ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΔΠΜΣ ΜΑΘΗΜΑΤΙΚΗ ΠΡΟΤΥΠΟΠΟΙΗΣΗ ΣΕ ΣΥΓΧΡΟΝΕΣ ΤΕΧΝΟΛΟΓΙΕΣ ΚΑΙ ΤΗ ΧΡΗΜΑΤΟΟΙΚΟΝΟΜΙΚΗ

ΑΠΟΔΟΤΙΚΟΙ ΑΛΓΟΡΙΘΜΟΙ ΓΙΑ ΤΟΝ ΥΠΟΛΟΓΙΣΜΟ ΤΩΝ ΒΕΛΤΙΣΤΩΝ ΜΟΝΤΕΛΩΝ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΠΡΟΒΛΗΜΑΤΑ ΜΕΓΑΛΩΝ ΔΙΑΣΤΑΣΕΩΝ

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΔΗΜΟΣΘΕΝΟΥΣ ΜΑΡΙΟΣ

ΑΜ: 09320010

Επιβλέπουσα: Χρυσηΐς Καρώνη

Καθηγήτρια Ε.Μ.Π

Καρώνη Χρυσηΐς	Παπανικολάου Βασίλης	Στεφανέας Πέτρος
Καθηγήτρια Ε.Μ.Π.	Καθηγητής Ε.Μ.Π.	Αναπλ. Καθηγητής Ε.Μ.Π.
……	……	……..

Αθήνα, Οκτώβριος 2022

NATIONAL TECHNICAL UNIVERSITY OF ATHENS

MATHEMATICAL MODELING IN MODERN TECHNOLOGIES AND FINANCIAL ENGINEERING

EFFICIENT ALGORITHMS FOR COMPUTING THE BEST SUBSET REGRESSION MODELS FOR HIGH DIMENSIONAL PROBLEMS

MASTER THESIS

DEMOSTHENOUS MARIOS

Supervisor: Chrysseis Caroni

Professor NTUA

Caroni Chrysseis	Papanicolaou Vassilis	Stefaneas Petros
Professor NTUA	Professor NTUA	Associate Professor NTUA
……	……	…….

Athens, October 2022

ΔΗΜΟΣΘΕΝΟΥΣ ΜΑΡΙΟΣ

Περίληψη

Η Στατιστική Μοντελοποίηση αποτελεί ένα Μαθηματικό εργαλείο μοντελοποίησης δεδομένων με σκοπό την πρόβλεψη ή τη μελέτη ενός φαινομένου. Με την εξέλιξη της τεχνολογίας είμαστε ικανοί να συλλέξουμε ολοένα και περισσότερα και πιο ακριβή δεδομένα. Αυτό συνεπάγεται στην αύξηση της διαστατικότητας στο πρόβλημα της μοντελοποίησης και την αύξηση του υπολογιστικού κόστους. Επομένως, κρίνεται όλο και πιο σημαντική η δημιουργία αποδοτικών μεθόδων υψηλής ακρίβειας για την προσαρμογή των βέλτιστων μοντέλων. Η παρούσα εργασία πραγματεύεται την προσαρμογή γραμμικών μοντέλων σε κάποιο σύνολο δεδομένων και μελετά ακριβείς και αποδοτικές μεθόδους για την προσαρμογή των μοντέλων καθώς και για την εύρεση βέλτιστων μοντέλων.

Στο ΚΕΦΑΛΑΙΟ 1 παρουσιάζουμε το πρόβλημα της γραμμικής παλινδρόμησης καθώς και βασικές μεθόδους προσαρμογής ενός μοντέλου. Στο ΚΕΦΑΛΑΙΟ 2 παρουσιάζουμε διάφορες μεθόδους επίλυσης του προβλήματος ελαχίστων τετραγώνων. Πιο συγκεκριμένα μελετούμε μεθόδους βασισμένες στην απαλοιφή Gauss και άλλες μεθόδους όπως την παραγοντοποίηση Cholesky, QR και SVD. Στο ΚΕΦΑΛΑΙΟ 3 μελετούμε το πρόβλημα της εύρεσης του βέλτιστου μοντέλου. Αρχικά ορίζουμε μερικά μέτρα σύγκρισης μοντέλων όπως το άθροισμα τετραγώνων των υπολοίπων, RSS, ο δείκτης C_p, τα AIC και BIC και άλλα. Στη συνέχεια παρουσιάζουμε μερικούς αλγόριθμους για την εύρεση του βέλτιστου μοντέλου. Συγκεκριμένα, παρουσιάζουμε τους αλγόριθμους Leaps-and-Bounds, Branch-and-Bound και lmSelect. Στο τέλος του κεφαλαίου μελετούμε την περίπτωση όπου έχουμε λιγότερες παρατηρήσεις στο σύνολο δεδομένων μας σε σύγκριση με το πλήθος των παραμέτρων του μοντέλου και προτείνουμε μία προσέγγιση σε αυτό το πρόβλημα για την εύρεση του βέλτιστου δυνατού μοντέλου.

Λέξεις – κλειδιά: γραμμική παλινδρόμηση, RSS, μέθοδος ελαχίστων τετραγώνων, επιλογή μοντέλου, SWEEP, απαλοιφή Gauss, παραγοντοποίηση QR, leaps-and-bounds, branch-and-bound, lmSelect.

Abstract

Statistical Modeling is a Mathematical tool which can be used to study a phenomenon or make predictions through statistical models. With the ongoing technological advancements, we are increasingly more able to produce a vast amount of data in more accurate ways. This results in an increase in dimensionality, for the statistical modeling problem, along with increased computational cost. Thus, it is important to develop accurate and efficient methods for extracting the best models. In this Thesis, we study different methods for fitting regression models and for selecting the best models.

In Chapter 1, we present the linear regression problem along with basic methods for fitting a regression model. In Chapter 2, we present different methods for solving the least-squares problem. More specifically, we study methods based on Gauss elimination and other methods such as the Cholesky, QR and SVD decompositions. In Chapter 3, we study the problem of finding the best model(s). Initially, we define some evaluation metrics for the regression models, such as the residual sum of squares, RSS, $C_p - Mallows$, AIC and BIC etc. Furthermore, we present different algorithms for finding the best model(s), such as the Leaps-and-Bounds, Branch-and-Bound and lmSelect algorithms. At the end of the Chapter, we study the case where we have fewer observations in our data than variables and we suggest an approach to find the best possible model in this case.

Key – words: linear regression, RSS, least squares method, model selection, SWEEP, Gauss elimination, QR factorization, leaps-and-bounds, branch-and-bound, lmSelect.

Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά την κυρία Χρυσηίς Καρώνη, Καθηγήτρια του Ε.Μ.Π. και επιβλέπουσα της Μεταπτυχιακής μου εργασίας για την ανάθεση της παρούσας εργασίας και την υποστήριξη που μου προσέφερε τόσο για την εκπόνησή της εργασίας όσο και κατά τη διάρκεια των σπουδών μου.

Θα ήθελα επίσης να ευχαριστήσω ιδιαιτέρως τον Καθηγητή Ερρίκο Κοντογιώργη και τον Καθηγητή Cristian Gatu για την καθοδήγηση και υποστήριξη που μου προσέφεραν, καθώς και για το χρόνο που αφιέρωσαν.

Τέλος, ευχαριστώ την οικογένεια και τους φίλους μου για τη στήριξη που μου δείχνουν σε ό,τι κάνω.

Περιεχόμενα

Λίστα Πινάκων

Λίστα Σχημάτων

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

1.1 Γραμμικό Μοντέλο Παλινδρόμησης

Ένα σύνηθες πρόβλημα σε πολλές επιστήμες είναι η μοντελοποίηση της σχέσης εξάρτησης μεταξύ μεταβλητών με στόχο την πρόβλεψη κάποιας κατάστασης και την περιγραφή του τρόπου εξάρτησης μεταξύ μεταβλητών. Μία μη στοχαστική σχέση εξάρτησης θα έχει τη μορφή

$$y = f(x_1, x_2, \dots, x_k)$$

όπου μία μεταβλητή y εξαρτάται από τις μεταβλητές $x_1, x_2, \dots, x_k$ μέσω μίας συνάρτησης f. Η μεταβλητή y ονομάζεται **εξαρτημένη μεταβλητή** (dependent variable) ή **μεταβλητή απόκρισης** (response variable). Οι μεταβλητές $x_1, x_2, \dots, x_k$ ονομάζονται **ανεξάρτητες μεταβλητές** (independent variables), **επεξηγηματικές μεταβλητές** (explanatory variables), **προβλέπουσες** (predictors) ή **συμμεταβλητές** (covariates).

Για παράδειγμα, στη Φυσική έχουμε τη σχέση

$$u = f(t) = u_0 + \alpha t$$

όπου η ταχύτητα u περιγράφεται ως μία συνάρτηση του χρόνου t με παραμέτρους την αρχική ταχύτητα u_0 και την επιτάχυνση α. Αυτή η σχέση είναι μη στοχαστική αφού, γνωρίζοντας τις ακριβείς τιμές των u_0 και α, μπορούμε να υπολογίσουμε χωρίς σφάλμα την ακριβή τιμή της ταχύτητας u σε δεδομένο χρόνο t. Φυσικά, στην πράξη η ακρίβεια του υπολογισμού της ποσότητας $u(t)$ εξαρτάται από την ακρίβεια των μετρήσεών μας για τις ποσότητες u_0 και α.

Στην περίπτωση ενός στατιστικού μοντέλου η σχέση εξάρτησης είναι εν μέρει στοχαστική (Καρώνη & Οικονόμου, 2017). Για να ενσωματώσουμε στο μοντέλο μας την αβεβαιότητα των μετρήσεών μας για τη μεταβλητή y καθώς και τον θόρυβο στα δεδομένα μας, προσθέτουμε τη μεταβλητή ε ως εξής

$$y = f(x_1, x_2, \dots, x_k) + \varepsilon \qquad (1.1)$$

όπου η τυχαία μεταβλητή ε ονομάζεται **τυχαίο σφάλμα** και αντιπροσωπεύει τη διαφορά μεταξύ της παρατήρησής μας για τη μεταβλητή y και της αληθινής τιμής της y, υπολογισμένη μέσω της σχέσης $y = f(x_1, x_2, \dots, x_k)$ εφόσον γνωρίζουμε τις ακριβείς, θεωρητικές τιμές των παραμέτρων του μοντέλου μας. Αν θεωρήσουμε ότι η αναμενόμενη τιμή της μεταβλητής ε είναι μηδέν, δηλαδή $E(\varepsilon) = 0$, τότε καταλήγουμε στο $E(y) = f(x_1, x_2, \dots, x_k)$ που αντιπροσωπεύει το συστηματικό ή μη στοχαστικό μέρος του μοντέλου.

Για σκοπούς στατιστικής συμπερασματολογίας θα κάνουμε μερικές υποθέσεις για τα τυχαία σφάλματα ε. Θεωρούμε ότι τα τυχαία σφάλματα $\varepsilon_i, \quad i = 1,2,\dots,n$, είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές με

- $E(\varepsilon_i) = 0, \forall\, i$
- $V(\varepsilon_i) = \sigma^2, \ \forall\, i$ (ομοσκεδαστικότητα)
- $cov(\varepsilon_i, \varepsilon_j) = 0\ , \forall\, i \neq j$ (τα ε_i είναι ασυσχέτιστα μεταξύ τους)

Οι παραπάνω υποθέσεις χρησιμοποιούνται στην εύρεση της αναμενόμενης τιμής και διασποράς των εκτιμητών ελαχίστων τετραγώνων τους οποίους θα περιγράψουμε στη συνέχεια. Επιπρόσθετα, για την κατασκευή ελέγχων υποθέσεων και διαστημάτων εμπιστοσύνης, θα χρησιμοποιήσουμε την υπόθεση της κανονικότητας των τυχαίων σφαλμάτων, ότι δηλαδή

$$\varepsilon_i \ \sim N(0, \sigma^2)$$

ή διαφορετικά,

$$\boldsymbol{\varepsilon} \sim N_n\big(\mathbf{0}, \boldsymbol{\sigma^2} I\big)$$

όπου $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$, $E(\boldsymbol{\varepsilon}) = \mathbf{0}$ και

$$V(\boldsymbol{\varepsilon}) = E\left[\big(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\big)'\big(\boldsymbol{\varepsilon} - E(\boldsymbol{\varepsilon})\big)\right] = E[\boldsymbol{\varepsilon}'\boldsymbol{\varepsilon}] = \sigma^2 \boldsymbol{I}$$

Για ένα μοντέλο γραμμικής παλινδρόμησης με k επεξηγηματικές μεταβλητές η σχέση (1.1) γράφεται

$$y|\boldsymbol{x} = y_{\boldsymbol{x}} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \varepsilon_{\boldsymbol{x}} \tag{1.2}$$

όπου $\boldsymbol{x} = (x_1, x_2, \dots, x_k)'$ αντιπροσωπεύει ένα διάνυσμα παρατηρήσεων με μία τιμή για κάθε μεταβλητή x_i. Οι $p = k + 1$ ποσότητες β_j, $j = 0,1,2,\dots,k$ αποτελούν τις **παραμέτρους** του μοντέλου ή αλλιώς τους **συντελεστές παλινδρόμησης**. Στην περίπτωση όπου έχουμε μόνο μία επεξηγηματική μεταβλητή ($k = 1$), έχουμε το **απλό γραμμικό μοντέλο**. Διαφορετικά, για $k > 1$ έχουμε το **πολλαπλό γραμμικό μοντέλ**ο. Να σημειωθεί ότι αναφερόμενοι στη γραμμικότητα του μοντέλου μιλάμε για τη γραμμικότητα ως προς τις παραμέτρους του μοντέλου, β_j και όχι ως προς τις μεταβλητές x_j. Για παράδειγμα, το μοντέλο

$$\begin{aligned} \mathrm{y} &= \beta_0 + \beta_1 x_1^2 + \beta_2 \ln(x_2) + \varepsilon \\ &= \beta_0 + \beta_1 z_1 + \beta_2 z_2 + \varepsilon \end{aligned}$$

είναι γραμμικό, ενώ το μοντέλο

$$y = \beta_0 + \exp(\beta_1 x_1 + \beta_2 x_2) + \varepsilon$$

δεν είναι γραμμικό ως προς τους συντελεστές β_j .

Στόχος μας είναι η προσαρμογή μίας ευθείας ή ενός υπερεπιπέδου, στην περίπτωση του απλού ή του πολλαπλού γραμμικού μοντέλου, αντίστοιχα, που να εξηγεί όσο το δυνατόν καλύτερα τη συμπεριφορά των δεδομένων μας. Ένα τέτοιο υπερεπίπεδο θα έχει τη μορφή

$$E(y|\boldsymbol{x}) = E(y_{\boldsymbol{x}}) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k = \mu_{\boldsymbol{x}} \tag{1.3}$$

Η προσαρμογή του βέλτιστου υπερεπιπέδου γίνεται λαμβάνοντας υπόψη n ανεξάρτητες παρατηρήσεις $(\boldsymbol{x_i}, y_i)$, $i = 1, 2, \dots, n$, όπου υποθέτουμε ότι τα x_i δεν υπόκεινται σε σφάλματα μέτρησης. Το διάνυσμα $\boldsymbol{x_i} = (x_{i0}, x_{i1}, x_{i2}, \dots, x_{ik})'$, όπου συνήθως $x_{i0} = 1$, αντιπροσωπεύει την i-οστή μας παρατήρηση. Τότε, για τα y_i και $i = 1, 2, \dots, n$ θα ισχύει

$$\begin{aligned} y_{x_i} = y_i &= \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_k x_{ik} + \varepsilon_{\boldsymbol{x_i}} \\ &= E(y_i) + \varepsilon_i \end{aligned}$$

Δηλαδή, έχουμε ένα σύστημα n εξισώσεων με p αγνώστους, τα β_j, $j = 0, 1, 2, \dots, k$. Σε μορφή πινάκων το σύστημα γράφεται

$$\boldsymbol{y} = \boldsymbol{X\beta} + \boldsymbol{\varepsilon} \tag{1.4}$$

όπου $\boldsymbol{y} = (y_1, y_2, \dots, y_n)'$ είναι το $n \times 1$ διάνυσμα τιμών της μεταβλητής απόκρισης y, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$ είναι το $p \times 1$ διάνυσμα των άγνωστων παραμέτρων του μοντέλου, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_n)'$ είναι το $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων και

$$\boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \dots & x_{1k} \\ 1 & x_{21} & x_{22} & \dots & x_{2k} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \dots & x_{nk} \end{pmatrix}$$

είναι ένας $n \times p$ πίνακας ο οποίος ονομάζεται **πίνακας σχεδιασμού** όπου x_{ij} είναι η i-οστή παρατήρηση της j-οστής μεταβλητής. Πολλές φορές το $\boldsymbol{X\beta}$ ονομάζεται συστηματικό ή μη-στοχαστικό μέρος του μοντέλου.

Υπό τις υποθέσεις που κάναμε για τα τυχαία σφάλματα ε, παρατηρούμε ότι η κατανομή της $\boldsymbol{y}$ είναι

$$\boldsymbol{y} \sim N_n(\boldsymbol{X\beta}, \sigma^2 \boldsymbol{I_n})$$

(Καρώνη & Οικονόμου, 2017) όπου

$$E(\boldsymbol{y}) = \boldsymbol{X\beta}$$

$$V(\boldsymbol{y}) = V(\boldsymbol{X\beta} + \boldsymbol{\varepsilon}) = V(\boldsymbol{\varepsilon}) = \sigma^2 \boldsymbol{I_n}$$

Επομένως για την παρατήρηση y_i θα ισχύει

$$y_i \sim N(\boldsymbol{x_i'\beta}, \sigma^2)$$

Από την προσαρμογή του μοντέλου προκύπτει το προσαρμοσμένο υπερεπίπεδο

$$\widehat{y_i} = \widehat{\beta_0} + \widehat{\beta_1} x_{i1} + \widehat{\beta_2} x_{i2} + \cdots + \widehat{\beta_k} x_{ik} \tag{1.5}$$

όπου $\widehat{y_i}$ είναι η εκτίμηση της μεταβλητής y με δεδομένη την παρατήρηση $\boldsymbol{x} = \boldsymbol{x_i} = (1, x_{i1}, x_{i2}, \dots, x_{ik})'$ και $\widehat{\beta_j}$ $(j = 0, 1, \dots, k)$ είναι η εκτίμηση της παραμέτρου β_j του μοντέλου . Το υπερεπίπεδο της σχέσης (1.5) διέρχεται από τα σημεία $(\boldsymbol{x_i}, \widehat{y_i})$, $i = 1, 2, \dots, n$.

Η εκτίμηση $\widehat{y_i}$ ενδέχεται να διαφέρει από την πραγματική τιμή y_i. Τη διαφορά αυτή την ονομάζουμε **υπόλοιπο** το οποίο γράφεται ως

$$e_i = y_i - \widehat{y_i} \tag{1.6}$$

Τα υπόλοιπα είναι μία χρήσιμη ποσότητα που μας βοηθούν στο να καταλήξουμε στο συμπέρασμα αν το μοντέλο επαρκεί για να περιγράψει τη σχέση μεταξύ των y και $\boldsymbol{x}$. Τα e_i μπορούν επίσης να θεωρηθούν ως οι εκτιμήσεις των άγνωστων, μη παρατηρήσιμων τυχαίων σφαλμάτων ε_i.

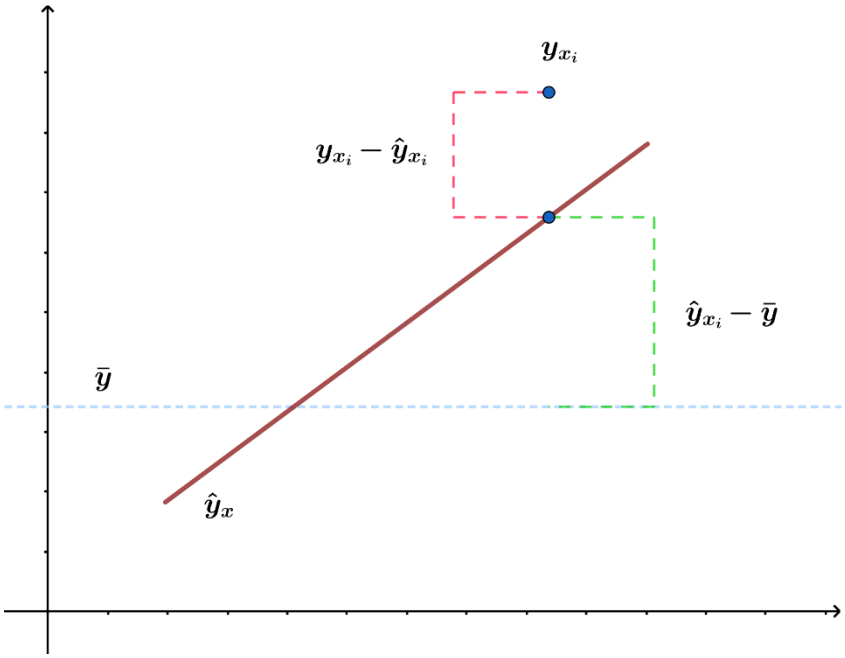
Χρησιμοποιώντας τα υπόλοιπα ορίζουμε το **άθροισμα τετραγώνων των υπολοίπων** (RSS - residual sum of squares) ως εξής

$$RSS = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \boldsymbol{e}'\boldsymbol{e} = (\boldsymbol{y} - \hat{\boldsymbol{y}})'(\boldsymbol{y} - \hat{\boldsymbol{y}}) \tag{1.7}$$

Το RSS είναι επίσης γνωστό ως **άθροισμα τετραγώνων λόγω σφάλματος** (SSE – sum of squares due to error). Ορίζουμε επίσης τη **άθροισμα τετραγώνων λόγω παλινδρόμησης** (SSR – sum of squares due to regression) και το **συνολικό άθροισμα τετραγώνων** (SST – sum of squares total) ως

$$SSR = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 \tag{1.8}$$

$$SST = \sum_{i=1}^{n}(y_i - \bar{y})^2 = SSR + RSS \tag{1.9}$$



Σχήμα 1

Με τη βοήθεια του Σχήματος 1 (*Σχήμα 1*) μπορούμε να κατανοήσουμε καλύτερα τις ποσότητες RSS, SSR και SST. Το SST μετρά την έμφυτη μεταβλητότητα στη μεταβλητή απόκρισης y. Το SSR μετρά την μεταβλητότητα που αναμένουμε να έχουμε η οποία οφείλεται στο μοντέλο που έχουμε προσαρμόσει. Τέλος, το RSS μετρά την μη αναμενόμενη μεταβλητότητα στην πρόβλεψή μας η οποία οφείλεται σε σφάλματα του προσαρμοσμένου μοντέλου.

Ορίζουμε επίσης τον **συντελεστή προσδιορισμού** (coefficient of determination)

$$\mathrm{R}^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} = 1 - \frac{RSS}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1.10}$$

ο οποίος εκφράζει το ποσοστό της μεταβλητότητας της τυχαίας μεταβλητής $\boldsymbol{y}$ που εξηγείται από τις επεξηγηματικές μεταβλητές $\boldsymbol{x}$. Ο συντελεστής προσδιορισμού μπορεί να πάρει τιμές $0 \leq R^2 \leq 1$. Στην περίπτωση του απλού γραμμικού μοντέλου ισχύει $R^2 = r_{xy}^2$ όπου

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

είναι ο **δειγματικός συντελεστής συχέτισης Pearson** που εκφράζει το βαθμο της γραμμικής συσχέτισης μεταξύ των μεταβλητών x και y. Ο συντελεστής συσχέτησης μπορεί να πάρει τιμές $-1 \leq r_{xy} \leq 1$.

1.2 Προσαρμογή Μοντέλου

Η προσαρμογή του μοντέλου (1.3), δηλαδή η εκτίμηση των παραμέτρων $\beta_j,\ j = 0, 1, \dots, k$, μπορεί να γίνει χρησιμοποιώντας πολλές μεθόδους (Seber & Lee, 2003; Rao et al. , 2008). Σε αυτήν την ενότητα θα περιγράψουμε τη μέθοδο ελαχίστων τετραγώνων και τη μέθοδο μεγίστης πιθανοφάνειας.

1.2.1 Μέθοδος Ελαχίστων Τετραγώνων

Για την εκτίμηση των παραμέτρων $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_k)'$, η **μέθοδος ελαχίστων τετραγώνων** (ordinary least squares – OLS) βασίζεται στην ελαχιστοποίηση του αθροίσματος τετραγώνων των υπολοίπων,

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} \varepsilon_i^2 = \boldsymbol{\varepsilon}'\boldsymbol{\varepsilon} = \big(\boldsymbol{y} - E(\boldsymbol{y})\big)'\big(\boldsymbol{y} - E(\boldsymbol{y})\big) = (\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) \tag{1.11}$$

με δεδομένα τα $\boldsymbol{y}$ και $\boldsymbol{X}$**.** Το αποτέλεσμα της μεθόδου ελαχίστων τετραγώνων είναι η εκτιμήτρια $\widehat{\boldsymbol{\beta}}$ η οποία ελαχιστοποιεί το RSS. Γράφουμε την $S(\boldsymbol{\beta})$ ως

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

και παραγωγίζοντας ως προς $\boldsymbol{\beta}$ (βλ. Παράρτημα Α, Ορισμός Α.1 & Θεώρημα Α.1) παίρνουμε

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

Θέτοντας την παραπάνω σχέση ίση με το μηδέν, καταλήγουμε στις **κανονικές εξισώσεις** (normal equations)

$$\boldsymbol{X'X\widehat{\beta}} = \boldsymbol{X'y} \tag{1.12}$$

Εάν ο πίνακας $\boldsymbol{X}$ έχει τάξη p (βλ. Παράρτημα Α, Ορισμός Α.2), δηλαδή $rank(\boldsymbol{X}) = p = k + 1$, τότε ο πίνακας $\boldsymbol{X'X}$ είναι θετικά ορισμένος και επομένως είναι αντιστρέψιμος (βλ. Παράρτημα Α, Θεώρημα Α.5). Τότε η μοναδική λύση για τη σχέση (1.12) δίνεται από τη σχέση

$$\boldsymbol{\widehat{\beta}} = \boldsymbol{(X'X)^{-1}X'y} \tag{1.13}$$

Σε μορφή πινάκων το προσαρμοσμένο μοντέλο γράφεται

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

όπου $\boldsymbol{H} = \boldsymbol{X(X'X)^{-1}X'}$ είναι ο **πίνακας προβολής** (hat matrix) ο οποίος είναι συμμετρικός και ταυτοδύναμος . Κάθε προσαρμοσμένη παρατήρηση $\widehat{y}_i$ εκφράζεται ως

$$\widehat{y}_i = \boldsymbol{x_i'\widehat{\beta}} = \sum_{i=1}^{n} h_{ij} y_j$$

όπου $h_{ij} = (\boldsymbol{H})_{ij} = \boldsymbol{x_i'(X'X)^{-1}x_j}$ είναι στοιχεία του πίνακα $\boldsymbol{H}$ και y_j είναι μία παρατήρηση για την μεταβλητή y.

Η αναμενόμενη τιμή του $\boldsymbol{\widehat{\beta}}$ είναι

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

δηλαδή η $\boldsymbol{\widehat{\beta}}$ είναι αμερόληπτη εκτιμήτρια της $\boldsymbol{\beta}$**.** Ο πίνακας διασποράς-συνδιασποράς του $\boldsymbol{\widehat{\beta}}$ είναι

$$\begin{aligned} V\big(\boldsymbol{\widehat{\beta}}\big) &= E\left[\Big(\boldsymbol{\widehat{\beta}} - E\big(\boldsymbol{\widehat{\beta}}\big)\Big)\Big(\boldsymbol{\widehat{\beta}} - E\big(\boldsymbol{\widehat{\beta}}\big)\Big)'\right] \\ &= V\big(\boldsymbol{(X'X)^{-1}X'y}\big) = \boldsymbol{(X'X)^{-1}X'}V(\boldsymbol{y})\big\{\boldsymbol{(X'X)^{-1}X'}\big\}' \\ &= \boldsymbol{(X'X)^{-1}X'}\sigma^2\boldsymbol{I(X'X)^{-1}X'} = \sigma^2\boldsymbol{(X'X)^{-1}X'X(X'X)^{-1}} \\ &= \sigma^2\boldsymbol{(X'X)^{-1}} = \sigma^2\boldsymbol{C} \end{aligned}$$

Εφόσον ισχύει η υπόθεση της Κανονικής κατανομής των $\boldsymbol{\varepsilon}$, η κατανομή του διανύσματος $\boldsymbol{\widehat{\beta}}$ είναι

$$\boldsymbol{\widehat{\beta}} \sim N_p(\boldsymbol{\beta}, \sigma^2\boldsymbol{C})$$

(Καρώνη & Οικονόμου, 2017). Για κάθε συνιστώσα $\hat{\beta}_j$ του $\boldsymbol{\widehat{\beta}}$ ισχύει $E\big(\hat{\beta}_j\big) = \beta_j$, δηλαδή αποτελεί αμερόληπτη εκτιμήτρια της αντίστοιχης παραμέτρου β_j. Επίσης ισχύει $V\big(\hat{\beta}_j\big) = \sigma^2 c_{jj}$, όπου $c_{jj} = \boldsymbol{(X'X)}_{jj}^{-1}$ είναι το j-οστό διαγώνιο στοιχείο του συμμετρικού πίνακα $\boldsymbol{(X'X)^{-1}}$**.**

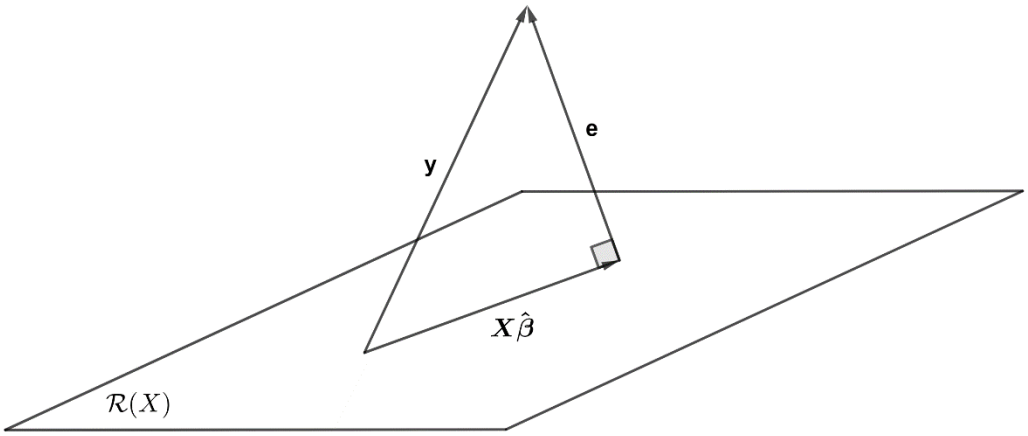
Σύμφωνα με το θεώρημα Markov-Gauss (Καρώνη & Οικονόμου, 2017), η εκτιμήτρια ελαχίστων τετραγώνων $\boldsymbol{\widehat{\beta}}$ είναι η βέλτιστη γραμμική, αμερόληπτη εκτιμήτρια του $\boldsymbol{\beta}$ με την ελάχιστη διασπορά (BLUE – Best Linear Unbiased Estimators). Το θεώρημα αυτό ισχύει εφόσον πληρούνται οι υποθέσεις της ενότητας 1.1, εξαιρώντας την υπόθεση της Κανονικότητας η οποία χρησιμοποιείται για εκτέλεση στατιστικών ελέγχων για το γραμμικό μοντέλο.

Στη συνέχεια θα περιγράψουμε τη γεωμετρική ερμηνεία της μεθόδου ελαχίστων τετραγώνων (Seber & Lee, 2003; Rao et al. , 2008) . Για ένα $n \times p$ πίνακα $\boldsymbol{X}$ ορίζουμε το χώρο στηλών

$$\mathcal{R}(\boldsymbol{X}) = \{\boldsymbol{\theta}: \boldsymbol{\theta} = \boldsymbol{X\beta}, \boldsymbol{\beta} \in \mathbb{R}^p\}$$

ο οποίος είναι υπόχωρος του $\mathbb{R}^n$. Επιλέγοντας τη νόρμα $\|\boldsymbol{x}\| = (\boldsymbol{x}'\boldsymbol{x})^{1/2}$ για $\boldsymbol{x} \in \mathbb{R}^n$, τότε η μέθοδος ελαχίστων τετραγώνων είναι ισοδύναμη με το πρόβλημα ελαχιστοποίησης

$$\min_{\theta \in \mathcal{R}(\boldsymbol{X})} \|\boldsymbol{y} - \boldsymbol{\theta}\|$$



Σχήμα 2 Στη μέθοδο ελαχίστων τετραγώνων θέλουμε να βρούμε $\boldsymbol{\beta} \in \mathbb{R}^{\boldsymbol{p}}$ ώστε να ελαχιστοποιείται το υπόλοιπο $\boldsymbol{e} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$

Όπως φαίνεται στο Σχήμα 2, η ποσότητα $\|\boldsymbol{y} - \boldsymbol{\theta}\|, \boldsymbol{\theta} \in \mathcal{R}(\boldsymbol{X})$, ελαχιστοποιείται στο $\widehat{\boldsymbol{\theta}} = \boldsymbol{X}\widehat{\boldsymbol{\beta}}$ για το οποίο ισχύει $\boldsymbol{e} = \left(\boldsymbol{y} - \widehat{\boldsymbol{\theta}}\right) \perp \mathcal{R}(\boldsymbol{X})$. Δηλαδή, το $\boldsymbol{y} - \widehat{\boldsymbol{\theta}}$ είναι κάθετο σε όλα τα διανύσματα του $\mathcal{R}(\boldsymbol{X})$, που σημαίνει ότι το $\widehat{\boldsymbol{\theta}}$ είναι η ορθογώνια προβολή του $\boldsymbol{y}$ στον $\mathcal{R}(\boldsymbol{X})$. Τότε θα ισχύει

$$\boldsymbol{X}'\left(\boldsymbol{y} - \widehat{\boldsymbol{\theta}}\right) = \boldsymbol{0} \;\Leftrightarrow \boldsymbol{X}'\widehat{\boldsymbol{\theta}} = \boldsymbol{X}'\boldsymbol{y} \Leftrightarrow \boldsymbol{X}'\boldsymbol{X}\widehat{\boldsymbol{\beta}} = \boldsymbol{X}'\boldsymbol{y}$$

δηλαδή καταλήγουμε στις κανονικές εξισώσεις (1.12).

1.2.2 Μέθοδος Μέγιστης Πιθανοφάνειας

Υπό τις υποθέσεις που κάναμε για τα σφάλματα του μοντέλου, έχουμε ότι

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

Επομένως

$$y_i = \boldsymbol{x_i'}\boldsymbol{\beta} + \varepsilon_i \sim N(\boldsymbol{x_i'}\boldsymbol{\beta}, \sigma^2)$$

Τότε η συνάρτηση πιθανοφάνειας για την μεταβλητή $\boldsymbol{y}$ είναι

$$L(\beta, \sigma^2) = (2\pi)^{-\frac{n}{2}} \frac{1}{|\sigma^2\boldsymbol{I}|^{\frac{1}{2}}} \exp\left\{-\frac{1}{2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\sigma^2\boldsymbol{I})^{-1}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})'(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\right\}$$

Η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας είναι

$$l = \ln L = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta})$$

Εάν δεν υπάρχουν περιορισμοί για τις παραμέτρους, τότε ο χώρος των παραμέτρων είναι

$$\Omega = \{\beta; \sigma^2 : \beta \in \mathbb{R}^p; \sigma^2 > 0\}$$

Για να βρούμε τις εκτιμήτριες μέγιστης πιθανοφάνειας (ε.μ.π.), εξισώνουμε με το μηδέν τις πρώτες παραγώγους της λογαριθμοποιημένης συνάρτησης πιθανοφάνειας

$$\frac{\partial l}{\partial \beta} = -\frac{1}{\sigma^2}(\boldsymbol{X'X\beta} - \boldsymbol{X'y}) = 0$$

$$\frac{\partial l}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4}(\boldsymbol{y} - \boldsymbol{X\beta})'(\boldsymbol{y} - \boldsymbol{X\beta}) = 0$$

Καταλήγουμε στις εξισώσεις

$$(I) \quad \boldsymbol{X'X\widehat{\beta}} = \boldsymbol{X'y}$$

$$(II) \quad \hat{\sigma}^2 = \frac{1}{n}\left(\boldsymbol{y} - \boldsymbol{X\widehat{\beta}}\right)'\left(\boldsymbol{y} - \boldsymbol{X\widehat{\beta}}\right) = \frac{\boldsymbol{e'e}}{n}$$

Παρατηρούμε ότι η εξίσωση (I) ταυτίζεται με τις κανονικές εξισώσεις (1.12). Αν για τον πίνακα σχεδιασμού $\boldsymbol{X}$ ισχύει $rank(\boldsymbol{X}) = p$ τότε η εκτιμήτρια μέγιστης πιθανοφάνειας για την παράμετρο $\boldsymbol{\beta}$ είναι

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

η οποία ταυτίζεται με την εκτιμήτρια ελαχίστων τετραγώνων.

Μπορεί να δειχθεί ότι η ε.μ.π. $\hat{\sigma}^2$ της σχέσης (II) δεν είναι μεροληπτική εκτιμήτρια της σ^2 δηλαδή ότι

$$E(\hat{\sigma}^2) = \frac{n-p}{n}\sigma^2 \neq \sigma^2$$

Για αυτόν τον λόγο η διασπορά των τυχαίων σφαλμάτων, σ^2, εκτιμάται γενικώς από την αμερόληπτη εκτιμήτρια

$$S^2 = \frac{\boldsymbol{e'e}}{n-p}$$

ΚΕΦΑΛΑΙΟ 2

Μέθοδοι Επίλυσης του Προβλήματος Ελαχίστων Τετραγώνων

Σε αυτό το Κεφάλαιο θα παρουσιάσουμε μερικές από τις βασικές άμεσες μεθόδους (direct methods) που μπορούν να χρησιμοποιηθούν για να λυθεί το πρόβλημα ελαχίστων τετραγώνων με σκοπό την προσαρμογή ενός γραμμικού μοντέλου παλινδρόμησης. Να σημειώσουμε ότι στην περίπτωση που ο πίνακας $\boldsymbol{X}$ είναι μεγάλος και αραιός τότε οι άμεσες μέθοδοι υστερούν επομένως προτιμώνται επαναληπτικές μέθοδοι όπως για παράδειγμα η μέθοδος συζυγών κλίσεων (Conjugate Gradient) (Kontoghiorghes, 2000).

Τρεις βασικές μέθοδοι χρησιμοποιούνται για την προσαρμογή ενός μοντέλου ελαχίστων τετραγώνων. Η πρώτη μέθοδος δημιουργεί τον πίνακα $\boldsymbol{X}'\boldsymbol{X}$ (SSCP, sum of squares and cross-products) και στην συνέχεια εφαρμόζει απαλοιφή Gauss, τη διαδικασία SWEEP ή την παραγοντοποίηση Cholesky για την επίλυση των κανονικών εξισώσεων (1.12). Η δεύτερη μέθοδος χρησιμοποιεί την παραγοντοποίηση QR για να υπολογίσει τον παράγοντα Cholesky του πίνακα $\boldsymbol{X}$. Η παραγοντοποίηση QR μπορεί να υπολογιστεί χρησιμοποιώντας τον τροποποιημένο αλγόριθμο Gram-Schmidt, ανακλάσεις (reflections) Householder ή περιστροφές (rotations) Givens. Η τρίτη μέθοδος χρησιμοποιεί τη μέθοδο Singular Value Decomposition (SVD) του πίνακα $\boldsymbol{X}$.

2.1 Υπολογισμός του πίνακα SSCP ($\boldsymbol{X}'\boldsymbol{X}$)

Υποθέτουμε ότι το μοντέλο μας συμπεριλαμβάνει το σταθερό όρο β_0 επομένως η πρώτη στήλη του πίνακα $\boldsymbol{X}$ αποτελείται από μονάδες. Αν έχουμε $k = p - 1$ μεταβλητές και n παρατηρήσεις τότε ο πίνακας $SSCP$ είναι

$$\boldsymbol{X}'\boldsymbol{X} = \begin{pmatrix} n & \sum_{i=1}^{n} x_{i1} & \cdots & \sum_{i=1}^{n} x_{ik} \\ \sum_{i=1}^{n} x_{i1} & \sum_{i=1}^{n} x_{i1}^2 & \cdots & \sum_{i=1}^{n} x_{i1}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ \sum_{i=1}^{n} x_{ik} & \sum_{i=1}^{n} x_{i1}x_{ik} & \cdots & \sum_{i=1}^{n} x_{ik}^2 \end{pmatrix} \tag{2.1}$$

Ο υπολογισμός του παραπάνω πίνακα χρειάζεται περίπου np^2 πράξεις εκ των οποίων σχεδόν οι μισές μπορούν να αποφευχθούν αν εκμεταλλευτούμε τη συμμετρία του πίνακα. Σημειώνουμε ότι ένα μέτρο της αποδοτικότητας ενός αλγόριθμου, ως προς το κόστος υπολογισμού, είναι οι **πράξεις κινητής υποδιαστολής** (floating-point operations, **flops**) οι οποίες μετρούν το πλήθος των προσθέσεων, αφαιρέσεων, πολλαπλασιασμών και διαιρέσεων ενός αλγορίθμου.

