

UNiversità degli studi di Salerno |DIEM – A.A. 2021/2022

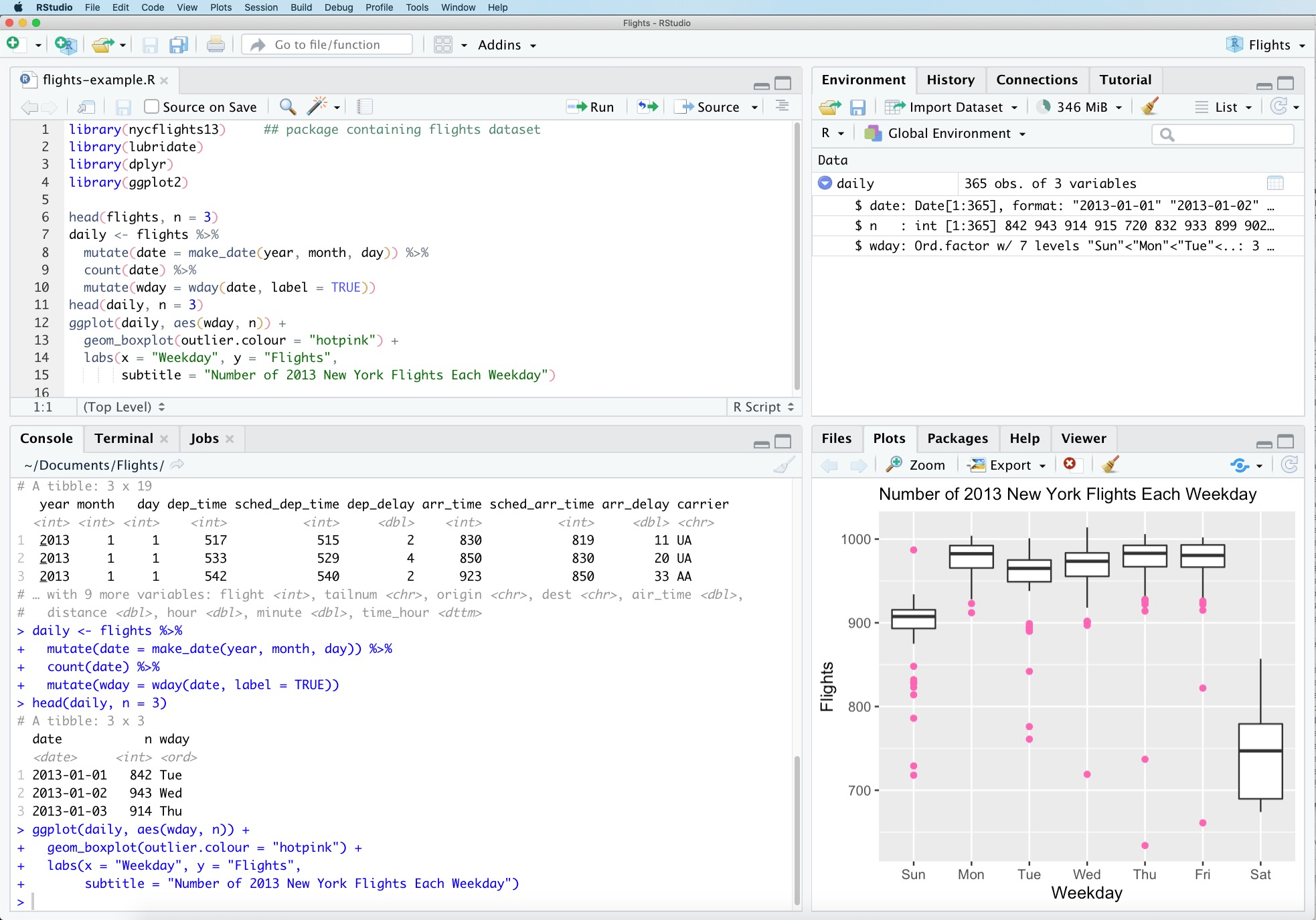
utilizzo della statistica per l’analisi di un elaboratore

LANZARA NICOLA, SARNO FABRIZIO, SQUITIERI BENIAMINO, VITALE ANTONIO

Analisi di regressione di un dataset tramite software R

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 nel corso di questa relazione.

