UNIVERSIDADE FEDERAL DO CEARÁ
CENTRO DE CIÊNCIAS
DEPARTAMENTO DE ESTATÍSTICA E MATEMÁTICA APLICADA
CURSO DE ESTATÍSTICA

RENATA SOARES DA COSTA

TESTE DE DIAGNÓSTICO BASEADO EM ANÁLISE DE REGRESSÃO LOGÍSTICA

FORTALEZA
2013

RENATA SOARES DA COSTA

TESTE DE DIAGNÓSTICO BASEADO EM ANÁLISE DE REGRESSÃO LOGÍSTICA

Monografia apresentada ao curso de Estatística do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará, como requisito parcial para obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. José Ailton Alencar Andrade

FORTALEZA

2013

RENATA SOARES DA COSTA

TESTE DE DIAGNÓSTICO BASEADO EM ANÁLISE DE REGRESSÃO LOGÍSTICA

Monografia apresentada ao curso de Estatística do Departamento de Estatística e Matemática Aplicada da Universidade Federal do Ceará, como requisito parcial para obtenção do título de Bacharel em Estatística.

Aprovada em __/__/____.

BANCA EXAMINADORA

Prof. Dr. José Ailton Alencar Andrade (Orientador)
Universidade Federal do Ceará (UFC)

Prof[a]. Dr[a]. Maria Jacqueline Batista
Universidade Federal do Ceará (UFC)

Prof. Dr. Rafael Bráz Azevedo Farias
Universidade Federal do Ceará (UFC)

À minha mãe Madalena.

A meu irmão João Victor (in memoriam).

A meus amigos que estiveram sempre ao meu lado.

AGRADECIMENTOS

Primeiramente agradeço a Deus e ao Universo por permitir minha entrada na universidade, encontrar pessoas especiais e me dar forças para enfrentar os momentos difíceis ao longo da minha graduação.

À minha mãe Madalena, que nunca mediu esforços para proporcionar-me a melhor educação possível, seja ela moral ou escolar. Por sempre se preocupar com o meu bem-estar, além do seu apoio em situações difíceis.

Ao meu namorado Bruno, que tenho tanto amor e carinho, por estar sempre ao meu lado e dar todo o apoio nos momentos difíceis, além do grande incentivo técnico e emocional para a conclusão da faculdade.

À Universidade Federal do Ceará, por ter me proporcionado esse aprendizado imensurável, especialmente ao Departamento de Estatística e Matemática Aplicada. Faço meu agradecimento especial à todos os professores que tive a oportunidade de estudar, por seus ensinamentos transmitidos em cada aula. Agradeço especialmente ao professor Shiguemoto, por ter acreditado em mim no início do curso, além dos professores Maurício e Juvêncio, pois com eles pude aprender bastante. Agradeço também às funcionárias Luiza e Márgeri, por serem sempre solícitas e simpáticas comigo.

Ao professor Ailton, pela orientação nesta monografia, além das oportunidades concedidas e o carinho que teve por mim.

À Gauss - Empresa Júnior de Estatística, onde foi uma fonte de conhecimento ímpar para meu desenvolvimento profissional e pessoal. Aos meus colegas Gaussianos que tive a oportunidade de trabalhar junto, pois com eles pude aprender a essência de trabalhar em equipe e lidar com situações difíceis.

Aos meus colegas da faculdade, em especial os meus amigos Amanda, Chico Cícero, Chico Welington e Márcio, pelo apoio e companherismo ao longo da faculdade.

“Num piscar de olhos tudo se transforma. Tá vendo? Já passou! Mas ao mesmo tempo fica o sentimento de um mundo sempre igual, igual ao que já era.”

Humberto Gessinger

“Nada disso é tudo. Tudo isso é fundamental.”

Humberto Gessinger

RESUMO

O modelo de regressão logística é frequentemente utilizado em situações em que a variável resposta é de natureza dicotômica. Este é um caso particular dos modelos lineares generalizados, com componente aleatório binomial e função de ligação *logit*, e modela a probabilidade de um evento ocorrer em função de outras variáveis preditoras. Este trabalho tem como finalidade abordar a metodologia do modelo logístico, bem como ajustar um modelo para a predição de pré-eclâmpsia em mulheres grávidas, além de levar em consideração as técnicas de qualidade de ajuste do modelo de regressão, os testes de diagnóstico e a avaliação dos possíveis pontos *outliers*. O modelo ajustado atendeu às expectativas de qualidade do ajuste, tendo uma eficiência de aproximadamente 83% em discriminar mulheres que têm ou não pré-eclâmpsia, sensibilidade de 74,2% e especificidade de 73,9%. Tendo essas taxas como referência, concluiu-se que pacientes submetidas ao modelo estatístico que produzam probabilidade superior a 0,06 são classificadas como doentes.

Palavras-Chave: Análise de Regressão Logística, Pré-eclâmpsia.

ABSTRACT

The logistic regression model is often used in situations where the response variable is dichotomous nature. This is a particular case of generalized linear models with binomial random component and *logit* link, and models the probability of an event occurring as a function of other predictors. This work aims to address the methodology of the logistic model and set a model for the prediction of preeclampsia in pregnant women, besides taking into account the technical quality of fit of the regression model, the diagnostic tests and evaluation points of possible *outliers*. The adjusted model met the expectations of quality adjustment, with an efficiency of approximately 83% in discriminating women who do not have or preeclampsia, a sensitivity of 74,2% and specificity of 73,9%. With these rates as a reference, it is concluded that patients subjected to statistical model producing higher probability than 0,06 are classified as ill.

Keywords: Logistic Regression Analysis, Pre-eclampsia.

LISTA DE FIGURAS

LISTA DE TABELAS

SUMÁRIO

1 INTRODUÇÃO

A análise de regressão é uma técnica estatística utilizada para investigar e modelar o relacionamento entre variáveis. É bastante empregada na área de negócios e em pesquisas acadêmicas e é utilizada principalmente com o propósito de previsão, consistindo em determinar uma função matemática que busca descrever o comportamento de uma variável de interesse (variável resposta ou dependente) em função de outras variáveis explicativas (ou independentes). Nessas circunstâncias a variável resposta pode assumir qualquer valor no conjunto dos números reais. Porém, existem casos em que ela só pode assumir um entre dois resultados, frequentemente de natureza qualitativa. Nesta situação, percebe-se que é inviável a utilização do modelo linear. De fato, ocorrem muitas situações em que a variável resposta é de natureza binária ou qualitativa, como por exemplo: um cliente pode se tornar inadimplente ou não; um paciente pode ter determinada doença ou não; uma empresa pode ingressar em estado de falência; dentre outros.

Fenômenos como estes exigem modelagens que só admitem uma entre duas alternativas do tipo "ocorre ou não ocorre", "sim ou não". Nos exemplos citados o objetivo é explicar ou predizer a ocorrência de um determinado evento em função de um conjunto de variáveis, podendo ser categóricas ou não. Além da variável resposta ser de natureza categórica, ela exige resultados que possam ser expressos em termos de probabilidade. Para resolver problemas desse tipo, a regressão logística foi desenvolvida.

1.1 Motivação

A regressão logística tem uma vasta aplicação em diversas áreas do conhecimento. Na área médica é preciso reconhecer a doença (variável de efeito, o desfecho) e quais fatores estão envolvidos no seu aparecimento e na sua evolução (variáveis preditivas, fatores de exposição ou simplesmente exposição). Esses fatores podem representar associação entre exposição e doença, sendo assim, fatores de risco ou de proteção.

A pré-eclâmpsia é uma doença que ocorre apenas durante a gravidez, principalmente no últimos três meses de gestação, onde seu diagnóstico é normalmente feito no acompanhamento pré-natal. Ela pode ser leve ou grave e, quando se torna grave, pode afetar vários sistemas do corpo, reduzindo o fluxo de sangue para a placenta e tornando-se perigosa para o bebê. Além disso, a pré-eclâmpsia pode evoluir para a eclâmpsia, colocando a mãe e o bebê em riscos que podem ser fatais (MELCA, 2007).

