PROGETTO DI STATISTICA E MODELLI STOCASTICI (II MODULO)

Università degli studi di Bergamo 2020/2021, 16 aprile 2021

Nome del gruppo: G25_MONACO

Componenti del gruppo:
Kenna Miriam Fatima (1058218), Dandis Iana (1065350),
Luca Lorenzi (1068520), Hurtado Beltran Ariana (1065528)

Docenti del corso:
Francesco Finazzi
Frank Yannick Massoda Tchoussi

Contenuto:

1. **Introduzione**
   - 1.2 Descrizione del dataset
   - 1.3 Obiettivo
   - 1.4 Analisi grafica e conclusioni
   - 1.5 Vettore delle medie e matrice varianza-covarianza

2. **Regressione multipla**
   - 2.2 Prima analisi del dataset - stepwise
   - 2.3 Ulteriori analisi e stime dei coefficienti
   - 2.4 Test t sui coefficienti
   - 2.5 Analisi dei residui
     - 2.5.1 Analisi grafica dei residui
     - 2.5.2 Il test Jarque-Bera
     - 2.5.3 Omoschedasticità dei residui

3. **Tecniche di regressione avanzata**
   - 3.1 Minimi Quadrati pesati
   - 3.2 Descrizione algoritmo interattivo

4. **Regressione non parametrica**
   - 4.1 Basi di Fourier
   - 4.2 Costruzione del modello tramite Cross-Validazione Generalizzata (GCV)

5. **Crossvalidazione**

6. **Simulazione del modello con il dataset di validazione**
   - 6.1 Confronto grafico finale

1.1 Introduzione

La seguente relazione contiene l'analisi del dataset scelto dal gruppo G25 Monaco per la realizzazione del progetto attinente al corso di Statistica e Modelli Stocastici (MODULO II). L'analisi è stata condotta con l'ausilio del software Matlab ed applicando le tecniche e metodi statistici introdotti a lezione, oltre alle normali funzioni della Toolbox Statistic, sono state utilizzate funzioni incluse nella toolbox FDA.

1.2 Descrizione del dataset

Il dataset è composto dai dati richiesti attraverso il sito dell'Arpa Lombardia: riguardano la stazione di rilevamento collocata presso Milano Città Studi Pascal e coprono il periodo che va dal 13/01/2020 al 01/07/2020 con cadenza giornaliera. In particolare, il dataset è composto da:

- **Inquinanti:** Ozono, Biossido di azoto, Benzene, PM10, PM2,5 , Ammoniaca;
- **Variabili metereologiche**: Temperatura, Umidità relativa.

Sempre inerenti allo stesso periodo, al dataset così ottenuto sono stati aggiunti i dati sugli spostamenti in macchina a Milano, secondo i dataset pubblicati da Apple, ricavati attraverso la richiesta di indicazioni stradali alle mappe della società di Cupertino.

1.3 Obiettivo

L'obiettivo finale dello studio è rispondere alla seguente domanda: qual è la relazione esistente tra la presenza dell'ozono nell'aria e gli altri inquinanti/ fattori? Rispondiamo utilizzando la statistica descrittiva.

1.4 Analisi grafica e conclusioni

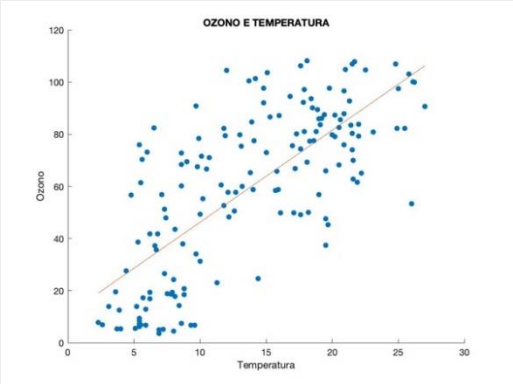
Il primo passo è caricare in Matlab il nostro database ['*database.xlsx*'] completo di tutte le variabili che andremo ad utilizzare, eliminando le celle prive di valore. Tramite l'interpretazione preliminare dei grafici è possibile mettere in evidenza l'eventuale correlazione tra la variabile risposta Ozono e i regressori presi singolarmente.

