

Ανάλυση Δεδομένων με Η/Υ

ΘΕΜΑ 1 (3.0 μονάδες):

Δυναδική αναζήτηση ονομάζεται ένας αναδρομικός αλγόριθμος αναζήτησης ενός στοιχείου (το λεγόμενο στοιχείο κλειδί) σε ένα ταξινομημένο κατά αύξουσα τάξη μεγέθους διάνυσμα μήκους έστω n . Θεωρήστε ότι το διάνυσμα δεν περιέχει ισοπαλίες (στοιχεία δηλαδή με την ίδια τιμή). Ο αλγόριθμος δέχεται ως είσοδο το αρχικό (μη ταξινομημένο) διάνυσμα και το στοιχείο κλειδί. **Επιστρέφει τη θέση του κλειδιού στο ταξινομημένο διάνυσμα** ή μια ειδική ένδειξη για να ενημερώσει ότι το κλειδί δεν υπάρχει στο διάνυσμα, αν αυτή είναι η περίπτωση. Δημιουργήστε μια δική σας συνάρτηση στην R για να υλοποιήσετε τον εν λόγω αλγόριθμο εφαρμόζοντας τα παρακάτω βήματα:

- i) Ξεκινήστε ταξινομώντας αρχικά το δοθέν διάνυσμα σε αύξουσα τάξη μεγέθους και καλέστε $low = 1$ και $high = n$. Αυτό θα έχει ως αποτέλεσμα η αρχική σας αναζήτηση να γίνει σε όλο το αρχικό διάνυσμα.
- ii) Συγκρίνετε εν συνεχεία το δοθέν κλειδί με την τιμή του στοιχείου που βρίσκεται στο μέσο του διανύσματος. Ως μέσο του διανύσματος θεωρήστε τη θέση $mid = \lfloor (low+high)/2 \rfloor$, όπου $\lfloor \cdot \rfloor$ δηλώνει το ακέραιο μέρος (χρησιμοποιείστε στην R την εντολή *floor*).
- iii) Αν είναι ίσα, τότε ο αλγόριθμος τερματίζει επιστρέφοντας τη θέση αυτή.
- iv) Αν το κλειδί είναι μικρότερο, τότε επαναλάβετε τα βήματα ii) και iii), όχι όμως για όλο το διάνυσμα αλλά για το πρώτο μισό του διανύσματος. Αυτό επιτυγχάνεται εύκολα αλλάζοντας την τιμή $high = mid - 1$ και διατηρώντας την προηγούμενη τιμή του low .
- v) Αν το κλειδί είναι μεγαλύτερο, τότε επαναλάβετε τα βήματα ii) και iii), όχι όμως για όλο το διάνυσμα αλλά για το δεύτερο μισό του διανύσματος. Αυτό επιτυγχάνεται εύκολα αλλάζοντας την τιμή $low = mid + 1$ και διατηρώντας την προηγούμενη τιμή του $high$.
- vi) Επαναλάβετε τα βήματα iv) και v) μέχρι να βρείτε είτε τη θέση του ζητούμενου στοιχείου, είτε μέχρι να ολοκληρωθεί η αναζήτηση (σε αυτή την περίπτωση το $low > high$) χωρίς αποτέλεσμα.

ΘΕΜΑ 2 (5.0 μονάδες): Υποθέστε ότι έχετε δεδομένα για 30 μέλη ΔΕΠ ενός Πανεπιστημίου της Ευρώπης, τα οποία συνοψίζονται στις ακόλουθες 5 τυχαίες μεταβλητές: **Salary** (ο ετήσιος μισθός του μέλους ΔΕΠ σε ευρώ), **Gender** (0: για γυναίκα, 1: για άντρα), **Dept** (το τμήμα στο οποίο ανήκει το μέλος ΔΕΠ, **1**: Κοινωνικές Επιστήμες, **2**: Φυσικές Επιστήμες, **3**: Οικονομικές Επιστήμες), **Rank** (η βαθμίδα του μέλους ΔΕΠ, **1**: για Επίκουρο Καθηγητή, **2**: για Αναπληρωτή Καθηγητή και **3**: για Καθηγητή) και **Years** (το χρονικό διάστημα σε έτη, ως συνεχή μεταβλητή, από την πρόσληψη στην τελευταία βαθμίδα που ανήκει). Σκοπός της μελέτης είναι να διερευνηθεί αν ο μισθός των μελών ΔΕΠ σχετίζεται με κάποιες από τις υπόλοιπες τ.μ. του συνόλου δεδομένων.

(A) Προσαρμόσαμε ένα μοντέλο γραμμικής παλινδρόμησης στην R (και το αποτέλεσμά του το καταχωρήσαμε σε ένα αντικείμενο με το όνομα **modell**) με μεταβλητή απόκρισης την τ.μ. **Salary** και επεξηγηματικές μεταβλητές τις τ.μ. **Dept**, **Gender**, **Rank** και **Years**. Με ποια εντολή στην R θα προσαρμόζατε το εν λόγω πολλαπλό γραμμικό μοντέλο; Ποιες είναι οι εικονικές μεταβλητές στο μοντέλο και με ποιες κατηγορίες αναφοράς;

(B) Με βάση τα παρακάτω αποτελέσματα ερμηνεύστε τους εκτιμητές των συντελεστών του γραμμικού μοντέλου.