Para auxiliar médicos no acompanhamento pré-natal, um modelo logístico que possibilite a previsão de pré-eclâmpsia pode ser construído como ferramenta auxiliar na detecção precoce da doença, podendo garantir uma melhor qualidade de vida para a mãe e para o bebê.

1.2 Objetivos

1.2.1 *Objetivo geral*

Utilizar o modelo de regressão logística para verificar a relação entre pré-eclâmpsia e variáveis preditivas suspeitas de serem fatores de risco para sua ocorrência.

1.2.2 *Objetivos específicos*

- Apresentar a metodologia da regressão logística;
- Descrever os principais testes de diagnóstico;
- Ajustar um modelo logístico que possibilite a previsão de pré-eclâmpsia;
- Avaliar a qualidade de ajuste do modelo através de testes e análise de resíduos e diagnóstico.

1.3 Estrutura do trabalho

Ao longo do trabalho são abordadas as metodologias utilizadas para o ajuste de modelos de regressão logística. Primeiramente, como é tratato no Capítulo 2, há uma abordagem dos Modelos Lineares Generalizados (MLG), na qual contém sua definição e principais propriedades, além da estimação de seus coeficientes, as principais técnicas de qualidade de ajuste, testes de hipóteses e seleção de modelos. Tais conceitos são fundamentais para o entendimento da regressão logística, já que ela é um caso particular

dos MLG. No Capítulo 3 é definido o conceito de regressão logística, com seu modelo matemático, a estimação e interpretação de seus coeficientes, as medidas de avaliação e análise de resíduos e diagnósticos do modelo logístico. Os testes de diagnóstico que são ferramentas auxiliares no ajuste do modelo (tais como a sensibilidade, a especifidade, a razão de probabildades e a curva ROC), são tratadas no Capítulo 4. No Capítulo 5 é realizada a aplicação da regressão logística para prever a ocorrência ou não de pré-eclâmpsia em pacientes, bem como a validação do modelo.

Os dados utilizados na aplicação foram retirados de Alves (2012), contando com o total de 487 pacientes atendidas no Laboratório de Medicina Materno-Fetal UECE/HGF, que se encontra no Hospital Geral de Fortaleza (HGF), em Fortaleza/Ceará. Para a escolha do modelo final foram utilizadas técnicas de seleção de variáveis, como o *stepwise*, *backward* e *forward*. Com o modelo selecionado, foi feita a interpretação dos resultados, bem como a avaliação e a devida análise de diagnóstico para que, no fim, o modelo ajustado possa ser utilizado como teste de diagnóstico para auxiliar no acompanhamento de mulheres grávidas. Por fim, serão apresentadas as considerações finais sobre esse estudo.

No Apêndice estão as tabelas e gráficos gerados na análise descritiva dos dados, além da tabela que contém os escores do modelo logístico ajustado e da tabela de classificação geral do modelo ajustado.

2 MODELOS LINEARES GENERALIZADOS

2.1 Introdução - Modelos Lineares Generalizados

Conforme Turkman e Silva (2000), por muito tempo os modelos normais lineares foram abordados para descrever a relação que uma ou mais variáveis explicativas têm sobre uma variável de interesse. Mesmo quando essa relação não apresentava uma resposta para a qual fosse razoável a suposição de normalidade, tentava-se algum tipo de transformação no sentido de alcançar a normalidade procurada. Porém, vários modelos não lineares ou não normais foram desenvolvidos para contornar situações que não eram adequadamente explicadas pelo modelo linear normal, tais como:

- O modelo complementar log-log para ensaios de diluição;
- Os modelos *probit* e *logit* para proporções;
- Os modelos log-lineares para dados de contagens;
- Os modelos de regressão para análise de sobrevivência, dentre outros.

Todos esses modelos apresentam uma estrutura não linear em um conjunto linear de parâmetros, mas que são linearizados, e têm em comum o fato da variável resposta seguir uma distribuição dentro de uma família de distribuições com propriedades muito específicas: a família exponencial.

Os Modelos Lineares Generalizados (MLG) introduzidos por Nelder e Wedderburn (1972) correspondem a uma síntese destes e de outros modelos, vindo assim unificar, tanto do ponto de vista teórico como conceitual, a teoria da modelação estatística até então desenvolvida. Os MLG desempenham um papel cada vez mais importante na análise estatística, devido ao grande número de modelos que englobam, apesar das limitações ainda

impostas, tais como as distribuições se restringem à família exponencial e por exigirem independência entre as observações da variável resposta (TURKMAN e SILVA, 2000).

2.2 Definição

Os MLG são uma extensão dos modelos lineares clássicos e contém os modelos mais importantes para respostas categóricas, bem como os modelos padrões de respostas contínuas. Eles são especificados por três componentes, que são comuns a todos eles: **(i)** uma componente aleatória, a qual identifica a distribuição de probabilidade da variável dependente; **(ii)** uma componente sistemática, que especifica uma função linear entre as variáveis independentes; **(iii)** e uma função de ligação que descreve a relação matemática entre a componente sistemática e o valor da componente aleatória.

As componentes dos MLG são dadas pelas seguintes definições (DEMÉTRIO, 2002):

(i) Componente aleatória

É representada pela variável resposta $(Y_1, Y_2, ..., Y_k)$ provenientes de uma mesma distribuição que faz parte da família exponencial na forma canônica com médias $\mu_1, \mu_2, ..., \mu_k$, ou seja:

$$\mathbb{E}(Y_i) = \mu_i, \ \ i = 1, 2, ..., n,$$

um parâmetro constante de escala, conhecido, $\phi > 0$ e que depende de um único parâmetro θ_i, chamado parâmetro canônico ou natural. A função densidade de probabilidade (f.d.p) de Y_i, que faz parte da família exponencial, tem a seguinte forma dada por McCullagh & Nelder (1989):

$$f_Y = (y_i; \theta_i, \phi) = \exp\left\{\frac{1}{a_i(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i; \phi)\right\}, \tag{2.1}$$

em que $b(\cdot)$ e $c(\cdot)$ são funções conhecidas. Em geral, $a_i = \frac{\phi}{w_i}$, sendo w_i os pesos a *priori*. Várias distribuições importantes são casos especiais desta família, sendo elas: Normal, Normal Inversa, Gama, Poisson, Binomial e Binomial Negativa.

A média e variância da componente aleatória é dada por:

$$\begin{aligned} \mu_i &= \mathbb{E}(Y) = b'(\theta_i) \\ \sigma^2 &= Var(Y) = a_i(\phi)b''(\theta_i) = a_i(\phi)V(\mu_i) = a_i(\phi)V_i, \end{aligned} \tag{2.2}$$

em que $b''(\theta_i) = \partial\mu_i/\partial\theta_i$ é uma função de μ_i e é representada por $V(\mu_i)$.

(ii) Componente sistemática

É definida pelas variáveis explicativas X_i's que entram na forma de uma soma linear de seus efeitos:

$$\eta_i = \sum_{i=1}^{n} \beta_j x_{ij} = \mathbf{x}_i^T \boldsymbol{\beta} \text{ ou } \boldsymbol{\eta} = \mathbf{X}\boldsymbol{\beta},$$

sendo $\mathbf{X} = (x_1, x_2, ..., x_n)^T$ a matriz do modelo, $\boldsymbol{\beta} = (\beta_1, \beta_2, ..., \beta_n)^T$ o vetor de parâmetros e $\boldsymbol{\eta} = (\eta_1, \eta_2, ..., \eta_n)^T$ o preditor linear.

