UNIVERSIDADE ESTADUAL DA PARAÍBA

CENTRO DE CIÊNCIAS E TECNOLOGIA

DEPARTAMENTO DE ESTATÍSTICA

Sandro Lins Lopes de Lucena

COMPARAÇÃO DE MODELO CLÁSSICO E BAYESIANO PARA DADOS DE ÓBITOS PERINATAIS NO ISEA, CAMPINA GRANDE-PB

Campina Grande - PB

Dezembro de 2018

Sandro Lins Lopes de Lucena

COMPARAÇÃO DE MODELO CLÁSSICO E BAYESIANO PARA DADOS DE ÓBITOS PERINATAIS NO ISEA, CAMPINA GRANDE-PB

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Orientador: Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros

Campina Grande - PB

Dezembro de 2018

L935c Lucena, Sandro Lins Lopes de.

Comparação de Modelo Clássico e Bayesiano para dados de óbitos perinatais no ISEA, Campina Grande - PB [manuscrito] / Sandro Lins Lopes de Lucena. - 2018.

45 p. : il. colorido.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia , 2018.

"Orientação : Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros , Coordenação do Curso de Estatística - CCT."

1. Modelos lineares generalizados. 2. Inferência Bayesiana. 3. Regressão logística. 4. Óbito perinatal. I. Título

21. ed. CDD 519.5

Elaborada por Giulianne M. Pereira - CRB - 15/714 BC/UEPB

Sandro Lins Lopes de Lucena

COMPARAÇÃO DE MODELO CLÁSSICO E BAYESIANO PARA DADOS DE ÓBITOS PERINATAIS NO ISEA, CAMPINA GRANDE-PB

Trabalho de Conclusão de Curso apresentado ao curso de Bacharelado em Estatística do Departamento de Estatística do Centro de Ciências e Tecnologia da Universidade Estadual da Paraíba em cumprimento às exigências legais para obtenção do título de bacharel em Estatística.

Trabalho aprovado em 12 de dezembro de 2018.

BANCA EXAMINADORA

Prof. Dr. Kleber Napoleão N. de O. Barros
Universidade Estadual da Paraíba
(Orientador)

Prof. Dr. Mácio Augusto de Albuquerque
Universidade Estadual da Paraíba

Agradecimentos

Obrigado a Senhora Luiza Lucena de Andrade Silva (Mãe) e ao Senhor Severino Lopes da Silva (Pai), sem os quais não conseguiria caminhar na vida.

Obrigado a todos os Professores que por meio do seu trabalho ajudaram-me na aquisição do conhecimento estatístico apurado e de grande valor.

Obrigado ao Departamento de Estatística que sempre está a disposição para nos apoiar.

Obrigado ao Professor Dr. Mácio Augusto, sou grato pela oportunidade de desenvolver projetos científicos.

Obrigado do Instituto de Saúde Elpídio de Almeida por ter liberado o acesso aos dados.

Obrigado aos amigos de curso, e agora colegas de profissão, Alisson de Lima Brito, Rodolfo Crystian Pereira Silva e Ivson Freires Monteiro por terem trabalhado arduamente na coleta dos dados, contribuindo assim para o desenvolvimento deste trabalho.

Um agradecimento especial ao Orientador Prof. Dr. Kleber Napoleão. Obrigado, especialmente pela amizade.

Obrigado meus amigos. Cursar Bacharelado em Estatística foi uma experiência de inestimável valor. Ser estatístico não é fácil, a 95% de confiança. Nunca temos certeza de nada, mas fiquem certo de que sou grato por tudo.

Obrigado a todos!

"Talvez seja bom ter uma mente bonita, mas um dom ainda maior é descobrir um coração bonito."

(John Nash)

Resumo

Modelos lineares generalizados são úteis, dentre outras situações, quando se quer ajustar modelos a dados que não seguem normalidade e não podem ser ajustados usando apenas a regressão linear simples. Outra ferramenta poderosa de estimação são os métodos Bayesianos, baseado em probabilidades condicionais. Neste trabalho apresenta-se um ajuste de modelos de regressão logístico com parâmetros estimado pelo método da máxima verossimilhança que é atualizado usando as técnicas da inferência Bayesiana. Tais métodos foram aplicados em dados obtidos no Instituto de Saúde Elpídio de Almeida que fica localizado no município de Campina Grande - PB. As informações referem-se a pacientes gestantes atendidas nesta unidade de saúde. Objetivou-se obter o melhor modelo possível que nos forneça informação sobre a chance de óbito de uma criança em função de variáveis maternas usando o método de estimação da máxima verossimilhança e o método Bayesiano. Os ajustes e diagnósticos dos modelos foram realizados com auxílio do software `R`. Contatou-se que o modelo estimado pela máxima verossimilhança é muito próximo do modelo Bayesiano.

Palavras-chave: Regressão Logística, Inferência Bayesiana, Óbito Perinatal.

Abstract

Generalized linear models are useful, among other situations, when you want to fit models to data that do not follow normality and can not be adjusted using only simple linear regression. Another powerful tool for estimation is Bayesian methods, based on conditional probabilities. In this work we present an adjustment of logistic regression models with parameters estimated by the maximum likelihood method that is updated using Bayesian inference techniques. Esses métodos foram aplicados aos dados obtidos no Instituto de Saúde Elpídio de Almenida which is located in the city of Campina Grande - PB. The information refers to pregnant patients that attended this health unit. he objective was to obtain the best possible model provide us with information about a child's chance of dying in function of maternal variables using the maximum likelihood estimation method and the Bayesian method. The adjustments and diagnoses of the models were carried out using the software `R`. It was contacted that the model estimated by the maximum likelihood is very close to the Bayesian model.

Key-words: Logistic Regression, Bayesian Inference, Perinatal death.

Lista de ilustrações

Lista de tabelas

Sumário

1 Introdução

Nos últimos anos o Brasil vem enfrentando mudanças sociais, principalmente nas ultimas três décadas. Por meio do desenvolvimento social, o país vem obtendo índices cada vez mais positivo, que quantificam a melhora de vida da população. Na saúde pública, por exemplo, há diversos estudos que constatam a influenciam de fatores sociais na vida de menores recém nascidos. Silva et al. (2006) relatam que os determinantes sociais repercutem na situação de saúde na população de menores de um ano de vida enquanto que Sarinho (1998) destaca sua atenção para a idade materna (abaixo de 20 anos e acima de 35 anos) está associada à maior probabilidade de morte neonatal. O maior risco de mortalidade perinatal em adolescentes tem sido atribuído a maior incidência de baixo peso ao nascer, parto prematuro e complicações na gestação (SARINHO, 1998).

A saúde infantil no Brasil vem apresentando avanços importantes principalmente a partir da década de 1990 e 2000, demonstrada por indicadores que mostram significativas melhorias nos determinantes sociais das doenças em conjunto a organização dos serviços de saúde (VICTORA et al., 2011). Esses avanços são promovidos, em parte, por estudos científicos que podem detectar resultados e oferecer possíveis soluções para problemas que a saúde pública venha a enfrentar.

Dentre os avanços científicos destacamos o desenvolvimento dos Modelos Lineares Generalizados (MLG's). Estes modelos foram propostos por Nelder e Wedderbun, em 1972, e apresentam uma componente sistemática, uma componente aleatória e uma função de ligação. Trabalha-se com o modelo logístico, que é um caso especial de modelo linear generalizado quando tem-se a função de probabilidade binomial para a variável resposta e a função de ligação *logit*. O modelo logístico é utilizado quando se tem variável resposta do tipo binária, por isso o uso da função binomial.

Dois métodos foram utilizados para estimar os parâmetros do modelo, o método da máxima verossimilhança, ver (MARDIA; MARSHALL, 1984), e o método Bayesiano. Para método da máxima verossimilhança a seleção das variáveis é feita pelo método `stepwise` que baseia-se na observação de que uma variável pode ser movida de um conjunto de variáveis para outro conjunto de variáveis que dá maior redução no resíduo. Em outras palavras, o método consiste em selecionar modelos onde as variáveis são adicionas e/ou excluídas uma de cada vez para que o modelo se aproxime ao máximo dos dados. Mais detalhes consultar Jennrich e Sampson (1968) e Derksen e Keselman (1992).

A adequação do modelo pode ser verificada por meio de vários critérios. O Critério de Informação de Akaike e o desvio residual pode ser utilizado para verificar o quão bom é o modelo ajustado pelo método da máxima verossimilhança e por meio da análise residual verifica-se a adequabilidade do modelo estimado. O estudo residual pode ser feito diretamente observando o envelope simulado dos resíduos conforme descreve Moral, Hinde

e Demétrio (2017).

Os métodos bayesianos foram desenvolvidos a partir do trabalho de Thomas Bayes (1702-1761), ministro presbiteriano inglês cujos escritos matemáticos lhe valeram um lugar como membro da Royal Society of London (PICHE; PENTTINEN et al., 2010). Ainda segundo Piche, Penttinen et al. (2010), uma das desvantagens históricas dos métodos bayesianos é que as fórmulas de inferência existiam apenas para modelos relativamente simples. Isso não é mais uma limitação desde que o desenvolvimento nos anos 90 de algoritmos e softwares de computador eficazes permitem análise de modelos mesmo muito complexos.

Na abordagem Bayesiana, assume-se que os parâmetros sobre os quais se deseja realizar inferências estão associados a uma distribuição de probabilidade, sendo a distribuição dos mesmos definidas utilizando as informações adicionais que se possa ter sobre as quantidades de interesse ou de estudos anteriores da população estudada (FONSECA; MARTINS, 1996).

Dentre os softwares que auxiliam na inferência bayesiana destaca-se o `OpenBUGS` que pode ser acessado indiretamente usando o `R` através do pacote `R2OpnBUGS` (KINAS; ANDRADE, 2010). Outros pacotes implementados diretamente no `R` auxilia nos cálculos bayesianos, é o caso do pacote `MCMCpack`.

Com apoio teórico dos modelos lineares generalizados e com apoio dos programas computacionais descritos neste trabalho há grande facilidade para a implementação das técnicas bayesianas para modelos complexos. Diante desse contexto, objetivamos obter o melhor modelo possível que nos forneça informação sobre a chance de óbito de uma criança em função de variáveis maternas usando o método de estimação da máxima verossimilhança e o método Bayesiano.

2 Revisão de Literatura

2.1 SUS

O Sistema Único de Saúde (SUS) é um dos maiores sistemas públicos de saúde do mundo. Ele abrange desde o simples atendimento ambulatorial até o transplante de órgãos, garantindo acesso integral, universal e gratuito para toda a população do país. Amparado por um conceito ampliado de saúde, o SUS foi criado, em 1988 pela Constituição Federal Brasileira, para ser o sistema de saúde de todos os brasileiros (MINISTÉRIO DA SAÚDE, 2006).

O SUS funciona descentralizado e hierarquizado. Nem todo município tem condições de ofertar integralmente os serviços de saúde. Por exemplo, nem todo município pode suprir as necessidades de gestantes de alto risco que precisam de acompanhamento integral e procedimentos cirúrgicos. Para atender a essa demanda há centros regionais especializados. Dessa forma qualquer cidadão que precise de acompanhamento especializado deve ser transferido para esses centros regionais que conta com mais recursos.

Esses recursos são ofertados pela União, pelos estados e municípios. A União formula as políticas de saúde nacionais e é o principal financiador da saúde pública no país. Basicamente, metade dos gastos é realizado pelo governo federal, a outra metade fica por conta dos estados e municípios. Estes implementam as políticas de saúde, junto com organizações não governamentais e entidades privadas (MINISTÉRIO DA SAÚDE, 2006).

Com a implementação do Pacto pela Saúde, os recursos federais, destinados ao custeio de ações e serviços da saúde, passaram a ser divididos em seis blocos de financiamento (assistência farmacêutica; atenção básica; média e alta complexidade; gestão; investimentos; e vigilância em saúde). A Tabela 1 apresenta a quantia do repasse financeiro destinada para o estado da Paraíba e para a cidade de Campina Grande nos últimos cinco anos.

Tabela 1 – Repasses financeiros do SUS (em R$).

Ano	Paraíba	Campina Grande
2013	1.154.189.466,11	159.739.393,60
2014	1.295.982.227,80	178.874.698,94
2015	1.394.955.624,33	181.669.377,49
2016	1.415.131.777,67	193.347.341,17
2017	1.493.860.319,42	192.540.093,48

2.2 Instituto de Saúde Elpídio de Almeida - ISEA

O Instituto de Saúde Elpídio de Almeida (ISEA) foi fundado em 05 de agosto de 1951 como Maternidade Elpídio de Almeida durante o governo estadual do Dr. José Américo de Almeida e a gestão municipal do Dr. Elpídio de Almeida. Em 27 de abril de 1992, a então maternidade passou a se chamar Instituto de Saúde Elpídio de Almeida (ISEA) (RETALHOS HISTÓRICOS DE CAMPINA GRANDE, 2011). O Instituto de Saúde Elpídio de Almeida (ISEA) fica situado na Rua Vila Nova da Rainha, nº 147, no centro de Campina Grande, cidade do estado da Paraíba, localizada a 130 km da capital do estado, João Pessoa.

