Case Study - Regressione lineare

Corso di Statistica e Modelli Stocastici (SMS1)

Anno Accademico 2021 - 2022

Gruppo G08 - Berlino

- *Zanotti Paolo* *Mat. 1074166*
- *De Duro Federico* *Mat. 1073477*
- *Ciullo Roberto* *Mat. 1074568*

UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

1. Introduzione

Il nostro gruppo ha ricevuto un dataset con i dati registrati presso la stazione di Casirate d'Adda, in provincia di Bergamo.

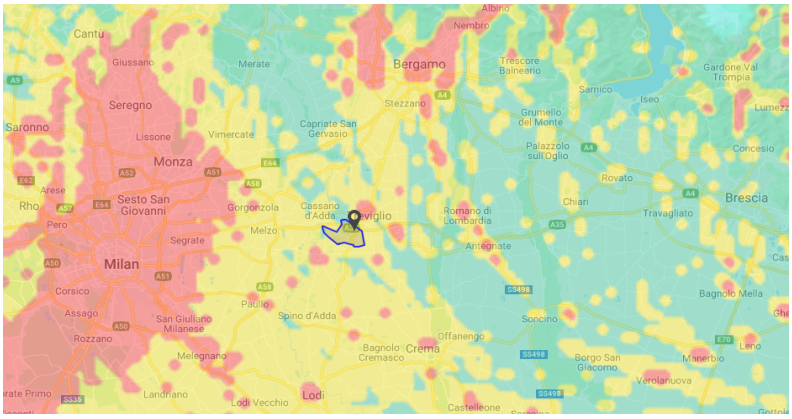
L'obiettivo principale è quello di verificare la presenza / assenza di correlazione lineare tra le variabili indipendenti (o Regressori X), tra cui eventi climatici (Umidità, pioggia, temperatura), inquinanti (Ossidi/Biossidi di azoto e Ozono) e la vendita di carburante per vetture o per il riscaldamento, e la variabile dipendente (Y), studiata in due casistiche diverse (PM_{10} e NO_x).

Abbiamo scelto il PM_{10} perché lo riteniamo un inquinante molto dannoso per la salute dell'uomo, studi epidemiologici infatti, confermati anche da analisi cliniche e tossicologiche, hanno dimostrato come l'inquinamento atmosferico abbia un impatto sanitario notevole; quanto più è alta la concentrazione di polveri fini nell'aria, infatti, tanto maggiore è l'effetto sulla salute della popolazione.

In riferimento alla pagina ARPA "PM10 - MEDIA GIORNALIERA IN µg/m3", si può notare come Casirate d'Adda sia in una zona in cui la concentrazione delle polveri sottili e degli ossidi di azoto possono essere molto dannose nel caso si superassero i limiti giornalieri e annuali:

Inquinante	Tipo di limite	Limite
PM_{10}	Limite giornaliero	50 µg/m³ da non superarsi per più di 35 giorni all'anno
	Limite annuale	40 µg/m³ media annua
NO_2	Limite orario	200 µg/m³ media oraria da non superare per più di 18 volte all'anno
	Limite annuale	40 µg/m³ media annua

Zona Stazione	Zona A (Pianura ad elevata urbanizzazione)
Tipo Stazione	RB (Rurale background)



2. Strategia e fasi iniziali

Inizialmente, abbiamo analizzato i dati raccolti tramite l'utilizzo di MatLab, aiutandoci con il comando *plotmatrix,* che fornisce una matrice di grafici in cui vengono correlati tra loro inquinanti, condizioni meteorologiche e vendita di carburanti. Grazie a questa matrice di grafici siamo stati in grado di capire subito l'andamento delle polveri sottili (PM_{10}) e degli ossidi di azoto (NO_x) al variare dei fattori.

Come si può vedere nella *Fig. 1.1*, sono stati sottolineati ed evidenziati, i grafici significativi che abbiamo voluto analizzare in modo più approfondito per il nostro case study. In particolare, si può notare con facilità che il PM_{10} è *fortemente correlato* con le precipitazioni, NO_x, NO_2, e la quantità di gasolio venduta per il riscaldamento. Inoltre, notiamo che NO_x e NO_2 sono molto correlati tra loro, tuttavia questo non stupisce, visto che l'NO_2 è un sottoinsieme di NO_x.

Per avere una seconda conferma, abbiamo inoltre costruito un vettore contenente tutti i coefficienti di correlazione con PM_{10}. (*Fig. 1.2*) [Analogo per NO_x].

Grazie a queste prime analisi, prima di stimare il modello migliore, possiamo già ipotizzare che nel modello per PM_{10} ci saranno O_3, il gasolio e uno tra NO_x e NO_2. La nostra strategia è quella di applicare una regressione graduale ai nostri dati (*Stepwise Backward Elimination*).