Υποθέτουμε ότι έχουμε υπολογίσει με ακρίβεια τους πίνακες $\boldsymbol{X'X}$ και $\boldsymbol{X'y}$. Στη συνέχεια θα περιγράψουμε μερικούς αλγόριθμους που μπορούμε να χρησιμοποιήσουμε για να επιλύσουμε τις κανονικές εξισώσεις

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

Για σκοπούς απλοποίησης των συμβολισμών, γράφουμε τις κανονικές εξισώσεις ως

$$\boldsymbol{Ax} = \boldsymbol{b} \tag{2.2}$$

όπου $\boldsymbol{A} = \boldsymbol{X'X},\ \boldsymbol{x} = \boldsymbol{\hat{\beta}},\ \boldsymbol{b} = \boldsymbol{X'y}$

2.2 Απαλοιφή Gauss

2.2.1 Υπολογισμός των Συντελεστών του Μοντέλου

Υποθέτουμε ότι ο πίνακας $\boldsymbol{A} = \boldsymbol{X'X}$ είναι ένας $p \times p$ αντιστρέψιμος τετραγωνικός πίνακας. Η μέθοδος απαλοιφής Gauss (Gaussian Elimination, GE) αποτελεί μία κλασική μέθοδο επίλυσης ενός συστήματος γραμμικών εξισώσεων. Ο στόχος της μεθόδου είναι να μειώσει ένα σύστημα p γραμμικών εξισώσεων με p αγνώστους σε τριγωνική μορφή χρησιμοποιώντας στοιχειώδεις γραμμοπράξεις. Για να το πετύχει αυτό χρειάζονται $p-1$ βήματα, αρχίζοντας με το να θέσουμε $\boldsymbol{A}^{(1)} = \boldsymbol{A} \in \mathbb{R}^{p\times p},\ \boldsymbol{b}^{(1)} = \boldsymbol{b}$ και τελειώνοντας με το άνω τριγωνικό σύστημα $\boldsymbol{A}^{(p)}\boldsymbol{x} = \boldsymbol{b}^{(p)}$. Στην αρχή του βήματος k της διαδικασίας, έχουμε ήδη μετατρέψει το αρχικό σύστημα στο σύστημα $\boldsymbol{A}^{(k)}\boldsymbol{x} = \boldsymbol{b}^{(k)}$, όπου

$$\boldsymbol{A}^{(k)} = \begin{array}{c} \\ k-1 \\ p-k+1 \end{array} \begin{array}{c} \begin{array}{cc} k-1 & p-k+1 \end{array} \\ \begin{bmatrix} \boldsymbol{A}_{11}^{(k)} & \boldsymbol{A}_{12}^{(k)} \\ \boldsymbol{0} & \boldsymbol{A}_{22}^{(k)} \end{bmatrix} \end{array}$$

όπου ο πίνακας $\boldsymbol{A}_{11}^{(k)}$ είναι άνω τριγωνικός. Στο βήμα k, ονομάζουμε το στοιχείο $a_{kk}^{(k)}$ του πίνακα $\boldsymbol{A}^{(k)}$ **οδηγό στοιχείο** (pivot element). Ο στόχος μας σε αυτό το βήμα είναι να μηδενίσουμε τα στοιχεία της στήλης k του πίνακα $\boldsymbol{A}^{(k)}$, που βρίσκονται κάτω από το οδηγό στοιχείο, χρησιμοποιώντας τις εξής πράξεις

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - m_{ik}a_{kj}^{(k)},\quad i = k+1{:}\,p,\ \ j = k{:}\,p$$
$$b_i^{(k+1)} = b_i^{(k)} - m_{ik}b_k^{(k)},\quad\ \ i = k+1{:}\,p$$

όπου οι ποσότητες

$$m_{ik} = \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}}\ ,\qquad i = k+1{:}\,p$$

ονομάζονται **πολλαπλασιαστές**. Προφανώς για $j = k$ έχουμε $a_{ij}^{(k+1)} = 0,\ i = k+1:p$, επομένως δε χρειάζεται να κάνουμε τους υπολογισμούς για αυτά τα στοιχεία. Παρατηρούμε ότι τα στοιχεία $a_k^{(k)}$ πρέπει να είναι διάφορα του μηδενός, κάτι που ισχύει εφόσον ο πίνακας $\boldsymbol{A}$ είναι αντιστρέψιμος. Στο τέλος του βήματος $p-1$ καταλήγουμε στο άνω τριγωνικό σύστημα $\boldsymbol{Ux} \equiv \boldsymbol{A}^{(p)}\boldsymbol{x} = \boldsymbol{b}^{(p)}$ το οποίο επιλύουμε με τη μέθοδο της προς τα πίσω αντικατάστασης χρησιμοποιώντας τις σχέσεις

$$x_p = \frac{b_p}{u_{pp}}$$

$$x_k = \frac{\left(b_k - \sum_{j=k+1}^{p} u_{kj}x_j\right)}{u_{kk}} \quad , \quad k = p-1, p-2, \dots, 1$$

Να σημειώσουμε επίσης ότι στην πράξη αυτό που κάνουμε είναι να ενώσουμε τους πίνακες $\boldsymbol{X'X}$ και $\boldsymbol{X'y}$ σε ένα πίνακα $[\boldsymbol{X'X}|\boldsymbol{X'y}] \equiv [\boldsymbol{A}|\boldsymbol{b}]$ και να εφαρμόσουμε σε αυτόν τον πίνακα τα $p-1$ βήματα της απαλοιφής Gauss καταλήγοντας στον πίνακα $[\boldsymbol{U}|\boldsymbol{c}]$, όπου $\boldsymbol{c} \equiv \boldsymbol{b}^{(p)}$. Μία σημαντική πτυχή των γραμμοπράξεων είναι ότι η εφαρμογή τους σε κάποιο πίνακα ισοδυναμεί με τον πολλαπλασιασμό του πίνακα από αριστερά με κάποιο άλλο πίνακα (Higham, 2011; Seber & Lee, 2003). Το αποτέλεσμα του βήματος k μπορεί να γραφτεί ως $\boldsymbol{A}^{(k+1)} = \boldsymbol{M}_k\boldsymbol{A}^{(k)}$ όπου $\boldsymbol{M}_k = \boldsymbol{I} - \boldsymbol{m}_k\boldsymbol{e}_k^T$ είναι ο πίνακας μετασχηματισμού για το βήμα k, $\boldsymbol{m}_k = \left[0,\cdots,0,m_{k+1,k},\cdots,m_{p,k}\right]^T$ είναι ένα $p \times 1$ διάνυσμα που περιέχει μηδενικά στις πρώτες k θέσεις και τους πολλαπλασιαστές του βήματος k στις υπόλοιπες θέσεις και $\boldsymbol{e}_k$ είναι ένα $p \times 1$ διάνυσμα που περιέχει παντού μηδενικά εκτός το στοιχείο στη θέση k το οποίο είναι η μονάδα (Golub & Van Loan, 2013). Αν $\boldsymbol{v} = \left[v_1, \dots, v_k, v_{k+1}, \dots, v_p\right]^T$ τότε

$$\boldsymbol{M}_k\boldsymbol{v} = \begin{bmatrix} 1 & \cdots & 0 & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & & \vdots \\ 0 & & 1 & 0 & & 0 \\ 0 & & -m_{k+1} & 1 & & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & -m_p & 0 & \cdots & 1 \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_k \\ v_{k+1} \\ \vdots \\ v_p \end{bmatrix} = \begin{bmatrix} v_1 \\ \vdots \\ v_k \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Για τα $p-1$ βήματα της απαλοιφής Gauss έχουμε

$$\boldsymbol{M}_{p-1}\boldsymbol{M}_{p-2}\cdots\boldsymbol{M}_1\boldsymbol{A} = \boldsymbol{MA} = \boldsymbol{A}^{(p)} =: \boldsymbol{U}$$

όπου

$$\boldsymbol{M} = \boldsymbol{M}_{p-1}\boldsymbol{M}_{p-2}\cdots\boldsymbol{M}_1 \tag{2.3}$$

είναι ένας $p \times p$ κάτω τριγωνικός πίνακας του οποίου τα διαγώνια στοιχεία είναι μονάδες.

Χρησιμοποιώντας το γεγονός ότι $\boldsymbol{M}_k^{-1} = \boldsymbol{I} + \boldsymbol{m}_k\boldsymbol{e}_k^T$ είναι εύκολο να δείξουμε ότι

$$\begin{aligned} \boldsymbol{A} &= \boldsymbol{M}_1^{-1}\boldsymbol{M}_2^{-1}\cdots\boldsymbol{M}_{p-1}^{-1}\boldsymbol{U} \\ &= (\boldsymbol{I} + \boldsymbol{m}_1\boldsymbol{e}_1^T)(\boldsymbol{I} + \boldsymbol{m}_2\boldsymbol{e}_2^T)\cdots\left(\boldsymbol{I} + \boldsymbol{m}_{p-1}\boldsymbol{e}_{p-1}^T\right)\boldsymbol{U} \\ &= \left(\boldsymbol{I} + \sum_{i=1}^{p-1}\boldsymbol{m}_i\boldsymbol{e}_i^T\right)\boldsymbol{U} \end{aligned}$$

$$= \begin{bmatrix} 1 & & & & \\ m_{21} & 1 & & & \\ \vdots & m_{32} & \ddots & & \\ \vdots & \vdots & & \ddots & \\ m_{p1} & m_{p2} & \dots & m_{p,p-1} & 1 \end{bmatrix} \boldsymbol{U} =: \boldsymbol{LU}$$

Συμπεραίνουμε ότι η μέθοδος απαλοιφής Gauss ισοδυναμεί με μία παραγοντοποίηση LU, όπου ο πίνακας $\boldsymbol{L}$ είναι κάτω τριγωνικός και ο πίνακας $\boldsymbol{U}$ είναι άνω τριγωνικός.

Συνοψίζοντας, για να μετατρέψουμε το σύστημα $[\boldsymbol{X}'\boldsymbol{X}|\boldsymbol{X}'\boldsymbol{y}]$ σε άνω τριγωνικό, πολλαπλασιάζουμε από αριστερά με $\boldsymbol{M}$ λαμβάνοντας $[\boldsymbol{M}\boldsymbol{X}'\boldsymbol{X}|\boldsymbol{M}\boldsymbol{X}'\boldsymbol{y}] = [\boldsymbol{U}|\boldsymbol{c}]$

Στη συνέχεια θα υπολογίσουμε τον αριθμό των πράξεων που χρειάζονται για την απαλοιφή Gauss. Για το σκοπό αυτό θα χρησιμοποιήσουμε τις σχέσεις

$$\sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{6}$$
$$\sum_{i=1}^{n} i = \frac{n(n+1)}{2}$$

Για την μετατροπή του $p \times p$ πίνακα $\boldsymbol{A}$ σε άνω τριγωνικό, στο βήμα k εκτελούμε

- $p-k$ διαιρέσεις για τον υπολογισμό των πολλαπλασιαστών m_{ik}, $i = k+1:n$
- 1 πολλαπλασιασμό και 1 αφαίρεση για κάθε στοιχείο $a_{ij}^{(k+1)}$, $i,j = k+1:p$. Συνολικά, χρειάζονται $2(p-k)^2$ πράξεις για τον υπολογισμό αυτών των στοιχείων.

Επομένως, για την μετατροπή του πίνακα $\boldsymbol{A}$ σε άνω τριγωνικό εκτελούμε συνολικά

$$\begin{aligned} \sum_{k=1}^{p-1} [(p-k) + 2(p-k)^2] &= \sum_{k=1}^{p} [2p^2 + p - (4p+1)k + 2k^2] \\ &= 2p^3 + p^2 - (4p+1)\sum_{k=1}^{p} k + 2\sum_{k=1}^{p} k^2 \\ &= 2p^3 + p^2 - \frac{p(p+1)(4p+1)}{2} + \frac{2p(p+1)(2p+1)}{6} \\ &= \frac{2}{3}p^3 - \frac{1}{2}p^2 - \frac{1}{6}p = \frac{2}{3}p^3 + O(p^2) \end{aligned}$$

πράξεις.

Για την μετατροπή του διανύσματος $\boldsymbol{b}$, στο βήμα k εκτελούμε

- $p-k$ διαιρέσεις για τον υπολογισμό των πολλαπλασιαστών m_{ik}, $i = k+1:n$.
- 1 πολλαπλασιασμό και 1 αφαίρεση για κάθε στοιχείο $b_i^{(k+1)}$, $i = k+1:p$. Συνολικά, χρειάζονται $2(p-k)$ πράξεις για τον υπολογισμό αυτών των στοιχείων.

Εξαιρώντας τις πράξεις για τους υπολογισμούς των m_{ik}, εφόσον λήφθηκαν υπόψη στις πράξεις για τη μετατροπή του πίνακα $\boldsymbol{A}$ όπως δείξαμε παραπάνω, για την μετατροπή του διανύσματος $\boldsymbol{b}$ εκτελούμε συνολικά

$$\sum_{k=1}^{p-1} 2(p-k) = 2\sum_{k=1}^{p}(p-k) = 2p^2 - 2\sum_{k=1}^{p} k = 2p^2 - p(p+1) = p(p-1)$$

πράξεις.

Για την προς τα πίσω αντικατάσταση εκτελούμε μία διαίρεση για το x_p και για κάθε x_k, $\mathrm{k} = p-1, \dots, 1$ εκτελούμε μία διαίρεση, $p-k$ πολλαπλασιασμούς και $p-k$ προσθέσεις . Επομένως για την προς τα πίσω αντικατάσταση εκτελούμε συνολικά

$$1 + \sum_{k=1}^{p-1}(1 + 2(p-k)) = 1 + p - 1 + 2\sum_{k=1}^{p}(p-k) = 2p^2 + p - 2\sum_{k=1}^{p} k$$
$$= 2p^2 + p - p(p+1) = p^2$$

πράξεις.

Συνοψίζοντας, μπορούμε να εφαρμόσουμε τον αλγόριθμο απαλοιφής Gauss στον πίνακα $[\boldsymbol{X'X}|\boldsymbol{X'y}] \equiv [\boldsymbol{A}|\boldsymbol{b}]$ και έπειτα εφαρμόζοντας προς τα πίσω αντικατάσταση να υπολογίσουμε την εκτιμήτρια ελαχίστων τετραγώνων $\widehat{\boldsymbol{\beta}}$ η οποία λύνει τις κανονικές εξισώσεις

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

Για τη διαδικασία αυτή θα χρειαστεί να εκτελέσουμε το πολύ

$$\left(\frac{2}{3}p^3 - \frac{1}{2}p^2 - \frac{1}{6}p\right) + p(p-1) + p^2 = \frac{4p^3 + 9p^2 - 7p}{6} = \frac{2}{3}p^3 + O(p^2)$$

πράξεις.

2.2.2 Υπολογισμός του RSS

Ο αλγόριθμος απαλοιφής Gauss μπορεί επίσης να χρησιμοποιηθεί για να προσαρμόσουμε ένα μοντέλο γραμμικής παλινδρόμησης επιλύοντας το πρόβλημα ελαχίστων τετραγώνων και υπολογίζοντας το άθροισμα τετραγώνων των υπολοίπων, RSS (Seber & Lee, 2003).

Θέτουμε $\boldsymbol{X_A} = (\boldsymbol{X}, \boldsymbol{y})$. Τότε εφαρμόζοντας τον αλγόριθμο GE στον $(p+1) \times (p+1)$ πίνακα

$$\boldsymbol{X_A'X_A} = \begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'y} \\ \boldsymbol{y'X} & \boldsymbol{y'y} \end{bmatrix} \tag{2.4}$$

λαμβάνουμε

$$\begin{bmatrix} \boldsymbol{M} & \boldsymbol{0} \\ \boldsymbol{m}' & 1 \end{bmatrix} \begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'y} \\ \boldsymbol{y'X} & \boldsymbol{y'y} \end{bmatrix} = \begin{bmatrix} \boldsymbol{U} & \boldsymbol{c} \\ \boldsymbol{0}' & d \end{bmatrix} \tag{2.5}$$

όπου $\boldsymbol{M}, \boldsymbol{U}$ είναι $p \times p$ πίνακες και $\boldsymbol{m}, \boldsymbol{c}$ είναι $p \times 1$ διανύσματα.

Πολλαπλασιάζοντας τους πίνακες και εξισώνοντας τα αντίστοιχα μέρη λαμβάνουμε

$$\boldsymbol{MX'X} = \boldsymbol{U}$$
$$\boldsymbol{MX'y} = \boldsymbol{c}$$
$$\boldsymbol{m'X'y} + \boldsymbol{y'y} = d$$
$$\boldsymbol{m'X'X} + \boldsymbol{y'X} = \boldsymbol{0'}$$

Επομένως, οι πίνακες $\boldsymbol{M}, \boldsymbol{U}$ και το διάνυσμα $\boldsymbol{c}$ είναι οι ίδιες ποσότητες που λαμβάνουμε από την εφαρμογή του αλγορίθμου στον πίνακα $[\boldsymbol{X'X}|\boldsymbol{X'y}]$. Λύνοντας τις δύο τελευταίες από τις παραπάνω εξισώσεις λαμβάνουμε

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

και

$$RSS \equiv d = \boldsymbol{y'y} - \widehat{\boldsymbol{\beta}}\boldsymbol{X'y}$$

(βλ. Παράρτημα Β, Β.1) όπου RSS είναι το άθροισμα τετραγώνων των υπολοίπων για το πλήρες μοντέλο με τις $k = p - 1$ μεταβλητές. Σημειώνουμε ότι η εφαρμογή του αλγόριθμου GE στον πίνακα της σχέσης (2.4) έχει αποτέλεσμα μόνο τον πίνακα στο δεξί μέρος της σχέσης (2.5) που σημαίνει ότι λαμβάνουμε μόνο το RSS του μοντέλου χωρίς να υπολογίζουμε στην πράξη την εκτιμήτρια $\widehat{\boldsymbol{\beta}}$.

Επίσης, για την εφαρμογή του αλγόριθμου GE στον πίνακα της σχέσης (2.4) χρειάζεται να εκτελέσουμε το πολύ

$$\frac{2}{3}(p+1)^3 - \frac{1}{2}(p+1)^2 - \frac{1}{6}(p+1) = \frac{2}{3}p^3 + \frac{3}{2}p^2 + \frac{11}{6}p = \frac{2}{3}p^3 + O(p^2)$$

πράξεις.

2.2.3 Προσαρμογή Υπομοντέλου

Αξίζει να σημειώσουμε ότι με τον αλγόριθμο απαλοιφής Gauss μπορούμε επίσης να προσαρμόσουμε ένα μοντέλο χρησιμοποιώντας ένα υποσύνολο από τις k μεταβλητές (στήλες) του πίνακα $\boldsymbol{X}$. Ας υποθέσουμε ότι χωρίζουμε τον πίνακα $\boldsymbol{X}$ ως $\boldsymbol{X} = (\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{X}_3)$, όπου οι πίνακες $\boldsymbol{X}_i$ περιέχουν οσεσδήποτε στήλες. Εάν θέλουμε να προσαρμόσουμε το μοντέλο που περιέχει τις μεταβλητές των πινάκων $\boldsymbol{X}_1$ και $\boldsymbol{X}_3$ τότε προφανώς μπορούμε να εφαρμόσουμε την απαλοιφή Gauss στον πίνακα $\big((\boldsymbol{X}_1, \boldsymbol{X}_3), \boldsymbol{y}\big)'\big((\boldsymbol{X}_1, \boldsymbol{X}_3), \boldsymbol{y}\big) = \widetilde{\boldsymbol{X}}_{\boldsymbol{A}}'\widetilde{\boldsymbol{X}}_{\boldsymbol{A}}$ όπως δείξαμε προηγουμένως. Ωστόσο, μπορεί να δειχθεί ότι αυτό είναι ισοδύναμο με το να εφαρμόσουμε τον αλγόριθμο σε ολόκληρο τον πίνακα $(\boldsymbol{X}, \boldsymbol{y})'(\boldsymbol{X}, \boldsymbol{y}) = \boldsymbol{X}_{\boldsymbol{A}}'\boldsymbol{X}_{\boldsymbol{A}}$ αγνοώντας τα βήματα που αντιστοιχούν στις μεταβλητές του πίνακα $\boldsymbol{X}_2$, τις οποίες δε θέλουμε να συμπεριλάβουμε στο μοντέλο (Seber & Lee, 2003). Για παράδειγμα, εάν έχουμε τρεις μεταβλητές και θέλουμε να προσαρμόσουμε το μοντέλο με την 1η και 3η μεταβλητή τότε θα εκτελέσουμε το βήμα 1 του αλγορίθμου, θα αγνοήσουμε το βήμα 2 και τέλος θα εκτελέσουμε το βήμα 3. Το ίδιο ισχύει και στην περίπτωση όπου διαχωρίζουμε τον πίνακα $\boldsymbol{X}$ σε περισσότερα μέρη. Το RSS του προσαρμοσμένου μοντέλου θα είναι το κάτω-δεξιά στοιχείο του αποτελέσματος του αλγορίθμου.

2.2.4 Ακρίβεια

Εάν σε κάποιο στάδιο της διαδικασίας απαλοιφής Gauss το οδηγό στοιχείο είναι μικρό, τότε, λόγω της διαίρεσης που εκτελούμε με αυτό το στοιχείο, στον τελικό άνω τριγωνικό πίνακα $\boldsymbol{U}$ θα υπάρχουν στοιχεία τα οποία θα είναι πολύ μεγάλα σε σύγκριση με τα υπόλοιπα στοιχεία του πίνακα. Αυτό θα οδηγήσει σε ανακρίβειες στη λύση μας όταν αυτή θα υπολογιστεί με την προς τα πίσω αντικατάσταση (Golub & Van Loan, 2013; Seber & Lee, 2003). Η ακρίβεια της μεθόδου απαλοιφής Gauss μπορεί να βελτιωθεί χρησιμοποιώντας μεθόδους οδήγησης. Μία τέτοια μέθοδος είναι αυτή της μερικής οδήγησης σύμφωνα με την οποία στην αρχή του βήματος j εναλλάσσουμε τη γραμμή j με τη γραμμή j' για την οποία ισχύει

$$a_{j'j}^{(j-1)} = \max_{l \geq j} \left| a_{lj}^{(j-1)} \right|$$

δηλαδή ως οδηγό στοιχείο χρησιμοποιούμε το μεγαλύτερο κατά μέτρο στοιχείο της στήλης j του υποπίνακα για το βήμα j.Ο υπολογιστικός χρόνος που χρειάζεται για τις εναλλαγές των γραμμών είναι μικρός συγκριτικά με το συνολικό χρόνο της απαλοιφής Gauss. Για την εναλλαγή των γραμμών εκτελούνται $O(p^2)$ συγκρίσεις στοιχείων ενώ για την απαλοιφή Gauss εκτελούνται περίπου $\frac{2}{3}p^3$ flops.

Ακόμη μία μέθοδος είναι η μέθοδος πλήρους οδήγησης όπου χρησιμοποιούμε σαν οδηγό στοιχείο το μεγαλύτερο κατά μέτρο στοιχείο του υποπίνακα για το βήμα j, εναλλάσσοντας τόσο τις γραμμές όσο και τις στήλες του πίνακα. Για την πλήρη οδήγηση ο επιπλέον υπολογιστικός χρόνος είναι σημαντικός εφόσον σε αυτήν την περίπτωση εκτελούνται $O(p^3)$ συγκρίσεις στοιχείων. Παρ'όλο που η μέθοδος απαλοιφής Gauss με πλήρη οδήγηση θεωρείται ευσταθής, γενικά δεν προτιμάται έναντι της μερικής οδήγησης εκτός στην περίπτωση όπου για τον πίνακα $\boldsymbol{A}$ ισχύει $rank(\boldsymbol{A}) < p$ δηλαδή ο πίνακας $\boldsymbol{A}$ έχει ανεπαρκή τάξη (rank deficient), και τότε η μέθοδος πλήρους οδήγησης μπορεί να χρησιμοποιηθεί για την εύρεση της τάξης του πίνακα (Golub & Van Loan, 2013).

2.3 Απαλοιφή Gauss-Jordan

Όπως είδαμε προηγουμένως, η μέθοδος απαλοιφής Gauss οδηγεί σε ένα άνω τριγωνικό πίνακα για τους συντελεστές των αγνώστων ενός γραμμικού συστήματος. Μία επέκταση της διαδικασίας αυτής είναι η μέθοδος απαλοιφής Gauss-Jordan η οποία αποτελεί μία μέθοδο διαγωνοποίησης του πίνακα των συντελεστών των αγνώστων. Γράφουμε το σύστημα της σχέσης (2.2) ως $[\boldsymbol{A}|\boldsymbol{b}]$. Αρχίζουμε θέτοντας $\left[\boldsymbol{A}^{(1)}\middle|\boldsymbol{b}^{(1)}\right] \equiv [\boldsymbol{A}|\boldsymbol{b}]$. Στο k βήμα του αλγόριθμου Gauss-Jordan χρησιμοποιούμε ως οδηγό στοιχείο το στοιχείο $a_{kk}^{(k)}$ του πίνακα $\boldsymbol{A}^{(k)}$, δηλαδή το στοιχείο που βρίσκεται στην k θέση της κύριας διαγωνίου του $\boldsymbol{A}^{(k)}$. Ο στόχος μας είναι να μηδενίσουμε όλα τα στοιχεία της στήλης k του πίνακα $\boldsymbol{A}^{(k)}$ που έχουμε λάβει από το προηγούμενο βήμα του αλγόριθμου και στη θέση του οδηγού στοιχείου να καταλήξουμε να έχουμε τη μονάδα. Συμβολίζουμε το αποτέλεσμα του βήματος k ως $\left[\boldsymbol{A}^{(k+1)}\middle|\boldsymbol{b}^{(k+1)}\right]$. Οι πράξεις που εκτελούμε στο βήμα k, $k = 1, \dots, p$ περιγράφονται από τις παρακάτω σχέσεις.

$$a_{kj}^{(k+1)} = \frac{a_{kj}^{(k)}}{a_{kk}^{(k)}} \quad , \quad j = k{:}\, p+1$$

$$a_{ij}^{(k+1)} = a_{ij}^{(k)} - a_{ik}^{(k)} a_{kj}^{(k+1)} \qquad , \qquad \forall\, i \neq k, j = k{:}\, p+1$$
$$= a_{ij}^{(k)} - \frac{a_{ik}^{(k)}}{a_{kk}^{(k)}} a_{kj}^{(k)}$$

όπου $a_{ij}^{(k)}$ είναι το (i,j) στοιχείο του πίνακα $\left[\boldsymbol{A}^{(k)} \middle| \boldsymbol{b}^{(k)}\right]$. Είναι εύκολο να δούμε ότι

$$a_{ik}^{(k+1)} = \begin{cases} 1, & i = k \\ 0, & i \neq k \end{cases}$$

επομένως στην πράξη δε χρειάζεται να εκτελέσουμε αυτούς τους υπολογισμούς.

Στη θέση του πίνακα $\boldsymbol{A}$, στο τέλος του αλγορίθμου θα καταλήξουμε να έχουμε τον μοναδιαίο πίνακα $\boldsymbol{I}$, δηλαδή καταλήγουμε σε ένα πίνακα-σύστημα $[\boldsymbol{I}|\boldsymbol{c}]$. Όπως αναφέραμε παραπάνω, η εκτέλεση των γραμμοπράξεων του αλγορίθμου ισοδυναμεί με τον πολλαπλασιασμό του πίνακα $[\boldsymbol{X}'\boldsymbol{X}|\boldsymbol{X}'\boldsymbol{y}] \equiv [\boldsymbol{A}|\boldsymbol{b}]$ από αριστερά με κάποιον άλλο πίνακα. Βάσει των παραπάνω, έχουμε ότι ο αλγόριθμος Gauss-Jordan ισοδυναμεί με τον πολλαπλασιασμό από αριστερά με τον πίνακα $(\boldsymbol{X}'\boldsymbol{X})^{-1}$, δηλαδή το αποτέλεσμα του αλγορίθμου είναι ο πίνακας $[\boldsymbol{I}|(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y}] \equiv \left[\boldsymbol{I}\middle|\widehat{\boldsymbol{\beta}}\right]$, όπου $\widehat{\boldsymbol{\beta}}$ είναι η εκτιμήτρια ελαχίστων τετραγώνων για το μοντέλο παλινδρόμησης που περιέχει τις μεταβλητές (στήλες) του πίνακα $\boldsymbol{X}$.

Για τη διαδικασία επίλυσης του συστήματος (2.2), στο βήμα k εκτελούμε

- $p-k+1$ διαιρέσεις για τον υπολογισμό των στοιχείων $a_{kj}^{(k+1)}$
- $(p-1)\times(p-k+1)$ πολλαπλασιασμούς και $(p-1)\times(p-k+1)$ αφαιρέσεις για τον υπολογισμό των στοιχείων $a_{ij}^{(k+1)}$

Συνολικά εκτελούμε

$$\sum_{k=1}^{p}(p-k+1) + 2\sum_{k=1}^{p}(p-1)(p-k+1) = (2p-1)\sum_{k=1}^{p}(p-k+1)$$
$$= p^3 + \frac{1}{2}p^2 - \frac{1}{2}p = p^3 + O(p^2)$$

πράξεις.

Εκτός από τον υπολογισμό της εκτιμήτριας $\widehat{\boldsymbol{\beta}}$, ο αλγόριθμος Gauss-Jordan μπορεί να χρησιμοποιηθεί για να υπολογίσουμε ταυτόχρονα και το RSS του προσαρμοσμένου μοντέλου (Goodnight, 1979). Εφαρμόζουμε τα p βήματα του αλγορίθμου στον $(p+1)\times(p+1)$ πίνακα

$$\boldsymbol{X}_{\boldsymbol{A}}'\boldsymbol{X}_{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{bmatrix}$$

της σχέσης (2.4), χρησιμοποιώντας ως οδηγό στοιχείο το στοιχείο a_{kk} του πίνακα $\boldsymbol{X}'\boldsymbol{X}$ για το βήμα k. Αυτό ισοδυναμεί με τον πολλαπλασιασμό του $\boldsymbol{X}_{\boldsymbol{A}}'\boldsymbol{X}_{\boldsymbol{A}}$ από αριστερά με τον πίνακα

$$\begin{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{0} \\ -\boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1} & \boldsymbol{I} \end{bmatrix}$$

Το αποτέλεσμα του αλγορίθμου είναι ο πίνακας

$$\begin{bmatrix} \boldsymbol{I} & (\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{0}' & \boldsymbol{y}'\boldsymbol{y} - \boldsymbol{y}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{y} \end{bmatrix} \equiv \begin{bmatrix} \boldsymbol{I} & \widehat{\boldsymbol{\beta}} \\ \boldsymbol{0}' & RSS_p \end{bmatrix}$$

Για αυτήν τη διαδικασία εκτελούμε συνολικά

$$p^3 + \frac{1}{2}p^2 - \frac{1}{2}p + 2\sum_{k=1}^{p}(p-k+1) = p^3 + \frac{1}{2}p^2 - \frac{1}{2}p + (p^2+p)$$
$$= p^3 + \frac{3}{2}p^2 + \frac{1}{2}p = p^3 + O(p^2)$$

πράξεις, όπου οι επιπλέον $p(p+1)$ πράξεις αναλογούν στην γραμμή $[\boldsymbol{y'X}|\boldsymbol{y'y}]$ του πίνακα $\boldsymbol{X_A'X_A}$.

Χρησιμοποιώντας τον αλγόριθμο Gauss-Jordan μπορούμε επίσης να υπολογίσουμε τον αντίστροφο $(\boldsymbol{X'X})^{-1}$ (Goodnight, 1979). Χρησιμοποιώντας ως οδηγό στοιχείο το στοιχείο a_{kk} του πίνακα $\mathbf{X'X}$ για το βήμα k, εφαρμόζουμε τον αλγόριθμο στον πίνακα

$$\begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'y} & \boldsymbol{I} \\ \boldsymbol{y'X} & \boldsymbol{y'y} & \mathbf{0'} \end{bmatrix}$$

και στο τέλος του αλγορίθμου μετά από p βήματα καταλήγουμε στον πίνακα

$$\begin{bmatrix} \boldsymbol{I} & \widehat{\boldsymbol{\beta}} & (\boldsymbol{X'X})^{-1} \\ \mathbf{0'} & \boldsymbol{RSS_p} & -\widehat{\boldsymbol{\beta}} \end{bmatrix}$$

Αλγόριθμος SWEEP

Ο αλγόριθμος SWEEP αποτελεί μία παραλλαγή της απαλοιφής Gauss-Jordan που μας επιτρέπει να υπολογίσουμε τις ποσότητες $\widehat{\boldsymbol{\beta}}, (\boldsymbol{X'X})^{-1}, RSS$ για την προσαρμογή ενός γραμμικού μοντέλου παλινδρόμησης (Goodnight, 1979; Seber & Lee, 2003). Επίσης, μέσω του αλγορίθμου SWEEP μπορούμε να προσθέσουμε ή να αφαιρέσουμε μεταβλητές σε ένα μοντέλο παλινδρόμησης με αποδοτικό τρόπο. Εφαρμόζουμε ένα SWEEP στη γραμμή (ή στήλη) k ενός τετραγωνικού πίνακα $\boldsymbol{A}$, μετασχηματίζοντάς τον σε ένα πίνακα $\boldsymbol{A}^*$, όπου τα στοιχεία του νέου πίνακα $\boldsymbol{A}^* = \left(a_{ij}^*\right)$ ορίζονται από τις σχέσεις

$$a_{kk}^* = \frac{1}{a_{kk}}$$
$$a_{ik}^* = -\frac{a_{ik}}{a_{kk}} \quad , \quad (i \neq k)$$
$$a_{kj}^* = \frac{a_{kj}}{a_{kk}} \quad , \quad (j \neq k)$$
$$a_{ij}^* = a_{ij} - \frac{a_{ik}a_{kj}}{a_{kk}} \quad , \quad (i \neq k, j \neq k)$$

Από τον παραπάνω ορισμό συμπεραίνουμε ότι ο αλγόριθμος SWEEP:

- είναι αντιστρέψιμος (δηλαδή, εφαρμόζοντας ένα SWEEP δύο φορές στη γραμμή k ισοδυναμεί με το να μην εφαρμόσουμε κανένα SWEEP)

- έχει την αντιμεταθετική ιδιότητα (δηλαδή, εφαρμόζοντας ένα SWEEP στη γραμμή r και μετά στη γραμμή s ισοδυναμεί με το να εφαρμόσουμε ένα SWEEP στη γραμμή s και μετά στη γραμμή r)

Εφαρμόζοντας διαδοχικά SWEEP στις γραμμές $k = 1,2, \dots, p$ του $(p+1) \times (p+1)$ πίνακα

$$\boldsymbol{X}_{\boldsymbol{A}}'\boldsymbol{X}_{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{bmatrix}$$

θα καταλήξουμε στον πίνακα

$$\begin{bmatrix} (\boldsymbol{X}'\boldsymbol{X})^{-1} & \widehat{\boldsymbol{\beta}} \\ -\widehat{\boldsymbol{\beta}}' & RSS_p \end{bmatrix}$$

Ένα SWEEP στη γραμμή k του πίνακα $\boldsymbol{X}_{\boldsymbol{A}}'\boldsymbol{X}_{\boldsymbol{A}}$ εισαγάγει στο μοντέλο τη μεταβλητή που αντιστοιχεί σε αυτήν τη γραμμή (και στήλη), εάν η μεταβλητή δε βρίσκεται ήδη στο μοντέλο. Στην περίπτωση που η μεταβλητή αυτή ήδη υπάρχει στο μοντέλο τότε ένα SWEEP στη γραμμή k θα αφαιρέσει τη μεταβλητή από το μοντέλο.

Για τη διαδικασία των p διαδοχικών SWEEP στον πίνακα $\boldsymbol{X}_{\boldsymbol{A}}'\boldsymbol{X}_{\boldsymbol{A}}$, για κάθε SWEEP εκτελούμε

- 1 διαίρεση για τον υπολογισμό του a_{kk}^*
- p διαιρέσεις για τον υπολογισμό των a_{ik}^*
- p διαιρέσεις για τον υπολογισμό των a_{kj}^*
- p^2 διαιρέσεις, p^2 πολλαπλασιασμούς και p^2 αφαιρέσεις για τον υπολογισμό των a_{ij}^*

Δηλαδή για κάθε SWEEP σε ένα $(p+1) \times (p+1)$ πίνακα χρειάζονται $3p^2 + 2p + 1$ πράξεις. Επομένως συνολικά για την παραπάνω διαδικασία προσαρμογής του μοντέλου με τις p παραμέτρους εκτελούμε

$$p(3p^2 + 2p + 1) = 3p^3 + 2p^2 + p = 3p^3 + O(p^2)$$

πράξεις.

2.4 Παραγοντοποίηση Cholesky και $\boldsymbol{LDL}^T$

Εφόσον έχουμε υποθέσει ότι για τον $n \times p$ πίνακα $\boldsymbol{X}$ ισχύει $rank(\boldsymbol{X}) = p$, τότε από το Θεώρημα Α.5 έχουμε ότι ο συμμετρικός πίνακας $\boldsymbol{X}'\boldsymbol{X}$ είναι θετικά ορισμένος και αντιστρέψιμος. Τότε από το Θεώρημα Α.7 έχουμε ότι ο πίνακας $\boldsymbol{A} \equiv \boldsymbol{X}'\boldsymbol{X}$ μπορεί να γραφτεί ως

$$\boldsymbol{A} = \boldsymbol{L}\boldsymbol{D}\boldsymbol{L}^{\boldsymbol{T}}$$

όπου $\boldsymbol{L}$ είναι ένας κάτω τριγωνικός πίνακας με μονάδες στην κύρια διαγώνιο και $\boldsymbol{D}$ είναι ένας διαγώνιος πίνακας με θετικά στοιχεία στην κύρια διαγώνιο.

Αφού τα στοιχεία της κύριας διαγωνίου του πίνακα $\boldsymbol{D}$ είναι θετικά, μπορούμε να γράψουμε

$$\boldsymbol{D}^{1/2} = \begin{bmatrix} \sqrt{d_1} & & & \\ & \sqrt{d_2} & & \\ & & \ddots & \\ & & & \sqrt{d_p} \end{bmatrix}$$

και τότε λαμβάνουμε

$$\boldsymbol{A} = \boldsymbol{LDL}^T = \left(\boldsymbol{LD}^{1/2}\right)\left(\boldsymbol{D}^{1/2}\boldsymbol{L}^T\right) = \boldsymbol{U}^T\boldsymbol{U} \equiv \boldsymbol{R}^T\boldsymbol{R} = \boldsymbol{GG}^T$$

που αποτελεί την **παραγοντοποίηση Cholesky**, όπου $\boldsymbol{G} = \boldsymbol{R}^{\boldsymbol{T}}$ είναι ένας κάτω τριγωνικός πίνακας και $\boldsymbol{U} \equiv \boldsymbol{R}$ είναι ένας άνω τριγωνικός πίνακας ο οποίος ονομάζεται **παράγοντας Cholesky**.

Αλγόριθμος 1 – Παραγοντοποίηση Cholesky (Seber & Lee, 2003)

Ο παρακάτω αλγόριθμος υπολογίζει την παραγοντοποίηση Cholesky $\boldsymbol{A} = \boldsymbol{R}^{\boldsymbol{T}}\boldsymbol{R}$.

Βήμα 1

$$r_{11} = \sqrt{a_{11}}$$
$$r_{1j} = \frac{a_{1j}}{r_{11}} \quad (j = 2,3,\dots,p)$$

Βήμα 2 Για $i = 2,3,\dots,p-1$,

$$\mathrm{r_{ij}} = 0 \quad (j = 1,\dots,i-1)$$
$$r_{ii} = \left(a_{ii} - \sum_{l=1}^{i-1} r_{li}^2\right)^{1/2}$$
$$r_{ij} = \frac{a_{ij} - \sum_{l=1}^{i-1} r_{li}r_{lj}}{r_{ii}} \qquad (j = i+1,\dots,p)$$

Βήμα 3

$$\mathrm{r}_{pp} = \left(a_{pp} - \sum_{l=1}^{p-1} r_{li}^2\right)^{1/2}$$

Διαφορετικά, μπορούμε να χρησιμοποιήσουμε τον παρακάτω αλγόριθμο (Golub & Van Loan, 2013) για να υπολογίσουμε την παραγοντοποίηση Cholesky $\boldsymbol{A} = \boldsymbol{G}\boldsymbol{G}^T$ αντικαθιστώντας τα στοιχεία a_{ij} του $\boldsymbol{A}$ με τα στοιχεία g_{ij} του $\boldsymbol{G}$ για $i \geq j$, εξοικονομώντας θέσεις μνήμης.

<u>Αλγόριθμος 2 – Παραγοντοποίηση Cholesky</u> (Golub & Van Loan, 2013)

Ο παρακάτω αλγόριθμος υπολογίζει την παραγοντοποίηση Cholesky $\boldsymbol{A} = \boldsymbol{G}\boldsymbol{G}^T$ αντικαθιστώντας τα στοιχεία a_{ij} του $\boldsymbol{A}$ με τα στοιχεία g_{ij} του $\boldsymbol{G}$ για $i \geq j$, εξοικονομώντας θέσεις μνήμης. Με $A(i{:}\, j, k{:}\, l)$ συμβολίζουμε τον υποπίνακα του $\boldsymbol{A}$ που απαρτίζεται από τις γραμμές i έως και j και τις στήλες k έως και l.

for $j = 1{:}\, p$

  if $j > 1$

    $A(j{:}\, p, j) = A(j{:}\, p, j) - A(j{:}\, p, 1{:}\, j-1) \cdot A(j, 1{:}\, j-1)^T$

  end

  $A(j{:}\, p, j) = A(j{:}\, p, j) / \sqrt{A(j, j)}$

end

Ο παραπάνω αλγόριθμος εκτελεί $p^3/3$ flops.

