UNIVERSITÀ DEGLI STUDI DI BERGAMO

Incidenza casi di ipertensione in Italia

Analisi di serie storiche multiple

De Duro Federico (1073477), *Medolago Emanuele* (1058907) e *Zanotti Paolo* (1074166)

Statistica e Modelli Stocastici - Modulo II

Anno Accademico 2021 - 2022

Modelli previsionali sull'andamento nel tempo dei casi di ipertensione sul territorio Italiano in funzione di alcuni tra i principali fattori di rischio, suddivisi per zona geografica, stimati a partire dai dati riferiti al periodo storico 1990 - 2014.

Indice

1. Introduzione

Il nostro gruppo ha analizzato l'incidenza dei casi di ipertensione sul territorio italiano dal 1990 al 2014. In particolare sono stati presi in considerazione alcuni dei principali fattori di rischio di ipertensione suddivisi per zone d'Italia. Abbiamo cercato quindi di individuare la relazione esistente tra i fattori di rischio e la percentuale di casi applicando ciò che è stato visto durante il corso.

2. Descrizione del Dataset

Il dataset utilizzato è stato composto partendo da serie storiche multiple differenti pubblicate dall'ISTAT; include 25 osservazioni fatte a partire dal 1990 fino al 2014, divise per zone geografiche: *Nord ovest* (Liguria, Lombardia, Piemonte e Valle d'Aosta), *Nord est* (Emilia-Romagna, Friuli-Venezia Giulia, Trentino-Alto Adige e Veneto), *Centro* (Lazio, Marche, Toscana ed Umbria), *Sud* (Abruzzo, Basilicata, Calabria, Campania, Molise e Puglia) ed *Isole* (Sicilia e Sardegna). Il dataset è costituito dai dati percentuali annuali riferiti a:

- Casi di ipertensione accertati (Variabile dipendente)
- Casi di diabete accertati
- Casi di malattie allergiche, che comprendono cause da allergeni, asma bronchiale, malattie atopiche…
- Casi di eccesso di peso nella popolazione
- Casi di sedentarietà in Italia, persone di 3 anni o più che non praticano sport
- Casi di malattie respiratorie, che comprendono insufficienza respiratoria, Bronchiectasie…

Sono stati scelti questi dati poiché, dopo un'analisi preliminare, abbiamo scoperto che sono tra i principali fattori di rischio di ipertensione primaria.

*3. Quesiti da verificare **(edited)***

1. *Quale metodo utilizzare per la stima dei dati mancanti?*
2. *Le variabili del dataset sono correlate tra loro?*
3. *Quali sono i regressori significativi utili per stimare l'impatto dell'ipertensione in Italia?*
4. *E' possibile creare modelli per stimare e prevedere l'incidenza dei casi di ipertensione?*
5. *I modelli stimati risultano essere buoni modelli?*
6. *Qual è il modello più adatto in funzione anche della sua capacità previsiva?*

*4. Strategia utilizzata **(edited)***

Si sono seguite le seguenti fasi di lavoro:

- *Stima dei dati mancanti:* verifica normalità della congiunta dei regressori e, in caso di esito positivo, sfruttare questa informazione per la stima oppure utilizzo della media mobile.
- *Studio della correlazione delle variabili*
- *Costruzione dei modelli di regressione lineare multipla*:
  - Crossvalidazione per la scelta dei regressori
  - Costruzione del modello di regressione lineare multipla (visto l'andamento lineare dei dati di ipertensione) con l'utilizzo del metodo dei minimi quadrati ordinari (*Ordinary Least Squares, OLS*), come prima ipotesi di metodo di stima ottimale, e successiva verifica della sua ottimalità rispetto al metodo generalizzato (verifica di omoschedasticità e di non autocorrelazione degli errori nell'analisi dei residui);
  - Verifica della non-multicollinearità all'interno del modello;
  - Analisi residui: studio della distribuzione, incorrelazione con i regressori, varianza omogenea, autocorrelazione, linearità e outliers.
- *Costruzione dei modelli di regressione dinamica*: stima dei parametri, filtro, smoothing e previsione a un passo, accompagnate da una rapida analisi dei residui (media nulla, istogramma, autocorrelazione e autocorrelazione parziale).
- *Costruzione dei modelli di regressione lineare con errori ARIMA*:
  - Scelta dei regressori in base alla loro significatività
  - Scelta di p e q in funzione di BIC (*Bayesian info criterion*), MSE e significatività dei coefficienti

