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

Διαφημιστική εταιρεία μελετά τους παράγοντες που επηρεάζουν το ύψος των χρημάτων που χρησιμοποιεί για αγορές ένας ιδιώτης κατά τις εορταστικές περιόδους. Για αυτό το λόγο, συλλέχθηκαν πληροφορίες σε δείγμα 53 διαφορετικών ανθρώπων για τις ακόλουθες μεταβλητές: την ηλικία (**age**) του ατόμου σε έτη, τη μοναδική εορταστική περίοδο (**holiday)** που πραγματοποιήθηκαν οι αγορές, με κατηγορίες τις "Christmas", "Easter", "Other", (Χριστούγεννα – Πάσχα – Λοιπές), το φύλο (**sex)** του ατόμου που έκανε τις αγορές, με κατηγορίες τις "Man", "Woman", (Άντρας – Γυναίκα), το μέσο χρόνο **(time)** σε λεπτά που θεωρεί πως γίνεται αποδέκτης διαφημιστικών προωθήσεων (από οποιοδήποτε μέσο) στη διάρκεια μιας ημέρας, το μηνιαίο του εισόδημα (**salary)** σε ευρώ και τέλος το ποσό (**spend**) σε ευρώ που ξόδεψε την εορταστική περίοδο. Κάθε άνθρωπος στο δείγμα έχει προβεί αναγκαστικά σε αγορές σε μία μόνο περίοδο από τις τρεις της μεταβλητής **holiday**.

Τα δεδομένα βρίσκονται στο αρχείο:

http://www.math.ntua.gr/~fouskakis/Data_Analysis/Exercises/gifts.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 **spend**, **age, time** και **salary**. Εξετάστε αν οι συντελεστές συσχέτισης είναι στατιστικά σημαντικοί για κάθε συνδυασμό.
ii. Προσαρμόστε το πολλαπλό γραμμικό μοντέλο με μεταβλητή απόκρισης την τ.μ. **spend** και επεξηγηματικές μεταβλητές τις **age, holiday, sex, time** και **salary.** Ερμηνεύστε τους εκτιμητές των συντελεστών του παραπάνω πολλαπλού γραμμικού μοντέλου. Δώστε 95% διαστήματα εμπιστοσύνης για τους συντελεστές του εν λόγω μοντέλου.
iii. 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 της κατηγορικής μεταβλητής **holiday**;
iv. Ελέγξτε τις προϋποθέσεις του πολλαπλού γραμμικού μοντέλου που προσαρμόσατε παραπάνω. Σχολιάστε λεπτομερώς τα ευρήματά σας.
v. Εκτιμήστε σημειακά και με τη βοήθεια ενός 90% διαστήματος εμπιστοσύνης, με βάση το παραπάνω πολλαπλό γραμμικό μοντέλο, το ποσό σε ευρώ που αναμένεται να ξοδέψει στην διάρκεια των εορτών του Πάσχα, μια 25-χρονή γυναίκα, που παρακολουθεί 13 λεπτά διαφημιστικά μηνύματα την ημέρα και αμείβεται με 575 ευρώ το μήνα.
vi. Θεωρήστε ως κατηγορία αναφοράς της κατηγορικής μεταβλητής **holiday** την εορταστική περίοδο του Πάσχα και χωρίς να προσαρμόσετε εκ νέου το πολλαπλό γραμμικό μοντέλο του ερωτήματος ii., σχολιάστε ποιοι συντελεστές θα αλλάξουν τιμή και υπολογίστε τους (χωρίς τη χρήση της R). Ερμηνεύστε εκ νέου όσους αλλάζουν τιμή.
vii. Με χρήση του νεπέριου λογάριθμου (μόνο για τις ποσοτικές μεταβλητές) προσαρμόστε το πολλαπλό πολλαπλασιαστικό μοντέλο με μεταβλητή απόκρισης την τ.μ. **spend** και επεξηγηματικές μεταβλητές τις **age, holiday, sex, time** και **salary.** Ερμηνεύστε τις εκτιμώμενες τιμές των συντελεστών των πέντε επεξηγηματικών μεταβλητών.
viii. Ελέγξτε τις προϋποθέσεις του πολλαπλού πολλαπλασιαστικού μοντέλου που προσαρμόσατε παραπάνω. Σχολιάστε λεπτομερώς τα ευρήματά σας.

ix. Απαντήστε στο ερώτημα v. με χρήση του πολλαπλού πολλαπλασιαστικού μοντέλου του ερωτήματος vii..

x. Αφού πρώτα ορίσετε τη νέα μεταβλητή **timesq** με χρήση της εντολής *timesq=time^2,* προσαρμόστε το πολλαπλό γραμμικό μοντέλο με μεταβλητή απόκρισης την τ.μ. **spend** και επεξηγηματικές μεταβλητές τις **age, holiday, sex, time, salary** και **timesq.** Ερμηνεύστε τις εκτιμώμενες τιμές των συντελεστών των έξι επεξηγηματικών μεταβλητών.

xi. Ελέγξτε τις προϋποθέσεις του πολλαπλού γραμμικού μοντέλου που προσαρμόσατε στο ερώτημα x.. Σχολιάστε λεπτομερώς τα ευρήματά σας.

xii. Απαντήστε στο ερώτημα v. με χρήση του πολλαπλού γραμμικού μοντέλου του ερωτήματος x.

xiii. Επικαλούμενοι τους ελέγχους προϋποθέσεων που κάνατε στα ερωτήματα iv., viii. και xi., καθώς και (αν χρειάζεται) μεθόδους σύγκρισης μοντέλων, όσον αφορά τα τρία μοντέλα που προσαρμόσατε στα ερωτήματα ii., vii. και x. ποια πρόβλεψη από αυτήν που υπολογίσατε στα ερωτήματα v., ix. και xii. θα εμπιστευόσασταν περισσότερο;

Οδηγίες

- Η εργασία θα πρέπει **να παραδοθεί ηλεκτρονικά** στον **Γιώργο Τζουμέρκα** στο email του, tzoum_giorgos@hotmail.gr.
- 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 Τρίτη 6 Ιουνίου 2023 στις 13:00. Καμιά 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