

Ανάλυση Δεδομένων με R/Y

ΘΕΜΑ 1 (2 μονάδες): Δημιουργήστε μια συνάρτηση στην R η οποία θα δέχεται ως όρισμα ένα αριθμητικό διάνυσμα και έναν φυσικό αριθμό (διάφορο του μηδενός) και θα επιστρέψει σε μία λίστα την k μεγαλύτερη καθώς και την k μικρότερη τιμή του διανύσματος. Η επιστρεφόμενη λίστα πρέπει να έχει κατάλληλα ονόματα. Αν το διάνυσμα δεν είναι αριθμητικό η συνάρτηση θα πρέπει να επιστρέψει κατάλληλο μήνυμα λάθους όπως επίσης και αν το μήκος του δοθέντος διανύσματος είναι μικρότερο του k. Τέλος αν το k δεν είναι φυσικός αριθμός ή είναι μηδέν η συνάρτηση και πάλι θα επιστρέψει κατάλληλο μήνυμα λάθους.

ΘΕΜΑ 2 (5 μονάδες): Διαθέτουμε ένα σύνολο ετήσιων μακροοικονομικών δεδομένων που αποτελείται από τις ακόλουθες 6 μεταβλητές: **GNP** (Ακαθάριστο Εθνικό Προϊόν), **Unemployed** (το πλήθος των ανέργων), **Armed.Forces** (το πλήθος των ατόμων στις Ένοπλες Δυνάμεις), **Population** (ο πληθυσμός > 14 ετών εξαιρουμένων των ψυχιατρικά έγκλειστων), **Decade** (40s, 50s, 60s: η δεκαετία στην οποία ανήκει η αντίστοιχη μέτρηση), **Employed** (το πλήθος των εργαζόμενων ατόμων). Σκοπός είναι να διερευνηθεί αν το ετήσιο πλήθος των εργαζόμενων ατόμων 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 **res1**) με μεταβλητή απόκρισης την τ.μ. **Employed** και επεξηγηματικές μεταβλητές τις **GNP**, **Unemployed**, **Armed.Forces**, **Population** και **Decade**. 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 έκτιμητές των συντελεστών του γραμμικού μοντέλου καθώς και την τιμή του συντελεστή προσδιορισμού.

Residuals:

	Min	1Q	Median	3Q	Max
	-0.52907	-0.26913	0.06869	0.23214	0.89748

Coefficients:

	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	68.219717	33.393963	2.043	0.0714 .
GNP	0.054388	0.023310	2.333	0.0445 *
Unemployed	-0.006448	0.004012	-1.607	0.1425
Decade50s	-0.259674	0.659494	-0.394	0.7029
Decade60s	-0.828213	0.878268	-0.943	0.3703
Armed.Forces	-0.005691	0.003465	-1.642	0.1349
Population	-0.171417	0.369945	-0.463	0.6541

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.4863 on 9 degrees of freedom

Multiple R-squared: 0.9885, Adjusted R-squared: 0.9808

F-statistic: 128.9 on 6 and 9 DF, p-value: 3.296e-08

(Γ) Τι διαφορά αναμένουμε στο μέσο πλήθος εργαζόμενων αν το αντίστοιχο GNP διαφέρει κατά 120 μονάδες;

(Δ) Δώστε την εκτιμώμενη ευθεία γραμμικής παλινδρόμησης για κάθε κατηγορία της μεταβλητής **Decade**.

(Ε) Ποιες είναι οι προϋποθέσεις του πολλαπλού γραμμικού μοντέλου που προσαρμόσατε προηγουμένως; Με ποιες εντολές της R μπορείτε να τις ελέγξετε;

(ΣΤ) Ο παρακάτω πίνακας δίνει τις τιμές του δειγματικού συντελεστή συσχέτισης ανάμεσα σε κάθε ζεύγος ποσοτικών επεξηγηματικών μεταβλητών. Τι παρατηρείτε;

	GNP	Unemployed	Armed.Forces	Population
GNP	1.00	0.60	0.45	0.99
Unemployed	0.60	1.00	-0.18	0.69
Armed.Forces	0.45	-0.18	1.00	0.36
Population	0.99	0.69	0.36	1.00

(Ζ) Προσαρμόσαμε ένα μοντέλο γραμμικής παλινδρόμησης στην R (και το αποτέλεσμά του το καταχωρίσαμε σε ένα αντικείμενο με το όνομα **res2**) με μεταβλητή απόκρισης την τ.μ. **Employed** και επεξηγηματικές μεταβλητές τις ίδιες με το ερώτημα (Α) εκτός της τ.μ. **Population**. Ποιες διαφορές διακρίνετε μεταξύ των μοντέλων **res1** και **res2** όσον αφορά τη στατιστική σημαντικότητα των επεξηγηματικών μεταβλητών; Λαμβάνοντας υπόψη τις παρατηρήσεις σας από το ερώτημα (ΣΤ), ποιο μοντέλο θα ήταν πιο πιθανό να επιλέξετε εκ των **res1** και **res2** ως πιο αξιόπιστο και γιατί;

Residuals:					
	Min	1Q	Median	3Q	Max
	-0.64023	-0.23012	0.04574	0.20069	0.90051
Coefficients:					
	Estimate	Std. Error	t value	Pr(> t)	
(Intercept)	52.752724	0.923043	57.151	6.53e-14 ***	
GNP	0.043680	0.002929	14.910	3.70e-08 ***	
Unemployed	-0.008006	0.002103	-3.807	0.00345 **	
Decade50s	-0.140515	0.582956	-0.241	0.81440	
Decade60s	-0.902979	0.828727	-1.090	0.30145	
Armed.Forces	-0.005972	0.003275	-1.823	0.09825	.

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1					
Residual standard error: 0.4668 on 10 degrees of freedom					
Multiple R-squared: 0.9882, Adjusted R-squared: 0.9823					
F-statistic: 167.8 on 5 and 10 DF, p-value: 2.621e-09					

ΘΕΜΑ 3 (3 μονάδες): Τα παρακάτω δεδομένα αφορούν την τιμή πώλησης (σε δολάρια) μιας φιάλης κρασιού Σαντορίνης σε 10 διαφορετικά καταστήματα για δύο διαφορετικές χρονιές.

2009	4.65	4.55	4.11	4.15	4.20	4.55	3.80	4.00	4.19	4.75
2016	4.73	5.29	4.89	4.95	4.25	4.90	5.15	5.30	4.29	4.95

Θέλουμε να ελέγξουμε, σε ε.σ. 3%, την υπόθεση ότι η μέση τιμή πώλησης της συγκεκριμένης μάρκας κρασιού δεν αυξήθηκε το 2016 σε σχέση με το 2009 με εναλλακτική ότι αυξήθηκε.

- Τι είδους στατιστική ανάλυση θα εφαρμόζατε και με ποιες εντολές στην R;
- Ποιες προϋποθέσεις θα ελέγχατε για την στατιστική ανάλυση του παραπάνω υποερωτήματος και με ποιες εντολές στην R;
- Αν δεν ισχυαν οι παραπάνω προϋποθέσεις ποιον έλεγχο θα εφαρμόζατε και με ποια εντολή στην R;