Εργασία στο μάθημα 'Ανάλυση Δεδομένων', Δεκέμβριος 2022

Δημήτρης Κουγιουμτζής

E-mail: dkugiu@auth.gr

24 Δεκεμβρίου 2022

Οδηγίες: Σχετικά με την παράδοση της εργασίας θα πρέπει:

- Για κάθε ζήτημα θα δημιουργήσετε ένα ή περισσότερα προγράμματα και συναρτήσεις Matlab. Τα ονόματα τους θα είναι ως εξής, όπου ως παράδειγμα δίνεται η ομάδα φοιτητών No 10 και το ζήτημα 5. Για τα προγράμματα τα ονόματα των αρχείων θα είναι Group10Exe5Prog1.m, Group10Exe5Prog2.m κτλ (αν χρειάζεται να δημιουργήσετε περισσότερα από ένα προγράμματα). Για τις συναρτήσεις τα ονόματα των αρχείων θα είναι Group10Exe5Fun1.m, Group10Exe5Fun2.m κτλ (αν χρειάζεται να δημιουργήσετε περισσότερες από μια συναρτήσεις). Στην αρχή κάθε προγράμματος και συνάρτησης θα υπάρχουν (σε σχολιασμό) τα ονοματεπώνυμα των μελών της ομάδας.

- Τα προγράμματα θα πρέπει να είναι εκτελέσιμα και η εκτέλεση τους να δίνει τις απαντήσεις που ζητούνται σε κάθε ζήτημα. Επεξηγήσεις, σχολιασμοί αποτελεσμάτων και συμπεράσματα, όπου ζητούνται, θα δίνονται με μορφή σχολίων στο πρόγραμμα (τα συμπεράσματα στο τέλος του προγράμματος). Τα σχόλια θα πρέπει να είναι γραμμένα στην Αγγλική γλώσσα ή στην Ελληνική με λατινικούς χαρακτήρες (Greeklish) για να αποφευχθεί τυχόν πρόβλημα στην ανάγνωση τους.

- Θα υποβληθούν μόνο τα αρχεία Matlab και τυχόν αρχεία δεδομένων άλλα από αυτά που σας δίνονται και που έχετε χρησιμοποιήσει (μέσω του elearning).

- Η κάθε εργασία (σύνολο προγραμμάτων και συναρτήσεων Matlab) θα πρέπει να συντάσσεται αυτόνομα από την ομάδα. Ομοιότητες εργασιών θα οδηγούν σε μοίρασμα της βαθμολογίας (δύο 'όμοιες' άριστες εργασίες θα μοιράζονται το βαθμό δια δύο, τρεις δια τρία κτλ.).

Περιγραφή εργασίας

Στο αρχείο `Heathrow.xlsx` που υπάρχει στην ιστοσελίδα του μαθήματος, δίνονται ετήσια δεδομένα μετεωρολογικών δεικτών για το αεροδρόμιο Heathrow ανά έτος την περίοδο 1949 – 1958 και 1973 – 2017. Οι δείκτες περιγράφονται στον παρακάτω πίνακα. Κάποιοι δείκτες παίρνουν ένα μικρό αριθμό διακεκριμένων τιμών, όπως οι μέρες του έτους με χαλάζι.

Α/Α	*Ακρωνύμιο*	*Περιγραφή*
1	T	Μέση ετήσια θερμοκρασία
2	TM	Μέση ετήσια μέγιστη θερμοκρασία
3	Tm	Μέση ετήσια ελάχιστη θερμοκρασία
4	PP	Συνολική ετήσια χιονόπτωση ή βροχόπτωση
5	V	Μέση ετήσια ταχύτητα αέρα
6	RA	Μέρες του έτους με βροχή
7	SN	Μέρες του έτους με χιόνι
8	TS	Μέρες του έτους με άνεμο
9	FG	Μέρες του έτους με ομίχλη
10	TN	Μέρες του έτους με ανεμοστρόβιλο
11	GR	Μέρες του έτους με χαλάζι

Ζητήματα εργασίας

1. Φτιάξε μια συνάρτηση που να δέχεται ως μεταβλητή εισόδου ένα διάνυσμα (παρατηρήσεις δείγματος), να μετράει τις διακεκριμένες τιμές στο διάνυσμα και να κάνει τα παρακάτω:

   (α') Αν το πλήθος των διακεκριμένων τιμών στο διάνυσμα είναι μεγαλύτερο από 10 να σχηματίζει το ιστόγραμμα για κατάλληλη ισομερή διαμέριση. Επίσης να κάνει έλεγχο X^2 καλής προσαρμογής σε κανονική κατανομή, καθώς και σε ομοιόμορφη κατανομή, και να δηλώνει τις δύο p-τιμές ελέγχου στο σχήμα (στον τίτλο ή σε κείμενο μέσα στο σχήμα) και να τις δίνει και στην έξοδο.

   (β') Αν το πλήθος των διακεκριμένων τιμών στο διάνυσμα είναι μικρότερο ή ίσο με 10 να σχηματίζει το ραβδόγραμμα. Επίσης να κάνει έλεγχο X^2 καλής προσαρμογής σε διωνυμική κατανομή, καθώς και σε διακριτή ομοιόμορφη κατανομή, και να δηλώνει τις δύο p-τιμές ελέγχου στο σχήμα (στον τίτλο ή σε κείμενο μέσα στο σχήμα) και να τις δίνει και στην έξοδο.

   Φτιάξε ένα πρόγραμμα που για κάθε έναν από τους δείκτες να καλεί αυτή τη συνάρτηση, θα συλλέγει τις δύο p τιμές για κάθε δείκτη σε έναν πίνακα, ώστε στο τέλος το πρόγραμμα θα δίνει έναν πίνακα που θα δηλώνεται αν ο δείκτης αντιμετωπίζεται ως συνεχής ή διακριτή μεταβλητή, στην περίπτωση που είναι συνεχής αν ακολουθεί κανονική ή ομοιόμορφη κατανομή και στην περίπτωση που είναι διακριτή αν ακολουθεί διωνυμική ή διακριτή ομοιόμορφη κατανομή.

2. Φτιάξε μια συνάρτηση που να δέχεται ως μεταβλητή εισόδου ένα διάνυσμα (παρατηρήσεις δείγματος) και να υπολογίζει 95% παραμετρικό διάστημα εμπιστοσύνης (δ.ε.) καθώς και 95% bootstrap δ.ε. για τη μέση τιμή. Στην έξοδο η συνάρτηση θα δίνει τα δύο δ.ε..

   Φτιάξε ένα πρόγραμμα που για κάθε έναν από τους 9 πρώτους δείκτες θα κάνει τα εξής. Θα καλεί την συνάρτηση δίνοντας τα δείγμα των παρατηρήσεων του δείκτη για έτη από το 1973 και μετά. Θα υπολογίζει τη δειγματική μέση τιμή του δείκτη στην περίοδο 1949 - 1958 και θα εξετάζει αν η τιμή αυτή είναι στα δύο διαστήματα εμπιστοσύνης (παραμετρικό και bootstrap). Σχολίασε αν τα δύο διαστήματα εμπιστοσύνης (παραμετρικό και bootstrap) διαφέρουν σημαντικά για κάποιον(ους) δείκτη(ες) και για ποιους δείκτες (αν

υπάρχουν) η μέση τιμή της πρώτης περιόδου 1949 - 1958 έχει αλλάξει στην περίοδο 1973 - 2017.

3. Φτιάξε μια συνάρτηση που να δέχεται ως μεταβλητή εισόδου δύο διανύσματα, το πρώτο να έχει τα έτη (πρώτη στήλη) και το δεύτερο τις αντίστοιχες παρατηρήσεις κάποιου δείκτη. Η συνάρτηση θα κάνει τα παρακάτω:

   (α΄) Θα εντοπίζει το πρώτο σημείο ασυνέχειας στο πρώτο διάνυσμα, δηλαδή το στοιχείο του διανύσματος που διακόπτεται η αύξουσα σειρά των τιμών κατά ένα. Αν δεν βρίσκει τέτοιο σημείο θα σταματάει εκεί και θα το δηλώνει ως σφάλμα, αλλιώς θα συνεχίζει.

