ΣΧΟΛΗ: Ε.Μ.Φ.Ε. **Ιανουάριος 2016**
ΕΞΑΜΗΝΟ: 5ο

Ανάλυση Δεδομένων με Η/Υ

ΘΕΜΑ 1 (2.0 μονάδες):

Γράψτε μια συνάρτηση στην R η οποία θα παίρνει ως όρισμα ένα διάνυσμα θετικών παρατηρήσεων X και θα επιστρέφει την ποσότητα

$$d = \sum_{i=1}^{n} c_i X_{(i)}, \text{ όπου } c_i = \left(\frac{n-i}{n}\right)^{1/2} - \left(\frac{n-i+1}{n}\right)^{1/2} - \frac{1}{n},$$

n είναι το πλήθος των παρατηρήσεων και $X_{(i)}$ είναι η i διατεταγμένη παρατήρηση, δηλαδή $X_{(1)}$ είναι η μικρότερη και $X_{(n)}$ η μεγαλύτερη παρατήρηση αντίστοιχα.

Προσοχή: Μη θετικές παρατηρήσεις θα πρέπει να αγνοηθούν για τους υπολογισμούς, δηλαδή το πλήθος των παρατηρήσεων θα μειωθεί.

ΘΕΜΑ 2 (5.5 μονάδες):

Με βάση μια ευρωπαϊκή έρευνα, διαθέτουμε στοιχεία για το μερίδιο προϋπολογισμού το οποίο 150 νοικοκυριά στην Ισπανία ξόδεψαν σε τρόφιμα κατά το έτος 1980. Συγκεκριμένα, τα δεδομένα περιέχουν το ποσοστό των συνολικών ετήσιων δαπανών που το κάθε νοικοκυριό ξόδεψε σε τρόφιμα (**wfood**, %), τη συνολική ετήσια δαπάνη του νοικοκυριού (**totexp**, σε χιλιάδες ισπανικές πεσέτες), το πλήθος των ατόμων που απαρτίζουν το νοικοκυριό (**size**), την ηλικία του προσώπου αναφοράς στο νοικοκυριό (**age**, σε έτη) και το φύλο του προσώπου αναφοράς (**gender**: 1 = Γυναίκα, 2 = Άνδρας).

(Α) 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. **wfood** και επεξηγηματικές μεταβλητές τις εξής: **totexp**, **size**, **age** και **gender**. Με ποια εντολή στην R θα προσαρμόζατε το εν λόγω πολλαπλό γραμμικό μοντέλο;

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

```
Residuals:
   Min       1Q        Median      3Q       Max
-0.39761  -0.08366  -0.00299  0.07726  0.37119

Coefficients:
                Estimate   Std. Error  t value   Pr(>|t|)
(Intercept)   3.829e-01  6.028e-02  6.353   2.57e-09 ***
totexp       -1.718e-04  1.917e-05 -8.963   1.44e-15 ***
size          1.700e-02  6.500e-03  2.615   0.00986  **
age           8.101e-04  7.966e-04  1.017   0.31087
gender2       4.619e-02  3.325e-02  1.389   0.16688
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.1293 on 145 degrees of freedom
Multiple R-squared:  0.3785,    Adjusted R-squared:  0.3614
F-statistic: 22.08 on 4 and 145 DF,  p-value: 2.996e-14
```

(Γ) Τι διαφορά αναμένουμε στο **ποσοστό ετήσιων δαπανών για τρόφιμα** ανάμεσα σε δύο νοικοκυριά ίδιου μεγέθους με άτομα αναφοράς ίδιου φύλου και ηλικίας αν η συνολική ετήσια δαπάνη τους διαφέρει κατά 50000 πεσέτες;

(Δ) Δώστε την εκτιμώμενη ευθεία γραμμικής παλινδρόμησης για κάθε κατηγορία της μεταβλητής **gender**.
(Ε) Ποια είναι η διαφορά μεταξύ του συντελεστή προσδιορισμού και του διορθωμένου συντελεστή προσδιορισμού; Στη συνέχεια, ερμηνεύστε την τιμή του συντελεστή προσδιορισμού στο παραπάνω μοντέλο.
(ΣΤ) 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;
(Ζ) Εκτιμήστε σημειακά και με τη βοήθεια ενός 95% διαστήματος εμπιστοσύνης το **ποσοστό ετήσιων δαπανών που ένα συγκεκριμένο νοικοκυριό αναμένεται να ξοδέψει σε τρόφιμα** για το οποίο γνωρίζουμε ότι απαρτίζεται από 6 άτομα (**size**), το άτομο αναφοράς είναι ένας άνδρας (**gender**) 50 ετών (**age**) και η συνολική ετήσια δαπάνη του (**totexp**) είναι ίση με 750[redacted] χιλιάδες πεσέτες. Δώστε την αντίστοιχη εντολή στην R. Ποιο διάστημα εμπιστοσύνης θα χρησιμοποιούσατε για την εν λόγω πρόβλεψη;
(Η) Ποιους ελέγχους υποθέσεων μας ενδιαφέρει να κατασκευάσουμε για κάθε έναν ξεχωριστά συντελεστή του παραπάνω μοντέλου; Με τη βοήθεια του παραπάνω πίνακα σχολιάστε τα αποτελέσματα των εν λόγω ελέγχων. Πώς με τη βοήθεια της R θα δημιουργούσατε 95% διαστήματα εμπιστοσύνης για τους εν λόγω συντελεστές;

ΘΕΜΑ 3 (2.5 μονάδες):

Στην απογραφή μιας κωμόπολης το 2011, κάθε συμμετέχων ταξινομήθηκε ανάλογα με την κατάσταση απασχόλησής του {εργαζόμενος, άνεργος, εκτός εργατικού δυναμικού} και ανάλογα με την οικογενειακή του κατάσταση {έγγαμος, διαζευγμένος/χήρος, άγαμος}. Τα παρακάτω καταγεγραμμένα αποτελέσματα αφορούν 1074 άνδρες 35-44 ετών. Ο σκοπός είναι να ερευνήσουμε κατά πόσο η κατάσταση απασχόλησης εξαρτάται από την οικογενειακή κατάσταση ενός άνδρα της συγκεκριμένης ηλικιακής κατηγορίας.

	Έγγαμος	Διαζευγμένος/Χήρος	Άγαμος
Εργαζόμενος	679	103	114
Άνεργος	63	10	20
Εκτός εργατικού δυναμικού	42	18	25

i) Ποιον έλεγχο θα εφαρμόζατε για να απαντήσετε στο παραπάνω ερευνητικό ερώτημα και ποιες είναι οι προϋποθέσεις του; Αν δεν ισχύουν οι προϋποθέσεις, ποιος είναι ο έλεγχος που θα εφαρμόζατε;
ii) Να υπολογιστούν οι αναμενόμενες συχνότητες, κάτω από την υπόθεση ανεξαρτησίας της κατάστασης απασχόλησης και της οικογενειακής κατάστασης, με βάση τον παραπάνω πίνακα συχνοτήτων.
iii) Με ποιες εντολές στην R θα ελέγχατε αν υπάρχει ανεξαρτησία σε επίπεδο σημαντικότητας 2% μεταξύ της κατάστασης απασχόλησης των ανδρών 35-44 ετών και της οικογενειακής τους κατάστασης (με εναλλακτική ότι υπάρχει εξάρτηση);

Διάρκεια Εξέτασης: 2 ½ ώρες ***ΕΥΧΟΜΑΙ ΕΠΙΤΥΧΙΑ***