O ISEA não faz apenas atendimentos de partos, oferece mais de cinquenta serviços. Dentre esses, pode-se citar o Teste da Orelhinha, cirurgias de vasectomia, Centro de Tratamento para Obesidade, Ouvidoria SUS, Núcleo de Prevenção de Violência contra as Mulheres, referência para vítimas de violência sexual, cartório de Registro Civil, no qual as crianças nascidas na maternidade já saem com o documento emitido gratuitamente, suítes PPP's (Pré-Parto, Parto e Pós-Parto) e Casa da Mãe Dr. Flaviano Xavier Guedes, para mães que precisam acompanhas os filhos que permanecem internados na maternidade (COSTA, 2012).

Costa (2012) ainda ressalta que o ISEA oferece partos de baixo e de alto risco, pré-natal, atendimentos psicológicos, de fisioterapia e serviço social, UTI Neonatal, imunização, Banco de Leite Humano, pediatria para nascidos na maternidade, tratamento odontológico para pacientes de pré-natal, planejamento familiar, e exames de ultrassonografia, raios x e laboratoriais.

Segundo o Jornal da Paraíba (2018) a maternidade atende gestantes de 170 municípios pactuados com Campina Grande e ainda continua recebendo gestantes de municípios que não destinam recursos de obstetrícia para a rede municipal de saúde campinense, até mesmo de cidades do Rio Grande do Norte e de Pernambuco.

2.3 Óbito Perinatal

O ser humano pode vim ao mundo com vida, o que é esperado, ou apresentar alguma complicação que impeça a ocorrência desse evento. Para entender melhor as variáveis de nosso interesse, inicialmente precisa-se compreender algumas definições que posteriormente servirão de base para compreendermos o estudo proposto neste trabalho. Essas definições são dadas pelos SUS, são elas:

i) O período perinatal começa em 22 semanas completas (154 dias) de gestação (época em que o peso de nascimento é normalmente de 500 g), e termina com sete dias completos após o nascimento.

ii) Óbito fetal é a morte de um produto da concepção, antes da expulsão ou da extração completa do corpo da mãe, independentemente da duração da gravidez; indica o óbito o fato do feto, depois da separação, não respirar nem apresentar nenhum outro sinal de vida, como batimentos do coração, pulsações do cordão umbilical ou movimentos efetivos dos músculos de contração voluntária.

iii) Nascimento vivo é a expulsão ou extração completa do corpo da mãe, independentemente da duração da gravidez, de um produto de concepção que, depois da separação, respire ou apresente qualquer outro sinal de vida, tal como batimentos do coração, pulsações do cordão umbilical ou movimentos efetivos dos músculos de contração voluntária, estando ou não cortado o cordão umbilical e estando ou não desprendida a placenta. Cada produto de um nascimento que reúna essas condições se considera como uma criança viva.

Dessa forma pode-se entender por óbito perinatal a morte de um produto da concepção no período entre 22 semanas completas de gestação e sete dias incompletos após ocorrido o parto, estando este produto não apresentando sinais vitais, tal como respiração, batimentos do coração ou movimentos musculares efetivos voluntários que devem ser validados por um médico.

2.4 Modelos Lineares Generalizados

A teoria dos Modelos Lineares Generalizados vem desempenhando um papel importante na Estatística moderna devido ao grande número de métodos estatísticos que engloba. Esses modelos foram propostos por Nelder e Wedderbun (1972) e apresentam duas componentes: uma aleatória e outra sistemática (DEMETRIO, 1989). Os modelos lineares generalizados podem ser usados quando se tem uma única variável aleatória Y associada a um conjunto de variáveis explanatórias $x_1, ..., x_p$ (CORDEIRO; DEMÉTRIO, 2008).

Cordeiro e Demétrio (2008) ainda ressaltam que para uma amostra de n observações (y_i, x_i) em que $\mathbf{X}_i = (x_{i1}, ..., x_{ip})^T$ é o vetor de variáveis explicativas, o MLG envolve os três componentes: componente aleatório, o componente sistemático e a função de ligação. Os modelos lineares generalizados permitem duas extensões; primeiro a distribuição no componente aleatório pode vir de uma família exponencial diferente da Normal, e em segundo lugar a função de ligação pode se tornar qualquer função monotônica diferenciável (MCCULLAGH; NELDER, 1989).

O componente aleatório representado por um conjunto de variáveis aleatórias independentes $Y_1, ..., Y_n$ provenientes de uma mesma distribuição que faz parte da família exponencial de distribuições com média $\mu_1, ..., \mu_n$ ou seja, $E(Y_i) = \mu_i, i = 1, ..., n$, sendo

$\phi > 0$ um parâmetro de dispersão e o parâmetros θ_i denominado parâmetro canônico. Então, a função densidade de probabilidade de Y_i é dada por

$$f(y_i, \theta_i, \phi) = \exp\{\phi^{-1}[y_i\theta_i - b(\theta_i)] + c(y_i, \phi)\}, \tag{2.1}$$

sendo $b(.)$ e $c(.)$ funções conhecidas(CORDEIRO; DEMÉTRIO, 2008).

Segundo Paula (2004), sob as condições usuais de regularidade que

$$E\left\{\frac{\partial \log f(Y_i; \theta_i, \phi)}{\partial \theta_i}\right\} = 0 \tag{2.2}$$

e

$$E\left\{\frac{\partial^2 \log f(Y_i; \theta_i, \phi)}{\partial \theta_i^2}\right\} = -E\left[\left\{\frac{\partial \log f(Y_i; \theta_i, \phi)}{\partial \theta_i}\right\}^2\right], \tag{2.3}$$

$\forall_i$, de onde segue que $E(Y_i) = \mu_i = b'(\theta_i)$ e $Var(Y_i) = \phi^{-1}V(\mu_i)$, em que $V_i = V(\mu_i) = d\mu_i/d\theta_i$ é a função de variância e $\phi^{-1} > 0$ é o parâmetro de dispersão.

Paula (2004) ainda comenta que os modelos lineares generalizados definidos por (2.1) precisam ainda ter um componente sistemático

$$g(\mu_i) = \eta_i = \mathbf{X}^T\boldsymbol{\beta} \tag{2.4}$$

sendo $\eta_i = \mathbf{X}^T\boldsymbol{\beta}$ é o preditor linear, $\boldsymbol{\beta} = (\beta_1, ..., \beta_p)^T$, $p < n$, é um vetor de parâmetros desconhecidos a serem estimados, $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$, representa os valores de variáveis explicativas e $g(.)$ é uma função monótona e diferenciável, denominada função de ligação (PAULA, 2004).

Para Cordeiro e Demétrio (2008) a variável resposta, componente aleatório do modelo, tem distribuição pertencente à família de distribuição que engloba as distribuições normal, gama e normal inversa para dados contínuos; binomial para proporções; Poisson, binomial negativa para contagens. As variáveis explanatórias ou independentes entram na forma de uma estrutura linear, constitui o componente sistemático do modelo e a ligação entre os componentes aleatório e sistemático é feita através de uma função de ligação, por exemplo, logarítmica para os modelos log-lineares, chama de função de ligação.

Outros modelos foram surgindo e os desenvolvimentos que levaram a essa visão geral da modelagem estatística, remontam a quase dois séculos. Assim, um MLG é definido por uma distribuição de probabilidade, membro da família de distribuições, para a variável resposta, um conjunto de variáveis independentes descrevendo a estrutura linear do modelo e uma função de ligação entre a média da variável resposta e a estrutura linear. Entre os métodos estatísticos para análise de dados univariados que são casos especiais de MLG, cita-se o modelo logístico (CORDEIRO; DEMÉTRIO, 2008).

2.4.1 Modelo Binomial e Família Exponencial

O modelo binomial é usado, principalmente, no estudo de dados na forma de proporções, como nos casos da análise probito (Finney, 1952), logística (ou "logit") (Ashton,

1972) e complemento log-log (Fisher, 1922), e na análise de dados binários, como na regressão logística linear (Cox, 1970) (CORDEIRO; DEMÉTRIO, 2008). A distribuição binomial pertence à família exponencial, pois pode ser escrita na forma (2.5), como pode-se ver adiante.

Segundo Bolfarine e Sandoval (2001) dizemos que a distribuição da variável aleatória (ou de um vetor aleatório) Y pertence à família exponencial de dimensão k se a função de densidade (ou de probabilidade) de Y é dada por

$$f(y|\theta) = e^{\sum_{j=1}^{k} c_j T_j(y) + d(\theta) + S(y)}, y \in A, \tag{2.5}$$

onde c_j, T_j, d e S são funções reais, $j = 1, ..., k$, e como no caso unidimensional, $d(\theta)$ está associado à constante de normalização de (2.5) e A não depende de θ. Note que (2.5) é equivalente a (2.1) quando $d(\theta) = b(\theta)$ e $T_j(y) = y_i$.

2.4.1.1 Modelo Binomial

Para Bolfarine e Sandoval (2001) uma variável aleatória Y tem distribuição binomial, com parâmetros n e θ, que denotamos por $Y \sim Binomial(n, \theta)$, se sua função de probabilidade é dada por

$$f(y \mid \theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}, y = 0, 1, ..., n,$$

em que $0 < \theta < 1$. Nesse caso o suporte é discreto e é dado por $A(y) = \{y, y = 0, 1, ..., n\}$ Temos também que $E[Y] = n\theta$ e $Var[Y] = n\theta(1-\theta)$.

O modelo Bernoulli(caso especial do modelos Binomial quando $n = 1$) é comumente empregado em situações em que associamos a cada observação da amostra dois tipos de resposta (como por exemplo, sim e não, ou sucesso e fracasso) aos quais associamos os valores 0 e 1 (BOLFARINE; SANDOVAL, 2001).

2.4.1.2 Modelo Binomial na Família Exponencial

Para escrever a função binomial na família exponencial procede-se da seguinte forma:

$$f(y \mid \theta) = \left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y}$$

$$f(y \mid \theta) = \exp\left\{ \log\left[\left(\begin{array}{c} n \\ y \end{array} \right) \theta^y (1-\theta)^{n-y} \right] \right\}$$

$$f(y|\theta) = \exp\left\{\log\left(\begin{array}{c} n \\ y \end{array}\right) + x\log\left(\frac{\theta}{1-\theta}\right) + n\log(1-\theta)\right\}$$

De onde pode-se tirar:

$$\mu = \log\left(\frac{\theta}{1-\theta}\right) \rightarrow \theta = \frac{e^{\mu}}{1+e^{\mu}}$$

$$\alpha(\phi) = \phi \rightarrow \phi = 1$$

$$g(\theta) = -n\log(1-\theta) \rightarrow g(\mu) = n\log(1+e^{\mu})$$

$$b(\phi, y) = log\left(\begin{array}{c} n \\ y \end{array}\right)$$

$$E(y) = g'(\mu) = n\left(\frac{e^{\mu}}{1+e^{\mu}}\right) = n\theta$$

$$\begin{aligned} Var(y) &= b''(\mu)\alpha(\phi) \\ &= n\left[\frac{e^{\mu}(1+e^{\mu}) - e^{\mu}e^{\mu}}{(1+e^{\mu})^2}\right] \\ &= n\theta(1-\theta) \end{aligned}$$

$$V(\theta) = \frac{e^{\mu}}{e^{\mu}+1} - \frac{e^{2\mu}}{(e^{2\mu}+1)^2} = \theta - \frac{2\theta}{\theta+1}$$

2.5 Modelo Logístico

Vamos considerar nosso vetor Y_i, variável binária, com a seguinte definição

$$Y_i = \begin{cases} 1, \text{ se o i-ésimo elemento amostral possuir determinada característica} \\ 0, \text{ se caso contrário} \end{cases}$$

Com essa característica cada Y_i segue uma distribuição Bernoulli com valor esperado $\pi(x_i)$, sendo $\pi(x_i) = P(Y_i = 1|x_i)$, que é a probabilidade de existir a característica de interesse associado ao valor dado de x_i. Então Y_i tem distribuição de probabilidade dada por

$$Y_i = \begin{cases} P(Y_i = 1|x_i) &= \pi(x_i) \\ P(Y_i = 0|x_i) &= 1 - \pi(x_i), i = 1, 2, ..., n. \end{cases} \tag{2.6}$$

Considera-se o modelo

$$\begin{aligned}\pi(x_i) &= P(Y_i - 1|x_i) \\ &= \frac{\exp\{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}\}}{1 + \exp\{\beta_0 + \beta_1 x_{i1} + ... + \beta_p x_{ip}\}} \\ &= \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1 - \exp(\mathbf{x}_i'\boldsymbol{\beta})}\end{aligned}$$

sendo $\pi(x_i)$ é a probabilidade de sucesso do evento para a i-ésimo elemento amostral Y_i é o valor da variável resposta binária do i-ésimo indivíduo e $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_{p-1})$ o vetor de parâmetros.

