Άλυτες ασκήσεις Κεφαλαίου 6: Παλινδρόμηση

ΒΕΡΥΚΙΟΣ ΒΑΣΙΛΗΣ, ΣΤΑΥΡΟΠΟΥΛΟΣ ΗΛΙΑΣ, ΚΩΤΣΙΑΝΤΗΣ ΣΩΤΗΡΗΣ, ΤΖΑΓΚΑΡΑΚΗΣ ΜΑΝΩΛΗΣ: Η Επιστήμη των Δεδομένων: Βασικές Αρχές, Θεωρία & Εφαρμογές με τη Γλώσσα R", Εκδόσεις Νέων Τεχνολογιών, 2019

v0.7
rd28/07/2020
tzagara@upatras.gr

Ασκήσεις

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

2. Ποιο από τα παρακάτω διαγράμματα είναι το καταλληλότερο για την εξερεύνηση της συσχέτισης μεταξύ δύο μεταβλητών (εξαρτημένης και ανεξάρτητης) σε ένα σύνολο δεδομένων; Τεκμηριώστε την απάντησή σας.

Α	Ένα διάγραμμα διασποράς (Scatter plot)
Β	Ένα ραβδόγραμμα (Barchart)
Γ	Ένα ιστόγραμμα (Histogram)
Δ	Ένα θηκόγραμμα (Box plot)

3. Δίνεται το παρακάτω σύνολο δεδομένων που έχει τις τιμές δύο μεταβλητών 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 ταιριάζει καλύτερα στα παραπάνω δεδομένα:

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

4. Ποια από τα παρακάτω μοντέλα παλινδρόμησης είναι γραμμικά και ποια μη γραμμικά; Τεκμηριώστε την απάντησή σας. Στα μοντέλα που παρουσιάζονται παρακάτω, θεωρείστε ότι β_i είναι οι συντελεστές του μοντέλου.

i) $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2$

ii) $y = \beta_0 + \beta_1 X_1^3 + \beta_2 X_2^5$

Iii) $y = 42$

vi) $\ln(y) = \beta_0 + \beta_1 \ln(X_1) + \beta_2 log_2(X_2)$

vii) $y = \frac{\beta_1 X_1}{\beta_2 + X_1}$

viii) $y = \beta_1 X^2 + \beta_2 X^3 + \beta_3 X^4 + \beta_3 X^5 + \beta_3 X^7 + \beta_3 X^9 + \beta_3 X^{10} + \beta_0$

iv) $y = \beta_0 + \beta_1 \ln(X_1) + \beta_2 \ln(X_2)$

v) $y = \beta_0 + e^{\beta_1 X_1} + \beta_2 X_2$

ix) $y = \dfrac{1}{\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3}$

x) $y = \beta_1 X_1 + \beta_2 \sqrt{X_2} + \beta_0$

5. Επιθυμούμε να κάνουμε χρήση μεθόδου παλινδρόμησης για την μελέτη της συσχέτισης μεταξύ των μεταβλητών που αναφέρονται στα παρακάτω σενάρια. Για κάθε μία από τις παρακάτω περιπτώσεις σεναρίων, προσδιορίστε ποια είναι η εξαρτημένη και ποιες οι ανεξάρτητες μεταβλητές. Για τις ανεξάρτητες μεταβλητές, αναφέρετε το είδος της (αν είναι ονομαστική, διατάξιμη, διαστήματος ή αναλογίας) και παραδείγματα τιμών που αυτές μπορούν να λάβουν:

i)	Μια ασφαλιστική εταιρεία θέλει να μελετήσει εάν η ηλικία, το φύλλο και εισόδημα επηρεάζουν το πλήθος αυτοκινητιστικών ατυχημάτων.
ii)	Ένα εργαστήριο θέλει να μελετήσει εάν η καθημερινή άσκηση, το φύλο και το πρόγραμμα διατροφής ενός ατόμου συμβάλλει στον έλεγχο της αρτηριακής πίεσης του ατόμου.
iii)	Ένα online shop θέλει να μελετήσει εάν το επίπεδο εκπαίδευσης των καταναλωτών και ο χρόνος που ξοδεύουν στο διαδίκτυο σχετίζεται με το πλήθος αγορών προϊόντων από τον ιστότοπο της επιχείρησης.
iv)	Η εφορία θέλει να μελετήσει εάν το ύψος των προστίμων στους ελεύθερους επαγγελματίες για φορολογικά παραπτώματα σχετίζεται με το επάγγελμά τους, το εισόδημά τους, την οικογενειακή τους κατάσταση και το εμβαδό του σπιτιού τους.
v)	Ένας ερευνητής θέλει να μελετήσει εάν το εισόδημα, η εκπαίδευση και το ζώδιο ενός ατόμου σχετίζεται με το ζώδιό του.
vi)	Μια επιχείρηση θέλει να μελετήσει εάν το ποσό που ξοδεύει για την διαφήμιση ενός προϊόντος κάθε μήνα καθώς και το εάν η διαφήμιση εμφανίζεται στην τηλεόραση, το ραδιόφωνο ή την εφημερίδα σχετίζεται με τις πωλήσεις του προϊόντος αυτού.

6. Γιατί η χρησιμοποιείται ο όρος «παλινδρόμηση» (“regression”) για τον χαρακτηρισμό τέτοιας ανάλυσης της συσχέτισης μεταξύ της ανεξάρτητης και των εξαρτημένων μεταβλητών; Δηλαδή, πως προέκυψε το όνομα «παλινδρόμηση»; Αναζητήστε στο διαδίκτυο την προέλευση του όρου.

7. Ο Francis Galton ήταν από τους πρώτους που έκανε χρήση των εννοιών της συσχέτισης μεταξύ μεταβλητών και παλινδρόμησης. Αναζητήστε στο διαδίκτυο το βιογραφικό του και αναφέρετε 2 (δύο) ιστορίες από τη ζωή του που σας έκαναν να πείτε “WTF” καθώς επίσης και “OMG και 3 LoL”.

