M.Sc. in Data Science

Course: Probability and Statistics for Data Analysis

Instructor: Ioannis Vrontos (vrontos@aueb.gr)

Grader: Constandina Koki (kokiconst@aueb.gr)

Assignment 3 Deadline: 9 January 2018

Άσκηση 1

Οι τιμές που δίνονται στον παρακάτω πίνακα προέρχονται από τα αρχεία τμήματος κάποιου πανεπιστημιακού ιδρύματος και αναφέρονται στον βαθμό εισαγωγής σε αυτό και στον βαθμό αποφοίτησης από αυτό 15 τυχαία επιλεγμένων αποφοίτων που εισήχθησαν την ίδια χρονιά.

Βαθμός Εισαγωγής	Βαθμός Αποφοίτησης
17.24	6.88
18.06	7.46
17.41	7.34
17.60	7.21
18.95	8.22
19.60	8.42
17.49	6.41
18.60	8.04
17.50	7.35
19.24	8.30
18.87	8.09
17.68	6.68
19.31	8.77
18.40	7.33
17.20	7.15

1) Να εκτιμηθεί το απλό γραμμικό μοντέλο για τις παραπάνω μεταβλητές, όπου εξαρτημένη μεταβλητή είναι ο Βαθμός Αποφοίτησης, και ανεξάρτητη μεταβλητή ο Βαθμός εισαγωγής των φοιτητών/τριών. Επίσης, να υπολογιστεί ο συντελεστής προσδιορισμού και ο συντελεστής συσχέτισης των δύο μεταβλητών. Να διατυπωθεί και να σχολιαστεί ο έλεγχος που δίνεται στον πίνακα ανάλυσης διακύμανσης του μοντέλου. Στην περίπτωση του απλού γραμμικού μοντέλου, ο έλεγχος αυτός με ποιον άλλο έλεγχο είναι ισοδύναμος;
2) Να αξιολογηθεί το απλό γραμμικό μοντέλο ως προς την ικανοποίηση των βασικών υποθέσεων του. Ποιες από αυτές ικανοποιούνται και ποιες όχι. Η αξιολόγηση να γίνει και με ελέγχους και διαγραμματικά.
3) Να υπολογιστούν 95% διαστήματα εμπιστοσύνης για τις τιμές των παραμέτρων του μοντέλου. Επίσης να προβλεφθεί ο βαθμός αποφοίτησης για έναν φοιτητή που εισήχθη στην σχολή με βαθμό 18.5. Τέλος, για τον ίδιο βαθμό εισαγωγής να υπολογιστούν 95% διαστήματα εμπιστοσύνης για τον μέσο και την πρόβλεψη.
4) Να εκτιμηθεί και να αξιολογηθεί το μοντέλο $Y=\alpha+\beta X+\gamma X^2+\varepsilon$. Θεωρείτε ότι είναι καλύτερο σε σχέση με το απλό γραμμικό μοντέλο για την περιγραφή της σχέσης των δύο μεταβλητών;
5) Να εκτιμηθεί το εκθετικό μοντέλο $\ln Y=\alpha+\beta X+\varepsilon$ και να υπολογιστεί το ποσοστό της μεταβλητότητας της Y που ερμηνεύεται από αυτό.

Άσκηση 2

Μεσιτικό γραφείο ενδιαφέρεται να ερευνήσει ποιοι παράγοντες επηρεάζουν την τιμή πώλησης των διαμερισμάτων σε μια μεγάλη πόλη. Από το αρχείο των πελατών της, στο οποίο καταγράφει στοιχεία για τις τιμές πώλησης αλλά και τα χαρακτηριστικά των διαμερισμάτων, επέλεξε τυχαίο δείγμα 30 διαμερισμάτων, που πουλήθηκαν κατά το τελευταίο έτος. Τα στοιχεία για τα διαμερίσματα αυτά παρουσιάζονται στον Πίνακα 1, όπου:

- Y: Τιμή πώλησης (σε 1000$)
- X_1: Μέγεθος διαμερίσματος (σε τετραγωνικά πόδια)
- X_2: Αριθμός υπνοδωματίων
- X_3: Συνολικός αριθμός δωματίων
- X_4: Ηλικία διαμερίσματος (σε έτη)
- X_5: Τοποθεσία (0: κέντρο, 1: προάστιο).