**Software utilizzato:**

Il software utilizzato è 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 funzionalità eventualmente non previste agli albori, il che permette 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 fornitoci, in modo da visualizzare le varie variabili in esame e capirne il tipo, al fine di poter studiare meglio le varie relazioni di correlazione presenti tra esse. Dopo aver caricato il dataset si è notato che tutte le variabili in esame 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 variabili indipendenti (ovvero i regressori) :

* Indice standardizzato e centrato di velocità di CPU
* Indice standardizzato e centrato di dimensioni HD
* Indice standardizzato e centrato legato al numero di processi SW
* Indice standardizzato e centrato legato all’aging SW
* Indice standardizzato e centrato legato alle prestazioni della scheda audio e infine
* Indice standardizzato e centrato legato alle prestazioni RAM

**Matrice di correlazione**

Successivamente si è studiato la correlazione tra le varie variabili indipendenti e quella dipendente tramite la matrice di correlazione rappresentata tramite scatterplot, al fine di vedere quanto le prime influiscano sull’ultima. Da quest’ultima possiamo notare come il primo, il terzo e il quarto regressore influenzino molto la variabile dipendente, mentre le rimanenti non tanto in quanto si presenta un indice di correlazione prossimo al valore nullo. In particolare il primo e il terzo regressore incidono sulla *y* in maniera positiva, ovvero all’aumentare di una aumenta anche l’altra, mentre il quarto in maniera negativa dato che l’indice di correlazione presenta valori negativi. Tramite un’attenta osservazione è possibile notare come questo scatterplot rappresenti esattamente una matrice simmetrica in quanto leggendola per righe o per colonne il risultato non varia.

**Istogramma**

In aggiunta si è fatto uso della rappresentazione tramite istogramma, con lo scopo di ottenere un ulteriore modo per attestare la relazione tra le variabili in esame. 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 relativa rappresentazione tabellare la quale, per motivi di praticità, non è conveniente da implementare nel software utilizzato.

**Boxplot**

Infine come ulteriore metodo di rappresentazione e analisi dei dati è stato utilizzato il boxplot. La linea centrale nella scatola rappresenta la mediana dei dati. La metà dei dati si trova sopra questo valore, l'altra metà sotto. Se i dati sono simmetrici, la mediana è al centro della scatola. Se, invece, i dati sono asimmetrici, la mediana sarà più vicina alla parte superiore o a quella inferiore della scatola. La parte inferiore e superiore della mostrano il venticinquesimo e il settantacinquesimo quantile, o percentile. Questi due quantili sono chiamati anche 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 a partire dalla scatola sono chiamate baffi. I baffi rappresentano la variazione dei dati attesa e si estendono per 1,5 volte dall'IQR dalla parte superiore e inferiore della scatola. Se i dati non arrivano fino alla fine dei baffi, significa che questi si estendono fino ai valori di dati minimi e massimi. Se, invece, i dati ricadono sopra o sotto la fine dei baffi, sono rappresentati come punti denominati spesso outlier. Un outlier è più estremo della variazione attesa. Vale la pena esaminare questi punti di dati per determinare se sono errori o outlier. I baffi non comprendono gli outlier. Successivamente si è osservato che soltanto per quanto riguarda il boxplot relativo alla relazione della variabile dipendente sono presenti degli outliers , 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 svolgere tale compito, 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, ma assume 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 per giunta tornerà utile per il continuo dell’esecuzione del progetto. La relazione tra variabile dipendente e la/le variabili indipendenti è specificata dalla tilde ~. Come output, verranno 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*) e il valore della statistica *t* (*t value*) e il *p-value* (*Pr(>|t|)*).  
Si dividerà 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 può notare dalla regressione polinomiale l’elevata incidenza che questo ha 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*, confermano tale affermazione.  
   Analizzando consecutivamente le relazioni presenti per la variabile *x1* elevata al quadrato, al cubo e alla quarta (si è imposto tale limite in quanto con gradi superiori si rischiava un crollo del software), si può concludere dicendo che tra queste solo la *x13* ha una minima rilevanza per il valore della *y*.
2. Per il secondo regressore, il procedimento svolto è stato identico a quello precedente. La differenza abissale rispetto alla *x1* è però l’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 *x3* di grado due. Questa infine, con l’aggiunta dei regressori di terzo e quarto grado, è andata ad annullarsi completamente lasciando la *y* completamente indipendente da tale variabile di ingresso.
4. Per la *x4* 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 questo simile alla variabile precedente.
5. Con il quinto regressore si è registrato un comportamento equivalente al secondo, ovvero l’assenza totale di incidenza sulla variabile dipendente.
6. Per l’ultima variabile d’ingresso *x6* l’andamento è stato opposto a quello di *x3*e *x4*: 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 *x2* e *x5*. Tutte le altre invece, in maniera marcata o meno, 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, quindi la influenzano, e inoltre anche della presenza di un coefficiente di determinazione R2 ottimale, 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 è una proporzione, il suo valore sarà sempre compreso tra 0 ed 1, oppure tra lo 0% e il 100% se lo si vuole esprimere in termini percentuali.