8. Αποδείξτε, ότι στη μέθοδο των ελαχίστων τετραγώνων (OLS), για ένα απλό γραμμικό μοντέλο παλινδρόμησης της μορφής

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

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

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}$$

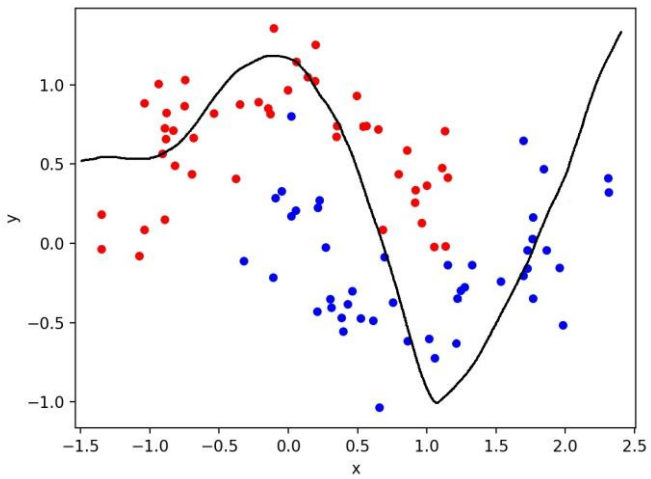
$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X}$$

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

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

10. Εξηγείστε γιατί η κανονική εξίσωση (normal equation), στη γενική περίπτωση, δεν μπορεί να χρησιμοποιηθεί για την εκτίμηση συντελεστών μη-γραμμικών μοντέλων παλινδρόμησης; Εξετάστε αν υπάρχουν περιπτώσεις μη-γραμμικών μοντέλων παλινδρόμησης όπου μπορεί να γίνει η χρήση της κανονικής εξίσωσης για την εκτίμηση των συντελεστών του μοντέλου.

11. 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 που δημιούργησε τη γραμμή παλινδρόμησης που φαίνεται στο διάγραμμα να είναι γραμμικό; Τεκμηριώστε την απάντησή σας.



12. Ένα γραμμικό μοντέλο παλινδρόμησης, μπορεί να έχει ανεξάρτητες μεταβλητές οι οποίες είναι κατηγορικές/ονομαστικές (διατάξιμες ή μη);

13. Με ποιον τρόπο χειρίζονται[1] οι κατηγορικές ανεξάρτητες μεταβλητές που υπάρχουν σε ένα γραμμικό μοντέλο παλινδρόμησης από τις μεθόδους εκτίμησης συντελεστών τέτοιων μοντέλων; Πως πρέπει να ερμηνεύονται οι συντελεστές που εκτιμώνται από τις μεθόδους τέτοιων κατηγορικών ανεξάρτητων μεταβλητών, οι οποίες είναι στατιστικά σημαντικές;

14. Κατεβάστε από τον ιστότοπο "UCI Machine Learning Repository", και ειδικότερα από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set το σύνολο δεδομένων που αναφέρει ορισμένα χαρακτηριστικά ακινήτων καθώς και τις τιμές των ακινήτων αυτών σε διάφορες περιοχές της πρωτεύουσας της Ταϊβάν, Ταϊπέϊ. Αφού εξοικειωθείτε με το σύνολο δεδομένων που έχετε κατεβάσει, απαντήστε στα παρακάτω ερωτήματα:

[1] Δηλαδή με ποιον τρόπο νοούνται οι τιμές τέτοιων κατηγορικών μεταβλητών, εφόσον οι μέθοδοι αυτές κάνουν χρήση μαθηματικών τύπων για την εύρεση των συντελεστών.

i. Συγγράψτε κώδικα R ο οποίος απεικονίζει γραφικά, τις τιμές των ακινήτων συναρτήσει της ηλικίας του ακινήτου. Απεικονίστε στο ίδιο γράφημα, τη μέση τιμή των τιμών των ακινήτων. Κοιτάζοντας το διάγραμμα που έχει δημιουργηθεί, τί έχετε να παρατηρήσετε σχετικά με τη συσχέτιση που υπάρχει μεταξύ της τιμής του ακινήτου και της ηλικίας του;

ii. Συγγράψετε πρόγραμμα σε R, το οποίο εκτιμά τους συντελεστές του παρακάτω απλού μοντέλου γραμμικής παλινδρόμησης με τη μέθοδο των ελαχίστων τετραγώνων (OLS) χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων:

$$House\ price\ of\ unit\ area = \beta_1 House\ Age + \beta_0$$

Αφού εκτιμήσετε τους συντελεστές, απαντήστε στις παρακάτω ερωτήσεις:

a) Εκφράστε με λόγια ποια συσχέτιση μεταβλητών επιχειρεί να εκτιμήσει το μοντέλο που αναφέρετε παραπάνω.

b) Προσθέστε τις κατάλληλες εντολές στο πρόγραμμά σας ώστε να εμφανίζονται μόνο οι συντελεστές που έχουν εκτιμηθεί. Αναφέρετε τις τιμές του συντελεστή του σταθερού όρου (β_0) και τον συντελεστή της μεταβλητής House Age (β_1) που έχουν προκύψει.

c) Προσθέστε τις κατάλληλες εντολές στο πρόγραμμα που εκτιμά τους συντελεστές του μοντέλου, για να εμφανιστούν τα 1) κατάλοιπα και 2) οι τιμές της εξαρτημένης μεταβλητής όπως τις υπολογίζει το μοντέλο που έχει εκτιμηθεί για όλες τις τιμές της εξαρτημένης μεταβλητής που υπάρχουν στο αρχείο.

d) Συγκρίνετε το άθροισμα των τιμών της εξαρτημένης μεταβλητής House price of unit area που υπολογίζει το μοντέλο που εκτιμήσατε για όλες τις τιμές της ανεξάρτητης μεταβλητής House Age του συνόλου δεδομένων, με το άθροισμα όλων των τιμών της εξαρτημένης μεταβλητής House price of unit area που υπάρχουν στο σύνολο δεδομένων. Τί έχετε να παρατηρήσετε; Αποδείξτε ότι η σχέση αυτή θα ισχύει για όλα τα απλά γραμμικά μοντέλα παλινδρόμησης.

15. Κατεβάστε από τον ιστότοπο "UCI Machine Learning Repository", και ειδικότερα 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. Απαντήστε στα παρακάτω ερωτήματα:

i. Αφού εξοικειωθείτε με τα γνωρίσματα που υπάρχουν στο αρχείο δεδομένων και τη σημασία τους, αναφέρετε τί εκφράζει κάθε μεταβλητή (εξαρτημένη και ανεξάρτητες) που υπάρχει στο παρακάτω πολλαπλό γραμμικό μοντέλο παλινδρόμησης:

$$ViolentCrimesPerPop = \beta_1 NumStreet + \beta_2 HousVacant + \beta_3 medIncome + \beta_4 whitePerCap + \beta_5 blackPerCap + \beta_6 HispPerCap + \beta_0$$

ii. Συγγράψτε πρόγραμμα σε R που εκτιμά με τη μέθοδο των ελαχίστων τετραγώνων (OLS) τους συντελεστές του πολλαπλού γραμμικού μοντέλου παλινδρόμησης που δίνεται στο υποερώτημα i) παραπάνω, χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων που κατεβάσατε ως σύνολο εκπαίδευσης.

iii. Προσθέστε στο πρόγραμμά σας την κατάλληλη εντολή για να εμφανιστούν μόνο οι συντελεστές των ανεξάρτητων μεταβλητών που έχουν εκτιμηθεί για το παραπάνω μοντέλο.

