1)-$διάστατος χώρος στον $\Re^p$ και $N(E) \cap N(X) = \{0\}$ εξ' ορισμού. Αφού όμως ο X έχει μέγιστη τάξη, ο $N(X)$ θα έχει διάσταση $p-N$. Θεωρούμε τώρα το μονοδιάστατο χώρο $S = \{\eta : \eta = \lambda \bar{\boldsymbol{\varepsilon}}, \lambda \in \Re\}$, όπου το $\bar{\boldsymbol{\varepsilon}} = \frac{1}{k} \sum_{j=1}^{k} \boldsymbol{\varepsilon}_j$ έχει συνιστώσες $\bar{e}_i = 1$ αν $\beta^*_i > 0$, $\bar{e}_i = -1$ αν $\beta^*_i < 0$ και $\bar{e}_i = 0$ αν $\beta^*_i = 0$. Είναι φανερό ότι $\bar{\boldsymbol{\varepsilon}}' \boldsymbol{\varepsilon} = p - l > 0$ για κάθε $\boldsymbol{\varepsilon} \in V(\beta^*)$, έτσι ώστε $S \cap N(E) = \{0\}$. Ισχύει, όμως, επίσης ότι $S \cap N(X) = \{0\}$, γιατί διαφορετικά, για $\lambda > 0$ αρκετά μικρό, το $\beta^* - \lambda \bar{\boldsymbol{\varepsilon}}$ θα ήταν λύση του (3.10) με L^1-νόρμα μικρότερη από αυτήν του β^*, άτοπο λόγω του θεωρήματος 1. Αφού οι χώροι $S, N(E)$ και $N(X)$ έχουν συνολικά διάσταση το πολύ p, θα ισχύει ότι $(p-l-1) + (p-N) + 1 \leq p$, δηλαδή $p - l \leq N$. □

Στη συνέχεια, οι *Osborne, Presnell* και *Turlach* (2000) υπολογίζουν τον πίνακα συνδιασποράς μιας λύσης β^* του (3.10) και καταλήγουν σε διαφορετικό τύπο από αυτόν του *Tibshirani* (1996). Η λύση β^* θα ικανοποιεί τη σχέση (3.18). Αφού το $X'r$ δεν εξαρτάται από τη συγκεκριμένη λύση β^*, το ίδιο θα ισχύει και για το $v = X'r/||X'r||_\infty$ στην (3.18). Επίσης, γνωρίζουμε ότι $\lambda = r'X\beta^*/||\beta^*||_1$. Η (3.18) λοιπόν γίνεται: $X'(y - X\beta^*) = \lambda v \Leftrightarrow X'y = A\beta^* + \lambda v \Leftrightarrow X'y = A\beta^* + \frac{r'X\beta^*}{||\beta^*||_1} \frac{X'r}{||X'r||_\infty} \beta^*$. Οπότε, τελικά προκύπτει η σχέση:

$$X'y = (A + W)\beta^* \tag{3.19}$$

όπου ο πίνακας $W = \frac{1}{||\beta^*||_1 ||X'r||_\infty} (X'r)(X'r)'$ έχει τάξη l. Επίσης, μπορούμε να γράψουμε:

$$A + W = X'(I + \frac{1}{||\beta^*||_1 ||X'r||_\infty} rr')X$$

Αυτό δείχνει ότι η τάξη του πίνακα $A+W$ είναι ίση με την τάξη του X και άρα ίση με την τάξη του A. Άρα, αν $p \leq N$, ο πίνακας συνδιασποράς των εκτιμητών δίνεται από τον τύπο:

$$Var(\beta^*) = (A+W)^{-1}A(A+W)^{-1}\hat{\sigma}^2 \tag{3.20}$$

όπου $\hat{\sigma}^2$ είναι ένας εκτιμητής της διασποράς των σφαλμάτων. Η σχέση αυτή έρχεται σε αντίθεση με την αντίστοιχη σχέση (3.6) του *Tibshirani* (1996).

Οι παραπάνω συγγραφείς, μας δίνουν έναν τρόπο εύρεσης του γ στη σχέση (3.3), η οποία μας δίνει τη λύση (3.10) στην περίπτωση που ο X είναι ορθοκανονικός, δηλαδή όταν ισχύει ότι $X'X = A = I$. Παρατηρούμε ότι:

$$\begin{aligned} t_0 - t &= \sum_{j=1}^{p} |\hat{\beta}_j^0| - \sum_{j=1}^{p} |\hat{\beta}_j| = \sum_{j=1}^{p} \{|\hat{\beta}_j^0| - (|\hat{\beta}_j^0| - \gamma)^+\} \\ &= \sum_{j=1}^{p} |\hat{\beta}_j^0| I(|\hat{\beta}_j^0| \leq \gamma) + \gamma \sum_{j=1}^{p} I(|\hat{\beta}_j^0| > \gamma) \\ &= \sum_{j=1}^{K} b_i + \gamma(p-K) \end{aligned}$$

όπου τα $b_1 \leq \ldots \leq b_p$ είναι τα $|\hat{\beta}_1^0|, \ldots, |\hat{\beta}_1^p|$ σε αύξουσα σειρά και $K = \max\{i : b_i \leq \gamma\}$. Αφού $t < t_0$, θα ισχύει ότι $K < p$ και $b_K \leq \gamma \leq b_{K+1}$. Έστω $c_0 = 0$ και $c_j = \sum_{i=1}^{j} b_i + b_j(p-j)$, για $j = 1, ..., p$ ώστε $0 = c_0 \leq c_1 \leq \ldots \leq c_p = t_0$. Τότε, $K = max\{i : c_i \leq t_0 - t\}$ και

$$\gamma = \frac{(t_0 - t) - \sum_{i=1}^{K} b_i}{p - K}$$

Κεφάλαιο 4

ΠΑΛΙΝΔΡΟΜΗΣΗ ΕΛΑΧΙΣΤΗΣ ΓΩΝΙΑΣ

Στο παρόν κεφάλαιο, αναλύεται μια νέα μέθοδος κατασκευής γραμμικών μοντέλων, η Παλινδρόμηση Ελάχιστης Γωνίας ($LARS$), σύμφωνα με το σχετικό άρθρο των $B.Efron$, $T.Hastie$, $I.Johnstone$, $R.Tibshirani$ (2004). Αρχικά, περιγράφεται ο αλγόριθμος της μεθόδου και έπειτα αποδεικνύεται ότι μέσω μιας τροποποίησης, η μέθοδος $LARS$ μπορεί να ταυτιστεί με τη μέθοδο *Stagewise* ή τη μέθοδο *Lasso* με την έννοια ότι δίνουν τις ίδιες λύσεις για τους εκτιμώμενους συντελεστές των επεξηγηματικών μεταβλητών του μοντέλου. Στη συνέχεια, προσεγγίζονται κάτω από ορισμένες προϋποθέσεις οι βαθμοί ελευθερίας του εκτιμητή της εξαρτημένης μεταβλητής και τέλος, υπολογίζεται το κόστος υπολογισμού για τις τρεις μεθόδους.

Ο στόχος των αλγορίθμων επιλογής μοντέλου, όπως η *All subsets*, η *Forward selection* και η *Backward elimination*, είναι να επιλέξουμε ένα γραμμικό μοντέλο με σκοπό την πρόβλεψη της τιμής μιας εξαρτημένης μεταβλητής με βάση τις τιμές των επεξηγηματικών μεταβλητών $x_1, x_2, ..., x_m$. Βασικά κριτήρια επιλογής είναι η ακρίβεια της πρόβλεψης αλλά και το κόστος υπολογισμού αφού οι ερευνητές προτιμούν τα πιο απλά μοντέλα για να έχουν στη διάθεσή τους μια σχέση ανάμεσα στην y και στις $x_1, x_2, ..., x_m$. Δύο πρόσφατοι αλγόριθμοι κατασκευής μοντέλων είναι η *Lasso*, που εξετάστηκε αναλυτικά στο προηγούμενο κεφάλαιο, και η *Forward Stagewise* γραμμική παλινδρόμηση. Αυτές οι δύο χρησιμοποιήθηκαν ώστε να δημιουργηθεί μια υπολογιστικά απλούστερη μέθοδος, η **Παλινδρόμηση Ελάχιστης Γωνίας** (*Least Angle Regression*) ή $LARS$.

Ο πίνακας 4.1 δείχνει ένα μέρος των αποτελεσμάτων μιας έρευνας για ασθενείς που πάσχουν από διαβήτη ($B.Efron$, $T.Hastie$, $I.Johnstone$, $R.Tibshirani$ (2004)). Δέκα είναι οι βασικές μεταβλητές που μετρήθηκαν για καθέναν από τους $n = 442$ διαβητικούς ασθενείς, ενώ η εξαρτημένη μεταβλητή y χαρακτηρίζει την εξέλιξη της ασθένειας. Οι στατιστικοί κλήθηκαν να κατασκευάσουν ένα μοντέλο που να προβλέπει την τιμή της y με βάση δοσμένες τιμές των μεταβλητών $x_1, x_2, ..., x_{10}$, έτσι ώστε να έχουν ακριβείς προβλέψεις για μέλλοντικούς ασθενείς όπως επίσης και να γνωρίζουν ποιές από τις επεξηγηματικές μεταβλητές είναι σημαντικοί παράγοντες για την εξέλιξη της ασθένειας.

Η Μέθοδος *Lasso*, όπως ξέρουμε ήδη, είναι μια παραλλαγή της μεθόδου των ελαχίστων τετραγώνων (OLS) κάτω από ένα περιορισμό. Έστω $x_1, x_2, ..., x_m$ τα διανύσματα μήκους n με συνιστώσες τις παρατηρήσεις για κάθεμία από τις επεξηγηματικές μεταβλητές. Στο παράδειγμά μας, $m = 10$ και $n = 442$ ενώ y είναι το διάνυσμα των τιμών της εξαρτημένης

	AGE	SEX	BMI	BP	··· Serum Measurements ···						Response
Patient	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	y
1	54	1	24.2	106	181	121	44.0	3.21	2.04	91	53
2	50	1	33.9	92	278	196	44.5	5.42	2.47	89	124
3	63	1	32.4	78	237	154	48.5	4.00	2.42	94	167
4	57	2	25.7	83	181	103	67.5	2.00	1.79	85	50
5	55	1	27.7	96	195	120	59.5	2.00	2.03	97	52
6	53	1	31.8	102	200	104	58.5	2.00	2.47	98	205
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
441	38	2	23.2	89	162	97	51.5	2.00	1.95	80	141
442	36	2	26.5	110	232	159	40.5	5.00	2.39	88	184

Πίνακας 4.1: Έρευνα για ασθενείς που πάσχουν από διαβήτη [*B.Efron*, *T.Hastie*, *I.Johnstone*, *R.Tibshirani* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), $p.407-499$]

μεταβλητής για τις n περιπτώσεις.

Υποθέτουμε ότι οι επεξηγηματικές μεταβλητές έχουν κανονικοποιηθεί ώστε να ισχύει:

$$\sum_{i=1}^{n} y_i = 0, \sum_{i=1}^{n} x_{ij} = 0, \sum_{i=1}^{n} x_{ij}^2 = 1 \tag{4.1}$$

για $j = 1, 2, ..., m$.

Το υποψήφιο διάνυσμα των συντελεστών παλινδρόμησης $\hat{\beta} = (\hat{\beta_1}, \hat{\beta_2}, ..., \hat{\beta_m})'$ δίνει ως πρόβλεψη το διάνυσμα $\hat{\mu}$:

$$\hat{\mu} = \sum_{j=1}^{m} x_j \hat{\beta}_j = X\hat{\beta}, \tag{4.2}$$

όπου X ο $n \times m$ πίνακας $(x_1, x_2, ..., x_m)$. Το συνολικό τετραγωνικό σφάλμα του y τότε είναι:

$$S(\hat{\beta}) = ||y - \hat{\mu}||^2 = \sum_{i=1}^{n} (y_i - \hat{\mu_i})^2 \tag{4.3}$$

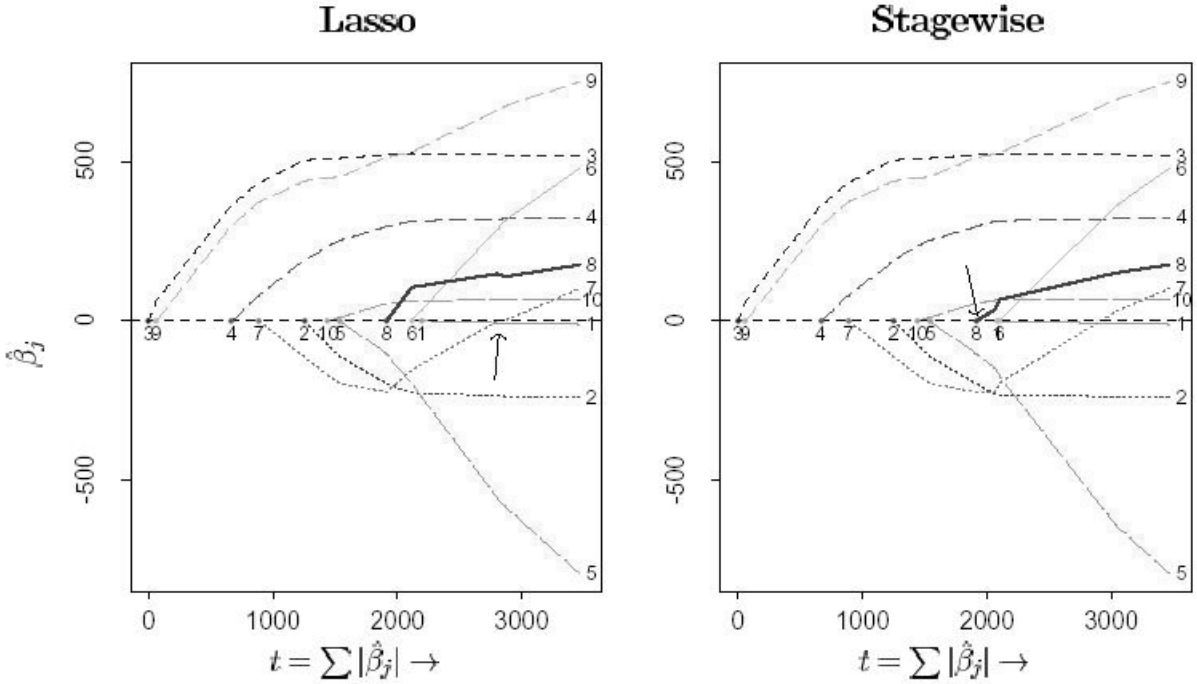
Έστω $T(\hat{\beta})$ η απόλυτη νόρμα του $\hat{\beta}$, δηλαδή:

$$T(\hat{\beta}) = \sum_{j=1}^{m} |\hat{\beta}_j| \tag{4.4}$$

Η *Lasso* διαλέγει το $\hat{\beta}$ αυτό που ελαχιστοποιεί το $S(\hat{\beta})$ και συγχρόνως η $T(\hat{\beta})$ να μην ξεπερνάει την τιμή t. Δηλαδή:

$$\begin{cases} minimize\ S(\hat{\beta}) \\ \\ \tau.\omega.\ T(\hat{\beta}) \leq t \end{cases} \tag{4.5}$$

Στο αριστερό γράφημα του σχήματος 4.1 φαίνονται όλες οι λύσεις $\hat{\beta}(t)$ της μεθόδου *Lasso* για το παράδειγμά μας, από το σημείο όπου $\hat{\beta} = 0$ μέχρι το $t = 3460$, όπου ο εκτιμητής $\hat{\beta}$ ισούται με τον OLS εκτιμητή. Παρατηρούμε ότι για κάθε $t < 3460$, κάποιοι συντελεστές θα είναι ίσοι με 0, ενώ οι υπόλοιποι θα είναι απολύτως μικρότεροι από την απολύτως μέγιστη τιμή τους, δηλαδή την τιμή των OLS εκτιμητών. Αν $t = 1000$ για παράδειγμα, μόνο οι μεταβλητές 3,9,4 και 7 μπαίνουν στο μοντέλο. Βλέπουμε λοιπόν ότι η *Lasso* τείνει να συρρικνώσει τους OLS συντελεστές προς το 0, γεγονός που μειώνει τη συνολική διασπορά του μοντέλου παλινδρόμησης, άρα βελτιώνει την ακρίβεια της πρόβλεψης. Όμως, αυξάνεται η μεροληψία του y.