ΠΙΝΑΚΑΣ 1

Y	X_1	X_2	X_3	X_4	X_5
84.00	13.80	3	7	10	0
93.00	19.00	2	7	22	1
83.10	10.00	2	7	15	1
85.20	15.00	3	7	12	1
85.20	12.00	3	7	8	1
85.20	15.00	3	7	12	1
85.20	12.00	3	7	8	1
63.30	9.10	3	6	2	1
84.30	12.50	3	7	11	1
84.30	12.50	3	7	11	1
77.40	12.00	3	7	5	0
92.40	17.90	3	7	18	0
92.40	17.90	3	7	18	0
61.50	9.50	2	5	8	0
88.50	16.00	3	7	11	0
88.50	16.00	3	7	11	0
65.00	8.00	2	5	5	0
81.60	11.80	3	7	8	1
86.70	16.00	3	7	9	0
89.70	16.80	2	7	12	0
86.70	16.00	3	7	9	0
89.70	16.80	2	7	12	0
75.90	9.50	3	6	6	1
78.90	10.00	3	6	11	0
87.90	16.50	3	7	15	0
91.00	15.10	3	7	8	1
92.00	17.90	3	8	13	1
87.90	16.50	3	7	15	0
90.90	15.00	3	7	8	1
91.90	17.80	3	8	13	1

Ο υπεύθυνος του μεσιτικού γραφείου ανέθεσε την ανάλυση των δεδομένων σε 3 διαφορετικούς ερευνητές και τα συμπεράσματα στα οποία κατέληξαν αναφορικά με τις μεταβλητές που επηρεάζουν την τιμή των διαμερισμάτων συνοψίζονται παρακάτω:

Ερευνητής	**Μεταβλητές**
$1^{ος}$	X_1, X_4, X_5
$2^{ος}$	X_1, X_3, X_4
$3^{ος}$	X_1, X_3

Να αναλύσετε τα δεδομένα, και να προτείνετε το καλύτερο κατά τη γνώμη σας μοντέλο για την περιγραφή της σχέσης των μεταβλητών. Να επισημανθούν τα συμπεράσματα που προκύπτουν από αυτήν και θα μπορούσαν να φανούν χρήσιμα στο μεσιτικό γραφείο. Με ποιον ή ποιους από τους παραπάνω ερευνητές συμφωνείτε ή διαφωνείτε και γιατί.

[Δείτε το αρχείο Assignment3-askisi2.txt]
Προκειμένου να διαβάσετε τα δεδομένα στο R, θα χρησιμοποιήσετε την ακόλουθη εντολή:
dataall<-read.table("E:/Vrontos/mathimata/A-mathimata-2017-2018/MSc-Informatics-Statistics-Full-Time/Assignments-Vrontos-Koki/Assignment3/Assignment3-askisi2.txt")

Άσκηση 3
Ο πίνακας που ακολουθεί περιέχει δεδομένα για το Ελληνικό ΑΕΠ (GDP), το Κεφάλαιο (C) και την Εργασία (L) για τα έτη 1961-2003 (δείτε το αρχείο Assignment3-askisi3.txt).

