UNIVERSITÀ DEGLI STUDI DI PADOVA

Facoltà di Scienze Statistiche

Corso di Laurea Specialistica in Statistica e Informatica

TESI DI LAUREA

Studio del Picco di Flusso Inspiratorio Nasale in Ambito Classico e Bayesiano

RELATORE: Ch.mo Prof. Stuart Coles

LAUREANDA: Lisa Pevarello

Anno Accademico 2005-2006

Ringraziamenti

In primis volevo ringraziare la mia famiglia per avermi sostenuto in questa nuova esperienza.

Un ringraziamento particolare va al professore Stuart Coles per l'estrema pazienza e competenza dimostratami in questi mesi.

Non dimentico, infine, il medico Giancarlo Ottaviano per avermi permesso di utilizzare i dati necessari per lo svolgimento della tesi e per la collaborazione prestata.

INTRODUZIONE

L'ostruzione nasale è uno dei più comuni problemi nella vita di tutti i giorni ad esempio è una manifestazione tipica della rinite allergica, ecco perchè la valutazione delle capacità respiratorie nasali è stato oggetto di ricerca per molti fisiologi e rinologi sin dal 1959.
La maggior parte di questi studi verteva sulla misura delle variazioni di pressione durante atti respiratori normali o durante respirazione forzata . Tra le metodiche più conosciute e tuttora in uso, per la valutazione registrazione delle resistenze nasali, vi è la rinomanometria ma nonostante sia una tecnica accettata e sicura, che possiede un limitato grado d'errore, presenta alcune limitazioni. Necessita infatti, di un'apparecchiatura ancora oggi alquanto costosa, richiede molta esperienza e dispendio di tempo, specie nel testare i bambini, spesso scarsamente collaboranti. Inoltre, per le sue dimensioni e per il numero di elementi ingombranti che lo compongono, non è facilmente trasportabile. Da anni, quindi, ricercatori e clinici a livello internazionale conducono studi e ricerche diverse tra loro, con un'unico minimo comune denominatore: perfezionare una metodica efficace, economica e semplice che consenta di valutare la pervietá nasale.
Tra queste tecniche troviamo il Picco di Fusso Inspiratorio Nasale (PNIF) che è stato presentato per la prima volta da Youlten nel 1980 e che consiste in una maschera facciale, a cui è collegato un misuratore, che il paziente deve applicare sulla faccia coprendo il naso, ma senza minimamente comprimerlo. Il soggetto da testare deve quindi inalare in maniera energica attraverso il naso, a bocca chiusa, partendo da un'espirazione profonda. Il picco di flusso sarà poi segnalato dal cursore sito allinterno del misuratore.
Il PNIF risulta, a differenza della rinomanometria, un metodo semplice, economico e facilmente eseguibile anche dai meno esperti per lo studio della pervietà nasale e con notevoli vantaggi specialmente se utilizzato per studi-

are i bambini, visto il basso grado di collaborazione richiesta.
Inoltre la misurazione del PNIF può essere un elemento utile nella valutazione della pervietà nasale e per tale motivo potrebbe risultare molto utile in campo rinologico ed allergologico. In particolare potrebbe essere utilizzato in maniera molto semplice nei test di provocazione nasale, nella valutazione dei flussi nasali pre- e post-operatori, anche avvalendosi di indici di pervietà nasale forzati espiratori ed inspiratori (rapporto tra picco di flusso nasale ed orale). Infine, data la facilità d'esecuzione, potrebbe essere utilizzato direttamente dal paziente a casa propria, ovviamente dopo aver ricevuto un'accurata spiegazione sulla maniera di adoperarlo. Ciò permetterebbe un'utile valutazione quotidiana dei risultati della terapia.
Il medico ha quindi deciso di eseguire uno studio pilota il cui scopo fosse quello di creare delle tabelle di normalità per l'utilizzo clinico quotidiano del PNIF.
Durante questo studio si è cercato di trovare valori normali di riferimento del PNIF nella popolazione adulta in relazione all'età, la statura e il sesso e successivamente si è valutato parallelamente al PNIF anche la funzione polmonare, cioè il PEF.
Lo studio di questo caso è stato effettuato su un campione costituito da soggetti sani, senza sintomi nasali (ostruzione, rinorrea, etc), non fumatori, senza asma e senza precedenti trattamenti chirurgici al distretto nasosinusale.
Ed anche se questo studio non è stato condotto su una popolazione di base, riteniamo che, essendo alquanto difficile trovare una popolazione ideale normale, sia stato un compromesso del tutto accettabile utilizzare il gruppo di volontari più accessibile.
Per tale motivo possiamo considerare i dati presentati in questo lavoro come preliminari, ma nondimeno di grande importanza poichè si focalizzano su un problema molto dibattuto in campo rinologico. Riteniamo che le tabelle pre-

sentate potranno essere un buon riferimento per tutti i medici che vogliano studiare la pervietà nasale in una popolazione caucasica.

Indice

1 Descrizione del problema **3**

2 Risoluzione del problema mediante R **7**

2.1 Primo esperimento 7

2.2 Secondo esperimento 14

2.3 Confronti tra i modelli stimati nei due esperimenti 21

3 MCMC e Gibbs Sampler **23**

3.1 Inferenza Bayesiana 24

3.2 Catene di Markov 28

3.3 Markov chain Monte Carlo (MCMC) e Gibbs Sampler 31

4 Winbugs **37**

5 Risoluzione del problema mediante Winbugs **51**

5.1 Primo esperimento 51

5.2 Secondo esperimento 56

5.3 Dati del primo e del secondo esperimento 61

5.4 Dati del primo e del secondo esperimento senza l'ipotesi di confrontabilità tra i modelli 65

6 Conclusioni **81**

Capitolo 1

Descrizione del problema

L'ostruzione nasale è un problema molto comune ai giorni nostri ecco perchè la valutazione delle capacità respiratorie nasali è oggetto di ricerca per molti fisiologi e rinologi.
Tra le varie tecniche in uso oggi vi è il Picco di Fusso Inspiratorio Nasale (PNIF) il quale consiste in una maschera facciale, a cui è collegato un misuratore, che il paziente deve applicare sulla faccia coprendo il naso, ma senza minimamente comprimerlo. Il soggetto da testare deve quindi inalare in maniera energica attraverso il naso, a bocca chiusa, partendo da un'espirazione profonda. Il picco di flusso sarà quindi segnalato dal cursore sito all'interno del misuratore.
Lo scopo del lavoro proposto dal medico, che si occupa di testare la validità di questo strumento, è di trovare valori normali di riferimento del PNIF nella popolazione adulta in relazione all'età, la statura e il sesso e successivamente di valutare parallelamente al PNIF anche la funzione polmonare, cioè PEF.
Per il primo esperimento sono state reclutate 170 persone di età compresa tra i 16 e gli 84 anni. Di queste persone però sono entrate nel nostro studio solo 137 volontari tutti soggetti sani, senza sintomi nasali (ostruzione, rinorrea, etc), non fumatori, senza asma e senza precedenti trattamenti chirurgici al

distretto naso-sinusale, ai quali prima dell'inizio del test sono state chieste informazioni circa l'età, la razza, il sesso, i medicinali usati e l'altezza.
Tutti i soggetti sono stati poi testati mentre erano seduti, ed è stata loro applicata sulla bocca la maschera dopo che hanno inalato aria in maniera energica dal naso. L'esperimento è stato poi ripetuto per altre due volte per avere conferma del risultato ottenuto.
Dopo un pò di tempo è stato effettuato un'altro esperimento in cui oltre a testare PNIF si è testato anche il flusso di espirazione (PEF) in quanto vi è un collegamento tra le due variabili.
Infatti se un soggetto ha volume polmonare (cioè PEF) grande allora ci saranno dei picchi di flusso (cioè PNIF) grandi. Questo significa che due soggetti che hanno la stessa età, la stessa altezza possono avere PNIF diversi perché hanno volumi polmonari diversi.
Anche in questo caso sono stati reclutati dei volontari di età compresa tra i 16 e gli 84 anni, e che sono 71, tutti soggetti sani, senza sintomi nasali (ostruzione, rinorrea, etc), non fumatori, senza asma e senza precedenti trattamenti chirurgici al distretto naso-sinusale, ai quali prima dell'inizio del test sono state chieste informazioni circa l'età, la razza, il sesso, i medicinali usati e l'altezza.
Anche in questo caso, tutti i soggetti sono stati testati mentre erano seduti ma a differenza del primo esperimento oltre ad inspirare dovevano anche espirare. L'esperimento è stato poi ripetuto per altre due volte.
Data la stretta relazione intercorrente tra PEF e PNIF, l'obiettivo del nostro studio è di vedere se il modello previsto per il primo esperimento possa essere applicato anche al secondo esperimento e se l'informazione contenuta in PEF possa dare maggiore informazione a PNIF.
Per lo sviluppo di questo studio si sono utilizzati due software applicativi: R e Winbugs.
Per quanto riguarda R tutte le analisi verranno fatte basandoci sul test stan-

dard dell'analisi della varianza e utilizzando come livello critico di significatività il 5 per cento.
Mentre in Winbugs tutte le analisi verranno fatte basandoci su distribuzioni a priori non informative.

Capitolo 2

Risoluzione del problema mediante R

2.1 Primo esperimento

Per procedere con l'elaborazione dei dati si è inizialmente provveduto all'acquisizione del data-set costituito da 137 unità statistiche.
Successivamente si è effettuato un esame grafico dei dati per vedere se esistono delle correlazioni fra le variabili esplicative e la variabile risposta.
Dall'analisi del grafico, in figura 2.1, si vede che PNIF aumenta con l'altezza mentre diminuisce con l'età ed inoltre mostra che c'è una leggera differenza tra i due sessi, sebbene ci sia un'ampia variabilità residua nelle variabili.
Questo suggerisce che stimare PNIF attraverso un modello di regressione lineare è inappropriato e inefficiente.
Per ridurre questo problema di eterogeneità nelle varibili si è , quindi, provato ad esaminare una trasformazione delle variabili e dopo molte prove, si è notato che la migliore trasformazione è quella che considera:

$$\text{MODPNIF}=(\text{PNIF})^{1/2}$$

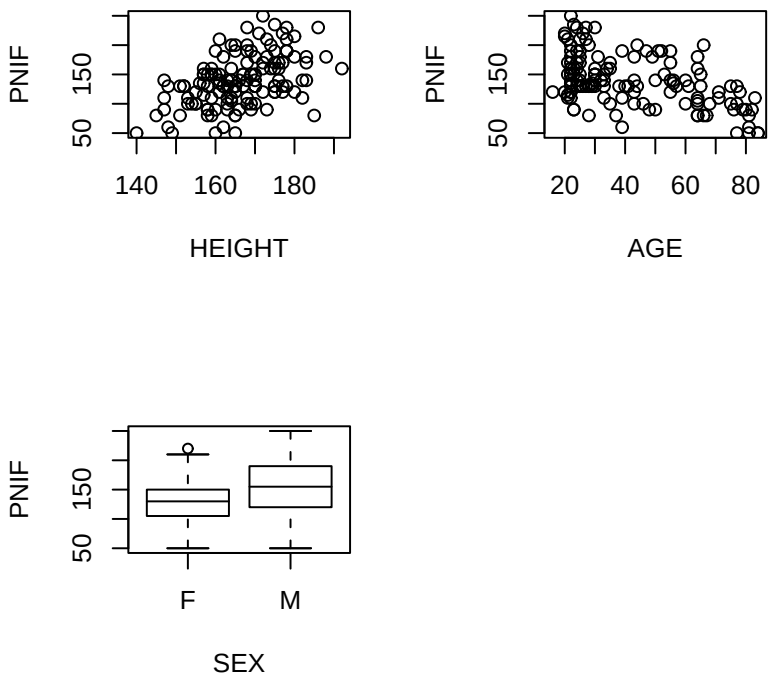

Figura 2.1: Diagramma di dispersione di PNIF in funzione di HEIGHT, AGE e SEX.

cioè la radice quadratica della variabile risposta.
Come si può notare nella Figura 2.2, utilizzando questa scala si è ridotta molto la variabilità delle variabili.

Successivamente si sono analizzati vari modelli:
Model 1: MODPNIF $\sim$ HEIGHT * AGE * SEX
Model2: MODPNIF $\sim$ HEIGHT *AGE + SEX
Model 3: MODPNIF $\sim$ HEIGHT + AGE + SEX
e dall'analisi della varianza

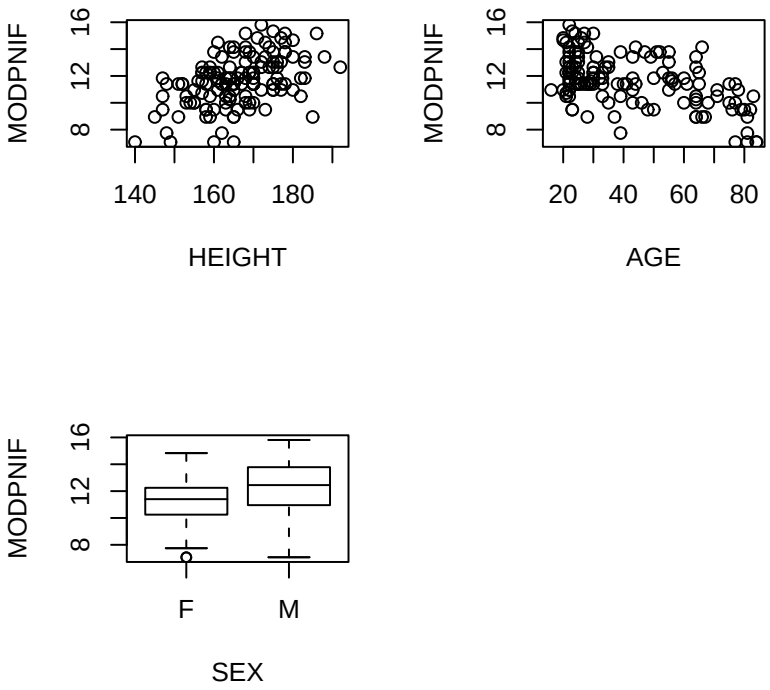

Figura 2.2: Diagramma di dispersione di MODPNIF in funzione di HEIGHT, AGE e SEX.

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	129	272.498				
2	132	285.796	-3	-13.298	2.0984	0.1036
3	133	289.106	-1	-3.310	1.5669	0.2129

si vede che il modello con grande complessità, cioè quello che include tutte le interazioni, e il modello che include un termine di interazione non comportano nessun miglioramento significativo.

Perciò il modello più appropriato per spiegare MODPNIF è:

$$\text{MODPNIF}=\alpha+\beta*\text{HEIGHT}+\gamma*\text{AGE}+\tau*\text{SEX}+\varepsilon$$

nella quale SEX è una variabile indicatrice che associa valore 1 ai maschi e valore 0 alle femmine, ed ε è una variabile normale.
Dalla sintesi prodotta dal summary per il modello preso in esame:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	7.29677	3.21672	2.268	0.0249
HEIGHT	0.03529	0.01877	1.880	0.0623
AGE	-0.04292	0.00758	-5.663	8.8e-08
SEXM	0.65957	0.33742	1.955	0.0527

si vede che sotto l'ipotesi di normalità la variabile AGE è molto significativa mentre le variabili HEIGHT e SEX sono marginalmente significative al livello 5%.
Si è, quindi, provato a togliere entrambi i parametri e poi ad inserire prima un parametro e poi l'altro e viceversa, ma si è ottenuto che il modello migliore è quello stimato prima anche se le due variabili sono significative solo marginalmente.
Successivamente, per avere una valutazione grafica della validità del modello stimato, si è costruito il grafico dei valori osservati di MODPNIF in funzione dei valori stimati dal modello.

Dall'analisi del grafico, in figura 2.3, si può osservare che i valori non tendono a distribuirsi lungo la bisettrice ma questo è in accordo con il fatto che R^2 non è molto elevato.
Infine, per verificare la bontà di adattamento del modello si è eseguita un'analisi dei residui, considerando il diagramma dei residui in funzione dei valori stimati e il diagramma Q-Q plot.

Dall'esame della Figura 2.4, che mostra il diagramma dei residui, non si

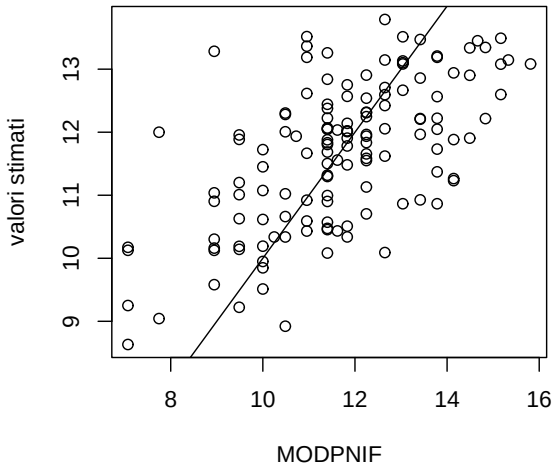

Figura 2.3: Grafico dei valori osservati in funzione dei valori stimati dal modello

evidenzia alcun andamento sistematico, mentre nella figura 2.5, che mostra il diagramma Q-Q plot, si osserva una leggera curvatura rispetto alla retta di interesse.
Concludendo il modello:

$$\text{MODPNIF}=\alpha+\beta\text{*HEIGHT}+\gamma\text{*AGE}+\tau\text{*SEX}+\varepsilon$$

è nel complesso abbastanza soddisfacente per spiegare PNIF a partire dalle variabili a disposizione.
Uno svantaggio di questo modello, però, consiste nel fatto che le medie stimate sono calcolate per la variabile MODPNIF invece di PNIF, cosa che invece ci interessa maggiormente.

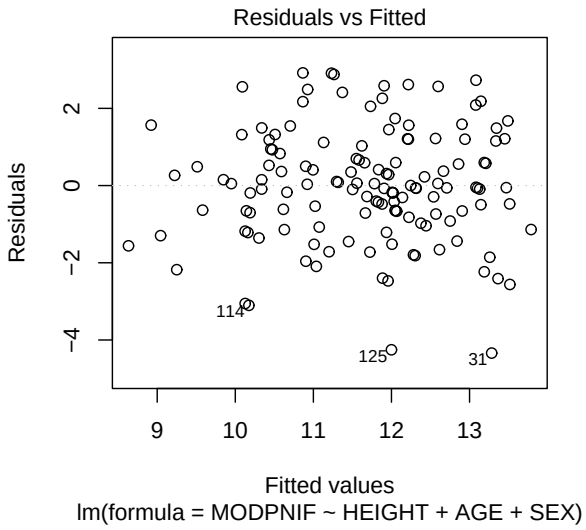

Figura 2.4: Grafico dei residui in funzione dei valori stimati

Comunque, osservando che per ogni variabile Z vale la seguente espressione:

$$\mathrm{Var(Z)=E[Z^2]\text{-}E[Z]^2}$$

si ha che:

$$\mathrm{E[Z^2]=E[Z]^2+Var(Z)}$$

Perciò, se μ_i e σ_i^2 rappresentano la media e la varianza di MODPNIF per ogni individuo, dall'osservazione precedente segue che la media stimata di PNIF è semplicemente $\mu_i^2 + \sigma_i^2$.
In pratica questo implica che le medie stimate di PNIF sono ottenute dalle medie stimate di MODPNIF per ogni sesso con specificata età e altezza attraverso la trasformazione, più un termine addizionale che è la stima di σ_i^2.

