ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ
ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

ΜΕΘΟΔΟΙ ΕΠΙΛΟΓΗΣ ΜΟΝΤΕΛΩΝ ΣΤΗ ΣΤΑΤΙΣΤΙΚΗ

ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του

Ρεπόπουλου Σοφοκλή

Επιβλέπων: **Φουσκάκης Δημήτριος**

Αθήνα, Ιούλιος 2021

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Δημήτρη Φουσκάκη για την μεγάλη βοήθεια και στήριξη που μου παρείχε καθ΄ όλη τη διάρκεια εκπόνησης της διπλωματικής μου εργασίας. Ένα μεγάλο ευχαριστώ επίσης στον αδελφικό μου φίλο και υποψήφιο διδάκτορα Ιωσήφ Λύτρα για τις επιστημονικές συμβουλές του. Τέλος, ένα θερμό ευχαριστώ στην οικογένειά μου για όλη την υποστήριξη που μου προσέφερε.

Abstract

In modern times we have access to a wealth of data which if used properly can lead to sound decision making and offer insight into problem solving. Statistical models try to explain the possible relationships between the variables-concepts being studied. One of the main issues of the model selection problem is to find the appropriate subset of explanatory variables which leads to models with high predictability and low computational cost. In the present work a number of methods and criteria for model selection are studied in order to effectively predict the variable of interest and detect the factors with the greatest influence.

In the 1^{st} chapter "Basic concepts" we present the fundamental principles of the model selection problem that every analyst must take into account. In addition, the concept of Kullback-Leibler distance is introduced, which is related to the information criteria studied in the next chapter. The linear regression model, which is the most widely used statistical model, is also presented and analyzed along with the R^2 and Mallow's C_p metrics. In chapter 2 "Information Theoretic criteria" we first connect the concept of the Kullback-Leibler distance with the information theoretic criteria, tools critical to model selection. Then we present and analyze the AIC and AIC_c information criteria that we apply to the linear regression model. We also present the Bayesian information criterion BIC and compare it with the previous criteria in a real lifetime data application. At the end of the chapter we analyze the problem of variable selection focusing on the method of full exploration of the modeling space and the stepwise procedures, stating their advantages and disadvantages. In chapter 3 "Information Criteria Properties" we introduce fundamental information criteria properties that are often required. In particular, we examine with the help of relative theorems the properties of the weak and strong consistency of the AIC, AIC_c and BIC criteria. In the 4th chapter "Cross validation and Bootstrap" we introduce the computational methods cross validation and Bootstrap and present the way they are used to estimate the prediction error of the models. We deal with several versions of the cross validation method that we evaluate together with the Boostrap method in a data example. In the 5th chapter "Lasso method" we first analyze the problem of multicollinearity that often appears in the available data and makes it difficult to find an optimal linear model. Next, we present the lasso penalization method which addresses the aforementioned problem and automatically selects variables. In the last chapter "Comparison of methods in simulated data" we perform two different data simulations from the multivariate normal distribution in order to evaluate the performance of the methods presented.

Περίληψη

Στη σύγχρονη εποχή έχουμε πρόσβαση σε πληθώρα δεδομένων τα οποία αν αξιοποιηθούν κατάλληλα μπορούν να οδηγήσουν στην ορθή λήψη αποφάσεων και να προσφέρουν διορατικότητα στην επίλυση προβλημάτων. Τα στατιστικά μοντέλα παλινδρόμησης προσπαθούν να εξηγήσουν τις πιθανές σχέσεις που υπάρχουν ανάμεσα στις μεταβλητές-έννοιες που μελετώνται. Ένα από τα κυριότερα ζητήματα του προβλήματος επιλογής μοντέλων είναι η εύρεση του κατάλληλου υποσυνόλου επεξηγηματικών μεταβλητών, το οποίο οδηγεί κατά επέκταση σε μοντέλα με υψηλή προβλεπτική ικανότητα και μειώνει το υπολογιστικό κόστος στην αξιοποίησή τους.

Στην παρούσα εργασία μελετάται ένα πλήθος από μεθόδους και κριτήρια επιλογής στατιστικών μοντέλων με στόχο την αποτελεσματική πρόβλεψη της μεταβλητής ενδιαφέροντος και την ανίχνευση των παραγόντων με τη μεγαλύτερη επιρροή.

Στο 1ο κεφάλαιο "Βασικές έννοιες" παρουσιάζουμε τις θεμελιώδεις αρχές του προβλήματος επιλογής μοντέλου που πρέπει να λαμβάνει υπόψιν του ο κάθε αναλυτής. Επιπλέον, εισάγεται η έννοια της Kullback-Leibler απόστασης με την οποία συνδέονται τα κριτήρια πληροφορίας που μελετώνται στο επόμενο κεφάλαιο. Επίσης, παρουσιάζεται και αναλύεται το γραμμικό μοντέλο παλινδρόμησης, που είναι το πιο ευρέως διαδεδομένο στατιστικό μοντέλο, μαζί με τα μέτρα καταλληλότητας R^2 και Mallow's-C_p.

Στο 2ο κεφάλαιο "Θεωρητικά κριτήρια πληροφορίας" συνδέουμε αρχικά την έννοια της Kullback-Leibler απόστασης με τα θεωρητικά κριτήρια πληροφορίας, εργαλεία κρίσιμης σημασίας στην επιλογή μοντέλων. Στη συνέχεια παρουσιάζουμε και αναλύουμε τα AIC και AIC_c κριτήρια πληροφορίας τα οποία εφαρμόζουμε στο γραμμικό μοντέλο παλινδρόμησης. Επίσης παρουσιάζουμε το Μπεϋζιανό κριτήριο πληροφορίας BIC και το συγκρίνουμε με τα προηγούμενα κριτήρια σε εφαρμογή με δεδομένα διάρκειας ζωής. Στο τέλος του κεφαλαίου αναλύουμε το πρόβλημα της επιλογής μεταβλητών εστιάζοντας στη μέθοδο της πλήρους εξερεύνησης του χώρου μοντελοποίησης και στις διαδικασίες κατά βήματα, αναφέροντας τα πλεονεκτήματα και τα μειονεκτήματά τους.

Στο 3ο κεφάλαιο "Ιδιότητες κριτηρίων πληροφορίας" εισάγουμε θεμελιώδεις ιδιότητες που είναι θεμιτό να κατέχουν τα κριτήρια πληροφορίας. Συγκεκριμένα, εξετάζουμε με τη βοήθεια αντίστοιχων θεωρημάτων τις ιδιότητες της ασθενούς και ισχυρής συνέπειας των κριτηρίων AIC, AIC_c και BIC.

Στο 4ο κεφάλαιο "Cross Validation και Bootstrap" εισάγουμε τις υπολογιστικές μεθόδους cross validation και Bootstrap και παρουσιάζουμε τον τρόπο

που χρησιμοποιούνται για την εκτίμηση του προβλεπτικού σφάλματος των μοντέλων. Ασχολούμαστε με αρκετές εκδοχές της cross validation μεθόδου τις οποίες αξιολογούμε μαζί με την Boostrap μέθοδο σε παράδειγμα δεδομένων.

Στο 5ο κεφάλαιο "Μέθοδος lasso" αναλύουμε αρχικά το πρόβλημα της πολυσυγγραμμικότητας που εμφανίζεται συχνά στα διαθέσιμα δεδομένα και δυσκολεύει την εύρεση του βέλτιστου γραμμικού μοντέλου. Στη συνέχεια παρουσιάζουμε τη μέθοδο ποινικοποίησης lasso η οποία αντιμετωπίζει το προαναφερθέν πρόβλημα και πραγματοποιεί με αυτόματο τρόπο την επιλογή μεταβλητών.

Στο τελευταίο κεφάλαιο "Σύγκριση μεθόδων σε προσομοιωμένα δεδομένα" πραγματοποιούμε δύο διαφορετικές προσομοιώσεις δεδομένων από την πολυδιάστατη κανονική κατανομή προκειμένου να αξιολογήσουμε την απόδοση των μεθόδων που παρουσιάστηκαν στα πλαίσια της εργασίας.

Περιεχόμενα

Βασικές έννοιες

1.1 Εισαγωγή

Στον χώρο της στατιστικής ανάλυσης καλούμαστε συχνά να πάρουμε αποφάσεις για διάφορα ερευνητικά προβλήματα που μας τίθενται έχοντας στη διάθεσή μας δεδομένα που έχουν συλλεχθεί με βάση το ερευνητικό ερώτημα. Η προσαρμογή και η εύρεση του κατάλληλου στατιστικού μοντέλου είναι ένα πολύ σημαντικό ζήτημα σε αυτές τις περιπτώσεις καθώς οι παράμετροι και οι σχέσεις μεταξύ των μεταβλητών του μπορεί να περιέχουν φυσικό νόημα και να συνεισφέρουν σε μεγάλο βαθμό στην κατανόηση του προβλήματος. Ας υποθέσουμε ότι έχουμε δοσμένου ενός συνόλου δεδομένων για το ζήτημα που μας ενδιαφέρει προσαρμόσει όλα τα δυνατά μοντέλα που μπορούν να προκύψουν μέσω διάφορων τεχνικών. Πως θα διαλέξουμε το βέλτιστο ή τα βέλτιστα εξ αυτών;

Τις τελευταίες δεκαετίες έχουν παρατηρηθεί ραγδαίες εξελίξεις στην ικανότητα προσαρμογής σύνθετων και ευέλικτων μοντέλων αλλά και στην θεωρητική κατανόηση της διαδικασίας με την οποία επιλέγονται τα μοντέλα. Στόχος της διπλωματικής εργασίας είναι να εξεταστούν βασικά κριτήρια και τεχνικές με τις οποίες γίνεται η επιλογή των μοντέλων **(model selection problem)**. Στο πρόβλημα της επιλογής μοντέλου είναι αναγκαία κάποια κριτήρια κατάταξης των εξεταζόμενων μοντέλων. Πολλές φορές ελλοχεύει ο κίνδυνος της υπερπροσαρμογής (overfitting) των μοντέλων στα διαθέσιμα δεδομένα. Ένα επιπλέον αντικείμενο του προβλήματος επιλογής μοντέλου είναι η επιλογή των μεταβλητών που εισάγονται. Απλά μοντέλα με μικρό αριθμό μεταβλητών είναι προτιμότερα από πιο σύνθετα καθώς η ερμηνεία των αποτελεσμάτων είναι ευκολότερη και έχουν χαμηλότερο υπολογιστικό κόστος. Ωστόσο, μπορεί να μην εξηγούν σε ικανοποιητικό βαθμό τη μεταβλητότητα του προβλήματος με συνέπεια να κριθούν αναξιόπιστα.

1.2 Αρχές των στατιστικών μοντέλων

Τα στατιστικά μοντέλα προσαρμόζονται εξ ολοκλήρου με βάση τυχαίο δείγμα από παρατηρήσεις των μεγεθών ενδιαφέροντος του πληθυσμού που μελετάται και για αυτό το λόγο σχεδόν ποτέ ή σπάνια ταυτίζονται με το μοντέλο που γέννησε τα πραγματικά δεδομένα. Αυτό γίνεται περισσότερο αντιληπτό εάν παρατηρήσουμε δεδομένα που προέρχονται παραδείγματος χάριν από βιολογικές διεργασίες όπου τα υπό μελέτη συστήματα είναι πολύπλοκα με πληθώρα αλληλεπιδράσεων και παραγόντων που είναι ανέφικτο να ποσοτικοποιηθούν ή να υπολογιστούν πλήρως. Εντούτοις, ένα στατιστικό μοντέλο βασίζεται σε ένα μικρό μόνο αριθμό παρατηρήσεων της πραγματικότητας που μελετάει, με αποτέλεσμα να την προσεγγίζει μέχρι ένα βαθμό. Από εδώ και στο εξής λοιπόν αναφερόμαστε στην επιλογή μοντέλου αποκλειστικά σε προσεγγιστικά μοντέλα τα οποία περιέχουν κάποια αβεβαιότητα στις μετρήσεις και την ακρίβειά τους.

Παρότι ένα μοντέλο απέχει από την υπό μελέτη πραγματικότητα, μπορούμε να το διακρίνουμε σε πολύ χρήσιμο, χρήσιμο, επαρκές και κακό. Ο Box (1976) διατύπωσε την χαρακτηριστική φράση ότι «*όλα τα μοντέλα είναι λανθασμένα, αλλά κάποια είναι χρήσιμα*». Οι τεχνικές επιλογής μοντέλων προσπαθούν να τα κατατάξουν με βάση το πόσο 'καλά' είναι, έννοια που συνδέεται άμεσα με την ποιότητα των διαθέσιμων δεδομένων αλλά και της *a priori* σκέψης που έχει καταβληθεί κατά τη διάρκεια της μοντελοποίησης.

Ακολουθούν μερικά βασικά στοιχεία τα οποία πρέπει κάθε αναλυτής να λαμβάνει υπόψιν όταν έρχεται σε επαφή με στατιστικά μοντέλα:

Στατιστικό Μοντέλο:

Αρκετά συχνά εμφανίζεται η ανάγκη έρευνας ταυτόχρονα δύο ή περισσότερων χαρακτηριστικών ενός προβλήματος με σκοπό την κατανόηση του τρόπου με τον οποίο τα χαρακτηριστικά αυτά αλληλεπιδρούν μεταξύ τους.

Ας υποθέσουμε για παράδειγμα ότι διαθέτουμε δεδομένα από ασθενείς της μολυσματικής νόσου COVID-19 οι οποίοι εισήχθησαν σε μονάδες εντατικής θεραπείας και θέλουμε να προβλέψουμε το μέσο χρονικό διάστημα σε μέρες παραμονής ενός νέου ασθενή. Τα δεδομένα των ασθενών μπορεί να είναι το ιατρικό τους ιστορικό, η ηλικία, το βιωτικό επίπεδο, η φυσική τους κατάσταση κ.α. τα οποία εκφράζονται με τη χρήση τυχαίων μεταβλητών $X_1, X_2 \dots X_p$, ενώ το χρονικό διάστημα παραμονής τους μέσω της μεταβλητής Y. Οι μεταβλητές X_i , $i = 1 \dots p$, συμβολιζόμενες εν συντομία μέσω του τυχαίου διανύσματος $\mathbf{X} = (X_1, X_2, \dots, X_p)^T$ καλούνται επεξηγηματικές μεταβλητές. Η μεταβλητή Y

καλείται μεταβλητή απόκρισης και θεωρούμε ότι εξαρτάται μέσω κάποιας αιτιακής σχέσης από τις επεξηγηματικές μεταβλητές $\mathbf{X}$. Μαθηματικά η πεποίθησή μας αυτή γράφεται με τη βοήθεια ενός στατιστικού μοντέλου παλινδρόμησης που έχει τη μορφή:

$$Y = g(\mathbf{X}; \boldsymbol{\theta}) + \varepsilon.$$

Σκοπός ενός στατιστικού μοντέλου παλινδρόμησης είναι να προσδιοριστεί η τιμή της τ.μ.(τυχαίας μεταβλητής) Y με βάση τις τιμές που λαμβάνουν οι τ.μ. $\mathbf{X}$. Για το λόγο αυτό θεωρούμε επιπλέον ότι η κατανομή της μεταβλητής Y εκφράζεται από μία δεσμευμένη ή υπό συνθήκη κατανομή, έστω $F(Y|\mathbf{X})$ και η δεσμευμένη μέση τιμή της μεταβλητής απόκρισης Y δίνεται για κάποια τιμή $\boldsymbol{x}$ του τυχαίου διανύσματος $\mathbf{X}$ με τη χρήση του στατιστικού μοντέλου ως:

$$\mathbb{E}[Y|\mathbf{X} = \mathbf{x}] = g(\mathbf{x}; \boldsymbol{\theta}).$$

Η συνάρτηση $g(\mathbf{X}; \boldsymbol{\theta})$ χαρακτηρίζεται από το διάνυσμα παραμέτρων $\boldsymbol{\theta} = (\theta_1, \theta_2 \dots \theta_m)^T$ και η μορφή της είναι άγνωστη. Συλλέγοντας τυχαίο δείγμα από παρατηρήσεις $(y_1, \mathbf{x}_1), \dots, (y_n, \mathbf{x}_n)$ και με βάση θεωρητικές ή εμπειρικές γνώσεις που έχουμε προσδιορίζουμε τη συναρτησιακή μορφή της g. Οι παράμετροι $\boldsymbol{\theta}$ εκτιμούνται από το τυχαίο δείγμα παρατηρήσεων που διαθέτουμε και αντιμετωπίζονται ως σταθερές ποσότητες. Στο μοντέλο έχει προστεθεί ο όρος ε ο οποίος συμβολίζει το τυχαίο σφάλμα που προκύπτει λόγω ελλιπούς πληροφορίας.

Το τυχαίο σφάλμα ε εκφράζει την αβεβαιότητα που έχουμε για τις τιμές της μεταβλητής Y που μας δίνει το μοντέλο g, υπόθεση λογική αν σκεφτούμε ότι η ανάλυσή μας στηρίζεται σε ένα μικρό μέρος του πληθυσμού που μελετάται και σε νέα ζεύγη τιμών $(y, \mathbf{x})$ πιθανότατα το μοντέλο να μην ικανοποιείται επακριβώς. Συνήθως θεωρούμε επιπλέον ότι το τυχαίο σφάλμα ε είναι τυχαία μεταβλητή που ακολουθεί κάποια γνωστή κατανομή με μέση τιμή $\mathbb{E}[\varepsilon] = 0$ και με άγνωστες παραμέτρους τις οποίες προσπαθούμε να εκτιμήσουμε πάλι μέσω του τυχαίου δείγματος.

Διαμάχη διασποράς - μεροληψίας:

Η ισορροπία και η κατάλληλη αλληλεπίδραση μεταξύ της διασποράς και της μεροληψίας μοντελοποίησης είναι ζωτικής σημασίας στο πρόβλημα επιλογής μοντέλου. Η μεροληψία ταυτίζεται με τη διαφορά ανάμεσα στο μέσο όρο της προβλεπόμενης τιμής του μοντέλου και την πραγματική τιμή που προσπαθούμε να προβλέψουμε. Η διασπορά μοντελοποίησης ισοδυναμεί με την αναμενόμενη τετραγωνική απόκλιση της προβλεπόμενης τιμής του μοντέλου γύρω από το μέσο όρο της. Παρακάτω οι δύο έννοιες αναπαρίστανται και μαθηματικά. Σε ένα μοντέλο με λίγες παραμέτρους προς εκτίμηση, η μεταβλητότητα της μεταβλητής

απόκρισης είναι χαμηλή και άρα η διασπορά είναι μικρή, ωστόσο υπεισέρχεται μεγάλη μεροληψία μοντελοποίησης. Αντίθετα, μοντέλα με πολλές παραμέτρους έχουν μικρή μεροληψία αλλά υψηλότερη διασπορά. Η μεγάλη μεροληψία υπεραπλουστεύει τα μοντέλα **(underfitting)** με αποτέλεσμα να γίνεται δύσκολο να γενικευθούν σε όλο τον πληθυσμό. Από την άλλη πλευρά αυξημένη διασπορά δυσκολεύει την ικανότητα εξαγωγής συμπερασμάτων από το μοντέλο καθώς το μοντέλο υπερεστιάζει στα δεδομένα που προσαρμόστηκε **(overfitting)**. Στην επιλογή μοντέλου επιδιώκεται η ισορροπία ανάμεσα στην υπερπροσαρμογή και την υποπροσαρμογή, έννοιες ταυτόσημες με τη διαμάχη απλότητας και πολυπλοκότητας.

Ας υποθέσουμε για λόγους διευκόλυνσης ότι έχουμε προσαρμόσει το μοντέλο $Y = f(X) + \varepsilon$ όπου το τυχαίο σφάλμα $\varepsilon \sim N(0, \sigma_\varepsilon^2)$ και θέλουμε να υπολογίσουμε το αναμενόμενο τετραγωνικό σφάλμα πρόβλεψης του μοντέλου σε δεδομένη σημείο x_0 με αντίστοιχη πραγματική τιμή της μεταβλητής Y την y. Μαθηματικά γράφουμε ότι:

$$\begin{aligned} Err(x_0) &= \mathbb{E}[(y - \hat{f}(x))^2 | X = x_0] = \mathbb{E}[(f(x_0) + \varepsilon - \hat{f}(x_0))^2] \\ &= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] + \mathbb{E}[\varepsilon^2] + 2\mathbb{E}[(f(x_0) - \hat{f}(x_0))\varepsilon]. \end{aligned}$$

Χρησιμοποιώντας την ανεξαρτήσια του τυχαίου σφάλματος ε από την προβλεπόμενη τιμή $\hat{f}(x_0)$ και τις σχέσεις $Var(\varepsilon) = \sigma_\varepsilon^2$, $\mathbb{E}[\varepsilon] = 0$, έχουμε:

$$\begin{aligned} Err(x_0) &= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] + \mathbb{E}[\varepsilon^2] + 2\mathbb{E}[(f(x_0) - \hat{f}(x_0))]\mathbb{E}[\varepsilon] \\ &= \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] + \sigma_\varepsilon^2. \end{aligned} \tag{1}$$

Μελετάμε τώρα τον πρώτο όρο της σχέσης (1):

$$\begin{aligned} \mathbb{E}[(f(x_0) - \hat{f}(x_0))^2] &= \mathbb{E}\left[\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)] - (\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)])\right)^2\right] \\ &= \mathbb{E}\left[\left(\mathbb{E}[\hat{f}(x_0)] - f(x_0)^2\right)\right] + \mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)^2\right] \\ &- 2\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\mathbb{E}\left[\left(\hat{f}(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\right] \\ &= bias^2[\hat{f}(x_0)] + Var(\hat{f}(x_0)) \\ &- 2\left(f(x_0) - \mathbb{E}[\hat{f}(x_0)]\right)\left(\mathbb{E}[\hat{f}(x_0)] - \mathbb{E}[\hat{f}(x_0)]\right) \\ &= bias^2[\hat{f}(x_0)] + Var(\hat{f}(x_0)). \end{aligned}$$

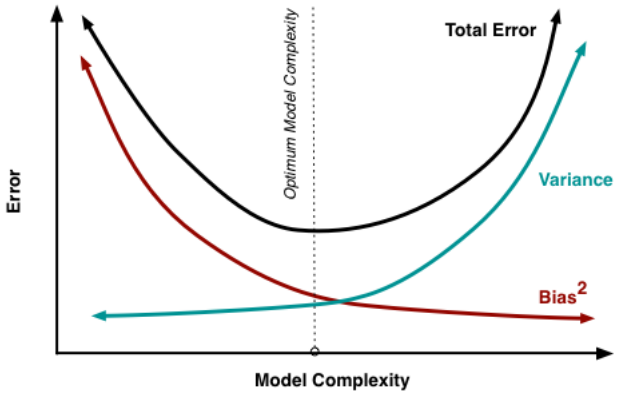
Επιστρέφοντας στη σχέση (1) προκύπτει ότι:

$$Err(x_0) = \sigma_\varepsilon^2 + bias^2[\hat{f}(x_0)] + Var(\hat{f}(x_0)). \quad (2)$$

Ο πρώτος όρος της σχέσης (2) είναι το αμείωτο σφάλμα (irreducible error) το οποίο προκύπτει από το υιοθετημένο μοντέλο και δεν μπορεί να μηδενιστεί καθώς πάντα αναμένουμε σφάλματα στις προβλέψεις μας. Οι υπόλοιποι δύο όροι είναι το τετράγωνο της μεροληψίας μοντελοποίησης και η διασπορά μοντελοποίησης, ποσότητες που λειτουργούν αντιστρόφως ανάλογα με την αύξηση ή μείωση των παραμέτρων του μοντέλου (model complexity).

Αρχή της φειδωλότητας:

Η έννοια της φειδωλότητας (parsimony) συναντάται σε πολλούς κλάδους όπως η φιλοσοφία, η φυσική, η τέχνη και τέλος ακόμα και στην στατιστική. Μια αντιπροσωπευτική φράση είναι το ρητό του Einstein «*όλα τα πράγματα οφείλουν να γίνονται απλά, αλλά όχι απλούστερα*». Στην επιλογή και προσαρμογή μοντέλων η αρχή συνιστά να περιλαμβάνονται οι παράμετροι που πραγματικά συνεισφέρουν στο μοντέλο και δεν προσθέτουν απλώς περισσότερο 'θόρυβο'. Εάν παραδείγματος χάριν η προσθήκη τετραγωνικού όρου σε ένα γραμμικό μοντέλο βελτιώνει σημαντικά την προβλεπτική του ικανότητα και αυτό είναι το κριτήριο επιλογής, τότε είναι σκόπιμο να συμπεριληφθεί αυτός ο όρος, διαφορετικά όχι. Στη διαμάχη διασποράς-μεροληψίας η αρχή της φειδωλότητας αποτελεί σημαντικό εργαλείο διότι μπορεί να οδηγήσει στην χρυσή τομή της αναζητούμενης ισορροπίας όπως φαίνεται και στο Διάγραμμα 1.1:



Διάγραμμα 1.1: Αρχή της φειδωλότητας στη διαμάχη διασποράς - μεροληψίας.

Ερευνητικό πλαίσιο:

Κάθε μοντελοποίηση προέρχεται από κατάλληλο επιστημονικό σχεδιασμό και εξυπηρετεί συγκεκριμένο σκοπό. Είναι σημαντικό να αντιληφθούμε ότι το πλαίσιο στο οποίο γίνεται ο σχεδιασμός των μοντέλων δεν είναι εξαρχής μια επακριβώς καθορισμένη έννοια και είναι πιθανό διαφορετικοί αναλυτές να ανακαλύψουν ή να οδηγηθούν σε διαφορετικά ευρήματα εργαζόμενοι με τα ίδια σύνολα δεδομένων.

Ο Breiman (2001) διαχωρίζει σε δύο «κουλτούρες» τις στατιστικές αναλύσεις. Από τη μία μεριά τα επιστημονικά ευρήματα είναι αντικείμενα πρόβλεψης (prediction) και διακριτοποίησης (classification), και από την άλλη αποσκοπούν σε βαθύτερη κατανόηση της ανάλυσης με χαρακτηριστικό παράδειγμα την εύρεση μιας σημαντικής μη μηδενικής παραμέτρου ακόμα και αν δεν βελτιώνεται η ακρίβεια των αποτελεσμάτων. Συνεπώς οι τεχνικές μοντελοποίησης είναι σχεδιασμένες να απαντούν σε ερωτήσεις που προέρχονται από διαφορετικές «κουλτούρες».

Ποσότητες ενδιαφέροντος:

Πολλές φορές στην επιστήμη της στατιστικής κάποιες ποσότητες, συναρτήσεις ή παράμετροι κρίνονται πιο σημαντικές από άλλες. Για αυτό το λόγο συχνά στην επιλογή μοντέλου οι τεχνικές και τα κριτήρια ευνοούν περισσότερο μοντέλα με αυτές τις ποσότητες. Η εστίαση σε διαφορετικές ποσότητες μπορεί λοιπόν να οδηγήσει σε διαφορετικά επιλεγμένα μοντέλα υπό την ίδια λίστα μοντέλων.

Στάθμιση μοντέλων:

Οι περισσότερες τεχνικές επιλογής μοντέλων εναποθέτουν μία τιμή σκορ σε κάθε υποψήφιο μοντέλο με βάση την οποία επιλέγεται το βέλτιστο. Κάποιες φορές είναι ξεκάθαρο ποιο μοντέλο υπερνικά τα υπόλοιπα και άλλες όχι. Στην δεύτερη περίπτωση που δεν υπάρχει καθαρή εικόνα του μοντέλου-«νικητή» μπορεί ο συνδυασμός όλων των καλών μοντέλων να αποβεί χρήσιμος για την εξαγωγή συμπερασμάτων.

1.3 Kullback-Leibler απόσταση

Ας υποθέσουμε ότι το μοντέλο που περιγράφει «αψεγάδιαστα» τη μεταβλητή, έστω X, που μελετάμε εκφράζεται από την συνάρτηση πυκνότητας πιθανότητας

$f(x;\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_k) \in \mathbb{R}^k$ και πως ένα προσεγγιστικό μοντέλο της κατανομής της X είναι το $g(x;\boldsymbol{\theta^*})$, $\boldsymbol{\theta^*} = (\theta_1^*, \theta_2^*, \ldots, \theta_m^*) \in \mathbb{R}^m$. Θα μας ήταν ιδιαίτερα χρήσιμο ένα μέτρο του πόσο κοντά ή μακριά είναι η κατανομή g από την f, ή αλλιώς πόσο καλά προσαρμόζεται το μοντέλο g στο f.

Ένα τέτοιο μέτρο αποτελεί η **K-L** απόσταση, ή απώλεια πληροφορίας όπως είναι γνωστή, μεταξύ δύο μοντέλων η οποία ορίζεται από τη σχέση:

$$KL(f,g) = \int f(x;\boldsymbol{\theta}) \log\left(\frac{f(x;\boldsymbol{\theta})}{g(x;\boldsymbol{\theta^*})}\right) dx.$$

Στην περίπτωση που οι κατανομές f, g είναι διακριτές η $K-L$ απόσταση γράφεται:

$$KL(f,g) = \sum_{i=1}^{K} p_i \log\left(\frac{p_i}{\pi_i}\right).$$

Ο όρος p_i συμβολίζει την πιθανότητα να συμβεί το $i-$ οστό ενδεχόμενο από τα K και $\pi_1, \pi_2, \ldots, \pi_K$ είναι οι αντίστοιχες πιθανότητες για τα K ενδεχόμενα από το προσεγγιστικό μοντέλο. Ισχύει προφανώς ότι $\sum_{i\in K} p_i = \sum_{i\in K} \pi_i = 1$ και αποδεικνύεται επιπλέον πως $KL(f,g) \geq 0$ για οποιοδήποτε ζεύγος κατανομών f, g, με την περίπτωση $KL(f,g) = 0$ να ισχύει μόνο όταν $f = g$ (Burnham, Anderson, David 2002).

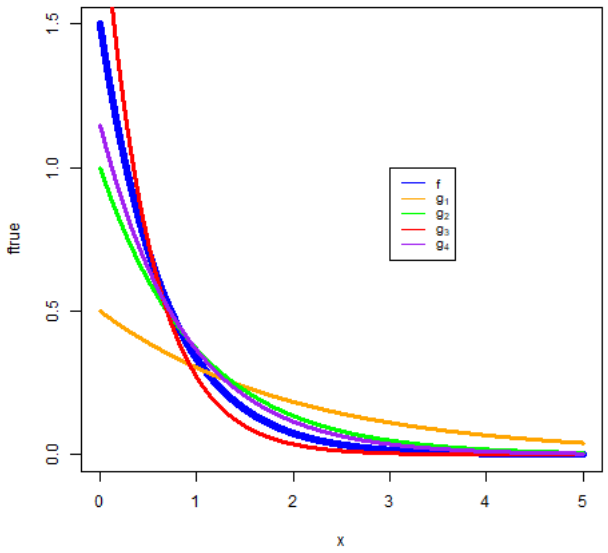
Η K-L απόσταση εκφράζει την πληροφορία που χάνεται όταν χρησιμοποιείται η κατανομή g για να προσεγγιστεί η f και αποτελεί μια επέκταση της εντροπίας που εισήχθη στη θεωρία πληροφορίας από τον ***Shannon***. Η εντροπία κατά $Shannon$ είναι ένα μέτρο της αβεβαιότητας ή της μεταβλητότητας που διακατέχει μία τυχαία μεταβλητή.

Θεμιτά είναι τα μοντέλα που ελαχιστοποιούν όσο περισσότερο γίνεται την απόσταση K-L. Ακολουθεί παράδειγμα αξιοποίησης της K-L απώλειας πληροφορίας:

Θεωρούμε ότι η τυχαία μεταβλητή X ακολουθεί την Εκθετική κατανομή με πεδίο τιμών το $\mathbb{R}^+$, συνάρτηση πυκνότητας πιθανότητας $f(x)$ και παράμετρο ρυθμού $\lambda = 1.5$ και θέλουμε να εξετάσουμε την προσαρμογή τεσσάρων κατανομών-μοντέλων στην προσπάθεια να την προσεγγίσουμε. Οι κατανομές αυτές είναι επίσης εκθετικές με διαφορετικές όμως παραμέτρους ρυθμού. Έχουμε λοιπόν ότι:

- $f(x) = 1.5 \cdot e^{-1.5x}$
- $g_1(x) = 0.5 \cdot e^{-0.5x}$
- $g_2(x) = e^{-x}$

- $g_3(x) = 2 \cdot e^{-2x}$
- $g_4(x) = 1.15 \cdot e^{-1.15x}$



Διάγραμμα 1.2: Καμπύλη συνάρτησης πυκνότητας πιθανότητας f της μεταβλητής X και καμπύλες των προσεγγιστικών συναρτήσεων g_1, g_2, g_3, g_4.

Παρατηρούμε από το Διάγραμμα 1.2 ότι η κατανομή με συνάρτηση πυκνότητας πιθανότητας τη g_1 φαίνεται να αποτελεί την χειρότερη εκ των τεσσάρων προσέγγιση και οι υπόλοιπες συναρτήσεις g_2, g_3, g_4 να συναγωνίζονται για την καλύτερη προσέγγιση της f, χωρίς όμως να είναι ξεκάθαρο ποια είναι η βέλτιστη γραφικά. Υπολογίζουμε στη συνέχεια την K-L απώλεια πληροφορίας κάθε προσεγγιστικής κατανομής από την κατανομή της τυχαίας μεταβλητής X. Έστω $f(x) = \lambda e^{-\lambda x}$ και $g_i(x) = \kappa_i e^{-\kappa_i x}, i = 1, \ldots, 4$. Τα αποτελέσματα δίνονται στον Πίνακα 1.1 και η απόσταση K-L υπολογίζεται ως εξής:

$$KL(f, g_i) = \int_0^{+\infty} \lambda e^{-\lambda x} \log\left(\frac{\lambda e^{-\lambda x}}{\kappa_i e^{-\kappa_i x}}\right) dx = \int_0^{+\infty} \lambda e^{-\lambda x} \log\left(\frac{\lambda}{\kappa_i} \cdot e^{-\lambda x + \kappa_i x}\right) dx$$
$$= -e^{-\lambda x} \log\left(\frac{\lambda}{\kappa_i}\right)\Big|_0^{+\infty} - e^{-\lambda x} x[-\lambda + \kappa_i]\Big|_0^{+\infty} - \left[e^{-\lambda x} \frac{-\lambda + \kappa_i}{\lambda}\right]\Big|_0^{+\infty}$$

Προσεγγιστικό μοντέλο-κατανομή	KL απόσταση	Κατάταξη
Εκθετική ($g_4 : \lambda = 1.15$)	0.03236	1
Εκθετική ($g_3 : \lambda = 2$)	0.04565	2
Εκθετική ($g_2 : \lambda = 1$)	0.07213	3
Εκθετική ($g_1 : \lambda = 0.5$)	0.43194	4

Πίνακας 1.1: Τιμές της K-L απώλειας πληροφορίας μεταξύ της πραγματικής συνάρτησης πυκνότητας πιθανότητας f και των προσεγγιστικών μοντέλων g_1, g_2, g_3, g_4.

Συμπεραίνουμε ότι παρόλο που οι εκθετικές κατανομές $g_3(\lambda = 2)$, $g_4(\lambda = 1.15)$ προσεγγίζουν με βάση το γράφημα κατανομών εξίσου ικανοποιητικά την f, η εκθετική κατανομή (g_4) με παράμετρο $\lambda = 1.15$ είναι αυτή που ελαχιστοποιεί την K-L απόσταση από την κατανομή f της X.

Σε αρκετές περιπτώσεις δεν είναι οφθαλμοφανές το ποιες κατανομές αποτελούν τις καλύτερες προσεγγίσεις και χρειάζονται άλλα μέτρα της καλής προσαρμογής τους. Στο γενικότερο πλαίσιο εφαρμογών οι διαστάσεις του προβλήματος που μελετάται είναι μεγάλες και δεν είναι καν εφικτή κάποια οπτικοποίησή του. Επιπλέον, η πραγματικότητα (f) είναι άγνωστης μορφής συνάρτηση με άγνωστες παραμέτρους και δεν μπορεί να υπάρξει άμεσος υπολογισμός της K-L απώλειας πληροφορίας των μοντέλων που χρησιμοποιούνται. Ωστόσο, παρακάτω θα δούμε ότι η K-L απόσταση αποτελεί την βάση αρκετών κριτηρίων πληροφορίας, κρίσιμων στο πρόβλημα της επιλογής μοντέλου.

1.4 Πολλαπλό γραμμικό μοντέλο

Το γραμμικό μοντέλο αποτελεί το πιο ευρέως διαδεδομένο στατιστικό μοντέλο με μεγάλη συνεισφορά στις εφαρμογές. Η γενική μορφή του πολλαπλού γραμμικού μοντέλου είναι:

$$Y = g(\boldsymbol{X}; \boldsymbol{\beta}) + \varepsilon = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon.$$

Με Y συμβολίζουμε την μεταβλητή απόκρισης η οποία εξαρτάται από το διάνυσμα επεξηγηματικών μεταβλητών $\boldsymbol{X} = (X_1, \ldots, X_p)$ μέσω της παραπάνω σχέσης. Διαθέτοντας δείγμα μεγέθους n από παρατηρήσεις $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$ η σχέση του γραμμικού μοντέλου για κάθε τιμή i της μεταβλητής απόκρισης Y δίνεται ως:

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + b_p x_{ip} + \varepsilon_i. \tag{3}$$

Η σχέση (3) μπορεί να γραφτεί υπό τη μορφή πινάκων ως εξής:

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

Το διάνυσμα $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)^T$ είναι το $n \times 1$ διάνυσμα παρατηρούμενων τιμών της μεταβλητής Y. Με $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$ συμβολίζουμε το $(p+1) \times 1$ διάνυσμα παραμέτρων του μοντέλου, $\boldsymbol{\varepsilon} = (\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_n)^T$ το $n \times 1$ διάνυσμα των τυχαίων σφαλμάτων και $\tilde{\boldsymbol{X}}$ είναι ο $n \times (p+1)$ πίνακας σχεδιασμού όπως ονομάζεται, με μορφή:

$$\tilde{\boldsymbol{X}} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \vdots & \ldots & \vdots \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np}. \end{pmatrix}$$

Η βασική υπόθεση του μοντέλου είναι ότι το διάνυσμα των τυχαίων σφαλμάτων $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\boldsymbol{I})$, όπου $\boldsymbol{I}$ ο $n \times n$ μοναδιαίος πίνακας και επομένως $\boldsymbol{Y}|\tilde{\boldsymbol{X}} \sim N_n(\tilde{\boldsymbol{X}}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$.

Οι άγνωστες παράμετροι $\boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ του μοντέλου πρέπει να εκτιμηθούν με βάση το δείγμα των παρατηρήσεων. Στη συνέχεια παρουσιάζεται η κλασσική μέθοδος εκτίμησής τους, η μέθοδος των ελαχίστων τετραγώνων.

Γραμμικό μοντέλο και μέθοδος ελαχίστων τετραγώνων

Συμβολίζοντας με $\mathbb{E}[y_i] = \mathbb{E}[y_i|\boldsymbol{X} = \boldsymbol{x}_i] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip}$, την αναμενόμενη τιμή της παρατηρούμενης τιμής y_i και ισοδύναμα με $\mathbb{E}[\boldsymbol{y}] = (\mathbb{E}[y_1], \ldots, \mathbb{E}[y_n])^T$, η μέθοδος ελαχίστων τετραγώνων έγκειται στην ελαχιστοποίηση της παράστασης:

$$S(\boldsymbol{\beta}) = \sum_{i=1}^{n} (y_i - \mathbb{E}[y_i])^2.$$

Η παράσταση $S(\boldsymbol{\beta})$ υπό μορφή πινάκων γράφεται:

$$\begin{aligned} S(\boldsymbol{\beta}) &= (\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}])^T (\boldsymbol{y} - \mathbb{E}[\boldsymbol{y}]) \\ &= (\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})^T (\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) \\ &= \boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta} - \boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\boldsymbol{y} + \boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta}. \end{aligned}$$

Παρατηρώντας ότι ο όρος $\boldsymbol{y}^T$ είναι $1 \times n$ διάνυσμα, ο όρος $\tilde{\boldsymbol{X}}$ ένας $n \times (p+1)$ πίνακας και ο όρος $\boldsymbol{\beta}$ ένα $(p+1) \times 1$ διάνυσμα, ο όρος $\boldsymbol{y}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta}$ θα είναι ένας

1×1 πίνακας και επομένως ισούται με τον ανάστροφό του:

$$\boldsymbol{y}^T \tilde{\boldsymbol{X}} \boldsymbol{\beta} = (\boldsymbol{y}^T \tilde{\boldsymbol{X}} \boldsymbol{\beta})^T = \boldsymbol{\beta}^T \tilde{\boldsymbol{X}}^T \boldsymbol{y}.$$

Η παράσταση συνεπώς παίρνει τη μορφή:

$$S(\boldsymbol{\beta}) = \boldsymbol{y}^T \boldsymbol{y} - 2\boldsymbol{\beta}^T \tilde{\boldsymbol{X}}^T \boldsymbol{y} + \boldsymbol{\beta}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{\beta}.$$

Παραγωγίζοντας την τελευταία σχέση ως προς $\boldsymbol{\beta}$ και χρησιμοποιώντας κανόνες παραγώγισης διανυσμάτων έχουμε:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = -2\tilde{\boldsymbol{X}}^T \boldsymbol{y} + 2\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{\beta} \text{ χρησιμοποιώντας τις σχέσεις:}$$
$$\frac{\partial(\boldsymbol{\beta}^T \tilde{\boldsymbol{X}}^T \boldsymbol{y})}{\partial \boldsymbol{\beta}} = \tilde{\boldsymbol{X}}^T \boldsymbol{y}, \quad \frac{\partial \boldsymbol{\beta}^T \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{\beta}}{\partial \boldsymbol{\beta}} = 2\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{\beta}.$$

Εξισώνοντας με το μηδέν έχουμε:

$$\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = 0 \leftrightarrow \tilde{\boldsymbol{X}}^T \boldsymbol{y} = \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{\beta}.$$

Αν ο πίνακας $\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}$ είναι αντιστρέψιμος τότε η εκτιμήτρια των παραμέτρων $\boldsymbol{\beta}$ από τη μέθοδο θα είναι:

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

Αν οι επεξηγηματικές μεταβλητές $X_1, \ldots, X_p$ δεν είναι τέλεια γραμμικά εξαρτημένες, ισοδύναμο με $rank(\tilde{\boldsymbol{X}}) = p + 1$, τότε ο πίνακας $\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}}$ θα αντιστρέφεται. Η πιο σημαντική ιδιότητα της εκτιμήτρια $\hat{\boldsymbol{\beta}}$ είναι ότι είναι αμερόληπτη ($\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}$), γεγονός που αποδεικνύεται ως εξής:

$$\begin{aligned} \mathbb{E}[\hat{\boldsymbol{\beta}}] &= \mathbb{E}[(\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{y}] \\ &= (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \mathbb{E}[\boldsymbol{y}] \\ &= (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \mathbb{E}[\tilde{\boldsymbol{X}} \boldsymbol{\beta} + \boldsymbol{\varepsilon}] \\ &= (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}} \boldsymbol{\beta} \\ &= \boldsymbol{\beta}. \end{aligned} \tag{4}$$

Καταλήγουμε λοιπόν στην εκτίμηση του προσαρμοσμένου γραμμικού μοντέλου με τις προβλεπόμενες τιμές $\mathbb{E}[\boldsymbol{y}] = \hat{\boldsymbol{y}} = (\hat{y_1}, \ldots, \hat{y_n})$ να δίνονται από τη

σχέση:

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

Ο πίνακας $\boldsymbol{H} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T)^{-1}\tilde{\boldsymbol{X}}^T$ καλείται πίνακας προβολής (hat matrix) και είναι συμμετρικός και ταυτοδύναμος. Επιπλέον ισχύει ότι:

$$tr(\boldsymbol{H}) = p + 1.$$

Οι ποσότητες $\boldsymbol{e} = (y_1 - \hat{y_1}, \ldots, y_n - \hat{y_n})^T = (e_1, \ldots, e_n)^T$ ονομάζονται υπόλοιπα και αποτελούν εκτιμήσεις των τυχαίων σφαλμάτων $\boldsymbol{\varepsilon} \sim N_n(0, \sigma^2\boldsymbol{I}_n)$. Επιπλέον, για το διάνυσμα των υπολοίπων έχουμε ότι:

$$\begin{aligned}\boldsymbol{e} &= \boldsymbol{y} - \hat{\boldsymbol{y}} = \boldsymbol{y} - \boldsymbol{H}\boldsymbol{y} = (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{y} \\ &= (\boldsymbol{I} - \boldsymbol{H})(\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \varepsilon) = \tilde{\boldsymbol{X}}\boldsymbol{\beta} - \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon} - \boldsymbol{H}\boldsymbol{\varepsilon} \\ &= (\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}.\end{aligned}$$

και

$$\begin{aligned}\mathbb{E}[\boldsymbol{e}] &= \mathbb{E}[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}] = 0, \\ V(\boldsymbol{e}) &= V[(\boldsymbol{I} - \boldsymbol{H})\boldsymbol{\varepsilon}] = (\boldsymbol{I} - \boldsymbol{H})V(\boldsymbol{\varepsilon})(\boldsymbol{I} - \boldsymbol{H})^T \\ &= \sigma^2(\boldsymbol{I} - \boldsymbol{H}), \quad \text{αφού ο } \boldsymbol{I} - \boldsymbol{H} \text{ είναι συμμετρικός και ταυτοδύναμος.}\end{aligned}$$

Συνεπώς $\boldsymbol{e} \sim N_n(0, \sigma^2(\boldsymbol{I} - \boldsymbol{H}))$ και $e_i \sim N(0, \sigma^2(1 - h_{ii}))$ όπου $h_{ii} = \boldsymbol{x}_i^T(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\boldsymbol{x}_i$ τα διαγώνια στοιχεία του πίνακα $\boldsymbol{H}$. Την άγνωστη διασπορά σ^2 των τυχαίων σφαλμάτων την εκτιμούμε από το μέσο τετραγωνικό σφάλμα το οποίο αποτελεί αμερόληπτη εκτιμήτρια της διασποράς σ^2 και ορίζεται ως:

$$s^2_{y|\boldsymbol{x}} = \frac{1}{n-p-1}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n-p-1}\sum_{i}^{n} e_i^2.$$

Η αμεροληψία της εκτιμήτριας $s^2_{y|\boldsymbol{x}}$ προκύπτει ως εξής:

$$\mathbb{E}[s^2_{y|\boldsymbol{x}}] = \frac{1}{n-p-1}\sum_{i=1}^{n}\mathbb{E}[e_i^2],$$

όμως $Var(e_i) = \mathbb{E}[e_i^2] - \mathbb{E}[e_i]^2 = \mathbb{E}[e_i^2]$, συνεπώς:

$$\begin{aligned}\mathbb{E}[s^2_{y|\boldsymbol{x}}] &= \frac{1}{n-p-1}\sum_{i=1}^{n}\sigma^2(1-h_{ii}) \\ &= \frac{1}{n-p-1}\sum_{i=1}^{n}(\sigma^2-\sigma^2 h_{ii}) \\ &= \frac{1}{n-p-1}\sigma^2(n-(p+1)) = \sigma^2.\end{aligned}$$

Το μέσο τετραγωνικό σφάλμα $s^2_{y|\boldsymbol{x}}$ συχνά χρησιμοποιείται για την αξιολόγηση των μοντέλων. Όσο μικρότερες τιμές παίρνει, τόσο περισσότερο θεωρούμε ότι βελτιώνεται η προβλεπτική ικανότητα των γραμμικών μοντέλων. Ωστόσο, μία περισσότερο αξιόπιστη εκτίμηση του προβλεπτικού σφάλματος θα δούμε ότι προκύπτει χρησιμοποιώντας υπολογιστικές μεθόδους επαναδειγματοληψίας.

1.4.1 Μέτρα καλής προσαρμογής

Συντελεστής $\mathbf{R}^2$

Ένα μέτρο αξιολόγησης της προσαρμογής του γραμμικού μοντέλου παλινδρόμησης είναι ο συντελεστής προσδιορισμού (coefficient of determination) R^2 που ορίζεται από τη σχέση:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{\sum_{i=1}^{n}(y_i-\bar{y})^2}. \tag{5}$$

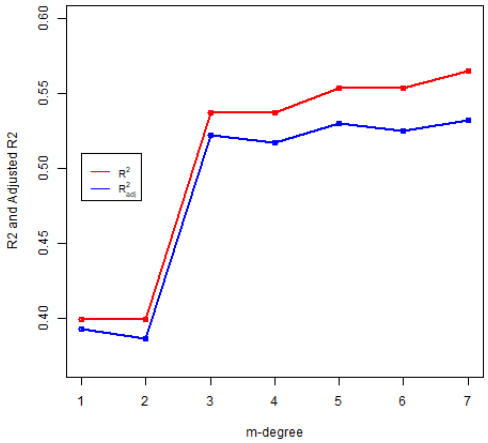
Ο συντελεστής R^2 λαμβάνει τιμές στο διάστημα $[0,1]$ και εκφράζει το ποσοστό της διασποράς της τ.μ. Y που εξηγείται με βάση το προσαρμοσμένο γραμμικό μοντέλο. Όσο πιο μεγάλες τιμές παίρνει τόσο ισχυρότερη είναι η γραμμική σχέση εξάρτησης των τ.μ. Y και $X_1, X_2, \ldots, X_p$ υπό την προϋπόθεση όμως ότι το γραμμικό μοντέλο είναι το κατάλληλο. Θα δούμε ωστόσο ότι ο συντελεστής προσδιορισμού εσφαλμένα χρησιμοποιείται ως μέτρο καλής προσαρμογής του μοντέλου στα δεδομένα ή ως μέτρο σύγκρισης δύο μοντέλων. Προσθέτοντας περισσότερες μεταβλητές σε ένα γραμμικό μοντέλο ο συντελεστής αυξάνεται (έστω και απειροελάχιστα) δίνοντάς μας την εσφαλμένη εντύπωση ότι το νέο πιο πολύπλοκο μοντέλο είναι καταλληλότερο.

