UNIVERSITÀ DEGLI STUDI DI SALERNO

DIEM|A.A. 2021/2022

ANALISI DI REGRESSIONE DI UN DATASET

TRAMITE L'IMPIEGO DEL SOFTWARE R

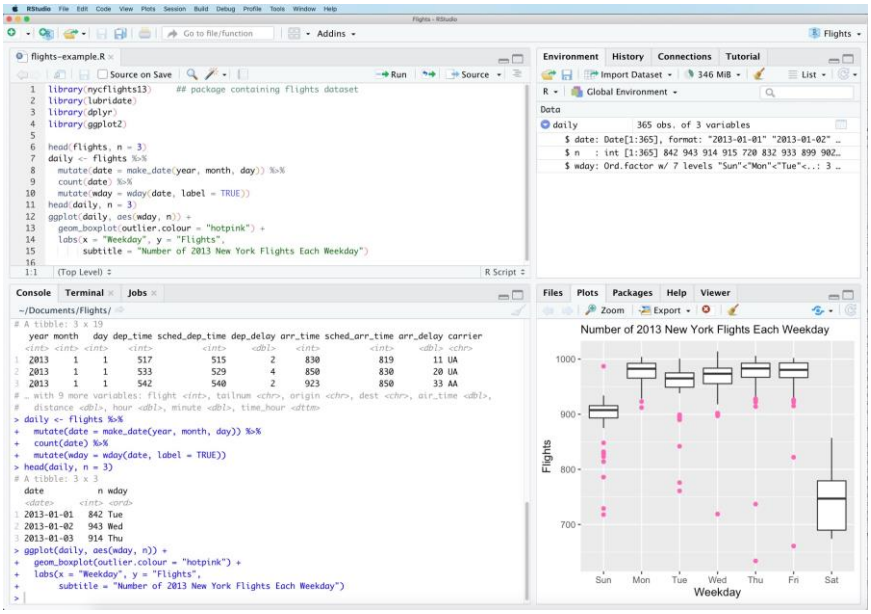
RELAZIONE A CURA DI:

LANZARA NICOLA
SARNO FABRIZIO
SQUITIERI BENIAMINO
VITALE ANTONIO

Obiettivo principale di questo progetto, eseguito in un gruppo di quattro componenti, è stato quello di analizzare la regressione correlata a un dataset inerente la scheda tecnica di un calcolatore. Sono state fondamentali le conoscenze acquisite durante il corso per l'analisi dei dati e la presentazione dei risultati come: la statistica descrittiva, l'analisi di correlazione, i test di ipotesi e altri aspetti che verranno poi approfonditi con il proseguire di questa relazione.

Software utilizzato:

Il software impiegato per tale lavoro è stato **RStudio**: un IDE per la programmazione utile al calcolo statistico e alla corrispettiva grafica. È in parte scritto in *C++* e utilizza il framework *Qt* per la sua interfaccia utente grafica. La maggior parte del codice è in *Java*, comprendendo anche *JavaScript* tra i linguaggi utilizzati. Nonostante le alte funzionalità di base presentate dal software, è ulteriormente possibile istanziare pacchetti aggiuntivi utili per svolgere operazioni non previste al momento della creazione, i quali permettono di rendere l'IDE in grado di soddisfare le nuove esigenze richieste dalle nuove tecnologie.



