

Ανάλυση Δεδομένων με Η/Υ

**ΘΕΜΑ 1 (2.5 μονάδες):** Το *jackknife* είναι μια μέθοδος επαναδειγματοληψίας ιδιαίτερα χρήσιμη για την εκτίμηση του τυπικού σφάλματος ενός εκτιμητή  $\hat{\theta}$ . Θεωρούμε ότι έχουμε ένα τυχαίο δείγμα αποτελούμενο από  $N$  παρατηρήσεις  $\mathbf{y} = (y_1, \dots, y_N)^T$  και επαναλαμβάνουμε  $N$  φορές τα παρακάτω βήματα:

- i. Αφαιρούμε την  $i$ -οστή παρατήρηση από το δείγμα (συμβολίζουμε το εναπομείναν δείγμα με  $\mathbf{y}_{-i}$ ).
- ii. Χρησιμοποιώντας το δείγμα  $\mathbf{y}_{-i}$  το οποίο αποτελείται από  $N-1$  παρατηρήσεις υπολογίζουμε την τιμή  $\hat{\theta}_{(i)}$  του εκτιμητή που μας ενδιαφέρει.

Μετά το πέρας της επαναληπτικής αυτής διαδικασίας, έχουμε αποθηκεύσει ένα σύνολο εκτιμητών *jackknife*  $\hat{\theta}_{(jack)} = (\hat{\theta}_{(1)}, \dots, \hat{\theta}_{(N)})^T$ . Τέλος, υπολογίζουμε το τυπικό σφάλμα *jackknife* του εκτιμητή  $\hat{\theta}$  από τον τύπο:

$$se(\hat{\theta}_{(jack)}) = \left[ \frac{N-1}{N} \sum_{i=1}^N (\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2 \right]^{1/2},$$

όπου  $\hat{\theta}_{(.)}$  η μέση τιμή των εκτιμητών *jackknife*, δηλαδή  $\hat{\theta}_{(.)} = n^{-1} \sum_{i=1}^N \hat{\theta}_{(i)}$ .

Να γραφτεί μια συνάρτηση στην R που να υλοποιεί τον αλγόριθμο *jackknife* αν θεωρήσουμε ότι ο εκτιμητής ενδιαφέροντος είναι ο δειγματικός συντελεστής μεταβλητότητας ( $CV = 100 \cdot (S/|\bar{X}|)\%$ , όπου  $S$  είναι η δειγματική τυπική απόκλιση και  $\bar{X}$  ο δειγματικός μέσος), με βάση ένα διάνυσμα παρατηρήσεων  $\mathbf{y}$  που θα δίνει ο χρήστης σαν παράμετρο εισόδου. Η συνάρτηση θα πρέπει να επιστρέφει σε μια λίστα με κατάλληλα ονόματα την τιμή του τυπικού σφάλματος *jackknife* καθώς και τις τιμές των εκτιμητών *jackknife*.

**ΘΕΜΑ 2 (4.5 μονάδες):** Η αλιωτίδα (abalone) είναι θαλάσσιο είδος που ανήκει στα μαλάκια και είναι γνωστή ως το "αφτί της θάλασσας" λόγω του σχήματός της καθώς το όστρακό της είναι πλατύ και ωοειδές. Απολιθωμένα λείψανα αλιωτίδων έχουν βρεθεί σε διάφορα σημεία της Μεσογείου, μεταξύ άλλων και στην περιοχή της διώρυγας της Κορίνθου. Η ηλικία σε έτη μιας αλιωτίδας υπολογίζεται αν στον αριθμό των δακτυλίων, δηλαδή των στρωμάτων που παρατηρούνται σε μια εγκάρσια τομή του όστρακου, προστεθεί ο αριθμός 1.5. Στα πλαίσια μιας έρευνας για τη μοντελοποίηση της ηλικίας των αλιωτίδων, διατίθενται στοιχεία για 4177 αλιωτίδες που περιέχουν, μεταξύ άλλων, τον αριθμό των δακτυλίων (**Rings**), το μέγιστο μήκος σε mm (**LongestShell**), το ύψος σε mm (**Height**) και το γένος **Type** ('F' = θηλυκό, 'M' = αρσενικό, 'I' = έμβρυο χωρίς εκφρασμένο -ακόμα- γένος).

(A) Προσαρμόσαμε το πολλαπλό μοντέλο γραμμικής παλινδρόμησης στην R (και το αποτέλεσμα καταχωρήθηκε σε ένα αντικείμενο με το όνομα **mymodel**) με μεταβλητή απόκρισης την τ.μ. **Rings** και επεξηγηματικές μεταβλητές τις εξής: **LongestShell**, **Height** και **Type**. Με ποια εντολή στην R θα προσαρμόζατε το εν λόγω πολλαπλό γραμμικό μοντέλο;

(B) Με βάση τα παρακάτω αποτελέσματα, ερμηνεύστε τις τιμές των εκτιμητών των συντελεστών του γραμμικού μοντέλου καθώς και την τιμή του συντελεστή προσδιορισμού. Ποιο είναι το τυπικό σφάλμα της παλινδρόμησης; Ποιες είναι οι εικονικές μεταβλητές στο μοντέλο και με ποιες κατηγορίες αναφορές;

