Regressione Lineare Multipla

Silviu Filote 1059252,
Nicolò Carissimi 1069015,
Jonathan Bommarito 1068755

January 5, 2021

Contents

1 Introduzione

Il dataset originale al centro della nostra analisi consiste in un elenco di osservazioni riguardanti dati sulla qualità dell'aria e metereologia misurata dal 2017 fino a settembre del 2020.
I dati sono stati confrontati tra due stazioni in due località lombarde.

L'obiettivo è stato quello di produrre i miglior modelli di Regressione delle 2 stazioni partendo dalle variabili metereologiche che siano in grado di spiegare la variabilità di un'inquinante dell'aria presente nel suddetto dataset.
Una volta ottenuti i due modelli si sono analizzati i risultati e tratte le principali correlazioni.

2 Risoluzione adottata

Si è partiti importando il dataset con le rispettive stazioni da analizzare, ossia quelle di Bergamo ed Erba, e si è scelta $\mathbf{NO_x}$ come variabile da studiare in funzione delle altre.
Una volta scelta la variabile dipendente si sono aggiunte tutte le variabili disponibili nel dataset e tramite il comando matlab "fitlm" si è calcolato il modello di regressione multipla.
Successivamente tramite passi iterativi si è stimato il miglior modello lineare:

- tutti i p-value dei regressori vengono confrontati con il livello di significatività al 95%, scartando quelli non significativi;
- si analizza nuovamente il modello di regressione con le covariate rimaste, tenendo conto del R-adjusted;
- le variabili in questo caso scartate mantengono invariato l'R-adjusted o al più ne causano l'aumento;
- se i regressori sono significavi il valore del R-adjusted risulta essere il più elevato, ciò implica che il modello è rappresentato nel migliore dei modi;

Il procedimento si ripete fino a quando tutte le covariate risultano essere, in base all'analisi del p-value, significative.

Dopo aver stimato i due modelli si è deciso di prendere in considerazione anche la poszione geografica delle due stazioni, nell'eventualità di poter effettuare ulteriori considerazioni.

3 Risultati ottenuti

Per giungere al miglior modello lineare per la stazione di Erba, sulla base dell'algoritmo precedentemente esplicato, si è deciso di eliminare le variabili metereologiche:

- Ozono (Pv : 0.72159 %);
- Pioggia (Pv : 0.51251 %);

```
erba =


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

Estimated Coefficients:
                    Estimate       SE        tStat       pValue
                    ________    ________    _______    __________

    (Intercept)      -38.252      5.9146    -6.4673    8.0692e-10
    NO2               2.0947     0.11411     18.357    4.0193e-44
    Temperatura      0.38166     0.15687     2.4329      0.015894
    Umidita          0.21933    0.054032     4.0592    7.1636e-05
    PM10             0.33343    0.089087     3.7428    0.00024032


Number of observations: 197, Error degrees of freedom: 192
Root Mean Squared Error: 8.44
R-squared: 0.929,  Adjusted R-Squared: 0.928
F-statistic vs. constant model: 632, p-value = 2.74e-109
```

Figure 1: fitlm - Erba

stazione di riferimento: 564

Indirizzo: Erba - via Battisti,279,CO,Erba
Coordinate: 45.80857380692296, 9.22177920448843

Per giungere al miglior modello lineare per la stazione di Bergamo, sulla base dell'algoritmo precedentemente esplicato, si è deciso di eliminare le variabili metereologiche:

- Ozono (Pv : 0.6072 %);
- Temperatura (Pv : 0.35616 %);
- Pioggia (Pv : 0.15267 %) → guardare considerazioni.

```
bg =


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

Estimated Coefficients:
                   Estimate      SE        tStat       pValue
                   ________    _______    _______    __________

    (Intercept)    -53.133      6.7156    -7.9119    1.9253e-13
    NO2             2.2638     0.12814     17.666    3.4867e-42
    Umidita        0.41824     0.10361     4.0367    7.8106e-05
    PM10           0.35046     0.11597     3.0221      0.002851


Number of observations: 197, Error degrees of freedom: 193
Root Mean Squared Error: 15.9
R-squared: 0.868,  Adjusted R-Squared: 0.866
F-statistic vs. constant model: 423, p-value = 1.56e-84
```

