Κεφάλαιο 6
Παλινδρόμηση

Στόχοι

Στόχος του κεφαλαίου είναι η εισαγωγή του αναγνώστη στις μεθόδους ανάλυσης παλινδρόμησης (regression analysis). Η ανάλυση παλινδρόμησης επιχειρεί να συλλάβει και να αποτυπώσει συσχετίσεις μεταξύ δεδομένων με στόχο τόσο την εξήγηση όσο και την πρόβλεψη των τιμών των μεταβλητών. Στο κεφάλαιο αυτό παρουσιάζονται οι περιπτώσεις στις οποίες μπορεί να χρησιμοποιηθεί η παλινδρόμηση, γεωμετρικές εξηγήσεις αυτής καθώς και μεθόδους και τεχνικές για τον ακριβή προσδιορισμό αυτών. Παρουσιάζονται επίσης και μέθοδοι γιαι την αξιολόγηση και ερμηνεία των εκτιμήσεών τους ανάλογο με τον στόχο του μοντέλου παλινδρόμησης. Όλα τα παραπάνω ο αναγνώστης θα έχει τη δυνατότητα να τα υλοποιήσει εφαρμόζοντας κατάλληλες εντολές που διαθέτει η γλώσσα R.

Προσδοκώμενα Αποτελέσματα

Με την ολοκλήρωση της μελέτης του κεφαλαίου, ο αναγνώστης θα είναι σε θέση να

- Αναγνωρίζει την ανάγκη της ανάλυσης παλινδρόμησης και να αναγνωρίζει τα προβλήματα στα οποία μπορεί να εφαρμοστεί
- Αναγνωρίζει τις διάφορες μορφές των μοντέλων παλινδρόμησης
- Κατανοήσει τον τρόπο λειτουργίας της ανάλυσης παλινδρόμησης για τον προσδιορισμό των συντελεστών

- Κατανοήσει τις μεθόδους που υπάρχουν με τις οποίες μοντέλα παλινδρόμησης μπορούν να προσδιοριστούν και να εκτιμηθούν οι παράμετροί τους και πως οι μέθοδοι αυτοί υλοποιούνται στο περιβάλλον της R.
- Αξιολογήσει και να ερμηνεύσει τα αποτελέσματα της ανάλυσης παλινδρόμησης
- Κατανοήσει τις διάφορες μεθόδους αξιολόγησης και ερμηνείας γραμμικών μοντέλων παλινδρόμησης που εξαρτάται από τον στόχο του μοντέλου.
- Χρησιμοποιεί την R και πραγματικά δεδομένα για να εκφράζει και να εκτιμά μοντέλα παλινδρόμησης

Έννοιες – Κλειδιά

Συνεχής μεταβλητή
Σχέσεις μεταβλητών
Εξαρτημένη μεταβλητή
Ανεξάρτητη μεταβλητή
Διάγραμμα διασποράς
Εξίσωση/μοντέλο παλινδρόμησης
Εξήγηση διακύμανσης
Πρόβελψη τιμής
Γραμμικό μοντέλο παλινδρόμησης
Μη-γραμμική μοντέλο παλινδρόμησης
Εκτίμηση συντελεστών γραμμικού μοντέλου παλινδρόμησης
Μέθοδος Ελαχίστων Τετραγώνων (Ordinary Least Squares - OLS)
Μέθοδος Σταδιακής Καθόδου Δέσμης (Batch Gradient Descent)
Μέθοδος Στοχαστικής Σταδιακής Καθόδου (Stochastic Gradient Descent)
Μέθοδος Σταδιακής Καθόδου Μικρών Δεσμών (Mini-batch Gradient Descent)
Αξιολόγηση γραμμικού μοντέλου παλινδρόμησης
Έλεγχος γραμμικότητας
Ομοσκεδαστικότητα
Διάγραμμα καταλοίπων
Πολυσυγγραμμικότητα
Μετρικές εκτίμησης σφάλματος πρόβλεψης
Διασταυρωμένη επικύρωση k-πτυχών (k-fold Cross Validation)
Υποπροσαρμογή
Υπερπροσαρμογή
Κανονικοποίηση
Παλινδρόμηση LASSO
Παλινδρόμηση Ridge
Παλινδρόμηση Elacticnet

Προαπαιτούμενες γνώσεις

Για την καλύτερη κατανόηση του κεφαλαίου κρίνεται απαραίτητο να έχει μελετηθεί το Κεφάλαιο 3 που περιγράφει τα περιβάλλοντα και τα εργαλεία της R που πρόκειται να χρησιμοποιήσουμε καθώς και το Κεφάλαιο 5 που εισάγει στην Περιγραφική στατιστική και οπτικοποίηση. Επίσης θα χρησιμοποιηθούν έννοιες της Γραμμικής Άλγεβρας.

Εισαγωγικές Παρατηρήσεις

Στο πλαίσιο της μελέτης πολλών σύγχρονων προβλημάτων απαιτείται να εξεταστεί η ακριβής ποσοτική σχέση μεταξύ των μεταβλητών ενός συνόλου δεδομένων, ώστε να συναχθούν χρήσιμα συμπεράσματα. Η σχέση τέτοιων μεταβλητών μελετάται είτε για να διευρευνηθεί πως μία μεταβλητή επηρεάζει μία άλλη μεταβλητή στόχου είτε για την πρόβλεψη της τιμής μιας μεταβλητής στόχου. Εάν η μεταβλητή στόχος που μελετάται λαμβάνει συνεχείς τιμές, η ανάλυση παλινδρόμησης αποτελεί τον κατάλληλο τρόπο ανάλυσης των σχέσεων αυτών. Η ανάλυση παλινδρόμησης έχει ως στόχο τον ποσοτικό προσδιορισμό των σχέσεων μεταξύ μεταβλητών. Παρέχει μεθόδους με τους οποίους η επίδραση μιας μεταβλητής σε μία άλλη μπορεί να προσιοριστεί με ακρίβεια καθώς επίσης να εξετάσει εάν την επηρεάζει σημαντικά. Οι μέθοδοι προσδιορισμού και αξιολόγησης των σχέσεων μεταξύ τέτοιων μεταβλητών χρησιμοποιούνται ευρέως σε πολλούς διαφορετικούς χώρους όπως τα οικονομικά, το μάρκετινγκ, την ιατρική και τις θετικές επιστήμες.

6.1 Βασικές έννοιες

Η ανάλυση παλινδρόμησης (regression analysis) είναι ένα σύνολο στατιστικών διαδικασιών που έχουν ως στόχο την εκτίμηση και αξιολόγηση της σχέσης μεταξύ μεταβλητών σε ένα σύνολο δεδομένων. Ειδικότερα, η ανάλυση παλινδρόμησης έχει ως στόχο τον προσδιορισμό της σχέσης μεταξύ μίας μεταβλητής που απαραιτήτως πρέπει να λαμβάνει συνεχείς αριθμητικές τιμές και καλείται εξαρτημένη μεταβλητή και μίας ή περισσοτέρων άλλων μεταβλητών, που καλούνται ανεξάρτητες μεταβλητές και οι οποίες μπορεί να είναι οποιουδήποτε τύπου δεδομένων. Η εκτίμηση της σχέσης των μεταβλητών αυτών αποσκοπεί στο να μελετήσει

εάν και πόσο οι τιμές της εξαρτημενης μεταβλητής επηρεάζονται από τις τιμές των ανεξάρτητων μεταβλητών με τρόπο που να είναι σύμφωνες με τις παρατηρούμενες - στον πραγματικό κόσμο - τιμές των μεταβλητών αυτών. Γι'αυτό η ανάλυση παλινδρόμησης απαιτεί να υπάρχουν διαθέσιμες πραγματικές, παρατηρούμενες τιμές των μεταβλητών αυτών, οι οποίες θα αναλυθούν και απ'όπου θα εξαχθεί η σχέση των υπο εξέταση μεταβλητών. Στα πλαίσια της βιβλιογραφίας οι μεταβλητές μιας παλινδρόμησης μπορεί να εμφανίζονται και με διαφορετική ορολογία: η εξαρτημένη μεταβλητή μπορεί να βρεθεί επίσης με τους όρους *αποκριτική (responsive)*, *παρατηρηθείσα* ή *μεταβλητή εξόδου*, ενώ οι ανεξάρτητες μεταβλητές μπορούν να εμφανιστούν με τους όρους *προβλέπουσες*, *παλινδρομούσες, επεξηγηματικές* ή *μεταβλητές εισόδου*. Ο στόχος της παλινδρόμησης είναι τόσο να περιγράψει και να εξηγήσει τη σχέση των τιμών μεταβλητών αυτών, και δη της εξαρτημένης και των ανεξάρτητων μεταβλητών, όσο και για να προβλέψει τις τιμές της εξαρτημένης μεταβλητής βάσει των τιμών των ανεξαρτήτων μεταβλητών.

Η σχέση της εξαρτημένης και των ανεξάρτητων μεταβλητών συλλαμβάνεται με τη μορφή εξίσωσης μεταξύ των μεταβλητών αυτών, που καλείται *εξίσωση παλινδρόμησης* ή *μοντέλο παλινδρόμησης*. Η σχέση που αναζητείται μεταξύ των μεταβλητών αυτών, συλλαμβάνεται από αυτήν ακριβώς την εξίσωση και κατά συνέπεια στόχος της παλινδρόμησης είναι η αναζήτηση της εξίσωσης εκείνης που καλύτερα απ'όλες συλλαμβάνει την πραγματική σχέση των μεταβλητών αυτών. Η εξίσωση αυτή μπορεί να λάβει οποιαδήποτε αλγεβρική μορφή· ωστόσο, δεν πρέπει να νοηθεί ως ντετερμινιστική συνάρτηση με την αυστηρή μαθηματική έννοια, αλλά ως μία *στατιστική σχέση* μεταξύ της εξαρτημένης και των ανεξάρτητων μεταβλητών. Αυτό σημαίνει για παράδειγμα ότι η εφαρμογή και η ερμηνεία των αποτελεσμάτων ενός μοντέλου παλινδρόμησης απαιτεί να γίνεται με πολύ συγκεκριμένο τρόπο. Στην εξίσωση παλινδρόμησης, συνηθίζεται η εξαρτημένη μεταβλητή να εμφανίζεται στο αριστερό σκέλος της ενώ οι ανεξάρτητες μεταβλητές εμφανίζονται στο δεξί. Παράδειγμα μίας απλής εξίσωσης/απλού μοντέλου παλινδρόμησης φαίνεται παρακάτω, όπου η μεταβλητή *Κατανάλωση* είναι η εξαρτημένη μεταβλητή που λαμβάνει συνεχείς τιμές, ενώ η μεταβλητή *Εισόδημα είναι* η ανεξάρτητη μεταβλητή:

$$Κατανάλωση = \beta_1 Εισόδημα + \beta_0$$

Στην παραπάνω εξίσωση παλινδρόμησης αυτό που επιχειρείται είναι να συλληφθεί η σχέση μεταξύ της ετήσιας κατανάλωσης ενός νοικοκυριού και του εισοδήματος με ένα μοντέλο παλινδρόμησης. Ειδικότερα επιχειρείται να εκφραστεί ότι η ετήσια κατανάλωση ενός νοικοκυριού (σε χιλιάδες ευρώ) προσδιορίζεται/εξαρτάται από το εισόδημα του νοικοκυριού, εκφρασμένο σε χιλιάδες ευρώ. Οι συντελεστές β_1, β_0 που εμφανίζονται στο παραπάνω μοντέλο παλινδρόμησης καλούνται *παράμετροι ή συντελεστές του μοντέλου παλινδρόμησης* και στόχος της ανάλυσης παλινδρόμησης είναι να προσδιοριστούν οι τιμές τους για το συγκεκριμένο μοντέλο από τα *δεδομένα εκπαίδευσης* (training data), που πρέπει οπωσδήποτε να υπάρχουν. Στην ανάλυση παλινδρόμησης, οι άγνωστες ποσότητες είναι οι συντελεστές β, ενώ οι τιμές των εξαρτημένων και ανεξάρτητων μεταβλητών είναι γνωστές από το σύνολο εκπαίδευσης. Οι συντελεστές γενικά επιχειρούν να συλλάβουν εάν και πόσο ισχυρά μία ανεξάρτητη μεταβλητή επηρεάζει την τιμή της εξαρτημένης. Ο συντελεστής β_0 καλείται επίσης και σταθερός όρος (intercept). Εξαιτίας της αναγκαίας ύπαρξης των δεδομένων εκπαίδευσης για τον προσδιορισμό των παραμέτρων παλινδρόμησης β, η παλινδρόμηση ανήκει στην κατηγορία *μοντέλων επιβλεπόμενης μάθησης*.

Το μοντέλο παλινδρόμησης παραπάνω καλείται και *απλό γραμμικό μοντέλο παλινδρόμησης*. Είναι *παλινδρόμηση*, καθότι συλλαμβάνει τη σχέση μιας εξαρτημένης μεταβλητής που λαμβάνει συνεχείς τιμές (Κατανάλωση) με μία άλλη (Εισόδημα) που είναι η ανεξάρτητη μεταβλητή. Χαρακτηρίζεται ως *απλό*, γιατί το μοντέλο συσχετίζει μόνο δύο μεταβλητές (μεταβλητές Κατανάλωση και Εισόδημα) και *γραμμικό* καθότι γίνεται η υπόθεση ότι η Κατανάλωση έχει μία γραμμική συσχέτιση με το Εισόδημα. Το *γραμμικό* εδώ νοείται με την μαθηματική έννοια και αναφέρει με ποιον τρόπο αλλάζει η Κατανάλωση αν αλλάξει το εισόδημα μιας οικογένειας. Συνηθίζεται επίσης μοντέλα παλινδρόμησης να γράφονται λαμβάνοντας υπόψιν τις παρατηρήσεις στο σύνολο δεδομένων εκπαίδευσης και τότε η μορφή της γίνεται

$$Κατανάλωση_i = \beta_1 Εισόδημα_i + \beta_0$$

όπου ο δείκτης i αναφέρεται στην i-οστή παρατήρηση του συνόλου δεδομένων εκπαίδευσης και εκφράζει το γεγονός ότι η τιμή της μεταβλητής Κατανάλωση στην παρατήρηση i στο σύνολο εκπαίδευσης συσχετίζεται με την τιμή της μεταβλητής Εισόδημα στην ίδια παρατήρηση.

Αν το πλήθος των ανεξαρτήτων μεταβλητών σε μία συνάρτηση παλινδρόμησης είναι παραπάνω από δύο, τότε μιλάμε για ένα *μοντέλο πολλαπλής παλινδρόμησης* (*multiple regression model*) παράδειγμα του οποίου φαίνεται παρακάτω

$$Κατανάλωση = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού + \beta_0$$

Σε αυτό το μοντέλο *πολλαπλής γραμμικής παλινδρόμησης,* επιχειρείται πάλι να καθοριστεί η σχέση της μεταβλητής κατανάλωση με άλλες ανεξάρτητες μεταβλητές, ωστόσο αυτή τη φορά το πλήθος των μεταβλητών είναι τρία: Κατανάλωση, Εισόδημα και Αριθμός ατόμων νοικοκυριού. Κατά συνέπεια πρόκειται για ένα μοντέλο *πολλαπλής γραμμικής παλινδρόμησης*.

Στα μοντέλα παλινδρόμησης, οι τιμές της εξαρτημένης και των ανεξαρτήτων μεταβλητών είναι εξ'αρχής γνωστές από το σύνολο εκπαίδευσης, ενώ άγνωστες είναι οι τιμές των συντελεστών β_i. Όλες οι μέθοδοι ανάλυσης παλινδρόμησης προσπαθούν να εκτιμήσουν όλους τους συντελεστές β του μοντέλου από τα δεδομένα του συνόλου εκπαίδευσης, με τρόπο ώστε το μοντέλο παλινδρόμησης να ταιριάζει όσο καλύτερα γίνεται στις παρατηρούμενες τιμές του συνόλου δεδομένων.

Άσκηση Αυτοαξιολόγησης 0.1

Σε κάθε μία από τις παρακάτω περιπτώσεις, εντοπίστε την εξαρτημένη και την ανεξάρτητη μεταβλητή, εάν πρόκειται να μελετηθούν οι σχέσεις των μεταβλητών αυτών με τη μέθοδο της ανάλυσης παλινδρόμησης:

i. Μία μελέτη προσπαθεί να εξακριβώσει εάν ηλικιωμένοι οδηγοί αυτοκινήτων εμπλέκονται σε περισσότερα ατυχήματα απ'ότι άλλοι οδηγοί. Ο αριθμός των ατυχημάτων ανά 100000 οδηγοί συγκρίνεται με την ηλικία του οδηγού.

ii. Μία μελέτη προσπαθεί να εξετάσει εάν το εβδομαδιαίο ποσό που ξοδεύει ένα νοικοκυριό στο super market μεταβάλλεται με τον αριθμού των ατόμων του νοικοκυριού.

iii. Ασφαλιστικές εταιρείες καθορίζουν το πόσο θα πληρώνεται κάθε μήνα σε ασφάλιστρα σε πολλά συμβόλαια βάσει της ηλικίας του ασφαλισμένου.

iv. Ο λογαριασμός ρεύματος κυμαίνεται ανάλογα με την κατανάλωση ε-νός νοικοκυριού.

v. Μία μελέτη προσπαθεί να εξακριβώσει εάν το επίπεδο εκαίδευσης των ατόμων (μετρούμενο σε έτη που βρίσκεται σε οποιαδήποτε εκπαιδευτική διαδικασία) μειώνει το ποσοστό εγκληματικότητας σε έναν πληθυσμό.

vi. Μία τράπεζα προσπαθεί να μελετήσει εάν το εισόδημα ενός ατόμου αποτελεί ένδειξη για το εάν θα πληρώνει το άτομο κανονικά τις δόσεις ενός δανείου ή όχι..

6.2 Διερεύνηση της σχέσης μεταξύ μεταβλητών

Όταν επιθυμείται να προσδιοριστεί η σχέση μεταξύ δύο ή περισσοτέρων μεταβλητών στα πλαίσια της ανάλυσης παλινδρόμησης, αυτό που αρχικά υπάρχει διαθέσιμο είναι ένα σύνολο δεδομένων με πραγματικές τιμές των υπό εξέταση μεταβλητών δηλαδή τιμές των μεταβλητών που έχουν παρατηρηθεί στον πραγματικό κόσμο. Το σύνολο αυτό των δεδομένων αποτελεί το σύνολο εκπαίδευσης της ανάλυσης παλινδρόμησης, το οποίο θα αναλυθεί και απ' όπου θα προσδιοριστεί η σχέση των μεταβλητών αυτών και κατά συνέπεια το ακριβές μοντέλο παλινδρόμησης.

Ως πρώτο βήμα της μελέτης της σχέσης μεταξύ των εξεταζόμενων μεταβλητών και της αποτύπωσης της τάσης που τις διέπει, γίνεται μία οπτική αναπαράσταση των υπαρχόντων δεδομένων με τη χρήση ενός διαγράμματος διασποράς (scatter plot). Ένα τέτοιο διάγραμμα δίνει μία πρώτη ένδειξη της τάσης ή συσχέτισης που υπάρχει μεταξύ των μεταβλητών που μελετώνται. Με τον όρο «τάση» ή «συσχέτιση» μεταξύ μεταβλητών εννοείται ο τρόπος με τον οποίο αλλάζουν οι τιμές της ανεξάρτητης μεταβλητής, αν μεταβληθούν οι τιμές των ανεξαρτήτων μεταβλητών και ειδικότερα αν αυξηθεί ή μειωθεί η τιμή μιας ανεξάρτητης μεταβλητής, πως θα μεταβληθεί η τιμή της εξαρτημένης. Το διάγραμμα διασποράς δίνει μία πρώτη ένδειξη του πως οι υπό εξέταση μεταβλητές συμμεταβάλλονται.

Για παράδειγμα, το αρχείο HouseholdData.csv, διατηρεί παρατηρήσεις σχετικά με το ετήσιο εισόδημα νοικοκυριών (μεταβλητή Income, σε ευρώ), το ετήσιο ποσό που καταναλώνουν σε τρόφιμα (μεταβλητή FoodExpenditure, σε ευρώ), το πλήθος των ατόμων του νοικοκυριού (μεταβλητή FamilySize), τα χρόνια εκπαίδευσης του επικεφαλής (Household Head) του νοικοκυριού (μεταβλητή YearsOfEducationHH – με τον όρο επικεφαλής εννοείται εκείνο το μέλος του νοι-

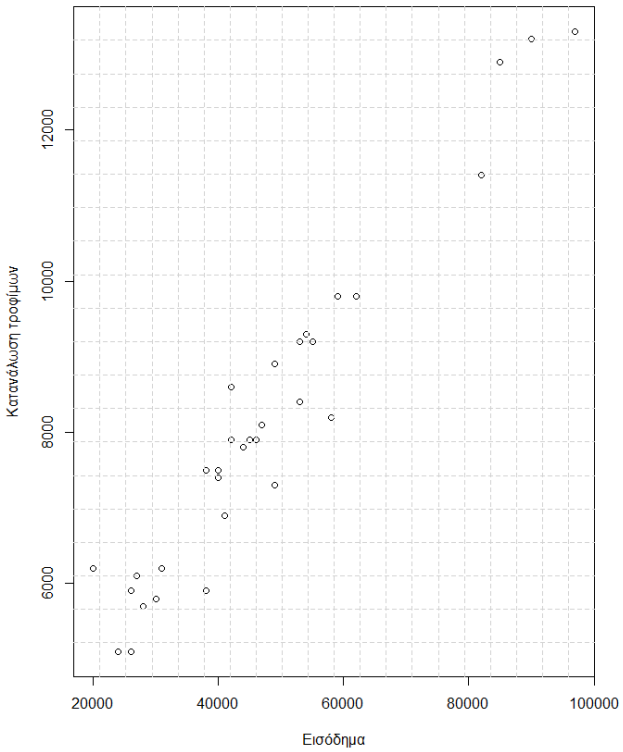
κοκυριού με το μεγαλύτερο εισόδημα) καθώς και το φύλο του επικεφαλής του νοικοκυριού (μεταβλητή GenderHH, που λαμβάνει τιμή 0 για αρσενικό και 1 για θηλυκό). Οι παραητήσεις που περιέχει το αρχείο αυτό χρησιμοποιούνται στο κεφάλαιο μπορούν να βρεθούν στο παράρτημα Α. Το ερώτημα που επιθυμείται να μελετηθεί στα πλαίσια των δεδομένων αυτών είναι η σχέση που υπάρχει μεταξύ του ποσού που καταναλώνει ένα νοικοκυριό σε τρόφιμα (μεταβλητή FoodExpenditure) και του εισοδήματος του ίδιου νοικοκυριού (μεταβλητή Income). Ειδικότερα επιχειρείται να μελετηθεί με ποιον τρόπο το ποσό που ξοδεύουν τα νοικοκυριά ετησίως για τρόφιμα σχετίζεται με/εξαρτάται από το εισόδημά του. Γι' αυτόν τον λόγο επιλέγεται η μεταβλητή που δηλώνει την κατανάλωση τροφίμων ως εξαρτημένη μεταβλητή και το εισόδημα ως η ανεξάρτητη[1]. Προς τούτο δημιουργείται ως πρώτο βήμα ένα διάγραμμα διασποράς μεταξύ των μεταβλητών κατανάλωση τροφίμων (στον άξονα y) και εισόδημα (στον άξονα x) των διαθέσιμων δεδομένων. Στο περιβάλλον της R ένα τέτοιο διάγραμμα διασποράς μπορεί να δημιουργηθεί για το δοθέν σύνολο δεδομένων με τον ακόλουθο κώδικα:

```
# Ρύθμιση του τρόπου εμφάνισης των αριθμών στις γραφικές παραστά-
σεις: εδώ καθορίζεται ότι αριθμοί δεν θα εμφανίζονται με την επι-
στημονική τους μορφή.
options(scipen = 999)


# Ανάγνωση αρχείου δεδομένων που περιέχει τα δεδομένα σε μορφή CSV
data<-read.csv("HouseholdData.csv ", sep=",", header=T)


# Δημιουργία διαγράμματος διασποράς με τη μεταβλητή Εισόδημα στον
άξονα x και τη μεταβλητή κατανάλωση τροφίμων στον άξονα y
```

[1] Εδώ γίνεται η υπόθεση ότι το ποσό που ξοδεύει ένα νοικοκυριό σε τρόφιμα *είναι αποτέλεσμα του εισοδήματος του νοικοκυριού* δηλαδή ότι το ποσό κατανάλωσης τροφίμων ενός νοικοκυριού είναι συνάρτηση του εισοδήματός του, γι'αυτό και θεωρείται η μεταβλητή της κατανάλωσης τροφίμων ως η εξαρτημένη μεταβλητή. Γενικά θα μπορούσε και η μεταβλητή του εισοδήματος να θεωρηθεί ως η εξαρτημένη μεταβλητή και μεταβλητή κατανάλωσης τροφίμων ως η ανεξάρτητη. Ωστόσο το ποια μεταβλητή θα επιλεγεί ως η εξαρτημένη μεταβλητή εξαρτάται από το ερώτημα που επιθυμείται να μελετηθεί. Μία πιο αναλυτική συζήτηση για τον τρόπο επιλογής της εξαρτημένης μεταβλητής στα πλαίσια της ανάλυσης παλινδρόμησης μπορεί να βρεθεί στο [Waugh, 1942]

```
plot(data$Income, data$FoodExpenditure, xlab="Εισόδημα",
ylab="Κατανάλωση τροφίμων")


# Προσθήκη πλέγματος με διακεκομμένη γραμμή κάθε 20 στιγμές στον
άξονα x

grid(20, 20, col='lightgrey', lty=2)
```

Το διάγραμμα διασποράς των δεδομένων που δημιουργείται εμφανίζεται στο παρακάτω σχήμα 6.1



Εικόνα 0.1 *Μία πρώτη απεικόνιση της σχέσης των μεταβλητών Κατανάλωση τροφίμων και Εισόδημα για το δοθέν σύνολο δεδομένων, απεικονισμένη ως διάγραμμα διασποράς.*

Κάθε σημείο στο παραπάνω διάγραμμα διασποράς αναπαριστά το ποσό που καταναλώνει ένα νοικοκυριό σε τρόφιμα ετησίως και το εισόδημά του. Όπως διαφαίνεται από το διάγραμμα διασποράς στο σχήμα 6.1, υπάρχει μία συσχέτιση μεταξύ των μεταβλητών Κατανάλωση τροφίμων και Εισόδημα στο σύνολο δεδομένων: όσο μεγαλύτερο το εισόδημα, τόσο μεγαλύτερη και η κατανάλωση. Ή διαφορετικά διαφαίνεται η τάση να αυξάνει το ποσό που καταναλώνουν οι οικογένειες για τρόφιμα, εάν αυξάνει το εισόδημά τους και η αύξηση αυτή φαίνεται να είναι ανάλογη της αύξησης τους εισοδήματος. Αυτό σημαίνει ότι η σχέση μεταξύ των μεταβλητών φαίνεται να υπάρχει *θετική συσχέτιση* μεταξύ των μεταβλητών αυτών και η συσχέτιση αυτή είναι γραμμική. Ο χαρακτηρισμός *θετική* για τη συσχέτιση σημαίνει ότι αύξηση της ανεξάρτητης μεταβλητής φαίνεται να οδηγεί και σε αύξηση της εξαρτημένης μεταβλητής. Εάν μία αύξηση της ανεξάρτητης μεταβλητής συνοδεύεται από μείωση της τιμής της εξαρτημένης μεταβλητής, τότε γίνεται λόγος για αρνητική συσχέτιση των μεταβλητών. Επιπλέον, στο σχήμα 6.2 επιχειρείται να αποτυπωθεί η σχέση και τάση των μεταβλητών αυτών με πιο ξεκάθαρο τρόπο ως μία ευθεία γραμμή που διαπερνά τα δεδομένα. Η ευθεία γραμμή που διαπερνά τα δεδομένα στο σχήμα 6.2, είναι ένα τρόπος αναπαράστασης της τάσης/σχέσης που υπάρχει μεταξύ των μεταβλητών κατανάλωση τροφίμων και εισόδημα. Η συγκεκριμένη γραμμή φαίνεται να συλλαμβάνει ικανοποιητικά τη σχέση που υπάρχει μεταξύ των μεταβλητών αυτών. Το πως ακριβώς σχεδιάστηκε η γραμμή αυτή θα αναλυθεί στις επόμενες ενότητες. Ωστόσο, αυτό που πρέπει να γίνει κατανοητό στο σημείο αυτό είναι ότι παρόλο που η σχέση των δύο αυτών μεταβλητών είναι εμφανής «με το μάτι» από το διάγραμμα διασποράς, η παλινδρόμηση καθορίζει μαθηματικά επακριβώς τη σχέση τους και την εξίσωση παλινδρόμησης. Και ακριβώς ο ποσοτικός καθορισμός της σχέσης τους στα πλαίσια της παλινδρόμησης, χρησιμοποιείται τόσο για την ερμηνεία όσο και για την πρόβλεψη των μελλοντικών τιμών της εξαρτημένης μεταβλητής, αν είναι γνωστές οι τιμές των ανεξάρτητων μεταβλητών.

Έτσι για παράδειγμα, επειδή όπως φαίνεται μία ευθεία γραμμή φαίνεται να προσεγγίζει καλά τη σχέση των υπό εξέταση μεταβλητών, μπορεί να γίνει η υπόθεση ότι ένα γραμμικό μοντέλο παλινδρόμησης θα αποτυπώνει καλά τη σχέση των μεταβλητών και θα ταιριάζει καλά με τα δεδομένα του συνόλου εκπαίδευσης, κι

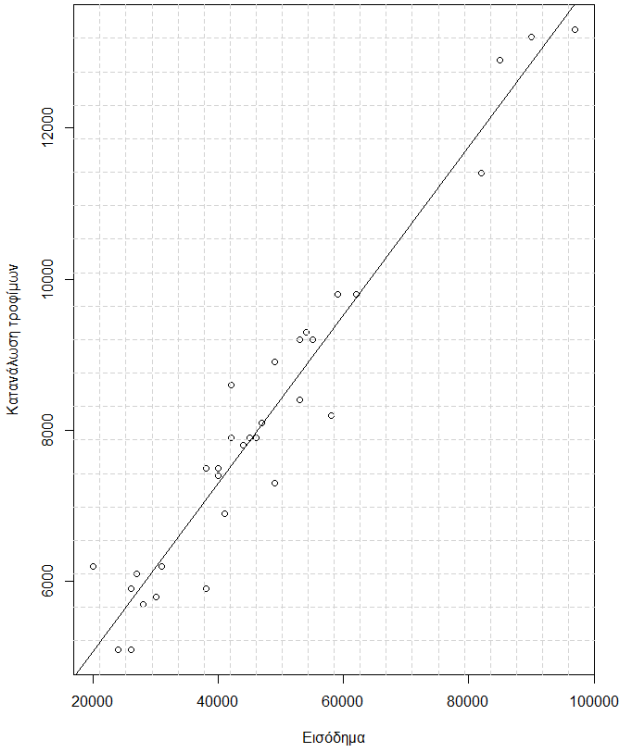
έτσι η σχέση μεταξύ των μεταβλητών μπορεί να συλληφθεί ικανοποιητικά από το εξής μοντέλο παλινδρόμησης[2]

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_0$$

όπου οι μεταβλητές Κατανάλωση τροφίμων και Εισόδημα είναι τιμές των παρατηρήσεων του συνόλου εκπαίδευσης και οι συντελεστές β_1 και β_0 σταθερές. Το σημαντικό σε ένα μοντέλο παλινδρόμησης είναι οι συντελεστές β, οι οποίοι θα πρέπει να προσδιοριστούν επακριβώς από τα δεδομένα εκπαίδευσης για το συγκεκριμένο μοντέλο.

Αφού προσδιοριστούν οι συντελεστές β ενός μοντέλου παλινδρόμησης μπορεί να γίνει και αξιολόγηση του μοντέλου αναφορικά με το πόσο καλά εξηγεί, προβλέπει και συλλαμβάνει τη σχέση των υπο εξέταση μεταβλητών. Αν και σε αυτό το σημείο δεν θα αναλυθούν λεπτομέρειες σχετικά με το πως εκτιμήθηκαν οι συντελεστές β, εφόσον αυτοί προσδιοριστούν, τότε μπορεί και το μοντέλο παλινδρόμησης να απεικονιστεί πάνω στο διάγραμμα διασποράς των δεδομένων, δείχνοντας τις τιμές που προβλέπει το μοντέλο, ώστε να μπορεί να αποτυπωθεί η σχέση των υπο εξέταση μεταβλητών πάνω στα ίδια τα δεδομένα και να φανεί «με το μάτι» πόσο καλά το μοντέλο παλινδρόμησης ταιριάζει στα δεδομένα. Τέτοια απεικόνιση των τιμών που προβλέπει το μοντέλο παλινδρόμησης καλείται και *γραμμή παλινδρόμησης* και φαίνεται στο παρακάτω διάγραμμα.

[2] Οι εξισώσεις της μορφής y= ax+b όπου a και b σταθερές, γραφικά απεικονίζουν μία ευθεία γραμμή με κλίση a και οι οποίες τέμνουν τον άξονα y στο σημείο (0, b).



Εικόνα 0.2 *Απεικόνιση της γραμμής παλινδρόμησης που προκύπτει από το μοντέλο παλινδρόμησης Κατανάλωση τροφίμων = β1Εισόδημα+ β0 επί του σύνολο δεδομένων για να φανεί πόσο καλά το μοντέλο συλλαμβάνει και ταιριάζει στη σχέση των υπό εξέταση μεταβλητών Κατανάλωση Τροφίμων*

Το παραπάνω μοντέλο παλινδρόμησης γραφικά απεικονίζει μία ευθεία γραμμή, της οποίας η κλίση -που συλλαμβάνει τον ρυθμό μεταβολής της σχέσης των μεταβλητών- εκφράζεται από τον συντελεστή β_1 και το σημείο που τέμνει τον άξονα Υ εκφράζεται από τον συντελεστή β_0. Με ένα τέτοιο μοντέλο παλινδρόμησης όπως το παραπάνω και αφού προσδιοριστούν οι συντελεστές β, μπορεί να μελετηθεί ποσοτικά η σχέση τους και από εκεί να εκτιμηθεί για παράδειγμα η τιμή της κατανάλωσης αν δοθεί τιμή στην ανεξάρτητη μεταβλητή εισόδημα.

Εν κατακλείδι, ένα διάγραμμα διασποράς μεταξύ των υπο εξέταση μεταβλητών αποτελεί χρήσιμο εργαλείο να προσκομιστούν ενδείξεις για τη συσχέτιση των

μεταβλητών αυτών όπως αποτυπώνονται στο διαθέσιμο σύνολο δεδομένων. Αναλυτικότερα, ένα διάγραμμα διασποράς δίνει ενδείξεις για να απαντηθούν ερωτήματα όπως:

- Υπάρχει σχέση μεταξύ της ανεξάρτητης και εξαρτημένων μεταβλητών και αν ναι ποια η φύση της;
- Είναι η σχέση των υπο εξέταση μεταβλητών γραμμική ή όχι;
- Πόσο ισχυρή είναι η σχέση των υπό εξέταση μεταβλητών;
- Η διακύμανση της εξαρτημένης μεταβλητής, εξαρτάται από την τιμή των ανεξαρτήτων μεταβλητών;
- Υπάρχουν ακραίες τιμές στο σύνολο δεδομένων;

Οι ενδείξεις που συλλέγονται από το διάγραμμα διασποράς αποτελούν τη βάση για τη έκφραση πιο αναλυτικών μοντέλων παλινδρόμησης μεταξύ των υπό εξέταση μεταβλητών.

Άσκηση Αυτοαξιολόγησης 0.2

Ποιες από τις παρακάτω προτάσεις είναι σωστές και ποιες λάθος;

(α) Η γραμμή παλινδρόμησης ενός απλού μοντέλου παλινδρόμησης που απεικονίζεται σε ένα διάγραμμα διασποράς, συλλαμβάνει τον ρυθμό μεταβολής της εξαρτημένης μεταβλητής ως προς την ανεξάρτητη μεταβλητή.

(β) Ένα διάγραμμα διασποράς που δημιουργείται με τη συνάρτηση plot() της R, μπορεί να χρησιμοποιηθεί για τη διαγραμματική διερεύνηση της σχέσης μεταξύ πέντε (5) μεταβλητών.

(γ) Ο σταθερός όρος β_0 (συντελεστής β_0) ενός απλού γραμμικού μοντέλου παλινδρόμησης εκφράζει την υπόθεση, ότι ο ρυθμός μεταβολής της ανεξάρτητης προς την εξαρτημένη μεταβλητή είναι σταθερός.

6.3 Εξερεύνηση σχέσεων μεταβλητών: Η μήτρα διαγραμμάτων διασποράς

Στην παραπάνω παρουσίαση του ρόλου του διαγράμματος διασποράς, η αφετηρία ήταν ένα πολύ συγκεκριμένο ερώτημα και δη, αν το ποσό που καταναλώνουν τα νοικοκυριά σε τρόφιμα σχετίζεται με το εισόδημά του.

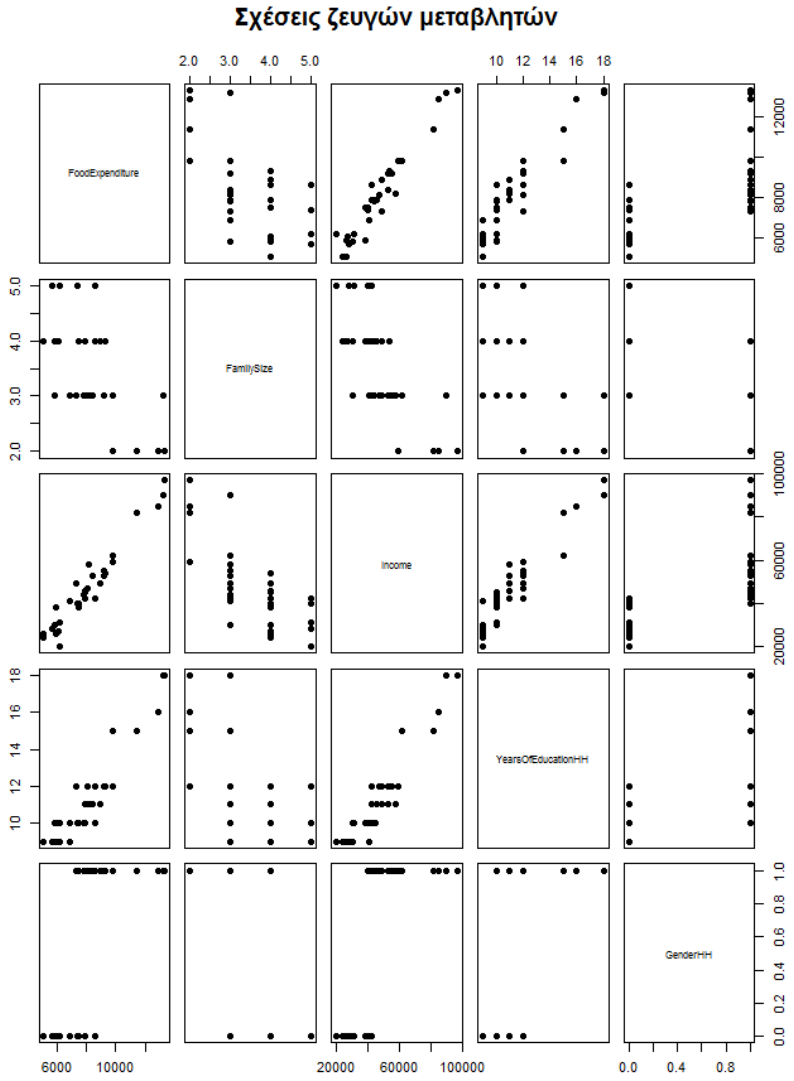
Ωστόσο, σε περιπτώσεις που το ερώτημα δεν έχει διατυπωθεί σαφώς εξαρχής και επιθυμείται μία εξερεύνηση των σχέσεων που τυχόν υπάρχουν *μεταξύ οποιουδήποτε ζεύγους μεταβλητών του συνόλου δεδομένων*, μπορεί να γίνει χρήση της μήτρας διαγράμματος διασποράς (scatter plot matrix). Η μήτρα διαγράμματος διασποράς δημιουργεί διαγράμματα διασποράς μεταξύ όλων συνδυασμών ζευγών των μεταβλητών του συνόλου δεδομένων (ή επιλεγμένου υποσυνόλου αυτών) και παραθέτει τα διαγράμματα αυτά σε μορφή μήτρας, ώστε να είναι εύκολη η ανάγνωση και σύγκρισή τους. Έτσι, αν το σύνολο δεδομένων έχει συνολικά n μεταβλητές, η μήτρα διαγράμματος διασποράς μεταξύ κάθε ζεύγους μεταβλητών, θα παραγάγει συνολικά $\frac{n!}{(n-2)!}$ σε πλήθος διαγράμματα διασποράς.

Στην R, η μήτρα διαγράμματος διασποράς μπορεί να γίνει με τη συνάρτηση pairs(), η οποία δέχεται ως όρισμα τις μεταβλητές που επιθυμείται ο συνδυασμός τους να απεικονιστεί ως διάγραμμα διασποράς καθώς και ορίσματα εμφάνισης (όπως είδος στιγμών και χρώματα) του διαγράμματος. Παρακάτω φαίνεται κώδικας σε R ο οποίος θα απεικονίσει τη μήτρα διαγράμματος διασποράς για όλες τις μεταβλητές (πλην της μεταβλητής Family) του αρχείου HouseholdData.csv :

```
options(scipen = 999)

# Ανάγνωση αρχείου δεδομένων

foodConsumptionData<-read.csv("HouseholdData.csv", sep=",",
header=T)

# Μήτρα διαγράμματος διασποράς για κάθε συνδυασμό ζεύγους μεταβλη-
τών

# στο σύνολο δεδομένων

# Το πρώτο όρισμα δηλώνει ονομαστικά ποιες μεταβλητές του συνόλου
δεδομένων να ληφθούν υπόψιν, ενώ
```

```
# το όρισμα pch καθορίζει το είδος στιγμών που θα χρησιμοποιηθεί
για την απεικόνιση των δεδομένων. Το

# όρισμα pch=19 είναι προσδιοριστής εμφάνισης: καθορίζει το είδος
των στιγμών που είναι κύκλος με πλήρες χρώμα

# Η εντολή pairs() μπορεί να συνταχθεί και με τον ακόλουθο τρόπο,
παράγοντας το ίδιο αποτέλεσμα:

# pairs( foodConsumptionData[, 2:6], main="Σχέσεις ζευγών μεταβλη-
τών", pch = 19)

# προσδιορίζονται με δείκτες τις υπο εξέταση μεταβλητές και όχι
ονομαστικά

pairs(~FoodExpenditure+FamilySize+Income++YearsOfEducationHH+Gender
HH, data=foodConsumptionData, main="Σχέσεις ζευγών μεταβλητών", pch
= 19)
```

Ο παραπάνω κώδικας, αν εκτελεστεί για το σύνολο δεδομένων, θα εμφανίσει την ακόλουθη μήτρα διαγράμματος διασποράς



Εικόνα 0.3: *Μήτρα διαγράμματος διασποράς, όπου εμφανίζονται διαγράμματα διασποράς μεταξύ όλων των ζευγών μεταβλητών ενός συνόλου δεδομένων με τη μορφή μήτρας. Χρησιμοποιείται για να ληφθεί μία πρώτη ένδειξη της σχέσης μεταξύ των μεταβλητών.*

Κάθε διάγραμμα διασποράς που εμφανίζεται στην παραπάνω μήτρα του σχήματος 6.3, είναι ένα διάγραμμα διασποράς μεταξύ κάθε ζεύγους μεταβλητων που

εμφανίζονται στο σύνολο δεδομένων. Το όνομα της μεταβλητής που εμφανίζεται σε κάθε γραμμή, δηλώνει τη μεταβλητή που υπάρχει στον άξονα y (ως εξαρτημένη μεταβλητή) σε κάθε διάγραμμα της ίδιας γραμμής, ενώ το όνομα της μεταβλητής σε κάθε στήλη, δηλώνει τη μεταβλητή που βρίσκεται στον άξονα X (ως ανεξάρτητη μεταβλητή) στα διαγράμματα της ίδιας στήλης. Για παράδειγμα το διάγραμμα που εμφανίζεται στην πρώτη γραμμή και δεύτερη στήλη εμφανίζει το διάγραμμα διασποράς μεταξύ των μεταβλητών FoodExpediture και FamilySize όπου η μεταβλητή FoodExpenditure είναι στον άξονα y (εξαρτημένη μεταβλητή) και η μεταβλητή FamilySize στον άξονα x (ανεξάρτητη μεταβλητή). Αντίστοιχα, το διάγραμμα διασποράς της τρίτης γραμμής και δεύτερης στήλης απεικονίζει τη σχέση μεταξύ της μεταβλητής Income (ως εξαρτημένη μεταβλητή) και FamilySize (ως ανεξάρτητης μεταβλητής). Σε μία μήτρα διαγράμματος διασποράς, εμφανίζονται όλες οι μεταβλητές τόσο ως εξαρτημένες όσο και ως ανεξάρτητες μεταβλητές. Για παράδειγμα, στην παραπάνω μήτρα, το διάγραμμα διασποράς στην δεύτερη γραμμή και πρώτη στήλη επιχειρεί να απεικονίσει τη συσχέτιση μεταξύ των ίδιων μεταβλητών (FoodExpenditure και FamilySize) όπως και το διάγραμμα της πρώτης γραμμής και δεύτερης στήλης, μόνο που στο πρώτο η εξαρτημένη μεταβλητή είναι. Μόνο που στο πρώτο διάγραμμα, στον άξονα y βρίσκεται η μεταβλητή FamilySize ενώ στο άλλο η μεταβλητή FoodExpenditure.

Η μήτρα διαγράμματος διασποράς αποτελεί έναν βολικό συνοπτικό τρόπο προκειμένου να απεικονιστούν και ταυτόχρονα να συγκριθούν οι συσχετίσεις μεταξύ όλων των ζευγών μεταβλητών ενός συνόλου δεδομένων. Γίνεται για παράδειγμα εύκολο αντιληπτό ότι υπάρχει θετική συσχέτιση μεταξύ των μεταβλητών FoodExpenditure και Income όπως επίσης και μεταξύ των μεταβλητών YearsOfEducationHH και Income. Αντιθέτως αν και διακρίνεται συσχέτιση μεταξύ των μεταβλητών FamilySize και Income αυτή φαίνεται να μην είναι ισχυρή.

Άσκηση Αυτοαξιολόγησης 0.3

Εάν η συσχέτιση μεταξύ δύο μεταβλητών που λαμβάνουν συνεχείς τιμές X, Y φαίνεται να είναι αρνητική, τί υποθέσεις μπορείτε να κάνετε για την τιμή που θα λάβει ο συντελεστής β_1 του ακόλουθου απλού μοντέλου παλινδρόμησης;

$$Y = \beta_1 X \quad + \beta_0$$

Δραστηριότητα 0.1

Δίνονται τα παρακάτω δεδομένα μιας ομάδας μπάσκετ, που αναφέρουν πόσο χρόνο κάνει προθέρμανση (σε λεπτά) η ομάδα πριν έναν αγώνα και τους τραυματισμούς των παικτών κατά τη διάρκεια του αγώνα:

Χρόνος προθέρμανσης	0	30	10	15	5	25	35	40
Τραυματισμοί	4	1	2	2	3	1	0	1

Συγγράψτε πρόγραμμα σε R, το οποίο να εξετάζει με τη χρήση διαγράμματος εάν υπάρχει συσχέτιση μεταξύ των μεταβλητών αυτών και εφόσον υπάρχει, αν αυτή είναι θετική ή αρνητική.

6.4 Στόχοι ενός μοντέλου παλινδρόμησης

Δύο είναι οι βασικοί λόγοι για τους οποίους συντάσσεται κι εκτιμάται ένα μοντέλο παλινδρόμησης που επιχειρεί να ταιριάξει καλά στα δεδομένα εκπαίδευσης:

(1) Για να *προβλέψει, με όση μεγαλύτερη ακρίβεια γίνεται,* την τιμή αυτής εξαρτημένης μεταβλητής, βάσει μόνο των τιμών των ανεξαρτήτων μεταβλητών. Εάν για ένα σύνολο δεδομένων έχει προσδιοριστεί το μοντέλο παλινδρόμησης, τότε αυτό μπορεί να χρησιμοποιηθεί για να υπολογίσει αυτής μελλοντικές τιμές αυτής εξαρτημένης μεταβλητής αν του δοθούν ως είσοδο οι τιμές των ανεξαρτήτων μεταβλητών.

Έτσι για παράδειγμα, εάν έχει προσδιοριστεί επακριβώς το μοντέλο παλινδρόμησης (και δη οι συντελεστές) του μοντέλου που συλλαμβάνει τη σχέση αυτής μεταβλητής κατανάλωσης τροφίμων με αυτής, μπορεί να χρησιμοποιηθεί για την πρόβλεψη αυτής τιμής αυτής μεταβλητής Κατανάλωση τροφίμων, αν δοθεί ως είσοδος η τιμή του εισοδήματος. Η τιμή αυτής εξαρτημένης μεταβλητής που υπολογίζεται από το μοντέλο παλινδρόμησης καλείται αυτής και *προβλεφθείσα τιμή* για αυτής τιμές των ανεξαρτήτων μεταβλητών. Προκειμένου να μπορούν να διακριθούν οι παρατηρούμενες/πραγματικές τιμές αυτής εξαρτημένης μεταβλητής που υπάρχουν στο σύνολο δεδομένων, από αυτής τιμές που προβλέπει το

μοντέλο παλινδρόμησης, οι προβλεφθείσες τιμές αυτής εξαρτημένης μεταβλητής που προκύπτουν από το μοντέλο παλινδρόμησης εμφανίζονται με τον χαρακτήρα καπελάκι (hat) ως εξής $\hat{y}$ ή $\widehat{Κατανάλωση\ τροφίμων}$.

