PROGETTO DI STATISTICA

Regressione lineare multipla sulla concentrazione di ozono nelle provincie di Monza e Bergamo

Università degli Studi di Bergamo a.a 2020/2021

Docenti del corso:

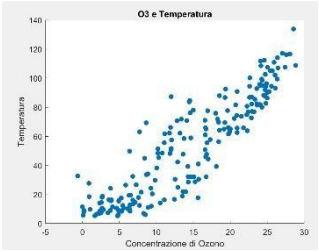
Alessandro Fassò e Paolo Maranzano

Componenti del gruppo:

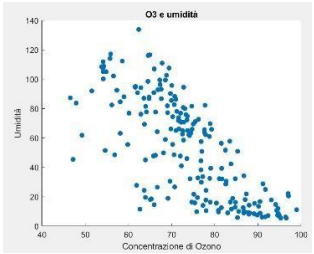
Kenna Miriam Fatima, Dandis Iana, Luca Lorenzi e Negro Matias

INTRODUZIONE E ANALISI DEI DATI

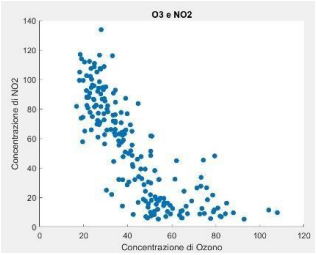
Al gruppo 'Monaco' è stato assegnato il dataset contenente i dati settimanali della qualità dell'aria e meteorologia di due stazioni: **Monza - via Machiavelli** ed **Bergamo- via Meucci.**
Il primo passo è stato l'ispezione del dataset, per poter individuare le statistiche descrittive minime (minimo, media e massimo); conseguentemente a queste prime individuazioni è stata scelta come variabile risposta la concentrazione di Ozono.
Tramite l'utilizzo di grafici a dispersione è stato possibile ottenere una prima visione delle relazioni tra l'ozono e le altre variabili, qui riportiamo alcuni grafici relativi alla provincia di Monza.



TEMPERATURA
Attraverso l'analisi di correlazione tra l'ozono e la temperatura si identifica (Come mostrato dal grafico) che vi è una diretta correlazione tra l'aumento della temperatura e l'aumento di ozono nell'aria (in fitlm = 2.8093).



UMIDITA'
Collegandosi all'umidità, in questo caso diminuisce la quantità di ozono presente nell'aria (in fitlm = -87424 < 0) questo fenomeno è dato dalla reazione chimica che avviene tra l'umidità stessa e l'ozono, portando una riduzione del suo rendimento per KWh.



NO2
Anche in questo caso, all'aumentare di NO2 diminuisce la concentrazione di ozono nell'aria, questo fenomeno è spiegabile attraverso la reazione chimica: O3 + NO2 → NO3 + O2.

Tramite appositi comandi è stata estrapolata una tabella secondaria contenente esclusivamente le variabili di nostro interesse (temperatura, concentrazione di NO2 ed NOx). Per poter identificare il modello utile a descrivere le concentrazioni, inizialmente è stato costruito m1:
Ozono ~ 1 + PM10 + Temperatura + Pioggia +Umidità.
Si ha la possibilità inoltre di visualizzare i coefficienti stimati, le statistiche di riepilogo del modello e le proprietà dei coefficienti. Proprio attraverso l'osservazione del p-value, ed una accettazione o rifiuto debole/medio/forte (o analogamente tramite le tStat ed il calcolo del valore critico del test d'ipotesi associato) si definisce la significatività, o non, degli stessi.

Successivamente abbiamo analizzato l'adattamento: le misure di bontà di adattamento in genere sintetizzano la discrepanza tra i valori osservati e i valori previsti dal modello in questione.

Per selezionare il più adatto abbiamo invece fatto uso del comando "stepwiselm", aggiungendo o togliendo variabili in modo automatizzato, ottenendo così: Ozono ~ 1 + PM10*Temperatura + Temperatura*Umidità. Da quest'ultima analisi è emerso che la temperatura sia influente sulla concentrazione di ozono nell'aria, il che non sorprende poiché questa è solita aumentare in estate, quando le temperature sono più alte.

