UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

Case study – Regressione lineare

Anno accademico: 2020-21

Gruppo di lavoro (LISBONA):

- Domenico Gaeni matricola n° 1065107
- Fabio Palazzi matricola n° 1066365

Introduzione

Il nostro gruppo ha ricevuto un dataset con i dati registrati presso la stazione di Mantova. L'obiettivo è quello di verificare la presenza/assenza di correlazione lineare fra le variabili indipendenti o regressori (X) tra cui *umidità*, *pioggia*, *temperatura*, *ossidi di azoto*, *ozono*, *biossido di azoto* e la variabile dipendente (Y) *PM10*. Si è scelto di estrapolare un modello per il pm10 perché riteniamo che sia un inquinante molto importante per la salute dell'uomo e sul quale oggigiorno si presta molta attenzione.

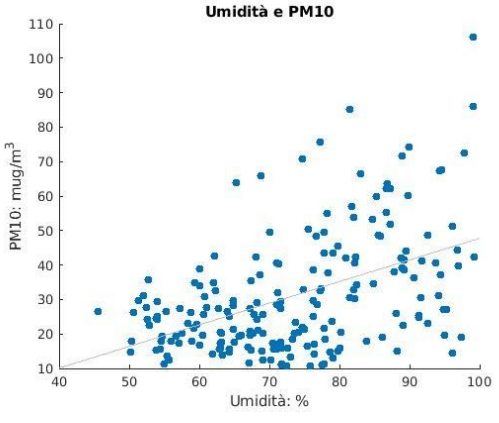
Strategia

La strategia utilizzata per identificare il modello più adatto è quella di verificare passo passo l'incidenza di ogni singolo regressore nel modello. Per prima cosa si mettono in relazione tutti i regressori e si osservano i parametri *stima*, *p-value* e *adattamento*. Se ci sono regressori con *p-value alto* (> 0.10) e *stima* prossima allo 0 si scarta quello con il p-value maggiore costruendo un nuovo modello senza quest'ultimo regressore, verificando che l'adattamento rimanga invariato o differisca di poco. E così via fino a quando si hanno solo regressori significativi. I regressori significativi saranno quelli con un p-value piccolo (almeno < 0.10 e con un parametro stima molto diverso da 0).
Così facendo eliminiamo i regressori non significativi, quelli cioè che soddisfano l'ipotesi nulla $H_0 : \beta_i = 0$, cioè quelli che hanno un coefficiente angolare per quel regressore pari a 0 e quindi non aggiungono informazioni al modello. Per determinare quali regressori sono meno significativi si osserva il p-value. Se è maggiore di 0,10 allora si accetta l'ipotesi nulla H_0 e quindi si tratta di un regressore non significativo. Eliminando questi regressori non significativi l'indice di adattamento non varierà di molto.
Questa strategia che abbiamo pensato in realtà esiste già e si chiama "***backward stepwise regression***": inizia considerando tutti i regressori e ad ogni step elimina il regressore meno significativo.

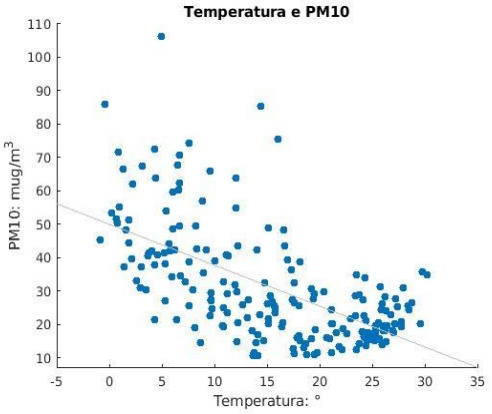
Svolgimento

Come prima cosa, per familiarizzare con i dati e fare un minimo di analisi, abbiamo disegnato un grafico (*scatter plot*) per la variabile risposta (*Y = PM10*) in funzione di ogni regressore per verificare graficamente se c'è o meno correlazione fra le due variabili.
Abbiamo ottenuto i seguenti risultati:

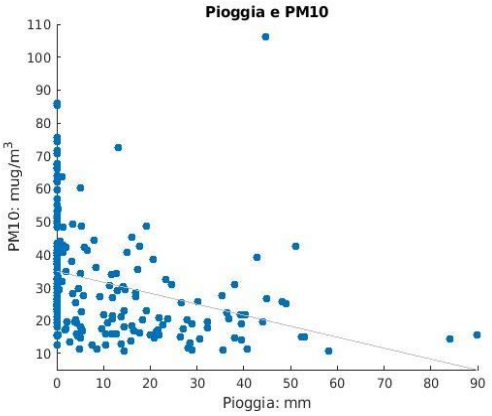
Umidità-PM10
Possiamo osservare che l'umidità e il PM10 sono correlati, infatti possiamo affermare che se l'umidità aumenta del 10%, il PM10 aumenta di circa 15 µg/m^3.
Inoltre, il coefficiente della retta trovata e visualizzabile nel grafico, è positivo.
La correlazione rimane comunque debole: alcuni dati si discostano significativamente dalla retta.
L'indice di correlazione è: 0.48



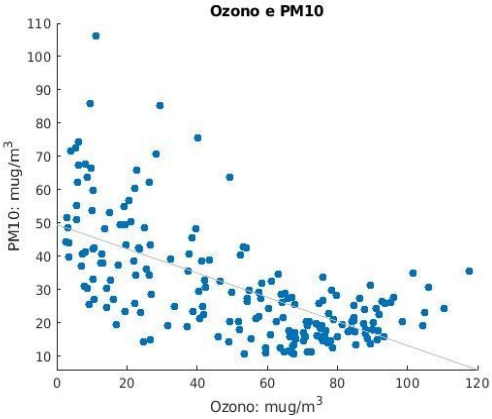
Temperatura-PM10
Dal grafico possiamo osservare che la temperatura e il PM10 sono correlati fra loro con il segno opposto rispetto al caso precedente: notiamo infatti che all'aumentare della temperatura il PM10 diminuisce, quindi la retta di regressione avrà coefficiente angolare <0. Anche in questo caso alcuni dati si discostano in maniera significativa dalla retta.
L'indice di correlazione vale -0.61



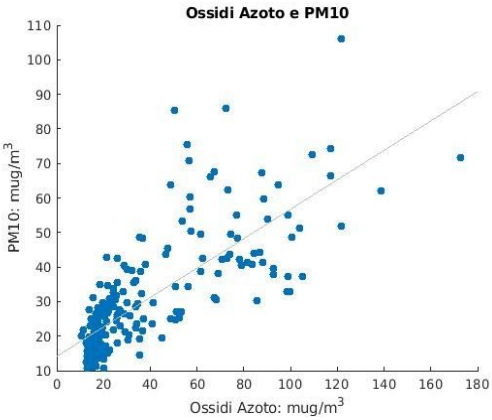
Pioggia-PM10
Possiamo affermare che la pioggia e il PM10 sono incorrelati o meglio debolmente correlati come si può notare dal grafico. A sostegno di ciò, si può notare che l'indice di correlazione lineare sarà basso e negativo.
L'indice di correlazione vale: -0.31

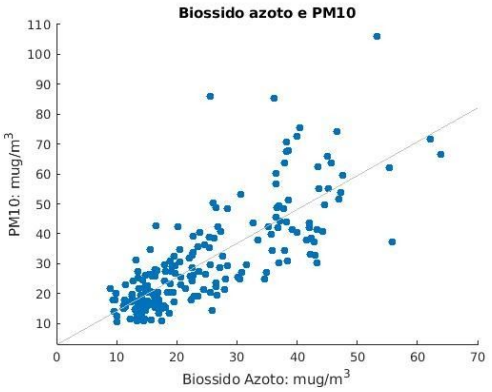


Ozono-PM10
Osserviamo che esiste una correlazione tra l'ozono e il PM10, con coefficiente angolare della retta <0 che sta quindi a confermare che all'aumento dell'ozono, diminuisce il PM10. Inoltre, è osservabile dal grafico che la correlazione è discreta, in quanto il coefficiente angolare non risulta particolarmente lontano da 0 ed inoltre i dati si discostano in modo significativo dalla retta.
L'indice di correlazione vale: -0.64



Ossido di Azoto-PM10
Possiamo notare l'esistenza di una buona correlazione tra gli ossidi di azoto e il PM10 in quanto il coefficiente angolare è >> 0 ed inoltre i dati sembrano addensarsi vicino alla retta.
L'indice di correlazione vale: 0.76



Biossido di Azoto-PM10 Osserviamo che i dati sono molto addensati in prossimità della retta, inoltre anche il coefficiente angolare >>0. Possiamo quindi affermare che esiste una buona correlazione tra Biossido di azoto e PM10. *L'indice di correlazione vale: 0.78*	

Abbiamo poi analizzato, come spiegato nel paragrafo "*Strategia*", un primo modello composto da tutti i regressori citati e poi passo dopo passo abbiamo eliminato i regressori non significativi (guardando principalmente il *p-value* e prestando attenzione al *coefficiente di determinazione*).