Ωστόσο, αν γίνει χρήση αυτής μοντέλου παλινδρόμησης για την πρόβλεψη αυτής τιμής αυτής εξαρτημένης μεταβλητής, το αποτέλεσμα που δίνει το μοντέλο δεν πρέπει να ερμηνευθεί ντετερμινιστικά αλλά στατιστικά. Για παράδειγμα αυτής για το μοντέλο παλινδρόμησης κατανάλωσης τροφίμων έχουν προσδιοριστεί οι συντελεστές β από το σύνολο δεδομένων εκπαίδευσης, και έχουν για παράδειγμα βρεθεί ότι είναι ίσοι με β_1=0.1786 και β_0 = -223.81[3] το μοντέλο παλινδρόμησης θα λάβει τη μορφή:

$$Κατανάλωση\ τροφίμων\ = 0.17816 Εισόδημα - \ 223.81$$

Τότε, για ετήσιο εισόδημα νοικοκυριού ίσο με 76700 ευρώ, η κατανάλωση τροφίμων του νοικοκυριού προβλέπεται από το μοντέλο να είναι

$$\widehat{Κατανάλωση\ τροφίμων}\ = 0.17816 * 76700 - \ 223.81 = 13441.06$$

δηλαδή 13441.06 ευρώ. Ωστόσο το αποτέλεσμα αυτό δεν πρέπει να ερμηνευθεί ντετερμινιστικά: δηλαδή δεν πρέπει να ερμηνευθεί ότι όποιο νοικοκυριό έχει εισόδημα 76700 ευρώ *με βεβαιότητα θα καταναλώνει το ακριβές ποσό των 13441.06 ευρώ ετησίως για τρόφιμα*. Και αυτό γιατί το ποσό που καταναλώνουν τα νοικοκυριά για τρόφιμα και έχουν το ίδιο εισόδημα μπορεί να παρουσιάζει μία διακύμανση: για παράδειγμα αυτής διαφαίνεται από το σύνολο δεδομένων υπάρχουν νοικοκυρά με εισόδημα 42000 Ευρώ αλλά ξοδεύουν για τρόφιμα το ένα 6900 και το άλλο 7200 Ευρώ. Η ερμηνεία αυτής προβλεφθείσας από το μοντέλο τιμής πρέπει να γίνεται στατιστικά: η τιμή 13441.06 σημαίνει ότι το ποσό που καταναλώνουν τα νοικοκυριά για τρόφιμα θα ακολουθεί μία κατανομή με συγκεκριμένη διακύμανση, *όπου η μέση τιμή αυτής κατανομής αυτής θα είναι 13441.06 ευρώ*. Ή πιο απλά, το αποτέλεσμα του μοντέλου παλινδρόμησης αναφέρει ότι ένα νοικοκυριό *με ετήσιο εισόδημα 76700 ευρώ θα ξοδεύει κατά μέσο όρο το ποσό των 13441.06 Ευρώ σε τρόφιμα*.

(2) Για να *εξηγήσει και να ερμηνεύσει* τη διακύμανση της τιμής της εξαρτημένης μεταβλητής, βάσει των τιμών των ανεξάρτητων μεταβλητών και να γενικεύσει την εξήγηση αυτή στο σύνολο του πληθυσμού. Η έμφαση εδώ είναι στην αποτύ-

[3] Οι τιμές των συντελεστών β σε ένα μοντέλο παλινδρόμησης μπορεί να είναι και αρνητικοί.

πωση της κατανόησης για το πως οι ανεξάρτητες μεταβλητές επηρεάζουν τη διακύμανση της εξαρτημένης.

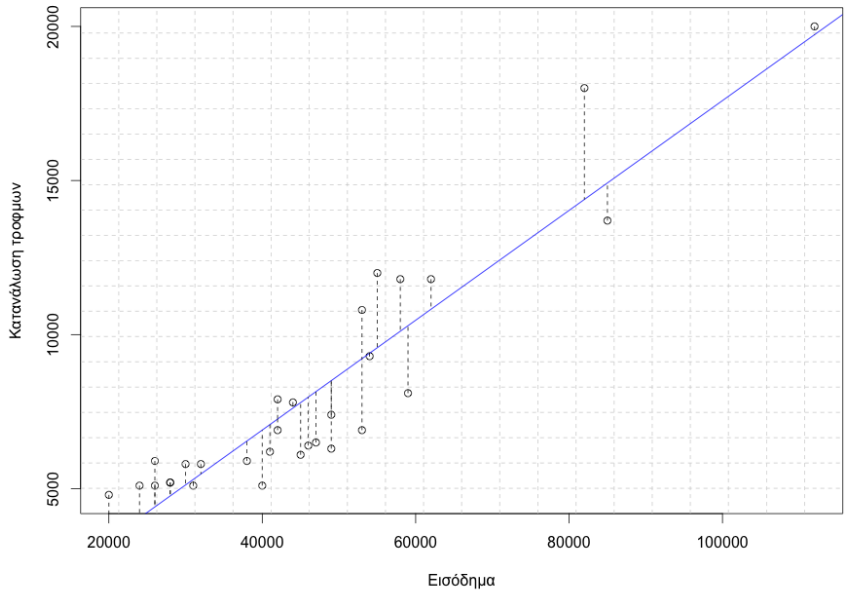
Γενικά, η διακύμανση των τιμών μιας μεταβλητής αποτελεί από τις σημαντικότερες πτυχές της και προσπάθειες ερμηνείας της, δηλαδή η μελέτη που οφείλονται οι διακυμάνσεις αυτές, μπορεί να γίνει με χρήση της μεθόδου της παλινδρόμησης. Σε μία τέτοια περίπτωση, επιχειρείται να μελετηθεί πως οι ανεξάρτητες μεταβλητές που εμφανίζονται σε ένα μοντέλο παλινδρόμησης μπορούν ή όχι να ερμηνεύσουν τη διακύμανση των τιμών της εξαρτημένης μεταβλητής.

Από το διάγραμμα διασποράς του σχήματος 6.2 αυτό που φαίνεται είναι ότι οι πραγματικές τιμές δεδομένων Κατανάλωσης τροφίμων και Εισοδήματος δεν πέφτουν ακριβώς πάνω στη γραμμή παλινδρόμησης με την οποία προσεγγίστηκε τυπικά η τάση/σχέση των μεταβλητών αυτών. Ορισμένες παρατηρήσεις είναι πάνω από τη γραμμή και άλλες παρατηρήσεις είναι κάτω από αυτήν. Δηλαδή, οι τιμές των δεδομένων παρουσιάζουν μία διασπορά ή διακύμανση γύρω από τις τιμές που προβλέπει το μοντέλο παλινδρόμησης. Αυτό επί της ουσίας σημαίνει ότι το συγκεκριμένο μοντέλο παλινδρόμησης που απεικονίστηκε δεν μπορεί να εξηγήσει πλήρως τη διακύμανση όλων των τιμών της εξαρτημένης μεταβλητής. Στην ιδανική περίπτωση, οι πραγματικές παρατηρήσεις θα πρέπει να πέφτουν ακριβώς πάνω στη γραμμή παλινδρόμησης που συλλαμβάνει τη σχέση των μεταβλητών αυτών, και τότε θα λέγονταν ότι η γραμμή παλινδρόμησης εξηγεί όλη τη διακύμανση της εξαρτημένης μεταβλητής. Στην πράξη ωστόσο, κάτι τέτοιο δεν συμβαίνει σχεδόν ποτέ.

Η διασπορά/διακύμανση που δεν μπορεί να εξηγηθεί από την γραμμή παλινδρόμησης μπορεί να μετρηθεί και ένας τρόπος μέτρησης είναι να βρεθεί η απόσταση κάθε σημείου πραγματικών δεδομένων από την τιμή που προβλέπει το μοντέλο παλινδρόμησης δηλαδή τη γραμμή παλινδρόμησης. Ως απόσταση εδώ νοείται ως η ευθεία η οποία είναι παράλληλη προς τον άξονα y, περνάει από το σημείο δεδομένων και τέμνει την γραμμή παλινδρόμησης[4]. Στο σχήμα 6.4 απεικονίζονται αυτές οι αποστάσεις –που φαίνονται με διακεκομμένη γραμμή- κάθε σημείου

[4] Υπάρχουν και άλλοι τρόποι να μετρηθεί η απόσταση του σημείου από τη γραμμή παλινδρόμησης όπως για παράδειγμα να γίνει η χρήση της ευθείας που περνάει από το σημείο δεδομένων και είναι κάθετη προς τη γραμμή τάσης. Ο τρόπος επιλογής του μέτρου απόστασης σε αυτό το σημείο δεν έχει μεγάλη σημασία καθότι ο στόχος είναι να δειχθεί πως νοείται γεωμετρικά η έννοια της εξήγησης της διακύμανσης της εξαρτημένης μεταβλητής.

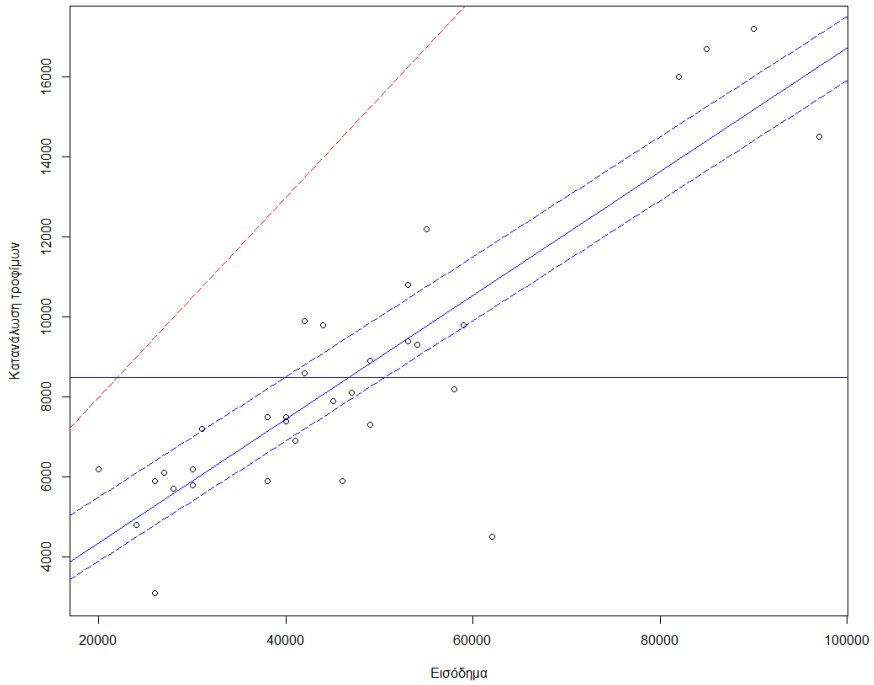
από τη γραμμή παλινδρόμησης του μοντέλου για την μελέτη της κατανάλωσης τροφίμων σε νοικοκυριά. Οι αποστάσεις αυτές από τη γραμμή παλινδρόμησης υπολογίζονται ως η διαφορά της παρατηρούμενης τιμής της εξαρτημένης μεταβλητής από την προβλεφθείσα τιμή του μοντέλου για την ίδια τιμή της ανεξάρτητης μεταβλητής $\left|Κατανάλωση\ τροφίμων_i - \widehat{Κατανάλωση}\ τροφίμων_i\right|$[5] όπου i η i-οστή παρατήρηση του συνόλου δεδομένων. Οι επιμέρους αποστάσεις κάθε σημείου μπορούν να αθροιστούν και να προκύψει έτσι μία μετρική για την διακύμανση που δεν μπορεί να ερμηνευτεί από το μοντέλο. Όσο μεγαλύτερος ο αριθμός αυτός, τόσο περισσότερη διακύμανση δεν μπορεί να ερμηνευτεί από το μοντέλο παλινδρόμησης.



Εικόνα 0.4 *Απόσταση δεδομένων από τη γραμμή παλινδρόμησης που αποτυπώνει τη διακύμανση των παρατηρούμενων τιμών που δεν μπορεί να εξηγήσει το μοντέλο παλινδρόμησης (ευθεία γραμμή).*

[5] Γενικά, οι έννοιες «απόσταση των δεδομένων από τη γραμμή παλνδρόμησης» και «διαφορά πραγματικής και προβλεφθείσας τιμής» θα θεωρηθούν ότι εκφράζουν το ίδιο πράγμα στο κείμενο που ακολουθεί. Επιπλέον, ο ακριβής τύπος υπολοφισμού της απόστασης αυτής θα παρουσιαστεί σε άλλη ενότητα.

Ωστόσο, για ένα σύνολο δεδομένων, μπορούν να χαραχθούν κι άλλες, πολλές γραμμές παλινδρόμησης που αποτυπώνουν την τάση των μεταβλητών αυτών όπως φαίνεται στο παρακάτω σχήμα 6.5. Για κάθε μία από τις γραμμές αυτές παλινδρόμησης, υπάρχει ένα ποσό της διακύμανσης της εξαρτημένης μεταβλητής που δεν μπορεί να εξηγηθεί από τη γραμμή και η οποία μπορεί να μετρηθεί από τις αποστάσεις των σημείων από την εκάστοτε γραμμή.



Εικόνα 0.5 *Υπάρχουν πολλές υποψήφιες γραμμές παλινδρόμησης που μπορούν να αποτυπωθούν για ένα σύνολο δεδομένων. Η γραμμή που ελαχιστοποιεί τις αποστάσεις λέγεται ότι συλλαμβάνει και εξηγεί τη διακύμανση καλύτερα από οποιοδήποτε άλλο μοντέλο*

Απ' όλες αυτές τις γραμμές που μπορούν να χαραχθούν προκειμένου να αποτυπώσουν την τάση των μεταβλητών αυτών, αυτή που καλύτερα απ'όλες εξηγεί τη διακύμανση της εξαρτημένης μεταβλητής είναι εκείνη όπου οι αποστάσεις των σημείων δεδομένων από τη γραμμή αυτή είναι οι μικρότερες δυνατές. Μία τέτοια γραμμή παλινδρόμησης, αφήνει ανεξήγητη τη μικρότερη διακύμανση και

εξηγεί το μεγαλύτερο μέρος της. Όπως θα δειχθεί σε επόμενες ενότητες, η μέθοδος της ανάλυσης παλινδρόμηση επιχειρεί πάντα να καταλήγει σε μοντέλα παλινδρόμησης που ελαχιστοποιούν τις αποστάσεις των σημείων και κατά συνέπεια εξηγούν την διακύμανση καλύτερα από οποιαδήποτε άλλο μοντέλο.

Ωστόσο, αυτό που πρέπει να μείνει από την παραπάνω παρατήρηση είναι το ότι συντελείται μία σύγκριση μοντέλων παλινδρόμησης με ένα μοντέλο παλινδρόμησης βάσης. Και από τη σύγκριση αυτή αξιολογούνται τα μοντέλα παλινδρόμησης σχετικά με την ικανότητά τους να εξηγήσουν τη διακύμανση της εξαρτημένης μεταβλητής. Το μοντέλο βάσης που αναφέρθηκε και με το οποίο συγκρίνονται τα όλα τα άλλα μοντέλα παλινδρόμησης είναι εκείνο όπου δεν εμφανίζεται καμία ανεξάρτητη μεταβλητή σε αυτό και η τιμή της εξαρτημένης μεταβλητής τίθεται ίση με τη μέση τιμή της εξαρτημένης μεταβλητής του συνόλου δεδομένων. Έτσι, για το σύνολο δεδομένων, υπολογίζεται η μέση τιμή της εξαρτημένης μεταβλητής Κατανάλωση τροφίμων ως εξής

```
data<-read.csv("HouseholdData.csv ", sep=",", header=T)
# Υπολογισμός μέσης τιμής κατανάλωσης τροφίμων
FoodExpenditure.mean <- mean(data$FoodExpenditure)
```

Εκτελώντας τον παραπάνω κώδικα, θα προκύψει μέση τιμή για την μεταβλητή Κατανάλωση τροφίμων ίση με 8477.143. Αυτό σημαίνει ότι το μοντέλο παλινδρόμησης, με το οποίο θα συγκριθούν όλα τα άλλα μοντέλα όσ'αναφορά την εξήγηση της διακύμανσης θα έχει τη μορφή:

$$Κατανάλωση\ τροφίμων = 8477.143$$

Το παραπάνω μοντέλο παλινδρόμησης απεικονίζεται γραφικά και στο σχήμα 6.5 ως η ευθεία γραμμή παράλληλη προς τον άξονα x. Και για το μοντέλο αυτό, το πόσο καλά ερμηνεύει τη διακύμανση των δεδομένων βρίσκεται υπολογίζονται τις αποστάσεις των σημείων των δεδομένων από τη συγκεκριμένη παράλληλη γραμμή ή διαφορετικά από τις τιμές που προβλέπει το μοντέλο βάσης. Το μοντέλο παλινδρόμησης αυτό αναφέρει ότι έαν στο μοντέλο δεν εισαχθεί καμία ανεξάρτητη μεταβλητή, τότε η καλύτερη τιμή της εξαρτημένης μεταβλητής που μπορεί να προβλεφθεί είναι ίση με 8477.143.

Αν συνταχθεί ένα νέο μοντέλο παλινδρόμησης για τα ίδια δεδομένα, όπου εισάγεται μία επιπλέον μεταβλητή όπως το Εισόδημα

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_0$$

και η γραμμή παλινδρόμησης μειώνει τις αποστάσεις των δεδομένων σε σχέση με τις αποστάσεις τους από το μοντέλο βάσης, τότε λέγεται ότι το νέο αυτό μοντέλο συλλαμβάνει και ερμηνεύει καλύτερα τη διακύμανση σε σχέση με τη μέση τιμή της εξαρτημένης μεταβλητής. Από την άλλη, αν η εισαγωγή της μεταβλητής Εισόδημα στο μοντέλο δεν συνεισφέρει στη σημαντική μείωση των αποστάσεων σε σχέση με το μοντέλο βάσης, τότε αυτό σημαίνει ότι η μεταβλητή Εισόδημα δεν μπορεί να ερμηνεύσει τη διακύμανση των δεδομένων και κατά συνέπεια η μεταβλητή Εισόδημα δεν προσφέρει τίποτε καινούργιο στο μοντέλο και θα πρέπει να αφαιρεθεί απ'αυτό. Προσθέτοντας περισσότερες μεταβλητές στο μοντέλο παλινδρόμησης μπορεί να οδηγήσει σε ακόμη μικρότερες αποστάσεις των σημείων από τη γραμμή παλινδρόμησης και κατά συνέπεια ακόμη καλύτερη εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής. Για παράδειγμα, αν εκτός του εισοδήματος προστεθεί και το πλήθος ατόμων του νοικοκυριού ως ανεξάρτητη μεταβλητή στο υπόδειγμα και προκύψει η πολλαπλή γραμμική παλινδρόμηση:

$$Κατανάλωση = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού + \beta_0$$

η σύγκριση θα γίνει και πάλι με το μοντέλο βάσης. Αν οι αποστάσεις του μοντέλου αυτού μειώνονται ακόμη πιο πολύ τις από τις αποστάσεις που προκύπτουν στο απλό μοντέλο παλινδρόμησης που έχει μόνο τη μεταβλητή Εισαόδημα ως ανεξάρτητη, τότε λέγεται ότι το πολλαπλό μοντέλο γραμμικής παλινδρόμησης αυτό ερμηνεύει ακόμη καλύτερα τη διακύμανση των δεδομένων. Αν όχι, τότε λέγεται ότι δεν έχει καμία επεξηγηματική δύναμη και η μεταβλητή Αριθμός ατόμων νοικοκυριού μπορεί να αφαιρεθεί από το μοντέλο καθότι δεν προσφέρει κάτι επιπλέον.

Ωστόσο, όσο καλά και να προσαρμόζεται στα δεδομένα και να εξηγεί τη διακύμανση ένα μοντέλο παλινδρόμησης, πάντα θα υπάρχει ένα ποσό διακύμανσης της εξαρτημένης μεταβλητής που δεν θα μπορεί να συλληφθεί από το μοντέλο. Κατα συνέπεια, οι πραγματικές τιμές της εξαρτημένης μεταβλητής θα «απέχουν» από τις τιμές που προβλέπει το μοντέλο παλινδρόμησης. Αυτή η απόσταση αποτυπώνεται ρητά και στο μοντέλο παλινδρόμησης με την εισαγωγή ενός νέου όρου που συλλαμβάνει ακριβώς αυτή την διακύμανση που δεν μπορεί να εξηγήσει

το μοντέλο. Έτσι, λαμβάνοντας υπόψιν αυτήν την παρατήρηση, η πλήρης μορφή του μοντέλου παλινδρόμησης για την Κατανάλωση τροφίμων λαμβάνει τη μορφή

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_0 + \varepsilon$$

όπου ο όρος ε που εμφανίζεται καλείται *σφάλμα*, *διαταρακτικός όρος* ή *θόρυβος* του μοντέλου παλινδρόμησης και εκφράζει τη μη ερμηνεύσιμη από το μοντέλο διακύμανση. Ή διαφορετικά, ο όρος ε συλλαμβάνει τη διακύμανση που δεν μπορεί να εξηγηθεί από τις μεταβλητές που υπάρχουν στο μοντέλο παλινδρόμησης. Έτσι κάθε μοντέλο παλινδρόμησης που συντάσσεται, απαρτίζεται από δύο παράγοντες που συμμετέχουν στην εκτίμηση της τιμής της εξαρτημένης μεταβλητής:

- Έναν ντετερμινιστικό παράγοντα που αναπαρίσταται από τις ανεξάρτητες μεταβλητές που περιέχει
- Έναν τυχαίο/στοχαστικό παράγοντα που αναπαρίσταται από τον όρο ε ο οποίος είναι τυχαίος (δηλαδή δεν είναι γνωστή η τιμή του) και συλλαμβάνει τη διακύμανση που δεν μπορεί να ερμηνεύσει η ανεξάρτητη μεταβλητή

Η μη ερμηνευόμενη διακύμανση που συλλαμβάνεται από το όρο ε, μπορεί να οφείλεται σε άλλη ή άλλες μεταβλητές που θα πρέπει να ενσωματωθούν στον μοντέλο ή από έναν μη γραμμικό όρο των υπαρχόντων που θα πρέπει να προστεθεί. Σε κάθε μοντέλο παλινδρόμησης υπάρχει ο διαταρακτικός όρος ε και σε μοντέλα όπου αυτός δεν απεικονίζεται ρητά, θα υπονοείται.

Επιπλέον, αν και παραπάνω αναφέρθηκε ως μέτρο εξήγησης και ερμηνείας της διακύμανσης η έννοια της απόστασης των δεδομένων από τη γραμμή παλινδρόμησης, το μέτρο αυτό στη συζήτηση που έγινε δεν έχει τυποποιηθεί μαθηματικά. Ο ακριβής τύπος που υπολογίζει το πόσο καλά ένα μοντέλο εξηγεί τη διακύμανση βάσει των αποστάσεων που συζητήθηκαν παραπάνω αποτελεί σημαντικό τμήμα της ανάλυσης παλινδρόμησης και θα παρουσιαστεί σε επόμενη ενότητα.

Οι δύο αυτοί στόχοι ενός μοντέλου παλινδρόμησης που αναλύθηκαν παραπάνω, είναι συνήθως αμοιβαία αποκλειόμενοι: ένα μοντέλο που εξηγεί καλά τη διακύμανση της εξαρτημένης μεταβλητής δεν μπορεί συνήθως να χρησιμοποιηθεί για την ακριβή πρόβλεψη της τιμής της για άγνωστα δεδομένα και αντιστρόφως. Αυτό γιατί οι διαφορετικοί αυτοί στόχοι απαιτούν διαφορετικές διαδικασίες που αφορούν τόσο τον τρόπο με τον οποίο επιλέγονται οι ανεξάρτητες μεταβλητές

του μοντέλου όσο και τον τρόπο με τον οποίο τα μοντέλα ελέγχονται και αξιολογούνται. Έτσι για παράδειγμα, μοντέλα που έχουν ως στόχο την εξήγηση της διακύμανσης υποβάλλονται σε διαφορετικούς ελέγχους από μοντέλα που έχουν ως στόχο την πρόβλεψη της ίδιας εξαρτημένης μεταβλητής. Γι' αυτό τον λόγο, πριν την σύνταξη ενός μοντέλου παλινδρόμησης πρέπει εξαρχής να γίνεται σαφές, ποιος είναι ο στόχος του.

Άσκηση Αυτοαξιολόγησης 0.4

Έστω δύο διαφορετικά μοντέλα παλινδρόμησης που επιχειρούν να εξηγήσουν την διακύμανση της ίδιας εξαρτημένης μεταβλητής. Με ποιον τρόπο μπορεί να γίνει η σύγκρισή τους, προκειμένου να εξεταστεί ποιο εξηγεί καλύτερα τη διακύμανση της εξαρτημένης μεταβλητής;

6.5 Γραμμική και μη-γραμμική παλινδρόμηση

Ένα μοντέλο ή εξίσωση παλινδρόμησης μπορεί να λάβει οποιαδήποτε αλγεβρική μορφή. Ανάλογα με τη μορφή του, ένα μοντέλο μπορεί να χαρακτηριστεί ως γραμμικό ή μη-γραμμικό μοντέλο παλινδρόμησης. Ο χαρακτηρισμός αυτός αναφέρεται κυρίως στον τρόπο με τον οποίο μεταβάλλεται η τιμή της εξαρτημένης μεταβλητής, εάν μεταβληθεί η τιμή μιας ή όλων των συντελεστών ή/και ανεξαρτήτων μεταβλητών.

Η γραμμική παλινδρόμηση αποτελεί μία από τις πιο διαδεδομένες και σημαντικές μορφές της όπου η τιμή της εξαρτημένης μεταβλητής είναι ένας γραμμικός συνδυασμός μιας ή περισσοτέρων ανεξαρτήτων μεταβλητών. Παράδειγμα ενός γραμμικού μοντέλου παλινδρόμησης φαίνεται παρακάτω:

$$Κατανάλωση = \beta_1 Εισόδημα + \beta_0 + \varepsilon$$

Συνηθίζεται συχνά ένα γραμμικό μοντέλο παλινδρόμησης να αναφέρεται στη γενική, αφηρημένη μορφή του ως εξής:

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \cdots + \beta_0 + \varepsilon$$

όπου β_i οι συντελεστές και X_i οι ανεξάρτητες μεταβλητές. Αυτό που κάνει ένα γραμμικό μοντέλο παλινδρόμησης όπως το παραπάνω, «γραμμικό» είναι ο τρόπος με τον οποίο αλλάζει η τιμή της ανεξάρτητης μεταβλητής αν αλλάξουν οι τι-

μές των συντελεστών ή/και των ανεξαρτήτων μεταβλητών. Γενικά λέγεται ότι μεταξύ δύο μεταβλητών Υ και Χ υπάρχει γραμμική σχέση - ή ότι *η μεταβλητή Υ είναι γραμμική ως προς τη μεταβλητή Χ* - εάν ο ρυθμός μεταβολής της εξαρτημένης μεταβλητής Υ ως προς την ανεξάρτητη Χ (δηλαδή ο λόγος $^{dY}/_{dX}$) δεν εξαρτάται από την τιμή της ανεξάρτητης μεταβλητής Χ. Ή διαφορετικά, ότι ο ρυθμός μεταβολής είναι σταθερός και η μεταβολή στην ανεξάρτητη μεταβλητή είναι ανάλογη της μεταβολής της ανεξάρτητης μεταβλητής $\Delta Y = \beta \Delta X$ όπου β μία σταθερά.

Για παράδειγμα, στο γραμμικό μοντέλο παλινδρόμησης
$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_0$, η εξαρτημένη μεταβλητή Κατανάλωση τροφίμων είναι γραμμική τόσο ως προς την παράμετρο β_1όσο και ως προς την ανεξάρτητη μεταβλητή Εισόδημα. Αυτό γιατί οι ρυθμοί μεταβολής της εξαρτημένης μεταβλητής Κατανάλωση ως προς τις μεταβολές της παραμέτρου β_1 ($\frac{\Delta Κατανάλωση}{\Delta \beta_1}$) και της μεταβλητής Εισόδημα $\left(\frac{\Delta Κατανάλωση}{\Delta Εισόδημα}\right)$ δεν εξαρτώνται από τις τιμές β_1 και του Εισοδήματος αντίστοιχα. Αν υποθέσουμε ότι τόσο η παράμετρος β_1 όσο και το Εισόδημα αυξηθούν κατά 1 μονάδα, οι παραπάνω ρυθμοί μεταβολής θα είναι:

$$\frac{\Delta Κατανάλωση}{\Delta \beta_1} = \frac{(\beta_1 + 1) Εισόδημα + \beta_0 - \beta_1 Εισόδημα - \beta_0}{1}$$
$$= Εισόδημα\ (ανεξάρτητο\ της\ τιμής\ \beta_1)$$

$$\frac{\Delta Κατανάλωση}{\Delta Εισόδημα} = \frac{\beta_1 (Εισόδημα + 1) + \beta_0 - \beta_1 Εισόδημα - \beta_0}{1}$$
$$= \beta_1\ (ανεξάρτητο\ της\ τιμής\ της\ μεταβλητής\ Εισόδημα)$$

Σε μοντέλα πολλαπλής γραμμικής παλινδρόμησης, όπως το παρακάτω:

$$Κατανάλωση = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού + \beta_0$$

η εξαρτημένη μεταβλητή είναι γραμμική τόσο ώς προς όλες τις παραμέτρους β όσο και ως προς όλες τις ανεξάρτητες μεταβλητές. Η εξέταση γραμμικότητας σε τέτοιο μοντέλο πολλαπλής παλινδρόμησης γίνεται θεωρώντας ότι μία μόνο ανεξάρτητη μεταβλητή ή παράμετρος μεταβάλλεται, ενώ όλες οι υπόλοιπες παράμετροι και ανεξάρτητες μεταβλητές παραμένουν σταθερές.

Για να χαρακτηριστεί ένα μοντέλο παλινδρόμησης ως *γραμμικό*, θα πρέπει η *εξαρτημένη μεταβλητή να είναι γραμμική μόνο ως προς όλους τους συντελεστές β*.

Δεν απαιτείται να είναι γραμμική και ως προς τις ανεξάρτητες μεταβλητές. Αυτό σημαίνει για παράδειγμα, ότι και όλα τα παρακάτω μοντέλα είναι μοντέλα γραμμικής παλινδρόμησης, αφού σε αυτά *η εξαρτημένη μεταβλητή είναι γραμμική ως προς τις παραμέτρους β_i, αλλά όχι ως προς τις ανεξάρτητες μεταβλητές*

$$Κατανάλωση = \beta_1 Εισόδημα^2 + \beta_0 \ (i)$$

$$Κατανάλωση = \beta_1 Εισόδημα^2 + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού^3 * Αριθμός\ πισίνων^3 + \beta_0 \quad (ii)$$

$$Αρτηριακή\ πίεση = \beta_1 Φύλλο + \beta_2 \sqrt{Ηλικία} + \beta_0 \quad (iii)$$

$$\ln(Εισόδημα) = \beta_1 Εμπειρία + \beta_2 Εμπειρία^2 + \beta_3 Εκπαίδευση + \beta_0 \quad (iv)$$

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_0 \quad (v)$$

Σημασιολογικά, ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης συλλαμβάνει το γεγονός ότι κάθε ανεξάρτητη μεταβλητή επηρεάζει και συνεισφέρει στην αναμενόμενη τιμή της εξαρτημένης μεταβλητής με κάποιο συντελεστή (ή βάρος) β, και η συνεισφορά των ανεξαρτήτων μεταβλητών είναι αθροιστική. Οι συντελεστές β καθορίζουν το πόσο συνεισφέρει κάθε ανεξάρτητη μεταβλητή.

Μοντέλα παλινδρόμησης στα οποία η εξαρτημένη μεταβλητή δεν είναι γραμμική ως προς τουλάχιστον έναν συντελεστή β καλούνται *μη-γραμμικά*. Παραδείγματα μη-γραμμικών μοντέλων παλινδρόμησης φαίνεται παρακάτω:

$$Κατανάλωση = \beta_1 Εισόδημα + \beta_2^3 Αριθμός\ ατόμων\ νοικοκυριού + \beta_0$$

$$Βάρος = \left(\frac{1}{\beta_1 \beta_2}\right)^2 Ύψος + \beta_0$$

$$y = e^{\beta_1} x + \beta_0$$

$$y = \sqrt{\beta_1}\, x_1 + \beta_2^5\, x_2 + \beta_0$$

$$y = \frac{\beta_1 x}{\beta_0 + x}$$

Η διαφοροποίηση μεταξύ γραμμικού και μη γραμμικού μοντέλου παλινδρόμησης είναι σημαντική καθότι καθορίζει τον τρόπο με τον οποίο θα γίνει η εκτίμηση των συντελεστών. Αν και υπάρχουν τρόποι με τους οποίους μη-γραμμικά μοντέλα μπορούν να μετασχηματιστούν σε γραμμικά (με μεθόδους που καλούνται γραμ-

μικοί μετασχηματισμοί), στη γενική περίπτωση γραμμικά μοντέλα παλινδράμησης χρησιμοποιούν διαφορετικές μεθόδους για την εκτίμηση των συντελεστών τους απ'ότι μη-γραμμικά μοντέλα. Έτσι για παράδειγμα, αν και υπάρχουν κλειστοί τύποι για την εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης, τέτοιοι δεν υφίστανται στην περίπτωση μη-γραμμικών, όπου κυρίως γίνεται χρήση αριθμητικών μεθόδων για τον προσδιορισμό τους.

Άσκηση Αυτοαξιολόγησης 0.5

Εξετάστε εάν το παρακάτω μοντέλο παλινδρόμησης είναι γραμμικό ή μη-γραμμικό. Θεωρείστε ότι β_i είναι οι συντελεστές του μοντέλου και Y, X η ανεξάρτητη και εξαρτημένη μεταβλητή αντίστοιχα:

$$Y = \beta_1 X^{\beta_2} + \beta_0$$

6.6 Μοντέλα γραμμικής παλινδρόμησης

Επειδή τα μοντέλα γραμμικής παλινδρόμησης αποτελούν από τα πιο δεδομένα μοντέλα παλινδρόμησης, έχουν μελετηθεί διεξοδικά, είναι απλά, εύκολα κατανοητά και πληθώρα προβλημάτων μπορούν να μοντελοποιηθούν με αυτά, στις επόμενες ενότητες θα παρουσιαστούν τρόποι αναλυτικού προσδιορισμού τέτοιων μοντέλων και ειδικότερα τρόποι εκτίμησης των συντελεστών τους.

6.6.1 Εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης

Εάν έχει συνταχθεί ένα μοντέλο παλινδρόμησης, το επόμενο σημαντικό βήμα είναι η εκτίμηση των παραμέτρων του μοντέλου (δηλαδή των συντελεστών) από το διαθέσιμο σύνολο δεδομένων ή το σύνολο εκπαίδευσης. Όπως έχει ήδη τονιστεί, σε ένα μοντέλο παλινδρόμησης, οι άγνωστες τιμές είναι οι τιμές των συντελεστών β ενώ οι τιμές της ανεξάρτητης και των εξαρτημένων μεταβλητών είναι γνωστές από το σύνολο εκπαίδευσης.

Γενικά, η προσέγγιση που ακολουθείται για την εκτίμηση των συντελεστών β ενός μοντέλου παλινδρόμησης που έχει συνταχθεί, είναι εκείνη της ελαχιστοποίησης μίας συνάρτησης κόστους (Loss function), η οποία επιχειρεί να συλλάβει το πόσο απέχουν οι προβλεφθείσες τιμές της εξαρτημένης μεταβλητής από το μοντέλο παλινδρόμησης από τις παρατηρούμενες τιμές της εξαρτημένης μεταβλη-

τής εντός του συνόλου δεδομένων δηλαδή την ποσότητα που σε προηγούμενη ενότητα αναφέρθηκε ως «απόσταση» του δεδομένου από την προβλεφθείσα τιμή και γραμμή (σε περίπτωση γραμμικής) ή καμπύλη παλινδρόμησης. Στην ουσία με τη συνάρτηση κόστους, μετράται πόσο λάθος κάνει το μοντέλο παλινδρόμησης να προσδιορίσει τις τιμές της ανεξάρτητης μεταβλητής αν συγκριθεί με τις τιμές που πραγματικά λαμβάνει. Υπάρχουν πολλοί τρόποι αυτό να μετρηθεί.. Η διαφορά της παρατηρούμενης τιμής της ανεξάρτητης μεταβλητής που συναντάται εντός του συνόλου δεδομένων από την τιμή που προβλέπει το μοντέλο παλινδρόμησης για τις ίδιες τιμές των ανεξαρτήτων μεταβλητών *καλείται και κατάλοιπο (residual)* Συλλαμβάνει το *παρατηρούμενο σφάλμα*[6] και γι' αυτό χρησιμοποιείται συχνά και ο όρος *συνάρτηση σφάλματος* αντί για συνάρτηση κόστους .

Τέτοιες συναρτήσεις κόστους εμφανίζονται σε όλες τις περιπτώσεις εκτίμησης συντελεστών μοντέλων παλινδρόμησης. Το μοντέλο παλινδρόμησης με συντελεστές β, οι οποίοι ελαχιστοποιούν τα κατάλοιπα (ή αν νοηθούν γεωμετρικά, τις αποστάσεις) θεωρείται εκείνο το μοντέλο παλινδρόμησης που περιγράφει και ταιριάζει καλύτερα απ' όλα τα άλλα στα δοθέν σύνολο δεδομένων εκπαίδευσης. Υπάρχουν διάφορες συναρτήσεις κόστους καθώς επίσης και διάφοροι τρόποι ελαχιστοποίησής αυτών των συναρτήσεων που χρησιμοποιούνται στα πλαίσια της παλινδρόμησης. Η χρήση της κατάλληλης συνάρτησης κόστους και μεθόδου ελαχιστοποίησης εξαρτάται τόσο από τη μορφή του μοντέλου παλινδρόμησης όσο και και από τα χαρακτηριστικά των διαθέσιμων δεδομένων. Ομοίως, οι μέθοδοι που χρησιμοποιούνται για την εκτίμηση των συντελεστών ενός μοντέλου παλινδρόμησης εξαρτάται από το εάν το μοντέλο παλινδρόμησης είναι γραμμικό ή μη.

[6] Στο πλαίσιο της ανάλυσης παλινδρόμησης είναι σημαντικό να κατανοηθεί η διαφορά μεταξύ της έννοιας του διαταρακτικού όρου/σφάλματος ε που εμφανίζεται σε ένα μοντέλο παλινδρόμησης και της έννοιας των καταλοίπων, που εσφαλμένα θεωρούνται ως συνώνυμες. Ο διαταρακτικός όρος ε που εμφανίζεται στα μοντέλα παλινδρόμησης ορίζεται ως η διαφορά μεταξύ της παρατηρούμενης τιμής της εξαρτημένης μεταβλητής στο δείγμα και της πραγματικής τιμής της μεταβλητής στον πληθυσμό, η οποία συμπίπτει με την θεωρητική τιμή που προβλέπει το ιδανικό μοντέλο παλινδρόμησης, όπου έχουν προσδιοριστεί οι πραγματικοί συντελεστές. Ως κατάλοιπο ωρίζεται η διαφορά μεταξύ της παρατηρούμενης τιμής της εξαρτημένης μεταβλητής στο δείγμα που υπάρχει διαθέσιμο και της τιμής που προβλέπει το μοντέλο παλινδρόμησης όπου έχουν εκτιμηθεί οι συντελεστές. Οι διαταρακτικοί όροι δεν μπορούν να παρατηρηθούν ποτέ και είναι άγνωστοι, σε αντίθεση με τα κατάλοιπα που προκύπτουν από το σύνολο εκπαίδευσης.

Επιπλέον πρέπει να τονιστεί, ότι οι συντελεστές β που υπολογίζονται από το σύνολο δεδομένων είναι εκτιμήσεις των συντελεστών και όχι οι πραγματικές τιμές αυτών. Και τούτο γιατί το σύνολο δεδομένων στο οποίο βασίζονται οι μέθοδοι για τον υπολογισμό τους είναι ένα δείγμα του πληθυσμού και κατά συνέπεια διαφορετικά δείγματα δεδομένων μπορούν να οδηγήσουν σε διαφορετικές εκτιμήσεις των συντελεστών β για το ίδιο μοντέλο παλινδρόμησης. Δηλαδή επειδή ακριβώς το διαθέσιμο σύνολο δεδομένων είναι ένα δείγμα του πληθυσμού, είναι επί της ουσίας άγνωστη η πραγματική τιμή της εξαρτημένης μεταβλητής στον πληθυσμό. Η πραγματική τιμή της εξαρτημένης μεταβλητής θα ήταν γνωστή, εάν ήταν δυνατόν να καταγραφεί η τιμή της ανεξάρτητης μεταβλητής *για κάθε δυνατές τιμές των ανεξαρτήτων μεταβλητών* κάτι το οποίο, σε πραγματικές συνθήκες, είναι αδύνατον να συμβεί. Προκειμένου να δειχθεί ότι οι συντελεστές β που υπολογίζονται είναι εκτιμήσεις βάσει του διαθέσιμου συνόλου δεδομένων που είναι δείγμα και όχι οι πραγματικές τιμές που θα προκύπταν από τον πληθυσμό, θα συμβολίζονται με ένα καπελάκι (hat) $\hat{\beta}$. ΄

Στις επόμενες ενότητες θα παρουσιαστούν ειδικότερα μέθοδοι εκτίμησης συντελεστών γραμμικών μοντέλων παλινδρόμησης μιας κι αυτά χρησιμοποιούνται ευρέως για την μελέτη της συσχέτισης μεταξύ μεταβλητών.

6.6.2 Εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης: Η μέθοδος των ελαχίστων τετραγώνων.

Μία από τις πιο διεδομένες μεθόδους εκτίμησης συντελεστών β ενός γραμμικού μοντέλου παλινδρόμησης είναι η μέθοδος των ελαχίστων τετραγώνων των καταλοίπων (Ordinary Least Squares – OLS) η οποία είναι μία –από πολλές άλλες- μεθόδους εκτίμηση των συντελεστών. Αυτό που χαρακτηρίζει τη μέθοδο των ελαχίστων τετραγώνων είναι η μορφή της συνάρτησης κόστους που χρησιμοποεί και επιχειρεί να ελαχιστοποιήσει.

Όπως έχει τονιστεί, οι συντελεστές β του μοντέλου υπολογίζονται βάσει του συνόλου εκπαίδευσης με τρόπο ώστε το μοντέλο παλινδρόμησης να έχει το καλύτερο δυνατό ταίριασμα με τα δεδομένα αυτά. Προκειμένου να δηλωθεί ότι το γραμμικό μοντέλο παλινδρόμησης θα πρέπει να ταιριάξει στα υπάρχοντα δεδομένα, αν στα δεδομένα υπάρχουν n παρατηρήσεις, το γραμμικό μοντέλο γράφεται και με τη μορφή

$$Y_i = \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \cdots + \beta_0 + \varepsilon_i, i = 1..n$$

όπου Y_i η τιμή της εξαρτημένης, X_{1i}, X_{2i}, X_{3i}, … οι τιμές ανεξάρτητων μεταβλητών και ε_i ο διαταρακτικός όρος, σφάλμα *για την i-οστή παρατήρηση του συνόλου δεδομένων*. Ο όρος i που εμφανίζεται σηματοδοτεί τις παρατηρήσεις του συνόλου δεδομένων. Έτσι πρέπει να νοηθεί ότι αν υπάρχουν n παρατηρήσεις στο σύνολο εκπαίδευσης, τότε μπορούν να σχηματιστούν n εξισώσεις παλινδρόμησης όπως η παραπάνω, μία για κάθε παρατήρηση του συνόλου δεδομένων με τους συντελεστές β_i να είναι κοινοί για όλες τις εξισώσεις και οι άγνωστοι για τους οποίους αναζητείται η κατάλληλη τιμή τους.

Συνηθίζεται επίσης, ένα γραμμικό μοντέλο παλινδρόμησης να γράφεται με τη μορφή μήτρας των δεδομένων (matrix form). Αν το σύνολο των διαθέσιμων δεδομένων έχει n παρατηρήσεις και το μοντέλο γραμμικής παλινδρόμησης έχει k ανεξάρτητες μεταβλητές, ορίζοντας τις μήτρες:

$$y = \begin{bmatrix} Y1 \\ Y2 \\ Y3 \\ \dots \\ Yn \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ 1 & X_{13} & X_{23} & \dots & X_{\kappa 3} \\ .. & & & & \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix} \quad \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \dots \\ \beta_k \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{bmatrix}$$

όπου y μήτρα διαστάσεων nx1 των τιμών της εξαρτημένης μεταβλητής στο σύνολο εκπαίδευσης, X η nx(k+1) μήτρα των τιμών των ανεξαρτήτων μεταβλητών[7], β μήτρα διαστάσεων (k+1) x 1 των συντελεστών και e το nx1 διάνυσμα των καταλοίπων του μοντέλου[8], τότε η εξίσωση παλινδρόμησης μπορεί να γραφτεί και με τη μορφή:

$$y = X\beta + e$$

Στην παραπάνω μορφή μήτρας της εξίσωσης παλινδρόμησης, αν γίνουν οι πράξεις στο δεξί σκέλος της ισότητας, θα προκύψει διάνυσμα με τις n εξισώσεις της αλγεβρικής μορφής που προαναφέρθηκε για όλες τις παρατηρήσεις του συνόλου δεδομένων. Όπως θα φανεί παρακάτω, η έκφραση του προβλήματος με τη μορ-

[7] Μία τιμή X_{ij} στη μήτρα X των τιμών των ανεξάρτητων μεταβλητών σηματοδοτεί την τιμή της ανεξάρτητης μεταβλητής X_i της j παρατήρησης του συνόλου δεδομένων. Επίσης στη μήτρα X, η πρώτη στήλη με τις μονάδες (1) υπάρχει για να εμφανιστεί ο σταθερός όρος β_0 στις εξισώσεις παλινδρόμησης.

[8] Τα κατάλοιπα συμβολίζονται συνήθως με το λατινικό e σε αντίθεση με τον διαταρακτικό όρο που συμολίζεται με το ελληνικό ε.

φή μήτρας βολεύει καθότι μπορούν να χρησιμοποιηθούν οι πράξεις της γραμμικής άλγεβρας προκειμένου να επιλυθεί το πρόβλημα της βελτιστοποίησης (στα οποία ανήκει η μέθοδος των ελαχίστων τετραγώνων) και οδηγούν σε περιεκτικούς κλειστούς τύπους για τον υπολογισμό όλων των συντελεστών.

Ωστόσο, επειδή όπως έχει αναφερθεί, δεν είναι δυνατόν να βρεθούν οι πραγματικές τιμές των συντελεστών β και αυτές μπορούν μόνο να εκτιμηθούν από το διαθέσιμο σύνολο δεδομένων, που αποτελεί δείγμα του πληθυσμού, στη παραπάνω μορφή μήτρας του γραμμικού μοντέλου χρησιμοποιούνται διαφορετικοί όροι για να σηματοδοτήσουν ακριβώς αυτό το γεγονός:

$$y = \begin{bmatrix} Y1 \\ Y2 \\ Y3 \\ \dots \\ Yn \end{bmatrix} \quad X = \begin{bmatrix} 1 & X_{11} & X_{21} & \dots & X_{k1} \\ 1 & X_{12} & X_{22} & \dots & X_{k2} \\ 1 & X_{13} & X_{23} & \dots & X_{k3} \\ .. & & & & \\ 1 & X_{1n} & X_{2n} & \dots & X_{kn} \end{bmatrix} \quad \hat{\beta} = \begin{bmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \\ \hat{\beta}_2 \\ \dots \\ \hat{\beta}_k \end{bmatrix} \quad e = \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{bmatrix}$$

όπου $\hat{\beta}$ το διάνυσμα των εκτιμήσεων των άγνωστων συντελεστών και e το διάνυσμα των καταλοίπων. Τότε η μορφή εξίσωσης του γραμμικού μοντέλου παλινδρόμησης λαμβάνει τη μορφή

$$y = X\hat{\beta} + e$$

Η μορφή μήτρας του γραμμικού μοντέλου παλινδρόμησης προτιμάται καθότι μπορεί, με τη χρήση των ιδιοτήτων μητρών, να υπολογίσει άμεσα τις εκτιμήσεις όλων των συντελεστών β που εμφανίζονται στο μοντέλο.

Σύμφωνα με τη μέθοδο των ελαχίστων τετραγώνων, αναζητούνται εκείνες οι τιμές των συντελεστών β_i για το δοθέν μοντέλο, οι οποίοι θα ελαχιστοποιήσουν μία πολύ συγκεκριμένη συνάρτηση κόστους και δη το άθροισμα των τετραγώνων των καταλοίπων (Sum of Squared Errors - SSE), δηλαδή θα ελαχιστοποιούν την παρακάτω συνάρτηση κόστους

$$S = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} \left(Y_i - \hat{Y}_i\right)^2 = \sum_{i=1}^{n} \left(Y_i - \hat{\beta}_1 X_{1i} - \hat{\beta}_2 X_{2i} - \hat{\beta}_3 X_{3i} - \dots - \hat{\beta}_0\right)^2$$

όπου n το πλήθος των παρατηρήσεων στο σύνολο δεδομένων, Y_i η παρατήρηση της εξαρτημένης μεταβλητής στο σύνολο δεδομένων και $\hat{Y}_i$ η τιμή που υπολογίζει το μοντέλο παλινδρόμησης για τις τιμές των ανεξαρτήτων μεταβλητών στην πα-

ρατήρηση i του συνόλου δεδομένων. Η παραπάνω συνάρτηση αυτό που επιχειρεί να συλλάβει είναι η «απόσταση» των δεδομένων γύρω από τη γραμμή παλινδρόμησης ή όπως αλλιώς αναφέρεται ως *το σφάλμα* μεταξύ της παρατηρουμενης και προβλεφθείσας τιμής από το μοντέλο και να την ελαχιστοποιήσει, όπως συζητήθηκε σε προηγούμενη ενότητα. Αυτός ο τρόπος για την μέτρηση της «απόστασης» μεταξύ της παρατηρούμενης και προβλεφθείσας τιμής *είναι χαρακτηριστικό της μεθόδου των ελαχίστων τετραγώνων*. Στα πλαίσια της ανάλυσης γραμμικής παλινδρόμησης μπορούν να χρησιμοποιηθούν και άλλες συναρτήσεις κόστους, αλλά σε τέτοιες περιπτώσεις η μέθοδος δεν καλείται ελαχίστων τετραγώνων. Για παράδειγμα στα πλαίσια γραμμικών μοντέλων μπορούν να χρησιμοποιηθούν άλλες συναρτήσεις κόστους όπως, η απόλυτη τιμή της διαφοράς μεταξύ πραγματικής και προβλεφθείσας τιμής $S = \sum_{i=1}^{n} \left| Y_i - \hat{Y}_i \right|$ ή σταθμισμένο άθροισμα τετραγώνου της διαφοράς μεταξύ παρατηρούμενης και προβλεφθείσας τιμής. Το ποια συνάρτηση κόστους είναι η κατάλληλη εξαρτάται τόσο από το ερώτημα που επιθυμείται να απαντηθεί όσο και από την ποιότητα των διαθέσιμων δεδομένων. Η μέθοδος των ελαχίστων τετραγώνων παρουσιάζει το ενδιαφέρον να παρέχει κλειστούς τύπους για τον υπολογισμών των εκτιμήσεων των συντελεστών.

Παρακάτω παρουσιάζεται αναλυτικά ο τρόπος με τον οποίο προκύπτουν οι εκτιμώμενοι συντελεστές με τη μέθοδο των ελαχίστων τετραγώνων, ώστε να κατανοηθούν οι μηχανισμοί και πράξεις που χρησιμοποιούνται ώστε να μπορεί να γίνει αξιολόγηση της επίδοσης της μεθόδου.