Cordeiro e Paula (1989), consideram o modelo logístico linear simples derivado da função matemática

$$f(x) = (1 + e^{-x})^{-1}, -\infty < x < \infty \tag{2.7}$$

que varia monoticamente de 0 a 1 à medida que x cresce, sendo simétrica em torno de $x = f^{-1}(1/2)$. O termo linear refere-se à propriedade da transformação logit em $f(x)$

$$\text{logit} f(x) = \log\left[\frac{f(x)}{1 - f(x)}\right] = x \tag{2.8}$$

que é linear em x. Então tem-se o modelo de regressão logístico linear para prever a probabilidade $\pi(x) = P\{y = 1|x\}$. Esse modelo é dado por

$$\text{logit}\pi(x) = \eta \tag{2.9}$$

ou

$$\begin{aligned}\pi(x) &= \exp(\eta)/\{1 + \exp(\eta)\} \\ &= \{1 + \exp(-\eta)\}^{-1},\end{aligned}$$

em que $\eta = \alpha + \beta x$ é chamado de preditor linear e α e β são parâmetros a serem estimados (CORDEIRO; PAULA, 1989).

Na maioria dos problemas, no entanto, a variável resposta y é expressa em função de um grande conjunto de informações, faz-se necessário então um ajuste de múltiplas variáveis regressoras.

Segundo Kinas e Andrade (2010) quando a variável y é discreta e segue uma distribuição Binomial, então o MLG é denominado regressão logística. As três equações que definem este modelo são definidas a seguir:

$$y_i \sim Bin(n_i, \theta_i)$$

$$\eta_i = g(\theta_i) = \log\left(\frac{\theta_i}{1 - \theta_i}\right)$$

$$\eta_i = \beta_0 + \beta_1 x_i$$

A função de ligação definida acima é denominada de função logit e transforma o parâmetro θ, restrito ao intervalo $[0, 1]$, para η que está definido em $\mathbb{R}$ (KINAS; ANDRADE, 2010).

2.6 Inferência Bayesiana

A inferência bayesiana é baseada em probabilidades subjetivas ou credibilidades a posteriori associadas com diferentes valores do parâmetro θ e condicionadas pelo particular valor de x observado (PAULINO; TURKMAN; MURTEIRA, 2003). O valor de x é dado, fixado, enquanto que há uma considerada variação no valor de θ. Ainda segundo Paulino, Turkman e Murteira (2003), para os bayesianos há apenas um estimador que é precisamente a distribuição a posteriori $p(\theta|x)$. Na estimação por intervalos, aos intervalos de confiança os Bayesianos contrapõem os intervalos de credibilidade. Observado y e determinada pelo mecanismo do Teorema de Bayes a distribuição a posteriori, um intervalo de credibilidade para o parâmetro θ é formado por um par de valores de Θ, sejam $[\theta^*(y), \overline{\theta}(y)]$, ou mais simplesmente, $(\theta^*, \overline{\theta})$, tais que,

$$P(\theta^* < \theta < \overline{\theta}|y) = \int_{\theta^*}^{\overline{\theta}} p(\theta|y)d\theta = 1 - \alpha,$$

onde $1 - \alpha$ (em geral, 0,90; 0,95 ou 0,99) é o nível de credibilidade desejado (PAULINO; TURKMAN; MURTEIRA, 2003).

2.6.1 A Priori e a Posteriori

A informação a priori quantifica a credibilidade que o pesquisador e/ou o estatístico tem diante das informações a respeito de θ antes dos acontecimentos. Para Paulino, Turkman e Murteira (2003), a informação a priori que se pretende incorporar na análise é a informação apriorística possuída por alguém, que se identifica como especialista (perito, expert) do problema concreto - seja ele o investigador, o estatístico ou outrem - e contém elementos subjetivos que, em geral, até são dominantes. A informação a priori pode traduzir-se formalmente por uma distribuição de probabilidade, geralmente subjetiva, para θ, seja $p(\theta)$, designada distribuição a priori. Quando não se tem informações prévias minimamente consistentes, utiliza-se uma priori nã-informativa para representar a falta de informação que se tem de θ (PAULINO; TURKMAN; MURTEIRA, 2003).

A posteriori de θ é a distribuição do parâmetro depois de já ter sabido o resultado dos valores da amostra. Assim tendo em conta as informações contida nos dados y a atitude inicial do investigador, caracterizada por $p(\theta)$, é modificada passando a nova atitude a traduzir-se por $p(\theta|y)$. Usando o Teorema de Bayes para atualizar a informação a priori,

temos:

$$p(\theta|y) = \frac{p(y|\theta)p(\theta)}{\int_\Theta p(y|\theta)p(\theta)d\theta},$$

em que $p(\theta|y)$ é a distribuição a posteriori de θ após o conhecimento que já aconteceu $Y = y$ (PAULINO; TURKMAN; MURTEIRA, 2003).

2.7 OpenBUGS

O `OpenBUGS` é um programa para análise bayesiana de modelos estatísticos complexos usando as técnicas de Monte Carlo via Cadeia de Markov (MCMC). O `OpenBUGS` permite que os modelos sejam descritos usando linguagem BUGS, ou Doodles (representação gráfica de modelos) que podem, se desejado, ser uma descrição baseada em texto. A linguagem BUGS é mais flexível do que a representação gráfica Doodles.

Kinas e Andrade (2010) informam que o analista dos dados pode escrevem seu próprio código para implementar o métodos MCMC, entretanto houve grande popularização e tais métodos já estão implementados em programas computacionais disponibilizados na internet. Por exemplo, o `OpenBUGS` pode fazer ligação com o `R`, ou seja, podemos simular algoritmos MCMC indiretamente no `R` usando o `OpenBUGS` através do pacote `R2WinBUGS` (KINAS; ANDRADE, 2010).

2.8 O Teste Qui-quadrado para Independência

O teste Qui-quadrado de independência, também conhecido como teste do Qui-quadrado de Pearson, é uma das estatísticas mais úteis para testar hipóteses quando as variáveis são nominais, como frequentemente acontece na pesquisa clínica (MCHUGH, 2013). É conveniente usar o teste Qui-quadrado quando os níveis de medição das variáveis é nominal ou ordinal, quando os dados nas tabelas são frequências e os grupos de estudo são independentes e cada indivíduo da amostra pertence a um único grupo.

As hipóteses mais comuns formalizam a afirmação de que não há conexão entre as categorizações. Esta é a hipótese de independência (MOORE, 1976). Para Firmino et al. (2015), as hipóteses em testes são as seguintes:

$$\begin{cases} H_0 : \text{As variáveis são independentes.} \\ H_1 : \text{As variáveis não são independentes.} \end{cases}$$

Vieira (2015) lembra que o nível de significância do teste, α, é a probabilidade de cometer erro tipo I, isto é, rejeitar H_0 quando H_0 é verdadeira. A autora também afirma que o valor p informa quão provável seria obter uma amostra tão ou mais extrema a que foi obtida, quando a hipótese de nulidade é verdadeira. Rejeita-se H_0 quando o valor p é

menor que o nível de significância adotado. Para maiores detalhes consultar (SHESKIN, 2003).

2.9 O Teste de Anderson-Darling

Existem vários procedimentos para averiguar a hipótese de normalidade para um conjunto de dados. Dentre esses procedimentos destaca-se o teste de Anderson-Darling (1954). Segundo Nelson (1998), o teste de Anderson-Darling tem a vantagem de que os valores críticos para cada tamanho de amostra não são necessários além de ser um pouco mais sensível a desvios nas caudas da distribuição. Nelson (1998) define a estatística de teste da forma:

$$A = \frac{-\{\sum_{i=1}^{n}(2i-1)[\ln(p_i) + \ln(1-p_{n+1-i})]\}}{n} - n,$$

em que p_i é a probabilidade acumulada da distribuição normal padrão. As hipóteses são:

$$\begin{cases} H_0 : \text{Os dados seguem distribuição normal;} \\ H_1 : \text{Os dados não seguem distribuição normal.} \end{cases}$$

Rejeita-se a hipótese H_0 quando o valor p é menor que o nível de significância adotado.

2.10 O Teste de Spearman

O coeficiente de correlação de Spearman é uma de várias medidas que podem representar a correlação entre variáveis. Para Sheskin (2003), o coeficiente de correlação de Spearman determina o grau em que existe uma relação monotônica entre duas variáveis. Uma relação monotônica pode ser descrita como crescente monotônica (que está associada a uma correlação positiva) ou decrescente monotônica (que está associada a uma correlação negativa). A hipóteses testadas, para um teste bilateral, são:

$$\begin{cases} H_0 : \text{As variáveis não são correlacionadas. } (r_s = 0) \\ H_1 : \text{As variáveis são correlacionadas. } (r_s \neq 0) \end{cases}$$

Calcula-se o coeficiente de correlação de Spearman por meio da equação

$$r_s = 1 - \frac{6\sum d^2}{n(n^2-1)}$$

Sheskin (2003) salienta que se a hipótese alternativa direcional $H_1 : r_s \neq 0$ for empregada, a hipótese nula pode ser rejeitado se o valor absoluto obtido de r_s for igual ou superior ao valor crítico tabelado calculado no nível de significância pré-especificado. A Equação (2.10) é uma alternativa quando não se tem acesso a tabela de valores críticos

e segundo Sheskin (2003) fornece uma boa aproximação da distribuição de amostragem quanto o tamanho da amostra for maior que 10.

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}} \tag{2.10}$$

2.11 Seleção de Modelos

Um modelo de regressão composto por k variáveis explicativas, sendo k um número alto, pode ser representativo tanto quanto um modelo ajustado com um número reduzido das k variáveis regressoras. Ajustando modelo completo, com todas as variáveis explicativas, torna-se difícil tirar qualquer conclusão a respeito das influências das variáveis explicativas sobre a variável resposta, além de poder conter problemas no ajuste do modelo. Para facilitar a explicação da variável resposta em função das variáveis regressoras deve-se optar por um modelo reduzido que tenha um alto poder explicativo (princípio da parcimônia).

Esse poder explicativo depende, dentre outros fatores, de quais variáveis estão compondo o modelo. Por isso, deve-se escolher as variáveis regressoras cuidadosamente. Existem vários critérios que podem ser seguidos para selecionar um subconjunto de variáveis que deve estruturar o modelo, a qualidade do ajuste também pode ser mensurada. Aqui discute-se a respeito do Critério de Informação de Akaike (AIC), o Critério de Informação do Desvio (DIC) e do método de seleção de variáveis `Stepwise`.

2.11.1 Critério de Informação de Akaike (AIC)

Proposto por Akaike (1974), esse método objetiva selecionar um modelo que esteja bem ajustado e tenha um número reduzido de parâmetros. Para os modelos lineares generalizados o método de Akaike pode ser expresso, segundo Paula, numa forma mais simples em função do desvio do modelo. Neste caso o critério consiste em encontra-se o modelo para o qual a equação abaixo seja minimizada (PAULA, 2004).

$$AIC = D^*(y;\mu) + 2p, \tag{2.11}$$

em que $D^*(y;\mu)$ denota o desvio do modelo e p o número de parâmetros (JAMES et al., 2013).

Segundo Paulino, Turkman e Murteira (2003), o AIC é dado por

$$\Delta AIC = -2\log\left[\frac{sup_{M_1} f(x|\theta_1, M_1)}{sup_{M_2} f(x|\theta_2, M_2)}\right] - 2(p_2 - p_1),$$

em que $p_i, i = 1, 2$ representa o número de parâmetros de cada modelo. Este critério é baseado em considerações frenquentistas de eficiência assintótica, mas só aproxima $-2\log BF$ se a informação contida na distribuição a priori aumentar na mesma razão que

a informação contida na verossimilhança, em que BF é o Fator de Bayes. Ver (LUNN et al., 2012). Esta situação não é realista do ponto de vista da metodologia bayesiana e, como tal, não é razoável a sua utilização neste contexto (PAULINO; TURKMAN; MURTEIRA, 2003).

