UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS EXATAS E DA NATUREZA

DEPARTAMENTO DE ESTATÍSTICA

Uso de Regressão Logística para Identificar os Fatores de Risco associados à Ocorrência de Anomalias Congênitas em Recém-nascidos

Monografia apresentada ao Departamento de Estatística da Universidade Federal da Paraíba - UFPB para a obtenção do grau de Bacharel em Estatística

Por **LIDIA DAYSE ARAUJO DE SOUZA**

Orientador: **JOAB DE OLIVEIRA LIMA**

João Pessoa - PB, Brasil.

Abril/2013

Dados Internacionais de Catalogação na Publicação (CIP)
Universidade Federal da Paraíba
Biblioteca Setorial do CCEN

S729u Souza, Lidia Dayse Araujo de.
Uso de regressão logística para identificar os fatores de risco Associados à ocorrência de anomalias congênitas em recém-nascidos / Lidia Dayse Araujo de Souza. – João Pessoa, 2013.
37p.

Monografia (Bacharelado em Estatística) – Universidade Federal da Paraíba.
Orientador: Profº. Joab de Oliveira Lima.

1. Regressão logística. 2. Malformação congênita. 3. Modelos de Regressão. I. Título.

UFPB/BS-CCEN CDU 519.246.8(043.2)

Monografia de Projeto Final de Graduação sob o título "**Uso de Regressão Logística para Identificar os Fatores de Risco associados à Ocorrência de Anomalias Congênitas em Recém-nascidos"**, defendida por Lidia Dayse Araujo de Souza e aprovada em 26 de Abril de 2013 em João Pessoa no Estado da Paraíba, sendo apreciada pela banca examinadora constituída pelos Professores:

Aprovada em ____/____/_____.

BANCA EXAMINADORA

_____

Prof. Dr. Joab de Oliveira Lima

Universidade Federal da Paraíba

_____

Prof. Dr. José Carlos de Lacerda Leite

Universidade Federal da Paraíba

_____

Prof. Ms. Izabel Cristina Alcantara de Souza

Universidade Federal da Paraíba

Resumo

Os modelos de regressão logística vêm sendo aplicado intensamente em várias áreas de conhecimento e, em especial, na área da saúde. A condição de se estudar variáveis respostas binárias em função de um conjunto de fatores explicativos tem se tornado cada vez mais comum em estudos epidemiológicos.

Assim, o presente estudo tem como objetivo utilizar um modelo de regressão logística para investigar os fatores de risco associados à ocorrência de malformação congênita em crianças de um hospital de João Pessoa – PB.

Os resultados mostraram que a idade, a escolaridade da mãe, o uso de corticoides durante a gravidez, o tipo de parto e as medidas de APGAR de 1 e 5 minutos estavam associadas com a probabilidade de nascimento de filhos com alguma anomalia congênita. Além disso, constatou-se que o modelo ajustado conseguiu classificar corretamente mais de 93% dos casos examinados.

Palavras-chave: Regressão Logística, Malformação Congênita.

Abstract

The logistic regression models have been extensively applied in several areas of knowledge, particularly in the area of health. The condition of studying binary response variables in terms of a set of explanatory factors have become increasingly common in epidemiological studies.

Thus, this study aims to use a logistic regression model to investigate the risk factors associated with the occurrence of congenital malformation in children from a hospital in João Pessoa - PB.

The results showed that age, mother's education, the use of corticosteroids during pregnancy, type of childbirth and APGAR measures 1 and 5 minutes were associated with the probability of birth of children with congenital abnormality. Furthermore, it was found that the adjusted model could correctly classify over 93% of the cases examined.

Keywords: Logistic Regression, Congenital Malformation

Dedico esta monografia a meu querido e amado irmão José Roberto Araujo de Souza (*in memoriam*).

Agradecimentos

» A Deus, Jesus Cristo e Maria Santíssima, por tudo.

» À minha grande família: meu pai Gilson, minha mãe Josefina, e meus irmãos Carlinhos e meu sobrinhos Felipe Anderson e Carlos Eduardo, ao meu namorado Leonardo Silva pessoas que tanto amo.

» Aos meus demais familiares e, especialmente, aos meus tios Jairo Procópio de Moura e Galega Araújo por me ajudarem durante todo esse tempo de estudo, o meu mais profundo e sincero obrigada.

» Ao Professor Dr. Joab de Oliveira Lima pela orientação, apoio, dedicação e compreensão para a construção deste trabalho.

» Aos professores José Carlos e Izabel Cristina por terem aceitado o convite para a banca examinadora.

» A todos os meus amigos de graduação e em especial: Ísis, Luana, Alisson, José Edson, Antonio, Wanessa, Amaury, Thiago Lima, Fabio, Daniel, Jessica, Iam, Paola, Tié, Telmo, Camila Ravena, Anna Paola, Lerivan, Juliana Tavares.

» A todos os meus professores que contribuíram para minha formação acadêmica e em especial: Ana Flávia Uzeda dos Santos Macambira, Andréa Vanessa Rocha, Antônio Marcos Moreira, Eufrásio de Andrade Lima Neto, Gilmara Alves Cavalcanti, Hemílio Fernandes Campos Coêlho, Joab de Oliveira Lima, João Batista Parente, José Carlos de Lacerda Leite, Luciano da Costa Silva, Marcelo Rodrigo Portela Ferreira, Neir Antunes Paes, Renata Patrícia Lima Jerônymo, Sydney Gomes da Silva, Turíbio José Gomes dos Santos e *in memoriam* Abdoral.

» Às Funcionárias Fátima e Renata.

Lista de Símbolos

β:	Parâmetros das variáveis explicativas do modelo logístico
RG:	Regressão Logística
MLG:	Modelos Lineares Generalizados
μ:	Média da População
ε:	Erro Aleatório
σ^2:	Variância da População
E(Y):	Valor Esperado de Y
Var(Y):	Variância de Y
Ω:	Experimento Aleatório
∅:	Probabilidade de "sucesso"
m:	Número de Repetições
MV:	Máxima Verossimilhança
ln:	Logaritmo Natural

Lista de Tabelas

Lista de Figuras

Sumário

1. INTRODUÇÃO

As anomalias congênitas, também chamadas de defeitos de nascimento, são anormalidades físicas presentes no momento do nascimento. Estatísticas nacionais apontam que, aproximadamente, 3% dos recém-nascidos têm algum defeito congênito grave. As anomalias mais frequentes são as cardiopatias, anencefalia, Trissomia, Síndrome de Down entre outras. Algumas dessas anomalias só serão descobertas a media que a criança cresce.

Muitas das malformações congênitas importantes podem ser diagnosticadas antes do nascimento, o que facilita tanto o tratamento quanto a reparação desses "defeitos de fábrica". Sabe-se, além disso, que algumas anomalias não necessitam de tratamento, enquanto outras não podem ser tratadas e, em consequência, a criança fica gravemente incapacitada de forma permanente.

Em geral, a ocorrência das malformações congênitas não surge apenas a partir de uma única causa, mas sim de vários fatores, ente eles, os aspectos sociodemográfico da mãe, os cuidados com a saúde do bebê durante a gestação, as condições do parto, a herança genética, as doenças preexistentes ou contraídas pela mãe nos primeiros meses de gravidez, a ingestão de medicamentos, o tabagismo, o alcoolismo.

A prevalência de malformações congênitas na Paraíba em 2012 é de 1,34% e, por fim, em João Pessoa, a proporção de malformação congênita é cerca de 0,71% (TABNET SAUDE PB, 2012).

