Modelagem Bayesiana

Caroline Vasconcelos

ÁCARO-VERMELHO-EUROPEU

O ácaro-vermelho-europeu possui cor vermelho-escura, medindo cerca de 0,7 mm de comprimento. Os machos são menores, mais delgados e de coloração mais clara do que as fêmeas. Vivem preferencialmente na parte inferior das folhas e tecem teia. Temperaturas quentes e clima seco favorecem a multiplicação. Seus prejuízos são provocados devido ao ataque nas folhas e brotações novas, causando inicialmente amarelamento seguido de bronzeamento, manchas necróticas e consequente redução do crescimento da planta. Quando o ataque é intenso o bronzeamento das folhas diminui a atividade fotossintética favorecendo a queda prematura das folhas.

Em 18 de julho de 1951, 25 folhas foram selecionadas aleatoriamente de cada uma das seis árvores McIntosh em um único pomar que recebeu o mesmo tratamento de pulverização, e o número de fêmeas adultas foi contado em cada folha. Vamos encontrar e estudar um modelo que nos permita fazer estimativas sobre o ácaro-vermelho-europeu com relação a sua frequência por folha.

No. ácaros/folha	Frequência
0	70
1	38
2	17
3	10
4	9
5	3
6	2
7	1
8	0

Como a distribuição Poisson é comumente usada para modelar dados de contagem, é natural que comecemos investigando se essa distribuição é adequada para o nosso problema.

Dada a verossimilhança da distribuição Poisson:

$$L(x;\lambda) = \frac{e^{-n\lambda}.\lambda^{\sum_{i=1}^{n}}}{\prod_{i=1}^{n} x_i!}$$

E utilizando a distribuição Gama como priori:

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

Usamos a seguinte posteriori para saber se o modelo Poisson é adequado:

$$p(\lambda|a,b) = \frac{b^a}{\Gamma(a)\prod_{i=1}^{n}}e^{-\lambda(b+n)}\lambda^{\sum_{i=1}^{n} x_i+a-1}$$

Como estamos utilizando uma priori conjugada, consequentemente, o parâmetro λ terá distribuição Gama:

$$\lambda \sim Gamma(\sum_{i=1}^{n} x_i + a, b+n)$$

Lançamos mão, então, da **Distribuição Preditiva da Posteriori** para saber se a escolha da distribuição poisson foi apropriada. Geradas as amostras via método de Monte Carlo simples, temos:

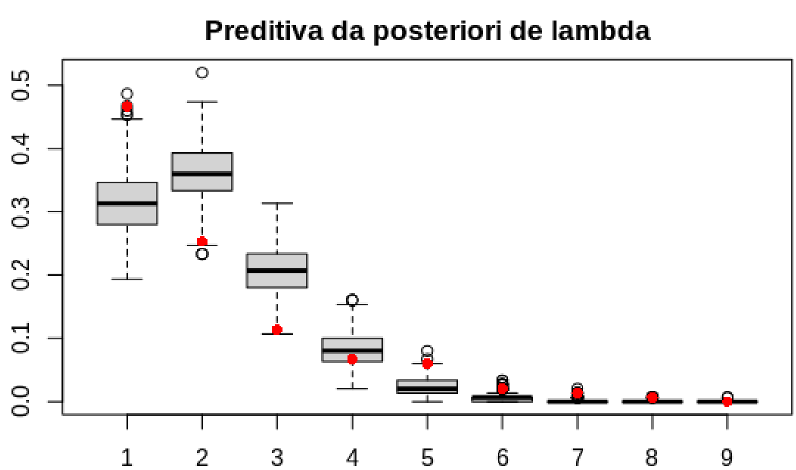


Figure 1: Preditiva da posteriori para distribuição Poisson.

Os pontos vermelhos são as frequências dos dados e os boxplots são as amostras geradas via método de Monte Carlo (MCMC) . Observamos, claramente, que os pontos não se ajustam às medianas dos boxplots, como deveriam, caso o modelo fosse adequado. Portanto, a distribuição poisson não é apropriada para esse conjunto de dados.

Outra distribuição que é usada para processos de contagem e também para modelar dados biológicos, é a **Distribuição Binomial Negativa**. Como os dados apresentam variação de taxas, pois nem todas as folhas recebem a mesma quantidade de pulverização ou não são pulverizadas com a mesma intensidade ou ainda, pode existir um efeito aleatório durante o processo, faz sentido usá-la como uma das fontes de informação da posteriori. Uma forma de contornar essa variação de taxa é utilizar a **Distribuição Beta**, também como uma das fontes de informação da posteriori, a priori.

