In dieser Übung wird der Datensatz *Donald.RData* verwendet. Für diese Daten wurde in der ersten Laborübung eine multiple lineare Regression durchgeführt.

A 1 Teile die Daten zufällig in zwei Teildatensätze vom Umfang 50 (Testdaten) und 100 (Trainingsdaten) auf. Wiederhole den Vorgang drei Mal und vergleiche die Verteilungen der 6 Variablen im jeweiligen Trainingsdatensatz mit den Verteilungen im Testdatensatz. Führe eine lineare Regression auf dem Trainingsdatensatz durch und bestimme für alle drei Aufteilungen den Test MSE und vergleiche diese. Interpretiere Deine Ergebnisse.
Verwendete Funktionen: *sample*, *predict*.

A 2

a) Führe eine 'Leave-One-Out' Cross Validation mit Hilfe einer *for* Schleife durch. Gehe folgendermaßen vor:

(i) Führe eine lineare Regression auf den Daten bis auf die erste Zeile durch.

(ii) Bestimme mit Hilfe von (i) einen prognostizierten Wert für die Variable *Trump* in der ersten Zeile und bestimme die quadratische Abweichung vom wahren Messwert.

(iii) Führe Schritt (i) und (ii) mit Hilfe einer *for* Schleife für die Zeilen 2-150 durch.

(iv) Berechne den MSE und Vergleiche das Ergebnis mit den Ergebnissen aus Aufgabe 1. Interpretiere den Vergleich.

b) Führe 'Leave-One-Out' Cross Validation automatisiert mit Hilfe der Funktionen *glm* und *cv.glm* (Paket *boot*) durch und vergleiche das Ergebnis mit dem Ergebnis aus a).

A 3 Führe jeweils 10 k-fache Cross Validations mit $k = 5$ und $k = 10$ mit Hilfe der Funktionen *glm* und *cv.glm* durch. Vergleiche die Ergebnisse sowohl untereinander als auch mit den Ergebnissen aus Aufgabe 1 und Aufgabe 2.