Για το λόγο αυτό είναι περισσότερο αξιόπιστος στην έκφραση του ποσοστού της διασποράς της τ.μ. Y που εξηγείται από το μοντέλο ο προσαρμοσμένος συντελεστής προσδιορισμού (adjusted coefficient of determination) ο οποίος

λαμβάνει υπόψιν και την πολυπλοκότητα του γραμμικού μοντέλου:

$$R^2_{adj} = 1 - \frac{(n-1)\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{(n-p-1)\sum_{i=1}^{n}(y_i - \bar{y})^2} = 1 - [1 - R^2]\frac{n-1}{n-p-1}. \qquad (6)$$

Με την προσθήκη περισσότερων επεξηγηματικών μεταβλητών στο μοντέλο ενδέχεται ο προσαρμοσμένος συντελεστής να έχει μικρότερη τιμή από ότι αρχικά.



Διάγραμμα 1.3: R^2 και R^2_{adj} συντελεστές προσδιορισμού συναρτήσει του βαθμού (m-degree) του πολυωνυμικού μοντέλου της Παραγράφου 4.5.

Στο διάγραμμα 1.3 απεικονίζονται οι R^2 και R^2_{adj} συντελεστές προσδιορισμού στα γραμμικά μοντέλα της Παραγράφου 4.5. Παρατηρούμε ότι από το μοντέλο με 3 παραμέτρους και μετά ο συντελεστής R^2 αυξάνεται απότομα, οδηγώντας ενδεχομένως εσφαλμένα στην επιλογή του πολυπλοκότερου μοντέλου M_7. Αντίθετα, ο συντελεστής R^2_{adj} μετά το μοντέλο M_3 δείχνει περισσότερη ανθεκτικότητα στην αύξηση της πολυπλοκότητας των μοντέλων.

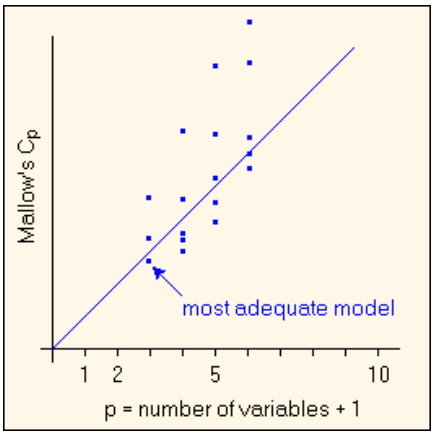
Mallows's C_p

Ένα δεύτερο μέτρο αξιολόγησης της καταλληλότητας του γραμμικού μοντέλου παλινδρόμησης είναι η στατιστική συνάρτηση του Mallows (1973):

$$\begin{aligned} C_p &= \frac{SSE_p}{\hat{\sigma}^2} + 2p' - n \quad \text{με εναλλακτικη μορφή:} \\ C_p &= \frac{SSE_p}{n} + \frac{2}{n}p'\hat{\sigma}^2. \end{aligned} \qquad (7)$$

Ο όρος $SSE_p = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ είναι το τετραγωνικό άθροισμα των υπολοίπων του μοντέλου με p επεξηγηματικές μεταβλητές, n είναι ο αριθμός των διαθέσιμων παρατηρήσεων και p' ο αριθμός των προς εκτίμηση συντελεστών του μοντέλου ο οποίος ισούται με $p' = p+1$. Η διασπορά $\hat{\sigma}^2$ εκτιμάται συνήθως από το μέσο τετραγωνικό σφάλμα $MSE = \frac{\sum_{i=1}^{n}(y_i-\hat{y}_i)^2}{n}$ του πλήρους μοντέλου με όλες τις επεξηγηματικές μεταβλητές. Οι δύο μορφές του κριτηρίου C_p δεν δίνουν τις ίδιες τιμές αλλά η εφαρμογή τους σε κοινό σύνολο επιλογής μοντέλων οδηγεί στο ίδιο αποτέλεσμα.

Προσαρμόζοντας και υπολογίζοντας το κριτήριο C_p για όλα τα υποψήφια μοντέλα, κάνουμε τη γραφική παράσταση των τιμών του C_p με τις αντίστοιχες τιμές p του αριθμού των επεξηγηματικών μεταβλητών. Το σημείο του διαγράμματος με την χαμηλότερη τιμή C_p η οποία είναι περίπου ίση με το αντίστοιχο p ισοδυναμεί με το καταλληλότερο μοντέλο. Για αυτό το λόγο προστίθεται στο Διάγραμμα 1.4 η ευθεία $y = x$ και αναζητούμε το σημείο με τη χαμηλότερη τιμή C_p που απέχει τη μικρότερη απόσταση από την ευθεία.



Διάγραμμα 1.4: Επιλογή του καταλληλότερου μοντέλου μέσω του κριτηρίου C_p.

Κεφάλαιο 2

Θεωρητικά κριτήρια πληροφορίας

Στο προηγούμενο κεφάλαιο ήρθαμε σε επαφή με την Kullback Leibler απώλεια πληροφορίας μεταξύ δύο μοντέλων και διαπιστώσαμε ότι για τον υπολογισμό της πρέπει να είναι οι γνωστές οι κατανομές f, g των μοντέλων καθώς και οι παράμετροι αυτών. Σπάνια ωστόσο είναι γνωστή η αναζητούμενη πραγματικότητα f και δεν μπορεί να οριστεί επακριβώς. Στο κεφάλαιο αυτό εισάγουμε και μελετάμε τα θεωρητικά κριτήρια πληροφορίας τα οποία αποτελούν ευρέως διαδεδομένα κριτήρια επιλογής μοντέλων με αρκετά να συνδέονται με την K-L απόσταση και να εμπεριέχουν την έννοια της φειδωλότητας στον τρόπο λειτουργίας τους. Τα περισσότερα κριτήρια πληροφορίας συνδέουν την μεγιστοποιημένη πιθανοφάνεια ως μέτρο καλής προσαρμογής των μοντέλων μαζί με κάποιους όρους ποινικοποίησης. Αρχικά παρουσιάζουμε τον τρόπο υπολογισμού των εκτιμητών μέγιστης πιθανοφάνειας μέσα από απλό παράδειγμα.

2.1 Εκτιμητές μέγιστης πιθανοφάνειας

Έστω τυχαίο δείγμα $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ όπου κάθε μεταβλητή ακολουθεί κατανομή με σ.π.π. ή σ.μ.π. (αν είναι διακριτή) την $p(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_m)^T \in \mathbb{R}^m$. Ορίζουμε ως συνάρτηση πιθανότητας του τυχαίου δείγματος την ποσότητα:

$$p(\mathbf{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta}).$$

Για γνωστές τιμές $x_i, i = 1, \ldots, n$ από τυχαίο δείγμα των τυχαίων μεταβλητών $X_i, i = 1, \ldots, n$, η συνάρτηση $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} p(x_i; \boldsymbol{\theta})$ καλείται συνάρτηση πιθανοφάνειας και μας δίνει την πιθανότητα το δείγμα μας να προέρχεται από

την κατανομή $p(\boldsymbol{x};\boldsymbol{\theta})$ με παράμετρο $\boldsymbol{\theta}$.

Η συνάρτηση $L(\boldsymbol{\theta})$ για δεδομένες τιμές x_i είναι συνάρτηση του διανύσματος παραμέτρων $\boldsymbol{\theta}$ και εκφράζει το πόσο σύμφωνες είναι οι διάφορες τιμές των παραμέτρων $\boldsymbol{\theta}$ με το συγκεκριμένο δείγμα παρατηρούμενων τιμών. Οι εκτιμήτριες μέγιστης πιθανοφάνειας (Ε.Μ.Π.) είναι οι εκτιμήσεις των παραμέτρων $\boldsymbol{\theta}$ οι οποίες μεγιστοποιούν την συνάρτηση $L(\boldsymbol{\theta})$. Για διευκόλυνση στις πράξεις αλλά και εξοικονόμηση κόστους συνήθως μεγιστοποιούμε την λογαριθμική (με βάση τον νεπέριο λογάριθμο) πιθανοφάνεια $\boldsymbol{\ell(\theta)} = \log L(\boldsymbol{\theta})$ η οποία είναι αύξουσα συνάρτηση της $L(\boldsymbol{\theta})$ και για αυτό το λόγο η μεγιστοποίησή της είναι ισοδύναμη με την μεγιστοποίηση της $L(\boldsymbol{\theta})$.

Παράδειγμα εύρεσης Ε.Μ.Π.:

Έστω τυχαίο δείγμα $\boldsymbol{x} = (x_1, \ldots, x_n)$ από ανεξάρτητες και ισόνομες τ.μ. $X_i, i = 1, \ldots, n$, που θεωρούμε ότι προέρχονται από την κατανομή $Poisson$ με $X_i \in \mathbb{N}_0$ και παράμετρο λ. Η συνάρτηση πιθανοφάνειας του $Poisson$ μοντέλου είναι:

$$L(\lambda) = \prod_{i=1}^{n} e^{-\lambda}\frac{\lambda^{x_i}}{x_i!} = e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}.$$

Η λογαριθμοποιημένη συνάρτηση πιθανοφάνειας είναι:

$$\boldsymbol{\ell}(\lambda) = \log\left\{e^{-n\lambda}\frac{\lambda^{\sum_{i=1}^{n} x_i}}{\prod_{i=1}^{n} x_i!}\right\} = -n\lambda + \sum_{i=1}^{n} x_i \log(\lambda) - \sum_{i=1}^{n} \log(x_i!).$$

Παραγωγίζοντας ως προς λ και εξισώνοντας με το 0 έχουμε:

$$\frac{\partial \boldsymbol{\ell}(\lambda)}{\partial \lambda} = -n + \frac{1}{\lambda}\sum_{i=1}^{n} x_i = 0 \quad \leftrightarrow \quad \hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Επιπλέον:

$$\frac{\partial^2 \boldsymbol{\ell}(\lambda)}{\partial \lambda^2} = -\frac{1}{\lambda^2}\sum_{i=1}^{n} x_i \quad \text{και} \quad \frac{\partial^2 \boldsymbol{\ell}(\lambda)}{\partial \lambda^2}\Big|_{\lambda=\hat{\lambda}} = -\frac{n^2}{\sum_{i=1}^{n} x_i} < 0.$$

Επομένως επιτυγχάνεται μέγιστο σημείο και η εκτιμήτρια λοιπόν της μέγιστης πιθανοφάνειας του μονοδιάστατου μοντέλου $Poisson(\lambda)$ είναι η $\hat{\lambda} = \frac{1}{n}\sum_{i=1}^{n} x_i = \bar{x}$.

2.2 Μέση λογαριθμική πιθανοφάνεια και K-L απόσταση

Υποθέτουμε ότι διαθέτουμε διάνυσμα $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ από συνεχείς τυχαίες μεταβλητές με σ.π.π. την συνάρτηση $f(x; \boldsymbol{\theta})$, $\boldsymbol{\theta} \in \mathbb{R}^k$ για κάθε μεταβλητή X. Σκοπός μας είναι να υπολογίσουμε την απόσταση του προσεγγιστικού μοντέλου $g(x; \boldsymbol{\theta}^*)$, $\boldsymbol{\theta}^* \in \mathbb{R}^p$ από την πραγματικότητα f. Η K-L απώλεια πληροφορίας γράφεται ως:

$$\begin{aligned} KL(f, g) &= \int f(x; \boldsymbol{\theta}) \log\left(\frac{f(x; \boldsymbol{\theta})}{g(x; \boldsymbol{\theta}^*)}\right) dx \\ &= \int f(x; \boldsymbol{\theta}) \log(f(x; \boldsymbol{\theta}))\, dx - \int f(x; \boldsymbol{\theta}) \log(g(x; \boldsymbol{\theta}^*))\, dx. \end{aligned}$$

Η ποσότητα $\mathbb{E}_f[\log(f(x; \boldsymbol{\theta}))] = \int f(x; \boldsymbol{\theta}) \log(f(x; \boldsymbol{\theta}))\, dx$ ονομάζεται κατά *Shannon* αρνητική εντροπία και είναι σταθερή για δεδομένη $f(x; \boldsymbol{\theta})$ με θετικές μόνο τιμές.

Σε πραγματικά προβλήματα η απόσταση $KL(f, g)$ δεν υπολογίζεται άμεσα διότι η μορφή της πραγματικότητας f και οι παράμετροι $\boldsymbol{\theta}, \boldsymbol{\theta}^*$ είναι άγνωστες. Με τη βοήθεια δεδομένων $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ από τυχαίο δείγμα ανεξάρτητων και ισόνομων τυχαίων μεταβλητών που θεωρούμε ότι προέρχονται από την $g(x; \boldsymbol{\theta}^*)$, ορίζουμε την μέση λογαριθμική πιθανοφάνεια του δείγματος, η οποία εκφράζει το πόσο καλά προσαρμόζεται το μοντέλο g στα δεδομένα αυτά για τις δάφορες τιμές του $\boldsymbol{\theta}^*$ ως:

$$I_n = \frac{1}{n}\boldsymbol{\ell}(\boldsymbol{\theta}^*) = \frac{1}{n}\log(L(\boldsymbol{\theta}^*)) = \frac{1}{n}\sum_{i=1}^{n} \log g(x_i; \boldsymbol{\theta}^*). \tag{8}$$

Από τη σχέση (8) και εφαρμόζοντας τον ισχυρό νόμο των μεγάλων αριθμών έχουμε πλέον ότι:

$$I_n(\boldsymbol{\theta}^*) = \frac{1}{n}\big[\log g(x_1; \boldsymbol{\theta}^*) + \cdots + \log g(x_n; \boldsymbol{\theta}^*)\big] \xrightarrow{n} \mathbb{E}_f[\log(g(X; \boldsymbol{\theta}^*))]. \tag{9}$$

Η σύγκλιση της σχέσης (9) ισχύει σχεδόν παντού με πιθανότητα 1.

Χρησιμοποιώντας επιπλέον τα δεδομένα $\mathbf{x}$ παίρνουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\theta}}^*$ του παραμετρικού διανύσματος $\boldsymbol{\theta}^*$, και μία εκτίμηση της $KL(f, g)$ απόστασης του μοντέλου g από το f είναι:

$$\hat{KL}(f, g) = C - I_n(\hat{\boldsymbol{\theta}}^*), \quad C = \int f(x; \boldsymbol{\theta}) \log(f(x; \boldsymbol{\theta}))\, d\mathbf{x}. \tag{10}$$

Βλέπουμε από την τελευταία σχέση ότι η μεγιστοποίηση της αναμενόμενης μέσης πιθανοφάνειας $I_n(\hat{\boldsymbol{\theta}}^*)$ είναι ισοδύναμη με την ελαχιστοποίηση της απόστασης $\hat{KL}(f,g)$. Ο όρος C είναι σταθερός και δεν μεταβάλλεται για οποιοδήποτε προσεγγιστικό μοντέλο g χρησιμοποιούμε. Επιπλέον, οι εκτιμητές $\hat{\boldsymbol{\theta}}^*$ που μεγιστοποιούν την μέση λογαριθμική πιθανοφάνεια, παρέχουν την καλύτερη παραμετρική προσέγγιση της πραγματικής σ.π.π. $f(x;\boldsymbol{\theta})$ υπό την υπόθεση του προσαρμοσμένου μοντέλου $g(x;\boldsymbol{\theta}^*)$.

Αν συμβολίσουμε με $\boldsymbol{\theta_0^*}$ το διάνυσμα των παραμέτρων της g που ελαχιστοποιεί την $KL(f,g)$ απόσταση από την f, τότε ασυμπτωτικά θα ισχύει ότι:

$$\hat{\boldsymbol{\theta}}^* \to \boldsymbol{\theta_0^*} = \arg\min_{\boldsymbol{\theta}^*}\{KL(f, g(x;\boldsymbol{\theta}^*))\}. \tag{11}$$

Σημειώνουμε ότι εάν το παραμετρικό μοντέλο g ταυτίζεται με την f τότε θα ισχύει ότι $f(x;\boldsymbol{\theta}) = g(x;\boldsymbol{\theta_0^*})$ και η απόσταση $KL(f,g)$ σε αυτή την περίπτωση ισούται με μηδέν.

Συνοψίζοντας, η εκτιμήτρια $\hat{KL}(f,g)$ ελαχιστοποιείται από τους εκτιμητές $\hat{\boldsymbol{\theta}}^*$ της μέγιστης λογαριθμικής πιθανοφάνειας, οι οποίοι όμως με τη σειρά τους εξαρτώνται από το εκάστοτε δείγμα που χρησιμοποιείται. Για αυτό το λόγο η εκτιμήτρια $\hat{KL}(f,g)$ εμπεριέχει σφάλμα το οποίο εξισορροπείται με όρους ποινικοποίησης στα κριτήρια πληροφορίας που αναπτύσσονται στη συνέχεια.

Στην αναζήτησή μας για το βέλτιστο προσεγγιστικό μοντέλο παρατηρούμε ότι είναι αναγκαίο να εκτιμήσουμε τις παραμέτρους των υποψήφιων μοντέλων από το δείγμα που διαθέτουμε. Όμως, ακόμα και αν για κάποιο προσεγγιστικό μοντέλο g ισχύει ότι $g(x;\boldsymbol{\theta}^*) = f(x;\boldsymbol{\theta})$, είναι σχεδόν βέβαιο ότι η εκτίμηση από το τυχαίο δείγμα του διανύσματος παραμέτρων του μοντέλου g θα διαφέρει της τιμής $\boldsymbol{\theta}^*$ αποτέλεσμα που φαίνεται και από την σχέση (11) η οποία ισχύει μόνο ασυμπτωτικά. Εντούτοις αναμένουμε πάντα η εκτιμήτρια $\hat{KL}(f,g)$ να μεροληπτεί της πραγματικής $KL(f,g)$ απόστασης του μοντέλου g από την πραγματικότητα f.

Αυτή τη συλλογιστική πορεία ακολούθησε ο **Akaike** (1974) από όπου προήλθε το κριτήριο πληροφορίας AIC (Akaike's Information Criterion). Σε αυτό το σημείο ορίζουμε κάποιες βασικές έννοιες-εργαλεία που θα χρειαστούμε για την περιγραφή των θεωρητικών κριτηρίων πληροφορίας.

Έστω συνεχής τυχαία μεταβλητή Y με σ.π.π. $g(y;\boldsymbol{\theta})$, $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p) \in \mathbb{R}^p$. Ορίζουμε τις τ.μ.:

$$u(Y,\boldsymbol{\theta}) = \frac{\partial \log g(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} \quad \text{και} \quad I(Y,\boldsymbol{\theta}) = \frac{\partial^2 \log g(Y;\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}. \tag{12}$$

Η πρώτη ποσότητα καλείται διάνυσμα επίδοσης (score vector) διάστασης p με πραγματικές τιμές και η δεύτερη ονομάζεται πίνακας πληροφορίας του μο-

ντέλου g με διαστάσεις $p \times p$ και πραγματικές τιμές που δίνονται από τα στοιχεία : $\frac{\partial^2 \log g(Y;\boldsymbol{\theta})}{\partial \theta_j \partial \theta_k} \quad j,k = 1,\ldots,p$. Οι δύο έννοιες χρησιμοποιούνται προκειμένου να βρεθούν αριθμητικά εκτιμητές μέγιστης πιθανοφάνειας και να μελετηθεί η συμπεριφορά τους. Χρησιμοποιώντας τις δύο ποσότητες θέτουμε:

$$J = -\mathbb{E}_f[I(Y,\boldsymbol{\theta}_0)] \quad \text{και} \quad K = Var_f u(Y,\boldsymbol{\theta}_0). \tag{13}$$

Το διάνυσμα $\boldsymbol{\theta}_0$ εκπροσωπεί τις τιμές των παραμέτρων που ελαχιστοποιούν την απώλεια πληροφορίας του μοντέλου f από το παραμετρικό μοντέλο g κατά $K - L$. Όταν η πραγματική κατανομή f συμπίπτει με την προσαρμοσμένη $g(y;\boldsymbol{\theta}_0)$ τότε οι πίνακες J, K είναι ίσοι και ορίζεται ο πίνακας πληροφορίας κατά **Fisher** του μοντέλου g η ποσότητα:

$$J(\boldsymbol{\theta}_0) = -\int g(y;\boldsymbol{\theta}_0) I(y,\boldsymbol{\theta}_0)\, dy. \tag{14}$$

Στην περίπτωση της παλινδρόμησης, όπου θεωρούμε ως πραγματική κατανομή την $f(y|\boldsymbol{X})$ που γέννησε τα δεδομένα από παρατηρήσεις $(y_i, \boldsymbol{x}_i), i = 1,\ldots,n$, για το προσεγγιστικό μοντέλο $g(y|\boldsymbol{X};\boldsymbol{\theta})$ ορίζουμε τους αντίστοιχους πίνακες:

$$u(y|\boldsymbol{X};\boldsymbol{\theta}) = \frac{\partial \log g(y|\boldsymbol{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}, \qquad I(y|\boldsymbol{X};\boldsymbol{\theta}) = \frac{\partial^2 \log g(y|\boldsymbol{X};\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T},$$

$$J_n = -n^{-1} \sum_{i=1}^{n} \int f(y|\boldsymbol{x}_i) I(y|\boldsymbol{x}_i,\boldsymbol{\theta}_{0,n})\, dy, \quad K_n = n^{-1} \sum_{i=1}^{n} Var_f u(Y_i|\boldsymbol{x}_i,\boldsymbol{\theta}_{0,n}).$$

Με $\boldsymbol{\theta}_{0,n}$ συμβολίζεται το διάνυσμα παραμέτρων ελάχιστης απόστασης $KL(f,g)$ στην περίπτωση της παλινδρόμησης. Εκτιμήσεις των ποσοτήτων K_n, Jn (όπου $\hat{\boldsymbol{\theta}}$ οι εκτιμητές μέγιστης πιθανοφάνειας) δίνονται από τις σχέσεις:

$$\hat{J}_n = -n^{-1} \frac{\partial^2 \boldsymbol{\ell}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} = -n^{-1} \sum_{i=1}^{n} I(y_i|\boldsymbol{x}_i,\hat{\boldsymbol{\theta}}) \tag{15α′}$$

$$\hat{K_n} = n^{-1} \sum_{i=1}^{n} u(y_i|\boldsymbol{x_i},\hat{\boldsymbol{\theta}}) u(y_i|\boldsymbol{x_i},\hat{\boldsymbol{\theta}})^T. \tag{15β′}$$

Οι εκτιμήσεις $\hat{J}_n, \hat{K_n}$ χρησιμοποιούνται παρακάτω στις επεκτάσεις του κριτηρίου AIC στην προσπάθεια να διορθωθεί η μεροληψία που εισάγεται από τους εκτιμητές μέγιστης πιθανοφάνειας.

2.3 AIC κριτήριο πληροφορίας

Είδαμε προηγουμένως ότι η εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\theta}}^*$ συγκλίνει ασυμπτωτικά στην παράμετρο $\boldsymbol{\theta}_0$ η οποία ελαχιστοποιεί την K-L απόσταση του προσεγγιστικού μοντέλου g από το f. Έστω $g(y;\boldsymbol{\theta}^*)$ η σ.π.π. κάθε μεταβλητής του τυχαίου δείγματος και $f(y;\boldsymbol{\theta})$ η πραγματική άγνωστη σ.π.π.. Η αντίστοιχη εκτίμηση της K-L απόστασης σε αυτή την περίπτωση θα γράφεται:

$$\begin{aligned} \hat{KL}(f,g) &= \int f(y;\boldsymbol{\theta}) \log(f(y;\boldsymbol{\theta}))\, dy - \int f(y;\boldsymbol{\theta}) \log\Big(g(y;\hat{\boldsymbol{\theta}}^*)\Big)\, dy \\ &= C - R_n. \end{aligned} \tag{16}$$

Εξετάζοντας διάφορα προς επιλογή προσεγγιστικά μοντέλα g_i διαπιστώνουμε ότι ο πρώτος όρος (C) της $\hat{KL}(f,g_i)$ απόστασης είναι σταθερός για όλα τα μοντέλα και για αυτό επικεντρωνόμαστε στον όρο R_n για τον οποίο έχουμε:

$$Q_n = \mathbb{E}_f[R_n] = \mathbb{E}_f \int f(y) \log g(y;\hat{\boldsymbol{\theta}}^*)\, dy. \tag{17}$$

Η στρατηγική του κριτηρίου AIC είναι να εκτιμηθεί για κάθε υποψήφιο προς επιλογή μοντέλο η ποσότητα Q_n και στη συνέχεια να επιλεχθεί το μοντέλο εκείνο με την υψηλότερη εκτίμηση Q_n, πράγμα που ισοδυναμεί με την επιλογή του μοντέλου που απέχει την ελάχιστη κατά K-L απόσταση από την πραγματικότητα f. Χρησιμοποιώντας την εμπειρική κατανομή του τυχαίου δείγματος δεδομένων $y_i, i = 1,\ldots,n$ μία λογική εκτίμηση του όρου Q_n είναι:

$$\hat{Q_n} = n^{-1} \sum_{i=1}^{n} \log g(y_i;\hat{\boldsymbol{\theta}}^*) = I_n(\hat{\boldsymbol{\theta}}^*). \tag{18}$$

Αποδεικνύεται ότι (Akaike, 1974) ο εκτιμητής $\hat{Q_n}$ υπερεκτιμά την τιμή-στόχο Q_n με $\mathbb{E}[\hat{Q_n} - Q_n] = \frac{p^*}{n}$, και $p^* = Tr(J_n^{-1}K_n)$, όπου p^* ονομάζεται γενικευμένη διάσταση του μοντέλου. Επομένως η μεροληψία του $\hat{Q_n}$ είναι $\frac{p^*}{n}$ και μία διορθωμένη εκτιμήτρια της ποσότητας Q_n θα είναι:

$$\hat{Q_n} - \frac{p^*}{n} = n^{-1}\big\{\boldsymbol{\ell}(\hat{\boldsymbol{\theta}}^*) - p^*\big\}. \tag{19}$$

Κάνοντας την παραδοχή ότι το υποψήφιο μοντέλο, έστω M, με συνάρτηση κατανομής για τις μεταβλητές του την g είναι το πραγματικό, δηλαδή ότι $p^* = p$, και πολλαπλασιάζοντας με $2n$ καταλήγουμε στον γενικό τύπο του κριτηρίου AIC:

$$AIC(M) = 2\boldsymbol{\ell}(\hat{\boldsymbol{\theta}}^*) - 2p. \tag{20}$$

Από τη σχέση (20) καταλήγουμε στη μορφή $AIC(M) = -2\boldsymbol{\ell}(\hat{\boldsymbol{\theta}}^*) + 2p$ η οποία επικρατεί στη βιβλιογραφία. Στις εφαρμογές διαθέτουμε ένα σύνολο μοντέλων και για καθένα από αυτά υπολογίζουμε την τιμή του κριτηρίου AIC. Ύστερα, από τις τιμές που προκύπτουν έχουμε μια εικόνα του πόσο απέχει το κάθε μοντέλο από την άγνωστη πραγματικότητα σε σχέση με τα υπόλοιπα μοντέλα. Μοντέλα με χαμηλή τιμή της συνάρτησης-*score* AIC θεωρούνται ότι προσαρμόζονται καλύτερα στα δεδομένα μας και για αυτό προτιμούνται.

Παρατηρήσεις:

- Εφαρμόζοντας το κριτήριο AIC σε ένα σύνολο μοντέλων δε μας ενδιαφέρει η τιμή αυτή καθαυτή που θα λάβει το κριτήριο σε κάθε προσαρμοσμένο μοντέλο. Αντίθετα, εστιάζουμε στις διαφορές των τιμών AIC ανάμεσα σε όλα τα μοντέλα, διότι αυτές μας δίνουν ένα ουσιαστικό μέτρο σύγκρισή τους. Αυτό ισχύει και στα υπόλοιπα κριτήρια πληροφορίας με τα οποία θα ασχοληθούμε.

- Λειτουργώντας κανείς με το κριτήριο AIC πρέπει πάντα να προσέχει ότι τα υποψήφια μοντέλα έχουν προσαρμοστεί στο ίδιο σύνολο δεδομένων, διαφορετικά τα αποτελέσματα πολύ πιθανό να είναι παραπλανητικά.

- Προσθέτοντας περισσότερες μεταβλητές στα μοντέλα παρατηρείται μια τάση μείωσης του όρου $-2\boldsymbol{\ell}(\hat{\boldsymbol{\theta}}^*)$ μέχρι όμως ένα σημείο πέρα από το οποίο υπερισχύει ο όρος $2p$ του κριτηρίου AIC. Βλέπουμε δηλαδή ότι το κριτήριο τιμωρεί εξίσου τα υπερπροσαρμοσμένα μοντέλα αλλά και τα μεροληπτικά με «φτωχή» προσαρμογή.

- Αρκετές φορές το μέγεθος του δείγματος n είναι πολύ μεγαλύτερο του αριθμού των προς εκτίμηση παραμέτρων p με αποτέλεσμα ο πρώτος όρος του κριτηρίου να υπερισχύει του 2ου και να ευνοούνται περισσότερο τα πιο σύνθετα μοντέλα. Θα δούμε παρακάτω μία διορθωμένη εκδοχή του κριτηρίου που λειτουργεί αποτελεσματικά σε αυτές τις περιπτώσεις.

2.3.1 Εφαρμογή στο γραμμικό μοντέλο

Θεωρούμε το πολλαπλό γραμμικό μοντέλο με p επεξηγηματικές μεταβλητές $X_1, \ldots, X_p$ και μεταβλητή απόκρισης Y, με γενική μορφή:

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2).$$

Το μοντέλο διαθέτοντας δείγμα παρατηρήσεων $(y_i, \boldsymbol{x}_i), i = 1, \ldots, n$ υπό την μορφή πινάκων γράφεται: $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$. Η δεσμευμένη συνάρτηση πυκνότητας πιθανότητας της παρατηρούμενης y_i τιμής είναι:

$$f(y_i|\boldsymbol{x}_i) = \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}}.$$

Συμβολίζουμε με $\boldsymbol{x}_i^T = (1, x_{i1}, x_{i2}, \ldots, x_{ip})$ την i-οστή γραμμή του πίνακα σχεδιασμού $\tilde{\boldsymbol{X}}$. Η συνάρτηση πιθανοφάνειας του γραμμικού μοντέλου είναι:

$$\begin{aligned} L(\boldsymbol{\beta}, \sigma^2) &= \prod_{i=1}^{n} \frac{1}{(2\pi\sigma^2)^{1/2}} e^{-\frac{(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} e^{-\frac{\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2}{2\sigma^2}}. \end{aligned}$$

Λογαριθμοποιώντας τη συνάρτηση πιθανοφάνειας παίρνουμε:

$$\boldsymbol{\ell} = \log L(\boldsymbol{\beta}, \sigma^2) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\big(\sigma^2\big) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\boldsymbol{\beta})^2.$$

Η τελευταία σχέση υπό μορφή πινάκων γράφεται:

$$\begin{aligned} \boldsymbol{\ell} &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\big(\sigma^2\big) - \frac{1}{2\sigma^2}(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta})^T(\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\big(\sigma^2\big) - \frac{1}{2\sigma^2}(\boldsymbol{y}^T\boldsymbol{y} - \boldsymbol{y}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta} - \boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\boldsymbol{y} + \boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta}). \end{aligned}$$

Παραγωγίζοντας τη λογαριθμοποιημένη πιθανοφάνεια ως προς το διάνυσμα των παραμέτρων $\boldsymbol{\beta}$ προκύπτει:

$$\frac{\partial\boldsymbol{\ell}}{\partial\boldsymbol{\beta}} = \frac{-2\tilde{\boldsymbol{X}}^T\boldsymbol{y} + 2\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta}}{2\sigma^2}. \tag{21}$$

Στη σχέση (21) χρησιμοποιήθηκαν οι εξισώσεις:

$$\boldsymbol{y}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta} = (\boldsymbol{y}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta})^T = \boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\boldsymbol{y},$$

$$\frac{\partial(\boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\boldsymbol{y})}{\partial\boldsymbol{\beta}} = \tilde{\boldsymbol{X}}^T\boldsymbol{y} \quad \text{και} \quad \frac{\partial(\boldsymbol{\beta}^T\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta})}{\partial\boldsymbol{\beta}} = 2\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\boldsymbol{\beta}.$$

Εξισώνοντας τη σχέση (21) με το μηδέν παίρνουμε:

$$0 = -2\tilde{\boldsymbol{X}}^T\boldsymbol{y} + 2\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} \;\leftrightarrow\; \hat{\boldsymbol{\beta}}_{ML} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}. \tag{22}$$

Παραγωγίζοντας τώρα την $\boldsymbol{\ell}$ ως προς σ^2 και εξισώνοντας μετά με το μηδέν έχουμε:

$$\begin{aligned} \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\sigma}^2} &= -\frac{n}{2}\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}[\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}]^T[\boldsymbol{y} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}] = 0 \quad \leftrightarrow \\ \hat{\sigma}^2_{ML} &= \frac{1}{n}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{ML})^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \frac{1}{n}\sum_{i=1}^{n} e_i^2. \end{aligned} \tag{23}$$

Παρατηρούμε ότι η εκτιμήτρια $\hat{\boldsymbol{\beta}}_{ML}$ που προκύπτει από τη μέθοδο μέγιστης πιθανοφάνειας είναι ίδια με την εκτιμήτρια της μεθόδους ελαχίστων τετραγώνων, με την ίδια πάλι προϋπόθεση αντιστρεψιμότητας του πίνακα $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$. Ωστόσο, η εκτιμήτρια $\hat{\sigma}^2_{ML}$ είναι μεροληπτική εκτιμήτρια της διασποράς σ^2 με:

$$\begin{aligned} \mathbb{E}[\hat{\sigma}^2_{ML}] &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[e_i^2] = \frac{1}{n}\sum_{i=1}^{n}\sigma^2(1 - h_{ii}) \\ &= \frac{\sigma^2}{n}[n - (p+1)] \neq \sigma^2. \end{aligned}$$

Συνεπώς η μεγιστοποιημένη συνάρτηση της λογαριθμοποιημένης πιθανοφάνειας αξιοποιώντας τη σχέση (23) θα γράφεται ως:

$$\begin{aligned} \boldsymbol{\ell}_{max} &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\big(\hat{\sigma}^2_{ML}\big) - \frac{1}{2\hat{\sigma}^2_{ML}}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{ML}) \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\big(\hat{\sigma}^2_{ML}\big) - \frac{n}{2\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{ML})^2}\sum_{i=1}^{n}(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}_{ML})^2 \\ &= -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\big(\hat{\sigma}^2_{ML}\big) - \frac{n}{2}. \end{aligned}$$

Σημείωση: Προκειμένου να εξασφαλιστεί σημείο μεγίστου πρέπει να υπολογιστεί ο Εσσιανός πίνακας (Hessian Matrix) $\frac{\partial^2 \boldsymbol{\ell}}{\partial \boldsymbol{\theta}\partial \boldsymbol{\theta}^T}\Big|_{pxp}$ (για διάνυσμα παραμέτρων $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_p)^T$) στο σημείο ακροτάτου της $\boldsymbol{\ell}$ και να είναι γνήσια αρνητικός, πράγμα που ισοδυναμεί με αρνητικές ιδιοτιμές.

Το κριτήριο AIC στο πολλαπλό γραμμικό μοντέλο λαμβάνοντας υπόψιν ότι εκτιμήσαμε $p+2$ το πλήθος παραμέτρους ($p+1$ συντελεστές παλινδρόμησης και η διασπορά σ^2) θα παίρνει τη μορφή:

$$\begin{aligned} AIC(M) &= -2\boldsymbol{\ell}_{max} + 2(p+2) \\ &= n\log(2\pi) + n\log\big(\hat{\sigma}^2_{ML}\big) + n + 2(p+2). \end{aligned} \tag{24}$$

Αγνοώντας τις σταθερές $n\log(2\pi)$ και n, από τα υποψήφια γραμμικά μοντέλα διαλέγουμε εκείνο (ή εκείνα) με χαμηλότερη τιμή κριτηρίου AIC, η οποία παρατηρούμε ότι στη συγκεκριμένη περίπτωση επηρεάζεται αποκλειστικά από τους όρους $n\log(\hat{\sigma}^2_{ML})$ και $2(p+2)$.

2.4 AIC_c κριτήριο πληροφορίας

Εξετάζοντας με προσοχή τη γενική έκφραση του κριτηρίου AIC διαπιστώνουμε ότι όσο αυξάνεται το μέγεθος n του δείγματος παρατηρήσεων που διαθέτουμε, το κριτήριο επιλέγει όλο και πιο σύνθετα μοντέλα (**overfitted** models) προφανώς υπό την προϋπόθεση ότι αυτά τα μοντέλα αυξάνουν την λογαριθμική πιθανοφάνεια, πράγμα που δείχνει ότι προσαρμόζονται καλύτερα στα πραγματικά δεδομένα. Το φαινόμενο αυτό συμβαίνει διότι η λογαριθμοποιημένη πιθανοφάνεια αυξάνεται γραμμικά με το μέγεθος n του δείγματος, ενώ ο όρος ποινικοποίησης $2p$ είναι ανάλογος του αριθμού των παραμέτρων των μοντέλων και συνήθως ισχύει ότι $n > p$.

Μελετώντας την περίπτωση των γραμμικών μοντέλων παλινδρόμησης θα δούμε πως μπορεί να προκύψει μια διορθωμένη εκδοχή του κριτηρίου AIC η οποία μεριμνά εξίσου για την προσαρμοστικότητα (goodness of fit) και την φειδωλότητα (parsimony) των μοντέλων, χωρίς να επηρεάζεται από την αδυναμία του AIC ως προς το μέγεθος του δείγματος n.

Στην Παράγραφο 2.3.1 θεωρήσαμε το γραμμικό μοντέλο $\boldsymbol{Y} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ για το οποίο η έκφραση του κριτηρίου AIC είναι:

$$AIC = n\log(2\pi) + n\log\left(\hat{\sigma^2}_{ML}\right) + n + 2(p+2).$$

Η διασπορά σ^2 εκτιμάται από την μέθοδο μέγιστης πιθανοφάνειας με $\hat{\sigma^2}_{ML} = \frac{\sum_{i=1}^n e_i^2}{n} = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$. Το κριτήριο AIC συμβουλεύει από τα υποψήφια μοντέλα να επιλεχθεί εκείνο που ελαχιστοποιεί την ποσότητα $n\log\left(\hat{\sigma^2}_{ML}\right) + 2p$ καθώς οι υπόλοιποι όροι στα υποψήφια γραμμικά μοντέλα παραμένουν σταθεροί. Στόχος του κριτηρίου είναι να εκτιμήσει την αναμενόμενη $K - L$ απώλεια πληροφορίας του προσεγγιστικού μοντέλου $g(y|\boldsymbol{X};\boldsymbol{\theta}), \boldsymbol{\theta} = (\boldsymbol{\beta}, \sigma^2)$ σε σχέση με το μοντέλο $f(y|\boldsymbol{X})$ που παρήγαγε τα πραγματικά δεδομένα. Στην περίπτωση των γραμμικών μοντέλων που εξετάζουμε θεωρούμε χωρίς βλάβη της γενικότητας ότι $\boldsymbol{\theta} = (\beta_1, \beta_2, \ldots, \beta_p, \sigma^2)$ και επομένως:

$$AIC = n\log(2\pi) + n\log\left(\hat{\sigma^2}_{ML}\right) + n + 2(p+1).$$

Με την υπόθεση ότι το υποψήφιο μοντέλο ταυτίζεται με το πραγματικό και θεωρώντας ότι το πραγματικό μοντέλο $f(y|\boldsymbol{X})$ έχει αναμενόμενη δεσμευμένη

μέση τιμή $\xi(\boldsymbol{X})$ και τυπική απόκλιση σ_0, έχουμε ότι $\xi_i = \boldsymbol{x}_i^T\boldsymbol{\beta}$ όπου $\boldsymbol{x}_i^T$ η i-οστή γραμμή του πίνακα σχεδιασμού $\tilde{\boldsymbol{X}}$ του γραμμικού μοντέλου $g(Y|\boldsymbol{X})$.

Επιπλέον, από το γεγονός ότι $\frac{\sum^n e_i^2}{\sigma^2} = \frac{SSE}{\sigma^2} \sim X^2_{n-p}$, είναι ορθό αντί της εκτιμήτριας $\hat{\sigma^2}_{ML}$ που προκύπτει από τη μέθοδο μέγιστης πιθανοφάνειας να χρησιμοποιήσουμε την αμερόληπτη $\hat{\sigma^2}_{u.e.} = \frac{SSE}{n-p} = \frac{1}{n-p}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2$.

Προκύπτουν λοιπόν δύο εκτιμήτριες για την παράμετρο διασποράς σ^2 του γραμμικού μοντέλου, οι οποίες στη γενική μορφή γράφονται:

$$\hat{\sigma^2} = \frac{1}{n-a}SSE = \frac{1}{n-a}\sum_{i=1}^n (y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2.$$

- με $a = 0$ να αντιστοιχεί στην εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\sigma^2}_{ML}$,
- και $a = p$ στην αμερόληπτη εκτιμήτρια $\hat{\sigma^2}_{u.e.}$.

Ακολουθώντας την ίδια συλλογιστική πορεία με αυτή της προέλευσης του κριτηρίου AIC έχουμε:

$$\begin{aligned}
\hat{Q}_n &= n^{-1}\sum_{i=1}^n \log\Big(g(y_i|\boldsymbol{x}_i;\hat{\boldsymbol{\beta}},\hat{\sigma^2})\Big) \\
&= -n^{-1}\sum_{i=1}^n \Big\{ -\frac{1}{2}\log(2\pi) - \log(\hat{\sigma}) - \frac{1}{2}\frac{(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2}{\hat{\sigma^2}}\Big\} \qquad (25)\\
&= -\log(\hat{\sigma}) - \frac{1}{2}\log(2\pi) - \frac{1}{2}\frac{n-a}{n}.
\end{aligned}$$

Σκοπός μας είναι να βρούμε κατά πόσο η εκτιμήτρια $\hat{Q}_n$ υπερεκτιμά την ποσότητα:

$$\begin{aligned}
R_n &= n^{-1}\sum_{i=1}^n \int f(y|\boldsymbol{x}_i)\log\Big(g(y|\boldsymbol{x}_i;\hat{\boldsymbol{\beta}},\hat{\sigma^2})\Big)\,dy \\
&= -\log\Big(\hat{\sigma^2}\Big) - \frac{1}{2}\log(2\pi) - n^{-1}\sum_{i=1}^n \frac{1}{2\hat{\sigma^2}}\int f(y|\boldsymbol{x}_i)(y - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2\,dy.
\end{aligned}$$

Η προέλευση της έκφρασης R_n είναι σύνθετη και ξεφεύγει των στόχων της εργασίας. Για περισσότερες τεχνικές πληροφορίες παραπέμπουμε στους Claeskens και Hjort (2008). Χρησιμοποιούμε τώρα την υπόθεση ότι τα δεδομένα προέρχονται από το πραγματικό μοντέλο $f(y|\boldsymbol{X})$ με μέση τιμή $\xi(\boldsymbol{X})$ και

διασπορά σ_0^2 καθώς και τη σχέση:

$$\begin{aligned}\mathbb{E}_f[(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2] &= Var_f(y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}) + (\mathbb{E}_f[y_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}}])^2 \\ &= \sigma_0^2 + (\xi_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2.\end{aligned}$$

Στη συνέχεια μετασχηματίζουμε την έκφραση της R_n στην:

$$R_n = -log(\hat{\sigma}) - \frac{1}{2}\log(2\pi) - \frac{1}{2}n^{-1}\sum_{i=1}^{n}\frac{(\xi_i - \boldsymbol{x}_i^T\hat{\boldsymbol{\beta}})^2 + \sigma_0^2}{\hat{\sigma^2}}. \tag{26}$$

Συνδυάζοντας τις σχέσεις (25) και (26), για την μεροληψία του $\hat{Q}_n$ πλέον έχουμε:

$$\mathbb{E}_f[\hat{Q}_n - R_n] = -\frac{1}{2}\frac{n-a}{n} + \frac{1}{2}\mathbb{E}_f\Big[\frac{\sigma_0^2}{\hat{\sigma^2}}\big\{n^{-1}\frac{\sum_{i=1}^{n}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \xi_i)^2}{\sigma_0^2} + 1\big\}\Big]. \tag{27}$$

Από την υπόθεση του πραγματικού μοντέλου f για τα δεδομένα αντικαθιστούμε με $\xi_i = \boldsymbol{x}_i^T\boldsymbol{\beta}, \quad \sigma_0 = \sigma$ και χρησιμοποιώντας γνωστό αποτέλεσμα του αθροίσματος τετραγώνων κανονικών μεταβλητών ισχύει ότι $\frac{\hat{\sigma^2}}{\sigma^2} \sim \frac{X^2_{n-p}}{n-a}$ (ανεξάρτητη τυχαία μεταβλητή των συντελεστών $\hat{\boldsymbol{\beta}}$).

Επιπλέον, για το προσαρμοσμένο μοντέλο έχουμε:

$$\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T(\tilde{\boldsymbol{X}}\boldsymbol{\beta} + \varepsilon) = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{H}\varepsilon.$$

Επομένως ο όρος: $n^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \xi_i)^2 = n^{-1}\sum_{i=1}^{n}(\boldsymbol{x}_i^T\hat{\boldsymbol{\beta}} - \boldsymbol{x}_i^T\boldsymbol{\beta})^2 = n^{-1}\Big\|\tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}} - \tilde{\boldsymbol{X}}\boldsymbol{\beta}\Big\|^2 = n^{-1}\boldsymbol{\varepsilon}^T\boldsymbol{H}\boldsymbol{\varepsilon}$, όπου $\|.\|$ η Ευκλείδεια νόρμα. Αποδεικνύεται ότι $\mathbb{E}_f[n^{-1}\boldsymbol{\varepsilon}^T\boldsymbol{H}\boldsymbol{\varepsilon}] = n^{-1}\mathbb{E}_f[tr(\boldsymbol{H}\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T)] = \frac{\sigma^2}{n}tr(\boldsymbol{H}) = p\frac{\sigma^2}{n}$. Επιπλέον έχουμε:

$$\mathbb{E}[\frac{1}{X^2_{n-p}}] = \frac{1}{n-p-2}, n > p+2 \quad \text{και ισοδύναμα } \mathbb{E}\big[\frac{n-a}{X^2_{n-p}}\big] = \frac{n-a}{n-p-2}.$$

Γυρνάμε στη σχέση (27) και πλέον έχουμε:

$$\begin{aligned}\mathbb{E}_f[\hat{Q}_n - R_n] &= -\frac{1}{2}\frac{n-a}{n} + \frac{1}{2}\frac{n-a}{n-p-2}\frac{p+n}{n} \\ &= \frac{1}{2}\frac{n-a}{n}\frac{2p+2}{n-p-2} = \frac{p+1}{n}\frac{n-a}{n-p-2}.\end{aligned} \tag{28}$$

Είμαστε πλέον σε θέση να προχωρήσουμε σε τροποποιήσεις της μορφής

του κριτηρίου AIC που ενσωματώνουν πιο ακριβείς όρους ποινικοποίησης. Η πρώτη τροποποίηση του κριτηρίου προτάθηκε από τους Sugiura(1978), Hurvich και Tsai (1989) και διατηρεί στην εξίσωση του AIC την εκτιμήτρια μέγιστης πιθανοφάνειας $\hat{\sigma^2}_{ML}$, αλλά χρησιμοποιεί τον πιο ακριβή όρο ποινικοποίησης εμπλέκοντας τον όρο της σχέσης (28) για $a = 0$ από όπου καταλήγουμε στη γενική σχέση (πολλαπλασιάζοντας πάλι με $2n$):

$$AIC_c^1 = -2\boldsymbol{\ell}(\hat{\boldsymbol{\beta}}, \hat{\sigma^2}_{ML}) + 2(p+1)\frac{n}{n-p-2}.$$

Η δεύτερη προσέγγιση είναι να χρησιμοποιηθεί η διορθωμένη εκτιμήτρια $\hat{\sigma^{2*}} = \frac{SSE}{n-p-2}$ για $a = p+2$ ούτως ώστε να διατηρηθεί η μορφή του AIC και να οδηγηθούμε στη σχέση:

$$AIC_c^2 = -2\boldsymbol{\ell}(\boldsymbol{\beta}, \hat{\sigma^{2*}}) + 2(p+1).$$

Από τους διορθωμένους συντελεστές AIC_c^1,AIC_c^2 προτιμάται η χρήση του AIC_c^1 καθώς γενικεύεται στην περίπτωση των παραμετρικών μοντέλων παλινδρόμησης και έχει πιο άμεσο υπολογισμό, με τη μορφή του κριτηρίου να γράφεται ως:

$$AIC_c = -2\boldsymbol{\ell}(\hat{\boldsymbol{\theta}}) + 2 \cdot dim(\boldsymbol{\theta})\frac{n}{n - dim(\boldsymbol{\theta}) - 1}, \tag{29}$$

όπου συμβολίζουμε με $\boldsymbol{\theta}$ το διάνυσμα παραμέτρων του μοντέλου και με $dim(\boldsymbol{\theta})$ τη διάσταση του παραμετρικού χώρου.