Σχήμα 4.1: Εκτιμητές των Συντελεστών Παλινδρόμησης για *Lasso* και *Stagewise* [*B.Efron*, *T.Hastie*, *I.Johnstone*, *R.Tibshirani* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), *p.*407 – 499]

Η *Forward Stagewise* είναι μια επαναληπτική μέθοδος που ξεκινά με $\hat{\mu} = 0$ και κατασκευάζει τη συνάρτηση παλινδρόμησης σε διαδοχικά βήματα. Αν $\hat{\mu}$ είναι ο εκτιμητής της y και $c(\hat{\mu})$ το διάνυσμα των συσχετίσεων των επεξηγηματικών μεταβλητών με την εκτιμώμενη τιμή της y, τότε ισχύει:

$$\hat{c} = c(\hat{\mu}) = X'(y - \hat{\mu}) \tag{4.6}$$

άρα το $\hat{c_j}$ είναι ανάλογο της συσχέτισης μεταξύ της x_j και του σφάλματος της πρόβλεψης. Το επόμενο βήμα λαμβάνεται κατά τη διεύθυνση της μεγαλύτερης συσχέτισης.

$$\begin{cases} \hat{j} = \arg\max |\hat{c_j}| \\ \\ \hat{\mu} \to \hat{\mu} + \varepsilon sign(\hat{c_{\hat{j}}}) x_{\hat{j}} \end{cases} \tag{4.7}$$

όπου ε κάποια μικρή σταθερά. Η "μεγάλη" επιλογή $\varepsilon = |\hat{c_{\hat{j}}}|$ στη συνήθη *Forward selection* τεχνική μπορεί να προκαλέσει την απαλοιφή των μεταβλητών που είναι συσχετισμένες με τη $x_{\hat{j}}$. Το δεξί γράφημα του σχήματος 4.1 απεικονίζει τους συντελεστές που λαμβάνονται κατά τη *Stagewise* μέθοδο ως προς t για τα δεδομένα των διαβητικών ασθενών. Παρατηρούμε ότι τα δύο γραφήματα του σχήματος είναι παρόμοια, παρόλο που οι συντελεστές της *Lasso* και της *Stagewise* υπολογίζονται διαφορετικά.

Το βασικό σημείο του κεφαλαίου αυτού είναι ότι οι *Lasso* και *Stagewise* μέθοδοι είναι παραλλαγές μιας διαδικασίας, της Παλινδρόμησης Ελάχιστης Γωνίας (*LARS*) που περιγράφεται αμέσως παρακάτω.

4.1. Αλγόριθμος *LARS*

Η Παλινδρόμηση Ελάχιστης Γωνίας (*LARS*) είναι μια παραλλαγή της *Stagewise* διαδικασίας και χρησιμοποιεί μια απλή μαθηματική διαδικασία ώστε να επιταχύνει τους υπολογισμούς. Μόνο m βήματα απαιτούνται για να υπολογιστεί το πλήρες σύνολο των λύσεων, όπου m είναι το πλήθος των επεξηγηματικών μεταβλητών.

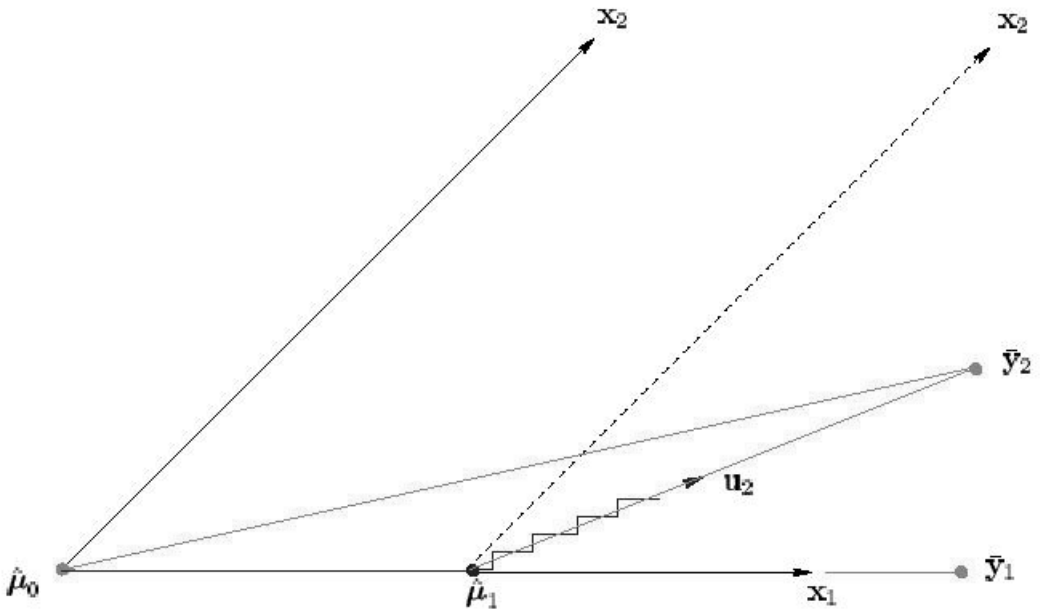
Η μέθοδος της *LARS* κατασκευάζει εκτιμητές $\hat{\mu} = X\hat{\beta}$, σε διαδοχικά βήματα και σε καθένα από αυτά προστίθεται μια μεταβλητή στο μοντέλο, έτσι ώστε μετά από k βήματα, μόνο k από τα $\hat{\beta_j}$ θα είναι μη-μηδενικά. Το Σχήμα 4.2 απεικονίζει τον αλγόριθμο στην περίπτωση όπου $m = 2$ μεταβλητές, δηλαδή $X = (x_1, x_2)$. Στην περίπτωση αυτή, οι συσχετίσεις (4.6) εξαρτώνται μόνο από την προβολή $\bar{y_2}$ του y πάνω στο γραμμικό χώρο $L(X)$ που δημιουργείται από τις x_1 και x_2, και δίνονται από τον τύπο:

$$c(\hat{\mu}) = X'(y - \hat{\mu}) = X'(\bar{y_2} - \hat{\mu}) \tag{4.8}$$

Ο αλγόριθμος ξεκινά με $\hat{\mu_0} = 0$. Στο Σχήμα 4.2 φαίνεται ότι η γωνία που σχηματίζει η $\bar{y_2} - \hat{\mu_0}$ με τη x_1 είναι μικρότερη από αυτήν που σχηματίζει με το x_2, δηλαδή $c_1(\hat{\mu_0}) > c_2(\hat{\mu_0})$. Η *LARS* τότε κατευθύνει το $\hat{\mu_0}$ κατά τη διεύθυνση του x_1, έστω:

$$\hat{\mu_1} = \hat{\mu_0} + \hat{\gamma_1} x_1 \tag{4.9}$$

Η *Stagewise* διαλέγει το $\hat{\gamma_1}$ ίσο με κάποια μικρή τιμή ε και επαναλαμβάνει τη διαδικασία πολλές φορές. Η παλιότερη μέθοδος *Forward selection* παίρνει τη $\hat{\gamma_1}$ αρκετά μεγάλη ώστε το $\hat{\mu_1}$ να γίνει ίσο με το $\bar{y_1}$, δηλαδή την προβολή του y στον $L(X_1)$. Η *LARS* όμως χρησιμοποιεί μια συγκεκριμένη τιμή για το $\hat{\gamma_1}$. Αυτή η τιμή θα είναι τέτοια ώστε η $\bar{y_2} - \hat{\mu}$



Σχήμα 4.2: Ο αλγόριθμος *LARS* στην περίπτωση όπου $m = 2$ [*B.Efron, T.Hastie, I.Johnstone, R.Tibshirani* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), $p.407 - 499$]

θα έχει συσχέτιση με τη x_1 ίση με αυτήν που έχει με τη x_2. Δηλαδή η $\bar{y_2} - \hat{\mu_1}$ διχοτομεί τη γωνία μεταξύ x_1 και x_2, ώστε $c_1(\hat{\mu_1}) = c_2(\hat{\mu_1})$.

Έστω u_2 το μοναδιαίο διάνυσμα πάνω στη διχοτόμο αυτή. Ο επόμενος *LARS* εκτιμητής θα είναι:

$$\hat{\mu_2} = \hat{\mu_1} + \hat{\gamma_2} u_2 \tag{4.10}$$

με το $\hat{\gamma_2}$ επιλεγμένο έτσι ώστε $\hat{\mu_2} = \bar{y_2}$ στην περίπτωση όπου $m = 2$. Για $m > 2$, το $\hat{\gamma_2}$ θα ήταν μικρότερο και θα οδηγούσε σε μια άλλη αλλαγή κατεύθυνσης, όπως φαίνεται στο Σχήμα 4.4.

Για $m > 2$, η μέθοδος *LARS* χρησιμοποιεί **ισογώνια διανύσματα**, (*equiangular vectors*) γενικεύοντας τη διχοτόμο u_2 του Σχήματος 4.2. Υποθέτουμε ότι τα διανύσματα $x_1, x_2, ..., x_m$ είναι γραμμικώς ανεξάρτητα. Αν A ένα υποσύνολο των δεικτών $\{1, 2, ..., m\}$, ορίζουμε τον πίνακα:

$$X_A = (.....s_j x_j.....)_{j \epsilon A} \tag{4.11}$$

όπου τα πρόσημα s_j ισούνται με ± 1. Έστω:

$$\begin{cases} G_A = X_A' X_A \\ \\ H_A = (1_A' G_A^{-1} 1_A)^{-1/2} \end{cases} \tag{4.12}$$

όπου 1_A είναι ένα μοναδιαίο διάνυσμα μήκους $|A|$, όπου $|A|$ το πλήθος των στοιχείων του A. Το ισογώνιο διάνυσμα:

$$u_A = X_A w_A, w_A = H_A G_A^{-1} 1_A \tag{4.13}$$

είναι το μοναδιαίο διάνυσμα που δημιουργεί ίσες οξείες γωνίες με τα διανύσματα που αντιστοιχούν στις στήλες του X_A, δηλαδή είναι τέτοιο ώστε:

$$X_A' u_A = H_A 1_A, ||u_A||^2 = 1 \tag{4.14}$$

Στη συνέχεια, περιγράφουμε πιο αναλυτικά τον αλγόριθμο $LARS$. Όπως και στη $Stagewise$ διαδικασία, ξεκινάμε με $\hat{\mu}_0 = 0$ και σε κάθε βήμα υπολογίζουμε το $\hat{\mu}$. Έστω ότι $\hat{\mu}_A$ είναι ο τρέχων $LARS$ εκτιμητής και ότι το:

$$\hat{c} = X'(y - \hat{\mu}_A) \tag{4.15}$$

είναι το διάνυσμα των συσχετίσεων (4.6). **Ενεργό σύνολο A** θα ονομάζουμε το σύνολο που περιέχει τους δείκτες που αντιστοιχούν στις μεταβλητές με τις απολύτως μεγαλύτερες συσχετίσεις:

$$\begin{cases} \hat{C} = \max_j\{|c_j|\} \\ \\ A = \{j : |\hat{c_j}| = \hat{C}\} \end{cases} \tag{4.16}$$

Έστω ότι:

$$s_j = sign\{\hat{c_j}\}, j \epsilon A \tag{4.17}$$

Υπολογίζουμε τότε τα X_A, H_A και u_A από τις σχέσεις (4.11)-(4.13) όπως επίσης και το εσωτερικό γινόμενο:

$$a \equiv X' u_A \tag{4.18}$$

Το επόμενο βήμα του αλγορίθμου $LARS$ τροποποιεί το $\hat{\mu_A}$ ως εξής:

$$\hat{\mu}_{A_+} = \hat{\mu}_A + \hat{\gamma} u_A \tag{4.19}$$

όπου

$$\hat{\gamma} = \min_{j \epsilon A^c}{}^+ \{\frac{\hat{C} - \hat{c_j}}{H_A - a_j}, \frac{\hat{C} + \hat{c_j}}{H_A + a_j}\} \tag{4.20}$$

Το "$\min^+$" υποδεικνύει ότι το ελάχιστο λαμβάνεται μόνο μεταξύ των θετικών συνιστωσών για κάθε επιλογή του j.

Οι τύποι (4.19) και (4.20) έχουν την ακόλουθη ερμηνεία. Έστω:

$$\mu(\gamma) = \hat{\mu}_A + \gamma u_A \tag{4.21}$$

με $\gamma > 0$. Τότε, η αντίστοιχη συσχέτιση $c_j(\gamma)$ δίνεται από τον τύπο:

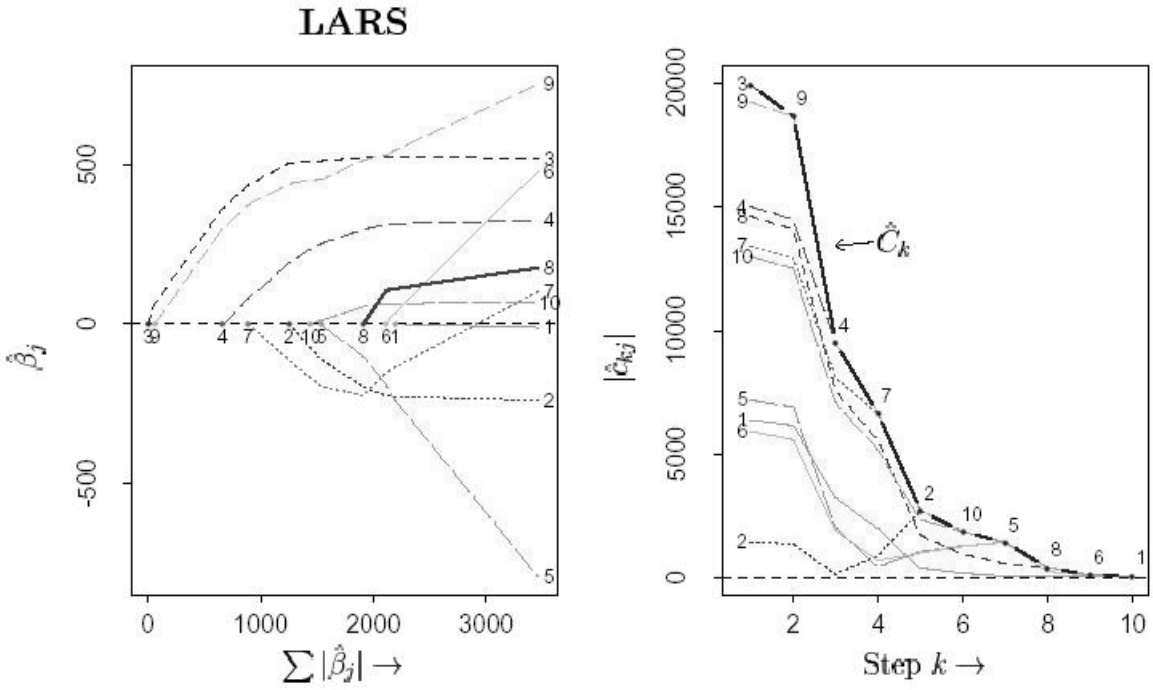
$$c_j(\gamma) = x_j'(y - \mu(\gamma)) = \hat{c}_j - \gamma a_j \tag{4.22}$$

Για $j \epsilon A$, οι (4.14)-(4.16) δίνουν:

$$|c_j(\gamma)| = \hat{C} - \gamma H_A \tag{4.23}$$

ώστε όλες οι μέγιστες τρέχουσες συσχετίσεις να γίνουν ίσες. Για $j \epsilon A^c$, εξισώνοντας τις (4.22) και (4.23), προκύπτει ότι η $c_j(\gamma)$ παίρνει τη μέγιστη τιμή όταν $\gamma = \frac{\hat{C} - \hat{c}_j}{H_A - a_j}$. Όμοια, η $-c_j(\gamma)$ γίνεται μέγιστη τιμή όταν $\gamma = \frac{\hat{C} + \hat{c}_j}{H_A + a_j}$. Επομένως, το $\hat{\gamma}$ στην (4.20) είναι η μικρότερη θετική τιμή του γ ώστε ο νέος δείκτης $\hat{j}$ να εισέλθει στο ενεργό σύνολο. Τότε το νέο ενεργό σύνολο θα είναι το $A_+ = A \bigcup \{\hat{j}\}$ και η νέα απολύτως μεγαλύτερη συσχέτιση θα είναι η $\hat{C}_+ = \hat{C} - \hat{\gamma} H_A$.

