Μεταπτυχιακό Υπολογιστικής Φυσικής, Φεβρουάριος 2019
Εξέταση στο μάθημα 'Ανάλυση Δεδομένων'

Δημήτρης Κουγιουμτζής

13 Φεβρουαρίου 2019

Οδηγίες: Όλα τα αρχεία συναρτήσεων και προγραμμάτων Matlab που σας ζητούνται, με τα ονόματα που σας ζητούνται, θα είναι μέσα σε έναν φάκελο με όνομα το ονοματεπώνυμο σας, τον οποίο θα παραδώσετε. Καλό είναι να υπάρχουν κάποια βασικά σχόλια (χρησιμοποιώντας τον χαρακτήρα %) στον κώδικα του Matlab.

Δεδομένα

Στο αρχείο `EnergyUS.xlsx` δίνονται ιστορικά δεδομένα από την περίοδο 3/1/2017 - 9/11/2018 για 10 δείκτες μετοχών εταιριών ενέργειας που συμμετέχουν στο New York Stock Exchange (NYSE) καθώς και για τον δείκτη NYSE (469 τιμές για κάθε δείκτη). Το αρχείο είναι σε μορφή excel και διαβάζεται με την εντολή Matlab `xlsread`. Η πρώτη γραμμή έχει την ετικέτα για την κάθε στήλη (για τις μετοχές εταιριών δίνονται τα ακρωνύμια τους). Στην πρώτη στήλη είναι η ημερομηνία (δε θα τη χρειαστείτε στα ερωτήματα), στη δεύτερη στήλη η τιμή του δείκτη NYSE και στις επόμενες 10 στήλες οι τιμές μετοχών των 10 εταιριών ενέργειας.

Η ανάλυση θα γίνει στη χρονοσειρά των μεταβολών του κάθε δείκτη (πρώτες διαφορές), που ορίζεται ως η διαφορά της τιμής του δείκτη από τη μια μέρα στην άλλη. Ειδικότερα αν y_t, $t = 1, \ldots, n'$, είναι η χρονοσειρά των $n' = 469$ τιμών κάποιου δείκτη, τότε η χρονοσειρά των μεταβολών είναι $x_t = y_t - y_{t-1}$, για $t = 2, \ldots, n'$ και έχει μήκος $n = n' - 1 = 468$.

Ερωτήματα

Θέτουμε δύο βασικά ερωτήματα με βάση αυτά τα δεδομένα: 1) αν οι μεταβολές του δείκτη NYSE για κοντινές μέρες συσχετίζονται, και β) αν μπορούμε να προβλέψουμε τη μεταβολή της τιμής του δείκτη NYSE από τις μεταβολές τιμών των άλλων δεικτών μετοχών από τις 10 εταιρίες ενέργειες και από ποιες.

1. Θα διερευνήσετε αν υπάρχουν συσχετίσεις στη χρονοσειρά των μεταβολών του δείκτη NYSE. Αυτό θα το κάνετε με δύο αναλύσεις που θα υλοποιηθούν σε δύο συναρτήσεις αντίστοιχα και σε ένα πρόγραμμα που θα τις καλεί:

   (α') *Υπολογισμός και έλεγχος σημαντικότητας της αυτοσυσχέτισης για συγκεκριμένες υστερήσεις.* Ο υπολογισμός της αυτοσυσχέτισης σε μια χρονοσειρά x_t, $t = 1, \ldots, n$, για

κάποια υστέρηση τ, $r(\tau)$, θα γίνει με τον συντελεστή συσχέτισης δύο μεταβλητών X και Y, $r(X, Y)$, όπου $X = X_t$ και $Y = X_{t+\tau}$. Για παράδειγμα, αν $\tau = 1$, το αντίστοιχο δείγμα των δύο μεταβλητών έχει τις ζευγαρωτές παρατηρήσεις