   (β΄) Θα χωρίζει τα στοιχεία του δεύτερου διανύσματος στα δύο με βάση το στοχείο ασυνέχειας (στα δεδομένα μας θα είναι το 1958), δηλαδή παρατηρήσεις ως και το 1958 θα ανήκουν στο ένα διάνυσμα (τ.μ. X_1) και τα υπόλοιπα στο άλλο διάνυσμα (τ.μ. X_2).

   (γ΄) Θα κάνει έλεγχο για τη διαφορά μέσων τιμών της τ.μ. στις δύο περιόδους. Αυτό θα γίνει με παραμετρικό έλεγχο (student) υποθέτοντας κανονική κατανομή και έλεγχο επαναδειγματοληψίας (ελεύθερη επιλογή από τους τρεις ελέγχους αυτής της κατηγορίας που έχουμε μελετήσει).

   (δ΄) Θα δίνει στην έξοδο τις δύο p τιμές των ελέγχων για διαφορά μέσων τιμών.

   Κάλεσε την συνάρτηση αυτή για τους 9 πρώτους δείκτες και εντόπισε τους δείκτες για τους οποίους φαίνεται να υπάρχει διαφορά στις δύο περιόδους. Εντόπισε το δείκτη που φαίνεται να έχει τη μεγαλύτερη διαφορά στις δύο περιόδους (με βάση τον κάθε τύπο ελέγχου). Συμφωνούν τα αποτελέσματα με τους δύο τύπους ελέγχου (παραμετρικό και επαναδειγματοληψίας);

4. Φτιάξε μια συνάρτηση που να δέχεται ως δύο μεταβλητές εισόδου δύο διανύσματα ίσου μήκους (ζευγαρωτά δείγματα δύο τυχαίων μεταβλητών, τ.μ.) και να κάνει τα παρακάτω:

   (α΄) Να βρίσκει τα κενά στοιχεία στα διανύσματα (NaN) και να αφαιρεί τα αντίστοιχα ζεύγη τιμών έτσι ώστε τα δύο διανύσματα (ενδεχομένως μικρότερου μήκους) να μην έχουν κενά στοιχεία.

   (β΄) Να υπολογίζει το 95% δ.ε. για το συντελεστή συσχέτισης των δύο τ.μ. χρησιμοποιώντας το μετασχηματισμό Fisher καθώς και τη μέθοδο bootstrap.

   (γ΄) Να κάνει παραμετρικό έλεγχο της υπόθεσης για μηδενική συσχέτιση των δύο τ.μ., χρησιμοποιώντας το στατιστικό της κατανομής Student, καθώς και μη-παραμετρικό έλεγχο χρησιμοποιώντας τη μέθοδο της τυχαιοποίησης.

   (δ΄) Να δίνει στην έξοδο τα δύο δ.ε. για το συντελεστή συσχέτισης, τις δύο p τιμές των δύο ελέγχων (παραμετρικό και τυχαιοποίησης) καθώς και το μήκος των διανυσμάτων χωρίς κενά.

   Κάλεσε την συνάρτηση αυτή για όλα τα ζεύγη των 9 πρώτων δεικτών και δημιούργησε μια λίστα με τα ζεύγη δεικτών που βρέθηκαν να έχουν γραμμική συσχέτιση. Αυτό θα το κάνεις με βάση το κάθε ένα από τα δύο δ.ε. (παραμετρικό και bootstrap) καθώς κάθε

έναν από τους δύο ελέγχους (παραμετρικό και τυχαιοποίησης σε επίπεδο σημαντικότητας $\alpha = 0.05$). Συμφωνούν οι τέσσερις προσεγγίσεις; Επίσης δήλωσε τα τρία ζεύγη με την πιο στατιστικά σημαντική συσχέτιση με βάση τους δύο ελέγχους (παραμετρικό και τυχαιοποίησης). Συμφωνούν οι δύο έλεγχοι;