  - costruzione del modello regArima prescelto
  - Calcolo delle radici del polinomio caratteristico, test di Dickey-Fuller (Verifica stazionarietà)
  - Bootstrap per stimare gli intervalli di confidenza dei coefficienti del modello
  - Analisi dei residui (media nulla, histfit, qqplot, outliers, autocorrelazione, autocorrelazione parziale, correlazione residui-regressori, omoschedasticità, test di Lujing-Box, JB Test via Monte Carlo)
- Valutazione della capacità previsiva dei modelli:
  - Suddivisione del dataset in dataset Training e dataset Test
  - Calcolo del MSE (*Mean Square Error*)
  - Calcolo degli IC sulla previsione
  - Plot previsione, osservazione, IC (Intervalli di confidenza)
- Confronto delle performance dei diversi modelli

Questo procedimento è stato fatto per i diversi territori, nel particolare viene mostrato lo sviluppo sui modelli costruiti per la zona geografica "Nord Est". Per tutti i test e gli IC è stato scelto un livello di significatività $\alpha = 0.05$.

5. Svolgimento

5.1 - Stima dei dati mancanti

Per quanto riguarda i dati mancanti si è voluto testare la possibilità di una distribuzione congiunta dei regressori di tipo normale, affinché si possa effettuare successivamente una stima utilizzando queste informazioni. Ciò è stato fatto partendo dall'ipotesi di normalità delle marginali, in quanto ogni singolo regressore si può vedere come una realizzazione di una binomiale con *n* prove, con una certa probabilità π_i di osservare l'evento. (es: P. di essere sovrappeso: π, P di non esserlo: $1-\pi$). Dato il grande numero di osservazioni Istat, per il TLC le variabili standardizzate tenderanno a distribuirsi normalmente. Essendo che la normalità delle marginali (i regressori) non implica la normalità della congiunta, si è testata tale ipotesi. Per fare ciò abbiamo utilizzato l'indice di *Mardia*, in modo da stimare Skewness e Kurtosis della congiunta e utilizzarli per verificarne la normalità. Dato che Skewness e Kurtosis risultano essere molto elevati si rifiuta l'ipotesi di normalità. Abbiamo quindi modificato strategia e optato per la *Media Mobile* per la stima dei dati mancanti, una tecnica molto utilizzata nelle serie storiche.

5.2 - Correlazione delle variabili

Per verificare la correlazione tra i regressori abbiamo calcolato le matrici di correlazione per ogni zona geografica e verificato gli andamenti tramite il comando matlab *plotmatrix.*
Dai grafici e dalle matrici di correlazione si possono notare delle relazioni tra i regressori; tra quelli inclusi nei modelli non risulta però esserci multicollinearità, infatti per tutte le zone geografiche il det(X'X) >> 0. Gli indici di correlazione più rilevanti tra regressori e variabile dipendente per il Nord Est risultano essere: *Ma. allergiche* ($\rho = 0{,}95$), *Peso* ($\rho = 0{,}94$), *Diabete* ($\rho = 0{,}93$) e *Sedentari* ($\rho = 0{,}903$).

5.3 - Modello di regressione lineare

5.3.1 - Regressori utili per studiare l'impatto dell'ipertensione

Abbiamo eseguito la *Cross Validazione* (k-fold, k = 5) per la scelta dei regressori utili. Gli MSE (*Mean Square Error*) riportati sono dati dalla media degli MSE di Cross Validazione calcolati per ciascun regressore aggiuntivo. Per quanto riguarda il Nord Est, si può notare come al terzo regressore l'MSE medio passi da 0,51 a 0,35 e l'R^2 salga da 0,944 a 0,964, mentre con l'aggiunta del quarto l'MSE scenda solamente a 0,32 e l'R^2 aumenti solo a 0,973. Per evitare *Overfitting* abbiamo deciso di non includere il quarto regressore.