<u>Αλγόριθμος 3 – Αλγόριθμος Παραγοντοποίησης $\boldsymbol{LDL^T}$</u> (Golub & Van Loan, 2013)

Σε αυτόν τον αλγόριθμο τα στοιχεία a_{ij} του πίνακα $\boldsymbol{A}$ αντικαθίστανται με τα στοιχεία l_{ij} του $\boldsymbol{L}$, εάν $i > j$ ή με τα στοιχεία d_i του $\boldsymbol{D}$ εάν $i = j$.

for $j = 1{:}\, p$

  for $i = 1{:}\, j-1$

    $v(i) = A(j, i) A(i, i)$

  end

  $A(j, j) = A(j, j) - A(j, 1{:}\, j-1) \cdot v(1{:}\, j-1)$

  $A(j+1{:}\, p, j) = \big(A(j+1{:}\, p, j) - A(j+1{:}\, p, 1{:}\, j-1) \cdot v(1{:}\, j-1)\big) / A(j, j)$

Ο Αλγόριθμος 3 (*Αλγόριθμος 3*) δεν υπολογίζει τετραγωνικές ρίζες, όπως ο αλγόριθμος Cholesky και εκτελεί περίπου $p^3/3$ flops, δηλαδή περίπου τα μισά σε σύγκριση με την απαλοιφή Gauss. Επίσης αναμένεται να είναι λίγο πιο ακριβής σε σύγκριση με την απαλοιφή Gauss-Jordan (Miller, 2002).

2.4.1 Προσαρμογή Μοντέλου Παλινδρόμησης

Χρησιμοποιώντας την παραγοντοποίηση Cholesky μπορούμε να υπολογίσουμε τους συντελεστές ενός μοντέλου παλινδρόμησης καθώς και το RSS του μοντέλου. Αρχικά υπολογίζουμε την παραγοντοποίηση του πίνακα $\boldsymbol{X'X}$ ως $\boldsymbol{A = X'X = R'R}$. Στη συνέχεια, οι κανονικές εξισώσεις γράφονται

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

Αφού ο $\boldsymbol{R}$ είναι άνω τριγωνικός, μπορούμε να χρησιμοποιήσουμε προς τα πίσω αντικατάσταση για να λύσουμε το σύστημα

$$\boldsymbol{R'z = X'y}$$

ως προς $\boldsymbol{z}$ και στη συνέχεια να λύσουμε το σύστημα

$$\boldsymbol{R\beta = z}$$

ως προς $\boldsymbol{\beta}$ λαμβάνοντας την εκτιμήτρια $\boldsymbol{\hat{\beta}}$**.**

Το RSS δίνεται από τη σχέση

$$\begin{aligned} RSS &= \boldsymbol{y'y - \hat{\beta}'X'X\hat{\beta}} \\ &= \boldsymbol{y'y - \left(R\hat{\beta}\right)'R\hat{\beta}} \\ &= \boldsymbol{y'y - z'z} \end{aligned}$$

Οι υπολογισμοί μπορούν να γίνουν αποδοτικά υπολογίζοντας την παραγοντοποίηση Cholesky για τον πίνακα

$$\boldsymbol{X_A'X_A} = \begin{bmatrix} \boldsymbol{X'X} & \boldsymbol{X'y} \\ \boldsymbol{y'X} & \boldsymbol{y'y} \end{bmatrix}$$

της σχέσης (2.4). Ο παράγοντας Cholesky θα έχει τη μορφή

$$\boldsymbol{R_A} = \begin{bmatrix} \boldsymbol{R} & \boldsymbol{z} \\ \boldsymbol{0'} & d \end{bmatrix}$$

όπου $d = \sqrt{RSS}$.

2.5 Παραγοντοποίηση QR

Υπενθυμίζουμε ότι προσαρμόζοντας ένα μοντέλο παλινδρόμησης ο στόχος μας είναι η ελαχιστοποίηση των υπολοίπων

$$\boldsymbol{e} = \boldsymbol{y} - \widehat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}$$

λύνοντας το πρόβλημα

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \|\boldsymbol{e}\|^2 = \underset{\beta}{\operatorname{argmin}} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2$$

όπου $\boldsymbol{X} \in \mathbb{R}^{n\times p}, n \geq p$ και $\boldsymbol{y}, \boldsymbol{\beta} \in \mathbb{R}^n$ και $\|\cdot\|$ είναι η Ευκλείδεια νόρμα με $\|z\|_2 = (z^T z)^{1/2}$ όπου $z \in \mathbb{R}^n$.

Έστω $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ ένας ορθογώνιος πίνακας (βλ. Παράρτημα Α, Ορισμός Α.3). Είναι γνωστό ότι οι ορθογώνιοι μετασχηματισμοί διατηρούν το εσωτερικό γινόμενο διανυσμάτων και συνεπώς το Ευκλείδειο μέτρο τους. Επομένως, το παραπάνω πρόβλημα ελαχιστοποίησης των τετραγώνων των υπολοίπων είναι ισοδύναμο με το πρόβλημα ελαχιστοποίησης

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}} \|\boldsymbol{Q}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|^2 \tag{2.6}$$

Στην Ενότητα αυτή θα παρουσιάσουμε τρόπους για να επιλέξουμε το $\boldsymbol{Q}$ ώστε το πρόβλημα (2.6) να επιλύεται εύκολα.

Από το Θεώρημα Α.8 έχουμε ότι ο πίνακας $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ μπορεί να παραγοντοποιηθεί ως

$$\boldsymbol{X} = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix} \tag{2.7}$$

όπου $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ είναι ένας ορθογώνιος πίνακας με $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{Q}\boldsymbol{Q}' = \boldsymbol{I}_n$ και $\boldsymbol{R} \in \mathbb{R}^{p\times p}$ είναι ένας άνω τριγωνικός πίνακας με μη-αρνητικά στοιχεία στην κύρια διαγώνιό του ο οποίος ονομάζεται **παράγοντας-$\boldsymbol{R}$** ($\boldsymbol{R}$-factor) του $\boldsymbol{X}$**.** Η παραπάνω παραγοντοποίηση ονομάζεται **παραγοντοποίηση QR**.

Παρατηρούμε ότι εάν ο πίνακας $\boldsymbol{R}$ έχει θετικά διαγώνια στοιχεία τότε

$$\boldsymbol{X}'\boldsymbol{X} = (\boldsymbol{R}' \quad \boldsymbol{0})\boldsymbol{Q}'\boldsymbol{Q}\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix} = \boldsymbol{R}'\boldsymbol{R}$$

και επομένως ο $\boldsymbol{R}$ είναι ο μοναδικός παράγοντας Cholesky του $\boldsymbol{X}'\boldsymbol{X}$**.**

Στη συνέχεια θεωρούμε ότι $rank(\boldsymbol{X}) = p$. Διαμερίζουμε τον πίνακα $\boldsymbol{Q}$ ως

$$\boldsymbol{Q} = (\boldsymbol{Q_1} \quad \boldsymbol{Q_2}),\ \ \boldsymbol{Q_1} \in \mathbb{R}^{n\times p}, \boldsymbol{Q_2} \in \mathbb{R}^{n\times(n-p)}$$

όπου οι στήλες του πίνακα $\boldsymbol{Q}_1$ είναι ορθοκανονικές, δηλαδή ο $\boldsymbol{Q}_1$ είναι ορθογώνιος και οι στήλες του έχουν μέτρο ίσο με τη μονάδα.

Τότε έχουμε

$$\boldsymbol{X} = (\boldsymbol{Q_1} \quad \boldsymbol{Q_2})\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix} = \boldsymbol{Q_1 R} \tag{2.8}$$

και εφόσον ο $\boldsymbol{R}$ είναι αντιστρέψιμος μπορούμε να εκφράσουμε το $\boldsymbol{Q_1}$ με μοναδικό τρόπο ως

$$\boldsymbol{Q_1} = \boldsymbol{X R^{-1}}$$

Η σχέση (2.8) ονομάζεται "**λεπτή**" (thin) ή **οικονομική παραγοντοποίηση QR**.

2.5.1 Προσαρμογή Μοντέλου Παλινδρόμησης

Οι κανονικές εξισώσεις (1.12) γράφονται

$$\boldsymbol{X^T X\beta} = \boldsymbol{X^T y} \Rightarrow \boldsymbol{R^T R\beta} = (\boldsymbol{R^T} \quad \boldsymbol{0})\boldsymbol{Q^T y} = \boldsymbol{R^T d}_1$$

όπου

$$\boldsymbol{Q^T y} = \begin{matrix}p\\ n-p\end{matrix}\begin{pmatrix}\boldsymbol{Q_1^T}\\ \boldsymbol{Q_2^T}\end{pmatrix}\boldsymbol{y} = \begin{pmatrix}\boldsymbol{Q_1^T y}\\ \boldsymbol{Q_2^T y}\end{pmatrix} \equiv \begin{matrix}\boldsymbol{d}_1\\ \boldsymbol{d}_2\end{matrix}\begin{matrix}p\\ n-p\end{matrix}$$

Επομένως, η εκτιμήτρια ελαχίστων τετραγώνων υπολογίζεται επιλύοντας το άνω τριγωνικό σύστημα

$$\boldsymbol{R}\widehat{\boldsymbol{\beta}} = \boldsymbol{d}_1$$

δηλαδή

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{R^{-1} d}_1$$

Μία εναλλακτική προσέγγιση (Golub & Van Loan, 2013; Kontoghiorghes, 2000) είναι μέσω του προβλήματος (2.6) γράφοντας

$$\begin{aligned}
\widehat{\boldsymbol{\beta}} &= \underset{\beta}{\operatorname{argmin}} \|\boldsymbol{Q^T y} - \boldsymbol{Q^T X\beta}\|^2 \\
&= \underset{\beta}{\operatorname{argmin}} \left\| \begin{pmatrix}\boldsymbol{d}_1\\ \boldsymbol{d}_2\end{pmatrix} - \boldsymbol{Q^T Q}\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix}\boldsymbol{\beta} \right\|^2 \\
&= \underset{\beta}{\operatorname{argmin}} \left\| \begin{pmatrix}\boldsymbol{d}_1 - \boldsymbol{R\beta}\\ \boldsymbol{d}_2\end{pmatrix} \right\|^2 \\
&= \underset{\beta}{\operatorname{argmin}} (\|\boldsymbol{d}_1 - \boldsymbol{R\beta}\|^2 + \|\boldsymbol{d}_2\|^2) \\
&= \underset{\beta}{\operatorname{argmin}} (\|\boldsymbol{d}_1 - \boldsymbol{R\beta}\|^2) \\
&= \boldsymbol{R^{-1} d}_1
\end{aligned}$$

Το άθροισμα τετραγώνων των υπολοίπων, RSS, υπολογίζεται από την ποσότητα

$$RSS = \|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\|_2^2 = \|\boldsymbol{Q}^T(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}})\|_2^2 = \left\|\begin{pmatrix}\boldsymbol{d}_1\\ \boldsymbol{d}_2\end{pmatrix} - \begin{pmatrix}\boldsymbol{R}\boldsymbol{R}^{-1}\boldsymbol{d}_1\\ \boldsymbol{0}\end{pmatrix}\right\|_2^2 = \left\|\begin{pmatrix}\boldsymbol{0}\\ \boldsymbol{d}_2\end{pmatrix}\right\|_2^2 = \|\boldsymbol{d}_2\|_2^2$$

Προσθέτουμε ότι, εάν υπολογίσουμε την παραγοντοποίηση QR για τον πίνακα $(\boldsymbol{X}, \boldsymbol{y})$, ο πίνακας $\boldsymbol{R}$ που λαμβάνουμε, τον οποίο θα συμβολίζουμε ως $\boldsymbol{R}^*$, έχει τη μορφή

$$\boldsymbol{R}^* = \begin{pmatrix}\boldsymbol{R} & \boldsymbol{d}_1\\ \boldsymbol{0} & s\end{pmatrix} \tag{2.9}$$

όπου $\boldsymbol{R}$ και $\boldsymbol{d}_1$ είναι οι ποσότητες που αναφέραμε παραπάνω και $s^2 = RSS$ για το προσαρμοσμένο μοντέλο (Seber & Lee, 2003).

2.5.2 Σχηματίζοντας την Παραγοντοποίηση QR

Διάφορες μέθοδοι έχουν προταθεί για τον σχηματισμό της παραγοντοποίησης QR (2.7) (Kontoghiorghes, 2000) η οποία μπορεί να γραφτεί ως

$$\boldsymbol{Q}^T\boldsymbol{X} = \begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix} \tag{2.10}$$

ή ως

$$\boldsymbol{X} = \boldsymbol{Q}_1\boldsymbol{R} \tag{2.11}$$

όπου $\boldsymbol{X} \in \mathbb{R}^{n\times p}$, $\boldsymbol{Q} = (\boldsymbol{Q}_1 \quad \boldsymbol{Q}_2) \in \mathbb{R}^{n\times n}$, $\boldsymbol{R} \in \mathbb{R}^{p\times p}$ και $n > p$. Θεωρούμε ότι $rank(\boldsymbol{X}) = p$. Οι πιο γνωστές μέθοδοι είναι οι ανακλάσεις Householder, οι περιστροφές Givens και ο αλγόριθμος Gram-Schmidt στην κλασική και την τροποποιημένη του μορφή.

Για το υπόλοιπο της Ενότητας θα χρησιμοποιήσουμε τον εξής συμβολισμό για να αναφερόμαστε σε τμήματα πινάκων και διανυσμάτων. Η k-οστή στήλη και γραμμή του πίνακα $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ συμβολίζονται ως $\boldsymbol{X}_{:,k}$ και $\boldsymbol{X}_{k,:}$ αντίστοιχα. Ο υποπίνακας $\boldsymbol{X}_{i:k,j:s}$ έχει διαστάσεις $(k - i + 1) \times (s - j + 1)$ και γράφεται

$$\boldsymbol{X}_{i:k,j:s} = \begin{pmatrix} x_{i,j} & x_{i,j+1} & \cdots & x_{i,s}\\ x_{i+1,j} & x_{i+1,j+1} & \cdots & x_{i+1,s}\\ \vdots & \vdots & \ddots & \vdots\\ x_{k,j} & x_{k,j+1} & \cdots & x_{k,s}\end{pmatrix}$$

Ομοίως, το $\boldsymbol{v}_{i:k}$ αντιπροσωπεύει ένα τμήμα του διανύσματος $\boldsymbol{v} \in \mathbb{R}^n$ που περιέχει $(k - i + 1)$ στοιχεία και γράφεται

$$\boldsymbol{v}_{i:k} = \begin{pmatrix} v_i\\ v_{i+1}\\ \vdots\\ v_k\end{pmatrix}$$

Εάν κάποιος δείκτης παραλείπεται τότε εννοούμε ότι στη θέση του μπαίνει αναλόγως είτε η μονάδα είτε η μεγαλύτερη τιμή που μπορεί να πάρει. Για παράδειγμα, $\boldsymbol{v}_{:j} \equiv \boldsymbol{v}_{1:j}$, $\boldsymbol{v}_{i:} \equiv \boldsymbol{v}_{i:n}$ και $\boldsymbol{v}_{:} \equiv \boldsymbol{v}_{1:n} \equiv \boldsymbol{v}$ για ένα διάνυσμα $\boldsymbol{v} \in \mathbb{R}^n$. Ανάλογα για τον συμβολισμό υποπινάκων. Μία οντότητα μηδενικών διαστάσεων αντιπροσωπεύει ένα κενό (null) πίνακα ή διάνυσμα. Ο πίνακας $\boldsymbol{X}_{i:k,j:s}$ είναι κενός εάν $k < i$ ή $s < j$. Τα διανύσματα θεωρούνται διανύσματα-στήλες. Δηλαδή,

$$\boldsymbol{X}_{k,:} = \begin{pmatrix} x_{k,1} \\ x_{k,2} \\ \vdots \\ x_{k,p} \end{pmatrix} \quad \text{και} \quad \boldsymbol{X}_{k,:}^T = (x_{k,1} \quad x_{k,2} \quad \dots \quad x_{k,p})$$

Σημειώνουμε ότι ο $\boldsymbol{X}_{i:k,j:s}^T$ ταυτίζεται με τον $\left(\boldsymbol{X}_{i:k,j:s}\right)^T$ και όχι με τον $(\boldsymbol{X}^T)_{i:k,j:s}$ ο οποίος αντιπροσωπεύει τον $(k-i+1) \times (s-j+1)$ υποπίνακα του $\boldsymbol{X}^T$.

2.5.2.1 Μέθοδος Householder

Ένας **μετασχηματισμός** (transformation), **πίνακας** (matrix) ή μία **ανάκλαση** (reflection) **Householder** διαστάσεων $n \times n$ γράφεται στη μορφή

$$\boldsymbol{H} = \boldsymbol{I}_n - 2\frac{\boldsymbol{h}\boldsymbol{h}^T}{\|\boldsymbol{h}\|^2} = \boldsymbol{I}_n - 2\frac{\boldsymbol{h}\boldsymbol{h}^T}{\boldsymbol{h}^T\boldsymbol{h}} = \boldsymbol{I}_n - b\boldsymbol{h}\boldsymbol{h}^T \tag{2.12}$$

όπου $\mathbf{h} \in \mathbb{R}^n$ καλείται το **διάνυσμα Householder**, $b = \frac{2}{\|\boldsymbol{h}\|^2} = \frac{2}{\boldsymbol{h}^T\boldsymbol{h}} \in \mathbb{R}$ και $\|\boldsymbol{h}\|^2 \neq 0$. Οι πίνακες Householder είναι συμμετρικοί και ορθογώνιοι, δηλαδή $\boldsymbol{H} = \boldsymbol{H}^T$ και $\boldsymbol{H}^2 = \boldsymbol{I}_n$. Οι πίνακες Householder είναι χρήσιμοι επειδή μπορούν να χρησιμοποιηθούν για να μηδενίσουμε συγκεκριμένα στοιχεία ενός διανύσματος ή ενός πίνακα (Golub & Van Loan, 2013; Seber & Lee, 2003; Kontoghiorghes, 2000).

Έστω $\boldsymbol{x} \in \mathbb{R}^n, \boldsymbol{x} \neq \mathbf{0}$. Ο στόχος μας είναι να βρούμε $\boldsymbol{h}$ ώστε $\boldsymbol{H}\boldsymbol{x} = a\boldsymbol{e}_1, a \in \mathbb{R}$ και $\boldsymbol{e}_1 \equiv \boldsymbol{I}_{:,1}$ είναι ένα $n \times 1$ διάνυσμα που περιέχει παντού μηδενικά εκτός στην πρώτη του θέση όπου περιέχει τη μονάδα.

Έχουμε ότι

$$\boldsymbol{H}\boldsymbol{x} = \left(\boldsymbol{I}_n - 2\frac{\boldsymbol{h}\boldsymbol{h}^T}{\boldsymbol{h}^T\boldsymbol{h}}\right)\boldsymbol{x} = \boldsymbol{x} - 2\frac{\boldsymbol{h}^T\boldsymbol{x}}{\boldsymbol{h}^T\boldsymbol{h}}\boldsymbol{h}$$

Θέτοντας

$$\boldsymbol{h} = \boldsymbol{x} + a\boldsymbol{e}_1$$

λαμβάνουμε

$$\boldsymbol{h}^T\boldsymbol{x} = \boldsymbol{x}^T\boldsymbol{x} + ax_1$$

και

$$\boldsymbol{h}^T\boldsymbol{h} = \boldsymbol{x}^T\boldsymbol{x} + 2ax_1 + a^2$$

και επομένως

$$\boldsymbol{Hx} = \left(1 - \frac{2(\boldsymbol{x}^T\boldsymbol{x} + ax_1)}{\boldsymbol{x}^T\boldsymbol{x} + 2ax_1 + a^2}\right)\boldsymbol{x} - 2a\frac{\boldsymbol{h}^T\boldsymbol{x}}{\boldsymbol{h}^T\boldsymbol{h}}\boldsymbol{e}_1$$

Για να μηδενίσουμε το συντελεστή του $\boldsymbol{x}$ στην παραπάνω σχέση θέτουμε $a = \pm\|\boldsymbol{x}\|_2$ και τότε λαμβάνουμε

$$\boldsymbol{h} = \boldsymbol{x} \pm \|\boldsymbol{x}\|_2\boldsymbol{e}_1$$

$$\Rightarrow \boldsymbol{Hx} = \mp\|\boldsymbol{x}\|_2\boldsymbol{e}_1 = \begin{pmatrix} \mp\|\boldsymbol{x}\|_2 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$$

Θέτοντας $h_1 = x_1 - \|\boldsymbol{x}\|_2$ λαμβάνουμε το ωραίο αποτέλεσμα ότι το $\boldsymbol{Hx}$ είναι θετικό πολλαπλάσιο του $\boldsymbol{e}_1$. Ωστόσο αυτό μπορεί να αποτελέσει πρόβλημα απώλειας σημαντικών ψηφίων στους υπολογισμούς μας αν το x_1 είναι θετικό και μεγάλο συγκριτικά με τα άλλα στοιχεία του $\boldsymbol{x}$ (Golub & Van Loan, 2013; Seber & Lee, 2003). Σε αυτήν την περίπτωση υπολογίζουμε το h_1 ως εξής

$$\mathrm{h}_1 = x_1 - \|\boldsymbol{x}\|_2 = \frac{x_1^2 - \|\boldsymbol{x}\|_2^2}{x_1 + \|\boldsymbol{x}\|_2} = \frac{-(x_2^2 + \cdots + x_n^2)}{x_1 + \|\boldsymbol{x}\|_2}$$

Στην πράξη είναι χρήσιμο να κανονικοποιήσουμε το διάνυσμα Householder ώστε $h_1 = 1$. Αυτό μας επιτρέπει να αποθηκεύσουμε τα στοιχεία $h_{2:n}$ στις θέσεις των στοιχείων του $\boldsymbol{x}$ που μηδενίζονται.

<u>Αλγόριθμος 4 – Υπολογισμός του διανύσματος Householder</u> (Golub & Van Loan, 2013)

Για $\boldsymbol{x} \in \mathbb{R}^n$, αυτή η συνάρτηση υπολογίζει το $\boldsymbol{h} \in \mathbb{R}^n$ με $h_1 = 1$ και $b \in \mathbb{R}$ ώστε ο $\boldsymbol{H} = \boldsymbol{I}_n - b\boldsymbol{h}\boldsymbol{h}^T \in \mathbb{R}^{n\times n}$ να είναι ορθογώνιος και $\boldsymbol{Hx} = \|\boldsymbol{x}\|_2\boldsymbol{e_1}$**.** Η συνάρτηση $length(\boldsymbol{x})$ επιστρέφει τη διάσταση ενός διανύσματος $\boldsymbol{x}$.

function $[\boldsymbol{h}, b] = \mathrm{house}(\boldsymbol{x})$

$n = length(\boldsymbol{x}),\ \ \sigma = \boldsymbol{x}_{2:n}^T\boldsymbol{x}_{2:n},\ \ \boldsymbol{h} = \begin{bmatrix} 1 \\ \boldsymbol{x}_{2:n} \end{bmatrix}$

if $\sigma = 0$ and $\boldsymbol{x}_1 \geq 0$

$b = 0$

elseif $\sigma = 0$ and $\boldsymbol{x}_1 < 0$

b=-2

else

$\mu = \sqrt{x_1^2 + \sigma}$

if $x_1 \leq 0$

$h_1 = x_1 - \mu$

(συνέχεια Αλγόριθμος 4)

else

$h_1 = -\sigma/(x_1 + \mu)$

end

$b = 2h_1^2/(\sigma + h_1^2)$

$\boldsymbol{h} = \boldsymbol{h}/h_1$

end

Ο παραπάνω αλγόριθμος εκτελεί περίπου $3n$ flops.

Θεωρούμε τον $n \times p$ πίνακα $\boldsymbol{X}$ με $rank(\boldsymbol{X}) = p$ ο οποίος έχει στήλες $\mathbf{x}^{(i)} \equiv \boldsymbol{X}_{:,i}$, $i = 1, \dots, p$. Εάν επιλέξουμε ένα πίνακα Householder $\boldsymbol{H}_1$ για να μηδενίσουμε όλα τα στοιχεία της στήλης $\boldsymbol{x}^{(1)}$, εκτός το πρώτο στοιχείο, τότε πολλαπλασιάζοντας από αριστερά τον $\boldsymbol{X}$ με τον $\boldsymbol{H}_1$ λαμβάνουμε

$$\boldsymbol{H}_1\boldsymbol{X} = \begin{pmatrix} r_{11} & r_{12} & \cdots & r_{1p} \\ 0 & & & \\ \vdots & & \boldsymbol{X_1} & \\ 0 & & & \end{pmatrix}$$

όπου $r_{11} = \left\|\boldsymbol{x}^{(1)}\right\|_2$. Γνωρίζουμε ότι $\left\|\boldsymbol{x}^{(1)}\right\|_2 \neq 0$ αφού διαφορετικά ο $\boldsymbol{X}$ θα είχε μία μηδενική στήλη και επομένως $rank(\boldsymbol{X}) < p$.

Στη συνέχεια θεωρούμε ένα πίνακα της μορφής

$$\boldsymbol{H_2} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & & & \\ \vdots & & \boldsymbol{K}_2 & \\ 0 & & & \end{pmatrix}$$

όπου $\boldsymbol{K}_2$ είναι ένας $(n-1) \times (n-1)$ πίνακας Householder τέτοιος ώστε να μηδενίζονται όλα τα στοιχεία, εκτός το πρώτο, της πρώτης στήλης του πίνακα $\boldsymbol{X_1}$. Τότε

$$\boldsymbol{H}_2\boldsymbol{H}_1\boldsymbol{X} = \begin{pmatrix} r_{11} & r_{12} & r_{13} & \cdots & r_{1p} \\ 0 & r_{22} & r_{23} & \cdots & r_{2p} \\ 0 & 0 & & & \\ \vdots & \vdots & & \boldsymbol{X}_2 & \\ 0 & 0 & & & \end{pmatrix}$$

όπου και πάλι έχουμε $r_{22} \neq 0$ διότι διαφορετικά η δεύτερη στήλη του $\boldsymbol{H_2H_1X}$ θα ήταν γραμμικώς εξαρτημένη με την πρώτη στήλη. Αυτό θα ερχόταν σε αντίθεση με το γεγονός ότι $rank(\boldsymbol{H_2H_1X}) = rank(\boldsymbol{X}) = p$ (Seber & Lee, 2003).

Συνεχίζοντας την παραπάνω διαδικασία λαμβάνουμε

$$\boldsymbol{H_pH_{p-1}} \cdots \boldsymbol{H_1X} = \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix}$$

όπου $\boldsymbol{R}$ είναι ένας $p \times p$ άνω τριγωνικός πίνακας με θετικά διαγώνια στοιχεία.

Θέτοντας $\boldsymbol{Q} = \left(\boldsymbol{H_p H_{p-1}} \cdots \boldsymbol{H_1}\right)' = \boldsymbol{H_1 H_2} \cdots \boldsymbol{H_p}$ λαμβάνουμε την παραγοντοποίηση QR

$$\boldsymbol{X} = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix}$$

εφόσον ο $\boldsymbol{Q}$ είναι ορθογώνιος, όντας γινόμενο ορθογώνιων πινάκων. Ο $\boldsymbol{Q}$ είναι ένας $n \times n$ πίνακας. Γράφοντας $\boldsymbol{Q} = \left(\boldsymbol{Q_p}, \boldsymbol{Q_{n-p}}\right)$, όπου ο $\boldsymbol{Q}_{n-p}$ είναι $n \times (n-p)$ λαμβάνουμε

$$\boldsymbol{X} = \boldsymbol{Q_p R}$$

δηλαδή την λεπτή παραγοντοποίηση της σχέσης (2.8).

Σημειώνουμε πως ο πίνακας Householder $\boldsymbol{H}$ δε χρειάζεται να σχηματιστεί. Στην πράξη, η συνάρτηση *house* εφαρμόζεται σε ένα τμήμα μίας στήλης ενός πίνακα και στη συνέχεια εφαρμόζουμε τον μετασχηματισμό $(\boldsymbol{I} - b\boldsymbol{h}^T\boldsymbol{h})$ σε ένα υποπίνακα.

<u>Αλγόριθμος 5 – Householder QR</u> (Golub & Van Loan, 2013)

Δεδομένου ενός πίνακα $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ με $n \geq p$, ο ακόλουθος αλγόριθμος βρίσκει πίνακες Householder $\boldsymbol{H_1}, \boldsymbol{H_2}, \ldots, \boldsymbol{H_p}$ ώστε αν $\boldsymbol{Q} = \boldsymbol{H_1 H_2} \cdots \boldsymbol{H_p}$ τότε ο $\boldsymbol{Q}^T\boldsymbol{X} = \boldsymbol{R}$ είναι άνω τριγωνικός. Το άνω τριγωνικό κομμάτι του $\boldsymbol{X}$ αντικαθίσταται με το άνω τριγωνικό κομμάτι του $\boldsymbol{R}$ και τα στοιχεία $j+1{:}\,n$ του j-οστού διανύσματος Householder αποθηκεύονται στον $\boldsymbol{X}_{j+1:n,j}$, $j < n$.

for j=1:n

[**h,**b] = house$\left(\boldsymbol{X}_{j:n\,,j}\right)$

$\boldsymbol{X}_{j:n,j:p} = (\boldsymbol{I} - b\boldsymbol{h}\boldsymbol{h}^T)\boldsymbol{X}_{j:n,j:p}$

if $j < n$

$\boldsymbol{X}_{j+1:n,j} = \boldsymbol{h}_{2:n-j+1}$

end

end

Ο παραπάνω αλγόριθμος εκτελεί $2p^2\left(n - \frac{p}{3}\right)$ flops.

Οι αποθηκεύσεις των διανυσμάτων Householder γίνονται με τον εξής τρόπο. Θεωρούμε ότι $\boldsymbol{X}$ είναι ένας 6×5 πίνακας. Αν το j-οστό διάνυσμα Householder είναι το

$$\boldsymbol{h}^{(j)} = \begin{matrix}j-1\left\{\right.\\ \\ \\ \\ \\ \\ \end{matrix}\begin{bmatrix}0\\ \vdots\\ 0\\ h_{j+1}^{(j)}\\ \vdots\\ h_n^{(j)}\end{bmatrix}$$

τότε με το πέρας του αλγορίθμου το αποτέλεσμα θα είναι

$$\boldsymbol{X} = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & r_{15} \\ h_2^{(1)} & r_{22} & r_{23} & r_{24} & r_{25} \\ h_3^{(1)} & h_3^{(2)} & r_{33} & r_{34} & r_{35} \\ h_4^{(1)} & h_4^{(2)} & h_4^{(3)} & r_{44} & r_{45} \\ h_5^{(1)} & h_5^{(2)} & h_5^{(3)} & h_5^{(4)} & r_{55} \\ h_6^{(1)} & h_6^{(2)} & h_6^{(3)} & h_6^{(4)} & h_6^{(5)} \end{bmatrix}$$

2.5.2.2 Μέθοδος Givens

Οι ανακλάσεις Householder είναι χρήσιμες όταν θέλουμε να μηδενίσουμε πολλά στοιχεία ενός πίνακα και πιο συγκεκριμένα όταν θέλουμε να μηδενίσουμε όλα τα στοιχεία ενός διανύσματος (στήλης ή γραμμής) εκτός το πρώτο στοιχείο. Ωστόσο, όταν θέλουμε να μηδενίσουμε πιο συγκεκριμένα στοιχεία, οι περιστροφές Givens μπορούν να μας φανούν ιδιαίτερα χρήσιμες.

Μία **περιστροφή Givens** (Givens rotation) είναι ένας $n \times n$ πίνακας της μορφής

$$\boldsymbol{G}_{ij} = \begin{bmatrix} & & & i & & j & & \\ 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 & \\ \vdots & \ddots & \vdots & & \vdots & & \vdots & \\ 0 & \cdots & c & \cdots & s & \cdots & 0 & i \\ \vdots & & \vdots & \ddots & \vdots & & \vdots & \\ 0 & \cdots & -s & \cdots & c & \cdots & 0 & j \\ \vdots & & \vdots & & \vdots & \ddots & \vdots & \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 & \end{bmatrix}$$

όπου $c = \cos(\theta)$ και $s = \sin(\theta)$ για κάποιο θ. Όλα τα στοιχεία που βρίσκονται εκτός της κύριας διαγωνίου, εκτός από τα s και $-s$, είναι μηδενικά. Οι περιστροφές Givens είναι ορθογώνιες, δηλαδή $\boldsymbol{G}_{ij}^T \boldsymbol{G}_{ij} = \boldsymbol{G}_{ij} \boldsymbol{G}_{ij}^T = \boldsymbol{I}_n$.

Πολλαπλασιάζοντας ένα διάνυσμα $\boldsymbol{x} \in \mathbb{R}^n$ από αριστερά με τον πίνακα $\boldsymbol{G}_{ij}$ ισοδυναμεί με μία αριστερόστροφη περιστροφή του διανύσματος κατά θ μοίρες ως προς το επίπεδο (i, j). Πιο συγκεκριμένα,

$$\boldsymbol{y} = \boldsymbol{G}_{ij} \boldsymbol{x}$$

όπου

$$y_k = \begin{cases} cx_i + sx_j, & k = i \\ cx_j - sx_i, & k = j \\ x_k, & k \neq i, j \end{cases}$$

Από την παραπάνω σχέση μπορούμε να επιλέξουμε τα c και s ώστε να μηδενίζεται το στοιχείο y_j δηλαδή επιλέγοντας

$$c = \frac{x_i}{\sqrt{x_i^2 + x_j^2}} = \frac{x_i}{t}, \quad s = \frac{x_j}{\sqrt{x_i^2 + x_j^2}} = \frac{x_j}{t}, \quad t = \left(x_i^2 + x_j^2\right)^{\frac{1}{2}} \neq 0 \tag{2.13}$$

Ομοίως, πολλαπλασιάζοντας ένα πίνακα $\boldsymbol{A} \in \mathbb{R}^{n\times p}$ με την περιστροφή $\boldsymbol{G}_{ij}$ επηρεάζονται μόνο τα στοιχεία στην i-οστή και j-οστή γραμμή του $\boldsymbol{A}$. Έστω οι πίνακες $\boldsymbol{A}, \widetilde{\boldsymbol{A}} \in \mathbb{R}^{n\times p}$ και η περιστροφή $\mathbf{G}_{ij}^{(k)}, 1 \le k \le p$ ώστε

$$\widetilde{\boldsymbol{A}} = \boldsymbol{G}_{ij}^{(k)}\boldsymbol{A}$$

δηλαδή

$$\widetilde{\boldsymbol{A}}_{p,:} = \begin{cases} c\boldsymbol{A}_{i,:} + s\boldsymbol{A}_{j,:}, & p = i \\ c\boldsymbol{A}_{j,:} - s\boldsymbol{A}_{i,:}, & p = j \\ \boldsymbol{A}_{p,:} & p = 1,\dots,n \text{ και } p \neq i,j \end{cases}$$

όπου ο $\mathbf{G}_{ij}^{(k)}$ είναι τέτοιος ώστε το στοιχείο $\tilde{a}_{j,k}$ να μηδενίζεται δηλαδή

$$\tilde{a}_{jk} = ca_{jk} - sa_{ik} = 0$$

Τότε αν $a_{ik} \neq 0$ και $a_{jk} \neq 0$, χρησιμοποιώντας την τριγωνομετρική σχέση $c^2 + s^2 = 1$ λαμβάνουμε ότι

$$c = a_{ik}/t, \quad s = a_{jk}/t \quad \text{και} \quad t^2 = a_{ik}^2 + a_{jk}^2$$

Στη συνέχεια παραθέτουμε ένα αλγόριθμο για τον αποδοτικό υπολογισμό των c, s και t.

<u>Αλγόριθμος 6</u> (Bjorck, 1996)

Ο παρακάτω αλγόριθμος υπολογίζει τα c, s και t ώστε $c\beta - s\alpha = 0$

$[c, s, t] = \text{givrot}(\alpha, \beta)$

if $\beta = 0$

$c = 1.0;\ \ s = 0.0;\ \ t = \alpha$

else if $|\beta| > |\alpha|$

$\tau = \alpha/\beta; \quad k = \sqrt{1+\tau^2};$

$s = 1/k; \quad c = \tau s; \quad t = k\beta;$

else

$\tau = \beta/\alpha; \quad k = \sqrt{1+\tau^2};$

$c = 1/k; \quad s = \tau c; \quad t = k\alpha;$

end

Χρησιμοποιώντας μία ακολουθία από περιστροφές Givens μπορούμε να υπολογίσουμε την παραγοντοποίηση QR της σχέσης (2.7). Μία τέτοια ακολουθία μπορεί να εκφραστεί ως

$$\boldsymbol{Q} = \left(\boldsymbol{G}_{n-1,n}^{(1)} \cdots \boldsymbol{G}_{1,2}^{(1)}\right)\left(\boldsymbol{G}_{n-1,n}^{(2)} \cdots \boldsymbol{G}_{2,3}^{(2)}\right) \cdots \left(\boldsymbol{G}_{n-1,n}^{(p)} \cdots \boldsymbol{G}_{p,p+1}^{(\boldsymbol{p})}\right)$$

Οι περιστροφές $\left(\boldsymbol{G}_{n-1,n}^{(i)} \cdots \boldsymbol{G}_{i,i+1}^{(i)}\right)$ μηδενίζουν τα στοιχεία $(n,i), \dots, (i+1,i)$ δηλαδή τα τελευταία $n-i$ στοιχεία της στήλης i και επιπρόσθετα διατηρούν τα στοιχεία που μηδενίζονται σε προηγούμενα στάδια. Η παραπάνω διαδικασία περιγράφεται στον παρακάτω Αλγόριθμο όπου ο πίνακας $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ σταδιακά αντικαθίσταται με τον $\begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix}$. Ο ορθογώνιος πίνακας $\boldsymbol{Q}$ δεν υπολογίζεται.

<u>Αλγόριθμος 7 – Givens QR</u> (Kontoghiorghes, 2000)

Ο παρακάτω αλγόριθμος υπολογίζει την παραγοντοποίηση QR για ένα πίνακα $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ χρησιμοποιώντας ακολουθία περιστροφών Givens βασισμένη σε στήλες (column-based).

for $i = 1,2, \dots, p$ **do**

  for $j = n, n-1, \dots, i+1$ **do**

    $\boldsymbol{X} \coloneqq \boldsymbol{G}_{j-1,j}^{(i)}\boldsymbol{X}$

  end for

end for

Στον παραπάνω Αλγόριθμο εκτελούνται συνολικά

$$\sum_{i=1}^{p}(n-i) = \frac{p(2n-p-1)}{2}$$

περιστροφές Givens.

Θα δώσουμε ένα παράδειγμα εφαρμογής του Αλγόριθμου. Για $\boldsymbol{X} \in \mathbb{R}^{4\times 3}$ το αποτέλεσμα του αλγόριθμου είναι ο πίνακας $\boldsymbol{X} \coloneqq \boldsymbol{G}_{3,4}^{(3)}\boldsymbol{G}_{2,3}^{(2)}\boldsymbol{G}_{3,4}^{(2)}\boldsymbol{G}_{1,2}^{(1)}\boldsymbol{G}_{2,3}^{(1)}\boldsymbol{G}_{3,4}^{(1)}\boldsymbol{X}$. Στη συνέχεια δείχνουμε τα βήματα που εκτελούνται όπου με $\boldsymbol{x}$ συμβολίζουμε τα δύο στοιχεία του εκάστοτε πίνακα που ορίζουν την κάθε περιστροφή Givens.

$$\boldsymbol{X} \equiv \begin{bmatrix} x & x & x\\ x & x & x\\ \boldsymbol{x} & x & x\\ \boldsymbol{x} & x & x\end{bmatrix} \xrightarrow{\boldsymbol{G}_{3,4}^{(1)}} \begin{bmatrix} x & x & x\\ x & x & x\\ \boldsymbol{x} & x & x\\ 0 & x & x\end{bmatrix} \xrightarrow{\boldsymbol{G}_{2,3}^{(1)}} \begin{bmatrix} \boldsymbol{x} & x & x\\ \boldsymbol{x} & x & x\\ 0 & x & x\\ 0 & x & x\end{bmatrix} \rightarrow$$

$$\xrightarrow{\boldsymbol{G}_{1,2}^{(1)}} \begin{bmatrix} x & x & x\\ 0 & x & x\\ 0 & \boldsymbol{x} & x\\ 0 & \boldsymbol{x} & x\end{bmatrix} \xrightarrow{\boldsymbol{G}_{3,4}^{(2)}} \begin{bmatrix} x & x & x\\ 0 & \boldsymbol{x} & x\\ 0 & \boldsymbol{x} & x\\ 0 & 0 & x\end{bmatrix} \xrightarrow{\boldsymbol{G}_{2,3}^{(2)}} \begin{bmatrix} x & x & x\\ 0 & x & x\\ 0 & 0 & \boldsymbol{x}\\ 0 & 0 & \boldsymbol{x}\end{bmatrix} \xrightarrow{\boldsymbol{G}_{3,4}^{(3)}} \begin{bmatrix} x & x & x\\ 0 & x & x\\ 0 & 0 & x\\ 0 & 0 & 0\end{bmatrix} \equiv \begin{pmatrix}\boldsymbol{R}\\ \boldsymbol{0}\end{pmatrix}$$

2.5.2.3 Αλγόριθμος Gram-Schmidt

Θα παρουσιάσουμε δύο εναλλακτικές μεθόδους που μπορούν να χρησιμοποιηθούν για να υπολογίσουμε την λεπτή παραγοντοποίηση QR, $\mathbf{X} = \boldsymbol{Q_1 R_1}$ (Kontoghiorghes, 2000).