Προσθαφαιρώντας μάλιστα τον όρο $2 \cdot dim(\boldsymbol{\theta})$ στη σχέση (29) προκύπτει η σύνδεση του κριτηρίου AIC_c με το κριτήριο AIC:

$$\begin{aligned} AIC_c &= -2\boldsymbol{\ell}(\hat{\boldsymbol{\theta}}) + 2 \cdot dim(\boldsymbol{\theta}) - 2 \cdot dim(\boldsymbol{\theta}) + 2 \cdot dim(\boldsymbol{\theta})\frac{n}{n - dim(\boldsymbol{\theta}) - 1} \\ &= AIC + 2 \cdot dim(\boldsymbol{\theta})\left[\frac{dim(\boldsymbol{\theta}) + 1}{n - dim(\boldsymbol{\theta}) - 1}\right]. \end{aligned} \tag{30}$$

Λαμβάνοντας υπόψιν λοιπόν τη μεροληψία που εισάγεται στο κριτήριο πληροφορίας AIC από την εκτιμήτρια $\hat{\sigma^2}_{ML}$ της μεθόδου μέγιστης πιθανοφάνειας, καταλήγουμε σε μία πιο ακριβή διορθωμένη (corrected) εκδοχή του, ονομαστικά την AIC_c, η οποία βασίζεται στη θεωρία των γραμμικών μοντέλων και μπορεί να γενικευθεί ακολουθώντας παρόμοιες συλλογιστικές πορείες και σε υπόλοιπα παραμετρικά μοντέλα παλινδρόμησης.

2.5 BIC κριτήριο πληροφορίας

Στην προηγούμενη ενότητα αναλύθηκε η προέλευση του κριτηρίου AIC και της επέκτασής-διόρθωσής του AIC_c μαζί με κάποιες εφαρμογές. Εισάγουμε σε αυτό το σημείο το Μπεϋζιανό κριτήριο πληροφορίας BIC το οποίο προέρχεται από τη θεωρία της Μπεϋζιανής Στατιστικής και ποινικοποιεί σε μεγαλύτερο βαθμό τα πολύπλοκα μοντέλα οδηγώντας σε φειδωλότερα μοντέλα.

Το κριτήριο BIC αναπτύχθηκε από τους Schwartz(1978) και Akaike(1977, 1978) ως προσπάθεια βελτίωσης του κριτηρίου AIC. Η μορφή του συνθέτεται από τη συνάρτηση μεγιστοποιημένης πιθανοφάνειας με όρο ποινικοποίησης το λογάριθμο του μεγέθους του δείγματος παρατηρήσεων που προσαρμόστηκε το μοντέλο, πολλαπλασιασμένο με τον αριθμό των εκτιμημένων παραμέτρων. Η ακριβής γραφή για κάθε υποψήφιο μοντέλο M είναι:

$$BIC(M) = -2\boldsymbol{\ell}_{max}(M) + \log(n) \cdot dim(M). \tag{31}$$

Με $\boldsymbol{\ell}_{max}(M)$ συμβολίζεται η τιμή της μεγιστοποιημένης πιθανοφάνειας όπως προκύπτει από τους ε.μ.π. του μοντέλου M, $dim(M)$ είναι ο αριθμός των παραμέτρων που εκτιμήθηκαν για το μοντέλο M και n ο αριθμός του μεγέθους δείγματος των δεδομένων που χρησιμοποιήθηκαν. Το μοντέλο με τη μικρότερη τιμή $BIC(M)$ προτιμάται ως βέλτιστο. Το κριτήριο προέρχεται από τη θεωρία της Μπεϋζιανής Στατιστικής η οποία δεν αποτελεί κομμάτι της εν λόγω εργασίας. Ωστόσο, η μορφή του είναι παρόμοια με αυτή του κριτηρίου AIC με τη βασική διαφορά όμως στον αυστηρότερο όρο ποινικοποίησης $\log(n) \cdot dim(M)$ της πολυπλοκότητας των μοντέλων και με την προϋπόθεση ότι το μέγεθος του δείγματος είναι $n \geq 8$.

2.6 Εφαρμογή κριτηρίων σε δεδομένα διάρκειας ζωής

Διαθέτουμε δεδομένα διάρκειας ζωής από 25 ρουλεμάν (Lieblein & Zelen 1956). Θα προσαρμόσουμε δύο μοντέλα στα δεδομένα με σκοπό την προσέγγιση της κατανομής των δεδομένων και θα συγκρίνουμε την απόδοση τους με βάση τα κριτήρια BIC και AIC.

Οι παρατηρήσεις με αστερίσκο($*$) στον Πίνακα 2.1 είναι από ρουλεμάν που ήταν ακόμη σε λειτουργία όταν πραγματοποιήθηκε η συλλογή των δεδομένων. Αναφερόμαστε σε αυτά με τον όρο δεξιά αποκομμένα δεδομένα και θεωρούμε ότι η αποκοπή είναι μη πληροφοριακή [1]. Έστω Y τυχαία μεταβλητή που εκφράζει

[1] Το πείραμα για τη συλλογή των δεδομένων διεξήχθη για συγκεκριμένο χρονικό διάστημα και είναι σύνηθες αρκετές μονάδες να συνεχίσουν να λειτουργούν μετά τον τερματισμό του. Αν και δε γνωρίζουμε τη διάρκεια ζωής μιας τέτοιας μονάδας, διαθέτουμε την πληροφορία ότι έχει ξεπεράσει τη χρονική διάρκεια κατά την οποία η μονάδα ήταν στο πείραμα. Με τον

17.88	28.92	33.00	41.52	42.12	45.60	48.48	51.84	51.96
54.12	55.56	67.80	67.80*	67.80*	68.64	68.64*	68.88*	84.12
93.12	98.64	105.12	105.84*	127.92	128.04	173.40*		

Πίνακας 2.1: Δεδομένα διάρκειας ζωής από 25 ρουλεμάν. Η μονάδα μέτρησης της διάρκειας ζωής είναι ο αριθμός των περιστροφών (σε εκατομμύρια) μέχρις ότου υποστούν βλάβη.

τη διάρκεια ζωής των δεδομένων με αντίστοιχες παρατηρούμενες τιμές $y_i, i = 1, \ldots, 25$. Οι τιμές της τ.μ. Y χωρίζονται σε δύο υποσύνολα:

$$U = \{\text{μη αποκομμένες παρατηρήσεις (uncensored data)}\}$$
$$C = \{\text{αποκομμένες παρατηρήσεις (censored data)}\}.$$

Θεωρώντας ότι τα δεδομένα της τ.μ. Y προέρχονται από κάποια συνεχή κατανομή με σ.π.π. $f(y)$ η συνάρτηση πιθανοφάνειας του μοντέλου f θα έχει τη μορφή:

$$L = \prod_{i \in U} f(y_i) \prod_{i \in C} S(y_i). \tag{32}$$

Με $S(y_i) = P[Y > y_i]$ συμβολίζεται η συνάρτηση επιβίωσης και εκφράζει την πιθανότητα η διάρκεια ζωής Y ενός μέλους του πληθυσμού να υπερβαίνει το χρόνο y_i. Για περισσότερες τεχνικές πληροφορίες παραπέμπουμε στην Καρώνη (2009).

Εκθετικό μοντέλο

Το πρώτο μοντέλο διάρκειας ζωής που προσαρμόζουμε είναι το εκθετικό. Υποθέτουμε λοιπόν ότι η τ.μ. Y ακολουθεί την εκθετική κατανομή με συνάρτηση πυκνότητας πιθανότητας:

$$f(y) = \lambda e^{-\lambda y}, \quad y > 0, \quad \lambda > 0 \quad (\text{παράμετρος ρυθμού}).$$

Υπολογίζουμε στη συνέχεια τη συνάρτηση πιθανοφάνειας για το εκθετικό μοντέλο με βάση τη σχέση (32). Η συνάρτηση επιβίωσης θα ισούται με:

όρο μη πληροφοριακή αποκοπή (uninformative censoring) εννοούμε ότι η αποκοπή πρέπει να είναι τυχαία, δηλαδή να μην σχετίζεται με τη μετέπειτα διάρκεια ζωής της μονάδας (Καρώνη, 2009).

$$
\begin{aligned}
&S(y) = P(Y > y) = 1 - P(Y \leq y) = 1 - F(y),\\
&F(y) = \int_0^y \lambda e^{-\lambda z}\, dz = 1 - e^{-\lambda y} \quad \text{και επομένως}\\
&S(y) = e^{-\lambda y}.
\end{aligned}
$$

Η πιθανοφάνεια του εκθετικού μοντέλου για το σύνολο των αποκομμένων και μη δεδομένων θα είναι:

$$L = \prod_{i \in U} (\lambda e^{-\lambda y_i}) \prod_{i \in C} e^{-\lambda y_i}.$$

Λογαριθμίζοντας την L έχουμε:

$$
\begin{aligned}
\boldsymbol{\ell} = \log L &= \sum_{i \in U} (\log(\lambda) - \lambda y_i) + \sum_{i \in C} (-\lambda y_i)\\
&= \sum_{i \in U} \log(\lambda) + \sum_{i \in U} (-\lambda y_i) + \sum_{i \in C} (-\lambda y_i)\\
&= \kappa \log(\lambda) - \lambda \sum_{i=1}^{n} y_i.
\end{aligned}
$$

Με $\kappa = 19$ συμβολίζεται το πλήθος των τιμών χωρίς αποκοπή και $n = 25$ είναι το μεγεθος του δείγματος των δεδομένων που συλλέχθηκαν συνολικά.

Παραγωγίζοντας ως προς λ τη λογαριθμοποιημένη συνάρτηση πιθανοφάνειας και εξισώνοντας με το μηδέν έχουμε:

$$
\begin{aligned}
&\frac{\partial \boldsymbol{\ell}}{\partial \lambda} = \frac{\kappa}{\lambda} - \sum_{i=1}^{n} y_i = 0 \leftrightarrow \hat{\lambda} = \frac{\kappa}{\sum_{i=1}^{n} y_i}\\
&\text{και } \frac{\partial^2 \boldsymbol{\ell}}{\partial \lambda^2} = -\frac{\kappa}{\lambda^2} < 0.
\end{aligned}
$$

Καταλήγουμε λοιπόν στην εκτιμήτρια μέγιστης πιθανοφάνειας της παραμέτρου λ του εκθετικού μοντέλου η οποία ισούται με $\hat{\lambda} = \frac{19}{1796{,}76} = 0.010574$.

Η τιμή του κριτηρίου BIC για το εκθετικό μοντέλο υπό το συγκεκριμένο δείγμα παρατηρούμενων τιμών είναι:

$$
\begin{aligned}
BIC(exp) &= -2\boldsymbol{\ell}_{max}(exp) + \log(n) \cdot dim(exp)\\
&= -2\Big[\kappa \log\left(\frac{\kappa}{\sum_{i=1}^{n} y_i}\right) - \frac{\kappa}{\sum_{i=1}^{n} y_i}\Big] + \log(n) \cdot 1
\end{aligned}
$$

$$= -2\big[\kappa \log(\kappa) - \kappa \log\left(\sum_{i=1}^{n} y_i\right) - \kappa\big] + \log(n).$$

Για τα δεδομένα που διαθέτουμε προκύπτει ότι $\sum_{i=1}^{n} y_i = 1796,76$ και επομένως η τιμή του Μπεύζιανού κριτηρίου για το εκθετικό μοντέλο είναι:

$$\begin{aligned} BIC(exp) &= -2\big[19 \cdot \log(19) - 19 \cdot \log(1796,76) - 19\big] + \log(25) \\ &= 214.09. \end{aligned}$$

Μοντέλο Weibull

Η κατανομή Weibull είναι η πιο διαδεδομένη κατανομή που χρησιμοποιείται στην ανάλυση αξιοπιστίας και επιβίωσης με εφαρμογές σε βιοϊατρικές επιστήμες μέχρι και στο κλάδο της οικονομετρίας. Υποθέτουμε ότι η τ.μ. Y των δεδομένων διάρκειας ζωής ακολουθεί την Weibull κατανομή με συνάρτηση πυκνότητας πιθανότητας που δίνεται από τη σχέση:

$$f(y) = \eta \alpha^{-\eta} y^{\eta-1} e^{-(\frac{y}{\alpha})^{\eta}},$$

όπου $\alpha > 0$ είναι η παράμετρος κλίμακας και $\eta > 0$ η παράμετρος σχήματος. Η συνάρτηση αξιοπιστίας - επιβίωσης για την κατανομή Weibull είναι:

$$\begin{aligned} S(y) = P(Y > y) &= \int_{y}^{\infty} \eta \alpha^{\eta} t^{\eta-1} e^{-(\frac{t}{\alpha})^{\eta}}\, dt \\ &= \int_{y}^{\infty} \eta \big(\frac{t}{\alpha}\big)^{\eta-1} \alpha^{-1} e^{-(\frac{t}{\alpha})^{\eta}}\, dt. \end{aligned}$$

Θέτουμε:

$$u = (\frac{t}{\alpha})^{\eta}, \quad \frac{\partial u}{\partial t} = \frac{\eta}{\alpha^{\eta}} t^{\eta-1} = \eta \big(\frac{t}{\alpha}\big)^{\eta-1} \alpha^{-1} \quad \text{και έχουμε:}$$

$$S(y) = \int_{(y/\alpha)^{\eta}}^{\infty} e^{-u}\, du = -e^{-u}\big|_{(y/\alpha)^{\eta}}^{\infty} = e^{-(\frac{y}{\alpha})^{\eta}}.$$

Η πιθανοφάνεια του συνόλου αποκομμένων και μη δεδομένων στο Weibull μοντέλο με βάση τα προηγούμενα και τη σχέση (32) είναι:

$$L = \prod_{i \in U} (\eta \alpha^{-\eta} y_i^{\eta-1} e^{-(\frac{y_i}{\alpha})^\eta}) \prod_{i \in C} e^{-(\frac{y_i}{\alpha})^\eta}.$$

Λογαριθμίζοντας έχουμε:

$$\begin{aligned}\boldsymbol{\ell} &= \sum_{i \in U} \log(\eta) - \eta \sum_{i \in U} \log(\alpha) + (\eta - 1) \sum_{i \in U} \log(y_i) - \sum_{i \in U} (\frac{y_i}{\alpha})^\eta - \sum_{i \in C} (\frac{y_i}{\alpha})^\eta \\ &= \kappa \log(\eta) - \kappa\eta \log(\alpha) + (\eta - 1) \sum_{i \in U} \log(y_i) - \sum_{i=1}^{n} (\frac{y_i}{\alpha})^\eta.\end{aligned}$$

Συμβολίζουμε με κ όπως και στο εκθετικό μοντέλο το πλήθος των μη αποκομμένων δεδομένων. Παραγωγίζοντας και εξισώνοντας με το μηδέν τις μερικές παραγώγους των παραμέτρων της $\boldsymbol{\ell}$ έχουμε:

- $$\begin{aligned}\frac{\partial \boldsymbol{\ell}}{\partial \alpha} &= -\frac{\kappa\eta}{\alpha} + \eta \alpha^{-\eta-1} \sum_{i=1}^{n} (y_i)^\eta = 0 \quad \leftrightarrow \\ \frac{\kappa\eta}{\alpha} &= \eta \sum_{i=1}^{n} (y_i)^\eta \frac{1}{\alpha^{\eta+1}} \\ \alpha^\eta &= \frac{\sum_{i=1}^{n} (y_i)^\eta}{\kappa} \quad \leftrightarrow \quad \hat{\alpha} = \Big(\frac{\sum_{i=1}^{n} (y_i)^{\hat{\eta}}}{\kappa} \Big)^{1/\hat{\eta}}. \end{aligned} \tag{33}$$

- $$\begin{aligned}\frac{\partial \boldsymbol{\ell}}{\partial \eta} &= \frac{\kappa}{\eta} - \kappa \log(\alpha) + \sum_{i \in U} \log(y_i) - \sum_{i=1}^{n} (\frac{y_i}{\alpha})^\eta \log\Big(\frac{y_i}{\alpha}\Big) = 0 \\ \frac{\kappa}{\eta} - \kappa \log(\alpha) + \sum_{i \in U} \log(y_i) - \frac{\sum_{i=1}^{n} (y_i)^\eta [\log(y_i) - \log(\alpha)]}{\alpha^\eta} &= 0 \\ \frac{\kappa}{\eta} - \kappa \log(\alpha) + \sum_{i \in U} \log(y_i) - \frac{\sum_{i=1}^{n} (y_i)^\eta \log(y_i) - \sum_{i=1}^{n} (y_i)^\eta \log(\alpha)}{\alpha^\eta} &= 0.\end{aligned}$$

  Χρησιμοποιώντας την εκτιμήτρια $\hat{\alpha}$ από τη σχέση (33) έχουμε:

$$\begin{aligned}&\frac{\kappa}{\eta} - \kappa \frac{1}{\eta} \log\Big(\frac{\sum_{i=1}^{n} (y_i)^\eta}{\kappa}\Big) + \sum_{i \in U} \log(y_i) - \frac{\sum_{i=1}^{n} (y_i)^\eta \log(y_i) - \frac{1}{\eta} \log\Big(\frac{\sum_{i=1}^{n} (y_i)^\eta}{\kappa}\Big) \sum_{i=1}^{n} (y_i)^\eta}{\frac{\sum_{i=1}^{n} (y_i)^\eta}{\kappa}} = 0 \\ &\frac{\kappa}{\eta} - \kappa \frac{1}{\eta} \log\Big(\frac{\sum_{i=1}^{n} (y_i)^\eta}{\kappa}\Big) + \sum_{i \in U} \log(y_i) - \kappa \Big[\frac{\sum_{i=1}^{n} (y_i)^\eta \log(y_i)}{\sum_{i=1}^{n} (y_i)^\eta} - \frac{1}{\eta} \log\Big(\frac{\sum_{i=1}^{n} (y_i)^\eta}{\kappa}\Big) \Big] = 0\end{aligned} \tag{34}$$

$$\frac{\kappa}{\eta}+\sum_{i\in U}\log(y_i)-\kappa\frac{\sum_{i=1}^{n}(y_i)^{\eta}\log(y_i)}{\sum_{i=1}^{n}(y_i)^{\eta}}=0 \quad \leftrightarrow \quad \frac{\sum_{i=1}^{n}(y_i)^{\eta}\log(y_i)}{\sum_{i=1}^{n}(y_i)^{\eta}}-\frac{1}{\eta}-\frac{\sum_{i\in U}\log(y_i)}{\kappa}=0. \tag{35}$$

Η επίλυση της εξίσωσης (35) οδηγεί στην εκτίμηση μέγιστης πιθανοφάνειας της παραμέτρου η και κατά επέκταση της παραμέτρου α. Προκειμένου να βρούμε ρίζα στην μη γραμμική εξίσωση που καταλήξαμε χρησιμοποιούμε την αριθμητική μέθοδο **Newton-Raphson**. Η αναδρομική σχέση της μεθόδου στο πρόβλημά μας έχει τη μορφή:

$$\eta_{n+1}=\eta_n-\frac{f(\eta_n)}{f'(\eta_n)}.$$

Ξεκινώντας με μία καλή αρχική λύση, έστω η_0 της εξίσωσης $f(\eta)=0$ υπολογίζουμε καινούργιες ρίζες της εξίσωσης με βάση την παραπάνω αναδρομική σχέση, έως ότου η διαφορά έστω $|\eta_k-\eta_{k-1}|$ μεταξύ δύο διαδοχικών επαναλήψεων-εκτιμήσεων να είναι μικρότερη από μια επιλεγμένη ακρίβεια (tolerance). Στην περίπτωσή μας θέτουμε:

$$f(\eta)=\frac{\sum_{i=1}^{n}(y_i)^{\eta}\log(y_i)}{\sum_{i=1}^{n}(y_i)^{\eta}}-\frac{1}{\eta}-\frac{\sum_{i\in U}\log(y_i)}{\kappa}$$

$$f'(\eta)=\frac{\sum_{i=1}^{n}(y_i)^{\eta}(\log(y_i))^2}{\sum_{i=1}^{n}(y_i)^{\eta}}-\frac{(\sum_{i=1}^{n}(y_i)^{\eta}\log(y_i))^2}{(\sum_{i=1}^{n}(y_i)^{\eta})^2}+\frac{1}{\eta^2}.$$

Για τα δεδομένα μας παρατηρούμε ότι $f(1)=-0.67$ και $f(2)=0.02$, $f(3)=0.34$ κ.ο.κ. Επομένως μια αρχικά καλή τιμή η_0 για τη μέθοδο *Newton-Raphson* είναι η $\eta_0=1.5$. Από την υλοποίηση της μεθόδου στο προγραμματιστικό περιβάλλον της R έχουμε ως έξοδο την τιμή $\hat{\eta}=1.9467$ και η αντίστοιχη τιμή της παραμέτρου α είναι $\hat{\alpha}=91.6383$.

Ο Εσσιανός πίνακας (Hessian matrix) υπολογισμένος στις τιμές $\hat{\eta},\hat{\alpha}$ έχει τη μορφή:

$$H=\begin{bmatrix}\frac{\partial^2\boldsymbol{\ell}}{\partial\alpha^2} & \frac{\partial^2\boldsymbol{\ell}}{\partial\alpha\partial\eta}\\ \frac{\partial^2\boldsymbol{\ell}}{\partial\eta\partial\alpha} & \frac{\partial^2\boldsymbol{\ell}}{\partial\eta^2}\end{bmatrix}_{\hat{\alpha},\hat{\eta}}=\begin{bmatrix}-0.0085 & 0.0744\\ 0.0234 & -8.5192\end{bmatrix}.$$

Οι αντίστοιχες ιδιοτιμές του πίνακα H είναι $\lambda_1=-8.5194<0$ και $\lambda_2=-0.0083<0$ συνεπώς είμαστε βέβαιοι ότι οι εκτιμήτριες της μεθόδου πιθανοφάνειας αποτελούν σημεία μεγιστοποίησης της συνάρτησης πιθανοφάνειας των δεδομένων μας.

Η τιμή της λογαριθμοποιημένης πιθανοφάνειας στο σημείο μεγίστου είναι:

$$\boldsymbol{\ell}_{max} = \kappa \log(\hat{\eta}) - \kappa\hat{\eta}\log(\hat{\alpha}) + (\hat{\eta}-1)\sum_{i\in U}\log(y_i) - \sum_{i=1}^{n}(\frac{y_i}{\hat{\alpha}})^{\hat{\eta}} = -100.3785.$$

Επομένως το κριτήριο BIC για το Weibull μοντέλο παίρνει την τιμή:

$$BIC(weib) = -2(-100.3785) + 2 \cdot \log(25) = 207.19.$$

Υπολογίζουμε επιπλέον τις τιμές του κριτηρίου AIC για το εκθετικό και το Weibull μοντέλο. Με βάση τις σχέσεις των λογαριθμοποιημένων πιθανοφανειών των δύο μοντέλων έχουμε ότι:

$$AIC(exp) = -2\boldsymbol{\ell}_{max}(exp) + 2 \cdot dim(exp) = 212.87$$
$$AIC(weib) = -2\boldsymbol{\ell}_{max}(weib) + 2 \cdot dim(weib) = 204.75.$$

	Εκθετικό μοντέλο	$Weibull$ μοντέλο
AIC	212.87	204.75
BIC	214.09	207.19

Πίνακας 2.2: Τιμές των AIC,BIC κριτηρίων πληροφορίας για το εκθετικό και το weibull μοντέλο.

Από τον Πίνακα 2.2 βλέπουμε ότι και τα δύο κριτήρια έχουν μικρότερη τιμή στο Weibull μοντέλο με αποτέλεσμα να προτιμάται του εκθετικού. Το κριτήριο BIC παρατηρούμε από τις μεγάλες τιμές που παίρνει ότι ποινικοποιεί με αυστηρότερο βαθμό τα δύο μοντέλα από ότι το AIC.

Το κριτήριο BIC αναπτύχθηκε θεωρητικά με βάση τον ισχυρισμό ότι το πραγματικό μοντέλο που περιγράφει πλήρως τα δεδομένα ανήκει στο σύνολο των υποψήφιων μοντέλων, ενώ το AIC και οι επεκτάσεις του δεν απαιτούν κάτι τέτοιο. Άμεσα σχετική ιδιότητα του BIC (την οποία θα δούμε και παρακάτω) είναι ότι με πιθανότητα να τείνει στο ένα καθώς το μέγεθος του δείγματος των δεδομένων αυξάνεται, το κριτήριο αναγνωρίζει το κατάλληλο μοντέλο με τις παραμέτρους που γέννησαν τα δεδομένα με την προϋπόθεση ότι αυτό υπάρχει (ιδιότητα συνέπειας). Για το λόγο αυτό στο κριτήριο BIC ευνοούνται περισσότερο τα φειδωλά μοντέλα και ποινικοποιείται σε μεγάλο βαθμό η πολυπλοκότητα των μοντέλων.

2.7 Επιλογή μεταβλητών

Έχοντας αναλύσει σε ένα βαθμό κάποια θεμελιώδη κριτήρια πληροφορίας και τον τρόπο αξιοποίησής τους σε μοντέλα παλινδρόμησης, συνεχίζουμε στα γραμμικά μοντέλα μελετώντας την εφαρμογή των κριτηρίων στο πρόβλημα της επιλογής μεταβλητών (variable selection).

Πλήρης εξερεύνηση

Εάν υποθέσουμε ότι διαθέτουμε δεδομένα για p υποψήφιες επεξηγηματικές μεταβλητές και προσπαθούμε να βρούμε το γραμμικό μοντέλο με την καλύτερη προσαρμογή, τότε η πιο αποτελεσματική στρατηγική που μας εξασφαλίζει την εύρεσή του είναι η πλήρης εξερεύνηση του χώρου μοντελοποίησης. Όλα τα πιθανά μοντέλα (χωρίς αλληλεπιδράσεις) που μπορούν να προκύψουν από τις p μεταβλητές είναι $\mathbf{2^p}$ το πλήθος. Προσαρμόζοντας και κατατάσσοντας τα μοντέλα με βάση κάποιο κριτήριο πληροφορίας που επιλέγουμε λαμβάνοντας υπόψιν και του σκοπούς της ανάλυσης, διαλέγουμε το μοντέλο με την ελάχιστη τιμή του κριτηρίου.

Η παραπάνω διαδικασία είναι εφικτή για σχετικά μικρό αριθμό επεξηγηματικών μεταβλητών (predictors). Όσο όμως αυξάνεται ο αριθμός τους εμφανίζεται το πρόβλημα του υπολογιστικού κόστους. Ενδεικτικά, εάν διαθέτουμε 15 ανεξάρτητες μεταβλητές τότε ο αριθμός των μοντέλων που πρέπει να εξετάσουμε είναι $2^{15} = 32768$. Για ακόμα μεγαλύτερο αριθμό του p η στρατηγική της πλήρους εξερεύνησης αλλά και διάφορων παραλλαγών της καθίσταται ανέφικτη.

Με αφορμή την παραπάνω δυσκολία αναπτύχθηκαν οι δημοφιλείς μέθοδοι επιλογής μεταβλητών κατά βήματα (stepwise procedures). Προκειμένου να αποφευχθεί η εξερεύνηση όλων των δυνατών μοντέλων, οι διαδικασίες κατά βήματα ξεκινούν με ένα μοντέλο αφετηρία, που είναι συνήθως το μοντέλο που περιέχει όλες τις επεξηγηματικές μεταβλητές (πλήρες) ή το μοντέλο με το σταθερό όρο β_0 μόνο, και σε κάθε βήμα εξετάζουν αν η προσθήκη ή η αφαίρεση μίας επεξηγηματικής μεταβλητής βελτιώνει την τιμή κάποιου κριτηρίου πληροφορίας μέχρις ότου να μην υπάρχει περαιτέρω βελτίωση στην τιμή του κριτηρίου. Οι τρεις πιο γνωστές μέθοδοι επιλογής μεταβλητών κατά βήματα είναι η διαδικασία της διαδοχικής αφαίρεσης, της διαδοχικής πρόσθεσης και η διαδικασία κατά βήματα που συνδυάζει τη λογική των προηγούμενων δύο.

Διαδικασία της διαδοχικής αφαίρεσης: Στη μέθοδο της διαδοχικής αφαίρεσης ξεκινάμε με το μοντέλο που περιέχει όλες τις ανεξάρτητες μεταβλητές και υπολογίζουμε την τιμή του στο χρησιμοποιούμενο κριτήριο πληροφορίας. Στη συνέχεια αφαιρώντας μία προς μία τις μεταβλητές και προσαρμόζοντας εκ νέου το γραμμικό μοντέλο, εξετάζουμε εάν βελτιώνεται η τιμή του κριτηρίου. Ε-

παναλαμβάνουμε την παραπάνω διαδικασία μέχρις ότου η αφαίρεση οποιασδήποτε μεταβλητής να μην βελτιώνει το κριτήριο. Τα παραπάνω βήματα δουλεύοντας με το κριτήριο AIC συνοπτικά είναι:

1. Εισάγουμε όλες τις διαθέσιμες επεξηγηματικές μεταβλητές στο μοντέλο.
2. Προσαρμόζουμε όλα τα μοντέλα που προκύπτουν με την αφαίρεση μίας από τις p μεταβλητές και υπολογίζουμε την τιμή τους στο κριτήριο πληροφορίας AIC.
3. Επιλέγουμε να αφαιρέσουμε την μεταβλητή που ανήκει στο μοντέλο του προηγούμενου βήματος 2 με τη μικρότερη τιμή του κριτηρίου AIC.
4. Προσαρμόζουμε εκ νέου το μοντέλο που προκύπτει χωρίς τη μεταβλητή του βήματος 3 και υπολογίζουμε την AIC τιμή του.
5. Επαναλαμβάνουμε τα βήματα 2 και 3 μέχρις ότου από την αφαίρεση οποιασδήποτε μεταβλητής να μην προκύψει μοντέλο με μικρότερη AIC τιμή.

Διαδικασία της διαδοχικής πρόσθεσης: Στην ίδια λογική στηρίζεται και η διαδικασία της διαδοχικής πρόσθεσης, με τη βασική όμως διαφορά ότι ξεκινάμε με το μοντέλο που περιέχει τον σταθερό όρο μόνο και σε κάθε βήμα ελέγχουμε εάν η πρόσθεση κάποιας μεταβλητής στο μοντέλο βελτιώνει την τιμή του κριτηρίου πληροφορίας.

Διαδικασία κατά βήματα: Η διαδικασία κατά βήματα συνδυάζει τις προηγούμενες δύο μεθόδους. Ξεκινάμε με το μοντέλο χωρίς καμία ανεξάρτητη μεταβλητή και εξετάζουμε η πρόσθεση ποιας μεταβλητής βελτιώνει το κριτήριο πληροφορίας. Από εκείνο το σημείο και μετά, εξετάζεται σε κάθε βήμα εκτός από την πρόσθεση επιπλέον μεταβλητών, εάν η αφαίρεση κάποιας μεταβλητής που εισήχθη βελτιώνει το μοντέλο.

Παρατηρήσεις:

- Από τις τρεις μεθόδους προτιμάται η διαδικασία κατά βήματα καθώς συνδυάζει την λογική της διαδοχικής αφαίρεσης και της διαδοχικής πρόσθεσης. Για μικρό αριθμό επεξηγηματικών μεταβλητών $p < n$ η μέθοδος κατά βήματα εφαρμόζεται με αρχικό μοντέλο το πλήρες, διαφορετικά για $p > n$ χρησιμοποιείται ως αφετηρία το μοντέλο με το σταθερό όρο.

- Η διαδικασία διαδοχικής πρόσθεσης είναι λιγότερο υπολογιστικά έντονη από τις άλλες δύο γιατί προσαρμόζονται λιγότερα μοντέλα συνολικά. Η διαδοχική αφαίρεση έχει το μειονέκτημα ότι οποιαδήποτε μεταβλητή αφαιρεθεί δεν μπορεί να επανενταχθεί στο μοντέλο σε κάποιο επόμενο βήμα της μεθόδου.

- Οι μέθοδοι κατά βήματα συνήθως επιλέγουν καλά μοντέλα αλλά όχι βέλτιστα. Αυτό δικαιολογείται από το γεγονός ότι με την προσθήκη και την αφαίρεση μόνο μεταβλητών ελοχεύει ο κίνδυνος να εγκλωβιστούν σε τοπικό μέγιστο του χώρου μοντελοποίησης. Η κατάληξη αυτή είναι σχεδόν βέβαιη εάν μάλιστα ο αριθμός p των μεταβλητών είναι πολύ μεγαλύτερος του πλήθους των διαθέσιμων παρατηρήσεων n.

- Στην περίπτωση μάλιστα που αρκετές μεταβλητές είναι υψηλά γραμμικά συσχετισμένες (φαινόμενο πολυσυγγραμμικότητας) η βέλτιστη επιλογή μεταβλητών γίνεται δύσκολη υπόθεση και οι μέθοδοι κατά βήματα δεν είναι αποτελεσματικές.

- Ωστόσο, ελλείψει των προηγούμενων προβλημάτων μπορούν να αποτελέσουν μία καλή αρχική βάση επιλογής επεξηγηματικών μεταβλητών.

Κεφάλαιο 3

Ιδιότητες κριτηρίων πληροφορίας

Στο παρόν κεφάλαιο συγκρίνουμε κάποια από τα κριτήρια πληροφορίας που αναπτύχθηκαν προηγουμένως ως προς την ασθενή και την ισχυρή συνέπεια οι οποίες αποτελούν κλασσικές έννοιες στη θεωρία της επιλογής μοντέλου. Αν κάνουμε την υπόθεση ότι ανάμεσα στα υποψήφια προς επιλογή μοντέλα υπάρχει εκείνο που γέννησε τα πραγματικά δεδομένα, τότε θα θέλαμε η χρησιμοποιούμενη μέθοδος επιλογής να μπορεί να αναγνωρίσει αυτό το μοντέλο.

Η συνέπεια χωρίζεται σε δύο μορφές, την ασθενή και την ισχυρή. Μία μέθοδος επιλογής μοντέλου είναι ασθενώς συνεπής όταν με πιθανότητα που πλησιάζει το ένα μπορεί και επιλέγει το μοντέλο που γέννησε τα δεδομένα, καθώς το μέγεθος του δείγματος $n \to \infty$. Η ισχυρή συνέπεια χαρακτηρίζει μια μέθοδο όταν η επιλογή του πραγματικού μοντέλου συμβαίνει σχεδόν σίγουρα (ή σχεδόν παντού με πιθανότητα 1). Συχνά δε θέλουμε να κάνουμε την υπόθεση ότι το πραγματικό μοντέλο υπάρχει στο σύνολο των προς εξέταση μοντέλων. Σε αυτή την περίπτωση υποθέτουμε ότι υπάρχει μοντέλο ανάμεσα στα υποψήφια με την ελάχιστη K-L απόσταση από το πραγματικό. Η ασθενής και η ισχυρή συνέπεια τότε υποδεικνύει ότι το μοντέλο που επιλέγεται, με πιθανότητα να τείνει στο ένα και αντίστοιχα σχεδόν σίγουρα, είναι το κοντινότερο από το πραγματικό κατά απόσταση K-L.

Ξεκινάμε με τους ακριβείς ορισμούς της ασθενούς και ισχυρής συνέπειας των κριτηρίων πληροφορίας (οι οποίες δεν πρέπει να συγχέονται με τις αντίστοιχες έννοιες της θεωρίας εκτιμητριών) και εξετάζουμε στη συνέχεια ως προς τη συνέπειά τους τα κριτήρια πληροφορίας AIC, AIC_c, BIC.

Η πλειοψηφία των κριτηρίων πληροφορίας έχουν κοινή μορφή την οποία αναγνωρίζουμε παραδείγματος χάριν από τα κριτήρια AIC, BIC:

$$AIC(M) = 2\boldsymbol{\ell}_{max} - 2 \cdot dim(M)$$
$$BIC(M) = 2\boldsymbol{\ell}_{max} - \log(n) \cdot dim(M).$$

Η μορφή τους συνθέτεται από το διπλάσιο της μεγιστοποιημένης λογαριθμικής πιθανοφάνειας του μοντέλου και έναν όρο ποινικοποίησης της πολυπλοκότητας του μοντέλου. Για τους σκοπούς του κεφαλαίου ακολουθούμε αυτή την γραφή των κριτηρίων πληροφορίας που ισοδυναμεί με την ***μεγιστοποίηση*** τους για την επιλογή του βέλτιστου μοντέλου.

Έστω τώρα ότι διαθέτουμε $\kappa = 1, 2, \ldots, K$ προς επιλογή μοντέλα και $i = 1, \ldots, n$ παρατηρήσεις δεδομένων. Συμβολίζουμε με $\boldsymbol{\theta}_\kappa$ το διάνυσμα των παραμέτρων του μοντέλου κ και τη συνάρτηση πυκνότητάς του για την i παρατήρηση με $f_{\kappa,i}$. Στην περίπτωση των μοντέλων παλινδρόμησης γράφουμε $f_{\kappa,i} = f_\kappa(y_i|\boldsymbol{X}_i; \boldsymbol{\theta}_\kappa)$. Τα κριτήρια AIC και BIC μπορούν να πάρουν με τους νέους συμβολισμούς τη γενική μορφή:

$$IC(M_\kappa) = 2\sum_{i=1}^{n} \log f_{\kappa,i}(y_i; \hat{\boldsymbol{\theta}}_\kappa) - c_{n,\kappa}. \tag{36}$$

Ο πρώτος όρος είναι η τιμή της μεγιστοποιημένης λογαριθμικής πιθανοφάνειας του μοντέλου M_κ, όπου $\hat{\boldsymbol{\theta}}_\kappa$ οι εκτιμήτριες μέγιστης πιθανοφάνειας του $\boldsymbol{\theta}_\kappa$. Ο πολλαπλασιαστής του πρώτου όρου υπάρχει για καθαρά ιστορικούς λόγους. Ο δεύτερος όρος $c_{n,\kappa}$ αποτελεί τον όρο ποινικοποίησης του μοντέλου M_κ που προσαρμόστηκε σε τυχαίο δείγμα δεδομένων μεγέθους n. Για το κριτήριο AIC θα ισχύει ότι $c_{n,\kappa} = 2 \cdot dim(M_\kappa)$ και στο BIC θα ισχύει ότι $c_{n,\kappa} = \log(n) \cdot dim(M_\kappa)$. Τη μορφή αυτή μπορεί να πάρει και το κριτήριο AIC_c.

Ακολουθούν τα θεωρήματα ασθενούς και ισχυρής συνέπειας που χαρακτηρίζουν τα κριτήρια πληροφορίας. Για τις αποδείξεις και περαιτέρω ανάλυση των θεωρημάτων παραπέμπουμε στους Sin και White (1996).

3.1 Ασθενής και ισχυρή Συνέπεια

Θεώρημα 1 (Ασθενής Συνέπεια)**.** *Έστω ότι ανάμεσα στα υποψήφια μοντέλα υπάρχει ακριβώς ένα μοντέλο M_{κ_0} που απέχει την ελάχιστη $K - L$ απόσταση από το πραγματικό μοντέλο g που γέννησε τα δεδομένα. Αυτό ισοδυναμεί με ότι:*

$$\lim_{n\to\infty} inf \ \min_{\kappa\neq\kappa_0} \left\{ n^{-1} \sum_{i=1}^{n} \{KL(f_{\kappa,i}, g) - KL(f_{\kappa_0,i}, g)\} \right\} > 0. \tag{37}$$

Έστω επιπλέον ότι ο όρος ποινικοποίησης είναι της τάξης $o_p(n)$ (σύγκλιση

κατά πιθανότητα). Τότε, με πιθανότητα που τείνει στο ένα το κριτήριο πληροφορίας επιλέγει το μοντέλο M_{κ_0} ως βέλτιστο.

Από το θεώρημα 1[2] προκύπτει άμεσα ότι προκειμένου ένα κριτήριο πληροφορίας να είναι ασθενώς συνεπές και κατά επέκταση να επιλέγει το μοντέλο με την ελάχιστη $K-L$ απόσταση με πιθανότητα να τείνει στο ένα, πρέπει ο όρος ποινικοποίησης $c_{n,\kappa}$ όταν διαιρείται με το n να τείνει στο 0 για $n \to \infty$.

Εφαρμόζοντας τη συνέπεια του θεωρήματος στα κριτήρια AIC, AIC$_c$, BIC έχουμε ότι:

$$\text{Για το } AIC: \quad c_{n,\kappa} = 2 \cdot dim(M_\kappa) \quad \text{και } \frac{c_{n,\kappa}}{n} = \frac{2 \cdot dim(M_\kappa)}{n} \xrightarrow[n\to\infty]{} 0$$

$$\text{Για το } AIC_c: \quad c_{n,\kappa} = 2 \cdot dim(M_\kappa) + 2 \cdot dim(M_\kappa)\frac{dim(M_\kappa)+1}{n - dim(M_\kappa) - 1}$$

και $\dfrac{c_{n,\kappa}}{n} \xrightarrow[n\to\infty]{} 0$ από ιδιότητες ορίων ρητής συνάρτησης.

Η οριακή περίπτωση όπου ένας παράγοντας του $c_{n,\kappa}$ είναι μηδέν, δεν υφίσταται στα κριτήρια AIC,AIC$_c$.

$$\text{Για το } BIC: \quad c_{n,\kappa} = \log(n) \cdot dim(M_\kappa) \quad \text{και } \frac{c_{n,\kappa}}{n} = \frac{\log(n) \cdot dim(M_\kappa)}{n} \xrightarrow[n\to\infty]{} 0.$$

Στην ακραία περίπτωση όπου $\log(n) = 0 \ \leftrightarrow n = 1$ έχουμε ότι $dim(M_\kappa) > 0$.

Βλέπουμε λοιπόν ότι τα κριτήρια AIC,AIC$_c$ και BIC είναι ασθενώς συνεπή και καθώς το $n \to \infty$ τείνουν να επιλέξουν μεταξύ των υποψήφιων μοντέλων ως καταλληλότερο εκείνο που ελαχιστοποιεί την απόσταση K-L από το πραγματικό μοντέλο.

Θεώρημα 2 (Ισχυρή Συνέπεια)**.** *Έστω ότι μεταξύ των υποψήφιων μοντέλων υπάρχει ακριβώς ένα μοντέλο M_{κ_0} με την ελάχιστη απώλεια πληροφορίας K-L από το πραγματικό μοντέλο. Ισοδύναμα:*

- $$\lim_{n\to\infty} \ inf \ \min_{\kappa \neq \kappa_0} \left\{ n^{-1} \sum_{i=1}^{n} \{KL(f_{\kappa,i}, g) - KL(f(\kappa_0, i), g)\} \right\} > 0.$$

[2]Η ασθενής συνέπεια ενός κριτηρίου πληροφορίας ισχύει και στην περίπτωση όπου ένας από τους όρους του $c_{n,\kappa}$ είναι 0, αρκεί όλοι οι υπόλοιποι όροι να είναι αυστηρά θετικοί.

- *Αν ο αυστηρά θετικός όρος ποινικοποίησης είναι τέτοιος ώστε* $c_{n,\kappa} = o(n)$ *σχεδόν παντού, τότε:*

$$P\Big\{ \min_{\ell \neq \kappa_0} (IC(M_{\kappa_0}) - IC(M_\ell)) > 0, \ \textit{σχεδόν για κάθε } n \Big\} = 1.$$

Ισοδύναμα, το κριτήριο πληροφορίας επιλέγει με πιθανότητα ένα για κάθε n το μοντέλο M_{κ_0} ως καταλληλότερο.

Είναι προφανές ότι η συνθήκη του Θεωρήματος 2 ικανοποιείται για τους όρους ποινικοποίησης των κριτηρίων AIC, AIC_c, BIC και επομένως τα κριτήρια αυτά είναι ισχυρά συνεπή.

3.2 Συνέπεια

Ας υποθέσουμε τώρα ότι στα προς επιλογή μοντέλα υπάρχουν περισσότερα από ένα που ελαχιστοποιούν την απώλεια πληροφορίας K-L. Πως διαλέγουμε το βέλτιστο σε αυτή την περίπτωση; Η αρχή της ***φειδωλότητας*** υποδεικνύει ως καταλληλότερο εξ αυτών το «απλούστερο» που ισοδυναμεί με το μοντέλο με τις λιγότερες παραμέτρους. Συχνά στη βιβλιογραφία η φειδωλότητα ταυτίζεται με την έννοια της συνέπειας των κριτηρίων πληροφορίας και αγνοείται δια τούτου η κατάσταση όπου υπάρχει ένα μοναδικό μοντέλο κοντά στο πραγματικό. Συνεχίζουμε με κάποιους συμβολισμούς που θα χρησιμοποιηθούν στον ορισμό της συνέπειας.

Έστω J το σύνολο δεικτών όλων των μοντέλων που απέχουν την ελάχιστη K-L απόσταση από το πραγματικό και $J_0 \subset J$ που περιλαμβάνει τα μοντέλα του J με την μικρότερη διάσταση (αριθμό παραμέτρων). Σημειώνουμε ότι μπορεί να υπάρχουν περισσότερα από ένα τέτοια μοντέλα.

Θεώρημα 3 (Συνέπεια)**.** *Οι δύο παρακάτω συνθήκες είναι ισοδύναμες:*

- *Υποθέτουμε ότι για κάθε* $\kappa_0, \ell_0 \in J$ *με* $\kappa_0 \neq \ell_0$ *ισχύει:*

$$\lim_{n \to \infty} \ sup(n^{-1} \sum_{i=1}^{n} \{KL(f_{\kappa_0,i}, g) - KL(f_{\ell_0,i}, g)\}) < \infty.$$

  Έστω επιπλέον ότι για κάθε $j_0 \in J_0$ *και* $\ell \in J \backslash J_0$*, ο όρος ποινικοποίησης είναι τέτοιος ώστε* $P\{(c_{n,\ell} - c_{n,j_0})/\sqrt{n} \to \infty\} = 1$.

- *Έστω ότι για κάθε* $\kappa_0, \ell_0 \in J$ *με* $\kappa_0 \neq \ell_0$ *ο λόγος των λογαριθμοποιημένων πιθανοφανειών είναι:*

$$\sum_{i=1}^{n} \log\left(\frac{f_{\kappa_0,i}(y_i;\boldsymbol{\theta}^*_{\kappa_0})}{f_{\ell_0,i}(y_i;\boldsymbol{\theta}^*_{\ell_0})}\right) = O_p(1),$$

και επιπλέον ότι για κάθε $j_0 \in J_0$ *και* $\ell \in J\backslash J_0$ *ισχύει ότι:* $P\{c_{n,\ell} - c_{n,j_0} \to \infty\} = 1$.

Αν ικανοποιείται οποιαδήποτε από τις παραπάνω δύο συνθήκες τότε με πιθανότητα που τείνει στο ένα το κριτήριο πληροφορίας θα επιλέξει το απλό μοντέλο M_{j_o} *που ελαχιστοποιεί την απόσταση K-L και έχει τις λιγότερες παραμέτρους:*

$$\lim_{n\to\infty} P\{\min_{\ell\in J\backslash J_0} (IC(M_{j_0}) - IC(M_\ell)) > 0\} = 1.$$

Η δεύτερη συνθήκη του θεωρήματος συνέπειας απαιτεί η κατανομή του στατιστικού του λόγου των λογαριθμοποιημένων πιθανοφανειών να είναι στοχαστικά φραγμένη. Η ασυμπτωτική κατανομή του στατιστικού αυτού μελετάται διεξοδικά από τον Vuong (1989) στον οποίο παραπέμπουμε. Αναφέρουμε ενδεικτικά ότι στη συνηθισμένη περίπτωση όπου τα εξεταζόμενα μοντέλα είναι εμφωλευμένα, είναι γνωστό ότι το διπλάσιο του στατιστικού του λόγου των μεγιστοποιημένων λογαριθμοποιημένων πιθανοφανειών ακολουθεί ασυμπτωτικά την X^2 κατανομή με βαθμούς ελευθερίας ίσους με τη διαφορά του αριθμού των παραμέτρων των δύο μοντέλων και επομένως ικανοποιείται τότε η πρώτη υποσυνθήκη.

Εξετάζουμε τώρα την συνέπεια των κριτηρίων AIC, AIC_c και BIC με βάση την ισχύ της δεύτερης συνθήκης του Θεωρήματος 3. Επικεντρωνόμαστε στη δεύτερη υπόθεση που εμπλέκει τον όρο ποινικοποίησης $c_{n,\kappa}$.

Για το κριτήριο AIC : $\quad c_{n,\kappa} = 2 \cdot dim(M_\kappa)$. Για $j_0 \in J_0$ και $\ell \in J\backslash J_0$ έχουμε:

$c_{n,j_0} = 2 \cdot dim(M_{j_0})$ και $c_{n,\ell} = 2 \cdot dim(M_\ell)$ με $dim(M_{j_0}) < dim(M_\ell)$.

Όμως για $n \to \infty \quad c_{n,\ell} - c_{n,j_0} = 2 \cdot (dim(M_\ell) - dim(M_{j_0})) =$ σταθερό > 0.

Επομένως $P\{c_{n,\ell} - c_{n,j_0} \to \infty\} \neq 1$ και το κριτήριο AIC **δεν** είναι συνεπές.

AIC_c : από τη σχέση (27): $c_{n,\kappa} = 2 \cdot dim(M_\kappa) + 2 \cdot dim(M_\kappa)\dfrac{dim(M_\kappa)+1}{n - dim(M_\kappa) - 1}$.

Για $j_0 \in J_0, \quad \ell \in J \backslash J_0$ με $dim(M_{j_0}) < dim(M_\ell)$ έχουμε:

$c_{n,j_0} = 2 \cdot dim(M_{j_0}) + 2 \cdot dim(M_{j_0})\dfrac{dim(M_{j_0})+1}{n - dim(M_{j_0}) - 1}$ και

$c_{n,\ell} = 2 \cdot dim(M_\ell) + 2 \cdot dim(M_\ell)\dfrac{dim(M_\ell)+1}{n - dim(M_\ell) - 1}$.

