ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Διπλωματική Εργασία

Μέθοδοι επιλογής μεταβλητών
για δεδομένα πωλήσεων από την εταιρεία **IRI**

ΚΕΦΑΛΑ ΑΝΑΣΤΑΣΙΑ

Επιβλέπων: ΦΟΥΣΚΑΚΗΣ ΔΗΜΗΤΡΙΟΣ,
Αναπληρωτής Καθηγητής Ε.Μ.Π.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον κ. Βίκτωρα Τραπουνζαλή ο οποίος σε συνεργασία με την εταιρεία **IRI** μου έδωσε το θέμα καθώς και τα δεδομένα της διπλωματικής μου εργασίας. Ακόμα θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Δημήτρη Φουσκάκη για την καθοδήγησή του και τη στήριξή του στην εκπόνηση της παρούσας διπλωματικής.

ΠΕΡΙΛΗΨΗ

Η γραμμική ανάλυση παλινδρόμησης αποτελείται από μια συλλογή από τεχνικές και μεθόδους οι οποίες χρησιμοποιούνται για να ερευνήσουν και να εξηγήσουν τις πιθανές σχέσεις μεταξύ μεταβλητών (χαρακτηριστικών). Ωστόσο, ένα από τα πλέον σημαντικά προβλήματα που απασχολεί την ανάλυση παλινδρόμησης είναι η επιλογή ενός μικρότερου συνόλου από το αρχικό σύνολο των ανεξάρτητων μεταβλητών που είναι διαθέσιμες κάθε φορά στο εκάστοτε γραμμικό μοντέλο. Με την δημιουργία αυτού του υποσυνόλου επιτυγχάνεται από τη μία εξοικονόμηση κόστους κατά την πρόβλεψη της εξαρτημένης μεταβλητής, και από την άλλη αποτρέπεται η μεγάλη απώλεια στην αποτελεσματικότητα του μοντέλου πρόβλεψης.

Στη παρούσα εργασία παρουσιάζεται ένα πλήθος από μεθόδους και κριτήρια για την επιλογή ενός «βέλτιστου» συνόλου ανεξάρτητων μεταβλητών για την πρόβλεψη μιας εξαρτημένης μεταβλητής μέσω των επεξηγηματικών μεταβλητών του εκάστοτε γραμμικού μοντέλου, με ιδιαίτερη έμφαση στις νέες ποινικοποιημένες μεθόδους.

Το πρώτο κεφάλαιο, αποτελεί μια εισαγωγή στο πολλαπλό γραμμικό μοντέλο και στη βασική τεχνική εκτίμησης των συντελεστών του γραμμικού μοντέλου. Στη συνέχεια, παρουσιάζεται η μέθοδος της εξέτασης όλων των δυνατών μοντέλων του γραμμικού χώρου που εξετάζεται, δηλαδή όλων των δυνατών συνδυασμών των διαθέσιμων επεξηγηματικών μεταβλητών (*All Possible Regressions* ή *Full Enumeration*), δίνοντας αναλυτικά τα διάφορα κριτήρια που έχουν προταθεί για τον εντοπισμό του κατάλληλου μοντέλου καθώς και μία παραλλαγή αυτής, τη μέθοδο επιλογής καλύτερου υποσυνόλου (*Best Subset Selection*).

Στη συνέχεια παρουσιάζονται οι πιο γνωστές μέθοδοι επιλογής μεταβλητών, οι διαδικασίες κατά βήματα. Με αυτές δημιουργείται μια

αλληλουχία γραμμικών μοντέλων εισάγοντας ή αφαιρώντας κάθε φορά μια επεξηγηματική μεταβλητή ή συνδυάζοντας και τα δύο, μέχρις ότου να φτάσουν σε κάποιο σημείο όπου σύμφωνα με ορισμένα κριτήρια που χρησιμοποιούμε να μην μπορούν να εισαχθούν ή να εξαχθούν πλεόν άλλες επεξηγηματικές μεταβλητές από το εξεταζόμενο μοντέλο.

Έπειτα, παρουσιάζουμε ένα συνηθισμένο φαινόμενο, όταν έχουμε μεγάλο αριθμό επεξηγηματικών μεταβλητών, το πρόβλημα της πολυσυγγραμμικότητας. Καταλήγουμε έτσι στην ανάλυση των ποινικοποιημένων μεθόδων ή αλλιώς μεθόδων συρρίκνωσης *Ridge* και *LASSO*, αναλύοντας την θεωρία που τις ορίζει και τις υλοποιεί.

Στο τελευταίο μέρος γίνεται η εφαρμογή των μεθόδων εντοπισμού του «βέλτιστου» συνόλου επεξηγηματικών μεταβλητών σε ένα γραμμικό μοντέλο και γίνεται σύγκριση των μεθόδων που παρουσιάστηκαν στα πλαίσια της παρούσας διπλωματικής. Το δείγμα που χρησιμοποιήσαμε αποτελείται από πραγματικά δεδομένα που αφορούν τον όγκο πωλήσεων ενός προϊόντος (μεταβλητή απόκρισης) και μεταβλητές που υποθέτουμε ότι τις ερμηνεύουν (τις πωλήσεις αυτές). Το πλήθος των διαθέσιμων παρατηρήσεων είναι 288 εβδομάδες και 141 το πλήθος των μεταβλητών που θα χρησιμοποιηθούν.

ABSTRACT

Linear regression analysis consists of a collection of techniques and methods used to investigate and explain the possible relationships between variables. However, one of the most important problems involved in balancing analysis is the selection of a smaller set of the original set of independent variables available each time in the particular linear model. By creating this subset, on the other hand, cost savings are made in the prediction of the dependent variable, and on the other it avoids a great loss in the efficiency of the prediction model.

This paper presents a set of methods and criteria for selecting an "optimal" set of independent variables to predict a dependent variable through the explanatory variables of the particular linear model, with particular emphasis on new penalized methods.

The first chapter is an introduction to the multiple linear model and the basic technique for estimating the coefficients of the linear model. Then, we present the method of "examining all possible models", all possible combinations of independent variables (*All Possible Regressions*), analyzing the various criteria proposed for identifying the appropriate model as well as a variation of this, the method of Best Subset Selection.

Below, we present the widely known methods of selecting variables, the procedures in steps. These procedures create a sequence of linear models by inserting or removing an explanatory variable each time or combining both until they reach a point where, according to certain criteria we use, no other explanatory variables can be imported or extracted by the model that we examine.

Then, we present a common phenomenon, when we have a large number of explanatory variables, the problem of multicollinearity. We conclude in the analysis of penalized or shrinking methods *Ridge* and *LASSO*, analyzing the theory that defines them and implements them.

In the last part, we apply the methods of locating the optimal set of explanatory variables in a linear model and compare the methods presented in the present thesis. The sample we used consists of actual sales volume data for a product (response variable) and variables that we assume that they interpret (these sales). The number of available observations is 288 weeks and 141 the number of variables to be used.

Περιεχόμενα

4 Τεχνικές με ποινή

5 Εφαρμογή μεθόδων με χρήση της γλώσσας στατιστικού προγραμματισμού *R*

ΚΕΦΑΛΑΙΟ 1

Θεωρητικό υπόβαθρο

1.1 Εισαγωγή

Κατά διάρκεια της ζωής του ο κάθε άνθρωπος καλείται να πάρει αποφάσεις και να κάνει επιλογές. Ωστόσο, για να παρθεί κάθε φορά η σωστότερη απόφαση πρέπει να λαμβάνεται υπόψιν το εκάστοτε πλαίσιο αναφοράς, οι συνθήκες και οι ζητούμενες ανάγκες. Από αυτά γίνεται αντιληπτό ότι το να πάρει κάποιος μια απόφαση δεν αποτελεί μία απλή διαδικασία μιας και όλα τα παραπάνω κριτήρια θα πρέπει να προσμετρώνται.

Η Στατιστική είναι η επιστήμη αυτή που επιχειρεί να εξάγει μία έγκυρη γνώση χρησιμοποιώντας εμπειρικά δεδομένα παρατηρήσεων ή πειραμάτων. Η χρήση της καθίσταται απαραίτητη γενικότερα στη Διοίκηση, όπου η ορθή λήψη αποφάσεων είναι άρρικτα συνδεδεμένη με την πρόοδο ενός κράτους ή ακόμα και μιας επιχείρησης. Ακόμα η έρευνα αγοράς, η οποία προϋποθέτει τη μελέτη της συμπεριφοράς και των συνηθειών των καταναλατών, έχει γίνει επιτακτική ανάγκη στη σύγχρονη κοινωνία. Πάνω σε αυτόν τον τομέα θα ασχοληθούμε στην παρούσα διπλωματική εργασία, καθώς τα δεδομένα που διαθέτουμε προέρχονται από τον κλάδο της αγοράς.

Με απώτερο στόχο λοιπόν την εξαγωγή βάσιμων συμπερασμάτων η οποία συνεπάγεται την ορθή λήψη αποφάσεων, ο ερευνητής καταφεύγει στην κατασκευή και ανάλυση στατιστικών μοντέλων, δηλαδή στην

δημιουργία και μελέτη κάποιων σχέσεων οι οποίες είναι κατασκευασμένες με τέτοιο τρόπο ώστε να περιγράφουν τη συμπεριφορά του προς μελέτη χαρακτηριστικού.

Ωστόσο, αρκετά συχνά εμφανίζεται η ανάγκη να μελετήσουμε ταυτόχρονα δύο ή περισσότερα χαρακτηριστικά, με στόχο να κατανοήσουμε τον τρόπο με τον οποίο τα χαρακτηριστικά αυτά συνδέονται μεταξύ τους. Για παράδειγμα:

- το ύψος του μισθού που λαμβάνει κάθε υπαλλήλος μιας εταιρείας εξαρτάται από το χρόνο υπηρεσίας του.

- το πλήθος των προιόντων που θα πουλήσει μια εταιρεία σχετίζεται με το μέγεθος της διαφημιστικής δαπάνης που θα κάνει για το συγκεκριμένο προιόν ώστε να προσελκύσει όσο το δυνατόν περισσότερο καταναλωτικό κοινό.

- η απόδοση (βαθμοί) ενός μαθητή εξαρτάται από τις ώρες που αφιερώνει στη μελέτη.

Στα προβλήματα αυτά παρουσιάζεται ενδιαφέρον να εξεταστεί πως ένα χαρακτηριστικό επηρεάζει ένα άλλο (θετικά ή αρνητικά). Ο τομέας αυτός της Στατιστικής που ασχολείται με την εξέταση της σχέσης μεταξύ δύο ή περισσότερων μεταβλητών με στόχο την πρόβλεψη μιας από αυτές με χρήση των τιμών μιας ή περισσοτέρων άλλων ονομάζεται Ανάλυση Παλινδρόμησης και τα μοντέλα που αναπτύσσονται μέσω αυτής καλούνται μοντέλα παλινδρόμησης.

1.2 Πολλαπλό γραμμικό μοντέλο

Τα μοντέλα παλινδρόμησης, όπως αναφέραμε, δημιούργηθηκαν και χρησιμοποιούνται για να περιγράψουν το πώς ένα χαρακτηριστικό εξαρτάται ή επηρεάζει ένα άλλο. Για την κατασκευή ενός μοντέλου παλινδρόμησης συνίσταται η εύρεση μιας μαθηματικής-συναρτησιακής σχέσης η οποία να συνδέει ένα χαρακτηριστικό X με ένα άλλο Y. Αυτός ο τύπος δίνεται μέσω μιας συνάρτησης $y = f(x)$ και εκφράζει τον τρόπο υπολογισμού της τιμής y που έχει το χαρακτηριστικό Y μέσω της τιμής x που λαμβάνει το χαρακτηριστικό X. Αυτές οι χαρακτηριστικές ιδιότητες

ενός πληθυσμού, με τη μελέτη των οποίων ασχολείται η Στατιστική, ονομάζονται μεταβλητές.

Η μεταβλητή X λέγεται ανεξάρτητη μεταβλητή ή επεξηγηματική μεταβλητή, αφού η τιμή της x καθορίζεται από εμάς και παρέχει πληροφορία για τη συμπεριφορά της μεταβλητής Y, η οποία καλείται εξαρτημένη μεταβλητή ή μεταβλητή απόκρισης, αφού η τιμή y που θα πάρει εξαρτάται από την τιμή που έχουμε δώσει στην ανεξάρτητη μεταβλητή X.

Για να εκφράσουμε την παραπάνω μορφή εξάρτησης μεταξύ των μεταβλητών Y και X (για παράδειγμα για να αποδώσουμε το ύψος του μισθού κάθε υπαλλήλου μιας εταιρείας σχετικά με το χρόνο υπηρεσίας του) θα γράφουμε:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

όπου:

x: η συγκεκριμένη τιμή που πήρε η μεταβλητή X (ο χρόνος υπηρεσίας υπαλλήλου),

y: τιμή της εξαρτημένης μεταβλητής που αντιστοιχεί στην τιμή x της X (ο μισθός του υπαλλήλου) και

ε: μια τυχαία μεταβλητή που περιγράφει την «απόκλιση» της y από το γραμμικό όρο $\beta_0 + \beta_1 x$.

Έτσι προκύπτει το απλό γραμμικό μοντέλο:

$$y = \beta_0 + \beta_1 x + \varepsilon$$

το οποίο περιέχει μία μόνο επεξηγηματική μεταβλητή x.

Ωστόσο, τις περισσότερες φορές καλούμαστε να μελετήσουμε περισσότερες από μία επεξηγηματικές μεταβλητές $x_j, j = 1, \dots, p$. Έτσι προκύπτει το πολλαπλό γραμμικό μοντέλο για την i παρατήρηση:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i, \forall i = 1, \dots, n.$$

$\boldsymbol{\beta}'$=$(\beta_0, \beta_1, \beta_2, \dots, \beta_p)$: άγνωστες παράμετροι του μοντέλου (ή συντελεστές) οι οποίες και πρέπει να εκτιμηθούν,

$\boldsymbol{y_i}$: τιμή της μεταβλητής απόκρισης (ή εξαρτημένης μεταβλητής) κατά την i επανάληψη του πειράματος,

$\boldsymbol{x}'$ =$(x_0, x_{i1}, x_{i2}, \dots, x_{ip})$ ένας $n \times (p+1)$ πίνακας με $x_0 = 1$: είναι οι τιμές των επεξηγηματικών μεταβλητών (γνωστοί αριθμοί) ή ανεξάρτητη μεταβλητή και

ε_i: τυχαία σφάλματα που ακολουθούν τη $N(0, \sigma^2)$. Ακόμα τα σφάλματα ε_i και ε_j που αντιστοιχούν σε διαφορετικές επαναλήψεις του πειράματος θωρούνται ασυσχέτιστα, δηλαδή ισχύει ότι $Cov(\varepsilon_i, \varepsilon_j) = 0$.

Τέλος, το πολλαπλό γραμμικό μοντέλο μπορεί να γραφεί υπό μορφή πινάκων ως:

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon}$$

$$\begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1n} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{pmatrix} \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix} + \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

όπου:

$\boldsymbol{X}$: ένας $n \times p$ πίνακας και ονομάζεται πίνακας σχεδιασμού,

$\boldsymbol{\beta}$=$(\beta_0, \beta_1, \dots, \beta_p)$': το $p \times 1$ διάνυσμα των άγνωστων παραμέτρων και

$\boldsymbol{\varepsilon}$=$(\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)$': το $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων.

Έχουν αναπτυχθεί διάφορες τεχνικές οι οποίες προσπαθούν να εκτιμήσουν τη τιμή του $\boldsymbol{\beta}$ μέσω μίας εκτιμήτριας $\widehat{\boldsymbol{\beta}}$. Πιο συγκεκριμένα, δεδομένου ενός συνόλου από n παρατηρήσεις, εφαρμόζουμε μία τεχνική η οποία παράγει μία εκτιμήτρια $\widehat{\boldsymbol{\beta}}$ του πραγματικού διανύσματος $\boldsymbol{\beta}$. Επειδή όμως η $\widehat{\boldsymbol{\beta}}$ εξαρτάται από το σύνολο των n παρατηρήσεων,

καταλαβαίνουμε ότι εάν το σύνολο αυτό υποστεί αλλαγές, θα αλλάξει και η τιμή της εκτιμήτριας αυτής. Υπό αυτή την έννοια, η εκτιμήτρια $\widehat{\boldsymbol{\beta}}$ αποτελεί τυχαία μεταβλητή.

Γενικά, κάθε ερευνητής μελετάει τη μεροληψία $(bias)$ της εκτιμήτριας, η οποία αποτυπώνει τη διαφορά μεταξύ της αναμενόμενης τιμής της εκτιμήτριας και της πραγματικής τιμής της παραμέτρου που εκτιμάει, δηλαδή:

$$bias = E\left(\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{\beta}.$$

(1.1)

Ωστόσο, παρακάτω θα δούμε την πιο συνηθισμένη τεχνική εκτίμησης η οποία παράγει αμερόληπτες εκτιμήτριες (όπου $bias = 0$ και από τον τύπο ***(1.1)*** έχουμε $E\left(\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{\beta}$), ώστε να καταλήξουμε (Κεφάλαιο 4) στην κατασκευή μεροληπτικών εκτιμητριών.

1.3 Εκτίμηση των παραμέτρων του μοντέλου με τη μέθοδο ελαχίστων τετραγώνων

Η εκτίμηση των παραμέτρων του μοντέλου, δηλαδή των β_j, $\forall j = 1, \dots, p$ μπορεί να γίνει με τη μέθοδο των ελαχίστων τετραγώνων (*least squares method*), η οποία προτάθηκε από τους *Legendre* και *Gauss*. Το 1821 ο *Gauss* εισήγαγε μια βελτιωμένη θεωρία για τη μέθοδο των ελαχίστων τετραγώνων, η οποία είναι γνωστή σήμερα ως θεώρημα του *Gauss-Markov*. Σύμφωνα με τη θεωρία αυτή οι εκτιμήτριες ελαχίστων τετραγώνων είναι *blue* (*best linear unbiased estimators*), δηλαδή αποτελούν τις καλύτερες (έχουν τη μικρότερη διασπορά στη κλάση των αμερόληπτων εκριμητριών) γραμμικές αμερόληπτες (απόδειξη παρακάτω) εκτιμήτριες. Το θεώρημα αυτό, αποτελεί θεμελιώδες θεώρημα στον τομέα των γραμμικών μοντέλων. Ωστόσο, όπως θα δούμε στο Κεφάλαιο 4, οι εκτιμήτριες αυτές αποτυγχάνουν να εκτιμήσουν με σωστό τρόπο τους συντελεστές του μοντέλου κατά την ύπαρξη του φαινομένου της πολυσυγγραμμικότητας, όπου και καταφεύγουμε στη κατασκευή και χρήση μεροληπτικών εκτιμητριών.

Το κριτήριο που εφαρμόζεται για τη μέθοδο ελαχίστων τετραγώνων είναι πολύ απλό και συνίσταται στην ελαχιστοποίηση του αθροίσματος τετραγώνων των σφαλμάτων :

$$SSE(\beta_0, \beta_1, \dots, \beta_p) = \sum_{i=1}^{n} \varepsilon_i^2$$

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

$$= \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}\right)^2,$$

y_i: οι παρατηρούμενες τιμές της μεταβλητής απόκρισης,

$\hat{y}_i$: οι προβλεπόμενες τιμές της.

Τελικά προκύπτουν οι εκτιμήτριες $\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p$.

Με τη μορφή πινάκων, για να βρεθεί η εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$, πρέπει να ελαχιστοποιηθεί το άθροισμα των τετραγώνων των σφαλμάτων:

$$\begin{aligned} SSE(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \varepsilon_i^2 \\ &= \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} \\ &= (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \\ &= (\boldsymbol{y}' - \boldsymbol{\beta}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X\beta}) \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X\beta}. \end{aligned}$$

(1.2)

Να σημειώσουμε εδώ ότι στην προτελευταία σειρά της ***(1.2)*** ο δεύτερος και ο τρίτος όρος είναι ίσοι (ένας 1×1 πίνακας είναι πάντα συμμετρικός) και μπορούν να αντικατασταθούν από τον όρο $-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$.

Παραγωγίζοντας έτσι την ***(1.2)*** ως προς $\boldsymbol{\beta}$ έχουμε:

$$\frac{\partial SSE(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta}.$$

(1.3)

Ο τελευταίος όρος στην ***(1.3)*** προκύπτει από το γεγόνός ότι σύμφωνα με τις ιδιότητες πινάκων για την παραγώγιση έχουμε:

$$\frac{\partial(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}.$$

Τέλος, θέτοντας την παράγωγο ίση με μηδέν, $\frac{\partial SSE(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0$, παίρνουμε:

$$\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} = \boldsymbol{X}'\boldsymbol{y}.$$

Αν ο πίνακας $\boldsymbol{X}'\boldsymbol{X}$ αντιστρέφεται τότε η εκτιμήτρια ελαχίστων τετραγώνων του πραγματικού διανύσματος **β**, δίνεται από τη σχέση:

$$\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Οπότε και το μοντέλο που εκτιμούμε δίνεται από τη σχέση:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_{LS}$$
$$= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}$$
$$= \boldsymbol{H}\boldsymbol{y},$$

με $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$.

Ο $\boldsymbol{H}$ ονομάζεται πίνακας προβολής ή *hat matrix* και αποτελεί συμμετρικό πίνακα ($\boldsymbol{H}' = \boldsymbol{H}$).

Ενώ οι βαθμοί ελευθερίας δίνονται από το ίχνος του πίνακα προβολής:

$$tr(\boldsymbol{H}) = p.$$

Τέλος, θα μελετήσουμε κάποιες ιδιότητες της εκτιμήτριας ελαχίστων τετραγώνων.

i) Αναμενόμενη τιμή

Η πιο σημαντική ιδιότητά της, όπως αναφέραμε, είναι ότι αποτελεί αμερόληπτη εκτιμήτρια του $\boldsymbol{\beta}$, δηλαδή $E(\widehat{\boldsymbol{\beta}}_{LS}) = \boldsymbol{\beta}$.

Απόδειξη:

$$E(\widehat{\boldsymbol{\beta}}_{LS}) = E((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y})$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{y})$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon})$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon})$$

$$= \boldsymbol{\beta},$$

αφού $(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$ και $E(\boldsymbol{\varepsilon})$= $\mathbf{0}$.

ii) Διασπορά

Η διασπορά της εκτιμήτριας ελαχίστων τετραγώνων δίνεται από τον τύπο:

$$Var(\widehat{\boldsymbol{\beta}}_{LS}) = E\left(\left(\widehat{\boldsymbol{\beta}}_{LS} - E(\widehat{\boldsymbol{\beta}}_{LS})\right)\left(\widehat{\boldsymbol{\beta}}_{LS} - E(\widehat{\boldsymbol{\beta}}_{LS})\right)'\right)$$

$$= E\left((\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta})'\right)$$

$$= E\left((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})((\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon})'\right),$$

αφού $\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'y - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'(\boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}) - \boldsymbol{\beta} = (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{\varepsilon}$.

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}')\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$= (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\sigma^2\boldsymbol{I}\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

1.4 Επιλογή μεταβλητών

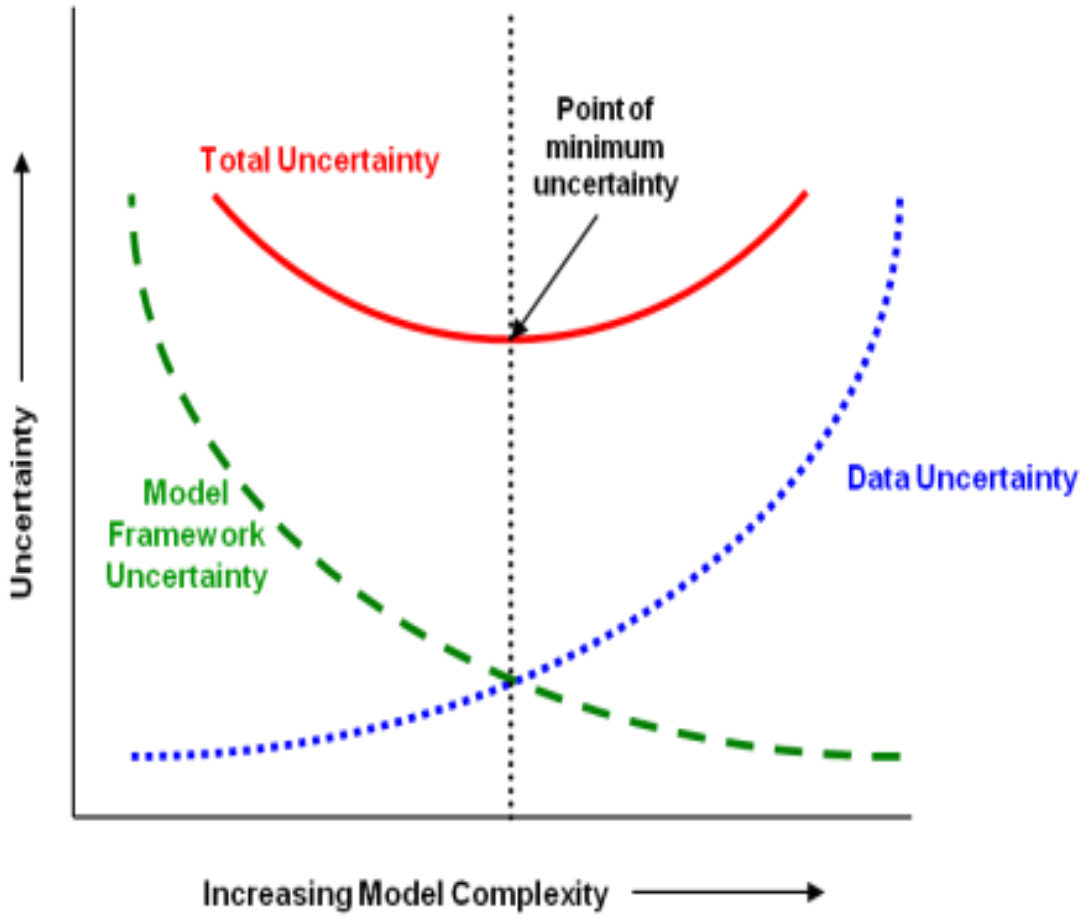
Η προσαρμογή του εκάστοτε στατιστικού μοντέλου, έγκειται στο σύνολο δεδομένων της εξαρτημένης μεταβλητής και των επεξηγηματικών μεταβλητών που έχουμε στη διάθεση μας. Όσο καλύτερη είναι η προσαρμογή του μοντέλου, με βάση τα στατιστικά δεδομένα που διαθέτουμε, τόσο καλύτερα εκτιμούνται οι συντελεστές του μοντέλου παλινδρόμησης.

Ωστόσο ένα από τα κύρια ζητήματα που προκύπτουν κατά τη μελέτη των γραμμικών μοντέλων είναι να καθορίσουμε το σύνολο των επεξηγηματικών μεταβλητών που θα εισαχθούν στο εκάστοτε μοντέλο έτσι ώστε να καταλήξουμε στο ποιο θα είναι το «καταλληλότερο». Αυτό αποτελεί και το κεντρικό θέμα της παρούσας διπλωματικής και στη συνέχεια θα παρουσιαστούν αναλυτικά όλα τα κριτήρια και οι μέθοδοι που θα χρησιμοποιήσουμε ώστε να φτάσουμε κοντά στο «βέλτιστο» μοντέλο.

Η επιλογή μεταβλητών κατέχει κεντρικό ρόλο στη σύγχρονη στατιστική μάθηση, αφού έχει ως σκοπό να εντοπίσει όλες τις πραγματικά σημαντικές μεταβλητές, δηλαδή τις μεταβλητές αυτές που οι συντελεστές τους παλινδρόμησης δεν εξαφανίζονται και να παρέχει αποτελεσματικές εκτιμήσεις για αυτούς τους συντελεστές. Σκοπός δηλαδή είναι να επιλεγούν οι παράγοντες που έχουν σημαντική επίδραση στην μεταβλητή απόκρισης. Με αυτόν τον τρόπο μας δίνεται η δυνατότητα να εξηγήσουμε τα δεδομένα με τον απλούστερο δυνατό τρόπο. Όμως, ενώ από τη μία η αύξηση των μεταβλητών οδηγεί ταυτόχρονα σε αύξηση της διασποράς των εκτιμήσεων, από την άλλη η μείωση του αριθμού τους έχει ως συνέπεια την αύξηση της μεροληψίας των εκτιμήσεων.

Η αρχή της φειδωλότητας (*Law of parsimony*) (***Σχήμα 1.1***) είναι αυτή που προσπαθεί να «παντρέψει» τις δύο προηγούμενες ακραίες περιπτώσεις και δηλώνει πως το τελικό μοντέλο θέλουμε να συμπεριλαμβάνει όσο το δυνατόν μικρότερο πλήθος επεξηγηματικών μεταβλητών, οι οποίες όμως ταυτόχρονα να περιγράφουν με τον καλύτερο δυνατό τρόπο τα δεδομένα μας. Άλλωστε, σύμφωνα με τον Άινσταιν «όλα πρέπει να γίνονται όσο πιο απλά είναι δυνατόν, αλλά όχι απλούστερα». Έτσι και η εφαρμογή της αρχής της φειδωλότητας στην ανάλυση παλινδρόμησης

αποδεικνύει ότι το μικρότερο μοντέλο που ταιριάζει με τα δεδομένα που έχουμε στη διάθεσή μας, δηλαδή εξηγεί σωστά την συμπεριφορά της μεταβλητής απόκρισης, είναι και το πιο κατάλληλο. Επομένως, λαμβάνοντας υπόψιν μόνο τους στατιστικά σημαντικούς παράγοντες γλιτώνουμε χρόνο και χρήμα κατά τη διάρκεια μιας μελέτης, καθώς ο αριθμός τους είναι αρκετά μικρότερος σε σχέση με το αρχικό σύνολό μας.



***Σχήμα 1.1*: Law of parsimony**

1.5 Σκοπός διπλωματικής

Σκοπός της παρούσας διπλωματικής είναι να εξεταστούν μέθοδοι επιλογής μεταβλητών για την ερμηνεία μιας μεταβλητής απόκρισης για την οποία υποθέτουμε ότι υπάρχει σχέση αιτίας-αιτιατού με τις επεξηγηματικές μεταβλητές.

Πιο συγκεκριμένα, θα εξεταστεί η απλή γραμμική σχέση μεταξύ τους, βάσει του κλασικού μοντέλου ανάλυσης παλινδρόμησης, καθώς και κάποιες μέθοδοι επιλογής επεξηγηματικών μεταβλητών. Οι μέθοδοι που θα εξεταστούν είναι:

- η μέθοδος της προς τα εμπρός επιλογής (*Forward Selection*)
- η μέθοδος της προς τα πίσω απαλοιφής (*Backward Elimination*)
- η μέθοδος της κατά βήματα παλινδρόμησης (*Stepwise Regression*)
- η *Ridge*
- η *LASSO*.

Οι τρεις πρώτες αποτελούν βηματικές μεθόδους, οι οποίες υλοποιούνται μέσω αλγορίθμων ενώ οι δύο τελευταίες είναι πιο βελτιωμένες μέθοδοι και αποτελούν τεχνικές της ποινικοποιημένης παλινδρόμησης. Δίνουμε περισσότερο έμφαση στις δύο τελευταίες μεθόδους και τις συγκρίνουμε ως προς τα αποτελέσματα που δίνουν βάσει στατιστικών που αφορούν την ερμηνευτική ικανότητα των μοντέλων που θα προκύψουν.

Θα χρησιμοποιηθούν για την διερεύνηση των παραπάνω, δεδομένα που αφορούν τον όγκο πωλήσεων ενός προϊόντος (μεταβλητή απόκρισης) και μεταβλητές που υποθέτουμε ότι τις ερμηνεύουν (τις πωλήσεις αυτές). Τα δεδομένα αυτά είναι πραγματικά και μου δόθηκαν από την εταιρεία *IRI* στα πλαίσια της πρακτικής μου άσκησης. Το πλήθος των διαθέσιμων παρατηρήσεων είναι 288 εβδομάδες και 141 το πλήθος των μεταβλητών που θα χρησιμοποιηθούν.

ΚΕΦΑΛΑΙΟ 2

Πλήρης εξερεύνηση του χώρου

2.1 Εισαγωγή

Πολλές φορές, ο αναλυτής καλείται να έρθει αντιμέτωπος με ένα αρκετά μεγάλο σύνολο από επεξηγηματικές μεταβλητές προκειμένου να μελετήσει την επίδρασή τους στην μεταβλητή απόκρισης. Αν, λοιπόν, επέλεγε να χρησιμοποιήσει όλες τις μεταβλητές θα κατέληγε σε μια πολυέξοδη και μακροσκελή διαδικασία, δεδομένου του όγκου των υπολογισμών που πρέπει να πράξει. Επομένως, εμφανίζεται μάλλον λογικότερο να αναζητήσουμε ένα μικρότερο υποσύνολο, σε σχέση με το αρχικό ογκώδες σύνολο, διατηρώντας ταυτόχρονα όσο γίνεται την ίδια λειτουργικότητα του τελικού αποτελέσματος.

Με στόχο την δημιουργία αυτών των μικρότερων υποσυνόλων, ο αναλυτής είναι αναγκασμένος να καταφύγει στην απαλοιφή ενός αριθμού επεξηγηματικών μεταβλητών από το αρχικό πλήρες μοντέλο. Το κίνητρο της διαδικασίας αυτής, εκτός του ότι είναι να αντιμετωπίσουμε μία οικονομικά ασύμφορη διαδικασία, είναι η βελτίωση της ακρίβειας των εκτιμήσεων των παραμέτρων των επεξηγηματικών μεταβλητών που παραμένουν στο μοντέλο, ακόμα και όταν κάποιες από τις μεταβλητές που αφαιρέθηκαν δεν είναι ασήμαντες.

Για την εύρεση αυτών των υποσυνόλων η χρήση των στατιστικών πακέτων για έναν αναλυτή αποτελεί μονόδρομο, ειδικότερα όταν έχει στη διάθεσή του μεγάλο αριθμό επεξηγηματικών μεταβλητών. Στις μέρες μας, τα περισσότερα στατιστικά πακέτα παρέχουν στον χρήστη την επιλογή βέλτιστου μοντέλου επεξηγηματικών μεταβλητών. Αυτή όμως η αυτόματη, άκοπη και εύκολη διαδικασία οδηγεί στο λεγόμενο «τυφλό σύστημα», όπου ο χρήστης δεν είναι γνώστης της υλοποίησης των μεθόδων. Αυτή η διαδικασία οδηγεί με τη σειρά της σε ένα μοντέλο που τις περισσότερες φορές δεν μοιάζει να είναι λειτουργικό και λογικό. Επομένως, κρίνεται αναγκαία η κατασκευή μιας ευρύτερης γκάμας εναλλακτικών μοντέλων, η οποία θα δίνει τη δυνατότητα στον χρήστη να επιλέξει, σύμφωνα με την εμπειρία και τη γνώση του, αυτό το μοντέλο που κρίνεται βέλτιστο για τις ανάγκες του και θα επεξηγεί όσο το δυνατόν καλύτερα την μεταβλητή απόκρισης που εξετάζει.

Γι΄αυτό το λόγο στη συνέχεια θα παρουσιάσουμε μεθόδους εντοπισμού κατάλληλων υποσυνόλων ανεξάρτητων μεταβλητών, η εφαρμογή των οποίων ναι μεν είναι εύκολη με τη χρήση στατιστικών πακέτων, αλλά ο χρήστης θα πρέπει να καταλήγει στην τελική επιλογή χρησιμοποιώντας και την κρίση του. Αρχικά θα αναπτύξουμε τη μέθοδο της εξέτασης όλων των δυνατών γραμμικών μοντέλων, η οποία εφοδιάζει τον αναλυτή με μία ποικιλία από μοντέλα και την μέθοδο καλύτερου υποσυνόλου. Στη συνέχεια (Κεφάλαιο 3) θα περιγραφούν οι πιο διαδεδομένες μέθοδοι, οι διαδικασίες κατά βήματα οι οποίες κατασκευάζουν αυτόματα υποσύνολα επεξηγηματικών μεταβλητών.

Κατά την υλοποίηση αυτών των μεθόδων όλες οι διαθέσιμες επεξηγηματικές μεταβλητές μπορούν να εισαχθούν καθώς και να εξαχθούν από το αρχικό μας μοντέλο, βασιζόμενοι σε κριτήρια που θα παρουσιάσουμε παρακάτω.

Συνοπτικά, να σημειώσουμε ότι οι διαδικασίες κατά βήματα είναι πιο γρήγορες από άποψη υπολογιστικού κόστους και σε αρκετές περιπτώσεις τα υποσύνολα που δημιουργούνται από την υλοποίησή τους προσεγγίζουν το βέλτιστο υποσύνολο που δίνει η μέθοδος της εξέτασης όλων των γραμμικών μοντέλων. Βέβαια, η μέθοδος αυτή έχει ένα σοβαρό μειονέκτημα, όπως θα δούμε και παρακάτω.

2.2 Κριτήρια επιλογής βέλτιστου στατιστικού μοντέλου

Ενώ η θεωρία μας δίνει μια γενική κατεύθυνση ως προς τις ποιες από τις υποψήφιες επεξηγηματικές μεταβλητές μπορούν να συμπεριληφθούν στο μοντέλο, το πραγματικό πλήθος τυχαίων μεταβλητών που χρησιμοποιούνται δίνεται αποκλειστικά από την ανάλυση δεδομένων και συγκεκριμένα από αυτό που ονομάζουμε επιλογή του καλύτερου γραμμικού μοντέλου με τη χρήση γραμμικής παλινδρόμησης.

Όμως όπως αναφέραμε στην αρχή της φειδωλότητας για να επιλέξουμε ένα γραμμικό μοντέλο θα πρέπει να επιτευχθεί η «χρυσή τομή» ανάμεσα στις δύο αλληλοσυγκρουόμενες απαιτήσεις που αφορούν τον αριθμό των επεξηγηματικών μεταβλητών που πρέπει να συμπεριληφθούν στο μοντέλο. Έτσι, επιδιώκουμε την εφαρμογή μεθόδων για την επιλογή του καλύτερου υποσυνόλου μεταβλητών, η οποία να στοχεύει στην ισορροπία μεταξύ της απλότητας του μοντέλου και της καλής προσαρμογής των δεδομένων που έχουμε στη διάθεσή μας.

Για να εντοπίσουμε αυτό το ιδανικό μοντέλο ή πιο σωστά για να φτάσουμε όσο το δυνατόν πιο κοντά στην εύρεση του «κατάλληλου» μοντέλου δεν υπάρχει μία μόνο διαδικασία. Αρκετά κριτήρια στα οποία ο ερευνητής βασίζεται για την επιλογή υποσυνόλων έχουν προταθεί. Τα κριτήρια αυτά θεωρούνται αποτελεσματικά καθώς βασίζονται στην αρχή της φειδωλότητας. Παρ' όλα αυτά, πρέπει να σημειώσουμε ότι η επιλογή του κάθε κριτηρίου ίσως οδηγεί σε διαφορετικά υποσύνολα.

Τα πιο συνηθισμένα κριτήρια που έχουν αναπτυχθεί για την υλοποίηση των μεθόδων που χρησιμοποιούμε για να αποφανθούμε για το ποιες μεταβλητές θα συμπεριλάβουμε στο εκάστοτε γραμμικό μοντέλο είναι τα ακόλουθα:

- Συντελεστής προσδιορισμού R^2
- Τροποποιημένος συντελεστής προσδιορισμού R^2_{Adj}
- Άθροισμα τετραγώνων των σφαλμάτων SSE
- Μέσο άθροισμα τετραγώνων των σφαλμάτων MSE
- Κριτήριο AIC

- Κριτήριο BIC

Σε αυτό το σημείο καλό θα ήταν να εισάγουμε κάποιες ποσότητες ώστε να ορίσουμε τα κριτήρια που αναφέραμε:

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

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

$$SST = \sum_{i=1}^{n} (y_i - \bar{y})^2, i = 1, \dots, n$$

όπου:

$\hat{y}_i$: η προβλεπόμενη τιμή της εξαρτημένης μεταβλητής και

$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$: η μέση τιμή της.

Η σχέση που συνδέει και τις τρεις αυτές ποσότητες είναι:

$$SST = SSE + SSR.$$

Το SSE, όπως έχουμε αναφέρει και στο πρώτο κεφάλαιο κατά τη μέθοδο ελαχίστων τετραγώνων, αποτελεί το άθροισμα τετραγώνων των σφαλμάτων. Λαμβάνει μη αρνητικές τιμές και γίνεται μηδέν όταν δεν υπάρχει σφάλμα πρόβλεψης δηλαδή όταν οι παρατηρούμενες τιμές συμπίπτουν με τις προβλεπόμενες. Ενώ το SSR αποτελεί το άθροισμα τετραγώνων της παλινδρόμησης.

Όμως το γεγονός ότι δεν υπάρχει κάποιο κριτήριο το οποίο να καθορίζει ποιες τιμές του SSE θεωρούνται μικρές και ποιες μεγάλες, οδηγεί στην ανάπτυξη ενός ποσοστιαίου δείκτη που μας βοηθάει να κρίνουμε την προσαρμογή και την καταλληλότητα ενός γραμμικού μοντέλου. Αυτό το μέτρο δίνεται από τον τύπο:

$$R^2 = \frac{SSR}{SST} = \frac{SST - SSE}{SST} = 1 - \frac{SSE}{SST},$$

που ονομάζεται συντελεστής προσδιορισμού και εκφράζει το ποσοστό της μεταβλητότητας της μεταβλητής απόκρισης που εξηγείται από τις

επεξηγηματικές μεταβλητές. Από τη μορφή της σχέσης που δίνεται, εύκολα μπορούμε να καταλάβουμε ότι παίρνει τιμές μεταξύ 0 και 1. 'Οσο πιο κοντά στη μονάδα είναι η τιμή του συντελεστή προσδιορισμού, τόσο «καλύτερη» είναι η εξάρτηση των επεξηγηματικών μεταβλητών με την μεταβλητή απόκρισης.

