# Análise bayesiana semiparamétrica de resposta binária com covariável contínua sujeita a omissão não aleatória

Frederico Z. Poleto
IME, Universidade de São Paulo, Brasil, *fpoleto@ime.usp.br*
Carlos Daniel Paulino
IST e CEAUL, Universidade de Lisboa, Portugal, *dpaulino@math.ist.utl.pt*
Julio M. Singer
IME, Universidade de São Paulo, Brasil, *jmsinger@ime.usp.br*
Geert Molenberghs
Hasselt University, Bélgica, *geert.molenberghs@uhasselt.be*

**Palavras–chave**: mistura por processo Dirichlet, dados incompletos, MNAR, regressão binária, análise bayesiana não paramétrica

**Resumo**: A omissão em variáveis explicativas requer um modelo para estas, mesmo que o interesse recaia apenas no modelo condicional para as respostas dadas as covariáveis. Uma especificação incorreta dos modelos para as covariáveis ou para o mecanismo de omissão pode levar a inferências enviesadas para os parâmetros de interesse. A literatura conhecida segue uma de duas vias: uso para as covariáveis de distribuições flexíveis, não paramétricas ou semiparamétricas, juntamente com uma suposição MAR, ou de distribuições paramétricas aliadas a um mecanismo de omissão mais geral de tipo MNAR. Considera-se aqui uma análise de variáveis respostas combinando um mecanismo MNAR com um modelo não paramétrico baseado numa mistura por processo Dirichlet para covariáveis contínuas sujeitas a omissão. A via descrita é ilustrada com dados simulados e também por análise de um conjunto de dados reais.

# 1 Introdução

Em muitos estudos surgem dados omissos para algumas das variáveis explicativas ($\mathbf{X}$) de tal modo que não se afigura conveniente excluir tais variáveis ou unidades amostrais da análise. Ainda que o interesse possa estar na distribuição condicional das variáveis respostas ($\mathbf{Y}$) dado $\mathbf{X}$, necessita-se de especificar também um modelo para a distribuição marginal das variáveis sofrendo omissão ou, pelo menos, para a distribuição condicional delas dadas as covariáveis que são sempre observadas.

Em casos onde pelo menos uma covariável é contínua, pode-se não ter *a priori* qualquer informação para um modelo paramétrico plausível. Suposições incorretas para o mecanismo de omissão ou para a distribuição das covariáveis podem gerar inferências enviesadas para a distribuição condicional das respostas dadas as covariáveis. Para fazer face à complexidade analítica adota-se pragmaticamente uma metodologia bayesiana para a análise de um modelo global composto de distribuições paramétricas condicionais para $\mathbf{Y}$ dado $\mathbf{X}$ e de um modelo flexível para $\mathbf{X}$ concretizado numa mistura não paramétrica por um processo Dirichlet (Ishwaran e James [2]), aliado a um mecanismo de omissão que se permite ser não aleatório. Para simplicidade restringir-se-á o modelo semiparamétrico ao caso de uma única covariável contínua sujeita a omissão.

O resto do artigo é desenhado da seguinte forma. Na Secção 2 introduz-se o modelo não paramétrico para uma variável contínua, o qual vai constituir uma componente dos modelos semiparamétricos abordados a seguir, em especial na Secção 3. O modelo semiparamétrico desta secção vai ser comparado com modelos paramétricos alternativos no quadro de um estudo de simulação na Secção 4. Um conjunto de dados reais é analisado na Secção 5 através de um modelo semiparamétrico que se propõe incorporar particularmente alguns juízos apriorísticos sobre o mecanismo de omissão que tornam a sua modelação diferente da referida nas secções anteriores. O artigo termina com uma referência a breves conclusões.

# 2 Modelo não paramétrico para dados completos contínuos

Seja $X_i$, $i = 1, \ldots, n$, uma amostra aleatória de tamanho $n$ de uma função de distribuição $F$. Num quadro paramétrico supõe-se uma forma conhecida para $F$, indexada por um parâmetro dimensionalmente finito especificado a priori mas geralmente desconhecido. Para permitir uma maior flexibilidade na modelação e robustez contra uma incorreta especificação de $F$, consideram-se modelos não paramétricos.

