Elenco delle figure

Elenco delle tabelle

CAPITOLO 1

Simulazione dominio a C

Prima di applicare il modello allo studio della produzione di rifiuti nella provincia di Venezia sono state eseguite simulazioni su un caso noto e più semplice. Si è scelto di analizzare il dominio a forma di C e la corrispondente funzione spaziale $g(\underline{p})$ (riportata in fig. 1.1) descritti in CITAZIONE NECESSARIA e nel pacchetto R *mgcv*. La funzione $g(\underline{p})$ è solo spaziale, quindi è stata introdotta una variazione temporale deformando con il coseno:

$$f(\underline{p}, t) = g(\underline{p})cos(t)$$

Su questo semplice caso sono stati eseguiti i primi tentativi per il modello STR-PDE sia senza covariate che con una covariata generata.

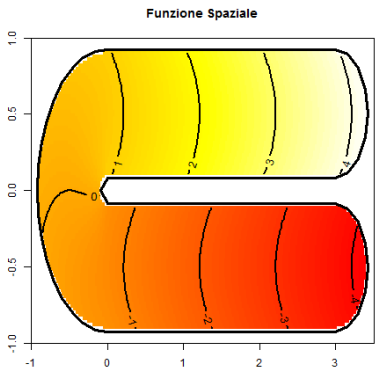


Figura 1.1: Funzione spaziale $g(\underline{p})$

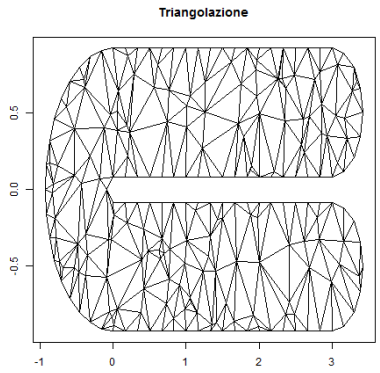


Figura 1.2: Triangolazione del dominio a forma di C

1.1 Triangolazione e istanti temporali

Nel dominio a forma di C non sono presenti punti spaziali definiti dalla natura del problema (come possono essere i comuni per la provincia di Venezia), quindi è stato necessario ricavarli. Sono stati generati casualmente 150 punti all'interno del rettangolo $(-1, +3.5) \times (-1, +1)$ e di questi sono stati considerati validi solo quelli che ricadevano all'interno del dominio. Non è stata usata la descrizione della frontiera presente in *mgcv*, ma una versione diversa che permette di avere punti anche nella parte rettilinea del bordo. In fig. 1.2 è riportata la triangolazione ottenuta grazie al pacchetto R *RTriangle*. Come basi in spazio sono stati usati gli elementi finiti lineari definiti su questa triangolazione. In tutti gli esempi che seguiranno sarà considerata questa descrizione del dominio, che è formata da 241 punti (pari anche al numero di basi spaziali N). Di questi, 108 sono di frontiera e i restanti 133 corrispondono agli n punti sui quali saranno disponibili i dati.

Come intervallo temporale di variazione dei dati è stato scelto $[0, 2\pi]$, per sfruttare la periodicità del coseno. All'interno di questo intervallo sono stati ricavati 9 istanti temporali equidistanti tra di loro, quindi uno ogni $\frac{\pi}{4}$. Si è scelto di fissare come basi in tempo le B-splines cubiche. Il numero di basi M è uguale al numero di istanti temporali a disposizione m, quindi 9.

1.2 Caso senza covariata

1.2.1 Ricerca del miglior $\underline{\lambda}$

Nei punti e negli istanti temporali disponibili i dati sono stati ricavati dalla funzione esatta con l'aggiunta del rumore:

$$z_{ij} = g(\underline{p}_i)cos(t_j) + \varepsilon_{ij} \qquad \forall i \in 1 \dots n, \forall j \in 1 \dots m$$

dove

$$\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, 0.5^2) \qquad \forall i \in 1 \dots n, \forall j \in 1 \dots m \; .$$

