ΣΧΟΛΗ: Ε.Μ.Φ.Ε. **Φεβρουάριος 2011**
ΕΞΑΜΗΝΟ: 5ο

Ανάλυση Δεδομένων με Η/Υ

ΘΕΜΑ 1 (3 μονάδες):
(Α) Η συνάρτηση πυκνότητας πιθανότητας της κατανομής Γάμμα είναι η

$$f(x) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, \; x > 0, \; a, b > 0.$$

Θεωρήστε ότι a = 2. Χωρίς να κάνετε χρήση της εντολής dgamma, να γράψετε μια δική σας συνάρτηση στην R, η οποία θα υπολογίζει την τιμή της f(x) ή της log(f(x)) για συγκεκριμένο b και για διάφορες τιμές του x. Η συνάρτηση που θα γράψετε θα πρέπει ως παραμέτρους εισόδου να δίνει το b τις τιμές του x (υπό μορφή διανύσματος) για τις οποίες θέλετε να γίνει ο υπολογισμός και το αν θέλετε να υπολογίσετε την f(x) ή την log(f(x)), με προκαθορισμένη (default) τιμή την f(x). Η συνάρτηση θα πρέπει επιπλέον να επιστρέφει ένα μήνυμα λάθους αν $b \leq 0$ ή αν τα στοιχεία του διανύσματος x δεν είναι όλα θετικά. Δίνεται ότι η τιμή της συνάρτησης Γ(a) στην R υπολογίζεται από την εντολή gamma(a) και ο λογάριθμος αυτής από την εντολή lgamma(a).

(Β) Δημιουργήστε στην R μια ακολουθία 10000 τιμών για την παράμετρο b από το 0.001 έως το 10. Να γράψετε μια συνάρτηση στην R, κάνοντας χρήση της συναρτήσεως του ερωτήματος (Α), η οποία για γνωστά δεδομένα θα απεικονίζει σε ένα διάγραμμα, για τις διάφορες τιμές της άγνωστης παραμέτρου b που δημιουργήσατε, την λογαριθμική πιθανοφάνεια της κατανομής Γάμμα για a = 2 και θα υπολογίζει προσεγγιστικά το b εκείνο που αντιστοιχεί στο μέγιστο της εν λόγω συνάρτησης (δηλ. την εκτιμήτρια μέγιστης πιθανοφάνειας του b).

ΘΕΜΑ 2 (5 μονάδες):
Μια ερευνητική ομάδα ενδιαφέρθηκε να μελετήσει τα επίπεδα λίπους στο σώμα ενήλικων ατόμων. Σε τυχαίο δείγμα 18 ατόμων, ηλικίας από 23 έως 61 ετών, συλλέχθηκαν οι παρακάτω μεταβλητές:

- Age : Ηλικία (σε χρόνια).
- Percent.Fat : Ποσοστό λίπους σώματος (%).
- Gender : Φύλο (0 = γυναίκα / 1 = άνδρας).
- Exercise : Ώρες εκγύμνασης την εβδομάδα.

(Α)

i) Έστω ότι τα δεδομένα, βρίσκονται υπό μορφή ενός πίνακα 18×4 στο αρχείο "data.txt". Με ποιον τρόπο μπορείτε να εισάγετε τα δεδομένα στην R και να δημιουργήσετε ένα πλαίσιο δεδομένων δίνοντας ονόματα στις 4 μεταβλητές, όπως αυτά που δόθηκαν στην αρχή της εκφώνησης;

ii) Με ποιους τρόπους (αριθμητικούς και γραφικούς) και με ποιες εντολές στην R θα περιγράφατε τις τιμές του τυχαίου δείγματος για όλες τις μεταβλητές;

(Β)
Χρησιμοποιώντας τα παραπάνω δεδομένα θέλουμε να εξετάσουμε αν το επίπεδο λίπους στο σώμα εξαρτάται από την ηλικία, το φύλο και την εβδομαδιαία άσκηση του ατόμου, προσαρμόζοντας το κάτωθι γενικό γραμμικό μοντέλο:

$$E[\text{Percent.Fat} \mid \text{Age, Gender, Exercise}] = a + \beta_1 \text{ Age} + \beta_2 \text{ Gender} + \beta_3 \text{ Exercise}.$$

i) Με ποια εντολή στην R θα προσαρμόζατε το παραπάνω γενικό γραμμικό μοντέλο;

ii) Ποιες είναι οι εικονικές μεταβλητές στο παραπάνω μοντέλο και με ποιες κατηγορίες αναφοράς;
iii) Ποιες προϋποθέσεις θα ελέγχατε για το παραπάνω γενικό γραμμικό μοντέλο και με ποιες εντολές στην R;
iv) Εξηγήστε πλήρως τα παρακάτω αποτελέσματα που παίρνετε από την R ύστερα από την προσαρμογή του παραπάνω γενικού γραμμικού μοντέλου.

```
Residuals:
   Min    1Q Median    3Q    Max
-6.580 -3.667 -1.187 3.703 9.446

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 13.37952   8.35721   1.601   0.1317
Age          0.36178   0.14259   2.537   0.0237 *
Gender1     -9.84677   3.81684  -2.580   0.0218 *
Exercise     0.09617   0.30424   0.316   0.7566
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.059 on 14 degrees of freedom
Multiple R-squared: 0.7479,    Adjusted R-squared: 0.6939
F-statistic: 13.85 on 3 and 14 DF,  p-value: 0.0001793
```

v) Με βάση τα παραπάνω αποτελέσματα ερμηνεύστε τους εκτιμητές των συντελεστών του γενικού γραμμικού μοντέλου.
vi) Εκτιμήστε το μέσο ποσοστό σωματικού λίπους (%) μιας γυναίκας ηλικίας 56 ετών που γυμνάζεται 2 ώρες την εβδομάδα.
vii) Συνοψίστε τα συμπεράσματα της παραπάνω ανάλυσης. ~ 3,4 γραμμές

ΘΕΜΑ 3 (2 μονάδες):
Τα παρακάτω δεδομένα αφορούν την τιμή πώλησης (σε δολάρια) μιας συγκεκριμένης μάρκας κρασιού σε 10 διαφορετικά καταστήματα για δύο διαφορετικές χρονιές.

2004	4.65	4.55	4.11	4.15	4.20	4.55	3.80	4.00	4.19	4.75
2007	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. 5%, την υπόθεση ότι η μέση τιμή πώλησης της συγκεκριμένης μάρκας κρασιού δεν αυξήθηκε το 2007 σε σχέση με το 2004 με εναλλακτική ότι αυξήθηκε.
i) Τι είδους στατιστική ανάλυση θα εφαρμόζατε και με ποιες εντολές στην R;
ii) Ποιες προϋποθέσεις θα ελέγχατε για την στατιστική ανάλυση του παραπάνω υποερωτήματος και με ποιες εντολές στην R;
iii) Αν δεν ίσχυαν οι παραπάνω προϋποθέσεις ποιον έλεγχο θα εφαρμόζατε και με ποια εντολή στην R;

Διάρκεια Εξέτασης: 2 ½ h

ΕΥΧΟΜΑΣΤΕ ΕΠΙΤΥΧΙΑ!