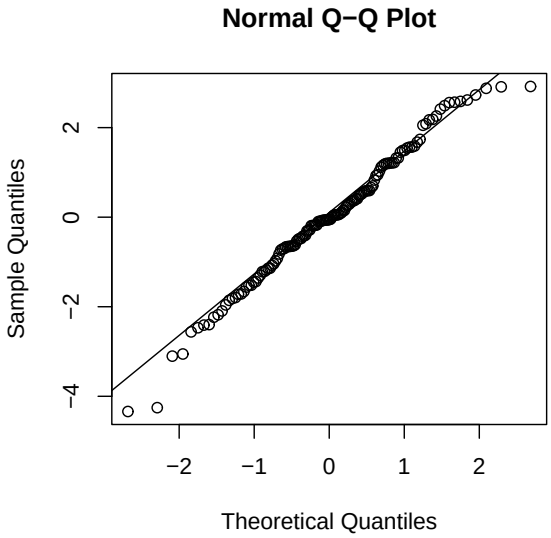

Figura 2.5: Grafico Q-Q normale

Dove, quest'ultima quantità si ottiene calcolando l'analisi della varianza del modello e risulta pari a 2.174.

2.2 Secondo esperimento

Analogamente a quanto si è effettuato per il primo esperimento si è inizialmente provveduto all'acquisizione del data-set costituito da 71 unità statistiche.
Successivamente si è effettuato un esame grafico dei dati per vedere se esistono delle correlazioni fra le variabili esplicative e la variabile risposta.
La Figura 2.6, infatti, mostra che PNIF aumenta con l'altezza e con PEF,

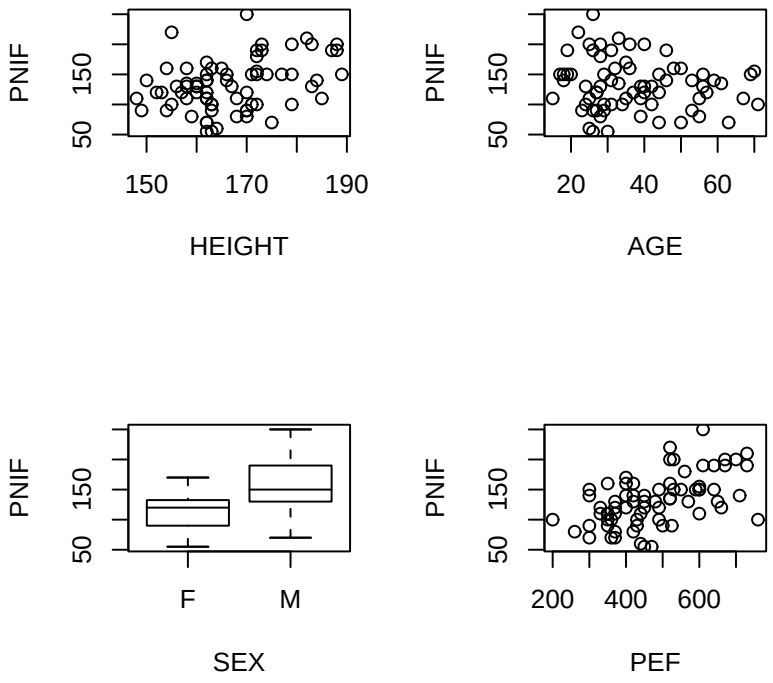

Figura 2.6: Diagramma di dispersione di PNIF in funzione di HEIGHT, AGE,SEX e PEF.

mentre diminuisce con l'età ed inoltre mostra che c'è una differenza tra i due sessi e che c'è un'ampia variabilità residua.
Da una prima analisi dei dati , inoltre, si vede che mentre le variabili HEIGHT, AGE e SEX hanno la stessa natura, cioè sono variabili esplicative, la variabile PEF ha la stessa natura di PNIF quindi dovremmo considerarla come variabile risposta.
Inizialmente, però, si è costruito un modello che considera PEF come variabile esplicativa e per ridurre il problema di eterogeneità nelle varibili si è provato ad esaminare una trasformazione delle variabili. Dopo molte prove si è notato che la migliore trasformazione è quella che considera:

$$\text{MODPNIF}=(\text{PNIF})^{1/2}$$
$$\text{MODPEF}=(\text{PEF})^{1/2}$$

cioè la radice quadratica della variabile risposta e di PEF.
Come si può notare nella Figura 2.7, utilizzando questa scala si è ridotta un pò la variabilità delle variabili.

Successivamente si sono analizzati diversi modelli:
Model 1: MODPNIF $\sim$ HEIGHT * AGE * SEX * MODPEF
Model 2: MODPNIF $\sim$HEIGHT * AGE + SEX * MODPEF
Model 3: MODPNIF $\sim$ HEIGHT + AGE + SEX +MODPEF
e tramite l'analisi della varianza di essi:

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	55	138.716				
2	64	151.546	-9	-12.831	0.5653	0.8194
3	66	156.542	-2	-4.996	0.9904	0.3780

si vede che il modello con grande complessità, cioè quello che include tutte le interazioni, e il modello che include due termini di interazione non com-

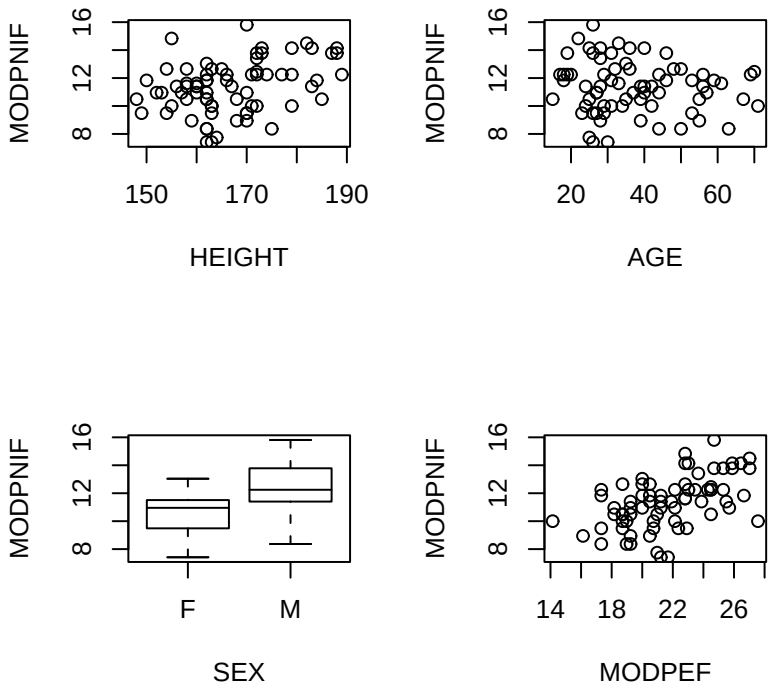

Figura 2.7: Diagramma di dispersione di MODPNIF in funzione di HEIGHT, AGE, SEX e MODPEF.

portano nessun miglioramento significativo.
Quindi il modello più appropriato per spiegare MODPNIF è:

$$\text{MODPNIF}=\alpha+\beta*\text{HEIGHT}+\gamma*\text{AGE}+\tau*\text{SEX}+\delta*\text{MODPEF}+\varepsilon$$

nella quale SEX è una variabile indicatrice che associa valore 1 ai maschi e valore 0 alle femmine, e ε è una variabile normale.
Dalla sintesi prodotta dal summary del modello preso in esame:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	9.789330	4.497567	2.177	0.03309
HEIGHT	-0.026579	0.026899	-0.988	0.32670
AGE	-0.000438	0.014765	-0.030	0.97642
SEXM	1.433730	0.539746	2.656	0.00990
MODPEF	0.247233	0.085844	2.880	0.00536

si vede che sotto l'ipotesi di normalità le variabili SEX e MODPEF sono molto significative mentre le variabili HEIGHT e AGE non sono significative.
Si è quindi provato a togliere i due parametri e poi ad inserire prima un parametro e poi l'altro e viceversa, e si è visto che il modello migliore è il seguente:

$$\text{MODPNIF}=\alpha+\tau*\text{SEX}+\delta*\text{MODPEF}+\varepsilon$$

il cui summary è:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	6.07742	1.52090	3.996	0.000161
SEXM	1.16783	0.43726	2.671	0.009459
MODPEF	0.21874	0.07422	2.947	0.004389

Successivamente, per avere una valutazione grafica della validità del modello stimato, si è quindi considerato il grafico dei valori osservati di MODPNIF in funzione dei valori stimati dal modello.

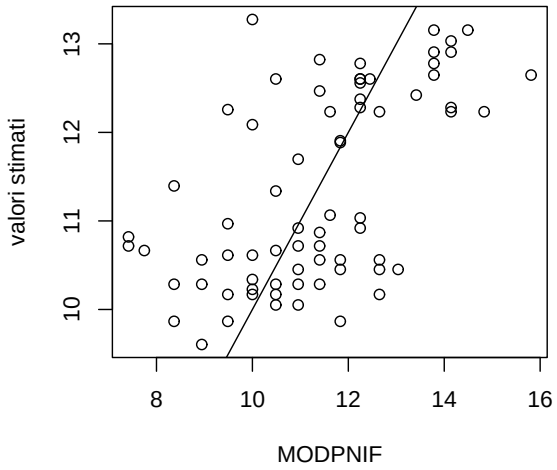

Figura 2.8: Grafico dei valori osservati in funzione dei valori stimati dal modello

Dall'analisi del grafico, in figura 2.8, si vede che i valori non tendono a distribuirsi lungo la bisettrice in accordo con il fatto che R^2 non è molto elevato (infatti è pari a 0.3413).
Infine, per verificare la bontà di adattamento del modello si è eseguita un'analisi dei residui, considerando il diagramma dei residui in funzione dei valori stimati e il diagramma Q-Q plot.

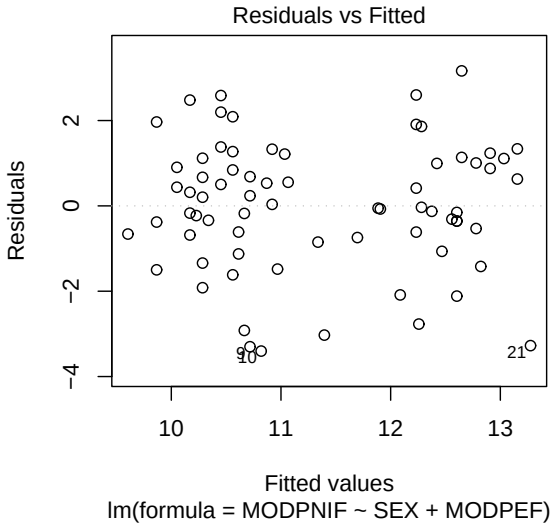

Figura 2.9: Grafico dei residui in funzione dei valori stimati

Dall'esame della Figura 2.9, che mostra il grafico dei residui, non si evidenzia alcun andamento sistematico, mentre nella figura 2.10, che mostra il diagramma Q-Q plot, si osserva una curvatura rispetto alla retta di interesse. Concludendo il modello:

$$\text{MODPNIF}=\alpha+\tau*\text{SEX}+\delta*\text{MODPEF}+\varepsilon$$

è nel complesso abbastanza soddisfacente per spiegare MODPNIF a partire dalle variabili a disposizione.
Anche in questo caso, però, quello che veramente ci interessa è la stima di PNIF, perciò utilizzando le osservazioni fatte precedentemente si vede che le medie stimate di PNIF sono ottenute dalle medie stimate di MODPNIF per ogni sesso con specificata PEF attraverso la trasformazione, più un termine addizionale che è la stima di σ_i^2.

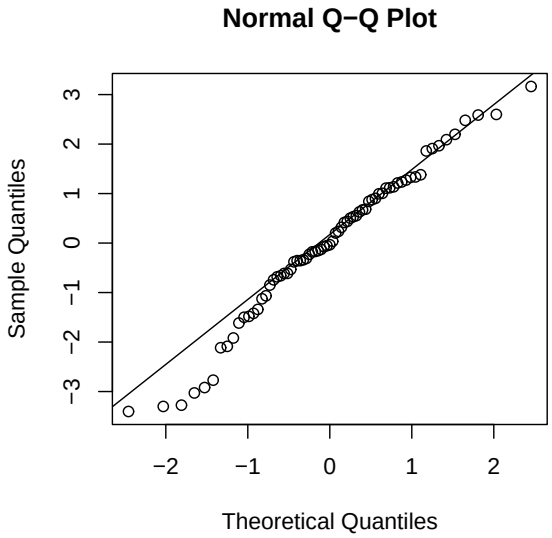

Figura 2.10: Grafico Q-Q normale

Dove, quest'ultima quantità si ottiene calcolando l'analisi della varianza del modello e risulta pari a 2.338.

2.3 Confronti tra i modelli stimati nei due esperimenti

Dalle analisi effettuate precedentemente si è visto che:

- per il primo esperimento il modello stimato è:

$$\text{MODPNIF}=\alpha+\beta*\text{HEIGHT}+\gamma*\text{AGE}+\tau*\text{SEX}+\varepsilon$$

- per il secondo esperimento il modello stimato è:

$$\text{MODPNIF}=\alpha_1+\tau_1*\text{SEX}+\delta_1*\text{MODPEF}+\varepsilon_1$$

Come si può vedere i due esperimenti a prima vista non sono tra loro confrontabili e quindi non è possibile stimare i dati del secondo esperimento attraverso il modello stimato per il primo esperimento.
Per vedere se questo è effettivamente vero, si è inizialmente supposto di non considerare la variabile MODPEF in modo da ottenere il confronto tra due modelli con le stesse variabili.
Da tale confronto può risultare o che i modelli non sono confrontabili a causa dell'inserimento nel secondo esperimento di una nuova variabile o che i modelli non sono confrontabili a causa di altri fattori come ad esempio il fatto che la numerosità dei campioni non è la stessa oppure che alcune variabili sono più significative in uno dei due esperimenti.
Per il confronto tra i due esperimenti, si è quindi creato un data-set che contiene i dati del primo e del secondo esperimento e si è creata una variabile indicatrice ESP che indica se il dato appartiene al primo o al secondo esperimento.
In questo caso il modello da stimare è il seguente:

$$\text{MODPNIF}=\alpha+\beta*\text{HEIGHT}+\gamma*\text{AGE}+\tau*\text{SEX}+\text{ESP}$$

e il relativo summary mostra che:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	9.041551	2.701371	3.347	0.000973
HEIGHT	0.022764	0.015788	1.442	0.150880
AGE	-0.038825	0.006839	-5.677	4.7e-08
SEXM	1.046004	0.291616	3.587	0.000419
ESP1	-0.512713	0.229030	-2.239	0.026266

la variabile ESP risulta significativa, quindi i due esperimenti non sono confrontabili.
Successivamente, per vedere se la causa della non confrontabilità tra gli esperimenti è causata dal fatto che alcune variabili sono più significative in uno dei due esperimenti, si è considerato il modello che contiene tutte le interazioni e dall'analisi di questo si è visto che:

	Estimate	Std. Error	t value	Pr(> |t|)
(Intercept)	20.905419	5.502829	3.799	0.000193
HEIGHT	-0.051871	0.034073	-1.522	0.129516
AGE	-0.325750	0.122353	-2.662	0.008394
SEXM	2.493080	0.681634	3.658	0.000326
ESP1	-2.144432	0.618264	-3.468	0.000642
HEIGHT:AGE	0.001846	0.000779	2.370	0.018747
AGE:SEXM	-0.040537	0.014756	-2.747	0.006563
AGE:ESP1	0.032539	0.014285	2.278	0.023800
SEXM:ESP1	0.700889	0.446869	1.568	0.118366

le variabili AGE e SEX interagiscono maggiormente con la variabile ESP1. Quindi i due esperimenti non sono tra loro confrontabili a causa del fatto che le variabili AGE e SEX sono più significative nel primo esperimento.

Capitolo 3

MCMC e Gibbs Sampler

Dalle analisi effettuate precedentemente si sono evidenziati alcuni problemi che sono difficilmente risolvibili tramite l'utilizzo di R.
Inanzitutto si è evidenziato il fatto che nel secondo esperimento la variabile PEF è una variabile esplicativa e quindi la variabile risposta è bidimensionale.
Successivamente, dal confronto tra i modelli si è tolta la variabile PEF in quanto nel primo esperimento questa variabile non appariva. Questo dimostra che nel caso volessimo confrontare i due modelli con tutte le variabili presenti otterremo un data-set con dati mancanti.
Per risolvere questi due problemi che sono molto importanti per caratterizzare in maniera più precisa i risultati ottenuti si sono utilizzati metodi MCMC, in particolare la variante che verrà utilizzata è quella del Gibbs Sampler.
Ma prima di procedere alla risoluzione del problema diamo alcune nozioni sulle tecniche che verranno poi utilizzate.

3.1 Inferenza Bayesiana

L'inferenza bayesiana è un approccio all'inferenza statistica in cui le probabilità non sono interpretate come frequenze, proporzioni o concetti analoghi, ma piuttosto come livelli di fiducia nel verificarsi di un dato evento.
Il nome deriva dal teorema di Bayes, che fornisce il fondamento di questo approccio, infatti esso fornisce un metodo per modificare il livello di confidenza di una data ipotesi alla luce di nuova informazione.
Dati due eventi qualsiasi A e B, il teorema di Bayes può essere enunciato come segue:

$$P(A|B) = \frac{P(B|A) * P(A)}{P(B)}$$

dove:

- $P(A)$ è la probabilià a priori dell'evento A.
- $P(B|A)$ rappresenta la funzione di verosimiglianza di A, ed è ciò su cui si fonda l'inferenza classica.
- $P(B)$ è detta probabilità marginale ed è una costante di normalizzazione.
- $P(A|B)$ è la probabilità a posteriori dell'evento B.

Il fattore di scala $P(B|A)/P(B)$ può essere interpretato come misura dell'impatto che l'osservazione di B ha sul grado di confidenza del ricercatore, rappresentato a sua volta da $P(A)$. La probabilità a posteriori, di conseguenza, combina le convinzioni che il ricercatore ha a priori con quelle derivanti dall'osservazione del nuovo evento.
Nel contesto della statistica bayesiana le probabilità si considerano, quindi, una misura del grado soggettivo di confidenza da parte del ricercatore e si

suppone che restringano le potenziali ipotesi a un insieme limitato inquadrato in un modello di riferimento.
Il teorema di Bayes dovrebbe, quindi, fornire un criterio razionale per valutare fino a che punto una data osservazione dovrebbe alterare le convinzioni del ricercatore; in questo caso tuttavia la probabilità rimane soggettiva dunque è possibile usare il teorema per giustificare razionalmente una qualche ipotesi, ma alle spese di rifiutare l'oggettività delle informazioni che ne derivano.
L'inferenza bayesiana ha a lungo rappresentato una corrente minoritaria nella teoria della statistica. Ciò è in larga parte dovuto alle difficoltà algebriche che essa pone; infatti se siamo in ambito continuo il teorema di Bayes diventa:

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta) * L(\theta;\mathbf{y})}{\int_\Theta \pi(\theta) * L(\theta;\mathbf{y}) d\theta}$$

dove:

- $\pi(\theta)$ è la probabilià a priori.
- $L(\theta;\mathbf{y})$ è la funzione di verosimiglianza.
- $\int_\Theta \pi(\theta) * L(\theta;\mathbf{y}) d\theta$ è la costante di normalizzazione.
- Θ rappresenta l'insieme dei valori assumibili dal parametro θ e θ è k-dimensinale.