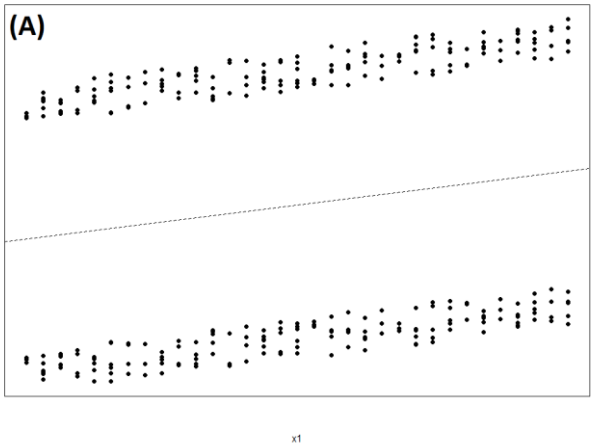
16. Συγγράψτε πρόγραμμα σε R που εκτιμά και πάλι τους συντελεστές του γραμμικού μοντέλου με τη μέθοδο των ελαχίστων τετραγώνων (OLS) που αναφέρεται στην άσκηση 15 και για το ίδιο σύνολο εκπαίδευσης αλλά αυτή τη φορά μην εκτιμήσετε τον σταθερό όρο β_0 στο πολλαπλό μοντέλο γραμμικής παλινδρόμησης. Δηλαδή συγγράψτε πρόγραμμα σε R που εκτιμά τους συντελεστές του παρακάτω πολλαπλού μοντέλου γραμμικής παλινδρόμησης:

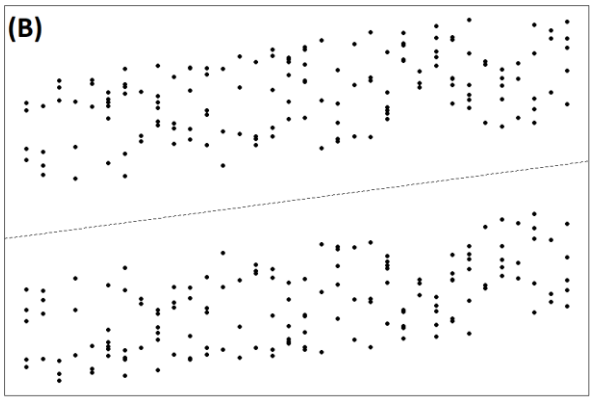
$$ViolentCrimesPerPop = \beta_1 NumStreet + \beta_2 HousVacant + \beta_3 medIncome + \beta_4 whitePerCap + \beta_5 blackPerCap + \beta_6 HispPerCap$$

χρησιμοποιώντας το ίδιο σύνολο δεδομένων της άσκησης 15. Το πρόγραμμα θα πρέπει και πάλι να εμφανίζει όλους τους συντελεστές που έχουν εκτιμηθεί.

17. Παρακάτω εμφανίζεται η απεικόνιση των τιμών δύο συνόλων δεδομένων (A) και (B) (μαύρες στιγμές) που καθένα έχει δύο μεταβλητές x1 και y και οι τιμές των οποίων δημιουργήθηκαν τυχαία με ομοιόμορφη κατανομή, μαζί με την γραμμή παλινδρόμησης (διακεκομμένη γραμμή) για κάθε ένα από τα δύο αυτά σύνολα δεδομένων, όπως προέκυψε κατά την εκτίμηση του απλού γραμμικού μοντέλου παλινδρόμησης με χρήση της μεθόδου των ελαχίστων τετραγώνων

$$y = \beta_1 x1 + \beta_0$$





Ποια από τις παρακάτω προτάσεις είναι αληθής για το *άθροισμα των καταλοίπων* της γραμμής παλινδρόμησης στις περιπτώσεις Α και Β;

Α	Το άθροισμα των καταλοίπων στην περίπτωση Α θα είναι μεγαλύτερο από την περίπτωση Β.
Β	Το άθροισμα των καταλοίπων στην περίπτωση Β θα είναι μεγαλύτερο από την περίπτωση Α.
Γ	Το άθροισμα των καταλοίπων και στις δύο περιπτώσεις Α και Β θα έχουν την ίδια τιμή δηλαδή είναι ίσα.
Δ	Καμία από τις προτάσεις Α, Β, Γ.

18. Συγγράψτε πρόγραμμα σε R, το οποίο εκτιμά τους συντελεστές του πολλαπλού γραμμικού μοντέλου παλινδρόμησης που εμφανίζεται στο υποερώτημα i) της άσκησης 15 αλλά κάνει χρήση της μεθόδου της Σταδιακής Καθόδου Δέσμης (Batch Gradient Descent) της ενότητας 6.6.3.4. Για την εκτίμηση των συντελεστών, κάνετε χρήση του ίδιου συνόλου δεδομένων που αναφέρει η άσκηση 15. Αφού έχετε εκτιμήσει τους συντελεστές, απαντήστε στις εξής ερωτήσεις:

i. Σε ποιες τιμές της παραμέτρου μάθησης α και πλήθους επαναλήψεων καταλήξατε για την εκτέλεση του αλγορίθμου της Σταδιακής Καθόδου Δέσμης; Να παραθέσετε τις ενδείξεις που τεκμηριώνουν ότι οι παράμετροι αυτές ήταν οι κατάλληλες.

ii. Εμφανίστε τους συντελεστές που έχουν εκτιμηθεί. Αναφέρετε την τιμή του συντελεστή για κάθε ανεξάρτητη μεταβλητή (και του σταθερού όρου) που έχει εκτιμηθεί.

iii. Συγκρίνετε τους συντελεστές του μοντέλου που προέκυψαν με χρήση της μεθόδου της Σταδιακής Καθόδου Δέσμης με τους συντελεστές που προέκυψαν για το ίδιο μοντέλο παλινδρόμησης (και το ίδιο σύνολο δεδομένων) με τη μέθοδο των ελαχίστων τετραγώνων (OLS) στο ερώτημα 15 Τί παρατηρήσεις μπορείτε να κάνετε;

19. Δίνεται το παρακάτω πρόγραμμα R

```
myData <- data.frame(y=numeric(0), x1=numeric(0),
                                   x2=numeric(0),
                                   x3=numeric(0),
                                   x4=numeric(0),
                                   x5=numeric(0),
                                   x6=numeric(0))


for (i in 1:4){
    myData[i,] <- runif(7, min=1, max=10)
}
rModel<-lm( y ~ ., data=myData)
print(rModel$coefficients)
```

Αφού εκτελέσετε το παραπάνω πρόγραμμα, απαντήστε στις παρακάτω ερωτήσεις:

i. Εξηγείστε τί ακριβώς κάνει το πρόγραμμα που έχετε εκτελέσει.
ii. Εξηγείστε τις τιμές των εκτιμώμενων συντελεστών που έχουν προκύψει. Ειδικότερα, σε ποιο γεγονός νομίζετε ότι οφείλονται κάποιες από τις τιμές των συντελεστών που έχουν προκύψει; Προκειμένου να μελετήσετε καλύτερα που οφείλονται οι τιμές ορισμένων συντελεστών που προκύπτουν, προσπαθήστε να επανεκτελέσετε το πρόγραμμα που δίνεται κάνοντας αλλαγές στο σύνολο δεδομένων που χρησιμοποιείται για την εκτίμηση των συντελεστών.
iii. Με αφορμή το πρόγραμμα που έχετε εκτελέσει και την ερμηνεία των αποτελεσμάτων που προέκυψαν, σε ποιο γενικό συμπέρασμα μπορείτε να καταλήξετε για τη μέθοδο των ελαχίστων τετραγώνων; Αποδείξτε μαθηματικά το συμπέρασμα αυτό στη γενική του περίπτωση.

