Università degli studi di Padova
Dipartimento di Scienze Statistiche
Corso di Laurea Magistrale in

Scienze Statistiche

Relazione Finale

Modello dose-risposta change-point bayesiano: applicazione ai dati sul covid-19

Relatore Prof.ssa Laura Ventura
Correlatore Prof. Paolo Girardi
Dipartimento di Scienze Statistiche

Laureando Lorenzo Bosetti
Matricola N 1211200

Anno Accademico 2020/2021

Indice

Elenco delle tabelle

Introduzione

I *modelli dose-risposta* (si veda Racugno e Ventura, 2017, Capitolo 10, come testo di riferimento) sono modelli utilizzati per descrivere l'effetto di una determinata sostanza su una variabile di interesse al variare della quantità o della concentrazione di tale sostanza. La variabile risposta, costituita da una misura dell'effetto della sostanza, dev'essere una variabile quantitativa ben definita, mentre la dose corrisponde alla quantità o alla concentrazione della sostanza ed è solitamente fissata dal disegno sperimentale. Questi modelli sono molto flessibili, poiché coinvolgono più parametri che consentono di cogliere al meglio anche un andamento non lineare dell'effetto della dose. Ideati per l'analisi dei dati in ambito biologico e tossicologico, i modelli dose-risposta possono trovare applicazioni anche in contesti diversi, come l'epidemiologia. Uno degli obiettivi di questa tesi è proprio quello di studiare l'applicazione di questi modelli ai dati riguardanti la pandemia di COVID-19; i conteggi cumulati giornalieri, ad esempio dei contagiati o dei deceduti, sono infatti ben descritti come variabile risposta di un modello che utilizza come dose il tempo: l'aumento della dose corrisponde quindi al passare dei giorni.
Lo schema della tesi è il seguente: nel Capitolo 1 vengono presentati i modelli dose-risposta, descrivendone la struttura e le basi teoriche su cui sono fondati. Vengono inoltre approfonditi gli aspetti riguardanti l'inferenza con approccio frequentista basati su funzioni di verosimiglianza composite (si veda Varin *et al.*, 2011), per poi mostrarne l'applicazione su due dataset di riferimento. Il Capitolo 2 è incentrato sull'approccio bayesiano ai modelli dose-risposta, che consente di includere nel modello eventuali conoscenza a priori sul fenomeno studiato; anche questo capitolo presenta gli aspetti salienti dell'inferenza e

si conclude con l'applicazione ai due dataset. Nel Capitolo 3 vengono discussi i *modelli change-point* (si veda Qiu, 2014, come testo di riferimento sui *change-point* e Greco *et al.*, 2021, *Submitted*, per una prima applicazione ai dati sul COVID-19), particolarmente utili nel contesto dei dati sul COVID-19 per analizzare il momento del passaggio da un'ondata a quella successiva. Infine, nel Capitolo 4 vengono applicati ai dati nazionali sul COVID-19 gli strumenti presentati in precedenza, con l'obiettivo di trarre conclusioni sulle caratteristiche delle diverse ondate e sul momento di passaggio da un'ondata all'altra. Al fine di validare la metodologia e i risultati ottenuti viene anche svolto uno studio di simulazione, che mostra come il modello stimato ben si adatti alla tipologia di dati in questione.
Mentre i primi due capitoli costituiscono una sorta di rassegna di quanto già presente in letteratura sulla teoria dei modelli dose-risposta, sia in ambito frequentista sia in ambito bayesiano, gli ultimi due capitoli rappresentano il cuore e la proposta innovativa di questa tesi: questo lavoro costituisce infatti un primo approccio bayesiano ai modelli dose-risposta con *change-point*, che trova un'immediata e rilevante applicazione ai dati sui deceduti giornalieri a causa del COVID-19. Un ulteriore aspetto di particolare interesse è la discussione sulle funzioni di verosimiglianza utilizzate per l'inferenza: a causa dell'autocorrelazione presente nei dati viene introdotto un particolare tipo di verosimiglianza composita, l'*independence likelihood*, ma viene proposto anche l'utilizzo di altre funzioni in grado di garantire robustezza al modello, come lo *score di Tsallis*; infine, vengono utilizzate la verosimiglianza profilo e la verosimiglianza profilo generalizzata all'interno di un approccio bayesiano ibrido, per dare maggior rilevanza al parametro ritenuto di maggior interesse.

Capitolo 1

I modelli dose-risposta

1.1 Introduzione

In un contesto in cui sono presenti p variabili esplicative $x_1, ..., x_p$ la soluzione più semplice per modellare una variabile risposta y è la regressione lineare multivariata. Il modello lineare multivariato (Grigoletto *et al.*, 2017) è della forma

$$y_i = x_i^T \beta + \varepsilon_i, \qquad i = 1, ..., n, \tag{1.1}$$

dove $\beta = (\beta_1, \beta_2, ..., \beta_p)$ è il vettore ignoto dei parametri di regressione e ε_i sono variabili aleatorie indipendenti ed identicamente distribuite a media nulla. Nel caso del modello di regressione lineare normale si assume $\varepsilon_i \sim N(0, \sigma^2)$, $i = 1, ..., n$, indipendenti.
In alcune applicazioni però la relazione tra le variabili esplicative e la variabile risposta non è di tipo lineare ed è quindi necessario fare ricorso a dei modelli più flessibili, come i modelli di regressione non lineare. Un modello di regressione non lineare si ottiene sostituendo il predittore lineare $x^T\beta$ della (1.1) con una funzione nota $\mu(x, \beta)$, detta *funzione media*. Il modello di regressione non lineare normale diventa quindi (Bates e Watts, 2007)

$$y_i = \mu(x_i, \beta) + \varepsilon_i, \tag{1.2}$$

dove β è un parametro ignoto p-dimensionale e $\varepsilon_i \sim N(0, \sigma^2)$ sono variabili aleatorie indipendenti ed identicamente distribuite, $i = 1, ..., n$.

Un esempio di relazione che può essere descritta da modelli di regressione non lineare è la relazione dose-risposta, su cui si basano i cosiddetti modelli dose-risposta. In questo tipo di modelli la variabile esplicativa (la dose) è la quantità o la concentrazione di una determinata sostanza che provoca una risposta biologica ben definita; la dose è una quantità non negativa ed è spesso assunta nota senza errori, poiché fissata dal disegno sperimentale. Talvolta viene utilizzato il tempo come dose, ad esempio negli studi sulla germinazione ma non solo. La variabile risposta è invece la misura dell'effetto della dose ed è quindi soggetta a variazione casuale. Nella maggior parte dei casi la risposta è una variabile continua, ma può essere pure una variabile dicotomica o di conteggio (Ritz e Van Der Vliet, 2009).
Nei modelli dose-risposta la funzione $\mu(x, \beta)$ più utilizzata è la log-logistica a cinque parametri, data da

$$\mu(x, \beta) = c + \frac{d - c}{(1 + \exp(b(\log(x) - \log(e))))^f}, \tag{1.3}$$

con $\beta = (b, c, d, e, f)$ parametri ignoti. Spesso la (1.3) viene semplificata nella versione a quattro parametri, che si ottiene ponendo $f = 1$, ossia

$$\mu(x, \beta) = c + \frac{d - c}{1 + \exp(b(\log(x) - \log(e)))}. \tag{1.4}$$

Il vantaggio della funzione log-logistica sta nell'interpretabilità dei parametri: c e d sono infatti gli asintoti, rispettivamente, inferiore e superiore della variabile risposta, b indica la pendenza della curva ed e, nella versione semplificata (1.4), corrisponde alla dose con la quale si ottiene un valore della risposta pari alla metà del valore massimo, chiamato ED50 e corrispondente al punto di flesso.

1.2 Inferenza

Utilizzando un approccio frequentista è possibile effettuare inferenza nell'ambito dei modelli dose-risposta basandosi sulla funzione di verosimiglianza. In generale, sia $\mathcal{F} = \{p_Y(y; \theta), \theta \in \Theta \subseteq \mathbb{R}^d\}$ un modello parametrico, dove θ è un parametro d-dimensionale che assume valori nello spazio parametrico $\Theta \subseteq \mathbb{R}^d$ e $p_Y(y; \theta)$ è la funzione di densità o di probabilità, che può essere

continua o discreta. Nel caso del modello dose-risposta con risposta normale e funzione media log-logistica a 4 parametri, $\theta = (b, c, d, e, \sigma^2) \in \Theta \subseteq \mathbb{R}^4 \times (0, +\infty)$.
La *funzione di verosimiglianza* per θ è definita da

$$L(\theta) = c(y)p_Y(y; \theta), \tag{1.5}$$

con $c(y)$ costante positiva dipendente dai dati ma non dai parametri. Spesso l'inferenza è basata sulla *funzione di log-verosimiglianza* $l(\theta)$ definita da

$$l(\theta) = \log L(\theta), \tag{1.6}$$

con $l(\theta) = -\infty$ se $L(\theta) = 0$. Se le osservazioni $y_1, ..., y_n$ sono realizzazioni di variabili aleatorie Y_i tra loro indipendenti e con densità marginali $p_{Y_i}(y_i; \theta)$, la funzione di verosimiglianza può essere scritta come

$$L(\theta) = \prod_{i=1}^{n} p_{Y_i}(y_i; \theta), \tag{1.7}$$

e la funzione di log-verosimiglianza diventa quindi

$$l(\theta) = \sum_{i=1}^{n} \log p_{Y_i}(y_i; \theta). \tag{1.8}$$

La stima di massima verosimiglianza di θ si ottiene massimizzando la funzione $L(\theta)$ o, in maniera equivalente, la funzione $l(\theta)$. La stima di massima verosimiglianza è quindi quel valore $\hat{\theta} \in \Theta$ tale che

$$L(\hat{\theta}) \geq L(\theta), \quad \forall \theta \in \Theta. \tag{1.9}$$

In un modello con verosimiglianza regolare (si veda Salvan *et al.*, 2020, capitolo 1, per la definizione) la stima di massima verosimiglianza è una soluzione dell'equazione di verosimiglianza

$$l_*(\theta) = \left(\frac{\partial l(\theta)}{\partial \theta_1}, ..., \frac{\partial l(\theta)}{\partial \theta_d} \right)^T = 0, \tag{1.10}$$

dove $l_*(\theta)$ è il vettore delle derivate prime parziali della funzione di log-verosimiglianza, detto *funzione di punteggio*. Ad eccezione di casi particolari,

l'equazione (1.10) non ammette soluzioni in forma esplicita; è quindi necessario ricorrere a metodi numerici per determinare la stima di massima verosimiglianza $\hat{\theta}$ (Salvan *et al.*, 2020).
Nel caso del modello non lineare normale con funzione media (1.4), la funzione di log-verosimiglianza è data da

$$l(\theta) = l(\overbrace{b, c, d, e}^{\beta}, \sigma^2) = -\frac{n}{2}\log\sigma^2 - \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i - \mu(x_i, \beta))^2}{\sigma^2}.$$

Le equazioni di verosimiglianza sono date da

$$\frac{\partial l(\theta)}{\partial \theta} = \begin{cases} \frac{\partial l(\beta,\sigma^2)}{\partial \beta} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu(x_i,\beta))\frac{\partial \mu(x_i,\beta)}{\partial \beta} = 0 \\ \frac{\partial l(\beta,\sigma^2)}{\partial \sigma^2} = -\frac{n}{2\sigma^2} + \frac{1}{2}\sum_{i=1}^{n}\frac{(y_i-\mu(x_i,\beta))^2}{(\sigma^2)^2} = 0 \end{cases}.$$

Dato che le equazioni di verosimiglianza per β non ammettono soluzione esplicita, la stima di massima verosimiglianza $\hat{\beta}$ deve essere ricavata numericamente tramite metodi computazionali iterativi. Al contrario, dall'ultima equazione è possibile ricavare in maniera esplicita la stima di massima verosimiglianza per σ^2, data da

$$\hat{\sigma}^2 = \sum_{i=1}^{n}\frac{(y_i - \mu(x_i, \hat{\beta}))^2}{n}.$$

Per ottenere una stima intervallare dei parametri o testarne la singola significatività è necessario introdurre la matrice di informazione osservata $j(\theta)$, definita come la matrice $d \times d$ delle derivate parziali seconde di $l(\theta)$ cambiate di segno, ossia

$$j(\theta) = -l_{**}(\theta) = -\frac{\partial^2 l(\theta)}{\partial\theta\partial\theta^T}. \tag{1.11}$$

Gli intervalli di confidenza e i test per verificare la significatività dei parametri si basano sui seguenti risultati asintotici, validi sotto l'ipotesi che θ sia il vero valore del parametro:

$$\hat{\theta} \dot{\sim} N_d(\theta, j(\hat{\theta})^{-1}) \tag{1.12}$$

$$W(\theta) = 2\{l(\hat{\theta}) - l(\theta)\} \dot{\sim} \chi^2_d, \tag{1.13}$$

dove $\dot{\sim}$ significa "è distribuito asintoticamente come", $N_d(\mu, \Sigma)$ indica la distribuzione normale d-variata con media μ e matrice di covarianza Σ e χ^2_d

indica la distribuzione chi-quadrato con d gradi di libertà.
Per un generico modello normale si può mostrare che β e σ^2 sono tra loro ortogonali e che quindi gli stimatori di massima verosimiglianza per β e σ^2 sono tra loro asintoticamente indipendenti. Ne consegue che, per effettuare procedure di inferenza su β, sia sufficiente calcolare il blocco $j_{\beta\beta}$ della matrice di informazione osservata. Dalla (1.12) si ottiene

$$\hat{\beta} \dot{\sim} N_4(\beta, j_{\beta\beta}(\hat{\beta}, \hat{\sigma}^2)^{-1}) \tag{1.14}$$

ed è quindi possibile calcolare gli intervalli di confidenza di livello approssimato $(1-\alpha)$ per i singoli elementi di β (intervalli alla Wald) come

$$\hat{\beta}_k \pm z_{(1-\alpha)/2}\sqrt{j_{\beta\beta}(\hat{\beta}, \hat{\sigma}^2)^{-1}_{kk}}, \tag{1.15}$$

dove $j_{\beta\beta}(\hat{\beta}, \hat{\sigma}^2)^{-1}_{kk}$ è l'elemento di posto (k,k) della matrice $j_{\beta\beta}(\hat{\beta}, \hat{\sigma}^2)^{-1}$ e $z_{(1-\alpha)/2}$ è il quantile di livello $(1-\alpha)/2$ della distribuzione normale standard. Dalla distribuzione approssimata dello stimatore di massima verosimiglianza (1.14) è possibile ricavare anche la statistica test per verificare la significatività dei singoli parametri, ovvero per verificare $H_0 : \beta_k = 0$ contro $H_1 : \beta_k \neq 0$, data da

$$z = \frac{\hat{\beta}_k}{\sqrt{j_{\beta\beta}(\hat{\beta}, \hat{\sigma}^2)^{-1}_{kk}}}. \tag{1.16}$$

Gli intervalli di confidenza basati sulla normalità asintotica dello stimatore di massima verosimiglianza hanno il vantaggio di essere semplici da calcolare, ma presentano alcuni difetti: ad esempio, possono escludere punti con verosimiglianza maggiore rispetto a punti che invece sono inclusi, oppure possono non essere compresi nello spazio parametrico. In alternativa, si possono costruire intervalli di confidenza utilizzando la quantità $W(\theta)$ definita in (1.13). Per il modello preso ad esempio, tale intervallo di confidenza di livello $1-\alpha$ per β è

$$\{\beta : W(\beta) < \chi^2_{p;1-\alpha}\} = \left\{\beta : l(\beta, \hat{\sigma}^2_\beta) > l(\hat{\beta}, \hat{\sigma}^2) - \frac{\chi^2_{p;1-\alpha}}{2}\right\}, \tag{1.17}$$

dove $\hat{\sigma}^2_\beta$ è la stima di massima verosimiglianza di σ^2 per un fissato β. È infine possibile confrontare due modelli annidati sempre tramite il test del rapporto

di verosimiglianza; se p_0 è il numero di parametri nel modello semplificato, ovvero sotto l'ipotesi nulla H_0, e p il numero di parametri nel modello completo, allora

$$W_P^{H_0} = 2\{l(\hat{\theta}) - l(\hat{\theta}_0)\} \overset{H_0}{\sim} \chi_{p-p_0}, \tag{1.18}$$

dove $\hat{\theta}_0$ è la stima di massima verosimiglianza di θ sotto i vincoli imposti dall'ipotesi nulla (Salvan *et al.*, 2020).