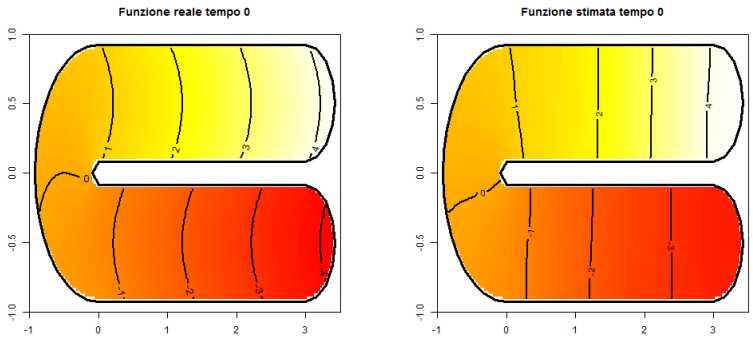
Per poter eseguire una analisi ottimale, come primo passo è necessario scegliere i valori per λ ottimizzando l'indice GCV($\underline{\lambda}$) come riportato in RIMANDO NECESSARIO. In queste analisi λ_S e λ_T sono sempre espressi in potenze di 10. Per trovare dei buoni valori per i parametri si procede per tentativi, creando due insiemi discreti di variazione per $\log_{10}\lambda_S$ e $\log_{10}\lambda_T$ e minimizzando sui $\underline{\lambda}$ corrispondenti al prodotto cartesiano tra di essi. Il procedimento viene iterato qualche volta (a causa dell'alto costo computazionale non è opportuno eseguire troppe iterazioni) rendendo la griglia sempre più fitta. In particolare, nel primo caso i valori sono distanziati di 1, poi (una volta che è possibile centrare gli intervalli in base al risultato precedente) di 0.25 e 0.125.

Intervalli per $\log_{10}\lambda_S$ e $\log_{10}\lambda_T$	Miglior valore
$\log_{10}\lambda_S \in \{-5, -4, \dots, +1\}$	$\underline{\lambda} = (10^0, 10^{-3})$
$\log_{10}\lambda_T \in \{-5, -4, \dots, +1\}$	
$\log_{10}\lambda_S \in \{-1, -0.75, \dots, +1\}$	$\underline{\lambda} = (10^{-0.5}, 10^{-3.25})$
$\log_{10}\lambda_T \in \{-4, -3.75, \dots, -2\}$	
$\log_{10}\lambda_S \in \{-1, -0.875, \dots, +0\}$	$\underline{\lambda} = (10^{-0.375}, 10^{-3.25})$
$\log_{10}\lambda_T \in \{-3.75, -3.625, \dots, -2.75\}$	

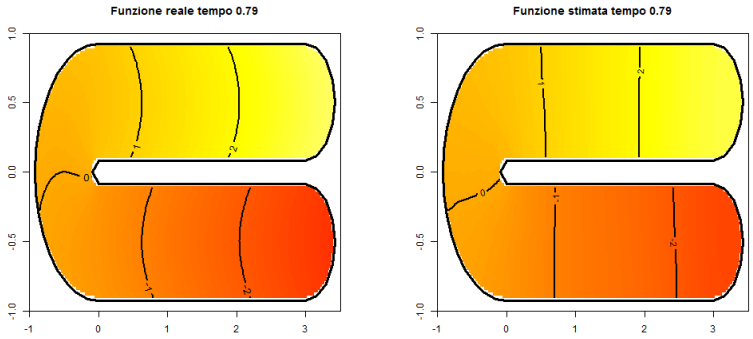
Tabella 1.1: Analisi di GCV($\underline{\lambda}$) per il dominio a forma di C, caso senza covariata

1.2.2 Risultati

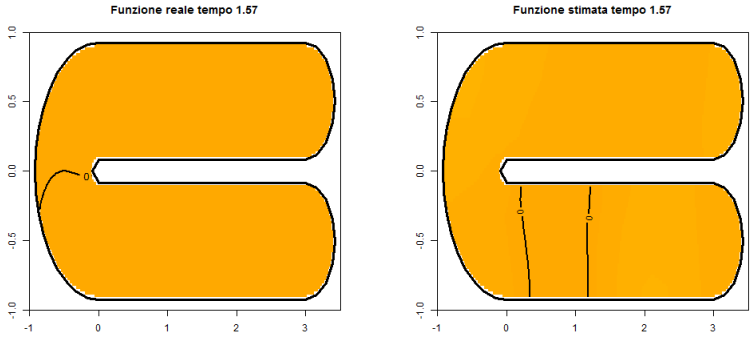
L'analisi è stata eseguita con $\underline{\lambda} = (10^{-0.375}, 10^{-3.25})$ e la stima della funzione si è rivelata molto buona. In fig. 1.3 sono riportati i confronti tra funzione reale e stimata nei primi istanti di tempo (la scala di colori è stata resa uniforme tra tutti i grafici). Si può notare come la funzione stimata sia effettivamente molto simile a quella reale.