20. Υλοποιήστε στο περιβάλλον της R τον αλγόριθμο της Σταδιακής Καθόδου Μικρών Δεσμών (Mini Batch Gradient Descent) από την αρχή (from scratch).

21. Υλοποιήστε στο περιβάλλον της R τον αλγόριθμο της Στοχαστικής Σταδιακής Καθόδου (Stochastic Gradient Descent) από την αρχή (from scratch). Υλοποιείστε τον αλγόριθμο της Στοχαστικής Σταδιακής Καθόδου βασισμένοι στον ψευδοκώδικα που υπάρχει στην ενότητα 6.6.3.6 και ο οποίος διαπερνά ολόκληρο το σύνολο δεδομένων μία μόνο φορά. Αφού έχετε υλοποιήσει τον αλγόριθμο, απαντήστε στα παρακάτω ερωτήματα:

i. Μπορείτε να κάνετε χρήση της μορφής της συνάρτησης κόστους J(θ) όπως αναφέρετε στην ενότητα 6.6.3.3 για να επιλέξετε την κατάλληλη τιμή της παραμέτρου μάθησης α στον αλγόριθμο της Στοχαστικής Σταδιακής Καθόδου; Τεκμηριώστε την απάντησή σας.
ii. Χρησιμοποιείστε τον αλγόριθμο της Στοχαστικής Σταδιακής Καθόδου που έχετε υλοποιήσει, για να εκτιμήσετε τους συντελεστές του πολλαπλού γραμμικού μοντέλου παλινδρόμησης της άσκησης 15 χρησιμοποιώντας το ίδιο σύνολο δεδομένων που αναφέρει η άσκηση 15. Το πρόγραμμά σας θα πρέπει να εμφανίζει τους συντελεστές που έχουν εκτιμηθεί. Με ποιον τρόπο επιλέξατε την τιμή της παραμέτρου μάθησης α;

22. Γιατί η παράμετρος α (alpha) στη μέθοδο της Σταδιακής Καθόδου Δέσμης δεν μπορεί να πάρει αρνητικές τιμές;

23. Στη παρουσίαση της μεθόδου Σταδιακής Καθόδου Δέσμης (και των διαφόρων εκδοχών της) έγινε η παρατήρηση και υπόθεση, ότι η παράμετρος μάθησης α (alpha) είναι μία σταθερά και καθορίζεται εξαρχής (και άπαξ) πριν την εκτέλεση του αλγορίθμου της Σταδιακής Καθόδου. Είναι δυνατόν η παράμετρος μάθησης α να αλλάζει δυναμικά κατά τη διάρκεια εκτέλεσης του αλγορίθμου της Σταδιακής Καθόδου Δέσμης; Σχολιάστε.

24. Κατεβάστε από τον ιστότοπο "UCI Machine Learning Repository", και ειδικότερα από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Real+estate+valuation+data+set το σύνολο δεδομένων που αναφέρει ορισμένα χαρακτηριστικά ακινήτων καθώς και τις τιμές των ακινήτων αυτών σε διάφορες περιοχές της πρωτεύουσας της Ταϊβάν, Ταϊπέϊ.
Δημιουργείται το παρακάτω απλό γραμμικό μοντέλο παλινδρόμησης, που στόχο έχει την εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής House price of unit area:

$$House\ price\ of\ unit\ area = \beta_1 House\ Age + \beta_0$$

Αφού εκτιμήσετε τους συντελεστές του παραπάνω απλού μοντέλου παλινδρόμησης με την κατάλληλη μέθοδο, απαντήσετε στα παρακάτω ερωτήματα:

i. Τεκμηριώνεται η γραμμική συσχέτιση μεταξύ των μεταβλητών που εμφανίζονται στο μοντέλο παλινδρόμησης; Τεκμηριώστε την απάντησή σας.
ii. Η μεταβλητή House Age είναι στατιστικά σημαντική, εάν το επιλεγμένο επίπεδο σημαντικότητας είναι 5%; Εάν ναι, τί ακριβώς σημαίνει αυτό για τη μεταβλητή House Age;
iii. Περιγράψτε τί ακριβώς εκφράζουν οι στήλες με επικεφαλίδες "Std. Error" και "t value" που εμφανίζονται στην ενότητα "Coefficients", εάν εκτελεστεί η εντολή summary() δίνοντας ως όρισμα το αποτέλεσμα της συνάρτησης lm().
iv. Πως ερμηνεύεται ο συντελεστής β_1 που έχει εκτιμηθεί για την ανεξάρτητη μεταβλητή House Age;
v. Τί ποσοστό της διακύμανσης της εξαρτημένης μεταβλητής μπορεί να ερμηνεύσει η ανεξάρτητης μεταβλητής του μοντέλου; Που νομίζετε ότι οφείλεται η μικρή τιμή του ποσοστού της διακύμανσης που μπορεί να ερμηνευτεί;

25. Συγγράψτε πρόγραμμα σε R το οποίο εκτιμά τους συντελεστές του απλού γραμμικού μοντέλου παλινδρόμησης της άσκησης 24 με το ίδιο σύνολο δεδομένων που αναφέρει η άσκηση κάνοντας χρήση της μεθόδου της Σταδιακής Καθόδου Δέσμης. Εκτελέστε τη μέθοδο της Σταδιακής Καθόδου Δέσμης με παράμετρο μάθησης α=0.001 και πλήθος επαναλήψεων 25000. Το πρόγραμμά σας θα πρέπει επίσης να περιέχει τις κατάλληλες εντολές που εξετάζουν εάν ο συντελεστής της ανεξάρτητης μεταβλητής HouseAge είναι στατιστικά σημαντικός στο επίπεδο σημαντικότητας 5% ή όχι.