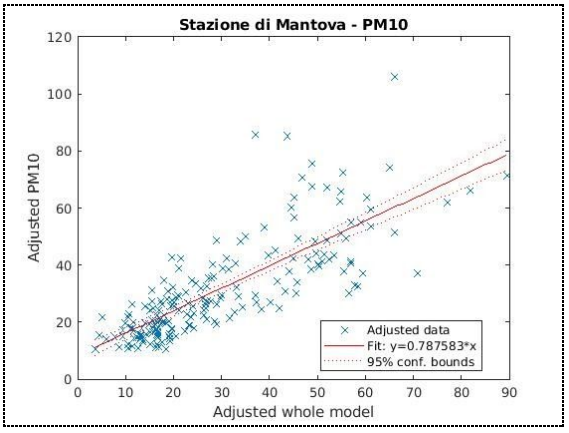
Stazione di Mantova

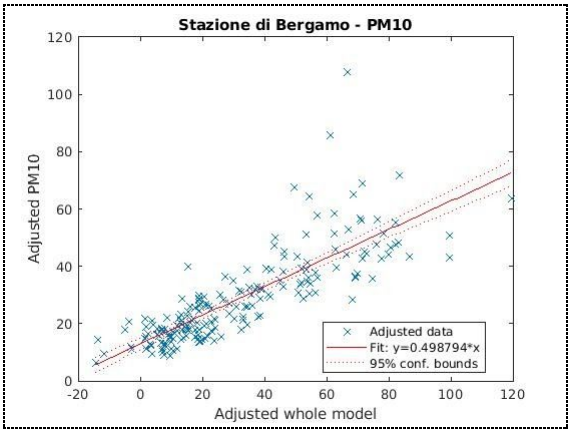
Dopo aver ricavato dal primo modello, contenente tutti i regressori, un R^2-aggiustato=0.626, abbiamo eliminato ben 4 regressori (*temperatura*, *ossidi di azoto*, *umidità*, *ozono*) ottenendo un coefficiente di determinazione aggiustato uguale a 0,62 e quindi pressochè uguale al primo. Dopo aver fatto ciò, abbiamo reinserito gli *ossidi di azoto* perché riteniamo che siano collegati con il biossido di azoto e anche perché nell'analisi preliminare ha evidenziato un indice di correlazione forte. *Questa aggiunta non rientra nella strategia, ma è frutto di un ragionamento fatto considerando il contesto del modello in questione*. Aggiungendo anche questo regressore al modello il suo p-value è conforme con la nostra strategia e R^2-aggiustato è diventato leggermente superiore (0.625).
Quindi possiamo affermare che nella stazione di Mantova il 62.5% della variabilità complessiva di PM10 è spiegata dalla relazione lineare con la *pioggia*, *ossidi di azoto* e *biossido di azoto*.

Stazione di Bergamo

Dal primo modello, contenente tutti i regressori, abbiamo ottenuto un R^2-aggiustato=0.659. Togliendo di volta in volta i regressori non significativi (*temperatura*, *ozono*, *umidità*), per via del loro valore del p-value (*> 0.10*) abbiamo ottenuto, a differenza del caso precedente, un aumento del valore di R^2-aggiustato a 0.663.
Quindi possiamo affermare che nella stazione di Bergamo il 66.3% della variabilità complessiva di PM10 è spiegata dalla relazione lineare con la *pioggia*, *ossidi di azoto* e *biossido di azoto*.
Come si può osservare i tre regressori in questione sono gli stessi utilizzati nel modello per la stazione di Mantova, per cui si può ipotizzare che, al di là del luogo di misurazione, il PM10 sia in qualche modo correlato con questi tre regressori, che se presi insieme ci forniscono un modello più o meno adatto per descrivere questo inquinante.

Grafico finale dei due modelli trovati:





Verifica Risultati

Per verificare la validità dei risultati ottenuti, oltre ad aver analizzato vari parametri (paragrafo "*Conclusioni*"), abbiamo sfruttato la potenza di MATLAB. Dopo una ricerca online, abbiamo trovato il comando stepwisefit. Questo comando si basa sulla stepwise ossia un metodo di selezione delle variabili indipendenti allo scopo di selezionare un set di predittori che abbiano la migliore relazione con la variabile dipendente. Esistono inoltre vari metodi di selezione delle variabili, tra cui il ***backward*** che si basa sull'algoritmo che abbiamo pensato e usato nella nostra risoluzione.