Το Σχήμα 4.3 απεικονίζει τη διαδικασία της μεθόδου *LARS* για το παράδειγμά μας. Ο πλήρης αλγόριθμος χρειάστηκε μόνο $m = 10$ βήματα, με τις μεταβλητές να εισέρχονται στο ενεργό σύνολο A κατά την ίδια σειρά με τη *Lasso*: 3,9,4,7,...,1. Οι συντελεστές $\hat{\beta}_j$ μοιάζουν πολύ αλλά δεν είναι ίδιοι με αυτούς των *Lasso* και *Stagewise* στο Σχήμα 4.1.



Σχήμα 4.3: Η διαδικασία του αλγορίθμου *LARS* για τα δεδομένα μας [*B.Efron, T.Hastie, I.Johnstone, R.Tibshirani* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), *p.*407 − 499]

Το δεξί γράφημα του Σχήματος 4.3 απεικονίζει τις απόλυτες συσχετίσεις:

$$|\hat{c}_{kj}| = |x_j'(y - \hat{\mu}_{k-1})| \tag{4.24}$$

για τις μεταβλητές $j = 1, 2, ..., 10$ ως συνάρτηση του $k-$οστού βήματος του αλγορίθμου $LARS$. Η μέγιστη συσχέτιση:

$$\hat{C}_k = \max\{|\hat{c}_{kj}|\} = \hat{C}_{k-1} - \hat{\gamma}_{k-1}H_{k-1} \tag{4.25}$$

μειώνεται καθώς αυξάνεται το k. Σε κάθε βήμα, για τη νέα μεταβλητή j που εισέρχεται στο ενεργό σύνολο θα ισχύει $|\hat{c}_{kj}| = \hat{C}_k$. Τα πρόσημα s_j των αντιστοίχων x_j στην (4.11) παραμένουν τα ίδια όσο στο ενεργό σύνολο εισέρχονται κι άλλοι δείκτες.

Υποθέτουμε τώρα ότι ο αλγόριθμος $LARS$ έχει συμπληρώσει τα πρώτα $k-1$ βήματα, δίνοντας ως εκτιμητή το $\hat{\mu}_{k-1}$, και προχωράει στο $k-$οστό βήμα. Το ενεργό σύνολο A_k θα έχει k στοιχεία τώρα και θα αντιστοιχούν σε αυτό συγκεκριμένες τιμές για τα X_k, G_k, H_k και u_k οι οποίες υπολογίζονται από τις σχέσεις (4.11)-(4.13). Έστω $\bar{y}_k$ η προβολή του y στο χώρο $L(x_k)$. Αφού $\hat{\mu}_{k-1}\epsilon L(x_{k-1})$ θα ισχύει ότι:

$$\bar{y}_k = \hat{\mu}_{k-1} + X_k G_k^{-1} X_k'(y - \hat{\mu}_{k-1}) = \hat{\mu}_{k-1} + \frac{\hat{C}_k}{H_k} u_k \tag{4.26}$$

Η τελευταία ισότητα συνεπάγεται από την (4.13) και από το γεγονός ότι οι συσχετίσεις των μεταβλητών των οποίων οι δείκτες ανήκουν στο A_k είναι όλες ίσες με $\hat{C}_k$, δηλαδή:

$$X_k'(y - \hat{\mu}_{k-1}) = \hat{C}_k 1_A \tag{4.27}$$

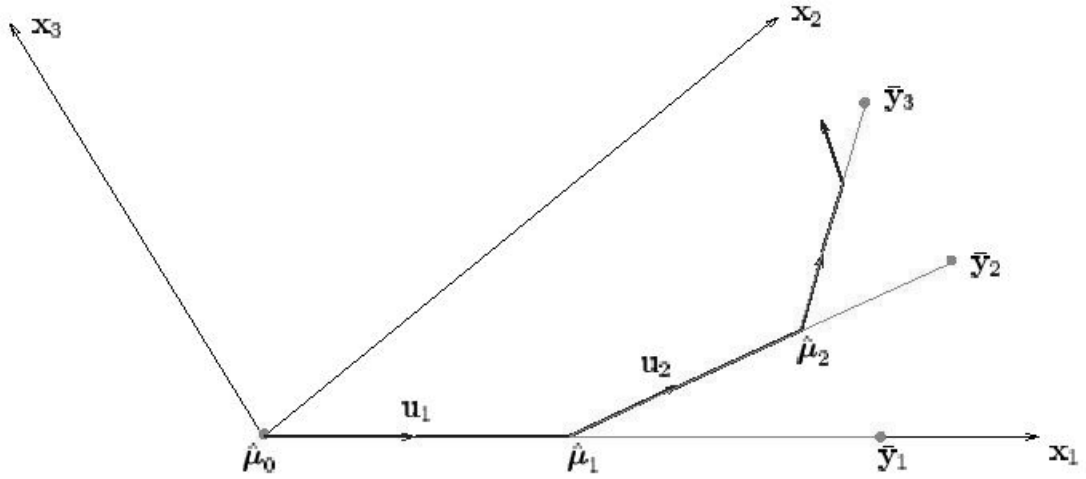
Αφού το u_k είναι μοναδιαίο διάνυσμα, η (4.26) δείχνει ότι η $\bar{y}_k - \hat{\mu}_{k-1}$ έχει μήκος:

$$\bar{\gamma_k} \equiv \frac{\hat{C}_k}{H_k} \tag{4.28}$$

Γράφοντας τη σχέση (4.19) ως $\hat{\mu}_k = \hat{\mu}_{k-1} + \hat{\gamma}_k u_k$ και συνδυάζοντας την με την (4.26) με τη βοήθεια της (4.28), προκύπτει ότι ο εκτιμητής $\hat{\mu}_k$ βρίσκεται στην ευθεία που ορίζουν τα $\hat{\mu}_{k-1}$ και $\bar{y}_k$, δηλαδή:

$$\hat{\mu}_k - \hat{\mu}_{k-1} = \frac{\hat{\gamma}_k}{\bar{\gamma}_k}(\bar{y}_k - \hat{\mu}_{k-1}) \tag{4.29}$$

Είναι εύκολο να δούμε ότι η $\hat{\gamma}_k$ είναι πάντα μικρότερη από τη $\bar{\gamma}_k$, οπότε το $\hat{\mu}_k$ βρίσκεται πιο κοντά στο $\hat{\mu}_{k-1}$ παρά στο $\bar{y}_k$. Το Σχήμα 4.4 δείχνει τους διαδοχικούς $LARS$ εκτιμητές $\hat{\mu}_k$ πάντα να πλησιάζουν τους OLS εκτιμητές $\bar{y}_k$ αλλά ποτέ να μην τους φτάνουν. Η εξαίρεση είναι στο τελευταίο στάδιο: εφόσον το A_m περιέχει όλες τις μεταβλητές, η (4.20) δεν ορίζεται. Σε αυτό το σημείο, ο αλγόριθμος θεωρεί ότι $\hat{\gamma}_m = \bar{\gamma}_m = \hat{C}_m/H_m$, άρα $\hat{\mu}_m = \bar{y}_m$ και ο $\hat{\beta}_m$ είναι ίσος με τον OLS εκτιμητή για το πλήρες σύνολο των m μεταβλητών.



Σχήμα 4.4: Ο Αλγόριθμος *LARS* για μεγαλύτερες διαστάσεις [*B.Efron*, *T.Hastie*, *I.Johnstone*, *R.Tibshirani* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), $p.407-499$]

4.2. Τροποποιημένες Εκδοχές της Μεθόδου *LARS*

Τα σχήματα 4.1 και 4.3 έδειξαν ότι η *Lasso*, η *Stagewise* και η *LARS* έδωσαν παρόμοιους εκτιμητές. Αυτό δεν είναι συμπτωματικό αφού όπως θα δούμε αμέσως παρακάτω, απλές τροποποιήσεις του αλγορίθμου *LARS* δίνουν τους εκτιμητές της *Lasso* ή της *Stagewise*. Και οι τρεις μέθοδοι μπορούν να θεωρηθούν ως πολυβηματικές διαδικασίες, οι οποίες λαμβάνουν υπ'όψη τους τις μεταβλητές με τη μεγαλύτερη συσχέτιση. Η *LARS* κινείται προς την κατεύθυνση του ισογώνιου διανύσματος, ενώ οι *Lasso* και *Stagewise* θέτουν κάποιους περιορισμούς σε αυτή τη στρατηγική, όπως θα δούμε αμέσως παρακάτω.

4.2.1. Η σχέση μεταξύ *LARS* και *Lasso*

Οι λύσεις της μεθόδου *Lasso* μπορούν να βρεθούν αν τροποποιήσουμε τον αλγόριθμο *LARS* ως εξής: Έστω $\hat{\beta}$ η λύση της *Lasso*, που υπολογίζεται από το πρόβλημα (4.5) και δίνει ως εκτιμητή το διάνυσμα $\hat{\mu} = X\hat{\beta}$. Τότε, το πρόσημο κάθε μη μηδενικού συντελεστή $\hat{\beta}_j$ πρέπει να συμφωνεί με το πρόσημο s_j της τρέχουσας συσχέτισης $\hat{c}_j = x_j'(y-\hat{\mu})$, δηλαδή:

$$sign(\hat{\beta}_j) = sign(\hat{c}_j) = s_j \tag{4.30}$$

Αυτό εξηγείται ως εξής: Έστω S_A ο διαγώνιος πίνακας με στοιχεία τα s_j και A το τρέχονενεργό σύνολο. Τότε, η σχέση $\hat{\mu} = X\hat{\beta}$, λόγω των σχέσεων (4.11) και (4.17), γράφεται $\hat{\mu} = X_A S_A \hat{\beta}$. Γράφουμε το πρόβλημα ελαχιστοποίησης (4.5) χρησιμοποιώντας το *Lagrange* πολλαπλασιαστή:

$$minimize \frac{1}{2}||y - X_A S_A \hat{\beta}_A||^2 + \lambda \sum_A s_j \hat{\beta}_j$$

Παίρνοντας τότε το κριτήριο της πρώτης παραγώγου, προκύπτει τελικά η σχέση:

$$S_A X_A'(y - X_A S_A \hat{\beta}_A) = \lambda S_A 1_A$$

Το $j-$οστό στοιχείο όμως του αριστερού μέλους ισούται με $\hat{c}_j$, ενώ το αντίστοιχο του δεξιού μέλους ισούται με $\lambda sign(\hat{\beta}_j)$, για $j \in A$. Επίσης, $\lambda = |\hat{c}_j| = \hat{C}$, οπότε προκύπτει η ζητούμενη σχέση.

Ο αλγόριθμος $LARS$ δεν απαιτεί τον περιορισμό (4.30), αλλά, όπως θα δούμε παρακάτω, μπορεί εύκολα να τροποποιηθεί ώστε να τον απαιτεί. Έστω ότι έχουμε συμπληρώσει ένα $LARS$ βήμα, το οποίο μας έδωσε ως ενεργό σύνολο το A και ότι ο αντίστοιχος εκτιμητής $\hat{\mu}_A$ αντιστοιχεί στη λύση της $Lasso$ $\hat{\mu} = X\hat{\beta}$. Έστω

$$w_A = H_A G_A^{-1} 1_A \tag{4.31}$$

ένα διάνυσμα μήκους ίσου με το πλήθος των στοιχείων του A. Ορίζουμε ως $\hat{d}$ το διάνυσμα μήκους m που ισούται με $s_j w_{A_j}$ για $j \epsilon A$ ή με 0 διαφορετικά. Κινούμενοι προς την κατεύθυνση του θετικού αριθμού γ κατά τη διαδικασία (4.21) της μεθόδου $LARS$, παρατηρούμε ότι:

$$\begin{cases} \mu(\gamma) = X\beta(\gamma) \\ \\ \beta_j(\gamma) = \hat{\beta}_j + \gamma \hat{d}_j \end{cases} \tag{4.32}$$

για $j \epsilon A$. Επομένως, το $\beta_j(\gamma)$ αλλάζει πρόσημο όταν:

$$\gamma_j = -\hat{\beta}_j / \hat{d}_j, \tag{4.33}$$

και η πρώτη αλλαγή θα εμφανιστεί στο

$$\tilde{\gamma} = \min_{\gamma_j > 0} \{\gamma_j\} \tag{4.34}$$

έστω για τη μεταβλητή $x_{\tilde{j}}$. Το $\tilde{\gamma}$ ισούται με το άπειρο εξ' ορισμού αν δεν υπάρχει j τέτοιο ώστε $\gamma_j > 0$.

Αν $\tilde{\gamma} < \hat{\gamma}$, όπου το $\hat{\gamma}$ δίνεται από την (4.20), τότε ο $\beta_j(\gamma)$ δεν μπορεί να είναι λύση της $Lasso$ αν $\gamma > \tilde{\gamma}$ αφού ο περιορισμός (4.30) παραβιάζεται επειδή ο $\beta_{\tilde{j}}(\gamma)$ αλλάζει πρόσημο ενώ το $c_{\tilde{j}}(\gamma)$ δεν αλλάζει. (Η συνεχής συνάρτηση $c_{\tilde{j}}(\gamma)$ δεν μπορεί να αλλάξει πρόσημο σε ένα $LARS$ βήμα, αφού $|c_{\tilde{j}}(\gamma)| = \hat{C} - \gamma H_A > 0$).

Lasso τροποποίηση: Αν $\tilde{\gamma} < \hat{\gamma}$, σταμάτα το επόμενο βήμα της μεθόδου $LARS$ στο σημείο όπου $\tilde{\gamma} = \hat{\gamma}$ και απάλειψε το $\tilde{j}$ από τον υπολογισμό του επόμενου ισογώνιου διανύσματος. Δηλαδή:

$$\hat{\mu}_{A_+} = \hat{\mu}_A + \tilde{\gamma} u_A, A_+ = A - \{\tilde{j}\}, \tag{4.35}$$

αντί για το (4.19).

