Modelli per le probabilità di concepimento giornaliere

Bruno Scarpa

7 luglio 2005

Motivazioni

- L'identificazione di variabili previsive per le probabilità di concepimento in relazione al tempo e alla frequenza di rapporti sessuali nel ciclo mestruale è importante sia per le coppie interessate al concepimento (*achievers*) sia per gli utilizzatori dei metodi naturali di pianificazione familiare (*avoiders*), sia ancora per diagnosticare clinicamente possibili cause di infertilità o per scopi di epidemiologia riproduttiva.

Motivazioni

- L'identificazione di variabili previsive per le probabilità di concepimento in relazione al tempo e alla frequenza di rapporti sessuali nel ciclo mestruale è importante sia per le coppie interessate al concepimento (*achievers*) sia per gli utilizzatori dei metodi naturali di pianificazione familiare (*avoiders*), sia ancora per diagnosticare clinicamente possibili cause di infertilità o per scopi di epidemiologia riproduttiva.
- Spesso l'interesse si focalizza su uno o più previsori relativi all'intero ciclo (come età della donna, dell'uomo, numero di gravidanze precedenti) o relativi ai singoli giorni del ciclo (*day-specific*), come il tipo di muco cervicale osservato nel giorno in cui avviene il rapporto, o il livello di un ormone

Motivazioni

- L'identificazione di variabili previsive per le probabilità di concepimento in relazione al tempo e alla frequenza di rapporti sessuali nel ciclo mestruale è importante sia per le coppie interessate al concepimento (*achievers*) sia per gli utilizzatori dei metodi naturali di pianificazione familiare (*avoiders*), sia ancora per diagnosticare clinicamente possibili cause di infertilità o per scopi di epidemiologia riproduttiva.
- Spesso l'interesse si focalizza su uno o più previsori relativi all'intero ciclo (come età della donna, dell'uomo, numero di gravidanze precedenti) o relativi ai singoli giorni del ciclo (*day-specific*), come il tipo di muco cervicale osservato nel giorno in cui avviene il rapporto, o il livello di un ormone
- Poiché in ogni ciclo ci possono essere rapporti multipli durante la fase potenzialmente fertile del ciclo, il concepimento rappresenta l'aggregazione di esperimenti Bernoulliani per ogni giorno del rapporto.

Il modello di Schwartz e altri

- Il modello di Schwartz, MacDonald e Heuchel (1980)

$$Pr\{Y_{ij} = 1|\mathbf{X}_{ij}\} = \omega\{1 - \prod_{k=1}^{K}(1 - \lambda_k)^{X_{ijk}}\},$$

dove

- Y_{ij} è un indicatore del concepimento nel ciclo j per la donna i,
- $\mathbf{X}_{ij} = [X_{ij1}, \ldots, X_{ijK}]^{\mathrm{T}}$ è un vettore di indicatori di rapporti per i giorni $1, \ldots, K$ (tipicamente indiciati relativamente al giorno dell'ovulazione),
- ω è la probabilità di "viabilità del ciclo" (*cycle viability*);
- λ_k è la probabilità di concepimento in un ciclo "viabile" (*viable*) con rapporti solo nel giorno k.

Generalizzazioni di Scwhartz et al.

- Aggiunta di effetti covariati legati all'intero ciclo su ω (Weinberg, Gladen e Wilcox, 1994, Colombo e Masarotto, 2000).
- Eterogeneità tra le coppie in ω attraverso effetti casuali (Dunson e Zhou, 2000, Zhou et al. 1996).
- Effetti covariati dipendenti dai singoli giorni del ciclo sui λ_k (Zhou e Weinberg, 1996, Colombo et al. 2004).
- Dati mancanti per i rapporti (Dunson e Weinberg, 2000),
- Errori di misurazione nell'identificare il vero giorno dell'ovulazione (Dunson et al. 2001, Dunson e Weinberg, 2000).
- Alternative parametriche al modello di Schwartz (Royston, 1982; Weinberg e Wilcox, 1995; Royston e Ferreira, 1999).

Discussione del modello di Schwartz et al.

Discussione del modello di Schwartz et al.

- Un ciclo mestruale è considerato "viabile" se tutti i fattori non legati al tempo e alla frequenza dei rapporti sono favorevoli al concepimento.
- Non solo, quindi, i fattori specifici della donna (come il rilascio di un ovulo viabile, conducibilità negli ovidotti, trasporto sicuro nell'utero e ricettività uterina per l'impianto), ma anche fattori maschili (come la produzione di spermatozoi capaci di progressiva motilità nel tratto riproduttivo femminile), o gli effetti di interazione (capacità dello spermatozoo di fertilizzare l'ovulo, sopravvivenza dell'embrione fino al momento della rilevazione...)

Discussione del modello di Schwartz et al.

Sfortunatamente non sono disponibili dati diretti per indagare sui diversi fattori ed è perfino difficile stimare separatamente ω e uno dei λ_k (ad esempio il più grande), infatti:

- L'incorporazione di fattori maschili e femminili sia in ω che in λ_k rende difficile la definizione dei fattori biologici direttamente legati a ciascun elemento
- Inoltre, tende a essere difficile la separazione delle stime di ω e di uno dei λ_k
- L'identificabilità stessa di tali parametri è molto debole: ad es. se tutti i cicli mestruali disponibili hanno un solo giorno con rapporto, il modello di Schwartz non è stimabile, e uno dei parametri deve essere fissato per adattare il modello.

Il modello di Dunson (2001)

- Poiché il più grande dei λ_k (nelle stime ottenute basandosi su diversi insiemi di dati) è sempre vicino a 1, una modifica ragionevole del modello di Schwartz che risolve il problema di non identificabilità (collinearità) cnsiste nel fissare il più grande dei λ_k pari a 1.

Il modello di Dunson (2001)

- Poiché il più grande dei λ_k (nelle stime ottenute basandosi su diversi insiemi di dati) è sempre vicino a 1, una modifica ragionevole del modello di Schwartz che risolve il problema di non identificabilità (collinearità) cnsiste nel fissare il più grande dei λ_k pari a 1.
- Tale soluzione proposta da Dunson (2001) consiste nell'utilizzare il modello

$$\omega \left\{ X_{ijM} + (1 - X_{ijM}) \left[1 - \prod_{l=-C}^{D} (1 - p_l)^{X_{ij,l+M}} \right] \right\},$$

dove M è il giorno più fertile relativamente all'ovulazione, e p_l è indicizzato relativamente a M quando $l = 0$, con $p_l = 0$ per $l \notin [-C, D]$.