	PM10
PM10	1
Temperatura	-0.76338
Pioggia	-0.48738
Umidita	0.59247
O3	-0.7433
NOx	0.73642
NO2	0.53127
Benzina	-0.041862
Gasolio_motori	0.095612
Gasolio_risc	0.7721

Figura 1.2

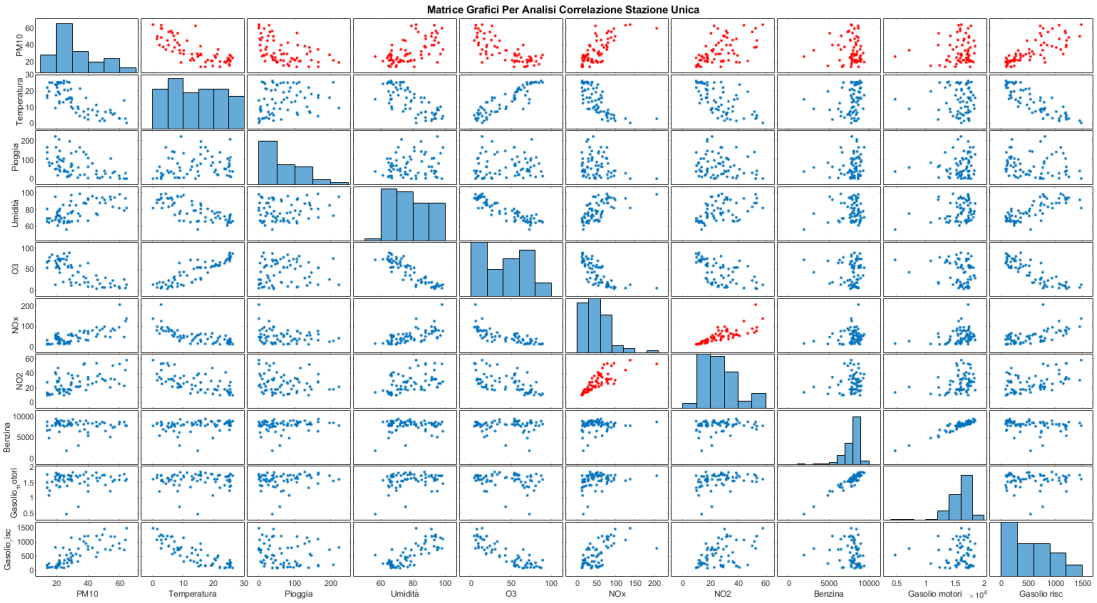
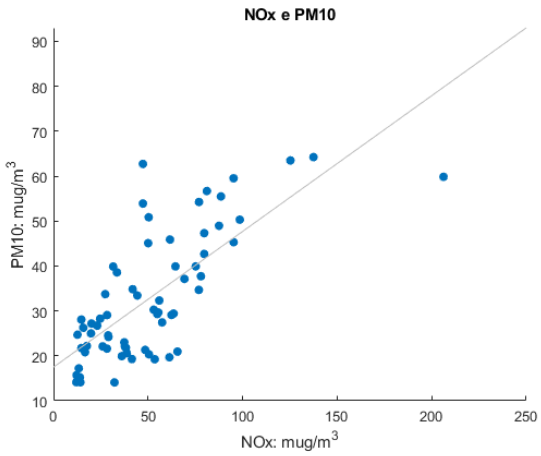


Figura 1.1 Grafici di correlazione per PM10

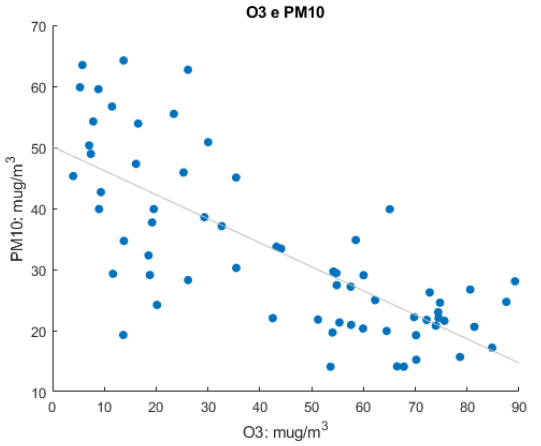
Analizzando i singoli grafici di correlazione dei regressori ipotizzati nel modello con le polveri sottili, abbiamo ottenuto i seguenti risultati:

NO_x - PM_{10}
Si può notare una forte correlazione tra le concentrazioni di ossidi di azoto e polveri sottili, infatti all'aumentare di uno, anche l'altro aumenterà di conseguenza.

L'indice di correlazione vale 0.73, i dati si addensano vicino alla retta e il coefficiente è >> 0.
Sarà uno dei regressori del nostro modello.

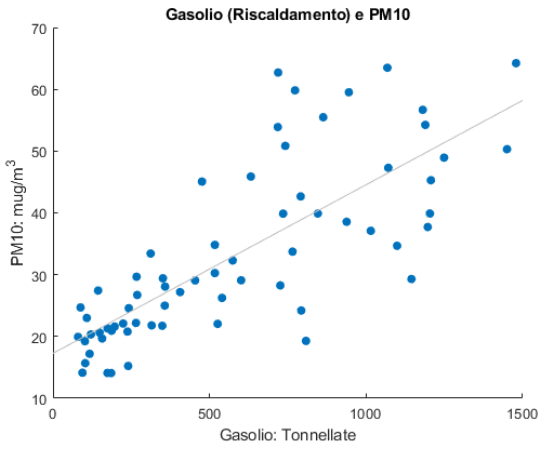


O_3 - PM_{10}
Si osserva che esiste una discreta correlazione tra O_3 e PM_{10}, con coefficiente angolare della retta < 0, che sta quindi a confermare che l'aumento di ozono, porta ad una diminuzione delle polveri sottili nell'aria.
L'indice di correlazione vale -0.74 e alcuni dati sembrano addensarsi vicino alla retta.