* R2=0(implica anche SQR=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 gli *y* stimati dal modello coincidono esattamente con la media di *y*. In questo caso anche il corrispondente indice di correlazione sarà nullo.
* R2=1(implica anche SQE=0) indica un modello le cui variabili indipendenti riescono a spiegare completamente la variabilità della *y* intorno alla sua media. Ovvero, conoscendo i valori delle variabili indipendenti si può predire esattamente quale sarà il valore della *y*. Questa situazione si verifica solo quando nel grafico a dispersione tutti punti si collocano esattamente sulla retta di regressione. Quando R2 è uguale ad 1 infatti anche l’indice di correlazione *r* sarà uguale ad 1 oppure a -1. In questo caso non c’è quindi nessun errore di previsione nell’utilizzare *x* per prevedere *y*. In altre parole, i valori osservati della *y* coincidono esattamente con i valori della *y* stimati dal modello. Di solito, più è grande il valore di R2, 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 R2 può infatti essere comunque errato.
* In alcuni casi non si può ottenere un R2 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 dipendente da tantissimi altri fattori, molti dei quali non sono stati misurati. D’altra parte un R2 elevato è condizione necessaria ma non sufficiente per poter effettuare delle previsioni precise.
* Per capire se il coefficiente di determinazione è statisticamente significativo si deve guardare il  
  *p-value* della statistica F. Il modello di regressione costruito infatti ha senso solo c’è almeno una variabile indipendente che riesca a spiegare i valori della *y*. Se il *p-value* relativo al test F è molto basso (spesso si considera come soglia α=0,05), allora si può affermare che R2 è statisticamente significativo. Se invece il valore del *p-value* del test F è oltre la soglia prefissata allora si afferma il contrario.
* Un ulteriore problema di R2 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 R2 corretto e di R2 predetto (*Adjusted*). Inoltre, se si costruisce una curva che si adatta “troppo” ai dati (ad esempio utilizzando dei termini polinomiali) si otterrà 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 poi passati alla realizzazione del modello. Come primo passo si è costruito un modello di regressione lineare tramite la funzione “lm" inserendo al suo interno tutti i regressori e la variabille dipendente (dando vita al modello base). Dal modello risultante si evince che il secondo e il quinto regressore non influenzano la variabile dipendente, come era stato gia 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 nella determinazione della variabile *y*. Il motivo principale di questa scelta era quello di aumentare il valore di R2 corretto, per i motivi elencati precedentemente. Nel secondo si evince però ancora la presenza degli stessi regressori non correlati con la variabile dipendente presenti al punto precedente, quindi nonostante il valore di R2 corretto sia migliorato, si può dedurre che esista un ulteriore modello ancora più ottimale di quello appena ottenuto. Sulla base di questi ragionamenti, tramite un approccio del tipo Hybrid stepwise si è giunti al modello di regressione migliore 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**

**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 per predire per fare in modo che il modello imparerà 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, in quanto bisogna sempre far attenzione a non ricadere nell’*Overfitting* accennato in precedenza, ed è qui che entrerà in gioco la fase seguente.
* Validation Set: Nel Training Set quindi il modello ha imparato 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 basterà confrontare 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 che il nostro modello riuscirà ad avere 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.