Il modello di Dunson (2001)

- Poiché il più grande dei λ_k (nelle stime ottenute basandosi su diversi insiemi di dati) è sempre vicino a 1, una modifica ragionevole del modello di Schwartz che risolve il problema di non identificabilità (collinearità) cnsiste nel fissare il più grande dei λ_k pari a 1.
- Tale soluzione proposta da Dunson (2001) consiste nell'utilizzare il modello

$$\omega \left\{ X_{ijM} + (1 - X_{ijM}) \left[1 - \prod_{l=-C}^{D} (1 - p_l)^{X_{ij,l+M}} \right] \right\},$$

dove M è il giorno più fertile relativamente all'ovulazione, e p_l è indicizzato relativamente a M quando $l = 0$, con $p_l = 0$ per $l \notin [-C, D]$.

- Dunson assume anche che

$$0 = p_{-C-1} \leq p_{-C} \leq \cdots \leq p_{-1} \leq p_0 \geq p_1 \geq \cdots \geq p_D \geq p_{D+1} = 0$$

Discussione sul modello di Dunson (2001)

- Il modello proposto permette di stimare sia effetti giornalieri sia l'eterogeneità dovuta alle coppie.

Discussione sul modello di Dunson (2001)

- Il modello proposto permette di stimare sia effetti giornalieri sia l'eterogeneità dovuta alle coppie.
- La stima avviene tramite un procedimento Bayesiano gerarchico, che una volta definite le distribuzioni a-priori utilizza algoritmi di tipo MCMC per ottenere le stime.

Discussione sul modello di Dunson (2001)

- Il modello proposto permette di stimare sia effetti giornalieri sia l'eterogeneità dovuta alle coppie.
- La stima avviene tramite un procedimento Bayesiano gerarchico, che una volta definite le distribuzioni a-priori utilizza algoritmi di tipo MCMC per ottenere le stime.
- Purtroppo l'algoritmo proposto da Dunson (2001) è computazionalmente molto intensivo, il che rappresenta una rilevante barriera all'utilizzo del modello nell'analisi routinaria di dati reali (oltre a presentare difficoltà nell'effettuare studi di simulazione per lo studio delle proprietà statistiche del modello).

Discussione sul modello di Dunson (2001)

- Il modello proposto permette di stimare sia effetti giornalieri sia l'eterogeneità dovuta alle coppie.
- La stima avviene tramite un procedimento Bayesiano gerarchico, che una volta definite le distribuzioni a-priori utilizza algoritmi di tipo MCMC per ottenere le stime.
- Purtroppo l'algoritmo proposto da Dunson (2001) è computazionalmente molto intensivo, il che rappresenta una rilevante barriera all'utilizzo del modello nell'analisi routinaria di dati reali (oltre a presentare difficoltà nell'effettuare studi di simulazione per lo studio delle proprietà statistiche del modello).
- Inoltre l'approccio non aiuta la selezione delle variabili previsive, né permette di fare inferenza sui trend rispetto alle probabilità giornaliere legate a variabili esplicative categoriali (come il tipo di muco).

Un nuovo modello: formulazione

- Sia $\mathbf{U}_{ij} = [\mathbf{u}_{ij1}^{\mathrm{T}}, \ldots, \mathbf{u}_{ijk}^{\mathrm{T}}]^{\mathrm{T}}$ una matrice di covariate per il ciclo j $(j = 1, \ldots, n_i)$ dalla coppia i $(i = 1, \ldots, n)$
- la probabilità di concepimento può essere modellata come

$$\begin{aligned} P(Y_{ij} = 1 | \xi_i, \mathbf{X}_{ij}, \mathbf{U}_{ij}) &= 1 - \prod_{K=1}^{K} (1 - \lambda_{ijk})^{X_{ijk}} \\ \lambda_{ijk} &= 1 - \exp\{-\xi_i \exp(\mathbf{u}_{ijk}^{\mathrm{T}} \beta)\} \\ \xi_i &\sim \mathcal{G}(\phi, \phi) \end{aligned}$$

dove

- λ_{ijk} è la probabilità giornaliera di concepimento nel ciclo j della coppia i dato un rapporto solo nel giorno k,
- ξ_i è un moltiplicatore di fecondabilità relativo alla coppia i
- $\mathcal{G}(a, b)$ denota la densità Gamma con media a/b e varianza a/b^2, e
- β è un vettore di coefficienti di regressione.

Alcuni commenti sul modello

- La scelta di fissare i due parametri della distribuzione Gamma uguali tra loro implica che $E(\xi_i) = 1$, il che evita il fenomeno di non-identificabilità tra $E(\xi_i)$ e i parametri giornalieri (*day-specific*).
- Il parametro comune ϕ è quindi interpretabile come $\phi = 1/V(\xi_i)$ ed è quindi una misura del livello della variabilità tra le coppie.

Alcuni commenti sul modello

- La scelta di fissare i due parametri della distribuzione Gamma uguali tra loro implica che $E(\xi_i) = 1$, il che evita il fenomeno di non-identificabilità tra $E(\xi_i)$ e i parametri giornalieri (*day-specific*).
- Il parametro comune ϕ è quindi interpretabile come $\phi = 1/V(\xi_i)$ ed è quindi una misura del livello della variabilità tra le coppie.
- In questo modello viene eliminato il parametro ω e l'effetto casuale ξ_i viene inserito nel termine giornaliero (*day-specific*).

Alcuni commenti sul modello

- La scelta di fissare i due parametri della distribuzione Gamma uguali tra loro implica che $E(\xi_i) = 1$, il che evita il fenomeno di non-identificabilità tra $E(\xi_i)$ e i parametri giornalieri (*day-specific*).
- Il parametro comune ϕ è quindi interpretabile come $\phi = 1/V(\xi_i)$ ed è quindi una misura del livello della variabilità tra le coppie.
- In questo modello viene eliminato il parametro ω e l'effetto casuale ξ_i viene inserito nel termine giornaliero (*day-specific*).
- Nel caso particolare in cui $\lambda_{ijk} = \lambda_k$ per ogni i e j il modello si semplifica nella forma proposta da Barrett e Marshall (1969).