Residuals:				
Min	1Q	Median	3Q	Max
-6.6731	-2.2659	-0.6828	2.1642	10.7739

Coefficients:				
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	27.8461	2.1637	12.870	5.40e-12 ***
Gender1	6.4814	1.9011	3.409	0.002404 **
Rank2	10.7552	2.0555	5.233	2.63e-05 ***
Rank3	15.5568	2.0487	7.593	1.04e-07 ***
Dept2	9.5739	2.3052	4.153	0.000384 ***
Dept3	16.6137	1.8597	8.934	6.14e-09 ***
Years	0.6709	0.2636	2.545	0.018085 *

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 4.279 on 23 degrees of freedom
Multiple R-squared: 0.8753, Adjusted R-squared: 0.8428
F-statistic: 26.91 on 6 and 23 DF, p-value: 2.641e-09

- (F) Με ποιον τρόπο θα αλλάζατε την κατηγορία αναφοράς της μεταβλητής Dept σε αυτή των Οικονομικών Επιστημών; Πως θα μεταβάλλονταν τότε οι τιμές των εκτιμητών των συντελεστών του ίδιου γραμμικού μοντέλου.
- (Δ) Ποιες είναι οι προϋποθέσεις του πολλαπλού γραμμικού μοντέλου που προσαρμόσατε προηγουμένως; Με ποιες εντολές της R μπορείτε να τις ελέγξετε;
- (E) Ποιον έλεγχο εξετάζουμε με το F-test στη παραπάνω πολλαπλή γραμμική παλινδρόμηση; Με βάση τα παραπάνω αποτελέσματα σε τι τελικά συμπεράσματα θα καταλήγατε;
- (ΣΤ) Δώστε τον τύπο του συντελεστή προσδιορισμού καθώς και τον τύπο του διορθωμένου συντελεστή προσδιορισμού στο παραπάνω μοντέλο. Ποια είναι η χρησιμότητα του τελευταίου; Ερμηνεύστε την τιμή του συντελεστή προσδιορισμού στο μοντέλο που προσαρμόσατε προηγουμένως.
- (Z) Με ποια εντολή της R θα εκτιμούσατε σημειακά και με τη βοήθεια ενός 95% διαστήματος εμπιστοσύνης τον αναμενόμενο ετήσιο μισθό σε ευρώ για μια επίκουρη καθηγήτρια των Κοινωνικών Επιστημών με τρία χρόνια υπηρεσίας στην τελευταία της βαθμίδα.
- (H) Έστω ότι προσαρμόζουμε ξανά το ίδιο μοντέλο παλινδρόμησης κεντράροντας όμως αυτή τη φορά τις τιμές της τ.μ. Years. Δοθέντος ότι η τιμή του δειγματικού μέσου για την εν λόγω μεταβλητή είναι 3.2 έτη, ποιοι συντελεστές θα αλλάζαν τιμή; Ερμηνεύστε εκ νέου τις τιμές των συντελεστών που έχουν μεταβληθεί.
- (Θ) Έστω ότι προσαρμόζουμε εναλλακτικά το αντίστοιχο πολλαπλασιαστικό μοντέλο και η τιμή του συντελεστή της τ.μ. Years είναι 0.04. Ερμηνεύστε την εν λόγω τιμή.

ΘΕΜΑ 3 (2.0 μονάδες): Στα πλαίσια μιας έρευνας για το ύψος των μηνιαίων μισθών σε ευρώ, σε μια πολυεθνική εταιρεία ρωτήθηκαν 20 άντρες (Α) και 20 γυναίκες (Γ). Τα δεδομένα φαίνονται παρακάτω.

	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Γ	1020	998	976	1091	678	865	1003	1005	987	934	1230	988	843	1067	895	1001	768	1251	1009	908
Α	1234	1456	890	756	789	890	1023	1178	1192	904	956	678	1008	1267	1123	1056	1303	1083	854	896

- Θέλουμε να ελέγξουμε σε ε.σ. 2% την υπόθεση ότι η μέση τιμή του μισθού των υπαλλήλων της εταιρείας δε διαφέρει ανάμεσα στα δύο φύλα με εναλλακτική ότι είναι χαμηλότερος στις γυναίκες.
- (α) Τι είδους στατιστική ανάλυση θα εφαρμόζατε και με ποιες εντολές στην R;
- (β) Ποιες προϋποθέσεις θα ελέγχατε για τη στατιστική ανάλυση του παραπάνω ερωτήματος και με ποιες εντολές στην R;
- (γ) Αν δεν ίσχυαν οι παραπάνω προϋποθέσεις, ποιον έλεγχο θα εφαρμόζατε και με ποια εντολή στην R;

Διάρκεια Εξέτασης: 2 ½ ώρες

ΕΥΧΟΜΑΙ ΕΠΙΤΥΧΙΑ