1.2.1 Un caso di studio in R

Il dataset `spinach` (Streibig *et al.*, 1999) contiene i dati provenienti da una sperimentazione per valutare l'inibizione della fotosintesi negli spinaci in risposta all'utilizzo di due erbicidi (Bentazone e Diuron). In particolare, sono stati effettuati cinque esperimenti in cui la risposta è costituita dal consumo di ossigeno nei cloroplasti (`SLOPE`), mentre la variabile esplicativa è la quantità di erbicida (`DOSE`). In questa sezione verranno utilizzati solamente i dati del primo dei cinque esperimenti (con il Bentazone come erbicida). Il dataset utilizzato e il diagramma di dispersione (Figura 1.1) si ottengono con:

```
> dati=spinach[spinach$CURVE==1,3:4]
> head(dati)
  DOSE   SLOPE
1 0.00 1.81295
2 0.00 1.86704
3 0.00 1.95606
4 0.62 1.39073
5 0.62 1.15721
6 0.62 1.06126
> attach(dati)
> plot(SLOPE~DOSE)
```

Per la costruzione del modello dose-risposta viene proposto inizialmente il modello normale con funzione media log-logistica a 3 parametri (la funzione (1.3) con $c = 0$ e $f = 1$). I parametri di un generico modello non lineare possono essere stimati nell'ambiente open-source `R` con la funzione `nls()`, la quale richiede come argomenti la formula del modello e i valori dei parametri

per l'algoritmo iterativo di stima. Un'alternativa specifica per i modelli dose-risposta è costituito dal pacchetto `drc` (Ritz *et al.*, 2015), che verrà utilizzato nel seguito della tesi. I modelli sono adattati con la funzione `drm()`, che richiede la specificazione della `formula` del modello e della media $\mu(x, \beta)$:

```
> ll3.spinach=drm(SLOPE~DOSE, fct=LL.3())
> summary(ll3.spinach)

Model fitted: Log-logistic (ED50 as parameter) with lower limit at 0 (3 parms)

Parameter estimates:

              Estimate Std. Error t-value   p-value
b:(Intercept) 0.530382   0.034939 15.1801 1.055e-11 ***
d:(Intercept) 1.879055   0.043789 42.9113 < 2.2e-16 ***
e:(Intercept) 1.733230   0.243197  7.1269 1.220e-06 ***
```

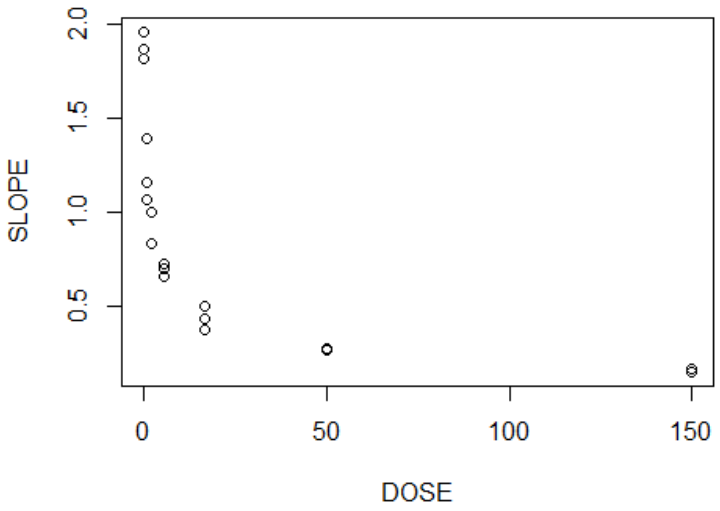


Figura 1.1: Diagramma di dispersione di `SLOPE` vs `DOSE`

```
Residual standard error:

0.07616962 (18 degrees of freedom)
> plot(ll3.spinach, log="",type = "confidence")
> plot(ll3.spinach, log="",type = "obs", add=T, pch=".", cex = 3)
```

L'output di `summary` fornisce le stime di massima verosimiglianza dei parametri, con i relativi standard error e significatività. Si può notare come tutti i parametri risultino significativi (`p-value` < 0.001) e che $\hat{b} = 0.53$, $\hat{d} = 1.88$ e $\hat{e} = 1.73$. L'asintoto superiore per il consumo di ossigeno è quindi 1.88, mentre $1.73\mu M$ è la concentrazione di erbicida che produce un effetto pari al 50% dell'effetto massimo. La funzione $\mu(x, \beta)$ stimata per il modello è quindi

$$\mu(x, \hat{\beta}) = \frac{1.88}{1 + \exp(0.53 \cdot (\log(x) - \log(1.73)))}. \tag{1.19}$$

In Figura 1.2 è rappresentata la curva per la media stimata dal modello e il relativo intervallo di confidenza. Si può notare un buon adattamento del modello ai dati. Con la funzione `confint` è inoltre possibile ricavare gli intervalli di confidenza alla Wald. Di seguito vengono riportati tali intervalli di livello 95% per i tre parametri del modello:

```
> confint(ll3.spinach)
                  2.5 %   97.5 %
b:(Intercept) 0.456977 0.603786
d:(Intercept) 1.787058 1.971053
e:(Intercept) 1.222292 2.244169
```

Dall'analisi dei residui emerge come l'assunto di normalità dei residui possa essere accettato (`p-value` del test di Shapiro-Wilk maggiore di 0.1). Inoltre, il secondo grafico in Figura 1.3 permette di accettare anche l'assunto di omoschedasticità.

```
> res=residuals(ll3.spinach)
> qqnorm(res)
> qqline(res)
> plot(res,fitted(ll3.spinach))
```

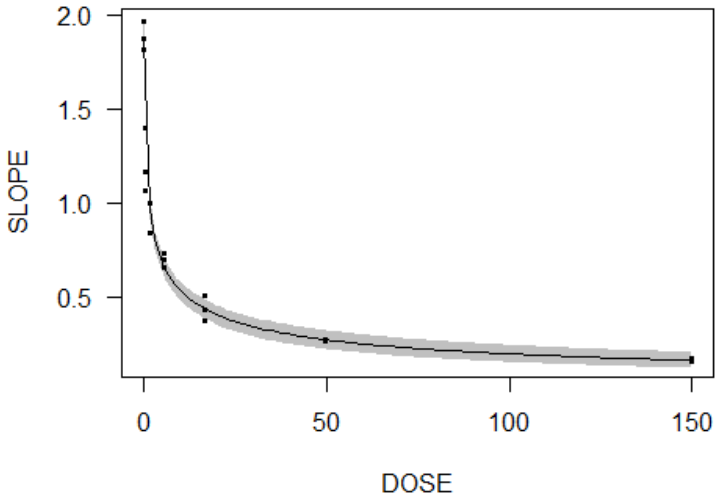


Figura 1.2: Modello stimato (e relativo intervallo di confidenza)

```
> shapiro.test(res)

Shapiro-Wilk normality test

data:  res
W = 0.9301, p-value = 0.1383
```

Infine, viene proposto il modello con $\mu(x, \beta)$ log-logistica a 5 parametri: si tratta quindi di un modello più complicato del precedente, poiché comprende due parametri in più a causa dell'assenza dei vincoli $c = 0$ e $f = 1$. Come si può notare dall'output di `summary`, i parametri c e f non solo non risultano significativi, ma la loro introduzione nel modello rende non significativi anche i parametri b e e:

```
> ll5.spinach=drm(SLOPE~DOSE, fct=LL.5())
> summary(ll5.spinach)

Model fitted: Generalized log-logistic (ED50 as parameter) (5 parms)
```

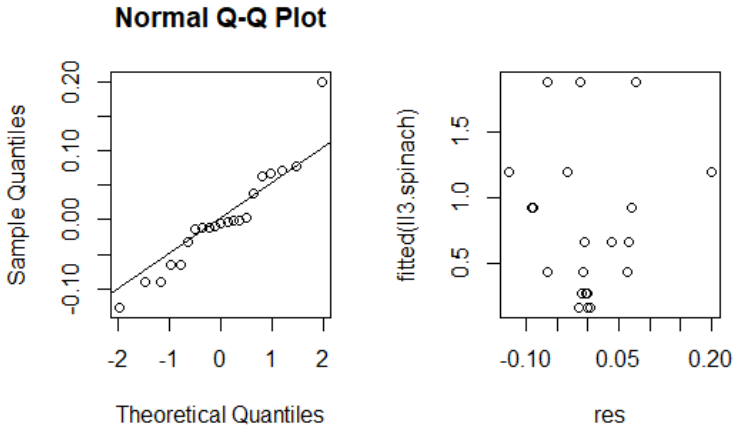


Figura 1.3: Analisi dei residui del modello (1.19)

```
Parameter estimates:

Estimate Std. Error t-value p-value
b:(Intercept)  0.767103   1.113423  0.6890  0.5007
c:(Intercept) -0.233241   0.773116 -0.3017  0.7668
d:(Intercept)  1.878906   0.046452 40.4487  <2e-16 ***
e:(Intercept)  0.232551   0.926591  0.2510  0.8050
f:(Intercept)  0.343573   1.009508  0.3403  0.7380

Residual standard error:

0.08047945 (16 degrees of freedom)
```

Infine, viene proposto il test presentato in (1.18) per modelli annidati, che si ottiene in `R` col comando `anova`; il test porta ad accettare l'ipotesi nulla del modello semplificato a 3 parametri (`p-value` > 0.1):

```
> anova(ll3.spinach,ll5.spinach,test="chisq")

1st model
```

```
fct:      LL.3()
2nd model
fct:      LL.5()

ANOVA-like table

ModelDf Loglik Df LR value p value
1st model       4 25.892
2nd model       6 25.972  2   0.1618  0.9223
```

Come ultima considerazione, l'AIC del modello semplificato risulta essere minore rispetto a quello del modello completo:

```
> AIC(ll3.spinach, ll5.spinach)
df       AIC
ll3.spinach  4 -43.78303
ll5.spinach  6 -39.94483
```

Tutte queste considerazioni portano a scegliere come media per il consumo di ossigeno la funzione (1.19).

1.3 Modelli per dati di conteggio

Molti studi hanno dimostrato come l'utilizzo di tecniche di regressione presenti diversi vantaggi (Ritz e Van Der Vliet, 2009) quando si vuole stimare un modello dose-risposta. Tuttavia, per utilizzare queste tecniche di regressione correttamente è necessario che le osservazioni rispettino gli assunti teorici: ad esempio, per una risposta continua si assume spesso una distribuzione normale degli errori e omoschedasticità degli errori per diversi valori della dose. Quando però la risposta è costituita da dati di conteggio l'assunto di normalità è violato, anche se in alcuni casi i modelli di regressione non lineare normali possono fornire una valida approssimazione. In particolare, una distribuzione di Poisson di parametro λ elevato può essere ben approssimata con una distribuzione normale, poiché presenta la caratteristica forma a campana ed è raro trovare valori ripetuti tra le osservazioni. Quando però

i conteggi sono prossimi allo zero, come nel caso di distribuzioni di Poisson di parametro λ piccolo, emerge la natura discreta dei dati, si possono osservare valori ripetuti e l'approssimazione normale non è più valida. Un'altra caratteristica che non consente ai dati di soddisfare le ipotesi del modello è l'eteroschedasticità, che viene riscontrata sia in dati di tipo continuo sia in dati di conteggio soprattutto in ambito tossicologico. Solitamente infatti vi è un'alta variabilità della risposta quando la dose è nulla o bassa, mentre all'aumentare della dose la variabilità cala. Per contrastare la non normalità e l'eteroschedasticità dei dati (Ritz e Van Der Vliet, 2009) propongono due diverse soluzioni: la trasformazione di Box-Cox e l'utilizzo della distribuzione di Poisson nei modelli dose-risposta.
La prima soluzione per ottenere normalità e omoschedasticità è applicare a entrambi i membri dell'equazione (1.4) una opportuna funzione $g(\cdot)$; solitamente si utilizza la radice quadrata o il logaritmo. Il modello risultante sarà quindi

$$g(y_i) = g(\mu(x_i, \beta)) = g\left(c + \frac{d-c}{1+\exp(b(\log(x_i) - \log(e)))}\right) + \varepsilon_i, \quad i = 1, ..., n. \tag{1.20}$$

Le trasformazioni di Box-Cox costituiscono un'ampia famiglia di funzioni spesso utilizzate, indicizzate da un parametro $\lambda > 0$. La generica trasformazione di Box-Cox è definita nel seguente modo

$$g_\lambda(x) = \frac{x^\lambda - 1}{\lambda}, \tag{1.21}$$

dove, con $\lim_{\lambda \to 0^+} g_\lambda(x) = log(x)$. Si noti come anche la radice quadrata rientri nelle trasformazioni di Box-Cox, poiché la si ottiene ponendo λ pari a 0.5. Un possibile metodo per stimare λ consiste nel minimizzare la somma dei quadrati dei residui (si veda sempre Ritz e Van Der Vliet (2009) per i dettagli), ossia

$$\sum_{i=1}^{n} \left(g_\lambda(y_i) - g_\lambda \left(c + \frac{d-c}{1+\exp(b(\log(x_i) - \log(e)))} \right) \right)^2.$$

Quando la risposta è una variabile di conteggio, è invece possibile considerare per i dati un modello di Poisson. Il modello prevede che le osservazioni siano

realizzazioni di una variabile di Poisson di media

$$\mu(x_i, \beta) = c + \frac{d - c}{1 + \exp(b(\log(x_i) - \log(e)))}, \qquad i = 1, ..., n.$$

Quindi $\mu(x, \beta)$, oltre ad essere la media della variabile risposta, è anche la sua varianza. Spesso però i dati presentano *sovradispersione*, ossia la varianza è maggiore della media. Si può quindi scrivere $var(Y_i) = \phi\mu(x_i, \beta)$, con $\phi > 1$ detto *parametro di dispersione* (Ritz e Van Der Vliet, 2009, McCullagh e Nelder, 1989). Una statistica utilizzata per valutare la presenza di sovradispersione è il test chi quadrato di Pearson, basato sul numero di casi osservati O_i e sul numero di casi attesi E_i, che in questo caso corrisponde a $\mu(x_i, \hat{\beta})$:

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i}, \tag{1.22}$$

che, sotto l'ipotesi di corretta specificazione del modello, ha distribuzione approssimata χ^2_{n-d}, con n numerosità campionaria e d dimensione nello spazio parametrico. Questa statistica test è utile più per evidenziare un eventuale allontanamento dalla distribuzione di Poisson, poiché un esito non significativo non comporta necessariamente la correttezza delle assunzioni fatte. È inoltre possibile ottenere una stima del parametro di dispersione come

$$\hat{\phi} = \frac{X^2}{n - d}. \tag{1.23}$$

Infine, è possibile aggiustare il modello e renderlo più robusto moltiplicando gli standard error dei parametri per un fattore di scala, calcolato come la radice quadrata di $\hat{\phi}$.

1.4 Un caso di studio: i dati sul COVID-19

Di seguito viene proposto un esempio di utilizzo dei modelli dose-risposta e della stima dei loro parametri tramite il pacchetto *drc*. I dati provengono dai report giornalieri della Protezione Civile sulla diffusione del COVID-19 in Italia (`https://github.com/pcm-dpc/COVID-19/`); in particolare, è stata scelta come variabile risposta il numero cumulato di morti giornaliero per COVID-19, mentre è il tempo, misurato in giorni, a fungere da dose. I dati

sono disponibili a partire dal 24 febbraio 2020; nel seguito verranno utilizzati a scopo illustrativo i dati aggiornati al 15 febbraio 2021, data scelta come approssimazione della fine della seconda ondata. L'obiettivo di questa prima analisi è duplice: mostrare un'applicazione pratica dei modelli dose-risposta e tentare di cogliere eventuali differenze nelle caratteristiche delle prime due ondate con cui il virus si è diffuso in Italia. Si è scelto di fissare il 15 agosto 2020 come fine della prima ondata: i modelli dei Capitoli successivi mostreranno che la data in cui è terminata la prima ondata dovrebbe essere posta qualche giorno più avanti, ma considerato lo scopo puramente illustrativo e i risultati ottenuti in altri articoli (Greco *et al.*, 2021, *Submitted*), la scelta del 15 agosto è accettabile.
Si noti come, vista la struttura dei dati, è possibile avere un solo valore della risposta per ciascun valore della dose: non si tratta infatti di uno studio sperimentale nel quale è possibile ripetere le misurazioni più volte per ciascun giorno. In Figura 1.4 è rappresentato, per ciascuna delle due ondate, l'andamento del numero di deceduti da inizio ondata. Si notano immediatamente alcune differenze tra le due ondate: innanzitutto la prima ha causato circa 36000 morti, mentre nella seconda si è arrivati a quasi 60000 deceduti. Inoltre, mentre nel periodo tra fine febbraio e inizio marzo 2020 c'è stato un rapido aumento del numero di deceduti, nel periodo autunnale sono passate alcune settimane prima di raggiungere un numero elevato di morti. Nei due grafici è riportata la retta che segnala il giorno in cui si sono superati i 5000 morti: per la prima ondata è il ventottesimo mentre per la seconda è l'ottantatreesimo.
Considerata la natura dei dati, con la variabile risposta che rappresenta dei conteggi non prossimi allo zero, ci si aspetta che la distribuzione normale possa essere una buona approssimazione della vera distribuzione dei dati. Per questo motivo inizialmente si è scelto di stimare, per ciascuna ondata, un modello di regressione non lineare normale (si veda Girardi *et al.*, 2020b). Si può notare immediatamente come l'ipotesi di indipendenza non sia rispettata: vale infatti $y_{i+1} \geq y_i, \forall i$; si rimanda alla parte conclusiva del capitolo per una discussione dettagliata di questa problematica e delle sue possibili soluzioni. Il modello completo è stato costruito utilizzando come funzione $\mu(x, \beta)$ la log-logistica a 5 parametri (1.3). Successivamente vengono proposti

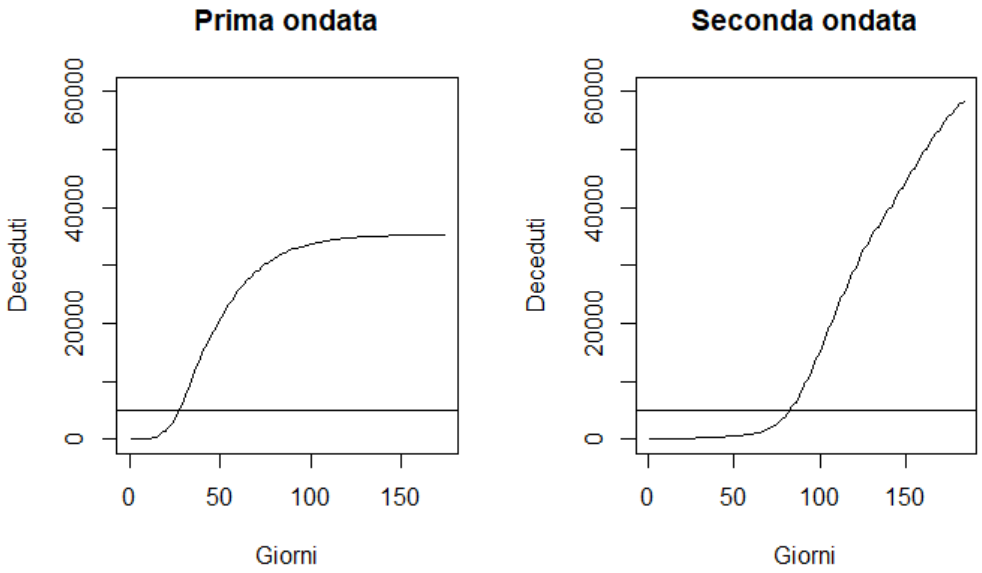


Figura 1.4: Andamento del numero cumulato di deceduti: sull'asse delle ascisse il numero di giorni trascorsi dall'inizio dell'ondata

e valutati alcuni modelli semplificati annidati, ottenuti ponendo dei vincoli sui parametri. In particolare, si è scelto di verificare le ipotesi nulle

- $H_0 : c = 0$ per la prima ondata, che costituisce il logico limite inferiore del numero cumulato di deceduti a inizio pandemia;
- $H_0 : f = 1$, che rende il modello simmetrico e il parametro e è di più semplice interpretazione.