Alcuni commenti sul modello

- La scelta di fissare i due parametri della distribuzione Gamma uguali tra loro implica che $E(\xi_i) = 1$, il che evita il fenomeno di non-identificabilità tra $E(\xi_i)$ e i parametri giornalieri (*day-specific*).
- Il parametro comune ϕ è quindi interpretabile come $\phi = 1/V(\xi_i)$ ed è quindi una misura del livello della variabilità tra le coppie.
- In questo modello viene eliminato il parametro ω e l'effetto casuale ξ_i viene inserito nel termine giornaliero (*day-specific*).
- Nel caso particolare in cui $\lambda_{ijk} = \lambda_k$ per ogni i e j il modello si semplifica nella forma proposta da Barrett e Marshall (1969).
- La distribuzione Gamma per ξ_i prevede la presenza di coppie non–fertili (tipicamente coppie che ci mettono più di un anno a concepire), ma non permette la presenza di una proporzione di coppie sterili con probabilità di concepimento pari a 0 indipendentemente da quando avvengono i rapporti. Per affrontare una tale eventualità si potrebbe incorporare un punto massa in 0 e usare una distribuzione mistura (Weinberg e Gladen, 1986; Dunson e Zhou, 2000).

La probabilità di concepimento marginale

- La probabilità di concepimento marginale, una volta integrato il parametro relativo all'effetto casuale (*frailty*) ξ_i, ha una forma chiusa:

$$\begin{aligned} P(Y_{ij} = 1 | \mathbf{X}_{ij}, \mathbf{U}_{ij}) &= 1 - \int_0^\infty \exp\left\{ -\xi_i \sum_{k=1}^{K} X_{ijk} \exp(\mathbf{u}_{ijk}^{\mathrm{T}} \beta) \right\} \mathcal{G}(\xi_i; \phi, \phi) d\xi_i \\ &= 1 - \left(\frac{\phi}{\phi + \sum_{k=1}^{K} X_{ijk} \exp(\mathbf{u}_{ijk}^{\mathrm{T}} \beta)} \right)^{\phi}. \end{aligned}$$

La probabilità di concepimento marginale

- La probabilità di concepimento marginale, una volta integrato il parametro relativo all'effetto casuale (*frailty*) ξ_i, ha una forma chiusa:

$$\begin{aligned} P(Y_{ij} = 1|\mathbf{X}_{ij}, \mathbf{U}_{ij}) &= 1 - \int_0^\infty \exp\left\{-\xi_i \sum_{k=1}^{K} X_{ijk} \exp(\mathbf{u}_{ijk}^{\mathrm{T}}\beta)\right\} \mathcal{G}(\xi_i; \phi, \phi) d\xi_i \\ &= 1 - \left(\frac{\phi}{\phi + \sum_{k=1}^{K} X_{ijk} \exp(\mathbf{u}_{ijk}^{\mathrm{T}}\beta)}\right)^{\phi}. \end{aligned}$$

- La probabilità di concepimento giornaliera (*day specific*) in un ciclo con rapporto solo nel giorno k e con previsori $\mathbf{u}$ è

$$P(Y = 1|\mathbf{u}) = 1 - \left(\frac{\phi}{\phi + \exp(\mathbf{u}_{ijk}^{\mathrm{T}}\beta)}\right)^{\phi},$$

Relazione con i GLM

- Questo modello si riduce al modello di regressione logistica quando $\phi = 1$. La scelta di ϕ incognita permette una classe più flessibile di funzioni legame che includono legami asimmetrici.
- Questo nuovo modo di derivare il modello in esame ci permette di dare a ϕ un nuovo significato ϕ, infatti, ha un ruolo duale, dauna parte come misura di eterogeneità e, dall'altra come parametro che identifica la diversità (*lack of fit*) rispetto al modello di regressione logistica.
- Si può sfruttare questa caratteristica per un calcolo efficiente delle distribuzioni a posteriori in applicazioni più generali.

Specificazione delle distribuzioni a priori

- L'obiettivo di interesse è la selezione di previsori per le probabilità giornaliere, quantificandone gli effetti.
  In particolare interessano previsori categoriali ordinati, come ad esempio il tipo di muco o l'età raggruppata in classi.

Specificazione delle distribuzioni a priori

- L'obiettivo di interesse è la selezione di previsori per le probabilità giornaliere, quantificandone gli effetti.
  In particolare interessano previsori categoriali ordinati, come ad esempio il tipo di muco o l'età raggruppata in classi.
- Si scelgono quindi delle distribuzioni a-priori opportune per i parametri di regressione β, che assegnino punti massa a 0 per permettere ai previsori di venire eliminati dal modello.

Specificazione delle distribuzioni a priori

- L'obiettivo di interesse è la selezione di previsori per le probabilità giornaliere, quantificandone gli effetti.
  In particolare interessano previsori categoriali ordinati, come ad esempio il tipo di muco o l'età raggruppata in classi.
- Si scelgono quindi delle distribuzioni a-priori opportune per i parametri di regressione β, che assegnino punti massa a 0 per permettere ai previsori di venire eliminati dal modello.
- In particolare ponendo $\gamma_h = \exp(\beta_h)$, le a-priori per i γ_h saranno della forma

$$\pi(\gamma) = \prod_h \delta_1 - \mathcal{G}_{\mathcal{A}h}(\gamma_h; p_h, a_h, b_h),$$

  dove $\delta_1 - \mathcal{G}_{\mathcal{A}h}(\gamma_h; p_h, a_h, b_h)$ è la densità della mistura con un punto massa a 1 (con probabilità p_h) e una densità $\mathcal{G}(a_h, b_h)$ troncata alla regione $\mathcal{A}_h$.

Specificazione delle distribuzioni a priori

- L'obiettivo di interesse è la selezione di previsori per le probabilità giornaliere, quantificandone gli effetti.
  In particolare interessano previsori categoriali ordinati, come ad esempio il tipo di muco o l'età raggruppata in classi.
- Si scelgono quindi delle distribuzioni a-priori opportune per i parametri di regressione β, che assegnino punti massa a 0 per permettere ai previsori di venire eliminati dal modello.
- In particolare ponendo $\gamma_h = \exp(\beta_h)$, le a-priori per i γ_h saranno della forma

$$\pi(\gamma) = \prod_h \delta_1 - \mathcal{G}_{\mathcal{A}h}(\gamma_h; p_h, a_h, b_h),$$

  dove $\delta_1 - \mathcal{G}_{\mathcal{A}h}(\gamma_h; p_h, a_h, b_h)$ è la densità della mistura con un punto massa a 1 (con probabilità p_h) e una densità $\mathcal{G}(a_h, b_h)$ troncata alla regione $\mathcal{A}_h$.
- Si sceglie come a-priori per ϕ una $\mathcal{G}(c_1, c_2)$.