(iii) Função de ligação

É a função que liga o componente aleatório ao componente sistemático do modelo linear, relacionando a média da resposta $\mu_i = \mathbb{E}(Y_i)$, ao preditor linear η_i, dada por:

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

A função $g(\cdot)$ é monótona e diferenciável. A função de ligação $g(\mu_i) = \mu_i$, chamada de função identidade, tem $\eta_i = \mu_i$. Esta especifica um modelo linear para a própria média e é a função de ligação para a regressão comum com Y normalmente distribuído. A função de ligação que transforma a média para o parâmetro natural é chamada de ligação canônica.

A escolha da função de ligação depende do tipo de resposta e do estudo em questão. Na Tabela 1 são apresentadas as principais funções de ligação utilizadas. A Tabela 2 apresenta algumas funções de ligação canônicas.

Tabela 1 - Funções de ligação

Funções de ligação	
Identidade	μ
Recíproca	$1/\mu$
Quadrática inversa	$1/\mu^2$
Raiz quadrada	$\sqrt{\mu}$
Expoente	$(\mu + c_1)^{c_2}$
Logarítmica	$\ln(\mu)$
Logit	$\ln[\mu/(1-\mu)]$
Complementar log-log	$\ln[-\ln(1-\mu)]$
Probit	$\Phi^{-1}(\mu)$

Fonte: Demétrio, 2002.

Tabela 2 - Funções de ligação canônicas

Distribuição	**Ligação canônica**
Normal	Identidade: $\eta = \mu$
Poisson	Logarítmica: $\eta = \ln(\mu)$
Binomial	Logística: $\eta = \ln\left(\frac{\pi}{1-\pi}\right) = \left(\frac{\mu}{m-\mu}\right)$
Gama	Recíproca: $\eta = \frac{1}{\mu}$
Normal inversa	Recíproca^2: $\eta = \frac{1}{\mu^2}$

Fonte: Demétrio, 2002.

2.3 Estimação dos coeficientes

Dada a definição de família exponencial em (2.1), pode-se denotar $\ell_i = \ln[f_Y(y_i; \theta_i, \phi)]$ indicando a contribuição de y_i no logaritmo da função de verossimilhança, isto é, o logaritmo da função de verossimilhança é $\ell = \sum_i \ell_i$. Então, de (2.1),

$$\ell_i = \frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i; \phi). \tag{2.3}$$

Para N observações independentes, de (2.3) o logaritmo da função de verossimilhança é

$$\ell(\boldsymbol{\beta}) = \sum_i \ell_i = \sum_i \ln[f(y_i; \theta_i, \phi)] = \sum_i \left\{ \frac{1}{a(\phi)}[y_i\theta_i - b(\theta_i)] + c(y_i; \phi) \right\}. \tag{2.4}$$

A notação $\ell(\boldsymbol{\beta})$ reflete a dependência de θ nos parâmetros $\boldsymbol{\beta}$ do modelo.

Conforme Demétrio (2002), uma propriedade da família exponencial de distribuições é que seus elementos satisfazem a condições de regularidade suficientes para assegurar que o máximo global do logaritimo da função de verossimilhança ℓ é dado unicamente pela solução do sistema de equações $\mathbf{U}_{\boldsymbol{\theta}} = \frac{\partial \ell}{\partial \boldsymbol{\theta}} = \mathbf{0}$ ou equivalentemente, $\mathbf{U}_{\boldsymbol{\beta}} = \frac{\partial \ell}{\partial \boldsymbol{\beta}} = \mathbf{0}$. Então, a função escore é dada por:

$$U_j = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_i \frac{\partial \ell_i}{\partial \beta_j} = 0, \forall\ j.$$

Para diferenciar (2.4), pode-se usar a regra da cadeia,

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{\partial \ell_i}{\partial \theta_i}\frac{\partial \theta_i}{\partial \mu_i}\frac{\partial \mu_i}{\partial \eta_i}\frac{\partial \eta_i}{\partial \beta_j}. \tag{2.5}$$

Já que $\partial \ell_i / \partial \theta_i = [y_i - b'(\theta_i)]/a_i(\phi)$ e que $\mu_i = b'(\theta_i)$ e $Var(Y_i) = b''(\theta_i)a_i(\phi)$ de (2.2),

$$\frac{\partial \ell_i}{\partial \theta_i} = (y_i - \mu_i)/a_i(\phi)$$

$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i)\frac{Var(Y_i)}{a_i(\phi)}.$$

Sabe-se também que $\eta_i = \sum_j \beta_j x_{ij}$,

$$\frac{\partial \eta_i}{\partial \beta_j} = x_{ij}.$$

Finalmente, sabendo que $\eta_i = g(\mu)$, $\partial \mu_i / \partial \eta_i$ depende da função de ligação para o modelo. Em resumo, substituindo (2.5), a seguinte forma é dada:

$$\frac{\partial \ell_i}{\partial \beta_j} = \frac{y_i - \mu_i}{a_i(\phi)}\frac{a_i(\phi)}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}x_{ij} = \frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i}. \tag{2.6}$$

Logo, as equações de verossimilhança são

$$U_j = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{N}\frac{(y_i - \mu_i)x_{ij}}{Var(Y_i)}\frac{\partial \mu_i}{\partial \eta_i} = 0,\ j = 1, ..., p. \tag{2.7}$$

Embora β não apareça nessas equações, ele está lá através de μ_i, já que $\mu_i = g^{-1}(\sum_j \beta_j x_{ij})$. Funções de ligação diferentes produzem diferentes conjuntos de equações. As equações de verossimilhança em (2.7) dependem da distribuição de Y_i através de μ_i e $Var(Y_i)$. A variância em si depende da média através da forma funcional particular $Var(Y_i) = \upsilon(\mu_i)$ (AGRESTI, 2002).

Em geral, as equações $U_j = 0$, $j = 1, 2, ...p$ não são lineares e têm que ser resolvidas numericamente por processos iterativos, por exemplo, o de Newton-Raphson, Escore de Fisher, Estimador de Momentos, entre outros.

O método escore de Fisher fornece a seguinte solução:

$$\widehat{\boldsymbol{\beta}}^{(m+1)} = (\mathbf{X'W}^{(m)}\mathbf{X})^{-1}\mathbf{X'W}^{(m)}\mathbf{z}^{(m)} \tag{2.8}$$

em que $\mathbf{X}$ é matriz do modelo; $\mathbf{W}$ é a matriz diagonal dos pesos, em que $W_i = \frac{w_i}{V(\mu_i)}\left(\frac{\partial \mu_i}{\partial \eta_i}\right)$ e w_i peso a priori; e $\mathbf{z}^{(m)} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{(m)} + \Delta^{(m)}(\mathbf{y} - \mu)^{(m)} = \eta^{(m)} + \Delta^{(m)}(\mathbf{y} - \mu)^{(m)}$, chamada de variável dependente ajustada no passo m do processo iterativo, onde $\Delta = diag\left(\frac{\partial \eta_i}{\partial \mu_i}\right)$.

A expressão (2.8) tem a forma da solução das equações normais, para o modelo linear obtida pelo método dos mínimos quadrados ponderados, exceto que nesse caso a solução $\widehat{\boldsymbol{\beta}} = \boldsymbol{\beta}^{(m+1)}$ é obtida por processo numérico iterativo (lembrando que a expressão (2.8) independe de ϕ).

2.4 Qualidade de ajustamento

2.4.1 *Função desvio*

Dadas n observações, a essas podem ser ajustados modelos contendo até n parâmetros. O modelo mais simples é o **modelo nulo** que tem um único parâmetro, representado por um valor μ comum a todos os dados. No outro extremo, está o **modelo saturado ou completo** que tem n parâmetros, um para cada observação e é composto por todas as combinações das variáveis explanatórias. Conforme Demétrio (2002), na prática o modelo nulo é simples demais e o modelo saturado não é informativo, pois não resume os dados, mas simplesmente os repetem. Existem, contudo, dois outros modelos limitantes, porém, menos extremos. O **modelo minimal** contém o menor número de termos necessários para o ajuste. Por outro lado, o modelo que contém o maior número de termos que podem ser considerados é chamado de **modelo maximal**.