Come si può vedere la difficoltà maggiore consiste nel calcolare l'integrale a denominatore che in quasi tutti i casi non ha una forma esplicita.
Questa difficoltà ha fino a pochi anni fa limitato la capacità della statistica bayesiana di riprodurre modelli realistici della realtà. Al fine di evitare di ricorrere in questo problema, gran parte dei risultati erano basati sulla teoria delle coniugate, che sono particolari famiglie di distribuzioni per cui la probabilità a posteriori risulta avere la stessa forma di quella a priori.

Successivamente, grazie alle maggiori disponibilità di risorse informatiche (anni '90) è stato possibile superare tali difficoltà. È infatti possibile risolvere gli integrali in via numerica, aggirando i problemi algebrici. Questa possibilità ha inoltre stimolato l'applicazione alla statistica bayesiana di metodi numerici sviluppati in altri contesti, come quelli basati sulla simulazione, tra cui il Metodo Monte Carlo, gli algoritmi del Gibbs Sampler e di Metropolis-Hastings, i quali permettono di analizzare le proprietà di $\pi(\theta)$ tramite simulazione ma che non possono essere applicati quando la distribuzione del parametro θ è multidimensionale o quando la costante di normalizzazione è ignota.
Inoltre si è avuto lo sviluppo di metodi nuovi nell'ambito della statistica bayesiana stessa, ad esempio i popolari metodi basati sul Markov Chain Monte Carlo (o MCMC), i quali permettono di calcolare la distribuzione a posteriori senza necessariamente conoscere la costante di normalizzazione.
Lo sviluppo di questi nuovi metodi ha notevolmente incrementato la popolarità dell'inferenza bayesiana tra gli statistici; sebbene i bayesiani costituiscano ancora una minoranza.
Oltre alla difficoltà di calcolare gli integrali, questo approccio prevede altri svantaggi come ad esempio la necessità di specificare la distribuzione a priori, $\pi(\theta)$, e il fatto che diverse scelte di $\pi(\theta)$ portano a calcolare diverse distribuzioni finali, $\pi(\theta|\mathbf{y})$.
Per specificare $\pi(\theta)$ si distinguono tre casi principali:

- Priori non informative: questo caso si ha quando non si dispone di nessuna informazione a priori e quindi per definirle ci si basa solamente sulla struttura statistica dell'esperimento.

- Conoscenza a priori vaga: questo caso si verifica quando, per la quantità di informazione posseduta, è irragionevole costruire una priori non

informativa e quindi la posteriori è essenzialmente la verosimiglianza normalizzata.

- Conoscenza a priori sostanziale: questo caso si verifica quando l'informazione a priori è abbastanza forte da far sì che la posteriori sia diversa dalla verosimiglianza.

3.2 Catene di Markov

Prima di descrivere i metodi MCMC facciamo una breve introduzione sulle catene di Markov in quanto questi metodi utilizzano la simulazione di catene di Markov per campionare distribuzioni di probabilità.
Se X_t denota il valore della variabile casuale al tempo t, e se l'insieme degli stati è discreto, finito e contiene tutti i possibili valori di X. La variabile casuale è un processo di Markov se la probabilità di transizione tra due stati differenti dipende solamente dallo stato attuale in cui la variabile si trova, cioè:

$$Pr(X_{t+1} = s_j | X_0 = s_k, ..., X_t = s_i) = Pr(X_{t+1} = s_j | X_t = s_i)$$

In altre parole, il futuro (X_{t+1}), condizionato al presente (X_t), è indipendente dal passato($X_0, ..., X_{t-1}$). Una catena di Markov è, quindi, una sequenza di variabili casuali ($X_0, ...X_n$) generata da un processo di Markov.
Una particolare catena è definita attraverso le sue probabilità di transizione, $p_{i,j}$, che rappresentano la probabilità che un processo allo stato s_i raggiunga lo stato s_j in un solo passo, cioè:

$$p_{i,j} = Pr(X_{t+1} = s_j | X_t = s_i)$$

e che sono gli elementi della matrice di probabilità di transizione (o nucleo di transizione):

$$P = \begin{pmatrix} p_{1,1} & \dots & p_{1,n} \\ & \vdots & \\ p_{n,1} & \dots & p_{n,n} \end{pmatrix}$$

Se

$$\pi_j(t) = Pr(X_t = s_j)$$

denota la probabilità che la catena sia allo stato j al tempo t, allora la

probabilità che la catena si trovi allo stato j al tempo$t+1$ è data da:
$\pi_j(t+1) = Pr(X_{t+1} = s_j) =$

$$= \sum_k Pr(X_{t+1} = s_j | X_t = s_k) \cdot Pr(X_t = s_k) = \sum_k p_{k,j} \cdot \pi_k(t)$$

Inoltre se lo stato iniziale della catena è definito da $\pi(0) = [\pi_1, ..., \pi_n]$, dove $\pi_j = Pr(X_0 = j)$, allora il vettore degli stati al tempo t è:

$$\pi(t) = \pi(t-1) * P = (\pi(t-2) * P) * P = ... = \pi(0) * P^t$$

La distribuzione π=[π_1,...,π_m] è definita invariante (o stazionaria) per la catena markoviana se:

$$\pi P = \pi$$

Cioè, se gli stati della catene hanno probabilità π al passo j, avranno la stessa distribuzione al passo $j+1$ e, per ricorsione, ad ogni passo successivo.
Una catena markoviana ha un'unica distribuzione invariante che coincide con la distribuzione limite della catena se la catena è:

- irriducibile: cioè se il movimento tra gli stati è sempre possibile
- aperiodica: cioè se il numero di passi richiesti per andare dallo stato i allo stato j non è un multiplo dello stesso intero
- ricorrente positiva: cioè se il numero atteso di transizioni di ritorno a qualsiasi stato è finito

cioè se la catena è ergodica.
Una condizione sufficiente per ottenere distribuzioni invarianti è che la catena sia reversibile. Una catena markoviana è reversibile se esiste una distribuzione $\pi = [\pi_1, ..., \pi_n]$ tale che;

$$\pi_i p_{i,j} = \pi_j p_{j,i}$$

per ogni scelta di i e j.
Di conseguenza, se riusciamo a trovare una distribuzione π che soddisfa l'equazione di reversibilità , e se la catena è ergodica, allora π sarà l'unica distribuzione invariante della catena e sarà inoltre uguale alla distribuzione limite.
L'idea base delle catene markoviane con spazio degli stati discreti può essere generalizzata al caso continuo dove le transizioni sono definite secondo transizioni kernel del tipo:

$$P(x, A) = Pr(X_{t+1} \in A | X_t = x)$$

Nel caso continuo rimangono comunque validi i concetti di irriducibilità, aperiodicità e ricorrenza positiva come nel caso discreto, sebbene siano più complicate le definizioni formali.
Una catena con queste proprietà formali è ancora detta ergodica e questa proprietà è strettamente sufficiente per la validità degli algoritmi MCMC.

3.3 Markov chain Monte Carlo (MCMC) e Gibbs Sampler

I metodi MCMC sono una classe di algoritmi utilizzati, anche nella statistica bayesiana, per simulare da una distribuzione π da cui sarebbe difficile simulare tramite altri metodi.
L'idea base dei metodi MCMC è semplice: nel caso in cui non si riesca a generare direttamente variabili aleatorie da una distribuzione π su A ,o perchè la dimensione dello spazio è molto grande, o perchè π è complicata , si costruisce una catena di Markov avente come unica distribuzione invariante proprio π e si simula l'andamento della catena.
Per costruire una catena di Markov che abbia π come distribuzione limite vi sono vari metodi tra cui l'algoritmo del Metropolis-Hastings, che funziona nel seguente modo:

1. Si sceglie un arbitrario punto iniziale $\theta^{(0)}$=i con t=0;
2. Si simula θ^* dalla funzione di densità $\pi(\theta^*|\theta^{(t)})$;
3. Si calcola
$$\alpha = \alpha(\theta^{(t)}, \theta^*) = \min\left\{1, \frac{\pi(\theta^*) * \pi(\theta^{(t)}|\theta^*)}{\pi(\theta^{(t)}) * \pi(\theta^*|\theta^{(t)})}\right\}$$
4. Si definisce:
$$\theta^{(t+1)} = \begin{cases} \theta^* & \text{con probabilità } \alpha \\ \theta^{(t)} & \text{con probabilità } 1-\alpha \end{cases}$$
5. Si pone i=$\theta^{(t+1)}$, t=t+1 e si torna al passo 2.

Questo algoritmo definisce, quindi, una catena markoviana in quanto la generazione di $\theta^{(t+1)}$ dipende solo da $\theta^{(t)}$ ed inoltre π è una distribuzione

invariante della catena.
Per dimostrare quest'ultima affermazione si deve dimostrare che la catena è reversibile rispetto a π cioè che vale:

$$\pi_i * Q(i,j) = \pi_j * Q(j,i) \qquad \forall \quad i,j$$

dove:

- $\pi_i = \pi(\theta^{(t)} = i) = \pi(\theta^{(t)})$
- $\pi_j = \pi(\theta^* = j) = \pi(\theta^*)$
- $Q(i,j) = \pi(\theta^* = j|\theta^{(t)} = i) * \alpha(\theta^{(t)}, \theta^*) = \pi(\theta^*|\theta^{(t)}) * \alpha(\theta^{(t)}, \theta^*)$
- $Q(j,i) = \pi(\theta^{(t)} = j|\theta^* = i) * \alpha(\theta^{(t)}, \theta^*) = \pi(\theta^{(t)}|\theta^*) * \alpha(\theta^{(t)}, \theta^*)$

Questo è triviale se $i = j$.
Quando $i \neq j$

$$\begin{aligned}
\pi_i * Q(i,j) &= \pi(\theta^{(t)}) * \pi(\theta^*|\theta^{(t)}) * \alpha(\theta^{(t)}, \theta^*) \\
&= \pi(\theta^{(t)}) * \pi(\theta^*|\theta^{(t)}) * \min\left\{1, \frac{\pi(\theta^*) * \pi(\theta^{(t)}|\theta^*)}{\pi(\theta^{(t)}) * \pi(\theta^*|\theta^{(t)})}\right\} \\
= \min\left\{\pi(\theta^{(t)}) * \pi(\theta^*|\theta^{(t)}), \frac{\pi(\theta^*) * \pi(\theta^{(t)}|\theta^*) * \pi(\theta^{(t)}) * \pi(\theta^*|\theta^t)}{\pi(\theta^{(t)}) * \pi(\theta^*|\theta^{(t)})}\right\} \\
&= \min\left\{\pi(\theta^{(t)}) * \pi(\theta^*|\theta^{(t)}), \pi(\theta^*) * \pi(\theta^{(t)})|\theta^*)\right\} \\
&= \pi(\theta^*) * \pi(\theta^{(t)})|\theta^*) * \min\left\{\frac{\pi(\theta^{(t)}) * \pi(\theta^*|\theta^{(t)})}{\pi(\theta^*) * \pi(\theta^{(t)}|\theta^*)}, 1\right\} \\
&= \pi_j * \pi(\theta^{(t)}|\theta^*) * \alpha(\theta^{(t)}, \theta^*) \\
&= \pi_j * Q(j,i)
\end{aligned}$$

L'algoritmo appena descritto rimane valido anche quando θ è multivariato, cioè quando $\theta = (\theta_1, \theta_2, \ldots, \theta_n)$, salvo che le transizioni vengono fatte in

uno spazio multivariato.
In quest'ultimo caso però ci sono due versioni dell'algoritmo:

- la prima utilizza l'algoritmo scritto precedentemente dove al posto di θ^* si considera $\underline{\theta^*}$ e al posto di θ si considera $\underline{\theta_i}$=$(\theta_{i,1}, \ldots, \theta_{i,n})$.
- la seconda applica l'algoritmo a ciascuna componente del vettore θ, una per volta, quindi la proposta per la componente θ_j dipende solo dal valore corrente di tale componente.

In pratica è molto utilizzata la seconda versione di questo algoritmo in quanto molte volte funziona bene, anche se, in teoria, la prima versione potrebbe essere più efficiente se il valore iniziale è scelto bene.
In ambito multivariato un algoritmo molto efficiente che però si può utilizzare solo se la simulazione dalle densità condizionate è possibile e non troppo difficile è il Gibbs Sampler il quale è un caso speciale della seconda versione del Metropolis-Hastings. In questo algoritmo $\alpha(\theta^{(t)}, \theta^*)$ è sempre uguale a 1 e questo significa che tutti i valori proposti sono sempre accettati.
Assumiamo, quindi, di essere interessati a calcolare $\pi(\theta)$ dove $\theta = (\theta_1, ..., \theta_d)'$, in cui ogni θ_i può essere un vettore o una matrice e supponiamo, inoltre, di conoscere tutte le distribuzioni condizionate $\pi(\theta_i) = \pi(\theta_i|\theta_{i-1})$, $i = 1, ..., d$.
In questo caso per le transizioni proposte c'è una scelta particolare:

$$\pi_j(\theta_j|\underline{\theta}) = \pi(\theta_j|\underline{\theta}_{(j)})$$

in cui $\underline{\theta}_{(j)}$ è il vettore θ senza la j-esima componente. Dunque, π_j è la densità di θ_j condizionata a tutte le altre componenti del vettore θ.
In questo caso:

$$\alpha = \min\left\{1, \frac{\pi(\underline{\theta^*}) * \pi_j(\theta_{(j)})|\underline{\theta^*})}{\pi(\underline{\theta}) * \pi_j(\theta_j^*|\underline{\theta})}\right\}$$
$$= \min\left\{1, \frac{\pi(\underline{\theta^*}) * \pi(\theta_{(j)})|\underline{\theta^*}_{(j)})}{\pi(\underline{\theta}) * \pi(\theta_j^*|\underline{\theta}_{(j)})}\right\}$$

$$= \min\left\{1, \frac{\pi(\underline{\theta}^*) * \pi(\theta_{(j)}|\underline{\theta}_{(j)})}{\pi(\underline{\theta}) * \pi(\theta_j^*|\underline{\theta^*}_{(j)})}\right\}$$
$$= \min\left\{1, \frac{\pi(\underline{\theta}^*) * \pi(\underline{\theta})/\pi(\underline{\theta}_{(j)})}{\pi(\underline{\theta}) * \pi(\underline{\theta^*})/\pi(\underline{\theta^*}_{(j)})}\right\}$$
$$= 1$$

dove si è usato il fatto che $\underline{\theta}^*_{(j)} = \underline{\theta}_{(j)}$.
Questo metodo verrà utilizzato per lo svolgimento della tesi in quanto consente di calcolare $\pi(\theta)$ generando una serie di valori dalle distribuzioni condizionate che sono sempre accettati; ed il suo metodo di funzionamento è descritto qui sotto:

1. Si sceglie un'arbitrario punto iniziale $\theta^0 = (\theta_1^0, ...\theta_d^0)$ e si pone $j = 1$.

2. Si ottiene un nuovo valore $\theta^j = (\theta_1^j, ..., \theta_d^j)'$ da θ^{j-1} attraverso successive generazioni di valori:
   $\theta_1^j \sim \pi(\theta_1|\theta_2^{j-1}, ..., \theta_d^{j-1})$
   $\theta_2^j \sim \pi(\theta_2|\theta_1^j, \theta_3^{j-1}..., \theta_d^{j-1})$
   $\vdots$
   $\theta_d^j \sim \pi(\theta_d|\theta_1^j, ..., \theta_{d-1}^j)$

3. Si pone $j = j + 1$ e si torna al passo 2 finchè non si raggiunge la convergenza.

Se la convergenza è stata raggiunta, allora il valore calcolato θ^j ha come distribuzione limite $\pi(\theta)$ che è proprio il valore che si voleva calcolare.
Il problema di fondo è quindi stabilire se la convergenza è stata raggiunta. Per studiare la convergenza ci sono due approcci molto validi.
Il primo approccio è molto teorico e consiste nel misurare la variazione totale della distanza tra la distribuzione della catena all'iterazione j e la distribuzione limite π. Questo approccio fu proposto da Meyer e Tweedie (1994), Polson (1996), Robert e Polson (1994). Questa è un'area in via di sviluppo

ma in questo momento i risultati a cui si è pervenuti hanno un piccolo impatto nella pratica.
Il secondo approccio è più empirico e consiste nell'analizzare le proprietà statistiche delle catene osservate. Tra i vari metodi che rientrano in questo approccio ci sono:

- Il metodo di Gelfan e Smith (1990) il quale suggerisce di analizzare la convergenza utilizzando tecniche grafiche.
- Il metodo di Gelman e Rubin (1992) il quale suggerisce di considerare più catene contemporaneamente e verificare se il loro comportamento è lo stesso.

La difficoltà di questo approccio è che non garantisce la convergenza perchè si basa solamente sulle osservazioni della catena.

Capitolo 4

Winbugs

Winbugs é un pacchetto del programma BUGS (Bayesian Inference Using Gibbs Sampling) che utilizza le tecniche MCMC per effettuare analisi bayesiane di modelli statistici complessi.
Per fare un'analisi utilizzando Winbugs si devono seguire i seguenti passi:

- Specificazione del modello
- Compilazione del modello
- Monitoraggio dei valori dei parametri stimati
- Visualizzazione della convergenza

illustriamo, quindi, i seguenti passi attraverso degli esempi pratici.
Gli esempi pratici che si utilizzeranno sono contenuti negli esempi di Winbugs e sono stati scelti, tra i tanti presenti, in quanto considerano degli aspetti che sono comuni al problema discusso nell'introduzione.
Tramite questi esempi si mostrerà com' è possibile affrontare problemi con variabili risposta univariata, multivariata e problemi in cui ci sono dati mancanti.

CASO UNIVARIATO

Supponiamo di considerare 30 topi e supponiamo di aver misurato per 5 settimane il loro peso, il nostro obiettivo è di trovare una relazione che lega il peso con i giorni.
Una parte del data-set è rappresentata qui sotto, dove Y_{ij} rappresenta il peso dell'i-esimo topo al tempo x_j.