$$\{(x_1, x_2), (x_2, x_3), \ldots, (x_{n-1}, x_n)\}.$$

Θα γίνει επίσης έλεγχος σημαντικότητας της αυτοσυσχέτισης με τυχαιοποίηση (ανακατεύοντας τυχαία τη χρονοσειρά x_t, $t = 1, \ldots, n$) για κάθε υστέρηση ξεχωριστά. Για αυτό θα φτιάξετε μια συνάρτηση με όνομα `MyAutocorrelation`, που ως είσοδο θα παίρνει ένα πρώτο διάνυσμα με τη χρονοσειρά των μεταβολών, ένα δεύτερο διάνυσμα θετικών ακεραίων με τις υστερήσεις (ο υπολογισμός θα γίνει για κάθε μια από αυτές), μια τιμή για το πλήθος των τυχαιοποιήσεων καθώς και μια τιμή για το επίπεδο σημαντικότητας του ελέγχου σημαντικότητας. Στην έξοδο θα δίνει ένα διάνυσμα με τις τιμές αυτοσυσχέτισης (για τις αντίστοιχες υστερήσεις, με τη σειρά που δόθηκαν στο διάνυσμα στην είσοδο) και ένα διάνυσμα με την απόφαση του ελέγχου σημαντικότητας (1 απόρριψη, 0 μη-απόρριψη), όπου και τα δύο διανύσματα θα έχουν μέγεθος όσο και το πλήθος των υστερήσεων που δίνεται στην είσοδο.

(β΄) *΄Ελεγχος ανεξαρτησίας για συγκεκριμένες υστερήσεις.* Εδώ επικεντρωνόμαστε στις αυξήσεις και μειώσεις και αγνοούμε τις αριθμητικές τιμές των μεταβολών. Για αυτό πρώτα θα μετατρέψετε τη χρονοσειρά των μεταβολών τιμών σε χρονοσειρά δύο τιμών, ένα (1) για αύξηση και μείον ένα (-1) για μείωση, δηλαδή κρατάτε μόνο το πρόσημο των τιμών μεταβολής και όχι την αριθμητική τιμή τους. Για κάθε υστέρηση τ, θα κάνετε έλεγχο ανεξαρτησίας μεταξύ των αντίστοιχων μεταβλητών (των δύο τιμών, 1 και -1), δηλαδή για $X = X_t$ και $Y = X_{t+\tau}$ για υστέρηση τ. Συγκεκριμένα, θα πρέπει να ελέγξετε αν ισχύει $f_{X_t, X_{t+\tau}}(x_t, x_{t+\tau}) = f_{X_t}(x_t) f_{X_{t+\tau}}(x_{t+\tau})$, όπου $f_{X,Y}(x, y)$ η από κοινού συνάρτηση μάζας πιθανότητας δύο μεταβλητών X και Y, δηλαδή η πιθανότητα οι X και Y να παίρνουν τιμές x και y, αντίστοιχα, και $f_X(x)$ η συνάρτηση μάζας πιθανότητας της X, δηλαδή η πιθανότητα η X να παίρνει την τιμή x. Θα χρησιμοποιήσετε έλεγχο καταλληλότητας $\mathcal{X}^2$ για το αν η από κοινού κατανομή που δίνεται από την $f_{X_t, X_{t+\tau}}(x_t, x_{t+\tau})$ προσαρμόζεται στο γινόμενο των περιθώριων κατανομών $f_{X_t}(x_t) f_{X_{t+\tau}}(x_{t+\tau})$, δηλαδή η μηδενική υπόθεση είναι

$$\mathrm{H}_0 : f_{X_t, X_{t+\tau}}(x_t, x_{t+\tau}) = f_{X_t}(x_t) f_{X_{t+\tau}}(x_{t+\tau})$$

Θα κάνετε έλεγχο σημαντικότητας της $\mathcal{X}^2$ με τυχαιοποίηση (ανακατεύοντας τυχαία τη χρονοσειρά x_t, $t = 1, \ldots, n$) για κάθε υστέρηση. Για αυτό θα φτιάξετε μια συνάρτηση με όνομα `MyLagChiSquare`, που θα έχει την ίδια είσοδο με την συνάρτηση στο ερώτημα 1α΄, δηλαδή το διάνυσμα της δίτιμης χρονοσειράς, το διάνυσμα των υστερήσεων, το πλήθος των τυχαιοποιήσεων και το επίπεδο σημαντικότητας. Στην έξοδο η συνάρτηση θα δίνει ένα διάνυσμα με τις τιμές του $\mathcal{X}^2$ και ένα διάνυσμα με την απόφαση του ελέγχου σημαντικότητας (1 απόρριψη, 0 μη-απόρριψη), όπου και τα δύο διανύσματα θα έχουν μέγεθος όσο και το πλήθος των υστερήσεων που δίνεται στην είσοδο.