Figure 2: fitlm - Bergamo

stazione di riferimento: 583

Indirizzo: Bergamo - via Meucci,249,BG
Coordinate: 45.69103740547214, 9.643650579461385

4 Considerazioni

4.1 Summary dei dati

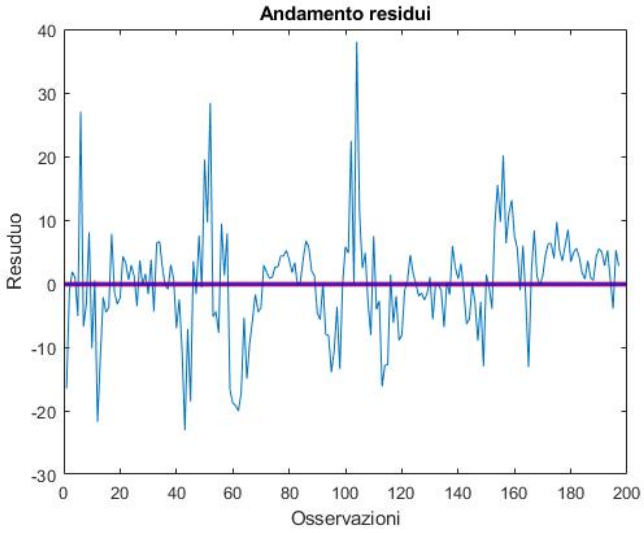
Erba							
Stime	NO2	NOx	PM10	Pioggia	Temperatura	Umidità	Ozono
Min	5.3228	6.2275	6	0	-0.89157	35.478	7.9571
Max	66.274	173.02	71.143	156.4	29.54	98.408	153.06
Mean	21.412	33.942	21.422	22.623	14.814	66.32	66.555

Bergamo							
Stime	NO2	NOx	PM10	Pioggia	Temperatura	Umidità	Ozono
Min	8.8423	9.3083	6.3333	0	-0.53938	38.417	3.15
Max	75.127	225.8	108	123.2	30.68	94.987	146.7
Mean	28.156	49.26	28.637	22.332	15.61	68.424	56.769

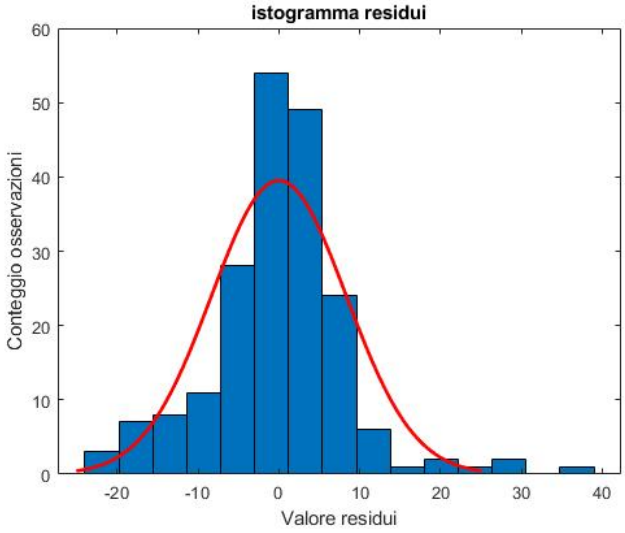
Analisi dei dati:

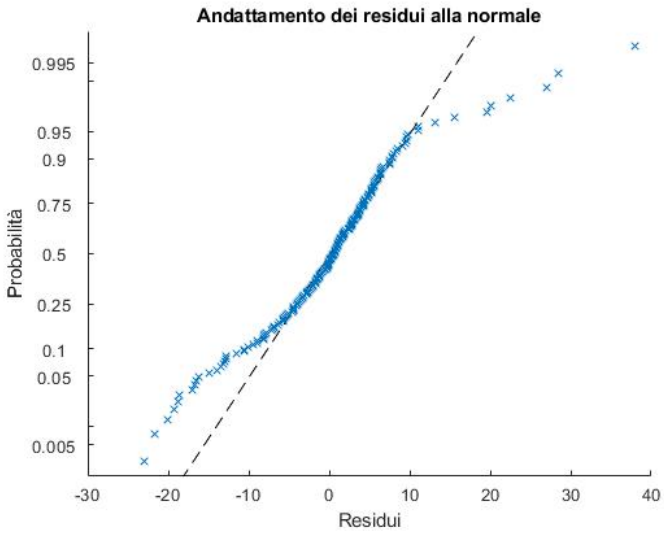
- L'inquinamento dell'aria di Bergamo risulta essere mediamente più elevato rispetto ad Erba, infatti si riscontra un aumento di NO2 e PM10, che a loro volta causano un incremento di NOx.
  Questa osservazione è compatibile con la realtà fortemente industrializzata di Bergamo;
- Le concentrazioni di ozono media di Erba sono maggiori rispetto a quelle di Bergamo. Non avendo nel dataset osservazioni riguardanti l'irraggiamento solare nei due comuni non è possibile fare considerazioni ulteriori che spieghino questo fenomeno.

4.2 Analisi dei residui - Erba



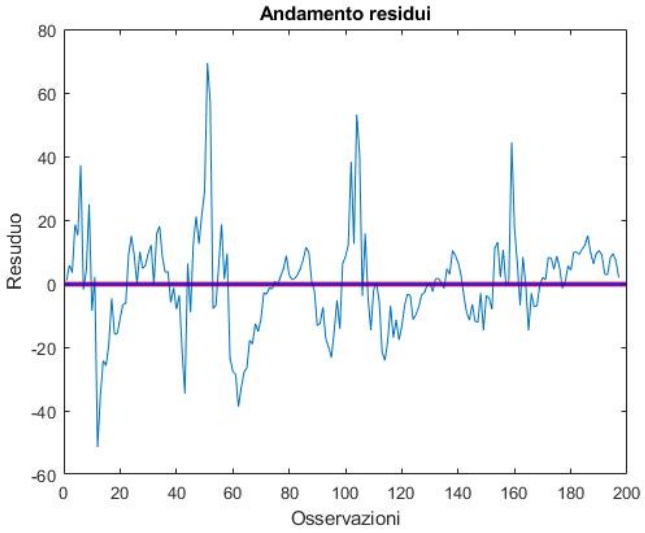
In merito al grafico di cui sopra si può osservare che, com'era prevedibile, avendo utilizzato il metodo dei minimi quadrati, si è ottenuta una media dei residui molto vicina a 0.