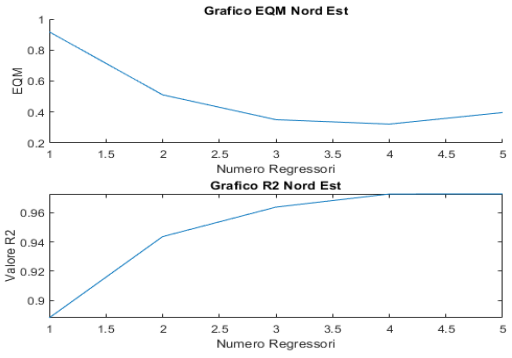


figura 5.3.1.1

5.3.2 - Costruzione dei modelli

I modelli sono stati stimati con il *metodo dei minimi quadrati ordinari (OLS)*, tramite l'utilizzo dei comandi *fitlm*, in modo da ottenere i p-value risultanti dai test sui coefficienti e l'R^2 aggiustato, e *regress*, utilizzato in fase di Cross Validazione e da cui abbiamo ricavato gli intervalli di confidenza dei coefficienti di regressione. Da qui poi sono state valutate l'ottimalità (tramite verifica di omoschedasticità e non autocorrelazione degli errori) e la non distorsione (media residua nulla) della stima *OLS* attraverso l'analisi dei residui.

5.3.3 - Analisi dei residui

Come ultimo passaggio, abbiamo analizzato i residui dei modelli costruiti in precedenza. Per il Nord Est si hanno i seguenti risultati. La media dei residui di regressione con metodo *OLS* è per definizione uguale a zero, questo si può verificare dalla Linea Blu (linea media) nella [*Figura 5.3.3.1] Grafico Residui vs Osservazioni*, la media calcolata è infatti praticamente uguale 0 ($2{,}34e^{-15}$).

La normalità è stata studiata tramite il *test di Jarque Bera*. I valori critici di JB_n per numerosità piccole sono stati calcolati via *simulazione Monte Carlo (MC)*. Dalla simulazione risulta che $JB_{CRIT} > JB_0$, con p-value di 0.82 e potenza del test pari a 0.72, il che porta ad accettare l'ipotesi nulla H_0*: ku = 3, sk = 0*. Si può dunque sostenere con una certa confidenza che i residui sono normalmente distribuiti.

L'andamento dei percentili viene osservato tramite il comando matlab *qqplot*, il quale mantiene sull'asse delle ascisse i quantili teorici di una distribuzione normale, mentre su quello delle ordinate i quantili dei residui standardizzati. Dal grafico si nota che i dati seguono la bisettrice confermando la distribuzione gaussiana.

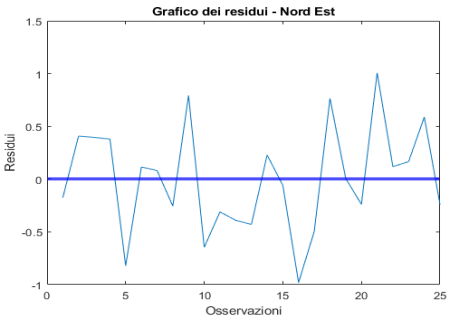


figura 5.3.3.1

L'ipotesi di varianza omogenea è stata esaminata, in prima battuta, tramite un grafico a dispersione dei valori dei residui contro i valori stimati. Il risultato ottenuto è una banda orizzontale di punti dispersi in modo casuale, il che porta all'ipotesi di omogeneità della varianza degli errori. L' omoschedasticità è stata verificata anche tramite il test *'Breush-Pagan*', con il quale si è accettata l'ipotesi nulla. Tale test, infatti, assume che gli errori siano indipendenti, normalmente distribuiti e con varianza funzione del tempo, secondo l'equazione: $ln(\sigma_t^2) = a + bt$, di conseguenza, si ha omoschedasticità se si realizza l'ipotesi nulla: H_0*: b = 0.*

