UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

CASE STUDY

REGRESSIONE LINEARE

Gruppo G27 - NICOSIA

Componenti del gruppo:

Carminati Luca	n°matricola: 1067252
Carminati Matteo	n°matricola: 1066354
Silvestro Giuseppe	n°matricola: 1068822
Torri Lorenzo	n°matricola: 1069047

DESCRIZIONE LAVORO SVOLTO

Introduzione

Abbiamo deciso di analizzare le concentrazioni di PM10 presenti nell'aria per la stazione di Morbegno e di Bergamo in funzione delle variabili atmosferiche a nostra disposizione e delle concentrazioni di Ozono presenti nell'aria.

Dopo aver importato i dati nel workspace di Matlab e aver ispezionato il dataset della tabella associata al nostro gruppo (G27), abbiamo ricavato alcune statistiche descrittive (media, minimo, massimo e standard error) riguardanti le concentrazioni di PM10 per le due stazioni. (Tabella 1)

Analisi preliminare (Tabella 2, Grafici delle Regressioni)

Come analisi preliminare dei dati abbiamo scelto di osservare qualitativamente i grafici che mettevano in relazione le concentrazioni di PM10 con ciascuna variabile atmosferica e successivamente con le concentrazioni di Ozono. In questo modo abbiamo potuto individuare sommariamente quali variabili risultano più influenti sul PM10.

Abbiamo osservato che sia per Morbegno sia per Bergamo l'umidità relativa non influenza in modo significativo le concentrazioni di PM10 nell'aria, di conseguenza riteniamo che questa covariata possa essere esclusa dal modello di correlazione multipla.

Inoltre, si osserva che temperatura e pioggia cumulata hanno un forte impatto sulle concentrazioni di PM10 e quindi sono parametri da tenere in considerazione nella valutazione dei modelli.

Studio modello di regressione

Come prima operazione abbiamo sfruttato l'algoritmo *stepwiselm* messo a disposizione dal software Matlab al fine di individuare il modello più adatto a descrivere le concentrazioni di PM10 sfruttando i dati a disposizione.

Il metodo *stepwiselm* è un algoritmo automatizzato che, partendo da un modello costante, aggiunge oppure rimuove covariate confrontando il nuovo modello con quello precedente e scegliendo quello più adatto sulla base dei seguenti parametri: coefficiente di determinazione multipla, p-value e distorsione.

A partire dal modello proposto da *stepwiselm* ne abbiamo ricavato uno approssimato.

Successivamente abbiamo confrontato tale modello con altri ottenuti con il comando *fitlm,* andando a valutare i valori del coefficiente di determinazione multipla, il p-value e osservando qualitativamente i grafici dei residui e delle interpolate.

Sono stati scelti, seguendo questi criteri, quelli che secondo noi erano i migliori modelli per ciascuna stazione.

Morbegno

Poiché il modello ricavato sfruttando il comando *stepwiselm* presenta prodotti e interazioni tra le variabili, la nostra scelta è ricaduta sul modello approssimato.

Tale modello è stato ottenuto a partire da *stepwiselm* mediante la rimozione dei prodotti tra le covariate, mantenendo però quelle selezionate dall'algoritmo.

In questo modo è stato ricavato il modello m5SO che inserisce come covariate pioggia cumulata, temperatura e concentrazione di ozono.

Abbiamo in seguito testato tale approssimazione provando ad elaborare altri modelli e valutandone i parametri.

Al termine del confronto abbiamo ritenuto che l'approssimazione dell'algoritmo *stepwiselm* (modello m5SO) fosse quella migliore e più accurata.

Bergamo

Anche per la stazione di Bergamo il procedimento iniziale è analogo a quello seguito per la stazione di Morbegno.

Tuttavia, in questo caso l'approssimazione del modello di *stepwiselm* (modello m5BG) non è risultata la scelta migliore poiché presentava una covariata (la temperatura) poco significativa. È stato analizzato dunque un modello che mantenesse come covariate esclusivamente la pioggia cumulata e la concentrazione di ozono (modello m6BG).

I parametri dei due modelli sono molto simili, tuttavia, seguendo il principio metodologico del Rasoio di Occam, la nostra scelta è ricaduta sul modello m6BG. Quest'ultimo mantenendo performances simili al modello m5BG risulta essere più "parsimonioso" in quanto usa solo due covariate.