(a) Funzione reale a $t = 0$ (b) Funzione stimata a $t = 0$



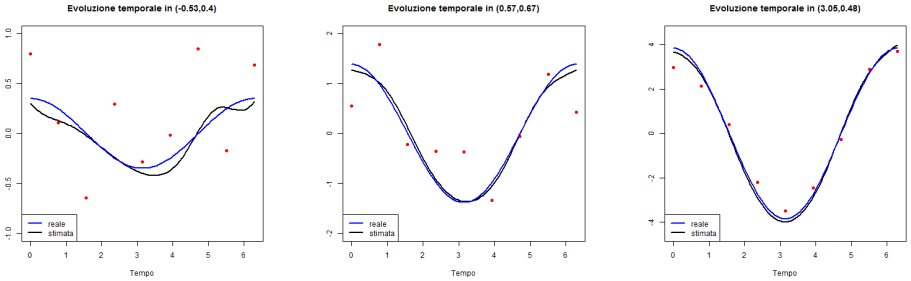
(c) Funzione reale a $t = \frac{\pi}{4}$ (d) Funzione stimata a $t = \frac{\pi}{4}$



(e) Funzione reale a $t = \frac{\pi}{2}$ (f) Funzione stimata a $t = \frac{\pi}{2}$

Figura 1.3: Stime della funzione $f(\underline{p}, t)$ ad alcuni istanti di tempo, caso senza covariata

In fig. 1.4 si ha il confronto dell'evoluzione temporale in alcuni nodi della triangolazione. Oltre alla curva stimata è tracciata la reale, che è una



(a) Ampiezza 0.3486 (b) Ampiezza 1.3819 (c) Ampiezza 3.8395

Figura 1.4: Evoluzione temporale in alcuni nodi della triangolazione, caso senza covariata

cosinusoide di ampiezza nota (grazie alla perfetta conoscenza di $g(\underline{p})$) riportata con i grafici. I punti rossi corrispondono al dato sporcato dal rumore. Tanto più è vicina a zero l'ampiezza della curva, tanto più il rumore influenza la stima poichè più rilevante (infatti, in fig. 1.3, le curve di livello della funzione stimata a $t = \frac{\pi}{2}$ sono le più distanti dalle reali). Tuttavia, anche nel caso con ampiezza vicina al valore massimo di $g(\underline{p})$ in fig. 1.4(c), la curva stimata non è una perfetta interpolazione dei dati e coglie il vero andamento temporale della funzione. Si può concludere quindi che la stima è molto buona, sebbene più confusa nella parte centrale del dominio a forma di C.

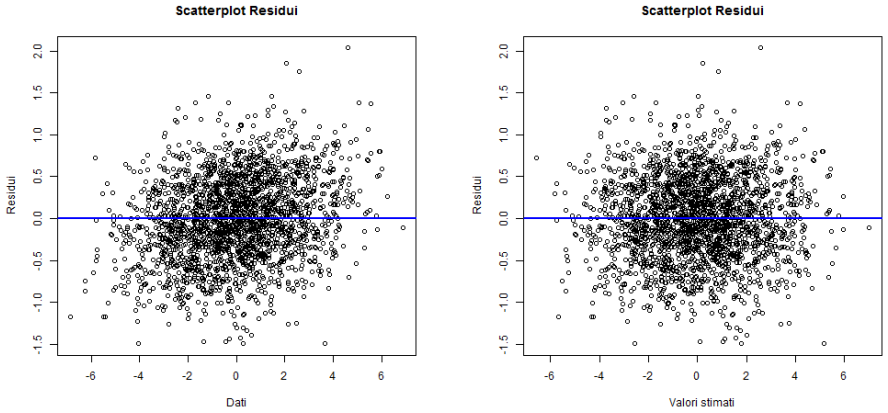
1.2.3 Analisi dei residui

Nella costruzione del modello spiegata nel dettaglio nel capitolo CITAZIONE NECESSARIA si ipotizza che il rumore aggiunto al dato funzionale sia generato da una variabile aleatoria di media nulla e varianza σ^2. Quindi si ha l'ipotesi di omoschedasticità che deve essere verificata per validare i risultati ottenuti, analogamente a quanto si fa per i modelli di regressione lineare. In questo caso il processo di generazione dei dati è perfettamente noto e il rumore rispetta questa ipotesi poichè generato da una normale con parametri fissi. Tuttavia, per introdurre una prassi che deve essere rispettata quando si controllano i risultati di questo algoritmo, anche in questo caso in cui già a priori si ha la certezza della validità dell'ipotesi di omoschedasticità è opportuno eseguire l'analisi dei residui.