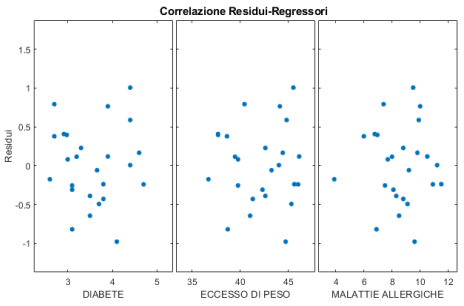


figura 5.3.3.2

Si determina in seguito che i regressori non sono correlati con i residui. Effettuando una verifica analitica dei coefficienti di correlazione tra residui e regressori risulta esserci una correlazione quasi nulla *[Figura 5.3.3.2]*. Inoltre, è stato impiegato un particolare test statistico utilizzato per rilevare la presenza di autocorrelazione dei residui in un'analisi di regressione. Il test in questione è noto come *'Statistica di Durbin-Watson'*, la quale restituisce un valore sempre compreso tra 0 e 4. Un valore di 2 indica che non appare alcuna autocorrelazione. Nel nostro caso, la statistica test ha assunto valore *2.15*, portando ad un p-value di 0.89, di conseguenza si accetta l'ipotesi nulla: i residui non sono autocorrelati.
Infine si è verificata la linearità dei residui, i quali infatti si distribuiscono in modo casuale attorno allo 0 nel grafico *'Residuals vs Fitted data'*, e l'assenza di outliers, in seguito all'analisi dello scatter plot *'residui studentizzati vs Fitted data'*, dove si nota che i residui studentizzati rimangono all'interno del range [-3, 3].

*5.4 - Modelli di regressione dinamica **(new)***

In questa fase si è cercato di stimare lo stato latente, ossia i coefficienti della regressione dinamica, tramite l'utilizzo di Kalman Filter e Kalman Smoother, previa stima delle matrici necessarie (matrici del modello state space). Successivamente, è stato ricostruito l'andamento dell'ipertensione, utilizzando i coefficienti ricavati dal Kalman Filter e a seguire dal Kalman Smoother, in modo da poterli poi confrontare con i valori osservati. Si è notato che in generale, in tutte le zone geografiche, la ricostruzione è quasi sovrapponibile all'osservazione, il che indica un'eccessiva variabilità dei modelli. Inoltre si osserva che le stime dei valori delle matrici e dello stato latente al tempo finale risultano non significative. Si può quindi affermare che i modelli così stimati non risultano validi. È stata comunque fatta una rapida analisi dei residui per studiarne le caratteristiche, con i seguenti risultati: media nulla, autocorrelazione e autocorrelazione parziale che tendono a zero e andamento casuale dei residui intorno allo zero.

*5.5 - Modelli di regressione con errori ARIMA **(new)***

*5.5.1 - Costruzione dei modelli **(new)***

Per la costruzione dei modelli di regressione con errori *ARIMA (Autoregressive Integrated Moving Average)* si è partiti da un'analisi preliminare per la determinazione dei regressori significativi (p-value<0.05 dai t test sui coefficienti) e dei valori di p (ordine della componente AR) e di q (ordine della componente MA) posto d=0 (ordine di differenziazione) (tranne nelle Isole dove si è fissato d=1 per ottenere un modello significativo). I valori di p e q sono stati scelti in modo da minimizzare BIC, MSE e complessità del modello (confronto tramite ciclo for e plot dei risultati [Figura 5.5.1.1), tenendo contemporaneamente in considerazione la significatività dei coefficienti, l'invertibilità della componente MA e la stazionarietà della componente AR. In particolare, la stazionarietà è stata in seguito verificata sul modello scelto tramite il calcolo delle soluzioni dell'equazione caratteristica della componente AR, in modo da accertare che fossero in modulo minori di 1, e il test di Dickey-Fuller, eseguito sulle osservazioni dell'ipertensione a meno della componente di regressione lineare Xβ*hat* stimata. Questo test assume che le radici della componente AR siano unitarie contro l'ipotesi alternativa che siano in modulo inferiori a 1. Il calcolo del MSE è stato fatto suddividendo il dataset in dataset training e dataset test (MSE sulla previsione). Quest'analisi è stata fatta per ogni zona geografica. Per quanto riguarda il Nord Est è stato scelto come modello un regARIMA(2,0,0) con diabete (x_1) ed eccesso di peso (x_2) come regressori, a cui è associato un BIC di 47.85 e un MSE di 0.779. Le soluzioni dell'equazione caratteristica della componente AR sono in modulo entrambe 0.871, quindi inferiori ad 1. Infine, il test di di Dickey-Fuller ha portato al rifiuto dell'ipotesi nulla (radici unitarie), confermando così la stazionarietà del processo.