Topi	x_j =8 15 22 29 36
Rat1	151 199 246 283 320
Rat2	145 199 249 293 354
...	... ...
Rat26	160 207 257 303 345
...	... ...
Rat30	153 200 244 286 324

Analizzando la figura 4.1, che mostra l'andamento delle curve dei primi sei topi, vediamo un evidente andatura curvilinea di queste. Il grafico mostra anche che è possibile stimare ogni curva attraverso un modello di regressione lineare con valori di α e di β diversi per ogni curva. Quindi in questo caso si assume che gli α_i e i β_i, per $i = 1, ..., 30$, non siano tra loro correlati.Per ottenere ciò, bisogna perciò standardizzare gli x_i attorno alla loro media ed inoltre non bisogna considerare il parametro che descrive la correlazione tra α_i e β_i.
Il modello che vogliamo stimare è quindi del seguente tipo:

$$\begin{aligned} Y_{ij} &\sim \text{Normal}(\alpha_i + \beta_j(x_j - \overline{x}), \tau_c) \\ \alpha_i &\sim \text{Normal}(\alpha_c, \tau_\alpha) \\ \beta_i &\sim \text{Normal}(\beta_c, \tau_\beta) \end{aligned}$$

dove $\overline{x} = 22$ rappresenta la media di tutti gli x_i e τ rappresenta la precisione di una distribuzione normale. Questo tipo di modello viene denominato

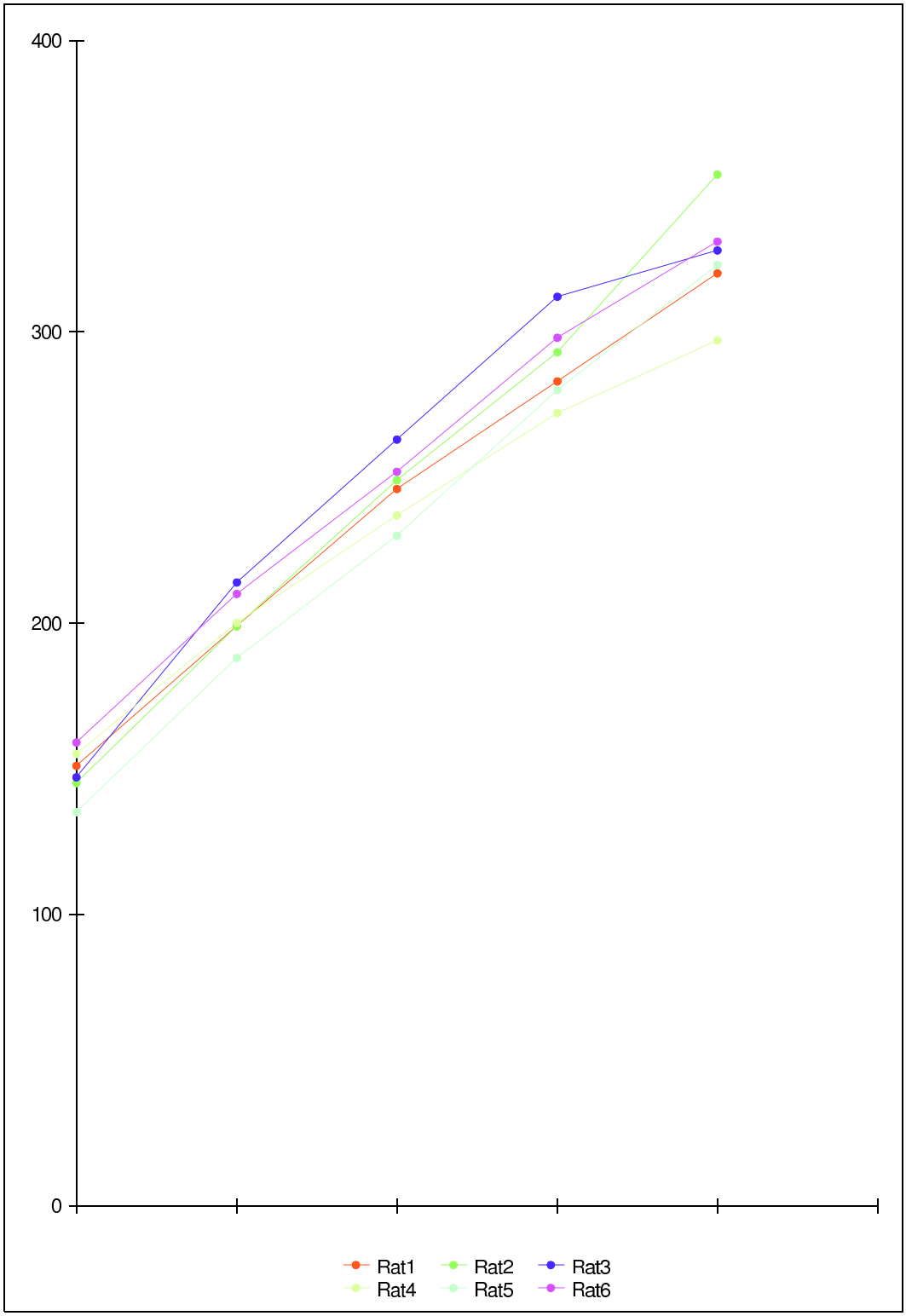

Figura 4.1: Grafico relativo ai primi 6 topi

"modello gerarchico" in quanto permette di costruire le priori di parametri che sono in qualche modo collegati tra di loro.
Supponiamo, inoltre, che $\alpha_c, \tau_\alpha, \beta_c, \tau_\beta, \tau_c$ siano delle distribuzioni a priori "non informative" ed indipendenti ed inoltre supponiamo di essere interessati a stimare l'intercetta al tempo zero, cioè $\alpha_0 = \alpha_c - \beta_c * \overline{x}$.
Dopo aver individuato il modello da stimare si deve specificarlo in linguaggio Winbugs. Per fare ciò esistono due metodi:

- creare un documento di testo nel quale viene utilizzato il simbolo $\sim$ per denotare le relazioni stocastiche e il simbolo $\leftarrow$ per denotare le relazioni logiche.

- utilizzare il Doodles-BUGS che permette di rappresentare graficamente il modello, specificando tutte le quantità con dei nodi che possono essere stocastici, deterministici oppure costanti.

Quindi il modello di prima nel linguaggio Winbugs diventa:

```
model
{
for(i in 1:N){
for(j in 1:T){
Y[i,j]~ dnorm(mu[i,j],tau.c)
mu[i,j]<- alpha[i]+beta[i]*(x[i]-xbar)
}
alpha[i]~ dnorm(alpha.c,alpha.tau)
beta[i]~dnorm(beta.c,beta.tau)
}
tau.c~dgamma(0.001,0.001)
sigma<- 1/sqrt(tau.c)
alpha.c~dnorm(0.0,1.0E-6)
alpha.tau~dgamma(0.001,0.001)
```

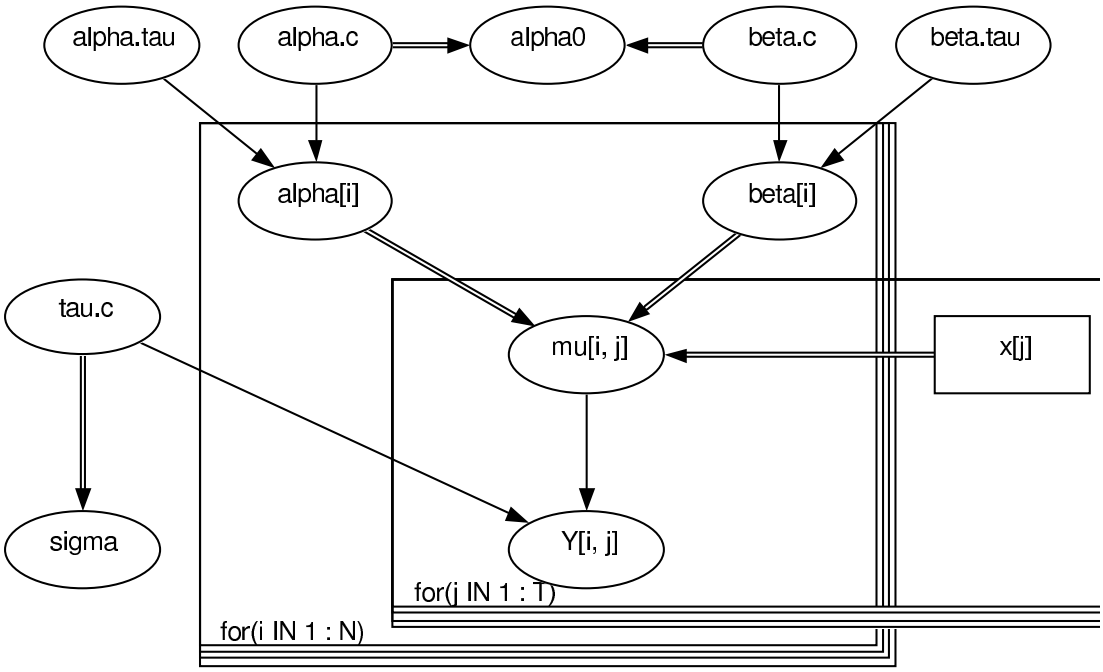

Figura 4.2: Modello grafico

beta.c∼dnorm(0.0,1.0E-6)
beta.tau∼dgamma(0.001,0.001)
alpha0← alpha.c-xbar*beta.c }
ed il relativo modello grafico è rappresentato in figura 4.2.
Dopo aver salvato il documento di testo appena creato si devono creare due file aggiuntivi , uno che contiene i dati e uno che contiene i valori iniziali dei nodi stocastici, che devono essere scritti in formato lista o in formato rettangolare.
Dopo questa fase iniziale si deve controllare che il modello scritto precedentemente non contenga errori. Per fare ciò si clicca sull'icona $Model$ e si sceglie l'opzione $Specification$ in questo modo comparirà una finestra, successivamente si seleziona la finestra contenente il modello e si clicca sulla parola *model* della finestra aperta. Nel caso in cui il modello non presenti errori sotto la finestra di lavoro comparirà un messaggio con scritto "model is syntactically correct".
Successivamente si apre il file contenente i dati e si clicca su *load data* della finestra aperta precedentemente e se i dati sono stati letti comparirà un messaggio con scritto "data loaded".
Successivamente si seleziona il numero di catene (di default è 1) e poi si compila il modello cliccando su *compile*, se tutto è andato a buon fine apparirà un messaggio con scritto "model compiled". Dopo aver compilato il modello si apre il file contenente i valori iniziali e si clicca su *inits*, se non ci sono errori comparirà un messaggio con scritto "initial value loaded: model initialized nodes". Nel caso in cui avessi selezionato due (o più) catene dovrei aprire il file che contiene i valori iniziali per l' altra catene oppure potrei cliccare su *gen inits* per generare valori casuali per ogni parametro stocastico. Dopo aver fatto ciò si chiude la finestra $Specification$ e si fa partire la simulazione selezionando da $Model$ l'opzione $Update$ in questo modo apparirà una finestra di dialogo nella quale dobbiamo scrivere il numero di simulazioni che

vogliamo effettuare e successivamente dobbiamo cliccare su *update*. Se non ci sono stati problemi comparirà un messaggio con scritto "updates took *** s". Successivamente si chiude la finestra *update* e si monitorizza i valori dei parametri.
Nel caso in cui ho pochi parametri si seleziona da $Interface$ l'opzione $Samples$ in questo modo comparirà una finestra di dialogo nella quale vengono scritti, uno alla volta, i nomi dei parametri che vogliamo monitorare e vengono poi salvati cliccando su *set*.
Nel caso in cui ho tanti parametri, invece, si seleziona da $Interface$ l'opzione $Summary$ e si procede come scritto precedentemente.
Nel nostro esempio i risultati ottenuti dopo 10000 iterazioni sono:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
alpha0	106.6	3.625	0.03477	99.32	106.6	113.6	1001	10000
beta.c	6.185	0.1068	0.001354	5.979	6.184	6.398	1001	10000
sigma	6.082	0.4714	0.007308	5.248	6.052	7.093	1001	10000

Per visualizzare la convergenza dei parametri devo costruire dei grafici. Per ottenere tali grafici si seleziona da $Interface$ l'opzione $Samples$ successivamente si scrive il nome del nodo e si clicca su *history*, in questo modo apparirà una finestra contenente il grafico che visualizza la convergenza del parametro di interesse.
Nel nostro caso i grafici che rappresentano la convergenza sono raffigurati in figura 4.3. Come possiamo vedere dalla figura tutti i parametri considerati convergono ad un certo valore.
Dopo aver finito si salva ogni file creato in Winbugs e si esce.

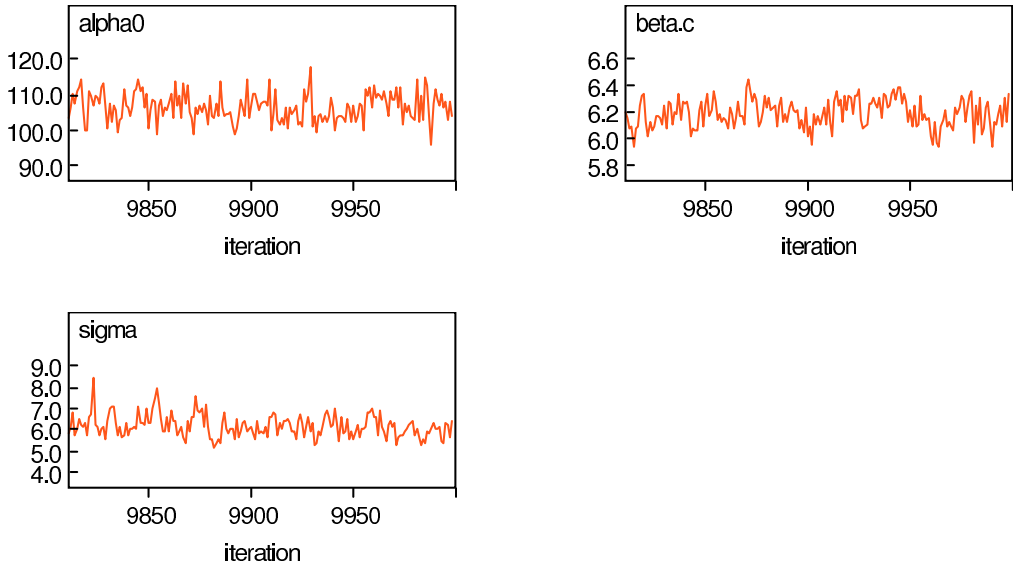

Figura 4.3: Convergenza dei parametri

CASO UNIVARIATO

CON DATI MANCANTI

In questa sezione mettiamo in evidenza il fatto che utilizzando algoritmi MCMC è possibile ottenere dei buoni risultati anche quando nel data-set ci sono dati mancanti; infatti tramite l'utilizzo di Winbugs si ottengono automaticamente le stime dei dati mancanti.
Riconsideriamo quindi l'esempio dei topi in cui si suppone di togliere l'ultima osservazione dai casi 6-10, le ultime due dai casi 11-20, le ultime tre dai casi 21-25 e le ultime quattro dai casi 26-30. Il data-set appropriato è quindi ottenuto semplicemente replicando i valori selezionati attaverso NA.
La specificazione del modello è la stessa dell'esempio precedente, quindi la distinzione tra quantità osservate e non osservate si vede solo nel data-set.
Il modello da stimare è quindi:

```
model
{
for(i in 1:N){
for(j in 1:T){
Y[i,j]~ dnorm(mu[i,j],tau.c)
mu[i,j]<- alpha[i]+beta[i]*(x[i]-xbar)
}
alpha[i]~ dnorm(alpha.c,alpha.tau)
beta[i]~dnorm(beta.c,beta.tau)
}
tau.c~dgamma(0.001,0.001)
sigma<- 1/sqrt(tau.c)
alpha.c~dnorm(0.0,1.0E-6)
alpha.tau~dgamma(0.001,0.001)
beta.c~dnorm(0.0,1.0E-6)
```

beta.tau∼dgamma(0.001,0.001)
alpha0← alpha.c-xbar*beta.c }

Procedendo nello stesso modo dell'esempio precedente e ponendo l'attenzione sulle stime ottenute per le quattro osservazioni finali del topo 26, si ottengono, dopo 10000 iterazioni, i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
Y[26,2]	204.5	8.74	0.1159	187.0	204.4	221.7	1001	10000
Y[26,3]	250.0	10.27	0.1642	229.7	249.9	270.1	1001	10000
Y[26,4]	295.4	12.64	0.2092	270.3	295.3	320.3	1001	10000
Y[26,5]	340.6	15.32	0.284	310.2	340.5	370.5	1001	10000
beta.c	6.575	0.1507	0.003708	6.281	6.573	6.875	1001	10000

Confrontando le stime ottenute per il topo 26 con i veri valori vediamo che:

Valore vero	Valore stimato
207	204
257	250
303	295
345	340

otteniamo valori molto simili. Questo dimostra che Winbugs è un ottimo strumento per stimare modelli in cui ci sono dati mancanti.

CASO MULTIVARIATO

Riconsideriamo l'esempio dei topi, e illustriamo l'uso della distribuzione normale multivariata per i coefficienti di regressione della curva di regressione per ogni topo. Questo modello è stato adottato da Gelfand (1990) per questi dati, e assume che a priori l'intercetta e il coefficiente angolare dei parametri per ogni topo siano correlati. Per esempio, una correlazione positiva indica che inizialmente i topi sono pesanti ma tendono molto velocemente a dimagrire.
Il modello appropriato è quindi:

$$\begin{aligned} \mathrm{Y}_{ij} &\sim \mathrm{Normal}(\mu_{ij}, \tau_C) \\ \mu_{ij} &= \beta_{1i} + \beta_{2i} * \mathrm{x}_j \\ \beta_i &\sim \mathrm{MVN}(\mu_\beta, \Omega) \end{aligned}$$

dove Y_{ij} è il peso dell'i-esimo topo al tempo x_j e β_i denota il vettore (β_{1i}, β_{2i}). In questo caso si assume una priori 'non-informativa' di variabili normali univariate e indipendenti per ciascuna delle componenti μ_{β_1} e μ_{β_2}. Per definire la priori di Ω si utilizza Wishart(R,ρ), dove la matrice R è specificata da:

$$\mathrm{R}= \begin{pmatrix} 200 & 0 \\ 0 & 0.2 \end{pmatrix}$$

mentre per definire τ_C si utilizza la priori 'non-informativa' Gamma(0.001,0.001). Quindi, questo modello in linguaggio Winbugs diventa:

```
    model
{
for(i in 1:N){
beta[i,1:2] ~ dmnorm(mu.beta[],R[,])
for(j in 1:T){
Y[i,j]~ dnorm(mu[i,j],tauC)
mu[i,j]← beta[i,1]+ beta[i,2]* x[j]
}
}
mu.beta[1:2]~ dmnorm(mean[],prec[,])
R[1:2,1:2]~ dwish(Omega[,],2)
tauC~ dgamma(0.001,0.001)
sigma←1/sqrt(tauC)
}
```

ed il relativo modello grafico è rappresentato in figura 4.4.

Innanzitutto osserviamo che nell'analisi precedente dei topi si era assunto che β_{1i} e β_{2i} fossero delle priori indipendenti e che la covarianza fosse centrata attorno alla media in modo da garantire che le probabilità di β_{1i} e β_{2i} fossero uguali. In questo caso, invece, non si è centrata la covarianza e si sono utilizzate delle priori normali multivariate per β_{1i} e β_{2i} in modo da aggiungere due forme addizionali di dipendenza. Procedendo nello stesso modo dell'esempio precedente, con lo stesso data-set e con valori iniziali diversi si sono ottenuti, dopo 10000 iterazioni, i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
mu.beta[1]	106.6	2.355	0.03929	102.0	106.6	111.3	1001	10000
mu.beta[2]	6.183	0.1077	0.001501	5.97	6.183	6.397	1001	10000
sigma	6.151	0.4735	0.008216	5.315	6.12	7.166	1001	10000

e la figura 4.5 mostra la convergenza dei parametri di interesse.

