

## Ανάλυση Δεδομένων με R/Y

**ΘΕΜΑ 1 (2 μονάδες):** Ένας απλός και αποδοτικός αλγόριθμος ταξινόμησης με χαμηλή πολυπλοκότητα είναι ο αλγόριθμος εισαγωγής (*insertion sort*). Η διαδικασία ταξινόμησης ενός διανύσματος (μεγέθους  $n$ ) ξεκινά από το δεύτερο στοιχείο του και έχει ως εξής: σε κάθε επανάληψη το  $i$ -οστό στοιχείο του υπό ταξινόμηση διανύσματος ( $i = 2, \dots, n$ ) μετακινείται αριστερά (αν αυτό είναι απαραίτητο) καταλαμβάνοντας τη σωστή του θέση ανάμεσα στα ( $i - 1$ ) προηγούμενα στοιχεία.

## Παράδειγμα:

Αρχικό διάνυσμα: 6 5 9 1 2

Πρώτη επανάληψη: 5 6 9 1 2

Δεύτερη επανάληψη: 5 6 9 1 2

Τρίτη επανάληψη: 1 5 6 9 2

Τέταρτη επανάληψη: 1 2 5 6 9

sumary (comphr, hist, plot)

var

for

hist

boxplot

~~studyhr~~studyhr  $\leftarrow$  os. factor (studyhr)

table

prop.table (table)

barplot (table)

plot (table)

print (table)

Να γραφτεί μια συνάρτηση στην R που θα υλοποιεί τον παραπάνω αλγόριθμο για την ταξινόμηση ενός διανύσματος  $x$  το οποίο θα δίνει ο χρήστης σαν παράμετρο εισόδου. Η συνάρτηση θα πρέπει να επιστρέφει σε μια λίστα με κατάλληλα ονόματα το αρχικό αταξινόμητο διάνυσμα, καθώς και το τελικό ταξινομημένο διάνυσμα.

**ΘΕΜΑ 2 (5 μονάδες):** Μια οφθαλμολογική κλινική παίδων θέλησε να εξετάσει τους παράγοντες από τους οποίους εξαρτάται η ανάπτυξη της παιδικής μυωπίας. Σε δείγμα 618 παιδιών ηλικίας από 5 έως 9 ετών, συλλέχθηκε πληροφορία για τις ακόλουθες μεταβλητές: **gender** = το φύλο (0: αγόρι, 1: κορίτσι), **comphr** = οι ανά εβδομάδα ώρες που το παιδί έπαιζε βιντεοπαιχνίδια ή καθόταν στον ηλεκτρονικό υπολογιστή, **studyhr** = οι ανά εβδομάδα ώρες διαβάσματος/μελέτης για το σχολείο και **parentmy** = αν τουλάχιστον ένας από τους γονείς είναι μύωπας (0: όχι, 1: ναι). Ακόμη μετρήθηκε ο δείκτης σφαιρικού ισοδύναμου διάθλασης (**SER**) με μονάδα μέτρησης τη διόπτρα (D). Όσο πιο μικρή αρνητική τιμή έχει ο εν λόγω δείκτης, τόσο πιο υψηλή είναι η μυωπία του ατόμου.

(Α) Με ποιους τρόπους (αριθμητικούς και γραφικούς) και με ποιες εντολές στην R θα περιγράφατε τις τιμές κάθε μεταβλητής στο δείγμα; **μόνο την πουστίκης;**

(Β) Προσαρμόσαμε το μοντέλο γραμμικής παλινδρόμησης στην R (και το αποτέλεσμα του το καταχωρίσαμε σε ένα αντικείμενο με το όνομα **model1**) με μεταβλητή απόκρισης τη **SER** και επεξηγηματικές μεταβλητές τις **gender**, **comphr**, **studyhr** και **parentmy**. Με ποια εντολή στην R θα προσαρμόσατε το εν λόγω πολλαπλό γραμμικό μοντέλο; Ποιες είναι οι εικονικές μεταβλητές στο μοντέλο και με ποιες κατηγορίες αναφοράς; **comphr**  $\leftarrow$  **studyhr**  $\leftarrow$  **parentmy**, **studyhr**  $\leftarrow$  **parentmy**

(Γ) Με βάση τα παρακάτω αποτελέσματα ερμηνεύστε τις τιμές των εκτιμητών των συντελεστών του γραμμικού μοντέλου καθώς και την τιμή του συντελεστή προσδιορισμού.

το D2: δεν ήντι το μοντέλο

Call:					
lm(formula = SER ~ gender + comphr + studyhr + parentmy)					
Residuals:					
Min	1Q	Median	3Q	Max	
-1.5258	-0.3795	-0.0645	0.2382	3.5211	
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t )	
(Intercept)	0.899058	0.051494	17.460	< 2e-16 ***	
gender1	0.034177	0.051770	0.660	0.509399	
comphr	-0.002564	0.008531	-0.301	0.763833	
studyhr	-0.016794	0.011383	-1.475	0.140655	
parentmy1	-0.166492	0.050051	-3.326	0.000932 ***	
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Residual standard error: 0.621 on 613 degrees of freedom					
Multiple R-squared: 0.02209, Adjusted R-squared: 0.01571					
F-statistic: 3.461 on 4 and 613 DF, p-value: 0.008265					