2.11.2 Critério de Informações do Desvio (DIC)

Existem diversos critérios baseados em aproximações ao fator de Bayes entre dois modelos M_1 e M_2, nomeadamente o critério AIC (Akaike Information Criterion), citado anteriormente, o critério BIC (Bayesian Information Criterion) e o critério DIC (Deviance Information Criterion) (PAULINO; TURKMAN; MURTEIRA, 2003). Dificuldades com critérios anteriores levou Spiegelhalter a sugerir uma generalização do critério AIC baseada na distribuição a posteriori de $D_i(\theta) = -2\log\frac{p(y|\theta,M_i)}{p(y)}$, onde $p(y)$ é uma função apenas dos dados que não tem impacto na escolha do modelo. Propõem como medida da adequabilidade do modelo o valor esperado a posteriori de $D_i(\theta)$ e como penalização $p_{D,i}$ associado à complexidade do modelo a diferença entre este valor esperado e o valor de $D_i(\theta_i)$ calculado no valor esperado a posteriori de θ_i. Assim

$$p_{D,i} = E_{\theta_i|x,M_i}[D_i(\theta_i)] - D_i(E_{(\theta_i|x,M_i)}[\theta_i])$$

e

$$DIC_i = E_{\theta_i|x,M_i}[D_i(\theta_i)] + p_{D,i} = 2E_{\theta_i|x,M_i}[D_i(\theta_i)] - D_i(E_{(\theta_i|x,M_i)}[\theta_i]).$$

2.11.3 Métodos de Seleção de Variáveis

2.11.3.1 Forward

Esse método considera inicialmente o modelo com apenas uma variável regressora. Para Demetrio (1989) a variável com maior coeficiente de correlação simples com a variável resposta, e na sequência adicionar outras variáveis regressoras, uma a uma, com o objetivo de encontrar um modelo com poucas variáveis. Para Paula (2004) inicia-se o método pelo modelo $\mu = \alpha$ e ajusta-se então, para cada variável explicativa o modelo, $\mu = \alpha + \beta_j x_j, (j = 1, ..., q)$.

2.11.3.2 Backward

Esse método, diferente do método anterior, considera inicialmente o modelo composto por todas as variáveis regressoras. Em seguida elimina variáveis, uma de cada vez, de forma que o resultado final seja um modelo parcimonioso. Paula (2004) comenta que se deve iniciar pelo modelo $\mu = \alpha + \beta_1 x_1 + ... + \beta_q x_q$.

2.11.3.3 Stepwise

Um dos métodos mais comuns em modelagem de regressões é o método `stepwise`. Esse método é uma mistura dos dois métodos anteriores, inclui-se e/ou exclui-se uma variável regressora dependendo do poder que a mesma tem na explicação da variável resposta. Para Paula (2004) é uma mistura dos dois procedimentos anteriores e iniciamos o processo com o modelo $\mu = \alpha$. Após duas variáveis terem sido incluídas no modelo, verificamos se a primeira não sai do modelo. O processo continua até que nenhuma variável seja incluída ou seja retirada do modelo.

Uma possível condição para a exclusão ou inclusão das variáveis regressoras no modelo é o nível de significância que esta possui. Paula (2004) ainda ressalta que esse grau de importância pode ser avaliado por exemplo, pelo nível de significância do teste da razão de verossimilhanças entre os modelos que incluem ou excluem as covariáveis em questão e que no caso de regressão logística o teste da razão de verossimilhança, pelo fato de ser obtido pela diferença de duas funções desvio, aparece como o mais indicado.

O teste da razão de verossimilhanças é usualmente realizado utilizando a estatística da razão de verossimilhanças em modelos encaixados (COX et al., 1977). Para Giolo e Colosimo (2006), o teste é realizado a partir dos seguintes dois ajustes: (i) modelo generalizado e obtenção do valor do logaritmo de sua função de verossimilhança; (ii) modelo de interesse e obtenção do valor do logaritmo de sua função de verossimilhança.

A partir destes valores é possível calcular a estatística da razão de verossimilhanças,

$$T = -2 \log \left[\frac{L_r}{L_s} \right] = 2[\log L_s - \log L_r],$$

em que L_s é o valor do logaritmo da função de verossimilhança do modelo saturado e L_r é o valor do logaritmo da função de verossimilhança do modelo restrito. T, sob H_0, tem aproximadamente uma distribuição qui-quadrado com graus de liberdade igual a diferença do número de parâmetros dos modelos sendo comparados (GIOLO; COLOSIMO, 2006).

2.12 Qualidade do Modelo e Resíduos

Para Nelder e Baker (1972) a função de desvio é mais diretamente útil, não como uma medida absoluta da adequação do ajuste, mas para comparar dois modelos e é definido como sendo o dobro da diferença entre a máxima verossimilhança alcançável e a obtida no modelo ajustado. A função desvio é escrita da forma

$$\begin{aligned} D(y; \hat{\theta}) &= 2l(\mu; y) - 2l(\hat{\mu}; y) \\ &= \sum_i \left\{ y_i \log(y_i/\hat{\mu}_i) + (m_i - y_i) \log \left(\frac{m_i - y_i}{m_i - \hat{\mu}_i} \right) \right\} \end{aligned}$$

A qualidade do ajuste de um MLG é avaliada através da função desvio que é a distância entre o logaritmo da função de verossimilhança do modelo saturado (com n parâmetros) e do modelo sob investigação (com p parâmetros) avaliado na estimativa de máxima verossimilhança β (PAULA, 2004). Menores valores para a função desvio indica que obtemos um melhor ajuste. Vale lembrar que a escolha do modelo definitivo depende de outros critérios que podem estar implícitos nos objetivos da pesquisa.

Paula (2004) cita que para o caso Binomial em que assumimos $Y \sim B(n_i, \mu_i), i = 1, 2, ..., k$, $\tilde{\theta}_i = \log\{y_i/(n_i - y_i)\}$ e $\hat{\theta}_i = \log\{\hat{\mu}_i/(1 - \hat{\mu}_i)\}$ para $0 < y_i < n_i$, o desvio assume a seguinte forma:

$$D(y; \hat{\mu}) = 2\sum_{i=1}^{k}[y_i \log(y_i/n_i\hat{\mu}_i) + (n_i - y_i)\log\{(1 - y_i/n_i)/(1 - \hat{\mu}_i)\}]$$

Graficamente também pode-se avaliar um modelo linear generalizado. O gráfico normal de probabilidade para t_{D_i}, definido a seguir, com envelope é uma dos gráficos mais utilizadas. Moral, Hinde e Demétrio (2017), descreveram o pacote `hnp` do software estatístico `R`, que pode ser usado para gerar o gráfico meio normal com envelope simulado para resíduos de diferentes tipos de modelos, incluindo para dados de proporção com distribuição binomial e quase-binomial. Ainda segundo Moral, Hinde e Demétrio (2017), parcelas meio-normais com envelopes simulados são úteis para avaliar a adequação do ajuste, especialmente ao analisar dados super dispersos e para um modelo bem ajustado, o envelope é tal que os disgnósticos de modelo são prováveis que caia dentro.

Os resíduos mais utilizados em modelos lineares generalizados são definidos a partir dos componentes da função desvio. A versão padronizada é a seguinte

$$t_{D_i} = \frac{d^*(y_i; \mu_i)}{\sqrt{(1 - \hat{h}_{ii})}} = \frac{\phi^{1/2} d(y_i; \hat{\mu}_i)}{\sqrt{(1 - \hat{h}_{ii})}},$$

em que $d(y_i, \hat{\mu}_i) = \pm\sqrt{2}\{y_i(\tilde{\theta}_i - \hat{\theta}_i) + (b(\hat{\theta}_i) - b(\tilde{\theta}_i))\}^{1/2}$. Em particular, segundo Paula (2004), para os modelos binomiais, esse resíduo é expresso, para $0 < y_i < n_i$, na forma

$$t_{D_i} = \pm\sqrt{\frac{2}{1 - \hat{h}_{ii}}}\left\{y_i \log\left(\frac{y_i}{n_i\hat{\pi}_i}\right) + (n_i - y_i)\log\left(\frac{n_i - y_i}{n_i - n_i\hat{\pi}_i}\right)\right\}^{1/2}$$

2.13 Matriz de Confusão

Para Congalton (1991), uma matriz de erro é uma matriz quadrada de números dispostos em linhas e colunas que expressam o número de unidades de amostra [...] atribuídos de uma categoria específica em relação à categoria real. Flach (2003) comenta que tabela de contingência (às vezes chamada de matriz de confusão) é uma maneira conveniente de tabular estatísticas para avaliar o qualidade de um modelo.

A Tabela 2 apresenta o esquema de uma matriz de erro para duas categorias. A é número da amostra predito na categoria 1 e que realmente pertence a categoria 1. B foram erroneamente preditos na categoria 2 mas pertencem no contexto real a categoria 1. C é o valor predito, também erroneamente, na categoria 1 mas que pertencem a categoria 2, e D é o valor que predito na categoria 2 e pertencem a categoria 2.

Tabela 2 – Matriz de confusão.

		Valores Preditos	
		Categoria 1	Categoria 2
Valores Reais	Categoria 1	a	b
	Categoria 2	c	d

Azevedo (2016) define a precisão (P) como o cálculo da proporção de itens que foram classificados como positivos, que são realmente positivos. A precisão é dada por $P = \frac{d}{b+d}$. Azevedo (2016) também define a acurácia (A_c), como quantidade de objetos corretamente classificados, dividido pelo total de itens presentes na base de testes. A acurácia é dada por $A_c = \frac{a+d}{a+b+c+d}$. A taxa positiva verdadeira (TP), é a proporção de casos positivos que foram corretamente identificados, dada por $TP = \frac{d}{c+d}$. A taxa de falsos positivos (FP) é a proporção dos casos negativos que foram incorretamente classificados como positivos, FP é dada pela equação $\frac{b}{a+b}$. A taxa de negativa verdadeira (TN) é definida como a proporção de casos negativos que foram classificados corretamente. Seu valor é dado por $\frac{a}{a+b}$ e falsos negativos (FN) é a proporção de casos positivos que foram incorretamente classificados como negativos, FN é dado por $\frac{c}{c+d}$.

3 Materiais e Métodos

Os dados foram coletados no Instituto de Saúde Elpídio de Almeida e referem-se a características de gestantes atendidas na maternidade no ano de 2013. Inicialmente tinha-se informações de 5604 pacientes no banco de dados, e para cada paciente coletou-se vinte variáveis, dentre características da mãe e da criança. Devido a ausência de informações no preenchimento dos formulários, dados ausentes, e após minuciosa mineração do banco de dados, a amostra ficou restrita e composta por informações de 2807 mulheres que foram atendidas nesta instituição no ano citado.

Dos prontuários, com as características das gestantes, trabalhamos com a idade de cada paciente, o grau de instrução, a situação conjugal, a raça declarada e o número de gestações e o mês do atendimento. Depois da mineração dos dados, fez-se uma análise descritiva buscando verificar as frequências de cada variável e entender o perfil das mães gestantes que passaram pela maternidade.

Trata-se na modelagem, com a variável resposta binária em que 0 indica que a mãe não sofreu com óbitos de seu filho em nenhuma de suas gestações e 1 indica que a mãe já perdeu seu filho no período perinatal, ou seja, a criança veio a óbito. Na sequência considera-se o modelo de regressão linear generalizado saturado, com as variáveis idade, o número de gestação, o grau de instrução, a situação conjugal e a raça declarada, com distribuição binomial e função de ligação *logit*, ajustando parâmetros pelo método da máxima verossimilhança, para verificar quais das variáveis podem ter relação com a chance de óbito dos produtos gestacionais.

Aplicando-se os métodos de seleção de modelos descritos na seção anterior, ajusta-se um modelo logístico com tendência linear e outro modelo logístico com tendência quadrática na variável idade. Em seguida verifica-se, por meio dos critérios de bondade do ajuste e da figura de envelope simulado, a adequação de cada um dos modelos ajustados.

Ao encontrar o melhor modelo aplica-se a inferência Bayesiana nos parâmetros do mesmo, por meio de simulação de Monte Carlos via Cadeias de Markov, tendo como distribuição a priori para os parâmetros distribuições normais com médias iguais a suas respectivas estimativas dada pela estimação máxima verossimilhança e credibilidade 0,01, obtendo-se as estimativas dos parâmetros atualizadas e a distribuição a posteriori com a aplicação direta do Teorema de Bayes.

4 Resultado e Discussão

Nesta seção vamos fazer uma breve análise descritiva das variáveis, mostrar os modelos resultantes com as estimativas dos seus respectivos parâmetros, apresentar suas respectivas matrizes de confusão, os cálculos das probabilidades para algumas situações simuladas e demais resultados desse estudo.

4.1 Análise Descritiva

Dos 2807 prontuários que compõem este estudo, 30,96% ($IC_{95\%} = [29,25\%; 32,67\%]$) das mães sofreram com o óbito de pelo menos um produto de sua(s) gestação(ões), enquanto que 69,04% ($IC_{95\%} = [67,33\%; 70,75\%]$) não registraram óbitos conforme observado na Tabela 3. Avaliando por variável, no tocante à situação conjugal das mães, há prevalência das solteiras, com percentual de 55,51% ($IC_{95\%} = [53,67\%; 57,35\%]$), seguido de mães casadas com 26,00% ($IC_{95\%} = [24,38\%; 27,62\%]$) na (Tabela 3). Das 1558 mães solteiras, 29,97% ($IC_{95\%} = [27,69\%; 32,24\%]$) tiveram pelo menos um insucesso em alguma de sua(s) gravidez. 50% dos produtos gerados pelas mulheres viúvas vieram a óbito (Tabela 3).

