# Πολλαπλή Παλινδρόμηση και Διαγνωστικά με την R

## Εργασία 3, 24/05/2019

### Άσκηση 1

Στην πρώτη άσκηση, θα ασχοληθούμε με τα δεδομένα anscombe, που δημοσιεύτηκαν απο τον Frank Anscombe το 1973.

(1) Πληκτρολογήστε το όνομα των δεδομένων αυτών για να δείτε τί περιέχουν και εξηγείστε σύντομα.
(2) Υπολογίστε τη μέση τιμή και τη διασπορά των $x_i$ και $y_i$, $i = 1 \dots 4$ με τη βοήθεια της συνάρτησης apply.
(3) Βρείτε τις εξισώσεις των ευθειών παλινδρόμησης των $y_i$ πάνω στα $x_i$, για $i = 1 \dots 4$. Τί παρατηρείτε ?
(4) Απεικονίστε και τα 4 γραφήματα στο ίδιο παράθυρο. Για να βάλετε έναν τίτλο χρησιμοποιείστε τη συνάρτηση mtext με την επιλογή outer=True. Οι γραφικές επιλογές είναι προσβάσιμες με τη βοήθεια της par, π.χ., par(mfrow=c(2, 2)). Τί παρατηρείτε ?
(5) Με χρήση της συνάρτησης summary ερμηνεύστε και συγκρίνετε τα αποτελέσματα και των 4 παλινδρομήσεων ?
(6) Κάντε τους διαγνωστικούς ελέγχους των παλινδρομήσεων απο 2 μέχρι 4. Για παράδειγμα, με την plot(reg) βγαίνει ένας αριθμός γραφημάτων. Εξηγείστε λεπτομερώς ποιός είναι ο ρόλος αυτών των γραφημάτων ? Τί συμπεράσματα βγάζετε απο τους διαγνωστικούς ελέγχους ? Ποιά τυποποιημένα κατάλοιπα χρησιμοποιούνται στα γραφήματα ? Ελέγξτε με τη βοήθεια της θεωρίας (rstudent(reg)). Υπάρχουν σημεία μοχλοί (hatvalues(reg)), σημεία επιροής ?

### Άσκηση 2

(Συνέχεια της εργασίας 2) Σε αυτήν την άσκηση ενδιαφερόμαστε να ερμηνεύσουμε τη μέγιστη συγκέντρωση του όζοντος ως συνάρτηση πολλών υποψήφιων ανεξάρτητων μεταβλητών. Συγκεκριμένα, στο αρχείο ozone.txt δίνονται δεδομένα απο τις εξής μεταβλητές: $T12, T15$ : θερμοκρασία στις $12h$ και $15h$; $Vx$: η ταχύτητα του ανέμου; $Ne12$ : νεφελότητα στις $12h$ ; $max03v$ : η μέγιστη συγκέντρωση του όζοντος O3 την προηγούμενη μέρα ; $N12$, $S12$, $E12$, $W12$ : ατμοσφαιρικός δείκτης της ποιότητας του αέρα στις 4 κατευθύνσεις στις $12h$.

(1) Κάντε την γραμμική παλινδρόμηση της $max03$ σε όλες τις άλλες επεξηγηματικές μεταβλητές, και ερμηνεύστε όλα τα αποτελέσματα της summary. Είναι όλες οι μεταβλητές στατιστικά σημαντικές σε επίπεδο 0.05 ?
(2) Κάντε μία επιλογή μεταβλητών με τη μέθοδο backward. Σε ποιό μοντέλο καταλήγετε ? Τί κάνει η συνάρτηση update ?
(3) Φορτώστε το πακέτο leaps. Εφαρμόστε τη συνάρτηση regsubsets επιλέγοντας nbest=1 (κρατάμε μόνο ένα καλύτερο μοντέλο για κάθε δυνατή τιμή του $p$), nvmax=10 (ο αριθμός των επιπέδων που ελέγχουμε), method='exhaustive'. Παρουσιάστε τα γραφικά αποτελέσματα ελέγχοντας τα κριτήρια BIC, Cp Mallow's, R2 και προσαρμοσμένο R2 [1]. Σχολιάστε τα αποτελέσματα.
(4) Κάντε μία επιλογή μεταβλητών με τη μέθοδο forward : Για το σκοπό αυτό χρησιμοποιείστε τη συνάρτηση step. Μετά απο όλους αυτούς τους ελέγχους καταλήξτε στο καλύτερο μοντέλο πρόβλεψης της συγκέντρωσης του όζοντος.

[1] επιλογή scale της plot στο αποτέλεσμα της regsubsets.