Θεώρημα 1:
Με την προσθήκη της *Lasso* τροποποίησης, και υποθέτοντας την "ένα τη φορά" συνθήκη που περιγράφεται παρακάτω, ο αλγόριθμος $LARS$ δίνει όλες τις λύσεις της *Lasso*.

Πριν αποδείξουμε το θεώρημα, δίνουμε δύο λήμματα που χρησιμοποιούνται στην απόδειξη. Υποθέτουμε ότι ο αλγόριθμος $LARS$ έχει συμπληρώσει $k-1$ βήματα, δίνοντας ως εκτιμητή τον $\hat{\mu}_{k-1}$ και ενεργό σύνολο το A_k, ενώ το x_k είναι η πιο πρόσφατη μεταβλητή που εισήλθε στο μοντέλο.

Λήμμα 1:
Αν το x_k ήταν η μοναδική μεταβλητή που εισήλθε στο ενεργό σύνολο στο $(k-1)-$οστό βήμα, τότε η $k-$οστή συνιστώσα w_{kk} του διανύσματος $w_k = H_k G_k^{-1} 1_k$, με αντίστοιχο ισογώνιο διάνυσμα το $u_k = X_k w_k$, έχει το ίδιο πρόσημο με την τρέχουσα συσχέτιση $c_{kk} = x_k'(y - \hat{\mu}_{k-1})$. Επιπλέον, η $k-$οστή συνιστώσα $\hat{\beta}_{kk}$ του διανύσματος $\hat{\beta}_k$, με εκτιμητή το $\hat{\mu}_k = X\hat{\beta}_k$, έχει το ίδιο πρόσημο με το c_{kk} □

Στη συνέχεια, αν συμβολίσουμε με S_A το σύνολο:

$$S_A = \{v = \sum_{j \in A} s_j x_j P_j : \sum_{j \in A} P_j = 1\}$$

έχουμε το παρακάτω λήμμα.

Λήμμα 2:
Το σημείο του S_A που βρίσκεται πιο κοντά στην αρχή των αξόνων είναι το:

$$v_A = H_A u_A = H_A X_A w_A = H_A X_A H_A G_A^{-1} 1_A$$

το οποίο έχει μήκος $||v_A|| = H_A$. Αν $A \subseteq B$, για κάποιο σύνολο δεικτών B, τότε $H_A \geq H_B$, ενώ η τιμή $H_A = 1$ είναι η μέγιστη δυνατή και επιτυγχάνεται όταν το ενεργό σύνολο A περιέχει ένα μόνο στοιχείο.

Το λήμμα αποδεικνύεται με απλή εφαρμογή του κριτηρίου της πρώτης παραγώγου και του *Lagrange* πολλαπλασιαστή. □

Στη συνέχεια, ορίζουμε τα σύνολα $A_1 = \{j : \hat{\beta}_j \neq 0\}$, $A_0 = \{j : \hat{\beta}_j = 0 \text{ και } |\hat{c}_j| = C\}$ και $A_{10} = A_1 \cup A_0$ και θέτουμε 4 περιορισμούς στην επιλογή του ενεργού συνόλου A κατά τη διαδικασία *Lasso*.

Περιορισμός 1: $A_1 \subseteq A$.
Περιορισμός 2: $A \subseteq A_{10}$.
Περιορισμός 3: Για το $w_A = H_A G_A^{-1} 1_A$, ισχύει $sign(w_j) = sign(\hat{c}_j)$, $\forall j \in A_0$.
Περιορισμός 4: Το A ελαχιστοποιεί το H_A.

Τα στοιχεία του ενεργού συνόλου A αυξάνουν όσο προχωράει ο αλγόριθμος $LARS$, αλλά η *Lasso* τροποποίηση επιτρέπει σε αυτά να μειώνονται. Η "ένα τη φορά" συνθήκη σημαίνει ότι σε κάθε βήμα δεν μπορούμε να προσθέσουμε ή να απαλείψουμε πάνω από

ένα δείκτη j.

Απόδειξη του θεωρήματος 1:

Ξεκινώντας από τον εκτιμητή $\hat{\beta}_0 = 0$, ακολουθούμε τον τροποποιημένο $LARS-Lasso$ αλγόριθμο και θα δείξουμε επαγωγικά ότι σε κάθε βήμα ο αλγόριθμος αυτός πρέπει να συμφωνεί με το πρόβλημα ελαχιστοποίησης (4.5) της μεθόδου *Lasso*. Έστω $\hat{\beta}$ η λύση της *Lasso* και της τροποποιημένης $LARS - Lasso$ μεθόδου, η οποία έχει προκύψει σε ένα μόνο συγκεκριμένο βήμα της μεθόδου. Τότε, το σύνολο A_0 είναι κενό, οπότε οι περιορισμοί 1 και 2 δείχνουν ότι το ενεργό σύνολο μένει σταθερό. Εξάλλου, η ισοδυναμία των δύο μεθόδων πρέπει να συνεχίσει να υπάρχει τουλάχιστον μέχρι το τέλος του βήματος.

Η "ένα τη φορά" υπόθεση του θεωρήματος σημαίνει ότι σε κάθε σημείο αλλαγής βήματος, το A_0 έχει ένα ακριβώς στοιχείο, έστω το j_0. Άρα, το A είναι ίσο είτε με το A_1 είτε με το A_{10}. Υπάρχουν δύο περιπτώσεις. Η πρώτη περίπτωση είναι το j_0 να έχει προστεθεί στο σύνολο $\{|\hat{c}_j| = \hat{C}\}$. Τότε το λήμμα 1 δείχνει ότι $sign(w_{j_0}) = sign(\hat{c}_{j_0})$, οπότε ικανοποιείται ο περιορισμός 3. Οι υπόλοιποι τρεις περιορισμοί και το λήμμα 2 δείχνουν ότι η επιλογή $A = A_{10}$ της μεθόδου *Lasso* συμφωνεί με τον $LARS - Lasso$ αλγόριθμο. Η δεύτερη περίπτωση είναι το j_0 να έχει απαλειφθεί από το ενεργό σύνολο, όπως υποδεικνύει η σχέση (4.35). Τότε, σύμφωνα με τον περιορισμό 3, το w_A θα παραμείνει το ίδιο και λόγω της σχέσης (4.35) αποφύγαμε την ασυμφωνία στα πρόσημα για το δείκτη j_0. Δηλαδή $A = A_1$, σύμφωνα με τη *Lasso* τροποποίηση (4.35). Αυτό συμπληρώνει την απόδειξη. □

Το γράφημα της *Lasso* στο Σχήμα 4.1 υπολογίστηκε χρησιμοποιώντας τον τροποποιημένο αλγόριθμο $LARS$. Η τροποποίηση (4.35) εφαρμόστηκε μόνο μία φορά, στο σημείο που δείχνει το βέλος στο αριστερό διάγραμμα. Στο σημείο αυτό, το A περιείχε και τους 10 δείκτες ενώ $A_+ = A - \{7\}$. Η μεταβλητή με δείκτη 7 επανήλθε στο ενεργό σύνολο ένα βήμα μετά. Η σύντομη απουσία της μεταβλητής αυτής επηρέασε την τιμή των $\hat{\beta}_j$, και ιδιαίτερα της $\hat{\beta}_8$. Επίσης, η *Lasso* χρειάστηκε 12 βήματα ενώ η $LARS$ μόνο 10.

4.2.2. Η σχέση μεταξύ *LARS* και *Stagewise*

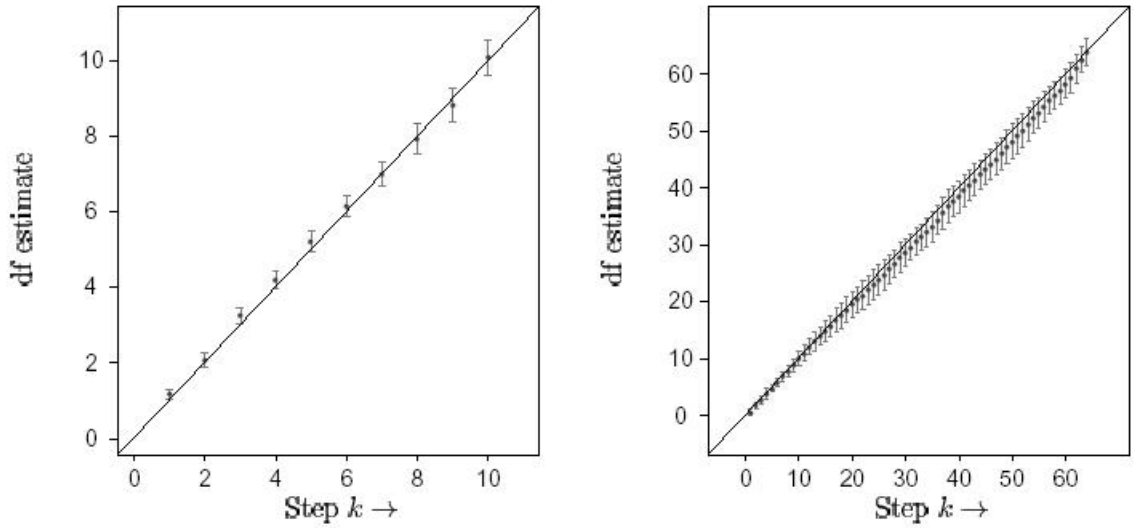
Το Σχήμα 4.2 δείχνει ότι ο *Stagewise* αλγόριθμος μπορεί να συνεχίσει από το $\hat{\mu_1}$, το οποίο είναι σημείο ίσων συσχετίσεων $\hat{c}_1 = \hat{c}_2$. Ως ένα μικρό πρώτο βήμα έχει επιλεχθεί τυχαία ο δείκτης $j = 1$, οπότε κατευθυνόμαστε προς το σημείο $\hat{\mu_1} + \varepsilon x_1$. Τώρα, η μεταβλητή με δείκτη 2 έχει μεγαλύτερη συσχέτιση, αφού:

$$x_2'(y - \hat{\mu}_1 - \varepsilon x_1) > x_1'(y - \hat{\mu}_1 - \varepsilon x_1) \tag{4.36}$$

οπότε το $j = 2$ θα είναι η επόμενη επιλογή της *Stagewise* κ.ο.κ.

Αν θεωρήσουμε μια ιδανική *Stagewise* διαδικασία κατά την οποία το ε είναι πολύ κοντά στο 0, τότε πηγαίνουμε προς την κατεύθυνση της διχοτόμου u_2 στο Σχήμα 4.2, κάνοντας τους *Stagewise* και $LARS$ εκτιμητές να συμφωνούν. Ειδικότερα, συμφωνούν πάντα για $m = 2$ μεταβλητές, αλλά μια άλλη τροποποίηση είναι απαραίτητη ώστε η $LARS$ να δίνει πάντα τους εκτιμητές της *Stagewise* για $m > 2$.

Έστω ότι η *Stagewise* διαδικασία χρειάστηκε N βήματα μεγέθους ε από κάποιο προηγούμενο εκτιμητή $\hat{\mu}$ και έστω N_j το πλήθος των βημάτων που επέλεξαν ως δείκτη το



Σχήμα 4.5: Βαθμοί Ελευθερίας για τους *LARS* εκτιμητές $\hat{\mu}_k$ [*B.Efron, T.Hastie, I.Johnstone, R.Tibshirani* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), $p.407-499$]

$j = 1, 2, ..., m$. Ισχύει ότι $N_j = 0$ για τα j που δεν ανήκουν στο ενεργό σύνολο A. Έστω

$$P \equiv (N_1, N_2, ..., N_m)/N \tag{4.37}$$

Αν συμβολίσουμε με P_A τις συντεταγμένες του P για $j \epsilon A$, ο νέος εκτιμητής θα είναι:

$$\mu = \hat{\mu} + N\varepsilon X_A P_A \tag{4.38}$$

Γνωρίζοντας ότι $X_A = (.....s_j x_j.....)_{j \epsilon A}$, παρατηρούμε ότι τα βήματα της *Stagewise* λαμβάνονται κατά τις διευθύνσεις των $s_j x_j$.

Ο αλγόριθμος *LARS* δίνει τον εκτιμητή:

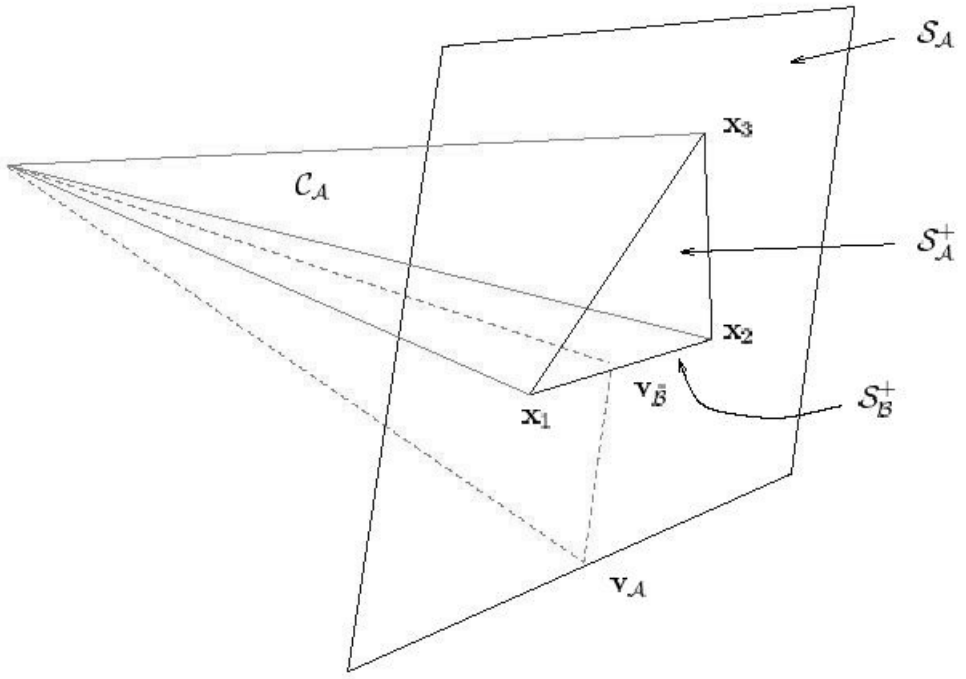
$$\mu_A + \gamma X_A w_A \tag{4.39}$$

όπου $w_A = H_A G_A^{-1} 1_A$. Συγκρίνοντας τις (4.38) και (4.39), παρατηρούμε ότι η *LARS* δεν μπορεί να συμφωνεί με τη *Stagewise* αν το w_A έχει αρνητικές συνιστώσες, αφού το P_A είναι μη αρνητικό. Με άλλα λόγια, η κατεύθυνση $X_A P_A$ της *Stagewise* πρέπει να βρίσκεται μέσα στον κυρτό κώνο που δημιουργείται από τις στήλες του X_A:

$$C_A = \{v = \sum_{j \in A} s_j x_j P_j, P_j \geq 0\}$$

Αν $u_A \epsilon C_A$ τότε δεν υπάρχει καμιά διαφορά μεταξύ των (4.38) και (4.39). Διαφορετικά, αντικαθιστούμε το u_A με την προβολή του στο C_A, δηλαδή το πιο κοντινό σημείο στον κώνο.

Stagewise τροποποίηση: Προχωράμε σύμφωνα με τις σχέσεις (4.15)-(4.20), με τη διαφορά ότι αντικαθιστούμε το u_A με το $u_{\hat{B}}$, το οποίο είναι το μοναδιαίο διάνυσμα που βρίσκεται πάνω στην προβολή του u_A στο C_A, όπως φαίνεται από το Σχήμα 4.6.