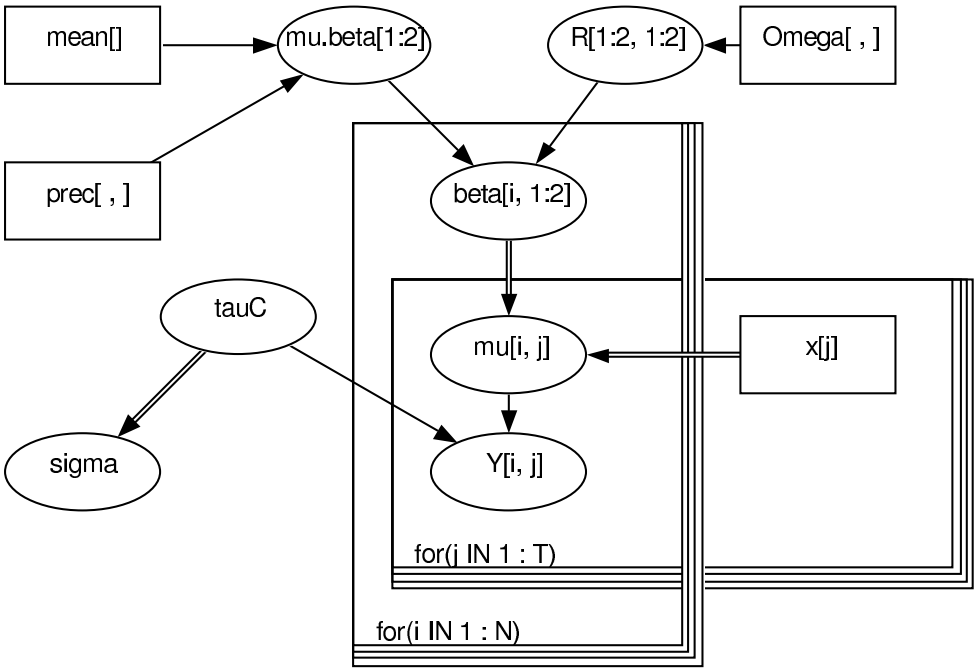

Figura 4.4: Modello grafico

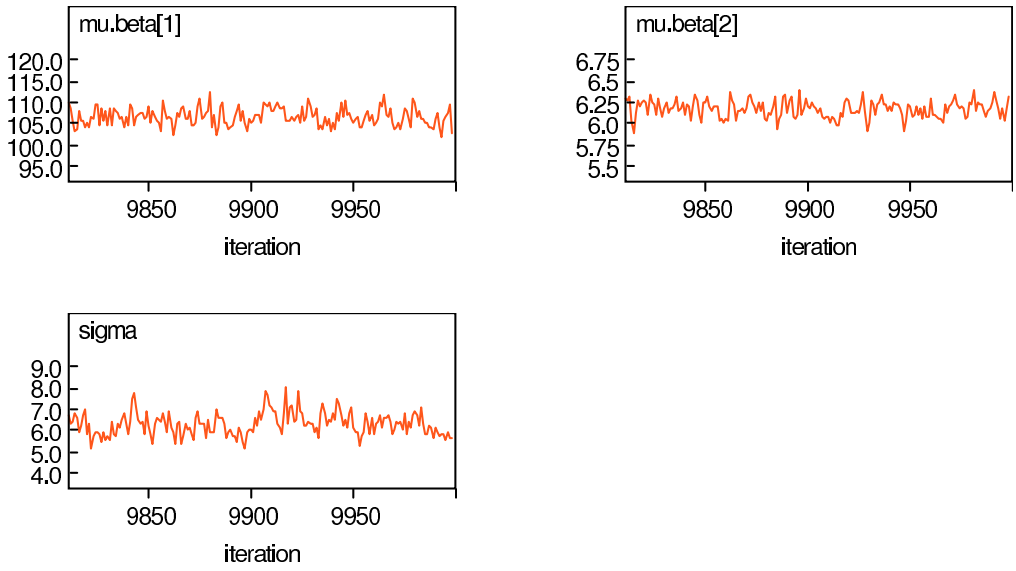

Figura 4.5: Convergenza dei parametri

Confrontando i risultati con quelli ottenuti nel caso precedente vediamo che:

Probabilità per β_{1i} e β_{2i}	Media a posteriori e s.d	Media a posteriori e sd
alpha0 $\equiv$ mu.beta[1]	106 — 3.6	106 — 2.3
mu.beta[2]$\equiv$ beta.c	6.18 — 0.11	6.18 — 0.11

l'assunzione di indipendenza non comporta una differenza nella stima dei parametri l'unica cosa che cambia è la loro precisione che nel secondo caso aumenta.

Capitolo 5

Risoluzione del problema mediante Winbugs

5.1 Primo esperimento

Riconsideriamo i dati del primo esperimento, in questo caso per spiegare MODPNIF si è assunto un modello di regressione lineare.
Specificatamente:

$$\text{MODPNIF} \sim \text{Normal}(\mu_i, \tau)$$
$$\mu_i = \beta_0 + \beta_1 * \text{HEIGHT[i]} + \beta_2 * \text{AGE[i]} + \beta_3 * \text{SEX[i]}$$

dove β_1, β_2, β_3 sono delle priori indipendenti "non-informative"di tipo normale mentre τ è una priori "non-informativa" di tipo gamma.
Quindi il modello nel linguaggio Winbugs diventa:
model
{for(i in 1:136)
{MODPNIF[i]∼ dnorm(mu[i],tau)
mu[i]← beta0+beta1*HEIGHT[i]+beta2*AGE[i]+beta3*SEX[i]
}

beta0∼ dnorm(0,0.00001)
beta1∼ dnorm(0,0.00001)
beta2∼ dnorm(0,0.00001)
beta3∼ dnorm(0,0.00001)
tau∼ dgamma(1.0E-3,1.0E-3)
sigma←sqrt(1/tau)
}
ed il relativo modello grafico è rappresentato in figura 5.1.

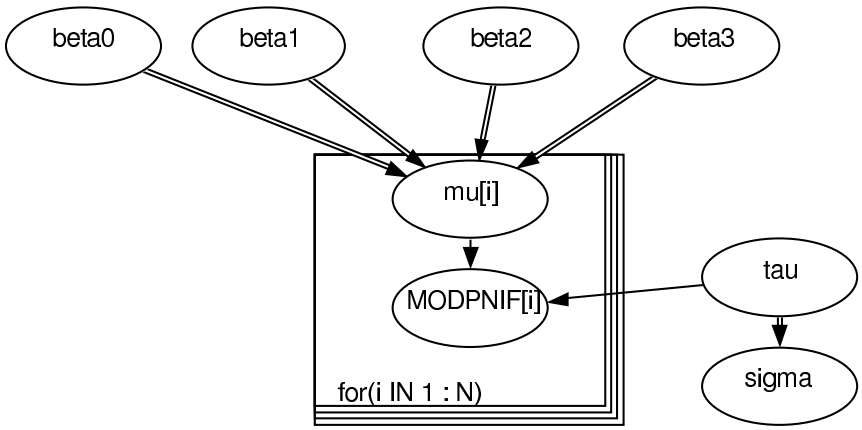

Figura 5.1: Grafico del modello

Procedendo allo stesso modo di quanto descritto nel capitolo 4, e considerando due catene aventi valori iniziali diversi, si ottengono, dopo 10000 iterazioni,

i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	7.46	3.237	0.0228	1.122	7.455	13.84	1	20000
beta1	0.03421	0.01889	1.34E-4	-0.002919	0.0342	0.07105	1	20000
beta2	-0.0422	0.007689	5.329E-5	-0.05729	-0.0422	-0.02704	1	20000
beta3	0.6513	0.3401	0.002427	-0.01625	0.65	1.318	1	20000
tau	0.4594	0.05669	4.056E-4	0.355	0.457	0.5763	1	20000

La figura 5.2 mostra come le due catene pur partendo da valori iniziali diversi convergono allo stesso valore.

Successivamente si è provato a cambiare i valori delle priori per vedere se

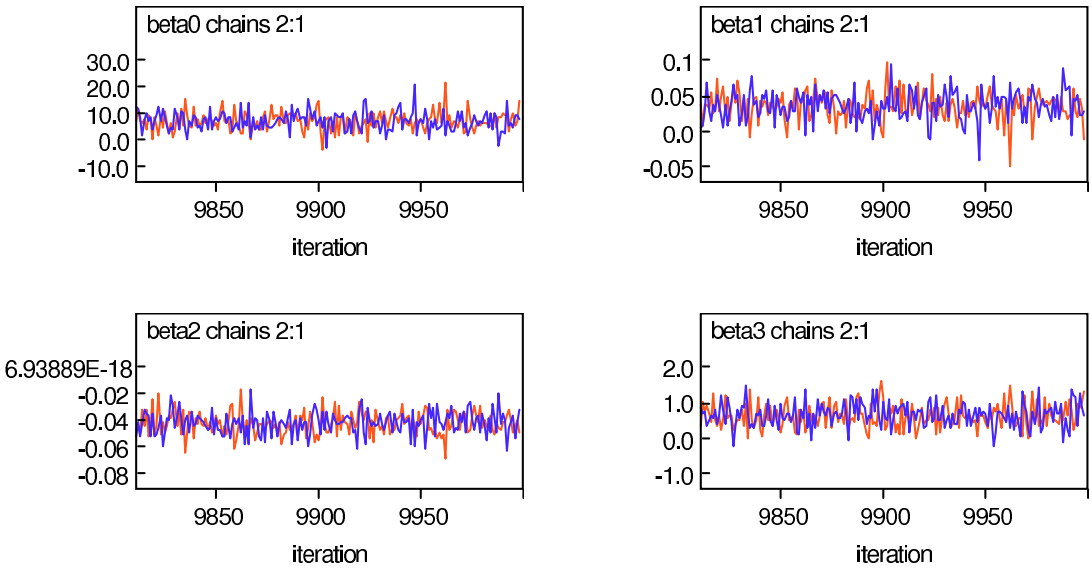

Figura 5.2: Convergenza delle catene

i risultati ottenuti precedentemente erano influenzati da queste assunzioni oppure no.

Si è quindi cosiderato il seguente modello:
model
{ for(i in 1:136)
{MODPNIF[i]∼ dnorm(mu[i],tau)
mu[i]← beta0+beta1*HEIGHT[i]+beta2*AGE[i]+beta3*SEX[i]
}
beta0∼ dnorm(0,0.001)
beta1∼ dnorm(0,0.0001)
beta2∼ dnorm(0,0.00001)
beta3∼ dnorm(0,0.001)
tau∼ dgamma(1.0E-3,1.0E-3)
sigma←sqrt(1/tau) }
e riutilizzando gli stessi valori iniziali del modello precedente si sono ottenuti, dopo 10000 iterazioni, i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	7.302	3.18	0.0669	1.228	7.271	13.54	1	20000
beta1	0.03513	0.01855	3.899E-4	-7.376E-4	0.03534	0.07053	1	20000
beta2	-0.04204	0.007729	1.579E-4	-0.05693	-0.04228	-0.02682	1	20000
beta3	0.6501	0.3439	0.00777	-0.01643	0.6476	1.334	1	20000
tau	0.4592	0.05758	0.001349	0.3563	0.4573	0.5759	1	20000

che sono simili a quelli ottenuti precedentemente, questo significa che il modello utilizzato non dipende dal valore dei parametri delle priori. Inoltre se confrontiamo le stime ottenute con R e con Winbugs :

Coefficienti	Stima con R	Stima con Winbugs
beta0	7.23	7.30
beta1	0.03	0.03
beta2	-0.04	-0.04
beta3	0.66	0.65

vediamo che i risultati sono molto simili, perciò il modello che si è utilizzato è un buon modello per stimare i nostri dati.

5.2 Secondo esperimento

Riconsideriamo ora i dati del secondo esperimento, in questo caso come si è già visto precedentemente la variabile risposta sarebbe (MODPNIF,MODPEF) ma supponiamo inizialmente di considerare MODPEF come variabile esplicativa e di avere quindi una variabile risposta unidimensionale. Anche in questo caso per spiegare MODPNIF si è assunto un modello di regressione lineare, del seguente tipo:

$$\text{MODPNIF} \sim \text{Normal}(\mu_i, \tau)$$
$$\mu_i = \beta_0 + \beta_1 * \text{HEIGHT[i]} + \beta_2 * \text{AGE[i]} + \beta_3 * \text{SEX[i]} + \beta_4 * \text{MODPEF[i]}$$

dove β_1, β_2, β_3 e β_4 sono delle priori indipendenti "non-informative"di tipo normale mentre τ è una priori "non-informativa" di tipo gamma.

Quindi il modello nel linguaggio Winbugs diventa:

```
model
{for(i in 1:71)
{MODPNIF[i]~dnorm(mu[i],tau)
mu[i]<-beta0+beta1*HEIGHT[i]+beta2*AGE[i]+beta3*SEX[i]+beta4*MODPEF[i]
}
beta0~dnorm(0,0.00001)
beta1~dnorm(0,0.00001)
beta2~dnorm(0,0.00001)
beta3~dnorm(0,0.00001)
beta4~dnorm(0,0.00001)
tau~dgamma(1.0E-3,1.0E-3)
sigma<-sqrt(1/tau) }
```

ed il relativo modello grafico è rappresentato in figura 5.3.

Procedendo allo stesso modo di quanto descritto nel capitolo 4, e considerando due catene aventi valori iniziali diversi si sono ottenuti, dopo 10000 iterazioni, i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	9.857	4.536	0.0319	0.8961	9.86	18.79	1	20000
beta1	-0.02681	0.02725	1.914E-4	-0.08039	-0.02687	0.02671	1	20000
beta2	-6.206E-4	0.01505	1.051E-4	-0.03027	-5.352E-4	0.02886	1	20000
beta3	1.437	0.5475	0.003825	0.3535	1.433	2.531	1	20000
beta4	0.2462	0.08749	6.373E-4	0.07483	0.2463	0.4181	1	20000
tau	0.4208	0.07387	5.445E-4	0.29	0.4162	0.5798	1	20000

La figura 5.4 mostra come le due catene pur partendo da valori iniziali diversi convergono allo stesso valore.
Successivamente si è provato a cambiare i valori delle priori per vedere se i risultati ottenuti precedentemente erano influenzati da queste scelte oppure no.
Si è quindi cosiderato il seguente modello:

```
model
{for(i in 1:71)
{MODPNIF[i]∼dnorm(mu[i],tau)
mu[i]←beta0+beta1*HEIGHT[i]+beta2*AGE[i]+beta3*SEX[i]+beta4*MODPEF[i]
}
beta0∼dnorm(0,0.01)
beta1∼dnorm(0,0.0001)
beta2∼dnorm(0,0.01)
beta3∼dnorm(0,0.001)
beta4∼dnorm(0,0.0001)
tau∼dgamma(1.0E-3,1.0E-3)
sigma←sqrt(1/tau)
}
```

e riutilizzando gli stessi valori iniziali si sono ottenuti, dopo 10000 iterazioni, i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	8.167	4.121	0.0291	0.01986	8.196	16.19	1	20000
beta1	-0.01775	0.02528	1.775E-4	-0.06688	-0.01788	0.03222	1	20000
beta2	0.00209	0.01473	1.029E-4	-0.02695	0.002142	0.03089	1	20000
beta3	1.302	0.5255	0.00369	0.2581	1.3	2.34	1	20000
beta4	0.2524	0.08718	6.366E-4	0.0818	0.2524	0.4241	1	20000
tau	0.4211	0.07383	5.453E-4	0.2906	0.4167	0.5807	1	20000

che sono simili a quelli ottenuti precedentemente, questo significa che il modello utilizzato non dipende dal valore dei parametri delle priori.
Inoltre confrontando le stime ottenute con R e quelle ottenute con Winbugs:

Coefficienti	Stime con R	Stime con Winbugs
beta0	9.79	9.85
beta1	-0.03(Non significativa)	-0.03
beta2	-0.0004(Non significativa)	-6.206E-4
beta3	1.43	1.43
beta4	0.25	0.25

vediamo che i risultati sono molto simili, perciò il modello che si è utilizzato è un buon modello per stimare i nostri dati.
Successivamente si è costruito un modello che considera PEF come variabile risposta. In questo caso il problema che si dovrà risolvere è del seguente tipo:

$$\text{MODPNIF} \sim \text{Normal}(\mu_{1i}, \tau_1)$$
$$\text{MODPEF} \sim \text{Normal}(\mu_{2i}, \tau_2)$$
$$\mu_{1i} = \beta_0 + \beta_1 * \text{HEIGHT[i]} + \beta_2 * \text{AGE[i]} + \beta_3 * \text{SEX[i]} + \delta[\text{i}]$$
$$\mu_{2i} = \beta_4 + \beta_5 * \text{HEIGHT[i]} + \beta_6 * \text{AGE[i]} + \beta_7 * \text{SEX[i]} + \delta[\text{i}]$$

doveβ_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7 sono delle priori "non-informative" di tipo normale mentre τ_1 e τ_2 sono delle priori "non-informative" di tipo gamma.

Nelle due equazioni inoltre si è aggiunto il fattore δ[i] in quanto le due variabili risposta sono in qualche modo legate tra loro e per esso si è supposto una priori di tipo normale standard.
In questa situazione il modello in linguaggio Winbugs diventa:

```
model {
for(i in 1:71)
{ MODPNIF[i]∼dnorm(mu1[i],tau1)
MODPEF[i]∼dnorm(mu2[i],tau2)
mu1[i]←beta0+beta1*HEIGHT[i]+beta2*AGE[i]+beta3*SEX[i]+delta[i]
mu2[i]←beta4+beta5*HEIGHT[i]+beta6*AGE[i]+beta7*SEX[i]+delta[i]
delta[i]∼dnorm(0,1)
}
beta0∼dnorm(0,0.00001)
beta1∼dnorm(0,0.00001)
beta2∼dnorm(0,0.00001)
beta3∼dnorm(0,0.00001)
beta4∼dnorm(0,0.00001)
beta5∼dnorm(0,0.00001)
beta6∼dnorm(0,0.00001)
beta7∼dnorm(0,0.00001)
tau1∼dgamma(1,0.01)
tau2∼dgamma(1,0.01)
sigma1←sqrt(1/tau1)
sigma2←sqrt(1/tau2)
}
```

ed il relativo modello grafico è rappresentato in figura 5.5.

Procedendo allo stesso modo di quanto descritto nel capitolo 4, e considerando due catene aventi valori iniziali diversi si sono ottenuti, dopo 10000 iterazioni, i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	sample
beta0	12.43	4.546	0.05127	3.467	12.45	21.32	20000
beta1	-0.008325	0.02691	3.06E-4	-0.06118	-0.008342	0.04475	20000
beta2	-0.01564	0.01424	1.475E-4	-0.04355	-0.01562	0.01257	20000
beta3	1.979	0.524	0.005718	0.9571	1.981	3.009	20000
beta4	10.27	6.093	0.05957	-1.652	10.24	22.21	20000
beta5	0.07635	0.03605	3.568E-4	0.005898	0.0766	0.1474	20000
beta6	-0.06148	0.01931	1.68E-4	-0.09915	-0.06152	-0.02353	20000
beta7	2.174	0.7049	0.006869	0.8046	2.17	3.548	20000
delta[1]	0.4608	0.7554	0.006547	-1.012	0.4628	1.949	20000
delta[2]	0.2875	0.7304	0.005368	-1.142	0.2853	1.722	20000
delta[3]	1.252	0.7473	0.005964	-0.2363	1.264	2.696	20000
sigma1	1.214	0.1631	0.002214	0.9173	1.206	1.555	20000
sigma2	1.897	0.1924	0.001953	1.553	1.885	2.309	20000

La figura 5.6 mostra come le due catene pur partendo da valori iniziali diversi convergono allo stesso valore.

In questo caso, però, non possiamo confrontare i risultati ottenuti con quelli ottenuti tramite R in quanto in R è difficile scrivere modelli con variabile risposta bivariata.

