ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Γεώργιος Α. Σταθόπουλος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούλιος 2017

ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

Σχολή Χρηματοοικονομικής και Στατιστικής

Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ **ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΣΤΑΤΙΣΤΙΚΕΣ ΤΕΧΝΙΚΕΣ ΠΑΛΙΝΔΡΟΜΗΣΗΣ ΓΙΑ ΤΗΝ ΑΝΑΛΥΣΗ ΜΕΓΑΛΩΝ ΔΕΔΟΜΕΝΩΝ

Γεώργιος Α. Σταθόπουλος

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην *Εφαρμοσμένη Στατιστική*

Πειραιάς
Ιούλιος 2017

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. ……. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- Καθηγητής Μ. Κούτρας (Επιβλέπων)
- Καθηγητής Γ. Θεοδωρίδης
- Επίκουρος Καθηγητής Ν. Πελέκης

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

School of Finance and Statistics

Department of Statistics and Insurance Science

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

REGRESSION TECHNIQUES FOR THE ANALYSIS OF BIG DATA

By

George A. Stathopoulos

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
July 2017

Στους γονείς μου

Αχιλλέα και Στέλλα

Ευχαριστίες

Θα ήθελα να ευχαριστήσω τον επιβλέποντα καθηγητή μου κ. Μ. Κούτρα για την ανάθεση του θέματος, την καθοδήγησή του και τη στήριξή του στην εκπόνηση της παρούσας εργασίας. Ακόμη θα ήθελα να ευχαριστήσω την οικογένειά μου, τη σύντροφό μου και τους φίλους μου που με στηρίζουν σε κάθε προσπάθειά μου.

Περίληψη

Στη σύγχρονη εποχή, η δυνατότητα συλλογής μεγάλων όγκων δεδομένων τα οποία συνήθως αφορούν πολλά χαρακτηριστικά έχει οδηγήσει στην ανάγκη εφαρμογής και ανάπτυξης ειδικών στατιστικών τεχνικών που θα βοηθήσουν στη μελέτη της δομής τους και την εξαγωγή χρήσιμων συμπερασμάτων. Έχει διαπιστωθεί ότι, για την ανάλυση τέτοιων δεδομένων, απαιτείται είτε κάποια ειδική προσαρμογή των διαθέσιμων κλασσικών στατιστικών τεχνικών είτε η ανάπτυξη εναλλακτικών τεχνικών.

Στα πλαίσια της παρούσας εργασίας θα γίνει σύντομη παρουσίαση τεχνικών που χρησιμοποιούνται για την ανάλυση μεγάλων δεδομένων, προγραμματισμός των αντίστοιχων αλγορίθμων σε περιβάλλον R και σύγκριση των διάφορων τεχνικών με τη χρήση πραγματικών δεδομένων.

Πιο συγκεκριμένα θα χρησιμοποιηθούν τεχνικές που βασίζονται στα δέντρα παλινδρόμησης και μέσω αλγορίθμων μηχανικής μάθησης βελτιώνεται η ακρίβεια των τους. Ακόμη θα χρησιμοποιηθούν τεχνικές που βασίζονται στο κλασσικό μοντέλο γραμμικής παλινδρόμησης, στο οποίο όμως εφαρμόζεται ένας περιορισμός (ποινή) στους συντελεστές του μοντέλου, ώστε αντιμετωπιστεί το φαινόμενο της πολυσυγγραμμικότητας, να μειωθεί η διασπορά των συντελεστών και να επιτευχθεί μεγαλύτερη ακρίβεια πρόβλεψης.

Abstract

In modern times, the collection of large amounts of data which usually involves a lot of features has led to the need for the development of specific statistical techniques that will help studying their structure and extracting useful conclusions for them. It has been found that, for such an analysis it is required either a special adaptation of available conventional statistical techniques or the development of alternative techniques.

In this paper we will provide a brief presentation of techniques used for the analysis of large amounts of data, and describe the respective algorithms along with the codes in R environment. In addition a comparison study of those techniques on real data is carried out.

Specifically there will be used tree based regression techniques which are improved by exploiting machine learning algorithms. Furthermore there be used techniques based on classical linear regression model, in which a restriction (penalty) is applied on the model coefficients in order to address the phenomenon of multicollinearity, reduce the variance of the coefficients and achieve greater prediction accuracy.

Περιεχόμενα

Κατάλογος Σχημάτων iii

1. Κεφάλαιο 1: Εισαγωγή 1

2. Κεφάλαιο 2: Δέντρα Παλινδρόμησης 5

2.1 Εισαγωγή στα Δέντρα Παλινδρόμησης 5

2.2 *Stochastic Gradient Boosting* 8

2.3 *Random Forests* 11

3. Κεφάλαιο 3: Εισαγωγή στις τεχνικές παλινδρόμησης με ποινή 15

3.1 Επιλογή μεταβλητών στο μοντέλο της γραμμικής παλινδρόμησης 15

3.2 Η τεχνική *non-negative garrote estimator* 20

4. Κεφάλαιο 4: Παλινδρόμηση τύπου *Ridge* 23

4.1 Εισαγωγή στην παλινδρόμηση τύπου *Ridge* 23

4.2 Η παράμετρος λ στην παλινδρόμηση τύπου *Ridge* 25

5. Κεφάλαιο 5: Μέθοδος *LASSO* 31

5.1 Εισαγωγή στη μέθοδο *LASSO* 31

5.2 Η παράμετρος λ στη μέθοδο *LASSO* 34

6. Κεφάλαιο 6: Σύνοψη 39

Βιβλιογραφία 41

Κατάλογος Σχημάτων

2.1.1	Διαμέριση του χώρου δύο μεταβλητών	5
2.1.2	Δέντρο παλινδρόμησης 2 μεταβλητών	6
2.2.1	Προσαρμογή δέντρων παλινδρόμησης με τη μέθοδο *Stochastic Gradient Boosting*	8
2.2.2	Εφαρμογή της μεθόδου *Stochastic Gradient Boosting*	10
2.3.1	Προσαρμογή δέντρων παλινδρόμησης με τη μέθοδο *Random Forests*	12
2.3.2	Εφαρμογή της μεθόδου *Random Forests*	13
3.1.1	Εφαρμογή της κατά βήματα παλινδρόμησης	20
4.1.1	Περιορισμός *Ridge* σε δύο διαστάσεις	24
4.2.1	Μεταβολές με βάση την παράμετρο λ	26
4.2.2	Μεταβολή των εκτιμητριών *Ridge* σε σχέση με την παράμετρο λ	27
4.2.3	Μεταβολή του *MSE* σε σχέση με την παράμετρο λ	28
4.2.4	Μεταβολή του *GCV* σε σχέση με την παράμετρο λ	29
5.1.1	Περιορισμός *LASSO* σε 2 διαστάσεις	32
5.2.1	Μεταβολή των εκτιμητριών *LASSO* σε σχέση με την παράμετρο λ	35
5.2.2	Σύγκριση της μεταβολής των εκτιμητριών *Ridge και LASSO* σε σχέση με την παράμετρο λ.	35
5.2.3	Μεταβολή του *MSE* σε σχέση με την παράμετρο λ	36

ΚΕΦΑΛΑΙΟ 1

Εισαγωγή

Η επιστήμη της Στατιστικής συνεχώς καλείται να δώσει απαντήσεις σε προβλήματα που τίθενται, τόσο σε επιστημονικό όσο και σε επιχειρηματικό επίπεδο. Παλαιότερα τα προβλήματα που καλούταν να αντιμετωπίσει προέρχονταν κυρίως από το χώρο της βιομηχανίας και της γεωργίας, τα οποία ήταν μικρής κλίμακας.

Ωστόσο τα τελευταία χρόνια με την ανάπτυξη της τεχνολογίας και των ηλεκτρονικών υπολογιστών, παρέχεται η δυνατότητα αποθήκευσης και επεξεργασίας μεγάλου όγκου δεδομένων. Αυτό έχει ως αποτέλεσμα τα προβλήματα που καλείται να αντιμετωπίσει η Στατιστική να έχουν αυξηθεί σημαντικά τόσο σε όγκο όσο και σε πολυπλοκότητα. Σε πολλούς κλάδους, όπως στις επιστήμες, στα οικονομικά και τις επιχειρήσεις, παράγονται μεγάλοι όγκοι δεδομένων, που με την κατάλληλη επεξεργασία και την εφαρμογή κατάλληλων στατιστικών τεχνικών είναι δυνατή η εξαγωγή σημαντικών αποτελεσμάτων.

Ένα από τα πιο συχνά προβλήματα που καλείται να αντιμετωπίσει η στατιστική, είναι η πρόβλεψη της τιμής ενός αντικειμένου, όταν είναι γνωστές οι τιμές άλλων αντικειμένων που πιθανώς σχετίζονται με αυτό.

Μία από τις πιο κλασσικές τεχνικές που εφαρμόζονται είναι η τεχνική της γραμμικής παλινδρόμησης. Με βάση την τεχνική αυτή εφαρμόζεται ένα γραμμικό μοντέλο που περιγράφει την επίδραση που έχουν κάποιες επεξηγηματικές μεταβλητές σε μία μεταβλητή απόκρισης. Οι συντελεστές αυτού του γραμμικού μοντέλου υπολογίζονται με την τεχνική των ελαχίστων τετραγώνων.

Ωστόσο, ενώ οι εκτιμήτριες ελαχίστων τετραγώνων είναι αμερόληπτες, υστερούν σε ακρίβεια πρόβλεψης. Επιπλέον, όταν το πλήθος των μεταβλητών είναι μεγάλο, παρουσιάζεται το πρόβλημα της επιλογής εκείνων των μεταβλητών που θα συμμετέχουν στο μοντέλο, καθώς υπάρχουν πάρα πολλοί δυνατοί συνδυασμοί μεταβλητών που μπορούν να χρησιμοποιηθούν. Ένα επιπρόσθετο πρόβλημα που προκύπτει όταν οι συσχετίσεις των μεταβλητών είναι υψηλές, είναι αυτό της πολυσυγγραμμικότητας.

Στα πλαίσια της παρούσας διπλωματικής εργασίας θα περιγράψουμε κάποιες τεχνικές οι οποίες επιχειρούν να διορθώσουν τα παραπάνω προβλήματα και είναι καταλληλότερες για την ανάλυση μεγάλων δεδομένων.

Μία προσέγγιση είναι η εφαρμογή τεχνικών που βασίζονται στα δέντρα παλινδρόμησης *(Tree-based methods)*. Οι τεχνικές αυτές είναι απλές, εύκολες στη χρήση και έχουν τη

δυνατότητα να προσαρμόζονται στα δεδομένα. Όμως, σε σχέση με άλλες τεχνικές τα δέντρα παλινδρόμησης υστερούν σε ακρίβεια.

Για τη βελτίωση της ακρίβειας των δέντρων παλινδρόμησης θα περιγράψουμε δύο μεθόδους που βασίζονται σε δύο διαφορετικές τεχνικές μηχανικής μάθησης. Η πρώτη μέθοδος είναι η *Stochastic Gradient Boosting,* η οποία ανήκει σε μία κατηγορία μεθόδων *(boosting methods),* οι οποίες βελτιώνουν την ακρίβεια του αρχικού μοντέλου κατασκευάζοντας διαδοχικά συμπληρωματικά μοντέλα που προστίθενται στο αρχικό. Η δεύτερη μέθοδος είναι η *Random Forests* η οποία ανήκει σε μία κατηγορία μεθόδων *(bagging methods),* με βάση τις οποίες συνδυάζονται πολλά μοντέλα που έχουν προσαρμοσθεί σε παραλλαγές του πρωτότυπου συνόλου δεδομένων.

Μία άλλη προσέγγιση του προβλήματος είναι η εφαρμογή τεχνικών παλινδρόμησης με ποινή. Σε αυτές τις τεχνικές εφαρμόζεται ένας περιορισμός στον υπολογισμό των εκτιμητριών της παλινδρόμησης, με βάση τον οποίο συρρικνώνονται οι τιμές τους, με σκοπό να επιτευχθεί μεγαλύτερη ακρίβεια πρόβλεψης.

Οι τεχνικές που θα αναλύσουμε στη συνέχεια είναι η τεχνική *non-negative garrote estimator*, η οποία συρρικνώνει τους συντελεστές παλινδρόμησης κατά έναν μη αρνητικό παράγονται, η τεχνική παλινδρόμησης τύπου *Ridge,* η οποία εφαρμόζει έναν περιορισμό στο άθροισμα των τετραγώνων των συντελεστών παλινδρόμησης και η τεχνική παλινδρόμησης *LASSO,* η οποία εφαρμόζει έναν περιορισμό στο άθροισμα των απολύτων τιμών των συντελεστών παλινδρόμησης.

Για να μελετήσουμε την εφαρμογή των παραπάνω τεχνικών σε πραγματικά δεδομένα και για να συγκρίνουμε την αποτελεσματικότητα τους, θα χρησιμοποιήσουμε ένα σύνολο δεδομένων, στο οποίο καταγράφονται κάποια χαρακτηριστικά κατοικιών, με σκοπό την πρόβλεψη της τιμής τους. Το σύνολο δεδομένων αναφέρεται σε πωλήσεις κατοικιών στην περιοχή *King Country, Washington* των Η.Π.Α, στο διάστημα από τον Μάιο του 2014 έως τον Μάιο του 2015 και αποτελείται από 21613 διαφορετικές εγγραφές ενώ περιλαμβάνει 18 επεξηγηματικές μεταβλητές και μία εξαρτημένη μεταβλητή.

Στον παρακάτω πίνακα περιγράφονται οι 19 μεταβλητές του συνόλου δεδομένων:

ΠΙΝΑΚΑΣ 1	
Όνομα μεταβλητής	Περιγραφή
price	Τιμή πώλησης της κατοικίας
bedrooms	Πλήθος των υπνοδωματίων
Bathrooms	Πλήθος των μπάνιων
sqft_living	Επιφάνεια της κατοικίας σε τετραγωνικά πόδια
sqft_lot	Επιφάνεια του οικοπέδου σε τετραγωνικά πόδια
floors	Πλήθος των ορόφων
waterfront	Δηλώνει αν οι κατοικία έχει θέα στην παραλία
view	Αριθμός των υποψήφιων αγοραστών που έχουν δει την κατοικία
condition	Συνολική κατάσταση της κατοικίας
Grade	Βαθμός με βάση το σύστημα βαθμολόγησης του *King Country*
sqft_above	Επιφάνεια της κατοικίας σε τετραγωνικά πόδια χωρίς το υπόγειο
sqft_basement	Επιφάνεια του υπογείου σε τετραγωνικά πόδια
yr_built	Έτος κατασκευής
yr_renovated	Έτος ανακαίνισης
zipcode	Ταχυδρομικός κώδικας
lat	Γεωγραφικό πλάτος
long	Γεωγραφικό μήκος
sqft_living15	Επιφάνεια της κατοικίας το 2015 (υπονοεί ανακαίνιση)
sqft_lot15	Επιφάνεια του οικοπέδου το 2015 (υπονοεί ανακαίνιση)

ΚΕΦΑΛΑΙΟ 2

Δέντρα Παλινδρόμησης

2.1 Εισαγωγή στα Δέντρα Παλινδρόμησης

Σε πολλές περιπτώσεις όπου ο όγκος των δεδομένων συνεχώς αυξάνεται και μεταβάλλεται, αναζητούμε τεχνικές οι οποίες είναι απλές, εύκολες στη χρήση και έχουν τη δυνατότητα να προσαρμόζονται στα δεδομένα.

