Εθνικο Μετσοβιο Πολυτεχνειο

Σχολη Εφαρμοσμενων Μαθηματικων και Φυσικων Επιστημων

Διπλωματικη Εργασια

Επιλογή Μεταβλητών στο Πολλαπλό Γραμμικό Μοντέλο: **Ridge και LASSO** με χρήση της **R**

Φοιτητής:
Γιάννης Παπαδογιαννάκης

Επιβλέπων Καθηγητής:
Δημήτρης Φουσκάκης

john.papadogiannakis@gmail.com

Οκτώβριος 2016

Περιεχόμενα

Εισαγωγή

Στην παρούσα διπλωματική εργασία θα μελετήσουμε το πρόβλημα της επιλογής μεταβλητών στο πολλαπλό γραμμικό μοντέλο, δίνοντας έμφαση στις μεθόδους συρρίκνωσης των συντελεστών του μοντέλου *Ridge* και $LASSO$. Αρχικά, στο Κεφάλαιο 1, θα αναπτύξουμε το θεωρητικό υπόβαθρο του γραμμικού μοντέλου παλινδρόμησης πάνω στο οποίο θα βασιστούμε στη συνέχεια. Στη συνέχεια, στο Κεφάλαιο 2, θα εξετάσουμε το πρόβλημα της πολυσυγγραμικότητας, η οποία αποτελεί κύρια αιτία έρευνας των μεθόδων επιλογής μεταβλητών, εξετάζοντας τις παρενέργειες που προκαλεί πιθανή παρουσία της και παραθέτοντας μεθοδοδολογίες εντοπισμού της. Στο Κεφάλαιο 3 θα εμβαθύνουμε στο πρόβλημα επιλογής μεταβλητών στο πολλαπλό γραμμικό μοντέλο, αναλύοντας κάποιες από τις πιο διαδεδομένες μεθοδολογίες που εφαρμόζονται για την ανάδειξη των πιο καίριων μεταβλητών που οφείλουμε να συμπεριλάβουμε στο πολλαπλό γραμμικό μοντέλο. Έπειτα, στο Κεφάλαιο 4, θα εισαγάγουμε τη μεθοδολογία του Cross Validation, η οποία αποτελεί τη χρηστικότερη και πρότιμότερη μέθοδο αξιολόγησης της ικανότητας πρόβλεψης ενός μοντέλου. Καταλήγουμε έτσι στα Κεφάλαια 5 και 6 όπου θα εξετάσουμε ενδελεχώς τις μεθόδους συρρίκνωσης *Ridge* και $LASSO$, αναλύοντας τη πλήρη θεωρία που τις ορίζει και τις υλοποιεί, μελετώντας παράλληλα τις ομοιότητες και τις διαφορές τους. Στο τέλος κάθε κεφαλαίου, παραθέτουμε εφαρμογές στο στατιστικό πακέτο R, υλοποιώντας τα όσα έχουμε αναπτύξει θεωρητικά, για την καλύτερη κατανόησή τους.

1

Γραμμικό μοντέλο παλινδρόμησης

1.1 Θεωρητικό υπόβαθρο

Συχνά στη στατιστική καλούμαστε να μελετήσουμε και να μοντελοποιήσουμε προβλήματα, με σκοπό την πρόβλεψη μιας ποσοτικής μεταβλητής $\boldsymbol{Y}$ την οποία καλούμε μεταβλητή απόκρισης (response) ή εξαρτημένη (independent) μεταβλητή, συναρτήσει κάποιων άλλων τυχαίων μεταβλητών, έστω p το πλήθος, $\boldsymbol{X}_1, \boldsymbol{X}_2, ..., \boldsymbol{X}_p$ τις οποίες καλούμε επεξηγηματικές (explanatory) μεταβλητές ή ανεξάρτητες (independent) μεταβλητές ή συμμεταβλητές (covariates). Έστω ότι έχουμε $i = 1, 2, .., n$ παρατηρήσεις από τη μεταβλητή απόκρισης, $\boldsymbol{y} = (y_1, y_2, ..., y_n)^T$ και από τις p επεξηγηματικές μεταβλητές, $\boldsymbol{x}_j = (x_{1j}, x_{2j}, ..., x_{nj})^T$, $j = 1, 2, .., p$. Η πιο απλή αλλά παράλληλα ικανοποιητικότατη, για πολλές μορφές προβλημάτων, προσέγγιση είναι να θεωρήσουμε ότι οι παρατηρήσεις της μεταβλητής απόκρισης και των επεξηγηματικών μας μεταβλητών συνδέονται με τη πραγματική σχέση:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon_i \quad i = 1, 2, ..., n$$

όπου β_0 η σταθερά του μοντέλου, β_j $j = 1, 2, .., p$ οι συντελεστές των επεξηγηματικών μας μεταβλητών και ε_i το σφάλμα της i παρατήρησης. Η σταθερά και οι συντελεστές των συμμεταβλητών είναι άγνωστες παράμετροι ως προς τις οποίες η σχέση με την μεταβλητή απόκρισης είναι γραμμική. Ο συντελεστής β_j δείχνει την επίδραση της συμμεταβλητής $\boldsymbol{X}_j$ στην μεταβλητή απόκρισης αν η $\boldsymbol{X}_j$ αυξηθεί κατά μία μονάδα και οι υπόλοιπες επεξηγηματικές μεταβλητές παραμείνουν σταθερές. Η σταθερά β_0 αποτυπώνει την τιμή της μεταβλητής απόκρισης αν και εφόσον όλες οι επεξηγηματικές μεταβλητές ήταν ίσες με το 0. Τα σφάλματα υποθέτουμε ότι είναι τυχαίες iid (identically independently distributed) μεταβλητές που ακολουθούν την κανονική κατανομή με μέση τιμή

0 και τυπική απόκλιση σ_ε:

$$\varepsilon_i \overset{iid}{\sim} N\Big(0, \sigma_\varepsilon^2\Big) \quad i = 1, 2, ..., n.$$

Με αυτά κατά νου, η πραγματική σχέση που συνδέει τα Y_i με τα x_{ij} μπορεί να γραφεί ισοδύναμα και ως εξής:

$$E\big[Y_i|\boldsymbol{X}\big] = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip}$$

και ορίζεται να είναι το μοντέλο πολλαπλής γραμμικής παλινδρόμησης που συνδέει την μεταβλητή απόκρισης $\boldsymbol{Y}$ με τις επεξηγηματικές μεταβλητές $\boldsymbol{X}_1, .., \boldsymbol{X}_p$. Σκοπός μας είναι να εκτιμήσουμε τις άγνωστες παραμέτρους $\beta_0, \beta_1, ..., \beta_p$ ώστε να αποτυπώσουμε την επίδραση της κάθε επεξηγηματικής μεταβλητής στην μεταβλητή απόκρισης. Προτού όμως το πράξουμε αυτό, οφείλουμε να ελέγξουμε ότι πληρούνται οι προϋποθέσεις του πολλαπλού γραμμικού μοντέλου παλινδρόμησης, οι οποίες είναι οι ακόλουθες:

- **Γραμμικότητα**: Το μοντέλο είναι γραμμικό ως προς τις παραμέτρους και σωστά ορισμένο:

$$E\big[\boldsymbol{Y}|\boldsymbol{X}\big] = \beta_0 + \beta_1 \boldsymbol{x}_1 + \beta \boldsymbol{x}_2 + ... + \beta_p \boldsymbol{x}_p.$$

- **Κανονικότητα σφαλμάτων**: Υποθέτουμε ότι τα σφάλματα ακολουθούν την κανονική κατανομή με μέση τιμή μηδέν.

- **Ομοσκεδαστικότητα**: Η διασπορά των σφαλμάτων είναι ίδια για κάθε τιμή των επεξηγηματικών μεταβλητών.

$$Var(\varepsilon_i) = \sigma_\varepsilon^2 \quad \forall i = 1, 2, ..., n$$

- **Ασυσχέτιστα σφάλματα**: Τα σφάλματα είναι ασυσχέτιστα μεταξύ τους.

$$cov(\varepsilon_i, \varepsilon_j) = 0 \quad \forall i \neq j$$

Για να προχωρήσουμε, γράφουμε το μοντέλο παλινδρόμησης στη μορφή πινάκων:

$$\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n\Big(\boldsymbol{0}, \sigma_\varepsilon^2 \boldsymbol{I}_p\Big),$$

όπου $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_p)^T$ το διάνυσμα των παραμέτρων-συντελεστών, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, ..., \varepsilon_n)^T$ το διάνυσμα των τυχαίων σφαλμάτων και:

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ 1 & x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \end{pmatrix}$$

ο πίνακας σχεδιασμού. Το διάνυσμα των σφαλμάτων ακολουθεί την n-διάστατη κανονική κατανομή:

$$\boldsymbol{\varepsilon} \sim N_n\Big(\mathbf{0}, \sigma_\varepsilon^2 \boldsymbol{I}_p\Big)$$

με $E[\boldsymbol{\varepsilon}] = \mathbf{0}$ την αναμενόμενη τιμή των σφαλμάτων και $Var[\boldsymbol{\varepsilon}] = \sigma_\varepsilon^2 \boldsymbol{I}_p$ τον διαγώνιο πίνακα διασπορών-συνδιασπορών των σφαλμάτων. Παραθέτουμε τώρα ένα πολύ χρήσιμο θεώρημα: Αν για μια m-διάστατη τυχαία μεταβλητή $\boldsymbol{v}$ ισχύει ότι $\boldsymbol{v} \sim N_m(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ τότε ο γραμμικός μετασχηματισμός $\boldsymbol{u} = \boldsymbol{b} + \boldsymbol{A}\boldsymbol{v}$, όπου $\boldsymbol{A}$ πίνακας $n \times m$ και $\boldsymbol{b} \in \mathbb{R}^n$, κατανέμεται σύμφωνα με την n-διάστατη Κανονική κατανομή:

$$\boldsymbol{u} \sim N_n\Big(\boldsymbol{b} + \boldsymbol{A}\boldsymbol{\mu}, \boldsymbol{A}\boldsymbol{\Sigma}\boldsymbol{A}^T\Big).$$

Έτσι λοιπόν ο γραμμικός μετασχηματισμός $\boldsymbol{Y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ ακολουθεί την n-διάστατη Κανονική κατανομή:

$$\boldsymbol{Y}|\boldsymbol{X} \sim N_n\Big(\boldsymbol{X}\boldsymbol{\beta}, \sigma_\varepsilon^2 \boldsymbol{I}_p\Big).$$

Για την εκτίμηση των άγνωστων παραμέτρων-συντελεστών χρησιμοποιούμε την μέθοδο ελαχίστων τετραγώνων και συγκεκριμένα την παράσταση του αθροίσματος των τετραγώνων των υπολοίπων:

$$\begin{aligned} RSS(\boldsymbol{\beta}) &= (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) \\ &= \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \end{aligned}$$

την οποία ελαχιστοποιούμε επιλύοντας το πρόβλημα:

$$\underset{\boldsymbol{\beta} \in \mathbb{R}^{p+1}}{minimize} \;\; RSS(\boldsymbol{\beta}).$$

Υπολογίζουμε τις πρώτες και δεύτερες μερικές παραγώγους ως προς $\boldsymbol{\beta}$:

$$\frac{\partial RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}),$$
$$\frac{\partial^2 RSS(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}^T} = 2\boldsymbol{X}^T\boldsymbol{X}.$$

Υποθέτοντας ότι ο $\boldsymbol{X}$ είναι πλήρους βαθμού, δηλαδή $rank(\boldsymbol{X}) = p+1$, ο πίνακας $\boldsymbol{X}^T\boldsymbol{X}$ είναι θετικά ημιορισμένος, άρα θέτοντας την πρώτη παράγωγο ίση με μηδέν λαμβάνουμε την εκτιμήτρια ελαχίστων τετραγώνων του διανύσματος $\boldsymbol{\beta}$:

$$\hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Μάλιστα για το $\hat{\boldsymbol{\beta}}^{OLS}$ ισχύει ότι:

$$\begin{aligned} E\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] &= E\Big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\Big] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \cdot E[\boldsymbol{y}] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &= \boldsymbol{\beta}, \end{aligned}$$

δηλαδή η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\boldsymbol{\beta}}^{OLS}$ είναι αμερόληπτη εκτιμήτρια για το $\boldsymbol{\beta}$. Επίσης για την διασπορά της εκτιμήτριας ελαχίστων τετραγώνων έχουμε ότι:

$$\begin{aligned} Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] &= Var\Big[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}\Big] \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \cdot Var[\boldsymbol{y}] \cdot \Big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\Big)^T \\ &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T \cdot \sigma_\varepsilon^2\boldsymbol{I}_p \cdot \Big((\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\Big)^T \\ &= \sigma_\varepsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma_\varepsilon^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}. \end{aligned}$$

Υπό την υπόθεση ότι η μεταβλητή απόκρισης $\boldsymbol{Y}|\boldsymbol{X} = \boldsymbol{x}$ ακολουθεί την πολυδιάστατη κανονική και χρησιμοποιώντας το θεώρημα που αναφέραμε παραπάνω για την κατανομή που ακολουθεί ο γραμμικός μετασχηματισμός μιας μεταβλητής που ακολουθεί την πολυδιάστατη κανονική κατανομή, για $\hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$ καταλήγουμε στο ότι η κατανομή του διανύσματος της εκτιμήτριας ελαχίστων τετραγώνων είναι:

$$\hat{\boldsymbol{\beta}}^{OLS} \sim N_p(\boldsymbol{\beta}, \sigma_\varepsilon^2\Big(\boldsymbol{X}^T\boldsymbol{X})^{-1}\Big).$$

Η εκτιμήτρια $\hat{\boldsymbol{\beta}}^{OLS}$ έχει την μικρότερη διασπορά ανάμεσα σε όλες τις αμερόληπτες γραμμικές εκτιμήτριες του διανύσματος των παραμέτρων $\boldsymbol{\beta}$. Έχοντας την εκτιμήτρια ελαχίστων τετραγώνων μπορούμε να εκτιμήσουμε και την τιμή της μεταβλητή απόκρισης για τις δεδομένες τιμές των επεξηγηματικών μεταβλητών:

$$\hat{\boldsymbol{y}} = \boldsymbol{X}\hat{\boldsymbol{\beta}}$$

η οποία γράφεται και ακολούθως:

$$\begin{aligned}\hat{\boldsymbol{y}} &= \boldsymbol{X}\hat{\boldsymbol{\beta}} \\ &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{H}\boldsymbol{y},\end{aligned}$$

όπου $\boldsymbol{H}$ ο πίνακας προβολής (hat matrix) που αποτελεί πολύ σημαντικό εργαλείο στην ανάλυση παλινδρόμησης. Τα υπόλοιπα $\boldsymbol{r}$ αποτελούν την εκτίμηση των σφαλμάτων και ορίζονται να είναι:

$$\boldsymbol{r} = \boldsymbol{y} - \hat{\boldsymbol{y}}.$$

Σκόπιμο είναι τώρα να ορίσουμε τις ποσότητες Explained Sum of Squares:

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

και Total Sum of Squares:

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

. Θυμίζοντας ότι το Residual Sum of Squares είναι:

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

ισχύει ότι $TSS = RSS + ESS$. Μπορούμε τώρα να ορίσουμε τον συντελεστή προσδιορισμού R^2:

$$\begin{aligned}R^2 &= \frac{ESS}{TSS} \\ &= 1 - \frac{RSS}{TSS},\end{aligned}$$

για τον οποίο ισχύει $0 \leq R^2 \leq 1$. Ο συντελεστής προσδιορισμού αποτυπώνει το ποσοστό διασποράς το αριστερού μέλους του μοντέλου παλινδρόμησης το οποίο εξηγείται από το δεξί μέλος.

1.2 Πολλαπλή γραμμική παλινδρόμηση στην R

Τα δεδομένα μας

Το πρόβλημα το οποίο θα μελετήσουμε καθόλη τη διάρκεια αυτής της εργασίας, παραθέτοντας εφαρμογές της θεωρίας που αναπτύσσουμε στην R, αφορά την ασθένεια του διαβήτη. Θα χρησιμοποιήσουμε τα δεδομένα που χρησιμοποίησαν οι Efron, Hastie, Johnstone και Tibshirani το 2004 στη δημοσίευσή τους, Least Angle Regression. Τα δεδομένα αυτά αποτελούνται από $n = 442$ παρατηρήσεις, $p = 10$ επεξηγηματικές μεταβλητές και 1 μεταβλητή απόκρισης. Συγκεκριμένα, από 442 ασθενείς που πάσχουν από διαβήτη έχουμε την ηλικία τους AGE, το φύλλο τους SEX, τον δείκτη μάζας σώματός τους BMI (Body Mass Index), την μέση πίεση του αίματός τους BP (Blood Pressure), και μετρήσεις σε 6 διαφορετικά συστατικά του ορού του αίματός τους $S1$, $S2$, $S3$, $S4$, $S5$, $S6$. Αυτές οι 10 θα αποτελέσουν τις επεξηγηματικές μας μεταβλητές και είναι όλες ποσοτικές πλήν του φύλλου SEX που είναι κατηγορική και αποτελείται προφανώς από 2 κατηγορίες. Η τιμή 0 αντιστοιχεί στις ασθενείς γυναικείου φύλλου ενώ η τιμή 1 στους ασθενείς ανδρικού φύλλου. Επίσης για κάθε ασθενή έχουμε μια ποσοτική μέτρηση της εξέλιξης του διαβήτη, έναν χρόνο μετά τις μετρήσεις των 10 επεξηγηματικών μεταβλητών, η οποία θα αποτελέσει την μεταβλητή απόκρισής μας. Όσο μεγαλύτερη είναι η τιμή της μέτρησης, τόσο περισσότερο έχει εξελιχθεί η νόσος.

```
1 diab<-read.table("http://www4.stat.ncsu.edu/~boos/var.
    select/diabetes.tab.txt",header=T)
2 X.diab<-as.data.frame(diab[,-ncol(diab)])
3 X.diab[,2]<-replace(X.diab[,2], X.diab[,2]==1, 0) #women
     --> zeros
4 X.diab[,2]<-replace(X.diab[,2], X.diab[,2]==2, 1) #men
    -->ones
5 Y.diab<-as.vector(diab[,ncol(diab)])
6 diab.full.data<-data.frame(Y.diab,X.diab)
```

Κώδικας 1.1: Φορτώνουμε τα δεδομένα μας

Στον Κώδικα 1.1 φορτώνουμε τα δεδομένα μας με την εντολή $read.table()$, στην οποία τοποθετούμε την ηλεκτρονική διεύθυνση στην οποία βρίσκεται το αρχείο με τα δεδομένα μας. Εκχωρούμε τις τιμές των επεξηγηματικών μας μεταβλητών $X.diab$ ως πλαίσιο δεδομένων $as.data.frame()$, στο οποίο αλλάζουμε τα στοιχεία της δεύτερης στήλης που αντιστοιχούν στη μεταβλητή SEX ούτως ώστε

να έχουν τις τιμές 0 για τις γυναίκες (αρχικά ήταν 1) και 1 για τους άνδρες (αρχικά ήταν 2). Εκχωρούμε επίσης τις τιμές της μεταβλητής απόκρισης $Y.diab$ τις οποίες τις κάνουμε διάνυσμα με την εντολή $as.vector()$. Τα δεδομένα μας έχουν την μορφή:

```
> head(diab.full.data)
  Y.diab AGE SEX  BMI  BP  S1    S2 S3 S4     S5 S6
1    151  59   1 32.1 101 157  93.2 38  4 4.8598 87
2     75  48   0 21.6  87 183 103.2 70  3 3.8918 69
3    141  72   1 30.5  93 156  93.6 41  4 4.6728 85
4    206  24   0 25.3  84 198 131.4 40  5 4.8903 89
5    135  50   0 23.0 101 192 125.4 52  4 4.2905 80
6     97  23   0 22.6  89 139  64.8 61  2 4.1897 68
```

Εικόνα 1.1: Πρώτες 6 παρατηρήσεις των δεδομένων μας με την εντολή $head()$.

Σκοπός μας είναι να εντοπίσουμε μοντέλα που θα αποτελέσουν όσο το δυνατόν ακριβέστερα εφαλτήρια πρόβλεψης της εξέλιξης της ασθένειας του διαβήτη, για μελλοντικούς ασθενείς.

Παλινδρόμηση και έλεγχος προϋποθέσεων

```
1 reg.diab<-lm(Y.diab~.,data=X.diab)
```

Κώδικας 1.2: Εκτέλεση παλινδρόμησης με την εντολή $lm()$.

Στον Κώδικα 1.2 εκτελούμε την πολλαπλή γραμμική παλινδρόμηση όλων των τιμών των επεξηγηματικών μας μεταβλητών με τις τιμές της μεταβλητής απόκρισης και εκχωρούμε τα αποτελέσματά της στο $reg.diab$. Προτού παραθέσουμε όμως την περίληψη των αποτελεσμάτων θα ελέγξουμε αν ισχύουν οι προϋποθέσεις της πολλαπλής γραμμικής παλινδρόμησης που αναφέραμε στην Παράγραφο 1.1.

Γραμμικότητα

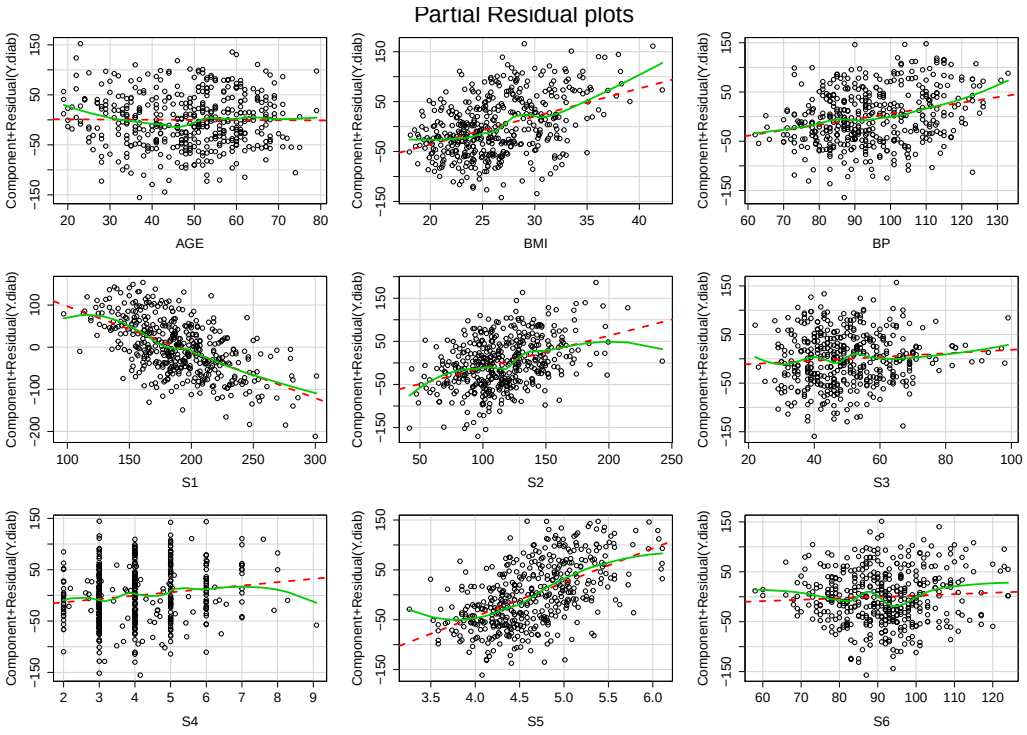
Για να ελέγξουμε την προϋπόθεση της γραμμικότητας του μοντέλου, κατασκευάζουμε τα διαγράμματα διασποράς, για κάθε j συμμεταβλητή, των μερικών υπολοίπων:

$$P_j = \hat{\beta}_j^{OLS} x_{ij} + r_i \quad i = 1, 2, .., n \ ,$$

με τις παρατηρήσεις της αντίστοιχης επεξηγηματικής μεταβλητής. Τοιουτοτρόπως ελέγχουμε αν η δεσμευμένη μέση τιμή της μεταβλητής απόκρισης συνδέεται γραμμικά με την κάθε επεξηγηματική μεταβλητή, δεδομένου του ότι οι υπόλοιπες επεξηγηματικές μεταβλητές συνδέονται και αυτές γραμμικά με την μεταβλητή απόκρισης. Θέλουμε λοιπόν να παρατηρήσουμε γραμμικές σχέσεις σε όλα τα διαγράμματα διασποράς τα οποία κατασκευάζουμε στην R με τον Κώδικα 1.3 και φαίνονται στην Εικόνα 1.2. Να σημειώσουμε εδώ ότι ο εν λόγω έλεγχος δεν έχει νόημα για κατηγορικές επεξηγηματικές μεταβλητές, οπότε παραλείπουμε τη συμμεταβλητή SEX.

```
install.packages("car")
library(car)
crPlots(reg.diab,terms=~.-SEX,main="Partial residual
   plots")
```

Κώδικας 1.3: Κατασκευάζουμε τα διαγράμματα διασποράς, της Εικόνας 1.2, των μερικών υπολοίπων κάθε συμμεταβλητής με τις αντίστοιχες παρατηρήσεις της.



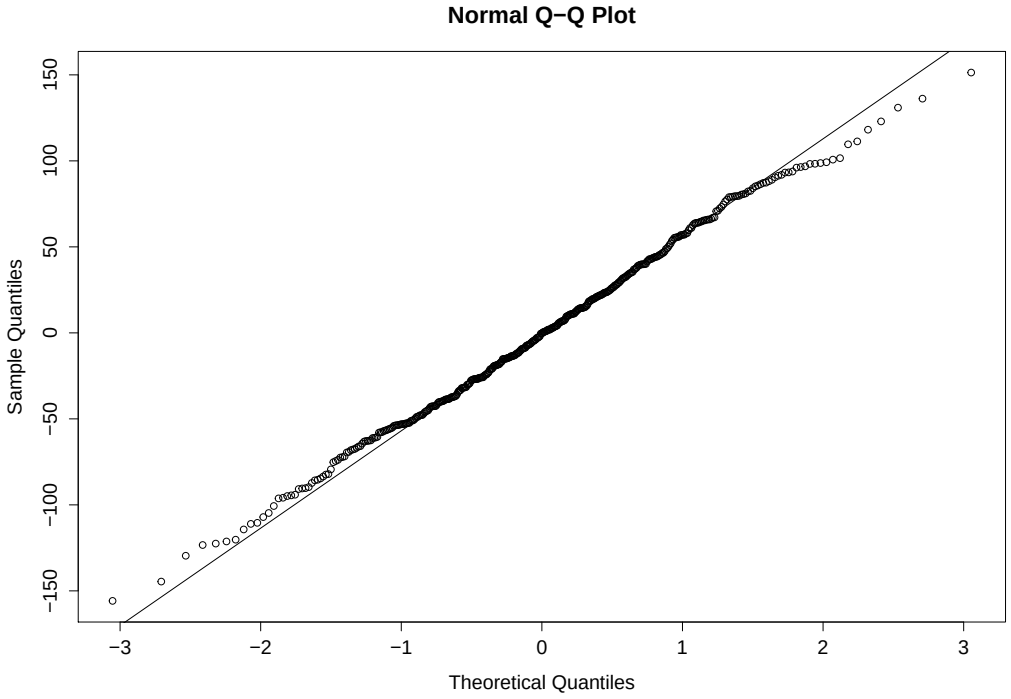
Εικόνα 1.2: Διαγράμματα διασποράς των μερικών υπολοίπων κάθε συμμεταβλητής με τις παρατηρήσεις της.

Παρατηρούμε να φαίνεται λογική η υπόθεση της γραμμικότητας για τα δεδομένα όλων των ποσοτικών επεξηγηματικών μεταβλητών, πλην αυτών της $S4$. Πιθανόν κάποιος μετασχηματισμός της ή μεγαλύτερο πλήθος παρατηρήσεων, να μας έδινε πιο επιθυμητά αποτελέσματα.

Κανονικότητα σφαλμάτων

```
1 install.packages("car")
2 library(car)
3 qqnorm(residuals(reg.diab))
4 qqline(residuals(reg.diab))
```

Κώδικας 1.4: Ελεγχουμε αν τα σφάλματα των παρατηρήσεων ακολουθούν κανονική κατανομή με τις εντολές $qqnorm()$ και $qqline()$.



Εικόνα 1.3: Έλεγχος της κανονικότητας των σφαλμάτων με $QQ-plot$.

Στον Κώδικα 1.4 εγκαθιστούμε και φορτώνουμε το πακέτο *car* για να χρησιμοποιήσουμε την εντολή $qqnorm()$, που μας παρέχει ένα διάγραμμα ποσοστημορίων των υπολοίπων (εκτίμηση των σφαλμάτων) του δείγματος, και την εντολή $qqline()$ που μας δείχνει σε μια ευθεία γραμμή τις θεωρητικές τιμές των

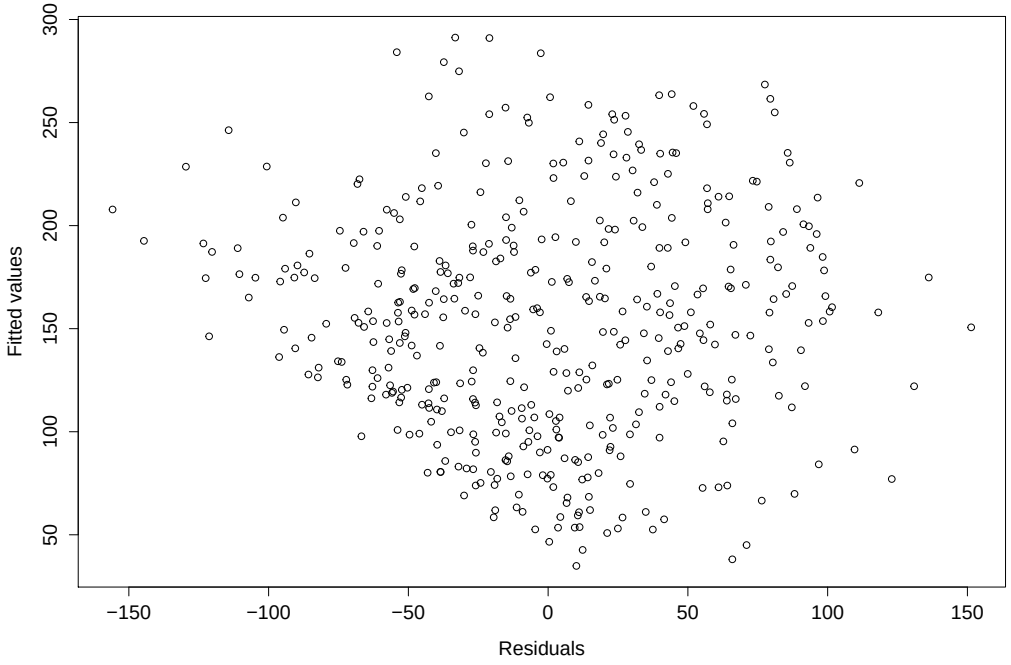
ποσοστημορίων της κανονικής κατανομής. Και στις δυο εντολές εισάγουμε τα σφάλματα $residuals()$ της παλινδρόμησης που εκτελέσαμε. Όσο λιγότερο αποκλίνουν οι τιμές των ποσοστημορίων των υπολοίπων από τις θεωρητικές (την γραμμή) τόσο πιο πεποισμένοι είμαστε ότι τα υπόλοιπα ακολουθούν την κανονική κατανομή. Στην Εικόνα 1.3 βλέπουμε ότι κατα κύριο λόγο οι τιμές του δείγματος συμπίπτουν με τις θεωρητικές, οπότε μπορούμε να αποφανθούμε ότι η προϋπόθεση ότι τα σφάλματα ακολουθούν την κανονική κατανομή ισχύει.

Ομοσκεδαστικότητα

```
plot(residuals(reg.diab),predict(reg.diab), xlab="
   Residuals", ylab="Fitted values")
```

Κώδικας 1.5: Ελέγχουμε αν η διαπορά των σφαλμάτων είναι ίδια για όλα τα σφάλματα κατασκευάζοντας το διάγραμμα διασποράς της Εικόνας 1.4.

Επόμενο βήμα είναι να ελέγξουμε την προϋπόθεση της ομοσκεδαστικότητας,



Εικόνα 1.4: Διάγραμμα διασποράς των υπολοίπων με τις εκτιμήσεις της μεταβλητής απόκρισης για έλεγχο ομοσκεδαστικότητας.

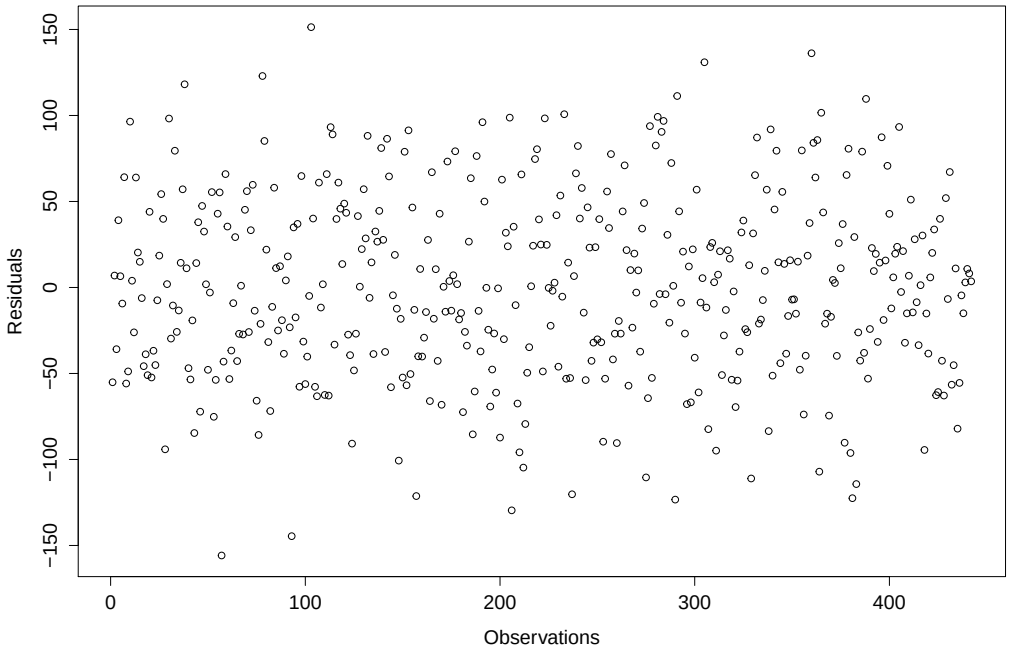
δηλαδή ότι η διασπορά των σφαλμάτων είναι ίδια για κάθε τιμή των επεξηγηματικών μεταβλητών. Θα δημιουργήσουμε λοιπόν το διάγραμμα διασποράς των

υπολοίπων $residuals()$ με τις εκτιμήσεις της μεταβλητής απόκρισης $predict()$ για τις δεδομένες τιμές των επεξηγηματικών μεταβλητών, με την εντολή $plot()$, Κώδικας 1.5. Θα θέλαμε να μην παρατηρήσουμε κάποια συστηματική συμπεριφορά, ώστε να ισχύει η προϋπόθεση. Παρατηρούμε στην Εικόνα 1.4 πως δεν διακρίνεται κάποια συστηματική συμπεριφορά ανάμεσα στα υπόλοιπα και στις εκτιμήσεις της μεταβλητής απόκρισης της παλινδρόμησης που εκτελέσαμε. Άρα μπορούμε να αποφανθούμε ότι η υπόθεση της ομοσκεδαστικότητας ισχύει, οπότε η διασπορά των σφαλμάτων είναι κοινή για κάθε παρατήρηση.

Ασυσχέτιστα σφάλματα

```
plot(1:nrow(X.diab),reg.diab$res,xlab="Observations",
   ylab="Residuals")
```

Κώδικας 1.6: Κατασκευάζουμε το διάγραμμα διασποράς της Εικόνας 1.5.



Εικόνα 1.5: Διάγραμμα διασποράς των υπολοίπων με τον αντίστοιχο αριθμό παρατήρησης για να ελέγξουμε αν τα σφάλματα είναι ανεξάρτητα.

Η τελευταία προϋπόθεση που μας απομένει να ελέγξουμε είναι το αν τα σφάλματα είναι ασυσχέτιστα μεταξύ τους, δηλαδή αν $cov(\varepsilon_i, \varepsilon_j) = 0 \; \forall i \neq j$. Μπορούμε

να ελέγξουμε λοιπόν αν είναι ανεξάρτητα. Σε περίπτωση που είναι ανεξάρτητα, τότε θα είναι και ασυσχέτιστα. Αν είναι εξαρτημένα όμως δεν μπορούμε να αποφανθούμε. Κατασκευάζουμε λοιπόν, με τον Κώδικα 1.6, ένα διάγραμμα διασποράς των υπολοίπων με τον αντίστοιχο αριθμό παρατήρησης, ελπίζοντας να μην παρατηρήσουμε κάποια συστηματική συμπεριφορά. Στην Εικόνα 1.5 φαίνεται ότι δεν υπάρχει κάποια συστηματική συμπεριφορά μεταξύ των σφαλμάτων, άρα μπορούμε να καταλήξουμε στο συμπέρασμα ότι τα σφάλματα είναι ανεξάρτητα μεταξύ τους, άρα και ασυσχέτιστα.

Αποτελέσματα παλινδρόμησης

```
> summary(reg.diab)

Call:
lm(formula = Y.diab ~ ., data = X.diab)

Residuals:
     Min       1Q   Median       3Q      Max
-155.827  -38.536   -0.228   37.806  151.353

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -357.42679   67.05807  -5.330 1.59e-07 ***
AGE           -0.03636    0.21704  -0.168 0.867031
SEX          -22.85965    5.83582  -3.917 0.000104 ***
BMI            5.60296    0.71711   7.813 4.30e-14 ***
BP             1.11681    0.22524   4.958 1.02e-06 ***
S1            -1.09000    0.57333  -1.901 0.057948 .
S2             0.74645    0.53083   1.406 0.160390
S3             0.37200    0.78246   0.475 0.634723
S4             6.53383    5.95864   1.097 0.273459
S5            68.48312   15.66972   4.370 1.56e-05 ***
S6             0.28012    0.27331   1.025 0.305990
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 54.15 on 431 degrees of freedom
Multiple R-squared:  0.5177, Adjusted R-squared:  0.5066
F-statistic: 46.27 on 10 and 431 DF,  p-value: < 2.2e-16
```

Εικόνα 1.6: Περίληψη του πλήρους πολλαπλού γραμμικού μοντέλου παλινδρόμησης που εξετάζουμε. Οι αστερίσκοι δηλώνουν τις στατιστικά σημαντικές επεξηγηματικές μεταβλητές.

Αφού λοιπόν ελέγξαμε τις προϋποθέσεις του πολλαπλού γραμμικού μοντέλου

παλινδρόμησης που εξετάζουμε και καταλήξαμε στο ότι ισχύουν, μπορούμε πλέον με ασφάλεια να παρουσιάσουμε τα αποτελέσματά του. Αυτό θα γίνει με χρήση της εντολής $summary()$ που δέχεται ως όρισμα το αντικείμενο της παλινδρόμησης $reg.diab$ που προσαρμόσαμε στον Κώδικα 1.2. Παρατηρούμε στην Εικόνα 1.6 πως στατιστικά σημαντικές σε επίπεδο σημαντικότητας 5% είναι οι επεξηγηματικές μεταβλητές SEX, BMI, BP και $S5$. Οι εκτιμήσεις τους φαίνονται στην στήλη $Estimate$. Για τις ποσοτικές συμμεταβλητές, δηλαδή όλες πλην του φύλου του ασθενή, δηλώνουν πόσο μεταβάλεται η εξέλιξη του διαβήτη όταν αυξηθεί κατά μια μονάδα η αντίστοιχη συμμεταβλητή και παραμείνουν οι υπόλοιπες σταθερές. Για την κατηγορική SEX δηλώνει πόσο μεταβάλλεται η εξέλιξη της νόσου για την κατηγορία 1, το ανδρικό φύλλο. Τις μεγαλύτερες επιδράσεις τις έχουν οι επεξηγηματικές μεταβλητές SEX, με -22.9 επίδραση στην εξέλιξη της ασθένειας για το ανδρικό φύλλο και η μέτρηση του ορού του αίματος $S5$, με $+68.5$ επίδραση στην εξέλιξη του διαβήτη για κάθε αυξημένη μονάδα της μέτρησης. Άρα ένα πρώτο συμπέρασμα που μπορούμε να εξάγουμε είναι ότι αν είσαι γυναίκα και έχεις διαβήτη, η ασθένεια εξελίσσεται πιο γρήγορα απότι αν είσαι άντρας. Η συμμεταβλητή $S1$ παρόλο που δεν εμφανίζεται να είναι στατιστικά σημαντική, έχει $p - value = 0.058$, το οποίο σημαίνει ότι σε επίπεδο σημαντικότητας 6% θα εμφανιζόταν ως στατιστικά σημαντική, με μικρή επίδραση παρόλα αυτά. Να αναφέρουμε επίσης ότι η σταθερά εμφανίζεται να είναι στατιστικά σημαντική, όμως δεν χρήζει αξιόλογης προσέγγισης καθώς η ερμηνεία της, ότι αυτή είναι η εξέλιξη της νόσου όταν όλες οι επεξηγηματικές είναι μηδενικές, δεν είναι λογική. Ο συντελεστής προσδιορισμού R^2 είναι ίσος με 51.77%, το οποίο σημαίνει ότι περίπου το 52% της διασποράς των δεδομένων της μεταβλητής απόκρισης εξηγείται από τη διασπορά του δεξιού μέλους του μοντέλου παλινδρόμησης, ποσοστό όχι τόσο ικανοποιητικό αφού ένα 48% φαίνεται να παραμένει ανεξήγητο. Με όλα αυτά κατά νου, θα προχωρήσουμε στην μελέτη της πιθανής ύπαρξης πολυσυγγραμικότητας στο μοντέλο, στο επόμενο Κεφάλαιο.

2

Πολυσυγγραμικότητα

Τα μοντέλα πολλαπλής γραμμικής παλινδρόμησης που εξετάζουμε έχουν ως απώτερο σκοπό τη πρόβλεψη της μεταβλητής απόκρισης $\boldsymbol{Y}$ μέσω διάφορων επεξηγηματικών μεταβλητών $\boldsymbol{X}_j$, μελετώντας τη γραμμική σχέση που τις συνδέει. Μπορούμε, να αντιληφθούμε διαισθητικά ότι η ύπαρξη συμμεταβλητών, οι οποίες είναι συσχετισμένες μεταξύ τους, θα μπορούσε να αποφέρει παραπλανητικά αποτελέσματα. Αρχικά λοιπόν μας ενδιαφέρει να μελετήσουμε τη πιθανή γραμμική σχέση ανάμεσα σε ορισμένες ή και όλες τις επεξηγηματικές μεταβλητές του μοντέλου, καθώς και τις επιπτώσεις της.

Τέλεια πολυσυγγραμικότητα ονομάζουμε το φαινόμενο κατά το οποίο δύο ή και περισσότερες επεξηγηματικές μεταβλητές του μοντέλου σχετίζονται γραμμικά, δηλαδή την ύπαρξη τουλάχιστον δύο σταθερών $a_i \in \mathbb{R}$, $i = 1, ..p$ μη μηδενικών, ούτως ώστε να ισχύει:

$$a_1\boldsymbol{X}_1 + a_2\boldsymbol{X}_2 + ... + a_p\boldsymbol{X}_p = 0.$$

Αντιλαμβανόμαστε ότι είναι πιθανό να συνυπάρχουν πολλές τέτοιες γραμμικές σχέσεις για διάφορους συνδυασμούς συμμεταβλητών σε ένα μοντέλο. Για παράδειγμα αν το μοντέλο παλινδρόμησης που εξετάζουμε είναι το:

$$\boldsymbol{Y} = \beta_0 + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \beta_3\boldsymbol{x}_3 + \beta_4\boldsymbol{x}_4 + \boldsymbol{\varepsilon}, \qquad \boldsymbol{\varepsilon} \sim N_n\Big(\boldsymbol{0}, \sigma_\varepsilon^2\boldsymbol{I}_p\Big),$$

μπορεί να ισχύει ταυτόχρονα και ότι $\boldsymbol{x}_1 = 5 + 4\boldsymbol{x}_2$ καθώς και $\boldsymbol{x}_3 = 2\boldsymbol{x}_4 - 1$.

Στην πράξη είναι δύσκολο εως αδύνατο να παρατηρηθούν τέλειες γραμμικές σχέσεις ανάμεσα στις εξαρτημένες μεταβλητές. Σφάλματα, υπήρχαν, υπάρχουν και θα συνεχίσουν να υπάρχουν σε όλες τις μετρήσεις. Για αυτόν ακριβώς το λόγο εμάς μας ενδιαφέρει η μελέτη των στατιστικά σημαντικών γραμμικών

σχέσεων που μπορεί να υπάρχουν, η ύπαρξη των οποίων θα αναφέρεται ως πολυσυγγραμικότητα.

Κατά την ύπαρξη τέλειας πολυσυγγραμικότητας οι εκτιμητές ελαχίστων τετραγώνων δεν υπάρχουν, όπως εξηγούμε παρακάτω. Όταν έχουμε στατιστικά σημαντικές γραμμικές σχέσεις, δηλαδή πολυσυγγραμικότητα, οι εκτιμήτριες ναι μεν θα υπάρχουν αλλά:

- Θα έχουν υψηλά τυπικά σφάλματα.
- Σημαντικές επιδράσεις θα εμφανίζονται ως ασήμαντες.
- Θα υπάρχει αλλοίωση των επιδράσεων σε βαθμό που η ερμηνεία τους μπορεί να εξάγει μέχρι και αντίθετα συμπεράσματα από τα πραγματικά.
- Οι επιδράσεις των συμμεταβλητών που σχετίζονται γραμμικά δεν θα μπορούν να διαχωριστούν.
- Θα υπάρχει εν γένει αστάθεια των εκτιμητριών.

Η ύπαρξη *τέλειας πολυσυγγραμικότητας* οδηγεί σε αδυναμία εκτίμησης των παραμέτρων του μοντέλου. Όταν μια συμμεταβλητή $\boldsymbol{X}_j$ είναι τέλειος γραμμικός συνδυασμός των υπολοίπων επεξηγηματικών μεταβλητών οι εκτιμητές ελαχίστων τετραγώνων δεν υπάρχουν, διότι ο πίνακας $\boldsymbol{X}^T\boldsymbol{X}$ είναι μη αντιστρέψιμος.

Όταν δύο επεξηγηματικές μεταβλητές είναι υψηλά συσχετισμένες, ουσιαστικά αποτυπώνουν παρόμοια πληροφορία. Εφόσον γνωρίζουμε την τιμή της μιας μπορούμε να προβλέψουμε την τιμή της άλλης. Έτσι λοιπόν, το να συμπεριλάβουμε και τις δύο αυτές μεταβλητές στο μοντέλο μας, δεν προσδίδει κάποια επιπλέον πληροφορία. Υπό αυτό το σκεπτικό, ίδιο είναι και το πρόβλημα που προκύπτει κατά την ύπαρξη πολυσυγγραμικότητας. Ας θεωρήσουμε το μοντέλο παλινδρόμησης που αναφέραμε στην αρχή του κεφαλαίου:

$$\boldsymbol{Y} = \beta_0 + \beta_1\boldsymbol{x}_1 + \beta_2\boldsymbol{x}_2 + \beta_3\boldsymbol{x}_3 + \beta_4\boldsymbol{x}_4 + \boldsymbol{\varepsilon}.$$

Αφότου εκτιμούσαμε τις παραμέτρους β_i θα μπορούσαμε να αποφανθούμε για την επίδραση της κάθε επεξηγηματικής μεταβλητής στη μεταβλητή απόκρισης. Έστω όμως ότι ισχύει $\boldsymbol{x}_1 = 5 + 4\boldsymbol{x}_2$ καθώς και $\boldsymbol{x}_3 = 2\boldsymbol{x}_4 - 1$. Τότε το μοντέλο μας μετασχηματίζεται ως εξής:

$$\boldsymbol{Y} = (\beta_0 + 5\beta_1 - \beta_3) + (4\beta_1 + \beta_2)\boldsymbol{x}_2 + (2\beta_3 + \beta_4)\boldsymbol{x}_4 + \boldsymbol{\varepsilon}$$

Με τα καινούργια δεδομένα, παρατηρούμε δύο πράγματα. Πρώτον, οι επεξηγηματικές μεταβλητές $\boldsymbol{X}_1$ και $\boldsymbol{X}_3$ πλέον δεν υπάρχουν στο μοντέλο, αφού η

πληροφορία που είχαν αποτυπώνεται μέσω των συμμεταβλητών $\boldsymbol{X}_2$ και $\boldsymbol{X}_4$ αντιστοίχως. Δεύτερον, οι επιδράσεις των επεξηγηματικών μεταβλητών $\boldsymbol{X}_2$ και $\boldsymbol{X}_4$ στη μεταβλητή απόκρισης $\boldsymbol{Y}$, δεν είναι πλέον οι παράμετροί τους, β_2 και β_4. Η επίδραση της $\boldsymbol{X}_2$ είναι $4\beta_1 + \beta_2$ ενώ η επίδραση της $\boldsymbol{X}_4$ είναι $2\beta_3 + \beta_4$. Αντιλαμβανόμαστε λοιπόν, ότι αν εφαρμόζαμε το πρώτο μοντέλο παλινδρόμησης, ενώ παράλληλα ίσχυαν οι γραμμικές σχέσεις που αναφέραμε, οι εκτιμήσεις που θα λαμβάναμε για τις επιδράσεις των επεξηγηματικών μεταβλητών θα ήταν παντελώς αποπροσανατολιστικές.

2.1 Μέθοδοι Εντοπισμού

Παρακάτω θα αναλύσουμε τα διαγνωστικά ελέγχου που μπορούμε να εφαρμόσουμε ώστε να εντοπίσουμε την ύπαρξη πολυσυγγραμικότητας.

2.1.1 Συντελεστής συσχέτισης

Ο συντελεστής συσχέτισης (pearson correlation) αποτελεί ένα μέτρο γραμμικής συσχέτισης ανάμεσα σε δύο μεταβλητές, έστω X και Y, και ορίζεται ως εξής για τον πληθυσμό:

$$\rho_{X,Y} = \frac{cov(X,Y)}{\sigma_X \sigma_Y}$$

Για ένα δείγμα n παρατηρήσεων, υπολογίζεται από τα δειγματικά ανάλογα της συνδιασποράς δυο μεταβλητών και των τυπικών αποκλίσεων κάθε μεταβλητής. Ο συντελεστής συσχέτισης μπορεί να λάβει τιμές από -1 εως 1. Τιμές κοντά στο 1 και στο -1 υποδηλώνουν ότι υπάρχει στατιστικά σημαντική γραμμική σχέση ανάμεσα στις μεταβλητές X και Y (απολύτως γραμμική για τιμές ακριβώς 1 ή -1), ενώ τιμές κοντά στο 0 δείχνουν ότι δεν υπάρχει στατιστικά σημαντική γραμμική συσχέτιση μεταξύ των μεταβλητών.