26. Κατεβάστε από τον ιστότοπο "UCI Machine Learning Repository", και συγκεκριμένα από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Wine+Quality δεδομένα που αφορούν τα χαρακτηριστικά μιας ποικιλίας πορτογαλικού κρασιού. Ειδικότερα, κατεβάστε τα δεδομένα που αφορούν μόνο το λευκό κρασί της ποικιλίας αυτής (αρχείο winequality-white.csv). Συγγράψτε πρόγραμμα στο περιβάλλον της R, το οποίο εκτιμά, με τη μέθοδο των ελαχίστων τετραγώνων (OLS) τους συντελεστές του παρακάτω πολλαπλού γραμμικού μοντέλου παλινδρόμησης που στόχος του είναι η εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής:

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

Επιπλέον, απαντήστε στα εξής ερωτήματα:

i. Τεκμηριώνεται η γραμμική συσχέτιση μεταξύ των μεταβλητών που εμφανίζονται στο μοντέλο παλινδρόμησης; Να παρατεθούν όλες οι ενδείξεις που τεκμηριώνουν την απάντησή σας.
ii. Τί ποσοστό της διακύμανσης μπορεί να εξηγήσει το παραπάνω γραμμικό μοντέλο παλινδρόμησης;
iii. Ποιες από τις ανεξάρτητες μεταβλητές του μοντέλου είναι στατιστικά σημαντικές εάν επιλεγεί ως επίπεδο σημαντικότητας 1%;
iv. Εξηγείστε πως ερμηνεύονται οι συντελεστές β_2 και β_3 που προέκυψαν για τις ανεξάρτητες μεταβλητές pH και density.

27. Για κάθε μία από τις παρακάτω περιπτώσεις, αναφέρετε τα συμπεράσματα που μπορείτε να βγάλετε για ένα απλό μοντέλο γραμμικής παλινδρόμησης δύο μεταβλητών (Y εξαρτημένης και X ανεξάρτητης), αν το επίπεδο στατιστικής σημαντικότητας τεθεί στο 5%:

i. Χαμηλή τιμή συντελεστή προσδιορισμού R^2 (π.χ. $R^2 < 0.03$) και p-τιμής < 0.05
ii. Χαμηλή τιμή συντελεστή προσδιορισμού R^2 (π.χ. $R^2 < 0.03$) και p-τιμής > 0.05
iii. Υψηλή τιμή συντελεστή προσδιορισμού R^2 (π.χ. $R^2 > 0.6$) και p-τιμή < 0.05
iv. Υψηλή τιμή συντελεστή προσδιορισμού R^2 (π.χ. $R^2 > 0.6$) και p-τιμή > 0.05

28. Από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset κατεβάστε το σύνολο δεδομένων που αφορά την ενοικίαση ποδηλάτων στην πόλη Πόρτο της Πορτογαλίας. Χρησιμοποιώντας το σύνολο δεδομένων που καταγράφει το πλήθος των ενοικιάσεων ποδηλάτων ανά ημέρα μαζί με τις μετεωρολογικές (καιρικές) συνθήκες κάθε ημέρας (αρχείο day.csv) και αφού μελετήσετε τις μεταβλητές που αυτό περιέχει, απαντήστε στα παρακάτω ερωτήματα:

i. Συγγράψτε πρόγραμμα σε R που εκτιμά, με τη μέθοδο των ελαχίστων τετραγώνων (OLS) τους συντελεστές του ακόλουθου πολλαπλού γραμμικού μοντέλου παλινδρόμηση προκειμένου να μελετηθεί εάν και με ποιον τρόπο οι καιρικές συνθήκες επηρεάζουν την ενοικίαση ποδηλάτων:

$$cnt = \beta_1 weathersit + \beta_2 temp + \beta_3 hum + \beta_4 windspeed + \beta_0$$

Αφού έχετε εκτιμήσει του συντελεστές απαντήστε στις ακόλουθες ερωτήσεις:

a. Εμφανίστε τους συντελεστές που έχουν εκτιμηθεί για κάθε μεταβλητή του μοντέλου. Τί έχετε να παρατηρήσετε σχετικά με τις μεταβλητές, των οποίων εκτιμήθηκαν οι συντελεστές;
b. Ποιες μεταβλητές είναι στατιστικά σημαντικές, εάν το επίπεδο σημαντικότητας είναι 5%;
c. Πως πρέπει να ερμηνευτούν οι συντελεστές που εκτιμήθηκαν για την ανεξάρτητη μεταβλητή weathersit;
d. Αν σας ρωτούσαν, ποια καιρική συνθήκη επηρεάζει περισσότερο απ'όλες και πως ακριβώς τις ενοικιάσεις ποδηλάτων, τί θα απαντούσατε;

ii. Συγγράψτε πρόγραμμα σε R που εκτιμά και πάλι τους συντελεστές του πολλαπλού γραμμικού μοντέλου παλινδρόμησης του υποερωτήματος i) αλλά αυτή τη φορά με τη μέθοδο της Σταδιακής Καθόδου Δέσμης (κάνετε χρήση της υλοποίησης που υπάρχει στην ενότητα 6.6.3.4). Επιλέξτε τις κατάλληλες τιμές για την παράμετρο μάθησης α και το πλήθος επαναλήψεων. Επίσης απαντήστε στα ακόλουθα ερωτήματα

   a. Εμφανίστε τους συντελεστές που έχουν εκτιμηθεί.
   b. Δείξτε ότι η παράμετρος μάθησης α που έχετε επιλέξει, είναι η κατάλληλη.
   c. Εξηγείστε τον τρόπο με τον οποίο έχετε χειριστεί την κατηγορική μεταβλητή που εμφανίζεται στο μοντέλου, ώστε να είναι δυνατή η εκτίμηση των συντελεστών με τη μέθοδο της Σταδιακής Καθόδου Δέσμης.

29. 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 . Απαντήστε στα παρακάτω ερωτήματα:

   a. Ποιες εξαρτημένες μεταβλητές αναφέρονται στο άρθρο και ποιες ανεξάρτητες;
   b. Ποια είναι η μορφή του/των μοντέλου (-ων) παλινδρόμησης που εκτιμήθηκε (-αν) στον άρθρο;
   c. Τί μέθοδο χρησιμοποιήθηκε για να εκτιμηθούν οι συντελεστές των μοντέλων παλινδρόμησης;
   d. Ο στόχος της μελέτης είναι η εξήγηση ή πρόβλεψη της εξαρτημένης μεταβλητής; Τεκμηριώστε την άποψή σας αναλύοντας τη μεθοδολογία που χρησιμοποιεί το άρθρο.

30. Κατεβάστε και μελετήστε το άρθρο "Redmond, M and Baveja, A.: A data-driven software tool for enabling cooperative information sharing among police departments, European Journal of Operational Research, Volume 141, Issue 3, September 2002, pp660-678" που μπορεί να βρεθεί από εδώ: https://www.sciencedirect.com/science/article/pii/S0377221701002648. Απαντήστε στα εξής ερωτήματα:

   a. Ποιες εξαρτημένες μεταβλητές κάνει χρήση το άρθρο και ποιες ανεξάρτητες;
   b. Ποια είναι η μορφή του/των μοντέλου (-ων) παλινδρόμησης που εκτιμήθηκε (-αν) στον άρθρο;
   c. Τί μέθοδο χρησιμοποιήθηκε για να εκτιμηθούν οι συντελεστές των μοντέλων παλινδρόμησης;
   d. Ο στόχος της μελέτης είναι η εξήγηση ή πρόβλεψη της εξαρτημένης μεταβλητής; Τεκμηριώστε την άποψή σας αναλύοντας τη μεθοδολογία που χρησιμοποιεί το άρθρο.