Medesimi passaggi logici e pratici sono stati effettuati sulle misurazioni della stazione di Bergamo.

Qui di seguito riportiamo l'output Matlab per il modello di regressione lineare:

```
Linear regression model:
    Ozono ~ 1 + PM10 + Temperatura + Pioggia + Umidita

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    _________

    (Intercept)      95.052      7.0602      13.463    1.5811e-29
    PM10            -0.10938    0.05765     -1.8973     0.059287
    Temperatura       2.8093     0.1302      21.576    3.2509e-53
    Pioggia         0.040446    0.032299     1.2523         0.212
    Umidita          -1.0777    0.084114    -12.812    1.4662e-27


Number of observations: 197, Error degrees of freedom: 192
Root Mean Squared Error: 10.2
R-squared: 0.912,  Adjusted R-Squared: 0.91
F-statistic vs. constant model: 498, p-value = 3.75e-100
```

e per il modello di regressione lineare realizzato con il commando "stepwise"

I coefficienti ottenuti sono significativi?
Analizziamo il P-value. Intercetta: PV=1.3393e-16, rifiutiamo l'ipotesi nulla, il coeff. è significativo. PM10: PV= 0.0077467 , rifiutiamo l'ipotesi nulla, il coeff. è significativo.
Temperatura: PV= 5.4476e-46, rifiutiamo l'ipotesi nulla, il coeff. è significativo.
Pioggia: PV= 0.42337, accettiamo l'ipotesi nulla, non è un coeff. significativo.
Umidità: PV= 8.9753e-16, rifiutiamo l'ipotesi nulla, il coeff. è significativo.

Una visione più ampia è possibile tramite l'osservazione di F-statistic: il nostro P-Value è -->0 , rifiuto quindi l'ipotesi nulla (la quale era: tutti i coeff. sono contemporaneamente non significativi). Si parla di RIGETTO FORTE.

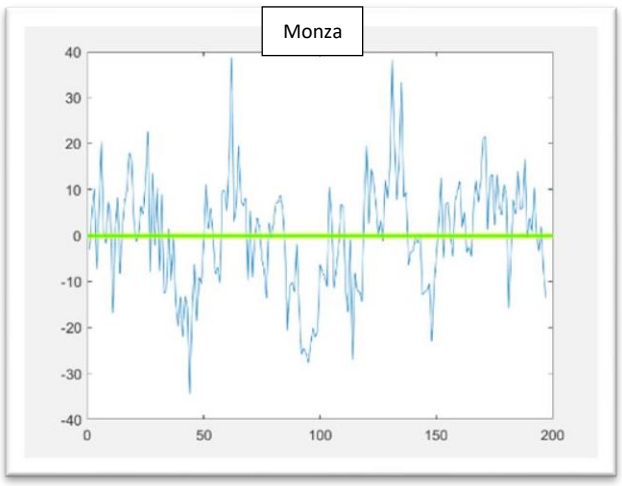
```
Linear regression model:
    Ozono ~ 1 + PM10 + Temperatura + Pioggia + Umidita

Estimated Coefficients:
                   Estimate       SE        tStat       pValue
                   ________    ________    _______    _________

    (Intercept)      95.052      7.0602      13.463    1.5811e-29
    PM10            -0.10938    0.05765     -1.8973     0.059287
    Temperatura       2.8093     0.1302      21.576    3.2509e-53
    Pioggia         0.040446    0.032299     1.2523         0.212
    Umidita          -1.0777    0.084114    -12.812    1.4662e-27


Number of observations: 197, Error degrees of freedom: 192
Root Mean Squared Error: 10.2
R-squared: 0.912,  Adjusted R-Squared: 0.91
F-statistic vs. constant model: 498, p-value = 3.75e-100
>> stepwiselm (tab)
1. Adding Temperatura, FStat = 862.4885, pValue = 1.627908e-73
2. Adding Umidita, FStat = 196.2025, pValue = 2.925751e-31
3. Adding Temperatura:Umidita, FStat = 10.8917, pValue = 0.00115053
4. Adding PM10, FStat = 14.8672, pValue = 0.000157305
5. Adding PM10:Temperatura, FStat = 9.0456, pValue = 0.0029878
Linear regression model:
    Ozono ~ 1 + PM10*Temperatura + Temperatura*Umidita

Estimated Coefficients:
                             Estimate       SE        tStat       pValue
                             ________    ________    _______    _________

    (Intercept)                61.297      9.6345      6.3622    1.4375e-09
    PM10                     -0.039818    0.075724    -0.52583       0.59962
    Temperatura                5.4881     0.61534      8.9188    3.8125e-16
    Umidita                   -0.64038     0.13101      -4.888    2.1521e-06
    PM10:Temperatura         -0.019966   0.0066386     -3.0076     0.0029878
    Temperatura:Umidita      -0.030693    0.009227     -3.3264     0.0010549


Number of observations: 197, Error degrees of freedom: 191
Root Mean Squared Error: 9.61
R-squared: 0.923,  Adjusted R-Squared: 0.921
F-statistic vs. constant model: 459, p-value = 2.57e-104
```