Di seguito vengono presentate le procedure che hanno portato alla selezione dei parametri per il miglior modello non lineare normale, effettuate utilizzando i dati della prima ondata. Come si può notare dall'output il parametro c non risulta essere significativo:

```
> m1.ll5=drm(prima~d1, fct=LL.5())
> summary(m1.ll5)

Model fitted: Generalized log-logistic (ED50 as parameter) (5 parms)
```

```
Parameter estimates:

                Estimate  Std. Error  t-value p-value
b:(Intercept) -3.1329e+00  3.9590e-02 -79.1324  <2e-16 ***
c:(Intercept) -4.9949e+01  5.5465e+01  -0.9005  0.3691
d:(Intercept)  3.6104e+04  4.7344e+01 762.5943  <2e-16 ***
e:(Intercept)  4.0566e+01  8.1051e-01  50.0499  <2e-16 ***
f:(Intercept)  1.3028e+00  5.9493e-02  21.8979  <2e-16 ***

Residual standard error:

199.08 (169 degrees of freedom)
```

Viene quindi costruito il modello semplificato che si ottiene fissando $c = 0$, attraverso il comando `LL.5(fixed=c(NA,0,NA,NA,NA))`. Un'ulteriore semplificazione della funzione $\mu(x, \beta)$ si ottiene col vincolo $f = 1$. Tale semplificazione porta però ad un notevole peggioramento del modello nell'adattamento ai dati, come mostrato dall'output del test del rapporto di verosimiglianza (`p-value` < 0.05):

```
> m1.ll3=drm(prima~d1, fct=LL.3())
> anova(m1.ll4.f,m1.ll3,test="Chisq")

1st model
fct:      LL.5(fixed = c(NA, 0, NA, NA, NA))
2nd model
fct:      LL.3()

ANOVA-like table

          ModelDf  Loglik Df LR value p value
1st model       5 -1166.6
2nd model       4 -1193.5  1   53.841       0
```

Per quanto riguarda la prima ondata il miglior modello non lineare normale risulta essere quello con funzione media log-logistica a 4 parametri (b, d, e e f). Stime, standard error e intervalli di confidenza alla Wald per i parametri di tale modello sono riportati di seguito:

```
> summary(m1.ll4.f)

Model fitted: Generalized log-logistic (ED50 as parameter) (4 parms)

Parameter estimates:

                Estimate  Std. Error t-value   p-value
b:(Intercept) -3.1178e+00  3.8466e-02 -81.055 < 2.2e-16 ***
d:(Intercept)  3.6116e+04  4.7459e+01 760.993 < 2.2e-16 ***
e:(Intercept)  4.0133e+01  7.4849e-01  53.619 < 2.2e-16 ***
f:(Intercept)  1.3380e+00  5.4906e-02  24.369 < 2.2e-16 ***

Residual standard error:

199.8043 (170 degrees of freedom)

> confint(m1.ll4.f)
                         2.5 %       97.5 %
b:(Intercept)    -3.193753    -3.041889
d:(Intercept) 36022.477844 36209.848484
e:(Intercept)    38.655847    41.610915
f:(Intercept)     1.229640     1.446410
```

La stima della funzione media $\mu(x, \beta)$ per la prima ondata è quindi

$$\mu(x, \hat{\beta}) = \frac{36116}{(1 + \exp(-3.12 \cdot (\log(x) - \log(40.13))))^{1.34}}.$$

Nel grafico di sinistra di Figura 1.5 è riportata la curva stimata dal modello per la prima ondata, confrontata con i dati osservati.
Per quanto riguarda il modello completo stimato sui dati della seconda ondata, risultano invece significativi tutti e 5 i parametri. In questo caso non

ha senso imporre il vincolo $c = 0$, poiché il numero cumulato di deceduti è riferito all'intera pandemia e non solo alla seconda ondata:

```
> summary(m2.ll5)

Model fitted: Generalized log-logistic (ED50 as parameter) (5 parms)

Parameter estimates:

Estimate  Std. Error  t-value   p-value
b:(Intercept) -2.7009e+00  5.4534e-02 -49.5277 < 2.2e-16 ***
c:(Intercept)  3.5651e+04  5.9636e+01 597.8136 < 2.2e-16 ***
d:(Intercept)  1.1674e+05  1.0070e+03 115.9300 < 2.2e-16 ***
e:(Intercept)  4.7880e+01  4.4984e+00  10.6438 < 2.2e-16 ***
f:(Intercept)  1.3167e+01  2.7914e+00   4.7172 4.798e-06 ***

Residual standard error:

477.4419 (179 degrees of freedom)
```

Si noti come, per questo modello, i parametri c e d rappresentino, rispettivamente, il numero stimato di deceduti a inizio seconda ondata e la somma prevista di deceduti causata dalle due ondate. Il modello normale per i dati della seconda ondata ha quindi funzione media stimata

$$\mu(x, \hat{\beta}) = 35651 + \frac{81085}{(1 + \exp(-2.70 \cdot (\log(x) - \log(47.88))))^{13.17}}.$$

Nel grafico di destra di Figura 1.5 è riportata la curva stimata dal modello per la seconda ondata, confrontata con i dati osservati.
Gli intervalli di confidenza alla Wald per i parametri di questo modello sono:

```
> confint(m2.ll5)
                       2.5 %        97.5 %
b:(Intercept)     -2.808546     -2.593322
c:(Intercept)  35533.491979  35768.851813
d:(Intercept) 114748.928385 118722.976693
```

```
e:(Intercept)     39.003574     56.757065
f:(Intercept)      7.659266     18.675648
```

Sono stati stimati per le due ondate anche i modelli di Poisson

$$Y_i \sim Po(\mu_i), \quad \mu_i = c + \frac{d-c}{(1+\exp(b(\log(x_i)-\log(e))))^f}, \quad i = 1, ..., n, \tag{1.24}$$

ponendo $c = 0$ per la prima ondata visti i risultati ottenuti con i modelli normali. In Tabella 1.1 viene riportato un confronto tra i parametri del modello log-logistico, sia normale sia di Poisson, per la prima e la seconda ondata. Sono riportate sia le stime di massima verosimiglianza sia i loro standard error. Innanzitutto, risulta evidente come per la prima ondata il modello normale e quello di Poisson forniscano stime dei parametri simili; al contrario, nella seconda ondata vi è una differenza tra le stime dei parametri fornite dai due modelli: si noti come le stime dei parametri del modello di Poisson non siano comprese negli intervalli di confidenza ottenuti con il modello normale. Inoltre, gli standard error dei modelli di Poisson sono sistematicamente inferiori rispetto a quelli dei modelli normali. Tuttavia, visti i risultati del test chi-quadrato di Pearson presentati in Tabella 1.2, gli standard error dei modelli di Poisson andrebbero moltiplicati per il fattore di scala $\sqrt{\hat{\phi}}$, pari a 7.23 per la prima ondata e 3.18 per la seconda ondata, al fine di tener conto della sovradispersione. Per quanto riguarda il confronto tra le ondate, si possono notare alcune differenze già evidenziate dai grafici precedenti; la seconda ondata presenta una maggiore differenza tra i coefficienti d e c e un valore più alto per il coefficiente e: ciò va interpretato come un maggior numero di decessi all'interno dell'ondata e un maggior numero di giorni prima di raggiungere un determinato numero di deceduti.

Per valutare la correttezza dell'assunto sulla distribuzione di Poisson si utilizza la statistica chi-quadrato di Pearson (1.22) che, sotto l'ipotesi nulla di corretta specificazione del modello, ha distribuzione χ^2_{n-d}, con n numerosità campionaria e d dimensione nello spazio parametrico. I valori osservati di tale statistica, i relativi gradi di libertà e la radice quadrata della stima del parametro di dispersione sono riportati in Tabella 1.2. Sia per la prima che per la seconda ondata il modello di Poisson non riesce a cogliere tutta la variabilità presente nei dati (`p-value`<0.001): c'è quindi sovradispersione

Tabella 1.1: Stime e standard error (tra parentesi) dei parametri della log-logistica nelle due ondate

	Prima ondata				**Seconda ondata**			
	Normale		Poisson		Normale		Poisson	
b	-3.12	(0.04)	-3.23	(0.02)	-2.70	(0.05)	-4.24	(0.04)
c	—	—	—	—	35651.17	(59.64)	35232.75	(24.80)
d	36116.16	(47.46)	35926.20	(34.77)	116735.95	(1007.00)	105092.14	(328.00)
e	40.13	(0.75)	40.80	(0.25)	47.88	(4.50)	116.50	(0.52)
f	1.34	(0.05)	1.29	(0.02)	13.17	(2.79)	1.43	(0.08)

Tabella 1.2: Test chi-quadrato di Pearson sui modelli di Poisson

	X^2_{obs}	Gradi di libertà	$\sqrt{\hat{\phi}}$	p-value
Prima ondata	8893.792	170	7.23	<0.001
Seconda ondata	1805.484	179	3.18	<0.001

(si veda Greco *et al.*, 2021, *Submitted*). In conclusione, i modelli di Poisson, anche se più corretti da un punto di vista degli assunti teorici su cui sono basati, si adattano peggio ai dati di quanto facciano i modelli normali.
In Figura 1.5 sono riportate le due curve stimate con i modelli normali. Nel complesso i modelli ben si adattano ai dati, seppur con delle difficoltà nel cogliere l'andamento nella prima parte della seconda ondata, come si può notare dalla Figura 1.6. Nonostante il buon adattamento, l'analisi dei residui rivela una mispecificazione dei modelli per entrambe le ondate. I test di Shapiro-Wilk portano a rifiutare l'ipotesi di normalità (`p-value`<0.05); la medesima conclusione può essere tratta a partire dai test grafici mostrati in Figura 1.7, che portano a rifiutare anche l'assunto di omoschedasticità.

```
> res=residuals(m1.ll4.f)
> shapiro.test(res)

Shapiro-Wilk normality test

data:  res
W = 0.97511, p-value = 0.003249
```

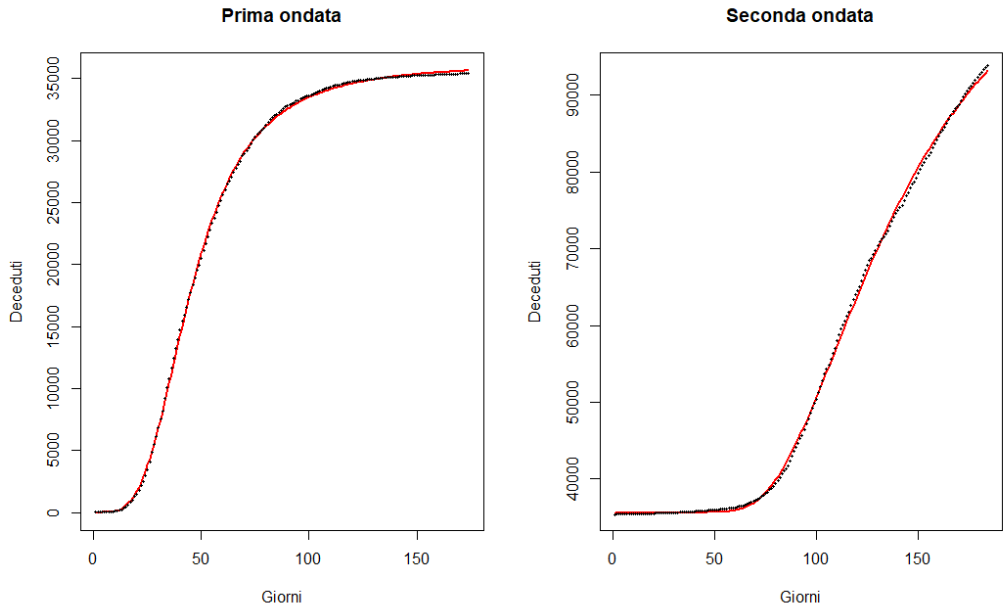


Figura 1.5: Dati osservati e curve stimate dal modello

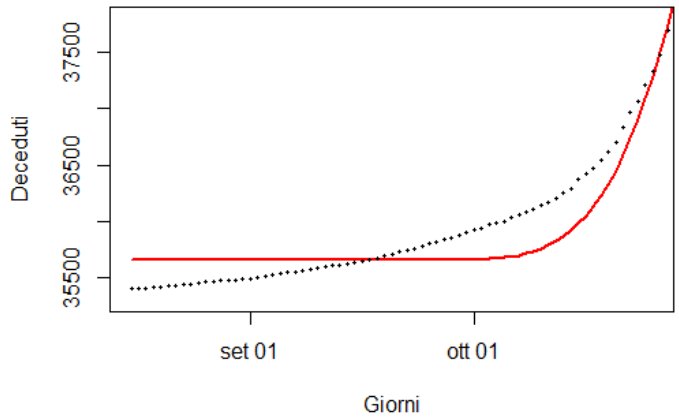


Figura 1.6: Modello normale per i dati della seconda ondata (dal 16/08/2020 al 25/10/2020)

```
> res2=residuals(m2.ll5)
> shapiro.test(res2)

Shapiro-Wilk normality test

data:  res2
W = 0.98389, p-value = 0.03261
```

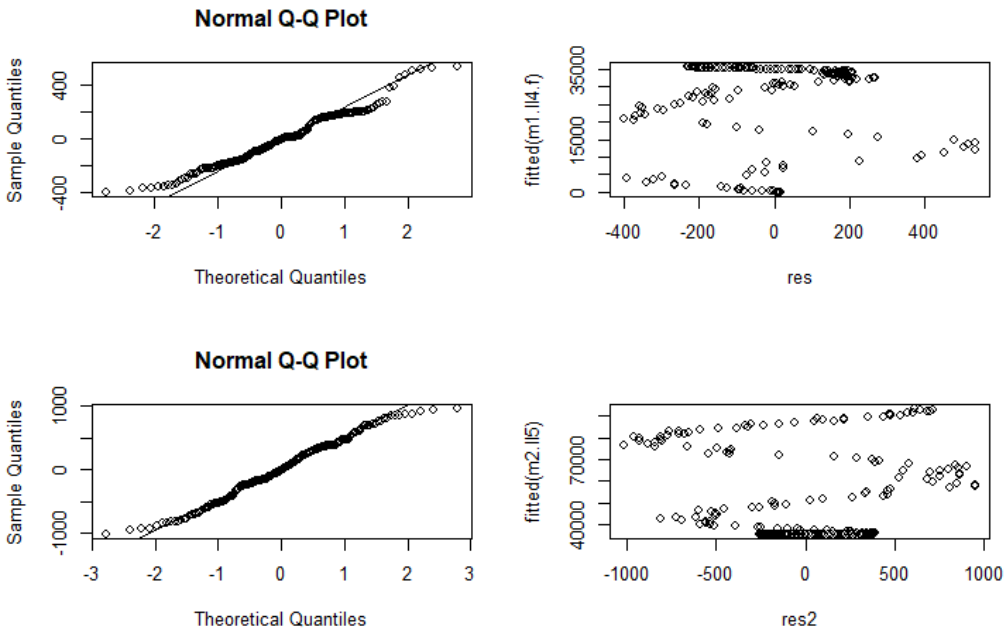


Figura 1.7: Analisi dei residui nei modelli normali: prima riga per la prima ondata, seconda riga per la seconda ondata

Inoltre, sia per i modelli normali sia per i modelli di Poisson, risulta evidente come l'assunto di indipendenza per i residui non venga rispettato. Infatti, vista la particolare natura dei dati, vale necessariamente $y_{i+1} \geq y_i, \forall i$. Per ovviare a questo, è possibile utilizzare un approccio basato su una particolare funzione di verosimiglianza *composita*, ovvero una funzione ottenuta moltiplicando un insieme di componenti individuali (Varin *et al.*, 2011). Nello specifico, quando queste componenti individuali corrispondono alle verosimiglianze

marginali, si ottiene

$$L_I(\theta) = \prod_{i=1}^{n} p(y_i; \theta),$$

con $L_I(\theta)$ che viene chiamata *independence likelihood* in letteratura (si veda Greco *et al.*, 2021, *Submitted*, e Varin *et al.*, 2011). La *stima di massima verosimiglianza composita* $\hat{\theta}_I$ si ottiene risolvendo l'*equazione score composita* definita da

$$u_I(\theta) = \sum_{i=1}^{n} u_I(y_i; \theta) = \frac{\partial l_I(\theta)}{\partial \theta} = 0,$$

dove $l_I(\theta) = \log L_I(\theta)$. Lo stimatore corrispondente è asintoticamente distribuito come una normale d-variata, ossia

$$\hat{\theta}_I \dot{\sim} N_d(\theta, G(\theta)^{-1}),$$

con $G(\theta)$ matrice d'informazione di Godambe (Godambe, 1960), definita da

$$G(\theta) = H(\theta) J(\theta)^{-1} H(\theta), \tag{1.25}$$

con $H(\theta) = \mathbb{E}(-\partial u_I(\theta)/\partial \theta^T)$ e $J(\theta) = \mathbb{E}(u_I(\theta) u_I(\theta)^T)$.

