ΣΧΟΛΗ: Ε.Μ.Φ.Ε. **Ιούνιος 2017**

ΕΞΑΜΗΝΟ: 6ο

→fact

$\binom{p-1}{j} = \frac{(p-1)!}{j!(p-1-j)!}$

Ανάλυση Δεδομένων με Η/Υ input: διάνυσμα p ακεραίων

ΘΕΜΑ 1 (2 μονάδες): Δίνεται η παρακάτω συνάρτηση: $f(p) = \frac{\pi^2}{6} + p \cdot w_p - p^2 \cdot c_p$, όπου

$w_p = \sum_{j=0}^{p-1} (-1)^j \cdot \binom{p-1}{j} \cdot \frac{\log(1+j)}{1+j}$ και $c_p = \sum_{j=0}^{p-1} (-1)^j \cdot \binom{p-1}{j} \cdot \frac{[\log(1+j)]^2}{1+j}$. Με τη βοήθεια της R να γράψετε μια συνάρτηση που θα υπολογίζει τις τιμές της f για διάφορες τιμές του p. Η συνάρτηση θα δέχεται ως όρισμα ένα διάνυσμα ακεραίων τιμών p και θα επιστρέφει το διάνυσμα τιμών της f. Εν συνεχεία χρησιμοποιώντας τη συνάρτηση που δημιουργήσατε δώστε την εντολή της R που θα έδινε το γράφημά της για τιμές του p από 2 μέχρι και 100. plot(funct(p[2:100]))

→log10

ΘΕΜΑ 2 (5.0 μονάδες): Στα πλαίσια της έρευνας μιας φοιτητικής κινηματογραφικής λέσχης της Αθήνας, επιλέχθηκαν 50 ταινίες από την Ελλάδα και το εξωτερικό (τ.μ. **Type**, όπου Gr: ελληνική ταινία και Int: διεθνής ταινία) και καταγράφηκε το σκορ κριτικής τους από το 1 ως το 10 (τ.μ. **imdb**) όπως εμφανίζεται στον γνωστό ιστότοπο κριτικής κινηματογραφικών ταινιών *imdb.com*. Εν συνεχεία, τριάντα άτομα από τα μέλη της λέσχης, αφού παρακολούθησαν τις ταινίες, τις βαθμολόγησαν σε μια κλίμακα από το 1 ως το 100. Εν τέλει, για κάθε ταινία προέκυψε η μέση βαθμολογία που έδωσαν τα τριάντα μέλη της λέσχης που συμμετείχαν στην έρευνα (τ.μ. **meanRatingAfterFilm**). Ο σκοπός ήταν να αξιολογήσουν κατά πόσο τα σκορ του δημοφιλούς ιστότοπου imdb.com ήταν συμβατά με τη δική τους αξιολόγηση των ταινιών.

(Α) Προσαρμόσαμε το μοντέλο γραμμικής παλινδρόμησης στην R (και το αποτέλεσμά του το καταχωρήσαμε σε ένα αντικείμενο με το όνομα **res1**) με μεταβλητή απόκρισης την **meanRatingAfterFilm** και επεξηγηματικές μεταβλητές τις **imdb** και **Type**. 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;

(Β) Με βάση τα παρακάτω αποτελέσματα ερμηνεύστε τους εκτιμητές των συντελεστών του γραμμικού μοντέλου καθώς και την τιμή του συντελεστή προσδιορισμού. R^2: τί σημαίνει

```
Residuals:
   Min     1Q  Median     3Q    Max
-40.810 -12.416   0.143  10.632  29.315

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept) -110.686     20.415  -5.422 2.00e-06 ***
imdb          25.052      2.754   9.098 6.18e-12 ***
typeInt        4.672      6.359   0.735    0.466
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 18.36 on 47 degrees of freedom
Multiple R-squared:  0.6396,Adjusted R-squared:  0.6243
F-statistic: 41.71 on 2 and 47 DF,  p-value: 3.836e-11
```

(Γ) Τι διαφορά αναμένουμε στην μέση αξιολόγηση των μελών της λέσχης ανάμεσα σε δύο ελληνικές ταινίες αν το αντίστοιχο imdb σκορ διαφέρει κατά 3 μονάδες;

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

(Ε) Στα πλαίσια της πρόβλεψης της μέσης βαθμολογίας μιας ταινίας με βάση το παραπάνω μοντέλο γραμμικής παλινδρόμησης, τρέξαμε την κάτωθι εντολή. Εξηγήστε πλήρως τον κώδικα που παρατίθεται και ερμηνεύστε τα αποτελέσματα.