Ο συντελεστής συσχέτισης φαινομενικά αποτελεί ένα αρκετά καλό διαγνωστικό ελέγχου για τον εντοπισμό πολυσυγγραμικότητας. Παρόλα αυτά, υπάρχει ένας πολύ σημαντικός, προφανής, περιορισμός. Εξετάζει τη γραμμική συσχέτιση μόνο μεταξύ δύο μεταβλητών, οπότε αποτυγχάνει στον εντοπισμό κάποιου γραμμικού συνδυασμού που περιλαμβάνει περισσότερες από δυο επεξηγηματικές μεταβλητές. Μπορεί για παράδειγμα να βρούμε μια τιμή του συντελεστή συσχέτισης κοντά στο 0, ανάμεσα σε δύο συμμεταβλητές και στην πραγματικότητα να υπάρχει στατιστικά σημαντικός γραμμικός συνδυασμός ανάμεσά τους, που να περιλαμβάνει και άλλες συμμεταβλητές.

2.1.2 Παράγοντας πληθωρισμού διασποράς

Ο παράγοντας πληθωρισμού διασποράς (Variance Inflation Factor) της εκτιμήτριας $\hat{\beta}_j$ ορίζεται ως εξής:

$$VIF_j = \frac{1}{1 - R_j^2}$$

όπου R_j ο συντελεστής προσδιορισμού του μοντέλου παλινδρόμησης με μεταβλητή απόκρισης την $\boldsymbol{X}_j$ και επεξηγηματικές μεταβλητές τις υπόλοιπες $p-1$ το πλήθος συμμεταβλητές του αρχικού μας μοντέλου, δηλαδή εξάγεται από το μοντέλο παλινδρόμησης:

$$\boldsymbol{X}_j = \delta_0 + \delta_1 \boldsymbol{x}_1 + .. + \delta_{j-1}\boldsymbol{x}_{j-1} + \delta_{j+1}\boldsymbol{x}_{j+1} + ... + \delta_p \boldsymbol{x}_p + \boldsymbol{u}, \quad \boldsymbol{u} \sim N_n\Big(\boldsymbol{0}, \sigma_u^2 \boldsymbol{I}_{p-1}\Big).$$

Να θυμηθούμε ότι ο πίνακας διασπορών συνδιασπορών των εκτιμητριών $\hat{\beta}_j^{OLS}$ για $j = 0, 1, ..., p$ είναι:

$$Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] = \sigma_\varepsilon^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}.$$

Άρα οι διασπορές των $\hat{\beta}_j^{OLS}$ είναι τα στοιχεία της διαγωνίου του πίνακα $Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big]$. Συγκεκριμένα έχουμε:

$$\begin{aligned} Var\Big[\hat{\beta}_j\Big] &= \sigma_\varepsilon^2 (\boldsymbol{X}^T \boldsymbol{X})^{-1}_{jj} \\ &= \frac{\sigma_\varepsilon^2}{(n-1)\hat{\sigma}^2_{\boldsymbol{x}_j}} \cdot \frac{1}{1 - R_j^2} \\ &= \frac{\sigma_\varepsilon^2}{(n-1)\hat{\sigma}^2_{\boldsymbol{x}_j}} \cdot VIF_j, \end{aligned}$$

όπου $\hat{\sigma}^2_{\boldsymbol{x}_j}$ η διασπορά των δεδομένων τιμών της συμμεταβλητής $\boldsymbol{X}_j$. Ουσιαστικά η τετραγωνική ρίζα του παράγοντα πληθωρισμού διασποράς μας δείχνει πόσο μεγαλύτερο είναι το τυπικό σφάλμα της εκτιμήτριας $\hat{\beta}_j^{OLS}$ συγκριτικά με την τιμή που θα είχε, το τυπικό σφάλμα, αν και εφόσον η συμμεταβλητή $\boldsymbol{X}_j$ ήταν γραμμικώς ασυσχέτιστη με τις υπόλοιπες επεξηγηματικές μεταβλητές.

Μπορούμε να αντιστοιχίσουμε τις τιμές του παράγοντα πληθωρισμού πληθωρισμού διασποράς με την ύπαρξη πολυσυγγραμικότητας ακολούθως:

- $VIF_j = 1 \implies R_j^2 = 0$: Δεν υπάρχει πολυσυγγραμικότητα.
- $2 \leq VIF_j \leq 5 \implies 0.5 \leq R_j^2 \leq 0.8$: Σημάδια πολυσυγγραμικότητας.

- $5 < VIF_j < 10 \implies 0.8 < R_j^2 < 0.9$: Πιθανή πολυσυγγραμικότητα.
- $10 \leq VIF_j \implies 0.9 \leq R_j^2$: Σχεδόν σίγουρη πολυσυγγραμικότητα.

Ο παράγοντας πληθωρισμού διασποράς είναι ένα καλό μέτρο για τον γενικότερο εντοπισμό ύπαρξης πολυσυγγραμικότητας. Η αδυναμία του έγκειται στο ότι δεν μπορεί να προσδιορίσει επακριβώς τις συνυπάρχουσες γραμμικές σχέσεις των επεξηγηματικών μεταβλητών.

2.1.3 Δείκτες κατάστασης

Ένα πολύ σημαντικό μέτρο για τον εντοπισμό της ύπαρξης γραμμικών σχέσεων ανάμεσα στις επεξηγηματικές μεταβλητές του μοντέλου που μελετάμε, είναι οι δείκτες κατάστασης (Condition Indices) του πίνακα σχεδιασμού $\boldsymbol{X}$. Αποτελούν ένα μέτρο ένδειξης του αν ο $\boldsymbol{X}$ είναι ill-conditioned, δηλαδή αν «πάσχει» από φαινόμενα πολυσυγγραμικότητας.

Οι δείκτες κατάστασης ορίζονται ακολούθως:

$$CI_j = \sqrt{\frac{max(d_j^2)}{d_j^2}} = \frac{max(d_j)}{d_j}$$

όπου d_j^2 οι ιδιοτιμές του πίνακα $\boldsymbol{X}^T\boldsymbol{X}$ και d_j οι ιδιάζουσες τιμές του πίνακα σχεδιασμού $\boldsymbol{X}$. Προφανώς ισχύει πως $CI_j \geq 1 \ \forall j = 1, .., p+1$. Ουσιαστικά οι δείκτες κατάστασης συγκρίνουν τις ιδιάζουσες τιμές του πίνακα σχεδιασμού με τη μέγιστη ιδιάζουσα τιμή του, με σκοπό να αποτυπώσουν την ύπαρξη και το πλήθος των γραμμικών σχέσεων ανάμεσα στις επεξηγηματικές μεταβλητές.

Το τελευταίο δεν προκύπτει διαισθητικά, αλλά βασίζεται στην παραγοντοποίηση σε ιδιάζουσες τιμές (Singular Value Decomposition - SVD) του πίνακα $\boldsymbol{X}$. Βάσει της SVD ο $\boldsymbol{X}$ παραγοντοποιείται ως εξής:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

όπου $\boldsymbol{U}_{n\times n}$ και $\boldsymbol{V}_{(p+1)\times(p+1)}$ οι ορθογώνιοι πίνακες που έχουν για στήλες τα ιδιοδιανύσματα των $\boldsymbol{X}\boldsymbol{X}^T$ και $\boldsymbol{X}^T\boldsymbol{X}$ αντιστοίχως, ενώ $\boldsymbol{D}_{n\times(p+1)}$ είναι ο πίνακας με τις ιδιάζουσες τιμές του $\boldsymbol{X}$ στη κύρια διαγώνιο και όλα τα υπόλοιπα στοιχεία μηδενικά. Να θυμίσουμε εδώ ότι ένας πίνακας $\boldsymbol{A}$ καλείται ορθογώνιος όταν $\boldsymbol{A}^T\boldsymbol{A} = \boldsymbol{A}\boldsymbol{A}^T = \boldsymbol{I}$. Υποθέτουμε πως ο $\boldsymbol{X}$ έχει ακριβώς $p+1-r$ τέλειους γραμμικούς συνδυασμούς μεταξύ των στηλών του και άρα $rank(\boldsymbol{X}) = r < p+1$. Από τη στιγμή που οι $\boldsymbol{U}$ και $\boldsymbol{V}$ είναι ορθογώνιοι, είναι εξ′ορισμού πλήρους βαθμού, δηλαδή $rank(\boldsymbol{U}) = n$ και $rank(\boldsymbol{V}) = p+1$. Οπότε πρέπει να ισχύει

πως $rank(\boldsymbol{D}) = rank(\boldsymbol{X}) = r$. Αυτό σημαίνει πως ο $\boldsymbol{D}$ έχει ακριβώς τόσα μηδενικά στοιχεία στη κύρια διαγώνιό του, όσοι είναι και οι γραμμικοί συνδυασμοί του πίνακα σχεδιασμού $\boldsymbol{X}$. Βέβαια επειδή στην πράξη όπως έχουμε προαναφέρει είναι αδύνατο να παρατηρήσουμε τέλειους γραμμικούς συνδυασμούς μεταξύ των επεξηγηματικών μεταβλητών, δεν πρόκειται να παρατηρήσουμε μηδενικές ιδιάζουσες τιμές του πίνακα σχεδιασμού $\boldsymbol{X}$. Εφόσον υπάρχουν στατιστικά σημαντικές γραμμικές σχέσεις μεταξύ των συμμεταβλητών, θα προκύψουν και αντίστοιχες το πλήθος, αρκετά μικρές, κοντά στο μηδέν, ιδιάζουσες τιμές. Επειδή όμως δεν υπάρχει κάποιο αντικειμενικό μέτρο για το πόσο κοντά στο μηδέν μια τιμή μπορεί να θεωρηθεί μικρή, συγκρίνουμε τις ιδιάζουσες τιμές με τη μεγαλύτερη και έχουμε μια πιο κατατοπιστική εικόνα.

Έτσι λοιπόν μπορούμε να αντιστοιχίσουμε τις τιμές των δεικτών κατάστασης με την ύπαρξη πολυσυγγραμικότητας ακολούθως, σύμφωνα με τους Belsley, Kuh and Welsch, 1980:

- $CI_j > 15 \implies$ Πιθανό πρόβλημα πολυσυγγραμικότητας
- $CI_j > 30 \implies$ Σίγουρο πρόβλημα πολυσυγγραμικότητας

Πρέπει να καταστήσουμε σαφές ότι οι δείκτες κατάστασης από μόνοι τους δεν μας δείχνουν το ποιές επεξηγηματικές μεταβλητές περιλαμβάνονται σε γραμμικές σχέσεις, αλλά αποτελούν μια ένδειξη για την ύπαρξη στατιστικά σημαντικών γραμμικών συνδυασμών. Για παράδειγμα αν είχαμε δύο δείκτες κατάστασης μεγαλύτερους του 30, θα είχαμε δύο σίγουρους γραμμικούς συνδυασμούς, αν είχαμε τρεις μεγαλύτερους του 30, θα είχαμε τρεις σίγουρους γραμμικούς συνδυασμούς, κ.ο.κ.

2.1.4 Ποσοστά διάσπασης διασποράς

Τα ποσοστά διάσπασης διασποράς (variance decomposition proportions) σε συνδυασμό με τους δείκτες κατάστασης, μπορούν να μας δώσουν μια σαφή εικόνα για το ποιές επεξηγηματικές μεταβλητές εμπεριέχονται σε πιθανό πρόβλημα πολυσυγγραμικότητας.

Για τον ορισμό τους θα χρησιμοποιήσουμε την παραγοντοποίηση σε ιδιάζουσες τιμές (SVD) του πίνακα σχεδιασμού $\boldsymbol{X}$, στην οποία αναφερθήκαμε στην Παράγραφο 2.1.3 . Βάσει της SVD, μπορούμε να γράψουμε τον πίνακα διασπορών συνδιασπορών της εκτιμήτριας ελαχίστων τετραγώνων $\hat{\boldsymbol{\beta}}^{OLS}$ ακολούθως:

$$\begin{aligned} Var\left[\hat{\boldsymbol{\beta}}^{OLS}\right] &= \sigma_{\varepsilon}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma_{\varepsilon}^2\boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{V}^T. \end{aligned}$$

Άρα για τον k-οστό συντελεστή της εκτιμήτριας ελαχίστων τετραγώνων $\hat{\boldsymbol{\beta}}$ έχουμε ότι η διασπορά του ισούται με:

$$Var\left[\hat{\beta}_k^{OLS}\right] = \sigma_\varepsilon^2 \sum_{j=0}^{p} \frac{v_{kj}^2}{d_j^2}$$

όπου d_j^2 οι ιδιοτιμές του $\boldsymbol{X}^T\boldsymbol{X}$ και $\boldsymbol{V} \equiv (v_{ij})$. Παρατηρούμε ότι αναλύσαμε τη διασπορά $Var\left[\hat{\beta}_k^{OLS}\right]$ σε ένα άθροισμα j στοιχείων, καθένα από τα οποία εμπεριέχει μόνο μια από τις $p+1$ ιδιοτιμές. Από τη στιγμή που οι ιδιοτιμές εμφανίζονται στον παρανομαστή, τα στοιχεία τα οποία σχετίζονται με πιθανή πολυσυγγραμικότητα θα είναι μεγαλύτερα σε σχέση με τα υπόλοιπα στοιχεία, διότι όπως αναφέραμε στην Παράγραφο 2.1.3 οι ιδιάζουσες τιμές d_j θα είναι μικρές κοντά στο 0.

Ορίζουμε το k,j-οστό ποσοστό διάσπασης διασποράς π_{kj} ως το ποσοστό της διασποράς του k-οστού συντελεστή $\hat{\beta}_k$ που σχετίζεται με το j-οστό στοιχείο της διάσπασης του. Άρα το k,j-οστό ποσοστό διάσπασης διασποράς είναι ίσο με:

$$\pi_{kj} = \frac{\phi_{kj}}{\phi_k}$$

όπου $\phi_{kj} = \dfrac{v_{kj}^2}{d_j^2}$ και $\phi_k = \sum_{j=0}^{p} \phi_{kj}$. Έτσι λοιπόν, ένα υψηλό ποσοστό ανάλυσης διασποράς δύο ή παραπάνω συντελεστών που συνδέεται με στοιχεία της ίδιας ιδιάζουσας τιμής, υποδεικνύει την ύπαρξη πιθανής πολυσυγγραμικότητας η οποία και δημιουργεί προβλήματα. Εμείς κοιτάμε λοιπόν για δύο ή παραπάνω υψηλά ποσοστά διάσπασης διασποράς που συνδέονται με έναν μεγάλο δείκτη κατάστασης. Τότε οι επεξηγηματικές μεταβλητές των οποίων οι συντελεστές έχουν τα υψηλά ποσοστά διάσπασης διασποράς είναι πιθανό να προκαλούν προβλήματα πολυσυγγραμικότητας. Υψηλό ποσοστό διάσπασης διασποράς σύμφωνα με τον *Belsley* είναι ένα ποσοστό της τάξεως του 50% και άνω.

2.2 Πολυσυγγραμικότητα στην R

Στο Κεφάλαιο 1, προσαρμόσαμε το πολλαπλό γραμμικό μοντέλο παλινδρόμησης εξετάζοντας τις προϋποθέσεις του και εξάγοντας κάποια πρώτα συμπεράσματα. Πλέον, μετά τη θεωρία που αναπτύξαμε σε αυτό το Κεφάλαιο, θα μπορέσουμε να αναλύσουμε καλύτερα το αν το μοντέλο που εφαρμόσαμε είναι κατάλληλο, μελετώντας την ύπαρξη στατιστικά σημαντικών γραμμικών σχέσεων, δηλαδή την ύπαρξη πολυσυγγραμικότητας. Θα εφαρμόσουμε όλες τις μεθόδους εντοπισμού που αναπτύξαμε στην Παράγραφο 2.1.

Πίνακας συσχετίσεων

Πρώτα θα υπολογίσουμε τον πίνακα συσχετίσεων (Correlation matrix) των τιμών των επεξηγηματικών μας μεταβλητών, που περιέχει στη διαγώνιό του τις διασπορές των τιμών των συμμεταβλητών μας και στις υπόλοιπες θέσεις του, τους συντελεστές συσχέτισεις των τιμών των επεξηγηματικών μεταβλητών, που αναπτύξαμε στην Παράγραφο 2.1.1. Θυμίζουμε αν εντοπίσουμε υψηλές κατ' απόλυτη τιμή συσχετίσεις έχουμε μια πρώτη ένδειξη γραμμικών σχέσεων, χωρίς βέβαια αυτό να σημαίνει ότι πιθανή απουσία υψηλών συσχετίσεων συνεπάγεται έλλειψη προβλήματος.

```
c<-as.data.frame(round(cor(X.diab),2))
install.packages("corrplot")
library(corrplot)
corrplot(as.matrix(c), method="color",tl.col=1,tl.cex
   =0.9)
```

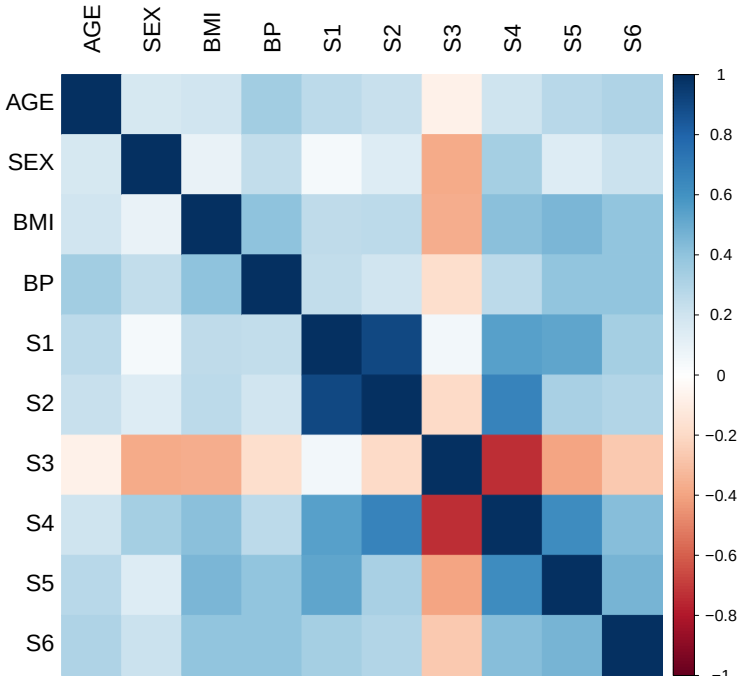
Κώδικας 2.1: Υπολογισμός του πίνακα συσχετίσεων με την εντολή cor() και γραφική του αναπαράσταση με την εντολή corrplot().

```
> c
      AGE   SEX   BMI    BP   S1    S2    S3    S4    S5    S6
AGE  1.00  0.17  0.19  0.34 0.26  0.22 -0.08  0.20  0.27  0.30
SEX  0.17  1.00  0.09  0.24 0.04  0.14 -0.38  0.33  0.15  0.21
BMI  0.19  0.09  1.00  0.40 0.25  0.26 -0.37  0.41  0.45  0.39
BP   0.34  0.24  0.40  1.00 0.24  0.19 -0.18  0.26  0.39  0.39
S1   0.26  0.04  0.25  0.24 1.00  0.90  0.05  0.54  0.52  0.33
S2   0.22  0.14  0.26  0.19 0.90  1.00 -0.20  0.66  0.32  0.29
S3  -0.08 -0.38 -0.37 -0.18 0.05 -0.20  1.00 -0.74 -0.40 -0.27
S4   0.20  0.33  0.41  0.26 0.54  0.66 -0.74  1.00  0.62  0.42
S5   0.27  0.15  0.45  0.39 0.52  0.32 -0.40  0.62  1.00  0.46
S6   0.30  0.21  0.39  0.39 0.33  0.29 -0.27  0.42  0.46  1.00
```

Εικόνα 2.1: Πίνακας συσχετίσεων των τιμών των επεξηγηματικών μας μεταβλητών

Στον Κώδικα 2.1 εκχωρούμε στο c το πλαίσιο δεδομένων που αποτελείται από τα στοιχεία του πίνακα συσχετίσεων των τιμών των συμμεταβλητών, στρογγυλοποιημένα στα 2 δεκαδικά ψηφία. Στην Εικόνα 2.1 βλέπουμε αυτόν τον πίνακα συσχετίσεων και παρατηρούμε ότι υπάρχουν κάποιες σχετικά υψηλές κατά απόλυτη τιμή συχετίσεις ανάμεσα στις τιμές των επεξηγηματικών μεταβλητών $S1$ και $S2$ ($corr(S1, S2 = 0.9)$), $S2$ και $S4$ ($corr(S2, S4 = 0.66)$), $S3$ και $S4$ ($corr(S3, S4 = -0.74)$), $S4$ και $S5$ ($corr(S4, S5 = 0.62)$). Άρα ένα πρώτο συμπέρασμα που μπορούμε να εξάγουμε είναι ότι η συμμεταβλητή $S4$ με μεγάλη πιθανότητα εμπεριέχεται σε έναν ή πολλούς γραμμικούς συνδυασμούς

και οι συμμεταβλητές $S1$ και $S2$ αλληλοεπηρεάζονται σε μεγάλο βαθμό. Τα αποτελέσματα του πίνακα συσχετίσεων οπτικοποιούνται στην Εικόνα 2.2, με την εντολή *corrplot*() του πακέτου *corrplot* στον Κώδικα 2.1, που δέχεται ως όρισμα τον πίνακα συσχετίσεων c. Οι θετικές συσχετίσεις απεικονίζονται στη κλίμακα του μπλέ ενώ οι αρνητικές συσχετίσεις στη κλίμακα του κόκκινου, με τις μεγαλύτερες κατά απόλυτη τιμή συσχετίσεις να είναι πιο σκούρες από τις μικρότερες.



Εικόνα 2.2: Οπτικοποίηση του πίνακα συσχετίσεων των δεδομένων των επεξηγηματικών μεταβλητών.

Παρόλα αυτά, όπως αναφέραμε και στο θεωρητικό κομμάτι του παρόντος κεφαλαίου, κανένα ασφαλές συμπέρασμα δεν μπορεί να εξαχθεί μόνο από τους συντελεστές συσχέτισης, οπότε προχωράμε και στις υπόλοιπες μεθόδους εντοπισμού.

Παράγοντες πληθωρισμού διασποράς

Οι παράγοντες πληθωρισμού διασποράς (Variance Inflation Factors) που αναπτύξαμε στην Παράγραφο 2.1.2, και συγκεκριμένα οι ρίζες τους, μας δείχνουν πόσο μεγαλύτερο είναι το τυπικό σφάλμα της εκτιμήτριας $\hat{\beta}_j^{OLS}$ συγκριτικά με

την τιμή που θα είχε αν και εφόσον η συμμεταβλητή $\boldsymbol{X}_j$ ήταν γραμμικώς ασυσχέτιστη με τις υπόλοιπες συμμεταβλητές. Στον Κώδικα 2.2 υπολογίζουμε αυτές τις τιμές με την εντολή $vif()$ του πακέτου car, που δέχεται ως όρισμα τη λίστα του μοντέλου παλινδρόμησης $reg.diab$ που προσαρμόσαμε στον Κώδικα 1.2. Υπολογίζουμε επίσης τους συντελεστές προσδιορισμού R_j^2 (των παλινδρομήσεων με μεταβλητή απόκρισης την μεταβλητή $\boldsymbol{X}_j$ και επεξηγηματικές μεταβλητές τις υπόλοιπες συμμεταβλητές) βάσει του ορισμού της Παραγράφου 2.1.2. Τιμές των $VIF > 10 \iff R^2 > 0.9$ δείχνουν σχεδόν σίγουρη ύπαρξη πολυσυγγραμικότητας.

```
1 VIF<-vif(reg.diab) #vif>10 potential collinearity
     problem
2 rsqrd<-1-(1/vif(reg.diab))#rsqrd of each regression
```

Κώδικας 2.2: Υπολογισμός των παραγόντων πληθωρισμού διασποράς με τη συνάρτηση $vif()$ και των συντελεστών προσδιορισμού βάσει του ορισμού των VIF

```
> round(VIF,1)
 AGE  SEX  BMI   BP   S1   S2   S3   S4   S5   S6
 1.2  1.3  1.5  1.5 59.2 39.2 15.4  8.9 10.1  1.5
> round(rsqrd,4)
   AGE    SEX    BMI     BP     S1     S2     S3     S4     S5     S6
0.1785 0.2176 0.3375 0.3148 0.9831 0.9745 0.9351 0.8875 0.9008 0.3264
```

Εικόνα 2.3: Οι τιμές των παραγόντων πληθωρισμού διασποράς στρογγυλοποιημένες στο 1 δεκαδικό ψηφίο και οι συντελεστές προσδιορισμού στρογγυλοποιημένοι στα 4 δεκαδικά ψηφία. Παρατηρούμε τιμές που μας υποδεικνύουν σχεδόν σίγουρη πολυσυγγραμικότητα.

Από την Εικόνα 2.3 βλέπουμε ότι οι παράγοντες πληθωρισμού διασποράς των συμμεταβλητών $S1$ και $S2$ είναι πολύ μεγαλύτερες του 10, 59.2 και 39.2 αντιστοίχως, με αντίστοιχους συντελεστές προσδιορισμού στο 98% και 97%. Μεγαλύτερος του 10 είναι και ο VIF της συμμεταβλητής $S3$, ίσος με 15, ίσος με 10 ο VIF της επεξηγηματικής $S5$ και λιγο μικρότερος του 10, ίσος με 9, ο VIF της επεξηγηματικής $S4$. Άρα μπορούμε να καταλήξουμε στο ασφαλές συμπέρασμα ότι το μοντέλο παλινδρόμησης που εξετάζουμε πάσχει από πολυσυγγραμικότητα, με τις μεταβλητές $S1$, $S2$, $S3$ και $S5$ να εμπλέκονται στις στατιστικά σημαντικές γραμμικές σχέσεις που υπάρχουν.

Δείκτες κατάστασης

Οι δείκτες κατάστασης Condition Indices -CI που αναπτύξαμε στην Παράγραφο 2.1.3 αποτυπώνουν το πλήθος των στατιστικά σημαντικών γραμμικών σχέσεων των συμμεταβλητών που υπάρχουν σε ένα μοντέλο παλινδρόμησης, ελέγχοντας πόσο κοντά στο 0 είναι οι ιδιάζουσες τιμές του πίνακα σχεδιασμού $\boldsymbol{X}$ καθώς τις συγκρίνουν με την μέγιστη ιδιάζουσα τιμή του $\boldsymbol{X}$. Εφόσον έχουμε κάποιον $CI > 15$ έχουμε πιθανή ύπαρξη μιας στατιστικά σημαντικής γραμμικής σχέσης ενώ αν $CI > 30$ έχουμε σίγουρη ύπαρξη μιας στατιστικά σημαντικής γραμμικής σχέσης. Στον Κώδικα 2.3 υπολογίζουμε του δείκτες κατάστασης του πίνακα σχεδιασμού $\boldsymbol{X}$ και αναλυτικά, όπως τους ορίσαμε στην Παράγραφο 2.1.3, και με την εντολή $colldiag()\$cond$ του πακέτου *perturb* που δέχεται ως όρισμα τη λίστα της παλινδρόμησής μας $reg.diab$ και το $scale = F$ για να μην τυποποιήσει τις τιμές του πίνακα σχεδιασμού.

```
X1<-model.matrix(reg.diab)
values<-sqrt(eigen(t(X1)% * %X1)$value) #square root of
   eigen values of (X^T)X
conditionindx<-max(values)/values #if >15 possible
   collinearity problem, if >30 certain
install.packages("perturb")
library(perturb)
ci<-colldiag(reg.diab,scale=F)$cond
```

Κώδικας 2.3: Υπολογισμός των δεικτών κατάστασης αναλυτικά και με την εντολή $colldiag()\$cond$

Η εντολή $model.matrix()$ δέχεται ως όρισμα το μοντέλο παλινδρόμησης και επιστρέφει τον πίνακα σχεδιασμού του. Η εντολή $eigen()\$value$ υπολογίζει τις ιδιοτιμές του πίνακα που δέχεται ως όρισμα. Ο ανάστροφος πίνακας δίνεται από την εντολή $t()$ και ο πολλαπλασιασμός πινάκων γίνεται με το $\% * \%$. Παραθέτουμε λοιπόν τα αποτελέσματα.

Στην Εικόνα 2.4 βλέπουμε τις τιμές των δεικτών κατάστασης, με τον αναλυτικό υπολογισμό στην πρώτη στήλη και με την έτοιμη συνάρτηση στην δεύτερη στήλη, οι οποίες είναι ίδιες. Όλοι, πλην των πρώτων δύο δεικτών είναι μεγαλύτεροι του 15, ενώ οι τελευταίοι 6 είναι και μεγαλύτεροι του 30, κατά πολύ μάλιστα. Διαπιστώνουμε λοιπόν ότι είναι σίγουρη η ύπαρξη αρκετών, περίπου 5, στατιστικά σημαντικών γραμμικών σχέσεων μεταξύ των επεξηγηματικών μας μεταβλητών, ενώ είναι πολύ πιθανό να υπάρχουν και περισσότερες. Το ποιές ακριβώς συμμεταβλητές εμπλέκονται θα το αποσαφηνίσουμε στην επόμενη μέθοδο, των ποσοστών ανάλυσης διασποράς, που συνδέεται άμεσα με τους

```
> data.frame(conditionindx,ci)
   conditionindx cond.index
1        1.00000    1.00000
2       10.88622   10.88622
3       17.15451   17.15451
4       23.88905   23.88905
5       27.73559   27.73559
6       36.97522   36.97522
7       70.82524   70.82524
8      469.24639  469.24639
9      621.62483  621.62483
10     993.30464  993.30464
11    7196.48954 7196.48954
```

Εικόνα 2.4: Οι τιμές των δεικτών κατάστασης, ίδιες και με τον αναλυτικό υπολογισμό (conditionindx) και με την έτοιμη συνάρτηση colldiag()$cond

δείκτες κατάστασης.

Ποσοστά διάσπασης διασποράς

Τα ποσοστά διάσπασης διασποράς (Variance Decomposition Proportions) που αναπτύξαμε στην Παράγραφο 2.1.4, μπορούν να μας δώσουν μια εικόνα, σε συνδυασμό με τους δείκτες κατάστασης, για το ποιές επεξηγηματικές μεταβλητές δημιουργούν προβλήματα πολυσυγγραμικότητας. Θεωρούμε υψηλά τα ποσοστά άνω του 50%. Συγκεκριμένα δύο ή παραπάνω υψηλά ποσοστά διάσπασης διασποράς που συνδέονται με έναν μεγάλο δείκτη κατάστασης υποδεικνύουν τις μεταβλητές που κατά πάσα πιθανότητα εμπλέκονται σε στατιστικά σημαντικές γραμμικές σχέσεις.

```
1 vdp<-round(colldiag(reg.diab,scale=F)$pi,3)
2 cd<-colldiag(reg.diab,scale=F)
```

Κώδικας 2.4: Υπολογισμός των ποσοστών διάσπασης διασποράς και των δεικτών κατάστασης με την εντολή colldiag

Στον Κώδικα 2.4 υπολογίζουμε τα ποσοστά διάσπασης διαποράς των εκτιμητριών με την εντολή *colldiag*()$*pi* και τους δείκτες κατάστασης του πίνακα σχεδιασμού. Στην Εικόνα 2.5 εμφανίζουμε τον πίνακα που περιέχει όλα τα ποσοστά ανάλυσης διασποράς ενώ στην Εικόνα 2.6 μόνο τα ποσοστά άνω του 50% και τους αντίστοιχους δείκτες κατάστασης στην δεύτερη στήλη.

Τα συμπεράσματά μας μπορούμε να τα εξάγουμε από την Εικόνα 2.6 που εμπεριέχονται και οι δείκτες κατάστασης ενώ εμφανίζονται μόνο τα ποσοστά διάσπασης διασποράς άνω του 50%.

```
> vdp
      intercept   AGE   SEX   BMI    BP    S1    S2    S3    S4    S5    S6
 [1,]     0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000 0.000
 [2,]     0.000 0.012 0.000 0.000 0.058 0.002 0.013 0.002 0.000 0.000 0.024
 [3,]     0.000 0.027 0.000 0.000 0.051 0.007 0.007 0.027 0.000 0.000 0.021
 [4,]     0.000 0.945 0.000 0.001 0.070 0.000 0.000 0.000 0.000 0.000 0.038
 [5,]     0.000 0.000 0.000 0.000 0.575 0.000 0.001 0.000 0.000 0.000 0.532
 [6,]     0.000 0.001 0.000 0.000 0.005 0.128 0.175 0.048 0.000 0.000 0.018
 [7,]     0.000 0.001 0.000 0.844 0.121 0.002 0.000 0.002 0.000 0.000 0.125
 [8,]     0.000 0.000 0.031 0.036 0.030 0.048 0.003 0.121 0.456 0.009 0.160
 [9,]     0.000 0.009 0.969 0.032 0.027 0.012 0.005 0.001 0.044 0.001 0.000
[10,]     0.001 0.002 0.000 0.083 0.053 0.085 0.172 0.010 0.333 0.304 0.069
[11,]     0.999 0.002 0.001 0.002 0.010 0.715 0.624 0.789 0.167 0.686 0.013
```

Εικόνα 2.5: Τα ποσοστά διάσπασης διασποράς που υπολογίσαμε. Παρατηρούμε ότι κατά στήλη αθροίζουν στη μονάδα.

```
> print(cd,fuzz=.5)
Condition
Index Variance Decomposition Proportions
             intercept AGE   SEX   BMI   BP    S1    S2    S3    S4 S5    S6
1      1.000  .          .     .     .     .     .     .     .     .  .     .
2     10.886  .          .     .     .     .     .     .     .     .  .     .
3     17.155  .          .     .     .     .     .     .     .     .  .     .
4     23.889  .        0.945   .     .     .     .     .     .     .  .     .
5     27.736  .          .     .     .   0.575   .     .     .     .  .   0.532
6     36.975  .          .     .     .     .     .     .     .     .  .     .
7     70.825  .          .     .   0.844   .     .     .     .     .  .     .
8    469.246  .          .     .     .     .     .     .     .     .  .     .
9    621.625  .          .   0.969   .     .     .     .     .     .  .     .
10   993.305  .          .     .     .     .     .     .     .     .  .     .
11  7196.490 0.999      .     .     .     .   0.715 0.624 0.789   . 0.686  .
```

Εικόνα 2.6: Τα ποσοστά διάσπασης διασποράς άνω του 50% και οι αντίστοιχοι δείκτες κατάστασης.

Παρατηρώντας την Εικόνα 2.6 βλέπουμε ότι στην γραμμή του 5ου δείκτη κατάστασης που είναι λίγο μικρότερος του 30, έχουμε 2 ποσοστά διάσπασης διασποράς μεγαλύτερα του 50%, αυτά της μεταβλητής BP (πίεση αίματος) και $S6$. Είναι αρκετά πιθανό αυτές οι δύο συμμεταβλητές να δημιουργούν πρόβλημα πολυσυγγραμικότητας. Στην γραμμή του τελευταίου δείκτη κατάστασης, που είναι και ο μεγαλύτερος όλων, έχουμε 4 ποσοστά διάσπασης διασποράς (δεν συμπεριλαμβάνουμε τη σταθερά) που είναι μεγαλύτερα του 50%, αυτά των επεξηγηματικών μεταβλητών $S1$, $S2$, $S3$ και $S5$. Μπορούμε να αποφανθούμε με σιγουριά ότι αυτές οι συμμεταβλητές εμπεριέχονται σε στατιστικά σημαντικές γραμμικές σχέσεις που δημιουργούν προβλήματα στο μοντέλο.

3

Επιλογή μεταβλητών

Η επιλογή μεταβλητών στο πολλαπλό γραμμικό μοντέλο δεν είναι τίποτα άλλο από την επιλογή του καλύτερου μοντέλου, δηλαδή ποιές επεξηγηματικές μεταβλητές χρειάζεται να χρησιμοποιηθούν ούτως ώστε το μοντέλο μας να εξυπηρετήσει το δυνατόν καλύτερα το σκοπό του, την πρόβλεψη της μεταβλητής απόκρισης. Το ποιό είναι το καλύτερο μοντέλο όμως είναι καθαρά υποκειμενικό. Έχουν αναπτυχθεί διάφορες διαδικασίες για την εύρεσή του και παρόλο που ορισμένες φορές τα αποτελέσματα τους συμφωνούν (πολλές φορές δεν συμφωνούν), δεν παύει το ζήτημα της επιλογής μεταβλητών να έγκειται κατά μεγάλο βαθμό στα όρια που θεσπίζει και στην ανάλυση που αναπτύσσει ο εκάστοτε ερευνητής καθώς και στη φύση του προβλήματος που μελετάται. Άλλωστε όπως έχει πεί και ο βρετανός στατιστικολόγος George E.P. Box, «όλα τα μοντέλα είναι λάθος, αλλά ορισμένα είναι χρήσιμα». Η βασική αρχή της επιλογής μεταβλητών αφορά την εξισορρόπηση της καλής προσαρμογής ενός μοντέλου (goodness of fit) και του πόσο φειδωλό (parsimonious), από πλευράς πλήθους των επεξηγηματικών μεταβλητών, είναι το μοντέλο. Αναζητούμε συνήθως "οικονομικά" μοντέλα, τα οποία προσαρμόζονται όσο πιο καλά γίνεται στα δεδομένα και εξυπηρετούν παράλληλα το σκοπό της πρόβλεψης της μεταβλητής απόκρισης. Η σημαντικότητα της επιλογής μεταβλητών πηγάζει σε μεγάλο βαθμό και από τα προβλήματα που δημιουργεί πιθανή ύπαρξη πολυσυγγραμικότητας σε ένα μοντέλο παλινδρόμησης. Εφόσον λοιπόν έχει ανιχνευθεί η ύπαρξη πολυσυγγραμικότητας, είναι φρόνιμο να προχωρήσει κανείς σε κάποια από τις διαδικασίες επιλογής μεταβλητών που έχουν αναπτυχθεί. Το μοντέλο με το οποίο θα καταλήξουμε παρ′όλα αυτά δεν μας εγγυάται κανείς ότι δεν θα πάσχει από πολυσυγγραμικότητα. Πολλές φορές το πρόβλημα της πολυσυγγραμικότητας δυστυχώς συνεχίζει να υπάρχει και καλό είναι να γίνουν πολλές δοκιμές, επανελέγχοντας κάθε φορά τις ιδότητες του μοντέλου με το οποίο καταλήγουμε. Η μόνη περίπτωση κατά την οποία οι διαδικασίες επιλογής μεταβλητών που θα περιγράψουμε μας προμηθεύουν με το βέλτιστο μοντέλο είναι

όταν ο πίνακας σχεδιασμού $\boldsymbol{X}$ είναι ορθογώνιος, κάτι το οποίο θα μπορούσε να επιτευχθεί με προσεκτικό σχεδιασμό του πειράματος συλλογής δεδομένων, όμως πολλές φορές είναι δύσκολο, χρονοβόρο και κοστοβόρο.

3.1 Κριτήρια πληροφορίας AIC και BIC

Τα κριτήρια πληροφορίας AIC ($Akaike\ Information\ Criterion$) και BIC ($Bayesian\ Information\ Criterion$) δεν αποτελούν μέθοδο επιλογής μεταβλητών, αλλά κριτήρια σύγκρισης μοντέλων που χρησιμοποιούνται στην επιλογή μεταβλητών και για αυτόν τον λόγο, πριν προχωρήσουμε στις μεθόδους επιλογής μεταβλητών, σκόπιμο είναι να τα αναπτύξουμε. Τα AIC και BIC εξετάζουν και εκτιμούν την απώλεια πληροφορίας που εμφανίζεται όταν ένα καινούργιο μοντέλο καλείται να αναπαραστήσει τη διαδικασία που γέννησε τα δεδομένα. Η πληροφορία αυτή υπεισέρχεται ως ένα $trade-off$ μεταξύ της καλής προσαρμογής του μοντέλου ($goodness\ of\ fit$) και της πολυπλοκότητας ($complexity$) του μοντέλου. Τα προηγούμενα γίνονται πιο κατανοητά κατά τον ορισμό των AIC και BIC:

$$AIC = -2ln(L(\boldsymbol{\beta}|\boldsymbol{X})) + 2df$$
$$BIC = -2ln(L(\boldsymbol{\beta}|\boldsymbol{X})) + ln(n)df$$

όπου $L(\boldsymbol{\beta}|\boldsymbol{X})$ η μέγιστη τιμή της πιθανοφάνειας, ήτοι η τιμή της για τους εκτιμητές μέγιστης πιθανοφάνειας των παραμέτρων του μοντέλου, $df = p + 2$ ο αριθμός των παραμέτρων του μοντέλου προς εκτίμηση που συναντώνται συχνά ως βαθμοί ελευθερίας του μοντέλου, δηλαδή οι συντελεστές $\beta_1, .., \beta_p$, η σταθερά β_0 και η διασπορά του τυχαίου σφάλματος σ_ε^2 και n ο αριθμός των παρατηρήσεων.

Μπορούμε να παρατηρήσουμε ότι και το AIC και το BIC ποινικοποιούν ένα παράγοντα πολλαπλάσιο του αριθμού των παραμέτρων προς εκτίμηση df, οι οποίοι αποτυπώνουν την πολυπλοκότητα, ως προς την πιθανοφάνεια, που αποτελεί ένα μέτρο καλής προσαρμογής του μοντέλου. Μπορούμε να παρατηρήσουμε ότι η ποινικοποίηση του αριθμού των παραμέτρων προς εκτίμηση των δύο κριτηρίων είναι ίδια όταν $ln(n) = 2 \Leftrightarrow n \simeq 8$. Όταν οι παρατηρήσεις μας είναι $n > 8$ το κριτήριο πληροφορίας BIC ποινικοποιεί πάντα έναν μεγαλύτερο παράγοντα πολλαπλάσιο του αριθμού των παραμέτρων προς εκτίμηση απ′ ότι το κριτήριο πληροφορίας AIC. Και τα δύο αυτά κριτήρια πληροφορίας αποτελούν μέτρα σύγκρισης μοντέλων καθώς αν εξετάσουμε την τιμή τους για ένα μόνο μοντέλο, δεν μπορούμε να εξάγουμε κάποιο ασφαλές συμπέρασμα. Διαλέγουμε τα

μοντέλα που έχουν τις μικρότερες τιμές AIC και BIC. Αξίζει επίσης να σημειώσουμε ότι τα AIC και BIC υπό την υπόθεση ότι τα σφάλματα του μοντέλου είναι iid και ακολουθούν την κανονική κατανομή, που ισχύει στο πολλαπλό γραμμικό μοντέλο, γράφονται ως εξής:

$$AIC = n \cdot \left(ln\Big(\frac{2\pi RSS}{n}\Big) + 1 \right) + 2df$$

$$BIC = n \cdot \left(ln\Big(\frac{2\pi RSS}{n}\Big) + 1 \right) + ln(n)df.$$

3.2 Πλήρης εξερεύνηση του χώρου των πιθανών μοντέλων

Η διαδικασία της πλήρους εξερεύνησης του χώρου όλων των πιθανών μοντέλων συναντάται συχνά ως Full Enumeration. Υποθέτοντας ότι έχουμε p υποψήφιες επεξηγηματικές μεταβλητές για το μοντέλο μας, ο χώρος όλων των πιθανών συνδυασμών των επεξηγηματικών μεταβλητών απαρτίζεται από 2^p μοντέλα. Επιλέγουμε ένα κριτήριο σύγκρισης των διαφορετικών μοντέλων, όπως τα κριτήρια πληροφορίας AIC ή BIC, υπολογίζουμε για καθένα από τα 2^p μοντέλα το κριτήριο σύγκρισης που επιλέξαμε και εν τέλει διαλέγουμε το μοντέλο με τη χαμηλότερη (για κριτήριο σύγκρισης το AIC ή το BIC) τιμή του κριτηρίου σύγκρισης των πιθανών μας μοντέλων. Το Full Enumeration είναι μια απλή διαδικασία επιλογής μεταβλητών η οποία όμως έχει το σημαντικό μειονέκτημα ότι για μεγάλο αριθμό επεξηγηματικών μεταβλητών p είναι υπολογιστικά αδύνατη. Για παράδειγμα για $p = 40$ επεξηγηματικές μεταβλητές, τα πιθανά μοντέλα είναι $2^{40} = 1.099512e{+}12$, αριθμός τεράστιος που απαγορεύει την πλήρη εξερεύνηση του χώρου. Άρα η πλήρης εξερεύνηση του χώρου των πιθανών μας μοντέλων προτιμάται όταν έχουμε μικρό αριθμό υποψήφιων επεξηγηματικών μεταβλητών, συνήθως $p \leq 15$.

3.3 Επιλογή του καλύτερου υποσυνόλου των συμμεταβλητών

Η διαδικασία της επιλογής του καλύτερου υποσυνόλου των επεξηγηματικών μεταβλητών συναντάται συχνά με την ονομασία Best-Subset Selection. Έστω $k \in \{0, 1, 2, .., p\}$ ο αριθμός των επεξηγηματικών που δύναται να χρησιμοποιηθούν για την προσαρμογή του μοντέλου. Άρα για κάθε k έχουμε $\binom{p}{k}$ πιθανά μοντέλα. Η διαδικασία της επιλογής του καλύτερου υποσυνόλου λειτουργεί ακολούθως:

1. Για $k = 0, 1, 2, .., p$:

   α΄ Προσαρμόζουμε όλα τα $\binom{p}{k}$ πιθανά μοντέλα που περιέχουν ακριβώς k επεξηγηματικές μεταβλητές.

   β΄ Ανάμεσα από αυτά τα $\binom{p}{k}$ μοντέλα διαλέγουμε εκείνο που έχει το μικρότερο RSS (άθροισμα τετραγώνων των υπολοίπων) και το καλούμε $\mathcal{M}_k$.

2. Έχοντας βρει πλέον τα καλύτερα, βάσει του RSS, μοντέλα $\mathcal{M}_0, \mathcal{M}_1, \mathcal{M}_2$, ..., $\mathcal{M}_p$ για κάθε αριθμό επεξηγηματικών μεταβλητών, επιλέγουμε ένα από αυτά βασιζόμενοι σε κάποιο κριτήριο σύγκρισης μοντέλων της επιλογής μας (AIC ή BIC).

Προφανώς η διαδικασία αυτή αν γινόταν αναλυτικά θα είχε ακριβώς το ίδιο υπολογιστικό κόστος με τη διαδικασία του Full Enumeration. Οι Furnival and Wilson το 1974 ανέπτυξαν τον αποδοτικό αλγόριθμο leaps and bounds, τον οποίον δεν θα αναπτύξουμε, ο οποίος κάνει τη διαδικασία της επιλογής του καλύτερου υποσυνόλου προσιτή για αριθμό επεξηγηματικών μεταβλητών p μέχρι 40.

3.4 Διαδοχική αφαίρεση συμμεταβλητών

Η διαδικασία της διαδοχικής αφαίρεσης επεξηγηματικών μεταβλητών, η οποία είναι γνωστή και ως Backward Stepwise selection, λειτουργεί ακολούθως:

1. Ξεκινάμε από το πλήρες μοντέλο που περιλαμβάνει όλες τις επεξηγηματικές μεταβλητές.

2. Σε κάθε βήμα ελέγχουμε ποιά επεξηγηματική μεταβλητή πρέπει να αφαιρεθεί από το μοντέλο (μόνο μία κάθε φορά). Η επιλογή αυτής της συμμεταβλητής γίνεται βάσει κάποιου κριτήριου σύγκρισης μοντέλων (ελάχιστο AIC ή BIC ή μέγιστο $p - value$). Δηλαδή επιλέγουμε να αφαιρέσουμε τη συμμεταβλητή με την αφαίρεση της οποίας βελτιώνεται βέλτιστα το κριτήριο σύγκρισης που επιλέξαμε.

3. Σταματάμε όταν δεν μπορούμε να αφαιρέσουμε καμία άλλη επεξηγηματική μεταβλητή, δηλαδή όταν δεν βελτιώνεται περαιτέρω το κριτήριο σύγκρισης.

Προφανώς όταν έχουμε αφαιρέσει μια επεξηγηματική μεταβλητή δεν μπορούμε να την ξανασυμπεριλάβουμε στο μοντέλο, ακόμη και αν αυτή εμφανίζεται ως στατιστικά σημαντική σε κάποιο βήμα.

3.5 Διαδοχική πρόσθεση συμμεταβλητών

Η διαδικασία της διαδοχικής πρόσθεσης επεξηγηματικών μεταβλητών, η οποία είναι γνωστή και ως Forward Stepwise selection, λειτουργεί ακολούθως:

1. Ξεκινάμε από το μοντέλο που περιλαμβάνει μόνο τη σταθερά.

2. Σε κάθε βήμα ελέγχουμε ποιά επεξηγηματική μεταβλητή πρέπει να προστεθεί στο μοντέλο (μόνο μία κάθε φορά). Η επιλογή αυτής της συμμεταβλητής γίνεται με κάποιο κριτήριο σύγκρισης μοντέλων (ελάχιστο AIC ή BIC ή ελάχιστο $p-value$). Δηλαδή διαλέγουμε να συμπεριλάβουμε εκείνη τη συμμεταβλητή για την οποία το κριτήριο σύγκρισης του μοντέλου γίνεται βέλτιστο.

3. Σταματάμε όταν δεν μπορούμε να προσθέσουμε καμία άλλη επεξηγηματική μεταβλητή, δηλαδή όταν δεν βελτιώνεται περαιτέρω το κριτήριο σύγκρισης.

Η διαδικασία της διαδοχικής πρόσθεσης συμμεταβλητών έχει υπολογιστικά πλεονεκτήματα συγκριτικά με τη διαδικασία της διαδοχικής αφαίρεσης καθώς προσαρμόζει λιγότερα στο πλήθος μοντέλα.

3.6 Διαδικασία κατά βήματα

Η διαδικασία κατά βήματα (Stepwise procedure) αποτελεί μια παραλλαγή της διαδικασίας διαδοχικής πρόσθεσης, διότι επιτρέπει και την αφαίρεση μεταβλητών. Συγκεκριμένα:

1. Ξεκινάμε από ένα μοντέλο της επιλογής μας. Συνήθως προτιμάται το μοντέλο που περιλαμβάνει μόνο τη σταθερά

2. Σε κάθε βήμα ελέγχουμε όπως και στο Forward Stepwise Selection ποιά επεξηγηματική μεταβλητή πρέπει να προστεθεί βάσει ενός κριτηρίου σύγκρισης μοντέλων της επιλογής μας (AIC, BIC, κ.ο.κ.).

3. Αφότου προσθέσουμε τη συμμεταβλητή που παρέχει τη βέλτιστη τιμή για το κριτήριο σύγκρισης μοντέλων, ελέγχουμε αν μειώθηκε η σημαντικότητα κάποιας άλλης επεξηγηματικής μεταβλητής που υπάρχει ήδη στο μοντέλο μας. Εφόσον όντως παρατηρήσουμε εξασθένηση της σημαντικότητας κάποιας συμμεταβλητής, εξετάζουμε την αξία παραμονής της μέσω του κριτηρίου σύγκρισης μοντέλων.

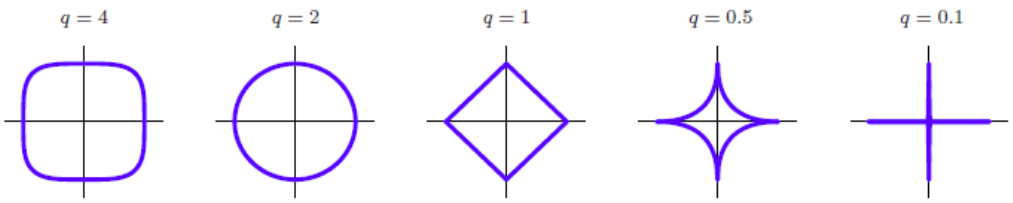
4. Σταματάμε όταν δεν μπορούμε να βελτιώσουμε περαιτέρω το μοντέλο μας.