1.5 Possibili sviluppi e miglioramenti

L'analisi che è stata effettuata sui dati del COVID-19 in Italia ha lo scopo di introdurre con un esempio pratico le peculiarità dei modelli dose-risposta presentati da un punto di vista teorico nelle sezioni precedenti. Non avendo l'obiettivo di trovare dei risultati innovativi dal punto di vista della modellazione e della previsione, quest'analisi risulta quindi piuttosto semplicistica e soggetta a diversi possibili miglioramenti. La natura stessa dei dati fa emergere delle problematiche la cui risoluzione necessita di strumenti statistici più complessi rispetto a quelli presentati finora. La principale di queste problematiche è costituita dall'errata specificazione del modello: come evidenziato dall'analisi dei residui, gli assunti di normalità, omoschedasticità e indipendenza non sono validi per i dati del COVID-19. Inoltre, considerata l'inevitabile mancanza di accuratezza nella fase di raccolta, non è possibile attribuire ai

dati un elevato livello di attendibilità. Per approfondimenti legati alla necessità di dati di alta qualità si rimanda al sito del gruppo di ricerca StatGroup-19 (https://statgroup-19.blogspot.com/2020/11/il-manifesto.html) e alla conferenza *Lotta al COVID-19 Dati di Alta Qualità per le Analisi e Competenze Adeguate* (https://www.youtube.com/watch?v=W1cy9uN-wCQ). Per affrontare i problemi di errata specificazione del modello e di scarsa attendibilità dei dati è consigliato l'utilizzo di un approccio robusto (Girardi *et al.*, 2020b, 2020a). Nello specifico, viene proposto l'utilizzo di un approccio robusto all'inferenza basato su una *funzione punteggio* (*scoring rule*), ovvero una funzione obiettivo che misura la qualità di una determinata distribuzione di probabilità per una variabile aleatoria Y, alla luce di una realizzazione $y = (y_1, ..., y_n)$ di Y (Dawid *et al.*, 2016). La funzione punteggio scelta in Girardi *et al.* (2020b) è la funzione score di Tsallis, definita da

$$S(y;\theta) = (\gamma - 1) \int p_Y(y;\theta)^\gamma dy \; - \; \gamma p_Y(y;\theta)^{\gamma-1}, \quad \gamma > 1, \tag{1.26}$$

dove γ è un parametro che regola il compromesso tra efficienza e robustezza. Nei modelli di regressione non lineari normali considerati in questa tesi, la (1.26) per $\theta = (\beta, \sigma^2)$ diventa

$$S(y;\theta) = \sum_{i=1}^{n} \left[\frac{\gamma - 1}{\sqrt{\gamma}(2\pi\sigma^2)^{(\gamma-1)/2}} - \gamma \left(\frac{1}{\sqrt{2\pi\sigma^2}} \right)^{\gamma-1} \exp\left(-\frac{\gamma - 1}{2\sigma^2}(y_i - \mu(x_i, \beta))^2 \right) \right]. \tag{1.27}$$

Dato $s(y;\theta)$ vettore delle derivate parziali di $S(y;\theta)$ rispetto alle d componenti di θ, la *funzione di stima* è

$$s(y;\theta) = \sum_{i=1}^{n} s(y_i;\theta), \tag{1.28}$$

da cui si ottiene una stima di θ come

$$\tilde{\theta} \in \Theta : s(y;\tilde{\theta}) = 0. \tag{1.29}$$

Lo stimatore $\tilde{\theta}$ ha distribuzione asintotica normale d-variata con media θ e matrice di covarianze $V(\theta)$, dove

$$V(\theta) = K(\theta)^{-1} J(\theta) (K(\theta)^{-1})^T, \tag{1.30}$$

con $K(\theta) = \mathbb{E}_\theta\left(\frac{\partial s(y;\theta)}{\partial \theta^T}\right)$ e $J(\theta) = \mathbb{E}_\theta(s(y;\theta)s(y;\theta)^T)$.
Le procedure di inferenza per θ, come intervalli di confidenza e test di verifica di ipotesi, possono essere effettuate partendo dalla statistica di tipo Wald

$$W_{S,e}(\theta) = (\tilde{\theta} - \theta)^T V(\tilde{\theta})^{-1}(\tilde{\theta} - \theta), \tag{1.31}$$

che è asintoticamente distribuita come una chi-quadrato con d gradi di libertà. In alternativa, è possibile utilizzare la statistica

$$W_S(\theta) = 2\{S(y;\theta) - S(y;\tilde{\theta})\}, \tag{1.32}$$

che però ha lo svantaggio di non avere una distribuzione asintotica standard, dato che

$$W_S(\theta) \xrightarrow{d} \sum_{j=1}^{d} \lambda_j Z_j^2, \tag{1.33}$$

dove λ_j sono gli autovalori di $J(\theta)K(\theta)^{-1}$ e Z_j sono variabili normali standard indipendenti (Girardi *et al.*, 2020b, Greco *et al.*, 2021, *Submitted*).
Utilizzando una generica funzione punteggio, le matrici $K(\theta)$ e $J(\theta)$ possono non essere derivabili analiticamente: in questo caso una loro stima, necessaria per la stima della varianza di $\tilde{\theta}$, può essere ottenuta con metodi di ricampionamento, ad esempio bootstrap o jackknife (Varin *et al.*, 2011). Nel caso dello score di Tsallis le matrici $K(\theta)$ e $J(\theta)$ sono però ricavabili in forma esplicita; in particolare, se sono verificati gli assunti del Teorema 3.1 di Ghosh e Basu (2013), vale

$$K(\theta) = \begin{pmatrix} \frac{\xi_\alpha}{n}\frac{\partial \mu}{\partial \beta}^T\frac{\partial \mu}{\partial \beta} & \mathbf{0} \\ \mathbf{0}^T & \varsigma_\alpha \end{pmatrix} \tag{1.34}$$

$$J(\theta) = \begin{pmatrix} \frac{\xi_{2\alpha}}{n}\frac{\partial \mu}{\partial \beta}^T\frac{\partial \mu}{\partial \beta} & \mathbf{0} \\ \mathbf{0}^T & \varsigma_{2\alpha} - \frac{\alpha^2\xi_\alpha}{4} \end{pmatrix}, \tag{1.35}$$

dove $\alpha = \gamma - 1$, $\frac{\partial \mu}{\partial \beta}^T = \left(\frac{\partial \mu(x_1,\beta)}{\partial \beta}, ..., \frac{\partial \mu(x_n,\beta)}{\partial \beta}\right)$ è una matrice $(d-1) \times n$, $\xi_\alpha = 2\pi^{(-\alpha/2)}\sigma^{-(\alpha+2)/2}(1+\alpha)^{-3/2}$ e $\varsigma_\alpha = \frac{1}{4}(2\pi)^{-\alpha/2}\sigma^{(-\alpha+4)/2}\frac{2+\alpha^2}{(1+\alpha)^{5/2}}$ (Girardi *et al.*, 2020b, Ghosh e Basu, 2013).
Un'alternativa all'utilizzo di funzioni score robuste per contrastare eteroschedasticità e autocorrelazione nei dati consiste nell'operare sulla matrice di covarianze dello stimatore per θ. Greco *et al.* (2021, *Submitted*) propongono due soluzioni di questo tipo: la prima consiste nel moltiplicare la matrice

di covarianze per un parametro di dispersione analogo a quello definito in (1.23), mentre la seconda proposta è quella di correggere la matrice di covarianze per eteroschedasticità e autocorrelazione (HAC, si veda Zeileis, 2004). Infine, una tipologia di modelli che tenga in considerazione la sovrapposizione dei dati riferiti alle due ondate è costituita dai modelli mistura; in alternativa, è possibile utilizzare un modello con punti di cambio per descrivere le due ondate, con l'istante temporale della transizione dalla prima alla seconda ondata che può essere stimato con delle procedure basate su quantità di verosimiglianza (Greco *et al.*, 2021, *Submitted*). Tale modello sarà presentato nel Capitolo 3.

Capitolo 2

Modelli dose-risposta: un approccio bayesiano

In questo capitolo vengono presentati i modelli dose-risposta nel contesto dell'inferenza bayesiana, sia attraverso le basi teoriche su cui si fonda tale approccio, sia attraverso degli esempi. Per gli approfondimenti teorici si fa riferimento, ad esempio, a Gelman *et al.* (2013), Sartori (2020) e Liseo (2008).

2.1 Introduzione all'inferenza bayesiana

L'approccio bayesiano costituisce la principale alternativa all'approccio frequentista per l'inferenza statistica. Mentre la statistica frequentista assume che i parametri di un modello siano dei valori fissati ma ignoti, la peculiarità della statistica bayesiana sta nell'attribuire un carattere di variabile casuale a qualsiasi oggetto ignoto, quindi anche ai parametri di un modello. L'obiettivo dell'inferenza bayesiana è quindi ottenere la distribuzione dei parametri del modello alla luce dei dati osservati.
L'elemento cardine della statistica bayesiana è il teorema di Bayes, che ha il seguente enunciato: dati due eventi A e B, la probabilità che si verifichi l'evento A condizionata al verificarsi dell'evento B è

$$P(A|B) = \frac{P(A)P(B|A)}{P(B)}. \tag{2.1}$$

Nell'ambito dell'inferenza bayesiana su θ la (2.1) viene utilizzata con la seguente notazione:

$$\pi(\theta|\mathbf{y}) = \frac{\pi(\theta)L(\mathbf{y}|\theta)}{\int_\Theta \pi(\theta)L(\mathbf{y}|\theta)\,d\theta}, \tag{2.2}$$

dove:

- $\pi(\theta|\mathbf{y})$ è la *distribuzione a posteriori* e indica la distribuzione del parametro θ condizionata alle osservazioni $\mathbf{y} = (y_1, ..., y_n)$;
- $\pi(\theta)$ è detta *distribuzione a priori* e rappresenta una sintesi dell'informazione che si ha sul parametro θ prima di conoscere l'esito delle osservazioni;
- $L(\mathbf{y}|\theta)$ indica la distribuzione dei dati condizionata al valore del parametro θ, ed è quindi la *verosimiglianza*;
- $\int_\Theta \pi(\theta)L(\mathbf{y}|\theta)\,d\theta$ è un fattore di normalizzazione che non dipende da θ.

Lo scopo dell'inferenza bayesiana è ricostruire la distribuzione a posteriori $\pi(\theta|\mathbf{y})$ partendo dalla distribuzione a priori $\pi(\theta)$, aggiornandola alla luce dei dati osservati. Si ottiene quindi un compromesso tra l'informazione in partenza, sintetizzata dalla distribuzione a priori, e l'informazione contenuta nei dati, sintetizzata dalla verosimiglianza.
Stabilire la distribuzione a priori $\pi(\theta)$ non è un compito semplice, poiché deve riflettere la conoscenza sui parametri prima dell'esperimento e può influenzare pesantemente la stima della distribuzione a posteriori $\pi(\theta|\mathbf{y})$, soprattutto nel caso in cui si scelgano a priori con bassa variabilità e concentrate in valori dei parametri distanti dalle stime di massima verosimiglianza ottenute con i dati. La distribuzione a posteriori contiene tutta l'informazione necessaria per fare inferenza sui parametri: è infatti possibile ottenere sia stime puntuali, ad esempio il valore medio $\mathbb{E}(\theta|\mathbf{y})$, sia stime intervallari, ad esempio utilizzando i quantili della distribuzione. È inoltre possibile effettuare verifiche di ipotesi e fare previsioni.
Esistono alcune distribuzioni a priori, dette *coniugate*, che consentono di ricavare $\pi(\theta|\mathbf{y})$ in forma esplicita senza calcolare $\int_\Theta \pi(\theta)L(\mathbf{y}|\theta)\,d\theta$: ciò accade quando il prodotto di $\pi(\theta)$ per $L(\mathbf{y}|\theta)$ assume la forma di una distribuzione

nota. Si prenda ad esempio $Y_1, ..., Y_n \sim Po(\lambda)$, indipendenti tra loro, e la a priori $\pi(\lambda) \sim Ga(a, b)$, dove $Ga(a, b)$ indica la distribuzione *Gamma*, con a e b detti *iperparametri*. Allora

$$L(\mathbf{y}|\lambda) = \prod_{i=1}^{n} \frac{e^{-\lambda}\lambda^{y_i}}{y_i!} = \frac{e^{-\lambda n}\lambda^{-\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!}$$

$$\pi(\lambda) = \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)}.$$

Ne segue che

$$\begin{aligned}
\pi(\lambda|\mathbf{y}) &\propto \pi(\lambda) L(\mathbf{y}|\lambda) \\
&\propto \frac{b^a \lambda^{a-1} e^{-b\lambda}}{\Gamma(a)} \cdot \frac{e^{-\lambda n}\lambda^{-\sum_{i=1}^{n} y_i}}{\prod_{i=1}^{n} y_i!} \\
&\propto \lambda^{a-1+\sum_{i=1}^{n} y_i} e^{-(b+n)\lambda}.
\end{aligned}$$

La distribuzione a posteriori può quindi essere ricondotta a una distribuzione Gamma di parametri

$$\pi(\lambda|\mathbf{y}) \sim Ga\left(a + \sum_{i=1}^{n} y_i, b + n\right).$$

Tuttavia, nella maggior parte dei casi, non è possibile ricavare $\pi(\theta|\mathbf{y})$ in forma esplicita, poiché non è possibile risolvere l'integrale $\int_\Theta \pi(\theta) L(\mathbf{y}|\theta)\, d\theta$. Ciò ha rappresentato un ostacolo per l'utilizzo nella pratica delle tecniche bayesiane fino agli inizi degli anni '90, quando lo sviluppo delle prestazioni computazionali ha consentito di ovviare al problema: attualmente è possibile calcolare $\int_\Theta \pi(\theta) L(\mathbf{y}|\theta)\, d\theta$ numericamente (se il parametro non ha dimensione troppo elevata) oppure simulare direttamente dalla distribuzione a posteriori per ottenerne un'approssimazione (Sartori, 2020).

2.2 Modelli dose-risposta

L'approccio bayesiano può essere particolarmente utile quando si vuole fare inferenza sui parametri dei modelli dose-risposta, vista l'interpretabilità dei suoi parametri. È infatti ragionevole che gli esperti, come tossicologi, biologi o botanici, abbiano delle informazioni sulle caratteristiche del processo

studiato anche prima di effettuare l'esperimento: ad esempio potrebbe essere nota un'approssimazione della risposta per valori nulli e/o particolarmente elevati della dose, così come il segno del coefficiente b, che indica se c'è un aumento o una diminuzione dell'effetto all'aumentare della dose. Le tecniche bayesiane permettono di includere questo tipo di informazioni nella stima dei parametri del modello, cosa che non è possibile fare con le tecniche frequentiste, poiché con queste ultime si utilizzano solamente i dati osservati per fare inferenza. Diventa quindi fondamentale il ruolo della distribuzione $\pi(\theta)$, che può essere più o meno informativa a seconda del grado di conoscenza a priori del processo analizzato.
Le distribuzioni a priori più utilizzate per i modelli dose-risposta sono quelle uniformi e quelle normali, univariate e indipendenti per ciascun parametro (si veda ad esempio Johnstone *et al.*, 2017 e Labelle *et al.*, 2019). In alternativa, è possibile utilizzare la distribuzione a priori di Jeffreys. Tale distribuzione è non informativa ed è definita come

$$\pi_J(\theta) \propto \sqrt{\det(I(\theta))}, \tag{2.3}$$

dove $I(\theta)$ indica la matrice di informazione attesa di Fisher, ovvero il valore atteso della matrice di informazione osservata $j(\theta)$ definita in (1.11). Ad esempio, se $y_1, ..., y_n$ sono realizzazioni indipendenti di una variabile aleatoria di Bernoulli di parametro p, si ha che

$$\begin{aligned} L(p) &= p^y(1-p)^{n-y} \\ l(p) &= y\log(p) + (n-y)\log(1-p) \\ \frac{\partial l(p)}{\partial p} &= \frac{y}{p} - \frac{n-y}{1-p} \\ j(p) &= -\frac{\partial^2 l(p)}{\partial p^2} = \frac{y}{p^2} + \frac{n-y}{(1-p)^2} \\ I(p) &= \mathbb{E}_p(j(p)) = \frac{1}{p(1-p)}. \end{aligned}$$

La distribuzione di Jeffreys per il caso preso ad esempio è quindi

$$\pi_J(p) \propto p^{-1/2}(1-p)^{-1/2},$$

ovvero è una Beta di parametri 1/2 e 1/2.

2.2.1 Analisi bayesiana del dataset `spinach`

Nel paragrafo 1.2.1 è stata effettuata una semplice analisi del dataset `spinach` per mostrare un'applicazione pratica dei modelli dose-risposta in ambito frequentista. Di seguito viene proposta l'analisi del medesimo dataset con lo scopo di mettere in risalto le peculiarità dell'approccio bayesiano. Viene assunto per i dati il modello selezionato nel Paragrafo 1.2.1; per motivi computazionali viene utilizzata la riparametrizzazione con i logaritmi di e e di σ, in modo tale da evitare vincoli sullo spazio parametrico (Sartori, 2020). Il modello risultante è

$$y_i = \frac{d}{1 + \exp(b(\log(x_i) - \bar{e}))} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \exp(\bar{\sigma})^2), \quad i = 1, ..., n,$$

con $\bar{e} = \log(e)$ e $\bar{\sigma} = \log(\sigma)$.
Nel seguito si illustrano due scenari, che si differenziano sulla base del livello di informazione a priori sui parametri; in entrambi i casi si assume che i parametri del modello (b, d, e e σ) siano a priori indipendenti. La distribuzione a posteriori, non essendo ricavabile in forma esplicita, viene ricostruita in maniera approssimata tramite simulazione; per fare ciò è stato implementato in `R` un algoritmo di Metropolis-Hastings (Metropolis *et al.*, 1953) con Random Walk uniforme su ciascuna componente del parametro $\theta = (b, d, \bar{e}, \bar{\sigma})$, che consente di simulare valori da una catena di Markov con *distribuzione limite* nota (si veda Brémaud, 2020, e Sartori, 2020). Seguendo le indicazioni presenti in letteratura, l'ampiezza della distribuzione uniforme con la quale effettuare il Random Walk è stata scelta in maniera tale da ottenere una probabilità di accettazione di poco superiore al 25%.
Nel primo scenario sono state scelte delle distribuzioni a priori normali con media pari alla stima di massima verosimiglianza e varianza pari a 10 per ciascun parametro:

$$\beta \sim N_d(\hat{\beta}, \Sigma), \quad \Sigma = diag(100, ..., 100).$$

L'alta variabilità delle distribuzioni è tipica dei casi nei quali vi è una scarsa conoscenza a priori sui parametri. Nel secondo scenario viene invece simulata una situazione in cui si conosce con poca incertezza il valore dei quattro

parametri del modello; nello specifico, vengono assunte per i parametri distribuzioni normali con media pari alla stima di massima verosimiglianza e varianza pari a 0.01:

$$\beta \sim N_d(\hat{\beta}, \Sigma), \qquad \Sigma = diag(0.01, ..., 0.01).$$

Gli istogrammi dei valori simulati dalle distribuzioni a posteriori nei due scenari sono riportati in Figura 2.1. Si noti come una minor variabilità delle distribuzioni a priori si traduca in una minor variabilità delle distribuzioni a posteriori. In Tabella 2.1 sono invece riportate le stime puntuali (medie del-