No âmbito local, o Instituto Cândida Vargas (ICV) representa o maior complexo materno-infantil do Estado da Paraíba, prestando assistência à saúde da mulher e do recém-nato da grande João Pessoa e cidades circunvizinhas. Atualmente o instituto contém 169 leitos e realiza, em média, 900 internações mensais, das quais 67% são para a realização de partos. Desses, aproximadamente 2,1% são de recém-nascidos que apresentam algum problema de anomalia congênita.

Para a maioria dos estudos epidemiológicos que tem como o objetivo investigar a relação entre algumas variáveis preditivas e uma variável-desfecho do tipo dicotômica,

a estratégia analítica mais simples e objetiva é o emprego dos Modelos de Regressão Logística, que vê sendo aplicados com sucesso há bastante tempo.

Por isso, o estudo aqui descrito pretende avaliar os fatores de risco associados com a ocorrência de malformação congênita em recém-nascidos em uma maternidade de João Pessoa – PB. Os detalhes da técnica estatística utilizadas, bem como os resultados serão discutidos nas seções que seguem.

2. OBJETIVOS

2.1. OBJETIVO GERAL

Este trabalho tem como objetivo geral identificar os principais fatores de risco associados à ocorrência de casos de recém-nascidos com malformação congênita.

2.2. OBJETIVOS ESPECÍFICOS

Além do objetivo geral, pretende-se, ainda, cumprir os seguintes objetivos específicos:

- Avaliar a consistência e a confiabilidade da base de dados utilizada no trabalho;

- Avaliar a qualidade do modelo logístico proposto em termos de adequabilidade e percentual de classificação correta.

3. REVISÃO DA LITERATURA

Existem na literatura inúmeras obras que tratam do ajuste e do uso dos Modelos de Regressão Logística para estudar variáveis respostas dicotômicas em função de um conjunto de variáveis explicativas.

Neto et al. (2009) relacionaram os fatores associados ao volume de líquido amniótico com a ocorrência de anomalias fetais. Os resultados mostraram que o líquido diminuído ou oligohi drâmico estava associado significativamente com as anomalias fetais. Já Brito et al. (2010) estudaram as associações entre as malformações congênitas e os seus principais fatores de risco materno em Campina Grande – PB. O estudo detectou 190 anomalias e os resultados indicaram que não houve associações significativas entre as variáveis sociais relativas às mães com a ocorrência de malformações congênitas.

Costa (2001), por outro lado, investigou o perfil das malformações congênitas em uma amostra de nascimentos no município do Rio de Janeiro. A amostra contemplou 9386 casos e se baseou em um estudo descritivo e seccional, cujos resultados revelaram uma prevalência de malformação congênita de 1,7%. Nessa mesma linha, Ramos, Oliveira e Cardoso (2008) investigaram a prevalência de malformações congênitas em recém-nascidos em um hospital da rede pública de Jequié – BA. A prevalência de malformação congênita observada no estudo foi de 3,1%, sendo mais frequente nas crianças do sexo masculino e prematuras de parto normal (vaginal).

Adicionalmente, o impacto das malformações congênitas na mortalidade perinatal e neonatal em uma maternidade-escola do Recife – PE foram o foco do estudo de Amorim et al. (2006) e os resultados encontrados pelos autores apontaram uma prevalência de malformações de 2,8% para a amostra total; de 2,7% entre os nativivos e de 6,7% entre os natimortos.

Por fim, através da aplicação de um modelo de regressão logística, Ferreira (2012) concluiu que existe uma associação estatisticamente significativa entre o desenvolvimento de leucemias na infância e antecedentes de exposição química materna durante a gestação e consumo de serviços de saúde durante a gestação.

4. METODOLOGIA

4.1. MODELOS LINEARES GENERALIZADOS – MLG

Hoje em dia nos estudos estatísticos se estar sempre tentando estudar correlações entre certas variáveis em estudo, mas precisamente a correlação de uma variável específica, chamada de variável resposta, em relação a outras variáveis, conhecidas como variáveis explicativas. Um dos métodos bastante utilizado para determinar essa relação entre as variáveis explicativas e a variável resposta é o modelo de regressão.

Uma classe de modelos de regressão, desenvolvido por Legendre e Gauss (TURKMAN; SILVA, 2000), que predominou no inicio do século XIX, foi a classe dos modelos de regressão linear normal. Nesse mesmo período outros modelos também foram criados para explicar situações em que o modelo linear normal não se adequava, como, por exemplo, os modelos não lineares.

A estrutura geral do modelo linear normal é da forma:

$$Y = \mu + \varepsilon \qquad [1]$$

em que,

Y é o vetor de dimensão $n \times 1$ da variável resposta;

$\mu = E(Y) = X\beta$ é componente sistemático;

X é a matriz de dimensão $n \times p$ do modelo;

$\beta = (\beta_1, \dots, \beta_p)^t$ é o vetor dos parâmetros;

$\varepsilon = (\varepsilon_1, \dots, \varepsilon_n)^t$ é o vetor de erros aleatórios com $\varepsilon_i \sim N(0, \sigma^2),\ i = 1, \dots, n$

O modelo acima, como já foi dito, para algumas situações, não era adequado justamente porque não atendia as premissas de sua aplicação. Por exemplo, os erros ε não eram identicamente distribuídos como uma normal ou não eram independentes, ou ainda, a suposição de homogeneidade da variância desses erros era violada.

Tendo em vista esta situação, Nelder e Wedderburn (1972) unificou todos os modelos propostos anteriormente baseados principalmente no modelo linear normal e o chamou de Modelos Lineares Generalizados ou simplesmente MLG (DEMÉTRIO, 2002). Com essa unificação, os modelos abaixo se tornaram casos particulares do MLG (TURKMAN; SILVA, 2000):

- Modelo de regressão linear normal,
- Modelos de análise de variância e covariância,
- Modelo de regressão logística,
- Modelo de regressão Poisson,
- Modelos Log-lineares para tabelas de contingências multidimensionais,
- Modelo probito para estudo de proporções, etc.

Assim, da mesma forma que o modelo linear normal "dominou" os estudos e publicações do século XIX, atualmente são os modelos lineares generalizados que dominam os estudos de modelagem estatística.

Como todos os modelos têm as suas suposições, com os MLG's não poderia ser diferente. O uso do MLG pressupõe que a variável resposta pertença à família exponencial.

DEFINIÇÃO 1 (Família Exponencial): Diz-se que uma variável aleatória Y tem distribuição pertencente à família exponencial se a sua função densidade e probabilidade (*fdp*) puder ser escrita na forma (DEMÉTRIO, 2002):

$$f(y|\theta,\phi) = exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\} \qquad [2]$$

em que,

θ e ϕ são parâmetros escalares

a(.), b(.) e c(.,.) são funções reais conhecidas

Logo, observa-se que a estrutura do MLG é composta por três partes:

- A componente aleatória: representada pela a variável resposta;
- A componente sistemática: representada pela combinação linear das variáveis explicativas;
- A função de ligação: função que linearizar a relação entre $E(Y)$ e $X\beta$.

Desse modo, a combinação das três estruturas acima determina o tipo de modelo que a filosofia MLG engloba. Por exemplo, se a componente aleatória for normal, a função de ligação for a identidade e as covariáveis forem contínuas, então estar-se diante do modelo de regressão linear. Na tabela abaixo são mostrados alguns tipos de MLG (TURKMAN; SILVA, 2000).

TABELA 1: Variações dos Modelos Lineares Generalizados

Componente Aleatória	**Tipos das Covariáveis**	**Função de Ligação**	**Tipo de Modelo**
Normal	Contínuas	Identidade	Regressão linear
Normal	Categorizadas	Identidade	Análise de Variância
Normal	Mistas	Identidade	Analise de Covariância
Binomial	Mistas	*Logit*	Regressão Logística
Poisson	Mistas	Logarítima	Log-linear

Como pode ser notado na tabela acima, uma das variações do MLG é o Modelo de Regressão Logística que será estudado em detalhes na próxima seção.