Θέτουμε $\boldsymbol{Q}_1 \equiv \widetilde{\boldsymbol{Q}}$ ώστε

$$\boldsymbol{X}_{:,i} = \widetilde{\boldsymbol{Q}}\boldsymbol{R}_{:,i} = \boldsymbol{R}_{i,i}\widetilde{\boldsymbol{Q}}_{:,i} + \widetilde{\boldsymbol{Q}}_{:,1:i-1}\boldsymbol{R}_{1:i-1,i}$$

Από την παραπάνω σχέση λαμβάνουμε ότι

$$\widetilde{\boldsymbol{Q}}_{:,i} = \boldsymbol{b}/\boldsymbol{R}_{i,i} \tag{2.14}$$

όπου $\boldsymbol{b} = \boldsymbol{X}_{:,i} - \widetilde{\boldsymbol{Q}}_{:,1:i-1}\boldsymbol{R}_{1:i-1,i}$. Πολλαπλασιάζοντας από αριστερά τη σχέση (2.14) με $\widetilde{\boldsymbol{Q}}_{:,i}^T$ λαμβάνουμε $\boldsymbol{R}_{i,i} = \widetilde{\boldsymbol{Q}}_{:,i}^T\boldsymbol{b}$ και στη συνέχεια αντικαθιστώντας το $\widetilde{\boldsymbol{Q}}_{:,i}$ από τη σχέση (2.14) έχουμε ότι $\boldsymbol{R}_{i,i} = \|\boldsymbol{b}\|$. Τα υπόλοιπα μη-μηδενικά στοιχεία του $\boldsymbol{R}_{:,i}$ μπορούν να υπολογιστούν από τη σχέση $\boldsymbol{R}_{1:i-1,i} = \widetilde{\boldsymbol{Q}}_{:,1:i-1}^T\boldsymbol{X}_{:,i}$.

Ο παρακάτω Αλγόριθμος δίνει τα βήματα μεθόδου ορθογωνιοποίησης γνωστή ως ο **κλασικός αλγόριθμος Gram-Schmidt** (Classical Gram-Schmidt – CGS Algorithm). Στο k-οστό βήμα του αλγορίθμου υπολογίζουμε τις k-οστές στήλες των πινάκων $\boldsymbol{Q_1}$ και $\boldsymbol{R}$ και ο πίνακας $\boldsymbol{X}$ σταδιακά αντικαθίσταται με τον πίνακα $\boldsymbol{Q}_1 \equiv \widetilde{\boldsymbol{Q}}$**.**

<u>Αλγόριθμος 8 – Κλασικός Αλγόριθμος Gram-Schmidt (CGS)</u> (Kontoghiorghes, 2000)

Ο παρακάτω αλγόριθμος υπολογίζει την λεπτή παραγοντοποίηση QR για ένα πίνακα $\boldsymbol{X} \in \mathbb{R}^{n\times p}$

$\boldsymbol{R}_{1,1} \coloneqq \|\boldsymbol{X}_{:,1}\|_2$

$\boldsymbol{X}_{:,1} \coloneqq \boldsymbol{X}_{:,1}/\boldsymbol{R}_{1,1}$

for $i = 2, \dots, p$ **do**

$\boldsymbol{R}_{1:i-1,i} \coloneqq \boldsymbol{X}_{:,1:i-1}^T\boldsymbol{X}_{:,i}$

$\boldsymbol{b} \coloneqq \boldsymbol{X}_{:,i} - \boldsymbol{X}_{:,1:i-1}\boldsymbol{R}_{1:i-1,i}$

$\boldsymbol{R}_{i,i} \coloneqq \|\boldsymbol{b}\|_2$

$\boldsymbol{X}_{:,i} \coloneqq \boldsymbol{b}/\boldsymbol{R}_{i,i}$

end for

Ο αλγόριθμος CGS έχει κακές αριθμητικές ιδιότητες και αυτό μπορεί να έχει ως αποτέλεσμα την απώλεια ορθογωνιότητας για τις στήλες του $\widetilde{\boldsymbol{Q}}$**.** Η ευστάθεια της μεθόδου CGS μπορεί να βελτιωθεί χρησιμοποιώντας τον **τροποποιημένο αλγόριθμο Gram-Schmidt** (Modified Gram-Schmidt – MGS). Με τη μέθοδο MGS, σε κάθε βήμα του αλγορίθμου υπολογίζεται μία στήλη του $\widetilde{\boldsymbol{Q}}$ και μία γραμμή του $\boldsymbol{R}$ (Golub & Van Loan, 2013; Kontoghiorghes, 2000)**.**

Αλγόριθμος 9 – Τροποποιημένος Αλγόριθμος Gram-Schmidt (MGS)

(Golub & Van Loan, 2013; Kontoghiorghes, 2000)

Ο παρακάτω αλγόριθμος υπολογίζει την λεπτή παραγοντοποίηση QR για ένα πίνακα $\boldsymbol{X} \in \mathbb{R}^{n \times p}$

for $i = 1, \dots, p$ **do**

$\boldsymbol{R}_{i,i} \coloneqq \left\| \boldsymbol{X}_{:,i} \right\|_{\mathbf{2}}$

$\boldsymbol{X}_{:,i} \coloneqq \boldsymbol{X}_{:,i} / \boldsymbol{R}_{i,i}$

for $j = i + 1, \dots, p$ **do**

$\boldsymbol{R}_{i,j} \coloneqq \boldsymbol{X}_{:,i}^{T} \boldsymbol{X}_{:,j}$

$\boldsymbol{X}_{:,j} \coloneqq \boldsymbol{X}_{:,j} - \boldsymbol{R}_{i,j} \boldsymbol{X}_{:,i}$

end for

end for

Ο παραπάνω αλγόριθμος εκτελεί $2np^2$ flops.

2.6 Παραγοντοποίηση SVD

Υποθέτουμε ότι $\boldsymbol{X} \in \mathbb{R}^{n \times p}$ και $rank(\boldsymbol{X}) = r$. Από το Θεώρημα Α.10 υπάρχει $n \times n$ ορθογώνιος πίνακας $\boldsymbol{U}$ και $p \times p$ ορθογώνιος πίνακας $\boldsymbol{V}$ ώστε

$$\boldsymbol{U}'\boldsymbol{X}\boldsymbol{V} = \begin{pmatrix} \boldsymbol{\Sigma} \\ \mathbf{0} \end{pmatrix}$$

όπου

$$\boldsymbol{\Sigma} = \begin{pmatrix} \sigma_1 & 0 & 0 & \cdots & 0 \\ 0 & \sigma_2 & 0 & \cdots & 0 \\ \vdots & 0 & \ddots & & \vdots \\ \vdots & \vdots & & \ddots & 0 \\ 0 & 0 & \cdots & 0 & \sigma_p \end{pmatrix}$$

Οι ποσότητες σ_j είναι οι ιδιάζουσες τιμές του $\boldsymbol{X}$ και ικανοποιούν

$$\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > \sigma_{r+1} = \cdots = \sigma_p = 0$$

και είναι μοναδικές.

Η εξίσωση

$$\boldsymbol{X} = \boldsymbol{U}\begin{pmatrix}\boldsymbol{\Sigma}\\ \boldsymbol{0}\end{pmatrix}\boldsymbol{V}' \tag{2.15}$$

ονομάζεται **παραγοντοποίηση ιδιαζουσών τιμών** (Singular Value Decomposition – SVD) του $\boldsymbol{X}$. Μερικές φορές γράφεται στη **λεπτή μορφή**

$$\boldsymbol{X} = \boldsymbol{U}_p\boldsymbol{\Sigma}\boldsymbol{V}' \tag{2.16}$$

όπου $\boldsymbol{U}_p$ περιέχει τις πρώτες p στήλες του $\boldsymbol{U}$.

2.6.1 Προσαρμογή Μοντέλου Παλινδρόμησης

Έστω ότι $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ και $rank(\boldsymbol{X}) = p$ ώστε όλα τα διαγώνια στοιχεία του $\boldsymbol{\Sigma}$ να είναι θετικά. Αντικαθιστώντας την λεπτή παραγοντοποίηση $\boldsymbol{X} = \boldsymbol{U}_p\boldsymbol{\Sigma}\boldsymbol{V}'$ στις κανονικές εξισώσεις και χρησιμοποιώντας το γεγονός ότι $\boldsymbol{U}_p'\boldsymbol{U}_p = \boldsymbol{I}_p$ λαμβάνουμε

$$\boldsymbol{V}\boldsymbol{\Sigma}^{\boldsymbol{2}}\boldsymbol{V}'\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\boldsymbol{\Sigma}\boldsymbol{U}_{\boldsymbol{1}}'\boldsymbol{y}$$

Αφού οι $\boldsymbol{V}$ και $\boldsymbol{\Sigma}$ είναι αντιστρέψιμοι, τότε

$$\widehat{\boldsymbol{\beta}} = \boldsymbol{V}\boldsymbol{\Sigma}^{\boldsymbol{-1}}\boldsymbol{U}_p'\boldsymbol{y}$$

Για τον υπολογισμό του RSS χρησιμοποιούμε το διαμερισμό $\boldsymbol{U} = \left(\boldsymbol{U}_p, \boldsymbol{U}_{n-p}\right)$ και λαμβάνουμε (Seber & Lee, 2003)

$$RSS = \left\|\boldsymbol{U}_{n-p}'\boldsymbol{y}\right\|^2$$

Υπάρχουν διάφοροι αλγόριθμοι για τον υπολογισμό της παραγοντοποίησης SVD (Bjorck, 1996; Golub & Van Loan, 2013) ωστόσο δε θα αναφερθούμε σε αυτούς στην παρούσα εργασία.

ΚΕΦΑΛΑΙΟ 3

Εύρεση του Βέλτιστου Μοντέλου

Ας υποθέσουμε ότι έχουμε στη διάθεσή μας σύνολο δεδομένων που περιέχει k ανεξάρτητες μεταβλητές και καλούμαστε να εφαρμόσουμε πολλαπλή γραμμική παλινδρόμηση για να μελετήσουμε μία διαδικασία ή ένα φαινόμενο. Συχνά υποψιαζόμαστε ότι ένα υποσύνολο από αυτές τις μεταβλητές είναι αρκετό για να εξηγήσει το φαινόμενο που μελετούμε. Η διαδικασία της επιλογής αυτού του υποσυνόλου μας βοηθάει να εντοπίσουμε τις μεταβλητές εκείνες οι οποίες παίζουν σημαντικό ρόλο στην επεξήγηση του φαινομένου ή της διαδικασίας υπό μελέτη.

Διαισθητικά, φαίνεται λογικό να χρησιμοποιήσουμε όλη τη διαθέσιμη πληροφορία για το μοντέλο μας, συμπεριλαμβάνοντας στο μοντέλο μας όλες τις επεξηγηματικές μεταβλητές που έχουμε συλλέξει. Ωστόσο θα δείξουμε στη συνέχεια ότι συχνά λαμβάνουμε καλύτερα αποτελέσματα εάν χρησιμοποιήσουμε ένα υποσύνολο από τις k επεξηγηματικές μεταβλητές. Επιλέγοντας λιγότερες μεταβλητές αφενός το μοντέλο γίνεται πιο ευσταθές και αφετέρου εξοικονομούμε κόστος από τη συλλογή των δεδομένων.

Στη στατιστική μοντελοποίηση συναντάμε το γνωστό δίλημμα μεροληψίας-διασποράς (bias-variance tradeoff) (Hastie et al. , 2021). Ας θεωρήσουμε ότι έχουμε προσαρμόσει ένα μοντέλο χρησιμοποιώντας ένα σύνολο δεδομένων. Αν μας δοθεί μία καινούργια παρατήρηση $\boldsymbol{x_0} = (x_{00}, x_{01}, x_{02}, \dots, x_{0p})'$ τότε η αναμενόμενη τιμή του τετραγωνικού σφάλματος πρόβλεψης είναι

$$\begin{aligned} E[(y_0 - \widehat{y_0})^2] &= E[(y_0 - E(y_0) + E(y_0) - \widehat{y_0})^2] = E[(\varepsilon_0 + E(y_0) - \widehat{y_0})^2] \\ &= E(\varepsilon_0^2) + E[(E(y_0) - \widehat{y_0})^2] \\ &= Var(\varepsilon_0) + E\left[\left(\widehat{y_0} - E(y_0)\right)^2\right] \end{aligned}$$

Για την εκτιμήτρια $\hat{\theta}$ μίας παραμέτρου θ ορίζουμε το **Μέσο Τετραγωνικό Σφάλμα**

$$\begin{aligned} MT\Sigma(\hat{\theta}) &= E\left[\left(\hat{\theta} - \theta\right)^2\right] = E\left[\left(\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta\right)^2\right] \\ &= E\left[\left(\hat{\theta} - E(\hat{\theta})\right)^2\right] + 2E\left[\hat{\theta} - E(\hat{\theta})\right]E\left[E(\hat{\theta}) - \theta\right] + E\left[\left(E(\hat{\theta}) - \theta\right)^2\right] \\ &= Var(\hat{\theta}) + \left[E(\hat{\theta}) - \theta\right]^2 = Var(\hat{\theta}) + \left[Bias(\hat{\theta})\right]^2 \end{aligned}$$

Τότε

$$\begin{aligned} E[(y_0 - \widehat{y_0})^2] &= [E(\widehat{y_0}) - E(y_0)]^2 + E[(E(\widehat{y_0}) - \widehat{y_0})^2] \\ &= [Bias(\widehat{y_0})]^2 + Var(\widehat{y_0}) + Var(\varepsilon_0) \end{aligned} \tag{3.1}$$

Στην σχέση (3.1) η διασπορά του τυχαίου σφάλματος, $Var(\varepsilon_0)$, είναι μία ποσότητα που δεν μπορούμε να μειώσουμε (irreducible error) και αποτελεί μέτρο του θορύβου στα δεδομένα μας. Η διασπορά $Var(\widehat{y_0})$ αφορά τη μεταβολή του προσαρμοσμένου μοντέλου $\hat{y}$ όταν χρησιμοποιούμε διαφορετικά σύνολα δεδομένων για την προσαρμογή του μοντέλου. Ιδανικά, θέλουμε η μεταβλητότητα στην προσαρμογή του μοντέλου να είναι μικρή μεταξύ διαφορετικών συνόλων

δεδομένων. Η μεροληψία $Bias(\widehat{y_0})$ αφορά το σφάλμα που μπορεί να προκύψει όταν προσπαθούμε να προσεγγίσουμε μια αληθινή, συχνά περίπλοκη σχέση χρησιμοποιώντας ένα πιο απλό μοντέλο.

Η σχέση (3.1) μας λέει ότι για να ελαχιστοποιήσουμε το αναμενόμενο τετραγωνικό σφάλμα πρέπει το μοντέλο μας να πετυχαίνει ταυτόχρονα μικρή διασπορά και μεροληψία για την πρόβλεψη $\widehat{y_0}$. Το δίλημμα μεροληψίας-διασποράς προκύπτει από το γεγονός ότι αν προσθέσουμε περισσότερες μεταβλητές στο μοντέλο μας, οι οποίες είναι γραμμικά ανεξάρτητες με τις μεταβλητές που ήδη υπάρχουν στο μοντέλο, τότε η μεροληψία μειώνεται όμως ταυτόχρονα αυξάνεται η διασπορά (Seber & Lee, 2003). Επομένως είναι σημαντικό να μπορούμε να επιλέξουμε το βέλτιστο υποσύνολο μεταβλητών για το μοντέλο μας. Μπορούμε να δούμε τη διαδικασία επιλογής των μεταβλητών ως μία διαδικασία η οποία μειώνει τη διασπορά του μοντέλου μας, ως αποτέλεσμα της μείωσης της διαστατικότητας, δηλαδή του πλήθους των μεταβλητών του μοντέλου, που έχει ως αντίκτυπο την αύξηση της μεροληψίας, ως αποτέλεσμα της αφαίρεσης μεταβλητών από το μοντέλο οι οποίες προσδίδουν κάποια χρήσιμη πληροφορία. Στόχος λοιπόν της διαδικασίας αυτής είναι η εύρεση της χρυσής τομής, δηλαδή του υποσυνόλου μεταβλητών που ελαχιστοποιούν το αναμενόμενο τετραγωνικό σφάλμα πρόβλεψης.

Σε αυτό το κεφάλαιο υποθέτουμε ότι έχουμε στη διάθεσή μας ένα πλήθος k επεξηγηματικών μεταβλητών και ότι τα μοντέλα που προσαρμόζουμε συμπεριλαμβάνουν τον σταθερό όρο επομένως έχουν $p = k + 1$ παραμέτρους. Θα περιγράψουμε τρόπους επιλογής των μεταβλητών που θα χρησιμοποιηθούν για την προσαρμογή του μοντέλου.

3.1 Κριτήρια Επιλογής Μοντέλου

Για να μπορούμε να συγκρίνουμε πιθανά μοντέλα πρέπει να ορίσουμε κάποιο κριτήριο αξιολόγησης το οποίο θα μετρά την απόδοση του κάθε μοντέλου.

Τα κριτήρια που χρησιμοποιούνται συχνά στην πράξη είναι βασισμένα (Seber & Lee, 2003):

- σε κριτήρια καλής προσαρμογής (goodness-of-fit measures)
- στην εκτίμηση του σφάλματος πρόβλεψης
- στην εκτίμηση της διαφοράς μεταξύ της πραγματικής κατανομής για την μεταβλητή απόκρισης $\boldsymbol{Y}$ και της κατανομής που ορίζεται από το μοντέλο μας
- στην εκτίμηση posterior πιθανοτήτων

3.1.1 Κριτήρια Καλής Προσαρμογής

Στην ενότητα 1.1 ορίσαμε το άθροισμα τετραγώνων των υπολοίπων, RSS. Το RSS αποτελεί ένα κριτήριο καλής προσαρμογής ενός μοντέλου και στόχος μας είναι η ελαχιστοποίησή του. Αν συγκρίνουμε μοντέλα που περιέχουν το ίδιο πλήθος μεταβλητών, τότε θα προτιμήσουμε το μοντέλο με το μικρότερο RSS. Παρ' όλα αυτά το RSS δεν μπορεί να χρησιμοποιηθεί ευθέως για τη σύγκριση μοντέλων που περιέχουν διαφορετικό πλήθος μεταβλητών. Αν υποθέσουμε ότι έχουμε δύο σύνολα μεταβλητών S_1 και S_2 ώστε $S_1 \subset S_2$ και ότι οι επιπλέον μεταβλητές του συνόλου S_2 δεν είναι γραμμικός συνδυασμός των μεταβλητών του συνόλου S_1. Τότε $RSS_2 \leq RSS_1$, όπου RSS_1, RSS_2 αντιπροσωπεύει το RSS του μοντέλου με τις μεταβλητές S_1, S_2 αντίστοιχα (Rao et al., 2008). Δηλαδή,

προσθέτοντας μεταβλητές στο μοντέλο μας μειώνεται το RSS χωρίς κατ' ανάγκη αυτό να σημαίνει ότι έχει βελτιωθεί το μοντέλο.

Χρειαζόμαστε λοιπόν κάποιο κριτήριο που να διορθώνει το RSS λαμβάνοντας υπόψη το πλήθος των μεταβλητών του μοντέλου. Ένα τέτοιο κριτήριο είναι η **εκτίμηση της διασποράς σ^2 των τυχαίων σφαλμάτων**

$$S^2 = \frac{RSS}{n-p} \tag{3.2}$$

όπου n είναι το πλήθος των παρατηρήσεων στο σύνολο δεδομένων και p το πλήθος των συντελεστών του μοντέλου, συμπεριλαμβανομένου του σταθερού όρου. ΄

Ο συντελεστής προσδιορισμού R^2,

$$\begin{aligned} \mathrm{R}^2 &= \frac{\sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} = \frac{SSR}{SST} \\ &= 1 - \frac{RSS}{SST} = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \end{aligned}$$

που ορίσαμε στην ενότητα 1.1 μπορεί να χρησιμοποιηθεί εν μέρει ως ένα κριτήριο αξιολόγησης ενός γραμμικού μοντέλου. Όσο πιο κοντά στη μονάδα είναι η τιμή του R^2 τόσο πιο ισχυρή είναι η γραμμική σχέση εξάρτησης μεταξύ των μεταβλητών y και $\boldsymbol{x}$. Ωστόσο, εν γένει το R^2 δεν αποτελεί καλό κριτήριο αξιολόγησης ενός μοντέλου. Αρχικά, παρατηρούμε ότι αν έχουμε δύο παρατηρήσεις με ίδιες τιμές $\boldsymbol{x}$ αλλά διαφορετικές τιμές y, κάτι που συνήθως αναμένεται λόγω του τυχαίου σφάλματος ε, το R^2 είναι αδύνατο να πάρει τη μέγιστη τιμή του $(R^2 = 1)$ αφού προφανώς ένα μοντέλο γραμμικής παλινδρόμησης δεν μπορεί να περάσει και από τις δύο τιμές y, για τη συγκεκριμένη τιμή $\boldsymbol{x}$. Επομένως, όταν έχουμε πολλαπλές επαναλήψεις παρατηρήσεων για τις μεταβλητές $\boldsymbol{x}$ τότε η μέγιστη τιμή που μπορεί να πάρει το R^2 είναι πολύ χαμηλότερη της μονάδας και έτσι η αξιολόγηση της τιμής του γίνεται δύσκολη. Επιπρόσθετα, όταν το πλήθος των παρατηρήσεων δεν είναι πολύ μεγαλύτερο του αριθμού των μεταβλητών του μοντέλου τότε η τιμή του R^2 ενδέχεται να είναι αρκετά υψηλή, χωρίς ωστόσο το μοντέλο να είναι καλό. Τέλος, για μοντέλα τα οποία δεν συμπεριλαμβάνουν τον σταθερό όρο β_0, δηλαδή διέρχονται από την αρχή των αξόνων, ο συντελεστής προσδιορισμού δεν αποτελεί καλό μέτρο για την προσαρμογή του μοντέλου (Eisenhauer, 2003).

Όπως έχουμε αναφέρει παραπάνω, προσθέτοντας μεταβλητές σε κάποιο μοντέλο μειώνεται το άθροισμα τετραγώνων των υπολοίπων RSS ακόμα και όταν οι μεταβλητές που προστίθενται δεν είναι στατιστικά σημαντικές για την εξήγηση της μεταβλητής απόκρισης y. Η μείωση του RSS συνεπάγεται με την αύξηση του R^2. Για να αντιμετωπίσουμε αυτό το πρόβλημα θα ορίσουμε τον **διορθωμένο** ή τροποποιημένο **συντελεστή προσδιορισμού $\bar{\boldsymbol{R}}^{\boldsymbol{2}}$** (adjusted-$R^2$) :

$$\begin{aligned}\overline{\mathrm{R}}^2 &= 1 - \frac{MSE}{MST} \\ &= 1 - \frac{RSS/(n-p)}{SST/(n-1)} \\ &= 1 - \frac{n-1}{n-p}\frac{RSS}{SST} \\ &= 1 - \frac{n-1}{n-p}(1-R^2)\end{aligned} \tag{3.3}$$

όπου $p = K + 1$ είναι το πλήθος των παραμέτρων του μοντέλου, $MSE = \frac{RSS}{n-p} = S^2$ είναι η εκτιμήτρια στη σ^2 που ορίσαμε στη σχέση (3.2) και $MST = \frac{SST}{n-1} = \frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n-1}$ είναι η **εκτιμήτρια της διασποράς του δείγματος τιμών y_i**.

Ο διορθωμένος συντελεστής προσδιορισμού προτιμάται από τον συντελεστή προσδιορισμού R^2 λόγω της σημαντικής ιδιότητάς του να αυξάνεται μόνο όταν προστίθεται στο μοντέλο κάποια μεταβλητή που πράγματι βελτιώνει το μοντέλο. Μπορούμε να δούμε την ιδιότητα αυτή μέσω του στατιστικού ελέγχου F που θα ορίσουμε ακολούθως.

Υποθέτουμε ότι έχουμε δύο μοντέλα M_0, M_1 ώστε $M_0 \subset M_1$ (το M_0 είναι **εμφωλευμένο** (nested) στο M_1), δηλαδή όλες οι μεταβλητές του μοντέλου M_0 συμπεριλαμβάνονται στο M_1 και επιπρόσθετα το M_1 περιλαμβάνει επιπλέον μεταβλητές οι οποίες δεν είναι γραμμικός συνδυασμός των μεταβλητών του M_0. Υποθέτουμε επίσης ότι το μοντέλο M_1 περιέχει k μεταβλητές, δηλαδή $p = k + 1$ παραμέτρους, συμπεριλαμβάνοντας και τον σταθερό όρο, ενώ το μοντέλο M_0 περιέχει $k' = k - q$ μεταβλητές, δηλαδή $p' = k - q + 1 = p - q$ παραμέτρους. Χωρίς βλάβη της γενικότητας θα υποθέσουμε ότι οι q μεταβλητές που περιλαμβάνονται στο M_1 αλλά δεν περιλαμβάνονται στο M_0 έχουν συντελεστές με δείκτη j, $0 \le k' < j \le k$. Μπορούμε να συγκρίνουμε τα δύο μοντέλα με τον ακόλουθο έλεγχο υποθέσεων

$$H_0: \beta_j = 0, \quad \forall\ 0 \le k' < j \le k$$
$$H_1: \beta_i \ \text{χωρίς περιορισμούς}, \ \forall\ 1 \le i \le k$$

Η ελεγχοσυνάρτηση για τις παραπάνω υποθέσεις είναι

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)} \sim F_{q,(n-p)} \tag{3.4}$$

όπου RSS_0, RSS_1 είναι το άθροισμα τετραγώνων των υπολοίπων για τα μοντέλα M_0, M_1 αντίστοιχα.

Επιστρέφουμε τώρα στο συντελεστή προσδιορισμού. Αφού $M_0 \subset M_1$, τότε $RSS_1 \leq RSS_0$ επομένως $R_1^2 \geq R_0^2$. Για τον διορθωμένο συντελεστή προσδιορισμού θα ισχύει

$$\bar{R}_1^2 > \bar{R}_0^2 \Rightarrow 1 - \frac{\frac{RSS_1}{n-p}}{\frac{SST}{n-1}} > 1 - \frac{\frac{RSS_0}{n-p'}}{\frac{SST}{n-1}}$$

$$\Rightarrow \frac{RSS_1}{n-p} < \frac{RSS_0}{n-p'}$$

$$\Rightarrow \frac{n-p'}{n-p} < \frac{RSS_0}{RSS_1} = \frac{RSS_0 - RSS_1}{RSS_1} + 1$$

$$\Rightarrow \frac{q}{n-p} < \frac{RSS_0 - RSS_1}{RSS_1}$$

$$\Rightarrow \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)} > 1$$

$$\Rightarrow F > 1$$

Επομένως το $\bar{R}^2$ αυξάνεται μόνο στην περίπτωση όπου $F > 1$ για τον παραπάνω έλεγχο υποθέσεων.

3.1.2 Κριτήρια Βασισμένα στο Σφάλμα Πρόβλεψης

Ακόμα ένα κριτήριο αξιολόγησης ενός μοντέλου είναι η στατιστική συνάρτηση C_p-Mallows (Mallows, 1973). Το κριτήριο αυτό βασίζεται στην ακρίβεια πρόβλεψης του μοντέλου και πιο συγκεκριμένα στο Μέσο Τετραγωνικό Σφάλμα (ΜΤΣ) που ορίσαμε στην αρχή αυτού του κεφαλαίου. Υποθέτουμε ότι έχουμε στη διάθεσή μας k επεξηγηματικές μεταβλητές και ότι έχουμε προσαρμόσει ένα μοντέλο με p παραμέτρους, συμπεριλαμβανομένου και του σταθερού όρου, χρησιμοποιώντας $p-1 < k$ μεταβλητές από αυτές που έχουμε στη διάθεσή μας. Έχοντας επίσης ένα σύνολο δεδομένων με n παρατηρήσεις, τότε το Μέσο Τετραγωνικό Σφάλμα της πρόβλεψης είναι

$$\begin{aligned}\sum_{i=1}^{n} MT\Sigma(\hat{y}_i) &= \sum_{i=1}^{n} E[\hat{y}_i - E(y_i)]^2 \\ &= \sum_{i=1}^{n} [E(y_i) - E(\hat{y}_i)]^2 + \sum_{i=1}^{n} Var(\hat{y}_i) \\ &= \sum_{i=1}^{n} bias^2 + \sum_{i=1}^{n} Var(\hat{y}_i)\end{aligned}$$

Για τον δεύτερο όρο ισχύει

$$\sum_{i=1}^{n} Var(\hat{y}_i) = tr\big(Var(\hat{\boldsymbol{y}})\big) = tr\big(Var(\boldsymbol{H}\boldsymbol{y})\big)$$
$$= tr(\boldsymbol{H}Var(\boldsymbol{y})\boldsymbol{H}') = tr(\boldsymbol{H}(\sigma^2\boldsymbol{I})\boldsymbol{H}')$$
$$= \sigma^2 tr\big(\boldsymbol{H^2}\big) = \sigma^2 tr(\boldsymbol{H}) = \sigma^2 p$$

Για τον πρώτο όρο έχουμε ότι (Καρώνη & Οικονόμου, 2017)

$$\sum_{i=1}^{n}[E(y_i) - E(\hat{y}_i)]^2 = \sum_{i=1}^{n} bias^2 = E\left[\sum_{i=1}^{n}(y_i - \hat{y}_i)^2\right] - (n-p)\sigma^2$$
$$= E\big[RSS_p\big] - (n-p)\sigma^2$$

Επομένως

$$\sum_{i=1}^{n} MT\Sigma(\hat{y}_i) = E\big[RSS_p\big] - (n-p)\sigma^2 + \sigma^2 p = E\big[RSS_p\big] + (2p-n)\sigma^2$$

Διαιρώντας την παραπάνω σχέση με το σ^2 λαμβάνουμε την ποσότητα

$$\Gamma_p = \frac{E\big[RSS_p\big]}{\sigma^2} + (2p-n)$$

Η ποσότητα Γ_p εκτιμάται από την

$$C_p = \frac{RSS_p}{\hat{\sigma}^2} + (2p-n) \qquad (3.5)$$

Στη σχέση (3.5) το $E\big[RSS_p\big]$ έχει αντικατασταθεί από το RSS_p του μοντέλου που έχουμε προσαρμόσει στο σύνολο δεδομένων μας. Να σημειωθεί ότι εάν χρησιμοποιήσουμε το ίδιο σύνολο δεδομένων για την προσαρμογή του μοντέλου και για την διαδικασία επιλογής μοντέλου τότε το RSS_p ενδέχεται να είναι μικρό εξαιτίας του γεγονότος ότι το μοντέλο έχει προσαρμοστεί στα δεδομένα αυτά. Το σ^2 έχει αντικατασταθεί με την αμερόληπτη εκτιμήτρια $\hat{\sigma}^2 = \frac{RSS_{p'}}{n-p'}$ όπου για τον υπολογισμό της εκτιμήτριας συνήθως χρησιμοποιείται το πλήρες μοντέλο το οποίο περιέχει $p' = k + 1$ παραμέτρους.

Αν για το υπό εξέταση μοντέλο έχουμε αμελητέα μεροληψία, δηλαδή $Bias \approx 0$ τότε

$$\sum_{i=1}^{n}[E(y_i) - E(\hat{y}_i)]^2 = \sum_{i=1}^{n} bias^2 \approx 0$$

και

$$E\big[C_p|\ Bias = 0\big] = \frac{E\big[RSS_p\big]}{\sigma^2} + (2p-n) = \frac{(n-p)\sigma^2}{\sigma^2} + (2p-n) = p$$

Επομένως το βέλτιστο μοντέλο είναι εκείνο για το οποίο ισχύει $C_p \approx p$. Στην περίπτωση που έχουμε περισσότερα του ενός μοντέλα με $C_p \approx p$ τότε επιλέγουμε εκείνο με το μικρότερο p, δηλαδή τις λιγότερες μεταβλητές. Να σημειώσουμε ότι στην περίπτωση που εξετάζουμε το πλήρες μοντέλο τότε πάντα θα ισχύει $C_{p'} = p'$ χωρίς όμως αυτό να σημαίνει απαραίτητα ότι αυτό είναι το βέλτιστο μοντέλο.

3.1.3 Κριτήρια Βασισμένα στην Απόκλιση Κατανομών

Ακόμη μία προσέγγιση για την εύρεση του βέλτιστου μοντέλου είναι η χρήση κριτηρίων όπως το AIC (Akaike's Information Criterion) (Akaike, 1973). Το AIC είναι βασισμένο στη Θεωρία Πληροφοριών και στην Θεωρία Πιθανοφάνειας. Συγκεκριμένα, βασίζεται στην απόκλιση Kullback-Leibler (Kullback & Leibler, 1951) η οποία εκφράζει ένα είδος "απόστασης" μεταξύ δύο συναρτήσεων. Στην περίπτωσή μας, εξετάζεται η απόσταση μεταξύ του υπό εξέταση μοντέλου και του αληθινού μοντέλου που μελετούμε. Το AIC ορίζεται από τη σχέση

$$AIC = 2p - 2\ln L \tag{3.6}$$

όπου p είναι το πλήθος των παραμέτρων του υπό εξέταση μοντέλου και L είναι η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμηθέν μοντέλο. Το βέλτιστο μοντέλο θεωρείται εκείνο με τη μικρότερη τιμή AIC.

Αξίζει να παρατηρήσουμε ότι προσθέτοντας μεταβλητές στο μοντέλο βελτιώνουμε την προσαρμογή του, κάτι που σημαίνει ότι αυξάνεται ο όρος $\ln L$ μειώνοντας την τιμή του AIC. Ωστόσο, ο πρώτος όρος $2p$ λειτουργεί ως αντίβαρο στην μείωση αυτή ή ως μία ποινή και αυξάνεται με την εισαγωγή επιπλέον μεταβλητών στο μοντέλο. Επομένως η τιμή του AIC μειώνεται μόνο όταν προσθέσουμε στο μοντέλο μεταβλητές οι οποίες βελτιώνουν επαρκώς το μοντέλο μας.

Στην περίπτωση του πολλαπλού γραμμικού μοντέλου οι εκτιμήτριες μέγιστης πιθανοφάνειας των παραμέτρων του μοντέλου μας είναι

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

$$\hat{\sigma}^2 = \frac{RSS_p}{n}$$

επομένως το AIC γράφεται

$$AIC = n\left[\ln\left(\frac{2\pi RSS_p}{n}\right) + 1\right] + 2(p+1)$$

3.1.4 Κριτήρια Βασισμένα στην Μεγιστοποίηση Posterior Πιθανοτήτων

Το BIC (Bayesian Information Criterion) (Schwartz, 1978) αποτελεί ακόμα ένα κριτήριο επιλογής του βέλτιστου μοντέλου και είναι βασισμένο στη μεγιστοποίηση εκ των υστέρων (posterior) πιθανοτήτων. Η μορφή και η χρήση του είναι παρόμοια με το κριτήριο AIC, με το BIC να θέτει

μεγαλύτερη ποινή σε μοντέλα με περισσότερες μεταβλητές, σε σύγκριση με το AIC. Το BIC ορίζεται από τη σχέση

$$\mathrm{B}IC = p \ln n - 2 \ln L \tag{3.7}$$

όπου p είναι το πλήθος των παραμέτρων του υπό εξέταση μοντέλου και L είναι η μεγιστοποιημένη τιμή της συνάρτησης πιθανοφάνειας για το εκτιμηθέν μοντέλο. Το βέλτιστο μοντέλο θεωρείται εκείνο με τη μικρότερη τιμή BIC.

Στην περίπτωση του πολλαπλού γραμμικού μοντέλου το BIC παίρνει τη μορφή

$$BIC = n\left[\ln\left(\frac{2\pi RSS_p}{n}\right) + 1\right] + (p+1)\ln n$$

3.2 Διαδικασίες Επιλογής Μοντέλου σε Βήματα

Ένας τρόπος εύρεσης ενός κατάλληλου μοντέλου είναι με τη χρήση μεθόδων οι οποίες επιλέγουν ένα καλό μοντέλο ακολουθώντας μία διαδικασία σε βήματα. Κάθε βήμα της διαδικασίας βασίζεται στον έλεγχο F

$$F = \frac{(RSS_0 - RSS_1)/q}{RSS_1/(n-p)} \sim F_{q,(n-p)}$$

που περιγράψαμε στην ενότητα 3.1.1 και αφορά την πρόσθεση ή αφαίρεση q μεταβλητών ενός μοντέλου. Συγκεκριμένα, σε κάθε βήμα εκτελείται ο έλεγχος για την πρόσθεση ή αφαίρεση μίας επεξηγηματικής μεταβλητής ($q = 1$) χρησιμοποιώντας τον έλεγχο

$$\begin{aligned} F &= \frac{RSS_0 - RSS_1}{RSS_1/(n-p')} \sim F_{1,(n-p')} \\ &= \frac{RSS_0 - RSS_1}{RSS_1/(n-k'-1)} \sim F_{1,(n-k'-1)} \end{aligned}$$

για να συγκρίνουμε τα μοντέλα M_0 και M_1, με $M_0 \subset M_1$, τα οποία περιέχουν $k'-1$ και k' μεταβλητές, αντίστοιχα και $p' = k' + 1$ είναι οι παράμετροι του μοντέλου M_1. Αν εξετάζουμε την αφαίρεση μίας μεταβλητής και αν η διαφορά $RSS_0 - RSS_1$ είναι μικρή τότε η μεταβλητή αυτή δε θεωρείται σημαντική για το μοντέλο και πρέπει να αφαιρεθεί. Αντιθέτως, αν η διαφορά είναι μεγάλη τότε η μεταβλητή πρέπει να παραμείνει στο μοντέλο. Στην περίπτωση που εξετάζουμε την πρόσθεση μιας μεταβλητής και αν η διαφορά $RSS_0 - RSS_1$ είναι μικρή τότε η μεταβλητή αυτή δε χρειάζεται να προστεθεί στο μοντέλο, ενώ εάν η διαφορά είναι μεγάλη τότε πρέπει να προσθέσουμε τη μεταβλητή στο μοντέλο.

Διαδικασία Διαδοχικής Πρόσθεσης

Στη διαδικασία της **διαδοχικής πρόσθεσης** ή της **προς τα εμπρός επιλογής** (forward selection) ξεκινάμε με το μοντέλο που περιέχει μόνο το σταθερό όρο, δηλαδή το μοντέλο $y = \beta_0$. Σε κάθε βήμα προσθέτουμε τη μεταβλητή εκείνη, που δεν είναι ήδη στο μοντέλο, η οποία μας δίνει τη μεγαλύτερη

τιμή για την ελεγχοσυνάρτηση F, δηλαδή τη μεταβλητή εκείνη η οποία έχει τη μεγαλύτερη στατιστική σημαντικότητα για το στάδιο της διαδικασίας που βρισκόμαστε. Η μεταβλητή αυτή θα δίνει τη μεγαλύτερη μείωση στο RSS του μοντέλου. Εάν η μείωση αυτή είναι στατιστικά σημαντική, δηλαδή η τιμή της F είναι στατιστικά σημαντική, τότε προσθέτουμε τη μεταβλητή στο μοντέλο, ξαναπροσαρμόζουμε το μοντέλο συμπεριλαμβάνοντας τη μεταβλητή αυτή και συνεχίζουμε την αναζήτηση άλλης υποψήφιας μεταβλητής που θα μπορούσε να εισαχθεί στο μοντέλο. Η διαδικασία σταματάει όταν η τιμή της F δεν είναι στατιστικά σημαντική για καμία υποψήφια μεταβλητή, δηλαδή όταν είναι μικρότερη από μία προκαθορισμένη τιμή F_{IN}.

Διαδικασία Διαδοχικής Αφαίρεσης

Στη διαδικασία **διαδοχικής αφαίρεσης** ή **προς τα πίσω απαλοιφής** (backward elimination) ξεκινάμε με το μοντέλο που περιέχει όλες τις διαθέσιμες μεταβλητές. Σε κάθε βήμα αφαιρούμε τη μεταβλητή εκείνη που αντιστοιχεί στη μικρότερη αύξηση του RSS, εφόσον η τιμής της ελεγχοσυνάρτησης F είναι στατιστικά μη σημαντική. Αυτή θα είναι η μεταβλητή που έχει τη μικρότερη συμβολή στο μοντέλο. Εάν βρεθεί μία τέτοια μεταβλητή, τότε την αφαιρούμε από το μοντέλο, ξαναπροσαρμόζουμε το μοντέλο εξαιρώντας αυτήν την μεταβλητή και συνεχίζουμε την αναζήτηση άλλης υποψήφιας προς αφαίρεση μεταβλητής. Η διαδικασία σταματάει όταν η αφαίρεση μίας οποιασδήποτε μεταβλητής είναι στατιστικά σημαντική, δηλαδή για όλες τις μεταβλητές η ελεγχοσυνάρτηση F παίρνει τιμές $F \geq F_{OUT}$, για κάποια προκαθορισμένη τιμή F_{OUT}.