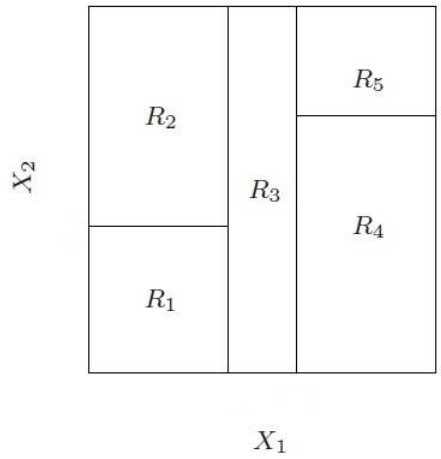
Μία τέτοια κατηγορία μεθόδων είναι οι Tree-based methods που βασίζονται στα δέντρα απόφασης για την ανάλυση των δεδομένων. Οι μέθοδοι αυτές είναι απλές και βοηθούν στην εύκολη ερμηνεία των αποτελεσμάτων. Χρησιμοποιούνται τόσο σε προβλήματα παλινδρόμησης όσο και σε προβλήματα κατηγοριοποίησης.

Οι μέθοδοι αυτές είναι απλές και εύκολες στην ερμηνεία γιατί έχουν διακριτά βήματα και κάθε βήμα ορίζεται από έναν απλό λογικό κανόνα.

Θεωρούμε ένα σύνολο δεδομένων της μορφής $\{(y_i, x_{i1}, \dots, x_{ip}), i = 1, \dots, n\}$, όπου $y_i, i = 1, \dots, n$ οι τιμές της μεταβλητής απόκρισης Y, και $x_{ik}, i = 1, \dots, n, k = 1, \dots, p$ οι αντίστοιχες τιμές των επεξηγηματικών μεταβλητών X_k.

Η κατασκευή ενός δέντρου παλινδρόμησης χωρίζεται πρακτικά σε δύο βήματα:

- Αρχικά διαμερίζεται ο χώρος των επεξηγηματικών μεταβλητών σε K μη επικαλυπτόμενους υποχώρους $R_1, R_2, \dots, R_K$.
- Για κάθε παρατήρηση η οποία ανήκει σε έναν υπόχωρο έχουμε την ίδια εκτίμηση για την εξαρτημένη μεταβλητή.



Σχήμα 2.1.1 Διαμέριση του χώρου δύο μεταβλητών

Για την κατασκευή των K μη επικαλυπτόμενων υποχώρων στόχος είναι η ελαχιστοποίηση μίας συνάρτησης κόστους. Η συνάρτηση κόστους που χρησιμοποιούμε συνήθως είναι το RSS που δίνεται από τον τύπο:

$$RSS = \sum_{k=1}^{K} \sum_{i \in R_k} (y_i - \hat{y}_{R_k})^2$$

όπου $\hat{y}_{R_k}$ είναι η εκτίμηση της μεταβλητής απόκρισης για τις παρατηρήσεις που βρίσκονται στον υπόχωρο R_k. Υπολογιστικά είναι ανέφικτο να θεωρήσουμε όλους τους δυνατούς υπόχωρους και για αυτό το λόγο επιλέγουμε μία διαφορετική προσέγγιση κατά την οποία σε κάθε βήμα ο χώρος χωρίζεται σε δύο υποχώρους μέσα από δύο νέα κλαδιά. Ο διαχωρισμός σε κάθε βήμα γίνεται επιλέγοντας μία επεξηγηματική μεταβλητή X_j και ένα σημείο διαχωρισμού s με βάση τα οποία επιτυγχάνεται η μεγαλύτερη δυνατή μείωση του RSS.

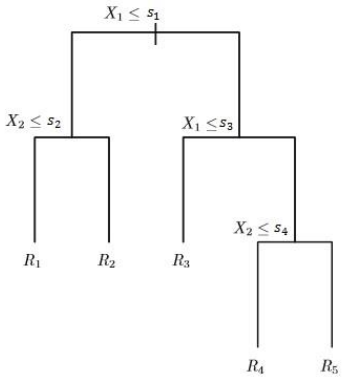
Πιο συγκεκριμένα αναζητούμε τα j και s για το οποία ελαχιστοποιείται η συνάρτηση:

$$\sum_{i:x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

όπου

$$R_1(j,s) = \{\boldsymbol{X} | X_j < s\} \text{ και } R_2(j,s) = \{\boldsymbol{X} | X_j \geq s\}.$$

Στη συνέχεια επαναλαμβάνεται η ίδια διαδικασία για κάθε έναν από τους υποχώρους που δημιουργήθηκαν, προσθέτοντας έτσι δύο νέα κλαδιά για κάθε υπόχωρο στο δέντρο. Η διαδικασία αυτή επαναλαμβάνεται μέχρι να ικανοποιηθεί κάποιο κριτήριο που σταματάει τον αλγόριθμο.



Σχήμα 2.1.2 Δέντρο παλινδρόμησης 2 μεταβλητών

Όταν έχουν δημιουργηθεί οι επιθυμητοί υπόχωροι δίνεται η εκτίμηση για την μεταβλητή απόκρισης χρησιμοποιώντας τις παρατηρήσεις που υπάρχουν σε κάθε υπόχωρο. Ως εκτίμηση

δίνεται ο μέσος όρος των τιμών της μεταβλητής απόκρισης για όλες τις παρατηρήσεις που βρίσκονται σε έναν υπόχωρο.

Συχνά η διαδικασία αυτή οδηγεί σε καλά αποτελέσματα για τα δεδομένα εκμάθησης του μοντέλου. Υπάρχει όμως η πιθανότητα της υπερπροσαρμογής του στα συγκεκριμένα δεδομένα, δηλαδή προσαρμόζοντας το μοντέλο σε ένα νέο σετ δεδομένων να μην έχει καλή απόδοση. Αυτό μπορεί να συμβαίνει γιατί το μοντέλο που έχει παραχθεί μπορεί να είναι ιδιαίτερα πολύπλοκο. Για αυτό το λόγο κατασκευάζουμε μικρότερα δέντρα με λιγότερα κλαδιά, δηλαδή λιγότερους τελικούς υποχώρους, για να έχουμε καλύτερη προσαρμοστικότητα και ερμηνεία του μοντέλου.

Ένα προτεινόμενο κριτήριο, ώστε να σταματήσει το δέντρο να μεγαλώνει είναι, όταν ο επόμενος διαχωρισμός που προκύπτει προσθέτοντας δύο νέα κλαδιά δεν προσφέρει κάποια σημαντική μείωση του *RSS*. Όμως με αυτόν τον τρόπο ένας διαχωρισμός που δεν είναι σημαντικός, μπορεί να οδηγεί σε ένα διαχωρισμό που προσφέρει σημαντική μείωση του RSS. Επιλέγουμε λοιπόν να κατασκευάζουμε μεγαλύτερα δέντρα και στην συνέχεια να τα ΄΄κλαδεύουμε΄΄, αφαιρώντας κλαδιά και κατασκευάζοντας δέντρα με λιγότερους κόμβους από το αρχικό.

Για να επιλέξουμε το πλέον κατάλληλο δέντρο χρησιμοποιούμε μία μέθοδο που προσθέτει κόστος για κάθε κόμβο που περιλαμβάνει το δέντρο. Έστω $|T|$ ο αριθμός των τελικών κόμβων σε ένα δέντρο και:

$$N_k = \#\{x_i \in R_k\},$$

$$\hat{y}_k = \frac{1}{N_k} \sum_{x_i \in R_k} y_i,$$

$$Q_k(T) = \frac{1}{N_k} \sum_{x_i \in R_k} (y_i - \hat{y}_k)^2.$$

Η συνάρτηση κόστους που επιθυμούμε να ελαχιστοποιήσουμε δίνεται από τη σχέση:

$$C_a(T) = \sum_{k=1}^{|T|} N_k \, Q_k(T) + a|T|.$$

Η παράμετρος $a \geq 0$ καθορίζει τον συμβιβασμό μεταξύ του μεγέθους ενός δέντρου και της προσαρμογής του στα δεδομένα. Μεγάλες τιμές της παραμέτρου a έχουν ως αποτέλεσμα μικρότερα δέντρα.

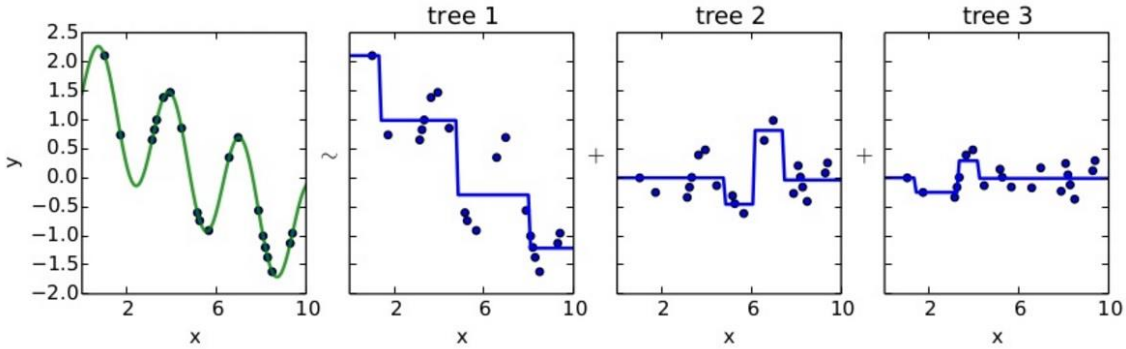
2.2 *Stochastic Gradient Boosting*

Ένα δέντρο παλινδρόμησης έχει το πλεονέκτημα ότι αποτελεί μία γρήγορη μέθοδο που προσαρμόζεται εύκολα στα δεδομένα, όμως δεν έχει την αποτελεσματικότητα που έχουν άλλες μέθοδοι όσο αφορά στην ακρίβεια, δηλαδή οι εκτιμήσεις του μοντέλου έχουν σημαντικές αποκλίσεις από τις πραγματικές τιμές της μεταβλητής απόκρισης. Για να βελτιώσουμε την ακρίβεια στα δένδρα παλινδρόμησης υπάρχουν διάφορες τεχνικές που εφαρμόζονται. Μία από τις πιο διαδεδομένες είναι η κατασκευή συμπληρωματικών δέντρων παλινδρόμησης τα οποία βελτιώνουν την ακρίβεια του μοντέλου, προσθέτοντας τα στο αρχικό μοντέλο *(boosting methods)*. Μία τέτοια μέθοδος είναι η *Stochastic Gradient Boosting*.

Σύμφωνα με αυτή την μέθοδο παράγουμε ένα δέντρο παλινδρόμησης ακολουθώντας τη διαδικασία που περιγράψαμε παραπάνω. Το μοντέλο που έχει παραχθεί δίνεται από τη συνάρτηση F και για κάθε παρατήρηση y_i παίρνουμε την εκτίμηση $F(\boldsymbol{x}_i) = \hat{y}_i$. Η απόκλιση της εκτίμησης από την πραγματική τιμή, υπολογίζεται από τα κατάλοιπα $\hat{y}_i - y_i$.

Για τη βελτίωση της ακρίβειας η μέθοδος προτείνει τη δημιουργία ενός νέου δέντρου παλινδρόμησης στα κατάλοιπα του προηγούμενου, δηλαδή χρησιμοποιώντας ως μεταβλητή απόκρισης τα κατάλοιπα $F(\boldsymbol{x_i}) - y_i$. Έτσι λοιπόν παράγεται ένα νέο δέντρο παλινδρόμησης h το οποίο όταν προστεθεί στο αρχικό μοντέλο βελτιώνει την ακρίβεια του.

Η νέα εκτίμηση που προκύπτει για κάθε παρατήρηση είναι της μορφής $F(\boldsymbol{x_i}) + \rho \cdot h(\boldsymbol{x_i})$, όπου η παράμετρος $0 < \rho \leq 1$ ορίζει το ρυθμό μάθησης του αλγορίθμου. Για να επιτύχουμε την ακρίβεια που επιθυμούμε μπορούμε να επαναλάβουμε τη διαδικασία, προσθέτοντας κάθε φορά στο μοντέλο ένα νέο δέντρο παλινδρόμησης που έχει προσαρμοστεί στα κατάλοιπα του προηγούμενου.



Σχήμα 2.2.1 Προσαρμογή δέντρων παλινδρόμησης με τη μέθοδο *Stochastic Gradient Boosting*

Στο παραπάνω σχήμα περιγράφεται η διαδικασία προσαρμογής των δέντρων παλινδρόμησης με τη μέθοδο *Stochastic Gradient Boosting*, για μία μεταβλητή απόκρισης Y και μία επεξηγηματική μεταβλητή X. Το πρώτο δέντρο προσαρμόζεται χρησιμοποιώντας ως μεταβλητή απόκρισης την μεταβλητή Y. Το δεύτερο δέντρο προσαρμόζεται χρησιμοποιώντας

ως μεταβλητή απόκρισης τα κατάλοιπα του πρώτου δέντρου και προστίθεται στο αρχικό. Τέλος το τρίτο δέντρο προσαρμόζεται στα κατάλοιπα του μοντέλου που παράγεται από το άθροισμα του πρώτου και του δεύτερου δέντρου, και προστίθεται και αυτό στο προηγούμενο μοντέλο.

Για να είναι αυτή η μέθοδος στοχαστική σε κάθε δέντρο που προσθέτουμε στο μοντέλο επιλέγουμε τυχαία τμήμα των δεδομένων. Για να επιτευχθεί αυτό πραγματοποιούμε μία τυχαία μετάθεση στις εγγραφές των δεδομένων και επιλέγουμε μικρότερο πλήθος εγγραφών για να προσαρμόσουμε το νέο δέντρο.

Έστω $\{y_i, \boldsymbol{x_i}\}_{i=1}^{N}$ ολόκληρο το σύνολο δεδομένων και $\{\pi(i)\}_{i=1}^{N}$ μία τυχαία μετάθεση των ακεραίων $\{1, \dots, N\}$. Επιλέγουμε λοιπόν ένα τυχαίο δείγμα $\widetilde{N} < N$, το οποίο δίνεται από την τυχαία μετάθεση, $\{y_{\pi(i)}, \boldsymbol{x}_{\pi(i)}\}_{i=1}^{\widetilde{N}}$ με βάση το οποίο προσαρμόζεται το νέο δέντρο το οποίο θα προστεθεί στο αρχικό μοντέλο.

Όσο πιο μικρό είναι το δείγμα $\widetilde{N}$ τόσο διαφορετικά θα είναι τα τυχαία δείγματα που χρησιμοποιούνται σε κάθε επανάληψη, για να δημιουργηθεί ένα νέο δέντρο. Έτσι εισάγεται μεγαλύτερη τυχαιότητα στο συνολικό μοντέλο, γεγονός που μπορεί να οδηγήσει σε καλύτερα αποτελέσματα όταν το μοντέλο εφαρμοστεί σε ένα νέο σύνολο δεδομένων. Χρησιμοποιώντας όμως πολύ μικρά δείγματα μειώνεται η διαθέσιμη πληροφορία για την προσαρμογή ενός μοντέλου με μεγαλύτερη ακρίβεια.