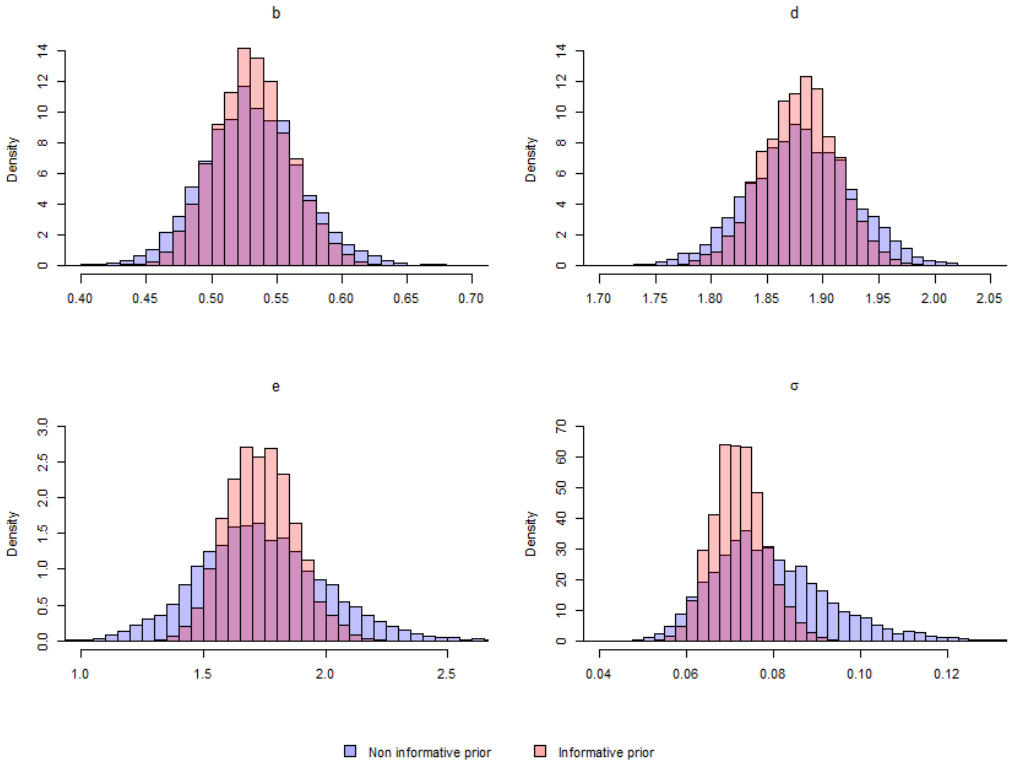


Figura 2.1: Istogrammi delle distribuzioni a posteriori simulate

le distribuzioni a posteriori) e gli intervalli di credibilità *equi-tailed* al 95% (quantili di ordine 2.5 e 97.5% delle distribuzioni a posteriori) per ciascun parametro. Si noti come le stime puntuali a posteriori dei parametri sostanzialmente coincidano con le stime di massima verosimiglianza, a prescindere

dalla variabilità della distribuzione a priori utilizzata. Per quanto riguarda le stime intervallari, gli intervalli di credibilità risultano essere tanto più larghi tanto più è variabile la distribuzione a priori.

Tabella 2.1: Stime puntuali e intervalli di credibilità per i parametri del modello

	Priori non informativa		**Priori informativa**			
	$\mathbb{E}(\theta\|\mathbf{y})$	$IC_{0.95}$	$\mathbb{E}(\theta\|\mathbf{y})$	$IC_{0.95}$	SMV	$IC_{0.95}$
b	0.532	$(0.461 - 0.610)$	0.531	$(0.476 - 0.591)$	0.530	$(0.457 - 0.604)$
d	1.879	$(1.787 - 1.975)$	1.879	$(1.812 - 1.944)$	1.879	$(1.787 - 1.971)$
e	1.726	$(1.270 - 2.321)$	1.729	$(1.481 - 2.019)$	1.733	$(1.222 - 2.244)$
σ	0.079	$(0.058 - 0.112)$	0.072	$(0.061 - 0.085)$	0.076	—

2.2.2 Analisi bayesiana dei dati sul COVID-19

Nel seguente paragrafo viene presentata l'analisi dei dati sul COVID-19, effettuata utilizzando gli strumenti forniti dalla statistica bayesiana. Come già visto nel Capitolo 1, viene considerata come variabile di interesse il numero cumulato di deceduti giornaliero, e sono utilizzati i dati dal 24 febbraio 2020 al 15 febbraio 2021. Vengono inoltre ripresi gli assunti presentati nel primo capitolo: i dati sono suddivisi in due ondate e viene fissato il 15 agosto 2020 come termine della prima ondata. Considerata la selezione del modello presentata nel paragrafo 1.4, per entrambe le ondate si assume

$$y_i = c + \frac{d-c}{(1+\exp(b(\log(x_i) - \log(e))))^f} + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2), \quad i = 1, ..., n,$$

con il vincolo $c = 0$ per i dati della prima ondata. Analogamente a quanto visto per l'analisi del dataset `spinach`, la distribuzione a posteriori viene ricostruita simulando i valori tramite un algoritmo di Metropolis-Hastings. Per ciascuna ondata sono state utilizzate due diverse distribuzioni a priori: la prima è totalmente non informativa; come seconda distribuzione a priori viene invece scelta una normale univariata per ciascun parametro, centrata nella stima di massima verosimiglianza e con deviazione standard pari alla deviazione standard della stima di massima verosimiglianza. La scelta delle distribuzioni a priori è stata effettuata puramente a scopo illustrativo, poiché l'obiettivo dell'analisi è mostrare le metodologie nell'ambito dei modelli

dose-risposta bayesiani e non quello di trovare risultati innovativi e accurati nella descrizione della diffusione della pandemia.
In Figura 2.2 sono rappresentati gli istogrammi dei valori simulati dalla distribuzione a posteriori per i parametri b, d, e e f nella prima ondata. Come si può notare, le distribuzioni a priori informative portano a distribuzioni a posteriori con minor variabilità. Le stime bayesiane dei parametri, sia puntuali

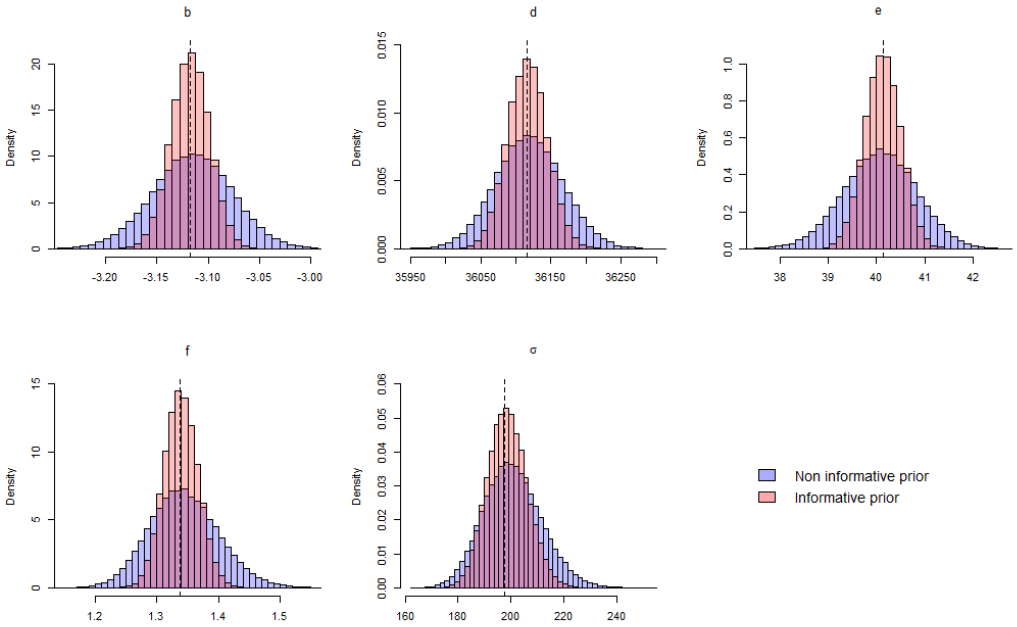


Figura 2.2: Istogrammi delle distribuzioni a posteriori simulate: prima ondata. Linea tratteggiata: stima di massima verosimiglianza

sia intervallari, sono riportate in Tabella 2.2. Come ci si poteva aspettare dai grafici in Figura 2.2, le stime puntuali, che corrispondono alla media dei valori simulati dalla distribuzione a posteriori, sono sostanzialmente identiche nel caso in cui la distribuzione a priori sia informativa e nel caso in cui non lo sia e, inoltre, coincidono con la stima di massima verosimiglianza. Risultano invece più stretti gli intervalli di credibilità ottenuti con distribuzione a priori informativa.
Per quanto riguarda la seconda ondata, la Figura 2.3 mostra come le distribuzioni a posteriori dei parametri (b, c, d, e, e f) siano molto più irregolari

Tabella 2.2: Stime puntuali e intervalli di credibilità per i parametri del modello: prima ondata

	Priori non informativa		**Priori informativa**			
	$\mathbb{E}(\theta\|\mathbf{y})$	$IC_{0.95}$	$\mathbb{E}(\theta\|\mathbf{y})$	$IC_{0.95}$	SMV	$IC_{0.95}$
b	-3.12	(-3.19 – -3.04)	-3.12	(-3.15 – -3.08)	-3.12	(-3.19 – -3.04)
d	36119.06	(36025.65 – 36213.86)	36116.50	(36061.06 – 36172.15)	36116.16	(36022.48 – 36209.85)
e	40.06	(38.56 – 41.54)	40.12	(39.39 – 40.86)	40.13	(38.66 – 41.61)
f	1.34	(1.24 – 1.46)	1.34	(1.29 – 1.39)	1.34	(1.23 – 1.45)
σ	200.66	(180.58 – 233.51)	198.62	(184.38 – 213.84)	197.50	—

rispetto alle medesime distribuzioni per la prima ondata. Risulta quindi difficile trovare una relazione tra la variabilità delle distribuzioni a priori e la forma delle distribuzioni a posteriori. Si noti come, nonostante le distribuzioni a priori siano centrate nella stima di massima verosimiglianza, le distribuzioni a posteriori non lo siano (ad eccezione di quelle del parametro c); di conseguenza, come riportato in Tabella 2.3, gli intervalli di credibilità in alcuni casi non comprendono la stima di massima verosimiglianza. Inoltre, per i parametri e e f, la distribuzione a posteriori ha una variabilità minore quando è ricavata a partire da una distribuzione a priori meno informativa: per questi parametri si hanno quindi intervalli di credibilità più larghi quando la distribuzione a priori ha minor variabilità. Questi risultati anomali sono una conseguenza della verosimiglianza poco regolare a causa della scarsa attendibilità dei dati, dovuta sia ad una mancanza accuratezza in fase di raccolta sia alla sovrapposizione con un'eventuale terza ondata che non è stata considerata in questa tesi.

Tabella 2.3: Stime puntuali e intervalli di credibilità per i parametri del modello: seconda ondata

	Priori non informativa		**Priori informativa**			
	$\mathbb{E}(\theta\|\mathbf{y})$	$IC_{0.95}$	$\mathbb{E}(\theta\|\mathbf{y})$	$IC_{0.95}$	SMV	$IC_{0.95}$
b	-2.63	(-2.71 – -2.56)	-2.68	(-2.74 – -2.63)	-2.70	(-2.81 – -2.59)
c	35661.25	(35548.72 – 35773.53)	35657.92	(35576.03 – 35739.43)	35651.17	(35533.49 – 35768.85)
d	117719.9	(115926.9 – 119539.6)	116880.0	(115661.6 – 118088.3)	116736.0	(114748.9 – 118723.0)
e	40.35	(37.27 – 44.84)	44.34	(40.23 – 49.70)	47.88	(39.00 – 56.76)
f	19.50	(15.32 – 22.77)	15.98	(12.10 – 19.69)	13.17	(7.66 – 18.68)
σ	470.03	(424.16 – 523.27)	475.02	(440.41 – 511.60)	477.44	—

Le analisi effettuate finora hanno come unico scopo l'illustrare le tecniche

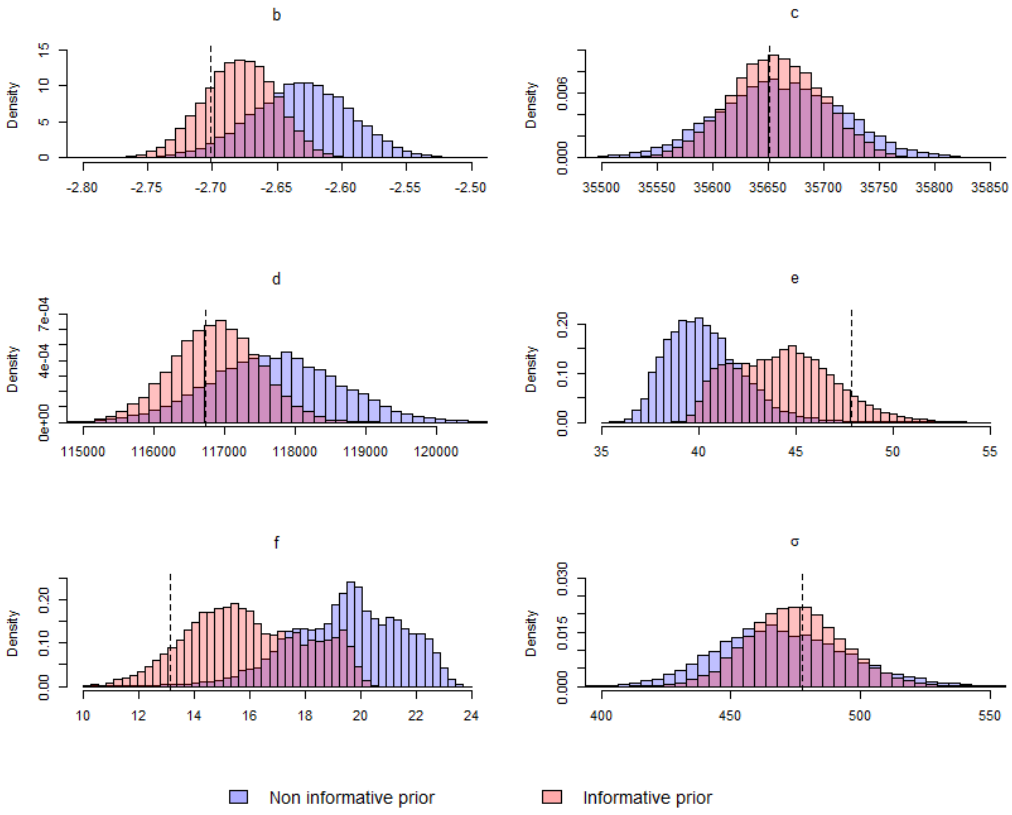


Figura 2.3: Istogrammi delle distribuzioni a posteriori simulate: seconda ondata. Linea tratteggiata: stima di massima verosimiglianza

di inferenza per i modelli dose-risposta sia in ambito frequentista sia in ambito bayesiano, utilizzando i dati sul COVID-19 a titolo esemplificativo. A partire dal capitolo successivo ci si focalizzerà maggiormente sull'analisi dei dati sul COVID-19 considerando in maniera approfondita le questioni tralasciate finora, come l'identificazione del momento del passaggio tra la prima e la seconda ondata e la scelta delle distribuzioni a priori per i parametri del modello.

Capitolo 3

Modelli *change-point*

Nei primi due capitoli di questa tesi sono stati presentati i modelli dose-risposta, inizialmente con la descrizione della struttura e lo studio dell'inferenza sia nel contesto frequentista sia nel contesto bayesiano, e successivamente con l'applicazione pratica di tali modelli su due dataset di riferimento. I seguenti due capitoli si focalizzano invece sull'analisi dei dati sul COVID-19 in Italia, attraverso la proposta di strumenti e idee specifici per questa tipologia di dataset. Vengono quindi presentati i cosiddetti *modelli change-point*, particolarmente adatti a descrivere i dati relativi a una pandemia che si sviluppa in più ondate. Più in generale, i modelli *change-point* vengono utilizzati quando si lavora con una sequenza di variabili casuali identicamente distribuite fino ad un certo istante τ, detto per l'appunto *change-point* (si veda, ad esempio, Qiu, 2014). A partire da quell'istante le variabili casuali sono sempre identicamente distribuite, ma con distribuzione diversa da quella di partenza. Le distribuzioni prima e dopo il *change-point* possono appartenere a due differenti famiglie di distribuzioni, così come possono appartenere alla stessa famiglia di distribuzioni, ma avere parametri diversi. All'interno di un'analisi basata su modelli di questo tipo, uno dei principali obiettivi è proprio l'individuazione del *change-point*. Questo compito assume notevole importanza nell'ambito in cui sono nati i modelli *change-point*: il *Controllo Statistico di Processo*, ovvero il monitoraggio di processi sequenziali per assicurarsi che tali processi funzionino stabilmente ed in maniera soddisfacente (Qiu, 2014).

In aggiunta all'utilizzo dei modelli *change-point*, si vuole condurre un'analisi che tenga in considerazione le problematiche presentate nei capitoli precedenti, ovvero la presenza di autocorrelazione e sovradispersione, così come la scarsa accuratezza in fase di raccolta dati.

