**Esercizio 1 – SAS**

Il dataset 'companies è una Matrice di dati che contiene alcune informazioni riguardanti 64 compagnie che compaiono nella top 500 redatta da Forbes per l’anno 1986.

Le variabili presenti nel dataset sono le seguenti:

* assets = attivo in bilancio (milioni di dollari);
* sales = fatturato relativo alle vendite (milioni di dollari);
* mark\_val = valore di mercato della compagnia (milioni di dollari);
* profits = profitto (milioni di dollari);
* cash = flusso di cassa (milioni di dollari);
* employ = numero complessivo di dipendenti (migliaia);
* sector = settore di mercato in cui opera la compagnia: Communication; Energy; Finance; HiTech; Manufacturing; Medical; Retail; Transportation; Other.

1. Si effettui un’analisi descrittiva e preliminare del dataset e si commentino i risultati
2. Si studi la correlazione tra le variabili numeriche del dataset e si commentino i risultati
3. Si producano i seguenti grafici a dispersione si commentino i risultati
   1. profits vs. cash
   2. assets vs. profits
   3. assets vs. employ
4. Si costruisca il modello di regressione lineare che combina la relazione tra profits(y) e cash(x) e se ne commentino i risultati
5. Si aggiunga la retta di regressione appena stimata al grafico a dispersione stima a punto 4 (facoltativo)
6. Si studi la presenza di eventuali valori estremi presenti nelle variabili profits e assets sia quantitativamente sia graficamente
7. Si stimino i modelli di regressione che modellizano la relazione tra assets(y) e employ (x)
   1. Lineare Semplice
   2. Parabolico (2° Livello)
   3. Cubico (3° Livello)
   4. Polinomio di grado 4
   5. Polinomio di grado 5

Si valuti per questi modelli l’adattamento ai dati sia quantitativamente che graficamente.

**Esercizio 2 – SAS**

Il data set PM\_10 contiene informazioni relative all’inquinamento dell’aria rilevato in corrispondenza delle strade e

degli snodi stradali principali, e contiene le seguenti variabili:

* PM10: concentrazione di particelle di pm10
* CARS: numero di auto che transitano in un’ora
* TEMP: temperatura misurata a 2 metri dal suolo
* WIND: velocità del vento (metri/secondo)
* D\_TEMP: differenza tra temperatura misurata a 25 metri dal suolo e a 2 metri dal suolo
* HOURS: numero di ore trascorse dalla mezzanotte del giorno di rilevazione

1. Si effettui un’analisi descrittiva e preliminare del dataset e si commentino i risultati
2. Si studi la correlazione tra le variabili numeriche del dataset e si commentino i risultati
3. Si studino i seguenti modelli di regressione
4. pm10 = cars temp wind
5. pm10 = cars temp wind D\_TEMP HOURS
6. Si effettui un’analisi sulla collinearità delle variabili inserite nel modello e si commentino gli opportuni indici
7. Si identifichino eventuali punti di influenza che impattano le performance del modello e si mostri il risultato graficamente
8. Leverage plot
9. Studentized residual
10. Cov Ratio plot
11. DFFITS plot
12. DFBETAS
13. Distanza di Cook
14. Si elimini l’unità 277 e si stimi il modello pm10 = cars temp wind . Si commentino i risultati comparando con il modello stimato al punto 3.a

**Per casa:**

1. Per il seguente modello completo si effettui
   1. un’analisi sulla normalità degli errori (Quantitativamente e graficamente)
   2. un’analisi sulla Omoschedasticità e Incorrelazione

**Esercizio 3 – SAS**

Il dataset ‘AEROBIC FITNESS PREDICTION

Le Variabili contenute nel dataset sono:

* Age: età (in anni)
* Weight: peso (in kg.)
* Oxygen: tasso di ossigenazione (ml/kg. al minuto)
* RunTime: tempo di percorrenza di 1.5 miglia (minuti)
* RestPulse: battito cardiaco a riposo
* RunPulse: battito cardiaco durante la corsa
* MaxPulse: battito cardiaco massimo durante la corsa

1. Si effettui un’analisi descrittiva e preliminare del dataset e si commentino i risultati
2. Si studi la correlazione tra le variabili numeriche del dataset e si commentino i risultati
3. Si studino i seguenti modelli di regressione
4. Oxygen= RunTime MaxPulse RunPulse
5. Oxygen= Age Weight RunTime RestPulse RunPulse MaxPulse
6. Si effettui un’analisi sulla collinearità delle variabili inserite nel modello e si commentino gli opportuni indici
7. Si identifichino eventuali punti di influenza che impattano le performance del modello e si mostri il risultato graficamente
8. Leverage plot
9. Studentized residual
10. Cov Ratio plot
11. DFFITS plot
12. DFBETAS
13. Distanza di Cook
14. Si elimini l’unità 10 e si stimi il modello Oxygen= RunTime MaxPulse RunPulse. SI commentino i risultati comparando con il modello stimato al punto 3.a
15. ripetere l'analisi sul dataset fitness ridotto eliminando l’osservazione anomala (unità 4)

**Per casa:**

1. Per il seguente modello (Oxygen=RunTime MaxPulse RunPulse) si effettui
   1. un’analisi sulla normalità degli errori (Quantitativamente e graficamente)
   2. un’analisi sulla Omoschedasticità e Incorrelazione

**Per Casa:**

1. Per il dataset 'fitness', si ripetano le analisi di cui sopra per il modello completo (mod2).
2. Per il Dataset 'companies', si proceda allo studio completo di diagnostica su più modelli costruiti a piacere.