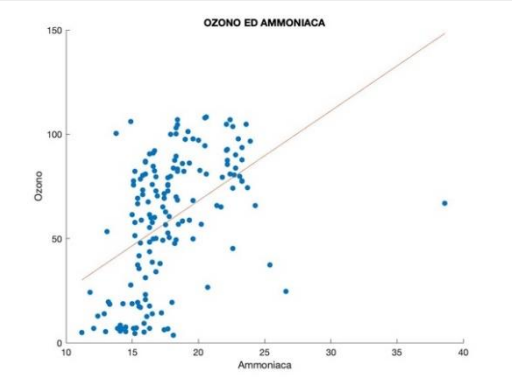
Quello che è stato possibile concludere dall'osservazione dei grafici riportati sulla pagina successiva è che sembra esserci una correlazione positiva tra Ozono e Temperatura, ed Ozono ed Ammoniaca; mentre sembriamo ritrovare una correlazione negativa tra Ozono e gli altri 6 regressori (benzene, NO2, PM10, PM2_5, spostamenti, umidità).

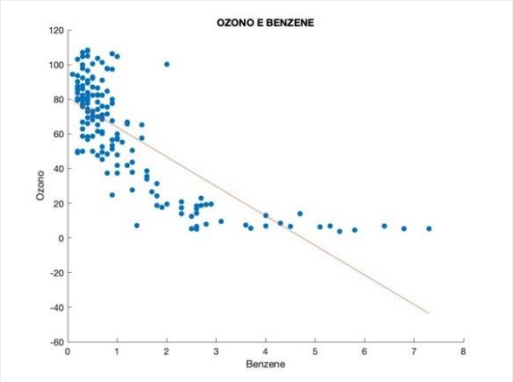
1.5 Vettore delle medie e matrice varianza-covarianza

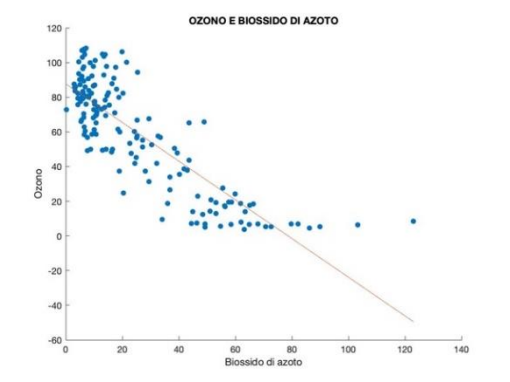
Successivamente calcoliamo la media dei valori osservati per ogni regressore, sempre per poter avere più informazioni possibili; otteniamo così un vettore di medie. Calcoliamo anche la matrice varianza-covarianza tramite l'apposito comando *cov(X)* . Essendo X una matrice le cui colonne rappresentano variabili casuali e le cui righe rappresentano osservazioni, otterremo la matrice di covarianza con le corrispondenti varianze di colonna lungo la diagonale.

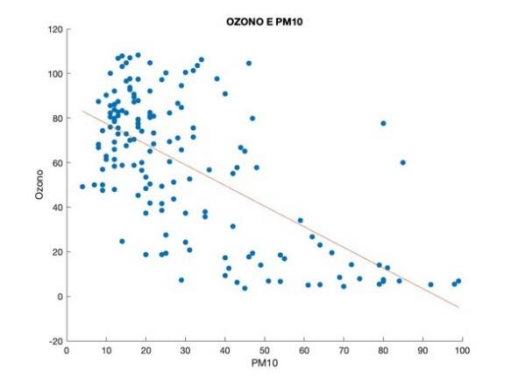
Affinché la matrice *varcov* sia ben definita, dobbiamo accertarci che essa sia una matrice (semi)definita positiva, e quindi che tutti gli autovalori siano strettamente positivi. Questo è il nostro caso, infatti l'autovalore minimo risulta essere pari a 0.3261.