Όμως για $n \to \infty$ $c_{n,\ell} - c_{n,j_0} = 2 \cdot (dim(M_\ell) - dim(M_{j_0})) =$ σταθερό > 0.

Επομένως $P\{c_{n,\ell} - c_{n,j_0} \to \infty\} \neq 1$ και το κριτήριο AIC_c **δεν** είναι συνεπές.

Για το BIC έχουμε: $c_{n,\kappa} = \log(n) \cdot dim(M_\kappa)$. Για $j_0 \in J_0$ και $\ell \in J \backslash J_0$ έχουμε:

$c_{n,j_0} = \log(n) \cdot dim(M_{j_0})$ και $c_{n,\ell} = \log(n) \cdot dim(M_\ell)$ με $dim(M_{j_0}) < dim(M_\ell)$.

Για $n \to \infty$, $\log(n) \to \infty$ άρα $c_{n,\ell} - c_{n,j_0} = \log(n) \underbrace{[dim(M_\ell) - dim(M_{j_0})]}_{\text{σταθερό}} \to \infty$

με αποτέλεσμα: $P\{c_{n,\ell} - c_{n,j_0} \to \infty\} = 1$.

Καταλήγουμε λοιπόν στο γενικό συμπέρασμα ότι το κριτήριο BIC **είναι** συνεπές, ενώ το AIC και το AIC_c όχι. Σε αντίθεση με το BIC τα κριτήρια AIC, AIC_c δεν επιλέγουν κάθε φορά το φειδωλότερο μοντέλο από τα υποψήφια μοντέλα που έχουν την ίδια ελάχιστη απόσταση K-L. Η πιθανότητα της υπερπροσαρμογής είναι υπαρκτή με τα κριτήρια AIC, AIC_c ειδικά στις περιπτώσεις που το μέγεθος του δείγματος n είναι μεγάλο. Γενικότερα, τα κριτήρια πληροφορίας με σταθερό όρο ποινικοποίησης που δεν εξαρτάται από το μέγεθος του δείγματος n δεν ικανοποιούν καμία από τις δύο συνθήκες του θεωρήματος της συνέπειας.

Κεφάλαιο 4

Cross Validation και Bootstrap

Στο προηγούμενο κεφάλαιο φτάσαμε στο συμπέρασμα ότι σε προβλήματα επιλογής μοντέλων πρόβλεψης επιθυμούμε τα κριτήρια πληροφορίας που χρησιμοποιούμε να είναι αποδοτικά. Να επιλέγουν δηλαδή για μεγάλα μεγέθη δείγματος μοντέλα με όσο το δυνατόν ελάχιστο σφάλμα πρόβλεψης. Σε αυτό το κεφάλαιο εισάγουμε και μελετάμε τις μη παραμετρικές μεθόδους Cross Validation (εν συντομία CV) και bootstrap οι οποίες εκτιμούν το αναμενόμενο σφάλμα πρόβλεψης με βάση το οποίο γίνεται έπειτα η επιλογή του καταλληλότερου (ή καταλληλότερων) μοντέλου. Η προβλεπτική ικανότητα αλλά και η δυνατότητα των μοντέλων να γενικευθούν μπορεί να αξιολογηθεί χρησιμοποιώντας μια συνάρτηση απώλειας που μετρά το σφάλμα μεταξύ των πραγματικών ($\boldsymbol{y}$) και των προβλεπόμενων τιμών $\hat{f}(\boldsymbol{x})$. Οι πιο συνηθισμένες μορφές της συνάρτησης αυτής είναι:

$$L(y, \hat{f}\left(\boldsymbol{x}\right)) = \begin{cases} (y - \hat{f}\left(\boldsymbol{x}\right))^2 \text{ τετραγωνικό σφάλμα} \\ \left|y - \hat{f}\left(\boldsymbol{x}\right)\right| \text{ απόλυτο σφάλμα} \end{cases} . \tag{38}$$

Ο όρος $\hat{f}\left(\boldsymbol{x}\right)$ υποδηλώνει το μοντέλο που προσαρμόστηκε σε κάποια παρατηρούμενα δεδομένα $\boldsymbol{y}$ που είναι οι τιμές της μεταβλητής απόκρισης (ή ενδιαφέροντος) Y. Στην πράξη η καλή προσαρμογή ενός μοντέλου μπορεί να αξιολογηθεί χρησιμοποιώντας μελλοντικά δεδομένα y_i ανεξάρτητα των παρατηρούμενων με τα οποία έγινε η προσαρμογή των μοντέλων, υπολογίζοντας τη συνάρτηση απώλειας $L(y_i, \hat{f}\left(\boldsymbol{x}_i\right))$ για κάθε μελλοντική παρατήρηση y_i και κατά επέκταση τον μέσο όρο $\frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}\left(\boldsymbol{x}_i\right))$. Αν όμως διαθέτουμε επιπλέον δεδομένα πέρα των παρατηρούμενων θα ήταν προτιμότερο να συνδυαστούν μεταξύ τους αποσκοπώντας σε καλύτερη προσαρμογή των υπό εξέταση μοντέλων δημιουργώντας όμως έλλειψη στην εκτίμηση του αναμενόμενου προβλεπτικού σφάλματος.

Η cross validation μέθοδος είναι μια τεχνική που στηρίζεται αποκλειστικά στα παρατηρούμενα δεδομένα και ξεπερνάει το προαναφερθέν πρόβλημα. Συνήθως ως συνάρτηση απώλειας χρησιμοποιείται το τετραγωνικό σφάλμα απώλειας και η αξιολόγηση της προβλεπτικής ικανότητας των μοντέλων γίνεται μέσω του μέσου τετραγωνικού προβλεπτικού σφάλματος (predictive mean squared error) το οποίο ορίζεται ως:

$$PSE = \frac{1}{n}\sum_{\alpha'=1}^{n}\mathbb{E}\left[\left(y_{\alpha'} - \hat{f}\left(\boldsymbol{x}_{\alpha'}\right)\right)^2\right], \tag{39}$$

όπου οι τιμές $y_{\alpha'}, \alpha = 1, \ldots, n$, είναι τυχαία επιλεγμένα μελλοντικά δεδομένα της μεταβλητής Y στα σημεία $\boldsymbol{x}_{\alpha'}$, ανεξάρτητα των παρατηρούμενων δεδομένων που χρησιμοποιήθηκαν για την προσαρμογή του μοντέλου $\hat{f}\left(\boldsymbol{x}\right)$. Αν χρησιμοποιήσουμε αποκλειστικά τα παρατηρούμενα δεδομένα, έστω $(y_\alpha, \boldsymbol{x}_\alpha), \alpha = 1, \ldots, n$, και το άθροισμα των τετραγωνικών υπολοίπων $y_\alpha - \hat{f}\left(\boldsymbol{x}_\alpha\right)$, τότε η εκτίμηση του PSE θα είναι:

$$RSS = \frac{1}{n}\sum_{\alpha=1}^{n}\left[y_\alpha - \hat{f}\left(\boldsymbol{x}_\alpha\right)\right]^2. \tag{40}$$

Η ποσότητα RSS (mean residual sum of squares) αναμένεται να υποεκτιμά το πραγματικό προβλεπτικό σφάλμα διότι τα ίδια δεδομένα χρησιμοποιούνται για την προσαρμογή και την αξιολόγηση των μοντέλων. Επιπλέον αρκετές φορές τα μοντέλα προσαρμόζονται ικανοποιητικά στα αρχικά δεδομένα με αποτέλεσμα η εκτίμηση RSS να δίνει την ψευδή εντύπωση ότι έχουν ικανοποιητική προβλεπτική ικανότητα.

Η μέθοδος CV υπολογίζει μία εκτίμηση του μέσου τετραγωνικού προβλεπτικού σφάλματος χωρίζοντας σε δύο μέρη τα διαθέσιμα δεδομένα. Το πρώτο μέρος (*training data*) χρησιμοποιείται για την προσαρμογή του υπό εξέταση μοντέλου και με το δεύτερο μέρος (*test data*) γίνεται η αξιολόγησή του. Πιο αναλυτικά, εξετάζουμε τη μέθοδο leave one out cross validation στην Παράγραφο 4.1, την K-fold cross validation στην Παράγραφο 4.2, και την Generalized cross validation στην παράγραφο 4.3.

4.1 Leave one out cross validation

Έστω ότι διαθέτουμε τα δεδομένα $(y_\alpha, \boldsymbol{x}_\alpha), \alpha = 1, \ldots, n$, και θέλουμε να εκτιμήσουμε το αναμενόμενο προβλεπτικό σφάλμα ενός μοντέλου $f(\boldsymbol{x})$:

- **Βήμα 1**: Από τις n διαθέσιμες παρατηρήσεις αφαιρούμε την α-οστή $(y_\alpha, \boldsymbol{x}_\alpha)$. Μετά, χρησιμοποιώντας τις υπόλοιπες $n-1$ παρατηρήσεις προ-

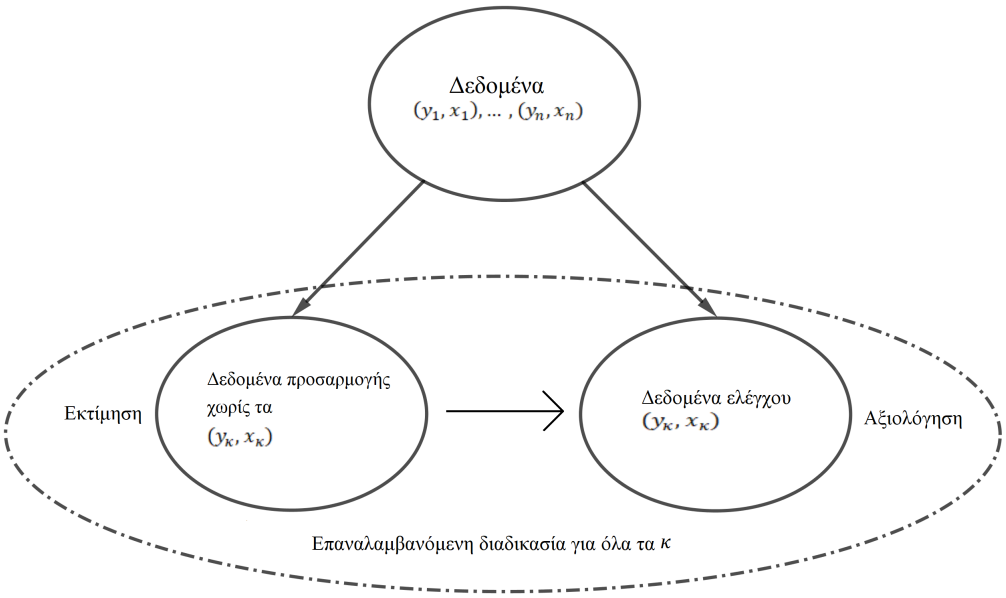
σαρμόζουμε και εκτιμούμε τις παραμέτρους του μοντέλου. Το μοντέλο αυτό το συμβολίζουμε με $\hat{f}^{-\alpha}(\boldsymbol{x})$.

- **Βήμα 2**: Για την α-οστή παρατήρηση που αφαιρέθηκε στο βήμα 1 υπολογίζουμε την τιμή του $\left[y_\alpha - \hat{f}^{-\alpha}(\boldsymbol{x}_\alpha)\right]^2$ τετραγωνικού σφάλματος.
- **Βήμα 3**: Επαναλαμβάνουμε τα βήματα 1 και 2 για όλες τις παρατηρήσεις $\alpha \in \{1, \ldots, n\}$.

Στο τέλος των τριών βημάτων υπολογίζουμε ως εκτίμηση του μέσου τετραγωνικού προβλεπτικού σφάλματος την ποσότητα:

$$CV = \frac{1}{n}\sum_{\alpha=1}^{n}\left(y_\alpha - \hat{f}^{-\alpha}(\boldsymbol{x}_\alpha)\right)^2. \tag{41}$$

Στο Διάγραμμα 4.1 βλέπουμε ποιοτικά πώς λειτουργεί η leave one out cross validation μέθοδος (LOOCV). Για όλα τα $\kappa \in \{1, \ldots, n\}$ χωρίζονται τα δεδομένα στις παρατηρήσεις χωρίς το ζεύγος (y_κ, x_κ) οι οποίες χρησιμοποιούνται για την εκτίμηση των παραμέτρων του μοντέλου και στη συνέχεια γίνεται η αξιολόγηση του μοντέλου στην παρατήρηση (y_κ, x_κ) με βάση το τετραγωνικό σφάλμα:



Διάγραμμα 4.1: Σχεδιάγραμμα υλοποίησης μεθόδου leave on out CV.

Πρόταση 1. *To leave on out cross validated σφάλμα CV αποτελεί υπό συγκεκριμένες προϋποθέσεις αμερόληπτο ασυμπτωτικά εκτιμητή του μέσου αναμενόμενου τετραγωνικού σφάλματος PSE.*

Απόδειξη. Έστω ότι το μοντέλο που εξετάζουμε είναι της μορφής $Y = f(\boldsymbol{X})+\varepsilon$ όπου ε το τυχαίο σφάλμα κανονικά κατανεμημένο με $\mathbb{E}[\varepsilon] = 0$ και $\mathbb{E}[\varepsilon^2] = \sigma^2 = Var(\varepsilon)$. Θεωρούμε επιπλέον μελλοντικές παρατηρήσεις της μεταβλητής Y έστω $y_\alpha^{'}, \alpha = 1, \ldots, n$ στα σημεία $\boldsymbol{x}_\alpha$ οι οποίες προέρχονται από το μοντέλο με τυχαίο τρόπο.

- Για το αναμενόμενο τετραγωνικό σφάλμα έχουμε ότι:

$$\begin{aligned} PSE &= \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(y_\alpha^{'} - \hat{f}(\boldsymbol{x}_a)\right)^2\right] \\ &= \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(y_\alpha^{'} - f\left(\boldsymbol{x}_\alpha\right) + f\left(\boldsymbol{x}_\alpha\right) - \hat{f}\left(\boldsymbol{x}_\alpha\right)\right)^2\right] \\ &= \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(y_\alpha^{'} - f(\boldsymbol{x}_\alpha)\right)^2\right] + \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(f(\boldsymbol{x}_\alpha) - \hat{f}(\boldsymbol{x}_\alpha)\right)^2\right] \\ &= \sigma^2 + \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(f(\boldsymbol{x}_\alpha) - \hat{f}(\boldsymbol{x}_\alpha)\right)^2\right]. \end{aligned} \tag{42}$$

- Το αναμενόμενο *cross validated* σφάλμα της σχέσης (41) γράφεται ως:

$$\begin{aligned} \mathbb{E}\left[CV\right] &= \mathbb{E}\left[\frac{1}{n}\sum_{\alpha=1}^{n}\left[y_\alpha^{'} - \hat{f}^{-\alpha}(\boldsymbol{x}_\alpha)\right]^2\right] \\ &= \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(y_\alpha^{'} - f(\boldsymbol{x}_\alpha) + f(\boldsymbol{x}_\alpha) - \hat{f}^{-\alpha}(\boldsymbol{x}_\alpha)\right)^2\right] \\ &= \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(y_\alpha^{'} - f(\boldsymbol{x}_\alpha)\right)^2\right] + \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(f(\boldsymbol{x}_\alpha) - \hat{f}^{-\alpha}(\boldsymbol{x}_\alpha)\right)^2\right] \\ &= \sigma^2 + \frac{1}{n}\sum_{\alpha=1}^{n}\mathbb{E}\left[\left(f(\boldsymbol{x}_\alpha) - \hat{f}^{-\alpha}(\boldsymbol{x}_\alpha)\right)^2\right]. \end{aligned} \tag{43}$$

Συνεπώς από τις σχέσεις (42) και (43) και δεδομένου ότι ασυμπτωτικά $\hat{f}^{-\alpha}(\boldsymbol{x}_\alpha) = \hat{f}(\boldsymbol{x}_\alpha)$ καταλήγουμε ότι $\lim_{n\to\infty}\mathbb{E}\left[CV|PSE\right] = PSE$. Επομένως, το leave one out CV σφάλμα είναι ασυμπτωτικά αμερόληπτος εκτιμητής του PSE σφάλματος. □

4.2 K-fold cross validation

Η leave on out cross validation μέθοδος μπορεί να γενικευθεί στην K-fold cross validation μέθοδο η οποία υλοποιείται ως εξής: Χωρίζουμε τα παρατηρούμενα δεδομένα σε K ισοπληθικά σύνολα-φακέλους ($folds$). Στον Πίνακα 5.1 απεικονίζεται ένα στιγμιότυπο του διαχωρισμού των δεδομένω όταν εφαρμόζεται η $5-fold$ cross validation μέθοδος.

1 Εκτίμηση	2 Εκτίμηση	3 Αξιολόγηση	4 Εκτίμηση	5 Εκτίμηση

Πίνακας 4.1: $5-fold$ CV στιγμιότυπο διαχωρισμού των δεδομένων σε 5 σύνολα-φακέλους εκ των οποίων οι 1, 2, 4 και 5 χρησιμοποιούνται για την προσαρμογή του μοντέλου και ο φάκελος 3 για την αξιολόγησή του.

Η διαδικασία περιγράφεται ως εξής: Για το k-οστό σύνολο δεδομένων προσαρμόζουμε το μοντέλο στα υπόλοιπα $K-1$ σύνολα δεδομένων και υπολογίζουμε το προβλεπτικό σφάλμα του μοντέλου για τα δεδομένα που περιέχονται στο k οστό σύνολο. Επαναλαμβάνουμε για όλα τα $k = 1, 2, \ldots, K$ και στο τέλος συνδυάζουμε τις εκτιμήσεις από τα K τμήματα δεδομένων για να προκύψει η εκτίμηση του μέσου τετραγωνικού προβλεπτικού σφάλματος.

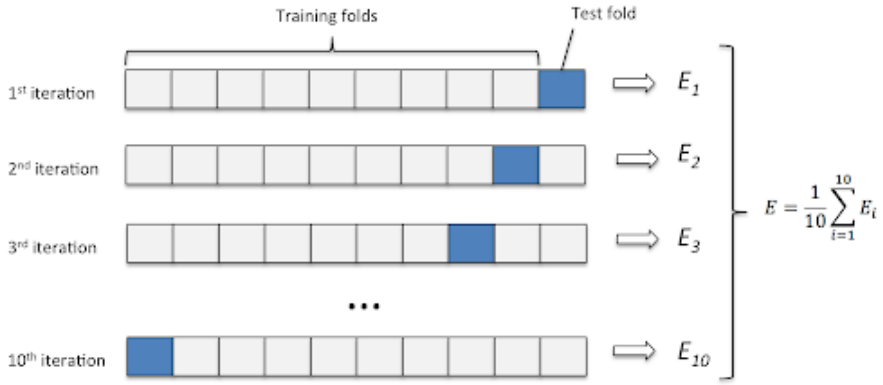
Συμβολικά έχουμε: Έστω T_k, $k = 1, 2, \ldots, K$ οι K φάκελοι με αντίστοιχο μέγεθος στοιχείων n_k (στην περίπτωση που μελετάμε ισχύει ότι $n_k =$ σταθερό και άρα $n_k \cdot K = n$). Ορίζουμε επίσης με $\hat{f}^{-k}(\boldsymbol{x})$ το μοντέλο που έχει προσαρμοστεί σε όλα τα δεδομένα εκτός αυτών που ανήκουν στο k-οστό φάκελο. Το μέσο τετραγωνικό σφάλμα του κάθε φακέλου είναι:

$$MSE(T_k) = \frac{1}{n_k} \sum_{i \in T_k} \left(y_i - \hat{f}^{-k}(\boldsymbol{x}_i) \right)^2 \tag{44}$$

Συνεπώς, συνδυάζοντας τις εκτιμήσεις $MSE(T_k)$ από κάθε φάκελο T_k έχουμε ως εκτίμηση του μέσου τετραγωνικού προβλεπτικού σφάλματος την ποσότητα:

$$CV_K = \overline{MSE} = \frac{1}{K} \sum_{k=1}^{K} MSE(T_k) \tag{45}$$

Ισοδύναμη έκφραση της σχέσης (45) προκύπτει ως εξής: έστω δείκτρια συνάρτηση $\kappa : \{1, 2, \ldots, n\} \to \{1, \ldots, K\}$ που υποδεικνύει σε ποιο διαχωρισμένο



Διάγραμμα 4.2: Σχεδιάγραμμα υλοποίησης μεθόδου 10 fold CV. Σε κάθε επανάληψη $k = 1, \ldots, 10$ υπολογίζεται το τετραγωνικό σφάλμα (E_k) μεταξύ των δεδομένων του τρέχοντα φακέλου και του μοντέλου που προσαρμόζεται στα υπόλοιπα δεδομένα και στο τέλος υπολογίζεται ο μέσος όρος αυτών των σφαλμάτων.

σύνολο δεδομένων ανήκει η i-οστή παρατήρηση. Ως K-fold cross validation εκτίμηση του μέσου τετραγωνικού προβλεπτικού σφάλματος ορίζουμε τότε την ποσότητα:

$$CV_K(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} L\left(y_i, \hat{f}^{-\kappa(i)}(\boldsymbol{x}_i)\right).$$

Εργαζόμενοι με το τετραγωνικό σφάλμα απώλειας καταλήγουμε στην έκφραση:

$$CV_K(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n} \left[y_i - \hat{f}^{-\kappa(i)}(\boldsymbol{x}_i)\right]^2. \tag{46}$$

Οι Breiman, Spector (1992) και Kohari (1995) προτείνουν η μέθοδος να εφαρμόζεται για $K = 5$ ή 10 ως εμπειρικό κανόνα. Στο Διάγραμμα 4.2 απεικονίζεται ο τρόπος με τον οποίο υπολογίζεται το 10-fold cross validation σφάλμα. Στην περίπτωση όπου $K = n$ καταλήγουμε στην $LOOCV$ μέθοδο και στη σχέση (46) έχουμε $\kappa(i) = i$.

Παρατήρηση:

- Για $K = n$ είδαμε ότι ασυμπτωτικά η εκτίμηση του αναμενόμενου προβλεπτικού σφάλματος που προτείνει η $LOOCV$ μέθοδος είναι αμερόληπτη. Ωστόσο το υπολογιστικό κόστος της μεθόδου δεν πρέπει να παραβλέπεται καθώς απαιτούνται n το πλήθος προσαρμογές του υπό εξέταση μοντέλου.

4.3 Generalized cross validation

Η γενικευμένη cross validation μέθοδος (GCV) αποτελεί μια βολική προσέγγιση της leave one out cross validation όταν εφαρμόζεται σε γραμμικές μεθόδους προσαρμογής με τετραγωνικό σφάλμα απώλειας.

Γραμμικές μέθοδοι προσαρμογής:

Μία γραμμική μέθοδος προσαρμογής μοντέλων προκειμένου να προβλέψει τις τιμές της μεταβλητής απόκρισης μετατρέπει τις παρατηρούμενες τιμές έστω $\boldsymbol{y} = (y_1, y_2, \ldots, y_n)$ σε προβλεπόμενες $\hat{\boldsymbol{y}}$ με βάση τη γραμμική σχέση:

$$\hat{\boldsymbol{y}} = \boldsymbol{S}\boldsymbol{y}$$

Η ποσότητα $\boldsymbol{S}$ (linear smoother) είναι ένας $n \times n$ πίνακας που εξαρτάται από τις τιμές των διανυσμάτων $\boldsymbol{x}_i$ των επεξηγηματικών μεταβλητών και όχι από τις τιμές y_i. Στις γραμμικές μεθόδους προσαρμογής εκτός από την περίπτωση της γραμμικής παλινδρόμησης με εκτιμητές της μεθόδου ελαχίστων τετραγώνων, ανήκουν και μέθοδοι εξομάλυνσης όπως είναι η ridge regression και οι cubic smoothing splines (Hastie, Tibshirani, Friedman, 2009).

Ορίζεται ως *αποτελεσματικός αριθμός παραμέτρων* (ή αποτελεσματικοί βαθμοί ελευθερίας) του μοντέλου η ποσότητα:

$$\mathrm{df}(\boldsymbol{S}) = \mathrm{trace}(\boldsymbol{S}) \tag{47}$$

Συμβολίζουμε με $\mathrm{trace}(\boldsymbol{S})$ το άθροισμα των διαγώνιων στοιχείων του πίνακα $\boldsymbol{S}$. Η **GCV** μέθοδος εκτίμησης του προβλεπτικού σφάλματος των γραμμικών μεθόδων προσαρμογής ορίζεται από τον τύπο:

$$GCV(\hat{f}) = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(\boldsymbol{x}_i)}{1 - trace(\boldsymbol{S})/n}\right]^2. \tag{48}$$

Πρόταση 2. *Στις γραμμικές μεθόδους προσαρμογής ισχύει ότι:*

$$\frac{1}{n}\sum_{i=1}^{n}\left[y_i - \hat{f}^{-i}(\boldsymbol{x}_i)\right]^2 = \frac{1}{n}\sum_{i=1}^{n}\left[\frac{y_i - \hat{f}(\boldsymbol{x}_i)}{1 - S_{ii}}\right]^2. \tag{49}$$

Απόδειξη.
Δείχνουμε το ζητούμενο για την περίπτωση της γραμμικής παλινδρόμησης χρησιμοποιώντας το γραμμικό μοντέλο $\boldsymbol{y} = \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\boldsymbol{\varepsilon} \sim N(0, \sigma^2\boldsymbol{I})$. Ο πίνακας $\boldsymbol{S}$ ταυτίζεται με τον πίνακα προβολής $\boldsymbol{H}$ και επομένως θα ισχύει ισοδύναμα ότι $\boldsymbol{S} = \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T$ και $\hat{\boldsymbol{y}} = \hat{\boldsymbol{f}}(\boldsymbol{x}) = \boldsymbol{S}\boldsymbol{y}$ όπως προκύπτει από τη μέθοδο των ελαχίστων τετραγώνων.

Έστω ότι στο διάνυσμα των παρατηρούμενων τιμών $\boldsymbol{y}$ αντικαθιστούμε το i-οστό στοιχείο y_i με τον όρο $\hat{f}^{-i}(\boldsymbol{x}_i)$ που είναι η πρόβλεψη της τιμής y_i από το γραμμικό μοντέλο που προσαρμόστηκε σε δεδομένα που δεν περιλαμβάνουν το ζεύγος $(y_i, \boldsymbol{x}_i)$. Προκύπτει τότε το τροποποιημένο διάνυσμα παρατηρούμενων τιμών $\boldsymbol{y}^*$:

$$\boldsymbol{y}^* = \boldsymbol{y} + \boldsymbol{e}_i(\hat{f}^{-i}(\boldsymbol{x}_i) - y_i).$$

Ο όρος $\boldsymbol{e}_i = (0, 0, \ldots, 1, \ldots, 0)^T$ είναι ένα $n \times 1$ διάνυσμα με μοναδικό μη μηδενικό στοιχείο το 1 στην i-οστή θέση. Η προσαρμογή του γραμμικού μοντέλου στις παρατηρήσεις $(\boldsymbol{y}^*, \boldsymbol{x})$ παράγει τους ίδιους συντελεστές $\hat{\boldsymbol{\beta}}$ με την προσαρμογή του μοντέλου στα δεδομένα χωρίς την παρατήρηση $(y_i, \boldsymbol{x}_i)$ (Hastie, Tibshirani, Friedman, 2009). Επομένως οι συντελεστές $\hat{\boldsymbol{\beta}}^{-i}$ που προκύπτουν από την προσαρμογή στα δεδομένα χωρίς το ζεύγος $(y_i, \boldsymbol{x}_i)$ θα είναι:

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

Το αντίστοιχο διάνυσμα των προβλέψεων θα γράφεται ως:

$$\begin{aligned}
\hat{\boldsymbol{f}}^{(-i)}(\boldsymbol{x}) &= \tilde{\boldsymbol{X}}\hat{\boldsymbol{\beta}}^{(-i)} = \tilde{\boldsymbol{X}}\left[(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}^*\right] \\
&= \tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\left[\boldsymbol{y} + \boldsymbol{e}_i(\hat{f}^{-i}(\boldsymbol{x}_i) - y_i)\right] \\
&= \boldsymbol{S}\left[\boldsymbol{y} + \boldsymbol{e}_i(\hat{f}^{-i}(\boldsymbol{x}_i) - y_i)\right].
\end{aligned} \tag{50}$$

Πολλαπλασιάζοντας τώρα με το διάνυσμα $\boldsymbol{e}_i^T = (0, 0, \ldots, 1\ , \ldots, 0)$ την τελευταία σχέση έχουμε:

$$\begin{aligned}
\hat{f}^{-i}(\boldsymbol{x}_i) &= \boldsymbol{e}_i^T\boldsymbol{S}\left[\boldsymbol{y} + \boldsymbol{e}_i(\hat{f}^{-i}(\boldsymbol{x}_i) - y_i)\right] \\
&= \boldsymbol{e}_i^T\boldsymbol{S}\boldsymbol{y} + S_{ii}(\hat{f}^{-i}(\boldsymbol{x}_i) - y_i).
\end{aligned}$$

Αφαιρούμε την y_i από τα δύο μέλη και παίρνουμε:

$$\begin{aligned} \hat{f}^{-i}(\boldsymbol{x}_i) - y_i &= \boldsymbol{e}_i^T \boldsymbol{S}\boldsymbol{y} + S_{ii}(\hat{f}^{-i}(\boldsymbol{x}_i) - y_i) - y_i \\ \Leftrightarrow (1 - S_{ii})(\hat{f}^{-i}(\boldsymbol{x}_i) - y_i) &= \boldsymbol{e}_i^T \boldsymbol{S}\boldsymbol{y} - y_i. \end{aligned}$$

Με την παρατήρηση τώρα ότι $\boldsymbol{S}\boldsymbol{y} = \hat{\boldsymbol{f}}(\boldsymbol{x})$ και $\boldsymbol{e}_i^T \boldsymbol{S}\boldsymbol{y} = \hat{f}(\boldsymbol{x}_i)$ έχουμε:

$$\begin{aligned} (1 - S_{ii})(\hat{f}^{-i}(\boldsymbol{x}_i) - y_i) &= \hat{f}(\boldsymbol{x}_i) - y_i \\ (\hat{f}^{-i}(\boldsymbol{x}_i) - y_i) &= \left(\frac{\hat{f}(\boldsymbol{x}_i) - y_i}{1 - S_{ii}} \right). \end{aligned}$$

Επομένως, η ζητούμενη σχέση (49) δείχθηκε. Ο τύπος της GCV μεθόδου προκύπτει αν θέσουμε τα διαγώνια στοιχεία S_{ii} ίσα με τον μέσο όρο τους $\frac{\sum_{i=1}^n S_{ii}}{n} = \text{trace}(\boldsymbol{S})/n$.

□

Η **GCV** μέθοδος μπορεί να έχει σημαντικά υπολογιστικά οφέλη σε σχέση με την **leave on out CV** ειδικά όταν το ίχνος του πίνακα $\boldsymbol{S}$ υπολογίζεται πιο εύκολα από τα μεμονωμένα στοιχεία S_{ii}. Στην περίπτωση της γραμμικής παλινδρόμησης ο υπολογισμός αυτός είναι άμεσος καθώς $\text{trace}(\boldsymbol{S}) = \text{trace}(\boldsymbol{H}) = p + 1$ (αριθμός των προς εκτίμηση συντελεστών παλινδρόμησης).

Δείχνουμε τώρα ότι η GCV μέθοδος, το μέτρο καταλληλότητας C_p και το κριτήριο AIC οδηγούν προσεγγιστικά στα ίδια αποτελέσματα στην περίπτωση των γραμμικών μοντέλων παλινδρόμησης.

Η GCV μέθοδος είδαμε ότι έχει τη μορφή:

$$GCV(\hat{f}) = \frac{1}{n} \sum_{i=1}^{n} \left[\frac{y_i - \hat{f}(\boldsymbol{x}_i)}{1 - trace(\boldsymbol{S})/n} \right]^2$$

Καθώς το $trace(\boldsymbol{S}) << n$, χρησιμοποιώντας την προσέγγιση $\frac{1}{(1-x)^2} \approx 1 + 2x$, $|x| < 1$ της γεωμετρικής σειράς $\frac{1}{(1-x)^2}$ έχουμε ότι:

$$\begin{aligned} GCV(\hat{f}) &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(\boldsymbol{x}_i))^2 \left(1 + \frac{2trace(\boldsymbol{S})}{n} \right) \\ &= \frac{1}{n} \sum_{i=1}^{n} (y_i - \hat{f}(\boldsymbol{x}_i))^2 + \frac{2}{n^2} trace(\boldsymbol{S}) \sum_{i=1}^{n} (y_i - \hat{f}(\boldsymbol{x}_i))^2. \end{aligned}$$

Ο πρώτος όρος στο δεξί μέλος ισούται με την ποσότητα SSE_p και το ίχνος του πίνακα $\boldsymbol{S}$ είναι ο αποτελεσματικός αριθμός των παραμέτρων με $trace(\boldsymbol{S}) = p + 1 = p^{'}$. Θεωρούμε επιπλέον ότι η διασπορά σ^2 του **πλήρους** μοντέλου εκτιμάται προσεγγιστικά από την ποσότητα:

$$\hat{\sigma}^2 \approx \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{f}(\boldsymbol{x}_i))^2$$

Επομένως, η GCV μέθοδος παίρνει τη μορφή:

$$GCV(\hat{f}) = \frac{SSE_p}{n} + \frac{2}{n}p'\hat{\sigma}^2 \tag{51}$$

Η σχέση (51) είναι ισοδύναμη με τη δεύτερη μορφή του κριτηρίου C_p.

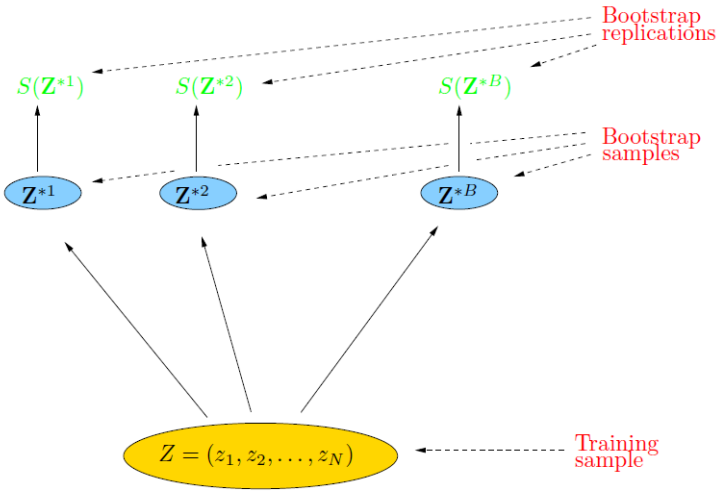
Σημείωση: Ο Attilgan (1996) αποδεικνύει ότι στη γραμμική παλινδρόμηση τα κριτήρια AIC και C_p είναι ισοδύναμα με την έννοια ότι προτείνουν το ίδιο καταλληλότερο μοντέλο και κατατάσσουν με την ίδια σειρά τα υποψήφια μοντέλα.

Αναμένουμε λοιπόν εφαρμόζοντας στη γραμμική παλινδρόμηση τις μεθόδους AIC, C_p και GCV να οδηγηθούμε στην ίδια επιλογή μοντέλων.

4.4 Bootstrap μέθοδος

Μία δεύτερη υπολογιστική μέθοδος αξιολόγησης των μοντέλων παλινδρόμησης είναι η μέθοδος επαναδειγματοληψίας Bootstrap (Efron,Tibshirani 1993). Αρχικά περιγράφουμε τη γενική μεθοδολογία και στη συνέχεια βλέπουμε πως εφαρμόζεται στην εκτίμηση του προβλεπτικού σφάλματος των μοντέλων.

Ας υποθέσουμε ότι διαθέτουμε ένα σύνολο από N δεδομένα που το συμβολίζουμε με $\boldsymbol{Z} = (z_1, z_2, \ldots, z_N)$ όπου $z_i = (\boldsymbol{x}_i, y_i)$, και ένα στατιστικό μοντέλο προς εξέταση. Η βασική ιδέα της μεθόδου είναι η δημιουργία νέων δειγμάτων διαλέγοντας τυχαία με επανατοποθέτηση από τις ήδη υπάρχουσες παρατηρήσεις $\boldsymbol{Z}$. Τα νέα δείγματα που κατασκευάζονται είναι B το πλήθος και το μέγεθος του καθενός συμπίπτει με το μέγεθος N του αρχικού δείγματος $\boldsymbol{Z}$. Σε κάθε ένα από τα B δείγματα προσαρμόζουμε το μοντέλο και υπολογίζουμε το στατιστικό αξιολόγησης με το οποίο εργαζόμαστε, έστω $S(\boldsymbol{Z}^{*B})$. Στο τέλος παίρνουμε τον μέσο όρο του στατιστικού $S(\boldsymbol{Z})$ των B δειγμάτων. Η παραπάνω διαδικασία αναπαρίσταται ποιοτικά στο Διάγραμμα 4.3.



Διάγραμμα 4.3: Από το αρχικό δείγμα (training sample) $\boldsymbol{Z}$ δημιουργούμε B τυχαία δείγματα με δειγματοληψία επανάθεσης, μεγέθους n το καθένα. Σε κάθε ένα από τα $\boldsymbol{Z}^{*1}, \boldsymbol{Z}^{*2}, \ldots, \boldsymbol{Z}^{*B}$ *Bootstrap* δείγματα υπολογίζεται η στατιστική συνάρτηση $S(\boldsymbol{Z}^{*i})$ για την αξιολόγηση του μοντέλου.

Παρατηρήσεις:

- Για την επιλογή των παρατηρήσεων σε κάθε *bootstrap* δείγμα χρησιμοποιείται η εμπειρική κατανομή των αρχικών δεδομένων, η οποία δίνει $\frac{1}{N}$ πιθανότητα σε κάθε παρατήρηση του αρχικού δείγματος να επιλεχθεί στο *bootstrap* δείγμα και 0 αλλού.

- Ένα δείγμα *bootstrap* μπορεί να περιέχει κάποια ή κάποιες τιμές z_i του αρχικού δείγματος περισσότερες από μία φορές ή να μην τις περιέχει καθόλου.

- Η βασική προϋπόθεση της μεθόδου είναι ότι θεωρούμε πως η εμπειρική κατανομή αποτελεί καλή προσέγγιση της κατανομής του χαρακτηριστικού με το οποίο αξιολογούμε τα μοντέλα. Επομένως αν π.χ. το μέγεθος του αρχικού δείγματος N είναι μικρό, ενδέχεται η μέθοδος να μην αποδώσει.

Η γενική *bootstrap* μεθοδολογία που περιγράφτηκε παραπάνω εφαρμόζεται σε πολλές πτυχές της στατιστικών αναλύσεων όπως είναι η εκτίμηση τυπικών σφαλμάτων, οι έλεγχοι υποθέσεων και η παλινδρόμηση. Επικεντρωνόμαστε στην περίπτωση των μοντέλων παλινδρόμησης όπου η μέθοδος προσφέρει μια μορφή αξιολόγησης της προβλεπτικής ικανότητας των μοντέλων.

4.4.1 Bootstrap εκτίμηση προβλεπτικού σφάλματος

Θέλουμε να εκτιμήσουμε το προβλεπτικό σφάλμα του μοντέλου παλινδρόμησης $y_i = f(\boldsymbol{x}_i) + \varepsilon_i$ όπου $\varepsilon_i \sim N(0, \sigma^2)$, διαθέτοντας τυχαίο δείγμα από παρατηρήσεις $(y_i, \boldsymbol{x}_i), i = 1, \ldots, N$. Δημιουργούμε από το αρχικό δείγμα B *bootstrap* δείγματα και σε κάθε ένα από αυτά προσαρμόζουμε το μοντέλο. Συμβολίζοντας με $\hat{f}^{*b}(\boldsymbol{x}_i)$ την προβλεπόμενη τιμή της y_i παρατήρησης από το προσαρμοσμένο στο b *bootstrap* δείγμα μοντέλο, ορίζεται ως *bootstrap* εκτίμηση του προβλεπτικού σφάλματος του μοντέλου η ποσότητα:

$$\hat{Err}_{boot} = \frac{1}{B}\frac{1}{N}\sum_{b=1}^{B}\sum_{i=1}^{N} L(y_i, \hat{f}^{*b}(\boldsymbol{x}_i)). \tag{52}$$

Συμβολίζουμε με $L(,)$ τη συνάρτηση απώλειας της πρόβλεψης της παρατήρησης y_i με συνηθισμένη επιλογή την $L(y_i, \hat{f}^{*b}(\boldsymbol{x}_i)) = (y_i - \hat{f}^{*b}(\boldsymbol{x}_i))^2$.

Εν γένει, ο απλός εκτιμητής $\hat{Err}_{boot}$ της σχέσης (52) δεν είναι πάντα αξιόπιστος, διότι είναι αρκετά πιθανό πολλές παρατηρήσεις του αρχικού δείγματος να περιέχονται εις διπλούν, τριπλούν, κ.ο.κ. στα B δείγματα που προσαρμόζεται το μοντέλο. Οι πολλές αλληλοκαλυπτόμενες παρατηρήσεις οδηγούν σε υπερβολικά αισιόδοξες προβλέψεις $\hat{f}^{*b}(\boldsymbol{x}_i)$ μεροληπτώντας έτσι προς τα κάτω το προβλεπτικό σφάλμα του μοντέλου.

Μία καλύτερη *bootstrap* εκτίμηση του σφάλματος προκύπτει ακολουθώντας το σκεπτικό της *cross validation* μεθόδου. Για κάθε παρατήρηση y_i κρατάμε στην τελική εκτίμηση του σφάλματος τις εκτιμήσεις των *bootstrap* δειγμάτων που δεν περιέχουν αυτή την παρατήρηση. Καλούμε τη νέα εκτίμηση leave one out bootstrap εκτίμηση του προβλεπτικού σφάλματος η οποία ορίζεται ως:

$$\hat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(\boldsymbol{x}_i)). \tag{53}$$

Ο όρος C^{-i} συμβολίζει το σύνολο των δεικτών των *bootstrap* δειγμάτων που δεν περιέχουν την παρατήρηση i και $|C^{-i}|$ είναι ο αντίστοιχος αριθμός-πλήθος αυτών των δειγμάτων. Προκειμένου να διασφαλιστεί ότι όλα τα $|C^{-i}| > 0$ πρέπει να επιλέξουμε αρκετά μεγάλο αριθμό B κατασκευασμένων *bootstrap* δειγμάτων ή να αγνοήσουμε στους υπολογισμούς μας όρους που αντιστοιχούν σε μηδενικά $|C^{-i}|$.

Η $\hat{Err}^{(1)}$ εκτίμηση λύνει το πρόβλημα της υπερπροσαρμοστικότητας που

αντιμετωπίζει η απλή εκτίμηση $\hat{Err}_{boot}$, αλλά μεροληπτεί προς τα πάνω λόγω του μειωμένου αριθμού των όρων $L(y_i, \hat{f}^{*b}(\boldsymbol{x}_i))$ που λαμβάνονται υπόψιν.

Μελετάμε τώρα την πιθανότητα η m παρατήρηση του αρχικού δείγματος να ανήκει στο b *bootstrap* δείγμα μέσω απλού αριθμητικού παραδείγματος: Έστω $S = (s_1, s_2, \ldots, s_n)$ δείγμα από n αριθμούς που έχουν προκύψει από ανεξάρτητες τυχαίες επιλογές αριθμών με επανάθεση απο το σύνολο $\{1 : n\}$. Θεωρούμε δείκτη $m \in \{1 : n\}$ και τότε έχουμε:

$$P\left(s_i = m\right) = \frac{1}{n} \quad \text{και} \quad p\left(s_i \neq m\right) = 1 - 1/n \quad \forall\ 1 \leq i \leq n.$$

Επομένως προκύπτει:

$$\begin{aligned} P(m \in S) &= 1 - P(m \notin S) = 1 - P(\cap_{i=1}^{n} s_i \neq m) \\ &= 1 - \prod_{i=1}^{n} P(s_i \neq m) = 1 - (1 - 1/n)^n \simeq 1 - e^{-1} \quad \forall n \geq 11. \end{aligned}$$

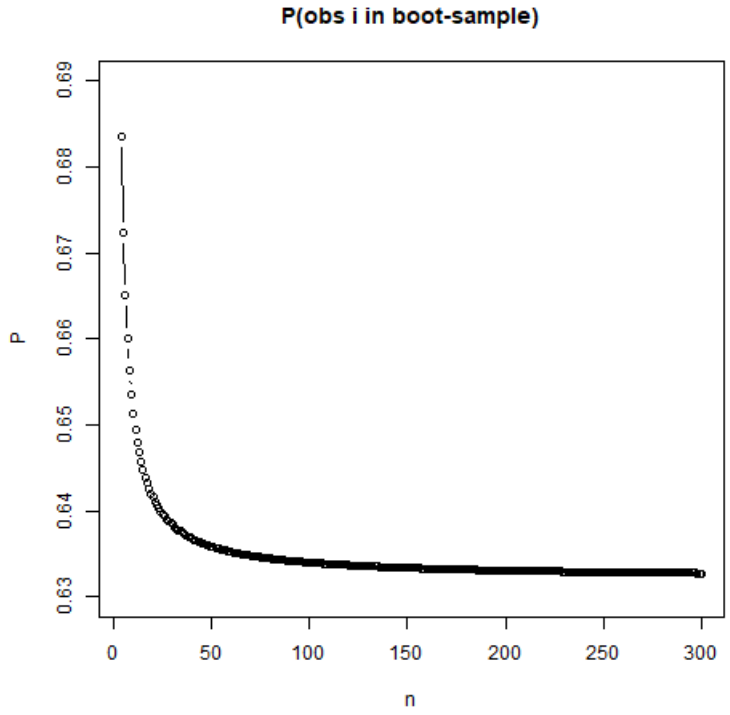
Η πιθανότητα λοιπόν η i παρατήρηση να περιέχεται στο b *bootsrap* δείγμα είναι $\simeq 1 - e^{-1} = 0.632$ όπως φαίνεται και από το Διάγραμμα 4.4, και σε κάθε b δείγμα αναμένεται να περιέχονται κατά μέσο όρο $0.632N$ διακριτές παρατηρήσεις από το αρχικό δείγμα.

Προκειμένου να αντιμετωπιστεί η προς τα πάνω μεροληψία της $\hat{Err}^{(1)}$ εκτίμησης, οι Efron, Tibshirani (1997) πρότειναν τον .632 εκτιμητή που ορίζεται ως:

$$\hat{Err}^{.632} = 0.368 \cdot \overline{err} + 0.632 \cdot \hat{Err}^{(1)}. \tag{54}$$

Με $\overline{err}$ συμβολίζουμε την «αφελή» εκτίμηση $\overline{err} = \frac{1}{N}\sum_{i=1}^{N} L(y_i, \hat{f}(\boldsymbol{x}_i))$ του προβλεπτικού σφάλματος που στηρίζεται μόνο στα αρχικά δεδομένα. Η προέλευση του $\hat{Err}^{(0.632)}$ εκτιμητή είναι περίπλοκη. Διαισθητικά ισοσταθμίζει κατά μέσο όρο τον μεροληπτικό προς τα κάτω εκτιμητή $\overline{err}$ με τον όρο 0.368 και τον προς τα πάνω εκτιμητή $\hat{Err}^{(1)}$ με τον όρο .632. Ο όρος .632 συνδέεται με την πιθανότητα της i παρατήρησης να βρίσκεται στο b δείγμα όπως ορίστηκε προηγουμένως.

Ο εκτιμητής $\hat{Err}^{.632}$ οδηγεί σε καλές εκτιμήσεις του προβλεπτικού σφάλματος όταν τα μοντέλα που εξετάζονται δεν υπερπροσαρμόζονται στα αρχικά δεδομένα. Αντιθέτως, σε περιπτώσεις υπερπροσαρμογής όπου $\overline{err} \approx 0$, ακόμα και ο εκτιμητής $\hat{Err}^{(.632)}$ μεροληπτεί προς τα κάτω.



Διάγραμμα 4.4: Διάγραμμα της πιθανότητας να ανήκει η i-οστή παρατήρηση των δεδομένων στο δείγμα *bootstrap* για διάφορα μεγέθη δείγματος (n). Για μεγέθη δείγματος ≥ 200 η πιθανότητα αυτή ισούται προσεγγιστικά με τον αριθμό 0.632.