Διαδικασία Εμπρός-Πίσω Επιλογής

Ένα αρνητικό χαρακτηριστικό των δύο παραπάνω διαδικασιών είναι ότι στην περίπτωση της διαδοχικής πρόσθεσης, εφόσον μία μεταβλητή έχει προστεθεί στο μοντέλο τότε είναι αδύνατο να αφαιρεθεί και στην περίπτωση της διαδοχικής αφαίρεσης, εφόσον μία μεταβλητή έχει αφαιρεθεί από το μοντέλο τότε είναι αδύνατο να ξαναπροστεθεί. Αυτό είναι αρνητικό διότι, στην περίπτωση της διαδοχικής πρόσθεσης, υπάρχει περίπτωση η πρόσθεση μίας νέας μεταβλητής στο μοντέλο να οδηγεί στην εξασθένηση της σημαντικότητας κάποιας άλλης μεταβλητής που είχε εισαχθεί νωρίτερα στο μοντέλο. Η διαδικασία της **εμπρός-πίσω επιλογής** (stepwise selection) συνδυάζει τις διαδικασίες διαδοχικής αφαίρεσης και πρόσθεσης. Αρχίζοντας από το μοντέλο που περιέχει μόνο το σταθερό όρο, η διαδικασία αυτή εκτελεί διαδοχικά ένα βήμα της διαδικασίας διαδοχικής πρόσθεσης και στη συνέχει ένα βήμα της διαδικασίας διαδοχικής αφαίρεσης. Δηλαδή, γίνεται έλεγχος εάν υπάρχει μεταβλητή που πρέπει να προστεθεί στο μοντέλο και εάν βρεθεί μία τέτοια μεταβλητή τότε προστίθεται στο μοντέλο και στης συνέχεια εκτελείται έλεγχος για το εάν πρέπει να αφαιρεθεί κάποια μεταβλητή από το τρέχον μοντέλο. Η διαδικασία αυτή σταματάει όταν δεν υπάρχουν άλλες στατιστικά σημαντικές μεταβλητές για να προσθέσουμε στο μοντέλο. Δεδομένου ότι $F_{OUT} \leq F_{IN}$, η διαδικασία της εμπρός-πίσω επιλογής θα τερματίσει (Seber & Lee, 2003).

Η χρήση των παραπάνω διαδικασιών βημάτων έχει μερικά μειονεκτήματα (Seber & Lee, 2003). Αρχικά, το αποτέλεσμα των διαδικασιών αυτών είναι ένα μόνο μοντέλο ενώ στην πράξη μπορεί να υπάρχουν πολλά μοντέλα με παρόμοια αποδοτικότητα τα οποία δεν έχουν εξεταστεί από τις διαδικασίες. Δεν υπάρχει κάποια εγγύηση ότι το μοντέλο που θα μας επιστρέψουν αυτές οι διαδικασίες θα είναι το ίδιο μοντέλο που θα λαμβάναμε στην περίπτωση που αξιολογούσαμε όλα τα δυνατά μοντέλα βάσει κάποιου κριτηρίου αξιολόγησης. Επιπρόσθετα, αφού σε κάθε βήμα η επιλογή βασίζεται σε ένα έλεγχο F, μπορεί αρχικά να φαίνεται ότι οι διαδικασίες θα βρουν το καλύτερο μοντέλο με μια πιθανότητα. Ωστόσο, σε κάθε βήμα η τιμή της ελεγχοσυνάρτησης F είναι το μέγιστο

ή ελάχιστο ενός συνόλου συσχετισμένων στατιστικών, όπου το κάθε ένα από αυτά σχετίζεται με τα προηγούμενα βήματα της διαδικασίας, με περίπλοκους τρόπους. Η πιθανότητα να επιλέξουμε το σωστό μοντέλο παραμένει άγνωστη.

3.3 Όλα τα Πιθανά Μοντέλα

Μία προφανής προσέγγιση για την επιλογή του βέλτιστου μοντέλου θα ήταν να προσαρμόσουμε όλα τα πιθανά μοντέλα και να επιλέξουμε το καλύτερο από αυτά βάσει κάποιου κριτηρίου. Με k μεταβλητές στη διάθεσή μας, μπορούμε να προσαρμόσουμε $2^k - 1$ μοντέλα εξαιρώντας το μοντέλο που δεν περιέχει καμία από αυτές τις μεταβλητές. Το πλήθος των πιθανών μοντέλων αυξάνεται εκθετικά με το k, κάτι που καθιστά την προσαρμογή όλων αυτών των μοντέλων σχεδόν αδύνατη, λόγω του υπολογιστικού κόστους, ακόμα και για σχετικά μικρές τιμές του k. Γι' αυτόν τον λόγο πολλοί ερευνητές έχουν προσπαθήσει να μειώσουν το υπολογιστικό κόστος που χρειάζεται για την προσαρμογή κάθε μοντέλου καθώς επίσης και να δημιουργήσουν διαδικασίες εύρεσης του βέλτιστου υποσυνόλου μεταβλητών χωρίς να χρειαστεί η προσαρμογή όλων των πιθανών μοντέλων.

3.3.1 Προσαρμογή Όλων των Πιθανών Μοντέλων

3.3.1.1 Χρησιμοποιώντας τη μέθοδο SWEEP

Στη συνέχεια θα περιγράψουμε μία διαδικασία με την οποία μπορούμε να προσαρμόσουμε όλα τα πιθανά μοντέλα βασισμένη στον αλγόριθμο SWEEP που ορίσαμε στην Ενότητα 2.3. Θα θεωρήσουμε ότι έχουμε στη διάθεσή μας K επεξηγηματικές μεταβλητές και ότι κάθε μοντέλο που θα προσαρμόζουμε θα περιέχει και το σταθερό όρο. Υπάρχουν 2^k πιθανά μοντέλα συμπεριλαμβανομένου του μοντέλου που περιέχει μόνο το σταθερό όρο. Υπενθυμίζουμε ότι εκτελώντας ένα SWEEP σε μία γραμμή του $(p+1) \times (p+1) = (k+2) \times (k+2)$ πίνακα (2.4)

$$\boldsymbol{X}_{\boldsymbol{A}}'\boldsymbol{X}_{\boldsymbol{A}} = \begin{bmatrix} \boldsymbol{X}'\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{y} \\ \boldsymbol{y}'\boldsymbol{X} & \boldsymbol{y}'\boldsymbol{y} \end{bmatrix}$$

προσθέτουμε τη μεταβλητή που αντιστοιχεί σε αυτή τη γραμμή (και στήλη) στο μοντέλο, εάν δεν βρίσκεται ήδη μέσα, ενώ στην περίπτωση που η μεταβλητή αυτή ανήκει στο μοντέλο, η εκτέλεση του SWEEP θα την αφαιρέσει.

Αρχικά, εκτελώντας ένα SWEEP στην πρώτη γραμμή του πίνακα έχει ως αποτέλεσμα να προσθέσουμε το σταθερό όρο στο μοντέλο και λαμβάνουμε τον πίνακα

$$\begin{bmatrix} \frac{1}{n} & \bar{\boldsymbol{X}}' & \bar{y} \\ -\bar{\boldsymbol{X}} & \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{X}}'\tilde{\boldsymbol{y}} \\ -\bar{y} & \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{X}} & \tilde{\boldsymbol{y}}'\tilde{\boldsymbol{y}} \end{bmatrix} \tag{3.8}$$

όπου

- $\overline{\boldsymbol{X}}$ είναι ένα $p \times 1$ διάνυσμα με $(\overline{\boldsymbol{X}})_j = \bar{x}_j = \frac{1}{n}\sum_{i=1}^{n} x_{ij}$, δηλαδή περιέχει τις μέσες τιμές των μεταβλητών (στηλών) του πίνακα $\boldsymbol{X}$.
- $\bar{y} = \sum_{i=1}^{n} y_i$
- $\widetilde{\boldsymbol{X}}_{n\times K}$ και $\widetilde{\boldsymbol{y}}_{n\times 1}$ είναι οι «κεντραρισμένες» εκδοχές των $\boldsymbol{X}$ και $\boldsymbol{y}$ αντίστοιχα, δηλαδή $\left(\widetilde{\boldsymbol{X}}\right)_{ij} = x_{ij} - \bar{x}_j \ \ , \ j = 1, \dots, K$ και $(\widetilde{\boldsymbol{y}})_i = y_i - \bar{y}$

Εφόσον θεωρούμε ότι όλα τα μοντέλα θα περιέχουν το σταθερό όρο, τότε η πρώτη γραμμή και στήλη του πίνακα (3.8) μπορεί να παραλειφθεί (Seber & Lee, 2003). Τα υπόλοιπα $2^k - 1$ μοντέλα μπορούν να υπολογιστούν εφαρμόζοντας μία σειρά από SWEEPS στις γραμμές του $(k+1) \times (k+1)$ πίνακα

$$\begin{bmatrix} \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{X}} & \widetilde{\boldsymbol{X}}'\widetilde{\boldsymbol{y}} \\ \widetilde{\boldsymbol{y}}'\widetilde{\boldsymbol{X}} & \widetilde{\boldsymbol{y}}'\widetilde{\boldsymbol{y}} \end{bmatrix} \tag{3.9}$$

Είναι σημαντικό να μπορέσουμε να υπολογίσουμε όλα τα πιθανά μοντέλα χρησιμοποιώντας όσο το δυνατόν λιγότερα SWEEPS γίνεται ώστε να εξοικονομήσουμε υπολογιστικό κόστος. Μία τέτοια ακολουθία SWEEPS πρέπει να αποφεύγει να υπολογίζει το ίδιο μοντέλο περισσότερες από μία φορές. Ο Garside (1965) πρότεινε τη βέλτιστη ακολουθία από SWEEPS και οι Schatzoff et al. (1968) απέδειξαν ότι αυτή η ακολουθία είναι όντως η βέλτιστη και έχει ως αποτέλεσμα τον υπολογισμό όλων των $2^k - 1$ μοντέλων. Η βέλτιστη ακολουθία μπορεί να οριστεί ως εξής. Εάν S_k είναι η βέλτιστη ακολουθία των SWEEPS έχοντας στη διάθεσή μας k μεταβλητές, τότε η βέλτιστη ακολουθία για $k+1$ μεταβλητές ορίζεται από $S_{k+1} = S_k \cup \{k+1\} \cup S_k$. Για παράδειγμα, για $\mathrm{k} = 1$ εκτελούμε SWEEP στην πρώτη γραμμή του (3.9) και λαμβάνουμε το μοντέλο $\{x_1\}$. Για $k = 2$ εκτελούμε SWEEPS σύμφωνα με την ακολουθία 1,2,1 λαμβάνοντας τα μοντέλα $\{x_1\}, \{x_1, x_2\}, \{x_2\}$ αντίστοιχα. Για $k = 3$ εκτελούμε SWEEPS σύμφωνα με την ακολουθία 1,2,1,3,1,2,1 λαμβάνοντας τα μοντέλα $\{x_1\}, \{x_1, x_2\}, \{x_2\}, \{x_2, x_3\}, \{x_1, x_2, x_3\}, \{x_1, x_3\}, \{x_3\}$. Η διαδικασία αυτή, για k μεταβλητές, χρειάζεται να εκτελέσει $2^k - 1$ SWEEPS στον πίνακα (3.9).

Έχουν γίνει πολλές προσπάθειες για να μειωθεί το υπολογιστικό κόστος που χρειάζεται για κάθε SWEEP. Οι Schatzoff et al. (Schatzoff, Tsao, & Fienberg, 1968) τροποποίησαν τον αλγόριθμο SWEEP ώστε να διατηρείται η συμμετρία στον πίνακα (3.9) μειώνοντας το υπολογιστικό κόστος εφόσον τα στοιχεία που βρίσκονται κάτω από την κύρια διαγώνιο δε χρειάζεται να υπολογίζονται. Επιπρόσθετα, αντί να εφαρμόζεται το SWEEP σε ολόκληρο τον SSCP πίνακα (3.9), εφαρμόζεται σε κάποιον υποπίνακα μειώνοντας ακόμα περισσότερο το υπολογιστικό κόστος. Με τη χρήση αυτών των τεχνικών οι Schatzoff et al. εκτελούν λιγότερες από 50% πράξεις σε σύγκριση με την απλή εφαρμογή των SWEEPS στον SSCP πίνακα, σύμφωνα με την ακολουθία του Garside. Το αντίτιμο για τη χρήση αυτών των υποπινάκων είναι ότι πρέπει να αποθηκεύουμε κάποιους πίνακες που δημιουργούνται σε προηγούμενα βήματα της διαδικασίας για να χρησιμοποιηθούν σε μετέπειτα στάδια, με το πλήθος των πινάκων που αποθηκεύονται να μην ξεπερνάει το $k+1$. Θα εξηγήσουμε τη βασική ιδέα χρησιμοποιώντας ένα παράδειγμα με $k = 2$ μεταβλητές. Στο βήμα 1 για να προσθέσουμε τη μεταβλητή x_1 στο μοντέλο χρειάζεται να εκτελέσουμε SWEEP μόνο στον υποπίνακα του SSCP που αντιστοιχεί στη μεταβλητή x_1, αγνοώντας εντελώς τα στοιχεία που αφορούν τη μεταβλητή x_2. Το αποτέλεσμα του βήματος 1 είναι το μοντέλο $\{x_1\}$. Για το βήμα 2 πρέπει να έχουμε αποθηκεύσει τον αρχικό $SSCP$ πίνακα, εφόσον η εφαρμογή του SWEEP στο βήμα 1 τον έχει μετασχηματίσει. Εφαρμόζουμε το SWEEP σε ολόκληρο τον πίνακα για να προσαρμόσουμε το μοντέλο $\{x_2\}$. Στο βήμα 3 της διαδικασίας θα χρησιμοποιήσουμε τον πίνακα που λάβαμε στο βήμα 2 για να προσθέσουμε τη

μεταβλητή x_1 στο μοντέλο, λαμβάνοντας το μοντέλο $\{x_1, x_2\}$. Αυτή η ιδέα εφαρμόστηκε και από τους Furnival (1971) και Morgan and Tatar (1972).

3.3.2 Leaps-and-Bounds

Πολλά από τα κριτήρια που έχουν προταθεί για την αξιολόγηση υποψήφιων μοντέλων, όπως αυτά που έχουμε αναφέρει στην Ενότητα 3.1, αποτελούν μονότονες συναρτήσεις του αθροίσματος τετραγώνων των υπολοίπων, RSS, για μοντέλα με το ίδιο πλήθος μεταβλητών (Hocking, 1972). Επομένως, το πρόβλημα επιλογής του βέλτιστου υποσυνόλου μεταβλητών μετατρέπεται στο πρόβλημα εύρεσης των υποσυνόλων μεγέθους $p = 1,2, \dots, k-1$ με το μικρότερο RSS.

Μπορούμε να εκμεταλλευτούμε μία βασική ιδιότητα του RSS για να αποφύγουμε την προσαρμογή πολλών μοντέλων, εξοικονομώντας έτσι υπολογιστικό κόστος. Έχουμε ήδη αναφέρει την ιδιότητα αυτή στην Ενότητα 3.1.1 ωστόσο θα την επαναλάβουμε τώρα εξηγώντας την λίγο πιο λεπτομερώς με ένα μικρό παράδειγμα. Εάν έχουμε δύο μοντέλα M_0, M_1 με $M_0 \subset M_1$, υπό την έννοια ότι όλες οι μεταβλητές του μοντέλου M_0 ανήκουν και στο μοντέλο M_1, τότε γνωρίζουμε ότι $RSS_1 \leq RSS_0$, όπου RSS_0, RSS_1 είναι το άθροισμα τετραγώνων των υπολοίπων για τα μοντέλα M_0, M_1 αντίστοιχα. Για παράδειγμα, εάν έχουμε $K = 4$ μεταβλητές και έχουμε προσαρμόσει τα μοντέλα $\{123\}$ και $\{4\}$ βρίσκοντας $RSS_{\{123\}} = 500$ και $RSS_{\{4\}} = 100$ τότε γνωρίζουμε ότι το καλύτερο μοντέλο που περιέχει μόνο μία μεταβλητή είναι το $\{4\}$. Αυτό ισχύει διότι $RSS_{\{1\}} \geq RSS_{\{123\}} = 500 > 100 = RSS_{\{4\}}$, ομοίως για τα $RSS_{\{2\}}, RSS_{\{3\}}$. Επομένως, σε αυτήν την περίπτωση μπορούμε να αποφύγουμε την προσαρμογή των μοντέλων $\{1\}, \{2\}, \{3\}$. Έχουν γίνει αρκετές μελέτες βασισμένες σε αυτήν την ιδέα (LaMotte & Hocking, 1970; Hocking & Leslie, 1967; Furnival & Wilson, 1974)

Θα αρχίσουμε περιγράφοντας τη μέθοδο **Leaps and Bounds** (LBA) των Furnival και Wilson (1974). Αρχικά, οι K μεταβλητές ταξινομούνται σε φθίνουσα σειρά βάσει της τιμής της ελεγχοσυνάρτησης t, δηλαδή η μεταβλητή x_1 έχει τη μεγαλύτερη στατιστική σημαντικότητα ενώ η μεταβλητή x_k τη μικρότερη. Η ελεγχοσυνάρτηση t χρησιμοποιείται για τον έλεγχο των υποθέσεων

$$\mathrm{H}_0: \beta_j = 0$$

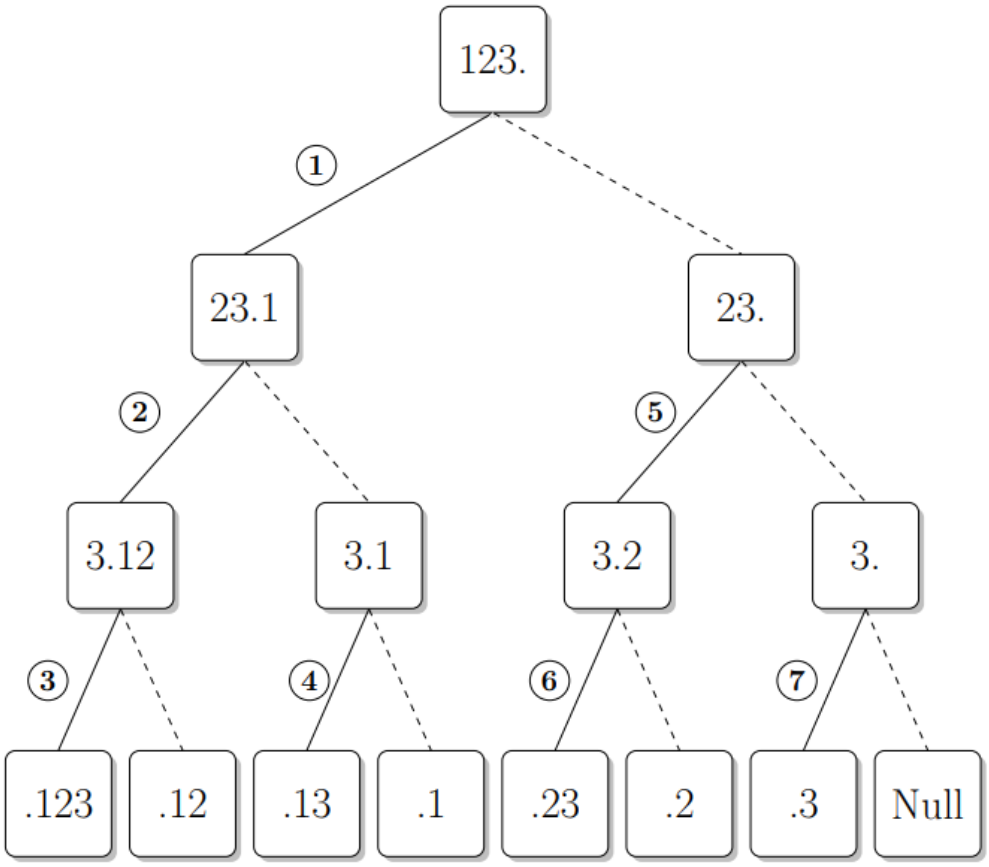
$$H_1: \beta_j \neq 0$$

για το μοντέλο που περιέχει όλες τις K επεξηγηματικές μεταβλητές και ισούται με

$$t = \frac{\hat{\beta}_j}{se(\hat{\beta}_j)} \sim t_{n-k-1}$$

όπου $se(\widehat{\beta_j}) = S\sqrt{c_{jj}}$ είναι η δειγματική τυπική απόκλιση της εκτιμήτριας ελαχίστων τετραγώνων $\widehat{\beta_j}$, $S^2 = \frac{RSS}{n-K-1}$ και c_{jj} είναι το j −οστό διαγώνιο στοιχείο του πίνακα $\boldsymbol{C} = (\boldsymbol{X}'\boldsymbol{X})^{-1}$.

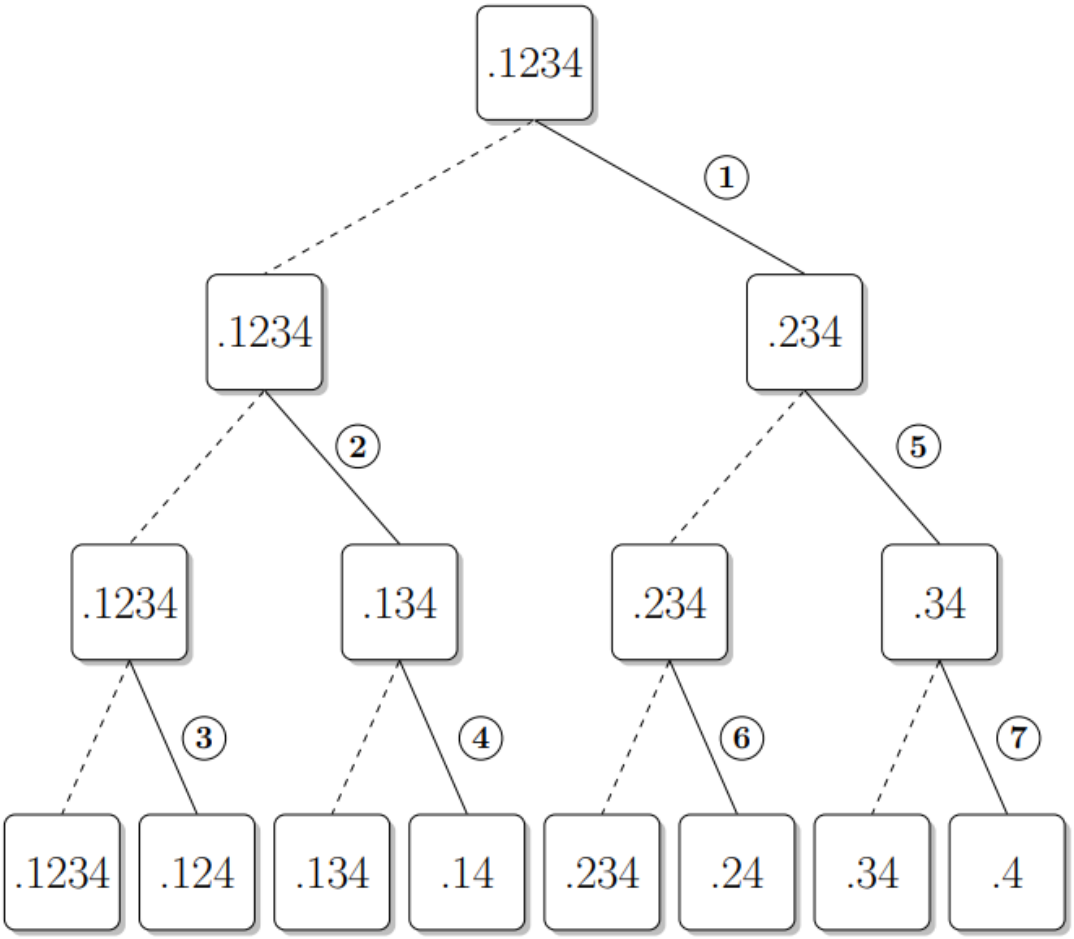
Στη συνέχεια κατασκευάζουμε ένα δέντρο παλινδρόμησης το οποίο αποτελεί ένα δυαδικό δέντρο που αντιστοιχεί την προσαρμογή των 2^{k-1} μοντέλων που δεν περιέχουν τη μεταβλητή x_k, συμπεριλαμβανομένου του μοντέλου που περιέχει μόνο το σταθερό όρο β_0. Για παράδειγμα, για $k = 4$, το δέντρο που αναπαριστά την προσαρμογή των $2^{4-1} = 2^3 = 8$ μοντέλων που δεν περιέχουν την μεταβλητή x_4 φαίνεται στο Σχήμα 3. Αρχίζοντας από τον κόμβο "123." και προχωρώντας προς τα κάτω, κάθε μονοπάτι αναλογεί σε μία ακολουθία από SWEEPS η οποία αντιστοιχεί στην προσαρμογή ενός από αυτά τα $2^3 = 8$ μοντέλα. Για παράδειγμα, συμβολίζοντας το SWEEP που αντιστοιχεί στη μεταβλητή x_r ως SWEEP(r), το μονοπάτι που βρίσκεται στο αριστερό άκρο του δέντρου αντιστοιχεί στην ακολουθία SWEEP(1), SWEEP(2), SWEEP(3) η οποία έχει ως αποτέλεσμα την προσαρμογή του μοντέλου $\{123\}$ το οποίο περιέχει τις μεταβλητές x_1, x_2, x_3. Οι συμπαγείς γραμμές του δέντρου αντιστοιχούν στην εκτέλεση ενός SWEEP. Αντιθέτως, οι διακεκομμένες γραμμές αντιστοιχούν στην παράλειψη ενός SWEEP. Για παράδειγμα, το μονοπάτι που ακολουθεί τους κόμβους "123." – "23." – "3.2" – ".23" αντιστοιχεί στην ακολουθία SWEEP(2), SWEEP(3), όπου στην αρχή της ακολουθίας έχουμε παραλείψει το SWEEP(1), και καταλήγουμε στην προσαρμογή του μοντέλου $\{23\}$. Για την ονομασία των κόμβων του δέντρου, οι μεταβλητές που βρίσκονται στα δεξιά της τελείας (".") είναι αυτές που έχουν ήδη προστεθεί στο μοντέλο σε προηγούμενο στάδιο ενώ οι μεταβλητές που βρίσκονται στα αριστερά της τελείας είναι αυτές που είναι διαθέσιμες να μπουν στο μοντέλο σε εκείνο το στάδιο της διαδικασίας. Οι μεταβλητές που δε συμπεριλαμβάνονται στην ονομασία ενός κόμβου είναι αυτές που έχουν διαγραφεί (παραλειφθεί) σε προηγούμενο στάδιο.



Σχήμα 3 - Δέντρο παλινδρόμησης για $k = 4$ μεταβλητές

Σημειώνουμε ότι χρησιμοποιώντας αυτή τη διαδικασία δεν αφαιρούνται ποτέ μεταβλητές από το μοντέλο. Για αυτόν το λόγο αντί SWEEPS μπορούμε να χρησιμοποιήσουμε απαλοιφή Gauss για να προσθέσουμε μεταβλητές στο μοντέλο εξοικονομώντας έτσι υπολογιστικό κόστος.

Εκτός από το δέντρο παλινδρόμησης χρησιμοποιείται και ένα δευτερεύον δέντρο φραγμάτων το οποίο φαίνεται στο Σχήμα 4 για την περίπτωση όπου $k = 4$. Το δέντρο αυτό αντιστοιχεί στην προσαρμογή των υπόλοιπων 2^{k-1} μοντέλων που περιέχουν τη μεταβλητή x_k. Η ρίζα του δέντρου, δηλαδή ο κόμβος ".1234" στο παράδειγμά μας, αντιπροσωπεύει το πλήρες μοντέλο που περιέχει όλες τις διαθέσιμες μεταβλητές. Οι συμπαγείς γραμμές αντιστοιχούν στην αφαίρεση μίας μεταβλητής από το μοντέλο χρησιμοποιώντας ένα SWEEP ενώ οι διακεκομμένες γραμμές αντιστοιχούν στην παράλειψη ενός SWEEP. Για την ονομασία των κόμβων, οι μεταβλητές εμφανίζονται μόνο στα δεξιά της τελείας "." και αντιπροσωπεύουν τις μεταβλητές που περιέχει το μοντέλο και είναι διαθέσιμες για αφαίρεση, εξαιρώντας τη μεταβλητή x_k η οποία δεν αφαιρείται ποτέ.



Σχήμα 4 - Δέντρο φραγμάτων για $k = 4$ μεταβλητές

Ένα SWEEP στο πρωτεύον δέντρο παλινδρόμησης αναλογεί στην κατάβαση από ένα κόμβο στον επόμενο μέσω του αριστερού κλάδου του κόμβου. Στο δευτερεύον δέντρο αυτό το SWEEP αντιστοιχεί σε ένα SWEEP το οποίο αναλογεί στην κατάβαση από τον αντίστοιχο κόμβο στον επόμενο μέσω του

δεξιού κλάδου του κόμβου. Για παράδειγμα, στο πρωτεύον δέντρο έχουμε την κατάβαση από τον κόμβο "123." από τον κόμβο "23.1" μέσω του SWEEP(1) και στο δευτερεύον δέντρο έχουμε την κατάβαση από τον κόμβο ".1234" προς τον κόμβο ".234" μέσω του SWEEP(1).

Η κατάβαση στο πρωτεύον δέντρο γίνεται με λεξικογραφική σειρά. Αυτό αναλογεί στην κατάβαση του αριστερότερου μονοπατιού του δέντρου, μέχρι τον κατώτερο κλάδο και στη συνέχεια την οπισθοδρόμηση κατά μία ελάχιστη απόσταση ώστε να συνεχίσουμε στο επόμενο αριστερότερο μονοπάτι. Σε κάθε κόμβο αποφασίζουμε αν είναι αναγκαία η κατάβαση σε κατώτερους κόμβους χρησιμοποιώντας το δευτερεύον δέντρο και την ιδιότητα του RSS για να βρούμε ένα κάτω φράγμα για τα RSS των μοντέλων που ακολουθούν.

Αρχίζουμε με το Βήμα 0 προσαρμόζοντας τα μοντέλα του αριστερότερου μονοπατιού του πρωτεύοντος δέντρου, δηλαδή τα μοντέλα $\{1\},\{1,2\},\dots,\{1,2,\dots,k-1\}$ και στη συνέχεια εφαρμόζουμε το SWEEP(k) για να προσαρμόσουμε και το μοντέλο $\{1,2,\dots,k\}$. Σε αυτό το σημείο έχουμε προσαρμόσει k μοντέλα. Έπειτα εκτελούμε τα αντίστοιχα SWEEPS για το δευτερεύον δέντρο. Αυτό θα μας δώσει τα μοντέλα $\{2,3,..,k\},\{1,3,4,\dots k\},\dots,\{1,2,\dots,r-1,r+1,\dots,k\},\dots,\{1,2,\dots,k-2,k\}$ για $1\le r\le k-1$. Σε αυτό το σημείο έχουμε προσαρμόσει $2k-1$ μοντέλα. Συγκεκριμένα, έχουμε 1 μοντέλο που περιέχει k' μεταβλητές για κάθε k' με $1\le k'\le K-1$, 1 μοντέλο που περιέχει όλες τις μεταβλητές και $k-1$ μοντέλα που περιέχουν $k-1$ μεταβλητές. Έπειτα συνεχίζουμε με το Βήμα 1 προσαρμόζοντας το επόμενο μοντέλο του αμέσως επόμενου αριστερότερου μονοπατιού του πρωτεύοντος δέντρου, αφού πρώτα ελέγξουμε το φράγμα για το RSS που δίνει ο αντίστοιχος κόμβος του δευτερεύοντος δέντρου για να αποφασίσουμε αν πρέπει να συνεχίσουμε σε αυτό το μονοπάτι ή να το παραλείψουμε.

Θα χρησιμοποιήσουμε ένα παράδειγμα με $k=4$ ώστε να γίνει πιο κατανοητή η παραπάνω διαδικασία. Στον Πίνακα 1 (*Πίνακας 1*) παρουσιάζουμε τα μοντέλα για το παράδειγμά μας, με τη σειρά που προσαρμόζονται ή ελέγχονται, καθώς και τα αντίστοιχα RSS. Τα κελιά που σκιαγραφούνται με γκρίζο χρώμα αντιστοιχούν στα μοντέλα των οποίων η προσαρμογή παραλείπεται. Τα κελιά που παρουσιάζονται με έντονο χρώμα γραμματοσειράς αντιστοιχούν στα καλύτερα μοντέλα 1,2 και 3 μεταβλητών για το τρέχον βήμα. Στο Βήμα 0 εκτελούμε τα SWEEPS του αριστερότερου μονοπατιού του δέντρου παλινδρόμησης προσαρμόζοντας τα μοντέλα $\{1\},\{1,2\},\{1,2,3\}$, έπειτα εκτελούμε το $SWEEP(4)$ για να προσαρμόσουμε το μοντέλο $\{1,2,3,4\}$ και στη συνέχεια εκτελούμε τα αντίστοιχα SWEEPS στο δέντρο φραγμάτων, προσαρμόζοντας τα μοντέλα $\{2,3,4\},\{1,3,4\}$ και $\{1,2,4\}$. Σε αυτό το σημείο, το καλύτερο μοντέλο μίας μεταβλητής είναι το $\{1\}$ με $RSS_{\{1\}}=82$, το καλύτερο μοντέλο δύο μεταβλητών είναι το $\{1,2\}$ με $RSS_{\{1,2\}}=10$ και το καλύτερο μοντέλο τριών μεταβλητών είναι το $\{1,2,3\}$ με $RSS_{\{1,2,3\}}=6$.

Εφόσον έχουμε φτάσει στο κατώτερο σημείο του αριστερότερου μονοπατιού του δέντρου παλινδρόμησης, πηγαίνουμε προς τα πίσω για να φτάσουμε στον επόμενο υποψήφιο κόμβο, του δεύτερου αριστερότερου μονοπατιού του δέντρου παλινδρόμησης. Αυτός είναι ο κόμβος "3.1" και το υποψήφιο μοντέλο για προσαρμογή είναι το $\{1,3\}$. Σε αυτό το σημείο ελέγχουμε το φράγμα που δίνει ο αντίστοιχος κόμβος του δέντρου φραγμάτων, δηλαδή ο κόμβος ".134". Αυτό το φράγμα είναι το $RSS_{\{1,3\}}\ge RSS_{\{1,3,4\}}=60$ και εφόσον το τρέχον καλύτερο μοντέλο δύο μεταβλητών έχει $RSS_{\{1,2\}}=10<60\le RSS_{\{1,3\}}$ τότε δε χρειάζεται να προσαρμόσουμε το μοντέλο $\{1,3\}$. Δηλαδή παραλείπουμε το SWEEP αυτού του βήματος τόσο στο δέντρο παλινδρόμησης όσο και στο δέντρο φραγμάτων. Στην περίπτωση όπου το φράγμα μας ήταν μικρότερο από το τρέχον καλύτερο RSS τότε θα έπρεπε να εφαρμόσουμε το SWEEP στα δέντρα παλινδρόμησης και φραγμάτων.

Συνεχίζουμε με τον επόμενο υποψήφιο κόμβο του δέντρου παλινδρόμησης ο οποίος είναι ο “23.” και τα υποψήφια μοντέλα για προσαρμογή είναι τα $\{2\}$ και $\{2,3\}$. Ο αντίστοιχος κόμβος στο δέντρο φραγμάτων είναι ο “.234” και το φράγμα είναι το $RSS_{\{2,3,4\}} = 157$. Δηλαδή

$$RSS_{\{2\}} \geq RSS_{\{2,3,4\}} = 157 > RSS_{\{1\}} = 82$$

$$RSS_{\{2,3\}} \geq RSS_{\{2,3,4\}} = 157 > RSS_{\{1,2\}} = 10$$

επομένως δε χρειάζεται να προσαρμοστούν ούτε αυτά τα μοντέλα ούτε τα αντίστοιχα μοντέλα του δέντρου φραγμάτων.

Τέλος, εξετάζουμε τον κόμβο “3.” με υποψήφιο μοντέλο το $\{3\}$ και φράγμα

$$RSS_{\{3\}} \geq RSS_{\{3,4\}} \geq RSS_{\{2,3,4\}} = 157 > RSS_{\{1\}} = 82$$

επομένως το μοντέλο $\{3\}$ δε χρειάζεται να προσαρμοστεί. Το ίδιο ισχύει και για το μοντέλο $\{4\}$ του δέντρου φραγμάτων.