4.1.1. REGRESSÃO LOGÍSTICA

4.1.1.1. ESTRUTURA DO MODELO

O Modelo de Regressão Logística é adequado para estudar situações em que existe um conjunto de variáveis explicativas que se correlacionam com uma variável resposta dicotômica. Para uma melhor compreensão do modelo de regressão logística, convém recorrer a um experimento descrito da seguinte forma:

Seja Ω um experimento aleatório composto por *m* repetições independentes de um evento dicotômico, isto é, contendo apenas dois resultados possíveis: **0** (fracasso) ou **1** (sucesso). Assim, diz-se que esse experimento é de natureza binária. Como as condições são as mesmas para todas as repetições, as probabilidades de cada resultado são $P(1) = \phi$ e $P(0) = 1 - \phi$ e constantes ao longo das *m* repetições. Chamando de Y a variável aleatória de interesse representando o número de vezes, nas *m* repetições, em que ocorre "sucesso", os valores que Y poderão assumir são 0, 1, 2, ..., *m*; a cada um desses valores está associada uma probabilidade $P(Y = y)$ de ocorrência, sendo y = 0, 1, 2, ..., *m* (BARRETO, 2011).

As definições do experimento Ω induz diretamente uma distribuição de probabilidade Binomial com parâmetros para *m* e ϕ para a variável aleatória Y. Com isso é possível determinar sua função de probabilidade, o valor esperado e sua variância como sendo:

$$f(y;\phi) = P(Y_i = y) = \binom{m}{y} \phi^y (1-\phi)^{m-y} \quad [3]$$

$$E(Y_i) = m\phi \quad [4]$$

$$Var(Y_i) = m\phi(1-\phi) \quad [5]$$

No caso particular do experimento acima em que só ocorra uma repetição, ou seja, $m = 1$, passa-se a trabalhar com uma distribuição de probabilidade de Bernoulli e a variável aleatória Y_i assumirá apenas os valores 0 ou 1.

No âmbito dos MLG's, para o caso de variáveis respostas contínuas, não há restrições quanto ao domínio da resposta esperada a ser estimada pelo modelo. Porém, no caso do experimento Ω, em que a variável aleatória Y_i pertence ao intervalo $[0, 1]$, o valor da resposta esperada será, na verdade, uma probabilidade, isto é, $P(Y_i = 1)$. Logo, para corrigir essa limitação será necessário utilizar uma transformação sobre a variável resposta Y, de modo a tornar o seu domínio contínuo na reta real. A transformação mais comumente utilizada para o caso de variáveis dicotômicas é a função exponencial que, por sua vez, transforme o valor esperado da variável Y na função de ligação logística (BARRETO, 2011):

$$E(Y_i) = \phi_1 = \frac{\exp(\eta_i)}{1 + \exp(\eta_i)} \qquad [6]$$

Nota-se, então, que o objetivo da transformação é plenamente atingido pela equação [6], uma vez que mesmo η_i podendo assumir qualquer valor real, ϕ_1, que é uma probabilidade, continuará restrito ao intervalo $[0,1]$, como era desejado. Dessa forma o valor estimado pelo modelo linear será η_i e, para a formalização do modelo de regressão logística, ele pode ser relacionado a um modelo linear, contendo $p - 1$ variáveis explicativas e p parâmetros, resultando em uma função conhecida como função resposta *logit*:

$$\eta_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ ... \ + \beta_{p-1} X_{ip-1} \ ; \ i = 1, 2, \cdots, n \qquad [7]$$

em que,

η_i é a resposta média *logit* para a *i-ésima* observação;

X_i é o valor da variável preditora para a *i-ésima* observação;

β_k $(k = 0, 1, \ldots, p - 1)$ é o coeficiente da regressão logística.

Substituindo a equação [7] em [6], o modelo de regressão logística múltipla pode ser formulado da seguinte forma:

$$E(Y_i) = \phi_i = \frac{exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{ip-1})}{1 + exp(\beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_{p-1} X_{ip-1})} \quad [8]$$

Este modelo descrito em [8] assume que os Y_i são variáveis aleatórias Bernoulli independentes com parâmetro $E(Y_i) = \phi_i$.

Um modelo de regressão logística simples seria denotado por uma única variável explicativa e dois parâmetros a serem estimados: β_0 e β_1.

Com o advento do modelo linear em [7], o valor esperado da resposta em [6] significa agora a probabilidade de que Y_i=1, dados os níveis para as preditoras em [7]; e η_i se vincula à resposta logística esperada pela conhecida função de ligação *logit*, pois explicando em [6] em termos de η_i, obtém-se:

$$\eta_i = ln\left(\frac{\phi_i}{1 - \phi_i}\right) \quad [9]$$

O termo entre parênteses em [9] é conhecido como *odds* de sucesso. Se a probabilidade de sucesso equivaler a 0,5, o *odds* vale 1; se a probabilidade de sucesso for inferior a 0,5, o *odds* é inferior a 1; e se a probabilidade de sucesso for superior a 0,5, o *odds* será maior do que 1.

A interpretação dos coeficientes β em regressão logística não é tão simples e nem de direta compreensão, como no caso de regressão linear, uma vez que se trata de uma função de resposta não-linear. Assim, no caso de regressão logística simples, um incremento unitário de X implicará em um efeito multiplicativo do *odds* estimado de sucesso, ou seja, $exp(b_1)$. Segundo Barreto (2011), no caso de uma regressão logística múltipla, um incremento unitário em X_1 ocasionaria esse mesmo efeito, mantidas constantes todas as demais variáveis do modelo $(X_2, X_3, \ldots, X_{p-1})$.

A prova da afirmação acima é desenvolvida no raciocínio matemático que segue. Assim, seja o $odds_1$ estimado definido por:

$$\widehat{odds}_1 = \frac{\hat{\phi}}{1 - \hat{\phi}} \quad [10]$$

$$\widehat{odds}_1 = \frac{\frac{exp(b_0 + b_1 X)}{1 + exp(b_0 + b_1 X)}}{1 - \frac{exp(b_0 + b_1 X)}{1 + exp(b_0 + b_1 X)}} \quad [11]$$

$$\widehat{odds}_1 = exp(b_0 + b_1 X) \quad [12]$$

De maneira análoga, seja o $odds_2$ definido como segue, para uma variação unitária em X:

$$\widehat{odds}_2 = \frac{\frac{exp(b_0 + b_1 (X + 1))}{1 + exp(b_0 + b_1 (X + 1))}}{1 - \frac{exp(b_0 + b_1 (X + 1))}{1 + exp(b_0 + b_1 (X + 1))}} \quad [13]$$

$$\widehat{odds}_2 = \frac{\frac{exp(b_0 + b_1 X + b_1)}{1 + exp(b_0 + b_1 X + b_1)}}{1 - \frac{exp(b_0 + b_1 X + b_1)}{1 + exp(b_0 + b_1 X + b_1)}} \quad [14]$$

$$\widehat{odds}_2 = \frac{exp(b_0 + b_1 X + b_1)}{1 + exp(b_0 + b_1 X + b_1) - exp(b_0 + b_1 X + b_1)} \quad [15]$$

$$\widehat{odds}_2 = exp(b_0 + b_1 X + b_1) \quad [16]$$

$$\widehat{odds}_2 = exp(b_0 + b_1 X)\, exp(b_1) \quad [17]$$

$$\widehat{odds}_2 = \widehat{odds}_1\, exp(b_1) \quad [18]$$

Logo, conclui-se que o *odds ratio* (também conhecido como razão entre os *odds*, ou razão das chances), que mensura a taxa de variação de *odds* de sucesso em função da variação em X, equivale a:

$$\frac{\widehat{odds}_2}{\widehat{odds}_1} = exp(b_1) \quad [19]$$

Um aspecto que deve ser acrescentado à discussão em relação ao modelo de regressão logística é que ao invés de formalizá-lo como em [6] e [7], poder-se-ia tentar utilizar o já conhecido modelo de regressão linear simples para modelar a resposta binária, isto é:

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i \quad [20]$$

em que: $Y_i = 0$ ou $Y_i = 1$.