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 R 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

Inoltre viene calcolato il valore dell’Errore Quadratico Medio (MSE, *Mean Square Error*), corrispondente a quello dei residui.  
Dopo aver stimato il modello di regressione è necessario verificare che le ipotesi di base legate a questo siano 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 t 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 viene verificata, quindi il test è considerato superato.
2. Test di *Shapiro-Wilk*: è uno dei test più potenti per la verifica della [normalità](https://it.wikipedia.org/wiki/Variabile_casuale_normale), soprattutto per piccoli campioni. Si tratta di un [test per la verifica di ipotesi statistiche](https://it.wikipedia.org/wiki/Test_di_verifica_d%27ipotesi). Viene analizzato il valore prodotto *W* e verificato che questo sia molto vicino a 1. Nel modello scelto l’ipotesi nulla viene accettata; anche questo test può definirsi superato.
3. Test di *Breusch-Pagan* (o di *Cook-Weisberg*): è un [test d'ipotesi](https://it.wikipedia.org/wiki/Test_di_verifica_d%27ipotesi) di [omoschedasticità](https://it.wikipedia.org/wiki/Omoschedasticit%C3%A0) in un modello di [regressione lineare](https://it.wikipedia.org/wiki/Regressione_lineare). Prende in ipotesi che il valore di *p* prodotto sia maggiore della soglia del 5% affinché possa essere superato. Condizione che viene superata dal modello analizzato.
4. Test di *Durbin-Watson*: è un test statistico utilizzato per rilevare la presenza di [autocorrelazione](https://it.wikipedia.org/wiki/Autocorrelazione) dei residui in un'analisi di [regressione](https://it.wikipedia.org/wiki/Regressione_lineare). 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, tale condizione viene soddisfatta e anche questo test può dirsi superato.

Tutti i test di specificazione del modello hanno dato esito positivo, di conseguenza si può affermare che le ipotesi alla base del modello di regressione OLS (Stima ai minimi quadrati) sono valide.  
Successivamente si è proceduto alla determinazione degli intervalli di confidenza, i quali 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 *x1*, *x2*, *x3*… quindi è stata utilizzata la prima funzione sia per il 95% che il 99%.