Ωστόσο, αξίζει να σημειωθεί ότι ο συντελεστής προσδιορισμού αποτελεί ένα ανεπαρκές μέτρο καλής προσαρμογής καθώς όταν σε ένα γραμμικό μοντέλο εισάγουμε επιπλέον μεταβλητές, τότε το R^2 τείνει να αυξάνεται. Ως εκ τούτου, δεν αποτελεί καλό μέτρο σύγκρισης για μοντέλα με διαφορετικό πλήθος μεταβλητών, καθώς δεν είναι «δίκαιο».

Γι' αυτό το λόγο, εισάγουμε τον προσαρμοσμένο (ή διορθωμένο) συντελεστή προσδιορισμού ο οποίος λαμβάνει υπόψιν του τον αριθμό των μεταβλητών του εκάστοτε μοντέλου και διορθώνει τον R^2 ως προς το πλήθος των μεταβλητών.

Ο προσαρμοσμένος συντελεστής προσδιορισμού δίνεται από τη σχέση :

$$R^2_{Adj} = R^2 - (1 - R^2)\frac{p}{n-p-1},$$

όπου:

n: το πλήθος των παρατηρήσεων και

p: το πλήθος των επεξηγηματικών μεταβλητών.

Ακόμα, το μέσο τετραγωνικό σφάλμα MSE ορίζεται ως:

$$MSE = \frac{1}{n-p}\sum(y_i - \hat{y}_i)^2 = \frac{SSE}{n-p}.$$

Το AIC (*Akaike's information criterion*) αποτελεί ένα κριτήριο επιλογής του βέλτιστου μοντέλου, το οποίο βασίζεται στην αρχή της φειδωλότητας. Αποτελεί ένα μέτρο της σχετικά καλής προσαρμογής ενός στατιστικού μοντέλου. Ωστόσο, αν κανένα από τα υποψήφια μοντέλα δεν προσαρμόζεται καλά, το AIC δεν παρέχει στον αναλυτή κάποια πληροφορία για το γεγονός αυτό, καθώς δεν πράττει κάποιον έλεγχο υποθέσεων. Στην πράξη το κριτήριο στηρίζεται στη χαμένη πληροφορία όταν ένα μοντέλο χρησιμοποιείται για να περιγράψει την

πραγματικότητα. Μπορούμε να πούμε ότι περιγράφει το «ζύγισμα» μεταξύ μεροληψίας και διακύμανσης κατά την κατασκευή του μοντέλου, δηλαδή μεταξύ ακρίβειας και πολυπλοκότητας. Ορίστηκε από τον *Akaike* (1971) και για το γραμμικό μοντέλο δίνεται από τη σχέση:

$$AIC = n \ln\left(\frac{SSE}{n}\right) + 2\ (p+1),$$

n: το πλήθος των παρατηρήσεων και

p: το πλήθος των ανεξάρτητων μεταβλητών στο γραμμικό μοντέλο.

Συγκρίνοντας όλα τα υποψήφια μοντέλα με βάση το παραπάνω κριτήριο φαίνεται να είναι βέλτιστο εκείνο με τη μικρότερη τιμή AIC.

Το κριτήριο AIC αποτελεί μία από τις πιο διαδεδομένες στρατηγικές ποινικοποίησης έχοντας εφαρμογές σε οποιοδήποτε τομέα της στατιστικής χρειάζεται σύγκριση μοντέλων. Παρ' όλα αυτά, εξαιτίας της τάσης που έχει να επιλέγει μοντέλα με μεγάλο πλήθος επεξηγηματικών μεταβλητών, ένα άλλο κριτήριο έχει επίσης αναπτυχθεί.

Το κριτήριο BIC (*Bayesian information criterion*) προτάθηκε από τον *Schwarz* (1978). Για το γραμμικό μοντέλο ορίζεται από τη σχέση:

$$BIC = n \ln\left(\frac{SSE}{n}\right) + \ln(n)\ (p+1),$$

n: το πλήθος των παρατηρήσεων και

p: το πλήθος των ανεξάρτητων μεταβλητών στο γραμμικό μοντέλο.

Όπως και το AIC, έτσι και αυτό προσπαθεί να παράξει το καλύτερο μοντέλο με όσο το δυνατόν μικρότερο αριθμό επεξηγηματικών μεταβλητών. Ωστόσο η κύρια διαφορά του με το κριτήριο AIC έγκειται στο γεγονός ότι το BIC ποινικοποιεί πιο αυστηρά την αύξηση του πλήθους των επεξηγηματικών μεταβλητών που εμπλέκονται στο γραμμικό μοντέλο. Επομένως, το BIC τείνει να επιλέγει μικρότερα και πιο απλά μοντέλα συγκριτικά με το AIC. Συγκεκριμένα, ο συντελεστής

των παραμέτρων είναι $\ln(n)$ εν αντιθέσει του συντελεστή 2 που εμφανίζεται στο κριτήριο AIC. Επομένως, δεδομένου αυτής της διαφοράς συντελεστών για τα δύο κριτήρια ισχύει ότι όταν έχουμε 8 ή λιγότερες παρατηρήσεις το κριτήριο AIC φαίνεται να είναι αυστηρότερο ενώ όταν οι παρατηρήσεις ξεπερνούν τις 7 το κριτήριο BIC είναι αυστηρότερο. Και σε αυτή την περίπτωση καλύτερο γραμμικό μοντέλο θεωρείται εκείνο που έχει μικρότερη τιμή.

Τέλος, να σημειώσουμε ότι τα παραπάνω κριτήρια επηρεάζονται, το καθένα με διαφορετικό τρόπο, από τον αριθμό των μεταβλητών που εμπλέκονται στο εκάστοτε μοντέλο. Πιο συγκεκριμένα, κάποια κριτήρια όταν εφαρμοστούν στο υπό εξέταση μοντέλο δηλώνουν καλύτερο γραμμικό μοντέλο όταν η τιμή τους είναι μεγαλύτερη. Για παράδειγμα, ο συντελεστής προσδιορισμού R^2 όσο πιο μεγάλος είναι, δηλαδή όσο πιο κοντά στη μονάδα βρίσκεται η τιμή του, τόσο καλύτερο θεωρείται το μοντέλο (ανάλογη λογική και για τον προσαρμοσμένο συντελεστή προσδιορισμού R^2_{Adj}). Ωστόσο, κάποια άλλα κριτήρια δηλώνουν καλύτερο γραμμικό μοντέλο όταν η τιμή τους είναι μικρότερη. Για παράδειγμα, σύμφωνα με το SSE, που ισούται με το άθροισμα των τετραγώνων των σφαλμάτων, ή αντίστοιχα το MSE, καλύτερο μοντέλο είναι αυτό που έχει τη μικρότερη τιμή. Το ίδιο ισχύει και για τα κριτήρια πληροφορίας AIC και BIC, όπως επισημάναμε.

2.3 Μέθοδος εξέτασης όλων των δυνατών γραμμικών μοντέλων (*All Possible Regressions*)

Ο μόνος τρόπος για να σιγουρευτούμε ότι έχουμε καταλήξει στο «βέλτιστο» μοντέλο είναι να υπολογίσουμε όλα τα πιθανά υποσύνολα του αρχικού μας γραμμικού μοντέλου παλινδρόμησης. Αυτός είναι και ο σκοπός της μεθόδου της εξέτασης όλων των γραμμικών μοντέλων που συχνά συναντάται και ως *All Possible Regressions* ή *Full Enumeration*, να μελετηθούν δηλαδή όλα τα δυνατά γραμμικά μοντέλα που μπορούν να δημιουργηθούν κάνοντας χρήση όλων των διαθέσιμων αρχικών μεταβλητών και ύστερα από αυτές να ορίσουμε το «κατάλληλο» υποσύνολο.

Πιο συγκεκριμένα, κατά την διαδικασία της πλήρους εξέτασης, προσαρμόζονται μοντέλα που αποτελούνται από κάθε δυνατό συνδυασμό από τις p υποψήφιες επεξηγηματικές μεταβλητές, δηλαδή

αναπτύσσονται ομάδες υποσυνόλων με καμία, μία, δύο κ.ο.κ ανεξάρτητες μεταβλητές με τέτοιο τρόπο ώστε όλα τα παραγόμενα υποσύνολα μοντέλων να διαφέρουν κατά μία μεταβλητή. Άρα, ο συνολικός αριθμός των δυνατών μοντέλων που θα προσαρμοστούν ανέρχεται στα 2^p μοντέλα! Επομένως η ιδέα είναι να εξεταστούν όλα τα υποψήφια μοντέλα που προκύπτουν από τον συνδυασμό όλων των ανεξάρτητων μεταβλητών που διαθέτει το αρχικό μας μοντέλο. Για να επιτευχθεί αυτή η σύγκριση μεταξύ τους χρησιμοποιούμε ένα κριτήριο, το οποίο μπορεί να εφαρμοστεί για τη σύγκριση μοντέλων που έχουν διαφορετικό αριθμό επεξηγηματικών μεταβλητών.

Τα διαφορετικά μοντέλα αξιολογούνται συνήθως χρησιμοποιώντας τα κριτήρια πληροφορίας AIC ή BIC καθώς και τον προσαρμοσμένο συντελεστή προσδιορισμού R^2_{Adj}, τα οποία αναλύσαμε και παραπάνω. Έτσι, υπολογίζουμε για καθένα από τα 2^p μοντέλα ξεχωριστά το κριτήριο σύγκρισης που επιλέξαμε και τέλος επιλέγουμε το μοντέλο που έχει τη μικρότερη τιμή του κριτηρίου σύγκρισης από τα διαθέσιμα μοντέλα (για κριτήριο σύγκρισης το AIC ή το BIC) ή τη μεγαλύτερη τιμή αν επιλέξουμε ως κριτήριο σύγκρισης τον προσαρμοσμένο συντελεστή προσδιορισμού R^2_{Adj}.

2.4 Μειονέκτημα μεθόδου

Από θεωρητική άποψη, η εξέταση όλων των δυνατών γραμμικών μοντέλων είναι καλύτερη αφού δίνει στον ερευνητή μία πλήρη εικόνα για όλα τα υποψήφια μοντέλα που κατασκευάζονται με τον συνδυασμό των p επεξηγηματικών μεταβλητών που διαθέτουμε και του παρέχει τη δυνατότητα να μελετήσει όλες τις δυνατές περιπτώσεις με αποτέλεσμα να είναι σε θέση να επιλέξει με μεγαλύτερη ακρίβεια το καλύτερο υποσύνολο μεταβλητών. Η μέθοδος αυτή ναι μεν βρίσκει το «βέλτιστο» μοντέλο μέσα από όλα τα μοντέλα, η εξέταση τους δε, προκύπτει από συγκεκριμένα κριτήρια που χρησιμοποιεί ο κάθε ερευνητής. Υπό αυτή την έννοια, η τελική επιλογή για τον αν το μοντέλο αυτό έχει νόημα και από θεωρητικής πλευράς, θα εξαρτάται από τον ίδιο τον ερευνητή.

Από πρακτική άποψη, δεδομένου ότι η μέθοδος αυτή δίνει στον χρήστη μια μεγάλη οικογένεια διαθέσιμων υποσυνόλων, ο υπολογιστικός χρόνος που απαιτείται για την προσαρμογή όλων των γραμμικών μοντέλων αποτελεί «σπατάλη». Αυτή η απαιτούμενη πλήρης φυσική

προσπάθεια για την εξέταση όλων των αποτελεσμάτων από τον υπολογιστή είναι τεράστια όταν εξετάζονται όλες οι μεταβλητές.

Όπως αναφέραμε, για p τυχαίες μεταβλητές το πλήθος των γραμμικών μοντέλων που προκύπτουν είναι 2^p και αυτό σημαίνει πως για προβλήματα γραμμικής παλινδρόμησης όπου υπάρχει μεγάλο πλήθος τυχαίων μεταβλητών είναι πολύ μεγάλος και ο αριθμός των γραμμικών μοντέλων που προκύπτουν, καθώς το πλήθος αυτών μεγαλώνει εκθετικά σε σχέση με το πλήθος των τυχαίων μεταβλητών. Έτσι, ένα από τα βασικά μειονεκτήματα της μεθόδου της εξέτασης όλων των γραμμικών μοντέλων είναι ότι για μεγάλο αριθμό επεξηγηματικών μεταβλητών p καθίσταται υπολογιστικά αδύνατη.

Ενδεικτικά σημειώνουμε πως για 10 μεταβλητές έχουμε $2^{10} = 1024$ γραμμικά μοντέλα, ή για 30 μεταβλητές έχουμε $2^{30} = 1.073.741.824$ γραμμικά μοντέλα, ή όπως τα δεδομένα που θα χρησιμοποιήσουμε στο τελευταίο κεφάλαιο όπου έχουμε **141 μεταβλητές** προκύπτουν $\mathbf{2^{141} \approx 2,8 \cdot 10^{42}}$ γραμμικά μοντέλα, αριθμός τεράστιος που απαγορεύει την πλήρη εξερεύνηση του χώρου. Άρα η πλήρης εξερεύνηση του χώρου των πιθανών μας μοντέλων προτιμάται μόνο όταν έχουμε μικρό αριθμό υποψήφιων επεξηγηματικών μεταβλητών.

2.5 Επιλογή του καλύτερου υποσυνόλου επεξηγηματικών μεταβλητών (*Best Subset Selection*)

Με στόχο, λοιπόν, την μείωση αυτού του προβλήματος που προκαλεί ο υπολογισμός της δύναμης 2^p έχει προταθεί η εξής απλή παραλλαγή της μεθόδου, η οποία αναφέρεται ως διαδικασία επιλογής του καλύτερου υποσυνόλου των επεξηγηματικών μεταβλητών και συναντάται συχνά με την ονομασία *Best Subset Selection*. Η μέθοδος αυτή δεν κάνει τίποτα άλλο πέρα από το να «σπάσει» τον παραπάνω αλγόριθμο, δηλαδή τον υπολογισμό των 2^p μοντέλων, σε δύο στάδια:

i) Στο πρώτο στάδιο προσαρμόζουμε όλους τους δυνατούς συνδυασμούς μοντέλων οι οποίοι περιέχουν k επεξηγηματικές μεταβλητές εκ των p μεταβλητών που απαρτίζεται το αρχικό μας σύνολο.

Για παράδειγμα, αν έχουμε ένα μοντέλο με $p = 4$ πλήθος μεταβλητών, προσαρμόζουμε όλα τα δυνατά μοντέλα που περιέχουν:

- $k = 0$ μεταβλητές, όπου και παίρνουμε μόνο τον σταθερό όρο,

- $k = 1$ μεταβλητές, με 4 διαφορετικά πιθανά μοντέλα που περιέχει το καθένα μόνο μια μεταβλητή από τις συνολικές 4 κ.ο.κ εώς $k = p = 4$.

Αφού προσαρμοστούν όλα τα δυνατά αυτά μοντέλα, διαλέγουμε το καλύτερο για κάθε ομάδα μοντέλων με k μεταβλητές. Για την επιλογή του καλύτερου μοντέλου χρησιμοποιούμε συνήθως είτε τον συντελεστή προσδιορισμού R^2, βάση του οποίου καλύτερο είναι αυτό με την μεγαλύτερη τιμή, είτε το άθροισμα τετραγώνων των σφαλμάτων SSE, σύμφωνα με το οποίο καλύτερο θεωρείται αυτό με την μικρότερη τιμή.

Αξίζει να σημειωθεί ότι ο λόγος που χρησιμοποιούμε αυτά τα κριτήρια στο συγκεκριμένο στάδιο είναι ότι η σύγκριση εδώ γίνεται μεταξύ μοντέλων με τον ίδιο αριθμό επεξηγηματικών μεταβλητών k.

ii) Στο δεύτερο στάδιο επιλέγουμε το καλύτερο μοντέλο μέσα από τα καλύτερα μοντέλα του προηγούμενου βήματος. Τώρα η σύγκρισή τους γίνεται με βάση τα κριτήρια που εφαρμόζονται σε μοντέλα που περιέχουν διαφορετικό αριθμό επεξηγηματικών μεταβλητών. Έτσι επιλέγουμε το μοντέλο με το μικρότερο AIC ή BIC, ή με τη μεγαλύτερη τιμή του R^2_{Adj}.

2.6 Συμπεράσματα

Προφανώς και η μέθοδος εύρεσης του καλύτερου υποσυνόλου που αποτελεί απλή διαδικασία ως παραλλαγή-επέκταση της μεθόδου της πλήρους εξερεύνησης όλων των δυνατών μοντέλων, συνεχίζει όμως να «πάσχει» από υπολογιστικούς περιορισμούς. Καθώς ο αριθμός των επεξηγηματικών μεταβλητών αυξάνεται, αυξάνεται ραγδαία και το πλήθος των πιθανών μοντέλων. Επομένως, γίνεται υπολογιστικά αδύνατη όταν έχουμε στη διάθεσή μας μεγάλο αριθμό επεξηγηματικών μεταβλητών (για $p \geq 40$).

Ωστόσο τόσο η μέθοδος της πλήρους εξερεύνησης του χώρου όλων των γραμμικών μοντέλων όπως και η παραλλαγή της, μέθοδος καλύτερου

υποσυνόλου, ανοίγουν το δρόμο για την ανάπτυξη των πιο δημοφιλών και ευρέως χρησιμοποιούμενων διαδικασιών που θα παρουσιάσουμε στο παρακάτω κεφάλαιο, τις διαδικασίες κατά βήματα ή αλλιώς επαναληπτικές μεθόδους.

ΚΕΦΑΛΑΙΟ 3

Διαδικασίες κατά βήματα

3.1 Εισαγωγή

Όπως έχει ήδη αναφερθεί, ο μεγάλος αριθμός διαφορετικών προσαρμογών που απαιτείται να γίνει στα διαθέσιμα δεδομένα κατά την πλήρη εξερεύνηση του χώρου των γραμμικών μοντέλων αποτελεί δαπάνη και καθιστά πολλές φορές τη μέθοδο αυτή ανέφικτη.

Η ανάγκη για να ξεπεραστεί το πρόβλημα αυτό οδήγησε στην αναζήτηση και ανάπτυξη εναλλακτικών μεθόδων εντοπισμού «βέλτιστων» υποσυνόλων επεξηγηματικών μεταβλητών και έτσι αναπτύχθηκαν κάποιες επαναληπτικές διαδικασίες. Οι διαδικασίες αυτές αφορούν την ανάπτυξη ορισμένων αλγορίθμων οι οποίοι επιλέγουν και εισάγουν αυτόματα και διαδοχικά μία μία τις επεξηγηματικές μεταβλητές που κρίνονται σημαντικές για την πρόβλεψη της εξαρτημένης μεταβλητής ή αντίστοιχα απορρίπτουν μία μία τις μη σημαντικές επεξηγηματικές μεταβλητές ή τέλος συνδυάζοντας και τις δύο αυτές τεχνικές.

Με αυτόν τον τρόπο τελικά δημιουργείται μια σειρά από γραμμικά μοντέλα. Σύμφωνα με τον ακολουθιακό τρόπο που παράγονται αντιλαμβανόμαστε ότι κάθε μοντέλο προκύπτει με διαφορά μίας μεταβλητής από αυτό του προηγούμενου βήματος. Επαναλαμβάνοντας την ίδια διαδικασία τελικά καταλήγουμε στο τελευταίο μοντέλο της αλληλουχίας, το οποίο θεωρείται το πλέον ικανό να προβλέψει την

εξαρτημένη μεταβλητή, καθώς σύμφωνα με αυτό επιτυγχάνεται η βέλτιστη τιμή του εκάστοτε κριτηρίου που χρησιμοποιούμε.

Πιο συγκεκριμένα, τα κριτήρια που χρησιμοποιούμε για να αποφανθούμε για την εισαγωγή ή την εξαγωγή μιας επεξηγηματικής μεταβλητής από το μοντέλο είναι τα AIC και BIC καθώς και η στατιστική συνάρτηση t.

Οι διαδικασίες κατά βήματα υπερτερούν έναντι της εξέτασης όλων των δυνατών μοντέλων. Ενώ κατά την εξέταση του πλήρη χώρου των γραμμικών μοντέλων εφαρμόζονται τα κριτήρια βελτιστότητας σε όλα τα μοντέλα, οι επαναληπτικές διαδικασίες έχουν το προνόμιο να παράγουν ένα υποσύνολο μεταβλητών και σε αυτό να εφαρμόζονται τα κριτήρια ώστε να καταλήξουν στο «βέλτιστο». Επομένως, γίνεται αντιληπτό ότι η μέθοδος της εξέτασης όλων των δυνατών γραμμικών μοντέλων θεωρείται ιδιαίτερα πολυδάπανη και γι' αυτό το λόγο δε συνίσταται όταν το αρχικό μας σύνολο αποτελείται από μεγάλο αριθμό επεξηγηματικών μεταβλητών.

Ωστόσο, η εξέταση όλων των δυνατών μοντέλων παρουσιάζει στον χρήστη μία πλήρη εικόνα μοντέλων και έτσι ο αναλυτής είναι σε θέση να επιλέξει, σύμφωνα με την εμπειρία του, μέσα από το πλήθος των «καλών» μοντέλων που του προτείνεται, αυτό που κρίνει ο ίδιος ως το «καλύτερο» που ανταποκρίνεται στα δεδομένα του. Αντίθετα, ο τρόπος με τον οποίο υλοποιούνται οι επαναληπτικές μέθοδοι έχει ως αποτέλεσμα να καταλήγουν σε ένα μόνο μοντέλο. Αυτό αποτελεί και ένα βασικό μειονέκτημα των επαναληπτικών μεθόδων αφού σε κάποιες περιπτώσεις η λανθασμένη επιλογή της αλληλουχίας υποσυνόλων μεταβλητών μπορεί να αποκλείσει κάποια γραμμικά μοντέλα που ενδεχομένως να είναι «καλά».

Επομένως, κάθε αναλυτής θα πρέπει να έχει υπόψιν του ότι αυτές οι διαδικασίες μπορούν να θεωρηθούν χρήσιμες στην περίπτωση όπου δίνεται μόνο μια αόριστη εικόνα των μεταβλητών που πρέπει να συμπεριληφθούν στο μοντέλο από τη θεωρία και την εμπειρία. Ωστόσο, εάν η θεωρία και η εμπειρία είναι «καλοί οδηγοί», είναι γενικά καλύτερα για έναν ερευνητή να τις προτιμήσει από τη χρήση κάθε αυτοματοποιημένης διαδικασίας, συμπεριλαμβανομένου και αυτής της μεθόδου του καλύτερου υποσυνόλου μεταβλητών. Επιπλέον, αν τελικά καταφεύγουμε στην ανάπτυξη αυτών των ευμετάβλητων διαδικασιών,

θα πρέπει να τις χρησιμοποιούμε ως μόνο το πρώτο βήμα της διαδικασίας επιλογής μοντέλου, καθώς δεν λαμβάνουν υπόψιν τους την ύπαρξη του συχνού φαινομένου της πολλαπλής παλινδρόμησης, την πολυσυγγραμμικότητα.

Στη συνέχεια θα περιγράψουμε αναλυτικά τις τρεις πιο δημοφιλείς επαναληπτικές τεχνικές επιλογής βέλτιστων υποσυνόλων ανεξάρτητων μεταβλητών:

- τη μέθοδο της προς τα εμπρός επιλογής (*Forward Selection*),
- τη μέθοδο της προς τα πίσω απαλοιφής (*Backward Elimination*) και
- τη μέθοδο της κατά βήματα παλινδρόμησης (*Stepwise Regression*).

3.2 Μέθοδος της προς τα εμπρός επιλογής (*Forward Selection*)

Η μέθοδος της προς τα εμπρός επιλογής ή διαδοχικής πρόσθεσης επεξηγηματικών μεταβλητών, γνωστή και ως *Forward Selection*, λειτουργεί ως εξής:

Αρχίζει από το μοντέλο που περιέχει μόνο τον σταθερό όρο. Συνεχίζει προσθέτοντας μία μία τις επεξηγηματικές μεταβλητές, ξεκινώντας από αυτή που μας δίνει τη βέλτιστη (μικρότερη) τιμή του κριτηρίου σύγκρισης AIC ή BIC ή τη μεγαλύτερη τιμή της στατιστικής συνάρτησης t. Η διαδικασία επανλαμβάνεται μέχρις ότου η τιμή του κριτηρίου σύγκρισης για την προσθήκη οποιασδήποτε από όλες τις άλλες μεταβλητές που περιλαμβάνει το αρχικό μας μοντέλο να μην βελτιώνεται περισσότερο, οπότε και τερματίζεται.

Είναι προφανές ότι με την υλοποίηση αυτής της μεθόδου δεν επιτυγχάνεται η επανεκτίμηση της κάθε εισερχόμενης επεξηγηματικής μεταβλητής από τη στιγμή που αυτή προστεθεί στο μοντέλο σε δεδομένο βήμα. Δηλαδή, κάθε επεξηγηματική μεταβλητή που εισέρχεται στο μοντέλο μένει εκεί μόνιμα, χωρίς να επανελέγχεται για τη σημαντικότητά της.

Πιο συγκεκριμένα χρησιμοποιώντας την στατιστική συνάρτηση t ως κριτήριο επιλογής ώστε να εισαχθεί μία μεταβλητή στο μοντέλο, η μέθοδος της μπρος τα εμπρός επιλογής ξεκινάει με την προσαρμογή όλων των γραμμικών μοντέλων που περιέχουν μόνο μία ανεξάρτητη μεταβλητή της μορφής:

$$y_i = \beta_0 + x_{ij}\beta_j + \varepsilon_{i,}\ i = 1, \dots, n, j = 1, \dots, p.$$

Στη συνέχεια αφού υπολογίσει τις τιμές της στατιστικής συνάρτησης:

$$t = \left|\frac{\widehat{\beta}_j}{s(\widehat{\beta}_j)}\right|,$$

(3.1)

εντοπίζει την πρώτη μεταβλητή που θα εισαχθεί στο μοντέλο με δείκτη j_1, η οποία μας δίνει τη μεγαλύτερη στατιστικά σημαντική τιμή της t, δηλαδή ισχύει:

$$\left|\frac{\widehat{\beta}_{j_1}}{s(\widehat{\beta}_{j_1})}\right| = max\left\{\left|\frac{\widehat{\beta}_j}{s(\widehat{\beta}_j)}\right|, j = 1,2, \dots, p\right\}$$

και

$$\left|\frac{\widehat{\beta}_{j_1}}{s(\widehat{\beta}_{j_1})}\right| > t_{n-2}(\alpha/2),$$

(3.2)

όπου *α* ένα επιθυμητό επίπεδο σημαντικότητας ($0 < a < 1$).

Στην περίπτωση όμως που δεν ικανοποιείται το παραπάνω κριτήριο, δηλαδή έχουμε:

$$\left|\frac{\widehat{\beta}_{j_1}}{s(\widehat{\beta}_{j_1})}\right| \leq t_{n-2}(\alpha/2),$$

(3.3)

καμία μεταβλητή δεν θεωρείται σημαντική και έχουμε τερματισμό αναζήτησης.

Στην περίπτωση που ικανοποιείται η ***(3.2)*** και έχουμε επιλέξει να εισάγουμε την μεταβλητή X_{j_1} ως πρώτη μεταβλητή, η μέθοδος συνεχίζει στην επιλογή της δεύτερης μεταβλητής κάνοντας προσαρμογή σε όλα τα γραμμικά μοντέλα με δύο μεταβλητές της μορφής:

$$y_i = \beta_0 + x_{ij_1}\beta_{j_1} + x_{ij}\beta_j + \varepsilon_{i,}\ i = 1, \dots, n, j = 1, \dots, p, j \neq j_1.$$

Στη συνέχεια αφού πάλι υπολογίσει τους λόγους ***(3.1)*** εντοπίζει την δεύτερη μεταβλητή που θα εισαχθεί στο μοντέλο, Xj_2, για την οποία ισχύει:

$$\left|\frac{\hat{\beta}_{j_2}}{s(\hat{\beta}_{j_2})}\right| = max\left\{\left|\frac{\hat{\beta}_j}{s(\hat{\beta}_j)}\right|, j = 1,2, \dots, p, \mu\varepsilon\ \text{j}\ \neq\ j_1\ \right\}$$

και

$$\left|\frac{\widehat{\beta}_{j_2}}{s(\widehat{\beta}_{j_2})}\right| > t_{n-3}(\alpha/2).$$

Επομένως, με την εισαγωγή της X_{j_2}, το προτεινόμενο μοντέλο θα είναι τώρα το εξής:

$$y_i = \beta_0 + x_{ij_1}\beta_{j_1} + x_{ij_2}\beta_{j_2} + \varepsilon_{i,}\ i = 1, \dots, n.$$

Αν όμως ισχύει:

$$\left|\frac{\widehat{\beta}_{j_2}}{s(\widehat{\beta}_{j_2})}\right| \leq t_{n-3}(\alpha/2),$$

σημαίνει ότι καμία από τις εναπομείναντες μεταβλητές δεν κρίνεται σημαντική και η μέθοδος σταματάει κρίνοντας ως βέλτιστο το μοντέλο που βρέθηκε στο προηγούμενο βήμα.

Η διαδικασία επαναλαμβάνεται με τον ίδιο τρόπο, μέχρι να ικανοποιηθεί μία ανισότητα της μορφής ***(3.3)***.

3.3 Μέθοδος της προς τα πίσω απαλοιφής (*Backward Elimination*)

Η μέθοδος της προς τα πίσω απαλοιφής ή διαδοχικής αφαίρεσης επεξηγηματικών μεταβλητών, γνωστή και ως *Backward Elimination* υλοποιείται με αντίστροφη φορά από ότι η μέθοδος της προς τα εμπρός επιλογής και υλοποιείται ως εξής:

Αρχίζει συμπεριλαμβάνοντας στο μοντέλο όλες τις διαθέσιμες επεξηγηματικές μεταβλητές p. Συνεχίζει αφαιρώντας μία μία τις επεξηγηματικές μεταβλητές, ξεκινώντας από αυτή που δίνει τη μικρότερη τιμή κριτηρίου σύγκρισης του μοντέλου, με βάση τα κριτήρια πληροφορίας AIC ή BIC ή της στατιστικής συνάρτησης t. Η διαδικασία αυτή επαναλαμβάνεται μέχρις ότου το κριτήριο σύγκρισης του μοντέλου για την αφαίρεση οποιασδήποτε άλλης από τις εναπομείναντες επεξηγηματικές μεταβλητές να μην βελτιώνεται περισσότερο οπότε και σταματάει.

Και σε αυτή την περίπτωση είναι προφανές ότι όταν έχουμε αφαιρέσει μια επεξηγηματική μεταβλητή δεν μπορούμε να την συμπεριλάβουμε πάλι στο μοντέλο, ακόμα και αν αυτή εμφανίζεται σε κάποιο βήμα ως στατιστικά σημαντική. Ακόμα δεδομένου ότι ξεκινάει από το μοντέλο που περιέχει όλες τις ανεξάρτητες μεταβλητές, η διαδικασία απαλοιφής θα είναι αρκετά αργή όταν έχουμε πολύ μεγάλο πλήθος επεξηγηματικών μεταβλητών.

Με την χρήση της σταστιστικής συνάρτησης t ως κριτήριο απόρριψης για να εξαχθεί μία μεταβλητή από το μοντέλο η μέθοδος της προς τα πίσω απαλοιφής ξεκινάει με την προσαρμογή του πλήρους μοντέλου:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_{i,}\ i = 1, \dots, n.$$

Στη συνέχεια αφού υπολογίσει τους λόγους ***(3.1)*** εντοπίζει την μεταβλητή που θα εξαχθεί από το μοντέλο, X_{j_1}, η οποία μας δίνει τη μικρότερη στατιστικά σημαντική τιμή της t, δηλαδή ισχύει:

$$\left|\frac{\widehat{\beta}_{j_1}}{s(\widehat{\beta}_{j_1})}\right| = min\left\{\left|\frac{\widehat{\beta}_j}{s(\widehat{\beta}_j)}\right|, j = 1,2, \dots, p\right\}$$

και

$$\left|\frac{\hat{\beta}_{j_1}}{s(\hat{\beta}_{j_1})}\right| \leq t_{n-p+1}(\alpha/2).$$

(3.4)

Σε περίπτωση όμως που δεν ικανοποιείται το παραπάνω κριτήριο, δηλαδή έχουμε:

$$\left|\frac{\hat{\beta}_{j_1}}{s(\hat{\beta}_{j_1})}\right| > t_{n-p+1}(\alpha/2),$$

(3.5)

σημαίνει ότι ολες οι επεξηγηματικές μεταβλητές θεωρούνται σημαντικές όπου έχουμε και τερματισμό αναζήτησης με βέλτιστο να είναι το πλήρες μοντέλο.

Στην περίπτωση που ικανοποιείται η ***(3.4)*** και έχουμε αφαιρέσει την μεταβλητή X_{j_1}, η μέθοδος συνεχίζει στην εξαγωγή της δεύτερης μεταβλητής, προσαρμόζοντας τα γραμμικά μοντέλα της μορφής:

$$y_i = \beta_0 + \sum_{j=1}^{p} x_{ij}\beta_j + \varepsilon_{i,}\ i = 1, \dots, n, j \neq j_1.$$

Στη συνέχεια αφού πάλι υπολογίσει τους λόγους ***(3.1)*** εντοπίζει την δεύτερη μεταβλητή που θα εξαχθεί από το μοντέλο, X_{j_2}, για την οποία ισχύει:

$$\left|\frac{\hat{\beta}_{j_2}}{s(\hat{\beta}_{j_2})}\right| = min\left\{\left|\frac{\hat{\beta}_j}{s(\hat{\beta}_j)}\right|, j = 1,2, \dots, p, \mu\varepsilon\ \mathrm{j}\ \neq\ j_1\ \right\}$$

και

$$\left|\frac{\hat{\beta}_{j_2}}{s(\hat{\beta}_{j_2})}\right| \leq t_{n-p}(\alpha/2).$$

Η διαδικασία επαναλαμβάνεται με τον ίδιο τρόπο, μέχρι να ικανοποιηθεί μία ανισότητα της μορφής ***(3.5)***.

3.4 Μέθοδος της κατά βήματα παλινδρόμησης (*Stepwise Regression*)

Οι δύο προηγούμενες μέθοδοι όπως επισημάναμε έχουν ένα σημαντικό μειονέκτημα. Κατά την υλοποίηση της μεθόδου της προς τα εμπρός επιλογής αν μία μεταβλητή έχει εισαχθεί στο μοντέλο αφού είχε θεωρηθεί σημαντική σε κάποιο βήμα δεν μπορεί να αφαιρεθεί από το μοντέλο σε επόμενο βήμα ακόμα και αν πλέον είναι ασήμαντη για τη πρόβλεψη της εξαρτημένης μεταβλητής. Αντίστοιχα, κατά την ανάπτυξη της μεθόδου της προς τα πίσω απαλοιφής αν μία μεταβλητή έχει αφαιρεθεί από το μοντέλο αφού είχε θεωρηθεί ασήμαντη σε κάποιο βήμα δεν μπορεί να εισαχθεί στο μοντέλο σε κάποιο επόμενο βήμα ακόμα και αν πλέον θεωρείται σημαντική για τη πρόβλεψη της εξαρτημένης μεταβλητής.

Εξαιτίας αυτής της «μία τη φορά» εισαγωγής ή εξαγωγής των επεξηγηματικών μεταβλητών έχει αναπτυχθεί μια συνδυαστική επαναληπτική διαδικασία γνωστή ως κατά βήματα παλινδρόμηση η οποία υλοποιεί ταυτόχρονα και τη διαδικασία της προς τα εμπρός επιλογής και αυτή της προς τα πίσω απαλοιφής. Ουσιαστικά, η μέθοδος αυτή ξεκινάει όπως ακριβώς η μέθοδος της προς τα εμπρός επιλογής και στη συνέχεια με την εισαγωγή κάθε μεταβλητής στο μοντέλο ελέγχεται ποιες από τις μεταβλητές που συμπεριλαμβάνονται ήδη στο μοντέλο μπορούν να εξαχθούν από αυτό. Η διαδικασία αυτή της ταυτόχρονης εισαγωγής-εξαγωγής επεξηγηματικών μεταβλητών επαναλαμβάνεται μέχρι, σύμφωνα με τα κριτήρια σύγκρισης που χρησιμοποιούμε, να μην υπάρχει κάποια μεταβλητή εκτός μοντέλου, η οποία θα έπρεπε να συμπεριληφθεί στο μοντέλο και καμία μεταβλητή μέσα στο μοντέλο, η οποία δεν πρέπει να προστεθεί.

Τα κριτήρια που χρησιμοποιούμε για να επιλέξουμε ή να απορρίψουμε τις επεξηγηματικές μεταβλητές και σε αυτή τη συνδυαστική μέθοδο είναι τα ίδια με τις δύο προηγούμενες μεθόδους που εξετάσαμε και βασίζονται στα κριτήρια πληροφορίας AIC και BIC.

3.5 Συμπεράσματα

Αξίζει να σημειωθεί ότι κατά την υλοποίηση των τριών επαναληπτικών μεθόδων στα ίδια δεδομένα δεν είναι απαραίτητο να οδηγήσουν όλες στο ίδιο σύνολο ανεξάρτητων μεταβλητών.

Και οι τρεις διαδικασίες είναι συγκριτικά γρήγορες αφού γίνεται αυτόματη επιλογή υποσυνόλων και η εφαρμογή τους είναι αρκετά απλή καθώς είναι διαθέσιμες σε όλα τα υπολογιστικά πακέτα. Ακόμα, με την αύξηση του πλήθους των επεξηγηματικών μεταβλητών το κόστος των υπολογισμών αυξάνεται πιο αργά συγκριτικά με τη μέθοδο της πλήρους εξερεύνησης και αυτή του καλύτερου υποσυνόλου.

Παρ΄όλα αυτά, δεν είναι σίγουρο ότι τελικά θα καταλήξουν στο «βέλτιστο» μοντέλο, πέραν του πρώτου βήματος στην διαδικασία της προς τα εμπρός επιλογής και της μεθόδου της κατά βήματα παλινδρόμησης, όπου οι αλγόριθμοι εντοπίζουν τον πρώτο σημαντικότερο όρο του μοντέλου. Αυτό οφείλεται στον τρόπο με τον οποίο είναι κατασκευασμένοι οι αλγόριθμοι των επαναληπτικών διαδικασιών. Σύμφωνα με αυτόν καταλαβαίνουμε ότι η συμβολή της κάθε επεξηγηματικής μεταβλητής εξαρτάται από τη σειρά εισαγωγής της στο μοντέλο. Άρα κατά την υλοποίηση των διαδικασιών κατά βήματα ο κάθε ερευνητής θα πρέπει να έχει κατά νου του ότι η σειρά με την οποία οι ανεξάρτητες μεταβλητές εισάγονται ή εξάγονται, δεν σημαίνει απαραίτητα ότι αυτή είναι και η σειρά σημαντικότητας των επεξηγηματικών μεταβλητών.

Επομένως, η περίπτωση που κυρίως προτιμούνται αυτές οι μέθοδοι είναι όταν έχουμε μεγάλο αριθμό παρατηρήσεων και επεξηγηματικών μεταβλητών, όπου οι πιο σπάταλες και χρονοβόρες διαδικασίες δεν κρίνονται εφικτές.

Τέλος, η διαδικασία της κατά βήματα παλινδρόμησης, αν και απαιτεί περισσότερο υπολογισμό από αυτόν της διαδοχικής πρόσθεσης και της διαδοχικής αφαίρεσης μεταβλητών, προτιμάται και από τις δύο αυτές διαδικασίες καθώς επιτυγχάνει δύο ελέγχους ταυτόχρονα. Ο διπλός αυτός έλεγχος έγκειται στο γεγονός ότι όταν μια μεταβλητή είναι σημαντική και μπαίνει στο μοντέλο, ταυτόχρονα ελέγχεται αν υπάρχουν στο μοντέλο μεταβλητές που είχαν προστεθεί στο μοντέλο και πλέον δεν είναι σημαντικές. Υπό αυτή την έννοια, είναι λογικό να αναμένουμε η

συνδυαστική αυτή διαδικασία να έχει μεγαλύτερη πιθανότητα να επιλέγει καλύτερα υποσύνολα.

ΚΕΦΑΛΑΙΟ 4

Τεχνικές με ποινή

4.1 Εισαγωγή

Όπως είδαμε οι πιο δημοφιλείς και απλές ως προς την υλοποίησή τους, μέθοδοι επιλογής μεταβλητών, ειδικά όταν διαθέτουμε μεγάλο αριθμό επεξηγηματικών μεταβλητών είναι οι διαδικασίες κατά βήματα. Ωστόσο αυτές οι παραδοσιακές μέθοδοι επιλογής μεταβλητών, λόγω των επαναλαμβανόμενων μεμονομένων βημάτων που πράττουν, όχι μόνο δεν εγγυώνται ότι θα φτάσουν στο βέλτιστο μοντέλο, αλλά τα εκάστοτε μοντέλα που παράγουν είναι ιδιαίτερα μεταβλητά. Εξάλλου, οι διαδικασίες αυτές χρησιμοποιούν για την υλοποίησή τους τις συνήθεις εκτιμήτριες ελαχίστων τετραγώνων, οι οποίες όπως έχουμε δει προκύπτουν ελαχιστοποιώντας το άθροισμα των τετραγώνων των υπολοίπων. Ωστόσο, η μέθοδος των ελαχίστων τετραγώνων δεν είναι πάντα αποτελεσματική για την εκτίμηση των παραμέτρων του μοντέλου, ειδικά όταν το πλήθος των ανεξάρτητων μεταβλητών είναι μεγάλο ή όταν υπάρχουν υψηλά συσχετισμένες μεταβλητές, όπως θα παρουσιάσουμε αναλυτικά παρακάτω κατά το φαινόμενο της πολυσυγγραμμικότητας.