Η αλγεβρική μορφή της συνάρτησης κόστους των ελαχίστων τετραγώνων, η οποία θα πρέπει να ελαχιστοποιηθεί, με τη μορφή μήτρας, μπορεί να γραφτεί ως

$$S = e^T e = [e_1, e_2, e_3, \dots, e_n] \begin{bmatrix} e_1 \\ e_2 \\ e_3 \\ \dots \\ e_n \end{bmatrix}$$

όπου e μήτρα των καταλοίπων και e^T ο ανάστροφος[9] της μήτρας καταλοίπων. Από την εξίσωση παλινδρόμησης μορφής μήτρας προκύπτει ότι η μήτρα καταλοίπων θα είναι ίση με

$$e = y - X\hat{\beta}$$

και κατά συνέπεια η συνάρτηση κόστους των ελαχίστων τετραγώνων με τη μορφή μήτρας και ως συνάρτηση των άγνωστων εκτιμήσεων των συντελεστών $\hat{\beta}$, S($\hat{\beta}$) λαμβάνει τη μορφή

$$S(\hat{\beta}) = e^T e = \left(y - X\hat{\beta}\right)^T \left(y - X\hat{\beta}\right)$$

όπου αν γίνουν οι πράξεις καταλήγει ως[10]

$$S(\hat{\beta}) = \left(y - X\hat{\beta}\right)^T \left(y - X\hat{\beta}\right) = y^T y - y^T X\hat{\beta} - \hat{\beta}^T X^T y + \hat{\beta}^T X^T X\hat{\beta}$$

Παρατηρώντας πιο προσεκτικά τους όρους της παραπάνω συνάρτησης $S(\hat{\beta})$, αυτή μπορεί να απλοποιηθεί λαμβάνοντας υπόψιν ότι οι όροι $y^T X\hat{\beta}$ και $\hat{\beta}^T X^T y$ δίνουν κάθε ένας ως αποτέλεσμα μία μήτρα διαστάσεων 1x1[11] που θα είναι συμμετρικές[12] και ίσες δηλαδή $\hat{\beta}^T X^T y = y^T X\hat{\beta}$ καθότι

$$\hat{\beta}^T X^T y = \left(X\hat{\beta}\right)^T y \overset{\lambda\acute{o}\gamma\omega\ συμμετρίας}{=} \left(\left(X\hat{\beta}\right)^T y\right)^T = y^T \left(\left(X\hat{\beta}\right)^T\right)^T = y^T X\hat{\beta}$$

Έτσι, αντικαθιστώντας στην $S(\hat{\beta})$ παραπάνω τον όρο $y^T X\hat{\beta}$ με $\hat{\beta}^T X^T y$, η σχέση λαμβάνει τη μορφή

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

Επειδή η $S(\hat{\beta})$ είναι η μορφή του τύπου των ελαχίστων τετραγώνων των καταλοίπων σε μορφή μήτρας με αγνώστους τους συντελεστές β η οποία θα πρέπει

---

[9] Ως ανάστροφος (transpose) μιας μήτρας mxn A ορίζεται μία άλλη μήτρα διστάσεων nxm που συμβολίζεται με A^T, η οποία προκύπτει από τη μήτρα A εάν οι γραμμές της μήτρας A γίνουν στήλες στη μήτρα A^T. Ή διαφορετικά το στοιχείο a_{ij}της μήτρας A θα βρεθεί στη θέση a'_{ji} στη μήτρα A^T έτσι ώστε $a_{ij} = a'_{ji}$. Έτσι η παράσταση $e^T e$ θα επιστρέψει μία μήτρα nx1 όπου κάθε όρος της είναι το τετράγωνο του αντίστοιχου κατάλοιπου.

[10] Αν A και B μήτρες τότε ισχύουν οι εξής ιδιότητες για τον ανάστροφο: $(A+B)^T = A^T + B^T$, $(AB)^T = B^T A^T$, $(A^T)^T = A$

[11] Αυτό μπορεί εύκολα να δειχθεί από τις διστάσεις των μητρών.

[12] Μία μήτρα A διαστάσεων nxn λέγεται συμμετρική εάν $a_{ij} = a_{ji}$ για κάθε i και j. Επίσης για συμμετρική μήτρα A ισχύει ότι $A^T = A$

να ελαχιστοποιηθεί, αυτό θα συμβεί εάν αν η πρώτη παράγωγος της $S(\hat{\beta})$ ως προς προς το διάνυσμα συντελεστών $\hat{\beta}$ είναι ίση με μηδέν δηλαδή

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = 0$$

Και επειδή

$$\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = -2X^{T}y + 2X^{T}X\,\hat{\beta}$$

προκύπτει ότι

$$-2X^{T}y + 2X^{T}X\,\hat{\beta} = 0 \iff 2X^{T}X\,\hat{\beta} = 2X^{T}y \iff$$

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

Ο παραπάνω τύπος είναι εκείνος που υπολογίζει αναλυτικά τις εκτιμήσεις των συντελεστών β ενός γραμμικού μοντέλου παλινδρόμησης που ελαχιστοποιούν το τετράγωνο του αθροίσματος καταλοίπων από το διαθέσιμο σύνολο εκπαίδευσης, βάσει μόνο των παρατηρούμενων τιμών των ανεξάρτητων μεταβλητών (μήτρα X) και των παρατηρούμενων τιμών της εξαρτημένης μεταβλητής (μήτρα y). Ο κλειστός αυτός τύπος εκτίμησης των συντελεστών είναι γνωστός ως η *κανονική εξίσωση (normal equation)* για τον υπολογισμό των συντελεστών γραμμικών μοντέλων παλινδρόμησης. Ο κλειστός αυτός τύπος επιστρέφει ένα διάνυσμα τιμών $\hat{\beta}$, όπου κάθε τιμή που εμφανίζεται αντιστοιχεί σε μία εκτίμησης του αντίστοιχου συντελεστή.

Άσκηση Αυτοαξιολόγησης 0.6

Γιατί πρέπει η πρώτη παράγωγος της συνάρτησης κόστους $\frac{\partial S(\hat{\beta})}{\partial \hat{\beta}} = 0$ να τεθεί ίση με το μηδέν προκειμένου να βρεθούν οι συντελεστές με τη μέθοδο των ελαχίστων τετραγώνων που ελαχιστοποιούν τη συνάρτηση κόστους;

Άσκηση Αυτοαξιολόγησης 0.7

Γιατί συνηθίζεται να εκφράζεται η κανονική εξίσωση εκτίμησης συντελεστών με τη μορφή μήτρας;

6.6.2.1 Εκτίμηση συντελεστών γραμμικού μοντέλου παλινδρόμησης: με τη μέθοδο των ελαχίστων τετραγώνων στο περιβάλλον της R.

Το περιβάλλον της R διαθέτει ειδική συνάρτηση η οποία επιτρέπει την εκτίμηση των συντελεστών β ενός γραμμικού μοντέλου παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων, η οποία είναι η συνάρτηση lm() (από τα αρχικά της φράσης "**l**inear **m**odel").

Η συνάρτηση lm() λαμβάνει ως ορίσματα τόσο το γραμμικό μοντέλο παλινδρόμησης με τη μορφή τύπου (formula) της R, το οποίο μπορεί να είναι απλό ή πολλαπλό, με ειδική σύνταξη για τον προσδιορισμό της εξαρτημένης και των ανεξαρτήτων μεταβλητών, το σύνολο εκπαίδευσης απ'όπου θα εκτιμηθούν οι συντελεστές και διάφορες άλλες παράμετροι που επιτρέπουν μεταξύ άλλων την ρύθμιση της συνάρτησης κόστους. Αν κληθεί επιτυχώς, η συνάρτηση lm() επιστρέφει ένα αντικείμενο το οποίο περιέχει μεταβλητές με τα αποτελέσματα της μεθόδου των ελαχίστων τετραγώνου μεταξύ των οποίων συγκαταλέγονται οι εκτιμήσεις των συντελεστών β καθώς και τα κατάλοιπα. Στον παρακάτω κώδικα R εκτελείται ένα γραμμικό μοντέλο παλινδρόμησης για το σύνολο δεδομένων του αρχείου HouseholdData.csv με στόχο να εκτιμηθούν οι συντελεστές του γραμμικού μοντέλου παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_0$$

```
foodConsumptionData<-read.csv("HouseholdData.csv ", sep=",", head-
er=T)

# Εκτίμηση συντελεστών του απλού γραμμικού μοντέλου παλινδρόμησης
- με μία ανεξάρτηση μεταβλητή

# Κατανάλωση τροφίμων = β1Εισόδημα + β0

# με τη χρήση της συνάρτησης lm() που κάνει χρήση της μεθόδου των
ελαχίστων τετραγώνων.

# Το πρώτο όρισμα είναι ένας ειδικός τύπος δεδομένων που καλείται
formula και αποτυπώνει το μοντέλο παλινδρόμησης. Εδώ δίνεται ως
όρισμα ο τύπος FoodExpenditure ~ Income που δηλώνει ότι η μεταβλητή
FoodExpenditure είναι η εξαρτημένη

# και Income η ανεξάρτητη.
```

```
# Τα αποτελέσματα της γραμμικής παλινδρόμησης αποθηκεύονται στη
μεταβλητή linear.regression.model

linear.regression.model <-  linear.regression.model<-
lm(FoodExpenditure ~ Income, data=foodConsumptionData)

#Εμφάνιση των συντελεστών του μοντέλου γραμμικής παλινδρόμησης,
εξετάζοντας

# τη μεταβλητή coefficients του αντικειμένου
linear.regression.model

print( linear.regression.model$coefficients )
```

Το αποτέλεσμα που εμφανίζει ο παραπάνω κώδικας R για το σύνολο δεδομένων, φαίνεται παρακάτω

```
(Intercept)       Income

2853.1014236    0.1112861
```

Οι τιμές που φαίνονται αναπαριστούν τις εκτιμήσεις συντελεστών β των ανεξαρτήτων μεταβλητών του γραμμικού μοντέλου παλινδρόμησης από το σύνολο εκπαίδευσης, οι οποίοι ελαχιστοποιούν τη διαφορά τετραγώνων των καταλοίπων. Αυτό σημαίνει ότι για το σύνολο δεδομένων, η εκτίμηση του συντελεστή $\hat{\beta}_0$ (σταθερός όρος που καλείται intercept) έχει λάβει τιμή 2853.1014[13] και η εκτίμηση του συντελεστή $\hat{\beta}_1$ της μεταβλητής εισόδημα έχει λάβει τιμή 0.1112. Έτσι, αφού για το μοντέλο γραμμικής παλινδρόμησης και το σύνολο δεδομένων έχουν εκτιμηθεί όλοι οι συντελεστές β, αυτό θα λάβει την ακόλουθη μορφή

$$Κατανάλωση\ τροφίμων = 0.1112861\ Εισόδημα + 2853.1014236$$

Με τον πλήρη προσδιορισμό του γραμμικού μοντέλου παλινδρόμησης, αυτό μπορεί να χρησιμοποιηθεί για τον σκοπό που έχει δημιουργηθεί και δη την εξήγηση της διακύμανσης ή την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής αν δοθεί ως είσοδο το εισόδημα.

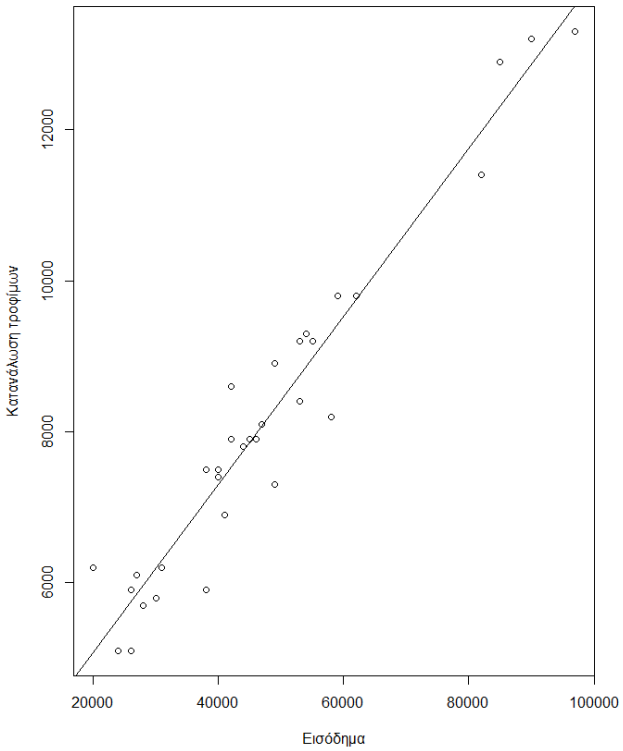
Συνηθίζεται επίσης, όποτε αυτό είναι εφικτό, να απεικονίζεται το προσδιορισμένο μοντέλο παλινδρόμησης στο διάγραμμα διασποράς των δεδομένων ώστε να

[13] Οι συντελεστές ενός μοντέλου παλινδρόμησης μπορούν να λάβουν και αρνητικές τιμές.

φανεί και οπτικά η τάση μεταξύ των μεταβλητών. Στην R, η απεικόνιση του γραμμικού μοντέλου παλινδρόμησης μπορεί να γίνει με την εντολή abline() η οποία προσθέτει μία γραμμή σε υπάρχον διάγραμμα.

```
options(scipen = 999)

foodConsumptionData<-read.csv("HouseholdData.csv ", sep=",", header=T)

# Εκτίμηση συντελεστών του απλού γραμμικού μοντέλου παλινδρόμησης - με μία ανεξάρτηση μεταβλητή

# Κατανάλωση τροφίμων = β1Εισόδημα + β0

# με τη χρήση της συνάρτησης lm() που βασίζεται στη μέθοδο των ελαχίστων τετραγώνων.

# Το πρώτο όρισμα FoodExpenditure ~ Income δηλώνει ότι η μεταβλητή FoodExpenditure του συνόλου

# δεδομένων είναι η εξαρτημένη μεταβλητή και Income η ανεξάρτηση.

# Τα αποτελέσματα της γραμμικής παλινδρόμησης αποθηκεύονται στο αντικείμενο linear.regression.model

linear.regression.model <-  linear.regression.model<-lm(FoodExpenditure ~ Income, data=foodConsumptionData)

#Εμφάνιση των συντελεστών του μοντέλου γραμμικής παλινδρόμησης, εξετάζοντας

# τη μεταβλητή coefficients του αντικειμένου linear.regression.model

print( linear.regression.model$coefficients )

# Δημιουργία διαγράμματος διασποράς των δεδομένων με καθορισμό των ετικετών αξόνων

plot(foodConsumptionData$Income, foodConsumptionData$FoodExpenditure, xlab="Εισόδημα", ylab="Κατανάλωση τροφίμων")

# Απεικόνιση πάνω στο διάγραμμα διασποράς της γραμμής γραμμικής παλινδρόμησης που έχει

# εκτιμηθεί

abline(linear.regression.model)
```

Η απεικόνιση του διαγράμματος διασποράς μαζί με την γραμμή παλινδρόμησης, που αποτυπώνει τις τιμές που προβλέπει το μοντέλο, φαίνεται στο παρακάτω σχήμα 6.6.



Εικόνα 0.6 *Απεικόνιση διαγράμματος διασποράς των μεταβλητών Κατανώλωση τροφίμων και Εισόδημα και της γραμμής παλινδρόμησης του μοντέλου Κατανάλωση τροφίμων = 0.1112861Εισόδημα +2853.1014.*

Άλλα δεδομένα που μπορούν να εξεταστούν από το αντικείμενο που επιστρέφεται από τη συνάρτηση lm() είναι:

- Το διάνυσμα καταλοίπων, μέσω της μεταβλητής $residuals του αντικειμένου αποτελέσματος της παλινδρόμησης, που για το παραπάνω παράδειγμα θα εμφανίσει – με τη μορφή διανύσματος στήλης και όχι γραμμής όπως συνήθως κάνει η R:

```
> cbind( linear.regression.model$residuals )
          [,1]
1    242.17449
2   -646.53943
3   -391.68374
4   -423.96728
5    437.45041
6    381.02002
7     50.31118
8   -391.68374
9    195.45549
10  -578.55976
11  -515.83059
12 -1107.69390
13  -269.11159
14  1121.17703
15   372.88333
16    16.45295
17  -347.85091
18   587.58201
19  -102.96982
20   153.46057
21   448.73648
22 -1006.11921
23   226.16433
24    95.45549
25    47.16179
26   -72.26098
27 -1181.97236
28    39.02510
29  -351.26352
30   593.88079
31  1072.88333
32  1072.88333
33   418.02764
34   331.15163
35  -515.83059
```

όπου η πρώτη τιμή (242.17449) είναι η διαφορά μεταξύ της πρώτης τιμής της εξαρτημένης μεταβλητής FoodExpenditure του αρχείου

HouseholdData.csv (6100) και της τιμής που προβλέπει το γραμμικό μοντέλο παλινδρόμησης για το εισόδημα της πρώτης γραμμής (0.1112861*27000 + 2853.1014236 = 5857.826) δηλαδή 6100 - 5857.826 = 242.174.

- Το διάνυσμα όλων των προβλεφθεισών τιμών της εξαρτημένης μεταβλητής του γραμμικού μοντέλου παλινδρόμησης, για κάθε τιμή της ανεξάρτητης τιμής του αρχείου δεδομένων, μέσω της μεταβλητής $fitted.values:

```
> cbind(linear.regression.model$fitted.values)
        [,1]
1   5857.826
2   5746.539
3   6191.684
4   5523.967
5   8862.550
6   9418.980
7   7749.689
8   6191.684
9   7304.545
10 11978.560
11  7415.831
12  9307.694
13  5969.112
14  5078.823
15  7527.117
16  8083.547
17 13647.851
18 12312.418
19  6302.970
```

```
20  5746.539
21  8751.264
22  8306.119
23  8973.836
24  7304.545
25  9752.838
26  7972.261
27  7081.972
28  7860.975
29  8751.264
30  8306.119
31  7527.117
32  7527.117
33  7081.972
34 12868.848
35  7415.831
```

Με τη συνάρτηση lm() μπορούν επίσης να εκτιμηθούν μοντέλα πολλαπλής γραμμικής παλινδρόμησης, δηλαδή γραμμικά μοντέλα παλινδρόμησης που έχουν πάνω από μία ανεξάρτητη μεταβλητή. Αυτό γίνεται χρησιμοποιώντας τον τελεστή + στον προσδιορισμό των ανεξάρτητων μεταβλητών. Για παράδειγμα, για την εκτίμηση των συντελεστών του πολλαπλού μοντέλου γραμμικής παλινδρόμησης

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού + \beta_0$$

γίνεται με τον ακόλουθο τρόπο στην R:

```
options(scipen = 999)

foodConsumptionData<-read.csv("HouseholdData.csv ", sep=",", head-
er=T)
```

```
# Εκτίμηση συντελεστών του πολλαπλού γραμμικού μοντέλου παλινδρόμη-
σης  - με δύο ανεξάρτητες μεταβλητες

# Κατανάλωση τροφίμων = β1Εισόδημα + β2Αριθμός ατόμων νοικοκυριού +
β0

# με τη χρήση της συνάρτησης lm() που βασίζεται στη μέθοδο των ελα-
χίστων τετραγώνων.

# Το πρώτο όρισμα FoodExpenditure ~ Income + FamilySize δηλώνει ότι
η μεταβλητή FoodExpenditure του # συνόλου  δεδομένων είναι η εξαρ-
τημένη μεταβλητή και ανεξάρτητες οι Income και FamilySize.

# Γενικά με τον τελεστή + μπορούν να προστεθούν ανεξάρτητες μετα-
βλητές στο γραμμικό μοντέλο

# παλινδρόμησης

# Τα αποτελέσματα της γραμμικής παλινδρόμησης αποθηκεύονται στο
αντικείμενο linear.regression.model

linear.regression.model <-  linear.regression.model<-
lm(FoodExpenditure ~ Income + FamilySize, data=foodConsumptionData)

#Εμφάνιση των συντελεστών του μοντέλου γραμμικής παλινδρόμησης,
εξετάζοντας

# τη μεταβλητή coefficients του αντικειμένου
linear.regression.model

print( linear.regression.model$coefficients )
```

Η εκτέλεση του παραπάνω κώδικα για το σύνολο δεδομένων, θα εμφανίσει ως αποτέλεσμα τις εκτιμήσεις των συντελεστών β για κάθε μία από τις ανεξάρτητες μεταβλητές του μοντέλου γραμμικής παλινδρόμησης

```
(Intercept)       Income   FamilySize

1182.1064194    0.1223607  325.7171067
```

με τις τιμές να αντιστοιχούν στους συντελεστές $\hat{\beta}_0 = 1182.1064194$, $\hat{\beta}_1 = 0.1223607$, $\hat{\beta}_2 = 325.7171067$. Έτσι, το μοντέλο πολλαπλής γραμμικής παλινδρόμησης θα λάβει τη μορφή:

$$Κατανάλωση\ τροφίμων = 0.1223607 * Εισόδημα + 325.7171067 * Αριθμός\ ατόμων\ νοικοκυριού + 1182.10644194$$

Άσκηση Αυτοαξιολόγησης 0.8

Τί τύπος δεδομένων (data type) της R πρέπει να είναι το όρισμα της συνάρτησης lm(), που αναπαριστά το γραμμικό μοντέλο παλινδρόμησης που πρέπει να εκτιμηθούν οι συντελεστές του;

Δραστηριότητα 0.2

Χρησιμοποιώντας το σύνολο δεδομένων του αρχείου HouseholdData.csv, συγγράψτε πρόγραμμα σε R που εκτιμά και εμφανίζει τους συντελεστές για το παρακάτω γραμμικό μοντέλο παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων:

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού + \beta_3 Αριθμός\ ατόμων\ νοικοκυριού^2 + \beta_0$$

όπου $Αριθμός\ ατόμων\ νοικοκυριού^2$ ένας όρος δευτέρου βαθμού. Ποιο ζήτημα προκύπτει και πως μπορεί αυτό να αντιμετωπιστεί;

6.6.3 Εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης: Η μέθοδος της Σταδιακής Καθόδου (Gradient Descent).

Μία διαφορετική μέθοδος εκτίμησης των συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης, είναι η μέθοδος της Σταδιακής Καθόδου[14] (Gradient Descent). Όπως και η μέθοδος των ελαχίστων τετραγώνων έτσι και η μέθοδος της Σταδιακής Καθόδου επιχειρεί να βρει εκείνους τους συντελεστές ενός γραμμικού μοντέλου παλινδρόμησης που ελαχιστοποιούν μία συγκεκριμένη συνάρτηση κόστους. Ωστόσο, ο τρόπος με τον οποίο βρίσκει τους συντελεστές η μέθοδος της Σταδιακής Καθόδου είναι εντελώς διαφορετικός: δεν παρέχει κλειστούς τύπους υπολογισμού των συντελεστών (όπως η κανονική εξίσωση στην μέθοδο των ελαχίστων

[14] Η μέθοδος εμφανίζεται και με το όνομα «Μέθοδος κλίσης» στην ελληνική βιβλιογραφία

τετραγώνων), αλλά βασίζεται σε επαναληπτική μέθοδο για την εύρεση των κατάλληλων τιμών τους.

Γενικά η μέθοδος της Σταδιακής Καθόδου είναι μία επαναληπτική διαδικασία βελτιστοποίησης συνάρτησης και στα ειδικότερα πλαίσια της ανάλυσης παλινδρόμησης, μια *επαναληπτική μέθοδος ελαχιστοποίησης μιας δοθείσης συνάρτησης κόστους*. Αναζητεί εκείνες τις τιμές που ελαχιστοποιούν την τιμή μιας συνάρτησης[15]. Η λέξη «επαναληπτική» χρησιμοποιείται για να τονιστεί ότι η μέθοδος αυτή δεν παρέχει κλειστούς τύπους αλλά επιχειρεί να «μαντέψει» τις κατάλληλες τιμές των συντελεστών που ελαχιστοποιούν την δοθείσα συνάρτηση, κάνοντας διαδοχικές εκτιμήσεις βάσει συγκεκριμένων κανόνων. Η μέθοδος αυτή προσπαθεί να «μάθει» με ποιον τρόπο πρέπει να αλλάξουν οι τιμές των συντελεστών ώστε να ελαχιστοποιηθεί η συνάρτηση κόστους, από το τρόπο με τον οποίο αλλάζει η συνάρτηση κόστους.

Η μέθοδος της Σταδιακής Καθόδου αποτελεί από τις πιο παλιές αλλά και πιο διαδεδομένες τεχνικές ελαχιστοποίησης συνάρτησης. Αποτέλεσε από τους θεμελιώδους αλγορίθμους της περιοχής της μηχανικής μάθησης που χρησιμοποιείται ακόμη και σήμερα σε διάφορες περιοχές όπως τα Νευρωνικά Δίκτυα.

Στα πλαίσια της εκτίμησης των συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης με τη μέθοδο της Σταδιακής Καθόδου, ορίζεται ως συνάρτηση κόστους η μέση τιμή του τετραγωνικού σφάλματος (Mean Squared Error) μεταξύ της προβλεφθείσας και πραγματικής τιμής της εξαρτημένης μεταβλητής. Η συνάρτησης κόστους, που συμβολίζεται $J(\theta_0, \theta_1, \theta_2, \dots, \theta_k)$, για την οποία η μέθοδος της Σταδιακής Καθόδου επιχειρεί να βρει τις τιμές των συντελεστών θ που ελαχιστοποιούν την τιμή της, για ένα πολλαπλό μοντέλο γραμμικής παλινδρόμησης λαμβάνει τη μορφή

$$J(\theta_0, \theta_1, \theta_2, \dots, \theta_k) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2$$

όπου θ_i ο άγνωστος συντελεστής i από συνολικά k+1 σε πλήθος (άγνωστους) συντελεστές του μοντέλου γραμμικής παλινδρόμησης, m το πλήθος των παρατηρήσεων στο σύνολο δεδομένων, $h_\theta\left(x^{(i)}\right)$ η τιμή που παράγει το γραμμικό μοντέλο

[15] Υπάρχει και η εκδοχή της επαναληπτικής μεθόδου που αναζητά τιμές που μεγιστοποιούν την τιμή μιας συνάρτησης που ονομάζεται μέθοδος Σταδιακής Ανόδου (Gradient Ascent).

παλινδρόμησης για τις τιμές των ανεξάρτητων μεταβλητών της παρατήρησης i στο σύνολο δεδομένων και $x^{(i)}$, $y^{(i)}$ οι τιμές των ανεξαρτήτων και της εξαρτημένης μεταβλητής της i παρατήρησης στο σύνολο δεδομένων[16]. Συνηθίζεται επίσης η συνάρτηση κόστους να συμβολίζεται πιο απλά ως J(θ), όπου θ το σύνολο όλων των συντελεστών του γραμμικού μοντέλου παλινδρόμησης.

Η συνάρτηση αυτή μοιάζει με τη συνάρτηση κόστους που χρησιμοποιείται και στη μέθοδο των ελαχίστων τετραγώνων (ειδικά το άθροισμα των τετραγώνων των καταλοίπων). Ωστόσο η συνάρτηση κόστους της μεθόδου Σταδιακής Καθόδου διαιρεί το άθροισμα τετραγώνων με τον όρο 2m. Ο όρος 2 που εμφανίζεται στον παρονομαστή εμφανίζεται κυρίως για να απλοποιηθεί ο τύπος ελαχιστοποίησης[17]. Ο όρος m, που σηματοδοτεί το πλήθος των παρατηρήσεων στο σύνολο δεδομένων υπάρχει για να εκφραστεί *ο μέσος όρος του τετραγωνικού σφάλματος* και επιπλέον για να μπορεί να αντιμετωπιστούν δύο σημαντικά ζητήματα: 1) επιτρέπει τη σύγκριση των τιμών της συνάρτησης κόστους, αν το μόνο που αλλάξει είναι το σύνολο δεδομένων εκπαίδευσης και 2) αντιμετωπίζει το πρόβλημα της υπερχείλισης αριθμών[18] μιας και στην Σταδιακή Κάθοδο χρησιμοποιούνται υποχρεωτικά υπολογιστικές μέθοδοι για την εκτίμηση των συντελεστών.

Άσκηση Αυτοαξιολόγησης 0.9

Γιατί η συνάρτηση κόστους της Σταδιακής Καθόδου επιτρέπει τη σύγκριση των τιμών της συνάρτησης κόστους για δύο διαφορετικά μοντέλα παλινδρόμησης με την ίδια εξαρτημένη μεταβλητή, αν το μόνο που αλλάζει είναι το σύνολο δεδομένων; Μπορεί να γίνει τέτοια σύγκριση των τιμών της συνάρτησης κόστους στην περίπτωση της μεθόδου των ελαχίστων τετραγώνων;

---

[16] Στο πλαίσιο της μεθόδου της Σταδιακής Καθόδου χρησιμοποιούνται άλλοι συμβολισμοί για τις ίδιες έννοιες που έχουν ήδη συναντηθεί στις προηγούμενες ενότητες. Έτσι χρησιμοποιείται το σύμβολο θ για τους συντελεστές αντί για β και ο συμβολισμός $h_\theta()$ για το γραμμικό μοντέλο παλινδρόμησης αντί για Υ. Αυτό γίνεται για να συμβαδίζει ο φορμαλισμός με αυτόν που συναντιέται στη υπάρχουσα βιβλιογραφία.

[17] Όπως θα φανεί παρακάτω, η πρώτη παράγωγος της συνάρτησης κόστους, λόγω του γραμμικού μοντέλου θα οδηγήσει στην απαλοιφή του 2.

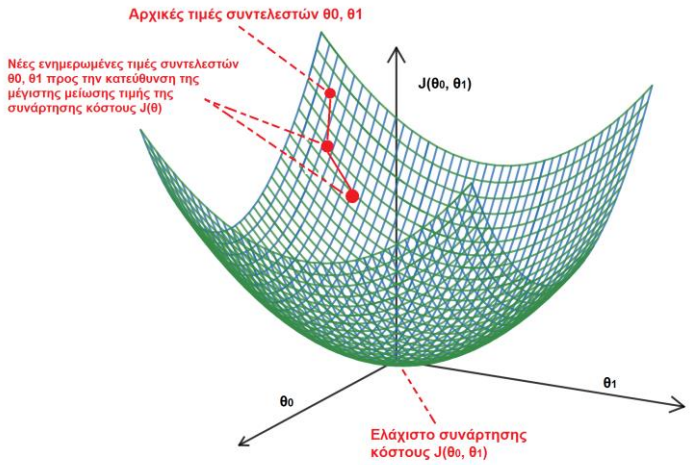
[18] Ο όρος υπερχείλιση χρησιμοποιείται για να περιγράψει την κατάσταση κατά την οποία, στα πλαίσια μιας πράξης προκύπτουν αριθμοί ,οι οποίοι είναι τόσο μεγάλοι ή μικροί, που δεν μπορούν να αναπαρασταθούν από το υπολογιστικό σύστημα.

6.6.3.1 Ο αλγόριθμος της Σταδιακής Καθόδου.

Ο αλγόριθμος της Σταδιακής Καθόδου βρίσκει εκείνους τους συντελεστές θ που ελαχιστοποιούν τη συνάρτηση κόστους J(θ) η οποία παρουσιάστηκε παραπάνω. Όπως αναφέρθηκε, ο αλγόριθμος το πετυχαίνει αυτό δίχως τη χρήση κλειστών τύπων αλλά με επαναληπτικές διαδικασίες εκτίμησης των συντελεστών.

Η βασική ιδέα του αλγορίθμου της Σταδιακής Καθόδου είναι να υπολογίζει την τιμή της συνάρτησης κόστους για κάποιες τιμές των συντελεστών θ του γραμμικού μοντέλου παλινδρόμησης (που αρχικά επιλέγονται τυχαία) και ακολούθως να αλλάζει τις τιμές των συντελεστών αυτών με τρόπο που θα οδηγήσει τη συνάρτηση σε ακόμη μικρότερη τιμή ή εκφρασμένα διαφορετικά η τιμή της συνάρτησης κόστους θα έχει τη μεγαλύτερη/πιο απότομη μείωση/κάθοδο της τιμής της. Ο αλγόριθμος δηλαδή επιλέγει κατευθύνσεις όπου εκτιμάται ότι θα οδηγήσουν σε ακόμη μικρότερες τιμές της συνάρτησης κόστους J(θ) και επαναληπτικά ενημερώνει τις τιμές των συντελεστών.

Οπτικό παράδειγμα για το πως λειτουργεί ο αλγόριθμος της Σταδιακής Καθόδου φαίνεται στο παρακάτω σχήμα 6.7. Το σχήμα απεικονίζει τη μορφή της συνάρτηση κόστους J(θ) για ένα απλό γραμμικό μοντέλο παλινδρόμησης της μορφής $h_\theta(x) = \theta_1 x + \theta_0$ για διάφορες τιμές των συντελεστών θ και για συγκεκριμένο σύνολο δεδομένων. Η συνάρτηση κόστους J(θ) που προκύπτει από ένα τέτοιο γραμμικό μοντέλο παλινδρόμησης θα είναι κυρτή (convex).



Εικόνα 0.7 *Οπτική αναπαράσταση του αλγορίθμου της Σταδιακής Καθόδου. Το διάγραμμα απεικονίζει τη γραφική παράσταση μιας συνάρτησης κόστους J(ϑ) ενός απλού γραμμικού μοντέλου παλινδρόμησης για συγκεκριμένο σύνολο δεδομένων και για όλες τις δυνατές τιμές ϑ0 και ϑ1. Η τιμή της συνάρτησης κόστους εμφανίζεται στον κάθετο άξονα. Ο αλγόριθμος της Σταδιακής Καθόδου ξεκινά από ένα τυχαίο σημείο (ϑ0, ϑ1) και ενημερώνει τις αρχικές τιμές αυτές συντελεστών καθέναν προς τη κατεύθυνση της πιο απότομης/μεγαλύτερης μείωσης της συνάρτησης κόστους J(ϑ).*

Ο αλγόριθμος αλλάζει την τιμή κάθε συντελεστή θ_i ανεξάρτητα από τους άλλους, με τρόπο ώστε να προκύψει μικρότερη τιμή της συνάρτησης κόστους J(θ). Η διαδικασία αυτή της ενημέρωσης των τιμών των συντελεστών θ εκτελείται επαναληπτικά και αποτελεί το κεντρικό στοιχείο του αλγορίθμου. Η κατεύθυνση προς την οποία θα κινηθεί ο αλγόριθμος και που καθορίζει τον τρόπο με τον οποίο θα ενημερωθούν οι συντελεστές καθορίζεται από την τιμή της πρώτης παραγώγου της συνάρτησης κόστους στο σημείο αυτό. Γενικά η τιμή της πρώτης παραγώγου σε κάποιο σημείο x καθορίζει αν η τιμή μιας συνάρτησης f(x) αυξάνεται, μειώνεται ή παραμένει η ίδια αν η τιμή x αυξηθεί κατά μία μικρή ποσότητα. Ειδικότερα αν η τιμή της πρώτης παραγώγου σε κάποιο σημείο x είναι θετική, αύξηση της τιμής x θα οδηγήσει σε αύξηση της τιμής της συνάρτησης f(x), αρνητική τιμή της πρώτης παραγώγου σημαίνει ότι αύξηση της τιμής x θα οδηγήσει σε μείωση της τιμής της συνάρτησης ενώ αν η τιμή της είναι 0 σημαίνει ότι αύξηση της τιμής της

x μπορεί να οδηγήσει είτε σε αύξηση είτε σε μείωση της τιμής της f(x), καθώς εμφανίζει στο σημείο αυτό ακρότατο.

Η συνάρτηση κόστους J(θ) είναι συνάρτηση πολλών μεταβλητών, με αγνώστους τους συντελεστές θ των οποίων αναζητείται η τιμή που την ελαχιστοποιούν. Οι ερμηνείες της τιμής της πρώτης παραγώγου που αναφέρθηκαν παραπάνω ισχύουν και για την περίπτωση συναρτήσεων πολλών μεταβλητών, αν αντί για τη πρώτη παράγωγο ληφθεί η μερική παράγωγος ως προς κάθε άγνωστο συντελεστή θ_i. Ειδικότερα τιμή της μερικής παραγώγου ως προς έναν συντελεστή θi στο σημείο $(\theta_0, \theta_1, \dots, \theta_k)$, αναφέρουν πως θα αλλάξει (αν θα αυξηθεί, μειωθεί ή παραμείνει σταθερή) η τιμή της συνάρτησης κόστους J(θ) αν η μεταβλητή θ_i αυξηθεί και οι υπόλοιπες μεταβλητές παραμείνουν σταθερές. Έτσι για παράδειγμα η τιμή της μερικής παραγώγου της συνάρτησης κόστους ως προς τη μεταβλητή θ_i $\frac{\partial}{\partial \theta_i} J(\theta)$ για κάποια συγκεκριμένη τιμή θ_i αναφέρει αν μία αύξηση της τιμής της θ_i θα οδηγήσει σε αύξηση, μείωση ή καμία μεταβολή στην τιμή της J(θ). Αν τέτοιες μερικές παράγωγοι ληφθούν για όλους τους συντελεστές θ_i της J(θ) και εκτιμηθεί η τιμή τους για κάποιο σημείο στο χώρο, αυτό θα δώσει την κατεύθυνση με την οποίο πρέπει να μεταβληθεί κάθε μεταβλητή θ_i (αν πρέπει να αυξηθεί ή να μειωθεί) ώστε να προκύψει μία ακόμη πιο μικρή τιμή της συνάρτησης κόστους J(θ).

Αν και η τιμή της μερικής παραγώγου ως προς έναν συντελεστή θ_i αναφέρει την απαραίτητη κατεύθυνση της μεταβολής της μεταβλητής θ_i για τη μείωση της συνάρτησης κόστους, δεν προσδιορίζει το πόσο πολύ πρέπει να μεταβληθεί η τιμή της θ_i για να επιτευχθεί μείωση της συνάρτησης κόστους. Για αυτόν τον λόγο, ο αλγόριθμος της σταδιακής καθόδου εισάγει και νέα μία παράμετρο, την παράμετρο α, που καλείται ρυθμός ή παράμετρος μάθησης (learning rate), η οποία αναφέρει κατά πόσο πρέπει να μεταβληθεί η τιμή της παραμέτρου προς τη κατεύθυνση της τιμής της μερικής παραγώγου. Ο ρυθμός μάθησης α είναι ένας θετικός αριθμός (> 0) που καθορίζεται εξ'αρχής και είναι σταθερός καθ'όλη τη διάρκεια εκτέλεσης του αλγορίθμου. Η επιλογή της κατάλληλης τιμής του είναι κρίσιμη πτυχή της επιτυχίας του αλγορίθμου για την ανεύρεση των συντελεστών

θ_i. Κάθε νέα τιμή καθενός από τους k+1 συντελεστές θ_j που εμφανίζεται στο μοντέλο παλινδρόμησης υπολογίζεται και ενημερώνεται από τον κάτωθι τύπο[19]

$$\theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

Έτσν, σε ψευδοκώδικα, η επαναληπτική διαδικασία ενημέρωσης των συντελεστών θ_i στον αλγόριθμο της Σταδιακής Καθόδου έχει ως εξής:

Αρχικοποίηση όλων των συντελεστών θ_j με τυχαίες τιμές
Ενόσω δεν πληρούνται τα κριτήρια τερματισμού {
Για όλους τους συντελεστές θ_j {
/ Υπολογισμός νέων τιμών συντελεστών προς τη κατεύθυνση μεγαλύτερης μείωσης της τιμής της J(ϑ). Οι νέες τιμές δεν ανατίθενται απ'ευθείας στους συντελεστές ϑ για να αποφευχθεί η χρήση των νέων τιμών στον υπολογισμό της τιμής της μερικής παραγώγου ως προς τις επόμενες μεταβλητές. Οι νέες τιμές αποθηκεύονται σε προσωρινές μεταβλητές new_ϑj και ακολούθως ανατίθενται στις μεταβλητές για τη χρήση τους στην επόμενη επανάληψη. */*

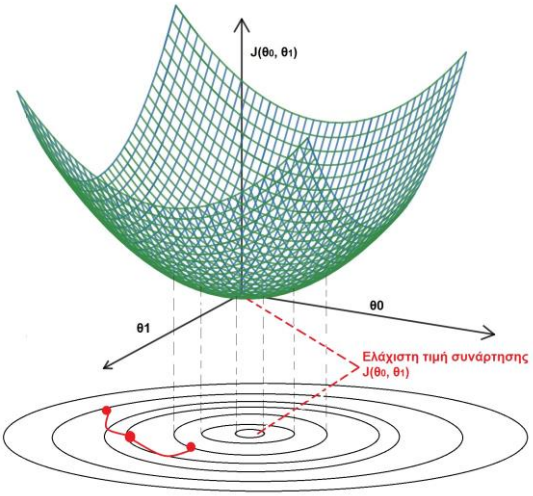
$new_{\theta_j} := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$

}
Για όλους τους συντελεστές θ_j
$\theta_j := new_{\theta_j}$ */* Ενημέρωση όλων των συντελεστών με τις νέες τιμές τους για την επόμενη επανάληψη */*

}

Συνηθίζεται επίσης, ο τρόπος με τον οποίο συγκλίνουν οι εκτιμήσεις των συντελεστών κατά την εκτέλεση του αλγορίθμου της Σταδιακής Καθόδου, ειδικά για απλά γραμμικά μοντέλα παλινδρόμησης, να απεικονίζεται με τη μορφή διαγράμματος ισοϋψών καμπύλων (contour plot). Ένα τέτοιο διάγραμμα απεικονίζει τις τιμές των συντελεστών για τις οποίες η συνάρτηση κόστους λαμβάνει την ίδια τιμή. Τέτοιες τιμές συντελεστών απεικονίζονται στο δισδιάστατο επίπεδο όπως φαίνεται από τους ομόκεντρους κύκλους στο σχήμα 6.8.

[19] Ο τύπος ενημέρωσης των συντελεστών θ στη βιβλιογραφία εμφανίζεται και με τη μορφή $\theta := \theta - \alpha \nabla J(\theta)$ όπου θ το διάνυσμα των συντελεστών και ∇ ο διανυσματικός διαφορικός τελεστής Ανάδελτα (gradient) που ορίζεται ως $\nabla J(\theta) = \left(\frac{\partial}{\partial \theta_0} J(\theta), \frac{\partial}{\partial \theta_1} J(\theta). \ldots, \frac{\partial}{\partial \theta_\kappa} J(\theta)\right)$. Επιπλέον, ο τελεστής := είναι ο τελεστής ανάθεσης τιμής, όχι ο τελεστής ελέγχου ισότητας.



Εικόνα 0.8 *Απεικόνιση της σύγκλισης των τιμών των συντελεστών ϑ0 και ϑ1 ενός απλού μοντέλου παλινδρόμησης με διάγραμμα ισοϋψών καμπύλων για ένα απλό μοντέλο παλινδρόμησης. Οι ισοϋψείς καμπύλες (οι ομόκεντροι κύκλοι) απεικονίζουν τις τιμές των συντελεστών ϑ0 και ϑ1 για τους οποίους η συνάρτηση κόστους J(ϑ) λαμβάβει την ίδια τιμή.*

Ο αλγόριθμος αρχικοποιεί τις τιμές των συντελεστών με τυχαίες τιμές των συντελεστών θ_j και ακολούθως ξεκινά την επαναληπτική διαδικασία ενημέρωσης των συντελεστών προς τη σωστή κατεύθυνση για τη βελτίωση της ελάχιστης τιμής της συνάρτησης κόστους J(θ). Είναι σημαντικό να τονιστεί ότι ο αλγόριθμος απαιτεί τον υπολογισμό όλων των νέων συντελεστών και η μαζική/ταυτόχρονη ενημέρωσή τους, ώστε η τιμή της μερικής παραγώγου να υπολογίζεται με τις κατάλληλες νέες εκτιμήσεις των συντελεστών. Η διαδικασία εκτελείται επαναληπτικά έως ότου ισχύσουν τα κριτήρια τερματισμού. Συνήθη κριτήρια τερματισμού του αλγορίθμου είναι τα ακόλουθα:

- Έχει συμπληρωθεί ένα προκαθορισμένο πλήθος επαναλήψεων. Η επαναληπτική διαδικασία του αλγορίθμου μπορεί να σταματήσει μετά από συγκεκριμένο αριθμό επαναλήψεων που καθορίζεται εξ’ αρχής. Το κριτήριο αυτό αποτελεί ίσως το πιο δημοφιλές στις υπάρχουσες υλοποιήσεις του αλγορίθμου καθότι δίνει ικανοποιητικές προσεγγίσεις των συντελεστών

σε σύντομο χρονικό διάστημα αν επιλεγεί σωστά το πλήθος επαναλήψεων και η παράμετρος μάθησης.

- Η βελτίωση/μείωση της συνάρτησης κόστους J(θ) είναι μικρότερη από ένα προκαθορισμένο όριο. Σε μία τέτοια περίπτωση, ο αλγόριθμος της Σταδιακής Καθόδου υπολογίζει, μετά από κάθε ενημέρωση των συντελεστών, την τιμή της συνάρτησης κόστους για τους νέους συντελεστές και τη συγκρίνει με την τιμή της J(θ) στην προηγούμενη επανάληψη. Αν η μείωση είναι μικρότερη από κάποιο αρχικά προκαθορισμένο όριο, που δίδεται παράμετρος, ο αλγόριθμος τερματίζει.
- Πρόωρη παύση. Το κριτήριο αυτό χρησιμοποιεί ένα ξεχωριστό σύνολο δεδομένων, το σύνολο επικύρωσης ή ελέγχου που δεν αποτελεί τμήμα του συνόλου εκπαίδευσης, προκειμένου να υπολογίσει τη συνάρτηση κόστους γι' αυτό το σύνολο δεδομένων σε κάθε επανάληψη του αλγορίθμου. Οι τιμές της συνάρτησης κόστους για το σύνολο επικύρωσης για δύο διαδοχικές επαναλήψεις συγκρίνονται και η διαδικασία τερματίζει εάν η συνάρτηση κόστους για το σύνολο επικύρωσης αυξάνεται. Το κριτήριο αυτό χρησιμοποιείται κυρίως για να αντιμετωπιστεί το πρόβλημα της υπερπροσαρμογής (overfitting) που θα αναλυθεί σε άλλη ενότητα.

Η φύση της προσέγγισης της μεθόδου της Σταδιακής Καθόδου είναι τέτοια που δεν εγγυάται πάντα την ίδια λύση (δηλαδή την εύρεση των ίδιων συντελεστών) αν εκτελεστεί για το ίδιο γραμμικό μοντέλο παλινδρόμησης και για το ίδιο σύνολο δεδομένων. Παράγοντες όπως οι αρχικές τιμές των μεταβλητών καθώς, η τιμή της παραμέτρου μάθησης α καθώς και τα κριτήρια τερματισμού (όπως το πλήθος των επαναλήψεων) επηρεάζουν τις τελικές εκτιμήσεις συντελεστών που θα προκύψουν από τη μέθοδο αυτή.

Άσκηση Αυτοαξιολόγησης 0.10

Εξηγήστε γιατί στον τύπο ενημέρωσης τιμής του συντελεστή, θα πρέπει ο όρος $\alpha \frac{\partial}{\partial \theta_j} J(\theta)$ να αφαιρεθεί από την τρέχουσα τιμή του συντελεστή θ_j και όχι να προστεθεί στον όρο θ_j .

6.6.3.2 Τύπος υπολογισμού συντελεστών για πολλαπλό γραμμικό μοντέλο παλινδρόμησης με τη μέθοδο της Σταδιακής Καθόδου

Σε περίπτωση μοντέλου πολλαπλής γραμμικής παλινδρόμησης με k ανεξάρτητες μεταβλητές της μορφής

$$h_\theta(X) = \theta_0 + \theta_1 X_1 + \theta_2 X_2 + \cdots + \theta_k X_k$$

όπου θ_j οι άγνωστοι συντελεστές που πρέπει να εκτιμηθούν και X_j η τιμή της j ανεξάρτητης μεταβλητής στο σύνολο δεδομένων, οι μερικές παράγωγοι της συνάρτησης κόστους J(θ) ως προς τους συντελεστές θ_j μπορούν να υπολογιστούν, με αποτέλεσμα οι τύποι ενημέρωσης του σταθερού όρου θ_0 και για κάθε έναν από τους συντελεστές θ_j να λαμβάνουν τη κλειστή μορφή

$$\theta_0 := \theta_0 - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)$$

$$\theta_j := \theta_j - \alpha \frac{1}{m}\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) {x_j}^{(i)}$$

όπου $x^{(i)}, y^{(i)}$ οι τιμές των ανεξαρτήτων μεταβλητών και η παρατηρηθείσα τιμή της εξαρτημένης μεταβλητής στην παρατήρηση i του συνόλου δεδομένων και ${x_j}^{(i)}$ η τιμή της ανεξάρτητης μεταβλητής j στην i παρατήρηση του συνόλου δεδομένων και m το πλήθος των παρατηρήσεων στο σύνολο δεδομένων.

Με μορφή μήτρας, οι παραπάνω τύποι ενημέρωσης όλων των συντελεστών θ_j ενός πολλαπλού γραμμικού μοντέλου παλινδρόμησης με k ανεξάρτητες μεταβλητές, λαμβάνουν την ακόλουθη μορφή

$$\theta := \theta - \alpha \frac{1}{m} X^T (X\theta - Y)$$

όπου θ η διαστάσεων (k+1) x 1 μήτρα των συντελεστών, X η m x (k+1) μήτρα των τιμών των ανεξάρτητων τιμών με την πρώτη στήλη της να περιέχει άσσους (1) και Y η mx1 μήτρα των τιμών της εξαρτημένης μεταβλητής στο σύνολο δεδομένων.

6.6.3.3 Ο ρόλος της παραμέτρου μάθησης α

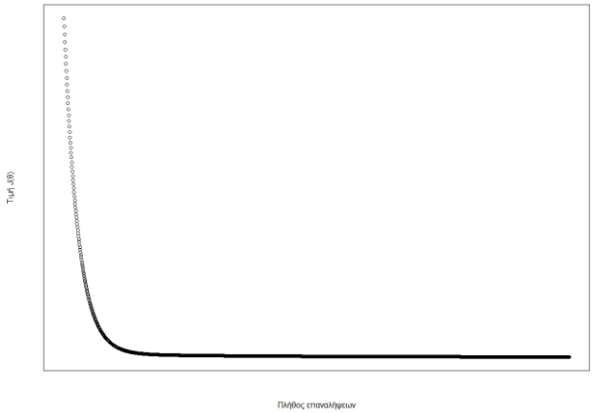
Η παράμετρος μάθησης α που εμφανίζεται στον τύπο ενημέρωσης των τιμών των συντελεστών καθορίζει το πόσο θα μεταβληθούν οι συντελεστές θ ώστε να προκύψει ακόμη μικρότερη τιμή της συνάρτησης κόστους. Η παράμετρος α μπορεί να θεωρηθεί ως το βήμα με τον οποίο αλλάζουν οι υποψήφιες τιμές των συντελεστών θ, είναι σταθερά και καθορίζεται εξαρχής.