31. Έστω ότι η ανεξάρτητη μεταβλητή X αποδεικνύεται στατιστικά σημαντική στο παρακάτω πολλαπλό γραμμικό μοντέλο παλινδρόμησης

$$Y = \beta_1 X + \beta_2 Z + \beta_3 L + \beta_4 M + \beta_0$$

ενώ η ίδια μεταβλητή X δεν προκύπτει να είναι στατιστικά σημαντική για το απλό πολλαπλό γραμμικό μοντέλο για την ίδια εξαρτημένη μεταβλητή Y και το ίδιο σύνολο δεδομένων

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

Τί συμπέρασμα μπορείτε να βγάλετε για τις μεταβλητές X, Z, L και M;

32. Τί πρόβλημα δημιουργεί η εμφάνιση πολυσυγγραμμικότητας σε ένα γραμμικό μοντέλο παλινδρόμησης εάν ο στόχος του μοντέλου είναι η εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής;

33. Συγγράψτε πρόγραμμα στο περιβάλλον της R το οποίο δημιουργεί ένα πλασματικό/συνθετικό σύνολο δεδομένων με τέτοιο τρόπο, ώστε να εμφανίζεται το φαινόμενο της πολυσυγγραμμικότητας μεταξύ τουλάχιστον 2 μεταβλητών σε ένα γραμμικό μοντέλο παλινδρόμησης. Το πρόγραμμα θα πρέπει να έχει τις κατάλληλες εντολές που αποδεικνύουν την πολυσυγγραμμικότητα μεταξύ των μεταβλητών αυτών.

34. Έστω ότι καμία ανεξάρτητη μεταβλητή δεν αποδεικνύεται στατιστικά σημαντική στο παρακάτω πολλαπλό γραμμικό μοντέλο παλινδρόμησης

$$Y = \beta_1 X + \beta_2 Z + \beta_3 L + \beta_4 M + \beta_0$$

ενώ η μεταβλητή X προκύπτει να είναι στατιστικά σημαντική για το απλό πολλαπλό γραμμικό μοντέλο για την ίδια εξαρτημένη μεταβλητή Y και το ίδιο σύνολο δεδομένων

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

Τί συμπέρασμα μπορείτε να βγάλετε για τις μεταβλητές X, Z, L και M;

35. Εξηγείστε γιατί όταν ο στόχος ενός μοντέλου παλινδρόμησης είναι η πρόβλεψη της τιμής της εξαρτημένης μεταβλητής, η πολυσυγγραμμικότητα δεν αποτελεί σημαντικό πρόβλημα[2] που χρήζει επείγουσας αντιμετώπισης ενώ αποτελεί σημαντικό πρόβλημα όταν ο στόχος του μοντέλου είναι η εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής. Αναζητείστε δημοσιεύσεις που εξηγούν και τεκμηριώνουν την θέση αυτή.

36. Ποιες από τις παρακάτω προτάσεις είναι αληθείς εάν προστεθεί μία επιπλέον ανεξάρτητη μεταβλητή σε ένα υπάρχον μοντέλο γραμμικής παλινδρόμησης;

1. Ο συντελεστής προσδιορισμού R^2 και ο διορθωμένος συντελεστής προσδιορισμού R^2 (adjusted R^2) αυξάνουν και οι δύο.

2. Ο συντελεστής προσδιορισμού R^2 αυξάνεται ενώ ταυτόχρονα ο διορθωμένος συντελεστής προσδιορισμού R^2 (adjusted R^2) μειώνεται.

3. Ο συντελεστής προσδιορισμού R^2 μειώνεται ενώ ταυτόχρονα και ο διορθωμένος συντελεστής προσδιορισμού R^2 (adjusted R^2) μειώνεται.

[2] Η πολυσυγγραμμικότητα αποτελεί πρόβλημα και σε τέτοιες περιπτώσεις, αλλά όχι σημαντικό.

4. Ο συντελεστής προσδιορισμού R^2 μειώνεται ενώ ταυτόχρονα ο διορθωμένος συντελεστής προσδιορισμού R^2 (adjusted R^2) αυξάνεται.

37. Εάν σε ένα πολλαπλό μοντέλο γραμμικής παλινδρόμησης με στόχο την εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής, ορισμένες μεταβλητές δεν είναι στατιστικά σημαντικές στο επιλεγμένο επίπεδο σημαντικότητας, μπορούν αυτές οι μεταβλητές να αφαιρεθούν από το μοντέλο παλινδρόμησης; Τεκμηριώστε την απάντησή σας.

38. 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 για τη μεταβλητή αυτή;

39. Για τον έλεγχο της στατιστικής σημαντικότητας μιας ανεξάρτητης μεταβλητής σε ένα γραμμικό μοντέλο παλινδρόμησης, γίνεται η υπόθεση ότι οι τιμές των συντελεστών ακολουθούν την κατανομή Student (Student's t-distribution). Για ποιον λόγο γίνεται χρήση της κατανομής Student;

40. Στη βιβλιογραφία χρησιμοποιείται το επίπεδο σημαντικότητας 5% ως το de facto όριο για την απόρριψη ή μη της μηδενικής υπόθεσης H_0 σχετικά με την αξιολόγηση της στατιστικής σημαντικότητας των ανεξάρτητων μεταβλητών ενός μοντέλου γραμμικής παλινδρόμησης. Απαντήστε και αναπτύξτε τις παρακάτω ερωτήσεις:

i. Γιατί επιλέχτηκε η τιμή 5%; Αναζητείστε στο διαδίκτυο για το πως καθιερώθηκε η τιμή του 5%.
ii. Σε ορισμένες επιστημονικές περιοχές, το όριο του 5% δεν κρίνεται ως επαρκές και τίθεται ως de facto όριο το 1% για την απόρριψη ή μη της μηδενικής υπόθεσης. Τί σημαίνει μια τέτοια μείωση του επιπέδου σημαντικότητας και σε ποιες περιοχές συναντιέται αυτό το όριο; Αναζητείστε στο διαδίκτυο πηγές που απαντούν στα παραπάνω ερωτήματα.