Um modo de evitar a especificação da forma de $F$ é empregar medidas de probabilidade aleatórias (RPM), que são distribuições de probabilidade sobre o espaço de medidas de probabilidade, tal como o chamado processo Dirichlet (DP) simbolizado por $F \sim \mathrm{DP}(\alpha, F_0)$. Este processo significa que, para qualquer partição mensurável do espaço amostral $A_1, \ldots, A_M$, o vetor probabilístico de componentes $F(A_j), j = 1, \ldots, M$ segue uma distribuição Dirichlet com vetor paramétrico $[\alpha F_0(A_1), \ldots, \alpha F_0(A_M)]$, onde $\alpha$ é um parâmetro de precisão e $F_0$ é uma distribuição de referência sobre o espaço amostral. Com esta parametrização, $F_0$ é a esperança *a priori* da distribuição $F$ e à medida que $\alpha$ aumenta há uma maior concentração de $F$ em torno de $F_0$. Todavia, sabe-se que o DP gera (talvez inesperadamente) uma distribuição discreta quase certamente, o que pode não ser apropriado para muitas aplicações.

Um modo de gerar uma RPM compatível com distribuições absolutamente contínuas é supor que $X_i$ segue uma distribuição absolutamente contínua dado um valor de um parâmetro específico $\theta_i$ e que, por sua vez, $\theta_i$, $i = 1, \ldots, n$, constitui uma amostra aleatória de um DP, i.e., $X_i|\theta_i \overset{\text{ind.}}{\sim} F_{\theta_i}$, $i = 1, \ldots, n$, $(\theta_1, \ldots, \theta_n)|G \overset{\text{i.i.d.}}{\sim} G$, $G|(\alpha, G_0) \sim \mathrm{DP}(\alpha, G_0)$. Admitindo que os parâmetros $\{\theta_i\}$ seguem uma distribuição *a priori* do tipo DP centrada em $G_0$, em vez da abordagem comum de supor que eles seguem diretamente uma distribuição paramétrica $G_0$, acrescenta uma flexibilidade desejável ao modelo. O termo mistura por processo Dirichlet (DPM) advém da formulação

hierárquica que implica que a distribuição marginal para $X_i$ é uma mistura, i.e., $f(x_i) = \int f(x_i|\theta_i)dG(\theta_i)$, $G|(\alpha,G_0) \sim \text{DP}(\alpha,G_0)$.
A definição construtiva do DP (Sethuraman [8]) mostra que $G|(\alpha,G_0) \sim \text{DP}(\alpha,G_0)$ pode ser representado por $G(A) = \sum_{j=1}^{\infty} p_j \delta_{\theta_j}(A)$ para qualquer subconjunto mensurável $A$ do espaço de valores de $\{\theta_j\}$, onde $p_1 = V_1$, $p_j = V_j \prod_{k=1}^{j-1}(1 - V_k)$,$j > 1$, $V_j \overset{\text{i.i.d.}}{\sim} \text{Beta}(1,\alpha)$, $j = 1,2,\ldots$, $\delta_{\theta_j}(A)$ é a medida de Dirac (i.e., igual a um se $\theta_j \in A$ e a zero, no caso contrário), e $\theta_j \overset{\text{i.i.d.}}{\sim} G_0$, $j = 1,2,\ldots$. Esta construção dos pesos aleatórios $\{p_j\}$ é rotulada de procedimento quebra-vara (*stick-breaking*). Tal representação do DP permite delinear algoritmos eficientes para ajustar modelos DPM reescrevendo a distribuição marginal da mistura como $f(x_i) = \sum_{j=1}^{\infty} p_j f(x_i|\theta_j)$, $i = 1,\ldots,n$.
Na prática, contudo, por razões de simplicidade é comum truncar a mistura a $M$ componentes (veja, e.g., Ishwaran e James [2]), o que equivale a aproximar $\text{DP}(\alpha,G_0)$ por um DP truncado (TDP), denotado por $\text{TDP}(\alpha,G_0,M)$. Neste caso, para obter os pesos $p_1,\ldots,p_M$, geram-se as variáveis $V_j \sim \text{Beta}(1,\alpha)$, $j = 1,\ldots,M-1$, e fixa-se $V_M = 1$. A opção por um TDP permite implementar o modelo em software disponível, como BUGS, JAGS e R; respetivamente, *vide, e.g.*, Lunn *et al.* [4], Plummer [5] e R Core Team [7]).
A escolha de $M$ é a questão-chave da via de recorrer a distribuições *a priori* TDP. Recorrendo a resultados de Antoniak [1] sobre o DP e à distribuição *a posteriori* do número de valores distintos dos $\{\theta_i\}$ e de $\alpha$ pode-se definir valores razoáveis para o ponto $M$ de truncatura (*vide* Poleto, Paulino, Singer e Molenberghs [6]).