Η διαδικασία κατά βήματα προτιμάται των διαδικασιών της διαδοχικής αφαίρεσης και της διαδοχικής πρόσθεσης λόγω του διπλού ελέγχου που πράττει. Και οι τρεις αυτές διαδικασίες επιλέγουν συνήθως αρεστά αλλά όχι βέλτιστα μοντέλα. Πολλές φορές «πέφτουν» στην παγίδα τοπικών μεγίστων ή ελαχίστων των κριτηρίων σύγκρισης για τον χώρο των πιθανών μοντέλων. Επίσης η αφαίρεση κάποιας συμμεταβλητής οδηγεί αυτομάτως στην αύξηση της σημαντικότητας κάποιας άλλης, με αποτέλεσμα η πραγματική της επίδραση πολλές φορές να υπερεκτιμάται. Δυστυχώς, οι διαδικασίες που έχουμε αναφέρει προς το παρόν δεν εφαρμόζονται σε δεδομένα υψηλών διαστάσεων λόγω υπολογιστικού κόστους.

3.7 Μέθοδοι Συρρίκνωσης

Οι μέθοδοι συρρίκνωσης των συντελεστών των επεξηγηματικών μεταβλητών, έχουν ευρεία εφαρμογή σε δεδομένα υψηλών διαστάσεων. Στις μεθόδους συρρίκνωσης, επιδιώκεται η εκτίμηση των συντελεστών των συμμεταβλητών μέσω της ελαχιστοποίησης του αθροίσματος των τετραγώνων των υπολοίπων (RSS) υπό κάποιον περιορισμό-ποινή που θεσπίζεται στην ℓ_q- νόρμα των συντελεστών. Τυποποιούμε τις τιμές των επεξηγηματικών μεταβλητών ώστε να έχουν μέση τιμή 0 και διασπορά 1, δηλαδή $\sum_{i=1}^{n} x_{ij} = 0$ και $\sum_{i=1}^{n} x_{ij}^2 = 1 \ \forall j = 1, 2, ..p$. Επίσης θεωρούμε ότι οι τιμές της μεταβλητής απόκρισης είναι κεντραρισμένες ώστε να έχουν μέση τιμή μηδέν και ότι η σταθερά παραλείπεται από το μοντέλο. Εφόσον θέλουμε να την συμπεριλάβουμε, η εκτίμησή της είναι η γνωστή: $\hat{\beta}_0 = \bar{\boldsymbol{y}}$ (όπου $\bar{\boldsymbol{y}}$ η μέση τιμή των μη κεντραρισμένων τιμών της μεταβλητής απόκρισης). Άρα καλούμαστε να επιλύσουμε το πρόβλημα ελαχιστοποίησης:

$$\begin{aligned} \underset{\boldsymbol{\beta} \in \mathbb{R}^p}{minimize} \quad & \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \\ s.t. \quad & \|\boldsymbol{\beta}\|_q \leq t, \end{aligned}$$



Εικόνα 3.1: Περιοχές των περιορισμών $\|\beta_j\|_q \leq t$ για $t = 1$ και $p = 2$ επεξηγηματικές μεταβλητές. Για $q < 1$ οι περιοχές των περιορισμών δεν είναι κυρτές.

Αναπτύσσοντας τις νόρμες και με χρήση των πολλαπλασιαστών *Lagrange*, το πρόβλημα γράφεται ισοδύναμα:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad \sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda\sum_{j=1}^{p}|\beta_j|^q$$

Η ποινή που θεσπίζεται στους συντελεστές ουσιαστικά ελέγχει τη διασπορά των εκτιμήσεων, ούτως ώστε να μην λάβει μεγάλες ανεπιθύμητες τιμές. Η τυποποίηση των τιμών των συμμεταβλητών γίνεται ώστε η ποινή να είναι ομοιόμορφη για τους συντελεστές όλων των επεξηγηματικών μεταβλητών. Στην Εικόνα 3.1 παρατηρούμε τις περιοχές των περιορισμών για κάποιες τιμές του q στο πολλαπλό γραμμικό μοντέλο δύο διαστάσεων. Οι πιο διαδεδομένες μεθοδολογίες συρρίκνωσης είναι οι *Ridge* ($q = 2$) και $LASSO$ ($q = 1$) και θα αποτελέσουν και το κύριο τμήμα μελέτης της παρούσας διπλωματικής εργασίας, στα Κεφάλαια 5 και 6 αντιστοίχως. Αποτελούν προέρτιο της διαδικασίας της επιλογής μεταβλητών καθώς εφαρμόζοντάς τες λαμβάνουμε μια πρώτη εικόνα του ποιές επεξηγηματικές μεταβλητές έχουν μεγαλύτερη επίδραση στην μεταβλητή απόκρισης και άρα προσεγγίζουν καλύτερα το επιθυμητό μοντέλο. Όπως θα δούμε παρακάτω, ενώ και η *Ridge* και η $LASSO$ συρρικνώνουν τους συντελεστές των επεξηγηματικών μεταβλητών, η $LASSO$ έχει την επιπλέον ιδιότητα να θέτει τους συντελεστές των λιγότερο στατιστικά σημαντικών συμμεταβλητών ίσους με μηδέν, διευκολύνοντας κατά αυτόν τον τρόπο την πρώτη αξιολόγηση (*screening*) που εφαρμόζουμε στις επεξηγηματικές μας μεταβλητές για να αποφανθούμε ποιές εξ αυτών θα συμπεριλάβουμε στο μοντέλο.

3.8 Επιλογή μεταβλητών στην R

Πλήρης εξερεύνηση του χώρου των πιθανών μοντέλων

Η διαδικασία της πλήρης εξερεύνησης του χώρου των πιθανών μοντέλων, για την οποία μιλήσαμε στην Παράγραφο 3.2, εφαρμόζεται σε προβλήματα που έχουν το πολύ 15 επεξηγηματικές μεταβλητές. Στην περίπτωσή μας, έχουμε $p = 10$ συμμεταβλητές, οπότε το πλήθος όλων των πιθανών μοντέλων είναι $2^{10} = 1024$ και μπορούμε εύκολα να τα εξερευνήσουμε όλα. Για να κάνουμε την πλήρη εξερεύνηση του χώρου των πιθανών μοντέλων χρειαζόμαστε μια συνάρτηση στην R η οποία θα παράγει μια κωδικοποίηση όλων των πιθανών 2^p μοντέλων. Την δουλειά αυτή την κάνει εν μέρει η συνάρτηση *integer.base.b*(), Κώδικας 3.1.

```
integer.base.b<-function(x,b=2){
  xi<-as.integer(x)
  if(any(is.na(xi) | ((x-xi)!=0)))
    print(list(ERROR="x is not integer", x=x))
  N<-length(x)
  xMax<-max(x)
  ndigits<-(floor(logb(xMax,base=2))+1)
  Base.b<-array(NA,dim=c(N,ndigits))
  for(i in 1:ndigits){
    #i<-1
    Base.b[,ndigits-i+1]<-(x%%b)
    x<-(x%/%b)
  }
  if(N==1) Base.b[1,] else Base.b
}
```

Κώδικας 3.1: Η συνάρτηση integer.base.b() κωδικοποιεί τον αριθμό x στο b-αδικό σύστημα. Αποτελεί βοηθητική συνάρτηση για την κωδικοποίηση στο δυαδικό σύστημα όλων των πιθανών μας μοντέλων.

Η συνάρτηση *integer.base.b*() του Κώδικα 3.1 δέχεται ως όρισμα έναν αριθμό x και έναν φυσικό αριθμό b και πραγματοποιεί την κωδικοποίηση του αριθμού x στο $b-$αδικό σύστημα. Ως προεπιλογή πραγματοποιεί την κωδικοποίηση του αριθμού στο δυαδικό σύστημα. Εμάς μας ενδιαφέρει να αποτυπώσουμε όλα τα πιθανά διαφορετικά μοντέλα που μπορούν να προκύψουν από οποιονδήποτε συνδυασμό των $p = 10$ επεξηγηματικών μας μεταβλητών. Άρα θέλουμε να παράξουμε μια $10-bit$ κωδικοποίηση, σε δυαδικό σύστημα, των 2^{10} διαφορετικών μοντέλων. Κατ' αυτόν τον τρόπο θα αποκτήσουμε έναν πίνακα $\boldsymbol{\gamma}$ του οποίου κάθε σειρά θα αντιστοιχεί σε κάποιο πιθανό μοντέλο. Στην j στήλη κάθε σειράς θα έχουμε είτε $\gamma_j = 1$, εφόσον η μεταβλητή X_j συμπεριλαμβάνεται στο μοντέλο, είτε $\gamma_j = 0$ εφόσον αυτή δεν συμπεριλαμβάνεται και αυτή ακριβώς θα είναι η αντιστοίχιση που θα χρησιμοποιήσουμε. Για παράδειγμα, το μοντέλο με τον αριθμό 1 θα είναι το:

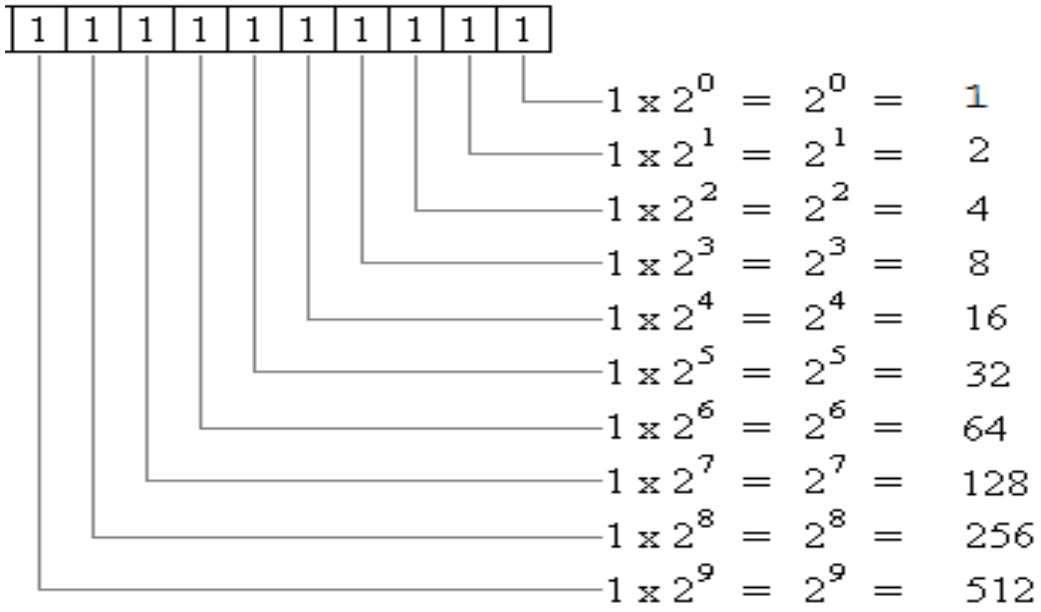
$$0\,0\,0\,0\,0\,0\,0\,0\,0\,1$$

και θα περιέχει μόνο τη μεταβλητή X_{10}, δηλαδή την $S6$. Το μοντέλο με τον αριθμό 2 θα είναι το:

$$0\,0\,0\,0\,0\,0\,0\,0\,1\,0$$

και θα περιέχει μόνο τη μεταβλητή X_9, δηλαδή την $S5$. Αντιστοίχως το μοντέλο με τον αριθμό $2^p - 1 = 1023$ θα είναι το:

$$1\,1\,1\,1\,1\,1\,1\,1\,1\,1$$



Εικόνα 3.2: 10-bit full model

και θα περιέχει όλες τις επεξηγηματικές μεταβλητές. Στην Εικόνα 3.2 παρουσιάζεται μάλιστα και το πως λειτουργεί η κωδικοποίηση ενός αριθμού στο δυαδικό σύστημα (η οποία εύκολα γενικεύεται στο b-αδικό). Σύμφωνα με τη γενικότερη μετατροπή ενός δυαδικού σε δεκαδικό αριθμό, για κάθε μία από τις δέκα θέσεις που περιέχει ψηφίο 1, αθροίζεται η αντίστοιχη δύναμη του δύο, με αρχική δύναμη τη 2^0 (πρώτη θέση) και τελική τη 2^9 (δέκατη θέση). Όντως παρατηρούμε ότι στη συγκεκριμένη περίπτωση ισχύει πως: 1+2+4+...+512=1023. Παρέχοντας λοιπόν, με τη συνάρτηση *integer.base.b*() του Κώδικα 3.1, τη δυαδική κωδικοποίηση όλων των αριθμών από το 1 μέχρι το $2^p - 1$ και προσθέτοντας το μηδενικό διάνυσμα (που αντιστοιχεί στο μοντέλο που περιέχει μόνο τη σταθερά) στη πρώτη σειρά του πίνακα $\boldsymbol{\gamma}$ που λαμβάνουμε, έχουμε στα χέρια μας όλα τα πιθανά μοντέλα που χρειάζεται να εξερευνήσουμε.

Κατασκευάζουμε στην R τη συνάρτηση *best.models*, Κώδικας 3.2, η οποία δεχόμενη ως ορίσματα τα *X.diab* και *Y.diab*, πραγματοποιεί την πλήρη εξερεύνηση του χώρου των πιθανών μοντέλων που χρειαζόμαστε. Υπολογίζει τα κριτήρια πληροφορίας AIC και BIC (το BIC από το πακέτο *nlme*) για κάθε μοντέλο και μας επιστρέφει μια λίστα δύο στοιχείων: τα μοντέλα σε κατάταξη βάσει του AIC και τα μοντέλα σε κατάταξη βάσει του BIC. Για αναλυτική περιγραφή των κριτηρίων πληροφορίας AIC και BIC βλέπε Παράγραφο 3.1.

```
1  best.models<-function(X,Y){
2    g<-integer.base.b(1:(2^ncol(X)-1)) #return all the
       possible models
3    g<-as.matrix(g)
4    g<-rbind(rep(0,ncol(X)),g)        #include the null
       model
5    AIC<-rep(NA,nrow(g))
6    BIC<-rep(NA,nrow(g))
7    reg0<-lm(Y~1,data=X) #null regression
8    AIC[1]<-AIC(reg0)
9    BIC[1]<-BIC(reg0)
10   for(i in 2:nrow(g)){
11     ones<-which(g[i,] %in% 1) #detect the positions of
         the variables included
12     data<-X[ones] #dataset of the variables included in
         each model
13     reg<-lm(Y~.,data=data)
14     AIC[i]<- AIC(reg)
15     BIC[i]<- BIC(reg)
16   }
17   models<-data.frame(g,AIC,BIC)
18   models.aic<-models[order(models$AIC),] # best models
       according aic
19   models.bic<-models[order(models$BIC),] # best models
       according bic
20   best.models<-list(models.aic,models.bic)
21   return(best.models)
22 }
```

Κώδικας 3.2: Η συνάρτηση best.models() πραγματοποιεί την πλήρη εξερεύνηση του χώρου όλων των πιθανών μοντέλων. Δέχεται ως ορίσματα τον πίνακα των επεξηγηματικών μεταβλητών και το διάνυσμα της μεταβλητής απόκρισης. Επιστρέφει τα καλύτερα μοντέλα σε κατάταξη βάσει των κριτηρίων πληροφορίας AIC και BIC.

Με την εντολή $which(g[i,]\%in\%1)$, στην γραμμή 11 του Κώδικα 3.2, λαμβάνουμε τις θέσεις στις οποίες βρίσκεται το 1 στη σειρά i του πίνακα g, ο οποίος περιέχει όλα τα πιθανά μοντέλα σε κωδικοποίηση στο δυαδικό σύστημα. Άρα αναγνωρίζουμε ποιές θα είναι οι επεξηγηματικές μας μεταβλητές κάθε φορά. Χρησιμοποιούμε εν συνεχεία αυτές τις θέσεις, δηλαδή τις αντίστοιχες επεξηγηματικές μεταβλητές, για να προσαρμόσουμε κάθε φορά το μοντέλο και να υπολογίσουμε τα κριτήρια πληροφορίας. Εκτελούμε λοιπόν τη συνάρτηση που κατασκευάσαμε για τα δεδομένα που εξετάζουμε και αποθηκεύουμε τα αποτελέσματά της στο *diab.models*, Κώδικας 3.3, και αυτά με τη σειρά τους τα

χωρίζουμε στο *diab.aic* (τα καλύτερα μοντέλα σε κατάταξη βάσει του AIC) και στο *diab.bic* (τα καλύτερα μοντέλα σε κατάταξη βάσει του BIC).

```
diab.models<-best.models(X.diab,Y.diab)
diab.aic<-diab.models[[1]]
names(diab.aic)<-c(names(X.diab),"AIC","BIC")
diab.bic<-diab.models[[2]]
names(diab.bic)<-c(names(X.diab),"AIC","BIC")
```

Κώδικας 3.3: Εκτελούμε τη συνάρτηση best.models του Κώδικα 3.2 για τα δεδομένα των ασθενών που πάσχουν από διαβήτη και εκχωρούμε τα αποτελέσματα στο diab.models και έπειτα τις κατατάξεις των μοντέλων στα diab.aic και diab.bic .

```
> head(diab.aic)
    AGE SEX BMI BP S1 S2 S3 S4 S5 S6      AIC      BIC
499   0   1   1  1  1  1  0  0  1  0 4790.603 4823.334
503   0   1   1  1  1  1  0  1  1  0 4791.320 4828.142
500   0   1   1  1  1  1  0  0  1  1 4791.374 4828.196
487   0   1   1  1  1  0  0  1  1  0 4791.917 4824.648
491   0   1   1  1  1  0  1  0  1  0 4792.122 4824.852
504   0   1   1  1  1  1  0  1  1  1 4792.241 4833.154
```

Εικόνα 3.3: Τα 6 καλύτερα μοντέλα βάσει του κριτηρίου πληροφορίας AIC

Στην Εικόνα 3.3 παραθέτουμε τα 6 καλύτερα μοντέλα βάσει του κριτηρίου πληροφορίας AIC. Το καλύτερο μοντέλο, με αύξοντα αριθμό 499, περιλαμβάνει τις συμμεταβλητές SEX, BMI, BP, $S1$, $S2$ και $S5$. Μάλιστα, όλες αυτές οι επεξηγηματικές μεταβλητές, εκτός των $S2$ και $S5$ εμπεριέχονται και στα 6 καλύτερα μοντέλα βάσει του AIC. Στην Εικόνα 3.4 βλέπουμε τα 6 καλύτερα μοντέλα βάσει του κριτηρίου πληροφορίας BIC. Το καλύτερο μοντέλο, βάσει του BIC, με αύξοντα αριθμό 459, περιλαμβάνει τις συμμεταβλητές SEX, BMI, BP, $S3$ και $S5$, και όλες αυτές οι συμμεταβλητές πλην της $S3$ περιλαμβάνονται και στα 6 καλύτερα μοντέλα. Τα κοινά μοντέλα ανάμεσα στα 6 καλύτερα των δύο κατατάξεων είναι αυτά με αύξοντα αριθμό 499, 487 και 491. Το μοντέλο 499 είναι το καλύτερο βάσει του AIC και το δεύτερο καλύτερο βάσει του BIC. Τα άλλα δύο κοινά μοντέλα περιέχουν τις συμμεταβλητές SEX, BMI, BP, $S1$, και $S5$. Το μοντέλο 487 περιέχει και την $S4$ ενώ το μοντέλο 491 αντί για την $S4$ περιέχει την $S3$.

```
> head(diab.bic)
    AGE SEX BMI BP S1 S2 S3 S4 S5 S6      AIC      BIC
459   0   1   1  1  0  0  1  0  1  0 4794.264 4822.903
499   0   1   1  1  1  1  0  0  1  0 4790.603 4823.334
487   0   1   1  1  1  0  0  1  1  0 4791.917 4824.648
491   0   1   1  1  1  0  1  0  1  0 4792.122 4824.852
475   0   1   1  1  0  1  1  0  1  0 4793.089 4825.819
463   0   1   1  1  0  0  1  1  1  0 4795.177 4827.907
```

Εικόνα 3.4: Τα 6 καλύτερα μοντέλα βάσει του κριτηρίου πληροφορίας BIC

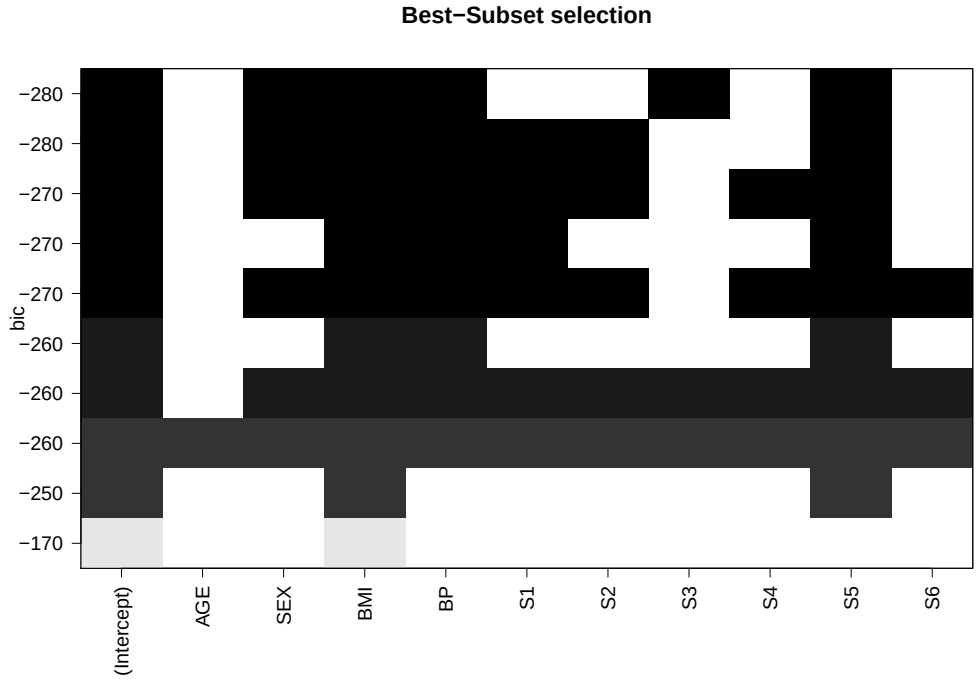
Επιλογή του καλύτερου υποσυνόλου των συμμεταβλητών

Η διαδικασία της επιλογής του καλύτερου υποσυνόλου των επεξηγηματικών μας μεταβλητών, που περιγράψαμε στην Παράγραφο 3.2, εφαρμόζεται στην R από το πακέτο *leaps* με την εντολή $regsubsets()$ και το όρισμα μεθόδου "$exhaustive$".

```
1 install.packages("leaps")
2 library(leaps)
3 best.sub.diab<-regsubsets(as.vector(Y.diab)~.,data=X.
    diab, nbest=1, nvmax=ncol(X.diab),intercept=TRUE,
    method="exhaustive")
4 plot(best.sub.diab,scale=c("bic"),main="Best-Subset
    selection")
```

Κώδικας 3.4: Εκτελούμε την διαδικασία της επιλογής του καλύτερου υποσυνόλου με την εντολή $regsubsets()$ που παίρνει ως ορίσματα τα δεδομένα, τη μέθοδο $method =$ "$exhaustive$", τον μέγιστο αριθμό μεταβλητών $nvmax$ για τα μοντέλα που θα εξετάσει και το πόσα μοντέλα θα κρατήσει από κάθε διαφορετικό πλήθος μεταβλητών $nbest$. Κατασκευάζουμε επίσης ένα γράφημα των αποτελεσμάτων που κατατάσσει τα μοντέλα βάσει του κριτηρίου πληροφορίας BIC.

Στην Εικόνα 3.5 βλεπουμε την κατάταξη των καλύτερων, από όλα τα διαφορετικά πλήθη επεξηγηματικών μεταβλητών, μοντέλων βάσει του κριτηρίου BIC και μάλιστα πιο συγκεκριμένα βάσει της διαφοράς της τιμής του BIC κάθε μοντέλου από την τιμή του BIC του μοντέλου που περιλαμβάνει μόνο τη σταθερά. Με σκούρο χρώμα δηλώνονται οι μεταβλητές που εμπεριέχονται στο εκάστοτε μοντέλο. Καλύτερο αναδεικνύεται αυτό που περιέχει τις επεξηγηματικές μεταβλητές SEX, BMI, BP, $S3$ και $S5$. Βλέπουμε λοιπόν ότι στο παρόν πρόβλημα, η διαδικασία του $Best-Subset\ selection$ συμφωνεί απόλυτα με την πλήρη εξερεύνηση του χώρου των μοντέλων που διεξάγαμε, καθώς βάσει του ίδιου κριτηρίου, επιλέγει το ίδιο καλύτερο μοντέλο.



Εικόνα 3.5: Κατάταξη των μοντέλων της διαδικασίας επιλογής του καλύτερου υποσυνόλου επεξηγηματικών μεταβλητών βάσει του κριτηρίου πληροφορίας BIC.

Διαδικασίες της διαδοχικής πρόσθεσης συμμεταβλητών, της διαδοχικής αφαίρεσης συμμεταβλητών και η διαδικασία κατά βήματα

Οι διαδικασίες της διαδοχικής πρόσθεσης σμμεταβλητών, της διαδοχικής αφαίρεσης συμμεταβλητών και η διαδικασία κατά βήματα που περιγράψαμε στις παραγράφους 3.4, 3.5 και 3.6 αντιστοίχως, εφαρμόζονται στην R με την εντολή $step()$ με ορίσματα το μοντέλο παλινδρόμησης από το οποίο ξεκινάνε και κατεύθυνση είτε "$backward$" είτε "$forward$" είτε "$both$", ανάλογα το ποια διαδικασία εφαρμόζουμε. Στον Κώδικα 3.5 εφαρμόζουμε το μοντέλο παλινδρόμησης $reg.diab$ με όλες τις επεξηγηματικές μεταβλητές και το μοντέλο παλινδρόμησης $reg.diab.null$ που περιλαμβάνει μόνο τη σταθερά. Από το πλήρες μοντέλο θα ξεκινήσει η διαδικασία της διαδοχικής αφαίρεσης συμμεταβλητών με όρισμα $direction =$ "$backward$" (γραμμή 3 Κώδικας 3.5) ενώ από το μοντέλο που περιλαμβάνει μόνο τη σταθερά η διαδικασία της διαδοχικής πρόσθεσης συμμε-

ταβλητών με όρισμα $direction$ = "$forward$" (γραμμή 4 Κώδικας 3.5) και η διαδικασία κατά βήματα με όρισμα $direction$ = "$both$" (γραμμή 5 Κώδικας 3.5). Οι διαδικασίες που ξεκινούν από το μοντέλο που περιλαμβάνει μόνο τη σταθερά απαιτούν και το όρισμα $scope = list(lower = reg.diab.null, upper = reg.diab)$ που ορίζει το εύρος των μοντέλων που θα εξετασθούν. Επιλέγουμε για κριτήριο σύγκρισης το BIC, οπότε βάζουμε σε όλες τις εντολές το επιπλέον όρισμα $k = log(nrow(X.diab))$ για να μετατρέψουμε το κριτήριο σύγκρισης από AIC σε BIC. Δυστυχώς στα αποτελέσματα συνεχίζει και αναγράφεται ως AIC και επίσης πολλές φορές ο υπολογισμός του αποκλίνει από τις τιμές των κριτηρίων πληροφορίας που υπολογίζουν άλλα πακέτα της R. Παρόλα αυτά, τα αποτελέσματα των καλύτερων μοντέλων δεν επηρεάζονται.

```
1 reg.diab<-lm(Y.diab~.,data=X.diab)
2 reg.diab.null<-lm(Y.diab~1,data=X.diab)
3 step(reg.diab,direction="backward",k=log(nrow(X.diab)))
4 step(reg.diab.null, scope=list(lower=reg.diab.null,upper
    =reg.diab),direction="forward"),k=log(nrow(X.diab))
5 step(reg.diab.null, scope=list(lower=reg.diab.null,upper
    =reg.diab),direction="both",k=log(nrow(X.diab)))
```

Κώδικας 3.5: Εκτελούμε τις διαδικασίες της διαδοχικής αφαίρεσης, της διαδοχικής πρόσθεσης και τη διαδικασία κατά βήματα.

```
Step:  AIC=3562.9
Y.diab ~ SEX + BMI + BP + S1 + S2 + S5
```

Εικόνα 3.6: Το μοντέλο στο οποίο κατέληξε η Διαδοχική Αφαίρεση.

```
Step:  AIC=3562.9
Y.diab ~ BMI + S5 + BP + S1 + SEX + S2
```

Εικόνα 3.7: Το μοντέλο στο οποίο κατέληξε η Διαδοχική Πρόσθεση.

Από τις Εικόνες 3.6, 3.7 και 3.8 παρατηρούμε ότι και οι 3 διαδικασίες κατέληξαν στο ίδιο μοντέλο βάσει του κριτηρίου BIC, που περιλαμβάνει τις επεξηγηματικές μεταβλητές SEX, BMI, BP, $S1$, $S2$ και $S5$. Δεν κατάφεραν δηλαδή να

```
Step:  AIC=3562.9
Y.diab ~ BMI + S5 + BP + S1 + SEX + S2
```

Εικόνα 3.8: Το μοντέλο στο οποίο κατέληξε η Διαδικασία κατά βήματα.

βρουν το βέλτιστο μοντέλο βάσει του BIC όπως το εντόπισε η διαδικασία του $Best-Subset\ selection$. Παρόλα αυτά να θυμηθούμε ότι το μοντέλο με τις εν λόγω επεξηγηματικές μεταβλητές, είχε αναδειχθεί δεύτερο καλύτερο βάσει του κριτηρίου BIC στην πλήρη εξερεύνηση του χώρου των πιθανών μοντέλων και βέλτιστο όταν χρησιμοποιήσαμε το κριτήριο πληροφορίας AIC. Άρα μπορούμε να καταλήξουμε στο συμπέρασμα ότι καλό θα ήταν να εξετάσουμε περαιτέρω τα δύο μοντέλα:

Μικρότερο AIC : $Y.diab \sim SEX + BMI + BP + S1 + S2 + S5$

Μικρότερο BIC : $Y.diab \sim SEX + BMI + BP + S3 + S5$

που έχουν ξεχωρίσει ως τα βέλτιστα στις μέχρι τώρα προσεγγίσεις μας και ει δυνατόν να τα συγκρίνουμε με κάποιο άλλο μέτρο που δεν έχουμε χρησιμοποιήσει μέχρι τώρα. προχωρώντας λοιπόν στο Κεφάλαιο 4, θα μελετήσουμε την πιο διαδεδομένη μεθοδολογία σύγκρισης και ελέγχου καλής προσαρμογής μοντέλων, αυτή του $Cross\ Validation$, της οποίας τις παραλλαγές θα χρησιμοποιήσουμε για να συγκρίνουμε τα δύο εν λογω μοντέλα.

4

Cross Validation

Μια πολύ διαδεδομένη μεθοδολογία για την εξέταση καλής προσαρμογής ενός μοντέλου στα δεδομένα, είναι η μέθοδος του Cross Validation. Αποτελεί ένα μέτρο αξιολόγησης, του πόσο καλά προσαρμόζονται τα αποτελέσματα που προέκυψαν από το εκάστοτε στατιστικό μοντέλο που εξετάζεται, σε καινούργια δεδομένα. Χρησιμοποιείται κυρίως σε μοντέλα που αποσκοπούν στην πρόβλεψη κάποιας τιμής, όπως τα μοντέλα πολλαπλής γραμμικής παλινδρόμησης που εξετάζουμε, ώστε να ελεγχθεί η αποτελεσματικότητά της εκτίμησης αυτής της πρόβλεψης στην πράξη. Κατ' επέκτασιν, μπορεί να χρησιμοποιηθεί και ως εργαλείο σύγκρισης διαφορετικών μοντέλων, ώστε να καταλήξουμε στο βέλτιστο επιθυμητό. Η διαδικασία του Cross Validation του μοντέλου γίνεται εν γένει σε δύο βασικά βήματα. Το πρώτο βήμα είναι αυτό της εκπαίδευσης (*training phase*), όπου εκτιμώνται οι άγνωστες παράμετροι του μοντέλου. Το δεύτερο βήμα είναι αυτό της επικύρωσης (*validation*), όπου γίνεται ο έλεγχος της αποδοτικότητας των εκτιμήσεων που προέκυψαν κατά την εκπαίδευση. Για τον έλεγχο αυτό χρησιμοποιούμε κάποιο στατιστικό μέτρο, όπως για παράδειγμα για τα πολλαπλά γραμμικά μοντέλα το R^2, το AIC, το BIC, ή το στατιστικό $PRESS$. Ιδανικά για τα βήματα της εκπαίδευσης και της επικύρωσης, πρέπει να χρησιμοποιήσουμε δυο διαφορετικά σετ δεδομένων, καθώς η χρήση των ιδίων δεδομένων δύο φορές στη στατιστική δεν ενδείκνυται. Αυτό όμως αποτελεί πρόβλημα διότι η συλλογή καινούργιων δεδομένων μπορεί να αποτελέσει μια χρονοβόρα, κοστοβόρα εως και αδύνατη διαδικασία. Αντ' αυτού λοιπόν το Cross Validation χωρίζει τα δεδομένα σε ένα *training set* και ένα *validation set*. Στη συνέχεια γίνονται πολλοί "γύροι" Cross Validation, ώστε να λάβουμε τους μέσους όρους των αποτελεσμάτων ελέγχου, οι οποίοι θα παρέχουν περισσότερη πληροφορία αφού θα έχουν προκύψει από διαφορετικές διαμερίσεις, με τον ίδιο πάντα τρόπο, του σετ δεδομένων που έχουμε στα χέρια μας. Θα αναλύσουμε τρεις μορφές του Cross Validation: το Leave one-out Cross Validation, το Leave k-out Cross Validation και το K-fold Cross Validation,

δίνοντας έμφαση στην εφαρμογή τους στα πολλαπλά γραμμικά μοντέλα παλινδρόμησης. Θεωρούμε ότι έχουμε n παρατηρήσεις για τη μεταβλητή μεταβλητή απόκρισης $\boldsymbol{Y}$, καθώς και για τις p επεξηγηματικές μεταβλητές $\boldsymbol{X}_j$ $j = 1, .., p$

4.1 Leave one-out Cross Validation

Στο Leave one-out Cross Validation, αφήνοντας εκτός μια παρατήρηση κάθε φορά όπως υποδηλώνει το όνομά του, χρησιμοποιούμε ως *training set* τις $n-1$ υπόλοιπες παρατηρήσεις και ως *validation set* την i-οστή παρατήρηση που έμεινε εκτός από το *training set*. Αρχικά, εκτελούμε την πολλαπλή γραμμική παλινδρόμηση για τις $n-1$ παρατηρήσεις του *training set*, λαμβάνοντας τις εκτιμήσεις των παραμέτρων που μας ενδιαφέρουν, $\hat{\boldsymbol{\beta}}_{-i} = (\hat{\beta}_{0-i}, \hat{\beta}_{1-i}, ..., \hat{\beta}_{p-i})$:

$$\hat{\boldsymbol{\beta}}_{-i} = (\boldsymbol{X}_{-i}^T \boldsymbol{X}_{-i})^{-1} \boldsymbol{X}_{-i}^T \boldsymbol{y}_{-i},$$

όπου $\boldsymbol{X}_{-i}$ και ο πίνακας των τιμών των επεξηγηματικών μεταβλητών χωρίς την i παρατήτρηση και $\boldsymbol{y}_{-i}$ το διάνυσμα των τιμών της μεταβλητής απόκρισης χωρίς την i παρατήρηση. Για το επόμενο βήμα, αυτό του *validation*, χρησιμοποιούμε την εκτίμηση της i-οστής τιμής της μεταβλητής απόκρισης που προέκυψε χωρίς την i-οστή παρατήρηση, έστω $\hat{y}_{i,-i}$, καθώς και τις i-οστές παρατηρήσεις των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης ώστε να υπολογίσουμε το i-οστό εκτιμώμενο σφάλμα:

$$\begin{aligned} r_{i,-i} &= y_i - \hat{y}_{i,-i} \\ &= y_i - \boldsymbol{x}_i \hat{\boldsymbol{\beta}}_{-i} \end{aligned}$$

Κατ' αυτόν τον τρόπο εξετάζουμε την προβλεπτική ικανότητα του μοντέλου για την παρατήρηση i. Αυτή η διαδικασία επαναλαμβάνεται n φορές, δηλαδή μέχρις ότου μείνει κάθε i παρατήρηση εκτός του *training set* ακριβώς μια φορά ή καλύτερα μέχρις ότου κάθε i παρατήρηση αποτελέσει το *validation set* ακριβώς μια φορά. Αθροίζοντας τα τετράγωνα όλων αυτών των n το πλήθος εκτιμώμενων σφαλμάτων για κάθε παρατήρηση, λαμβάνουμε το στατιστικό $PRESS$ (Predicted Residual Error Sum of Squares):

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

Τελικά το Leave one-out Cross Validation για το μοντέλο θα είναι ο μέσος όρος του $PRESS$, δηλαδή:

$$CV_{LeaveOneOut} = \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{y}_{i,-i})^2$$

Βέβαια το i-οστό εκτιμώμενο σφάλμα $r_{i,-i}$ μπορεί να γραφεί και διαφορετικά στην περίπτωση του πολλαπλού γραμμικού μοντέλου. Θα χρησιμοποιήσουμε για αυτόν τον σκοπό τον τύπο $Sherman\text{-}Morisson\text{-}Woodbury$, σύμφωνα με τον οποίο για κάθε αντιστρέψιμο και τετραγωνικό πίνακα $\boldsymbol{A}$ και για κάθε διανύσματα στήλες $\boldsymbol{u}$, $\boldsymbol{v}$, που ικανοποιούν τη σχέση $\mathbf{1} + \boldsymbol{v^T A^{-1} u} \neq \mathbf{0}$, ισχύει ότι

$$(\boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T)^{-\mathbf{1}} = \boldsymbol{A^{-1}} - \frac{\boldsymbol{A^{-1}uv^TA^{-1}}}{\mathbf{1} + \boldsymbol{v^TA^{-1}u}}.$$

Για $\boldsymbol{A} = \boldsymbol{X^TX}$, $\boldsymbol{u} = -\boldsymbol{x}_i$ και $\boldsymbol{v} = \boldsymbol{x}_i$ και παρατηρώντας πως $\boldsymbol{X}_{-i}^T\boldsymbol{X}_{-i} = \boldsymbol{X}^T\boldsymbol{X} - \boldsymbol{x}_i\boldsymbol{x}_i^T = \boldsymbol{A} + \boldsymbol{u}\boldsymbol{v}^T$ και $\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i = h_i$, με h_i τα στοιχεία της διαγωνίου του πίνακα προβολής $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^TX})^{-1}\boldsymbol{X}^T$, προκύπτει ότι

$$(\boldsymbol{X}_{-i}^T\boldsymbol{X}_{-i})^{-1} = (\boldsymbol{X}^T\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{1 - h_i}.$$

Επίσης, παρατηρούμε ότι $\boldsymbol{X}_{-i}^T\boldsymbol{y}_{-i} = \boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{x_i}y_i$.

Άρα το διάνυσμα των εκτιμήσεων των παραμέτρων χωρίς την i-οστή παρατήρηση, $\hat{\boldsymbol{\beta}}_{-i}$, γράφεται ακολούθως:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{-i} &= (\boldsymbol{X}_{-i}^T\boldsymbol{X}_{-i})^{-1}\boldsymbol{X}_{-i}^T\boldsymbol{y}_{-i} \\
&= \left[(\boldsymbol{X}^T\boldsymbol{X})^{-1} + \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T(\boldsymbol{X}^T\boldsymbol{X})^{-1}}{1 - h_i}\right](\boldsymbol{X}^T\boldsymbol{y} - \boldsymbol{x}_iy_i) \\
&= \hat{\boldsymbol{\beta}} - (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_iy_i + \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}}{1 - h_i} - \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_ih_iy_i}{1 - h_i} \\
&= \hat{\boldsymbol{\beta}} - \left[\frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_i}{1 - h_i}\right]\left[y_i(1 - h_i) - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} + h_iy_i\right] \\
&= \hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{x}_ir_i}{1 - h_i}.
\end{aligned}$$

Οπότε το i-οστό εκτιμώμενο σφάλμα γράφεται:

$$\begin{aligned} r_{-i} &= y_i - \boldsymbol{x}_i^T \hat{\boldsymbol{\beta}}_{-i} \\ &= y_i - \boldsymbol{x}_i^T \left[\hat{\boldsymbol{\beta}} - \frac{(\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{x}_i r_i}{1 - h_i} \right] \\ &= r_i + \frac{h_i r_i}{1 - h_i} \\ &= \frac{r_i}{1 - h_i} \\ &= \frac{y_i - \hat{y}_i}{1 - h_i}. \end{aligned}$$

Άρα τελικά το $CV_{LeaveOneOut}$ στο πολλαπλό γραμμικό μοντέλο είναι ίσο και με:

$$CV_{LeaveOneOut} = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - h_i} \right)^2,$$

όπου h_i τα στοιχεία της διαγωνίου του *hat matrix* $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X^T X})^{-1} \boldsymbol{X}^T$. Εδώ να προσθέσουμε πως το $CV_{LeaveOneOut}$ είναι περίπου ίσο με το *Generalized Cross Validation*:

$$CV_{LeaveOneOut} \approx GCV = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - \frac{Tr(\boldsymbol{H})}{n}} \right)^2.$$

Προφανώς κατά τη σύγκριση διαφορετικών μοντέλων, προτιμάμε το μοντέλο με το μικρότερο $CV_{LeaveOneOut}$, από τη στιγμή που η συνάρτηση ελέγχου αναφέρεται σε εκτιμώμενα σφάλματα. Τέλος να αναφέρουμε ότι θα μπορούσαμε επίσης αντί των τετραγώνων των εκτιμώμενων σφαλμάτων να αθροίζαμε τα απόλυτά τους, ή να μετρούσουμε την αποδοτικότητα του μοντέλου με κάποιον άλλο τρόπο αντί των σφαλμάτων.

4.2 Leave k-out Cross Validation

Το Leave k-out Cross Validation αποτελεί μια επέκταση της ιδέας του Leave one-out Cross Validation, αφού χωρίζει τα δεδομένα τυχαία σε ένα *training set* με $n-k$ παρατηρήσεις και σε ένα *validation set* με k παρατηρήσεις, όπου προφανώς ισχύει $k < n$. Συνηθίζεται, χωρίς να είναι απαραίτητο, οι k παρατηρήσεις να είναι περίπου το $\frac{1}{3}$ του δείγματος. Στη συνέχεια εκτιμώνται οι συντελεστές του μοντέλου βάσει των $n-k$ παρατηρήσεων του *training set* και γίνεται το

validation του μοντέλου με αντίστοιχο τρόπο με πριν, υπολογίζοντας το μέσο τετραγωνικό σφάλμα (Mean Squared Error) των εκτιμώμενων σφαλμάτων:

$$MSE = \frac{1}{k}\sum_{i=1}^{k}(y_i - \hat{y}_{i,-k})^2$$

όπου πλέον $\hat{y}_{i,-k}$ είναι η εκτίμηση της τιμής της μεταβλητής απόκρισης στην i παρατήρηση χωρίς τις k παρατηρήσεις που δεν χρησιμοποιήθηκαν για την εκτίμηση των συντελεστών του μοντέλου. Εφόσον επαναλάβουμε τη παραπάνω διαδικασία εώς ότου λάβουμε όλους τους πιθανούς συνδυασμούς «k παρατηρήσεων εκτός του *training set*», οι οποίοι είναι $\binom{n}{k}$=b, το Leave k-out Cross Validation για το μοντέλο θα είναι ο μέσος όρος όλων των διαφορετικών MSE, δηλαδή:

$$CV_{k-out} = \frac{1}{b}\sum_{j=1}^{b} MSE_j.$$

Εδώ πρέπει να καταστήσουμε σαφές ότι όλοι οι πιθανοί συνδυασμοί «k παρατηρήσεων εκτός του *training set*», τις περισσότερες των περιπτώσεων, θα είναι ένας πολύ μεγάλος αριθμός, το οποίο συνεπάγεται τεράστιο υπολογιστικό κόστος και εξ′ αυτού το Leave k-out Cross Validation δεν προτιμάται. Βέβαια, μπορούμε να διενεργήσουμε το Leave k-out Cross Validation για ορισμένους μόνο συνδυασμούς «k παρατηρήσεων εκτός του *training set*», έστω m με $m << \binom{n}{k}$=b, το οποίο είναι υπολογιστικά εφικτό.

4.3 K-fold Cross Validation

Στο K-fold Cross Validation τα δεδομένα αρχικά χωρίζονται τυχαία σε K φακέλους $(folds)$ $F_1, F_2, .., F_k$ και ο κάθε φάκελος περιέχει τον ίδιο αριθμό παρατηρήσεων. Οπότε για n παρατηρήσεις και K φακέλους, θα έχουμε σε κάθε φάκελο $\dfrac{n}{K} = \ell \in \mathbb{N}$ παρατηρήσεις. Έπειτα, επιλέγουμε τυχαία ένα φάκελο του οποίου οι παρατηρήσεις αποτελούν το *validation set*, ενώ όλες οι υπόλοιπες $n - l$ παρατηρήσεις των $K - 1$ φακέλων αποτελούν το *training set*. Στη συνέχεια γίνεται η εκτίμηση των συντελεστών του μοντέλου από τις $n - l$ παρατηρήσεις του *training set* και αυτές χρησιμοποιούνται στον υπολογισμό του μέσου τετραγωνικού εκτιμώμενου σφάλματος των l παρατηρήσεων του φακέλου που επιλέχθηκε στην αρχή ως *validation set*:

$$MSE(F_k) = \frac{1}{\ell}\sum_{i\in F_k}(y_i - \hat{y}_{i,-\ell_k})^2$$

όπου $i \in F_k$ οι δείκτες των $\ell_k = \ell$ παρατηρήσεων που ανήκουν στον φάκελο F_k και $\hat{y}_{i,-l_k}$ η εκτιμώμενη τιμή της y_i που προέκυψε μέσω των εκτιμήσεων των συντελεστών του μοντέλου από το *training set*, δηλαδή από όλα τα δεδομένα χωρίς τις l_k παρατηρήσεις του φακέλου F_k. Η παραπάνω διαδικασία επαναλαμβάνεται Κ φορές, λαμβάνοντας κάθε φορά διαφορετικό φάκελο ως *validation set* και όλους τους υπόλοιπους ως *training set*. Τοιουτοτρόπως όλοι οι φάκελοι στους οποίους χωρίστηκαν τυχαία τα δεδομένα λειτουργούν ως *validation set* ακριβώς μια φορά. Τελικά, το K-fold Cross Validation του μοντέλου είναι ίσο με τον μέσο όρο των μέσων τετραγωνικών εκτιμώμενων σφαλμάτων:

$$CV_{K-fold} = \overline{MSE} = \frac{1}{K}\sum_{k=1}^{K} MSE(F_k.)$$

Προφανώς κατά τη σύγκριση διαφορετικών μοντέλων, επιλέγουμε εκείνο που έχει το μικρότερο CV_{K-fold}. Αξίζει να σημειωθεί ότι αυτή η μορφή του Cross Validation χρησιμοποιείται αρκετά συχνά και ορισμένες φορές η συνάρτηση ελέγχου επιδέχεται και παραλλαγές όπως η ακόλουθη:

$$CV_{K-fold} = \overline{RMSE} = \frac{1}{K}\sum_{k=1}^{K} \sqrt{MSE(F_k)}.$$

Η επιλογή της συνάρτησης ελέγχου έγκειται στη διακριτική ευχέρεια του καθενός. Προφανώς στο K-fold Cross Validation αναφερόμαστε σε αριθμό φακέλων $K \geq 2$, με τις συνηθέστερες επιλογές να είναι είτε 5 φάκελοι είτε 10. Στην περίπτωση που ο ο αριθμός των φακέλων K δεν διαιρεί ακριβώς τον αριθμό των παρατηρήσεων n, οι παρατηρήσεις που περισσεύουν ισομοιράζονται τυχαία σε κάποιους από τους K φακέλους και η διαδικασία που περιγράφτηκε ακολουθείται αναλόγως.

4.4 Cross Validation στην R

Σε αυτήν την ενότητα θα μελετήσουμε πως εφαρμόζονται οι 3 παραλλαγές του $Cross\,Validation$ που μελετήσαμε στην R. Θα το πράξουμε αυτό συγκρίνοντας τα 2 καλύτερα μοντέλα που βρήκαμε, βάσει των κριτηρίων AIC και BIC, με την εξερεύνηση του χώρου όλων των πιθανών μας μοντέλων στο Κεφάλαιο 3. Τα μοντέλα αυτά είναι:

Μικρότερο AIC : $Y.diab \sim SEX + BMI + BP + S1 + S2 + S5$

Μικρότερο BIC : $Y.diab \sim SEX + BMI + BP + S3 + S5$

οπότε ορίζουμε τους αντίστοιχους πίνακες των επεξηγηματικών μας μεταβλητών στον Κώδικα 4.1.

```
Xmodelaic<-X.diab[,c(2,3,4,5,6,9)]
Xmodelbic<-X.diab[,c(2,3,4,7,9)]
```

Κώδικας 4.1: Ορίζουμε τους πίνακες των τιμών των επεξηγηματικών μεταβλητών των 2 μοντέλων που θα συγκρίνουμε

Leave one out CV

Το Leave one out Cross Validation, Παράγραφος 4.1, είναι πολύ εύκολο να εφαρμοστεί, αφού υπάρχει έτοιμη η στατιστική συνάρτηση $press()$, στο πακέτο *asbio*, που δέχεται ως όρισμα τη λίστα που επιστρέφει η προσαρμογή ενός μοντέλου παλινδρόμησης στην *R*. Αρκεί λοιπόν να το διαιρέσουμε με τον αριθμό των παρατηρήσεων και λαμβάνουμε το $CV_{LeaveOneOut}$ που μας ενδιαφέρει.

```
install.packages("asbio")
library(asbio)
l1o.aic<-press(lm(Y.diab~.,data=Xmodelaic))/nrow(
   Xmodelaic)
l1o.bic<-press(lm(Y.diab~.,data=Xmodelbic))/nrow(
   Xmodelbic)
```

Κώδικας 4.2: Υπολογίζουμε το $CV_{LeaveOneOut}$ για τα δύο μοντέλα που συγκρίνουμε.

```
> l1o.aic
[1] 2967.821
> l1o.bic
[1] 2992.415
```

Εικόνα 4.1: Αποτελέσματα του $CV_{LeaveOneOut}$ για τα 2 μοντέλα.

Παρατηρούμε στην Εικόνα 4.1 ότι το $CV_{LeaveOneOut}$ του μοντέλου με το χαμηλότερο AIC είναι μικρότερο, αν και όχι για πολύ, από το $CV_{LeaveOneOut}$ του μοντέλου με το χαμηλότερο BIC. Άρα το μοντέλο με το βέλτιστο AIC έχει καλύτερο σφάλμα πρόβλεψης βάσει της μεθόδου Leave one out Cross Validation και στην προκειμένη περίπτωση είναι προτιμητέο.