Confronto modelli

In conclusione del lavoro abbiamo confrontato il modello scelto per la stazione di Morbegno con quello adottato per la stazione di Bergamo valutando i seguenti aspetti:

- covariate;
- coefficiente di determinazione multipla;
- p-value del modello;
- valutazione dei residui;
- grafico delle interpolate;

ANALISI DATI

Tabella 1	Min ($\mu g/m^3$)	Max ($\mu g/m^3$)	Med ($\mu g/m^3$)	STD ($\mu g/m^3$)
Bergamo	6,3333	108	28,637	15,69
Morbegno	6,5714	54	19,782	9,8848

Analisi preliminare

Tabella 2		Valore associato alla Conc. Min PM10	Valore associato alla Conc. Max PM10
Bergamo	**Temperatura (°C)**	16,86	4,71
	Pioggia (mm)	123	34,8
	Umidità (%)	70,321	94,097
	Ozono ($\mu g/m3$)	37,947	14,171
Morbegno	**Temperatura (°C)**	11,079 e 16,359	2,8299
	Pioggia (mm)	37,4 e 35,4	0
	Umidità (%)	68,379 e 61,395	33,813
	Ozono ($\mu g/m3$)	61,801 e 57,405	23,054

Osservando i grafici ipotizziamo che nei periodi più caldi dell'anno le concentrazioni di PM10 sono più basse rispetto a quelle registrate nei periodi freddi, probabilmente a causa delle emissioni dovute ai riscaldamenti domestici.
Le precipitazioni abbattono le concentrazioni di PM10, infatti quando i mm di pioggia cumulata sono considerevoli i livelli di PM10 sono esigui, mentre quando le precipitazioni sono scarse la qualità dell'aria ne risente negativamente.

Studio modello di regressione

Per entrambe le stazioni abbiamo preso in esame diversi modelli basandoci sia su quello proposto da *stepwiselm* sia sulle considerazioni fatte a partire dall'analisi preliminare.

Morbegno

Modelli analizzati: (Dati fitlm)

- m1SO: Temperatura, Pioggia, Umidità e Ozono;
- m2SO: Pioggia e Umidità;
- m3SO: Temperatura e Pioggia;
- m4SO: Umidità e Ozono;
- m5SO: Temperatura, Pioggia e Ozono (approssimazione da *stepwiselm*).

Valutando i parametri di ciascun modello di regressione, (R^2, p-value covariate, p-value modello, grafico residui, grafico interpolate) poiché il modello m5SO è quello con R^2 maggiore, p-value delle covariate significativi, p-value totale più basso e grafici qualitativamente più accurati, concludiamo che è il migliore per spiegare le concentrazioni di PM10.
Nonostante il modello m5SO sia a nostro avviso il migliore, il grafico dei residui (Grafici dei residui) sembra rappresentare una sorta di parabola.
Riteniamo che probabilmente nei residui vi sia una parte deterministica che non è stata catturata completamente dalla nostra soluzione, cosa che il modello di stepwiselm sembra

riuscire a fare. In questo caso infatti i residui sono compresi in una banda orizzontale in modo quasi simmetrico rispetto all'asse x.
A supporto del modello proposto, come si può verificare graficamente (Grafici modelli interpolati), lo scarto tra valori interpolati e reali è contenuto.

Si osservi che il modello individuato tramite stepwiselm risulta essere ancora più significativo di m5SO, ma introduce considerazioni non affrontate in aula (rapporti tra covariate).

Bergamo

Modelli analizzati: (Dati fitlm)

- m1BG: Temperatura, Pioggia, Umidità e Ozono;
- m2BG: Pioggia e Umidità;
- m3BG: Temperatura e Pioggia;
- m4BG: Umidità e Ozono;
- m5BG: Temperatura, Pioggia e Ozono (approssimazione da *stepwiselm*);
- m6BG: Pioggia e Ozono.

Con lo stesso ragionamento adottato per la stazione di Morbegno e avvalendosi del principio metodologico di Occam abbiamo ritenuto il modello m6BG il più adatto.
Il grafico dei residui (Grafici dei Residui) ha valori che si posizionano sopra e sotto l'asse delle x e sono contenuti in una banda orizzontale, quindi non appaiono evidenti difetti nel modello.
Anche per la stazione di Bergamo, come si può verificare graficamente (Grafici modelli interpolati), lo scarto tra valori interpolati e reali è contenuto.

Si osservi che, analogamente alla stazione di Morbegno, il modello individuato tramite stepwiselm risulta essere ancora più significativo di m6BG, ma introduce considerazioni non affrontate in aula (rapporti tra covariate).