Em geral, trabalha-se com modelos encaixados e o conjunto de matrizes dos modelos pode, então, ser formado pela adição sucessiva de termos ao modelo minimal até se chegar ao modelo maximal. Qualquer modelo com p parâmetros linearmente independentes, situado entre os modelos minimal e maximal, é chamado **modelo corrente ou modelo sob pesquisa**. O problema é determinar a utilidade de um parâmetro extra no modelo corrente (sob pesquisa) ou, então, verificar a falta de ajuste induzida pela omissão dele. A fim de discriminar entre modelos, medidas de discrepância devem ser introduzidas para medir o ajuste de um modelo (DEMÉTRIO, 2002).

Proposto por Nelder e Wedderburn (1972), a medida de discrepância *deviance* (traduzida como desvio por Cordeiro (1986)), com expressão dada por:

$$S_p = 2(\widehat{\ell}_n - \widehat{\ell}_p),$$

em que $\widehat{\ell}_n$ é o máximo do logaritmo da função de verossimilhança para o modelo saturado e $\widehat{\ell}_p$ para o modelo corrente. Quanto menor for o valor de S_p, melhor será o ajuste do modelo aos dados. Na prática, contenta-se em testar a adequação de um modelo linear generalizado, sem muito rigor, comparando-se o valor S_p com os percentis da distribuição qui-quadrado com $n - p$ graus de liberdade e nível de significância α, ou seja, $\chi^2_{n-p;\alpha}$. Assim, nos casos em que é possível a aproximação de uma $\chi^2_{n-p;\alpha}$, tem-se que se

$$S_p \leq \chi^2_{n-p;\alpha},$$

pode-se considerar que existem evidências, a um nível aproximado de $100\alpha\%$ de proba-

bilidade, que o modelo proposto está bem ajustado aos dados (DEMÉTRIO, 2002).

2.4.2 *Estatística de Pearson generalizada*

A estatística de Pearson X^2 generalizada é outra medida da discrepância de ajuste de um modelo a um conjunto de dados e é expressa por:

$$X^2 = \sum_{i=1}^{n} w_i \frac{(y_i - \widehat{\mu}_i)^2}{V(\widehat{\mu}_i)}$$

sendo w_i os pesos a *priori* e $V(\widehat{\mu}_i)$ a função de variância estimada sob o modelo que está sendo ajustado. Nos casos de dados provenientes das distribuições binomial e de Poisson, em que $\phi = 1$, X^2 é a estatística original de Pearson, muito utilizada na análise dos modelos logístico e log-linear para tabelas multidimensionais e que pode ser escrita na forma

$$X^2 = \sum_{i=1}^{n} \frac{(O_i - E_i)^2}{E_i},$$

em que O_i é a frequência observada e E_i a frequência esperada.

2.5 Testes de hipóteses

Nos MLG os métodos de inferência são baseados, fundamentalmente, na teoria de máxima verossimilhança. De acordo com esta teoria, existem três estatísticas para testar hipóteses relativas aos parâmetros $\boldsymbol{\beta}$'s, que são deduzidas de distribuições assintóticas de funções adequadas das estimativas dos $\boldsymbol{\beta}$'s. São elas:

1. Teste da razão de verossimilhanças;
2. Teste de Wald;
3. Teste de escore.

Nas quais são assintoticamente equivalentes e, sob H_0 e para ϕ conhecido, convergem para uma variável com distribuição χ^2_p. No caso, a razão de verossimilhanças é o critério que define um teste uniformemente mais poderoso.

Supondo o interesse em testar as seguintes hipóteses:

$$H_0 : \boldsymbol{\beta} = \boldsymbol{\beta}_0 \text{ contra } H_1 : \boldsymbol{\beta} \neq \boldsymbol{\beta}_0,$$

em que $\boldsymbol{\beta}_0$ é um vetor p-dimensional conhecido e ϕ é também assumido conhecido. As três estatísticas são:

1. A estatística da razão de verossimilhança que é definida como

$$\Lambda = -2\ln\left[\frac{L(\boldsymbol{\beta}_0)}{L(\widehat{\boldsymbol{\beta}})}\right] = 2[\ell(\boldsymbol{\beta}) - \ell(\boldsymbol{\beta}_0)],$$

   em que $\widehat{\boldsymbol{\beta}}$ é o estimador da máxima verossimilhança sob todo o espaço paramétrico.

2. A estatística Wald que é dada por:

$$W = (\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0)^T \mathbf{I}(\widehat{\boldsymbol{\beta}})(\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0),$$

   em que $\mathbf{I}(\widehat{\boldsymbol{\beta}})$ é a matriz de informação de Fisher avaliada em $\widehat{\boldsymbol{\beta}}$.

3. A estatística Escore que é dada por:

$$E_S = \mathbf{U}^T(\boldsymbol{\beta}_0) I(\boldsymbol{\beta}_0)^{-1} \mathbf{U}(\boldsymbol{\beta}_0),$$

   em que $I(\boldsymbol{\beta}_0)^{-1}$ é a matriz de informação avaliada em $\boldsymbol{\beta}_0$.

2.6 Seleção de modelos

Um problema importante em muitas aplicações da análise de regressão envolve selecionar o conjunto de variáveis regressoras a ser usado no modelo. O interesse é filtrar as variáveis candidatas para obter um modelo que contenha o melhor subconjunto de regressores, logo, o ideal é que o modelo final tenha regressores suficientes de modo que ele desempenhe satisfatoriamente seu uso pretendido, que é o da previsão. Existem vários procedimentos para a seleção de modelos de regressão, dentre eles os mais conhecidos são: maior R_p^2, menor s_p^2, C_p, *forward*, *backward*, *stepwise*, *AIC* e *BIC*. Alguns deles serão descritos brevemente a seguir.

2.6.1 *Método forward*

Esse método parte da suposição de que não há variável no modelo, apenas o intercepto. A ideia do método é adicionar uma variável de cada vez. Segundo Paula (2010), o método é iniciado pelo modelo $\mu = \beta_0$. Para cada variável explicativa é ajustado o modelo

$$\mu = \beta_0 + \beta_j x_j, (j = 1, ..., q).$$

É testada a hipótese $H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$. Seja P o menor nível descritivo dentre os q testes e P_E o nível descritivo de entrada. Se $P \leq P_E$, a variável correspondente entra no modelo. Supondo que X_1 tenha sido selecionada, então, no passo seguinte são ajustados os modelos

$$\mu = \beta_0 + \beta_1 x_1 + \beta_j x_j, (j = 2, .., q).$$

Testa-se a hipótese $H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$. Sendo P o menor nível descritivo dentre os $(q-1)$ testes. Se $P \leq P_E$, a variável correspondente entra no modelo. O procedimento é repetido até que ocorra $P > P_E$.

2.6.2 *Método backward*

Enquanto o método *forward* começa sem nenhuma variável no modelo e adiciona variáveis a cada passo, o método *backward* faz o caminho oposto. Ele incorpora inicialmente todas as variáveis e depois, por etapas, cada uma pode ser ou não eliminada. Logo, o procedimento é iniciado pelo modelo

$$\mu = \beta_0 + \beta_1 x_1 + ... + \beta_q x_q.$$

Testa-se a hípótese $H_0 : \beta_j = 0$ contra $H_1 : \beta_j \neq 0$ para $j = 1, ..., q$. Seja P o maior nível descritivo dentre os q testes e P_S o nível descritivo de saída. Se $P > P_S$, a variável correspondente sai do modelo. Supondo que X_1 tenha saído do modelo. Então, ajusta-se o modelo

$$\mu = \beta_0 + \beta_2 x_2 + ... + \beta_q x_q.$$

A hipótese $H_0 : \beta_j$ contra $H_1 : \beta_j \neq 0$ é testada, para $j = 2, ..., q$. Seja P o maior nível descritivo dentre os $(q-1)$ testes. Se $P > P_S$, então a variável correspondente sai do modelo. O procedimento é repetido até que ocorra $P \leq P_S$.