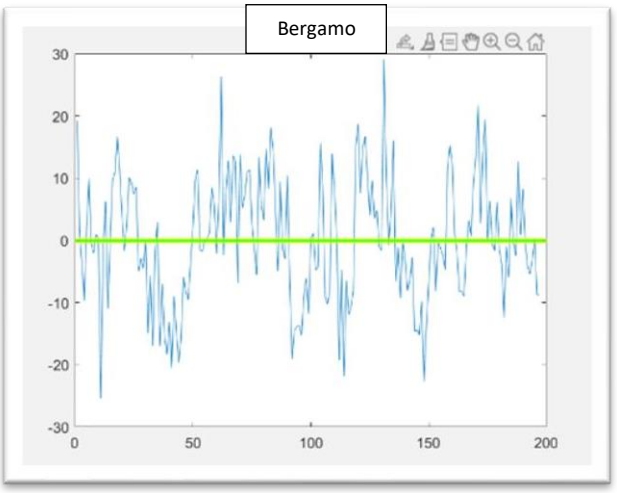
ANALISI DEI RESIDUI:

Dopo l'analisi di regressione abbiamo eseguito alcuni test sui residui per avere un'ulteriore conferma della validità del modello. Considerando che il modello di regressione lineare si regge sull'ipotesi che la media dei residui sia pari a zero, abbiamo controllato la validità di tale ipotesi sia attraverso metodi grafici che un metodo analitico.

METODO 1: PLOT

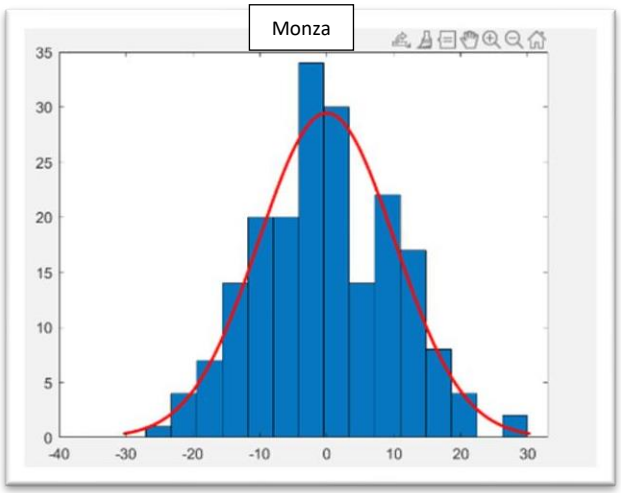
In questo caso abbiamo considerato la distribuzione dei residui, ovvero il plot dei punti, rispetto alla retta y = 0 (gialla nel grafico). Abbiamo osservato che si ottiene una disposizione omogenea dei punti e questo significa che il modello è corretto. Calcolando poi la media di tale distribuzione abbiamo ottenuto la retta verde che si sovrappone alla retta y = 0, dunque la media dei residui risulta essere effettivamente nulla.