Σύμφωνα με τον *Tibshirani* (1996) το κύριο πρόβλημα που δημιουργείται με τις εκτιμήτριες ελαχίστων τετραγώνων σχετίζεται με την ακρίβεια πρόβλεψης. Οι εκτιμήτριες αυτές είναι αμερόληπτες αλλά έχουν συχνά μεγάλη διασπορά, η οποία με τη σειρά της οδηγεί σε μείωση της ακρίβειας πρόβλεψης των εκτιμητριών του μοντέλου. Αν ο αναλυτής αποδεχόταν κάποια μεροληψία προκειμένου να μειωθεί κατά ένα

ποσοστό η διασπορά των προβλεπόμενων τιμών, θα είχε πετύχει ταυτόχρονα και την βελτίωση της ακρίβειας πρόβλεψης του μοντέλου. Ένα τρόπος προκειμένου να επιτευχθεί ο παραπάνω συλλογισμός θα ήταν η συρρίκνωση κάποιων συντελεστών προς το μηδέν, ή ακόμα και ο μηδενισμός κάποιων.

Για τον λόγο αυτό δημιουργήθηκε η ανάγκη για την κατασκευή εκτιμητριών και τεχνικών που θα χρησιμοποιούσαν τις εκτιμήτριες αυτές, προκειμένου να αντιμετωπίσουν το πρόβλημα της μεθόδου ελαχίστων τετραγώνων, αλλά παράλληλα και να εκτιμήσουν τους συντελεστές παλινδρόμησης και να συμβάλλουν και αυτές με τη σειρά τους στην επιλογή μεταβλητών για εύρεση κατάλληλων υποσυνόλων.

Πάνω σε αυτό εργάστηκαν οι *Fan* και *Li* (2001) οι οποίοι πρότειναν μια καινούρια μεθοδολογία που στηρίζεται στα ποινικοποιημένα ελάχιστα τετράγωνα (*penalized least squares*). Έτσι αναπτύσσονται οι ποινικοποιήμενες μέθοδοι παλινδρόμησης, οι οποίες χρήζουν ιδιαίτερης προσοχής κυρίως την τελευταία δεκαετία, λόγω της υψηλής ακρίβειας πρόβλεψης και της υπολογιστικής αποτελεσματικότητας που πετυχαίνουν. Ουσιαστικά αυτό που τελικά επιτυγχάνεται με τις ποινικοποιημένες μεθόδους είναι ότι διατηρούν όλες τις αρχικές μεταβλητές στο μοντέλο αλλά ταυτόχρονα ποινικοποιούν, βάζοντας περιορισμούς, τους συντελεστές παλινδρόμησης συρρικνώνοντάς τους κοντά στο μηδέν. Βέβαια, αν η συρρίκνωση είναι αρκετά μεγάλη τότε αυτές οι μέθοδοι συντελούν στην επιλογή μεταβλητών, συρρικνώνοντάς τους στο μηδέν.

Θα παρουσιάσουμε λοιπόν στη συνέχεια δύο ποινικοποιημένες τεχνικές, την παλινδρόμηση κορυφογραμμής ή διασέλου (*Ridge Regression*) και τη *LASSO* (*Least Absolute Shrinkage and Selection Operator*). Ως μέθοδοι μοιάζουν στο γεγονός ότι και οι δύο τεχνικές συρρικνώνουν τους συντελεστές στο μοντέλο σύμφωνα με ένα συγκεκριμένο κριτήριο, έτσι ώστε οι συντελεστές να είναι κατάλληλοι όχι μόνο να ελαχιστοποιούν το άθροισμα των τετραγώνων των σφαλμάτων (SSE), αλλά να προσπαθούν να πετύχουν ένα χαμηλό SSE με λίγες μεταβλητές και με χαμηλούς συντελεστές. Με άλλα λόγια, ο αλγόριθμος προσαρμογής τιμωρείται για το μέγεθος των συντελεστών, επιβάλλοντας μία ποινή στους συντελεστές παλινδρόμησης. Ωστόσο, η βασική τους διαφορά έγκειται στη μορφή της ποινής αυτής. Ακόμα, η πρώτη λειτουργεί έμμεσα ως επιλογή

μεταβλητών, ενώ η δεύτερη αποτελεί άμεση μέθοδο επιλογής μεταβλητών.

4.2 Φαινόμενο πολυσυγγραμμικότητας

Τα πολλαπλά γραμμικά μοντέλα που συνήθως καλείται να μελετήσει ένας ερευνητής αποτελούνται από ένα μεγάλο πλήθος επεξηγηματικών μεταβλητών. Επομένως, αυξάνεται και η πιθανότητα ύπαρξης συσχετίσεων μεταξύ των μεταβλητών αυτών του υπό εξέταση μοντέλου. Όταν οι ανεξάρτητες μεταβλητές δεν είναι ορθογώνιες μεταξύ τους, δηλαδή όταν υπάρχει κάποια γραμμική σχέση μεταξύ τους, είναι ενδεχόμενο οι εκτιμούμενοι συντελεστές παλινδρόμησης να είναι εξαιρετικά ασταθείς και οι τιμές τους να υφίστανται δραματικές αλλαγές όταν κάποια νέα μεταβλητή προστίθεται ή απομακρύνεται. Δυστυχώς, στις περισσότερες εφαρμογές της παλινδρόμησης, παρατηρείται έλλειψη ορθογωνιότητας. Οπότε έχουμε το γνωστό πρόβλημα πολυσυγγραμμικότητας (*multicollinearity*), το οποίο θα πρέπει να αντιμετωπιστεί και αυτός είναι ο λόγος που μελετάμε τις ποινικοποιημένες μεθόδους.

Σε μια τέτοια περίπτωση, όπου μία ανεξάρτητη μεταβλητή είναι γραμμική συνάρτηση των υπολοίπων ή κάποιων ανεξάρτητων μεταβλητών, η αστάθεια των διαδικασιών κατά βήματα μπορεί να είναι ιδιαίτερα προβληματική. Η χρήση των εκτιμητριών ελαχίστων τετραγώνων για τη προσαρμογή του μοντέλου με μειωμένο αριθμό μεταβλητών καθώς και η προσθήκη ή η αφαίρεση μίας επεξηγηματικής μεταβλητής επιφέρουν μεγάλες αλλαγές στις τιμές των εκτιμώμενων συντελεστών του μοντέλου της γραμμικής παλινδρόμησης. Επομένως, το μοντέλο δεν είναι σε καμία περίπτωση αποτελεσματικό και η ερμηνεία της επίδρασης μιας ανεξάρτητης μεταβλητής στην εξαρτημένη μεταβλητή είναι παραπλανητική, δεδομένου ότι η ανεξάρτητη μεταβλητή στο μοντέλο μπορεί να ερμηνεύει άλλες επιδράσεις από αυτές που αναμένονται. Η ποινικοποιημένη παλινδρόμηση μπορεί να αντιμετωπίσει την αστάθεια αυτή μειώνοντας τη διασπορά αυτή που εμπλέκεται στην εκτίμηση των συντελεστών.

Η έλλειψη ορθογωνιότητας των μεταβλητών συνεπάγεται την ύπαρξη κάποιας «τέλειας» ή «σχεδόν τέλειας» γραμμικής σχέσης μεταξύ των στηλών που περιέχουν τις τιμές ορισμένων ανεξάρτητων μεταβλητών και οδηγούν αντίστοιχα στην τέλεια ή σχεδόν τέλεια πολυσυγγραμμικότητα.

Προκειμένου να αντιληφθούμε, από μαθηματικής άποψης, την γραμμική συσχέτιση των μεταβλητών και το πώς αυτή επηρεάζει τον υπολογισμό της εκτιμήτριας ελαχίστων τετραγώνων θα παραθέσουμε αρχικά τον ορισμό της γραμμικής εξάρτησης.

Ορισμός 4.1(Γραμμική εξάρτηση):

Οι στήλες $X_1, \dots, X_p$ του πίνακα σχεδιασμού $\boldsymbol{X}$ θα είναι γραμμικώς εξαρτημένες αν υπάρχουν σταθεροί αριθμοί $c_1, \dots, c_p$ όχι όλοι μηδέν ώστε:

$$\sum_{j=1}^{p} c_j X_j = 0$$

ή

$$c_1 X_1 + c_2 X_2 + \cdots + c_p X_p = 0. \qquad \textbf{(4.1)}$$

Αν ο τύπος ***(4.1)*** ισχύει για τουλάχιστον δύο από τις στήλες του πίνακα $\boldsymbol{X}_{(n\times p)}$, τότε η τάξη του πίνακα $\boldsymbol{X'X}_{(p\times p)}$ θα είναι μικρότερη από τη τιμή p και δε θα ορίζεται ο $(\boldsymbol{X'X})^{-1}$, με αποτέλεσμα να μη μπορεί να εκτιμηθεί το διάνυσμα $\boldsymbol{\beta}$ κατά τη χρήση της εκτιμήτριας ελαχίστων τετραγώνων αφού υπενθυμίζουμε ότι $\widehat{\boldsymbol{\beta}}_{LS} = (\boldsymbol{X'X})^{-1}\boldsymbol{X'y}$.

Ωστόσο, για ορισμένες στήλες του πίνακα $\boldsymbol{X}$ ο ***(4.1)*** δίνει τιμές όχι ακριβώς μηδέν αλλά πολύ κοντά στο μηδέν, δηλαδή:

$$\sum_{j=1}^{p} c_j X_j \simeq 0$$

ή

$$\sum_{j=1}^{p} c_j X_j + d_i = 0,$$

όπου d_i στοχαστικό σφάλμα.

Επομένως, σε αυτή τη περίπτωση (όπου υπάρχει μία σχεδόν τέλεια γραμμική σχέση μεταξύ των επεξηγηματικών μεταβλητών) ναι μεν ορίζεται ο $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, αλλά η ορίζουσα του $\boldsymbol{X}'\boldsymbol{X}$ είναι πολύ κοντά στο μηδέν, γεγονός που ευθύνεται για τις μεγάλες απόλυτες τιμές των εκτιμητριών ελαχίστων τετραγώνων $\hat{\beta}_j, j = 1, \dots, p$ που παράγονται.

Αυτό αποδεικνύεται αν υπολογίσουμε το μέσο τετραγωνικό σφάλμα (MSE). Εδώ να σημειώσουμε ότι το MSE οποιουδήποτε εκτιμητή $\widehat{\boldsymbol{\beta}}$ ορίζεται ως η αναμενόμενη τετραγωνική απόσταση του $\widehat{\boldsymbol{\beta}}$ από το διάνυσμα των πραγματικών συντελεστών $\boldsymbol{\beta}$, δηλαδή:

i) $MSE = E\left(\hat{\beta} - \beta\right)^2$ για ένα συντελεστή β:

$$MSE = E\left(\hat{\beta} - \beta\right)^2$$

$$= E\left(\left(\hat{\beta} - E\left(\hat{\beta}\right) + E\left(\hat{\beta}\right) - \beta\right)^2\right)$$

$$= E\left(\left(\hat{\beta} - E\left(\hat{\beta}\right)\right)^2\right) + E\left(\left(E\left(\hat{\beta}\right) - \beta\right)^2\right) + 2E\left(\left(\hat{\beta} - E\left(\hat{\beta}\right)\right)\left(E\left(\hat{\beta}\right) - \beta\right)\right)$$

$$= Var\left(\hat{\beta}\right) + bias(\hat{\beta})^2 + 2E\left(\hat{\beta}E\left(\hat{\beta}\right) - \left(E\left(\hat{\beta}\right)\right)^2 - \hat{\beta}\beta + E\left(\hat{\beta}\right)\beta\right)$$

$$= Var\left(\hat{\beta}\right) + bias(\hat{\beta})^2 + 2\left(\left(E\left(\hat{\beta}\right)\right)^2 - \left(E\left(\hat{\beta}\right)\right)^2 - \beta E\left(\hat{\beta}\right) + \beta E\left(\hat{\beta}\right)\right)$$

$$= Var\left(\hat{\beta}\right) + bias(\hat{\beta})^2.$$

(4.2)

ii) $MSE = E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)$, όταν πρόκειται για διάνυσμα τιμών, το οποίο ισούται και με το ίχνος (*trace*) της αναμενόμενης

τιμής του πίνακα διασποράς των σφαλμάτων, δηλαδή $MSE = trM(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})$, όπου $M(\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta}) = E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right)$.

Πιο συγκεκριμένα χρησιμοποιώντας τον πίνακα διασποράς των σφαλμάτων όπου δίνεται από:

$$M\left((\widehat{\boldsymbol{\beta}}, \boldsymbol{\beta})\right) = E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\right)$$

$$= E\left((\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}}) + E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}}) + E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})'\right)$$

$$= E\left(\left(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right)\left(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right)'\right)$$
$$+2E\left(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right)E(E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})' + (E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})(E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})'$$

$$= Var(\widehat{\boldsymbol{\beta}}) + bias(\widehat{\boldsymbol{\beta}})bias(\widehat{\boldsymbol{\beta}})', \qquad \textbf{(4.3)}$$

το μέσο τετραγωνικό σφάλμα ορίζεται ως:

$$MSE = E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right)$$

$$= E\left((\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}}) + E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}}) + E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})\right)$$

$$= E\left(\left(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right)'\left(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right)\right) + 2E(E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})'E\left(\widehat{\boldsymbol{\beta}} - E(\widehat{\boldsymbol{\beta}})\right)$$
$$+(E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})'(E(\widehat{\boldsymbol{\beta}}) - \boldsymbol{\beta})$$

$$= tr\left(Var(\widehat{\boldsymbol{\beta}})\right) + bias(\widehat{\boldsymbol{\beta}})'bias(\widehat{\boldsymbol{\beta}}).$$

Άρα:

$$MSE = tr(M)$$

$$= tr\left(Var(\widehat{\boldsymbol{\beta}})\right) + tr\left(bias(\widehat{\boldsymbol{\beta}})bias(\widehat{\boldsymbol{\beta}})'\right)$$

$$= tr\left(Var(\widehat{\boldsymbol{\beta}})\right) + tr\left(bias(\widehat{\boldsymbol{\beta}})' bias(\widehat{\boldsymbol{\beta}})\right)$$

$$= tr\left(Var(\widehat{\boldsymbol{\beta}})\right) + bias(\widehat{\boldsymbol{\beta}})' bias(\widehat{\boldsymbol{\beta}}).$$

(4.4)

Συγκεκριμένα, για την εκτιμήτρια ελαχίστων τετραγώνων έχουμε από τον τύπο ***(4.4)*** αντικαθιστώντας $bias(\widehat{\boldsymbol{\beta}})' bias(\widehat{\boldsymbol{\beta}}) = 0$, καθώς αποτελεί αμερόληπτη εκτιμήτρια:

$$MSE(\widehat{\boldsymbol{\beta}}_{LS}) = E({L_1}^2)$$

$$= tr\left(Var(\widehat{\boldsymbol{\beta}}_{LS})\right)$$

$$= \sigma^2 tr(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Πράγματι,

$$MSE(\widehat{\boldsymbol{\beta}}_{LS}) = E({L_1}^2)$$

$$= E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right),$$
$$\text{αφού } L_1 = \widehat{\boldsymbol{\beta}} - \boldsymbol{\beta} \text{ και } {L_1}^2 = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})$$

$$= \sum_{j=1}^{p} E(\hat{\beta}_j - \beta_j)^2$$

$$= \sum_{j=1}^{p} Var(\hat{\beta}_j)$$

$$= \sigma^2 tr(\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Προκειμένου να καταλάβουμε το αποτέλεσμα στο οποίο φτάσαμε, καλό θα ήταν να το γράψουμε υπό μορφή ιδιοτιμών, καθώς γνωρίζουμε ότι το άθροισμα των διαγώνιων στοιχείων ενός πίνακα (*trace*) είναι ίσο με το άθροισμα των ιδιοτιμών του. Είναι χρήσιμο λοιπόν, να γράψουμε τον

πίνακα $\boldsymbol{X'X}$ (και στη συνέχεια τον αντίστροφό του, $(\boldsymbol{X'X})^{-1}$) με τη μορφή διαγώνιου πίνακα.

Ορισμός 4.2(Διαγωνοποιημένη μορφή):

Έστω ένας τετραγωνικός πίνακας $\boldsymbol{A}_{(p\times p)}$ και έστω ότι έχει ιδιοτιμές $v_1, \dots, v_p$ και αντίστοιχα ιδιοδιανύσματα $p_1, \dots, p_p$. Έστω τώρα $\boldsymbol{P}$ ο πίνακας που έχει στήλες τα ιδιοδιανύσματα και $\boldsymbol{V}$ ο διαγώνιος πίνακας με στοιχεία του τις ιδιοτιμές του $\boldsymbol{A}$. Τότε ισχύει:

$$\boldsymbol{A} = \boldsymbol{PVP^{-1}},$$

όπου P ορθογώνιος πίνακας ($\boldsymbol{P'} = \boldsymbol{P^{-1}}$).

Στην περίπτωσή μας:

$$\boldsymbol{X'X} = \boldsymbol{PVP^{-1}},$$

όπου $\boldsymbol{V}$ διαγώνιος πίνακας του $\boldsymbol{X'X}$ ($\boldsymbol{V} = diag(\boldsymbol{X'X})$) με

$$\boldsymbol{V} = \begin{pmatrix} v_1 & & \\ & \ddots & \\ & & v_p \end{pmatrix}.$$

Τώρα δεδομένου ότι ο αντίστροφος πίνακας αντιστοιχεί σε αντιστροφή των ιδιοτιμών, ο όρος $(\boldsymbol{X'X})^{-1}$ γράφεται ως:

$$(\boldsymbol{X'X})^{-1} = \boldsymbol{PV^{-1}P'} \text{ (με } \boldsymbol{P'} = \boldsymbol{P^{-1}}),$$

όπου $\boldsymbol{V^{-1}}$ διαγώνιος πίνακας του $(\boldsymbol{X'X})^{-1}$ ($\boldsymbol{V^{-1}} = diag(\,(\boldsymbol{X'X})^{-1})$) με

$$\boldsymbol{V^{-1}} = \begin{pmatrix} 1/v_1 & & \\ & \ddots & \\ & & 1/v_p \end{pmatrix}.$$

Έτσι λοιπόν θα έχουμε:

$$MSE(\widehat{\boldsymbol{\beta}}_{LS}) = E(L_1{}^2) = trVar(\widehat{\boldsymbol{\beta}}_{LS}) = \sigma^2 \sum_{j=1}^{p} \frac{1}{v_j}.$$

(4.5)

Όμως, κατά την ύπαρξη του φαινομένου της πολυσυγγραμμικότητας, τουλάχιστον μία από τις ιδιοτιμές του $\boldsymbol{X'X}$ θα είναι πολύ μικρή (κοντά στο μηδέν). Αυτό γιατί δεδομένου ότι οι ιδιοτιμές v_j είναι της μορφής:

$$v_j = p_j{}'\boldsymbol{X'X}p_j = (\boldsymbol{X}p_j)'(\boldsymbol{X}p_j)\,, j = 1, \dots, p,$$

για μικρές ιδιοτιμές v_j του $\boldsymbol{X'X}$ παίρνουμε:

$$(\boldsymbol{X}p_j)'(\boldsymbol{X}p_j) \simeq 0$$

ή

$$\boldsymbol{X}p_j \simeq 0.$$

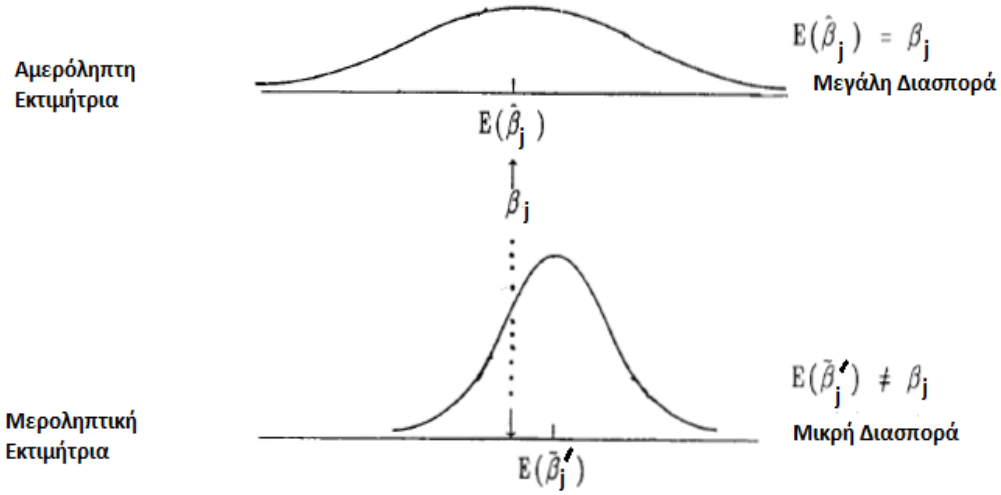
Γεγονός που σημαίνει ότι υπάρχει μία τουλάχιστον γραμμική σχέση μεταξύ των στηλών του $\boldsymbol{X}$ (φαινόμενο πολυσυγγραμμικότητας).

Άρα από τον τύπο ***(4.5)*** αντιλαμβανόμαστε ότι κατά την παρουσία της πολυσυγγραμμικότητας οι εκτιμήτριες ελαχίστων τετραγώνων είναι αρκετά μεγάλες σε απόλυτη τιμή (έχουν αρκετά μεγάλες διασπορές). Αυτό συμβαίνει από την απαίτηση η εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$ να είναι αμερόληπτη εκτιμήτρια του συντελεστή $\boldsymbol{\beta}$, καθώς σύμφωνα με το θεώρημα των *Gauss-Markov* ναι μεν η εκτιμήτρια ελαχίστων τετραγώνων έχει την ελάχιστη διασπορά στη κλάση των αμερόληπτων γραμμικών εκτιμητριών, δε συνεπάγεται όμως ότι θα είναι και μικρή η τιμή της, με αποτέλεσμα τα διαστήματα εμπιστοσύνης για το $\boldsymbol{\beta}$ να έχουν μεγάλο εύρος (***Σχήμα 4.1***). Αυτή η απότομη αύξηση της διασποράς των τιμών, μπερδεύει κάπως τον αναλυτή καθώς είναι πιο δύσκολος ο εντοπισμός των στατιστικά σημαντικών μεταβλητών.

Ο αναλυτής λοιπόν για να αντιμετωπίσει το πρόβλημα που δημιουργεί η παρουσία της πολυσυγγραμμικότητας, προτιμάει να «θυσιάσει» την αμεροληψία που του προσφέρουν οι εκτιμήτριες ελαχίστων τετραγώνων πετυχαίνοντας ταυτόχρονα μεγαλύτερη μείωση στη διασπορά. Έτσι

καταφεύγει στην εύρεση πιο «καλών»-μεροληπτικών εκτιμητριών όσον αφορά την ακρίβεια πρόβλεψης.

Γενικά, καλοί εκτιμητές θεωρούνται αυτοί που έχουν μικρή μεροληψία και ταυτόχρονα μικρή διασπορά. Από τον τύπο ***(4.2)*** παρατηρούμε ότι αν επιτρέψουμε ένα ποσοστό μεροληψίας στην μεροληπτική εκτιμήτρια έστω $\widehat{\boldsymbol{\beta}}'$, η διασπορά της μπορεί να γίνει τόσο μικρή όσο τελικά να πετύχουμε και μικρότερο μέσο τετραγωνικό σφάλμα συγκριτικά με αυτό της αμερόληπτης εκτιμήτριας, έστω $\widehat{\boldsymbol{\beta}}$ της οποίας το μέσο τετραγωνικό σφάλμα θα ισούται με την διασπορά της εκτιμήτριας. Άρα, αυτό που επιδιώκουμε σε τελική ανάλυση είναι το $MSE(\widehat{\boldsymbol{\beta}}') = Var(\widehat{\boldsymbol{\beta}}') + bias(\widehat{\boldsymbol{\beta}}')^2$ να είναι μικρότερο από το $MSE(\widehat{\boldsymbol{\beta}}_{LS}) = Var(\widehat{\boldsymbol{\beta}}_{LS})$.



***Σχήμα 4.1*: Διασπορά και μεροληψία**

Επομένως, όταν τα δεδομένα μας χαρακτηρίζονται από πολυσυγγραμμικότητα, η κλασική μέθοδος των ελαχίστων τετραγώνων δεν μπορεί να προσφέρει ικανοποιητικές εκτιμήσεις για τους συντελεστές παλινδρόμησης των μεταβλητών μας. Η υψηλή συσχέτιση των συμμεταβλητών έχει ως αποτέλεσμα οι εκτιμήτριες των παραμέτρων του μοντέλου παλινδρόμησης να έχουν μεγάλη διασπορά. Επομένως, οι εκτιμήτριες αυτές τείνουν να διαφέρουν εξαιρετικά από ένα δείγμα σε

άλλο. Αυτό με τη σειρά του έχει ως αποτέλεσμα ανακριβείς πληροφορίες για τους συντελεστές παλινδρόμησης, καταλήγοντας σε λάθος συμπερασματολογία. Ακόμα, η ερμηνεία των συντελεστών παλινδρόμησης ότι μετρούν την μεταβολή της αναμενόμενης τιμής της μεταβλητής απόκρισης όταν η αντίστοιχη επεξηγηματική μεταβλητή αυξηθεί κατά μία μονάδα, όταν οι υπόλοιπες επεξηγηματικές μεταβλητές παραμένουν ίδιες, δεν μπορούμε να την χρησιμοποιήσουμε.

4.3 Παλινδρόμηση Κορυφογραμμής ή Διασέλου (*Ridge Regression*)

4.3.1 Περιγραφή της μεθόδου

Μία από τις μεθόδους που αναπτύχθηκε για την εύρεση αυτών των μεροληπτικών εκτιμητριών για τους συντελεστές παλινδρόμησης, οι οποίες αντιμετωπίζουν το πρόβλημα της πολυσυγγραμμικότητας είναι και η παλινδρόμηση κορυφογραμμής ή διασέλου (*Ridge Regression*). Ουσιαστικά, αποτελεί μία μέθοδο συρρίκνωσης των συντελεστών των επεξηγηματικών μεταβλητών του μοντέλου μας προς το μηδέν. Η ανάπτυξη αυτής της μεθόδου οδηγεί σε άλλες πιο αποτελεσματικές μεθόδους συρρίκνωσης που έχουν την δυνατότητα να θέτουν συντελεστές ακριβώς ίσους με το μηδέν.

4.3.2 Ποινικοποίηση της L_2-νόρμας

Αυτό που ουσιαστικά θέλουμε να πετύχουμε με τις τεχνικές της ποινικοποιημένης παλινδρόμησης είναι να ελέγξουμε τη διασπορά των συντελεστών και γι΄αυτό το λόγο βάζουμε περιορισμούς στους συντελεστές ώστε να θέσουμε ένα όριο για το πόσο μεγάλοι μπορεί να γίνουν.

Η παλινδρόμηση *Ridge* στοχεύει στην ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων ενώ ταυτόχρονα ποινικοποιεί το τετράγωνο της L_2-νόρμας, το οποίο εκφράζεται από το αθροίσμα των τετραγώνων των συντελεστών παλινδρόμησης και μετράει την απόσταση των συντελεστών αυτών από το μηδέν:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2$$

υπό τον περιορισμό:

$$\sum_{j=1}^{p} \beta_j^2 \leq t,$$

(4.6)

όπου t είναι μια ρυθμιστική παράμετρος.

Για να λύσουμε αυτήν την ελαχιστοποίηση με τον περιορισμό χρησιμοποιούμε τη μέθοδο πολλαπλασιαστών *Lagrange* και το πρόβλημα γίνεται ισοδύναμο με ελαχιστοποίηση της ποσότητας ***(4.7)***:

$$\sum_{i=1}^{n} \left(y_i - \sum_{j=1}^{p} x_{ij}\, \beta_j \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

(4.7)

όπου $\lambda \geq 0$ είναι μια σταθερά η οποία επιλέγεται από τον αναλυτή και ονομάζεται παράμετρος ποινής.

Η έννοια και η εισαγωγή του φράγματος t στην ***(4.6)*** ή της παραμέτρου ποινής λ στον όρο $\lambda \sum_{j=1}^{p} \beta_j^2$ στην ***(4.7)*** λειτουργούν ως περιορισμοί και ελέγχουν το μέγεθος στο άθροισμα των τετραγώνων των συντελεστών, έτσι ώστε να αποτρέπονται οι μη λογικές μεγάλες τιμές που μπορούν να πάρουν κατά την παρουσία του φαινομένου της πολυσυγγραμμικότητας.

Ακόμα, η σχέση της παραμέτρου ποινής λ με το φράγμα t είναι αντίστροφη. Από τη σχέση ***(4.7)*** αντιλαμβανόμαστε ότι εάν η τιμή της παραμέτρου ποινής είναι μεγάλη τότε το άθροισμα τετραγώνων των συντελεστών πρέπει να είναι μικρό, προκαλώντας έτσι την συρρίκνωσή τους προς το μηδέν. Με αυτόν τον τρόπο μειώνεται και η διασπορά των συντελεστών και κατά συνέπεια επιτυγχάνεται μεγάλυτερη ακρίβεια στο μοντέλο. Αντίθετα, από τη σχέση ***(4.6)*** διαισθητικά καταλαβαίνουμε ότι για να πετύχουμε την συρρίκνωση των συντελεστών επιδιώκουμε μικρή τιμή του φράγματος t.

Η ελαχιστοποίηση της ποσότητας ***(4.7)*** δίνεται παραγωγίζοντάς τη και προκύπτει με αυτόν τον τρόπο η εκτιμήτρια *Ridge* η οποία δίνεται από τον τύπο:

$$\widehat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}.$$

Πιο συγκεκριμένα, γράφοντας τον τύπο ***(4.7)*** υπό μορφή πινάκων έχουμε:

$$\begin{aligned} L(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}'\boldsymbol{\beta} \\ &= (\boldsymbol{y}' - \boldsymbol{\beta}'\boldsymbol{X}')(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + \lambda \boldsymbol{\beta}'\boldsymbol{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \lambda \boldsymbol{\beta}'\boldsymbol{\beta} \\ &= \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \lambda \boldsymbol{\beta}'\boldsymbol{\beta}. \end{aligned}$$

(4.8)

Να σημειώσουμε εδώ ότι στην πρότελευταια σειρά της ***(4.8)*** ο δεύτερος και ο τρίτος όρος είναι ίσοι (ένας 1×1 πίνακας είναι πάντα συμμετρικός) και μπορούν να αντικατασταθούν από τον όρο $-2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y}$.

Παραγωγίζοντας την ***(4.8)*** θα έχουμε:

$$\begin{aligned} \frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} &= -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + 2\lambda\boldsymbol{\beta} \\ &= -2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta}. \end{aligned}$$

(4.9)

Ο πρότελευταιος όρος στην πρώτη σειρά της ***(4.9)*** προκύπτει από το γεγονός ότι σύμφωνα με τις ιδιότητες πινάκων για την παραγώγιση έχουμε:

$$\frac{\partial(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}.$$

Με το ίδιο σκεπτικό προκύπτει και ο τελευταίος όρος ($2\lambda\boldsymbol{\beta}$) αφού δεδομένου ότι $\boldsymbol{\beta}'\boldsymbol{\beta} = \boldsymbol{\beta}'I\boldsymbol{\beta}$, θα έχουμε:

$$\frac{\partial(\boldsymbol{\beta}'\boldsymbol{I}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 2\boldsymbol{\beta}.$$

Τέλος, θέτοντας την παράγωγο ίση με μηδέν καταλήγουμε στην λύση της εκτιμήτριας *Ridge*:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 0$$

ή

$$-2\boldsymbol{X}'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{\beta} = 0$$

ή

$$\widehat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}. \qquad \textbf{(4.10)}$$

Από την σχέση ***(4.10)*** παρατηρούμε ότι η $\widehat{\boldsymbol{\beta}}_R$ είναι και αυτή γραμμική συνάρτηση του $\boldsymbol{y}$, όπως και η εκτιμήτρια ελαχίστων τετραγώνων. Ωστόσο η λύση της προσθέτει μία θετική σταθερά στα διαγώνια στοιχεία του $\boldsymbol{X}'\boldsymbol{X}$ πριν αντιστραφεί. Σε αυτή την περίπτωση για κάθε πίνακα $\boldsymbol{X}$, η ποσότητα $\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I}$ είναι πάντα αντιστρέψιμη και έτσι υπάρχει πάντα μία μοναδική λύση $\widehat{\boldsymbol{\beta}}_R$. Ακόμα από τη λύση της εκτιμήτριας *Ridge* μπορούμε έυκολα να συμπεράνουμε ότι στην περίπτωση όπου $\lambda = 0$, δηλαδή όταν δεν θέτουμε καμία ποινή, η εκτιμήτρια *Ridge* συμπίπτει με αυτή της μεθόδου ελαχίστων τετραγώνων. Ενώ όσο μεγαλύτερη γίνεται η τιμή του λ, τόσο αυξάνεται η μεροληψία της εκτιμήτριας καθώς ταυτόχρονα μειώνεται η διασπορά της. Επομένως, δεδομένου του γενικού τύπου του μέσου τετραγωνικού σφάλματος ***(4.2)*** που δείξαμε:

$$MSE = E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2 = Var(\widehat{\boldsymbol{\beta}}) + bias(\widehat{\boldsymbol{\beta}})^2,$$

η τιμή της παραμέτρου ποινής θα πρέπει να επιλεχθεί έτσι ώστε να «ισορροπεί» τη μεροληψία με τη διασπορά ώστε τελικά το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge* να είναι πολύ μικρότερο από αυτό της εκτιμήτριας ελαχίστων τετραγώνων.

Γενικά υπάρχει μια βέλτιστη τιμή της παραμέτρου ποινής για κάθε πρόβλημα αλλά είναι καλό αρχικά να εξετάσουμε γραφικά τη λύση που δίνει η παλινδρόμηση *Ridge* (θα δούμε παρακάτω) για ένα εύρος αποδεκτών τιμών. Όπως είδαμε και παραπάνω, αποδεκτή τιμή σημαίνει τιμή για την οποία η αντίστοιχη εκτιμήτρια έχει μικρότερο μέσο τετραγωνικό σφάλμα από αυτό της αντίστοιχης εκτιμήτριας ελαχίστων τετραγώνων.

Τέλος, σε αντίθεση με τη μέθοδο ελαχίστων τετραγώνων, η οποία παράγει μόνο ένα σύνολο εκτιμητριών για τους συντελεστές παλινδρόμησης, η παλινδρόμηση *Ridge* σχηματίζει διαφορετικά σύνολα εκτιμητριών, για κάθε διαφορετική τιμή της παραμέτρου ποινής που δίνει ο αναλυτής. Ουσιαστικά, οι μέθοδοι ποινικοποιημένης παλινδρόμησης παράγουν μια αλληλουχία από μοντέλα π.χ $M_0, \dots, M_\lambda$ για τις αντίστοιχες διαφορετικές τιμές της παραμέτρου ποινής που δίνονται από τον αναλυτή. Το να διαλέξουμε μία ιδανική τιμή της παραμέτρου ποινής αποτελεί το πιο σημαντικό σήμειο, βάση της οποίας γίνεται και η συρρίκνωση των συντελεστών του μοντέλου παλινδρόμησης που επιδιώκουμε.

4.3.3 Ιδιότητες της εκτιμήτριας *Ridge*

Προκειμένου να αντιληφθούμε το λόγο που μελετάμε την εκτιμήτρια *Ridge* καθώς και την «υπεροχή» που έχει η κατασκευή και η χρήση της έναντι της εκτιμήτριας ελαχίστων τετραγώνων, αρκεί να μελετήσουμε ορισμένες ιδιότητές της, να την συγκρίνουμε ως προς την αποδοτικότητά της με την εκτιμήτρια ελαχίστων τετραγώνων, ώστε τελικά να καταλήξουμε αν και για ποιες τιμές της αγνώστου παραμέτρου λ, η $\widehat{\boldsymbol{\beta}}_R$ αποτελεί «καλύτερη» εκτιμήτρια.

Αρχικά, καλό θα ήταν να εκφράσουμε την $\widehat{\boldsymbol{\beta}}_R$ συναρτήσει της εκτιμήτριας ελαχίστων τετραγώνων, καθώς θα την χρειαστούμε παρακάτω για να δείξουμε τις ιδιότητες της $\widehat{\boldsymbol{\beta}}_R$:

$$\widehat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-\boldsymbol{1}}\boldsymbol{X}'\boldsymbol{y}$$

$$= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})\widehat{\boldsymbol{\beta}}_{LS}.$$

Θέτοντας:

$$\boldsymbol{D}_{\lambda} = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X}),$$

(4.11)

προκύπτει τελικά:

$$\widehat{\boldsymbol{\beta}}_{R} = \boldsymbol{D}_{\lambda}\widehat{\boldsymbol{\beta}}_{LS}.$$

(4.12)

Δηλαδή, από την ***(4.12)*** διαπιστώνουμε ότι η σχέση που συνδέει την $\widehat{\boldsymbol{\beta}}_{R}$ με την εκτιμήτρια ελαχίστων τετραγώνων είναι γραμμική.

i) Αναμενόμενη τιμή

Βέβαια, όπως έχουμε ήδη πει, το γεγονός που καταφύγαμε στην κατασκευή της εκτιμήτριας *Ridge* είναι ότι αποτελεί μεροληπτική εκτιμήτρια. Αυτό αποδεικνύεται εύκολα παίρνοντας την αναμενόμενη τιμή της $\widehat{\boldsymbol{\beta}}_{R}$ μέσω της σχέσης ***(4.12)***:

$$\begin{aligned} E\left(\widehat{\boldsymbol{\beta}}_{R}\right) &= E(\boldsymbol{D}_{\lambda}\widehat{\boldsymbol{\beta}}_{LS}) \\ &= \boldsymbol{D}_{\lambda}E(\widehat{\boldsymbol{\beta}}_{LS}) \\ &= \boldsymbol{D}_{\lambda}\boldsymbol{\beta}. \end{aligned}$$

Όπως έχουμε ορίσει τον $\boldsymbol{D}_{\lambda}$ από τη σχέση ***(4.11)***, παρατηρούμε ότι για $\lambda > 0$ έχουμε ότι $E\left(\widehat{\boldsymbol{\beta}}_{R}\right) = \boldsymbol{D}_{\lambda}\boldsymbol{\beta} \neq \boldsymbol{\beta}$. Άρα η $\widehat{\boldsymbol{\beta}}_{R}$ είναι πράγματι μεροληπτική εκτιμήτρια του $\boldsymbol{\beta}$.

Το γεγονός που καθιστά την $\widehat{\boldsymbol{\beta}}_{R}$ μεροληπτική είναι η ποσότητα $\lambda \boldsymbol{I}$ που προστίθεται στον πίνακα $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ και η εκτίμηση εδώ βασίζεται πλέον στον πίνακα $(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}$, $\lambda > 0$. Επομένως, αφού εκφράσουμε τον πίνακα $(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}$ με ιδιοτιμές και τις συγκρίνουμε με τις ιδιοτιμές του $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, εξετάζουμε την «προσφορά» της παραμέτρου λ και το πώς

η εισαγωγή της βοηθάει ώστε να αντιμετωπιστεί το πρόβλημα της πολυσυγγραμμικότητας.

Εδώ να θυμήσουμε ότι οι ιδιοτιμές του $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ είναι $\frac{1}{v_j}, j = 1, \dots, p$ με v_j να είναι οι ιδιοτιμές του $\boldsymbol{X}'\boldsymbol{X}$.

Για τις ιδιοτιμές του $(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$, εκφράζοντας τον πίνακα $(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$ με την διαγωνοποιημένη μορφή ($\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{P}\boldsymbol{V}\boldsymbol{P}'$) έχουμε:

$$(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} = (\boldsymbol{P}\boldsymbol{V}\boldsymbol{P}' + \lambda\boldsymbol{P}\boldsymbol{P}')^{-1}$$

$$= \boldsymbol{P}(\boldsymbol{V} + \lambda\boldsymbol{I})^{-1}\boldsymbol{P}'$$

$$= \boldsymbol{P}diag\left(\frac{1}{v_j + \lambda}\right)\boldsymbol{P}'. \qquad \textbf{(4.13)}$$

Ο $diag(\frac{1}{v_j+\lambda})$ παριστάνει έναν διαγώνιο πίνακα με j −οστό διαγώνιο στοιχείο το $\frac{1}{v_j+\lambda}$.

Επομένως, οι ιδιοτιμές του $(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$ είναι $\frac{1}{v_j+\lambda}, j = 1, \dots, p$.

Δηλαδή παρατηρούμε ότι η μόνη διαφορά των ιδιοτιμών του $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ που εμπλέκεται στη λύση των εκτιμητριών ελαχίστων τετραγώνων με αυτές του $(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$, ο οποίος εμπλέκεται στη λύση της εκτιμήτριας *Ridge*, είναι η προσθήκη της παραμέτρου λ. Επομένως, αυτό που κάνει η *Ridge* είναι ότι ουσιαστικά «στέλνει» όλες τις ιδιοτιμές μακριά από το μηδέν, αφού έχει ιδιοτιμές $\frac{1}{v_j+\lambda}$ με $v_j + \lambda > \lambda > 0$.

Γι΄αυτό το λόγο το να διαλέξουμε θετική παράμετρο ποινής $(\lambda > 0)$, όπως έχουμε αναφέρει, κάνει τον πίνακα να αντιστρέφεται.

Στην ιδανική περίπτωση όπου έχουμε ορθοκανονικό πίνακα σχεδιασμού, δηλαδή $\boldsymbol{X}'\boldsymbol{X} = (\boldsymbol{X}'\boldsymbol{X})^{-1} = \boldsymbol{I}$, η σχέση που συνδέει της εκτιμήτρια ελαχίστων τετραγώνων με τη *Ridge* είναι:

$$\widehat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$= (\boldsymbol{I} + \lambda \boldsymbol{I})^{-1} \boldsymbol{X}' \boldsymbol{y}$$

$$= (1 + \lambda)^{-1} \boldsymbol{I}\, \boldsymbol{X}' \boldsymbol{y}$$

$$= (1 + \lambda)^{-1} (\boldsymbol{X}' \boldsymbol{X})^{-1}\, \boldsymbol{X}' \boldsymbol{y}$$

$$= \frac{\widehat{\boldsymbol{\beta}}_{LS}}{1 + \lambda}.$$

Και από εδώ είναι σαφές ότι όσο αυξάνεται το λ, η $\widehat{\boldsymbol{\beta}}_R$ συρρικνώνεται και έχουμε ότι $\left|\widehat{\boldsymbol{\beta}}_{LS}\right| > \left|\widehat{\boldsymbol{\beta}}_R\right|$.