Ad ogni dato è possibile associare il residuo

$$e_{ij} = z_{ij} - \hat{z}_{ij}$$

che, tramite opportuni scatterplot, è impiegato nella verifica dell'ipotesi di omogeneità della varianza.



(a) z_{ij} vs e_{ij} (b) $\hat{z}_{ij}$ vs e_{ij}

Figura 1.5: Scatterplot dei residui, caso senza covariata

Dall'analisi dei grafici in fig. 1.5 si può facilmente capire che non si ha una cattiva dispersione dei residui attorno allo zero. Quindi, come ci si aspettava, l'ipotesi di omoschedasticità può essere considerata valida.

Inoltre, il grafico in fig. 1.5(a) permette di evidenziare una delle particolarità dello smoothing associato all'algoritmo. I residui negativi più bassi sono in corrispondenza dei dati minori e, analogamente, i residui positivi più alti sono in corrispondenza dei dati maggiori. Questo è dovuto allo smoothing imposto nei massimi e nei minimi della funzione $f(\underline{p}, t)$: come si può notare in fig. 1.4(c) in questi punti si ha la maggior distanza tra la curva reale e quella stimata.

1.3 Caso con covariata

1.3.1 Generazione della covariata e ricerca del miglior $\underline{\lambda}$

Nel problema della stima della funzione $f(\underline{p}, t) = g(\underline{p})cos(t)$ non sono presenti covariate. Quindi per poter provare il modello in questo caso, è stato necessario generare valori da assumere come covariata in ogni punto spaziale ed istante temporale in cui si hanno le misurazioni della risposta.

In definitiva i dati sono così formati:

$$z_{ij} = g(\underline{p}_i)cos(t_j) + \beta w_{ij} + \varepsilon_{ij} \qquad \forall i \in 1 \dots n, \forall j \in 1 \dots m$$

dove covariata e rumore sono generate da due normali tra loro indipendenti:

$$w_{ij} \overset{\text{iid}}{\sim} N(0,1) \qquad \forall i \in 1 \dots n, \forall j \in 1 \dots m$$

$$\varepsilon_{ij} \overset{\text{iid}}{\sim} N(0, 0.5^2) \qquad \forall i \in 1 \dots n, \forall j \in 1 \dots m$$

e β è fissato a 1. Se il modello è buono, c'è da aspettarsi che la parte di funzione stimata senza covariata sia vicina a $f(\underline{p}, t)$ e che $\hat{\beta}$ si avvicini a 1.

Anche in questo caso è necessaria una analisi preliminare per fissare i valori per λ ottimizzando l'indice GCV($\underline{\lambda}$), che nel caso con covariata si differenzia dal precedente solo per la forma della *smoothing matrix*. In tab. 1.2 sono riportati i risultati ricavati adattando lo stesso approccio del caso senza covariata.

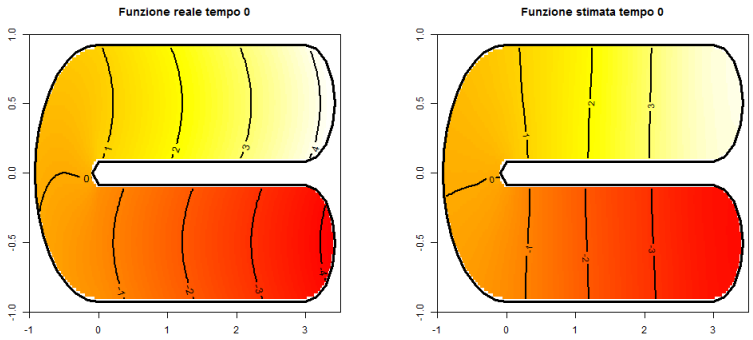
Intervalli per $\log_{10}\lambda_S$ e $\log_{10}\lambda_T$	Miglior valore
$\log_{10}\lambda_S \in \{-5, -4, \dots, +1\}$ $\log_{10}\lambda_T \in \{-5, -4, \dots, +1\}$	$\underline{\lambda} = (10^{0}, 10^{-4})$
$\log_{10}\lambda_S \in \{-1, -0.75, \dots, +1\}$ $\log_{10}\lambda_T \in \{-5, -4.75, \dots, -3\}$	$\underline{\lambda} = (10^{0.25}, 10^{-3.75})$
$\log_{10}\lambda_S \in \{-0.25, -0.125, \dots, +0.75\}$ $\log_{10}\lambda_T \in \{-4.25, -4.125, \dots, -3.25\}$	$\underline{\lambda} = (10^{-0.125}, 10^{-3.25})$

