UNIVERSITÀ
DEGLI STUDI
DI BERGAMO

PROGETTO SMS2

Componenti del gruppo:

Carminati Luca n°matricola: 1067252

Carminati Matteo n°matricola: 1066354

Silvestro Giuseppe n°matricola: 1068822

Torri Lorenzo n°matricola: 1069047

DESCRIZIONE DATASET

Il dataset scelto dal nostro gruppo contiene le misurazioni effettuate a ogni ora per tutto il corso dell'anno 2011 attraverso la stazione meteo mobile INAIL collocata a Villa Pamphili (RM).

I dati rilevati riguardano: VV02mAs (velocità vento media a 2mt.), W02mAs (velocità vento verticale a 2mt.), RADGLAs (radiazione solare globale media), PIOGG (mm di pioggia cumulata), T02mAv (temperatura aria media a 2mt.), UMRELAs (umidità relativa media), RADNTAs (radiazione solare netta media), TSOIAs (temperatura al suolo media), PRESSAs (pressione atm media).

DESCRIZIONE DEI QUESITI

Ci siamo posti l'obiettivo di trovare un modello che potesse spiegare la temperatura del suolo in funzione di uno o più regressori tra quelli a disposizione.

L'intenzione era quella di osservare la matrice di correlazione per vedere quali parametri fossero più correlati con la nostra variabile di interesse; successivamente di sviluppare un modello di regressione lineare valutando tutte le sue limitazioni, per poi passare a metodi più sofisticati come B-Spline.

DESCRIZIONE DEL LAVORO SVOLTO

Innanzitutto precisiamo che nella cartella di lavoro di matlab sono presenti tre tabelle: una è il dataset completo descritto sopra (tabella Meteo-Completo), un'altra è lo stesso dataset modificato e privato di alcune misurazioni che abbiamo considerato essere degli errori di rilevazione (tabella Meteo2011) e l'ultima è il dataset modificato di prima ma riordinato rispetto ai valori di una variabile utilizzata per la stima B-Spline (tabella Meteo2011-Ordinato).

Inoltre sono presenti altre funzioni utilizzate per vari scopi (regf e clustf).

Come primo passo abbiamo deciso di considerare una sola misurazione, avvenuta alle ore 14:00, per ogni giorno dell'anno.

Questo perché, trattandosi di dati meteorologici, ogni misurazione effettuata a un tempo t generico dipende molto da ciò che è accaduto al tempo t-1, di conseguenza i dati sarebbero troppo correlati tra loro.

Inoltre il vettore dei valori che abbiamo considerato non contiene 365 elementi, ma solo 355; questa discrepanza di 10 valori è dovuta al fatto che, a volte, le misurazioni non state riportate, altre volte invece come valori della misurazione compare un -999 ipotizzabile come errore di misura. Abbiamo quindi pensato di eliminare dal dataset queste misurazioni (tabella Meteo2011) e di utilizzare il metodo di stima OLS. Successivamente abbiamo però anche deciso di studiare il dataset completo (tabella Meteo-Completo) utilizzando una stima WLS che andasse a ridurre l'impatto di questi valori anomali sul modello. Si ipotizza che le due soluzioni portino a risultati molto simili.

In seguito sfruttando il test di Jarque-Bera abbiamo verificato se la variabile dipendente provenisse da una distribuzione normale. Avendo ottenuto un risultato del *JBtest* pari a 31, e quindi lontano da 0, concludiamo che i nostri dati non possano provenire da una distribuzione normale.

Osservando la matrice di correlazione si individuano i regressori che potrebbero meglio spiegare la variabile dipendente. La nostra scelta è ricaduta su quei regressori con una buona correlazione con la temperatura del suolo, ma che, allo stesso tempo, non fossero troppo correlati tra di loro.

	VV02mAs	W02mAs	RADGLAs	PIOGG	T02mAv	UMRELAs	RADNTAs	TSOIAs	PRESSAs
VV02mAs	1	0.30254	0.50992	-0.048878	0.42851	-0.28749	0.53304	0.48527	-0.29621
W02mAs	0.30254	1	0.49861	-0.083431	0.75302	-0.3154	0.5713	0.68235	-0.086873
RADGLAs	0.50992	0.49861	1	-0.24053	0.74057	-0.69184	0.97334	0.77605	-0.097573
PIOGG	-0.048878	-0.083431	-0.24053	1	-0.13608	0.31818	-0.21226	-0.11049	-0.08181
T02mAv	0.42851	0.75302	0.74057	-0.13608	1	-0.54895	0.7978	0.94351	-0.18145
UMRELAs	-0.28749	-0.3154	-0.69184	0.31818	-0.54895	1	-0.61169	-0.49448	-0.11404
RADNTAs	0.53304	0.5713	0.97334	-0.21226	0.7978	-0.61169	1	0.84481	-0.19413
TSOIAs	0.48527	0.68235	0.77605	-0.11049	0.94351	-0.49448	0.84481	1	-0.2994
PRESSAs	-0.29621	-0.086873	-0.097573	-0.08181	-0.18145	-0.11404	-0.19413	-0.2994	1