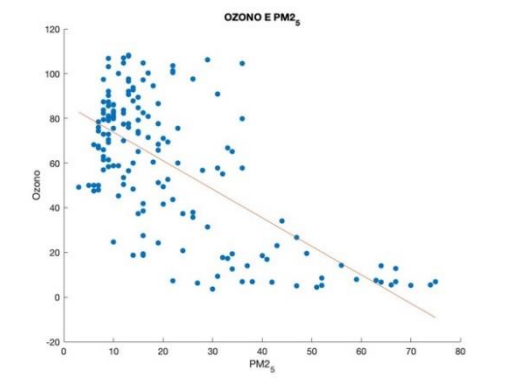

OZONO E TEMPERATURA
Ozono
Temperatura

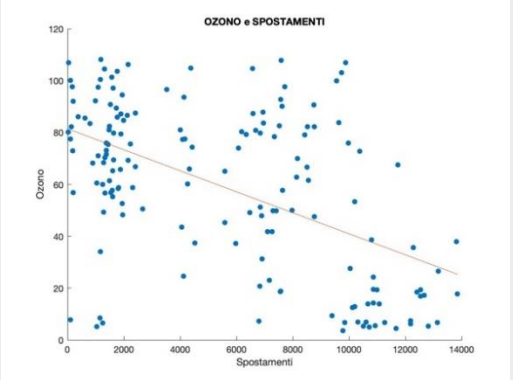

OZONO ED AMMONIACA
Ozono
Ammoniaca

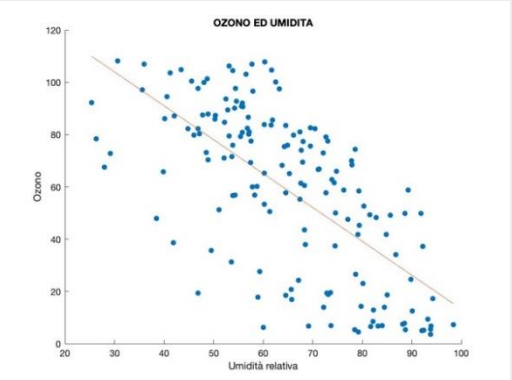

OZONO E BENZENE
Ozono
Benzene


OZONO E BIOSSIDO DI AZOTO
Ozono
Biossido di azoto


OZONO E PM10
Ozono
PM10


OZONO E PM2.5
Ozono
PM2.5


OZONO e SPOSTAMENTI
Ozono
Spostamenti


OZONO ED UMIDITA
Ozono
Umidità relativa

2 Regressione Multipla

2.1 Prima analisi del dataset: stepwise

Per poter comprendere in maniera analitica la capacità dei vari inquinanti/fattori di influenzare la concentrazione di Ozono nell'aria abbiamo deciso di aggiungergli e togliergli singolarmente dal modello calcolando di volta in volta (con il commando *fitlm(…)* di Matlab) l'indice $\mathbf{R^2}$ e $\mathbf{R^2_{corr}}$.

Infatti, l'indice $\mathbf{R^2}$ è uno dei più immediati indicatori della bontà della regressione poiché esprime quanto della variabilità complessiva della variabile dipendente **Y** nel nostro caso **Ozono** si può attribuire al legame lineare stimato mediante la retta di regressione, il resto infatti esprime la parte non spiegata e verrà aggiunto alla componente che esprime complessivamente gli errori.

L'indice $\mathbf{R^2_{corr}}$ invece è stato calcolato per tenere conto anche del numero delle variabili esplicative **X** incluse nel modello.

La combinazione che meglio spiega il modello è quella che include le seguenti variabili indipendenti:*'Temperatura','SpostamentiInMacchina','Umidit_Relativa','Ammoniaca','PM10','PM2_5', 'BiossidoDiAzoto'.* In questo modo otteniamo un $R^2 = 0.891$, valore molto vicino ad *0.909* ottenuto usando la 'stepwise' automatica (comando *stepwise(…)*). Quest'ultimo aspetto verrà ripreso e approfondito, tramite la tecnica di Cross-Validazione.

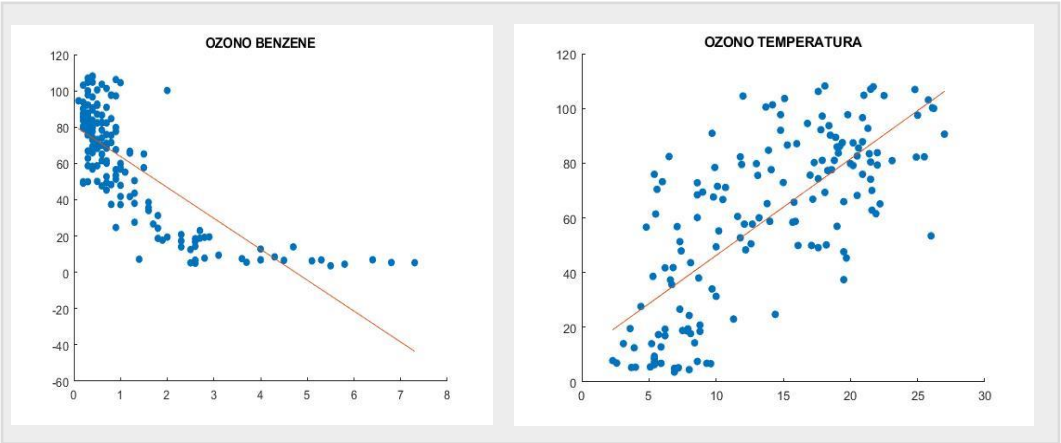
2.2 Ulteriori analisi e stime dei coefficienti

