

Στατιστική Μοντελοποίηση

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
displacement	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 παρατήρησης Y (με τιμές του x_0 της επιλογής σας) και τέλος να δοθούν ερμηνείες.

B) Αρχείο canary.txt

Εξετάζεται ο αριθμός παλμών/1'' κελαηδήματος Y δύο ειδών καναρινιών A και B σε σχέση με διαφορετικές τιμές θερμοκρασίας x_1 (σε $^{\circ}\text{C}$). Στόχος της ανάλυσης είναι να εξετάσουμε αν τα δύο είδη καναρινιών διαφοροποιούνται μεταξύ τους ως προς τους ρυθμούς των παλμών. Έστω δείκτρια μεταβλητή x_2 ($x_2=0$ - αν ομάδα B, $x_2=1$ - αν ομάδα A).

- 1) Περιγράψτε πώς μέσω του μοντέλου παλινδρόμησης $E(y_x) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$, μπορούμε να ελέγξουμε αν χρειάζεται να προσαρμοστούν **(I)** δύο διαφορετικές ευθείες, **(II)** δύο παράλληλες ευθείες, ή **(III)** μια κοινή ευθεία και για τα δύο είδη καναρινιών, όπου $x_3 = x_1 x_2$, η μεταβλητή που εκφράζει την αλληλεπίδραση μεταξύ των μεταβλητών x_1 και x_2 .
 - 2) Να γίνουν αυτοί οι έλεγχοι, η γραφική παράσταση και να δοθούν ερμηνείες για το τελικό μοντέλο
-

Χ. Καρώνη
Χειμερινό Εξάμηνο 2021-22