Σχήμα 4.6: Η γεωμετρία της $LARS - Stagewise$ τροποποίησης [*B.Efron, T.Hastie, I.Johnstone, R.Tibshirani* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), $p.407 - 499$]

Θεώρημα 2:

Με την προσθήκη της *Stagewise* τροποποίησης, ο αλγόριθμος *LARS* δίνει όλες τις λύσεις της *Stagewise*.

Απόδειξη:

Πρέπει να δείξουμε ότι οι διαδοχικοί *Stagewise* εκτιμητές του β υπολογίζονται σύμφωνα με τον τροποποιημένο $LARS - Stagewise$ αλγόριθμο. Για λόγους απλότητας, υποθέτουμε ότι τα πρόσημα $s_j = sign(\hat{c}_j)$ είναι όλα μη αρνητικά. Αυτό επιτυγχάνεται αν θέσουμε όπου x_j το $-x_j$ για τα j, όπου $s_j < 0$. Υποθέτουμε επίσης, σύμφωνα με τις σχέσεις (4.37) και (4.38), ότι η διαδικασία (4.7) της μεθόδου *Stagewise* έχει κάνει N επιπλέον βήματα από το σημείο όπου $\hat{\mu} = X\hat{\beta}$, οπότε το τρέχον διάνυσμα του εκτιμητή του μ συμβολίζεται με $\hat{\mu}(N)$ ή $\hat{\mu}(\gamma)$ όπου $\gamma \equiv N\varepsilon$.

Τότε, μπορούμε να γράψουμε:

$$\hat{\mu}(\gamma) = \hat{\mu} + \gamma v$$

όπου $v = X_A P_A$, με $P_A = N_j/N > 0$, για $j \in A$.

Η διαδικασία της *Stagewise* θέτει τρεις περιορισμούς:

Περιορισμός 1: $v \in S_A^+$, όπου

$$S_A^+ = \{v = \sum_{j \in A} x_j P_j, P_j \geq 0, \sum_{j \in A} P_j = 1\}$$

Ισοδύναμα, $\gamma v \in C_A$. Αν $B \subseteq A$ και B είναι το σύνολο των δεικτών για τους οποίους τα αντίστοιχα P_j είναι μη-μηδενικά, η *Stagewise* μπορεί να χρησιμοποιήσει το B ως ενεργό σύνολο αντί του A, άρα $v \in L(X_B)$.

Περιορισμός 2: Το v είναι ανάλογο του u_B, δηλαδή:

$$v = v_B = H_B^2 X_B G_B^{-1} 1_B = H_B u_B$$

Επομένως, $X_B' v_B = H_B^2 1_B$, δηλαδή $x_j' v_B = H_B^2$ για $j \in B$.

Περιορισμός 3: Ισχύει $x_j' v_B \geq H_B^2$, για $j \in A - B$.

Αν συμβολίσουμε με $\hat{\beta}(z)$ την οικογένεια των λύσεων της *Stagewise*, τότε αυτή είναι παραγωγίσιμη από τα δεξιά με παράγωγο:

$$\hat{v} = \frac{d\hat{\beta}(z)}{dz}$$

Θέλουμε να δούμε τώρα αν το $\hat{v}$ πληρεί τους τρεις περιορισμούς. Αν $u_A \in C_A$, τότε το $v = v_A$ προφανώς ικανοποιεί τους περιορισμούς. Παίρνουμε τώρα την περίπτωση όπου το u_A δεν ανήκει στο C_A. Παρατηρούμε τότε ότι η προβολή του u_A πάνω στον κώνο που ορίζεται από το σύνολο $B \subset A$, όπου $P_j > 0$ για $j \in B$ και 0 αλλού, είναι ανάλογη του u_B. Άρα και η προβολή του v_A στο B είναι ανάλογη του v_B. Το πιο κοντινό σημείο του u_A στο C_A, έστω το $\hat{u}_A$, έχει τη μορφή $\sum_A x_j \hat{P}_j$, με $\hat{P}_j \geq 0$. Επομένως, το $\hat{u}_A$ βρίσκεται στον κώνο που ορίζει το $\hat{B} = \{j : \hat{P}_j > 0\}$ και θα είναι ίσο με την προβολή του u_A στον $\hat{B}$. Επομένως, θα είναι ανάλογο του $u_{\hat{B}}$, άρα και του $v_{\hat{B}}$. Δηλαδή, το $v_{\hat{B}}$ ικανοποιεί τους δύο πρώτους περιορισμούς και αποδεικνύεται ότι ικανοποιεί και τον τρίτο.

Παρατηρούμε τώρα ότι η $\hat{v} = v_{\hat{B}}$ εξαρτάται μόνο από το σύνολο $\hat{A} = \{j : |\hat{c}_j| = \hat{C}\}$. Όμως, οι εκτιμητές $\hat{\beta}$ που παράγει ο τροποποιημένος $LARS - Stagewise$ αλγόριθμος έχουν παραγώγους που εξαρτώνται και αυτοί μόνο από το $\hat{A}$. Όλα τα παραπάνω δείχνουν ότι ο τροποποιημένος $LARS - Stagewise$ αλγόριθμος είναι μια εκδοχή της *Stagewise* μεθόδου και αποτελούν μια περιγραφή της απόδειξης του θεωρ.2. □

Το διάνυσμα $u_{\hat{B}}$ (Σχήμα 4.6) στην *Stagewise* τροποποίηση παίζει το ρόλο του ισογώνιου διανύσματος (4.13) αλλά για το σύνολο $\hat{B} \subseteq A$ που αντιστοιχεί στην πλευρά του κώνου C_A στην οποία βρίσκεται η προβολή $u_{\hat{B}}$. Η *Stagewise* είναι ένας αλγόριθμος τύπου $LARS$, ο οποίος όμως επιτρέπει στο ενεργό σύνολο να μειωθεί κατά έναν ή περισσότερους δείκτες. Αυτό συνέβει στο σημείο όπου δείχνει το βέλος στο δεξί γράφημα του Σχήματος 4.1. Το σύνολο $A = \{3, 9, 4, 7, 2, 10, 5, 8\}$ έχει μετατραπεί στο $\hat{B} = A - \{3, 7\}$. Χρειάστηκαν 13 βήματα τελικά για να φτάσουμε στη λύση ελαχίστων

τετραγώνων $\bar{\beta}_m = (X'X)^{-1}X'y$. Οι τρεις μέθοδοι: *LARS*, *Lasso* και *Stagewise*, πάντα καταλήγουν στον $\bar{\beta}_m$ τελικά, αλλά η *LARS* το πετυχαίνει σε m μόνο βήματα, ενώ η *Lasso* και ειδικά η *Stagewise* χρειάζονται περισσότερα.

Σύμφωνα με το Θεώρημα 2, η διαφορά μεταξύ δύο διαδοχικών εκτιμητών της *Stagewise*-τροποποιημένης *LARS* μεθόδου είναι:

$$\hat{\mu}_{A_+} - \hat{\mu}_A = \hat{\gamma} u_{\hat{B}} = \hat{\gamma} X_{\hat{B}} w_{\hat{B}}$$

όπως και στην (4.39). Εφόσον το $u_{\hat{B}}$ βρίσκεται στον κυρτό κώνο C_A, το $w_{\hat{B}}$ πρέπει να έχει μη αρνητικές συνιστώσες. Αυτό σημαίνει ότι η διαφορά των διαδοχικών εκτιμητών των συντελεστών για τη συντεταγμένη $j \in \hat{B}$ ικανοποιεί τη σχέση:

$$sign(\hat{\beta}_{+j} - \hat{\beta}_j) = s_j, \tag{4.40}$$

όπου $s_j = sign\{x_j'(y - \hat{\mu})\}$.

Μπορούμε τώρα να δώσουμε τα βασικά χαρακτηριστικά των τριών μεθόδων:

- *Stagewise*: Διαδοχικές διαφορές των $\hat{\beta}_j$ έχουν το ίδιο πρόσημο με την τρέχουσα συσχέτιση $\hat{c}_j = x_j'(y - \hat{\mu})$.
- *Lasso*: Η $\hat{\beta}_j$ έχει το ίδιο πρόσημο με τη $\hat{c}_j$.
- *LARS*: Κανένας περιορισμός στα πρόσημα.

Η ιδιότητα (4.40) φανερώνει ότι οι εκτιμητές $\hat{\beta}_j$ της *Stagewise* κινούνται μονότονα απομακρυνόμενοι από το 0. Αναστροφές είναι πιθανές μόνο αν το $\hat{c}_j$ αλλάζει πρόσημο ενώ το $\hat{\beta}_j$ μένει σταθερό μεταξύ δύο περιόδων αλλαγής. Αυτό συνέβη στη μεταβλητή με δείκτη 7 στο Σχήμα 4.1 μεταξύ 8ου και 10ου βήματος της *Stagewise*- τροποποιημένης *LARS* μεθόδου.

4.3. Βαθμοί Ελευθερίας και C_p Εκτιμητές

Τα Σχήματα 4.1 και 4.3 δείχνουν όλους τους πιθανούς *Lasso*, *Stagewise* και *LARS* εκτιμητές του διανύσματος β για τα δεδομένα μας. Για την κατασκευή ενός μοντέλου όμως, θέλουμε ένα μόνο $\hat{\beta}$, οπότε χρησιμοποιούμε ως κριτήριο επιλογής το C_p ειδικά μεταξύ των *LARS* εκτιμητών.

Έστω ότι το $\hat{\mu} = g(y)$ αποτελεί το μοναδικό τύπο για να εκτιμήσουμε το μ σε συνάρτηση με το διάνυσμα y. Εδώ, έχουμε στη διάθεσή μας τις επεξηγηματικές μεταβλητές $x_1, x_2, ..., x_m$ και την εξαρτημένη μεταβλητή y, για την οποία ισχύει:

$$y \sim (\mu, \sigma^2 I) \tag{4.41}$$

δηλαδή οι συνιστώσες y_i είναι ασυσχέτιστες με μέση τιμή μ_i και διασπορά σ^2. Προφανώς, ισχύει ότι:

$$(\hat{\mu}_i - \mu_i)^2 = (y_i - \hat{\mu}_i)^2 - (y_i - \mu_i)^2 + 2(\hat{\mu}_i - \mu_i)(y_i - \mu_i)$$

Αν πάρουμε μέσες τιμές και αθροίσουμε ως προς i, έχουμε:

$$E\{\frac{||\hat{\mu}-\mu||^2}{\sigma^2}\} = E\{\frac{||y-\hat{\mu}||^2}{\sigma^2} - n\} + 2\sum_{i=1}^{n} cov(\hat{\mu}_i, y_i)/\sigma^2 \tag{4.42}$$

Ο τελευταίος όρος της (4.42) αποτελεί τους βαθμούς ελευθερίας για ένα εκτιμητή $\hat{\mu} = g(y)$, δηλαδή:

$$df_{\mu,\sigma^2} = \sum_{i=1}^{n} cov(\hat{\mu}_i, y_i)/\sigma^2 \tag{4.43}$$

ενώ ο τύπος για την εκτίμηση ως προς C_p του κινδύνου είναι:

$$C_p(\hat{\mu}) = \frac{||y-\hat{\mu}||^2}{\sigma^2} - n + 2df_{\mu,\sigma^2} \tag{4.44}$$

Αν τα σ^2 και df_{μ,σ^2} είναι άγνωστα, το $C_p(\hat{\mu})$ είναι ένας αμερόληπτος εκτιμητής του κινδύνου $E\{||\hat{\mu}-\mu||^2/\sigma^2\}$. Για γραμμικούς εκτιμητές $\hat{\mu} = My$, το μοντέλο (4.41) δίνει $df_{\mu,\sigma^2} = trace(M)$.

Πρακτική χρήση του τύπου (4.44) προϋποθέτει να έχουμε εκτιμητές των μ και σ^2 και να ισχύει $df_{\mu,\sigma^2} = trace(M)$. Έστω ότι $\bar{\mu}$ και $\bar{\sigma}^2$ είναι οι OLS εκτιμητές και ότι τα δείγματα y^* και οι αντίστοιχοι εκτιμητές $\hat{\mu}^*$ κατασκευάστηκαν σύμφωνα με:

$$y^* \sim N(\bar{\mu}, \bar{\sigma}^2), \hat{\mu}^* = g(y^*) \tag{4.45}$$

Αν εκτελέσουμε B ανεξάρτητες επαναλήψεις για τη σχέση (4.45), παίρνουμε εκτιμητές για τις συνδιασπορές στην (4.43):

$$\hat{cov}_i = \frac{\sum_{b=1}^{B} \hat{\mu}_i^*(b)[y_i^*(b) - y_i^*(.)]}{B-1}$$

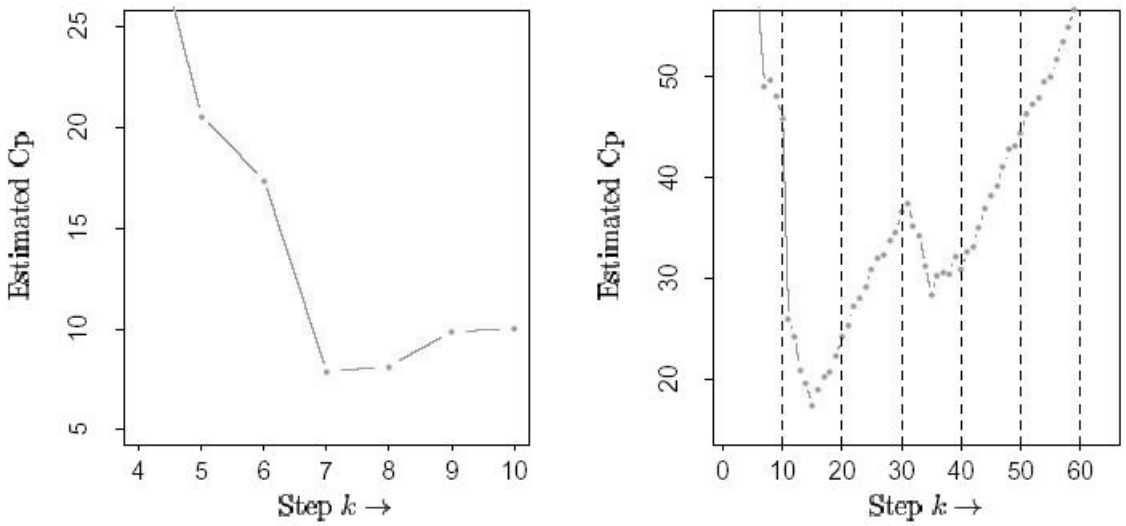
όπου $y^*(.) = \frac{\sum_{b=1}^{B} y^*(b)}{B}$. Τότε ως εκτιμητή $\hat{df}$ του $df_{\mu,\sigma^2} = trace(M)$ παίρνουμε το:

$$\hat{df} = \sum_{i=1}^{n} \hat{cov}_i/\bar{\sigma}^2$$

Το αριστερό γράφημα του Σχήματος 4.7 απεικονίζει τους εκτιμητές $\hat{df}_k$ που υπολογίστηκαν με τη βοήθεια των εκτιμητών $\hat{\mu}_k$ της μεθόδου $LARS$ για τα δεδομένα μας, με $k = 1, 2, ..., m = 10$. Εκφράζει μια σχετικά απλή σχέση, την οποία θα καλούμε *απλή προσέγγιση*:

$$df(\hat{\mu}_k) = k \tag{4.46}$$

Το δεξί γράφημα παριστάνει ξανά τα $\hat{df}_k$, αλλά αυτή τη φορά για $m = 64$ επεξηγηματικές μεταβλητές, οι οποίες περιλαμβάνουν τα γινόμενα και τα τετράγωνα των 10 βασικών μεταβλητών, δηλαδή εκτός από τις 10 βασικές μεταβλητές, χρησιμοποιήθηκαν τα 45 γινόμενα και τα 9 τετράγωνα, με εξαίρεση το τετράγωνο της διχοτόμου x_2. Παρατηρούμε ότι η απλή προσέγγιση (4.46) είναι και εδώ ακριβής μέσα στα όρια του υπολογισμού



Σχήμα 4.7: C_p Εκτιμητές του κινδύνου για τις δύο καταστάσεις του Σχήματος 4.5 [*B.Efron, T.Hastie, I.Johnstone, R.Tibshirani* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), *p.*407 – 499]

$\hat{df} = \sum_{i=1}^{n} \hat{cov}_i/\bar{\sigma}^2$, όπου οι B=500 επαναλήψεις έχουν χωριστεί σε 10 ομάδες των 50, ώστε να υπολογιστούν τα διαστήματα εμπιστοσύνης.