Porém, se o modelo descrito em [20] fosse adotado, certamente ocorreria dois problemas em relação aos pressupostos de regressão linear. O primeiro deles está relacionado ao fato de se modelar uma variável resposta binária Y_i, assumindo os valores zero ou um, o que implicaria que os termos de erro, necessariamente, também assumiriam apenas dois valores distintos, quais sejam, $(1 - \beta_0 + \beta_1 X_i)$ e $(-\beta_0 - \beta_1 X_i)$. Esse fato torna inviável considerar os erros como normalmente distribuídos. Além disso, há a heterocedasticidade dos termos de erro que variam à medida que os níveis para X se alteram. Esses problemas, adicionados ao fato da resposta esperada

$E(Y_i)$, no caso do modelo em [10], pode assumir valores fora do intervalo [0,1], acabam por tornar mais interessante modelar a resposta binária por meio de [6] e [7] (BARRETO, 2011).

4.1.1.2. ESTIMAÇÃO DOS PARÂMETROS DO MODELO

Um método de estimação de parâmetros bastante utilizado, quando a forma funcional da distribuição de probabilidade das variáveis aleatórias envolvidas na modelagem é conhecida, é o de Máxima Verossimilhança (MV). A ideia básica do método, como diz seu nome, é encontrar um conjunto de estimadores que maximizem os parâmetros populacionais à luz da amostra dos dados observados. Para isso, a função de probabilidade conjunta da amostra, sob o modelo especificado, é avaliada para cada uma das observações da variável resposta, sendo tratada como uma função dos parâmetros do modelo.

De maneira geral, a verossimilhança é dada pela função de probabilidade conjunta das n observações amostrais, em função do vetor de parâmetros do modelo β, sendo denotada por $L(\beta)$. O método escolhe como estimador de MV um vetor b que forneça o maior valor possível para a função $L(\beta)$.

O método de MV pode ser utilizado para a estimação em regressão logística, já que se conhece a distribuição de probabilidade associada à variável resposta binária. Como comentado anteriormente, se houver apenas uma repetição para cada variável aleatória Y_i, então se está trabalhando sob a distribuição de probabilidade de Bernoulli. Com isso, a função de probabilidade de cada Y_i é dada por:

$$P(Y_i = k) = \phi_i^k (1 - \phi_i)^{1-k} \; ; \; k = 0, 1 \qquad [21]$$

Com base nesse modelo probabilístico, e assumindo-se que as variáveis aleatórias Y_i são independentes, a função de probabilidade conjunta dos Y_i equivale à:

$$g(Y_i, \dots, Y_n; \phi_i, \dots, \phi_n) = \prod_{i=1}^{n} \phi_i^{k_i}(1-\phi_i)^{1-k_i} \;\; ; \;\; k = 0,1 \; e \; i = 1,2,\cdots,n \qquad [22]$$

Aplicando o logaritmo natural (ln) na função de probabilidade conjunta, tem-se:

$$ln[g(Y_i, \dots, Y_n; \phi_i, \dots, \phi_n)] = ln \prod_{i=1}^{n} \phi_i^{k_i}(1-\phi_i)^{1-k_i} \qquad [23]$$

Utilizando as propriedades do logaritmo natural em relação ao produto e à diferença, e sabendo-se que para cada observação da amostra de dados tem-se que $Y_i = k_i$, pode-se escrever:

$$\ln[g(Y_i, \dots, Y_n; \phi_i, \dots, \phi_n)] = \sum_{i=1}^{n} Y_i \, ln\left(\frac{\phi_i}{1-\phi_i}\right) + \sum_{i=1}^{n} ln\,(1-\phi_i) \qquad [24]$$

Hipoteticamente, considerando um caso de regressão logística simples, utilizando o *logit* nos moldes como foi definido em [9], com $ln\left(\frac{\phi_i}{1-\phi_i}\right) = \eta_i = \beta_0 + \beta_1 X_i$, e sabendo-se que $E(Y_i) = \phi_i$ e $ln[g(Y_i, \dots, Y_n; \phi_i, \dots, \phi_n)] = L(\beta_0; \beta_1)$ é a função de log-verossimilhança dos parâmetros $(\beta_0; \beta_1)$ a serem estimados com base nos dados observados, a equação [6] pode ser reescrita sob a forma:

$$L(\beta_0, \beta_1) = \sum_{i=1}^{n} Y_i(\beta_0 + \beta_1 X_i) - \sum_{i=1}^{n} ln[1 + exp\,(\beta_0 + \beta_1 X_i)] \qquad [25]$$

A equação [25] é justamente a função de log-verossimilhança aplicável nos casos de regressão logística simples. A partir de uma amostra de observações, as estimativas de máxima verossimilhança dos parâmetros são dadas pelos valores de β_0 e β_1 que maximizam essa função (BARRETO, 2011).

Assim, uma vez obtidas as estimativas de MV para β_0 e β_1, no caso de uma regressão logística simples, pode-se obter a resposta logística estimada (ou valor ajustado) para o *i-ésimo* caso, qual seja:

$$\hat{\phi}_i = \frac{exp(b_0 + b_1 X)}{1 + exp(b_0 + b_1 X)} \qquad [26]$$

A equação [26] pode, obviamente, ser estendida para o caso de regressão logística múltipla. Um outro importante passo na modelagem de Regressão Logística é a predição dos valores futuros. Essa explicação é detalhada a seguir.

5. RESULTADOS E DISCUSSÃO

5.1. BASE DE DADOS

Para a realização desse estudo foi utilizado o banco de dados, referente aos anos de 2011 e 2012, da Unidade de Cuidados Intermediários Neonatal – UCIN do Instituto Cândida Vargas do município de João Pessoa – PB. Essa maternidade realiza mensalmente uma média de 600 partos de mães provenientes da capital (48%) ou de cidades circunvizinhas (52%).

O registro das informações tanto das mães quanto dos recém-nascidos é realizado a partir de uma ficha própria que está estrutura no software Epi Info™. As informações contidas nessa plataforma referem-se aos dados demográficos, epidemiológicos e clínicos das gestantes, além das características do parto e do recém-nascido.

Depois de realizar a consistência do banco de dados, restaram 543 registros, sendo esse total o tamanho da amostra considerada. Além disso, das variáveis disponíveis, decidiu-se manter apenas aquelas que apresentaram correlações ou associação forte com o indicativo de malformação congênita, que é a variável resposta (dicotômica) de interesse. As variáveis selecionadas foram: Idade da mãe, Escolaridade da mãe, Uso de corticoide no pré-natal, Tipo de parto e Índices APGAR de 1 minuto e 5 minutos. Mais especificamente, os índices de APGAR medidos no 1º e no 5º minutos tratam-se de um método simples e eficiente para medir a saúde do recém-nascido e de determinar se ele precisa ou não de alguma assistência médica imediata. Desde que a Dra. norte-americana Virginia Apgar o desenvolveu este teste, em 1952, o procedimento passou a ser rotineiro após os partos.

