



UNIVERSIDADE ESTADUAL DA PARAÍBA  
CENTRO DE CIÊNCIAS E TECNOLOGIA  
DEPARTAMENTO DE ESTATÍSTICA

Adriana Dionísio Coêlho

**Modelo de regressão logístico clássico e  
Bayesiano aplicado a dados de partos  
prematurados no Instituto de Saúde Elpídeo de  
Almeida, Campina Grande - PB**

Campina Grande - PB

Dezembro de 2018

Adriana Dionísio Coêlho

**Modelo de regressão logístico clássico e Bayesiano  
aplicado a dados de partos prematuros no Instituto de  
Saúde Elpídeo de Almeida, 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

É expressamente proibido a comercialização deste documento, tanto na forma impressa como eletrônica. Sua reprodução total ou parcial é permitida exclusivamente para fins acadêmicos e científicos, desde que na reprodução figure a identificação do autor, título, instituição e ano do trabalho.

C672m Coêlho, Adriana Dionísio.

Modelo de regressão logístico clássico e Bayesiano aplicado a dados de partos prematuros no Instituto de Saúde Elpídeo de Almeida, Campina Grande - PB [manuscrito] / Adriana Dionísio Coelho. - 2018.

33 p.

Digitado.

Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2019.

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

1. Modelo de Regressão logístico. 2. Inferência Bayesiana.  
3. Prematuridade. I. Título

21. ed. CDD 519.5

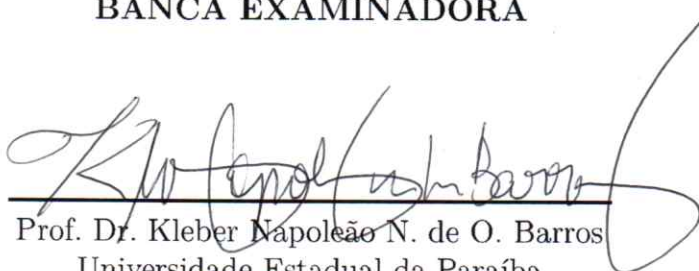
Adriana Dionísio Coêlho

**Modelo de regressão logístico clássico e Bayesiano  
aplicado a dados de partos prematuros no Instituto de  
Saúde Elpídeo de Almeida, Campina Grande - PB**

Trabalho de Conclusão de Curso apresentado  
ao curso de Bacharelado em Estatística do  
Departamento de Estatística do Centro de Ci-  
ências e Tecnologia da Universidade Estadual  
da Paraíba em cumprimento às exigências le-  
gais para obtenção do título de bacharel em  
Estatística.

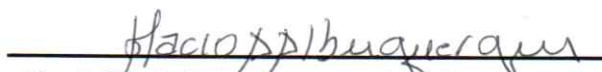
Trabalho aprovado em 18 de dezembro de 2018.

**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



---

Prof. Dr. Gustavo Henrique Esteves  
Universidade Estadual da Paraíba

# Agradecimentos

Dedico este trabalho, primeiramente a Deus, pois sem ele, não teria chegado até aqui. A minha mãe Maria Zuleide, que sempre me apoiou, confiou em mim e não mediu esforços pra que este sonho se realizasse, pois sem a compreensão, ajuda e confiança dela nada disso seria possível hoje.

Ao meu pai Manuel Dionísio (in memorian), que infelizmente não pode estar presente neste momento tão especial na minha vida, mas que não poderia deixar de dedicar a ele, pois se hoje estou aqui, devo muitas coisas a ele e por seus ensinamentos e valores passados. Obrigada por tudo! Saudades eternas!

Agradeço, especialmente ao meu esposo Diego Adolfo que entendeu as minhas ausências e me apoiou, incentivou-me nos momentos em que tive dificuldades para dar continuidade a essa caminhada. E aos meus filhos Enzo Gabriel e Alexia, que foram fundamentais, para que esse trabalho pudesse ser realizado.

Ao meu orientador, Prof. Dr. Kleber Napoleão Nunes de Oliveira Barros, o meu sincero agradecimento, pela orientação valiosa, confiança e amizade, antes de tudo por ter acreditado nesse trabalho, e ter me ajudado a realizar um sonho.

Aos professores do curso que foram tão essenciais na minha vida acadêmica.

A banca examinadora Prof. Dr. Mácio Augusto de Albuquerque e Prof. Dr. Gustavo Henrique Esteves, por ter aceito o convite e dividirem comigo esse momento tão importante.

Não poderia deixar de agradecer a minha amiga Rafaella Vitorino, que esteve comigo nessa longa caminhada.

E para finalizar, agradeço aos meus amigos, que fizeram parte da minha formação e continuará presente em minha vida.

*"Deus dá as batalhas mais difíceis aos seus melhores soldados".  
(Papa Francisco)*

# Resumo

A prematuridade se destaca como um dos maiores problemas de saúde pública em virtude das altas taxas de mortalidade neonatal, infantil e na vida adulta. É definido prematuro, toda criança que nasce com idade gestacional menor que 37 semanas. O objetivo do presente trabalho é investigar quais fatores estão associados à prematuridade das crianças no Instituto de Saúde Elpídeo de Almeida, no município de Campina Grande - PB. Foi realizada a análise dos dados no modelo de regressão logístico, aplicando o modelo linear generalizado considerando a distribuição binomial com função *logit*. No entanto, o modelo empregado foi adequado para explicar a prematuridade das crianças. Posteriormente, foram realizadas inferência clássica e Bayesiana, verificou-se que as estimativas tanto da clássica, quanto da Bayesiana foram muito próximos. A variável dependente em estudo foi a prematuridade, seguindo das variáveis independentes: idade materna, tipo de gravidez, número de consultas, número de filho(s) vivo(s) e peso da criança ao nascer, que foram estatisticamente significativas. As análises foram implementadas através do *software R* (R Core Team, 2018) e *OpenBUGS* (Thomas, 2004).

**Palavras-chaves:** Modelo de Regressão logístico. Inferência Bayesiana. Prematuridade.

# Abstract

Prematurity stands out as one of the greatest public health problems due to the high rates of neonatal, infant and adult mortality. It is defined as premature, every child born with gestational age less than 37 weeks. The objective of the present study is to investigate which factors are associated with the prematurity of the children at the Elpídeo de Almeida Health Institute, Campina Grande - PB. Data analysis was performed in the logistic regression model, applying the generalized linear model considering the binomial distribution with *logit* function. However, the model employed was adequate to explain the prematurity of the children. Afterwards, classical and Bayesian inference were performed, it was verified that the estimates of both the classical and Bayesian were very close. The dependent variable was preterm birth, followed by the following independent variables: maternal age, type of pregnancy, number of consultations, number of live children and birth weight, which were statistically significant. The analyzes were implemented through *software R* (R Core Team, 2018) and *OpenBUGS* (Thomas, 2004).

**Key-words:** Logistic Regression Model. Bayesian Inference. Prematurity.



# Lista de ilustrações

- Figura 1 – Envelope simulado para os resíduos do modelo com distribuição binomial. 27
- Figura 2 – Representação gráfica do traço e da função densidade posteriori . . . . 29

# Lista de tabelas

Tabela 1 – Características de algumas distribuições pertencentes à família exponencial	14
Tabela 2 – Notação para estudo de coorte ou caso-controle . . . . .	15
Tabela 3 – Matriz de confusão . . . . .	21
Tabela 4 – Distribuição das características sócio-demográfico das mães relacionado à prematuridade assistido no ISEA . . . . .	24
Tabela 5 – Distribuição das características obstétricas relacionados à prematuridade assistido no ISEA . . . . .	25
Tabela 6 – Resumo das estatísticas descritivas de algumas das variáveis explicativas	26
Tabela 7 – Descrição das covariáveis utilizadas no estudo sobre a prematuridade .	26
Tabela 8 – Seleção de covariáveis considerando o modelo binomial . . . . .	27
Tabela 9 – Estimativas dos parâmetros do modelo via análise clássica . . . . .	28
Tabela 10 – Estimativas dos parâmetros do modelo via análise bayesiana . . . . .	28
Tabela 11 – Matriz de confusão do modelo clássico . . . . .	29
Tabela 12 – Matriz de confusão do modelo bayesiano . . . . .	30
Tabela 13 – Indicadores do modelo ajustado . . . . .	30