Commenti sulle a-priori

- I valori di $\gamma_h = 1$ corrispondono ai $\beta_h = 0$ e all'eliminazione dal modello del previsore $\mathbf{u}_{ijk}$

Commenti sulle a-priori

- I valori di $\gamma_h = 1$ corrispondono ai $\beta_h = 0$ e all'eliminazione dal modello del previsore $\mathbf{u}_{ijk}$
- La probabilità a priori che l'h-esimo previsore sia incluso nel modello sarà quindi $1 - p_h$

Commenti sulle a-priori

- I valori di $\gamma_h = 1$ corrispondono ai $\beta_h = 0$ e all'eliminazione dal modello del previsore $\mathbf{u}_{ijk}$
- La probabilità a priori che l'h-esimo previsore sia incluso nel modello sarà quindi $1 - p_h$
- Se il previsore è incluso, allora il coefficiente γ_h è vincolato ad appartenere alla regione $\mathcal{A}_h$ che generalmente viene scelta come $\mathbb{R}^+$ cioè senza alcun vincolo, $[1, +\infty)$ cioè con un possibile effetto solo positivo del previsore sulle probabilità di concepimento o $(0, 1)$ in corrispondenza ad un effetto negativo.
- La scelta appropriata di $\mathcal{A}_h$ permette una diminuzione dell'incertezza a-posteriori attraverso l'inserimento nell'a-priori di informazioni sulla direzione dell'associazione.

Commenti sulle a-priori

- I valori di $\gamma_h = 1$ corrispondono ai $\beta_h = 0$ e all'eliminazione dal modello del previsore $\mathbf{u}_{ijk}$
- La probabilità a priori che l'h-esimo previsore sia incluso nel modello sarà quindi $1 - p_h$
- Se il previsore è incluso, allora il coefficiente γ_h è vincolato ad appartenere alla regione $\mathcal{A}_h$ che generalmente viene scelta come $\mathbb{R}^+$ cioè senza alcun vincolo, $[1, +\infty)$ cioè con un possibile effetto solo positivo del previsore sulle probabilità di concepimento o $(0, 1)$ in corrispondenza ad un effetto negativo.
- La scelta appropriata di $\mathcal{A}_h$ permette una diminuzione dell'incertezza a-posteriori attraverso l'inserimento nell'a-priori di informazioni sulla direzione dell'associazione.
- Poiché un punto massa è associato all'ipotesi nulla di nessuna associazione è immediato impostare una verifica d'ipotesi sia unilaterale che bilaterale.

Inferenza

- L'inferenza è fatta usando procedure MCMC.

Inferenza

- L'inferenza è fatta usando procedure MCMC.
- Viene aumentato lo spazio parametrico per includere delle variabili ausiliarie e rendere più efficiente l'algoritmo MCMC proposto.

Inferenza

- L'inferenza è fatta usando procedure MCMC.
- Viene aumentato lo spazio parametrico per includere delle variabili ausiliarie e rendere più efficiente l'algoritmo MCMC proposto.
- Il modello può venire ri-espresso nella seguente forma

$$Y_{ij} = \mathbb{I}\left(\sum_{k=1}^{K} X_{ijk} Z_{ijk} > 0\right)$$

con $\mathbb{I}(\cdot)$ è la funzione indicatrice e $\mathbf{Z}_{ij} = [Z_{ij1}, \ldots, Z_{ijK}]^{\mathrm{T}}$ è un vettore di variabili latenti di Poisson, condizionatamente indipendenti dati i parametri degli effetti casuali ξ_i

$$Z_{ijk} \sim \text{Poisson}\left(\xi_i \exp(\mathbf{u}_{ijk}^{\mathrm{T}} \beta)\right), \quad k = 1, \ldots, K$$

Variabili ausiliarie

- Se si itegra questo modello, eliminando le variabili ausiliarie di Poisson $Z_{ij1}, \ldots, Z_{ijK}$, si ottiene il modello proposto.

Variabili ausiliarie

- Se si itegra questo modello, eliminando le variabili ausiliarie di Poisson $Z_{ij1}, \ldots, Z_{ijK}$, si ottiene il modello proposto.
- In particolare in questo modello $Y_{ij} = 0$ se e solo se $Z_{ijk} = 0$ per ogni k tale che $X_{ijk} = 1$, cioè in tutti i giorni con rapporti nel j-esimo ciclo della i-esima coppia. Poiché $P(Z_{ijk} = 0|\xi_i, \mathbf{u}_{ijk}) = \exp\{-\xi_i \exp(\mathbf{u}_{ijk}^{\mathrm{T}}\beta)\}$ e le variabili di Poisson sono indipendenti, abbiamo

$$\begin{aligned} P(Y_{ij} = 0|\xi_i, \mathbf{X}_{ij}, \mathbf{U}_{ij}) &= \prod_{k:X_{ijk}=1} P(Z_{ijk} = 0|\xi_i, \mathbf{u}_{ijk}) \\ &= \prod_{k=1}^{K} \exp\{-\xi_i \exp(\mathbf{u}_{ijk}^{\mathrm{T}}\beta)\}^{X_{ijk}} \end{aligned}$$

che è nella forma $\prod_k (1 - \lambda_{ijk})^{X_{ijk}}$. Da questo è immediato ottenere $P(Y_{ij} = 1|\xi_i, \mathbf{X}_{ij}, \mathbf{U}_{ij})$, nella forma del modello proposto.

L'a-posteriori

- Con le variabili ausiliarie introdotte la distribuzione a-posteriori congiunta risulta

$$\left(\prod_{i=1}^{n} \mathcal{G}(\xi_i;\phi,\phi) \prod_{j=1}^{n_i} \left\{ \mathbb{I}\left(\sum_{k=1}^{K} X_{ijk} Z_{ijk} > 0\right) Y_{ij} + \mathbb{I}\left(\sum_{k=1}^{K} X_{ijk} Z_{ijk} = 0\right)(1 - Y_{ij}) \right\} \right.$$
$$\left. \times \left[\prod_{k=1}^{K} \frac{\{\xi_i \exp(\mathbf{u}_{ijk}^{\mathrm{T}}\beta)\}^{Z_{ijk}} \exp\{-\xi_i \exp(\mathbf{u}_{ijk}^{\mathrm{T}}\beta)\}}{Z_{ijk}!} \right] \right) \times \pi(\beta)\pi(\phi)$$

- Da questa distribuzione congiunta si possono ottenere tramite passaggi standard di algebra le "*full conditional*" per ciascuno dei parametri e per le variabili latenti.