Αν δεχτούμε την (4.46), μπορούμε να εκτιμήσουμε τον κίνδυνο ενός $LARS$ εκτιμητή $\hat{\mu}_k$:

$$C_p(\hat{\mu}_k) = ||y - \hat{\mu}_k||^2/\bar{\sigma}^2 - n + 2k \tag{4.47}$$

Η σχέση αυτή, που είναι όμοια με τον C_p εκτιμητή του κινδύνου για έναν OLS εκτιμητή βασισμένο σε ένα υποσύνολο k προεπιλεγμένων επεξηγηματικών μεταβλητών, έχει το πλεονέκτημα ότι δεν απαιτεί άλλους υπολογισμούς εκτός αυτών που χρειάζονται για να υπολογιστούν οι OLS εκτιμητές. Ο τύπος εφαρμόζεται μόνο για τη $LARS$ αλλά όχι και για τη *Lasso* ή τη *Stagewise*.

Το Σχήμα 4.7 παριστάνει το $C_p(\hat{\mu}_k)$ σε συνάρτηση με το k και για τις δύο καταστάσεις του Σχήματος 4.5. Παρατηρούμε ότι το ελάχιστο C_p επιτυγχάνεται στα βήματα $k = 7$ και $k = 16$ αντίστοιχα. Και τα δύο μοντέλα που αντιστοιχούν στα ελάχιστα C_p δείχνουν λογικά και οι αρχικές επιλογές των σημαντικών επεξηγηματικών μεταβλητών συμφωνούν με ένα μοντέλο που βασίστηκε σε λεπτομερή εξέταση των δεδομένων με τη βοήθεια ιατρικής εξειδίκευσης.

Θεώρημα 3:
Αν οι επεξηγηματικές μεταβλητές $x_1, x_2, ..., x_m$ είναι ανά δύο ορθογώνιες, τότε ο εκτιμητής $\hat{\mu}_k$ της $LARS$ για το k βήμα έχει β.ε. ίσους με $df(\hat{\mu}_k) = k$.

Για να διατυπώσουμε μια γενικότερη παραλλαγή του Θεωρήματος 3, ας παρουσιάσουμε τη:

Συνθήκη Θετικού Κώνου: Για όλους τους υποπίνακες X_A του πλήρους πίνακα σχεδιασμού X, ισχύει $G_A^{-1}1_A > 0$.

Η Συνθήκη Θετικού Κώνου ισχύει αν ο X είναι ορθογώνιος. Είναι αυστηρά πιο γενική από την ορθογωνιότητα, αλλά παραδείγματα δείχνουν ότι δεν την ικανοποιούν όλοι οι πίνακες σχεδιασμού X.

Μπορεί να αποδειχθεί ότι η *LARS*, η *Lasso* και η *Stagewise* ταυτίζονται όταν ισχύει η Συνθήκη Θετικού Κώνου, οπότε η (4.46) ισχύει και για τις τρεις μεθόδους σε αυτήν την περίπτωση, δηλαδή:

Θεώρημα 4:
Υπό την υπόθεση της "Συνθήκης Θετικού Κώνου", ισχύει $df(\hat{\mu}_k) = k$.

Αφού η Συνθήκη Θετικού Κώνου ισχύει όταν ο X είναι ορθογώνιος, η απόδειξη του θεωρήματος 4 αρκεί για να αποδειχθεί το θεώρημα 3. Πριν όμως περιγράψουμε την απόδειξη του θεωρήματος 4, παραθέτουμε παρακάτω τρία λήμματα.

Λήμμα 3:
Έστω ότι ο X είναι διαγώνιος, δηλαδή ότι $x_j = e_j$, $j = 1, ..., n$, όπου το διάνυσμα e_j έχει μηδενικά στοιχεία εκτός το $j-$οστό στοιχείο που είναι ίσο με τη μονάδα. Έστω επίσης ότι οι απόλυτες τιμές των συνιστωσών του y διατάσσονται κατά φθίνουσα σειρά ως εξής:

$$|y|_{(1)} \geq |y|_{(2)} \geq \ldots \geq |y|_{(n)} \geq |y|_{n+1} := 0$$

Τότε, ο $k-$οστός εκτιμητής της *LARS* με $0 \leq k \leq n$, δίνεται από τη σχέση:

$$\hat{\mu}_{k,i}(y) = \begin{cases} y_i - |y|_{(k+1)}, & \text{αν } y_i > |y|_{(k+1)} \\ 0, & \text{αν } y_i \leq |y|_{(k+1)} \\ y_i - |y|_{(k+1)}, & \text{αν } y_i < -|y|_{(k+1)} \end{cases}$$

Για το παρακάτω λήμμα, δίνουμε τους εξής ορισμούς:

- Ένας εκτιμητής $\hat{\mu}(y)$ είναι τοπικά γραμμικός σε ένα σημείο y_0 όταν υπάρχει περιοχή του y_0 στην οποία το $\hat{\mu}(y)$ να είναι γραμμικός, δηλαδή $\exists M : \hat{\mu}(y) = My$.
- Ένα σύνολο G έχει πλήρες μέτρο, όταν το συμπλήρωμά του έχει *Lebesgue*-μέτρο ίσο με 0.
- Για κάθε σχεδόν παραγωγίσιμη συνάρτηση $g : \Re^n \to \Re^n$ η κλίση ∇g είναι ίση με $\sum_{i=1}^n \partial g_i / \partial x_i$.

Λήμμα 4:
Υπάρχει ένα ανοικτό σύνολο G_k πλήρους μέτρου τέτοιο ώστε για κάθε $y \in G_k$, ο $\hat{\mu}_k(y)$ είναι τοπικά γραμμικός και μάλιστα $\nabla \hat{\mu}_k(y) = k$. □

Λήμμα 5:
Υπό τη συνθήκη του θετικού κώνου, ο $\hat{\mu}_k(y)$ είναι μια συνεχής και σχεδόν παραγωγίσιμη συνάρτηση. □

Απόδειξη του θεωρήματος 4:
Αν τώρα $y \sim N(\mu, \sigma^2 I)$, τότε λόγω του λήμματος 5, ο τύπος του *Stein*(1981) γίνεται:

$$\sum_{i=1}^{n} cov(\hat{\mu}_{k,i}, y_i)/\sigma^2 = E[\nabla \hat{\mu}_k(y)]$$

Όμως, το αριστερό μέλος ισούται με τους βαθμούς ελευθερίας του $\hat{\mu}_k$, δηλαδή το $df(\hat{\mu}_k)$. Αρκεί λοιπόν να δείξουμε ότι $\nabla \hat{\mu}_k(y) = k$. Στην ορθογώνια περίπτωση, λόγω του λήμματος 3 ισχύει ότι: $\nabla \hat{\mu}_k = \sum_i \frac{\partial \hat{\mu}_{k,i}}{\partial y_i}(y) = \sum_i I\{|y_i| > |y|_{(k+1)}\} = k$. Στη γενική περίπτωση, λόγω του λήμματος 4 ισχύει ότι $\nabla \hat{\mu}_k = k$. Οπότε, τελικά, για κάθε περίπτωση, αποδείξαμε το θεώρημα 4, δηλαδή ότι $df(\hat{\mu}_k) = k$. □

4.4. Κόστος Υπολογισμών

Τα βήματα του αλγορίθμου *LARS* με $m < n$ μεταβλητές απαιτούν συνολικά $O(m^3 + nm^2)$ υπολογισμούς, δηλαδή όσους απαιτεί και η μεθόδος ελαχίστων τετραγώνων για m μεταβλητές. Πιο συγκεκριμένα, στο $k-$οστό βήμα της μεθόδου, υπολογίζουμε τα $m-k$ εσωτερικά γινόμενα c_{jk} των μη-ενεργών x_j ενώ τα τρέχοντα υπόλοιπα καθορίζουν την επόμενη ενεργή μεταβλητή. Έπειτα, αναστρέφουμε τον $k \times k$ πίνακα $G_k = X_k' X_k$ για να βρούμε την επόμενη *LARS* κατεύθυνση. Αυτό το κάνουμε ενημερώνοντας την *Cholesky* παραγοντοποίηση R_{k-1} του G_{k-1} που βρέθηκε στο προηγούμενο βήμα. Στο τελευταίο βήμα m, έχουμε υπολογίσει με την *Cholesky* παραγοντοποίηση τον πλήρη πίνακα $R = R_m$, που είναι ο κύριος υπολογισμός στην μέθοδο ελαχίστων τετραγώνων. Επομένως, η διαδικασία *LARS* μπορεί να θεωρηθεί ως μια παραγοντοποίηση *Cholesky* με καθοδηγούμενη ταξινόμηση των μεταβλητών. Οι υπολογισμοί μπορούν να μειωθούν περισσότερο παρατηρώντας ότι τα εσωτερικά γινόμενα παραπάνω μπορούν να μεταβάλλονται σε κάθε επανάληψη χρησιμοποιώντας τον πίνακα $X'X$ και τις πρόσφατες κατευθύνσεις.

Για τη *Lasso* τροποποίηση, οι υπολογισμοί είναι παρόμοιοι εκτός του ότι συχνά πρέπει να απαλείφουμε κάποια μεταβλητή, και άρα να ενημερώνουμε εκ νέου το R_k (το οποίο θα μας κοστίσει το πολύ $O(m^2)$ πράξεις για κάθε ενημέρωση). Για την *Stagewise* τροποποίηση της *LARS*, πρέπει να ελέγχουμε σε κάθε επανάληψη αν τα στοιχεία του w είναι όλα θετικά. Αν όχι, μία ή περισσότερες μεταβλητές απαλείφονται, γεγονός που απαιτεί ξανά ενημέρωση του R_k. Όταν υπάρχουν πολλές συσχετισμένες μεταβλητές, η *Stagewise* εκδοχή μπορεί να χρειαστεί πολύ περισσότερα βήματα από τη *LARS* λόγω της συχνής απαλοιφής και εισαγωγής μεταβλητών, η οποία αυξάνει τους υπολογισμούς κατά παράγοντα ίσο ή ακόμα και μεγαλύτερο του 5.

Ο αλγόριθμος *LARS* συμπεριφέρεται ικανοποιητικά ακόμα και στην περίπτωση που οι μεταβλητές είναι πολύ περισσότερες από τις παρατηρήσεις, δηλαδή: $m >> n$. Σε αυτή την περίπτωση, ο αλγόριθμος *LARS* τερματίζει στον υπολογισμό των *OLS* εκτιμητών αφότου $n-1$ μεταβλητές εισέλθουν στο ενεργό σύνολο, γεγονός που στοιχίζει $O(n^3)$

πράξεις. Οι μεταβλητές που θα εισέλθουν είναι $n-1$ και όχι n, επειδή οι στήλες του X έχουν κανονικοποιηθεί με κέντρο τη μέση τιμή τους και επομένως ο πίνακας έχει τάξη $n-1$. Τέλος, κάνουμε μερικές παρατηρήσεις για τη μέθοδο *Lasso* στην περίπτωση όπου $m >> n$:

- Ο αλγόριθμος $LARS$ συνεχίζει να δίνει τις λύσεις που δίνει η μέθοδος *Lasso* και η τελική λύση επιβεβαιώνει το γεγονός ότι η εφαρμογή της *Lasso* δεν μπορεί να έχει περισσότερες από $n-1$ μεταβλητές με μη-μηδενικούς συντελεστές.

- Παρόλο που το μοντέλο δεν περιλαμβάνει ποτέ περισσότερες από $n-1$ μεταβλητές, το πλήθος των *διαφορετικών* μεταβλητών που εισήλθαν στο μοντέλο κατά τη διαδικασία μπορεί να είναι, και τυπικά είναι, μεγαλύτερος του $n-1$.

- Το μοντέλο, ειδικά κοντά στο τέλος της διαδικασίας, τείνει να είναι αρκετά μεταβαλλόμενο όταν το y μεταβάλλεται ελάχιστα.

Κεφάλαιο 5

ΣΤΑΤΙΣΤΙΚΗ ΕΦΑΡΜΟΓΗΣ ΜΕΘΟΔΟΥ *LARS*

Σε αυτό το κεφάλαιο, εφαρμόζουμε τη μέθοδο $LARS$ που περιγράψαμε στο προηγούμενο κεφάλαιο, αλλά και τις μεθόδους *Lasso* και *Forward Stagewise* ώστε να συγκρίνουμε τα αποτελέσματά τους. Για το σκοπό αυτό, έχουμε στη διάθεσή μας δεδομένα τα οποία θα χρησιμοποιήσουμε μέσα από το στατιστικό πακέτο R για να εφαρμόσουμε τις μεθόδους. Επιλέγουμε λοιπόν τις κατάλληλες μεταβλητές και εκτιμούμε τους συντελεστές ώστε να προσδιορίσουμε το κατάλληλο μοντέλο που θα προσαρμόσουμε στα δεδομένα μας.

Η έρευνα, σύμφωνα με την οποία καταγράφηκαν τα δεδομένα που διαθέτουμε, αφορά βρέφη που γεννήθηκαν με βάρος μικρότερο των 1500 γραμμαρίων. Έχουμε πάρει $n = 39$ δείγματα από τις $m = 13$ επεξηγηματικές μεταβλητές x_1, x_2, ..., x_{13} αλλά και από την εξαρτημένη μεταβλητή y. Οι επεξηγηματικές μεταβλητές εκφράζουν τις δοσοληψίες σε διάφορα φάρμακα που πήρε κάθε γυναίκα κατά τη διάρκεια της εγκυμοσύνης της, το βάρος του νεογέννητου τον πρώτο μήνα και άλλα σωματικά χαρακτηριστικά του. Η εξαρτημένη μεταβλητή εκφράζει το βάρος του βρέφους κατά τον 6ο μήνα της ζωής του. Τα δεδομένα φαίνονται στο Πίνακα 5.1 (*Loui*, *A.*, *Tsalikaki*, *E.*, *Maier*, *K.*, *Walch*, *E.*, *Kamarianakis*, *Y.*, (2006)).

Σκοπός μας είναι να επιλέξουμε ένα κατάλληλο υποσύνολο των επεξηγηματικών μεταβλητών και παράλληλα να εκτιμήσουμε τους αντίστοιχους συντελεστές. Με τη βοήθεια δύο συγκεκριμένων κριτηρίων επιλογής, το C_p στατιστικό του *Mallows* και το $cross - validation$ σφάλμα πρόβλεψης (το οποίο αναφέρθηκε στο κεφάλαιο 3), κατασκευάζουμε ένα βέλτιστο γραμμικό μοντέλο, με το οποίο θα μπορούμε να προβλέψουμε το βάρος ενός βρέφους στον έκτο μήνα της ζωής του αν διαθέτουμε τις τιμές για τις m επεξηγηματικές μεταβλητές.