# Sumário

<b>1</b>	<b>INTRODUÇÃO</b>	<b>10</b>
<b>2</b>	<b>REVISÃO DE LITERATURA</b>	<b>11</b>
<b>2.1</b>	<b>SUS</b>	<b>11</b>
<b>2.2</b>	<b>Prematuridade em recém nascidos</b>	<b>11</b>
<b>2.3</b>	<b>ISEA</b>	<b>12</b>
<b>2.4</b>	<b>MLG</b>	<b>13</b>
<b>2.5</b>	<b>Regressão Logística</b>	<b>14</b>
<b>2.6</b>	<b>Inferência Bayesiana</b>	<b>16</b>
2.6.1	MLG Bayesiano	16
2.6.1.1	MLG	16
2.6.1.2	Regressão logística	17
2.6.2	OpenBugs	17
<b>2.7</b>	<b>Seleção de Modelos</b>	<b>18</b>
2.7.1	Coefficiente de Determinação ( $R^2$ )	18
2.7.2	Critério de Informação de Akaike - AIC	18
2.7.3	Critério de Informação Bayesiano - BIC	19
2.7.4	Critério de Informação de Desvio - DIC	19
2.7.5	<i>Stepwise (Forward, Backward, Bidirecional)</i>	19
2.7.5.1	<i>Forward</i>	20
2.7.5.2	<i>Backward</i>	20
2.7.5.3	<i>Bidirecional</i>	20
<b>2.8</b>	<b>Matriz de confusão</b>	<b>21</b>
<b>3</b>	<b>MATERIAL E MÉTODOS</b>	<b>22</b>
<b>3.1</b>	<b>Estrutura dos dados</b>	<b>22</b>
<b>3.2</b>	<b>Modelo Empregado</b>	<b>22</b>
<b>4</b>	<b>RESULTADOS E DISCUSSÃO</b>	<b>23</b>
<b>5</b>	<b>CONCLUSÃO</b>	<b>31</b>
	<b>REFERÊNCIAS</b>	<b>32</b>

# 1 Introdução

A prematuridade é um dos grandes desafios para a saúde pública, em virtude das elevadas taxas de mortalidade neonatal, infantil e na vida adulta. A prematuridade representou 15,31% dos nascimentos em 2013, no Instituto Elpídeo de Almeida. E dentre os fatores de risco, podemos encontrar, idade materna, tipo de gravidez, número de filhos, controle pré-natal e peso da criança ao nascer.

A regressão logística é uma técnica estatística que tem uma vasta aplicação em diferentes áreas do conhecimento. É frequentemente empregado em situações em que a variável dependente é de natureza binária (sucesso ou fracasso) e seus valores são conclusivos em termos de probabilidade. Sendo que, na área médica é preciso entender a respeito da doença e os fatores que contribuíram para o seu aparecimento e sua evolução.

Os Modelos Lineares Generalizados são ferramentas importantíssimas na análise de dados, tendo em vista, o interesse em estabelecer de forma significativa a relação entre uma variável resposta, em função das variáveis preditoras. Em alguns casos, quando se aplica determinada metodologia de análise, são requeridas alguns pressupostos que nem sempre poderão ser totalmente atendidas, ou seja, muitas das vezes a suposição de normalidade para variável resposta, não se chega a resultados de forma coerente. Visando estes problemas, Nelder e Wedderburn (1972), analisaram uma classe de modelos que compreendesse melhor essa relação de não normalidade e propuseram outras distribuições que pertencessem à família exponencial.

Na abordagem Bayesiana, a distribuição à priori requer um conhecimento prévio sobre o parâmetro de interesse  $\theta$  antes da observação dos dados, a verosimilhança estima os parâmetros de interesse com base nos dados e a distribuição à posteriori, através do Teorema de Bayes, armazena toda informação a respeito do parâmetro.

Dentre os critérios a serem utilizados para verificação da adequação do modelo podemos destacar, o Critério de Informação de Akaike e o desvio residual utilizado para aferir o quão bom o modelo se ajusta aos dados e mediante o estudo residual, verificar por meio do envelope simulado a adequação do modelo.

O objetivo do presente trabalho é investigar quais fatores estão associados a prematuridade e por meio da análise do modelo de regressão clássico e Bayesiano, ajustar os modelos de regressão logístico via modelos lineares generalizados considerando a distribuição binomial com ligação *logit*, verificando a adequação do modelo com dados de Prematuridade.

## 2 Revisão de Literatura

### 2.1 SUS

A implementação do Sistema Único de Saúde (SUS) é considerado um dos maiores movimentos de inclusão social conquistada pela sociedade brasileira, nas últimas décadas, e criado desde a Constituição em 1988, para garantir o direito ao atendimento público de saúde a população do país (AGUIAR, 2011).

O sistema de saúde trouxe muitos avanços, para a sociedade brasileira, visando ofertar uma assistência de melhor qualidade na saúde da criança e da mulher, com diversos programas de controle e prevenção a doenças, projetos, serviços especializados na área ginecológica e obstetrícia e das políticas públicas para garantir os direitos de todos os cidadãos, entretanto ainda enfrenta muitos desafios, de promover um modelo de atenção integral a saúde para toda população (SOUZA; COSTA, 2010).

A assistência pré-natal abrange uma série de cuidados, portanto é um dos maiores desafios, prestar um atendimento de qualidade, na atenção a mulher grávida, uma vez que, na ausência ou baixa qualidade no atendimento, pode estar associada as altas taxas de mortalidade neonatal e baixo peso ao nascer. Por isso a necessidade de uma boa assistência para identificar possíveis doenças durante a gestação e também cuidados recebidos durante o parto, pois certas complicações pode colocar em risco a saúde da mulher e da criança. Espera-se que toda criança tenha direito a uma assistência de qualidade do nascimento ao desenvolvimento, visto que, foi fundamental a elaboração das leis e políticas públicas, para atenderam as mulheres desde aquelas que planejam a gestação, como também as que se recuperam no pós-parto (NETO et al., 2008).

### 2.2 Prematuridade em recém nascidos

A prematuridade é considerada um dos maiores problemas de saúde pública e está ligada aos altos índices de mortalidade neonatal, mortalidade infantil e na vida adulta. Considera-se prematuridade, os recém nascidos vivos, com idade gestacional inferior a 37 semanas (SALGE et al., 2009).

Prematuridade é classificada segundo a idade gestacional: sendo o prematuro limítrofe aqueles bebês que nascem entre 37 e 38 semanas; os moderados estão entre 31 e 36 semanas e os prematuros extremos aqueles nascidos entre 24 e 30 semanas (BOTÊLHO et al., 2012).

As mais diversas causas podem estar associadas a prematuridade, envolvendo uma relação mútua entre os fatores fetais, uterinos, placentários e maternos (KLIEGMAN et

al., 2014). Entre as causas maternas mais comuns estão a infecção urinária, pressão alta, diabetes, descolamento precoce de placenta, a idade materna (abaixo de 16 anos e acima de 35 anos), alterações de tireoide, consumo excessivo de bebidas alcoólicas e drogas, já as causas dos bebês são as malformações e as síndrome genética. Entretanto, há causa desconhecida (CAB, 2014).

Podemos caracterizar os recém nascidos prematuros, aqueles com baixo peso ao nascer, normalmente inferior a 2,5kg, mas por vezes inferiores a este valor, podendo chegar menos de 1kg, aquele com pele fina, brilhante e rosada, sendo possível observar as veias subjacentes, pouco cabelo, cabeça proporcionalmente maior em relação ao corpo, musculatura mais relaxada e poucos reflexos de sugar e deglutir (DELLAQUA; CARDOSO, 2012).

Entre as complicações mais comuns da prematuridade incluem as dificuldades respiratórias, problemas cardíacos, paralisia cerebral, problemas de visão, surdez, anemia, refluxo e infecções no intestino, contudo podendo trazer prejuízos a curto e longo prazo (LEONE et al., 2012).

A avaliação da atenção do acompanhamento pré-natal é um excelente instrumento para prevenção de possíveis nascimentos prematuros e mortalidade tanto materna como neonatal, controlando diversos fatores de riscos envolvidos na gravidez. Realizar um acompanhamento médico pré-natal de qualidade é fundamental para detecção de fatores de riscos, minimizando assim possíveis complicações.

Com os avanços obtidos, no decorrer dos anos, podemos observar que houve uma melhora na assistência ao recém nascido prematuro, contribuindo assim, para o crescimento, desenvolvimento e sobrevivência de bebês ainda mais prematuros.

## 2.3 ISEA

O Instituto de Saúde Elpídio de Almeida (ISEA), foi fundado em 05 de agosto de 1951 e posteriormente ficou conhecido como Maternidade Elpídio de Almeida, em homenagem ao prefeito Dr. Elpídio de Almeida. Mas somente em 27 de abril de 1992, após a reforma, a então maternidade passou a se chamar Instituto de Saúde Elpídio de Almeida (ISEA) (RETALHOS HISTÓRICOS DE CAMPINA GRANDE, 2011).

O ISEA é uma maternidade pública municipal, que atende não só a comunidade de Campina Grande, como também as cidades vizinhas e oferece serviços além de partos, procedimentos como, cirurgias, pré-natal para casos de alto risco, teste da Orelhinha, Pezinho, Linguinha, posto de vacinação e a casa da gestante, que acolhe as mães e bebês.