εσ. ΔΕ

```
> predict(res1, list(imdb=7.7, type='Int'), int="p", level=0.90 )
     fit        lwr        upr
  86.88241   55.37372   118.3911
```

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

κανονικ. residuals residuals(mymodel,"partial",1:2)

(ΣΤ) Έστω ότι ο δειγματικός μέσος της τ.μ. **imdb** είναι 6.83. Αφού κεντράρουμε τα δεδομένα της τ.μ. **imdb** (**imdbCentered**) και ξαναπροσαρμόσουμε το ίδιο πολλαπλό γραμμικό μοντέλο, η σταθερά του μοντέλου έχει τώρα την τιμή 60.516. Δώστε την ερμηνεία της εν λόγω τιμής. Ποιες θα είναι τώρα οι τιμές των άλλων δύο συντελεστών του μοντέλου;

x_1 και x_2

N(0,4)

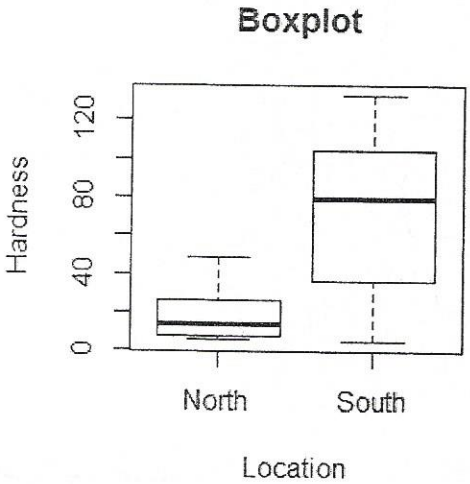
ΘΕΜΑ 3 (3 μονάδες): Στα πλαίσια μιας περιβαλλοντικής μελέτης, εξετάστηκε κατά πόσο η υψηλή περιεκτικότητα του πόσιμου νερού σε ασβέστιο (*σκληρότητα*) σχετίζεται με τη γεωγραφική θέση των πόλεων. Τα δεδομένα αφορούν 19 Βρετανικές πόλεις με καταγεγραμμένες μεταβλητές τη σκληρότητα (*hardness*) του νερού (σε *ppt: parts per million*) και μια δίτιμη μεταβλητή γεωγραφικής θέσης (*location*) της εκάστοτε πόλης (στον βορρά - *north* ή στον νότο - *south*). Για να ελέγξουμε κατά πόσο το μέσο επίπεδο σκληρότητας του νερού διαφοροποιείται μεταξύ των βόρειων και νότιων πόλεων της Βρετανίας; εφαρμόσαμε τρεις διαφορετικούς ελέγχους υποθέσεων. Ποιοι είναι οι εν λόγω έλεγχοι και ποιες οι διαφορές τους; Ποιες είναι οι προϋποθέσεις των εν λόγω ελέγχων; Καλείστε να ερμηνεύσετε πλήρως τα ακόλουθα αποτελέσματα των τριών αυτών ελέγχων.

προϋποθέσεις, υποθέσεις, τί δέχουμε να ελέγξουμε

Two Sample t-test	Welch Two Sample t-test	Wilcoxon rank sum test with continuity correction
Data: hardness by location t = -3.6377, df = 17, p-value = 0.002035 95 percent confidence interval: -83.07487 -22.08423 sample estimates: mean in group North mean in group South 19.54545 72.12500	Data: hardness by location t = -3.1789, df = 8.168, p-value = 0.01267 95 percent confidence interval: -90.58511 -14.57398 sample estimates: mean in group North mean in group South 19.54545 72.12500	Data: hardness by location W = 16, p-value = 0.0228

2 δαφ. ειν. ε. μη παραμετρ.

var.equal=true
var.equal=False



Με ποια εντολή της R θα δημιουργούσατε το αριστερό γράφημα; Λαμβάνοντας υπόψη το εν λόγω γράφημα, ποιος έλεγχος εκ των τριών που προαναφέρθηκαν φαίνεται να είναι καταλληλότερος για το εν λόγω ερευνητικό πρόβλημα και γιατί; Ποια είναι τα τελικά σας συμπεράσματα;

boxplot(hardness ~ location)

→ Ποιος απορ. σαν 2°.
→ τελικό συμπ.

Διάρκεια Εξέτασης: 2 ½ ώρες

ΕΥΧΟΜΑΙ ΕΠΙΤΥΧΙΑ