L'a-posteriori

- Dopo aver introdotto le variabili ausiliarie, se i previsori sono variabili categoriali, le a-priori scelte sono condizionatamente coniugate ad eccezione di ϕ
- È immediato quindi ottenere dei campioni dalle distribuzioni condizionate
- Per il calcolo dell'a-posteriori è quindi opportuno utilizzare un semplice algoritmo ibrido di *Gibbs sampling* e Metropolis, con un unico passo di Metropolis per aggiornare ϕ
- In questo caso la scelta di a-priori coniugate, per quanto non cruciale vista la disponibilità di algoritmi MCMC come Metropolis-Hastings, è rilevante per aumentare l'efficienza computazionale in problemi Bayesiani di selezione di variabili.

L'algortimo MCMC

Let $\mathrm{Z_{ij}} = \sum_{\mathrm{k=1}}^{\mathrm{K}} \mathrm{X_{ijk}} \mathrm{Z_{ijk}}$

Step 1. Sample from the full conditional distribution of $\mathbf{Z}_{ij}$ by setting $\mathbf{Z}_{ij} = \mathbf{0}$ if $Y_{ij} = 0$

and otherwise sampling sequentially from

$$\pi(Z_{ij} \,|\, Y_{ij} = 1, \boldsymbol{\theta}, \boldsymbol{\xi}, \text{data}) = \text{Poisson}(\mathbf{X}'_{ij}\boldsymbol{\mu}_{ij}) \text{ truncated so that } Z_{ij} > 0,$$

$$\pi(\mathbf{Z}_{ij} \,|\, Z_{ij}, Y_{ij}, \boldsymbol{\theta}, \boldsymbol{\xi}, \text{data}) = \text{Multinomial}(Z_{ij}; X_{ij1}\alpha_{ij1}, \ldots, X_{ijK}\alpha_{ijK}).$$

Step 2. Sample the elements of $\boldsymbol{\lambda}$ from their conjugate full conditional distributions:

$$\pi(\lambda_k \,|\, \mathbf{Z}_{[\mathbf{X}=\mathbf{1}]}, \boldsymbol{\theta}_{(-\lambda_k)}, \boldsymbol{\xi}, \text{data}) = \mathcal{G}\Big(\lambda_k; a_{0k} + \sum_{ij:X_{ijk}=1} Z_{ijk}, b_{0k} + \sum_{i,j:X_{ijk}=1} \xi_i \prod_{h=1}^{w_{ijk}-1} \gamma_h\Big),$$

where $\mathbf{Z}_{[\mathbf{X}=\mathbf{1}]} = \{Z_{ijk} : X_{ijk} = 1\}$ and we integrate out $\{Z_{ijk} : X_{ijk} = 0\}$.

Step 4. Sample the elements of $\boldsymbol{\gamma}$ from their conjugate full conditional distributions:

$$\pi(\gamma_h \,|\, \mathbf{Z_{[X=1]}}, \boldsymbol{\theta}_{(-\gamma_h)}, \boldsymbol{\xi}, \text{data}) \;=\; \text{I}_1\text{-}\mathcal{G}_{[1,\infty)}(\gamma_h; \widetilde{\pi}_h, \widetilde{a}_h, \widetilde{b}_h),$$

where $\widetilde{a}_h = a_h + \sum_{i,j,k:X_{ijk}=1} 1_{(h<w_{ijk})} Z_{ijk}$, $\widetilde{b}_h = b_h + \sum_{i,j,k:X_{ijk}=1} 1_{(h<w_{ijk})} \xi_i \lambda_k \prod_{l:l\neq h}^{w_{ijk}-1} \gamma_l$,

$$\widetilde{\pi}_h = \frac{\pi_{0h} \exp\Big(- \sum_{i,j,k:X_{ijk}=1} 1_{(h<w_{ijk})} \xi_i \lambda_k \prod_{l:l\neq h}^{w_{ijk}-1} \gamma_l \Big)}{\pi_{0h} \exp\Big(- \sum_{i,j,k:X_{ijk}=1} 1_{(h<w_{ijk})} \xi_i \lambda_k \prod_{l:l\neq h}^{w_{ijk}-1} \gamma_l \Big) + (1-\pi_{0h}) \frac{C(a_h,b_h)}{C(\widetilde{a}_h,\widetilde{b}_h)} \frac{1-F(1;\widetilde{a}_h,\widetilde{b}_h)}{1-F(1;a_h,b_h)}}.$$

Step 5. Sample ξ_i, for $i = 1, \ldots, n$, from its full conditional distribution, which is

$$\pi(\xi_i \,|\, \mathbf{Z_{[X=1]}}, \boldsymbol{\theta}, \text{data}) = \mathcal{G}\Big(\xi_i; \nu^{-1} + \sum_{j,k:X_{ijk}=1} Z_{ijk}, \nu^{-1} + \sum_{j,k:X_{ijk}=1} \lambda_k \prod_{h=1}^{w_{ijk}-1} \gamma_h\Big),$$

Step 6. Update ν using a Metropolis step.

Step 7. Repeat steps 1-6 until apparent convergence and calculate posterior summaries based on a large number of additional iterations.

Selezione del modello

- L'algoritmo può essere utilizzato per selezionare il miglior modello
- Sia $M_h = \mathbb{I}(\beta_h \neq 0)$, per $h = 1, \ldots, q$ il modello viene indicizzato dal vettore $\mathbf{M} = [M_1, \ldots, M_q]^{\mathrm{T}}$
- Ponendo $\mathbf{M}^{(s)}$ il valore di $\mathbf{M}$ all'iterazione s-esima dell'algoritmo MCMC dopo un primo periodo di *burn-in*, possiamo ottenere le probabilità a posteriori dei modelli usando

$$\hat{P}(\mathbf{M} = \mathbf{m}|dati) = \frac{1}{S}\sum_{s=1}^{S}\mathbb{I}(\mathbf{M}^{(s)} = \mathbf{m}).$$

Verifica di ipotesi di associazione

- Oltre alla selezione delle variabili rilevanti è spesso di interesse capire se un dato previsore è associato o meno alla probabilità di concepire.
- Per un semplice previsore dicotomico, u_{ijkh}, la probabilità dell'ipotesi nulla di nessuna associazione, $H_{0h} : \gamma_h = 1$ può essere stimata direttamente dalle iterazioni MCMC usando lo stimatore Rao-Blackwellizzato

$$\hat{P}(H_{0h}|dati) = \frac{1}{S}\sum_{s=1}^{S} \tilde{p}_h^{(s)}$$

dove $\tilde{p}$ è la probabilità a posteriori condizionata di $\gamma_h = 1$ ottenuta all'iterazione s

Variabile ordinale

- Spesso è di interesse cercare evidenze di una possibile associazione tra un previsore categoriale ordinale $w_{ijk} \in \{1, \ldots, d\}$ e il concepimento.