Για να μελετήσουμε την ακρίβεια της μεθόδου και να την συγκρίνουμε με άλλες τεχνικές χρησιμοποιούμε το σύνολο δεδομένων για την πρόβλεψη της τιμής των κατοικιών. Η υλοποίηση της μεθόδου πραγματοποιήθηκε σε περιβάλλον *R* με την χρήση του πακέτου *'gbm'*. Επίσης χρησιμοποιήθηκε το πακέτο *'caret'* με τη βοήθεια του οποίου επιλέγουμε τις βέλτιστες τιμές των παραμέτρων της μεθόδου. Το πακέτο *'caret'* εφαρμόζει διαδοχικά όλους τους συνδυασμούς για τις τιμές των παραμέτρων που έχουμε ορίσει και επιλέγει τις τιμές των παραμέτρων για τις οποίες επιτυγχάνεται η μέγιστη δυνατή ακρίβεια του μοντέλου που εφαρμόζεται.

Η ακρίβεια του μοντέλου αξιολογείται με βάση το *RMSE*, που είναι η τετραγωνική ρίζα του μέσου τετραγωνικού σφάλματος, και δίνεται από τον τύπο:

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

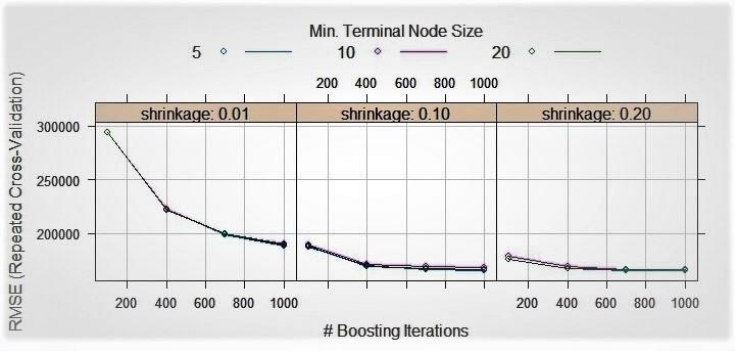
Οι παράμετροι της μεθόδου που πρέπει να ορίσουμε είναι το πλήθος των δέντρων που κατασκευάζει η μέθοδος, τον ρυθμό μάθησης και τον ελάχιστο αριθμό τελικών κόμβων. Στον παρακάτω πίνακα δίνονται οι τιμές των παραμέτρων για τις οποίες εφαρμόζεται η μέθοδος:

Πίνακας 2	
Πλήθος δέντρων	100 300 700 1000
Ρυθμός μάθησης	0,01 0,1 0,2
Ελάχιστος αριθμός τελικών κόμβων	5 10 20

Ο αλγόριθμος ο οποίος υλοποιεί την μέθοδο είναι ο εξής:

```
library(caret)
library(gbm)
trainControl<-trainControl(method="repeatedcv",
                           number=10,repeats=1,
                           classProbs=F)

gbmGrid<-expand.grid(n.trees = seq(100,1000, by=300),
                     shrinkage = c(0.01,0.1,0.2),
                     n.minobsinnode = c(5,10,20))

gbmFit<-train(price~.,data = houses,
              method="gbm",
              metric="RMSE",
              trControl=trainControl,
              tuneGrid=gbmGrid,
              verbose=FALSE)
gbmFit
plot(gbmFit)
gbmpred<-predict(gbmFit,houses[,-1])
DMwR::regr.eval(houses[,1],gbmpred)
```

Ο παραπάνω αλγόριθμος εμφανίζει το παρακάτω σχήμα:



Σχήμα 2.2.2 Εφαρμογή της μεθόδου *Stochastic Gradient Boosting*

Στο παραπάνω σχήμα βλέπουμε την ακρίβεια που επιτυγχάνει η μέθοδος για τις διάφορες τιμές των παραμέτρων που την επηρεάζουν, δηλαδή τον αριθμό των δέντρων που κατασκευάζονται, τον ρυθμό μάθησης και τον ελάχιστο αριθμό τελικών κόμβων. Παρατηρούμε ότι όταν ο αριθμός των δέντρων και ο ρυθμός μάθησης του μοντέλου είναι μικρός το μοντέλο που κατασκευάζεται έχει μεγαλύτερη απόκλιση. Ακόμη μπορούμε να παρατηρήσουμε ότι η αύξηση του ρυθμού μάθησης από $\rho = 0{,}1$ σε $\rho = 0{,}2$ δεν βελτιώνει σημαντικά την ακρίβεια.

Η μικρότερη τιμή του μέσου τετραγωνικού σφάλματος που επιτυγχάνεται με την μέθοδο *Stochastic Gradient Boosting* είναι $MSE = 2{,}244 \cdot 10^{10}$.

2.3 Random Forests

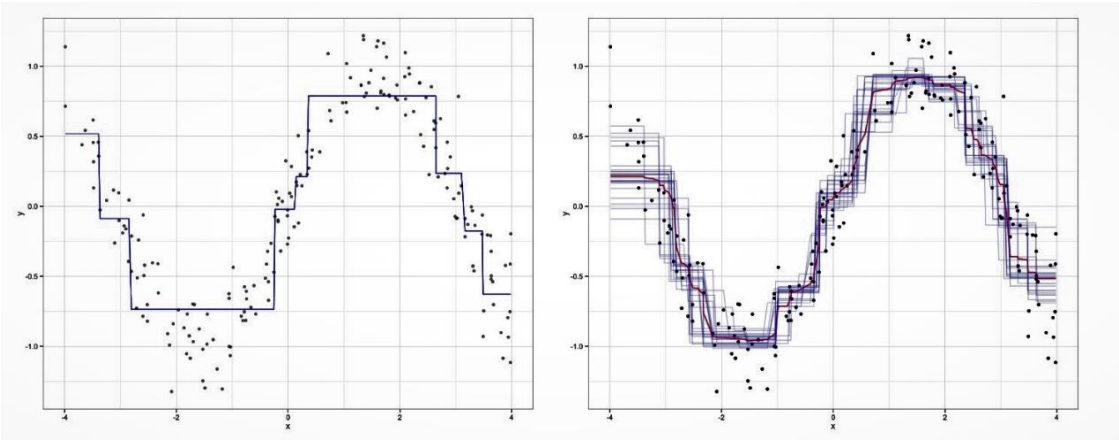
Μία άλλη διαδικασία που εφαρμόζεται για να πετύχουμε μεγαλύτερη ακρίβεια είναι η κατασκευή πολλών παραλλαγών του πρωτότυπου αρχείου δεδομένων, στα οποία προσαρμόζονται δέντρα παλινδρόμησης και το τελικό μοντέλο παράγεται συνδυάζοντας τα δέντρα που έχουν παραχθεί *(bagging methods)*. Μία τέτοια διαδικασία είναι η τεχνική *Random Forests*.

Αρχικά επιλέγεται ένα τυχαίο δείγμα από το σύνολο δεδομένων με την μέθοδο Bootstrap. Με βάση την μέθοδο αυτή από ένα σύνολο δεδομένων με N το πλήθος παρατηρήσεις επιλέγεται τυχαία ένα δείγμα από N παρατηρήσεις με επανάθεση, δηλαδή στο δείγμα των δεδομένων μπορεί να συμμετέχει μία παρατήρηση περισσότερες από μία φορά. Στη συνέχεια με βάση το τυχαίο δείγμα κατασκευάζουμε ένα δέντρο παλινδρόμησης, όμως σε κάθε κόμβο του δέντρου δεν χρησιμοποιούνται όλες οι διαθέσιμες μεταβλητές. Επιλέγουμε τυχαία m από τις συνολικά p μεταβλητές και βρίσκουμε τον καλύτερο διαχωρισμό με βάση τις m μεταβλητές που έχουμε επιλέξει. Στον επόμενο κόμβο επιλέγουμε τυχαία m διαφορετικές μεταβλητές και βρίσκουμε τον επόμενο καλύτερο διαχωρισμό. Με αυτόν τον τρόπο κατασκευάζουμε ένα δέντρο παλινδρόμησης στο οποίο σε κάθε κόμβο έχουν χρησιμοποιηθεί διαφορετικές μεταβλητές για να δημιουργηθούν τα δύο νέα κλαδιά.

Επαναλαμβάνουμε αυτή τη διαδικασία και κατασκευάζουμε πολλά δέντρα παλινδρόμησης, τα οποία είναι διαφορετικά μεταξύ τους καθώς διαφέρει το αρχικό δείγμα από το οποίο κατασκευάζουμε το κάθε ένα από αυτά ενώ παράλληλα σε κάθε κόμβο χρησιμοποιούνται τυχαία διαφορετικές μεταβλητές.

Για να καταλήξουμε στο τελικό μοντέλο, για κάθε εγγραφή στα δεδομένα μας χρησιμοποιούμε ως εκτίμηση τον μέσο όρο των εκτιμήσεων που προκύπτουν από τα δέντρα που έχουμε κατασκευάσει.

Σε αυτή την μέθοδο η μόνη παράμετρος που πρέπει να οριστεί εξ αρχής είναι το πλήθος m των μεταβλητών που επιλέγονται σε κάθε κόμβο. Η συνήθης επιλογή που γίνεται για το m είναι $m = \sqrt{p}$.



Σχήμα 2.3.1 Προσαρμογή δέντρου παλινδρόμησης με τη μέθοδο *Random Forests*

Στο παραπάνω σχήμα παρουσιάζεται η προσαρμογή ενός δέντρου παλινδρόμησης με τη μέθοδο *Random Forests* για μία μεταβλητή απόκρισης Y και μία επεξηγηματική μεταβλητή X. Στην αριστερή εικόνα προσαρμόζεται ένα απλό δέντρο παλινδρόμησης στα ζεύγη των παρατηρήσεων, ενώ στη δεξιά παρουσιάζονται τα δέντρα που παράγει η μέθοδος και με κόκκινη γραμμή φαίνεται το μοντέλο που δίνει την εκτίμηση για τη μεταβλητή απόκρισης.

Για να μελετήσουμε την ακρίβεια της μεθόδου και να την συγκρίνουμε με άλλες τεχνικές χρησιμοποιούμε το σύνολο δεδομένων για την πρόβλεψη της τιμής των κατοικιών. Η υλοποίηση της μεθόδου πραγματοποιήθηκε σε περιβάλλον *R* με την χρήση του πακέτου *'randomForest'* και του πακέτου *'caret'*.

Στον παρακάτω πίνακα παρουσιάζονται οι τιμές της παραμέτρου *m* για τις οποίες εφαρμόζεται η μέθοδος:

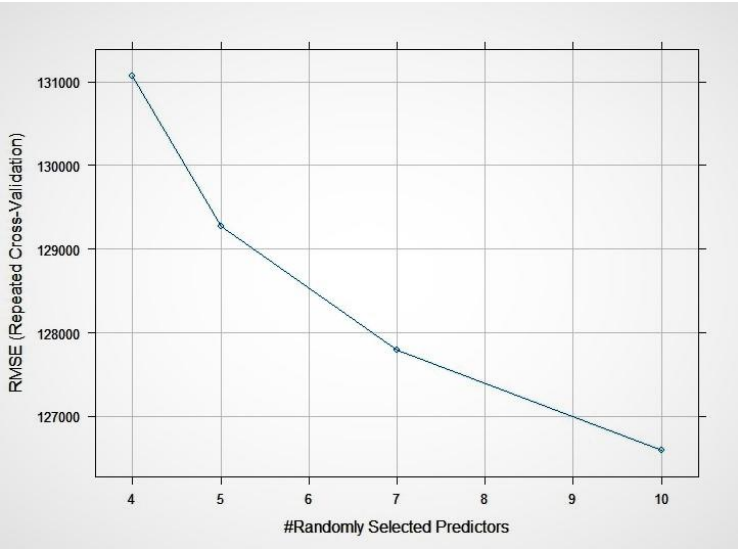
Πίνακας 3				
Πλήθος μεταβλητών που χρησιμοποιούνται σε κάθε κόμβο	4	5	7	10

Ο αλγόριθμος ο οποίος υλοποιεί την μέθοδο είναι ο εξής:

```
library(caret)
library(randomForest)
trainControl<-trainControl(method="repeatedcv",
                           number=10,repeats=1,
                           classProbs=F)

RFGrid<-expand.grid(mtry=c(4,5,7,10))

RFFit<-train(price~.,data = houses,
             method="rF",
             metric="RMSE",
             preProcess=c("center","scale"),
             trControl=trainControl,
             tuneGrid=RFGrid,
             verbose=FALSE)
RFFit
```

```
plot(RFFit)

RFpred<-predict(RFFit,houses[,-1])
DMwR::regr.eval(houses[,1],RFpred)
```

Ο παραπάνω αλγόριθμος εμφανίζει το παρακάτω σχήμα:



Σχήμα 2.3.2 Εφαρμογή της μεθόδου Random Forests

Στο παραπάνω σχήμα βλέπουμε πως μεταβάλλεται η ακρίβεια της μεθόδου σε σχέση με την τιμή της παραμέτρου m που έχει επιλεγεί. Παρατηρούμε ότι για μεγαλύτερες τιμές την παραμέτρου η μέθοδος επιτυγχάνει μεγαλύτερη ακρίβεια.

Η μικρότερη τιμή του μέσου τετραγωνικού σφάλματος που επιτυγχάνεται με την μέθοδο *Random Forests* είναι $MSE = 2{,}845 \cdot 10^9$.

ΚΕΦΑΛΑΙΟ 3

Εισαγωγή στις τεχνικές παλινδρόμησης με ποινή

3.1 Επιλογή μεταβλητών στο μοντέλο της γραμμικής παλινδρόμησης

Μία από τις πιο διαδεδομένες στατιστικές τεχνικές ανάλυσης δεδομένων είναι η γραμμική παλινδρόμηση. Θεωρούμε το σύνολο δεδομένων της μορφής $\{(y_i, x_{i1}, \dots, x_{ip}), i = 1, \dots, n\}$, όπου $y_i, i = 1, \dots, n$ οι τιμές της μεταβλητής απόκρισης Y, και $x_{ik}, i = 1, \dots, n, k = 1, \dots, p$ οι αντίστοιχες τιμές των επεξηγηματικών μεταβλητών X_k και το μοντέλο γραμμικής παλινδρόμησης:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i , \qquad i = 1, \dots, n$$

Στο πρότυπο αυτό τα $\beta_k, k = 0, \dots, p$ είναι οι άγνωστες παράμετροι του μοντέλου και τα ε_i , $i = 1, \dots, n$ είναι τα τυχαία σφάλματα του μοντέλου, για τα οποία υποθέτουμε ότι ακολουθούν κανονική κατανομή με μέση τιμή 0 και διασπορά σ^2 και ότι είναι μεταξύ τους ασυσχέτιστα.

Με χρήση πινάκων το γραμμικό μοντέλο γράφεται στη μορφή:

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

όπου $\boldsymbol{X}$ είναι ο πίνακας σχεδιασμού:

$$\boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \cdots & x_{1p} \\ 1 & x_{21} & \cdots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{np} \end{bmatrix}$$

και Y, $\boldsymbol{\beta}$, $\boldsymbol{\varepsilon}$ τα διανύσματα:

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}, \boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Οι εκτιμήσεις των παραμέτρων β_k υπολογίζονται με την μέθοδο ελαχίστων τετραγώνων και δίνονται από τον τύπο:

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

όπου $\widehat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \dots, \hat{\beta}_p)'$ το διάνυσμα το οποίο περιέχει τις εκτιμήσεις των παραμέτρων.