A frequência, segundo o grau de instrução, foi maior para as mulheres que possuem apenas o Ensino Fundamental II com 1167 mães. Em segundo lugar registrou-se mãe com Ensino Médio, com frequência de 590 mães, e em terceiro lugar 430 mães com Ensino Fundamental I (Tabela 3). Constatou-se ainda 1,61% ($IC_{95\%} = [1,14\%; 2,07\%]$) das mães não possuem escolaridade. Dessas 44,44% ($IC_{95\%} = [29,92\%; 58,96\%]$) perderam pelo menos um de seu(s) filho(s). No caso de mães com ensino superior, completo ou incompleto, apenas 22,77% ($IC_{95\%} = [14,59\%; 30,95\%]$) tiveram registros de óbitos para pelo menos um de seus filhos.

No quesito racial mães pardas são ampla maioria. Registrou-se que 93,30% ($IC_{95\%} = [92,37\%; 94,22\%]$) das mulheres são pardas. Mães declaradas brancas foram apenas 5,38% ($IC_{95\%} = [4,54\%; 6,21\%]$) , pretas 1,28%($IC_{95\%} = [1,07\%; 1,49\%]$) e indígena apenas uma (Tabela 3). Pode-se concluir que 31,16% ($IC_{95\%} = [29,38\%; 32,93\%]$) das mães pardas não obtiveram êxito em pelo menos uma gestação. Esse número é de 38,89% das mães declaradas pretas.

Quanto a idade verificou-se que, em média, as mães tem 27,63 anos ($IC_{95\%} = [27,43; 27,82]$) com números variando dos 13 anos e 50 anos de idade. Os registros foram alocados em quatro faixas etária conforme observado na Tabela 3. A classe de mulheres com idade entre 23 e 27 anos obteve maior frequência, foram 814 pacientes, ou seja 29% ($IC_{95\%} = [27,32\%; 30,68\%]$) do total, seguida de perto pelo grupo de mulheres que possuem

entre 32 e 50 anos de idade, que registrou 769 casos, 27,40% ($IC_{95\%} = [25,75\%; 29,04\%]$) do total (Tabela 3). Registrou-se 103 casos de mães menores de idade, ou seja entre 13 e 17 anos, das quais 46 não tiveram êxito em pelo menos uma gestação.

Tabela 3 – Número de mães atendidas no ISEA segundo a situação conjugal, o grau de instrução, raça e grupo de idade em 2013.

	Óbito			
	Não	**Sim**	**Total**	**Teste Qui-Quadrado**
	$n(\%)$	$n(\%)$	$N(\%)$	Valor p
Situação Conjugal				0,367
Solteira	1091 (38,87)	467 (16,64)	1558 (55,51)	
Casada	498 (17,74)	232 (8,26)	730 (26,00)	
Viúva	5 (0,18)	5 (0,18)	10 (0,36)	
Divorciada	7 (0,25)	6 (0,21)	13 (0,46)	
União Estável	337 (12,01)	159 (5,66)	496 (17,67)	
Grau de Instrução				0,018
Sem Escolaridade	25 (0,89)	20 (0,71)	45 (1,61)	
Fundamental I	430 (15,32)	225 (8,02)	655 (23,33)	
Fundamental II	815 (29,03)	352 (12,54)	1167 (41,57)	
Ensino Médio	590 (21,02)	249 (8,87)	839 (29,89)	
Superior Incompleto	31 (1,10)	5 (0,18)	36 (1,28)	
Superior Completo	47 (1,67)	18 (0,64)	65 (2,32)	
Raça				0,324
Branca	112 (3,99)	39 (1,39)	151 (5,38)	
Preta	22 (0,78)	14 (0,50)	36 (1,28)	
Parda	1803 (64,23)	816 (29,07)	2619 (93,30)	
Indígena	1 (0,04)	0 (0,00)	1 (0,04)	
Faixa Etária				0,001
$[13 - 22]$	477 (16,99)	190 (6,77)	667 (23,76)	
$[23 - 27]$	593 (21,13)	221 (7,87)	814 (29,00)	
$[28 - 31]$	365 (13,00)	192 (6,84)	557 (19,84)	
$[32 - 50]$	503 (17,92)	266 (9,48)	769 (27,40)	
Total	1938 (69,04)	869 (30,96)	2807 (100,00)	

O teste para a normalidade de Anderson Darling aponta para a não normalidade para todas as variáveis incluídas no estudo, a 0,05% de significância, dado que o valor p é menor que 0,001. O teste de qui-quadrado foi realizado para verificar a hipótese de que as variáveis estejam interferindo diretamente no óbito do produto das gestações. Os resultados estão descritos na Tabela 3. Quanto a situação conjugal, de acordo com o teste, não pode-se rejeitar a hipótese, a 0,05% de significância dado valor p = 0,367, de que a

taxa de óbito independe da situação conjugal da mãe. Conclusão semelhante pode-se ter com relação a variável raça da mãe. De acordo com teste, com valor $p = 0,324$, não há indícios para afirmar, a 0,05 de significância, que a chance de óbito tenha dependência com a variável raça. Para o grau de instrução da mãe, pode-se afirmar que não há indícios para rejeitar a hipótese de que esta variável interfira diretamente na chance da criança nascer viva a 0,05 de significância, o teste aponta valor p = 0,018. A idade da mãe interfere diretamente na chance da criança vir a óbito, a 5% significância. O teste qui-quadrado obteve valor p = 0,001 (Tabela 3).

Na Figura 1 é apresentado o número de atendimentos, no ISEA, no ano de 2013 cujas informações fazem parte do banco de dados. Observa-se que o mês de Maio com 305 atendimentos dos quais 87 produtos de gestações vieram a óbito. Em seguida tem-se Julho com 279 casos registrados dos quais 66 foram óbito. Junho foi o terceiro mês com mais registros, foram 260 atendimentos registrados sendo 194 partos bem sucedidos e 66 casos de óbito. O mês com menor número de atendimento foi Março, com 161 registros. Desses 62 resultaram em óbito. Dezembro e Fevereiro foram os outros dois meses com menor número de partos, com frequências 200 e 201, respectivamente. No mês de dezembro registrou-se 61 óbitos e no mês de fevereiro foram registrados 77 óbitos.

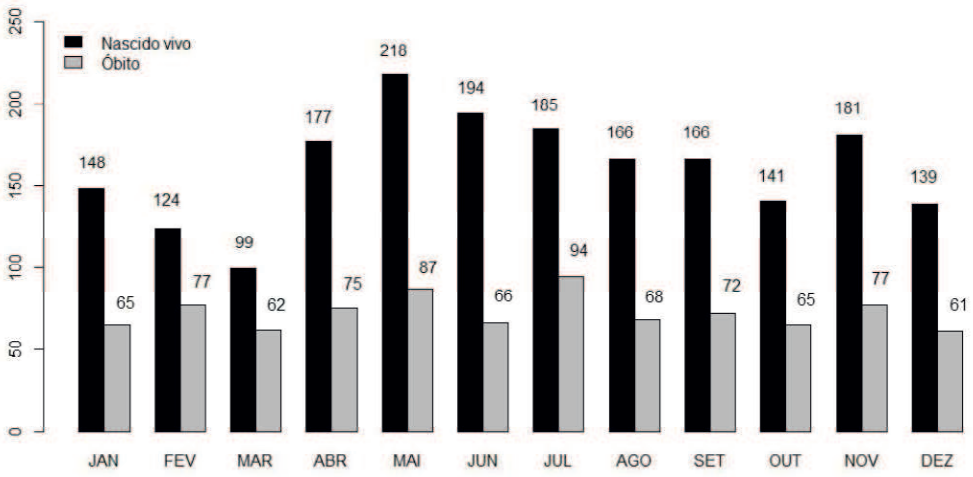

Figura 1 – Número de partos em que a criança nasceu viva e casos de óbitos por mês em 2013.

4.2 Escolha do Modelo

O vetor de respostas, $\mathbf{Y}$, neste caso é binária e assume valor 0 quando a mãe atendida declara não ter perdido seu filho em nenhuma de suas gestações e valor 1 quando

a mãe declara que não obteve sucesso em alguma de suas gestações. As variáveis explicativas são: a idade da mãe, a situação conjugal nos níveis solteira, casada, viúva e união estável. A raça autodeclarada pela paciente nos níveis branca, preta, parda e indígena. O grau de instrução nos níveis sem escolaridade, fundamental I, fundamental II, ensino médio, ensino superior incompleto e superior completo. Outra variável utilizada é o número de gestações.

Ajusta-se inicialmente o modelo saturado com tendência linear na variável idade, usando o método da máxima verossimilhança obtém-se as estimativas dos parâmetros. Observa-se por meio da Tabela 4 os resultados encontrados, a significância para as estimativas é de $0,05$. Observa-se que a variável raça nã foi significativa. O modelo saturado registrou $AIC = 3121,3$ e desvio residual de 3091,3.

Tabela 4 – Estimativas de máxima verossimilhança dos parâmetros do modelo saturado com tendência linear na variável idade e intervalo de confiança.

Parâmetro	**Estimativa**	**E. P.**	**valor t**	$IC_{2,5\%}$	$IC_{97,5\%}$	**Pr(>|t|)**
Intercepto	-2,0500	0,509	-4,003	-3,060	-1,06	0,001***
Idade (I)	-0,0476	0,008	-5,590	-0,060	-0,03	0,001***
Número de gestações (ν)	0,5973	0,036	16,440	0,530	0,670	0,001***
Situação Conjugal						
Casada (τ_2)	0,3514	0,111	-0,880	0,130	0,570	0,002**
Viúva (τ_3)	0,3234	0,708	0,460	-1,120	1,720	0,648
Divorciada (τ_4)	0,3925	0,629	0,620	-0,880	1,630	0,533
União estável (τ_5)	0,1493	0,120	1,250	-0,090	0,380	0,213
Raça/Cor						
Preta (ω_2)	0,2078	0,438	0,470	-0,680	1,050	0,635
Parda (ω_3)	0,0690	0,199	0,350	-0,310	0,470	0,729
Indígena (ω_4)	-11,4744	324,744	-0,040	-	54,700	0,972
Grau de escolaridade						
E. fundamental I (γ_2)	0,6910	0,397	1,740	-0,080	1,490	0,082.
E. fundamental II (γ_3)	0,9788	0,401	2,440	0,200	1,780	0,015*
E. Médio (γ_4)	1,2468	0,408	3,060	0,460	2,060	0,002**
E. S. Incompleto (γ_5)	0,4190	0,633	0,660	-0,90	1,620	0,508
E. S. Completo (γ_6)	1,3616	0,496	2,750	0,390	2,340	0,006**

*** Significante a 0,001
** Significante a 0,01
* Significante a 0,05
. Significante a 0,10

A Figura 2 observa-se o gráfico de envelope simulado dos cálculos teóricos versus os resíduos. Observe que mesmo a variável raça não sendo significativa para o modelo, a $0,05$ de significância, o gráfico mostra apenas 3,49% dos pontos fora do envelope.

Usando o método `stepwise`, observando os critérios AIC e o desvio residual, o

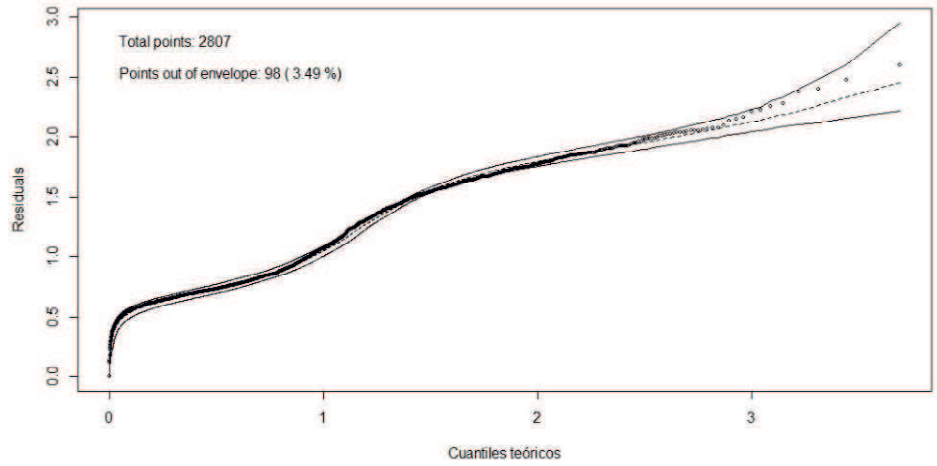

Figura 2 – Envelope simulado para os resíduos do modelo saturado com tendência linear na variável idade.

modelo foi reajustado para a função

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = -1,2695 - 0,0479I + 0,5914\nu + 0,3063\tau_2 + 0,3278\gamma_3 + 0,5906\gamma_4 + 0,7203\gamma_6 \tag{4.1}$$

O modelo (4.1) registrou $AIC = 3116,1$ e desvio residual 3092,1. Houve, portando, uma diminuição no valor de AIC e um pequeno aumento no valor do desvio residual. Na Tabela 5 observa-se as estimativas dos parâmetros e seus respectivos intervalos de confiança. Detecta-se que, com relação aos níveis de situação conjugal, mulheres casadas se diferencia estatisticamente dos demais níveis quanto a chance de óbito do filho. Em relação aos graus de instrução, mulheres com ensino fundamental II, ensino médio e ensino superior completo se diferenciam das mulheres analfabetas, com ensino fundamental I e/ou com ensino superior completo.