Leave k-out CV

Το Leave k-out Cross Validation, Παράγραφος 4.2, όπως έχουμε αναφέρει δεν προτιμάται ως μέθοδος λόγω του μεγάλου υπολογιστικού του κόστος. Θυμίζουμε ότι η διαδικασία πρέπει να επαναληφθεί $\binom{n}{k}$ φορές και συνήθως αυτός ο αριθμός είναι αρκετά μεγάλος. Συγκεκριμένα, στο πρόβλημα των ασθενών που πάσχουν από διαβήτη που εξετάζουμε, η μόνη σχετικά προσιτή, αλλά άσκοπη, επιλογή είναι να διαλέξουμε για $k = 2$ και άρα να χρειαστεί να επαναληφθεί η διαδικασία $\binom{442}{2}$=97461 φορές! Παρόλα αυτά, όπως αναφέραμε, μπορούμε να διενεργήσουμε το Leave k-out Cross Validation για ορισμένους μόνο συνδυασμούς «k παρατηρήσεων εκτός του *training set*», έστω m, με $m << \binom{n}{k}$=b, το οποίο είναι υπολογιστικά εφικτό. Εμείς θα διαλέξουμε για k το $\frac{1}{3}$ των παρατηρήσεών μας, δηλαδή $k = 147$ και θα διαλέξουμε μόνο $m = 1000$ από τους $\binom{442}{147}$=$4.878257 \cdot 10^{120}$ συνδυασμούς «147 παρατηρήσεων εκτός του *training set*» για να πραγματοποιήσουμε το Leave k-out Cross Validation.

```
1  cv.leave.k.out<-function(k=1,X,Y,m=nrow(X)){
2    data<-data.frame(Y,X)
3    mse<-rep(NA,length=m)
4    ind<-replicate(m,sample(1:nrow(data),k,replace=FALSE))
5    for(i in 1:m){
6      trainData<-data[-ind[,m],]
7      testData<-data[ind[,m],]
8      reg<-lm(Y~.,data=trainData)
9      pred<-predict(reg,testData)
10     mse[i]<-sum((testData$Y-pred)^2)/nrow(testData)
11     }
12   m<-mean(mse)
13   return(m)
14 }
```

Κώδικας 4.3: Η συνάρτηση cv.leave.k.out() που υλοποιεί το Leave k-out Cross Validation. Δέχεται ως ορίσματα τις k τιμές που αφήνει εκτός του training set σε κάθε επανάληψη, τον πίνακα των τιμών των επεξηγηματικών μεταβλητών X, τις τιμές της μεταβλητή απόκρισης Y και τον αριθμό m των συνδυασμών $\binom{n}{k}$ για τους οποίους θα εφαρμόσει τη μέθοδο. Επιστρέφει το CV_{k-out}.

Στον Κώδικα 4.3 λοιπόν δίνεται η συνάρτηση *cv.leave.k.out*() που δέχεται ως ορίσματα το πλήθος k των παρατηρήσεων που θα αφήνει εκτός του *training set* σε κάθε επανάληψη, τα δεδομένα των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης και m τον αριθμό των συνδυασμών «k παρατηρήσεων εκτός του *training set*». Υπολογίζει το CV_{k-out}. Με την εντολή *replicate*() στη γραμμή 4 του Κώδικα 4.3, δημιουργούμε τον πίνακα *ind* που έχει m δια-

φορετικές στήλες, στην καθεμία εκ των οποίων περιλαμβάνονται 147 αριθμοί, τυχαία διαλεγμένων, παρατηρήσεων. Θα χρησιμοποιήσουμε τους αριθμούς των παρατηρήσεων της κάθε στήλης του πίνακα ind από μια φορά, ως $validation\ set$, για τον υπολογισμό του CV_{k-out}, των μοντέλων που θέλουμε να εξετάσουμε. Στον Κώδικα 4.4 εκτελούμε τη συνάρτηση $cv.leave.k.out()$ για τα δεδομένα των βέλτιστων μοντέλων βάσει του AIC και του BIC που βρήκαμε στην πλήρη εξερεύνηση του χώρου των μοντέλων μας στο Κεφάλαιο 3, $k = 147$ και $m = 1000$.

```
1 lko.aic<-cv.leave.k.out(k=round(nrow(X.diab)/3),
     Xmodelaic,Y.diab,m=1000)
2 lko.bic<-cv.leave.k.out(k=round(nrow(X.diab)/3),
     Xmodelbic,Y.diab,m=1000)
```

Κώδικας 4.4: Εκτελούμε την συνάρτηση cv.leave.k.out για k=147 και $m = 1000$.

```
> lko.aic
[1] 3057.28
> lko.bic
[1] 3116.268
```

Εικόνα 4.2: Αποτελέσματα του CV_{k-out} για τα 2 μειωμένα σε παρατηρήσεις μοντέλα.

Στην Εικόνα 4.2 παρατηρούμε τα CV_{k-out} για τα 2 μοντέλα που συγκρίνουμε. Και εδώ η προβλεπτική ικανότητα του μοντέλου με το βέλτιστο AIC φαίνεται να είναι καλύτερη αφού έχει μικρότερο σφάλμα πρόβλεψης CV_{k-out}. Παρ'όλα αυτά και πάλι οι τιμές των σφαλμάτων CV δεν αποκλίνουν πολύ μεταξύ τους.

K-fold CV

Για την εφαρμογή του K-fold Cross Validation, που αναπτύξαμε στην Παράγραφο 4.3, κατασκευάζουμε στην R τη συνάρτηση $CV.K.fold()$ που δέχεται ως ορίσματα τον αριθμό των φακέλων K, τον πίνακα των επεξηγηματικών μεταβλητών του μοντέλου που ελέγχουμε X και τη μεταβλητή απόκρισης Y. Υπολογίζει και επιστρέφει το CV_{K-fold} του μοντέλου. Την παραθέτουμε στον Κώδικα 4.5. Με την εντολή $split(x, f)$ χωρίζουμε τα δεδομένα μας $x = data$ σε $groups$ βάσει της διαδικασίας που περιγράφει η συνάρτηση f. Ξεκινάμε ελέγχοντας στην γραμμή 3 του Κώδικα 4.5 αν το υπόλοιπο(%%) της διαίρεσης του αριθμού των παρατηρήσεων με τον αριθμό των φακέλων K είναι 0 ώστε να ισομοιραστούν οι παρατηρήσεις στους φακέλους. Αν είναι 0, τότε η f βάσει της οποίας γίνεται το $split$ είναι η $sample(rep(1 : K, nrow(data)\%/\%K))$, ένα διάνυσμα μήκους όσο και ο αριθμός των παρατηρήσεων, με τιμές 1,2,..,K ίσες

σε πλήθος, ανακατεμένες, που αντιστοιχούν στον φάκελο που θα τοποθετηθεί η κάθε παρατήρηση. Εφόσον ο αριθμός των παρατηρήσεων δεν διαιρείται ακριβώς από τον αριθμό των φακέλων η f βάσει της οποίας γίνεται το *split* είναι η $sample(c(rep(1:K, nrow(data)\%/\%K), seq(1, nrow(data)\%\%K, by = 1)))$ πάλι ένα διάνυσμα μήκους όσο και ο αριθμός των παρατηρήσεων, με τιμές 1,2,..,K ανακατεμένες, αλλά όχι ίσες σε πλήθος πλέον, που αντιστοιχούν στον φάκελο που θα τοποθετηθεί η κάθε παρατήρηση. Αυτό που μόλις περιγράψαμε είναι και το «δύσκολο» κομμάτι της συνάρτησης. Κατά τα άλλα, αφότου χωρίσουμε τις παρατηρήσεις στους φακέλους, προχωράμε κατά τα γνωστά που έχουμε περιγράψει στη θεωρία. Σε κάθε επανάληψη κρατάμε έναν φάκελο ως *validation set* και οι υπόλοιποι ενώνονται και γίνονται το *training set* βάσει του οποίου προσαρμόζεται το μοντέλο. Εν συνεχεία υπολογίζουμε τις προβλέψεις της μεταβλητής απόκρισης για τα δεδομένα του *validation set* και τέλος το μέσο τετραγωνικό εκτιμώμενο σφάλμα του φακέλου που λειτούργησε ως *validation set*. Επαναλαμβάνουμε την διαδικασία μέχρις ότου όλοι οι φάκελοι αποτελέσουν το *validation set* ακριβώς μια φορά, υπολογίζουμε το μέσο όρο των μέσων τετραγωνικών εκτιμώμενων σφαλμάτων και έτσι υπολογίζουμε εν τέλει το CV_{K-fold} του μοντέλου.

```
CV.K.fold<-function(K=5,X,Y){
  data<-data.frame(Y,X)
  if(nrow(data)%%K==0){
    folds<-split(data, sample(rep(1:K, nrow(data)%/%K)))
        #create a list of k folds
  }
  else{
    folds<-split(data, sample(c(rep(1:K, nrow(data)%/%K)
       ,seq(1,nrow(data)%%K,by=1)))) #create a list of k
        folds
  }
  cv<-rep(NA,K)
  for(i in 1:K){
    test<-as.data.frame(folds[[i]])
    ind <- as.integer(row.names(test))
    trainData<-data[-ind,]  #remove the rows of the
       current fold for train
    testData<-data[ind,]   #keep the rows of current
       fold for test
    reg<-lm(Y~.,data=trainData) #regression based on
       train
    pred<-predict(reg,testData) #prediction based on
```

```
        test
17    cv[i]<-sum((test$Y-pred)^2)/nrow(testData) #cv for
        each fold /k
18
19  }
20  m<-mean(cv)
21  return(m)
22 }
```

Κώδικας 4.5: Η συνάρτηση $CV.K.fold()$ δέχεται ως ορίσματα τον αριθμό των φακέλων K, τον πίνακα με τα δεδομένα των επεξηγηματικών μεταβλητών του μοντέλου που ελέγχουμε και τα δεδομένα της μεταβλητή απόκρισης. Επιστρέφει το CV_{K-fold} του μοντέλου.

Αφού λοιπόν περιγράψαμε πως λειτουργεί η συνάρτηση $CV.K.fold()$, την εφαρμόζουμε για να συγκρίνουμε τα 2 μοντέλα που εξετάζουμε. Χρησιμοποιούμε αριθμό φακέλων $K = 10$. Στην Εικόνα 4.3 παρατηρούμε τα CV_{K-fold} των 2 αυτών μοντέλων. Πάλι το μοντέλο με το βέλτιστο AIC δείχνει να έχει καλύτερη ικανότητα πρόβλεψης απότι το «αντίπαλό» του με το βέλτιστο BIC, όμως και πάλι τα CV_{K-fold} δεν απέχουν πολύ μεταξύ τους. Εν γένει λοιπόν, αν έπρεπε να διαλέξουμε ένα εκ των δύο μοντέλων, βασιζόμενοι απόλυτα στον κανόνα του μικρότερου CV_{K-fold}, θα διαλέγαμε το μοντέλο με το βέλτιστο AIC. Αυτό όμως δεν σημαίνει ότι το άλλο μοντέλο δεν ανταποκρίνεται καλά στον σκοπό του: την πρόβλεψη της εξέλιξης της ασθένειας του διαβήτη για κάποιον ασθενή. Μάλιστα όπως έχουμε ήδη αναφέρει στην εισαγωγή του παρόντος Κεφαλαίου, μπορούμε να επιλέξουμε το μοντέλο με το μικρότερο BIC, το οποίο αποτελεί ένα πιο φειδωλό μοντέλο από αυτό με το μικρότερο AIC, με εξίσου καλό σφάλμα πρόβλεψης, άρα και πιο προτιμητέο.

```
> cvk.aic<-CV.K.fold(K=10,Xmodelaic,Y.diab)
> cvk.aic
[1] 2954.827
> cvk.bic<-CV.K.fold(K=10,Xmodelbic,Y.diab)
> cvk.bic
[1] 3017.445
```

Εικόνα 4.3: Αποτελέσματα του CV_{K-fold} για τα 2 μοντέλα που συγκρίνουμε.

5

Ridge

Η παλινδρόμηση *Ridge* αποτελεί την πρώτη εκ των μεθόδων συρρίκνωσης των συντελεστών προς εκτίμηση του πολλαπλού γραμμικού μοντέλου που θα εξετάσουμε. Εισήχθηκε από τους Hoerl and Kennard, 1970. Αποτελεί μια μέθοδο συρρίκνωσης, προς το μηδέν, των συντελεστών των επεξηγηματικών μεταβλητών του μοντέλου μας. Ουσιαστικά δεν αποτελεί μεθοδολογία επιλογής μεταβλητών, αλλά μπορεί να την απλουστεύσει, όπως θα δούμε παρακάτω. Ενδείκνυται η χρησιμοποίησή της σε δεδομένα που «πάσχουν» απο πολυσυγγραμικότητα και άρα ο πίνακας σχεδιασμού $\boldsymbol{X}$ είναι ill-conditioned, δηλαδή έχει μικρές ιδιάζουσες τιμές (βλέπε Παράγραφο 2.1.3). Επίσης είναι αποτελεσματική σε περιπτώσεις όπου $p >> n$. Ο $\boldsymbol{X}$ όπως γνωρίζουμε είναι ο πίνακας που έχει για στήλες τις τιμές των επεξηγηματικών μεταβλητών που χρησιμοποιούμε για το μοντέλο μας. Ας θεωρήσουμε ότι τυποποιούμε αυτές τις τιμές στην ίδια κλίμακα (*centered and scaled*), δηλαδή τις μετατρέπουμε έτσι ώστε να έχουν μέση τιμή 0 ($\sum_{i=1}^{n} x_{ij} = 0 \ \forall j$) και διασπορά 1 ($\sum_{i=1}^{n} x_{ij}^2 = 1 \ \forall j$). Η τυποποίηση γίνεται ούτως ώστε οι επεξηγηματικές μεταβλητές να συνεισφέρουν με το ίδιο μέγεθος στην ποινικοποίηση των συντελεστών τους, για την οποία θα γίνει λόγος παρακάτω, και άρα να είναι πιο άμεσα συγκρίσιμα τα αποτελέσματα. Στην ειδική περίπτωση όπου οι επεξηγηματικές μας μεταβλητές έχουν τις ίδιες μονάδες μέτρησης, η τυποποίησή τους δεν είναι απαραίτητη. Επίσης ας θεωρήσουμε ότι οι τιμές της μεταβλητής απόκρισης, $\boldsymbol{y}$, κεντράρονται, ώστε να έχουν μέση τιμή μηδέν, άρα δεν συμπεριλαμβάνουμε τη σταθερά στο μοντέλο μας. Η μέθοδος των ελαχίστων τετραγώνων (Ordinary Least Squares) μας υποδεικνύει την ελαχιστοποίηση των τετραγώνων των υπολοίπων:

$$\begin{aligned} &\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad \sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 \\ \Longleftrightarrow &\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \end{aligned}$$

ούτως ώστε να αποκομίσουμε τους εκτιμητές ελαχίστων τετραγώνων των συντελεστών των συμμεταβλητών του μοντέλου μας $\hat{\boldsymbol{\beta}}^{OLS}$.

5.1 Ποινικοποίηση της ℓ_2-νόρμας και η εκτιμήτρια Ridge

Η μέθοδος της αμφικλινούς παλινδρόμησης, δηλαδή της παλινδρόμησης *Ridge*, μας υποδεικνύει την ελαχιστοποίηση των τετραγώνων των υπολοίπων, υπό κάποιον περιορισμό του τετραγώνου της ℓ_2-νόρμας, δηλαδή του αθροίσματος των τετραγώνων, των συντελεστών $\boldsymbol{\beta}$ του γραμμικού μας μοντέλου:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad \sum_{i=1}^{n}\Big(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2 \; s.t \; \sum_{j=1}^{p}\beta_j^2 \le t$$

$$\Longleftrightarrow \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}) \; s.t \; \sum_{j=1}^{p}\beta_j^2 \le t$$

Εφόσον οι συντελεστές β_j δεν υπόκειντων σε κάποιο περιορισμό, η ελαχιστοποίηση του RSS θα μπορούσε να οδηγήσει σε «έκρηξη» των εκτιμήσεων σε μη λογικές, μεγάλες τιμές και άρα οι εκτιμήσεις να είχαν μεγάλη διασπορά. Για να ελεγχθεί αυτή η διασπορά, τείθεται ένα άνω φράγμα t στο τετράγωνο της ℓ_2-νόρμας του διανύσματος $\boldsymbol{\beta}$. Η ποινικοποίηση της ℓ_2-νόρμας σε προβλήματα ελαχίστων τετραγώνων είναι γνωστή και ως κανονικοποίηση κατά $Tikhonov$. Τοιουτοτρόπως, κανονικοποιούνται οι συντελεστές με την έννοια του ελέγχου του πόσο μεγάλες τιμές μπορούν να πάρουν στο σύνολό τους.

Το πρόβλημα ελαχιστοποίησης υπό τον περιορισμό *Ridge* που παραθέσαμε, μπορεί να γραφεί ακολούθως, με τη μορφή του ποινικοποιημένου, ως προς την l_2-νόρμα, αθροίσματος τετραγώνων των υπολοίπων:

$$\begin{aligned}\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad PRSS(\boldsymbol{\beta})_{l_2} &= \sum_{i=1}^{n}\Big(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2+\lambda\sum_{j=1}^{p}\beta_j^2 \\ &= \underbrace{\|\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2}_{\text{Loss}}+\lambda\underbrace{\|\boldsymbol{\beta}\|_2^2}_{\text{Penalty}} \\ &= (\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y}-\boldsymbol{X}\boldsymbol{\beta})+\lambda\boldsymbol{\beta}^T\boldsymbol{\beta}.\end{aligned}$$

Το άνω φράγμα t του τετραγώνου της ℓ_2-νόρμας μετατράπηκε με 1-1 αντιστοιχία στην παράμετρο ποινής, ή συρρίκνωσης, λ, με χρήση των πολλαπλασιαστών

Lagrange. Μάλιστα η σχέση που συνδέει τα t και λ είναι αντιστρόφως ανάλογη. Οπότε η παράμετρος ποινής λ ελέγχει και αυτή με τη σειρά της το μέγεθος κανονικοποίησης, δηλαδή το πόσο μεγάλοι θα γίνουν οι συντελεστές των επεξηγηματικών μεταβλητών του μοντέλου μας. Όσο μεγαλύτερη είναι η παράμετρος ποινής λ, τόσο μεγαλύτερη συρρίκνωση των συντελεστών επιτυγχάνεται. Όταν $\lambda = 0$ δεν έχουμε συρρίκνωση και λαμβάνουμε την εκτιμήτρια ελαχίστων τετραγώνων. Καλούμαστε λοιπόν πλέον να ελαχιστοποιήσουμε το $PRSS(\boldsymbol{\beta})_{l_2}$. Το παρών αποτελεί ένα κυρτό πρόβλημα ελαχιστοποίησης και άρα η λύση θα είναι μοναδική.

Ένα πρόβλημα βελτιστοποίησης καλείται κυρτό όταν η συνάρτηση που καλούμαστε να βελτιστοποιήσουμε και οι συναρτήσεις των περιορισμών στους οποίους καλούμαστε να υπακούσουμε, είναι κυρτές συναρτήσεις. Στη συγκεκριμένη περίπτωση γνωρίζουμε ήδη από το Κεφάλαιο 1 πως το $RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ είναι αυστηρώς κυρτό στον $\mathbb{R}^p$ δοθέντος του ότι ο πίνακας $\boldsymbol{X}^T\boldsymbol{X}$ είναι θετικά ορισμένος, το οποίο προκύπτει από το ότι θεωρούμε ότι ο πίνακας $\boldsymbol{X}$ είναι πλήρους βαθμού (αυτό δεν συμβαίνει μόνο στην σχεδόν αδύνατη περίπτωση της τέλειας πολυσυγγραμικότητας). Απομένει λοιπόν να μελετήσουμε τη συνάρτηση του περιορισμού:

$$g(\boldsymbol{\beta}) := \boldsymbol{\beta}^T\boldsymbol{\beta} = \boldsymbol{\beta}^T\boldsymbol{I}_p\boldsymbol{\beta}.$$

Για οποιονδήποτε τετραγωνικό πίνακα $\boldsymbol{A}$ ισχύει πως:

$$\frac{\partial}{\partial \boldsymbol{\theta}}(\boldsymbol{\theta}^T\boldsymbol{A}\boldsymbol{\theta}) = \boldsymbol{A}\boldsymbol{\theta} + \boldsymbol{A}^T\boldsymbol{\theta} = (\boldsymbol{A} + \boldsymbol{A}^T)\boldsymbol{\theta}.$$

Στην περίπτωσή μας λοιπόν έχουμε:

$$\frac{\partial}{\partial \boldsymbol{\beta}}(\boldsymbol{\beta}^T\boldsymbol{I}_p\boldsymbol{\beta}) = (\boldsymbol{I}_p + \boldsymbol{I}_p)\boldsymbol{\beta} = 2\boldsymbol{I}_p\boldsymbol{\beta}$$

και:

$$\frac{\partial^2}{\partial \boldsymbol{\beta}^2}(\boldsymbol{\beta}^T\boldsymbol{I}_p\boldsymbol{\beta}) = 2\boldsymbol{I}_p \ > \ 0$$

το οποίο αποδεικνύει ότι η συνάρτηση περιορισμού είναι αυστηρώς κυρτή στον $\mathbb{R}^p$ και άρα τελικά το πρόβλημα ελαχιστοποίησης είναι αυστηρώς κυρτό στον $\mathbb{R}^p$.

Για να βρούμε λοιπόν τη μοναδική λύση, υπολογίζουμε την μερική παράγωγο του ποινικοποιημένου με την ℓ_2-νόρμα αθροίσματος των τετραγώνων των υπολοίπων ως προς $\boldsymbol{\beta}$:

$$\frac{\partial PRSS(\boldsymbol{\beta})_{\ell_2}}{\partial \boldsymbol{\beta}} = -2\boldsymbol{X}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}) + 2\lambda\boldsymbol{I}_p\boldsymbol{\beta}$$

και καταλήγουμε με τη λύση του παρόντος προβλήματος ελαχιστοποίησης, δηλαδή την εκτιμήτρια $Ridge$ των συντελεστών των επεξηγηματικών μεταβλητών σε κλειστή μορφή:

$$\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

που συνήθως δίνει καλύτερα σφάλματα πρόβλεψης απότι η εκτιμήτρια ελαχίστων τετραγώνων. Για $\lambda > 0$ υφίσταται λύση ακόμη και αν ο πίνακας $\boldsymbol{X}^T\boldsymbol{X}$ δεν είναι αντιστρέψιμος, κάτι το οποίο συμβαίνει στην πολύ σπάνια περίπτωση της τέλειας πολυσυγγραμικότητας όπως έχουμε αναφέρει στο Κεφάλαιο 2.

5.2 Η προσέγγιση της επαύξησης των δεδομένων

Στο πρόβλημα που μελετάμε, το ποινικοποιημένο ως προς την l_2-νόρμα άθροισμα τετραγώνων των υπολοίπων μπορεί να γραφεί και ως εξής:

$$\begin{aligned} PRSS(\boldsymbol{\beta})_{\ell_2} &= \sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda\sum_{j=1}^{p}\beta_j^2 \\ &= \sum_{i=1}^{n}(y_i - \boldsymbol{x}_i\boldsymbol{\beta})^2 + \sum_{j=1}^{p}(0 - \sqrt{\lambda}\beta_j)^2. \end{aligned}$$

Άρα είναι σαν να θεωρούμε p επιπλέον παρατηρήσεις με μηδενικές τιμές για τη μεταβλητή απόκρισης και τιμές $\sqrt{\lambda}$ στη διαγώνιο του πίνακα των επεξηγηματικών μεταβλητών των επιπλέον παρατηρήσεων. Κατ′ αυτόν τον τρόπο ουσιαστικά μπορούμε να προσεγγίσουμε το πρόβλημα με τη μέθοδο των ελαχίστων τετραγώνων και άρα να ελαχιστοποιήσουμε το άθροισμα τετραγώνων

των υπολοίπων για τα εξής πλέον δεδομένα:

$$\boldsymbol{X}_\lambda = \begin{pmatrix} x_{11} & x_{12} & x_{13} & \dots & x_{1p} \\ x_{21} & x_{22} & x_{23} & \dots & x_{2p} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & x_{n3} & \dots & x_{np} \\ \sqrt{\lambda} & 0 & 0 & \dots & 0 \\ 0 & \sqrt{\lambda} & 0 & \dots & 0 \\ 0 & 0 & \sqrt{\lambda} & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & \sqrt{\lambda} \end{pmatrix} \quad \boldsymbol{y}_\lambda = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \\ 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

δηλαδή σε πιο συνοπτική μορφή για τα:

$$\boldsymbol{X}_\lambda = \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda}\boldsymbol{I}_p \end{pmatrix} \quad \boldsymbol{y}_\lambda = \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix}.$$

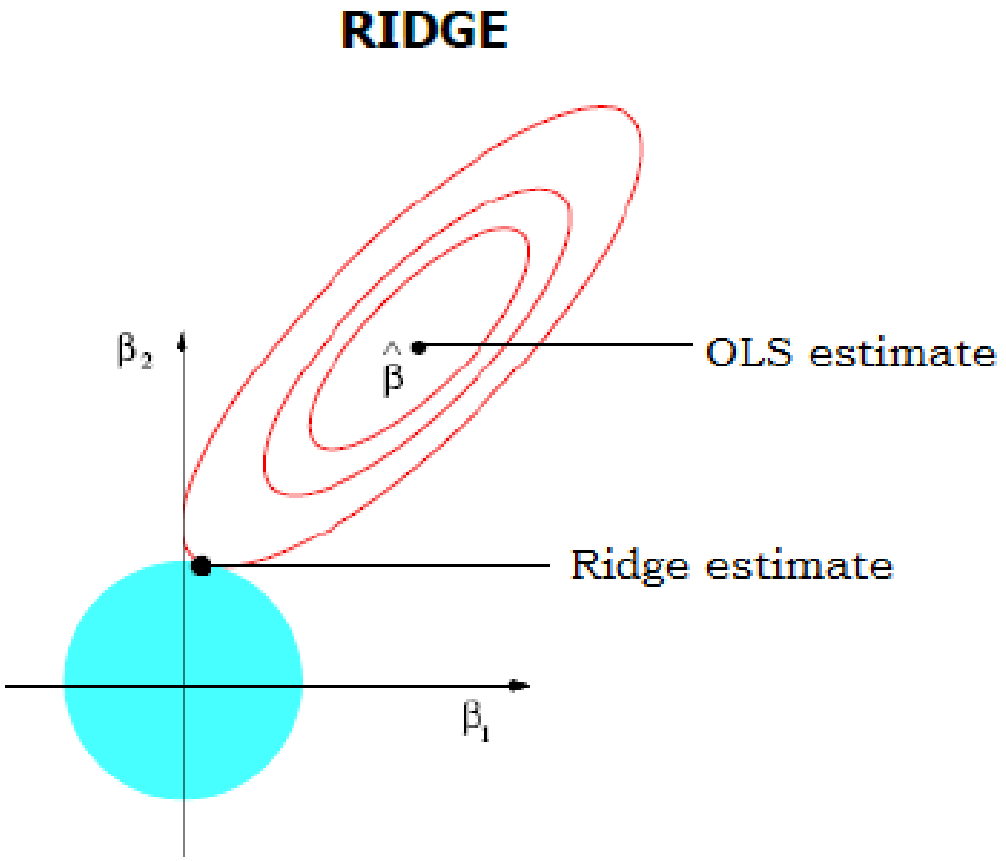
Βάσει της μεθόδου ελαχίστων τετραγώνων, η εκτιμήτρια για τα επαυξημένα δεδομένα θα είναι της μορφής:

$$\begin{aligned} (\boldsymbol{X}_\lambda^T \boldsymbol{X}_\lambda)^{-1} \boldsymbol{X}_\lambda^T \boldsymbol{y}_\lambda &= \left((\boldsymbol{X}^T \sqrt{\lambda}\boldsymbol{I}_p) \begin{pmatrix} \boldsymbol{X} \\ \sqrt{\lambda}\boldsymbol{I}_p \end{pmatrix} \right)^{-1} (\boldsymbol{X}^T \sqrt{\lambda}\boldsymbol{I}_p) \begin{pmatrix} \boldsymbol{y} \\ \boldsymbol{0} \end{pmatrix} \\ &= (\boldsymbol{X}^T \boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} \boldsymbol{X}^T \boldsymbol{y} \\ &= \hat{\boldsymbol{\beta}}_\lambda^{Ridge}. \end{aligned}$$

Παρατηρούμε λοιπόν ότι και με την προσέγγιση των επαυξημένων δεδομένων και χρησιμοποιώντας τη μέθοδο ελαχίστων τετραγώνων καταλήγουμε στον ίδιο ακριβώς τύπο για την εκτιμήτρια $Ridge$ με αυτόν που καταλήξαμε επιλύοντας το κυρτό πρόβλημα ελαχιστοποίησης της παραγράφου 5.1 .

5.3 Γεωμετρία της παλινδρόμησης Ridge

Στην Εικόνα 5.1 παρατηρούμε πως συσχετίζονται οι εκτιμητές OLS με τους



Εικόνα 5.1: Η περιοχή του περιορισμού $\sum_{j=1}^{p} \beta_j^2 \leq t$ *και οι ισοϋψείς ελλείψεις RSS στο πολλαπλό γραμμικό μοντέλο δύο διαστάσεων* ($p = 2$).

εκτιμητές *Ridge* στο πολλαπλό γραμμικό μοντέλο δύο διαστάσεων ($p = 2$). Οι ισοϋψείς ελλείψεις απεικονίζουν το άθροισμα των τετραγώνων των υπολοίπων RSS ($Residual\ Sum\ of\ Squares$) που καλούμαστε να ελαχιστοποιήσουμε ως προς τις παραμέτρους $\boldsymbol{\beta}$. Ο κύκλος απεικονίζει την μπάλα στον $\mathbb{R}^2$ με την ℓ_2-νόρμα, δηλαδή τον περιορισμό που έχουμε για τις παραμέτρους $\boldsymbol{\beta}$. Το εφαπτόμενο σημείο του κύκλου στις ισοϋψείς ελλείψεις αποτελεί την εκτιμήτρια *Ridge*, $\hat{\boldsymbol{\beta}}^{Ridge}$, καθώς ικανοποιεί βέλτιστα και την ελαχιστοποίηση του RSS και τον περιορισμό του αθροίσματος των τετραγώνων των β_j, άρα την ελαχιστοποίηση του $PRSS(\boldsymbol{\beta})_{\ell_2}$. Τα παραπάνω εύκολα γενικεύονται, αλλά δεν απεικονίζονται, στον $\mathbb{R}^p$. Όσο αυξάνεται η διάσταση του προβλήματος p, η μπάλα στον $\mathbb{R}^p$ με την ℓ_2-νόρμα θα είναι μια πολυδιάστατη σφαίρα. Άρα αυξάνεται η πιθανότητα, κατά το πρόβλημα ελαχιστοποίησης που παρουσιάζουμε, να πλησιάσουν σημαντικά ορισμένοι συντελεστές το μηδέν, χωρίς όμως να λάβουν

τη μηδενική τιμή. Η παλινδρόμηση *Ridge* πραγματοποιεί συρρίκνωση των συντελεστών, χωρίς βέβαια να θέτει κάποιον ίσο με μηδέν. Παρόλαυτά δίνει μια πρώτη εικόνα του ποιοί συντελεστές είναι περισσότερο στατιστικά σημαντικοί μετά τη συρρίκνωση, αυτοί δηλαδή που δεν έλαβαν μικρές τιμές, και άρα μπορεί να αποτελέσει ένα πρώτο βήμα σε κάποια διαδικασία επιλογής μεταβλητών.

5.4 Μεροληψία

Θα μελετήσουμε πρώτα την μεροληψία της εκτιμήτριας *Ridge*. Στους παρακάτω υπολογισμούς θεωρούμε ότι ο πίνακας $\boldsymbol{X}^T\boldsymbol{X}$ είναι αντιστρέψιμος, δηλαδή ότι δεν είμαστε αντιμέτωποι με τη σπάνια εως αδύνατη περίπτωση του τέλειου γραμμικού συνδυασμού μεταξύ των στηλών του πίνακα των επεξηγηματικών μας μεταβλητών. Έστω $\boldsymbol{R} = \boldsymbol{X}^T\boldsymbol{X}$, $\boldsymbol{W}_\lambda = (\boldsymbol{R} + \lambda \boldsymbol{I}_p)^{-1}$. Για τον $\boldsymbol{W}_\lambda$ ισχύει πως:

$$\begin{aligned}
\boldsymbol{W}_\lambda &= (\boldsymbol{R} + \lambda \boldsymbol{I}_p)^{-1} \\
&= (\boldsymbol{R}\boldsymbol{I}_p + \lambda \boldsymbol{R}\boldsymbol{R}^{-1})^{-1} \\
&= (\boldsymbol{R}(\boldsymbol{I}_p + \lambda \boldsymbol{R}^{-1})^{-1} \\
&= (\boldsymbol{I}_p + \lambda \boldsymbol{R}^{-1})^{-1}\boldsymbol{R}^{-1} \\
&= \boldsymbol{L}_\lambda \boldsymbol{R}^{-1}
\end{aligned}$$

όπου $\boldsymbol{L}_\lambda = (\boldsymbol{I}_p + \lambda \boldsymbol{R}^{-1})^{-1} = \boldsymbol{W}_\lambda \boldsymbol{R} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}$. Τώρα ας θυμηθούμε ότι $\hat{\boldsymbol{\beta}}^{OLS} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$. Τότε έχουμε:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\lambda^{Ridge} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\
&= (\boldsymbol{R} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{R}(\boldsymbol{R}^{-1}\boldsymbol{X}^T\boldsymbol{y}) \\
&= \boldsymbol{W}_\lambda \boldsymbol{R}[(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}] \\
&= \boldsymbol{W}_\lambda \boldsymbol{R}\hat{\boldsymbol{\beta}}^{OLS} \\
&= \boldsymbol{L}_\lambda \hat{\boldsymbol{\beta}}^{OLS} \\
&= (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}^{OLS}.
\end{aligned}$$

Υπολογίζοντας τώρα την μέση τιμή της εκτιμήτριας *Ridge* και ενθυμούμενοι

ότι η εκτιμήτρια ελαχίστων τετραγώνων ειναι αμερόληπτη, λαμβάνουμε:

$$\begin{aligned} E\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right] &= E\left[\boldsymbol{L}_{\lambda}\hat{\boldsymbol{\beta}}^{OLS}\right] \\ &= \boldsymbol{L}_{\lambda} \cdot E\left[\hat{\boldsymbol{\beta}}^{OLS}\right] \\ &= \boldsymbol{W}_{\lambda}\boldsymbol{R}\boldsymbol{\beta} \\ &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta} \\ &\overset{\lambda\neq 0}{\neq} \boldsymbol{\beta}. \end{aligned}$$

Οπότε η εκτιμήτρια $Ridge$ είναι μεροληπτική για $\lambda \neq 0$ και συγκεκριμένα ισχύει:

$$\begin{aligned} Bias\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right] &= E\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right] - \boldsymbol{\beta} \\ &= (\boldsymbol{L}_{\lambda} - \boldsymbol{I}_p)\boldsymbol{\beta} \\ &= (\boldsymbol{W}_{\lambda}\boldsymbol{R} - \boldsymbol{I}_p)\boldsymbol{\beta} \\ &= (\boldsymbol{W}_{\lambda}\boldsymbol{R} - \boldsymbol{W}_{\lambda}\boldsymbol{W}_{\lambda}^{-1})\boldsymbol{\beta} \\ &= \boldsymbol{W}_{\lambda}(\boldsymbol{R} - \boldsymbol{W}_{\lambda}^{-1})\boldsymbol{\beta} \\ &= \boldsymbol{W}_{\lambda}(\boldsymbol{R} - (\boldsymbol{R} + \lambda\boldsymbol{I}_p))\boldsymbol{\beta} \\ &= -\lambda\boldsymbol{W}_{\lambda}\boldsymbol{\beta} \\ &= -\lambda(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{\beta}. \end{aligned}$$

5.5 Διασπορά

Το γεγονός της έλλειψης αμεροληψίας μας προβληματίζει εκ πρώτης όψεως διότι έχουμε συνηθίσει να χρησιμοποιούμε εν γένει αμερόληπτες εκτιμήτριες. Ας μελετήσουμε λοιπόν και το πίνακα διασπορών συνδιασπορών της εκτιμήτριας $Ridge$, διότι όπως προαναφέραμε, οι συντελεστές ποινικοποιήθηκαν ούτως ώστε να ελεγχθεί το πόσο μεγάλες τιμές θα λάβουν, άρα να ελεγχθεί και η διασπορά τους. Θυμίζοντας ότι $Var\left[\hat{\boldsymbol{\beta}}^{OLS}\right] = \sigma_{\varepsilon}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1}$, $\boldsymbol{L}_{\lambda} = \boldsymbol{W}_{\lambda}\boldsymbol{R}$ καθώς και το ότι για οποιονδήποτε μη τυχαίο πίνακα $\boldsymbol{A}$ ισχύει πως $Var[\boldsymbol{A}\boldsymbol{\theta}] = \boldsymbol{A}Var[\boldsymbol{\theta}]\boldsymbol{A}^T$, υπολογίζουμε το πίνακα διασπορών συνδιασπορών της εκτιμήτριας $Ridge$:

$$
\begin{aligned}
Var\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] &= Var\Big[\boldsymbol{L}_{\lambda}\hat{\boldsymbol{\beta}}^{OLS}\Big] \\
&= \boldsymbol{L}_{\lambda} \cdot Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] \cdot \boldsymbol{L}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{L}_{\lambda}(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}\boldsymbol{L}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{W}_{\lambda}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{R}^{T}\boldsymbol{W}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{W}_{\lambda}\boldsymbol{R}^{T}\boldsymbol{W}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{W}_{\lambda}\boldsymbol{X}^{T}\boldsymbol{X}\boldsymbol{W}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}(\boldsymbol{X}^{T}\boldsymbol{X} + \lambda\boldsymbol{I}_{p})^{-1}\boldsymbol{X}^{T}\boldsymbol{X}[(\boldsymbol{X}^{T}\boldsymbol{X} + \lambda\boldsymbol{I}_{p})^{-1}]^{T}.
\end{aligned}
$$

Επόμενο βήμα είναι να συγκρίνουμε τους δύο πίνακες διασπορών συνδιασπορών, της εκτιμήτριας ελαχίστων τετραγώνων και της εκτιμήτριας $Ridge$, ώστε να εξετάσουμε αν αποκομίζουμε κάποιο όφελος από το ότι η εκτιμήτρια $Ridge$ δεν είναι αμερόληπτη. Για το σκοπό αυτό θα χρησιμοποιήσουμε ότι $Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] = \sigma_{\varepsilon}^{2}(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}$ και για την εκτιμήτρια $Ridge$ τη διασπορά της στη μορφή $Var\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] = \sigma_{\varepsilon}^{2}\boldsymbol{L}_{\lambda}(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}\boldsymbol{L}_{\lambda}^{T}$. Κατ' αυτούς τους συμβολισμούς έχουμε πως:

$$
\begin{aligned}
&Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] - Var\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] \\
&= \sigma_{\varepsilon}^{2}(\boldsymbol{X}^{T}\boldsymbol{X})^{-1} - \sigma_{\varepsilon}^{2}\boldsymbol{L}_{\lambda}(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}\boldsymbol{L}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}(\boldsymbol{R}^{-1} - \boldsymbol{L}_{\lambda}\boldsymbol{R}^{-1}\boldsymbol{L}_{\lambda}^{T}) \\
&= \sigma_{\varepsilon}^{2}(\boldsymbol{L}_{\lambda}\boldsymbol{L}_{\lambda}^{-1}\boldsymbol{R}^{-1}(\boldsymbol{L}_{\lambda}^{T})^{-1}\boldsymbol{L}_{\lambda}^{T} - \boldsymbol{L}_{\lambda}\boldsymbol{R}^{-1}\boldsymbol{L}_{\lambda}^{T}) \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{L}_{\lambda}\big(\boldsymbol{L}_{\lambda}^{-1}\boldsymbol{R}^{-1}(\boldsymbol{L}_{\lambda}^{T})^{-1} - \boldsymbol{R}^{-1}\big)\boldsymbol{L}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{L}_{\lambda}\big([\boldsymbol{I}_{p} + \lambda\boldsymbol{R}^{-1}]\boldsymbol{R}^{-1}[\boldsymbol{I}_{p} + \lambda\boldsymbol{R}^{-1}]^{T} - \boldsymbol{R}^{-1}\big)\boldsymbol{L}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{L}_{\lambda}\big([\boldsymbol{R}^{-1} + \lambda\boldsymbol{R}^{-2}][\boldsymbol{I}_{p} + \lambda\boldsymbol{R}^{-1}]^{T} - \boldsymbol{R}^{-1}\big)\boldsymbol{L}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{L}_{\lambda}\big(2\lambda\boldsymbol{R}^{-2} + \lambda^{2}\boldsymbol{R}^{-3}\big)\boldsymbol{L}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}\boldsymbol{W}_{\lambda}[2\lambda\boldsymbol{I}_{p} + \lambda^{2}\boldsymbol{R}^{-1}]\boldsymbol{W}_{\lambda}^{T} \\
&= \sigma_{\varepsilon}^{2}[\boldsymbol{X}^{T}\boldsymbol{X} + \lambda\boldsymbol{I}_{p}]^{-1}[2\lambda\boldsymbol{I}_{p} + \lambda^{2}(\boldsymbol{X}^{T}\boldsymbol{X})^{-1}][(\boldsymbol{X}^{T}\boldsymbol{X} + \lambda\boldsymbol{I}_{p})^{-1}]^{T}.
\end{aligned}
$$

Υπολογίσαμε πως η διαφορά των πινάκων διασπορών συνδιασπορών των εκτιμητριών ελαχίστων τετραγώνων και $Ridge$ είναι ένας πίνακας πολλαπλασιασμένος με έναν μη αρνητικό αριθμό. Ο πίνακας αυτός είναι μη αρνητικά ορισμένος διότι καθένας από τους πίνακες του γινομένου από το οποίο προκύπτει είναι μη αρνητικά ορισμένος. Άρα ορίζοντας η ανισότητα μεταξύ δύο πινάκων $\boldsymbol{A}$, $\boldsymbol{B}$: $\boldsymbol{B} \geq \boldsymbol{A}$ να συμβολίζει ότι τα στοιχεία του $\boldsymbol{A}$ είναι μικρότερα ή ίσα από τα αντίστοιχα στοιχεία του $\boldsymbol{B}$, καταλήγουμε στο εξής συμπέρασμα:

$$
Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] \geq Var\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big].
$$

Συμπεραίνουμε λοιπόν ότι παρότι υπάρχει έλλειψη αμεροληψίας της εκτιμήτριας *Ridge*, κερδίζουμε ένα συγκριτικό πλεονέκτημα όσον αφορά τον πίνακα διασπορών συνδιασπορών της, ο οποίος θα είναι πάντα μικρότερος ή ίσος στις επιμέρους τιμές του από αυτές του πίνακα διασπορών συνδιασπορών της εκτιμήτριας ελαχίστων τετραγώνων.

Σκόπιμο θα ήταν να αναφερθούμε και στην ολική διασπορά της εκτιμήτριας *Ridge*. Η ολική διασπορά ενός πίνακα διασπορών συνδιασπορών ορίζεται να είναι ίση με το ίχνος του εν λόγω πίνακα. Για να υπολογίσουμε λοιπόν το ίχνος των πινάκων $Var\left[\hat{\boldsymbol{\beta}}^{OLS}\right]$ και $Var\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right]$, θα χρησιμοποιήσουμε αρχικά την παραγοντοποίηση σε ιδιάζουσες τιμές (SVD) του πίνακα $\boldsymbol{X}$, για την οποία έγινε λόγος στην Παράγραφο 2.1.3, και στην συνέχεια το γεγονός ότι το ίχνος ενός πίνακα ισούται με το άθροισμα των ιδιοτιμών του. Βάσει της SVD ο $\boldsymbol{X}$ παραγοντοποιείται ως εξής:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

με τον πίνακα $\boldsymbol{D}$, θυμίζουμε, να περιέχει τις ιδιάζουσες τιμές d_j του $\boldsymbol{X}$ στην κύρια διαγώνιο και τους $\boldsymbol{U}$ και $\boldsymbol{V}$ να είναι ορθογώνιοι, $\boldsymbol{V}^T = \boldsymbol{V}^{-1}$, $\boldsymbol{U}^T = \boldsymbol{U}^{-1}$. Άρα ο πίνακας $\boldsymbol{X}^T\boldsymbol{X}$ γράφεται:

$$\begin{aligned}\boldsymbol{X}^T\boldsymbol{X} &= \boldsymbol{V}\boldsymbol{D}^T\boldsymbol{U}^T\boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T \\ &= \boldsymbol{V}\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}^T\end{aligned}$$

οπότε για τον πίνακα διασπορών συνδιασπορών της εκτιμήτριας ελαχίστων τετραγώνων έχουμε:

$$\begin{aligned}Var\left[\hat{\boldsymbol{\beta}}^{OLS}\right] &= \sigma_{\varepsilon}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma_{\varepsilon}^2(\boldsymbol{V}\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}^T)^{-1} \\ &= \sigma_{\varepsilon}^2\boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D})^{-1}\boldsymbol{V}^T.\end{aligned}$$

Ο πίνακας $Var\left[\hat{\boldsymbol{\beta}}^{OLS}\right]$ είναι διαγωνοποιήσιμος ως προς τον πίνακα $\boldsymbol{V}$ αφού:

$$\boldsymbol{V}^{-1}Var\left[\hat{\boldsymbol{\beta}}^{OLS}\right]\boldsymbol{V} = \sigma_{\varepsilon}^2(\boldsymbol{D}^T\boldsymbol{D})^{-1}$$

ο οποίος είναι ένας διαγώνιος πίνακας μεγέθους $p \times p$ που έχει στη διαγώνιό του τις ιδιοτιμές του πίνακα $Var\left[\hat{\boldsymbol{\beta}}^{OLS}\right]$. Αφού λοιπόν το ίχνος ενός πίνακα είναι ίσο με το άθροισμα των ιδιοτιμών του, για την ολική διασπορά της εκτιμήτριας ελαχίστων τετραγώνων ισχύει πως:

$$tr\Big(Var\left[\hat{\boldsymbol{\beta}}^{OLS}\right]\Big) = \sigma_{\varepsilon}^2\sum_{j=1}^{p}\frac{1}{d_j^2}.$$

Με αντίστοιχο τρόπο εργαζόμαστε για να υπολογίσουμε την ολική διασπορά της εκτιμήτριας *Ridge*. Ο πίνακας $(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}$ μέσω της SVD του $\boldsymbol{X}$ γράφεται:

$$\begin{aligned}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} &= (\boldsymbol{V}\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}^T + \lambda \boldsymbol{I}_p)^{-1} \\ &= \boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{V}^T\end{aligned}$$

άρα κατ' επέκτασιν ο πίνακας διασπορών συνδιασπορών της εκτημήτριας *Ridge* γράφεται ως εξής:

$$\begin{aligned}Var\Big[\hat{\boldsymbol{\beta}}_\lambda^{Ridge}\Big] &= \sigma_\varepsilon^2(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}[(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}]^T \\ &= \sigma_\varepsilon^2\boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^T\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{V}^T \\ &= \sigma_\varepsilon^2\boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{D}^T\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{V}^T.\end{aligned}$$

Ο πίνακας $Var\Big[\hat{\boldsymbol{\beta}}_\lambda^{Ridge}\Big]$ είναι διαγωνοποιήσιμος ως προς τον πίνακα $\boldsymbol{V}$ αφού:

$$\boldsymbol{V}^{-1}Var\Big[\hat{\boldsymbol{\beta}}_\lambda^{Ridge}\Big]\boldsymbol{V} = \sigma_\varepsilon^2(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{D}^T\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}$$

ο οποίος είναι ένας διαγώνιος πίνακας μεγέθους $p \times p$ που έχει στη διαγώνιό του τις ιδιοτιμές του πίνακα $Var\Big[\hat{\boldsymbol{\beta}}_\lambda^{Ridge}\Big]$. Αφού λοιπόν το ίχνος ενός πίνακα είναι ίσο με το άθροισμα των ιδιοτιμών του, για την ολική διασπορά της εκτιμήτριας *Ridge* ισχύει πως:

$$tr\Big(Var\Big[\hat{\boldsymbol{\beta}}_\lambda^{Ridge}\Big]\Big) = \sigma_\varepsilon^2\sum_{j=1}^{p}\frac{d_j^2}{(d_j^2 + \lambda)^2}.$$

Συγκρίνοντας λοιπόν τις ολικές διασπορές των εκτιμητριών ελαχίστων τετραγώνων και *Ridge*, μπορεί εύκολα να δει κανείς πως για κάθε $\lambda \geq 0$ ισχύει ότι:

$$tr\Big(Var\Big[\hat{\boldsymbol{\beta}}_\lambda^{Ridge}\Big]\Big) \leq tr\Big(Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big]\Big),$$

δηλαδή η ολική διασπορά της εκτιμήτριας *Ridge* είναι πάντα μικρότερη ή ίση με την ολική διασπορά της εκτιμήτριας ελαχίστων τετραγώνων.