Παίρνοντας το όριο της αναμενόμενης τιμής της $\widehat{\boldsymbol{\beta}}_R$ καθώς το λ τείνει στο άπειρο θα διαπιστώσουμε ότι:

$$\lim_{\lambda \to \infty} E\left(\widehat{\boldsymbol{\beta}}_R\right) = \lim_{\lambda \to \infty} (\boldsymbol{X}' \boldsymbol{X} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{X}' \boldsymbol{X}) \boldsymbol{\beta} = \boldsymbol{0}.$$

Γεγονός που αποδεικνύει ότι καθώς αυξάνεται η τιμή της παραμέτρου ποινής λ, οι συντελεστές παλινδρόμησης συρρικνώνονται προς το μηδέν. Βέβαια, αυτή η συμπεριφορά δεν είναι αυστηρά μονότονη στο λ. Δηλαδή αν $\lambda_a > \lambda_b$ δεν συνεπάγεται απαραίτητα ότι $\left|\widehat{\boldsymbol{\beta}}_j(\lambda_a)\right| < \left|\widehat{\boldsymbol{\beta}}_j(\lambda_b)\right|$.

Τέλος, η αναμενόμενη τιμή της $\widehat{\boldsymbol{\beta}}_R$ στην ορθοκανονική περίπτωση είναι:

$$E(\widehat{\boldsymbol{\beta}}_R) = E\left(\frac{\widehat{\boldsymbol{\beta}}_{LS}}{1 + \lambda}\right)$$

$$= \frac{1}{1 + \lambda} E(\widehat{\boldsymbol{\beta}}_{LS})$$

$$= \frac{1}{1 + \lambda} \boldsymbol{\beta}.$$

ii) Μεροληψία

Εφ’όσον δείξαμε ότι η εκτιμήτρια *Ridge* είναι μεροληπτική θα πρέπει να υπολογίσουμε την μεροληψία της.

Υπενθυμίζοντας ότι η μεροληψία ενός εκτιμητή $\widehat{\boldsymbol{\beta}}$ δίνεται από τον τύπο:

$$bias = E\left(\widehat{\boldsymbol{\beta}}\right) - \boldsymbol{\beta},$$

για την $\widehat{\boldsymbol{\beta}}_R$ θα έχουμε:

$$\begin{aligned} bias\left(\widehat{\boldsymbol{\beta}}_R\right) &= E\left(\widehat{\boldsymbol{\beta}}_R\right) - \boldsymbol{\beta} \\ &= \boldsymbol{D}_\lambda \boldsymbol{\beta} - \boldsymbol{\beta} \\ &= ((\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{X} - \boldsymbol{I})\boldsymbol{\beta} \qquad \textbf{(4.14 α)} \\ &= (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X} - (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I}))\boldsymbol{\beta} \\ &= -\lambda(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{\beta}. \qquad \textbf{(4.14 β)} \end{aligned}$$

Από την ***(4.14 β)*** βλέπουμε ότι η μεροληψία είναι ανάλογη του λ. Όσο αυξάνεται η τιμή του λ, μεγαλώνει ταυτόχρονα και η μεροληψία της $\widehat{\boldsymbol{\beta}}_R$ συναρτήσει του $\boldsymbol{\beta}$.

Ο τύπος της μεροληψίας ***(4.14 β)*** μέσω του τύπου ***(4.13)*** παίρνει την ακόλουθη μορφή με τη χρήση ιδιοτιμών:

$$bias\left(\widehat{\boldsymbol{\beta}}_R\right) = -\lambda\, \boldsymbol{P} diag\left(\frac{1}{v_j + \lambda}\right)\boldsymbol{P}'\boldsymbol{\beta}, j = 1, \dots, p.$$

Ωστόσο, όπως δείξαμε στην ***(4.2)*** $MSE = E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2 = Var\left(\widehat{\boldsymbol{\beta}}\right) + bias(\widehat{\boldsymbol{\beta}})^2$, επομένως αυτό που έχει ενδιαφέρον να μελετήσουμε είναι η τετραγωνισμένη μεροληψία $bias\left(\widehat{\boldsymbol{\beta}}_R\right)^2$. Όταν έχουμε διανυσματικές τιμές η τετραγωνισμένη μεροληψία παίρνει τη μορφή:

$$bias\left(\widehat{\boldsymbol{\beta}}_R\right)bias\left(\widehat{\boldsymbol{\beta}}_R\right)' = (\boldsymbol{D}_\lambda - \boldsymbol{I})\boldsymbol{\beta}\boldsymbol{\beta}'(\boldsymbol{D}_\lambda - \boldsymbol{I})', \qquad \textbf{(4.15 α)}$$

δεδομένης της σχέσης ***(4.14 α)***

ή

$$bias\left(\widehat{\boldsymbol{\beta}}_R\right)bias\left(\widehat{\boldsymbol{\beta}}_R\right)' = (-\lambda(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\beta})(-\lambda(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\beta})'$$

$$= \lambda^2(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}\boldsymbol{\beta}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1},$$

(4.15 β)

από τη σχέση ***(4.14 β)***.

Ο $(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1}$ είναι συμμετρικός δηλαδή:

$$((\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1})' = (\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-1},$$

δεδομένου ότι ο $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ είναι συμμετρικός, αφού:

$$((\boldsymbol{X}'\boldsymbol{X})^{-1})' = ((\boldsymbol{X}'\boldsymbol{X})')^{-1} = (\boldsymbol{X}'\boldsymbol{X})^{-1}.$$

Επιπλέον, η συνολική τετραγωνισμένη μεροληψία $tr\left(bias\left(\widehat{\boldsymbol{\beta}}_R\right)bias\left(\widehat{\boldsymbol{\beta}}_R\right)'\right)$ όπως δείξαμε από τον την σχέση ***(4.4)*** θα είναι:

$$bias\left(\widehat{\boldsymbol{\beta}}_R\right)'bias\left(\widehat{\boldsymbol{\beta}}_R\right) = \boldsymbol{\beta}'(\boldsymbol{D}_\lambda-\boldsymbol{I})'(\boldsymbol{D}_\lambda-\boldsymbol{I})\boldsymbol{\beta},$$

από τη σχέση ***(4.14 α)***

(4.16 α)

ή

$$bias\left(\widehat{\boldsymbol{\beta}}_R\right)'bias\left(\widehat{\boldsymbol{\beta}}_R\right) = \lambda^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X}+\lambda\boldsymbol{I})^{-2}\boldsymbol{\beta},$$

από τη σχέση ***(4.14 β)***.

(4.16 β)

Μέσω της ***(4.16 β)*** και με τη βοήθεια της ***(4.13)*** μπορούμε να γράψουμε τη συνολική τετραγωνισμένη μεροληψία με μορφή ιδιοτιμών:

$$bias\left(\widehat{\boldsymbol{\beta}}_R\right)'bias\left(\widehat{\boldsymbol{\beta}}_R\right) = \lambda^2\boldsymbol{\beta}'\boldsymbol{P}diag\left(\frac{1}{(v_j+\lambda)^2}\right)\boldsymbol{P}'\boldsymbol{\beta}, j = 1,\dots,p$$

$$= \lambda^2\boldsymbol{\alpha}'diag\left(\frac{1}{(v_j+\lambda)^2}\right)\boldsymbol{\alpha}, j = 1,\dots,p$$

$$= \lambda^2 \sum_{j=1}^{p} \frac{{\alpha_j}^2}{(v_j + \lambda)^2},$$

(4.17)

με $\boldsymbol{\alpha} = \boldsymbol{P}'\boldsymbol{\beta}$, με $\boldsymbol{\alpha}$ να είναι το διάνυσμα του συντελεστή παλινδρόμησης με μορφή $\boldsymbol{y} = \boldsymbol{XP\alpha} + \varepsilon$.

Στην ορθοκανονική περίπτωση θα έχουμε:

$$bias\left(\widehat{\boldsymbol{\beta}}_R\right)' bias\left(\widehat{\boldsymbol{\beta}}_R\right) = \lambda^2 \boldsymbol{\beta}' \left(\boldsymbol{I} + \lambda \boldsymbol{I}\right)^{-2} \boldsymbol{\beta}$$

$$= \frac{\boldsymbol{\lambda^2}}{(1+\lambda)^{\boldsymbol{2}}} \boldsymbol{\beta}'\boldsymbol{\beta}.$$

(4.18)

Τέλος, αξίζει να σημειώσουμε ότι σύμφωνα με τους *Hoerl* και *Kennard* (1970) η συνολική τετραγωνισμένη μεροληψία αποτελεί μία μονότονα αύξουσα συνάρτηση του λ:

$$\frac{d\left(bias\left(\widehat{\boldsymbol{\beta}}_R\right)' bias\left(\widehat{\boldsymbol{\beta}}_R\right)\right)}{d\lambda} = 2\lambda \sum_{j=1}^{p} \frac{{\alpha_j}^2}{\left(v_j + \lambda\right)^2} - 2\lambda^2 \sum_{j=1}^{p} \frac{{\alpha_j}^2}{\left(v_j + \lambda\right)^3}$$

$$= 2\lambda \sum_{j=1}^{p} \frac{v_j {\alpha_j}^2}{\left(v_j + \lambda\right)^3} > 0, \forall \lambda > 0$$

και έχει εύρος $(0, \sum_{j=1}^{p} {\beta_j}^2)$:

$$\lim_{\lambda \to 0} \left(bias\left(\widehat{\boldsymbol{\beta}}_R\right)' bias\left(\widehat{\boldsymbol{\beta}}_R\right)\right) = \lim_{\lambda \to 0} \lambda^2 \sum_{j=1}^{p} \frac{{\alpha_j}^2}{(v_j + \lambda)^2} = 0$$

$$\lim_{\lambda \to \infty} \left(bias\left(\widehat{\boldsymbol{\beta}}_R\right)' bias\left(\widehat{\boldsymbol{\beta}}_R\right)\right) = \lim_{\lambda \to \infty} \lambda^2 \sum_{j=1}^{p} \frac{{\alpha_j}^2}{(v_j + \lambda)^2}$$

$$= \lim_{\lambda \to \infty} \sum_{j=1}^{p} \frac{{\alpha_j}^2}{(\frac{v_j}{\lambda} + 1)^2}$$

$$= \sum_{j=1}^{p} {a_j}^2$$

$$= \boldsymbol{\alpha' a}$$

$$= \boldsymbol{\beta' P' P \beta}$$

$$= \boldsymbol{\beta' \beta}$$

$$= \sum_{j=1}^{p} {\beta_j}^2.$$

Για να έχουμε «κερδίσει» λοιπόν μία καλύτερη εκτιμήτρια από αυτή των ελαχίστων τετραγώνων θα πρέπει αυτό το ποσό της συνολικής τετραγωνισμένης μεροληψίας να αντισταθμίζεται από τη μείωση της συνολικής διασποράς της εκτιμήτριας *Ridge* συγκριτικά με αυτή που έχει η εκτιμήτρια ελαχίστων τετραγώνων. Γι΄αυτό το λόγο, θα υπολογίσουμε τη διασπορά και τη συνολική διασπορά της εκτιμήτριας *Ridge.*

iii) Διασπορά

Για τον υπολογισμό της διασποράς έχουμε:

$$Var(\widehat{\boldsymbol{\beta}}_R) = E\left(\left(\widehat{\boldsymbol{\beta}}_R - E(\widehat{\boldsymbol{\beta}}_R)\right)\left(\widehat{\boldsymbol{\beta}}_R - E(\widehat{\boldsymbol{\beta}}_R)\right)'\right),$$

με $\widehat{\boldsymbol{\beta}}_R - E(\widehat{\boldsymbol{\beta}}_R) = (\boldsymbol{X'X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X'y} - \sigma^2(\boldsymbol{X'X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X'X\beta}$

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

$$= (\boldsymbol{X'X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X'}\varepsilon,$$

αφού $\varepsilon \sim N(0, \sigma^2)$.

Άρα $Var(\widehat{\boldsymbol{\beta}}_R) = (\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'Var(\varepsilon)\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}.$$

(4.19 α)

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$$

$$= \sigma^2\boldsymbol{D}_\lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}_\lambda'.$$

(4.19 β)

Ενώ αν γράψουμε την ***(4.19 α)*** με μορφή ιδιοτιμών θα έχουμε:

$$Var(\widehat{\boldsymbol{\beta}}_R) = \sigma^2(\boldsymbol{P}\boldsymbol{V}\boldsymbol{P}' + \lambda\boldsymbol{P}\boldsymbol{P}')^{-1}\boldsymbol{P}\boldsymbol{V}\boldsymbol{P}'(\boldsymbol{P}\boldsymbol{V}\boldsymbol{P}' + \lambda\boldsymbol{P}\boldsymbol{P}')^{-1}$$

$$= \sigma^2\boldsymbol{P}(\boldsymbol{V} + \lambda\boldsymbol{I})^{-1}\boldsymbol{V}(\boldsymbol{V} + \lambda\boldsymbol{I})^{-1}\boldsymbol{P}'$$

$$= \sigma^2\boldsymbol{P}diag\left(\frac{v_j}{(v_j + \lambda)^2}\right)\boldsymbol{P}', j = 1 \dots, p.$$

(4.20)

Τώρα χρησιμοποιώντας την ***(4.20)*** εύκολα μπορούμε να γράψουμε τη συνολική διασπορά ως εξής:

$$\sum_{j=1}^{p} Var\left((\hat{\beta}_R)_j\right) = tr\left(Var(\widehat{\boldsymbol{\beta}}_R)\right)$$

$$= \sigma^2 tr\left(\boldsymbol{P}diag\left(\frac{v_j}{(v_j + \lambda)^2}\right)\boldsymbol{P}'\right)$$

$$= \sigma^2 tr\left(diag\left(\frac{v_j}{(v_j + \lambda)^2}\right)\boldsymbol{P}'\boldsymbol{P}\right)$$

$$= \sigma^2\sum_{j=1}^{p}\frac{v_j}{(v_j + \lambda)^2}.$$

(4.21)

Παίρνοντας το όριο της διασποράς $\widehat{\boldsymbol{\beta}}_R$ με το λ να τείνει στο άπειρο καταλήγουμε στο ότι με την αύξηση της τιμής του λ, η διασπορά των εκτιμώμενων συντελεστών της εκτιμήτριας *Ridge* πηγαίνει προς το μηδέν:

$$\lim_{\lambda\to\infty} Var\left(\widehat{\boldsymbol{\beta}}_R\right) = \lim_{\lambda\to\infty} \sigma^2 \boldsymbol{D}_\lambda (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{D}_\lambda' = 0.$$

Στην ορθοκανονική περίπτωση θα έχουμε:

$$Var\left(\widehat{\boldsymbol{\beta}}_R\right) = \sigma^2 (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1} (\boldsymbol{X}'\boldsymbol{X}) (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}$$

$$= \sigma^2 (\boldsymbol{I} + \lambda \boldsymbol{I})^{-1}\, \boldsymbol{I}\, (\boldsymbol{I} + \lambda \boldsymbol{I})^{-1}$$

$$= \sigma^2 (1+\lambda)^{-2}\, \boldsymbol{I}$$

και η συνολική διασπορά θα ισούται με:

$$tr\left(Var\left(\widehat{\boldsymbol{\beta}}_R\right)\right) = \frac{\sigma^2}{(1+\lambda)^2} tr\{\boldsymbol{I}\} = \frac{p\sigma^2}{(1+\lambda)^2}. \qquad \textbf{(4.22)}$$

Τέλος, να επισημάνουμε ότι σε αντίθεση με τη συνολική τετραγωνισμένη μεροληψία, η συνολική διασπορά αποτελεί μία μονότονα φθίνουσα συνάρτηση του λ:

$$\frac{dtr\left(Var\left(\widehat{\boldsymbol{\beta}}_R\right)\right)}{d\lambda} = -2\sigma^2 \sum_{j=1}^{p} \frac{v_j}{\left(v_j+\lambda\right)^3} < 0, \forall \lambda > 0,$$

με εύρος $(\sigma^2 \sum_{j=1}^{p} \frac{1}{v_j}, 0)$. Συγκεκριμένα λόγω της σχέσης ***(4.21)*** έχουμε:

$$\lim_{\lambda\to 0} tr\left(Var\left(\widehat{\boldsymbol{\beta}}_R\right)\right) = \lim_{\lambda\to 0} \sigma^2 \sum_{j=1}^{p} \frac{v_j}{(v_j+\lambda)^2} = \sigma^2 \sum_{j=1}^{p} \frac{1}{v_j}.$$

$$\lim_{\lambda\to\infty} tr\left(Var\left(\widehat{\boldsymbol{\beta}}_R\right)\right) = \lim_{\lambda\to\infty} \sigma^2 \sum_{j=1}^{p} \frac{v_j}{(v_j+\lambda)^2} = 0.$$

Ωστόσο, το αξιοσημείωτο με τη *Ridge* είναι ότι η μείωση του ποσοστού της συνολικής διασποράς καθώς το λ τείνει στο μηδέν που είναι:

$$\lim_{\lambda \to 0} \left| \frac{d}{d\lambda} \sum_{j=1}^{p} tr\left(Var\left(\widehat{\boldsymbol{\beta}}_R\right)\right) \right| = 2\sigma^2 \sum_{j=1}^{p} \frac{1}{{v_j}^2},$$

μπορεί να είναι πολύ μεγάλη για προβλήματα με μεγάλο βαθμό πολυσυγγραμμικότητας. Επομένως, αυτός είναι και ο κύριος λόγος που η *Ridge* είναι περισσότερο αποτελεσματική για προβλήματα με μεγάλο αριθμό γραμμικών συσχετίσεων.

iv) Μέσο τετραγωνικό σφάλμα

Οι *Hoerl* και *Kennard* (1970) έδειξαν ότι χρησιμοποιώντας την $\widehat{\boldsymbol{\beta}}_R$ μπορεί να επιτευχθεί βελτίωση στο μέσο τετραγωνικό σφάλμα.

Από την ***(4.4)*** δείξαμε ότι το MSE ορίζεται ως:

$$tr(M) = tr\left(Var\left(\widehat{\boldsymbol{\beta}}\right)\right) + bias\left(\widehat{\boldsymbol{\beta}}\right)' bias\left(\widehat{\boldsymbol{\beta}}\right),$$

και αντικαθιστώντας τις ***(4.19 β)*** και ***(4.14 β)*** αντίστοιχα έχουμε:

$$MSE\left(\widehat{\boldsymbol{\beta}}_R\right) = \sigma^2 tr(\boldsymbol{D}_\lambda (\boldsymbol{X}'\boldsymbol{X})^{-1} \boldsymbol{D}_\lambda') + \lambda^2 \boldsymbol{\beta}' (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1} \boldsymbol{\beta}.$$

(4.23)

Πράγματι, θα δείξουμε αναλυτικά ότι ξεκινώντας από τον ορισμό του μέσου τετραγωνικού σφάλματος, αναμενόμενη τιμή της τετραγωνικής απόστασης της $\widehat{\boldsymbol{\beta}}$ από την πραγματική της τιμή, καταλήγουμε στην ***(4.23)***:

$$MSE\left(\widehat{\boldsymbol{\beta}}_R\right) = E\left({L_1}^2\right)$$

$$= E\left(\left(\widehat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}\right)' \left(\widehat{\boldsymbol{\beta}}_R - \boldsymbol{\beta}\right)\right)$$

$$= E\left(\left(\boldsymbol{D}_\lambda \widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)' \left(\boldsymbol{D}_\lambda \widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)\right)$$

$$= E\left(\widehat{\boldsymbol{\beta}}_{LS}{}'\boldsymbol{D}_\lambda{}'\boldsymbol{D}_\lambda\widehat{\boldsymbol{\beta}}_{LS}\right) - E\left(\boldsymbol{\beta}'\boldsymbol{D}_\lambda\widehat{\boldsymbol{\beta}}_{LS}\right) - E\left(\widehat{\boldsymbol{\beta}}_{LS}{}'\boldsymbol{D}_\lambda{}'\boldsymbol{\beta}\right) + E(\boldsymbol{\beta}'\boldsymbol{\beta})$$

$$= E\left(\widehat{\boldsymbol{\beta}}_{LS}{}'\boldsymbol{D}_\lambda{}'\boldsymbol{D}_\lambda\widehat{\boldsymbol{\beta}}_{LS}\right) - E\left(\boldsymbol{\beta}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\widehat{\boldsymbol{\beta}}_{LS}\right) - E\left(\widehat{\boldsymbol{\beta}}_{LS}{}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\boldsymbol{\beta}\right)$$
$$+E(\boldsymbol{\beta}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\boldsymbol{\beta}) - E(\boldsymbol{\beta}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\boldsymbol{\beta}) + E\left(\boldsymbol{\beta}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\widehat{\boldsymbol{\beta}}_{LS}\right)$$
$$+E\left(\widehat{\boldsymbol{\beta}}_{LS}{}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\boldsymbol{\beta}\right) - E\left(\boldsymbol{\beta}'\boldsymbol{D}_\lambda\widehat{\boldsymbol{\beta}}_{LS}\right) - E\left(\widehat{\boldsymbol{\beta}}_{LS}{}'\boldsymbol{D}_\lambda{}'\boldsymbol{\beta}\right) + E(\boldsymbol{\beta}'\boldsymbol{\beta})$$

$$= E\left(\left(\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)'\boldsymbol{D}_\lambda{}'\boldsymbol{D}_\lambda\left(\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)\right)$$
$$-\boldsymbol{\beta}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\boldsymbol{\beta} + \boldsymbol{\beta}'\boldsymbol{D}_\lambda'\boldsymbol{D}_\lambda\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{D}_\lambda\boldsymbol{\beta} - \boldsymbol{\beta}'\boldsymbol{D}_\lambda'\boldsymbol{\beta}$$
$$+\boldsymbol{\beta}'\boldsymbol{\beta}$$

$$= E\left(\left(\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)'\boldsymbol{D}_\lambda{}'\boldsymbol{D}_\lambda\left(\widehat{\boldsymbol{\beta}}_{LS} - \boldsymbol{\beta}\right)\right) + \boldsymbol{\beta}'(\boldsymbol{D}_\lambda - \boldsymbol{I})'(\boldsymbol{D}_\lambda - \boldsymbol{I})\boldsymbol{\beta}$$

$$= \sigma^2 tr(\boldsymbol{D}_\lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}_\lambda') + \lambda^2\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-2}\boldsymbol{\beta}.$$

(4.23)

Στην τελευταία σειρά της ***(4.23)*** για τον πρώτο όρο χρησιμοποιήσαμε ότι $\widehat{\boldsymbol{\beta}}_{LS}{\sim}N(\boldsymbol{\beta}, \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1})$ και ότι η αναμενόμενη τιμή μίας τυχαίας πολυμεταβλητής $\boldsymbol{\varepsilon}{\sim}N(\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}, \boldsymbol{\Sigma}_\varepsilon)$ είναι $E(\boldsymbol{\varepsilon}'\boldsymbol{\Lambda}\boldsymbol{\varepsilon}) = tr(\boldsymbol{\Lambda}\boldsymbol{\Sigma}_\varepsilon) + \boldsymbol{\mu}_{\boldsymbol{\varepsilon}}{}'\boldsymbol{\Lambda}\boldsymbol{\mu}_{\boldsymbol{\varepsilon}}$, αντικαθιστώντας το $\boldsymbol{\varepsilon}$ με $\widehat{\boldsymbol{\beta}}_{LS}$. Ενώ για τον δεύτερο όρο κάναμε χρήση της ***(4.14 β)***.

Με μορφή ιδιοτιμών η ***(4.23)*** μπορει να γραφεί με τη βοήθεια των ***(4.21)*** και ***(4.17)*** αντίστοιχα ως εξής:

$$MSE\left(\widehat{\boldsymbol{\beta}}_R\right) = \sigma^2\sum_{j=1}^{p}\frac{v_j}{(v_j+\lambda)^2} + \lambda^2\sum_{j=1}^{p}\frac{{\alpha_j}^2}{(v_j+\lambda)^2}.$$

(4.24)

Τέλος για την ορθοκανονική περίπτωση το μέσο τετραγωνικό σφάλμα θα δίνεται μέσω των **(4.22)** και **(4.18)** αντίστοιχα από:

$$MSE\left(\widehat{\boldsymbol{\beta}}_R\right) = \frac{p\sigma^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\boldsymbol{\beta}'\boldsymbol{\beta}.$$

(4.25)

4.3.4 Σύγκριση *Ridge* με τη μέθοδο ελαχίστων τετραγώνων

i) Η παλινδρόμηση *Ridge* παράγει μικρότερες διασπορές για τους συντελεστές της από ότι αυτές που παράγουν τα ελάχιστα τετράγωνα.

Χρησιμοποιώντας τους τύπους των διασπορών όπως έχουμε δείξει για κάθε εκτιμήτρια θα αποδείξουμε ότι όντως ισχύει $Var(\widehat{\boldsymbol{\beta}}_{LS}) \geq Var(\widehat{\boldsymbol{\beta}}_{R})$:

$$Var(\widehat{\boldsymbol{\beta}}_{LS}) - Var(\widehat{\boldsymbol{\beta}}_{R}) = \sigma^2((\boldsymbol{X}'\boldsymbol{X})^{-1} - \boldsymbol{D}_\lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}_\lambda{}')$$

$$= \sigma^2\boldsymbol{D}_\lambda((\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{X})^{-1}(\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1})' - (\boldsymbol{X}'\boldsymbol{X})^{-1})\boldsymbol{D}_\lambda'$$

$$= \sigma^2\boldsymbol{D}_\lambda\big(((\boldsymbol{X}'\boldsymbol{X})^{-1} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-2})(\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1})' - (\boldsymbol{X}'\boldsymbol{X})^{-1}\big)\boldsymbol{D}_\lambda{}'$$

$$= \sigma^2\boldsymbol{D}_\lambda(2\lambda(\boldsymbol{X}'\boldsymbol{X})^{-2} + \lambda^2(\boldsymbol{X}'\boldsymbol{X})^{-3})\boldsymbol{D}_\lambda{}'$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(2\lambda\boldsymbol{I} + \lambda^2(\boldsymbol{X}'\boldsymbol{X})^{-1})(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} \geq \boldsymbol{0},$$

ως γινόμενο μη αρνητικών πινάκων.

(4.26)

Επιπλέον συγκρίνοντας και τις συνολικές διασπορές εύκολα βλέπουμε ότι:

$$tr\left(Var(\widehat{\boldsymbol{\beta}}_{LS})\right) \geq tr\left(Var(\widehat{\boldsymbol{\beta}}_{R})\right).$$

Απόδειξη:

Από τις ***(4.5)*** και ***(4.21)*** έχουμε αντίστοιχα:

$$tr\left(Var(\widehat{\boldsymbol{\beta}}_{LS})\right) = \sigma^2\sum_{j=1}^{p}\frac{1}{v_j}$$

και

$$tr\left(Var(\widehat{\boldsymbol{\beta}}_{R})\right) = \sigma^2\sum_{j=1}^{p}\frac{v_j}{(v_j+\lambda)^2},$$

με $\sigma^2 \sum_{j=1}^{p} \frac{1}{v_j} \geq \sigma^2 \sum_{j=1}^{p} \frac{v_j}{(v_j+\lambda)^2}$.

Επομένως, η εκτιμήτρια *Ridge* είναι λιγότερο ευαίσθητη σε αλλαγές του δείγματος συγκριτικά με την εκτιμήτρια ελαχίστων τετραγώνων. Παρ'όλα αυτά, η συμπεριφορά της *Ridge* δεν είναι το ίδιο καλή όσον αφορά τον συντελεστή προσδιορισμού.

ii) Η παλινδρόμηση *Ridge* δίνει μικρότερο μέσο τετραγωνικό σφάλμα σε σχέση με αυτό που δίνουν τα ελάχιστα τετράγωνα.

Το κριτήριο της παλινδρόμησης *Ridge* βασίζεται σε μία διαδικασία εκτίμησης η οποία ελαχιστοποιεί το μέσο τετραγωνικό σφάλμα, το οποίο όπως έχουμε αναφέρει αποτελεί την αναμενόμενη τετραγωνική απόσταση του εκάστοτε εκτιμητή από την πραγματική του τιμή ($MSE = E(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})^2$). Υπάρχει ένα μεγάλο εύρος τιμών της παραμέτρου ποινής το οποίο παρέχει ένα μεγάλο σύνολο από εκτιμήτριες *Ridge* και παράγει μικρότερο μέσο τετραγωνικό σφάλμα από αυτό της εκτιμήτριας ελαχίστων τετραγώνων. Υπό αυτή την έννοια η εκτιμήτρια *Ridge* αποτελεί καλύτερη εκτιμήτρια από αυτή των ελαχίστων τετραγώνων. Περισσότερα για τις τιμές της παραμέτρου ποινής όπου επιτυγχάνεται $MSE(\widehat{\boldsymbol{\beta}}_R) < MSE(\widehat{\boldsymbol{\beta}}_{LS})$ θα δούμε στα παρακάτω θεωρήματα (υποκεφάλαια **4.3.5.2**, **4.3.5.3** και **4.3.5.4**).

4.3.5 Μέθοδοι επιλογής της παραμέτρου ποινής

Η επιλογή της τιμής της παραμέτρου ποινής παίζει τον σημαντικότερο ρόλο στην παλινδρόμηση *Ridge*. Παρακάτω θα δούμε μία τεχνική καθώς και μερικές τιμές που έχουν προταθεί για την τιμή του λ.

4.3.5.1 Ίχνος Κορυφογραμμής (*Ridge Trace*)

Αρχικά, μπορεί να γίνει με γραφικό τρόπο παρατηρώντας το ίχνος κορυφογραμμής. Η διαδικασία του ίχνους κορυφογραμμής προτάθηκε από τον *Hoerl* (1962). Πρόκειται για μία γραφική παράσταση όπου παρουσιάζεται με μία καμπύλη (ίχνος) ο κάθε συντελεστής παλινδρόμησης του μοντέλου ως συνάρτηση της παραμέτρου ποινής. Ο αναλυτής θα πρέπει να επιλέξει μια τιμή της παραμέτρου ποινής όπου

παρατηρείται κάποια σταθεροποίηση των τιμών των συντελεστών αλλά ταυτόχρονα να μην δημιουργείται μεγάλη μεροληψία. Τελικά, η επιλογή αυτή θα πρέπει να έχει ως αποτέλεσμα να προκύψει ένα σύνολο σταθερών εκτιμητριών με μικρότερο μέσο τετραγωνικό σφάλμα σε σχέση με τις εκτιμήτριες ελαχίστων τετραγώνων, καθώς η ελάττωση που επιτυγχάνουν στη διασπορά σφάλματος θα είναι μεγαλύτερη από την αντιστάθμιση για την μεροληψία που εισάγεται.

Βέβαια για να είναι ευδιάκριτο το γράφημα και να καταλήξει ο αναλυτής στην επιλογή της παραμέτρου ποινής θα πρέπει στην γραφική παράσταση να μην υπάρχει μεγάλο πλήθος συντελεστών. Για το λόγο αυτό έγιναν προσπάθειες ώστε να εντοπίζουμε την κατάλληλη τιμή της παραμέτρου ποινής αριθμητικά, η οποία θα έχει ως αποτέλεσμα όχι μόνο σταθεροποίηση αλλά και την επιζητούμενη συρρίκνωση ορισμένων συντελεστών προς το μηδέν.

4.3.5.2 Θεώρημα ύπαρξης *Hoerl* και *Kennard*

Μία από αυτές τις υπολογιστικές μεθόδους για την επιλογή της παραμέτρου ποινής είναι αυτή που πρότειναν οι *Hoerl* και *Kennard* (1970). Σύμφωνα με το θεώρημα ύπαρξης που ανέπτυξαν, όπου μελέτησαν τα μέσα τετραγωνικά σφάλματα της εκτιμήτριας ελαχίστων τετραγώνων και της *Ridge*, απέδειξαν ότι:

Θεώρημα 4.1:

Υπάρχει πάντα ένα $\lambda > 0$ τέτοιο ώστε:

$$MSE\left(\widehat{\boldsymbol{\beta}}_R\right) < MSE\left(\widehat{\boldsymbol{\beta}}_{LS}\right) = \sigma^2 \sum_{j=1}^{p} \frac{1}{v_j}.$$

Ορίζοντας $MSE\left(\widehat{\boldsymbol{\beta}}_R\right) = f(\lambda)$ και $MSE\left(\widehat{\boldsymbol{\beta}}_{LS}\right) = f(0)$ αρκεί να δείξουμε ότι $\exists\ \lambda_0$ τέτοιο ώστε $\forall \lambda \epsilon (0, \lambda_0)$ να ισχύει:

$$f'(\lambda) < 0$$

Θυμίζοντας από τη σχέση ***(4.24)*** ότι:

$$f(\lambda) = MSE\left(\widehat{\boldsymbol{\beta}}_R\right) = \sigma^2 \sum_{j=1}^{p} \frac{v_j}{(v_j+\lambda)^2} + \lambda^2 \sum_{j=1}^{p} \frac{{\alpha_j}^2}{(v_j+\lambda)^2},$$

τότε παίρνοντας την πρώτη παράγωγο ως προς λ θα έχουμε:

$$f'(\lambda) = \frac{dMSE\left(\widehat{\boldsymbol{\beta}}_R\right)}{d\lambda} = -2\sigma^2 \sum_{j=1}^{p} \frac{v_j}{\left(v_j+\lambda\right)^3} + 2\lambda \sum_{j=1}^{p} \frac{v_j {\alpha_j}^2}{\left(v_j+\lambda\right)^3}. \quad \textbf{(4.27)}$$

Επομένως, για να αποδείξουμε το θεώρημα αρκεί να δείξουμε ότι πάντα υπάρχει ένα $\lambda > 0$ τέτοιο ώστε η πρώτη παράγωγος του $MSE\left(\widehat{\boldsymbol{\beta}}_R\right)$ ($f'(\lambda)$) να είναι αρνητική, δηλαδή:

$$f'(\lambda) = \frac{dMSE\left(\widehat{\boldsymbol{\beta}}_R\right)}{d\lambda} < \frac{dMSE\left(\widehat{\boldsymbol{\beta}}_{LS}\right)}{d\lambda} = 0,$$

δηλαδή ότι η συνάρτηση $MSE\left(\widehat{\boldsymbol{\beta}}_R\right)$ είναι φθίνουσα.

Πράγματι, αφού η $f'(\lambda)$ είναι συνεχής συνάρτηση προκύπτει ότι:

$$\lim_{\lambda \to 0} f'(\lambda) = f'(0).$$

Άρα λαμβάνοντας υπόψιν ότι:

$$\lim_{\lambda \to 0} \frac{dMSE\left(\widehat{\boldsymbol{\beta}}_R\right)}{d\lambda} = -2\sigma^2 \sum_{j=1}^{p} \frac{1}{{v_j}^2} < 0,$$

παίρνουμε ότι $f'(0) < 0$.

Από το τελευταίο και τη συνέχεια της $f'(\lambda) = \frac{dMSE\left(\widehat{\boldsymbol{\beta}}_R\right)}{d\lambda}$ προκύπτει ότι θα υπάρχει $\lambda_0 > 0$ ώστε $\forall \lambda \epsilon (0, \lambda_0)$ η $f'(\lambda) < 0$.

Ακόμα, έδειξαν ότι η παραπάνω ανισότητα ισχύει για όλα τα $0 < \lambda < \lambda_{max} = \frac{\sigma^2}{{a^2}_{max}}$.

Υπολογίζοντας τη δεύτερη παράγωγο της ***(4.27)*** έχουμε:

$$\frac{d^2MSE(\widehat{\boldsymbol{\beta}}_R)}{d\lambda^2} = 6\sigma^2 \sum_{j=1}^{p} \frac{v_j}{(v_j+\lambda)^4} + 2\sum_{j=1}^{p} \frac{v_j {\alpha_j}^2}{(v_j+\lambda)^3} - 6\lambda \sum_{j=1}^{p} \frac{v_j {\alpha_j}^2}{(v_j+\lambda)^4}.$$

Παίρνοντας το όριο της με το λ να τείνει στο μηδέν έχουμε:

$$\lim_{\lambda\to 0} \frac{d^2MSE(\widehat{\boldsymbol{\beta}}_R)}{d\lambda^2} = 6\sigma^2 \sum_{j=1}^{p} \frac{1}{{v_j}^3} + 2\sum_{j=1}^{p} \frac{{\alpha_j}^2}{{v_j}^2} > 0.$$

Από θεώρημα ελαχίστου έχουμε ότι αφού $\lim_{\lambda\to 0} \frac{d^2MSE(\widehat{\boldsymbol{\beta}}_R)}{d\lambda^2} > 0$, θα υπάρχει μία ελάχιστη τιμή για το MSE την οποία και υπολογίζουμε.

Από την ***(4.27)*** προκύπτει ότι:

$$\begin{aligned} f'(\lambda) &= 2\sum_{j=1}^{p} \frac{v_j(\lambda {\alpha_j}^2 - \sigma^2)}{(v_j+\lambda)^3} \\ &\leq 2(\lambda {\alpha_{max}}^2 - \sigma^2)\sum_{j=1}^{p} \frac{v_j}{(v_j+\lambda)^3}, \end{aligned}$$

με α_{max} να είναι η μεγαλύτερη τιμή του $\boldsymbol{\alpha} = \boldsymbol{P}'\boldsymbol{\beta}$.

Επομένως, για να είναι η ***(4.27)*** αρνητική θα πρέπει:

$$\lambda < \frac{\sigma^2}{{\alpha_{max}}^2}.$$

4.3.5.3 Θεώρημα *Theobald*

Παρ'όλα αυτά, ο *Theobald* (1974) σχολίασε τον τρόπο με τον οποίο οι *Hoerl* και *Kennard* (1970) κατέληξαν στην τιμή της παραμέτρου ποινής (βλ. Θεώρημα ύπαρξης) και πρότεινε ένα πιο γενικό κριτήριο.

Πιο συγκεκριμένα, ο *Theobald* (1974) πρότεινε την χρήση του σταθμισμένου (*weighted*) μέσου τετραγωνικού σφάλματος $(WMSE)$, το οποίο δίνεται από τον τύπο:

$$WMSE(\widehat{\boldsymbol{\beta}}) = E\left((\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})'\boldsymbol{B}(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta})\right),$$

για κάθε μη αρνητικά ορισμένο πίνακα $\boldsymbol{B}$. Βέβαια, παρατηρούμε ότι όταν $\boldsymbol{B} = \boldsymbol{I}$ έχουμε το MSE.

Προτού αναπτύξουμε το θεώρημα του *Theobald* (1974) θα παραθέσουμε μία πρόταση πάνω στην οποία βασίστηκε για να καταλήξει στην επιζητούμενη τιμή του λ:

Πρόταση 4.1 (Farebrother, 1976):

Έστω $\boldsymbol{B}$ ένας $(p \times p)$ θετικά ορισμένος πίνακας, $\boldsymbol{b}$ ένα μη μηδενικό διάνυσμα p διάστασης και $c \epsilon \mathbb{R}_+$. Τότε:

$$c\boldsymbol{B} - \boldsymbol{b}\boldsymbol{b}' > 0 \text{ αν και μόνο αν } \boldsymbol{b}'\boldsymbol{B}^{-1}\boldsymbol{b} > c.$$

Ο *Theobald* (1974) προκειμένου να καταλήξει στην τιμή του λ για την οποία το σταθμισμένο μέσο τετραγωνικό σφάλμα της εκτιμήτριας ελαχίστων τετραγώνων είναι μεγαλύτερο από αυτό της εκτιμήτριας *Ridge* χρησιμοποίησε τους πίνακες διασποράς των σφαλμάτων. Συγκεκριμένα έδειξε ότι:

Θεώρημα 4.2:

Έστω ότι έχουμε δύο εκτιμήτριες $\widehat{\boldsymbol{\beta}}_1$ και $\widehat{\boldsymbol{\beta}}_2$ μιας παραμέτρου $\boldsymbol{\beta}$, οι οποίες έχουν πίνακες διασποράς των σφαλμάτων:

$M_k = E\left((\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})'\right)$, για $k = 1,2$

και σταθμισμένο μέσο τετραγωνικό σφάλμα:

$m_k = WMSE(\widehat{\boldsymbol{\beta}}_k) = E\left((\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})'\boldsymbol{B}(\widehat{\boldsymbol{\beta}}_k - \boldsymbol{\beta})\right)$, για $k = 1,2$

με $\boldsymbol{B} \geq \boldsymbol{0}$ και $m_k = tr\boldsymbol{B}M_k$.

Τότε $M_1 - M_2 \geq 0$ αν και μόνο αν $m_1 - m_2 \geq 0$ (ή $WMSE(\widehat{\boldsymbol{\beta}}_1) - WMSE(\widehat{\boldsymbol{\beta}}_2) \geq 0$), για όλα τα $\boldsymbol{B} \geq \mathbf{0}$.

Θα δείξουμε λοιπόν για ποια τιμή του λ ($\lambda > 0$) ισχύει ότι $M_1 - M_2 \geq 0$:

Λόγω της ***(4.3)*** θα έχουμε για την εκτιμήτρια ελαχίστων τετραγώνων και την εκτιμήτρια *Ridge* αντίστοιχα ότι:

$$M_1 = Var(\widehat{\boldsymbol{\beta}}_{LS}) = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$$

και

$$M_2 = Var(\widehat{\boldsymbol{\beta}}_R) + bias(\widehat{\boldsymbol{\beta}}_R)bias(\widehat{\boldsymbol{\beta}}_R)'$$

$$= \sigma^2\boldsymbol{D}_\lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}_\lambda' + (\boldsymbol{D}_\lambda - \boldsymbol{I})\boldsymbol{\beta}\boldsymbol{\beta}'(\boldsymbol{D}_\lambda - \boldsymbol{I})',$$

από τις σχέσεις ***(4.19 β)*** και ***(4.15 α)***.