Seja X$|\rho, \phi$ o número de ácaros/folha, onde x$|\rho, \phi$ ~ Binomial Negativa(ρ, ϕ), cuja função de probabilidade é dada por

$$p(x|\rho,\phi) = \frac{\Gamma(\phi + x)}{x!\Gamma(\phi)} \rho^{\phi}(1-\rho)^{x}$$

Onde $\rho \in$ (0,1), $\phi > 0$ e x $\in \mathbb{N}$.

Supondo que ϕ é conhecido e Beta é uma priori conjugada para este modelo, teremos:

$$f(x) = \frac{1}{B(\alpha,\beta)} x^{\alpha-1}(1-x)^{\beta-1}$$

A fonte de informação da verossimilhança é Binomial Negativa:

$$L(x;\rho) = \frac{\Gamma(\phi + x)^n}{\prod_{i=1}^{n} x_i \Gamma(\phi)^n} \rho^{n\phi}(1-\rho)^{\sum_{i=1}^{n} x_i}$$

Logo, a posteriori é dada por:

$$p(x|\rho,\phi,\alpha,\beta) = \frac{x_i^{\alpha-1}(1-x_i)^{\beta-1}}{B(\alpha,\beta)} \frac{\Gamma(\phi + x)^n}{\prod_{i=1}^{n} \Gamma(\phi)^n} \rho^{n\phi}(1-\rho)^{\sum_{i=1}^{n} x_i}$$

$$p(x|\rho,\phi,\alpha,\beta) = \frac{\Gamma(\alpha+\beta)^{\alpha-1} x_i (1-x_i)^{\beta-1}}{\Gamma(\alpha)+\Gamma(\beta)\prod_{i=1}^{n} x_i \Gamma(\phi)^n} \Gamma(\phi + x_i)^n \rho^{n\phi+\alpha}(1-\rho)^{\sum_{i=1}^{n} +\beta}$$

Isto posto, a distribuição da posteriori de ρ é:

$$\rho(\mathbf{x},\phi) \sim Beta(n\phi + \alpha, \sum_{i=1}^{n} +\beta)$$

Sendo $\pi(\phi)$ a priori para ϕ:

$$\pi(\rho,\phi|\mathbf{x}) \prec \prod_{i-1}^{n} \frac{\Gamma(\phi+x_i)}{\Gamma(\phi^n)} \rho^{n\phi}(1-\rho)^{\sum_{i=1}^{n}} \rho^{\alpha-1}(1-\rho)^{\beta-1}\pi(\phi)$$

$$\pi(\rho,\phi|\mathbf{x}) \prec \prod_{i-1}^{n} \frac{\Gamma(\phi+x_i)}{\Gamma(\phi^n)} \rho^{n\phi+\alpha-1}(1-\rho)^{\beta+\sum_{i-1}^{n}+1}\pi(\phi)$$

$$\pi(\rho,\phi|\mathbf{x}) \prec \prod_{i-1}^{n} \frac{\Gamma(\phi+x_i)}{\Gamma(\phi^n)} B(n\phi+\alpha,\beta+\sum_{i=1}^{n})\rho^{n\phi+\alpha-1}(1-\rho)^{\beta+\sum_{i-1}^{n}+1}\pi(\phi)$$

Onde $B(n\phi+\alpha,\beta+\sum_{i=1}^{n}) = \pi(\rho|x,\phi)$, temos que:

$$\pi(\rho,\phi|\mathbf{x}) \prec \pi(\rho|\mathbf{x},\phi)\pi(\phi|\mathbf{x})$$