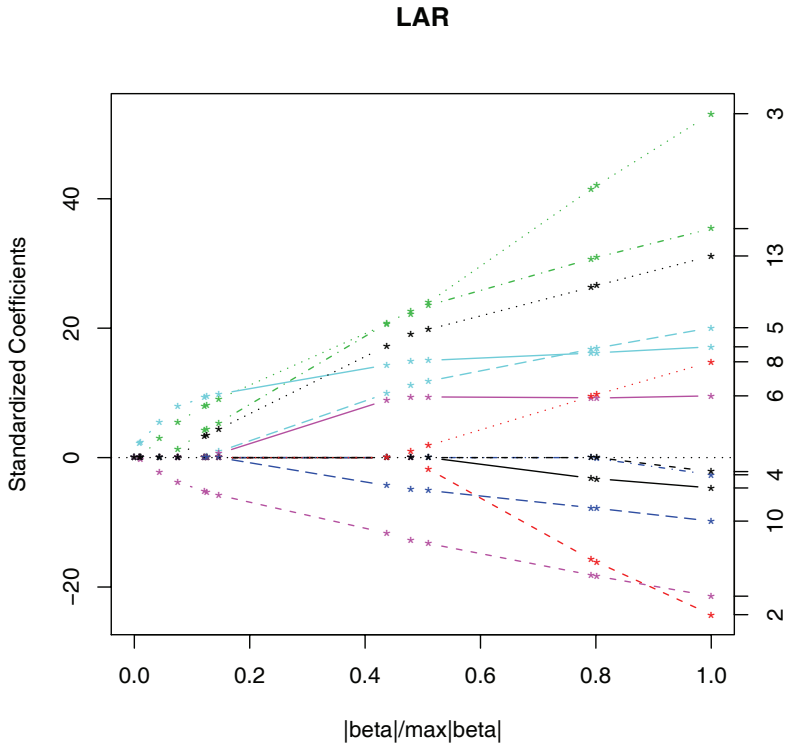
Όλα τα παραπάνω θα γίνουν με τη βοήθεια της μεθόδου της Ελάχιστης Γωνίας Παλινδρόμησης ($LARS$). Παράλληλα, θα εφαρμόσουμε τη *Lasso* και τη *Forward Stagewise* και θα συγκρίνουμε τα αποτελέσματα που έδωσαν αυτές οι δύο μέθοδοι σε σχέση με τη $LARS$. Τα δεδομένα έχουν κανονικοποιηθεί σύμφωνα με τη σχέση (4.1).

	x1	x2	x3	x4	x5	x6	x7	x8	x9	x10	x11	x12	x13	y
	gew1	k3	L1	ku1	ku.cm/Wo	Antib_T	Coffein_T	PE_T	EnterE_LT	GG_WE.LT	AVE.ml/kg/d	AVE.Pr.ges.g/kg/d	AVE.kcal/kg/d	gew6Mo
1	1350	66,412	41,2	29	0,1	14	35	35	28	19	184	3,2	100	8200
2	1410	66,72	40	28	0,8	0	0	14	24	9	133	2,3	98	7870
3	1310	63,732	39	27,4	0,52	0	23	14	21	13	142	2,3	100	7020
4	1120	57,922	37	25	0,4	0	35	28	32	12	163	2,6	102	6990
5	1440	68,506	36,5	26	0,9	0	7	15	16	11	156	3,0	116	8600
6	1475	72,322	40	30	0,4	0	0	3	5	8	164	3,2	117	7450
7	1015	52,196	35	26,5	0,46	7	8	15	11	13	167	2,8	122	4890
8	1100	56,194	38	25	0,4	14	35	19	19	17	147	3,2	113	6280
9	1350	66,73	39	29	0,7	0	0	16	15	11	160	3,0	112	6355
10	1070	59,064	39	25,5	0,46	7	25	25	26	11	175	2,3	104	9880
11	1455	62,096	39	30	0,6	0	10	8	10	13	153	3,6	119	8000
12	1210	65,136	38	28	0,76	13	0	35	12	10	148	2,3	81	8540
13	1340	60,41	40	25,5	0,8	6	32	19	26	16	157	3,2	103	7650
14	1455	66,254	36,5	26,5	0,6	7	17	30	25	18	176	2,4	113	8690
15	1210	60,616	34,5	26	0,2	47	32	89	78	23	181	2,1	96	5995
16	1320	60,794	36	29	1	0	14	15	17	9	148	5,0	112	6270
17	1340	62,86	41	28	0,4	7	0	45	6	9	171	1,9	93	6610
18	1445	68,466	41	29,3	0,74	0	0	21	19	11	164	2,8	113	6050
19	1040	60,192	35,4	26	0,7	7	20	23	21	10	181	3,0	114	6550
20	950	49,572	32,5	25	0,94	13	24	32	14	13	160	3,5	119	5740
21	820	53,624	32	25,7	0,66	11	6	23	13	12	168	2,8	117	5740
22	890	52,102	35	25,4	0,64	6	40	18	17	12	163	3,8	116	6650
23	745	53,402	34,5	24,5	0,94	0	0	14	18	4	151	3,5	112	6450
24	865	49,668	34	25	0,69	15	26	32	31	13	174	2,7	98	7120
25	995	56,542	36	24,5	0,66	7	54	28	31	14	170	2,9	102	9999
26	765	53,992	34	22,2	0,88	17	56	21	18	8	176	2,9	118	9999
27	860	49,254	33	23,7	0,69	20	54	24	23	9	181	3,0	117	9999
28	565	37,236	29	21,5	0,69	19	56	19	19	11	179	5,4	122	5920
29	874	49,572	31,5	23,7	0,56	37	36	56	13	15	190	3,3	116	7960
30	840	60,414	36,2	25	0,69	10	39	23	13	1	153	2,3	100	6640
31	935	52,95	35	24	0,5	13	25	28	29	16	174	2,8	111	6175
32	465	35,44	27,5	20,5	0,51	30	35	43	17	7	190	3,3	119	4540
33	780	49,832	35	24	0,38	23	31	67	17	7	181	3,3	118	6160
34	590	40,158	31,5	21,8	0,46	28	40	38	18	7	195	3,1	118	5320
35	790	52,618	33,5	25	0,3	14	56	31	24	12	149	2,8	103	6060
36	645	49,614	30	22	0,54	21	49	43	22	7	145	2,4	107	5629
37	970	53,75	35	25	0,63	0	56	27	25	18	171	3,6	110	6820
38	835	47,608	34,5	23	0,69	18	56	50	53	10	175	3,4	105	7810
39	765	44,628	34	25	0,56	13	56	40	40	16	181	3,4	102	6950

Πίνακας 5.1: Αποτελέσματα Έρευνας για το βάρος νεογέννητων μωρών

5.1. Εφαρμογή της $LARS$

Αρχικά, χρησιμοποιούμε τη μέθοδο $LARS$ για να βρούμε τη σειρά κατά την οποία εισέρχονται οι μεταβλητές στο μοντέλο. Το Σχήμα 5.1 μας δείχνει το γράφημα των κανονικοποιημένων εκτιμώμενων συντελεστών παλινδρόμησης $\hat{\beta}_i$ ως προς την $L1$ νόρμα του διανύσματος $\hat{\beta}$ των συντελεστών $\hat{\beta}_i$, $i = 1, 2, ..., 13$.



Σχήμα 5.1: Εκτιμήσεις των συντελεστών παλινδρόμησης $\hat{\beta}_i$, $i = 1, 2, ..., 13$ για τη $LARS$

Η σειρά τότε κατά την οποία μπαίνουν οι μεταβλητές στο μοντέλο είναι η εξής:

$$11 \to 12 \to 3 \to 9 \to 13 \to 5 \to 6 \to 10 \to 8 \to 2 \to 1 \to 4 \to 7$$

Το διάνυσμα $\tilde{\beta}$ των εκτιμώμενων συντελεστών για το μοντέλο που περιέχει όλες τις μεταβλητές ταυτίζεται με αυτό των συντελεστών που προκύπτουν από την εφαρμογή της μεθόδου των ελαχίστων τετραγώνων στα δεδομένα μας και είναι σε ακρίβεια τριών δεκαδικών ψηφίων ίσο με:

$$\tilde{\beta} = \begin{pmatrix} -0.076 \\ -0.357 \\ 1.276 \\ -0.067 \\ 0.479 \\ 0.353 \\ -0.058 \\ 0.343 \\ 0.786 \\ -0.357 \\ 0.248 \\ -0.305 \\ 0.443 \end{pmatrix} \tag{5.1}$$

και άρα χρησιμοποιούμε το μοντέλο:

$$\hat{Y} = X\tilde{\beta} \tag{5.2}$$

όπου $X = (X_1, X_2, ..., X_{13})'$ το διάνυσμα των τιμών των μεταβλητών που εισάγουμε και $\hat{Y}$ η εκτιμώμενη τιμή της εξαρτημένης μεταβλητής για τη συγκεκριμένη τιμή του διανύσματος X.

5.2. Εφαρμογή της *Lasso* και της *Stagewise*

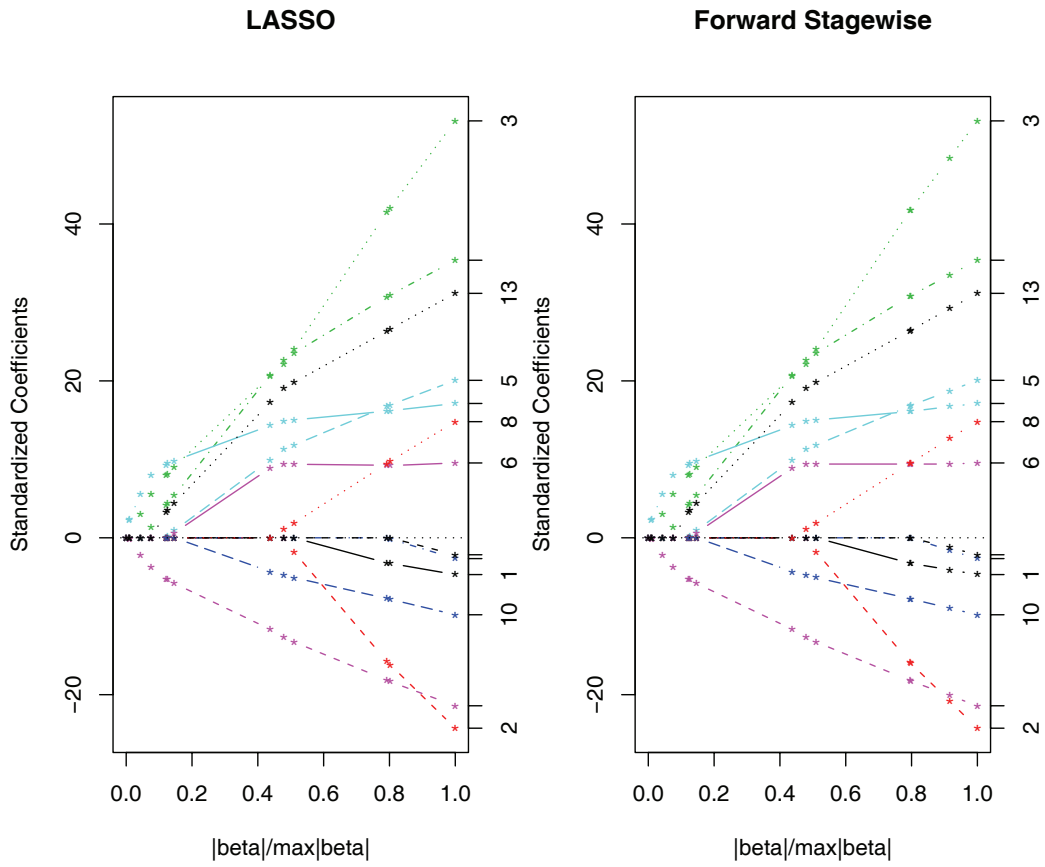
Στη συνέχεια, θα χρησιμοποιήσουμε το R για να βρούμε τη σειρά με την οποία εισέρχονται οι μεταβλητές στο μοντέλο κατά τις μεθόδους *Lasso* και *Stagewise*. Τα γραφήματα των εκτιμήσεων των συντελεστών παλινδρόμησης $\hat{\beta}_i$, $i = 1, 2, ..., 13$ για τη *Lasso* και τη *Stagewise* απεικονίζονται στο Σχήμα 5.2.

Η *Lasso* έδωσε ακριβώς τα ίδια αποτελέσματα με τη *LARS* γι' αυτό και τα γραφήματά τους ταυτίζονται. Παρόλο που η *Lasso* γενικά χρειάζεται περισσότερα βήματα από τη *LARS* για να τερματιστεί η διαδικασία, εδώ χρειάστηκαν μόνο 13 βήματα, όσα δηλαδή χρειάστηκαν και για τη *LARS*. Υπενθυμίζουμε ότι η *LARS* χρειάζεται πάντα τόσα βήματα όσα και το πλήθος των επεξηγηματικών μεταβλητών. Παρατηρούμε ότι σε κανένα βήμα της *Lasso* δεν εξήλθε κάποια μεταβλητή. Επίσης, αφού το άθροισμα των απολύτων τιμών των εκτιμώμενων συντελεστών στη μέθοδο των ελαχίστων τετραγώνων είναι ίσο με 5,148, η μέθοδος *Lasso* (3.10) ταυτίζεται με αυτήν των ελαχίστων τετραγώνων για $t \geq 5,148$ και χρησιμοποιεί όλες τις επεξηγηματικές μεταβλητές στο μοντέλο.

Η *Stagewise* έδωσε παρόμοια αποτελέσματα αλλά όχι ακριβώς τα ίδια. Η σειρά με την οποία οι μεταβλητές εισήλθαν στο μοντέλο ή εξήλθαν από αυτό είναι η εξής:

$$11 \to 12 \to 3 \to 9 \to 13 \to 5 \to 6 \to 10 \to 8 \to$$

$$2, -6 \to 1 \to 4 \to 7 \to 6$$

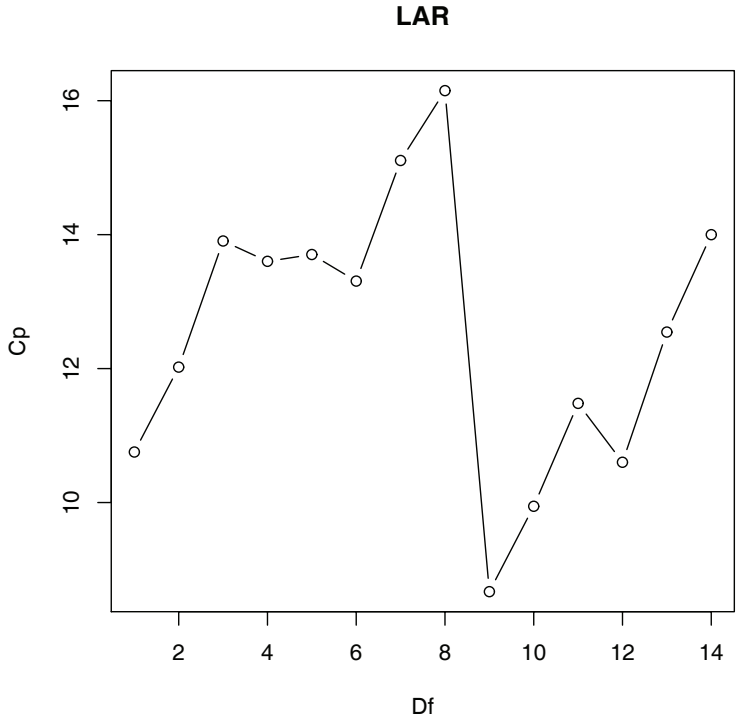


Σχήμα 5.2: Εκτιμήσεις των συντελεστών παλινδρόμησης $\hat{\beta}_i$, $i = 1, 2, ..., 13$ για τη *Lasso*

όπου το "-" δηλώνει ότι η εν λόγω μεταβλητή εξέρχεται από το μοντέλο. Η *Stagewise* λοιπόν χρειάστηκε 14 βήματα για να τερματιστεί, δηλαδή περισσότερα από αυτά που χρειάστηκε η *LARS* ή η *Lasso*. Το γράφημά της ταυτίζεται με αυτό των *LARS* και *Lasso* μέχρι και το 9ο βήμα ενώ υπάρχουν κάποιες μικρές διαφορές από το 10ο βήμα και μετά.