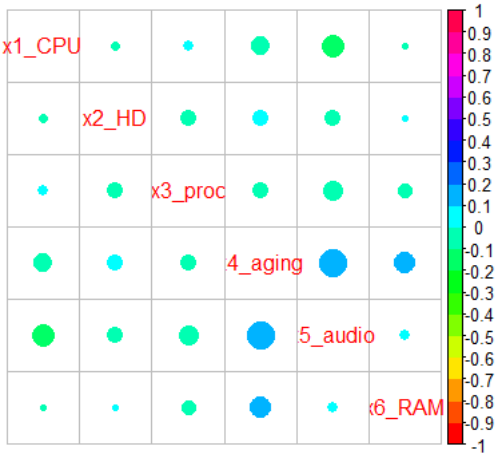
Analisi preliminare dei dati

La prima operazione effettuata è stata la **visualizzazione del dataset fornito**, in modo da avere una visione delle diverse variabili in esame e comprenderne il tipo, al fine di poter poi studiare meglio le varie relazioni di correlazione presenti tra esse. Dopo l'upload del file si è notato che tutte queste sono di tipo quantitativo discreto e constatata la presenza di una sola variabile dipendente, detta *Indice standardizzato e centrale delle prestazioni SW di calcolo*, e di sei invece indipendenti (ovvero i regressori) qui elencate:

- Indice standardizzato e centrato di velocità di CPU (variabile x_1)
- Indice standardizzato e centrato di dimensioni HD (variabile x_2)
- Indice standardizzato e centrato legato al numero di processi SW (variabile x_3)
- Indice standardizzato e centrato legato all'aging SW (variabile x_4)
- Indice standardizzato e centrato legato alle prestazioni della scheda audio (variabile x_5)
- Indice standardizzato e centrato legato alle prestazioni RAM (variabile x_6)

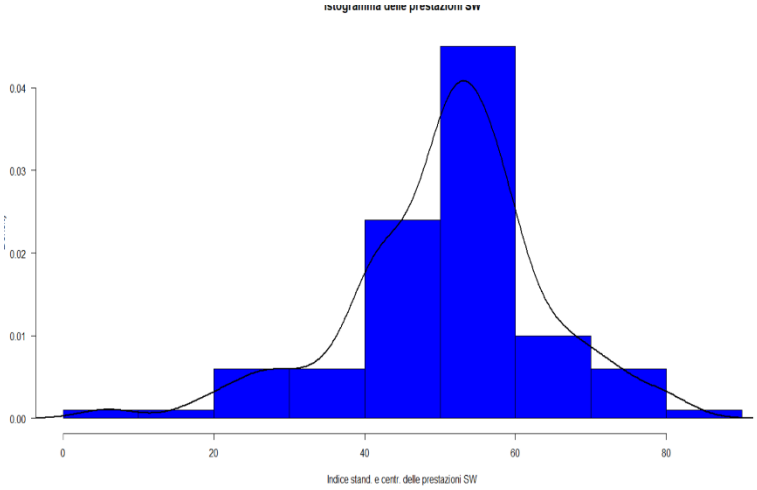
Matrice di correlazione

Successivamente si è studiata la **correlazione** tra le variabili indipendenti e quella dipendente tramite la matrice apposita rappresentata da uno **scatterplot**, al fine di vedere quanto le prime influissero sull'ultima. Da questa è stato possibile osservare come il primo, il terzo e il quarto regressore **influenzino molto** la *y*, a differenza degli altri, in quanto si presenta un indice di correlazione prossimo al **valore nullo**. In particolare il primo e il terzo incidono in maniera **positiva**, ovvero all'aumentare di una aumenta anche l'altra, mentre il quarto in maniera opposta dato che l'indice di correlazione presenta valori negativi. Tramite un'attenta osservazione è stato possibile notare come questo grafico rappresenti esattamente una **matrice simmetrica** in quanto leggendola per righe o per colonne il risultato non varia. Come materiale aggiuntivo è stata elaborata anche una rappresentazione grafica, grazie alla funzione *corrplot* di *RStudio*, riportata qui affianco.