Άρα:

$$M_1 - M_2 = \sigma^2(\boldsymbol{X}'\boldsymbol{X})^{-1} - \sigma^2\boldsymbol{D}_\lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{D}_\lambda' - (\boldsymbol{D}_\lambda - \boldsymbol{I})\boldsymbol{\beta}\boldsymbol{\beta}'(\boldsymbol{D}_\lambda - \boldsymbol{I})'$$

$$= \sigma^2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(2\lambda\,\boldsymbol{I} + \lambda^2(\boldsymbol{X}'\boldsymbol{X})^{-1})(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1} - \lambda^2(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{\beta}\boldsymbol{\beta}'(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}$$

$$= \lambda(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}(\sigma^2(2\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}) - \lambda\boldsymbol{\beta}\boldsymbol{\beta}')(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}.$$

Η δεύτερη ισότητα προκύπτει από τις σχέσεις ***(4.26)*** και της ***(4.16 β)***.

Για $\lambda > 0$ η ποσότητα αυτή είναι θετική αν και μόνο αν η:

$$\sigma^2(2\boldsymbol{I} + \lambda(\boldsymbol{X}'\boldsymbol{X})^{-1}) - \lambda\boldsymbol{\beta}\boldsymbol{\beta}'$$

είναι θετική και αυτό ισχύει αν η:

$$2\sigma^2\boldsymbol{I} - \lambda\boldsymbol{\beta}\boldsymbol{\beta}'$$

είναι θετική.

Από την *Πρόταση 4.1* θα πρέπει το λ να ικανοποιεί τη σχέση:

$$2\sigma^2(\boldsymbol{\beta}\boldsymbol{\beta}')^{-1} > \lambda.$$

Επομένως, σύμφωνα με τον *Theobald* $M_1 - M_2 \geq 0$ και ισοδύναμα $m_1 - m_2 \geq 0$ δηλαδή $WMSE(\widehat{\boldsymbol{\beta}}_{LS}) - WMSE(\widehat{\boldsymbol{\beta}}_R) \geq 0$ για:

$$\lambda < \frac{2\sigma^2}{\boldsymbol{\beta}\boldsymbol{\beta}'}.$$

4.3.5.4 Πρόταση από τους *Hoerl, Kennard* και *Baldwin* για βέλτιστη τιμή παραμέτρου ποινής

Ωστόσο, όπως βλέπουμε από τα Θεωρήματα 4.1 και 4.2 (Θεώρημα ύπαρξης *Hoerl* και *Kennard* και το Θεώρημα του *Theobald*) η τιμή της παραμέτρου ποινής εξαρτάται από τις ποσότητες σ^2,α_j και β_j με $j = 1, \dots, p$ που είναι άγνωστες.

Γι΄αυτό το λόγο οι *Hoerl*, *Kennard* και *Baldwin* (1975) πρότειναν μία βέλτιστη τιμή του λ, δείχνοντας μέσω προσομοιώσεων ότι υπάρχει μη μηδενική τιμή της παραμέτρου ποινής, με την ιδιότητα το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge* να είναι μικρότερο από αυτό των εκτιμητριών ελαχίστων τετραγώνων. Η τιμή αυτής της παραμέτρου ποινής δίνεται από τον τύπο:

$$\lambda = \frac{p\,\hat{\sigma}_R{}^2}{\widehat{\boldsymbol{\beta}}_R{}'\widehat{\boldsymbol{\beta}}_R},$$

όπου τα $\widehat{\boldsymbol{\beta}}_R$ και $\hat{\sigma}_R{}^2$βρίσκονται βάσει της μεθόδου ελαχίστων τετραγώνων και *p* είναι ο αριθμός των παραμέτρων στο μοντέλο.

Την πρόταση αυτή την δικαιολόγησαν από το γεγονός ότι στην ορθοκανονική περίπτωση το $MSE(\widehat{\boldsymbol{\beta}}_R)$, που υπενθυμίζουμε ότι από τη σχέση ***(4.25)*** είναι:

$$MSE\left(\widehat{\boldsymbol{\beta}}_R\right) = \frac{p\sigma^2}{(1+\lambda)^2} + \frac{\lambda^{\mathbf{2}}}{(1+\lambda)^{\mathbf{2}}}\boldsymbol{\beta}'\boldsymbol{\beta},$$

πετυχαίνει το ελάχιστο όταν $\lambda = \frac{p\sigma^2}{\boldsymbol{\beta}'\boldsymbol{\beta}}$.

Απόδειξη:

$$\frac{dMSE\left(\widehat{\boldsymbol{\beta}}_R\right)}{d\lambda} = \frac{d}{d\lambda}\left(\frac{p\sigma^2}{(1+\lambda)^2} + \frac{\lambda^{\mathbf{2}}}{(1+\lambda)^{\mathbf{2}}}\boldsymbol{\beta}'\boldsymbol{\beta}\right)$$

$$= \frac{d}{d\lambda}\left(\frac{p\sigma^2 + \lambda^{\mathbf{2}}\boldsymbol{\beta}'\boldsymbol{\beta}}{(1+\lambda)^2}\right)$$

$$= \frac{2\lambda\boldsymbol{\beta}'\boldsymbol{\beta}(1+\lambda)^2 - (p\sigma^2 + \lambda^{\mathbf{2}}\boldsymbol{\beta}'\boldsymbol{\beta})2(1+\lambda)}{(1+\lambda)^4}$$

$$= \frac{2\lambda\boldsymbol{\beta}'\boldsymbol{\beta}(1+\lambda) - 2p\sigma^2 - 2\lambda^{\mathbf{2}}\boldsymbol{\beta}'\boldsymbol{\beta}}{(1+\lambda)^3}$$

$$= \frac{2\lambda\boldsymbol{\beta}'\boldsymbol{\beta} + 2\lambda^{\mathbf{2}}\boldsymbol{\beta}'\boldsymbol{\beta} - 2p\sigma^2 - 2\lambda^{\mathbf{2}}\boldsymbol{\beta}'\boldsymbol{\beta}}{(1+\lambda)^3}$$

$$= \frac{2\lambda\boldsymbol{\beta}'\boldsymbol{\beta} - 2p\sigma^2}{(1+\lambda)^3}.$$

Για να έχουμε ελάχιστο πρέπει:

$$\frac{d}{d\lambda}\left(\frac{2\lambda\boldsymbol{\beta}'\boldsymbol{\beta}-2p\sigma^2}{(1+\lambda)^3}\right)=0.$$

Οπότε έχουμε:

$$2\lambda\boldsymbol{\beta}'\boldsymbol{\beta}-2p\sigma^2=0$$

ή

$$\lambda=\frac{p\sigma^2}{\boldsymbol{\beta}'\boldsymbol{\beta}}.$$

Ένας άλλος τρόπος για να υπολογίσουμε την κατάλληλη τιμή της παραμέτρου ποινής είναι με τη χρήση έμμεσων μεθόδων που χρησιμοποιούν κριτήρια σύγκρισης μοντέλων, όπως έχουμε δει στο Κεφάλαιο 2. Στην παρούσα διπλωματική θα χρησιμοποιήσουμε τα κριτήρια πληροφορίας AIC και BIC, καθώς και μία νέα μεθοδολογία, αυτή του *Cross Validation*, την οποία θα αναπτύξουμε πιο κάτω κατά την εξέταση της *LASSO* και θα την εφαρμόσουμε και για τις δύο μεθόδους με τη βοηθεία του πακέτου *glmnet*.

4.3.5.5 Κριτήρια πληροφορίας AIC και BIC

Στην περίπτωση που ο αναλυτής επιλέγει να χρησιμοποιήσει τα κριτήρια πληροφορίας, οφείλει να υπολογίσει τις τιμές τους που δίνονται από τις σχέσεις που είχαμε δει στο υποκεφάλαιο **2.2**:

$$AIC=n\ln(\frac{SSE}{n})+2\,df$$

$$BIC=n\,ln(\frac{SSE}{n})+\ln(n)\,df$$

όπου:

n: το πλήθος των παρατηρήσεων του μοντέλου,

SSE: το άθροισμα τετραγώνων των σφαλμάτων και

df: είναι οι βαθμοί ελευθερίας.

Υπενθυμίζοντας ότι η εκτιμήτρια *Ridge* δίνεται από τον τύπο:

$$\widehat{\boldsymbol{\beta}}_R = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y},$$

το μοντέλο που εκτιμούμε θα δίνεται από τη σχέση:

$$\widehat{\boldsymbol{y}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}_R$$

$$= \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'\boldsymbol{y}$$

$$= \boldsymbol{H}_\lambda \boldsymbol{y},$$

όπου $\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}'$.

Οι βαθμοί ελευθερίας (df) της παλινδρόμησης *Ridge* δίνονται από:

$$df = tr\boldsymbol{H}_\lambda = tr(\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-1}\boldsymbol{X}').$$

Αφού υπολογιστούν τα κριτήρια σύγκρισης, ο αναλυτής τελικά επιλέγει αυτή την τιμή της παραμέτρου που δίνει τη μικρότερη τιμή του εκάστοτε κριτηρίου πληροφορίας.

4.3.6 Τυποποίηση μεταβλητών

Θεωρούμε ότι έχουμε τα δεδομένα :

$$(\boldsymbol{x_i}, y_i), i = 1, \ldots, n,$$

$\boldsymbol{x_i}$= $(x_{i1}, x_{i2}, \ldots, x_{ip})'$: οι τιμές των ανεξάρτητων μεταβλητών και

y_i: οι τιμές της μεταβλητής αποκρίσης.

Πριν χρησιμοποιήσουμε την παλινδρόμηση *Ridge* (και γενικά κάθε ποινικοποιημένη μέθοδο) είναι απαραίτητο να προχωρήσουμε στην τυποποίηση των επεξηγηματικών μεταβλητών.

Αρχικά, καλό θα ήταν να διευκρινίσουμε ότι η συρρίκνωση εφαρμόζεται σε όλους τους συντελεστές των p επεξηγηματικών μεταβλητών και όχι στον σταθερό όρο, καθώς αποτελεί ένα μέτρο για τη μέση τιμή της μεταβλητής απόκρισης όταν οι υπόλοιπες επεξηγηματικές μεταβλητές είναι μηδέν. Επομένως υποθέτοντας ότι οι μεταβλητές, οι οποίες είναι οι στήλες του πίνακα σχεδιασμού $\boldsymbol{X}$, κεντράρονται ώστε να έχουν μέση τιμή μηδέν:

$$\sum_{i=1}^{n} \frac{x_{ij}}{n} = 0,$$

η εκτιμώμενη σταθερά θα έχει τη μορφή:

$$\hat{\beta}_0 = \bar{y} = \sum_{i=1}^{n} \frac{y_i}{n}.$$

Τέλος, αν γενικά δε θέλουμε να συμπεριλάβουμε την σταθερά στο μοντέλο παλινδρόμησης τότε κεντράρουμε και τις τιμές της εξαρτημένης μεταβλητής ώστε:

$$\sum_{i=1}^{n} \frac{y_i}{n} = \bar{y} = 0.$$

Ακόμα, πριν εφαρμόσουμε την παλινδρόμηση *Ridge* θα πρέπει να προχωρήσουμε στην τυποποίηση των μεταβλητών. Η παλινδρόμηση *Ridge*, όπως είδαμε, ποινικοποιεί την γραμμική παλινδρόμηση θέτοντας μια ποινή στους συντελεστές της. Με αυτόν τον τρόπο οι συντελεστές συρρικνώνονται τόσο προς το μηδέν όσο και μεταξύ τους. Όμως όταν συμβαίνει αυτό και στην περίπτωση που οι επεξηγηματικές μεταβλητές δεν έχουν την ίδια κλίμακα μέτρησης, η συρρίκνωση που υπόκεινται οι συντελεστές δε θα είναι «δίκαιη». Αυτό γιατί η κλίμακα των επεξηγηματικών μεταβλητών επηρεάζει το ποσοστό της ποινής που θα εφαρμοστεί στη συγκεκριμένη μεταβλητή. Με άλλα λόγια, δεδομένου ότι η ποινή αποτελεί άθροισμα τετραγώνων όλων των μεταβλητών, η

ύπαρξη μεταβλητών με διαφορετική κλίμακα θα έχει ως αποτέλεσμα και διαφορετική συνεισφορά της κάθε μίας στον περιορισμό ποινικοποίησης. Επομένως, για να αποφευχθεί αυτή η άνιση ποινή που δημιουργείται από τις μεταβλητές με διαφορετικές κλίμακες θα πρέπει να καταφεύγουμε στην μετατροπή των μεταβλητών ώστε να έχουν διασπορά 1.

4.3.7 Συμπεράσματα

Η παλινδρόμηση *Ridge* είναι πιο αποτελεσματική σε περιπτώσεις όπου οι εκτιμήτριες ελαχίστων τετραγώνων έχουν μεγάλη διασπορά (κατά την ύπαρξη του φαινομένου πολυσυγγραμμικότητας) λόγω της ειδικής κατασκευής της εκτιμήτριας *Ridge*, με την εισαγωγή της παραμέτρου ποινής. Ακόμα προσφέρει σημαντικό υπολογιστικό πλεονέκτημα έναντι της μεθόδου εύρεσης όλων των δυνατών γραμμικών μοντέλων και της μεθόδου επιλογής καλύτερου υποσυνόλου. Αυτό οφείλεται στο γεγονός ότι η παλινδρόμηση *Ridge* για κάθε δεδομένη τιμή της παραμέτρου ποινής προσαρμόζει μόνο ένα μοντέλο, η διαδικασία της οποίας υλοποιείται πολύ γρήγορα.

Ένα βασικό μειονέκτημα της παλινδρόμησης *Ridge*, ειδικά στην περίπτωση που το εξεταζόμενο μοντέλο αποτελείται από μεγάλο αριθμό επεξηγηματικών μεταβλητών είναι το γεγονός ότι δεν κάνει άμεση επιλογή μεταβλητών, αφού μέσω της διαδικασίας της γίνεται μόνο συρρίκνωση των συντελεστών (κυρίως των μη σημαντικών) προς το μηδέν, χωρίς να τις μηδενίζει. Ωστόσο από την άλλη, οι διαδικασίες κατά βήματα που έχουν την παραπάνω ικανότητα, να πετυχαίνουν δηλαδή τη δημιουργία ενός υποσυνόλου μεταβλητών, δεν εγγυώνται ότι αυτό θα είναι και τα «βέλτιστο», καθώς δεν είναι σίγουρο ότι οι εναπομείναντες μεταβλητές θα είναι αυτές που επιδρούν πιο έντονα στην μεταβλητή απόκρισης. Εξάλλου, σύμφωνα με τους *Marquardt* και *Snee* (1975) είναι προτιμότερο να χρησιμοποιούμε ένα μέρος της πληροφορίας σε όλες τις επεξηγηματικές μεταβλητές, όπως κάνει η *Ridge*, παρά όλη την πληροφορία μόνο σε ένα σύνολο των επεξηγηματικών μεταβλητών, όπως κάνουν οι διαδικασίες κατά βήματα.

Τελικά, η παλινδρόμηση *Ridge* μπορούμε να πούμε ότι πιθανόν να δίνει μια πρώτη εικόνα για το ποιες επεξηγηματικές μεταβλητές είναι λιγότερο σημαντικές, αυτές δηλαδή που συρρικνώνονται προς το μηδέν και

συντελεί στην δημιουργία μιας πιο αποτελεσματικής τεχνικής η οποία ακολουθώντας παρόμοια λογική δρα ως μέθοδος επιλογής μεταβλητών.

4.4 *LASSO*

4.4.1 Περιγραφή της μεθόδου

Όπως είδαμε στη παλινδρόμηση *Ridge* όσο και να αυξάνουμε τη τιμή της παραμέτρου ποινής ναι μεν πετυχαίνουμε τη μείωση του μεγέθους των συντελεστών αφού επιτυγχάνεται η συρρίκνωσή τους, αλλά ποτέ δεν καταλήγουμε στην διαγραφή κάποιας μεταβλητής από το αρχικό μοντέλο. Για την αντιμετώπιση αυτού του «προβλήματος», ο *Tibshirani* (1996) πρότεινε μια νέα και πιο αποτελεσματική μέθοδο, τη λεγόμενη *LASSO* (*Least Absolute Shrinkage and Selection Operator*). Η μέθοδος αυτή συρρικνώνει κάποιους συντελεστές ενώ παράλληλα άλλους τους θέτει ίσους με μηδέν, λειτουργώντας έτσι και ως μέθοδος επιλογής μεταβλητών. Και η μέθοδος αυτή εντάσσεται στο γενικότερο σύνολο των ποινικοποιημένων μεθόδων. Για το λόγο αυτό είναι αναγκαία η εφαρμογή της όταν παρατηρείται έντονη συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών του εκάστοτε υπό εξέταση μοντέλου.

Τέλος, καθώς και η *LASSO* αποτελεί μέθοδο με ποινή, η τυποποίηση των μεταβλητών συνίσταται πριν τη χρήση της, όπως ακριβώς αναφέραμε και για την παλινδρόμηση *Ridge* στο υποκεφάλαιο **4.3.6**.

4.4.2 Ποινικοποίηση της L_1-νόρμας

Με τη μέθοδο *LASSO*, όπως και με τη *Ridge*, επιτυγχάνεται η ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων ποινικοποιώντας όμως στην περίπτωση αυτή την L_1-νόρμα, δηλαδή το άθροισμα των απολύτων των συντελεστών παλινδρόμησης:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2$$

υπό τον περιορισμό:

$$\sum_{j=1}^{p}|\beta_j| \le t,$$

(4.28)

ή ισοδύναμα την ποσότητα:

$$\sum_{i=1}^{n}\left(y_i - \sum_{j=1}^{p} x_{ij}\,\beta_j\right)^2 + \lambda \sum_{j=1}^{p}\left|\beta_j\right|,$$

(4.29)

ή υπό μορφή πινάκων:

$$L(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1.$$

(4.30)

Η έννοια και η εισαγωγή της ρυθμιστικής παραμέτρου *t* στην ***(4.28)*** ή της παραμέτρου ποινής λ στην ***(4.29)*** λειτουργούν και έδω ως περιορισμοί στο άθροισμα των απολύτων των συντελεστών παλινδρόμησης, έτσι ώστε να ελεγχθούν οι μη λογικές μεγάλες τιμές που μπορούν να λάβουν κατά την παρουσία του φαινομένου της πολυσυγγραμμικότητας.

Όπως και στη παλινδρόμηση *Ridge*, όσο μεγαλύτερη είναι η τιμή της παραμέτρου ποινής τόσο πιο γρήγορα συρρικνώνονται οι συντελεστές προς το μηδέν. Ωστόσο, η κύρια διαφορά με τη *Ridge*, έγκειται στο γεγονός ότι κάποιοι συντελεστές γίνονται ακριβώς μηδέν. Αυτό είναι ιδιαίτερα βολικό όταν θέλουμε να πετύχουμε κάποια αυτόματη επιλογή μεταβλητών, όπως και στην παρούσα διπλωματική. Από την άλλη, όταν $t = max\sum_{j=1}^{p}\left|\beta_j\right|$ ή αντίστοιχα $\lambda = 0$, δεν επιτυγχάνουμε καμία συρρίκνωση στους συντελεστές οπότε και παίρνουμε τις εκτιμήτριες ελαχίστων τετραγώνων.

Η ελαχιστοποίηση της ποσότητας ***(4.29)*** δίνει την εκτιμήτρια *LASSO* ($\widehat{\boldsymbol{\beta}}_L$) με τον τρόπο που είδαμε κατά την παλινδρόμηση *Ridge* (πολλαπλασιαστές *Lagrange*). Σε αντίθεση με την εκτιμήτρια *Ridge*, εδώ δεν μπορούμε να βρούμε έναν σαφή τύπο για την *LASSO*, καθώς για την εύρεση της λύσης σε αυτό το πρόβλημα ελαχιστοποίησης θα ήταν να υπολογίσουμε την πρώτη παράγωγο ως προς $\boldsymbol{\beta}$ και να τη θέσουμε ίση με μηδέν. Όμως η απόλυτη τιμή της συνάρτησης $|\boldsymbol{\beta}|$ δεν έχει παράγωγο στο $\boldsymbol{\beta} = 0$. Ωστόσο, το πρόβλημα της *LASSO* αποτελεί ένα κυρτό πρόβλημα, ειδικά ένα πρόβλημα τετραγωνικού προγραμματισμού με κυρτό και μη

διαφορίσιμο περιορισμό. Γι΄αυτό και η εκτιμήτρια *LASSO* υπολογίζεται αποκλειστικά μέσα από στατιστικά πακέτα. Προκειμένου όμως να αντιληφθούμε τον τρόπο με τον οποίο δουλεύει η *LASSO* θα παραθέσουμε κάποιες προτάσεις σχετικά με τις κυρτές και μη διαφορίσιμες συναρτήσεις (βλ. ΠΑΡΑΡΤΗΜΑ Ι).

4.4.3 1η συνθήκη βελτιστότητας

Για την εύρεση της γενικής λύσης *LASSO* θα βασιστούμε στην 1η συνθήκη βελτιστότητας για κυρτές συναρτήσεις, σύμφωνα με την οποία:

Πρόταση 4.4:

Έστω μία κυρτή συνάρτηση $f: \mathbb{R}^n \to \mathbb{R}$, τότε το x_* ελαχιστοποιεί την $f(x)$ αν και μόνο αν το μηδέν είναι στοιχείο του υποδιαφορικού $\partial f(x_*)$, δηλαδή:

$$x_* \epsilon argmin_{x \epsilon \mathbb{R}^n} f(x) \Leftrightarrow 0 \epsilon \partial f(x_*).$$

Εξ'ορισμού για την εκτιμήτρια *LASSO* ισχύει:

$$\widehat{\boldsymbol{\beta}}_L \epsilon argmin_{\beta \epsilon \mathbb{R}^p} L(\boldsymbol{\beta}).$$

Επομένως, συμπεραίνουμε από τη 1η συνθήκη βελτιστότητας ότι $0 \epsilon \partial L\left(\widehat{\boldsymbol{\beta}}_L\right)$.

Από την ***(4.30)*** έχουμε:

$$L(\boldsymbol{\beta}) = \boldsymbol{y}'\boldsymbol{y} - 2\boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{y} + \boldsymbol{\beta}'\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1.$$

Άρα:

$$\frac{\partial L(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\boldsymbol{\beta} + \lambda\|\boldsymbol{\beta}\|_1,$$

υπενθυμίζοντας ότι $\frac{\partial(\boldsymbol{x}'\boldsymbol{A}\boldsymbol{x})}{\partial \boldsymbol{x}} = 2\boldsymbol{A}\boldsymbol{x}$.

Επομένως, αφού $0 \epsilon \partial L\left(\widehat{\boldsymbol{\beta}}_L\right) \Leftrightarrow \exists\ \boldsymbol{z} \epsilon \partial \left\|\widehat{\boldsymbol{\beta}}_L\right\|_1$ τέτοιο ώστε:

$$-2\boldsymbol{X}'\boldsymbol{y} + 2\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}_L + \lambda\boldsymbol{z} = 0$$

δηλαδή

$$\boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}}_L = \boldsymbol{X}'\boldsymbol{y} - {}^{\lambda}/_{2}\,\boldsymbol{z}.$$

(4.31)

Να σημειώσουμε ότι οι συνιστώσες του $\boldsymbol{z}$=$(z_1, \dots, z_p)$ είναι της μορφής:

$$\boldsymbol{z} = sign(\hat{\beta}_L) = \begin{cases} z_j = 1, \alpha\nu\ (\hat{\beta}_L)_j > 0 \\ z_j = -1, \alpha\nu\ (\hat{\beta}_L)_j < 0 \\ z_j \epsilon [-1,1], \alpha\nu\ (\hat{\beta}_L)_j = 0. \end{cases}$$

4.4.4 Περίπτωση μίας ανεξάρτητης μεταβλητής

Αρχικά θα μελετήσουμε το απλό μοντέλο (μία μεταβλητή):

$$\mathbf{y} = \boldsymbol{X}_1\beta_1 + \boldsymbol{\varepsilon}.$$

Το πρόβλημα ελαχιστοποίησης ορίζεται ως:

$$min_{\beta_1 \epsilon \mathbb{R}}\{\|\boldsymbol{y} - \boldsymbol{X_1}\beta_1\|_2^2 + \lambda\|\beta_1\|_1\}$$

και έστω $(\hat{\beta}_L)_1$ η λύση του.

Τότε από την ***(4.31)*** θα έχουμε:

$$\boldsymbol{X}_1{}'\boldsymbol{X}_1(\hat{\beta}_L)_1 = \boldsymbol{X}_1{}'\boldsymbol{y} - {}^{\lambda}/_{2}\,sign((\hat{\beta}_L)_1)$$

δηλαδή

$$(\hat{\beta}_L)_1 = (\hat{\beta}_{LS})_1 - {}^{\lambda}/_{2}\,sign\big((\hat{\beta}_L)_1\big),$$

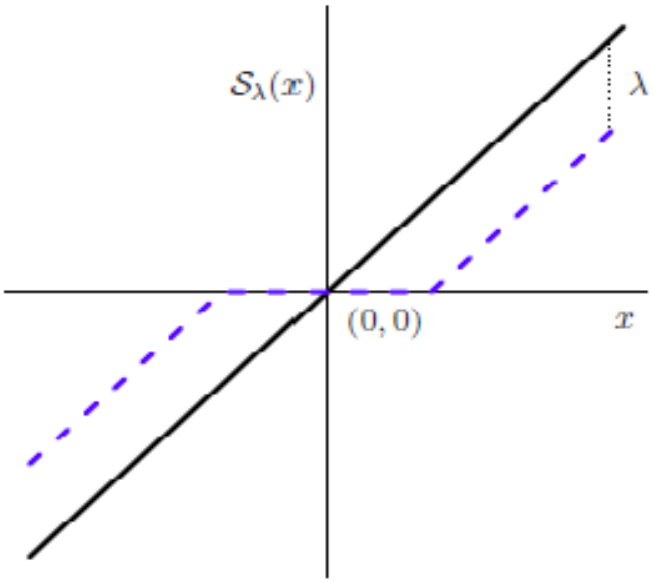
αφού $\boldsymbol{X}_1{}'\boldsymbol{X}_1 = 1$, λόγω κανονικοποίησης και $\boldsymbol{X}_1{}'\boldsymbol{y} = \big(\hat{\beta}_{LS}\big)_1$.

Άρα με τη βοήθεια λοιπόν της *soft-thresholding* συνάρτησης (***Σχήμα 4.2***) η οποία ορίζεται ως:

$$S_\lambda(x) = \begin{cases} x - \lambda, \alpha\nu\ x > \lambda \\ 0, \alpha\nu\ |x| \leq \lambda \\ x + \lambda, \alpha\nu\ x < -\lambda, \end{cases}$$

ο εκτιμητής $(\hat{\beta}_L)_1$ ισοδυναμεί με $(\hat{\beta}_L)_1 = S_{\lambda/2}((\hat{\beta}_{LS})_1)$, δηλαδή:

$$(\hat{\beta}_L)_1 = \begin{cases} (\hat{\beta}_{LS})_1 - \lambda/2\,, \alpha\nu\ (\hat{\beta}_{LS})_1 > \lambda/2 \\ 0, \alpha\nu\ |(\hat{\beta}_{LS})_1| \leq \lambda/2 \\ (\hat{\beta}_{LS})_1 + \lambda/2\,, \alpha\nu\ \ (\hat{\beta}_{LS})_1 < -\lambda/2\,. \end{cases}$$



Σχήμα 4.2: **Η *soft-thresholding* συνάρτηση απεικονίζεται με τις διακεκομμένες γραμμές.**

Ωστόσο έχει νόημα να δούμε πιο αναλυτικά τη λύση στην οποία φτάσαμε προκειμένου να καταλάβουμε το λόγο για τον οποίο η *LASSO* μηδενίζει συντελεστές.

Το πρόβλημα ελαχιστοποίησης, υπό μορφή πινάκων γράφεται ως εξής:

$$(\hat{\beta}_L)_1 = min_{\beta_1 \epsilon \mathbb{R}}\{\|\boldsymbol{y} - \boldsymbol{X}_1\beta_1\|_2^2 + \lambda\|\beta_1\|_1\}$$

$$= \min\{(\boldsymbol{y} - \boldsymbol{X}_1\beta_1)'(\boldsymbol{y} - \boldsymbol{X}_1\beta_1) + \lambda|\beta_1|\}$$

$$= \min\{\boldsymbol{y}'\boldsymbol{y} - 2\beta_1{}'\boldsymbol{X}_1{}'\boldsymbol{y} + \beta_1{}'\boldsymbol{X}_1{}'\boldsymbol{X}_1\beta_1 + \lambda|\beta_1|\}.$$

(4.32)

Έστω ότι $\boldsymbol{X}_1{}'\boldsymbol{y} > 0$ που συνεπάγεται $|\beta_1| = \beta_1$. Επομένως παραγωγίζοντας την ***(4.32)*** ως προς β_1 θα έχουμε:

$$-2\boldsymbol{X}_1{}'\boldsymbol{y} + 2\boldsymbol{X}_1{}'\boldsymbol{X}_1\beta_1 + \lambda,$$

η οποία έχει λύση:

$$\left(\hat{\beta}_L\right)_1 = \frac{\boldsymbol{X}_1{}'\boldsymbol{y} - {}^{\lambda}/_2}{\boldsymbol{X}_1{}'\boldsymbol{X}_1}.$$

Προφανώς, αυξάνοντας το ${}^{\lambda}/_2$ μπορούμε να οδηγήσουμε το $\left(\hat{\beta}_L\right)_1$ στο μηδέν (για ${}^{\lambda}/_2 = \boldsymbol{X}_1{}'\boldsymbol{y}$). Παρ'όλα αυτά, μόλις μηδενιστεί το $\left(\hat{\beta}_L\right)_1$, η επιπλέον αύξηση του ${}^{\lambda}/_2$ δε θα το κάνει αρνητικό αφού γράφοντας την παράγωγο της αντικειμενικής συνάρτησης ως προς β_1 θα έχουμε:

$$-2\boldsymbol{X}_1{}'\boldsymbol{y} + 2\boldsymbol{X}_1{}'\boldsymbol{X}_1\beta_1 - \lambda,$$

όπου η αλλαγή στο πρόσημο του λ οφείλεται στη φύση της απόλυτης τιμής της ποινής. Αυτό οδηγεί στη λύση:

$$\left(\hat{\beta}_L\right)_1 = \frac{\boldsymbol{X}_1{}'\boldsymbol{y} + {}^{\lambda}/_2}{\boldsymbol{X}_1{}'\boldsymbol{X}_1},$$

η οποία για $(\hat{\beta}_L)_1 < 0$ είναι ασυνεπής. Επομένως, γίνεται αντιληπτό ότι η $(\hat{\beta}_L)_1$ «κολλάει» στο μηδέν.

Η ίδια λογική ισχύει και όταν έχουμε αρνητική λύση ($\boldsymbol{X}_1{}'\boldsymbol{y} < 0$).

Από την άλλη, με την *Ridge* και την L_2 ποινή, για την παράγωγο της αντικειμενικής συνάρτησης θα έχουμε:

$$-2\boldsymbol{X}_1{}'\boldsymbol{y} + 2\boldsymbol{X}_1{}'\boldsymbol{X}_1\beta_1 + 2\lambda\beta_1,$$

η οποία έχει λύση:

$$(\hat{\beta}_R)_1 = \frac{\boldsymbol{X}_1{}'\boldsymbol{y}}{\boldsymbol{X}_1{}'\boldsymbol{X}_1 + \boldsymbol{\lambda}}.$$

Προφανώς το $(\hat{\beta}_R)_1$ δε θα γίνει ποτέ μηδέν όσο και να αυξηθεί το λ.

Ωστόσο, να σημειώσουμε ότι τα πράγματα μπορεί να αλλάξουν όταν έχουμε πολλαπλό μοντέλο, αλλά η γενική αρχή παραμένει ίδια.

4.4.5 Περίπτωση ορθοκανονικού πίνακα σχεδιασμού

Και στην ειδική περίπτωση που ο $\boldsymbol{X}$ είναι ορθογώνιος ($\boldsymbol{X}'\boldsymbol{X} = \boldsymbol{I}$), η εκτιμήτρια έχει σαφή λύση η οποία εκφράζεται με την μορφή των ελαχίστων τετραγώνων.

Λήμμα 4.2 (Tibshirani, 1996):

Έστω $\boldsymbol{X}$ ορθογώνιος πίνακας. Τότε ο $\boldsymbol{X}'\boldsymbol{X}$ αποτελεί έναν διαγώνιο πίνακα και η λύση του προβλήματος ελαχιστοποίησης:

$$\min\left\{\sum_{i=1}^{n}(y_i - \sum_{j=1}^{p} x_{ij}\beta_j)^2 + \lambda\sum_{j=1}^{p}|\beta_j|\right\},$$

μπορεί να γραφεί ως:

$$(\hat{\beta}_L)_j = sign\big((\hat{\beta}_{LS})_j\big)\Big(\big|(\hat{\beta}_{LS})_j\big| - {}^{\lambda}/_{2}\Big)^{+}, j = 1, \dots, p.$$

Ο συμβολισμός + στο $sign$ σημαίνει ότι $(\hat{\beta}_L)_j = 0$ για $\big|(\hat{\beta}_{LS})_j\big| \leq {}^{\lambda}/_{2}$.

Με τη βοήθεια και πάλι της *soft-thresholding* συνάρτησης ο εκτιμητής *LASSO* ισοδυναμεί με $(\hat{\beta}_L)_j = S_{{}^{\lambda}/_{2}}\big(\big(\hat{\beta}_{LS}\big)_j\big)$, δηλαδή:

$$(\hat{\beta}_L)_j = \begin{cases} (\hat{\beta}_{LS})_j - {}^{\lambda}/_{2}\,, \alpha\nu\ (\hat{\beta}_{LS})_j > {}^{\lambda}/_{2} \\ 0, \alpha\nu\ |(\hat{\beta}_{LS})_j| \leq {}^{\lambda}/_{2} \\ (\hat{\beta}_{LS})_j + {}^{\lambda}/_{2}\,, \alpha\nu\ \ (\hat{\beta}_{LS})_j < -{}^{\lambda}/_{2}\,. \end{cases}$$

Απόδειξη:

Η απόδειξη είναι άμεση από τον τύπο ***(4.31)*** αντικαθιστώντας το $\boldsymbol{X'X} = \boldsymbol{I}$ και $\boldsymbol{X'y} = \widehat{\boldsymbol{\beta}}_{LS}$. Αναλυτικά, το πρόβλημα ελαχιστοποίησης, υπό μορφή πινάκων, και κατ'επέκταση οι εκτιμήτριες στη μέθοδο *LASSO* υπολογίζονται ως εξής:

$$\widehat{\boldsymbol{\beta}}_L = \min\{\|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1\}$$

$$= \min\left\{(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

$$= \min\left\{\boldsymbol{y'y} - 2\boldsymbol{\beta'X'y} + \boldsymbol{\beta'X'X\beta} + \lambda\sum_{j=1}^{p}|\beta_j|\right\}$$

$$= \min\left\{\sum_{j=1}^{p} -2(\boldsymbol{X'y})_j\beta_j + {\beta_j}^2 + \lambda|\beta_j|\right\}.$$

Κάθε όρος του αθροίσματος αντιστοιχεί σε ξεχωριστή μεταβλητή β_j από τους υπόλοιπους. Άρα για να βρεθεί η τιμή της κάθε μεταβλητής, κάθε ένας όρος μπορεί να ελαχιστοποιηθεί ξεχωριστά. Επομένως:

$$L_j = -2(\boldsymbol{X'y})_j\beta_j + {\beta_j}^2 + \lambda|\beta_j|, j = 1, \dots, p.$$

- Αν για την εκτιμήτρια $(\boldsymbol{X'y})_j$, έχουμε $(\boldsymbol{X'y})_j > 0$, θα πρέπει και $\beta_j > 0$.

Άρα για $\boldsymbol{X'y} > 0$ η σχέση που θέλουμε να ελαχιστοποιήσουμε είναι:

$$L_j = -2(\boldsymbol{X}'\boldsymbol{y})_j\beta_j + {\beta_j}^2 + \lambda\beta_j.$$

Παραγωγίζοντάς τη και θέτοντάς τη ίση με μηδέν έχουμε:

$$-2(\boldsymbol{X}'\boldsymbol{y})_j + 2\beta_j + \lambda = 0.$$

Άρα:

$$(\hat{\beta}_L)_j = (\boldsymbol{X}'\boldsymbol{y})_j - {\lambda}/{2},$$

το οποίο είναι εφικτό όταν $(\hat{\beta}_L)_j > 0$ ή $(\boldsymbol{X}'\boldsymbol{y})_j - {\lambda}/{2} > 0$ ή $(\boldsymbol{X}'\boldsymbol{y})_j > {\lambda}/{2}$.

- Αντίστοιχα αν $(\boldsymbol{X}'\boldsymbol{y})_j < 0$ θα πρέπει $\beta_j < 0$. Άρα:

$$L_j = -2(\boldsymbol{X}'\boldsymbol{y})_j\beta_j + {\beta_j}^2 - \lambda\beta_j.$$

Και παραγωγίζοντάς τη και θέτοντάς τη ίση με μηδέν έχουμε:

$$-2(\boldsymbol{X}'\boldsymbol{y})_j + 2\beta_j - \lambda = 0$$

ή

$$(\hat{\beta}_L)_j = (\boldsymbol{X}'\boldsymbol{y})_j + {\lambda}/{2},$$

που είναι εφικτό όταν $(\hat{\beta}_L)_j < 0$, δηλαδή $(\boldsymbol{X}'\boldsymbol{y})_j + {\lambda}/{2} < 0$ ή $(\boldsymbol{X}'\boldsymbol{y})_j < -{\lambda}/{2}$.

Αντικαθιστώντας $\widehat{\boldsymbol{\beta}}_{LS} = \boldsymbol{X}'\boldsymbol{y}$ έχουμε τον τύπο για την ορθοκανονική περίπτωση.

Όπως είδαμε από τον παραπάνω τύπο η εκτιμήτρια *LASSO* ισούται με το μηδέν όταν $\left|(\hat{\beta}_{LS})_j\right| \leq {\lambda}/{2}$, δηλαδή για πολύ μικρές απόλυτες τιμές της εκτιμήτριας ελαχίστων τετραγώνων. Διαφορετικά η παράμετρος ποινής

προκαλεί συρρίκνωση των συντελεστών προς το μηδέν και το πρόσημο της εκτιμήτριας *LASSO* αντιστοιχεί στο πρόσημο της εκτιμήτριας ελαχίστων τετραγώνων.

4.4.6 Πολυμεταβλητή περίπτωση

Έστω ότι ο $\boldsymbol{X}$ είναι πλήρους μεγέθους, επομένως ο $\boldsymbol{X'X}$ είναι αντιστρέψιμος. Έστω ακόμα ότι με $\boldsymbol{X}_{-j}$ συμβολίζουμε όλες τις στήλες εκτός από την j-οστή στήλη και με β_{-j} όλους τους συντελεστές εκτός από τον β_j και έστω ότι η λύση της j-οστής συνιστώσας του β_j είναι $(\hat{\beta}_L)_j$. Τότε από την ***(4.33)*** έχουμε:

$$\boldsymbol{X'}_j\boldsymbol{X}_j(\hat{\beta}_L)_j = (\boldsymbol{X'}_j y - \boldsymbol{X'}_j\boldsymbol{X}_{-j}\beta_{-j} - {}^{\lambda}/_{2}\, sign((\hat{\beta}_L)_j)$$

ή

$$(\hat{\beta}_L)_j = (\boldsymbol{X'}_j y - \boldsymbol{X'}_j\boldsymbol{X}_{-j}\beta_{-j} - {}^{\lambda}/_{2}\, sign((\hat{\beta}_L)_j),$$

αφού $\boldsymbol{X'}_j\boldsymbol{X}_j = 1.$

Από την παραπάνω σχέση παρατηρούμε ότι η λύση του ενός β_j εξαρτάται από όλα τα άλλα στοιχεία $\beta_{i\neq j}$, γεγονός που σημαίνει ότι δεν υπάρχει ένας σαφής τύπος λύσης της εκτιμήτριας *LASSO* στην πολυμεταβλητή περίπτωση. Να σημειώσουμε ότι στην ορθοκανονική περίπτωση που είδαμε παραπάνω έχουμε σαφή λύση καθώς ο όρος $\boldsymbol{X'}_j\boldsymbol{X}_{-j}$ εξαφανίζεται λόγω της ορθογωνιότητας.

4.4.7 *Coordinate Descent*

Όπως είδαμε στη πολυμεταβλητή περίπτωση δεν υπάρχει μία λύση κλειστού τύπου, ωστόσο λόγω της κυρτότητας της αντικειμενικής συνάρτησης μπορούμε να υπολογίσουμε τη λύση ακόμα και στη πολυμεταβλητή αυτή περίπτωση με τη βοήθεια ενός αποδοτικού αλγόριθμου που χρησιμοποιεί το πακέτο *glmnet* της *R*, *Coordinate Descent*.

Η κεντρική ιδέα του αλγόριθμου *Coordinate Descent* είναι να ελαχιστοποιήσουμε διαδοχικά κάθε παράμετρο, κρατώντας τις άλλες σταθερές. Για κάθε υποπρόβλημα συνιστώσας (*coordinate*) φιξάρουμε

όλες τις συνιστώσες του $\boldsymbol{\beta}$ εκτός από την j-οστή συνιστώσα β_j. Επομένως, το πρόβλημα ελαχιστοποίησης που καλούμαστε να λύσουμε είναι:

$$min_{\beta_j \epsilon \mathbb{R}} \left\{ \left\| \boldsymbol{y} - \boldsymbol{X}_{-j}\beta_{-j} - \beta_j \boldsymbol{X}_j \right\|_2^2 + \lambda \sum_{i \neq j}^{p} |\beta_i| + \lambda |\beta_j| \right\},$$

όπου $r_j = \boldsymbol{y} - \boldsymbol{X}_{-j}\beta_{-j}$ είναι η διαφορά των πραγματικών τιμών και του μέρους του προσαρμοσμένου μοντέλου εκτός της $\boldsymbol{X}_j$ μεταβλητής.