# 3 Um modelo semiparamétrico para respostas binárias com uma covariável contínua sujeita a omissão não aleatória

Seja $Y_i$ uma resposta binária sempre observada, $X_i$ uma covariável contínua com valores potencialmente omissos e $R_i$ uma variável in-

dicadora assumindo o valor 1 se $X_i$ é observado e 0, se $X_i$ é omisso, $i = 1,\ldots,n$. Embora o interesse se concentre na distribuição condicional de $Y_i$ dado $X_i$, é necessário considerar um modelo para $X_i$ pois não se deseja desprezar a porção da amostra em que $X_i$ é omisso. Como se admite que o mecanismo gerador de dados omissos pode depender dos próprios valores não observados, necessita-se de modelar $R_i$. Adotando a denominada fatorização em modelo de seleção, considera-se o modelo

$$R_i|(Y_i,X_i,\delta_0,\delta_1,\delta_2,\delta_3) \overset{\text{ind.}}{\sim} \text{Bern}(\theta_i),\ \text{logito}(\theta_i) = \delta_0 + \delta_1 X_i + \delta_2 Y_i + \delta_3 X_i Y_i, \quad (1)$$

$$Y_i|(X_i,\beta_0,\beta_1) \overset{\text{ind.}}{\sim} \text{Bern}(\pi_i),\ \text{logito}(\pi_i) = \beta_0 + \beta_1 X_i, \quad (2)$$

$$X_i|(\mu_i,V) \overset{\text{ind.}}{\sim} N(\mu_i,V), \quad (3)$$

em que $\text{Bern}(\theta_i)$ denota a distribuição Bernoulli com probabilidade de sucesso $\theta_i$, $i = 1,\ldots,n$, juntamente com as distribuições *a priori* mutuamente independentes $\delta_j|(\mu_{\delta_j},\sigma_{\delta_j}) \overset{\text{ind.}}{\sim} N(\mu_{\delta_j},\sigma_{\delta_j})$, $j = 0,1,2,3$, $\beta_j|(\mu_{\beta_j},\sigma_{\beta_j}) \overset{\text{ind.}}{\sim} N(\mu_{\beta_j},\sigma_{\beta_j})$, $j = 0,1$, $(\mu_1,\ldots,\mu_n)|G \overset{\text{i.i.d.}}{\sim} G$, $G|\alpha,G_0,M \sim \text{TDP}(\alpha,G_0,M)$, $V|T \sim \text{Unif}[0,T]$, $\alpha|(\lambda_1,\lambda_2) \sim Ga(\lambda_1,\lambda_2)$, $G_0\ |(\mu_0,\tau) = N(\mu_0,\tau)$, $\mu_0|(a,A) \sim N(a,A)$. Recorde-se que o acrónimo TDP designa um processo Dirichlet truncado em que o ponto de truncatura $M$ foi baseado no argumento avançado por Antoniak [1] e Ishwaran e James [2]. Uma justificativa para o uso da distribuição *a priori* uniforme para a variância $V$, em vez da comummente empregada distribuição gama, é apresentada por Ishwaran e James [2] e Poleto *et al.* [6].

O modelo é considerado semiparamétrico porque emprega uma estrutura dita não paramétrica (na realidade, massivamente paramétrica) para o modelo marginal de $X_i$ e estruturas paramétricas convencionais para as distribuições condicionais de $Y_i$ dado $X_i$ e $R_i$ dado $Y_i$ e $X_i$.

O mecanismo (1) é do tipo omissão não ao acaso (MNAR de *missing not at random*) porque considera que a probabilidade de ocorrerem

covariáveis omissas pode depender dos seus próprios valores não observados. Por outro lado, se se incluir a suposição de omissão ao acaso (MAR de *missing at random*) $\delta_1 = \delta_3 = 0$, o mecanismo torna-se ignorável do ponto de vista de inferências bayesianas para $\beta_0$ e $\beta_1$ devido à suposta independência apriorística entre ($\delta_0$,$\delta_2$) e os outros parâmetros (Little e Rubin [3]). Uma subclasse do modelo MAR é o mecanismo de omissão completamente ao acaso (MCAR de *missing completely at random*) que pode ser formulado fixando $\delta_1 = \delta_2 = \delta_3 = 0$.