Η ακριβής τιμή της παραμέτρου α αποτελεί κρίσιμο στοιχείο του αλγορίθμου καθότι μπορεί να επηρεάσει τόσο την ταχύτητα εύρεσης των κατάλληλων τιμών των συντελεστών θ όσο και το εαν θα συγκλίνει ή θα αποκλίνει η συνάρτηση κόστους J(θ) από την ελάχιστη τιμή. Γενικά, η εύρεση της κατάλληλης τιμής της παραμέτρου μάθησης α δεν είναι εύκολη διαδικασία μιας και δεν υπάρχουν τυποποιημένες προσεγγίσεις και απαιτεί συνήθως αρκετές δοκιμές και ελέγχους.

Εάν η παράμετρος μάθησης έχει την κατάλληλη τιμή, τότε η τιμή της συνάρτησης κόστους J(θ) θα μειώνεται με κάθε επανάληψη υπολογισμού των νέων συντελεστών θ με ικανοποιητικό ρυθμό. Εμπειρικά, δύο είναι τα κριτήρια για την αξιολόγηση μιας τιμής μάθησης α: 1) σε κάθε επανάληψη του αλγορίθμου της Σταδιακής Καθόδου να μειώνεται η τιμή της συνάρτησης κόστους J(θ) και 2) η ταχύτητα με την οποία επιτυγχάνει τη μείωση της συνάρτησης κόστους.

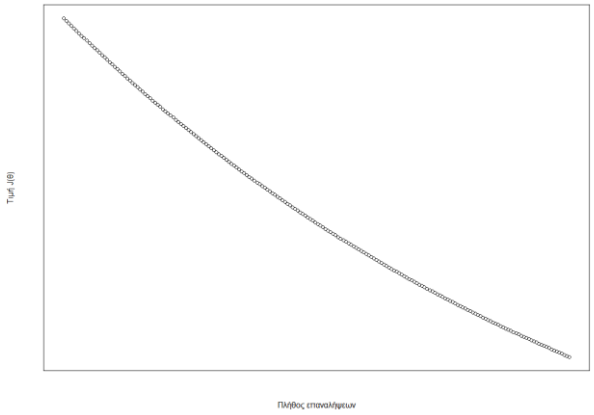
Ένας τρόπος για να εξεταστούν τα παραπάνω κριτήρια είναι να απεικονιστεί η τιμή της συνάρτησης κόστους ως προς το πλήθος επαναλήψεων του αλγορίθμου προκειμένου να εξεταστεί εάν η τιμή της παραμέτρου μάθησης είναι η κατάλληλη. Η μορφή της απεικόνισης θα επιτρέψει να αξιολογηθεί η τρέχουσα τιμή της παραμέτρου μάθησης α.

Έτσι, μία κατάλληλη τιμή της παραμέτρου μάθησης α θα οδηγήσει η απεικόνιση της τιμής της J(θ) συναρτήσει του πλήθους επαναλήψεων να έχει τη μορφή της εικόνας 6.9.



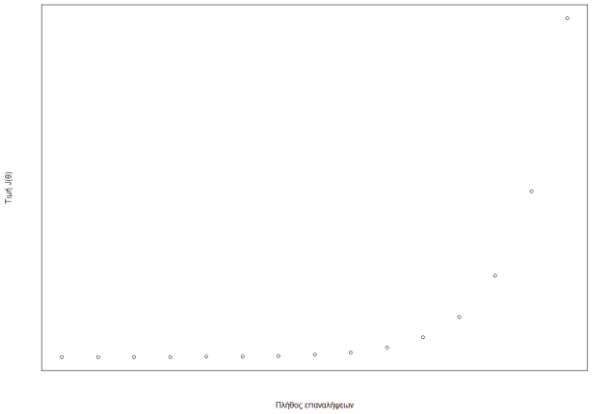
Εικόνα 0.9 *Μορφή της γραφικής παράστασης της συνάρτησης κόστους J(θ) ως προς το πλήθος επαναλήψεων με μία κατάλληλη τιμή της παραμέτρου μάθησης α.*

Εαν η τιμή της παραμέτρου μάθησης α είναι πολύ μικρή, τότε η συνάρτηση κόστους μπορεί να μειώνεται με κάθε επανάληψη, αλλά η μείωση της τιμής της θα είναι αργή με αποτέλεσμα ο αλγόριθμος της Σταδιακής Καθόδου να έχει πολύ χαμηλή ταχύτητα σύγκλισης προς την ελάχιστη τιμή και να απαιτείται περισσότερος χρόνος (ή πλήθος επαναλήψεων) προκειμένου αυτή να βρεθεί. Εαν επιλεγεί μια πολύ μικρή τιμή παραμέτρου μάθησης η απεικόνιση της τιμής J(θ) ως προς τον αριθμό επαναλήψεων θα έχει την χαρακτηριστική μορφή της εικόνας 6.10



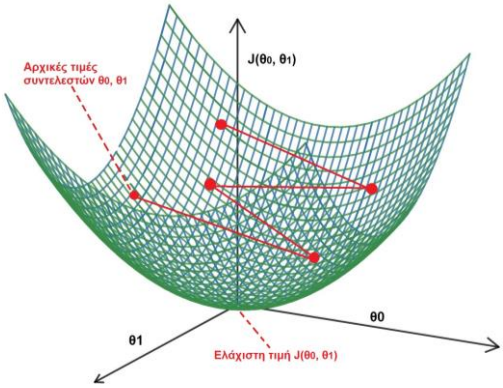
Εικόνα 0.10 *Μορφή της γραφικής παράστασης της συνάρτησης κόστους J(θ) ως προς το πλήθος επαναλήψεων με μία πολύ μικρή τιμή της παραμέτρου μάθησης α.*

Εαν επιλεγεί μία πολύ μεγάλη τιμή για την παράμετρο μάθησης α, τότε η συνάρτηση κόστους μπορεί να μην συγκλίνει προς την ελάχιστη τιμή της αλλά να αποκλίνει απ'αυτήν. Αυτό σημαίνει ότι η ελάχιστη τιμή της συνάρτησης κόστους δεν θα βρεθεί ποτέ όσες επαναλήψεις και αν κάνει ο αλγόριθμος της Σταδιακής Καθόδου. Σε μία τέτοια περίπτωση, η απεικόνιση της τιμής της συνάρτησης κόστους J(θ) ως προς το πλήθος επαναλήψεων θα αυξάνει καθώς αυξάνεται το πλήθος επαναλήψεων και θα έχει τη χαρακτηριστική μορφή του σχήματος 6.11 .



Εικόνα 0.11 *Μορφή της γραφικής παράστασης της συνάρτησης κόστους J(θ) ως προς το πλήθος επαναλήψεων με μία πολύ μεγάλη τιμή της παραμέτρου μάθησης α.*

Μία γεωμετρική εξήγηση της συμπεριφοράς της συνάρτησης κόστους για μεγάλες τιμές της παραμέτρου α φαίνεται στο σχήμα 6.12. Μεγάλες τιμές της παραμέτρου μάθησης α θα οδηγήσουν τους συντελεστές να αυξάνουν με πολύ μεγάλο βήμα κάτι που μπορεί να οδηγήσει στην υπερπήδηση των κατάλληλων τιμών τους, με αποτέλεσμα η συνάρτηση κόστους να αποκλίνει από -και όχι να συγκλίνει προς- την ελάχιστη τιμή της.



Εικόνα 0.12 *Απεικόνιση των τιμών των συντελεστών ϑ που προκύπτουν, όταν επιλεγεί πολύ μεγάλη τιμή της παραμέτρου μάθησης α. Σε μία τέτοια περίπτωση, μπορεί να υπάρξει υπερπήδηση των κατάλληλων τιμών των συντελεστών ϑ κατά τη φάση της ενημέρωσης με αποτέλεσμα η τιμή της συνάρτησης κόστους να αποκλίνει από την ελάχιστη τιμή της με κάθε επανάληψη.*

Χαρακτηριστικές τιμές της παραμέτρου μάθησης α, που συνήθως χρησιμοποιούνται στον αλγόριθμο της Σταδιακής Καθόδου, είναι 0.0001, 0.001, 001, 0.1, 1 ή και ακόμη μικρότερες τιμές. Το πλήθος επαναλήψεων μπορεί να κυμανθεί από 50 έως 50000 ή ακόμη και μεγαλύτερες τιμές, ανάλογα με τον ρυθμό σύγκλισης του αλγορίθμου.

Γενικά, η μεθοδολογία που ακολουθείται για τον προσδιορισμό τόσο της κατάλληλης τιμής της παραμέτρου μάθησης α όσο και του κατάλληλου πλήθους επαναλήψεων είναι να δοκιμάζεται μία τιμή α και ένα πλήθος επαναλήψεων n και να απεικονίζεται η τιμή της συνάρτησης κόστους J(θ) συναρτήσει του πλήθους επαναλήψεων. Αν η απεικόνιση μοιάζει με αυτή του σχήματος 6.9 τότε μπορεί να θεωρηθεί ότι οι τιμές αυτές είναι ικανοποιητικές. Αν ωστόσο η απεικόνιση αυτή μοιάζει με εκείνη του σχήματος 6.10, τότε μπορεί να αυξηθεί η παράμετρος μάθησης α π.χ. από 0.01 σε 0.03 ή και το πλήθος επαναλήψεων και να εκτελεστεί και πάλι ο αλγόριθμος της Σταδιακής Καθόδου. Όπως έχει αναφερθεί, η διαδικασία εύρεσης των κατάλληλων τιμών της παραμέτρου μάθησης και του πλήθους των επαναλήψεων απαιτεί μία ευρετική διαδικασία με αρκετές δοκιμές και αξιολόγηση των αποτελεσμάτων.

Άσκηση Αυτοαξιολόγησης 0.11

Με ποιον τρόπο θα ελέγξετε αν η παράμετρος μάθησης α έχει την κατάλληλη τιμή;

Άσκηση Αυτοαξιολόγησης 0.12

Παρόλο που η παράμετρος μάθησης α λαμβάνει τιμές μεγαλύτερες από το 0, εξηγείστε πως θα επηρεαστεί ο αλγόριθμος της Σταδιακής Καθόδου σε κάθε μία από τις παρακάτω περιπτώσεις:

i. Η παράμετρος μάθησης α λάβει τιμή ίση με το 0

ii. Η παράμετρος μάθησης α λάβει τιμή αρνητική (< 0).

6.6.3.4 Η μέθοδος της Σταδιακής Καθόδου στο περιβάλλον της R

Στο περιβάλλον της R υπάρχουν βιβλιοθήκες που παρέχουν τον αλγόριθμο της Σταδιακής Καθόδου. Ωστόσο, για να δειχθεί ότι η υλοποίησή του συγκεκριμένου αλγορίθμου δεν είναι δύσκολη, δίνεται παρακάτω μία υλοποίηση του αλγορίθμου της Σταδιακής Καθόδου που εκτιμά τους συντελεστές ενός πολλαπλού γραμμικού μοντέλου παλινδρόμησης χρησιμοποιώντας τους τύπους με τη μορφή μήτρας.

```
# calculateCost
# Υπολογίζει και επιστρέφει την τιμή της συνάρτησης κόστους J(θ)
# για τις τιμές των ανεξάρτητων και εξαρτημένων μεταβλητών του συνόλου δεδομένων
# και των τρεχουσών τιμών των συντελεστών θ.
# Ορίσματα
# X: η n x (k+1) μήτρα των τιμών των ανεξάρτητων μεταβλητών
# y: η (k+1) x 1 μήτρα των τιμών της εξαρτημένης μεταβλητής
# theta: η (k+1) x 1 μήτρα των τρεχουσών τιμών των συντελεστών.
# Επιστρέφει πραγματικό αριθμό (βαθμωτό μέγεθος) που σηματοδοτεί
# την τρέχουσα τιμή της συνάρτησης κόστους
calculateCost<-function(X, y, theta){
```

```
  # Πλήθος παρατηρήσεων
  m <- length(y)
  return( sum((X%*%theta- y)^2) / (2*m) )
  } # calculateCost
# gradientDescent
# Υλοποίηση του αλγορίθμου της Σταδιακής Καθόδου, που εκτιμά του συντελεστές θ ενός πολλαπλού
# γραμμικού μοντέλου παλινδρόμησης. Η συνάρτηση κάνει χρήση της μορφής μήτρας του τύπου ενημέρωσης
# των συντελεστών.
# Ορίσματα συνάρτησης
# X : μήτρα m x (k+1) των τιμών των ανεξαρτήτων μεταβλητών. m=πλήθος παρατηρήσεων στο σύνολο δεδομένων και
#     k=πλήθος ανεξάρτητων μεταβλητών. Η πρώτη στήλη της μήτρας X πρέπει να περιέχει μόνο άσσους (1) για
#     την αναπαράσταση του σταθερού όρου
# y : mx1 μήτρα των τιμών της εξαρτημένης μεταβλητής
# theta: (k+1) x 1 διάνυσμα συντελεστών, με τυχαίες τιμές αρχικοποίησης. theta[0] είναι ο συντελεστής του
#         σταθερού όρου, theta[1] ο συντελεστής της πρώτης ανεξάρτησης μεταβλητής κοκ.
# alpha: η παράμετρος μάθησης. Λαμβάνει την τιμή 0.01 αν δεν δοθεί τιμή στο όρισμα αυτό
# numIters: πλήθος επαναλήψεων. Λαμβάνει την τιμή 90 αν δεν δοθεί τιμή στο όρισμα αυτό
# Η συνάρτηση επιστρέφει λίστα με τις εξής μεταβλητές:
# coefficients: διάνυσμα των συντελεστών που έχουν εκτιμηθεί
# costs: διάνυσμα με την τιμή της συνάρτησης κόστους σε κάθε επανάληψη του αλγορίθμου.
gradientDescent<-function(X, y, theta, alpha=0.01, numIters=90){
  # Πλήθος παρατηρήσεων
  m <- length(y)
```

```
  # Διάνυσμα (της R) όπου θα αποθηκευτεί κάθε τιμή
  # της συνάρτησης κόστους J(θ) που υπολογίζεται
  # σε κάθε επανάληψη του αλγορίθμου της Σταδιακής Καθόδου.
  # Έτσι η θέση costHistory[1] θα έχει την τιμή της συνάρτησης κόστους στην πρώτη επανάληψη,
  # η θέση costHistory[2] την τιμή της συνάρτησης κόστους στη δεύτερη επανάληψη κοκ.
  #
  # Αυτό γίνεται ώστε να μπορέσει να απεικονιστεί σε γράφημα ο τρόπος
  # με τον οποίο μειώνεται η τιμή της συνάρτησης κόστους και να αξιολογηθούν τα ορίσματα.
  # Το διάνυσμα αρχικοποιείται με 0
  costHistory <- rep(0, numIters)
  # Έναρξη επαναληπτικής διαδικασίας. Θα γίνουν numIters σε πλήθος επαναλήψεις
  # για την εκτίμηση των συντελεστών θ.
  for(i in 1:numIters){
    # Ταυτόχρονος υπολογισμός (και ανάθεση) όλων των νέων τιμών των συντελεστών θ με
    # τη χρήση του τύπου ενημέρωσης με τη μορφή μήτρας.
    # alpha: η παράμετρος μάθησης
    # theta: διάνυσμα των τρεχουσών συντελεστών θ
    # m : το πλήθος των παρατηρήσεων
    # X : η μήτρα των τιμών των ανεξάρτητων συντελεστών
    # t() : ο τελεστής αναστροφής μήτρας. t(X) θα επιστρέψει την ανάστροφη μήτρα της X
    # y: οι τιμές της εξαρτημένης μεταβλητής
    # %*%: η πράξη πολλαπλασιαμού μητρών


    # Ο τύπος ενημέρωσης των συντελεστών με τη μορφή μήτρας
```

```
    theta <- theta - alpha*(1/m)*(t(X)%*%(X%*%theta - y))

    # Κλήση της συνάρτησης calculateCost για τον υπολογισμός της
συνάρτησης κόστους

    # για τους συντελεστές που προέκυψαν παραπάνω.

    # Η τιμή της συνάρτησης κόστους αποθηκεύεται στο διάνυσμα
costHistory, στην

    # κατάλληλη θέση (i σηματοδοτεί την επανάληψη όπου προέκυψε το
συγκεκριμένο κόστος).

    costHistory[i]  <- calculateCost(X, y, theta)


  } # for

    # Σε αυτό το σημείο έχει ολοκληρωθεί το πλήθος επαναλήψεων και
η εκτίμηση

  # των συντελεστών έχει ολοκληρωθεί.

  # Η συνάρτηση επιστρέφει το διάνυσμα των συντελεστών και το διά-
νυσμα με τις τιμές της

  # συνάρτησης κόστους. Τα στοιχεία αυτά επιστρέφονται ως μέλη μιας
λίστας.

  # Έτσι δημιουργείται μία λίστα όπου ονοματίζονται τα μέλη της:
coefficients είναι το διάνυσμα των συντελεστών

  # και costHistory το διάνυσμα τιμών της συνάρτησης κόστους J(θ)
που  προέκυψαν σε κάθε επανάληψη.

  gdResults<-list("coefficients"=theta, "costs"=costHistory)

  return(gdResults)

} # gradientDescent
```

Ο παραπάνω κώδικας R υλοποιεί τον αλγόριθμο της Σταδιακής Καθόδου όπως έχει περιγραφεί στις προηγούμενες ενότητες του οποίου το πλήρες όνομα είναι μέθοδος Σταδιακής Καθόδου Δέσμης (Batch Gradient Descent). Και τούτο γιατί κάνει χρήση ολόκληρου του συνόλου δεδομένων εκπαίδευσης (θεωρώντας το σύνολο δεδομένων ως μία δέσμη) για την εκτίμηση των νέων τιμών των συντελεστών σε κάθε επανάληψη του αλγορίθμου. Υπάρχουν άλλες εκδοχές του αλγορίθμου, οι οποίες χειρίζονται διαφορετικά το σύνολο δεδομένων εκπαίδευσης

προκειμένου να αντιμετωπίσουν ζητήματα σχετικά με το μέγεθός του και οι οποίοι θα παρουσιαστούν παρακάτω.

6.6.3.5 Χρήση της μεθόδου Σταδιακής Καθόδου για την εκτίμηση συντελεστών

Στο παρακάτω παράδειγμα γίνεται μία εφαρμογή της μεθόδου της Σταδιακής Καθόδου για την εκτίμηση των συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης με τη χρήση των συναρτήσεων που ορίστηκαν προηγουμένως. Θα χρησιμοποιηθούν τα δεδομένα του αρχείο IcecreamRevenues.csv τα οποία έχουν παρατηρήσεις σχετικά με την μέση θερμοκρασία ανα ημέρα (μεταβλητή Temperature) και τα έσοδα από πωλήσεις παγωτών (μεταβλητή Revenue). Τα περιεχόμενα του αρχείου φαίνονται και στο παράρτημα Β. Στόχος είναι να μελετηθεί η σχέση που υπάρχει μεταξύ της θερμοκρασίας και των εσόδων και ειδικότερα πως τα έσοδα επηρεάζονται από την ημερήσια μέση θερμοκρασία. Έτσι το γραμμικό μοντέλο παλινδρόμησης που θα εκτιμηθεί θα έχει τη μορφή

$$Έσοδα = \theta_0 + \theta_1 Θερμοκρασία$$

Η εκτίμηση των συντελεστών του παραπάνω γραμμικού μοντέλου παλινδρόμησης θ θα γίνει με τη μέθοδο της Σταδιακής Καθόδου. Θεωρώντας ότι έχουν οριστεί οι συναρτήσεις κόστους (calculateCost) και Σταδιακής Καθόδου (gradientDescent) που έχουν παρουσιαστεί παραπάνω, ο υπολογισμός των εκτιμήσεων των συντελεστών β για τα δεδομένα του αρχείου icecream.csv στο περιβάλλον της R φαίνεται παρακάτω

```
#Ανάγνωση δεδομένων από το αρχείο

icecreamRevenues<-read.csv("IcecreamRevenues.csv", sep=",",
header=T)

# Τιμές της εξαρτημένης μεταβλητής (Revenue)

revenue<- icecreamRevenues [, 1]

# Δημιουργία μήτρας με δύο στήλες των ανεξάρτητων μεταβλητών

# Η πρώτη στήλη της μήτρας έχει μονάδες (1) για την αναπαράσταση
του σταθερού όρου και η δεύτερη

# στήλη περιέχει τις τιμές της ανεξάρτητης μεταβλητής θερμοκρασίας
(Temperature) του αρχείου δεδομένων

indVariables<- cbind( rep(1, 35), icecreamRevenues [, 2] )
```

```
# Δημιουργία διανύσματος με αρχικές τιμές συντελεστών θ.

# Γίνεται με επιλογή 2 τυχαίων τιμών θ από το διάστημα (0,1), που θα είναι οι αρχικές τιμές

# των θ0 και θ1. Η τιμή στη θέση initialThetas[1] είναι η αρχική τιμή του συντελεστή θ0 και initialThetas[2]

# η αρχική τιμή του συντελεστή θ1

initialThetas<-rep(runif(1), 2)

# Εκτέλεση του αλγορίθμου της Σταδιακής Καθόδου, με παράμετρος μάθησης α=0.00199 και

# πλήθος επαναλήψεων 65000. Η μεταβλητή gdOutput είναι λίστα που περιέχει τους συντελεστές

# ($coefficients) καθώς και τις τιμές της συνάρτησης κόστους σε κάθε επανάληψη ($costs).

gdOutput<-gradientDescent(indVariables, revenue, initialThetas, 0.00199, 65000)

#Απεικόνιση τιμών της συνάρτησης κόστους J(θ) συναρτήσει των επαναλήψεων

plot(gdOutput $costs, xlab="Επαναλήψεις", ylab="J(θ)" )
```

Η εκτέλεση του κώδικα θα δώσει τους παρακάτω συντελεστές

```
> gdOutput$coefficients

          [,1]

[1,] 12.9730634

[2,]  0.4377312
```

το οποίο είναι το διάνυσμα των συντελεστών θ που έχει εκτιμηθεί, όπου ο σταθερός όρος/συντελεστής θ0 έχει λάβει την τιμή 12.9730634 και ο συντελεστής θ1 έχει λάβει την τιμή 0.4377312[20].

[20] Οι τιμές μπορεί να διαφέρουν σε διαδοχικές εκτελέσεις του κώδικα, εξαιτίας της τυχαίας επιλογής αρχικών τιμών των συντελεστών θ. Άλλες τιμές παραμέτρων θα προκύψουν επίσης αν τροποποιηθούν η παράμετρος μάθησης α ή/και το πλήθος επαναλήψεων.

Αν εκτιμηθούν οι συντελεστές -για το ίδιο γραμμικό μοντέλο παλινδρόμησης και τα ίδια δεδομένα – με τη μέθοδο των ελαχίστων τετραγώνων, εκτελώντας την εντολή της R

```
ols.linear.regression<-lm(Revenue ~ Temperature, da-
ta=icecreamRevenues)
```

θα προκύψουν οι συντελεστές

```
Call:

lm(formula = Revenue ~ Temperature, data = icecreamRevenues)

Coefficients:

(Intercept)  Temperature

    13.0882       0.4341
```

όπου ο σταθερός όρος/συντελεστής θ0 έχει λάβει την τιμή 13.0882 ενώ ο συντελεστής θ1 (της μεταβλητής θερμοκρασία) την τιμή 0.4341.

Μεταξύ των μεθόδων αυτών παρατηρείται μία διαφορά στις εκτιμήσεις των συντελεστών που παράγουν. Η διαφορά που παρατηρείται οφείλεται κυρίως στο γεγονός ότι η μέθοδος της Σταδιακής Καθόδου προσεγγίζει τους συντελεστές με επαναληπτική διαδικασία όπως έχει αναφεθεί και και κατά συνέπεια το πλήθος των επαναλήψεων θα έχει επίπτωση στις τελικές εκτιμήσεις όπως επίσης και οι τιμές αρχικοποιήσεων των συντελεστών. Διαφορετικό πλήθος επαναλήψεων και διαφορετικές αρχικές τιμές συντελεστών μπορεί να καταλήξουν σε διαφορετικές εκτιμήσεις.

Επιπλέον, ο αλγόριθμος της Σταδιακής Καθόδου παρουσιάζει και ένα ζήτημα που αφορά τον ρυθμό σύγκλισης των συντελεστών σε κάθε επανάληψη. Ειδικότερα, οι συντελεστές του μοντέλου μπορεί να μην προσεγγίζουν τις τιμές στις οποίες συγκλίνουν με τον ίδια ρυθμό σε κάθε επανάληψη του αλγορίθμου. Έτσι σε κάθε επανάληψη, ορισμένοι συντελεστές θα φανούν ότι συγκλίνουν πιο «γρήγορα» στις πραγματικές τους τιμές που ελαχιστοποιούν τη συνάρτηση κόστους απ’ότι άλλοι. Στο παραπάνω παράδειγμα πωλήσεων παγωτών φαίνεται ότι η εκτίμηση του συντελεστή θ_1 προσεγγίζει πιο γρήγορα την πραγματική τιμή απ’ό,τι ο σταθερός όρος/συντελεστής θ_0. Η διαφορά στον ρυθμό σύγκλισης οφείλεται στο γε-

γονός, ότι ο τύπος ενημέρωσης των συντελεστών, και κατ' επέκταση ο ρυθμός αλλαγής της τιμής του συντελεστή, για τους μη-σταθερούς συντελεστές, εξαρτάται και από την τιμή της ανεξάρτητης μεταβλητής, όπως φαίνεται από τον όρο $x_k{}^{(i)}$ στον παρακάτω τύπο

$$\theta_k := \theta_k - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta \left(x^{(i)} \right) - y^{(i)} \right) x_k{}^{(i)}$$

Αυτό σημαίνει ότι σε περιπτώσεις πολλαπλής γραμμικής παλινδρόμησης, όπου τιμές των ανεξάρτητων μεταβλητών στο σύνολο δεδομένων και μετέχουν στο μοντέλο παλινδρόμησης, παρουσιάζουν διαφορές στην κλίμακά τους (π.χ. ορισμένες ανεξάρτητες μεταβλητές λαμβάνουν πολύ μεγάλες ενώ άλλες λαμβάνουν πολύ μικρές τιμές) οι αντίστοιχοι συντελεστές θα συγκλίνουν με διαφορετικό ρυθμό στην κατάλληλη τιμή τους και η οποία ελαχιστοποιεί τη συνάρτηση κόστους. Η ταχύτητα σύγκλισης των συντελεστών στη μέθοδο της Σταδιακής Καθόδου είναι γενικά ευαίσθητη στην κλίμακα μέτρησης των δεδομένων.

Σε δεδομένα με τέτοια χαρακτηριστικά, μία αντιμετώπιση είναι η κανονικοποίηση των δεδομένων (normalization), η οποία γίνεται στη φάση της προεπεξεργασίας, όπου τα δεδομένα υποβάλλονται σε μετασχηματισμούς ώστε οι τιμές όλων των ανεξαρτήτων μεταβλητών να έχουν την ίδια κλίμακα μέτρησης και οι τιμές τους να βρίσκονται στο ίδιο εύρος τιμών. Συνήθεις μέθοδοι κανονικοποίησης δεδομένων είναι η μέθοδος του Ελαχίστου-Μέγιστου που έχει σαν αποτέλεσμα όλες οι τιμές μιας μεταβλητής να απεικονίζονται στο εύρος τιμών [0,1] είτε η μέθοδος z-score που συνήθως απεικονίζει τις τιμές στο διάστημα [-3,3] (αν και άλλα εύρη είναι εφικτά να εμφανιστούν) και στις οποίες υποβάλλονται οι τιμές των ανεξαρτήτων μεταβλητών πριν την εκτέλεση της μεθόδου της Σταδιακής Καθόδου. Εάν γίνουν τέτοιοι μετασχηματισμοί, ο αλγόριθμος της Σταδιακής Καθόδου αντιμετωπίζει πολύ καλά τον ρυθμό σύγκλισης για όλους τους συντελεστές και απαιτεί συνήθως πολύ λιγότερες επαναλήψεις.

Παρακάτω παρουσιάζεται κώδικας R, ο οποίος εκτιμάει τους συντελεστές για το ίδιο μοντέλο παλινδρόμησης με παραπάνω, όπου ωστόσο οι τιμές της ανεξάρτητης μεταβλητής θερμοκρασία έχουν κανονικοποιηθεί με τη μέθοδο z-score.

```
# Κανοκοποίηση με τη μέθοδο zscore.

# Η συνάρτηση θα  επιστρέψει για κάθε τιμή που υπάρχει στο διάνυσμα x, πόσες τυπικές

# αποκλίσεις απέχει από τη μέση τιμή των τιμών.

zscore<-function(x){

  return( (x-mean(x))/sd(x) )

}

#Ανάγνωση δεδομένων από το αρχείο

icecreamRevenues<-read.csv("IcecreamRevenues.csv ", sep=",", header=T)

# Κανονικοποίηση z-score της ανεξάρτητης μεταβλήτής Θερμοκρασία.

icecreamRevenues[, "Temperature"]<- zscore(icecreamRevenues[, "Temperature"])

# Τιμές της εξαρτημένης μεταβλητής (Revenue)

revenue<- icecreamRevenues [, 1]

# Δημιουργία μήτρας με δύο στήλες των ανεξάρτητων μεταβλητών

# Η πρώτη στήλη της μήτρας έχει μονάδες (1) για την αναπαράσταση του σταθερού όρου και η δεύτερη

# στήλη περιέχει τις τιμές της ανεξάρτητης μεταβλητής θερμοκρασίας (Temperature) του αρχείου δεδομένων

indVariables<- cbind( rep(1, 35), icecreamRevenues [, 2] )

# Δημιουργία διανύσματος με αρχικές τιμές συντελεστών θ.

# Γίνεται με επιλογή 2 τυχαίων τιμών θ από το διάστημα (0,1), που θα είναι οι αρχικές τιμές

# των θ0 και θ1. Η τιμή στη θέση initialThetas[1] είναι η αρχική τιμή του συντελεστή θ0 και initialThetas[2]

# η αρχική τιμή του συντελεστή θ1

initialThetas<-rep(runif(1), 2)

# Εκτέλεση του αλγορίθμου της Σταδιακής Καθόδου, με παράμετρος μάθησης α=0.01 και
```

```
# πλήθος επαναλήψεων 10000. Η μεταβλητή gdOutput είναι λίστα που
περιέχει τους συντελεστές

# ($coefficients) καθώς και τις τιμές της συνάρτησης κόστους σε
κάθε επανάληψη ($costs).

gdOutput<-gradientDescent(indVariables, revenue, initialThetas,
0.01, 10000)
```

Άσκηση Αυτοαξιολόγησης 0.13

Εάν η συνάρτηση gradientDescent() που παρουσιάστηκε στην ενότητα αυτή κληθεί με τον ακόλουθο τρόπο για την εκτίμηση των συντελεστών του γραμμικού μοντέλου *Έσοδα= ϑ_0+ ϑ_1 Θερμοκρασία* :

```
icecreamRevenues <- read.csv("IcecreamRevenues.csv", sep=",",
header=T)
revenue <- icecreamRevenues [, 1]
indVariables <- icecreamRevenues [, 2]
initialThetas <- rep(runif(1), 2)
gdOutput <- gradientDescent(indVariables, revenue, initial-
Thetas, 0.00199, 65000)
```

τί πρόβλημα θα προκύψει;

Δραστηριότητα 0.3

Στον κώδικα R που εκτιμά του συντελεστές του γραμμικού μοντέλου παλινδρόμησης δίχως κανονικοποίηση z-score *Έσοδα= ϑ_0+ ϑ_1 Θερμοκρασία* για το σύνολο δεδομένων IcecreamRevenue.csv και παρουσιάστηκε στην ενότητα αυτή, να τροποποιήσετε τις παραμέτρους α (την παράμετρος μάθησης) και πλήθος επαναλήψεων που δίνονται ως όρισμα στη συνάρτηση gradientDescent() με τρόπο, ώστε η συνάρτηση κόστους να αποκλίνει από την ελάχιστη τιμή της. Δώστε το διάγραμμα διασποράς της συνάρτησης κόστους ως συνάρτηση του πλήθους επαναλήψεων όπου φαίνεται η απόκλιση.

6.6.3.6 Εκδοχές της μεθόδου Σταδιακής Καθόδου: Χρήση σε Περιβάλλοντα Μεγάλων Δεδομένων

Ο αλγόριθμος της Σταδιακής Καθόδου έρχεται σε διάφορες εκδοχές, οι οποίες κυρίως έχουν ως στόχο να βελτιώσουν την επίδοσή του σε περιπτώσεις που το σύνολο δεδομένων είναι πάρα πολύ μεγάλο (είναι της τάξεως των δεκάδων ή εκατοντάδων εκατομμυρίων παρατηρήσεων και άνω). Σε τέτοιες περιπτώσεις μπορούν να προκύψουν διάφορα ζητήματα όπως για παράδειγμα το σύνολο εκπαίδευσης να μην χωράει στη διαθέσιμη μνήμη RAM του υπολογιστή που είναι μία βασική υπόθεση του αλγορίθμου της Σταδιακής Καθόδου Δέσμης που παρουσιάστηκε παραπάνω. Κατά συνέπεια, ο υπολογισμός της ενημέρωσης των νέων τιμών των συντελεστών μπορεί να είναι πολύ χρονοβόρα ή ακόμη και αδύνατη διαδικασία, ειδικά αν χρησιμοποιείται η μορφή μήτρας του τύπου ενημέρωσης των συντελεστών.

Αυτό οφείλεται στο γεγονός ότι το ο υπολογισμός του αθροίσματος της τιμής της μερικής παραγώγου της συνάρτησης κόστους, δηλαδή ο όρος $\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right)-y^{(i)}\right)x_k{}^{(i)}$ που εμφανίζεται στον τύπο ενημέρωσης για τον συντελεστή θ_k, θα πρέπει να κάνει χρήση όλων των παρατηρήσεων του συνόλου δεδομένων για να υπολογίσει το άθροισμα σε κάθε επανάληψη και για κάθε συντελεστή ανεξάρτητης μεταβλητής. Για να φανεί καλύτερα η επίπτωση του πλήθους των παρατηρήσεων, εαν για παράδειγμα το σύνολο δεδομένων αποτελείται από 10.000.000 παρατηρήσεις (m=10000000) και πρέπει να εκτιμηθούν οι συντελεστές πέντε (5) ανεξάρτητων μεταβλητών, σε κάθε επανάληψη του αλγορίθμου της Σταδιακής Καθόδου, θα διαπεραστούν και οι 10000000 παρατηρήσεις μία φορά για κάθε έναν από τους πέντε συντελεστές ώστε να υπολογιστεί το άθροισμα σε μία μόνο επανάληψη και *για να προκύψει ένα μικρό βήμα προς τη σωστή κατεύθυνση* και κατά συνέπεια *μία μόνο νέα εκτίμηση* των τιμών για κάθε έναν από τους πέντε συντελεστές. Η διαδικασία αυτή είναι αρκετά χρονοβόρα με αποτέλεσμα ο αλγόριθμος της Σταδιακής Καθόδου να αποβαίνει πολύ αργός. Επιπλέον, εάν γίνεται χρήση του τύπου ενημέρωσης με τη μορφή μήτρας θα πρέπει τα δεδομένα να έχουν φορτωθεί στη μνήμη RAM του υπολογιστή κάτι που λόγω του μεγέθους τους μπορεί να μην είναι εφικτό.

Προκειμένου να αντιμετωπιστούν τα ζητήματα ενημέρωσης των συντελεστών σε τέτοιες περιπτώσεις μεγάλων δεδομένων, υπάρχουν εκδοχές του αλγορίθμου

της Σταδιακής Καθόδου, οι οποίες διαφοροποιούνται κυρίως στο πόσα δεδομένα λαμβάνονται υπόψη για την ενημέρωση των νέων τιμών των συντελεστών θ. Οι πιο διαδεδομένες εκδοχές του αλγορίθμου της Σταδιακής Καθόδου είναι οι εξής:

- Η μέθοδος της Στοχαστικής Σταδιακής Καθόδου (Stochastic Gradient Descent - SGD). Σε αυτήν την εκδοχή του αλγορίθμου, ο τύπος ενημέρωσης των συντελεστών δεν χρησιμοποιεί το άθροισμα της τιμής της μερικής παραγώγου για κάθε παρατήρηση του συνόλου δεδομένων. Αντ' αυτού ο τύπος ενημέρωσης κάνει χρήση μίας παρατήρησης σε κάθε επανάληψη (αντί του αθροίσματος της μερικής παραγώγου όλων των παρατηρήσεων στην κλασική υλοποίηση του αλγορίθμου) για την ενημέρωση των συντελεστών και ο ψευδοκώδικας της εκδοχής αυτής θα έχει ως εξής:

Αρχικοποίηση όλων των συντελεστών θ_j με τυχαίες τιμές
Δημιουργία τυχαίας διάταξη του συνόλου δεδομένων
Ενόσω δεν πληρούνται τα κριτήρια τερματισμού {
  Για κάθε παρατήρηση i του συνόλου δεδομένων {

    Για όλους τους συντελεστές θ_j {
      $new_{\theta_j} := \theta_j - \alpha \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$
    }
  }
  Για όλους τους συντελεστές θ_j
    $\theta_j := new_{\theta_j}$ */* Ενημέρωση όλων των συντελεστών με τις νέες τιμές τους */*

}

Αυτό που κάνει την εκδοχή της Στοχαστικής Σταδιακής Καθόδου κατάλληλη σε περιβάλλοντα μεγάλων δεδομένων – και την διαφοροποιεί από την παραδοσιακή εκδοχή του αλγορίθμου - είναι το γεγονός ότι δεν θα «σαρώσει» ολόκληρο το σύνολο δεδομένων σε κάθε επανάληψη για να υπολογίσει το άθροισμα της τιμής της μερικής παραγώγου για μία ενημέρωση τιμής κάθε συντελεστή. Αντ' αυτού κάνει χρήση μίας μόνο παρατήρησης για μία ενημέρωση της τιμής του συντελεστή όπως φαίνεται από τον όρο $\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$ (και όπου λείπει το άθροισμα της μερικής παραγώγου για όλες τις παρατηρήσεις) και έτσι ολόκληρο το σύνο-

λο δεδομένων θα «σαρωθεί» μία μόνο φορά κατά την εκτέλεση αυτής της εκδοχής του αλγορίθμου της Σταδιακής Καθόδου. Κατά συνέπεια αποτελεί μία πολύ πιο γρήγορη διαδικασία εκτίμησης των συντελεστών..

Επιπλέον, η εκδοχή αυτή δεν απαιτεί να έχει φορτωθεί ολόκληρο το σύνολο δεδομένων στη μνήμη εφόσον απαιτεί μία παρατήρηση τη φορά. Τέλος, η εκδοχή της Στοχαστικής Σταδιακής Καθόδου αποτελεί μία κατάλληλη προσέγγιση σε περιπτώσεις όπου το σύνολο δεδομένων δεν είναι γνωστό εξ'αρχής και τα δεδομένα καταφθάνουν/γίνονται διαθέσιμα σε πραγματικό χρόνο ακριβώς επειδή επεξεργάζεται μία μόνο παρατήρηση τη φορά και βάση αυτής ενημερώνει τις νέες τιμές των συντελεστών (online algorithm).

- Η μέθοδος της Σταδιακής Καθόδου Μικρών Δεσμών (Mini-Batch Gradient Descent - MBGD). Η εκδοχή αυτή αποτελεί μία ενδιάμεση προσέγγιση σε σχέση με τις άλλες δύο: δεν χρησιμοποιεί όλο το σύνολο δεδομένων ως μία δέσμη (όπως κάνει η κλασική εκδοχή της Σταδιακής Καθόδου Δέσμης) αλλά ούτε και μία μόνο παρατήρηση του συνόλου δεδομένων (όπως κάνει η Στοχαστική Σταδιακή Κάθοδος) για την ενημέρωση των τιμών των συντελεστών.

  Η μέθοδος Σταδιακής Καθόδου Μικρών Δεσμών χωρίζει το σύνολο δεδομένων σε μικρότερα υποσύνολα συγκεκριμένου σταθερού μεγέθους (που αποτελούν μικρές δέσμες) και κάνει χρήση του τύπου ενημέρωσης συντελεστών της Σταδιακής Καθόδου Δέσμης, αθροίζοντας ωστόσο τις τιμές της πρώτης παραγώγου των δεδομένων κάθε μικρής δέσμης κι όχι ολόκληρου του συνόλου δεδομένων. Ο αλγόριθμος δηλαδή θεωρεί κάθε μικρή δέσμη που δημιουργείται ότι είναι το σύνολο δεδομένων και χρησιμοποιεί όλες τις δέσμες για την εκτίμηση των συντελεστών. Για παράδειγμα, εάν το σύνολο δεδομένων περιέχει 5.000.000 παρατηρήσεις, και αυτό χωριστεί σε μικρές δέσμες των 1000 παρατηρήσεων, θα υπάρχουν συνολικά 5000 δέσμες. Σε μία τέτοια περίπτωση ο ψευδοκώδικας της Σταδιακής Καθόδου Μικρών Δεσμών για την ενημέρωση των συντελεστών ενός πολλαπλού γραμμικού μοντέλου παλινδρόμησης θα έχει ως εξής:

Αρχικοποίηση όλων των συντελεστών θ_j με τυχαίες τιμές
Χωρισμός του συνόλου δεδομένων σε μικρές δέσμες π.χ. των 1000 παρατηρήσεων έκαστη
Ενόσω δεν πληρούνται τα κριτήρια τερματισμού {
*/*για κάθε μικρή δέσμη t */*
Για t=1…5000 {
Για όλους τους συντελεστές θ_j {

$$new_{\theta_j} := \theta_j - \alpha \sum_{i=1}^{1000} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) x_j^{(i)}$$

}
}
Για όλους τους συντελεστές θ_j
$\theta_j := new_{\theta_j}$ / Ενημέρωση όλων των συντελεστών με τις νέες τιμές τους **
}

όπου $x^{(i)}$ και $y^{(i)}$ οι τιμές των ανεξαρτήτων και εξαρτημένης μεταβλητής της i-οστής παρατήρηση στην τρέχουσα δέσμη t αντίστοιχα. Το ενδιαφέρον που παρουσιάζει η μέδοθος της Σταδιακής Μεθόδου Μικρών Δεσμών είναι ότι ο τύπος ενημέρωσης των συντελεστών μπορεί να κάνει χρήση μητρών για τον υπολογισμό αυτών και κατά συνέπεια να εκμεταλλευτεί μεθόδους βελτιστοποίησης ακόμη και εάν χρησιμοποιηθούν μεγάλα δεδομένα. Η χρήση μητρών μπορεί να γίνει καθότι οι δέσμες έχουν λίγες σε πλήθος παρατηρήσεις και κατά συνέπεια μπορούν να χωρέσουν στη μνήμη του υπολογιστή.

Το κατάλληλο μέγεθος (σε πλήθος παρατηρήσεων) της δέσμης είναι θέμα δοκιμών και ελέγχου της τιμής της συνάρτησης κόστους. Σε ακραίες τιμές του πλήθους παρατηρήσεων στις δέσμες η εκδοχή των Μικρών Δεσμών καταλήγει να γίνεται η Στοχαστική ή η εκδοχή Δέσμης. Έτσι, αν το μέγεθος της δέσμης τεθεί ίσο με το πλήθος των παρατηρήσεων στο σύνολο των δεδομένων, τότε ο αλγόριθμος καταλήγει να γίνεται ο αλγόριθμος της Σταδιακής Καθόδου Δέσμης. Αν το μέγεθος της δέσμης τεθεί ίσο με ένα (1) δηλαδή κάθε δέσμη να περιέχει μία μόνο παρατήρηση, τότε ο αλγόριθμος της Σταδιακής Καθόδου Μικρών Δεσμών καταλήγει να γίνεται η εκδοχή της Στοχαστικής Σταδιακής Καθόδου. Γενικά, χαρακτηριστικά μεγέθη δεσμών είναι 32, 64 ή και παραπάνω ενώ δεν συνηθίζονται

μεγέθη πάνω από 2000. Συνηθίζεται επίσης το πλήθος παρατηρήσεων στις δέσμες να είναι δύναμη του 2, καθότι αυτό επιτρέπει να βελτιστοποιηθούν διαδικασίες πρόσβασης των δεδομένων στη μνήμη και την βέλτιστη χρήση των πυρήνων του επεξεργαστή.

Στο περιβάλλον της R υπάρχουν βιβλιοθήκες που παρέχουν όλες τις εκδοχές του αλγορίθμου της Σταδιακής Καθόδου όπως για παράδειγμα η βιβλιοθήκη gradDescent.

6.6.4 Σύγκριση μεθόδων Ελαχίστων Τετραγώνων και Σταδιακής Καθόδου

Τόσο η μέθοδος των Ελαχίστων Τετραγώνων όσο και η μέθοδος της Σταδιακής Καθόδου και οι διάφορες εκδοχές της μπορούν να χρησιμοποιηθούν για να εκτιμηθούν οι συντελεστές ενός πολλαπλού γραμμικού μοντέλου παλινδρόμησης. Ωστόσο, κάθε μέθοδος έχει τα προτερήματά της και είναι η καλύτερη προσέγγιση σε συγκεκριμένες περιπτώσεις. Παρακάτω θα συζητηθούν και θα συγκριθούν οι μέθοδοι αυτές.

Τρεις είναι οι σημαντικότερες διαφορές των δύο αυτών μεθόδων. Η πρώτη είναι ότι η μέθοδος των ελαχίστων τετραγώνων θα υπολογίζει πάντα τις βέλτιστες, αμερόληπτες εκτιμήσεις συντελεστών για ένα σύνολο δεδομένων, αν ισχύσουν ορισμένες προϋποθέσεις, όσες φορές και αν εκτελεστεί η μέθοδος για το ίδιο γραμμικό μοντέλο παλινδρόμησης και για τα ίδια δεδομένα. Το βέλτιστες εδώ αναφέρεται στο ότι οι συντελεστές που θα υπολογιστούν εξασφαλίζουν δώσουν τις μικρότερες διακυμάνσεις απ'όλες τις άλλες τιμές συντελεστών. Αυτό οφείλεται στο γεγονός ότι βασίζεται στην κανονική εξίσωση για τον υπολογισμό των συντελεστών. Αντιθέτως, η μέθοδος της Σταδιακής Καθόδου μπορεί να οδηγήσει σε διαφορετικές εκτιμήσεις συντελεστών αν εκτελεστεί πάνω στα ίδια δεδομένα. Αυτό οφείλεται κυρίως στο γεγονός ότι οι αρχικές τιμές των συντελεστών επιλέγονται τυχαία καθώς επίσης και ότι η σύγκλιση της μεθόδου εξαρτάται τόσο από την παράμετρο μάθησης α και το πλήθος επαναλήψεων. Εαν αλλάξουν αυτές οι παράμετροι και εκτελεστεί ο αλγόριθμος της Σταδιακής Καθόδου πάνω στα ίδια δεδομένα, το αποτέλεσμα μπορεί να είναι διαφορετικό.

Η δεύτερη διαφορά σχετίζεται με περιπτώσεις, όπου το διαθέσιμο σύνολο εκπαίδευσης δεδομένων είναι πολύ μεγάλο τόσο με όρους παρατηρήσεων όσο και με όρους μεταβλητών. Τέτοιες περιπτώσεις χαρακτηρίζονται από πλήθος παρα-

τηρήσεων της τάξεως των εκατοντάδων χιλιάδων και μεγάλο πλήθος ανεξάρτητων μεταβλητών (που μπορεί να είναι πάνω από 500). Εαν τέτοια είναι τα χαρακτηριστικά των δεδομένων και του γραμμικού μοντέλου παλινδρόμησης, η χρήση της μεθόδου των ελαχίστων τετραγώνων είναι υπολογιστικά μία πολύ αργή διαδικασία επειδή η κανονική εξίσωση απαιτεί την αντιστροφή της μήτρας $X^T X$ όπως προκύπτει από την κανονική εξίσωση $\hat{\beta} = (X^T X)^{-1} X^T y$, όπου X η μήτρα τιμών των ανεξάρτητων μεταβλητών. Η πράξη της αντιστροφής μήτρας έχει πολυπλοκότητα χρόνου $O(n^3)$ όπου n το πλήθος στηλών τετραγωνικής μήτρας και κατά συνέπεια εαν υπάρχει μεγάλο πλήθος ανεξάρτητων μεταβλητών, η πράξη της αντιστροφής είναι ιδιαιτέρως χρονοβόρα διαδικασία. Επιπλέον, η κανονική εξίσωση απαιτεί ολόκληρη τη μήτρα τιμών X να φορτωθεί στη μνήμη RAM του υπολογιστή κάτι το οποίο -σε συνθήκες μεγάλων δεδομένων μπορεί να μην είναι εφικτό. Υπό τέτοιες συνθήκες ο αλγόριθμος της Σταδιακής Καθόδου (και ειδικά οι διάφορες εκδοχές του) δίνει πολύ καλύτερα αποτελέσματα σε πολύ σύντομο χρονικό διάστημα σε σχέση με τη μέθοδο των ελαχίστων τετραγώνων. Η μέθοδος της Σταδιακής Καθόδου αποτελεί υπό τέτοιες συνθήκες δεδομένων και μοντέλων την ιδανικότερη προσέγγιση.

Τέλος είναι επίσης σημαντικό να τονιστεί ότι η μέθοδος των ελαχίστων τετραγώνων μπορεί να χρησιμοποιηθεί μόνο για την εκτίμηση συντελεστών γραμμικών μοντέλων παλινδρόμησης. Αντιθέτως, η μέθοδος της Σταδιακής Καθόδου μπορεί να χρησιμοποιηθεί για την εκτίμηση συντελεστών και μη-γραμμικών μοντέλων παλινδρόμησης.

Μία πιο αναλυτική σύνοψη των δυο αυτών μεθόδων αυτών παρουσιάζεται στον παρακάτω πίνακα.