Assim, um minuto após nascer e novamente aos cinco minutos de vida fora do útero, o recém-nascido será avaliado a partir de um conjunto de 5 aspectos:

- Frequência cardíaca;
- Respiração;
- Tônus muscular;
- Reflexos;
- Cor da pele

Cada um destes itens recebe uma nota entre 0 e 2 para se chegar a um total geral de 10 pontos. Quando avaliação total fica entre 8 e 10 mostra que recém-nascido está com um excelente estado de saúde, e que, provavelmente, não vai precisar de cuidados extras; quando o escore total está entre 5 e 7 indica um estado de saúde regular e pode haver necessidade de ajuda de aparelhos para respirar e, por fim, índices abaixo de 5 pontos indicam que a saúde geral do bebê exige auxílio médico especial.

Estudos mostram que o índice APGAR medido no 5º é, geralmente, maior que àquele medido no primeiro minuto, uma vez que o estresse do parto, para o bebê, vai se esvaindo com o tempo. Para os propósitos desse estudo, resolveu-se classificar o índice APGAR como "Normal" para valores iguais ou superiores a 5 e "Alterado" caso contrário.

Adicionalmente, conforme é discutido a seguir, os resultados foram divididos em dois grupos: *(a)* Resumo Estatístico e *(b)* Ajuste do Modelo Logístico. Na seção de Resumo Estatístico serão apresentados os resultados descritivos e exploratórios obtidos a partir da amostra, enquanto a seção de Ajuste do Modelo Logístico será dedicada à avaliação dos fatores de risco que significativos para prevê a proporção de crianças com malformação congênita, além da análise de resíduos e do estudo da qualidade preditiva do modelo.

5.2. RESUMO ESTATÍSTICO

A TABELA 2 mostra a distribuição de frequência dos 543 registros das crianças que nasceram no Instituto Cândida Vargas entre os anos de 2011 e 2012, segundo alguns fatores de risco que foram disponibilizados para a investigação. Inicialmente, observa-se que 24,86% (135) dos bebês nasceram com alguma malformação congênita e que a idade média das mães que tiveram filhos com alguma anomalia é de 35 anos aproximadamente. Essa última informação, talvez, seja um alerta da relação entre a idade avançada da mãe e a possibilidade de geração de filhos com problemas congênitos.

Uma análise mais detalhada revela que dentre as crianças que nasceram com alguma anomalia congênita, 71,11% das mães eram analfabeta ou cursaram, no máximo, até o ensino fundamental; já entre os filhos nascidos sem qualquer malformação congênita,

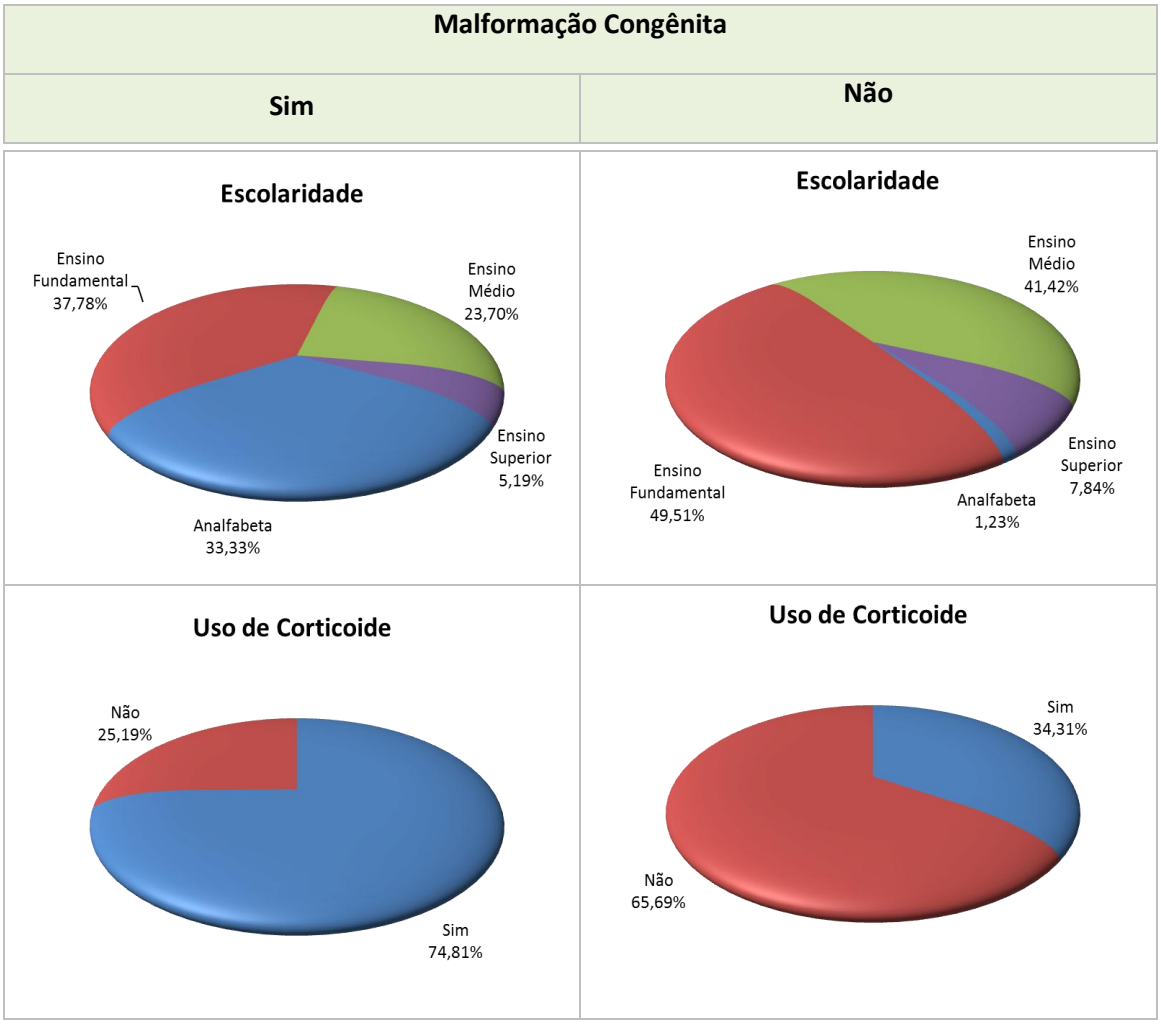
apenas 1,23% das mães eram analfabetas. Isso presume que o grau de escolaridade da mãe é um fator preponderante na redução da chance de conceber filhos com anomalias, uma vez que quanto maior for o grau de instrução da mãe, mais consciente ela será com relação aos cuidados com a gestação (pré-natal), com a alimentação, com a sua própria saúde; além de propiciar um acesso melhor aos serviços de saúde, já que a escolaridade elevada favorece a um padrão de rendimento familiar mais alto, o que explica a maior facilidade ao uso de sistema de saúde mais apropriado ou particular.

Nota-se ainda na TABELA 2, que os nascimentos de crianças com malformação congênita estão relacionados ao uso de corticoide no pré-natal; aos partos naturais (varginais) e às medições do índice APGAR tanto no primeiro quanto no quinto minutos. Em especial, foi comentado anteriormente que o índice APGAR medido no quinto minuto tem a tendência de fornecer uma avaliação da criança melhor que a sua primeira medição. No entanto, para uma criança, cujo índice APGAR medido no quinto minuto confirmou o comprometimento de sua saúde em relação à primeira medição desse índice, pode-se afirmar que será quase certo que essa criança apresentará, em um futuro próximo, alguma anomalia congênita, tal como, síndrome de down ou uma hidrocefalia. Os gráficos apresentados nas FIGURAS 2 e 3 mostram mais informações sobre esse resumo descritivo das variáveis envolvidas no estudo.