4.5 Χρήση εκτιμητών προβλεπτικού σφάλματος σε δεδομένα

Προχωράμε εφαρμόζοντας τις μεθόδους εκτίμησης του προβλεπτικού σφάλματος του εν λόγω κεφαλαίου σε προσομοιωμένα δεδομένα. Θεωρούμε ότι το πραγματικό μοντέλο που «γεννάει» τα δεδομένα έχει τη μορφή:

$$\boldsymbol{y} = \beta_0 1_n + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2 \mathbf{1}).$$

Το μέγεθος του δείγματος είναι $n = 100$ και θεωρούμε ότι η μεταβλητή απόκρισης Y συνδέεται συναρτησιακά μόνο με μία επεξηγηματική μεταβλητή την X, η οποία θεωρούμε ότι ακολουθεί την ομοιόμορφη κατανομή στο διάστημα $[0, 20]$ ($X \sim U[0, 20]$). Το διάνυσμα $\boldsymbol{\varepsilon}$ ακολουθεί την πολυδιάστατη κανονική κατανομή με $\boldsymbol{\varepsilon} \sim N_{100}(\mathbf{0}_{100}, \sigma^2 \mathbf{1}_n), \sigma^2 = 350$. Με 1_n συμβολίζεται το μοναδιαίο διάνυσμα και $\mathbf{1}_n$ είναι ο $n \times n$ μοναδιαίος πίνακας. Το διάνυσμα $\boldsymbol{\beta}$ αποτελείται από τα στοιχεία: $\boldsymbol{\beta} = (120, -12, 0.4)^T$ και $\beta_0 = 100$. Ο πίνακας $\boldsymbol{X}$ έχει διαστάσεις

100×3 και δεσμεύοντας στον $\boldsymbol{X}$ από αριστερά τη μοναδιαία στήλη $(1, \ldots, 1)^T$ και στο διάνυσμα $\boldsymbol{\beta}$ τον σταθερό όρο β_0,το μοντέλου από όπου προσομοιώνονται τα δεδομένα υπό μορφή πινάκων γράφεται:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & x_1^2 & x_1^3 \\ 1 & x_2 & x_2^2 & x_2^3 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & x_n^3 \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Η γενική μορφή του μοντέλου είναι:

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon.$$

Ισοδύναμα για κάθε παρατήρηση $i = 1, \ldots, n$ θα ισχύει ότι:

$$y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \beta_3 x_i^3 + \varepsilon_i.$$

Μοντέλα που έχουν την παραπάνω μορφή καλούνται πολυωνυμικά μοντέλα παλινδρόμησης με μία επεξηγηματική μεταβλητή X. Αντιμετωπίζοντας τους όρους X, X^2, X^3 ως ξεχωριστές επεξηγηματικές μεταβλητές βλέπουμε ότι καταλήγουμε σε ένα γραμμικό μοντέλο του οποίου η λύση με τη μέθοδο ελαχίστων τετραγώνων ξέρουμε ότι είναι $\hat{\boldsymbol{\beta}} = (\tilde{\boldsymbol{X}}^T \tilde{\boldsymbol{X}})^{-1} \tilde{\boldsymbol{X}}^T \boldsymbol{y}$.

Το Διάγραμμα 4.5 είναι το διάγραμμα διασποράς των παρατηρήσεων $(y, \boldsymbol{x}_i)$ που προκύπτουν από την προσομοίωση 100 ζευγών δεδομένων.

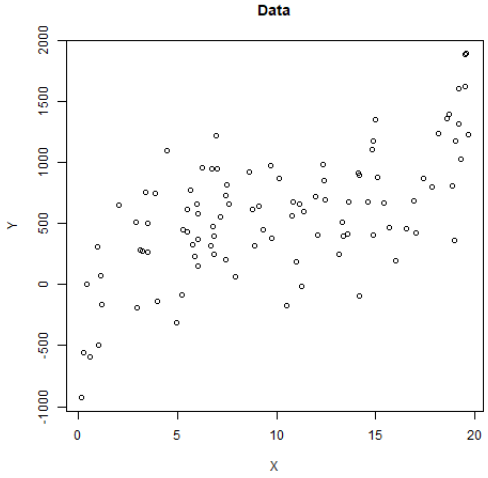
Σκοπός μας είναι να προσαρμόσουμε και να επιλέξουμε ένα μοντέλο παλινδρόμησης που θα προβλέπει αποτελεσματικά την μεταβλητή απόκρισης Y για μελλοντικές παρατηρήσεις της μεταβλητής X.

Τα μοντέλα που χρησιμοποιούμε έχουν τη γενική μορφή:

$$\begin{aligned} &Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m + \varepsilon \\ &\text{και} \quad \mathbb{E}[Y|X] = \beta_0 + \beta_1 X + \beta_2 X^2 + \cdots + \beta_m X^m. \end{aligned}$$

Το τυχαίο σφάλμα ε ακολουθεί την κανονική κατανομή $N(0, \sigma^2)$. Τα πολυωνυμικά μοντέλα που προσαρμόζουμε με αύξουσα σειρά βαθμού m είναι:

$$\begin{aligned} M_1 \quad & Y = \beta_0 + \beta_1 X + \varepsilon \\ M_2 \quad & Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \varepsilon \\ M_3 \quad & Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon \end{aligned}$$



Διάγραμμα 4.5: 100 προσομοιωμένα ζεύγη παρατηρήσεων από το πραγματικό μοντέλο. Παρατηρούμε ότι η σχέση που συνδέει την μεταβλητή Y με την X δεν είναι γραμμική και ενδείκνυται για αυτό το λόγο η χρήση πολυωνυμικών μοντέλων.

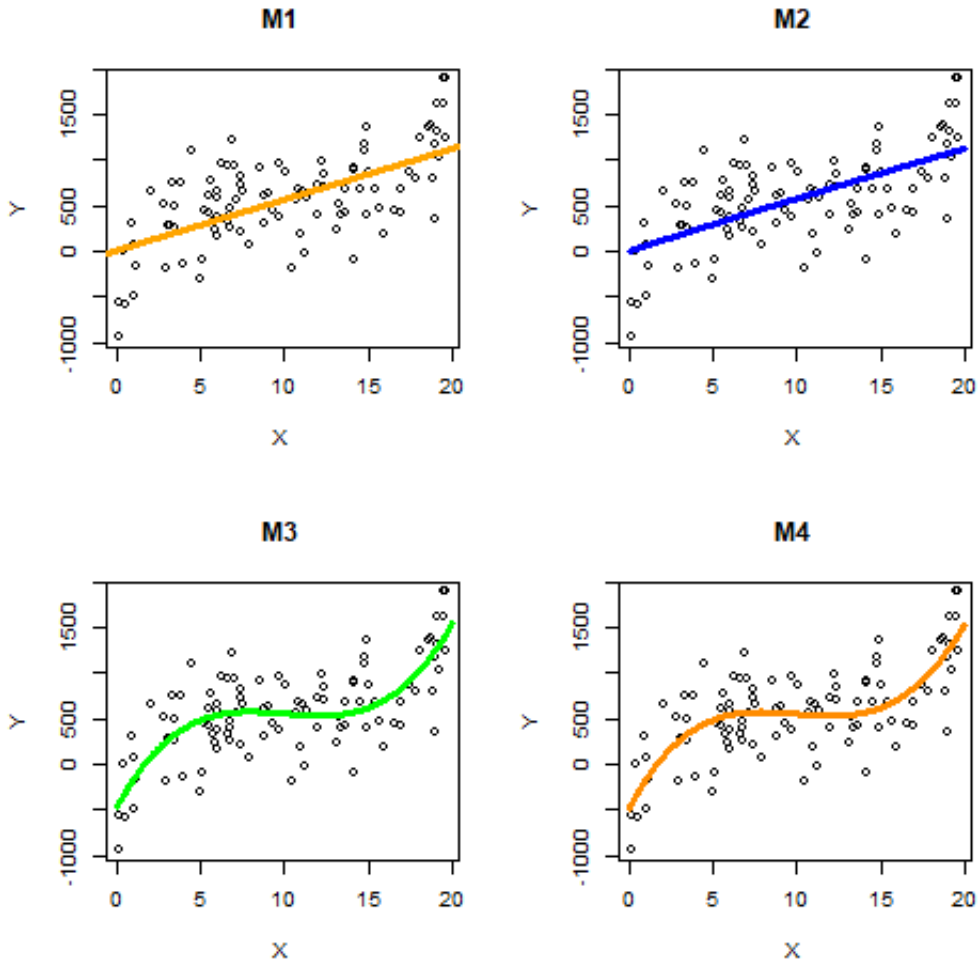
$$
\begin{aligned}
M_4 \quad & Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \varepsilon \\
M_5 \quad & Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \varepsilon \\
M_6 \quad & Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6 + \varepsilon \\
M_7 \quad & Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \beta_4 X^4 + \beta_5 X^5 + \beta_6 X^6 + \beta_7 X^7 + \varepsilon.
\end{aligned}
$$

Σημειώνουμε ότι εφόσον η μεταβλητή απόκρισης Y είναι συνάρτηση μίας μόνο επεξηγηματικής μεταβλητής X, είναι δυνατή η αναπαράσταση σε ένα δυσδιάστατο γράφημα των παρατηρούμενων τιμών (y_i, x_i) και των αναμενόμενων κατά μέσο όρο τιμών $\mathbb{E}[Y|X = x]$ που δίνει κάθε προσαρμοσμένο πολυωνυμικό μοντέλο. Τα Διαγράμματα 4.6 και 4.7 απεικονίζουν την προσαρμογή των επτά πολυωνυμικών μοντέλων στα προσομοιωμένα δεδομένα.

Παρατηρούμε ότι το απλό γραμμικό μοντέλο M_1 αλλά και το πολυωνυμικό μοντέλο M_2 δεν προσαρμόζονται ικανοποιητικά στα δεδομένα μας. Τα μοντέλα M_3, M_4 προσαρμόζονται πολύ ικανοποιητικά και από το μοντέλο M_5 και μετά υπάρχουν ενδείξεις υπερπροσαρμογής (overfitting).

Στη συνέχεια υπολογίζουμε με βάση τα αρχικά μας δεδομένα εκτιμήσεις του προβλεπτικού σφάλματος που προκύπτουν από τις μεθόδους:

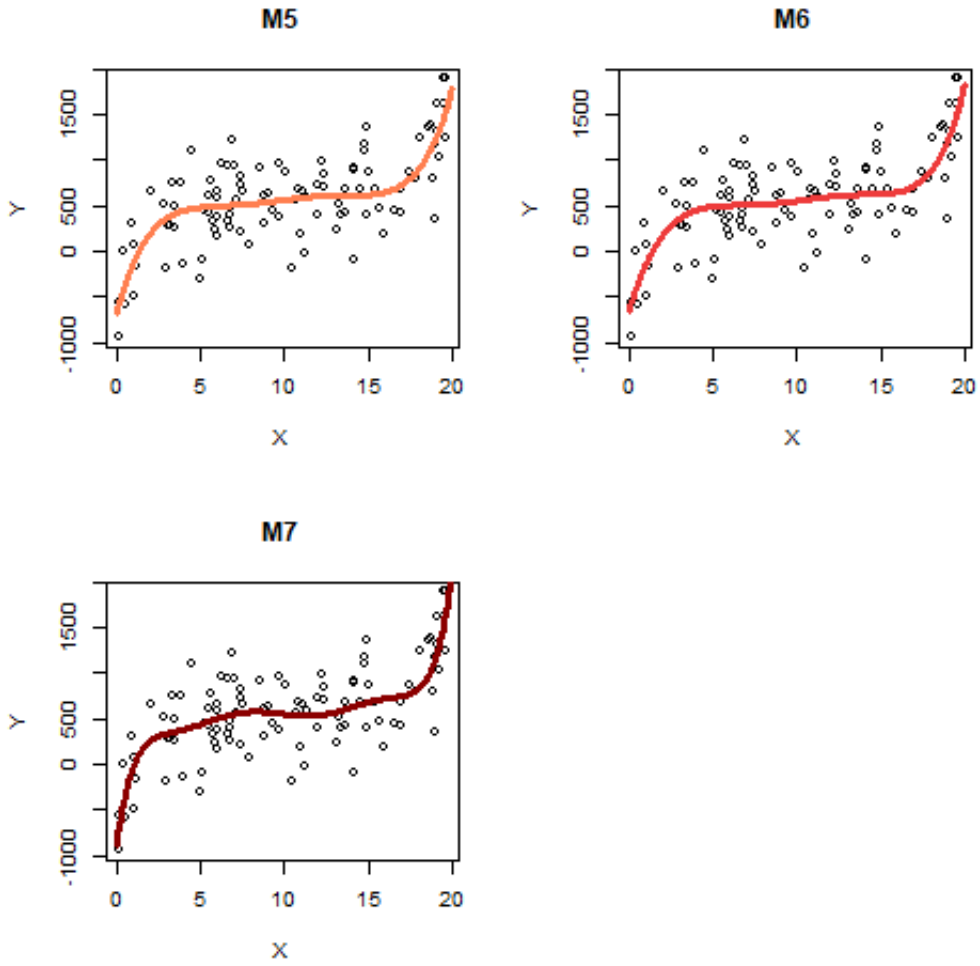
- μέσο τετραγωνικό σφάλμα MSE= $\frac{\sum_{i=1}^{n}(y_i - \hat{f}(x_i))^2}{n}$,



Διάγραμμα 4.6: Προσαρμογή των μοντέλων M_1, M_2, M_3, M_4 στα αρχικά δεδομένα.

- leave one out cross validation,
- generalized cross validation,
- 10-fold cross validation,
- leave one out bootstrap $\hat{Err}^{(1)} = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{|C^{-i}|}\sum_{b\in C^{-i}} L(y_i, \hat{f}^{*b}(\boldsymbol{x}_i))$,
- .632 bootstrap $\hat{Err}^{.632} = 0.368 \cdot \overline{err} + 0.632 \cdot \hat{Err}^{(1)}$.

Οι τιμές των εκτιμητών εκτελώντας τον κώδικα του παραρτήματος **B** στο περιβάλλον της R για κάθε μοντέλο δίνονται στον Πίνακα 4.2.



Διάγραμμα 4.7: Προσαρμογή των μοντέλων M_5, M_6, M_7 στα αρχικά δεδομένα.

Παρατηρήσεις:

1. Από τα αποτελέσματα του Πίνακα 4.2 παρατηρούμε ότι όσο αυξάνεται ο βαθμός των πολυωνυμικών μοντέλων ελαττώνεται το μέσο τετραγωνικό σφάλμα MSE. Αυτό συμβαίνει διότι προσθέτοντας περισσότερες μεταβλητές σε γραμμικά μοντέλα προκαλείται μία τάση βελτίωσης του MSE, το οποίο επιλέγει ως καταλληλότερο το μοντέλο M_7.

2. Η μέθοδος leave one out CV προτείνει ως μοντέλο με μικρότερο προβλεπτικό σφάλμα το μοντέλο M_5 και μετά το μοντέλο M_3 με μικρές διαφορές στις τιμές τους. Η GCV μέθοδος προτείνει ως καταλληλότερο το μοντέλο M_3 και μετά το M_5. Παρατηρούμε επίσης την ασυμπτωτική ισοδυναμία της

$\times 10^{-4}$	M_1	M_2	M_3	M_4	M_5	M_6	M_7
MSE	15.39	15.39	11.85	11.85	11.43	11.42	**11.12**
LooCV	16.12	16.61	12.94	13.25	**12.93**	13.25	13.22
GCV	16.02	16.35	**12.86**	13.13	12.93	13.20	13.14
10foldCV	14.95	18.68	13.41	13.89	**12.16**	13.12	13.96
$\hat{Err}^{(1)}$	16.49	17.11	**13.40**	13.98	13.70	14.16	14.57
$\hat{Err}^{.632}$	16.09	16.47	**12.83**	13.20	12.86	13.15	13.30

Πίνακας 4.2: Εκτιμήσεις του προβλεπτικού σφάλματος εφαρμόζοντας τις cross validation και bootstrap υπολογιστικές μεθόδους στα αρχικά δεδομένα.

μεθόδου με την LooCV λόγω των πολύ κοντινών τιμών τους. Η 10-fold CV μέθοδος προτείνει το μοντέλο M_5 με αμέσως επόμενο το M_6.

3. Οι μέθοδοι Leave one out bootstrap και .632 bootstrap επιλέγουν ως καταλληλότερο το μοντέλο M_3.

Από τα παραπάνω βλέπουμε ότι ορισμένες φορές η τάση επιλογής υπερπροσαρμοσμένων (overfitted) μοντέλων υπάρχει. Αυτό το αποτέλεσμα ωστόσο είναι αναμενόμενο καθώς οι εκτιμήσεις αυτές είναι δειγματοσυναρτήσεις και επομένως εμπεριέχουν σφάλμα στην εκτίμηση του πραγματικού προβλεπτικού σφάλματος των μοντέλων. Επίσης, οι εν λόγω υπολογιστικές μέθοδοι δεν λαμβάνουν υπόψιν την πολυπλοκότητα των μοντέλων (με εξαίρεση την GCV). Σκοπός τους βέβαια δεν είναι καθαρά η επιλογή ενός μοντέλου με μικρό προβλεπτικό σφάλμα **και** με τις λιγότερες δυνατές παραμέτρους-μεταβλητές.

Όταν όμως υπάρχει ανάγκη εύρεσης μοντέλου με όσο το δυνατόν χαμηλότερο σφάλμα πρόβλεψης οι μέθοδοι αυτές δίνουν ακριβείς εκτιμήσεις, φαινόμενο που μπορεί να επαληθευτεί με τη χρήση προσομοιωμένων δεδομένων.

Προσομοιώνουμε λοιπόν στη συνέχεια 100 νέα δείγματα παρατηρήσεων ανεξάρτητα του αρχικού (έστω τα $T_j = \{(y'_1, x'_1), (y'_2, x'_2), \ldots, (y'_n, x'_n)\}, j = 1, \ldots, 100$) και σε κάθε μοντέλο $M_z, z = 1, \ldots, 7$ και για κάθε δείγμα T_j υπολογίζουμε την ποσότητα:

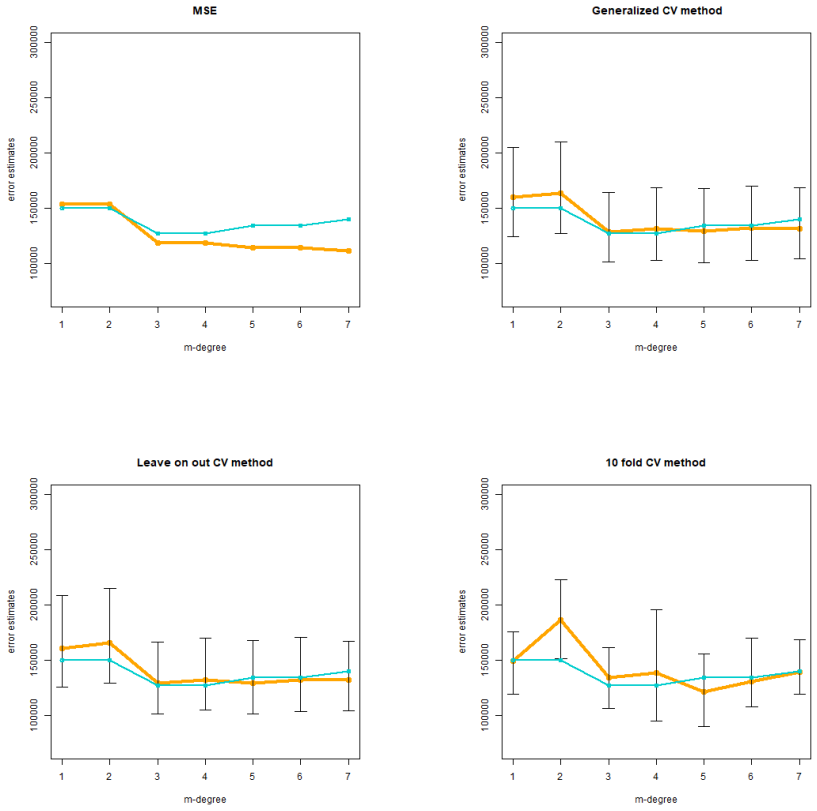
$$SampError[j]_z = \frac{1}{n}\sum_{i=1}^{n}(y'_i - \hat{f}_z(x'_i))^2.$$

Συμβολίζουμε με $\hat{f}_z(x'_i)$ την προβλεπόμενη τιμή του M_z μοντέλου για την παρατήρηση y'_i το οποίο όμως προσαρμόστηκε στα αρχικά δεδομένα. Τέλος, θεωρούμε ως **πραγματικό** προβλεπτικό σφάλμα των μοντέλων την ποσότητα:

$$TrueError_z = \frac{1}{100}\sum_{j=1}^{100} SampError[j]_z. \tag{55}$$

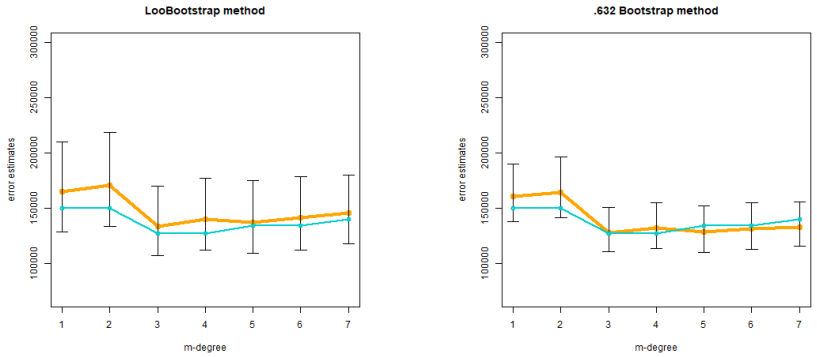
Στα Διαγράμματα 4.8 και 4.9 απεικονίζονται οι εκτιμήσεις της κάθε μεθόδου και τα πραγματικά σφάλματα των μοντέλων όπως προκύπτουν από τη σχέση (55). Τα σημεία με πορτοκαλί χρώμα απεικονίζουν τις εκτιμήσεις του προβλεπτικού σφάλματος των μεθόδων, ενώ με μπλε χρώμα απεικονίζεται το πραγματικό σφάλμα των μοντέλων.

Τα διαστήματα που φαίνονται με μαύρο χρώμα είναι BC_a διαστήματα εμπιστοσύνης των σφαλμάτων των υπολογιστικών μεθόδων και ο υπολογισμός τους περιγράφεται αναλυτικά στο παράρτημα **A**.



Διάγραμμα 4.8: Εκτιμήσεις MSE, GCV, LooCV,και 10-foldCV συναρτήσει του πραγματικού προβλεπτικού σφάλματος των πολυωνυμικών μοντέλων.

Βλέπουμε αρχικά ότι το μέσο τετραγωνικό σφάλμα ως εκτιμητής του πραγματικού προβλεπτικού σφάλματος μεροληπτεί σε μεγάλο βαθμό προς τα κάτω



Διάγραμμα 4.9: Εκτιμήσεις LooBootstrap και .632 bootsrap συναρτήσει του πραγματικού προβλεπτικού σφάλματος των πολυωνυμικών μοντέλων.

ενώ παράλληλα το πραγματικό σφάλμα δείχνει μία μικρή τάση αύξησης από το πολυωνυμικό μοντέλο βαθμού 5 και μετά.

Δεύτερον, από τις τρεις **cross validation** μεθόδους η leave one out CV και η GCV προσεγγίζουν σε πολύ ικανοποιητικό βαθμό το πραγματικό προβλεπτικό σφάλμα των μοντέλων, ενώ η 10-fold CV φαίνεται να είναι η πιο ασταθής.

Από τις **bootstrap** μεθόδους ο .632 bootstrap εκτιμητής φαίνεται να έχει τη μεγαλύτερη ακρίβεια, ενώ η leave one out bootstrap μέθοδος παρουσιάζει μια μικρή μεροληψία προς τα πάνω στην εκτίμηση του προβλεπτικού σφάλματος.

Εν γένει μπορούν να προκύψουν **φειδωλότερα** μοντέλα από τα υποψήφια εάν εφαρμόσουμε τον κανόνα «**one standard error rule**». Εργαζόμενοι παραδείγματος χάριν με την 10-fold CV μέθοδο ο κανόνας χρησιμοποιείται ως εξής:

1. Βρίσκουμε το μοντέλο στο οποίο παρατηρείται το ελάχιστο 10-fold CV σφάλμα έστω $Err10CV_{min}$, και το τυπικό του σφάλμα $se(Err10CV_{min}) = sd(Err10CV_{min})/\sqrt{10}$, όπου $sd(.)$ η τυπική απόκλιση του εκτιμητή.

2. «Κινούμενοι» προς τα μοντέλα μικρότερης διάστασης επιλέγουμε ως βέλτιστο το μοντέλο βαθμού m για το οποίο εξακολουθεί να ισχύει η συνθήκη: $Err10CV(m) \leq Err10CV_{min} + se(Err10CV_{min})$ μέχρι να μην είναι αληθής.

Εφαρμόζοντας τον κανόνα, οι επιλογές που μεταβάλλονται προκύπτουν ως εξής:

$$LooCV(m) \leq 12.93 + 1.65 = 14.58$$
$$Err10CV(m) \leq 12.16 + 1.76 = 13.92$$

Επομένως, οι μέθοδοι LooCV, 10foldCV επιλέγουν ως τελικό μοντέλο το M_3.

Παρατήρηση: Αυξάνοντας περαιτέρω τον βαθμό των πολυωνυμικών μοντέλων προκύπτουν προβλήματα πολυσυγγραμικότητας με συνέπεια να μην μπορούν να προσδιοριστούν πολλοί από τους συντελεστές $\beta_j, j = 1 \ldots, p$ και η εκτίμηση του προβλεπτικού τους σφάλματος να καθίσταται αδύνατη. Στο κεφάλαιο που έπεται εξηγούμαι αρχικά το συγκεκριμένο πρόβλημα και στη συνέχεια εξετάζουμε μία μέθοδο που αντιμετωπίζει αποτελεσματικά τέτοια συχνά φαινόμενα.

Κεφάλαιο 5

Στο μεγαλύτερο κομμάτι της εργασίας η εκτίμηση των παραμέτρων των μοντέλων παλινδρόμησης γίνεται με τη μέθοδο ελαχίστων τετραγώνων, ελαχιστοποιώντας το τετραγωνικό σφάλμα των υπολοίπων. Τα αποτελέσματα της μεθόδου δεν είναι πάντα ικανοποιητικά κυρίως για δύο λόγους. Ο πρώτος είναι η προβλεπτική ακρίβεια των μοντέλων. Είναι ευρέως γνωστό στη βιβλιογραφία ότι οι εκτιμήτριες των ελαχίστων τετραγώνων στο γραμμικό μοντέλο είναι βέλτιστες αμερόληπτες εκτιμήτριες των συντελεστών $\boldsymbol{\beta}$ του μοντέλου με την έννοια ότι $\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] = \boldsymbol{\beta}$ και η διασπορά τους $Var(\hat{\boldsymbol{\beta}}_{ols})$ είναι η ελάχιστη δυνατή στο σύνολο των αμερόληπτων εκτιμητριών. Ωστόσο, όπως θα δούμε παρακάτω πολλές φορές ο όρος $Var(\hat{\boldsymbol{\beta}}_{ols})$ αποκτά μεγάλες τιμές και υπερισχύει του όρου της μεροληψίας αυξάνοντας σημαντικά με αυτό τον τρόπο το προβλεπτικό σφάλμα του μοντέλου.

Ο δεύτερος λόγος αναποτελεσματικότητας της μεθόδου ελαχίστων τετραγώνων συνδέεται με την εξαγωγή συμπερασμάτων ύστερα από την προσαρμογή των μοντέλων. Όταν έχουμε ένα μεγάλο το πλήθος σύνολο από υποψήφιες επεξηγηματικές μεταβλητές (predictors), επιδιώκουμε συνήθως να κρατήσουμε ένα υποσύνολο αυτών με τις ισχυρότερες επιδράσεις στη μεταβλητή απόκρισης. Η μέθοδος ελαχίστων τετραγώνων δεν διαχωρίζει με κάποιο αυτόματο τρόπο τις μεταβλητές αυτές στο τελικό μοντέλο.

Στο παρόν κεφάλαιο εισάγουμε τη μέθοδο συρρίκνωσης lasso (least absolute shrinkage and selection operator) η οποία συρρικνώνει κάποιους συντελεστές των μοντέλων και άλλους τους εξισώνει με το μηδέν, με στόχο την βελτίωση της προβλεπτικής ικανότητας και της αποτελεσματικής επιλογής επεξηγηματικών μεταβλητών. Η μέθοδος στηρίζεται στην ιδέα ότι θυσιάζοντας ένα μικρό μέρος της αμεροληψίας των εκτιμητών των συντελεστών $\boldsymbol{\beta}$, μειώνεται σημαντικά η διασπορά τους σε βαθμό όπου βελτιώνεται η προβλεπτική ικανότητα των μοντέλων. Τα βασικά προτερήματα της μεθόδου είναι κυρίως ότι αντιμετωπίζει σε μεγάλο βαθμό το πρόβλημα της πολυσυγγραμικότητας το οποίο αναλύεται παρακάτω, και ότι πραγματοποιεί με αυτόματο τρόπο αποτελεσματική επιλογή μεταβλητών (variable selection) οδηγώντας σε φειδωλά μοντέλα.

5.1 Πολυσυγγραμικότητα

Στη θεωρία του πολλαπλού γραμμικού μοντέλου είδαμε ότι προκειμένου η μέθοδος ελαχίστων τετραγώνων (αλλά και η μέθοδος μεγιστοποιημένης πιθανοφάνειας) να εκτιμήσει τους συντελεστές $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$, πρέπει ο πίνακας σχεδιασμού $\tilde{\boldsymbol{X}}_{n\times p+1}$ να είναι πλήρους τάξης, δηλαδή τα $p+1$ διανύσματα-στήλες του να μην είναι γραμμικά εξαρτημένα. Έστω $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ τα διανύσματα-στήλες του πίνακα σχεδιασμού. Τα $\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_p$ είναι γραμμικώς εξαρτημένα εάν υπάρχουν σταθερές $\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_p$, όχι όλες μηδέν, έτσι ώστε:

$$\alpha_0\boldsymbol{x}_0 + \alpha_1\boldsymbol{x}_1 + \alpha_2\boldsymbol{x}_2 + \cdots + \alpha_p\boldsymbol{x}_p = 0.$$

Εάν τουλάχιστον δύο διανύσματα-στήλες $\boldsymbol{x}_i, \boldsymbol{x}_j$ του πίνακα σχεδιασμού $\tilde{\boldsymbol{X}}$ είναι γραμμικά εξαρτημένα, τότε η τάξη του πίνακα θα είναι μικρότερη από p και ο πίνακας $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ δε θα ορίζεται από τη θεωρία γραμμικής άλγεβρας πινάκων. Σε αυτή την περίπτωση λοιπόν η εκτιμήτρια ελαχίστων τετραγώνων $\hat{\boldsymbol{\beta}}_{ols} = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}$ θα είναι μη υπολογίσιμη.

Όταν τα διανύσματα-στήλες του πίνακα $\tilde{\boldsymbol{X}}$ είναι γραμμικά εξαρτημένα (αγνοώντας την πρώτη στήλη που αντιστοιχεί στο σταθερό όρο του μοντέλου), λέμε ότι οι αντίστοιχες επεξηγηματικές μεταβλητές $X_1, X_2, \ldots, X_p$ είναι γραμμικά συσχετισμένες και έχουμε να αντιμετωπίσουμε το πρόβλημα της «τέλειας» πολυσυγγραμικότητας. Στην πράξη στις εφαρμογές σχεδόν ποτέ δεν συναντάμε γραμμικά συσχετισμένες επεξηγηματικές μεταβλητές. Ωστόσο, όταν ορισμένες στήλες του πίνακα $\tilde{\boldsymbol{X}}$ είναι σχεδόν γραμμικά εξαρτημένες ($\alpha_0\boldsymbol{x}_0+\alpha_1\boldsymbol{x}_1+\alpha_2\boldsymbol{x}_2+\cdots+\alpha_p\boldsymbol{x}_p \approx 0$), ισοδύναμα η ορίζουσα του πίνακα $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ είναι κοντά στο μηδέν και επομένως κάποιες ιδιοτιμές προσεγγίζουν το μηδέν. Αυτό το φαινόμενο καλείται ασθενής πολυσυγγραμικότητα.

Εξετάζουμε τώρα το μέσο τετραγωνικό σφάλμα της εκτιμήτριας $\hat{\boldsymbol{\beta}}_{ols}$ της μεθόδου ελαχίστων τετραγώνων στο γραμμικό μοντέλο, τη συμπεριφορά του οποίου θα συνδέσουμε με το φαινόμενο της πολυσυγγραμικότητας:

$$\begin{aligned}
MSE(\hat{\boldsymbol{\beta}}_{ols}) &= \mathbb{E}[(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})] \\
&= \mathbb{E}\left[(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] + \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] + \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] + \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] - \boldsymbol{\beta})\right] \\
&= \mathbb{E}\Big[(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])^T(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}]) + (\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])^T(\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] - \boldsymbol{\beta}) \\
&\quad + (\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] + \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}]) + (\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] + \boldsymbol{\beta})^T(\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] - \boldsymbol{\beta})\Big]
\end{aligned}$$

$$
\begin{aligned}
&= \mathbb{E}\Big[(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])^T(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])\Big] + \mathbb{E}\Big[(\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] + \boldsymbol{\beta})^T(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])\Big] \\
&+ \mathbb{E}\Big[(\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] + \boldsymbol{\beta})^T(\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] - \boldsymbol{\beta})\Big] \\
&= \mathbb{E}\Big[(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])^T(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])\Big] + \mathbb{E}\Big[2\boldsymbol{\beta}^T(\hat{\boldsymbol{\beta}}_{ols} - \boldsymbol{\beta})\Big] \\
&= \mathbb{E}\Big[(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])^T(\hat{\boldsymbol{\beta}}_{ols} - \mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}])\Big].
\end{aligned}
$$

Στα παραπάνω χρησιμοποιήσαμε το γεγονός ότι $\mathbb{E}[\hat{\boldsymbol{\beta}}_{ols}] = \boldsymbol{\beta}$. Ισοδύναμα, από την ιδιότητα ότι για κάθε διάνυσμα $\boldsymbol{x} = (x_1, \ldots, x_p) \in \mathbb{R}^p : \ \boldsymbol{x}^T\boldsymbol{x} = \sum_{j=1}^p x_j^2$ η έκφραση του $MSE(\hat{\boldsymbol{\beta}}_{ols})$ παίρνε τη μορφή:

$$
\begin{aligned}
MSE(\hat{\boldsymbol{\beta}}_{ols}) &= \mathbb{E}\left[\sum_{j=1}^{p}(\hat{\beta}_{ols,j} - \mathbb{E}[\hat{\beta}_{ols,j}])^2\right] = \sum_{j=1}^{p}\mathbb{E}\left[(\hat{\beta}_{ols,j} - \mathbb{E}[\hat{\beta}_{ols,j}])^2\right] \\
&= \sum_{j=1}^{p} Var(\hat{\beta}_{ols,j}) \\
&= trace(Var(\hat{\boldsymbol{\beta}}_{ols})).
\end{aligned} \tag{56}
$$

Για την διασπορά της εκτιμήτριας $\hat{\boldsymbol{\beta}}_{ols}$ έχουμε:

$$
\begin{aligned}
Var(\hat{\boldsymbol{\beta}}_{ols}) &= Var\left((\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\boldsymbol{y}\right) = (\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\sigma^2\boldsymbol{I}_{n\times n}\left[(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\right]^T \\
&= \sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1} \\
&= \sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}.
\end{aligned}
$$

Εκφράζοντας τον πίνακα $(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}$ με τη βοήθεια του προσαρτημένου πίνακα $adj(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})$ προκύπτει ότι:

$$
Var(\hat{\boldsymbol{\beta}}_{ols}) = \sigma^2\frac{1}{det(\tilde{\boldsymbol{X}}\tilde{\boldsymbol{X}})} \cdot adj(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}) \tag{57}
$$

Βλέπουμε από την σχέση (57) ότι όσο πλησιάζει στο μηδέν η ορίζουσα του πίνακα $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$, οι τιμές του πίνακα διασποράς $Var(\hat{\boldsymbol{\beta}}_{ols})$ γίνονται ακραία μεγάλες και κατά επέκταση από τη σχέση (56) το αντίστοιχο αναμενόμενο μέσο

τετραγωνικό σφάλμα της εκτιμήτριας $\hat{\boldsymbol{\beta}}_{ols}$ αποκτά μεγάλες τιμές.

Κατά το φαινόμενο της ασθενούς πολυσυγγραμικότητας οι μεγάλες διασπορές που εμφανίζουν οι εκτιμήτριες ελαχίστων τετραγώνων οδηγούν αντίστοιχα σε μεγάλα τυπικά σφάλματα με αποτέλεσμα να γίνεται δύσκολο να αναδειχθούν οι επεξηγηματικές μεταβλητές που είναι πραγματικά σημαντικές και συνεισφέρουν στην εξήγηση της διασποράς της μεταβλητής απόκρισης Y.

Μια απλή προσέγγιση αντιμετώπισης του φαινομένου είναι για κάθε μία από τις $i = 1, \ldots, p$ επεξηγηματικές μεταβλητές να προσαρμοστεί το γραμμικό μοντέλο με μεταβλητή απόκρισης την i-οστή επεξηγηματική μεταβλητή και επεξηγηματικές μεταβλητές τις υπόλοιπες $p-1$ επεξηγηματικές μεταβλητές. Έπειτα, από τα μοντέλα που εμφανίζουν υψηλό συντελεστή προσδιορισμού ($R^2 > 0.90$) θα πρέπει να εξετασθεί εάν η αφαίρεσή τους από το αρχικό μοντέλο με μεταβλητή απόκρισης την Y λύνει το πρόβλημα της πολυσυγγραμικότητας.

Ωστόσο, η διαδικασία αυτή μπορεί να αποβεί χρονοβόρα και υπολογιστικά έντονη για αυτό προτιμάμε την μέθοδο *lasso*.

5.2 Γενική μέθοδος

Έστω ότι διαθέτουμε n παρατηρήσεις $(\boldsymbol{x}_i, y_i)$ και θέλουμε να εκτιμήσουμε τις παραμέτρους $(\beta_o, \boldsymbol{\beta})$ των συντελεστών του γραμμικού μοντέλου $Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p + \varepsilon,\ \varepsilon \sim N(0, \sigma^2)$, με προσαρμοσμένη ευθεία παλινδρόμησης $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{i1} + \cdots + \hat{\beta}_p x_{ip} = \hat{\beta}_0 + \sum_{j=1}^{p} x_{ij}\hat{\beta}_j$. Η μέθοδος **lasso** (least absolute shrinkage operator) εκτιμάει τους συντελεστές ελαχιστοποιώντας όπως και στη μέθοδο ελαχίστων τετραγώνων την αντικειμενική συνάρτηση:

$$\underset{\beta_0, \boldsymbol{\beta}}{\text{minimize}} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (y_i - \beta_0 - \sum_{j=1}^{p} x_{ij}\beta_j)^2 \right\} \tag{58}$$

Προστίθεται ωστόσο ο περιορισμός:

$$\sum_{j=1}^{p} |\beta_j| \leq t \tag{59}$$

Θέτοντας $\boldsymbol{y} = (y_1, \ldots, y_n)^T$ το διάνυσμα των παρατηρούμενων τιμών της μεταβλητής απόκρισης Y, $\tilde{\boldsymbol{X}}$ τον $n \times p$ πίνακα με στήλες-διανύσματα (χωρίς τη στήλη του σταθερού όρου του μοντέλου) τις τιμές των επεξηγηματικών μεταβλητών $\boldsymbol{x}_j,\ j = 1, \ldots, p$ και χρησιμοποιώντας την l_1 νόρμα και την Ευκλείδεια

νόρμα διανυσμάτων l_2, προκύπτει η ισοδύναμη συμπαγής μορφή:

$$\begin{aligned} &\underset{\beta_0,\boldsymbol{\beta}}{\text{minimize}} \left\{ \frac{1}{2n} \left\| \boldsymbol{y} - \beta_0 \mathbf{1} - \tilde{\boldsymbol{X}}\boldsymbol{\beta} \right\|_2^2 \right\}, \\ &\text{με περιορισμό } \|\boldsymbol{\beta}\|_1 \leq t \end{aligned} \tag{60}$$

Με $\mathbf{1}$ συμβολίζεται το μοναδιαίο διάνυσμα διάστασης n. Το φράγμα $t > 0$ περιορίζει το άθροισμα των απολύτων τιμών των εκτιμητών $\boldsymbol{\beta}$, συρρικνώνοντας με αυτό τον τρόπο προς το μηδέν ορισμένες παραμέτρους που παίρνουν ακραίες τιμές, και οδηγώντας σε πιο περιορισμένα μοντέλα.

Εάν τα δεδομένα των επεξηγηματικών μεταβλητών $x_{i1}, \ldots, x_{ip}$ δεν έχουν την ίδια μονάδα μέτρησης χρειάζονται κανονικοποίηση ούτως ώστε για κάθε στήλη j του πίνακα σχεδιασμού $\tilde{\boldsymbol{X}}$ να ισχύει $\frac{1}{n}\sum_{i=1}^{n} x_{ij} = 0$ και $\frac{1}{n}\sum_{i=1}^{n} x_{ij}^2 = 1$. Αυτό γίνεται αφαιρώντας την μέση τιμή σε κάθε διάνυσμα επεξηγηματικής μεταβλητής και διαιρώντας με την αντίστοιχη (μη διορθωμένη) τυπική απόκλιση $\sqrt{\frac{\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}{n}}$. Τον κανονικοποιημένο πίνακα σχεδιασμού τον συμβολίζουμε με $\boldsymbol{X}$ και αυτόν χρησιμοποιούμε μέχρι το τέλος του Κεφαλαίου 5.

Κεντράρουμε επιπλέον τις παρατηρήσεις της μεταβλητής απόκρισης Y αφαιρώντας το μέσο τους όρο ούτως ώστε $\frac{1}{n}\sum_{i=1}^{n} y_i = 0$. Με αυτό τον τρόπο μπορούμε να απαλείψουμε το σταθερό όρο β_0. Το πρόβλημα βελτιστοποίησης χρησιμοποιώντας πλέον τις κανονικοποιημένες και κεντραρισμένες παρατηρήσεις θα παίρνει τη μορφή:

$$\begin{aligned} &\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \right\}, \\ &\text{με περιορισμό } \|\boldsymbol{\beta}\|_1 \leq t. \end{aligned} \tag{61}$$

Η ισοδύναμη *Lagrange* μορφή της σχέσης (61) είναι:

$$\underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda \|\boldsymbol{\beta}\|_1 \right\} \tag{62}$$

Ο όρος $\lambda > 0$ καλείται παράμετρος ποινικοποίησης. Η ελαχιστοποίηση της σχέσης (61) είναι ισοδύναμη με την ελαχιστοποίηση της (62). Αυτό σημαίνει ότι για κάθε τιμή της παραμέτρου t στο εύρος τιμών που ικανοποιείται ο περιορισμός $\|\boldsymbol{\beta}\|_1 \leq t$, υπάρχει αντίστοιχη τιμή της παραμέτρου λ η οποία οδηγεί στην ίδια λύση.

Πρόταση 3. *Δοσμένης της λύσης, έστω $\hat{\boldsymbol{\beta}}$, από τη μέθοδο lasso στα κεντρα-*

ρισμένα και κανονικοποιημένα δεδομένα μπορούμε να εξάγουμε την αντίστοιχη λύση για τα αρχικά δεδομένα. Συγκεκριμένα:

$$\hat{\beta}_{0,initial} = \bar{y} - \sum_{j=1}^{p} \bar{x}_j \hat{\beta}_j, \quad \hat{\beta}_{j,initial} = \frac{\hat{\beta}_j}{S_j}$$

Οι μέσοι όροι $\bar{y}, \bar{x}_j$ να υπολογίζονται στα αρχικά δεδομένα.

Απόδειξη. Θεωρούμε τα κεντραρισμένα δεδομένα $Y_{center} = Y - \bar{y}$ και τα κανονικοποιημένα $z_j = (x_j - \bar{x}_j)/S_j, j = 1, \ldots, p$ όπου S_j η τυπική απόκλιση των δεδομένων της j επεξηγηματικής μεταβλητής. Εφαρμόζοντας τη μέθοδο *lasso* προκύπτει η εκτιμώμενη ευθεία παλινδρόμησης:

$$\begin{aligned} \hat{Y}_{center} &= \hat{\beta}_0 + \sum_{j=1}^{p} \frac{(x_j - \bar{x}_j)\hat{\beta}_j}{S_j} \\ &= \hat{\beta}_0 - \sum_{j=1}^{p} \frac{\bar{x}_j \hat{\beta}_j}{S_j} + \sum_{j=1}^{p} \frac{x_j \hat{\beta}_j}{S_j} \end{aligned}$$

Επιστρέφοντας στις αρχικές τιμές έχουμε:

$$\hat{Y} = \hat{Y}_{center} + \bar{y} = \hat{\beta}_0 + \bar{y} - \sum_{j=1}^{p} \frac{\bar{x}_j \hat{\beta}_j}{S_j} + \sum_{j=1}^{p} \frac{x_j \hat{\beta}_j}{S_j}$$

Ο σταθερός όρος ενός γραμμικού μοντέλου αποτελεί μέτρο της μέσης τιμής της μεταβλητής απόκρισης όταν οι επεξηγηματικές μεταβλητές είναι μηδέν. Στην περίπτωσή μας ο όρος $\hat{\beta}_0$ έχει προκύψει από την παλινδρόμηση των κεντραρισμένων δεδομένων επομένως επειδή $\bar{Y}_{center} = 0 \to \hat{\beta}_0 = 0$ και καταλήγουμε στη σχέση:

$$\hat{Y} = \bar{y} - \sum_{j=1}^{p} \frac{\bar{x}_j \hat{\beta}_j}{S_j} + \sum_{j=1}^{p} \frac{x_j \hat{\beta}_j}{S_j}$$

Βλέπουμε λοιπόν ότι ο σταθερός όρος του γραμμικού μοντέλου που χρησιμοποιείται για την πρόβλεψη τιμών με βάση την κλίμακα των αρχικών δεδομένων ισούται με $\bar{y} - \sum_{j=1}^{p} \frac{\bar{x}_j \hat{\beta}_j}{S_j}$ και οι p συντελεστές ισούνται με $\frac{\hat{\beta}_j}{S_j}$. □

Η εύρεση του διανύσματος $\boldsymbol{\beta} \in \mathbb{R}^p$ που ελαχιστοποιεί τη συνάρτηση της σχέσης (62) αποτελεί πρόβλημα τετραγωνικού προγραμματισμού με κυρτή αλλά μη διαφορίσιμη αντικειμενική συνάρτηση. Σε αντίθεση με την μέθοδο ελαχίστων τετραγώνων, δεν μπορούμε απευθείας να παραγωγίσουμε την αντικειμενική συνάρτηση ως προς το διάνυσμα $\boldsymbol{\beta}$ για να βρούμε το σημείο ελαχίστου, διότι η

παράγωγος του μέρους $\lambda \sum_{j=1}^{p} |\beta_j|$ δεν ορίζεται στο 0. Παραθέτουμε στη συνέχεια χρήσιμους ορισμούς και λήμματα που χρησιμοποιούνται μετέπειτα στην επίλυση της μεθόδου:

Ορισμός 1 (Κυρτό σύνολο)**.** *Ένα σύνολο $C \subseteq \mathbb{R}^p$ είναι κυρτό εάν για κάθε $\boldsymbol{\beta}, \boldsymbol{\beta}' \in C$ και για κάθε $s \in [0,1]$, όλα τα διανύσματα της μορφής $s\boldsymbol{\beta} + (1-s)\boldsymbol{\beta}'$ ανήκουν στο C.*

Ορισμός 2 (Κυρτή συνάρτηση)**.** *Μία βαθμωτή συνάρτηση $f : C \to \mathbb{R}$, $C \subseteq \mathbb{R}^p$ με p μεταβλητές καλείται κυρτή εάν το πεδίο ορισμού της C είναι κυρτό σύνολο και για κάθε $\boldsymbol{\beta}, \boldsymbol{\beta}' \in C$, $s \in (0,1)$ ισχύει ότι: $f(s\boldsymbol{\beta} + (1-s)\boldsymbol{\beta}') \leq sf(\boldsymbol{\beta}) + (1-s)f(\boldsymbol{\beta}')$.*

Πρόταση 4. *Ένα τοπικό ελάχιστο μίας κυρτής συνάρτησης ορισμένης σύμφωνα με τον ορισμό 2 είναι πάντα ολικό ελάχιστο.*

Απόδειξη. Έστω $f : S \to \mathbb{R}$ κυρτή συνάρτηση με κυρτό πεδίο ορισμού και x τοπικό ελάχιστο της f. Ισοδύναμα θα υπάρχει ανοιχτό σύνολο-γειτονιά U του x με $f(x) \leq f(u)$ για κάθε $u \in U$. Θα δείξουμε ότι $f(x) \leq f(y)\ \forall y \in S$:

Για $0 \leq t \leq 1$ και καθώς το $t \to 0$, ο συνδυασμός $ty + (1-t)x$ θα ανήκει στη γειτονιά U του x και τότε θα έχουμε:

$$\begin{aligned} f(x) &\leq f\left(ty + (1-t)x\right) \\ &\leq tf(y) + (1-t)f(x), \quad \text{αφού } f \text{ κυρτή} \end{aligned}$$

από όπου προκύπτει ότι: $f(x) \leq f(y)$. □

Ορισμός 3 (Υποδιαφορικό συνάρτησης)**.** *Έστω $C \subseteq \mathbb{R}^p$ κυρτό σύνολο και $f : C \to \mathbb{R}$ κυρτή συνάρτηση. Για κάθε $\boldsymbol{\beta} \in C$ το υποδιαφορικό της f στο $\boldsymbol{\beta}$ είναι το σύνολο των $z \in \mathbb{R}^p$ για τα οποία ισχύει:*

$$f(\boldsymbol{\beta}') \geq f(\boldsymbol{\beta}) + \langle z, \boldsymbol{\beta}' - \boldsymbol{\beta} \rangle \quad \forall \boldsymbol{\beta}' \in \mathbb{R}^p$$

Με $\langle ., . \rangle$ συμβολίζουμε το εσωτερικό γινόμενο διανυσμάτων και για το υποδιαφορικό γράφουμε:

$$\partial f(\boldsymbol{\beta}) := \left\{ z \in \mathbb{R}^p : \ f(\boldsymbol{\beta}') \geq f(\beta) + \langle z, \boldsymbol{\beta}' - \boldsymbol{\beta} \rangle, \ \boldsymbol{\beta}' \in \mathbb{R}^p \right\}$$

Λήμμα 1. *(Ελάχιστο μη διαφορίσιμης συνάρτησης). Το σημείο $\boldsymbol{\beta}^* \in C$ αποτελεί ελάχιστο της συνάρτησης $f : C \to \mathbb{R}$ (όχι απαραίτητα κυρτής) αν και μόνο αν η f είναι υποδιαφορίσιμη*[3] *στο σημείο $\boldsymbol{\beta}^*$ και:*

$$0 \in \partial f(\boldsymbol{\beta}^*)$$

Λήμμα 2. *(Κυρτότητα και Εσσιανός πίνακας). Μία διπλά διαφορίσιμη συνάρτηση $f : C \to \mathbb{R},\ C \in \mathbb{R}^p$ είναι κυρτή αν και μόνο αν ο Εσσιανός πίνακας $\boldsymbol{H}_{p\times p}$ είναι θετικά ημιορισμένος για κάθε $\boldsymbol{\beta} \in C$.*

Σημείωση: Ο πίνακας $\boldsymbol{H}_{p\times p}$ είναι θετικά ημιορισμένος αν για κάθε διάνυσμα $\boldsymbol{w} \in \mathbb{R}^p$ ισχύει ότι: $\boldsymbol{w}^T\boldsymbol{H}\boldsymbol{w} \geq 0$.