5.6 Μέσο τετραγωνικό σφάλμα

Στις Παραγράφους 4.4 και 4.5 αναλύσαμε την μεροληψία και τη διασπορά της εκτιμήτριας *Ridge* παρατηρώντας ότι το γεγονός της ύπαρξης μεροληψίας αντισταθμίζεται από το ότι η διασπορά και δη η ολική διασπορά της εκτιμήτριας

Ridge είναι πάντα μικρότερη ή ίση από την ολική διασπορά της εκτιμήτριας ελαχίστων τετραγώνων. Θα υπολογίσουμε τώρα το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge*, δηλαδή την ευκλείδεια απόσταση της εκτιμήτριας *Ridge* από την πραγματική τιμή, καθώς και το μέσο τετραγωνικό σφάλμα της εκτιμήτριας ελαχίστων τετραγώνων, ώστε να προβούμε σε συγκρίσεις. Για τον υπολογισμό της ευκλείδειας απόστασης μιας οποιαδήποτε εκτιμήτριας $\hat{\boldsymbol{\beta}}$ από την πραγματική τιμή $\boldsymbol{\beta}$ θα χρησιμοποιήσουμε αρχικά τον πίνακα συνδιασπορών των σφαλμάτων:

$$\begin{aligned}\boldsymbol{M}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta}) &= E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})^T] \\ &= E[(\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}})+E(\hat{\boldsymbol{\beta}})-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}})+E(\hat{\boldsymbol{\beta}})-\boldsymbol{\beta})^T] \\ &= E[(\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}}))(\hat{\boldsymbol{\beta}}-E(\hat{\boldsymbol{\beta}}))^T]+(E(\hat{\boldsymbol{\beta}})-\boldsymbol{\beta}))(E(\hat{\boldsymbol{\beta}})-\boldsymbol{\beta}))^T \\ &= Var[\hat{\boldsymbol{\beta}}]+(Bias[\hat{\boldsymbol{\beta}}])(Bias[\hat{\boldsymbol{\beta}}])^T\end{aligned}$$

και το γεγονός ότι το μέσο τετραγωνικό σφάλμα μιας εκτιμήτριας $\hat{\boldsymbol{\beta}}$ ισούται με το ίχνος του πίνακα $\boldsymbol{M}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})$:

$$\begin{aligned}MSE[\hat{\boldsymbol{\beta}}] &= tr(\boldsymbol{M}(\hat{\boldsymbol{\beta}},\boldsymbol{\beta})) \\ &= tr\big(Var[\hat{\boldsymbol{\beta}}]\big)+tr\big((Bias[\hat{\boldsymbol{\beta}}])(Bias[\hat{\boldsymbol{\beta}}])^T\big) \\ &= tr\big(Var[\hat{\boldsymbol{\beta}}]\big)+tr\big((Bias[\hat{\boldsymbol{\beta}}])^T(Bias[\hat{\boldsymbol{\beta}}])\big) \\ &= tr\big(Var[\hat{\boldsymbol{\beta}}]\big)+(Bias[\hat{\boldsymbol{\beta}}])^T(Bias[\hat{\boldsymbol{\beta}}]).\end{aligned}$$

Στο προτελευταίο βήμα χρησιμοποιήσαμε το γεγονός ότι το ίχνος ενός πίνακα ισούται με το ίχνος του αναστρόφου του, ενώ στο τελευταίο βήμα το γεγονός ότι το γινόμενο $(Bias[\hat{\boldsymbol{\beta}}])^T(Bias[\hat{\boldsymbol{\beta}}])$ είναι διάστασης 1×1. Άρα το μέσο τετραγωνικό σφάλμα της εκτιμήτριας ελαχίστων τετραγώνων είναι ίσο με:

$$\begin{aligned}MSE\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] &= tr\Big(Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big]\Big)+\Big(Bias\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big]\Big)^T\Big(Bias\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big]\Big) \\ &= \sigma_\varepsilon^2\sum_{j=1}^{p}\frac{1}{d_j^2},\end{aligned}$$

αφού η εκτιμήτρια ελαχίστων τετραγώνων είναι αμερόληπτη. Θυμίζουμε ότι d_j^2 είναι οι ιδιοτιμές του πίνακα $\boldsymbol{X}^T\boldsymbol{X}$. Χρησιμοποιώντας τα αποτελέσματα των παραγράφων 5.4 και 5.5, βρίσκουμε ότι το μέσο τετραγωνικό σφάλμα της

εκτιμήτριας *Ridge* ισούται με:

$$\begin{aligned} MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] &= tr\Big(Var\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]\Big) + \Big(Bias\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]\Big)^T\Big(Bias\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]\Big) \\ &= \sigma_{\varepsilon}^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^2} + (-\lambda \boldsymbol{W}_{\lambda}\boldsymbol{\beta})^T(-\lambda \boldsymbol{W}_{\lambda}\boldsymbol{\beta}) \\ &= \sigma_{\varepsilon}^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^2} + \lambda^2 \boldsymbol{\beta}^T \boldsymbol{W}_{\lambda}^T \boldsymbol{W}_{\lambda}\boldsymbol{\beta} \\ &= \sigma_{\varepsilon}^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^2} + \lambda^2 \boldsymbol{\beta}^T (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-2}\boldsymbol{\beta} \\ &= \gamma_1(\lambda) + \gamma_2(\lambda), \end{aligned}$$

όπου $\gamma_1(\lambda)$ είναι η ολική διασπορά της εκτιμήτριας *Ridge* και $\gamma_2(\lambda)$ η μεροληψία της στο τετράγωνο. Θέλουμε αρχικά να υπολογίσουμε την ποσότητα $\gamma_2(\lambda)$, οπότε θα χρησιμοποιήσουμε την SVD του πίνακα $\boldsymbol{X}$ και συγκεκριμένα το ότι:

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1} = \boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{V}^T.$$

Άρα η ποσότητα $\gamma_2(\lambda)$ είναι ίση με:

$$\begin{aligned} \gamma_2(\lambda) &= \lambda^2 \boldsymbol{\beta}^T (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-2}\boldsymbol{\beta} \\ &= \lambda^2 \boldsymbol{\beta}^T \boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{V}^T\boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{V}^T\boldsymbol{\beta} \\ &= \lambda^2 \boldsymbol{\beta}^T \boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda \boldsymbol{I}_p)^{-2}\boldsymbol{V}^T\boldsymbol{\beta} \\ &= \lambda^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(d_j^2+\lambda)^2} \end{aligned}$$

όπου $\alpha_j \in \boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{\beta}$. Άρα το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge* ισούται με:

$$\begin{aligned} MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] &= \gamma_1(\lambda) + \gamma_2(\lambda) \\ &= \sigma_{\varepsilon}^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^2} + \lambda^2 \sum_{j=1}^{p} \frac{\alpha_j^2}{(d_j^2+\lambda)^2}. \end{aligned}$$

Η ολική διασπορά $\gamma_1(\lambda)$ της εκτιμήτριας *Ridge* είναι μια γνησίως φθίνουσα συνάρτηση του λ αφού:

$$\frac{d\gamma_1(\lambda)}{d\lambda} = -2\sigma_{\varepsilon}^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^3} \; < \; 0, \quad \forall \lambda \geq 0,$$

ενώ η τετραγωνισμένη μεροληψία $\gamma_2(\lambda)$ της εκτιμήτριας $Ridge$ είναι μια γνησίως άυξουσα συνάρτηση του λ αφού:

$$\begin{aligned}\frac{d\gamma_2(\lambda)}{d\lambda} &= 2\lambda\sum_{j=1}^{p}\frac{\alpha_j^2}{(d_j^2+\lambda)^2} - 2\lambda^2\sum_{j=1}^{p}\frac{\alpha_j^2}{(d_j^2+\lambda)^3} \\ &= 2\lambda\sum_{j=1}^{p}\frac{\alpha_j^2 d_j^2 + \lambda\alpha_j^2 - \lambda\alpha_j^2}{(d_j^2+\lambda)^3} \\ &= 2\lambda\sum_{j=1}^{p}\frac{\alpha_j^2 d_j^2}{(d_j^2+\lambda)^3} \;>\; 0, \quad \forall\lambda \geq 0.\end{aligned}$$

Οπότε παρατηρούμε το προαναφερθέν $tradeoff$ μεταξύ μεροληψίας και διασποράς. Όσο το λ λαμβάνει μεγαλύτερες τιμές, η μεροληψία αυξάνεται και η διασπορά μειώνεται. Θέλουμε τώρα να συγκρίνουμε τα μέσα τετραγωνικά σφάλαματα των εκτιμητριών $Ridge$ και OLS. Συγκεκριμένα θα αποδείξουμε ότι $\exists\ \lambda > 0$:

$$MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] < MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda=0}^{Ridge}\Big] = MSE\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] = \sigma_{\varepsilon}^2\sum_{j=1}^{p}\frac{1}{d_j^2}.$$

Το $MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{OLS}\Big]$ είναι μια σταθερή συνάρτηση ως προς λ, παράλληλη στον άξονα των λ, ενώ το $MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]$ μια συνάρτηση του λ που ξεκινάει από την τιμή του $MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{OLS}\Big]$. Οπότε για να αποδείξουμε το ζητούμενο, αρκεί να δείξουμε ότι υπάρχει διάστημα [0,δ) όπου το $MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]$ είναι γνησίως φθίνουσα συνάρτηση του λ:

$$\frac{dMSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]}{d\lambda} < 0.$$

Για την παράγωγο του $MSE\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right]$ ως προς λ έχουμε ότι:

$$\begin{aligned}
\frac{dMSE\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right]}{d\lambda} &= \frac{d\gamma_1(\lambda)}{d\lambda} + \frac{d\gamma_2(\lambda)}{d\lambda} \\
&= -2\sigma_{\varepsilon}^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^3} + 2\lambda \sum_{j=1}^{p} \frac{\alpha_j^2 d_j^2}{(d_j^2+\lambda)^3} \\
&= 2\sum_{j=1}^{p} \frac{d_j^2(\lambda\alpha_j^2 - \sigma_{\varepsilon}^2)}{(d_j^2+\lambda)^3} \\
&\leq 2(\lambda\alpha_{max}^2 - \sigma_{\varepsilon}^2) \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^3} \\
&< 0
\end{aligned}$$

για $\lambda \in \left[0, \dfrac{\sigma_{\varepsilon}^2}{\alpha_{max}^2}\right)$ ενώ όπου α_{max}^2 το τετράγωνο της μεγαλύτερης τιμής $\alpha_j \in \boldsymbol{\alpha} = \boldsymbol{V}^T\boldsymbol{\beta}$. Άρα αποδείξαμε ότι:

$$\exists\, \lambda > 0 : MSE\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right] < MSE\left[\hat{\boldsymbol{\beta}}_{\lambda}^{OLS}\right].$$

Η δυσκολία του παραπάνω αποτελέσματος έγκειται στο γεγονός ότι οι τιμές του λ για τις οποίες το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge* είναι μικρότερο από το μέσο τετραγωνικό σφάλμα της εκτιμήτριας ελαχίστων τετραγώνων, εξαρτώνται από τις ποσότητες σ_{ε}^2 και $\boldsymbol{\beta}$. Άρα, ενώ οι τιμές αυτές του λ υπάρχουν, δεν γνωρίζουμε αν σε συγκεκριμένα πρακτικά προβλήματα ανταποκρίνονται στο επιθυμητό αποτέλεσμα, της εύρεσης μικρότερου μέσου τετραγωνικού σφάλματος.

5.7 Η ειδική περίπτωση του ορθογώνιου πίνακα σχεδιασμού

Ας θεωρήσουμε ότι ο πίνακας σχεδιασμού $\boldsymbol{X}$ είναι ορθογώνιος, δηλαδή:

$$\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p = (\boldsymbol{X}^T\boldsymbol{X})^{-1}.$$

Τότε για την εκτιμήτρια *Ridge* θα ισχύει:

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge} &= (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}^{OLS} \\
&= \frac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{OLS}.
\end{aligned}$$

Σε αυτή την ειδική περίπτωση, γίνεται λίγο πιο κατανοητή η έννοια της συρρίκνωσης των συντελεστών ως προς την παράμετρο ποινής λ. Όσο αυξάνει το λ η εκτιμήτρια $Ridge$ μικραίνει διότι μια μεγάλη τιμή του λ συρρικνώνει την εκτιμήτρια ελαχίστων τετραγώνων. Ο πίνακας διασπορών συνδιασπορών της εκτιμήτριας $Ridge$ όταν έχουμε ορθογώνιο πίνακα σχεδιασμού θα είναι:

$$\begin{aligned} Var\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] &= \sigma_{\varepsilon}^2(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{X}[(\boldsymbol{X}^T\boldsymbol{X}+\lambda\boldsymbol{I}_p)^{-1}]^T \\ &= \sigma_{\varepsilon}^2\boldsymbol{I}_p\frac{1}{(1+\lambda)^2}. \end{aligned}$$

Για κάθε $\lambda \geq 0$, φαίνεται έυκολα ότι θα είναι μικρότερος ή ίσος κατά στοιχείο από τον πίνακα διασπορών συνδιασπορών της εκτιμήτριας ελαχίστων τετραγώνων, όπως έχουμε ήδη αποδείξει στην Παράγραφο 5.5 στη γενική περίπτωση, για ορθογώνιο πίνακα σχεδιασμού:

$$\begin{aligned} Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big] &= \sigma_{\varepsilon}^2(\boldsymbol{X}^T\boldsymbol{X})^{-1} \\ &= \sigma_{\varepsilon}^2\boldsymbol{I}_p. \end{aligned}$$

Το ίδιο θα ισχύει και για την ολική διασπορά της εκτιμήτριας $Ridge$ η οποία για ορθογώνιο πίνακα σχεδιασμού γίνεται ίση με:

$$tr\Big(Var\Big[\hat{\boldsymbol{\beta}}^{Ridge}\Big]\Big) = \frac{p\sigma_{\varepsilon}^2}{(1+\lambda)^2},$$

μικρότερη ή ίση με την ολική διασπορά της εκτιμήτριας ελαχίστων τετραγώνων:

$$tr\Big(Var\Big[\hat{\boldsymbol{\beta}}^{OLS}\Big]\Big) = p\sigma_{\varepsilon}^2.$$

Όσο για το μέσο τετραγωνικό σφάλμα της εκτιμήτριας $Ridge$ στην περίπτωση του ορθογώνιου πίνακα σχεδιασμού έχουμε ότι:

$$MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] = \frac{p\sigma_{\varepsilon}^2}{(1+\lambda)^2} + \frac{\lambda^2}{(1+\lambda)^2}\boldsymbol{\beta}^T\boldsymbol{\beta}.$$

Η ειδική περίπτωση του ορθογώνιου πίνακα σχεδιασμού αναφέρεται διότι αποτελεί την βέλτιστη προσέγγιση συλλογής δεδομένων για ένα πείραμα, με σκοπό τη χρησιμοποίησή τους σε ένα πρόβλημα παλινδρόμησης. Δυστυχώς όμως, πολλές φορές, η πραγματοποίηση μιας τέτοιας προσέγγισης δεν είναι εφικτή. Ακόμα και όταν είναι εφικτό να μετατραπεί ένας πίνακας σχεδιασμού σε ορθογώνιο, με χρήση ορθογώνιων μετασχηματισμών, οι επεξηγηματικές μεταβλητές και οι επιδράσεις τους παύουν να είναι άμεσα ερμηνεύσιμες.

5.8 Βαθμοί ελευθερίας

Οι βαθμοί ελευθερίας, degrees of freedom, είναι ο αριθμός των παραμέτρων προς εκτίμηση στο πολλαπλό γραμμικό μοντέλο. Ας θυμηθούμε ότι για την εκτίμηση των τιμών της μεταβλητής απόκρισης, στην μέθοδο ελαχίστων τετραγώνων ισχύει:

$$\begin{aligned}\hat{\boldsymbol{y}}^{OLS} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{H}\boldsymbol{y},\end{aligned}$$

όπου $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ είναι ο πίνακας προβολής ($hat\ matrix$), ο οποίος είναι συμμετρικός, δηλαδή $\boldsymbol{H}^T = \boldsymbol{H}$ και ταυτοδύναμος, δηλαδή $\boldsymbol{H}\boldsymbol{H} = \boldsymbol{H}$. Ο αριθμός των βαθμών ελευθερίας df του μοντέλου είναι ίσος με τον βαθμό του πίνακα προβολής άρα και με το ίχνος του επειδή είναι ταυτοδύναμος. Οπότε έχουμε πως:

$$df = rank(\boldsymbol{H}) = tr(\boldsymbol{H}).$$

Αναλόγως, στην παλινδρόμηση $Ridge$, για τις εκτιμήσεις της μεταβλητής απόκρισης ισχύει ότι:

$$\begin{aligned}\hat{\boldsymbol{y}}^{Ridge} &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{H}_\lambda\boldsymbol{y}\end{aligned}$$

όπου $\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T$ είναι ο πίνακας προβολής $Ridge$. Οπότε οι βαθμοί ελευθερίας df_λ της παλινδρόμησης $Ridge$, αναλόγως την τιμή της παραμέτρου ποινής λ, θα είναι:

$$df_\lambda = tr(\boldsymbol{H}_\lambda).$$

Για να υπολογίσουμε το ίχνος του πίνακα $\boldsymbol{H}_\lambda$ θα χρησιμοποιήσουμε αρχικά την παραγοντοποίηση σε ιδιάζουσες τιμές (SVD) του πίνακα $\boldsymbol{X}$. Βάσει της SVD έχουμε ότι:

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T$$

και:

$$(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1} = \boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{V}^T.$$

Έτσι λοιπόν τον $\boldsymbol{H}_\lambda$ μπορούμε να τον γράψουμε στην μορφή:

$$\begin{aligned}\boldsymbol{H}_\lambda &= \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T \\ &= \boldsymbol{U}\boldsymbol{D}\boldsymbol{V}^T\boldsymbol{V}(\boldsymbol{D}^T\boldsymbol{D} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{V}^T\boldsymbol{V}\boldsymbol{D}^T\boldsymbol{U}^T \\ &= \boldsymbol{U}\boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{D} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{D}^T\boldsymbol{U}^T.\end{aligned}$$

Ο πίνακας $\boldsymbol{H}_\lambda$ είναι διαγωνοποιήσιμος ως προς τον πίνακα $\boldsymbol{U}$, αφού:

$$\boldsymbol{U}^{-1}\boldsymbol{H}_\lambda\boldsymbol{U} = \boldsymbol{D}(\boldsymbol{D}^T\boldsymbol{D} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{D}^T,$$

ο οποίος είναι ένας διαγώνιος πίνακας μεγέθους $p \times p$ που έχει στη διαγώνιό του τις ιδιοτιμές του πίνακα $\boldsymbol{H}_\lambda$. Αφού λοιπόν το ίχνος ενός πίνακα είναι ίσο με το άθροισμα των ιδιοτιμών του, για τους βαθμούς ελευθερίας df_λ της παλινδρόμησης $Ridge$ θα ισχύει τελικά πως:

$$\begin{aligned} df_\lambda &= tr(\boldsymbol{H}_\lambda) \\ &= \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + \lambda}, \end{aligned}$$

όπου d_j^2 οι ιδιοτιμές του πίνακα $\boldsymbol{X}^T\boldsymbol{X}$. Παρατηρούμε ότι οι βαθμοί ελευθερίας df_λ είναι μια γνησίως φθίνουσα συνάρτηση του λ, κάτι το οποίο φωτογραφίζει από την οπτική των παραμέτρων προς εκτίμηση, την έννοια της συρρίκνωσης μέσω της παραμέτρου ποινής λ. Με λίγα λόγια δηλαδή, όσο μεγαλύτερη η παράμετρος ποινής, τόσο λιγότερες οι παράμετροι προς εκτίμηση άρα και τόσο πιο συρρικνωμένο το μοντέλο (τόσο πιο συρρικνωμένοι οι συντελεστές των επεξηγηματικών μεταβλητών).

5.9 Η παράμετρος ποινής λ και μέθοδοι επιλογής της

Η ποινικοποίηση των συντελεστών των επεξηγηματικών μας μεταβλητών έγινε με τη βοήθεια της παραμέτρου ποινής λ και η επίλυση του προβλήματος ελαχιστοποίησης μας οδήγησε στην εκτιμήτρια $Ridge$:

$$\hat{\boldsymbol{\beta}}_\lambda^{Ridge} = (\boldsymbol{X}^T\boldsymbol{X} + \lambda\boldsymbol{I}_p)^{-1}\boldsymbol{X}^T\boldsymbol{y}.$$

Μπορούμε να παρατηρήσουμε ότι για κάθε λ λαμβάνουμε και διαφορετική εκτιμήτρια, οπότε οι διαφορετικές εκτιμήσεις $Ridge$ για τα διάφορα λ μπορούν να απεικονισθούν στα μονοπάτια κανονικοποίησης, που δεν είναι τίποτα άλλο από το κοινό διάγραμμα των τιμών που λαμβάνουν οι εκτιμώμενοι συντελεστές $\hat{\beta}_j^{Ridge}$ για κάθε λ, συναρτήσει προφανώς των τιμών του λ. Όσο αυξάνει η παράμετρος ποινής, επιτυγχάνουμε μεγαλύτερη συρρίκνωση των συντελεστών των συμμεταβλητών του μοντέλου μας. Σε αυτό το σημείο αξίζει να παραθέσουμε δυο ακραίες περιπτώσεις του παράγοντα ποινής λ και το πώς επιδρούν στην εκτιμήτρια $Ridge$:

$$\begin{aligned} \lambda \downarrow 0 &\iff \hat{\boldsymbol{\beta}}_{\lambda\downarrow 0}^{Ridge} \longrightarrow \hat{\boldsymbol{\beta}}^{OLS} \\ \lambda \uparrow \infty &\iff \hat{\boldsymbol{\beta}}_{\lambda\uparrow\infty}^{Ridge} \longrightarrow 0. \end{aligned}$$

Άρα στην πρώτη περίπτωση, όσο το λ τείνει στο μηδέν, τείνουμε να λάβουμε την εκτιμήτρια ελαχίστων τετραγώνων, ενώ στη δεύτερη περίπτωση, όσο το λ τείνει στο άπειρο, τείνουμε να λάβουμε το μοντέλο που περιλαμβάνει μόνο τη σταθερά. Έχουν προταθεί διάφοροι τρόποι για την επιλογή της κατάλληλης τιμής της παραμέτρου ποινής λ και δυστυχώς όλοι είναι καθαρά υποκειμενικοί.

5.9.1 Κριτήρια πληροφορίας AIC και BIC

Η πρώτη μέθοδος επιλογής της παραμέτρου ποινής λ στην οποία θα αναφερθούμε εξαρτάται από τις τιμές των κριτηρίων πληροφορίας AIC και BIC, που αναπτύξαμε στην Παράγραφο 3.1. Θυμίζουμε για το πολλαπλό γραμμικό μοντέλο ισχύει:

$$AIC = n \cdot \left(ln\Big(\frac{2\pi RSS}{n}\Big) + 1 \right) + 2df$$

$$BIC = n \cdot \left(ln\Big(\frac{2\pi RSS}{n}\Big) + 1 \right) + ln(n)df.$$

Εφόσον κάποιος λοιπόν θέλει να επιλέξει την παράμετρο ποινής λ βασιζόμενος σε ένα από τα δύο κριτήρια πληροφορίας που παραθέσαμε, δεν έχει παρά να υπολογίσει τους βαθμούς ελευθερίας του μοντέλου, Παράγραφος 5.8, για διάφορες τιμές του λ, τις αντίστοιχες τιμές του RSS και άρα στη συνέχεια να υπολογίσει τις αντίστοιχες τιμές του κριτηρίου πληροφορίας που επέλεξε να χρησιμοποιήσει και εν τέλει να επιλέξει την τιμή εκείνη της παραμέτρου ποινής που ελαχιστοποιεί τις τιμές του κριτηρίου πληροφορίας που υπολόγισε.

5.9.2 Δύο συγκεκριμένες προτάσεις

Οι Hoerl, Kennard and Baldwin το 1975 πρότειναν για παράμετρο ποινής $\lambda = k_{HKB}$ την ποσότητα:

$$k_{HKB} = \frac{p\hat{\sigma}_\varepsilon^2}{\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}}$$

όπου p ο αριθμός των επεξηγηματικών μεταβλητών, $\hat{\sigma}_\varepsilon^2$ η εκτίμηση της διασποράς των υπολοίπων από τη μέθοδο ελαχίστων τετραγώνων και $\hat{\boldsymbol{\beta}}$ η εκτιμήτρια ελαχίστων τετραγώνων των συντελεστών των επεξηγηματικών μεταβλητών. Η πρόταση αυτή έγινε με την δικαιολόγηση ότι η τιμή της παραμέτρου ποινής λ που ελαχιστοποιεί το άθροισμα των μέσων τετραγωνικών σφαλμάτων όταν ισχύει πως $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}_p$, είναι ίση με την ποσότητα $\dfrac{p\sigma_\varepsilon^2}{\boldsymbol{\beta}^T\boldsymbol{\beta}}$.

Ο Thisted το 1976 παρατήρησε ότι η συγκεκριμένη τιμή της παραμέτρου ποινής προκαλεί υπερσυρρίκνωση προς το μηδέν των συντελεστών των επεξηγηματικών μεταβλητών και πρότεινε την παραλλαγή της:

$$k_{HKB} = \frac{(p-2)\hat{\sigma}_{\varepsilon}^2}{\hat{\boldsymbol{\beta}}^T \hat{\boldsymbol{\beta}}}$$

η οποία και έχει καθιερωθεί.

Οι Lawless and Wang το 1976 πρότειναν για παράμετρο ποινής $\lambda = k_{LW}$ την ελαφρώς παραλλαγμένη ποσότητα:

$$k_{LW} = \frac{p\hat{\sigma}_{\varepsilon}^2}{\hat{\boldsymbol{\beta}}^T (\boldsymbol{Z}^T \boldsymbol{Z}) \hat{\boldsymbol{\beta}}} = \frac{p\hat{\sigma}_{\varepsilon}^2}{\hat{\boldsymbol{y}}^T \hat{\boldsymbol{y}}}$$

η οποία διορθώθηκε στη συνέχεια, για τον ίδιο λόγο που διορθώθηκε και η προτεινόμενη τιμή k_{HKB}, στην ποσότητα:

$$k_{LW} = \frac{(p-2)\hat{\sigma}_{\varepsilon}^2}{\hat{\boldsymbol{y}}^T \hat{\boldsymbol{y}}}.$$

Κατά καιρούς έχουν γίνει διάφορες προτάσεις για την κατάλληλη επιλογή της παραμέτρου ποινής, εδώ όμως αναφέρουμε μονάχα αυτές που χρησιμοποιούνται κατά κόρον από τα διαθέσιμα πακέτα της R.

5.9.3 Cross Validation

Η πιο διαδεδομένη μεθοδολογία για την επιλογή της παραμέτρου ποινής λ είναι αυτή του Cross Validation την οποία έχουμε περιγράψει στο Κεφάλαιο 4, και δη αυτή του K-fold Cross Validation, Παράγραφος 4.3, με συνάρτηση ελέγχου το μέσο τετραγωνικό σφάλμα (MSE):

$$CV_{K-fold} = \overline{MSE} = \frac{1}{K} \sum_{k=1}^{K} MSE(F_k).$$

Αρκετά συχνά βέβαια χρησιμοποιείται και το Generalized Cross Validation, Παράγραφος 4.1:

$$GCV = \frac{1}{n} \sum_{i=1}^{n} \left(\frac{y_i - \hat{y}_i}{1 - \frac{Tr(\boldsymbol{H})}{n}} \right)^2.$$

Όπως έχουμε προαναφέρει, για κάθε τιμή της παραμέτρου ποινής λ λαμβάνουμε και διαφορετική εκτιμήτρια *Ridge*. Άρα για κάθε διαφορετική εκτιμήτρια *Ridge* λαμβάνουμε και διαφορετικό μοντέλο. Η προσέγγιση λοιπόν είναι να δώσουμε ένα ικανοποιητικά μεγάλο πλήθος τιμών για την παράμετρο ποινής και να εφαρμόσουμε τη μεθοδολογία του Cross Validation, όποια εμείς επιλέξουμε, για καθένα από τα διαφορετικά μοντέλα που θα προκύψουν. Στη συνέχεια επιλέγουμε το μοντέλο εκείνο με το μικρότερο Cross Validation, άρα και την παράμετρο ποινής λ που το παρήγαγε.

5.10 Ridge στην R

Αφού μελετήσαμε αναλυτικά τη θεωρία της παλινδρόμησης *Ridge*, απομένει να δούμε πως εφαρμόζεται και στην πράξη με τη βοήθεια της *R*. Θα συνεχίσουμε να χρησιμοποιούμε τα δεδομένα των ασθενών που πάσχουν από διαβήτη, και θα μελετήσουμε τη συμπεριφορά των διαφορετικών μοντέλων που λαμβάνουμε για διαφορετικές τιμές της παραμέτρου ποινής λ. Για αυτόν τον λόγο δημιουργούμε πρώτα, στον Κώδικα 5.1 μια ακολουθία τιμών της παραμέτρου ποινής τέτοια ώστε να προσεγγίσουμε οποιοδήποτε πιθανό μοντέλο.

```
1 l.diab<-seq(0,2000,length.out=100000)
```

Κώδικας 5.1: Παράγουμε 100000 τιμές της παραμέτρου ποινής λ, από το 0 μέχρι το 2000.

Tradeoff Μεροληψίας - Διασποράς της εκτιμήτριας Ridge

Προτού προσαρμόσουμε τα διαφορετικά μοντέλα που λαμβάνουμε από τις διαφορετικές τιμές της παραμέτρου ποινής, σκόπιμο είναι να προσπαθήσουμε να αναπαραστήσουμε πρώτα γραφικά τις ιδιότητες της εκτιμήτριας *Ridge*. Όπως έχουμε αποδείξει στην Παράγραφο 5.4 η εκτιμήτρια *Ridge* είναι μεροληπτική. Μάλιστα όπως αποδείξαμε στην Παράγραφο 5.6, η τετραγωνισμένη μεροληψία είναι μια γνησίως αύξουσα συνάρτηση της παραμέτρου ποινής λ. Η ύπαρξη μεροληψίας όμως της εκτιμήτριας *Ridge* αντισταθμίζεται από την ολική διασπορά της, η οποία, όπως αποδείξαμε στην Παράγραφο 5.5, είναι πάντα μικρότερη η ίση της ολικής διασποράς της εκτιμήτριας ελαχίστων τετραγώνων. Στην Παράγραφο 5.6 αποδείξαμε ότι η ολική διαπορά της εκτιμήτριας *Ridge* είναι μια γνησίως φθίνουσα συνάρτηση της παραμέτρου ποινής λ. Σκοπός μας είναι να αποτυπώσουμε σε ένα γράφημα τα προαναφερθέντα, άρα και το $tradeoff$ ανάμεσα στην μεροληψία και την ολική διασπορά της εκτιμήτριας *Ridge* καθώς και να

οπτικοποιήσουμε το αρκετά χρήσιμο θεώρημα που αποδείξαμε στην Παράγραφο 5.6 για το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge*:

$$\exists\, \lambda > 0 : MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] < MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{OLS}\Big].$$

Για να επιτύχουμε το σκοπό μας κατασκευάζουμε στον Κώδικα 5.2 τη συνάρτηση $bvmse()$, που δέχεται ως ορίσματα τον πίνακα με τα δεδομένα των επεξηγηματικών μεταβλητών X, τα δεδομένα της μεταβλητής απόκρισης Y και τις τιμές της παραμέτρου ποινής λ. Υπολογίζει και επιστρέφει την ολική διασπορά, την τετραγωνισμένη μεροληψία και το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge*. Θυμίζουμε ότι η ολική διασπορά της εκτιμήτριας *Ridge* είναι:

$$tr\Big(Var\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]\Big) = \sigma_{\varepsilon}^2 \sum_{j=1}^{p} \frac{d_j^2}{(d_j^2+\lambda)^2},$$

οπότε στη γραμμή 2 του Κώδικα 5.2 προσαρμόζουμε το πολλαπλό γραμμικό μοντέλο με την μέθοδο ελαχίστων τετραγώνων για να αποκομίσουμε τις εκτιμήτριες ελαχίστων τετραγώνων των συντελεστών μας, που αποθηκεύουμε στη γραμμή 3 του Κώδικα 5.2 και την εκτίμηση του σ_{ε}^2, της διασποράς των σφαλμάτων του μοντέλου, την οποία αποθηκεύουμε στην γραμμή 6 του Κώδικα 5.2. Στις γραμμές 7,8,9 του Κώδικα 5.2 δημιουργούμε διανύσματα με αγνοούμενες μήκους ίσου με το μήκος του διανύσματος των τιμών της παραμέτρου ποινής για να εκχωρήσουμε σε αυτά στη συνέχεια τις εκτιμήσεις της τετραγωνισμένης μεροληψίας, της ολικής διασποράς και του μέσου τετραγωνικού σφάλματος αντιστοίχως. Στη γραμμή 10 του Κώδικα 5.2 υπολογίζουμε τις ιδιοτιμές d_j^2 του πίνακα $\boldsymbol{X}^T\boldsymbol{X}$ τις οποίες θα χρησιμοποιήσουμε στον υπολογισμό της ολικής διασποράς. Στη γραμμή 11 φτιάχνουμε έναν πίνακα με αγνοούμενες τιμές, με αριθμό στηλών όσες και οι επεξηγηματικές μεταβλητές και αριθμό γραμμών όσες και το μήκος του διανύσματος της παραμέτρου ποινής που παρέχουμε στη συνάρτηση, ώστε να εκχωρήσουμε στη κάθε γραμμή του μετέπειτα τα στοιχεία του αθροίσματος της ολικής διασποράς της εκτιμήτριας *Ridge*. Στη γραμμή 12 του Κώδικα 5.2 ξεκινάμε έναν βρόγχο for για όλα τα λ που έχουμε στη διάθεσή μας, δηλαδή για όλα τα διαφορετικά μοντέλα που προκύπτουν από τα διαφορετικά λ, και υπολογίζουμε τα ακόλουθα: Στη γραμμή 13 του Κώδικα 5.2 υπολογίζουμε τα στοιχεία του αθροίσματος της ολικής διασποράς της εκτιμήτριας *Ridge* και στη γραμμή 14 την ίδια την ολική διασπορά της εκτιμήτριας *Ridge*.

```
1 bvmse<-function(X,Y,l){
2   reg<-lm(Y~.,data=as.data.frame(X))
3   cf<-coef(reg)
```

```
4   p<-ncol(X)
5   Z<-scale(X)
6   s2<-summary(reg)$s^2
7   bias<-rep(NA,length(l))
8   variance<-rep(NA,length(l))
9   mse<-rep(NA,length(l))
10  d<-eigen(t(Z)% * %Z)$values
11  d_down<-matrix(rep(NA,length(l)*p),ncol=p,nrow=length(
        l))
12  for(i in 1:length(l)){
13    d_down[i,]<-d/(d+l[i])^2
14    variance[i]<-s2*sum(d_down[i,])
15    W<-solve(t(Z)% * %Z + l[i]*diag(p))
16    L<-W% * %t(Z)% * %Z
17    bias2<-t(cf[-1])% * %t(L-diag(p))% * %(L-diag(p))% *
          %cf[-1]
18    bias[i]<-bias2
19    mse[i]<-variance[i]+bias2
20  }
21  results<-list(variance,bias,mse)
22  names(results)<-c("variance","bias","mse")
23  return(results)
24 }
```

Κώδικας 5.2: Η συνάρτηση bvmse() που δέχεται ως ορίσματα τον πίνακα με τα δεδομένα των επεξηγηματικών μεταβλητών X, τα δεδομένα της μεταβλητή απόκρισης Y και τις τιμές της παραμέτρου ποινής. Υπολογίζει και επιστρέφει την τετραγωνισμένη μεροληψία, την ολική διασπορά και το μέσο τετραγωνικό σφάλμα της εκτιμήτριας Ridge

Στις επόμενες 3 γραμμές (15,16,17) του Κώδικα 5.2 υπολογίζουμε την τετραγωνισμένη μεροληψία της εκτιμήτριας *Ridge*. Χρησιμοποιούμε την έκφανση της μεροληψίας:

$$Bias\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right] = (\boldsymbol{L}_{\lambda} - \boldsymbol{I}_{p})\boldsymbol{\beta},$$

από την απόδειξη της Παραγράφου 5.4. Το $\boldsymbol{\beta}$ θα το εκτιμήσουμε με την αμερόληπτη εκτιμήτρια ελαχίστων τετραγώνων (γραμμή 3 Κώδικας 5.2) και θυμίζουμε ότι $\boldsymbol{L}_{\lambda} = \boldsymbol{W}_{\lambda}\boldsymbol{X}^{T}\boldsymbol{X}$, με $\boldsymbol{W}_{\lambda} = (\boldsymbol{X}^{T}\boldsymbol{X} + \lambda\boldsymbol{I}_{p})^{-1}$. Έτσι λοιπόν, στη γραμμή 15 του Κώδικα 5.2 υπολογίζουμε τον $\boldsymbol{W}_{\lambda}$, στη γραμμή 16 του Κώδικα 5.2 τον $\boldsymbol{L}_{\lambda}$ και στη γραμμή 17 του Κώδικα 5.2 τη τετραγωνισμένη μεροληψία της εκτιμήτριας *Ridge*:

$$\left(Bias\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right]\right)^{T}\left(Bias\left[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\right]\right) = \boldsymbol{\beta}^{T}(\boldsymbol{L}_{\lambda} - \boldsymbol{I}_{p})^{T}(\boldsymbol{L}_{\lambda} - \boldsymbol{I}_{p})\boldsymbol{\beta}.$$

Τέλος στη γραμμή 19 του Κώδικα 5.2 υπολογίζουμε το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge*:

$$MSE\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big] = tr\Big(Var\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]\Big) + \Big(Bias\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]\Big)^{T}\Big(Bias\Big[\hat{\boldsymbol{\beta}}_{\lambda}^{Ridge}\Big]\Big).$$

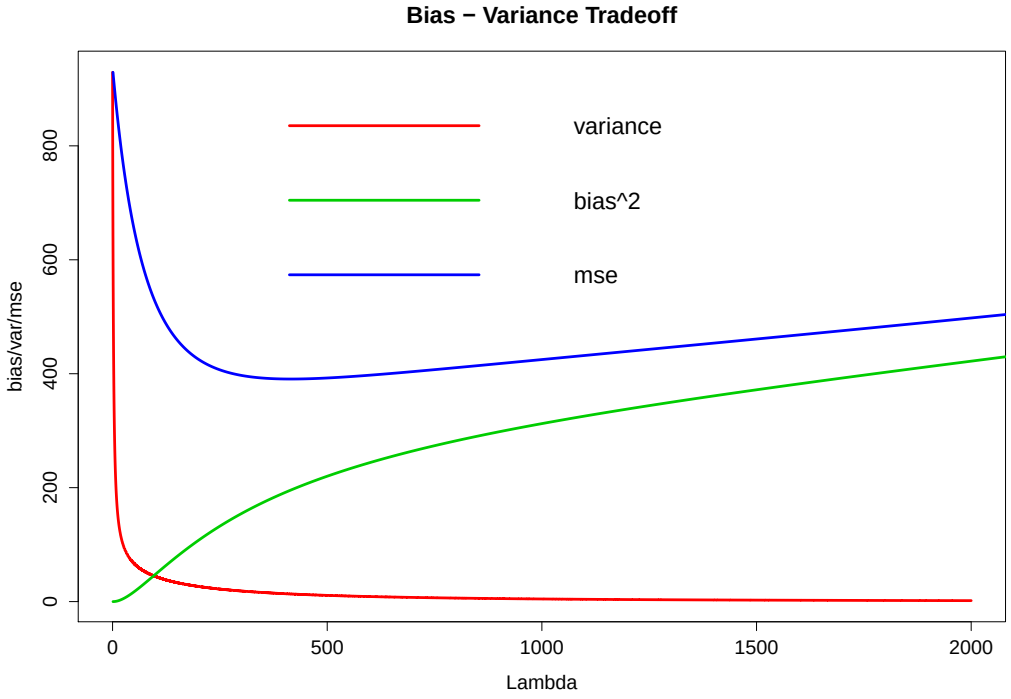
Στην γραμμή 21 του Κώδικα 5.2 εκχωρούμε τα διανύσματα των τιμών της ολικής διασποράς, της τετραγωνισμένης μεροληψίας και του μέσου τετραγωνικού σφάλαματος των εκτιμητριών *Ridge* που υπολογίσαμε για τα διαφορετικά μας μοντέλα, τα οποία και επιστρέφει η συνάρτηση $bvmse()$ σε μορφή λίστας.

```
1 bvm.diab<-bvmse(as.matrix(X.diab),as.matrix(Y.diab),l.
     diab)
2 variance<-bvm.diab[[1]]
3 bias<-bvm.diab[[2]]
4 mse<-bvm.diab[[3]]
5 plot(l.diab,variance,lwd=3,type="l",col=2,main="Bias -
     Variance Tradeoff",xlab="Lambda",ylab="bias/var/mse")
6 lines(bias,type="l",col=3,lwd=3)
7 lines(mse,type="l",col=4,lwd=3)
8 legend("topright",col=c(2,3,4),lwd=3,legend=c("variance"
     ,"bias^2","mse"),cex=1,bty="n")
```

Κώδικας 5.3: Εκτελούμε τη συνάρτηση bvmse(), εκχωρούμε τα αποτελέσματα και κατασκευάζουμε το κοινό γράφηματου μέσου τετραγωνικού σφάλματος, της ολικής διαποράς και της τετραγωνισμένης μεροληψίας.

Στον Κώδικα 5.3 εκτελούμε τη συνάρτηση $bvmse()$ για τα δεδομένα των ασθενών που πάσχουν από διαβήτη και τις τιμές της παραμέτρου ποινής λ που παράξαμε στον Κώδικα 5.1. Αποθηκεύουμε τα αποτελέσματα και κατασκευάζουμε με τις εντολές $plot()$ και $lines()$ το κοινό γράφημα που θέλουμε.

Στην Εικόνα 5.2 φαίνεται το γράφημα που κατασκευάσαμε. Παρατηρούμε πως ενώ η τετραγωνισμένη μεροληψία αυξάνεται όσο αυξάνεται η παράμετρος ποινής λ, η ολική διασπορά της εκτιμήτριας *Ridge* φθίνει αποτυπώνοντας το προαναφερθέν $tradeoff$. Επίσης παρατηρούμε ότι το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge* μετά το $\lambda = 0$ φθίνει μόνο μέχρι ένα σημείο, απεικονίζοντας αυτό που αποδείξαμε στην Παράγραφο 5.6, ότι δηλαδή υπάρχει λ θετικό έτσι ώστε το μέσο τετραγωνικό σφάλμα της εκτιμήτριας *Ridge* να είναι μικρότερο ή ίσο από το μέσο τετραγωνικό σφάλμα της εκτιμήτριας ελαχίστων τετραγώνων.



Εικόνα 5.2: Κοινό γράφημα της τετραγωνισμένης μεροληψίας, της ολικής διασποράς και του μέσου τετραγωνικού σφάλματος της εκτιμήτριας Ridge συναρτήσει της παράμετρου ποινής λ.

Η παλινδρόμηση Ridge

Στην *R* μπορούμε να εφαρμόσουμε την παλινδρόμηση *Ridge* με την εντολή *lm.ridge()* του πακέτου *MASS*, που δέχεται ως ορίσματα το γραμμικό μοντέλο που προσαρμόζουμε, τα δεδομένα μας και τις τιμές της παραμέτρου ποινής που θέλουμε.

```
1 install.packages("MASS")
2 library(MASS)
3 dridge<-lm.ridge(Y.diab~., data=X.diab,lambda=l.diab)
```

Κώδικας 5.4: Η παλινδρόμηση Ridge με την εντολή lm.ridge() του πακέτου MASS. Δέχεται ως ορίσματα το μοντέλο που εφαρμόζουμε, τα δεδομένα μας και τις τιμές της παραμέτρου ποινής. Επιστρέφει μια λίστα το περιεχόμενο της οποίας φαίνεται στην Εικόνα 5.3.

Στον Κώδικα 5.4 εκτελούμε την παλινδρόμηση *Ridge* για τα δεδομένα των ασθενών που πάσχουν από διαβήτη και για τις διαφορετικές τιμές της παρα-

μέτρου ποινής λ που παράξαμε στον Κώδικα 5.1. Στην Εικόνα 5.3 με την εντολή $names()$ βλέπουμε τί περιέχει η λίστα $dridge$ του Κώδικα 5.4, όπου αποθηκεύσαμε την παλινδρόμηση $Ridge$ που εκτελέσαμε. Ανάμεσά τους είναι:

- "$coef$": οι εκτιμήτριες $Ridge$ των συντελεστών όλων των διαφορετικών μοντέλων για όλες τις διαφορετικές παραμέτρου ποινής λ. Αν θέλουμε να λάβουμε τις εκτιμήτριες $Ridge$ για ένα συγκεκριμένο μοντέλο γράφουμε:

$$coef(lm.ridge(Y.diab \sim ., data = X.diab, lambda =))$$

  και όπου $lambda =$, βάζουμε την τιμή της παραμέτρου ποινής που αντιστοιχεί στο μοντέλο του οποίου τους συντελεστές θέλουμε να υπολογίσουμε.

- "GCV": τα Generalized Cross Validation, Παράγραφος 5.9.3, όλων των διαφορετικών μοντέλων για όλες τις διαφορετικές παραμέτρου ποινής λ. Για να βρούμε την παράμετρο ποινής που παράγει το μοντέλο με το μικρότερο Generalized Cross Validation, γράφουμε:

$$dridge\$lambda[dridge\$GCV == min(dridge\$GCV)].$$

- "$kHKB$" και "kLW: οι δύο συγκεκριμένες προτάσεις της παραμέτρου ποινής λ, Παράγραφος 5.9.2 . Τις καλούμε με τις εντολές $dridge\$kHKB$ και $dridge\$kLW$.

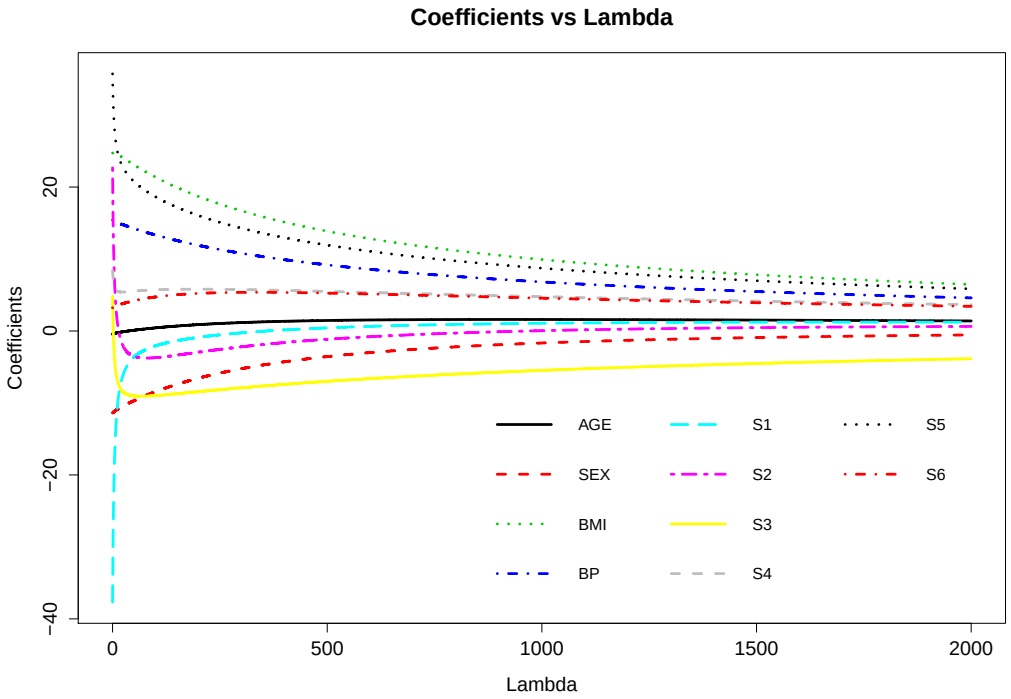
```
> names(dridge)
[1] "coef"   "scales" "Inter"  "lambda" "ym"     "xm"     "GCV"    "kHKB"   "kLW"
```

Εικόνα 5.3: Περιεχόμενα της λίστας dridge του Κώδικα 5.4.

Στον Κώδικα 5.5 κατασκευάζουμε τα μονοπάτια κανονικοποίησης των εκτιμητριών $Ridge$ των συντελεστών των επεξηγηματικών μας μεταβλητών, δηλαδή το κοινό γράφημα των διαφορετικών τιμών που λαμβάνουν οι εκτιμήτριες $Ridge$ για κάθε διαφορετικό λ.

```
1 plot(l.diab,dridge$coef[1,],ylim=range(dridge$coef),type
    ="l",ylab="Coefficients",xlab="Lambda",main="
    Coefficients vs Lambda",lwd=3)
2 for(j in 2:ncol(X.diab)) lines(l.diab,dridge$coef[j,],
    col=j,lwd=3,lty=j)
3 legend("bottomright",legend=paste(names(X.diab),sep=''),
    ncol=3,lwd=3,col=1:ncol(X.diab),bty="n",lty=1:ncol(X.
    diab),cex=0.8)
```

Κώδικας 5.5: Κατασκευή των μονοπατιών κανονικοποίησης των εκτιμητριών Ridge συναρτήσει της παραμέτρου ποινής λ.



Εικόνα 5.4: Μονοπάτια κανονικοποίησης των εκτιμητριών Ridge συναρτήσει της παραμέτρου ποινής λ. Οι εκτιμήτριες Ridge συρρικνώνονται όσο αυξάνει το λ.

Το κοινό αυτό γράφημα των εκτιμητριών *Ridge* συναρτήσει του λ φαίνεται στην Εικόνα 5.4. Παρατηρούμε ότι όσο αυξάνει η παράμετρος ποινής λ, τόσο περισσότερο συρρικνώνονται οι εκτιμήτριες των συντελεστών του μοντέλου μας, χωρίς όμως να μηδενίζονται, ενώ για $\lambda = 0$, έχουμε τις τιμές των εκτιμητριών ελαχίστων τετραγώνων.

Θέλουμε τώρα να υπολογίσουμε τους βαθμούς ελευθερίας, Παράγραφος 5.8, της παλινδρόμησης *Ridge* για όλα τα διαφορετικά μοντέλα που προκύπτουν από τις διαφορετικές τιμές της παραμέτρου ποινής λ καθώς και τα κριτήρια πληροφορίας AIC και BIC, για αυτά τα διαφορετικά μοντέλα. Για τον σκοπό αυτό κατασκευάζουμε τη συνάρτηση $stats()$, Κώδικας 5.6, που δέχεται ως ορίσματα τα δεδομένα των επεξηγηματικών μας μεταβλητών X, τα δεδομένα της μεταβλητής απόκρισης Y και τις τιμές της παραμέτρου ποινής. Υπολογίζει και επι-

στρέφει τους βαθμούς ελευθερίας των διαφορετικών μοντέλων και τα κριτήρια πληροφορίας τους AIC και BIC. Σις γραμμές 2,3,4 ορίζουμε τα διανύσματα με αγνοούμενες τιμές, μήκους όσο και τα διαφορετικά λ, των ποσοτήτων που θα υπολογίσουμε. Έπειτα αποθηκέυουμε τον αριθμό των επεξηγηματικών μεταβλητών $ncol(X)$, τον αριθμό των παρατηρήσεων $nrow(X)$ και τυποποιούμε τα δεδομένα των επεξηγηματικών μεταβλητών, κεντράροντας επίσης τις τιμές της μεταβλητής απόκρισης. Στη γραμμή 8 του Κώδικα 5.6 ξεκινάμε έναν βρόγχο for για να μελετήσουμε όλα τα διαφορετικά μοντέλα που έχουμε για τις διάφορες τιμές της παραμέτρου ποινής. Στη γραμμή 9 υπολογίζουμε τον πίνακα $\boldsymbol{W}_\lambda = (\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}$, στη γραμμή 10 τον πίνακα προβολής $Ridge$ $\boldsymbol{H}_\lambda = \boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X} + \lambda \boldsymbol{I}_p)^{-1}\boldsymbol{X}^T$ και στη γραμμή 11 τις εκτιμήσεις $Ridge$ των τιμών της μεταβλητής απόκρισης $\hat{\boldsymbol{y}}^{Ridge} = \boldsymbol{H}_\lambda \boldsymbol{y}$. Έτσι με όλα αυτά γνωστά υπολογίζουμε στις υπόλοιπες γραμμές του Κώδικα 5.6 το άθροισμα τετραγώνων των υπολοίπων κάθε μοντέλου, τους βαθμούς ελευθερίας του ως το ίχνος του πίνακα $\boldsymbol{H}_\lambda$, όπως έχουμε ορίσει στη Παράγραφο 5.8, και τα κριτήρια πληροφορίας του, AIC και BIC, Παράγραφος 5.9.1.

```
1  stats<-function(X,Y,l){
2    df<-rep(NA,length(l))
3    AIC<-rep(NA,length(l))
4    BIC<-rep(NA,length(l))
5    p<-ncol(X)
6    n<-nrow(X)
7    Z<-scale(X)
8    for(i in 1:length(l)){
9      W<-solve(t(Z)% * %Z + l[i]*diag(p))
10     H<-Z% * % W % * % t(Z)
11     yhat<-H% * %Y
12     RSS<-sum((yhat-Y)^2)
13     df[i]<-sum(diag(H)) #df1<-rankMatrix(H)
14     AIC[i]<-n*(log(2*pi*RSS/n)+1)+2*df[i]
15     BIC[i]<-n*(log(2*pi*RSS/n)+1)+log(n)*df[i]
16   }
17   results<-list(df,AIC,BIC)
18   return(results)
19 }
```

Κώδικας 5.6: Η συνάρτηση stats() που δέχεται ως ορίσματα τα δεδομένα των επεξηγηματικών μας μεταβλητών X, τα δεδομένα της μεταβλητής απόκρισης Y και τις τιμές της παραμέτρου ποινής, ενώ υπολογίζει και επιστρέφει τους βαθμούς ελευθερίας των διαφορετικών μοντέλων και τα κριτήρια πληροφορίας τους AIC και BIC.