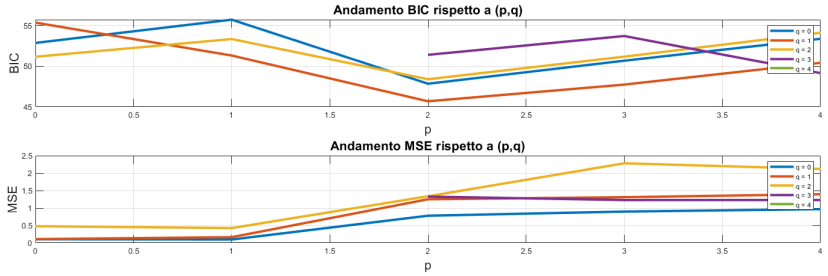


Figura 5.5.1.1: plot di BIC e MSE rispetto ai valori di p e q. Dove la funzione non esiste, il processo con i parametri indicati è non stazionario e/o non invertibile.

*5.5.2 - Stima intervalli di confidenza (IC) sui parametri tramite Bootstrap **(new)***

Gli intervalli di confidenza sui coefficienti dei modelli regARIMA sono stati calcolati tramite bootstrap parametrico, supponendo quindi valida l'ipotesi di normalità delle innovazioni (verificata nell'analisi dei residui).

Eccezione fatta per le Isole, dove i residui hanno portato al rifiuto dell'ipotesi di normalità con il JB test via simulaz. MC e, di conseguenza, la stima degli IC è stata effettuata tramite bootstrap semi-parametrico lineare.

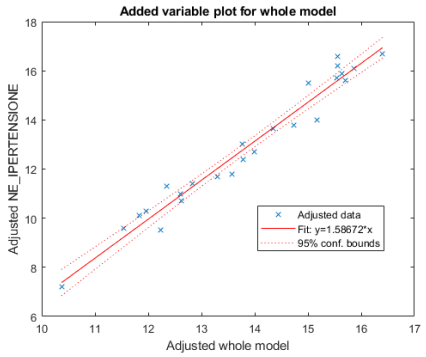
5.5.3 - Analisi dei residui ***(new)***

Nell'analisi dei residui si è andati a studiare: media, distribuzione, outliers, autocorrelazione, autocorrelazione parziale, correlazione residui-regressori, omoschedasticità, indipendenza e normalità.
I risultati ottenuti sui residui del modello regARIMA(2,0,0) nel Nord Est sono i seguenti: media nulla (media calcolata: 0.003, t test → accetta H_0: media = 0, p-value = 0.98); funzioni di autocorrelazione e autocorrelazione parziale campionarie intorno allo zero ed entro gli IC; nessun outlier (residui studentizzati compresi nel range [-3, 3] nello scatter plot *'residui studentizzati vs Fitted data'*); coefficienti di correlazione campionari residui-regressori vicini a zero ($\rho(e,x_1) = -0.07$ e $\rho(e,x_2) = -0.05$); omoschedasticità (test di *'Breush-Pagan'* → accetta H_0: varianza costante nel tempo, p-value = 0.134); residui che corrispondono a delle innovazioni iid (test di *Lujing Box* → accetta H_0: innovazioni iid, p-value = 0.458) e, infine, normalità (JB test via simulazione MC → accetta H_0: *ku = 3, sk = 0*, p-value = 0.279, inoltre si nota che i residui seguono la bisettrice nel qqplot).

