Ε Θ Ν Ι Κ Ο Μ Ε Τ Σ Ο Β Ι Ο Π Ο Λ Υ Τ Ε Χ Ν Ε Ι Ο
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ & ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
Τομέας Μαθηματικών
Πολυτεχνειούπολη – Ζωγράφου ΑΘΗΝΑ - 157 80
ΤΗΛ. : 772 1774
FAX : 772 1775

ΜΑΘΗΜΑ: *Ανάλυση Δεδομένων με Η/Υ (6ο εξάμηνο)*
ΔΙΔΑΣΚΩΝ: *Δημήτρης Φουσκάκης*

ΕΡΓΑΣΙΑ 3η

Θέμα Εργασίας: Ανάλυση Παλινδρόμησης

Άσκηση 1

Μια ερευνήτρια θέλει να εξετάσει τους παράγοντες που επηρεάζουν τον χρόνο που περνούν οι ενήλικες στο τηλέφωνό τους ανά ημέρα. Μέσω ενός ερωτηματολογίου, συλλέχθηκαν πληροφορίες από 70 διαφορετικούς ενήλικες. Οι μεταβλητές που συλλέχθηκαν στο δείγμα για κάθε ενήλικα είναι οι εξής: (α) η ηλικία (**Age**) σε έτη, (β) το φύλο **(Gender)** οπού παρατηρήθηκαν οι κλάσεις "M" (αρσενικό) και "F" (θηλυκό), (γ) το λογισμικό που χρησιμοποιούν **(Operating_System)** στο κινητό τους τηλέφωνο, οπού παρατηρήθηκαν οι κλάσεις "IOS" και "Android", (δ) ο τύπος εργασίας τους **(Occupation)** χωρισμένος σε κλάσεις "Self_Employed" (αυτοαπασχολούμενος), "Wage Worker" (μισθωτός) και "Other" (άλλο), (ε) η ηλικία του κινητού τους σε ημέρες **(Phone_Age_in_Days)**, (στ) ο χρόνος σε ώρες που περάσαν στο κινητό για λόγους ευχαρίστησης (μη εργασιακούς) **(Screen_Time_Leasure)** την προηγούμενη μέρα, (ζ) ο χρόνος σε ώρες που περάσαν στο κινητό για λόγους εργασιακούς **(Screen_Time_Business)** την προηγούμενη μέρα.

Τα δεδομένα βρίσκονται στο αρχείο:

http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/Phone.txt

Στα δεδομένα μας, με τον χαρακτήρα "$" συμβολίζουμε τις αγνοούμενες τιμές. Εισάγετε τα δεδομένα στην R με χρήση της εντολής `read.table`, αλλάζοντας κατάλληλα το σύμβολο για τις αγνοούμενες τιμές (για αυτόματη αλλαγή, ελέγξτε στο

μενού `help` της R τα δυνατά ορίσματα της εντολής `read.table`). Ποια είναι η δομή του αντικειμένου που δημιουργείται από την παραπάνω εντολή; Στη συνέχεια αφαιρέστε οποιαδήποτε γραμμή περιέχει αγνοούμενη τιμή. Επιπλέον αφαιρέστε, αν υπάρχει, οποιαδήποτε γραμμή περιέχει τιμές από μη ενήλικα άτομα (ηλικία μικρότερη των 18 ετών).