5.3 Dati del primo e del secondo esperimento

Supponiamo inizialmente che i due esperimenti siano tra loro confrontabili e costruiamo un modello che descriva questa situazione. Creiamo, quindi, un data-set che contiene i dati del primo e del secondo esperimento,senza però considerare la variabile PEF, e costruiamo una variabile indicatrice ESP che indica se i dati fanno parte del primo o del secondo esperimento.
In questa situazione il modello in linguaggio Winbugs diventa:
model
{for(i in 1:208)
{MODPNIF[i]∼dnorm(mu[i],tau)
mu[i]←beta0+beta1*HEIGHT[i]+beta2*AGE[i]+beta3*SEX[i]+beta4*ESP[i]
}
beta0∼dnorm(0,0.00001)
beta1∼dnorm(0,0.00001)
beta2∼dnorm(0,0.00001)
beta3∼dnorm(0,0.00001)
beta4∼dnorm(0,0.00001)
tau∼dgamma(1.0E-3,1.0E-3)
sigma←sqrt(1/tau)
}
ed il relativo modello grafico è rappresentato in figura 5.7.
Considerando due catene aventi valori iniziali diversi si sono ottenuti, dopo 10000 iterazioni, i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	8.946	2.722	0.0194	3.565	8.954	14.24	1	20000
beta1	0.02334	0.01591	1.142E-4	-0.007623	0.02327	0.05458	1	20000
beta2	-0.03853	0.00689	4.901E-5	-0.05197	-0.03854	-0.02504	1	20000
beta3	1.016	0.2945	0.002175	0.4365	1.015	1.597	1	20000
beta4	-0.4776	0.2306	0.001541	-0.9326	-0.4757	-0.02525	1	20000

La figura 5.8 mostra come le due catene pur partendo da valori iniziali diversi convergono allo stesso valore.
Inoltre se confrontiamo le stime ottenute con R e con Winbugs :

Coefficienti	Stime con R	Stime con Winbugs
beta0	9.041551	8.946
beta1	0.022764(Non significativa)	0.02334
beta2	-0.038825	-0.03853
beta3	1.046004	1.016
beta4	-0.512713	-0.4776

vediamo che i risultati sono molto simili, perciò il modello che si è utilizzato è un buon modello per stimare i nostri dati.
Successivamente, come si è già fatto nel capitolo quattro, si è costruito un modello che contiene le interazioni più significative tra le variabili.
In questa situazione il modello in linguaggio Winbugs diventa:
model
{for(i in 1:208)
{MODPNIF[i]∼dnorm(mu[i],tau) mu[i]←beta0+beta1*HEIGHT[i]+beta2*AGE[i]+
+beta3*SEX[i]+beta4*ESP[i]+beta5*(HEIGHT[i]*AGE[i])+beta6*(AGE[i]*SEX[i])+
+beta7*(AGE[i]*ESP[i])+beta8*(SEX[i]*ESP[i])
}
beta0∼dnorm(0,0.00001)
beta1∼dnorm(0,0.00001)

beta2∼dnorm(0,0.00001)
beta3∼dnorm(0,0.00001)
beta4∼dnorm(0,0.00001)
beta5∼dnorm(0,0.00001)
beta6∼dnorm(0,0.00001)
beta7∼dnorm(0,0.00001)
beta8∼dnorm(0,0.00001)
tau∼dgamma(1.0E-3,1.0E-3)
sigma←sqrt(1/tau)
}
ed il relativo modello grafico è rappresentato in figura 5.9.
Considerando due catene aventi valori iniziali diversi si ottengono, dopo 10000 iterazioni, i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	start	sample
beta0	20.97	5.543	0.04113	10.01	20.99	31.73	1	20000
beta1	-0.05237	0.03434	2.548E-4	-0.1189	-0.05258	0.01569	1	20000
beta2	-0.3299	0.1233	9.162E-4	-0.5719	-0.3301	-0.0877	1	20000
beta3	2.516	0.6875	0.00485	1.164	2.519	3.867	1	20000
beta4	-2.115	0.6185	0.004477	-3.342	-2.108	-0.9061	1	20000
beta5	0.001876	7.849E-4	5.833E-6	3.31E-4	0.00188	0.003414	1	20000
beta6	-0.04141	0.01489	1.051E-4	-0.07064	-0.04141	-0.01198	1	20000
beta7	0.03327	0.01427	1.036E-4	0.005361	0.03312	0.06154	1	20000
beta8	0.6436	0.4466	0.003143	-0.2191	0.6398	1.524	1	20000

La figura 5.10 mostra come le due catene pur partendo da valori iniziali diversi convergono allo stesso valore.
Inoltre se confrontiamo le stime ottenute con R e con Winbugs :

Coefficienti	Stime con R	Stime con Winbugs
beta0	20.905419	20.97
beta1	-0.051871	-0.05237
beta2	-0.325750	-0.3299
beta3	2.493080	2.516
beta4	-2.144432	-2.115
beta5	0.001846	0.001876
beta6	-0.040537	-0.04141
beta7	0.032539	0.03327
beta8	0.700889	0.6436

vediamo che i risultati sono molto simili, perciò il modello che si è utilizzato è un buon modello per stimare i nostri dati.

5.4 Dati del primo e del secondo esperimento senza l'ipotesi di confrontabilità tra i modelli

Finora abbiamo supposto che i due esperimenti fossero tra loro confrontabili, in realtà, come si è dimostrato nel capitolo quattro, i due esperimenti non sono confrontabili. Sulla base di questa considerazione si è quindi costruito un data-set che contiene i dati del primo e del secondo esperimento e che quindi contiene dati mancanti in quanto nel primo esperimento non ci sono valori per la variabile PEF.
Inoltre si è già sottolineato il fatto che PEF e PNIF hanno la stessa natura e quindi si dovrà risolvere un problema con variabile risposta bidimensionale del seguente tipo:

$$\text{MODPNIF} \sim \text{Normal}(\mu_{1i}, \tau_1)$$
$$\text{MODPEF} \sim \text{Normal}(\mu_{2i}, \tau_2)$$
$$\mu_{1i} = \beta_0 + \beta_1 * \text{HEIGHT[i]} + \beta_2 * \text{AGE[i]} + \beta_3 * \text{SEX[i]} + \delta\text{[i]}$$
$$\mu_{2i} = \beta_4 + \beta_5 * \text{HEIGHT[i]} + \beta_6 * \text{AGE[i]} + \beta_7 * \text{SEX[i]} + \delta\text{[i]}$$

doveβ_0, β_1, β_2, β_3, β_4, β_5, β_6, β_7 sono delle priori "non-informative" di tipo normale mentre τ_1 e τ_2 sono delle priori "non-informative" di tipo gamma.
Nelle due equazioni inoltre si è aggiunto il fattore δ[i] in quanto le due variabili risposta sono in qualche modo legate tra loro e per esso si è supposto una priori di tipo normale standard.
In questa situazione il modello in linguaggio Winbugs diventa:

```
    model {
for(i in 1:208)
{ MODPNIF[i]~dnorm(mu1[i],tau1)
MODPEF[i]~dnorm(mu2[i],tau2)
mu1[i]<-beta0+beta1*HEIGHT[i]+beta2*AGE[i]+beta3*SEX[i]+delta[i]
mu2[i]<-beta4+beta5*HEIGHT[i]+beta6*AGE[i]+beta7*SEX[i]+delta[i]
delta[i]~dnorm(0,1)
}
beta0~dnorm(0,0.00001)
beta1~dnorm(0,0.00001)
beta2~dnorm(0,0.00001)
beta3~dnorm(0,0.00001)
beta4~dnorm(0,0.00001)
beta5~dnorm(0,0.00001)
beta6~dnorm(0,0.00001)
beta7~dnorm(0,0.00001)
tau1~dgamma(1,0.01)
tau2~dgamma(1,0.01)
sigma1<-sqrt(1/tau1)
sigma2<-sqrt(1/tau2)
}
```

ed il relativo modello grafico è rappresentato in figura 5.11.

In questo caso visto che il data-set contiene dati mancanti per generare i valori iniziali delle catane si è utilizzata una funzione del programma che genera automaticamente i dati, e dopo 10000 iterazioni si sono ottenuti i seguenti risultati:

node	mean	sd	MC error	2.5%	median	97.5%	sample
beta0	8.561	2.695	0.0284	3.265	8.56	13.79	20000
beta1	0.02423	0.01578	1.668E-4	-0.006361	0.02419	0.05535	20000
beta2	-0.03696	0.006842	7.252E-5	-0.05038	-0.03693	-0.02371	20000
beta3	1.019	0.2899	0.00316	0.446	1.02	1.585	20000
beta4	8.663	5.987	0.0529	-2.966	8.648	20.52	20000
beta5	0.08979	0.03538	3.125E-4	0.01981	0.08984	0.1585	20000
beta6	-0.07006	0.01862	1.52E-4	-0.1065	-0.06997	-0.03319	20000
beta7	1.784	0.683	0.00619	0.4582	1.779	3.137	20000
delta[1]	-0.1366	0.773	0.005052	-1.67	-0.1355	1.383	20000
delta[2]	0.2769	0.7719	0.005693	-1.245	0.2824	1.772	20000
delta[3]	-0.482	0.7723	0.005869	-1.987	-0.486	1.025	20000
MODPEF[1]	24.42	2.112	0.01494	20.28	24.41	28.55	20000
MODPEF[2]	21.56	2.107	0.0145	17.47	21.55	25.69	20000
MODPEF[3]	21.0	2.11	0.01412	16.83	21.0	25.19	20000
...	...	...	...	...	...	...	...
MODPEF[135]	25.07	2.121	0.01713	20.86	25.06	29.3	20000
MODPEF[136]	21.66	2.095	0.01573	17.55	21.66	25.75	20000
MODPEF[137]	15.71	2.232	0.01646	11.33	15.71	20.07	20000

Come si può vedere dalla tabella, Winbugs produce le stime anche dei dati mancanti.

La figura 5.12 mostra come le due catene pur partendo da valori iniziali diversi convergono allo stesso valore.

In questo caso, però, non possiamo confrontare i risultati ottenuti con quelli ottenuti tramite R in quanto in R è difficile scrivere modelli quando la variabile risposta è bivariata.

Per verificare se il modello stimato è un buon modello abbiamo, quindi, costruito un grafico per vedere se c'è dipendenza tra MODPNIF e MOD-

PEF.
Analizzando il grafico, in figura 5.13, si vede che per ogni valore stimato di MODPEF esiste il corrispettivo valore di MODPNIF, questo significa che tra le due variabili esiste una certa dipendenza; quindi le assunzioni prese inizialmente sono verificate e il modello stimato è un buon modello per spiegare i nostri dati.

