

ΣΤΑΤΙΣΤΙΚΑ ΜΟΝΤΕΛΑ

ΣΕΙΡΑ 2

A) Στο αρχείο **vehicles.txt** παρουσιάζονται τα αποτελέσματα μιας μελέτης για 32 τύπους αυτοκινήτων. Η πρώτη στήλη δίνει τη λίστα των τύπων αυτοκινήτων (**car**).

Ακολουθούν οι μεταβλητές

mpg	Κατανάλωση βενζίνης Miles/(US) gallon
cyl	Αριθμός κυλίνδρων
disp	Μετατόπιση (Displacement) (cu.in.)
hp	Μικτή υποδύναμη (Gross horsepower)
drat	Αναλογία οπίσθιου άξονα (Rear axle ratio)
wt	Βάρος (1000 lbs)
qsec	1/4 mile time
vs	Διάταξη κινητήρα (0 = V, 1 = straight)
am	Κιβώτιο ταχυτήτων (0 = automatic, 1 = manual)
gear	Αριθμός προς τα εμπρός ταχυτήτων (forward gears)
carb	Αριθμός καρμπυρατέρ

1) Να προσαρμοστεί ένα μοντέλο πολλαπλής γραμμικής παλινδρόμησης στα δεδομένα του αρχείου σχετίζοντας τα μίλια/gallon *mpg* (*Y*) με τις δέκα παραπάνω επεξηγηματικές μεταβλητές. Να εξετάσετε αν υπάρχουν συσχετίσεις μεταξύ των μεταβλητών X_j , αν υπάρχει πολυσυγγραμμικότητα και αν τηρούνται οι προϋποθέσεις του μοντέλου με βάση την εξέταση των υπολοίπων. Να γίνει χρήση διαγνωστικών ελέγχων π.χ. για την πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής, αξιοποιώντας μέτρα όπως τα h_{ii} , απόσταση Cook, DFFITS, DFBETAS .

2) Να εξεταστεί αν το μοντέλο με τις δέκα επεξηγηματικές μεταβλητές είναι το βέλτιστο και αν όχι, να επιλέξετε ανάμεσα σε όλα τα δυνατά μοντέλα το βέλτιστο (να αξιοποιηθούν τεχνικές με βήματα με ελέγχους F και t , τα κριτήρια R^2 , \bar{R}^2 , R^2_{predict} , C_p και AIC).

3) Με χρήση διαγνωστικών τεχνικών, καθώς και με γραφικές παραστάσεις των πρόσθετων μεταβλητών και μερικών υπολοίπων, να εξεταστεί η καταλληλότητα του τελικού μοντέλου, αν χρειάζεται μετασχηματισμούς ή περαιτέρω βελτιώσεις. Εξετάζεται πάλι η πιθανή παρουσία άτυπων σημείων ή σημείων επιρροής. Επίσης να βρεθούν 95% Δ.Ε. για τους συντελεστές του τελικού μοντέλου, καθώς και για τη πρόβλεψη μιας άγνωστης Y (με τιμές του \mathbf{x}_0 της επιλογής σας) και τέλος να δοθούν ερμηνείες.

B) Αρχείο bacteria.txt

Τα δεδομένα αφορούν την επιβίωση ενός είδους βακτηρίου, με Y τον αριθμό αυτών που επέζησαν μετά από $X=1, 2, 3, \dots, 15$ λεπτά συνεχούς ακτινοβολίας και αντίστοιχα για μια δεύτερη ομάδα μετά από $X=1, 2, \dots, 10$ λεπτά συνεχούς ακτινοβολίας. Εδώ είναι απαραίτητη η χρήση του μετασχηματισμού $Y^*=\ln(Y)$.

1) Κατασκευάστε ένα διάγραμμα διασποράς για τις δύο ομάδες πριν και μετά από το μετασχηματισμό.

2) Περιγράψτε πώς μέσω μιας δείκτριας μεταβλητής Z σε ένα μοντέλο παλινδρόμησης μπορούμε να ελέγξουμε αν στα δεδομένα χρειάζεται να προσαρμοστούν

(Α) δύο διαφορετικές ευθείες για τις δύο ομάδες ή (Β) δύο παράλληλες ευθείες ή (Γ) μια ευθεία.

3) Να γίνουν αυτοί οι έλεγχοι.

Χ. Καρώνη