Variabile ordinale

- Spesso è di interesse cercare evidenze di una possibile associazione tra un previsore categoriale ordinale $w_{ijk} \in \{1, \ldots, d\}$ e il concepimento.
- Se si parametrizza il modello in modo tale che i primi $d-1$ elementi di $\mathbf{u}_{ijk}$ abbiano la struttura $[\mathbb{I}(w_i \geq 2), \ldots, \mathbb{I}(w_i \geq d)]^{\mathrm{T}}$ e i rimanenti elementi si riferiscano ad altre covariate, l'ipotesi nulla di nessuna associazione può essere espressa come

$$H_0 : \gamma_1 = \cdots = \gamma_{d-1} = 1.$$

Variabile ordinale

- Spesso è di interesse cercare evidenze di una possibile associazione tra un previsore categoriale ordinale $w_{ijk} \in \{1, \ldots, d\}$ e il concepimento.
- Se si parametrizza il modello in modo tale che i primi $d-1$ elementi di $\mathbf{u}_{ijk}$ abbiano la struttura $[\mathbb{I}(w_i \geq 2), \ldots, \mathbb{I}(w_i \geq d)]^{\mathrm{T}}$ e i rimanenti elementi si riferiscano ad altre covariate, l'ipotesi nulla di nessuna associazione può essere espressa come

$$H_0 : \gamma_1 = \cdots = \gamma_{d-1} = 1.$$

- La probabilità a priori di H_0 è allora $\prod_{h=1}^{d-1} p_h$. Scegliendo ad esempio $\mathcal{A}_h = (1, +\infty)$ si ottiene un vincolo di relazione non decrescente, con un'ipotesi alternativa unilaterale corrispondente ad almeno una crescita nelle probabilità di concepimento al variare di w_{ijk}.

Applicazione: l'effetto del muco

- Il modello viene applicato allo Studio Europeo di Foncdabilità (Colombo e Masarotto, 2000)
- 782 coppie selezionate da 7 Centri di Regolazione Naturale delle Nascite
- Donne tra i 18 e i 40 anni, senza evidenze precedenti di infertilità
- Informazioni giornaliere su BBT, perdite, muco, rapporti
- Ovulazione stimata attraverso BBT (regola del 3/6)
- Informazioni sul Muco. Punteggio su 4 livelli ordinali: 1=secco (assenza, meno fertile), ..., 4=tipo più fertile

- Sono stati esclusi dall'analisi i cicli per i quali

- Sono stati esclusi dall'analisi i cicli per i quali
  - c'erano insufficienti informazioni per stimare il giorno dell'ovulazione,

- Sono stati esclusi dall'analisi i cicli per i quali
  - c'erano insufficienti informazioni per stimare il giorno dell'ovulazione,
  - non erano riportati rapporti nell'intervallo, considerato fertile, di 6 giorni (intorno all'ovulazione),

- Sono stati esclusi dall'analisi i cicli per i quali
  - c'erano insufficienti informazioni per stimare il giorno dell'ovulazione,
  - non erano riportati rapporti nell'intervallo, considerato fertile, di 6 giorni (intorno all'ovulazione),
  - se in uno dei giorni con rapporti non era stata raccolta l'informazione sul tipo di muco

- Sono stati esclusi dall'analisi i cicli per i quali
  - c'erano insufficienti informazioni per stimare il giorno dell'ovulazione,
  - non erano riportati rapporti nell'intervallo, considerato fertile, di 6 giorni (intorno all'ovulazione),
  - se in uno dei giorni con rapporti non era stata raccolta l'informazione sul tipo di muco
- Dopo queste esclusioni si hanno 1473 cicli da 516 donne con 343 concepimenti

Obiettivo

- Si vuole utilizzare la metodologia proposta per studiare la relazione tra osservazioni giornaliere del muco cervicale e le probabilità giornaliere (*day specific*) di concepimento.
- Sia w_{ijk} l'indicatore del tipo di muco osservato nel giorno k del ciclo j della coppia i, con $w_{ijk} = 1$ che indica secco,..., fino a $w_{ijk} = 4$ che indica il tipo di muco più fertile.
- Si può considerare quindi la relazione tra tipo di muco (previsore di interesse) e probabilità di concepimento come monotona crescente.

Il modello e la specificazione a-priori

- Definiamo il vettore di previsori ponendo

$$\mathbf{u}_{ijk} = [\mathbb{I}(k = 1), \mathbb{I}(k = 2), \ldots, \mathbb{I}(k = 6), \\ \mathbb{I}(w_{ijk} \geq 2), \mathbb{I}(w_{ijk} \geq 3), \mathbb{I}(w_{ijk} = 4)]^{\mathrm{T}},$$

dove k indica il giorno nell'intervallo fertile, con $k = 1$ corrispondente a 5 giorni prima dell'ovulazione e $k = 6$ al giorno dell'ovulazione.

Il modello e la specificazione a-priori

- Definiamo il vettore di previsori ponendo

$$\mathbf{u}_{ijk} = [\mathbb{I}(k=1), \mathbb{I}(k=2), \ldots, \mathbb{I}(k=6), \\ \mathbb{I}(w_{ijk} \geq 2), \mathbb{I}(w_{ijk} \geq 3), \mathbb{I}(w_{ijk} = 4)]^{\mathrm{T}},$$

  dove k indica il giorno nell'intervallo fertile, con $k = 1$ corrispondente a 5 giorni prima dell'ovulazione e $k = 6$ al giorno dell'ovulazione.

- I coefficienti di regressione sono divisi in due sottovettori $\beta = [\beta_1^{\mathrm{T}}, \beta_2^{\mathrm{T}}]^{\mathrm{T}}$, con β_1^{T} che caratterizza i cambiamenti di base nelle probabilità di concepimento dovuti ai diversi giorni nell'intervallo fertile e β_2^{T} che caratterizza i cambiamenti tra i diversi livelli di muco.