Gasolio (Riscaldamento) - PM_{10}
Anche qui è presente una forte correlazione tra la vendita di gasolio per il riscaldamento e l'aumento delle polveri sottili, infatti il coefficiente angolare è >> 0, i dati si addensano vicino alla retta e *l'indice di correlazione è molto alto (0.77)*

Possiamo già immaginare che questo regressore sarà all'interno del modello finale.



3. Svolgimento

Per prima cosa, abbiamo deciso di applicare la *Stepwise Backward Elimination* per entrambi gli inquinanti studiati (PM_{10} e NO_x), partendo da un modello comprendente tutti gli inquinanti ed eliminando poi iterativamente tutti quelli non significativi, basandoci principalmente sul loro p-value e il loro coefficiente di determinazione.

Stepwise Backward Elimination per le *polveri sottili* (PM_{10})

Fissato α = 5% (0.05), abbiamo subito analizzato i *p-value* dei regressori, ed eliminato dal nostro modello la *Temperatura* (Pv: 0.71), perché non è *statisticamente significativa* (Un regressore, per essere significativo assumiamo che p-value ≤ α). Continuando il procedimento pensato, abbiamo rimosso dal modello *l'umidità dell'aria* (Pv: 0.33) e le tonnellate di *benzina* (Pv: 0.27) e di *gasolio* (Pv: 0.48) vendute su rete ordinaria.

A questo punto, nel nuovo modello il p-value relativo alle concentrazioni di *biossidi di azoto* è superiore al nostro α scelto e, di conseguenza, abbiamo rimosso anch'esso dalla nostra traccia. Ciò era prevedibile, infatti all'interno dei dati si trovano concentrazioni simili dei due inquinanti, perciò l'NO_x spiega gran parte dell' NO_2 presente (Essendo NO_2 un sottoinsieme di NO_x, i due dati sono *linearmente dipendenti*).

Alla fine del processo, siamo riusciti a trovare un modello del tipo PM10 ~ 1 + Pioggia + NO_x + O3 + Gasolio_risc, con un *indice di determinazione* molto vicino a quello del modello completo e dove tutti i coefficienti sono molto significativi (p-value molto bassi).

	R-squared	***R-squared Aggiustato***
Modello completo	*0.833*	*0.806*
Modello trovato	*0.820*	*0.808*

Per evitare *l'overfitting* dei dati, avremmo potuto anche rimuovere l'ozono dal nostro modello, infatti l'indice di determinazione è molto simile al modello trovato (80%). Tuttavia abbiamo preferito mantenerlo nel modello perché ci siamo basati sul p-value inferiore al nostro α = 5%.

Per una verifica più accurata, abbiamo esaminato tramite il comando *stepwiselm* (adattato per fornire un modello lineare) che il modello ottenuto fosse effettivamente il migliore, e infatti abbiamo trovato come regressori principali gli stessi del nostro modello (*Figura 3*).

Figura 3

```
Linear regression model:
    PM10 ~ 1 + Pioggia + NOx + O3 + Gasolio_risc

Estimated Coefficients:
                    Estimate        SE         tStat       pValue
                    ________    _________    _______    __________

    (Intercept)       33.411       6.1606     5.4234    1.0606e-06
    Pioggia        -0.093262     0.014684    -6.3515    2.9762e-08
    NOx               0.1013     0.034234      2.959     0.0043888
    O3             -0.13953     0.066956     -2.084      0.041358
    Gasolio_risc    0.011901    0.0038517     3.0899     0.0030148
```

Come fasi finali, abbiamo analizzato i *residui* del modello, verificando:

- La loro media pari a 0
- La loro distribuzione gaussiana
- L'andamento dei percentili
- L'incorrelazione dei regressori con i residui
- La loro varianza omogenea
- Eventuali *Outliers*

Tramite il comando MatLab *histfit* abbiamo verificato la distribuzione *Normale* dei residui, mentre con un normale comando *plot* ci siamo accertati di avere una media dei residui pari a zero, e che i residui si distribuiscano attorno ad essa (*Figura 4*).

