# Homework di Data Science & Data Analysis

## Laurea Magistrale in Ingegneria Informatica – Università degli Studi di Salerno

### Quesito su regressione lineare

Si analizzi in ambiente R il data set `Regression2024.csv` che consta di $n = 60$ osservazioni di una variabile dipendente $Y$ e di $p = 30$ regressori $X_j (j = 1, 2, \ldots, p)$, potenzialmente utili alla predizione di $Y$. L'obiettivo dell'analisi è, dopo aver confrontato diverse tecniche di regressione presentate al corso, la determinazione del modello empirico lineare che minimizza l'errore di predizione su un test set. Si chiede quindi di:

**a)** valutare la correlazione e la *multicollinearità* tra i regressori;

**b)** stimare i parametri $\beta_j$ $(j = 0, 1, \ldots, p)$ implementando lo stimatore ai minimi quadrati della regressione multipla **senza usare il comando** `lm()` **e** calcolare i *p-value* per i test sui parametri **senza usare il comando** `summary()`;

**c)** stimare i parametri del modello di regressione multipla tramite il comando `lm()`, calcolare i *p-value* tramite il comando `summary()` e confrontare i risultati con quelli ottenuti al punto b);

**d)** selezionare la strategia che permette di costruire il modello di regressione che minimizza l'errore di test sulla variabile $Y$, scegliendo tra

*i) stepwise*, utilizzando tutti gli approcci (*forward*, *backward* e *ibridi*, provando le metriche *AIC* e *BIC* e gli approcci basati su *cross-validation*);

*ii) ridge regression* e *iii) LASSO*;

**e)** individuare i regressori significativi per la predizione di $Y$ e fornire la stima dei loro coefficienti $\beta_j$ con la strategia determinata al punto d).

Si richiede che l'80% delle osservazioni del data set per la regressione venga utilizzato per il *training* dei modelli e la scelta dei loro parametri, mentre il test si basi sul restante 20% dei dati forniti.