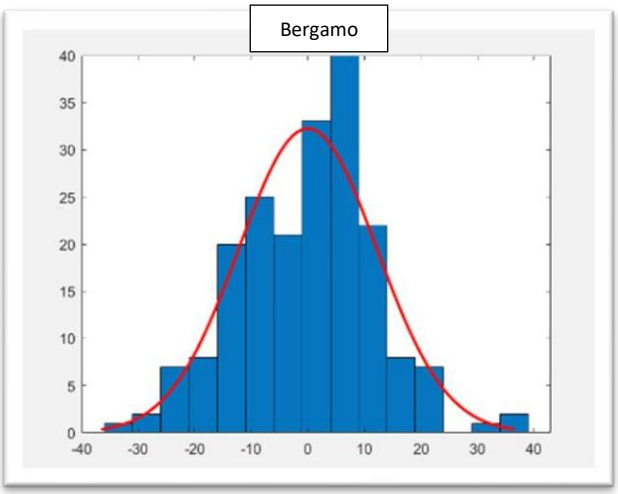




METODO 2 HISTFIT

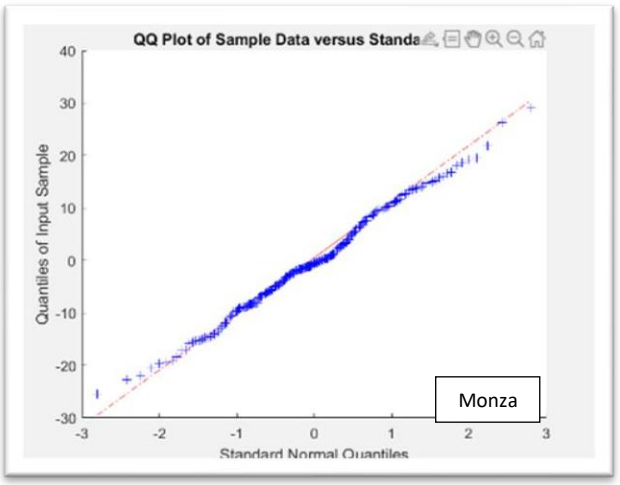
Abbiamo fatto poi un'altra semplice verifica sulla distribuzione dei dati attraverso il comando "histfit" di matlab, il quale traccia il diagramma dei residui e sovrappone ad esso una distribuzione normale e abbiamo osservato che i dati seguono bene quest'ultima distribuzione.

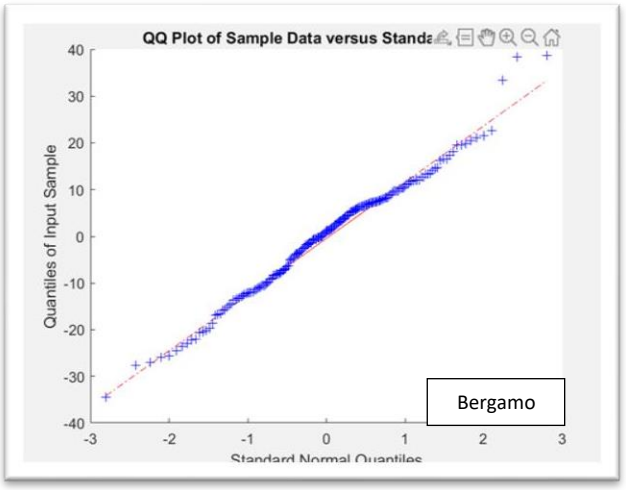




METODO 3: Q-Q PLOT

In un Q-Q plot, i quantili osservati vengono confrontati con i quantili attesi nel caso la distribuzione fosse normale su uno stesso diagramma cartesiano. Se i punti si dispongono lungo una retta, la distribuzione approssima bene la normale. Nel nostro caso il grafico q-q dei dati grezzi segue bene l'andamento della retta e quindi la regressione rappresenta un modello adeguato.