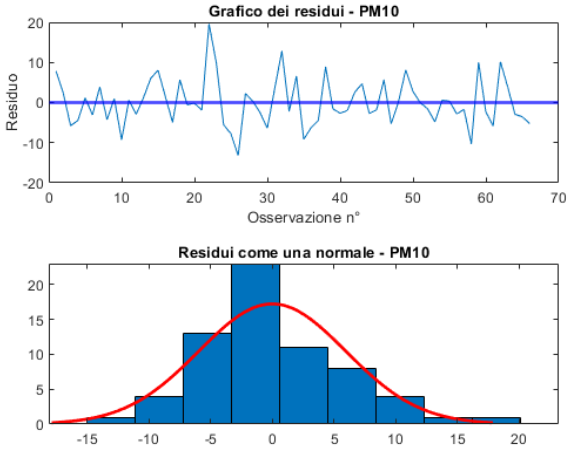


Figura 4

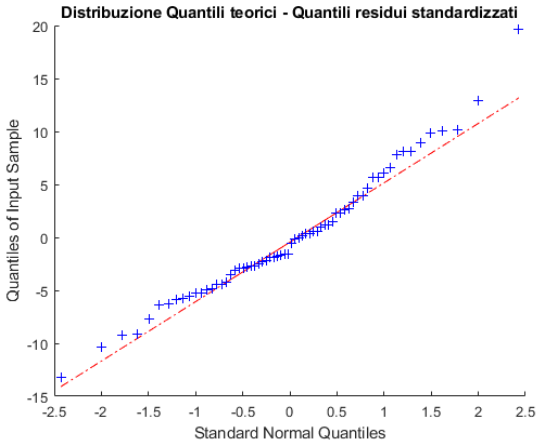


Figura 5

Abbiamo approfondito anche un altro aspetto dei residui, infatti, tramite il comando MatLab *qqplot*, che mantiene sull'asse delle ascisse i *quantili teorici di una distribuzione normale,* mentre su quello delle ordinate *i quantili dei residui standardizzati*, abbiamo verificato l'andamento dei percentili residui (*Figura 5*). L'ennesima dimostrazione che si distribuiscono normalmente.

Si verifica in seguito che i regressori non sono correlati con i residui. Questo punto è molto importante, perché la componente di errore di un modello deve essere *imprevedibile.* Abbiamo anche effettuato una verifica analitica tramite il coefficiente di correlazione tra residui e regressori (Tutti chiaramente con valori molto bassi) (*Figura 6*).

I residui hanno inoltre una varianza omogenea. La varianza deve essere uguale per tutti i residui. Questa ipotesi può essere verificata utilizzando un *grafico a dispersione* dei *valori residui* (asse y) e dei *valori stimati* (asse x). Il risultato finale dovrebbe apparire come una banda orizzontale di punti tracciati in modo casuale. In modo più semplice e intuitivo, abbiamo sfruttato il comando MatLab *plotResiduals*, creato apposta per questa esigenza (*Figura 7*).

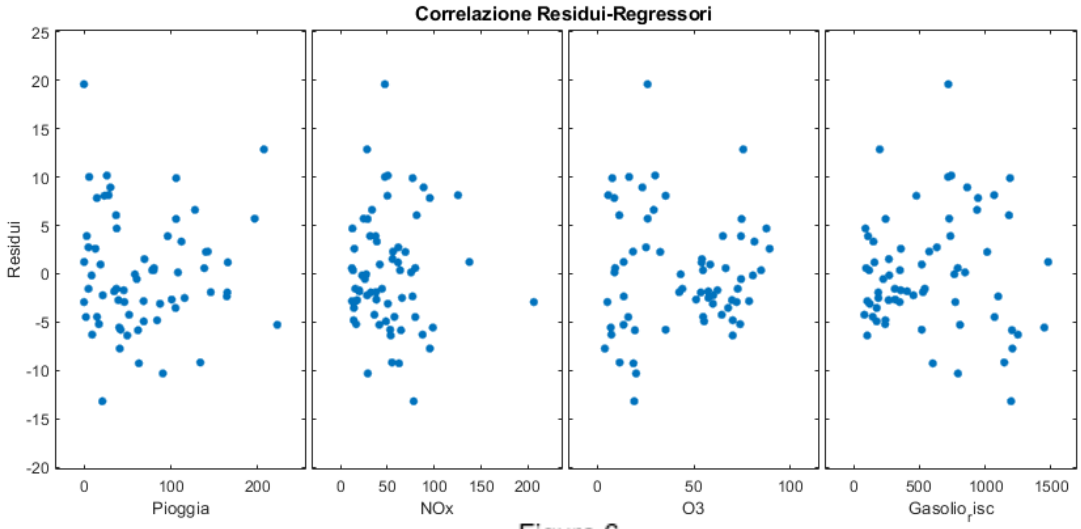


Figura 6

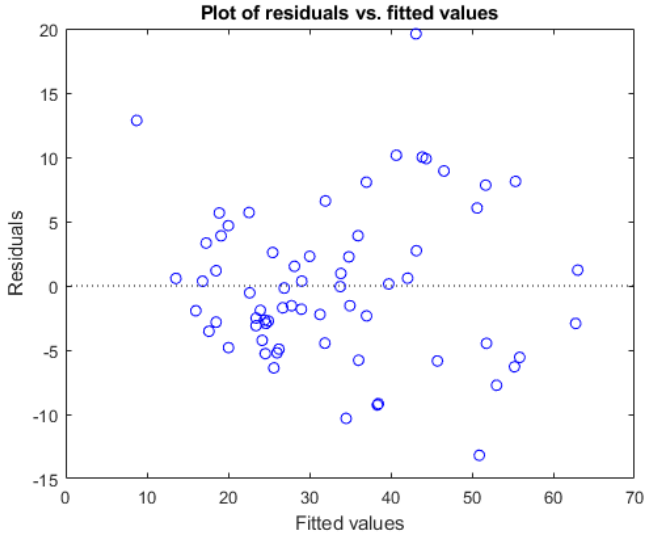
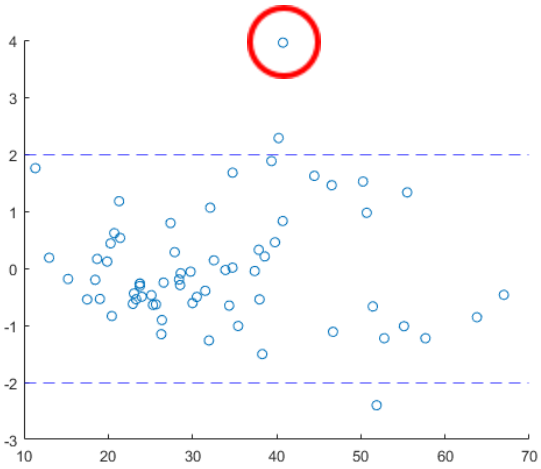