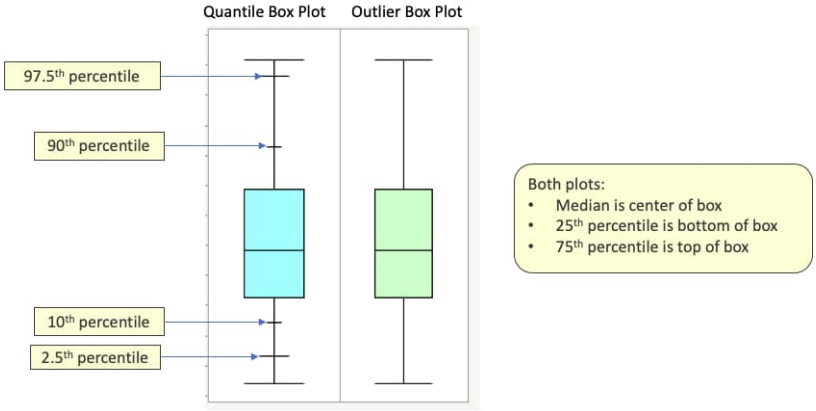
Istogramma

In aggiunta si è fatto uso dell'**istogramma**, con lo scopo di ottenere un ulteriore modo per attestare la relazione tra le variabili. Si tratta di uno dei grafici più conosciuti e utilizzati nell'ambito statistico ed è impiegato per **descrivere la distribuzione di una variabile quantitativa** con molti valori diversi tra loro. Nello specifico aiuta a **esplorare i dati**, **capire dove si trova il punto centrale** di una variabile e **quanta variabilità c'è** nei dati così come se ci sono dei **valori anomali** e se i dati sono **approssimativamente simmetrici** o **chiaramente asimmetrici**. Nello specifico, un istogramma della distribuzione di frequenza è un grafico cartesiano che si ottiene riportando, in corrispondenza di ciascuna classe, un rettangolo la cui altezza è proporzionale alla frequenza o alla frequenza relativa associata a quella classe. Nel caso in esame si sono integrati gli istogrammi con le varie classi relative ad ogni regressore, al fine di mettere in relazione ognuno di questi con la sua rappresentazione tabellare la quale, per motivi di praticità, non è conveniente da implementare nel software utilizzato. Sono stati poi calcolati i valori di grandezze quali la **media**, **mediana**, **varianza** e altre relative alla variabile di uscita *y*.



Boxplot

Infine come ulteriore metodologia di rappresentazione e analisi dei dati è stato utilizzato il **boxplot**, o *diagramma a scatola e baffi*. La linea centrale rappresenta la **mediana** degli stessi dati, la cui metà di questi si trova al di sopra di questo valore, l'altra metà al di sotto. Se i dati sono **simmetrici**, la mediana è al centro della scatola; viceversa sarà più



vicina alla parte superiore o inferiore se **asimmetrici**. Queste parti in particolare mostrano il venticinquesimo e il settantacinquesimo quantile, detto anche **percentile**, e sono chiamate in alternativa *quartili*, poiché ciascuno di essi esclude un quarto (25%) dei dati. La lunghezza della scatola è la differenza tra i due percentili e si chiama **range interquartile** (*IQR*). Le linee che si estendono sono chiamate **baffi** e rappresentano la variazione dei dati attesa estendendosi per 1,5 volte dall'*IQR* dalla parte superiore e inferiore. Se i dati non arrivano alla fine dei baffi, significa che questi si estendono fino ai valori di dati minimi e massimi. Se, al contrario, ricadono sopra o sotto la fine dei baffi allora sono rappresentati come punti denominati spesso **outlier**, i quali sono più estremi della variazione attesa. Vale la pena esaminare questi punti di dati per determinare se sono errori o outlier, cosa che i baffi non comprendono.
Successivamente si è osservato che soltanto per quanto riguarda il boxplot relativo alla relazione della variabile dipendente sono presenti degli outlier, il che permette di affermare che i regressori si presentano con valori abbastanza regolari tra di loro.