Πλέον λοιπόν, στον Κώδικα 5.7, εκτελούμε τη συνάρτηση $stats()$ και αποθηκεύουμε τις τιμές των βαθμών ελευθερίας, των AIC και των BIC των διαφορετικών μοντέλων. Έπειτα κατασκευάζουμε το κοινό γράφημα των τιμών των εκτιμητριών *Ridge* των συντελεστών των συμμεταβλητών μας συναρτήσει των βαθμών ελευθερίας για τα διαφορετικά μοντέλα, που φαίνεται στην Εικόνα 5.5.

```
models<-stats(X=as.matrix(X.diab),Y=as.matrix(Y.diab),l=
   l.diab)
df<-models[[1]]
AIC<-models[[2]]
BIC<-models[[3]]

#regularization paths with degrees of freedom
plot(df,dridge$coef[1,],ylim=range(dridge$coef),type="l"
   ,ylab="Coefficients",xlab="Degrees of freedom",main="
   Coefficients vs Degrees of freedom",lwd=3)
for(j in 2:ncol(X.diab)) lines(df,dridge$coef[j,],col=j,
   lwd=3,lty=j)
legend("bottomleft",legend=paste(names(X.diab),sep=''),
   ncol=3,lwd=3,col=1:ncol(X.diab),bty="n",lty=1:ncol(X.
   diab),cex=0.8)
```

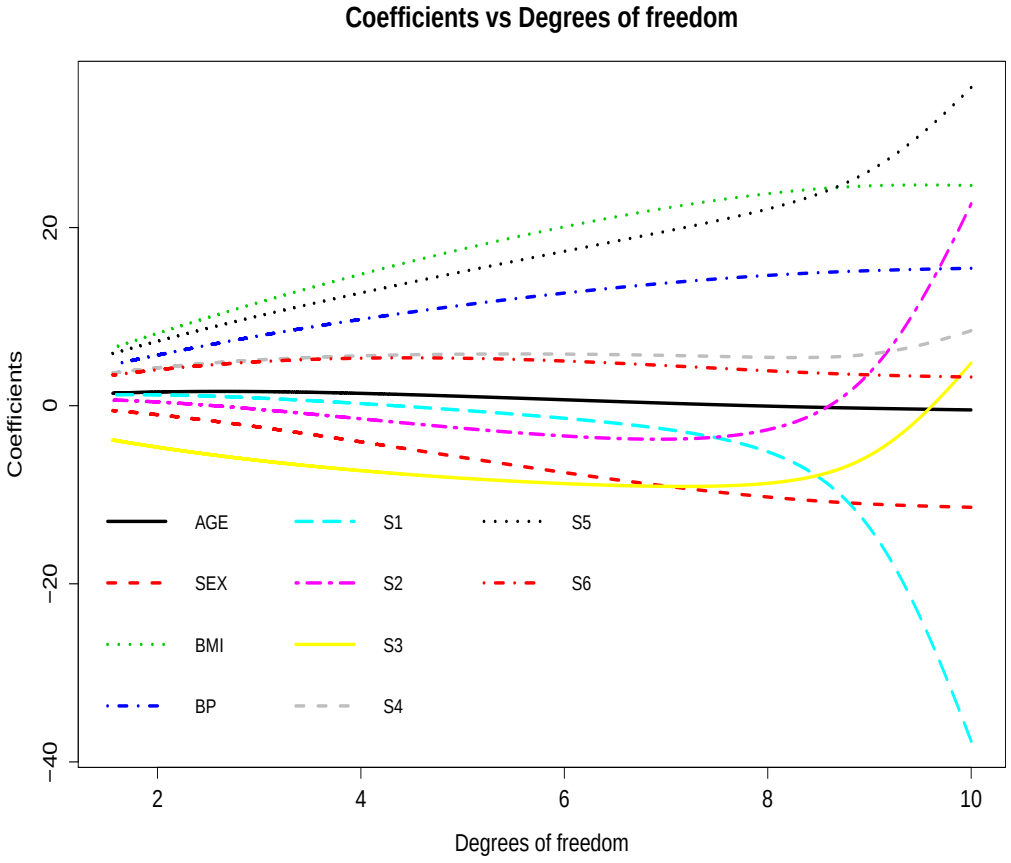
Κώδικας 5.7: Εκτελούμε τη συνάρτηση $stats()$ για τα δεδομένα των ασθενών που πάσχουν από διαβήτη και τις τιμές της παραμέτρου ποινής που έχουμε παράξει. Αποθηκεύουμε τα αποτελέσματα και κατασκευάζουμε το κοινό γράφημα των τιμών των εκτιμητριών Ridge των συντελεστών των συμμεταβλητών μας συναρτήσει των βαθμών ελευθερίας για τα διαφορετικά μοντέλα.

Στην Εικόνα 5.5 παρατηρούμε πως για μηδενικούς βαθμούς ελευθερίας, έχουμε το μοντέλο με τους πιο συρρικνωμένους συντελεστές, οι τιμές των οποίων αυξάνουν όσο αυξάνονται οι βαθμοί ελευθερίας.

Επιλογή της παραμέτρου ποινής λ

Θα εντοπίσουμε τώρα τις διαφορετικές παραμέτρους ποινής λ που παράγουν «βέλτιστα», βάσει της θεωρίας που αναπτύξαμε στην Παράγραφο 5.9, μοντέλα. Συγκεκριμένα θα βρούμε:

- το λ που ελαχιστοποιεί το AIC, Παράγραφος 5.9.1,
- το λ που ελαχιστοποιεί το BIC, Παράγραφος 5.9.1,
- το λ που ελαχιστοποιεί το GCV, Παράγραφος 5.9.3,



Εικόνα 5.5: Κοινό γράφημα των τιμών των εκτιμητριών Ridge των συντελεστών των επεξηγηματικών μας μεταβλητών, συναρτήσει των βαθμών ελευθερίας, για όλα τα διαφορετικά μοντέλα.

- το λ που ελαχιστοποιεί το $CV_{10-fold}$, Παράγραφος 5.9.3, το οποίο θα το υπολογίσουμε με την εντολή $ridge.cv()$ του πακέτου $parcor$,
- το λ που ισούται με την ποσότητα $k_{HKB} = \frac{(p-2)\hat{\sigma}_\varepsilon^2}{\hat{\boldsymbol{\beta}}^T\hat{\boldsymbol{\beta}}}$, Παράγραφος 5.9.2, και τέλος
- το λ που ισούται με την ποσότητα $k_{LW} = \frac{(p-2)\hat{\sigma}_\varepsilon^2}{\hat{\boldsymbol{y}}^T\hat{\boldsymbol{y}}}$, Παράγραφος 5.9.2.

Για το σκοπό αυτό, κατασκευάζουμε στην R τη συνάρτηση $diff.lambdas()$, Κώδικας 5.8, που δέχεται ως ορίσματα τη λίστα που επιστρέφει η R όταν εφαρμόζουμε τη παλινδρόμηση $Ridge$ για τις διαφορετικές παραμέτρους ποινής, τα δεδομένα μας και τις τιμές των AIC και BIC που έχουμε υπολογίσει για τα διαφορετικά μοντέλα. Υπολογίζει και επιστρέφει τις τιμές της παραμέτρου

ποινής που αναφέραμε, εκτελώντας παράλληλα και 10-fold Cross Validation με την εντολή *ridge.cv*() του πακέτου *parcor*. Έπειτα εκτελούμε αυτή τη συνάρτηση για τα δεδομένα των ασθενών που πάσχουν από διαβήτη, και λαμβάνουμε τις διαφορετικές τιμές της παραμέτρου ποινής που θέλουμε, οι οποίες φαίνονται στην Εικόνα 5.6.

```
install.packages("parcor")
library(parcor)
diff.lambdas<-function(ridge,X,Y,AIC,BIC){
  lambdas<-rep(NA,6)
  lambdas[1]<-ridge$lambda[AIC==min(AIC)]
  lambdas[2]<-ridge$lambda[BIC==min(BIC)]
  lambdas[3]<-ridge$lambda[ridge$GCV==min(ridge$GCV)]
  lambdas[4]<-ridge.cv(as.matrix(X),Y,lambda=seq(0,1000,
     length.out=100000),k=10,plot.it=F)$lambda.opt
  lambdas[5]<-ridge$kHKB
  lambdas[6]<-ridge$kLW
  names(lambdas)<-c("AIC","BIC","GCV","10-CV","HKB","LW"
     )
  return(lambdas)
}

dlambdas<-diff.lambdas(dridge,X.diab,Y.diab,AIC,BIC)
```

Κώδικας 5.8: Η συνάρτηση $diff.lambdas()$ *που δέχεται ως ορίσματα τα δεδομένα των επεξηγηματικών μεταβλητών* X *και της μεταβλητής απόκρισης* Y*, τη λίστα ridge που επιστρέφει η R από την παλινδρόμηση Ridge που έχουμε προσαρμόσει και τις τιμές των κριτηρίων πληροφορίας* AIC *και* BIC*. Εκτελεί επιπλέον* $10 - foldCrossValidation$ *και υπολογίζει και επιστρέφει τις 6 τιμές της παραμέτρου ποινής που παράγουν «βέλτιστα» μοντέλα.*

```
> dlambdas
      AIC       BIC       GCV     10-CV       HKB       LW
3.000030 77.260773  3.240032  4.160042  5.462343  7.641698
```

Εικόνα 5.6: Οι τιμές της παραμέτρου ποινής βάσει των 6 διαφορετικών μεθόδων επιλογής που έχουμε αναπτύξει.

Παρατηρούμε στην Εικόνα 5.6 τις τιμές της παραμέτρου ποινής βάσει των διαφορετικών μεθοδολογιών επιλογής που εφαρμόσαμε. Οι τιμές του λ που ελαχιστοποιούν το AIC, το GCV και το $CV_{10-fold}$ απέχουν λίγο μεταξύ τους και κυμαίνονται μεταξύ του 3 και του 4.16. Ακολουθούν, σε εξίσου κοντινή

απόσταση οι προτάσεις k_{HKB} και k_{LW} που δίνουν τιμές της παραμέτρου ποινής 5.46 και 7.64 αντιστοίχως. Η μεγαλύτερη τιμή παραμέτρου ποινής εντοπίζεται από την ελαχιστοποίηση του κριτηρίου πληροφορίας BIC και είναι ίση με 77.26, αρκετά μεγαλύτερη από τις υπόλοιπες. Επιτυγχάνει δηλαδή τη μεγαλύτερη συρρίκνωση των συντελεστών των επεξηγηματικών μεταβλητών. Οι τιμές των συντελεστών για αυτές τις παραμέτρους ποινής υπολογίζονται στον Κώδικα 5.9 με την εντολή $coef()$ που δέχεται ως όρισμα την παλινδρόμηση $Ridge$ $lm.ridge(Y.diab\ ., data = X.diab, lambda = dlambdas)$ για τις διαφορετικές τιμές $dlambdas$ της παραμέτρου ποινής που υπολογίσαμε.

```
coefficients<-coef(lm.ridge(Y.diab~.,data=X.diab, lambda
   =dlambdas))
coefficients<-as.data.frame(coefficients,row.names=names
   (dlambdas))
```

Κώδικας 5.9: Υπολογίζουμε τις τιμές των εκτιμητριών Ridge για τις 6 διαφορετικές τιμές της παραμέτρου ποινής.

```
> coefficients
              V1          AGE        SEX      BMI        BP          S1
AIC   -309.9969 -0.02843855 -22.48966 5.615692 1.1068608 -0.6252718
BIC   -230.6245  0.02273529 -17.98090 5.010963 0.9936935 -0.0746860
GCV   -307.9158 -0.02801516 -22.46629 5.615153 1.1062590 -0.6052263
10-CV -301.0573 -0.02653136 -22.38084 5.612077 1.1040786 -0.5395839
HKB   -293.5747 -0.02471004 -22.26826 5.605712 1.1012448 -0.4689454
LW    -284.6369 -0.02211958 -22.09371 5.591855 1.0969112 -0.3866405
              S2         S3       S4       S5        S6
AIC    0.3270717 -0.1727168 5.093179 56.66191 0.2909052
BIC   -0.1238695 -0.7007978 4.388790 37.34789 0.3940862
GCV    0.3090103 -0.1958520 5.035568 56.14001 0.2915721
10-CV  0.2499100 -0.2712090 4.851827 54.41695 0.2939955
HKB    0.1864286 -0.3514033 4.664865 52.53075 0.2971545
LW     0.1127534 -0.4430762 4.467521 50.26754 0.3019789
```

Εικόνα 5.7: Εκτιμήτριες Ridge των συμμεταβλητών που στοχεύουν στην πρόβλεψη της εξέλιξης της νόσου του διαβήτη, για τις 6 διαφορετικές τιμές της παραμέτρου ποινής στις οποίες καταλήξαμε. Στην πρώτη στήλη περιλαμβάνονται οι εκτιμήσεις της σταθεράς.

Στην Εικόνα 5.7 παρατηρούμε τις εκτιμήτριες $Ridge$ των συντελεστών των συμμεταβλητών μας στη μελέτη της εξέλιξης του διαβήτη, για τις διαφορετικές τιμές της παραμέτρου ποινής που επιλέξαμε. Λαμβάνουμε 6 διαφορετικά μοντέλα. Σε όλα τα διαφορετικά μοντέλα παρατηρούμε ότι οι συντελεστές των επεξηγηματικών μεταβλητών AGE, $S1$, $S2$, $S3$ και $S6$ είναι πολύ κοντά στο

μηδέν, αλλά όχι ακριβώς μηδέν, υποδεικνύοντας ότι αυτές οι συμμεταβλητές δεν έχουν μεγάλη επίδραση στην εξέλιξη της νόσου του διαβήτη. Οι υπόλοιπες συμμεταβλητές φαίνεται να έχουν σημαντικές επιδράσεις στην εξέλιξη της νόσου. Συμπεραίνουμε ότι για το ανδρικό φύλο ο διαβήτης εξελίσσεται με πιο ήπιο ρυθμό από ότι για το γυναικείο φύλο, ενώ ο αυξημένος δείκτης μάζας σώματος BMI, η αυξημένη αρτηριακή πίεση BP και οι αυξημένες μετρήσεις του ορού του αίματος $S4$ και $S5$ δείχνουν να επιβαρύνουν σημαντικά την εξέλιξη της νόσου.

6

LASSO

Η μεθοδολογία *LASSO* (Least Absolute Shrinkage and Selection Operator) αποτελεί την δεύτερη μέθοδο συρρίκνωσης των συντελεστών προς εκτίμηση του πολλαπλού γραμμικού μοντέλου που θα εξετάσουμε. Προτάθηκε από τον Tibshirani, 1996. Προς το παρόν, όσον αφορά τη βελτίωση της εκτιμήτριας ελαχίστων τετραγώνων, έχουμε αναφέρει την επιλογή του καλύτερου υποσυνόλου των επεξηγηματικών μεταβλητών με τις διαδικασίες κατα βήματα, Κεφάλαιο 3, και έχουμε αναπτύξει και τη παλινδρόμηση *Ridge*, Κεφάλαιο 5. Και οι δύο αυτές εναλλακτικές όμως έχουν μειονεκτήματα. Οι διαδικασίες κατα βήματα ναι μεν μας παρέχουν μοντέλα τα οποία είναι ερμηνεύσιμα, αλλά είναι και ευμετάβλητα διότι προκύπτουν από διακριτή διεργασία, αφού οι επεξηγηματικές μεταβλητές είτε προστείθενται είται αφαιρούνται από το μοντέλο. Η μέθοδος της παλινδρόμησης *Ridge* είναι μια συνεχής διεργασία, συρρίκνωσης των συντελεστών, η οποία όμως δεν θέτει κάποιον συντελεστή ίσο με το μηδέν και άρα δεν μας παρέχει εύκολα ερμηνεύσιμα μοντέλα. Η μεθοδολογία *LASSO* αποτελεί μια μέθοδο συρρίκνωσης, των συντελεστών των επεξηγηματικών μεταβλητών του μοντέλου μας, θέτοντας μάλιστα ορισμένους συντελεστές ακριβώς ίσους με το μηδεν, εξού και το *Selection* στο όνομά της. Κατα αυτόν τον τρόπο προσπαθεί να διατηρήσει τα θετικά στοιχεία των διαδικασιών κατά βήματα και της παλινδρόμησης *Ridge*. Ενδείκνυται η χρησιμοποίησή της σε δεδομένα που «πάσχουν» απο πολυσυγγραμικότητα, επίσης στην περίπτωση που αριθμός των παρατηρήσεων του δείγματός μας ξεπερνά τον αριθμό των επεξηγηματικών μεταβλητών που χρησιμοποιούμε καθώς και σε δεδομένα υψηλών διαστάσεων. Θεωρούμε, όπως και στην παλινδρόμηση *Ridge* ότι οι τιμές των συμμεταβλητών είναι τυποποιημένες ώστε να έχουν μέση τιμή 0 και διασπορά 1, δηλαδή $\sum_{i=1}^{n} x_{ij} = 0$ και $\sum_{i=1}^{n} x_{ij}^2 = 1 \ \forall j = 1, 2, ..p$. Επίσης θεωρούμε ότι οι τιμές της μεταβλητής απόκρισης είναι κεντραρισμένες ώστε να έχουν μέση τιμή μηδέν και ότι η σταθερά παραλείπεται από το μοντέλο. Εφόσον θέλουμε να την συμπεριλάβουμε η εκτίμησή της είναι η γνωστή: $\hat{\beta}_0 = \bar{\boldsymbol{y}}$ (όπου $\bar{\boldsymbol{y}}$ η μέση τιμή

των μη κεντραρισμένων τιμών της μεταβλητής απόκρισης).

6.1 Ποινικοποίηση της ℓ_1-νόρμας

Η μέθοδος $LASSO$, μας υποδεικνύει την ελαχιστοποίηση του αθροίσματος των τετραγώνων των υπολοίπων (RSS) υπό κάποιον περιορισμό του τετραγώνου της l_1-νόρμας, δηλαδή του αθροίσματος των απολύτων, των συντελεστών $\boldsymbol{\beta}$ του γραμμικού μας μοντέλου:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad \sum_{i=1}^{n}\Big(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2 \; s.t \; \sum_{j=1}^{p}|\beta_j|\leq t$$

$$\Longleftrightarrow \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad (\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta}) \; s.t \; \sum_{j=1}^{p}|\beta_j|\leq t$$

Τοιουτοτρόπως, κανονικοποιούνται οι συντελεστές με την έννοια του ελέγχου του πόσο μεγάλες τιμές μπορούν να πάρουν στο σύνολό τους, αφού τείθεται ένα άνω φράγμα t στο τετράγωνο της ℓ_1-νόρμας του διανύσματος $\boldsymbol{\beta}$.

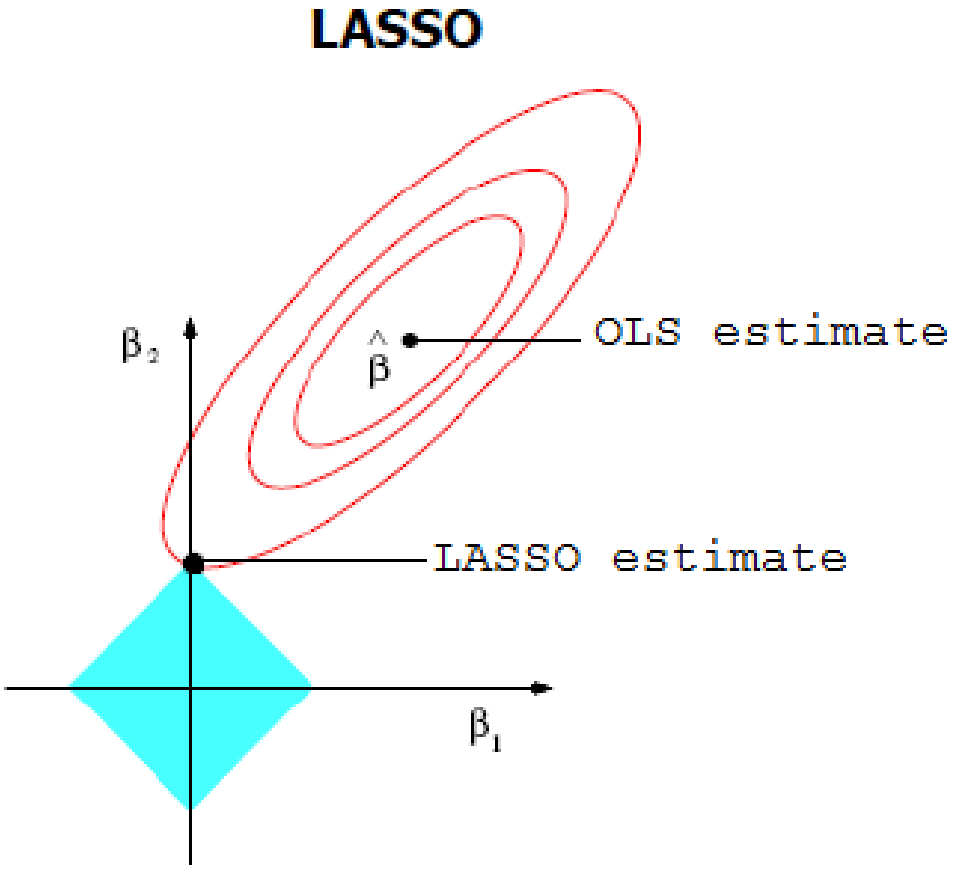
Το πρόβλημα ελαχιστοποίησης υπό τον περιορισμό $LASSO$ που παραθέσαμε, μπορεί να γραφεί ακολούθως, με τη μορφή του ποινικοποιημένου, ως προς την l_1-νόρμα, αθροίσματος τετραγώνων των υπολοίπων:

$$\begin{aligned}\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad PRSS(\boldsymbol{\beta})_{\ell_1} &= \sum_{i=1}^{n}\Big(y_i-\sum_{j=1}^{p}x_{ij}\beta_j\Big)^2+\lambda\sum_{j=1}^{p}|\beta_j| \\ &= \underbrace{\|\boldsymbol{y}-\boldsymbol{X\beta}\|_2^2}_{\text{Loss}}+\lambda\,\underbrace{\|\boldsymbol{\beta}\|_1}_{\text{Penalty}} \\ &= (\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})+\lambda\|\boldsymbol{\beta}\|_1\end{aligned}$$

Το άνω φράγμα t του τετραγώνου της ℓ_1-νόρμας μετατράπηκε με 1-1 αντιστοιχία στην παράμετρο ποινής, ή συρρίκνωσης, λ, με χρήση των πολλαπλασιαστών $Lagrange$. Μάλιστα η σχέση που συνδέει τα t και λ είναι αντιστρόφως ανάλογη. Οπότε η παράμετρος ποινής λ ελέγχει και αυτή με τη σειρά της το μέγεθος κανονικοποίησης, δηλαδή το πόσο μεγάλοι θα γίνουν οι συντελεστές των επεξηγηματικών μεταβλητών του μοντέλου μας. Σκοπός μας πλέον είναι να ελαχιστοποιήσουμε το $PRSS(\boldsymbol{\beta})_{l_1}$. Δυστυχώς κάτι τέτοιο δεν μπορεί να γίνει αναλυτικά για τη γενική περίπτωση που δεν γνωρίσουμε το πρόσημο των συντελεστών $\boldsymbol{\beta}$, διότι η συνάρτηση $|\cdot|$ δεν είναι παραγωγίσιμη στο μηδέν. Παρόλα αυτά, γνωρίζουμε ήδη ότι το άθροισμα τετραγώνων των υπολοίπων

$RSS(\boldsymbol{\beta}) = (\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})$ είναι αυστηρώς κυρτό στον $\mathbb{R}^p$, ενώ η συνάρτηση $g(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$ είναι κυρτή αλλά δεν είναι παραγωγίσιμη στα σημεία όπου τουλάχιστον ένας συντελεστής β_j είναι ίσος με 0. Στην Παράγραφο 6.3 θα μελετήσουμε αυτό το πρόβλημα, υπό τη σκοπιά των κυρτών προβλημάτων ελαχιστοποίησης με περιορισμούς όπου κάποιες εκ των συναρτήσεων που τα αποτελούν είναι μη διαφορίσιμες, ώστε να καταλήξουμε στην ικανή και αναγκαία συνθήκη ύπαρξης λύσης στο πρόβλημα $LASSO$. Έπειτα, στην Παράγραφο 6.4 θα μελετήσουμε τους δύο πιο διαδεδομένους και αποδοτικούς αλγορίθμους για την εύρεση των λύσεων $LASSO$ για διάφορες τιμές της παραμέτρου ποινής λ.

6.2 Γεωμετρία της LASSO



Εικόνα 6.1: Η περιοχή του περιορισμού $\sum_{j=1}^{p} |\beta_j| \leq t$ και οι ισοϋψείς ελλείψεις RSS στο πολλαπλό γραμμικό μοντέλο δύο διαστάσεων ($p = 2$).

Στην Εικόνα 6.1 παρατηρούμε πως συσχετίζονται οι εκτιμητές OLS με τους ε-

κτιμητές $LASSO$ στο πολλαπλό γραμμικό μοντέλο δύο διαστάσεων ($p = 2$). Οι ισοϋψείς ελλείψεις απεικονίζουν το άθροισμα των τετραγώνων των υπολοίπων RSS ($Residual\ Sum\ of\ Squares$) που καλούμαστε να ελαχιστοποιήσουμε ως προς τις παραμέτρους $\boldsymbol{\beta}$. Ο ρόμβος απεικονίζει τον περιορισμό που έχουμε για την l_1-νόρμα των παραμέτρων $\boldsymbol{\beta}$ στον $\mathbb{R}^2$. Το εφαπτόμενο σημείο του ρόμβου στις ισοϋψείς ελλείψεις αποτελεί την εκτιμήτρια $LASSO$ $\hat{\boldsymbol{\beta}}^{LASSO}$, καθώς ικανοποιεί βέλτιστα και την ελαχιστοποίηση του RSS και τον περιορισμό του αθροίσματος των απολύτων των β_j. Τα παραπάνω εύκολα γενικεύονται, αλλά δεν απεικονίζονται, στον $\mathbb{R}^p$. Όσο αυξάνεται η διάσταση του προβλήματος p, ο περιορισμός της l_1-νόρμας στον $\mathbb{R}^p$ θα είναι ένα πολυδιάστατο διαμάντι με αυξανόμενο αριθμό γωνιών. Άρα αυξάνεται η πιθανότητα, κατά το πρόβλημα ελαχιστοποίησης που παρουσιάζουμε, να τεθούν κάποιοι συντελεστές ίσοι με μηδέν. Ουσιαστικά λοιπόν η $LASSO$ πραγματοποιεί συρρίκνωση των συντελεστών, θέτοντας κάποιους ίσους με μηδεν, οπότε συνεπαγωγικά δρα και ως μέθοδος επιλογής μεταβλητών.

6.3 Κυρτά προβλήματα ελαχιστοποίησης

Ας θεωρήσουμε αρχικά το γενικό κυρτό πρόβλημα ελαχιστοποίησης:

$$\begin{aligned} \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{minimize} \quad & f(\boldsymbol{\beta}) \\ s.t. \quad & g_i(\boldsymbol{\beta}) \leq 0, \quad i = 1, 2, .., m \end{aligned}$$

όπου $f : \mathbb{R}^p \to \mathbb{R}$ η κυρτή αντικειμενική συνάρτηση και $g_i : \mathbb{R}^p \to \mathbb{R}$ οι κυρτές συναρτήσεις που εκφράζουν τους m ανισοτικούς περιορισμούς που θέλουμε να ικανοποιούνται. Η $Lagrangian$, $\mathcal{L} : \mathbb{R}^p \times \mathbb{R}^m_+ \to \mathbb{R}$, αυτού του προβλήματος ορίζεται να είναι:

$$\mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\lambda}) = f(\boldsymbol{\beta}) + \sum_{i=1}^{m} \lambda_i g_i(\boldsymbol{\beta})$$

όπου $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, ..., \lambda_m)$ το διάνυσμα των μη αρνητικών πολλαπλασιαστών $Lagrange$. Έστω $h : \mathbb{R}^m \to \mathbb{R}$ η δυϊκή συνάρτηση $h(\boldsymbol{\lambda}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{min} \mathcal{L}(\boldsymbol{\beta}; \boldsymbol{\lambda})$. Τότε το δυϊκό πρόβλημα, του πρωτεύοντος που παραθέσαμε, ορίζεται να είναι:

$$\begin{aligned} \underset{\boldsymbol{\lambda}\in\mathbb{R}^m}{maximize} \quad & h(\boldsymbol{\lambda}) \\ s.t. \quad & \lambda_i \geq 0, \quad i = 1, 2, .., m. \end{aligned}$$

Αν $f^* := f(\boldsymbol{\beta}^*)$ η λύση του πρωτεύοντος προβλήματος και $h^* := h(\boldsymbol{\lambda}^*)$ η λύση του δυϊκού του, τότε ισχύει πάντα ότι $h^* \leq f^*$. Όταν ισχύει ότι $f^* = h^*$ λέμε ότι το πρωτεύον πρόβλημα έχει ισχυρή δυϊκότητα (strong duality). Η ισχυρή δυϊκότητα σε κυρτά προβλήματα εξασφαλίζεται όταν το πρωτεύον πρόβλημα ικανοποιεί την ασθενή συνθήκη του *Slatter*:

$$\exists \;\; \boldsymbol{\beta}_0 : g_i(\boldsymbol{\beta}_0) \leq 0 \quad i = 1, 2, .., m$$

6.3.1 Συνθήκες KKT για διαφορίσιμα προβλήματα

Στην ελαχιστοποίηση κυρτών προβλημάτων, υπό την ύπαρξη ισχυρής δυϊκότητας, οι συνθήκες KKT (Karush-Kuhn-Tucker) είναι ικανές και αναγκαίες ώστε το διάνυσμα $\boldsymbol{\beta}^*$ που τις ικανοποιεί να είναι λύση του ολικού ελαχίστου που αναζητάμε. Υποθέτοντας πως οι συναρτήσεις f και $\{g_i\}$ είναι συνεχώς διαφορίσιμες, το βέλτιστο πρωτεύον διάνυσμα $\boldsymbol{\beta}^*$ και το βέλτιστο δυϊκό διάνυσμα $\boldsymbol{\lambda}^*$ ικανοποιούν τις συνθήκες KKT:

- **Stationarity**: $0 = \nabla_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}^*; \boldsymbol{\lambda}^*) = \nabla f(\boldsymbol{\beta}^*) + \sum_{i=1}^{m} \lambda_i^* \nabla g_i(\boldsymbol{\beta}^*)$
- **Complementary slackness**: $\lambda_i^* g_i(\boldsymbol{\beta}^*) = 0 \; \forall i = 1, 2, , .., m$
- **Primal feasibility**: $g_i(\boldsymbol{\beta}^*) \leq 0 \; \forall i = 1, 2, , .., m$
- **Dual feasibility**: $\lambda_i^* \geq 0 \; \forall i = 1, 2, , .., m$

Η πρώτη συνθήκη αναφέρεται στη στασιμότητα του υποψήφιου ελαχίστου. Η συνθήκη του Complementary slackness δηλώνει ότι ο πολλαπλασιαστής λ_i πρέπει να είναι μηδέν όταν ο περιορισμός $g_i(\boldsymbol{\beta}) \leq 0$ είναι ανενεργός στο βέλτιστο διάνυσμα, δηλαδή όταν ισχύει ότι $g_i(\boldsymbol{\beta}^*) < 0$. Οι δύο τελευταίες συνθήκες, Primal and Dual feasibility, εξασφαλίζουν ότι δεν παραβιάζονται οι περιορισμοί του πρωτεύοντος και του δυϊκού προβλήματος. Το σύστημα εξισώσεων-ανισώσεων που αντιστοιχεί στις συνθήκες KKT συνήθως δεν επιλύεται άμεσα, πλην της περίπτωσης όπου μπορούμε να λύσουμε το πρωτεύον κυρτό πρόβλημα ελαχιστοποίησης αναλυτικά, καταλήγοντας σε κλειστό τύπο λύσης. Πολλοί αλγόριθμοι βελτιστοποίησης μπορούν να ερμηνευθούν ως μέθοδοι αριθμητικής προσέγγισης της λύσης του συστήματος εξισώσεων-ανισώσεων KKT.

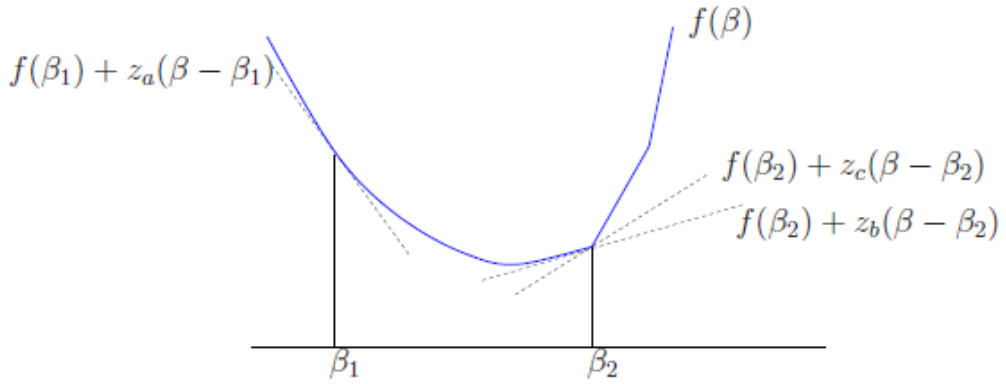
6.3.2 Μη διαφορίσιμα προβλήματα και υποδιαφορικά

Στην πράξη, πολλά κυρτά προβλήματα ελαχιστοποίησης αποτελούνται από συναρτήσεις οι οποίες δεν είναι διαφορίσιμες. Για παράδειγμα η συνάρτηση $g(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j|$ είναι κυρτή αλλά δεν είναι διαφορίσιμη στα σημεία όπου τουλάχιστον

ένας συντελεστής β_j είναι μηδενικός. Για τέτοιου είδους προβλήματα η συνθήκη στασιμότητας (*Stationarity*) παύει να είναι άμεσα εφαρμόσιμη, αφού περιέχει την κλίση της αντικειμενικής συνάρτησης και τις κλίσεις των συναρτήσεων περιορισμών. Παρόλα αυτά, για τις κυρτές συναρτήσεις, υπάρχει μια γενίκευση της κλίσης που επιτρέπει την ύπαρξη γενικευμένης θεωρίας βελτιστοποίησης.

Για τις διαφορίσιμες κυρτές συναρτήσεις, η εφαπτομένη σε οποιοδήποτε σημείο παρέχει ένα «κάτω φράγμα» για τη συνάρτηση σε εκείνο το σημείο. Ο ορισμός της υποκλίσης (*subgradient*) βασίζεται σε μια φυσική γενίκευση αυτής της ιδέας. Συγκεκριμένα, δοθείσης μιας κυρτής συνάρτησης $f : \mathbb{R}^p \to \mathbb{R}$, ένα διάνυσμα $\boldsymbol{z} \in \mathbb{R}^p$ ορίζεται να είναι η υποκλίση της f στο $\boldsymbol{\beta}$ αν:

$$f(\boldsymbol{\beta}') \geq f(\boldsymbol{\beta}) + \langle \boldsymbol{z}, \boldsymbol{\beta}' - \boldsymbol{\beta} \rangle$$



Εικόνα 6.2: Μια κυρτή συνάρτηση $f : \mathbb{R} \to \mathbb{R}$, μαζί με μερικά παραδείγματα υποκλίσεων στα β_1 και β_2.

Από γεωμετρική σκοπιά, το διάνυσμα υποκλίσης $\boldsymbol{z}$ είναι το διάνυσμα που ορίζει το μη κατακόρυφο επίπεδο το οποίο στηρίζει το επιγράφημα της f. Το σύνολο όλων των υποκλίσεων της f στο $\boldsymbol{\beta}$ ονομάζεται υποδιαφορικό ($subdifferential$) και συμβολίζεται $\partial f(\boldsymbol{\beta})$. Όταν η f είναι διαφορίσιμη στο $\boldsymbol{\beta}$ το υποδιαφορικό είναι το μονοσύνολο της κλίσης, $\partial f(\boldsymbol{\beta}) = \{\nabla f(\boldsymbol{\beta})\}$. Σε σημεία μη διαφορισιμότητας, το υποδιαφορικό είναι το κυρτό σύνολο όλων των δυνατών υποκλίσεων. Στην Εικόνα 6.2 δίνεται μια κυρτή συνάρτηση $f : \mathbb{R} \to \mathbb{R}$, μαζί με μερικά παραδείγματα υποκλίσεων στα β_1 και β_2. Στο β_1 είναι διαφορίσιμη και η υποκλίση ισούται με το μονοσύνολο της κλίσης. Στο β_2 δεν είναι διαφορίσιμη και παρατείθονται δύο παραδείγματα υποκλίσεων. Ένα άλλο παράδειγμα, που

σχετίζεται άμεσα με το πρόβλημα $LASSO$, αφορά τη συνάρτηση της απόλυτης τιμής $g(\beta) = |\beta|$, της οποίας το υποδιαφορικό είναι:

$$\partial g(\beta) = \begin{cases} \{+1\} & \text{αν} \quad \beta > 0 \\ \{-1\} & \text{αν} \quad \beta < 0 \\ [-1, +1] & \text{αν} \quad \beta = 0 \end{cases}$$

Συχνά γράφουμε $z \in sign(\beta)$ για να δηλώσουμε ότι z ανήκει στο υποδιαφορικό της συνάρτησης απόλυτης τιμής στο β.

Έτσι λοιπόν, η γενικευμένη θεωρία KKT μπορεί και πάλι να εφαρμοσθεί χρησιμοποιώντας τη τροποποιημένη συνθήκη:

$$\mathbf{0} \ \in \ \partial f(\boldsymbol{\beta}^*) + \sum_{i=1}^{m} \lambda_i^* \partial g_i(\boldsymbol{\beta}^*)$$

όπου αντικαταστήσαμε τις κλίσεις στη συνθήκη στασιμότητας ($stationarity$), που παραθέσαμε στη προηγούμενη Παράγραφο, με υποδιαφορικά. Από τη στιγμή που το υποδιαφορικό ($subdifferential$) είναι σύνολο, η σχέση που λάβαμε δηλώνει ότι το μηδενικό διάνυσμα ανήκει στο άθροισμα των υποδιαφορικών της αντικειμενικής συνάρτησης και των συναρτήσεων περιορισμών, ορίζοντας το άθροισμα δύο υποσυνόλων A και B του $\mathbb{R}^p$ να είναι $A + B := \{\alpha + \beta | \alpha \in A, \beta \in B\}$.

6.3.3 Ικανή και αναγκαία συνθήκη ύπαρξη λύσης LASSO

Επιστρέφοντας στη μελέτη του προβλήματος $LASSO$ που παραθέσαμε στη Παράγραφο 6.1, έχουμε ως αντικειμενική συνάρτηση το άθροισμα τετραγώνων των υπολοίπων, $f(\boldsymbol{\beta}) = RSS(\boldsymbol{\beta})$, που είναι κυρτή και διαφορίσιμη συνάρτηση καθώς και μόνο έναν περιορισμό, $g(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j| - t \leq 0$ για μια θετική σταθερά t. Άρα ο περιορισμός $g(\boldsymbol{\beta}) \leq 0$ υποδηλώνει ότι το $\boldsymbol{\beta}$ ανήκει στο p-διάστατο διαμάντι του $\mathbb{R}^p$ με την ℓ_1-νόρμα, ακτίνας t. Το πρόβλημα $LASSO$ έχει ισχυρή δυϊκότητα αφού ικανοποιείται η ασθενής συνθήκη του $Slatter$ (δες Παράγραφο 6.3) για $\boldsymbol{\beta}_0 = \mathbf{0}$, οπότε χρησιμοποιώντας τον ορισμό του υποδιαφορικού για τη συνάρτηση απόλυτης τιμής της Παραγράφου 6.3.2, η συνθήκη στασιμότητας της γενικευμένης θεωρίας KKT γράφεται:

$$\nabla f(\boldsymbol{\beta}^*) + \lambda^* z^* = \mathbf{0}$$

όπου τα στοιχεία z_j^* του διανύσματος υποκλίσης $\boldsymbol{z}^*$ ικανοποιούν:

$$z_j^* \in \begin{cases} sign(\beta_j^*) & \text{αν } \beta_j \neq 0 \\ [-1,1] & \text{αν } \beta_j = 0. \end{cases}$$

Από τη στιγμή που $f(\boldsymbol{\beta}) = (\boldsymbol{y}-\boldsymbol{X\beta})^T(\boldsymbol{y}-\boldsymbol{X\beta})$, η ικανή και αναγκαία συνθήκη ύπαρξης λύσης στο πρόβλημα $LASSO$ που μελετάμε παίρνει τη μορφή:

$$\begin{aligned} & -2\boldsymbol{X}^T(\boldsymbol{y}-\boldsymbol{X\beta}) + \lambda \boldsymbol{z} = \boldsymbol{0} \\ \iff & -2\langle \boldsymbol{x}_j, \boldsymbol{y}-\boldsymbol{X\beta}\rangle + \lambda z_j = 0 \qquad j = 1,2,...,p \\ \iff & \langle \boldsymbol{x}_j, \boldsymbol{y}-\boldsymbol{X\beta}\rangle = \frac{1}{2}\lambda z_j \qquad j = 1,2,...,p \end{aligned}$$

όπου z_j τα στοιχεία του διανύσματος υποκλίσης ίσα με $sign(\beta_j)$ αν $\beta_j \neq 0$ και ίσα με κάποια τιμή στο $[-1,1]$ διαφορετικά. Οι λύσεις $\hat{\boldsymbol{\beta}}$ στο πρόβλημα $LASSO$ της Παραγράφου 6.1, είναι οι ίδιες με τις λύσεις $(\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{z}})$ στην ικανή και αναγκαία συνθήκη που παραθέσαμε και θα τις συμβολίζουμε $\hat{\boldsymbol{\beta}}^{Lasso}$. Εκφράζοντας το πρόβλημα $LASSO$ υπό αυτήν την αναγκαία και ικανή συνθήκη που περιέχει τις υποκλίσεις της συνάρτησης περιορισμού, μπορεί να φανεί πολύ χρήσιμο στον σχεδιασμό αλγορίθμων που υπολογίζουν τις λύσεις του.

Σκόπιμο είναι να γενικεύσουμε την ικανή και αναγκαία συνθήκη που περιγράψαμε, για τις ενεργές (ο συντελεστής τους μη μηδενικός) και ανενεργές (ο συντελεστής τους μηδενικός) συμμεταβλητές του μοντέλου μας, για κάποιο δοθέν λ. Έστω $\mathcal{B}$ το σύνολο των δεικτών των ενεργών επεξηγηματικών μεταβλητών για κάποιο συγκεκριμένο λ. Τότε έχουμε:

Ενεργές επεξηγηματικές: $\forall j \in \mathcal{B} \iff \beta_j \neq 0 \iff z_j = sign(\beta_j)$

$$\begin{aligned} \langle \boldsymbol{x}_j, \boldsymbol{y}-\boldsymbol{X\beta}\rangle &= \frac{1}{2}\lambda z_j \qquad \forall j \in \mathcal{B} \\ \iff \langle \boldsymbol{x}_j, \boldsymbol{y}-\boldsymbol{X\beta}\rangle &= \frac{1}{2}\lambda \cdot sign(\beta_j) \quad \forall j \in \mathcal{B}. \end{aligned}$$

Ανενεργές επεξηγηματικές: $\forall j \notin \mathcal{B} \iff \beta_j = 0 \iff z_j \in [-1,1]$

$$\begin{aligned} \langle \boldsymbol{x}_j, \boldsymbol{y}-\boldsymbol{X\beta}\rangle &= \frac{1}{2}\lambda z_j \quad \forall j \notin \mathcal{B} \\ \iff |\langle \boldsymbol{x}_j, \boldsymbol{y}-\boldsymbol{X\beta}\rangle| &\leq \frac{1}{2}\lambda \quad \forall j \notin \mathcal{B}. \end{aligned}$$

6.4 Πολλές επεξηγηματικές μεταβλητές - Αλγόριθμοι εύρεσης της λύσης LASSO

6.4.1 Least Angle Regression

Ο αλγόριθμος *LAR* (Least Angle Regression) προτάθηκε από τους Efron, Hastie, Johnstone και Tibshirani το 2004 για την προσαρμογή του πολλαπλού γραμμικού μοντέλου παλινδρόμησης σε *high dimensional data*. Μια παραλλαγή του αλγορίθμου υπολογίζει τη λύση $LASSO$, όμως πριν αναφερθούμε σε αυτήν σκόπιμο είναι να περιγράψουμε τον LAR. Ο LAR κατασκευάζει τις εκτιμήσεις της τιμής της μεταβλητής απόκρισης, $\hat{\boldsymbol{y}} = X\hat{\boldsymbol{\beta}}$, σε διαδοχικά βήματα, προσθέτοντας σε κάθε βήμα μια επεξηγηματική μεταβλητή στο μοντέλο, έτσι ώστε μετά από k βήματα μόνο k εκτιμήτριες $\hat{\beta}_j$ να είναι μη μηδενικές. Στο πρώτο βήμα αναγνωρίζει τη συμμεταβλητή που έχει τη μεγαλύτερη κατ′ απόλυτη τιμή συσχέτιση (*correlation*) με το αρχικό υπόλοιπο (τη μεταβλητή απόκρισης) και την εισάγει στο μοντέλο. Αντί να προσαρμόσει εξ′ ολοκλήρου το μοντέλο αυτής της συμμεταβλητής με τη μεταβλητή απόκρισης ώστε να εκτιμήσει τον συντελεστή της, ο LAR μετακινεί τον συντελεστή αυτόν συνεχώς προς την τιμή της εκτιμήτριας ελαχίστων τετραγώνων που προκύπτει παλινδρομώντας την ενεργή συμμεταβλητή στο υπόλοιπο. Κατά την μετακίνηση αυτή, μειώνεται η συχέτιση της ενεργής συμμεταβλητής με το συνεχώς εξελισσόμενο υπόλοιπο. Όταν ανιχνευθεί ότι μια άλλη επεξηγηματική μεταβλητή έχει την ίδια συχέτιση με το τρέχον εξελισσόμενο υπόλοιπο, εισάγεται και αυτή στο μοντέλο και πλέον οι δύο ενεργές συμμεταβλητές έχουν την ίδια συσχέτιση κατά απόλυτη τιμή με το τρέχον υπόλοιπο. Στη συνέχεια μετακινούνται μαζί οι συντελεστές των συμμεταβλητών που έχουν εισηχθεί με κατεύθυνση προς τις τιμές των εκτιμητριών ελαχίστων τετραγώνων που προκύπτουν από την παλινδρόμηση των ενεργών συμμεταβλητών στο καινούργιο υπόλοιπο, μέχρις ότου μια τρίτη συμμεταβλητή βρεθεί να έχει την ίδια συσχέτιση με το εξελισσόμενο υπόλοιπο με αυτήν που έχουν οι ενεργές συμμεταβλητές. Εισάγεται και αυτή στο μοντέλο και η διαδικασία συνεχίζεται αντιστοίχως μέχρις ότου εισαχθούν όλες οι επεξηγηματικές μεταβλητές στο μοντέλο και προκύψει η λύση των εκτιμητριών ελαχίστων τετραγώνων. Παρόλο που ο LAR έχει αναπτυχθεί με τη χρήση των συσχετίσεων που αναφέραμε, από τη στιγμή που κανονικοποιούμε τα δεδομένα μας, είναι ισοδύναμο και ευκολότερο να δουλέψουμε με εσωτερικά γινόμενα $\langle \cdot, \cdot \rangle$. Άρα λειτουργούμε ακολούθως:

1. Τυποποιούμε τις τιμές των επεξηγηματικών μεταβλητών ώστε να έχουν μέση τιμή μηδέν και μοναδιαία l_2-νόρμα (ισοδύναμο με μοναδιαία διασπορά) και κεντράρουμε τις τιμές της μεταβλητής απόκρισης.

2. Ξεκινάμε με το διάνυσμα των συντελεστών των συμμεταβλητών να είναι $\boldsymbol{\beta}^0 = (\beta_1, \beta_2, ..., \beta_p) = \mathbf{0}$ και άρα με το αρχικό υπόλοιπο να είναι $\boldsymbol{r}_0 = \boldsymbol{y}$.

3. Βρίσκουμε την επεξηγηματική μεταβλητή $\boldsymbol{x}_j$ με τη μεγαλύτερη τιμή της ποσότητας $|\langle \boldsymbol{x}_j, \boldsymbol{r}_0 \rangle| = |\boldsymbol{x}_j^T \boldsymbol{r}_0|$ την οποία καλούμε αρχικό κόμβο λ_0. Εισάγουμε την $\boldsymbol{x}_j$ στο μοντέλο, ορίζουμε το ενεργό σύνολο $\boldsymbol{\mathcal{A}} = \{j\}$ και $\boldsymbol{X}_{\mathcal{A}}$ τον πίνακα σχεδιασμού που περιέχει μόνο αυτή τη συμμεταβλητή.

4. Για $k = 1, 2, ..., K = min(n-1, p)$ επαναλαμβάνουμε:

   α΄ Ορίζουμε το διάνυσμα κατεύθυνσης που προκύπτει από τη παλινδρόμηση ελαχίστων τετραγώνων των εισηγμένων συμμεταβλητών με το υπόλοιπο:

   $$\boldsymbol{\delta} = \frac{1}{\lambda_{k-1}} (\boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{X}_{\mathcal{A}})^{-1} \boldsymbol{X}_{\mathcal{A}}^T \boldsymbol{r}_{k-1}.$$

   Ορίζουμε επίσης το διάνυσμα συνολικής κατεύθυνσης p-διάστασης $\boldsymbol{\Delta}$ έτσι ώστε $\boldsymbol{\Delta}_{\mathcal{A}} = \boldsymbol{\delta}$, δηλαδή στις τιμές που αντιστοιχούν στους δείκτες των εισηγμένων συμμεταβλητών να έχει τις αντίστοιχες τιμές του $\boldsymbol{\delta}$ και οι υπόλοιπες τιμές του να είναι 0.