Επομένως, το πρόβλημα ισοδυναμεί με το πρόβλημα μίας μεταβλητής:

$$min_{\beta_j \epsilon \mathbb{R}} \left\{ \left\| r_j - \beta_j \boldsymbol{X}_j \right\|_2^2 + \lambda \sum_{i \neq j}^{p} |\beta_i| + \lambda |\beta_j| \right\}.$$

Άρα, από την ***(4.31)*** έχουμε:

$$\left(\hat{\beta}_L\right)_j = r_j' \boldsymbol{X}_j - {}^{\lambda}/_{2}\, sign\left(\left(\hat{\beta}_L\right)_j\right),$$

η οποία έχει ως αποτέλεσμα:

$$\left(\hat{\beta}_L\right)_j \leftarrow S_{{}^{\lambda}/_{2}}\left(r_j' \boldsymbol{X}_j\right).$$

(4.33)

Ο συνολικός αλγόριθμος λειτουργεί εφαρμόζοντας την ***(4.33)*** επαναλαμβανόμενα με κυκλικό τρόπο, ανανεώνοντας τις συνιστώσες της $\hat{\beta}_L$.

4.4.8 Μέθοδοι επιλογής της παραμέτρου ποινής

Όπως και στην παλινδρόμηση *Ridge*, έτσι και στη *LASSO* η τιμή που θα δώσει ο αναλυτής στην παράμετρο ποινής διαδραματίζει τον σημαντικότερο ρόλο για το ποσοστό συρρίκνωσης των συντελεστών ώστε να καταλήξουμε στην τελική επιλογή του μοντέλου.

4.4.8.1 Κριτήρια πληροφορίας AIC και BIC

Στην περίπτωση που ο αναλυτής επιλέγει να χρησιμοποιήσει τα κριτήρια πληροφορίας, οφείλει να υπολογίσει τις τιμές τους που δίνονται από τις σχέσεις:

$$AIC = n\ln(\frac{SSE}{n}) + 2\,df$$

$$BIC = n\ln\left(\frac{SSE}{n}\right) + \ln(n)\,df,$$

όπου εδώ οι βαθμοί ελευθερίας αναφέρονται στον αριθμό των μη μηδενικών συντελεστών του μοντέλου.

Αφού υπολογιστούν τα κριτήρια σύγκρισης, ο αναλυτής τελικά επιλέγει αυτή την τιμή της παραμέτρου που δίνει τη μικρότερη τιμή του εκάστοτε κριτηρίου πληροφορίας.

Ωστόσο, να σημειώσουμε ότι οι βαθμοί ελευθερίας έχουν μελετηθεί για γραμμικές διαδικασίες (για παράδειγμα οι βαθμοί ελευθερίας στην πολλαπλή γραμμική παλινδρόμηση είναι ίσοι με τον αριθμό των επεξηγηματικών μεταβλητών). Ακόμα μία γενίκευση έχει γίνει και για τις περιπτώσεις όπου η εκτιμώμενη τιμή μπορεί να γραφεί $\widehat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$ (*linear smoothers*), όπως είδαμε και στην περίπτωση της *Ridge* όπου $\widehat{\boldsymbol{y}} = \boldsymbol{H}_\lambda \boldsymbol{y}$, με $\boldsymbol{H}_\lambda = (\boldsymbol{X}'\boldsymbol{X} + \lambda \boldsymbol{I})^{-\mathbf{1}}$,όπου εδώ έχουμε ότι $df(\boldsymbol{S}) = tr(\boldsymbol{S})$. Δυστυχώς, για τη *LASSO* δεν είναι διαθέσιμη μία σαφής έκφραση εξαιτίας της μη γραμμικής φύσης της. Γι'αυτό το λόγο είναι προτιμότερο να καταφύγουμε σε άλλα κριτήρια που δεν απαιτούν τον υπολογισμό των βαθμών ελευθεριών, όπως το μέσο τετραγωνικό σφάλμα. Αν γνωρίζαμε το MSE συναρτήσει του λ, τότε θα διαλέγαμε απλώς το λ που ελαχιστοποιεί το MSE. Για να το κάνουμε αυτό πρέπει να εκτιμήσουμε το MSE εισάγοντας μία δημοφιλή μέθοδο αυτή του *Cross Validation* την οποία θα αναπτύξουμε παρακάτω.

4.4.8.2 *Cross Validation*

Όπως έχουμε δει μέχρι στιγμής η προσαρμογή του εκάστοτε μοντέλου στα διαθέσιμα δεδομένα αποτελεί τη βάση για τις μεθόδους επιλογής μεταβλητών που έχουμε αναπτύξει. Παρ' όλα αυτά, δε θα ήταν σωστό να ισχυριστούμε ότι ένα μοντέλο που προσαρμόζεται κατάλληλα στα δεδομένα μας, θα είναι εξίσου καλό αν εφαρμοστεί σε διαφορετικά

δεδομένα. Για να ελέγξουμε αυτή την επιπλέον δυνατότητα του μοντέλου, δηλαδή το να μπορεί να βασιστεί ο αναλυτής στο ίδιο μοντέλο και για μελλοντικές προβλέψεις λόγω της καλής προβλεπτικής του ικανότητας, θα πρέπει να συλλέξουμε καινούρια δεδομένα και να ξανακάνουμε την προσαρμογή μοντέλου. Επειδή όμως η συλλογή δεδομένων είναι πολυδάπανη και χρονοβόρα, προτείνεται μια λογική εναλλακτική μεθοδολογία, η οποία είναι γνωστή με το όνομα *Cross Validation.*

Σύμφωνα με τη μεθοδολογία αυτή χωρίζονται τυχαία τα διαθέσιμα δεδομένα σε δύο μέρη, αυτό της εκτίμησης και αυτό της πρόβλεψης. Τα δεδομένα που περιλαμβάνονται στο στάδιο της εκτίμησης, γνωστό και ως *training set*, χρησιμοποιούνται για εκτιμηθούν οι άγνωστοι παράμετροι του μοντέλου και τα δεδομένα της πρόβλεψης, γνωστό και ως *validation* ή *test set*, χρησιμοποιούνται στη συνέχεια για να ελεγχθεί η προβλεπτική ικανότητα του μοντέλου που προσαρμόσαμε με τα δεδομένα της εκτίμησης. Η διαδικασία του *Cross Validation* επαναλαμβάνεται για αρκετές το πλήθος διαμερίσεις και τελικά παίρνουμε τους μέσους όρους των αποτελεσμάτων.

Η πιο συνηθισμένη περίπτωση του *Cross Validation* είναι η *K-fold Cross Validation* (***Σχήμα 4.3***), η οποία χρησιμοποιείται και για την εύρεση της κατάλληλης παραμέτρου ποινής. Σύμφωνα με τη μεθολογία αυτή το αρχικό μας δείγμα δεδομένων χωρίζεται σε *K* υποσύνολα-φακέλους (*folds*), τα οποία αποτελούνται από ίδιο αριθμό παρατηρήσεων. Σε κάθε στάδιο αυτής της διαδικασίας ο ένας φάκελος παίζει πάντα τον ρόλο του validation set, ενώ οι υπόλοιποι *K-1* φάκελοι αποτελούν το *training set*. Η *K-fold Cross Validation* εφαρμόζεται *K* φορές, μέχρι όλοι οι φάκελοι να χρησιμοποιηθούν ακριβώς μία φορά ο καθένας ως *validation set*. Στην πράξη, ο αριθμός των φακέλων που χρησιμοποιείται συνήθως είναι 5 ή 10.

Στη συνέχεια διαλέγουμε ένα διάστημα τιμών για την παράμετρο ποινής π.χ $\lambda = [\lambda_t]$ και υπολογίζουμε τους συντελεστές παλινδρόμησης για κάθε τιμή της παραμέτρου ποινής. Στη συνέχεια, υπολογίζουμε το άθροισμα τετραγώνων των σφαλμάτων:

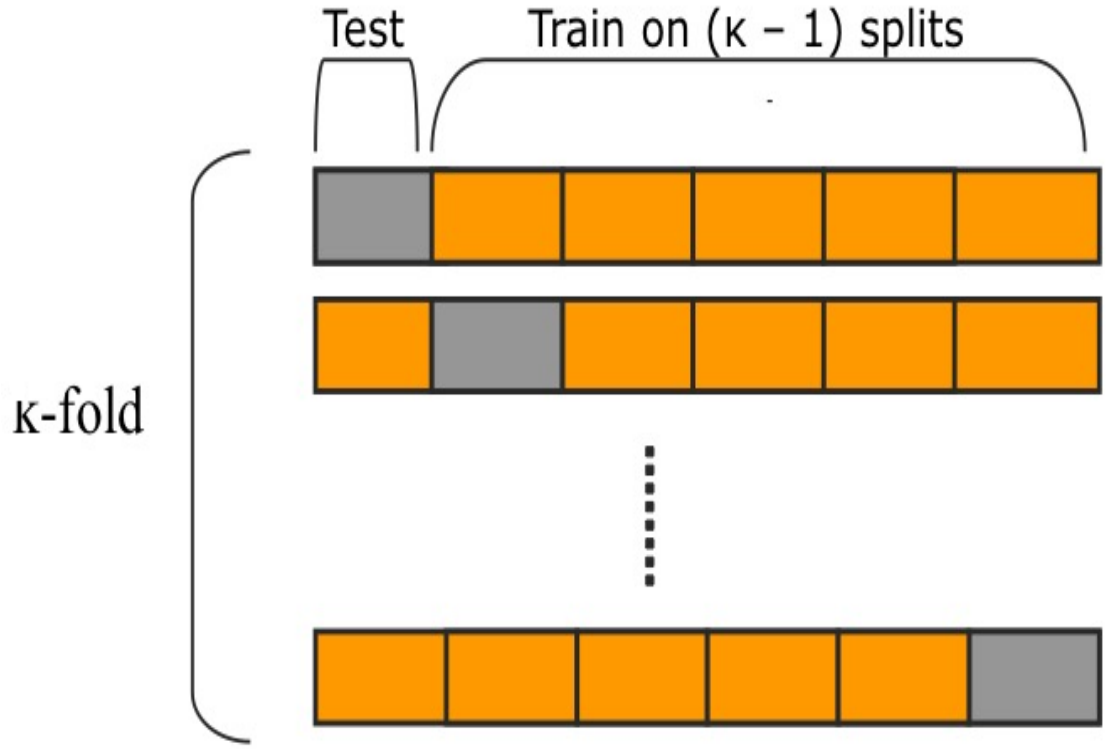
$$SSE_{\lambda_t}^{k} = \sum_{i=1}^{n}\left(y_{i-}\sum_{j=1}^{p}\hat{\beta}_j(k,\lambda_t)x_{ij}\right)^2,$$

όπου k είναι ο δείκτης του φακέλου που επιλέχθηκε ως το *validation set*.

Ωστόσο, κάποιος μπορεί να πάρει το μέσο όρο από αυτές τις τιμές του SSE για όλους τους φακέλους:

$$MSE_{\lambda_t} = \frac{1}{K}\sum_{k=1}^{K} SSE_{\lambda_t}^{k}.$$

Τελικά η βέλτιστη τιμή της παραμέτρου ποινής είναι αυτή που δίνει το μικρότερο MSE_{λ_t} και επιλέγεται εκείνο το μοντέλο που αντιστοιχεί στη συγκεκριμένη τιμή.



Σχήμα 4.3:* Διαδικασία του *K-fold Cross Validation

Τέλος, ενώ από την παραπάνω διαδικασία μπορούμε να αντιληφθούμε ότι το πλεονέκτημα αυτής της μεθοδολογίας είναι ότι όλες οι παρατηρήσεις χρησιμοποιούνται τόσο ως *validation* όσο και ως *training set*, η κατανομή των δεδομένων σε φακέλους, ωστόσο, δεν παύει να είναι τυχαία. Επομένως, θα πρέπει να γνωρίζουμε ότι κάθε MSE που προκύπτει από το *Cross Validation* έχει διασπορά (αφού είναι τυχαίο όπως και τα δεδομένα). Αυτό συνεπάγεται ότι από το μέσο όρο των MSE των *K*-φακέλων μπορούμε να λάβουμε την τυπική απόκλιση για κάθε

μοντέλο. Όμως επειδή και η ελάχιστη τιμή του MSE έχει διασπορά, αναπτύχθηκε ο κανόνας της μίας τυπικής απόκλισης «*1-standard error rule*» σύμφωνα με τον οποίο επιλέγεται το μοντέλο που έχει MSE εντός ενός τυπικού σφάλματος από το ελάχιστο MSE. Επομένως, σύμφωνα με τον κανόνα αυτόν για κάθε MSE_{λ_t} υπολογίζεται το τυπικό σφάλμα του μέσου όρου και στη συνέχεια επιλέγεται η μεγαλύτερη τιμή της παραμέτρου ποινής για την οποία το MSE_{λ_t} είναι εντός του ενός τυπικού σφάλματος. Κατά αυτόν τον τρόπο λοιπόν δημιουργείται ένα περισσότερο φειδωλό μοντέλο.

Τη μεθοδολογία αυτή θα εφαρμόσουμε τόσο για τη *LASSO* όσο και για την παλινδρόμηση *Ridge* με τη βοήθεια του πακέτου *glmnet*, όπως θα δούμε παρακάτω.

4.4.9 Σύγκριση *Ridge* με *LASSO*

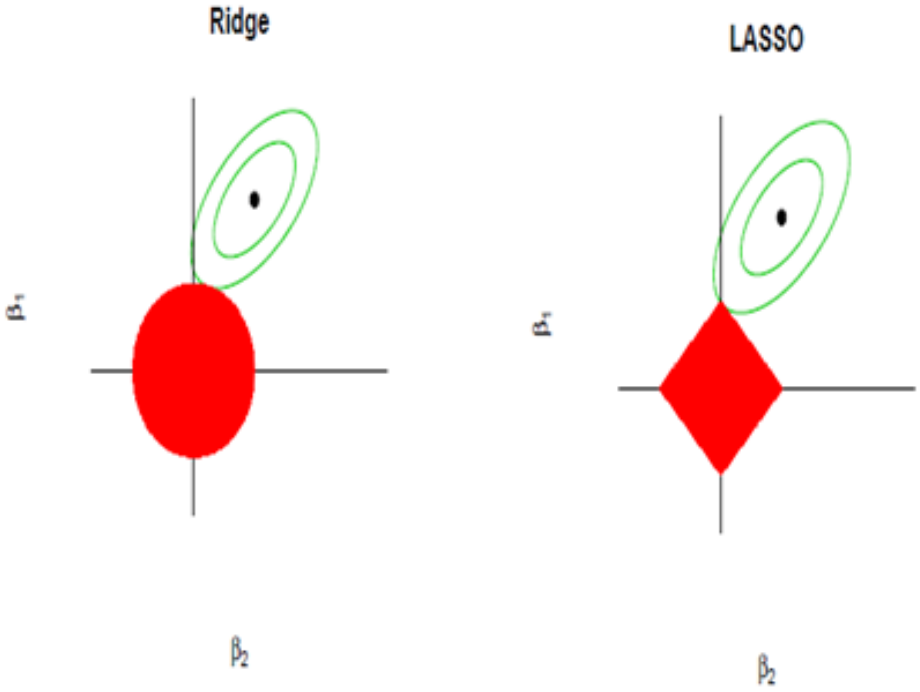
Στο ***Σχήμα 4.4*** παρουσιάζεται γραφικά η συσχέτιση των εκτιμητριών ελαχίστων τετραγώνων με τις εκτιμήτριες *Ridge* και *LASSO* αντίστοιχα στο πολλαπλό γραμμικό μοντέλο στις δύο διαστάσεις (δύο μεταβλητές).

Οι ελλείψεις οι οποίες σχηματίζονται γύρω από το κεντρικό σημείο $\widehat{\boldsymbol{\beta}}$, το οποίο αποτελεί την εκτιμήτρια ελαχίστων τετραγώνων, απεικονίζουν το άθροισμα των τετραγώνων των σφαλμάτων που καλούμαστε να ελαχιστοποιήσουμε ως προς τους συντελεστές παλινδρόμησης. Όπως είναι αναμενόμενο η τιμή του αθροίσματος των τετραγώνων των σφαλμάτων είναι ελάχιστη όσο πιο κοντά βρίσκεται η έλλειψη στην εκτιμήτρια ελαχίστων τετραγώνων ενώ αυξάνεται όσο οι ελλείψεις αυτές απομακρύνονται από αυτή.

Ο άλλος περιορισμός, δηλαδή αυτός που θέτουμε στο άθροισμα των τετραγώνων για τη *Ridge* ή των απολύτων για τη *LASSO* των συντελεστών του μοντέλου, απεικονίζεται από την περιοχή που έχει ως κέντρο της την αρχή των αξόνων. Στη συγκεκριμένη περίπτωση όπου βρισκόμαστε στον R^2, η περιοχή αυτή για την *Ridge* αντιστοιχεί στον περιορισμό ${\beta_1}^2 + {\beta_2}^2 \leq t$, γι΄αυτό και περιστάνεται από την εξίσωση κύκλου και με τον περιορισμό $|\beta_1| + |\beta_2| \leq t$, για τη *LASSO* όπου περιστάνεται με ρόμβο.

Με την παλινδρόμηση *Ridge* και τη *LASSO* αυτό που ουσιαστικά θέλουμε να πετύχουμε είναι η ταυτόχρονη ικανοποίηση τόσο της ελαχιστοποίησης του αθροίσματος των τετραγώνων των σφαλμάτων όσο

και του απαιτούμενου περιορισμού για την συρρίκνωση συντελεστών. Αυτό το σημείο δίνεται από εκεί που τέμνονται οι δύο προηγούμενες απαιτήσεις, το οποίο δίνει την εκτιμήτρια *Ridge* και *LASSO* αντίστοιχα. Εύκολα μπορούμε να διαπιστώσουμε ότι όσο αυξάνεται η διάσταση του χώρου ο περιορισμός του αθροίσματος των τετραγώνων των συντελεστών από κύκλο ή ρόμβο μετατρέπεται σε πολυδιάστατη σφαίρα ή πολυδιάστατο διαμάντι.

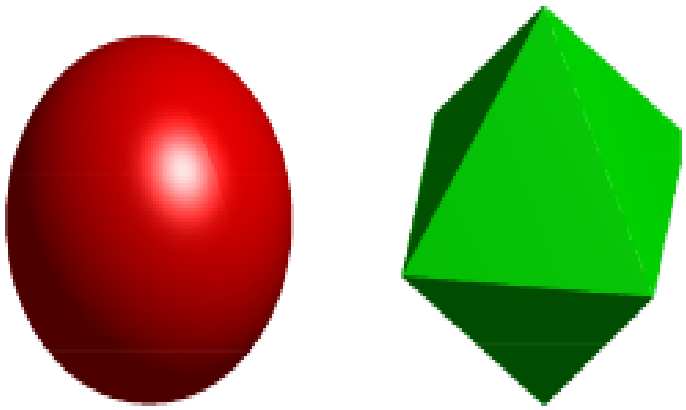


***Σχήμα 4.4*: Γεωμετρική ερμηνεία της *Ridge* και *LASSO* στις 2 διαστάσεις**

Τέλος, αξίζει να σημειώσουμε ότι λόγω της κυκλικής μορφής του περιορισμού των συντελεστών, η τομή των δύο περιορισμών που θέλουμε να πετύχουμε, δεν μπορεί να γίνεται πάνω στους άξονες και ως εκ τούτου με την παλινδρόμηση *Ridge* δεν επιτυγχάνεται κάποιος μηδενισμός των συντελεστών παλινδρόμησης. Αντιθέτως, όπως παρατηρούμε από το σχήμα η *LASSO* εμφανίζει «γωνίες», με αποτέλεσμα οι ελλείψεις που δημιουργούνται από τον περιορισμό του αθροίσματος των τετραγώνων των σφαλμάτων να τείνουν να τέμνουν αυτά τα σημεία πάνω σε ένα άξονα αρκετά συχνά. Για παράδειγμα, στο ***Σχήμα 4.4*** βλέπουμε ότι η ικανοποίηση των δύο περιορισμών επιτυγχάνεται στο σημείο $\beta_2 = 0$ και ως εκ τούτου από το τελικό μοντέλο απαλοίφεται η

μεταβλητή που έχει συντελεστή το β_2 και συμπεριλαμβάνεται μόνο η μεταβλητή με τον συντελεστή β_1.

Επεκτείνοντας το πρόβλημα σε περισσότερες διαστάσεις (***Σχήμα 4.5***), μπορούμε να καταλάβουμε ότι όλο και περισσότερες γωνίες σχηματίζονται λόγω της εμφάνισης του πολυδιάστατου διαμαντού, το οποίο με τη σειρά του οδηγεί στον ταυτόχρονο μηδενισμό όλο και περισσότερων συντελεστών του μοντέλου παλινδρόμησης. Σε αυτή την γεωμετρική ιδιότητα στηρίζεται και η επιπλέον δυνατότητα της *LASSO* να αποτελεί μέθοδο επιλογής μεταβλητών.



***Σχήμα 4.5*: Γεωμετρία *Ridge* (αριστερά) και *LASSO* (δεξιά) σε περισσότερες διαστάσεις**

4.4.10 Συμπεράσματα

Η *LASSO* προτιμάται έναντι όλων των άλλων μεθόδων-τεχνικών που έχουμε αναπτύξει μέχρι τώρα στην παρούσα διπλωματική καθώς συνδυάζει δύο σημαντικά πλεονέκτηματα όσον αφορά την προβλεπτική ακρίβεια και ερμηνεία των μοντέλων. Αρχικά, με τη βοήθεια της *LASSO* δημιουργούνται μοντέλα με πολύ μεγάλη ακρίβεια πρόβλεψης, αφού η συρρίκνωση και η απαλοιφή συντελεστών που επιτυγχάνεται μπορεί να μειώσει τη διασπορά εκτιμήσεων ενώ ταυτόχρονα δεν αυξάνει σημαντικά την μεροληψία.

Ακόμα, οδηγεί σε περισσότερο ερμηνεύσιμα μοντέλα καθώς καταλήγει σε ένα μικρότερο υποσύνολο επεξηγηματικών μεταβλητών, αφαιρώντας τις μεταβλητές αυτές που δεν εμφανίζουν υψηλή συσχετίση με τη

μεταβλητή απόκρισης, λαμβάνοντας παράλληλα υπόψιν της το πρόβλημα της πολυσυγγραμμικότητας.

4.5 *Elastic net*

Ωστόσο, μία από τις ελλείψεις της *LASSO* είναι ότι μέσω αυτής επιλέγεται μόνο μία μεταβλητή να εισαχθεί στο μοντέλο από το σύνολο των υψηλά συσχετισμένων επεξηγηματικών μεταβλητών. Ακόμα όταν ο αριθμός των παρατηρήσεων είναι πολύ μεγαλύτερος από αυτόν των παραμέτρων η *LASSO* δεν φέρει ικανοποιητικά αποτελέσματα ως μέθοδος επιλογής μεταβλητών. Μία καινούρια μέθοδος που προτάθηκε από τους *Zou* και *Hastie* (2005) αντιμετώπιζει αυτές τις ελλείψεις της *LASSO*. Πρόκειται για μία συνδυαστική μέθοδο της παλινδρόμησης *Ridge* και *LASSO*, με την *Elastic net*, καθώς συνδυάζει τις ποινές L_2 και L_1 αντίστοιχα.

Επομένως, οι τεχνικές *Ridge*, *LASSO* και *Elastic net* εντάσσονται στην ίδια οικογένεια με τον όρο ποινής:

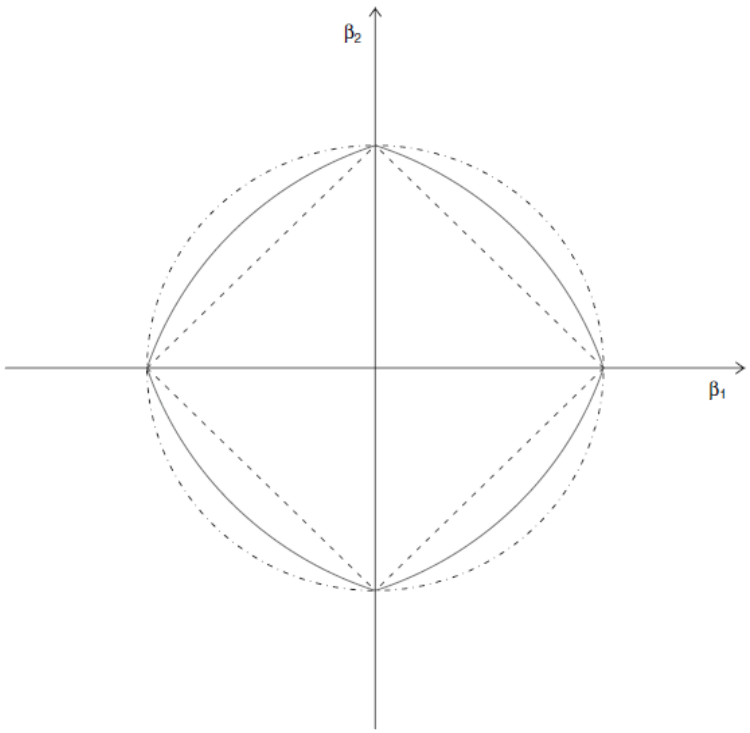
$$\sum_{j=1}^{p}(1-a)\beta_j^2 + \alpha\left|\beta_j\right|,$$

- Αν $a = 0$ τότε έχουμε την *Ridge* (κύκλος με διακεκομμένες γραμμές)
- Αν $a = 1$ τότε έχουμε τη *LASSO* (ρόμβος)
- Αν $0 < a < 1$ τότε έχουμε την *Elastic net* (κύκλος),

όπως βλέπουμε και στο ***Σχήμα 4.6***.

Παρ' όλα αυτά, την μέθοδο *Elastic net* δε θα την αναπτύξουμε στην παρούσα εργασία, απλώς την αναφέρουμε για να εισάγουμε τη χρήση ενός διαδεδομένου πακέτου στην *R*, του *glmnet*. Το πακέτο αυτό είναι πιο φιλικό στη χρήση του καθώς προτείνει άμεσα τις βέλτιστες τιμές της παραμέτρου ποινής. Θα προσδιορίσουμε το βέλτιστο μοντέλο τόσο για την *Ridge* όσο και για τη *LASSO* ώστε να έχουν μία υψηλή τιμή παραμέτρου ποινής, χρησιμοποιώντας τη μεθοδολογία του *Cross Validation*.

Ακόμα, στο πακέτο *R* του *glmnet* που θα χρησιμοποιήσουμε η τυποποίηση των μεταβλητών γίνεται αυτόματα κατά την προσαρμογή του μοντέλου. Τέλος, θα γίνει η σύγκριση των αποτελεσμάτων ώστε να προσδιορίσουμε ποια τεχνική λειτούργησε καλύτερα.



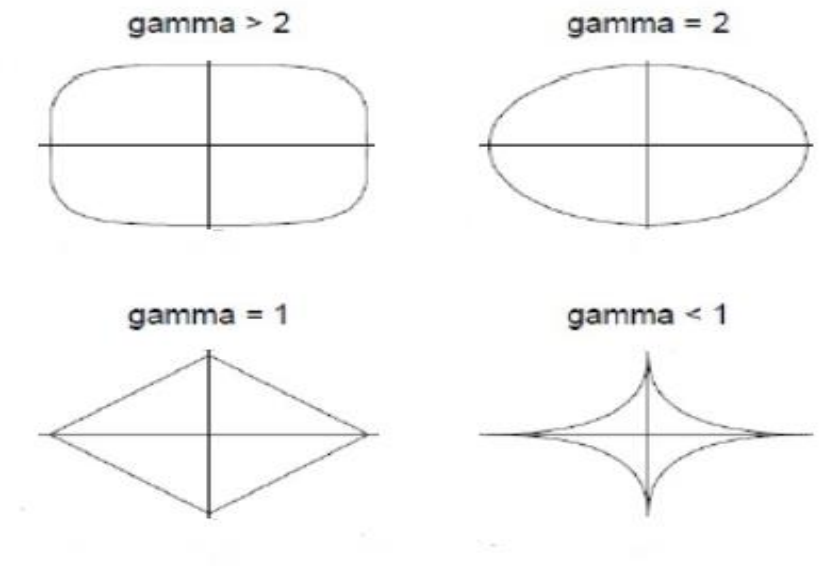
Σχήμα 4.6*: *Ridge, LASSO, Elastic net

4.6 *Bridge*

Τέλος, αξίζει να σημειωθεί ότι οι μέθοδοι *Ridge* και *LASSO* εντάσσονται στην οικογένεια της γενικότερης παλινδρόμησης *Bridge* η οποία παρουσιάστηκε από τους *Frank* και *Friedman* (1993) και έχει συνάρτηση ποινής $\sum|\beta_j|^\gamma \;\; 1 \leq \gamma \leq 2$. Και η μέθοδος *Bridge* υποδεικνύει την ελαχιστοποίηση του αθροίσματος των τετραγώνων των σφαλμάτων υπό τον περιορισμό:

$$\sum_{j=1}^{p}|\beta_j|^\gamma \leq t.$$

Στο ***Σχήμα 4.7*** φαίνεται η μορφή του περιορισμού της παλινδρόμησης *Bridge* για τις διάφορες τιμές της παραμέτρου γ, παρατηρώντας ότι για $\gamma = 1$ η μέθοδος αντιστοιχεί στη *LASSO*, ενώ για $\gamma = 2$ στη *Ridge*. Ωστόσο για $\gamma < 1$, το πρόβλημα δεν είναι κυρτό και αυτό κάνει το πρόβλημα ελαχιστοποίησης υπολογιστικά δύσκολο. Η $\gamma = 1$ είναι η μικρότερη τιμή όπου έχουμε κυρτό πρόβλημα.



Σχήμα 4.7:Παλινδρόμηση *Bridge*

ΚΕΦΑΛΑΙΟ 5

Εφαρμογή μεθόδων με χρήση της γλώσσας στατιστικού προγραμματισμού R

5.1 Περιγραφή των δεδομένων

Τα δεδομένα μας αποτελούν πραγματικά εβδομαδιαία στοιχεία και αφορούν τον όγκο πωλήσεων ενός προϊόντος (μεταβλητή απόκρισης) που απευθύνεται σε καταναλωτικό κοινό των Η.Π.Α και μεταβλητές που υποθέτουμε ότι ερμηνεύουν τις πωλήσεις αυτές. Το πλήθος των διαθέσιμων παρατηρήσεων είναι $n = 288$ εβδομάδες και $p = 141$ το πλήθος των επεξηγηματκών μεταβλητών που θα χρησιμοποιηθούν. Να σημειώσουμε ότι από αυτές τις 141 επεξηγηματικές μεταβλητές, οι 16 έχουν μηδενικές τιμές, τις οποίες τις εξαιρούμε από το μοντέλο, όπως θα δούμε παρακάτω. Η περιγραφή των τελικών 125 μεταβλητών φαίνεται στον ***Πίνακα 5.1***. Γενικά, από 288 εβδομάδων πωλήσεων του προιόντος έχουμε τις τιμές του (ανά ml, ανά συσκευασία,προσφορές κλπ), τη διανομή του προϊόντος στα καταστήματα κ.α (*TR* μεταβλητές), μεταβλητές που αφορούν μακροοικονιμικές συνιστώσες (*MA* μεταβλητές), μεταβλητές που αφορούν τα ανταγωνιστικά προιόντα (*Cross* μεταβλητές), κέρδη από διαφημιστικές καμπάνιες (*TV* μεταβλητές)

και τέλος κάποιες *dummies* μεταβλητές που αντιπροσωπεύουν κυρίως μέρες γιορτών των Η.Π.Α.

Μεταβλητές	***Περιγραφή***
Y	Πωλήσεις προϊόντος σε ml
TR1	Συνολική τιμή ανά ml
TR2	Βασική τιμή ανά ml (χωρίς προσφορά)
TR3	Τιμή ανά ml (με προσφορά)
TR4	Τιμή ανά ml με βάση το 1ο είδος προσφοράς
TR5	Τιμή ανά ml με βάση το 2ο είδος προσφοράς
TR6	Τιμή ανά ml με βάση το 3ο είδος προσφοράς
TR7	Τιμή ανά ml με βάση το 4ο είδος προσφοράς
TR8	Συνολική τιμή ανά συσκευασία
TR9	Τιμή ανά συσκευασία (χωρίς προσφορά)
TR10	Τιμή ανά συσκευασία (με προσφορά)
TR11	Τιμή ανά συσκευασία με βάση το 1ο είδος προσφοράς
TR12	Τιμή ανά συσκευασία με βάση το 2ο είδος προσφοράς
TR13	Τιμή ανά συσκευασία με βάση το 3ο είδος προσφοράς
TR14	Τιμή ανά συσκευασία με βάση το 4ο είδος προσφοράς
TR15	Συνολικός αριθμός καταστημάτων διάθεσης του προιόντος
TR16	Συνολική διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος
TR17	Μέση εβδομαδιαία διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος
TR18	Μέσος όρος προιόντων που πωλούνται ανά κατάστημα
TR19	Εβομαδιαίος μέσος όρος προιόντων που πωλούνται ανά κατάστημα
TR20	Μέσο σταθμισμένο ποσοστό συνολικής μείωσης τιμής
TR21	Μέσο σταθμισμένο ποσοστό συνολικής μείωσης τιμής (χωρίς επιπλέον προσφορά)
TR23	Μέσο σταθμισμένο ποσοστό συνολικής μείωσης τιμής με βάση το 1ο είδος προσφοράς
TR24	Μέσο σταθμισμένο ποσοστό συνολικής μείωσης τιμής με βάση το 2ο είδος προσφοράς
TR25	Μέσο σταθμισμένο ποσοστό συνολικής μείωσης τιμής με βάση το 3ο είδος προσφοράς
TR26	Μέσο σταθμισμένο ποσοστό συνολικής μείωσης τιμής με βάση το 4ο είδος προσφοράς
TR27	Πωλήσεις σε δολάρια (χωρίς προσφορά)
TR28	Πωλήσεις σε δολάρια (με προσφορά)
TR29	Πωλήσεις σε δολάρια με βάση το 1ο είδος προσφοράς
TR30	Πωλήσεις σε δολάρια με βάση το 2ο είδος προσφοράς
TR31	Πωλήσεις σε δολάρια με βάση το 3ο είδος προσφοράς

TR32	Πωλήσεις σε δολάρια με βάση το 4ο είδος προσφοράς
TR33	Πωλήσεις σε συσκευασίες (χωρίς προσφορά)
TR34	Πωλήσεις σε συσκευασίες (με προσφορά)
TR35	Πωλήσεις σε συσκευασίες με βάση το 1ο είδος προσφοράς
TR36	Πωλήσεις σε συσκευασίες με βάση το 2ο είδος προσφοράς
TR37	Πωλήσεις σε συσκευασίες με βάση το 3ο είδος προσφοράς
TR38	Πωλήσεις σε συσκευασίες με βάση το 4ο είδος προσφοράς
TR39	Πωλήσεις σε ml (χωρίς προσφορά)
TR40	Πωλήσεις σε ml (με προσφορά)
TR41	Πωλήσεις σε ml με βάση το 1ο είδος προσφοράς
TR42	Πωλήσεις σε ml με βάση το 2ο είδος προσφοράς
TR43	Πωλήσεις σε ml με βάση το 3ο είδος προσφοράς
TR44	Πωλήσεις σε ml με βάση το 4ο είδος προσφοράς
TR45	Συνολικός αριθμός διανομής προιόντος (χωρίς προσφορά)
TR46	Συνολικός αριθμός διανομής προιόντος (με προσφορά)
TR47	Συνολικός αριθμός διανομής προιόντος με βάση το 1ο είδος προσφοράς
TR48	Συνολικός αριθμός διανομής προιόντος με βάση το 2ο είδος προσφοράς
TR49	Συνολικός αριθμός διανομής προιόντος με βάση το 3ο είδος προσφοράς
TR50	Συνολικός αριθμός διανομής προιόντος με βάση το 4ο είδος προσφοράς
TR51	Διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος (χωρίς προσφορά)
TR52	Διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος (με προσφορά)
TR53	Διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος και του 1ου είδους προσφοράς
TR54	Διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος και του 2ου είδους προσφοράς
TR55	Διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος και του 3ου είδους προσφοράς
TR56	Διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος και του 4ου είδους προσφοράς
TR57	Μέση εβδομαδιαία διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος (χωρίς προσφορά)
TR58	Μέση εβδομαδιαία διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος (με προσφορά)
TR59	Μέση εβδομαδιαία διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος και του 1ου είδους προσφοράς
TR60	Μέση εβδομαδιαία διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος και του 2ου είδους προσφοράς
TR61	Μέση εβδομαδιαία διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος και του 3ου είδους προσφοράς
TR62	Μέση εβδομαδιαία διανομή προιόντος σταθμισμένη με βάση τις πωλήσεις του κάθε καταστήματος και του 4ου είδους προσφοράς
Seasonality	Εποχικότητα

MA1	Εισόδημα: Προσωπικό αναλώσιμο
MA2	Εθνικό Εισόδημα και μετρήσεις προιόντων (ΕΕΜΠ): Προσωπικά έξοδα κατανάλωσης (για μη αναλώσιμα αγαθά)
MA3	ΕΕΜΠ: Μεταβολή ιδιωτικών αποθεμάτων
MA4	ΕΕΜΠ: Μεικτή αποταμίευση
MA5	Δείκτης τιμών μετοχών
MA6	Εισόδημα: Συνολικό προσωπικό
MA7	Εισόδημα: Ημερομίσθια και μισθοί ιδιωτικών επιχειρήσεων
MA8	ΕΕΜΠ: Προσωπικά έξοδα κατανάλωσης (για αναλώσιμα αγαθά)
MA9	Δείκτης εμπιστοσύνης καταναλωτών (ΔΕΚ)
MA10	ΕΕΜΠ: Ακαθάριστο εγχώριο προιόν
MA11	ΔΕΚ: Αστικός καταναλωτής-εμπορεύματα φαγητού και ποτού
MA12	Συνολικές ιδιωτικές μέσες ωριαίες απολαβές
MA13	ΔΕΚ: Αστικός καταναλωτής-φαγητό στο σπίτι
MA14	ΔΕΚ: Αστικός καταναλωτής-φαγητό και ποτά
MA15	ΕΕΜΠ: Προσωπικά έξοδα κατνάλωσης-συνολικά
MA16	Δείκτης οικονομικής πολυπλοκότητας(ΔΟΠ): Συνολική αποζημίωση ιδιωτικών επιχειρήσεων-όλοι οι εργάτες
MA17	ΔΟΠ: Ημερομίσθια και μισθοί ιδιωτικών επιχειρήσεων-όλοι οι εργάτες
MA18	Υπάρχουσες πωλήσεις νοικοκυριών
MA19	Τιμές βενζίνης
MA20	Έρευνα νοικοκυριών: Ποσοστά ανεργίας
Dummy1	Πρωτοχρονιά:1 αλλιώς:0
Dummy2	Γενέθλια του Μάρτιν Λούθερ:1 αλλιώς:0
Dummy3	Γενέθλια του Ουάσινγκτον:1 αλλιώς:0
Dummy4	Εθνική ημέρα μνήμης (Memorial Day):1 αλλιώς:0
Dummy5	Ημέρα Ανεξαρτησίας (Indepedence Day):1 αλλιώς:0
Dummy6	Ημέρα των εργατών (Labor Day):1 αλλιώς:0
Dummy7	Ημέρα του Κολόμβου (Columbus Day):1 αλλιώς:0
Dummy8	Ημέρα των βετεράνων (Veterans Day):1 αλλιώς:0
Dummy9	Ημέρα των Ευχαριστιών (Thanksgiving Day):1 αλλιώς:0
Dummy10	Χριστούγεννα:1 αλλιώς:0
Dummy11	Πάσχα:1 αλλιώς:0
Dummy12	Σούπερ Μπόουλ (Super Bowl):1 αλλιώς:0
Dummy13	Halloween:1 αλλιώς:0
Cross1	Προσωρινή μείωση τιμής (προσφορά) ανταγωνιστικού προϊόντος 1
Cross2	Τιμή ανά ml ανταγωνιστικού προϊόντος 1
Cross3	Μέση σταθμισμένη τιμή ανταγωνιστικού προϊόντος 1
Cross4	Τιμή ανά συσκευασία ανταγωνιστικού προϊόντος 1
Cross5	Η διαφορά της προσωρινής μείωσης τιμής του ανταγωνιστικού προϊόντος 2 με εκείνου του προϊόντος 1