A Figura 3 mostra o envelope simulado dos resíduos para o modelo reduzido com tendência linear na variável idade. Observa-se que o modelo é significativo, a $0,05$ de significância, pois há apenas $3,95\%$ dos pontos simulados fora do envelope. Em outras palavras, o modelo se ajusta aos dados e todas as suas variáveis são significativas. Neste caso, pode-se conservar a hipótese de que o modelo se ajusta aos dados.

Espera-se que, naturalmente por motivos biológicos, a mulher ainda com pouca idade tenha chance zero de êxito na gestação e que atingido o período fértil, geralmente no início da adolescência, essa chance seja positiva e aumente consideravelmente com o passar dos anos atingindo um valor máximo. Posteriormente, quando a carga hormonal feminina tende a baixar, a chance da mulher ter sucesso no processo de gestação tende a diminuir. Esse comportamento não é linear.

Tabela 5 – Estimativas de máxima verossimilhança dos parâmetros do modelo reduzido com tendência linear na variável idade e respectivos intervalos de confiança.

Parâmetro	**Estimativa**	**E. P.**	**valor z**	$IC_{2,5\%}$	$IC_{97,5\%}$	**Pr(>|z|)**
Intercepto	-1,2695	0,241	-5,27	-1,741	-0,797	0,001***
Idade (I)	-0,0479	0,008	-5,65	-0,064	-0,031	0,001***
Número de Gestações (ν)	0,5914	0,036	16,60	0,522	0,661	0,001***
Situação Conjugal						
Casada (τ_2)	0,3063	0,106	2,88	0,098	0,515	0,004**
Grau de Instrução						
E. Fundamental II (γ_3)	0,3278	0,119	2,75	0,094	0,561	0,006**
E. Médio (γ_4)	0,5906	0,130	4,53	0,335	0,846	0,001***
E. S. Completo (γ_6)	0,7203	0,311	2,32	0,111	1,329	0,020*

*** Significante a 0,001
** Significante a 0,01
* Significante a 0,05

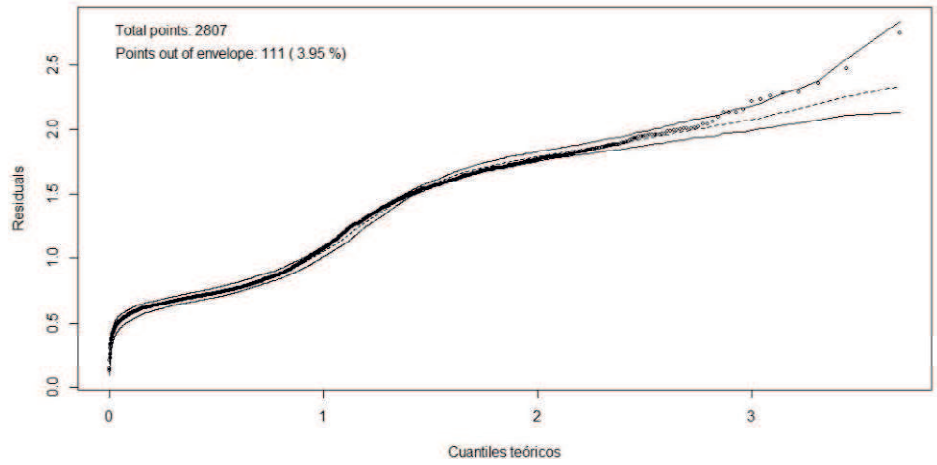

Figura 3 – Envelope simulado para os resíduos do modelo reduzido com tendência linear na variável idade.

Adicionando-se a tendência quadrática no modelo tem-se novas estimativas, que estão descritas na Tabela 6. O modelo possui $AIC = 3112,3$ e desvio residual de 3086,3. Com relação ao modelo reduzido com tendência linear na variável idade verifica-se uma queda de 3,8 no valor do AIC e uma redução de 5,8 no valor do desvio residual. Essa diminuição no valor do desvio é um indicativo de melhoramento no ajuste do modelo aos dados.

O modelo quadrático ajustado tem a equação $\log\left(\frac{\theta_i}{1-\theta_i}\right) = -0,1781I + 0,0023I^2 + 0,6007\nu + 0,3583\tau_2 + 1,0268\gamma_3 + 1,3180\gamma_4 + 1,3901\gamma_6$. O modelo apresentou o menor AIC e o menor valor para o desvio. Aparentemente é o modelo que melhor explica a relação dos dados com a chance de óbito. As estimativas dos parâmetros do modelo podem ser

vistas na Tabela 6.

Tabela 6 – Estimativas de máxima verossimilhança dos parâmetros do modelo reduzido com tendência quadrática na variável idade e intervalo de confiança.

Parâmetro	**Estimativa**	**E. P.**	**valor z**	$IC_{2,5\%}$	$IC_{97,5\%}$	**Pr(>|z|)**
Intercepto	-0,2805	0,843	-0,330	-1,940	1,370	0,739
Idade I	-0,1781	0,054	-3,270	-0,280	-0,070	0,001**
Idade^2	0,0023	0,001	2,430	0,001	0,004	0,015*
Número de Gestações (ν)	0,6007	0,036	16,470	0,5300	0,670	0,001***
Situação Conjugal						
Casada (τ_2)	0,3583	0,112	3,210	0,140	0,580	0,002**
Viúva (τ_3)	0,3311	0,708	0,470	-1,120	1,730	0,640
Divorciada (τ_4)	0,4299	0,629	0,680	-0,840	1,660	0,494
União Estável (τ_5)	0,1503	0,120	1,250	-0,090	0,380	0,209
Grau de Instrução						
E. Fundamental I (γ_2)	0,7372	0,404	1,820	-0,040	1,550	0,068.
E. Fundamental II (γ_3)	1,0268	0,408	2,520	0,240	1,850	0,012*
E. Médio (γ_4)	1,3180	0,415	3,180	0,520	2,150	0,002**
E. S. Incompleto (γ_5)	0,5043	0,638	0,790	-0,820	1,720	0,429
E. S. Completo (γ_6)	1,3901	0,501	2,770	0,4100	2,380	0,006**

*** Significante a 0,001
** Significante a 0,01
* Significante a 0,05
. Significante a 0,10

A Figura 4 mostra o envelope dos quantis teóricos versus os resíduos. Neste modelo, reduzido com tendência quadrática na variável idade, a porcentagem de pontos fora do envelope aumentou quando comparado com o modelo reduzido e apenas tendência linear na idade. Foram registrados 245, ou seja, $8,73\%$ dos pontos fora do envelope. Significa que não pode-se admitir que o modelo é significativo, a $0,05$ de significância, para explicar a relação entre os dados e a chance de óbito.

Apesar do valor do AIC e do desvio residual serem menores no modelo reduzido quadrático, nos resíduos houve uma piora, ou seja, parece ter havido um aumento significativo na distância entre o modelo e os dados em alguma parte da curva. Foi observado que o modelo (4.1) é o que nos dá a melhor relação das variáveis explicativas idade, número de gestações, grau de instrução e situação conjugal com a variável resposta $\mathbf{Y}$.

A Tabela 7 nos ajuda a compreender a seleção do modelo, nela podemos ver os valores do AIC, dos desvios residuais e da porcentagem de pontos fora do envelope simulado para cada modelo.

Selecionado o modelo que tem as melhores condições para modelar a relação entre a chance de óbito e as variáveis explicativas fizemos a atualização dos parâmetros por meio do Teorema de Bayes.

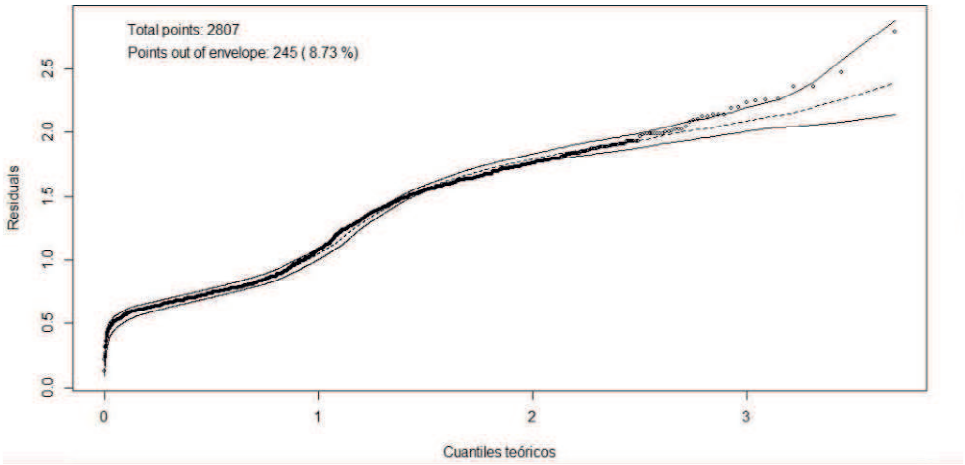

Figura 4 – Envelope simulado para os resíduos do modelo reduzido com tendência quadrática na variável idade.

Tabela 7 – Tabela resumo dos valores de AIC, desvio residual e percentual de pontos fora do envelope simulado.

Modelo	**AIC**	**Desvio Residual**	**Percentual(%)**
Linear Saturado	3121,3	3091,3	3,49%
Linear Reduzido	3116,1	3092,1	3,95%
Quadrático Reduzido	3112,3	3086,3	8,73%

Cada parâmetro tem distribuição a priori $N \sim (\mu_{MV}, 100)$, em que μ_{MV} é sua respectiva estimativa de máxima verossimilhança dada no modelo clássico. Isso reflete a informação a pouca informação que temos sobre os valores da variabilidade de cada parâmetro. Simulando 11 mil vezes, com saltos de 10 estimativas, sendo que as primeiras mil simulações são descartadas obtemos as atualizações dos parâmetros do modelo. A Tabela 8 apresenta a média e o desvio padrão e o intervalo de credibilidade de cada parâmetro.

Tabela 8 – Estimativas dos parâmetros do modelo bayesiano e intervalos de credibilidade.

Parâmetro	**Média**	**Desvio Padrão**	$Ic_{2,5\%}$	$Ic_{97,5\%}$
Intercepto	-1,26011	0,245500	-1,73867	-0,79012
Idade	-0,04864	0,008742	-0,06544	-0,03154
Número de Gestações (ν)	0,59535	0,036691	0,52452	0,66571
Situação Conjugal				
Casada (τ_2)	0,30285	0,107853	0,07748	0,50499
Grau de Instrução				
E. Fundamental II (γ_3)	0,32936	0,124712	0,06613	0,58345
E. Médio (γ_4)	0,59074	0,134722	0,33933	0,85858
E. Superior Completo (γ_6)	0,72752	0,312674	0,09130	1,31497

O modelo bayesiano é dado pela equação (4.2) a seguir.

$$\log\left(\frac{\theta_i}{1-\theta_i}\right) = -1,26011 - 0,04864I + 0,59535\nu + 0,30285\tau_2 + 0,32936\gamma_3 + 0,59074\gamma_4 + 0,72752\gamma_6 \tag{4.2}$$

A sequência de figuras a seguir, Figura 5 até a Figura 18, mostra as séries temporais dos parâmetros e as suas respectivas densidades a posterioris. Podemos ver a convergência dos parâmetros nas figuras.

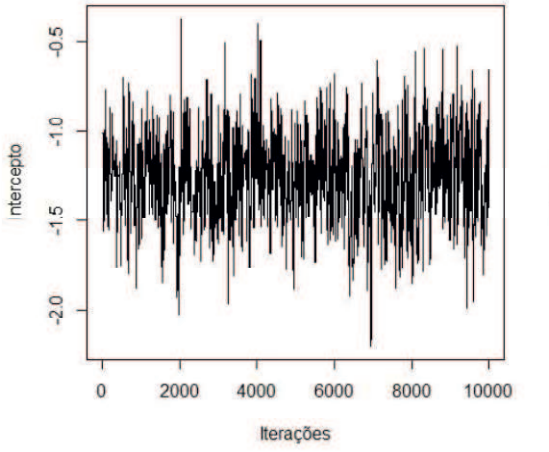

Figura 5 – Valores estimados nas iterações do Intercepto.

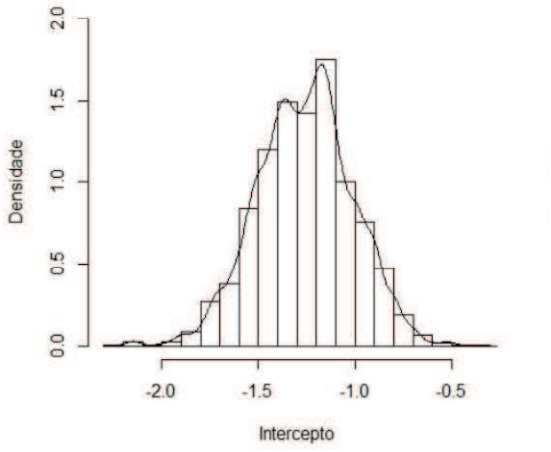

Figura 6 – Densidade a posteriori do Intercepto.