   β΄ Μετακινούμε το διάνυσμα των συντελεστών $\boldsymbol{\beta}$ από το $\boldsymbol{\beta}^{k-1}$ κατά τη συνολική κατεύθυνση $\boldsymbol{\Delta}$:

   $$\boldsymbol{\beta}(\lambda) = \boldsymbol{\beta}^{k-1} + (\lambda_{k-1} - \lambda)\boldsymbol{\Delta}$$

   για $0 < \lambda < \lambda_{k-1}$, παρακολουθώντας παράλληλα το συνεχώς εξελισσόμενο υπόλοιπο:

   $$\begin{aligned} \boldsymbol{r}(\lambda) &= \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}(\lambda) \\ &= \boldsymbol{r}_{k-1} - (\lambda_{k-1} - \lambda)\boldsymbol{X}\boldsymbol{\Delta}. \end{aligned}$$

   Να διευκρινίσουμε ότι λ είναι η **κοινή τιμή** του κατά απόλυτη τιμή εσωτερικού γινομένου καθεμίας εκ των ενεργών συμμεταβλητών με το εξελισσόμενο υπόλοιπο, δηλαδή:

   $$|\langle \boldsymbol{x}_j, \boldsymbol{r}(\lambda) \rangle| = \lambda \qquad \forall j \in \boldsymbol{\mathcal{A}}.$$

   Αυτό ισοδύναμα σημαίνει ότι οι συχετίσεις των ενεργών συμμεταβλητών παραμένουν ίσες και μειώνονται προς το 0 όσο μειώνεται το λ.

γ΄ Παρακολουθούμε για κάθε $\ell \notin \mathcal{A}$ τα, κατά απόλυτη τιμή, εσωτερικά γινόμενα των, ανενεργών, συμμεταβλητών με το εξελισσόμενο υπόλοιπο $|\langle \boldsymbol{x}_\ell, \boldsymbol{r}(\lambda)\rangle|$. Όταν κάποια ανενεργή επεξηγηματική μεταβλητή, έστω $\boldsymbol{x}_m$ αποκτήσει το ίδιο κατ' απόλυτη τιμή εσωτερικό γινόμενο με την κοινή τιμή λ, δηλαδή όταν $|\langle \boldsymbol{x}_m, \boldsymbol{r}(\lambda)\rangle| = \lambda$, έχουμε τον καινούργιο κόμβο λ_k. Ανανεώνουμε το διάνυσμα των συντελεστών $\boldsymbol{\beta}^k = \boldsymbol{\beta}^{k-1} + (\lambda_{k-1} - \lambda_k)\boldsymbol{\Delta}$ και το υπόλοιπο $\boldsymbol{r}^k = \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}^k$.

δ΄ Εισάγουμε την επεξηγηματική μεταβλητή $\boldsymbol{x}_m$ στο μοντέλο, ανανεώνουμε το ενεργό σύνολο $\mathcal{A} = \mathcal{A} \cup \{m\}$ και τον πίνακα $\boldsymbol{X}_\mathcal{A}$.

5. Επιστρέφουμε την ακολουθία $\{\lambda_k, \boldsymbol{\beta}^k\}_0^K$.

Είναι αναγκαίο να διευκρινίσουμε εδώ ότι στο βήμα $4\gamma'$ δεν χρειάζεται να μειώνουμε σε μικρά βήματα τη **κοινή τιμή** λ, και να επανελέγχουμε τα εσωτερικά γινόμενα των ανενεργών συμμεταβλητών με το υπόλοιπο. Μπορούμε αντ' αυτού να λύσουμε τις εξισώσεις $|\langle \boldsymbol{x}_\ell, \boldsymbol{r}(\lambda)\rangle| = \lambda$ ξεχωριστά για κάθε $\ell \notin \mathcal{A}$, οι οποίες είναι κατά τμήματα γραμμικές ως προς λ, να επιλέξουμε για λ_k τη μεγαλύτερη εξ αυτών των λύσεων και να προχωρήσουμε όπως περιγράφουμε.

Τροποποίηση του LAR για LASSO: Αν ο συντελεστής κάποιας ενεργής επεξηγηματικής μεταβλητής περάσει το 0 κατά τη μετακίνηση του διανύσματος των συντελεστών των ενεργών συμμεταβλητών, σταματάμε τον αλγόριθμο, αφαιρούμε την εν λόγω επεξηγηματική μεταβλητή από το μοντέλο και προχωράμε υπολογίζοντας ξανά το διάνυσμα κατεύθυνσης που προκύπτει από τη παλινδρόμηση ελαχίστων τετραγώνων των εισηγμένων συμμεταβλητών με το τρέχον υπόλοιπο. Η αφαίρεση της εν λόγω συμμεταβλητής δεν εμποδίζει την είσοδο της σε κάποιο επόμενο βήμα.

Για την τροποποίηση LASSO στον αλγόριθμο που παραθέσαμε λοιπόν, τροποποιούμε το βήμα $4\gamma'$. Αφού ανανεώσουμε το διάνυσμα των συντελεστών $\boldsymbol{\beta}^k$, το συγκρίνουμε κατά θέση με το διάνυσμα των συντελεστών $\boldsymbol{\beta}^{k-1}$. Συγκεκριμένα, για τις μη μηδενικές θέσεις του διανύσματος $\boldsymbol{\beta}^{k-1}$ ελέγχουμε αν κάποια τιμή των αντίστοιχων θέσεων του διανύσματος $\boldsymbol{\beta}^k$ έχει αλλάξει πρόσημο, δηλαδή αν ο συντελεστής κάποιας ενεργής, στο βήμα $k-1$, επεξηγηματικής μεταβλητής έχει περάσει από το 0 στην ολοκλήρωση του βήματος k. Αν όχι συνεχίζουμε κατά τα γνωστά. Σε περίπτωση που εντοπίσουμε αλλαγή προσήμου, βρίσκουμε την τιμή εκείνη του λ στην οποία μηδενίζεται ο συντελεστής της ενεργής συμμεταβλητής που άλλαξε πρόσημο και ορίζουμε αυτή η τιμή να είναι ο καινούργιος κόμβος λ_k. Ανανεώνουμε τα $\boldsymbol{\beta}^k$ και $\boldsymbol{r}_k$. Αφαιρούμε τη συμμεταβλητή

της οποίας ο συντελεστής μηδενίστηκε ανανεώνοντας το ενεργό σύνολο, παραλείπουμε το βήμα $4\delta'$ και συνεχίζουμε κανονικά.

Μπορούμε να παραθέσουμε ένα ευριστικό επιχείρημα του γιατί οι διαδικασίες LAR και $LASSO$ είναι τόσο παρόμοιες. Όπως παρατηρήσαμε, στον LAR, για τις ενεργές συμμεταβλητές, ισχύει καθόλη τη διάρκεια του αλγορίθμου ότι:

$$\begin{aligned} &|\langle \boldsymbol{x}_j, \boldsymbol{r}(\lambda)\rangle| = \lambda && \forall j \in \mathcal{A} \\ \Longleftrightarrow &\langle \boldsymbol{x}_j, \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}(\lambda)\rangle = \lambda \cdot sign(\langle \boldsymbol{x}_j, \boldsymbol{r}(\lambda)\rangle) && \forall j \in \mathcal{A} \end{aligned}$$

ενώ για τις ανενεργές συμμεταβλητές ισχύει πως:

$$|\langle \boldsymbol{x}_j, \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}(\lambda)\rangle| \leq \lambda \quad \forall j \notin \mathcal{A}.$$

Ας θυμηθούμε τώρα το κριτήριο $LASSO$:

$$R(\boldsymbol{\beta}) = \frac{1}{2}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1$$

όπου τοποθετήσαμε τον αριθμό $\frac{1}{2}$ μπροστά από το άθροισμα τετραγώνων των υπολοίπων ώστε να συμφωνήσουν τα αποτελέσματά μας με αυτά του LAR. Ουσιαστικά, η μόνη αλλαγή στα αποτελέσματα των προηγούμενων παραγράφων υπεισέρχεται στην παράμετρο ποινής λ, η οποία πλέον διπλασιάζεται. Βάσει λοιπόν αυτής της παραδοχής και χωρίς βλάβη της γενικότητας, αν $\mathcal{B}$ είναι το σύνολο των δεικτών των ενεργών συμμεταβλητών του μοντέλου μας, δείξαμε ήδη στην Παράγραφο 6.3.3 ότι η συνθήκη στασιμότητας για τις ενεργές επεξηγηματικές μεταβλητές μας δίνει:

$$\langle \boldsymbol{x}_j, \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\rangle = \lambda \cdot sign(\beta_j) \quad \forall j \in \mathcal{B},$$

το οποίο είναι ταυτόσημο με την συνθήκη που ικανοποιούν οι ενεργές συμμεταβλητές του LAR μόνο όταν το πρόσημο του συντελεστή β_j είναι ίδιο με το πρόσημο του εσωτερικού γινομένου $\langle \boldsymbol{x}_j, \boldsymbol{r}(\lambda)\rangle$. Για αυτόν ακριβώς το λόγο ο αλγόριθμος LAR και η διαδικασία $LASSO$ διαφοροποιούνται όταν μια ενεργή επεξηγηματική μεταβλητή περάσει από το μηδέν. Τα πρόσημα του συντελεστή αυτής της συμμεταβλητής και του εσωτερικού γινομένου παύουν να είναι ίδια, οπότε αφαιρούμε τη συγκεκριμένη επεξηγηματική από το ενεργό σύνολο στο βήμα $4\gamma'$ όπως περιγράψαμε στην τροποποίηση του LAR για $LASSO$. Για τις ανενεργές συμμεταβλητές της $LASSO$, η συνθήκη στασιμότητας (Παράγραφος 6.3.3), δίνει:

$$|\langle \boldsymbol{x}_j, \boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\rangle| \leq \lambda \quad \forall j \notin \mathcal{B},$$

που συμφωνεί απόλυτα με τη συνθήκη των ανενργών επεξηγηματικών του αλγορίθμου LAR. Ο LAR υπολογίζει το πλήρες φάσμα των λύσεων $LASSO$, το οποίο είναι κατά τμήματα γραμμικό για όλους τους συντελεστές.

6.4.2 Coordinate Descent

Συγκεκριμένες ομάδες προβλημάτων, ανάμεσά τους και το πρόβλημα $LASSO$ καθώς και παραλλαγές του, έχουν μια επιπλέον ιδιότητα διαχωρισιμότητας, μέσω της οποίας προκύπτει, με φυσικό τρόπο, ένας αλγόριθμος ελαχιστοποίησης κατά συντεταγμένη. Ο αλγόριθμος Coordinate Descent είναι μια τέτοιου είδους επαναληπτική διαδικασία η οποία ανανεώνει το διάνυσμα των συντελεστών από το $\boldsymbol{\beta}^t$ στο $\boldsymbol{\beta}^{t+1}$ διαλέγοντας μόνο μια συντεταγμένη του διανύσματος των συντελεστών για ανανέωση κάθε φορά. Συγκεκριμένα, αν στην επανάληψη $t+1$ επιλεχθεί ο συντελεστής j για ανανέωση, το διάνυσμα των συντελεστών ανανεώνεται ακολούθως:

$$\beta_j^{t+1} = \underset{\beta_j \in \mathbb{R}}{argmin}\, f(\beta_1^t, \beta_2^t .., \beta_{j-1}^t, \beta_j, \beta_{j+1}^t, .., \beta_p^t)$$
$$\beta_k^{t+1} = \beta_k^t \quad \forall k \neq j$$

Μια συνηθισμένη επιλογή είναι σε κάθε γύρο επαναλήψεων να κάνουμε έναν κύκλο γύρω από τους συντελεστές όλων των επεξηγηματικών μεταβλητών με μια σταθερή σειρά, έτσι ώστε μετά από p επαναλήψεις να έχουν ανανεωθεί όλοι οι συντελεστές από μια φορά, μετά από $2p$ επαναλήψεις να έχουν ανανεωθεί όλοι οι συντελεστές δυο φορές, κ.ο.κ. Για αυτόν τον λόγο ο εν λόγω αλγόριθμος συναντάται πολλές φορές με την ονομασία Cyclical Coordinate Descent ή Pathwise Coordinate Descent. Η χρησιμοποίησή του στο πρόβλημα $LASSO$ είχε υποτιμηθεί εώς ότου μελετήθηκε περαιτέρω η αποδοτικότητά του και η υπολογιστική του δεινότητα από τους Friedman, Hastie, Hofling και Tibshirani το 2007. Η ιδιότητα διαχωρισιμότητας η οποία πρέπει να ικανοποιείται ώστε να είναι αποδοτικός και εν τέλει να συγκλίνει ο αλγόριθμος είναι η ακόλουθη. Έστω f η συνάρτηση που θέλουμε να ελαχιστοποιήσουμε. Αν η f γράφεται ως εξής:

$$f(\boldsymbol{\beta}) = g(\boldsymbol{\beta}) + \sum_{j=1}^{p} h_j(\beta_j)$$

όπου $g : \mathbb{R}^p \to \mathbb{R}$ διαφορίσιμη και κυρτή και οι $h_j : \mathbb{R} \to \mathbb{R}$ κυρτές αλλά όχι απαραίτητα διαφορίσιμες, τότε ικανοποιείται η ιδιότητα διαχωρισιμότητας που αναφέραμε και μπορούμε να προβούμε στην ανανέωση των συντελεστών ελαχιστοποιώντας κατά συντεταγμένη κάθε φορά, καθώς εξασφαλίζεται ότι ο αλγόριθμος θα συγκλίνει και δεν θα «εγκλωβιστούμε» σε κάποιο τοπικό ελάχιστο. Το πρόβλημα $LASSO$ που μελετάμε μπορεί να γραφτεί κατά αυτόν τον τρόπο, όπου $f(\boldsymbol{\beta}) = PRSS(\boldsymbol{\beta})_{\ell_1}$, $g(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ και $h_j(\beta_j) = \lambda|\beta_j|$. Οπότε επόμενό μας βήμα είναι να υπολογίσουμε ποιά θα είναι η ανανέωση της

εκτίμησης για τον τυχόντα συντελεστή β_j. Στόχος μας είναι αρχικά να διαχωρίσουμε τον β_j από τους υπόλοιπους συντελεστές. Θα χρησιμοποιήσουμε για αυτόν τον σκοπό το μερικό υπόλοιπο:

$$r_i^{(j)} = y_i - \sum_{k \neq j} x_{ik}\hat{\beta}_k,$$

της i παρατήρησης χωρίς την j μεταβλητή. Το ποινικοποιημένο ως προς την ℓ_2-νόρμα άθροισμα τετραγώνων των υπολοίπων μπορεί να γραφεί ως εξής:

$$\begin{aligned} PRSS(\boldsymbol{\beta})_{\ell_1} &= \sum_{i=1}^{n}\Big(y_i - \sum_{j=1}^{p} x_{ij}\beta_j\Big)^2 + \lambda\sum_{j=1}^{p}|\beta_j| \\ &= \sum_{i=1}^{n}\Big(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\Big)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda|\beta_j| \end{aligned}$$

.

1η περίπτωση: Αν $\beta_j > 0$ έχουμε:

$$PRSS(\boldsymbol{\beta})_{\ell_1} = \sum_{i=1}^{n}\Big(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\Big)^2 + \lambda\sum_{k\neq j}|\beta_k| + \lambda\beta_j$$

οπότε μπορούμε να παραγωγίσουμε ως προς β_j και λαμβάνουμε:

$$\frac{\partial PRSS(\boldsymbol{\beta})_{\ell_1}}{\partial\beta_j} = -2\sum_{i=1}^{n}\Big(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\Big)x_{ij} + \lambda.$$

Θέτοντας την παραπάνω ποσότητα ίση με μηδέν και λύνοντας ως προς β_j λαμβάνουμε την εκτίμηση:

$$\begin{aligned} &-2\sum_{i=1}^{n}\Big(y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k - x_{ij}\hat{\beta}_j\Big)x_{ij} + \lambda = 0 \\ \iff &-2\sum_{i=1}^{n}\Big(r_i^{(j)} - x_{ij}\hat{\beta}_j\Big)x_{ij} + \lambda = 0 \\ \iff &\hat{\beta}_j = \frac{\sum_{i=1}^{n} x_{ij}r_i^{(j)}}{\sum_{i=1}^{n} x_{ij}^2} - \frac{\lambda}{2\sum_{i=1}^{n} x_{ij}^2}. \end{aligned}$$

Όμως $\sum_{i=1}^{n} x_{ij}r_i^{(j)} = \boldsymbol{x}_j^T\boldsymbol{r}^{(j)} = \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)}\rangle$ και $\sum_{i=1}^{n} x_{ij}^2 = 1$ αφού έχουμε τυποποιήσει τα δεδομένα μας, άρα:

$$\hat{\beta}_j = \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)}\rangle - \frac{\lambda}{2}.$$

Μάλιστα, για να είναι αυτή η λύση εφικτή, πρέπει $\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle > \dfrac{\lambda}{2}$, αφού θεωρήσαμε ότι $\beta_j > 0$. Άρα γράφουμε την ανανέωση της εκτίμησης του β_j στη μορφή:

$$\begin{aligned}\hat{\beta}_j &= \Big(\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle - \frac{\lambda}{2}\Big)_+ \\ &= sign\Big(\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle\Big)\Big(|\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle| - \frac{\lambda}{2}\Big)_+.\end{aligned}$$

2η περίπτωση: Αν $\beta_j < 0$ έχουμε:

$$PRSS(\boldsymbol{\beta})_{\ell_1} = \sum_{i=1}^{n}\Big(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\Big)^2 + \lambda\sum_{k\neq j}|\beta_k| - \lambda\beta_j$$

οπότε μπορούμε να παραγωγίσουμε ως προς β_j και λαμβάνουμε:

$$\frac{\partial PRSS(\boldsymbol{\beta})_{\ell_1}}{\partial \beta_j} = -2\sum_{i=1}^{n}\Big(y_i - \sum_{k\neq j} x_{ik}\beta_k - x_{ij}\beta_j\Big)x_{ij} - \lambda.$$

Θέτοντας την παραπάνω ποσότητα ίση με μηδέν και λύνοντας ως προς β_j λαμβάνουμε την εκτίμηση:

$$\begin{aligned}&-2\sum_{i=1}^{n}\Big(y_i - \sum_{k\neq j} x_{ik}\hat{\beta}_k - x_{ij}\hat{\beta}_j\Big)x_{ij} - \lambda = 0 \\ \iff &-2\sum_{i=1}^{n}\Big(r_i^{(j)} - x_{ij}\hat{\beta}_j\Big)x_{ij} - \lambda = 0 \\ \iff &\hat{\beta}_j = \frac{\sum_{i=1}^{n} x_{ij}r_i^{(j)}}{\sum_{i=1}^{n} x_{ij}^2} + \frac{\lambda}{2\sum_{i=1}^{n} x_{ij}^2}.\end{aligned}$$

Όμως $\sum_{i=1}^{n} x_{ij}r_i^{(j)} = \boldsymbol{x}_j^T\boldsymbol{r}^{(j)} = \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle$ και $\sum_{i=1}^{n} x_{ij}^2 = 1$ αφού έχουμε τυποποιήσει τα δεδομένα μας, άρα:

$$\hat{\beta}_j = \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle + \frac{\lambda}{2}.$$

Μάλιστα, για να είναι αυτή η λύση εφικτή, πρέπει $\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle < -\dfrac{\lambda}{2}$, αφού θεωρήσαμε ότι $\beta_j < 0$. Άρα γράφουμε την ανανέωση της εκτίμησης του β_j στη μορφή:

$$\begin{aligned}\hat{\beta}_j &= \Big(\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle + \frac{\lambda}{2}\Big)_- \\ &= sign\Big(\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle\Big)\Big(|\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle| - \frac{\lambda}{2}\Big)_+.\end{aligned}$$

Και για τις δύο περιπτώσεις που διακρίναμε καταλήξαμε στον ίδιο τύπο για την ανανέωση της εκτίμησης του συντελεστή β_j. Οπότε, γενικεύοντας, η ανανέωση του συντελεστή β_j είναι ίση με

$$\hat{\beta}_j = sign\Big(\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle\Big)\Big(|\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle| - \frac{\lambda}{2}\Big)_+$$

$$= \begin{cases} \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle - \dfrac{\lambda}{2} & \text{αν} \quad \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle > \dfrac{\lambda}{2} \\ 0 & \text{αν} \quad |\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle| \leq \dfrac{\lambda}{2} \\ \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle + \dfrac{\lambda}{2} & \text{αν} \quad \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle < -\dfrac{\lambda}{2}. \end{cases}$$

Άρα τελικά η ανανέωση του συντελεστή β_j στον αλγόριθμο Coordinate Descent είναι:

$$\hat{\beta}_j = \mathcal{S}\Big(\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle, \frac{\lambda}{2}\Big)$$

όπου:

$$\mathcal{S}(\theta, \lambda) = sign(\theta)(|\theta| - \lambda)_+ = \begin{cases} \theta - \lambda & \text{αν} \quad \theta > \lambda \\ 0 & \text{αν} \quad |\theta| \leq \lambda \\ \theta + \lambda & \text{αν} \quad \theta < -\lambda \end{cases}$$

ο τελεστής $soft - thresholding$ (Donoho and Johnstone, 1994). Μάλιστα αξίζει να παρατηρήσουμε ότι:

$$\begin{aligned} \langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle &= \sum_{i=1}^{n} x_{ij}\Big(y_i - \sum_{k \neq j} x_{ij}\hat{\beta}_k\Big) \\ &= \sum_{i=1}^{n} x_{ij}\Big(y_i - \sum_{k \neq j} x_{ij}\hat{\beta}_k - x_{ij}\hat{\beta}_j + x_{ij}\hat{\beta}_j\Big) \\ &= \sum_{i=1}^{n} x_{ij}(r_i + x_{ij}\hat{\beta}_j) \\ &= \hat{\beta}_j \sum_{i=1}^{n} x_{ij}^2 + \boldsymbol{x}_j^T \boldsymbol{r} \\ &= \hat{\beta}_j + \langle \boldsymbol{x}_j, \boldsymbol{r} \rangle. \end{aligned}$$

Οπότε η ανανέωση του αλγορίθμου Coordinate Descent μπορεί ισοδύναμα να είναι και η εξής:

$$\hat{\beta}_j \leftarrow \mathcal{S}\Big(\hat{\beta}_j + \langle \boldsymbol{x}_j, \boldsymbol{r} \rangle, \frac{\lambda}{2}\Big).$$

Να ξεκαθαρίσουμε εδώ, ότι και στις δύο ισοδύναμες ανανεώσεις, αν είμαστε στην επανάληψη $t+1$ και έχει επιλεχθεί ο συντελεστής β_j για ανανέωση, τα δεξιά μέλη των ανανεώσεων εμπεριέχουν τις τιμές που έχουν οι συντελεστές στην επανάληψη t και τα αριστερά μέλη αναφέρονται στην εκτίμηση $\hat{\beta}_j^{t+1}$:

$$\begin{aligned} \hat{\beta}_j^{t+1} &\leftarrow \mathcal{S}\Big(\hat{\beta}_j^{t} + \langle \boldsymbol{x}_j, \boldsymbol{r}^t \rangle, \frac{\lambda}{2}\Big) \\ \hat{\beta}_k^{t+1} &\leftarrow \beta_k^t \quad \forall k \neq j. \end{aligned}$$

Ο Coordinate Descent υπολογίζει τις λύσεις $LASSO$ για συγκεκριμένες τιμές της παραμέτρου ποινής, σε αντίθεση με τον Least Angle Regression που υπολογίζει το πλήρες φάσμα των λύσεων $LASSO$.

6.5 Ειδικές περιπτώσεις μοντέλων

Σε αυτή τη Παράγραφο θα μελετήσουμε τις ειδικές περιπτώσεις της μιας επεξηγηματικής μεταβλητής και του ορθογώνιου πίνακα σχεδιασμού. Θα χρησιμοποιήσουμε για αυτόν τον σκοπό την ανανέωση του συντελεστή β_j από τον αλγόριθμο Cordinate Descent:

$$\hat{\beta}_j = \mathcal{S}\Big(\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle, \frac{\lambda}{2}\Big)$$

και θα δούμε ότι και στις δυο περιπτώσεις η έννοια της ανανέωσης παύει να υφίσταται, οπότε καταλήγουμε σε τύπο κλειστής μορφής για την εκτιμήτρια $Lasso$.

6.5.1 Μια επεξηγηματική μεταβλητή

Ας υποθέσουμε ότι το μοντέλο μας αποτελείται από μια επεξηγηματική μεταβλητή έστω $\boldsymbol{X}$ με τιμές $\boldsymbol{x}$ και συντελεστή β. Σε αυτήν την περίπτωση το μερικό υπόλοιπο δεν ορίζεται, καθώς δεν έχουμε άλλες επεξηγηματικές μεταβλητές, και την θέση του στο εσωτερικό γινόμενο παίρνει το διάνυσμα τιμών της μεταβλητής απόκρισης, $\boldsymbol{y}$. Όλα τα στοιχεία που εμπεριέχονται στον τελεστή $soft-thresholding$ είναι γνωστά και δεν επιδέχονται ανανέωσης. Οπότε καταλήγουμε κατευθείαν στον κλειστό τύπο της εκτιμήτριας $Lasso$ για τη μοναδική

μας επεξηγηματική μεταβλητή:

$$\hat{\beta}^{Lasso} = \mathcal{S}\Big(\langle \boldsymbol{x}, \boldsymbol{y} \rangle, \frac{\lambda}{2}\Big).$$

Αξίζει να παρατηρήσουμε ότι όταν έχουμε μόνο μια επεξηγηματική μεταβλητή και τυποποιημένα δεδομένα, η εκτιμήτρια ελαχίστων τετραγώνων είναι ίση με:

$$\begin{aligned}
\hat{\beta}^{OLS} &= (\boldsymbol{x}^T \boldsymbol{x})^{-1} \boldsymbol{x}^T \boldsymbol{y} \\
&= (\sum_{i=1}^{n} x_i^2)^{-1} \boldsymbol{x}^T \boldsymbol{y} \\
&= \boldsymbol{x}^T \boldsymbol{y} \\
&= \langle \boldsymbol{x}, \boldsymbol{y} \rangle.
\end{aligned}$$

Οπότε μπορούμε να γράψουμε το αποτέλεσμα της εκτιμήτριας $LASSO$ για το συντελεστή της μοναδικής μας συμμεταβλητής και ακολούθως:

$$\hat{\beta}^{Lasso} = \mathcal{S}\Big(\hat{\beta}^{OLS}, \frac{\lambda}{2}\Big)$$

όπου: $\mathcal{S}(\theta, \lambda) = sign(\theta)(|\theta| - \lambda)_+$.

6.5.2 Ορθογώνιος πίνακας σχεδιασμού

Ας υποθέσουμε ότι ο πίνακας σχεδιασμού $\boldsymbol{X}$ του μοντέλου μας είναι ορθογώνιος, δηλαδή ισχύει ότι $\boldsymbol{X}^T \boldsymbol{X} = \boldsymbol{I}_p$. Ισοδύναμα, ισχύει ότι:

$$\langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle = 0 \qquad \forall j \neq k.$$

Έτσι λοιπόν το εσωτερικό γινόμενο της μεταβλητής $\boldsymbol{x}_j$ με το μερικό υπόλοιπο $\boldsymbol{r}^{(j)}$ απλοποιείται:

$$\begin{aligned}
\langle \boldsymbol{x}_j, \boldsymbol{r}^{(j)} \rangle &= \sum_{i=1}^{n} x_{ij} \Big(y_i - \sum_{k \neq j} x_{ik} \beta_k \Big) \\
&= \sum_{i=1}^{n} x_{ij} y_i - \sum_{k \neq j} \sum_{i=1}^{n} x_{ij} x_{ik} \beta_k \\
&= \langle \boldsymbol{x}_j, \boldsymbol{y} \rangle - \sum_{k \neq j} \beta_k \langle \boldsymbol{x}_j, \boldsymbol{x}_k \rangle \\
&= \langle \boldsymbol{x}_j, \boldsymbol{y} \rangle.
\end{aligned}$$

Παρατηρούμε ότι τα στοιχεία του εσωτερικού γινομένου είναι σταθερά και δεν επιδέχονται ανανέωσης, οπότε καταλήγουμε κατευθείαν σε κλειστό τύπο λύσης για την εκτιμήτρια $LASSO$ του συντελεστή της μεταβλητής $\boldsymbol{x}_j$:

$$\hat{\beta}_j^{Lasso} = \mathcal{S}\Big(\langle \boldsymbol{x}_j, \boldsymbol{y}\rangle, \frac{\lambda}{2}\Big).$$

Μάλιστα, στην περίπτωση του ορθογώνιου πίνακα σχεδιασμού, η εκτιμήτρια ελαχίστων τετραγώνων είναι ίση με:

$$\begin{aligned}\hat{\boldsymbol{\beta}}^{OLS} &= (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y} \\ &= \boldsymbol{X}^T\boldsymbol{y}\end{aligned}$$

άρα ισοδύναμα:

$$\hat{\beta}_j^{OLS} = \langle \boldsymbol{x}_j, \boldsymbol{y}\rangle.$$

Οπότε το αποτέλεσμα της εκτιμήτριας $Lasso$ για τον συντελεστή β_j στην περίπτωση του ορθογώνιου πίνακα σχεδιασμού γράφεται και ακολούθως:

$$\hat{\beta}_j^{Lasso} = \mathcal{S}\Big(\hat{\beta}_j^{OLS}, \frac{\lambda}{2}\Big)$$

όπου: $\mathcal{S}(\theta, \lambda) = sign(\theta)(|\theta| - \lambda)_+$.

Παρατηρούμε ότι η εκτιμήτρια $LASSO$ όταν έχουμε ορθογώνιο πίνακα σχεδιασμού, συρρικνώνει τους μεγαλύτερους (κατ' απόλυτη τιμή) συντελεστές, κατά έναν σταθερό παράγοντα $\dfrac{\lambda}{2}$, ενώ θέτει τους μικρότερους (πάλι κατ' απόλυτη τιμή) συντελεστές ίσους με μηδέν. Βέβαια η αναγνώριση των μεγαλύτερων και μικρότερων συντελεστών είναι εξαρτημένη, όπως βλέπουμε παραπάνω, από τον παράγοντα συρρίκνωσης λ, για διαφορετικές τιμές του οποίου λαμβάνουμε διαφορετικές εκτιμήσεις $LASSO$, άρα και διαφορετικά μοντέλα.

Γίνεται λοιπόν σαφές το συγκριτικό πλεονέκτημα της μεθοδολογίας $Lasso$ έναντι της παλινδρόμησης $Ridge$. Θέτει τους λιγότερο στατιστικά σημαντικούς συντελεστές ίσους με το μηδέν διευκολύνοντας την επιλογή μεταβλητών, σε αντίθεση με την $Ridge$ $\left(\hat{\boldsymbol{\beta}}_\lambda^{Ridge} = \dfrac{1}{1+\lambda}\hat{\boldsymbol{\beta}}^{OLS}\right)$ η οποία ναι μεν συρρικνώνει τους συντελεστές, αλλά δεν θέτει κάποιον ίσον με το μηδέν.

6.6 Η παράμετρος ποινής λ, ο παράγοντας συρρίκωσης s και μέθοδοι επιλογής τους

Η μέθοδος $LASSO$ ποινικοποιεί τους συντελεστές του γραμμικού μοντέλου θέτοντας ένα άνω φράγμα t στην l_1-νόρμα του διανύσματος των συντελεστών $\boldsymbol{\beta}$. Το t και κατ' επέκτασιν, με 1-1 αντιστοιχία, το λ αποτελούν τις παραμέτρους ποινικοποίησης. Όσο το λ αυξάνει, δηλαδή όσο το t τείνει στο μηδέν, ισχύει πως:

$$\|\boldsymbol{\beta}\|_1 = \sum_{j=1}^{p} |\beta_j| \longrightarrow 0$$

με τους συντελεστές των επεξηγηματικών μεταβλητών που δεν είναι στατιστικά σημαντικές να τείνουν στο μηδέν γρηγορότερα. Αν θέσουμε το t ίσο με $max\|\boldsymbol{\beta}\|_1$, ή αντίστοιχα το λ ίσο με μηδέν, δεν επιτυγχάνουμε καμία συρρίκνωση των συντελεστών οπότε λαμβάνουμε τους εκτιμητές ελαχίστων τετραγώνων. Άρα ισχύει πως:

$$max\|\boldsymbol{\beta}\|_1 = \|\hat{\boldsymbol{\beta}}^{OLS}\|_1$$

Αξίζει να εισάγουμε σε αυτό το σημείο τον παράγοντα συρρίκνωσης s που παίρνει τιμές στο $[0, 1]$:

$$s = \frac{\|\boldsymbol{\beta}\|_1}{max\|\boldsymbol{\beta}\|_1}$$

Ουσιαστικά ο παράγοντας συρρίκνωσης s παρουσιάζει πόσο μικρότερο είναι το άθροισμα των απολύτων των τυχόντων εκτιμήσεων από το άθροισμα των απολύτων των εκτιμήσεων που λαμβάνουμε με τη μέθοδο των ελαχίστων τετραγώνων. Όταν $s = 1$ δεν έχουμε ποινικοποίηση άρα αποκομίζουμε την εκτιμήτρια ελαχίστων τετραγώνων ενώ όταν $s = 0$, $\beta_j = 0\ \forall j = 1, , p$.

Για διαφορετικές τιμές της παραμέτρου ποινής λ, άρα και με 1-1 αντιστοιχία του παράγοντα συρρίκνωσης s, λαμβάνουμε διαφορετικές εκτιμήσεις για τους συντελεστές του μοντέλου μας, οι οποίες απεικονίζονται γραφικώς μέσω των μονοπατιών κανονικοποίησης, που αναφέραμε και στη παλινδρόμηση $Ridge$. Άρα σκοπός μας είναι να επιλέξουμε την κατάλληλη τιμή της παραμέτρου ποινής (ή του παράγοντα συρρίκνωσης), μέσω των μεθοδολογιών που αναφέρουμε παρακάτω.

6.6.1 Mallows' C_p

Το στατιστικό Mallows' C_p αποτελεί ένα μέτρο για την αξιολόγηση της καταλληλότητας ενός μοντέλου παλινδρόμησης (Mallows 1973). Χρησιμοποιείται ως κριτήριο σύγκρισης διαφορετικών μοντέλων αποτυπώνοντας την προβλεπτική τους ικανότητα. Έστω p_m ένα υποσύνολο των p επεξηγηματικών μεταβλητών. Τότε για το μοντέλο που εμπεριέχει μόνο αυτές τις p_m συμμεταβλητές, το στατιστικό Mallows' C_p ισούται με:

$$C_{p_m} = \frac{RSS_{p_m}}{\hat{\sigma}_{\varepsilon}^{full}} + 2p_m - n$$

όπου RSS_{p_m} το άθροισμα τετραγώνων των υπολοίπων του μοντέλου με τις p_m επεξηγηματικές μεταβλητές, p_m το πλήθος αυτών των συμμεταβλητών και $\hat{\sigma}_{\varepsilon}^{full}$ η εκτίμηση της διασποράς του μοντέλου που εμπεριέχει όλες τις p επεξηγηματικές μεταβλητές. Παρέχοντας λοιπόν ένα ικανοποιητικά μεγάλο εύρος τιμών της παραμέτρου ποινής λ, λαμβάνουμε διαφορετικά μοντέλα για τα οποία υπολογίζουμε το στατιστικό Mallows' C_p και διαλέγουμε την παράμετρο ποινής εκείνη που παρήγαγε το μοντέλο με το μικρότερο C_p. Παρ' όλα αυτά, δυστυχώς το Mallows' C_p δεν μπορεί να αποτυπώσει καλά την προβλεπτική ικανότητα περίπλοκων εκ φύσεως μοντέλων (πολυσυγγραμικότητα, $p >> n$, κ.ο.κ.), γι'αυτό και δεν προτιμάται συχνά.

6.6.2 Cross Validation

Όπως και στην παλινδρόμηση *Ridge*, η πιο διαδεδομένη μεθοδολογία για την επιλογή της παραμέτρου ποινής λ και κατ' αντιστοιχία της παραμέτρου συρρίκνωσης s, είναι αυτή του Cross Validation που έχουμε περιγράψει στο Κεφάλαιο 4. Συγκεκριμένα, για κάθε διαφορετική τιμή της παραμέτρου ποινής λαμβάνουμε και διαφορετικό μοντέλο στο οποίο εφαρμόζουμε K-fold Cross Validation (Παράγραφος 4.3):

$$CV_{K-fold} = \overline{MSE} = \frac{1}{K}\sum_{k=1}^{K} MSE(F_k),$$

με αριθμό φακέλων K της επιλογής μας. Διαλέγουμε στη συνέχεια την παράμετρο ποινής εκείνη που παρήγαγε το μοντέλο με το μικρότερο CV_{K-fold}, άρα και το μοντέλο με την καλύτερη «ικανότητα πρόβλεψης» βάσει του μέτρου που θεσπίσαμε. Προς το παρόν λοιπόν κάνουμε ο,τι ακριβώς και στην *Ridge* για την επιλογή της παραμέτρου ποινής. Παρόλα αυτά, εφαρμόζοντας K-fold Cross Validation, μπορούμε να αποκτήσουμε ένα ακόμα πιο φειδωλό μοντέλο, από αυτό που μας δίνει το μικρότερο CV_{K-fold}, με εξίσου ικανοποιητική προβλεπτική

ικανότητα. Κάθε CV_{K-fold} έχει και ένα τυπικό σφάλμα:

$$\begin{aligned}
\hat{se}(CV_{K-fold}) &= \sqrt{Var(CV_{K-fold})} \\
&= \sqrt{Var\left(\frac{1}{K}\sum_{k=1}^{K} MSE(F_k)\right)} \\
&= \sqrt{\frac{1}{K^2} Var\left(\sum_{k=1}^{K} MSE(F_k)\right)} \\
&= \sqrt{\frac{1}{K^2}\sum_{k=1}^{K} Var\Big(MSE(F_k)\Big)} \\
&= \sqrt{\frac{1}{K} Var\Big(MSE(F_k)\Big)} \\
&= \frac{1}{\sqrt{K}}\sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\Big(MSE(F_k) - \overline{MSE}\Big)^2}
\end{aligned}$$

Αφότου λοιπόν εντοπίσουμε το μικρότερο CV_{K-fold}, έστω $min(cv)$, εξετάζουμε όλα τα μοντέλα που έχουν CV_{K-fold} εντός του εύρους τιμών που ορίζεται από το τυπικό σφάλμα του μικρότερου:

$$\Big[min(cv) - \hat{se}(min(cv)), min(cv) + \hat{se}(min(cv))\Big]$$

και επιλέγουμε το μοντέλο εκείνο που παρήχθει από την μεγαλύτερη παράμετρο ποινής λ ή αντίστοιχα την μικρότερη παράμετρο συρρίκνωσης s. Η εν λόγω επιλογή είναι γνωστή ως "1 standard error rule" (1se) και μας δίνει ένα μοντέλο με λιγότερες επεξηγηματικές μεταβλητές του οποίου το CV_{K-fold} δεν απέχει πολύ από το ελάχιστο, άρα ανταποκρίνεται εξίσου καλά στην απόδοση πρόβλεψης.

6.7 LASSO στην R

Στην R μπορούμε να εφαρμόσουμε τη μεθοδολογία $LASSO$ με δύο πακέτα. Το πακέτο $lars$ που εφαρμόζει τον αλγόριθμο Least Angle Regression και το πακέτο $glmnet$ που εφαρμόζει τον αλγόριθμο Coordinate Descent. Παρακάτω θα εξετάσουμε τα δύο πακέτα ξεχωριστά, εφαρμόζοντας τη μεθοδολογία $LASSO$ για τη μελέτη του προβλήματος της εξέλιξης της νόσου του διαβήτη.

LARS package

Το πακέτο $lars$ εφαρμόζει τον αλγόριθμο Least Angle Regression, που αναλύσαμε στην Παράγραφο 6.4.1. Παραθέτουμε τις πιο σημαντικές εντολές του πακέτου $lars$:

- $lars()$: Υλοποιεί τον αλγόριθμο Least Angle Regression. Δέχεται ως ορίσματα τον πίνακα με τα δεδομένα των επεξηγηματικών μεταβλητών και τα δεδομένα της μεταβλητής απόκρισης. Δέχεται επίσης ως όρισμα τη μέθοδο που θα χρησιμοποιήσει στην εκτέλεση του αλγορίθμου Least Angle Regression, με επιλογές τις "$lasso$", "lar", "$forward.stagewise$", και "$stepwise$". Αυτόματη προεπιλογή είναι η "$lasso$". Επιστρέφει ως αποτέλεσμα το ποιά συμμεταβλητή προσέθεσε ή αφαίρεσε ο αλγόριθμος σε κάθε βήμα καθώς και μια λίστα στοιχείων που μεταξύ άλλων περιέχει τους κόμβους της παραμέτρου ποινής λ_k, το διάνυσμα των συντελεστών $\boldsymbol{\beta}^k$ καθώς και το στατιστικό ελέγχου Mallows' C_p σε κάθε k βήμα του αλγορίθμου.

- $cv.lars()$: Πραγματοποιεί K- fold Cross Validation για τα διαφορετικά μοντέλα που προκύπτουν από 100 διαφορετικές τιμές του παράγοντα συρρίκνωσης s. Η ακολουθία των τιμών του παράγοντα συρρίκνωσης μπορεί να δοθεί και από τον χρήστη με το όρισμα $index$. Επιστρέφει μια λίστα στοιχείων που μεταξύ άλλων περιέχει τα CV_{K-fold}, τις τυπικές αποκλίσεις τους, και τις τιμές του παράγοντα συρρίκνωσης s για τις οποίες παρήχθησαν τα διαφορετικά μοντέλα.

- $coef()$: Υπολογίζει τις εκτιμήτριες $LASSO$ των συντελεστών των επεξηγηματικών μεταβλητών. Δέχεται ως ορίσματα την υλοποίηση του αλγορίθμου, δηλαδή την εντολή $lars()$ που εκτελέσαμε, $s =$ τις τιμές που ορίζουν το μονοπάτι $LASSO$ στις οποίες θέλουμε να υπολογίσουμε τις εκτιμήτριες $LASSO$ και $mode =$ "$fraction$" αν αυτές οι τιμές αναφέρονται στον παράγοντα συρρίκνωσης, $mode =$ "$step$" αν αυτές οι τιμές αναφέρονται στα βήματα του αλγορίθμου, $mode =$ "$norm$" αν αναφέρονται στις τιμές της ℓ_1 νόρμας των συντελεστών ενώ $mode =$ "$lambda$" αν αναφέρονται στις τιμές της παραμέτρου ποινής λ.

Κατασκευάζουμε λοιπόν στον Κώδικα 6.1 τη συνάρτηση $lasso.lars()$ που δέχεται ως ορίσματα τα δεδομένα των συμμεταβλητών και της μεταβλητής απόκρισης και υλοποιεί τον αλγόριθμο Least Angle Regression. Εντοπίζει το βήμα στο οποίο ελαχιστοποιείται το κριτήριο $Mallows'\ C_p$, Παράγραφος 6.6.1, υπολογίζει τις εκτιμήτριες $LASSO$ που του αντιστοιχούν και έπειτα τον παράγοντα συρρίκνωσης που παρήγαγε το εν λόγω μοντέλο. Στη συνέχεια εφαρμόζει r

(επίσης όρισμα της συνάρτησης lasso.lars) γύρους 5-fold Cross Validation αποθηκεύοντας τις τιμές των CV_{5-fold} κάθε γύρου και τα σφάλματά τους για κάθε μοντέλο ώστε να λάβουμε στη συνέχεια τις μέσες τιμές τους και τις διασπορές τους και να εντοπίσουμε τους παράγοντες συρρίκνωσης που αντιστοιχούν στο μικρότερο μέσο CV_{5-fold} και στον κανόνα «1 standard error» που αναπτύξαμε στην Παράγραφο 6.6.2. Υπολογίζει τις εκτιμήτριες $LASSO$ για αυτές τις τιμές του παράγοντα συρρίκνωσης.

```
#lasso with lars---> function
install.packages("lars")
library(lars)

lasso.lars<-function(X,Y,r){
  reg<- lm(Y~., data=X)
  #ols coefficients for original data
  b.ols<- coef(reg)[-1]
  #ols coefficients for standardized data
  zb.ols<- b.ols*apply(X,2,sd)

  #lasso
  lasso<- lars(as.matrix(X),Y)

  #coefficients for s that minimizes Cp original data
  b.lasso<-coef(lasso, s=which.min(lasso$Cp), mode="step
     ")
  #coefficients for s that minimizes Cp standardized
     data
  zb.lasso<-b.lasso*apply(X,2,sd)
  #s that minimizes Cp
  min.cp.s<- sum(abs(zb.lasso))/sum(abs(zb.ols))

  cvk<- matrix(rep(NA,r*100),ncol=100,nrow=r)
  var<- matrix(rep(NA,r*100),ncol=100,nrow=r)
  #r repetitions of 5-fold cv for various s
  for(i in 1:r){
    rescv<-cv.lars(as.matrix(X),Y,plot.it=F,K=5)
    cvk[i,]<-rescv$cv
    var[i,]<-rescv$cv.error^2
  }
  cv<-apply(cvk,2,mean)
  sds<-apply(var,2,mean)
  sds<-sqrt(sds)
```

```
33
34    #s that minimizes the average 5-fold cv
35    min.cv.s<-lambda[which.min(cv)]
36    #coefficients for s that minimizes the average 5-fold
         cv
37    coef.cv.s<-coef(lasso,s=min.cv.s,mode="fraction")
38
39    #1 standard error rule
40    range<-c(min(cv)-sds[lambda==min.cv.s],min(cv)+sds[
         lambda==min.cv.s])
41    j<-which(cv>=range[1]&cv<=range[2])
42    #s of 1 standard error rule
43    s1se<-min(lambda[j])
44    #coefficients for s of 1 standard error rule
45    coef1se<-coef(lasso,s=s1se,mode="fraction")
46
47    results<-list(lasso,min.cv.s,coef.cv.s,s1se,coef1se,
         min.cp.s,b.lasso,cv,sds,lambda)
48    names(results)<-c("lasso","min.cv.s","coef.cv.s","s1se
         ","coef1se","min.cp.s","coef.cp.s","cv","sds","s")
49    return(results)
50  }
```

Κώδικας 6.1: Συνάρτηση lasso.lars για την εφαρμογή του αλγορίθμου LAR με την τροποποίηση LASSO και την εφαρμογή πολλών γύρων 5-fold Cross Validation. Επιστρέφει τη λίστα του αλγορίθμου LAR με τα βήματα που έγιναν, τους επιθυμητούς παράγοντες συρρίκνωσης s βάσει του μικρότερου μέσου 5-fold Cross Validation, του 1 standard error rule και του στατιστικού Mallows' C_p καθώς και τους συντελεστές του μοντέλου για αυτούς τους παράγοντες συρρίκνωσης. Επίσης επιστρέφει τους παράγοντες συρρίκνωσης s που χρησιμοποιήθηκαν για τους διαφορετικούς γύρους του 5-fold Cross Validation, τους μέσους όρους των 5-fold Cross Validation κάθε μοντέλου και τα σφάλματά τους.

```
> lar<-lasso.lars(X.diab,Y.diab,r=1000)
> names(lar)
 [1] "lasso"     "min.cv.s"  "coef.cv.s" "s1se"      "coef1se"   "min.cp.s"
 [7] "coef.cp.s" "cv"        "sds"       "s"
```

Εικόνα 6.3: Εκτελούμε τη συνάρτηση lasso.lars() για τα δεδομένα των ασθενών που πάσχουν από διαβήτη και λαμβάνουμε τα ονόματα των στοιχείων της λίστας που επιστρέφει.

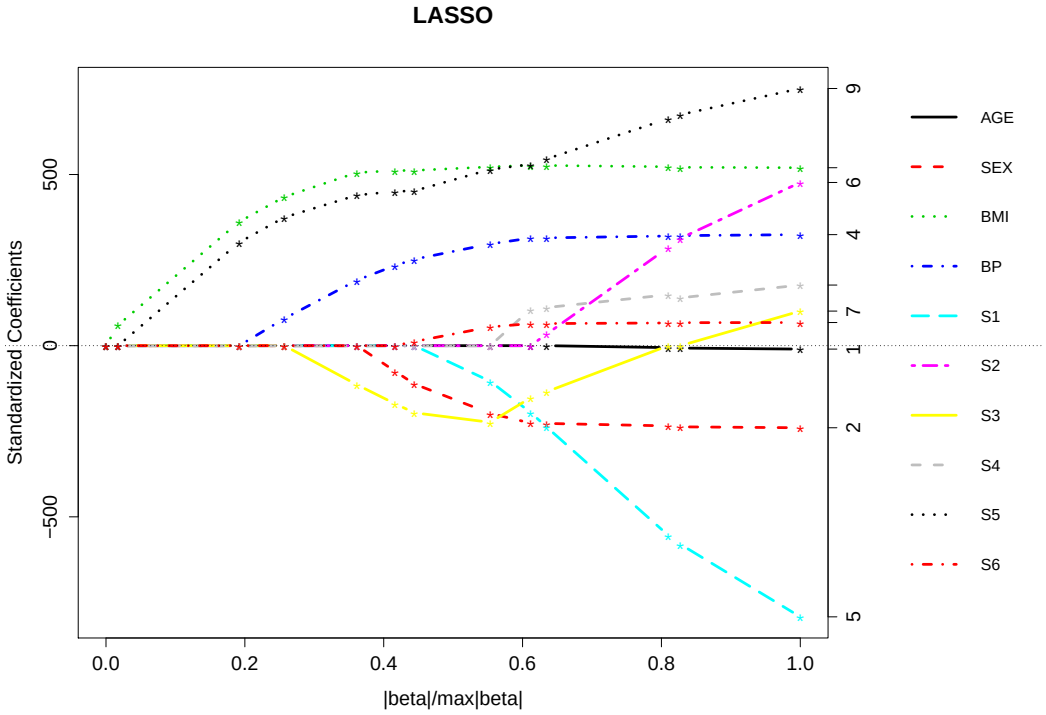
Στην Εικόνα 6.3 εκτελούμε τη συνάρτηση *lasso.lars*() του Κώδικα 6.1 για τα

δεδομένα των ασθενών που έχουν διαβήτη και $r = 1000$ γύρους 5-fold Cross Validation και βλέπουμε τα στοιχεία της λίστας που επιστρέφει η συνάρτηση. Στην Εικόνα 6.4 παρατηρούμε τα βήματα που έκανε ο αλγόριθμος Least Angle Regression και το ποιές συμμεταβλητές εισήγαγε ή αφαίρεσε σε κάθε βήμα, με τις πιο στατιστικά σημαντικές μεταβλητές να εισάγονται στα πρώτα βήματα. Στο 1ο βήμα εισήχθη η 3η συμμεταβλητή BMI, στο 2ο βήμα εισήχθη η 9η συμμεταβλητή $S5$, στο 3ο βήμα εισήχθη η 4η συμμεταβλητή BP, κ.ο.κ.

```
> lar$lasso

Call:
lars(x = as.matrix(X), y = Y)
R-squared: 0.518
Sequence of LASSO moves:
      BMI S5 BP S3 SEX S6 S1 S4 S2 AGE S3 S3
Var     3  9  4  7   2 10  5  8  6   1 -7  7
Step    1  2  3  4   5  6  7  8  9  10 11 12
```

Εικόνα 6.4: Τα βήματα που έκανε ο αλγόριθμος Least Angle Regression.