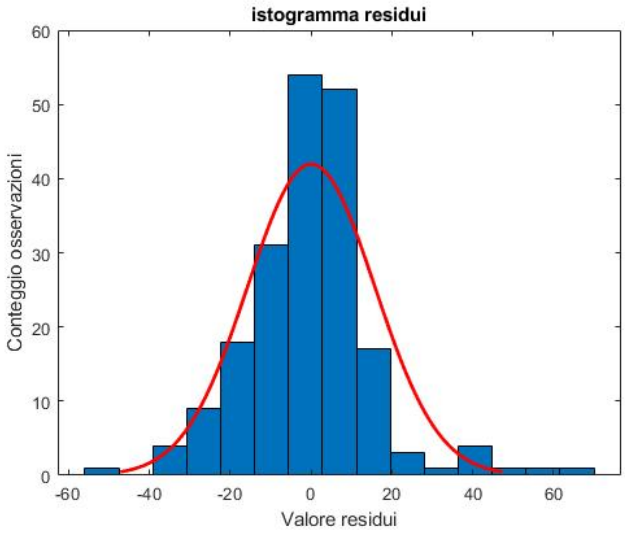


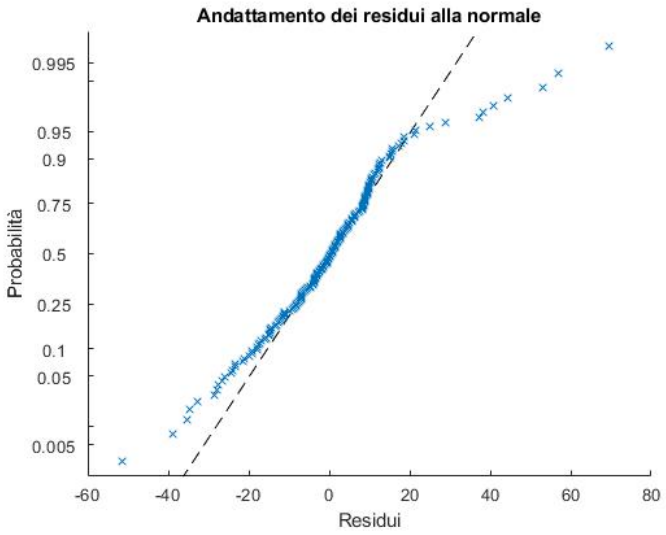
Come visibile da entrambi i grafici si denota una skewness positiva, (i valori di destra sono più distanti dalla media rispetto a quelli di sinistra) mentre per quanto riguarda la coda di sinistra vi sono più punti che si discostino dall'andamento normale.

4.3 Analisi dei residui - Bergamo



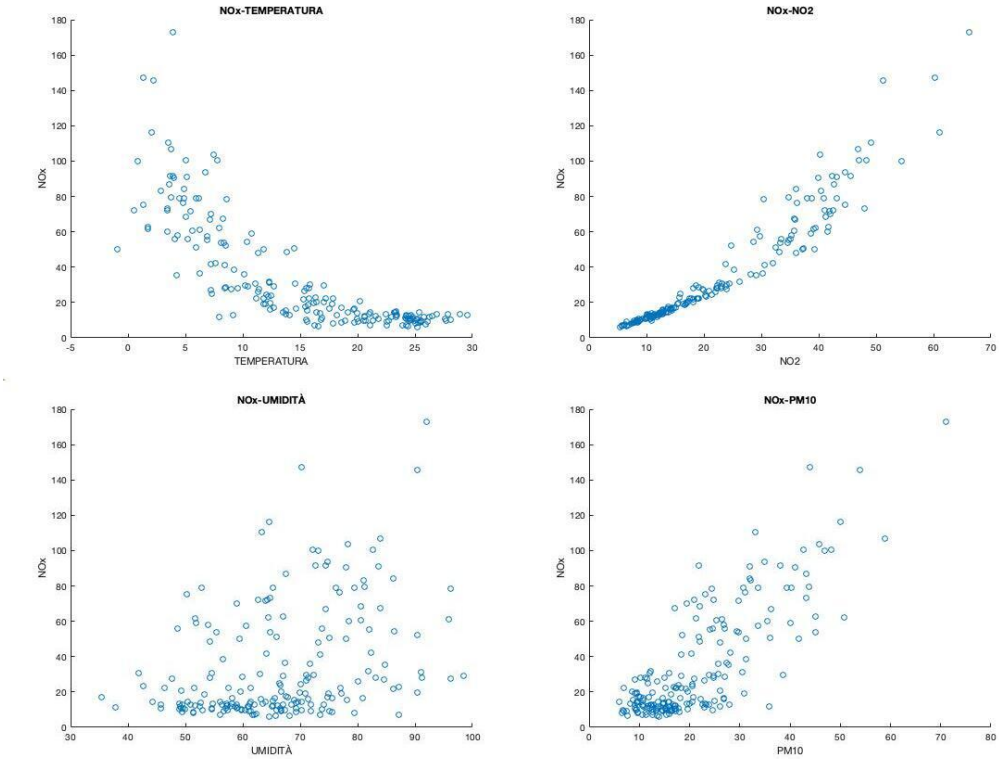
Come nel caso precedente si è ottenuta una media dei residui pari a 0 per le medesime considerazioni suddette.





Anche in questo caso la skewness risulta essere positiva, mentre per quanto concerne la coda di sinistra è possibile affermare che quest'ultima si adatta leggermente meglio alla distribuzione normale rispetto al grafico di Erba, infatti i punti risultano essere più vicini alla retta.