6/10

dix

QB

+

ta Dohm

=

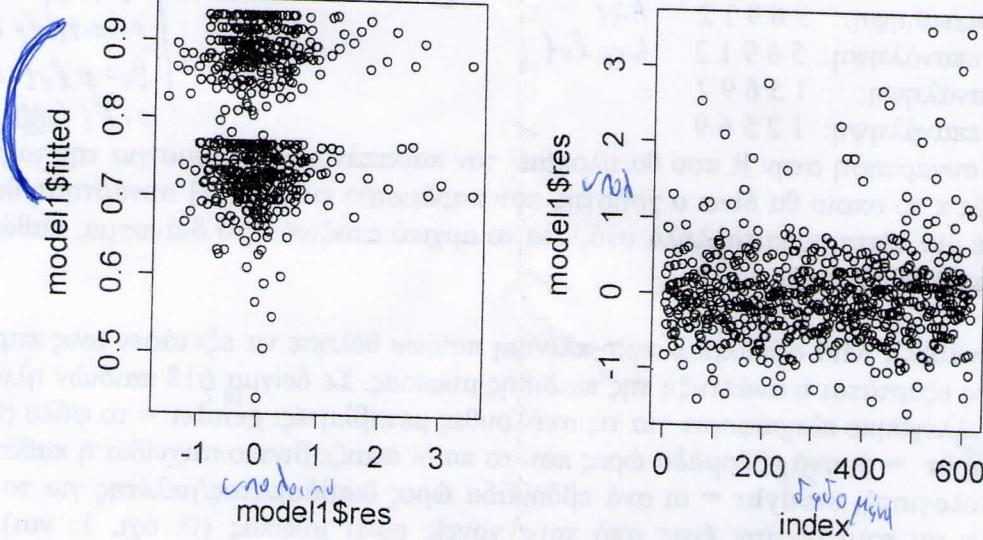
=

=

=

(Δ) Εκτιμήστε την αναμενόμενη τιμή του δείκτη SER ενός αγοριού που ξοδεύει περίπου 14 ώρες την εβδομάδα μπροστά στον υπολογιστή και άλλες 10 ώρες την εβδομάδα για μελέτη, δεδομένου ότι η μητέρα του έχει μυωπία. Δώστε και την αντίστοιχη εντολή στην R.

(Ε) Τα κάτωθι διαγράμματα απεικονίζουν τις προσαρμοσμένες τιμές του μοντέλου έναντι των υπολογίων (αριστερά) και τα υπόλοιπα του μοντέλου έναντι της σειράς των δεδομένων (δεξιά). Δώστε τις εντολές της R που δημιουργούν τα εν λόγω διαγράμματα. Αναφέρατε όλες τις προϋποθέσεις που πρέπει να ισχύουν για το παραπάνω μοντέλο, και σχολιάστε (χρησιμοποιώντας την πληροφορία που σας δίνεται από τα παρακάτω διαγράμματα και την περιγραφή του προβλήματος) αν το μοντέλο που προσαρμόσαμε είναι κατάλληλο για τον σκοπό της έρευνας.



### ΘΕΜΑ 3 (3 μονάδες):

Περιοδικά ποικίλης ύλης κατατάχτηκαν από το αναγνωστικό κοινό ανάλογα με το περιεχόμενό τους σε περιοδικά υψηλού εκπαιδευτικού επιπέδου (E1) και περιοδικά χαμηλού εκπαιδευτικού περιεχομένου (E2). Επιλέξαμε τυχαία 10 περιοδικά από κάθε κατηγορία και ένα άρθρο παρόμοιου μεγέθους μέσα σε κάθε περιοδικό εξ αυτών. Ένας διαδεδομένος δείκτης αναγνωσιμότητας (readability index) είναι το πλήθος των λέξεων ενός κειμένου που αποτελούνται από τρεις και πλέον συλλαβές. Ο δείκτης αυτός υπολογίστηκε για το άρθρο κάθε περιοδικού.

	E1	34	21	37	31	10	24	39	10	17	18	$r_1$
	E2	3	8	16	9	10	12	41	14	24	15	$r_2$

Θέλουμε να ελέγξουμε, σε ε.σ. 10%, την υπόθεση ότι η μέση τιμή του συγκεκριμένου δείκτη δε διαφέρει σημαντικά ανάμεσα σε περιοδικά υψηλού και χαμηλού εκπαιδευτικού περιεχομένου, με εναλλακτική ότι είναι υψηλότερος στα περιοδικά της πρώτης κατηγορίας.

Μια εντολή στην R που παρεμβάλλεται στη στατιστική ανάλυση του ερευνητικού αυτού ερωτήματος είναι η εξής: `var.test(E1, E2)`. Τα βασικά αποτελέσματα της εντολής αυτής είναι τα παρακάτω:

Βαθμοί ελαυνθρίδας των F κατανομής

$F = 1.0071$ , num df = 9, denom df = 9, p-value = 0.9917.

ρ σημειώσεις

$$F = \frac{G_1^2 / S_1^2}{G_2^2 / S_2^2} \rightarrow H_0: G_1^2 / S_1^2 = G_2^2 / S_2^2$$

$$H_1: G_1^2 / S_1^2 \neq G_2^2 / S_2^2$$

- Εξηγήστε την παραπάνω εντολή και τα αποτελέσματα που δίνονται.
- Ποια προϋπόθεση πρέπει να ικανοποιείται για να είναι έγκυρα τα αποτελέσματα του παραπάνω ελέγχου. Με ποιους τρόπους και εντολές στην R θα ελέγχατε την εν λόγω προϋπόθεση;
- Τι έπεται της παραπάνω εντολής ώστε να ολοκληρώθει σωστά η ανάλυση;

Διάρκεια Εξέτασης: 2 ½ ώρες

fuller  
tissue

EYXOMAI EPITYXIA

greater less