TABELA 2: Resumo descritivo dos fatores de risco, segundo a presença de malformação congênita.

Fatores de Risco		Resposta				Total		Valor-p
		Sim		Não				
		Freq.	%	Freq.	%	Freq.	%	
Idade da Mãe (média ± dp)		35,13 ± 4,48		25,55 ± 6,83		27,94 ± 7,56		0,0000
Escolaridade da Mãe	Analfabeta	45	33,33	5	1,23	**50**	**9,21**	0,0000
	Ensino Fundamental	51	37,78	202	49,51	**253**	**46,59**	
	Ensino Médio	32	23,70	169	41,42	**201**	**37,02**	
	Ensino Superior	7	5,19	32	7,84	**39**	**7,18**	
	Total	**135**	**100,00**	**408**	**100,00**	**543**	**100,00**	
Uso de Corticoide no Pré-Natal	Sim	101	74,81	140	34,31	**241**	**44,38**	0,0000
	Não	34	25,19	268	65,69	**302**	**55,62**	
	Total	**135**	**100,00**	**408**	**100,00**	**543**	**100,00**	

Fatores de Risco		Resposta				Total		Valor-p
		Sim		Não				
		Freq.	%	Freq.	%	Freq.	%	
APGAR de 1 min.	Alterada	90	66,67	51	12,50	141	25,97	
	Normal	45	33,33	357	87,50	**402**	**74,03**	0,0000
	Total	**135**	**100,00**	**408**	**100,00**	**543**	**100,00**	
APGAR de 5 min.	Alterada	84	62,22	8	1,96	**92**	**16,94**	
	Normal	51	37,78	400	98,04	**451**	**83,06**	0,0000
	Total	**135**	**100,00**	**408**	**100,00**	**543**	**100,00**	



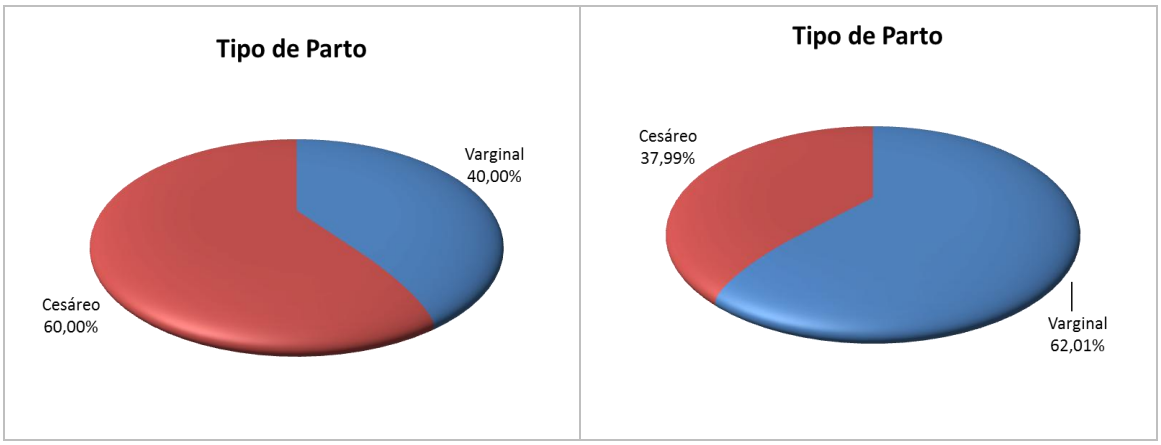


FIGURA 1: Distribuição amostral do grau de escolaridade da mãe, uso de corticoides e tipo de parto

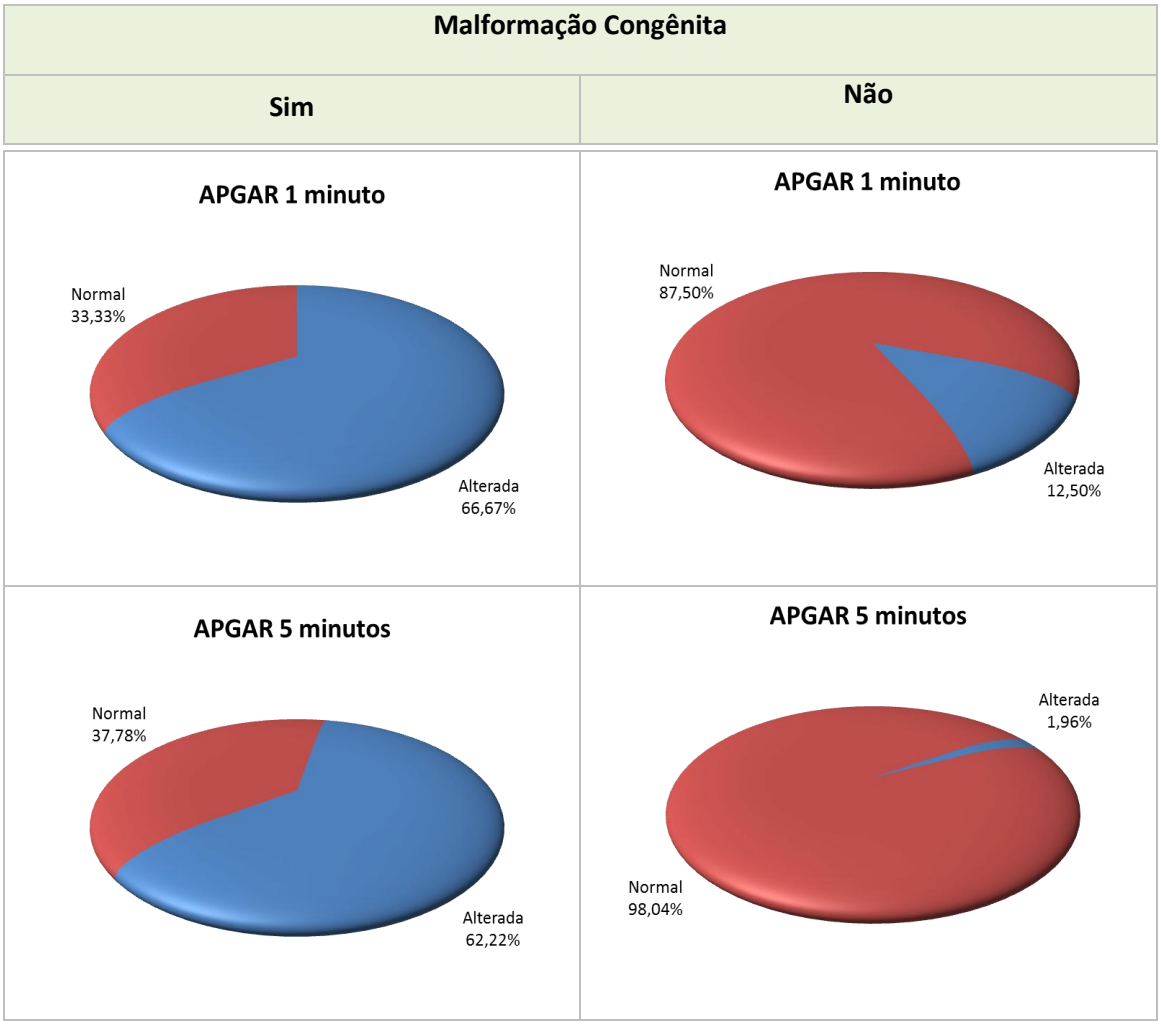


FIGURA 2: Distribuição amostral segundo as classificações de APGAR de 1 e de 5 minutos

5.3. AJUSTE DO MODELO LOGÍSTICO

A TABELA 3 apresenta os fatores de risco que foram significativos para prevê o nascimento de um bebê com alguma anomalia congênita. Nessa tabela, as linhas marcadas com "(R)" na descrição dos parâmetros referem-se às categorias de referência das variáveis categóricas. Os resultados demonstram que as mães analfabetas apresentam, aproximadamente, 29 vezes mais chances de gerarem filhos com malformação congênita do que a mães com ensino superior, quando mantêm-se os outros fatores constantes. Além disso, mães que usaram corticoides durante a gravidez estão 5 vezes mais propensas a terem filhos com alguma anomalia e os partos vaginais (normais) aumentam cerca de 3 vezes os riscos de se conceber filhos com problemas congênitos.