i. Εκτιμήστε σημειακά τους συντελεστές συσχέτισης όλων των δυνατών (ανά δύο) συνδυασμών των μεταβλητών **Age**, **Screen_Time_Leasure, Screen_Time_Business** και **Phone_Age_in_Days**. Εξετάστε αν οι συντελεστές συσχέτισης είναι στατιστικά σημαντικοί για κάθε συνδυασμό.
ii. Προσαρμόστε το απλό γραμμικό μοντέλο με μεταβλητή απόκρισης την τ.μ. **Screen_Time_Business** και επεξηγηματική μεταβλητή την τ.μ. **Phone_Age_in_Days**. Δώστε το γράφημα διασποράς των παρατηρήσεων μαζί με την ευθεία ελαχίστων τετραγώνων και σχολιάστε. Ερμηνεύστε τους εκτιμητές των συντελεστών του παραπάνω απλού γραμμικού μοντέλου. Ερμηνεύστε τέλος την τιμή του συντελεστή προσδιορισμού και του διορθωμένου συντελεστή προσδιορισμού για το εν λόγω μοντέλο. Συγκρίνετε τα αποτελέσματα σας με τα αποτελέσματα του ερωτήματος i.
iii. Ονομάστε και εν συντομία εξηγήστε τις προϋποθέσεις ενός απλού γραμμικού μοντέλου. Εν συνεχεία με γραφικούς τρόπους ελέγξτε τις για το γραμμικό μοντέλο που προσαρμόσατε παραπάνω. Σχολιάστε λεπτομερώς τα ευρήματά σας.
iv. Προσαρμόστε το πολλαπλό γραμμικό μοντέλο με μεταβλητή απόκρισης την τ.μ. **Screen_Time_Leasure** και επεξηγηματικές μεταβλητές τις **Age, Gender, Phone_Age_in_Days, Occupation, Operating_System.** Ερμηνεύστε τους εκτιμητές των συντελεστών του εν λόγω πολλαπλού γραμμικού μοντέλου. Δώστε 95% διαστήματα εμπιστοσύνης για τους συντελεστές του εν λόγω μοντέλου και σχολιάστε ποιες επεξηγηματικές μεταβλητές είναι στατιστικά σημαντικές σε ε.σ. 5%. Ερμηνεύστε τέλος την τιμή του συντελεστή προσδιορισμού και του διορθωμένου συντελεστή προσδιορισμού για το εν λόγω μοντέλο.
v. Με τη χρήση της εντολής `summary` εξηγήστε πλήρως τα αποτελέσματα που παίρνετε από την R στο παραπάνω πολλαπλό γραμμικό μοντέλο. Ποια είναι η κατηγορία αναφοράς των κατηγορικών μεταβλητών του μοντέλου;

vi. Ελέγξτε με γραφικούς τρόπους τις προϋποθέσεις του πολλαπλού γραμμικού μοντέλου που προσαρμόσατε στο ερώτημα iv. Σχολιάστε λεπτομερώς τα ευρήματά σας.

vii. Εκτιμήστε σημειακά και με τη βοήθεια ενός 95% διαστήματος εμπιστοσύνης, με βάση το παραπάνω πολλαπλό γραμμικό μοντέλο, τον αναμενόμενο χρόνο που ξοδεύει για διασκέδαση σε μία μέρα ένας 35 ετών άντρας, αυτοαπασχολούμενος, με κινητό λογισμικού "IOS", ηλικίας 300 ημερών.

viii. Θεωρήστε ως κατηγορία αναφοράς της κατηγορικής μεταβλητής **Occupation** τους μισθωτούς και χωρίς να προσαρμόσετε εκ νέου το πολλαπλό γραμμικό μοντέλο του ερωτήματος iv., σχολιάστε ποιοι εκτιμητές των συντελεστών θα αλλάξουν τιμή και υπολογίστε τις νέες τους τιμές. Ερμηνεύστε εκ νέου τους εκτιμητές των συντελεστών που αλλάζουν τιμή.

ix. Αν στο μοντέλο του ερωτήματος iv. είχαν κεντραριστεί οι τιμές των ποσοτικών επεξηγηματικών μεταβλητών **Age** και **Phone_Age_in_Days**, αφού αναφέρεται τις εντολές με τις οποίες υπολογίζουμε τις τιμές των κεντραρισμένων μεταβλητών, να υπολογιστούν οι εκτιμήσεις των συντελεστών του μοντέλου, χωρίς να προσαρμόσετε εκ νέου το πολλαπλό γραμμικό μοντέλο, και να ερμηνευτούν εκ νέου όσες μόνο αλλάζουν τιμή.

x. Προσαρμόστε εκ νέου το πολλαπλό γραμμικό μοντέλο του ερωτήματος iv. με μεταβλητή απόκρισης την τ.μ. **Screen_Time_Leasure** προσθέτοντας την μεταβλητή **Screen_Time_Business**. Σχολιάστε τα αποτελέσματα της `summary` εντολής της R για το νέο μοντέλο και αναφερθείτε στα αποτελέσματα της συσχέτισης του **Screen_Time_Leasure** και **Screen_Time_Business** από το πρώτο ερώτημα. Ελέγξτε εκ νέου τις προϋποθέσεις του πολλαπλού γραμμικού μοντέλου.