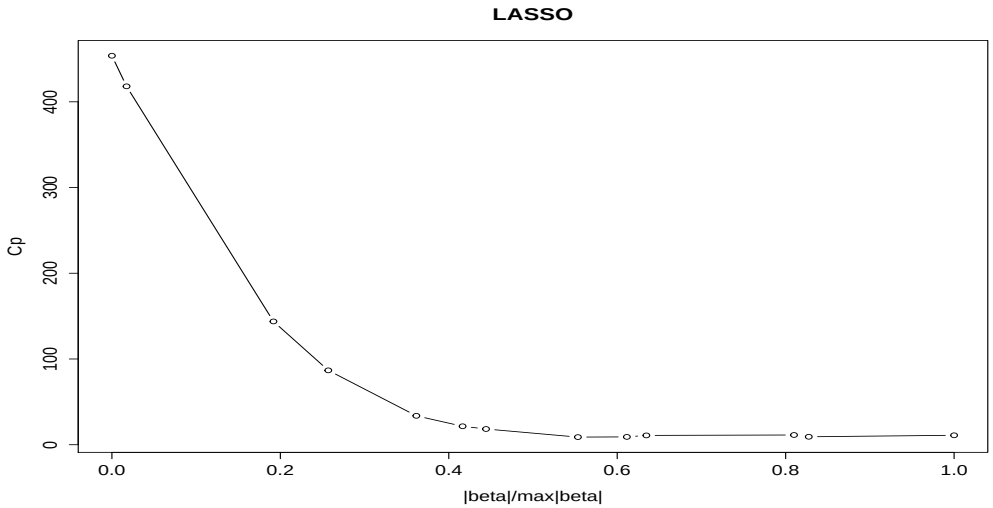


Εικόνα 6.5: Κοινό διάγραμμα των τιμών των συντελεστών των επεξηγηματικών μας μεταβλητών συναρτήσει του παράγοντα συρρίκνωσης.

Στην Εικόνα 6.5 παραθέτουμε το κοινό διάγραμμα των τιμών των συντελεστών των επεξηγηματικών μας μεταβλητών συναρτήσει του παράγοντα συρρίκνωσης s, το οποίο κατασκευάσαμε στον Κώδικα 6.2 στις γραμμές 2,3,4. Παρατηρούμε ότι όσο πιο μικρός είναι ο παράγοντας συρρίκνωσης τόσο πιο μικρές είναι οι τιμές των συντελεστών ενώ για $s = 1$ έχουμε τους εκτιμητές ελαχίστων τετραγώνων.

```
1 #plot the regularization paths
2 par(mar=c(5, 4, 4, 12)+0.1,xpd=TRUE)
3 plot(lar$lasso,breaks=F,xvar="norm",plottype="
     coefficients",col=1:ncol(X.diab),lty=1:ncol(X.diab),
     lwd=3)
4 legend("topright",inset=c(-0.3,0),seg.len=1.8,bty="n",
     legend=paste(names(X.diab),sep=''),ncol=1,lwd=3,col
     =1:ncol(X.diab),lty=1:ncol(X.diab),cex=0.8)
5
6 #plot Cp vs shrinkage factor
7 par(mar=c(5, 4, 4, 5)+0.1,xpd=F)
8 plot(lar$lasso,breaks=F,xvar='n',plottype='Cp')
```

Κώδικας 6.2: Κατασκευή των μονοπατιών κανονικοποίησης των συντελεστών της Εικόνας 6.5 και του διαγράμματος των τιμών του στατιστικού Mallows' Cp συναρτήσει του παράγοντα συρρίκνωσης, της Εικόνας 6.6.



Εικόνα 6.6: Διάγραμμα των τιμών του στατιστικού Mallows' Cp συναρτήσει του παράγοντα συρρίκνωσης.

Στην Εικόνα 6.6 παρατηρούμε το διάγραμμα των τιμών του στατιστικού Mallows' Cp συναρτήσει του παράγοντα συρρίκνωσης s που κατασκευάσαμε στον Κώδικα 6.2 στις γραμμές 7 και 8. Η τιμή του παράγοντα συρρίκνωσης για την οποία λαμβάνουμε τη μικρότερη τιμή του Mallows' Cp είναι $s = 0.5534$, όπως φαίνεται στην Εικόνα 6.7, όπου λαμβάνουμε και τις τιμές των συντελεστών για τον εν λόγω παράγοντα συρρίκνωσης. Παρατηρούμε ότι έχουν μηδενιστεί οι συντελεστές των επεξηγηματικών μεταβλητών AGE, $S2$ και $S4$, οι οποίες είχαν φανεί να μην είναι στατιστικά σημαντικές και κατά την μελέτη των άλλων μεθοδολογιών επιλογής μεταβλητών στο Κεφάλαιο 3. Οι συντελεστές των συμμεταβλητών $S1$, $S2$ και $S6$ δείχνουν ότι οι εν λόγω συμμεταβλητές έχουν μικρές επιδράσεις στην εξέλιξη της νόσου του διαβήτη.

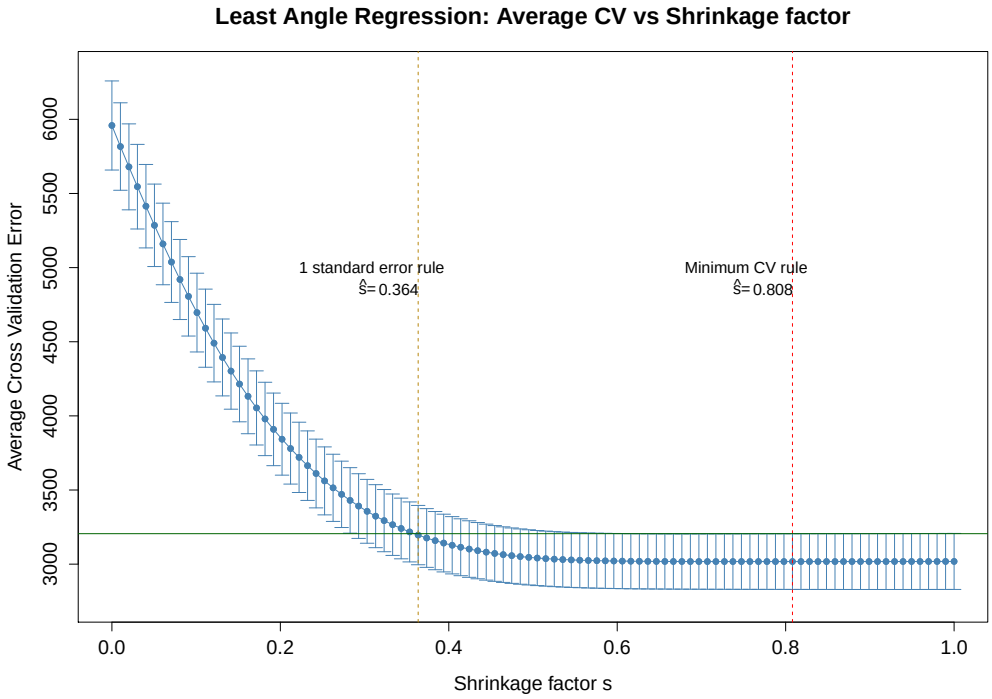
```
> lar$min.cp.s
[1] 0.5533458
> lar$coef.cp.s
        AGE         SEX         BMI          BP          S1          S2
  0.0000000 -18.8502075   5.6290895   1.0230567  -0.1430241   0.0000000
         S3          S4          S5          S6
 -0.8244074   0.0000000  46.9223824   0.2268591
```

Εικόνα 6.7: Ο παράγοντας συρρίκνωσης που ελαχιστοποιεί το στατιστικό Mallows' Cp και οι τιμές των συντελεστών των συμμεταβλητών που του αντιστοιχούν.

```
1  install.packages("ggplot2")
2  library(ggplot2)
3  install.packages("Hmisc")
4  library(Hmisc)
5  help(plot)
6  plot(lar$s,lar$cv,lwd=1,ylab="Average Cross Validation
       Error",xlab="Shrinkage parameter s",main="Least Angle
        Regression: Average CV vs Shrinkage factor",type="o"
       ,col="steelblue",pch=16,ylim=c(min(cvk)-min(sds),max(
       cvk)+max(sds)))
7  errbar(lar$s, lar$cv, lar$cv+lar$sds, lar$cv-lar$sds,
       add=T, pch=16,errbar.col="steelblue",col="steelblue",
       lwd=1)
8  abline(v=lar$s1se,lty=2,col="darkgoldenrod" )
9  abline(v=lar$min.cv.s,lty=2,col=2)
10 abline(h=lar$cv[lar$s==lar$min.cv.s]+lar$sds[lar$s==lar$
       min.cv.s],col="darkgreen")
11 text(lar$s1se-0.055,5000,"1 standard error rule",cex
       =0.8)
12 text(lar$s1se-0.065,4900,"^",cex=0.8)
13 text(lar$s1se-0.06,4850,"s=",cex=0.8)
```

```
14 text(lar$s1se-0.023,4850,paste(round(lar$s1se,3)),cex
     =0.8)
15 text(lar$min.cv.s-0.055,5000,"Minimum CV rule",cex=0.8)
16 text(lar$min.cv.s-0.065,4900,"^",cex=0.8)
17 text(lar$min.cv.s-0.06,4850,"s=",cex=0.8)
18 text(lar$min.cv.s-0.023,4850,paste(round(lar$min.cv.s,3)
     ),cex=0.8)
```

Κώδικας 6.3: Κατασκευή του διαγράμματος των μέσων όρων των 5-fold Cross Validation με τα σφάλματά τους συναρτήσει του παράγοντα συρρίκνωσης.



Εικόνα 6.8: Διάγραμμα των μέσων όρων των 5-fold Cross Validation με τα σφάλματά τους συναρτήσει του παράγοντα συρρίκνωσης.

Στην Εικόνα 6.8 παρατηρούμε το διάγραμμα των μέσων όρων των 5-fold Cross Validation με τα σφάλματά τους συναρτήσει του παράγοντα συρρίκνωσης, που κατασκευάσαμε στον Κώδικα 6.3. Στην Εικόνα 6.9 βλέπουμε την τιμή του παράγοντα συρρίκνωσης που δίνει τον μικρότερο μέσο όρο των 5-fold Cross Validation, $s = 0.808$ και τις τιμές των συντελεστών που του αντιστοιχούν. Παρατηρούμε ότι δεν έχει μηδενιστεί ο συντελεστής κάποιας επεξηγηματικής

μεταβλητής, αφού ο παράγοντας συρρίκνωσης που αντιστοιχεί στον ελάχιστο μέσο όρο των 5-fold Cross Validation είναι κοντά στο 1, παρέχοντάς μας ένα μοντέλο με όχι πολύ συρρικνωμένους συντελεστές. Απο την άλλη, με τον κανόνα 1 standard error, όπως βλέπουμε στην Εικόνα 6.9, λαμβάνουμε πολύ μικρότερο παράγοντα συρρίκνωσης, $s = 0.364$, ο οποίος μας παρέχει με ένα μοντέλο πιο φειδωλό και αντιστοίχως καλή προβλεπτική ικανότητα όπως ήδη έχουμε αναπτύξει στη θεωρία. Σε αυτό το μοντέλο, οι συντελεστές των συμμεταβλητών AGE, $S1$, $S2$, $S4$ και $S6$ έχουν μηδενιστεί ως μη στατιστικά σημαντικές. Εν ολίγοις λαμβάνουμε το μοντέλο που περιέχει τις ίδιες συμμεταβλητές με το μοντέλο που έχει το μικρότερο BIC, όπως υπολογίσαμε στην Παράγραφο 3.8 με την πλήρη εξερεύνηση του χώρου των πιθανών μοντέλων. Το όφελος που αποκομίζουμε όμως με τη μεθοδολογία $LASSO$ έγκειται στο ότι έχουν συρρικνωθεί ομοιόμορφα οι συντελεστές όλων των επεξηγηματικών μεταβλητών, αποτυπώνοντας καλύτερα τις επιδράσεις των στατιστικά σημαντικών συμμεταβλητών που συμπεριλαμβάνονται εν τέλει στο μοντέλο.

```
> lar$min.cv.s
[1] 0.8080808
> lar$coef.cv.s
         AGE          SEX          BMI           BP           S1
-0.020546819 -22.335593481   5.633660916   1.102674179  -0.758027012
          S2           S3           S4           S5           S6
 0.444759474  -0.005239178   5.479923731  60.325841564   0.274679270
> lar$s1se
[1] 0.3636364
> lar$coef1se
       AGE        SEX        BMI         BP         S1         S2         S3
 0.0000000 -0.2808606  5.4525154  0.6643128  0.0000000  0.0000000 -0.4281314
        S4         S5         S6
 0.0000000 40.1175189  0.0000000
```

Εικόνα 6.9: Οι τιμές του παράγοντα συρρίκνωσης βάσει του ελάχιστου 5-fold Cross Validation και του κανόνα 1 standard error και οι εκτιμήτριες LASSO των συντελεστών των συμμεταβλητών που τους αντιστοιχούν.

GLMNET package

Το πακέτο *glmnet* εφαρμόζει τον αλγόριθμο Coordinate Descent που αναλύσαμε στην Παράγραφο 6.4.2. Παρακάτω παραθέτουμε τις πιο σημαντικές εντολές του πακέτου *glmnet*:

- *glmnet*(): Υλοποιεί τον αλγόριθμο Coordinate Descent. Δέχεται ως ορίσματα τον πίνακα με τα δεδομένα των επεξηγηματικών μεταβλητών και τα δεδομένα της μεταβλητής απόκρισης. Επιστρέφει μια λίστα τα ονόματα των στοιχείων της οποίας μπορεί να δει κάποιος με την εντολή *names*().

- *cv.glmnet*(): Πραγματοποιεί K- fold Cross Validation για τα διαφορετικά μοντέλα που προκύπτουν από διαφορετικές τιμές της παραμέτρου ποινής λ, που ορίζει από μόνη της η εντολή. Η ακολουθία των τιμών της παραμέτρου ποινής μπορεί να δοθεί και από τον χρήστη με το όρισμα λ. Επιστρέφει μια λίστα στοιχείων που μεταξύ άλλων περιέχει τα CV_{K-fold}, τις τυπικές αποκλίσεις τους, τις τιμές της παραμέτρου ποινής λ για τις οποίες παρήχθησαν τα διαφορετικά μοντέλα, την παράμετρο ποινής που αντιστοιχεί στο μικρότερο CV_{K-fold} και την παράμετρο ποινής που αντιστοιχεί στο κανόνα 1 standard error.

- *coef*(): Υπολογίζει τις εκτιμήτριες $LASSO$ των συντελεστών των επεξηγηματικών μεταβλητών. Δέχεται ως ορίσματα είτε την υλοποίηση του αλγορίθμου, δηλαδή την εντολή *glmnet*() είτε την εντολή *cv.glmnet*() και $s =$ την τιμή της παραμέτρου ποινής λ στην οποία θέλουμε να υπολογίσουμε τις εκτιμήτριες $LASSO$.

Με χρήση λοιπόν των εν λόγω εντολών, στον Κώδικα 6.4 κατασκευάζουμε τη συνάρτηση *lasso.coord*() η οποία δέχεται ως ορίσματα τα δεδομένα των επεξηγηματικών μεταβλητών και της μεταβλητής απόκρισης καθώς και r τον αριθμό των γύρων του 5-fold Cross Validation που θέλουμε να εκτελεστεί για κάθε διαφορετικό μοντέλο. Πραγματοποιεί τους r γύρους 5-fold Cross Validation και εν συνεχεία υπολογίζει και επιστρέφει τις τιμές της παραμέτρου ποινής λ βάσει του μικρότερου μέσου όρου των 5-fold Cross Validation και του κανόνα 1 standard error. Υπολογίζει επίσης και επιστρέφει τις εκτιμήτριες $LASSO$ για αυτές τις παραμέτρους ποινής και τους παράγοντες συρρίκνωσης που τους αντιστοιχούν. Να διευκρινίσουμε σε αυτό το σημείο ότι επειδή η εντολή *cv.glmnet*() για συγκεκριμένη ακολουθία τιμών της παραμέτρου ποινής ορισμένες φορές «κολλάει» και υπολογίζει τα CV_{K-fold} για μια παράμετρο ποινής λιγότερη, δηλαδή ένα μοντέλο λιγότερο, προτού εκτελέσουμε τους r γύρους 5-fold Cross Validation εκτελούμε ένα βοηθητικό 5-fold Cross Validation, αποθηκεύοντας τα CV_{5-fold} και τα σφάλματά τους, ώστε όταν εμφανιζεται αυτό το πρόβλημα κατά

τη διάρκεια των r επαναλήψεων, να τοποθετούμε στα διανύσματα των CV_{5-fold} και των σφαλμάτων τους τις τιμές που λείπουν.

```
install.packages("glmnet")
library(glmnet)

lasso.coord<-function(X,Y,r){
  reg<-lm(Y~.,data=X)  #ols regression
  bols<-coef(reg)[-1]
  #ols coefficients for standardized data
  zbols<-coef(reg)[-1]*apply(X,2,sd)

  lasso<-glmnet(as.matrix(X),Y)   #lasso

  #5 fold cross validation to obtain lambdas,cvs and sds
      to help debug the multiple rounds of 5 fold CV
  lasso1<-cv.glmnet(as.matrix(X.diab),Y.diab,nfolds=5)
  l.glmnet<-lasso1$lambda
  cv.help<-lasso1$cvm
  cvsd.help<-lasso1$cvsd


  cvk<-matrix(rep(NA,r*length(l.glmnet)),ncol=length(l.
     glmnet),nrow=r)
  var<-matrix(rep(NA,r*length(l.glmnet)),ncol=length(l.
     glmnet),nrow=r)
  #r repetitions of 5 fold CV for various lambdas
  for(i in 1:r){
    rescv<-cv.glmnet(as.matrix(X),Y,nfolds=5,lambda=l.
       glmnet)
    a<-rescv$cvm
    b<-rescv$cvsd
    #debug
    if (length(a)<length(l.glmnet)){
      a<-c(a,cv.help[length(l.glmnet)])
      b<-c(b,cvsd.help[length(l.glmnet)])
    }
    cvk[i,]<-a
    var[i,]<-b^2
  }
  cv<-apply(cvk,2,mean)
  sds<-apply(var,2,mean)
  sds<-sqrt(sds)
```

```
37
38    #lambda that minimizes the average 5-fold cv
39    min.cv.l<-l.glmnet[which.min(cv)]
40    #coefficients for this lambda original data
41    blasso<-coef(lasso,s=min.cv.l)
42    #coefficients for this lambda standardized data
43    zblasso<-blasso[-1]*apply(X,2,sd)
44    #s that minimizes the average 5 fold cv
45    s.mincv<-sum(abs(zblasso))/sum(abs(zbols))
46
47    #1 standard error rule
48    range<-c(min(cv)-sds[l.glmnet==min.cv.l],min(cv)+sds[l
          .glmnet==min.cv.l])
49    j<-which(cv>=range[1]&cv<=range[2])
50    #lambda of 1 standard error rule
51    l.1se<-max(l.glmnet[j])
52    #coefficients for this lambda original data
53    blasso1<-coef(lasso,s=l.1se)
54    #coef for this lambda standardized data
55    zblasso1<-blasso1[-1]*apply(X,2,sd)
56    #s of 1 standard error rule
57    s.1se<-sum(abs(zblasso1))/sum(abs(zbols))
58
59    results<-list(lasso,cv,sds,l.glmnet,min.cv.l,s.mincv,
          blasso,l.1se,s.1se,blasso1)
60    names(results)<-c("lasso","cv","sds","lambdas","min.cv
          .l","s.mincv","coef.min.cv.l","l.1se","s.1se","coef
          .1se")
61    return(results)
62  }
```

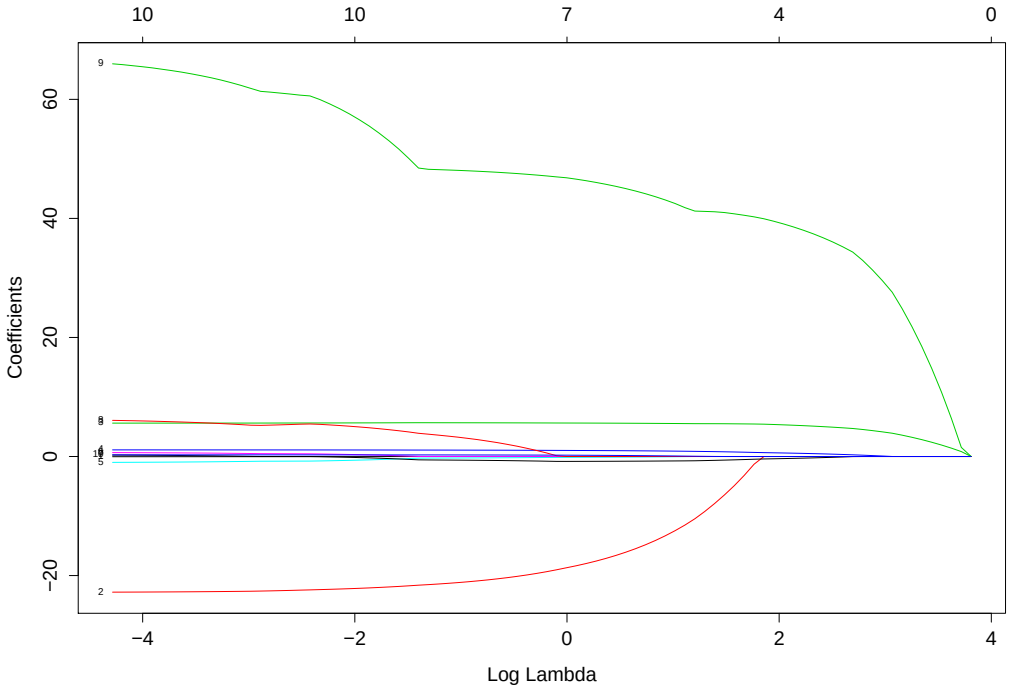
Κώδικας 6.4: Συνάρτηση lasso.coord() για την υλοποίηση του αλγορίθμου Coordinate Descent και την εφαρμογή πολλών γύρων 5-fold Cross Validation. Επιστρέφει τις επιθυμητές παραμέτρους ποινής λ και τους αντίστοιχους παράγοντες συρρίκνωσης s βάσει των 5-fold Cross Validation, 1 standard error rule καθώς και τις εκτιμήτριες LASSO για αυτές τις παραμέτρους ποινής.

Στην Εικόνα 6.10 βλέπουμε, αφότου εκτελέσουμε τη συνάρτηση *lasso.coord()* του Κώδικα 6.4 για τα δεδομένα των ασθενών που πάσχουν από διαβήτη, τα ονόματα των στοιχείων της λίστας που λαμβάνουμε. Θα τα χρησιμοποίησουμε στην πορεία για την ανάλυσή μας.

```
> coord<-lasso.coord(X.diab,Y.diab,r=1000)
> names(coord)
 [1] "lasso"         "cv"            "sds"           "lambdas"       "min.cv.l"

 [6] "s.mincv"       "coef.min.cv.l" "l.1se"         "s.1se"         "coef.1se"
```

Εικόνα 6.10: Εκτελούμε τη συνάρτηση coord.lasso() για τα δεδομένα των ασθενών που πάσχουν από διαβήτη και λαμβάνουμε τα ονόματα των στοιχείων της λίστας που επιστρέφει.



Εικόνα 6.11: Κοινό διάγραμμα των εκτιμητριών LASSO συναρτήσει του λογαρίθμου της παραμέτρου ποινής.

Στην Εικόνα 6.11 παρατηρούμε το κοινό διάγραμμα των εκτιμητριών *LASSO* των συντελεστών των συμμεταβλητών που χρησιμοποιούμε για την μελέτη της εξέλιξης της νόσου του διαβήτη, συναρτήσει του λογαρίθμου της παραμέτρου ποινής, το οποίο κατασκευάσαμε στη γραμμή 1 του Κώδικα 6.5. Παρατηρούμε ότι όσο αυξάνεται ο λογάριθμος της παραμέτρου ποινής, άρα και η ίδια η παράμετρος ποινής λ, τόσο περισσότερο συρρικνώνονται οι συντελεστές των επεξηγηματικών μας μεταβλητών.

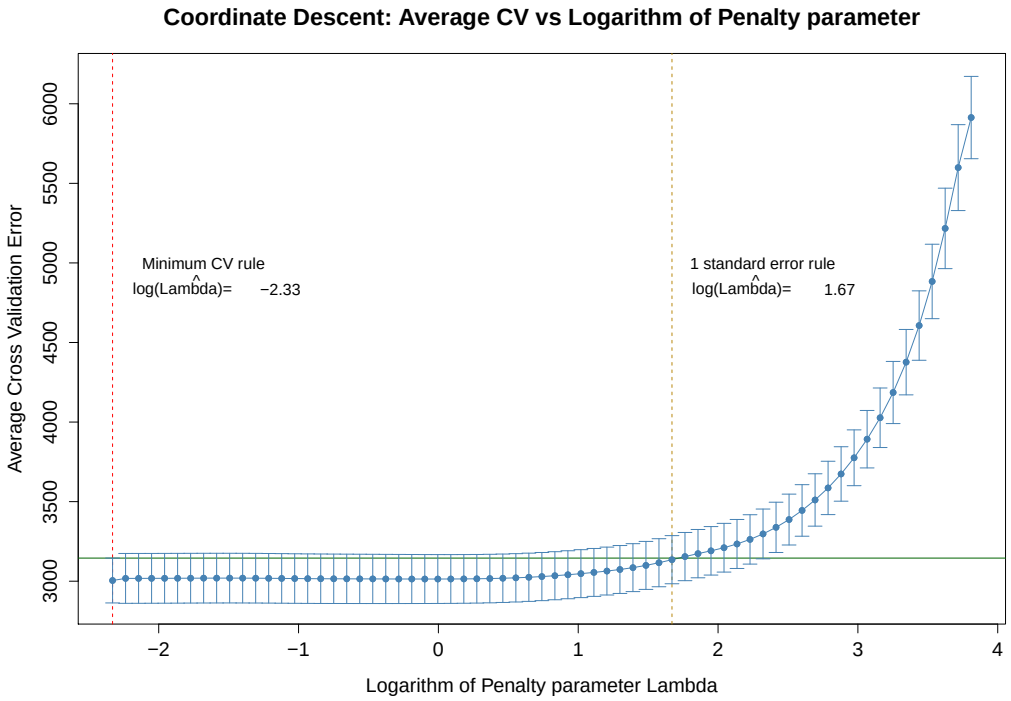
```
plot(coord$lasso,label=T,xvar="lambda")

install.packages("ggplot2")
library(ggplot2)
install.packages("Hmisc")
library(Hmisc)
plot(log(coord$lambdas),coord$cv,lwd=1,ylab="Average
   Cross Validation Error",xlab="Logarithm of Penalty
   parameter Lambda", main="Coordinate Descent: Average
   CV vs Logarithm of Penalty parameter",type="o",col="
   steelblue",pch=16,ylim=c(min(coord$cv)-min(coord$sds)
   ,max(coord$cv)+max(coord$sds)))
errbar(log(coord$lambdas), coord$cv, coord$cv+coord$sds,
    coord$cv-coord$sds, add=T, pch=16,errbar.col="
   steelblue",col="steelblue",lwd=1)
abline(v=log(coord$l.1se),lty=2,col="darkgoldenrod" )
abline(v=log(coord$min.cv.l),lty=2,col=2)
abline(h=coord$cv[coord$lambdas==coord$min.cv.l]+coord$
   sds[coord$lambdas==coord$min.cv.l],col="darkgreen")
text(log(coord$l.1se)+0.65,5000,"1 standard error rule",
   cex=0.8)
text(log(coord$l.1se)+0.6,4900,"^",cex=0.8)
text(log(coord$l.1se)+0.5,4830,"log(Lambda)=",cex=0.8)
text(log(coord$l.1se)+1.2,4830,paste(round(log(coord$l.1
   se),3)),cex=0.8)
text(log(coord$min.cv.l)+0.65,5000,"Minimum CV rule",cex
   =0.8)
text(log(coord$min.cv.l)+0.6,4900,"^",cex=0.8)
text(log(coord$min.cv.l)+0.5,4830,"log(Lambda)=",cex
   =0.8)
text(log(coord$min.cv.l)+1.2,4830,paste(round(log(coord$
   min.cv.l),3)),cex=0.8)
```

Κώδικας 6.5: Στη γραμμή 1 κατασκευάζουμε το κοινό διάγραμμα των εκτιμητριών LASSO των συντελεστών των συμμεταβλητών μας συναρτήσει του λογαρίθμου της παραμέτρου ποινής λ, που φαίνεται στην Εικόνα 6.11. Σε όλες τις υπόλοιπες γραμμές κατασκευάζουμε το διάγραμμα των μέσων όρων των 5-fold Cross Validation με τα σφάλματά τους συναρτήσει του λογαρίθμου της παραμέτρου ποινής λ, που φαίνεται στην Εικόνα 6.12.

Στην Εικόνα 6.12 βλέπουμε το διάγραμμα των μέσων όρων των 5-fold Cross Validation με τα σφάλματά τους, συναρτήσει του λογαρίθμου της παραμέτρου ποινής καθώς και το ποιές είναι οι τιμές του λογαρίθμου της παραμέτρου ποινής

που αντιστοιχούν στο μικρότερο μέσο όρο των 5-fold Cross Validation και στον κανόνα 1 standard error.



Εικόνα 6.12: Διάγραμμα των μέσων όρων των 5-fold Cross Validation με τα σφάλματά τους συναρτήσει του λογαρίθμου της παραμέτρου ποινής.

Στην Εικόνα 6.13 βλέπουμε την παράμετρο ποινής λ, τον παράγοντα συρρίκνωσης s και τις εκτιμήτριες $LASSO$ που αντιστοιχούν στο μοντέλο με το μικρότερο μέσο όρο των 5-fold Cross Validation που λάβαμε από τη συνάρτηση $lasso.coord()$. Παρατηρούμε ότι η παράμετρος ποινής είναι αρκετά χαμηλή, λ=0.097, και αντίστοιχα ο παράγοντας συρρίκνωσης υψηλός, $s = 0.8$, άρα οι συντελεστές δεν έχουν υποστεί μεγάλη συρρίκνωση, κάτι το οποίο γίνεται σαφές κοιτώντας τις εκτιμήτριες $LASSO$ για το μοντέλο που προκύπτει από την εν λόγω παράμετρο ποινής. Δεν έχει μηδενιστεί ο συντελεστής κάποιας επεξηγηματικής μεταβλητής, άρα έχουν θεωρηθεί όλες οι συμμεταβλητές στατιστικά σημαντικές, έστω και αν ορισμένες φαίνεται να έχουν μικρές επιδράσεις. Από την άλλη, στην Εικόνα 6.14, βλέπουμε την παράμετρο ποινής λ, τον παράγοντα συρρίκνωσης s και τις εκτιμήτριες $LASSO$ που αντιστοιχούν στο μοντέλο με τον κανόνα 1 standard error, που λάβαμε από τη συνάρτηση $lasso.coord()$. Εδώ η παράμετρος ποινής είναι αισθητά πιο μεγάλη, λ=5.31, και αντίστοιχα ο παράγοντας συρρίκνωσης πιο μικρός, σ=0.39, και όπως φαίνεται από τις τι-

μές των εκτιμητριών $LASSO$ για το εν λόγω μοντέλο, η συρρίκνωση που έχει επιτευχθεί στους συντελεστές είναι υψηλότερη. Μηδενίστηκαν οι τιμές των συντελεστών των συμμεταβλητών AGE, $S1$, $S2$, $S4$ και $S6$, με τις υπόλοιπες συμμεταβλητές να εμφανίζουν στατιστικά σημαντικές επιδράσεις στην εξέλιξη της νόσου του διαβήτη. Επιλέχθηκε δηλαδή το μοντέλο που περιέχει τις ίδιες συμμεταβλητές με το μοντέλο που έχει το μικρότερο BIC, όπως υπολογίσαμε στην Παράγραφο 3.8 με την πλήρη εξερεύνηση του χώρου των πιθανών μοντέλων. Παρατηρούμε τέλος ότι τα αποτελέσματα των συναρτήσεων που κατασκευάσαμε για την υλοποίηση του αλγορίθμου Least Angle Regression και Coordinate Descent, συμφωνούν, όχι σε ακρίβεια δεκαδικού ψηφίου μεν, αλλά είναι πολύ κοντινά ώστε τα συμπεράσματα που εξάγουμε για το ποιες συμμεταβλητές θα ήταν φρόνιμο να συμπεριλάβουμε στο μοντέλο μας, να είναι ίδια.

```
> coord$min.cv.l
[1] 0.09729434
> coord$s.mincv
[1] 0.8029463
> coord$coef.min.cv.l
11 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -322.93139592
AGE           -0.02128086
SEX          -22.36053022
BMI            5.63506605
BP             1.10324422
S1            -0.74351108
S2             0.43278207
S3            -0.02555918
S4             5.40275279
S5            59.97342633
S6             0.27514617
```

Εικόνα 6.13: Η παράμετρος ποινής, ο παράγοντας συρρίκνωσης και οι εκτιμήτριες LASSO που αντιστοιχούν στο μοντέλο με το μικρότερο μέσο όρο των 5-fold Cross Validation.

```
> coord$l.1se
[1] 5.314486
> coord$s.1se
[1] 0.3859181
> coord$coef.1se
11 x 1 sparse Matrix of class "dgCMatrix"
                      1
(Intercept) -222.0270585
AGE            .
SEX           -3.1789604
BMI            5.4768318
BP             0.7241772
S1             .
S2             .
S3            -0.5111684
S4             .
S5            40.5270386
S6             .
```

Εικόνα 6.14: Η παράμετρος ποινής, ο παράγοντας συρρίκνωσης και οι εκτιμήτριες LASSO που αντιστοιχούν στον κανόνα 1 standard error.

7

Σύνοψη

Συνοψίζοντας, θέλουμε να εξάγουμε διάφορα συμπεράσματα βάσει της θεωρίας που αναπτύξαμε και την οποία εφαρμόσαμε στη μελέτη του προβλήματος της επιλογής μεταβλητών για την πρόβλεψη της εξέλιξης της νόσου του διαβήτη. Είχαμε στη διάθεσή μας 442 παρατηρήσεις μιας ποσοτικής μέτρησης της εξέλιξης της νόσου η οποία αποτέλεσε τη μεταβλητή απόκρισης $Y.diab$ και 10 επεξηγηματικών μεταβλητών: AGE, SEX, BMI, BP, $S1$, $S2$, $S3$, $S4$, $S5$ και $S6$. Ξεκινήσαμε τη μελέτη μας παραθέτοντας στο Κεφάλαιο 1 το θεωρητικό υπόβαθρο του πολλαπλού γραμμικού μοντέλου παλινδρόμησης, το οποίο και προσαρμόσαμε για τα δεδομένα μας, αφότου ελέγξαμε τις προϋποθέσεις του. Συνεχίσαμε στο Κεφάλαιο 2 μελετώντας το πρόβλημα της πολυσυγγραμικότητας, της ύπαρξης δηλαδή στατιστικά σημαντικών σχέσεων μεταξύ των επεξηγηματικών μεταβλητών, καθώς και μεθόδους εντοπισμού της. Εφαρμόζοντας αυτές τις μεθόδους εντοπισμού στα δεδομένα μας, καταλήξαμε στο ότι το πλήρες μοντέλο, δηλαδή αυτό που περιλαμβάνει όλες τις συμμεταβλητές, πάσχει από πολυσυγγραμικότητα και ότι οι επεξηγηματικές μεταβλητές $S1$, $S2$, $S3$, και $S5$ εμπεριέχονται σε ορισμένες από τις υπάρχουσες στατιστικά σημαντικές γραμμικές σχέσεις. Προχωρήσαμε λοιπόν στο Κεφάλαιο 3 όπου αναφερθήκαμε στα κριτήρια πληροφορίας AIC και BIC και έπειτα αναπτύξαμε και εφαρμόσαμε ορισμένες διαδεδομένες μεθόδους επιλογής μεταβλητών. Συγκεκριμένα διενεργήσαμε πρώτα την πλήρη εξερεύνηση του χώρου των πιθανών μοντέλων θέλοντας να εντοπίσουμε τα μοντέλα που έχουν το μικρότερο AIC και το μικρότερο BIC, καταλήγοντας στα κάτωθι αποτελέσματα:

Μικρότερο AIC : $Y.diab \sim SEX + BMI + BP + S1 + S2 + S5$

Μικρότερο BIC : $Y.diab \sim SEX + BMI + BP + S3 + S5.$

Στη συνέχεια εφαρμόσαμε τις μεθόδους της επιλογής του καλύτερου υποσυνόλου των συμμεταβλητών, της διαδοχικής πρόσθεσης συμμεταβλητών, της διαδοχικής αφαίρεσης συμμεταβλητών και τη διαδικασία κατά βήματα, χρησιμοποι-

ώντας ως κριτήριο σύγκρισης μοντέλων για όλες αυτές, το κριτήριο πληροφορίας BIC. Και με τις 4 αυτές μεθόδους, «καλύτερο» αναδείχθηκε το μοντέλο:

$$Y.diab \sim SEX + BMI + BP + S1 + S2 + S5,$$

το οποίο εμπεριέχει ακριβώς τις ίδιες συμμεταβλητές με το μοντέλο που έχει το μικρότερο AIC, και το δεύτερο μικρότερο BIC, ανάμεσα σε όλα τα πιθανά μοντέλα. Ναι μεν οι 4 αυτές μέθοδοι δεν κατάφεραν να εντοπίσουν το μοντέλο με το μικρότερο BIC, αλλά έφτασαν κοντά, στο μοντέλο με το δεύτερο μικρότερο BIC. Με τη μέχρι στιγμής ανάλυσή μας λοιπόν, είχαμε ξεχωρίσει ως «καλύτερα» τα μοντέλα που είχαν το μικρότερο AIC και το μικρότερο BIC. Μάλιστα, και τα δύο αυτά μοντέλα, περιείχαν τις συμμεταβλητές SEX, BMI, BP και $S5$ μεταξύ άλλων, κάτι το οποίο μας προϊδέασε για τη στατιστική σημαντικότητά των συγκεκριμένων συμμεταβλητών. Προχωρήσαμε στο Κεφάλαιο 4, εισάγοντας τη μεθοδολογία του $Cross\ Validation$, που αποτελεί ένα μέτρο καλής προσαρμογής του εκάστοτε μοντέλου που εξετάζουμε, στα δεδομένα μας. Αναπτύξαμε τις παραλλαγές του Cross Validation: Leave One Out, Leave k out και K-fold Cross Validation τις οποίες και εφαρμόσαμε, για $k = 147$ και για $K = 10$, ώστε να συγκρίνουμε τα μοντέλα που βρήκαμε ότι έχουν το μικρότερο AIC και το μικρότερο BIC. Τα αποτελέσματα αυτά φαίνονται στον Πίνακα 7.1 και όπως παρατηρούμε και οι 3 παραλλαγές του Cross Validation ανέδειξαν «καλύτερο» το μοντέλο με το μικρότερο AIC. Παρόλα αυτά, οι τιμές των Cross Validation κάθε παραλλαγής είναι αρκετά κοντινές για τα δύο μοντέλα, το οποίο σημαίνει ότι κάποιος θα μπορούσε να διαλέξει ως «καλύτερο» το μοντέλο με το μικρότερο BIC αφού έχει αντίστοιχη προβλεπτική ικανότητα. Μάλιστα, το εν λόγω μοντέλο είναι και πιο φειδωλό από αυτό με το μικρότερο AIC, περιέχοντας μια λιγότερη συμμεταβλητή και άρα, όπως έχουμε αναφέρει, συχνά προτιμάμε πιο «οικονομικά» μοντέλα.

Cross Validation			
	$\text{CV}_{LeaveOneOut}$	$\text{CV}_{147-out}$	$\text{CV}_{10-fold}$
Μοντέλο με το μικρότερο AIC	2967.82	3057.28	2954.83
Μοντέλο με το μικρότερο BIC	2992.42	3116.27	3017.445

Πίνακας 7.1: Cross Validation για τα μοντέλα με το μικρότερο AIC και το μικρότερο BIC.

Στη συνέχεια, στο Κεφάλαιο 5, παρουσιάσαμε και αναπτύξαμε τη πρώτη μέθοδο συρρίκνωσης των συντελεστών των συμμεταβλητών ενός πολλαπλού γραμμικού

μοντέλου παλινδρόμησης, που εξετάσαμε, την παλινδρόμηση *Ridge*. Η μέθοδος *Ridge* ποινικοποιεί την ℓ_2-νόρμα των συντελεστών των επεξηγηματικών μεταβλητών. Έπειτα από εκτενή μελέτη των ιδιοτήτων της σε θεωρητικό επίπεδο, αποφανθήκαμε ότι η παλινδρόμηση *Ridge* εμπεριέχει άκρως θετικά χαρακτηριστικά, με κυριότερο αυτό του ελέγχου της διασποράς των εκτιμητριών *Ridge* των συντελεστών του μοντέλου μας. Όμως έχει και το αρνητικό στοιχείο του ότι δεν μηδενίζει κάποιον συντελεστή κατά τη συρρίκνωση των συντελεστών. Παρατηρήσαμε το $tradeoff$ ανάμεσα στην φθίνουσα διασπορά και την αύξουσα τετραγωνισμένη μεροληψία των εκτιμητριών *Ridge* των συντελεστών των συμμεταβλητών του μοντέλου, όσο αυξάνεται η παράμετρος ποινής λ δηλαδή όσο συρρικνώνονται περισσότερο οι συντελεστές. Εφαρμόζοντας στα δεδομένα μας τη παλινδρόμηση *Ridge* και εν συνεχεία τις διαφορετικές μεθόδους επιλογής της παραμέτρου ποινής λ που αναπτύξαμε, καταλήξαμε στις εκτιμήτριες *Ridge* για κάθε διαφορετική μέθοδο με την αντίστοιχη παράμετρο ποινής που βρήκαμε, οι οποίες φαίνονται στον Πίνακα 7.2.

Ridge						
	AIC λ=3.00	BIC λ=77.26	GCV λ=3.24	$CV_{10-fold}$ λ=4.16	HKB λ=5.46	LW λ=7.64
AGE	-0.03	0.02	-0.03	-0.03	-0.03	-0.02
SEX	-22.49	-17.98	-22.47	-22.38	-22.26	-22.09
BMI	5.62	5.01	5.62	5.61	5.61	5.59
BP	1.10	0.99	1.10	1.10	1.10	1.10
S1	-0.62	-0.08	-0.61	-0.54	-0.47	-0.39
S2	0.33	-0.12	0.31	0.25	0.19	0.11
S3	-0.17	-0.70	-0.20	-0.27	-0.35	-0.44
S4	5.09	4.39	5.03	4.85	4.66	4.47
S5	56.66	37.35	56.14	54.42	52.53	50.27
S6	0.29	0.39	0.29	0.29	0.30	0.30

Πίνακας 7.2: Εκτιμήτριες Ridge των συντελεστών των συμμεταβλητών για τις παραμέτρους ποινής στις οποίες καταλήξαμε με τις διαφορετικές μεθόδους επιλογής.

Παρατηρούμε κατά κύριο λόγο δύο πράγματα. Πρώτον για μεγαλύτερες τιμές της παραμέτρου ποινής λαμβάνουμε πιο συρρικνωμένους συντελεστές των συμμεταβλητών. Δεύτερον, δεν έχει μηδενιστεί ο συντελεστής κάποιας επεξηγηματικής μεταβλητής και εκεί έγκειται και το μειονέκτημα της παλινδρόμησης $Ridge$, αφού δεν μπορεί να δράσει, εξ′ αυτού του λόγου, άμεσα ως μέθοδος επιλογής μεταβλητών. Παρόλα αυτά, μπορεί να αποτελέσει προεόρτιο για κάποια άλλη μέθοδο επιλογής επεξηγηματικών μεταβλητών αφού μας παρέχει τις εκτιμήτριες των συντελεστών, για τις επιθυμητές παραμέτρους ποινής, δηλαδή για το επιθυμητό επίπεδο ελέγχου της διασποράς των συντελεστών των συμμεταβλητών. Όπως βλέπουμε στον Πίνακα 7.2 οι συμμεταβλητές AGE, $S1$, $S2$, $S3$ και $S6$ έχουν μικρές επιδράσεις, μικρότερες κατά απόλυτη τιμή από τη μονάδα, κάτι το οποίο μας οδήγησε στο συμπέρασμα ότι οι υπόλοιπες συμμεταβλητές οφείλουν σίγουρα να συμπεριληφθούν στο μοντέλο για την πρόβλεψη της εξέλιξης της νόσου του διαβήτη, ως στατιστικά σημαντικές με υψηλές επιδράσεις. Αυτό όμως δε σημαίνει ότι οι συμμεταβλητές με τις μικρές επιδράσεις δεν πρέπει να συμπεριληφθούν στο μοντέλο μας, αφού η παλινδρόμηση $Ridge$ δεν υποδεικνύει κάτι τέτοιο.

Με όλα αυτά κατά νού, προχωρήσαμε στο Κεφάλαιο 6 στη μελέτη της μεθοδολογίας $LASSO$, τη δεύτερη μέθοδο συρρίκνωσης των συντελεστών των συμμεταβλητών που εξετάσαμε. Αναπτύξαμε τις ιδιότητές της και τους πιο διαδεδομένους αλγορίθμους εύρεσης των εκτιμητριών $LASSO$. Η $LASSO$ ποινικοποιεί την ℓ_2-νόρμα των συντελεστών των επεξηγηματικών μεταβλητών, ελέγχοντας έτσι και αυτή, όπως και η $Ridge$, το πόσο μεγάλες τιμές μπορούν να λάβουν οι συντελεστές των συμμεταβλητών. Η $LASSO$ έχει το πλεονέκτημα έναντι της $Ridge$, ότι μηδενίζει τους συντελεστές των μη στατιστικά σημαντικών, για τον αντίστοιχο παράγοντα συρρίκνωσης, επεξηγηματικών μεταβλητών. Τοιουτοτρόπως δρα άμεσα ως μέθοδος επιλογής μεταβλητών. Εμείς εφαρμόσαμε τους δύο αλγορίθμους εύρεσης των λύσεων $LASSO$ για τα δεδομένα των ασθενών που πάσχουν από διαβήτη και επιλέξαμε τους επιθυμητούς παράγοντες συρρίκνωσης βάσει των μεθόδων που αναπτύξαμε στη θεωρία. Τα αποτελέσματα αυτά φαίνονται στον Πίνακα 7.3. Η πρώτη μας παρατήρηση είναι ότι τα αποτελέσματα των δύο αλγορίθμων για τις διαφορετικές μεθόδους επιλογής του παράγοντα συρρίκνωσης, παράγουν σχεδόν ίδια μοντέλα, άρα μπορούμε να εξάγουμε συμπεράσματα χωρίς να αναλύσουμε τα επιμέρους αποτελέσματα των αλγορίθμων. Στο μοντέλο με το μικρότερο $\overline{CV_{5-fold}}$, στο οποίο αντιστοιχεί παράγοντας συρρίκνωσης $s \simeq 0.8$, δηλαδή όχι μεγάλη συρρίκνωση των συντελεστών, δεν μηδενίστηκε κάποιος συντελεστής ενώ στο μοντέλο που αντιστοιχεί στον κανόνα του 1 τυπικού σφάλματος και έχει παράγοντα συρρίκνωσης $s \simeq 0.37$, με αντίστοιχη προβλεπτική ικανότητα με το μοντέλο που έχει το μικρότερο $\overline{CV_{5-fold}}$, η $LASSO$ μηδένισε τους συντελεστές των συμμεταβλη-

τών AGE, $S1$, $S2$, $S4$ και $S6$. Στο μοντέλο που αντιστοιχεί στο μικρότερο $Mallows'\ C_p$ η $LASSO$ μηδένισε τους συντελεστές των συμμεταβλητών AGE, $S2$ και $S4$.

LASSO					
	Μικρότερο C_p	Μικρότερο $\overline{CV_{5-fold}}$		Κανόνας 1 τυπικού σφάλματος	
	LAR s=0.5334	LAR s=0.8080	GLMNET s=0.8029	LAR s=0.3636	GLMNET s=0.3859
AGE	0	-0.02	-0.02	0	0
SEX	-18.85	-22.34	-22.36	-0.28	-3.18
BMI	5.63	5.63	5.64	5.45	5.48
BP	1.02	1.10	1.10	0.66	0.72
S1	-0.14	-0.75	-0.74	0	0
S2	0	0.45	0.43	0	0
S3	-0.82	-0.01	-0.02	-0.43	-0.51
S4	0	5.48	5.40	0	0
S5	46.92	60.33	59.97	40.12	40.53
S6	0.23	0.28	0.28	0	0

Πίνακας 7.3: Εκτιμήτριες LASSO των συντελεστών των συμμεταβλητών για τους παράγοντες συρρίκνωσης στους οποίους καταλήξαμε με τις διαφορετικές μεθόδους επιλογής.

Καταλήξαμε δηλαδή σε 3 διαφορετικά μοντέλα κατά την εφαρμογή της $LASSO$, με εξίσου καλό σφάλμα πρόβλεψης της εξέλιξης της νόσου του διαβήτη. Παρόλα αυτά, αναζητούμε φειδωλά μοντέλα και για αυτόν τον λόγο διαλέγουμε το μοντέλο που αντιστοιχεί στον κανόνα του ενός τυπικού σφάλματος:

$$Y.diab \sim SEX + BMI + BP + S3 + S5.$$

Το μοντέλο με αυτές τις συμμεταβλητές θυμίζουμε ότι είχε το μικρότερο BIC στην πλήρη εξερεύνηση του χώρου των πιθανών μοντέλων που εκτελέσαμε στο

Κεφάλαιο 3. Επίσης περιέχει όλες τις επεξηγηματικές μεταβλητές που κρίναμε, καθόλη τη διάρκεια αυτής της εργασίας, απαραίτητο να συμπεριλάβουμε στο τελικό μας μοντέλο. Αποτελεί την τελική μας πρόταση για την πρόβλεψη της εξέλιξης της ασθένειας του διαβήτη σε μελλοντικούς ασθενείς και θα συνιστούσαμε ως τελικό βήμα, για την εξασφάλιση της ακρίβειας του μοντέλου, τη συμμετοχή κάποιου γνώστη της ιατρικής σκοπιάς του προβλήματος, στην εξαγωγή των τελικών συμπερασμάτων.

Βιβλιογραφία

Φουσκάκης, Δ., (2013), *Ανάλυση Δεδομένων με Χρήση της R*, Εκδόσεις Τσότρας, Αθήνα.

Belsley, D., Kuh, E., Welsch, R., (2004), *Regression Diagnostics: Identifying Influential Data and Sources of Collinearity*, Wiley.

Efron, B., Hastie, T., Johnstone, I. and Tibshirani, R., (2004), Least angle regression (with discussion), *Annals of Statistics* **32**(2): 407–499.

Friedman, J., Hastie, T., Hoefling, H. and Tibshirani, R., (2007), Pathwise coordinate optimization, *Annals of Applied Statistics* **1**(2), 302–332.

James, G., Witten, D., Hastie, T., Tibshirani, R.,(2013), *An Introduction to Statistical Learning with Applications in R*, Springer, New York.

Hastie, T., Tibshirani, R. and Friedman, J. (2009), *The Elements of Statistical Learning: Data Mining, Inference and Prediction*, second edn, Springer, New York

Hastie, T., Tibshirani T., Wainwright, M. (2015), *Statistical Learning with Sparsity: The Lasso and Generalizations*, Chapman and Hall CRC.

Hoerl, A. E., and Kennard, R. W., (1970), Ridge regression: biased estimation for nonorthogonal problems, *Technometrics*, **12**, 55-67.

Ryan, T. P., (2008), *Modern Regression Methods,* 2nd Edition, Wiley.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistic Society, Series B* **58**, 267–288.