Per effettuare le successive indagini verifichiamo tramite la funzione *"numel()"* che tutte le nostre X abbiano la stessa dimensione, troviamo un n=160 e con questo dato possiamo calcolare devianza e varianza totale. Esse se scomposte in residua e spiegata, portano a notare che:
la varianza residua è decisamente più piccola rispetto alla spiegata (rispettivamente *var_sp=928,1932* e *var_res=112,7793*) quindi è possibile già aspettarci un coefficiente di determinazione elevato.
Il passo successivo, quindi, risulta essere il calcolo manuale dell'indice di determinazione del modello, includendo quindi tutte le variabili indipendenti. Otteniamo un indice pari a $0{,}8917$ molto vicino a 1, segno che i regressori descrivono bene l'andamento della nostra variabile dipendente Ozono.
Dal coefficiente di determinazione otteniamo il coefficiente di correlazione che risulta $0{,}944$ quindi vi è una correlazione positiva tra le variabili. Infine estrapoliamo i coefficienti beta dal modello e troviamo l'intervallo di confidenza al 95% tramite la funzione "model.coefCI".

2.3 Test-t sui coefficienti

Il nostro interesse successivo è verificare che i nostri coefficienti siano significativi nel modello: effettuiamo quindi un test-t. Questo test serve a confermare o rifiutare l'ipotesi nulla secondo cui il nostro coefficiente β_j sia significativo o meno. Esso include la creazione di una regione di accettazione data da $\pm t_{n-k-1}$,entro cui se rientra la nostra statistica *t* siamo portati all'accettazione dell'ipotesi nulla (cioè il coefficiente preso in esame non è significativo). Dai risultati osserviamo che benzene, PM10, PM2_5 e ammoniaca godono di una statistica *t* che rientra nell'intervallo di accettazione, di conseguenza accettiamo l'ipotesi nulla (beta=0) e possiamo dire che questi coefficienti non sono significativi. Questo è confermato anche da un punto di vista grafico, infatti, se tracciamo la retta di regressione ,ad esempio, tra ozono e benzene notiamo che i valori osservati non seguono la nostra retta (*fig 1a*), al contrario il grafico tra ozono e temperatura la quale ha superato il test-t mostra come i valori osservati seguano la nostra retta di regressione (fig 1b).

fig 1a *fib 1b*



2.4 Analisi dei residui

2.4.1 Analisi grafica dei residui

Come passo successivo alla stima del modello abbiamo voluto verificare la normalità dei residui. L'analisi grafica dei residui consente di valutare, a posteriori, se il modello ipotizzato è corretto, se questo fosse vero, infatti, gli errori dovrebbero distribuirsi secondo una normale. Nel nostro caso attraverso l'applicazione in Matlab dei comandi ,rispettivamente, *plot(…)*, *normplot(…)* e *histfit(…)* ai residui abbiamo potuto avere una prima conferma grafica sulla normalità dei residui.

Grafici riportati di seguito nella Figura 2.

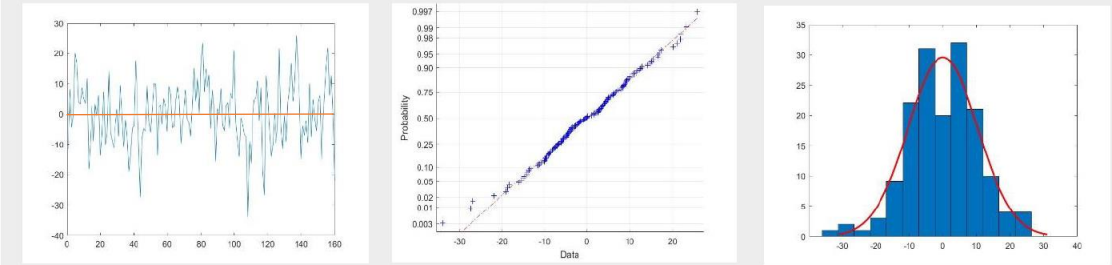


Figura 2: Analisi grafica dei residui

2.4.2 Il test di Jarque-Bera

Per avere un'ulteriore conferma della normalità, abbiamo eseguito in Matlab, con il comando *jbtest(…),* il test Jarque – Bera sui residui. Si tratta di un test di normalità che verifica come ipotesi nulla,simultaneamente, se l'asimmetria e la curtosi siano rispettivamente $k = 0$ e $sk = 3$ (valori che assumono sotto l'ipotesi nulla di normalità). Nel nostro caso implementando direttamente il jb-test sui residui, l'output risulta essere uguale a zero e quindi i residui si distribuiscono come una normale. In alternativa abbiamo calcolato direttamente l'asimmetria con il comando *skewness()* che risulta essere uguale a $k = -0{,}1557$ e la curtosi con il comando *kurtosis()* che risulta essere uguale a $sk = 3{,}3322$.