Abbiamo quindi scelto come possibili variabili indipendenti del modello la RADNTAs, la PRESSAs e l'UMRELAs. Sfruttando il comando matlab *fitlm* abbiamo potuto constatare che i modelli migliori di regressione lineare per spiegare la temperatura fossero: uno spiegato da radianza e pressione e l'altro con la sola radianza. Dato

che i valori dei coefficienti di determinazione dei modelli erano tra loro molto simili, ci siamo basati sul principio della parsimonia o di Occam e abbiamo quindi scelto il modello che teneva in considerazione solo la radianza.

Linear regression model:
TSOIAs ~ 1 + UMRELAs + RADNTAs

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	6.9192	1.5047	4.5984	5.9453e-06
UMRELAs	0.020742	0.020976	0.98884	0.32342
RADNTAs	0.044475	0.0018476	24.072	2.2387e-76

Number of observations: 355, Error degrees of freedom: 352
Root Mean Squared Error: 4.98
R-squared: 0.714, Adjusted R-Squared: 0.713
F-statistic vs. constant model: 440, p-value = 1.54e-96

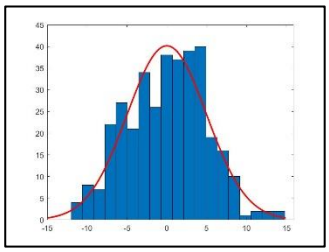
Linear regression model:
TSOIAs ~ 1 + RADNTAs

Estimated Coefficients:

	Estimate	SE	tStat	pValue
(Intercept)	8.3148	0.5216	15.941	1.776e-43
RADNTAs	0.043358	0.0014616	29.665	6.7324e-98

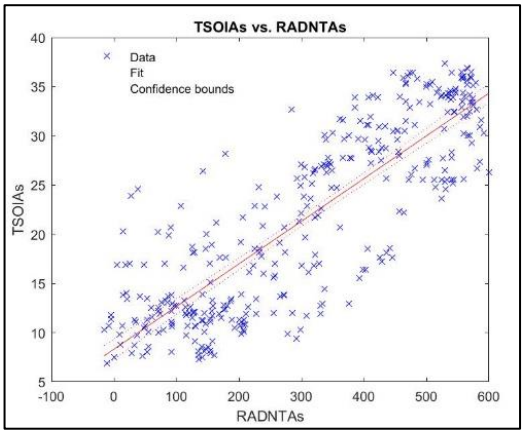
Number of observations: 355, Error degrees of freedom: 353
Root Mean Squared Error: 4.98
R-squared: 0.714, Adjusted R-Squared: 0.713
F-statistic vs. constant model: 880, p-value = 6.73e-98

In seguito si verificano le proprietà dei residui del modello scelto: si nota che la loro media è zero, e che sono incorrelati con i regressori, quindi nei residui non ci sono informazioni ulteriori che non sono state catturate dal modello (come ci aspettavamo dato che abbiamo fatto una stima a minimi quadrati).



Sfruttando ancora il test di Jarque-Bera, abbiamo osservato che i residui hanno distribuzione normale.
Poiché gli errori di un modello sono una stima della parte stocastica, possiamo affermare che per la nostra stima anche la parte stocastica dovrebbe avere le stesse proprietà dei residui e quindi avere distribuzione normale.
Facendo ciò viene verificata la condizione per affermare che il nostro stimatore beta ha anch'esso una distribuzione normale e di conseguenza si può fare inferenza statistica su quest'ultimo.

I dati ottenuti dal nostro modello con *fitlm* non sono comunque ottimali. Il coefficiente di determinazione è di solo 0.7 quindi c'è una buona relazione tra variabile dipendente ed indipendente, ma questa non è ottimale.



Facendo un test t di Student sui coefficienti del modello si ottiene che sia per l'intercetta sia per il regressore il pvalue del coefficiente relativo è bassissimo e approssimabile a 0. Dunque si conclude che entrambi sono statisticamente significativi.

