ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

Ανάλυση Δεδομένων με Η/Υ- 28 Σεπτεμβρίου 2012

Άσκηση 1 (35 μονάδες)

Στην κανονική εξεταστική προσήλθαν στις εξετάσεις του μαθήματος 66 φοιτητές. Για κάθε φοιτητή που προσήλθε στις εξετάσεις καταγράφηκε ο αριθμός μητρώου, ο βαθμός του σε καθεμιά από τις 3 εργασίες που είχαν ανατεθεί, και ο βαθμός του στην τελική εξέταση. Οι βαθμοί στις εργασίες που παραδόθηκαν και στην τελική εξέταση είναι ακέραιοι αριθμοί από το 1 ως το 100. Αν ένας φοιτητής δεν είχε παραδώσει κάποια εργασία τότε στη θέση του αντίστοιχου βαθμού μπήκε το σύμβολο *. Τα δεδομένα αυτά βρίσκονται σε μορφή πίνακα 66×5 στο αρχείο `grades.txt`.

α) Με ποιον τρόπο θα εισαγάγετε τα δεδομένα στην R και θα τα αποθηκεύσετε σε ένα πλαίσιο δεδομένων gframe, αντικαθιστώντας το σύμβολο * για τους ελλείποντες βαθμούς με τον αριθμό 0, και δίνοντας στις μεταβλητές τα ονόματα am για τον αριθμό μητρώου, h1, h2, h3 για τους βαθμούς των εργασιών, exam για τον βαθμό της εξέτασης;

Ορίζουμε τις μεταβλητές final, class, hw ως εξής. Ο τελικός βαθμός final υπολογίζεται από τον τύπο

$$\text{final} = 0.7 \times \text{exam} + 0.1 \times (\text{ h1} \vee \text{exam}) + 0.1 \times (\text{h2} \vee \text{exam}) + 0.1 \times (\text{h3} \vee \text{exam})$$

όπου $x \vee y = \max\{x, y\}$. Η class παίρνει την τιμή F αν ο τελικός βαθμός είναι μικρότερος του 45 και την τιμή P διαφορετικά. Τέλος, η hw παίρνει την τιμή 1 αν ο φοιτητής έχει παραδώσει τουλάχιστον μια εργασία και την τιμή 0 διαφορετικά.

β) Φτιάξτε μια συνάρτηση στην R που θα δέχεται σαν είσοδο ένα διάνυσμα βαθμών (h1, h2, h3, exam) και θα επιστρέφει σαν έξοδο τη λίστα (final, class, hw).

γ) Φτιάξτε μια ρουτίνα στην R που θα επισυνάπτει στο πλαίσιο gframe τις μεταβλητές final, class και hw.

δ) Με ποια εντολή της R μπορείτε να παρουσιάσετε σε μια εικόνα τα θηκογραφήματα για τον τελικό βαθμό εκείνων που δεν παρέδωσαν καμμία εργασία και εκείνων που παρέδωσαν τουλάχιστον μία εργασία, δίνοντας τους τα ονόματα "ΧΩΡΙΣ" και "ΜΕ" αντίστοιχα;

```
> t<-table(gframe$class,gframe$hw)
> margin.table(t,1)
 F  P
28 38
> margin.table(t,2)
 0  1
30 36
> sum(gframe$hw[gframe$class=="P"])
25
```

ε) Στο διπλανό πλαίσιο εμφανίζονται μια σειρά από εντολές που δώσαμε στην R και τα αντίστοιχα αποτελέσματα που μας επέστρεψε. Βάσει των δεδομένων αυτών, ποιος είναι ο πίνακας συνάφειας των μεταβλητών class, hw;

στ) Σχεδιάστε με βάση τα διπλανά δεδομένα ποιο θα είναι το αποτέλεσμα που θα πάρουμε αν εκτελέσουμε την εντολή

```
> barplot(t,col=c("grey","white"),legend=c("F","P"))
```

και εξηγήστε πολύ σύντομα πώς θα περιγράφατε το αποτέλεσμα.