Στην περίπτωση όπου τα διανύσματα των μεταβλητών X_k είναι ορθογώνια, ο πίνακας $\boldsymbol{X}'\boldsymbol{X}$ είναι ίσος με τον ταυτοτικό πίνακα και οι εκτιμήσεις των παραμέτρων θα υπολογίζονται από την σχέση:

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

Στο μοντέλο της γραμμικής παλινδρόμησης ένα από τα προβλήματα που καλούμαστε να αντιμετωπίσουμε είναι αυτό της επιλογής των μεταβλητών που θα συμμετέχουν στο μοντέλο. Όταν οι επεξηγηματικές μεταβλητές δεν είναι πολλές κατασκευάζουμε όλα τα δυνατά μοντέλα και στη συνέχεια επιλέγουμε το μοντέλο που ικανοποιεί κάποιο κριτήριο που έχει τεθεί, όπως για παράδειγμα:

- ο συντελεστής προσδιορισμού R^2,
- το μέσο τετραγωνικό σφάλμα *MSE*,
- ο συντελεστής προσδιορισμού C_p του Mallows.

Όταν όμως το πλήθος των μεταβλητών είναι σχετικά μεγάλο είναι δύσκολο υπολογιστικά να κατασκευαστούν όλα τα δυνατά μοντέλα. Για την επιλογή των μεταβλητών που θα συμμετέχουν στο μοντέλο, οι συνήθεις τεχνικές που προτείνονται είναι η προς τα εμπρός επιλογή των μεταβλητών (Forward Selection), η προς τα πίσω απαλοιφή (Backward Elimination) και η κατά βήματα παλινδρόμηση (Stepwise Regression).

Στην τεχνική της προς τα εμπρός επιλογής οι μεταβλητές που συμμετέχουν στο μοντέλο επιλέγονται διαδοχικά ξεκινώντας με την επιλογή της πρώτης μεταβλητής. Προσαρμόζονται δηλαδή όλα τα απλά γραμμικά μοντέλα της μορφής:

$$Y_i = \beta_0 + \beta_k x_{ik} + \varepsilon_i \,, \qquad i = 1, \dots, n$$

για $k = 1, \dots, p$ και υπολογίζουμε τη στατιστική συνάρτηση t:

$$\frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \,, \qquad k = 1, \dots, p$$

όπου $\hat{\beta}_k$ είναι οι εκτιμητές ελαχίστων τετραγώνων.

Στην συνέχεια επιλέγουμε ως πρώτη μεταβλητή που συμμετέχει στο μοντέλο την μεταβλητή X_{k_1} για την οποία ισχύει:

$$\left|\frac{\hat{\beta}_{k_1}}{s(\hat{\beta}_{k_1})}\right| = \max\left\{\left|\frac{\hat{\beta}_k}{s(\hat{\beta}_k)}\right| \,, k = 1, \dots, p\right\}$$

και

$$\left|\frac{\hat{\beta}_{k_1}}{s(\hat{\beta}_{k_1})}\right| > t_{n-2}(a/2)$$

όπου $0 < \alpha < 1$ ένα επιθυμητό επίπεδο σημαντικότητας. Για την επιλογής της δεύτερης μεταβλητής προσαρμόζουμε όλα τα γραμμικά μοντέλα με δύο μεταβλητές της μορφής:

$$y_i = \beta_0 + \beta_{k_1} x_{ik_1} + \beta_k x_{ik} + \varepsilon_i\,, \qquad i = 1, \dots, n$$

για κάθε $k = 1, \dots, p$ με $k \neq k_1$ και επιλέγουμε την δεύτερη μεταβλητή X_{k_2}για την οποία ισχύει:

$$\left|\frac{\hat{\beta}_{k_2}}{s(\hat{\beta}_{k_2})}\right| = \max\left\{\left|\frac{\hat{\beta}_k}{s(\hat{\beta}_k)}\right|, k = 1, \dots, p\ \mu\varepsilon\ k \neq k_1\right\}$$

και

$$\left|\frac{\hat{\beta}_{k_2}}{s(\hat{\beta}_{k_2})}\right| > t_{n-3}(a/2).$$

Επαναλαμβάνουμε την ίδια διαδικασία για την επιλογή των υπόλοιπων μεταβλητών του μοντέλου, και σταματάμε όταν πλέον δεν ικανοποιείται το παραπάνω κριτήριο οπότε και επιλέγουμε το αμέσως προηγούμενο μοντέλο.

Στην τεχνική της προς τα πίσω απαλοιφής ξεκινάμε από το πλήρες μοντέλο:

$$Y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i\,, \qquad i = 1, \dots, n$$

και επιλέγουμε την μεταβλητή X_{k_1} που θα αφαιρέσουμε από το μοντέλο για την οποία ισχύει:

$$\left|\frac{\hat{\beta}_{k_1}}{s(\hat{\beta}_{k_1})}\right| = min\left\{\left|\frac{\hat{\beta}_k}{s(\hat{\beta}_k)}\right|, k = 1, \dots, p\right\}$$

και

$$\left|\frac{\hat{\beta}_{k_1}}{s(\hat{\beta}_{k_1})}\right| \leq t_{n-p+1}(a/2).$$

Στην περίπτωση που για καμία μεταβλητή δεν ικανοποιούνται τα παραπάνω κριτήρια τότε το βέλτιστο μοντέλο είναι το πλήρες μοντέλο. Σε αντίθετη περίπτωση αφαιρούμε την μεταβλητή που πληροί τα παραπάνω κριτήρια.

Στην συνέχεια προσαρμόζουμε τα γραμμικά μοντέλα:

$$Y_i = \beta_0 + \sum_{\substack{1 \le k \le p \\ k \ne k_1}} \beta_k x_{ik} + \varepsilon_i \,, \qquad i = 1, \dots, n$$

και επιλέγουμε την μεταβλητή X_{k_2} που θα αφαιρέσουμε από το μοντέλο για την οποία ισχύει:

$$\left| \frac{\hat{\beta}_{k_2}}{s(\hat{\beta}_{k_2})} \right| = min \left\{ \left| \frac{\hat{\beta}_k}{s(\hat{\beta}_k)} \right| , k = 1, \dots, p \; \mu\varepsilon \; k \ne k_1 \right\}$$

και

$$\left| \frac{\hat{\beta}_{k_2}}{s(\hat{\beta}_{k_2})} \right| \le t_{n-p}(a/2).$$

Επαναλαμβάνουμε την ίδια διαδικασία για την επιλογή των υπόλοιπων μεταβλητών που θα αφαιρεθούν από το μοντέλο, μέχρι να μην ικανοποιείται το παραπάνω κριτήριο.

Στην τεχνική της κατά βήματα παλινδρόμησης γίνεται σταδιακά η επιλογή των μεταβλητών και κάθε φορά που εισάγεται μία μεταβλητή στο μοντέλο, εξετάζεται επίσης αν θα μπορούσε να αφαιρεθεί κάποια από τις ήδη υπάρχουσες. Η διαδικασία επιλογής τερματίζεται όταν δεν μπορεί να εισαχθεί καμία μεταβλητή στο μοντέλο, ούτε και να εξαχθεί από αυτό.

Αρχικά ορίζουμε δύο τιμές F_{in} και F_{out} με βάση τις οποίες θα εισάγουμε και θα εξάγουμε μεταβλητές στο μοντέλο. Στη συνέχεια υπολογίζουμε τις στατιστικές συναρτήσεις:

$$F_k^* = \frac{SSR(X_k)}{MSE(X_k)}, \qquad k = 1, \dots, p$$

και επιλέγουμε για να εισάγουμε στο μοντέλο την μεταβλητή X_{k_1}, για την οποία ισχύει:

$$F_{k_1}^* = max \; F_k^*, \qquad k = 1, \dots, p$$

και

$$F_{k_1}^* > F_{in} \,.$$

Όταν σε κάποιο βήμα έχουμε επιλέξει τις $m-1$ μεταβλητές $X_{k_1}, X_{k_1}, \dots, X_{k_{m-1}}$, υπολογίζουμε τους λόγους:

$$F_k^* = \frac{SSR(X_k | X_{k_1}, X_{k_1}, \dots, X_{k_{m-1}})}{MSE(X_k, X_{k_1}, X_{k_1}, \dots, X_{k_{m-1}})} =$$

$$\frac{SSR\left(X_k, X_{k_1}, X_{k_1}, \dots, X_{k_{m-1}}\right) - SSR(X_{k_1}, X_{k_1}, \dots, X_{k_{m-1}})}{MSE(X_k, X_{k_1}, X_{k_1}, \dots, X_{k_{m-1}})}$$

για $k \neq k_1, \dots, k_{m-1}$ και επιλέγουμε για να εισάγουμε στο μοντέλο την μεταβλητή X_{k_m}, για την οποία ισχύει:

$$F^*_{k_m} = max\, F^*_k, \qquad k = 1, \dots, p \; \kappa\alpha\iota \; k \neq k_1, \dots, k_{m-1}$$

και

$$F^*_{k_m} > F_{in}\,.$$

Αντίστοιχα όταν σε κάποιο βήμα έχουμε επιλέξει τις $m-1$ μεταβλητές $X_{k_1}, X_{k_1}, \dots, X_{k_{m-1}}$, και η τελευταία ανεξάρτητη μεταβλητή που έχουμε εισάγει στο μοντέλο είναι η X_{k_1}, τότε οι υπόλοιπες μεταβλητές του μοντέλου γίνονται υποψήφιες για εξαγωγή. Επιλέγουμε να εξάγουμε από το μοντέλο την μεταβλητή X_{k_m} για την οποία ισχύει:

$$F^*_{k_m} = min\, F^*_k, \qquad k = k_2, \dots, k_{m-1}$$

και

$$F^*_{k_m} < F_{out}\,.$$

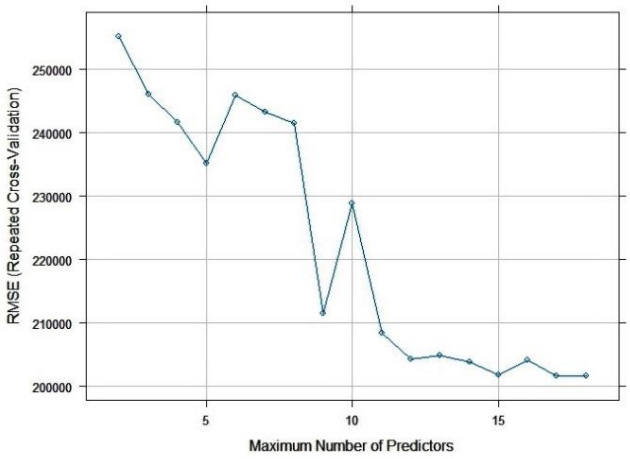
Εφαρμόζοντας τις παραπάνω τεχνικές στο ίδιο σύνολο δεδομένων, δεν είναι απαραίτητο οι μεταβλητές που θα επιλεγούν να είναι οι ίδιες, για κάθε τεχνική που εφαρμόζεται. Επιπλέον όσο μεγαλύτερο είναι το πλήθος των μεταβλητών τόσο μεγαλύτερο χρόνο χρειαζόμαστε για την επιλογή του τελικού συνόλου μεταβλητών που θα συμμετέχουν στο μοντέλο, καθώς σε κάθε βήμα εισάγουμε ή εξάγουμε μία μεταβλητή.

Για να μελετήσουμε την ακρίβεια της μεθόδου και να την συγκρίνουμε με άλλες τεχνικές χρησιμοποιούμε το σύνολο δεδομένων για την πρόβλεψη της τιμής των κατοικιών με βάση τα διάφορα χαρακτηριστικά τους. Η υλοποίηση της μεθόδου πραγματοποιήθηκε σε περιβάλλον *R* με την χρήση του πακέτου *'MASS'* και του πακέτου *'caret'*. Χρησιμοποιούμε την τεχνική της κατά βήματα παλινδρόμησης για την επιλογή των μεταβλητών, καθώς με αυτή την τεχνική ελέγχεται σε κάθε βήμα αν θα προστεθεί ή θα αφαιρεθεί κάποια μεταβλητή από το μοντέλο.

Ο αλγόριθμος που υλοποιεί την μέθοδο είναι ο εξής:

```
library(caret)
library(MASS)
trainControl<-trainControl(method="repeatedcv",
                           number=10,repeats=1,
                           classProbs=F)

LinearGrid<-expand.grid(nvmax=c(2:18))

LinearFit<-train(price~.,data = houses,
                 method="leapSeq",
                 metric="RMSE",
                 preProcess=c("center","scale"),
                 trControl=trainControl,
                 tuneGrid=LinearGrid)

LinearFit
plot(LinearFit)
Linearpred<-predict(LinearFit,houses[,-1])
DMwR::regr.eval(houses[,1],Linearpred)
```

Ο παραπάνω αλγόριθμος εμφανίζει το παρακάτω σχήμα:



Σχήμα 3.1.1 Εφαρμογή της κατά βήματα παλινδρόμησης

Στο παραπάνω σχήμα παρουσιάζεται η ακρίβεια που επιτυγχάνει η μέθοδος της γραμμικής παλινδρόμησης με την τεχνική της κατά βήματα επιλογής των μεταβλητών. Η μεγαλύτερη ακρίβεια επιτυγχάνεται με την συμμετοχή 15 μεταβλητών στο μοντέλο.

Η μικρότερη τιμή του μέσου τετραγωνικού σφάλματος που επιτυγχάνεται με την μέθοδο της κατά βήματα παλινδρόμησης είναι $MSE = 4{,}082 \cdot 10^{10}$.

3.2 Η τεχνική non-negative garrote estimator

Στο μοντέλο της γραμμικής παλινδρόμησης όταν το πλήθος των μεταβλητών είναι πολύ μεγάλο παρουσιάζεται το πρόβλημα της επιλογής εκείνων των μεταβλητών που θα συμμετέχουν στο μοντέλο, καθώς υπάρχουν πάρα πολλοί δυνατοί συνδυασμοί μεταβλητών που μπορούν να χρησιμοποιηθούν. Ακόμη οι εκτιμήτριες ελαχίστων τετραγώνων ενώ είναι αμερόληπτες, υστερούν σε ακρίβεια. Δηλαδή η πρόβλεψη που παράγει ένα μοντέλο, στο οποίο έχουν χρησιμοποιηθεί οι εκτιμήτριες ελαχίστων τετραγώνων μπορεί να έχει μεγάλη απόκλιση από την πραγματική τιμή. Η ακρίβεια πρόβλεψης μπορεί κάποιες φορές να βελτιωθεί συρρικνώνοντας τους συντελεστές της παλινδρόμησης ή θέτοντάς τους ίσους με το μηδέν. Με αυτό τον τρόπο αποδεχόμαστε κάποια μεροληψία προκειμένου να μειώσουμε την διασπορά των προβλεπόμενων τιμών και άρα να βελτιωθεί η συνολική ακρίβεια του μοντέλου.

Στην ίδια προσέγγιση του προβλήματος βασίζεται και η επιλογή των μεταβλητών στο μοντέλο της γραμμικής παλινδρόμησης με τις τεχνικές που περιγράψαμε παραπάνω. Με τις παραπάνω τεχνικές, ωστόσο, κάποιες μεταβλητές είτε χρησιμοποιούνται είτε αποκλείονται από το μοντέλο. Με αυτόν τον τρόπο μικρές μεταβολές στα δεδομένα μπορεί να οδηγήσουν σε

διαφορετική επιλογή των μεταβλητών που συμμετέχουν στο μοντέλο και αυτό μπορεί να οδηγήσει σε μικρή ακρίβεια πρόβλεψης.