Neste cenário com omissão em variáveis explicativas, é importante notar que a chamada análise de casos completos (CCA de *complete case analysis*), na qual se desprezam as unidades com dados omissos, gera usualmente inferências não enviesadas para $\beta_0$ e $\beta_1$, não só sob o mecanismo MCAR mas também sob quaisquer outros mecanismos que não dependem de $Y_i$ tais como na versão reduzida do mecanismo de omissão não ao acaso, $\text{MNAR}_{\text{red}} : \delta_2 = \delta_3 = 0$. A CCA em dados gerados sob o mecanismo $\text{MNAR}_{\text{red}}$ resulta em inferências enviesadas para a distribuição marginal de $X_i$, mas não para a distribuição condicional de $Y_i$ dado $X_i$. Note-se que a CCA não requer a especificação dum modelo marginal para $X_i$ se o interesse jaz apenas na distribuição condicional de $Y_i$ dado $X_i$.

# 4 Alguns resultados de um estudo de simulação

Considera-se as seguintes distribuições para a variável explicativa: $X^N \sim N(12,3^2)$, $X^L \sim$ Log-normal($2.45,0.246^2$), $X^C = 0.8X^{C1} + 0.2X^{C2}$, $X^{C1} \sim$ Unif[8,12], $X^{C2} \sim$ Log-normal($2.79,0.642^2$), em que Log-normal($\mu$,$\sigma^2$) denota uma distribuição log-normal e $\mu$ e $\sigma$ são respetivamente a média e o desvio padrão da variável subjacente na escala logarítmica. A média e o desvio padrão de $X^L$ e $X^C$ coincidem com os correspondentes parâmetros de $X^N$, embora as densidades sejam muito diferentes conforme ilustrado na Figura 1.

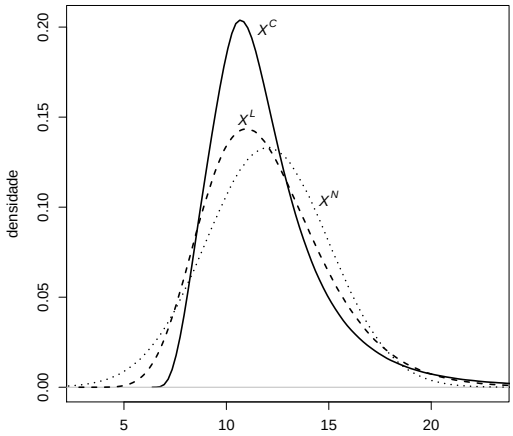


Figura 1: Densidades de distribuições normal ($X^N$), log-normal ($X^L$) e combinação linear ($X^C$) de uma uniforme e uma log-normal, com a mesma média e desvio padrão.

Com o propósito de avaliar o impacte de resultados obtidos sob diferentes suposições distribucionais para a covariável, simulou-se uma amostra de $X$ de tamanho $n = 10000$ de cada uma de três distribuições $X^N$, $X^L$ e $X^C$; de seguida, para cada valor simulado sob cada uma das distribuições da covariável gerou-se $Y$ de (2) com $\beta_0 = 6$ e $\beta_1 = -0.5$; finalmente, gerou-se $R$ de (1) com $\delta_0 = -3$, $\delta_1 = 0.5$ e $\delta_2 = \delta_3 = 0$. Para cada um dos três conjuntos de dados simulados (com $X^N$, $X^L$ e $X^C$), ajustou-se o modelo semiparamétrico da secção anterior bem como os modelos paramétricos normal e log-normal. Para estes últimos o modelo não paramétrico com o 1º nível (3) é substituído por $X_i|\mu_0,\tau \overset{\text{i.i.d.}}{\sim} N(\mu_0,\tau)$ e $X_i|\mu_0,\tau \overset{\text{i.i.d.}}{\sim}$ Log-normal$(\mu_0,\tau)$, $i = 1,\ldots,n$, dotado de distribuições