3.1 Il contesto

La pandemia di COVID-19, così come altre pandemie che hanno colpito l'umanità nel corso della storia, si è diffusa in diverse ondate. In Italia, da febbraio 2020 fino a maggio 2021, si sono registrate tre ondate, come riportato da diversi media e autorità sanitarie (si veda, ad esempio, il bollettino di aggiornamento nazionale pubblicato dall'Istituto Superiore di Sanità in data 9 aprile 2021 https://www.epicentro.iss.it/coronavirus/bollettino/Bollettino-sorveglianza-integrata-COVID-19_7-aprile-2021.pdf).
Per comprendere meglio le caratteristiche delle tre ondate viene riportato il grafico in Figura 3.1, in cui è rappresentata la media mobile di ordine 7 del numero di deceduti giornaliero, ovvero

$$mm_i = \frac{1}{7} \sum_{j=i-3}^{i+3} y_j,$$

dove y_j è il numero di deceduti giornaliero. A differenza delle analisi effettuate nei capitoli precedenti, per questo grafico sono stati utilizzati i dati giornalieri, ovvero il numero di nuove morti giorno per giorno, e non i dati cumulati. Si noti come sia possibile ricavare i dati giornalieri da quelli cumulati semplicemente calcolandone la differenza prima: indicando con y_i i dati giornalieri e con y_i^C i dati giornalieri cumulati, vale

$$y_i = y_i^C - y_{i-1}^C, \qquad i = 2, ..., n.$$

Sfruttando lo sviluppo in serie di Taylor, è inoltre possibile ricavare la media per i dati non cumulati come la derivata prima della media $\mu(x, \beta)$ per i dati cumulati. Vale infatti

$$\mu(x, \beta) = \mu(x_0, \beta) + \mu'(x_o, \beta)(x - x_0) + o(|x - x_0|),$$

da cui $\mu'(x-1,\beta) \approx \mu(x,\beta) - \mu(x-1,\beta)$.
L'utilizzo dei dati non cumulati consente di cogliere altre caratteristiche rispetto a quelle evidenziate finora: ad esempio, il picco è stato raggiunto rispettivamente a fine marzo 2020, a fine novembre 2020 e ad inizio aprile 2021; si noti inoltre come la prima ondata sia quella che ha raggiunto il picco più elevato, anche se la seconda ha provocato un maggior numero di deceduti. Infine è possibile descrivere l'andamento generale del numero di morti giornalieri per COVID-19: al primo rapido aumento del mese di marzo 2020 è seguito un altrettanto rapido calo nei due mesi successivi; nel corso di tutta l'estate si è registrato un numero esiguo di morti giornaliere, mentre a partire da ottobre la curva è tornata a salire rapidamente, per raggiungere il picco di novembre. Da novembre 2020 a febbraio 2021 la curva è tornata a scendere, ad eccezione del periodo compreso tra metà dicembre e metà gennaio, in cui si sono registrati costantemente circa 500 morti giornalieri. Infine, a partire da metà febbraio 2021 il numero di deceduti giornalieri è tornato ad aumentare, ha raggiunto il picco ad inizio aprile ed è poi calato nel periodo successivo. Si noti come la terza ondata sia iniziata quando la seconda ondata era ancora lontana dal giungere al termine: nel momento in cui la curva è tornata a salire si registravano infatti oltre 250 morti giornalieri; c'è quindi stato un ampio periodo in cui le due ondate sono state sovrapposte, cosa che non è accaduta tra la prima e la seconda ondata.
A prescindere dall'utilizzo dei dati cumulati o giornalieri, è necessario introdurre degli strumenti specifici per descrivere lo sviluppo su più ondate della pandemia. Singhal *et al.* (2020) propongono l'utilizzo di un modello mistura gaussiano, che ben si adatta all'andamento plurimodale del numero di decessi giornaliero; il modello è stimato sui dati giornalieri non cumulati e prevede la sovrapposizione delle tre ondate su tutto il periodo considerato: ciò significa che ogni ondata contribuisce ai decessi di ciascun giorno, anche se in maniera trascurabile per i giorni più distanti dal picco.
Un'alternativa ai modelli mistura è costituita dai *modelli change-point*, tema centrale del prosieguo della tesi. Per semplicità, ci si limita a presentare il modello per due ondate, e quindi con un solo *change-point*, che presenta la

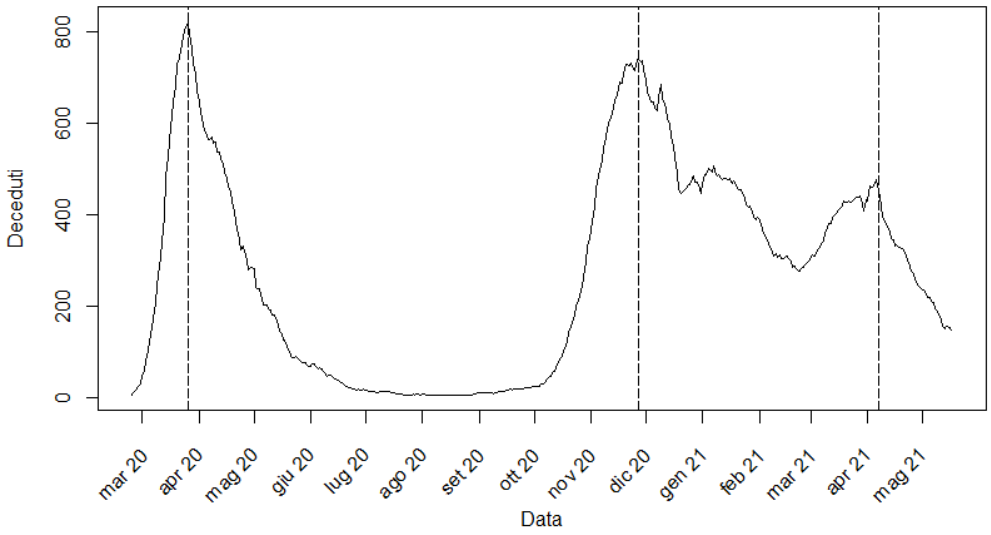


Figura 3.1: Andamento del numero giornaliero di deceduti: media mobile di 7 giorni

seguente struttura:

$$\mu(x,\beta) = \begin{cases} \mu(x,\beta_1), & x \leq \tau \\ \mu(x-\tau,\beta_2), & x > \tau \end{cases}. \tag{3.1}$$

Il parametro θ è quindi costituito da $\beta = (\beta_1, \beta_2)$ e da τ; i vari β_i possono poi a loro volta essere formati da più parametri: ad esempio, se μ è la funzione log-logistica a 5 parametri allora $\beta_i = (b_i, c_i, d_i, e_i, f_i)$, $i = 1, 2$. Nel seguito verranno stimati i modelli *change-point* sui dati cumulati, ma ciò non toglie che possano essere stimati anche sui dati giornalieri.
Si noti come, a differenza dei modelli mistura, nei modelli *change-point* le ondate non sono sovrapposte ma separate e l'inferenza riguarda anche l'istante τ in cui avviene il passaggio tra un'ondata e quella successiva; ne segue che i decessi di ciascun giorno sono attribuiti interamente ad un'unica ondata. Nella parte restante del capitolo viene introdotta la teoria sull'inferenza per i modelli appena presentati; .

3.2 Inferenza frequentista

Siano $Y_1, ..., Y_n$ una sequenza di variabili aleatorie e sia $\tau \in \{1, ..., n-1\}$ tale per cui

$$\begin{cases} p_{Y_i}(y_i; \beta_1), & i = 1, ..., \tau \\ p_{Y_i}(y_i; \beta_2), & i = \tau + 1, ..., n \end{cases}, \quad \beta_1 \neq \beta_2,$$

dove τ è detto *change-point* e costituisce, insieme a β_1 e β_2, il parametro di interesse. Utilizzando un approccio di tipo frequentista le procedure di inferenza si basano sulla *independence likelihood* (Greco *et al.*, 2021, *Submitted*), che assume la forma

$$L_I(\theta) = \prod_{i=1}^{n} p(y_i; \theta_1)^{z_{1i}} \cdot p(y_i; \theta_2)^{1-z_{1i}}, \tag{3.2}$$

$$z_{1i} = \begin{cases} 1 & i = 1, ..., \tau \\ 0 & i = \tau + 1, ..., n \end{cases},$$

da cui si ricava la *independence log-likelihood*

$$l_I(\theta) = \sum_{i=1}^{n} z_{1i} \log p(y_i; \theta_1) + (1 - z_{1i}) \log p(y_i; \theta_2). \tag{3.3}$$

Definendo $\beta = (\beta_1, \beta_2)$, la stima di massima verosimiglianza composita di $\theta = (\beta, \tau)$ avviene attraverso la massimizzazione della *profile independence likelihood* per τ, definita come

$$l_{I_P}(\tau) = l_I(\hat{\beta}^I_\tau, \tau), \tag{3.4}$$

dove $\hat{\beta}^I_\tau$ è la stima vincolata di massima verosimiglianza di β per τ fissato (Greco *et al.*, 2021, *Submitted*). Si ottiene così

$$\hat{\tau}^I = \underset{\tau}{\operatorname{argmax}}\, l_{I_P}(\tau)$$

e la stima di massima verosimiglianza composita per θ è

$$\hat{\theta}^I = (\hat{\beta}^I_\tau, \hat{\tau}^I).$$

Per costruire intervalli di confidenza ed effettuare test di verifica di ipotesi è necessario ricavare gli standard error delle componenti dei $\hat{\beta}^I_j$ per

$j = 1, 2$; ciò può essere fatto utilizzando la distribuzione asintotica di $\hat{\beta}_j^I$ condizionatamente a $\hat{\tau}^I$, vale a dire

$$\hat{\beta}_j^I \dot{\sim} N_p\left(\beta_j, \sqrt{Var(\hat{\beta}_j^I)}\right). \tag{3.5}$$

Per ciascuno dei due blocchi in cui sono divise le osservazioni, una stima della matrice di covarianze $Var(\hat{\beta}_j^I)$ può essere ottenuta come l'inversa della matrice di Godambe definita in (1.25): è quindi necessaria una stima delle matrici $H(\beta_j)$ e $J(\beta_j)$. In generale, se la numerosità campionaria n è sufficientemente grande, è possibile stimare le due matrici nella seguente maniera:

$$\hat{H}(\beta) = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial u_I(y_i;\beta)}{\partial \beta^T}\Big|_{\beta=\hat{\beta}^I}$$

$$\hat{J}(\beta) = \frac{1}{n}\sum_{i=1}^{n} u_I(y_i;\hat{\beta}^I) u_I(y_i;\hat{\beta}^I)^T,$$

con $u_I(y_i;\beta) = \partial l_I(\beta)/\partial\beta$. In alternativa, qualora sia possibile simulare dal modello, le matrici $H(\beta)$ e $J(\beta)$ possono essere ricavate tramite metodi di simulazione Monte Carlo (Greco *et al.*, 2021, *Submitted*). Si noti come, nel caso del modello *change-point*, il concetto di numerosità campionaria sufficientemente elevata debba essere valido per ogni intervallo in cui il *change-point* divide le osservazioni. Si rimanda il lettore a Varin *et al.* (2011) per un'analisi più approfondita e per la discussione sulla stima di $H(\beta)$ e $J(\beta)$ nel caso di scarsa numerosità campionaria.
La scelta di una funzione punteggio diversa da $u_I(y_i;\beta)$, definita come la derivata della *independence log-likelihood*, permette di effettuare procedure di inferenza più robuste: ad esempio è possibile utilizzare la derivata della funzione score di Tsallis definita nel Capitolo 1; tutte le procedure appena descritte rimangono inalterate, è sufficiente sostituire $l_I(\theta)$ con $S(\theta)$ definita in (1.26).

3.3 Inferenza bayesiana

Nel capitolo 2 è stata presentata l'inferenza in ambito bayesiano, con particolare riguardo al suo utilizzo nei modelli dose-risposta. Questo paragrafo

ha l'obiettivo di ampliare la presentazione dell'approccio bayesiano ai modelli con un *change-point*. Si considerino, come nel caso frequentista, una sequenza di variabili aleatorie $Y_1, ..., Y_n$ con distribuzione

$$\begin{cases} p_{Y_i}(y_i; \beta_1), & i = 1, ..., \tau \\ p_{Y_i}(y_i; \beta_2), & i = \tau + 1, ..., n \end{cases}, \quad \beta_1 \neq \beta_2,$$

dove β_1 è quindi il parametro β che caratterizza la distribuzione del primo gruppo di variabili, mentre β_2 è lo stesso parametro per il secondo gruppo. L'obiettivo dell'inferenza con un approccio bayesiano è derivare la distribuzione a posteriori del parametro $\theta = (\beta_1, \beta_2, \tau)$ a partire dalla distribuzione a priori e dalla verosimiglianza, o da una funzione che svolge un ruolo analogo, ad esempio lo score di Tsallis. Per quanto riguarda la discussione sulla verosimiglianza in presenza di un *change-point* si vedano le considerazioni del paragrafo precedente, mentre per la scelta della distribuzioni a priori per β si veda il Capitolo 2. Per riguarda τ, invece, una a priori uniforme costituisce la distribuzione non informativa che verrà utilizzata in seguito, mentre verrà considerata anche una distribuzione normale per gli scenari in cui si vuole considerare a priori informative.

Capitolo 4

Applicazione ai dati sul COVID-19

Nei capitoli precedenti sono stati illustrati alcuni modelli statistici adatti a descrivere la diffusione del COVID-19. In particolare, l'andamento del numero cumulato di deceduti giornaliero è ben approssimato da una funzione log-logistica a 5 parametri, mentre con i modelli *change-point* è possibile rappresentare l'evoluzione della pandemia su più ondate. Infine, l'utilizzo dell'*independence likelihood* permette di ovviare al problema dell'autocorrelazione presente nei dati.
In questo capitolo viene costruito un modello riassuntivo specifico per descrivere l'andamento del numero cumulato di decessi giornalieri per COVID-19 in Italia. Il modello completo è un modello non lineare normale con due *change-point*:

$$Y_i \sim N(\mu(x_i,\beta),\sigma^2), \quad \mu(x_i,\beta) = \begin{cases} \mu(x_i,\beta_1), & x_i \le \tau_1 \\ \mu(x_i-\tau_1,\beta_2), & \tau_1 < x_i \le \tau_2 \\ \mu(x_i-\tau_2,\beta_3), & x_i > \tau_2 \end{cases}, \quad i=1,...,n, \tag{4.1}$$

dove la funzione media $\mu(\cdot)$ è la log-logistica a 5 parametri (1.3). L'*independence likelihood* è

$$L_I(\theta) = \prod_{i=1}^{n} p(y_i;\theta_1)^{z_1} \cdot p(y_i;\theta_2)^{z_2} \cdot p(y_i;\theta_3)^{1-z_1-z_2}, \tag{4.2}$$

$$z_1 = \begin{cases} 1 & x_i \le \tau_1 \\ 0 & altrimenti \end{cases} \qquad z_2 = \begin{cases} 1 & \tau_1 < x_i \le \tau_2 \\ 0 & altrimenti \end{cases},$$

da cui si ricava l'*independence log-likelihood*

$$l_I(\theta) = \sum_{i=1}^{n} z_1 \log p(y_i;\theta_1) + z_2 \log p(y_i;\theta_2) + (1 - z_1 - z_2) \log p(y_i;\theta_3). \quad (4.3)$$

Le analisi svolte e presentate in seguito si concentrano solamente sulle prime due ondate e sull'inferenza relativa al *change-point* τ_1 che le divide, che per comodità verrà chiamato τ da ora in avanti. Il modello utilizzato è dunque

$$Y_i \sim \begin{cases} N(\mu_1(x_i,\beta_1),\sigma_1^2), & x_i \leq \tau \\ N(\mu_2(x_i-\tau,\beta_2),\sigma_2^2), & x_i > \tau \end{cases}, \qquad i = 1,...,n, \qquad (4.4)$$

con

$$\mu_1(x,\beta_1) = \frac{d_1}{(1+\exp(b_1(\log(x)-\log(e_1))))^{f_1}}$$

$$\mu_2(x,\beta_2) = c_2 + \frac{d_2 - c_2}{(1+\exp(b_2(\log(x)-\log(e_2))))^{f_2}}.$$

Si è scelto di lavorare con la media mobile di ordine 7 del numero di deceduti giornaliero, ovvero

$$\bar{y}_i = \frac{1}{7}\sum_{j=i-3}^{i+3} y_j,$$

dove y_j è il numero di deceduti giornaliero, in modo tale da annullare l'effetto dei trend settimanali presenti nei dati originali.
La peculiarità dell'analisi condotta è l'uso di un approccio bayesiano, che costituisce un'innovazione nello studio dei modelli *change-point* per descrivere il passaggio da un'ondata della pandemia a quella successiva.

4.1 Analisi delle distribuzioni a posteriori

La ricostruzione delle distribuzioni a posteriori è stata effettuata tramite l'algoritmo di Metropolis Hastings, implementato in `R` nel pacchetto `MCMCpack` e impostato in maniera tale da ottenere una probabilità di accettazione compresa tra il 20 e il 50%.
Vengono presentate 3 situazioni che si differenziano in base al grado di informatività delle distribuzioni a priori. Nel primo caso è stata utilizzata una distribuzione a priori totalmente non informativa, con la distribuzione a posteriori che risulta quindi proporzionale alla verosimiglianza. Nel secondo caso

è stata utilizzata, per ciascun parametro, una distribuzione a priori normale centrata nella stima di massima verosimiglianza e con deviazione standard pari a 3 volte lo standard error della stima di massima verosimiglianza; fa eccezione il parametro relativo al *change-point*, per il quale si è scelta una distribuzione a priori uniforme continua su un intervallo molto ampio (corrispondente all'intervallo di date compreso tra il 22 giugno 2020 e il 19 dicembre 2020). Si tratta quindi di distribuzioni quasi per nulla informative, che riflettono una scarsa conoscenza a priori del fenomeno studiato. Infine, nel terzo caso sono state scelte distribuzioni a priori molto informative, propendendo per distribuzioni normali centrate nella stima di massima verosimiglianza e con deviazione standard pari allo standard error della stima di massima verosimiglianza. Per il parametro τ è stata invece utilizzata una distribuzione uniforme continua sull'intervallo che corrisponde alle date che vanno dall'11 agosto 2020 al 20 ottobre 2020.
In Figura 4.1 sono riportati gli istogrammi delle distribuzioni a posteriori per i parametri della curva log-logistica per la prima ondata, sia con la distribuzione a priori non informativa sia con la distribuzione a priori molto informativa. Come era logico attendersi, le distribuzioni a priori più informative portano a distribuzioni a posteriori meno variabili.
In Figura 4.2 sono riportati gli istogrammi per gli analoghi parametri della seconda ondata. Si noti come, in questo caso, le distribuzioni a posteriori non siano centrate negli stessi punti e non presentino la classica forma a campana, soprattutto quando la distribuzione a priori è non informativa. Inoltre, non sembra esserci una relazione tra la variabilità delle distribuzioni a priori e la variabilità delle distribuzioni a posteriori.
Il grafico più interessante è riportato in Figura 4.3, dove sono rappresentate le distribuzioni a posteriori del parametro τ, ossia del *change-point*. Si noti come utilizzando una distribuzione a priori molto informativa la distribuzione a posteriori sia concentrata nel periodo di fine agosto, mentre se non vengono considerate informazioni a priori sul *change-point* la distribuzione a posteriori non è unimodale ma presente un picco minore a fine agosto e uno decisamente maggiore intorno alla metà di settembre. In tutte le figure non sono state riportate le distribuzioni a posteriori generate con le distribuzioni a priori poco informative, allo scopo di rendere i grafici più facilmente