## 2.4 MLG

A importância dos Modelos lineares Generalizados (MLG) não é apenas de ordem prática. Do ponto de vista teórico, a sua importância advém, essencialmente, do fato de que a metodologia destes modelos constituem uma abordagem unificada de muitos procedimentos estatísticos correntemente usados nas aplicações (TURKMAN; SILVA, 2000). Nelder e Wedderburn (1972), propuseram os MLG's como uma extensão dos modelos lineares de regressão simples e múltipla, para dados não normalmente distribuídos, em que se teria mais opções para a distribuição da variável resposta pertencer à família exponencial de distribuições, além de dar maior flexibilidade para a relação funcional entre a média da variável resposta univariada com o preditor linear ( $\eta$ ).

Considere  $Y = \beta_0 + \beta_1 X + \epsilon$  como a variável resposta ou dependente do modelo de interesse do experimento associada a um conjunto de variáveis explicativas  $x_1, x_2, \dots, x_p$ . McCullagh e Nelder (1989) definem os três elementos que compõem o modelo linear generalizado:

- Componente Aleatório (variável resposta): É representado por um conjunto de variáveis aleatórias independentes  $Y_1, \dots, Y_n$  e que seguem uma mesma distribuição pertencente a família exponencial de distribuições com médias  $\mu_1, \dots, \mu_n$ . Portanto, a função de densidade de probabilidade  $Y_i$  é dada por

$$f(y_i; \theta_i, \phi) = \exp\{\phi[y\theta_i - b(\theta_i)] + c(y_i, \phi)\},$$

sendo  $y$  a variável resposta,  $\theta_i$  o parâmetro canônico,  $\phi > 0$  o parâmetro de dispersão conhecido e  $b(\theta_i)$  e  $c(y_i, \phi)$  funções específicas.

- Componente Sistemático ou Estrutural: As variáveis explicativas entram na forma de um modelo linear

$$\eta = X\beta,$$

sendo  $X = (x_1, \dots, x_n)^t$  a matriz do modelo,  $\beta = (\beta_1, \dots, \beta_n)^t$  o vetor dos parâmetros,  $\eta$  o preditor linear e  $c(y_i, \phi)$  é conhecido.

- Função de ligação: Responsável por realizar a ligação entre os componentes aleatório (variável resposta) e explicativa (variáveis explanatórias).

Na Tabela 1 são apresentadas as principais distribuições pertencentes da família exponencial que a variável resposta  $Y$  pode seguir.

Tabela 1 – Características de algumas distribuições pertencentes à família exponencial

Distribuição	$\phi$	$\theta$	$b(\theta)$	$c(y, \phi)$	$V(\mu)$
Normal: $N(\mu, \sigma^2)$	$\sigma^{-2}$	$\mu$	$\frac{\sigma^2}{2}$	$\frac{1}{2} \left( \frac{\log \phi}{2\pi} - \frac{\phi y^2}{2} \right)$	1
Poisson: $P(\mu)$	1	$\log \mu$	$e^\theta$	$-\log y!$	$\mu$
Binomial: $B(n, \mu)$	n	$\log \left( \frac{\mu}{1-\mu} \right)$	$\log[1 + e^\theta]$	$\log \left( \frac{\phi}{\phi y} \right)$	$\mu(1 - \mu)$
Gama: $G(\mu, \phi)$	1	$-\frac{1}{\mu}$	$-\log(-\theta)$	$(\phi - 1) \log(y_i) +$ $\phi \log(\phi) - \log[\Gamma(\phi)]$	$\mu^2$
Normal Inversa: $IG(\mu, \phi)$	$\phi$	$-\frac{1}{2\mu^2}$	$-(-2\theta)^{\frac{1}{2}}$	$\frac{1}{2} \left( \log \left( \frac{\phi}{2\pi y^3} \right) \right) - \frac{\phi}{2y}$	$\mu^3$

Fonte: Paula (2004)

## 2.5 Regressão Logística

A regressão logística é uma ferramenta analítica, baseada nos princípios de regressão múltipla e, procura prever a relação existente entre a variável resposta com as variáveis explanatórias, sendo bastante utilizada para estimar associações por meio da medida de razão de chance.

A técnica de regressão logística é amplamente difundida não somente nas áreas econômica, como também na área médica, seu êxito se dá sobretudo nas numerosas ferramentas que permitem interpretar de modo aprofundado os resultados obtidos.

A variável resposta tem distribuição Bernoulli, são considerados valores do tipo sucesso ou fracasso e com  $n$  ensaios independentes desta distribuição, chega-se a distribuição Binomial. De um modo geral, o modelo de regressão logística, está baseado na transformação *logit* para proporções.

Inicialmente vamos considerar, um modelo logístico simples, tal que  $\pi(x)$  é a probabilidade de “sucesso” sendo  $x$  da variável explicativa qualquer. O modelo logístico é definido por:

$$\log \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\} = \beta_0 + \beta_1 x,$$

em que  $\beta_0$  e  $\beta_1$  são parâmetros desconhecidos. Aplicando exponencial em ambos os membros, temos:

$$\frac{\pi(x)}{1 - \pi(x)} = e^{\beta_0 + \beta_1 x},$$

em que  $e^{\beta_0}$  é o risco basal em escala exponencial e  $e^{\beta_1 x}$  será o tamanho do efeito de risco associado à variável  $x$ , em escala exponencial.

Como exemplo, tem-se uma associação entre determinada doença e a ocorrência ou não de um fator particular, com amostras independentes, teremos  $n_1$  indivíduos com presença do fator ( $x = 1$ ), e  $n_2$  indivíduos na ausência do fator ( $x = 0$ ),  $\pi(x)$  a probabilidade da doença desenvolver após fixar um certo período. Assim, a chance de desenvolvimento da doença para certo indivíduo com a presença de fator será a seguinte:



$$\frac{\pi(1)}{1 - \pi(1)} = e^{\beta_0 + \beta_1},$$

enquanto, a chance de desenvolvimento da doença para certo indivíduo na ausência de fator será:

$$\frac{\pi(0)}{1 - \pi(0)} = e^{\beta_0}.$$

Logo, a razão de chances será:

$$\psi = \frac{\pi(1)\{1 - \pi(0)\}}{\pi(0)\{1 - \pi(1)\}} = e^{\beta_1}.$$

Agora, utilizando com dois estratos  $x_1$  ( $x_1 = 0$ , estrato 1) e  $x_2$  ( $x_2 = 1$ , estrato 2), sendo para estratos 1,  $n_{11}$  indivíduos na presença e  $n_{21}$  indivíduos na ausência e,  $n_{12}$  e  $n_{22}$  respectivamente, do estrato 2. Para o desenvolvimento da doença, a probabilidade será  $\pi(x_1, x_2)$ , com  $x_2 = 1$  presença do fator e  $x_2 = 0$  ausência do fator. Logo, teremos quatro parâmetros para estimar,  $\pi(0, 0), \pi(0, 1), \pi(1, 0), \pi(1, 1)$ . Os dados estão apresentados na Tabela 2, para um estudo comparativo de coorte ou um estudo caso-controle.

Tabela 2 – Notação para estudo de coorte ou caso-controle

		Estratos		
		1	2	
<b>Doença</b>	Presença	$n_{11}$	$n_{12}$	$n_{1n}$
	Ausência	$n_{21}$	$n_{22}$	$n_{2n}$
		$n_{n1}$	$n_{n2}$	$n_{nn}$

Por consequência, qualquer reparametrização deverá ter no máximo quatro parâmetros. Consideramos a seguinte reparametrização:

$$\log \left\{ \frac{\pi(x_1, x_2)}{1 - \pi(x_1, x_2)} \right\} = \beta_0 + \gamma x_1 + \beta x_2 + \delta x_1 x_2,$$

em que  $\gamma$  o efeito do estrato,  $\beta$  o efeito do fator e  $\delta$  a interação entre fator e o estrato. Assim, a razão de chances para cada estrato será:

$$\psi_1 = \frac{\pi(0, 1)\{1 - \pi(0, 0)\}}{\pi(0, 0)\{1 - \pi(0, 1)\}} = e^{\beta}$$

e

$$\psi_2 = \frac{\pi(1, 1)\{1 - \pi(1, 0)\}}{\pi(1, 0)\{1 - \pi(1, 1)\}} = e^{\beta + \delta}.$$

De acordo com Paula (2004), considera-se um modelo de regressão logística múltipla, nos casos em que utiliza-se mais de uma variável explicativa para ajustar o modelo, seguindo a equação geral:

$$\log \left\{ \frac{\pi(x)}{1 - \pi(x)} \right\} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p,$$

em que  $X = (1, x_1, \dots, x_p)^T$  assume valores de variáveis explicativas.