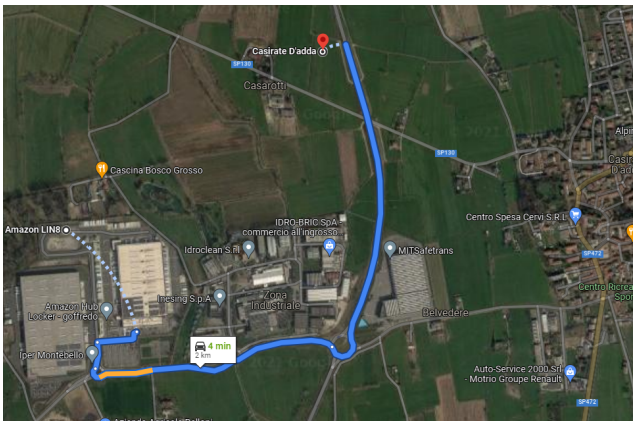
Figura 7

Analizzando gli *Outliers* del nostro modello, si può notare che un valore specifico si discosta particolarmente dagli altri.

Questo dato è stato registrato il 1° Ottobre 2017, il quale ha evidenziato un picco di polveri sottili (62.71 $\mu g/m^3$). E' un dato che non stupisce nei mesi invernali, infatti nel periodo tra Gennaio e Febbraio, durante gli altri anni sono stati rilevati picchi di 64.3, 63.48, 55.48 e 53.90 $\mu g/m^3$. Tuttavia, il dettaglio evidente è il mese, ossia Ottobre.

Informandoci su giornali del periodo, abbiamo scoperto che a Settembre 2017 sono cominciati i lavori della nuova sede operativa *Amazon*, i quali probabilmente hanno influito sulle concentrazioni di polveri sottili nell'aria, difatti, confrontando la posizione della nuova hub e della nostra stazione di controllo, abbiamo osservato che sono molto vicine tra loro.





Sono presenti altri Outliers all'interno del nostro modello, tuttavia sono relativamente marginali e non influiscono particolarmente sulla pendenza della retta.

Stepwise Backward Elimination per gli ossidi di azoto (NO_x)

Anche in questo caso siamo partiti da un α fissato al 5% (0.05) e abbiamo costruito un modello lineare completo, con tutti i regressori disponibili (*lm_completo_NO_x*). Guardando poi il *p-value* abbiamo deciso di eliminare dal modello tutti i regressori *non statisticamente significativi* (quando *p-value* ≥ *α*):

- Pioggia (*pv: 0.95*)
- O3 (*pv: 0.51*)
- Temperatura (*pv: 0.18*)
- Gasolio Riscaldamento (*pv: 0.33*)

Il modello finale trovato è quindi

$$NO_x \sim 1 + Umidità + PM_{10} + NO_2 + Benzina + Gasolio\ Motori$$

Ne conferma l'attendibilità la seguente tabella:

	R-squared	***R-squared Aggiustato***
Modello completo	*0.844*	*0.819*
Modello trovato	*0.835*	*0.821*

Per una verifica completa abbiamo utilizzato il comando "*stepwiselm*".

```
Linear regression model:
    NOx ~ 1 + Umidita + PM10 + NO2

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   --------    -------    -------    ----------

    (Intercept)     -66.173      14.392    -4.5978    2.1546e-05
    Umidita         0.66806      0.2193     3.0463     0.0034006
    PM10            0.80302     0.18432     4.3565    5.0541e-05
    NO2              1.4362     0.18824     7.6298    1.7377e-10
```

Dalla figura si può notare che non sono state incluse due variabili molto importanti, ovvero la *Benzina* e il *Gasolio Motori*.

Abbiamo deciso ugualmente di inserirle nel nostro modello in quanto *statisticamente significative* (*p-value* ≤ *α*) e perché riteniamo che la loro presenza possa influenzare notevolmente la quantità degli ossidi di azoto nell'aria.

Ulteriore verifica può essere fatta con il comando *stepwisefit*.