Μέθοδος Ελαχίστων Τετραγώνων (OLS)	**Μέθοδος Σταδιακής Καθόδου (Gradient Descent)**
Καταλήγει πάντα στις ίδιες, βέλτιστες εκτιμήσεις συντελεστών αν εφαρμοστεί για το ίδιο γραμμικό μοντέλο παλινδρόμησης και για τα ίδια δεδομένα εκπαίδευσης εφόσον πληρούνται οι ορισμένες προϋποθέσεις.	Δεν καταλήγει ούτε στις ίδιες ούτε στις βέλτιστες εκτιμήσεις συντελεστών αν εφαρμοστεί για το ίδιο γραμμικό μοντέλο παλινδρόμησης και για τα ίδια δεδομένα εκπαίδευσης. Οι τιμές των συντελεστών που εκτιμώνται εξαρτώνται από τις παρα-

	μέτρους της μεθόδου.
Είναι υπολογιστικά αργή διαδικασία για μεγάλα σύνολα δεδομένων και μεγάλο πλήθος ανεξάρτητων μεταβλητών. Δεν μπορεί να χρησιμοποιηθεί σε περιβάλλοντα μεγάλων δεδομένων (Big data).	Υπολογίζει πολύ γρήγορα ικανοποιητικές προσεγγίσεις συντελεστών εάν το σύνολο δεδομένων και το πλήθος των ανεξάρτητων μεταβλητών είναι μεγάλο. Εκδοχές της μεθόδου μπορούν να χρησιμοποιηθούν σε περιβάλλοντα μεγάλων δεδομένων.
Μπορεί να χρησιμοποιηθεί για την εκτίμηση συντελεστών μόνο γραμμικών μοντέλων παλινδρόμησης.	Μπορεί να χρησιμοποιηθεί για την εκτίμηση συντελεστών τόσο για γραμμικά όσο και μη-γραμμικά μοντέλα παλινδρόμησης.
Παρέχει κλειστούς τύπους (κανονική εξίσωση) για τον υπολογισμό των εκτιμήσεων συντελεστών του μοντέλου της γραμμικής παλινδρόμησης	Δεν παρέχει κλειστούς τύπους για τον υπολογισμό των εκτιμήσεων συντελεστών. Αποτελεί επαναληπτικό αλγόριθμο βελτιστοποίησης.
Απαιτεί ολόκληρο το σύνολο δεδομένων να είναι γνωστό και διαθέσιμο εξ'αρχής.	Εκδοχές της μπορούν να χρησιμοποιηθούν (η Στοχαστική Σταδιακή Καθόδου) σε περίπτωση που τα δεδομένα δεν είναι εξ'αρχής διαθέσιμα και προσέρχονται προς επεξεργασία σε πραγματικό χρόνο (online).
Διδάσκεται και χρησιμοποιείται κυρίως στις κοινωνικές επιστήμες όπου η έμφαση είναι στην εξήγηση της τιμής της εξαρτημένης μεταβλητής.	Διδάσκεται και χρησιμοποιείται κυρίως στις επιστήμες πληροφορικής και πολυτεχνικά τμήματα όπου η έμφαση είναι στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής.

Άσκηση Αυτοαξιολόγησης 0.14

Θέλετε να εκτιμήσετε τους συντελεστές ενός γραμμικού μοντέλου παλινδρόμησης. Ποια μέθοδος θα χρησιμοποιήσετε σε κάθε μία από τις παρακάτω περιπτώσεις;

i. Τα δεδομένα εκπαίδευσης είναι τόσες πολλές παρατηρήσεις, που δεν χωράνε στην κεντρική μνήμη του υπολογιστή.

ii. Τα δεδομένα εκπαίδευσης δεν είναι διαθέσιμα εξ'αρχής και προσέρχονται μία παρατήρηση τη φορα με άγνωστο ρυθμό.

iii. Επιθυμείται η βέλτιστη εκτίμηση συντελεστώνΜε χρήση κατάλληλων εντολών στην R, ορίστε ένα αντικείμενο για κάθε βασική κλάση αντικειμένου και εμφανίστε την τιμή του και την κλάση στην οποία ανήκει.

6.7 Αξιολόγηση και ερμηνεία μοντέλων γραμμικής παλινδρόμησης

Αυτό που έχει παρουσιαστεί μέχρι αυτό το σημείο είναι μέθοδοι εκτίμησης των συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης βάσει ενός συνόλου εκπαίδευσης. Η εκτίμηση των συντελεστών ενός μοντέλου παλινδρόμησης ωστόσο δεν ολοκληρώνει τη διαδικασία προσδιορισμού του. Αφού έχουν εκτιμηθεί οι συντελεστές πρέπει να γίνουν ορισμένοι έλεγχοι προκειμένου να εξεταστεί εάν ορισμένες υποθέσεις που προϋποθέτουν τα μοντέλα γραμμικής παλινδρόμησης ισχύουν, να αξιολογηθούν οι συντελεστές που εκτιμήθηκαν και να γίνει μία ερμηνεία του μοντέλου ώστε να βεβαιωθεί ότι τα συμπεράσματα που βγαίνουν απ' αυτό είναι έγκυρα και έχουν νόημα. Για παράδειγμα, κατά τη σύνταξη ενός γραμμικού μοντέλου παλινδρόμησης έγινε η υπόθεση ότι η εξαρτημένη μεταβλητή είναι γραμμική ως προς τους συντελεστές. Μία τέτοια υπόθεση υποστηρίχθηκε μόνο διαγραμματικά κάτι που αποτελεί μία ασθενή απόδειξη. Θα πρέπει να εξεταστεί ότι η υπόθεση της γραμμικής συσχέτισης μεταξύ των μεταβλητών και όπως αυτή αποτυπώνεται στο μοντέλο παλινδρόμησης είναι όντως δικαιολογημένη. Η αξιολόγηση και ερμηνεία μοντέλων γραμμικής παλινδρόμησης είναι ιδιαίτερα σημαντικές διαδκασίες καθότι αυτές αποδίδουν νόημα και σημασία στο μοντέλο και αποφαίνονται για το εάν μπορούν να βγουν χρήσιμα συμπεράσματα απ'αυτό. Και αυτό διότι, εκτίμηση των συντελεστών μπορεί να γίνει για οποιοδήποτε γραμμικό μοντέλο παλινδρόμησης, που μπορεί να περιέχει οποιεσδήποτε μεταβλητές, ανεξάρτητα από το εαν η μελέτη της σχέσης των μεταβλητών αυτών έχει νόημα ή όχι[21].

[21] Για παράδειγμα, μπορούν κάλλιστα να εκτιμηθούν οι συντελεστές ενός γραμμικού μοντέλου παλινδρόμησης που προσπαθεί να συλλάβει τη σχέση μεταξύ του μισθού εργαζομένων και το πλήθος των ηλιακών κηλίδων κάθε μήνα ή το ζώδιο του εργαζομένου. Το εάν ένα τέτοιο μοντέλο έχει νόημα, είναι αντικείμενο των διαδικασιών αξιολόγησης και ερμηνείας του.

Οι έλεγχοι αυτοί μπορούν να γίνουν μόνο μετά την εκτίμηση των συντελεστών, μιας και κάνουν χρήση ορισμένων στοιχείων που προκύπτουν κατά τη διάρκεια εκτίμησή τους. Σε αυτό το σημείο πρέπει να τονιστεί ότι το ποιοι έλεγχοι θα πρέπει να πραγματοποιηθούν εξαρτάται κυρίως από τον σκοπό του μοντέλου παλινδρόμησης. Έτσι, εάν ένα μοντέλο γραμμικής παλινδρόμησης συντάσσεται με στόχο την εξήγηση της διακύμανσης της τιμής της ανεξάρτητης μεταβλητής, θα πρέπει να γίνει χρήση διαφορετικών μεθόδων για τον έλεγχο των υποθέσεων, την αξιολόγηση και την ερμηνεία του απ'ο,τι στην περίπτωση που ο στόχος του μοντέλου είναι η πρόβλεψη. Έτσι στις επόμενες ενότητες, παρουσιάζονται οι τρόποι αξιολόγησης και ερμηνείας βάσει του στόχου του γραμμικού μοντέλου παλινδρόμησης.

6.7.1 Μοντέλα γραμμικής παλινδρόμησης με στόχο την εξήγηση της διακύμανσης

Μοντέλα που δημιουργούνται με στόχο την εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής αποσκοπούν κυρίως να εκτιμήσουν με ακρίβεια πως οι μεταβολές στις τιμές των ανεξαρτήτων μεταβλητών θα επηρεάσουν (και αν) την τιμή της εξαρτημένης μεταβλητής. Η έμφαση εδώ είναι στον ακριβή προσδιορισμό της επίδρασης αυτής από κάθε ανεξάρτητη μεταβλητή του μοντέλου δίχως να αφήνεται καμία υποψία ότι κάποιες τέτοιου είδους επιδράσεις έχουν παραληφθεί ή έχουν επισκιαστεί ή ότι η επίδραση δεν μπορεί να συλληφθεί με τη γραμμικότητα του μοντέλου.

6.7.1.1 Τεκμηρίωση του γραμμικού μοντέλου παλινδρόμησης

Το πρώτο βήμα που συντελείται αφότου έχουν εκτιμηθεί οι συντελεστές ενός γραμμικού μοντέλου παλινδρόμησης είναι να εξεταστεί εάν η γραμμικότητα του μοντέλου, έτσι όπως έχει προσδιοριστεί, είναι δικαιολογημένη και είναι μία βάσιμη υπόθεση για τα δεδομένα του συνόλου εκπαίδευσης.

Στα μοντέλα γραμμικής παλινδρόμησης που έχουν παρουσιαστεί έως τώρα, η υπόθεση ότι μία γραμμική συσχέτιση διέπει τις υπό εξέταση μεταβλητές του μοντέλου δεν είχε εξεταστεί αναλυτικότερα. Ως ένδειξη για τη γραμμική συσχέτιση έχει παρουσιαστεί μόνο το διάγραμμα διασποράς σε προηγούμενη ενότητα, το οποίο χρησιμοποιήθηκε για να συνταχθεί ένα γραμμικό μοντέλο μεταξύ των μεταβλητών. Ωστόσο, ένα τέτοιο διάγραμμα διασποράς δεν αποτελεί παρά μόνο

μία πρώτη ένδειξη της γραμμικής συσχέτισης μεταξύ των μεταβλητών. Επιπλέον, η χρήση τέτοιων διαγραμμάτων διασποράς δεν είναι προφανής όταν πρέπει να μελετηθούν οι μεταβλητές ενός πολλαπλού μοντέλου παλινδρόμησης. Σε μία τέτοια περίπτωση, δεν είναι δυνατή η πολυδιάστατη απεικόνιση του χώρου των δεδομένων.

Η τεκμηρίωση του γραμμικού μοντέλου είναι σημαντική διαδικασία καθότι αν αυτό δεν δικαιολογείται θα προκύψουν εκτιμήσεις συντελεστών, οι οποίες θα είναι ανακριβείς και θα οδηγήσει σε δυσκολίες ερμηνείας των αποτελεσμάτων και κατά συνέπεια σε λανθασμένα συμπεράσματα. Η τεκμηρίωση του γραμμικού μοντέλου εξασφαλίζει ότι ένα γραμμικό μοντέλο αποτελεί την καλυτερη προσαρμογή στο σύνολο δεδομένων εκπαίδευσης εξασφαλίζοντας την αξιοπιστία των αποτελεσμάτων και των ερμηνειών.

Τέτοια τεκμηρίωση και δικαιολόγηση ενός γραμμικού μοντέλου παλινδρόμησης γίνεται με τον έλεγχο συγκεκριμένων υποθέσεων. Εάν οι υποθέσεις αυτές ισχύουν για ένα μοντέλο παλινδρόμησης, τότε τεκμηριώνεται η γραμμική του φύση. Εαν κάποιες από αυτές δεν πληρούνται τότε προκύπτουν προβλήματα στην μοντελοποίηση της σχέσης των μεταβλητών με ένα τέτοιο γραμμικό μοντέλο και θα προκύψουν προβλήματα στην ερμηνεία των αποτελεσμάτων. Στη βιβλιογραφία αναφέρονται συχνά μία μεγάλη λίστα τέτοιων υποθέσεων που θα πρέπει να ελεγχθούν. Εδώ ωστόσο θα παρουσιαστούν οι τέσσερεις (4) πιο σημαντικές απ'αυτές και οι οποίες πρέπει να πάντα ελέγχονται στα πλαίσια μιας γραμμικής παλινδρόμησης που στόχο έχει την εξήγηση της διακύμανσης της τιμής της ανεξάρτητης μεταβλητής. Παρακάτω εξηγούνται οι υποθέσεις αυτές που πρέπει να πληρούνται ώστε να τεκμηριώνεται το γραμμικό μοντέλο, παρουσιάζεται πως μπορούν να ελεγχθούν με την χρήση της R και πόσο σοβαρή είναι η παραβίασή τους. Συνηθίζεται οι υποθέσεις αυτές να ελέγχονται αφού έχουν εκτιμηθεί οι συντελεστές του γραμμικού μοντέλου -και όχι πριν - καθότι οι υποθέσεις ελέγχουν και τις ιδιότητες των καταλοίπων, τα οποία είναι γνωστά αφότου προσδιοριστούν οι συντελεστές.

- ***Έλεγχος υπόθεσης γραμμικότητας***

Ο έλεγχος της υπόθεσης γραμμικότητας έχει ως στόχο να επιβεβαιώσει ότι η τιμή της εξαρτημένης μεταβλητής εξαρτάται γραμμικά από τις τιμές των ανεξαρτήτων μεταβλητών και η επίδραση της τιμής της κάθε ανεξάρτητης μεταβλητής στην

τιμή της εξαρτημένης είναι προσθετική και ανεξάρτητη από τις τιμές των άλλων ανεξάρτητων μεταβλητών.

Ο έλεγχος της υπόθεσης γραμμικότητας γίνεται με ένα διάγραμμα παρατηρούμενων και προβλεπόμενων τιμών[22] ή ένα διάγραμμα καταλοίπων ως προς τις προβλεφθείσες τιμές από το μοντέλο. Το σχήμα των διαγραμμάτων αυτών θα δώσει τις ενδείξεις για να τεκμηριωθεί ή όχι η υπόθεση γραμμικότητας. Τέτοια διαγράμματα μπορούν να ελέγξουν αν η υπόθεση αυτή ισχύει τόσο για απλά όσο και πολλαπλά γραμμικά μοντέλα παλινδρόμησης. Παρακάτω φαίνεται ο κώδικας σε R που δημιουργούν τα διαγράμματα αυτά και που επιτρέπουν το έλεγχο γραμμικότητας για το ακόλουθο γραμμικό μοντέλο παλινδρόμησης

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_0$$

και για το σύνολο δεδομένων HouseholdData.csv

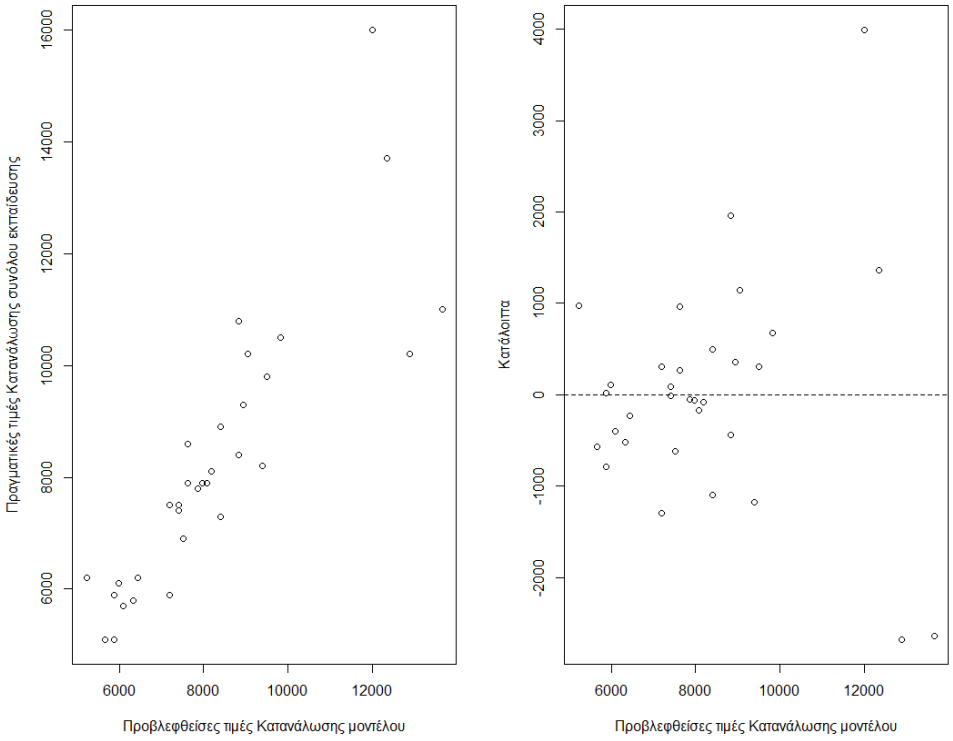
```
foodConsumptionData<-read.csv("HouseholdData.csv ", sep=",", head-
er=T)

# Εκτίμηση συντελεστών του απλού γραμμικού μοντέλου παλινδρόμησης
- με μία ανεξάρτηση μεταβλητή

# Κατανάλωση τροφίμων = β1Εισόδημα + β0

# με τη χρήση της συνάρτησης lm() που βασίζεται στη μέθοδο των ελα-
χίστων τετραγώνων (OLS).

linear.regression.model<-lm(FoodExpenditure ~ Income, da-
ta=foodConsumptionData)

# Έλεγχος υπόθεσης γραμμικότητας

# Θα απεικονιστούν δύο διαγράμματα:  διάγραμμα παρατηρούμενων και
προβλεπόμενων τιμών και το

# διάγραμμα καταλοίπων ως προς τις προβλεφθείσες τιμές από το μο-
ντέλο.

# Η εντολή par() επιτρέπει τον συνδυασμό και την συνεμφάνιση πολλών
διαγραμμάτων.
```

[22] Με τον όρο προβλεπόμενες τιμές μοντέλου παλινδρόμησης νοούνται οι τιμές τις εξαρτημένης μεταβλητής που προκύπτουν από γραμμικό μοντέλο παλινδρόμησης εάν δοθούν συγκεκριμένες τιμές των ανεξαρτήτων μεταβλητών.

```
# Η κλήση της par() δημιουργεί μία μήτρα με 1 γραμμή και 2 στήλες, όπου

# σε κάθε θέση/κελί της μήτρας θα απεικονιστεί ένα διαφορετικό διάγραμμα.

# Η συνάρτηση par() αλλάζει τις καθολικές ρυθμίσεις/προτιμήσεις των συναρτήσεων σχεδιασμού όπως της

# plot() που θα χρησιμοποιηθεί παρακάτω.

par( mfrow=c(1,2) )

#Απεικόνιση προβλεφθεισών και πραγματικών τιμών εξαρτημένης μεταβλητής

plot(linear.regression.model$fitted.values, foodConsumptionData$FoodExpenditure, xlab="Προβλεφθείσες τιμές Κατανάλωσης μοντέλου", ylab="Πραγματικές τιμές Κατανάλωσης συνόλου εκπαίδευσης")

#Απεικόνιση προβλεπόμενων τιμών εξαρτημένης μεταβλητής και καταλοίπων

plot(linear.regression.model$fitted.values, linear.regression.model$residuals, xlab="Προβλεφθείσες τιμές Κατανάλωσης μοντέλου", ylab="Κατάλοιπα")

# Απεικόνιση οριζόντιας γραμμής που διέρχεται από το σημείο y=0, για την επισήμανση των σημείων όπου τα

# κατάλοιπα είναι # ίσα με 0.

# Θα εμφανιστεί μία διακεκομμένη γραμμή (όρισμα lty=2)

abline(0,0, lty=2)
```

Η εκτέλεση του παραπάνω κώδικα R θα εμφανίσει τα δύο διαγράμματα όπως παρουσιάζονται στο σχήμα 6.13. Εάν το διάγραμμα παρατηρούμενων και προβλεπόμενων τιμών της εξαρτημένης μεταβλητής ακολουθεί μία ευθεία γραμμή (αριστερό διάγραμμα σχήματος 6.13) τότε υποστηρίζεται η υπόθεση της γραμμικότητας. Επιπλέον, εάν το διάγραμμα καταλοίπων ως προς τις προβλεπόμενες τιμές δεν αναδεικνύει καμία δομή και σχήμα και τα κατάλοιπα φαίνονται να είναι τυχαία κατανεμημένα γύρω από την τιμή 0 και να μην εξαρτώνται από την προβλεφθείσα τιμή του μοντέλου όπως στο δεξιό διάγραμμα του σχήματος 6.13, τό-

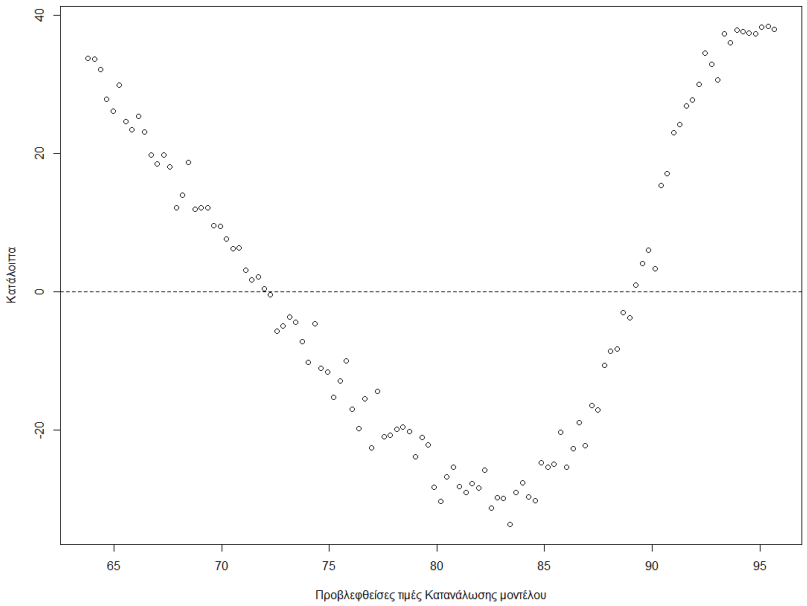
τε και πάλι υποστηρίζεται η γραμμική υπόθεση του μοντέλου. Το διάγραμμα καταλοίπων είναι ιδιαίτερα χρήσιμο εργαλείο για τον έλεγχο της υπόθεσης αυτής.



Εικόνα 0.13 *Διαγράμματα για τον έλεγχο της υπόθεσης γραμμικότητας. Αριστερά η απεικόνιση παρατηρούμενων και προβλεφθεισών τιμώ και δεξιά το διάγραμμα καταλοίπων ως προς τις προβλεφθείσες τιμές. Το σχήμα των διαγραμμάτων αυτών συνηγορεί υπέρ στο ότι τεκμηριώνεται η γραμμικότητα μεταξύ των μεταβλητών του μοντέλου παλινδρόμησης.*

Για την υποστήριξη της υπόθεσης γραμμικότητας, τα κατάλοιπα θα πρέπει να εμφανίζονται κατανεμημένα με τυχαίο τρόπο όταν απεικονιστούν ως προς τις προβλεφθείσες τιμές του γραμμικού μοντέλου παλινδρόμησης. Όπως έχει αναφερθεί, οι τιμές των ανεξαρτήτων μεταβλητών πρέπει να εξηγούν τη διακύμανση της τιμής της εξαρτημένης μεταβλητής, με τα κατάλοιπα να συλλαμβάνουν τη διακύμανση που δεν μπορεί να εξηγηθεί από τις μεταβλητές που έχουν συμπεριληφθεί στο μοντέλο. Τα κατάλοιπα δεν θα πρέπει να περιέχουν καμία προβλε-

πτική δύναμη της τιμής της ανεξάρτητης μεταβλητής και κατά συνέπεια δεν θα πρέπει να εξαρτώνται από τις τιμές που προβλέπει το μοντέλο. Εαν ωστόσο το διάγραμμα καταλοίπων εμφανίζει κάποιου είδους δομή ή σχήμα, όπως φαίνεται στο παρακάτω σχήμα 6.14, αυτό σημαίνει ότι υπόθεση της γραμμικότητας παραβιάζεται και δεν τεκμηριώνεται. Στο σχήμα 6.14 τα κατάλοιπα δεν φαίνονται να έχουν τυχαία κατανομή γύρω από την τιμή 0 και εμφανίζουν μία δομή ή μορφή καμπύλης. Φαίνεται για παράδειγμα ότι η τιμή των καταλοίπων εξαρτάται από την τιμή που προβλέπει το γραμμικό μοντέλο παλινδρόμησης. Κατά συνέπεια, τα κατάλοιπα δεν εμφανίζουν τυχαιότητα και μπορούν να χρησιμοποιηθούν για να προβλέψουν την τιμή της ανεξάρτητης μεταβλητής κι έτσι έχουν προβλεπτική δύναμη – κάτι το οποίο δεν θα έπρεπε να συμβαίνει. Μία τέτοια απεικόνιση των καταλοίπων δεν τεκμηριώνει το γραμμικό μοντέλο παλινδρόμησης.



Εικόνα 0.14 *Διάγραμμα καταλοίπων που δεν τεκμηριώνει τη γραμμική υπόθεση ενός γραμμικού μοντέλου παλινδρόμησης. Το διάγραμμα εμφανίζει τα κατάλοιπα να έχουν μία δομή ή μορφή και να σχετίζονται με τις τιμές της εξαρτημένης μεταβλητής – κάτι που αντίκειται στη στοχαστική τους φύση.*

Η παραβίαση της υπόθεσης γραμμικότητας ενός θεωρούμενου γραμμικού μοντέλου παλινδρόμησης είναι πολύ σοβαρή κατάσταση. Σηματοδοτεί το γεγονός ότι δεν μπορεί να τεκμηριωθεί το γραμμικό μοντέλο και θα πρέπει να εξεταστεί το ενδεχόμενο να συλληφθεί η σχέση μεταξύ των μεταβλητών αυτών με ένα μη-γραμμικό μοντέλο.

- ***Έλεγχος ομοσκεδαστικότητας καταλοίπων***

Με τον όρο ομοσκεδαστικότητα (homoscedasticity) νοείται η ιδιότητα μεταβλητών ενός συνόλου δεδομένων να έχουν την ίδια διασπορά/διακύμανση. Η απουσία τέτοιας ιδιότητας, δηλαδή όταν η διασπορά /διακύμανση των μεταβλητών είναι διαφορετική καλείται ετεροσκεδαστικότητα. Προκειμένου να μπορεί να τεκμηριωθεί ένα γραμμικό μοντέλο παλινδρόμησης θα πρέπει τα κατάλοιπα να παρουσιάζουν ομοσκεδαστικότητα εάν απεικονιστούν ως προς τις προβλεφθείσες τιμές του μοντέλου, το οποίο ερμηνεύεται ως ότι θα πρέπει να παρουσιάζουν την ίδια διακύμανση για τις διάφορες τιμές της εξαρτημένης μεταβλητής.

Υπάρχουν εξειδικευμένοι στατιστικοί έλεγχοι για την ομοσκεδαστικότητα των καταλοίπων. Ωστόσο συνηθίζεται η ομοσκεδαστικότητα να ελέγχεται διαγραμματικά, με τη χρήση του διαγράμματος καταλοίπων. Ο τρόπος με τον οποίο ελέγχεται η ομοσκεδαστικότητα των καταλοίπων είναι να απεικονιστούν τα κατάλοιπα ως προς τις προβλεφθείσες από το μοντέλο τιμές της εξαρτημένης μεταβλητής και εκτελώντας μία παλινδρόμηση[23] μεταξύ των τιμών των καταλοίπων και των τιμών που προβλέπει το μοντέλο. Εάν η γραμμή ζυγισμένης παλινδρόμησης είναι οριζόντια, αυτό σημαίνει ότι τα κατάλοιπα παρουσιάζουν την ίδια διακύμανση για όλες τις τιμές της ανεξάρτητης μεταβλητής και κατά συνέπεια παρουσιάζουν ομοσκεδαστικότητα. Αν η γραμμή παλινδρόμησης αποκλείνει από οριζόντια ευθεία, αυτό σημαίνει ότι τα κατάλοιπα παρουσιάζουν ετεροσκεδαστικότητα δηλαδή έχουν διαφορετική διακύμανση για τις διάφορες τιμές της ανεξάρτητης μεταβητής.

[23] Συγκεκριμένα εκτελείται μία μορφή παλινδρόμησης που καλείται ζυγισμένη παλινδρόμηση ελαχίστων τετραγώνων (locally weighted regression). Αυτή η μορφή της παλινδρόμησης προσπαθεί να ταιριάξει τιμές της ανεξάρτητης μεταβλητής που βρίσκονται κοντά η μία στην άλλη, ενώ επιχειρεί να αγνοήσει τιμές που βρίσκονται μακρυά. Το «ζυγισμένη» αναφέρεται στο γεγονός ότι η συνάρτηση κόστους εισάγει ένα βάρος (για κάθε διαφορά πραγματικής και προβλεφθείσας τιμής) το οποίο επιλέγεται με τρόπο ώστε να λαμβάνει διαφορετικές τιμές ανάλογα με το πόσο κοντά βρίσκονται οι τιμές της ανεξάρτητης μεταβλητής. Λαμβάνει την τιμή μηδέν για τιμές που βρίσκονται πολύ μακρυά.

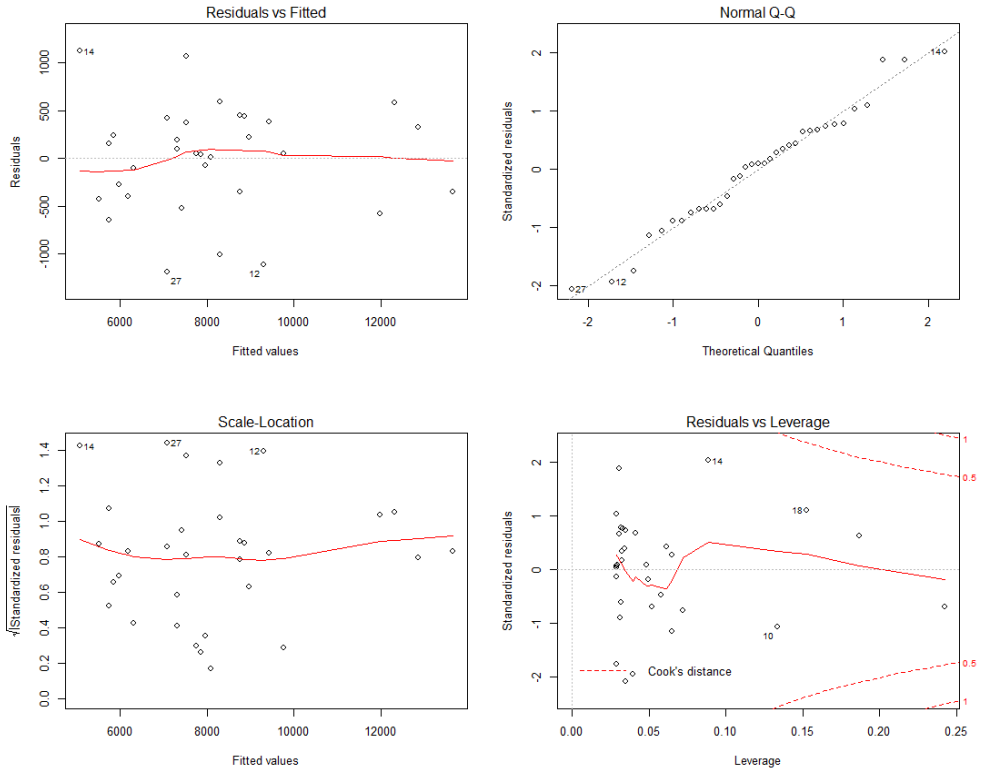
Στην R, ο διαγραμματικός έλεγχος ομοσκεδαστικότητας των καταλοίπων για το γραμμικό μοντέλο παλινδρόμησης

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_0$$

μπορεί να γίνει με τον κάτωθι κώδικα

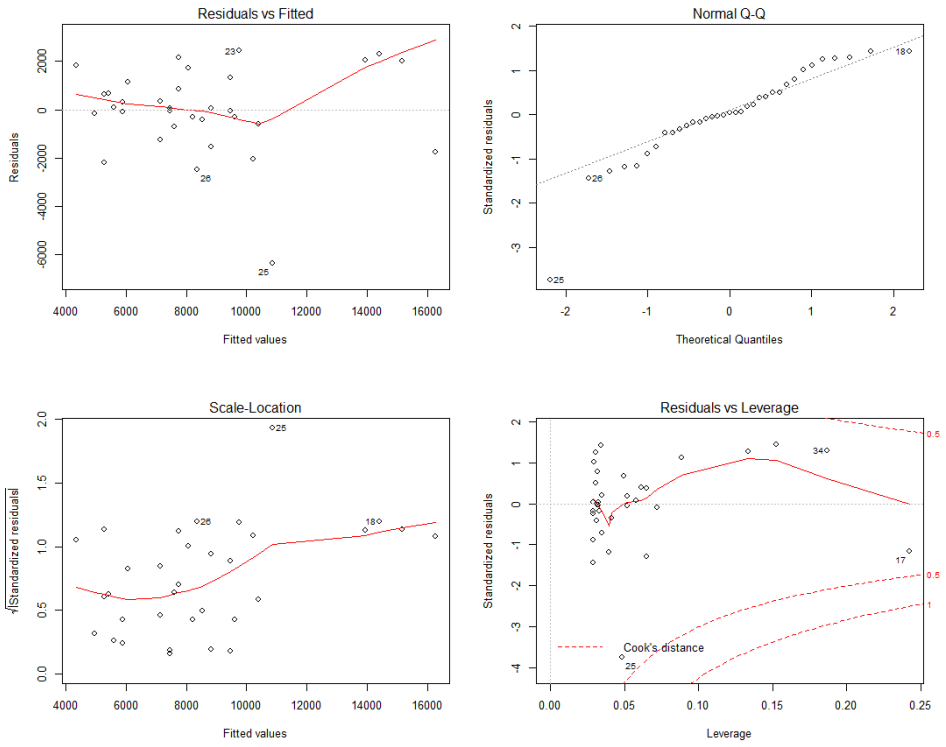
```
foodConsumptionData<-read.csv("HouseholdData. csv", sep=",", head-
er=T)

# Εκτίμηση συντελεστών του απλού γραμμικού μοντέλου παλινδρόμησης
- με μία ανεξάρτηση μεταβλητή

# Κατανάλωση τροφίμων = β1Εισόδημα + β0

# με τη χρήση της συνάρτησης lm() που βασίζεται στη μέθοδο των ελα-
χίστων τετραγώνων (OLS).

linear.regression.model<-lm(FoodExpenditure ~ Income, da-
ta=foodConsumptionData)

# Έλεγχος ομοσκεδαστικότητας καταλοίπων με τη βοήθεια του διαγράμ-
ματος καταλοίπων

# Η εντολή par() επιτρέπει τον συνδυασμό και την συνεμφάνιση πολλών
διαγραμμάτων.

# Η παρακάτω κλήση της par() δημιουργεί μία μήτρα με 2 γραμμές και
2 στήλες, όπου

# σε κάθε θέση/κελί της μήτρας θα απεικονιστεί ένα διαφορετικό διά-
γραμμα.

# Η συνάρτηση par() αλλάζει τις καθολικές ρυθμίσεις/προτιμήσεις των
συναρτήσεων σχεδιασμού όπως της

# plot() που θα χρησιμοποιηθεί παρακάτω.

par( mfrow=c(2,2) )

#Θα απεικονιστούν 4 διαγράμματα. Το διάγραμμα που εμφανίζεται πάνω
αριστερά απεικονίζει τα κατάλοιπα

#της γραμμικής παλινδρόμησης ως προς τις προβλεφθείσες τιμές του
μοντέλου. Επιπλέον, στο διάγραμμα θα

#εμφανιστεί και η ζυγισμένη παλινδρόμηση των καταλοίπων.

plot(linear.regression.model)
```

Το αποτέλεσμα του παραπάνω κώδικα R για το σύνολο εκπαίδευσης HouseholdData.csv φαίνεται στο σχήμα 6.15. Το διάγραμμα που εμφανίζεται στην πάνω αριστερή γωνία απεικονίζει το διάγραμμα καταλοίπων ως προς τις τιμές της ανεξάρτητης μεταβλητής που προβλέπει το μοντέλο. Στο ίδιο διάγραμμα εμφανίζεται και η γραμμή της ζυγισμένης παλινδρόμησης η οποία αποτελεί ένδειξη για το εάν τα κατάλοιπα εμφανίζουν ομοσκεδαστικότητα ή όχι. Επειδή η γραμμή ζυγισμένης παλινδρόμησης τείνει να είναι ευθεία, τα κατάλοιπα του γραμμικού μοντέλου παλινδρόμησης έχουν την ίδια διακύμανση για τις διάφορες τιμές της ανεξάρτητης μεταβλητής και κατά συνέπεια παρουσιάζουν ομοσκεδαστικότητα.



Εικόνα 0.15 *Έλεγχος ομοσκεδαστικότητας καταλοίπων που τεκμηριώνει τη γραμμική υπόθεση του μοντέλου.Η γραμμή ζυγισμένης παλινδρόμησης είναι ευθεία που σημαίνει ότι τα κατάλοιπα παρουσιάζουν την ίδια διακύμανση για τις διάφορες τιμές της εξαρτημένης μεταβλητής.*

Εάν τα κατάλοιπα παρουσιάζουν ετεροσκεδαστικότητα, το διάγραμμα καταλοίπων θα λάβει τη μορφή της πάνω αριστερής εικόνας του σχήματος 6.16. Η γραμμή ζυγισμένης παλινδρόμησης θα αποκλίνει σημαντικά από οριζόντια γραμμή, υποδεικνύοντας ότι η διακύμανση των καταλοίπων δεν είναι η ίδια για τις διάφορες τιμές της ανεξάρτητης μεταβλητής και κατά συνέπεια παρουσιάζουν ετεροσκεδαστικότητα.



Εικόνα 0.16 *Έλεγχος ομοσκεδαστικότητας καταλοίπων που δεν τεκμηριώνει τη γραμμική υπόθεση του μοντέλου. Η γραμμή ζυγισμένης παλινδρόμησης δεν είναι ευθεία.*

Εάν τα κατάλοιπα ενος γραμμικού μοντέλου παλινδρόμησης παρουσιάζουν ετεροσκεδαστικότητα, αυτό θα κάνει δύσκολη την εκτίμηση του τυπικού σφάλματος των τιμών που προβλέπονται από το μοντέλο. Ωστόσο, υπάρχουν τρόποι για να αντιμετωπιστεί η ετεροσκεδαστικότητα των καταλοίπων, αλλάζοντας τις προδιαγραφές του μοντέλου παλινδρόμησης. Ειδικότερα, εάν η τιμή της ανεξάρτητης

μεταβλητής είναι θετική και τα κατάλοιπα είναι ανάλογα του μεγέθους της ανεξάρτητης μεταβλητής (δηλαδή το σφάλμα, όπως αυτό εκφράζεται από τα κατάλοιπα, είναι σταθερό ως ποσοστό και όχι ως απόλυτη τιμή) τότε μπορεί να γίνει ένας λογαριθμικός μετασχηματισμός της εξαρτημένης μεταβλητής, αφήνοντας το υπόλοιπο γραμμικό μοντέλο ως έχει.

Άσκηση Αυτοαξιολόγησης 0.15

Με τον όρο λογαριθμικός μετασχηματισμός της εξαρτημένης μεταβλητής για την αντιμετώπιση της ετεροσκεδαστικότητας των καταλοίπων, τί ακριβώς εννοείται;

- ***Έλεγχος κανονικής κατανομή των καταλοίπων***

Τα κατάλοιπα ενός γραμμικού μοντέλου παλινδρόμησης θα πρέπει να είναι κανονικά κατανεμημένα ανεξάρτητα από το πως είναι κατανεμημένες η εξαρτημένη και οι ανεξάρτητες μεταβλητές. Η υπόθεση της κανονικής κατανομής των καταλοίπων απορρέι από το Κεντρικό Οριακό Θεώρημα. Αν και εδώ υπάρχουν στατιστικοί έλεγχοι για το εάν τα κατάλοιπα ακολουθούν την κανονική κατανομή συνηθίζεται και εδώ ο έλεγχος να γίνεται διαγραμματικά με τη χρήση ενός διαγράμματος δειγματικών ποσοστημορίων που καλείται Normal Q-Q plot (από το Quantile-Quantile Plot). Γενικά, ένα διάγραμμα ποσοστημορίων χρησιμοποιείται όταν επιθυμείται να συγκριθούν δύο κατανομές. Αυτό που κάνει ένα τέτοιο διάγραμμα είναι να συγκρίνει τα ποσοστημόρια δύο κατανομών και εξετάζει αν αυτά συμπίπτουν ή όχι. Στην περίπτωση του ελέγχου της κατανομής των καταλοίπων, τα ποσοστημόρια των καταλοίπων θα συγκριθούν με τα ποσοστημόρια της (υποθετικής και θεωρητικής) κανονικής κατανομής. Εάν η γραφική παράσταση ακολουθεί την ευθεία y=x, αυτό σημαίνει ότι τα κατάλοιπα ακολουθούν μία κανονική κατανομή. Αν η γραφική παράσταση αποκλείνει από την ευθεία y=x αυτό σημαίνει ότι τα κατάλοιπα δεν ακολουθούν την κανονική κατανομή.

Το διάγραμμα των δειγματικών ποσοστημορίων του μοντέλου παλινδρόμησης κατανάλωσης τροφίμων εμφανίζεται στο σχήμα 6.15 στο διάγραμμα που βρίσκεται στην πάνω δεξιά γωνία. Στον οριζόντιο άξονα εμφανίζονται τα θεωρητικά ποσοστημόρια της κανονικής κατανομής, ενώ στον κάθετο άξονα τα ποσοστημόρια των καταλοίπων. Η διακεκομμένη γραμμή σηματοδοτεί την περίπτωση y=x και στην οποία θα πρέπει να πέφτουν οι τιμές αν τα κατάλοιπα ακολουθούν την κα-

νονική κατανομή. Από το διάγραμμα διαφαίνεται ότι τα κατάλοιπα ακολουθούν την γραμμή y=x και κατά συνέπεια έχουν κανονική κατανομή.

Αν τα κατάλοιπα δεν ακολουθούν κανονική κατανομή, αυτό μπορεί να είναι αποτέλεσμα πολλών διαφορετικών αιτιών. Μπορεί να οφείλεται στο γεγονός ότι η υπόθεση γραμμικότητας δεν στέκει το οποίο μπορεί να αντιμετωπιστεί είτε με ένα μη-γραμμικό μοντέλο είτε με έναν λογαριθμικό μετασχηματισμό του υπάρχοντος γραμμικού μοντέλου. Μπορεί επίσης να οφείλεται στα χαρακατηριστικά του συνόλου εκπαίδευσης και και ειδικά όταν υποσύνολά του έχουν διαφορετικά στατιστικά χαρακτηριστικά. Σε μία τέτοια περίπτωση θα πρέπει είτε να χρησιμοποιηθούν διαφορετικά μοντέλα παλινδρόμησης για τα υποσύνολα αυτά είτε να αφαιρεθούν κάποιες παρατηρήσεις από το σύνολο εκπαίδευσης, αν αυτό είναι εφικτό.

❖ ***Έλεγχος πολυσυγγραμμικότητας***

Η υπόθεση που γίνεται σε ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης είναι ότι κάθε μία από τις ανεξάρτητες μεταβλητές που εμφανίζονται σε αυτό «συνεισφέρουν» στην τιμή της εξαρτημένης μεταβλητής και ότι η επίδραση των επιμέρους ανεξάρτητων στην τιμή της εξαρτημένης μεταβλητής είναι αθροιστική. Επιπλέον, μία άλλη σημαντική υπόθεση που γίνεται είναι οτι οι ανεξάρτητες μεταβλητές που εμφανίζονται σε ένα μοντέλο γραμμικής παλινδρόμησης δεν πρέπει να σχετίζονται μεταξύ τους. Αυτό σημαίνει ότι η τιμή μιας ανεξάρτητης μεταβλητής δεν σχετίζεται με την τιμή καμίας άλλης ανεξάρτητης μεταβλητής που εμφανίζονται και έτσι δεν μπορεί να εκτιμηθεί - με ορισμένη ακρίβεια - από τις τιμές των άλλων ανεξάρτητων μεταβλητών ενός πολλαπλού μοντέλου γραμμικής παλινδρόμησης. Εάν η τιμή μιας ανεξάρτητης μεταβλητής σχετίζται με τις τιμές μιας ή περισσοτέρων άλλων ανεξάρτητων μεταβλητών τότε λέγεται ότι το μοντέλο παλινδρόμησης παρουσιάζει πολυσυγγραμμικότητα (multicollinearity). Η πολυσυγγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών πρέπει να αποφεύγεται ειδικά όταν ο στόχος του μοντέλου είναι η εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής επειδή i) εισάγει πελονασμό στο μοντέλο παλινδρόμησης και ii) επηρεάζει τις τιμές των εκτιμώμενων συντελεστών κάτι το οποίο δεν επιτρέπει την ασφαλή εξαγωγή συμπερασμάτων για τον τρόπο με τον οποίο οι ανεξάρτητες μεταβλητές επηρεάζουν την τιμή της εξαρτημένης μεταβλητής.

Ειδικότερα, όταν ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης παρουσιάζει πολυσυγγραμμικότητα μεταξύ των ανεξάρτητων μεταβλητών αυτό θα οδηγήσει σε αύξηση της διακύμανσης των εκτιμώμενων συντελεστών που σχετίζονται με τις ανεξάρτητες που παρουσιάζουν πολυσυγγραμικότητα. Ωστόσο, υπάρχουν τρόποι να ελεγχθεί η πολυσυγγραμιότητα των ανεξάρτητων μεταβλητών ενός μοντέλου. Ο έλεγχος πολυσυγγραμμικότητας μπορεί να γίνει με διάφορους τρόπους κι ένας από τους πιο δημοφιλής είναι ο υπολογισμός του Εκτιμητή Διόγκωσης της Διακύμανσης (Variance Inflation Factor – VIF). Ο εκτιμητής αυτός υπολογίζει κατά πόσο θα αυξηθεί η διακύμανση ενός εκτιμώμενου συντελεστή εάν η αντίστοιχη ανεξάρτητη μεταβλητή παρουσιάζει πολυσυγγραμικότητα.

Ο τρόπος με τον οποίο υπολογίζεται ο εκτιμητής VIF είναι να εκτελείται μία γραμμική παλινδρόμηση θέτοντας κάθε ανεξάρτητη μεταβλητή του πολλαπλού μοντέλου γραμμικής παλινδρόμησης ως την εξαρτημένη μεταβητή και όλες τις υπόλοιπες ως τις ανεξάρτητες. Με αυτόν τον τρόπο υπολογίζει μία τιμή VIF για κάθε μία από τις ανεξάρτητες μεταβλητές που συμμετέχουν στο πολλαπλό μοντέλο γραμμικής παλινδρόμησης, η οποία ποσοτικοποιεί τη συσχέτιση της συγκεκριμένης μεταβλητής με τις άλλες. Η ελάχιστη τιμή που μπορεί να λάβει ο δείκτης VIF είναι 1 (ένα) η οποία σηματοδοτεί πλήρης απουσία πολυσυγγραμμικότητας, ενώ μπορεί να λάβει οποιαδήποτε τιμή. Όσο πιο υψηλή είναι η τιμή VIF για μία ανεξάρτητη μεταβλητή του μοντέλου παλινδρόμησης, τόσο μεγαλύτερη η συσχέτιση της μεταβλητής αυτής με κάποια (ή κάποιες) από τις άλλες ανεξάρτητες μεταβλητές. Η τιμή VIF εδικότερα εκφράζει πόσο πιο μεγάλη είναι η διακύμανση του εκτιμώμενου συντελεστή, εξαιτίας της πολυσυγγραμμικότητας, αν συγκριθεί με την περίπτωση απουσίας πολυσυγγραμμικότητας. Για παράδειγμα, αν η τιμή VIF για μία ανεξάρτηση μεταβλητή είναι τρία (3), αυτό σημαίνει ότι η διακύμανση του αντίστοιχου συντελεστή είναι τρεις (3) φορές μεγαλύτερη από την περίπτωση που δεν εμφανίζεται πολυσυγγραμμικότητα. Συνήθως, το ποια τιμή του δείκτη VIF θεωρείται προβληματική εξαρτάται από το προς μελέτη πρόβλημα. Ωστόσο, υπάρχουν εμπειρικοί κανόνες για την ερμηνεία των τιμών του δείκτη VIF που σηματοδοτούν τον βαθμό πολυσυγγραμμικότητας. Έτσι, αν ο δείκτης VIF για μία μεταβλητή είναι μεγαλύτερος από 10, αυτό σηματοδοτεί ισχυρή πολυσυγγραμμικότητα του μοντέλου και απαιτεί τη λήψη συγκεκριμένων μέτρων. Μία τιμή του δείκτη VIF μεγαλύτερη από 4 σηματοδοτεί βαθμό πολυσυγγραμμικότητας που θα πρέπει να μελετηθεί περαιτέρω. Τιμή VIF μικρότερη του 4

σηματοδοτεί ασθενή βαθμό πολυσυγγραμμικότητας που είναι ανεκτή και δεν απαιτεί την αντιμετώπισή της. Σε μοντέλα γραμμικής παλινδρόμησης, ασθενής βαθμός πολυσυγγραμμικότητας των ανεξάρτητων μεταβλητών είναι ανεκτός.

Παρακάτω φαίνεται πως ο έλεγχος πολυσυγγραμμικότητας ενός μοντέλου παλινδρόμησης μπορεί να γίνει στην R, με χρήση της συνάρτησης vif() που υπολογίζει τον δείκτη VIF πολυσυγγραμμικότητας για το πολλαπλό μοντέλο γραμμικής παλινδρόμησης

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού + \beta_0$$

```
# Η βιλιοθήκη car παρέχει τη συνάρτηση vif() που υπολογίζει τον
εκτιμητή διόγκωσης της διακύμανσης VIF

library(car)

foodConsumptionData<-read.csv("HouseholdData.csv", sep=",", head-
er=T)

linear.regression.model<-lm(FoodExpenditure ~ Income + FamilySize,
data=foodConsumptionData)

# Υπολογισμός δείκτη VIF για κάθε ανεξάρτητη μεταβλητή του πολλα-
πλού μοντέλου παλινδρόμησης.

vif( linear.regression.model )
```

Η εκτέλεση του παραπάνω κώδικα R θα δώσει το εξής αποτέλεσμα

```
   Income     FamilySize

   2.05552    2.05552
```

Η τιμή 2.05552 που εμφανίζεται κάτω από τη μεταβλητή Income είναι ο βαθμός πολυσυγγραμμικότητας (δείκτης VIF) της συγκεκριμένης μεταβλητής με τις υπόλοιπες του μοντέλου. . Επειδή το γραμμικό μοντέλο παλινδρόμησης έχει μόνο δύο ανεξάρτητες μεταβλητές, εμφανίζεται και η ίδια τιμή VIF για την μεταβλητή FamilySize, αφού η πολυσυγγραμμικότητα είναι συμμετρική. Όπως φαίνεται, η τιμή του δείκτη VIF είναι μικρότερη από 4 κι αυτό σημαίνει ότι υπάρχει χαμηλός βαθμός πολυσυγγραμμικότητας κι έτσι δεν αποτελεί πρόβλημα για το μοντέλο παλινδρόμησης.