Tabella 1.2: Analisi di GCV($\underline{\lambda}$) per il dominio a forma di C, caso con covariata

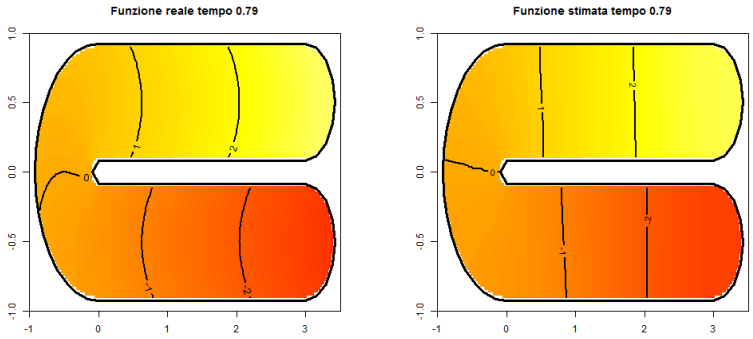
Procedendo per tentativi, si può notare come i valori si sono rivelati molto simili al caso senza covariata riportato in 1.1, facendo pensare che la stima della funzione $f(\underline{p}, t)$, cioè della parte non spiegata dalla covariata, possa essere molto vicina a quella stimata nel caso senza covariata e, quindi, a quella reale.

1.3.2 Risultati

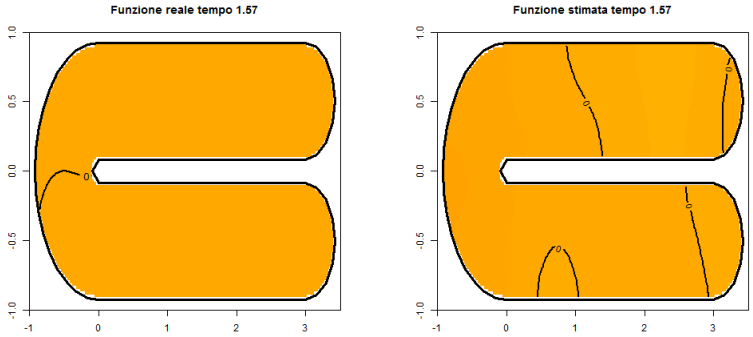
Eseguendo l'analisi con $\underline{\lambda} = (10^{-0.125}, 10^{-3.25})$ questa ipotesi è confermata e si trova una buona stima della funzione. In fig. 1.6 si hanno i grafici dei primi istanti di tempo (si ricorda che la funzione tracciata non contiene la parte spiegata dalla covariata, ma solo la stima di $f(\underline{p}, t)$).



(a) Funzione reale a $t = 0$ (b) Funzione stimata a $t = 0$



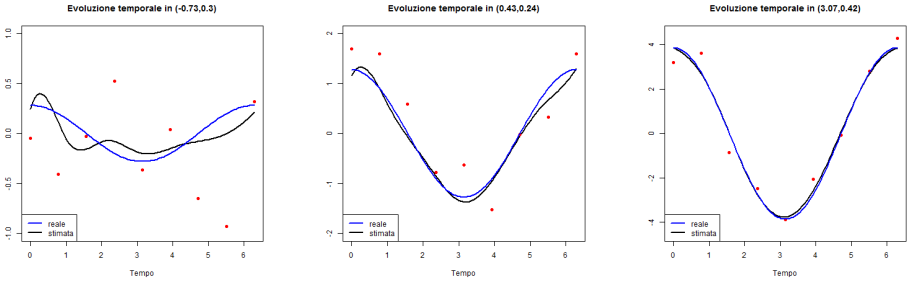
(c) Funzione reale a $t = \frac{\pi}{4}$ (d) Funzione stimata a $t = \frac{\pi}{4}$



(e) Funzione reale a $t = \frac{\pi}{2}$ (f) Funzione stimata a $t = \frac{\pi}{2}$

Figura 1.6: Stime della funzione $f(\underline{p}, t)$ ad alcuni istanti di tempo, caso con covariata

In fig. 1.7, analogamente a quanto fatto nel caso senza covariata, si hanno i grafici dell'evoluzione temporale della funzione in alcuni punti



(a) Ampiezza 0.2810 (b) Ampiezza 1.2761 (c) Ampiezza 3.8577

Figura 1.7: Evoluzione temporale in alcuni nodi della triangolazione, caso con covariata

fissati. I punti rossi tracciati corrispondono alla parte di dato senza il termine dovuto alla covariata. Le conclusioni sono le stesse del caso senza covariata: la stima è ben riuscita e la vera variazione temporale è stata colta dal modello. Tuttavia, avvicinandosi alla parte centrale del dominio a C sia una maggiore influenza del rumore.