Tabela 1: Valores esperados (VE), desvios padrões (DP) e intervalos de credibilidade equicaudais (IC) a 95% da distribuição *a posteriori*.

| Análise de Casos Disponíveis | | | | | | | |
|---|---|---|---|---|---|---|---|
| Distr. | Covariável | $\beta_0$ | | | $\beta_1$ | | |
| Gerado | Suposto | VE | DP | IC 95% | VE | DP | IC 95% |
| $X^N$ | Normal | 6.22 | 0.15 | [5.93; 6.51] | −0.515 | 0.012 | [−0.538; −0.491] |
| | Log-norm. | 6.37 | 0.14 | [6.09; 6.66] | −0.525 | 0.012 | [−0.549; −0.502] |
| | Não-Param. | 6.21 | 0.15 | [5.92; 6.51] | −0.514 | 0.012 | [−0.538; −0.491] |
| $X^L$ | Normal | 5.06 | 0.13 | [4.82; 5.32] | −0.428 | 0.011 | [−0.449; −0.407] |
| | Log-norm. | 6.01 | 0.14 | [5.73; 6.29] | −0.501 | 0.012 | [−0.525; −0.478] |
| | Não-Param. | 6.00 | 0.14 | [5.71; 6.28] | −0.500 | 0.012 | [−0.524; −0.477] |
| $X^C$ | Normal | 4.72 | 0.12 | [4.49; 4.96] | −0.395 | 0.010 | [−0.416; −0.375] |
| | Log-norm. | 5.08 | 0.13 | [4.83; 5.34] | −0.425 | 0.011 | [−0.447; −0.404] |
| | Não-Param. | 5.78 | 0.15 | [5.49; 6.08] | −0.481 | 0.013 | [−0.505; −0.456] |
| Análise de Casos Completos | | | | | | | |
| $X^N$ | | 6.22 | 0.15 | [5.93; 6.52] | −0.515 | 0.012 | [−0.539; −0.492] |
| $X^L$ | | 6.02 | 0.14 | [5.74; 6.30] | −0.502 | 0.012 | [−0.525; −0.479] |
| $X^C$ | | 5.83 | 0.15 | [5.54; 6.12] | −0.484 | 0.012 | [−0.509; −0.460] |
| Valores Verdadeiros | | $\beta_0 = 6.00$ | | | $\beta_1 = -0.500$ | | |

*a priori* vagas. Para todos os modelos as distribuições *a priori* vagas para $\delta_j$ e $\beta_j$ envolveram os hiperparâmetros $\mu_{\delta_j} = \mu_{\beta_j} = 0$ e $\sigma_{\delta_j} = \sigma_{\beta_j} = 10^3$, $j = 0{,}1$. Analogamente, $M = 10$, $T = s_x^2$ (variância empírica dos valores de $X$), $\tau = 16s_x^2$, $\lambda_1 = \lambda_2 = 2$, $a = 0$ e $A = 10^3$. Além disso, supôs-se sempre a estrutura correta para o mecanismo de omissão, i.e., $\delta_2 = \delta_3 = 0$, de modo que as únicas componentes que variavam no estudo eram a distribuição usada para gerar a covariável e a distribuição admitida para esta na análise.

De acordo com a Tabela 1, as amostras obtidas das distribuições *a posteriori* dos parâmetros $\beta_0$ e $\beta_1$ indicam que o modelo não paramétrico para a covariável gera resultados muito próximos dos obtidos com o correspondente modelo paramétrico verdadeiro sob qualquer das distribuições normal e log-normal. Nesses casos, os intervalos de credibilidade contêm os verdadeiros valores de $\beta_0$ e $\beta_1$; isto não ocorre nas análises sob modelos paramétricos incorretos para $X^N$ e $X^L$. Por outro lado, no caso de $X^C$, só os intervalos de credibilidade

da análise sob o modelo não paramétrico para a covariável continham os verdadeiros valores de $\beta_0$ e $\beta_1$. A CCA produz resultados para os parâmetros do modelo logístico muito próximos dos obtidos com todos os dados disponíveis, o que poderia ser antecipado para este modelo MNAR identificável.