Il modello e la specificazione a-priori

- Per scegliere le a-priori per i coefficienti $\lambda_k = \exp(\beta_{1k})$ per $k = 1, \ldots, 6$

Il modello e la specificazione a-priori

- Per scegliere le a-priori per i coefficienti $\lambda_k = \exp(\beta_{1k})$ per $k = 1, \ldots, 6$
  - si fissa la probabilità del punto massa a zero, $p_k = 0$, visto che la probabilità di concepimento deve essere positiva nell'intervallo ristretto di 6 giorni intorno all'ovulazione (Wilcox, Winberg e Baird, 1995)

Il modello e la specificazione a-priori

- Per scegliere le a-priori per i coefficienti $\lambda_k = \exp(\beta_{1k})$ per $k = 1, \ldots, 6$
  - si fissa la probabilità del punto massa a zero, $p_k = 0$, visto che la probabilità di concepimento deve essere positiva nell'intervallo ristretto di 6 giorni intorno all'ovulazione (Wilcox, Winberg e Baird, 1995)
  - si sceglie una a-priori diffusa per λ ponendo $a_{0k} = b_{0k} = 0.1$ per $k = 1, \ldots, 6$.

Il modello e la specificazione a-priori

- Per scegliere le a-priori per i coefficienti $\lambda_k = \exp(\beta_{1k})$ per $k = 1, \ldots, 6$
  - si fissa la probabilità del punto massa a zero, $p_k = 0$, visto che la probabilità di concepimento deve essere positiva nell'intervallo ristretto di 6 giorni intorno all'ovulazione (Wilcox, Winberg e Baird, 1995)
  - si sceglie una a-priori diffusa per λ ponendo $a_{0k} = b_{0k} = 0.1$ per $k = 1, \ldots, 6$.
- I coefficienti di regressione $\lambda_h = \exp(\beta_{2h})$ per $h = 1, 2, 3$ misurano i cambiamenti nelle probabilità di concepimento associate con il passaggio, rispettivamente dal tipo di muco 1 al tipo 2, dal tipo 2 al tipo 3 e dal tipo 3 al tipo 4.

Il modello e la specificazione a-priori

- Per scegliere le a-priori per i coefficienti $\lambda_k = \exp(\beta_{1k})$ per $k = 1, \ldots, 6$
  - si fissa la probabilità del punto massa a zero, $p_k = 0$, visto che la probabilità di concepimento deve essere positiva nell'intervallo ristretto di 6 giorni intorno all'ovulazione (Wilcox, Winberg e Baird, 1995)
  - si sceglie una a-priori diffusa per λ ponendo $a_{0k} = b_{0k} = 0.1$ per $k = 1, \ldots, 6$.
- I coefficienti di regressione $\lambda_h = \exp(\beta_{2h})$ per $h = 1, 2, 3$ misurano i cambiamenti nelle probabilità di concepimento associate con il passaggio, rispettivamente dal tipo di muco 1 al tipo 2, dal tipo 2 al tipo 3 e dal tipo 3 al tipo 4.
- quando il muco cresce la probabilità di concepimento non dovrebbe abbassarsi, dato il ruolo del muco nello spiegare il passaggio degli spermatozoi.

Il modello e la specificazione a-priori

Il modello e la specificazione a-priori

- Si può incorporare questo vincolo ponendo la regione $\mathcal{A}_h = (1, +\infty)$ nell'a-priori per λ_h.

Il modello e la specificazione a-priori

- Si può incorporare questo vincolo ponendo la regione $\mathcal{A}_h = (1, +\infty)$ nell'a-priori per λ_h.
- Incorporando un punto massa a $\gamma_h = 1$ per $h = 1, 2, 3$ si assegnano probabilità positive a-priori all'ipotesi che non ci sia alcun aumento nel passaggio da muco di tipo meno fertile a uno più fertile.

Il modello e la specificazione a-priori

- Si può incorporare questo vincolo ponendo la regione $\mathcal{A}_h = (1, +\infty)$ nell'a-priori per λ_h.
- Incorporando un punto massa a $\gamma_h = 1$ per $h = 1, 2, 3$ si assegnano probabilità positive a-priori all'ipotesi che non ci sia alcun aumento nel passaggio da muco di tipo meno fertile a uno più fertile.
- Si pone $p_h = 0.5^{1/3}$ per assegnare una a-priori di 0.5 all'ipotesi nulla di nessuna associazione tra tipo di muco e probabilità di concepimento giornaliere.

Il modello e la specificazione a-priori

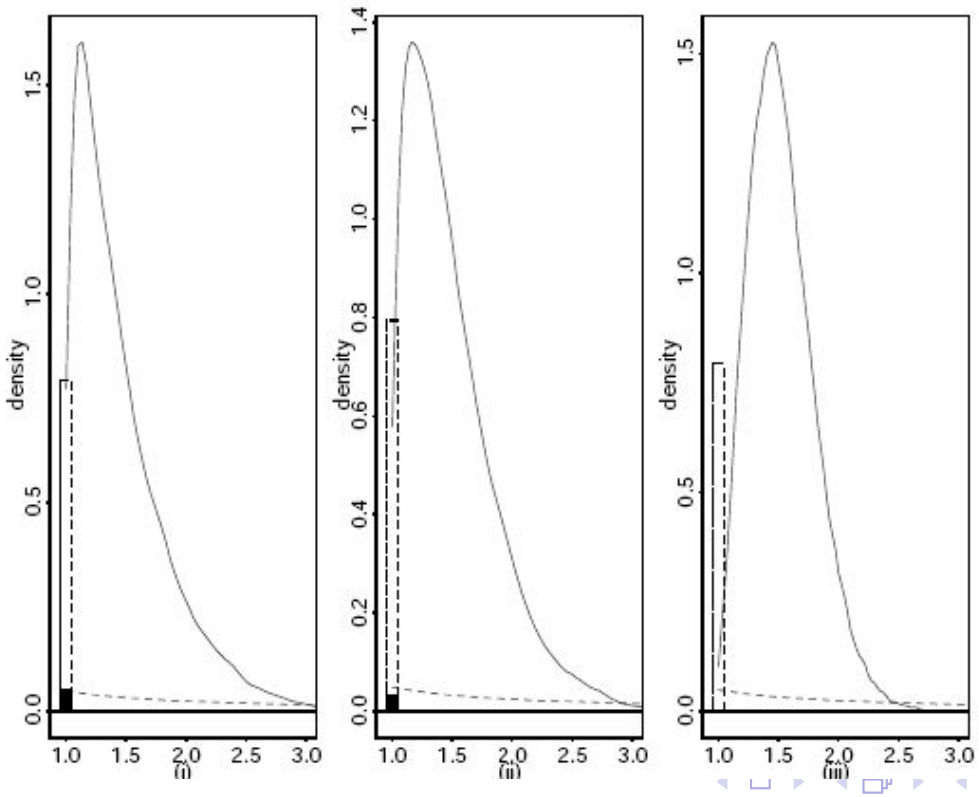
- Si può incorporare questo vincolo ponendo la regione $\mathcal{A}_h = (1, +\infty)$ nell'a-priori per λ_h.
- Incorporando un punto massa a $\gamma_h = 1$ per $h = 1, 2, 3$ si assegnano probabilità positive a-priori all'ipotesi che non ci sia alcun aumento nel passaggio da muco di tipo meno fertile a uno più fertile.
- Si pone $p_h = 0.5^{1/3}$ per assegnare una a-priori di 0.5 all'ipotesi nulla di nessuna associazione tra tipo di muco e probabilità di concepimento giornaliere.
- Si pone $a_h = b_h = 0.01$ per permettere un alto grado di incertezza per i valori di γ_h sotto l'ipotesi alternativa