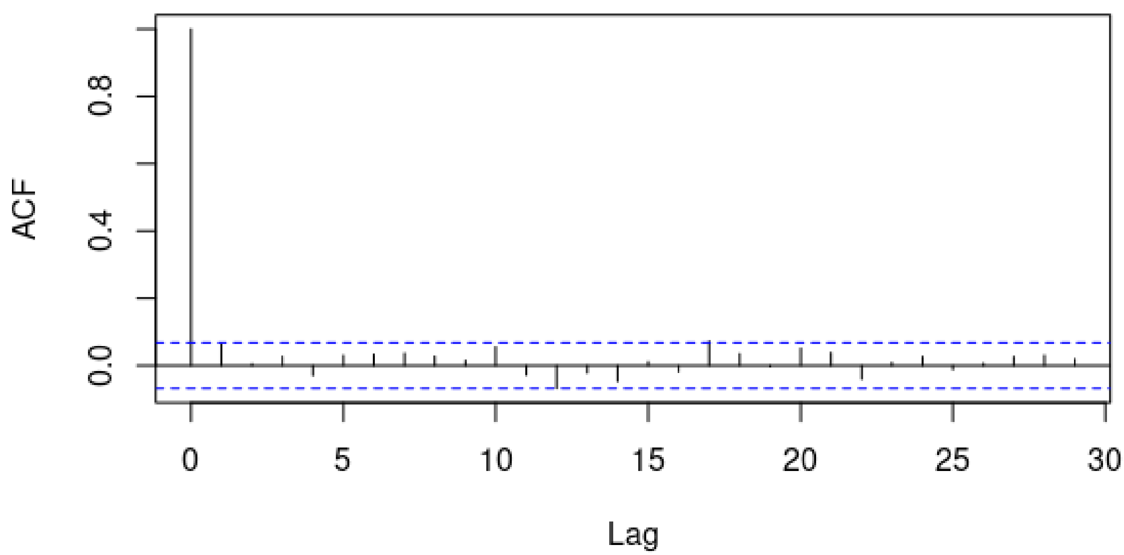
$$\cdots$$

$$\pi(\phi,\mathbf{x}) \prec \pi(\phi)B(\phi n+\alpha,\sum_{i=1}^{n}+\beta)\Gamma(\phi)^{-n}\prod_{i=1}^{n}\Gamma(\phi+x_i)$$

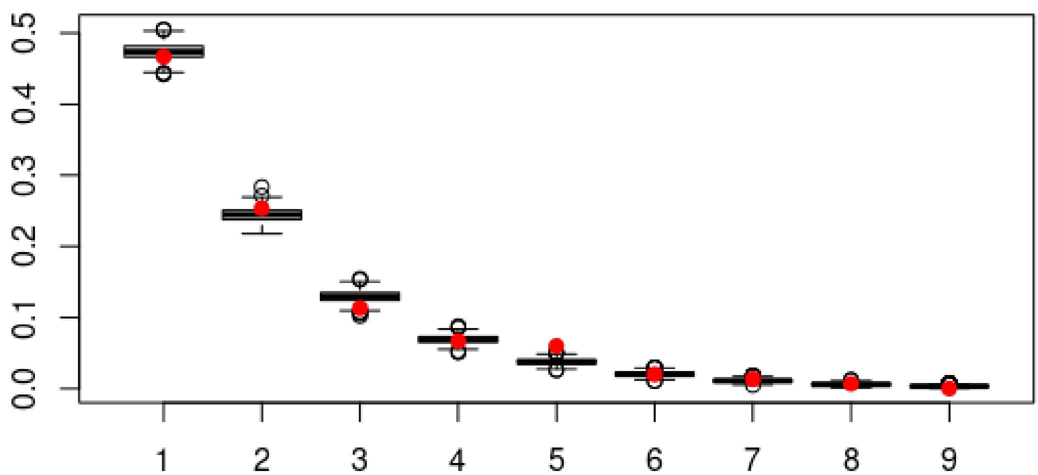
A priori escolhida para ϕ foi Gama, com hiperparâmetros $\alpha = 0.1$ e $\beta = 0.1$.

Verificamos, então, se as novas distribuições escolhidas para compor as informações da posteriori são pertinentes para o conjunto de dados através da preditiva da posteriori. O algoritmo **Metropolis-Hastings**, que também é um dos métodos de MCMC é uma forma conveniente de obter uma amostra simulada, a partir do uso de uma Cadeia de Markov generalizada para um espaço de estado contínuo, logo, uma desvantagem do M-H é que as amostras geradas são autocorrelacionadas. Como o objetivo é gerar um conjunto de amostras independentes e que reflete corretamente a distribuição, uma parte da amostra gerada é descartada e o restante passar por um refinamento para diminiur a autocorrelação.

A figura 2 mostra o correlograma das amostras geradas pelo algoritmo M-H. No gráfico, o eixo vertical indica a autocorrelação e o horizontal a defasagem. A linha tracejada azul indica onde é significativamente diferente de zero. Como é possível ver na imagem, praticamente todos os valores ACF estão dentro do limite da linha tracejada azul, indicando que a série é aleatória, conforme o esperado.



Quanto a preditiva da posteriori, os pontos vermelhos são as frequências dos dados e os boxplots são as amostras geradas, o gráfico da figura 3 mostra que os pontos se ajustam aos boxplots de maneira aceitável, portanto, o modelo se mostra adequado.



Com o modelo se mostrando adequado, podemos fazer **estimativas pontuais** e **intervalares** para os dados:

Table 2: Intervalos de confiança e médias estimadas para ρ e ϕ.

	$\overline{X}$	Limite inferior	Limite superior
ρ	0.442	0.3935	0.5240
ϕ	0.901	0.8335	1.0472

A tabela 2 mostra as médias estimadas para ρ e ϕ assim como também os limites intervalares estimados com 95% de nível de credibilidade.