5.3. C_p Κριτήριο

Πρέπει τώρα να χρησιμοποιήσουμε ένα κριτήριο ώστε να επιλέξουμε ένα κατάλληλο υποσύνολο των επεξηγηματικών μεταβλητών που θα συμπεριλάβουμε στο μοντέλο. Αρχικά, θα χρησιμοποιήσουμε το *κριτήριο C_p του Mallows*. Το Σχήμα 5.3 παριστάνει το γράφημα των τιμών του στατιστικού C_p για κάθε βήμα της μεθόδου *LARS* σε συνάρτηση με τους βαθμούς ελευθερίας του εκτιμητή της y για κάθε βήμα, τους οποίους θεωρούμε, σύμφωνα με την "άπλή προσέγγιση" που εξετάσαμε στο κεφάλαιο 4, ότι είναι ίσοι με το αντίστοιχο βήμα.



Σχήμα 5.3: C_p εκτιμητές για τη *LARS* μέθοδο

Βλέπουμε στο Σχήμα 5.3 ότι το ελάχιστο C_p επιτυγχάνεται στο μοντέλο που προκύπτει στο 8ο βήμα της μεθόδου (το *R* θεωρεί ότι το πρώτο βήμα είναι αυτό που μας δίνει μηδενικό διάνυσμα συντελεστών). Αυτό το μοντέλο θα θεωρήρουμε ότι είναι και το βέλτιστο. Περιλαμβάνει τις μεταβλητές x_{11}, x_{12}, x_3, x_9, x_{13}, x_5, x_6 και x_{10}. Τα μοντέλα που προκύπτουν στο 9ο, και 11ο βήμα μπορούν να χαρακτηριστούν επίσης καλά μοντέλα μιας και η τιμή του C_p είναι κοντά στο p, δηλαδή κοντά στο 9 και το 11 αντιστοίχως.

Υπολογίζοντας, με τη βοήθεια του R, τους συντελεστές των μεταβλητών όταν η μέθοδος φτάσει στο 8ο βήμα, βρίσκουμε το διάνυσμα $\hat{\beta}_{Cp}$ των εκτιμώμενων κανονικοποιημένων συντελεστών παλινδρόμησης:

$$\hat{\beta}_{Cp} = \begin{pmatrix} 0 \\ 0 \\ 0.496 \\ 0 \\ 0.237 \\ 0.329 \\ 0 \\ 0 \\ 0.460 \\ -0.157 \\ 0.207 \\ -0.166 \\ 0.246 \end{pmatrix} \tag{5.3}$$

ενώ το μοντέλο που χρησιμοποιούμε για να προβλέψουμε την τιμή $\hat{Y}$ με διάνυσμα εισόδου το $X = (X_1, X_2, ..., X_{13})'$ είναι το εξής:

$$\hat{Y}_{Cp} = X\hat{\beta}_{Cp} = 0.496X_3 + 0.237X_5 + 0.329X_6 + 0.460X_9 - 0.157X_{10} + 0.207X_{11} - 0.166X_{12} + 0.246X_{13} \tag{5.4}$$

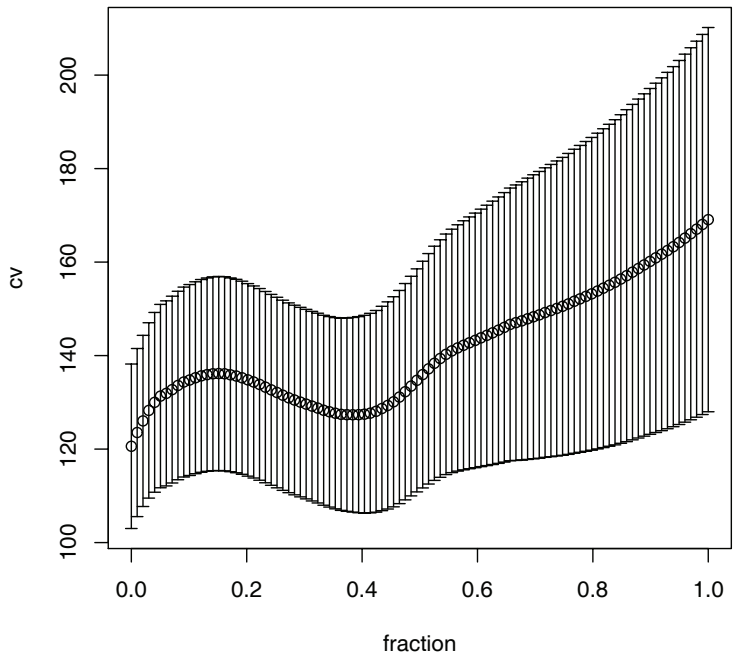
Εδώ ισχύει ότι $t < 5,148$ αφού $\sum_{i=1}^{13} |\hat{\beta}_{Cp,i}| = 2,298$. Παρατηρούμε ότι οι μεταβλητές x_3 και x_9 είναι αυτές που επηρεάζουν σε μεγαλύτερο βαθμό την τιμή της εξαρτημένης μεταβλητής y, αφού έχουν εκτιμώμενους συντελεστές ίσους με 0.496 και 0.460 αντίστοιχα. Δεν είναι τυχαίο λοιπόν που αυτές ήταν από τις πρώτες που μπήκαν στο μοντέλο. Φαίνεται ότι οι μεταβλητές x_{10} και x_{12} ασκούν τη μικρότερη επιρροή στην τιμή της y και μάλιστα αρνητική, μιας οι οι συντελεστές τους έχουν τιμές -0.157 και -0.166 αντίστοιχα. Η x_{10} μάλιστα ήταν η τελευταία που εισήλθε στο μοντέλο.

5.4. CV Κριτήριο

Στη συνέχεια, θα χρησιμοποιήσουμε ένα άλλο κριτήριο επιλογής του βέλτιστου υποσυνόλου, το $cross-validation$ *σφάλμα πρόβλεψης*, όπως αναφέρθηκε στο Κεφάλαιο 3. Το Σχήμα 5.4 απεικονίζει τη γραφική παράσταση του $cross-validation$ σφάλματος πρόβλεψης σε συνάρτηση με το $fraction$, το οποίο εκφράζει το ποσοστό της μεταβλητότητας που έχει χρησιμοποιηθεί.

Όπως υπολογίστηκε και όπως φαίνεται εξάλλου από το Σχήμα 5.4, το ελάχιστο CV σφάλμα επιτυγχάνεται στο σημείο όπου $fraction = 0.384$. Το ελάχιστο C_p που είχαμε εντοπίσει πριν, βρισκόταν στο σημείο όπου $fraction = 8/13 = 0.615$. Αν υπολογίσουμε με τη βοήθεια του R ποιές μεταβλητές θα χρησιμοποιήσουμε αυτή τη φορά στο μοντέλο μας αλλά και τους αντίστοιχους συντελεστές τους, διαπιστώνουμε ότι και αυτό το κριτήριο

μας προτείνει ακριβώς τις ίδιες μεταβλητές αλλά οι τιμές των εκτιμώμενων συντελεστών παλινδρόμησης είναι διαφορετικές. Πιο συγκεκριμένα, το διάνυσμα $\hat{\beta}_{CV}$ των συντελεστών αυτών είναι το εξής:



Σχήμα 5.4: CV προσέγγιση για τη $LARS$ μέθοδο

$$\hat{\beta}_{CV} = \begin{pmatrix} 0 \\ 0 \\ 0.444 \\ 0 \\ 0.197 \\ 0.273 \\ 0 \\ 0 \\ 0.397 \\ -0.128 \\ 0.195 \\ -0.151 \\ 0.212 \end{pmatrix} \tag{5.5}$$

δηλαδή το μοντέλο μας σε αυτή την περίπτωση είναι το:

$$\hat{Y}_{CV} = X\hat{\beta}_{CV} = 0.444X_3 + 0.197X_5 + 0.273X_6 + 0.397X_9 - 0.128X_{10} + 0.195X_{11} - 0.151X_{12} + 0.212X_{13} \tag{5.6}$$

Παρατηρούμε, λοιπόν, ότι παρόλο που τα μοντέλα αυτά έχουν ελαφρώς διαφορετικούς εκτιμώμενους συντελεστές, περιλαμβάνουν ακριβώς τις ίδιες μεταβλητές και οι αντίστοιχοι συντελεστές έχουν τα ίδια πρόσημα.

5.5. Συμπεράσματα

Ανακεφαλαιώνοντας, παρατηρούμε ότι στην προσπάθεια μας να επιλέξουμε ένα βέλτιστο υποσύνολο του συνόλου των επεξηγηματικών μεταβλητών, διαπιστώσαμε ότι και οι τρεις μέθοδοι: *LARS*, *Lasso* και *Stagewise* πρότειναν τις ίδιες μεταβλητές όταν χρησιμοποιήσαμε ως κριτήριο επιλογής το C_p.

Επιπλέον, έχοντας στη διάθεσή μας αυτό το βέλτιστο υποσύνολο μεταβλητών, εκτιμήσαμε σύμφωνα με το κριτήριο C_p τους συντελεστές αυτών των μεταβλητών και ως εκ τούτου κατασκευάσαμε ένα γραμμικό μοντέλο πρόβλεψης για την εξαρτημένη μεταβλητή Y.

Τέλος, χρησιμοποιήσαμε το CV-κριτήριο ώστε να κατασκευάσουμε ένα κατάλληλο μοντέλο. Είδαμε ότι το CV κριτήριο συμφωνεί με το C_p ως προς το ποιές μεταβλητές θα επιλέξουμε αλλά διαφέρει κατά ένα μικρό βαθμό ως προς την εκτίμηση των συντελεστών παλινδρόμησης.

ΒΙΒΛΙΟΓΡΑΦΙΑ

1. *Bancroft, T.A.*, (1944), *On Biases in Estimation due to the Use of Preliminary Tests of Significance*, *The Annals of Mathematical Statistics*, *Vol.*15, *No.*2, *pp.*190 − 204

2. *Breiman, L.* (1993) *Better subset selection using the non − negative garotte. Technical report*, *University of California*, *Berkeley*

3. *Draper, N.R.*, *Smith, H.* (1981), *Applied Regression Analysis* (2*nd ed.*), *Wiley*

4. *Efron, B.*, *Hastie, T.*, *Johnstone, I.*, *Tibshirani, R.* (2004), *Least Angle Regression*, *Annals of Statistics*, 32(2), *pp.*407 − 499

5. *Farebrother, R.W.* (1975) *The minimum mean square error linear estimator and ridge regression. Technometrics* 17, *pp.*127−8 ,ιτεμ *Frank, I. and Friedman, J.* (1993) *A statistical view of some chemometrics regression tools* (*with discussion*) *Technometrics* 35, *pp.*109 − 148

6. *Gorman, J.W.*, *Toman, R.J.* (1966), *Selection of Variables for Fitting Equations to Data*, *Technometrics*, *Vol.*8, *No.*1

7. *Hocking, R.R.* (1976), *The Analysis and Selection of Variables in Linear Regression*, *Biometrics* 32, *pp.*1 − 49

8. *Hocking, R.R.*, *Leslie, R.N.* (1967), *Selection of the Best Subset in Regression Analysis*, *Technometrics*, *Vol.*9, *No.*4

9. *Hoerl, A.E. and Kennard, R.W* (1970) *Ridge Regression*: *biased estimation for non − orthogonal problems*, *Technometrics* 12, *pp.*55 − 67

10. *Hoerl, A.E. and Kennard, R.W* (1975) *Ridge regression* : *iterative estimation of the biasing parameter*(*Preliminary report*) *IMS Bull* (*Abstract*) 4, 135

11. *Hoerl, A.E. and Kennard, R.W and Baldwin, K.F.* (1975) *Ridge regression* : *Some simulations. Comm.in Statist.* 4, *pp.*105 − 123

12. *Kennedy, W.J.*, *Bancroft, T.A.*, (1971), *Model building for prediction in regression based upon repeated significance tests*, *Annals of Mathematical Statistics*, 42(4), *pp.*1273 − 1284

13. *LaMotte, L.R. and Hocking, R.R.* (1970) *Computational efficiency in the selection of regression variables. Technometrics* 12, *pp.*83 − 93

14. *Larson, H.J.*, *Bancroft, T.A.* (1963), *Biases in Prediction by Regression for Certain Incompletely Specified Models*, *Biometrica*, 50, 3 *and* 4, *pp.*391, *Printed in Great Britain*

15. *Larson, H.J., Bancroft, T.A.* (1963), *Sequential Model Building for Prediction in Regresiion Analysis, I, Journal Paper No.J* $-$ 4567 *of the Iowa Agricultar and Home Economics Experiment Station, Ames, Iowa, Project* 169

16. *Loui, A., Tsalikaki, E., Maier, K., Walch, E., Kamarianakis, Y.,* (2006) *Growth in infants* $<$ 1500*g birthweight during the first* 5 *weeks. Submitted to the Archives of Disease in childhood.*

17. *Mallows.C.L.* (1964) *Choosing variables in a linear regression : a graphical aid. Presented at the Central Regional Meeting of the Inst. of Math.Statist., Manhattan, Kansas*

18. *Mallows, C.L.* (1973) *Some comments on Cp. Technometrics* $15, pp.661 - 75$

19. *Marquant, D.W. and Snee, R.D.* (1973) *Ridge regression, Proc.of Univ.of Kentucky conference on regression with a large number of predictor variables. Thompson, W.O and Cady, F.B.(eds.) Dept.of Statist.Univ.of Kentucky, Lexington, Kentucky*

20. *Osborne, M.R.,* (1985) *Finite Algorithms in optimization and Data Analysis, Wiley Series in Probability and Mathematical Statistics, Chichester : Wiley*

21. *Osborne, M.R., Presnell, B., Turlach, B.A.* (2000) *Journal of Computational and Graphical Statistics, Volume* $9, Number$ 2, $pp.319 - 337$

22. *Rao, C.R., Toutenburg, H.,* (1999), *Linear Models: Least Squares and Alternatives,* (2*nd ed.*), *Springer*

23. *Stein, C.M.* (1960) *Multiple regression. Contributions to Probability and Statistics. Essays in Honor of Harold Hotelling, Olkin, I(ed.), Stanford Univ. Press,* $pp.424 - 43$

24. *Stein, C.M.* (1981) *Estimation of the mean of a multivariate normal distribution. Ann.Statist.,* $9, pp.1135 - 1151$

25. *Tibshirani, R.* (1996), *Regression Shrinkage and Selection via the Lasso, J.R.Statist. Soc.B,* $58, No.1,$ $pp.267 - 288$

26. *Webster, J.T., Gunst, R.F., and Mason, R.L.* (1974) *Latent root regression analysis. Technometrics* 16, $pp.513 - 522$

27. *Weisberg, S.,* (2005) *Applied Linear Regression, third edition, Wiley*