# 5 Análise de dados de embolia pulmonar

Wicki *et al.* [9] analisaram dados de 1090 pacientes que foram consecutivamente admitidos no banco de urgência do Hospital Universitário de Genebra por suspeita de embolia pulmonar, i.e., bloqueio da artéria pulmonar ou de alguma das suas ramificações. O objetivo do seu estudo era desenvolver um sistema de pontuação que indicasse a probabilidade de ocorrência desta doença cardiovascular baseado em testes de diagnóstico e outra informação facilmente obtida. Por simplicidade, considera-se aqui só algumas das variáveis explicativas incluídas no modelo final apresentado por estes autores.

O indicador da presença de embolia pulmonar (variável resposta), bem como quatro variáveis explicativas (idade, embolia pulmonar prévia ou trombose venosa profunda, cirurgia recente e frequência cardíaca), foram observadas para todos os pacientes, enquanto duas variáveis indicando presença de certas caraterísticas (atelectasia laminar e elevação do hemidiafragma) apresentaram valores faltantes para um único paciente que, por esta razão, foi removido do conjunto de dados. Por outro lado, a pressão parcial do dióxido de carbono($PaCO_2$), obtida por gasometria arterial, surgiu omissa para 103 (9%) pacientes.

Análises preliminares permitiram mostrar que os dados observados para $PaCO_2$ parecem ser melhor acomodados pela distribuição preditiva *a posteriori* do modelo não paramétrico do que pelas correspondentes densidades dos modelos normal, log-normal e gama (vide Figura 2 para o modelo gama, que forneceu um melhor ajuste do que o normal e o log-normal). Por outro lado, elas não mostraram evidência de associação entre $PaCO_2$ e as outras variáveis explicativas.

Tendo isto em mente considerou-se um modelo não paramétrico marginal em vez de condicional para $PaCO_2$, do género daquele descrito na Secção 3.

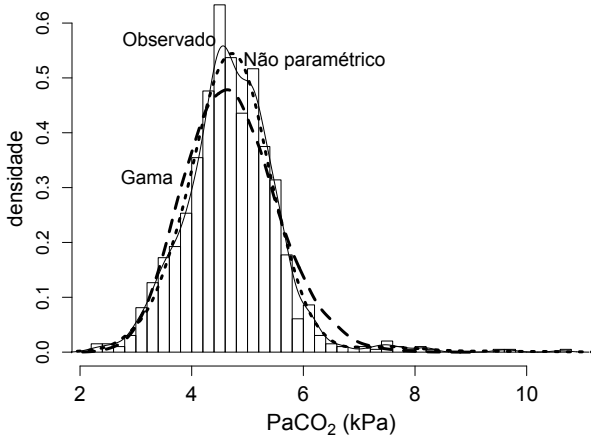


Figura 2: Histograma de dados observados para $X = PaCO_2$ e estimativas da densidade pelo método de núcleo gaussiano baseado em dados e valores amostrados da distribuição preditiva *a posteriori* decorrente do ajuste dos modelos não paramétrico (pontilhado) e gama (tracejado).

Wicki *et al.* [9] mencionam que $PaCO_2$ estava em falta para alguns pacientes porque a gasometria arterial não foi realizada ou foi executada enquanto os pacientes estavam respirando oxigénio. Perrier (comunicação pessoal) afirma que isso ocorreu para pacientes que estavam muito pouco doentes ou tão doentes que necessitavam da administração de oxigénio. Contudo, não foi registrado em qual dos dois casos os pacientes com dados de $PaCO_2$ faltantes seriam classi-

Tabela 2: Valores esperados (VE) e desvios padrões (DP) *a posteriori* e intervalos de credibilidade equicaudais (IC) a 95%.