Além do mais, constata-se que as crianças que tiverem uma avaliação "Alterada" no teste APGAR no primeiro minuto têm 10 vezes mais chances de apresentarem alguma malformação congênita do que àquelas que apresentaram uma avaliação "Normal". Já o resultado do teste APGAR efetuado após 5 minutos é muito mais contundente que o primeiro, no sentido de ser um fator realmente de risco na previsão de malformação congênita. Para se ter uma ideia, uma criança que apresentou uma saúde comprometida, segundo o teste APGAR de 5 minutos, terá, em média, 182 vezes mais chances de desenvolver alguma anomalia congênita do que àquelas que não apresentaram comprometimento.

Por fim, a TABELA 3 ainda revela que o aumento de 1 ano na idade da mãe provoca um aumento de 36% no risco dessa mãe gerar um bebê com alguma deficiência congênita. Esse resultado, de certa forma, está coerente com as orientações da maioria dos ginecologistas, quando sugerem que gestações tardias podem potencializar a geração de filhos com problemas de malformação congênita.

TABELA 3: Estimativas dos parâmetros do modelo logístico e avaliação dos riscos

Parâmetros	Estimativa (β)	Erro Padrão	Estatística de Wald	g.l.	Valor-p	Risco - Exp(β)	IC 95% para Exp(β)
Intercepto	-15,115	1,952	59,965	1	0,0000	0,00	--
Ensino Superior (R)	--	--	--	--	--	--	--

Parâmetros	Estimativa (β)	Erro Padrão	Estatística de Wald	g.l.	Valor-p	Risco - Exp(β)	IC 95% para Exp(β)
Ensino Médio	-0,032	0,834	0,001	1	0,9697	0,97	(0,19 ; 4,97)
Não Usou Corticóide (R)	--	--	--	--	--	--	--
Usou Corticóide	1,653	0,469	12,409	1	0,0004	5,22	(2,08 ; 13,09)
Parto Cesárea (R)	--	--	--	--	--	--	--
Parto Normal	1,175	0,455	6,650	1	0,0099	3,24	(1,33 ; 7,90)
APGAR de 1 min. Normal (R)	--	--	--	--	--	--	--
APGAR de 1 min. Alterada	2,383	0,478	24,897	1	0,0000	10,83	(4,25 ; 27,62)
APGAR de 5 min. Normal (R)	--	--	--	--	--	--	--
APGAR de 5 min. Alterada	5,208	0,820	40,375	1	0,0000	182,64	(36,64 ; 910,33)
Idade da Mãe	0,310	0,048	41,403	1	0,0000	1,36	(1,24 ; 1,50)

Os resultados apresentados na TABELA 3 só serão válidos se as premissas do modelo logístico forem atendidas. Tais premissas se referem às características dos resíduos, tais como independência, normalidade e variância constante. As FIGURAS 3 E 4 confirmam essas suposições, isto é, os resíduos do modelo logístico proposto parecem ser aleatórios, normais (com limites entre ± 3) e com um padrão de variabilidade constante para os níveis das variáveis explicativas examinadas. Especificamente, para a variável idade, observa-se que há uma pequena inflação na variância dos resíduos para as idades mais elevadas, mas entende-se que esse efeito de heterocedasticidade é insipiente.

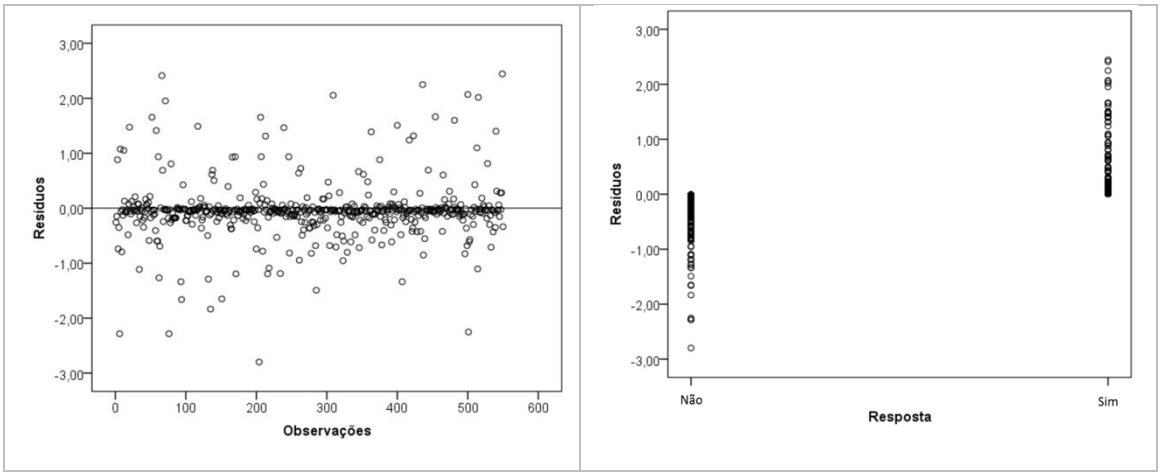


FIGURA 3: Distribuição dos resíduos *versus* a ordem das observações e à ocorrência de malformação congênita

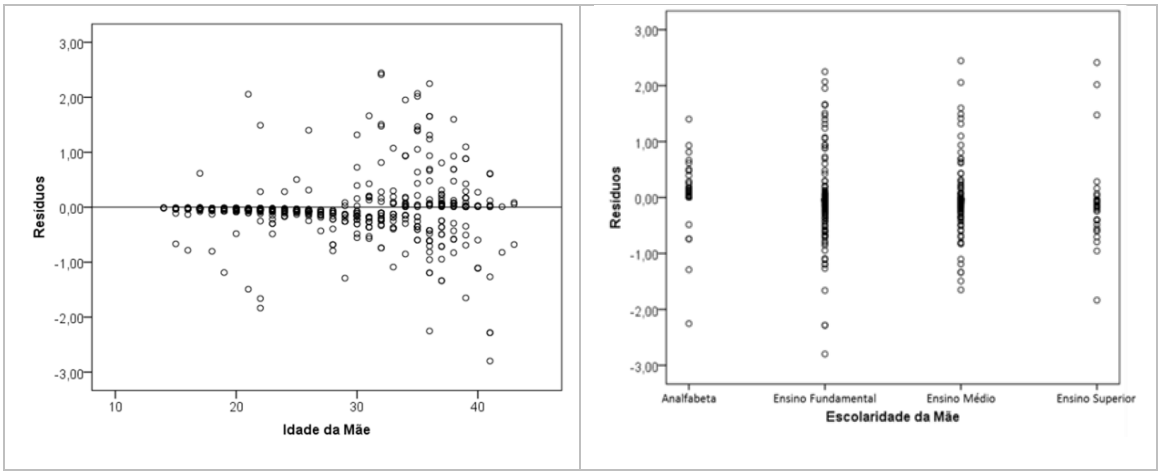


FIGURA 4: Distribuição dos resíduos versus algumas variáveis explicativas

Assim, uma vez que o modelo logístico ajustado mostrou-se adequado, uma medida que avalia a qualidade desse modelo é a comparação das taxas de acerto em relação aos grupos de crianças com ou sem malformação congênita. Desse modo, a TABELA 4 mostra que dos 135 registros de crianças com malformação congênita, 117 casos (86,67%) foram classificados corretamente pelo modelo logístico proposto; e dos 408 casos sem malformação congênita, 398 casos (97,55%) foram classificados corretamente. No geral o modelo logístico ajustado apresentou uma taxa de acerto de 94,84%, demonstrando uma elevada qualidade preditiva.