Il modello e la specificazione a-priori

- Si può incorporare questo vincolo ponendo la regione $\mathcal{A}_h = (1, +\infty)$ nell'a-priori per λ_h.
- Incorporando un punto massa a $\gamma_h = 1$ per $h = 1, 2, 3$ si assegnano probabilità positive a-priori all'ipotesi che non ci sia alcun aumento nel passaggio da muco di tipo meno fertile a uno più fertile.
- Si pone $p_h = 0.5^{1/3}$ per assegnare una a-priori di 0.5 all'ipotesi nulla di nessuna associazione tra tipo di muco e probabilità di concepimento giornaliere.
- Si pone $a_h = b_h = 0.01$ per permettere un alto grado di incertezza per i valori di γ_h sotto l'ipotesi alternativa
- Si sceglie $c_1 = 1$ e $c_2 = 2$ per specificare una a-priori debolmente informativa per la varianza del parametro dell'effetto casuale

Risultati

- L'algoritmo MCMC è stato fatto procedere per 45.000 iterazioni di cui 5000 sono state escluse dalle analisi come *burn-in*
- La figura seguente mostra le densità a priori e a posteriori per i parametri legati al muco $\gamma_1, \gamma_2, \gamma_3$



Risultati

- Sebbene si sia usata una a-priori che assegna una moderatamente alta probabilità a $\gamma_h = 1$ per aggiustare possibili effetti di aumento degli errori di tipo I dovuto al fatto che si stanno facendo test multipli, c'è una chiara evidenza nei dati in favore di $\gamma_h > 1$.
- le probabilità a posteriori $P(\gamma_h = 1)$ sono rispettivamente 0.05, 0.03 e < 0.01.

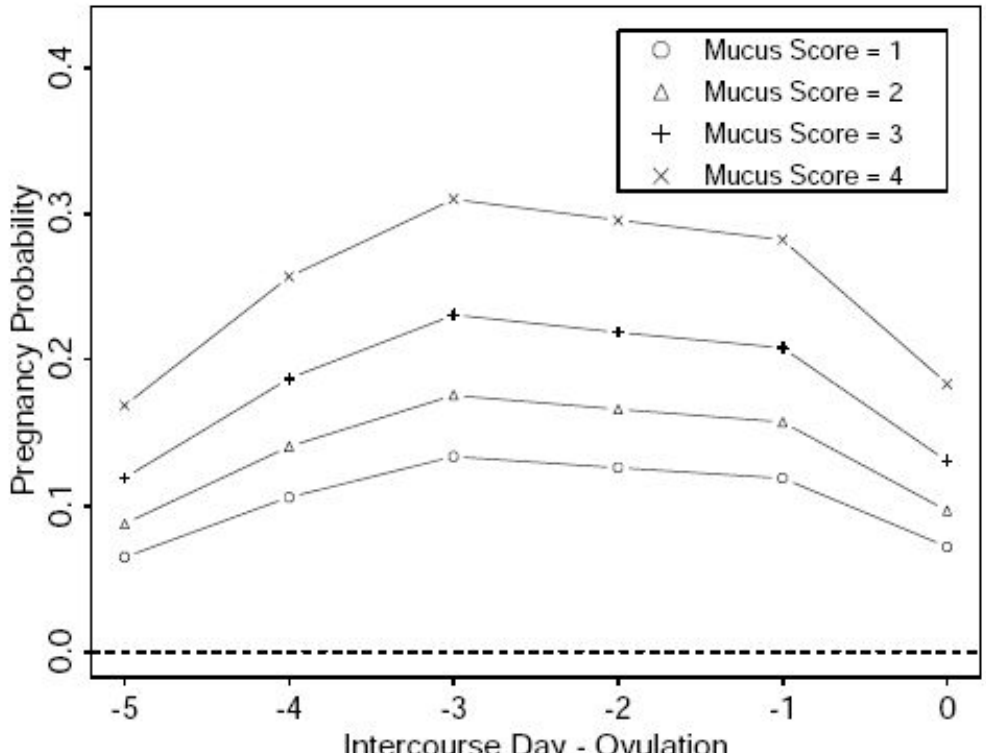
Risultati

- Ogni cambiamento del tipo di muco risulta quindi in un aumento significativo delle probabilità di concepimento giornaliere.
- Un riassunto delle distribuzioni a posteriori è presentato nella tabella seguente

Parametro	Media	Mediana	DS	intervallo di credibilità 95%
λ_1	0.07	0.07	0.02	[0.03, 0.12]
λ_2	0.12	0.12	0.03	[0.06, 0.20]
λ_3	0.16	0.15	0.05	[0.08, 0.26]
λ_4	0.15	0.14	0.04	[0.08, 0.25]
λ_5	0.14	0.13	0.04	[0.07, 0.22]
λ_6	0.08	0.07	0.03	[0.04, 0.14]
γ_1	1.43	1.32	0.40	[1, 2.47]
γ_2	1.47	1.38	0.38	[1, 2.44]
γ_3	1.53	1.50	0.27	[1.09, 2.15]
ϕ^{-1}	0.94	0.93	0.16	[0.65, 1.28]

Risultati

- Il tipo di muco sembra avere una capacità previsiva delle probabilità di concepimento maggiore del giorno relativo all'ovulazione come si può osservare chiaramente dal grafico seguente



...non senza fadiga si giunge al...

Fine