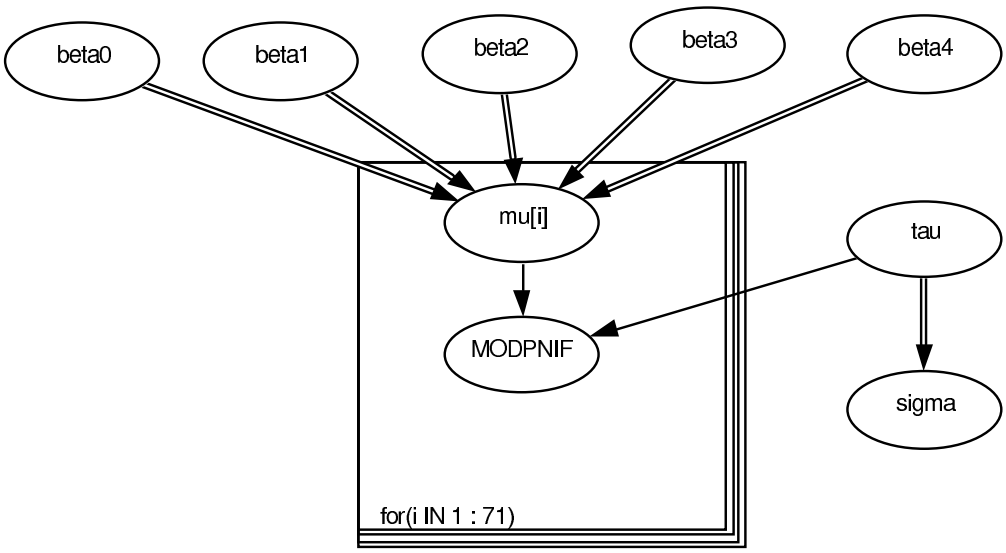

Figura 5.3: Grafico del modello

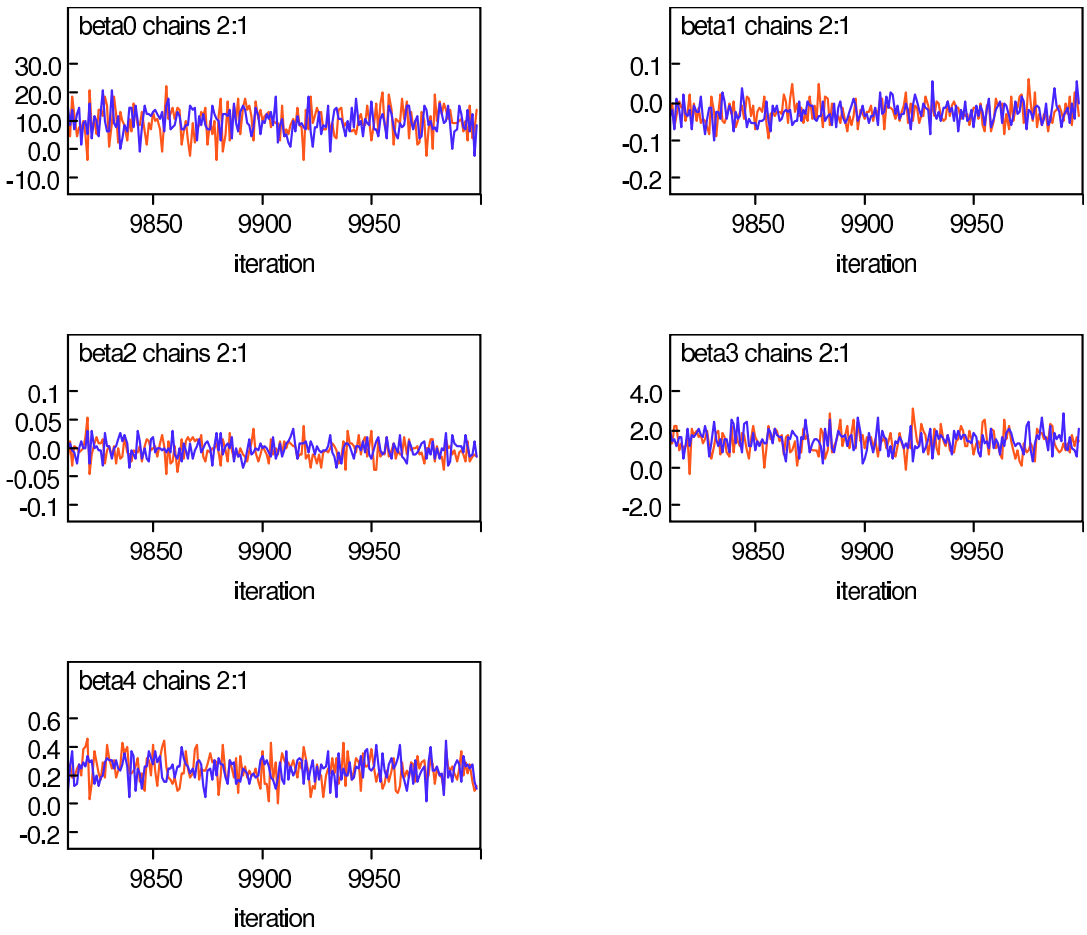

Figura 5.4: Convergenza delle catene

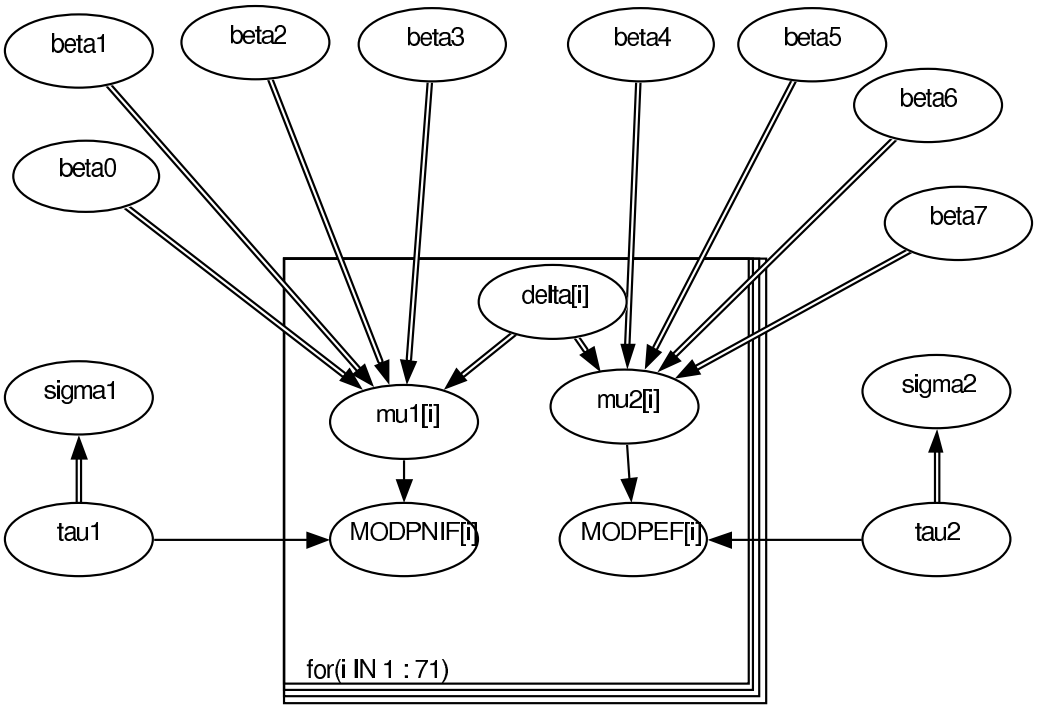

Figura 5.5: Grafico del modello

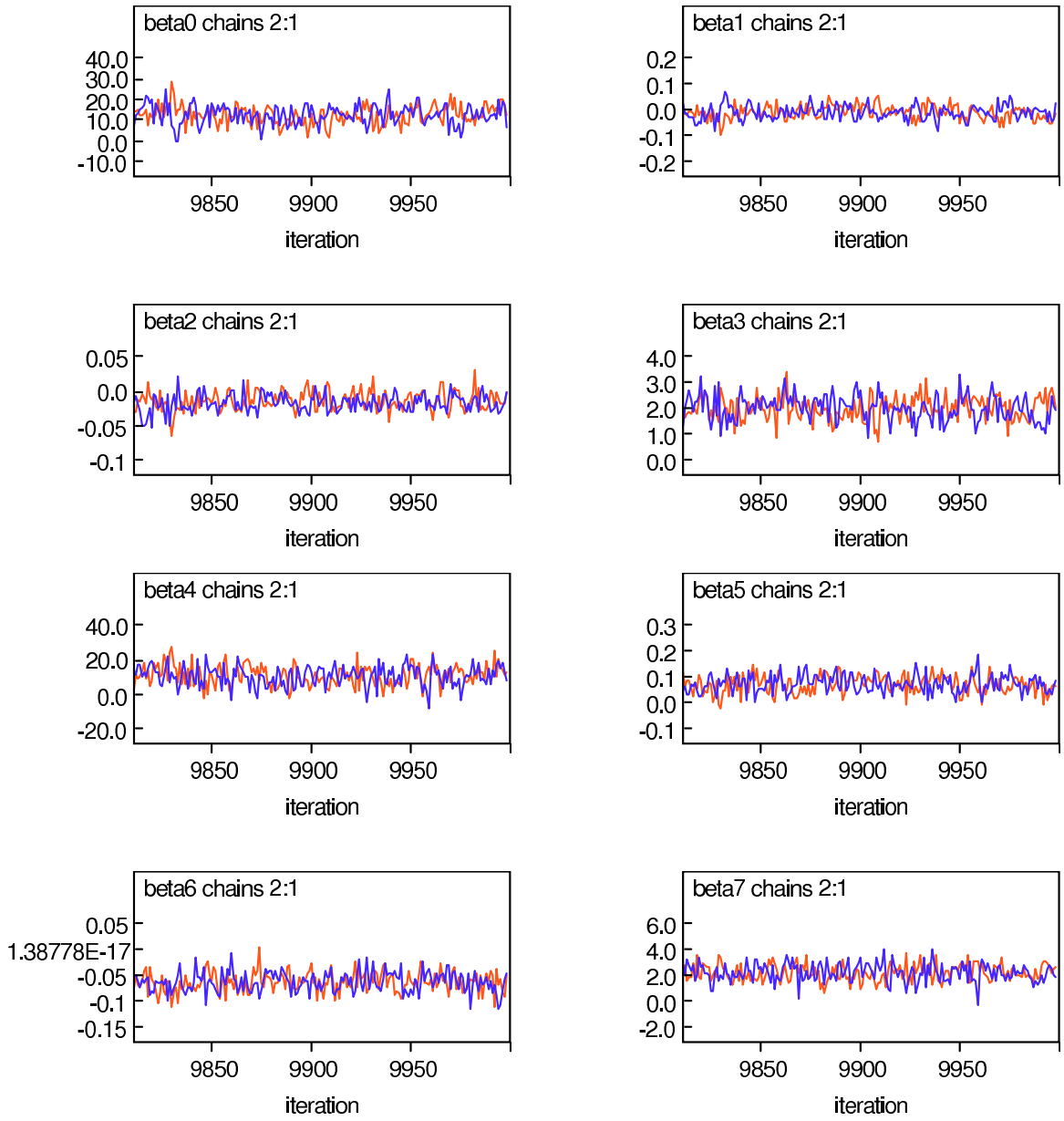

Figura 5.6: Convergenza delle catene

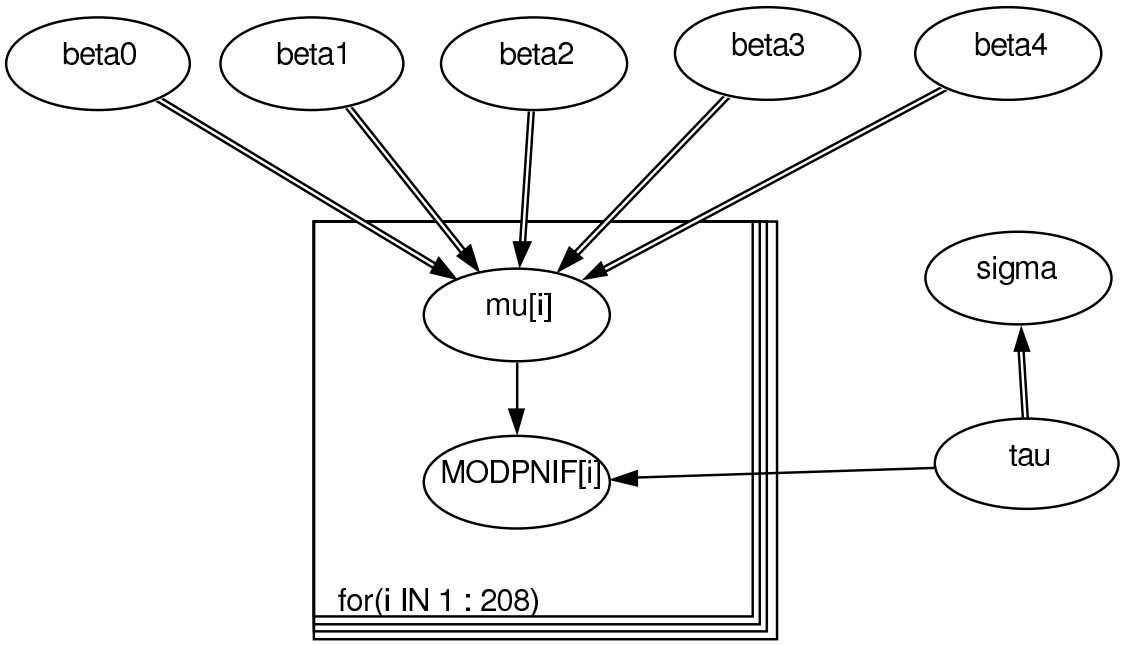

Figura 5.7: Grafico del modello

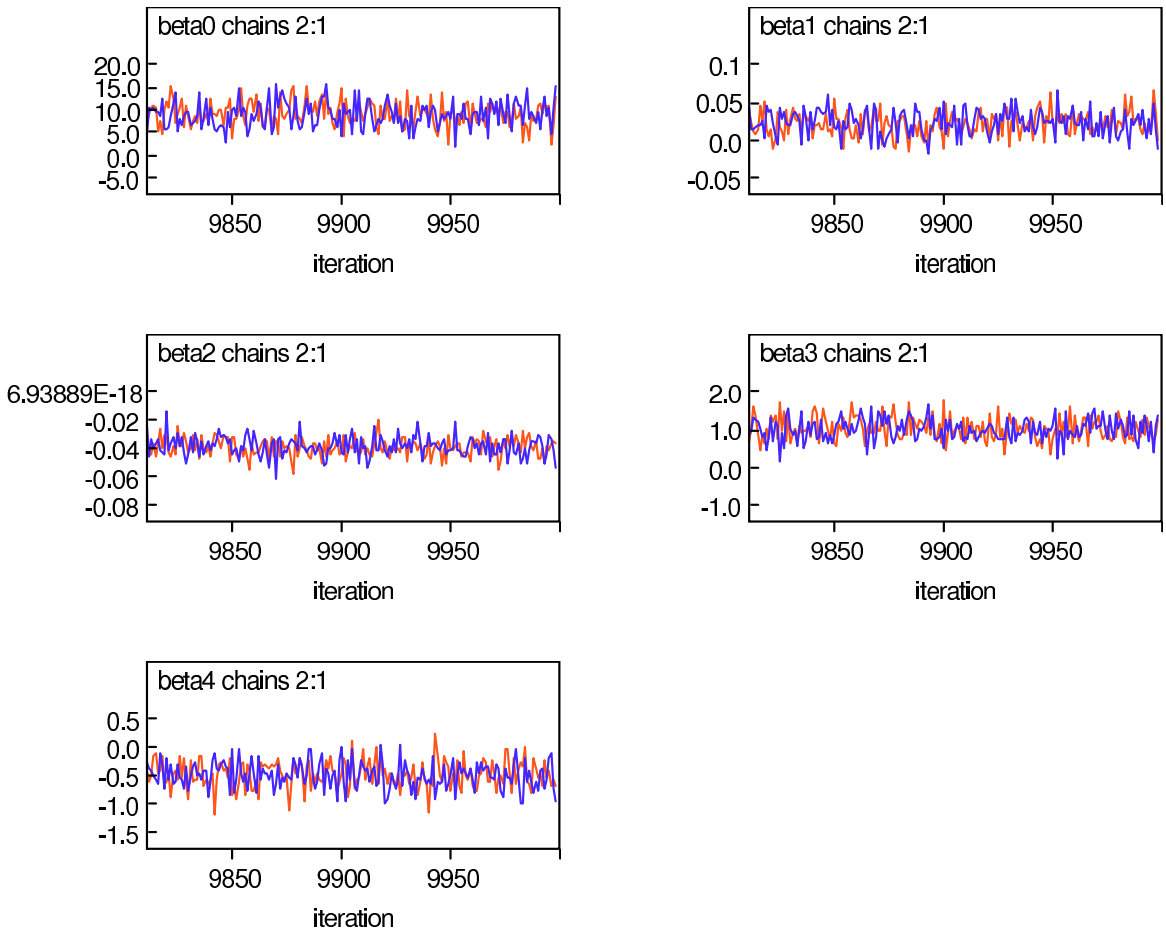

Figura 5.8: Convergenza delle catene

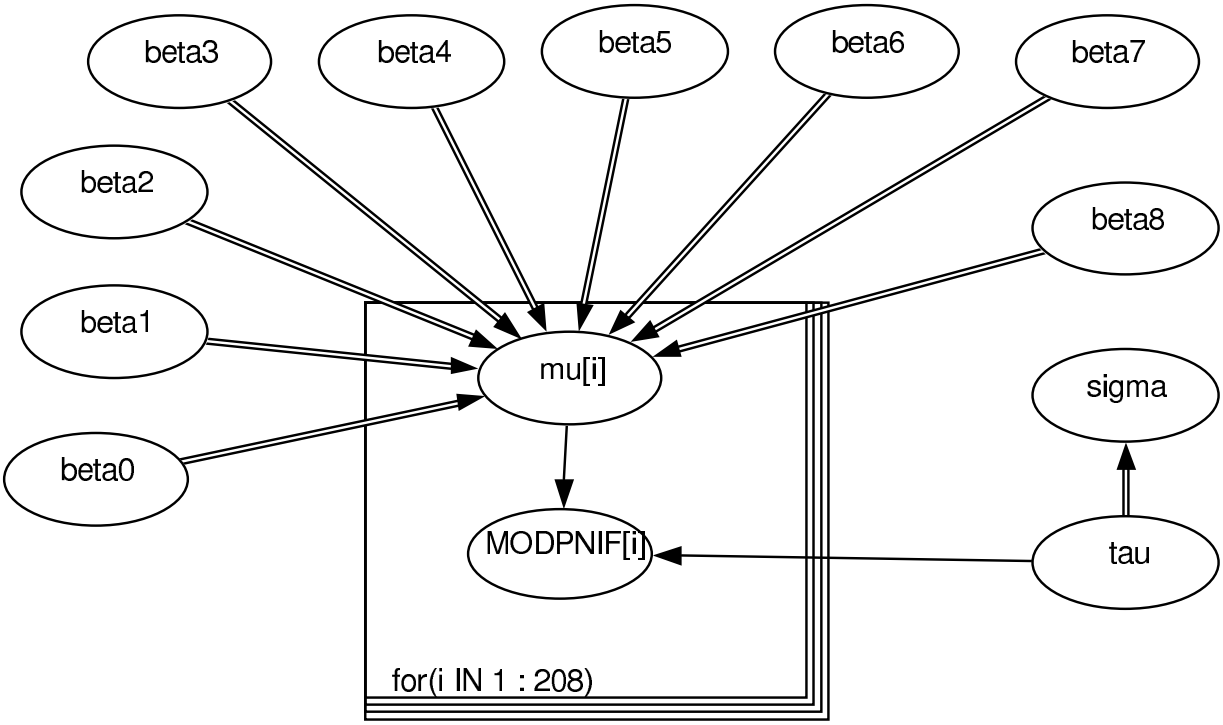

Figura 5.9: Grafico del modello

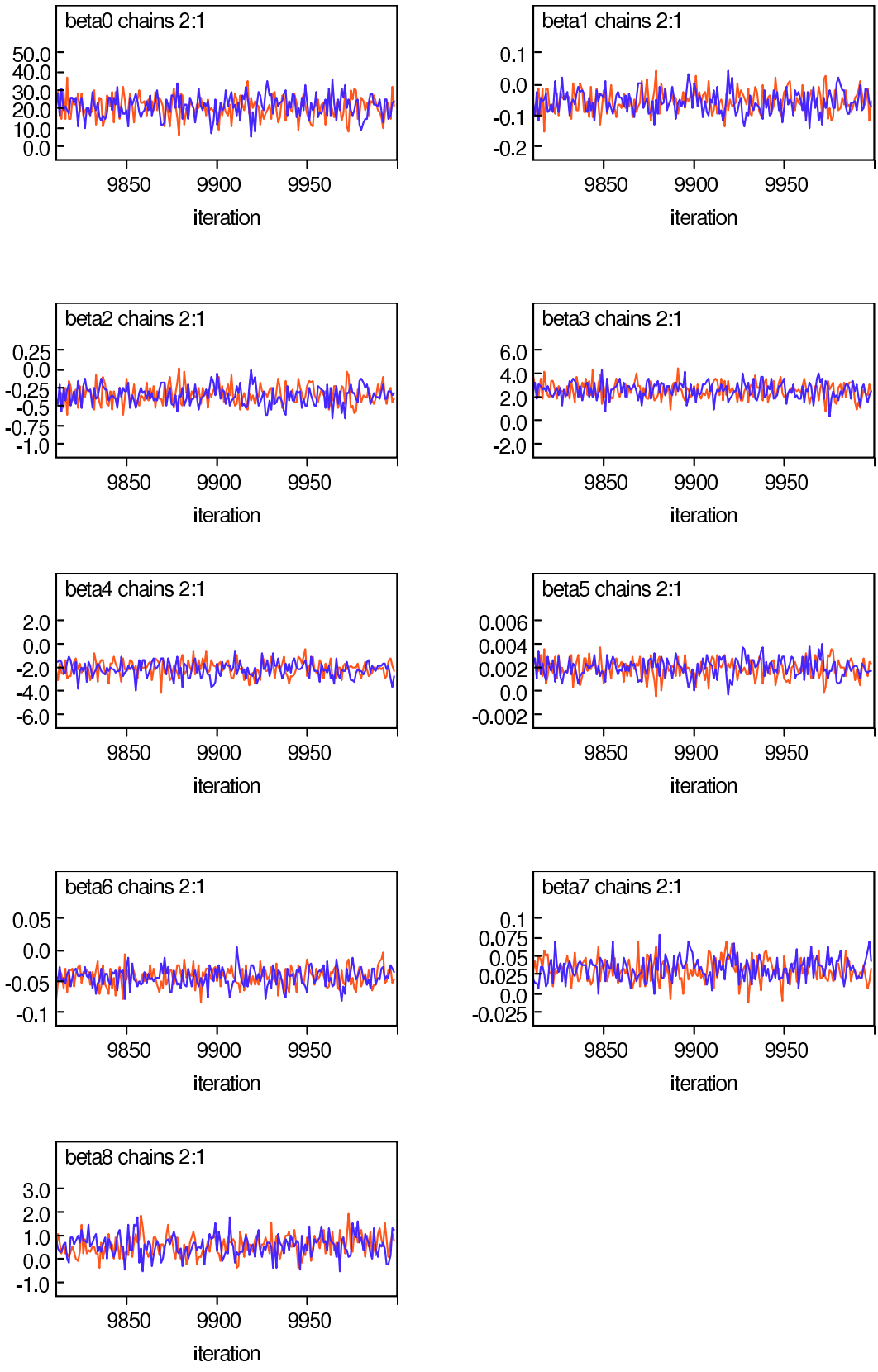

Figura 5.10: Convergenza delle catene

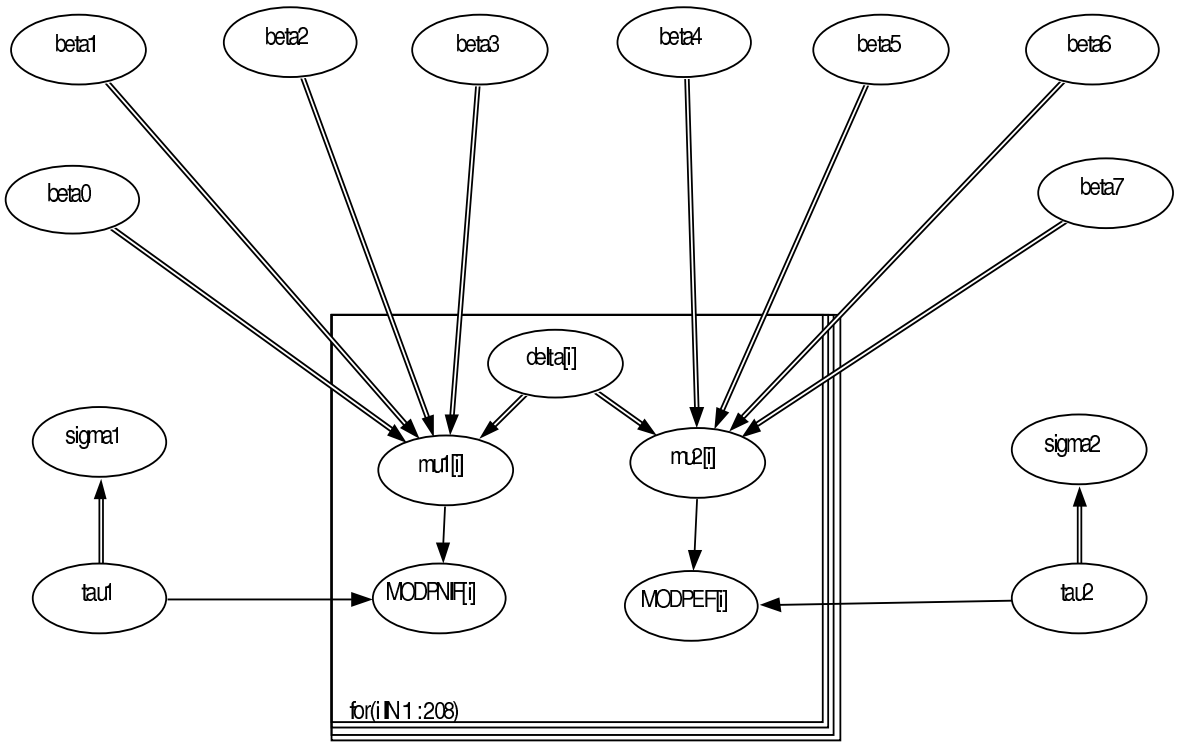

Figura 5.11: Grafico del modello

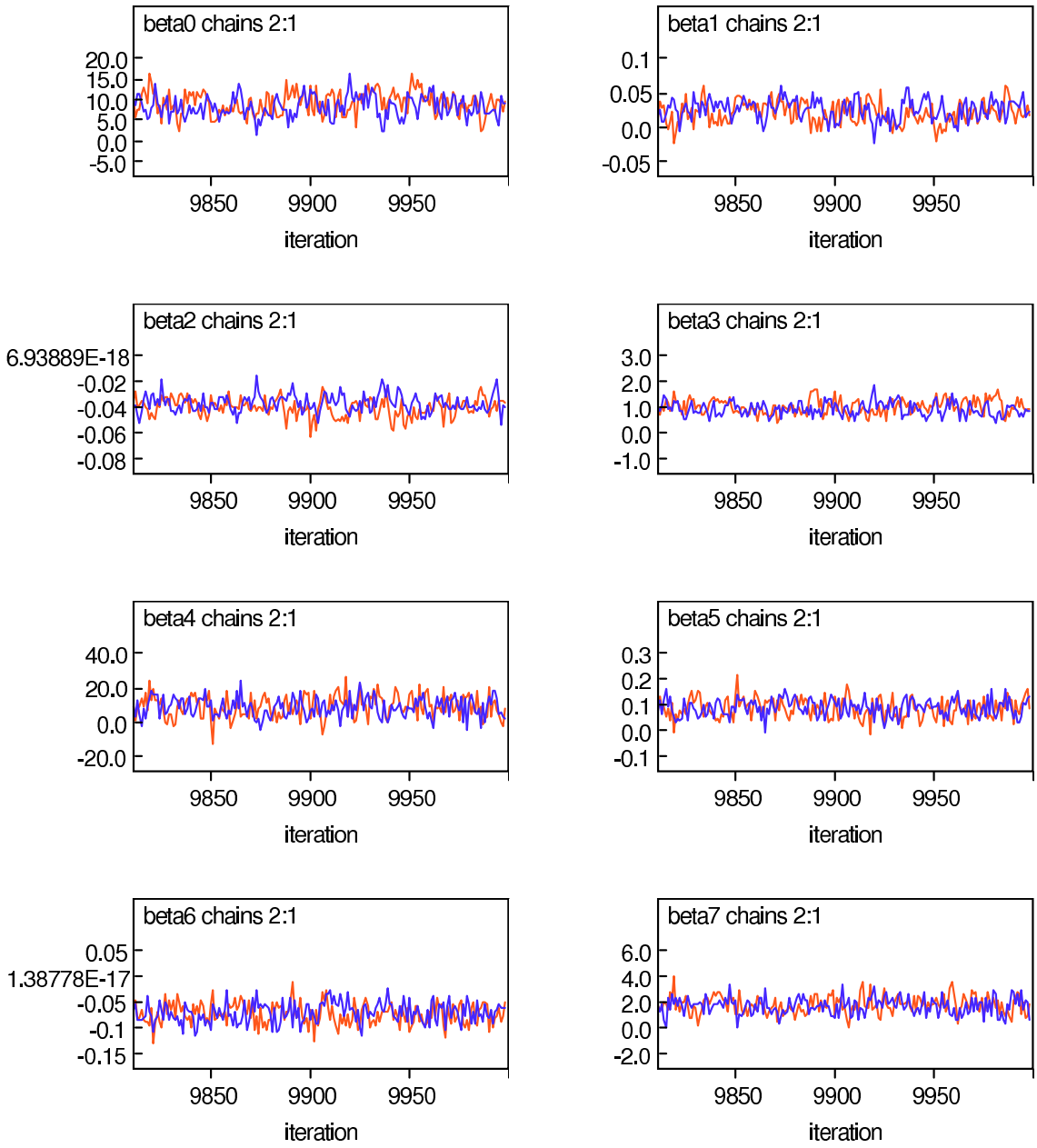

Figura 5.12: Convergenza delle catene

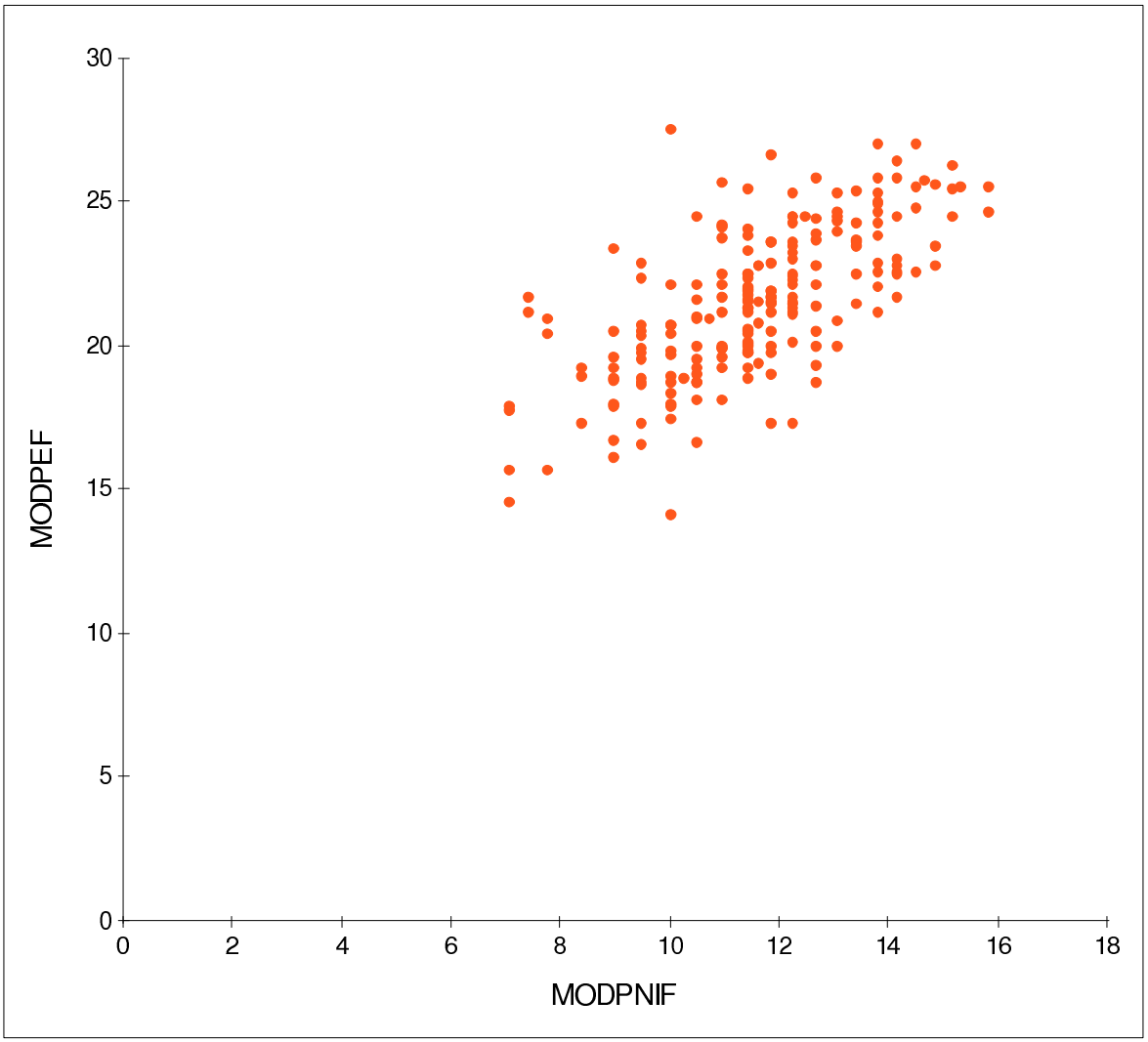

Figura 5.13: Grafico tra MODPNIF e MODPEF stimato

Capitolo 6

Conclusioni

L'obiettivo di questo studio era di vedere se esisteva una relazione tra i due esperimenti. Tramite questo studio, però, si è visto che non esiste nessuna relazione tra i modelli stimati infatti:

- nel primo esperimento il modello stimato è:
  MODPNIF=$\alpha + \beta$*HEIGHT+γ*AGE+τ*SEX+ε

- nel secondo esperimento il modello stimato è:
  MODPNIF=$\alpha_1 + \tau_1$*SEX+δ*MODPEF+ε_1

Utilizzando i modelli stimati si può, quindi, vedere che nel primo esperimento la media di PNIF con specificato sesso, età e altezza è visibile nel grafico 6.1 e nel grafico 6.2, mentre nel secondo esperimento la media di PNIF con specificato sesso, età, altezza e PEF è visibile nel grafico 6.3.