Αντιθέτως, αν εισαχθεί ως επιπλέον ανεξάρτητη μεταβλητή η μεταβλητή χρόνια εκπαίδευσης του επικεφαλής του νοικοκυριού (YearsOfEducationHH) έτσι ώστε η γραμμική παλινδρόμηση να λάβει τη μορφή

$$\begin{aligned} &Κατανάλωση\ τροφίμων \\ &\quad = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού \\ &\quad + \beta_3 Χρόνια\ εκπαίδευσης\ επικεφαλή + \beta_0 \end{aligned}$$

αν τυπολογιστούν με τον ίδιο τρόπο οι δείκτες VIF των τριών μεταβλητών, θα προκύψει προκύπτει το εξής αποτέλεσμα:

```
Income        FamilySize     YearsOfEducationHH
13.550411     2.290961       10.372593
```

Εδώ οι τιμές δείχνουν ότι για τις μεταβλητές Income και YearsOfEducationHH υπάρχει υψηλός βαθμός πολυσυγγραμμικότητας αφού παρουσιάζουν υψηλό δείκτη VIF (>10). Αυτό σημαίνει ότι οι δύο αυτές ανεξάρτητες μεταβλητές συσχετίζονται και επηρεάζει η μία μεταβλητή την τιμή της άλλης. Στο συγκεκριμένο μοντέλο κάτι τέτοιο είναι απολύτως λογικό: υπάρχει έρευνα που τεκμηριώνει με στοιχεία, ότι τα χρόνια εκπαίδευσης ενός ατόμου συσχετίζονται θετικά με τον μισθό του.

Σε περίπτωση που ο βαθμός πολυσυγγραμμικότητας είναι μεγάλος σε ένα μοντέλο παλινδρόμησης, υπάρχουν διάφοροι τρόποι αντιμετώπισης. Μπορούν να αφαιρεθούν από το μοντέλο παλινδρόμησης οι ανεξάρτητες μεταβλητές που έχουν μεγάλη τιμή του δείκτη VIF (στο παραπάνω παράδειγμα θα μπορούσε να αφαιρεθεί η μεταβλητή YearsOfEducationHH). Επίσης, μπορεί να αυξηθεί το σύνολο εκπαίδευσης με επιπλέον παρατηρήσεις. Τέτοια αύξηση του συνόλου εκπαίδεσης μπορεί να μειώσει σημαντικά τον βαθμό πολυσυγγραμμικότητας.

6.7.1.2 Αξιολόγηση και ερμηνεία γραμμικού μοντέλου παλινδρόμησης

Με τον όρο αξιολόγηση ενός γραμμικού μοντέλου παλινδρόμησης εννοείται η διαδικασία ελέγχου των αποτελεσμάτων της εκτίμησης συντελεστών ενός γραμμικού μοτνέλου παλινδρόμησης προκειμένου να αποφανθεί πόσο καλά το μοντέλο πετυχαίνει τον σκοπό της εξήγησης της διακύμανσης της τιμής της εξεταρτημένης μεταβλητής.

Όταν η εκτίμηση ενός μοντέλου παλινδρόμησης έχει ως στόχο την εξήγηση και ερμηνεία της διακύμανσης της τιμής της εξαρτημένης μεταβλητής, το κεντρικό σημείο της αξιολόγησης είναι η μελέτη του κατά πόσο οι επιλεγμένες ανεξάρτητες μεταβλητές του μοντέλου παλινδρόμησης επιτυγχάνουν αυτόν ακριβώς τον στόχο. Έτσι, σε μία τέτοια περίπτωση η αξιολόγηση εστιάζει κυρίως σε δύο πτυχές: 1) Πόσο καλά το γραμμικό μοντέλο που έχει εκτιμηθεί μπορεί να εξηγήσει τη διακύμανση της εξαρτημένης μεταβλητής στο σύνολο δεδομένων εκπαίδευσης και 2) εαν κάθε μεταβλητή που εμφανίζεται στο γραμμικό μοντέλο έχει όντως συσχέτιση με την τιμή της εξαρτημένης μεταβλητής ή όχι και κατά συνέπεια εάν μεταβολές στην τιμή της ανεξάρτητης μεταβλητής έχει σημαντική επίδραση στην τιμή της εξαρτημένης.

❖ ***Συντελεστής προσδιορισμού R^2***

Η αξιολόγηση του πόσο καλά ένα μοντέλο γραμμικής παλινδρόμησης εξηγεί τη διακύμανση της τιμής της εξαρτημένης μεταβλητής γίνεται με υπολογισμό του συντελεστή προσδιορισμού (coefficient of determination) R^2 (R-squared)[24]. Ο συντελεστής αυτός για ένα μοντέλο γραμμικής παλινδρόμησης λαμβάνει τιμές από το 0 έως 1 και εκφράζει το ποσοστό της διακύμανσης που κατορθώνει να εξηγήσει το γραμμικό μοντέλο παλινδρόμησης. Ο συντελεστής R^2 δεν έχει μονάδες μέτρησης καθότι εκφράζει ποσοστό. Αν η τιμή του συντελεστή προσδιορισμού R^2 για μοντέλο γραμμικής παλινδρόμησης είναι για παράδειγμα 0.48 αυτό σημαίνει ότι το γραμμικό μοντέλο παλινδρόμησης είναι ικανό να εξηγήσει το 48% της διακύμανσης της εξαρτημένης μεταβλητής στο σύνολο εκπαίδευσης. Γενικά, όσο πιο υψηλή η τιμή του συντελεστή R^2 τόσο καλύτερα επιτυγχάνει το μοντέλο να εξηγήσει τη διακύμανση της ανεξάρτητης μεταβλητής.

Ο τρόπος με τον οποίο υπολογίζεται ο συντελεστής προσδιορισμού R^2 για ένα γραμμικό μοντέλο παλινδρόμησης είναι να συγκριθεί το άθροισμα τετραγώνων των καταλοίπων του συγκεκριμένου μοντέλου (που επί της ουσίας μετρά τη διακύμανση που δεν μπορεί να εξηγήσει το μοντέλο) με τη διακύμανση που δεν μπορεί να εξηγήσει ένα μοντέλο παλινδρόμησησης βάσης. Το μοντέλο παλινδρόμησης βάσης με το οποίο συγκρίνεται ένα μοντέλο παλινδρόμησης που έχει εκτιμηθεί είναι εκείνο, το οποίο δεν έχει καμία ανεξάρτητη μεταβλητή και η τιμή της

[24] Εμφανίζεται και με τον όρο «συντελεστής πολλαπλού προσδιορισμού (coefficient of multiple determination)» για μοντέλο πολλαπλής γραμμικής παλινδρόμησης.

ανεξάρτητης μεταβλητής είναι σταθερή και ίση με τη μέση τιμή της, όπως αυτό απορρέει από το σύνολο δεδομένων εκπαίδευσης. Επί της ουσίας, αυτό που συμβαίνει για τον υπολογισμό του συντελεστή προσδιορισμού R^2 είναι να εξεταστεί εάν η εισαγωγή των ανεξάρτητων μεταβλητών που εμφανίζονται στο μοντέλο παλινδρόμησης ερμηνεύουν καλύτερα τη διακύμανση της εξαρτημένης μεταβλητής απ' ό,τι ένα μοντέλο παλινδρόμησης το οποίο δεν έχει καμία ανεξάρτητη μεταβλητή.

Ο συντελεστής προσδιορισμού R^2 υπολογίζεται από τον παρακάτω τύπο

$$R^2 = 1 - \frac{SS_{res}}{SS_{total}}$$

όπου SS_{res} (Sum of Squares of residuals) το άθροισμα τετραγώνων των καταλοίπων του γραμμικού μοντέλου και SS_{total} (total Sum of Squares) το άθροισμα τετραγώνων των καταλοίπων ενός μοντέλου, όπου δεν εμφανίζεται καμία ανεξάρτητη μεταβλητή και η τιμή της εξαρτημένης είναι ίση με τη μέση τιμή της εξαρτημένης μεταβλητής στο σύνολο εκπαίδευσης που αποτελεί και το μοντέλο βάσης. Στον παραπάνω τύπο, η τιμή SS_{res} επί της ουσίας μετρά τη διακύμανση που δεν μπορεί να εξηγήσει το μοντέλο παλινδρόμησης ενώ η τιμή SS_{total} τη διακύμανση που δεν μπορεί να εξηγήσει το μοντέλο βάσης το οποίο δεν περιέχει καμία ανεξάρτητη μεταβλητή.

Χρησιμοποιώντας το περιβάλλον R, ο συντελεστής προσδιορισμού R^2 υπολογίζεται αυτόματα με τη εκτίμηση των συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης και η τιμή του μπορεί να προσπελαστεί μέσω της εντολής summary() όπως φαίνεται στον παρακάτω κώδικα.

```
foodConsumptionData<-read.csv("HouseholdData.csv", sep=",", head-
er=T)

# Μοντέλο γραμμικής παλινδρόμησης με τις μεταβλητές Εισόδημα και
Πληθος ατόμων νοικοκυριού ως

# ανεξάρτητες μεταβλητές.

linear.regression.model<-lm(FoodExpenditure ~ Income + FamilySize,
data=foodConsumptionData)

# Εμφάνιση του συντελεστή προσδιορισμού R-squared για το παραπάνω
μοντέλο γραμμικής παλινδρόμησης
```

```
summary( linear.regression.model )$r.squared
```

Η εκτέλεση του παραπάνω κώδικα R θα δώσει ως τιμή του συντελεστή προσδιορισμού R^2

```
[1] 0.9381515
```

το οποίο σημαίνει ότι το πολλαπλό μοντέλο γραμμικής παλινδρόμησης μπορεί να εξηγήσει το 93.8% της διακύμανσης της μεταβλητής Εισόδημα στο σύνολο εκπαίδευσης.

Ειδικά για ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης συνηθίζεται αντί για τον συντελεστή προσδιορισμού, να αναφέρεται ο διορθωμένος ή προσαρμοσμένος συντελεστής προσδιορισμού (Adjusted R^2) που συμβολίζεται με $\bar{R}^2$ ή R^2_{adj}. Το ζήτημα που υπάρχει με τη χρήση του συντελεστή προσδιορισμού R^2 είναι ότι η τιμή του πάντα θα αυξάνεται αν προστεθούν κι άλλες ανεξάρτητες μεταβλητές στο μοντέλο, ανεξάρτητα εάν αυτές οι μεταβλητές όντως έχουν νόημα να προστεθούν στο μοντέλο ή όχι. Η τιμή του συντελεστή προσδιορισμού δεν μειώνεται ποτέ με την εισαγωγή μιας νέας (οποιασδήποτε) ανεξάρτητης μεταβλητής στο μοντέλο και κατά συνέπεια μπορεί να δώσει την εσφαλμένη εντύπωση ότι το μοντέλο παλινδρόμησης βελτιώνεται την εξήγηση και είναι δικαιολογημένη την εισαγωγή της νέας μεταβλητής. Προκειμένου να αντιμετωπιστεί τέτοιο ζήτημα σε ένα πολλαπλό μοντέλο γραμμικής παλινδρόμησης, στον υπολογισμό του συντελεστή προσδιορισμού λαμβάνεται υπόψη τόσο το πλήθος των ανεξάρτητων μεταβλητών του μοντέλου όσο και το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης. Τέτοιος τρόπος υπολογισμού του συντελεστή προσδιορισμού χαρακτηρίζει τον διορθωμένο ή προσαρμοσμένο συντελεστή προσδιορισμού, ο οποίος υπολογίζεται από τον τύπο

$$R^2_{adj} = 1 - \left(\frac{(1 - R^2)(n - 1)}{(n - k - 1)}\right)$$

όπου R^2 ο συντελεστής προσδιορισμού, n το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης και k το πλήθος των ανεξάρτητων μεταβλητών στο πολλαπλό μοντέλο γραμμικής παλινδρόμησης. Ο τύπος αυτός έχει το χαρακτηριστικό, καθώς αυξάνει το πλήθος των ανεξάρτητων μεταβλητών στο μοντέλο k, ο διορθω-

μένος συντελεστής προσδιορισμού θα μειωθεί, εκτός εάν οι νέες προσθήκες ανεξάρτητων μεταβλητών σημαντικά βελτιώνουν τον συντελεστή προσδιορισμού R^2 . Στο περιβάλλον της R, η τιμή του διορθωμένου συντελεστή προσδιορισμού μπορεί να ανακτηθεί μέσω της εντολής summary() και της μεταβλητής $adj.squared ως ακολούθως

```
foodConsumptionData<-read.csv("HouseholdData.csv", sep=",", head-
er=T)

# Μοντέλο γραμμικής παλινδρόμησης με τις μεταβλητές Εισόδημα και
Πληθος ατόμων νοικοκυριού ως

# ανεξάρτητες μεταβλητές.

linear.regression.model<-lm(FoodExpenditure ~ Income + FamilySize,
data=foodConsumptionData)

# Εμφάνιση του διορθωμένου συντελεστή προσδιορισμού (adjusted R-
squared) για το παραπάνω μοντέλο

# πολλαπλής γραμμικής παλινδρόμησης

summary( linear.regression.model ) $adj.r.squared
```

Η εκτέλεση του παραπάνω κώδικα R θα δώσει ως τιμή του διορθωμένου συντελεστή προσδιορισμού R^2

```
[1] 0.9342859
```

Ο διορθωμένος συντελεστής προσδιορισμού θα είναι πάντα μικρότερος από τον συντελεστή προσδιορισμού R^2 .

Το ποια τιμή του συντελεστή προσδιορισμού R^2 καταστά ένα γραμμικό μοντέλο παλινδρόμησης να πετυχαίνει τον στόχο του και κατά συνέπεια χρήσιμο για την εξαγωγή συμπερασμάτων, εξαρτάται κυρίως από τη φύση του προβλήματος που μελετάται. Για παράδειγμα εάν το μοντέλο γραμμικής παλινδρόμησης χρησιμοποιείται στα πλαίσια των κοινωνικών επιστημών (όπως π.χ. οικονομικά, κοινωνιολογία, ψυχολογία κλπ) τότε μία τιμή του συντελεστή (ή διορθωμένου συντελεστή) προσδιορισμού R^2 ίση με 0.3 (δηλαδή όταν το μοντέλο εξηγεί το 30% της διακύμανσης) θεωρείται καλή και αποδεκτή για οποιαδήποτε περαιτέρω ανάλυση. Αντιθέτως, στις θετικές επιστήμες, μία τιμή του συντελεστή προσδιορισμού

κάτω από 0.7 εκλαμβάνεται ως αποτυχία του μοντέλου να επιτελέσει τον στόχο του, δηλαδή έχει μικρή ερμηνευτική δύναμη της διακύμανσης της εξαρτημένης μεταβλητής. Η μεγάλη αυτή διαφορά οφείλεται κυρίως στο γεγονός ότι στις κοινωνικές επιστήμες, που αντικείμενό τους είναι η ανθρώπινη συμπεριφορά, αυτή παρουσιάζει πολύ μεγαλύτερη διακύμανση καθότι επηρεάζεται από πολλές μεταβλητές που μπορεί να μην έχουν ή να μην μπορούν να συλληφθούν. Στις θετικές επιστήμες, όπου οι συνθήκες για παράδειγμα πειραμάτων μπορούν να ελεγχθούν και να ποσοτικοποιηθούν με την επιθυμητή ακρίβεια, αναμένεται ένα γραμμικό μοντέλο παλινδρόμησης να μπορεί να εξηγήσει μεγάλο μέρος της διακύμανσης.

Άσκηση Αυτοαξιολόγησης 0.16

Κατά την εκτίμηση των συντλεστών ενός γραμμικού μοντέλου παλινδρόμησης, προέκυψε μία τιμή του συντελεστή προσδιορισμού R^2 ίση με 0.05. Πως ερμηνεύεται/τί σημαίνει η τιμή αυτή;

Άσκηση Αυτοαξιολόγησης 0.17

Σε ποια περίπτωση η τιμή του συντελεστή προσδιορισμού R^2 είναι ίση με την τιμή του διορθωμένου συντελεστή προσδιορισμού (Adjusted R^2);

- ***Στατιστική σημαντικότητα μεταβλητών***

Εφόσον έχει αξιολογηθεί η ερμηνευτική ικανότητα της διακύμανσης του γραμμικού μοντέλου παλινδρόμησης, πρέπει ακολούθως να μελετηθεί εάν κάθε μία ανεξάρτητη μεταβλητή που εμφανίζεται στο μοντέλο έχει επίδραση και πόσο σημαντική, στην τιμή της εξαρτημένης μεταβλητής. Αυτό κατά συνέπεια θα δείξει εάν υπάρχει πράγματι μία σχέση μεταξύ εκάστης των ανεξάρτητων και της εξαρτημένης μεταβλητής. Αυτό που βασικά πρέπει να δειχθεί είναι ότι η παρατηρηθείσα σχέση μεταξύ ανεξάρτητων και εξαρτημένης μεταβλητής, και η οποία συλλαμβάνεται από τους συντελεστές που έχουν εκτιμηθεί, έχει πολύ μικρή πιθανότητα να οφείλεται σε τύχαίους παράγοντες. Εάν μία σχέση μεταξύ ανεξάρτητης και εξαρτημένης μεταβλητής έχει πολύ μικρή πιθανότητα να οφείλεται στην τύχη, τότε λέγεται ότι η συγκεκριμένη ανεξάρτητη μεταβλητή είναι στατιστικά σημαντική (statistical significant). Ο έλεγχος της στατιστικής σημαντικότητας των ανεξάρτητων μεταβλητών ενός μοντέλου παλινδρόμησης γίνεται εξετάζοντας την

πιθανότητα εμφάνισης των εκτιμημένων τιμών των συντελεστών αυτών εάν υποτεθεί ότι ισχύουν συγκεκριμένες υποθέσεις.

Ο έλεγχος για τη στατιστική σημαντικότητα των ανεξάρτητων μεταβλητών γίνεται για κάθε έναν συντελεστή ξεχωριστά, εκτελώντας έναν έλεγχο υποθέσεων. Συγκεκριμένα η υπόθεση που ελέγχεται για κάθε εκτιμημένο συντελεστή ξεχωριστά είναι η πιθανότητα να ληφθεί η τιμή του συντελεστή που έχει εκτιμηθεί, εάν ισχύει η υπόθεση ότι η πραγματική τιμή του συντελεστή είναι ίση με το μηδέν (0). Δηλαδή η βασική υπόθεση που γίνεται για κάθε ανεξάρτητη μεταβλητή είναι ότι αυτή δεν έχει καμία επίπτωση στη διακύμανση της εξαρτημένης μεταβλητής, το οποίο σημαίνει ότι η πραγματική τιμή του αντίστοιχου συντελεστή είναι ίση με το μηδέν Αυτό που επι της ουσίας ελέγχεται είναι πόσο καλά υποστηρίζει τη μηδενική αυτή υπόθεση η τιμή του συντελεστή που εκτιμήθηκε. Αυτή η υπόθεση αποτελεί τη μηδενική υπόθεση H_0 του ελέγχου, με την εναλλακτική υπόθεση H_1 να είναι ότι η πραγματική τιμή του συντελεστή δεν είναι μηδέν. Η ιδέα είναι ότι εάν η πιθανότητα εμφάνισης της συγκεκριμένης τιμής του συντελεστή είναι πολύ μικρή υπό την υπόθεση ότι η πραγματική του τιμή είναι μηδέν (0), τότε απουσιάζουν τα στοιχεία να υποστηριχθεί η μηδενική υπόθεση αυτή. Σε μία τέτοια περίπτωση λέγεται ότι απορρίπτεται η μηδενική υπόθεση ότι ο συντελεστής δεν έχει τιμή ίση με το μηδέν και κατά συνέπεια η αντίστοιχη ανεξάρτητη μεταβλητή είναι στατιστικά σημαντική.

Επειδή ωστόσο η κατανομή των συντελεστών που έχουν εκτιμηθεί είναι άγνωστη, υπολογίζεται η τιμή της στατιστικής συνάρτησης ελέγχου της κατανομής Student για την τιμή του συντελεστή και εκτιμάται η πιθανότητα να ληφθεί η τιμή αυτή ή ακόμη πιο ακραία, εάν ισχύει η μηδενική υπόθεση. Η πιθανότητα εμφάνισης της τιμής αυτής ή πιο ακραίας είναι γνωστή με τον όρο p-τιμή (p-value). Εάν η p-τιμή ενός συντελεστή είναι μικρότερη από μία κρίσιμη τιμή, τότε η μηδενική υπόθεση απορρίπτεται. Η κρίσιμη τιμή στην οποία λέγεται ότι η μηδενική υπόθεση απορρίπτεται καλείται επίπεδο σημαντικότητας (significance level) και συνήθως ορίζεται ως 0.05 ή 5%. Αυτό σημαίνει ότι εάν η p-τιμή ενός συντελεστή έχει τιμή μικρότερη από 0.05 εάν ισχύει η μηδενική υπόθεση (δηλαδή η πιθανότητα εμφάνισής της τιμής αυτής είναι μικρότερη από 5%), τότε απορρίπτεται η μηδενική υπόθεση ότι η πραγματική τιμή του συντελεστή είναι μηδέν και η αντίστοιχη τιμή του συντελεστή λέγεται ότι είναι στατιστικά σημαντική και κατ' επέκταση ότι η αντίστοιχη ανεξάρτητη μεταβλητή είναι στατιστικά σημαντική. Σε μία

τέτοια περίπτωση αυτό σημαίνει ότι η ανεξάρτητη μεταβλητή έχει όντως μία επίδραση σημαντική στη διακύμανση της εξαρτημένης μεταβλητής. Εάν η p-τιμή μιας μεταβλητής έχει τιμή μεγαλύτερη από 0.05 τότε η συγκεκριμένη ανεξάρτητη μεταβλητή δεν είναι στατιστικά σημαντική, που σημαίνει ότι δεν έχει σημαντική επίδραση στην τιμή της εξαρτημένης μεταβλητής.

Στο περιβάλλον της R, ο υπολογισμός των p-τιμών για κάθε συντελεστή γίνεται αυτόματα και μπορεί να εμφανιστεί όπως φαίνεται παρακάτω

```
foodConsumptionData<-read.csv("HouseholdData.csv", sep=",", head-
er=T)

# Μοντέλο γραμμικής παλινδρόμησης με τις μεταβλητές Εισόδημα και
Πληθος ατόμων νοικοκυριού ως

# ανεξάρτητες μεταβλητές.

linear.regression.model<-lm(FoodExpenditure ~ Income + FamilySize,
data=foodConsumptionData)

# Εμφάνιση του διορθωμένου συντελεστή προσδιορισμού (adjusted R-
squared) για το παραπάνω μοντέλο

# πολλαπλής γραμμικής παλινδρόμησης

summary( linear.regression.model )$coefficients
```

Η εκτέλεση του παραπάνω κώδικα θα εμφανίσει τους συντελεστές (Estimate) και για κάθε συντελεστή την τιμή της στατιστικής ελέγχου της κατανομής Student (στήλη t value) καθώς και την p-τιμή (στήλη Pr(>|t|)):

```
               Estimate   Std. Error   t value     Pr(>|t|)
(Intercept) 1182.1064194 8.268801e+02  1.429598 1.625234e-01
Income         0.1223607 7.275833e-03 16.817412 1.910288e-17
FamilySize   325.7171067 1.533442e+02  2.124091 4.148596e-02
```

Από τα παραπάνω αποτελέσματα διαφαίνεται ότι οι p-τιμές των συντελεστών Income και FamilySize έχουν τμή μικρότερη από αποδεκτό επίπεδο σημαντικότητας του 5% (1.190288e-17 και 4.148596e-02 αντίστοιχα) και κατά συνέπεια οι

ανεξάρτητες μεταβλητές αυτές είναι στατιστικά σημαντικές δηλαδή έχουν σημαντική επίδραση στην διακύμανση της εξαρτημένης μεταβλητής FoodExpenditure.

Άσκηση Αυτοαξιολόγησης 0.18

Εάν η p-τιμή είναι ίση με 0.03, τί συμπέρασμα μπορείτε να βγάλετε για τον συντελεστή, εάν το επίπεδο σημαντικότητας είναι 5%;

Άσκηση Αυτοαξιολόγησης 0.19

Εάν η p-τιμή είναι ίση με 0.02 τί συμπέρασμα μπορείτε να βγάλετε για την αντίστοιχη ανεξάρτητη μεταβλητή, εάν το επίπεδο σημαντικότητας είναι 1% ;

Δραστηριότητα 0.4

Δίνεται το παρακάτω σύνολο δεδομένων:

Χρόνος προθέρμανηςης	0	30	10	15	5	25	35	40
Τραυματισμοί	4	1	2	2	3	1	0	1

Συγγράψτε κώδικα R, ο οποίος υπολογίζει τις απαντήσεις στα εξής ερωτήματα:

(α) Εκτιμήστε τους συντελεστές του γραμμικού μοντέλου παλινδρόμησης

$$\text{Τραυματισμοί} = \beta_1 \text{ Χρόνος Προθέρμανσης} + \beta_0$$

με τη μέθοδο των ελαχίστων τετραγώνων. Εμφανίστε τους συντελεστές.

(β) Τί ποσοστό της διακύμανσης της μεταβλητής Τραυματισμοί μπορεί να εξηγήσει η διακύμανση της μεταβλητής Χρόνος προθέρμανσης;

(γ) Η μεταβλητή Χρόνος προθέρμανσης είναι στατιστικά σημαντική στο επίπεδο σημαντικότητας 5% ή όχι;

- ***Ερμηνεία συντελεστών γραμμικού μοντέλου παλινδρόμησης***

Οι συντελεστές που εμφανίζονται και έχουν αξιολογηθεί σε ένα μοντέλο γραμμικής παλινδρόμησης, εκφράζουν την μεταβολή, που θα συμβεί κατά μέσο όρο,

στην τιμή της εξαρτημένης μεταβλητής εάν μία ανεξάρτητη μεταβλητή αυξηθεί κατά μία μονάδα και οι υπόλοιπες παραμείνουν σταθερές. Η μεταβολή της εξαρτημένης μεταβλητής εκφράζεται στις μονάδες μέτρησης της μεταβλητής. Το σημαντικό εδώ είναι να τονιστεί ότι το μοντέλο γραμμικής παλινδρόμησης δεν θα υπολογίσει με ακρίβεια πόση θα είναι η μεταβολή της εξαρτημένης μεταβλητής αλλά ποιά θα είναι η μεταβολή της κατά μέσο όρο.

Έτσι, για το γραμμικό μοντέλο παλινδρόμησης της κατανάλωσης τροφίμων του οποίου εκτιμήθηκαν οι συντελεστές και αξιολογήθηκε παραπάνω και έλαβε τη μορφή

$$FoodExpenditure = 0.1223607 * Income + 325.7171067 * FamilySize + 1182.1064194$$

οι συντελεστές του ερμηνεύονται ως ακολούθως: εάν τα μέλη μιας οικογένεια αυξηθούν κατά ένα (1) άτομο (δηλαδή η ανεξάρτητη μεταβλητή FamilySize αυξηθεί κατά ένα) και όλες οι υπόλοιπες ανεξάρτητες μεταβλητές διατηρήσουν σταθερές τις τιμές τους, η κατανάλωση τροφίμων θα αυξηθεί, κατά μέσο όρο, κατά 325.7171067 Ευρώ. Παρομοίως, εάν το εισόδημα αυξηθεί κατά ένα (1) Ευρώ, τότε η κατανάλωση τροφίμων θα αυξηθεί, κατά μέσο όρο, κατά 0.1223607 Ευρώ εάν οι τιμές των υπολοίπων ανεξάρτητων μεταβλητών παραμείνουν σταθερές.

Το πρόσημο του συντελεστή καθορίζει εάν μία μεταβολή στην τιμή της εξαρτημένης μεταβλητής οδηγήσει σε αύξηση ή μείωση της τιμής της εξαρτημένης μεταβλητής.

Άσκηση Αυτοαξιολόγησης 0.20

Πώς πιστεύεται ότι μπορεί να ερμηνευτεί η τιμή του σταθερού όρου που εκτιμήθηκε στο μοντέλο παλινδρόμησης της ενότητας 6.7.1.2.3 (και που είναι ίση με β0=1182.1064194);

6.7.2 Γραμμικά μοντέλα γραμμικής παλινδρόμησης με στόχο την πρόβλεψη

Γραμμικά μοντέλα παλινδρόμησης που δημιουργούνται με στόχο την πρόβλεψη αποσκοπούν κυρίως στο να εκτιμήσουν, με όση μεγαλύτερη ακρίβεια γίνεται, την

τιμή της εξαρτημένης μεταβλητής για άγνωστες, νέες τμές των ανεξάρτητων μεταβλητών. Εδώ ο όρος «άγνωστες. νεες τιμές» χρησιμοποιείται για να υποδηλώσει τιμές των ανεξαρτήτων μεταβλητών, οι οποίες δεν εμφανίζονται στο σύνολο εκπαίδευσης.

Εάν ο στόχος του μοντέλου είναι η πρόβλεψη, ζητήματα όπως η διερεύνηση του ακριβούς ρόλου που παίζει κάθε μία από τις ανεξάρτητες μεταβλητές στην τιμή της εξαρτημένης μεταβλητής, είναι δευτερευούσης σημασίας. Η έμφαση στην ακρίβεια της πρόβλεψης και όχι στον ρόλο που παίζει κάθε μεταβλητή, σημαίνει ότι έλεγχοι που είναι απαραίτητοι όταν ο στόχος είναι η εξήγηση της διακύμανσης, δεν χρειάζεται να εκτελεστούν. Έτσι για παράδειγμα, έλεγχοι όπως η τεκμηρίωση της γραμμικότητας ή η αποφυγή της πολυσυγγραμμικότητας, που επιχειρούν να εξασφαλίσουν σημαντικές προϋποθέσεις του μοντέλου όταν ο στόχος είναι η εξήγηση, δεν απαιτείται να γίνονται στην περίπτωση της πρόβλεψης εφόσον οι τιμές που προβλέπει το μοντέλο είναι ικανοποιητικές. Σε μια τέτοια περίπτωση η έμφαση είναι στην αξιολόγηση της πρόβλεψης δηλαδή πόσο καλά το μοντέλο προβλέπει τις τιμές της εξαρτημένης μεταβλητής για άγνωστες τιμές των ανεξαρτήτων μεταβλητών. Εάν μια επιπλέον ανεξάρτητη μεταβλητή στο μοντέλο βελτιώνει την πρόβλεψη, δεν έχει σημασία αν η προθήκη της μεταβλητής αυτής μπορεί να τεκμηριωθεί ή όχι.

6.7.2.1 Αξιολόγηση και ερμηνεία γραμμικού μοντέλου παλινδρόμησης με στόχο τη πρόβλεψη

Η αξιολόγηση ενος μοντέλου παλινδρόμησης με στόχο τη πρόβλεψη γίνεται με τη χρήση ορισμένων μετρικών οι οποίες εκτιμούν πόσο κοντά στην πραγματική τιμή βρίσκεται η τιμή που προβλέπει το μοντέλο παλινδρόμησης. Οι μετρικές αυτές επι της ουσίας μετρούν το σφάλμα (που σηματοδοτείται από τη διαφορά μεταξύ προβλεπόμενης και πραγματικής τιμής της εξαρτημένης μεταβλητής) που κάνει το μοντέλο παλινδρόμησης στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής και γι’ αυτό οι μετρικές αυτές καλούνται και μετρικές σφάλματος. Παρακάτω παρουσιάζονται οι πιο συχνά χρησιμοποιούμενες μετρικές σφάλματος για την αξιολόγηση μοντέλων παλινδρόμησης.

- ***Μετρικές εκτίμησης σφάλματος πρόβλεψης***

Οι μετρικές εκτίμησης του σφάλματος πρόβλεψης ενός γραμμικού μοντέλου παλινδρόμησης εξετάζουν κυρίως τις πραγματικές τιμές της εξαρτημένης μεταβλη-

τής y, όπως αυτή προκύπτει από το σύνολο δεδομένων, και την τιμή που υπολογίζει το μοντέλο παλινδρόμησης $\hat{y}$ για τις αντίστοιχες τιμές των ανεξαρτήτων μεταβλητών. Οι πιο συχνές μετρικές για την εκτίμηση του σφάλματος της πρόβλεψης είναι οι εξής – στους παρακάτω τύπους, ο όρος n σηματοδοτεί το πλήθος των παρατηρήσεων με τους οποίους συγκρίνεται η πρόβλεψη:

Μέσο Απόλυτο Σφάλμα (Mean Absolute Error - MAE): Υπολογίζει τον μέσο όρο του μεγέθους του σφάλματος, που ορίζεται ως η απόλυτη τιμή της διαφοράς μεταξύ πραγματικής και προβλεπόμενης τιμής της εξαρτημένης μεταβλητής. Η διαφορά αυτή είναι επί της ουσίας τα κατάλοιπα. Το Μέσο Απόλυτο Σφάλμα δίνεται από τον τύπο:

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Μέσο Τετραγωνισμένο Σφάλμα (Mean Squred Error – MSE): Υπολογίζει το σφάλμα ως τον μέσο όρο του αρθοίσματος του τετραγώνου των καταλοίπων

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

Μέσο Τετραγωνικό Σφάλμα (Root Mean Squared Error – RMSE): Υπολογίζει το σφάλμα ως την τετραγωνική ρίζα του μέσου όρου του τετραγώνου των καταλοίπων

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

Μέσο Απόλυτο Εκατοστιαίο Σφάλμα (Mean Absolute Percentage Error – MAPE): Υπολογίζει το σφάλμα με τη μορφή ποσοστού του Μέσου Απόλυτου Σφάλματος (MAE):

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

Όλες οι παραπάνω μετρικές δίνουν τιμές σφάλματος μεταξύ 0 και άπειρο με εξαίρεση τη μετρική του Μέσου Απόλυτου Εκατοστιαίου Σφάλματος που κυμαίνε-

ται μεταξύ 0 και 1. Προφανώς, όσο πιο μικρή η τιμή του σφάλματος τόσο πιο καλή η πρόβλεψη που κάνει το μοντέλο. Η επιλογή της κατάλληλης μετρικής εξαρτάται από το πρόβλημα που μελετάται καθότι οι μετρικές αυτές χειρίζονται τα σφάλματα πρόβλεψης με διαφορετικό τρόπο.

Το Μέσο Απόλυτο Σφάλμα δίνει την ίδια βαρύτητα σε όλα τα κατάλοιπα που προκύπτουν και δεν διακρίνει εάν η προβλεπόμενη τιμή υπερτιμά ή όχι την πραγματική τιμή ή όχι. Το Μέσο Τετραγωνισμένο Σφάλμα δίνει διαφορετική βαρύτητα στο μέγεθος των καταλοίπων με μεγαλύτερα κατάλοιπα συνεισφέρουν στο συνολικό σφάλμα παραπάνω εξαιτίας της ύψωσης στο τετράγωνο και προτιμάται εάν επιθυμείται να τιμωρηθούν μεγάλες τιμές καταλοίπων και ακραίες προβλεπόμενες τιμές. Το Μέσο Τετραφωνικό Σφάλμα από την άλλη έχει καλύτερη ερμηνευτική δύναμη καθότι το σφάλμα εκφράζεται σε μονάδες της ανεξάρτητης μεταβλητής – σε αντίθεση με το Μέσο Τετραγωνισμένο Σφάλμα. Τέλος παρόλο που το Μέσο Απόλυτο Εκατοστιαίο Σφάλμα έχει εύκολη ερμηνεία τα προβλήματα που παρουσιάζει είναι ότι δεν μπορεί να χρησιμοποιηθεί εάν η εξαρτημένη μεταβλητή μπορεί να λάβει την τιμή μηδέν (0) εξαιτίας της πράξης της διαίρεσης που εμφανίζεται. Επιπλέον, η μετρική αυτή μεροληπτεί υπέρ προβλεπόμενων τιμών που είναι συστηματικά μικρότερες από την πραγματική τιμή. Ειδικότερα, η μετρική αυτή θα είναι μικρότερη όπου οι προβλεπόμενες τιμές είναι μικρότερες από την πραγματική εάν συγκριθεί με άλλο μοντέλο το οποίο παρουσιάζει τιμές που προβλέπει τιμές μεγαλύτερες από τις πραγματικές κατά το ίδιο μέγεθος. Τα ζητήματα αυτά που παρουσιάζει η μετρική του Μέσου Απόλυτου Εκατοστιαίου Σφάλματος την κάνει τη λιγότερη δημοφιλής μεταξύ των σχετικών μετρικών.

Άσκηση Αυτοαξιολόγησης 0.21

Αναφέρθηκε ότι το μέσο τετραγωνισμένο σφάλμα (Mean Squared Error) χρησιμοποιείται ότι επιθυμείται να τιμωρηθούν μεγάλες τιμές καταλοίπων και ακραίες τιμές. Εξηγείστε γιατί συμβαίνει αυτό.

- ***Μεθοδολογία αξιολόγησης ακρίβειας πρόβλεψης μοντέλου παλινδρόμησης.***

Για την αξιολόγηση της πρόβλεψης, απαιτείται ένα σύνολο δεδομένων με το οποίο μπορούν να συγκριθούν οι πραγματικές τιμές της εξαρτημένης μεταβλητής και οι τιμές που προβλέπει το μοντέλο γι’ αυτήν, με τον τρόπο που ορίζουν οι

μετρικές που παρουσιάστηκαν παραπάνω. Ένα τέτοιο σύνολο με το οποίο ελέγχεται και αξιολογείται η ακρίβεια της πρόβλεψης ενός μοντέλου καλείται σύνολο ελέγχου (test set). Υπάρχουν διάφοροι τρόποι με τους οποίους μπορεί να δημιουργηθεί ένα τέτοιο σύνολο.

Ένας τρόπος δημιουργίας τέτοιου συνόλου ελέγχου είναι, να επιλεγεί με τυχαίο τρόπο ένα πλήθος παρατηρήσεων (ένα υποσύνολο) από ίδιο σύνολο δεδομένων το οποίο χρησιμοποιήθηκε για την εκτίμηση των συντελεστών του μοντέλου παλινδρόμησης. Με το υποσύνολο αυτό αξιολογείται η πρόβλεψη του μοντέλου με μία από τις μετρικές εκτίμησης σφάλματος που παρουσιάστηκαν παραπάνω. Ο τρόπος ελέγχου του μοντέλου με ένα τέτοιο σύνολο καλείται εντός-δείγματος έλεγχος (in-sample testing) καθότι τα δεδομένα του συνόλου ελέγχου (το υποσύνολο ή δείγμα) προέρχονται από το ίδιο σύνολο με το οποίο εκτιμήθηκαν οι συντελεστές. Το σφάλμα του μοντέλου που υπολογίζεται από αυτό το σύνολο καλείται και σφάλμα εκπαίδευσης (training error). Η βασική υπόθεση που γίνεται σε τέτοιο έλεγχο είναι, ότι το σφάλμα εκπαίδευσης που θα υπολογιστεί είναι μία καλή εκτίμηση του σφάλματος που θα προκύψει εάν το ίδιο μοντέλο χρησιμοποιηθεί για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής για νέα/άγνωστα δεδομένα. Όπως θα συζητηθεί αργότερα, η υπόθεση αυτή δεν είναι πάντα αξιόπιστη και η γενίκευση ενός μοντέλου μπορεί να οδηγεί σε πολύ μεγαλύτερα σφάλματα πρόβλεψης από τα σφάλματα εκπαίδευσης.

Ένας δεύτερος τρόπος επιχειρεί να αποφύγει την επικάλυψη των συνόλων εκπαίδευσης και ελέγχου. Ο τρόπος αυτός δίνει έμφαση στην εξασφάλιση άγνωστων δεδομένων, δηλαδή δεδομένων που δεν έχει «δει»/επεξεργαστεί το μοντέλο κατά τη διαδικασία εκτίμησης των συντελεστών από το σύνολο εκπαίδευσης. Το σφάλμα ενός μοντέλου στην πρόβλεψη της τιμή της εξαρτημένης μεταβλητής από ένα άγνωστο σύνολο ελέγχου καλείται και σφάλμα γενίκευσης του μοντέλου (generalization error). Στην ιδανική περίπτωση ένα τέτοιο σύνολο άγνωστων δεδομένων ελέγχου θα πρέπει συλλεχθεί εκ νέου πέραν από το υπάρχον διαθέσιμο σύνολο δεδομένων. Ωστόσο, κάτι τέτοιο σε πολλές περιπτώσεις δεν είναι είναι δυνατό ή εφικτό να πραγματοποιηθεί.

Σε τέτοιες περιπτώσεις συνηθίζεται να γίνεται ένας διαχωρισμός, μία τμηματοποίηση του αρχικού διαθέσιμου συνόλου των δεδομένων σε δύο, ξένα μεταξύ τους, υποσύνολα: ένα υποσύνολο που θα χρησιμοποιηθεί για την εκτίμηση των

συντελεστών του γραμμικού μοντέλου παλινδρόμησης (δηλαδή την εκπαίδευση του μοντέλου) και που καλείται σύνολο εκπαίδευσης (training set) και ένα άλλο υποσύνολο, τα δεδομένα του οποίου δεν εμφανίζονται στο σύνολο εκπαίδευσης και είναι άγνωστα για το μοντέλο, με το οποίο θα γίνει η αξιολόγηση της προβλεπτικής δύναμης του μοντέλου παλινδρόμησης και θα παίξει τον ρόλο του σύνολο ελέγχου. Τέτοιος τρόπος ελέγχου του μοντέλου παλινδρόμησης ανήκει στην κατηγορία ελέγχου γνωστή ως διασταυρωμένη επικύρωση (Cross Validation - CV).

Η μέθοδος της διασταυρωμένης επικύρωσης έρχεται σε διάφορες εκδοχές. Μία εκδοχή της, η μέθοδος γνωστή με το όνομα μέθοδος παρακράτησης (holdout method), ολόκληρο το αρχικό σύνολο δεδομένων χωρίζεται τυχαία σε δύο ξένα μεταξύ τους υποσύνολα: ένα υποσύνολο που θα παίξει τον ρόλο του συνόλου εκπαίδευσης και ένα υποσύνολο που θα λειτουργήσει ως σύνολο ελέγχου. Σε τέτοιες περιπτώσεις συνηθίζεται το 70% ή 80% του πλήθους των παρατηρήσεων στο αρχικό σύνολο δεδομένων να χρησιμοποιείται για την εκτίμηση των συντελεστών (σύνολο εκπαίδευσης) με το υπόλοιπο 30% ή 20% αντίστοιχα να χρησιμοποιείται για τον έλεγχο του μοντέλου (σύνολο ελέγχου). Το σφάλμα πρόβλεψης που θα προκύψει για το σύνολο ελέγχου θεωρείται μία εκτίμηση για το σφάλμα γενίκευσης του μοντέλου.

Από τις πιο δημοφιλείς μεθόδους ελέγχου της ακρίβειας ενός μοντέλου παλινδρόμησης είναι μία εκδοχή της διασταυρωτικής επικύρωσης που καλείται διασταυρωμένη επικύρωση k-πτυχών (k-fold Cross Validation) που έχει ως κύριο στόχο να κάνει μία ακόμη καλύτερη εκτίμηση του σφάλματος πρόβλεψης του μοντέλου για άγνωστα δεδομένα. Η μέθοδος της διασταυρωμένης επικύρωσης k-πτυχών παρουσιάζεται αναλυτικότερα στην επόμενη ενότητα.

Άσκηση Αυτοαξιολόγησης 0.22

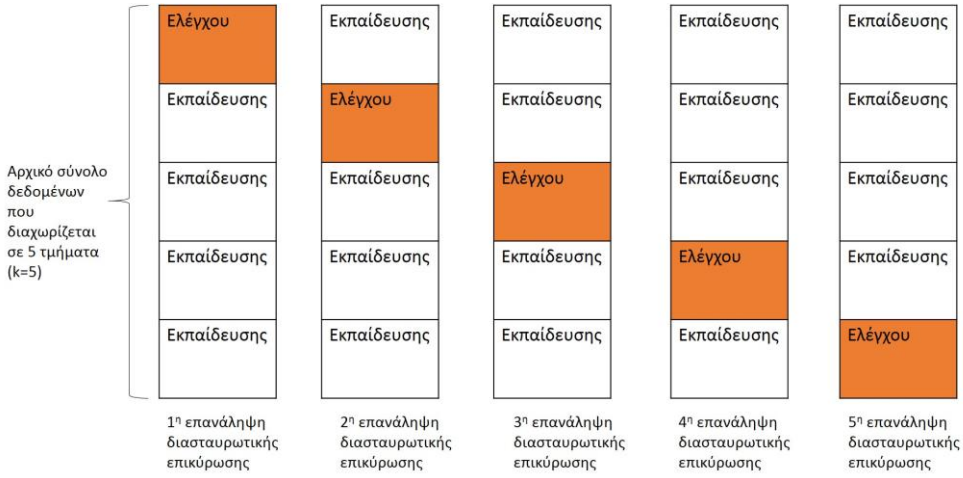
Ο τρόπος εκτίμησης των συντελεστών (μέθοδος Ελαχίστων Τετραγώνων ή Σταδιακής Καθόδου) που χρησιμοποιήθηκε για ένα μοντέλο παλινδρόμησης επηρεάζει τον τρόπο αξιολόγησης του μοντέλου, εάν ο στόχος είναι η πρόβλεψη;

- ***Διασταυρωμένη Επικύρωση k-Πτυχών (k-fold Cross Validation).***

Η διασταυρωμένη επικύρωση k-πτυχών επιχειρεί να πετύχει μια ακόμη καλύτερη εκτίμηση του σφάλματος γενίκευσης του μοντέλου παλινδρόμησης, με το να δη-

μιουργεί πολλά διαφορετικά σύνολα ελέγχου από το ίδιο, αρχικά διαθέσιμο σύνολο δεδομένων. Κατά τη μέθοδο αυτή, ολόκληρο το σύνολο δεδομένων διαχωρίζεται σε k ξένα υποσύνολα ή τμήματα (ή πτυχές), με κάθε ένα απ'ο αυτά να έχει περίπου ίσο αριθμό παρατηρήσεων. Ακολούθως, σε μία επαναληπτική διαδικασία, επιλέγεται διαδοχικά κάθε ένα από τα k αυτά τμήματα ως σύνολο ελέγχου με τα υπόλοιπα k-1 να παίζουν τον ρόλο του συνόλου εκπαίδευσης και με τα οποία θα εκτιμηθούν οι συντελεστές του μοντέλου. Μετά από κάθε εκτίμηση συντελεστών χρησιμοποιώντας το σύνολο εκπαίδευσης, αξιολογείται η ακρίβεια του μοντέλου κάνοντας χρήση το ένα τμήμα που δεν λήφθηκε υπόψη κατά την εκτίμηση των συντελεστών. Για την εκτίμηση του σφάλματος χρησιμοποιείται μία από τις μετρικές που παρουσιάστηκαν. Η διαδικασία αυτή επαναλαμβάνεται για το ίδιο μοντέλο παλινδρόμησης έως ότου κάθε ένα από τα k τμήματα να έχει λειτουργήσει ως σύνολο ελέγχου και τα υπόλοιπα ως σύνολο εκπαίδευσης. Για ένα μοντέλο παλινδρόμησης θα προκύψουν έτσι k σε πλήθος τιμές σφαλμάτων (ένα για κάθε τμήμα ελέγχου) και η τελική εκτίμηση για το σφάλμα του μοντέλου μπορεί να υπολογιστεί ως ο μέσος όρος των σφαλμάτων για κάθε τμήμα που προέκυψαν. Η υπόθεση που γίνεται είναι ότι η τιμή αυτή είναι μία πολύ καλύτερη εκτίμηση του σφάλματος του μοντέλου εάν προβλέψει τιμές για άγνωστα δεδομένα. Από την τιμή αυτή μπορεί να εξεταστεί εάν το συγκεκριμένο γραμμικό μοντέλο παλινδρόμησης έχει την επιθυμητή ακρίβεια στην πρόβλεψη άγνωστων τιμών. Στο σχήμα 6.17 φαίνεται ο τρόπος με τον οποίο προχωρά τον έλεγχο η διασταυρωμένη επικύρωση k-πτυχών.

Η τιμή k για το πλήθος των τμημάτων είναι παράμετρος της μεθόδου και καθορίζεται στην αρχή της διαδικασίας. Χαρακτηριστικές τιμές για την παράμετρο k είναι 5 ή 10. Στην περίπτωση όπου k=2 τότε το σύνολο δεδομένων χωρίζεται σε δύο μόνο τμήματα εκπαίδευσης και ελέγχου με τα οποία ελέγχεται και εκπαιδεύεται το μοντέλο και καθίσταται όμοιο με τη μέθοδο παρακράτησης. Εάν k=n, όπου n το πλήθος των παρατηρήσεων στο αρχικό σύνολο δεδομένων, τότε το τμήμα που χρησιμοποιείται ως σύνολο ελέγχου θα περιέχει μόνο μία παρατήρηση και ο τρόπος αυτός καλείται διασταυρωμένη επικύρωση με παράλειψη ενός (Leave-one-out cross validation - LOOCV). Ο τρόπος αυτός χρησιμοποιείται εάν το πλήθος παρατηρήσεων στο σύνολο δεδομένων είναι μικρό.



Εικόνα 0.17 *Διασταυρωμένη επικύρωση 5-πτυχών όπου k=5 που σημαίνει ότι το αρχικό σύνολο δεδομένων θα χωριστεί σε 5 τμήματα. Σε κάθε επανάληψη χρησιμοποιείται διαφορετικό τμήμα δεδομένων για έλεγχο και τα υπόλοιπα για την εκπαίδευση του ίδιου μοντέλου. Μετά από κάθε διαδικασία ελέγχου, θα υπολογιστεί με την επιλεγμένη μετρική ακρίβειας το σφάλμα. Η διαδικασία τερματίζει εάν κάθε ένα από τα 5 τμήματα έχει χρησιμοποιηθεί ως σύνολο ελέγχου.*

Η διασταυρωμένη επικύρωση k-πτυχών χρησιμοποιείται και σε περιπτώσεις για να αξιολογηθούν διαφορετικά μοντέλα παλινδρόμησης ως προς την ακρίβεια της πρόβλεψής τους. Έτσι για παράδειγμα, εάν υπάρχουν δύο (ή και παραπάνω) υποψήφια γραμμικά μοντέλα παλινδρόμησης για την πρόβλεψη της ίδιας εξαρτημένης μεταβλητής, όπου κάθε ένα από τα μοντέλα αυτά χρησιμοποιεί διαφορετικές ανεξάρτητες μεταβλητές, μπορεί να γίνει η χρήση της διασταυρωμένης επικύρωσης k-πτυχών προκειμένου να επιλεγεί εκείνο το μοντέλο που κάνει τις πιο ακριβείς προβλέψεις. Μετά την διασταυρωμένη επικύρωση των υποψήφιων μοντέλων, επιλέγεται εκείνο το μοντέλο με το μικρότερο σφάλμα ως το καλύτερο και το τελικό γραμμικό μοντέλο εκτιμάται χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων ως σύνολο εκπαίδευσης.