Είμαστε πλέον σε θέση να αποδείξουμε ότι η συνάρτηση ελαχιστοποίησης της μεθόδου *lasso* είναι κυρτή. Γράφουμε αρχικά τη συνάρτηση στη μορφή:

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

Εάν δείξουμε ότι οι επί μέρους όροι $f(\boldsymbol{\beta}), g(\boldsymbol{\beta})$ είναι κυρτές συναρτήσεις, τότε από θεμελιώδη ιδιότητα του αθροίσματος κυρτών συναρτήσεων και η $L(\boldsymbol{\beta})$ θα είναι κυρτή.

- Η $f(\boldsymbol{\beta})$ είναι διπλά διαφορίσιμη για κάθε $\boldsymbol{\beta} \in \mathbb{R}^p$ και ο Εσσιανός της πίνακας είναι:

$$\begin{aligned}\nabla^2 f(\boldsymbol{\beta}) = \frac{\partial^2 f(\boldsymbol{\beta})}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} &= \frac{\partial^2\left[\frac{1}{2n}[\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}]\right]}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}^T} \\ &= 2\boldsymbol{X}^T\boldsymbol{X}\frac{1}{2n} = \frac{\boldsymbol{X}^T\boldsymbol{X}}{n}\end{aligned}$$

  Για κάθε διάνυσμα $\boldsymbol{z}_{p\times 1} \in \mathbb{R}^p$ έχουμε:

$$n^{-1}\boldsymbol{z}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{z} = n^{-1}(\boldsymbol{X}\boldsymbol{z})^T\boldsymbol{X}\boldsymbol{z} = n^{-1}\|\boldsymbol{X}\boldsymbol{z}\|_2^2 \geq 0$$

  Συνεπώς ο πίνακας $\nabla^2 f(\boldsymbol{\beta})$ είναι θετικά ημιορισμένος και από το Λήμμα 2 η συνάρτηση $f(\boldsymbol{\beta}) = \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2$ είναι κυρτή.

[3]Μία συνάρτηση είναι υποδιαφορίσιμη στο $\boldsymbol{\beta}^*$ εάν υπάρχει τουλάχιστον ένα $z \in \partial f(\boldsymbol{\beta}^*)$. Στην περίπτωση που η f είναι κυρτή και διαφορίσιμη, η συνθήκη του λήμματος $0 \in \partial f(\boldsymbol{\beta}^*)$ μετατρέπεται στην $\nabla f(\boldsymbol{\beta}^*) = 0$.

- Για τη συνάρτηση $g : C \to \mathbb{R},\ C \subseteq \mathbb{R}^{'}$ και για κάθε $\boldsymbol{\beta}, \boldsymbol{\beta}^{'} \in C,\ s \in (0,1)$ έχουμε:

$$g\left(s\boldsymbol{\beta} + (1-s)\boldsymbol{\beta}^{'}\right) = \left\|s\boldsymbol{\beta} + (1-s)\boldsymbol{\beta}^{'}\right\|_1 \leq \|s\boldsymbol{\beta}\|_1 + \left\|(1-s)\boldsymbol{\beta}^{'}\right\|_1$$
$$\leq s\|\boldsymbol{\beta}\|_1 + (1-s)\left\|\boldsymbol{\beta}^{'}\right\|_1 = sg(\boldsymbol{\beta}) + (1-s)g(\boldsymbol{\beta}^{'})$$

  Επομένως, από τον ορισμό της κυρτής συνάρτησης η $g(\boldsymbol{\beta})$ είναι κυρτή. Η $L(\boldsymbol{\beta})$ λοιπόν ως άθροισμα κυρτών συναρτήσεων είναι κυρτή στο $C \subseteq \mathbb{R}^p$.

Επιστρέφοντας στο αρχικό πρόβλημα, αναζητούμε λύση $\hat{\boldsymbol{\beta}}_L \in \arg\min_{\boldsymbol{\beta}\in\mathbb{R}^p} L(\boldsymbol{\beta})$. Το λήμμα 1 αποτελεί ικανή συνθήκη για την εύρεση του ολικού ελαχίστου της $L(\boldsymbol{\beta})$ και η ισοδύναμη μορφή της συνθήκης $0 \in \partial L(\hat{\boldsymbol{\beta}}_L)$ είναι:

$0 \in \partial f(\hat{\boldsymbol{\beta}}_L) + \lambda \sum_{j=1}^{p} \partial g_j(\hat{\boldsymbol{\beta}}_L)$ (1^{η} Karush-Kuhn-Tucker συνθήκη),

με $f(\hat{\boldsymbol{\beta}}_L) = \left\|\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right\|_2^2,\ g(\hat{\boldsymbol{\beta}}_L) = \left\|\hat{\boldsymbol{\beta}}_L\right\|_1$ και υπάρχει $\boldsymbol{w} \in \partial g(\hat{\boldsymbol{\beta}}_L)$ με:

$$\boldsymbol{w} = \begin{cases} w_j = 1, & \hat{\beta}_{Lj} > 0 \\ w_j = -1, & \hat{\beta}_{Lj} < 0 \\ w_j \in [-1,1], & \hat{\beta}_{Lj} = 0 \end{cases}$$

Επομένως για την ελαχιστοποίηση της $L(\boldsymbol{\beta})$ προκύπτει η συνθήκη:

$$\nabla f(\hat{\boldsymbol{\beta}}_L) + \lambda \boldsymbol{w} = 0 \tag{63}$$

5.2.1 Περίπτωση μίας επεξηγηματικής μεταβλητής

Μελετάμε αρχικά την περίπτωση μίας ανεξάρτητης μεταβλητής X βασιζόμενοι στα δεδομένα $(x_i, y_i), i = 1, \ldots, n$, τα οποία είναι κανονικοποιημένα και κεντραρισμένα αντίστοιχα. Η συνάρτηση ελαχιστοποίησης $L(\beta)$ θα γράφεται ως:

$$\underset{\beta\in\mathbb{R}}{\text{minimize}} \left\{ \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}_1\beta\|_2^2 + \lambda|\beta| \right\}, \boldsymbol{X}_1 = (x_1, x_2, \ldots, x_n)^T$$

Θεωρώντας $f(\beta) = \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}_1\beta\|_2^2$ προκύπτει ότι:

$$\begin{aligned}\frac{\partial f(\beta)}{\partial \beta} &= \frac{\partial\left(\frac{1}{2n}[\boldsymbol{y}^T\boldsymbol{y} - 2\beta\boldsymbol{X}_1^T\boldsymbol{y} + \beta^2\boldsymbol{X}_1^T\boldsymbol{X}_1]\right)}{\partial\beta} \\ &= -\frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} + \frac{\beta\boldsymbol{X}_1^T\boldsymbol{X}_1}{n}\end{aligned}$$

Εφαρμόζοντας τη σχέση (63) έχουμε:

$$\begin{aligned}-\frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} + \frac{\hat{\beta}_L\boldsymbol{X}_1^T\boldsymbol{X}_1}{n} + \lambda w = 0 \\ \frac{\hat{\beta}_L\boldsymbol{X}_1^T\boldsymbol{X}_1}{n} = \frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} - \lambda w\end{aligned} \tag{64}$$

Με $\boldsymbol{X}_1^T\boldsymbol{X}_1 = n$ αφού η μεταβλητή X είναι κανονικοποιημένη και $\hat{\beta}_{ols} = (\boldsymbol{X}_1^T\boldsymbol{X}_1)^{-1}\boldsymbol{X}_1^T\boldsymbol{y} = \frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n}$ έχουμε πλέον ότι:

$$\hat{\beta}_L = \hat{\beta}_{ols} - \lambda w, \tag{65}$$

$$w = \begin{cases} 1 & ,\hat{\beta}_L > 0 \\ -1 & ,\hat{\beta}_L < 0 \\ [-1,1] & ,\hat{\beta}_L = 0 \end{cases}$$

Διακρίνουμε τώρα τις εξής περιπτώσεις:

- Εάν $\hat{\beta}_L > 0$ τότε από την (65) $w = 1$ και $\frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} > \lambda$.
  Συνεπώς $\hat{\beta}_L = \frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} - \lambda$.

- Αν $\hat{\beta}_L < 0$ τότε $w = -1$ και $\frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} < -\lambda$.
  Επομένως $\hat{\beta}_L = \frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} + \lambda$.

- Επειδή επιπλέον $-1 \leq w \leq 1$ και $\lambda > 0$ προκύπτει ότι:

  $$\begin{aligned}-\lambda \leq \lambda w \leq \lambda \\ \frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} - \lambda \leq \hat{\beta}_L \leq \frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} + \lambda\end{aligned}$$

  Εξετάζουμε την τελευταία περίπτωση όπου $\left|\frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n}\right| \leq \lambda$ που ισοδυναμεί με $\frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} - \lambda < 0$ και $\frac{\boldsymbol{X}_1^T\boldsymbol{y}}{n} + \lambda > 0$.

Έστω ότι $\hat{\beta}_L > 0$. Τότε έχουμε: $\hat{\beta}_L = \frac{\boldsymbol{X}_1^T \boldsymbol{y}}{n} - \lambda$ το οποίο εξ υποθέσεως είναι αρνητικό και καταλήγουμε σε άτοπο.

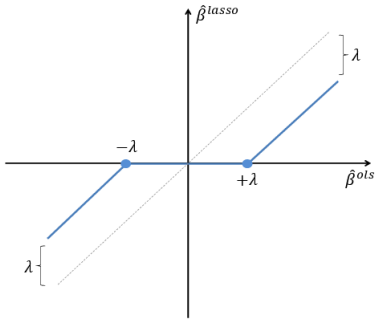
Ομοίως, εάν $\hat{\beta}_L < 0$ τότε $\hat{\beta}_L = \frac{\boldsymbol{X}_1^T \boldsymbol{y}}{n} + \lambda$ το οποίο εξ υποθέσεως είναι θετικό και πάλι καταλήγουμε σε άτοπο. Στη τελευταία περίπτωση λοιπόν το $\hat{\beta}_L$ «εγκλωβίζεται» στο μηδέν.

Συγκεντρώνοντας τα παραπάνω προκύπτει η κλειστή μορφή της λύσης της μεθόδου *lasso* με μία επεξηγηματική μεταβλητή:

$$\hat{\beta}_L = \begin{cases} \hat{\beta}_{ols} - \lambda & , \hat{\beta}_{ols} > \lambda \\ 0 & , \left|\hat{\beta}_{ols}\right| \leq \lambda \\ \hat{\beta}_{ols} + \lambda & , \hat{\beta}_{ols} < -\lambda \end{cases} \tag{66}$$

Με τη βοήθεια της soft-thresholding συνάρτησης $S_\lambda : \mathbb{R} \to \mathbb{R}$ η σχέση (66) παίρνει τη μορφή:

$$\hat{\beta}_L = S_\lambda(\hat{\beta}_{ols}), \quad S_\lambda(x) = \begin{cases} x - \lambda & , x > \lambda \\ 0 & , |x| < \lambda \\ x + \lambda & , x < -\lambda \end{cases}$$



Διάγραμμα 5.1: Η λύση της μεθόδου lasso με μία επεξηγηματική μεταβλητή στην soft-thresholding κλειστή μορφή της για διάφορες τιμές της εκτιμήτριας ελαχίστων τετραγώνων και σταθερό $\lambda > 0$.

Από τη σχέση (66) και το Διάγραμμα 5.1 παρατηρούμε ότι η εκτιμήτρια lasso $\hat{\beta}_L$ συρρικνώνει προς το μηδέν την εκτίμηση $\hat{\beta}_{ols}$ όταν αυτή είναι θετική ή αρνητική και υπερβαίνει κατά απόλυτη τιμή την παράμετρο ποινικοποίησης λ. Επιπλέον διατηρείται στην περίπτωση αυτή το πρόσημο της εκτιμήτριας $\hat{\beta}_{ols}$.

Για πολύ μικρές απόλυτες τιμές της εκτιμήτριας $\hat{\beta}_{ols}$ η εκτιμήτρια lasso εξισώνεται με το μηδέν, πράγμα που φανερώνει το λόγο για τον οποίο η μέθοδος πραγματοποιεί επιλογή μεταβλητών.

5.2.2 Περίπτωση ορθοκανονικού σχεδιασμού

Τα δεδομένα μας είναι ορθοκανονικά σχεδιασμένα όταν ο κανονικοποιημένος πίνακας σχεδιασμού $\boldsymbol{X}$ είναι ορθογώνιος και επομένως $\boldsymbol{X}^T\boldsymbol{X} = \boldsymbol{I}n$, όπου $\boldsymbol{I}$ ο $p \times p$ μοναδιαίος πίνακας. Ισοδύναμα, για κάθε στήλη-διάνυσμα i, j του $\boldsymbol{X}$ ισχύει ότι:

$$\boldsymbol{X}_i^T\boldsymbol{X}_j = \begin{cases} 0 & , i \neq j \\ n & , i = j \end{cases}.$$

Θα δείξουμε ότι στον ορθοκανονικό σχεδιασμό η μέθοδος lasso επιδέχεται λύση **κλειστής** μορφής. Η συνάρτηση ελαχιστοποίησης γράφεται:

$$L(\boldsymbol{\beta}) = \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2n}\|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \right\}$$

$$\begin{aligned}
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{1}{2n}(\boldsymbol{y}^T\boldsymbol{y} - 2\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y} + \boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{X}\boldsymbol{\beta}) + \lambda\sum_{j=1}^{p}|\beta_j| \right\} \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{\|\boldsymbol{y}\|_2^2}{2n} - \frac{\boldsymbol{\beta}^T\boldsymbol{X}^T\boldsymbol{y}}{n} + \frac{\boldsymbol{\beta}^T\boldsymbol{\beta}}{2} + \lambda\sum_{j=1}^{p}|\beta_j| \right\} \qquad (67) \\
&= \underset{\boldsymbol{\beta}\in\mathbb{R}^p}{\text{minimize}} \left\{ \frac{\|\boldsymbol{y}\|_2^2}{2n} + \sum_{j=1}^{p}\left(-\frac{(\boldsymbol{X}^T\boldsymbol{y})_j\beta_j}{n} + \frac{\beta_j^2}{2} + \lambda|\beta_j| \right) \right\}
\end{aligned}$$

Το πρώτο μέλος της συνάρτησης $L(\boldsymbol{\beta})$ είναι πάντα θετικό και επομένως επικεντρωνόμαστε στο δεύτερο. Ελαχιστοποιώντας κάθε μέρος του αθροίσματος που εμφανίζεται στην τελευταία σχέση (67), ελαχιστοποιείται και η $L(\boldsymbol{\beta})$. Αναζητούμε λοιπόν ελάχιστο σημείο των επί μέρους συναρτήσεων:

$$L_j(\beta_j) = -\frac{(\boldsymbol{X}^T\boldsymbol{y})_j\beta_j}{n} + \frac{\beta_j^2}{2} + \lambda|\beta_j|$$

Χρησιμοποιώντας τη συνθήκη (63) προκύπτει ότι:

$$-\frac{(\boldsymbol{X}_1^T\boldsymbol{y})_j}{n}+\hat{\beta}_j+\lambda w_j=0$$
$$\hat{\beta}_j=\frac{(\boldsymbol{X}_1^T\boldsymbol{y})_j}{n}-\lambda w_j$$

Καταλήγουμε λοιπόν στην περίπτωση της μίας ανεξάρτητης μεταβλητής από όπου έχουμε:

$$\hat{\beta}_j=\hat{\beta}_{ols,j}-\lambda w_j$$

$$\hat{\beta}_j=\begin{cases}\hat{\beta}_{ols,j}-\lambda & ,\hat{\beta}_{ols,j}>\lambda\\ 0 & ,\left|\hat{\beta}_{ols,j}\right|\leq\lambda\\ \hat{\beta}_{ols,j}+\lambda & ,\hat{\beta}_{ols,j}<-\lambda\end{cases} \tag{68}$$

$$\hat{\beta}_j=S_\lambda(\hat{\beta}_{ols,j})$$

5.2.3 Πολυμεταβλητή περίπτωση

Εξετάζουμε τη γενική πολυμεταβλητή περίπτωση υποθέτοντας ότι ο πίνακας σχεδιασμού $\boldsymbol{X}$ είναι πλήρους τάξης με αποτέλεσμα ο πίνακας $\boldsymbol{X}^T\boldsymbol{X}$ να είναι αντιστρέψιμος. Συμβολίζουμε επιπλέον με $\boldsymbol{X}_{-j}$ τον πίνακα με όλες τις στήλες του $\boldsymbol{X}$ εκτός από την j-οστή, $\boldsymbol{\beta}_{-j}$ το διάνυσμα με όλα τα στοιχεία του $\boldsymbol{\beta}_{p\times 1}$ εκτός από το j-οστό. Η συνάρτηση ελαχιστοποίησης γράφεται ως:

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

Χρησιμοποιώντας τη συνθήκη (63) έχουμε ότι:

$$-\frac{\boldsymbol{X}^T\boldsymbol{y}}{n}+\frac{\boldsymbol{X}^T\boldsymbol{X}\hat{\boldsymbol{\beta}}_L}{n}+\lambda\boldsymbol{w}=0$$
$$-\frac{\boldsymbol{X}^T}{n}\left[\boldsymbol{y}-\boldsymbol{X}\hat{\boldsymbol{\beta}}_L\right]+\lambda\boldsymbol{w}=0$$

Το διάνυσμα $\boldsymbol{w}$ έχει τη μορφή $\boldsymbol{w}=(w_1,w_2,\ldots,w_p)^T$. Ελαχιστοποιώντας ως προς κάθε συνιστώσα $\hat{\beta}_{L,j}$ του διανύσματος $\hat{\boldsymbol{\beta}}_L$ έχουμε:

$$-\frac{1}{n}\boldsymbol{X}_j^T\left[\boldsymbol{y}-\boldsymbol{X}_j\hat{\beta}_{L,j}-\boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}\right]+\lambda w_j=0$$

$$-\frac{1}{n}\boldsymbol{X}_j^T\boldsymbol{y}+\frac{1}{n}\boldsymbol{X}_j^T\boldsymbol{X}_j\hat{\beta}_{L,j}+\frac{1}{n}\boldsymbol{X}_j^T\boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}+\lambda w_j=0$$

Όμως $\boldsymbol{X}_j^T\boldsymbol{X}_j=n, \forall j=1,\ldots,p$ και η τελευταία εξίσωση γράφεται:

$$-\frac{1}{n}\boldsymbol{X}_j^T\boldsymbol{y}+\hat{\beta}_{L,j}+\frac{1}{n}\boldsymbol{X}_j^T\boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}+\lambda w_j=0$$

$$\hat{\beta}_{L,j}=\frac{\boldsymbol{X}_j^T\boldsymbol{y}}{n}-\frac{\boldsymbol{X}_j^T\boldsymbol{X}_{-j}\hat{\boldsymbol{\beta}}_{-j}}{n}+\lambda w_j$$

Από την παραπάνω σχέση που καταλήξαμε διαπιστώνουμε ότι η λύση της συνιστώσας $\hat{\beta}_{L,j}$ του διανύσματος $\hat{\boldsymbol{\beta}}_L$ εξαρτάται από όλες τις άλλες συνιστώσες $i\neq j$ και επομένως δεν υπάρχει λύση κλειστής μορφής της μεθόδου lasso στην πολυμεταβλητή περίπτωση.

Ωστόσο λόγω της κυρτότητας της συνεχούς συνάρτησης ελαχιστοποίησης ορισμένης σε συμπαγή διανυσματικό υπόχωρο, το σημείο ολικού ελαχίστου μπορεί να βρεθεί αποτελεσματικά με τη χρήση αλγοριθμικών υπολογιστικών μεθόδων. Μια απλή μέθοδος που εξετάζουμε και εξυπηρετεί τον σκοπό αυτό είναι ο Cyclic Coordinate Descent αλγόριθμος. Είδαμε προηγουμένως ότι στην περίπτωση μίας επεξηγηματικής μεταβλητής η λύση που επιτυγχάνεται έχει κλειστή μορφή και αυτή την ιδιότητα εκμεταλλεύεται ο αλγόριθμος όπως θα δούμε στη συνέχεια.

5.3 Αλγόριθμος Cyclic Coordinate Descent

Ο Cyclic Coordinate Descent αλγόριθμος είναι μια επαναληπτική μέθοδος που λύνει το πρόβλημα ελαχιστοποίησης της $L(\boldsymbol{\beta})$ ως προς μία συνιστώσα (coordinate) του διανύσματος $\boldsymbol{\beta}$ κρατώντας τις υπόλοιπες $p-1$ συνιστώσες σταθερές κάθε φορά. Η διαδικασία επαναλαμβάνεται κυκλικά για κάθε $j=1,2,\ldots,p$, μέχρι να επιτευχθεί σύγκλιση της συνάρτησης $L(\boldsymbol{\beta})$. Συμβολίζοντας λοιπόν με $\boldsymbol{\beta}_{-j}$ το διάνυσμα που περιέχει όλες τις στήλες του κανονικοποιημένου πίνακα σχεδιασμού εκτός της j-οστής, το υποπρόβλημα σε κάθε βήμα του αλγόριθμου έχει τη μορφή:

$$\underset{\beta_j\in\mathbb{R}}{\arg\min}\left\{\frac{1}{2n}\|\boldsymbol{y}-\boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}-\beta_j\boldsymbol{X}_j\|_2^2+\lambda\sum_{i\neq j}|\beta_i|+\lambda|\beta_j|\right\}$$

Θέτοντας τώρα $\boldsymbol{r}_j:=\boldsymbol{y}-\boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}$ το j-οστό μερικό υπόλοιπο του γραμμικού

μοντέλου που προσαρμόστηκε χωρίς τη μεταβλητή X_j έχουμε:

$$\begin{aligned}
&\arg\min_{\beta_j \in \mathbb{R}} \left\{ \frac{1}{2n} \|\boldsymbol{r}_j - \beta_j \boldsymbol{X}_j\|_2^2 + \lambda|\beta_j| + \lambda \sum_{i \neq j} |\beta_i| \right\} \\
&\arg\min_{\beta_j \in \mathbb{R}} \left\{ \frac{1}{2n} \left(\boldsymbol{r}_j^T \boldsymbol{r}_j - 2\beta_j \boldsymbol{X}_j^T \boldsymbol{r}_j + \beta_j^2 \boldsymbol{X}_j^T \boldsymbol{X}_j \right) + \lambda|\beta_j| + \sum_{i \neq j} |\beta_i| \right\} \\
&\arg\min_{\beta_j \in \mathbb{R}} \left\{ \frac{\boldsymbol{r}_j^T \boldsymbol{r}_j}{2n} - \frac{\beta_j \boldsymbol{X}_j^T \boldsymbol{r}_j}{n} + \frac{\beta_j^2}{2} + \lambda|\beta_j| + \lambda \sum_{i \neq j} |\beta_i| \right\}
\end{aligned}$$

Εφαρμόζοντας την συνθήκη (63) στην τελευταία σχέση ως προς β_j παίρνουμε:

$$\begin{aligned}
-\frac{\boldsymbol{X}_j^T \boldsymbol{r}_j}{n} + \hat{\beta}_j + \lambda w_j = 0 \\
\hat{\beta}_j = \frac{\boldsymbol{X}_j^T \boldsymbol{r}_j}{n} - \lambda w_j
\end{aligned}$$

Η τελευταία εξίσωση γράφεται με βάση την περίπτωση μίας επεξηγηματικής μεταβλητής στη συμπαγή μορφή:

$$\hat{\beta}_j = S_\lambda \left(\frac{\boldsymbol{r}_j^T \boldsymbol{X}_j}{n} \right) = \begin{cases} \frac{\boldsymbol{r}_j^T \boldsymbol{X}_j}{n} - \lambda & , \frac{\boldsymbol{r}_j^T \boldsymbol{X}_j}{n} > \lambda \\ 0 & , \left| \frac{\boldsymbol{r}_j^T \boldsymbol{X}_j}{n} \right| < \lambda \\ \frac{\boldsymbol{r}_j^T \boldsymbol{X}_j}{n} + \lambda & , \frac{\boldsymbol{r}_j^T \boldsymbol{X}_j}{n} < -\lambda \end{cases}$$

Σημείωση: Η συνάρτηση ελαχιστοποίησης της μεθόδου lasso έχει τη διαχωρίσιμη μορφή:

$$f(\beta_1, \ldots, \beta_p) = g(\beta_1, \ldots, \beta_p) + \sum_{j=1}^{p} h_j(\beta_j)$$

με $g(\boldsymbol{\beta}) = \frac{1}{2n} \|\boldsymbol{y} - \boldsymbol{X}\boldsymbol{\beta}\|$ και $h_j(\beta_j) = \lambda|\beta_j|$. Ο Tseng (1988, 2001) απέδειξε ότι για οποιαδήποτε κυρτή αντικειμενική συνάρτηση που έχει την παραπάνω μορφή, ο αλγόριθμος coordinate descent συγκλίνει εγγυημένα στο ολικό ελάχιστο της f.

Στο ακόλουθο σχεδιάγραμμα (Αλγόριθμος 5.1) βλέπουμε σε μορφή ψευδογλώσσας τον τρόπο με τον οποίο λειτουργεί ο αλγόριθμος cyclic coordinate descent εφαρμοσμένος στην πολυμεταβλητή περίπτωση της μεθόδου lasso.

Algorithm: Cyclic Coordinate Descent

Input: Data $(\boldsymbol{y}, \boldsymbol{X})$
Output: $\hat{\boldsymbol{\beta}}$ lasso estimator of coefficients
begin
  Start with random initial values $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$
  repeat
    for *each* $j \in \{1, \ldots, p\}$ **do**
      Calculate the partial residual $\boldsymbol{r}_j$:
      $\boldsymbol{r}_j = \boldsymbol{y} - \boldsymbol{X}_{-j}\boldsymbol{\beta}_{-j}$ and
      $\frac{\boldsymbol{r}_j^T \boldsymbol{X}_j}{n}$
      Update lasso coordinate β_j as:
      $\beta_j = S_\lambda \left(\frac{\boldsymbol{r}_j^T \boldsymbol{X}_j}{n} \right)$
  until *convergence of* $L(\boldsymbol{\beta})$;
  $\hat{\boldsymbol{\beta}}_L = \boldsymbol{\beta}$
  return $\hat{\boldsymbol{\beta}}_L$

Αλγόριθμος 5.1: Σχεδιάγραμμα υλοποίησης του αλγόριθμου Cyclic Coordinate Descent στην πολυμεταβλητή περίπτωση της μεθόδου lasso.

5.4 Παράμετρος ποινής λ και φράγμα t

Η παράμετρος ποινής λ και το φράγμα t αντίστοιχα διαδραματίζουν τον πιο σημαντικό ρόλο στη μέθοδο ποινικοποίησης lasso. Θέτοντας $t = max \sum_{j=1}^{p} |\beta_j| = \max \|\boldsymbol{\beta}\|_1$ ή $\lambda = 0$, δεν υφίσταται συρρίκνωση προς το μηδέν σε κανένα συντελεστή βj και η λύση της μεθόδου lasso συμπίπτει με την εκτιμήτρια της μεθόδου ελαχίστων τετραγώνων. Περιορίζοντας τις απόλυτες τιμές των συντελεστών του μοντέλου μπορεί να προκύψουν μηδενικές τιμές β_j γεγονός που οφείλεται στη φύση της ποινικοποίησης της ℓ_1 νόρμας στον περιορισμό της μεθόδου. Υπό το πλαίσιο αυτό η μέθοδος lasso παρέχει έναν αυτόματο τρόπο επιλογής μεταβλητών στο πρόβλημα επιλογής μοντέλου και επιπλέον αντιμετωπίζει το πιθανό πρόβλημα πολυσυγγραμικότητας ελέγχοντας και περιορίζοντας τις μεγάλες μη λογικές τιμές των συντελεστών $\boldsymbol{\beta}$.

Μικρές τιμές της παραμέτρου λ (μεγάλες αντίστοιχα τιμές του t) «απελευθερώνουν» περισσότερες μεταβλητές-παραμέτρους και επιτρέπουν στο μοντέλο να προσαρμοστεί «πιο κοντά» στα δεδομένα μας. Ωστόσο, πολύ μικρές τιμές της παραμέτρου λ μπορεί να οδηγήσουν σε υπερπροσαρμοσμένα μοντέλα (over-

fitting), ενώ πολύ μεγάλες τιμές σε υποπροσαρμοσμένα μοντέλα (underfitting). Και στις δύο περιπτώσεις η προβλεπτική ικανότητα των μοντέλων αξιολογούμενη σε ανεξάρτητο των αρχικών σύνολο δεδομένων δε θα είναι ικανοποιητική.

5.4.1 Cross-Validation εκτίμηση

Η βέλτιστη τιμή της παραμέτρου λ που επιτυγχάνει την ισορροπία στις προηγούμενες δύο ακραίες περιπτώσεις μοντελοποίησης, μπορεί να εκτιμηθεί με βάση την προβλεπτική ικανότητα των μοντέλων χρησιμοποιώντας την cross validation μέθοδο του Κεφαλαίου 4. Η διαδικασία εφαρμόζεται ως εξής:

Αρχικά διαλέγουμε σε πόσους ισοπληθικούς φακέλους K θα διαχωρίσουμε τυχαία τα δεδομένα μας με πιο συνηθισμένες επιλογές τις $K = 5, 10$. Για κάθε φάκελο (test set) $k = 1, \ldots, K$ εφαρμόζουμε τη μέθοδο lasso στα υπόλοιπα σύνολα δεδομένων των $K - 1$ φακέλων (training set) για ένα εύρος τιμών της παραμέτρου λ και χρησιμοποιούμε κάθε προσαρμοσμένο μοντέλο για να προβλέψουμε τις αναμενόμενες τιμές της μεταβλητής απόκρισης στα δεδομένα του φακέλου k.

Αν λοιπόν $\lambda_z, z = 1, \ldots, Z$ είναι οι διάφορες υποψήφιες τιμές της παραμέτρου λ, για κάθε τιμή λ_z υπολογίζουμε στον k φάκελο την ποσότητα:

$$T_{mse}(\lambda_z, k) = \frac{1}{n_k} \sum_{i \in T_k} \left(y_i - \sum_{j=1}^{p} \hat{\beta}_j(\lambda_z, k) x_{ij} \right)^2$$

Συμβολίζουμε με $\hat{\beta}_j(\lambda_z, k)$ τους συντελεστές της μεθόδου lasso που προσαρμόζεται για $\lambda = \lambda_z$ στα δεδομένα των υπόλοιπων $K-1$ φακέλων και n_k είναι ο αριθμός των στοιχείων που περιέχει κάθε φάκελος. Στη συνέχεια σε κάθε τιμή λ_z υπολογίζουμε το μέσο τετραγωνικό σφάλμα που προκύπτει θεωρώντας τον μέσο όρο του τετραγωνικού σφάλματος $T_{mse}(\lambda_z, k)$ των K φακέλων:

$$MSE_{\lambda_z} = \frac{1}{K} \sum_{k=1}^{K} T_{mse}(\lambda_z, k) \tag{69}$$

Η τιμή της παραμέτρου λ που επιλέγεται είναι εκείνη που ελαχιστοποιεί το CV μέσο τετραγωνικό σφάλμα σύμφωνα με τη σχέση (69). Συνήθως ωστόσο επειδή από κάποιες τιμές της παραμέτρου λ και μετά δεν παρατηρείται μεγάλη διαφορά στο μέσο τετραγωνικό σφάλμα MSE_λ και επειδή προτιμούνται φειδωλά μοντέλα (όσο αυτό είναι εφικτό), εφαρμόζουμε τον κανόνα της cross-validation μεθόδου «**one-standard-error-rule**» όπου:

Ξεκινάμε με την παράμετρο $\hat{\lambda}$ για την οποία $\hat{\lambda} = \arg\min_{\lambda_z} MSE_{\lambda_z}$

και υπολογίζουμε το τυπικό σφάλμα $SE(\hat{\lambda}) = SD(\hat{\lambda})/\sqrt{K}$

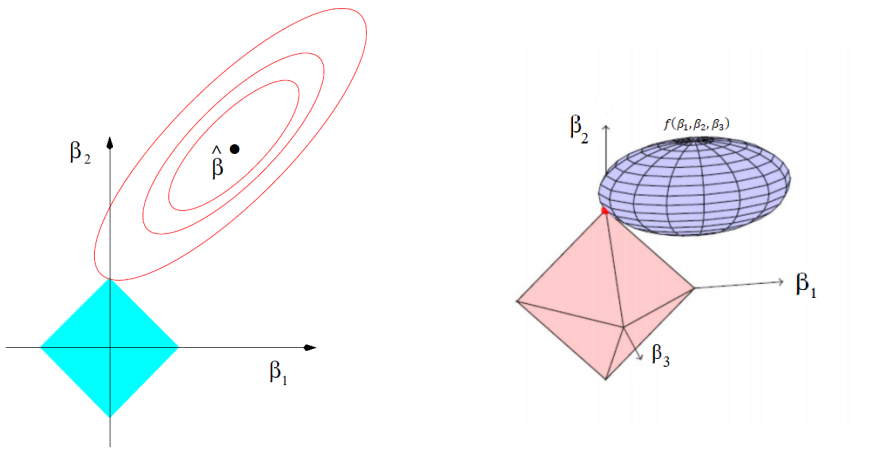
$$\text{με } SD(\hat{\lambda}) = \sqrt{\frac{\sum_{k=1}^{K}(T_{mse}(\hat{\lambda}, k) - MSE_{\hat{\lambda}})^2}{K-1}}.$$

Αρχίζοντας μετά από το $\hat{\lambda}$ και κατευθυνόμενοι σε μεγαλύτερες τιμές, το λ που επιλέγεται ως παράμετρος ποινής είναι αυτό για το οποίο παύει να ισχύει ότι:

$$MSE_{\lambda} \leq MSE_{\hat{\lambda}} + SE(\hat{\lambda})$$

Αυξάνοντας με αυτό τον τρόπο την τιμή της παραμέτρου λ επιτυγχάνεται χωρίς να χειροτερέψει δραματικά το προβλεπτικό σφάλμα MSE, η προτίμηση φειδωλότερου μοντέλου.

Ο ρόλος της παραμέτρου ποινής λ και του φράγματος t μπορεί να γίνει ακόμα περισσότερο κατανοητός μελετώντας γραφικά την περίπτωση δύο και τριών επεξηγηματικών μεταβλητών.



(α′) 2 επεξηγηματικές μεταβλητές (β′) 3 επεξηγηματικές μεταβλητές

Διάγραμμα 5.2: Γραφική αναπαράσταση της εκτιμήτριας lasso με δύο και τρεις επεξηγηματικές μεταβλητές.

Στο Διάγραμμα 5.2 (α′) ο οριζόντιος και ο κατακόρυφος άξονας αντιστοιχούν στις πιθανές τιμές των συντελεστών β_1, β_2 του γραμμικού μοντέλου. Το

σημείο $\hat{\boldsymbol{\beta}}$ απεικονίζει τη λύση της μεθόδου ελαχίστων τετραγώνων και οι ελλείψεις αντιστοιχούν στις τιμές που παίρνει η συνάρτηση τετραγωνικού σφάλματος υπολοίπων $f(\boldsymbol{\beta}) = f(\beta_1, \beta_2) = \sum_{i=1}^{n}(y_i - \sum_{j=1}^{2} \beta_j x_{ij})$ για διάφορες τιμές των συντελεστών β_1, β_2. Ο περιορισμός $g(\boldsymbol{\beta}) = \sum_{j=1}^{p} |\beta_j| = |\beta_1| + |\beta_2| \leq t$ απεικονίζεται από το δισδιάστατο ρόμβο με κέντρο την αρχή των αξόνων και πλευρά ίση με t.

Η λύση της μεθόδου lasso πρέπει να ικανοποιεί εξίσου τον περιορισμό $g(\boldsymbol{\beta})$ και την ελαχιστοποίηση του αθροίσματος των τετραγωνικών υπολοίπων και επομένως αντιστοιχεί στο πρώτο σημείο όπου έρχονται σε επαφή οι ελλείψεις της $f(\boldsymbol{\beta})$ με τον δισδιάστατο ρόμβο. Στο συγκεκριμένο παράδειγμα παρατηρούμε ότι η λύση της μεθόδου βρίσκεται σε γωνία του ρόμβου του περιορισμού $g(\boldsymbol{\beta})$ με αποτέλεσμα να μηδενίζεται ο συντελεστής β_1. Συνεπώς φαίνεται και γεωμετρικά πλέον ο λόγος για τον οποίο η μέθοδος πραγματοποιεί αυτόματα επιλογή μεταβλητών (variable selection).

Για $p > 2$ ο παραμετρικός χώρος πληθαίνει και ο περιορισμός $g(\boldsymbol{\beta})$ παίρνει τη μορφή ρομβοειδούς σε μεγαλύτερες διαστάσεις με περισσότερες γωνίες με αποτέλεσμα να αυξάνονται οι πιθανοί μηδενισμοί συντελεστών του μοντέλου. Στο Διάγραμμα 5.2 (β΄) αναπαρίσταται η περίπτωση $p = 3$ με το συντελεστή β_1 να μηδενίζεται λόγω της τομής της σφαίρας της συνάρτησης τετραγωνικών υπολοίπων και της πάνω γωνίας της πυραμίδας του περιορισμού $g(\boldsymbol{\beta})$.

Κεφάλαιο 6

Μέχρι στιγμής στο μεγαλύτερο κομμάτι της εργασίας αναλύσαμε αρχικά τη θεωρία και τις ιδιότητες των πιο γνωστών κριτηρίων πληροφορίας με συνέχεια τη μελέτη υπολογιστικών μεθόδων προσδιορισμού του προβλεπτικού σφάλματος ενός στατιστικού μοντέλου, καταλήγοντας στην μέθοδο ποινικοποίησης και αυτόματης επιλογής μεταβλητών lasso.

Σε αυτό το τελευταίο κεφάλαιο αξιολογούμε στο πρώτο μέρος την απόδοση και τις θεωρητικές ιδιότητες των κριτηρίων πληροφορίας AIC, AIC_c, BIC μαζί με διάφορα μέτρα καλής προσαρμογής, χρησιμοποιώντας προσομοιωμένα δεδομένα της πολυδιάστατης κανονικής κατανομής. Έπειτα, με μία δεύτερη περισσότερο εξειδικευμένη προσομοίωση η οποία ωστόσο προσεγγίζει πολύ συχνές καταστάσεις προβλημάτων ανάλυσης δεδομένων, εξετάζουμε την αποτελεσματικότητα της μεθόδου lasso στην κατάλληλη επιλογή μεταβλητών στο πρόβλημα της επιλογής μοντέλου.

6.1 Προσομοίωση 1

Δημιουργούμε 100 διαφορετικά τυχαία δείγματα με μέγεθος δείγματος $n = 40$ και $p = 8$ ανεξάρτητες τυχαίες μεταβλητές (predictors). Οι τιμές των 8 ανεξάρτητων τυχαίων μεταβλητών προκύπτουν προσομοιώνοντας από την πολυδιάστατη κανονική κατανομή $N(\mathbf{0}_8, \boldsymbol{\Sigma})$, όπου $\boldsymbol{\Sigma}_{p\times p}$ είναι ο συμμετρικός πίνακας συνδιακύμανσης με στοιχεία $\Sigma_{i,j} = (0.5)^{|i-j|}$. Από τη μορφή του πίνακα Σ, ο οποίος ταυτίζεται με τον πίνακα συσχέτισης των ανεξάρτητων μεταβλητών, συμπεραίνουμε ότι έχουμε πρόβλημα **ασθενούς πολυσυγγραμμικότητας**. Για την επίδραση των επεξηγηματικών μεταβλητών που συνεισφέρουν στο πραγματικό μοντέλο θέσαμε $(\beta_1, \beta_2, \beta_5) = (3, 1.5, 2)^T$ και οι υπόλοιποι συντελεστές του διανύσματος $\boldsymbol{\beta}_{8\times 1}$ είναι ίσοι με το μηδέν. Υποθέτουμε εν τέλει ότι οι τιμές της μεταβλητής απόκρισης Y δίνονται από το γραμμικό μοντέλο:

$$\boldsymbol{Y} = \beta_0 \cdot 1_n + \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

Για το σταθερό όρο θεωρούμε ότι $\beta_0 = 0.8$ και για το διάνυσμα των τυχαίων σφαλμάτων ισχύει ότι: $\boldsymbol{\varepsilon} \sim N_{40}(\mathbf{0}_{40}, \sigma^2\mathbf{1}_n)$, $\sigma^2 = 1$, και $\mathbf{1}_n$ μοναδιαίος

$n \times n$ πίνακας. Ο πίνακας σχεδιασμού $\tilde{\boldsymbol{X}}$ με διαστάσεις 40×8 περιέχει τις παρατηρούμενες τιμές των μεταβλητών-στηλών X_i, $i = 1, \ldots, 8$.

Σε κάθε ένα από τα 100 ανεξάρτητα δείγματα δεδομένων:

1. Πραγματοποιούμε πλήρη εξερεύνηση του χώρου μοντελοποίησης προσαρμόζοντας όλα τα δυνατά (χωρίς αλληλεπιδράσεις) $2^8 = 256$ το πλήθος γραμμικά μοντέλα με μεταβλητή απόκρισης την Y και επεξηγηματικές μεταβλητές στοιχεία από το δυναμοσύνολο του συνόλου $\boldsymbol{X} = (X_1, X_2, \ldots, X_p)$. Στη συνέχεια σε κάθε μοντέλο που προκύπτει υπολογίζουμε τις τιμές των κριτηρίων AIC, AIC_c, BIC και τις τιμές των μέτρων καλής προσαρμογής R^2, R^2_{adj}. Κρατάμε κάθε φορά το μοντέλο που ικανοποιεί τη συνθήκη βελτιστότητας του εκάστοτε κριτηρίου.

2. Εφαρμόζουμε τις τρεις μεθόδους κατά βήματα, με τη διαδικασία διαδοχικής αφαίρεσης να ξεκινάει από το πλήρες μοντέλο που περιέχει και τις 8 επεξηγηματικές μεταβλητές, τη διαδικασία της διαδοχικής πρόσθεσης να αρχίζει από το μοντέλο με καμία επεξηγηματική μεταβλητή, και τη διαδικασία κατά βήματα να ξεκινά από το πλήρες μοντέλο και να προσθαφαιρεί μεταβλητές σε κάθε βήμα της.

3. Αξιοποιούμε επιπλέον τη μέθοδο ποινικοποίησης lasso για την οποία η παράμετρος ποινικοποίησης λ βρίσκεται χρησιμοποιώντας τη 10-fold cross validation μέθοδο μέσω του «one standard error rule» όπως εφαρμόζεται στην Παράγραφο 5.4.1.

Οι σχετικές συχνότητες επιλογής του πραγματικού μοντέλου με επεξηγηματικές μεταβλητές τις X_1, X_2, X_5 εκτελώντας την παραπάνω διαδικασία και στα 100 προσομοιωμένα τυχαία δείγματα είναι:

	Σχετική συχνότητα
BIC	0.73
AIC_c	0.6
$Lasso$	0.56
$Stepforward$	0.43
AIC	0.40
$Stepbackward$	0.40
$Stepboth$	0.39
R^2_{adj}	0.18
R^2	0

Πίνακας 6.1: Σχετικές συχνότητες επιλογής του πραγματικού μοντέλου από τις μεθόδους της προσομοίωσης 1.

Από τα αποτελέσματα του Πίνακα 6.1 της $1^{ης}$ προσομοίωσης δεδομένων προκύπτουν οι εξής παρατηρήσεις:

- Τα μέτρα καλής προσαρμογής R^2_{adj} και R^2 ανιχνεύουν και επιλέγουν το πραγματικό μοντέλο που γέννησε τα δεδομένα μόλις 18 στις 100 φορές και 0 στις 100 φορές αντίστοιχα. Για το λόγο αυτό βλέπουμε ότι δε θα πρέπει να χρησιμοποιούνται ως κριτήρια επιλογής του βέλτιστου μοντέλου. Η εφαρμογή τους έπεται της διαδικασίας επιλογής μοντέλου και ο ρόλος τους είναι να ελέγχουν την προσαρμογή του μοντέλου που προέκυψε ότι είναι βέλτιστο με βάση άλλα πιο αξιόπιστα κριτήρια.

- Η διαδικασία κατά βήματα (Stepboth) επιλέγει 39 στις 100 φορές το πραγματικό μοντέλο. Την ίδια σχεδόν συχνότητα (40/100) σημειώνει και η διαδικασία της διαδοχικής αφαίρεσης (Stepbackward) μαζί με το κριτήριο πληροφορίας AIC.

- Η διαδικασία της διαδοχικής πρόσθεσης Stepforward βρίσκει το πραγματικό μοντέλο 43 στις 100 φορές, ενώ η μέθοδος Lasso είναι ανώτερη επιλέγοντας 56 στις 100 φορές το σωστό μοντέλο.

- Τέλος, η πλήρης εξερεύνηση του χώρου μοντελοποίησης με τα κριτήρια AIC_c και BIC σημειώνει τα καλύτερα αποτελέσματα επιλέγοντας το πραγματικό μοντέλο 60 και 73 στις 100 φορές αντίστοιχα. Παρατηρούμε μάλιστα ότι η διορθωμένη εκδοχή AIC_c επιλέγει αρκετά περισσότερες φορές το σωστό μοντέλο σε σχέση με το κριτήριο AIC. Επιπλέον, επιβεβαιώνεται ότι η ιδιότητα της συνέπειας (σύμφωνα με την οποία προτιμούνται φειδωλότερα μοντέλα), που χαρακτηρίζει το κριτήριο BIC παίζει καθοριστικό ρόλο στην εύρεση του σωστού μοντέλου.

Από τα παραπάνω συμπεραίνουμε ότι για σχετικά μικρό αριθμό διαθέσιμων επεξηγηματικών μεταβλητών $p < 15$ (στην περίπτωσή μας $p = 8$), η στρατηγική εύρεσης του πραγματικού μοντέλου με τα βέλτιστα αποτελέσματα είναι η πλήρης εξερεύνηση του χώρου μοντελοποίησης εφαρμόζοντας κάποιο κριτήριο πληροφορίας με καλές ιδιότητες όπως το BIC και το AIC_c.

6.2 Προσομοίωση 2

Στη $2^η$ προσομοίωση δημιουργούμε 100 διαφορετικά τυχαία δείγματα με μέγεθος $n = 25$ και $p = 50$ ανεξάρτητες επεξηγηματικές μεταβλητές (predictors), οι τιμές των οποίων προκύπτουν προσομοιώνοντας από την πολυδιάστατη κανονική κατανομή $N(\mathbf{0}_{50}, \mathbf{\Sigma})$. Ο συμμετρικός πίνακας συνδιακύμανσης $\mathbf{\Sigma}_{p\times p}$ έχει στοιχεία που ορίζονται μέσω της σχέσης $\Sigma_{i,j} = (0.75)^{|i-j|}$. Παρατηρούμε ότι

στη 2^η προσομοίωση το πρόβλημα ασθενούς πολυσυγγραμμικότητας είναι ισχυρότερο. Στο πραγματικό μοντέλο συνεισφέρουν μόνο οι μεταβλητές X_1, X_2, X_{10} με αντίστοιχους συντελεστές (coefficients) $(\beta_1, \beta_2, \beta_{10}) = (2, 0.8, 1.5)^T$ και οι συντελεστές των υπόλοιπων ανεξάρτητων μεταβλητών είναι ίσοι με το μηδέν. Θεωρούμε επιπλέον ότι ο σταθερός όρος είναι $\beta_0 = 0.6$ και η μεταβλητή απόκρισης συνδέεται με τις 50 επεξηγηματικές μεταβλητές μέσω της γραμμικής σχέσης:

$$\boldsymbol{Y} = \beta_0 \cdot 1_n + \tilde{\boldsymbol{X}}\boldsymbol{\beta} + \boldsymbol{\varepsilon}.$$

Ο πίνακας σχεδιασμού $\tilde{\boldsymbol{X}}_{25\times 50}$ περιέχει τις 25 παρατηρούμενες τιμές των ανεξάρτητων μεταβλητών $X_i, i = 1, \ldots, 50$ για κάθε ανεξάρτητο προσομοιωμένο τυχαίο δείγμα δεδομένων, και τέλος υποθέτουμε ότι για το διάνυσμα των τυχαίων σφαλμάτων ισχύει ότι $\boldsymbol{\varepsilon} \sim N_{25}(\boldsymbol{0}_{25}, \sigma^2 \boldsymbol{1}_n), \sigma^2 = 1.5$, και $\boldsymbol{1}_n$ μοναδιαίος πίνακας.