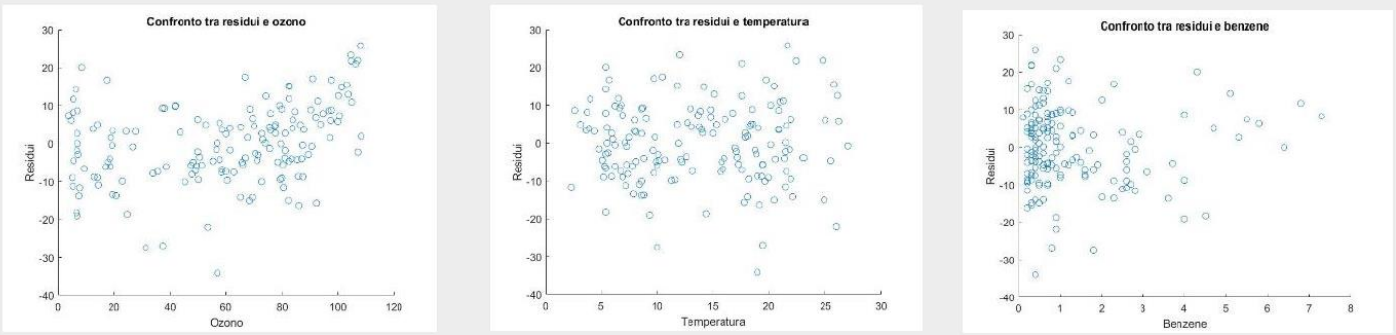
2.4.3 Omoschedasticità dei residui

Oltre ad assumere l'ipotesi che gli errori siano normali indipendentemente distribuiti e con $\mu = 0$, abbiamo anche assunto l'ipotesi che la loro varianza sia costante, per evitare i problemi che potrebbero derivare nel caso in cui quest'ultima ipotesi non fosse rispettata. Ovvero:
- le stime dei minimi quadrati sono corrette ma non sono efficienti;

- la stima della varianza, e quindi dell'errore standard è distorta e può compromettere i test di significatività.
Per verificare la varianza costante dei termini di errore abbiamo utilizzato il metodo grafico.
In presenza di omoschedasticità il grafico dei residui, infatti, dovrebbe presentarsi approssimativamente come una *nuvola di punti* che si dispone in modo casuale all'interno di una fascia orizzontale, esattamente quello che osserviamo nella distribuzione dei nostri dati.

Di seguito riportiamo alcuni confronti grafici nella Figura 3:

Figura 3



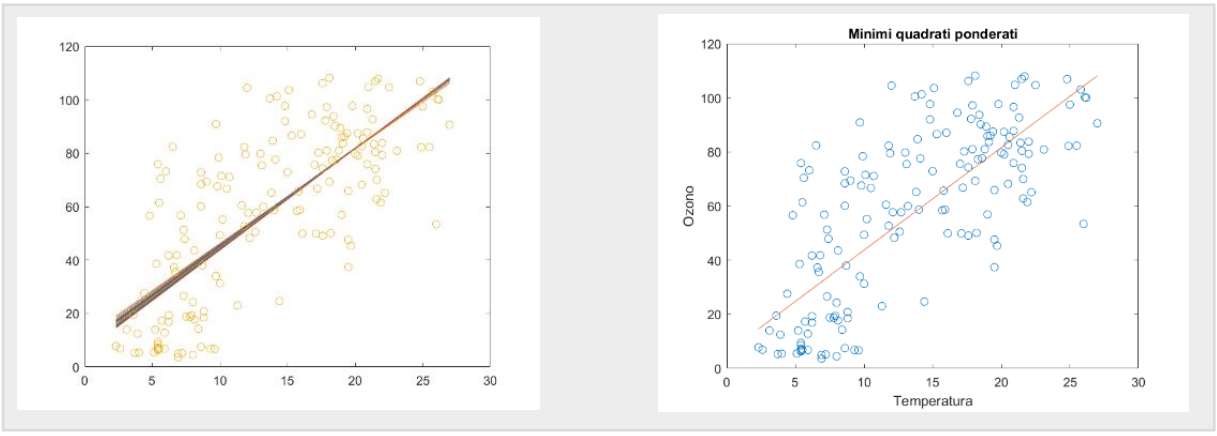
3. Tecniche di regressione avanzata

3.1 Minimi quadrati pesati

Successivamente a quanto già visto abbiamo utilizzato un metodo di regressione avanzata, in particolare: il Metodo dei Minimi Quadrati Pesati, per ottenere una retta di regressione più precisa. Per applicare in pratica questa tecnica di regressione avanzata abbiamo deciso di scegliere come variabile dipendente l'ozono e come variabile indipendente la temperatura omettendo per semplicità le altre variabili indipendenti, mentre il criterio utilizzato è stato il **'Criterio dei residui'**: è assegnato un peso maggiore ai punti con varianza minore, quindi più vicini alla retta e un peso minore ai punti più distanti dalla retta con la varianza maggiore.