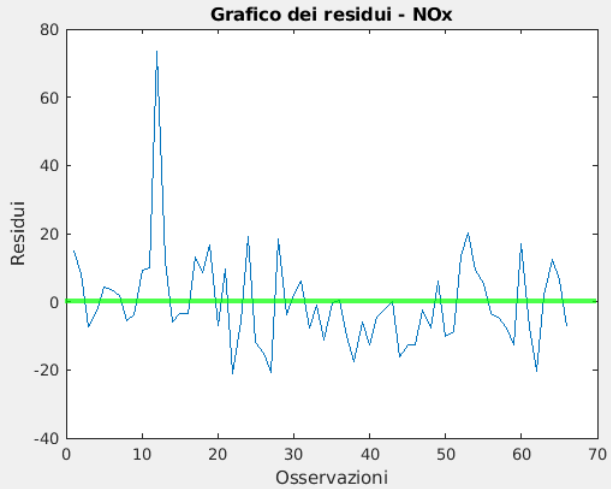
{'Coeff'}	{'Std.Err.'}	{'Status'}	{'P'}
{[-0.0146]}	{[0.0514]}	{'Out'}	{[0.7777]}
{[0.6681]}	{[0.2193]}	{'In'}	{[0.0034]}
{[0.2764]}	{[0.4713]}	{'Out'}	{[0.5597]}
{[0.0392]}	{[0.2098]}	{'Out'}	{[0.8525]}
{[0.8030]}	{[0.1843]}	{'In'}	{[5.0541e-05]}
{[1.4362]}	{[0.1882]}	{'In'}	{[1.7377e-10]}

Il risultato, come da aspettativa, è lo stesso ottenuto con il comando *stepwiselm*.

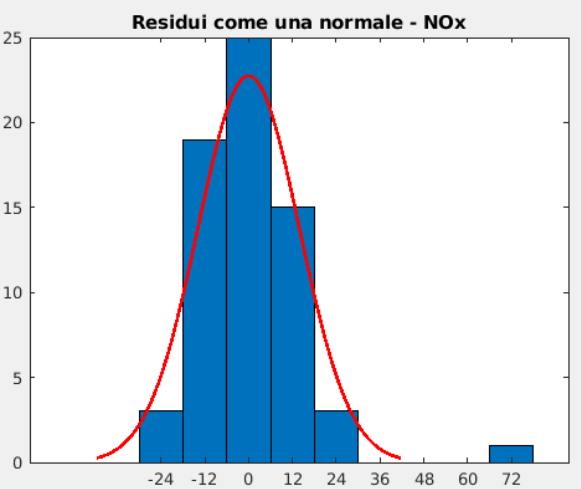
Per quanto riguarda l'analisi dei residui (calcolati molto semplicemente tramite Matlab con "*lm_4_NOx.Residuals.Raw*") sono stati controllati i 6 punti visti precedentemente:

- *Media pari a 0 e distribuzione normale*
- *Andamento percentili e Incorrelazione regressori con i residui*
- *Varianza omogenea e Outliers*

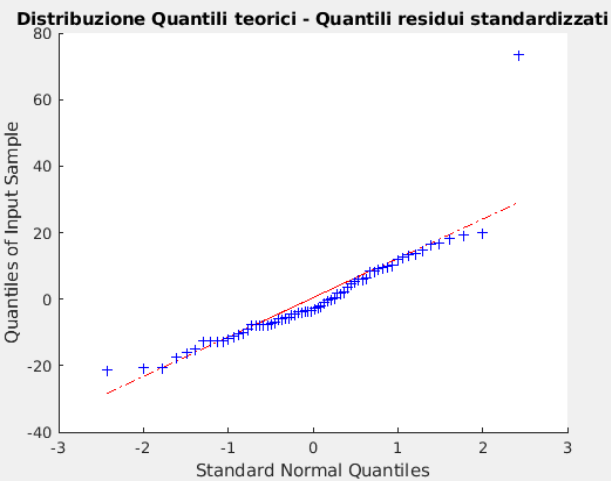
(1)



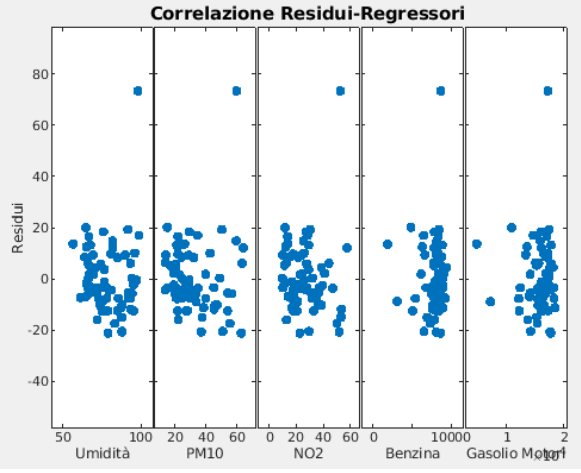
(2)



(3)