Άσκηση 2 (30 μονάδες)

Ένας οινοπαραγωγός θέλει να ελέγξει αν η χρήση ενός νέου τύπου λιπάσματος ενισχύει την απόδοση των αμπελιών του. Έχει 20 αμπελώνες σε διαφορετικά σημεία της χώρας που όλοι παράγουν την ίδια ποικιλία σταφυλιού. Εξετάζει δύο διαφορετικές μεθόδους: 1) να καλλιεργήσει 10 αμπελώνες με τον παλιό τύπο λιπάσματος και 10 με τον νέο ή 2) να χωρίσει κάθε αμπελώνα στα δύο και να καλλιεργήσει κάθε κομμάτι με διαφορετικό τύπο λιπάσματος. Σε κάθε περίπτωση σκέφτεται να καταγράψει τις αποδόσεις σε kg/ρίζα και να τις συγκρίνει.

α) Ποια στατιστική ανάλυση θα επιλέγατε για να συγκρίνετε τους δύο τύπους λιπάσματος για καθεμιά από τις δύο παραπάνω μεθόδους, και με ποιες εντολές στην R θα την υλοποιούσατε;

β) Θα προτείνατε μια από τις δύο μεθόδους ως καταλληλότερη και γιατί;

Άσκηση 3 (35 μονάδες)

Σε τυχαίο δείγμα 27 εργαζομένων σε μία εταιρεία πληροφορικής συλλέχθηκαν οι παρακάτω μεταβλητές:

- Y : Μηνιαίο καθαρό εισόδημα σε €.
- X_1: Επίπεδο Εκπαίδευσης (X_{11} = Απόφοιτος Λυκείου, X_{12} = Απόφοιτος ΑΕΙ, X_{13} = Μεταπτυχιακές Σπουδές).
- X_2: Έτη προϋπηρεσίας (X_{21}: ≤ 5 έτη, X_{22}: Από 5 μέχρι και 10 έτη, X_{23} : > 10 έτη).

Χρησιμοποιώντας τα παραπάνω δεδομένα θέλουμε να εξετάσουμε αν το καθαρό μηνιαίο εισόδημα των εργαζομένων στην εν λόγω εταιρία εξαρτάται από το επίπεδο μόρφωσης και τα έτη προϋπηρεσίας. Προσαρμόζουμε το κάτωθι γενικό γραμμικό μοντέλο:

$$E[Y \mid X_1,X_2]=\alpha+\beta_1X_{12}+\beta_2X_{13}+\beta_3X_{22}+\beta_4X_{23}$$

i) Με ποια εντολή στην R θα προσαρμόζατε το παραπάνω γενικό γραμμικό μοντέλο;
ii) Εξηγήστε πλήρως τα παρακάτω αποτελέσματα που παίρνετε από την R ύστερα από την προσαρμογή του εν λόγω γενικού γραμμικού μοντέλου.

```
Residuals:
    Min      1Q  Median      3Q     Max
-81.111 -16.667  -2.222  17.222  54.444

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    726.67      14.80  49.099  < 2e-16 ***
X12            107.78      16.21   6.648 1.10e-06 ***
X13            195.56      16.21  12.062 3.60e-11 ***
X22             58.89      16.21   3.632  0.00147 **
X23            184.44      16.21  11.377 1.10e-10 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34.39 on 22 degrees of freedom
Multiple R-squared: 0.9274,     Adjusted R-squared: 0.9142
F-statistic: 70.26 on 4 and 22 DF,  p-value: 3.305e-12
```

iii) Ποιες είναι οι εικονικές μεταβλητές στο παραπάνω μοντέλο και με ποιες κατηγορίες αναφοράς;
iv) Με βάση τα παραπάνω αποτελέσματα ερμηνεύστε τους εκτιμητές των συντελεστών του γενικού γραμμικού μοντέλου.
v) Εκτιμήστε το καθαρό μηνιαίο εισόδημα ενός εργαζόμενου στην εν λόγω εταιρεία, που είναι απόφοιτος Λυκείου και έχει 8 έτη προϋπηρεσίας.

Διάρκεια Εξέτασης 2 ώρες και 30 λεπτά

ΚΑΛΗ ΕΠΙΤΥΧΙΑ!