2.6.3 *Método stepwise*

Esse procedimento constrói iterativamente uma sequência de regressão pela adição ou remoção de variáveis em cada etapa. Ou seja, é uma mistura dos métodos *forward* e *backward*.

O processo é iniciado com o modelo $\mu = \beta_0$. Após duas variáveis terem sido incluídas no modelo, é verificado se a primeira não sai do modelo. O processo continua até que nenhuma variável seja incluída ou excluída do modelo. Geralmente é adotado $0,15 \leq P_E$,

$P_S \leq 0,25$, porém uma sugestão seria usar $P_E = P_S = 0,20$ (PAULA, 2010).

2.6.4 *Critério de Informação de Akaike*

O Critério de Informação de Akaike (AIC) foi proposto por Akaike (1974). A ideia básica deste método é selecionar um modelo que seja parcimonioso. Ele se diferencia dos procedimentos anteriores por ser um processo de minimização que não envolve testes estatísticos e retorna um valor exato, no qual é uma medida para avaliar a qualidade do ajuste de um modelo de regressão. Quanto menor for esse valor encontrado, melhor será o ajuste do modelo. O AIC é baseado no logaritmo da função de verossimlhança, sendo representado por

$$AIC = -2\ell(\widehat{\boldsymbol{\beta}}) + 2p,$$

em que p é o número de parâmetros no modelo e $\ell(\widehat{\boldsymbol{\beta}})$ representa o logaritmo da função de verossimilhança.

2.6.5 *Critério de Informação Bayesiano*

O Critério de Informação Bayesiano (BIC) foi proposto por Schwarz (1978), tendo esse nome por possuir argumentos bayesianos em sua forma de avaliação. Ele é bem parecido com o AIC, já que também retorna um valor exato, sendo este uma medida para avaliar a qualidade do ajuste do modelo. Ele também é baseado na função de verossimilhança e não envolve testes estatísticos. Porém, ao analisar a estrutura de penalidade, ele é mais rigoroso que o AIC, pois em alguns modelos o BIC é sensível ao aumento da verossimilhança. Portanto, ele penaliza mais fortemente que o AIC a introdução adicional de parâmetros. O critério bayesiano é dado por

$$BIC = -2\ell(\widehat{\boldsymbol{\beta}}) + p \ln n,$$

sendo p o número de parâmetros do modelo, $\ell(\widehat{\boldsymbol{\beta}})$ representa o logaritmo da função de verossimilhança e n é o número de observações na amostra.

3 MODELO DE REGRESSÃO LOGÍSTICA

A regressão logística é uma forma de modelagem estatística que é frequentemente utilizada em situações em que a variável resposta é de natureza dicotômica. Isso exige que o resultado da análise possibilite associações a certas categorias, descrevendo a relação entre a variável resposta categórica e um conjunto de variáveis explicativas, nas quais podem ser categóricas ou métricas. De acordo com Figueira (2006), a variável resposta é geralmente dicotômica, mas pode ser politômica, isto é, têm mais do que dois níveis de resposta. As categorias (ou valores) que a variável dependente assume podem ser de natureza nominal ou ordinal. Na situação de natureza ordinal, há uma ordem natural entre as possíveis categorias e, então, se tem o contexto da Regressão Logística Ordinal. Quando não existe esta ordem entre as categorias da variável dependente, tem-se o contexto da Regressão Logística Nominal.

O modelo de regressão logística é uma extensão da análise de tabelas de múltipla entrada para a estrutura de análise de regressão, na qual se modelam os resultados de probabilidades binomiais, além de que é um caso particular dos modelos lineares generalizados com componente aleatório binomial e função de ligação *logit*.

3.1 Modelo de Regressão Logística Simples

Inicialmente considera-se que $\pi(x)$ é uma função monotônica com valores entre zero e um, quando x varia na reta real, ou seja, $\pi(x)$ é uma função de distribuição de probabilidade, na qual representa a probabilidade de "sucesso", dado o valor de x de uma variável explicativa qualquer. A variável resposta Y é dicotômica, assumindo o valor 1 para o evento de interesse (sucesso) e o valor 0 para evento complementar (fracasso).

Considera-se uma série de eventos binários, onde $(Y_1, Y_2, ..., Y_n)$ são variáveis aleatórias independentes com distribuição Bernoulli, com probabilidade de sucesso $\pi(x)$, isto é,

$Y_i \sim Ber(\pi(x))$, como visto em Souza (2006). Desta forma, $\pi(x) = P(Y = 1|X = x) = 1 - P(Y = 0|X = x)$ e $0 < \pi(x) < 1$. Como $\pi(x)$ varia entre zero e um, uma representação linear para ela sobre todos os valores de x não é adequada, então se considera a transformação logística de $\pi(x)$ sob a forma linear:

$$\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \text{logit}[\pi(x)] = \beta_0 + \beta_1 x, \tag{3.1}$$

ou equivalentemente

$$\pi(x) = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}, \tag{3.2}$$

sendo β_0 e β_1 parâmetros desconhecidos. Conforme $x \to \infty$, $\pi(x) \downarrow 0$ quando $\beta_1 < 0$, e $\pi(x) \uparrow 1$ quando $\beta_1 > 0$. Enquanto $\pi(x)$ deve cair no intervalo $(0, 1)$, o logit$[\pi(x)]$ pode ser qualquer número real (AGRESTI, 2002).

Conforme Souza (2006), em qualquer problema de regressão a quantidade a ser modelada é a esperança da variável aleatória dependente, dado o valor da variável independente, ou seja, $\mathbb{E}(Y|X = x)$. Devido à natureza da variável resposta, $0 \leq \mathbb{E}(Y|X = x) \leq 1$ na regressão logística, enquanto que na regressão linear $-\infty \leq \mathbb{E}(Y|X = x) \leq \infty$. Na regressão linear, $\mathbb{E}(Y|X = x) = \beta_0 + \beta_1 x$ e na regressão logística, usando a definição de variáveis aleatórias discretas, tem-se que

$$\mathbb{E}(Y|X = x) = 1P(Y = 1|X = x) + 0P(Y = 0|X = x) = \pi(x).$$

Outra diferença importante entre o modelo de regressão linear e o modelo de regressão logístico refere-se à distribuição condicional de Y. No modelo linear assume-se que uma observação possa ser expressa como $y = \mathbb{E}(Y|x) + \epsilon$, com a suposição de que o erro ϵ tenha distribuição Normal com média zero e variância constante. Este não é o caso quando Y é dicotômica. Desta maneira, uma observação pode ser expressa como $y = \pi(x) + \epsilon$, em que ϵ pode assumir uma de duas possibilidades: se $y = 1$, então $\epsilon = 1 - \pi(x)$ com probabilidade $\pi(x)$, e se $y = 0$, $\epsilon = -\pi(x)$ com probabilidade $1 - \pi(x)$. Portanto, ϵ tem distribuição Bernoulli com média zero e variância $\pi(x)[1 - \pi(x)]$.