Per ricavare i pesi assegnati abbiamo sviluppato in Matlab un algoritmo che tramite iterazioni successive ci restituisse il grafico con la retta che meglio stima il modello corrispondente, tenendo conto anche dei pesi assegnati alle varie misurazioni.

Nel nostro caso sono state sufficienti 28 interazioni per ottenere la reta finale, che riportiamo nelle figure:



4. Regressione non parametrica

4.1 Basi di Fourier

Lo scopo principale della regressione non parametrica è la modellazione degli effetti anche non lineari di variabili indipendenti sulla variabile dipendente. Vogliamo quindi stimare tramite l'utilizzo delle basi di Fourier, l'andamento della variabile 'Temperatura' e della variabile 'Umidità Relativa' in un dato periodo di tempo, che abbiamo deciso essere di 10 giorni.

La scelta ricade su queste variabili in quanto erano le uniche dotate di misurazioni con cadenza oraria. Scegliamo un periodo di t = 10 giorni (01:00 del 13/01/2020 – 00:00 del 23/01/2020) con osservazioni effettuate ogni 1h in modo da avere un andamento delle misurazioni con aspetto periodico, di periodo T = 24h = 1gg. Per quanto riguarda la Temperatura, l'accortezza è convertire tutte le misurazioni in Kelvin per poter lavorare con una scala di vettori che non includa dati negativi o nulli.

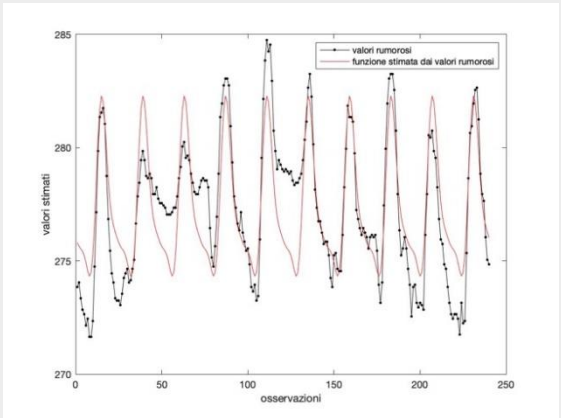
4.2 Costruzione del modello tramite Cross-Validazione Generalizzata (GCV)

Dal momento che sia la temperatura che l'umidità relativa presentano un andamento periodico, scegliamo di costruire un modello tramite l'utilizzo di basi di Fourier:

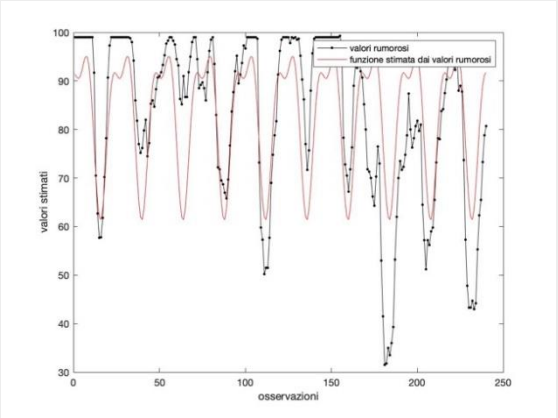
$\phi(x) \sim (1, \sin(\omega x), \cos(\omega x), \sin(2\omega x), \cos(2\omega x), \ldots, \sin(n\omega x), \cos(n\omega x))$ con $\omega = \frac{2\pi}{T}$

Tramite cicli iterativi, viene selezionato il numero di basi più adatto attraverso una cross-validazione generalizzata. Troviamo che il numero di funzioni di base ottimale (per entrambe le variabili studiate) è 5 (ordine 3, m = 2) e, a questo punto, creiamo il modello di stima confrontandolo con le misurazioni a nostra disposizione.

