

ΑΡΙΣΤΟΤΕΛΕΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ ΘΕΣΣΑΛΟΝΙΚΗΣ
ΠΟΛΥΤΕΧΝΙΚΗ ΣΧΟΛΗ - ΤΜΗΜΑ ΜΗΧΑΝΟΛΟΓΩΝ ΜΗΧΑΝΙΚΩΝ
ΑΚΑΔΗΜΑΪΚΟ ΕΤΟΣ 2023-24

ΕΞΑΜΗΝΟ: 8ο

ΜΑΘΗΜΑ: ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

ΔΙΔΑΣΚΟΥΣΑ: ΣΟΦΙΑ ΠΑΝΑΓΙΩΤΙΔΟΥ, Αν. Καθηγήτρια

1^η ΕΡΓΑΣΙΑ

Όνοματεπώνυμο φοιτητή : _____

A.E.M. : _____

Θεσσαλονίκη, Μάρτιος 2024

Εργασία στο μάθημα "Ανάλυση Δεδομένων" 2023-24

Τα δεδομένα που σας δίνονται αφορούν τη διάρκεια ζωής (life expectancy) του πληθυσμού διαφόρων χωρών, όπως αυτά έχουν καταγραφεί από τον Παγκόσμιο Οργανισμό Υγείας για τα έτη 2000 - 2015. Συγκεκριμένα, τα δεδομένα απαρτίζονται από τις ακόλουθες μεταβλητές:

No.	Variable	Description
1	Country	Country
2	Year	Year
3	Status	Developed or Developing status
4	Life expectancy	Life Expectancy in age
5	Adult Mortality	Adult Mortality Rates of both sexes (probability of dying between 15 and 60 years per 1000 population)
6	Alcohol	Alcohol, recorded per capita (15+) consumption (in litres of pure alcohol)
7	percentage expenditure	Expenditure on health as a percentage of Gross Domestic Product per capita (%)
8	Hepatitis B	Hepatitis B (HepB) immunization coverage among 1-year-olds (%)
9	Measles	Measles - number of reported cases per 1000 population
10	BMI	Average Body Mass Index of entire population
11	under-five deaths	Number of under-five deaths per 1000 population
12	Polio	Polio (Pol3) immunization coverage among 1-year-olds (%)
13	Total expenditure	General government expenditure on health as a percentage of total government expenditure (%)
14	Diphtheria	Diphtheria tetanus toxoid and pertussis (DTP3) immunization coverage among 1-year-olds (%)
15	HIV/AIDS	Deaths per 1 000 live births HIV/AIDS (0-4 years)
16	GDP	Gross Domestic Product per capita (in USD)
17	Population	Population of the country
18	thinness 5-9 years	Prevalence of thinness among children for Age 5 to 9 (%)
19	Income composition of resources	Human Development Index in terms of income composition of resources (index ranging from 0 to 1)
20	Schooling	Number of years of Schooling (years)

Να χρησιμοποιηθούν τα δεδομένα αυτά προκειμένου να διερευνηθεί κατά πόσο κάποιες από τις μεταβλητές μπορούν να χρησιμοποιηθούν για την πρόβλεψη/ερμηνεία της διάρκειας ζωής του πληθυσμού (Life expectancy). Συγκεκριμένα να διερευνηθούν τα ακόλουθα:

1. Να προσδιοριστούν τα μοντέλα απλής γραμμικής παλινδρόμησης που συνδέουν τη μεταβλητή απόκρισης με κάθε μία από τις υποψήφιες προβλεπτικές μεταβλητές (predictors), εξαιρώντας τις μεταβλητές Country και Status από την ανάλυση. Με ποιες από τις μεταβλητές υπάρχουν στατιστικά σημαντικές ενδείξεις εξάρτησης της διάρκειας ζωής; Στις περιπτώσεις αυτές να κατασκευαστούν τα αντίστοιχα διαγράμματα διασποράς.

2. Να προσδιοριστεί το μοντέλο πολλαπλής γραμμικής παλινδρόμησης που συνδέει τη μεταβλητή απόκρισης με όλες τις υποψήφιες προβλεπτικές μεταβλητές (predictors), εξαιρώντας τις μεταβλητές Country και Status από την ανάλυση. Για ποιες από τις μεταβλητές απορρίπτεται η μηδενική υπόθεση $H_0 : \beta_j = 0$;
3. Να προσδιοριστεί το βέλτιστο μοντέλο πολλαπλής γραμμικής παλινδρόμησης σύμφωνα με τη μέθοδο της πρόσω επιλογής μεταβλητών (forward selection), εξαιρώντας τις μεταβλητές Country και Status από την ανάλυση. Ποιες μεταβλητές περιλαμβάνονται στο προτεινόμενο μοντέλο και ποιος είναι ο συντελεστής προσδιορισμού που προκύπτει;
4. Υπάρχουν ενδείξεις μη γραμμικής σχέσης μεταξύ της μεταβλητής απόκρισης και κάποιας εκ των μεταβλητών Year, Alcohol, BMI και Schooling; Για κάθε μία από τις 4 αυτές προβλεπτικές μεταβλητές να προσδιοριστεί το καταλληλότερο μοντέλο παλινδρόμησης της μορφής

$$Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$$

και να σχολιαστούν τα αποτελέσματα.