Confronto modelli

Per confrontare i modelli m5SO e m6BG abbiamo seguito lo stesso ragionamento sfruttato per scegliere il modello più adatto per ciascuna stazione.

	BERGAMO	***MORBEGNO(SO)***
Covariate	Pioggia e Ozono.	Temperatura, Pioggia Ozono.
Coefficienti	Tutti negativi. Quello dell'intercetta è invece positivo.	Tutti negativi. Quello dell'intercetta è invece positivo.
P-value	Tutti statisticamente significativi.	Tutti statisticamente significativi.
P-value modello totale	Migliore tra i due.	Peggiore di quello di Bergamo ma comunque molto basso.
Performance di adattamento (R^2)	È più alto rispetto a quello della stazione di Morbegno.	È più basso rispetto a quello della stazione di Bergamo.
Valutazione dei residui	Grafico migliore tra i due. Somma dei residui quasi nulla.	Grafico peggiore di quello di Bergamo. Somma dei residui quasi nulla.
Grafico modello interpolato	Grafico simile a quello di Morbegno.	Grafico simile a quello di Bergamo.

Il modello adottato per la stazione di Bergamo sembra essere quello più efficace per descrivere l'andamento delle concentrazioni di PM10 presenti nell'aria.

Stepwiselm SO (Torna indietro)

```
Linear regression model:
    y ~ 1 + x1*x2 + x1*x4

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	45.565	1.719	26.507	6.5645e-66
x1	-2.0398	0.18036	-11.31	4.9441e-23
x2	-0.12501	0.02831	-4.4159	1.6817e-05
x4	-0.35464	0.040557	-8.7441	1.156e-15
x1:x2	0.0058917	0.0022333	2.6381	0.0090251
x1:x4	0.023337	0.0026713	8.7362	1.2152e-15

```
Number of observations: 197, Error degrees of freedom: 191
Root Mean Squared Error: 5.89
R-squared: 0.654,  Adjusted R-Squared: 0.645
F-statistic vs. constant model: 72.2, p-value = 3.75e-42
```

m5SO

```
Linear regression model:
    PM10_tG1 ~ 1 + O3_tG1 + Pioggia_cum_tG1 + Temperatura_tG1

Estimated Coefficients:
```

	Estimate	SE	tStat	pValue
(Intercept)	32.914	1.2347	26.657	1.3035e-66
O3_tG1	-0.080537	0.030224	-2.6646	0.0083596
Pioggia_cum_tG1	-0.076058	0.014862	-5.1175	7.4507e-07
Temperatura_tG1	-0.62557	0.11306	-5.5332	1.0153e-07

```
Number of observations: 197, Error degrees of freedom: 193
Root Mean Squared Error: 7.04
R-squared: 0.5,  Adjusted R-Squared: 0.492
F-statistic vs. constant model: 64.4, p-value = 6.73e-29
```

Stepwiselm BG (Torna indietro)

```
Linear regression model:
    y ~ 1 + x2 + x1*x4
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	57.648	2.5429	22.67	3.5012e-56
x1	-0.95603	0.25396	-3.7645	0.0002217
x2	-0.17583	0.027614	-6.3676	1.3833e-09
x4	-0.44117	0.081713	-5.399	1.9644e-07
x1:x4	0.012998	0.0032355	4.0173	8.4436e-05

```
Number of observations: 197, Error degrees of freedom: 192
Root Mean Squared Error: 10.3
R-squared: 0.577,  Adjusted R-Squared: 0.568
F-statistic vs. constant model: 65.3, p-value = 8.44e-35
```

m5BG

```
Linear regression model:
    PM10_BG ~ 1 + O3_BG + Pioggia_cum_BG + Temperatura_BG
```

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	49.827	1.6989	29.33	4.8425e-73
O3_BG	-0.18014	0.051457	-3.5008	0.00057609
Pioggia_cum_BG	-0.19941	0.028021	-7.1166	2.1165e-11
Temperatura_BG	-0.41706	0.22392	-1.8625	0.064053

```
Number of observations: 197, Error degrees of freedom: 193
Root Mean Squared Error: 10.7
R-squared: 0.541,  Adjusted R-Squared: 0.534
F-statistic vs. constant model: 75.8, p-value = 1.93e-32
```

m6BG

```
Linear regression model:
    PM10_BG ~ 1 + O3_BG + Pioggia_cum_BG
```

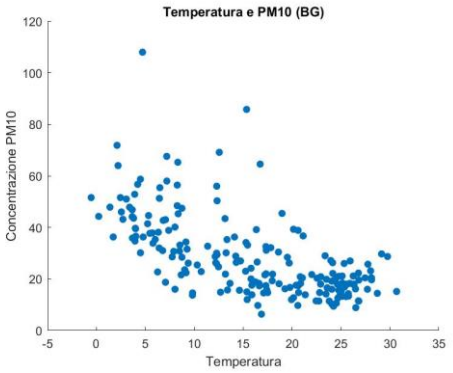
Estimated Coefficients:

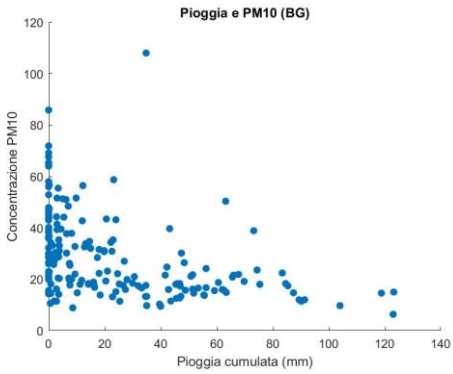
	Estimate	SE	tStat	pValue
(Intercept)	48.5	1.5521	31.248	1.2091e-77
O3_BG	-0.26698	0.021914	-12.183	9.9933e-26
Pioggia_cum_BG	-0.21079	0.027521	-7.659	8.6737e-13

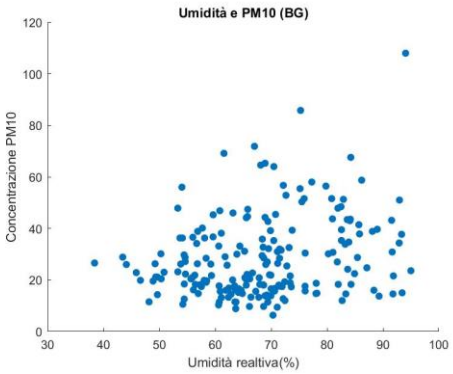
```
Number of observations: 197, Error degrees of freedom: 194
Root Mean Squared Error: 10.8
R-squared: 0.533,  Adjusted R-Squared: 0.528
F-statistic vs. constant model: 111, p-value = 8.97e-33
```

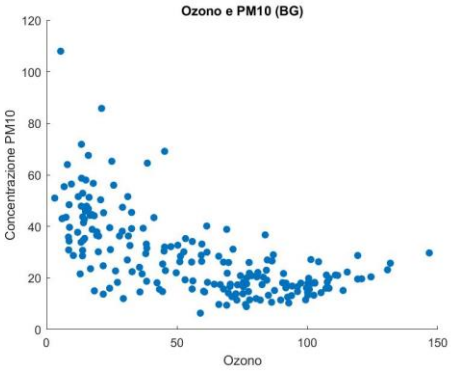
GRAFICI REGRESSIONI (Torna indietro)