Σημείωση: Δεδομένα με πολλές επεξηγηματικές μεταβλητές p που υπερβαίνουν το μέγεθος του δείγματος των παρατηρήσεων n, $(p > n)$ είναι αρκετά σύνηθες φαινόμενο στις στατιστικές αναλύσεις και μοντελοποιήσεις, όπως για παράδειγμα είναι η εξέταση των 30.000 γονιδίων του ανθρώπινου οργανισμού που συνδέονται με την εμφάνιση καρκίνου ή οι εφαρμογές αναγνώρισης προσώπων με πολλές μεταβλητές εισόδου (features).

Η πλήρης εξερεύνηση του χώρου μοντελοποίησης στην 2^η προσομοίωση είναι απαγορευτική καθώς όλα τα δυνατά γραμμικά μοντέλα είναι $2^{50} = 1.1259 \cdot 10^{15}$ το πλήθος. Η ανάλυσή μας αρχίζει εξετάζοντας το πλήρες γραμμικό μοντέλο που περιέχει και τις 50 επεξηγηματικές μεταβλητές χρησιμοποιώντας ένα εκ των 100 προσομοιωμένων τυχαίων δειγμάτων. Στο προγραμματιστικό περιβάλλον της R με την εντολή summary(mfull) πραγματοποιείται η προσαρμογή του πλήρους μοντέλου και τα αποτελέσματα φαίνονται παρακάτω:

```
> summary(mfull)

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

Residuals:
ALL 25 residuals are 0: no residual degrees of freedom!

Coefficients: (26 not defined because of singularities)
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.30545         NA      NA       NA
X1           2.00936         NA      NA       NA
X2           0.01516         NA      NA       NA
X3           1.69159         NA      NA       NA
X4          -1.40871         NA      NA       NA
X5           0.21198         NA      NA       NA
X6           1.03221         NA      NA       NA
X7          -1.00849         NA      NA       NA
X8           0.15226         NA      NA       NA
```

```
X9          -0.37971         NA       NA       NA
X10          1.23571         NA       NA       NA
X11          0.46970         NA       NA       NA
X12         -2.64272         NA       NA       NA
X13          2.84653         NA       NA       NA
X14         -2.10264         NA       NA       NA
X15          0.51910         NA       NA       NA
X16         -1.39049         NA       NA       NA
X17          1.33716         NA       NA       NA
X18         -0.35231         NA       NA       NA
X19         -0.13443         NA       NA       NA
X20         -0.23252         NA       NA       NA
X21          1.90343         NA       NA       NA
X22         -0.78834         NA       NA       NA
X23         -0.04688         NA       NA       NA
X24          1.11414         NA       NA       NA
X25               NA         NA       NA       NA
X26               NA         NA       NA       NA
X27               NA         NA       NA       NA
X28               NA         NA       NA       NA
X29               NA         NA       NA       NA
X30               NA         NA       NA       NA
X31               NA         NA       NA       NA
X32               NA         NA       NA       NA
X33               NA         NA       NA       NA
X34               NA         NA       NA       NA
X35               NA         NA       NA       NA
X36               NA         NA       NA       NA
X37               NA         NA       NA       NA
X38               NA         NA       NA       NA
X39               NA         NA       NA       NA
X40               NA         NA       NA       NA
X41               NA         NA       NA       NA
X42               NA         NA       NA       NA
X43               NA         NA       NA       NA
X44               NA         NA       NA       NA
X45               NA         NA       NA       NA
X46               NA         NA       NA       NA
X47               NA         NA       NA       NA
X48               NA         NA       NA       NA
X49               NA         NA       NA       NA
X50               NA         NA       NA       NA

Residual standard error: NaN on 0 degrees of freedom
Multiple R-squared:      1,     Adjusted R-squared:    NaN
F-statistic:   NaN on 24 and 0 DF,  p-value: NA
```

- Τα υπόλοιπα $y_i - \hat{f}(\boldsymbol{x}_i), i = 1, \ldots, 25$, δεν είναι υπολογίσιμα και κατά επέκταση το άθροισμα των τετραγώνων των υπολοίπων είναι μη προσδιορίσιμο (Residual Standard error: NaN).

- Από τις 50 επεξηγηματικές μεταβλητές του μοντέλου, 26 εκτιμήσεις των συντελεστών $(\hat{\beta}_i, i = 1, \ldots, 50)$ τους δεν κατάφεραν να προσδιοριστούν από τη μέθοδο των ελαχίστων τετραγώνων. Επιπλέον, για κανένα συντελεστή β_i δεν υπολογίστηκαν τα αντίστοιχα τυπικά σφάλματα $se(\hat{\beta}_i) = \sqrt{Var(\beta_i)} = \sqrt{\sigma^2(\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}})^{-1}_{ii}}, i = 1, \ldots, 50$, με αποτέλεσμα οι στατιστικοί έλεγχοι υποθέσεων $(H_0 : \beta_i = 0, H_1 : \beta i \neq 0)$ των συντελεστών να

είναι ανέφικτοι.

- Ο στατιστικός έλεγχος F που ελέγχει εάν όλες οι επεξηγηματικές μεταβλητές είναι μη στατιστικά σημαντικές (με εναλλακτική ότι κάποια είναι) δεν μπορεί να πραγματοποιηθεί. Ο συντελεστής προσδιορισμού R^2 είναι ίσος με ένα, ενώ ο διορθωμένος συντελεστής R^2_{adj} δεν ορίζεται, φαινόμενα που ενισχύουν τις μεγάλες αμφιβολίες μας για την προσαρμογή του μοντέλου.

Λόγω της απροσδιοριστίας των τυπικών σφαλμάτων των εκτιμητών των συντελεστών του μοντέλου βλέπουμε ότι ο πίνακας $\tilde{\boldsymbol{X}}^T\tilde{\boldsymbol{X}}$ είναι μη αντιστρέψιμος με αποτέλεσμα να έχει ορίζουσα ίση με το μηδέν και κάποιες στήλες-επεξηγηματικές μεταβλητές του να είναι γραμμικά συσχετισμένες. Καλούμαστε να αντιμετωπίσουμε λοιπόν το φαινόμενο της πολυσυγγραμικότητας παράλληλα με την εύρεση του πραγματικού μοντέλου που γέννησε τα δεδομένα.

Σε κάθε ένα από τα 100 τυχαία δείγματα προσομοιωμένων δεδομένων εκτελούμε τα ακόλουθα:

- Εφαρμόζουμε την διαδικασία διαδοχικής πρόσθεσης και την διαδικασία κατά βήματα. Η διαδικασία της διαδοχικής αφαίρεσης είναι αδύνατο να χρησιμοποιηθεί καθώς το πλήρες μοντέλο, που αποτελεί το μοντέλο εκκίνησής της, δεν είναι πλήρως προσδιορισμένο. Η διαδικασία κατά βήματα ξεκινάει με το μοντέλο που περιέχει μόνο τον σταθερό όρο.

- Αξιοποιούμε τη μέθοδο ποινικοποίησης Lasso με παράμετρο ποινής που βρίσκεται εφαρμόζοντας 5-fold cross validation με τον «one standard error» κανόνα.

Το πραγματικό μοντέλο που γέννησε τα δεδομένα και περιέχει τις μεταβλητές X_1, X_2, X_{10} ανιχνεύεται από τις μεθόδους με τις εξής συχνότητες (βλ. Πίνακα 6.3):

Μέθοδος	Σχετική συχνότητα
Lasso	0.13
Stepboth	0
$Stepforward$	0

Πίνακας 6.3: Σχετικές συχνότητες επιλογής του πραγματικού μοντέλου από τις μεθόδους της προσομοίωσης 2.

Επιπρόσθετα, παρουσιάζουμε τις σχετικές συχνότητες επιλογής μοντέλων με το πολύ 5 επεξηγηματικές μεταβλητές στις οποίες συμπεριλαμβάνονται οι μεταβλητές X_1, X_2, X_{10} του πραγματικού μοντέλου (βλ. Πίνακα 6.4):

Μέθοδος	Σχετική συχνότητα
Lasso	0.71
Stepboth	0.02
$Stepforward$	0.02

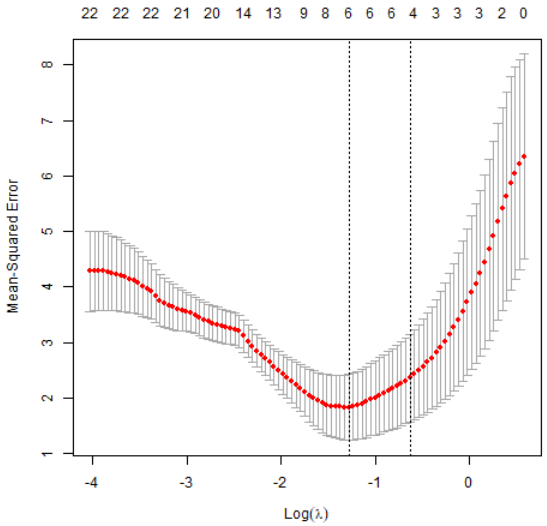
Πίνακας 6.4: Σχετικές συχνότητες επιλογής μοντέλων με το πολύ 5 επεξηγηματικές μεταβλητές συμπεριλαμβανομένων των (X_1, X_2, X_{10}) από τις μεθόδους της προσομοίωσης 2.

Με βάση τον Πίνακα 6.3 παρατηρούμε ότι η μέθοδος Lasso εντοπίζει το πραγματικό μοντέλο που γέννησε τα δεδομένα 13 στις 100 φορές, ενώ οι δύο μέθοδοι κατά βήματα (Stepboth, Stepforward) δεν καταφέρνουν ποτέ να εντοπίσουν το σωστό μοντέλο.

Επιπλέον, όπως φαίνεται στον Πίνακα 6.4, με τη μέθοδο Lasso επιλέγονται αρκετά συχνά φειδωλά μοντέλα που περιέχουν τις επεξηγηματικές μεταβλητές X_1, X_2, X_{10} του πραγματικού μοντέλου. Συγκεκριμένα επιλέγονται τέτοια μοντέλα με το πολύ 5 μεταβλητές 71 στις 100 φορές. Οι Stepwise τεχνικές καταφέρνουν μόλις 2 στις 100 φορές να βρίσκουν αυτά τα μοντέλα.

Βλέπουμε λοιπόν ότι η μέθοδος Lasso συρρικνώνει προς το μηδέν ένα πολύ μεγάλο ποσοστό των 50 ανεξάρτητων μεταβλητών επιτυγχάνοντας ταυτόχρονα την ανίχνευση των 3 μεταβλητών που πραγματικά συνεισφέρουν στο μοντέλο που γέννησε τα δεδομένα. Η ακρίβεια του συμπεράσματος αυτού μπορεί να αξιολογηθεί περαιτέρω χρησιμοποιώντας το διάγραμμα των εκτιμητών των συντελεστών $\hat{\boldsymbol{\beta}}$ συναρτήσει των τιμών της παραμέτρου ποινής λ της μεθόδου Lasso. Αρχικά δημιουργούμε την παράσταση του μέσου 5-fold cross validated σφάλματος για διάφορες τιμές της παραμέτρου ποινής λ σε ένα εκ των 100 προσομοιωμένων τυχαίων δειγμάτων δεδομένων.

Στο Διάγραμμα 6.1 απεικονίζεται το μέσο cross validated σφάλμα που προκύπτει για μια ακολουθία τιμών της παραμέτρου ποινής λ οι τιμές της οποίας είναι λογαριθμοποιημένες στο σχήμα. Από αριστερά προς τα δεξιά φαίνεται ο αριθμός των επεξηγηματικών μεταβλητών που συμμετέχουν στο γραμμικό μοντέλο που προσαρμόζει η μέθοδος Lasso για αντίστοιχες τιμές της παραμέτρου λ και οι οποίες μειώνονται όσο αυξάνεται η παράμετρος. Η πρώτη από αριστερά κάθετη γραμμή αντιστοιχεί στην τιμή της παραμέτρους λ στην οποία εμφανίζεται το ελάχιστο μέσο cross validated σφάλμα, ενώ στη δεύτερη κάθετη εντοπίζεται η παράμετρος ποινικοποίησης της οποίας το μέσο cross

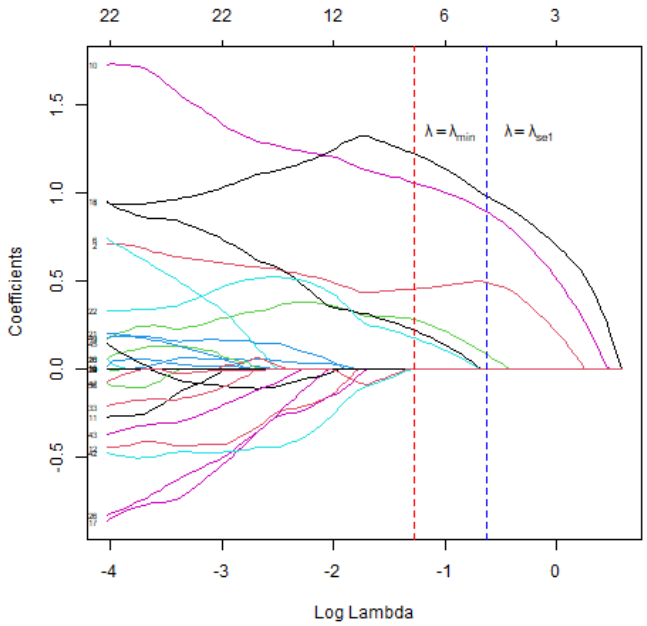


Διάγραμμα 6.1: Μέσο cross validated σφάλμα συναρτήσει διάφορων τιμών της παραμέτρου λ της μεθόδου lasso σε δεδομένα της προσομοίωσης 2.

validated σφάλμα απέχει το πολύ ένα τυπικό σφάλμα από το ελάχιστο μέσο cross validated σφάλμα («one standard error rule») και την οποία διαλέγουμε καθώς οδηγεί σε φειδωλότερα μοντέλα. Τα διαστήματα εμπιστοσύνης για κάθε mean cross-validated τιμή σφάλματος υπολογίζονται με βάση τη σχέση: $[MSE_\lambda - SE(\lambda), MSE_\lambda + SE(\lambda)]$.

Επιπροσθέτως, σχεδιάζουμε το διάγραμμα των εκτιμητών των συντελεστών του μοντέλου που προτείνει η μέθοδος Lasso για διάφορες τιμές της παραμέτρου λ.

Στο Διάγραμμα 6.2 η κόκκινη κάθετη γραμμή αντιστοιχεί στην επιλογή της παραμέτρου ποινής που ελαχιστοποιεί το μέσο 5-fold cross validated σφάλμα με αποτέλεσμα να επιλέγονται 6 μεταβλητές στο μοντέλο συμπεριλαμβανόμενων των X_1, X_2, X_{10}. Η μπλε κάθετη γραμμή αντιστοιχεί στην μεγαλύτερη τιμή της παραμέτρου που απέχει το πολύ ένα τυπικό σφάλμα από το ελάχιστο μέσο cross validated σφάλμα και οδηγεί στην επιλογή 4 μεταβλητών μαζί με τις μεταβλητές X_1, X_2, X_{10} του πραγματικού μοντέλου.



Διάγραμμα 6.2: Εκτιμητές των συντελεστών γραμμικής παλινδρόμησης συναρτήσει μίας ακολουθίας τιμών της παραμέτρου λ εφαρμόζοντας της μέθοδο Lasso σε δεδομένα της $2^{ης}$ προσομοίωσης.

Παράρτημα Α΄

BC_a Διαστήματα εμπιστοσύνης

Τα BC_a (Bias Correction and Acceleration) διαστήματα εμπιστοσύνης χρησιμοποιήθηκαν στις cross validation και bootstrap υπολογιστικές μεθόδους εκτίμησης του προβλεπτικού σφάλματος μοντέλων της Παραγράφου 4.5.

Υποθέτουμε ότι διαθέτουμε παραδείγματος χάριν τους $K = 10$ όρους από τους οποίους προκύπτει η 10-fold cross validation εκτιμήτρια του σφάλματος $CV_{10} = \frac{1}{10}\sum_{k=1}^{10} MSE(T_k)$ ενός μοντέλου. Δημιουργούμε στη συνέχεια πολλά ($B \geq 10000$) bootstrap στοιχεία διαλέγοντας τυχαία με επανάθεση από τα $K = 10$ αρχικά στοιχεία και από κάθε bootstrap δείγμα κρατάμε τον μέσο όρο έστω $\theta^{*b} = CV_{10}^{*b}, b = 1, \ldots, B$, σχηματίζοντας με αυτό τον τρόπο την λεγόμενη bootstrap κατανομή.

Τα BC_a διαστήματα εμπιστοσύνης έχουν τη γενική μορφή:

$$\left[\theta^{*(\alpha_1)}, \theta^{*(\alpha_2)}\right]$$

Ο όρος $\theta^{*(a)}$ είναι το α-ποσοστιαίο σημείο της bootstrap κατανομής των νέων προσομοιωμένων τιμών. Οι ποσότητες α_1, α_2 υπολογίζονται με βάση τους τύπους:

$$\alpha_1 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(\alpha)}}{1 - \hat{\alpha}(\hat{z}_0 + z^{(\alpha)})}\right),$$

και

$$\alpha_2 = \Phi\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(1-\alpha)}}{1 - \hat{\alpha}(\hat{z}_0 + z^{(1-\alpha)})}\right).$$

Συμβολίζουμε με $z^{(\alpha)}$ το α ποσοστιαίο σημείο της τυποποιημένης κανονικής κατανομής και $\Phi(\alpha)$ τη συνάρτηση κατανομής της τυποποιημένης κανονικής κατανομής. Επομένως θα ισχύει ότι $\Phi^{-1}(\alpha) = z^{(\alpha)}$. Οι ποσότητες $\hat{z}_0, \hat{\alpha}$ είναι άγνωστες και πρέπει να εκτιμηθούν. Η πρώτη διορθώνει την ενδεχόμενη μεροληψία της εκτιμήτριας του προβλεπτικού σφάλματος και η δεύτερη διορθώνει

ως προς την απόκλιση από την κανονική κατανομή. Οι τύποι υπολογισμού τους είναι οι εξής:

$$\hat{z}_0 = \Phi\left(\frac{\#\hat{\theta}_i^* < \hat{\theta}}{B}\right),$$

Ο όρος $\#\hat{\theta}_i^* < \hat{\theta}$ είναι ο αριθμός των bootstrap τιμών που είναι μικρότερες της εκτιμήτριας $\hat{\theta}$ του σφάλματος. Για την παράμετρο $\hat{\alpha}$ (acceleration parameter) έχουμε:

$$\hat{a} = \frac{\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^3}{6\left[\sum_{i=1}^{n}\left(\hat{\theta}_{(.)} - \hat{\theta}_{(i)}\right)^2\right]^{3/2}}$$

Με $\hat{\theta}_i$ συμβολίζεται η jackknife εκτίμηση του προβλεπτικού σφάλματος που προκύπτει αφαιρώντας την i παρατήρηση από τις N αρχικές. $\hat{\theta}_{(.)}$ είναι ο μέσος όρος των jackknife εκτιμήσεων.

Παράρτημα Β΄

Κώδικες Δεδομένων - R

Παράγραφος 1.3 - K-L απόσταση:

```
#exponential distributions K-L divergence calculation:


x = seq(0, 5, by =0.0005 )
#exponential true model:
ftrue = dexp(x,rate=1.5)

#exponential 1:
g1=dexp(x,rate=0.5)

#exponential 2:
g2=dexp(x,rate=1)

#exponential 3:
g3=dexp(x,rate=2)

#exponential 4:
g4=dexp(x,rate=1.15)

plot(x,ftrue,col="blue",lwd=1)
lines(x,g1,col="orange",lwd=3)
lines(x,g2,col="green",lwd=3)
lines(x,g3,col="red",lwd=3)
lines(x,g4,col="purple",lwd=3)
legend(3,1, legend=c("f",expression(g[1]),expression(g[2]),expression(g[3]),expression(g[4])),
       col=c("blue","orange","green","red","purple"), lty=1:1, cex=0.8)
#K-L divergence of f-g for the specific x (in [0,5]) values:
KLdiv=function(λ,κ){
  return(  log(λ/κ)-(-(-λ+κ)/1.5 )  )
}

#KL f,g1:
KLdiv(1.5,0.5)
#KL f,g2:
KLdiv(1.5,1)
#KL f,g3:
KLdiv(1.5,2)
#KL f,g4:
KLdiv(1.5,1.15)
```

Παράγραφος 2.6 - Δεδομένα διάρκειας ζωής:

```
## Δεδομένα παραδείγματος  ρουλεμάν - εκατομμύρια στροφές # #
Y<-c(17.88,28.92,33.00,41.52,42.12,45.60,48.48,51.84,51.96,54.12,55.56,
     67.80,67.80,67.80,68.64,68.64,68.88,84.12,93.12,98.64,105.12,105.84,127.92,128.04,173.40)
hist(Y,freq=F)


## Newton-Raphson: weibull model

funf<-function(x,Y){     # # f(η) # #
  s1<-0
  s2<-0
  for (i in 1:length(Y)){
    s1<-s1+(Y[i])^(x)*log(Y[i])
    s2<-s2+(Y[i])^(x)
  }
  s3<-log(Y[1])+log(Y[2])+log(Y[3])+log(Y[4])+log(Y[5])+log(Y[6])+log(Y[7])
  +log(Y[8])+log(Y[9])+log(Y[10])+log(Y[11])+log(Y[12])+log(Y[15])+log(Y[18])
  +log(Y[19])+log(Y[20])+log(Y[21])+log(Y[23])+log(Y[24])   #μη αποκομμενα δεδομενα

  f<-(s1/s2)-x^(-1)-s3/19
  return(f)
}

funfdev<-function(x,Y){   # # f'(η) # #
  s1<-0
  s2<-0
  s3<-0
  for (i in 1:length(Y)){
    s1<-s1+(Y[i])^(x)*(log(Y[i]))^2
    s2<-s2+(Y[i])^(x)
    s3<-s3+(Y[i])^(x)*log(Y[i])
  }
  f<-(s1/s2)-((s3)^(2)/(s2)^(2))+x^(-2)
  return(f)
}

## N R method # #
ηk<-1.5                #αρχική τιμή η0 που προκύπτει δοκιμάζοντας τιμές στην f(η) και παρατηρώντας τα πρόσημά της
iter<-1
ηklist<-c()
for (i in 1:1000){
  ηk1<-ηk-funf(ηk,Y)/funfdev(ηk,Y)
  ηklist[i]<-ηk1
  if (abs(ηk1-ηk)<10^(-5)){
    return(ηk1)
  }
  ηk<-ηk1
  iter<-iter+1
}
```

```
##τιμή μεγιστοποιημένης λογαριθμικής πιθανοφάνειας
#υπολογισμός ΕΜΠ της α:
s<-0
for (i in 1:length(Y)){
  s<-s+(Y[i])^(ηk1)
}
α<-(s/19)^(1/ηk1)
s1<-log(Y[1])+log(Y[2])+log(Y[3])+log(Y[4])+log(Y[5])+log(Y[6])+log(Y[7])
+log(Y[8])+log(Y[9])+log(Y[10])+log(Y[11])+log(Y[12])+log(Y[15])+log(Y[18])
+log(Y[19])+log(Y[20])+log(Y[21])+log(Y[23])+log(Y[24])

## μεγιστοποιημένη πιθανοφάνεια weibull μοντέλου:
loglike<-19*log(ηk1)-19*ηk1*log(α)+(ηk1-1)*s1-(s/α^(ηk1))

##τιμή BIC weibull:
BICweib<- -2*loglike +log(25)*2

# # Εκθετικό μοντέλο για τα δεδομένα 1.1.1 :
sum(Y)
λ<-19/sum(Y)
BICexp<- -2*(19*log(19)-19*log(1796.76)-19)+log(25)

## AIC για εκθετικό και weibull μοντέλο: ##
AICexp<- -2*(19*log(19)-19*log(1796.76)-19)+2*1
AICweib<- -2*(loglike)+2*2

## Weibull Hessian matrix and eigenvalues:
alphahat=91.6383
etahat=1.9467
s1<-0
s2<-0
s3<-(log(Y[1])+log(Y[2])+log(Y[3])+log(Y[4])+log(Y[5])+log(Y[6])+log(Y[7])
     +log(Y[8])+log(Y[9])+log(Y[10])+log(Y[11])+log(Y[12])+log(Y[15])
     +log(Y[18])+log(Y[19])+log(Y[20])+log(Y[21])+log(Y[23])+log(Y[24]))
for (i in 1:length(Y)){
  s1<-s1+((Y[i])^(etahat))*log(Y[i])
  s2<-s2+(Y[i])^(etahat)
}

s5=0
for (i in 1:length(Y)){
  s5<-s5+(Y[i])^(etahat)*((log(Y[i]/alphahat))^(2))
}

s4=0
for (i in 1:length(Y)) {
  s4=s4+(Y[i]/alphahat)^(etahat)*((log(Y[i]/alphahat))^(2))
}

Hessian=matrix(0,ncol = 2,nrow = 2)
Hessian[1,1]=(19*etahat)/(alphahat^2)-etahat*(etahat+1)*alphahat^(-etahat-2)*s2
Hessian[1,2]=-19/alphahat+s2/(alphahat^(etahat+1))+(etahat*s5)/(alphahat^(etahat+1))
Hessian[2,1]=(-(19/alphahat)+etahat*(alphahat^(-etahat-1))*s1
             +(alphahat^(-etahat-1))*s2-etahat*alphahat^(-etahat-1)*log(alphahat)*s2)
Hessian[2,2]=-19/(etahat^2)-s4
eigen(Hessian)
```

Παράγραφος 4.5:

```
#####  Simulated data for polynomial regression prediction errors :-------------
require(boot)
require(bootstrap)
require(MASS)

X=runif(100,0,20)
epsilon=rnorm(100,0,350)
Y=500+(X-10)^3+epsilon
dat=cbind(X,Y)
dat=as.data.frame(dat)

X=dat[,1]
Y=dat[,2]
png(filename="polysimdata.png")
plot(X,Y,main="Data")
dev.off()
Xtest=seq(0,20,by=0.05)

# M1: simple linear model:
png(filename="polysim_models1.png")
par(mfrow=c(2,2))
lm1=lm(Y~.,data = dat)
plot(X,Y,main = "M1")
abline(lm1,col="orange",lwd=3)

# M2
pm2=lm(Y~X+I(X^2),data = dat)
y_hatpm2=predict(pm2,list(X=Xtest))
plot(X,Y,main = "M2")
lines(Xtest,y_hatpm2,col="blue",lwd=3)

# M3
pm3=lm(Y~X+I(X^2)+I(X^3),data = dat)
y_hatpm3=predict(pm3,list(X=Xtest))
plot(X,Y,main = "M3")
lines(Xtest,y_hatpm3,col="green",lwd=3)

# M4
pm4=lm(Y~X+I(X^2)+I(X^3)+I(X^4),data = dat)
y_hatpm4=predict(pm4,list(X=Xtest))
plot(X,Y,main = "M4")
lines(Xtest,y_hatpm4,col="darkorange",lwd=3)
dev.off()

png(filename="polysim_models2.png")
par(mfrow=c(2,2))
# M5
pm5=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5),data = dat)
y_hatpm5=predict(pm5,list(X=Xtest))
plot(X,Y,main = "M5")
lines(Xtest,y_hatpm5,col="coral",lwd=3)

# M6
pm6=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5)+I(X^6),data = dat)
y_hatpm6=predict(pm6,list(X=Xtest))
plot(X,Y,main = "M6")
lines(Xtest,y_hatpm6,col="brown2",lwd=3)
```

```
# M7
pm7=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5)+I(X^6)+I(X^7),data = dat)
y_hatpm7=predict(pm7,list(X=Xtest))
plot(X,Y,main = "M7")
lines(Xtest,y_hatpm7,col="darkred",lwd=3)
dev.off()

# MSE error:
MSEerrlm1=sum((lm1$residuals)^2)/100
MSEerrpm2=sum((pm2$residuals)^2)/100
MSEerrpm3=sum((pm3$residuals)^2)/100
MSEerrpm4=sum((pm4$residuals)^2)/100
MSEerrpm5=sum((pm5$residuals)^2)/100
MSEerrpm6=sum((pm6$residuals)^2)/100
MSEerrpm7=sum((pm7$residuals)^2)/100


#Leave one out CV:---
# lm1:
looCVse=c(rep(0,7))


SEk=c(rep(0,100))
for (k in 1:100) {
  train=dat[-k,]
  test=dat[k,]
  mtrain=lm(Y~X,data = train)
  predictions=predict(mtrain,newdata=test)
  SEk[k]=sum((test$Y-predictions)^2)
}
LooCVerrlm1=sum(SEk)/100
testbcalm1cv=bcanon(SEk,10000,mean,alpha = c(0.025,0.975))
CI.lm1Loocv=testbcalm1cv$confpoints
CI.lm1Loocv=as.matrix(CI.lm1Loocv)
looCVse[1]=sd(SEk)/sqrt(100)

# pm2:
SEk=c(rep(0,100))
for (k in 1:100) {
  train=dat[-k,]
  test=dat[k,]
  mtrain=lm(Y~X+I(X^2),data = train)
  predictions=predict(mtrain,newdata=test)
  SEk[k]=sum((test$Y-predictions)^2)
}
LooCVerrpm2=sum(SEk)/100
testbcapm2cv=bcanon(SEk,10000,mean,alpha = c(0.025,0.975))
CI.pm2Loocv=testbcapm2cv$confpoints
CI.pm2Loocv=as.matrix(CI.pm2Loocv)
looCVse[2]=sd(SEk)/sqrt(100)
```

```
# pm3:
SEk=c(rep(0,100))
for (k in 1:100) {
  train=dat[-k,]
  test=dat[k,]
  mtrain=lm(Y~X+I(X^2)+I(X^3),data = train)
  predictions=predict(mtrain,newdata=test)
  SEk[k]=sum((test$Y-predictions)^2)
}
LooCVerrpm3=sum(SEk)/100
testbcapm3cv=bcanon(SEk,10000,mean,alpha = c(0.025,0.975))
CI.pm3Loocv=testbcapm3cv$confpoints
CI.pm3Loocv=as.matrix(CI.pm3Loocv)
looCVse[3]=sd(SEk)/sqrt(100)


# pm4:
SEk=c(rep(0,100))
for (k in 1:100) {
  train=dat[-k,]
  test=dat[k,]
  mtrain=lm(Y~X+I(X^2)+I(X^3)+I(X^4),data = train)
  predictions=predict(mtrain,newdata=test)
  SEk[k]=sum((test$Y-predictions)^2)
}
LooCVerrpm4=sum(SEk)/100
testbcapm4cv=bcanon(SEk,10000,mean,alpha = c(0.025,0.975))
CI.pm4Loocv=testbcapm4cv$confpoints
CI.pm4Loocv=as.matrix(CI.pm4Loocv)
looCVse[4]=sd(SEk)/sqrt(100)


# pm5:
SEk=c(rep(0,100))
for (k in 1:100) {
  train=dat[-k,]
  test=dat[k,]
  mtrain=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5),data = train)
  predictions=predict(mtrain,newdata=test)
  SEk[k]=sum((test$Y-predictions)^2)
}
LooCVerrpm5=sum(SEk)/100
testbcapm5cv=bcanon(SEk,10000,mean,alpha = c(0.025,0.975))
CI.pm5Loocv=testbcapm5cv$confpoints
CI.pm5Loocv=as.matrix(CI.pm5Loocv)
looCVse[5]=sd(SEk)/sqrt(100)
```

```
# pm6:
SEk=c(rep(0,100))
for (k in 1:100) {
  train=dat[-k,]
  test=dat[k,]
  mtrain=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5)+I(X^6),data = train)
  predictions=predict(mtrain,newdata=test)
  SEk[k]=sum((test$Y-predictions)^2)
}
LooCVerrpm6=sum(SEk)/100
testbcapm6cv=bcanon(SEk,10000,mean,alpha = c(0.025,0.975))
CI.pm6Loocv=testbcapm6cv$confpoints
CI.pm6Loocv=as.matrix(CI.pm6Loocv)
looCVse[6]=sd(SEk)/sqrt(100)


# pm7:
SEk=c(rep(0,100))
for (k in 1:100) {
  train=dat[-k,]
  test=dat[k,]
  mtrain=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5)+I(X^6)+I(X^7),data = train)
  predictions=predict(mtrain,newdata=test)
  SEk[k]=sum((test$Y-predictions)^2)
}
LooCVerrpm7=sum(SEk)/100
testbcapm7cv=bcanon(SEk,10000,mean,alpha = c(0.025,0.975))
CI.pm7Loocv=testbcapm7cv$confpoints
CI.pm7Loocv=as.matrix(CI.pm7Loocv)
looCVse[7]=sd(SEk)/sqrt(100)




# 10fold cv:---
fold10CVse=c(rep(0,7))
#10fold lm1:
mseTk=c(rep(0,10))
for (k in 1:10) {
  smp_size=floor(0.9*nrow(dat))
  train_ind=sample(seq_len(nrow(dat)),size = smp_size)
  train=dat[train_ind,]
  test=dat[-train_ind,]
  mtrain=lm(Y~.,data = train)
  predictions=predict(mtrain,newdata=test)
  mseTk[k]=sum((test$Y-predictions)^2)/10
}
fold10CVlm1=sum(mseTk)/10
testbcalm1_10foldcv=bcanon(mseTk,10000,mean,alpha = c(0.025,0.975))
CI.lm1_10foldcv=testbcalm1_10foldcv$confpoints
CI.lm1_10foldcv=as.matrix(CI.lm1_10foldcv)
fold10CVse[1]=sd(mseTk)/sqrt(10)
```

```r
#10fold pm2:
mseTk=c(rep(0,10))
for (k in 1:10) {
  smp_size=floor(0.9*nrow(dat))
  train_ind=sample(seq_len(nrow(dat)),size = smp_size)
  train=dat[train_ind,]
  test=dat[-train_ind,]
  mtrain=lm(Y~X+I(X^2),data = train)
  predictions=predict(mtrain,newdata=test)
  mseTk[k]=sum((test$Y-predictions)^2)/10
}
fold10CVpm2=sum(mseTk)/10
testbcapm2_10foldcv=bcanon(mseTk,10000,mean,alpha = c(0.025,0.975))
CI.pm2_10foldcv=testbcapm2_10foldcv$confpoints
CI.pm2_10foldcv=as.matrix(CI.pm2_10foldcv)
fold10CVse[2]=sd(mseTk)/sqrt(10)


#10fold pm3:
mseTk=c(rep(0,10))
for (k in 1:10) {
  smp_size=floor(0.9*nrow(dat))
  train_ind=sample(seq_len(nrow(dat)),size = smp_size)
  train=dat[train_ind,]
  test=dat[-train_ind,]
  mtrain=lm(Y~X+I(X^2)+I(X^3),data = train)
  predictions=predict(mtrain,newdata=test)
  mseTk[k]=sum((test$Y-predictions)^2)/10
}
fold10CVpm3=sum(mseTk)/10
testbcapm3_10foldcv=bcanon(mseTk,10000,mean,alpha = c(0.025,0.975))
CI.pm3_10foldcv=testbcapm3_10foldcv$confpoints
CI.pm3_10foldcv=as.matrix(CI.pm3_10foldcv)
fold10CVse[3]=sd(mseTk)/sqrt(10)



#10fold pm4:
mseTk=c(rep(0,10))
for (k in 1:10) {
  smp_size=floor(0.9*nrow(dat))
  train_ind=sample(seq_len(nrow(dat)),size = smp_size)
  train=dat[train_ind,]
  test=dat[-train_ind,]
  mtrain=lm(Y~X+I(X^2)+I(X^3)+I(X^4),data = train)
  predictions=predict(mtrain,newdata=test)
  mseTk[k]=sum((test$Y-predictions)^2)/10
}
fold10CVpm4=sum(mseTk)/10
testbcapm4_10foldcv=bcanon(mseTk,10000,mean,alpha = c(0.025,0.975))
CI.pm4_10foldcv=testbcapm4_10foldcv$confpoints
CI.pm4_10foldcv=as.matrix(CI.pm4_10foldcv)
fold10CVse[4]=sd(mseTk)/sqrt(10)
```

```
#10fold pm5:
mseTk=c(rep(0,10))
for (k in 1:10) {
  smp_size=floor(0.9*nrow(dat))
  train_ind=sample(seq_len(nrow(dat)),size = smp_size)
  train=dat[train_ind,]
  test=dat[-train_ind,]
  mtrain=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5),data = train)
  predictions=predict(mtrain,newdata=test)
  mseTk[k]=sum((test$Y-predictions)^2)/10
}
fold10CVpm5=sum(mseTk)/10
testbcapm5_10foldcv=bcanon(mseTk,10000,mean,alpha = c(0.025,0.975))
CI.pm5_10foldcv=testbcapm5_10foldcv$confpoints
CI.pm5_10foldcv=as.matrix(CI.pm5_10foldcv)
fold10CVse[5]=sd(mseTk)/sqrt(10)


#10fold pm6:
mseTk=c(rep(0,10))
for (k in 1:10) {
  smp_size=floor(0.9*nrow(dat))
  train_ind=sample(seq_len(nrow(dat)),size = smp_size)
  train=dat[train_ind,]
  test=dat[-train_ind,]
  mtrain=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5)+I(X^6),data = train)
  predictions=predict(mtrain,newdata=test)
  mseTk[k]=sum((test$Y-predictions)^2)/10
}
fold10CVpm6=sum(mseTk)/10
testbcapm6_10foldcv=bcanon(mseTk,10000,mean,alpha = c(0.025,0.975))
CI.pm6_10foldcv=testbcapm6_10foldcv$confpoints
CI.pm6_10foldcv=as.matrix(CI.pm6_10foldcv)
fold10CVse[6]=sd(mseTk)/sqrt(10)


#10fold pm7:
mseTk=c(rep(0,10))
for (k in 1:10) {
  smp_size=floor(0.9*nrow(dat))
  train_ind=sample(seq_len(nrow(dat)),size = smp_size)
  train=dat[train_ind,]
  test=dat[-train_ind,]
  mtrain=lm(Y~X+I(X^2)+I(X^3)+I(X^4)+I(X^5)+I(X^6)+I(X^7),data = train)
  predictions=predict(mtrain,newdata=test)
  mseTk[k]=sum((test$Y-predictions)^2)/10
}
fold10CVpm7=sum(mseTk)/10
testbcapm7_10foldcv=bcanon(mseTk,10000,mean,alpha = c(0.025,0.975))
CI.pm7_10foldcv=testbcapm7_10foldcv$confpoints
CI.pm7_10foldcv=as.matrix(CI.pm7_10foldcv)
fold10CVse[7]=sd(mseTk)/sqrt(10)
```

```
#Generalized Cross Validation error GCV:
#GCV for lm1 :
sum=0
GCVelemlm1=c()
for (i in 1:nrow(dat)) {
  GCVelemlm1[i]=( lm1$residuals[i]/(1-length(names(coef(lm1)))/nrow(dat)) )^2
  sum=sum+( lm1$residuals[i]/(1-length(names(coef(lm1)))/nrow(dat)) )^2
}
GCVerrlm1=sum/100
GCVlm1se=sd(GCVelemlm1)/sqrt(100)
testbcalm1_GCV=bcanon(GCVelemlm1,10000,mean,alpha = c(0.025,0.975))
CI.lm1_GCV=testbcalm1_GCV$confpoints
CI.lm1_GCV=as.matrix(CI.lm1_GCV)

#GCV for pm2 :
sum=0
GCVelempm2=c()
for (i in 1:nrow(dat)) {
  GCVelempm2[i]=( pm2$residuals[i]/(1-length(names(coef(pm2)))/nrow(dat)) )^2
  sum=sum+( pm2$residuals[i]/(1-length(names(coef(pm2)))/nrow(dat)) )^2
}
GCVerrpm2=sum/100
GCVpm2se=sd(GCVelempm2)/sqrt(100)
testbcapm2_GCV=bcanon(GCVelempm2,10000,mean,alpha = c(0.025,0.975))
CI.pm2_GCV=testbcapm2_GCV$confpoints
CI.pm2_GCV=as.matrix(CI.pm2_GCV)

#GCV for pm3 :
sum=0
GCVelempm3=c()
for (i in 1:nrow(dat)) {
  GCVelempm3[i]=( pm3$residuals[i]/(1-length(names(coef(pm3)))/nrow(dat) ))^2
  sum=sum+( pm3$residuals[i]/(1-length(names(coef(pm3)))/nrow(dat) ))^2
}
GCVerrpm3=sum/100
GCVpm3se=sd(GCVelempm3)/sqrt(100)
testbcapm3_GCV=bcanon(GCVelempm3,10000,mean,alpha = c(0.025,0.975))
CI.pm3_GCV=testbcapm3_GCV$confpoints
CI.pm3_GCV=as.matrix(CI.pm3_GCV)

#GCV for pm4 :
sum=0
GCVelempm4=c()
for (i in 1:nrow(dat)) {
  GCVelempm4[i]=( pm4$residuals[i]/(1-length(names(coef(pm4)))/nrow(dat)) )^2
  sum=sum+( pm4$residuals[i]/(1-length(names(coef(pm4)))/nrow(dat)) )^2
}
GCVerrpm4=sum/100
GCVpm4se=sd(GCVelempm4)/sqrt(100)
testbcapm4_GCV=bcanon(GCVelempm4,10000,mean,alpha = c(0.025,0.975))
CI.pm4_GCV=testbcapm4_GCV$confpoints
CI.pm4_GCV=as.matrix(CI.pm4_GCV)
```

```
#GCV for pm5 :
sum=0
GCVelempm5=c()
for (i in 1:nrow(dat)) {
  GCVelempm5[i]=( pm5$residuals[i]/(1-length(names(coef(pm5)))/nrow(dat)) )^2
  sum=sum+( pm5$residuals[i]/(1-length(names(coef(pm5)))/nrow(dat)) )^2
}
GCVerrpm5=sum/100
GCVpm5se=sd(GCVelempm5)/sqrt(100)
testbcapm5_GCV=bcanon(GCVelempm5,10000,mean,alpha = c(0.025,0.975))
CI.pm5_GCV=testbcapm5_GCV$confpoints
CI.pm5_GCV=as.matrix(CI.pm5_GCV)


#GCV for pm6 :
sum=0
GCVelempm6=c()
for (i in 1:nrow(dat)) {
  GCVelempm6[i]=( pm6$residuals[i]/(1-length(names(coef(pm6)))/nrow(dat)) )^2
  sum=sum+( pm6$residuals[i]/(1-length(names(coef(pm6)))/nrow(dat)) )^2
}
GCVerrpm6=sum/100
GCVpm6se=sd(GCVelempm6)/sqrt(100)
testbcapm6_GCV=bcanon(GCVelempm6,10000,mean,alpha = c(0.025,0.975))
CI.pm6_GCV=testbcapm6_GCV$confpoints
CI.pm6_GCV=as.matrix(CI.pm6_GCV)


#GCV for pm7 :
sum=0
GCVelempm7=c()
for (i in 1:nrow(dat)) {
  GCVelempm7[i]=( pm7$residuals[i]/(1-length(names(coef(pm7)))/nrow(dat)) )^2
  sum=sum+( pm7$residuals[i]/(1-length(names(coef(pm7)))/nrow(dat)) )^2
}
GCVerrpm7=sum/100
GCVpm7se=sd(GCVelempm7)/sqrt(100)
testbcapm7_GCV=bcanon(GCVelempm7,10000,mean,alpha = c(0.025,0.975))
CI.pm7_GCV=testbcapm7_GCV$confpoints
CI.pm7_GCV=as.matrix(CI.pm7_GCV)
```

```
# leave one out boot err code and .632 boot err :
loobooterrse=c(rep(0,7))
boot632se=c(rep(0,7))
#lm1 LooBoot error:
X=dat[,1]
datb=cbind(X,Y)
datb=as.data.frame(datb)
SEb=matrix(0,nrow = 100,ncol = 200)
C=c(rep(0,100))

for (b in 1:200) {
  smp_size=100
  bootsample_ind=sample(seq_len(nrow(datb)),size = smp_size,replace = TRUE)
  bootsample=datb[bootsample_ind,]
  mboot=lm(Y~.,data=bootsample)
  for (i in 1:100) {
    in_index=1
    for (z in 1:100) {
      if(Y[i]==bootsample$Y[z]){
        in_index=0
      }
    }
    if(in_index==1){
      C[i]=C[i]+1
      SEb[i,b]=( Y[i]-predict(mboot,list(X=datb[i,1])) )^2
    }
  }
}
sumN=0
for (i in 1:100) {
  sumN=sumN+sum(SEb[i,])*(C[i])^(-1)
}
LooBootlm1=sumN/100
#for  LooBoot lm1 confidence interval:
sumbC=c(rep(0,100))
for (i in 1:100) {
  sumbC[i]=sum(SEb[i,])*(C[i])^(-1)
}
testbcalm1_LooBooterr=bcanon(sumbC,10000,mean,alpha = c(0.025,0.975))
CI.lm1_LooBooterr=testbcalm1_LooBooterr$confpoints
CI.lm1_LooBooterr=as.matrix(CI.lm1_LooBooterr)
loobooterrse[1]=sd(sumbC)/sqrt(100)

#lm1 .632 boot:
apperrlm1=sum((lm1$residuals)^2)/100
Err632bootlm1=0.368*apperrlm1+0.632*LooBootlm1
elements632=c(rep(0,100))
for (i in 1:100) {
  elements632[i]=0.632*sumbC[i]+0.368*apperrlm1
}
testbcalm1_632Booterr=bcanon(elements632,10000,mean,alpha = c(0.025,0.975))
CI.lm1_632Booterr=testbcalm1_632Booterr$confpoints
CI.lm1_632Booterr=as.matrix(CI.lm1_632Booterr)
boot632se[1]=sd(elements632)/sqrt(100)
```

```r
#pm2  Looboot error:
X=cbind(dat[,1],dat[,1]^2)
datb=cbind(X,Y)
datb=as.data.frame(datb)
SEb=matrix(0,nrow = 100,ncol = 200)
C=c(rep(0,100))

for (b in 1:200) {
  smp_size=100
  bootsample_ind=sample(seq_len(nrow(datb)),size = smp_size,replace = TRUE)
  bootsample=datb[bootsample_ind,]
  mboot=lm(Y~.,data=bootsample)
  for (i in 1:100) {
    in_index=1
    for (z in 1:100) {
      if(Y[i]==bootsample$Y[z]){
        in_index=0
      }
    }
    if(in_index==1){
      C[i]=C[i]+1
      SEb[i,b]=( Y[i]-predict(mboot,list(V1=datb[i,1],V2=datb[i,2])) )^2
    }
  }
}
sumN=0
for (i in 1:100) {
  sumN=sumN+sum(SEb[i,])*(C[i])^(-1)
}
LooBootpm2=sumN/100
#for  LooBoot pm2 confidence interval:
sumbC=c(rep(0,100))
for (i in 1:100) {
  sumbC[i]=sum(SEb[i,])*(C[i])^(-1)
}
testbcapm2_LooBooterr=bcanon(sumbC,10000,mean,alpha = c(0.025,0.975))
CI.pm2_LooBooterr=testbcapm2_LooBooterr$confpoints
CI.pm2_LooBooterr=as.matrix(CI.pm2_LooBooterr)
loobooterrse[2]=sd(sumbC)/sqrt(100)

#pm2 .632 boot:
apperrpm2=sum((pm2$residuals)^2)/100
Err632bootpm2=0.368*apperrpm2+0.632*LooBootpm2
elements632=c(rep(0,100))
for (i in 1:100) {
  elements632[i]=0.632*sumbC[i]+0.368*apperrpm2
}
testbcapm2_632Booterr=bcanon(elements632,10000,mean,alpha = c(0.025,0.975))
CI.pm2_632Booterr=testbcapm2_632Booterr$confpoints
CI.pm2_632Booterr=as.matrix(CI.pm2_632Booterr)
boot632se[2]=sd(elements632)/sqrt(100)
```

```
#pm3  Looboot error:
X=cbind(dat[,1],dat[,1]^2,dat[,1]^3)
datb=cbind(X,Y)
datb=as.data.frame(datb)
SEb=matrix(0,nrow = 100,ncol = 200)
C=c(rep(0,100))

for (b in 1:200) {
  smp_size=100
  bootsample_ind=sample(seq_len(nrow(datb)),size = smp_size,replace = TRUE)
  bootsample=datb[bootsample_ind,]
  mboot=lm(Y~.,data=bootsample)
  for (i in 1:100) {
    in_index=1
    for (z in 1:100) {
      if(Y[i]==bootsample$Y[z]){
        in_index=0
      }
    }
    if(in_index==1){
      C[i]=C[i]+1
      SEb[i,b]=( Y[i]-predict(mboot,list(V1=datb[i,1],V2=datb[i,2],V3=datb[i,3])) )^2
    }
  }
}
sumN=0
for (i in 1:100) {
  sumN=sumN+sum(SEb[i,])*(C[i])^(-1)
}
LooBootpm3=sumN/100
#for  LooBoot pm3 confidence interval:
sumbC=c(rep(0,100))
for (i in 1:100) {
  sumbC[i]=sum(SEb[i,])*(C[i])^(-1)
}
testbcapm3_LooBooterr=bcanon(sumbC,10000,mean,alpha = c(0.025,0.975))
CI.pm3_LooBooterr=testbcapm3_LooBooterr$confpoints
CI.pm3_LooBooterr=as.matrix(CI.pm3_LooBooterr)
loobooterrse[3]=sd(sumbC)/sqrt(100)

#pm3 .632 boot:
apperrpm3=sum((pm3$residuals)^2)/100
Err632bootpm3=0.368*apperrpm3+0.632*LooBootpm3
elements632=c(rep(0,100))
for (i in 1:100) {
  elements632[i]=0.632*sumbC[i]+0.368*apperrpm3
}
testbcapm3_632Booterr=bcanon(elements632,10000,mean,alpha = c(0.025,0.975))
CI.pm3_632Booterr=testbcapm3_632Booterr$confpoints
CI.pm3_632Booterr=as.matrix(CI.pm3_632Booterr)
boot632se[3]=sd(elements632)/sqrt(100)
```

```
#pm4  Looboot error:
X=cbind(dat[,1],dat[,1]^2,dat[,1]^3,dat[,1]^4)
datb=cbind(X,Y)
datb=as.data.frame(datb)
SEb=matrix(0,nrow = 100,ncol = 200)
C=c(rep(0,100))

for (b in 1:200) {
  smp_size=100
  bootsample_ind=sample(seq_len(nrow(datb)),size = smp_size,replace = TRUE)
  bootsample=datb[bootsample_ind,]
  mboot=lm(Y~.,data=bootsample)
  for (i in 1:100) {
    in_index=1
    for (z in 1:100) {
      if(Y[i]==bootsample$Y[z]){
        in_index=0
      }
    }
    if(in_index==1){
      C[i]=C[i]+1
      SEb[i,b]=( Y[i]-predict(mboot,list(V1=datb[i,1],V2=datb[i,2],V3=datb[i,3],V4=datb[i,4])) )^2
    }
  }
}
sumN=0
for (i in 1:100) {
  sumN=sumN+sum(SEb[i,])*(C[i])^(-1)
}
LooBootpm4=sumN/100
#for  LooBoot pm4 confidence interval:
sumbC=c(rep(0,100))
for (i in 1:100) {
  sumbC[i]=sum(SEb[i,])*(C[i])^(-1)
}
testbcapm4_LooBooterr=bcanon(sumbC,10000,mean,alpha = c(0.025,0.975))
CI.pm4_LooBooterr=testbcapm4_LooBooterr$confpoints
CI.pm4_LooBooterr=as.matrix(CI.pm4_LooBooterr)
loobooterrse[4]=sd(sumbC)/sqrt(100)

#pm4 .632 boot:
apperrpm4=sum((pm4$residuals)^2)/100
Err632bootpm4=0.368*apperrpm4+0.632*LooBootpm4
elements632=c(rep(0,100))
for (i in 1:100) {
  elements632[i]=0.632*sumbC[i]+0.368*apperrpm4
}
testbcapm4_632Booterr=bcanon(elements632,10000,mean,alpha = c(0.025,0.975))
CI.pm4_632Booterr=testbcapm4_632Booterr$confpoints
CI.pm4_632Booterr=as.matrix(CI.pm4_632Booterr)
boot632se[4]=sd(elements632)/sqrt(100)
```

```
#pm5  Looboot error:
X=cbind(dat[,1],dat[,1]^2,dat[,1]^3,dat[,1]^4,dat[,1]^5)
datb=cbind(X,Y)
datb=as.data.frame(datb)
SEb=matrix(0,nrow = 100,ncol = 200)
C=c(rep(0,100))

for (b in 1:200) {
  smp_size=100
  bootsample_ind=sample(seq_len(nrow(datb)),size = smp_size,replace = TRUE)
  bootsample=datb[bootsample_ind,]
  mboot=lm(Y~.,data=bootsample)
  for (i in 1:100) {
    in_index=1
    for (z in 1:100) {
      if(Y[i]==bootsample$Y[z]){
        in_index=0
      }
    }
    if(in_index==1){
      C[i]=C[i]+1
      SEb[i,b]=( Y[i]-predict(mboot,list(V1=datb[i,1],V2=datb[i,2],V3=datb[i,3],V4=datb[i,4],V5=datb[i,5])) )^2
    }
  }
}
sumN=0
for (i in 1:100) {
  sumN=sumN+sum(SEb[i,])*(C[i])^(-1)
}
LooBootpm5=sumN/100
#for  LooBoot pm5 confidence interval:
sumbC=c(rep(0,100))
for (i in 1:100) {
  sumbC[i]=sum(SEb[i,])*(C[i])^(-1)
}
testbcapm5_LooBooterr=bcanon(sumbC,10000,mean,alpha = c(0.025,0.975))
CI.pm5_LooBooterr=testbcapm5_LooBooterr$confpoints
CI.pm5_LooBooterr=as.matrix(CI.pm5_LooBooterr)
loobooterrse[5]=sd(sumbC)/sqrt(100)

#pm5 .632 boot:
apperrpm5=sum((pm5$residuals)^2)/100
Err632bootpm5=0.368*apperrpm5+0.632*LooBootpm5
elements632=c(rep(0,100))
for (i in 1:100) {
  elements632[i]=0.632*sumbC[i]+0.368*apperrpm5
}
testbcapm5_632Booterr=bcanon(elements632,10000,mean,alpha = c(0.025,0.975))
CI.pm5_632Booterr=testbcapm5_632Booterr$confpoints
CI.pm5_632Booterr=as.matrix(CI.pm5_632Booterr)
boot632se[5]=sd(elements632)/sqrt(100)
```

```
#pm6  Looboot error:
X=cbind(dat[,1],dat[,1]^2,dat[,1]^3,dat[,1]^4,dat[,1]^5,dat[,1]^6)
datb=cbind(X,Y)
datb=as.data.frame(datb)
SEb=matrix(0,nrow = 100,ncol = 200)
C=c(rep(0,100))

for (b in 1:200) {
  smp_size=100
  bootsample_ind=sample(seq_len(nrow(datb)),size = smp_size,replace = TRUE)
  bootsample=datb[bootsample_ind,]
  mboot=lm(Y~.,data=bootsample)
  for (i in 1:100) {
    in_index=1
    for (z in 1:100) {
      if(Y[i]==bootsample$Y[z]){
        in_index=0
      }
    }
    if(in_index==1){
      C[i]=C[i]+1
      SEb[i,b]=( Y[i]-predict(mboot,list(V1=datb[i,1],V2=datb[i,2],V3=datb[i,3],
                                         V4=datb[i,4],V5=datb[i,5],V6=datb[i,6])) )^2
    }
  }
}
sumN=0
for (i in 1:100) {
  sumN=sumN+sum(SEb[i,])*(C[i])^(-1)
}
LooBootpm6=sumN/100
#for  LooBoot pm6 confidence interval:
sumbC=c(rep(0,100))
for (i in 1:100) {
  sumbC[i]=sum(SEb[i,])*(C[i])^(-1)
}
testbcapm6_LooBooterr=bcanon(sumbC,10000,mean,alpha = c(0.025,0.975))
CI.pm6_LooBooterr=testbcapm6_LooBooterr$confpoints
CI.pm6_LooBooterr=as.matrix(CI.pm6_LooBooterr)
loobooterrse[6]=sd(sumbC)/sqrt(100)

#pm6 .632 boot:
apperrpm6=sum((pm6$residuals)^2)/100
Err632bootpm6=0.368*apperrpm6+0.632*LooBootpm6
elements632=c(rep(0,100))
for (i in 1:100) {
  elements632[i]=0.632*sumbC[i]+0.368*apperrpm6
}
testbcapm6_632Booterr=bcanon(elements632,10000,mean,alpha = c(0.025,0.975))
CI.pm6_632Booterr=testbcapm6_632Booterr$confpoints
CI.pm6_632Booterr=as.matrix(CI.pm6_632Booterr)
boot632se[6]=sd(elements632)/sqrt(100)
```

```
#pm7  Looboot error:
X=cbind(dat[,1],dat[,1]^2,dat[,1]^3,dat[,1]^4,dat[,1]^5,dat[,1]^6,dat[,1]^7)
datb=cbind(X,Y)
datb=as.data.frame(datb)
SEb=matrix(0,nrow = 100,ncol = 200)
C=c(rep(0,100))

for (b in 1:200) {
  smp_size=100
  bootsample_ind=sample(seq_len(nrow(datb)),size = smp_size,replace = TRUE)
  bootsample=datb[bootsample_ind,]
  mboot=lm(Y~.,data=bootsample)
  for (i in 1:100) {
    in_index=1
    for (z in 1:100) {
      if(Y[i]==bootsample$Y[z]){
        in_index=0
      }
    }
    if(in_index==1){
      C[i]=C[i]+1
      SEb[i,b]=( Y[i]-predict(mboot,list(V1=datb[i,1],V2=datb[i,2],V3=datb[i,3],
                                         V4=datb[i,4],V5=datb[i,5],V6=datb[i,6],V7=datb[i,7])) )^2
    }
  }
}
sumN=0
for (i in 1:100) {
  sumN=sumN+sum(SEb[i,])*(C[i])^(-1)
}
LooBootpm7=sumN/100
#for  LooBoot pm7 confidence interval:
sumbC=c(rep(0,100))
for (i in 1:100) {
  sumbC[i]=sum(SEb[i,])*(C[i])^(-1)
}
testbcapm7_LooBooterr=bcanon(sumbC,10000,mean,alpha = c(0.025,0.975))
CI.pm7_LooBooterr=testbcapm7_LooBooterr$confpoints
CI.pm7_LooBooterr=as.matrix(CI.pm7_LooBooterr)
loobooterrse[7]=sd(sumbC)/sqrt(100)


#pm7 .632 boot:
apperrpm7=sum((pm7$residuals)^2)/100
Err632bootpm7=0.368*apperrpm7+0.632*LooBootpm7
elements632=c(rep(0,100))
for (i in 1:100) {
  elements632[i]=0.632*sumbC[i]+0.368*apperrpm7
}
testbcapm7_632Booterr=bcanon(elements632,10000,mean,alpha = c(0.025,0.975))
CI.pm7_632Booterr=testbcapm7_632Booterr$confpoints
CI.pm7_632Booterr=as.matrix(CI.pm7_632Booterr)
boot632se[7]=sd(elements632)/sqrt(100)
```

```
## independent test set for true error generation:
true_error_lm1=c(rep(0,100))
true_error_pm2=c(rep(0,100))
true_error_pm3=c(rep(0,100))
true_error_pm4=c(rep(0,100))
true_error_pm5=c(rep(0,100))
true_error_pm6=c(rep(0,100))
true_error_pm7=c(rep(0,100))
for (i in 1:100) {

  Xt=runif(100,0,20)
  epsilon=rnorm(100,0,350)
  Yt=500+(Xt-10)^3+epsilon
  dattest=cbind(Xt,Yt)
  dattest=as.data.frame(dattest)

  lm1hat=predict(lm1,list(X=Xt))
  true_error_lm1[i]=sum((Yt-lm1hat)^2)/100

  pm2hat=predict(pm2,list(X=Xt))
  true_error_pm2[i]=sum((Yt-pm2hat)^2)/100

  pm3hat=predict(pm3,list(X=Xt))
  true_error_pm3[i]=sum((Yt-pm3hat)^2)/100

  pm4hat=predict(pm4,list(X=Xt))
  true_error_pm4[i]=sum((Yt-pm4hat)^2)/100

  pm5hat=predict(pm5,list(X=Xt))
  true_error_pm5[i]=sum((Yt-pm5hat)^2)/100

  pm6hat=predict(pm6,list(X=Xt))
  true_error_pm6[i]=sum((Yt-pm6hat)^2)/100

  pm7hat=predict(pm7,list(X=Xt))
  true_error_pm7[i]=sum((Yt-pm7hat)^2)/100

}
true_error_lm1=sum(true_error_lm1)/100
true_error_pm2=sum(true_error_pm2)/100
true_error_pm3=sum(true_error_pm3)/100
true_error_pm4=sum(true_error_pm4)/100
true_error_pm5=sum(true_error_pm5)/100
true_error_pm6=sum(true_error_pm6)/100
true_error_pm7=sum(true_error_pm7)/100

# Plots and  BCa confidence intervals for the prediction accuracy of the error estimates:-----------------------------
MSE=c(MSEerrlm1,MSEerrpm2,MSEerrpm3,MSEerrpm4,MSEerrpm5,MSEerrpm6,MSEerrpm7)
LooCV=c(LooCVerrlm1,LooCVerrpm2,LooCVerrpm3,LooCVerrpm4,LooCVerrpm5,LooCVerrpm6,LooCVerrpm7)
fold10CV=c(fold10CVlm1,fold10CVpm2,fold10CVpm3,fold10CVpm4,fold10CVpm5,fold10CVpm6,fold10CVpm7)
GCV=c(GCVerrlm1,GCVerrpm2,GCVerrpm3,GCVerrpm4,GCVerrpm5,GCVerrpm6,GCVerrpm7)
LooBooterr=c(LooBootlm1,LooBootpm2,LooBootpm3,LooBootpm4,LooBootpm5,LooBootpm6,LooBootpm7)
Boot632err=c(Err632bootlm1,Err632bootpm2,Err632bootpm3,Err632bootpm4,Err632bootpm5,Err632bootpm6,Err632bootpm7)
trueError=c(true_error_lm1,true_error_pm2,true_error_pm3,true_error_pm4,true_error_pm5,true_error_pm6,
            true_error_pm7)
```

```
p=1:7
#MSE and true error:
png(filename="errorgraph1.png")
plot(MSE,type="o",xlab="m-degree",ylab="error estimates",main="MSE",ylim=c(70000,300000),col="orange",lwd=4)
points(p,trueError,col="cyan3",type="o",lwd=2)
dev.off()

# GCV error estimates: different units but same selection with LooCV
png(filename="errorgraph2.png")
CI.up=c(CI.lm1_GCV[2,2],CI.pm2_GCV[2,2],CI.pm3_GCV[2,2],CI.pm4_GCV[2,2],CI.pm5_GCV[2,2],
        CI.pm6_GCV[2,2],CI.pm7_GCV[2,2])
CI.down=c(CI.lm1_GCV[1,2],CI.pm2_GCV[1,2],CI.pm3_GCV[1,2],CI.pm4_GCV[1,2],CI.pm5_GCV[1,2],
          CI.pm6_GCV[1,2],CI.pm7_GCV[1,2])
plot(GCV,type = "o",xlab = "m-degree",ylab = "error estimates",main=" Generalized CV method",
     ylim = c(70000,300000),col="orange",lwd=4)
arrows(p,CI.down,p,CI.up,code = 3,length = 0.1,angle = 90,col = "black")
points(p,trueError,col="cyan3",type="o",lwd=2)
dev.off()

#Leave one out Cross validation:
png(filename="errorgraph3.png")
CI.up=c(CI.lm1Loocv[2,2],CI.pm2Loocv[2,2],CI.pm3Loocv[2,2],CI.pm4Loocv[2,2],CI.pm5Loocv[2,2],
        CI.pm6Loocv[2,2],CI.pm7Loocv[2,2])
CI.down=c(CI.lm1Loocv[1,2],CI.pm2Loocv[1,2],CI.pm3Loocv[1,2],CI.pm4Loocv[1,2],CI.pm5Loocv[1,2],
          CI.pm6Loocv[1,2],CI.pm7Loocv[1,2])
plot(LooCV,type = "o",xlab = "m-degree",ylab = "error estimates",main="Leave on out CV method",
     ylim = c(70000,300000),col="orange",lwd=4)
arrows(p,CI.down,p,CI.up,code = 3,length = 0.1,angle = 90,col = "black")
points(p,trueError,col="cyan3",type="o",lwd=2)
dev.off()

# 10 fold Cross validation:
png(filename="errorgraph4.png" )
CI.up=c(CI.lm1_10foldcv[2,2],CI.pm2_10foldcv[2,2],CI.pm3_10foldcv[2,2],
        CI.pm4_10foldcv[2,2],CI.pm5_10foldcv[2,2],CI.pm6_10foldcv[2,2],CI.pm7_10foldcv[2,2])
CI.down=c(CI.lm1_10foldcv[1,2],CI.pm2_10foldcv[1,2],CI.pm3_10foldcv[1,2],
          CI.pm4_10foldcv[1,2],CI.pm5_10foldcv[1,2],CI.pm6_10foldcv[1,2],CI.pm7_10foldcv[1,2])
plot(fold10CV,type = "o",xlab = "m-degree",ylab = "error estimates",main="10 fold CV method",
     ylim = c(70000,300000),col="orange",lwd=4)
arrows(p,CI.down,p,CI.up,code = 3,length = 0.1,angle = 90,col = "black")
points(p,trueError,col="cyan3",type="o",lwd=2)
dev.off()

# Leave one out bootstrap error estimates:
png(filename="errorgraph5.png")
CI.up=c(CI.lm1_LooBooterr[2,2],CI.pm2_LooBooterr[2,2],CI.pm3_LooBooterr[2,2],
        CI.pm4_LooBooterr[2,2],CI.pm5_LooBooterr[2,2],CI.pm6_LooBooterr[2,2],CI.pm7_LooBooterr[2,2])
CI.down=c(CI.lm1_LooBooterr[1,2],CI.pm2_LooBooterr[1,2],CI.pm3_LooBooterr[1,2],
          CI.pm4_LooBooterr[1,2],CI.pm5_LooBooterr[1,2],CI.pm6_LooBooterr[1,2],CI.pm7_LooBooterr[1,2])
plot(LooBooterr,type = "o",xlab = "m-degree",ylab="error estimates",main="LooBootstrap method",
     ylim=c(70000,300000),col="orange",lwd=4)
arrows(p,CI.down,p,CI.up,code = 3,length = 0.1,angle = 90,col = "black")
points(p,trueError,col="cyan3",type="o",lwd=2)
dev.off()
```

```
# .632 bootstrap error estimates:
png(filename="errorgraph6.png")
CI.up=c(CI.lm1_632Booterr[2,2],CI.pm2_632Booterr[2,2],CI.pm3_632Booterr[2,2],
        CI.pm4_632Booterr[2,2],CI.pm5_632Booterr[2,2],CI.pm6_632Booterr[2,2],CI.pm7_632Booterr[2,2])
CI.down=c(CI.lm1_632Booterr[1,2],CI.pm2_632Booterr[1,2],CI.pm3_632Booterr[1,2],
          CI.pm4_632Booterr[1,2],CI.pm5_632Booterr[1,2],CI.pm6_632Booterr[1,2],CI.pm7_632Booterr[1,2])
plot(Boot632err,type = "o",xlab = "m-degree",ylab = "error estimates",main=".632 Bootstrap method",
     ylim = c(70000,300000),col="orange",lwd=4)
arrows(p,CI.down,p,CI.up,code = 3,length = 0.1,angle = 90,col = "black")
points(p,trueError,col="cyan3",type="o",lwd=2)
dev.off()


# R^2 and AdjR^2:
p=1:7
png(filename="R2R2adj.png")
R2=c(summary(lm1)$r.squared,summary(pm2)$r.squared,summary(pm3)$r.squared,summary(pm4)$r.squared
     ,summary(pm5)$r.squared,summary(pm6)$r.squared,summary(pm7)$r.squared)
adjR2=c(summary(lm1)$adj.r.squared,summary(pm2)$adj.r.squared,summary(pm3)$adj.r.squared
        ,summary(pm4)$adj.r.squared,summary(pm5)$adj.r.squared,summary(pm6)$adj.r.squared,
        summary(pm7)$adj.r.squared)
plot(R2,type = "o",xlab = "m-degree",ylab="R2 and Adjusted R2",ylim = c(0.37,0.60),col="red",lwd=2)
points(p,adjR2,col="blue",type = "o",lwd=2)
legend(1,0.51,expression(R^2,R[adj]^2),col=c("red","blue"), cex=0.8,lty = 1)
dev.off()
```