Lo scarso risultato di questo studio potrebbe essere causato dal modo in cui sono stati reperiti i dati nei due esperimenti e dal fatto che il numero di persone sottoposte al primo esperimento è maggiore rispetto a quelle sottoposte al secondo esperimento.
Per future ricerche potrebbe essere interessante:

- studiare il caso in cui oltre alle variabili esplicative già presenti si introduca l'effetto della variabile etnia.
- riesaminare il problema nel caso in cui le unità statistiche nel primo e nel secondo esperimento fossero le stesse.

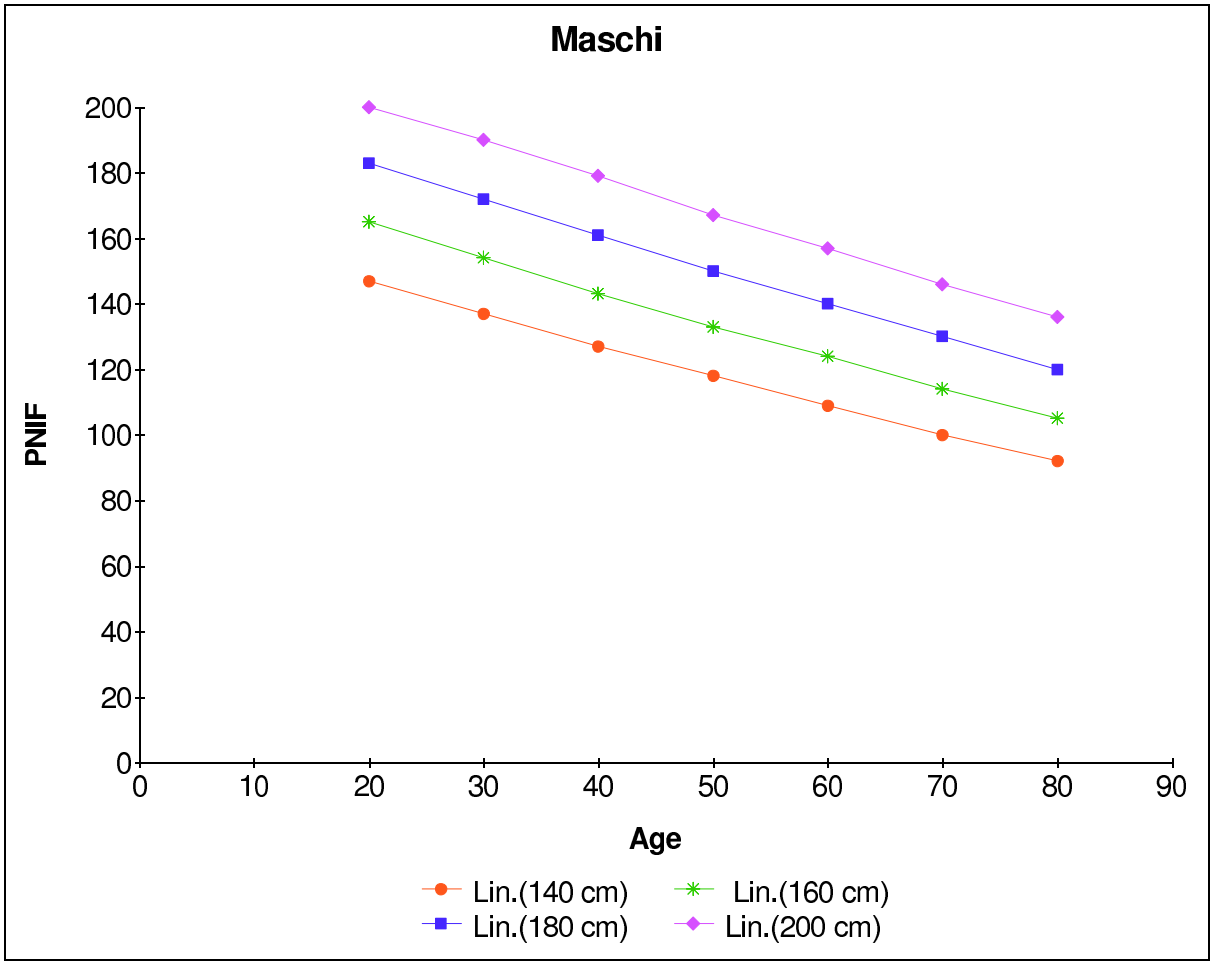

Figura 6.1: Grafico della stima di PNIF nei maschi

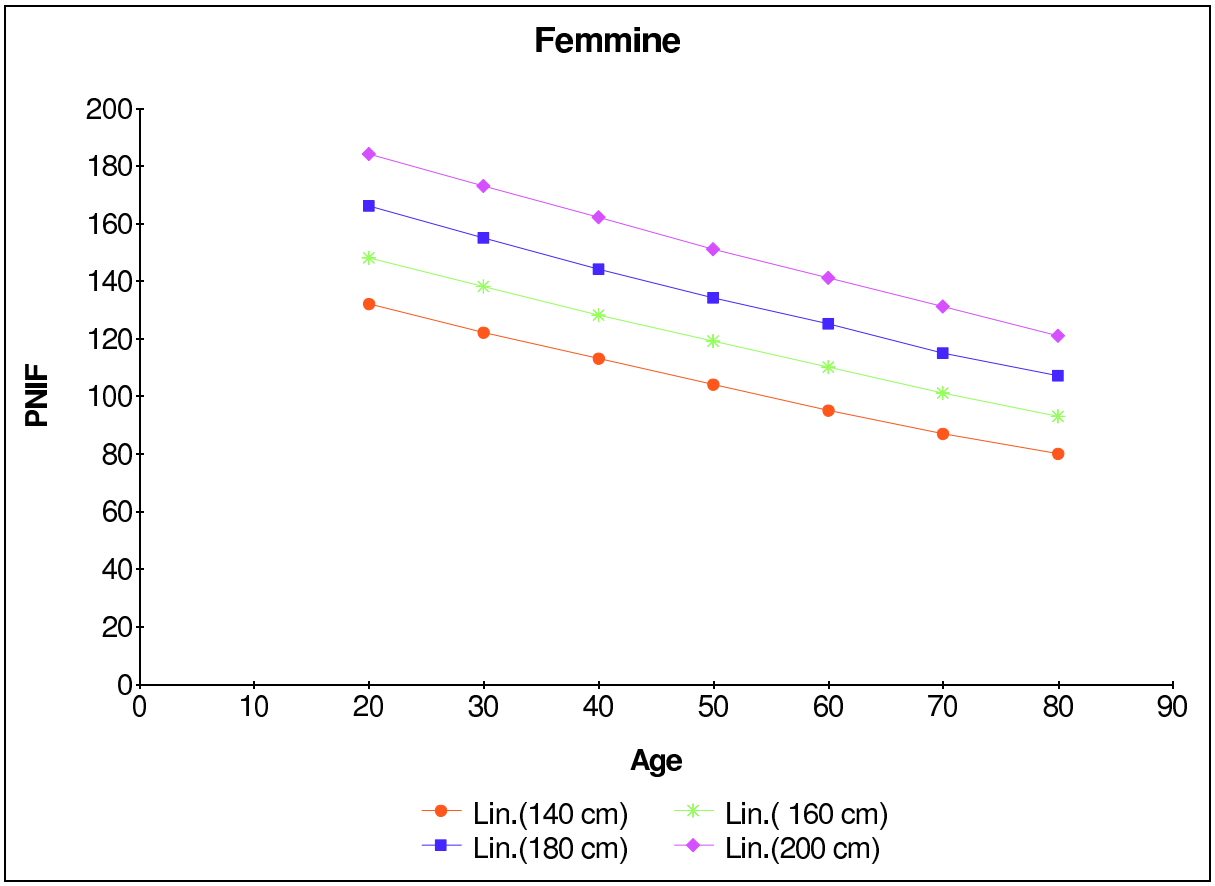

Figura 6.2: Grafico della stima di PNIF nelle femmine

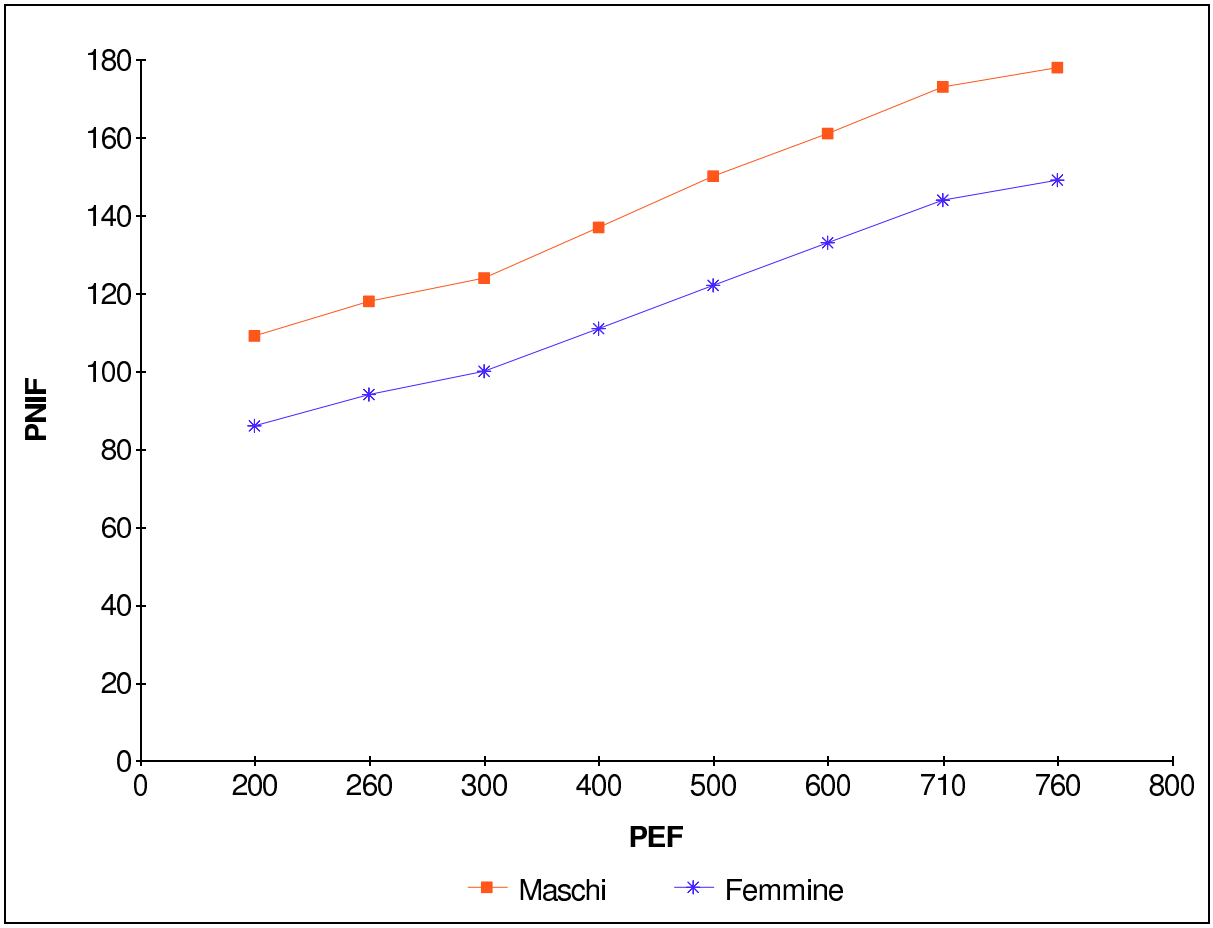

Figura 6.3: Grafico di PNIF nei maschi e nelle femmine

Bibliografia

[1] Liseo B.: *Introduzione alla statistica Bayesiana*, Dispensa utilizzata nel corso di Statistica cp, Anno Accademico 2004/2005.

[2] Stuart Coles: *Appunti del corso di Statistica Computazionale II*, Anno Accademico 2004/2005.

[3] *Manuale di Winbugs*, Versione 1.4, Gennaio 2003.

[4] German Dani: *Markov chain Monte Carlo: Stochastic simulation for Bayesian inference*, 1997.

[5] Giancarlo Ottaviano, Glenis K. Scadding, Valerie J. Lund: *Peak Nasal Inspiratory Floow. Normal Range in Adult Population.*