Para cada x_i, para $i \in \{1, ..., n\}$, segue-se que $\mathbb{E}(Y_i|x_i) = \pi(x_i)$. Cada observação y_i pode ser interpretada, para cada $i \in \{1, ..., n\}$, como

$$y_i = \frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)} + \epsilon_i,$$

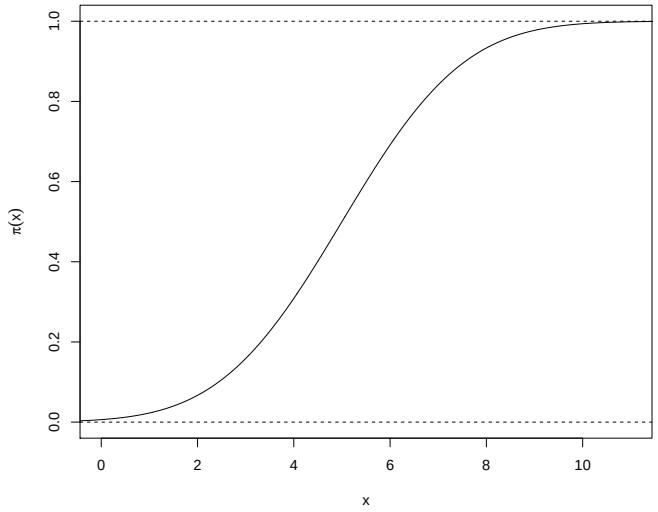
onde os erros ϵ_i seguem as seguintes suposições, para todo $i, l \in \{1, ..., n\}$ (FIGUEIRA,

2006):

$$(i)\ \mathbb{E}(\epsilon_i|x_i) = 0.$$
$$(ii)\ Var(\epsilon_i|x_i) = \pi(x_i)[1 - \pi(x_i)].$$
$$(iii)\ Cov(\epsilon_i, \epsilon_l) = 0, se\ i \neq l.$$

Devido à estimação de probabilidades estar compreendida nos limites (0,1), é de se esperar que as mudanças ocorridas na variável independente produzam efeitos cada vez menores sobre a variável resposta à medida que ela assuma valores mais próximos dos extremos. A curva de regressão monótona apresentada na Figura 1 tem o formato de uma função distribuição acumulada (f.d.a) para uma variável aleatória contínua quando $\beta_1 > 0$. Isso sugere um modelo para resposta binária tendo a forma $\pi(x) = F_X(x)$ de alguma f.d.a F_X.

Figura 1 - Curva do modelo logístico



Quando $\beta_1 > 0$, a curva de regressão logística é uma função de distribuição acumulada da distribuição logística. A f.d.a de uma distribuição logística com média μ e parâmetro de dispersão $\tau > 0$ é da seguinte forma:

$$F_X(x) = \frac{\exp[(x-\mu)/\tau]}{1 + \exp[(x-\mu)/\tau]}, -\infty < x < \infty.$$

A forma padronizada da f.d.a logística tem $\mu = 0$ e $\tau = 1$. Seja $\Phi(\cdot)$ denotando uma f.d.a padrão, então $\Phi(x) = e^x/(1 + e^x)$. Para esta função, a curva de regressão logística tem a forma $\pi(x) = \Phi(\beta_0 + \beta_1 x)$. A transformação *logit* é simplesmente a função inversa da

função distribuição acumulada logística padrão; isto é, quando $\Phi(x) = \pi(x) = e^x/(1+e^x)$, então $x = \Phi^{-1}[\pi(x)] = \ln[\pi(x)/(1-\pi(x))]$ (AGRESTI, 2002).

3.1.1 *Estimação dos coeficientes*

Identificada a equação que permite calcular a probabilidade relativa à ocorrência de determinado evento, resta estimar os seus coeficientes. Devido à variância não constante, o estimador de máxima verossimilhança é mais eficiente do que o estimador de mínimos quadrados. O mecanismo da estimação do logaritmo da função de verossimilhança e ajuste de modelo para regressão logística são casos especiais dos resultados de ajuste para MLG na Seção 2.3.

Supõe-se uma amostra de n observações independentes de pares (x_i, y_i) para $i \in \{1, ..., n\}$, em que y_i representa o valor da variável aleatória Y e x_i o valor da variável independente para a i-ésima observação. Seja $\boldsymbol{\beta} = (\beta_0, \beta_1)$ o vetor de parâmetros relacionado com a probabilidade condicional $P(Y_i = 1|x_i) = \pi(x_i)$. Conforme Figueira (2006), a função de verossimilhança para o modelo logístico é dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(x_i)^{y_i}[1 - \pi(x_i)]^{(1-y_i)}. \tag{3.3}$$

O princípio da máxima verossimilhança é estimar o valor de $\boldsymbol{\beta}$ que maximiza $L(\boldsymbol{\beta})$. O logaritmo da função de verossimilhança é dado por:

$$\begin{aligned}
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} [y_i \ln \pi(x_i) + (1 - y_i) \ln(1 - \pi(x_i))] \\
&= \sum_{i=1}^{n} [y_i \ln(\pi(x_i)) + \ln(1 - \pi(x_i)) - y_i \ln(1 - \pi(x_i))] \\
&= \sum_{i=1}^{n} \left[y_i \ln \left(\frac{\pi(x_i)}{1 - \pi(x_i)} \right) + \ln(1 - \pi(x_i)) \right]
\end{aligned}$$

$$
\begin{aligned}
&= \sum_{i=1}^{n} \left\{ y_i \ln \left[\frac{\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}}{1-\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}} \right] + \ln \left[1 - \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} \right] \right\} \\
&= \sum_{i=1}^{n} \left\{ y_i \ln \left[\frac{\frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}}{\frac{1}{1+\exp(\beta_0+\beta_1 x_i)}} \right] + \ln \left(\frac{1}{1+\exp(\beta_0+\beta_1 x_i)} \right) \right\} \\
&= \sum_{i=1}^{n} \{ y_i \ln[\exp(\beta_0+\beta_1 x_i)] - \ln[1+\exp(\beta_0+\beta_1 x_i)] \} \\
\ell(\boldsymbol{\beta}) &= \sum_{i=1}^{n} \{ y_i(\beta_0+\beta_1 x_i) - \ln[1+\exp(\beta_0+\beta_1 x_i)] \} .
\end{aligned}
$$

Para encontrar o valor de $\boldsymbol{\beta}$ que maximiza $\ell(\boldsymbol{\beta})$, deriva-se $\ell(\boldsymbol{\beta})$ em relação a cada parâmetro (β_0, β_1), obtendo-se as equações

$$
\begin{aligned}
\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0} &= \sum_{i=1}^{n} \left[y_i - \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)} \right] \\
\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1} &= \sum_{i=1}^{n} \left[y_i x_i - \frac{\exp(\beta_0+\beta_1 x_i) x_i}{1+\exp(\beta_0+\beta_1 x_i)} \right] .
\end{aligned}
$$

Então o estimador de $\boldsymbol{\beta}$ pelo método da máxima verossimilhança, denotado por $\widehat{\boldsymbol{\beta}}$, é a solução das equações de verossimilhança

$$
\sum_{i}^{n} [y_i - \pi(x_i)] = 0 \tag{3.4}
$$

$$
\sum_{i=1}^{n} x_i [y_i - \pi(x_i)] = 0. \tag{3.5}
$$

Como as equações apresentadas em (3.4) e (3.5) não são lineares, para solucioná-las faz-se necessário o uso de algum método de aproximação, sendo o método interativo de Newton-Raphson utilizado para tal fim.