METODO 4: Test di Shapiro-Wilk

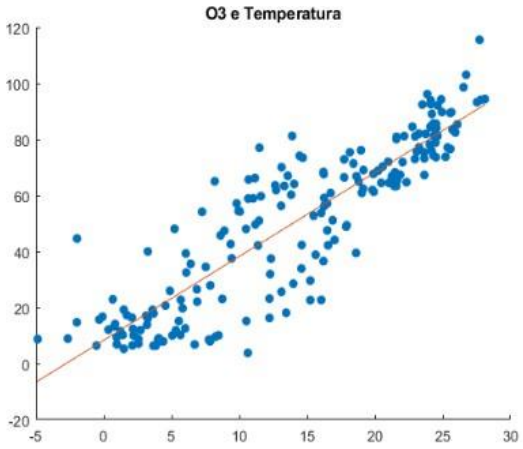
Abbiamo deciso di utilizzare questo metodo analitico perché questo test risulta particolarmente accurato nel caso di un "esiguo" numero di campioni (197 nel nostro caso). Il test pone come ipotesi nulla la distribuzione normale dei dati.
Utilizzando la funzione "swtest" abbiamo ottenuto come output del test:

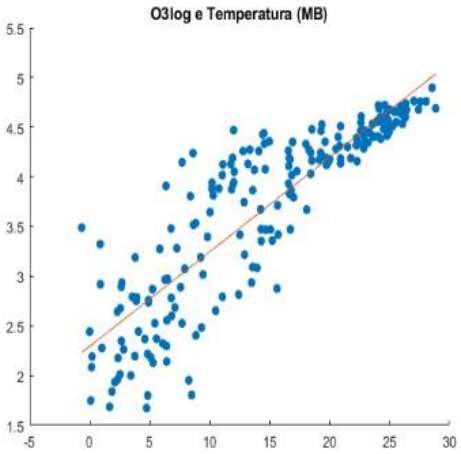
- per Monza: p-value pari a 0.6327 e il valore del test W = 0.9941
- per Bergamo: p – value pari a 0.0513 e il valore del test W = 0.9863

In entrambi i casi possiamo concludere che il p-value è maggiore del livello di significatività alpha pari a 0.05 e i valori della statistica test W molto vicini ad 1, dunque possiamo concludere che non rifiutiamo l'ipotesi nulla del test, e quindi il campione analizzato proviene da una popolazione con la distribuzione normale.

Correlazione Logaritmica:

Il modello logaritmico risulta particolarmente efficace nel momento in cui i dati non seguono una distribuzione normale (non nel nostro caso) per correggerne la distribuzione. Confrontando il modello di regressione tra le due provincie risulta chiaro come in entrambi i casi la quantità di ozono presente nell'aria risulti crescente o meno in maniera assai similare. Questa conclusione risulta in linea con le caratteristiche geografiche dei due luoghi, posti considerevolmente vicini tra loro. Le considerazioni ricavate sono comunque le medesime riscontrate nell'analisi sui dati grezzi, a conferma di ciò che è già stato esposto. Va comunque detto come dall'analisi dei dati sembrerebbe esistere una relazione tra PM10 ed ozono, la quale è solo casuale, poiché non esiste correlazione scientifica tra i due.