5. Η αμοιβαία πληροφορία $I(X, Y)$ δύο τ.μ. X και Y χρησιμοποιείται και ως μέτρο γραμμικής και μη-γραμμικής συσχέτισης και ορίζεται ως $I(X, Y) = H(X) + H(Y) - H(X, Y)$, όπου $H(X) = -\sum_x f_X(x) \log f_X(x)$ είναι η εντροπία (του Shannon) της X και $f_X(x)$ η συνάρτηση μάζας πιθανότητας της διακριτοποιημένης X (αν η X είναι συνεχής). Όμοια ορίζεται η κοινή εντροπία των X και Y, $H(X, Y) = -\sum_{x,y} f_{X,Y}(x, y) \log f_{X,Y}(x, y)$, όπου $f_{X,Y}(x, y)$ είναι η από κοινού συνάρτηση μάζας πιθανότητας των διακριτοποιημένων X και Y (αν η X ή και η Y είναι συνεχής). Για την υλοποίηση της εκτίμησης αμοιβαίας πληροφορίας $I(X, Y)$ θα θεωρήσετε τη διακριτοποίηση των X και Y σε δύο τιμές, π.χ. 0 και 1, με βάση τη διάμεσο των τιμών της X και αντίστοιχα της Y, π.χ. τιμές μικρότερες της διαμέσου αντιστοιχίζονται στο 0 και μεγαλύτερες της διαμέσου στο 1. Η αμοιβαία πληροφορία θεωρητικά παίρνει την τιμή 0 όταν οι X και Y είναι ανεξάρτητες ενώ όταν είναι εξαρτημένες (γραμμικά ή μη-γραμμικά συσχετισμένες) παίρνει θετική τιμή που δηλώνει την ένταση της εξάρτησης. Φτιάξε μια συνάρτηση που να δέχεται ως δύο μεταβλητές εισόδου δύο διανύσματα ίσου μήκους (ζευγαρωτά δείγματα δύο τυχαίων μεταβλητών, τ.μ.) και να κάνει τα παρακάτω:

   (α') Να βρίσκει τα κενά στοιχεία στα διανύσματα (NaN) και να αφαιρεί τα αντίστοιχα ζεύγη τιμών έτσι ώστε τα δύο διανύσματα (ενδεχομένως μικρότερου μήκους) να μην έχουν κενά στοιχεία.

   (β') Να υπολογίζει την εκτίμηση αμοιβαίας πληροφορίας των δύο μεταβλητών όπως περιγράφεται παραπάνω.

   (γ') Να κάνει μη-παραμετρικό έλεγχο χρησιμοποιώντας τη μέθοδο της τυχαιοποίησης (δες άσκηση 5.2).

   (δ') Να δίνει στην έξοδο την τιμή της αμοιβαίας πληροφορίας και την p τιμή του ελέγχου καθώς και το μήκος των διανυσμάτων χωρίς κενά.

   Επίλεξε κάποια ζεύγη δεικτών που μπορεί να έχουν μη-γραμμική συσχέτιση (ελεύθερη επιλογή). Κάνε ένα πρόγραμμα που για κάθε ένα από αυτά τα ζεύγη υπολογίζει τον συντελεστή συσχέτισης Pearson και την p-τιμή του ελέγχου σημαντικότητας (ελεύθερη επιλογή ελέγχου). Το πρόγραμμα θα καλεί επίσης την παραπάνω συνάρτηση και θα εκτυπώνει τα αποτελέσματα από τον συντελεστή συσχέτισης Pearson και την αμοιβαία πληροφορία. Σχολίασε τα αποτελέσματα ως προς την στατιστική σημαντικότητα των μέτρων (με βάση τις p-τιμές τους) και τα στατιστικό συμπέρασμα για ύπαρξη γραμμικής και μη-γραμμικής συσχέτισης.

6. Φτιάξε μια συνάρτηση που να δέχεται ως δύο μεταβλητές εισόδου δύο διανύσματα ίσου μήκους (ζευγαρωτά δείγματα δύο τυχαίων μεταβλητών, τ.μ.) και να κάνει τα παρακάτω:

   (α') Να βρίσκει τα κενά στοιχεία στα διανύσματα (NaN) και να αφαιρεί τα αντίστοιχα ζεύγη τιμών έτσι ώστε τα δύο διανύσματα (ενδεχομένως μικρότερου μήκους) να μην έχουν κενά στοιχεία.