| Parâ- | Modelo não paramétrico | | | Análise Casos Completos | | |
|---|---|---|---|---|---|---|
| metros | VE | DP | IC 95% | VE | DP | IC 95% |
| $\beta_0$ | $-2.476$ | 0.642 | $[-3.727; -1.192]$ | $-2.585$ | 0.672 | $[-3.901; -1.273]$ |
| $\beta_1$ | 1.375 | 0.268 | $[\ 0.852;\ 1.903]$ | 1.512 | 0.293 | $[\ 0.943;\ 2.086]$ |
| $\beta_2$ | 1.080 | 0.177 | $[\ 0.735;\ 1.429]$ | 1.087 | 0.187 | $[\ 0.724;\ 1.453]$ |
| $\beta_3$ | 0.706 | 0.187 | $[\ 0.339;\ 1.070]$ | 0.732 | 0.200 | $[\ 0.343;\ 1.126]$ |
| $\beta_4$ | 0.590 | 0.189 | $[\ 0.221;\ 0.962]$ | 0.591 | 0.201 | $[\ 0.198;\ 0.990]$ |
| $\beta_5$ | 0.268 | 0.046 | $[\ 0.179;\ 0.359]$ | 0.288 | 0.048 | $[\ 0.195;\ 0.384]$ |
| $\beta_6$ | 1.158 | 0.331 | $[\ 0.508;\ 1.809]$ | 1.221 | 0.352 | $[\ 0.538;\ 1.914]$ |
| $\beta_7$ | $-0.405$ | 0.101 | $[-0.609; -0.209]$ | $-0.429$ | 0.102 | $[-0.631; -0.231]$ |
| $\delta_0$ | 2.624 | 0.223 | $[\ 2.200;\ 3.077]$ | | | |
| $\delta_1$ | 0.482 | 0.210 | $[\ 0.082;\ 0.914]$ | | | |
| $\delta_2$ | $-0.112$ | 0.250 | $[-0.587;\ 0.399]$ | | | |

ficados. Os comentários de Perrier sugerem que é razoável supor que a probabilidade de se observar $PaCO_2$ ($\theta_i$) pode (i) ser máxima para pacientes com probabilidade de embolia pulmonar ($\pi_i$) próxima da prevalência de embolia pulmonar ($\pi$) e (ii) diminuir, à medida que a probabilidade de embolia pulmonar fica mais distante da prevalência. Tendo isto em vista, propõe-se um modelo de omissão com uma regressão segmentada que permite que $\theta_i$ decaia com velocidade diferente à medida que $\pi_i \longrightarrow 0$ e $\pi_i \longrightarrow 1$ e, assim, espera-se que $\delta_1 > 0$ e $\delta_2 < 0$. Este modelo juntamente com um modelo condicional para a resposta indicando embolia pulmonar dadas as covariáveis são indicados como segue: $R_i|(\delta_0,\delta_1,\delta_2,\mathrm{LN}_i,\mathrm{LP}_i) \overset{\text{ind.}}{\sim} \mathrm{Bern}(\theta_i)$, $\mathrm{logito}(\theta_i) = \delta_0 + \delta_1\mathrm{LN}_i + \delta_2\mathrm{LP}_i$, $Y_i|(\beta_0,\{X_{ji},\beta_j,j = 1,\ldots,7\}) \overset{\text{ind.}}{\sim} \mathrm{Bern}(\pi_i)$, $\mathrm{logito}(\pi_i) = \beta_0 + \sum_{j=1}^{7}\beta_j X_{ji}$, para $i = 1,\ldots,n$, onde $Y_i$ é a variável indicadora de embolia pulmonar, as variáveis explicativas $X_{1i},\ldots,X_{7i}$ são, respetivamente (i) um indicador de recente cirurgia, (ii) um indicador de anterior embolia pulmonar ou de profunda trombose venosa , (iii) um indicador de atelectasia pulmonar em radiografia peitoral (AL-RX), (iv) um indicador de elevação dum hemidiafragma em radiografia peitoral (EH-RX), (v) idade, em déca-

das, (vi) frequência cardíaca em centenas de batimentos por minuto (bpm) e (vii) pressão parcial de dióxido de carbono ($PaCO_2$) em kPa, $R_i$ é o indicador de observação de $PaCO_2$ ($X_{7i}$), $LN_i = LC_i$, se $LC_i < 0$, e $LN_i = 0$, caso contrário, $LP_i = LC_i$, se $LC_i > 0$, e $LP_i = 0$, caso contrário e $LC_i = \text{logito}(\pi_i) - \text{logito}(\hat{\pi})$, em que $\hat{\pi}$ é a prevalência de embolia pulmonar no hospital, supostamente conhecida e dada pela proporção de embolia pulmonar na amostra (27%). A componente DPM para $PaCO_2$ do modelo global é idêntica à referida no modelo da Secção 3.

A Tabela 2 exibe alguns resultados *a posteriori* das análises de casos disponíveis mediante o modelo TDP para $X_{7i}$ e de casos completos para propósitos comparativos.