## 2.6 Inferência Bayesiana

Segundo Kinas e Andrade (2017), a informação que dispomos sobre  $\theta$ , resumida probabilisticamente através de  $p(\theta)$ , pode ser aumentada, observando-se uma quantidade aleatória  $X$  relacionada com  $\theta$ . Dado que, a distribuição amostral  $p(x|\theta)$ , define esta relação. É evidente a teoria de que após analisar  $X=x$ , a quantidade de informação sobre  $\theta$  aumenta. A fórmula de Bayes é definida da seguinte forma:

$$p(\theta|x) = \frac{p(\theta, x)}{p(x)} = \frac{p(x|\theta)p(\theta)}{p(x)} = \frac{p(x|\theta)p(\theta)}{\int p(\theta, x)d\theta}.$$

Como  $1/p(x)$  não depende de  $\theta$  funciona como uma constante normalizadora de  $p(\theta|x)$ . Para um valor fixo de  $x$ , a distribuição observacional de uma amostra fornece a função de verossimilhança  $L(\theta; x) = p(x|\theta)$  de cada um dos prováveis valores de  $\theta$ , enquanto  $p(\theta)$  é chamada distribuição a priori de  $\theta$ . Estas duas informações, a priori e verossimilhança, são combinadas e levam à distribuição a posteriori de  $\theta$ ,  $p(\theta|x)$ . Portanto, a distribuição a posteriori pode ser escrita pela seguinte forma  $p(\theta|x) \propto L(\theta; x)p(\theta)$  (KINAS; ANDRADE, 2017).

De acordo com Kinas e Andrade (2017), a informação que se tem acerca de uma quantidade  $\theta$  é desconhecida, portanto a ideia é reduzir esse valor e quantificar os graus de incerteza a respeito de  $\theta$  representados através do modelo probabilístico para  $\theta$ .

### 2.6.1 MLG Bayesiano

#### 2.6.1.1 MLG

Para Kinas e Andrade (2017), os modelos lineares são estruturados como uma função linear da variável resposta  $y$ , com média condicionada às covariáveis  $x$  e com desvio constante  $\sigma$ . Não satisfazendo a esses pressupostos é necessário há deve-se procurar alternativas. Em certos casos, basta a transformação da variável  $y$ . Nos modelos lineares generalizados (MLG's) é essencial que a variável resposta siga uma distribuição de probabilidade pertencente à família exponencial, que inclua como um caso específico à distribuição normal, envolvendo algumas distribuições contínuas e discretas. A equação geral do modelo é definida em três partes:

$$y_i \sim p(y|\theta)$$

tal que

$$E(y_i) = \mu_i$$

$$\eta_i = g(\mu_i)$$

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

Na primeira equação,  $y_i$  segue um modelo probabilístico  $p(y|\theta)$  pertencente a família exponencial de distribuições e média  $\mu_i$ . Na segunda equação a função de ligação é conhecida  $g(\cdot)$ , as médias  $\mu_i$  se relacionam com os referentes valores transformados  $\eta_i$ . A terceira equação determina a relação linear entre  $\eta_i$  e a covariável  $x_i$  (KINAS; ANDRADE, 2017).

Kinas e Andrade (2017) descreve a equação do modelo linear simples como um caso particular do modelo linear generalizado:

$$\begin{aligned}y_i &\sim N(\mu_i|\sigma) \\ \eta_i &= g(\mu_i) = \mu_i \\ \eta_i &= \beta_0 + \beta_1 x_i\end{aligned}$$

Na primeira equação a distribuição normal pertence a família exponencial de distribuição, validando a equação. Na segunda equação, a função de ligação pode ser a de identidade. Na terceira equação refere-se ao componente determinístico (KINAS; ANDRADE, 2017).

### 2.6.1.2 Regressão logística

De acordo com Kinas e Andrade (2017), os modelos lineares generalizados são conhecidos por regressão logística, em situações que a variável resposta é discreta e que segue distribuição binomial. As equações que definem o modelo:

$$\begin{aligned}y_i &\sim Bin(n_i, \theta_i) \\ \eta_i &= g(\theta_i) = \log\left(\frac{\theta_i}{1 - \theta_i}\right) \\ \eta_i &= \beta_0 + \beta_1 x_i\end{aligned}$$

Kinas e Andrade (2017) determina a função de ligação acima, de função *logit* e transforma o parâmetro  $\theta$ , limitado ao intervalo  $[0,1]$ , para  $\eta$  definido em  $(-\infty, +\infty)$ .

## 2.6.2 OpenBUGS

O *OpenBUGS* é um programa para realização de análise bayesiana de modelos estatísticos complexos, utilizando-se as técnicas de Monte Carlo via Cadeia de Markov (MCMC). O *OpenBUGS* possibilita que os modelos sejam explicado empregando linguagem *BUGS*, a mais flexível ou *Doodles*, a representação gráfica.

Segundo Kinas e Andrade (2017), o analista pode-se executar o método MCMC, criando seu próprio código, no entanto possuem métodos já implementados em programas

computacionais. Por exemplo, podemos descrever o *OpenBUGS* utilizado como aplicativo independente, como também integrado ao software estatístico *R* (KINAS; ANDRADE, 2017).

## 2.7 Seleção de Modelos

Após ser definido o conjunto de covariáveis a serem incluídas no modelo logístico, deve-se encontrar a melhor maneira de incluir apenas covariáveis e interações mais importantes em um modelo reduzido para explicar a probabilidade de sucesso  $\pi(x)$ . Os métodos usuais de seleção de modelos que podem resolver os problemas envolvendo modelos logísticos são os métodos Stepwise (Forward, Backward, Bidirecional) e os Critérios de Coeficiente de Determinação, Critério de Informação de Akaike e Critério de Informação de Desvio.

### 2.7.1 Coeficiente de Determinação ( $R^2$ )

O coeficiente de determinação, também chamado de  $R^2$ , é uma medida de ajustamento de um modelo estatístico linear generalizado, como a regressão linear, em relação aos valores observados, quantifica a proporção da variabilidade da variável resposta que é explicada por um modelo de aproximação, dada por:

$$R^2 = 1 - \frac{SQRes}{SQTotal}.$$

O  $R^2$  pode assumir valores no intervalo  $[0,1]$ , indicando, o quanto o modelo consegue explicar os valores observados, sendo que valores próximos de 1, dá indícios de uma boa relação entre a variável resposta e as  $p$  variáveis preditoras. Já, valores próximos ou iguais a 0 (zero), indicam que o modelo não é superior a média amostral (DRAPER; SMITH, 1998). De acordo com Rencher e Schaalje (2008), há um aumento do coeficiente de determinação com a inclusão de variáveis regressoras, fazendo com que o erro diminua com a soma de quadrados.

### 2.7.2 Critério de Informação de Akaike - AIC

O Critério de Informação de Akaike (AIC) é baseado na teoria de informação, à qual foi desenvolvido por (AKAIKE, 1974). Para os casos de estimação por mínimos quadrados, o valor do AIC é simples de ser obtido, de uma forma geral, para casos de análises baseadas na estimativas de verossimilhança, e do ajuste de modelo de regressão (BURNHAM; ANDERSON, 2004).

Sua fórmula é definida por:

$$AIC = -2 \log L(\theta) + 2n,$$

em que  $L(\theta) = f(x_n | \theta)$  é a função de verossimilhança e  $n$  número de parâmetros ajustados.

### 2.7.3 Critério de Informação Bayesiano - BIC

O Critério de Informação Bayesiano (BIC), proposto por (SCHWARZ et al., 1978), consiste em uma estatística utilizada na comparação de modelos, dado por:

$$BIC = -2 \log L(\theta) + p \log n,$$

em que  $L(\theta) = f(x_n | \theta)$  é o modelo escolhido,  $p$  o número de parâmetros a serem estimados e  $n$  número de observações da amostra. O menor valor BIC, é considerado o melhor ajuste no modelo.

### 2.7.4 Critério de Informação de Desvio - DIC

O Critério de Informação de Desvio (DIC) é composto pela média a posteriori do desvio penalizado pelo número de parâmetros do modelo. Este critério é particularmente usual nos problemas Bayesianos de seleção de modelos para os quais amostras da distribuição a posteriori dos parâmetros dos modelos foram obtidas por simulação de Monte Carlo em Cadeias de Markov (MCMC). Semelhante aos outros critérios é uma aproximação assintótica para grandes amostras e é válido quando a distribuição a posteriori é aproximadamente uma distribuição normal multivariada (ARRABAL et al., 2012). Quanto menor for o valor para o DIC, melhor será seu ajuste. O DIC é dado por:

$$DIC = \bar{D}(\beta, M_i) + p_{di},$$

sendo,  $p_{di} = \bar{D}(\beta, M_i) - D(\bar{\beta}, M_i)$  mede a complexidade do modelo  $i$ . O critério sugere uma comparação entre o desvio médio e o desvio aplicado na média a posteriori.

### 2.7.5 Stepwise (Forward, Backward, Bidirecional)

Consiste na junção dos métodos *forward* e *backward* e inicia-se com o modelo simples  $\mu = \beta_0$ . Após a inclusão de duas variáveis no modelo, verifica-se se a primeira permanece no modelo, o processo continua até que nenhuma variável seja incluída ou excluída do modelo. Geralmente adota-se  $p_E = p_S = 0, 20$  como escolha de entrada e saída de variáveis, respectivamente, (PAULA, 2004).