Στη συνέχεια θα παρουσιάσουμε μερικές τεχνικές παλινδρόμησης, στις οποίες εφαρμόζεται κάποια ποινή στη εκτίμηση των συντελεστών.

Η τεχνική *non-negative (nn) garrote estimator (Breiman, 1995)* είναι μία τεχνική με ποινή που χρησιμοποιείται με σκοπό να επιτευχθεί μεγαλύτερη ακρίβεια.

Θεωρούμε το πλήρες μοντέλο γραμμικής παλινδρόμησης:

$$y_i = \beta_0 + \sum_{k=1}^{p} \beta_k x_{ik} + \varepsilon_i \,, \qquad i = 1, \dots, n.$$

Οι εκτιμήτριες *non-negative garrote* δίνονται από τον τύπο:

$$\hat{\beta}_{k,G} = c_k \hat{\beta}_k, \qquad k = 0,1, \dots, p$$

όπου $\hat{\beta}_k$ είναι οι εκτιμήτριες ελαχίστων τετραγώνων και c_k είναι μία παράμετρος συρρίκνωσης η οποία επιλέγεται ελαχιστοποιώντας την παράσταση:

$$\sum_{i=1}^{n} \hat{\varepsilon}_i{}^2 = \sum_{i=1}^{n} (y_i - \sum_{k=0}^{p} c_k \hat{\beta}_k x_{ik})^2$$

υπό τους περιορισμούς $c_k \geq 0$ και $\sum_{k=0}^{p} c_k \leq s$.

Παρατηρούμε ότι όσο μικρότερος είναι ο αριθμός s τόσο μικρότερες τιμές τείνουν να πάρουν οι συντελεστές c_k. Δηλαδή με την τεχνική των εκτιμητριών *non-negative garrote* καταλήγουμε σε ένα μοντέλο το οποίο έχει μεγαλύτερη ακρίβεια από το μοντέλο της συνήθους παλινδρόμησης, όπου η επιλογή των μεταβλητών έχει πραγματοποιηθεί με τις συνήθεις τεχνικές, καθώς συμμετέχουν όλες οι μεταβλητές στο μοντέλο και άρα δεν μειώνεται η διαθέσιμη πληροφορία. Σύμφωνα με πειραματισμούς του *Breiman* τόσο σε πραγματικά όσο και σε προσομοιωμένα δεδομένα η μέθοδος παράγει μικρότερα σφάλματα πρόβλεψης.

Στην περίπτωση που τα διανύσματα των μεταβλητών X_k είναι ορθογώνια και ο πίνακας $\boldsymbol{X'X}$ είναι ίσος με τον ταυτοτικό πίνακα, η παράσταση $\sum_{i=1}^{n}(y_i - \sum_{k=0}^{p} c_k \hat{\beta}_k x_{ik})^2$ ελαχιστοποιείται υπό τους περιορισμούς $c_k \geq 0$ και $\sum_{k=0}^{p} c_k = s$. Και οι συντελεστές c_k δίνονται από τον τύπο:

$$c_k = \left(1 - \frac{\lambda^2}{\hat{\beta}_k{}^2}\right)^+$$

όπου το λ καθορίζεται από το s υπό τον περιορισμό $\sum_{k=0}^{p} c_k = s$ και α^+ είναι το θετικό μέρος του α. Συνεπώς οι εκτιμήτριες *non-negative garrote* δίνονται από τον τύπο:

$$\hat{\beta}_{k,G} = \left(1 - \frac{\lambda^2}{\hat{\beta}_k{}^2}\right)^+ \cdot \hat{\beta}_k$$

Η τεχνική των εκτιμητριών *non-negative garrote* έχει μικρότερο σφάλμα πρόβλεψης από τις συνήθεις τεχνικές επιλογής μεταβλητών και ανταγωνίζεται την τεχνική της παλινδρόμησης τύπου *ridge*, την οποία θα προσεγγίσουμε σε επόμενο κεφάλαιο. Ένα μειονέκτημα της τεχνικής αυτής είναι ότι επηρεάζεται από το μέγεθος και το πρόσημο των εκτιμητριών ελαχίστων τετραγώνων. Επίσης η τεχνική παρουσιάζει προβλήματα όταν υπάρχουν υψηλές συσχετίσεις μεταξύ των επεξηγηματικών μεταβλητών.

Χρησιμοποιούμε το σύνολο δεδομένων για την πρόβλεψη της τιμής των κατοικιών με βάση τα διάφορα χαρακτηριστικά τους, για να συγκρίνουμε την ακρίβεια της μεθόδου με τις υπόλοιπες τεχνικές. Για την εφαρμογή των τεχνικών με ποινή κανονικοποιούμε τα δεδομένα, καθώς οι επεξηγηματικές μεταβλητές με διαφορετική διασπορά θα έχουν διαφορετική συνεισφορά στον περιορισμό της ποινής. Η υλοποίηση της μεθόδου πραγματοποιήθηκε σε περιβάλλον *R* με την χρήση του πακέτου *'lqa'*, το οποίο δεν υποστηρίζεται από το πακέτο *'caret'*.

Ο αλγόριθμος που υλοποιεί την μέθοδο είναι ο εξής:

```
library(lqa)
nng<-cv.nng(y.train=houses[,1],
            x.train=houses[,-1],
            intercept=TRUE,
            family = gaussian(),
            lambda.nng = list(0.1,0.5,1,2,5,10),
            n.fold = 10,
            standardize=TRUE,
            penalty = NULL)
nng
```

Η μικρότερη τιμή του μέσου τετραγωνικού σφάλματος που επιτυγχάνεται με την μέθοδο *non-negative garrote estimator* είναι $MSE = 4{,}069 \cdot 10^{10}$.

ΚΕΦΑΛΑΙΟ 4

Παλινδρόμηση τύπου *Ridge*

4.1 Εισαγωγή στην παλινδρόμηση τύπου *Ridge*

Όταν σε ένα σύνολο δεδομένων παρουσιάζεται υψηλή συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών, εμφανίζεται το φαινόμενο της πολυσυγγραμμικότητας. Στο μοντέλο της γραμμικής παλινδρόμησης η ύπαρξη πολυσυγγραμμικότητας οδηγεί σε υψηλά τυπικά σφάλματα για τις εκτιμήτριες ελαχίστων τετραγώνων και γι' αυτό το λόγο είναι δύσκολο να βρεθούν οι στατιστικά σημαντικές μεταβλητές που θα συμμετέχουν στο μοντέλο.

Στις περιπτώσεις που έχουμε ένα μεγάλο σύνολο δεδομένων το οποίο αποτελείται από πολλές μεταβλητές, συχνά παρατηρούνται υψηλές συσχετίσεις. Μία μέθοδος που προτείνεται για την ανάλυση τέτοιων συνόλων δεδομένων είναι η παλινδρόμηση τύπου *Ridge (Hoerl & Kennard, 1970)*, με βάση την οποία συρρικνώνονται κάποιοι συντελεστές παλινδρόμησης.

Οι εκτιμήτριες της παλινδρόμησης *ridge*, $\hat{\beta}^{ridge}$, υπολογίζονται ελαχιστοποιώντας την ποσότητα:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik}\right)^2 + \lambda\sum_{k=1}^{p}{\beta_k}^2$$

όπου $\lambda \geq 0$ μία ρυθμιστική *(tuning)* παράμετρος. Ένας εναλλακτικός τρόπος έκφρασης του προβλήματος είναι να επιτευχθεί η ελαχιστοποίηση της παράστασης:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik}\right)^2$$

υπό τον περιορισμό

$$\sum_{k=1}^{p}{\beta_k}^2 \leq t$$

όπου το t εξαρτάται από το λ.

Όσο αυξάνεται η τιμή της παραμέτρου λ οι συντελεστές παλινδρόμησης συρρικνώνονται. Με αυτόν τον τρόπο μειώνεται και η διασπορά των συντελεστών και κατά συνέπεια επιτυγχάνεται μεγαλύτερη ακρίβεια στο μοντέλο.

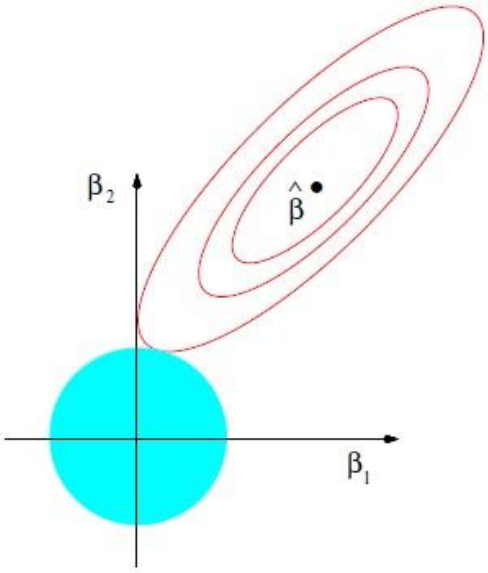
Πιο συγκεκριμένα αν:

$$\hat{\beta}^* = \frac{1}{1+\alpha}\hat{\beta}, \qquad \alpha > 0\,,$$

θα ισχύει ότι:

$$V[\hat{\beta}^*] = V\left[\frac{1}{1+\alpha}\hat{\beta}\right] = \left(\frac{1}{1+\alpha}\right)^2 V[\hat{\beta}] < V[\hat{\beta}].$$

Στο παρακάτω διάγραμμα παρουσιάζεται γεωμετρικά η επίδραση που έχει ο περιορισμός $\sum_{k=1}^{p} {\beta_k}^2 \leq t$ στην ελαχιστοποίηση της παράστασης $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik}\right)^2$, για $k = 2$:



Σχήμα 4.1.1 Περιορισμός *Ridge* σε 2 διαστάσεις.

Ο περιορισμός $\sum_{k=1}^{2} {\beta_k}^2 \leq t$ απεικονίζεται από τον κύκλο, ενώ οι ελλείψεις του σχήματος αντιπροσωπεύουν τα σημεία στα οποία η αντικειμενική συνάρτηση $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{k=1}^{2}\beta_k x_{ik})^2$ έχει την ίδια τιμή, η οποία γίνεται μικρότερη προς το κέντρο των ελλείψεων. Παρατηρούμε ότι η ελαχιστοποίηση της αντικειμενικής συνάρτησης επιτυγχάνεται για κάποιο σημείο της κυκλικής περιοχής που ορίζει ο περιορισμός $\sum_{k=1}^{2} {\beta_k}^2 \leq t$.

Ορίζουμε για ένα διάνυσμα, $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$:

$$\|\boldsymbol{x}\|_2 = \sqrt{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}.$$

Οι εκτιμήτριες της μεθόδου *Ridge* υπολογίζονται ως εξής:

$$\widehat{\boldsymbol{\beta}}^{ridge} = \min_{\boldsymbol{\beta}}\{\|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2{}^2 + \lambda\|\boldsymbol{\beta}\|_2\}$$
$$= \min_{\boldsymbol{\beta}}\{\boldsymbol{Y'Y} - 2\boldsymbol{Y'X\beta} + \boldsymbol{\beta'\beta} + \lambda\|\boldsymbol{\beta}\|_2\}$$
$$= \min_{\boldsymbol{\beta}}\{-2\boldsymbol{Y'X\beta} + \|\boldsymbol{\beta}\|_{\mathbf{2}}{}^{\mathbf{2}} + \lambda\|\boldsymbol{\beta}\|_2\}$$

Στην περίπτωση όπου τα διανύσματα των μεταβλητών X_k είναι ορθογώνια οι εκτιμήτριες ελαχίστων τετραγώνων υπολογίζονται από τη σχέση $\widehat{\boldsymbol{\beta}} = \boldsymbol{X'Y}$, οπότε η παραπάνω σχέση γράφεται ισοδύναμα:

$$\min_{\beta}\sum_{k=1}^{p} - 2\hat{\beta}_{\boldsymbol{k}}\beta_{\boldsymbol{k}} + \beta_{\boldsymbol{k}}{}^{\mathbf{2}} + \lambda\beta_k{}^2$$

Η παραπάνω σχέση αποτελείται από ένα άθροισμα όρων ο καθένας από τους οποίους αντιστοιχεί σε ξεχωριστή μεταβλητή $\beta_{\boldsymbol{k}}$ από τους υπόλοιπους. Κατά συνέπεια κάθε ένας από τους όρους μπορεί να ελαχιστοποιηθεί ξεχωριστά για να βρεθεί η τιμή της κάθε μεταβλητής.

Παραγωγίζοντας τις σχέσεις:

$$\mathcal{L}_k = -2\hat{\beta}_{\boldsymbol{k}}\beta_{\boldsymbol{k}} + \beta_{\boldsymbol{k}}{}^{\mathbf{2}} + \lambda\beta_k{}^2, \qquad k = 1, \dots, p$$

βρίσκουμε ότι:

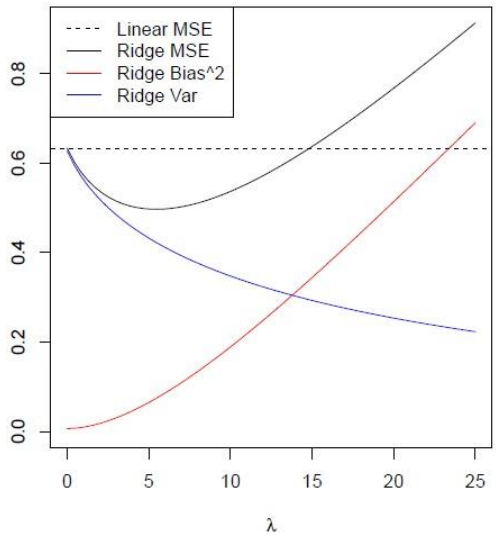
$$\hat{\beta}_k^{ridge} = \frac{1}{1+\lambda}\cdot\hat{\beta}_k$$

όπου $\hat{\beta}_k$ είναι οι εκτιμήτριες ελαχίστων τετραγώνων.

Οι εκτιμήτριες $\hat{\beta}^{ridge}$ είναι μεροληπτικές, ωστόσο σε προβλήματα όπου οι εκτιμήτριες ελαχίστων τετραγώνων δεν αποδίδουν καλά, έχουν δηλαδή μεγάλη διασπορά και κατ' επέκταση μεγάλα τετραγωνικά σφάλματα, επιλέγουμε ένα μοντέλο που εισάγει κάποια μεροληψία με σκοπό την μεγαλύτερη ακρίβεια.

4.2 Η παράμετρος λ στην παλινδρόμηση τύπου *Ridge*

Στο παρακάτω διάγραμμα παρουσιάζεται η μεταβολή του μέσου τετραγωνικού σφάλματος, της μεροληψίας και της διασποράς των εκτιμητριών με την μέθοδο *Ridge* σε σχέση με την παράμετρο λ:



Σχήμα 4.2.1 Μεταβολές με βάση την παράμετρο λ.