		Δέντρο Παλινδρόμησης			**Δέντρο Φραγμάτων**			
Βήμα	**Δείκτης Οδηγού Στοιχείου**	**Πίνακας**	**Μοντέλο (Πίνακας)**	***RSS***	**Πίνακας**	**Μοντέλο (Πίνακας)**	***RSS***	**Φράγμα Βήματος**
0			Null	320				
	1	123.	$\{\mathbf{1}\}$ **(23.1)**	**82**	.1234	$\{2,3,4\}$ (.234)	157	
	2	23.1	$\{\mathbf{1},\mathbf{2}\}$ **(3.12)**	**10**	.1234	$\{1,3,4\}$ (.134)	60	
	3	3.12	$\{\mathbf{1},\mathbf{2},\mathbf{3}\}$ **(.123)**	**6**	.1234	$\{1,2,4\}$ (.124)	10	
	4				(.123)	$\{1,2,3,4\}$ (.1234)	5	
1	3	3.1	$\{1,3\}$ (.13)	62				60
2	2	23.	$\{2\}$ (3.2)	186				157
	3	3.2	$\{2,3\}$ (.23)	157				157
3	3	3.	$\{3\}$ (.3)	229				157

Πίνακας 1 - Παράδειγμα εφαρμογής του αλγόριθμου Leaps-and-Bounds

3.3.3 Branch and Bound

Θεωρούμε τον πίνακα $\boldsymbol{X} \in \mathbb{R}^{n\times p}$ με $rank(\boldsymbol{X}) = p$. Από την Ενότητα 2.5 έχουμε ότι η παραγοντοποίηση QR για τον πίνακα $\boldsymbol{X}$ δίνεται από τη σχέση

$$\boldsymbol{Q}^T\boldsymbol{X} = \begin{matrix}\boldsymbol{R}\\ \boldsymbol{0}\end{matrix}\begin{matrix}p\\ n-p\end{matrix} \tag{3.10}$$

όπου ο $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ είναι ορθογώνιος και $\boldsymbol{R} \in \mathbb{R}^{p\times p}$ είναι ένας αντιστρέψιμος, άνω τριγωνικός πίνακας. Θέτουμε

$$\boldsymbol{Q}^T\boldsymbol{y} = \begin{matrix}\tilde{\boldsymbol{y}}_1\\ \tilde{\boldsymbol{y}}_2\end{matrix}\begin{matrix}p\\ n-p\end{matrix}$$

Τότε η εκτιμήτρια ελαχίστων τετραγώνων για την $\boldsymbol{\beta}$ δίνεται από τη σχέση

$$\widehat{\boldsymbol{\beta}} = \underset{\beta}{\operatorname{argmin}}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|^2 = \underset{\beta}{\operatorname{argmin}}\|\boldsymbol{Q}^T(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta})\|^2 = \boldsymbol{R}^{-1}\tilde{\boldsymbol{y}}_1$$

όπου η $\|\cdot\|$ αντιπροσωπεύει την Ευκλείδεια νόρμα. Το άθροισμα τετραγώνων των υπολοίπων (RSS) υπολογίζεται από τη σχέση

$$RSS = \left\|\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right\|^2 = \left\|\boldsymbol{Q}^T\left(\boldsymbol{y} - \boldsymbol{X}\widehat{\boldsymbol{\beta}}\right)\right\|^2 = \|\tilde{\boldsymbol{y}}_2\|^2$$

Έστω ότι $\boldsymbol{S}$ αντιπροσωπεύει ένα $p \times d$ πίνακα επιλογής, ο οποίος αποτελείται από d στήλες του $p \times p$ μοναδιαίου πίνακα $\boldsymbol{I}_p$**.** Ας θεωρήσουμε το τροποποιημένο μοντέλο παλινδρόμησης

$$\boldsymbol{y} = \boldsymbol{X}_{(\boldsymbol{S})}\boldsymbol{\beta}_{(\boldsymbol{S})} + \varepsilon \tag{3.11}$$

όπου $\boldsymbol{X}_{(\boldsymbol{S})} = \boldsymbol{X}\boldsymbol{S} \in \mathbb{R}^{n\times d}$ και $\boldsymbol{\beta}_{(\boldsymbol{S})} = \boldsymbol{S}^T\boldsymbol{\beta} \in \mathbb{R}^d$**.** Η εκτιμήτρια ελαχίστων τετραγώνων του $\boldsymbol{\beta}_{(\boldsymbol{S})}$ γράφεται

$$\widehat{\boldsymbol{\beta}}_{(\boldsymbol{S})} = \underset{\beta_{(S)}}{\operatorname{argmin}}\left\|\boldsymbol{y} - \boldsymbol{X}_{(\boldsymbol{S})}\boldsymbol{\beta}_{(\boldsymbol{S})}\right\|^2 = \underset{\beta_{(S)}}{\operatorname{argmin}}\left\|\boldsymbol{Q}^T\left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{S}\boldsymbol{\beta}_{(\boldsymbol{S})}\right)\right\|^2 = \underset{\beta_{(S)}}{\operatorname{argmin}}\left\|\tilde{\boldsymbol{y}}_1 - \boldsymbol{R}\boldsymbol{S}\boldsymbol{\beta}_{(\boldsymbol{S})}\right\|^2$$

Η παραγοντοποίηση QR για τον $p \times d$ πίνακα $\boldsymbol{R}\boldsymbol{S}$ γράφεται

$$\boldsymbol{Q}_{(\boldsymbol{S})}^T\boldsymbol{R}\boldsymbol{S} = \begin{matrix}\boldsymbol{R}_{(\boldsymbol{S})}\\ \boldsymbol{0}\end{matrix}\begin{matrix}d\\ p-d\end{matrix} \tag{3.12}$$

και

$$\boldsymbol{Q}_{(\boldsymbol{S})}^T\tilde{\boldsymbol{y}}_1 = \begin{matrix}\hat{\boldsymbol{y}}_1\\ \hat{\boldsymbol{y}}_2\end{matrix}\begin{matrix}d\\ p-d\end{matrix}$$

Η εκτιμήτρια ελαχίστων τετραγώνων και το RSS για το τροποποιημένο μοντέλο (3.11) δίνονται, αντίστοιχα, από τις σχέσεις

$$\widehat{\boldsymbol{\beta}}_{(\boldsymbol{S})} = \underset{\beta_{(S)}}{\operatorname{argmin}} \left\| \tilde{\boldsymbol{y}}_1 - \boldsymbol{RS}\boldsymbol{\beta}_{(\boldsymbol{S})} \right\|^2 = \underset{\beta_{(S)}}{\operatorname{argmin}} \left\| \boldsymbol{Q}_{(\boldsymbol{S})}^T \left(\tilde{\boldsymbol{y}}_1 - \boldsymbol{RS}\boldsymbol{\beta}_{(\boldsymbol{S})} \right) \right\|^2$$
$$= \underset{\beta_{(S)}}{\operatorname{argmin}} \left\| \begin{pmatrix} \widehat{\boldsymbol{y}}_1 - \boldsymbol{R}_{(\boldsymbol{S})}\boldsymbol{\beta}_{(\boldsymbol{S})} \\ \widehat{\boldsymbol{y}}_2 \end{pmatrix} \right\|^2 = \underset{\beta_{(S)}}{\operatorname{argmin}} \left\| \widehat{\boldsymbol{y}}_1 - \boldsymbol{R}_{(\boldsymbol{S})}\boldsymbol{\beta}_{(\boldsymbol{S})} \right\|^2 = \boldsymbol{R}_{(\boldsymbol{S})}^{-1}\widehat{\boldsymbol{y}}_1$$

και

$$RSS_{(\boldsymbol{S})} = \left\| \boldsymbol{y} - \boldsymbol{X}_{(\boldsymbol{S})}\widehat{\boldsymbol{\beta}}_{(\boldsymbol{S})} \right\|^2 = \left\| \boldsymbol{Q}^T \left(\boldsymbol{y} - \boldsymbol{X}_{(\boldsymbol{S})}\widehat{\boldsymbol{\beta}}_{(\boldsymbol{S})} \right) \right\|^2 = \left\| \boldsymbol{Q}^T \left(\boldsymbol{y} - \boldsymbol{X}\boldsymbol{S}\boldsymbol{R}_{(\boldsymbol{S})}^{-1}\widehat{\boldsymbol{y}}_1 \right) \right\|^2$$
$$= \left\| \begin{pmatrix} \tilde{\boldsymbol{y}}_1 - \boldsymbol{RS}\boldsymbol{R}_{(\boldsymbol{S})}^{-1}\widehat{\boldsymbol{y}}_1 \\ \tilde{\boldsymbol{y}}_2 \end{pmatrix} \right\|^2 = \|\tilde{\boldsymbol{y}}_2\|^2 + \left\| \tilde{\boldsymbol{y}}_1 - \boldsymbol{RS}\boldsymbol{R}_{(\boldsymbol{S})}^{-1}\widehat{\boldsymbol{y}}_1 \right\|^2$$
$$= \|\tilde{\boldsymbol{y}}_2\|^2 + \left\| \boldsymbol{Q}_{(\boldsymbol{S})}^T \left(\tilde{\boldsymbol{y}}_1 - \boldsymbol{RS}\boldsymbol{R}_{(\boldsymbol{S})}^{-1}\widehat{\boldsymbol{y}}_1 \right) \right\|^2 = \|\tilde{\boldsymbol{y}}_2\|^2 + \left\| \begin{pmatrix} \widehat{\boldsymbol{y}}_1 - \boldsymbol{R}_{(\boldsymbol{S})}\boldsymbol{R}_{(\boldsymbol{S})}^{-1}\widehat{\boldsymbol{y}}_1 \\ \widehat{\boldsymbol{y}}_2 \end{pmatrix} \right\|^2$$
$$= \|\tilde{\boldsymbol{y}}_2\|^2 + \|\widehat{\boldsymbol{y}}_2\|^2 = RSS + \widehat{\boldsymbol{y}}_2^T\widehat{\boldsymbol{y}}_2$$

Η παραγοντοποίηση (3.12) είναι ισοδύναμη με την επανα-τριγωνοποίηση του πίνακα $\boldsymbol{R}$ στη σχέση (3.13) έπειτα από διαγραφή ή εναλλαγή στηλών (Gatu & Kontoghiorghes, 2006; Gatu & Kontoghiorghes, 2003).

Σημειώνουμε ότι στην περίπτωση όπου ο πίνακας $\boldsymbol{S}$ αποτελείται από τις πρώτες d $(d = 1,2, \dots, p)$ στήλες του πίνακα $\boldsymbol{I}_p$ τότε $\boldsymbol{Q}_{(\boldsymbol{S})}^T \equiv \boldsymbol{I}_p$ και ο $\boldsymbol{R}_{(\boldsymbol{S})}$ αντιστοιχεί στον άνω αριστερά $d \times d$ υποπίνακα του $\boldsymbol{RS}$ (leading submatrix). Επομένως, δεδομένου του πίνακα $\boldsymbol{R}$, τα υπομοντέλα που απαρτίζονται από τις μεταβλητές $[1], [1,2], [1,2,3], \dots, [1,2, \dots, p]$ λαμβάνονται από τους αντίστοιχους $1 \times 1, 2 \times 2, 3 \times 3, \dots, p \times p$ άνω αριστερά τριγωνικούς υποπίνακες του $\boldsymbol{R}$.

Υπενθυμίζουμε ότι, εάν υπολογίσουμε την παραγοντοποίηση QR για τον πίνακα $(\boldsymbol{X}, \boldsymbol{y})$, ο πίνακας $\boldsymbol{R}$ που λαμβάνουμε, τον οποίο θα συμβολίζουμε ως $\boldsymbol{R}^*$, έχει τη μορφή

$$\boldsymbol{R}^* = \begin{pmatrix} \boldsymbol{R} & \boldsymbol{d}_1 \\ \boldsymbol{0} & s \end{pmatrix}$$

όπου $\boldsymbol{R}$ και $\boldsymbol{d}_1$ είναι οι ποσότητες που αναφέραμε παραπάνω και $s^2 = RSS$ για το προσαρμοσμένο μοντέλο.

Ο πίνακας $\boldsymbol{R}^*$ περιέχει όλη την πληροφορία που χρειάζεται για να υπολογίσουμε το RSS για όλα τα πιθανά υπομοντέλα (Smith & Bremner, 1989). Η τελευταία στήλη του $\boldsymbol{R}^*$ δίνει το RSS του προσαρμοσμένου μοντέλου καθώς και για άλλα μοντέλα που προσαρμόζονται σε φθίνουσα σειρά. Για παράδειγμα αν έχουμε προσαρμόσει ένα μοντέλο με τέσσερεις μεταβλητές, συμπεριλαμβανομένης της μεταβλητής $x \equiv 1$ που αντιστοιχεί στο σταθερό όρο, δηλαδή $p = 4$ και αν συμβολίσουμε αυτό το μοντέλο ως $ABCD$ με τις μεταβλητές να έχουν αποθηκευτεί με αυτή τη σειρά, τότε μπορούμε να υπολογίσουμε τα RSS για μερικά υπομοντέλα όπως φαίνεται στον παρακάτω πίνακα, συμβολίζοντας $\boldsymbol{d}_1 \equiv \tilde{\boldsymbol{d}}$:

Μοντέλο	***RSS***
$ABCD$	s^2
ABC	$s^2 + \tilde{d}_4^2$
AB	$s^2 + \tilde{d}_4^2 + \tilde{d}_3^2$
A	$s^2 + \tilde{d}_4^2 + \tilde{d}_3^2 + \tilde{d}_2^2$

Πίνακας 2 – Υπολογισμός του RSS υπομοντέλων

όπου $\tilde{d}_i$, $i = 2,..,4$ είναι τα στοιχεία του $\boldsymbol{d}_1$.

Παρόλο που το s θα είναι το ίδιο για το μοντέλο $DCBA$, τα στοιχεία του $\boldsymbol{d}_1$ καθώς και τα υπομοντέλα για τα οποία μπορούμε να υπολογίσουμε το RSS θα είναι διαφορετικά. Επομένως, μεταθέτοντας τις στήλες του $\boldsymbol{R}^*$ και εφαρμόζοντας ξανά τριγωνοποίηση στον τροποποιημένο πίνακα μπορούμε να υπολογίσουμε το RSS για άλλα μοντέλα. Για παράδειγμα, εάν έχουμε τον πίνακα $\boldsymbol{R}^*$ για το μοντέλο $ABCD$ τότε μπορούμε να ανταλλάξουμε θέσεις στις στήλες 3 και 4 και να τριγωνοποιήσουμε των πίνακα που λαμβάνουμε καταλήγοντας στον πίνακα $\boldsymbol{R}^*$για το μοντέλο $ABDC$. Τότε το RSS για το μοντέλο ABD υπολογίζεται από τη σχέση $s^2 + \tilde{d}_4^2$ του καινούργιου πίνακα $\boldsymbol{R}^*$. Η παραπάνω διαδικασία για τον υπολογισμό των RSS για όλα τα πιθανά μοντέλα είναι πιο αποδοτική από το να υπολογίσουμε την παραγοντοποίηση QR για κάθε μοντέλο ξεχωριστά ωστόσο κάθε φορά που ξαναυπολογίζουμε τα στοιχεία του $\boldsymbol{R}^*$ οδηγούμαστε σε κάποια απώλεια στην ακρίβεια λόγω της συσσώρευσης σφαλμάτων στρογγυλοποίησης.

Εφαρμόζοντας περιστροφές Givens

Ο αρχικός πίνακας $\boldsymbol{R}^*$ μπορεί να υπολογιστεί χρησιμοποιώντας οποιαδήποτε μέθοδο από αυτές που περιγράψαμε στην Ενότητα 2.5.2. Οι στήλες του $\boldsymbol{R}^*$ μπορούν να μετατοπιστούν κατά περισσότερες από μία θέσεις και στη συνέχεια να εφαρμοστεί τριγωνοποίηση του καινούργιου πίνακα χρησιμοποιώντας τον τροποποιημένο αλγόριθμο Gram-Schmidt ή ανακλάσεις Householder. Ωστόσο είναι πιο απλό να μετατοπίσουμε στήλες κατά μία θέση και να χρησιμοποιήσουμε περιστροφές Givens για την τριγωνοποίηση του καινούργιου πίνακα.

Η κατασκευή μίας περιστροφής Givens ισοδυναμεί με τον υπολογισμό των ποσοτήτων s, c και r ώστε να ισχύει

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} r \\ 0 \end{bmatrix}$$

όπου οι ποσότητες a και b είναι γνωστές και $c^2 + s^2 = 1$. Δηλαδή θέλουμε να βρούμε ένα πίνακα ώστε να μετασχηματίζουμε ένα διάνυσμα $\begin{bmatrix} a \\ b \end{bmatrix}$ σε ένα διάνυσμα $\begin{bmatrix} r \\ 0 \end{bmatrix}$. Η εφαρμογή μίας περιστροφής Givens αποτελείται από τον εξ αριστερά πολλαπλασιασμό ενός $2 \times j$ πίνακα με τον 2×2 πίνακα που υπολογίσαμε παραπάνω ο οποίος περιέχει τα c και s :

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} x_{11} & \cdots & x_{1j} \\ x_{21} & \cdots & x_{2j} \end{bmatrix}$$

Στη συνέχεια θεωρούμε το παράδειγμα που δώσαμε παραπάνω όπου έχουμε προσαρμόσει το μοντέλο $ABCD$ λαμβάνοντας τον πίνακα

$$\boldsymbol{R}^* = \begin{bmatrix} r_{11} & r_{12} & r_{13} & r_{14} & \tilde{d}_1 \\ 0 & r_{22} & r_{23} & r_{24} & \tilde{d}_2 \\ 0 & 0 & r_{33} & r_{34} & \tilde{d}_3 \\ 0 & 0 & 0 & r_{44} & \tilde{d}_4 \\ 0 & 0 & 0 & 0 & s \end{bmatrix}$$

Θα δείξουμε πώς μπορούμε υπολογίσουμε το RSS για το μοντέλο ACB χρησιμοποιώντας μεταθέσεις διαδοχικών στηλών και τριγωνοποιώντας τον πίνακα που προκύπτει χρησιμοποιώντας περιστροφές Givens. Αρχικά ανταλλάζουμε θέση στις στήλες 2 και 3 λαμβάνοντας τον πίνακα

$$\begin{bmatrix} r_{11} & r_{13} & r_{12} & r_{14} & \tilde{d}_1 \\ 0 & r_{23} & r_{22} & r_{24} & \tilde{d}_2 \\ 0 & r_{33} & 0 & r_{34} & \tilde{d}_3 \\ 0 & 0 & 0 & r_{44} & \tilde{d}_4 \\ 0 & 0 & 0 & 0 & s \end{bmatrix}$$

Υπολογίζουμε μία περιστροφή Givens ώστε

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} r_{23} \\ r_{33} \end{bmatrix} = \begin{bmatrix} r_{22}^* \\ 0 \end{bmatrix}$$

Όπως έχουμε αναφέρει στην Ενότητα 2.5.2 από την παραπάνω περιστροφή μόνο τα στοιχεία της δεύτερης και τρίτης γραμμής του πίνακα θα επηρεαστούν δηλαδή η εφαρμογή της περιστροφής Givens που υπολογίσαμε ισοδυναμεί με τον πολλαπλασιασμό

$$\begin{bmatrix} c & s \\ -s & c \end{bmatrix} \begin{bmatrix} 0 & r_{23} & r_{22} & r_{24} & \tilde{d}_2 \\ 0 & r_{33} & 0 & r_{34} & \tilde{d}_3 \end{bmatrix}$$

Μετά την εφαρμογή της περιστροφής Givens λαμβάνουμε ένα καινούργιο πίνακα $\boldsymbol{R}^*$ που αντιστοιχεί στο μοντέλο $ACBD$. Οι ποσότητες c, s και r για την περιστροφή Givens δίνονται από τις σχέσεις

$$r = SIGN \cdot (a^2 + b^2)^{1/2}$$

όπου $SIGN$ αντιπροσωπεύει το πρόσημο του a, αν $|a| \geq |b|$ ή το πρόσημο του b, αν $|b| \geq |a|$

και

$$c = \frac{a}{r}, \qquad s = \frac{b}{r}, \text{ αν } r \neq 0$$

ή

$$c = 1, \qquad s = 0, \text{ αν } r = 0$$

Υπολογισμός όλων των πιθανών μοντέλων – DCA

Για τον υπολογισμό όλων των πιθανών μοντέλων ο Clarke (1981) ανέπτυξε ένα αλγόριθμο ο οποίος βασίζεται στην εναλλαγή στηλών. Συγκεκριμένα, σε κάθε βήμα του αλγορίθμου εναλλάσσονται δύο διαδοχικές στήλες και στη συνέχεια εφαρμόζεται τριγωνοποίηση στον πίνακα που λαμβάνουμε

χρησιμοποιώντας περιστροφές Givens. Για την παραπάνω διαδικασία εφαρμόζεται μία συγκεκριμένη ακολουθία βημάτων ώστε να επιτευχθεί ο ελάχιστος αριθμός μεταθέσεων και περιστροφών Givens. Συνολικά εφαρμόζονται $2^p - p - 1$ περιστροφές Givens και από κάθε περιστροφή Givens λαμβάνουμε ένα μοντέλο. Τα υπόλοιπα p μοντέλα που απομένουν είναι τα μοντέλα που λαμβάνουμε από την αρχική παραγοντοποίηση QR.

Οι Smith και Bremner (1989) ανέπτυξαν τον αλγόριθμο DCA (Dropping Columns Algorithm) βασισμένο σε παρόμοια λογική με τον Clarke. Ο αλγόριθμος DCA εφαρμόζει περιστροφές Givens σε πίνακες μικρότερων διαστάσεων εξοικονομώντας υπολογιστικό κόστος ωστόσο χρειάζεται περισσότερο αποθηκευτικό χώρο.

Στη συνέχεια θα περιγράψουμε τη διαδικασία που ακολουθεί ο αλγόριθμος DCA η οποία παρουσιάζεται εκτενώς από τους Gatu and Kontoghiorghes (2003). Έστω ότι $M_{k,\lambda}^{\upsilon}$ αντιπροσωπεύει τον άνω τριγωνικό πίνακα $\boldsymbol{R}$ της παραγοντοποίησης QR ενός πίνακα $\boldsymbol{X}$ που απαρτίζεται από τις στήλες (μεταβλητές) $\upsilon_1, \dots, \upsilon_{k+\lambda}$. Το ζεύγος (k, λ) υποδεικνύει ότι οι στήλες $k+1, \dots, k+\lambda-1$ θα διαγραφούν μία-μία από τον άνω τριγωνικό πίνακα ώστε να λάβουμε καινούργια μοντέλα. Επιπρόσθετα, ορίζουμε το $(\lambda - 1)$ – δέντρο παλινδρόμησης $T_{k,\lambda}^{\upsilon}$ με κόμβο-ρίζα το $M_{k,\lambda}^{\upsilon}$ και παιδιά τα δέντρα $T_{k+i-1,\lambda-i}^{\upsilon^{(k+i)}}$ για $i = 1, \dots, \lambda - 1$. Ο συμβολισμός $\upsilon^{(k+i)}$ αντιπροσωπεύει το διάνυσμα $\boldsymbol{\upsilon} = [\upsilon_1, \dots, \upsilon_p]$ χωρίς το $(k+i)$-οστό στοιχείο του. Όπως έχουμε αναφέρει, χρησιμοποιώντας τον πίνακα $M_{k,\lambda}^{\upsilon} \equiv \boldsymbol{R}$ καθώς και το τροποποιημένο διάνυσμα απόκρισης $\boldsymbol{d}_1 \equiv \boldsymbol{Q}^T \boldsymbol{y}$ μπορούμε να υπολογίσουμε το RSS για τα υπομοντέλα $(\upsilon_1), (\upsilon_1, \upsilon_2), \dots, (\upsilon_1, \upsilon_2, \dots, \upsilon_{k+\lambda})$. Τα μοντέλα $(\upsilon_1), (\upsilon_1, \upsilon_2), \dots, (\upsilon_1, \dots, \upsilon_k)$ μπορούν να εξαχθούν από τον κόμβο-γονέα του δέντρου παλινδρόμησης ενώ τα μοντέλα $(\upsilon_1, \dots, \upsilon_{k+1}), \dots, (\upsilon_1, \dots, \upsilon_{k+\lambda})$ μπορούν να εξαχθούν από το $M_{k,\lambda}^{\upsilon}$. Για να παράξουμε ένα κόμβο-παιδί από κάποιο γονέα διαγράφουμε μία στήλη από τον άνω τριγωνικό πίνακα και στη συνέχεια εφαρμόζουμε τριγωνοποίηση στον πίνακα που λαμβάνουμε χρησιμοποιώντας περιστροφές Givens. Οι περιστροφές Givens εφαρμόζονται ταυτόχρονα και στο τροποποιημένο διάνυσμα $\boldsymbol{d_1}$. Στο Σχήμα 4 φαίνεται η ακολουθία περιστροφών Givens για την τριγωνοποίηση ενός 5×5 άνω τριγωνικού πίνακα αφού διαγραφεί η δεύτερη στήλη. Τα στοιχεία $\boldsymbol{x}$ αντιπροσωπεύουν τα στοιχεία του πίνακα που επηρεάζονται από την εκάστοτε περιστροφή Givens σε κάθε βήμα της τριγωνοποίησης.

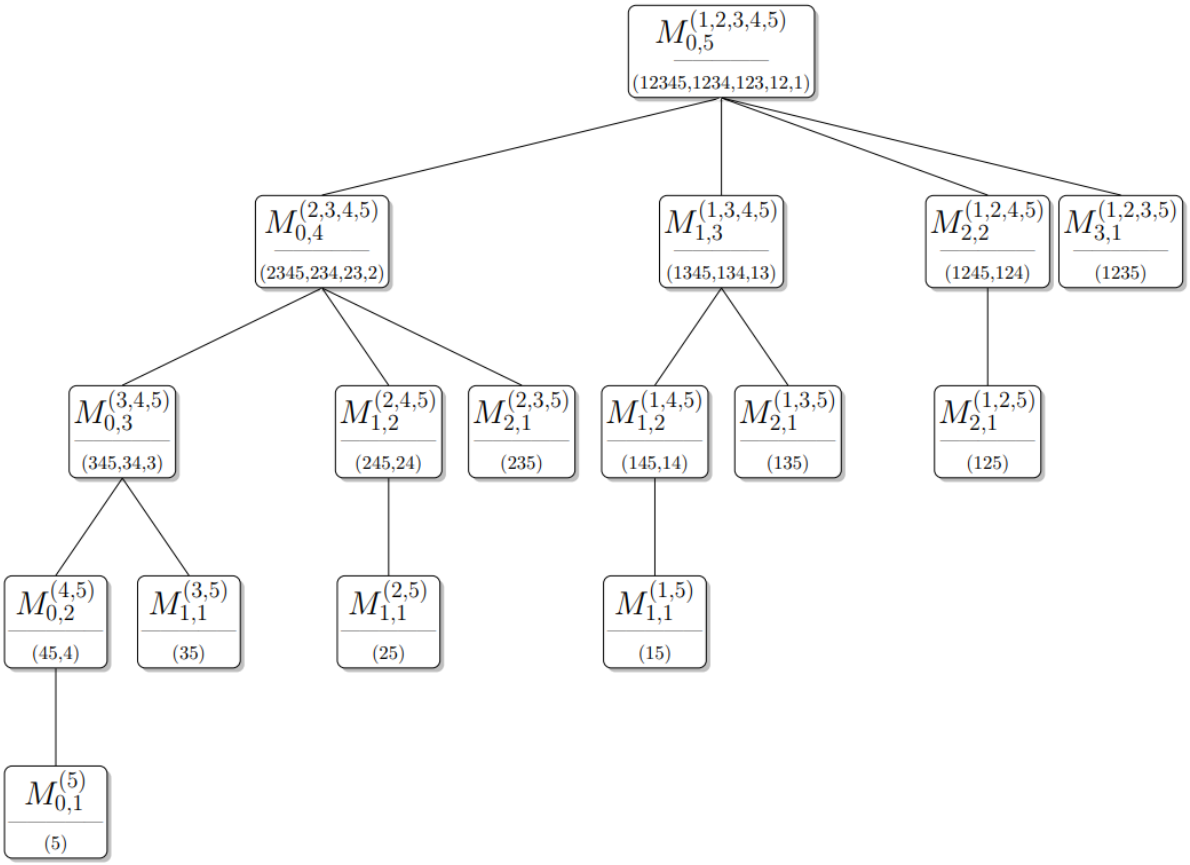
$$\boldsymbol{R} \equiv \begin{bmatrix} x & x & x & x & x \\ 0 & x & x & x & x \\ 0 & 0 & x & x & x \\ 0 & 0 & 0 & x & x \\ 0 & 0 & 0 & 0 & x \end{bmatrix} \xrightarrow{\textbf{Διαγραφή 2ης στήλης}} \begin{bmatrix} x & x & x & x \\ 0 & \boldsymbol{x} & \boldsymbol{x} & \boldsymbol{x} \\ 0 & \boldsymbol{x} & \boldsymbol{x} & \boldsymbol{x} \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \end{bmatrix} \xrightarrow{G_{2,3}^{(2)}} \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & \boldsymbol{x} & \boldsymbol{x} \\ 0 & 0 & \boldsymbol{x} & \boldsymbol{x} \\ 0 & 0 & 0 & x \end{bmatrix} \rightarrow$$

$$\xrightarrow{G_{3,4}^{(3)}} \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & \boldsymbol{x} \\ 0 & 0 & 0 & \boldsymbol{x} \end{bmatrix} \xrightarrow{G_{4,5}^{(4)}} \begin{bmatrix} x & x & x & x \\ 0 & x & x & x \\ 0 & 0 & x & x \\ 0 & 0 & 0 & x \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

Σχήμα 5 - Τριγωνοποίηση ενός $\boldsymbol{p \times p}$ τριγωνικού πίνακα έπειτα από διαγραφή της k-οστής στήλης, χρησιμοποιώντας περιστροφές Givens (για $\mathbf{p = 5}, \boldsymbol{k = 2}$)

Η εφαρμογή του αλγόριθμου DCA στο μοντέλο παλινδρόμησης (1.4) είναι ισοδύναμη με τον «αριστερότερο περίπατο» (leftmost walk) στο δέντρο παλινδρόμησης $T_{0,p}^{\upsilon}$, όπου $M_{0,p}^{\upsilon} \equiv \boldsymbol{R}$ από την παραγοντοποίηση (3.10) και $\upsilon_i = i,\ \ i = 1,2,\dots,p$. Στο Σχήμα 6 παρουσιάζεται το δέντρο $T_{0,p}^{\upsilon}$, για $\mathrm{p} = 5$ και $\upsilon = (1,\dots,5)$ καθώς και τα υπομοντέλα που μπορούν να εξαχθούν από κάθε κόμβο. Ένα υπομοντέλο ορίζεται από μία ακολουθία αριθμών που αντιστοιχεί στους δείκτες των μεταβλητών που περιέχονται στο μοντέλο. Για την εξαγωγή ενός παιδιού-κόμβου από κάποιον γονέα-κόμβο χρησιμοποιούνται οι διαδικασίες $Drop$ και $Shift$. Η διαδικασία του Σχήματος 5 (*Σχήμα 5*) αντιστοιχεί στην εφαρμογή της διαδικασίας $Drop$ στο $M_{1,4}^{\upsilon}$ η οποία επιστρέφει το $M_{1,3}^{\upsilon^{(2)}}$. Δεδομένου του $M_{k,\lambda}^{\upsilon}$, η διαδικασία $Shift$ επιστρέφει το $M_{k+1,\lambda-1}^{\upsilon}$. Πιο συγκεκριμένα, ένα $Shift$ στο $M_{k,\lambda}^{\upsilon}$ αυξάνει το δείκτη της πρώτης στήλης που θα διαγραφεί, αυξάνοντας το k και ταυτόχρονα μειώνει και το λ. Από ένα γονέα-κόμβο $M_{k,\lambda}^{\upsilon}$ μπορούμε να εξάγουμε το i-οστό παιδί κόμβο εφαρμόζοντας $(i-1)\ Shifts$ και στη συνέχεια εφαρμόζεται ένα $Drop$. Για παράδειγμα στο Σχήμα 6 εξάγουμε το παιδί-κόμβο $M_{1,2}^{(2,4,5)}$ εφαρμόζοντας ένα $Shift$ και ένα $Drop$ στον γονέα-κόμβο $M_{0,4}^{(2,3,4,5)}$. Συνεπώς, τα υπομοντέλα που μπορούν να εξαχθούν από το υπο-δέντρο $T_{k+i-1,\lambda-i}^{\upsilon^{(k+i)}}$ θα περιέχουν πάντα τις μεταβλητές $\upsilon_1,\dots,\upsilon_{k+i-1}$.

Η διαδικασία $SubTree$ στον Αλγόριθμο 10 (*Αλγόριθμος 10*) παράγει το δέντρο παλινδρόμησης $\mathrm{T}_{k,\lambda}^{\upsilon}$ όταν δώσουμε σαν παράμετρο εισόδου τον κόμβο-ρίζα $M_{k,\lambda}^{\upsilon}$. Επομένως, ο αλγόριθμος DCA για το μοντέλο παλινδρόμησης p παραμέτρων (1.4) είναι ισοδύναμος με το $SubTree\left(M_{0,p}^{\upsilon}\right)$ όπου $\upsilon = (1,2,\dots,p)$ και $M_{0,p}^{\upsilon} \equiv \boldsymbol{R}$ από την παραγοντοποίηση (3.10).



Σχήμα 6 – Δέντρο Παλινδρόμησης $T_{0,p}^{\upsilon}$ για $p = 5$ και $\upsilon = (1,\dots,5)$

<u>Αλγόριθμος 10 – Παραγωγή δέντρου παλινδρόμησης</u> (Gatu & Kontoghiorghes, 2003)

Ο παρακάτω αλγόριθμος παράγει το δέντρο παλινδρόμησης $T_{k,\lambda}^{\upsilon}$ από τον κόμβο-ρίζα $M_{k,\lambda}^{\upsilon}$

procedure $SubTree\left(M_{k,\lambda}^{\upsilon}\right)$

From $M_{k,\lambda}^{\upsilon}$ obtain the RSS of the submodels $(\upsilon_1, \dots, \upsilon_{k+1}), \dots, (\upsilon_1, \dots, \upsilon_{k+\lambda})$

for $i = 1, \dots, \lambda - 1$ **do**

Store $M_{k,\lambda}^{\upsilon}$

$M_{k+i-1,\lambda-i+1}^{\upsilon} \leftarrow$ Apply $i-1$ $Shifts$ on $M_{k,\lambda}^{\upsilon}$

$M_{k+i-1,\lambda-i}^{\upsilon} \leftarrow$ Apply $Drop$ on $M_{k+i-1,\lambda-i+1}^{\upsilon}$

$SubTree\left(M_{k+i-1,\lambda-i}^{\upsilon}\right)$

end for

end procedure

Ο αριθμός των πράξεων που εκτελούνται για την κατασκευή του δέντρου $T_{k,\lambda}^{\upsilon}$ με κόμβο-ρίζα $M_{k,\lambda}^{\upsilon}$ δίνεται από τη σχέση (Gatu & Kontoghiorghes, 2003)

$$C(\lambda) = 3t2^{\lambda} - \frac{t(\lambda+2)(\lambda+3)}{2}$$

όπου $t = 6$ είναι ο αριθμός των flops που χρειάζονται για να κατασκευάσουμε μία περιστροφή Givens. Επομένως, ο αριθμός των πράξεων για τον αλγόριθμο DCA είναι $O(2^p)$ και πιο συγκεκριμένα δίνεται από τη σχέση

$$C_{DCA}(p) = C(p) \approx 3t2^p$$

Ο Αλγόριθμος BBA

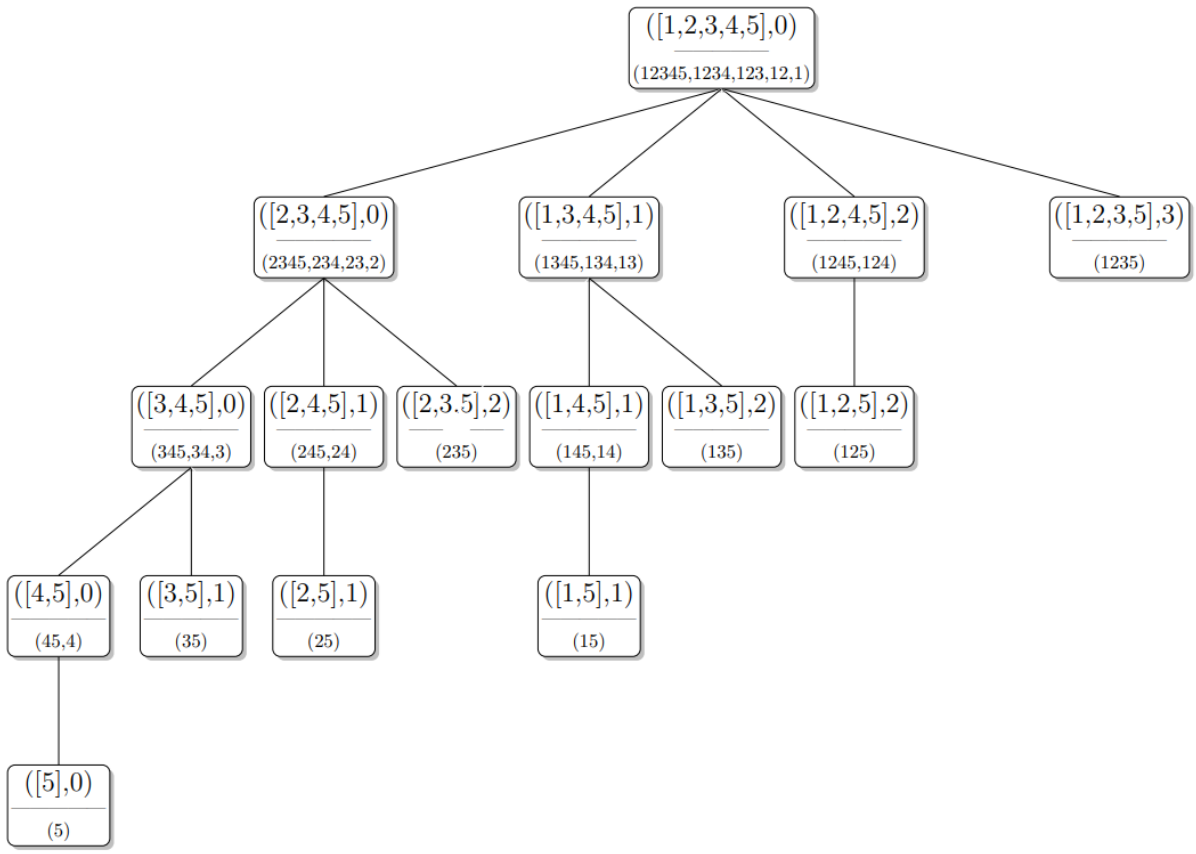
Ο αλγόριθμος BBA βασίζεται στην κατασκευή του δέντρου παλινδρόμησης χρησιμοποιώντας τη διαδικασία DCA που περιγράψαμε παραπάνω. Για χάρη συμφωνίας συμβολισμών, θα περιγράψουμε την κατασκευή του δέντρου παλινδρόμησης σύμφωνα με το συμβολισμό που χρησιμοποιείται από τους Gatu and Kontoghiorghes (2006).

Ένας κόμβος του δέντρου παλινδρόμησης είναι μία πλειάδα (V, k) όπου V είναι ένα σύνολο δεικτών μεταβλητών και το k $(k = 0, \dots, |V| - 1)$ υποδεικνύει ότι τα παιδιά-κόμβοι αυτού του κόμβου θα περιέχουν τις πρώτες k μεταβλητές. Αν $V = [\upsilon_1, \upsilon_2, \dots, \upsilon_p]$ με $\upsilon_i = i, i = 1,2, \dots, p,$ τότε το δέντρο παλινδρόμησης $T(V, k)$ είναι ένα $(p-1)$-δέντρο με κόμβο-ρίζα τον (V, k) για $k = 0, \dots, p-1$. Τα παιδιά-κόμβοι ορίζονται από τις πλειάδες $(Drop(V, i), i-1)$ για $i = k+1, \dots, p-1$ όπου

$$Drop(V,i) = [v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_p]$$

Συνοπτικά, ορίζουμε

$$T(V,k) = \begin{cases} (V,k), & \text{αν } k = p-1 \\ \big((V,k), T(Drop(V,k+1),k), \dots, T(Drop(V,p-1),p-2)\big), & \text{αν } k < p-1 \end{cases}$$



Σχήμα 7 – Δέντρο παλινδρόμησης $T(V,k)$ για $V = [1,2,3,4,5], k = 0$

Ο υπολογισμός όλων των πιθανών υπομοντέλων παλινδρόμησης, αρχίζοντας με ένα μοντέλο p παραμέτρων, ισοδυναμεί με την κατασκευή του δέντρου παλινδρόμησης $T(V,0)$, όπου $V = [1,2,\dots,p]$. Χρησιμοποιώντας τον παραπάνω συμβολισμό και για $V = [1,2,3,4,5]$, $k = 0$ στο Σχήμα 7 φαίνεται το δέντρο $T(V,k)$ το οποίο ταυτίζεται με το δέντρο του Σχήματος 6 (*Σχήμα 6*).

Όπως και ο αλγόριθμος Leaps-and-Bounds, έτσι και ο αλγόριθμος BBA είναι βασισμένος στην ιδιότητα

$$RSS(M_1) \leq RSS(M_0)$$

όπου M_0, M_1 δύο μοντέλα παλινδρόμησης με $M_0 \subseteq M_1$ υπό την έννοια ότι όλες οι μεταβλητές του M_0 περιέχονται και στο M_1.

Χρησιμοποιώντας την παραπάνω ιδιότητα μπορούμε να περιορίσουμε το πλήθος των μοντέλων που πρέπει να αξιολογήσουμε κατά την αναζήτησή μας για τα καλύτερα μοντέλα. Θεωρώντας $V = [1,2,\dots,p]$ ορίζουμε $r_j^{(g)}$ $(j = 1,\dots,p$ και $g = 1,\dots,2^{p-1})$ να αντιπροσωπεύει την τρέχουσα ελάχιστη τιμή RSS για τα μοντέλα με j παραμέτρους μετά τον υπολογισμό των πρώτων g κόμβων του δέντρου παλινδρόμησης $T(V,0)$. Σημειώνουμε ότι η σειρά κατασκευής των κόμβων δεν είναι

σημαντική. Ο κόμβος-ρίζα $(V, 0)$ παρέχει τις τιμές $r_1^{(1)}, r_2^{(1)}, \dots, r_p^{(1)}$ των RSS για τα μοντέλα $[1], [1,2], \dots, [1,2, \dots, p]$, αντίστοιχα. Μετά τον υπολογισμό του g-οστού κόμβου $([v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_d], i-1)$, τα μοντέλα $[v_1, \dots, v_{i-1}, v_{i+1}], \dots, [v_1, \dots, v_{i-1}, v_{i+1}, \dots, v_d]$ είναι διαθέσιμα και οι τιμές $r_j^{(g)}$ $(j = i, \dots, d-1)$ ενημερώνονται. Για $j = 1, \dots i-1$ και $j = d, \dots, p$ οι τιμές $r_j^{(g)}$ δεν ενημερώνονται και ισχύει $r_j^{(g)} = r_j^{(g-1)}$. Αφού κατασκευαστεί ολόκληρο το δέντρο παλινδρόμησης $T(V, 0)$, το ελάχιστο RSS που αντιστοιχεί στα καλύτερα μοντέλα παλινδρόμησης, για κάθε πλήθος παραμέτρων, δίνεται από τα $r_1^{(2^{p-1})}, \dots, r_p^{(2^{p-1})}$.

<u>Λήμμα 1</u> (Gatu & Kontoghiorghes, 2006)

$$r_j^{(g)} \geq r_{j+1}^{(g)}$$

όπου $g = 1, \dots, 2^{p-1}$ και $j = 1, \dots, p-1$

<u>Λήμμα 2</u> (Gatu & Kontoghiorghes, 2006)

$$\mathrm{r}_{|W|}^{(g)} \leq RSS(W)$$

όπου W είναι κάποιο υποσύνολο (υπομοντέλο) που λαμβάνεται από κάποιο κόμβο του δέντρου $T(V, i-1)$ με $r_i^{(g)} \leq RSS(V)$ και $1 \leq i < |V|$.

<u>Πρόταση 1</u> (Gatu & Kontoghiorghes, 2006)

$$(1-\tau) r_{|W|}^{(g)} \leq RSS(W)$$

όπου W είναι κάποιο υποσύνολο (υπομοντέλο) που λαμβάνεται από κάποιο κόμβο του δέντρο $T(V, i-1)$, $0 \leq \tau < 1$, $(1-\tau) r_i^{(g)} \leq RSS(V)$ και $1 \leq i < |V|$.

Στη συνέχεια θα παρουσιάσουμε τη διαδικασία που ακολουθεί ο αλγόριθμος BBA ώστε να περιορίσει τον αριθμό των μοντέλων που θα εξεταστούν. Θα χρησιμοποιήσουμε το δέντρο του Σχήματος 7 (*Σχήμα 7*) ώστε να γίνει πιο κατανοητή η διαδικασία και στη συνέχεια θα δώσουμε ένα πιο συγκεκριμένο, αριθμητικό παράδειγμα.

1. Έστω τώρα ότι (V, k) είναι ένας από τους g κόμβους που έχουμε υπολογίσει για το δέντρο παλινδρόμησης $T([1,2, \dots, p], 0)$ όπου $0 \leq k < |V| \leq p$ και $1 \leq g \leq 2^{p-1}$.

   <u>π.χ</u>. Για $V = [2,3,4,5], k = 0, p = 5$,
   Θεωρούμε ότι έχουμε υπολογίσει τον κόμβο $(V, k) \equiv ([2,3,4,5], 0)$ του δέντρου $T([1,2,3,4,5], 0)$.

2. Για το δέντρο $T(V, k)$ απομένει να υπολογίσουμε τα υπο-δέντρα $T\big((Drop(V, i), i-1)\big)$ για $i = k+1, \dots, |V|-1$.