Apenas para expressar um resultado mais prático, a FIGURA 5 compara as probabilidades estimadas de ocorrência de malformação congênita das mães analfabetas e com ensino fundamental para as variações de idade, e ainda considerando constantes as demais variáveis. Observa-se, nesse gráfico, confirmando os resultados prévios, que as mães sem instrução, na faixa etária entre 20 e 40 anos, têm um risco (probabilidade) de gerar filhos com alguma anomalia congênita bem superior às mães com ensino superior, mostrando que o grau de instrução é fator decisivo para a ocorrência de bebês com malformação congênita.

TABELA 4: Qualidade do modelo logístico em termos do percentual de classificação correta

Malformação Congênita Observada	Malformação Congênita Prevista		**Total**	% de Acerto
	Sim	Não		
Sim	117	18	**135**	
Não	10	398	**408**	**94,84%**
Total	**416**	**127**	**543**	

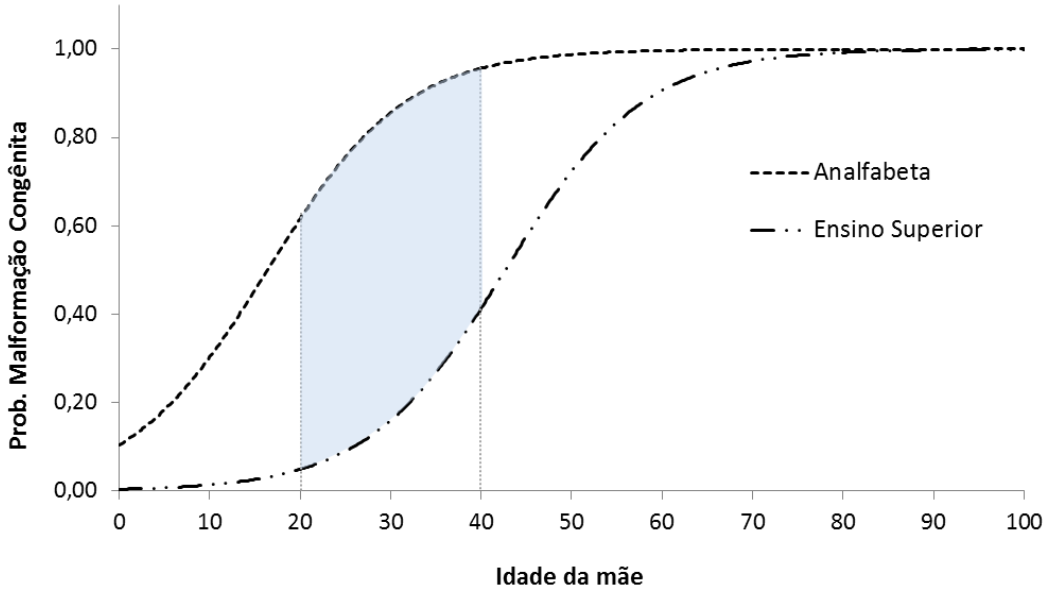


FIGURA 5: Comparação das probabilidades de ocorrer malformação congênita por idade para os grupos de mães analfabetas e com ensino superior

6. SUGESTÕES DE TRABALHOS FUTUROS

Para trabalhos futuros pretende-se reajustar o modelo de regressão logística, considerando agora outros fatores de risco ligados à mãe, tais como diabetes mellitus, hipertensão, tabagismo, alcoolismo, número de consultas pré-natal, entre outros. Esse novo modelo poderá servir, por exemplo, como um modelo classificador do risco do recém-nascido apresentar alguma malformação congênita.

7. CONSIDERAÇÕES FINAIS

Essa monografia apresentou uma aplicação dos modelos de regressão logística para avaliar os fatores de risco associados com a ocorrência de malformação congênita em crianças nascidas em um hospital de João Pessoa.

Dos fatores de risco disponíveis na base de dados utilizada nesse estudo, os resultados mostraram que a idade, a escolaridade da mãe, o uso de corticoides durante a gravidez, o tipo de parto e as medidas de APGAR de 1 e 5 minutos estavam associadas com a probabilidade de nascimento de filhos com alguma anomalia congênita. Adicionalmente, do ponto de vista estatístico, verificou-se, através da análise de resíduos, que o ajuste do modelo foi adequado e, além disso, que o modelo classificou corretamente mais de 93% dos casos examinados de malformação congênita.

Por fim, embora represente uma simples aplicação de um modelo de regressão logística, entende-se que o estudo desenvolvido nessa monografia poderá, sobretudo, fomentar aplicações mais sofisticadas para o tema abordado.

8. REFERÊNCIAS BIBLIOGRÁFICAS

[1] BARRETO, A. S. Modelos de Regressão: Teoria e Aplicação com o Programa Estatístico R, Edição do Autor, 1 Edição, p. 109-114, Brasília 2011.

[2] DEMETRIO, C. G. B. Modelos Lineares Generalizados em Experimentação Agronômica, Piracicaba – São Paulo, 2002.

[3] SECRETARIA DE SAUDE DA PARAIBA/– Sistema de Informação de Nascidos Vivos – TABNETSINASC. PB (http://www.saude.pb.gov.br).

[4] TURKMAN, M. A. A.; SILVA, G. L. Modelos Lineares Generalizados – da teoria à prática – Lisboa, 2000.

[5] NELDER, JOHN A;WEDDERBUM,ROBERT W.Generalized linear models". Journal of the Royal Statistical Society Series A (Journal of the Royal Statistical Society, Series A, Vol. 135, No. 3) (1972)

[6] FERREIRA, J.D. Exposições ambientais e leucemias na infância no Brasil: uma análise exploratória de sua associação Rev. bras. estud. popul. vol.29 no. 2 São Paulo July/Dec. 2012

[7] NETO, N.C, Souza ASR, Moraes Filho OB, Noronha AMB. Volume do líquido amniótico associado às anomalias fetais diagnosticadas em um centro de referência do nordeste brasileiro. Rev. Bras Ginecol. Obstet. 2009; 31(4): 164-70

[8] BRITO, V.R.S. Malformações Congênitas e Fatores de Risco Materno em Campina Grande - Paraíba Rev. Rene. Fortaleza, v. 11, n. 2, p. 27-36, abr./jun.

[9] AMORIM, M.M.R. Impacto das malformações congênitas na mortalidade perinatal e neonatal em uma maternidade-escola do Recife. Rev. Bras. Saúde Mater. Infanta. vol.6 suppl.1 Recife May 2006

[10] COSTA ,C.M.S.Perfil das Malformações congênitas em uma amostra de nascimentos no município do Rio de Janeiro 1999-2001. Cad. Saúde Pública vol.22 no.11 Rio de Janeiro Nov. 2006