Dai grafici precedenti si può concludere che la stima della parte funzionale della risposta sia effettivamente una buona approssimazione della reale. Tuttavia occorre verificare anche che il contributo delle covariata sia ben riconosciuto dal modello, e per questo basta controllare il valore stimato di β. Si ha:

$$\hat{\beta} \approx 1.005,$$

valore vicinissimo al reale.

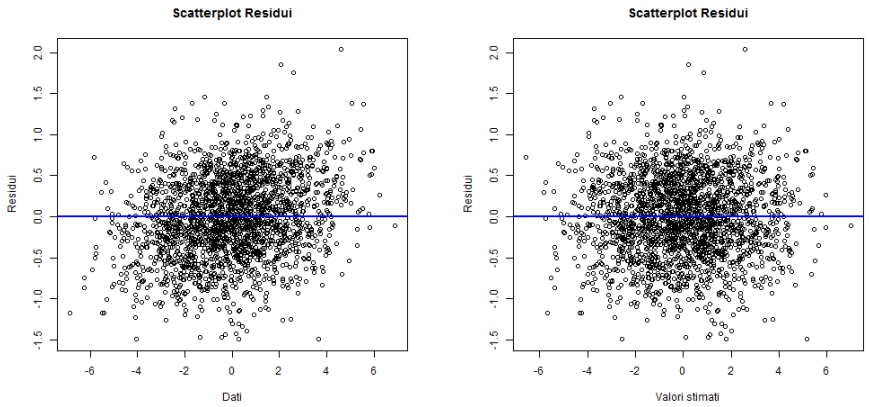
1.3.3 Analisi dei residui e test d'ipotesi per β

Sarebbe interessante eseguire un test del tipo

$$\begin{cases} H_0 : & \beta = \hat{\beta} \\ H_1 : & \beta \neq \hat{\beta} \end{cases}$$

per poter controllare con una data significatività se il valore stimato corrisponde a quello reale. Tuttavia è necessario controllare prima le ipotesi del modello e verificare se è possibile attribuire la normalità ai residui. Analogamente a quanto fatto nel caso senza covariata, siamo già certi che tutte queste ipotesi siano valide per come sono stati costruiti i dati. Tuttavia, per completezza, è necessario eseguire l'analisi dei residui anche in questo caso.

Per quanto riguarda l'omoschedasticità, in fig. 1.8 si hanno gli scatterplot dei residui. Non si hanno problemi riguardo all'ipotesi di varian-



(a) z_{ij} vs e_{ij} (b) $\hat{z}_{ij}$ vs e_{ij}

Figura 1.8: Scatterplot dei residui, caso con covariata

za uniforme e valgono le stesse considerazioni riportate nel caso senza covariata.

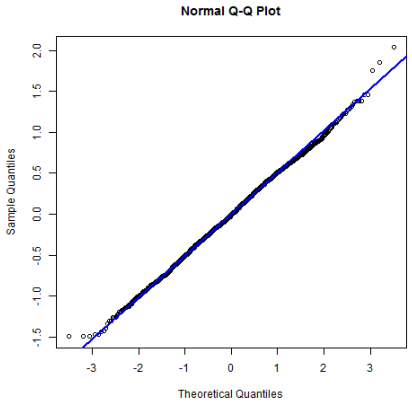


Figura 1.9: QQplot dei residui, caso con covariata

Riguardo alla normalità, sia in base al QQplot dei residui riportato in fig. 1.9 (che si adatta alla retta eccetto per qualche punto nelle code) sia in base al p-value del Shapiro-Wilk test (0.1259) si può concludere che i residui possono essere considerati gaussiani. Quindi è possibile costruire un intervallo di confidenza approssimato al 95% per β (eliminando il termine di distorsione) con quanto ricavato in CITAZIONE NECESSARIA. Ne risulta:

$$\beta \in [0.9843; 1.0259]$$

che contiene 1. L'ipotesi nulla è accettata.