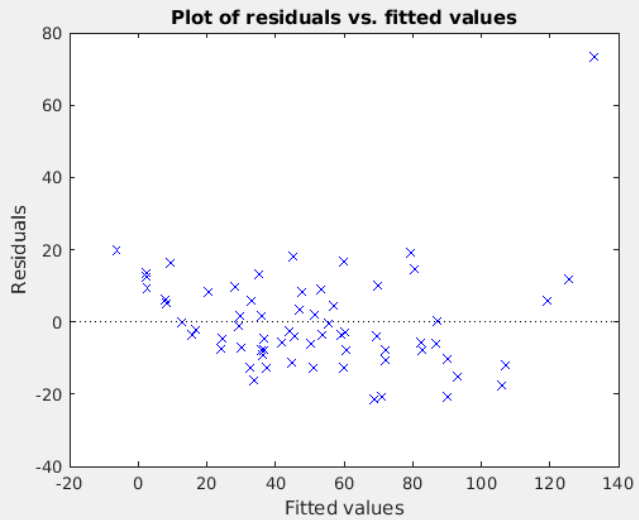


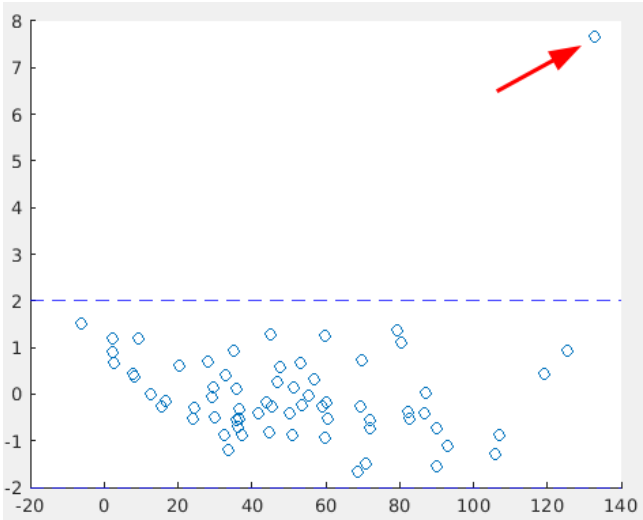
(4)



1) Abbiamo verificato il valore della media dei residui di NO_x (*Linea verde*)
2) Per quanto riguarda la distribuzione normale dei residui, la linea di adattamento migliore *è ottimizzata al centro* e non presenta una distorsione verso alcuni dati e non è lontana dagli altri.
3) L'andamento dei quantili è corretto. Infatti se la distribuzione è di tipo normale allora il *grafico sarà lineare*. Si può già notare un valore anomalo che verrà analizzato successivamente.
4) *I regressori non sono correlati con i residui*, dunque la componente di errore è imprevedibile, come ci aspettiamo in un modello valido.

(5)





(6)

5) La varianza deve essere la stessa per tutti i residui. Come già scritto prima, per verificare questo aspetto si può usare il comando "*plotResiduals*" e vedere se il grafico è una banda orizzontale.
6) Si può vedere chiaramente la presenza di un dato con un valore anomalo. Infatti i residui studentizzati dovrebbero essere compresi tra 2 e -2 nel nostro caso. Analizzando bene il dataset abbiamo scoperto che l'outlier si tratta del mese di Dicembre 2016, ovvero la quantità massima di NO_x registrata (206.23 µg/m3). In quel periodo la qualità dell'aria era pessima nella bassa bergamasca, come confermato da un articolo dell'Eco di Bergamo *[Sitografia, punto 4]*.

4. Risultati

Da un'osservazione approfondita dei dati relativi alla stazione ARPA collocata nel territorio di Casirate d'Adda, nel periodo compreso dal 1° Gennaio 2016 al 1° Giugno 2021, possiamo appurare l'incidenza del PM_{10} e dell' NO_x sull'ambiente.

PM_{10}

Possiamo notare come la media annuale della sostanza abbia mantenuto un livello compreso tra un minimo di 30,13 µg/m^3 ed un massimo di 39,96 µg/m^3, pertanto *sempre al di sotto del limite massimo di 40 µg/m3 imposto dalla legge*.

I picchi si sono registrati sempre tra Gennaio e Febbraio ad eccezione dell'anno 2017, nel quale il culmine si è raggiunto nel mese di Ottobre, quando si è arrivati ad una quantità di particolato pari a 62,71 µg/m^3. Questo aumento è stato registrato per diversi fattori tra cui, come già detto, le nuove costruzioni edili e la scarsità di pioggia caduta in quel periodo.

Negli anni in esame possiamo notare come le punte minime siano state raggiunte nei mesi di Maggio, in cui i dati oscillavano tra un minimo di 14,07 µg/m^3 e un massimo di 21,32 µg/m^3. Generalmente i dati sono sempre risultati essere più bassi nei mesi primaverili ed estivi per salire gradualmente da inizio autunno fino a pieno inverno.

Altri dati d'interesse che influiscono sulla concentrazione di polveri sottili sono la quantità di pioggia che cade nell'anno, il minor consumo di gasolio per riscaldamento, che influisce in senso positivo sulla qualità dell'aria e la quantità di ozono (O_3) presente nell'atmosfera, la quale aiuta a contenere la quantità di polveri sottili nell'ambiente. Influisce positivamente nel mantenere un livello basso di PM_{10} anche la temperatura, in quanto come si può notare dai dati raccolti un aumento della stessa influisce in modo positivo sul contenimento degli elementi dannosi per l'ambiente.