(β΄) Να προσαρμόζει γραμμικό μοντέλο παλινδρόμησης της δεύτερης τ.μ. ως προς την πρώτη τ.μ. με τη μέθοδο ελαχίστων τετραγώνων και να υπολογίζει το συντελεστή προσδιορισμού.

(γ΄) Να σχηματίζει το διάγραμμα διασποράς και την εκτιμώμενη ευθεία (ως στοιχείο σε πίνακα γραφημάτων), και να φαίνεται στο σχήμα ο συντελεστής προσδιορισμού.

(δ΄) Να δίνει στην έξοδο τον συντελεστή προσδιορισμού.

Κάλεσε την συνάρτηση αυτή για όλα τα ζεύγη των 10 δεικτών εκτός του δείκτη για τις μέρες του έτους με ανεμοστρόβιλο (TN). Η σειρά θα είναι ως εξής: θα θεωρείς έναν από τους 10 δείκτες ως εξαρτημένη τ.μ. και κάθε έναν από τους άλλους 9 δείκτες ως ανεξάρτητη. Τα 9 σχήματα των διαγραμμάτων διασποράς θα πρέπει να δίνονται σε ένα παράθυρο (πίνακας 9 σχημάτων). Στο τέλος το πρόγραμμα θα δίνει στην έξοδο, π.χ. σε μορφή πίνακα, για κάθε εξαρτημένη μεταβλητή (από τις 10 μεταβλητές) τα δύο μοντέλα (αρκεί μόνο την κάθε ανεξάρτητη μεταβλητή, όχι την εξίσωση του μοντέλου) με το μεγαλύτερο συντελεστή προσδιορισμού. Σχολίασε ποιοι δείκτες φαίνεται να μπορούν να εξηγηθούν καλύτερα με γραμμικό μοντέλο από κάποιον άλλο δείκτη.

7. Μας ενδιαφέρει να δούμε αν μπορούμε να προσδιορίσουμε τον δείκτη για τις μέρες του έτους με ομίχλη (FG) με ένα κατάλληλο μοντέλο ως προς έναν άλλο δείκτη (για κάθε ένα δείκτη εκτός του δείκτη για τις μέρες του έτους με ανεμοστρόβιλο, TN). Γι αυτό, φτιάξε μια συνάρτηση που να δέχεται ως δύο μεταβλητές εισόδου δύο διανύσματα ίσου μήκους (ζευγαρωτά δείγματα δύο τυχαίων μεταβλητών, τ.μ.), π.χ. η πρώτη ως ανεξάρτητη μεταβλητή και η δεύτερη ως εξαρτημένη (FG) και να κάνει τα παρακάτω:

   (α΄) Να βρίσκει τα κενά στοιχεία στα διανύσματα (NaN) και να αφαιρεί τα αντίστοιχα ζεύγη τιμών έτσι ώστε τα δύο διανύσματα (ενδεχομένως μικρότερου μήκους) να μην έχουν κενά στοιχεία.

   (β΄) Να προσαρμόζει με τη μέθοδο ελαχίστων τετραγώνων πολυωνυμικά μοντέλα παλινδρόμησης πρώτου, δευτέρου και τρίτου βαθμού, καθώς και μη-γραμμικά αλλά εγγενή γραμμικά μοντέλα (ελεύθερη επιλογή) και να υπολογίζει τον προσαρμοσμένο συντελεστή προσδιορισμού για κάθε μοντέλο.

   (γ΄) Να σχηματίζει το διάγραμμα διασποράς και την εκτιμώμενη καμπύλη για κάθε μοντέλο (να φαίνεται στο σχήμα ο προσαρμοσμένος συντελεστής προσδιορισμού για κάθε μοντέλο). Όλα τα σχήματα θα πρέπει να είναι σε ένα παράθυρο σχημάτων (χρήση subplot).

   (δ΄) Με βάση τον προσαρμοσμένο συντελεστή προσδιορισμού να επιλέγει το κατάλληλο μοντέλο.

   (ε΄) Να δίνει στην έξοδο τον τύπο του κατάλληλου μοντέλου (μπορεί να έχετε προσδιορίσει με κάποιο κωδικό ή αύξοντα αριθμό τον κάθε τύπο μοντέλου που δοκιμάζετε) μαζί με τον προσαρμοσμένο συντελεστή προσδιορισμού.