π.χ. Για $T(V,k) \equiv T([2,3,4,5],0)$ απομένει να υπολογίσουμε τα υπο-δέντρα $T(Drop(V,1),0) \equiv T([3,4,5],0)$, $T(Drop(V,2),1) \equiv T([2,4,5],1)$ και $T(Drop(V,3),2) \equiv T([2,3,5],2)$

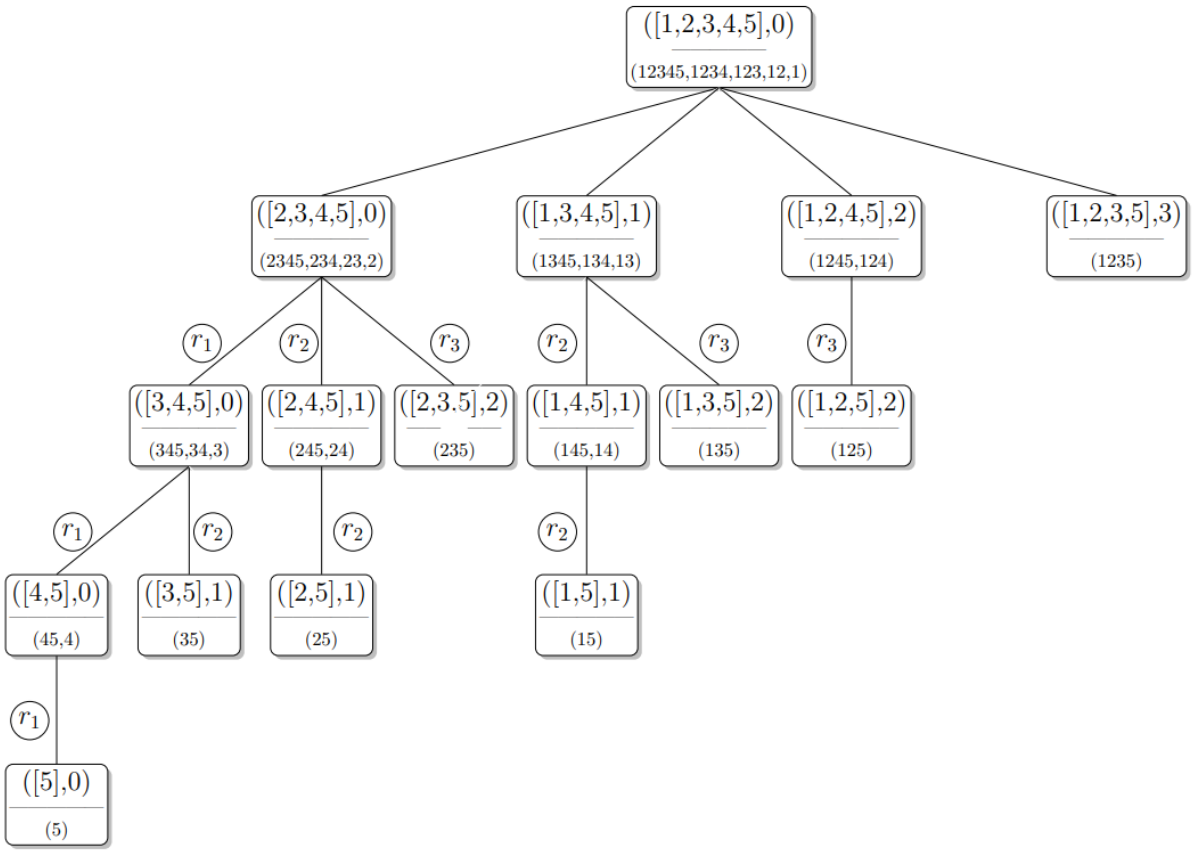
3. Αν $r_{(k+1)}^{(g)} \leq b_V$ τότε, από το Λήμμα 2, έχουμε ότι $r_{|W|}^{(g)} \leq RSS(W)$, όπου W είναι κάποιο υπομοντέλο που λαμβάνουμε από κάποιο κόμβο του δέντρου $T(V,k)$. Τα υπο-δέντρα του $T(V,k)$ είναι τα $T(Drop(V,i),i-1)$ όπου $i = k+1,\dots,|V|-1$. Αυτό υποδηλώνει ότι τα υπόλοιπα υπο-δέντρα του κόμβου (V,k) δεν μπορούν να βελτιώσουν τις τιμές $r_j^{(g)}$ όπου $j = 1,\dots,p$.
Εκτελούμε τον παρακάτω βρόγχο επαναλήψεων:

Για $i = k+1,\dots,|V|-1$,
Προτού υπολογιστεί ο κόμβος $(Drop(V,i),i-1)$, συγκρίνουμε την τιμή $r_i^{(g)}$ με το φράγμα b_V. Συγκεκριμένα,

- ➔ Εάν $r_i^{(g)} \leq b_V$, τότε όλα τα υπο-δέντρα $T(Drop(V,\rho),\rho-1)$, $\rho = i,\dots,|V|-1$, δε χρειάζεται να υπολογιστούν και ο βρόγχος τερματίζει έχοντας "κλαδέψει" αυτά τα υπο-δέντρα.
- ➔ Διαφορετικά, υπολογίζεται ο κόμβος $(Drop(V,i),i-1)$ και οι τιμές $r_j^{(g+1)}$ $(j = i,\dots,|V|-1)$ ενημερώνονται.

Η παραπάνω διαδικασία επαναλαμβάνεται για τον επόμενο παιδί-κόμβο $(Drop(V,i+1),i)$, δηλαδή αυξάνοντας την τιμή του i κατά 1.

Στο Σχήμα 8 παρουσιάζουμε το δέντρο $T([1,2,3,4,5],0)$ καθώς και το εκάστοτε $r_j^{(g)}$ το οποίο συγκρίνουμε με το φράγμα b_V. Για παράδειγμα, για να ελέγξουμε αν χρειάζεται να υπολογιστεί ο κόμβος $([3,4,5],0)$, και ως επακόλουθο και οι κόμβοι $([2,4,5],1)$ και $([2,3,5],2)$, συγκρίνουμε την τιμή $b_V = RSS([2,3,4,5])$ με την τρέχουσα τιμή r_1.



Σχήμα 8 - Δέντρο παλινδρόμησης $T(V,k)$ για $V=[1,2,3,4,5], k=0$ συμπεριλαμβάνοντας τα $r_j^{(g)}$

π.χ. Όπως έχουμε πει έχουμε υπολογίσει τον κόμβο $(V,k) \equiv ([2,3,4,5],0)$. Εκτελούμε τον ακόλουθο βρόγχο επαναλήψεων για $i=1,2,3$.

Για $i=1$,

Προτού υπολογιστεί ο κόμβος $(Drop(V,1),0) \equiv ([3,4,5],0)$ ελέγχουμε εάν $r_1^{(g)} \leq b_V \equiv RSS([2,3,4,5])$. Δηλαδή αν

$$r_1^{(g)} \leq RSS([2,3,4,5])$$

Τότε αφού

$$RSS([2,3,4,5]) \leq RSS(W), \quad \forall\, W \subseteq [2,3,4,5]$$

από το Λήμμα 2 έχουμε ότι

$$r_{|W|}^{(g)} \leq r_1^{(g)} \leq RSS([2,3,4,5]) \leq RSS(W), \forall\, W \subseteq [2,3,4,5]$$

που σημαίνει δεν υπάρχει υπομοντέλο που παράγεται από το δέντρο $T([2,3,4,5],0)$ το οποίο μπορεί να βελτιώσει τις τιμές $r_j^{(g)}, j=1,2,\dots,5$. Αυτό σημαίνει ότι δε χρειάζεται να υπολογίσουμε τα δέντρα $T([3,4,5],0), T([2,4,5],1)$ και $T([2,3,5],2)$ επομένως τα κλαδεύουμε και ο βρόγχος επαναλήψεων διακόπτεται.

Διαφορετικά, εάν $r_1^{(g)} > b_V \equiv RSS([2,3,4,5])$ τότε υπολογίζουμε τον κόμβο $([3,4,5],0)$ και εφαρμόζουμε την διαδικασία του BBA για να κρίνουμε εάν χρειάζεται να υπολογιστούν τα υπο-δέντρα του κόμβου αυτού. Επίσης, σε αυτό το βήμα είμαστε ακόμα μέσα στο βρόγχο

επανάληψης όπου τώρα έχουμε $i = 2$ και θα ασχοληθούμε με το αν πρέπει να υπολογίσουμε τον κόμβο $([2,4,5], 1)$ ελέγχοντας αν $r_2^{(g)} \leq b_V \equiv RSS([2,3,4,5])$ κάτι που θα σήμαινε ότι κλαδεύουμε τα δέντρα $, T([2,4,5], 1)$ και $T([2,3,5], 2)$. Διαφορετικά προχωρούμε στο επόμενο βήμα του βρόγχου για $i = 3$ και ελέγχουμε αν πρέπει να υπολογιστεί ο κόμβος $([2,3,5], 2)$ χρησιμοποιώντας τον έλεγχο $r_3^{(g)} \leq b_V \equiv [2,3,4,5]$ και έτσι φτάνουμε στο τέλος του βρόγχου.

Όπως έχουμε δείξει στην παραπάνω διαδικασία, η σειρά επεξεργασίας των κόμβων $(Drop(V, k+1), k)$,..., $(Drop(V, |V|-1), |V|-2)$ είναι η πιο αποδοτική αφού εάν το υπο-δέντρο $T(Drop(V, i), i-1)$ κλαδευτεί τότε και τα υπόλοιπα υπο-δέντρα $T(Drop(V, \rho), \rho - 1)$ για $\rho = i, \dots, |V| - 1$ θα κλαδευτούν.

Η διαδικασία BBA η οποία βρίσκει τα καλύτερα μοντέλα παλινδρόμησης περιγράφεται από τον Αλγόριθμο 11 (*Αλγόριθμος 11*). Ο αλγόριθμος χρησιμοποιεί την επαναληπτική διαδικασία $ProcessSubtree$ η οποία επεξεργάζεται ένα υπο-δέντρο παλινδρόμησης με κόμβο-ρίζα τον (V, k).

<u>Αλγόριθμος 11 - Διαδικασία BBA για την εύρεση των καλύτερων μοντέλων παλινδρόμησης</u> (Gatu & Kontoghiorghes, 2006)

1. Υπολόγισε την παραγοντοποίηση QR του πίνακα $\boldsymbol{X} \in \mathbb{R}^{n \times p}$: $\boldsymbol{Q}^T \boldsymbol{X} = \begin{pmatrix} \boldsymbol{R} \\ \boldsymbol{0} \end{pmatrix} \begin{matrix} p \\ n-p \end{matrix}$ και $\boldsymbol{Q}^T \boldsymbol{y} = \tilde{\boldsymbol{y}}$
2. Έστω $V = [1,2, \dots, p], k = 0$ και $r_j = RSS([1,2, \dots, j]) \equiv \sum_{i=j+1}^{n} \tilde{y}_i^{\,2}$ όπου $j = 1, \dots, p$
3. **call** $ProcessSubtree(V, k)$
4. **def** $ProcessSubtree(V, k)$ = **do**
5. **for** $i = k+1, \dots, |V|-1$ **do**
6. **if** $\left(r_i > RSS(V)\right)$ **then**
7. $V^{(i)} \leftarrow Drop(V, i)$
8. $r_j = \min\left(r_j, RSS\left(\left[v_1^{(i)}, v_2^{(i)}, \dots, v_j^{(i)}\right]\right)\right)$, όπου $j = i, \dots, |V|-1$
9. **else**
10. Κλάδεψε τα υπόλοιπα υπο-δέντρα και πήγαινε στο Βήμα 13
11. **end if**
12. **end for**
13. **call** $ProcessSubtree\left(V^{(j)}, j-1\right)$, όπου $j = k+1, \dots, \min(i, |V|-2)$
14. **end def**

Η αποδοτικότητα του αλγόριθμου BBA μπορεί να βελτιωθεί όταν κλαδεύονται περισσότεροι κόμβοι, δηλαδή όταν μεγαλύτερα υπο-δέντρα φράσσονται από μεγαλύτερες τιμές (Furnival & Wilson, 1974; Gatu & Kontoghiorghes, 2006). Αυτό μπορεί να επιτευχθεί εάν ταξινομήσουμε (preorder) τις μεταβλητές στο αρχικό σύνολο ώστε για τον κόμβο $(V, k) = ([1,2, \dots, p], 0)$ να ισχύει

$$RSS\big(Drop(V,1)\big) \geq RSS\big(Drop(V,2)\big) \geq \cdots \geq RSS\big(Drop(V,p-1)\big) \geq RSS(V)$$

Δηλαδή, ταξινομούμε τις μεταβλητές ώστε η περισσότερο και η λιγότερο σημαντική μεταβλητή εμφανίζεται στην πρώτη και στην τελευταία θέση του V, αντίστοιχα. Τον αλγόριθμο BBA που χρησιμοποιεί την παραπάνω ταξινόμηση συμβολίζουμε με BBA-1. Με μία άλλη προσέγγιση, τον

αλγόριθμο BBA-2, μπορεί να κλαδέψουμε μεγαλύτερα υπο-δέντρα σε νωρίτερο στάδιο της διαδικασίας. Ο BBA-2 επεξεργάζεται πρώτα τους κόμβους με το μικρότερο φράγμα. Δηλαδή, αν συμβολίσουμε με $(V_1, k_1), \dots, (V_d, k_d)$, $(0 < d \leq 2^{p-1})$ τους κόμβους που μπορούμε να επεξεργαστούμε σε κάποιο στάδιο της διαδικασίας BBA, τότε ο BBA-2 θα επεξεργαστεί πρώτα τον κόμβο που αντιστοιχεί στο φράγμα $\min\big(RSS(V_1), \dots, RSS(V_d)\big)$. Ο αλγόριθμος BBA μπορεί να τροποποιηθεί ώστε να επιστρέφει μία λίστα με τα καλύτερα μοντέλα για κάθε πλήθος μεταβλητών, αντί να επιστρέφει μόνο ένα μοντέλο για κάθε πλήθος μεταβλητών. Σε αυτήν την περίπτωση, η τρέχουσα ελάχιστη τιμή r_j που χρησιμοποιεί ο BBA αντικαθίσταται από την r_{jn} $(n = 1, \dots, \gamma)$, όπου γ είναι το πλήθος των καλύτερων μοντέλων μεγέθους j $(j = 1,2, \dots, p)$ μεταβλητών που θέλουμε να επιστρέφουμε. Ο έλεγχος κλαδέματος branch-and-bound γίνεται χρησιμοποιώντας την υποδεέστερη τιμή που έχουμε για τη λίστα των καλύτερων μοντέλων μεγέθους j, δηλαδή την $\max\big(r_{j1}, \dots, r_{j\gamma}\big)$.

Όταν το πλήθος των μεταβλητών είναι αρκετά μεγάλο η εφαρμογή του BBA ενδέχεται να είναι υπολογιστικά ανέφικτη. Σε αυτές τις περιπτώσεις μπορούμε να χρησιμοποιήσουμε διάφορες ευρετικές μεθόδους (heuristic methods) για να υπολογίσουμε προσεγγιστικά τη βέλτιστη λύση (Gatu & Kontoghiorghes, 2006). Μία τέτοια μέθοδος είναι η HBBA σύμφωνα με την οποία για την αποκοπή υπο-δέντρων χρησιμοποιούμε μία παράμετρο ανοχής δηλαδή χρησιμοποιούμε τον κανόνα $(1 - \tau)r_i^{(g)} \leq RSS(V)$ όπου τ είναι η παράμετρος ανοχής με $0 \leq \tau < 1$. Για $\tau = 0$ ο αλγόριθμος HBBA ταυτίζεται με τον αλγόριθμο BBA. Όσο πιο κοντά στο 0 είναι η τιμή της τ τόσο μεγαλύτερη είναι η πιθανότητα να λάβουμε λύση η οποία είναι κοντά στη βέλτιστη.

Εάν υπολογιστούν όλοι οι κόμβοι του δέντρου παλινδρόμησης τότε ο αλγόριθμος BBA ισοδυναμεί με τον αλγόριθμο DCA. Σε αυτήν την περίπτωση, συμβολίζοντας με LBA τον αλγόριθμο Leaps-and-Bounds (Furnival & Wilson, 1974) και C_{LBA}, C_{BBA} την πολυπλοκότητα των αλγορίθμων LBA και BBA αντίστοιχα, μπορούμε να συγκρίνουμε το άνω φράγμα για τις πολυπλοκότητες αυτών των αλγορίθμων (Gatu & Kontoghiorghes, 2006) βρίσκοντας ότι

$$\frac{C_{LBA}(p)}{C_{BBA}(p)} \approx 0.0058 * (2p^3 + 15p^2 + 13p - 6) \equiv O(p^3)$$

Συγκρίνοντας τους υπολογιστικούς χρόνους για τους αλγόριθμους LBA και BBA (Gatu & Kontoghiorghes, 2006) μπορούμε να δούμε ότι ο BBA είναι κατά 4 έως 30 φορές πιο αποδοτικός από τον LBA και ότι χρησιμοποιώντας ταξινόμηση, όπως αναφέραμε προηγουμένως, ο αλγόριθμος BBA-1 βελτιώνει σημαντικά τον υπολογιστικό χρόνο.

Το πακέτο lmSubsets και η διαδικασία lmSelect

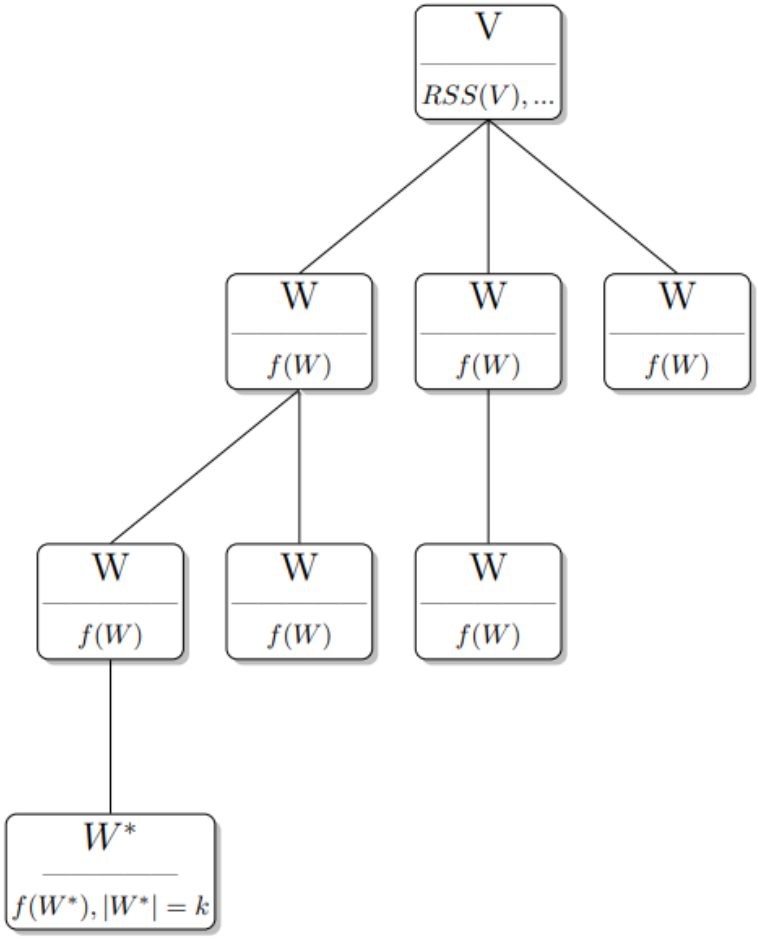
Οι Hofmann et al. (2020) δημιούργησαν το πακέτο $lmSubsets$, για την γλώσσα R, το οποίο, μεταξύ άλλων, δίνει τη δυνατότητα εφαρμογής των αλγόριθμων BBA, HBBA (Gatu & Kontoghiorghes, 2006) και των αλγόριθμων RangeBBA, RadiusBBA (Hofmann, Gatu, & Kontoghiorghes, 2007). Οι παραπάνω αλγόριθμοι αποτελούν παραλλαγές του αλγόριθμου BBA. Επιπρόσθετα, αναπτύσσεται η διαδικασία $lmSelect$ η οποία είναι διαθέσιμη στο πακέτο $lmSubsets$ και θα περιγράψουμε στη συνέχεια.

Η διαδικασία $lmSelect$ αποτελεί ένα αλγόριθμο επιλογής του βέλτιστου μοντέλου παλινδρόμησης χρησιμοποιώντας ως κριτήριο επιλογής κάποιο κριτήριο της οικογένειας IC (Information Criteria), σε αντίθεση με τους υπόλοιπους αλγόριθμους που έχουμε περιγράψει μέχρι στιγμής οι οποίοι

χρησιμοποιούν το άθροισμα τετραγώνων των υπολοίπων, RSS, ως κριτήριο. Η οικογένεια των κριτηρίων IC περιγράφεται από τη σχέση

$$\begin{aligned} \mathrm{AIC_k} &= n + n\log 2\pi + n\log\left(\frac{RSS}{n}\right) + k(p+1) \\ &= n\left[\log\left(\frac{2\pi RSS}{n}\right) + 1\right] + k(p+1) \end{aligned} \tag{3.13}$$

όπου n είναι το πλήθος των παρατηρήσεων, p είναι το πλήθος των παραμέτρων του μοντέλου και k $(k > 0)$ είναι η ποινή για κάθε παράμετρο του μοντέλου. Σημειώνουμε ότι για $k = 2$ και $k = \log n$ λαμβάνουμε τα κριτήρια AIC και BIC, αντίστοιχα, που περιγράψαμε στις ενότητες 3.1.3 και 3.1.4.



Σχήμα 9 – Υπο-δέντρο παλινδρόμησης

Ο αλγόριθμος $lmSelect$ έχει παρόμοια λειτουργία με τον αλγόριθμο BBA. Για την εύρεση του βέλτιστου μοντέλου εργαζόμαστε στο δέντρο παλινδρόμησης που παράγεται χρησιμοποιώντας τη διαδικασία DCA. Για να εξοικονομήσουμε υπολογιστικό κόστος, ο αλγόριθμος $lmSelect$ εφαρμόζει παρόμοια στρατηγική αποκοπής υποδέντρων όπως αυτήν που χρησιμοποιεί ο αλγόριθμος BBA. Αρχικά, ας παρατηρήσουμε ότι, για δεδομένη τιμή για τη μεταβλητή k, το κριτήριο AIC_k είναι μία συνάρτηση με παραμέτρους το RSS και το p και επίσης ότι είναι μονότονη ως προς την κάθε μία από αυτές τις παραμέτρους ξεχωριστά. Θέτουμε $AIC_k \equiv f(RSS_p, p)$ και $f^* = f(RSS^*_{p^*}, p^*)$ είναι η βέλτιστη τιμή που έχουμε βρει μέχρι στιγμής για το κριτήριο f. Θα εξηγήσουμε τον έλεγχο αποκοπής χρησιμοποιώντας το υποδέντρο του Σχήματος 9 (*Σχήμα 9*). Αρχικά υπολογίζουμε τον κόμβο-ρίζα

αυτού του υποδέντρου χρησιμοποιώντας το σύνολο μεταβλητών V. Συμβολίζουμε οποιοδήποτε μοντέλο παράγεται από αυτόν τον κόμβο ως $f(W)$ όπου W είναι το σύνολο μεταβλητών που χρησιμοποιείται από τον εκάστοτε κόμβο. Όπως έχουμε αναφέρει και στον αλγόριθμο BBA, θα ισχύει ότι $RSS(V) \leq RSS(W)\ \forall\ W$, αφού $W \subset V$. Θέτουμε ως k το πλήθος των μεταβλητών του κατώτερου κόμβου που βρίσκεται στο αριστερότερο μονοπάτι του υποδέντρου που μελετούμε. Το πλήθος των μεταβλητών του μοντέλου που παράγεται από αυτόν τον κόμβο είναι μικρότερο ή ίσο με το πλήθος των μεταβλητών οποιουδήποτε άλλου μοντέλου του υποδέντρου. Δηλαδή έχουμε ότι $k \leq |W|$. Βάσει των παραπάνω έχουμε βρει ένα κάτω φράγμα για όλα τα μοντέλα του υποδέντρου. Πιο συγκεκριμένα έχουμε ότι

$$f(RSS(V), k) \leq f(RSS(W), |W|) \quad \forall\ W$$

Επομένως, αν

$$f^* \leq f(RSS(V), k)$$

τότε

$$f^* \leq f(RSS(W), |W|)\ \ \forall\ W$$

που σημαίνει ότι κανένα μοντέλο του υπό μελέτη υποδέντρου δεν βελτιώνει την βέλτιστη τιμή f^* που έχουμε υπολογίσει μέχρι στιγμής και επομένως το υποδέντρο αυτό δεν υπολογίζεται, δηλαδή το κλαδεύουμε.

Στην περίπτωση όπου

$$f^* > f(RSS(V), k)$$

τότε υπολογίζουμε τον επόμενο κόμβο στη σειρά και εφαρμόζουμε τον έλεγχο αποκοπής.

Οι Hofmann et al. (2020) χρησιμοποίησαν προσομοιωμένα δεδομένα για να συγκρίνουν, μεταξύ άλλων, τη μέθοδο $lmSelect$ και τη μέθοδο Leaps-and-Bounds (LBA) με τις διάφορες μεθόδους του πακέτου $lmSubsets$. Τα αποτελέσματα έδειξαν ότι οι μέθοδοι του $lmSubsets$ είναι πιο αποδοτικοί, ως προς το χρόνο υπολογισμού, σε σύγκριση με τη μέθοδο LBA ενώ η μέθοδος $lmSelect$ βελτίωνε ακόμα περισσότερο την αποδοτικότητα σε σύγκριση με το $lmSubsets$.

Εφαρμογή στην R

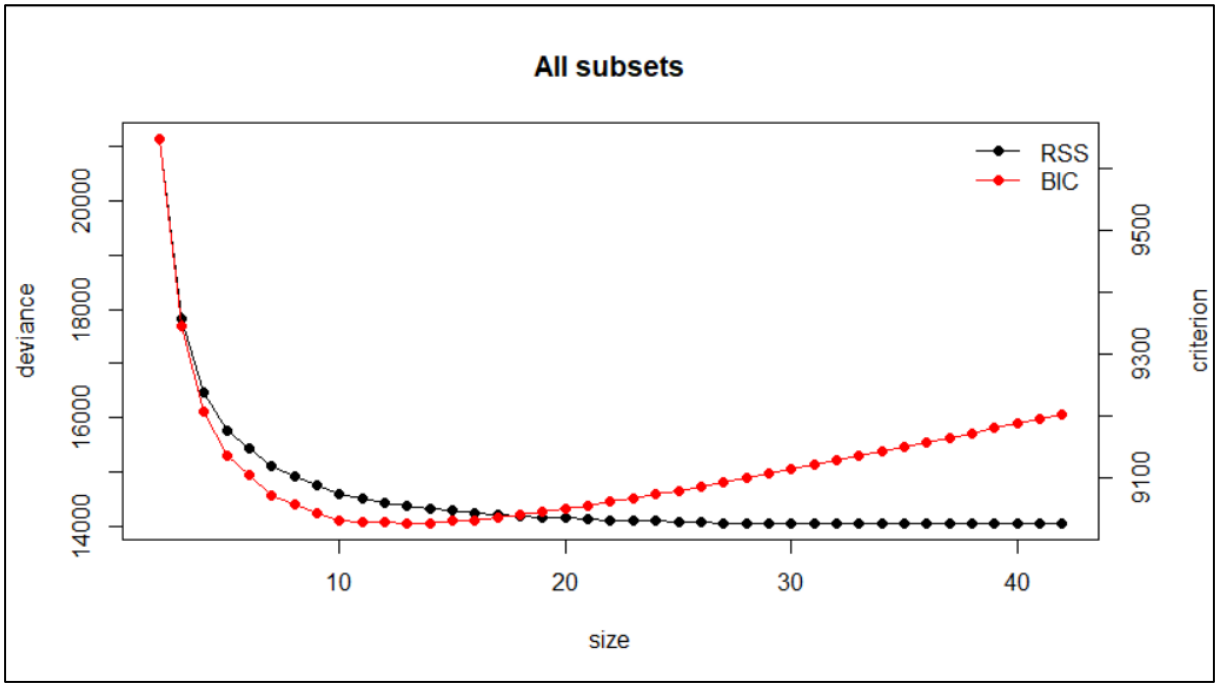
Το πακέτο $lmSubsets$ συμπεριλαμβάνει το σύνολο δεδομένων $IbkTemperature$ το οποίο περιέχει 1824 παρατηρήσεις από 42 μεταβλητές μετεωρολογικών δεδομένων. Χρησιμοποιώντας το πακέτο $lmSubsets$ θα βρούμε τα βέλτιστα γραμμικά μοντέλα παλινδρόμησης που έχουν ως μεταβλητή απόκρισης τη μεταβλητή $temp$ που αντιπροσωπεύει τη θερμοκρασία στο αεροδρόμιο του Innsbruck.

Αρχικά φορτώνουμε το πακέτο $lmSubsets$ και το σύνολο δεδομένων $IbkTemperature$.

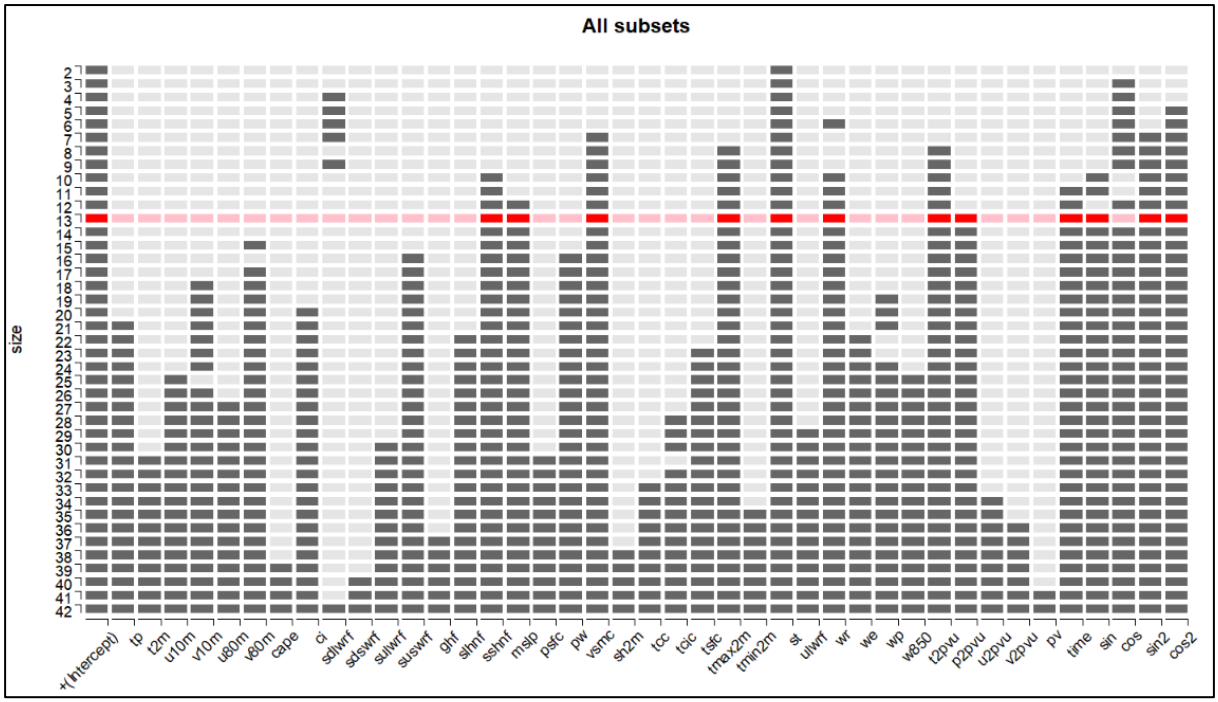
```
R> library(lmSubsets)
R> data("IbkTemperature", package = "lmSubsets")
R> IbkTemperature <- na.omit(IbkTemperature)
```

Χρησιμοποιούμε την εντολή $lmSubsets(\ \ )$ για να βρούμε τα βέλτιστα μοντέλα για κάθε πλήθος μεταβλητών, χρησιμοποιώντας ως κριτήριο σύγκρισης το RSS.

```
R> MOS1_all <- lmSubsets(temp ~ ., data = IbkTemperature)
R> plot(MOS1_all)
R> image(MOS1_all, hilite=1 , hilite_penalty = "BIC", size=2:42)
```



Σχήμα 10 – Τιμές *RSS* και *BIC* των βέλτιστων μοντέλων για κάθε πλήθος μεταβλητών



Σχήμα 11 – Οι μεταβλητές που περιέχονται στα βέλτιστα μοντέλα για κάθε πλήθος μεταβλητών

Στο Σχήμα 10 φαίνονται τα RSS και BIC των βέλτιστων μοντέλων και στο Σχήμα 11 φαίνονται οι μεταβλητές που περιέχει το κάθε ένα από αυτά τα μοντέλα όπου με κόκκινο χρώμα σημειώνουμε το μοντέλο με την μικρότερη τιμή BIC το οποίο ονομάζουμε $MOS1$ (Model Output Statistics). Μπορούμε να λάβουμε το μοντέλο $MOS1$ χρησιμοποιώντας την εντολή

R> MOS1 <- refit(lmSelect(MOS1_all))

ή την εντολή

R> MOS1 <- refit(lmSelect(temp ~., data=IbkTemperature))

Το μοντέλο $MOS1$ περιλαμβάνει 12 μεταβλητές και το σταθερό όρο και η τιμή του κριτηρίου BIC για το μοντέλο αυτό είναι $BIC = 9025.267$. Παρακάτω μπορούμε να δούμε περισσότερες πληροφορίες για το μοντέλο αυτό.

R> summary(MOS1)

R> BIC(MOS1)

```
Call:
lm(formula = temp ~ sshnf + mslp + vsmc + tmax2m + st + wr +
    t2pvu + p2pvu + time + sin + sin2 + cos2, data = IbkTemperature)

Residuals:
     Min       1Q   Median       3Q      Max
-10.9711  -1.7662  -0.0769   1.4689  14.1253

Coefficients:
              Estimate Std. Error t value Pr(>|t|)
(Intercept) -6.617e+02  9.523e+01  -6.949 5.13e-12 ***
sshnf        1.803e-02  3.535e-03   5.099 3.77e-07 ***
mslp        -2.928e-04  8.854e-05  -3.307 0.000961 ***
vsmc         2.018e+01  3.106e+00   6.497 1.06e-10 ***
tmax2m       1.815e-01  2.270e-02   7.997 2.26e-15 ***
st           1.142e+00  4.277e-02  26.701  < 2e-16 ***
wr           5.052e-01  1.032e-01   4.894 1.08e-06 ***
t2pvu        1.486e-01  2.776e-02   5.351 9.86e-08 ***
p2pvu       -1.245e-04  3.850e-05  -3.235 0.001240 **
time         1.469e-01  4.683e-02   3.137 0.001734 **
sin          8.113e-01  1.202e-01   6.749 1.99e-11 ***
sin2        -8.701e-01  1.181e-01  -7.369 2.60e-13 ***
cos2        -1.128e+00  9.749e-02 -11.565  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.82 on 1806 degrees of freedom
Multiple R-squared:  0.8549,    Adjusted R-squared:  0.8539
F-statistic: 886.4 on 12 and 1806 DF,  p-value: < 2.2e-16
```

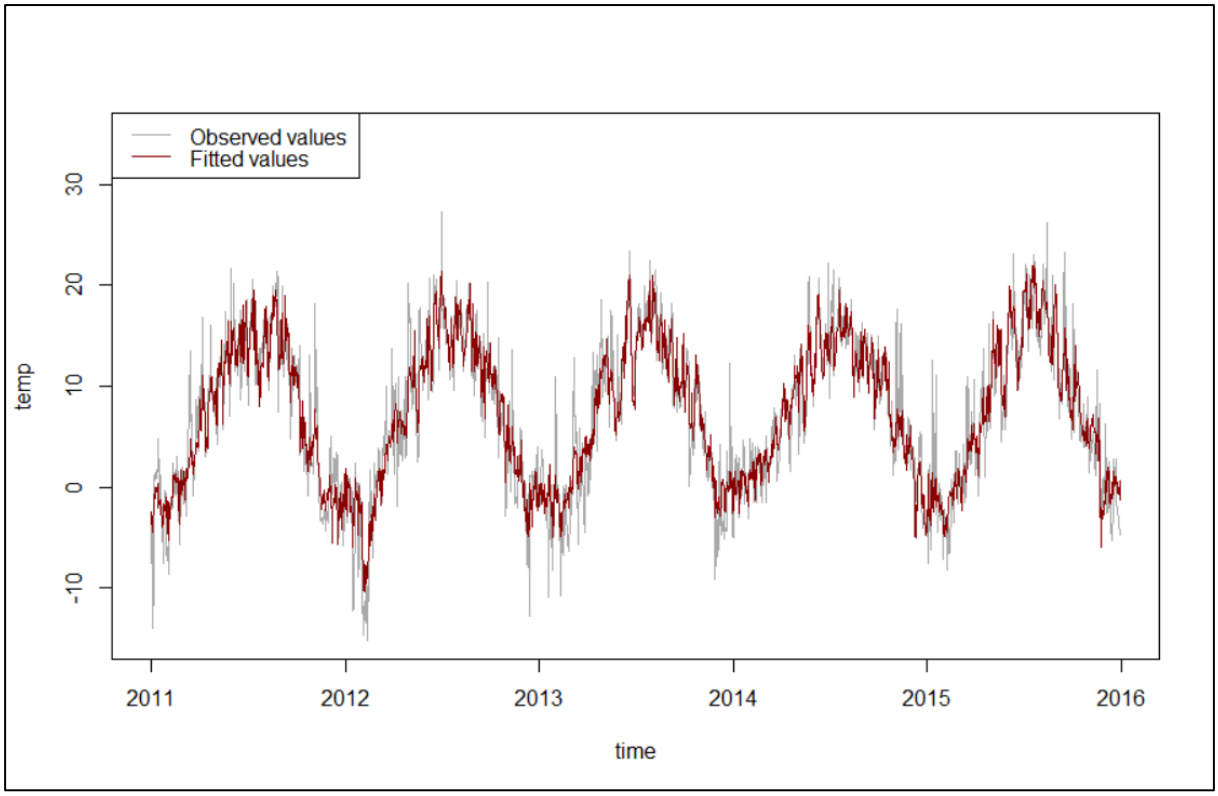
Επιπρόσθετα, στο Σχήμα 12 συγκρίνουμε τις παρατηρούμενες τιμές της θερμοκρασίας με τις αντίστοιχες προσαρμοσμένες τιμές του μοντέλου $MOS1$.

R> plot(temp ~ time, data = IbkTemperature, type = "l", col = "darkgray",ylim=c(-15,35))

R> lines(fitted(MOS1) ~ time, data = IbkTemperature, col = "darkred")

R> legend(x="topleft",legend=c("Observed values", "Fitted values"), col=c("darkgray",

+ "darkred"), lty=1)



Σχήμα 12 – Παρατηρούμενες και προσαρμοσμένες τιμές για τη θερμοκρασία

3.3.4 Μελλοντική Έρευνα

Οι μέθοδοι που περιγράψαμε στην παρούσα εργασία εφαρμόζονται στην περίπτωση όπου το πλήθος των παρατηρήσεών μας στο σύνολο δεδομένων που χρησιμοποιούμε είναι τουλάχιστον ίσο με το πλήθος των παραμέτρων του μοντέλου, δηλαδή $n \geq p$. Σε προβλήματα μεγάλων διαστάσεων συχνά θα έχουμε λιγότερες παρατηρήσεις απ'ότι μεταβλητές, δηλαδή $n < p$. Στην παρούσα Ενότητα θα μελετήσουμε αυτήν την περίπτωση και θα προτείνουμε μία προσέγγιση για την εύρεση των βέλτιστων μοντέλων χρησιμοποιώντας οποιοδήποτε συνδυασμό k, ή λιγότερων, μεταβλητών, όπου $k < n < p$.

Δέντρα Παλινρόμησης

Η προσέγγισή μας βασίζεται στα δέντρα παλινδρόμησης που παράγονται σύμφωνα με τη διαδικασία του αλγορίθμου DCA. Για να παράξουμε ένα δέντρο παλινδρόμησης αρχικά πρέπει να επιλέξουμε k από τις p μεταβλητές οι οποίες θα αποτελούν τον κόμβο-ρίζα του δέντρου. Υπάρχουν $\binom{p}{k}$ τέτοιοι συνδυασμοί μεταβλητών και ως επακόλουθο μπορούμε να δημιουργήσουμε $\binom{p}{k}$ δέντρα παλινδρόμησης.

Το κάθε δέντρο παράγει $2^k - 1$ μοντέλα επομένως από όλα τα δέντρα θα λάβουμε συνολικά $\binom{p}{k}(2^k - 1)$ μοντέλα. Αφού κάποιοι κόμβοι-ρίζες των δέντρων περιέχουν κοινές μεταβλητές με άλλους κόμβους-ρίζες, είναι λογικό ότι τα δέντρα που θα παραχθούν από αυτούς θα περιέχουν μερικά κοινά μοντέλα. Το συνολικό πλήθος των μοναδικών μοντέλων που θα παραχθούν από όλα τα δέντρα παλινδρόμησης δίνεται από το $M \equiv \sum_{i=1}^{k} \binom{p}{i}$. Δεν υπάρχει κλειστή μορφή για το M ωστόσο μπορούμε να βρούμε κάποιο εύρος τιμών στο οποίο ανήκει. Ένα προφανές εύρος τιμών είναι το $L_1 \equiv 2^k - 1 \leq M \leq \binom{p}{k}(2^k - 1) \equiv U_1$.