6. Risultati ***(edited)***

Risultati regressione lineare

La stima dei modelli di regressione lineare multipla con il metodo *'OLS'* risulta essere *ottimale* ($var(\beta hat)=\sigma^2(X'X)^{-1}$) e *non distorta* ($E(\beta hat)=\beta$). Questa è una conseguenza dell'analisi dei residui, difatti è stato verificato con un buon grado di confidenza che essi sono *indipendenti* e *identicamente distribuiti*, con distribuzione $N(0, \sigma^2)$ (σ^2 stimato=0.29). Risultati dati dal modello di regressione lineare via *'OLS'* Nord Est:




```
Estimated Coefficients:
                       Estimate       SE       tStat      pValue
    (Intercept)          -9.082     2.7074   -3.3545    0.0030021
    NE_DIABETE           1.4616     0.36161   4.0418    0.0005879
    NE_MA_ALLERGICHE      0.547     0.1598    3.4231    0.0025556
    NE_ECCESSO_PESO      0.28678    0.094988  3.0191    0.0065299

Number of observations: 25, Error degrees of freedom: 21
Root Mean Squared Error: 0.539
R-squared: 0.964,  Adjusted R-Squared: 0.959
F-statistic vs. constant model: 186, p-value = 2.72e-15
```

Intervalli di confidenza sui Beta del modello (matrice dati *bint3NE* in Matlab):
Intercetta : [-14.71; -3.45], *Diabete :* [0.71; 2.21], *Malattie allergiche*: [0.21; 0.88], *Eccesso di peso:* [0.09;0.48]
I risultati dell'analisi dei residui (residui iid $N(0, \sigma^2)$), ci permettono di conoscere la distribuzione degli stimatori, $\beta hat \sim N_4(\beta, \sigma^2(X'X)^{-1})$ (→ IC sui Beta affidabili), consentendoci di effettuare i test sul modello (*t test e test F*), da cui risulta che i coefficienti, sia presi singolarmente che nel loro insieme, sono significativi (p-value < 0.05).

Risultati regressione dinamica

Var. of innovation	Coeff	p-value
V. of innovation (1)	0.99943	0.99847
V. of innovation (2)	0.99344	0.98470
V. of innovation (3)	0.95405	0.91389
V. of innovation (4)	0	1

Final state:	Coeff	p-value
x(1)	4.14065	0.96720
x(2)	-0.58331	0.94045
x(3)	1.10750	0.72195
x(4)	0.02079	0.99342

Logarithmic likelihood: -88.7061
Akaike info criterion: 185.412
Bayesian info criterion: 189.395

Con varianza dell'errore di misura pari a 0.95.

I risultati riportati derivano dal comando matlab *estimate*, il quale utilizza Kalman filter e massima verosimiglianza per la stima dei parametri. Come si può notare, sia i valori stimati per la matrice delle innovazioni che la ricostruzione dello stato latente (Coefficienti di regressione) risultano non significativi, di conseguenza il modello stesso risulta essere privo di significatività.

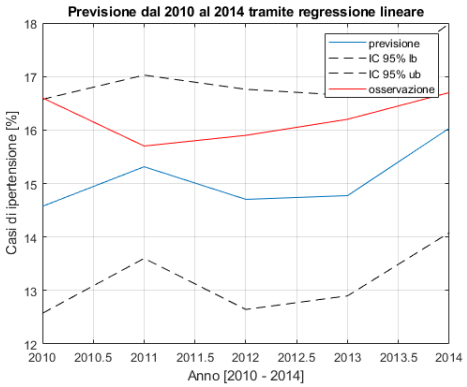
Risultati regressione con errori ARIMA

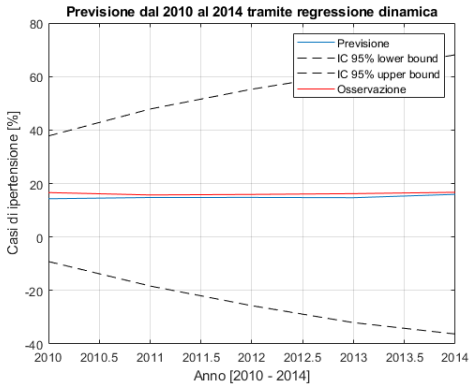
Regression with ARMA(2,0) Error Model (Gaussian Distribution):