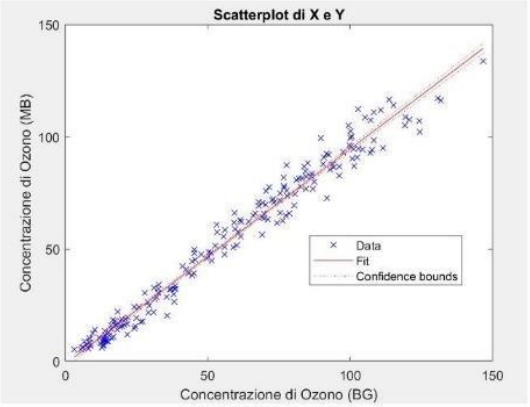
INTERPRETAZIONE MODELLO LOG-LINEARE

Esso ci permette di linearizzarla e ricondurre quindi il modello ad una normale.
Come forma ha: ln(Y)==β0+β1X+e.
La trasformazione in logaritmo della nostra variabile dipendente ha come effetto quello che all'aumentare di 1 di X la nostra Y aumenterà del 100%*β1.
Ad esempio, se β1=0.2 e la nostra X aumenta di 1 unità la nostra Y aumenterà del 20%.
Questo a differenza del modello Lineare-Lineare dove all'aumentare di 1 unità di X la nostra Y aumenterà di β1*X.
Qui di seguito riportiamo il modello estratto dal commando fitlm di matlab:

```
Linear regression model:
    logO3_tG1 ~ 1 + PM10_tG1 + Temperatura_tG1 + Pioggia_cum_tG1 + Umidita_relativa_tG1

Estimated Coefficients:
                           Estimate        SE         tStat       pValue
                           ________    _________    _______    __________

    (Intercept)              5.4725      0.19072     28.693    2.3343e-71
    PM10_tG1               -0.010057    0.0015573   -6.4579    8.4946e-10
    Temperatura_tG1         0.055707    0.0035173    15.838    1.0928e-36
    Pioggia_cum_tG1        0.0020565    0.0008725     2.357       0.01943
    Umidita_relativa_tG1   -0.031426    0.0022722    -13.83    1.2244e-30


Number of observations: 197, Error degrees of freedom: 192
Root Mean Squared Error: 0.277
R-squared: 0.905,  Adjusted R-Squared: 0.903
F-statistic vs. constant model: 458, p-value = 5.96e-97
```

Correlazione tra le provincie di Bergamo e Monza e Brianza:

Il lavoro da noi svolto intende verificare l'eventuale correlazione tra le provincie di Monza e Brianza e Bergamo nel campo della concentrazione di ozono nell'aria.
Con correlazione definiamo la misura statistica esprimente la relazione lineare due o più variabili. In particolare faremo affidamento all'indice di correlazione "r di Pearson",
attraverso il quale siamo in grado di determinare la forza e la direzione della una relazione lineare tra due variabili continue, rappresentandone il grado di concordanza e discordanza.
L'analisi di correlazione tra le due provincie in questione avviene attraverso lo studio della covarianza.



MODELLO GRAFICO DEL COMANDO FITLM

Difatti, attraverso il comando "Varcov" di matlab siamo stati in grado di verificare una covarianza pari a 1.1893; ciò indica come BG ed MB covariano positivamente.
Da definizione, una relazione positiva significa che gli individui che ottengono valori elevati in una variabile tendono ed ottenere valori elevati sulla seconda variabile. Ed è vero anche viceversa, cioè coloro che hanno bassi valori su una variabile tendono ed avere bassi valori sulla seconda variabile.
E' doveroso fare menzione del fatto che non ci aspettiamo, però, un rapporto causale tra i dati.
Per ovviare all'eventuale problema dei dati mancanti abbiamo utilizzato le funzioni disponibili del comando "Corr" di Matlab, attraverso il quale abbiamo ricavato una correlazione al 98.74% tra i dati, rendendo chiara la similitudine tra i due.
Queste conclusioni possono ottenere ulteriore significato se considerata la "misera" distanza, a livello geografico, che separa le due provincie.

Bibliografia:

• http://www.arpa.piemonte.it/approfondimenti/temi-ambientali/aria/aria/cartella-qualitaepisodi-acuti-di-inquinamento-daozono#:~:text=Le%20condizioni%20meteorologiche%20di%20intenso,)%20la%20formazione%20dell'Ozono.&text=A%20parit%C3%A0%20di%20irraggiamento%20solare%20un%20aumento%20della%20temperatura,aumento%20della%20formazione%20di%20Ozono.

• https://www.arpa.puglia.it/c/document_library/get_file?uuid=4ef9f653-3fcb-45a8-9a6fe1669715f220&groupId=13879

• https://www.lenntech.it/biblioteca/ozono/generazione/ozono-produzione.htm

•https://www.mathworks.com/matlabcentral/mlcdownloads/downloads/submissions/13964/versions/2/previews/swtest.m/index.html

•https://www.lenntech.it/biblioteca/ozono/generazione/ozonoproduzione.htm#:~:text=L'aria%20ambiente%20contiene%20elevata,rendimento%20dell'ozono%20per%20KWh.&text=Inoltre%2C%20si%20formano%20idrossido%2Dradicali,ozono%20%5B3%2C 5%5D.

•https://www.arpalombardia.it/qariafiles/RelazioniMM/RMM_Sovere_20181115.pdf

•https://link.springer.com/chapter/10.1007%2F88-470-0384-9_2