(γ΄) Θα φτιάξετε ένα πρόγραμμα με όνομα `Exercise1` που θα διαβάζει τη χρονοσειρά NYSE και θα υπολογίζει την αυτοσυσχέτιση μαζί με έλεγχο σημαντικότητας τυχαιοποίησης για υστερήσεις 1, 2 και 3 (συνάρτηση `MyAutocorrelation`) καθώς και τον έλεγχο $\mathcal{X}^2$ με τυχαιοποίηση για υστερήσεις 1, 2 και 3 (συνάρτηση

`MyLagChiSquare`). Το πρόγραμμα θα εμφανίζει στη γραμμή εντολών Matlab σε μορφή πίνακα τις τιμές αυτοσυσχέτισης και $\mathcal{X}^2$ για κάθε υστέρηση και θα δηλώνει αν υπάρχουν συσχετίσεις και για ποιες υστερήσεις στη χρονοσειρά των αριθμητικών μεταβολών και στη χρονοσειρά που μετράει αύξηση (1) και μείωση (-1). Σχολιάστε αν συμφωνούν τα αποτελέσματα με τις δύο αναλύσεις.

2. Θα διερευνήσετε αν οι μεταβολές του δείκτη NYSE μπορούν να καθοριστούν, και σε ποιο βαθμό, από τις μεταβολές των δεικτών των δέκα εταιριών ενέργειας ή και μόνο κάποιων από αυτές. Για αυτό θα κάνετε τα παρακάτω:

   (α') Θα φτιάξετε μια συνάρτηση με όνομα `MyRegress`, που θα παίρνει ως είσοδο το διάνυσμα της εξαρτημένης μεταβλητής (π.χ. μεταβολή τιμής NYSE) και τον πίνακα των ανεξάρτητων μεταβλητών (π.χ. μεταβολή τιμής για κάθε μια από τις 10 εταιρίες). Στην έξοδο θα δίνει τον προσαρμοσμένο συντελεστή προσδιορισμού adjusted-R^2 για το πλήρες μοντέλο με όλες τις 10 μεταβλητές, καθώς και για το μοντέλο που καταλήγει η μέθοδος της βηματικής παλινδρόμησης. Θα δίνει επίσης και ένα διάνυσμα με τους δείκτες των ανεξάρτητων μεταβλητών που επιλέχτηκαν από το μοντέλο της βηματικής παλινδρόμησης. Η συνάρτηση θα σχηματίζει το διάγραμμα διασποράς της κάθε μιας ανεξάρτητης μεταβλητής με την εξαρτημένη μεταβλητή, καθώς και το διάγραμμα των κανονικοποιημένων σφαλμάτων ως προς την εξαρτημένη μεταβλητή (με τα όρια για κανονική κατανομή σε επίπεδο σημαντικότητας $\alpha = 0.05$) για το πλήρες μοντέλο παλινδρόμησης και για το μοντέλο από τη βηματική παλινδρόμηση.

   (β') Θα φτιάξετε ένα πρόγραμμα με όνομα `Exercise2` που θα καλεί τη συνάρτηση `MyRegress`, δίνοντας στην είσοδο τη μεταβολή τιμής NYSE ως εξαρτημένη μεταβλητή και τον πίνακα των μεταβολών τιμής για κάθε μια από τις 10 εταιρίες ως ανεξάρτητες μεταβλητές. Το πρόγραμμα θα δίνει τους όρους του μοντέλου βηματικής παλινδρόμησης και θα εμφανίζει στη γραμμή εντολών του Matlab το adjusted-R^2 για τα δύο μοντέλα.

   (γ') Στη συνέχεια το ίδιο πρόγραμμα `Exercise2` θα τεμαχίζει τις 468 παρατηρήσεις σε μη-επικαλυπτόμενα παράθυρα των 25 παρατηρήσεων (χοντρικά ένας μήνας). Σε κάθε παράθυρο των 25 παρατηρήσεων το πρόγραμμα θα καλεί τη συνάρτηση `MyRegress` όπως και παραπάνω (εξαρτημένη μεταβλητή η μεταβολή τιμής NYSE και ανεξάρτητες οι μεταβολές τιμής για κάθε μια από τις 10 εταιρίες περιορίζοντας στις τιμές του παραθύρου). Θα κρατά τα αποτελέσματα της συνάρτησης από όλα τα παράθυρα σε κατάλληλες μεταβλητές πινάκων ή/και διανυσμάτων. Στο τέλος θα δίνει σε ένα σχήμα το προφίλ του adjusted-R^2 ως προς τη χρονική περίοδο (αύξων αριθμός του κυλιόμενου χρονικού παραθύρου, όχι ημερολογιακός χρόνος) για το πλήρες μοντέλο και το μοντέλο της βηματικής παλινδρόμησης. Επίσης θα βρίσκει το πιο συχνό μοντέλο βηματικής παλινδρόμησης που βρέθηκε στα χρονικά παράθυρα και θα το εμφανίζει στη γραμμή εντολών του Matlab. Συγκρίνετε τα αποτελέσματα από τα χρονικά παράθυρα με αυτά σε όλη την καταγραφή.