Κεφάλαιο 6 - Προσομοίωση 1:

```
#all methods simulation test:
#packages: boot,bootstrap,corrplot,dplyr,glmnet,JWileymisc,leaps,MASS

#matrixes for the variable names of the models selected by each method from all 100 simulations!
γBICexhaust=matrix(0,nrow = 100,ncol=8)
γAICexhaust=matrix(0,nrow = 100,ncol=8)
γAIC_cexhaust=matrix(0,nrow = 100,ncol = 8)
γstepcomb=matrix(0,nrow = 100,ncol=8)
γstepforw=matrix(0,nrow = 100,ncol=8)
γstepback=matrix(0,nrow = 100,ncol=8)
γlasso=matrix(0,nrow = 100,ncol=8)
γadjR2exhaust=matrix(0,nrow = 100,ncol=8)
γR2exhaust=matrix(0,nrow = 100,ncol=8)
γMSEexhaust=matrix(0,nrow = 100,ncol=8)
γMSEcorrectedexhaust=matrix(0,nrow = 100,ncol=8)

megaX=NULL
megaY=NULL
n=150
p=8
Sigma=matrix(0,8,8)
for(i in 1:8){
  for(j in 1:8){
    Sigma[i,j]=0.5^abs(i-j)
  }
}
### 100 samples ###
for (z in 1:100) {
  mu=rep(0,8)
  X=mvrnorm(n=40,mu,Sigma)
  b1=3
  b2=1.5
  b5=2
  b0=0.8
  epsilon=mvrnorm(n=1,rep(0,40),diag(1,40))
  Y=b0+X%*%c(b1,b2,rep(0,2),b5,rep(0,3))+epsilon
  megaX=rbind(megaX,X)
  megaY=rbind(megaY,Y)
}
megaX=as.data.frame(megaX)

a=1
b=40 #-------------------------------------------------------------------------
for (z in 1:100) {

  X=megaX[a:b,]
  Y=megaY[a:b]
  datafull=data.frame(X,Y)
  colnames(datafull)=c("X1","X2","X3","X4","X5","X6","X7","X8","Y")
  Xdat=datafull[,1:8]
  #full model:
  mfull=lm(Y~.,data = datafull)
  #null model:
  mnull=lm(Y~1,data=datafull)
```

```
# custom full enumeration mega loop :::
Y=expand.grid(c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1),c(0,1))
Y=Y[-1,]
Y=as.matrix(Y)
bic=rep(0,255)
aic=rep(0,255)
aic_c=rep(0,255)
MSEerr=rep(0,255)
MSEcorrected=rep(0,255)
adjR2exhaust=rep(0,255)
R2exhaust=rep(0,255)
coef_ofmodels=matrix(1,nrow = 255,ncol = 8)
colnames(coef_ofmodels)=colnames(Xdat)
for (i in 1:255) {
  Xsub=Xdat
  for (j in 1:8) {
    if(Y[i,j]==0){
      Xsub[,j]=0
    }
  }
  Xsub=Xsub[, colSums(Xsub == 0) !=nrow(Xsub),drop=FALSE]
  m=lm(Y~.,data=Xsub)
  mcoefnames=names(coef(m))[-1]
  for (k in 1:length(names(Xdat))) {
    c=0
    for (j in 1:length(mcoefnames)) {
      if(names(Xdat)[k]==mcoefnames[j]){
        c=1
      }
    }
    if(c==0){
      coef_ofmodels[i,k]=0
    }
  }
  bic[i]=BIC(m)
  aic[i]=AIC(m)
  aic_c[i]=AIC(m)+2*(length(mcoefnames)+2)*(  (length(mcoefnames)+2+1)/(40-length(mcoefnames)-2-1) )
  MSEerr[i]=sum((m$residuals)^2)/40
  MSEcorrected[i]=sum((m$residuals)^2)/(40-length(mcoefnames)-1)
  adjR2exhaust[i]=summary(m)$adj.r.squared
  R2exhaust[i]=summary(m)$r.squared

}
posBIC=which.min(bic)
posAIC=which.min(aic)
posAICc=which.min(aic_c)
posMSE=which.min(MSEerr)
posMSEcor=which.min(MSEcorrected)
posadjR2=which.max(adjR2exhaust)
posR2=which.max(R2exhaust)
```

```
#for correct model frequency selection!
γBICexhaust[z,]=coef_ofmodels[posBIC,]
γAICexhaust[z,]=coef_ofmodels[posAIC,]
γAIC_cexhaust[z,]=coef_ofmodels[posAICc,]
γMSEexhaust[z,]=coef_ofmodels[posMSE,]
γMSEcorrectedexhaust[z,]=coef_ofmodels[posMSEcor,]
γadjR2exhaust[z,]=coef_ofmodels[posadjR2,]
γR2exhaust[z,]=coef_ofmodels[posR2,]


#stepwise combination selection:
stepcombm=step(mfull,direction = "both",trace = 0)
coef(stepcombm)
#finding the design matrix of step model:
namecoefstep=names(stepcombm$coefficients)[-1]
Xdatastepcomb=Xdat
for (j in 1:length(names(Xdat))) {
  c=0
  for (i in 1:length(namecoefstep)) {
    if(names(Xdat)[j]==namecoefstep[i]){
      c=1
    }
  }
  if(c==0){
    Xdatastepcomb[,j]=0
  }
}
for (j in 1:8) {
  if(Xdatastepcomb[1,j]!=0){
    γstepcomb[z,j]=1
  }
}

#stepwise forward selection:
stepforw=step(mnull,scope=list(lower=mnull,upper=mfull),direction = "forward",trace = 0)
coef(stepforw)
#finding the design matrix of step model:
namecoefstep=names(stepforw$coefficients)[-1]
Xdatastepforw=Xdat
for (j in 1:length(names(Xdat))) {
  c=0
  for (i in 1:length(namecoefstep)) {
    if(names(Xdat)[j]==namecoefstep[i]){
      c=1
    }
  }
  if(c==0){
    Xdatastepforw[,j]=0
  }
}
for (j in 1:8) {
  if(Xdatastepforw[1,j]!=0){
    γstepforw[z,j]=1
  }
}
```

```
  #stepwise backward selection:
  stepback=step(mfull,direction = "backward",trace = 0)
  coef(stepback)
  #finding the design matrix of step model:
  namecoefstep=names(stepback$coefficients)[-1]
  Xdatastepback=Xdat
  for (j in 1:length(names(Xdat))) {
    c=0
    for (i in 1:length(namecoefstep)) {
      if(names(Xdat)[j]==namecoefstep[i]){
        c=1
      }
    }
    if(c==0){
      Xdatastepback[,j]=0
    }
  }
  for (j in 1:8) {
    if(Xdatastepback[1,j]!=0){
      ystepback[z,j]=1
    }
  }



  #lasso model selector:
  X=as.matrix(Xdat)
  mlasso=glmnet(X,Y,alpha=1)
  #cross validation for lambda estimation choose 5fold or 10fold, default is nfolds=10
  cv.lasso = cv.glmnet(X,Y,alpha=1)
  coeflasso1lambda1se=predict(mlasso,s=cv.lasso$lambda.1se,newx=X,type="coefficients")
  p=nnzero(as.matrix(coeflasso1lambda1se))-1   #minus the intercept
  yhat_lasso=cbind(1,X)%*%coeflasso1lambda1se
  #mse_lasso[z]=sum((Y-yhat_lasso)^2)/(150)
  namecoeflasso=as.matrix(coeflasso1lambda1se)[-1]
  for (j in 1:8) {
    if(namecoeflasso[j]!=0){
      ylasso[z,j]=1
    }
  }




  a=a+40
  b=b+40

}
```

```
# Run previous code simultaneously ----------                          ------------------------------
#------------------------------------------- SIMULATIONS END HERE ------------------------------

#frequencies of variables125model selection by methods:

counterbic=0
counteradjR2=0
counterstepcomb=0
counterstepforw=0
counterstepback=0
counterlasso=0
counterR2=0
counterAIC=0
counterAICc=0
counterMSE=0
counterMSEcorrected=0


for (i in 1:100) {
  if(yBICexhaust[i,1]==1 && yBICexhaust[i,2]==1 && yBICexhaust[i,5]==1
     && yBICexhaust[i,3]==0 && yBICexhaust[i,4]==0
     && yBICexhaust[i,6]==0 && yBICexhaust[i,7]==0 && yBICexhaust[i,8]==0){
    counterbic=counterbic+1
  }
  if(yadjR2exhaust[i,1]==1 && yadjR2exhaust[i,2]==1 && yadjR2exhaust[i,5]==1
     && yadjR2exhaust[i,3]==0 && yadjR2exhaust[i,4]==0
     && yadjR2exhaust[i,6]==0 && yadjR2exhaust[i,7]==0 && yadjR2exhaust[i,8]==0){
    counteradjR2=counteradjR2+1
  }
  if(ystepcomb[i,1]==1 && ystepcomb[i,2]==1 && ystepcomb[i,5]==1 && ystepcomb[i,3]==0
     && ystepcomb[i,4]==0 && ystepcomb[i,6]==0
     && ystepcomb[i,7]==0 && ystepcomb[i,8]==0){
    counterstepcomb=counterstepcomb+1
  }
  if(ystepforw[i,1]==1 && ystepforw[i,2]==1 && ystepforw[i,5]==1 && ystepforw[i,3]==0
     && ystepforw[i,4]==0 && ystepforw[i,6]==0
     && ystepforw[i,7]==0 && ystepforw[i,8]==0){
    counterstepforw=counterstepforw+1
  }
  if(ystepback[i,1]==1 && ystepback[i,2]==1 && ystepback[i,5]==1 && ystepback[i,3]==0
     && ystepback[i,4]==0 && ystepback[i,6]==0
     && ystepback[i,7]==0 && ystepback[i,8]==0){
    counterstepback=counterstepback+1
  }
  if(ylasso[i,1]==1 && ylasso[i,2]==1 && ylasso[i,5]==1 && ylasso[i,3]==0
     && ylasso[i,4]==0 && ylasso[i,6]==0
     && ylasso[i,7]==0 && ylasso[i,8]==0){
    counterlasso=counterlasso+1
  }
  if(yR2exhaust[i,1]==1 && yR2exhaust[i,2]==1 && yR2exhaust[i,5]==1 && yR2exhaust[i,3]==0
     && yR2exhaust[i,4]==0
     && yR2exhaust[i,6]==0 && yR2exhaust[i,7]==0 && yR2exhaust[i,8]==0){
    counterR2=counterR2+1
  }
```

```
  if(yAICexhaust[i,1]==1 && yAICexhaust[i,2]==1 && yAICexhaust[i,5]==1 && yAICexhaust[i,3]==0
      && yAICexhaust[i,4]==0
      && yAICexhaust[i,6]==0 && yAICexhaust[i,7]==0 && yAICexhaust[i,8]==0){
    counterAIC=counterAIC+1
  }
  if(yAIC_cexhaust[i,1]==1 && yAIC_cexhaust[i,2]==1 && yAIC_cexhaust[i,5]==1
      && yAIC_cexhaust[i,3]==0 && yAIC_cexhaust[i,4]==0
      && yAIC_cexhaust[i,6]==0 && yAIC_cexhaust[i,7]==0 && yAIC_cexhaust[i,8]==0){
    counterAICc=counterAICc+1
  }
  if(yMSEexhaust[i,1]==1 && yMSEexhaust[i,2]==1 && yMSEexhaust[i,5]==1 && yMSEexhaust[i,3]==0
      && yMSEexhaust[i,4]==0 && yMSEexhaust[i,6]==0
      && yMSEexhaust[i,7]==0 && yMSEexhaust[i,8]==0){
    counterMSE=counterMSE+1
  }
  if(yMSEcorrectedexhaust[i,1]==1 && yMSEcorrectedexhaust[i,2]==1 && yMSEcorrectedexhaust[i,5]==1
      && yMSEcorrectedexhaust[i,3]==0
      && yMSEcorrectedexhaust[i,4]==0 && yMSEcorrectedexhaust[i,6]==0 && yMSEcorrectedexhaust[i,7]==0
      && yMSEcorrectedexhaust[i,8]==0){
    counterMSEcorrected=counterMSEcorrected+1
  }
}


freq=matrix(0,ncol = 11,nrow=1)
colnames(freq)=c("BIC_exhaust","AIC_exhaust","AdjR2_exhaust","StepComb","StepForw",
                 "StepBack","Lasso","R2_exhaust","AICc_exhaust","MSE","MSEcorrected")
freq[1,1]=counterbic/100
freq[1,2]=counterAIC/100
freq[1,3]=counteradjR2/100
freq[1,4]=counterstepcomb/100
freq[1,5]=counterstepforw/100
freq[1,6]=counterstepback/100
freq[1,7]=counterlasso/100
freq[1,8]=counterR2/100
freq[1,9]=counterAICc/100
freq[1,10]=counterMSE/100
freq[1,11]=counterMSEcorrected/100
freq125=freq
freq125

#model125 contained frequencies:
counterbic=0
counteradjR2=0
counterstepcomb=0
counterstepforw=0
counterstepback=0
counterlasso=0
counterR2=0
counterAIC=0
counterAICc=0
counterMSE=0
counterMSEcorrected=0
```

```
for (i in 1:100) {
  if(yBICexhaust[i,1]==1 && yBICexhaust[i,2]==1 && yBICexhaust[i,5]==1 && sum(yBICexhaust[i,])<=5){
    counterbic=counterbic+1
  }
  if(yadjR2exhaust[i,1]==1 && yadjR2exhaust[i,2]==1 && yadjR2exhaust[i,5]==1
     && sum(yadjR2exhaust[i,])<=5){
    counteradjR2=counteradjR2+1
  }
  if(ystepcomb[i,1]==1 && ystepcomb[i,2]==1 && ystepcomb[i,5]==1 && sum(ystepcomb[i,])<=5){
    counterstepcomb=counterstepcomb+1
  }
  if(ystepforw[i,1]==1 && ystepforw[i,2]==1 && ystepforw[i,5]==1 && sum(ystepforw[i,])<=5){
    counterstepforw=counterstepforw+1
  }
  if(ystepback[i,1]==1 && ystepback[i,2]==1 && ystepback[i,5]==1 && sum(ystepback[i,])<=5){
    counterstepback=counterstepback+1
  }
  if(ylasso[i,1]==1 && ylasso[i,2]==1 && ylasso[i,5]==1 && sum(ylasso[i,])<=5){
    counterlasso=counterlasso+1
  }
  if(yR2exhaust[i,1]==1 && yR2exhaust[i,2]==1 && yR2exhaust[i,5]==1 && sum(yR2exhaust[i,])<=5){
    counterR2=counterR2+1
  }
  if(yAICexhaust[i,1]==1 && yAICexhaust[i,2]==1 && yAICexhaust[i,5]==1 && sum(yAICexhaust[i,])<=5){
    counterAIC=counterAIC+1
  }
  if(yAIC_cexhaust[i,1]==1 && yAIC_cexhaust[i,2]==1 && yAIC_cexhaust[i,5]==1 && sum(yAIC_cexhaust[i,])<=5){
    counterAICc=counterAICc+1
  }
  if(yMSEexhaust[i,1]==1 && yMSEexhaust[i,2]==1 && yMSEexhaust[i,5]==1 && sum(yMSEexhaust[i,])<=5){
    counterMSE=counterMSE+1
  }
  if(yMSEcorrectedexhaust[i,1]==1 && yMSEcorrectedexhaust[i,2]==1 && yMSEcorrectedexhaust[i,5]==1
     && sum(yMSEcorrectedexhaust[i,])<=5){
    counterMSEcorrected=counterMSEcorrected+1
  }
}

freq=matrix(0,ncol = 11,nrow=1)
colnames(freq)=c("BIC_exhaust","AIC_exhaust","AdjR2_exhaust","StepComb","StepForw",
                 "StepBack","Lasso","R2_exhaust","AICc_exhaust","MSE","MSEcorrected")
freq[1,1]=counterbic/100
freq[1,2]=counterAIC/100
freq[1,3]=counteradjR2/100
freq[1,4]=counterstepcomb/100
freq[1,5]=counterstepforw/100
freq[1,6]=counterstepback/100
freq[1,7]=counterlasso/100
freq[1,8]=counterR2/100
freq[1,9]=counterAICc/100
freq[1,10]=counterMSE/100
freq[1,11]=counterMSEcorrected/100
freq125contained=freq
freq125contained
```

Κεφάλαιο 6 - Προσομοίωση 2:

```
#load package MASS, glmnet
library(MASS)
library(glmnet)
megaX=NULL
megaY=NULL

n=25
p=50
Sigma=matrix(0,50,50)
for(i in 1:50){
  for(j in 1:50){
    Sigma[i,j]=0.75^abs(i-j)
  }
}
###100 deigmata###
for (k in 1:100){
  mu=rep(0,50)
  X=mvrnorm(n = 25, mu, Sigma)
  b1=2
  b2=0.8
  b10=1.5
  b0=0.6
  epsilon=mvrnorm(n = 1, rep(0,25), 1.5*diag(1,25))
  Y=b0+X%*%c(b1,b2,rep(0,7),b10,rep(0,40))+epsilon
  megaX=rbind(megaX,X)
  megaY=rbind(megaY,Y)
}
###1o deigma###
#y=megaY[1:25]
#X=megaX[1:25,]

#-----------------------------------------------------------------------------------------------------
ystepcomb=matrix(0,nrow = 100,ncol=50)
ylasso=matrix(0,nrow = 100,ncol=50)
ystepforw=matrix(0,nrow = 100,ncol=50)
megaX=as.data.frame(megaX)

a=1
b=25
for (z in 1:100) {

  X=megaX[a:b,]
  Y=megaY[a:b]
  datafull=data.frame(X,Y)
  colnames(datafull)=c("X1","X2","X3","X4","X5","X6","X7","X8","X9","X10","X11","X12",
                       "X13","X14","X15","X16","X17","X18","X19","X20","X21","X22","X23",
                       "X24","X25","X26","X27","X28","X29","X30","X31","X32","X33","X34",
                       "X35","X36","X37","X38","X39","X40","X41","X42","X43","X44","X45",
                       "X46","X47","X48","X49","X50","Y")
  Xdat=datafull[,1:50]
  #full model :
  mfull=lm(Y~.,data=datafull)
  #null model :
  mnull=lm(Y~1,data=datafull)
```

```r
#stepwise comb ( both directions starting from null model) selection:
stepcombm=step(mnull,scope=list(lower=mnull,upper=mfull),direction = "both",trace = 0)
coef(stepcombm)
#finding the design matrix of step model:
namecoefstep=names(stepcombm$coefficients)[-1]
Xdatastep=Xdat
for (k in 1:length(names(Xdat))) {
  c=0
  for (m in 1:length(namecoefstep)) {
    if(names(Xdat)[k]==namecoefstep[m]){
      c=1
    }
  }
  if(c==0){
    Xdatastep[,k]=0
  }
}
for (k in 1:50) {
  if(Xdatastep[1,k]!=0){
    ystepcomb[z,k]=1
  }
}
Xdatastep=Xdatastep[, colSums(Xdatastep == 0) !=nrow(Xdatastep)]



# forward stepwise (starting from null model) selection:
stepforwm=step(mnull,scope=list(lower=mnull,upper=mfull),direction = "forward",trace=0)
#finding the design matrix of stepforward  model:
namecoefstep=names(stepforwm$coefficients)[-1]
Xdatastepforw=Xdat
for (k in 1:length(names(Xdat))) {
  c=0
  for (m in 1:length(namecoefstep)) {
    if(names(Xdat)[k]==namecoefstep[m]){
      c=1
    }
  }
  if(c==0){
    Xdatastepforw[,k]=0
  }
}
for (k in 1:50) {
  if(Xdatastepforw[1,k]!=0){
    ystepforw[z,k]=1
  }
}
Xdatastepforw=Xdatastepforw[, colSums(Xdatastepforw == 0) !=nrow(Xdatastepforw)]
```

```
  #lasso model selector:
  X=as.matrix(Xdat)
  mlasso=glmnet(X,Y,alpha=1)
  #cross validation for lambda estimation choose 5fold! default is nfolds=10
  cv.lasso = cv.glmnet(X,Y,alpha=1,nfolds = 5)
  coeflasso1lambda1se=predict(mlasso,s=cv.lasso$lambda.1se,newx=X,type="coefficients")
  p=nnzero(as.matrix(coeflasso1lambda1se))-1   #minus the intercept
  yhat_lasso=cbind(1,X)%*%coeflasso1lambda1se
  #mse_lasso[z]=sum((Y-yhat_lasso)^2)/(150)
  namecoeflasso=as.matrix(coeflasso1lambda1se)[-1]
  for (k in 1:50) {
    if(namecoeflasso[k]!=0){
      ylasso[z,k]=1
    }
  }

  a=a+25
  b=b+25
}
#------------------------------------------------  SIMULATIONS END HERE  -------------------------

#frequencies of variables 1,2,10 model selection by methods:


counter12_10lasso=0
counter12_10stepforw=0
counter12_10stepcomb=0
for (i in 1:100) {
  if(ylasso[i,1]==1 && ylasso[i,2]==1 && ylasso[i,5]==0 && ylasso[i,3]==0 && ylasso[i,4]==0
      && ylasso[i,6]==0 && ylasso[i,7]==0 && ylasso[i,8]==0 && ylasso[i,9]==0
      && ylasso[i,10]==1 && ylasso[i,11]==0 && ylasso[i,12]==0 && ylasso[i,13]==0
      && ylasso[i,14]==0 && ylasso[i,15]==0 && ylasso[i,16]==0 && ylasso[i,17]==0
      && ylasso[i,18]==0 && ylasso[i,19]==0 && ylasso[i,20]==0 && ylasso[i,21]==0
      && ylasso[i,22]==0 && ylasso[i,23]==0 && ylasso[i,24]==0 && ylasso[i,25]==0
      && ylasso[i,26]==0 && ylasso[i,27]==0 && ylasso[i,28]==0 && ylasso[i,29]==0
      && ylasso[i,30]==0 && ylasso[i,31]==0 && ylasso[i,32]==0 && ylasso[i,33]==0
      && ylasso[i,34]==0 && ylasso[i,35]==0 && ylasso[i,36]==0 && ylasso[i,37]==0
      && ylasso[i,38]==0 && ylasso[i,39]==0 && ylasso[i,40]==0 && ylasso[i,41]==0
      && ylasso[i,42]==0 && ylasso[i,43]==0 && ylasso[i,44]==0 && ylasso[i,45]==0
      && ylasso[i,46]==0 && ylasso[i,47]==0 && ylasso[i,48]==0 && ylasso[i,49]==0
      && ylasso[i,50]==0){
    counter12_10lasso=counter12_10lasso+1
  }
```

```
  if(ystepforw[i,1]==1 && ystepforw[i,2]==1 && ystepforw[i,5]==0 && ystepforw[i,3]==0
      && ystepforw[i,4]==0 && ystepforw[i,6]==0 && ystepforw[i,7]==0
      && ystepforw[i,8]==0 && ystepforw[i,9]==0 && ystepforw[i,10]==1 && ystepforw[i,11]==0
      && ystepforw[i,12]==0 && ystepforw[i,13]==0 && ystepforw[i,14]==0
      && ystepforw[i,15]==0 && ystepforw[i,16]==0 && ystepforw[i,17]==0 && ystepforw[i,18]==0
      && ystepforw[i,19]==0 && ystepforw[i,20]==0 && ystepforw[i,21]==0
      && ystepforw[i,22]==0 && ystepforw[i,23]==0 && ystepforw[i,24]==0 && ystepforw[i,25]==0
      && ystepforw[i,26]==0 && ystepforw[i,27]==0 && ystepforw[i,28]==0
      && ystepforw[i,29]==0 && ystepforw[i,30] && ystepforw[i,31]==0 && ystepforw[i,32]==0
      && ystepforw[i,33]==0 && ystepforw[i,34]==0 && ystepforw[i,35]==0
      && ystepforw[i,36]==0 && ystepforw[i,37]==0 && ystepforw[i,38]==0 && ystepforw[i,39]==0
      && ystepforw[i,40]==0 && ystepforw[i,41]==0 && ystepforw[i,42]==0
      && ystepforw[i,43]==0 && ystepforw[i,44]==0 && ystepforw[i,45]==0 && ystepforw[i,46]==0
      && ystepforw[i,47]==0 && ystepforw[i,48]==0 && ystepforw[i,49]==0
      && ystepforw[i,50]==0){
    counter12_10stepforw=counter12_10stepforw+1
  }
  if(ystepcomb[i,1]==1 && ystepcomb[i,2]==1 && ystepcomb[i,5]==0 && ystepcomb[i,3]==0
      && ystepcomb[i,4]==0 && ystepcomb[i,6]==0 && ystepcomb[i,7]==0
      && ystepcomb[i,8]==0 && ystepcomb[i,9]==0 && ystepcomb[i,10]==1 && ystepcomb[i,11]==0
      && ystepcomb[i,12]==0 && ystepcomb[i,13]==0 && ystepcomb[i,14]==0
      && ystepcomb[i,15]==0 && ystepcomb[i,16]==0 && ystepcomb[i,17]==0 && ystepcomb[i,18]==0
      && ystepcomb[i,19]==0 && ystepcomb[i,20]==0 && ystepcomb[i,21]==0
      && ystepcomb[i,22]==0 && ystepcomb[i,23]==0 && ystepcomb[i,24]==0 && ystepcomb[i,25]==0
      && ystepcomb[i,26]==0 && ystepcomb[i,27]==0 && ystepcomb[i,28]==0
      && ystepcomb[i,29]==0 && ystepcomb[i,30] && ystepcomb[i,31]==0 && ystepcomb[i,32]==0
      && ystepcomb[i,33]==0 && ystepcomb[i,34]==0 && ystepcomb[i,35]==0
      && ystepcomb[i,36]==0 && ystepcomb[i,37]==0 && ystepcomb[i,38]==0 && ystepcomb[i,39]==0
      && ystepcomb[i,40]==0 && ystepcomb[i,41]==0 && ystepcomb[i,42]==0
      && ystepcomb[i,43]==0 && ystepcomb[i,44]==0 && ystepcomb[i,45]==0 && ystepcomb[i,46]==0
      && ystepcomb[i,47]==0 && ystepcomb[i,48]==0 && ystepcomb[i,49]==0
      && ystepcomb[i,50]==0){
    counter12_10stepcomb=counter12_10stepcomb+1
  }

}



freq12_10=matrix(0,nrow=1,ncol = 3)
freq12_10[1,1]=counter12_10lasso/100
freq12_10[1,2]=counter12_10stepcomb/100
freq12_10[1,3]=counter12_10stepforw/100
colnames(freq12_10)=c("Lasso","StepComb","StepForward")
freq12_10
```

```r
# frequencies of close models to the truth:
counterContainedModellasso=0
counterContainedModelstepforw=0
counterContainedModelstepcomb=0
for (i in 1:100) {
  if(ylasso[i,1]==1 && ylasso[i,2]==1 && ylasso[i,10]==1 && sum(ylasso[i,]<=5)){
    counterContainedModellasso=counterContainedModellasso+1
  }
  if(ystepforw[i,1]==1 && ystepforw[i,2]==1 && ystepforw[i,10]==1 && sum(ystepforw[i,])<=5){
    counterContainedModelstepforw=counterContainedModelstepforw+1
  }
  if(ystepcomb[i,1]==1 && ystepcomb[i,2]==1 && ystepcomb[i,10]==1 && sum(ystepcomb[i,])<=5){
    counterContainedModelstepcomb=counterContainedModelstepcomb+1
  }
}


freq12_10con=matrix(0,nrow=1,ncol=3)
freq12_10con[1,1]=counterContainedModellasso/100
freq12_10con[1,2]=counterContainedModelstepcomb/100
freq12_10con[1,3]=counterContainedModelstepforw/100
colnames(freq12_10con)=c("Lasso","StepComb","StepForward")
freq12_10con


#coefficients and log lambdas
plot(mlasso, xvar = "lambda", label = TRUE)
abline(v=-0.632304,col="blue",lty=2)
text(-0.25,1.35,expression(λ==λ[se1]))
abline(v=-1.283540,col="red",lty=2)
text(-0.95,1.35,expression(λ==λ[min]))


#Cross Validation Error for a sequence of lambda values:
cv.lasso$lambda.1se
cv.lasso$lambda.min
plot(cv.lasso)
```

Διεθνής Βιβλιογραφία

Akaike, H. (1974). A new look at the statistical model identification. *IEEE transactions on automatic control*, **19**, p.716-723

Atilgan, T. (1996). Selection of dimension and basis for density estimation and selection of dimension, basis and error distribution for regression. *Communications in Statistics-Theory and Methods*, **25**, p.1-28

Box, G. E. (1976). Science and statistics. *Journal of the American Statistical Association*, **71**, p.791-799

Breiman, L. (2001). Stasticial modelling: The two cultures (with comments and a rejoinder by the author). *Statistical science*, **16**, p.199-231

Burnham, K. P. & Anderson, D. R. (2002). Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach, 2ed. New York

Claeskens, G. & Hjort, N. (2008). *Model Selection and Model Averaging* Cambridge Books, Cambridge University Press

Efron, B. & Gong, G. (1983). A Leisurely Look at the Bootstrap, the Jackknife, and the Cross-Validation. *The American Statistician,* **37**, p.36-48

Hastie, T., Tibshirani, R., & Friedman, J. (2009). *The elements of statistical learning: data mining, inference, and prediction.* Springer Science & Business Media, New York

Hastie, T., Tibshirani, R., & Wainwright, M. (2015). *Statistical learning with sparcity: the lasso and generalizations*. CRC press, New York

Hurvich, C. M., & Tsai, C. L. (1991). Bias of the corrected AIC criterion for underfitted regression and time series models. *Biometrica*, **78**, p.499-509

Konishi, S., & Kitagawa, G. (2008). *Information criteria and statistical modeling*. Springer Science & Business Media, New York

Rawlings, J. O., Pantula, S. G., & Dickey, D. A. (2001). *Applied regression analysis: a research tool.* Springer Science & Business Media, New York

Sin, C. Y., & White, H. (1996). Information criteria for selecting possibly misspecified parametric models. *Journal of Econometrics*, **71**, p.207-225

Tseng, P. (2001). Convergence of a block coordinate descent method for nondifferentiable minimization. *Journal of optimization theory and applications*, **109**, p.475-494

Vuong, Q.H. (1989). Likelihood ratio tests for model selection and non-nested hypotheses. *Econometrica: Journal of the Econometric Society,* **57**, p.307-333

Ελληνική Βιβλιογραφία

Καρώνη, Χ. (2009). Μοντέλα Αξιοπιστίας και Επιβίωσης, Εκδόσεις Συμεών, Αθήνα

Καρώνη, Χ. & Οικονόμου, Π. (2010). Στατιστικά Μοντέλα Παλινδρόμησης, Εκδόσεις Συμεών, Αθήνα

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