Valutazione della relazione tra la variabile dipendente e le singole variabili indipendenti

Per un'analisi approfondita, oltre all'applicazione e rappresentazione di forme grafiche come gli scatterplot, si è deciso di ricorrere alla **regressione polinomiale** tra la variabile dipendente e le singole variabili indipendenti contenute nel dataset. Questa utilizza lo stesso metodo della regressione lineare assumendo che la funzione che meglio descrive l'andamento dei dati non sia una retta, bensì un **polinomio**. Quindi è adatta quando lo scatterplot di una **relazione bivariata**, ad esempio, mostra una forma diversa da quella della retta. Si basa sul collegamento presente tra l'uscita e, nel caso in esame, i molteplici ingressi esistenti presentati in maniera alternata; questi per giunta verranno poi elevati a gradi superiori al primo per analizzare l'incidenza che hanno proprio sulla *y*.
Dopo aver rimosso dal dataset tutti gli eventuali valori non disponibili (NA, *Not Available*), è stata utilizzata la funzione *lm* fornita dall'ambiente di programmazione la quale tornerà utile per il continuo dell'esecuzione del progetto. La relazione tra variabile dipendente e la/le variabili d'ingresso è specificata dalla tilde ~. Come output, sono restituite le principali **statistiche descrittive** sui residui, il **test di significatività** sui parametri e infine **gli indici di adattamento e di significatività** dell'intero modello. La tabella *Coefficients* riporta, per ogni parametro: il valore *Estimate*, l'errore standard *Std. Error*, il valore della statistica *t* (*t value*) e il *p-value* (*Pr(>|t|)*).
Si è poi divisa la valutazione nelle sei casistiche legate alle variabili *x* fornite dal dataset:

1) Per il **primo regressore**, ovvero l'indice standardizzato e centrato di velocità CPU, si è potuto notare dalla regressione polinomiale l'elevata incidenza che questo ha avuto sull'indice standardizzato e centrato delle prestazioni SW di calcolo. Tale robustezza viene indicata dalla presenza di tre asterischi i quali, consultando la legenda della funzione *summary*, hanno confermato tale supposizione.
   Analizzando consecutivamente le relazioni presenti per la variabile x_1 elevata al quadrato, al cubo e alla quarta (si è imposto tale limite in quanto con gradi superiori ci si poteva avventare in un crollo del software), si è potuto concludere dicendo che tra queste solo la x_1^3 ha una minima rilevanza per il valore della *y*.
2) Per il **secondo regressore**, così come per i rimanenti, il procedimento svolto è stato identico a quello precedente. La differenza è stata però nell'assenza totale di relazione con la variabile di uscita per tutti i gradi in cui si è svolta tale analisi.
3) Per l'indice standardizzato e centrato legato al numero di processi SW, ovvero il **terzo regressore**, si è registrata la particolarità di **un'influenza massima al primo grado** per poi diminuire a favore della x_3 di grado due. Questa infine, con l'aggiunta dei regressori di terzo e quarto grado, è andata ad **annullarsi completamente** lasciando la *y* del tutto indipendente da tale variabile.

4) Per la x_4 l'influenza sulla variabile di uscita era al massimo per il primo e il secondo grado, per poi calare al minimo con l'aggiunta del regressore di grado tre e quello di grado quattro. Un comportamento simile a quello della variabile precedente.
5) Con il **quinto regressore** si è registrato uno scenario equivalente al secondo, ovvero l'assenza totale di incidenza sulla variabile dipendente.
6) Per **l'ultima variabile d'ingresso x_6** l'andamento è stato opposto a quello di x_3 e x_4: una relazione con la *y* **nulla per i primi gradi polinomiali** (primo e secondo) per poi presentarsi, seppur con incidenza minima, con l'aggiunta dei regressori elevati alla terza e alla quarta.

In conclusione, le uniche due variabili d'ingresso che **non** hanno avuto alcuna risonanza su quella di uscita sono state x_2 e x_5. Tutte le altre invece, in maniera più o meno marcata, sono state rilevanti per il valore della *y* all'interno del dataset e quindi dei valori prodotti.

Definizione del modello statistico dei dati

È noto che un modello di regressione lineare multipla è ottimale quando si ha la **presenza dei soli regressori che effettivamente sono correlati con la variabile dipendente** e inoltre di un solido coefficiente di determinazione R^2, ovvero un indice di misura che informa della **bontà della previsione**. Questo indica la proporzione di varianza totale dei valori di *y* intorno alla media della stessa variabile che risulta spiegata dal modello di regressione. Proprio perché una proporzione, il suo valore è sempre compreso tra 0 ed 1, oppure tra lo 0% e il 100% se lo si vuole esprimere in termini percentuali.