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

$$\begin{aligned}\mathit{Κατανάλωση\ τροφίμων} &= 0.1405087\ \mathit{Εισόδημα} + 819.8224270\ \mathit{Αριθμός\ ατόμων\ νοικοκυριού} \\ &+ 760.5908490\end{aligned}$$

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

42. Γιατί είναι καλή ιδέα/χρήσιμο, εάν γίνεται χρήση της μεθόδου της Σταδιακής Καθόδου Δέσμης (Batch Gradient Descent) για την εκτίμηση συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης, να προηγείται η κανονικοποίηση των τιμών των ανεξαρτήτων μεταβλητών κατά την προεπεξεργασία του συνόλου δεδομένων εκπαίδευσης;

43. 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 του Παραρτήματος Α

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

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;

44. 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 Δέσμης δίνεται από τον κλειστό τύπο:

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

45. Κατά τη διαδικασία κανονικοποίησης των τιμών των εξαρτημένων μεταβλητών στο σύνολο εκπαίδευσης πριν την χρήση της μεθόδου της Σταδιακής Καθόδου Δέσμης για την εκτίμηση των συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης, γιατί συνήθως δεν απαιτείται να κανονικοποιηθούν και οι τιμές της εξαρτημένης μεταβλητής του μοντέλου στο σύνολο εκπαίδευσης; Μπορείτε να σκεφτείτε μία περίπτωση στην οποία χρειάζονται κανονικοποίηση και οι τιμές της εξαρτημένης μεταβλητής κατά την εκτίμηση συντελεστών με τη μέθοδο της Σταδιακής Καθόδου;

46. 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):

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

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

47. Κατεβάστε από την ιστοσελίδα https://archive.ics.uci.edu/ml/datasets/Forest+Fires σύνολο δεδομένων για πυρκαγιές από περιοχές της Πορτογαλίας. Τα δεδομένα περιέχουν γεωγραφικά και μετεωρολογικά στοιχεία όταν εκδηλώθηκαν πυρκαγιές καθώς επίσης και την επιφάνεια που κάηκε που μετριέται σε εκτάρια[3] (hectars). Έχοντας ως στόχο την πρόβλεψη της επιφάνειας που θα καεί βασει των μετεωρολογικών συνθηκών που επικρατούν συγγράψτε κώδικα R που αξιολογεί το παρακάτω μοντέλου παλινδρόμησης με τους τρόπους που ζητούνται παρακάτω:

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

i. Χρησιμοποιώντας όλες τις παρατηρήσεις του συνόλου δεδομένων που έχετε κατεβάσει, κάντε χρήση διασταυρωτικής επικύρωση 10-πτυχών (10-Fold Cross Validation) για την αξιολόγηση του

[3] 1 εκτάριο = 10 στρέμματα

παραπάνω μοντέλου γραμμικής παλινδρόμησης. Η εκτίμηση των συντελεστών του μοντέλου θα πρέπει να γίνεται με τη μέθοδο των Ελαχίστων Τετραγώνων (OLS) και θα πρέπει να υπολογίζεται το μέσο τετραγωνικό σφάλμα (Root Mean Squared Error – RMSE) της πρόβλεψης. Κάντε χρήση της υλοποίησης της μεθόδου της διασταυρωτικής επικύρωσης k-πτυχών που υπάρχει στην ενότητα 6.7.2.1 Το πρόγραμμά σας θα πρέπει να εμφανίζει τον μέσο όρο των μέσων τετραγωνικών σφαλμάτων (RMSE) που έχουν προκύψει κατά την εκτέλεση της διασταυρωτικής επικύρωση 10-πτυχών ως μέτρο αξιολόγησης της πρόβλεψης του μοντέλου. Τί ακριβώς εκφράζει η τιμή του σφάλματος που προέκυψε;

ii. Αξιολογείστε και πάλι το ίδιο μοντέλο παλινδρόμησης, αλλά αυτή τη φορά μην χρησιμοποιείστε ολόκληρο το σύνολο δεδομένων αλλά μόνο εκείνες τις παρατηρήσεις όπου η τιμή της εξαρτημένης μεταβλητής (μεταβλητή area) είναι μικρότερη από 3.2 εκτάρια (δηλαδή area < 3.2) και χαρακτηρίζει μικρές πυρκαγιές. Η εκτίμηση των συντελεστών του μοντέλου θα πρέπει να γίνεται με τη μέθοδο των Ελαχίστων Τετραγώνων (OLS) και θα πρέπει να υπολογίζεται το μέσο τετραγωνικό σφάλμα (Root Mean Squared Error – RMSE) της πρόβλεψης. Κάντε χρήση της υλοποίησης της μεθόδου της διασταυρωτικής επικύρωσης k-πτυχών που υπάρχει στην ενότητα 6.7.2.1 Το πρόγραμμά σας θα πρέπει να εμφανίζει τον μέσο όρο των μέσων τετραγωνικών σφαλμάτων (RMSE) που έχουν προκύψει κατά την εκτέλεση της διασταυρωτικής επικύρωση 10-πτυχών ως μέτρο αξιολόγησης της πρόβλεψης του μοντέλου. Τί ακριβώς εκφράζει η τιμή του σφάλματος που προέκυψε;

iii. Τί συμπέρασμα μπορείτε να βγάλετε σχετικά με την πρόβλεψη, αν συγκρίνετε τους μέσους όρους των μέσων τετραγωνικών σφαλμάτων (RMSE) που προέκυψαν στα υποερωτήματα i) και ii) παραπάνω;

48. Αξιολογείστε και πάλι την πρόβλεψη του μοντέλου παλινδρόμησης της άσκησης 47 με τις συνθήκες που περιγράφονται στο υποερώτημα i) αλλά αυτή τη φορά κάνετε χρήση των κατάλληλων συναρτήσεων για διασταυρωτική επικύρωση k-πτυχών που παρέχει η βιβλιοθήκη caret της R και όχι την υλοποίηση που υπάρχει στην ενότητα 6.7.2.1. Να εμφανιστεί και πάλι ο μέσος όρος των μέσων τετραγωνικών σφαλμάτων (RMSE) που έχουν προκύψει κατά την εκτέλεση της διασταυρωτικής επικύρωση 10-πτυχών.

49. Εκτιμήστε και πάλι τους συντελεστές του γραμμικού μοντέλου παλινδρόμησης της άσκησης 47 χρησιμοποιώντας ολόκληρο το σύνολο δεδομένων αλλά αυτή τη φορά κάνετε χρήση του μέσου απόλυτου σφάλματος (MAE – Mean Absolute Error) για την εκτίμηση του σφάλματος πρόβλεψης. Χρησιμοποιείστε κι εδώ τη μέθοδο ελαχίστων τετραγώνων καθώς και διασταυρωτική επικύρωση 10-πτυχών για την εκτίμηση του σφάλματος. Κάνετε χρήση της υλοποίησης που υπάρχει στην ενότητα 6.7.2.1 κάνοντας τις κατάλληλες αλλαγές. Συγκρίνετε το σφάλμα που προέκυψε με το σφάλμα που προέκυψε στο υποερώτημα i) της άσκησης 47. Τί παρατηρήσεις/σχόλια μπορείτε να κάνετε;

50. Ένα γραμμικό μοντέλο παλινδρόμησης που είναι πολύ καλό στην εξήγηση της διακύμανσης της εξαρτημένης μεταβλητής, θα είναι επίσης και καλό στην πρόβλεψη της τιμής της εξαρτημένης μεταβλητής; Σχολιάστε.