xi. Επικαλούμενοι τους ελέγχους προϋποθέσεων που υλοποιήσατε στα ερωτήματα vi. και x. καθώς και την τιμή του κριτηρίου AIC για τα δύο μοντέλα που προσαρμόσατε στα ερωτήματα iv. και x., ποιο από τα δύο μοντέλα θα χρησιμοποιούσατε τελικά για να προβλέψετε την αναμενόμενη τιμή της μεταβλητής απόκρισης; Θα επηρεαζόταν η απόφαση σας αν είχατε μόνο στην διάθεση σας τον συντελεστή προσδιορισμού; Σχολιάστε τα ευρήματα σας.

xii. Προσαρμόστε εκ νέου το πολλαπλό γραμμικό μοντέλο του ερωτήματος x. με μεταβλητή απόκρισης την τ.μ. **Screen_Time_Leasure** προσθέτοντας και την

αλληλεπίδραση των μεταβλητών **Age** και **Gender**. Ερμηνεύστε τους εκτιμητές των συντελεστών του εν λόγω μοντέλου.

xiii. Προσαρμόστε το πολλαπλό γραμμικό μοντέλο με μεταβλητή απόκρισης την τ.μ. log(**Screen_Time_Leasure)** και επεξηγηματικές μεταβλητές τις log(**Age)** και **Gender**, με log να δηλώνει τον νεπέριο λογάριθμο. Ερμηνεύστε τους εκτιμητές των συντελεστών του εν λόγω μοντέλου.

Οδηγίες

- Η εργασία θα πρέπει **να παραδοθεί ηλεκτρονικά** στον **Σωτήρη Ζαμπέλη** στο email του, szampelis.emp@gmail.com.
- Η εργασία που θα παραδώσετε πρέπει να είναι **σε pdf μορφή**. Παρακαλώ χρησιμοποιήστε τον **ακόλουθο τίτλο στο pdf αρχείο σας**: Surname-Name-Ex3.pdf, όπου Surname είναι το επώνυμό σας (με λατινικούς χαρακτήρες) και Name το όνομα σας (με λατινικούς χαρακτήρες). Π.χ. αν παρέδιδα εγώ εργασία θα την ονόμαζα ως εξής: Fouskakis-Dimitris-Ex3.pdf.
- Παρακαλώ χρησιμοποιήστε **ένα εξώφυλλο στο pdf αρχείο σας**, στο οποίο να αναγράφεται ο τίτλος της εργασίας (Ανάλυση Παλινδρόμησης), **το ονοματεπώνυμό σας, το email σας, η Σχολή σας και ο αριθμός μητρώου σας**.
- Θα πρέπει να **αποστείλετε ένα μόνο αρχείο**. Η εργασία θα πρέπει να περιλαμβάνει τους κώδικες της R, όχι σε παράρτημα αλλά στην απάντηση του κάθε ερωτήματος.
- Η εργασία θα πρέπει να αποσταλεί **μέχρι την Παρασκευή 23 Μαΐου στις 13:00. Καμία εργασία δεν θα γίνει δεκτή μετά την ώρα αυτή.**
- Η εργασία θα πρέπει να είναι σε μορφή **αναφοράς** και να περιλαμβάνει τους κώδικες της R με πλήρη επεξήγηση, γραφήματα και πίνακες με κατάλληλους τίτλους και πλήρη επεξήγηση των αποτελεσμάτων. Επίσης, δε θα πρέπει να υπερβαίνει τις **15 σελίδες** με μέγεθος γραμματοσειράς **12**.
- Θα δοθεί ιδιαίτερη σημασία **στην παρουσίαση της εργασίας**. Η εργασία πρέπει να είναι κατανοητή και να περιγράφει οτιδήποτε χρησιμοποιήσατε πειστικά για κάποιον που δεν γνωρίζει πολλά για το αντικείμενο.

Εύχομαι Επιτυχία