- R^2**=0** indica un modello le cui variabili predittive non spiegano per nulla la variabilità della *y* intorno alla sua media. Se invece delle variabili indipendenti inserite nel modello si utilizzasse solo la **media della *y*** si otterrebbe in pratica lo stesso valore esplicativo. Questa situazione si verifica quando le *y* stimate dal modello **coincidono** esattamente con la media di *y*. In questo caso anche il corrispondente indice di correlazione è nullo.
- R^2***=1*** indica un modello le cui variabili indipendenti riescono a spiegare completamente la variabilità della *y* intorno alla sua media. Ovvero, conoscendo i valori delle *x* si può prevedere esattamente quale sarà quello della *y*. Questa situazione si verifica solo quando nel grafico a dispersione tutti i punti si collocano esattamente sulla retta di regressione. Quando R^2 è uguale ad 1 infatti anche l'indice di correlazione *r* risulta essere ±1. In questo caso non c'è quindi nessun errore nell'utilizzare *x* per prevedere *y*. In altre parole, i valori osservati della *y* **coincidono** esattamente **con quelli stimati** dal modello. Di solito, più è grande il valore di R^2, più il modello ha un alto potere predittivo. Tuttavia, questa interpretazione in alcune situazioni può essere fuorviante: un modello che presenta un valore alto di R^2 può infatti essere comunque errato.
- In alcuni casi non si può ottenere un R^2 abbastanza elevato. Ciò non dipende dal fatto che il modello di regressione costruito non vada bene ma solo che, per sua natura, la variabile dipendente analizzata sia legata a tantissimi altri fattori, molti dei quali non misurati. D'altra parte un R^2 elevato è **condizione necessaria ma non sufficiente** per poter effettuare delle previsioni precise.
- Un ulteriore problema di R^2 è che aumenta ogni volta che si aggiunge una variabile indipendente al modello, anche se questa variabile non è per nulla esplicativa. Non è infatti possibile spiegare meno della variazione osservata per la variabile dipendente aggiungendo delle variabili esplicative al modello. Per evitare questa situazione, nei modelli di regressione con molte variabili indipendenti si preferisce interpretare il valore di R^2 *corretto* e di R^2 *predetto* (*Adjusted*). Inoltre, se si costruisce una curva che si adatta "troppo" ai dati (ad esempio utilizzando dei termini polinomiali) si ottiene probabilmente un modello con un coefficiente di determinazione molto alto. Tuttavia, un modello che si adatta troppo ad uno specifico set di dati, seguendone ogni minima variazione, risulta poi **poco generalizzabile** e con **basso potere predittivo**. In statistica, in questi casi, si parla di problemi di *Overfitting*.

Consapevoli di queste premesse, si è passati alla realizzazione del modello tramite la funzione *lm* inserendo al suo interno tutti i regressori e la variabille dipendente; dando vita al **modello base**. Dall'analisi associata si è evinto che il **secondo** e il **quinto regressore non influenzassero la variabile dipendente**, come era stato già dedotto dall'analisi di correlazione precedente, motivo per il quale sono stati eliminati. Successivamente si è costruito un secondo modello di regressione nel quale sono stati aggiunti ulteriori termini determinati dai regressori che influivano maggiormente sulla variabile *y*. Il motivo principale di questa scelta era quello di aumentare il valore di R^2 *corretto* per raggiungere gli obiettivi sopra indicati. Si è notata tuttavia ancora la presenza degli stessi regressori non correlati con la variabile dipendente presenti al punto precedente; quindi, nonostante il valore di R^2 *corretto* fosse migliorato, si è potuto dedurre che esistesse un **ulteriore modello migliore** di quello appena ottenuto. Sulla base di questi ragionamenti, tramite un approccio del tipo *Hybrid Stepwise* si è giunti al modello di regressione ottimale - denominato come *fit4* - all'interno del file inerente il progetto nel quale sono stati ottenuti tutti i risultati prefissati.