3.1.2 *Interpretação dos coeficientes*

Paula (2010) sugere que o modelo em (3.1) poderia, por exemplo, ser aplicado para analisar a associação entre uma determinada doença e a ocorrência ou não de um fator particular. Seriam então amostrados, independentemente, n_1 indivíduos com presença do fator $(x = 1)$ e n_2 indivíduos com ausência do fator $(x = 0)$ e $\pi(x)$ seria a probabilidade de desenvolvimento da doença após certo período fixo. Dessa forma, a chance de

desenvolvimento da doença para um indivíduo com presença do fator fica dada por

$$\frac{\pi(1)}{1-\pi(1)} = e^{\widehat{\beta}_0+\widehat{\beta}x},$$

enquanto que a chance de desenvolvimento da doença para um indivíduo com ausência do fator é simplesmente

$$\frac{\pi(0)}{1-\pi(0)} = e^{\widehat{\beta}_0}.$$

A razão de chances é um método bastante utilizado em tabelas de contingência, sendo seu objetivo estabelecer a relação de duas chances sobre uma determinada característica em comum. Logo, a razão de chances (ψ) é dada por

$$\psi = \frac{\frac{\pi(1)}{1-\pi(1)}}{\frac{\pi(0)}{1-\pi(0)}} = \frac{\pi(1)[1-\pi(0)]}{\pi(0)[1-\pi(1)]} = e^{\widehat{\beta}_1}.$$

Dependendo apenas do parâmetro $\widehat{\beta}_1$. Assim, sendo $\psi = e^{\widehat{\beta}_1}$, ele dá o quanto o evento $x = 1$ contribui a mais para um aumento/decréscimo de $\pi(x)$, em relação ao evento $x = 0$.

O intervalo de confiança para a razão de chances é dado por

$$IC_{1-\alpha}(\psi) = \exp[\widehat{\beta}_1 \pm z_{\alpha/2} EP(\widehat{\beta}_1)],$$

em que $z_{\alpha/2}$ é o quantil da distribuição normal e $EP(\widehat{\beta}_1)$ o erro padrão de $\widehat{\beta}_1$.

3.2 Modelo de Regressão Logística Múltipla

Aqui será generalizado o modelo logístico para o caso de mais de uma variável independente, ou seja, o caso múltiplo. Considera-se um conjunto de p variáveis independentes denotadas por $\mathbf{X} = (X_1, ..., X_p)^T$, em que $\mathbf{x} = (x_1, ..., x_p)^T$ é um valor particular e uma v.a dependente binária Y. Denotando por $\boldsymbol{\beta} = (\beta_0, \beta_1, ..., \beta_p)^T$ o vetor de parâmetros desconhecidos e β_j sendo o j-ésimo parâmetro associado à variável explicativa x_j, com $j = 0, 1, ..., p$, então a transformação logística de $\pi(\mathbf{x})$ sob a forma linear é dada por

$$\ln\left[\frac{\pi(\mathbf{x})}{1-\pi(\mathbf{x})}\right] = \text{logit}[\pi(\mathbf{x})] = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \tag{3.6}$$

Então, o modelo para predizer $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$ é

$$\pi(\mathbf{x}) = \frac{\exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}{1 + \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p)}. \tag{3.7}$$

Conforme Figueira (2006), dadas n observações independentes de Y, denotadas por $(y_1, ..., y_n)$, associadas aos valores de $\mathbf{x}_i = (x_{i1}, ..., x_{ip})$, para $i \in \{1, ..., n\}$, o *logit* dado pela equação (3.6) apresenta-se da forma

$$\text{logit}[\pi(\mathbf{x_i})] = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \epsilon_i,$$

onde os erros ϵ_i seguem as seguintes suposições, para todo $i, l \in \{1, ..., n\}$:

$$\begin{aligned} &(i)\ \mathbb{E}(\epsilon_i|\mathbf{x}_i) = 0. \\ &(ii)\ Var(\epsilon_i|\mathbf{x}_i) = \pi(\mathbf{x}_i)[1 - \pi(\mathbf{x}_i)]. \\ &(iii)\ Cov(\epsilon_i, \epsilon_l) = 0, se\ i \neq l. \end{aligned} \tag{3.8}$$

Desta forma, as v.a's $Y_1, ..., Y_n$ satisfazem um modelo logístico múltiplo se uma amostra de tamanho um de cada Y_i pode ser expressa como

$$y_i = \frac{\exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})}{1 + \exp(\beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip})} + \epsilon_i,$$

para qual x_{ij} é constante conhecida, β_j é parâmetro desconhecido do modelo e os erros ϵ_i possuem as suposições dadas em (3.8).

3.2.1 *Estimação dos coeficientes*

Para a estimação dos parâmetros foi utilizado o método de máxima verossimilhança similar ao caso do modelo logístico simples. No caso do modelo múltiplo, a função de verossimilhança é igual à expressão (3.3), com a modificação de que $\pi(\cdot)$ é dada pela expressão (3.7). Seja $\boldsymbol{\beta}$ o vetor de parâmetros relacionado com a probabilidade condicional $P(Y_i = 1|\mathbf{x}_i) = \pi(\mathbf{x}_i)$ para $i \in \{1, ..., n\}$. Então, para uma amostra de tamanho n, tem-se que

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}[1 - \pi(\mathbf{x}_i)]^{(1-y_i)}, \text{com } y_i \in \{0, 1\}, \tag{3.9}$$

e o logaritmo da função de verossimilhança é

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} [y_i \ln \pi(\mathbf{x}_i) + (1 - y_i) \ln(1 - \pi(\mathbf{x}_i))]. \tag{3.10}$$

Para encontrar o valor de $\boldsymbol{\beta}$ que maximiza $\ell(\boldsymbol{\beta})$, foi utilizado o processo iterativo de Newton-Raphson, e para isso fez-se necessário derivar $\ell(\boldsymbol{\beta})$ em relação a cada parâmetro,

ou seja,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \pi(\mathbf{x}_i) = 0$$
$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} x_{ij} y_i - \sum_{i=1}^{n} x_{ij} \pi(\mathbf{x}_i) = 0, \text{para } j \in \{1, ..., p\}. \tag{3.11}$$

Com base nas propriedades de somatório, a partir de (3.11), segue-se que o estimador pelo método da máxima verossimilhança $\widehat{\boldsymbol{\beta}}$, é a solução das equações de verossimilhança

$$\sum_{i=1}^{n} (y_i - \pi(\mathbf{x}_i)) = 0$$
$$\sum_{i=1}^{n} x_{ij} (y_i - \pi(\mathbf{x}_i)) = 0, \text{para } j \in \{1, ..., p\}.$$

Desta forma, o vetor escore $\mathbf{U}(\boldsymbol{\beta})$ pode ser escrito como

$$\mathbf{U}(\boldsymbol{\beta}) = \mathbf{X}^T \mathbf{y} - \mathbf{X}^T \boldsymbol{\pi} = \mathbf{X}^T (\mathbf{y} - \boldsymbol{\pi}).$$

e a matriz de informação de Fisher é dada por:

$$\mathbf{I}(\boldsymbol{\beta}) = \mathbb{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] = \mathbf{X}^T \mathbf{Q} \mathbf{X}$$

sendo $\mathbf{Q} = diag[\boldsymbol{\pi}(\mathbf{x})(1 - \boldsymbol{\pi}(\mathbf{x}))]$ e $\mathbf{X}$ a matriz de dados, e sua inversa $[\mathbf{I}(\boldsymbol{\beta})]^{-1}$, a matriz de variâncias e covariâncias das estimativas de máxima verossimilhança dos parâmetros (SILVA, 1992).

De acordo com Souza (2006), a solução para as equações de verossimilhança é obtida usando o método iterativo de Newton-Raphson, sendo o conjunto de equações iterativas dado por:

$$\begin{aligned} \boldsymbol{\beta}^{(t+1)} &= \boldsymbol{\beta}^{(t)} + [\mathbf{I}(\boldsymbol{\beta}^{(t)})]^{-1} \mathbf{U}(\boldsymbol{\beta}^{(t)});\ t = 0, 1, 2... \\ &= \boldsymbol{\beta}^{(t)} + [\mathbf{X}^T \mathbf{Q}^{(t)} \mathbf{X}]^{-1} \mathbf{X}^T (\mathbf{y} - \boldsymbol{\pi}^{(t)}). \end{aligned} \tag{3.12}$$

sendo que $\boldsymbol{\beta}^{(t)}$ e $\boldsymbol{\beta}^{(t+1)}$ são vetores dos parâmetros estimados nos passos t e $t+1$, respectivamente.