Cross6	Ο λόγος της προσωρινής μείωσης τιμής του ανταγωνιστικού προϊόντος 2 με εκείνου του προϊόντος 1
Cross7	Η διαφορά της τιμής ανά ml του ανταγωνιστικού προϊόντος 2 με εκείνου του προϊόντος 1
Cross8	Ο λόγος τιμής ανά ml του ανταγωνιστικού προϊόντος 2 με εκείνου του προϊόντος 1
Cross9	Η διαφορά της μέσης σταθμισμένης τιμής του ανταγωνιστικού προϊόντος 2 με εκείνου του προϊόντος 1
Cross10	Ο λόγος της μέσης σταθμισμένης τιμής του ανταγωνιστικού προϊόντος 2 με εκείνου του προϊόντος 1
Cross11	Η διαφορά της τιμής ανά συσκευασία του ανταγωνιστικού προϊόντος 2 με εκείνου του προϊόντος 1
Cross12	Ο λόγος τιμής ανά συσκευασία του ανταγωνιστικού προϊόντος 2 με εκείνου του προϊόντος 1
Cross13	Συνολικός πληθυσμός
TR75	Πλήθος υποκατηγοριών που περιλαμβάνει το προϊόν για τις πωλήσεις του οποίου τρέξαμε την ανάλυση
TV1	Συνολική δαπάνη διαφήμισης (σε δολάρια) για το ανταγωνιστικό προϊόν 2
TV2	Συνολική δαπάνη για TV διαφήμιση (σε δολάρια) για το ανταγωνιστικό προϊόν 2
TV3	Συνολική δαπάνη για διαφήμιση όχι σε TV (σε δολάρια) για το ανταγωνιστικό προϊόν 2
TV4	Συνολική δαπάνη για διαφήμιση (σε δολάρια) για διάρκεια 4 εβδομάδων για το ανταγωνιστικό προϊόν 2
TV5	Συνολική δαπάνη για διαφήμιση (σε δολάρια) για διάρκεια 8 εβομάδων για το ανταγωνιστικό προϊόν 2
TV6	Συνολική δαπάνη για διαφήμιση TV (σε δολάρια) για διάρκεια 4 εβδομάδων (σε τοπικό επίπεδο) για το ανταγωνιστικό προϊόν 2
TV7	Συνολική δαπάνη για διαφήμιση TV(σε δολάρια) για διάρκεια 8 (σε τοπικό επίπεδο) για το ανταγωνιστικό προϊόν 2
TV8	Συνολική δαπάνη για διαφήμιση TV (σε δολάρια) για διάρκεια 4 εβδομάδων (σε παγκόσμιο επίπεδο) για το ανταγωνιστικό προϊόν 2
TV9	Συνολική δαπάνη για διαφήμιση TV (σε δολάρια) για διάρκεια 8 εβδομάδων (σε παγκόσμιο επίπεδο) για το ανταγωνιστικό προϊόν 2
TVC1	Συνολική δαπάνη διαφήμισης (σε δολάρια) για το ανταγωνιστικό προϊόν 1
TVC2	Συνολική δαπάνη για TV διαφήμιση (σε δολάρια) για το ανταγωνιστικό προϊόν 1
TVC3	Συνολική δαπάνη για διαφήμιση όχι σε TV (σε δολάρια) για το ανταγωνιστικό προϊόν 1
TVC4	Συνολική δαπάνη για διαφήμιση (σε δολάρια) για διάρκεια 4 εβδομάδων για το ανταγωνιστικό προϊόν 1
TVC5	Συνολική δαπάνη για διαφήμιση (σε δολάρια) για διάρκεια 8 εβδομάδων για το ανταγωνιστικό προϊόν 1
TVC8	Συνολική δαπάνη για διαφήμιση TV (σε δολάρια) για διάρκεια 4 εβδομάδων (σε παγκόσμιο επίπεδο) για το ανταγωνιστικό προϊόν 1
TVC9	Συνολική δαπάνη για διαφήμιση TV (σε δολάρια) για διάρκεια 8 εβδομάδων (σε παγκόσμιο επίπεδο) για το ανταγωνιστικό προϊόν 1

TR76	Το σύνολο των promotion (προωθητικών ανεργιών) για το ανταγωνιστικό προϊόν 2

***Πίνακας 5.1*: Περιγραφή μεταβλητών**

Αρχικά εγκαθιστούμε την βιβλιοθήκη *"xlsx"* στο περιβάλλον της *R* και στη συνέχεια φορτώνουμε τα δεδομένα με την εντολή *read.xlsx*. Στη συνέχεια, αφαιρούμε όλες τις μεταβλητές που έχουν μηδενικές τιμές σε όλες τις παρατηρήσεις, καθώς δεν συνεισφέρουν στο μοντέλο μας και τελικά παραμένουν 125 μεταβλητές.

```
install.packages("openxlsx")

library(openxlsx)

Data <- read.xlsx("C:/Users/hp/Downloads/Data.xlsx", sheet = 1
)

non_zeros <- sapply(Data,function(x) sum(x==0)!=length(x))

Data <- Data[,non_zeros]

dummies <- Data[,sapply(Data,function(x) length(unique(x))==2)]

Data[,sapply(Data,function(x)length(unique(x))==2)]<-
data.frame(sapply(dummies,function(x)
as.factor(as.character(x))))
```

Στη συνέχεια, προσαρμόζουμε την πολλαπλή γραμμική παλινδρόμηση χρησιμοποιώντας όλες τις επεξηγηματικές μας μεταβλητές και με την εντολή *summary* βλέπουμε τα αποτελέσματα της παλινδρόμησης.

```
fit <- lm(Y~.,data = Data[,-c(1,2)]

summary(fit)
```

```
Call:
lm(formula = Y ~ ., data = Data[, -c(1, 2)])

Residuals:
       Min         1Q     Median         3Q        Max
-2.677e-09 -1.610e-10 -9.770e-12  1.731e-10  2.285e-09

Coefficients: (23 not defined because of singularities)
             Estimate Std. Error    t value Pr(>|t|)
(Intercept) -5.145e-07  3.108e-07 -1.656e+00  0.09948 .
TR1          5.008e-09  1.434e-08  3.490e-01  0.72735
TR2         -2.770e-11  7.792e-10 -3.600e-02  0.97167
TR3          2.977e-09  2.544e-09  1.170e+00  0.24339
TR4          1.809e-10  4.858e-10  3.720e-01  0.71005
TR5          1.030e-10  1.947e-10  5.290e-01  0.59764
TR6         -4.580e-11  2.566e-10 -1.780e-01  0.85857
```

```
TR7         -1.645e-11  1.158e-10 -1.420e-01  0.88718
TR8         -1.561e-09  6.500e-09 -2.400e-01  0.81044
TR9         -2.635e-10  3.017e-10 -8.730e-01  0.38358
TR10         1.339e-09  9.084e-10  1.474e+00  0.14225
TR11         1.356e-10  1.812e-10  7.480e-01  0.45534
TR12         3.224e-11  1.509e-10  2.140e-01  0.83100
TR13        -1.229e-10  1.145e-10 -1.073e+00  0.28481
TR14         3.258e-11  7.979e-11  4.080e-01  0.68351
TR15        -2.857e-12  8.945e-11 -3.200e-02  0.97456
TR16        -1.215e-10  1.119e-10 -1.086e+00  0.27886
TR17                NA         NA         NA       NA
TR18        -2.198e-09  8.743e-09 -2.510e-01  0.80177
TR19                NA         NA         NA       NA
TR20         2.766e-10  8.533e-11  3.242e+00  0.00141 **
TR21        -9.804e-11  3.039e-11 -3.226e+00  0.00149 **
TR23         3.596e-11  1.778e-11  2.023e+00  0.04451 *
TR24         1.318e-11  7.546e-12  1.746e+00  0.08243 .
TR25        -3.161e-12  7.601e-12 -4.160e-01  0.67797
TR26        -1.833e-12  3.449e-12 -5.310e-01  0.59580
TR27         4.989e-15  6.373e-15  7.830e-01  0.43474
TR28         5.939e-15  8.976e-15  6.620e-01  0.50898
TR29         5.954e-15  6.786e-15  8.770e-01  0.38145
TR30         4.015e-15  7.233e-15  5.550e-01  0.57947
TR31         2.667e-15  7.933e-15  3.360e-01  0.73714
TR32                NA         NA         NA       NA
TR33        -1.426e-14  1.218e-14 -1.171e+00  0.24310
TR34        -3.664e-15  1.550e-14 -2.360e-01  0.81344
TR35        -2.289e-15  1.074e-14 -2.130e-01  0.83152
TR36         5.338e-15  1.626e-14  3.280e-01  0.74306
TR37        -1.499e-14  1.171e-14 -1.281e+00  0.20196
TR38                NA         NA         NA       NA
TR39         1.000e+00  1.275e-14  7.846e+13  < 2e-16 ***
TR40         1.000e+00  1.736e-14  5.760e+13  < 2e-16 ***
TR41        -1.178e-14  1.180e-14 -9.980e-01  0.31940
TR42        -1.350e-14  1.640e-14 -8.230e-01  0.41162
TR43        -4.702e-15  1.276e-14 -3.680e-01  0.71294
TR44                NA         NA         NA       NA
TR45         3.701e-11  6.480e-11  5.710e-01  0.56861
TR46                NA         NA         NA       NA
TR47        -4.170e-11  6.424e-11 -6.490e-01  0.51708
TR48        -3.638e-11  6.749e-11 -5.390e-01  0.59055
TR49         4.897e-12  6.382e-11  7.700e-02  0.93892
TR50                NA         NA         NA       NA
TR51        -2.759e-11  2.165e-11 -1.274e+00  0.20423
TR52        -1.728e-11  1.940e-11 -8.910e-01  0.37425
TR53         1.576e-11  2.152e-11  7.330e-01  0.46476
TR54        -1.574e-12  2.386e-11 -6.600e-02  0.94747
TR55        -6.130e-11  3.818e-11 -1.606e+00  0.11004
TR56        -6.158e-11  1.151e-10 -5.350e-01  0.59338
TR57                NA         NA         NA       NA
TR58                NA         NA         NA       NA
TR59                NA         NA         NA       NA
TR60                NA         NA         NA       NA
TR61                NA         NA         NA       NA
TR62                NA         NA         NA       NA
Seasonality -3.214e-16  6.818e-16 -4.710e-01  0.63792
MA1          1.466e-11  5.155e-12  2.843e+00  0.00497 **
MA2          1.018e-11  1.621e-11  6.280e-01  0.53089
MA3          6.489e-12  4.619e-12  1.405e+00  0.16175
MA4          6.708e-12  2.707e-12  2.478e+00  0.01410 *
MA5         -8.004e-12  7.438e-12 -1.076e+00  0.28334
MA6         -1.058e-11  7.345e-12 -1.440e+00  0.15148
```

```
MA7         -1.778e-11  5.787e-12 -3.072e+00  0.00245 **
MA8         -4.026e-11  2.918e-11 -1.380e+00  0.16931
MA9         -1.216e-11  4.724e-11 -2.570e-01  0.79711
MA10        -1.574e-11  7.750e-12 -2.031e+00  0.04373 *
MA11        -8.025e-10  3.236e-10 -2.480e+00  0.01403 *
MA12        -1.775e-08  8.215e-09 -2.161e+00  0.03199 *
MA13         7.991e-10  7.151e-10  1.117e+00  0.26525
MA14        -1.614e-09  1.390e-09 -1.161e+00  0.24717
MA15         1.735e-11  1.211e-11  1.432e+00  0.15373
MA16        -7.729e-09  3.374e-09 -2.291e+00  0.02310 *
MA17         1.009e-08  4.626e-09  2.180e+00  0.03052 *
MA18         8.207e-11  3.301e-10  2.490e-01  0.80392
MA19         3.560e-09  1.812e-09  1.964e+00  0.05097 .
MA20        -1.319e-09  1.635e-09 -8.060e-01  0.42101
Dummy11      1.343e-10  2.660e-10  5.050e-01  0.61434
Dummy21      4.560e-10  2.240e-10  2.036e+00  0.04318 *
Dummy31      5.505e-12  2.082e-10  2.600e-02  0.97893
Dummy41     -5.882e-12  1.955e-10 -3.000e-02  0.97603
Dummy51      7.269e-11  1.996e-10  3.640e-01  0.71616
Dummy61     -1.925e-10  2.147e-10 -8.960e-01  0.37117
Dummy71      1.303e-10  2.138e-10  6.090e-01  0.54296
Dummy81      3.942e-11  2.217e-10  1.780e-01  0.85907
Dummy91      6.761e-11  2.129e-10  3.180e-01  0.75117
Dummy101    -2.014e-10  2.724e-10 -7.390e-01  0.46061
Dummy111     4.776e-11  2.100e-10  2.270e-01  0.82034
Dummy121    -3.379e-11  2.096e-10 -1.610e-01  0.87208
Dummy131    -7.313e-11  2.146e-10 -3.410e-01  0.73370
Cross1       1.922e-10  6.380e-11  3.013e+00  0.00295 **
Cross2       3.300e-06  1.069e-05  3.090e-01  0.75799
Cross3       2.526e-11  2.667e-11  9.470e-01  0.34478
Cross4      -2.206e-06  7.129e-06 -3.090e-01  0.75738
Cross5              NA         NA         NA       NA
Cross6       1.166e-09  3.427e-10  3.402e+00  0.00082 ***
Cross7              NA         NA         NA       NA
Cross8      -1.562e-08  2.499e-08 -6.250e-01  0.53270
Cross9              NA         NA         NA       NA
Cross10      2.218e-11  1.894e-11  1.171e+00  0.24299
Cross11             NA         NA         NA       NA
Cross12      3.054e-09  1.754e-08  1.740e-01  0.86199
Cross13      3.319e-15  1.834e-15  1.810e+00  0.07195 .
TR75         1.418e-11  3.438e-11  4.130e-01  0.68044
TV1         -8.703e-13  6.835e-13 -1.273e+00  0.20452
TV2          9.268e-13  7.108e-13  1.304e+00  0.19390
TV3                 NA         NA         NA       NA
TV4         -1.137e-13  2.054e-12 -5.500e-02  0.95590
TV5         -1.903e-12  3.390e-12 -5.610e-01  0.57519
TV6         -1.140e-13  2.091e-12 -5.500e-02  0.95659
TV7          1.682e-12  3.451e-12  4.870e-01  0.62666
TV8                 NA         NA         NA       NA
TV9                 NA         NA         NA       NA
TVC1         1.149e-12  5.434e-12  2.110e-01  0.83281
TVC3                NA         NA         NA       NA
TVC4         1.874e-12  1.616e-11  1.160e-01  0.90782
TVC5         8.203e-12  3.003e-11  2.730e-01  0.78501
TVC8                NA         NA         NA       NA
TVC9                NA         NA         NA       NA
TR76         1.070e-13  9.805e-14  1.091e+00  0.27656
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 4.033e-10 on 185 degrees of freedom
Multiple R-squared:      1,    Adjusted R-squared:      1
```

```
F-statistic: 2.174e+29 on 102 and 185 DF,  p-value: < 2.2e-16
 Warning message:
 In summary.lm(fit) : essentially perfect fit: summary may be unreliable
```

Παρατηρούμε ότι βάζοντας στο μοντέλο όλες τις επεξηγηματικές μεταβλητές έχουμε κάποια προβλήματα, όπως φαίνεται και από το μήνυμα που δίνει η *R*.

i) Η παλινδρόμηση δίνει συντελεστή προσδιορισμού ίσο με 1, γεγονός που δείχνει ότι υπάρχει τέλεια προσαρμογή.

ii) Η παλινδρόμηση δεν κατάφερε να βρει συντελεστές σε κάποιες μεταβλητές (*NA*), λόγω πολυσυγραμμικότητας.

iii) Για τις μεταβλητές *TR39* και *TR40* έχει δώσει συντελεστές ίσους με 1, το οποίο είναι και λογικό εφόσον όπως βλέπουμε από τον ***Πίνακα 5.1*** οι μεταβλητές αυτές μετράνε ότι και η εξαρτημένη μεταβλητή.

Για όλους τους παραπάνω λόγους θα πρέπει να αφαιρεθούν οι «προβληματικές» μεταβλητές προκειμένου να προχωρήσουμε στην εφαρμογή των μεθόδων επιλογής μεταβλητών.

Αρχικά αφαιρούμε όλες τις μεταβλητές με τις απούσες τιμές, τις μεταβλητές *TR39* και *TR40*, καθώς και τις 13 *dummies* εφόσον έχουν ελάχιστες μονάδες (δεν συνεισφέρει κάπου η εισαγωγή τους).

```
data1<-Data[,-c(1,19,21,33,39,40,41,42,45,47,51,58:63,85:97,102,104,
106,108,114,119,120,122,125,126)]
```

Εκτελώντας πάλι την παλινδρόμηση, τώρα πια δεν υπάρχουν μεταβλητές με απούσες τιμές αλλά ο συντελεστής προσδιορισμού συνεχίζει να είναι πολύ υψηλός (0.9995).

Επομένως πρέπει να αφαιρεθούν και άλλες «προβληματικές» μεταβλητές. Γι' αυτό το λόγο αν ονομάσουμε p τον αριθμό των επεξηγηματικών μεταβλητών που εν τέλει βρίσκονται στο μοντέλο ξεκινάμε να κάνουμε παλινδρομήσεις με κάθε μία από αυτές τις p ως μεταβλητή απόκρισης και τις υπόλοιπες $p-1$ ως επεξηγηματικές. Αν κάποιο μοντέλο έχει R^2 ίσο με 1 ή πολύ υψηλό, άνω του 0.999 αφαιρούμε τις εν λόγω μεταβλητές απόκρισης.

```
data2<-data1[,-c(2,9,16,18,25:27,30,31,38,39,49,50,52:56,58,60:65,67
,68,70,72,74,76,77,82,84)]
```

Αφού αφαιρέσουμε αυτές τις μεταβλητές, έχοντας τελικά 53 μεταβλητές, εκτελούμε πάλι την παλινδρόμηση με τις πωλήσεις ως μεταβλητή απόκρισης και τις εναπομείναντες ως επεξηγηματικές και με την εντολή *summary(fit)* βλέπουμε τα αποτελέσματα της παλινδρόμησης.

```
fit <- lm(Y~.,data = data2)
summary(fit)
```

```
Call:
lm(formula = Y ~ ., data = data2)

Residuals:
    Min      1Q  Median      3Q     Max
-105449  -20408    -934   22969  111020

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  4.179e+05  4.943e+05   0.845  0.39872
TR2         -3.348e+04  4.902e+04  -0.683  0.49528
TR3         -2.758e+05  5.275e+04  -5.228 3.80e-07 ***
TR4          4.658e+04  3.247e+04   1.434  0.15285
TR5          1.767e+04  1.376e+04   1.283  0.20059
TR6          1.454e+04  1.940e+04   0.749  0.45432
TR7         -5.901e+03  8.902e+03  -0.663  0.50806
TR9         -1.704e+04  2.065e+04  -0.825  0.40995
TR10        -6.985e+04  2.181e+04  -3.203  0.00155 **
TR11        -2.141e+03  1.305e+04  -0.164  0.86986
TR12        -4.230e+03  1.033e+04  -0.409  0.68259
TR13        -2.911e+03  7.692e+03  -0.378  0.70544
TR14         4.906e+03  5.958e+03   0.824  0.41105
TR16        -2.506e+03  2.600e+03  -0.964  0.33619
TR20        -1.984e+04  4.429e+03  -4.479 1.17e-05 ***
TR21         8.639e+03  2.196e+03   3.935  0.00011 ***
TR23        -2.769e+03  1.240e+03  -2.233  0.02649 *
TR24         1.103e+02  5.980e+02   0.184  0.85384
TR25         6.337e+02  5.866e+02   1.080  0.28111
TR26        -1.541e+02  2.732e+02  -0.564  0.57318
TR30         2.074e-01  2.314e-01   0.896  0.37095
TR31        -5.463e-01  4.811e-01  -1.136  0.25726
TR35        -4.489e-01  4.885e-01  -0.919  0.35912
TR36        -1.787e-01  6.746e-01  -0.265  0.79130
TR37        -9.072e-01  6.099e-01  -1.488  0.13822
TR41         8.081e-01  3.493e-01   2.313  0.02157 *
TR42         6.574e-01  5.547e-01   1.185  0.23719
TR43         2.004e+00  6.737e-01   2.975  0.00324 **
TR48        -2.316e+03  9.407e+02  -2.462  0.01452 *
TR49        -9.878e+01  1.306e+03  -0.076  0.93977
TR51        -2.454e+01  1.421e+03  -0.017  0.98624
TR52         6.242e+03  1.103e+03   5.660 4.41e-08 ***
TR53        -2.267e+03  1.282e+03  -1.768  0.07841 .
TR54        -5.290e+02  1.631e+03  -0.324  0.74596
TR55        -2.086e+03  2.831e+03  -0.737  0.46199
TR56         1.992e+03  1.800e+03   1.107  0.26945
```

```
Seasonality  3.552e-01  4.241e-02   8.375 5.12e-15 ***
MA3          1.629e+02  4.938e+01   3.299  0.00112 **
MA9         -1.011e+03  4.382e+02  -2.307  0.02192 *
MA11        -4.681e+03  3.347e+03  -1.399  0.16322
MA18         5.227e+04  1.163e+04   4.496 1.09e-05 ***
Cross1       1.907e+04  1.929e+03   9.889  < 2e-16 ***
Cross3      -5.076e+01  8.658e+02  -0.059  0.95329
Cross6       1.018e+05  9.748e+03  10.443  < 2e-16 ***
Cross10      1.392e+03  1.440e+03   0.967  0.33463
TR75        -3.384e+03  2.182e+03  -1.551  0.12229
TV1          3.073e+01  4.966e+01   0.619  0.53661
TV2         -2.481e+01  5.156e+01  -0.481  0.63088
TV4          2.442e+02  1.163e+02   2.100  0.03678 *
TV6         -2.191e+02  1.195e+02  -1.834  0.06797 .
TVC1         1.606e+02  3.939e+02   0.408  0.68394
TVC4         1.540e+02  1.261e+03   0.122  0.90292
TVC5        -3.326e+03  1.975e+03  -1.684  0.09349 .
TR76         4.865e+01  5.568e+00   8.738 4.66e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 36050 on 234 degrees of freedom
Multiple R-squared:  0.9157,   Adjusted R-squared:  0.8966
F-statistic: 47.95 on 53 and 234 DF,  p-value: < 2.2e-16
```

Οι αστερίσκοι μας υποδεικνύουν ότι οι επεξηγηματικές μεταβλητές *TR3, TR10, TR20, TR21, TR23, TR41, TR43, TR48, TR52, TR53, Seasonality, MA3, MA9, MA18, Cross1, Cross6, TV4, TV6, TVC5* και *TR76* θεωρούνται στατιστικά σημαντικές.

Στη συνέχεια θα προχωρήσουμε στη μελέτη πιθανής ύπαρξης πολυσυγγραμμικότητας στο μοντέλο.

```
correlation <- sapply
(
     1:ncol(data2),
     function(i)
{
          names(data2[, -i])[which(cor(data2[, i], data2[, -i])
> 0.90)]
}
)
 correlation <- unique(unlist(correlation))
 correlation
```

```
"TR36" "TR42" "TR37" "TR43" "TR49" "TR55" "TR41" "TR30" "TR31" "TR35
" "TR54" "TR48" "TV2"  "TV1"  "TV6"  "TV4"
```

Το ότι υπάρχει, λοιπόν, έντονη συσχέτιση (> 0.9) μεταξύ των παραπάνω μεταβλητών, και μιας και όπως έχουμε δει προηγουμένως, έχουμε αρκετές μεταβλητές στατιστικά μη σημαντικές, καταλαβαίνουμε ότι το

μοντέλο μπορεί να έχει περιθώρια βελτίωσης. Γι΄αυτό θα προχωρήσουμε στην υλοποίηση των τριών βηματικών διαδικασιών αλλά και των ποινικοποιημένων τεχνικών ώστε να έρθουμε κοντά στην επιλογή καταλληλότερου μοντέλου.

5.2 Εφαρμογή των διαδικασιών κατά βήματα στην *R*

Αρχικά, θα εφαρμόσουμε τις κατά βήματα διαδικασίες επιλογής μοντέλου, χρησιμοποιώντας ως κριτήρια σύγκρισης τόσο το κριτήριο *AIC* καθώς και το *BIC*. Πριν εισάγουμε τις εντολές για τις επαναληπτικές μεθόδους, πρέπει να ορίσουμε τα μοντέλα από τα οποία ξεκινούν. Έτσι, θέτουμε το κενό μοντέλο παλινδρόμησης που περιέχει μόνο τη σταθερά (*null*), το οποίο χρησιμοποιεί ως αφετηρία του η μέθοδος της προς τα εμπρός επιλογής καθώς και η συνδυαστική μέθοδος και το πλήρες μοντέλο παλινδρόμησης που συμπεριλαμβάνει όλες τις επεξηγηματικές μεταβλητές (*full*), από το οποίο ξεκινάει η μέθοδος της προς τα πίσω απαλοιφής.

```
null <-lm(Y ~1, data = data2)

full <-lm(Y ~., data = data2)
```

Για την υλοποίηση και των τριών βηματικών διαδικασιών χρησιμοποιούμε την εντολή *step(. , direction= ".")*. Στο *direction* βάζουμε είτε "*backward*", είτε "*forward*", είτε "*both*", ανάλογα με το ποια διαδικασία καλούμαστε να χρησιμοποιήσουμε κάθε φορά.

Έτσι, εκτελούμε τις μεθόδους της προς τα πίσω απαλοιφής, της προς τα εμπρός επιλογής και της συνδυαστικής διαδικασίας αντίστοιχα παρουσιάζοντας τα τελευταία μοντέλα που παράχθηκαν με την υλοποίηση της κάθε διαδικασίας, εφαρμόζοντας τα δύο κριτήρια πληροφορίας *AIC* και *BIC*:

i) Κριτήριο *AIC*

```
step(full,direction="backward")
```

```
Step:  AIC=6052.26
Y ~ TR3 + TR5 + TR10 + TR20 + TR21 + TR23 + TR31 + TR41 + TR42 +
    TR43 + TR48 + TR52 + TR53 + TR56 + Seasonality + MA3 + MA9 +
    MA18 + Cross1 + Cross6 + TR75 + TV4 + TV6 + TVC5 + TR76
```

```
step(null, scope=list(lower=null, upper=full), direction="forw
ard")

Step:  AIC=6060.03
Y ~ Seasonality + TR76 + TR42 + TR52 + TR13 + TR6 + TV6 + TVC5 +
    TR21 + TR16 + TR20 + TR3 + TR41 + TR43 + TR4 + TR10 + TR56 +
    MA18 + TR48 + TR5 + TR23 + TV1 + Cross6 + Cross1 + TR31 +
    MA3 + TR75 + TR51 + MA9 + TV4


step(null, scope=list(upper=full), direction="both")

Step:  AIC=6052.11
Y ~ Seasonality + TR76 + TR42 + TR52 + TV6 + TVC5 + TR21 + TR20 +
    TR3 + TR41 + TR43 + TR10 + TR56 + MA18 + TR48 + TR5 + TR23 +
    Cross6 + Cross1 + TR31 + MA3 + TR75 + TR51 + MA9 + TV4
```

ii) Κριτήριο *BIC*

Επιλέγοντας τώρα για κριτήριο σύγκρισης το *BIC,* βάζουμε σε όλες τις εντολές το επιπλέον όρισμα *k=log(nrow(data2))* για να μετατρέψουμε το κριτήριο σύγκρισης από *AIC* σε *BIC*.

```
step ( full , direction ="backward",k=log ( nrow (data2)))


Step:  AIC=6132.83
Y ~ TR3 + TR10 + TR20 + TR21 + TR23 + TR31 + TR41 + TR42 + TR43 +
    TR47 + TR48 + TR52 + Seasonality + MA3 + MA9 + MA18 + Cross1 +
    Cross6 + TV4 + TR76

step (null , scope = list ( lower =null ,
upper=full),direction ="forward", k=log( nrow (data2)))

Step:  AIC=6245.85
Y ~ Seasonality + TR76 + TR42 + TR52 + TR13 + TR6 + TV6 + TVC5 +
    TR21 + TR16 + TR20


step ( null , scope = list ( lower =null ,
upper=full),direction ="both",k=log( nrow (data2)))

Step:  AIC=6245.85
Y ~ Seasonality + TR76 + TR42 + TR52 + TR13 + TR6 + TV6 + TVC5 +
    TR21 + TR16 + TR20
```

Αξίζει να παρατηρήσουμε ότι το κριτήριο *BIC* δίνει πιο φειδωλά μοντέλα σε σύγκριση με τα μοντέλα που παράγονται με τη χρήση του *AIC,* αφού

όπως έχουμε επισημάνει στη θεωρία το *BIC* ποινικοποιεί αυστηρότερα την αύξηση των παραμέτρων στο μοντέλο. Ακόμα, να σημειώσουμε ότι το κριτήριο *BIC* καταλήγει στο ίδιο μοντέλο για τις *forward* και *stepwise* μεθόδους.

	AIC			***BIC***		
Μεταβλητές	**Backward**	**Forward**	**Stepwise**	**Backward**	**Forward**	**Stepwise**
TR2						
TR3	✓	✓	✓	✓		
TR4		✓				
TR5	✓	✓	✓			
TR6		✓			✓	✓
TR7						
TR9						
TR10	✓	✓	✓	✓		
TR11						
TR12						
TR13		✓			✓	✓
TR14						
TR16		✓			✓	✓
TR20	✓	✓	✓	✓	✓	✓
TR21	✓	✓	✓	✓	✓	✓
TR23	✓	✓	✓	✓		
TR24						
TR25						
TR26						
TR30						
TR31	✓	✓	✓	✓		
TR35						
TR36						
TR37						
TR41	✓	✓	✓	✓		
TR42	✓	✓	✓	✓	✓	✓
TR43	✓	✓	✓	✓		
TR48	✓	✓	✓	✓		
TR49						
TR51		✓	✓			
TR52	✓	✓	✓	✓	✓	✓
TR53	✓			✓		

TR54						
TR55						
TR56	✓	✓	✓			
Seasonality	✓	✓	✓	✓	✓	✓
MA3	✓	✓	✓	✓		
MA9	✓	✓	✓	✓		
MA11						
MA18	✓	✓	✓	✓		
Cross1	✓	✓	✓	✓		
Cross3						
Cross6	✓	✓	✓	✓		
Cross10						
TR75	✓	✓	✓			
TV1		✓				
TV2						
TV4	✓	✓	✓	✓		
TV6	✓	✓	✓		✓	✓
TVC1						
TVC4						
TVC5	✓	✓	✓		✓	✓
TR76	✓	✓	✓	✓	✓	✓
Αριθμός μεταβλητών	25	30	25	20	11	11

***Πίνακας 5.2*: Συνοπτικός πίνακας επιλογής επεξηγηματικών μεταβλητών σύμφωνα με τις επαναληπτικές διαδικασίες με τη βοήθεια των κριτηρίων σύγκρισης**

Στη συνέχεια υπολογίζουμε το μέσο τετραγωνικό σφάλμα για κάθε μέθοδο (βλ. ΠΑΡΑΡΤΗΜΑ ΙΙ) και παρατηρούμε ότι έχουμε μικρότερο κατά την υλοποίηση της *Forward* με το κριτήριο *AIC* και μεγαλύτερο με το κριτήριο *BIC* με τις μεθόδους *Forward* και *Stepwise* όπως παρουσιάζονται στον ***Πίνακα 5.3***.

MSE					
AIC			BIC		
Backward	Forward	Stepwise	Backward	Forward	Stepwise
4.6×10^7	2.1×10^7	2.1×10^7	2.2×10^7	3.9×10^7	3.9×10^7

***Πίνακας 5.3*: Μέσο τετραγωνικό σφάλμα για τις επαναληπτικές μεθόδους**

Οι επαναληπτικές μέθοδοι κάνουν αποκλειστική χρήση των εκτιμητριών ελαχίστων τετραγώνων που όπως έχουμε δει στη θεωρία η κατασκευή τους αγνοούν το πρόβλημα της πολυσυγγραμμικότητας. Επομένως, σε καμία περίπτωση δεν είμαστε βέβαιοι, δεδομένου και ότι έχουμε δει ότι αρκετές τυχαίες μεταβλητές εμπλέκονται σε γραμμικές σχέσεις, ότι οι μέθοδοι αυτοί έχουν καταλήξει στο κατάλληλο μοντέλο.

Για να οδηγηθούμε σε πιο έγκυρα αποτελέσματα θα χρησιμοποιήσουμε τις δύο τεχνικές με ποινή, οι οποίες αντιμετωπίζουν το πρόβλημα της πολυσυγγραμμικότητας, τη *Ridge* και τη *LASSO*.

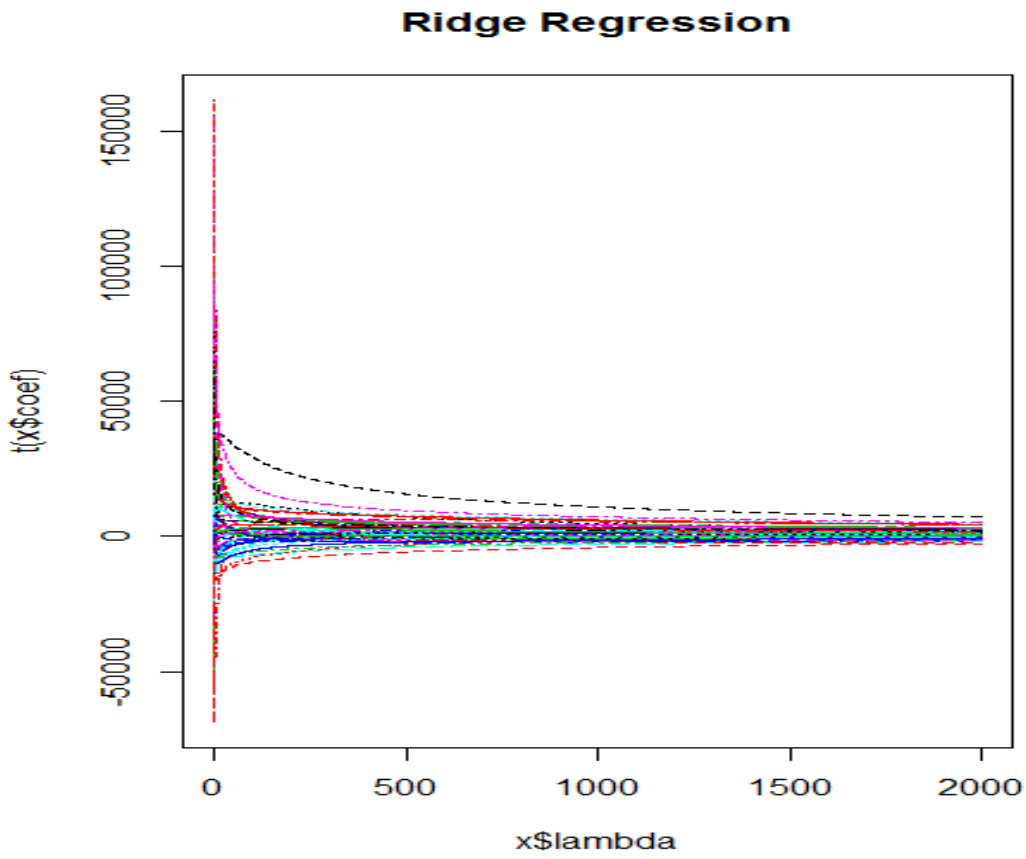
5.3 Εφαρμογή της *Ridge* στην *R*

Αρχικά εφαρμόζοντας την παλινδρόμηση *Ridge*, θα μελετήσουμε την συμπεριφορά των διαφορετικών μοντέλων που λαμβάνουμε για διαφορετικές τιμές της παραμέτρου ποινής. Εφαρμόζουμε την παλινδρόμηση *Ridge* με την εντολή *lm.ridge()* του πακέτου *MASS* που δέχεται ως ορίσματα το γραμμικό μοντέλο που προσαρμόζουμε και το εύρος τιμών που έχουμε δώσει στην παράμετρο ποινής.

```
install.packages("MASS")
library ( MASS )
l.data<-seq (0 ,2000 , length.out =100000)
ridge1 <-lm.ridge (Y~., data =data2, lambda =l.data)
plot(ridge1)
title("Ridge Regression")
```

Με την εντολή *plot()* παίρνουμε το γράφημα των εκτιμητριών *Ridge* συναρτήσει της παραμέτρου ποινής (***Γράφημα 5.1***). Παρατηρούμε ότι όσο αυξάνεται η παράμετρος ποινής, τόσο περισσότερο συρρικνώνονται οι εκτιμήτριες των συντελεστών του μοντέλου μας, χωρίς όμως να μηδενίζονται (σε αντίθεση με τη *LASSO* που θα δούμε παρακάτω), ενώ όταν η παράμετρος ποινής είναι μηδενική έχουμε τις εκτιμήτριες ελαχίστων τετραγώνων.

Στη συνέχεια, θα εντοπίσουμε τις διαφορετικές παραμέτρους ποινής που παράγουν «βέλτιστα» μοντέλα.



***Γράφημα 5.1*: Εκτιμήτριες *Ridge* συναρτήσει λ (πακέτο "MASS")**

i) Επιλογή της παραμέτρου ποινής χρησιμοποιώντας τα κριτήρια πληροφορίας *AIC* και *BIC* (βλ. ΠΑΡΑΡΤΗΜΑ ΙΙ)

Στη συνέχεια υπολογίζουμε τις τιμές των κριτηρίων με τις ακόλουθες εντολές στην *R*:

```
ridge2$lambda[ AIC==min(AIC) ]
```

```
[1] 0.2
```

```
ridge2$lambda[ BIC==min(BIC) ]
```

```
[1] 0.8545455
```

ii) Επιλογή της παραμέτρου ποινής χρησιμοποιώντας την τιμή της παραμέτρου που πρότειναν οι *Hoerl, Kennard* και *Baldwin*:

Δίνεται με την εντολή στην *R*:

```
ridge1$kHKB
```

```
[1] 0.7162074
```

AIC	BIC	HKB
0.2000000	0.8545455	0.7162074

***Πίνακας 5.4*: Σύνοψη τιμών παραμέτρου ποινής**

Από τον ***Πίνακα 5.4*** παρατηρούμε ότι η λίγο μεγαλύτερη τιμή παραμέτρου εντοπίζεται από το κριτήριο *BIC* που σημαίνει ότι επιτυγχάνει τη μεγαλύτερη συρρίκνωση των συντελεστών των επεξηγηματικών μεταβλητών.

Τέλος, οι τιμές των συντελεστών για αυτές τις παραμέτρους ποινής υπολογίζονται παρακάτω με την εντολή *coef()* που δέχεται ως όρισμα την παλινδρόμηση *Ridge* για τις 3 διαφορετικές τιμές της παραμέτρου ποινής:

```
coefficients <-coef (lm.ridge (Y~., data =data2 , lambda= lamb
da))
coefficients      <-as.data.frame(coefficients,row.names      =
names(lambda))


coefficients
```

	AIC	BIC	HKB
V1	338631.7	158170.4	189263.1
TR2	-40859.02	-48173.53	-47463.43
TR3	-271734.4	-261025.1	-263012.5
TR4	42467.13	37799.93	38240.55
TR5	18432.72	20094.06	19827.63
TR6	7735.54	-3923.593	-2155.318
TR7	-5894.589	-6297.071	-6201.931
TR9	-11105.36	-2873.255	-3911.652
TR10	-70637.41	-70742.02	-70876.70
TR11	-520.1315	1678.1110	1395.0993
TR12	-4894.654	-6239.374	-6025.599
TR13	-2352.898	698.6687	55.14519
TR14	4945.416	5191.771	5136.312
TR16	-1711.630	147.6264	-180.7893
TR20	-19409.31	-18212.97	-18466.56
TR21	8639.945	8473.738	8525.933
TR23	-2843.659	-2831.066	-2847.187
TR24	17.9363	-128.889	-108.447
TR25	524.6677	453.1834	469.5237
TR26	-147.1018	-122.1087	-126.9726
TR30	0.2005467	0.1764343	0.1803328
TR31	-0.34133804	-0.09384965	-0.12305365
TR35	-0.248531	0.0340759	0.0021931
TR36	0.00587955	0.24994479	0.22033581
TR37	-0.882277	-0.703614	-0.741601
TR41	0.6645480	0.4665240	0.4874721
TR42	0.5356080	0.3952873	0.4119969
TR43	1.636206	1.097154	1.171854
TR48	-2061.422	-1652.116	-1710.827
TR49	-14.85547	181.06134	147.72312
TR51	-115.1285	-132.3367	-139.7992
TR52	6235.939	6154.731	6174.119
TR53	-1937.070	-1468.469	-1532.441
TR54	-690.1928	-624.2856	-894.7014
TR55	-2379.898	-2633.463	-2618.316
TR56	2260.656	2783.463	2701.176
Seasonality	0.3722709	0.4019082	0.3973641
MA3	161.5459	149.6711	152.3189
MA9	-970.6429	-861.8610	-881.4109
MA11	-3629.018	1550.519	1888.816
MA18	46057.16	42204.02	43348.13
Cross1	17519.54	14094.96	14682.94
Cross3	-37.96587	-43.68950	-40.77437
Cross6	94365.50	77603.39	80509.20
Cross10	1469.153	1474.260	1484.587
TR75	-3452.435	-3559.969	-3543.810
TV1	26.69854	19.87534	20.89866
TV2	-20.36715	-13.02747	-14.11531

TV4	179.59352	95.88123	106.27563
TV6	-152.87162	66.71607	77.46097
TVC1	181.3598	208.0402	204.8569
TVC4	225.5417	295.0701	289.6653
TVC5	-3534.779	-3966.364	-3895.002
TR76	46.03285	40.67306	41.56297

Πίνακας 5.5: **Εκτιμήτριες *Ridge* των επεξηγηματικών μεταβλητών για τις 3 διαφορετικές τιμές του λ. Είναι σημειωμένες αυτές που συρρικνώνονται προς το μηδέν.**

Με αυτόν τον τρόπο λαμβάνουμε τις εκτιμήτριες *Ridge* των επεξηγηματικών μεταβλητών για τις 3 διαφορετικές τιμές της παραμέτρου ποινής, καταλήγοντας έτσι σε 3 διαφορετικά μοντέλα.

Παραπάνω (***Πίνακας 5.5***) έχουμε επισημάνει τις επεξηγηματικές μεταβλητές των οποίων οι συντελεστές συρρικνώνονται προς το μηδέν, υποδηλώνοντας ότι αυτές οι μεταβλητές πιθανόν να μην είναι σημαντικές για την πρόβλεψη των πωλήσεων του προϊόντος.