Stima ai minimi quadrati dei parametri del modello e determinazione degli intervalli di confidenza
Derivante dall'applicazione della funzione *lm* nel terzo punto è la digressione fatta all'interno di questo. Per motivi puramente legati all'approfondimento del concetto e in certi aspetti anche per la successiva applicazione, sono stati calcolati tramite opportune funzioni fornite dall'ambiente *RStudio* i seguenti valori:

- Lista attributi dell'oggetto
- Vettore dei coefficienti di regressione
- Devianza dei residui
- Vettore dei residui del modello
- Valori della risposta stimati dal modello
- Matrice X del modello di regressione
- Correlazioni tra i coefficienti
- Matrice delle varianze e covarianze dei coefficienti di regressione

Dopo aver stimato il modello di regressione è stato necessario verificare che le ipotesi di base legate a questo fossero valide, tramite opportuni test statistici. Di questi ne esistono quattro, ed è necessario che tutti siano considerati **superati** per far sì che la **stima ai minimi quadrati** (*OLS*) possa essere calcolata; altrimenti è richiesta l'applicazione di un metodo differente. Vengono elencati qui di seguito:

1) **Test *t di Student*** (oppure ***t test***): è uno strumento per valutare le medie di una o due popolazioni tramite verifica d'ipotesi. Il test può essere usato per determinare se un singolo gruppo differisce da un valore conosciuto (*test t* a un campione), se due gruppi differiscono l'uno dall'altro (*test t* a due campioni indipendenti), o se c'è una differenza significativa nelle misure appaiate (*test t* a campioni dipendenti, o appaiati). In questo caso si verifica che la media degli errori non sia significativamente diversa da zero tramite la condizione che il valore di *p-value* sia **maggiore di quello limite**, il quale nel caso di *RStudio* è impostato automaticamente ad α=0,05. Per il modello preso in esame l'ipotesi è stata verificata, quindi il test superato.
2) **Test di *Shapiro-Wilk***: è uno dei più potenti per la **verifica della normalità**, soprattutto per piccoli campioni. Si tratta di un test per la verifica di ipotesi statistiche. Viene analizzato il valore prodotto *W* e verificato che questo sia **molto vicino** a 1. Nel modello scelto l'ipotesi nulla è stata accettata; anche questo test dunque è stato superato.
3) **Test di *Breusch-Pagan*** (o di ***Cook-Weisberg***): è un test d'**ipotesi di omoschedasticità** in un modello di regressione lineare. Prende in ipotesi che il valore di *p* prodotto sia **maggiore della soglia del 5%** affinché possa essere superato. Condizione che è venuta soddisfatta dal modello analizzato.
4) **Test di *Durbin-Watson***: è un test statistico utilizzato per rilevare la **presenza di autocorrelazione** dei residui in un'analisi di regressione. Il valore risultante *d* deve essere **compreso** tra 0 e 4, nello specifico **molto vicino** a 2, affinché si possa affermare l'assenza di

correlazione all'interno del modello scelto. Nel caso elencato, anche questo test è potuto dirsi passato.

Tutti i test di specificazione del modello hanno dato esito positivo, di conseguenza si è potuto affermare che le ipotesi alla base del modello di regressione *OLS* (Stima ai minimi quadrati) fossero valide. Successivamente si è proceduto alla **determinazione degli intervalli di confidenza**. Trattasi questi di alcuni intervalli ottenuti mediante una procedura che ha, nella ripetizione dell'esperimento, una probabilità pari a 1− α di generarne uno che contiene il valore vero di un fissato parametro incognito. Possono essere calcolati per i regressori coinvolti nel modello oppure per la variabile dipendente *y* a seconda della funzione utilizzata: *confint* o *predict*. Le specifiche richiedevano un calcolo degli intervalli per i regressori x_1, x_2, x_3… quindi è stata utilizzata la prima funzione sia al 95% che al 99%.