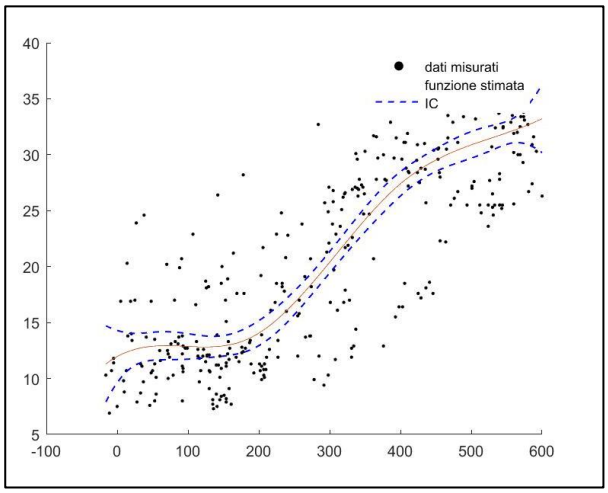
Terminato lo studio con la stima OLS, abbiamo deciso di utilizzare anche la stima WLS per la situazione con i valori -999.

	stima OLS	stima WLS
Intercetta	8.3148	5.6502
Coeff regressore	0.043358	0.049999

Utilizzando un algoritmo iterativo per dare un peso differente alle varie misurazioni, abbiamo ottenuto un modello molto simile a quello ottenuto con la stima OLS, come previsto dalle nostre supposizioni iniziali.

In generale il modello di regressione lineare è il metodo in assoluto più semplice per modellizzare dei dati, esso però presenta anche delle notevoli limitazioni in termini di qualità della stima.
Oltre alle limitazioni strettamente legate al modello sui nostri dati (coefficiente di determinazione non molto alto, dati con una varianza abbastanza elevata...), in generale i dati meteorologici rimangono comunque abbastanza correlati tra loro, quindi la nostra stima avrà sicuramente una certa distorsione.
Per valutare la bontà del modello scelto e la sua capacità predittiva abbiamo ritenuto necessario valutare i relativi EQM di training e soprattutto l'EQM di test.
Per trovare l'EQM di training abbiamo sfruttato il comando *MSE* dal modello ottenuto con *fitlm*, mentre per calcolare l'errore quadratico medio relativo ai dati di test abbiamo usato il metodo della cross-validazione,

sfruttando il comando *crossval* di Matlab si può infatti effettuare una cross-validazione in 10 fold sui dati disponibili.
Come ci aspettavamo, dato che il modello lineare proposto è piuttosto semplicistico, non riesce a catturare tutte le regolarità dei dati e l'EQM relativo ai dati di test è piuttosto elevato, abbiamo quindi deciso di cercare un modello più sofisticato sfruttando B-Spline.
Provando a eseguire i vari comandi matlab per la creazione del modello B-Spline, ci siamo accorti che il modello creato era troppo caotico e irregolare.
Successivamente abbiamo intuito che tale problema era dovuto a un errore nell'ordinamento dei nostri dati che, essendo stati campionati nel tempo, sono disposti in base a quest'ultimo.
Abbiamo dunque riorganizzato le misurazioni mettendole in ordine crescente rispetto alla variabile indipendente (RADNTAs), in questo modo lo Spline funziona correttamente in quanto necessita di valori ordinati rispetto ai range degli intervalli associati ad ogni singolo nodo.
Per poter scegliere il numero di nodi migliore, abbiamo deciso di basarci sul concetto della U-shape dell'errore di test.
In generale tale forma è determinata dal fatto che se aumentando il numero di variabili indipendenti l'EQM di training tende via via a diminuire, quello di test, invece, diminuisce fino al raggiungimento di un certo valore di minimo per poi riaumentare. Nel caso di uno Spline, i valori stimati non dipendono tanto dai regressori quanto dal numero basi, quindi l'andamento dell'errore complessivo legato al modello dipenderà perlopiù dal numero di nodi considerati.
Basandoci su questa idea abbiamo deciso di elaborare un semplice algoritmo iterativo che sfruttando la cross validazione generalizzata potesse andare ad individuare il numero di nodi ideale.
Dunque, partendo da un numero di nodi pari a uno e andandoli ad incrementare ad ogni iterazione, si calcola ogni volta il valore dell'EQM di test e lo si confronta con quello dell'iterazione precedente.



Se l'errore diminuisce allora il ciclo continua mantenendo l'ultimo valore ottenuto come parametro ottimale e di confronto, se invece si ottiene un valore maggiore significa che il valore minimo di EQM di test è stato oltrepassato e quindi il ciclo termina.
Mediante l'algoritmo siamo giunti alla conclusione che il modello sarebbe stato ottimizzato utilizzando 4 nodi.
In seguito, con le stesse considerazioni fatte sull'inferenza statistica riguardo la stima OLS, siamo passati a osservare il comportamento dei residui del modello B-Spline e abbiamo osservato la normalità di questi ultimi. Abbiamo quindi calcolato gli IC del modello utilizzando i dati ottenuti per la varianza dei residui e costruito il grafico del modello.
Confrontando i valori degli EQM di training e di test del modello di regressione lineare con quelli ottenuti dal modello elaborato con B-Spline si può concludere che quest'ultimo ha un migliore adattamento ai dati di training (EQM training minore) e anche una migliore capacità predittiva (EQM test minore).
In generale però anche il modello proposto da B-Spline, dato che i valori degli EQM sono piuttosto elevati, non è un modello ottimale.

	Regressione lineare	B-spline
EQM training	24.646	21.679
EQM test	24.796	22.431