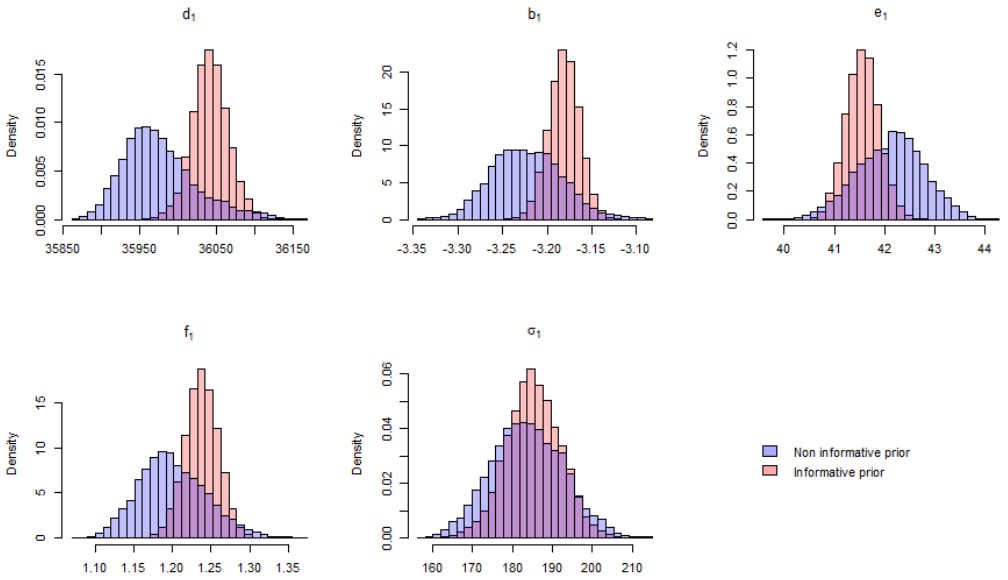


Figura 4.1: Istogrammi delle distribuzioni a posteriori simulate: prima ondata.

leggibili. Tali distribuzioni risultano comunque molto simili alle distribuzioni a posteriori ottenute dalle a priori non informative. Nonostante la forma irregolare, la distribuzione a posteriori assume valori coerenti con i risultati presentati nei primi due capitoli. In particolare,

In Tabella 4.1 sono riportate le stime e gli intervalli di credibilità per il *change-point* τ nei tre casi considerati; si noti come le distribuzioni a priori non informative e quelle poco informative forniscano sostanzialmente le medesime stime per τ, sia puntuali sia intervallari. Come si poteva evincere dai grafici, la stima di τ in mancanza di informazioni a priori cade intorno a metà settembre (11/09), mentre la stessa stima ottenuta con una distribuzione a priori molto informativa è fissata alla fine di agosto (24/08). Per quanto riguarda gli intervalli di credibilità, essi risultano più stretti quando la distribuzione a priori utilizzata è molto informativa.

In *Misspecified modeling of subsequent waves during COVID-19 outbreak: A change point growth model* (Greco *et al.*, 2021, *Submitted*) vengono presentati i risultati ottenuti tramite un approccio frequentista. Gli autori hanno lavorato con i dati originali anziché con la media mobile di ordine 7 e hanno

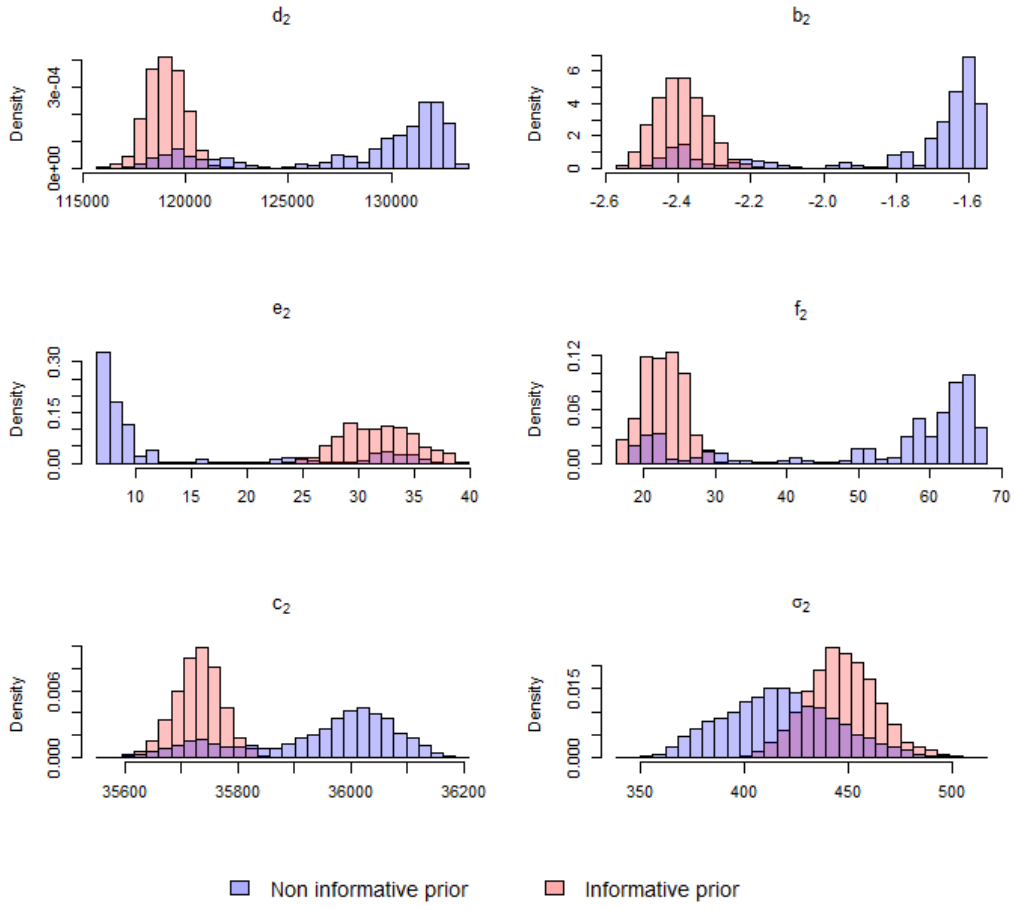


Figura 4.2: Istogrammi delle distribuzioni a posteriori simulate: seconda ondata.

Tabella 4.1: Stime puntuali e intervalli di credibilità per il *change-point* τ.
Le date sono in formato giorno/mese

Distribuzione a priori	$\mathbb{E}(\tau\|\mathbf{y})$	$IC_{0.95}$ (quantili)	$IC_{0.95}$ (HPD)
Priori non informativa	11/09	(22/08 – 18/09)	(22/08 – 18/09)
Priori poco informativa	11/09	(23/08 – 17/09)	(24/08 – 17/09)
Priori molto informativa	24/08	(21/08 – 28/08)	(21/08 – 27/08)

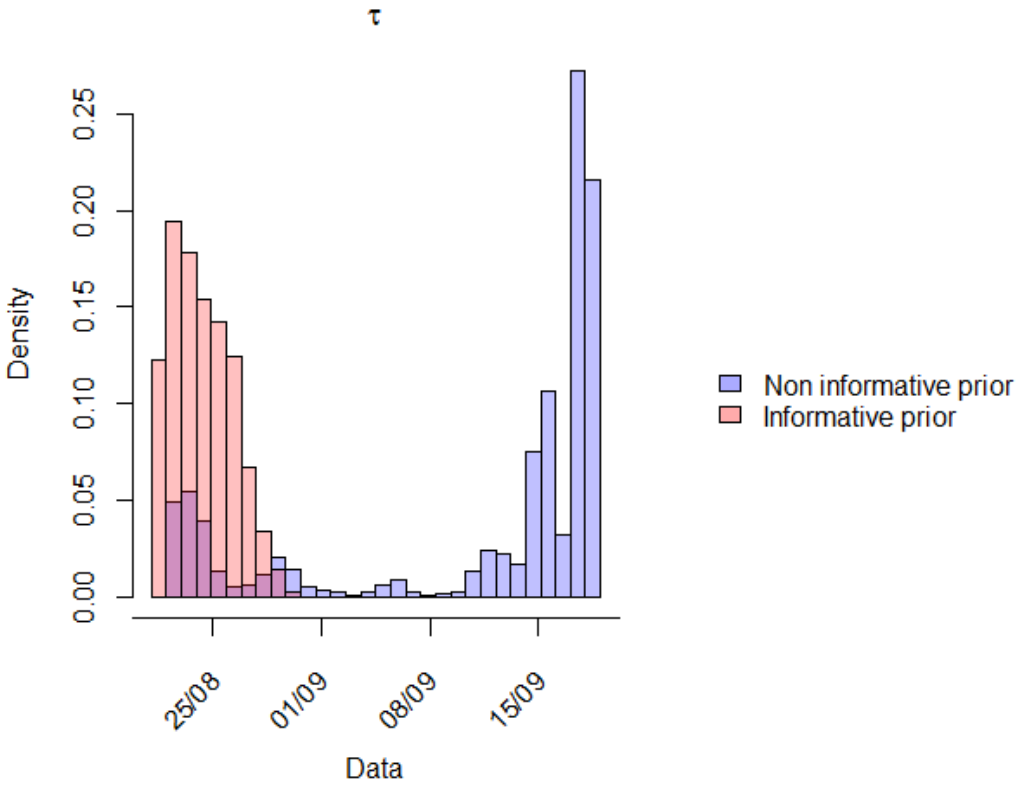


Figura 4.3: Istogrammi delle distribuzioni a posteriori del *change-point*

assunto il modello di Poisson invece di quello normale, proponendo due diverse vie per contrastare la sovradispersione e l'autocorrelazione presenti nei dati (si veda il Paragrafo 1.5 di questa tesi per i dettagli). Il *change-point* viene stimato al 12 agosto, con gli intervalli di confidenza che vanno dal 30 luglio al 25 agosto e dall'11 luglio al 12 settembre, a seconda della soluzione utilizzata per contrastare le problematiche appena presentate. La stima del *change-point* con l'approccio frequentista risulta quindi antecedente rispetto a quella ottenuta con l'approccio bayesiano. Inoltre, gli intervalli di confidenza frequentisti, a differenza degli intervalli di credibilità bayesiani, sono simmetrici e si espandono fino a metà/fine luglio.

4.1.1 Verosimiglianza profilo generalizzata

I risultati presentati finora sono basati su una distribuzione a posteriori ottenuta partendo dalla verosimiglianza (3.2), eventualmente moltiplicata per un'opportuna distribuzione a priori. La verosimiglianza coinvolge però numerosi parametri correlati, ed è perciò molto irregolare e frastagliata; ne consegue che, soprattutto per i parametri della seconda ondata, pure le distribuzioni a posteriori abbiano una forma irregolare, come si può notare dai grafici.
Nel seguito viene percorsa una via alternativa, un approccio bayesiano ibrido basato sull'utilizzo della *log-verosimiglianza profilo generalizzata* per τ, motivato dal ruolo primario del *change-point* rispetto a quello degli altri parametri. Per un approfondimento sull'utilizzo di verosimiglianze alternative in ambito bayesiano si rimanda a Racugno *et al.* (2010) e a Ventura e Racugno (2016); un'introduzione e una giustificazione all'utilizzo della verosimiglianza profilo generalizzata si veda, ad esempio, Severini (1998), Severini e Wong (1992), e Pace e Salvan (2006). La log-verosimiglianza profilo generalizzata per τ è definita come la funzione di log-verosimiglianza per τ con i restanti parametri stimati, in questo caso fissati pari alle stime di massima verosimiglianza. La forma assunta dalla *profile independence likelihood* in (3.4) viene quindi modificata in

$$l_{I_P}(\tau) = l_I(\hat{\beta}^I, \tau),$$

dove $\hat{\beta}^I$ è la stima di massima verosimiglianza per β non vincolata a τ. Anche in questo caso vengono considerati 3 scenari a seconda del grado di informatività della distribuzione a priori; in particolare, sono state scelte:

- $\pi(\tau) \sim U[1; n-1]$, dove $n-1$ è il massimo valore ipotetico per τ, ovvero il penultimo giorno presente nel dataset.Si tratta quindi di una distribuzione uniforme su tutti i possibili valori per il *change-point* ed è di conseguenza una *distribuzione a priori non informativa*;
- $\pi(\tau) \sim N(\hat{\tau}, 10)$, ovvero una *distribuzione a priori poco informativa* a causa dell'elevata variabilità;
- $\pi(\tau) \sim N(\hat{\tau}, 2.5)$ ovvero una *distribuzione a priori molto informativa* a causa della ridotta variabilità.

Si noti come, considerando una log-verosimiglianza profilo, la distribuzione a priori è univariata per l'unico parametro τ. Le distribuzioni a posteriori vengono ricostruite con le Catene di Markov che, tramite l'algoritmo di Metropolis-Hastings, consentono di ottenere una sequenza di valori estratti da una distribuzione nota; per ciascuno dei 3 scenari, sono simulati dei campioni di 50000 osservazioni dalla distribuzione a posteriori.
In Figura 4.4 sono messi a confronto gli istogrammi dei valori della distribuzione a posteriori negli scenari con distribuzione a priori non informativa e molto informativa; come in precedenza, si è scelto di non rappresentare anche i valori ottenuti utilizzando la distribuzione a priori poco informativa per rendere il grafico più leggibile, anche in considerazione del fatto che tali valori sono molto simili a quelli ottenuti con la a priori non informativa. In Tabella 4.2 sono riportate le stime puntuali e intervallari per il *change-point* nei tre scenari considerati. In tutti e tre i casi la media della distribuzione a posteriori è il 22 agosto, risultato coerente con la moda dell'istogramma in Figura 4.4.
Il grafico e la tabella evidenziano le differenze rispetto ai risultati ricavati lavorando con la verosimiglianza. L'utilizzo della verosimigilanza profilo generalizzata consente di ottenere distribuzioni a posteriori più regolari e le cui variabilità sono strettamente correlate con le variabilità delle rispettive distribuzioni a priori. Come conseguenza, gli intervalli di credibilità per il

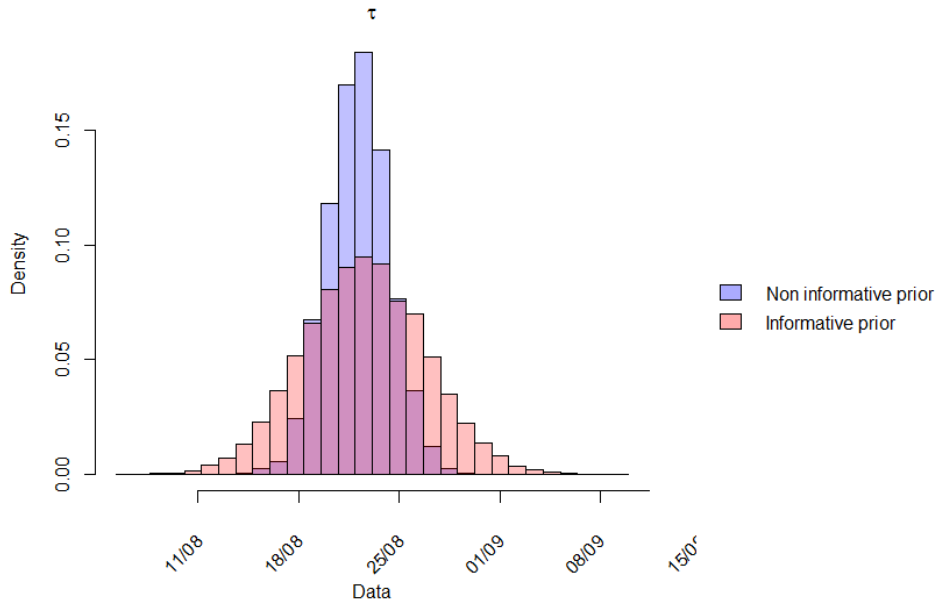


Figura 4.4: Istogrammi delle distribuzioni a posteriori del *change-point* τ: risultati ottenuti con la verosimiglianza profilo generalizzata.

Tabella 4.2: Stime puntuali e intervalli di credibilità per il *change-point* τ: risultati ottenuti con la verosimiglianza profilo generalizzata.
Le date sono in formato giorno/mese

Distribuzione a priori	$\mathbb{E}(\tau\|\mathbf{y})$	$IC_{0.95}$ (*equi-tailed*)	$IC_{0.95}$ (HPD)
Priori non informativa	22/08	(14/08 – 30/08)	(14/08 – 30/08)
Priori poco informativa	22/08	(15/08 – 30/08)	(14/08 – 29/08)
Priori molto informativa	22/08	(18/08 – 26/08)	(17/08 – 26/08)

change-point sono più simmetrici e, nel caso delle distribuzioni a priori non o poco informative, anche più corti di circa una settimana. La differenza principale tra i due metodi sta però nella stima puntuale del *change-point*, e di conseguenza nella posizione degli intervalli di credibilità: le medie delle distribuzioni a posteriori ottenute a partire dalla verosimiglianza profilo generalizzata cadono al 22 agosto per tutti e 3 gli scenari, ovvero circa 20 giorni prima rispetto a quanto ottenuto con la verosimiglianza e distribuzioni a priori non o poco informative. Le differenze sono meno marcate quando la distribuzione a priori considerata è molto informativa.