	2.5 %	97.5 %
(Intercept)	55.39652912	57.877184
data$x1_CPU	5.64703367	7.329500
data$x3_proc	2.49973522	4.270414
I(data$x3_proc^2)	-6.73840092	-4.874236
data$x4_aging	-7.05224961	-5.338054
data$x5_audio	-0.07072542	1.646124
data$x6_RAM	1.20725848	2.885099

	0.5 %	99.5 %
(Intercept)	54.9943316	58.279382
data$x1_CPU	5.3742493	7.602284
data$x3_proc	2.2126488	4.557500
I(data$x3_proc^2)	-7.0406446	-4.571992
data$x4_aging	-7.3301783	-5.060125
data$x5_audio	-0.3490844	1.924483
data$x6_RAM	0.9352241	3.157134

Training set e Test set

Dal momento che nell'analisi dei dati si ha l'esigenza di avere un modello in grado di fare previsioni accurate dei dati, una buona tecnica per costruirlo consiste nell'usare la sequenza di fasi/operazioni che prendono il nome di *Training Set*, *Validation Set* e *Test Set*.

- **Training Set**: consiste nel prendere una parte dei dati appartenenti e **stressarli**, ovvero allenarli a predire per fare in modo che il modello apprenda la relazione tra le variabili indipendenti e quella dipendente. Questa fase nota anche con il nome di *Learning* è il cuore della maggior parte dei processi statistici e si trova alla base del Machine Learning in quanto i modelli di IA osservano, studiano, analizzano e poi cercano di fare previsioni su quanto imparato. Questa fase nonostante tutto non si esime dall'essere senza errori: bisogna sempre far attenzione a non ricadere nell'*Overfitting* accennato in precedenza, ed è qui che entra in gioco la fase seguente.
- **Validation Set**: Nel *Training Set* quindi il modello ha memorizzato le relazioni tra input e output. Per evitare l'*Overfitting* ed avere quindi una reale capacità predittiva, si forniscono a questo dei dati che non ha mai visto e gli si comunica di compiere una previsione. Questi dati devono essere necessariamente etichettati, ovvero devono essere esattamente come quelli del set precedente. A questo punto si confrontano gli output predetti con quelli reali per vedere quanto il modello riesca a prevedere con buona approssimazione la variabile di output.
- **Test Set**: Infine, una volta ottenute delle buone performance sul *Training Set* e soprattutto sul *Validation Set*, è possibile testare il modello su altre osservazioni che quest'ultimo non ha mai visto.

Calcolo del coefficiente di determinazione e grafici diagnostici dei modelli

Sebbene il valore di R^2 fosse calcolato dalla funzione *lm* e stampato a video tramite la *summary* come informazione rilevante correlata, è stata comunque applicata la formula per determinarlo. Questa risulta essere:

$$R^2=SQR/SQ_{TOT}$$

Dove

$$SQR=SQ_{TOT}-SQE$$

Tale approccio è stato scelto anche come verifica del valore del coefficiente derivante dalle informazioni citate in precedenza, avendo riscontro positivo in quanto i due valori si equivalevano.

In alternativa, si è utilizzata la funzione apposita fornita da *RStudio* chiamata *rsquare* e implementabile grazie alla libreria *modelr*. I grafici diagnostici sono stati elaborati dalla stessa funzione *plot* per ogni modello presente nel progetto, e riportati all'ultima pagina del presente documento. Si consiglia comunque una visualizzazione migliore all'interno dello stesso ambiente di programmazione.

Scelta del modello che meglio si adatta ai dati forniti

Per quanto riguarda la scelta del modello esistono vari tipi di criteri che permettono di sceglierne uno rispetto a un altro, i più utilizzati sono *BIC* e *AIC*. Grazie a questi si è infatti preso come ottimale il modello con valore minimo tra tutti quelli analizzati:

- *AIC* (*Akaike Information Criterion*)*=2k-2logn(L)*, dove *k* è il **numero di parametri** nel modello statistico e *L* il **valore massimizzato della funzione di verosimiglianza** del modello stimato. Fornisce una misura della qualità della stima di un modello statistico tenendo conto sia della bontà di adattamento che della complessità del modello. È basato sul concetto di entropia come misura di informazione, tramite cui valuta la **quantità di informazione persa quando un dato modello è usato per descrivere la realtà**, inoltre la regola su cui si basa è quella di preferire i modelli con l'AIC più basso: un criterio di valutazione molto utile perché **permette di confrontare tra loro anche modelli non annidati**. È usato perché consente di selezionare il modello non conosciuto più vicino a quello reale ed è di facile interpretazione, mentre per quanto riguarda i suoi svantaggi si può dire che non è consigliato per campioni di grandi dimensioni.
- *BIC* (*Bayesian Information Criterion*)*=-2logn(L)+k*logn(N)*. Questo criterio, come si evince dalla formula, non è relazionato con il numero di variabili presenti nel modello ma dalla **dimensione dei dati** in esame. Come nel caso dell'AIC, si tende a selezionare il modello con il valore minore quando si deve scegliere quello vero che rappresenta i dati. Per cui se tale modello si trova tra quelli testati il criterio seleziona esattamente quello. Inoltre questo criterio è utile quando si deve **selezionare un modello con pochi parametri**, dato che questi non aumentano insieme alla dimensione del campione. Nonostante ciò come per l'AIC anche il BIC presenta uno svantaggio che si manifesta nel momento in cui il vero modello non è presente tra quelli testati dato che il criterio ha un comportamento imprevedibile.
- *MSE* (*Mean Square Error*, Errore Quadratico Medio): Indica la **discrepanza quadratica tra i valori dei dati osservati e i valori di quelli stimati**. Si calcola come:

  $$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}\left[(\hat{\theta} - \theta)^2\right]$$

  La sua radice quadrata rappresenta un ulteriore indice statistico detto *RMSE* (*Root Mean Square Error*, Radice dell'Errore Quadratico Medio) oppure *RMSD* (*Root Mean Square Deviation*) corrispondente alla **varianza interna**. Nell'ambito della scelta di un modello si è orientati a scegliere quello rappresentato da un valore di questo indice minore rispetto ad altri, in quanto rappresenta il modello con la minor discrepanza.

Osservando i risultati proposti dall'applicazione di ogni criterio sullo stesso modello si è potuto asserire che la scelta è ricaduta sul quarto.
Un'importante osservazione va fatta però per quanto riguarda il modello che valutava il minimo *MSE*: il criterio preso singolarmente avrebbe portato a preferire il secondo rispetto agli altri dato che registrava il minor valore di tutti. Tuttavia, sulla base di quanto constatato nei punti precedenti, quest'ultimo possedeva anche un elevato numero di regressori con il conseguente eccessivo avvicinarsi all'*Overfitting*. Si è dunque ripiegati su *fit4*, ovvero il modello con il maggior equilibrio tra i due valori.

Per quanto riguarda l'algoritmo di selezione per la determinazione automatica del modello, si è utilizzato quello noto come *Stepwise* (in tutte le tre varianti *Hybrid*, *Forward* e *Backward*) tramite la funzione *step* la quale ha confermato il risultato ricavato, sottilineando il fatto che non fosse possibile migliorare ulteriormente quello ottenuto in precedenza nel terzo punto.

Conclusioni

Grazie alle conoscenze acquisite e all'ausilio dell'ambiente di programmazione, il lavoro di analisi compiuto sul dataset assegnato è risultato essere stimolante e soddisfacente. Non negando qualche difficoltà iniziale legata all'implementazione di nuove librerie e alla comprensione della logica di alcune funzioni aggiuntive utilizzate autonomamente, c'è da affermare che il progetto è stato di maneggevole complessità e ha spinto ogni membro del gruppo a impegnarsi per la sua buona riuscita, cercando di trovare sempre un fattore aggiuntivo che potesse sposarsi al meglio con lo spirito di approfondimento legato all'intero lavoro e anche ad una presentazione migliore dei dati risultanti.

Grafici diagnostici dei modelli ottenuti