NO_x

Analizzando invece l'NO_x, si può notare che le medie annuali vanno da un minimo di 31,99 µg/m^3 nel 2021 ad un massimo di 69,75 µg/m^3 nel 2016, questo può indicare una *riduzione dell'inquinante negli anni, dovuto probabilmente alle leggi sempre più stringenti riguardanti le emissioni dei veicoli a motore e i limiti industriali.*

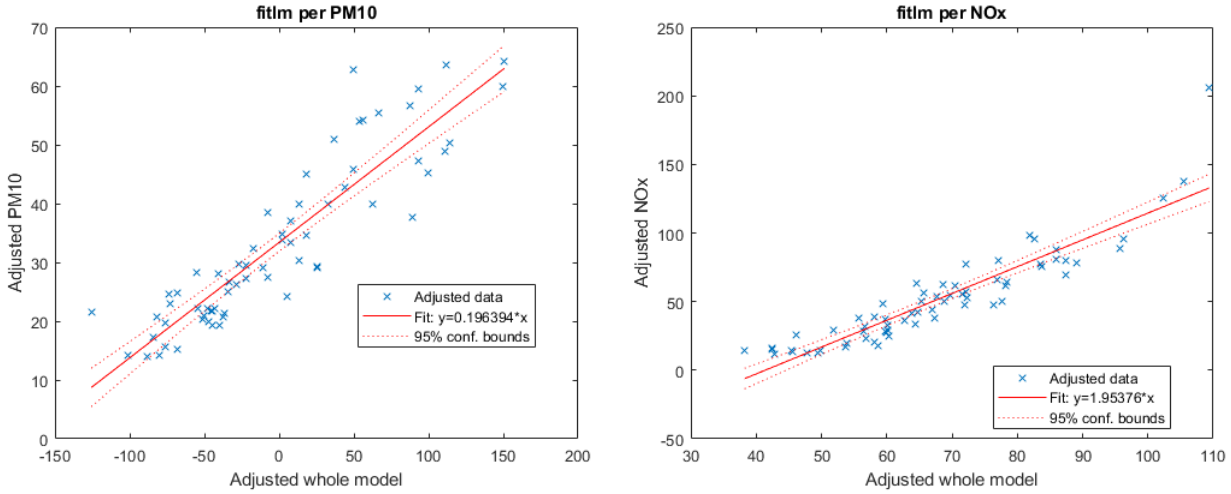
I picchi più importanti si sono rilevati a Dicembre 2016, con un record di 206,23 µg/m^3 e un minimo di 12,06 µg/m^3 nell'agosto 2020, a conferma del

fatto che le emissioni siano diminuite negli anni. Un'altra causa della probabile diminuzione degli ossidi di azoto, in particolare tra il 2020 e il 2021, è sicuramente la *riduzione drastica del traffico*, conseguenza dei lockdown istituiti per prevenire l'aumentare della pandemia.

5. Conclusioni

Come punto finale di questo report, possiamo quindi stabilire che nella stazione di Casirate d'Adda, l'*81%* della variabilità complessiva delle polveri sottili (PM_{10}) è spiegata dalla relazione lineare con la pioggia, le concentrazioni di NO_x e O_3 e la vendita di gasolio per il riscaldamento.

Similmente, l'*82%* della variabilità complessiva degli ossidi di azoto (NO_x) è spiegata dalla relazione lineare con umidità, PM_{10} e la vendita di carburante (sia benzina che gasolio).



PM_{10} ~ 1 + Pioggia + Ossidi di azoto + Ozono + Gasolio_risc

NO_x ~ 1 + Umidità + PM_{10} + NO_2+ Benzina + Gasolio_motori

Possiamo affermare che i due modelli presentano all'incirca le stesse *performances di adattamento* (81% e 82%). Abbiamo deciso di utilizzare l'indice di determinazione corretto perché tiene conto anche del numero di regressori impiegati nel modello e dell'ampiezza del campione. Questo indice ci indica fino a che punto il modello consente di approssimare la realtà dei dati osservati.

I regressori utilizzati sono tutti significativi, con p-value << 0.01 (Rigetto forte), e i residui sono uniformi, di conseguenza possiamo affermare che i due modelli trovati sono *validi*.

6. Sitografia

- Script di MatLab usato per i calcoli
- https://doc.arcgis.com/it/insights/latest/analyze/regression-analysis.htm
- https://paolapozzolo.it/analisi-dei-residui-regressione
- https://www.mathworks.com/help/stats/qqplot.html
- https://www.ecodibergamo.it/stories/bergamo-citta/dopo-25-giorni-finalmente-si-respirabergamo-smog-sotto-la-soglia-dallerta_1216926_11
- https://www.ecodibergamo.it/stories/Economia/amazon-cantiere-a-tempo-di-recordcasirate-avvio-dellattivita-il-28-ottobre_1287425_11/
- https://www.mathworks.com/help/matlab/ref/plotmatrix.html
- https://www.dsu.univr.it/documenti/OccorrenzaIns/matdid/matdid324170.pdf
- https://www.arpalombardia.it/Pages/Aria/qualita-aria.aspx
- https://www.arpalombardia.it/Pages/Meteorologia/Previsioni-e-Bollettini.aspx#/topPagina
- https://it.mathworks.com/help/matlab/ref/scatter.html
- https://it.mathworks.com/help/stats/fitlm.html
- https://it.mathworks.com/help/stats/histfit.html
- https://www.qualtrics.com/support/stats-iq/analyses/regression-guides/interpreting-residual-plots-improve-regression/
- https://it.mathworks.com/help/stats/residuals.html