Temperatura



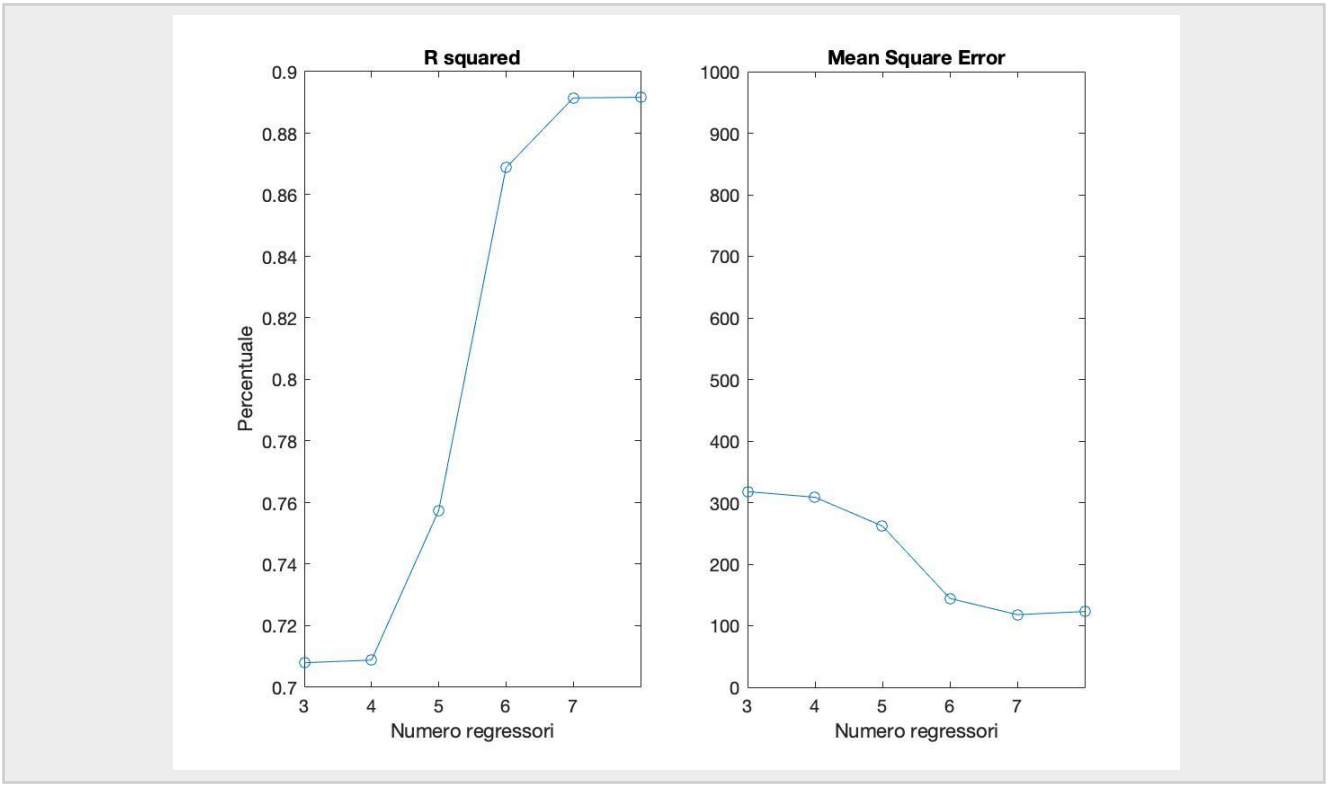
Umidità relativa



5. Cross-validazione

L'obiettivo di tale paragrafo è verificare l'efficacia e l'accuratezza del modello, tramite la valutazione progressiva dei valori di R-squared e del Mean Square Error: $MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$. Tramite il comando Matlab *crossval()* con l'applicazione di un metodo di ***crossvalidazione di tipo k-fold*** a 10 elementi abbiamo ottenuto un $\mathrm{MSE} \approx 317.8$ per il modello a due variabili e un $\mathrm{MSE} \approx 123.1$ per il modello a otto variabili, corrispondente a circa un terzo del valore iniziale; concludiamo quindi che convenga aggiungere al nostro modello i primi 7 regressori poichè sono in grado di spiegare in maniera sempre più ottimale i dati osservati.

Aggiungendo un regressore alla volta (nell'ordine poniamo: PM 2.5, PM 10, temperatura, umidità, spostamenti in macchina ed ammoniaca) stimiamo R^2 ed MSE per ogni modello. Come si evince dai grafici ottenuti, il modello migliore risulta essere quello a sette regressori, per il quale otteniamo $\mathrm{R2} \approx 0.89$ e $\mathrm{MSE} \approx 118$, entrambi migliori dei valori corrispondenti del modello iniziale a due regressori.