Residuals:				
Min	1Q	Median	3Q	Max
-22.8613	-1.6047	-0.5768	0.8415	16.5238
Coefficients:				

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.071273	0.242629	16.78	<2e-16 ***
TypeI	-1.210061	0.118351	-10.22	<2e-16 ***
TypeM	-0.170038	0.097705	-1.74	0.0819 .
LongestShell	0.032125	0.003069	10.47	<2e-16 ***
Height	0.105605	0.008599	12.28	<2e-16 ***
---				
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1				
Residual standard error: 2.585 on 4172 degrees of freedom				
Multiple R-squared: 0.3578, Adjusted R-squared: 0.3572				
F-statistic: 581.1 on 4 and 4172 DF, p-value: < 2.2e-16				

(Γ) Δώστε την εκτιμώμενη ευθεία γραμμικής παλινδρόμησης για κάθε κατηγορία της μεταβλητής Type.

(Δ) Δώστε τον τύπο του συντελεστή προσδιορισμού καθώς και τον τύπο του διορθωμένου συντελεστή προσδιορισμού στο πολλαπλό γραμμικό μοντέλο. Ποια είναι η χρησιμότητα του τελευταίου; Στη συνέχεια, ερμηνεύστε την τιμή του συντελεστή προσδιορισμού στο παραπάνω μοντέλο.

(Ε) Ποιον έλεγχο εξετάζουμε με το F-test στην παραπάνω πολλαπλή γραμμική παλινδρόμηση; Με βάση τα παραπάνω αποτελέσματα του F-test σε τι τελικά συμπεράσματα θα καταλήγατε;

(ΣΤ) Εκτιμήστε σημειακά και με τη βοήθεια ενός 95% διαστήματος εμπιστοσύνης την αναμενόμενη ηλικία σε έτη μιας αλιωτίδας για την οποία γνωρίζουμε ότι έχει μέγιστο μήκος (LongestShell) 130mm, ύψος (Height) 40mm και είναι γένους (Type) θηλυκού. Δώστε την αντίστοιχη εντολή στην R. Ποιο διάστημα εμπιστοσύνης θα χρησιμοποιούσατε για την εν λόγω πρόβλεψη;

### ΘΕΜΑ 3 (3 μονάδες):

(Α) Σε μια άτυπη δημοσκόπηση, πολίτες ερωτήθηκαν κατά πόσο συμφωνούσαν με τους χειρισμούς της κυβέρνησης σε θέματα αγροτικής πολιτικής της χώρας και η απάντησή τους δόθηκε με βάση μια κλίμακα από το 1 (= καθόλου) ως και το 10 (= πάρα πολύ). Την ίδια στιγμή, στα πλαίσια συλλογής δημογραφικών πληροφοριών για το δείγμα, οι συμμετέχοντες δήλωσαν την κατηγορία της οικονομικής τους κατάστασης σε μια κλίμακα από το 1 (= χαμηλή) ως και το 3 (= υψηλή).

- Με ποιον τρόπο και ποιες εντολές στην R θα εκτιμούσατε τον συντελεστή συσχέτισης μεταξύ του οικονομικού επιπέδου (X) και του βαθμού συμφωνίας με τους χειρισμούς της κυβέρνησης σε θέματα αγροτικής πολιτικής της χώρας (Y);
- Πώς μπορείτε να ελέγξετε στην R σε ε.σ. 5% κατά πόσο οι δύο τυχαίες μεταβλητές X και Y είναι ασυσχέτιστες ή όχι;

(Β) Σκοπός μιας έρευνας είναι να ελέγξει κατά πόσο το φύλο επηρεάζει την πρόθεση των φοιτητών να ακολουθήσουν μεταπτυχιακές σπουδές. Συλλέχθηκαν στοιχεία σε τυχαίο δείγμα 23 τελειόφοιτων φοιτητών αποτελούμενο από 12 άνδρες και 11 γυναίκες. Ο παρακάτω πίνακας δίνει τις συχνότητες ανά φύλο σχετικά με την πρόθεση συνέχισης σπουδών σε μεταπτυχιακό επίπεδο, με πιθανές απαντήσεις "Ναι", "Όχι", "Δεν ξέρω".

	Ναι	Όχι	Δεν ξέρω
Άνδρες	4	5	3
Γυναίκες	6	3	2

- Να υπολογιστούν οι αναμενόμενες συχνότητες, κάτω από την υπόθεση ανεξαρτησίας του φύλου και της πρόθεσης συνέχισης σπουδών, με βάση τον παραπάνω πίνακα συχνοτήτων.
- Να ελεγχθεί σε επίπεδο σημαντικότητας 1% αν υπάρχει ανεξαρτησία μεταξύ του φύλου των φοιτητών και της πρόθεσης για μεταπτυχιακές σπουδές (με εναλλακτική ότι υπάρχει εξάρτηση).

Διάρκεια Εξέτασης: 2 ½ ώρες

ΕΥΧΟΜΑΙ ΕΠΙΤΥΧΙΑ