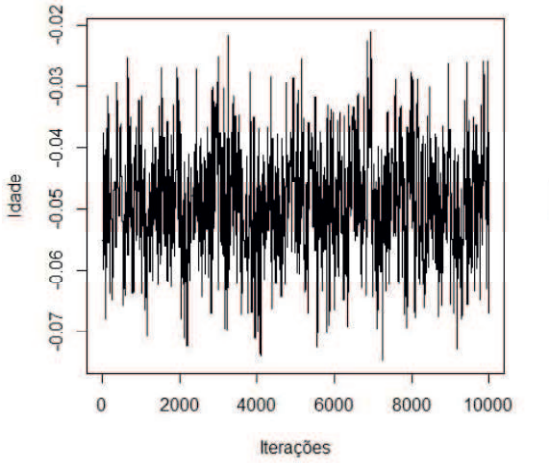

Figura 7 – Valores estimados nas iterações do parâmetro (I).

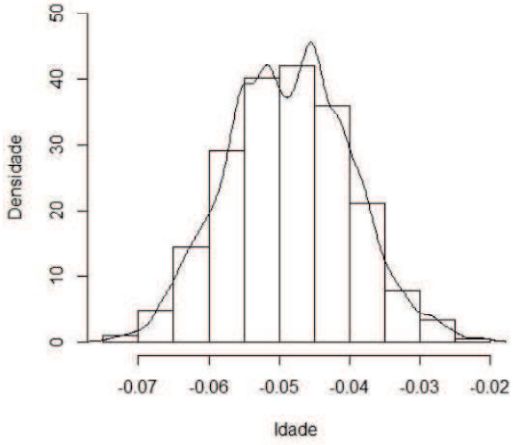

Figura 8 – Densidade a posteriori do parâmetro (I).

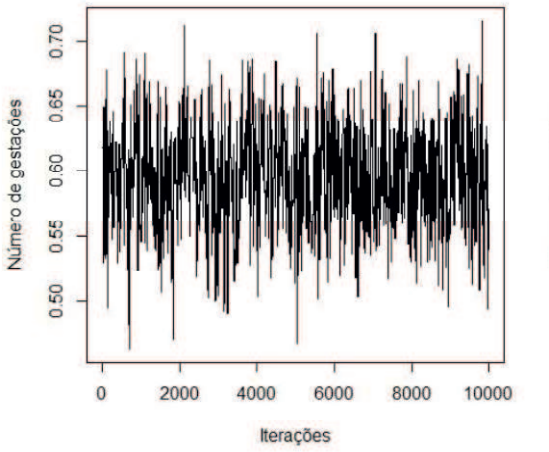

Figura 9 – Valores estimados nas iterações do parâmetro ν.

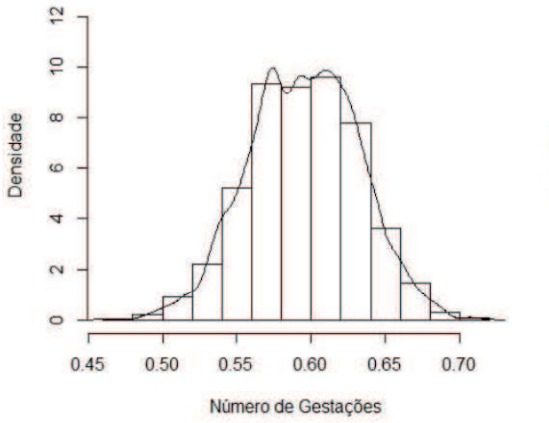

Figura 10 – Densidade a posteriori do parâmetro ν.

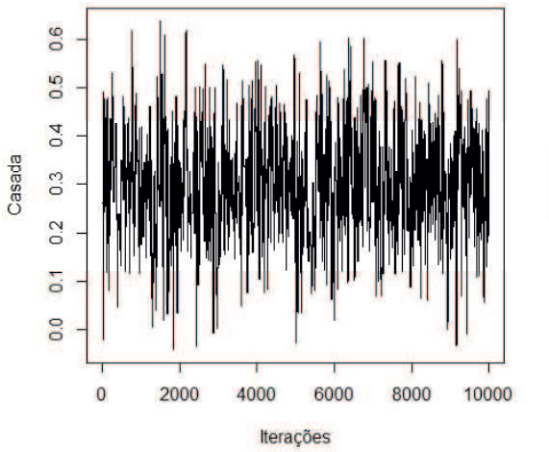

Figura 11 – Valores estimados nas iterações do parâmetro τ_2.

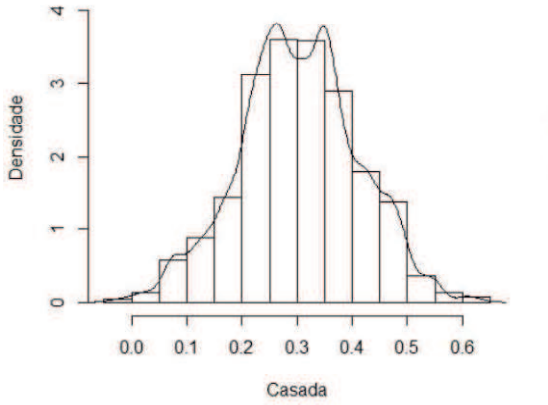

Figura 12 – Densidade a posteriori do parâmetro τ_2.

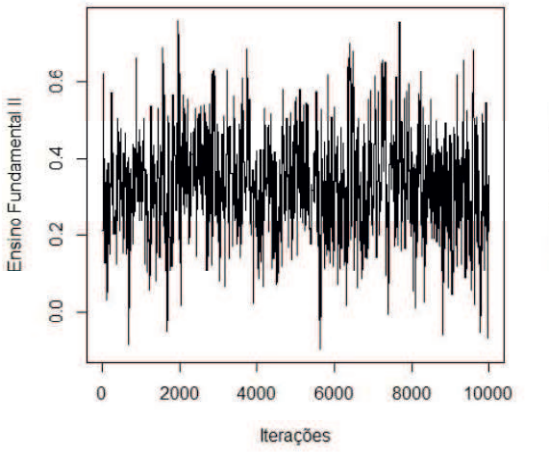

Figura 13 – Valores estimados nas iterações do parâmetro γ_3.

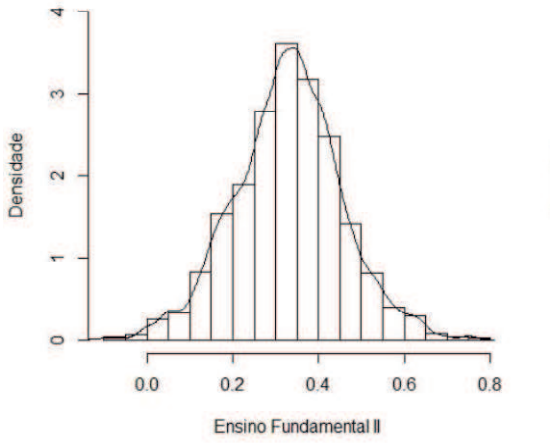

Figura 14 – Densidade a posteriori do do parâmetro γ_3.

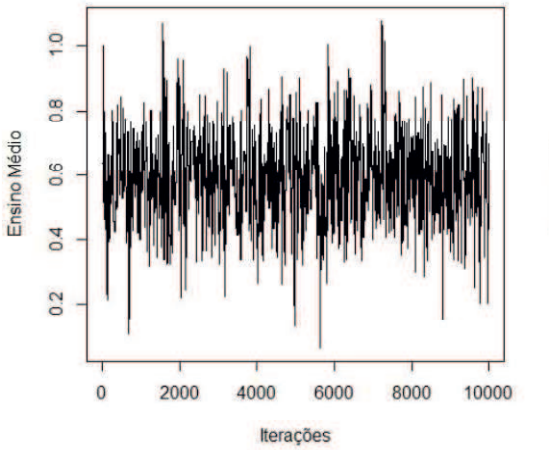

Figura 15 – Valores estimados nas iterações do parâmetro γ_4.

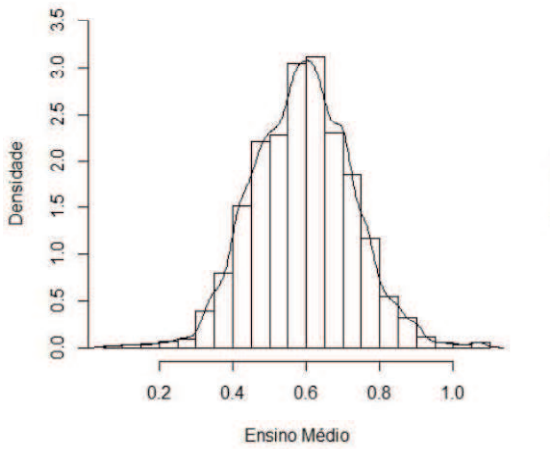

Figura 16 – Densidade a posteriori do parâmetro γ_4.

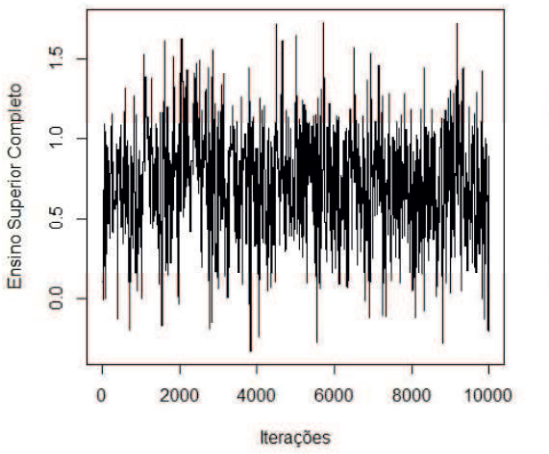

Figura 17 – Valores estimados nas iterações do parâmetro γ_6.

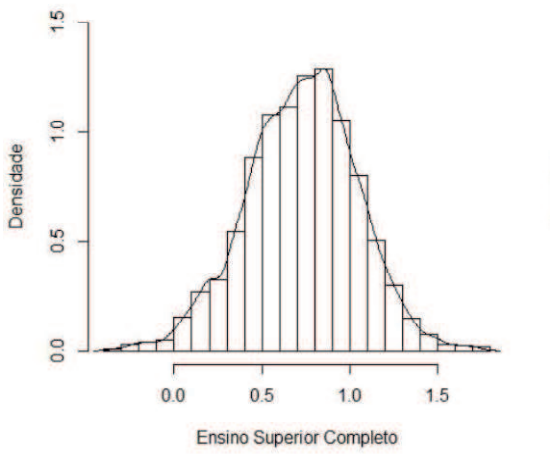

Figura 18 – Densidade a posteriori do parâmetro γ_6.

4.2.1 Interpretação dos modelos e predições

O modelo ajustado por verossimilhança, (4.1), é bastante semelhante ao modelo Bayesiano, (4.2). Observa-se essa semelhança pela proximidade dos valores dos parâmetros. Os valores preditos pelos dois modelos também se assemelham, conforme mostra a Tabela 9 e a Tabela 10. O modelo (4.1) mostra 1881 casos de crianças nascidas vivas, confirmado pelos dados, e 127 casos de óbitos dados que nasceram vivas. O mesmo modelo também nos diz que 649 casos de crianças nascidas vivas quando na verdade elas foram casos de óbito e 220 casos de óbitos confirmados pelos dados.

Tabela 9 – Matriz de confusão do modelo (4.1).

		Valores Preditos	
		Nascido vivo	Óbito
Valores Reais	Nascido vivo	1811	127
	Óbito	649	220

O modelo Bayesiano, (4.2), nos mostra resultado bastante semelhante. Esse modelo nos mostra 1810 casos de crianças que nasceram vivas, casos estes confirmados pelos dados, enquanto que 128 casos foram registrados como óbito erroneamente. Portanto o modelo (4.1) obteve 1 acerto a mais que o modelo Bayesiano. O modelo (4.2) ainda detecta 649 casos de nascidos vivos quando na verdade os indivíduos foram a óbito e 220 casos de óbitos quando realmente os indivíduos foram a óbito.

Na Tabela 11 tem-se a acurácia, a taxa positiva verdadeira (TP), a taxa de falsos positivos (FP), a taxa negativa verdadeira (TN), a taxa de falsos negativos (FN), e a

Tabela 10 – Matriz de confusão do modelo (4.2).

		Valores Preditos	
		Nascido vivo	Óbito
Valores Reais	Nascido vivo	1810	128
	Óbito	649	220

precisão para os dois modelos ajustados. Os valores são praticamente iguais, alguns diferem muito pouco. O modelo (4.2) tem uma acurácia um pouco maior enquanto que o modelo (4.1) tem maior precisão.