Παρακάτω φαίνεται ο κώδικας σε R, ο οποίος εκτελεί διασταυρωμένη επικύρωση k-πτυχών προκειμένου να αξιολογηθούν τέσσερα υποψήφια γραμμικά μοντέλα παλινδρόμησης, τα οποία επιχειρούν να προβλέψουν την κατανάλωση καυσίμου αυτοκινήτων. Το σύνολο δεδομένων που χρησιμοποιείται για την εκπαίδευση και τον έλεγχο υπάρχει στο αρχείο auto-mpg.csv και περιέχει στοιχεία διαφόρων ο-

χημάτων όπως η ιπποδύναμη, το βάρος, τον κυβισμό κ.α. Το αρχείο με τα δεδομένα εκπαίδευσης και η περιγραφή των μεταβλητών μπορεί να βρεθεί από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/auto+mpg [25]. Ο στόχος της διαδικασίας αυτής είναι να βρεθεί εκείνο το μοντέλο που προβλέπει την κατανάλωση με τη μεγαλύτερη ακρίβεια. Τα υποψήφια μοντέλα παλινδρόμησης που αξιολογούνται με τη μέθοδο της διασταυρωμένης επικύρωσης k-πτυχών ως προς την ικανότητα πρόβλεψης της κατανάλωσης είναι τα εξής:

$$Κατανάλωση\ καυσίμου\ ανα\ μίλι = \beta_1 Ιπποδύναμη + \beta_2 Βάρος + \beta_0$$

$$Κατανάλωση\ καυσίμου\ ανα\ μίλι = \beta_1 Ιπποδύναμη + \beta_2 Επιτάχυνση + \beta_3 Κυβισμός + \beta_0$$

$$Κατανάλωση\ καυσίμου\ ανά\ μίλι = \beta_1 Ιπποδύναμη + \beta_2 Κυβισμός + \beta_3 Βάρος + \beta_0$$

$$Κατανάλωση\ καυσίμου\ ανά\ μίλι = \beta_1 Ιπποδύναμη + \beta_2 Κυβισμός + \beta_3 Βάρος + \beta_4 Βάρος^2 + \beta_0$$

Κάθε ένα από τα παραπάνω γραμμικά μοντέλα παλινδρόμησης θα ελεγχθεί με τη μέθοδο της διασταυρωμένης επικύρωσης 10-πτυχών και για κάθε ένα μοντέλο θα υπολογιστεί η μέση τιμή του μέσου τετραγωνικού σφάλματος ως κριτήριο της ακρίβειάς του για άγνωστα δεδομένα. Το μοντέλο με τη μικρότερη μέση τιμή του μέσου τετραγωνικού σφάλματος θα θεωρηθεί και το καλύτερο για την πρόβλεψη της εξαρτημένης μεταβλητής (κατανάλωση καυσίμου). Για το καλύτερο μοντέλο, οι συντελεστές του θα εκτιμηθούν λαμβάνοντας υπόψη ολόκληρο το σύνολο δεδομένων ως σύνολο εκπαίδευσης.

```
# Διασταυρωμένη Επικύρωση k-Πτυχών
#
# Συνάρτηση που υπολογίζει και επιστρέφει το Μέσο Τετραγωνικό Σφάλμα (Root Mean Squared Error
# - RMSE)
# predictedValues: διάνυσμα με τιμές της εξαρτημένης μεταβλητής που προβλέπει το μοντέλο
```

[25] Στον ιστότοπο το αρχείο έχει όνομα auto-mpg.data. Για τις ανάγκες του κεφαλαίου, το αρχείο έχει μετονομαστεί σε auto-mpg.csv.

```
#                  παλινδρόμησης

# actualValues: διάνυσμα με τις πραγματικές τιμές της εξαρτημένης μεταβλητής

# Επιστρέφει το Μέσο Τετραγωνικό Σφάλμα πρόβλεψης

calculateRMSE<-function(predictedValues, actualValues){

  err<- sqrt( mean((actualValues - predictedValues)^2)  )

  return( err )

}

# Συνάρτηση που υλοποιεί τον αλγόριθμο της διασταρωμένης επικύρωσης k-πτυχών.

# Η συνάρτηση κάνει χρήση του μέσου τετραγωνικού σφάλματος (RMSE) ως μετρική σφάλματος.

# Παράμετροι συνάρτησης:

# data: το σύνολο δεδομένων που θα χωριστεί σε τμήματα ελέγχουν κα εκπαίδευσης

# frml: το γραμμικό μοντέλο παλινδρόμησης που θα αξιολογηθεί η α-κρίβεια πρόβλεψής του

# k: η τιμή k της διασταυρωμένης επικύρωσης k-πτυχών που δηλώνει σε πόσα τμήματα θα διαχωριστεί

#    το αρχικό σύνολο δεδομένων

# Επιστρεφόμενη τιμή:

# Η συνάρτηση επιστρέφει τον μέσο όρο του μέσου τετραγωνικού σφάλ-ματος

kFoldCrossValidation<-function(data, frml, k){

  # Τυχαία αναφιάταξη των παρατηρήσεων του συνόλου δεδομένων

  dataset<-data[sample(nrow(data)),]

   #Δημιουργία k σε πλήθος τμημάτων του συνόλου δεδομένων με περί-που ίσο πλήθος

  # παρατηρήσεων σε κάθε τμήμα.

  folds <- cut(seq(1,nrow(dataset)), breaks=k, labels=FALSE)
```

```
  # Διάνυσμα όπου αποθηκεύεται το Μέσο Τε

  RMSE<-vector()

  # Επαναλαπτική διαδικασία όπου κάθε ένα από τα k τμήματα θα χρη-
σιμοποιηθεί διαδοχικά

  # ως σύνολο ελέγχου για το μοντέλο παλινδρόμησης και όλα τα υπό-
λοιπα ως σύνολο εκπαίδευσης.

  # Η διαδικασία θα τερματίσει εάν όλα τα τμήματα έχουν χρησιμοποι-
ηθεί ως σύνολο σλέγχου.

  for(i in 1:k){

    # Καθορισμός του τμήματος ελέγχου για την τρέχουσα επανάληψη

    testIndexes <- which(folds==i,arr.ind=TRUE)

    # Καθορισμός συνόλου ελέγχου μοντέλου

    testData <- dataset[testIndexes, ]

    # Καθορισμός συνόλου εκπαίδευσης μοντέλου, που θα είναι όλα τα
υπόλοιπα

   # πλην των δεδομένων που χρησιμοποιηθούν για έλεγχο

    trainData <- dataset[-testIndexes, ]

    # Εκτίμηση συντελεστών του μοντέλου παλινδρόμησης χρησιμοποιώ-
ντας το σύνολο εκπαίδευσης

    candidate.linear.model<-lm( frml, data = trainData)

    # Υπολογισμός των τιμών της εξαρτημένης μεταβλητής που προβλέ-
πει το μοντέλο

    # για τις τιμές του τρέχοντος συνόλου ελέγχου

    predicted<-predict(candidate.linear.model, testData)

    # Υπολογισμός σφάλματος RMSE

    error<-calculateRMSE(predicted, testData[, "mpg"])

    # Αποθήκευση τιμής σφάλματος

    RMSE<-c(RMSE, error)

  }
```

```
  # Επιστροφή μέσης τιμής των σφαλμάτων που προέκυψαν απ'όλα τα
τμήματα ελέγχου

  return( mean(RMSE) )

}

# Ανάγνωση δεδομένων

carData<-read.csv("auto-mpg.csv", sep=";", header=T, stringsAsFac-
tors = F, quote = "\"")

# Τα τέσσερα υποψήφια μοντέλα των οποίων θα αξιολογηθεί η ικανότητα
πρόβλεψης

# με τη μέθοδο της διασταυρωτικής επικύρωσης 10-φορές.

# Τα μοντέλα παλινδρόμησης αποθηκεύονται στο διάνυσμα ως συμβολο-
σειρές και θα μετατραπούν

# σε τύπους της R (formula)

predictionModels<-vector()

predictionModels[1]<-"mpg ~ horsepower+weight"

predictionModels[2]<-"mpg ~ horsepower+acceleration+displacement"

predictionModels[3]<-"mpg ~ horsepower+displacement+weight"

# Συμπεριλαμβάνεται και μοντέλο που εισάγει πολυωνυμικό όρο βαθμού
2 weight2. Η εισαγωγή τέτοιων

# όρων στο μοντέλο απαιτεί τη χρήση της συνάρτησης I() που είναι η
συνάρτηση Inhibit Interpretation και

# έχει σαν αποτέλεσμα να μην ερμηνευτεί ο τελεστής ^ στα πλαίσια
του τύπου.

# Αυτό γιατί ο τελεστής ^ έχει ειδική σημασία για τύπους και αν χρη-
σιμοποιείται σε τέτοιους δίχως τη

# χρήση της I() δεν θα ερμηνευτεί ως ο τελεστής ύψωση σε δύναμη.

predictionModels[4]<-"mpg ~ horsepower+displacement+weight +
I(weight^2)"

# Μετά από κάθε διασταυρωμένη επικύρωση, ο μέσος όρος του μέσου
τετραγωνικού
```

```
# σφάλματος κάθε μοντέλου θα αποθηκευτεί σε διάνυσμα.

modelMeanRMSE<-vector()

# Διασταυρωμένη επικύρωση 10-φορές για κάθε ένα από τα

# τέσσερα υποψήφια μοντέλα.

for (k in 1:length(predictionModels)){

     # Δισταυρωμένη επικύρωση 10-πτυχών για το γραμμικό μοντέλο
παλινδρόμησης k

     modelErr<-kFoldCrossValidation(carData,
as.formula(predictionModels[k]), 10)

     # Αποθήκευση του μέσου σφάλματος

     modelMeanRMSE<-c(modelMeanRMSE, modelErr)

}

# Ποιο μοντέλο είχε το χαμηλότερο μέσο τετραγωνικό σφλάμα;

bestModelIndex<-which( modelMeanRMSE == min(modelMeanRMSE) )

# Εμφάνιση μοντέλου με το μικρότερο μέσο τετραγωνικό σφάμα δηλαδή
τη μεγαλύτερη ακρίβεια

print( sprintf("Model with best accuracy was: [%s] error: [%f]",
predictionModels[bestModelIndex], modelMeanRMSE[bestModelIndex]) )

# Για το μοντέλο με το χαμηλότερο μέσο σφάλμα, εκτιμώνται οι συντε-
λεστές του

# λαμβάνοντας υπόψη ολόκληρο το σύνολο δεδομένων ως σύνολο εκπαί-
δευσης

final.linear.model<-lm(
as.formula(predictionModels[bestModelIndex]), data=carData )
```

Άσκηση Αυτοαξιολόγησης 0.23

Στο κώδικα R της ενότητας αυτής που υλοποιεί τη μέθοδο της διασταυρωτικής επικύρωσης k-πτυχών χρησιμοποιήθηκε η ακόλουθη εντολή:

```
folds <- cut(seq(1,nrow(dataset)), breaks=k, labels=FALSE)
```

Ποιος ο ρόλος της και πώς ακριβώς λειτουργεί στα πλαίσια του προγράμματος R;

- ***Ερμηνεία αποτελεσμάτων γραμμικού μοντέλου παλινδρόμησης με στόχο την πρόβλεψη.***

Εάν έχει διαμορφωθεί το τελικό μοντέλο γραμμικής παλινδρόμησης με στόχο την πρόβλεψη, τότε μπορεί αυτό να χρησιμοποιηθεί προκειμένου να υπολογιστούν οι τιμές της εξαρτημένης μεταβλητής για νέες, άγνωστες τιμές των ανεξαρτήτων μεταβλητών. Ωστόσο, οι τιμές της εξαρτημένης μεταβλητής που προκύπτουν πρέπει να ερμηνευτούν στατιστικά και όχι ντετερμινιστικά. Ειδικότερα, οι τιμές που προκύπτουν για την εξαρτημένη μεταβλητή ερμηνεύονται ως ο μέσος όρος της τιμής εξαρτημένης μεταβλητής που θα προκύψει για τις συγκεκριμένες τιμές τιμές των ανεξαρτήτων μεταβλητών.

Έτσι για παράδειγμα, αν η διασταυρωμένη επικύρωση οδήγησε στο εξής μοντέλο γραμμικής παλινδρόμησης ως το πιο ακριβές για την την κατανάλωση καυσίμων αυτοκινήτων

$$\begin{aligned} Κατανάλωση\ καυσίμου\ ανά\ μίλι \\ = -0.04167 * Ιπποδύναμη - 0.005768 * Κυβισμός \\ - 0.005351 * Βάρος + 44.8559 \end{aligned}$$

τότε η κατανάλωση καυσίμου ενός αυτοκινήτου με ιπποδύναμη ίση με 125, κυβισμό ίσο με 1400 cc και βάρος 745 κιλά, σύμφωνα με το μοντέλο παλινδρόμησης θα είναι 34.8531 μίλια το γαλόνι. Η τιμή αυτή πρέπει να ερμηνευθεί ως ότι 34.8531 μίλια το γαλόνι θα είναι κατά μέσο όρο η κατανάλωση καυσίμου του αυτοκινήτου με τέτοια χαρακτηριστικά.

6.7.2.2 Υποπροσαρμογή, Υπερπροσαρμογή και Κανονικοποίηση μοντέλου παλινδρόμησης

Ο στόχος ενός μοντέλου παλινδρόμησης είναι να παρουσιάζει όσο το δυνατό μικρότερο σφάλμα γενίκευσης προκειμένου να είναι χρήσιμο στην πρόβλεψη των τιμών της εξαρτημένης μεταβλητής. Ωστόσο, η πραγματική τιμή του σφάλματος γενίκευσης ενός μοντέλου δεν μπορεί να γίνει ποτέ γνωστή και μπορεί μόνο να εκτιμηθεί με τις μεθόδους που παρουσιάστηκαν σε προηγούμενη ενότητα.

Από το σφάλμα εκπαίδευσης ενός μοντέλου, που είναι πάντα γνωστό και εκτιμάται κατά την εκπαίδευση του μοντέλου, δεν μπορεί να εξαχθεί με ασφάλεια κα-

μία εκτίμηση για το σφάλμα γενίκευσης του μοντέλου. Ωστόσο, από της σχέση που υπάρχει μεταξύ του σφάλματος εκπαίδευσης και του σφάλματος γενίκευσης ενός μοντέλου, μπορούν να βγουν ορισμένα χρήσιμα συμπεράσματα για το ίδιο το μοντέλο.

Έτσι, εαν τόσο το σφάλμα εκπαίδευσης όσο και το σφάλμα γενίκευσης ενός μοντέλου είναι μικρά, τότε λέγεται ότι το μοντέλο κατορθώνει να προσαρμοστεί καλά στα δεδομένα εκπαίδευσης. Αποτελεί την ιδανική κατάσταση ενός μοντέλου και σημαίνει ότι ο τρόπος με τον οποίο έχει προσδιοριστεί (που αναφέρεται στις ανεξάρτητες μεταβλητές που περιέχει, ο βαθμός των μονωνύμων που εμφανίζει κλπ.) είναι ικανός να προβλέψει με ακρίβεια την τιμή της ανεξάρτητης μεταβλητής με ακρίβεια.

Σε περίπτωση που το σφάλμα εκπαίδευσης και το σφάλμα γενίκευσης είναι μεγάλα, τότε λέγεται ότι το μοντέλο παρουσιάζει την κατάσταση της υποπροσαρμογής (underfitting). Σημαίνει ότι το μοντέλο παλινδρόμησης, με τον τρόπο με τον οποίο έχει προσδιοριστεί, δεν κατορθώνει να προσαρμοστεί καθόλου καλά στα δεδομένα εκπαίδευσης. Η αδυναμία προσαρμογής του στα δεδομένα εκπαίδευσης οφείλεται κυρίως στην απλότητα του μοντέλου. Όταν το μοντέλο παραείναι απλό, δεν έχει την ευελιξία να προσαρμοστεί στις μεταβολές και διακυμάνσεις των δεδομένων εκπαίδευσης και κατά συνέπεια αδυνατεί να συλλάβει πλήρως σχέση μεταξύ της εξαρτημένης και ανεξαρτήτων μεταβλητών. Αυτό έχει ως αποτέλεσμα οι προβλέψεις για άγνωστα δεδομένα να έχουν κι αυτές μεγάλο σφάλμα.

Τέλος, στη περίπτωση που το σφάλμα εκπαίδευσης είναι μικρό, ενώ το σφάλμα γενίκευσης μεγάλο, τότε λέγεται ότι το μοντέλο παρουσιάζει την κατάσταση της υπερπροσαρμογής (overfitting). Υπερπροσαρμογή σημαίνει ότι το μοντέλο παλινδρόμησης προσαρμόζεται τόσο καλά στα δεδομένα εκπαίδευσης, που δεν μπορεί να προβλέψει ορθά την τιμή της εξαρτημένης μεταβλητής για κανένα άλλο άγνωστο δεδομένο – εξ ού και ο όρος υπερπροσαρμογή που ερμηνεύεται ως υπερπροσαρμογή στα δεδομένα εκπαίδευσης. Και μάλιστα, το μεγάλο σφάλμα στην πρόβλεψη για άγνωστα δεδομένα οφείλεται ακριβώς στο γεγονός ότι προβλέπει πάρα πολύ καλά μόνο αν χρησιμοποιηθούν δεδομένα εκπαίδευσης. Τέτοια αδυναμία του μοντέλου οφείλεται κυρίως στο γεγονός ότι εμφανίζει μεγάλη πολυπλοκότητα με αποτέλεσμα να συλλαμβάνει πέραν του δέοντος καλά όχι μό-

νο τις σχέσεις που υπάρχουν μεταξύ των μεταβλητών στα δεδομένα εκπαίδευσης αλλά και τις τυχαίες διακυμάνσεις τους δηλαδή τον θόρυβο που υπάρχει σε αυτά.

Γενικά, η κατάσταση της υποπροσαρμογής αυξάνει την αμεροληψία του μοντέλου (οδηγεί δηλαδή οι προβλεπόμενες από το μοντέλο τιμές να απέχουν πολύ από τις πραγματικές) ενώ η κατάσταση της υπερπροσαρμογής αυξάνει τη διακύμανση των προβλεπόμενων τιμών του μοντέλου (δηλαδή οι προβλεπόμενες τιμές εμφανίζουν πολύ μεγάλη διασπορά γύρω από την πραγματική τιμή της εξαρτημένης μεταβλητής).

Η υπερπροσαρμογή ενός μοντέλου αποτελεί σημαντικό πρόβλημα το οποίο ωστόσο μπορεί να αντιμετωπιστεί. Ένας τρόπος αντιμετώπισης είναι η κανονικοποίηση (regularization) του μοντέλου παλινδρόμησης.

Άσκηση Αυτοαξιολόγησης 0.24

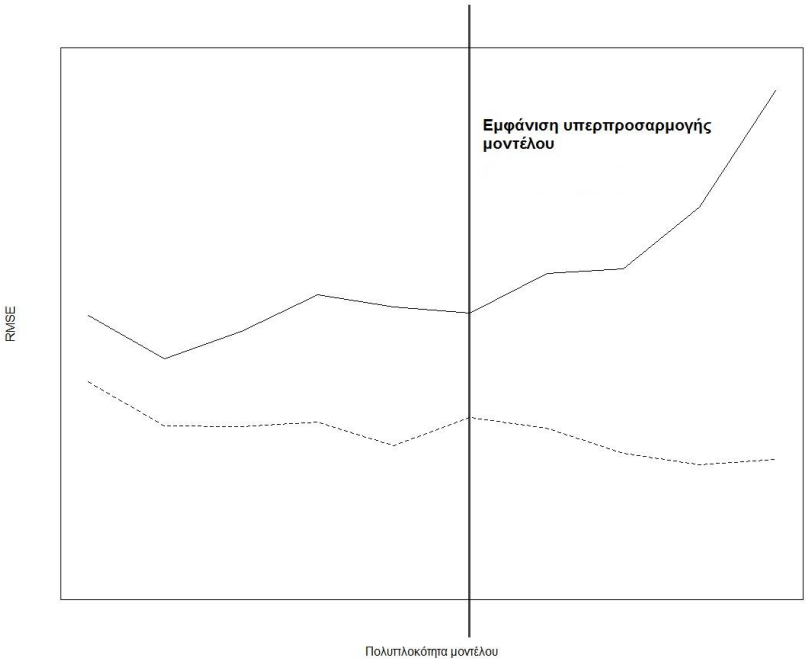
Εταιρεία προσπαθεί να αντικαταστήσει υπάλληλο μηχανικό λογισμικού που αποχώρησε με συνταξιοδότηση. Επειδή ο συγκεκριμένος μηχανικός αποδείχθηκε πάρα πολύ καλός, η επιχείρηση προσπαθεί να βρει νέο μηχανικό για την πλήρωση της κενής θέσης παρόμοιο με εκείνον που αποχώρησε και κάνει την εξής δημοσίευση σε εφημερίδες και σε κοινωνικά δίκτυα:

“Ζητείται μηχανικός λογισμικού με βαθμό πτυχίου 8.3/10 από πανεπιστήμιο της δυτικής ακτής των ΗΠΑ, με άριστες γνώσεις στη γλώσσα προγραμματισμού Java και εμπειρία τουλάχιστον 5 ετών σε έργα λογισμικού άνω των 200.000 Ευρώ και στα χόμπι του να είναι τα κόμικ, η κτηνοτροφία και η συλλογή βιβλίων του 15ου αιώνα. Να έχει ξανθά μαλλιά, καστανά μάτια, ελιά στο μάγουλο και στη δεξιά μασχάλη, να συχαίνεται τους κύκλους, να κατηγορεί τα κάστανα ότι είναι νωθρά και να ισχυρίζεται ότι έχει ανακαλύψει το ερωτηματικό. Όταν τρέχει το πρόγραμμα που έγραψε να αναφωνεί δυνατά «Hadouuuuuuuuuuken!» “

Η παραπάνω περιγραφή της θέσης, αποτελεί παράδειγμα υποπροσαρμογής ή υπερπροσαρμογής; Τεκμηριώστε την απάντησή σας.

- ***Αντιμετώπιση υπερπροσαρμογής μοντέλου με χρήση της κανονικοποίησης (regularization)***

Όπως αναφέρθηκε παραπάνω, η υπερπροσαρμογή οφείλεται στο γεγονός ότι το μοντέλο παλινδρόμησης είναι πολύ πολύπλοκο με αποτέλεσμα όχι μόνο να προσαρμόζεται καλά στα δεδομένα, αλλά και στον θόρυβο ή την τυχαία διακύμανση που αυτά εμπεριέχουν. Αυτός είναι και ο λόγος που μοντέλα με υπερπροσαρμογή παρουσιάζουν μεγάλο σφάλμα για άγνωστα δεδομένα, μιας και έχουν προσαρμοστεί απολύτως στα δεδομένα εκπαίδευσης.



Εικόνα 0.18 *Το φαινόμενο της υπερπροσαρμογής μοντέλου ως συνάρτηση της πολυπλοκότητάς του. Από ένα σημείο πολυπλοκότητας του μοντέλου και πέρα, όσο αυτή αυξάνεται, το σφάλμα εκπαίδευσης μειώνεται (διακεκομμένη γραμμή) ενώ ταυτόχρονα το σφάλμα γενίκευσης (συνεχή γραμμή) του μοντέλου μεγαλώνει. Η κατάσταση αυτή χαρακτηρίζει την υπερπροσαρμογή του μοντέλου..*

Με τον όρο πολυπλοκότητα ενός μοντέλου εννοείται ο τρόπος με τον οποίο έχει προσδιοριστεί και αναφέρεται συνήθως τόσο στο πλήθος των ανεξαρτήτων μεταβλητών που περιέχει όσο και στον βαθμό των μεταβλητών αυτών. Ένα μοντέλο με περισσότερες ανεξάρτητες μεταβλητές είναι πιο πολύπλοκο από ένα μοντέλο

με λιγότερες τέτοιες μεταβλητές. Ένα μοντέλο με όρους μονωνύμων με μεγαλύτερο βαθμό από άλλο, θεωρείται επίσης πιο πολύπλοκο. Γενικότερα, όσο πιο πολύπλοκο είναι ένα μοντέλο παλινδρόμησης, τόσο το σφάλμα εκπαίδευσης θα μειώνεται ενώ το σφάλμα γενίκευσης θα αυξάνεται. Η σχέση αυτή μεταξύ της πολυπλοκότητας του μοντέλου και της του σφάλματος εκπαίδευσης και γενίκευσης όπως φαίνεται στο σχήμα 6.18.

Ο έλεγχος για το εάν ένα μοντέλο παλινδρόμησης που έχει εκτιμηθεί, υπόκειται σε υπερπροσαρμογή, μπορεί να γίνει με τη μέθοδο της διασταυρωμένης επικύρωσης κ-πτυχών το οποίο θα δώσει το σφάλμα γενίκευσης του μοντέλου. Το σφάλμα γενίκευσης που προκύπτει μπορεί να συγκριθεί με το σφάλμα εκπαίδευσης του μοντέλου και εαν η διαφορά τους είναι πολύ μεγάλη, τότε υπάρχουν οι ενδείξεις ότι το μοντέλο παρουσιάζει υπερπροσαρμογή.

Χαρακτηριστικό ενός μοντέλου παλινδρόμησης που παρουσιάζει υπερπροσαρμογή είναι επίσης και οι πολύ μεγάλες τιμές συντελεστών που θα προκύψουν κατά την εκτίμησή του. Εξαιτίας των πολύ μεγάλων τιμών των συντελεστών του μοντέλου, η επίδραση των ανεξάρτητων μεταβλητών υπερκετιμάται, και εμφανίζεται έτσι μεγάλη διακύμανση στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής, μιας και ακόμη και μικρές μεταβολές στις τιμές των ανεξαρτήτων μεταβλητών θα οδηγήσει σε μεγάλες τιμές της εξαρτημένης.

Γενικά, δύο είναι οι τρόποι αντιμετώπισης ενός μοντέλου που παρουσιάζει το πρόβλημα της υπερπροσαρμογής. Ο πρώτος τρόπος στοχεύει στο να κάνει το μοντέλο πιο απλό: όπως συζητήθηκε, η υπερπροσαγμογή σχετίζεται άμεσα με την πολυπλοκότητα του μοντέλου. Μια αντιμετώπιση του προβλήματος είναι η μείωση της πολυπλοκότητάς του είτε με την αφαίρεση ορισμένων ανεξαρτήτων μεταβλητών ή τη μείωση του βαθμού των όρων μονωνύμων του μοντέλου. Ένα απλούστερο μοντέλο θα βελτιώσει το σφάλμα πρόβλεψης και το ποιες μεταβλητές θα πρέπει να αφαιρεθούν είναι ζήτημα μελέτης και ελέγχων.

Ένας δεύτερος τρόπος αντιμετώπισης ξεκινά από την παρατήρηση, ότι η υπερπροσαρμογή ενός μοντέλου θα επηρεάσει τη διακύμανση της προβλεπόμενης τιμής αλλά όχι την μεροληψία της. Επειδή η μεγάλη διακύμανση οφείλεται στις πολύ μεγάλες τιμές των συντελεστών που προκύπτουν σε μοντέλα με υπερπροσαρμογή, ο τρόπος αυτός εστιάζει σε μεθόδους που επιχειρούν να περιορίσουν τεχνητά το μέγεθος των συντελεστών που προκύπτουν. Τέτοιος περιορισμός των

τιμών των συντελεστών μπορεί να γίνει με διάφορους τρόπους και τέτοιες μέθοδοι καλούνται με τον όρο κανονικοποίηση του μοντέλου (regularization).

❖ ***Κανονικοποίηση συντελεστών μοντέλου παλινδρόμησης***

Κατά την κανονικοποίηση ενός γραμμικού μοντέλου παλινδρόμησης που παρουσιάζει υπερπροσαρμογή, τίθενται περιορισμοί, οι οποίοι προσπαθούν να τιμωρήσουν την υπέρμετρη προσαρμογή στα δεδομένα εκπαίδευσης με το να θέτουν περιορισμό στις τιμές που μπορούν να λάβουν οι συντελεστές του μοντέλου κατά τη διάρκεια εκτίμησής τους.

Κατά την κανονικοποίηση, αυτό που αλλάζει είναι η μορφή της συνάρτησης κόστους ώστε να θέσει και περιορισμούς σχετικά με την τιμή που μπορούν να λάβουν οι συντελεστές. Ειδικότερα, όταν επιχειρείται κανονικοποίηση του μοντέλου παλινδρόμησης, η συνάρτηση κόστους λαμβάνει την ακόλουθη γενική μορφή[26]:

$$J(\beta_0, \beta_1, \beta_2, \dots, \beta_k) = \frac{1}{2m}\sum_{i=1}^{m}\left(h_\beta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\left|\beta_j\right|^q$$

Στην συνάρτηση κόστους αυτό που αλλάζει είναι ότι προστίθεται και ο όρος $\lambda \sum_{j=1}^{n}\left|\beta_j\right|^q$ που είναι το άθροισμα δυνάμεων των εκτιμώμενων συντελεστών. Αποτελεί όρος που εισάγει ένα πέναλτυ στη συνάρτηση κόστους, ο οποίος καλείται και όρος κανονικοποίησης (regularization term). Ο λόγος που προστίθεται ο όρος αυτός στη συνάρτηση κόστους μπορεί να εξηγηθεί διαισθητικά ως εξής: επειδή ο στόχος στην ανάλυση παλινδρόμησης είναι να βρεθούν οι συντελεστές που ελαχιστοποιούν τη συνάρτηση κόστους, με την παραπάνω μορφή της θα οδηγήσει αναγκαστικά στην ελαχιστοποίηση και των δύο όρων του αθροίσματος που την απαρτίζουν. Δηλαδή θα εκτιμηθούν και οι συντελεστές που ελαχιστοποιούν και την παράστασης $\lambda \sum_{j=1}^{n}\left|\beta_j\right|^q$ η οποία είναι το άθροισμα των εκτιμωμένων συντελεστών. Αυτό σημαίνει ότι η ελαχιστοποίηση και της παράστασης αυτής δεν θα επιτρέψει τους συντελεστές β_j να λάβουν οποιαδήποτε (μεγάλη)

[26] Εδώ γίνεται χρήση της συνάρτησης κόστους που χρησιμοποιείται στη μέθοδο της σταδιακής καθόδου. Αντί αυτής θα μπορούσε να χρησιμοποιηθεί οποιαδήποτε άλλη που εμφανίζεται στις μεθόδους εκτίμησης συντελεστών μοντέλων παλινδρόμησης όπως για παράδειγμα η συνάρτηση κόστους της μέθόδου ελαχίστων τετραγώνων. Οι αλλαγές στη συνάρτηση κόστους θα ήταν ακριβώς οι ίδιες.

τιμή. Η παράμετρος λ που εμφανίζεται καλείται παράμετρος κανονικοποίησης (regularization parameter), είναι μία σταθερά και καθορίζει την ισχύ του πέναλτι που εισάγεται στη συνάρτηση κόστους. Η τιμή της παραμέτρου αυτής καθορίζεται εξαρχής και παίζει σημαντικό ρόλο στον τρόπο με τον οποίο θα λειτουργήσει η ανάλυση παλινδρόμησης. Η δύναμη q που εμφανίζεται είναι και αυτή μία σταθερά, η τιμή της οποίας καθορίζει το είδος της κανονικοποίησης και την έμφαση που θα δοθεί. Ειδικότερα, ανάλογα με την τιμή της δύναμης q στην οποία υψώνονται οι συντελεστές, διακρίνονται οι εξής τύποι κανονικοποίησης:

- Εάν η παράμετρος q λάβει την τιμή 1 (q=1), τότε η συνάρτηση κόστους λαμβάνει τη μορφή $\frac{1}{2m}\sum_{i=1}^{m}\left(h_\beta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\left|\beta_j\right|$ και η παλινδρόμηση καλείται Lasso regression ή L1 regression. Η παλινδρόμηση Lasso, με τον τρόπο που προσδιορίζει τη συνάρτηση κόστους, εκτός του ότι μειώνει τους συντελεστές είναι και ικανή να κάνει επιλογή χαρακτηριστικών (feature selection). Αυτό γιατί, όσο πιο μεγάλη είναι η τιμή λ, τόσο περισσότερες εκτιμήσεις συντελεστών θα οδηγηθούν να έχουν τιμή ίση με το μηδέν. Αυτό πολύ απλά σημαίνει, ότι συντελεστές που θα λάβουν τιμή ίση με το μηδέν σηματοδοτούν ανεξάρτητες μεταβλητές που δεν είναι σημαντικές για την πρόβλεψη της τιμής της εξαρτημένης μεταβλητής. Κατά συνέπεια, η παλινδρόμηση Lasso επιτρέπει την προσδιορισμό των σημαντικών χαρακτηριστικών (μεταβλητών) στο μοντέλο παλινδρόμησης.
- Εάν η παράμετρος q λάβει την τιμή 2 (q=2), τότε η συνάρτηση κόστους λαμβάνει τη μορφή $\frac{1}{2m}\sum_{i=1}^{m}\left(h_\beta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{j=1}^{n}\left|\beta_j\right|^2$ και η παλινδρόμηση καλείται Ridge regression ή L2 regression ή Thikonov regularization. Αποτελεί μία δημοφιλή μέθοδο κανονικοποίησης, η οποία καταφέρνει να μειώσει το μέγεθος των συντελεστών, όσο πιο μεγάλη είναι η τιμή της παραμέτρου λ, δίχως ωστόσο να οδηγεί τους συντελεστές να λάβουν την τιμή μηδέν. Έτσι, αν και κατορθώνει να μειώσει τη διακύμανση (με τη μείωση των συντελεστών) δεν μπορεί να υποδείξει ποιες μεταβλητές είναι σημαντικές και ποιες όχι όπως η παλινδρόμηση Lasso.

Υπάρχει και άλλη μία εκδοχή κανονικοποίησης που καλείται Elasticnet, η οποία επιχειρεί να αποτελέσει μία συμβιβαστική λύση μεταξύ Lasso και Ridge παλιν-

δρόμησης. Η μέθοδος αυτή εισάγει δύο όρους κανονικοποίησης στη συνάρτηση κόστους, ένας όμοιος με τον όρο της παλινδρόμησης Lasso και ο άλλος όμοιος με τον όρο της παλινδρόμησης Ridge στους οποίους ωστόσο προσδίδει διαφορετική βαρύτητα.

Εάν η τιμή της παραμέτρου κανονικοποίησης λ λάβει τιμή ίση με το μηδέν, τότε η συνάρτηση κόστους λαμβάνει την παραδοσιακή της μορφή και δεν συμβαίνει καμία κανονικοποίηση. Η κατάλληλη τιμή της παραμέτρου λ πρέπει να προσδιοριστεί και αυτό μπορεί να γίνει υποβάλλοντας το μοντέλο παλινδρόμησης σε διασταυρωτική επικύρωση k-πτυχών για διάφορες υποψήφιες τιμές της παραμέτρου λ και αξιολόγηση των αποτελεσμάτων.

Αν επιθυμείται η υλοποίηση παλινδρόμησης Ridge με τη μέθοδο της Σταδιακής Καθόδου στο περιβάλλον της R, ο τύπος ενημέρωσης των συντελεστών θα λάβει τη μορφή

$$\theta_j := \theta_j - \alpha \frac{1}{m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) {x_j}^{(i)} + \lambda\theta_j\right]$$

ενώ η συνάρτηση κόστους, όπου πρέπει να ενσωματωθεί και ο όρος κανονικοποίησης θα έχει ως εξής:

```
calculateCostRidge<-function(X, y, theta, lambda=0){
  # Πλήθος παρατηρήσεων
  m <- length(y)
  return( sum((X%*%theta- y)^2) / (2*m) + lambda*sum(theta^2))
} # calculateCostRidge
```

Στο περιβάλλον της R υπάρχουν εξειδικευμένες βιβλιοθήκες που υλοποιούν όλες τις εκδοχές παλινδρόμησης με κανονικοποίηση. Τέτοια βιβλιοθήκη στην R είναι για παράδειγμα η glmnet, που παρέχει υλοποιήσεις των παλινδρομήσεων Lasso, Ridge και Elasticnet.

Άσκηση Αυτοαξιολόγησης 0.25

Τί επίδραση θα έχει στην εκτίμηση των συντελεστών θ ενός μοντέλο παλινδρόμησης κατά την παλινδρόμηση Ridge, εάν επιλεγεί μία τιμή της παραμέτρου κανονικοποίησης λ που είναι πάρα πολύ μεγάλη;

6.8 Σύνοψη

Στο κεφάλαιο αυτό παρουσιάστηκε η μέθοδος της ανάλυσης παλινδρόμησης, η οποία είναι μία στατιστική μέθοδος που επιχειρεί να εκτιμήσει της συσχέτιση μεταξύ μεταβλητών ενός σύνολο δεδομένων με τη μορφή εξίσωσης. Αρχικά ορίστηκαν οι βασικές έννοιες που συναντούνται στην ανάλυση παλινδρόμησης και παρουσιάστηκαν διαγραμματικοί τρόποι για τη διερεύνηση της σχέσης μεταξύ των μεταβλητών σε ένα σύνολο δεδομένων και πως αυτοί μπορούν να γίνουν στο περιβάλλον της R. Ακολούθως, παρουσιάστηκαν οι στόχοι ενός μοντέλου παλινδρόμησης, οι οποίοι οδηγούν σε διαφορετικούς τρόπους αξιολόγησής του. Ορίστηκαν επίσης και παρουσιάστηκαν τα κριτήρια για να χαρακτηριστεί ένα μοντέλο παλινδρόμησης γραμμικό και μη-γραμμικό. Το κεφάλαιο επικεντρώθηκε στη συνέχεια σε γραμμικά μοντέλα παλινδρόμησης και τους τρόπους εκτίμησης των συντελεστών τους. Παρουσιάστηκαν δύο διαφορετικοί τρόποι εκτίμησης συντελεστών, η μέθοδος των ελαχίστων τετραγώνων και η μέθοδος της σταδιακής καθόδου και επισημάνθηκαν οι διαφορές τους και περιπτώσεις όπου είναι χρήσιμες. Δόθηκε κώδικας σε R, ο οποίος υλοποιεί τις μεθόδους αυτές και τις χρησιμοποεί για την εκτίμηση συντελεστών σε παραδειγματικά δεδομένα. Για τη μέθοδο της σταδιακής καθόδου, παρουσιάστηκαν επιπλέον δύο εκδοχές της οι οποίες μπορούν να λειτουργήσουν και σε περιβάλλοντα μεγάλων δεδομένων. Ακολούθως το κεφάλαιο προυσίασε τρόπος για την αξιολόγηση και την ερμηνεία μοντέλων γραμμικής παλινδρόμησης. Η ενότητα αυτή διαχώρησε τον τρόπο αξιολόγησης και ερμηνείας ανάλογα με τον στόχο του μοντέλου παλινδρόμησης. Στην περίπτωση που στόχος ενός μοντέλου παλινδρόμησης είναι η εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής, παρουσιάστηκαν οι απαραίτητοι έλεγχοι που πρέπει να γίνουν όπως έλεγχος γραμμικότητας, έλεγχος ομοσκεδαστικότητας καταλοίπων, έλεγχος κανονικής κατανομής καταλοίπων και έλεγχος πολυσυγγραμμικότητας και παρουσιάστηκε με παραδείγματα πως αυτοί οι έλεγχοι μπορούν να υλοποιηθούν στην R. Στην περίπτωση που ο στόχος του μοντέλου είναι η πρόβλεψη της τιμής της εξαρτημένης μεταβλητής, δόθηκε έμφαση στην

εκτίμηση του σφάλματος πρόβλεψης και μεθοδολογίες ελέγχου όπως η διασταυρωτική επικύρωση k-πτυχών. Με κώδικα R παρουσιάστηκε πως τέτοιοι έλεγχοι μπορούν να προγραμματιστούν στο περιβάλλον της R. Ειδικά για τον στόχο της πρόβλεψης, παρουσιάστηκε το μείζον πρόβλημα της υπερπροσαρμογής και τρόποι αντιμετώπισής του μέσω των διαφόρων εκδοχών της κανονικοποίησης: την κανονικοποίηση Lasso, Ridge και Elasticnet.

6.9 Κατάλογος Πακέτων της R

Πακέτο	Εγχειρίδιο Χρήσης
car	https://cran.r-project.org/web/packages/car/car.pdf
gradDescent	https://cran.r-project.org/web/packages/gradDescent/gradDescent.pdf
glmnet	https://cran.r-project.org/web/packages/glmnet/glmnet.pdf

6.10 Βιβλιογραφικές Πηγές

Waugh, F. V.: Choice of the Dependent Variable in Regression Analysis, Journal of the American Statistical Association, Vol. 38, No. 222, June., 1943, pp. 210-216

Shmueli, G.: To Explain or to Predict? Statistical Science 25(3), January 2011

Peng, R. D. *R Programming for Data Science*. Lean Publishing. Ανακτήθηκε στις 19 Νοεμβρίου 2018, από: https://leanpub.com/rprogramming

Draper, N. R. and Smith, H.: Applied Regression Analysis, Wiley-Interscience; Third edition, 1998

Brian, C.: *Regression Models for Data Science in R*. Lean Publishing, 2015, Ανακτήθηκε στις Ιούνιο 2019, από: https://leanpub.com/regmods

Nilsson, N. J.: *Introduction to Machine Learning*, 1998, Ανακτήθηκε στις Ιούνιο 2019, από: http://robotics.stanford.edu/~nilsson/MLBOOK.pdf

Wikipedia, *Classification*. Ανακτήθηκε Ιούνιο 2019, από: https://en.wikipedia.org/wiki/Classification

Yanchang, Z.: *Regression and Classification with R*, 2015, Ανακτήθηκε στις Ιούλιο 2018, από: http://www.rdatamining.com/docs/regression-and-classification-with-r

Kiefer, J. and Wolfowitz, J.: Stochastic Estimation of the Maximum of a Regression Function, Annals of Mathematical Statistics, Volume 23, Number 3, 1952, pp 462-466.

Bottou, L., Curtis, F. E., Nocedal, J.: Optimization Methods for Large-Scale Machine Learning, SIAM Review, Volume 60 (2), 2018. Διαθέσιμο από http://arxiv.org/abs/1606.04838

Tibshirani, R.: Regression Shrinkage and Selection via the Lasso, Journal Royal Statistical Society, Volume 58 (1), 1996, pp267-288

Hoerl, A.E. and Kennard, R.: Ridge regression: Biasedestimation for nonorthogonal problems.Technometrics, 12, 1970, pp55-67

Zou, H. and Hastie, T.: Regularization and variableselection via the elastic net, Journal of the Royal StatisticalSociety, Series B.67, 2005, pp. 301–320.

Stanton, J. M.: Galton, Pearson, and the Peas: A Brief History of Linear Regression for Statistics Instructors, Journal of Statistics Education, Volume 9, Issue 3, 2001

Poole, M. A. and O'Farrell, P. N.: The Assumptions of the Linear Regression Model, Transactions of the Institute of British Geographers, No. 52, 1971, pp. 145-158 .

Ασκήσεις

1. Εξηγείστε γιατί συνηθίζεται σε μία μήτρα διαγράμματος διασποράς για ένα σύνολο δεδομένων με n μεταβλητές να μην απεικονίζονται ακριβώς $\frac{n!}{(n-2)!}$ σε πλήθος διαγράμματα ζευγών μεταβλητών αλλά $\frac{n!}{2(n-2)!}$ σε πλήθος διαγράμματα.

2. Δίνεται το παρακάτω σύνολο δεδομένων που έχει τις τιμές δύο μεταβλητών με όνομα x και y:

X	**Y**
15	22
17	35
33	48
5	18

Συγγράψτε κώδικα σε R που εξετάζει ποια από τα δύο παρακάτω μοντέλα παλινδρόμησης ερμηνεύει καλύτερα τη διακύμανση της μεταβλητής y:

i. y = 15.67x + 9.87
ii. y = 38.6x – 24.098

3. Ποια από τα παρακάτω μοντέλα παλινδρόμησης είναι γραμμικά; Τεκμηριώστε την απάντησή σας. Στα παρακάτω μοντέλα θεωρείστε ότι β_i είναι οι συντελεστές του μοντέλου.

$y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$	$\ln(y) = \beta_0 + \beta_1 \ln(X_1) + \beta_2 log_2(X_2)$
$y = \beta_0 + \beta_1 X_1^3 + \beta_2 X_2^5$	$y = \frac{\beta_1 X_1}{\beta_2 + X_1}$
$y = 42$	$y = \beta_1 X_1 + \beta_2 X_2$
$y = \beta_0 + \beta_1 \ln(X_1) + \beta_2 \ln(X_2)$	$y = \frac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}$
$y = \beta_0 + e^{\beta_1 X_1} + \beta_2 X_2$	$y = \beta_1 X_1 + \beta_2 \sqrt{X_2} + \beta_0$

4. Να αποδείξτε ότι σε περιπτώσεις όπου το πλήθος των μεταβλητών σε ένα σύνολο εκπαίδευσης n, είναι μεγαλύτερο από το πλήθος των παρατηρήσεων m στο ίδιο σύνολο εκπαίδευσης, δηλαδή n > m, τότε δεν μπορεί να υπολογιστεί η αντίστροφη μήτρα $(X^T X)^{-1}$ που εμφανίζεται στον τύπο της κανονικής εξίσωσης όταν γίνεται χρήση η μέθοδος των ελαχίστων τετραγώνων για την εκτίμηση των συντελεστών ενός μοντέλου γραμμικής παλινδρόμησης.

5. Αποδείξτε, χρησιμοποιώντας τη μέθοδο των ελαχίστων τετραγώνων, ότι για ένα απλό γραμμικό μοντέλο παλινδρόμησης της μορφής

$$Y = \beta_1 X + \beta_0$$

οι εκτιμήσεις των συντελεστών $\hat{\beta}_i$ υπολογίζονται από τους κάτωθι κλειστούς τύπους:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

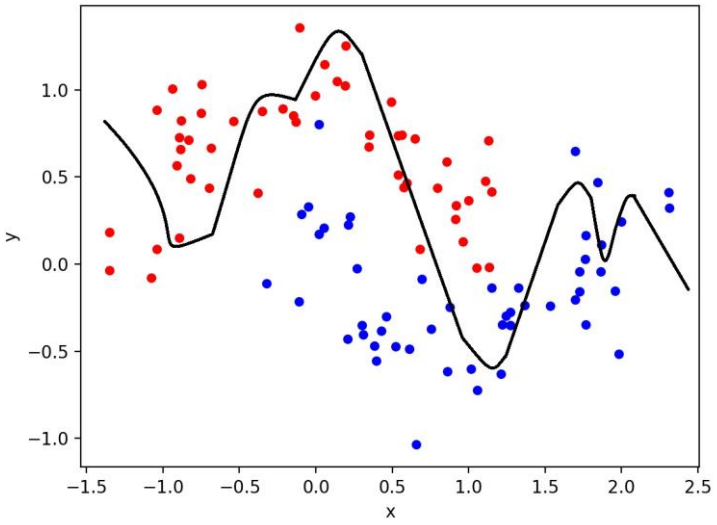
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

όπου $\bar{y}$ και $\bar{x}$ οι μέσες τιμές των μεταβλητών y και x αντίστοιχα στο σύνολο δεδομένων.

6. Εξηγείστε τη διαφορά μεταξύ της έννοια του διαταρακτικού όρου (ή σφάλμα ή θορύβου) και της έννοιας του καταλοίπου.

7. Εξηγείστε γιατί η μέθοδος των ελαχίστων τετραγώνων, και ειδικότερα η κανονική εξίσωση (normal equation) δεν μπορεί να χρησιμοποιηθεί για την εκτίμηση συντελεστών μη-γραμμικών μοντέλων παλινδρόμησης.

8. Δίνεται η παρακάτω γραφική παράσταση που απεικονίζει τις παρατηρήσεις ενός συνόλου δεδομένων ως σημεία και την γραμμή παλινδρόμησης που προέκυψε από την εκτίμηση συντελεστών ενός μοντέλου παλινδρόμησης για το ίδιο σύνολο δεδομένων. Μπορεί το μοντέλο παλινδρόμησης να είναι γραμμικό; Τεκμηριώστε την απάντησή σας.



9. Κατεβάστε από τον ιστότοπο "UCI Machine Learning Repository", από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Communities+and+Crime το σύνολο δεδομένων που περιέχει παρατηρήσεις σχετικά με την εγκληματικότητα ανά 100000 κατοίκους σε περιοχές των ΗΠΑ (μεταβλητή ViolentCrimesPerPop) μαζί με κοινωνικοοικονομικά στοιχεία για την κάθε περιοχή. Αφού εξοικειωθείτε με τα γνωρίσματα και τη σημασία τους, συγγράψτε πρόγραμμα σε R που εκτιμά με τη μέθοδο των ελαχίστων τετραγώνων τους συντελεστές του ακόλουθου πολλαπλού γραμμικού μοντέλου παλινδρόμησης:

$$ViolentCrimesPerPop = \beta_1 NumStreet + \beta_2 HousVacant + \beta_0$$

Εμφανίστε τους συντελεστές που προέκυψαν για τις ανεξάρτητες μεταβλητές του παραπάνω μοντέλου.

10. Κατεβάστε από τον ιστότοπο "UCI Machine Learning Repository", από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Wine+Quality δεδομένα που αφπρούν τα χαρακτηριστικά μιας ποικιλίας πορτογαλικού κρασιού. Ειδικότερα, κατεβάστε τα δεδομένα που αφορούν μόνο το λευκό κρασί της ποικιλίας αυτής (αρχείο winequality-white.csv). Με τη μέθοδο των ελαχίστων τετραγώνων εκτιμήστε τους συντελεστές του παρακάτω γραμμικού μοντέλου παλινδρόμησης:

$$alcohol = \beta_1 residual\ sugar + \beta_2 pH + \beta_3 density + \beta_4 fixed\ acidity + \beta_0$$

Επιπλέον, απαντήστε στα εξής ερωτήματα:

i. Τί ποσοστό της διακύμανσης μπορεί να εξηγήσει το παραπάνω γραμμικό μοντέλο παλινδρόμησης;
ii. Ποιες από τις ανεξάρτητες μεταβλητές του μοντέλου είναι στατιστικά σημαντικές;
iii. Τεκμηριώνεται ένα γραμμικό μοντέλο μεταξύ των μεταβλητών που εμφανίζει το μοντέλο;

11. Ο διορθωμένος συντελεστής προσδιορισμού (Adjusted R^2) μπορεί να πάρει αρνητικές τιμές; Τεκμηριώστε την απάντησή σας.