### 2.7.5.1 Forward

O método *forward* inicia-se pelo modelo  $\mu = \beta_0$  e pressupondo que para cada uma, seja consideradas  $q$  variáveis explicativas, ajusta se o modelo

$$\mu = \beta_0 + \beta_i x_j, \quad j = 1, \dots, q.$$

As hipóteses a serem testadas serão  $H_0 : \beta_j = 0$  vs  $H_1 : \beta_j \neq 0$ . Sendo  $p$  o menor nível descritivo entre os  $q$  testes realizados. Se caso  $p \leq p_E$ , então a variável correspondente entra no modelo, sendo  $p_E$  um nível descritivo crítico, usado como critério de entrada. Portanto, uma vez selecionada a variável ao modelo, está não será mais descartada, até que  $p > p_E$  ocorra (PAULA, 2004).

### 2.7.5.2 Backward

O método *backward*, inicia-se a partir do modelo completo, ou seja, serão incluídas todas as variáveis explicativas consideradas

$$\mu = \beta_0 + \beta_1 x_1 + \dots + \beta_q x_q,$$

Testa-se as hipóteses  $H_0 : \beta_j = 0$  vs  $H_1 : \beta_j \neq 0$ , com  $j = 0, \dots, q$ . Sendo  $p$  o maior nível descritivo entre os  $q$  testes. Nos casos que  $p > p_S$ , descarta-se a correspondente variável do modelo, em que  $p_S$  é o nível descritivo, escolhido como critério de descarte realizados, de forma que nenhuma variável excluída seja reconsiderada, até que não ocorra descarte, verificando quando ocorra  $p \leq p_S$ .

### 2.7.5.3 Bidirecional

Giolo e Colosimo (2006) utilizam o método bidirecional em análise de sobrevivência considerando o tempo até o desmame de bebês humanos considerando diversas covariáveis e uma distribuição gama generalizada para o tempo. O critério de seleção de covariáveis foi baseado na medida de verossimilhança.

Para Giolo e Colosimo (2006), o método bidirecional, é realizado em alguns passos. Primeiramente deve-se ajustar todos os modelos com uma única covariável e incluir todas que forem significativas ao nível de 0,10%. Na presença de outras covariáveis, elas podem deixar de serem significativas, não permanecendo no modelo. Após ajustar um novo modelo, depois da retirada das covariáveis não significativas, as covariáveis voltam para confirmar se realmente não são significativas. Essas covariáveis são incluídas com as do passo anterior, e as covariáveis excluídas no primeiro passo, voltam pra confirmar se realmente não são significativas. No quinto passo, ajusta-se o modelo com as covariáveis do quarto passo, e verifica se alguma pode ser retirada. Por último, ajusta-se o modelo com as covariáveis do quinto passo e verifica a possibilidade de inclusão de termos interação, testando cada interação duas a duas possíveis entre as covariáveis inclusas no modelo.

## 2.8 Matriz de confusão

A matriz de confusão exhibe o número de classificações corretas contra as classificações preditas para cada classe, sobre conjuntos de exemplos (GODBOLE, 2002).

A Tabela 3 apresenta o esquema de uma matriz de erro para duas classes. A é o valor predito falso e que realmente é falso. B é o valor predito verdadeiro, dado que é falso. C é o valor predito falso, dado que é verdadeiro. D é o valor predito verdadeiro e que realmente é verdadeiro.

Tabela 3 – Matriz de confusão

		Valores Preditos	
		Falso	Verdadeiro
Valores Reais	Falso	a	b
	Verdadeiro	c	d

Está localizado na diagonal principal da matriz, o número de acertos, para cada classe, também chamado de positivo verdadeiro e os demais elementos, que representam os erros na classificação, chamado de falso positivo. A matriz de confusão de um classificador ideal, possui fora da diagonal principal, todos os elementos iguais a zero (KIL; SHIN, 1998; OLSON; DELEN, 2008). Para o número de classes classificadas corretamente como negativa, chamamos de verdadeiro negativo e para o número de classes classificadas como negativas, chamamos falso negativo. A acurácia  $A_c$  estima o número de classes que foram classificadas corretamente divididas pelo conjunto de todas as classes. A acurácia é dada por  $A_C = \frac{a+d}{a+b+c+d}$ . A Precisão (P) estima o número de classes que foram classificadas corretamente como positiva dividida por todas as classes classificadas como positiva. A precisão é dada por  $P = \frac{d}{b+d}$ . E a Verdadeira Positiva (TP) estima o número de classes positivas que foram classificadas corretamente dividida pelas classes que são positivas. A verdadeira positiva é dada por  $TP = \frac{d}{c+d}$ . O falso positivo (FP) estima a quantidade de casos negativos que foram incorretamente identificados como positivo, FP dada por  $FP = \frac{b}{c+d}$ . A negativa verdadeira (TN) proporção de casos negativos, classificado corretamente. Dada por  $TN = \frac{a}{a+b}$  e Falso negativo FN definida como positivo que foram classificadas como negativo, dado por:  $FN = \frac{c}{c+d}$ .

## 3 Material e Métodos

A coleta de dados foi realizada no Instituto de Saúde Elpídio de Almeida, abrangendo todas as gestantes assistidas na maternidade. Foram consideradas apenas informações de 4257 mulheres, após utilização de técnica de mineração, devido a ausência de informações preenchidas nos formulários.

### 3.1 Estrutura dos dados

Os pacientes em estudo, foram avaliados segundo as covariáveis socioeconômica e obstétrica: idade da mãe, situação conjugal, raça/cor, grau de escolaridade, número de gestações, filhos nascidos vivos, número de consultas pré-natal, tipo de gravidez, sexo e peso das crianças ao nascer.

### 3.2 Modelo Empregado

Para o modelo logístico, apenas dois resultados podem ser empregados, na variável resposta. Sendo um desses conhecido de "premature", que corresponde ao resultado que se pretende chegar e o outro "não premature". No entanto, podemos caracterizar  $Y_i$  igual a 1, se caso ocorre prematuridade, ou igual a 0 para não prematuridade. Assim, aplicando-se a técnica *stepwise*, implementados no *software R* (R Core Team, 2018) é definido o conjunto de covariáveis inseridos no modelo linear generalizado, considerando a distribuição binomial e função de ligação *logit*, para descrever quais variáveis estão relacionados com a incidência de prematuridade. Após verificar o melhor modelo, faz-se uma análise, aplicando-se a Inferência clássica e Bayesiana.



## 4 Resultados e Discussão

Nesta seção serão apresentadas as análises descritivas das variáveis em estudo, seguido do resultado do ajuste do modelo Binomial que melhor explique a Prematuridade das crianças.

De acordo com a análise dos dados, observa-se que do total geral de 4257 prontuários utilizados, apenas 15,31% das mulheres tiveram filhos prematuros, enquanto 84,69% não possuíram.

Na tabela 4 são apresentadas as distribuições de frequência de prematuridade em função das características sócio-demográficas da mãe. Observa-se que a frequência da prematuridade segundo a situação conjugal, houve um predomínio no grupo de solteiras registrou 14,02% (342/2439), seguido das casadas 17,45% (182/1043), para as mães com união estável, o percentual foi de 16,87% (126/747), para as viúvas registrou-se 16,67% (2/12) e para as divorciadas não houve registros de nascimento de filhos prematuros.

Em relação a idade materna na prematuridade, os números variam entre 13 a 50 anos, sendo estes registros agrupados em três faixas etárias como observa-se na tabela 4. Para a classe da faixa etária entre 12 a 24, predominou 14,18% (314/2169), seguindo de 15,78% (295/1869) entre 25 a 37 anos e por último entre 38 a 50 com percentual de 19,63% (43/219).

No que se diz respeito a prevalência de raça/cor das mães na prematuridade, 15,02% (591/3934) das mulheres foram declaradas pardas, 20,94% (58/277) corresponde a cor branca, o percentual da cor preta foram apenas 4,44% (2/45) e para a raça indígena 100% (1/1).

De acordo com grau de escolaridade das mães em relação a prematuridade, 16,45% (264/1605) das mães cursaram até o Ensino Médio, seguido de 14% (241/1721) para aquelas que cursaram até o Ensino Fundamental II, 12,9% (92/713) concluíram até o Ensino Fundamental I, 30,11% (28/93) possuem Ensino Superior Completo, 25,31% (20/79) Ensino Superior Incompleto e 15,22% (7/46) não foram alfabetizadas Tabela 4.

Para a análise dos dados foi utilizado o teste qui-quadrado, para verificar se cada variável independente parece está associada à prematuridade. Segundo, os resultados obtidos na Tabela 4, podemos verificar quanto a situação conjugal, que rejeita-se a hipótese a 0,05% de significância, com um  $p = 0,0254$ , ou seja, interfere na prematuridade. Com relação a variável idade da mãe, com valor  $p = 0,2517$ , a 0,05% de significância, não há indícios que a chance de uma criança nascer prematura tenha dependência com a idade da mãe. Para a variável raça/cor, o teste aponta um  $p < 0,001$ , afirmando que há indícios para rejeitar a hipótese a 0,05% de significância, sendo assim, a variável interfere no nascimento de crianças prematuras. O grau de instrução, com valor  $p < 0,001$ , rejeita-se a hipótese a

0,05% de significância, de que a variável interfira na prematuridade.