   Κάλεσε την συνάρτηση αυτή για εξαρτημένη μεταβλητή τις μέρες του έτους με ομίχλη (FG) και ανεξάρτητη μεταβλητή να είναι κάθε ένας δείκτης εκτός του δείκτη για τις μέρες του έτους με ανεμοστρόβιλο (TN). Με βάση τα αποτελέσματα σχολίασε ποιοι δείκτες φαίνεται

να μπορούν να εξηγήσουν καλύτερα τον δείκτη για τις μέρες του έτους με ομίχλη (FG) και με ποιο μοντέλο.

8. Θέλουμε να ελέγξουμε με βάση τον (προσαρμοσμένο) συντελεστή προσδιορισμού αν το μοντέλο απλής παλινδρόμησης, όπως αυτά στο προηγούμενο ζήτημα, είναι στατιστικά σημαντικό. Ένας τρόπος είναι να κάνουμε έλεγχο τυχαιοποίησης για τον συντελεστή προσδιορισμού (δες άσκηση 5.2). Κάνε μια συνάρτηση που τον υλοποιεί. Η συνάρτηση θα πρέπει να δέχεται ως δύο μεταβλητές εισόδου δύο διανύσματα ίσου μήκους (ζευγαρωτά δείγματα δύο τυχαίων μεταβλητών, τ.μ.), π.χ. η πρώτη ως ανεξάρτητη μεταβλητή και η δεύτερη ως εξαρτημένη και να κάνει τα παρακάτω:

   (α′) Να βρίσκει τα κενά στοιχεία στα διανύσματα (NaN) και να αφαιρεί τα αντίστοιχα ζεύγη τιμών έτσι ώστε τα δύο διανύσματα (ενδεχομένως μικρότερου μήκους) να μην έχουν κενά στοιχεία.

   (β′) Να προσαρμόζει με τη μέθοδο ελαχίστων τετραγώνων ένα μη-γραμμικό μοντέλο παλινδρόμησης, δηλαδή πολυωνυμικό μοντέλο κάποιου βαθμού ή μη-γραμμικό αλλά εγγενή γραμμικό μοντέλο (ελεύθερη επιλογή), και να υπολογίζει τον προσαρμοσμένο συντελεστή προσδιορισμού του μοντέλου.

   (γ′) Να κάνει έλεγχο τυχαιοποίησης για τον τον προσαρμοσμένο συντελεστή προσδιορισμού του μοντέλου και να υπολογίζει την p τιμή του ελέγχου (το πλήθος των τυχαιοποιημένων δειγμάτων να είναι τουλάχιστον 1000).

   (δ′) Να δίνει στην έξοδο τον προσαρμοσμένο συντελεστή προσδιορισμού και την p τιμή του ελέγχου σημαντικότητας του για το επιλεγμένο μοντέλο.

   Κάλεσε την συνάρτηση αυτή για εξαρτημένη μεταβλητή τις μέρες του έτους με ομίχλη (FG) και ανεξάρτητη μεταβλητή να είναι κάθε ένας δείκτης εκτός του δείκτη για τις μέρες του έτους με ανεμοστρόβιλο (TN). Με βάση τα αποτελέσματα για το επιλεγμένο μοντέλο σχολίασε ποιοι δείκτες φαίνεται να μπορούν να εξηγήσουν τον δείκτη για τις μέρες του έτους με ομίχλη (FG).

9. Μας ενδιαφέρει να δούμε αν μπορούμε να προσδιορίσουμε τον δείκτη για τις μέρες του έτους με ομίχλη (FG) με ένα κατάλληλο γραμμικό μοντέλο που να συμπεριλαμβάνει τους πιο σχετικούς από τους άλλους δείκτες. Το ίδιο για το δείκτη για τις μέρες του έτους με χαλάζι (GR). Η ανάλυση θα γίνει για τα ιστορικά δεδομένα μετά το 1973. Το παρακάτω πρόγραμμα θα εκτελεστεί δύο φορές, για το δείκτη FG και το δείκτη GR. Το πρόγραμμα θα πρέπει:

   (α′) Να βρίσκει τα έτη που υπάρχουν παρατηρήσεις για όλους τους δείκτες και να χρησιμοποιεί στην ανάλυση μόνο αυτά.