Συμβολίζουμε τις p μεταβλητές ως $x_i, i = 1,2, \dots, p$. Τα δέντρα παλινδρόμησης που αναφέραμε προηγουμένως θα παραχθούν χρησιμοποιώντας τους κόμβους-ρίζες με την εξής σειρά:

$[x_1, \dots, x_{k-2}, x_{k-1}, \boldsymbol{x_k}], [x_1, \dots, x_{k-2}, x_{k-1}, \boldsymbol{x_{k+1}}], [x_1, \dots, x_{k-2}, x_{k-1}, \boldsymbol{x_{k+2}}], \dots, [x_1, \dots, x_{k-2}, x_{k-1}, \boldsymbol{x_p}]$,
$[x_1, \dots, x_{k-2}, \boldsymbol{x_k}, \boldsymbol{x_{k+1}}], [x_1, \dots, x_{k-2}, \boldsymbol{x_k}, \boldsymbol{x_{k+2}}], \dots, [x_1, \dots, x_{k-2}, \boldsymbol{x_k}, \boldsymbol{x_p}], [x_1, \dots, x_{k-2}, \boldsymbol{x_{k+1}}, \boldsymbol{x_{k+2}}], \dots,$
$[x_1, \dots, x_{k-2}, \boldsymbol{x_{p-1}}, \boldsymbol{x_p}], [x_1, \dots, x_{k-3}, \boldsymbol{x_{k-1}}, \boldsymbol{x_k}, \boldsymbol{x_{k+1}}], \dots, [\boldsymbol{x_{p-k}}, \dots, \boldsymbol{x_{p-2}}, \boldsymbol{x_{p-1}}, \boldsymbol{x_p}]$

Για παράδειγμα, για $p = 7$ και $k = 4$ η ακολουθία των κόμβων-ρίζες είναι

$$[123\mathbf{4}], [123\mathbf{5}], [123\mathbf{6}], [123\mathbf{7}], [12\mathbf{45}], [12\mathbf{46}], [12\mathbf{47}],$$
$$[12\mathbf{56}], [12\mathbf{57}], [12\mathbf{67}], [1\mathbf{345}], \dots, [\mathbf{3567}], [\mathbf{4567}]$$

Αν χρησιμοποιήσουμε την παραπάνω ακολουθία για την παραγωγή των δέντρων παλινδρόμησης τότε τα δέντρα που θα παράξουμε μπορούν να χωριστούν σε $k + 1$ ομάδες. Θα συμβολίζουμε τον αριθμό της ομάδας ως g. Για τις ομάδες αυτές ισχύουν τα ακόλουθα:

- Η Ομάδα 1 ($g = 1$) περιέχει ένα δέντρο (το πρώτο στη σειρά), από το οποίο παράγονται $2^k - 1$ μοναδικά μοντέλα.

- Η Ομάδα 2 $(g = 2)$ περιέχει τα επόμενα $(p - k)$ δέντρα, όπου το κάθε δέντρο παράγει 2^{k-1} μοναδικά μοντέλα.
- Η Ομάδα g $(3 \le g \le k + 1)$ περιέχει τα επόμενα $\sum_{l=0}^{p-k-1} \sum_{j=l}^{p-k-1} \cdots \sum_{i=z}^{p-k-1} (p - k - i)$ δέντρα, όπου το κάθε δέντρο παράγει 2^{k-g+1} μοναδικά μοντέλα. Το παραπάνω πολλαπλό άθροισμα περιέχει $g - 2$ αθροίσματα.

Σημειώνουμε ότι σύμφωνα με τα παραπάνω θα ισχύει

$$M = \sum_{i=1}^{k} \binom{p}{i} = \left(2^k - 1\right) + (p - k) \cdot \left(2^{k-1}\right) + \sum_{g=3}^{k+1} \left[2^{k-g+1} \cdot \sum_{l=0}^{p-k-1} \sum_{j=l}^{p-k-1} \cdots \sum_{i=z}^{p-k-1} (p - k - i) \right]$$

Βελτίωση Φραγμάτων

Χρησιμοποιώντας τα παραπάνω αποτελέσματα για τις Ομάδες των δέντρων παλινδρόμησης μπορούμε να βρούμε ένα άλλο άνω φράγμα για το πλήθος των μοναδικών μοντέλων, M. Για σκοπούς απλοποίησης, θέτουμε $\mathrm{d} = p - k$. Η Ομάδα g περιέχει συνολικά $2^{k-g+1} \cdot \sum_{l=0}^{d-1} \sum_{j=l}^{d-1} \cdots \sum_{i=z}^{d-1} (d - i)$ μοναδικά μοντέλα. Όμως

$$\sum_{l=0}^{d-1} \sum_{j=l}^{d-1} \cdots \sum_{i=z}^{d-1} (d - i) < \sum_{l=0}^{d-1} \sum_{j=0}^{d-1} \cdots \sum_{i=0}^{d-1} (d - i) = d^{g-3} \cdot \sum_{i=0}^{d-1} (d - i)$$

και

$$\sum_{i=0}^{d-1} (d - i) = d^2 - \sum_{i=0}^{d-1} i = d^2 - \frac{d(d-1)}{2} = \frac{d(d+1)}{2}$$

επομένως

$$2^{k-g+1} \cdot \sum_{l=0}^{d-1} \sum_{j=l}^{d-1} \cdots \sum_{i=z}^{d-1} (d - i) < 2^{k-g+1} \cdot d^{g-3} \cdot \frac{d(d+1)}{2} = \frac{2^k (d+1)}{d} \left(\frac{d}{2}\right)^{g-1}$$

Συνεπώς, για το συνολικό πλήθος μοναδικών μοντέλων ισχύει

$$M < \left(2^k - 1\right) + d \cdot \left(2^{k-1}\right) + \sum_{g=3}^{k+1} \left[\frac{2^k (d+1)}{d} \left(\frac{d}{2}\right)^{g-1} \right]$$

$$= \left(2^k - 1\right) + d \cdot \left(2^{k-1}\right) + \frac{2^k (d+1)}{d} \cdot \sum_{g=3}^{k+1} \left(\frac{d}{2}\right)^{g-1}$$

$$= \left(2^k - 1\right) + d \cdot \left(2^{k-1}\right) + \frac{2^k (d+1)}{d} \cdot \left[\sum_{g=1}^{k+1} \left(\frac{d}{2}\right)^{g-1} - 1 - \frac{d}{2} \right]$$

Όμως

$$\sum_{g=1}^{k+1}\left(\frac{d}{2}\right)^{g-1}=\frac{1-\left(\frac{d}{2}\right)^{k+1}}{1-\left(\frac{d}{2}\right)}$$

Τότε

$$M<\left(2^{k}-1\right)+d\cdot\left(2^{k-1}\right)+\frac{2^{k}(d+1)}{d}\cdot\left[\frac{1-\left(\frac{d}{2}\right)^{k+1}}{1-\left(\frac{d}{2}\right)}-\left[1+\frac{d}{2}\right]\right]\equiv U_{2}$$

Μπορούμε να υπολογίσουμε ακόμα ένα άνω φράγμα για το πλήθος M χρησιμοποιώντας τους αριθμούς Bell. Παρατηρούμε ότι

$$M=\sum_{i=1}^{k}\binom{p}{i}<\sum_{i=0}^{k}\binom{p}{i}<\sum_{i=0}^{p}\binom{p}{i}<\sum_{i=0}^{p}\binom{p}{i}B_{i}=B_{p+1}\equiv U_{3}$$

αφού για τους αριθμούς Bell $B_i, i\geq 0$ ισχύει $B_i>0$.

Συγκρίνοντας τα άνω φράγματα U_1, U_2, U_3, μπορούμε να παρατηρήσουμε ότι $U_2<U_1$ για μικρές τιμές των p και k. Ωστόσο στην περίπτωση προβλημάτων με δεδομένα μεγάλων διαστάσεων θα έχουμε μεγάλες τιμές για το p και παρατηρούμε ότι τότε $U_2>U_1$. Επιπρόσθετα, $U_2<U_3$ για οποιεσδήποτε τιμές των p και k. Τελικά συμπεραίνουμε ότι τα άνω φράγματα U_2, U_3 δε βελτιώνουν το άνω φράγμα U_1. Χρησιμοποιώντας παρόμοια προσέγγιση ίσως μπορέσουμε να υπολογίσουμε καλύτερα άνω και κάτω φράγματα που βελτιώνουν τα U_1 και L_1.

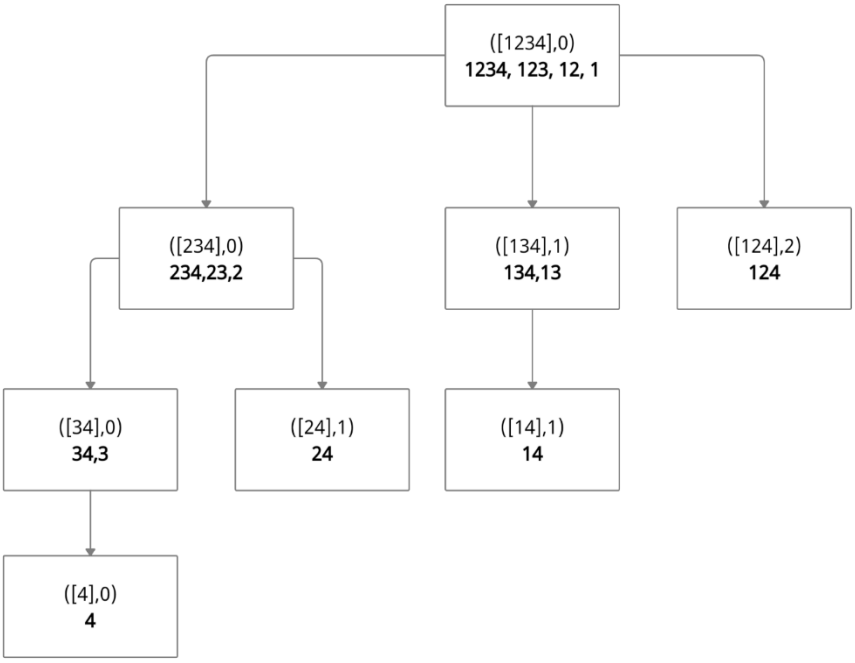
Παράδειγμα Εφαρμογής

Ας μελετήσουμε την περίπτωση όπου $p=7$ και $k=4$ ώστε να κατανοήσουμε σε περισσότερο βάθος την προτεινόμενη προσέγγιση. Παράγουμε τα δέντρα παλινδρόμησης με τη σειρά που αναφέραμε προηγουμένως. Πέραν από το πλήθος των μοναδικών μοντέλων, το οποίο είναι το ίδιο για δέντρα που ανήκουν στην ίδια ομάδα, παρατηρούμε μερικές επιπλέον ομοιότητες μεταξύ αυτών των δέντρων. Για ένα δέντρο που ανήκει στην ομάδα g $(g>1)$, τα 2^{k-g+1} μοναδικά μοντέλα που παράγονται από αυτό το δέντρο λαμβάνονται από 2^{k-g+1} διαφορετικούς κόμβους του δέντρου, δηλαδή 2^{k-g+1} κόμβοι του δέντρου περιέχουν ένα μοναδικό μοντέλο ο καθένας. Οι κόμβοι που περιέχουν κάποιο μοναδικό μοντέλο βρίσκονται στην ίδια θέση του δέντρου παλινδρόμησης, για κάθε δέντρο που ανήκει στην ομάδα g. Στα Σχήματα 13-17 παρουσιάζουμε ένα δέντρο παλινδρόμησης από κάθε ομάδα g για το παράδειγμά μας. Με έντονη γραμματοσειρά παρουσιάζονται τα μοναδικά μοντέλα της κάθε ομάδας, δηλαδή τα μοντέλα που δεν έχουν παραχθεί από κάποιο προηγούμενο δέντρο.

- **Ομάδα 1**

Κόμβοι-ρίζες : $[1234]$

Αυτή η ομάδα περιέχει 1 δέντρο με $2^k - 1 = 2^4 - 1 = 15$ μοναδικά μοντέλα.

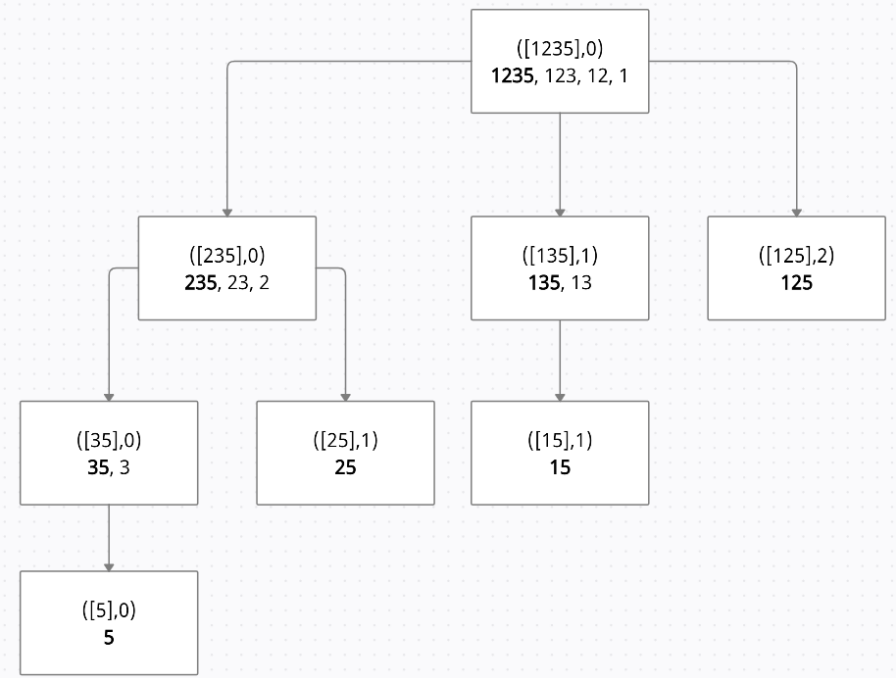


Σχήμα 13 – Δέντρο παλινδρόμησης Ομάδας 1 ($k = 4, p = 7$)

- **Ομάδα 2**

Κόμβοι-ρίζες : $[1235], [1236], [1237]$

Αυτή η ομάδα περιέχει $p - k = 7 - 4 = 3$ δέντρα με $2^{k-1} = 2^3 = 8$ μοναδικά μοντέλα το καθένα.

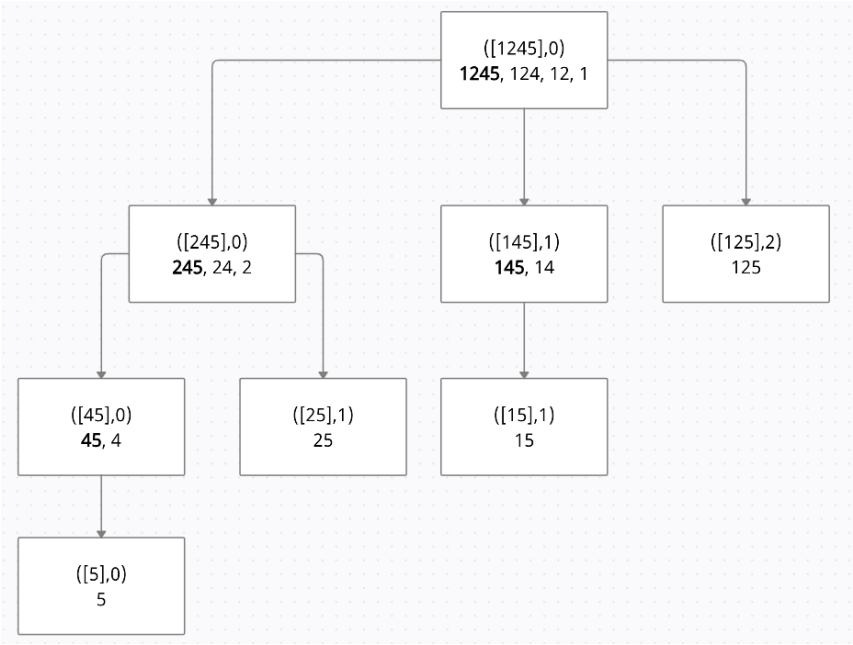


Σχήμα 14 - Δέντρο παλινδρόμησης Ομάδας 2 ($k = 4, p = 7$)

- **Ομάδα 3**

Κόμβοι-ρίζες : $[1245], [1246], [1247], [1256], [1257], [1267]$

Αυτή η ομάδα περιέχει $\sum_{i=0}^{p-k-1}(p-k-i) = \sum_{i=0}^{2}(3-i) = 6$ δέντρα με $2^{k-2} = 2^2 = 4$ μοναδικά μοντέλα το καθένα.

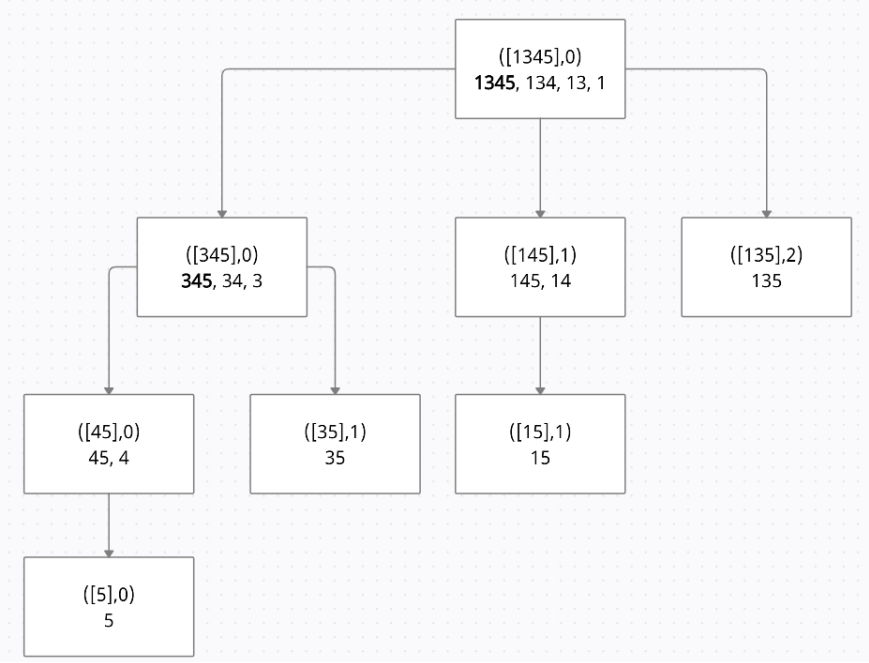


Σχήμα 15 - Δέντρο παλινδρόμησης Ομάδας 3 ($k = 4, p = 7$)

- **Ομάδα 4**

Κόμβοι-ρίζες : $[1345], [1346], [1347], [1356], [1357], [1367], [1456], [1457], [1467], [1567]$

Αυτή η ομάδα περιέχει $\sum_{j=0}^{p-k-1}\sum_{i=j}^{p-k-1}(p-k-i) = \sum_{j=0}^{2}\sum_{i=j}^{2}(3-i) = 10$ δέντρα με $2^{k-3} = 2$ μοναδικά μοντέλα το καθένα.

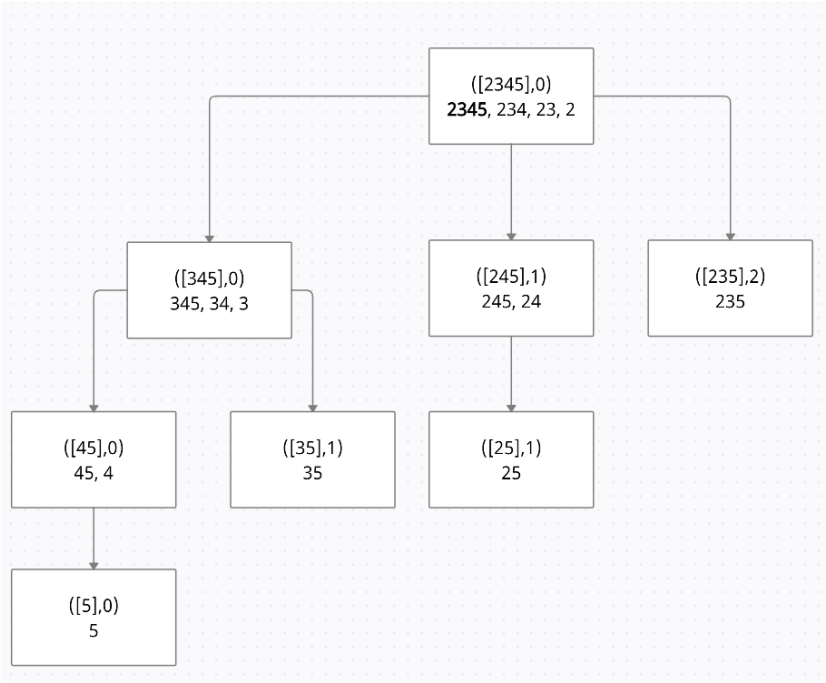


Σχήμα 16 - Δέντρο παλινδρόμησης Ομάδας 4 ($k = 4, p = 7$)

- **Ομάδα 5**

Κόμβοι-ρίζες : $[2345], [2346], [2347], [2356], [2357], [2367], [2456], [2457], [2467],$
$[2567], [3456], [3457], [3467], [3567], [4567]$

Αυτή η ομάδα περιέχει $\sum_{l=0}^{p-k-1}\sum_{j=l}^{p-k-1}\sum_{i=j}^{p-k-1}(p-k-i) = \sum_{l=0}^{2}\sum_{j=l}^{2}\sum_{i=j}^{2}(3-i) = 15$ δέντρα με $2^{k-4} = 2^0 = 1$ μοναδικό μοντέλο το καθένα.



Σχήμα 17 - Δέντρο παλινδρόμησης Ομάδας 5 ($k = 4, p = 7$)

Μείωση Υπολογιστικού Κόστους

Για να μειώσουμε το υπολογιστικό κόστος μπορούμε να αποφύγουμε τον υπολογισμό κάποιων κόμβων των δέντρων παλινδρόμησης χρησιμοποιώντας τους κανόνες των αλγόριθμων BBA και lmSelect. Μία προσέγγιση θα ήταν να χρησιμοποιούσαμε ολόκληρα τα δέντρα παλινδρόμησης και να εφαρμόζαμε κανονικά τους αλγόριθμους BBA και lmSelect όπως περιγράψαμε στην Ενότητα 3.3.3.

Θα μπορούσαμε να μειώσουμε πιθανό ακόμα περισσότερο τους κόμβους που πρέπει να υπολογίσουμε, βασισμένοι στην εξής παρατήρηση. Παρατηρούμε ότι καθώς ο αριθμός της ομάδας, g, αυξάνεται, κάποιοι κόμβοι στο κάτω και στο δεξί μέρος των δέντρων παλινδρόμησης καθίστανται περιττοί εφόσον έχουν ήδη ληφθεί υπόψη σε προηγούμενο δέντρο. Αν μπορέσουμε να βρούμε κάποιο γενικό κανόνα ώστε να γνωρίζουμε ποιοι ακριβώς κόμβοι των δέντρων κάθε ομάδας είναι περιττοί, τότε θα μπορέσουμε να μειώσουμε περισσότερο το υπολογιστικό κόστος.

Μπορούμε επίσης να μειώσουμε το υπολογιστικό κόστος χρησιμοποιώντας την παραγοντοποίηση QR κάποιου βήματος της διαδικασίας για να υπολογίσουμε την παραγοντοποίηση για κάποιο μετέπειτα στάδιο. Για παράδειγμα, στο παραπάνω παράδειγμά μας ο παράγοντας R, $\boldsymbol{R}_1$, της παραγοντοποίησης QR για τον κόμβο-ρίζα $[1245]$ μπορεί να χρησιμοποιηθεί για τον υπολογισμό του παράγοντα R, $\boldsymbol{R}_2$, για τον κόμβο $[1246]$. Πιο συγκεκριμένα, ας ονομάσουμε ως $\boldsymbol{R}$ τον παράγοντα R

της παραγοντοποίησης QR για τον $n \times p$ πίνακα $\boldsymbol{X}$ που περιέχει όλες τις μεταβλητές. Για να υπολογίσουμε το $\boldsymbol{R}_2$ θα μπορούσαμε να χρησιμοποιήσουμε τις στήλες $1,2,4$ και 6 του $\boldsymbol{R}$, για να ορίσουμε ένα πίνακα $\boldsymbol{R}^*$, και στη συνέχεια εφαρμόζοντας περιστροφές Givens στην 3η και 4η στήλη του $\boldsymbol{R}^*$ για να τον τριγωνοποιήσουμε θα βρίσκαμε τον $\boldsymbol{R}_2$. Διαφορετικά, θα μπορούσαμε να χρησιμοποιήσουμε τις πρώτες 3 στήλες του $\boldsymbol{R_1}$ και τη στήλη 6 του $\boldsymbol{R}$, για να ορίσουμε τον $\boldsymbol{R}^*$, και στη συνέχεια εφαρμόζοντας περιστροφές Givens μόνο στην 4η στήλη του $\boldsymbol{R}^*$ των τριγωνονοποιούμε και βρίσκουμε τον $\boldsymbol{R_2}$**.** Χρησιμοποιώντας τη δεύτερη προσέγγιση μειώνουμε το υπολογιστικό κόστος. Παρόμοια στρατηγική αναμένεται να μπορεί να εφαρμοστεί σε διάφορα στάδια της διαδικασίας. Εφαρμόζοντας αυτήν την μέθοδο για τη μείωση του υπολογιστικού κόστους απαιτεί τη χρήση περισσότερης μνήμης στον υπολογιστή αφού θα χρειάζεται να αποθηκεύουμε κάποιους πίνακες $\boldsymbol{R}$ προηγούμενων σταδίων.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Akaike, H. (1973). Information Theory and an Extension of the Maximum Likelihood Principle. *2nd International Symposium on Information Theory* (σσ. 267-281). Budapest: Akadémiai Kiadó.

Bjorck, A. (1996). *Numerical Methods for Least Squares Problems.* SIAM.

Clarke, M. R. (1981). Algorithm AS 163: A Givens Algorithm for Moving from One Linear Model to Another Without Going Back to the Data. *Journal of the Royal Statistics Society, 30*(2), 198-203.

Eisenhauer, J. G. (2003). Regression through the origin. *Teaching Statistics, 25*(3), 76-80.

Furnival, G. M. (1971). All Possible Regressions with Less Computation. *Technometrics, 13*(2), 403-408.

Furnival, G. M., & Wilson, R. W. (1974). Regressions by Leaps and Bounds. *Technometrics, 16*(4), 499-511.

Garside, M. J. (1965). The Best Sub-Set in Multiple Regression Analysis. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 196-200.

Gatu, C., & Kontoghiorghes, E. J. (2003). Parallel algorithms for computing all possible subset regression models using the QR decomposition. *Parallel Computing, 29*(4), 505-521.

Gatu, C., & Kontoghiorghes, E. J. (2006). Branch-and-Bound Algorithms for Computing the Best-Subset Regression Models. *Journal of Computational and Graphical Statistics, 15*, 139-156.

Golub, G. H., & Van Loan, C. F. (2013). *Matrix Computations.* The Johns Hopkins University Press.

Goodnight, J. H. (1979). A Tutorial on the SWEEP Operator. *The American Statistician, 33*(3), 149-158.

Hastie, T., Tibshirani, R., James, G., & Witten, D. (2021). *An Introduction to Statistical Learning with applications in R* (2 εκδ.). New York, NY: Springer.

Higham, N. J. (2011). Gaussian Elimination. *WIREs Computational Statistics, 3*(3), 230-238.

Hocking, R. R. (1972). Criteria for Selection of a Subset Regression: Which One Should Be Used? *Technometrics, 14*(4), 967-976.

Hocking, R. R., & Leslie, R. N. (1967). Selection of the Best Subset in Regression Analysis. *Technometrics, 9*(4), 531-540.

Hofmann, M., Gatu, C., & Kontoghiorghes, E. J. (2007). Efficient algorithms for computing the best subset regression models for large-scale problems. *Computational Statistics & Data Analysis, 52*(1), 16-29.

Hofmann, M., Gatu, C., Kontoghiorghes, E. J., Colubi, A., & Zeileis, A. (2020). lmSubsets: Exact Variable-Subset Selection in Linear Regression for R. *Journal of Statistical Software, 93*(3), 1-21.

Kontoghiorghes, E. J. (2000). *Parallel Algorithms for Linear Models: Numerical Methods and Estimation Problems.* New York, NY: Springer. doi:https://doi.org/10.1007/978-1-4615-4571-2

Kullback, S., & Leibler, R. A. (1951). On Information and Sufficiency. *The Annals of Mathematical Statistics, 22*(1), 79-86. doi:10.1214/aoms/1177729694

LaMotte, L. R., & Hocking, R. R. (1970). Computational Efficiency in the Selection of Regression Variables. *Technometrics, 12*(1), 83-93.

Mallows, C. L. (1973). Some Comments on Cp. *Technometrics, 15*, 661-675.

Miller, A. (2002). *Subset Selection in Regression* (Τόμ. 2). Chapman & Hall/CRC.

Morgan, J. A., & Tatar, J. F. (1972). Calculation of the Residual Sum of Squares for allPossible Regressions. *Technometrics, 14*(2), 317-325.

Rao , C. R., Shalabh, Heumann, C., & Toutenburg, H. (2008). *Linear Models and Generalizations* (3 εκδ.). Springer.

Schatzoff, M., Tsao, R., & Fienberg, S. (1968). Efficient Calculation of All Possible Regressions. *Technometrics, 10*(4), 769-779.

Schwartz, G. (1978). Estimating the Dimension of a Model. *The Annals of Statistics, 6*(2), 461-464.

Seber, G. A., & Lee, A. J. (2003). *Linear Regression Analysis* (2η εκδ.). John Wiley & Sons.

Smith, D. M., & Bremner, J. M. (1989). All possible subset regressions using the QR decomposition. *Computational Statistics & Data Analysis, 7*(3), 217-235.

Καρώνη, Χ., & Οικονόμου, Π. (2017). *Στατιστικά Μοντέλα Παλινδρόμησης* (2 εκδ.). Αθήνα: Εκδόσεις ΣΥΜΕΩΝ.

ΠΑΡΑΡΤΗΜΑ Α

Ορισμός Α.1

Αν η $f(\boldsymbol{X})$ είναι μία πραγματική συνάρτηση ενός $m \times n$ πίνακα $\boldsymbol{X} = (x_{ij})$ τότε η μερική παράγωγος της f ως προς $\boldsymbol{X}$ ορίζεται ως τον $m \times n$ πίνακα μερικών παραγώγων $\partial f / \partial x_{ij}$:

$$\frac{\partial f(\boldsymbol{X})}{\partial \boldsymbol{X}} = \begin{pmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{1n}} \\ \vdots & & \vdots \\ \frac{\partial f}{\partial x_{m1}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{pmatrix}$$

Θεώρημα Α.1

Έστω $\boldsymbol{A}$ ένας $n \times n$ πίνακας, $\boldsymbol{\gamma}$ ένα $n \times 1$ διάνυσμα σταθερών και $\boldsymbol{x}$ ένα $n \times 1$ διάνυσμα μεταβλητών. Τότε ισχύει

- $\frac{\partial \boldsymbol{\gamma}' \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{\gamma}$
- $\frac{\partial \boldsymbol{x}' \boldsymbol{x}}{\partial \boldsymbol{x}} = 2\boldsymbol{x}$
- $\frac{\partial \boldsymbol{x}' \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{A}'\boldsymbol{x}$

Αν επιπλέον ο πίνακας $\boldsymbol{A}$ είναι συμμετρικός τότε $\frac{\partial \boldsymbol{x}' \boldsymbol{A} \boldsymbol{x}}{\partial \boldsymbol{x}} = \boldsymbol{2A}\boldsymbol{x}$

Ορισμός Α.2 - Εικόνα, Πυρήνας, Τάξη/Βαθμός Πίνακα

Έστω ένας $m \times n$ πίνακας $\boldsymbol{A}$.

Η **εικόνα** (range) του πίνακα $\boldsymbol{A}$ ορίζεται ως

$$range(\boldsymbol{A}) = \{\boldsymbol{y} \in \mathbb{R}^m : \boldsymbol{y} = \boldsymbol{A}\boldsymbol{x}, \text{ για κάποιο } \boldsymbol{x} \in \mathbb{R}^n \}$$

Ο **πυρήνας** (nullspace) του πίνακα $\boldsymbol{A}$ ορίζεται ως

$$\mathrm{n}ull(\boldsymbol{A}) = \{\boldsymbol{x} \in \mathbb{R}^n : \boldsymbol{A}\boldsymbol{x} = 0\}$$

Αν $\boldsymbol{A} = [a_1|\cdots|a_n]$ είναι ένας διαμερισμός των στηλών του πίνακα $\boldsymbol{A}$ τότε

$$range(\boldsymbol{A}) = span\{a_1, \dots, a_n\} = \left\{ \sum_{j=1}^{n} \beta_j a_j : \beta_j \in \mathbb{R} \right\}$$

όπου $span\{a_1, \dots, a_n\}$ είναι ένας υπόχωρος του $\mathbb{R}^m$ που περιέχει όλους τους γραμμικούς συνδυασμούς των διανυσμάτων $a_1, \dots, a_n \in \mathbb{R}^m$.

Η **τάξη** ή ο **βαθμός** του πίνακα $\boldsymbol{A}$ ορίζεται ως

$$rank(\boldsymbol{A}) = \dim\big(range(\boldsymbol{A})\big)$$

και αποτελεί το πλήθος των γραμμικώς ανεξάρτητων στηλών (ή γραμμών) του πίνακα.

Ισχύει ότι

$$\dim\big(null(\boldsymbol{A})\big) + rank(\boldsymbol{A}) = n$$

Θεώρημα Α.2

Για ένα $m \times n$ πίνακα $\boldsymbol{A}$ ισχύει

$$0 \leq rank(\boldsymbol{A}) \leq \min\{m, n\}$$

Λέμε ότι ο πίνακας $\boldsymbol{A}$ έχει υστέρηση τάξης (είναι rank deficient) αν

$$rank(\boldsymbol{A}) < \min\{m, n\}$$

Θεώρημα Α.3

Έστω ο $n \times n$ πίνακας $\boldsymbol{A}$ και $\lambda_1, \lambda_2, \dots, \lambda_n$ οι ιδιοτιμές του. Τότε η ορίζουσα του πίνακα $\boldsymbol{A}$ δίνεται από τη σχέση

$$\det(\boldsymbol{A}) = \lambda_1 \lambda_2 \cdots \lambda_n$$

Θεώρημα Α.4

Έστω ο συμμετρικός και θετικά ορισμένος $n \times n$ πίνακας $\boldsymbol{A}$ και $\lambda_1, \lambda_2, \dots, \lambda_n$ οι ιδιοτιμές του. Τότε

$$\lambda_i > 0 \;\; \forall\, i = 1,2, \dots, n$$

Θεώρημα Α.5

Αν για τον $n \times p$ πίνακα $\boldsymbol{X}$ ισχύει $rank(\boldsymbol{X}) = p$, τότε ο πίνακας $\boldsymbol{X}'\boldsymbol{X}$ είναι θετικά ορισμένος και αντιστρέψιμος.

Απόδειξη

Έστω $\boldsymbol{x} \in \mathbb{R}^p - \{\boldsymbol{0}\}$. Τότε $\boldsymbol{x}'(\boldsymbol{X}'\boldsymbol{X})\boldsymbol{x} = (\boldsymbol{X}\boldsymbol{x})'(\boldsymbol{X}\boldsymbol{x}) = \|\boldsymbol{X}\boldsymbol{x}\|^2 \geq \boldsymbol{0}$ με την ισότητα να ισχύει αν και μόνο αν

$$\boldsymbol{Xx} = \mathbf{0} \overset{rank(\boldsymbol{X})=p}{\Longleftrightarrow} \boldsymbol{x} = \mathbf{0}$$

Εφόσον ο πίνακας $\boldsymbol{X'X}$ είναι συμμετρικός και θετικά ορισμένος, τότε $\det(\boldsymbol{X'X}) \neq 0$ (Θεώρημα Α.3 & Θεώρημα Α.4), δηλαδή ο πίνακας $\boldsymbol{X'X}$ είναι αντιστρέψιμος.

Θεώρημα Α.6 – Παραγοντοποίηση $\boldsymbol{LU}$

Κάθε αντιστρέψιμος $p \times p$ πίνακας $\boldsymbol{A}$ μπορεί να παραγοντοποιηθεί στη μορφή

$$\boldsymbol{PA} = \boldsymbol{LU}$$

όπου $\boldsymbol{P}$ είναι ένας πίνακας μεταθέσεων, $\boldsymbol{L}$ είναι ένας κάτω τριγωνικός πίνακας με μονάδες στην κύρια διαγώνιο και $\boldsymbol{U}$ είναι ένας άνω τριγωνικός πίνακας.

Θεώρημα Α.7 – Παραγοντοποίηση $\boldsymbol{LDL^T}$

Κάθε συμμετρικός, θετικά ορισμένος πίνακας $\boldsymbol{A}$ μπορεί να παραγοντοποιηθεί στη μορφή

$$\boldsymbol{A} = \boldsymbol{LDL}^T$$

όπου $\boldsymbol{L}$ είναι ένας κάτω τριγωνικός πίνακας με μονάδες στην κύρια διαγώνιο και $\boldsymbol{D}$ είναι ένας διαγώνιος πίνακας με θετικά στοιχεία στην κύρια διαγώνιο.

Ορισμός Α.3 – Ορθογώνιος Πίνακας

Ένας πίνακας $\boldsymbol{Q}$ ονομάζεται ορθογώνιος εάν ισχύουν:

i) Ο $\boldsymbol{Q}$ είναι τετραγωνικός και
ii) $\boldsymbol{Q}^{-1} = \boldsymbol{Q}^T$ ή διαφορετικά, $\boldsymbol{QQ}^T = \boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$

Θεώρημα Α.8 – Παραγοντοποίηση $\boldsymbol{QR}$ (Golub & Van Loan, 2013; Bjorck, 1996)

Αν $\boldsymbol{A} \in \mathbb{R}^{n\times p}, n \geq p$, τότε υπάρχει ορθογώνιος πίνακας $\boldsymbol{Q} \in \mathbb{R}^{n\times n}$ και άνω τριγωνικός πίνακας $\boldsymbol{R} \in \mathbb{R}^{p\times p}$ ώστε

$$\boldsymbol{A} = \boldsymbol{Q}\begin{pmatrix}\boldsymbol{R}\\ \mathbf{0}\end{pmatrix}\begin{matrix}p\\ n-p\end{matrix}$$

Ο πίνακας $\boldsymbol{R}$ περιέχει μη-αρνητικά στοιχεία στην κύρια διαγώνιό του και ονομάζεται ο παράγοντας-$\boldsymbol{R}$ ($\boldsymbol{R}$-factor) του $\boldsymbol{A}$.

Θεώρημα Α.9 (Bjorck, 1996)

Έστω ο πίνακας $\boldsymbol{A} \in \mathbb{R}^{n \times p}$ με $rank(\boldsymbol{A}) = p$. Τότε εάν ο παράγοντας-$\boldsymbol{R}$ του $\boldsymbol{A}$ έχει θετικά διαγώνια στοιχεία τότε ισούται με τον παράγοντα Cholesky του $\boldsymbol{A}^T\boldsymbol{A}$.

Θεώρημα Α.10 – Singular Value Decomposition (SVD) (Seber & Lee, 2003)

Έστω $\boldsymbol{X} \in \mathbb{R}^{n \times p}$. Τότε ο $\boldsymbol{X}$ μπορεί να εκφραστεί στη μορφή

$$\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}'$$

όπου

$\boldsymbol{U}$ είναι ένας $n \times p$ πίνακας που περιέχει p ορθοκανονικά ιδιοδιανύσματα που αντιστοιχούν στις p μεγαλύτερες ιδιοτιμές του $\boldsymbol{X}\boldsymbol{X}'$

$\boldsymbol{V}$ είναι ένας $p \times p$ ορθογώνιος πίνακας που περιέχει τα ορθοκανονικά ιδιοδιανύσματα του $\boldsymbol{X}\boldsymbol{X}'$ και

$\boldsymbol{\Sigma} = diag(\sigma_1, \sigma_2, \dots, \sigma_p)$ είναι ένας $p \times p$ διαγώνιος πίνακας. Οι ποσότητες $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ ονομάζονται ιδιάζουσες τιμές (singular values) του $\boldsymbol{X}$ και αποτελούν τις τετραγωνικές ρίζες των (μη-αρνητικών) ιδιοτιμών του $\boldsymbol{X}'\boldsymbol{X}$.

ΠΑΡΑΡΤΗΜΑ Β

Β.1

$$RSS = \boldsymbol{e}^T\boldsymbol{e} = (\boldsymbol{y} - \hat{\boldsymbol{y}})^T(\boldsymbol{y} - \hat{\boldsymbol{y}}) = (\boldsymbol{y}^T - \boldsymbol{y}^T\boldsymbol{H}^T)(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y}) = (\boldsymbol{y}^T - \boldsymbol{y}^T\boldsymbol{H})(\boldsymbol{y} - \boldsymbol{H}\boldsymbol{y})$$

$$= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} = \boldsymbol{y}^T(\boldsymbol{I} - 2\boldsymbol{H} + \boldsymbol{H}\boldsymbol{H})\boldsymbol{y}$$

$$= \boldsymbol{y}^T(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \ \ (\boldsymbol{H} \ \textit{είναι συμμετρικος και ταυτοδύναμος})$$

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