Tabela 4 – Distribuição das características sócio-demográfico das mães relacionado à prematuridade assistido no ISEA

	Prematuridade		Total N	Teste Qui-Quadrado Valor p
	Sim Frequência (%)	Não Frequência(%)		
<b>Situação Conjugal</b>				0,0254
Solteira	342 (14,02)	2097 (85,98)	2439	
Casada	182 (17,45)	861 (82,55)	1043	
Viúva	2 (16,67)	10 (83,33)	12	
Divorciada	0 (0)	16 (100)	16	
União Estável	126 (16,87)	621 (83,13)	747	
<b>Idade da mãe (anos)</b>				0,2517
12-24	314 (14,48)	1855 (85,52)	2169	
25-37	295 (15,78)	1574 (84,22)	1869	
38-50	43 (19,63)	176 (80,37)	219	
<b>Raça/cor</b>				<0,001
Branca	58 (20,94)	219 (79,06)	277	
Preta	2 (4,44)	43 (95,56)	45	
Parda	591 (15,02)	3343 (84,98)	3934	
Indígena	1 (100)	0 (0)	1	
<b>Grau de instrução</b>				<0,001
Não alfabetizada	7 (15,22)	39 (84,78)	46	
Ensino fundamental I	92 (12,9)	621 (87,1)	713	
Ensino fundamental II	241 (14)	1480 (86)	1721	
Ensino médio	264 (16,45)	1341 (83,55)	1605	
Ensino superior Incompleto	20 (25,31)	59 (74,69)	79	
Ensino superior completo	28 (30,11)	65 (69,89)	93	

Com base nas características obstétricas, das mães com filhos prematuros, 16,09% (287/1784) não tiveram gestações anteriores, seguido de 16,58% (186/1122) com apenas uma gestação. Segundo o número de filhos vivos prematuro anteriores, de 17,05% (337/1977) não possuíram filhos, com 14,87% (177/1189) com apenas um filho vivo e 12,95% (75/579) com dois filhos nascidos vivos. Em relação ao número de consultas, das mães que tiveram filhos prematuros a porcentagem 20,95% (426/2033) realizaram entre 1 a 6 consultas e 9,44% (194/2056) representa entre 7 a 12 consultas e 20,71% (29/140) não realizou nenhuma consulta pré-natal, ou seja, quanto maior o número de consultas, menor será o nascimento de bebês prematuros Tabela 5.

Segundo o tipo de gestação, constatou que a prematuridade foi prevalente na gestação única com 13,84% (568/4105), enquanto os relatos na gestação dupla verificou 54,54% (78/143) dos casos. Com relação ao sexo do recém-nascido, as proporções de prematuros foram de 15,46% (343/2218) para o sexo masculino e 15,15% (309/2039) para o sexo feminino. Ao analisar o peso dos bebês prematuros, observou-se que 41,77% (406/972) nasceram com peso entre 1,612 a 2,823kg, seguido de 93,98% (125/133) entre 0,400 a 1,611kg Tabela 5.

Para os resultados obtidos na Tabela 5, verifica-se que o número de gestações com um valor  $p = 0,3818$ , não se rejeita a hipótese, a 0,05% de significância, ou seja, não interfere na prematuridade. Para variável número de filhos vivos, com valor  $p = 0,1337$ , ao nível de 0,05% de significância, a taxa de prematuridade independe do número de filhos vivos. Quanto ao número de consultas pré-natal, pode-se afirmar que há indícios para rejeitar a hipótese, com valor  $p < 0,001$ , com 0,05% de significância de que esta variável interfira na chance da criança nascer prematura. De acordo com o teste, para a variável tipo de gravidez, com valor  $p < 0,001$ , há indícios para afirmar, 0,05% de significância, que a prematuridade tenha dependência com a variável. Em relação ao sexo da criança, o teste apontou valor  $p = 0,812$ , pode-se afirmar que não há indícios para rejeitar a hipótese de que esta variável interfira na prematuridade, a 0,05% de significância. O peso da criança, obteve valor  $p < 0,001$ , rejeitando a 0,05% de significância que a variável interfira diretamente na prematuridade.

Tabela 5 – Distribuição das características obstétricas relacionados à prematuridade assistido no ISEA

	Prematuridade			Teste Qui-Quadrado Valor p
	Sim Frequência (%)	Não Frequência(%)	Total N	
<b>Gestações</b>				0,3813
0	287 (16,09)	1497 (83,91)	1784	
1	186 (16,58)	936 (83,42)	1122	
2	80 (12,64)	553 (87,36)	633	
3	44 (13,46)	283 (86,54)	327	
4 ou mais	55 (14,07)	336 (85,93)	391	
<b>Filhos vivos</b>				0,1337
0	337 (17,05)	1640 (82,95)	1977	
1	177 (14,87)	1012 (85,13)	1189	
2	75 (12,95)	504 (87,05)	579	
3	29 (12,08)	211 (87,92)	240	
4 ou mais	34 (12,5)	238 (87,5)	272	
<b>Número de consultas pré-natais</b>				<0.001
Nenhuma	29 (20,71)	111 (79,29)	140	
1-6	426 (20,95)	1607 (79,05)	2033	
7-12	194 (9,44)	1862 (90,56)	2056	
13 ou mais	3 (10,71)	25 (89,29)	28	
<b>Tipos de Gravidez</b>				<0.001
Única	568 (13,84)	3537 (86,16)	4105	
Dupla	78 (54,54)	65 (45,46)	143	
Tripla	3 (33,33)	6 (66,67)	9	
<b>Sexo da criança</b>				0,812
Feminino	309 (15,15)	1730 (84,85)	2039	
Masculino	343 (15,46)	1875 (84,54)	2218	
<b>Peso da criança</b>				<0,001
0.400-1.611	125 (93,98)	8 (6,02)	133	
1.612-2.823	406 (41,77)	566 (58,23)	972	
2.824-4.035	116 (4,02)	2765 (95,98)	2886	
4.036-5.247	5 (1,88)	266 (98,12)	266	

Na Tabela 6, encontram-se um resumo das estatísticas descritivas de algumas covariáveis estudadas e através dessas medidas é possível analisar o comportamento dos dados. Verifica-se que os valores da mediana, estão todos próximos da média, indicando simetria. Quanto a idade materna, observa-se que as mães tem em média 26 anos, seguindo a idade mínima 12 anos e a máxima de 50 anos. Ao investigar o número de gestações, percebe-se que as mães tiveram em média 2 gestações, como também tiveram mães que chegaram um total de 14 gestações e há registro de mães que não tiveram filhos anteriores. Segundo a quantidade de filhos vivos, verificou-se em média, que as mães tiveram 2 (dois) filhos vivos, sendo que há registros de mães que chegaram a ter 13 filhos vivos e como também houve mães que não tiveram filhos vivos. Quanto ao número de consultas, em média foram realizadas 7 consultas durante o pré-natal, entretanto há registro de mães que não tiveram acompanhamento pré-natal. E por fim, o peso da criança ao nascer, que apresentou uma média de 3,140kg, seguido de um peso mínimo de 0,400kg e a máxima de 5,245kg.

Tabela 6 – Resumo das estatísticas descritivas de algumas das variáveis explicativas

Características	Mínimo	Primeiro Quartil	Mediana	Média	Terceiro Quartil	Máximo
Idade da mãe	12	20	24	25,12	30	50
Gestações	0	0	1	1,31	2	14
Filhos vivos	0	0	1	1,081	2	13
Número de consultas	0	5	6	6,28	8	37
Peso da criança	0,400	2,805	3,195	3,140	3,550	5,245

Inicialmente, para o estudo foram selecionadas 10 (dez) covariáveis para serem incluídas no modelo, afim de descrever o melhor comportamento da variável resposta, conforme Tabela 7.

Tabela 7 – Descrição das covariáveis utilizadas no estudo sobre a prematuridade

Código	Descrição
V1	Situação Conjugal
V2	Idade da mãe
V3	Raça/cor
V4	Grau de instrução
V5	Número de gestações
V6	Número de filhos vivos
V7	Número de consultas pré-natais
V8	Tipos de gravidez
V9	Sexo da criança
V10	Peso da criança

A regressão logística foi ajustada utilizando-se todas as covariáveis em estudo e através da técnica de *stepwise* bidirecional, implementadas pelo *software R* (R Core Team, 2018), verificou-se que, dentre as 10 (dez) covariáveis inseridas, apenas 5 (cinco) foram significativas, incluindo também o intercepto na Tabela 8. Para a escolha do modelo final, levou em consideração, os critérios com menor valor de AIC (Critério de Informação de Akaike), para concluir se realmente, esse foi o melhor modelo que se ajustou aos dados, aplicou-se o gráfico de envelope simulado para os resíduos.