6. Simulazione del modello con il dataset di validazione

Come ultima cosa abbiamo voluto testare la validità del nostro modello simulando, sempre sulla base del dataset che abbiamo avuto a disposizione, dei dati per poter creare così una sorta di ***'dataset di validazione'***. L'obiettivo è quello di mettere a confronto la variabile dipendente Y ottenute dalle rilevazioni della stazione e successivamente i valori di Y ottenute dall'applicazione del modello al dataset di validazione.

Abbiamo dunque calcolato il valore minimo e il valore massimo per ciascuno dei regressori, definendo cosi un range di valori entro il quale generare i dati randomici. I valori ottenuti sono stati vettori colonna di 160 unità, i quali successivamente abbiamo raggruppato in una matrice, ottenendo così il nostro dataset di validazione. Procediamo con la validazione del modello.

Il modello da testare è:

$$\hat{y} = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \beta_6 X_6 + \beta_7 X_7 + \beta_8 X_8 + \beta_9 X_9$$

Dove:

- $\hat{y}$ è la risposta ai valori, ossia rappresenta il risultato previsto dal modello;
- β_0 è l'intercetta, ossia il valore di $\hat{y}$ quando gli X_n sono tutti uguali a 0;
- da β_1 *a* β_9 sono i coefficienti del modello;
- da X_1 a X_9 sono le variabili dipendenti.

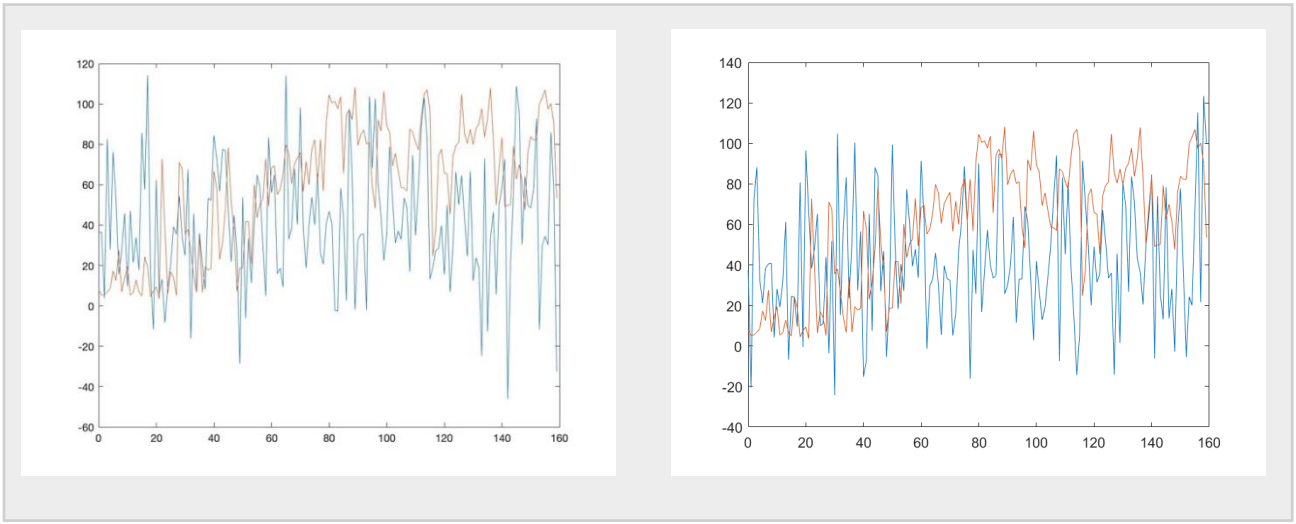
Per quanto riguarda i coefficienti da β_1 *a* β_9 sono stati ricavanti con il comando fitlm usando il dataset originale, riportati qui di seguito:

(Intercept)	107.28
BiossidoDiAzoto	-0.56948
Benzene	-1.3095
PM10	-0.12807
PM2_5	0.40762
Temperatura	1.5416
Umidit_Relativa	-0.7297
SpostamentiInMacchina	-0.001392
Ammoniaca	-0.16143

6.1 Confronto grafico finale

Nei seguenti grafici riportiamo il risultato dell'analisi:

in blu: la variabile ŷ in risposta alle variabili dipendenti del dataset di validazione.
in rosso: la variabile ŷ in risposta alle variabili dipendenti del modello originale;



Essendo una generazione casuale di dati, ne consegue che il grafico in rosso varierà ad ogni ri-esecuzione dello script. Quello che però si può notare è che l'andamento stimato approssima in maniera discreta l'andamento reale ed effettivo. Ne consegue che il nostro modello sia corretto per la descrizione dell'andamento della variabile risposta Ozono.