   (β′) Να υπολογίζει το μοντέλο πολλαπλής γραμμικής παλινδρόμησης με όλους τους δείκτες και να εκτυπώνει τη διασπορά των σφαλμάτων, το συντελεστή προσδιορισμού και τον προσαρμοσμένο συντελεστή προσδιορισμού, και να δηλώνει για ποιους δείκτες ο αντίστοιχος συντελεστής στο μοντέλο είναι στατιστικά σημαντικός (σε επίπεδο σημαντικότητας $\alpha = 0.05$).

(γ΄) Να κάνει το ίδιο για το μοντέλο που προκύπτει από τη μέθοδο της βηματικής παλινδρόμησης και κάποιο άλλο μοντέλο που εφαρμόζει μείωση διάστασης.

Σχολιάστε για κάθε έναν από τους δύο δείκτες (FG και GR) τα τρία μοντέλα (το μοντέλο με όλους τους δείκτες, το μοντέλο με τους επιλεγμένους δείκτες από βηματική παλινδρόμηση ως ανεξάρτητες μεταβλητές και το άλλο μοντέλο μείωσης διάστασης). Θα μπορούσαμε να εξηγήσουμε ικανοποιητικά το δείκτη FG (και το ίδιο για τον δείκτη GR) με κάποιους από τους άλλους δείκτες και ποιους;

10. Θέλουμε να διερευνήσουμε αν η μέθοδος LASSO μας δίνει το βέλτιστο μοντέλο και για ποια παράμετρο ποινής. Αυτό θα το διερευνήσουμε στο πρόβλημα γραμμικής πολλαπλής παλινδρόμησης για δείκτες που έχουμε δεδομένα. Φτιάξε μια συνάρτηση που παίρνει ως είσοδο ένα διάνυσμα εξαρτημένης μεταβλητής και έναν πίνακα που έχει σε στήλες τις ανεξάρτητες μεταβλητές και θα πρέπει να κάνει τα παρακάτω:

    (α΄) Να βρίσκει τις γραμμές (τα έτη) που υπάρχουν παρατηρήσεις για όλους τους δείκτες και να χρησιμοποιεί στην ανάλυση μόνο αυτά.

    (β΄) Να υπολογίζει τον προσαρμοσμένο συντελεστή προσδιορισμού για όλα τα δυνατά μοντέλα γραμμικής παλινδρόμησης (από αυτό της μέσης τιμής με κανένα δείκτη ως το πλήρες μοντέλο με όλους τους δείκτες), που για 8 ανεξάρτητες μεταβλητές είναι $2^8 = 256$ μοντέλα. Να επιλέγει ως βέλτιστο μοντέλο αυτό με τον υψηλότερο προσαρμοσμένο συντελεστή προσδιορισμού.

    (γ΄) Να εφαρμόζει τη μέθοδο LASSO και να εξετάζει για όλο το εύρος των τιμών του συντελεστή ποινής αν για κάποια τιμή του συντελεστή επιτυγχάνεται μοντέλο ίδιο με αυτό του βέλτιστου μοντέλου που βρέθηκε στο προηγούμενο βήμα.

    (δ΄) Να δίνει στην έξοδο το βέλτιστο μοντέλο και τον συντελεστή ποινής της μεθόδου LASSO που το πετυχαίνει (αλλιώς να δίνει κενό).

    Κάλεσε την συνάρτηση αυτή για εξαρτημένη μεταβλητή τις μέρες του έτους με ομίχλη (FG) και ανεξάρτητες μεταβλητές όλες τις άλλες εκτός των δεικτών TN και GR. Κάνε το ίδιο για εξαρτημένη μεταβλητή τις μέρες του έτους με χαλάζι (GR) και ανεξάρτητες μεταβλητές όλες τις άλλες εκτός των δεικτών TN και FG. Πετυχαίνει η μέθοδος LASSO να βρίσκει το βέλτιστο μοντέλο και για ποιον συντελεστή ποινής στις δύο περιπτώσεις;