I risultati ottenuti per la stazione di Mantova attraverso il comando matlab sono:

{'Coeff' }	{'Std.Err.'}	{'Status'}	{'P' }	
{[-0.1106]}	{[0.0487]}	{'In' }	{[0.0242]}	Pioggia
{[0.1273]}	{[0.0820]}	{'Out' }	{[0.1220]}	Umidità
{[-0.0023]}	{[0.0471]}	{'Out' }	{[0.9618]}	Ozono
{[0.0607]}	{[0.1516]}	{'Out' }	{[0.6894]}	Temperatura
{[1.0861]}	{[0.0659]}	{'In' }	{[1.0519e-38]}	NO2
{[0.1314]}	{[0.0704]}	{'Out' }	{[0.0634]}	NOX

Mentre per la stazione di Bergamo sono:

{'Coeff' }	{'Std.Err.'}	{'Status'}	{'P' }	
{[-0.1233]}	{[0.0241]}	{'In' }	{[7.7712e-07]}	Pioggia
{[0.0661]}	{[0.0731]}	{'Out' }	{[0.3668]}	Umidità
{[-0.0125]}	{[0.0343]}	{'Out' }	{[0.7168]}	Ozono
{[-0.0262]}	{[0.1417]}	{'Out' }	{[0.8533]}	Temperatura
{[0.4713]}	{[0.1141]}	{'In' }	{[5.3894e-05]}	NO2
{[0.1072]}	{[0.0390]}	{'In' }	{[0.0066]}	NOX

In entrambi i casi, i regressori significativi, sottolineati in rosso, coincidono con quelli illustrati nel paragrafo "*Svolgimento*".

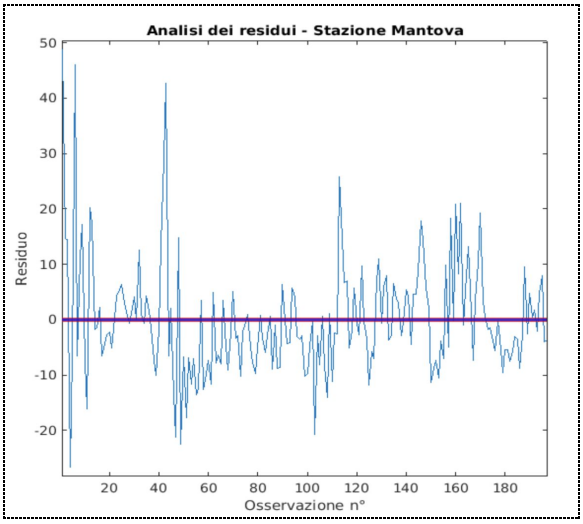
Conclusioni e commenti

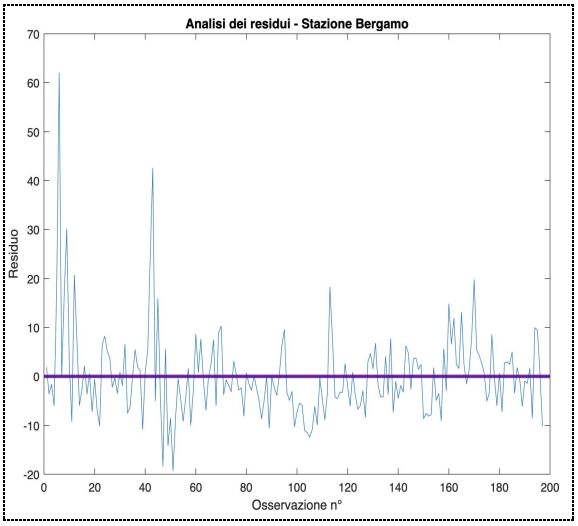
Il modello di regressione lineare trovato per entrambe le stazioni è:

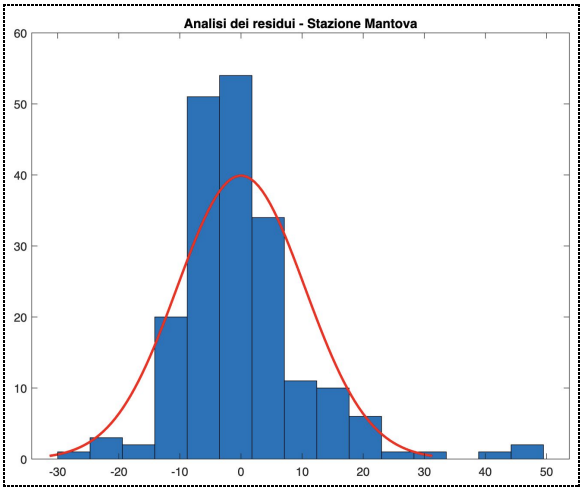
$$PM10 \sim 1 + Pioggia + Ossidi\ di\ Azoto + Biossido\ Azoto$$

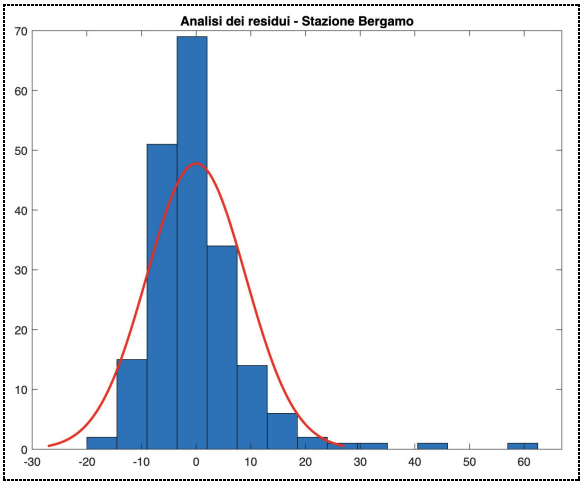
La ***stima dei coefficienti*** dei singoli regressori presentano lo stesso segno nei due modelli e sono in linea con quanto analizzato nelle singole regressioni lineare semplici (*scatter plot*).

Per esempio per la pioggia il coefficiente ha un valore negativo, in conformità con l'indice di correlazione trovato. Inoltre sempre per la pioggia i coefficienti differiscono di alcuni centesimi. Anche per gli ossidi di azoto i coefficienti differiscono solo di alcuni centesimi. Per quanto riguarda invece il biossido di azoto i due coefficienti differiscono di alcune decine. Presumiamo che questa differenza sia dovuta al fatto che le rilevazioni dei dati della stazione di Mantova vengono effettuate in prossimità del polo industriale e quindi in qualche modo vengono influenzate da questo fattore. Per verificare se la nostra ipotesi sia vera si può verificare il valore del coefficiente per le altre stazioni. Comunque in generale possiamo affermare che i coefficienti presentano tutti gli stessi segni e differiscono di poco.
Possiamo affermare inoltre che i due modelli presentano all'incirca le *stesse* ***performances di adattamento*** ovvero il 62.5% per Mantova e il 66.3% per Bergamo. Abbiamo deciso di utilizzare l'indice di determinazione corretto, perchè tiene conto anche del numero di regressori impiegati nel modello e dell'ampiezza del campione. Questo indice ci dice fino a che punto il modello lineare consente di approssimare la realtà dei dati osservati.
Le ***covariate*** dei modelli presentano gli stessi segni e non differiscono di molto, il che significa che in generale le variabili nei singoli modelli differiscono tra di loro nello stesso modo, al di là che il modello sia per la stazione di Mantova o di Bergamo.
I regressori utilizzati sono tutti ***significativi***, cioè con un p-value basso << 0.01 (rigetto forte) e con la statistica test alta. Questo è infatti l'obiettivo della strategia che abbiamo utilizzato volta a raggiungere un modello che presenti dei regressori significativi e che rappresenti al meglio la realtà dei dati.
Infine per avere un'ulteriore conferma della validità dei modelli si può effettuare l'***analisi dei residui*** e vedere se si distribuiscono uniformemente intorno all'asse di ascissa 0 e se hanno un andamento gaussiano. Come si può osservare dai seguenti grafici i residui di entrambi i modelli trovati hanno media uguale a 0 e si distribuiscono come una normale, per cui si può concludere che i due modelli trovati (Mantova e Bergamo) sono validi.






Analisi dei residui - Stazione Mantova
60
50
40
30
20
10
0
-30
-20
-10
0
10
20
30
40
50


Analisi dei residui - Stazione Bergamo
70
60
50
40
30
20
10
0
-30
-20
-10
0
10
20
30
40
50
60