Immagine che contiene testo

Descrizione generata automaticamenteImmagine che contiene testo

Descrizione generata automaticamente

**Calcolo del coefficiente di determinazione e grafici diagnostici dei modelli**

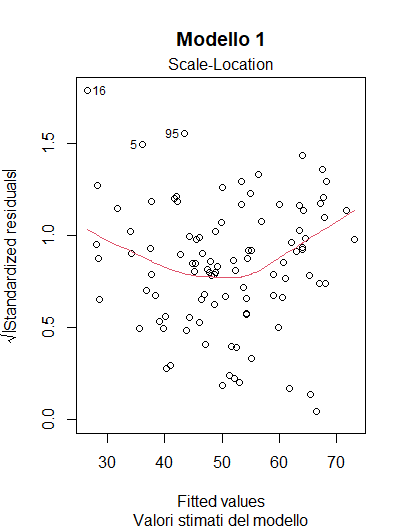
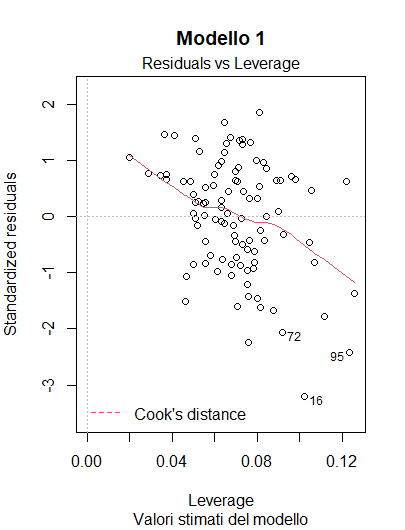
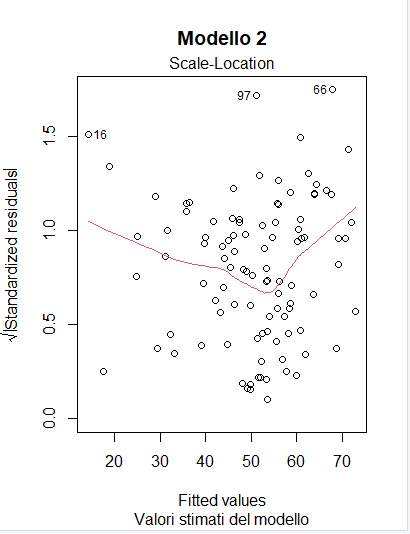
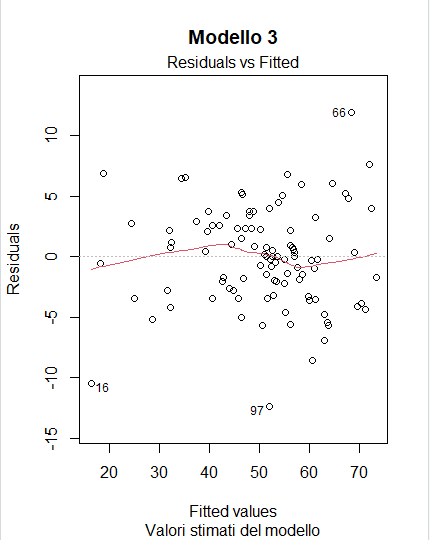
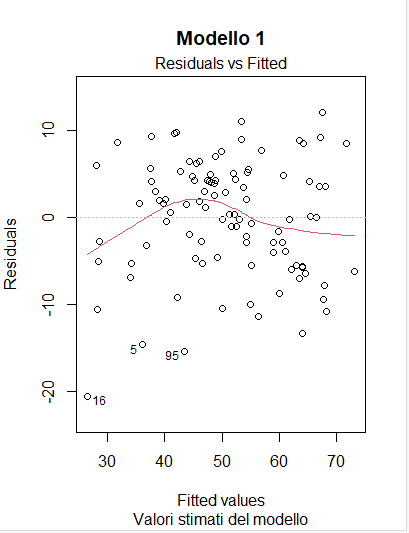
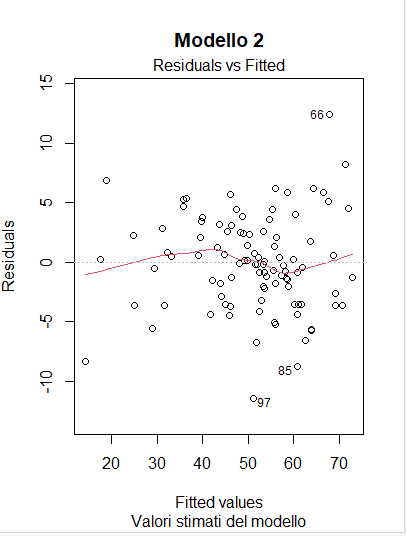
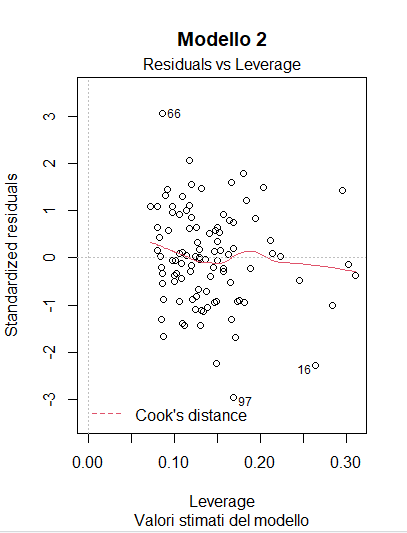
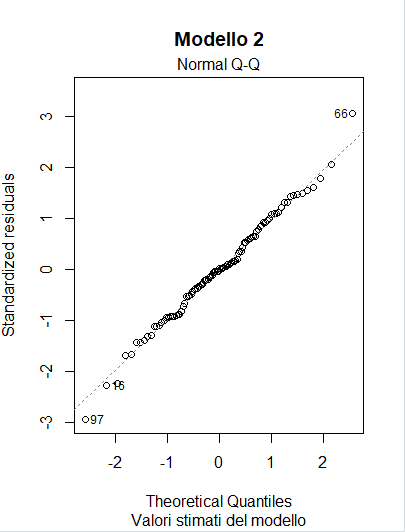