4.1.2 Studio di simulazione

Per validare la metodologia proposta e i risultati ottenuti, si è scelto di proseguire con uno studio di simulazione. Il modello di riferimento da cui si è simulato ha la forma (4.4), con i valori dei parametri pari alle stime di massima verosimiglianza:

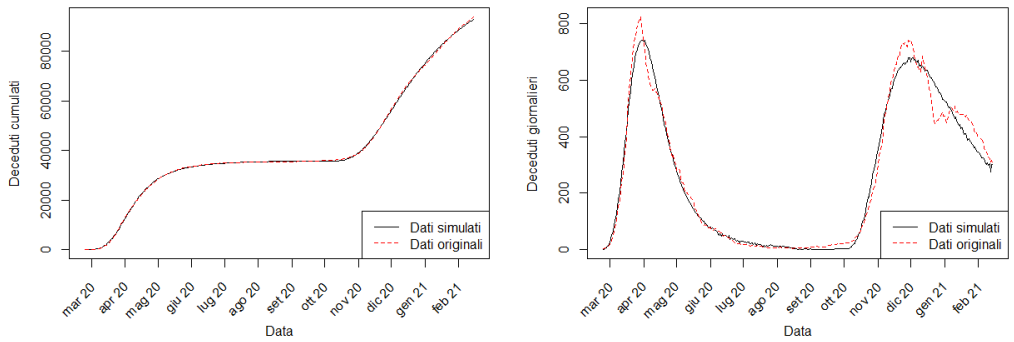
$$Y_i \sim \begin{cases} N(\mu_1(x_i, \hat{\beta}_1), \hat{\sigma}_1^2), & x_i \leq 181 \\ N(\mu_2(x_i - 181, \hat{\beta}_2), \hat{\sigma}_2^2), & x_i > 181 \end{cases}, \qquad i = 1, ..., n, \tag{4.5}$$

con

$$\mu_1(x_i, \hat{\beta}_1) = \frac{36043}{(1 + \exp(-3.18 \cdot (\log(x_i) - 3.72)))^{1.24}}, \quad \hat{\sigma}_1 = 5.22,$$

$$\mu_2(x_i, \hat{\beta}_2) = 35702 + \frac{81345}{(1 + \exp(-2.50 \cdot (\log(x_i) - 3.59)))^{18.93}}, \quad \hat{\sigma}_2 = 6.11.$$

La stima di massima verosimiglianza per τ è $\hat{\tau} = 181$: il *change-point* viene quindi stimato al 181-*esimo* giorno a partire dal 24 febbraio 2020, ovvero al 22 agosto 2020. Sono state simulate 1000 sequenze di osservazioni. In Figura 4.5 è riportato il confronto tra i dati originali e quelli ottenuti tramite una delle 1000 simulazioni, sia per il numero cumulato di deceduti [4.5a] sia per quello giornaliero [4.5b]. Come si può notare l'andamento delle due curve è molto simile, soprattutto per quanto riguarda il conteggio cumulato dei deceduti, mentre c'è una leggero e comprensibile discostamento per i dati giornalieri nel periodo corrispondente all'inizio del 2021, quando si è verificata un'interruzione della decrescita del numero di deceduti che il modello non



(a) Numero di deceduti cumulato **(b)** Numero di deceduti cumulato

Figura 4.5: Confronto tra i dati originali e i dati ottenuti tramite simulazione dal modello

è in grado di cogliere.

Per ciascuna delle 1000 simulazioni è stata calcolata una stima del *change-point*, ottenuto massimizzando la verosimiglianza profilo (3.4) per τ. Il boxplot in Figura 4.6 rappresenta le stime per τ ottenute dalle simulazioni: si noti come la stima di massima verosimiglianza $\hat{\tau} = 181$ sia abbastanza centrale rispetto alla distribuzione delle stime simulate, rientrando nel range interquartile. In particolare, c'è solo un giorno di differenza tra la mediana delle 1000 simulazioni (23 agosto) e la stima di massima verosimiglianza (22 agosto). Lo studio di simulazione è stato applicato anche all'analisi bayesiana, attraverso un approccio ibrido che coinvolge nuovamente la verosimiglianza profilo per τ. Per ognuna delle 1000 sequenze di dati simulate è stata ricostruita la distribuzione a posteriori tramite la generazione di 50000 valori con l'algoritmo di Metropolis-Hastings. Si tratta di un approccio ibrido perché la distribuzione a posteriori è stata ottenuto dal prodotto della verosimiglianza profilo (3.4) per un'opportuna distribuzione a priori. Successivamente è stata ottenuta per ciascuna distribuzione a posteriori una stima puntuale, calcolata come la media dei valori: ciò ha permesso di ricavare 1000 stime puntuali del *change-point*, ottenute tramite un approccio bayesiano. Il procedimento è stato effettuato utilizzando come distribuzioni a priori per τ la distribuzione a priori non informativa e quella molto informativa utilizzate nel paragrafo precedente nell'analisi con la verosimiglianza profilo generaliz-

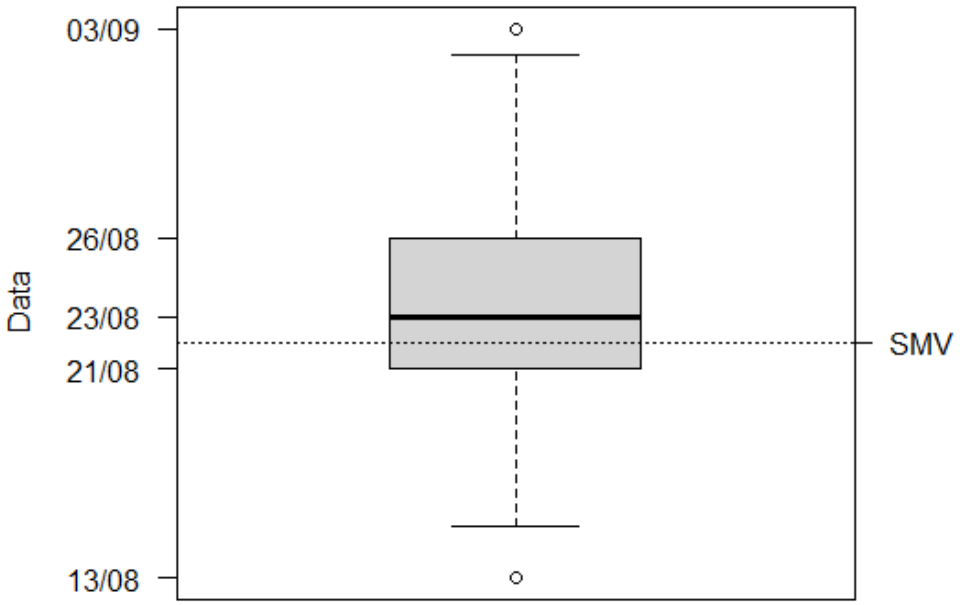


Figura 4.6: Boxplot delle stime per il *change-point* ottenute dalle simulazioni. Linea punteggiata: $\hat{\tau}$

zata. In Figura 4.7 sono rappresentati i boxplot delle 1000 stime puntuali per i due scenari considerati: l'utilizzo di una a priori informativa fa sì che i valori siano maggiormente concentrati intorno alla stima di massima verosimiglianza, mentre l'uso della a priori non informativa porta ad ottenere gli stessi risultati ottenuti con l'approccio frequentista.

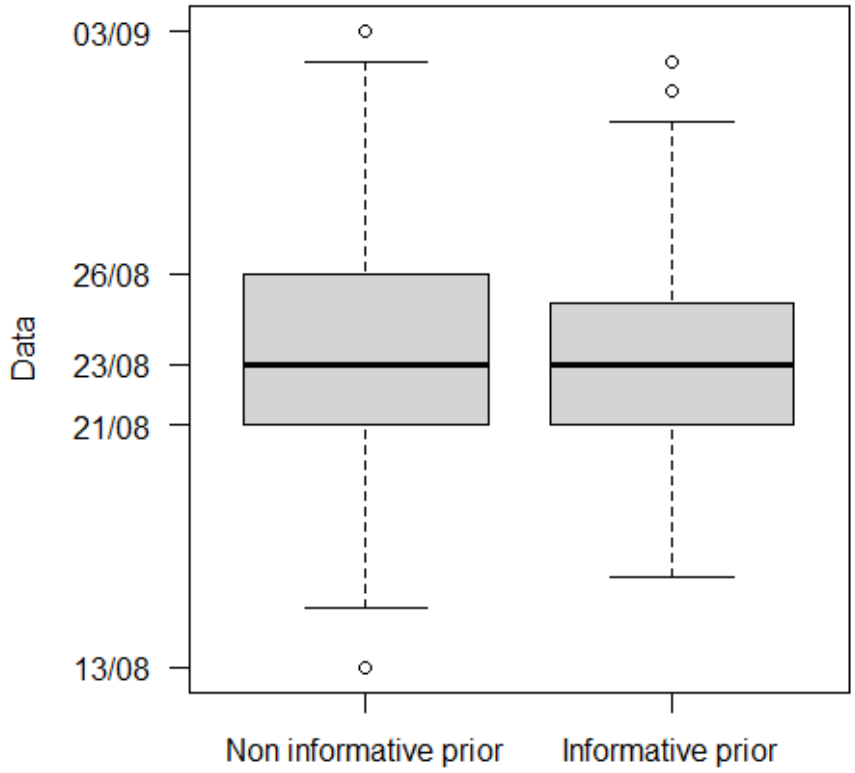


Figura 4.7: Boxplot delle stime puntuali bayesiane per il *change-point* ottenute dalle simulazioni

4.2 Discussione e possibili miglioramenti

I *modelli change-point* sono ampiamente utilizzati per analizzare una sequenza di osservazioni originata da un fenomeno che, in un certo istante

solitamente ignoto, subisce una modifica delle proprie caratteristiche. In letteratura sono presenti numerosi esempi dell'applicazione di tali modelli soprattutto nell'ambito del controllo statistico della qualità, sia con l'approccio frequentista all'inferenza sia con quello bayesiano. L'applicazione dei *modelli change-point* ai dati sul COVID-19 trova un ampio riscontro in letteratura, e non solamente per descrivere il passaggio da un'ondata a quella successiva. Ad esempio, Coughlin *et al.* (2021) utilizzano i *change-point* per modellare i cambiamenti nel trend del numero di nuovi casi, dovuti ad esempio alle misure di contenimento applicate dai governi. Greco *et al.* (2021, *Submitted*) sfruttano per primi le caratteristiche dei *modelli change-point* per analizzare il passaggio tra un'ondata e la successiva, attraverso una metodologia basata su un approccio frequentista e sulle curve di crescita log-logistiche. Nel loro articolo gli autori considerano un modello di Poisson e prestano particolare attenzione all'utilizzo di una metodologia robusta, che consenta di ottenere dei risultati validi anche a fronte della violazione dell'assunto di indipendenza e della presenza di sovradispersione.

Questa tesi, pur prendendo spunto dal suddetto articolo, si differenzia da quanto già presente in letteratura per la metodologia proposta; una delle principali differenze è l'assunzione di normalità per i dati, come in Girardi *et al.* (2020b), giustificata dal fatto che i conteggi sono quasi sempre elevati e possono quindi essere approssimati con una distribuzione continua. L'assunto di normalità prevede l'utilizzo di un parametro apposito per la variabilità dei dati slegato dalla media $\mu(x, \beta)$. L'altra innovazione da rimarcare è l'uso di un approccio bayesiano, che permette di ottenere l'intera distribuzione a posteriori del *change-point* e non solamente le stime puntuali e intervallari disponibili con l'approccio frequentista. Viene inoltre discussa la violazione dell'assunto di indipendenza, giustificando la metodologia proposta con l'utilizzo dell'*independence likelihood*.

Un possibile sviluppo di quanto proposto in questa tesi è l'applicazione delle tecniche bayesiane ai modelli con verosimiglianza di Poisson, essendo questi ultimi più corretti dal punto di vista degli assunti teorici su cui sono basati. Greco *et al.* (2021, *Submitted*) hanno proposto una prima modellazione dei dati del COVID-19 con modelli *change-point* assumendo la distribuzione di Poisson per il numero di deceduti, analizzando il problema da un punto di

vista prettamente frequentista.
Tra le ulteriori naturali continuazioni del lavoro presentato vi è senz'altro l'applicazione dei modelli ai dati sul COVID-19 maggiormente aggiornati, con l'obiettivo di ottenere risultati analoghi ma relativi al *change-point* tra seconda e terza ondata. Considerando le tre ondate, l'espansione del modello assumerebbe la forma

$$\mu(x,\beta) = \begin{cases} \mu(x,\beta_1), & x \leq \tau_1 \\ \mu(x-\tau_1,\beta_2), & \tau_1 < x \leq \tau_2 \\ \mu(x-\tau_2,\beta_3), & x > \tau_2 \end{cases},$$

e le successive analisi presenterebbero le medesime caratteristiche di quelle presentate in questa tesi. Alcune complicazioni potrebbero sorgere in uno studio in cui il numero di *change-point* è ignoto: si veda Barry e Hartigan (1993) e Siems (2020) per un approccio bayesiano a questo tipo di problematica.
In aggiunta, un'interessante strada esplorabile è quella dell'utilizzo di funzioni alternative alla verosimiglianza, ad esempio lo *score di Tsallis*; in questa tesi sono stati descritti gli aspetti teorici legati all'utilizzo di questa funzione, tralasciando l'applicazione ai dati sul COVID-19, onerosa da un punto di vista computazionale. Tuttavia, tali funzioni garantirebbero robustezza al modello, una proprietà particolarmente ricercata in un contesto in cui vengono analizzati dati di scarsa attendibilità come quelli relativi al COVID-19.
Infine, l'uso della verosimiglianza profilo generalizzata costituisce un'innovazione non solo nello studio dell'evoluzione del COVID-19 con modelli *change-point* basati su curve log-logistiche, ma più in generale nelle analisi con approccio bayesiano: meriterebbe senz'altro un approfondimento l'utilizzo di tale funzione in ambito bayesiano, soprattutto in contesti che presentano numerosi parametri di disturbo.

Bibliografia

Barry, D. e Hartigan, J. A. (1993). A Bayesian Analysis for Change Point Problems. *Journal of the American Statistical Association* **88**, pp. 309–319.

Bates, D. e Watts, D. (2007). *Nonlinear Regression Analysis and Its Applications*. Wiley series in probability and mathematical statistics. J. Wiley.

Brémaud, P. (2020). *Markov Chains Gibbs Fields, Monte Carlo Simulation and Queues: Gibbs Fields, Monte Carlo Simulation and Queues*. Springer.

Coughlin, S. S., Yigiter, A., Xu, H., Berman, A. E. e Chen, J. (2021). Early detection of change patterns in COVID-19 incidence and the implementation of public health policies: A multi-national study. *Public Health in Practice* **2**.

Dawid, A. P., Musio, M. e Ventura, L. (2016). Minimum Scoring Rule Inference. *Scandinavian Journal of Statistics* **43**, pp. 123–138.

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A. e Rubin, D. (2013). *Bayesian Data Analysis, Third Edition*. Taylor & Francis.

Ghosh, A. e Basu, A. (2013). Robust estimation for independent non-homogeneous observations using density power divergence with applications to linear regression. *Electronic Journal of Statistics* **7**, pp. 2420–2456.

Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E. e Ventura, L. (2020a). Robust Bayesian modelling for Covid-19 data in Italy. *Significance*.

Girardi, P., Greco, L., Mameli, V., Musio, M., Racugno, W., Ruli, E. e Ventura, L. (2020b). Robust inference for non-linear regression models from

the Tsallis score: Application to coronavirus disease 2019 contagion in Italy. *Stat* **9**, pp. 1–9.

Godambe, V. P. (1960). An Optimum Property of Regular Maximum Likelihood Estimation. *The Annals of Mathematical Statistics* **31**, pp. 1208–1211.

Greco, L., Girardi, P. e Ventura, L. (2021, *Submitted*). *Misspecified modeling of subsequent waves during COVID-19 outbreak: A change point growth model*.

Grigoletto, M., Ventura, L. e Pauli, F. (2017). *Modello lineare: teoria e applicazioni con R*. G. Giappichelli: Torino.

Johnstone, R. H., Bardenet, R., Gavaghan, D. J. e Mirams, G. R. (2017). Hierarchical Bayesian inference for ion channel screening dose-response data. *Wellcome Open Research* **1**.

Labelle, C., Marinier, A. e Lemieux, S. (2019). Enhancing the drug discovery process: Bayesian inference for the analysis and comparison of dose-response experiments. *Bioinformatics* **35**, pp. 464–473.

Liseo, B. (2008). *Dispensa corso bayesiana*.

McCullagh, P. e Nelder, J. A. (1989). *Generalized Linear Models, Second Edition*. Chapman e Hall.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H. e Teller, E. (1953). Equation of State Calculations by Fast Computing Machines. *jcp* **21**, pp. 1087–1092.

Pace, L. e Salvan, A. (2006). Adjustments of the profile likelihood from a new perspective. *Journal of Statistical Planning and Inference* **136**, pp. 3554–3564.

Qiu, P. (2014). *Introduction to statistical process control*. Texts in statistical science. CRC Press: Boca Raton.

Racugno, W., Salvan, A. e Ventura, L. (2010). Bayesian Analysis in Regression Models Using Pseudo-Likelihoods. *Communications in Statistics - Theory and Methods* **39**, pp. 3444–3455.

Racugno, W. e Ventura, L. (2017). *Biostatistica. Casi di Studio in R*. Egea: Milano.

Ritz, C., Baty, F., Streibig, J. C. e Gerhard, D. (2015). Dose-response analysis using R. *PLoS ONE* **10**, pp. 1–13.

Ritz, C. e Van Der Vliet, L. (2009). Handling nonnormality and variance heterogeneity for quantitative sublethal toxicity tests. *Environmental Toxicology and Chemistry* **28**, pp. 2009–2017.

Salvan, A., Sartori, N. e Pace, L. (2020). *Modelli lineari generalizzati*. Springer: Milano.

Sartori, N. (2020). *Statistica Computazionale (progredito): appunti delle lezioni*.

Severini, T. A. (1998). Likelihood Functions for Inference in the Presence of a Nuisance Parameter. *Biometrika* **85**, pp. 507–522.

Severini, T. A. e Wong, W. H. (1992). Profile Likelihood and Conditionally Parametric Models. *The Annals of Statistics* **20**, pp. 1768–1802.

Siems, T. (2020). *Bayesian Changepoint Analysis*.

Singhal, A., Singh, P., Lall, B. e Joshi, S. (2020). Modeling and prediction of COVID-19 pandemic using Gaussian mixture model. *Chaos, Solitons & Fractals* **138**.

Streibig, J. C., Dayan, F. E., Rimando, A. M. e Duke, S. O. (1999). Joint action of natural and synthetic photosystem II inhibitors. *Pesticide Science* **55**, pp. 137–146.

Varin, C., Reid, N. e Firth, D. (2011). An overview of composite likelihood methods. *Statistica Sinica* **21**, pp. 5–42.

Ventura, L. e Racugno, W. (2016). Pseudo-Likelihoods for Bayesian Inference. *Studies in Theoretical and Applied Statistics, Selected Papers of the Statistical Societies*, pp. 205–220.

Zeileis, A. (2004). Econometric Computing with HC and HAC Covariance Matrix Estimators. *Journal of Statistical Software* **11**, pp. 1–17.