51. Σε ποιες περιπτώσεις ***δεν*** μπορεί να χρησιμοποιηθεί το μέσο απόλυτο εκατοστιαίο Σφάλμα (Mean Absolute Percentage Error – MAPE) για την εκτίμηση του σφάλματος πρόβλεψης ενός γραμμικού μοντέλου παλινδρόμησης;

52. Τί εννοείται όταν αναφέρεται ο όρος «Στάθμιση Μεροληψίας-Διακύμανσης» (Bias-variance tradeoff); Αναζητείστε στο διαδίκτυο για την εύρεση και μελέτη του όρου αυτού. Ποια η σχέση του όρουμε την κανονικοποίηση (regularization) μοντέλου γραμμικής παλινδρόμησης;

53. Ποιες από τις παρακάτω προτάσεις είναι αληθείς για την παλινδρόμηση Ridge;

1. Εάν η παράμετρος λ λάβει τιμή 0, τότε το μοντέλο καταλήγει να γίνεται ένα γραμμικό μοντέλο παλινδρόμησης
2. Εάν η παράμετρος λ λάβει τιμή 0, τότε το μοντέλο δεν καταλήγει να γίνεται ένα γραμμικό μοντέλο παλινδρόμησης
3. Εάν η παράμετρος λ τείνει στο άπειρο, τότε θα προκύψουν πάρα πολύ μικρές τιμές συντελεστών που τείνουν στο 0 δίχως ωστόσο να λάβουν την τιμή 0.
4. Εάν η παράμετρος λ τείνει στο άπειρο, τότε θα προκύψουν πάρα πολύ μεγάλες τιμές συντελεστών που τείνουν στο άπειρο.

54. Υποθέστε ότι ένα γραμμικό μοντέλο παλινδρόμησης παρουσιάζει υποπροσαρμογή (underfitting). Με ποιον από τους παρακάτω τρόπους θα αντιμετωπίσετε μια τέτοια κατάσταση;

1. Πρόσθεση περισσότερων ανεξάρτητων μεταβλητών στο μοντέλο παλινδρόμησης.
2. Προσθήκη ανεξάρτητων μεταβλητών πολυωνυμικού βαθμού στο μοντέλο παλινδρόμησης.
3. Αφαίρεση κάποιων από τις ανεξάρτητες μεταβλητές του μοντέλου παλινδρόμησης.

55. Υποθέστε ότι ένα γραμμικό μοντέλο παλινδρόμησης παρουσιάζει υποπροσαρμογή (underfitting). Ποιον τρόπο κανονικοποίσης (regularization) του μοντέλου θα προτιμήσετε;

1. Παλινδρόμηση LASSO
2. Παλινδρόμηση Ridge
3. Οποιονδήποτε από τους αλγορίθμους παλινδρόμησης LASSO ή Rindge
4. Κανένα από τα παραπάνω.

56. Σε ποιες περιπτώσεις πρέπει να γίνεται κανονικοποίηση του μοντέλου παλινδρόμησης (regularization); Με ποιον τρόπο θα ελέγχατε εάν ένα μοντέλο γραμμικής παλινδρόμησης πρέπει να υποστεί κανονικοποίηση;

57. Πως επηρεάζει το πλήθος των παρατηρήσεων σε ένα σύνολο δεδομένων την τάση ενός γραμμικού μοντέλου παλινδρόμησης προς υπερπροσαρμογή (overfitting); Θεωρείστε ότι όλα τα υπόλοιπα στοιχεία του μοντέλου παραμένουν τα ίδια. Επιλέξτε ποιες από τις παρακάτω προτάσεις είναι αληθείς:

1. Εάν το σύνολο δεδομένων είναι μικρό, είναι εύκολο να παρουσιαστεί υπερπροσαρμογή.

2. Εάν το σύνολο δεδομένων είναι μικρό, είναι δύσκολο/απίθανο να παρουσιαστεί υπερπροσαρμογή.
3. Εάν το σύνολο δεδομένων είναι μεγάλο, είναι εύκολο να παρουσιαστεί υπερπροσαρμογή.
4. Εάν το σύνολο δεδομένων είναι μεγάλο, είναι δύσκολο/απίθανο να παρουσιαστεί υπερπροσαρμογή.

58. Συγγράψτε πρόγραμμα σε R το οποίο δημιουργεί συνθετικά/πλασματικά/τυχαία δεδομένα με τέτοιο τρόπο, ώστε να παρουσιάζεται το φαινόμενο της υπερπροσαρμογής (overfitting) εάν τα δεδομένα αυτά χρησιμοποιηθούν για την εκτίμηση συντελεστών ενός γραμμικού μοντέλου παλινδρόμησης. Το πρόγραμμά σας θα πρέπει να εμφανίζει όλες τις ενδείξεις που υποστηρίζουν ότι το μοντέλο έχει υποστεί υπερπροσαρμογή.

59. Αναφέρετε τις διαφορές της παλινδρόμησης Ridge και Lasso.

60. Αναζητήστε στο διαδίκτυο πληροφορίες για το Akaike Information Criterion (AIC) και δώστε τον ορισμό του. Ακολούθως, απαντήστε στις εξής ερωτήσεις:

i. Ποιος είναι ο στόχος του AIC;

ii. Σε ποια διαδικασία θα εντάσσατε τη χρήση του AIC: a) στη διαδικασία εκτίμησης συντελεστών, b) στη διαδικασία αξιολόγησης μοντέλων παλινδρόμησης ή c) στη διαδικασία αντιμετώπισης προβλημάτων υπερπροσαρμογής ενός μοντέλου;

61. Βρείτε τον ορισμό, αναζητώντας στο διαδίκτυο, της βηματικής γραμμικής παλινδρόμησης (stepwise linear regression). Σε ποιες περιπτώσεις χρησιμοποιείται και για την αντιμετώπιση ποιων ζητημάτων;

62. Βρείτε τον ορισμό, αναζητώντας στο διαδίκτυο, της τεταρτημοριακής γραμμικής παλινδρόμησης ή αλλιώς γραμμικής παλινδρόμησης κατά εκατοστημόριο (quantile linear regression). Σε ποιες περιπτώσεις χρησιμοποιείται και για την αντιμετώπιση ποιων ζητημάτων;

63. Βρείτε τον ορισμό, αναζητώντας στο διαδίκτυο, του πολυμεταβλητού πολλαπλού γραμμικού μοντέλου γραμμικής παλινδρόμησης (multivariate linear regression model). Που είναι χρήσιμη τέτοιας μορφής γραμμικής παλινδρόμηση;