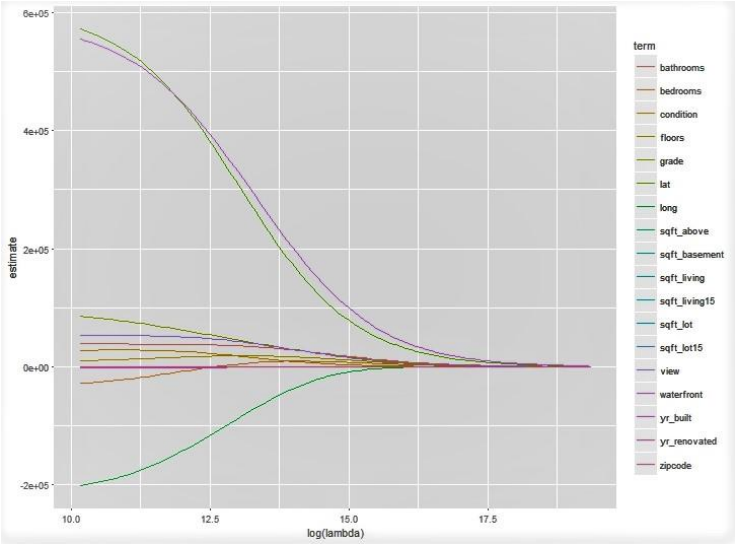
Συρρικνώνοντας δηλαδή τις τιμές των εκτιμητριών εισάγουμε κάποια μεροληψία στο μοντέλο με σκοπό να επιτύχουμε μεγαλύτερη ακρίβεια. Η μεροληψία του μοντέλου εκφράζεται από την τιμή της εκτιμήτριας $\hat{\beta}_0$ του σταθερού όρου, η οποία δεν επηρεάζεται από τον περιορισμό *Ridge*. Η τιμή της εκτιμήτριας $\hat{\beta}_0$ αυξάνεται όσο συρρικνώνονται οι τιμές των άλλων εκτιμητριών *Ridge*. Ακόμη από το παραπάνω διάγραμμα βλέπουμε πως μεταβάλλεται η τιμή του μέσου τετραγωνικού σφάλματος με την μεταβολή της παραμέτρου λ, ενώ δίνεται και ως μέτρο σύγκρισης η τιμή του μέσου τετραγωνικού σφάλματος, του πλήρους μοντέλου της γραμμικής παλινδρόμησης. Επιπροσθέτως παρουσιάζεται η μείωση της διασποράς των συντελεστών με την αύξηση της τιμής της παραμέτρου λ.

Για να εξετάσουμε την επίδραση που έχει η παράμετρος λ στο μοντέλο που παράγεται από την μέθοδο *Ridge* χρησιμοποιούμε το αρχείο δεδομένων που αφορά την πρόβλεψη της τιμής των κατοικιών με βάση τα διάφορα χαρακτηριστικά τους. Η υλοποίηση της μεθόδου πραγματοποιήθηκε σε περιβάλλον *R* με την χρήση του πακέτου *'glmnet'* και του πακέτου *'caret'*.

Με τον παρακάτω αλγόριθμο κατασκευάζουμε ένα διάγραμμα, στο οποίο παρουσιάζεται η μεταβολή της τιμής των συντελεστών $\hat{\beta}_k^{ridge}$, σε σχέση με την τιμή της παραμέτρου λ:

```
library(glmnet)
library(ggplot2)
library(broom)
ridge.fit<-glmnet(houses[,-1], houses[,1],
                  alpha=0,
                  standardize = TRUE)
ridge.td<-tidy(ridge.fit)
ggplot(ridge.td, aes(log(lambda), estimate, color = term)) + geom_line(size=0.8)
```

Ο παραπάνω αλγόριθμος εμφανίζει το παρακάτω διάγραμμα:



Σχήμα 4.2.2 Μεταβολή των εκτιμητριών *Ridge* σε σχέση με την παράμετρο λ.

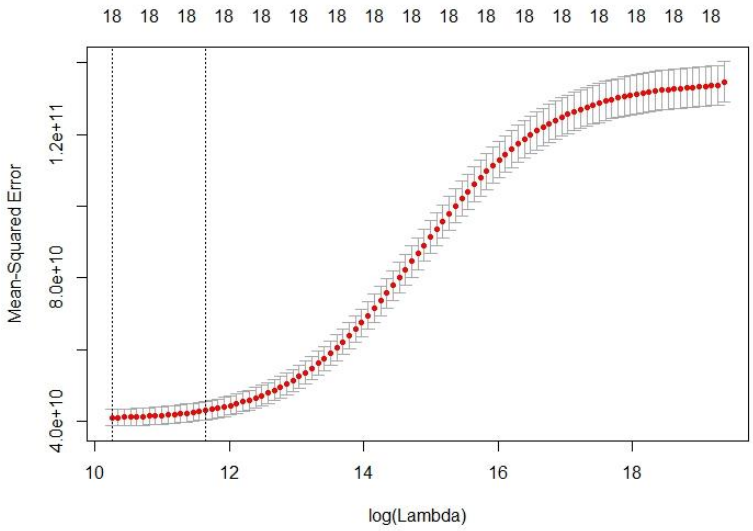
Παρατηρούμε ότι οι τιμές των εκτιμητριών Ridge συρρικνώνονται σταδιακά με την αύξηση της παραμέτρου λ και κάποιες τείνουν να γίνουν ίσες με το μηδέν.

Στόχος μας είναι να επιλέξουμε την τιμή της παραμέτρου λ για την οποία επιτυγχάνεται η μέγιστη δυνατή ακρίβεια. Για την επιλογή της βέλτιστης τιμής της παραμέτρου λ μπορούμε να χρησιμοποιήσουμε την τεχνική *Cross-Validation*, με βάση την οποία χωρίζουμε το σύνολο των δεδομένων σε δύο τμήματα, όπου το ένα τμήμα χρησιμοποιείται για τον υπολογισμό των εκτιμητριών $\hat{\beta}^{ridge}$ και το άλλο για την επιλογή της τιμής της παραμέτρου λ για την οποία επιτυγχάνεται η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος. Επειδή όμως με την τεχνική *Cross-Validation* ένα τμήμα των δεδομένων δεν χρησιμοποιείται καθόλου για τον υπολογισμό των $\hat{\beta}_k^{ridge}$ και άρα μειώνεται η διαθέσιμη πληροφορία που χρησιμοποιείται, επιλέγουμε την τεχνική *K-fold Cross-Validation* κατά την οποία χρησιμοποιείται όλο το σύνολο των δεδομένων, με βάση την οποία χωρίζεται το σύνολο των δεδομένων σε K τμήματα και χρησιμοποιούνται τα K-1 τμήματα για τον υπολογισμό των εκτιμητριών και το τελευταίο τμήμα χρησιμοποιείται για την επιλογή της παραμέτρου λ. Η διαδικασία αυτή επαναλαμβάνεται K φορές, και κάθε φορά χρησιμοποιείται διαφορετικό τμήμα των δεδομένων για την επιλογή της παραμέτρου λ. Η συνήθης τιμή που επιλέγεται για το K είναι 5 ή 10. Τελικά η επιλογή της παραμέτρου λ γίνεται υπολογίζοντας τον μέσο όρο των τιμών της παραμέτρου που έχουν προκύψει σε κάθε επανάληψη.

Για τον υπολογισμό της βέλτιστης τιμής της παραμέτρου λ, χρησιμοποιούμε το πακέτο *'glmnet'* εφαρμόζοντας τον παρακάτω αλγόριθμο:

```
library(glmnet)
cv.ridge<-cv.glmnet(houses[,-1], houses[,1],
alpha=0,
standardize=TRUE)
plot(cv.ridge)
cv.ridge$lambda.min
```

Ο παραπάνω αλγόριθμος εμφανίζει το παρακάτω διάγραμμα, στο οποίο παρουσιάζεται η μεταβολή του μέσου τετραγωνικού σφάλματος σε σχέση με την τιμή της παραμέτρου λ:



Σχήμα 4.2.3 Μεταβολή του *MSE* σε σχέση με την παράμετρο λ.

Στο παραπάνω σχήμα παρουσιάζεται η μεταβολή του μέσου τετραγωνικού σφάλματος σε σχέση με τις τιμές της παραμέτρου λ. Παρατηρούμε ότι το μέσο τετραγωνικό σφάλμα δεν μεταβάλλεται για κάποιες τιμές παραμέτρου λ και πάνω από κάποια τιμή της παραμέτρου το μέσο τετραγωνικό σφάλμα αυξάνεται.

Ένα άλλο κριτήριο που μπορούμε να χρησιμοποιήσουμε είναι το *GCV (Generalized Cross-validation)*, με βάση το οποίο επιλέγουμε την τιμή της παραμέτρου λ για την οποία ελαχιστοποιείται η παράσταση:

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

όπου $\hat{y}_i = \hat{f}(x_i)$ είναι η εκτίμηση για την τιμή y_i της μεταβλητής απόκρισης Y με την μέθοδο *Ridge* και $tr(\boldsymbol{S})$ είναι το ίχνος του πίνακα $\boldsymbol{S}$ που ικανοποιεί την σχέση $\hat{\boldsymbol{Y}} = \boldsymbol{SY}$. Ο πίνακας $\boldsymbol{S}$ στο μοντέλο της γραμμικής παλινδρόμησης είναι ο *hat matrix* $\boldsymbol{H}$, ο οποίος αν πολλαπλασιαστεί με το διάνυσμα της μεταβλητής απόκρισης δίνει το διάνυσμα της εκτίμησης της μεταβλητής

απόκρισης. Ο πίνακας $\boldsymbol{H}$ δίνεται από τη σχέση $\boldsymbol{H} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. Αντίστοιχα ο πίνακας $\boldsymbol{S}$ στην μέθοδο *Ridge* υπολογίζεται από την παρακάτω σχέση:

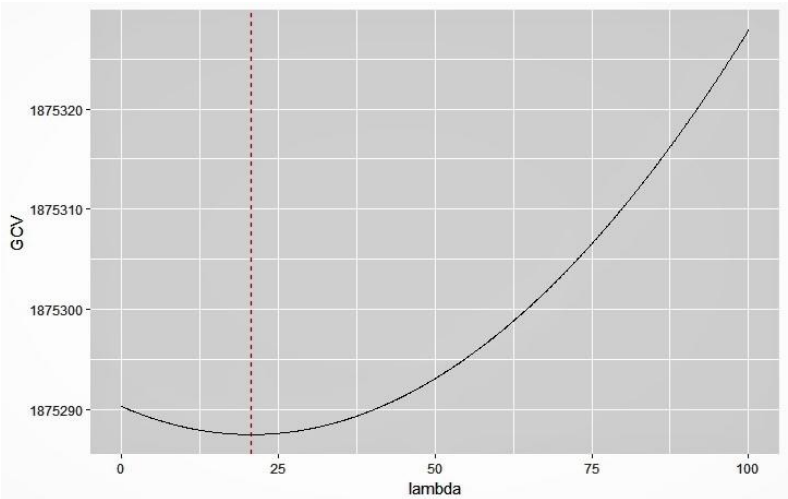
$$\boldsymbol{S} = \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X} + \lambda\boldsymbol{I})^{-1}\boldsymbol{X}'$$

Το παραπάνω κριτήριο είναι μία παραλλαγή του μέσου τετραγωνικού σφάλματος, το οποίο συνυπολογίζει το πλήθος των παραμέτρων που εκτιμώνται στο μοντέλο, το οποίο ισούται με το $tr(\boldsymbol{S})$, έτσι ώστε να επιλέγεται ένα απλούστερο μοντέλο. Με βάση την παραπάνω συνάρτηση, για δύο μοντέλα που έχουν το ίδιο άθροισμα τετραγωνικών αποκλίσεων, η τιμή του GCV θα είναι μικρότερη για το μοντέλο με τη μικρότερη πολυπλοκότητα.

Ο υπολογισμός του *GCV* πραγματοποιείται με τη χρήση του πακέτου *'MASS'* από τον παρακάτω αλγόριθμο:

```
library(MASS)
library(broom)
library(ggplot2)
lmridge.fit <- lm.ridge(price ~ ., houses, lambda = seq(0, 100, .01))
td <- tidy(lmridge.fit)
g<- glance(lmridge.fit)

ggplot(td, aes(lambda, GCV))
       + geom_line()
       + geom_vline(xintercept=g$lambdaGCV, col = "red", lty = 2)
```

Ο παραπάνω αλγόριθμος εμφανίζει το παρακάτω διάγραμμα:



Σχήμα 4.2.4 Μεταβολή του *GCV* σε σχέση με την παράμετρο *λ*.

Για να μελετήσουμε την ακρίβεια της μεθόδου και να την συγκρίνουμε με άλλες τεχνικές πραγματοποιούμε την υλοποίηση της μεθόδου σε περιβάλλον R με την χρήση του πακέτου *'elasticnet'* και του πακέτου *'caret'*.

Ο αλγόριθμος που υλοποιεί την μέθοδο είναι ο εξής:

```
library(caret)
library(elasticnet)
trainControl<-trainControl(method="repeatedcv",
                           number=10,repeats=1,
                           classProbs=F)

RidgeGrid<-expand.grid(lambda=(seq(0,100,0.1)))

RidgeFit<-train(price~.,data = houses,
                method="ridge",
                metric="RMSE",
                preProcess=c("center","scale"),
                trControl=trainControl)

Ridgepred<-predict(RidgeFit,houses[,-1])

DMwR::regr.eval(houses[,1],Ridgepred)
RidgeFit
plot(RidgeFit)
```

Η μικρότερη τιμή του μέσου τετραγωνικού σφάλματος που επιτυγχάνεται με τη μέθοδο *Ridge* είναι $MSE = 4{,}070 \cdot 10^{10}$.

ΚΕΦΑΛΑΙΟ 5

Μέθοδος *LASSO*

5.1 Εισαγωγή στη μέθοδο *LASSO*

Ένα πρόβλημα που παρουσιάζεται με την χρήση της μεθόδου *Ridge*, είναι ότι δεν γίνεται επιλογή των μεταβλητών που συμμετέχουν στο μοντέλο. Στην περίπτωση που το πλήθος των επεξηγηματικών μεταβλητών είναι μεγάλο τότε το μοντέλο που προκύπτει δεν είναι εύκολα ερμηνεύσιμο. Στόχος μας είναι να επιλέξουμε ένα μοντέλο το οποίο χρησιμοποιεί τις μεταβλητές εκείνες που έχουν την πιο σημαντική επίδραση στην εξαρτημένη μεταβλητή.

Για την αντιμετώπιση του παραπάνω προβλήματος χρησιμοποιούμε την μέθοδο *LASSO (Least Absolute Shrinkage and Selection Operator – Tibshirani, 1996)* με βάση την οποία συρρικνώνονται κάποιοι συντελεστές παλινδρόμησης ενώ κάποιοι άλλοι μηδενίζονται.

Οι εκτιμήτριες της παλινδρόμησης *LASSO*, $\hat{\beta}^{LASSO}$, υπολογίζονται ελαχιστοποιώντας την σχέση:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik}\right)^2 + \lambda\sum_{k=1}^{p}|\beta_k|$$

όπου $\lambda \geq 0$ μία ρυθμιστική παράμετρος.