Tabela 8 – Seleção de covariáveis considerando o modelo binomial

Código	AIC
V1	2294,8
V2	2317,3
V3	2311,8
V4	2284,9
V5	2317,8
V6	2320,8
V7	2327,6
V8	2316,1
V9	2319,1
V10	3452,7
V1+V2+V3+V5+V6+V7+V8+V9+V10	2258,48
V2+V3+V5+V6+V7+V8+V9+V10	2249,18
V2+V5+V6+V7+V8+V9+V10	2244,68
V2+V5+V6+V7+V9+V10	2239,81
V2+V6+V7+V9+V10	2237,16
V2+V6+V7+V10	2205,4
V2+V6+V7+V8+V10	2195,3

Pode-se verificar na Figura 1 que o gráfico de envelope simulado dos resíduos não apresenta indícios de que a distribuição utilizada para este modelo seja inadequada, ao nível de 95% de confiança, mesmo que embora apresente a proporção de 1,57%(67) dos pontos fora da banda de confiança. Logo, o envelope é bastante útil para verificar a qualidade do ajuste, sendo assim, o referido modelo é adequado à variável resposta em questão.

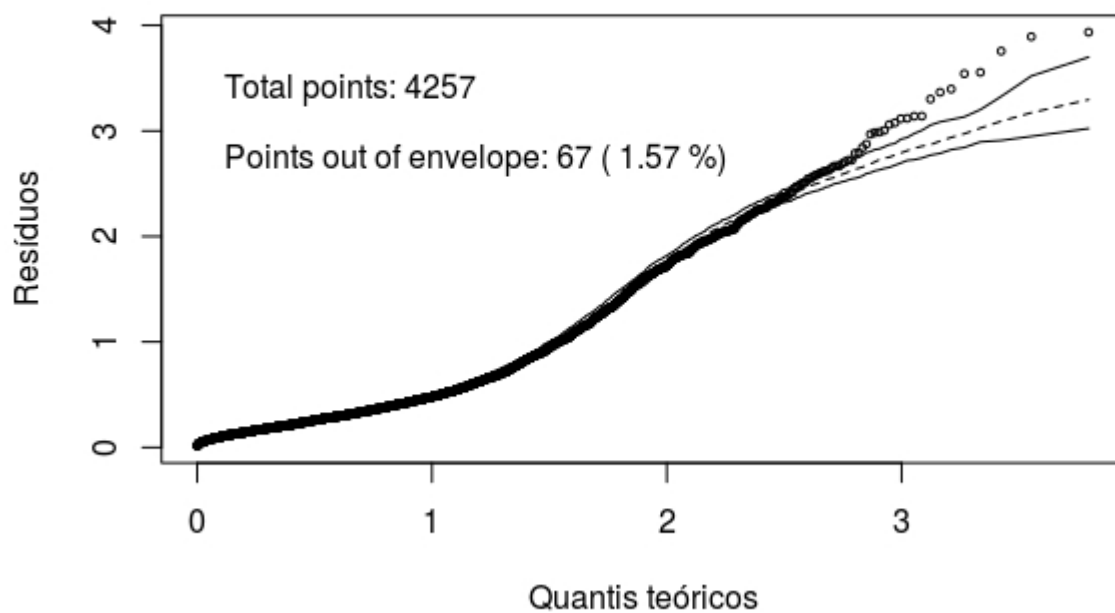


Figura 1 – Envelope simulado para os resíduos do modelo com distribuição binomial.

Observamos na Tabela 9 são apresentadas as estimativas dos parâmetros do modelo

Binomial, considerando a função de ligação *logit*.

A partir dos resultados obtidos, podemos verificar através do ajuste do modelo, quais fatores foram significativos para prever o nascimento de crianças prematuras, como mostra na Tabela 9. Sendo assim, podemos descrever o seguinte modelo:

$$\text{logit}(\hat{\pi}) = 6,83291 + 0,02855V_2 - 0,15332V_6 - 0,07608V_7 + 0,82205V_8 - 3,0630V_{10},$$

sendo o valor 6,83291, representado pelo intercepto,  $V_2$  idade da mãe,  $V_6$  número de filhos vivos,  $V_7$  número de consultas pré-natal,  $V_8$  tipos de gravidez e  $V_{10}$  peso da criança ao nascer. Com base na análise do modelo proposto, podemos inferir cinco fatores com maior relevância para determinar a prematuridade de uma criança. Sendo que, uma gestação tardia e/ou múltipla, aumentam as chances do feto nascer prematuro. Outros fatores externos podem ser determinantes para a prematuridade de uma criança, tais como: quantidade de consultas pré-natal realizadas, o número de filhos vivos e peso da criança ao nascer.

Este modelo registrou AIC=2195,3 e desvio residual de 2183,3 para 4251 graus de liberdade, evidenciando assim, que o modelo é válido, pois o desvio residual é menor que o grau de liberdade.

Tabela 9 – Estimativas dos parâmetros do modelo via análise clássica

Coefficientes	Estimativa	Erro padrão	Valor p
Intercepto	6,831964	0,402700	<0,001 ***
Idade da mãe	0,028501	0,009055	<0,001 **
Número de nascidos vivos	-0,151788	0,047159	<0,001 **
Número de consultas	-0,078162	0,022960	<0,001 ***
Tipo de gravidez	0,792209	0,226507	<0,001 ***
Peso da criança	-3,056865	0,123878	<0,001 ***

Códigos de significância: 0 '\*\*\*\*' 0,001 '\*\*\*' 0,01 '\*\*' 0,05 '' 0,1 ' ' 1

Observamos na Tabela 10 as estimativas dos parâmetros das abordagens Bayesianas. É possível verificar que as estimativas tanto da Clássica, quanto da Bayesiana foram muito próximos. Sendo o erro padrão da Clássica mais alta.

Tabela 10 – Estimativas dos parâmetros do modelo via análise bayesiana

Coefficientes	Estimativas	Erro Padrão	IC <sub>2.5%</sub>	IC <sub>50%</sub>	IC <sub>97.5%</sub>
Intercepto	6,83291	0,40670	6,03121	6,84275	7,62759
Idade da mãe	0,02855	0,00877	0,01168	0,02841	0,04456
Número de nascidos vivos	-0,15332	0,04637	-0,24781	-0,15019	-0,06786
Número de consultas	-0,07608	0,02246	-0,12218	-0,07495	-0,03461
Tipo de gravidez	0,82205	0,22089	0,38180	0,81533	1,24277
Peso da criança	-3,06306	0,12330	-3,31395	-3,05595	-2,81403

De acordo com a Figura 2 podemos observar que os parâmetros utilizados na cadeia convergiram, ou seja, não apresentaram nenhuma tendência, positiva ou negativa. Para o gráfico de densidade, percebemos que os parâmetros estão aleatoriamente em torno de uma média, indicando a convergência.

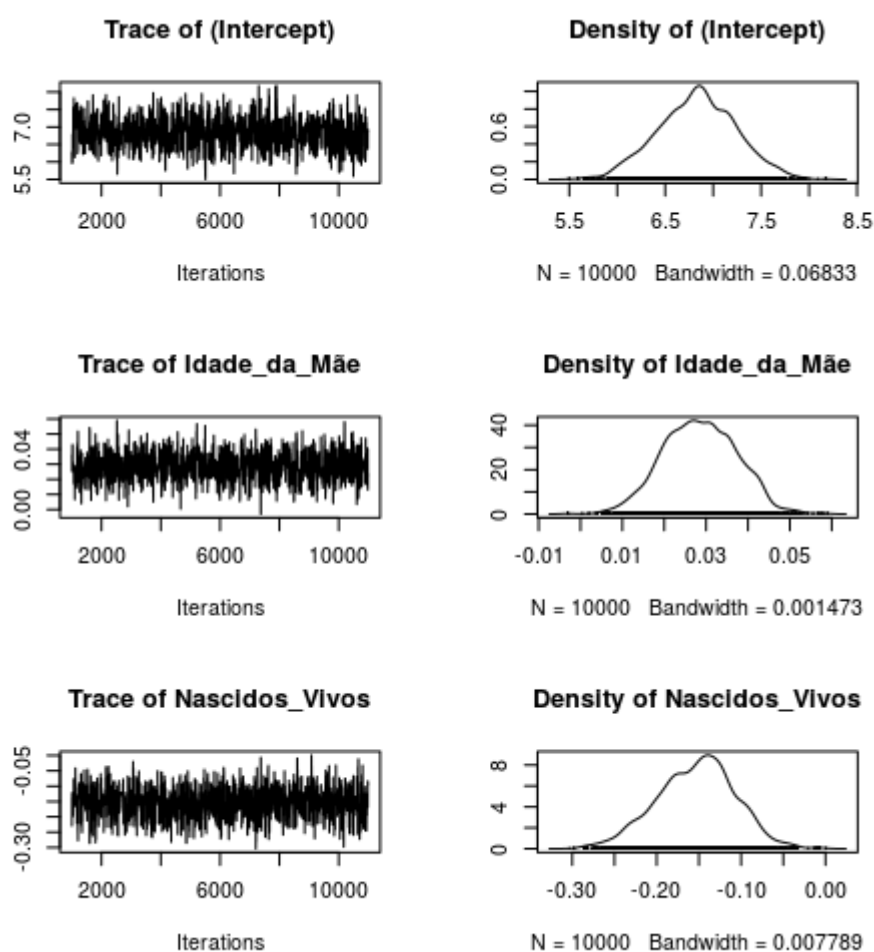


Figura 2 – Representação gráfica do traço e da função densidade posteriori

O modelo Bayesiano, assemelhasse ao modelo ajustado por verossimilhança, pela proximidade dos valores dos parâmetros. De acordo com a Tabela 11, o modelo clássico identifica 3524 nascimentos não prematuros, seguido de 81 nascimentos prematuros, dado que não nasceram prematuros. O modelo também mostra que 319 não nasceram prematuros, dado que foram prematuros e 333 confirmaram prematuros.