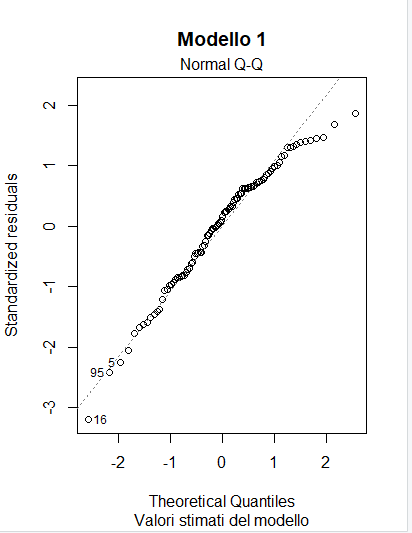
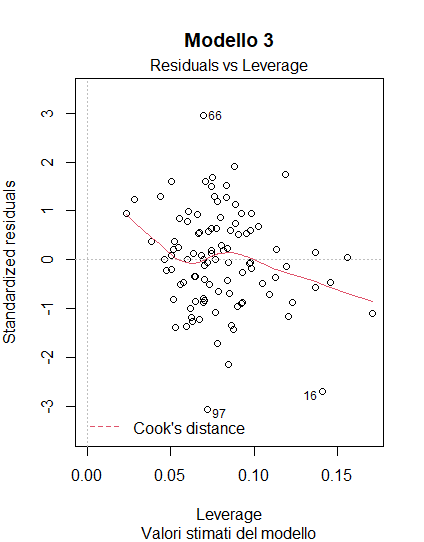
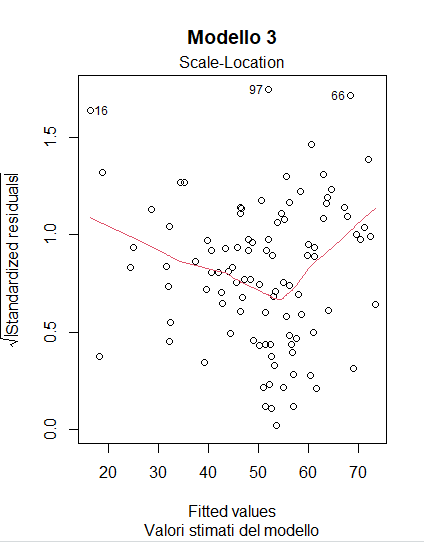
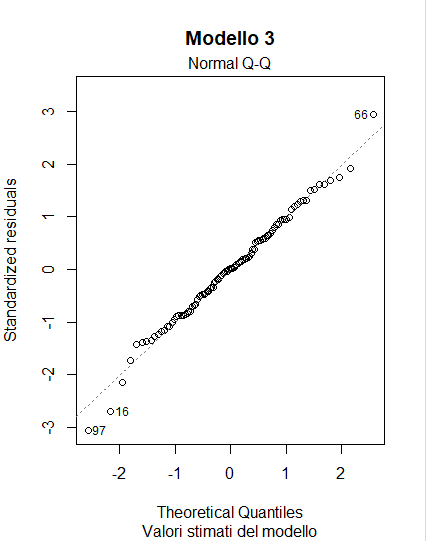
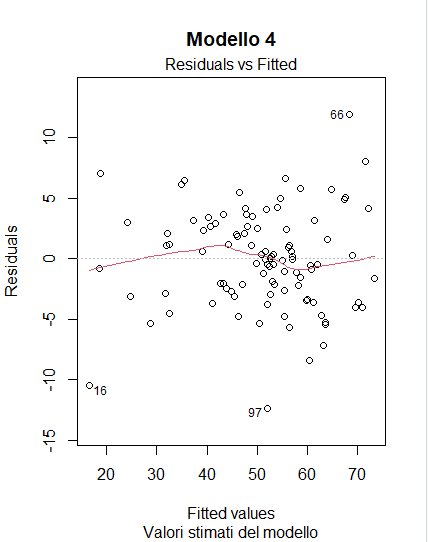
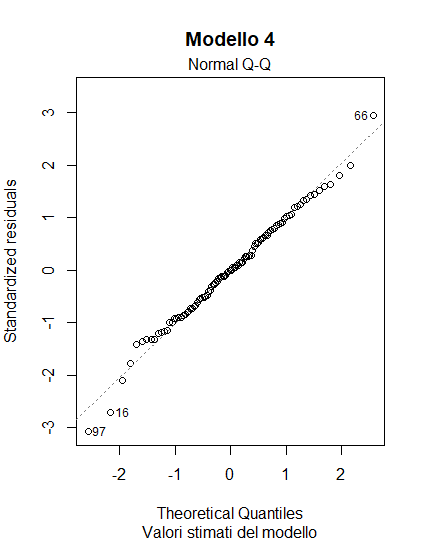
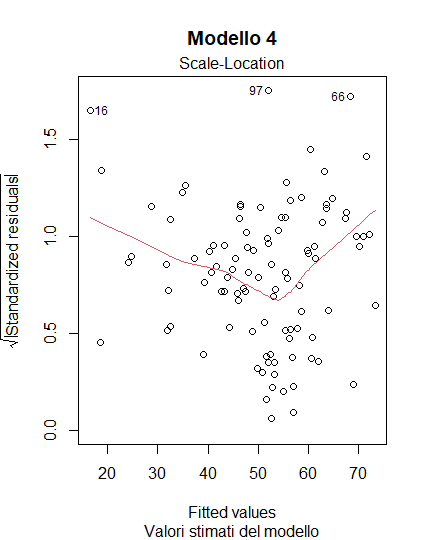
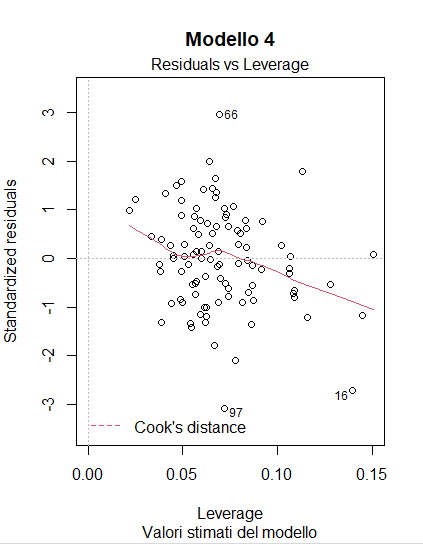
Sebbene il valore di R2 sia calcolato dalla funzione *lm* e stampato a video tramite la *summary* come informazione rilevante correlata, è stata comunque applicata la formula per calcolarlo. Questa risulta essere:

*R2=SQR/SQE*

Dove

*SQR=SQTOT– 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 equivalgono. 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 sono riportati nella pagina seguente. Si consiglia comunque una visualizzazione migliore all’interno dello stesso ambiente di programmazione.



**Scelta del modello che meglio si adatta ai dati forniti**

L scelta del modello è un’operazione caratterizzata dall’esistennza di vari tipi di criteri che permettono di scegliere un modello rispetto all’altro, i più utilizzati sono *BIC* e *AIC e MSE*. Con questi modelli si sceglierà il modello con valore minimo tra tutti quelli analizzati:

* *AIC=2k-2logn(L)*, dove *k* è il numero di parametri nel modello statistico e *L* è 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](https://it.wikipedia.org/wiki/Entropia_(teoria_dell%27informazione)) 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=-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 dimensioni 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 troverà tra quelli testati il criterio selezionerà 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 questo criterio 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*(errore quadratico medio): Indica la discrepanza quadratica tra i valori dei dati osservati e i valori dei dati stimati.

Si calcola come:MSE(thetha^)= E[theta-thetha^]=Var(thetha^)+bias(thetha^).La sua radice quadrata rappresenta un ulteriore indice statistico detto RMSE/RMSD(radice quadratico dell’errore medio/root mean square deviation),corrispondente alla varianza interna .Nella ‘ambito della scelta di un modello si è orientati a scegliere il modello rappresentato da un minor valore di questo indice relativo in quanto rappresenta il modello con la minor discrepanza tra i valori stimati e quelli osservati.

Osservando i risultati proposti dall’applicazione di ogni criterio sullo stesso modello si è potuto asserire che il modello da noi scelto rappresenta quello più ottimale.

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 modello 2 rispetto agli altri dato che rappresentava un valore minore rispetto agli altri. Tuttavia sulla base di quanto constatato nei punti precedenti(riguarda l’R squared), quest’ultimo rappresentava quello con il maggior numero di regressori con il conseguente eccessivo avvicinarsi all’ “overfitting” rischiando quindi di compromettere le prestazioni del software.

Per quanto riguarda la procedura di selezione utilizzata con tutti questi criteri , si è approcciati ad ogni modello con quella nota come “stepwise”(hybrid ,forward o backward).

Questo passaggio è stato rilevante per avere una conferma del modello ottenuto precedentemente. Tramite la funzione step(con tutti gli approcci possibili) ,infatti è stata lasciata la determinazione del modello ottimale all’ambiente RStudio, il quale ha confermato lo stessa risultato ricavato , in quanto non era possibile migliorarlo ulteriormente.