Ορίζουμε για ένα διάνυσμα, $\boldsymbol{x} = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}$:

$$\|\boldsymbol{x}\|_1 = \sum_{i=1}^{n}|x_i|.$$

Ισοδύναμα με την μορφή πινάκων το διάνυσμα των εκτιμητριών $\hat{\beta}^{LASSO}$, υπολογίζονται από την σχέση:

$$\widehat{\boldsymbol{\beta}}^{LASSO} = \min_{\boldsymbol{\beta}}\left\{\|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2{}^2 + \lambda\|\boldsymbol{\beta}\|_1\right\}$$

Ένας εναλλακτικός τρόπος έκφρασης του προβλήματος είναι να επιδιωχθεί η ελαχιστοποίηση της παράστασης:

$$\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik}\right)^2$$

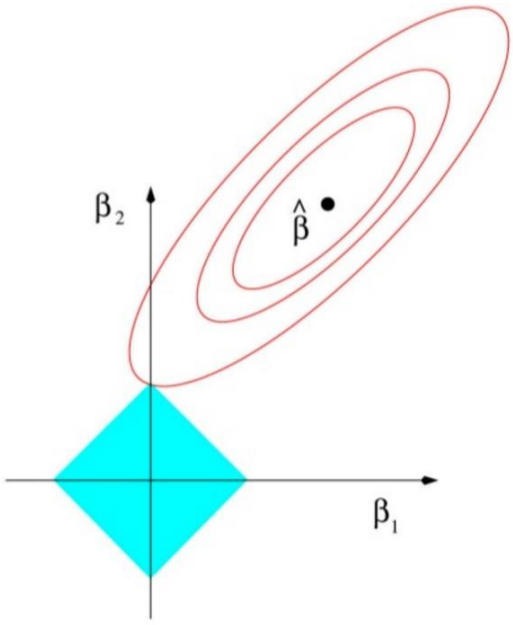
υπό τον περιορισμό:

$$\sum_{k=1}^{p}|\beta_k| \leq t$$

όπου το t εξαρτάται από το λ.

Όπως και στην παλινδρόμηση τύπου Ridge έτσι και στη μέθοδο LASSO με την αύξηση της τιμής της παραμέτρου λ οι συντελεστές παλινδρόμησης συρρικνώνονται. Το πλεονέκτημα όμως της μεθόδου αυτής σε σχέση με την μέθοδο Ridge είναι ότι κάποιοι συντελεστές παλινδρόμησης μπορούν να γίνουν ακριβώς ίσοι με το μηδέν. Αυτό επιτυγχάνεται λόγω του περιορισμού $\sum_{k=1}^{p}|\beta_k| \leq t$ στην ελαχιστοποίηση της παράστασης. Γεωμετρικά ο περιορισμός $\sum_{k=1}^{p}|\beta_k| \leq t$ παριστάνει ένα πολύτοπο *(polytope)* με πολλές κορυφές. Οι κορυφές αυτές είναι τα σημεία όπου κάποιοι συντελεστές παίρνουν την τιμή 0. Άμεσο επακόλουθο των παραπάνω είναι ότι η ελαχιστοποίηση της παράστασης $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik}\right)^2$ υπό τον περιορισμό $\sum_{k=1}^{p}|\beta_k| \leq t$, για κάποια τιμή του t, μπορεί να επιτυγχάνεται σε μία κορυφή του πολύτοπου, συνεπώς κάποιοι συντελεστές είναι ίσοι με το 0.

Η επίδραση που έχει ο περιορισμός $\sum_{k=1}^{p}|\beta_k| \leq t$ στην ελαχιστοποίηση της παράστασης $\sum_{i=1}^{n}\left(y_i - \beta_0 - \sum_{k=1}^{p}\beta_k x_{ik}\right)^2$ παρουσιάζεται γεωμετρικά στο παρακάτω διάγραμμα:



Σχήμα 5.1.1 Περιορισμός *LASSO* σε 2 διαστάσεις.

Ο περιορισμός $\sum_{k=1}^{2}|\beta_k| \leq t$ απεικονίζεται από τον ρόμβο ενώ οι ελλείψεις του σχήματος αντιπροσωπεύουν τα σημεία όπου η αντικειμενική συνάρτηση $\sum_{i=1}^{n}(y_i - \beta_0 - \sum_{k=1}^{2}\beta_k x_{ik})^2$ έχει την ίδια τιμή η οποία γίνεται μικρότερη προς το κέντρο των ελλείψεων. Παρατηρούμε ότι η ελαχιστοποίηση της αντικειμενικής συνάρτησης επιτυγχάνεται για κάποιο σημείο της περιοχής που ορίζει ο περιορισμός $\sum_{k=1}^{2}|\beta_k| \leq t$, που παριστάνει ρόμβο. Από το παραπάνω διάγραμμα φαίνεται ότι οι τιμές κάποιων παραμέτρων μπορούν να γίνουν ίσες με το μηδέν όταν η ελαχιστοποίηση της αντικειμενικής συνάρτησης επιτυγχάνεται για κάποια κορυφή του ρόμβου.

Οι εκτιμήτριες στη μέθοδο *Lasso* υπολογίζονται ως εξής:

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{LASSO} &= \min_{\boldsymbol{\beta}}\{\|\boldsymbol{Y} - \boldsymbol{X\beta}\|_2{}^2 + \lambda\|\boldsymbol{\beta}\|_1\} \\ &= \min_{\boldsymbol{\beta}}\{\boldsymbol{Y'Y} - 2\boldsymbol{Y'X\beta} + \boldsymbol{\beta'\beta} + \lambda\|\boldsymbol{\beta}\|_1\} \\ &= \min_{\boldsymbol{\beta}}\{-2\boldsymbol{Y'X\beta} + \|\boldsymbol{\beta}\|_{\mathbf{2}}{}^{\mathbf{2}} + \lambda\|\boldsymbol{\beta}\|_1\}\end{aligned}$$

Στην περίπτωση όπου τα διανύσματα των μεταβλητών X_k είναι ορθογώνια οι εκτιμήτριες ελαχίστων τετραγώνων υπολογίζονται από τη σχέση $\widehat{\boldsymbol{\beta}} = \boldsymbol{X'Y}$, οπότε η παραπάνω σχέση γράφεται ισοδύναμα:

$$\min_{\beta}\sum_{k=1}^{p}(-2\hat{\beta}_{\boldsymbol{k}}\beta_{\boldsymbol{k}} + \beta_{\boldsymbol{k}}{}^{\mathbf{2}} + \lambda|\beta_{\boldsymbol{k}}|).$$

Η παραπάνω σχέση αποτελείται από ένα άθροισμα όρων ο καθένας από τους οποίους αντιστοιχεί σε ξεχωριστή μεταβλητή $\beta_{\boldsymbol{k}}$ από τους υπόλοιπους. Κατά συνέπεια κάθε ένας από τους όρους μπορεί να ελαχιστοποιηθεί ξεχωριστά για να βρεθεί η τιμή της κάθε μεταβλητής.

Οι σχέσεις που θέλουμε να ελαχιστοποιήσουμε είναι οι εξής:

$$\mathcal{L}_k = -2\hat{\beta}_{\boldsymbol{k}}\beta_{\boldsymbol{k}} + \beta_{\boldsymbol{k}}{}^{\mathbf{2}} + \lambda|\beta_{\boldsymbol{k}}|, \qquad k = 1, \dots, p$$

Αν για την εκτιμήτρια $\hat{\beta}_{\boldsymbol{k}}$, έχουμε $\hat{\beta}_{\boldsymbol{k}} > 0$, θα πρέπει και $\beta_{\boldsymbol{k}} \geq 0$, γιατί αλλιώς θα μπορούσαμε να αντιστρέψουμε το πρόσημό της και να πετύχουμε μικρότερη τιμή για την αντικειμενική συνάρτηση. Αντίστοιχα αν $\hat{\beta}_{\boldsymbol{k}} < 0$, θα πρέπει και $\beta_{\boldsymbol{k}} \leq 0$.

Στην περίπτωση όπου $\hat{\beta}_{\boldsymbol{k}} > 0$ η σχέση που θέλουμε να ελαχιστοποιήσουμε έχει την μορφή:

$$\mathcal{L}_k = -2\hat{\beta}_{\boldsymbol{k}}\beta_{\boldsymbol{k}} + \beta_{\boldsymbol{k}}{}^{\mathbf{2}} + \lambda\beta_{\boldsymbol{k}}.$$

Παραγωγίζοντας την παραπάνω σχέση και θέτοντας την ίση με το 0, βρίσκουμε $\beta_{\boldsymbol{k}} = \hat{\beta}_{\boldsymbol{k}} - \frac{\lambda}{2}$, το οποίο είναι εφικτό όταν $\beta_{\boldsymbol{k}} \geq 0$.

Στην περίπτωση όπου $\hat{\beta}_{\boldsymbol{k}} < 0$ η σχέση που θέλουμε να ελαχιστοποιήσουμε έχει την μορφή:

$$\mathcal{L}_k = -2\hat{\beta}_{\boldsymbol{k}}\beta_{\boldsymbol{k}} + \beta_{\boldsymbol{k}}{}^{\mathbf{2}} - \lambda\beta_{\boldsymbol{k}}.$$

Παραγωγίζοντας την παραπάνω σχέση και θέτοντας την ίση με το 0, βρίσκουμε $\beta_{\boldsymbol{k}} = \hat{\beta}_{\boldsymbol{k}} + \frac{\lambda}{2}$, το οποίο είναι εφικτό όταν $\beta_{\boldsymbol{k}} \leq 0$.

Άρα η εκτιμήτριες LASSO, δίνονται από τον τύπο:

$$\hat{\beta}_k{}^{LASSO} = sgn(\hat{\beta}_{\boldsymbol{k}})(|\hat{\beta}_{\boldsymbol{k}}| - \lambda)^+.$$

Το πλεονέκτημα της μεθόδου *LASSO* σε σχέση με την μέθοδο *Ridge* είναι ότι με την κατάλληλη επιλογή της παραμέτρου λ μπορεί να πραγματοποιηθεί η επιλογή των μεταβλητών που συμμετέχουν στο μοντέλο, καθώς οι συντελεστές της παλινδρόμησης μπορούν να γίνουν ίσοι με το μηδέν.

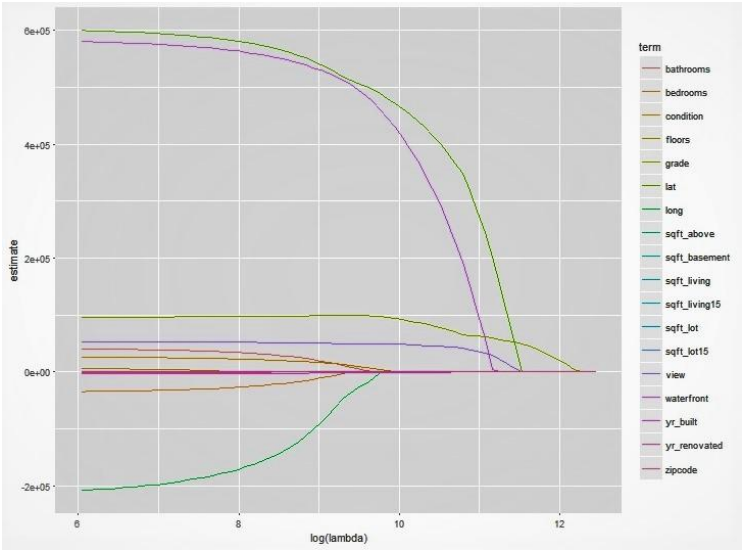
5.2 Η παράμετρος λ στην μέθοδο *LASSO*

Για να εξετάσουμε την επίδραση που έχει η παράμετρος λ στο μοντέλο που παράγεται από την μέθοδο *LASSO* χρησιμοποιούμε το αρχείο δεδομένων που χρησιμοποιούμε για την πρόβλεψη της τιμής των κατοικιών με βάση τα διάφορα χαρακτηριστικά τους. Η υλοποίηση της μεθόδου πραγματοποιήθηκε σε περιβάλλον *R* με την χρήση του πακέτου *'glmnet'*.

Με τον παρακάτω αλγόριθμο κατασκευάζουμε ένα διάγραμμα, στο οποίο παρουσιάζεται η μεταβολή της τιμής των συντελεστών $\hat{\beta}_k^{LASSO}$, σε σχέση με την τιμή της παραμέτρου λ:

```
library(glmnet)
library(ggplot2)
library(broom)
lasso.fit<-glmnet(houses[,-1], houses[,1],
                  alpha=1,
                  standardize = TRUE)
lasso.td<-tidy(lasso.fit)
ggplot(lasso.td, aes(log(lambda), estimate, color = term)) + geom_line(size=0.8)
```

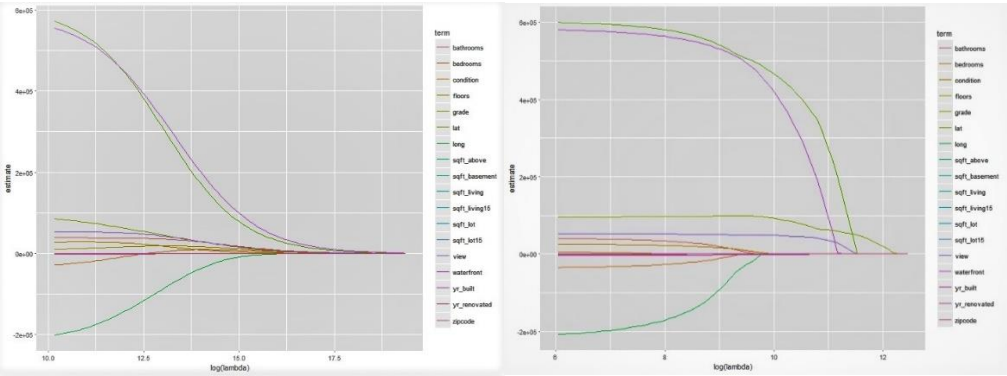
Στο παρακάτω διάγραμμα παρουσιάζεται η μεταβολή των τιμών των εκτιμητριών $\hat{\beta}_k{}^{LASSO}$ για τις διάφορες τιμές της παραμέτρου λ:



Σχήμα 5.2.1 Μεταβολή των εκτιμητριών *LASSO* σε σχέση με την παράμετρο λ.

Παρατηρούμε ότι οι τιμές των εκτιμητριών *LASSO* συρρικνώνονται με την αύξηση της τιμής της παραμέτρου λ και πολλές από αυτές γίνονται ίσες με το μηδέν. Ακόμη μπορούμε να παρατηρήσουμε ότι οι τιμές κάποιων εκτιμητριών μπορούν να αυξηθούν όταν συρρικνώνονται οι τιμές κάποιων άλλων.

Παρακάτω παρουσιάζονται συγκριτικά τα διαγράμματα που μας δίχνουν πως μεταβάλλονται οι συντελεστές $\hat{\beta}_k^{ridge}$ και $\hat{\beta}_k^{LASSO}$, με την μεταβολή της παραμέτρου λ:



Σχήμα 5.2.2 Σύγκριση της μεταβολής των εκτιμητριών *Ridge και LASSO* σε σχέση με την παράμετρο λ.

Από το παραπάνω σχήμα παρατηρούμε την διαφορετική επίδραση που έχει ο περιορισμός της μεθόδου *Ridge* από τον περιορισμό της μεθόδου *LASSO* στην τιμή των συντελεστών της παλινδρόμησης. Παρατηρούμε ότι ο περιορισμός *LASSO* μπορεί να θέσει κάποιους συντελεστές ίσους με το 0, ενώ δεν επιτρέπει την εναλλαγή στο πρόσημο κάποιου συντελεστή, σε αντίθεση με τον περιορισμό *Ridge*.