Στη συνέχεια, θα χρησιμοποιήσουμε το πακέτο *"glmnet"* ως έναν ακόμα τρόπο να καταλήξουμε κοντά στο «βέλτιστο» μοντέλο, κάνοντας χρήση της παλινδρόμησης *Ridge*. Αρχικά, δημιουργούμε ένα πίνακα που περιέχει όλα τα δεδομένα για τις επεξηγηματικές μεταβλητές και για την μεταβλητή απόκρισης:

```
install.packages("glmnet")
library(glmnet)
x <- model.matrix(Y ~ . - 1, data = data2)
y <- data2$Y
```

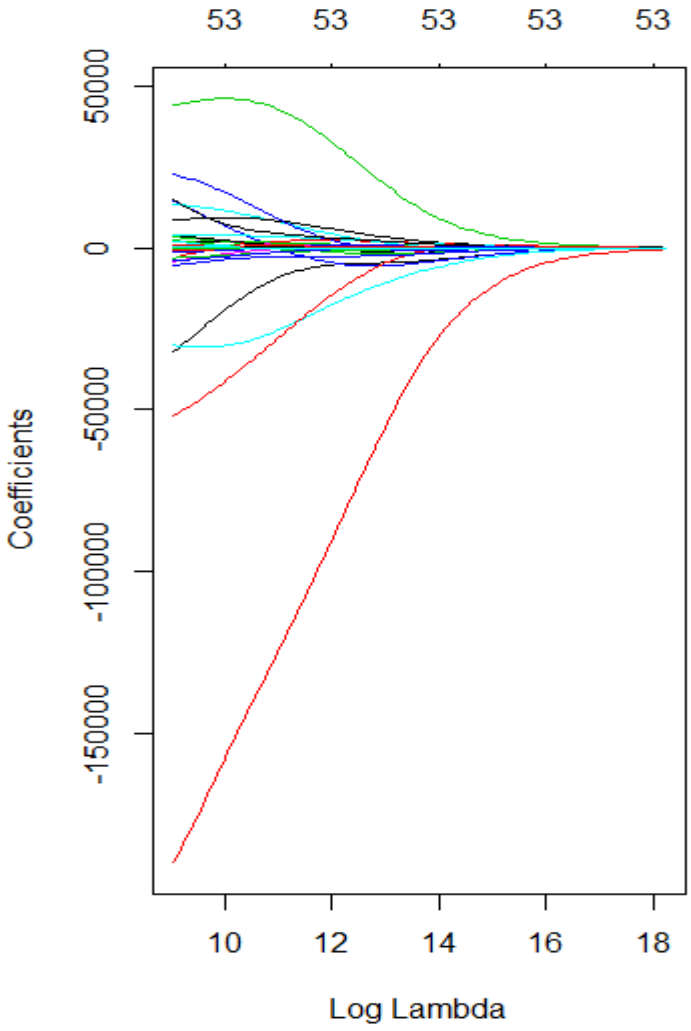
Πραγματοποιούμε την παλινδρόμηση *Ridge* κάνοντας χρήση της συνάρτησης *glmnet()* με όρισμα *alpha=0*.

Αξίζει να σημειώσουμε ότι η συνάρτηση *glmnet()* τυποποιεί τα δεδομένα πριν την εφαρμογή της παλινδρόμησης, δηλαδή τόσο οι επεξηγηματικές μεταβλητές όσο και η μεταβλητή απόκρισης κανονικοποιούνται και κεντράρονται. Όμως, τα αποτελέσματα δίνονται βάση της αρχικής κλίμακας. Στη συνέχεια, κάνουμε το κοινό γράφημα των τιμών των συντελεστών των επεξηγηματικών μεταβλητών συναρτήσει της

παραμέτρου ποινής. Χρησιμοποιούμε την εντολή *plot()* με όρισμα *xvar="lambda"* για να ορίσουμε ότι στον άξονα x αντιστοιχεί η τιμή της παραμέτρου ποινής, ενώ για να δώσουμε τα ονόματα στους άξονες χρησιμοποιούμε το όρισμα *label= TRUE*:

```
fit.ridge <- glmnet(x, y, alpha = 0)
plot(fit.ridge, xvar = "lambda", label = TRUE)
```

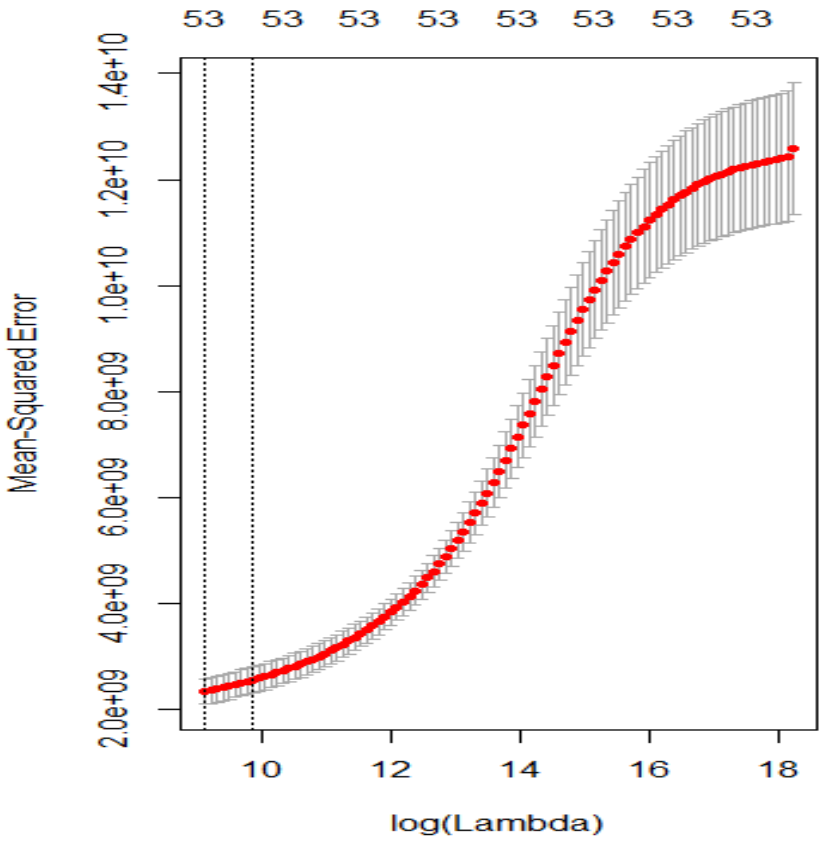
Από το ***Γράφημα 5.2*** παρατηρούμε ότι όσο αυξάνεται η τιμή της παραμέτρου ποινής, οι συντελεστές του μοντέλου συρρικνώνονται χωρίς όμως να μηδενίζεται κάποιος, όπως βλέπουμε και στο πάνω μέρος του γραφήματος όπου ο αριθμός των μη μηδενικών συντελεστών παραμένει σταθερός.



***Γράφημα 5.2*: Εκτιμήτριες *Ridge* συναρτήσει του logλ (glmnet)**

Για την επιλογή της βέλτιστης τιμής της παραμέτρου ποινής χρησιμοποιούμε τη διαδικασία του *Cross Validation*. Εφαρμόζουμε τη διαδικασία αυτή, με την χρήση της συνάρτησης *cv.glmnet()*. Ακολούθως, κατασκευάζουμε το γράφημα που δίνει το μέσο τετραγωνικό σφάλμα (*MSE*) συναρτήσει του λογαρίθμου της παραμέτρου ποινής. Η πρώτη κάθετη διακεκομμένη γραμμή δίνει το σημείο όπου το μέσο τετραγωνικό σφάλμα έχει τη μικρότερη τιμή, ενώ η δεύτερη δείχνει το σημείο που έχει επιλεχθεί από τον κανόνα του «*1-standard error*» (***Γράφημα 5.3***).

```
cv.ridge <- cv.glmnet(x, y, alpha = 0)
plot(cv.ridge)
```



***Γράφημα 5.3*: *MSE* συναρτήσει του λογαρίθμου της παραμέτρου ποινής**

Η τιμή της επιλεγμένης παραμέτρου ποινής του *Cross Validation* είναι:

```
cv.ridge$lambda.1se
[1] 21271.79
```

Τέλος, υπολογίζουμε τους συντελεστές παλινδρόμησης:

```
coef(cv.ridge$lambda.1se)
```

```
54 x 1 sparse Matrix of class "dgCMatrix"
                        s0
(Intercept)  8.888016e+04
TR2         -4.695733e+04
TR3         -2.571792e+05
TR4          3.665402e+04
TR5          2.041084e+04
TR6         -7.571308e+03
TR7         -6.542171e+03
TR9         -7.187244e+02
TR10        -7.082579e+04
TR11         1.691855e+03
TR12        -6.307796e+03
TR13         2.391274e+03
TR14         5.309070e+03
TR16         8.908636e+02
TR20        -1.793754e+04
TR21         8.421426e+03
TR23        -2.782715e+03
TR24        -1.643611e+02
TR25         4.182309e+02
TR26        -1.085529e+02
TR30         1.454377e-01
TR31        -6.756966e-02
TR35         9.360934e-02
TR36         3.410537e-01
TR37        -5.840445e-01
TR41         4.217064e-01
TR42         3.879218e-01
TR43         9.833855e-01
TR48        -1.516415e+03
TR49         2.602426e+02
TR51        -7.962097e+01
TR52         6.072032e+03
TR53        -1.363435e+03
TR54        -9.997425e+02
TR55        -2.649914e+03
TR56         3.000471e+03
Seasonality  4.106794e-01
MA3          1.429322e+02
MA9         -8.165736e+02
MA11        -8.255885e+02
MA18         3.958193e+04
Cross1       1.282571e+04
Cross3      -5.350416e+01
Cross6       7.129614e+04
Cross10      1.445345e+03
TR75        -3.586311e+03
TV1          1.790854e+01
TV2         -1.097283e+01
TV4          7.692574e+01
TV6         -4.700599e+01
TVC1         2.160317e+02
TVC4         3.009189e+02
TVC5        -4.106634e+03
TR76         3.866193e+01
```

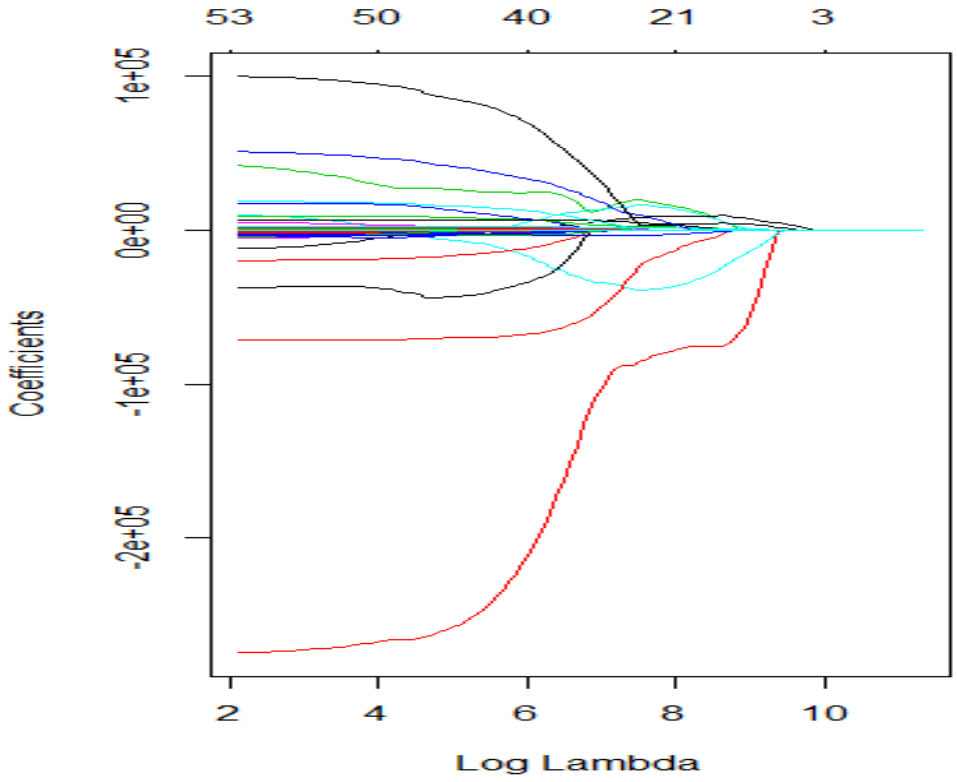
Πίνακας 5.6*: Εκτιμώμενοι συντελεστές παλινδρόμησης *Ridge

Από τον ***Πίνακα 5.6*** παρατηρούμε ότι κανένας συντελεστής δε μηδενίστηκε, εν τούτοις 9 συντελεστές έχουν συρρικνωθεί προς το μηδέν: οι *TR30*, *TR31*, *TR35*, *TR36*, *TR37*, *TR41*, *TR42*, *TR43* και *Seasonality*.

5.4 Εφαρμογή της *LASSO* στην *R*

Εφαρμόζουμε στην *R* τη μεθοδολογία *LASSO* με τη βοήθεια του πακέτου *glmnet* όπως ακριβώς το υλοποιήσαμε και στην παλινδρόμηση *Ridge* αντικαθιστώντας μόνο το όρισμα *alpha=0* με *alpha=1*. Επομένως έχουμε:

```
fit.lasso <- glmnet(x, y, alpha = 1)
plot(fit.lasso, xvar = "lambda", label = TRUE)
```



***Γράφημα 5.4*: Εκτιμήτριες *LASSO* συναρτήσει του logλ**

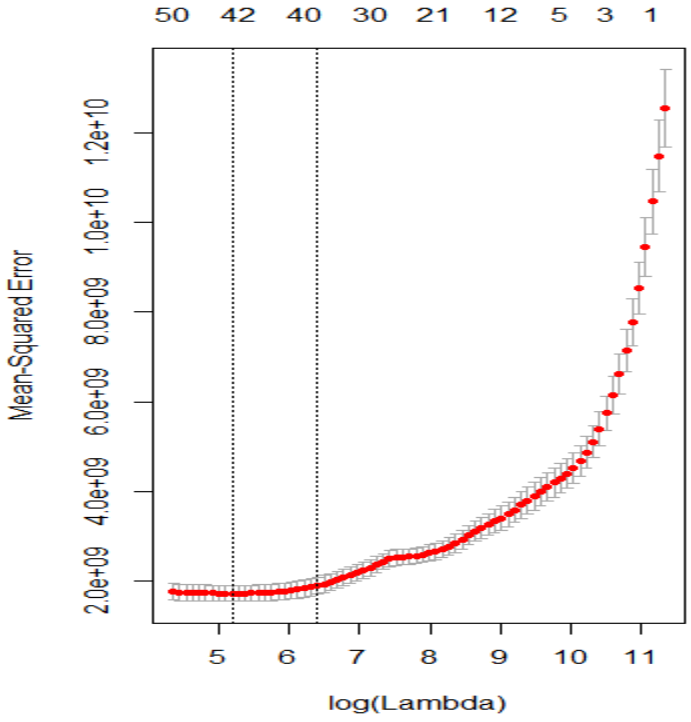
Παρατηρούμε (***Γράφημα 5.4***) ότι όσο αυξάνεται η τιμή της παραμέτρου ποινής, οι συντελεστές του μοντέλου συρρικνώνονται με ταυτόχρονο

μηδενισμό κάποιων. Στο πάνω μέρος του γραφήματος δίνεται ο αριθμός των μη μηδενικών συντελεστών για τις αντίστοιχες τιμές της παραμέτρου ποινής και όπως παρατηρούμε ο αριθμός τους συνεχώς μειώνεται με την αύξηση της παραμέτρου ποινής. Τέλος διακρίνουμε, ότι σε αντίθεση με τη *Ridge*, η *LASSO* δεν επιτρέπει την εναλλαγή προσήμων.

Χρησιμοποιώντας και πάλι την διαδικασία του *Cross Validation* για να επιλέξουμε την κατάλληλη τιμή της παραμέτρου ποινής (***Γράφημα 5.5***) έχουμε:

```
cv.lasso <- cv.glmnet(x, y, alpha = 1)

plot(cv.lasso)
```



***Γράφημα 5.5*: *MSE* συναρτήσει του λογαρίθμου του λ**

Η τιμή της επιλεγμένης παραμέτρου ποινής είναι:

```
cv.lasso$lambda.1se
[1] 552.008
```

Τέλος για τους εκτιμώμενους συντελεστές παλινδρόμησης προκύπτει:

```
coef(cv.lasso$lambda.1se)
```

```
54 x 1 sparse Matrix of class "dgCMatrix"
                      s0
(Intercept) -1.210884e+05
TR2         -2.760991e+04
TR3         -1.844770e+05
TR4          2.337510e+04
TR5          5.628076e+03
TR6         -2.360498e+04
TR7          .
TR9          .
TR10        -6.637872e+04
TR11         .
TR12         .
TR13         8.324004e+03
TR14         1.400340e+03
TR16         1.859548e+03
TR20        -9.303292e+03
TR21         5.642580e+03
TR23        -1.782647e+03
TR24         .
TR25         .
TR26        -7.241716e+01
TR30         2.200945e-01
TR31         .
TR35         .
TR36         2.854695e-01
TR37        -6.908198e-02
TR41         2.570062e-01
TR42         5.132086e-02
TR43         3.637903e-01
TR48        -1.092363e+03
TR49         .
TR51        -1.971408e+02
TR52         6.419002e+03
TR53        -5.605532e+02
TR54        -5.841412e+02
TR55        -1.176048e+03
TR56         2.221960e+03
Seasonality  4.352610e-01
MA3          9.381854e+01
MA9         -5.918894e+02
MA11         .
MA18         3.002841e+04
Cross1       1.075949e+04
Cross3       .
Cross6       6.024180e+04
Cross10      9.840194e+02
TR75        -2.938652e+03
TV1          8.984035e+00
TV2          .
TV4          3.395159e+01
TV6          .
TVC1         1.299517e+02
TVC4         .
TVC5        -3.249068e+03
TR76         3.927487e+01
```

Πίνακας 5.7*: Εκτιμώμενοι συντελεστές *LASSO

Από τον ***Πίνακα 5.7*** παρατηρούμε ότι έχουν μηδενιστεί οι συντελεστές των επεξηγηματικών μεταβλητών *TR7, TR9, TR11, TR12, TR24, TR25, TR31, TR35, TR49, MA11, Cross3, TV2, TV6* και *TVC4*.

5.5 Σύγκριση *Ridge-LASSO*

Στον ***Πίνακα 5.8*** παρουσιάζουμε ποιες επεξηγηματικές μεταβλητές έχουν επιλεχθεί μετά την υλοποίηση των ποινικοποιημένων μεθόδων, αφαιρώντας στη *Ridge* τις μεταβλητές που έχουν συρρικνωθεί προς το μηδέν. Καταλήγουμε ότι η *LASSO* απαλοίφει περισσότερες μεταβλητές.

Μεταβλητές	**Ridge**	**LASSO**
TR2	✓	✓
TR3	✓	✓
TR4	✓	✓
TR5	✓	✓
TR6	✓	✓
TR7	✓	
TR9	✓	
TR10	✓	✓
TR11	✓	
TR12	✓	
TR13	✓	✓
TR14	✓	✓
TR16	✓	✓
TR20	✓	✓
TR21	✓	✓
TR23	✓	✓
TR24	✓	
TR25	✓	
TR26	✓	✓
TR30		✓
TR31		
TR35		
TR36		✓
TR37		✓
TR41		✓
TR42		✓
TR43		✓

TR48	✓	✓
TR49	✓	
TR51	✓	✓
TR52	✓	✓
TR53	✓	✓
TR54	✓	✓
TR55	✓	✓
TR56	✓	✓
Seasonality		✓
MA3	✓	✓
MA9	✓	✓
MA11	✓	
MA18	✓	✓
Cross1	✓	✓
Cross3	✓	
Cross6	✓	✓
Cross10	✓	✓
TR75	✓	✓
TV1	✓	✓
TV2	✓	
TV4	✓	✓
TV6	✓	
TVC1	✓	✓
TVC4	✓	
TVC5	✓	✓
TR76	✓	✓
Αριθμός μεταβλητών	44	39

***Πίνακας 5.8*: Συνοπτικός πίνακας επιλογής επεξηγηματικών μεταβλητών με τη βοήθεια των ποινικοποιημένων μεθόδων**

Για την τελική επιλογή του μοντέλου θα ελέγξουμε ποια μέθοδος πετυχαίνει το μικρότερο μέσο άθροισμα τετραγώνων των σφαλμάτων:

```
coef_rid <- as.matrix(coef(cv.ridge))

coef_las <- as.matrix(coef(cv.lasso))

yhat_rid<-cbind(1, x) %*% coef_rid

yhat_las<-cbind(1, x) %*% coef_las

n<-length(x)
```

```
mse_rid<-sum((yhat_rid - y)^2)/n-53

mse_rid
[1] 35465096

mse_las<-sum((yhat_las - y)^2)/n-39

mse_las
[1] 24748405

(mse_rid - mse_las) / mse_las
[1] 0.4330255
```

MSE	
Ridge	LASSO
3.5×10^7	2.5×10^7

Βλέπουμε ότι σε όλα τα δεδομένα η *LASSO* κάνει καλύτερη δουλειά από ότι η *Ridge* πετυχαίνοντας και απαλοιφή μεταβλητών (μηδενίζει 14 μεταβλητές) και καλύτερη ακρίβεια πρόβλεψης.

ΣΥΝΟΨΗ

Συνοψίζοντας, ξεκίνησαμε με ένα μοντέλο που περιείχε 141 επεξηγηματικές μεταβλητές εκ των οποίων οι 16 ήταν μηδενικές και έτσι αφαιρέθηκαν εξ'αρχής από το μοντέλο. Έτσι καταλήξαμε να έχουμε 125 μεταβλητές. Κάνοντας την παλινδρόμηση παρατηρούμε ότι εκτός του ότι έχουμε τέλεια προσαρμογή ($R^2 = 1$), σε μερικές μεταβλητές δεν υπολογίζονται οι συντελεστές τους, λόγω πολυσυγγραμμικότητας. Αφαιρώντας τις μεταβλητές αυτές, τις *dummies* καθώς και 2 μεταβλητές που ταυτίζονται με την μεταβλητή απόκρισης, καταλήγουμε στις τελικές 53 όπου εφαρμόσαμε τις πέντε μεθόδους επιλογής μεταβλητών, με βάση τα διάφορα κριτήρια που αναπτύξαμε στην θεωρία.

Αρχικά εκτελέσαμε τις τρεις επαναληπτικές μεθόδους χρησιμοποιώντας τα κριτήρια πληρόφοριας *AIC* και *BIC*, επιβεβαιώνοντας ότι το *BIC* δίνει πιο φειδωλά μοντέλα σε σύγκριση με τα μοντέλα που παράγονται με τη χρήση του *AIC.* Ενώ παράλληλα παρατηρήσαμε ότι το κριτήριο *BIC* δίνει το ίδιο μοντέλο για τις *forward* και *stepwise* μεθόδους. Ωστόσο, αν και η *stepwise* μέθοδος προτιμάται έναντι της *backward* και της *forward*, λόγω του διπλού ελέγχου που πράττει, οι τρεις αυτές διαδικασίες επιλέγουν συνήθως αρεστά αλλά όχι βέλτιστα μοντέλα. Επίσης η αφαίρεση κάποιας συμμεταβλητής οδηγεί αυτομάτως στην αύξηση της σημαντικότητας κάποιας άλλης, με αποτέλεσμα η πραγματική της επίδραση πολλές φορές να υπερεκτιμάται.

Για τον λόγο αυτό, καθώς και για το γεγονός ότι υπάρχει έντονη συσχέτιση μεταξύ κάποιων επεξηγηματικών μεταβλητών που μας οδηγεί σε όχι και τόσο έγκυρα αποτελέσματα, χρησιμοποιήσαμε στη συνέχεια τις δύο τεχνικές με ποινή, καθώς αντιμετωπίζουν το πρόβλημα της πολυσυγγραμμικότητας, τη *Ridge* και τη *LASSO*.

Εφαρμόζουμε στα δεδομένα μας με την παλινδρόμηση *Ridge*, η οποία ναι μεν ελέγχει τη διασπορά των συντελεστών του μοντέλου, αλλά δεν μηδενίζει κάποιον συντελεστή κατά την συρρίκνωση των συντελεστών. Στη συνέχεια κάνοντας χρήση των τριών διαφορετικών μεθόδων επιλογής της παραμέτρου ποινής με το πακέτο *MASS*, καθώς και της *10-fold Cross Validation* με τη βοήθεια του πακέτου *glmnet*, καταλήγουμε στις εκτιμήτριες *Ridge*, υποδεικνύοντας τις συμμεταβλητές των οποίων οι συντελεστές έχουν συρρικνωθεί κοντά στο μηδέν. Με αυτό το

σκεπτικό θα μπορούσαμε να θεωρήσουμε όλες τις υπόλοιπες μεταβλητές ως σημαντικές για να μπουν στο μοντέλο, από τη στιγμή όμως που η *Ridge* δεν δρα άμεσα ως μέθοδος επιλογής μεταβλητών αυτό δε θα ήταν απόλυτα σωστό.

Έτσι καταλήξαμε στη *LASSO*, η οποία έχει το πλεονέκτημα, έναντι της *Ridge*, ότι μηδενίζει τους συντελεστές των μη στατιστικά σημαντικών επεξηγηματικών μεταβλητών, δρώντας έτσι ως μέθοδος επιλογής μεταβλητών. Εφαρμόσαμε το πακέτο *glmnet* και επιλέξαμε τους επιθυμητούς παράγοντες συρρίκνωσης βάσει της *10-fold Cross Validation*, απαλοίφοντας 14 μεταβλητές.

Τελικά μετά την εφαρμογή της *LASSO* καταλήγουμε στο παρακάτω μοντέλο:

```
> datalasso<-data2[,-c(7,8,10,11,18,19,22,23,30,40,43,48,50,52)]
> fitlasso<-lm(Y~.,data = datalasso)
> summary(fitlasso)

Call:
lm(formula = Y ~ ., data = datalasso)

Residuals:
    Min      1Q  Median      3Q     Max
-111127  -20398   -1017   21728  108183

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.048e+04  1.967e+05  -0.308 0.758684
TR2         -3.720e+04  4.404e+04  -0.845 0.399088
TR3         -2.902e+05  5.032e+04  -5.768 2.38e-08 ***
TR4          1.934e+04  2.816e+04   0.687 0.492906
TR5          1.141e+04  8.995e+03   1.268 0.205922
TR6         -4.323e+03  1.462e+04  -0.296 0.767772
TR10        -7.373e+04  1.762e+04  -4.183 3.99e-05 ***
TR13        -7.658e+03  6.365e+03  -1.203 0.230029
TR14         1.294e+03  2.655e+03   0.487 0.626516
TR16        -1.758e+03  2.358e+03  -0.746 0.456552
TR20        -1.962e+04  4.193e+03  -4.679 4.75e-06 ***
TR21         8.751e+03  1.768e+03   4.948 1.38e-06 ***
TR23        -2.985e+03  9.031e+02  -3.305 0.001090 **
TR26        -1.674e+02  2.545e+02  -0.658 0.511292
TR30         1.331e-01  2.129e-01   0.625 0.532432
TR36         3.438e-01  3.382e-01   1.017 0.310311
TR37        -1.103e+00  4.347e-01  -2.537 0.011795 *
TR41         5.179e-01  1.129e-01   4.587 7.13e-06 ***
TR42         4.415e-01  4.738e-01   0.932 0.352401
TR43         1.235e+00  2.665e-01   4.633 5.84e-06 ***
TR48        -1.816e+03  8.359e+02  -2.172 0.030801 *
TR51        -7.280e+02  1.332e+03  -0.547 0.585125
TR52         6.596e+03  1.024e+03   6.440 6.14e-10 ***
TR53        -1.311e+03  1.166e+03  -1.125 0.261703
TR54        -5.469e+02  1.568e+03  -0.349 0.727477
```

```
TR55         -3.640e+03  1.716e+03  -2.121 0.034946 *
TR56          2.879e+03  1.514e+03   1.902 0.058389 .
Seasonality   3.901e-01  3.914e-02   9.967  < 2e-16 ***
MA3           1.659e+02  4.421e+01   3.752 0.000218 ***
MA9          -1.015e+03  4.114e+02  -2.467 0.014291 *
MA18          4.196e+04  9.040e+03   4.642 5.59e-06 ***
Cross1        1.747e+04  1.699e+03  10.278  < 2e-16 ***
Cross6        9.677e+04  9.094e+03  10.642  < 2e-16 ***
Cross10       1.590e+03  1.108e+03   1.435 0.152480
TR75         -4.287e+03  1.997e+03  -2.147 0.032784 *
TV1           7.915e+00  7.369e+00   1.074 0.283768
TV4           3.100e+01  1.094e+01   2.835 0.004965 **
TVC1          2.286e+02  3.757e+02   0.609 0.543354
TVC5         -2.710e+03  1.465e+03  -1.850 0.065558 .
TR76          4.513e+01  4.984e+00   9.056  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 35730 on 248 degrees of freedom
Multiple R-squared:  0.9122,   Adjusted R-squared:  0.8984
F-statistic:  66.1 on 39 and 248 DF,  p-value: < 2.2e-16
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Α) Διεθνής Βιβλιογραφία

Bertsekas, D.P. (1995). Nonlinear Programming. Athena Scientific, Belmont, MA.

Farebrother, R.W. (1984). The Restricted Least Squares Estimator and Ridge Regression. Communications in Statistics - Theory and Methods, 13:2, p. 191-196

Hastie, T., Tibshirani, R. & Friedman, J. (2008). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. 2nd Ed., New York: Springer-Verlag.

Hastie, T., Tibshirani, R. & Wainwright, M. (2016). Statical Learning with Sparsity-The Lasso and Generalizations.

Hocking, R.R. & Leslie, R.N. (1967). Selection of the Best Subset in Regression Analysis.

Hoerl, A.E. & Kennard, R.W. (1970). Ridge Regression. Biased estimation for nonorthogonal problems. Technometrics 12, 55-67.

Hoerl, A.E, Kennard, R.W. & Baldwin, K.F. (1975). Ridge Regression Some Simulation. Communications in Statistics Theory and Methods, Vol. **4**, p. 105-123.

James G., Witten, D., Hastie, T., Tibshirani, R. (2013). An Introduction to Statistical Learning with Applications in R, Springer, New York.

Kasarda, John D. & Shih, Wen-Fu P. (1977). Optimal Bias in Ridge Regression Approaches to Multicollinearity. Sociological Methods & Research, Vol. **5**, p. 461-469.

Konishi, S. & Kitagama, G. (2008). Information Criteria and Statistical Modeling, Springer, New York.

Marquardt, Donald W. & Snee, Ronald D. (1975). Ridge Regression in Practice. The American Statistician, Vol. **29**, No. 1, p. 3-20.

Miller, A.J (2002). Subset Selection in Regression, 2nd Edition, Champan and Hall New York.

Obsorne, M.R. (1985). Finite Algorithms in optimization and Data Analysis, Wiley Series in Probability and Mathematical Statistics, Chichester: Wiley.

Rawlings, John O., Pantula, Sastry G. & Dickey, David A. (2001). Applied Regression Analysis. A Research Tool, Springer.

Seber, George A. F. & Lee, Alan J. (2003). Linear Regression Analysis, Wiley.

Theobald, C. M. (1974). Generalizations of Mean Square Error Applied to Ridge Regression, Journal of the Royal Statistical Society, Series B (Methodological), Vol. **36**, p. 103-106.

Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. Journal of the Royal Statistic Society. Series B (Methological), Vol. **58**, Issue 1, p.267-288.

Wenjiang, J. F. (1998) Penalized Regressions: The Bridge versus the Lasso. Journal of Computational and Graphical Statistics, 7:3, p. 397-416.

Wu, T. & Lange, K. (2008). Coordinate Descent Algorithms for Lasso Penalized Regression. The Annals of Applied Statistics, Vol. **2**, No. 1, p. 224-244.

Zou, Hui & Hastie, Trevor (2005). Regularization and Variable Selection via the elastic net. Journal of the Royal Statistical Society, Series B: p.301–320.

Β) Ελληνική Βιβλιογραφία

Καρώνη, Χ. & Οικονόμου, Π. (2010). Στατιστικά Μοντέλα Παλινδρόμησης, Εκδόσεις Συμεών, Αθήνα.

Κούτρας, Μ. (2014). Ανάλυση Παλινδρόμησης και Ανάλυση Διακύμανσης, Πανεπιστημιακές Σημειώσεις, ΠΜΣ «Εφαρμοσμένη Στατιστική».

Φουσκάκης, Δ. (2013). Ανάλυση Δεδομένων με Χρήση της R, Εκδόσεις Τσότρας, Αθήνα.

ΠΑΡΑΡΤΗΜΑ Ι

Κυρτές συναρτήσεις και υποκλίσεις

Ορισμός 4.3(Κυρτότητα):

Μία περιοχή D είναι κυρτή αν $\forall x_1, x_2 \epsilon D$ και $\forall \alpha \epsilon [0,1]$ ισχύει:

$$\alpha x_1 + (1-\alpha) x_2 \epsilon D.$$

Μία συνάρτηση $f(x)$ είναι κυρτή αν:

i) το πεδίο ορισμού της D είναι κυρτό,

ii) $f(\alpha x_1 + (1-\alpha) x_2) \leq \alpha f(x_1) + (1-\alpha) f(x_2)$.

Ορισμός 4.4(Θετικά ημιορισμένος πίνακας):

Ένας $p \times p$ πίνακας $\boldsymbol{H}$ είναι θετικά ημιορισμένος αν για όλα τα $p \times 1$ διανύσματα w έχουμε $w' \boldsymbol{H} w \geq 0$.

Πρόταση 4.2(Εσσιανός πίνακας και κυρτότητα):

Έστω $\boldsymbol{x}$ ένα $p \times 1$ διάνυσμα και $f(\boldsymbol{x})$ μία βαθμωτή συνάρτηση p μεταβλητών με συνεχείς δεύτερες παραγώγους ορισμένες σε κυρτό πεδίο ορισμού D. Αν ο Εσσιανός πίνακας $\nabla^2 f(\boldsymbol{x})$ είναι θετικά ημιορισμένος για όλα τα $\boldsymbol{x} \epsilon D$, τότε η f είναι κυρτή.

Έχοντας υπόψιν τα παραπάνω μπορούμε να δείξουμε ότι η αντικειμενική συνάρτηση της *LASSO* είναι κυρτή.

Απόδειξη:

Μπορούμε να γράψουμε την αντικειμενική συνάρτηση της *LASSO* ως:

$$h(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta}),$$

όπου $f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X\beta}\|_2^2$ και $g(\boldsymbol{\beta}) = \lambda \|\boldsymbol{\beta}\|_1$, με κυρτό πεδίο ορισμού και οι δύο, τον $\mathbb{R}^p$.

- Για τον Εσσιανό πίνακα της $f(\boldsymbol{\beta})$ έχουμε:

$$\nabla^2 f(\boldsymbol{\beta}) = 2\boldsymbol{X}'\boldsymbol{X}.$$

Για κάθε $p \times 1$ διάνυσμα w έχουμε:

$$w' 2\boldsymbol{X}'\boldsymbol{X} w = \|\boldsymbol{X}w\|_2^2 \geq 0.$$

Άρα ο Εσσιανός πίνακας είναι θετικά ημιορισμένος, επομένως από την *Πρόταση 4.2*, η $f(\boldsymbol{\beta})$ είναι κυρτή.

- Η συνάρτηση $g(\boldsymbol{\beta})$ είναι επίσης κυρτή.

Για κάθε β_1, β_2 και για κάθε $\alpha\epsilon(0,1)$, με $\beta = \alpha\beta_1 + (1-\alpha)\beta_2$ έχουμε:

$$g(\boldsymbol{\beta}) = \lambda\|\alpha\beta_1 + (1-\alpha)\beta_2\|_1$$

$$\leq \lambda\|\alpha\beta_1\|_1 + \lambda\|(1-\alpha)\beta_2\|_1$$

$$= \lambda\alpha\|\beta_1\|_1 + \lambda(1-\alpha)\|\beta_2\|_1$$

$$= \alpha g(\beta_1) + (1-\alpha) g(\beta_2).$$

Πρόταση 4.3(Άθροισμα κυρτών συναρτήσεων):

Αν $f(\boldsymbol{x})$ και $g(\boldsymbol{x})$ είναι κυρτές συναρτήσεις με πεδίο ορισμού D κυρτό, τότε το άθροισμά τους είναι επίσης κυρτή συνάρτηση στο D.

Επομένως, αφού και $f(\boldsymbol{\beta})$ και $g(\boldsymbol{\beta})$ κυρτές από την *Πρόταση 4.3* θα έχουμε ότι και η $h(\boldsymbol{\beta}) = f(\boldsymbol{\beta}) + g(\boldsymbol{\beta})$ είναι επίσης κυρτή.

Εφ'όσον μελετήσαμε την κυρτότητα θα προχωρήσουμε στην διαφορισιμότητα των συναρτήσεων.

Λήμμα 4.1:

Αν η f είναι παραγωγίσιμη στο x_1, τότε ισχύει:

$$f(x_2) \geq f(x_1) + < \nabla f(x_1), x_2 - x_1 >, \forall x_2 \epsilon \mathbb{R}^n.$$

Στην περίπτωση όμως, όπως και στο πρόβλημα της *LASSO*, όπου η f δεν είναι παραγωγίσιμη στο x_1, εισάγουμε το υποδιαφορικό ∂f *της* f το οποίο ορίζεται ως:

$$\partial f(x_1) = \{\omega \epsilon \mathbb{R}^n, f(x_2) \geq f(x_1) + < \omega, x_2 - x_1 >, \forall x_2 \epsilon \mathbb{R}^n\}.$$

Ένα διάνυσμα $\omega \epsilon \partial f(x_1)$ ονομάζεται υπόκλιση της f στο x_1.

Το υποδιαφορικό της L_1-νόρμας είναι:

$$\partial |x_1| = sign(x_1) = \begin{cases} \omega_j = 1, \gamma \iota \alpha \; x_j > 0 \\ \omega_j = -1, \gamma \iota \alpha \; x_j < 0 \\ \omega_j \epsilon [-1,1], \gamma \iota \alpha \; x_j = 0 \end{cases},$$

με $\omega \; \epsilon \mathbb{R}^n$.

Τέλος να σημειώσουμε ότι το υποδιαφορικό κυρτής συνάρτησης αποτελεί μονότονη συνάρτηση.

Πραάγματι αφού,

$$f(x_2) \geq f(x_1) + < \omega_{x_1}, x_2 - x_1 >$$
και
$$f(x_1) \geq f(x_2) + < \omega_{x_2}, x_1 - x_2 >,$$

αθροίζοντας έχουμε:

$$< \omega_{x_1} - \omega_{x_2}, x_1 - x_2 > \geq 0, \forall \omega_{x_1} \epsilon \partial f(x_1), \forall \omega_{x_2} \epsilon \partial f(x_2).$$

ΠΑΡΑΡΤΗΜΑ ΙΙ

Υπολογισμός *MSE* για τις επαναληπτικές μεθόδους:

```
n<-nrow(data2)
y <- data2$Y
databa<-data2[,c(2,4,6:8,10:14,18,19,20,21,23,24,25,27,30,31,3
4,35,40,43,45,47,48,51,52)]
fitba <- lm(Y~.,data = databa)
xba <- model.matrix(Y ~ . - 1, data = databa)
coef_ba <- as.matrix(coef(fitba))
yhat_ba<-cbind(1, xba) %*% coef_ba
mse_ba<-sum((yhat_sb - y)^2)/n-25
mse_ba
[1] 45858644


datafa<-data2[,-c(2,7,8,10,11,13,18:21,23:25,33:35,40,43,45,48
,51,52)]
fitfa <- lm(Y~.,data = datafa)
xfa <- model.matrix(Y ~ . - 1, data = datafa)
coef_fa <- as.matrix(coef(fitfa))
yhat_fa<-cbind(1, xfa) %*% coef_fa
mse_fa<-sum((yhat_fa - y)^2)/n-30
mse_fa
 [1] 20902416


datasa<-data2[,-c(2,4,6:8,10:14,18:21,23:25,33:35,40,43,45,47,
48,51,52)]
fitsa <- lm(Y~.,data = datafa)
xsa <- model.matrix(Y ~ . - 1, data = datasa)
coef_sa <- as.matrix(coef(fitsa))
yhat_sa<-cbind(1, xsa) %*% coef_sa
mse_sa<-sum((yhat_sa - y)^2)/n-25
mse_sa
 [1] 21059620


databb<-data2[,-c(2,4:8,10:14,18:21,23:25,31,35:36,40,43,45:48
,50:53)]
fitbb <- lm(Y~.,data = databb)
xbb <- model.matrix(Y ~ . - 1, data = databb)
coef_bb <- as.matrix(coef(fitbb))
yhat_bb<-cbind(1, xbb) %*% coef_bb
mse_bb<-sum((yhat_bb - y)^2)/n-20
mse_bb
 [1] 22483112


datafb<-data2[,-c(2:5,7:11,13,17:26,28:31,33:36,38:49,51,52)]
fitfb <- lm(Y~.,data = datafb)
xfb <- model.matrix(Y ~ . - 1, data = datafb)
coef_fb <- as.matrix(coef(fitfb))
yhat_fb<-cbind(1, xfb) %*% coef_fb
mse_fb<-sum((yhat_fb - y)^2)/n-11
mse_fb
[1] 39064122
```

Επιλογή της παραμέτρου ποινής χρησιμοποιώντας τα κριτήρια πληροφορίας *AIC* και *BIC:*

```
n<-nrow(data2)
l<-seq(0,0.9, length.out=100 )
ridge2 <-lm.ridge( Y~.,data=data2, lambda=l )
df <-numeric(length(l))
AIC <-numeric(length(l))
BIC <-numeric(length(l))
p<-53
y<-scale(data2$Y, scale=F)
for (i in 1:length(l)){
Z <-scale(data2[,-1])
#(X'X + λI)^-1
A <-solve(t(Z)%*% Z + l[i]*diag(p))
#(X'X + λI)^-1(X'X)
H <-Z %*% A %*% t(Z)
#ŷ_R = Hy
yhat<-H%*%y
#y-yhat=sfalmata
SSE <-sum((yhat-y)^2)
df[i] <-sum(diag(H)) AIC[i]<-n*ln(SSE/n)+df[i]*2
BIC[i]<-n*ln(SSE/n)+df[i]*log(n)
          }
```

TR43		✓
TR48	✓	✓
TR49	✓	
TR51	✓	✓
TR52	✓	✓
TR53	✓	✓

TR54	✓	✓
TR55	✓	✓
TR56	✓	✓
Seasonality		✓
MA3	✓	✓
MA9	✓	✓
MA11	✓	
MA18	✓	✓
Cross1	✓	✓
Cross3	✓	
Cross6	✓	✓
Cross10	✓	✓
TR75	✓	✓
TV1	✓	✓
TV2	✓	
TV4	✓	✓
TV6	✓	
TVC1	✓	✓
TVC4	✓	
TVC5	✓	✓
TR76	✓	✓
Αριθμός μεταβλητών	44	39