12. Εάν στα πλαίσια μιας ανάλυσης παλινδρόμησης ορίζεται ως επίπεδο σημαντικότητας 5%, και για μία ανεξάρτητη μεταβλητή ενός γραμμικού μοντέλου παλινδρόμησης προέκυψε η p-τιμή ίση με 0.06, τί συμπέρασμα μπορείτε να βγάλετε;

13. Για κάθε μία από τις παρακάτω περιπτώσεις, αναφέρετε τα συμπεράσματα που μπορείτε να βγάλετε για ένα απλό μοντέλο γραμμικής παλινδρόμησης αν το επίπεδο σημαντικότητας τεθεί στο 5%:

i. Χαμηλή τιμή συντελεστή προσδιορισμού R^2 και p-τιμής < 0.05
ii. Χαμηλή τιμή συντελεστή προσδιορισμού R^2 και p-τιμής > 0.05
iii. Υψηλή τιμή συντελεστή προσδιορισμού R^2 και p-τιμή < 0.05
iv. Υψηλή τιμή συντελεστή προσδιορισμού R^2 και p-τιμή > 0.05

14. Ποια μορφή θα έχει η απεικόνιση του πολλαπλού γραμμικού μοντέλου παλινδρόμησης

$$Κατανάλωση\ τροφίμων = 0.1405087\ Εισόδημα + 819.8224270\ Αριθμός\ ατόμων\ νοικοκυριού + 760.5908490$$

αν απεικονιστεί πάνω στο διάγραμμα διασποράς των δεδομένων του συνόλου εκπαίδευσης;

15. Εκτιμήστε και με τη μέθοδο των Ελαχίστων Τετραγώνων (OLS) και με τη μέθοδο της Σταδιακής Καθόδου Δέσμης (Batch Gradient Descent) τους συντελεστές του παρακάτω πολλαπλού γραμμικού μοντέλου παλινδρόμησης, χρησιμοποιώντας το σύνολο δεδομένων εκπαίδευσης HouseholdData.csv

$$Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_2 Αριθμός\ ατόμων\ νοικοκυριού + \beta_0$$

i. Συγκρίνετε τους συντελεστές που προέκυψαν από τις δύο αυτές μεθόδους. Τί παρατηρήσεις μπορείτε να κάνετε;

ii. Που πιστεύετε ότι οφείλεται η απόκλιση που παρατηρείται στις τιμές ορισμένων συντελεστών που εκτιμώνται με τη μέθοδο της Σταδιακής Καθόδου και πως μπορεί αυτή να αντιμετωπιστεί;

16. Εξηγείστε γιατί στη μέθοδο της Σταδιακής Καθόδου, η συνάρτηση κόστους J(θ) ενός απλού γραμμικού μοντέλου παλινδρόμησης της μορφής

$$y_i := \theta_0 + \theta_1 x_i$$

θα είναι κυρτή.

17. Αποδείξτε ότι η ενημέρωση της τιμής θ_k εάν ο συντελεστή θ_k δεν είναι ο σταθερός όρος, για ένα πολλαπλό γραμμικό μοντέλο παλινδρόμησης όταν χρησιμοποείται η μέθοδος της Σταδιακής Καθόδου δίνεται από τον κλειστό τύπο:

$$\theta_k := \theta_k - \alpha \frac{1}{m} \sum_{i=1}^{m} \left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) {x_k}^{(i)}$$

18. Υλοποιείστε τον αλγόριθμο της Σταδιακής Καθόδου Δέσμης σε R, που εκτιμά τους συντελεστές ενός γραμμικού μοντέλου παλινδρόμησης με k σε πλήθος ανεξάρτητες μεταβλητές και ο οποίος κάνει χρήση της ακόλουθης συνάρτησης κόστους (που καλείται L1 νόρμα):

$$J(\theta_0, \theta_1, \theta_2, \ldots, \theta_k) = \sum_{i=1}^{m} \left| y^{(i)} - h_\theta\left(x^{(i)}\right) \right|$$

Δίνεται ότι: $\frac{d}{dx}|f| = \frac{f}{|f|} * \frac{df}{dx}$

19. Υλοποιήστε σε R τον αλγόριθμο της Στοχαστικής Σταδιακής Καθόδου (Stochastic Gradient Descent).

20. Εξηγείστε γιατί όταν ο στόχος ενός μοντέλου παλινδρόμησης είναι η πρόβλεψη της τιμής της εξαρτημένης μεταβλητής, η πολυσυγγραμμικότητα δεν αποτελεί πρόβλημα που χρήζει αντιμετώπισης.

21. Κατεβάστε και διαβάστε το άρθρο «Anwar, S., Bayer, P., Hjalmarsson, R.: The Impact of Jury Race in Criminal Trials The Quarterly Journal of Economics, Volume 127, Issue 2, May 2012, Pages 1017–1055» που μπορεί να βρεθεί εδώ: https://academic.oup.com/qje/article/127/2/1017/1826107 . Ο στόχος της μελέτης είναι η εξήγηση ή πρόβλεψη; Τεκμηριώστε την άποψή σας.

22. Κατεβάστε από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Forest+Fires σύνολο δεδομένων για πυρκαγιές από περιοχές στην Πορτογαλία. Τα δεδομένα περιέχουν γεωγραφικά και μετερεωλογικά στοιχεία όταν εκδηλώνθηκαν πυρκαγιές καθώς επίσης και την επιφάνεια που κάηκε που μετριέται σε εκτάρια[27] (hectars). Έχοντας ως στόχο την πρόβλεψη της επιφάνειας που θα καεί βασει των μετεωρολογικών συνθηκών που επικρατούν συγγγράψτε κώδικα R που εκτιμάει τους συντελεστές του παρακάτω μοντέλου παλινδρόμησης με τους τρόπους που ζητείται:

$$area = \beta_1 temp + \beta_2 wind + \beta_3 rain + \beta_0$$

i. Εκτιμήστε τους συντελεστές τους συντελεστές του παραπάνω μοντέλου χρησιμοποιώντας ολόκληρο το σύνολο των δεδομένων, με τη μέθοδο των ελαχίστων τετραγώνων. Ακολούθως χρησιμοποιείστε διασταυρωτική επικύρωη 10-πτυχών και εκτιμήστε το μέσο τετραγωνικό σφάλμα (Root Mean Squared Error – RMSE) της πρόβλεψης. Τί συμπέρασμα μπορείτε να βγάλετε για την πρόβλεψη;

ii. Εκτιμήστε πάλι τους συντελεστές του παραπάνω μοντέλου παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων, αλλά αυτή τη φορά χρησιμοποιείστε όχι ολόκληρο το σύνολο δεδομένων αλλά μόνο εκείνες τις παρατηρήσεις όπου η τιμή επιφάνειας (μεταβλητή area) είναι

[27] 1 εκτάριο = 10 στρέμματα

μικρότερη από 3.2 εκτάρια (area < 3.2) και χαρακτηρίζει μικρές πυρκαγιές. Χρησιμοποείστε και πάλι διασταυρωτική επικύρωση 10-πτυχών και εκτιμήστε το μέσο τετραγωνικό σφάλμα (Root Mean Squared Error – RMSE) της πρόβλεψης. Τί συμπέρασμα μπορείτε να βγάλετε για την πρόβλεψη;

23. Υλοποιήστε σε R τον αλγόριθμο της Σταδιακής Καθόδου Μικρών Δεσμών (Mini Batch Gradient Descent).

24. Υλοποιήστε σε R τον αλγόριθμο παλινδρόμησης Ridge (Ridge regression), αν γίνει χρήση της μεθόδου ελαχίστων τετραγώνων.

25. Σχεδιάστε και υλοποιείστε στο περιβάλλον της R πρόγραμμα, το οποίο επιχειρεί να βρει τις βέλτιστες τιμές λ όταν χρησιμοποιείται παλινδρόμηση Ridge με τη μέθοδο της Σταδιακής Καθόδου Δεσμών για την εκτίμηση των συντελεστών του γραμμικού μοντέλου παλινδρόμησης $Κατανάλωση\ τροφίμων = \beta_1 Εισόδημα + \beta_0$. Χρησιμοποιείστε ως σύνολο εκπαίδευσης το σύνολο δεδομένων το αρχείο HouseholdData.csv .

ΠΑΡΑΡΤΗΜΑ Α

Περιεχόμενα του αρχείου HouseholdData.csv που περιέχει το σύνολο δεδομένων κατανάλωσης τροφίμων νοικοκυριών σε μορφή CSV που χρησιμοποιείται στον κώδικα R του κεφαλαίου.

```
Family, FoodExpenditure, Income, FamilySize, YearsOfEducationHH, GenderHH
1, 6100, 27000, 4, 9, 0
2, 5100, 26000, 4, 9, 0
3, 5800, 30000, 4, 10, 0
4, 5100, 24000, 4, 9, 0
5, 9300, 54000, 4, 12, 1
6, 9800, 59000, 2, 12, 1
7, 7800, 44000, 3, 10, 1
8, 5800, 30000, 3, 9, 0
9, 7500, 40000, 4, 10, 1
10, 11400, 82000, 2, 15, 1
11, 6900, 41000, 3, 10, 0
12, 8200, 58000, 3, 11, 1
13, 5700, 28000, 5, 9, 0
14, 6200, 20000, 5, 9, 0
15, 7900, 42000, 3, 11, 0
16, 8100, 47000, 3, 12, 1
17, 13300, 97000, 2, 18, 1
18, 12900, 85000, 2, 16, 1
19, 6200, 31000, 5, 10, 0
20, 5900, 26000, 4, 9, 0
21, 9200, 53000, 3, 12, 1
22, 7300, 49000, 3, 12, 1
23, 9200, 55000, 3, 12, 1
24, 7400, 40000, 5, 10, 0
25, 9800, 62000, 3, 15, 1
26, 7900, 46000, 4, 11, 1
27, 5900, 38000, 4, 10, 0
28, 7900, 45000, 4, 10, 1
29, 8400, 53000, 3, 11, 1
30, 8900, 49000, 4, 11, 1
31, 8600, 42000, 4, 10, 1
```

32, 8600, 42000, 5, 12, 0
33, 7500, 38000, 4, 10, 0
34, 13200, 90000, 3, 18, 1
35, 6900, 41000, 3, 9, 0

ΠΑΡΑΡΤΗΜΑ Β

Περιεχόμενα του αρχείου IcecreamRevenues.csv που περιέχει το σύνολο δεδομένων με τη μορφή CSV των εσόδων από πωλήσεις παγωτών και την ημερήσια θερμοκρασία που χρησιμοποιείται στον κώδικα R του κεφαλαίου.

Revenue, Temperature
26, 28.2
23, 21.4
32.9, 43
25.2, 30
31.4, 41
22.3, 19
21.5, 21
24, 27
24.9, 28.2
29.7, 36
27.4, 34.1
27.1, 32.5
23.8, 25
23.9, 22.5
26.1, 31.3
28.8, 38.1
29.8, 39.2
28.8, 36.7
27.7, 31.5
27.1, 29.3
27.8, 34.4
22.9, 23.3
23.2, 22.3
23.2, 26
23.7, 26.8
24.1, 27.7

27.2, 29.7
28.1, 36.2
26.9, 34.2
26.1, 30.1
25, 26.2
29.8, 37.2
28.5, 34.2
27.6, 33
26.9, 33.3

6.11 Απαντήσεις Ασκήσεων Αυτοαξιολόγησης

Άσκηση Αυτοαξιολόγησης 0.1

i. Μία μελέτη προπαθεί να εξακριβώσει εάν ηλικιωμένοι οδηγοί αυτοκινήτων εμπλέκονται σε περισσότερα ατυχήματα απ'ότι άλλοι οδηγοί. Ο αριθμός των ατυχημάτων ανά 100000 οδηγούς συγκρίνεται με την ηλικία του οδηγού.
Εξαρτημένη μεταβλητή: Ατυχήματα ανά 100000 κατοίκους
Ανεξάρτητη μεταβλητή: Ηλικία οδηγού

ii. Μία μελέτη προπαθεί να εξετάσει εάν το εβδομαδιαίο ποσό που ξοδεύει ένα νοικοκυριό στο super market μεταβάλλεται με τον αριθμού των ατόμων του νοικοκυριού.
Εξαρτημένη μεταβλητή: Εβδομαδιαίο ποσό που ξοδεύεται στο super market από το νοικοκυριό
Ανεξάρτητη μεταβλητή: αριθμός ατόμων νοικοκυριού

iii. Ασφαλιστικές εταιρείες καθορίζουν το πόσο θα πληρώνεται κάθε μήνα σε αφάλιστρα σε πολλά συμβόλαια βάσει της ηλικίας του ασφαλισμένου.
Εξαρτημένη μεταβλητή: Μηνιαίο ποσό ασφαλίστρων
Ανεξάρτητη μεταβλητή: Ηλικία ασφαλισμένου

iv. Ο λογαρασμός ρεύματος κυμαίνεται ανάλογα με την κατανάλωση ενός νοικοκυριού.
Εξαρτημένη μεταβλητή: Λογαριασμός ρεύματος
Ανεξάρτητη μεταβλητή: Κατανάλωση ρεύματος νοικοκυριού

v. Μία μελέτη προσπαθεί νε εξακριβώσει εάν το επίπεδο εκαπίδευσης των ατόμων (μετρούμενο σε έτη που βρίσκεται σε οποιαδήποτε εκπαιδευτική διαδικασία) μειώνει το ποσοστό εγκληματικότητας σε έναν πληθυσμό.

Εξαρτημένη μεταβλητή: Ποσοστό εγκληματικότητας
Ανεξάρτητη μεταβλητή: Επίπεδο εκπαίδευσης (σε έτη Π.χ. απόφοιτος λυκείου θα έχει 12, απόφοιτος πανεπιστημίου τετραετούς τμήματος 16 κλπ).

vi. Μία τράπεζα προσπαθεί να μελετήσει εάν το εισόδημα ενός ατόμου αποτελεί ένδειξη για το εάν θα πληρώνει το άτομο κανονικά τις δόσεις ενός δανείου ή όχι.

Αν θεωρηθεί ότι η εξαρτημένη μεταβλητή είναι εάν το άτομο θα πληρώνει κανονικά τις δόσεις δανείου ή οχι, το πρόβλημα αυτό δεν μπορεί να αναλυθεί με τη μέθοδο της πλαινδρόμησης. Και τούτο γιατί η εξαρτημένη μεταβλητή (αν το άτομο θα πληρώνει κανονικά τις δόσεις δανείου ή όχι) δεν είναι συνεχής μεταβλητή αλλά διακριτή που μάλιστα λαμβάνει δύο μη-αριθμητικές τιμές: Yes/No.

Άσκηση Αυτοαξιολόγησης 0.2

Πρόταση	**Σωστό**	**Λάθος**
Η γραμμή παλινδρόμησης ενός απλού μοντέλου παλινδρόμησης που απεικονίζεται σε ένα διάγραμμα διασποράς, συλλαμβάνει τον ρυθμό μεταβολής της εξαρτημένης μεταβλητής ως προς την ανεξάρτητη μεταβλητή.	X	
Ένα διάγραμμα διασποράς που δημιουργείται με τη συνάρτηση plot() της R, μπορεί να χρησιμοποιηθεί για την διαγραμματική διερεύνηση της σχέσης μεταξύ πέντε (5) μεταβλητών.		X
Ο σταθερός όρος β0 (συντελεστής β0) ενός απλού γραμμικού μοντέλου παλινδρόμησης εκφράζει την υπόθεση, ότι ο ρυθμός μεταβολής της ανεξάρτητης προς την εξαρτημένη μεταβλητή είναι σταθερός.		X

Άσκηση Αυτοαξιολόγησης 0.3

Εάν η συσχέτιση μεταξύ των μεταβλητών X, Y του απλού γραμμικού μοντέλου παλινδρόμησης είναι αρνητική, αυτό σημαίνει ότι όσο αυξάνεται η τιμή της ανεξάρτησης μεταβλητής X, τόσο μειώνεται η τιμή της εξαρτημένης Y. Κατά συνέπεια

η τιμή του συντελεστή β_1 που συλλαμβάνει ακριβώς αυτήν την σχέση πρέπει να είναι αρνητική ($\beta_1 < 0$).

Άσκηση Αυτοαξιολόγησης 0.4

Η σύγκριση των δύο αυτών μοντέλων θα γίνει συγκρίνοντας κάθε ένα από αυτά με το μοντέλο παλινδρόμησης βάσης, όπου η εξαρτημένη μεταβλητή είναι σταθερή και ίση με την μέση τιμή της ανεξάρτητης μεταβλητής στο σύνολο δεδομένων. Όποιο μοντέλο ερμηνεύει μεγαλύτερη διακύμανση σε σχέση με το μοντέλο αυτό βάσης, θεωρείται ότι εξηγεί καλύτερα τη διακύμανση της εξαρτημένης μεταβλητής.

Άσκηση Αυτοαξιολόγησης 0.5

Προκειμένου ένα μοντέλο παλινδρόμησης να χαρακτηριστεί ως γραμμικό, θα πρέπει ο ρυθμός μεταβολής της εξαρτημένης μεταβλητής να είναι γραμμικός ως προς όλους τους συντελεστές β του μοντέλου. Για να εξακριβωθεί αυτό, θα πρέπει να μελετηθούν οι ρυθμοί μεταβολής του μοντέλου παλινδρόμησης ως προς κάθε έναν συντελεστή (θεωρώντας τους υπόλοιπους σταθερούς) και ο ρυθμός μεταβολής να μην εξαρτάται από τον συντελεστή που μεταβάλλεται. Έτσι, αν γίνει η υπόθεση ότι οι συντελεστές μετβάλλονται κατά μία ποσότητα ε:
Για τον συντελεστή β_1, θεωρώντας τους συντελεστές β_2 και β_0 σταθερές:

$$\frac{\Delta Y}{\Delta \beta_1} = \frac{(\beta_1 + \varepsilon)X^{\beta_2} + \beta_0 - \left(\beta_1 X^{\beta_2} + \beta_0\right)}{\varepsilon} = \frac{\varepsilon X^{\beta_2}}{\varepsilon} = X^{\beta_2}$$

Επειδή ο ρυθμός μεταβολής είναι σταθερός (ανεξάρτητος του β_1), σημαίνει ότι το μοντέλο είναι γραμμικό ως προς τον συντελεστή β_1.
Για τον συντελεστή β_2, θεωρώντας τώρα τους συντελεστές β_1 και β_2 σταθερούς:

$$\frac{\Delta Y}{\Delta \beta_2} = \frac{\beta_1 X^{(\beta_2+\varepsilon)} + \beta_0 - \left(\beta_1 X^{\beta_2} + \beta_0\right)}{\varepsilon} = \frac{\beta_1 X^{\beta_2}(X^{\varepsilon} - 1)}{\varepsilon}$$

Ο ρυθμός μεταβολής του μοντέλου ως προς τον συντελεστή β_2 δεν είναι σταθερός, αφού εξαρτάται από την τιμή του συντελεστή β_2. Έτσι το μοντέλο δεν είναι γραμμικό ως προς τον συντελεστή β_2
Κατά συνέπεια το μοντέλο παλινδρόμησης που δίνεται δεν είναι γραμμικό προς όλους τους συντελεστές β κι έτσι είναι μη-γραμμικό.

Άσκηση Αυτοαξιολόγησης 0.6

Η τιμή της πρώτης παραγώγου μιας οποιασδήποτε παραγωγίσιμης συνάρτησης f(x) αναφέρει τον τρόπο με τον οποίο θα αλλάξουν οι τιμές της συνάρτησης φ(χ) αν μεταβληθούν οι τιμές της μεταβλητής x. Έτσι:

- αν η τιμή της πρώτης παραγώγου σε κάποιο σημείο x είναι θετική (>0), αυτό σημαίνει ότι αύξηση της τιμής της x θα οδηγήσει σε αύξηση της τιμής της συνάρτησης f(x).
- αν η τιμή της πρώτης παραγώγου είναι αρνητική (<0) σε κάποιο σημείο x, αυτό σημαίνει ότι αύξηση της τιμής της x θα οδηγήσει σε μείωση της τιμής της συνάρτησης f(x).
- αν η τιμή της πρώτης παραγώνου είναι ίση με το μηδέν (=0) σε κάποιο σημείο x, αυτό σημαίνει ότι η συνάρτηση f(x) εμφανίζει ακρότατο στο σημείο αυτό: δηλαδή τη μέγιστη ή ελάχιστη τιμή της συνάρτησης. Σε τέτοια περίπτωση αύξηση της τιμής της x θα έχει σαν αποτέλεσμα είτε αύξηση είτε μείωση της τιμής της συνάρτησης f(x).

Επειδή η συνάρτηση κόστους στη μέθοδο των ελαχίστων τετραγώνων είναι δευτέρου βαθμού, θα έχει ένα ακρότατο που είναι και η ελάχιστη τιμή της συνάρτησης. Θέτοντας την πρώτη παράγωγο της συνάρτησης κόστους ίση με το μηδέν, θα βρεθούν οι τιμές β που ελαχιστοποιούν τη συνάρτηση κόστους.

Άσκηση Αυτοαξιολόγησης 0.7

Η κανονική εξίσωση εκφράζεται με τη μορφή μήτρας, προκειμένου να διευκολυνθεί ο υπολογισμός των συντελεστών σε υπολογιστικά περιβάλλοντα. Όλα τα περιβάλλοντα προγραμματισμού παρέχουν βιβλιοθήκες με αποδοτικές μεθόδους για την εκτέλεση τέτοιων πράξεων.

Άσκηση Αυτοαξιολόγησης 0.8

Το όρισμα που αναπαριστά το γραμμικό μοντέλο παλινδρόμησης στη συνάρτηση R πρέπει να δοθεί ως τύπος δεδομένων (data type) formula της R και όχι π.χ. ως συμβολοσειρά (String). Η R παρέχει τον ειδικό αυτόν τύπο δεδομένων formula για τέτοιες περιπτώσεις ο οποίος ακολουθεί συγκεκριμένη σύνταξη. Είναι σημαντικό επίσης να τονιστεί, ότι ο τύπος δεδομένων formula ερμηνεύει τους τελεστές όπως +, ^ με διαφορετικό τρόπο από τον παραδοσιακό.

Άσκηση Αυτοαξιολόγησης 0.9

Στη μέθοδο Σταδιακής Καθόδου η συνάρτηση κόστους J(), επειδή διαιρεί με το πλήθος των παρατηρήσεων στο σύνολο εκπαίδευσης m, υπολογίζει επί της ουσίας το σφάλμα κατά μέσο όρο (το μέσο σφάλμα), και όχι ως απόλυτο αριθμό. Επειδή ακριβώς υπολογίζει το μέσο σφάλμα, μπορούν οι τιμές δύο διαφορετικών συναρτήσεων κόστους να συγκριθούν. Κάτι τέτοιο δεν μπορεί να γίνει για τη μέθοδο των ελαχίστων τετραγώνων, που υπολογίζει το απόλτο σφάλμα, και όχι το μέσο σφάλμα.

Άσκηση Αυτοαξιολόγησης 0.10

Όπως αναφέρθηκε, η τιμή της πρώτης παραγώγου μιας συνάρτησης σε ένα σημείο x αναφέρει προς ποια κατεύθυνση πρέπει να κινηθεί η τιμή της μεταβλητής x προκειμένου να αυξηθεί η να μειωθεί η τιμή της συνάρτησης. Έτσι για παράδειγμα, αν για τις τρέχουσες τιμές συντελεστών θi η τιμή της πρώτης παραγώγου $\frac{\partial}{\partial \theta_j} J(\theta)$ είναι θετική (>0), αυτό σημαίνει ότι αν αυξηθούν οι τιμές των θ_i τότε θα αυξηθεί και η τιμή J(θ), ενώ αν μειωθεί η τιμή των θi τότε θα μειωθεί και τιμή της συνάρτησης κόστους J(θ). Κατά συνέπεια, σε τέτοια περίπτωση πρέπει να μειωθούν οι συντελεστές θ_i εφόσον η παράσταση $\alpha \frac{\partial}{\partial \theta_j} J(\theta)$ είναι θετική αφού α > 0. Αντιθέτως, εάν η τιμή της πρώτης παραγώγου $\frac{\partial}{\partial \theta_j} J(\theta)$ της συνάρτησης κόστους είναι αρνητική, τότε πρέπει να αυξηθούν οι τιμές των συντελεστών θ_i προκειμένου να μειωθεί η συνάρτηση κόστους J(θ). Σε τέτοια περίπτωση η παράσταση $\alpha \frac{\partial}{\partial \theta_j} J(\theta)$ είναι αρνητική και κατά συνέπεια η $-\alpha \frac{\partial}{\partial \theta_j} J(\theta)$ θετική, που θα οδηγήσει σε αύξηση των τιμών των συντελεστών θ_i που είναι το ζητούμενο στην περίπτωση αυτή. Τέτοια συμπεριφορά επιτυχγάνεται μόνο αν ο συγκεκριμένος όρος αφαιρείται από τις τρέχουσες τιμές θ_i.

Άσκηση Αυτοαξιολόγησης 0.11

Ο έλεγχος θα γίνει, θέτοντας την παράμετρος α στην επιθυμητή τιμή και την εκτέλεση της μεθόδου της Σταδιακής Καθόδου, κρατώντας τις τιμές της συνάρτησης κόστους σε κάθε επανάληψη του αλγορίθμου. Ακολούθως θα αναπαρασταθεί γραφικά η τιμή της συνάρτησης κόστους ως συνάρτηση του πλήθους των ε-

παναλήψεων και εάν η απεικόνισή της έχει τη μορφή της εικόνας 6.9, τότε μπορεί να εξαχθεί το συμπέρασμα ότι η τιμή α που επιλέχθηκε είναι η κατάλληλη.

Άσκηση Αυτοαξιολόγησης 0.12

Έαν η τιμή της παραμέτρου μάθησης α λάβει τιμή ίση με το 0 (=0) αυτό θα έχει ως αποτέλεσμα ο όρος $\alpha \frac{\partial}{\partial \theta_j} J(\theta)$ να μηδενιστεί με αποτέλεσμα ο τύπος ενημέρωσης των συντελεστών να λάβει τη μορφή $\theta_j := \theta_j$ και κατά συνέπεια δεν θα γίνει καμία ενημέρωση των συντελεστών. Ο αλγόριθμος δεν θα τροποποιεί τους συντελεστές κατά τις επαναλήψεις του.
Εάν η τιμή της παραμέτρου μάθησης α λάβει τιμή μικρότερη από το μηδέν (<0) αυτό θα έχει ως αποτέλεσμα να αυξάνεται η τιμή των συντελεστών θ_i ενώ θα έπρεπε να μειώνεται, και να μειώνονται όταν θα πρέπει να αυξάνονται. Κατά συνέπεια, οι τιμές θ_i δεν θα συγκλίνουν προς και θα αποκλίνουν από την ελάχιστη τιμή της συνάρτηση J(θ).

Άσκηση Αυτοαξιολόγησης 0.13

Ο κώδικας R δεν θα εκτελεστεί κανονικά. Αυτό οφείλεται σε δύο λόγους:
Στη μήτρα με τις τιμές των ανεξαρτήτων μεταβλητών, θα πρέπει να υπάρχει στήλη με όλες τις τιμές της ίση με 1. Αυτό χρειάζεται για την αναπαράσταση του σταθερού όρου. Τετοια στήλη λείπει εντελώς στη μεταβλητή indVariables του κώδικα.
Επιπλέον, το πρώτο όρισμα της συνάρτησης gradientDescent() που παρουσιάστηκε στην ενότητα αυτή και αναπαριστά τις τιμές των ανεξαρτήτων μεταβλητών μαζί με τη στήλη των σταθερών όρων, πρέπει να είναι τύπου δεδομένων μήτρας (matrix) της R. Στο συγκεκριμένο κώδικα που δίνεται, το πρώτο όρισμα είναι τύπου δεδομένων διάνυσμα (Vector) και έτσι υπάρχει ασυμφωνία τύπων και δεν μπορούν να εφαρμοστούν οι πράξεις μητρών που έχει η συνάρτηση.

Άσκηση Αυτοαξιολόγησης 0.14

i. Τα δεδομένα εκπαίδευσης είναι τόσες πολλές παρατηρήσεις, που δεν χωράνε στην κεντρική μνήμη του υπολογιστή.
   Μπορεί να γίνει η χρήση της μεθόδου της Σταδιακής Καθόδου Μικρών Δεσμών ή της Στοχαστικής Σταδιακής Καθόδου.

ii. Τα δεδομένα εκπαίδευσης δεν είναι διαθέσιμα εξ'αρχής και προσέρχονται μία παρατήρηση τη φορα με άγνωστο ρυθμό.
Μπορεί να γίνει η χρήση μόνο της μεθόδου της Στοχαστικής Σταδιακής Καθόδου.

iii. Επιθυμείται η βέλτιστη εκτίμηση συντελεστών
Μπορεί να επιτευχθεί μόνο με τη μέθοδο των ελαχίστων τετραγώνων.

Άσκηση Αυτοαξιολόγησης 0.15

Εννοείται, αντί να χρησιμοποιηθεί η εξαρτημένη μεταβλητή ως έχει, να αντικατασταθεί από τον λογάριθμό της. Δηλαδή, αν το μοντέλο παλινδρόμησης έχει τη μορφή

$$Y = \beta_1 X_1 + \beta_2 X_2 + \beta_0$$

Να μετασχηματιστεί ως

$$\log(Y) = \beta_1 X_1 + \beta_2 X_2 + \beta_0$$

Αυτό μπορεί να συμβεί μόνο εφόσον η εξαρτημένη μεταβλητή λαμβάνει μόνο θετικές τιμές.

Άσκηση Αυτοαξιολόγησης 0.16

Μία τιμή του συντελεστή προσδιορισμού R2 ίση με 0.05 σημαίνει ότι το γραμμικό μοντέλο παλινδρόμησης μπορεί να ερμηνεύσει το 5% της διακύμανσης της εξαρτημένης μεταβλητής

Άσκηση Αυτοαξιολόγησης 0.17

Οι τιμές του συντελεστή προσδιορισμού R^2 και διορθωμένου συντελεστή προσδιορισμού R^2_{adj} είναι ίσες μόνο στην περίπτωση όπου το πλήθος των μεταβλητών στο μοντέλο παλινδρόμησης είναι ίσο με το 0, δηλαδή δεν έχει καμία ανεξάρτητη μεταβλητή και η τιμή της εξαρτημένης μεταβλητής είναι μία σταθερά. Αυτό προκύπτει από την εξής συνθήκη:

$$R^2 = R^2_{adj} \Leftrightarrow R^2 = 1 - \left(\frac{(1-R^2)(n-1)}{(n-k-1)}\right) \Leftrightarrow k = 0$$

Άσκηση Αυτοαξιολόγησης 0.18

Η p-τιμή ίση με 0.03 με επίπεδο σημαντικότητας 5% σημαίνει ότι μπορεί να απορριφθεί η μηδενική υπόθεση ότι ο αντίστοιχος συντελεστής είναι ίσος με το μηδέν (0).

Άσκηση Αυτοαξιολόγησης 0.19

Μία p-τιμή ίση με 0.02 στο επίπεδο σημαντικότητας 1% σημαίνει ότι η μηδενική υπόθεση ότι ο συντελεστής είναι ίσος με το μηδέν δεν μπορεί να απορριφθεί και κατά συνέπεια ότι η αντίστοιχη εξαρτημένη μεταβλητή δεν έχει σημαντική επίδραση στην τιμή της εξαρτημένης μεταβλητής. Δηλαδή υπάρχει σχέση μεταξύ των μεταβλητών αλλά αυτή δεν είναι στατιστικά σημαντική και θα μπορούσε κάλλιστα να οφείλεται στην τύχη.

Άσκηση Αυτοαξιολόγησης 0.20

Αν και η ερμηνεία του σταθερού όρου δεν είναι πάντα εφικτή και εξαρτάται από το πρόβλημα που μελετάται, στη συγκεκριμένη περίπτωση θα μπορούσε να εκφράζει το ποσό που ξοδεύει ένα άτομο που δεν έχει εισόδημα (Εισόδημα = 0) και δεν έχει νοικοκυριό (FamilySize=0) δηλαδή ζει μόνος του. Επειδή και τέτοια άτομα πρέπει να τραφούν για να ζήσουν θα πρέπει να ξοδεύουν ένα μέρος χρημάτων για την κατανάλωση τροφίμων. Ο συντελεστής β_0 θα μπορούσε να ερμηνευτεί με τέτοιο τρόπο στην περίπτωση αυτή κάνοντας ορισμένες υποθέσεις ειδικά για το πως ερμηνεύεται μία τιμή της μεταβλητής FamilySize ίση με το 0. Αυτά είναι ζητήματα της μεθοδολογίας και της κωδικοποίησης των τιμών των μεταβλητών που θα πρέπει να έχουν οριστεί εξ'αρχής.

Άσκηση Αυτοαξιολόγησης 0.21

Όταν λέγεται ότι επιθυμείται η τιμωρία των μεγάλων τιμών των καταλοίπων σημαίνει ότι, όταν εμφανιστούν τέτοιες τιμές, αυτές να αυξάνουν κατά πολύ τη συνάρτηση κόστους. Η τιμωρία εδώ έρχεται με την πολύ μεγάλη αύξηση της συνάρτησης κόστους. Επειδή οι μεγάλες τιμές καταλοίπων, όταν υψωθούν στο τετράγωνο, μεγαλώνουν ακόμη πιο πολύ, θα μεγαλώσουν ακόμη περισσότερο τη συνάρτηση κόστους και κατά συνέπεια θα προσθέσουν μία μεγάλη τιμή «τιμωρί-

ας». Έτσι, από την τιμή της συγκεκριμένης συνάρτησης κόστους μπορούν να βγουν συμπεράσματα για το εάν υπάρχουν μεγάλες τιμές καταλοίπων.

Άσκηση Αυτοαξιολόγησης 0.22

Όχι, η μέθοδος εκτίμησης των συντελεστών δεν επηρεάζει τον τρόπο αξιολόγησης του μοντέλου, εάν ο στόχος είναι η πρόβλεψη. Για παράδειγμα, η μέθοδος της Διασταυρωτικής επικύρωσης k- πτυχών μπορεί να χρησιμοποιηθεί τόσο αν γίνει εκτίμηση συντελεστών με τη μέθοδο ελαχίστων τετραγώνων όσο και με οποιαδήποτε εκδοχή της μεθόδου της Σταδιακής Καθόδου.

Άσκηση Αυτοαξιολόγησης 0.23

Η συνάρτηση

```
folds <- cut(seq(1,nrow(dataset)), breaks=k, labels=FALSE)
```

θα δημιουργήσει ένα νέο διάνυσμα που θα έχει τις ακόλουθες τιμές
1,1,1,1,1,2,2,2,2,2,3,3,3,3,3,...
Το πλήθος των τιμών στο διάνυσμα αυτό θα είναι όσο και το πλήθος του συνόλου δεδομένων dataset, έτσι ώστε κάθε μία τιμή της folds να αντιστοιχίζεται με μία παρατήρηση στο σύνολο δεδομένων dataset. Εδώ π.χ. στην πρώτη παρατήρηση του data.frame dataset θα δοθεί η τιμή 1, στη δεύτερη επίσης η τιμή 1, ενώ στην έκτη, έβδομη θα αντιστοιχιθούν η τιμή 2. Με τον τρόπο αυτόν επί της ουσίας ομαδοποιούνται οι παρατηρήσεις του συνόλου dataset σε k ομάδες κάθε μία από τις οποίες λαμβάνει μοναδικό αριθμό. Με αυτόν τον τρόπο είναι εύκολο να απομονωθεί το υποσύνολο ελέγχου από το σύνολο εκπαίδευσης

Άσκηση Αυτοαξιολόγησης 0.24

Η συγκεκριμένη περιγραφή θέσης είνα παράδειγμα υπερπροσαρμογής (overfitting). Για την εταιρεία, το σύνολο εκπαίδευσης είναι ο μηχανικός που συνταξιοδοτήθηκε και επειδή έμεινε ευχαριστημένη από τις επιδόσεις του προσπάθησε να δημιουργήσει ένα μοντέλο μηχανικού (που είναι η περιγραφή των προσόντων όπως φαίνεται στην ανάρτηση) όσο γίνεται πιο κοντά στον μηχανικό που συνταξιοδοτήθηκε. Ωστόσο, το μοντέλο που δημιουργήθηκε (τα προσόντα) συλλαμβάνουν και άσχετες με το αντικείμενο πτυχές του εργαζόμενου. Αυτό έχει σαν απο-

τέλεσμα, η περιγραφή αυτή να συλλαμβάνει τέλεια τα χαρακτηριστικά του εργαζομένου που έφυγε (μικρό σφάλμα εκπαίδευσης – εδώ ως σφάλμα νοείται πόσο κοντά ο μηχανικός είναι στα ζητούμενα προσόντα) και ταυτόχρονα να μην μπορεί να ταιριάξει με τα προσόντα κανενός άλλου μηχανικού λογισμικού που αναζητούν εργασία (μεγάλο σφάλμα ελέγχου) αφού η περιγραφή της θέσης είναι τόσο αναλυτική ακόμη και σε άσχετες με το αντικείμενο πτυχές.

Άσκηση Αυτοαξιολόγησης 0.25

Εάν κατά την παλινδρόμηση Ridge επιλεγεί μία τιμή της παραμέτρου κανονικοποίησης λ πολύ μεγάλη, αυτό θα έχει σαν αποτέλεσμα οι συντελεστές β να λάβουν πολύ μικρές τιμές και να τείνουν προς το 0. Αυτό γιατί η ελαχιστοποίηση της συνάρτησης κόστους θα πρέπει να ελαχιστοποιήσει και τον όρο κανονικοποίησης $\lambda \sum_{j=1}^{n} \left|\beta_j\right|^2$ και επειδή η τιμή λ είναι πάρα πολύ μεγάλη, η ελαχιστοποίηση του όρου αυτού θα συμβεί μόνο με πολύ μικρές τιμές των συντελεστών β_j.

Ενδεικτικές Απαντήσεις Δραστηριοτήτων

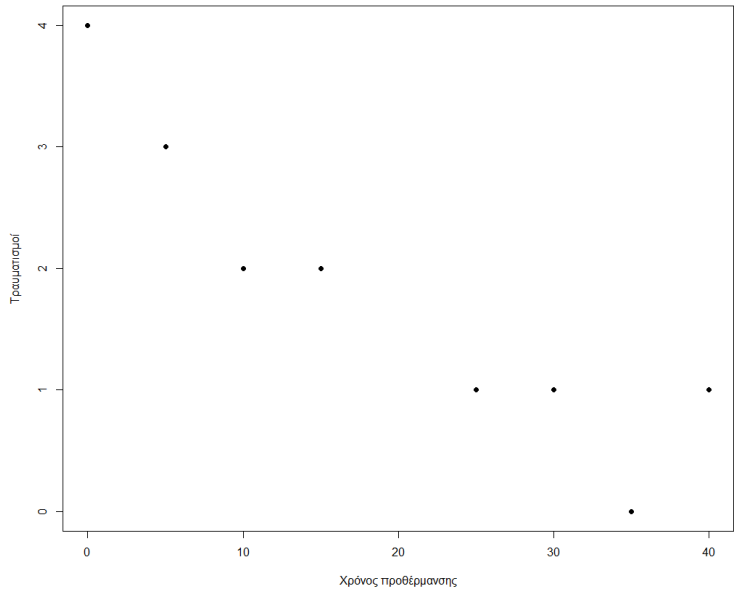
Δραστηριότητα 0.1

Η σχέση/συσχέτιση μεταξύ των μεταβλητών μπορεί να μελετηθεί με τη δημιουργία διαγράμματος διασποράς, που θα δώσει τις ενδείξεις για το είδος της συσχέτισής τους. Παρακάτω φαίνεται ο κώδικας σε R:

```
# Δημιουργία data.frame με τις δύο μεταβλητές. Οι τιμές

# των μεταβλητών δίνονται στην μορφή διανυσμάτων.

# WarmupTime = Χρόνος προθέρμανσης

# Injuries = Τραυματισμοί

injuryData <- data.frame("WarmupTime" = c(0, 30, 10,
15,5,25,35,40), "Injuries" = c(4,1,2,2,3,1,0,1))

# Εμφάνιση data.frame για την επιβεβαίωση ότι τα δεδομένα έχουν
καταχωρηθεί ορθά


print(injuryData)
```

```
# Δημιουργία διαγράμματος διασποράς για την απεικόνιση της σχέσης μεταξύ των δύο

# αυτών μεταβλητών.

# Στον άξονα X εμφανίζεται ο χρόνος προθέρμανσης και στον άξονα y το πλήθος τραυματισμών.

# Από το διάγραμμα προκύπτει μία αρνητική σχέση μεταξύ των μεταβλητών αυτών.

# Το διάγραμμα διασποράς θα μπορούσε επίσης να απεικονιστεί έχοντας στο άξονα X το πλήθος

# τραυματισμών και στο άξονα y τον χρόνο προθέρμανσης. Και σε ένα τέτοιο διάγραμμα θα

# φαίνονταν πάλι η αρνητική συσχέτιση.

plot( injuryData$WarmupTime, injuryData$Injuries, xlab="Χρόνος προθέρμανσης", ylab="Τραυματισμοί")
```

Η εκτέλεση του παρπάνω κώδικα θα απεικονίσει το διάγραμμα διασποράς που φαίνεται παρακάτω:



Από το διάγραμμα διασποράς φαίνεται να υπάρχει συσχέτιση μεταξύ των μεταβλητών αυτών και η συσχέτιση των μεταβλητών είναι αρνητική. Αυτό γιατί όσο αυξάνεται η μεταβλήτή Χρόνος προπόνησης τόσο μειώνεται το πλήθος τραυματισμών.

Δραστηριότητα 0.2

Το πρόβλημα που παρουσιάζεται είναι ότι οι πολυωνυμικοί όροι που εμφανίζονται σε ένα μοντέλο παλινδρόμησης πρέπει να δοθούν με ειδικό τρόπο, αν χρησιμοποιούνται σε τύπους (formula) στην R. Ειδικότερα, εάν το μοντέλο παλινδρόμησης προσδιοριστεί στην εντολή lm() με τον ακόλουθο τρόπο:

```
houseHoldData<-read.csv("HouseholdData.csv", sep=",", header=T)

linear.regression<-lm(FoodExpenditure ~ Income + FamilySize + Fami-
lySize^2, data=houseHoldData)

print(linear.regression$coefficients)
```

θα εμφανιστεί το εξής αποτέλεσμα:

```
(Intercept)      Income  FamilySize

436.9299573   0.1601037 159.6608893
```

Όπως φαίνεται, απουσιάζει παντελώς ο συντελεστής για τον όρο FamilySize^2 από το αποτέλεσμα. Αυτό οφείλεται στο γεγονός, ότι ο τελεστής ^ έχει ειδική ερμηνεία στα πλαίσια των τύπων (formula) στην R και κατά συνέπεια δεν ερμηνεύεται ως ο τελεστής ύψωσης σε δύναμη.

Ο ορθός τρόπος προσδιορισμού του γραμμικού μοντέλου παλινδρόμησης θα πρέπει να κάνει χρήση της ειδικής συνάρτησης I() (Inhibit Interpretation /Conversion of Objects) η οποία αναστέλλει την ερμηνεία του τελεστή ^ από τον τύπο και έτσι θα ερμηνευτεί ως ο τελεστής ύψωσης σε δύναμη που είναι το επιθυμητό. Στα πλαίσια τύπων, ο τελεστής ^ ερμηνεύεται ως ο τελεστής επιπτώσεων αλληλεπιδράσεων (interaction effects) μεταξύ μεταβλητών.

Έτσι, το μοντέλο παλινδρόμησης θα πρέπει να προσδιοριστεί όπως φαίνεται παρακάτω, για να εκληφθεί από τον τύπο ο όρος FamilySize^2 ως πολυώνυμικός:

```
houseHoldData<-read.csv("HouseholdData.csv", sep=",", header=T)

linear.regression<-lm(FoodExpenditure ~ Income + FamilySize +
I(FamilySize^2), data=houseHoldData)

print(linear.regression$coefficients)
```

Η εκτέλεση του κώδικα θα δώσει σαν αποτέλεσμα

```
(Intercept)        Income      FamilySize        I(FamilySize^2)

7593.5532927       0.1463769   -3523.5760350     490.7063296
```

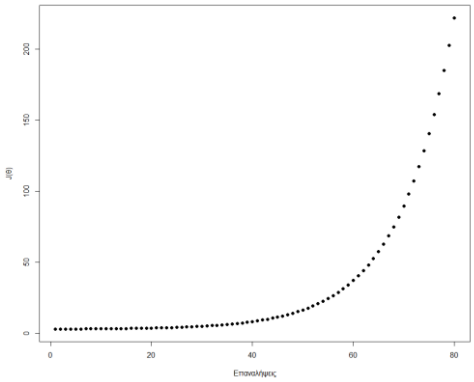
Αφού εμφανίζεται ο όρος I(FamilySize^2) στους συντελεστές, αυτό σημαίνει ότι η δήλωση ερμηνεύτηκε ορθά ως πολυωνυμικός όρος. Για περισσότερες πληροφορίες για τον τύπο δεδομένων formula της R, ανατρέξτε στη σχετική σελίδα βοήθειας.

Δραστηριότητα 0.3

Εάν τιμές της παραμέτρου μάθησης α και του πλήθους επαναλήψεων λάβουν τις τιμές α=0.002111 και πλήθος επαναλήψεων = 80 και η κλήση της συνάρτησης gradientDescent() στον κώδικα R της ενότητας γίνει με αυτές ως εξής:

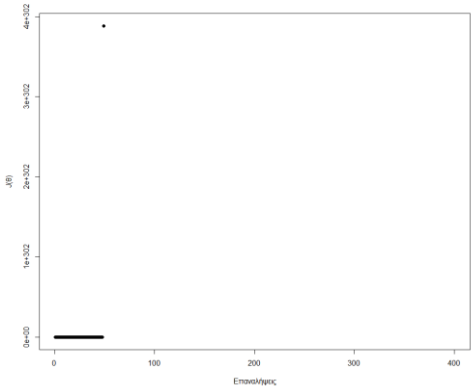
```
gdOutput<-gradientDescent(indVariables, revenue, initialThetas,
0.002111, 80)

# Αλλαγή του τρόπου εμφάνισης των στιγμών (pch=16) ώστε να εμφανί-
ζονται με μαύρο

# χρώμα για την καλύτερη απεικόνιση των σημείων.

plot(gdOutput$costs, xlab="Επαναλήψεις", ylab="J(θ)", pch=16)
```

θα εμφανιστεί η απόκλιση της συνάρτησης κόστους από την ελάχιστη τιμή της όπως φαίνεται από το παρακάτω διάγραμμα διασποράς της συνάρτησης κόστους. Όπως φαίνεται, ακόμη και για μικρές τιμές των παραμέτρων αυτών, μπορεί να παρατηρηθεί η απόκλιση της συνάρτησης κόστους. Γενικά, η μέθοδος της Σταδιακής Καθόδου είναι ευαίσθητη στις τιμές των παραμέτρων αυτών.



Προφανώς υπάρχουν και άλλες τιμές των παραμέτρων αυτών που οδηγούν στο ίδιο αποτέλεσμα.

Εάν η αναζήτηση των κατάλληλων τιμών των παραμέτρων α και πλήθους επαναλήψεων οδήγησε σε διάγραμμα διασποράς όμοιο με το παρακάτω



αυτό και πάλι σημαίνει απόκλιση της συνάρτησης κόστους. Τέτοια εμφάνιση οφείλεται στο γεγονός, ότι η τιμή της παραμέτρου μάθησης α ή/και το πλήθος επαναλήψεων ήταν πάρα πολύ μεγάλες και οδήγησαν σε τεράστιες τιμές της συνάρτησης κόστους που δεν μπορούν να αναπαρασταθούν ή να εμφανιστούν (αυτό φαίνεται και στις τιμές της συνάρτησης κόστους στον άξονα y). Σε τέτοια περίπτωση αν γίνει η εμφάνιση των τιμών της συνάρτησης κόστους που προέκυψαν μέσω της μεταβλητής gdOutput$costs θα εμφανιστούν πέραν ορισμένων αριθμητικών τιμών και οι τιμές Inf ή/και NaN. Αυτό σηματοδοτεί την υπερχείλιση της τιμής της συνάρτησης κόστους.

Δραστηριότητα 0.4

Με τον παρακάτω κώδικα R, εκτιμώνται οι συντελεστές του μοντέλου παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων (OLS) και εμφανίζονται τόσο οι συντελεστές που εκτιμήθηκαν όσο και τα στοιχεία που ζητούνται:

```
injuryData <- data.frame("WarmupTime" = c(0, 30, 10,
15,5,25,35,40), "Injuries" = c(4,1,2,2,3,1,0,1))

# Εκτίμηση συντελεστών του μοντέλου

# Injuries = β1Χρόνος προθέρμανσης + b0

# με τη μέθοδο ελαχίστων τετραγώνων (OLS)

linear.model<- lm( Injuries ~ WarmupTime, data =  injuryData)

# Εμφάνιση συντελεστή προσδιορισμού R^2 και στατιστικής σημαντικό-
τητας ανεξάρτητης μεταβλητής

summary(linear.model)
```

Η εκτέλεστη του κώδικα R θα δώσει ως αποτέλεσμα:

```
Call:

lm(formula = Injuries ~ WarmupTime, data = injuryData)

Residuals:

   Min     1Q Median     3Q    Max

 -0.55  -0.40  -0.05   0.20   0.85

Coefficients:

            Estimate Std. Error t value Pr(>|t|)

(Intercept)  3.35000    0.35218   9.512  7.7e-05 ***

WarmupTime  -0.08000    0.01453  -5.506  0.00151 **

---

Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.5627 on 6 degrees of freedom
```

```
Multiple R-squared:  0.8348,     Adjusted R-squared:  0.8072
F-statistic: 30.32 on 1 and 6 DF,  p-value: 0.001506
```

Από τα παραπάνω στοιχεία προκύπτει ότι:

1) Οι τιμές των συντελεστών είναι: β_1=-0.08 για την μεταβλητή Χρόνος προπόνησης και β_0=3.35 για τον σταθερό όρο (Στήλη Estimate των αποτελεσμάτων). Οι τιμές των συντελεστών μπορούν επίσης να εμφανιστούν μέσω της μεταβλητής $coefficients: print(linear.model$coefficients) .
2) Το γραμμικό μοντέλο παλινδρόμησης είναι ικανό να ερμηνεύσει το 83.48% της διακύμανσης της εξαρτημένης μεταβλητής (Multiple R-squared)
3) Η μεταβλητή Χρόνος προθέρμανσης είναι στατιστικά σημαντική αφού η p-τιμή της είναι 0.00151 (<5%). Τα δύο αστεράκια που υπάρχουν δίπλα στην τιμή του συντελεστή δηλώνουν το επίπεδο σημαντικότητας και ειδικά ότι η μεταβλητή είναι στατιστικά σημαντική στο επίπεδο σημαντικότητας 0.001 (ή 0.1%) όπως φαίνεται από τη γραμμή με ετικέτα Signif. codes .