4.4 Analisi coefficienti dei regressori - Erba



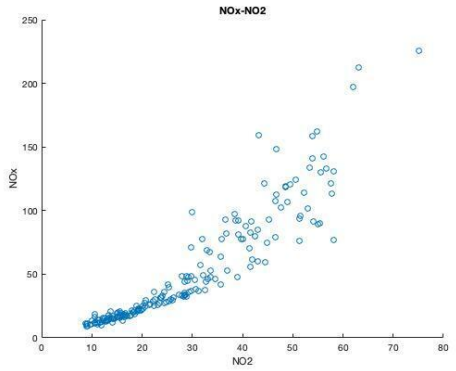
I grafici precedenti ci consentono di stimare se esistono delle relazioni tra variabili indipendenti, prese singolarmente, e l'NOx.

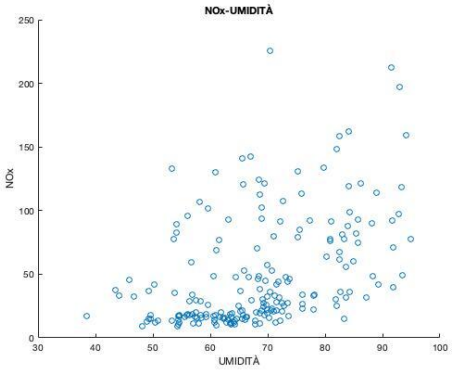
Osservando il primo ci si aspetterebbe una correlazione negativa, quindi all'aumentare della temperatura corrisponderebbe una diminuzione di NOx. *É* interessante notare come questo non accada se si inserisce la variabile temperatura all'interno del modello di regressione multipla :

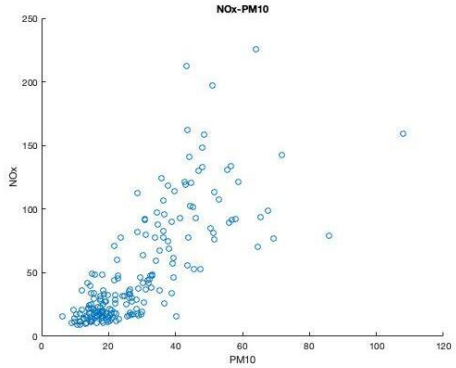
$$NOx \sim 1 + NO2 + temperatura + umidita + PM10$$

Questo cambiamento di segno è dovuto alla correlazione che esiste tra i regressori "temperatura" e "NO2", fenomeno che prende il nome di **"multicollinearità"**.

4.5 Analisi coefficienti dei regressori - Bergamo







Per quanto riguarda il comune di Bergamo si è deciso di non utilizzare nuovamente il regressore pioggia e nemmeno il regressore temperatura.

La variabile pioggia merita una trattazione ulteriore, in quanto la sua aggiunta al modello provoca un debole innalzamento del R-adjusted. Benchè questo aumento sia generalmente segno di miglioria, si è deciso di non utilazzare tale variabile a causa di un P-Value decisamente non significativo.

bg =

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

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	-55.511	6.8989	-8.0463	8.6587e-14
NO2	2.2473	0.12831	17.514	1.158e-41
Pioggia	-0.075276	0.052426	-1.4358	0.15267
Umidita	0.51033	0.12161	4.1964	4.1432e-05
PM10	0.28839	0.12346	2.3359	0.020531

Number of observations: 197, Error degrees of freedom: 192
Root Mean Squared Error: 15.8
R-squared: 0.869, Adjusted R-Squared: 0.867
F-statistic vs. constant model: 319, p-value = 1.25e-83

Figure 3: fitlm - Bergamo con pioggia

4.6 Correlazione tra i due comuni

In entrambi i comuni la covariata più significativa è l'NO2, seguita dall'umidità e dal PM10.

I coefficienti di correlazione in entrambi i modelli presentano lo stesso segno.

Per quanto riguarda le performance di adattamento del modello del comune di Erba risulta essere leggermente migliore:

- Erba : 0.928
- Bergamo : 0.866