Tabela 11 – Indicadores dos modelos ajustados.

	Acurácia	TP	FP	TN	FN	Precisão
Modelo (4.1)	0,7232	0,2532	0,0655	0,9344	0,7468	0,6340
Modelo (4.2)	0,7235	0,2532	0,0660	0,9339	0,7468	0,6322

A Figura 19 mostra em preto uma amostra observada de 100 desfechos (nascidos vivos ou óbitos), em vermelho a previsão Bayesiana. O valores, em vermelho, são obtidos probabilisticamente

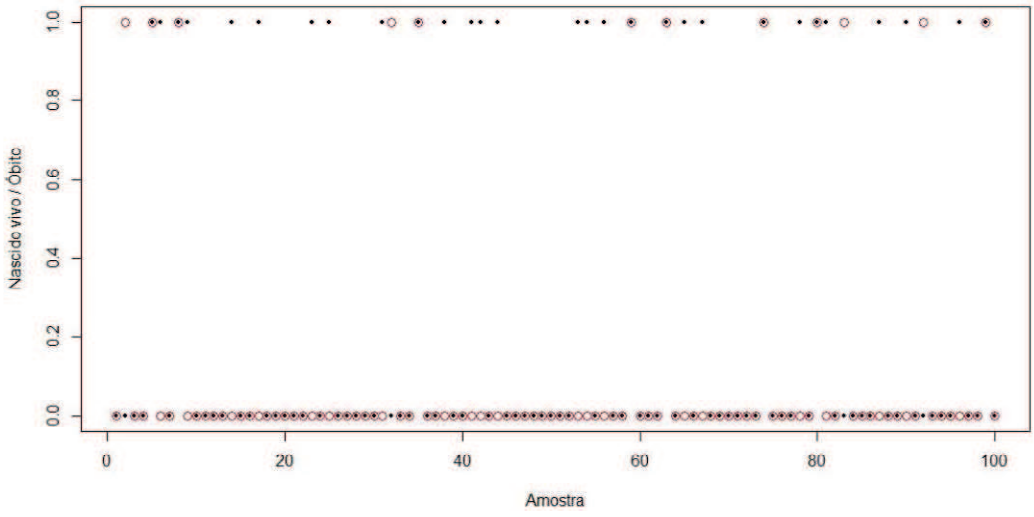

Figura 19 – Pontos preditos pelo modelo Bayesiano para uma amostra de 100 mulheres.

Simulando alguns resultados possíveis valores para as variáveis do modelo, e adotando o modelo Bayesianos, calculou-se a probabilidade estimada ($\hat{\theta}$) de óbito para cada situação. Observe a Tabela 12, destaca-se três situações. Na linha três, tem-se a situação em que a mãe de primeira viagem tem 13 anos de idade, não é casada e não tem ensino possui o fundamental II, a probabilidade de óbito da criança é de $0,215$. Na linha sete, a mãe têm 45 anos de idade, está em sua sétima gestação, não é casada e possui

ensino médio. Nesta situação a probabilidade de desfecho trágico de sua gestação é de $0,740$. Por fim, na linha nove, tem-se uma mãe que possui ensino superior completo, não é casada e está gerando seu primeiro filho com 50 anos de idade, a probabilidade de óbito da criança é de apenas $0,085$.

Tabela 12 – Cálculo da probabilidade de óbito para 10 situações aleatórios possíveis.

Idade (I)	Número de Gestação (ν)	Situação Conjugal	Grau de Instrução	$\hat{\theta}$
26	2	Outros	E. Fundamental II	0,268
29	2	Outros	Ensino Médio	0,291
13	1	Outros	Outros	0,215
38	1	Outros	E. Fundamental II	0,101
28	1	Casada	Outros	0,199
40	4	Outros	Outros	0,305
45	7	Outros	E. Médio	0,740
26	3	Outros	Outros	0,323
50	1	Outros	E. S. Completo	0,085
24	1	Outros	Outros	0,138

5 Conclusão

Neste trabalho pôde-se constatar que a idade da mãe é um fator que interfere diretamente na chance de uma criança vir a óbito no período perinatal assim como o número de gestações. As outras variáveis que interferem são o fato da mãe ser casada e ter ensino fundamental II, ter ensino médio ou curso superior completo.

O modelo estimado por máxima verossimilhança e mais adequado foi o modelo $\log\left(\frac{\theta_i}{1-\theta_i}\right) = -1,2695 - 0,0479I + 0,5914\nu + 0,3063\tau_2 + 0,3278\gamma_3 + 0,5906\gamma_4 + 0,7203\gamma_6$, que foi atualizado, usando as técnicas Bayesianas, para o modelo $\log\left(\frac{\theta_i}{1-\theta_i}\right) = -1,26011 - 0,04864I + 0,59535\nu + 0,30285\tau_2 + 0,32936\gamma_3 + 0,59074\gamma_4 + 0,72752\gamma_6$. Os modelos, tanto estimado pelo método da máxima verossimilhança quanto o modelo atualizado pelo Teorema de Bayes predizem resultados muito próximos um do outro. Os parâmetros do modelo Bayesiano não apresentou grande diferença para os valores dos parâmetros do modelo clássico.

O modelo quadrático, apesar de teoricamente ser mais plausível, não apresentou bons resultados nos resíduos. Os resíduos dos modelos lineares apresentaram melhores comportamentos. No entanto não quer dizer que os modelos (4.1) e (4.2) sejam os melhores modelos para representar essa relação entre variáveis maternas e a chance de óbito de uma criança, foram os melhores que encontramos neste momento.

Há uma diversidade de modelos que podem ser testados. Pode-se usar outra distribuição, por exemplo a distribuição Quase-Binomial, pode-se estruturar modelos mais complexos que modelam a assimetria e a curtose além da média e da variância. Fica a sugestão para os futuros trabalho, estruturar um modelo que minimize a distância entre a equação ajustada e os dados e supere em termos de qualidade os modelos (4.1) e (4.2).

Referências

AZEVEDO, L. P. de. *APLICAÇÃO DE REDES NEURAIS ARTIFICIAIS NO PROCESSO DE CLASSIFICAÇÃO DE ORQUÍDEAS DO GÊNERO CATTLEYA*. Dissertação (Mestrado) — INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DE MINAS GERAIS, dez. 2016. Citado na página 26.

BOLFARINE, H.; SANDOVAL, M. C. *Introdução à inferência estatística*. Rio de Janeiro: [s.n.], 2001. v. 2. Citado na página 16.

CONGALTON, R. G. A review of assessing the accuracy of classifications of remotely sensed data. *Remote sensing of environment*, Elsevier, v. 37, n. 1, p. 35–46, 1991. Citado na página 25.

CORDEIRO, G. M.; DEMÉTRIO, C. G. Modelos lineares generalizados e extensões. *Sao Paulo*, 2008. Citado 3 vezes nas páginas 14, 15 e 16.

CORDEIRO, G. M.; PAULA, G. A. *Modelos de regressão para análise de dados univariados*. São Paulo: IMPA, 1989. Citado na página 18.

COSTA, A. P. *Análise das Ações Essenciais Preconizadas Pelo Programa de Humanização do Pré-Natal e Nascimento a Partir do Cartão da Gestante*. Dissertação (Mestrado) — Universidade Federal da Paraíba, 2012. Citado na página 13.

COX, D. R. et al. The role of significance tests [with discussion and reply]. *Scandinavian Journal of Statistics*, JSTOR, p. 49–70, 1977. Citado na página 24.

DEMETRIO, C. B. Modelos lineares generalizados e extensões. *Statistical modelling*, Springer Verlag, v. 57, p. 95, 1989. Citado 2 vezes nas páginas 14 e 23.

DERKSEN, S.; KESELMAN, H. J. Backward, forward and stepwise automated subset selection algorithms: Frequency of obtaining authentic and noise variables. *British Journal of Mathematical and Statistical Psychology*, Wiley Online Library, v. 45, n. 2, p. 265–282, 1992. Citado na página 10.

FIRMINO, M. J. d. A. C. et al. *Testes de hipóteses: uma abordagem não paramétrica*. Tese (Doutorado), 2015. Citado na página 20.

FLACH, P. A. The geometry of roc space: understanding machine learning metrics through roc isometrics. In: *Proceedings of the 20th International Conference on Machine Learning (ICML-03)*. [S.l.: s.n.], 2003. p. 194–201. Citado na página 25.

FONSECA, J. S. d.; MARTINS, G. d. A. *Curso de estatística*. Dissertação (Mestrado), 1996. Citado na página 11.

GIOLO, S. R.; COLOSIMO, E. A. Análise de sobrevivência aplicada. *Edgard Blucher*, 2006. Citado na página 24.

JAMES, G. et al. *An introduction to statistical learning*. New York: Springer, 2013. v. 112. Citado na página 22.

JENNRICH, R.; SAMPSON, P. Application of stepwise regression to non-linear estimation. *Technometrics*, Taylor & Francis, v. 10, n. 1, p. 63–72, 1968. Citado na página 10.

KINAS, P. G.; ANDRADE, H. A. *Introdução à análise bayesiana (com R)*. São Paulo: maisQnada, 2010. ISBN 978-85-61797-10-2. Citado 4 vezes nas páginas 11, 18, 19 e 20.

LUNN, D. et al. *The BUGS book: A practical introduction to Bayesian analysis*. [S.l.]: Chapman and Hall/CRC, 2012. Citado na página 23.

MARDIA, K. V.; MARSHALL, R. J. Maximum likelihood estimation of models for residual covariance in spatial regression. *Biometrika*, Oxford University Press, v. 71, n. 1, p. 135–146, 1984. Citado na página 10.

MCCULLAGH, P.; NELDER, J. A. *Generalized linear models*. New York: CRC press, 1989. v. 37. Citado na página 14.

MCHUGH, M. L. The chi-square test of independence. *Biochemia medica: Biochemia medica*, Medicinska naklada, v. 23, n. 2, p. 143–149, 2013. Citado na página 20.

MINISTÉRIO DA SAÚDE. *Entendendo o SUS*. Brasília, 2006. Disponível em: <http://portalsaude.saude.gov.br/index.php/cidadao/entenda-o-sus>. Citado na página 12.

MOORE, D. S. *Chi-Square Tests*. Virginia, 1976. Citado na página 20.

MORAL, R. A.; HINDE, J.; DEMÉTRIO, C. G. Half-normal plots and overdispersed models in r: The hnp package. *Journal of Statistical Software*, v. 81, n. 10, p. 1–23, 2017. Citado 2 vezes nas páginas 11 e 25.

NELDER, J. A.; BAKER, R. J. *Generalized linear models*. New York: Wiley Online Library, 1972. Citado na página 24.

NELSON, L. S. The anderson-darling test for normality. *Journal of Quality Technology*, Taylor & Francis Ltd., v. 30, n. 3, p. 298, 1998. Citado na página 21.

PARAÍBA, J. da. *Mais de 5 mil partos feitos no Isea-CG são de outras cidades e caso vai ao MFP*. Campina Grande: Reportagem, 2018. Citado na página 13.

PAULA, G. A. *Modelos de regressão: com apoio computacional*. São Paulo: IME-USP São Paulo, 2004. Citado 5 vezes nas páginas 15, 22, 23, 24 e 25.

PAULINO, C. D. M.; TURKMAN, M. A. A.; MURTEIRA, B. *Estatística bayesiana*. Lisboa: FUNDAÇÃO CALOUSTE GULBENKIAN, 2003. ISBN 972-31-1043-1. Citado 4 vezes nas páginas 19, 20, 22 e 23.

PICHE, R.; PENTTINEN, A. et al. Bayesian methods. 2010. Citado na página 11.

RETALHOS HISTÓRICOS DE CAMPINA GRANDE. *Memória Fotográfica: A Maternidade Elpídio de Almeida (ISEA)*. Campina Grande, 2011. Disponível em: <http://cgretalhos.blogspot.com/search?q=ISEA#.WxIE-kgvzIV>. Citado na página 13.

SARINHO, S. W. Mortalidade neonatal na cidade do recife: um estudo caso-controle. 1998. Citado na página 10.

SHESKIN, D. J. *Handbook of parametric and nonparametric statistical procedures*. New York: crc Press, 2003. Citado 2 vezes nas páginas 21 e 22.

SILVA, C. F. d. et al. Fatores de risco para mortalidade infantil em município do nordeste do brasil: linkage entre bancos de dados de nascidos vivos e óbitos infantis-2000 a 2002. *Revista Brasileira de Epidemiologia*, SciELO Public Health, v. 9, p. 69–80, 2006. Citado na página 10.

VICTORA, C. G. et al. Saúde de mães e crianças no brasil: progressos e desafios. Lancet, 2011. Citado na página 10.

VIEIRA, S. *Introdução à bioestatística*. Rio de Janeiro: Elsevier Brasil, 2015. Citado na página 20.