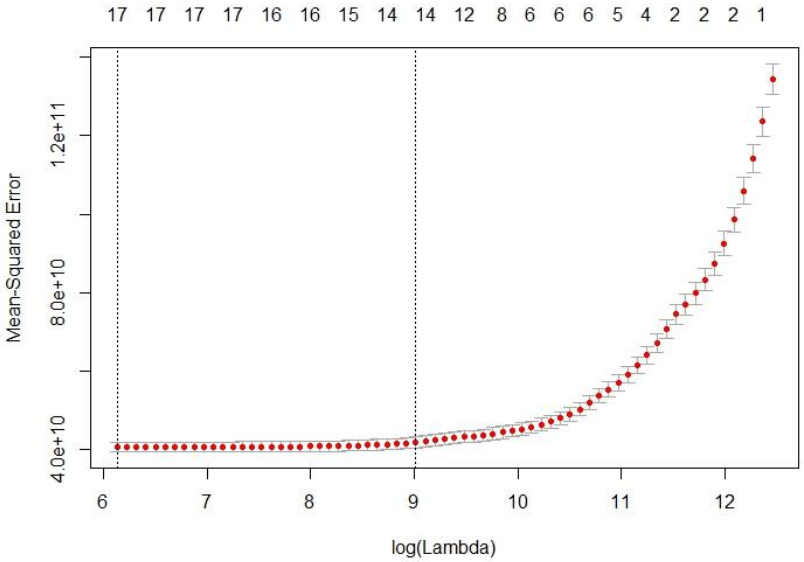
Για την επιλογή της βέλτιστης τιμής της παραμέτρου λ μπορούμε να χρησιμοποιήσουμε την τεχνική *Cross-Validation* χωρίζοντας το αρχείο των δεδομένων σε δύο τμήματα όπου το πρώτο χρησιμοποιείται για τον υπολογισμό των εκτιμητριών και το άλλο για την επιλογή της παραμέτρου λ για την οποία επιτυγχάνεται η ελαχιστοποίηση του μέσου τετραγωνικού σφάλματος. Επειδή όμως με την τεχνική *Cross-Validation* ένα τμήμα των δεδομένων δεν χρησιμοποιείται καθόλου για τον υπολογισμό των $\hat{\beta}_k^{LASSO}$ και άρα μειώνεται η διαθέσιμη πληροφορία που χρησιμοποιείται, επιλέγουμε την τεχνική *K-fold Cross-Validation* κατά την οποία χρησιμοποιείται όλο το σύνολο των δεδομένων.

Για τον υπολογισμό της βέλτιστης τιμής της παραμέτρου λ, χρησιμοποιούμε το πακέτο *'glmnet'* εφαρμόζοντας τον παρακάτω αλγόριθμο:

```
library(glmnet)
cv.lasso<-cv.glmnet(houses[,-1], houses[,1],
alpha=1,
standardize=TRUE)
plot(cv.lasso)
cv.lasso$lambda.min
```

Ο παραπάνω αλγόριθμος εμφανίζει το παρακάτω διάγραμμα:

Στο παρακάτω διάγραμμα παρουσιάζεται η μεταβολή του μέσου τετραγωνικού σφάλματος με την σε σχέση με την μεταβολή της παραμέτρου λ:



Σχήμα 5.2.3 Μεταβολή του *MSE* σε σχέση με την παράμετρο λ.

Παρατηρούμε ότι το μέσο τετραγωνικό σφάλμα δεν μεταβάλεται για κάποιες τιμές παραμέτρου λ και πάνω από κάποια τιμή της παραμέτρου το μέσο τετραγωνικό σφάλμα αυξάνεται. Από το παραπάνω διάγραμμα, μπορούμε επίσης να δούμε το πλήθος των

μεταβλητών που χρησιμοποιεί το μοντέλο για τις διάφορες τιμές του λ. Οι δύο κάθετες διακεκομμένες γραμμές μας δίνουν την τιμή λ_{min} της παραμέτρου για την οποία επιτυγχάνεται το μικρότερο μέσο τετραγωνικό σφάλμα καθώς και την τιμή λ_{1se} που απέχει ένα τυπικό σφάλμα από την τιμή λ_{min}. Η τιμή λ_{1se} μας δίνει ένα απλούστερο μοντέλο, καθώς σε αυτό συμμετέχουν λιγότερες μεταβλητές, χωρίς να αυξάνεται σημαντικά το μέσο τετραγωνικό σφάλμα.

Για να μελετήσουμε την ακρίβεια της μεθόδου και να την συγκρίνουμε με άλλες τεχνικές πραγματοποιούμε την υλοποίηση της μεθόδου σε περιβάλλον R με την χρήση του πακέτου *'elasticnet'* και του πακέτου *'caret'*.

Ο αλγόριθμος που υλοποιεί την μέθοδο είναι ο εξής:

```
library(caret)
library(elasticnet)
trainControl<-trainControl(method="repeatedcv",
                           number=10,repeats=1,
                           classProbs=F)

LassoGrid<-expand.grid(lambda=(seq(0,100,0.1)))

LassoFit<-train(price~.,data = houses,
                method="lasso",
                metric="RMSE",
                preProcess=c("center","scale"),
                trControl=trainControl)

Lassopred<-predict(LassoFit,houses[,-1])

DMwR::regr.eval(houses[,1],Lassopred)
LassoFit
plot(LassoFit)
```

Για την τιμή λ_{min} της παραμέτρου τιμή του μέσου τετραγωνικού σφάλματος που επιτυγχάνεται με την μέθοδο *LASSO* είναι $MSE_{\lambda_{min}} = 4{,}047 \cdot 10^{10}$ και για την τιμή λ_{1se} είναι: $MSE_{\lambda_{1se}} = 4{,}187 \cdot 10^{10}$.

ΚΕΦΑΛΑΙΟ 6

Σύνοψη

Η προσπάθεια που έγινε στα πλαίσια της παρούσας εργασίας είναι η παρουσίαση στατιστικών τεχνικών παλινδρόμησης, οι οποίες είναι κατάλληλες για την ανάλυση μεγάλου όγκου δεδομένων.

Οι τεχνικές που επιλέχθηκαν να χρησιμοποιηθούν ακολουθούν δύο διαφορετικές προσεγγίσεις του προβλήματος. Η πρώτη προσέγγιση έγινε με τη χρήση δέντρων παλινδρόμησης ενώ η δεύτερη πραγματοποιήθηκε με τη χρήση παραλλαγών της κλασσικής γραμμικής παλινδρόμησης.

Η φιλοσοφία των δέντρων παλινδρόμησης βασίζεται στον διαδοχικό διαχωρισμό χώρου των επεξηγηματικών μεταβλητών, και στη συνέχεια η πρόβλεψη δίνεται από ένα απλό μοντέλο, όπως ο μέσος όρος των τιμών της μεταβλητής απόκρισης, για τις παρατηρήσεις που έχουν καταλήξει στον εκάστοτε υπόχωρο. Ωστόσο, ενώ τα δέντρα παλινδρόμησης είναι απλά και γρήγορα στην εφαρμογή τους, έχουν ως μειονέκτημα ότι παράγουν υψηλά σφάλματα πρόβλεψης. Για την αντιμετώπιση αυτού του προβλήματος χρησιμοποιήθηκαν δύο τεχνικές μηχανικής μάθησης, η τεχνική *Stochastic Gradient Boosting* και η τεχνική *Random Forests*.

Η τεχνική *Stochastic Gradient Boosting* προσθέτει διαδοχικά στο αρχικό μοντέλο δέντρα παλινδρόμησης τα οποία έχουν προσαρμοστεί στα κατάλοιπα του προηγούμενου μοντέλου.

Η τεχνική *Random Forests* προσαρμόζει πολλά δέντρα παλινδρόμησης, παράλληλα, σε παραλλαγές του αρχικού αρχείου δεδομένων και δίνει ως πρόβλεψη τον μέσο όρο των προβλέψεων όλων των δέντρων που έχουν κατασκευαστεί.

Οι παραλλαγές της γραμμικής παλινδρόμησης που έχουν αναλυθεί είναι τεχνικές που εφαρμόζουν κάποιον περιορισμό (ποινή) στους συντελεστές παλινδρόμησης. Ο περιορισμός σε κάθε περίπτωση έχει ως αποτέλεσμα τη συρρίκνωση των συντελεστών παλινδρόμησης. Όταν ο αριθμός των μεταβλητών είναι πολύ μεγάλος ή όταν υπάρχει υψηλή συσχέτιση μεταξύ των επεξηγηματικών μεταβλητών, η διασπορά των συντελεστών παλινδρόμησης είναι μεγάλη. Συρρικνώνοντας τους συντελεστές παλινδρόμησης επιτυγχάνεται μικρότερη διασπορά και κατ' επέκταση καλύτερη ακρίβεια πρόβλεψης, εισάγοντας ωστόσο κάποια μεροληψία στο μοντέλο. Οι τεχνικές παλινδρόμησης με ποινή που χρησιμοποιήθηκαν είναι η τεχνική *non-negative garrote estimator,* η τεχνική παλινδρόμησης τύπου *Ridge*, και η τεχνική παλινδρόμησης *LASSO*.

Η τεχνική *non-negative garrote estimator* συρρικνώνει τους συντελεστές παλινδρόμησης κατά έναν μη αρνητικό παράγοντα. Το μειονέκτημα αυτής της μεθόδου είναι, ότι εξαρτάται από το μέγεθος και το πρόσημο των εκτιμητριών ελαχίστων τετραγώνων.

Η τεχνική παλινδρόμησης τύπου *Ridge* εφαρμόζει έναν περιορισμό στο άθροισμα των τετραγωνικό των συντελεστών παλινδρόμησης. Οι εκτιμήτριες των συντελεστών παλινδρόμησης με την τεχνική *Ridge* έχουν το πλεονέκτημα ότι δεν εξαρτώνται από τις εκτιμήτριες ελαχίστων τετραγώνων.

Η τεχνική παλινδρόμησης *LASSO* εφαρμόζει έναν περιορισμό στο άθροισμα των απολύτων τιμών των συντελεστών παλινδρόμησης. Η τεχνική αυτή έχει παρόμοια φιλοσοφία με την τεχνική *Ridge*, έχει όμως το πλεονέκτημα ότι μπορεί να θέσει κάποιους συντελεστές ίσους με μηδέν, και άρα να επιλέξει τις μεταβλητές που συμμετέχουν στο μοντέλο.

Στον παρακάτω πίνακα παρουσιάζεται η ακρίβεια που επιτυγχάνουν οι παραπάνω τεχνικές στην πρόβλεψη της μεταβλητής απόκρισης στο αρχείο δεδομένων που έχουμε χρησιμοποιήσει, καθώς και ο χρόνος υλοποίησής τους σε περιβάλλον *R*, για τις βέλτιστες τιμές των παραμέτρων που τις επηρεάζουν.

ΠΙΝΑΚΑΣ 4

	MAE	*MSE*	*RMSE*	*MAPE*	*TIME*
SGB	$9{,}182 \cdot 10^4$	$2{,}244 \cdot 10^{10}$	$5{,}334 \cdot 10^4$	$1{,}791 \cdot 10^{-1}$	6,27 *sec*
RF	$2{,}843 \cdot 10^4$	$2{,}845 \cdot 10^9$	$5{,}333 \cdot 10^4$	$5{,}457 \cdot 10^{-2}$	84,65 *sec*
NNG	$1{,}247 \cdot 10^5$	$4{,}069 \cdot 10^{10}$	$2{,}017 \cdot 10^5$	$2{,}578 \cdot 10^{-1}$	1,21 *sec*
Ridge	$1{,}241 \cdot 10^5$	$4{,}070 \cdot 10^{10}$	$2{,}017 \cdot 10^5$	$2{,}490 \cdot 10^{-1}$	0,30 *sec*
LASSO	$1{,}257 \cdot 10^5$	$4{,}047 \cdot 10^{10}$	$2{,}011 \cdot 10^5$	$2{,}549 \cdot 10^{-1}$	0,27 *sec*

Παρατηρούμε ότι οι τεχνικές παλινδρόμησης με ποινή επιτυγχάνουν παρόμοια αποτελέσματα σε ό,τι αφορά την ακρίβεια πρόβλεψης. Ακόμη παρατηρούμε ότι ο χρόνος υλοποίησης των τεχνικών αυτών είναι πολύ σύντομος. Αισθητά σημαντική βελτίωση όσο αφορά στην ακρίβεια πρόβλεψης παρουσιάζουν οι δύο τεχνικές που βασίζονται στα δέντρα παλινδρόμησης, με την τεχνική *Random Forests* να επιτυγχάνει τη μεγαλύτερη δυνατή ακρίβεια. Ωστόσο για να επιτευχθεί αυτή η βελτίωση, χρειάζεται περισσότερος χρόνος στην προσαρμογή του μοντέλου.

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Κούτρας, Μ. (2014). Ανάλυση παλινδρόμησης και ανάλυση διακύμανσης, Πανεπιστημιακές Σημειώσεις, ΠΜΣ «Εφαρμοσμένη Στατιστική».

Ξένη

Breiman, L., (1995). Better Subset Regression Using the Nonnegative Garrote. *Technometrics,* Vol. 37, No. 4, 373-384.

Breiman, L., (2001). Random Forests. *Machine Learning,* Vol. 45, Issue 1, 5-32.

Dean, J., (2014). *Big Data, Data Mining and Machine Learning*, Wiley.

Friedman, J., (1999). Stochastic Gradient Boosting. *Computational Statistics & Data Analysis,* Vol. 38, Issue 4, 367-378.

Friedman, J., Hastie, T., Tibsirani, R., (2008). *The Elements of Statistical Learning, 2nd ed.*, Springer.

Harrell, F., (2015). *Regression Modeling Strategies, 2nd ed.*, Springer.

Hastie, T., James, G., Tibshirani, R., Witten, D. (2014). *An Introduction to Statistical Learning*, Springer.

Hoerl, A., Kennard, R., (1970). Ridge Regression: Applications to Nonorthogonal Problems. *Technometrics*, Vol. 12, No. 1, 69-82

Johnson, K., Kuhn, M., (2013). *Applied Predictive Modeling*. Springer.

Jones, Z., Linder, F., (2015). Exploratory Data Analysis using Random Forests. Prepared for the 73rd annual MPSA conference.

Tibshirani, R., (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Society. Series B (Methodological),* Volume 58, Issue 1, 267-288.

Yuan, M., (2007) On the non-negative garrotte estimator. *J.R. Statist. Soc.* B 69, Part 2, 143-161

R Packages

Bates, D., Firth, D., Gebhardt, A., Hornik, K., Ripley, B., Venables, B., (2017). Support Functions and Datasets for Venables and Ripley's MASS. Package ‘MASS’. CRAN repository.

Benesty, M., Candan, C., Cooper, T., Engelhardt, Hunt, T., A., Keefer, Kenkel, B., C., Kuhn, Lescarbeau, R., M., Mayer, Z., R Core Team, Scrucca, L., Tang, Y., Weston, S., Williams, A., Wing, J., Ziem, A., (2017). Classification and Regression Training. Package ‘caret’. CRAN repository.

Breiman, L., Cutler, A., Liaw, A., Wiener, M., (2015). Breiman and Cutler's Random Forests for Classification and Regression. Package ‘randomForest’. CRAN repository.

Friedman, J., Hastie. T., Simon, N., Tibshirani, R., (2017). Lasso and Elastic-Net Regularized Generalized Linear Models. Package ‘glmnet’. CRAN repository.

Ridgeway, G., (2017). Generalized Boosted Regression Models. Package ‘gbm’. CRAN repository.

Ulbricht, J., (2015). Penalized Likelihood Inference for GLMs. Package ‘lqa’. CRAN repository.

Zou, H., (2015). Elastic-Net for Sparse Estimation and Sparce PCA. Package ‘elasticnet’. CRAN repository.

Dataset

House Sales in King County, USA. Retrieved from: https://www.kaggle.com/harlfoxem/housesalesprediction