As análises de todos os dados disponíveis baseados no modelo global referido acima acabam por ser mais adequadas do que as análises de casos completos porque, ao incorporarem suposições sobre o modo de ocorrência dos dados omissos, devem proporcionar resultados menos enviesados sobre a associação entre embolia pulmonar e $PaCO_2$, e gerar resultados mais precisos para as outras associações.

# 6 Conclusões

Este artigo centra-se na modelação de respostas binárias quando há uma covariável contínua sujeita a omissão informativa (MNAR). Mostra-se que uma abordagem bayesiana com um modelo semiparamétrico baseado numa mistura por processo Dirichlet para a distribuição marginal daquela covariável é uma alternativa viável para evitar eventuais vieses em inferências de interesse introduzidos pela adoção de uma distribuição paramétrica incorreta.

Algumas extensões podem ser tomadas em consideração como a possibilidade de se ter duas ou mais covariáveis contínuas sujeitas a omissão. Nesse caso, o modelo para o mecanismo de omissão pode assentar num produto de distribuições bernoullianas univariadas e o modelo para as covariáveis pode ser baseado em DPM unidimensionais. Havendo possibilidade de omissão também na variável res-

posta, tal pode ser manuseado através de modelação simultânea dos respetivos indicadores de observação/omissão.

Com ou sem as extensões referidas, os maiores desafios na aplicação dos modelos contemplados são provavelmente os casos em que suposições para o mecanismo de omissão originam modelos inidentificáveis e em que o tamanho amostral é grande em oposição a escassa informação *a priori*. Nessas circunstâncias, torna-se lento o processo computacional de gerar as distribuições *a posteriori* de interesse via MCMC, nomeadamente pelo elevado número de componentes do TDP e pela autocorrelação das respectivas cadeias. Nesses casos especiais, algumas das análises como as realizadas por Poleto *et al.* [6] podem levar alguns dias para serem produzidas, até mesmo para modelos paramétricos. Por outro lado, análises de modelos identificáveis como as realizadas neste artigo podem ser produzidas em algumas horas.

# Agradecimentos

Expressa-se com gratidão os apoios financeiros concedidos a este trabalho de investigação: Frederico Z. Poleto e Julio M. Singer, pela Coordenação de Aperfeiçoamento de Pessoal de Nível Superior (CAPES), Brasil, Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), Brasil, e Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), Brasil; Carlos Daniel Paulino, pela Fundação para a Ciência e Tecnologia (FCT) através da unidade CEAUL-FCUL, Portugal e projetos Pest-OE/MAT/UI0006 de 2011 e 2014; Geert Molenberghs, por IAP research network P6/03 do Governo Belga (Belgian Science Policy). Os autores agradecem ao Dr. Arnaud Perrier e ao Dr. Henri Bounameaux do Hospital Universitário de Genebra por fornecerem o conjunto de dados.

# Referências


[1] Antoniak, C.E. (1974). Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems. *The Annals of Statistics* 2, 1152–1174.

[2] Ishwaran, H., James, L.F. (2002). Approximate Dirichlet process computing finite normal mixtures: smoothing and prior information. *Journal of Computational and Graphical Statistics* 11, 508–532.

[3] Little, R.J.A., Rubin, D.B. (2002). *Statistical Analysis with Missing Data*, 2nd ed. John Wiley & Sons, New York.

[4] Lunn, D.J., Spiegelhalter, D., Thomas, A. e Best, N. (2009). The BUGS project: evolution, critique and future directions (with discussion). *Statistics in Medicine* 28, 3049–3082.

[5] Plummer, M. (2003). JAGS: A program for analysis of Bayesian graphical models using Gibbs sampling. *Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003)*, 20–22.

[6] Poleto, F., Paulino, C.D., Singer, J. e Molenberghs, G. (2014). Semiparametric Bayesian analysis of binary responses with a continuous covariate subject to non-random missingness. *Statistical Modeling* (no prelo).

[7] R Core Team (2014). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna.

[8] Sethuraman, J. (1994). A constructive definition of Dirichlet priors. *Statistica Sinica* 4, 639–650.

[9] Wicki, J., Perneger, T.V., Junod, A.F., Bounameaux, H., e Perrier, A. (2001). Assessing clinical probability of pulmonary embolism in the emergency ward. *Archives of Internal Medicine* 161, 92–97.