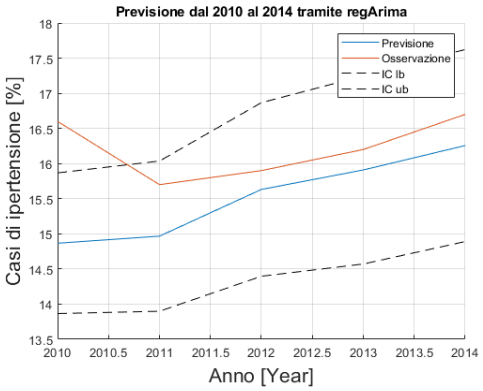
	Value	StandardError	TStatistic	PValue
Intercept	-17.2904176299697	2.22180208339645	-7.78215924774808	7.12970405036379e-15
AR{1}	0.374142734488574	0.179485291950917	2.08453144222475	0.0371118532643622
AR{2}	-0.75813574279424	0.214119677858774	-3.54071027182415	0.000399051517254835
Beta(1)	0.991373088130508	0.425000597641847	2.33263928011214	0.0196670855165973
Beta(2)	0.629582466861833	0.0808456492325839	7.78746256401003	6.83682878475506e-15
Variance	0.260737131381797	0.131469377558338	1.98325371447125	0.0473390966839039

Con regressore x_1 = Diabete e regressore x_2 = eccesso di peso. Tramite bootstrap parametrico si hanno i seguenti IC : *Intercetta*: [-20.74; -13.91], *Diabete:* [0.25 ;1.55], *Eccesso di peso:* [0.52;0.77], *AR(1)*: [0.03;0.71], *AR(2)*:[-0.99;-0.53] I risultati dell'analisi dei residui (residui iid *N (0, σ^2)*), ci permettono anche qui di effettuare i test sul modello (*t test*) con risultati affidabili: coefficienti significativi (p-value < 0.05).

Previsione







MSE regressione lineare: 1.63, MSE regressione dinamica: 2.07, MSE regARIMA: 0.78.

Per fare previsione il dataset è stato suddiviso in dataset training (dati dal 1990 al 2009), su cui sono stati stimati i parametri dei 3 modelli, e dataset test (dati dal 2010 al 2014), utilizzato per il confronto previsione-osservazione e per il calcolo del MSE. Qui sopra sono riportati i grafici previsione vs osservazione con relativi IC riferiti ai 3 modelli.

*7. Conclusioni **(New)***

Come evidenziato dalle analisi precedenti, i modelli di regressione lineare multipla e di regressione con errori ARIMA risultano essere significativi e correttamente specificati, a differenza invece del modello di regressione dinamica. Inoltre, anche dal punto di vista previsivo quest'ultimo modello risulta non affidabile (si noti l'eccessiva ampiezza degli intervalli di confidenza sulle previsioni). Per quanto riguarda la regressione lineare multipla e la regressione ARIMA, le previsioni risultano per entrambi i modelli buone. Il modello regArima, tuttavia, mostra intervalli di confidenza più stringenti e un minore MSE, denotando quindi una migliore capacità previsiva. In definitiva, dunque, la regressione con errori ARIMA permette di ottenere un modello con performance migliori.

8. Sitografia

- *Script Matlab & Dataset utilizzato (Repository Github)*
- *Mardia (EN)*
- *Indice di Mardia (IT)*
- *Correlazione tra ipertensione e bronchite cronica*
- *Test di Breusch-Pagan - Wikipedia*
- *Analisi tecnica: medie mobili e trading - Borsa Italiana*
- *Statistica di Durbin-Watson - Wikipedia*
- *Serie Storiche (istat.it)*
- *Due facce della stessa moneta: diabete e ipertensione - Paginemediche*