Πίνακας *:* Ελληνικό ΑΕΠ (GDP), Κεφάλαιο (C) και Εργασία (L) για τα έτη 1961-2003

	Ελληνικό ΑΕΠ (GDP)	Κεφάλαιο (C)	Εργασία (L)
1961	11706212.29	2290131.38	3401000.00
1962	12549494.92	2401897.68	3375214.13
1963	14000676.76	2243636.03	3349428.26
1964	15251018.50	2663762.56	3323571.74
1965	16446877.85	3071789.90	3297785.87
1966	18117905.50	3223234.09	3272000.00
1967	19584075.02	3176443.24	3246214.13
1968	18323205.16	3919155.11	3220357.61
1969	19489808.57	4536246.53	3194571.74
1970	20825230.37	4396758.43	3168785.87
1971	21437700.75	4925347.72	3143000.00
1972	22991261.62	6122219.13	3162869.53
1973	23745853.27	6478847.59	3182684.76
1974	23906643.82	4307664.68	3202500.00
1975	23535201.22	4810478.13	3222315.24
1976	23268630.35	5138386.92	3242184.76
1977	23017649.70	5773492.41	3262000.00
1978	23480438.05	6482002.43	3276000.00
1979	24069692.92	6759740.86	3311000.00
1980	24194292.11	5686644.18	3356000.00
1981	23647776.08	5132341.68	3531000.00

1982	24661760.01	4966824.37	3502000.00
1983	25598906.89	5282601.95	3540000.00
1984	25598906.89	4424784.69	3553000.00
1985	26392473.00	4838471.51	3588000.00
1986	26577220.32	4843611.26	3601000.00
1987	26151984.79	4570782.46	3598000.00
1988	26675024.49	4688397.23	3657000.00
1989	27235205.00	4974813.46	3671000.00
1990	27877490.00	5196645.96	3719000.00
1991	28891408.00	5417470.18	3632000.00
1992	29863171.00	5228582.83	3685000.00
1993	30885829.00	5019754.32	3715363.00
1994	32217205.00	4864973.20	3786157.00
1995	33487975.30	5065971.00	3820510.00
1996	34814546.31	5490784.00	3806800.00
1997	36304858.21	5866600.49	3784000.00
1998	11706212.29	6486510.57	3940400.00
1999	12549494.92	6887056.87	3909900.00
2000	14000676.76	7422269.15	3897593.95
2001	15251018.50	8052752.48	3913730.21
2002	16446877.85	8756416.11	3942723.33
2003	18117905.50	9554382.78	3983595.60

Να χρησιμοποιηθούν αυτά τα δεδομένα για να εκτιμηθεί κατάλληλο υπόδειγμα για το Ελληνικό ΑΕΠ.

(Α) Θεωρήστε το υπόδειγμα:

$$\ln(GDP_t) = \beta_0 + \beta_1 \ln(C_t) + \varepsilon_t ,$$

όπου ln(GDP) και ln(C) είναι ο λογάριθμος του Ελληνικού ΑΕΠ (GDP) και ο λογάριθμος του κεφαλαίου (C), αντίστοιχα.

(1) Να εκτιμήσετε τις παραμετρους του παραπάνω υποδείγματος και να ερμηνεύσετε τα αποτελέσματα.

(2) Να ελεγχθεί η στατιστική σημαντικότητα των παραμέτρων β_0 και β_1 σε επίπεδο σημαντικότητας α=5%.

(3) Να εξετάσετε αν ισχύουν οι υποθέσεις που αφορούν τα κατάλοιπα του υποδείγματος όσον αφορά την συσχέτιση, την ομοσκεδαστικότητα και την κανονικότητα των καταλοίπων. Θεωρείτε ότι το υπόδειγμα είναι αξιόπιστο;

(Β) Θεωρείστε τα παρακάτω υποδείγματα:

$$\mathrm{M1}: \ln(GDP_t) = \beta_0 + \beta_1 \ln(C_t) + \varepsilon_t$$

$$\mathrm{M2}: \ln(GDP_t) = \gamma_0 + \gamma_1 \ln(L_t) + \varepsilon_t$$

$$\mathrm{M3}: \ln(GDP_t) = \delta_0 + \delta_1 \ln(C_t) + \delta_2 \ln(L_t) + \varepsilon_t$$

(1) Να επιλέξετε το καταλληλότερο υπόδειγμα από τα παραπάνω υποδείγματα Μ1-Μ3 για το Ελληνικό ΑΕΠ.

(2) Να γίνει πρόβλεψη του ΑΕΠ για K=5000000 και L=3500000