BERGAMO

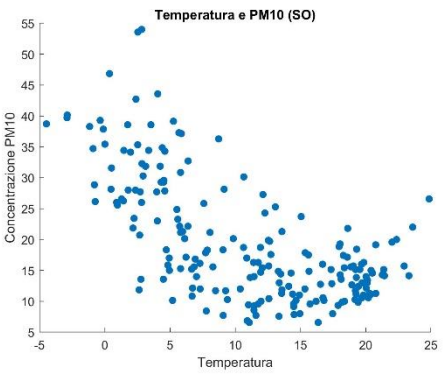


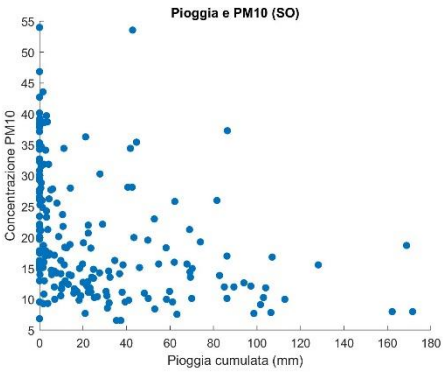


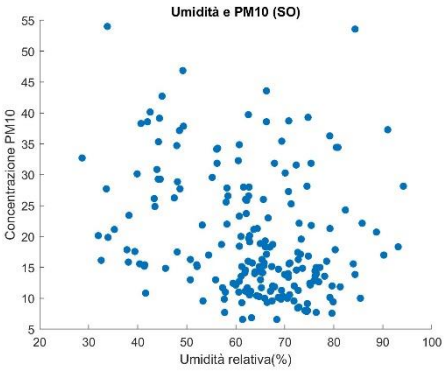


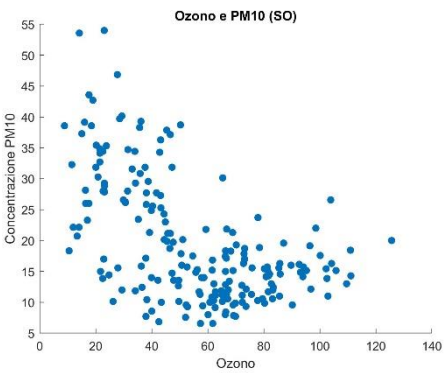


MORBEGNO (SO)



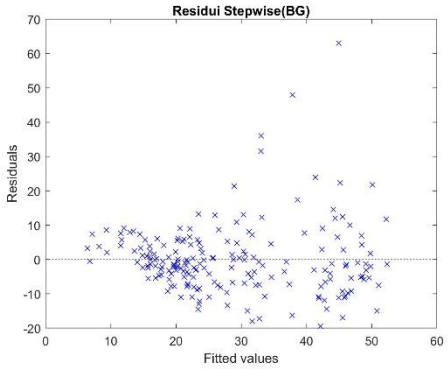


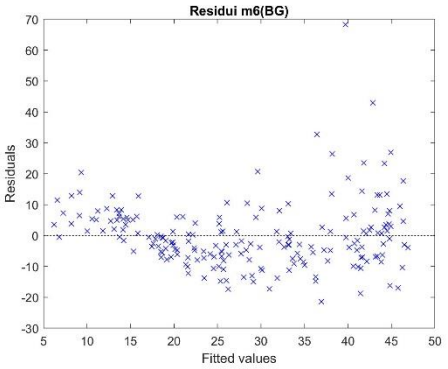




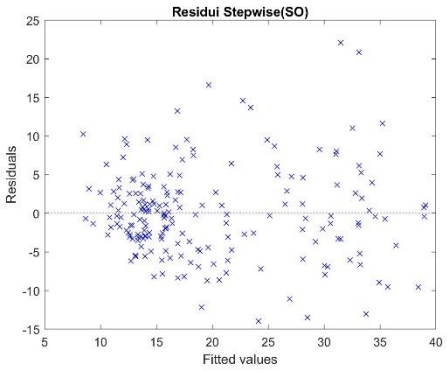
GRAFICI RESIDUI

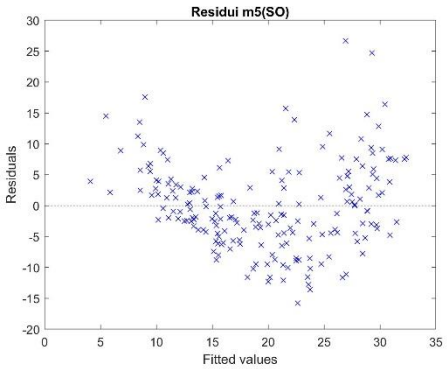
BERGAMO (Torna indietro)





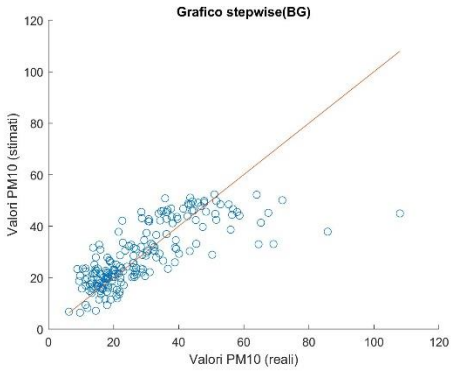
MORBEGNO (SO) (Torna indietro)

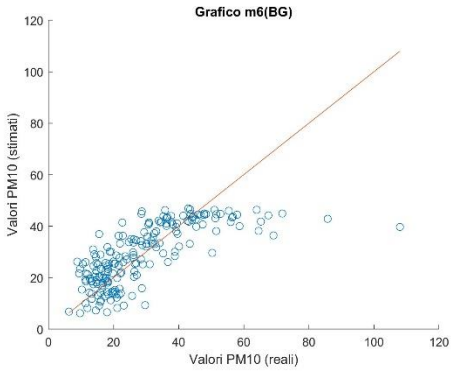




GRAFICI MODELLI INTERPOLATI

BERGAMO (Torna indietro)





MORBEGNO (SO) (Torna indietro)