Tabela 11 – Matriz de confusão do modelo clássico

		Valores Preditos	
		Não prematuro	Prematuro
Valores Reais	Não Prematuro	3524	81
	Prematuro	319	333

No modelo Bayesiano, na Tabela 12 os resultados são bem semelhantes ao modelo clássico. Esse modelo registra 3522 nascimentos não prematuros, no entanto 83 nasceram prematuros, dado que não nasceram. Todavia, o modelo clássico conquistou a mais 2 acertos que o modelo Bayesiano. O modelo registra 319 nascimentos não prematuros, dado que nasceram prematuros e 333 realmente nasceram prematuros.

Tabela 12 – Matriz de confusão do modelo bayesiano

		Valores Preditos	
		Não prematuro	Prematuro
Valores Reais	Não Prematuro	3522	83
	Prematuro	319	333

Na Tabela 13 tem-se para o modelo clássico e Bayesiano, a classificação a partir das taxas de a acurácia ( $A_c$ ), a positiva verdadeira (TP), de falsos positivos (FP), a negativa verdadeira (TN), a de falsos negativos (FN) e a precisão. E verifica-se que há valores idênticos e também semelhantes, em ambos os modelos.

Tabela 13 – Indicadores do modelo ajustado

	Acurácia	TP	FP	TN	FN	Precisão
Modelo clássico	0,9060	0,5107	0,1242	0,9775	0,4892	0,8043
Modelo Bayesiano	0,9055	0,5107	0,1273	0,9769	0,4892	0,8004



## 5 Conclusão

Para a seleção do modelo, foram utilizadas técnicas estatísticas computacionais no R, (R Core Team, 2018) e através de técnica *stepwise*, foi possível verificar que as covariáveis idade da mãe, número de nascidos vivos, número de consultas pré-natal, tipo de gravidez e peso da criança ao nascer, foram estatisticamente significativas para descrever o modelo, portanto podemos concluir que as covariáveis do ajuste interferiram diretamente na prematuridade das crianças.

Foi aplicado um estudo teórico dos modelos lineares generalizados considerando a distribuição binomial e função de ligação *logit* com ênfase na regressão logística.

Por meio do gráfico de envelope simulado, podemos observar que apenas 67 (1,57%) dos pontos encontram-se fora da banda de confiança, logo o modelo é adequado para explicar o comportamento dos dados.

## Referências

- AGUIAR, Z. N. Sus-sistema único de saúde. *Antecedentes, percurso*, 2011. Citado na página 11.
- AKAIKE, H. A new look at the statistical model identification. *IEEE transactions on automatic control*, Ieee, v. 19, n. 6, p. 716–723, 1974. Citado na página 18.
- ARRABAL, C. T. et al. Estimaco clssica e bayesiana para relao espcierea com distribuices truncadas no zero. Universidade Federal de So Carlos, 2012. Citado na pgina 19.
- BOTLHO, S. M. et al. O cuidar materno diante do filho prematuro: um estudo das representaes sociais. *Revista da Escola de Enfermagem da USP*, v. 46, n. 4, p. 929–934, 2012. Citado na pgina 11.
- BURNHAM, K. P.; ANDERSON, D. R. Multimodel inference: understanding aic and bic in model selection. *Sociological methods & research*, Sage Publications Sage CA: Thousand Oaks, CA, v. 33, n. 2, 2004. Citado na pgina 18.
- CAB, M. *Rezende J. Obstetrcia fundamental*. Rio de Janeiro: Guanabara Koogan, 2014. Citado na pgina 12.
- DELLAQUA, D. C.; CARDOSO, F. S. Assistncia de enfermagem ao recm-nascido prematuro extremo. *Revista Eletrnica da Faculdade Evanglica do Paran*, Curitiba, v. 2, n. 4, p. pg–02, 2012. Citado na pgina 12.
- DRAPER, N. R.; SMITH, H. *Applied Regression Analysis*. New York: John Wiley & Sons, 1998. v. 326. Citado na pgina 18.
- GIOLO, S. R.; COLOSIMO, E. A. Anlise de sobrevivncia aplicada. *Edgard Blucher*, 2006. Citado na pgina 20.
- GODBOLE, S. Exploiting confusion matrices for automatic generation of topic hierarchies and scaling up multi-way classifiers. *Annual Progress Report, Indian Institute of Technology–Bombay, India*, 2002. Citado na pgina 21.
- KIL, D. H.; SHIN, F. B. *Pattern Recognition and Prediction with Applications to Signal Processing (Aip Series in Modern Acoustics and Signal Processing)*. New York: Springer-Verlag New York, Inc., 1998. Citado na pgina 21.
- KINAS, P. G.; ANDRADE, H. A. *Introduo  anlise bayesiana (com R)*. So Paulo: Consultor Editorial, 2017. Citado 3 vezes nas pginas 16, 17 e 18.
- KLIEGMAN, R. et al. *Nelson tratado de pediatria*. Rio de Janeiro: Elsevier Brasil, 2014. Citado na pgina 12.
- LEONE, C. R. et al. *Assistncia integrada ao recm-nascido de baixo risco*. So Paulo: Atheneu, 2012. Citado na pgina 12.

- MCCULLAGH, P.; NELDER, J. A. *Generalized Linear Models*. New York: CRC press, 1989. v. 37. Citado na página 13.
- NELDER, J. A.; WEDDERBURN, R. W. *Generalized Linear Models*. New York: CRC press, 1972. v. 135. Citado 2 vezes nas páginas 10 e 13.
- NETO, F. R. G. X. et al. Qualidade da atenção ao pré-natal na estratégia saúde da família em sobral, ceará. *Revista Brasileira de Enfermagem*, Associação Brasileira de Enfermagem, v. 61, n. 5, 2008. Citado na página 11.
- OLSON, D. L.; DELEN, D. *Advanced data mining techniques*. [S.l.]: Springer Science & Business Media, 2008. Citado na página 21.
- PAULA, G. A. *Modelos de regressão: com apoio computacional*. São Paulo: [s.n.], 2004. Citado 4 vezes nas páginas 14, 15, 19 e 20.
- R Core Team. *R: A Language and Environment for Statistical Computing, R Foundation for Statistical Computing, Austria, 2015*. [S.l.]: ISBN 3-900051-07-0: URL <http://www.R-project.org>, 2018. Citado 5 vezes nas páginas 5, 6, 22, 26 e 31.
- RENCHER, A. C.; SCHAALJE, G. B. *Linear models in statistics*. New Jersey: John Wiley & Sons, 2008. Citado na página 18.
- RETALHOS HISTÓRICOS DE CAMPINA GRANDE. *Retalhos Históricos de Campina Grande*. 2011. Disponível em: <[http://cgretalhos.blogspot.com/2010/04/memoria-fotografica-maternidade-elpidio.html#.W\\_zZuHWPI8o](http://cgretalhos.blogspot.com/2010/04/memoria-fotografica-maternidade-elpidio.html#.W_zZuHWPI8o)>. Acesso em: 26 nov. 2018. Citado na página 12.
- SALGE, A. K. M. et al. Fatores maternos e neonatais associados à prematuridade. 2009. Citado na página 11.
- SCHWARZ, G. et al. Estimating the dimension of a model. *The annals of statistics*, Institute of Mathematical Statistics, v. 6, n. 2, p. 461–464, 1978. Citado na página 19.
- SOUZA, G. C. d. A.; COSTA, I. d. C. C. O sus nos seus 20 anos: reflexões num contexto de mudanças. *Saude e sociedade*, SciELO Public Health, v. 19, p. 509–517, 2010. Citado na página 11.
- Thomas. *OpenBUGS*. 2004. Citado 2 vezes nas páginas 5 e 6.
- TURKMAN, M. A. A.; SILVA, G. L. Modelos lineares generalizados: da teoria à prática. In: *VIII Congresso Anual da Sociedade Portuguesa de Estatística*. Lisboa: [s.n.], 2000. Citado na página 13.