O chute inicial é dado com todos os coeficientes igualados a zero. Esses valores iniciais são substituídos no lado direito da equação (3.12), que dará o resultado para a primeira iteração $\beta^{(1)}$. Em seguida, os valores são novamente substituídos no lado direito, $\mathbf{U}(\boldsymbol{\beta})$ e $\mathbf{I}(\boldsymbol{\beta})$ são recalculados, encontrando assim $\beta^{(2)}$. Esse processo é repetido, até que a máxima

mudança em cada parâmetro estimado do próximo passo seja menor que um critério, ou seja, o processo é repetido até a convergência, obtendo-se $\boldsymbol{\beta} = \boldsymbol{\beta}^{(t+1)}$. Dentre muitos critérios existentes, um deles para verificar a convergência poderia ser:

$$\sum_{j=1}^{p} \left(\frac{\boldsymbol{\beta}_j^{(t)} - \boldsymbol{\beta}_j^{(t+1)}}{\boldsymbol{\beta}_j^{(t)}} \right)^2 < \xi,$$

tomando-se para ξ um valor suficientemente pequeno. Em geral, esse algoritmo é robusto e converge rapidamente.

Como geralmente não é possível encontrar distribuições exatas para os estimadores dos parâmetros, trabalha-se com resultados assintóticos, em que considera-se que o modelo escolhido irá satisfazer as condições de regularidade. Conforme Souza (2006), em problemas regulares a função *Escore* $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$ tem valor esperado igual a zero e a estrutura de covariância é igual a matriz de informação de Fisher $\mathbf{I}(\boldsymbol{\beta}) = \mathbb{E}\left[-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}^T} \right] = \mathbf{X}^T \mathbf{Q} \mathbf{X}$. Sendo assim, a distribuiçao assintótica dos $\boldsymbol{\beta}$ é dada por:

$$\widehat{\boldsymbol{\beta}} \sim N_p(\boldsymbol{\beta}, I(\boldsymbol{\beta})^{-1}).$$

Considerando o modelo dado em (3.6), pode-se calcular a razão de chances referente a cada nível, o qual corresponderá a razão de chances de cada variável dicotômica x_j ($j = 1, ..., p$). Logo, $\psi(x = j) = e^{\widehat{\beta}_j}$. Essa estimativa intervalar depende dos valores estimados dos β's e seus respectivos erros padrões, sendo que estes valores estão relacionados somente com a categoria j de interesse. O intervalo de confiança é dado por

$$IC_{1-\alpha}(\psi_j) = \exp[\widehat{\beta}_j \pm z_{\alpha/2} EP(\widehat{\beta}_j)],$$

em que $z_{\alpha/2}$ é o quantil da distribuição normal e $EP(\widehat{\beta}_j)$ o erro padrão do coeficiente.

3.3 Medidas de avaliação do modelo

3.3.1 *O likelihood value*

É uma das principais medidas de avaliação geral do modelo de regressão logística, sendo que busca aferir a capacidade do modelo estimar a probabilidade associada à ocorrência de determinado evento. Ele tem sido representado pela expressão $-2LL$, que é o logaritmo natural do *likelihood value* multiplicado por -2, seguindo uma distribuição Qui-quadrado. Quanto mais próximo de zero, maior o poder preditivo do modelo como

um todo. Seu principal objetivo é verificar se a regressão como um todo é estatisticamente significante e facilitar comparações entre modelos alternativos, servindo para verificar se o modelo melhora com a inclusão ou exclusão de uma ou mais variáveis independentes (CORRAR *et al*, 2009).

3.3.2 *O teste de Hosmer e Lemeshow*

Segundo, por exemplo, Corrar *et al* (2009), a finalidade desse teste é verificar se existem diferenças significativas entre as classificações realizadas pelo modelo e a realidade observada. A certo nível de significância busca-se aceitar a hipótese de que não existem diferenças entre os valores preditos e observados. Caso existam diferenças significativas entre essas classificações, o modelo não representa a realidade de forma satisfatória.

A sua estatística é definida comparando-se o número observado com o número esperado de sucessos de g grupos formados. Conforme Paula (2010), o primeiro grupo deverá conter $n_1^{'}$ elementos correspondentes às $n_1^{'}$ menores probabilidades ajustadas, as quais serão denotadas por

$$\widehat{\pi}_{(1)} \leq \widehat{\pi}_{(2)} \leq \ldots \leq \widehat{\pi}_{(n_1^{'})}.$$

O segundo grupo deverá conter os $n_2^{'}$ elementos correspondentes às seguintes probabilidades ajustadas

$$\widehat{\pi}_{(n_1^{'}+1)} \leq \widehat{\pi}_{(n_1^{'}+2)} \leq \ldots \leq \widehat{\pi}_{(n_1^{'}+n_2^{'})}.$$

E assim, sucessivamente, até o último grupo que deverá conter as $n_g^{'}$ maiores probabilidades ajustadas

$$\widehat{\pi}_{(n_1^{'}+\ldots+n_{g-1}^{'}+1)} \leq \widehat{\pi}_{(n_1^{'}+\ldots+n_{g-1}^{'}+2)} \leq \ldots \leq \widehat{\pi}_{(n)}.$$

O número observado de sucessos no primeiro grupo formado será dado por

$$O_1 = \sum_{j=1}^{n_1^{'}} y_{(j)},$$

em que $y_{(j)} = 0$ se o elemento correspondente é fracasso e $y_{(j)} = 0$ se é sucesso. Generalizando, tem-se

$$O_i = \sum_{j=n_1^{'}+\ldots+n_{i-1}^{'}+1}^{n_1^{'}+\ldots+n_i^{'}} y_{(j)},\ 2 \leq i \leq g.$$

A estatística é definida por

$$\widehat{C} = \sum_{i=1}^{g} \frac{(O_i - n_i^{'}\overline{\pi}_i)^2}{n_i^{'}\overline{\pi}_i(1-\overline{\pi}_i)},$$

em que

$$\overline{\pi}_1 = \frac{1}{n_1^{'}} \sum_{j=1}^{n_1^{'}} \widehat{\pi}_{(j)} \; e \; \overline{\pi}_i = \frac{1}{n_i^{'}} \sum_{j=n_1^{'}+\ldots+n_{i-1}^{'}+1}^{n_1^{'}+\ldots+n_i^{'}} \widehat{\pi}_{(j)},$$

para $2 \leq i \leq g$. Hosmer e Lemeshow sugerem a formação de $g = 10$ grupos de mesmo tamanho (aproximadamente), de modo que o primeiro grupo contenha $n_i^{'}$ elementos correspondentes às $[n/10]$ menores probabilidades ajustadas e assim por diante até o último grupo com $n_{10}^{'}$ elementos correspondentes às $[n/10]$ maiores probabilidades ajustados. Verificaram também através de simulações que a distribuição nula assintótica de $\widehat{C}$ pode ser bem aproximada por uma distribuição $\chi^2_{(g-2)}$.

Desta forma, se o nível de significância adotado α for menor que o nível descritivo do teste (valor-p), aceita-se a hipótese nula de que não existem diferenças significativas entre os valores preditos e observados, concluíndo assim que o modelo representa a realidade de forma satisfatória.

3.4 Análise de Resíduos e Diagnósticos

A análise de diagnósticos é uma etapa importante na análise de um ajuste de regressão, pois ela auxilia na verificação de possíveis afastamentos das suposições feitas para o modelo, bem como a existência de observações discrepantes com alguma interferência desproporcional ou inferencial nos resultados do ajuste. Um outro tópico importante na análise de diagnóstico é a detecção de observações influentes, isto é, pontos que exercem um peso desproporcional nas estimativas dos parâmetros do modelo.

Segundo Souza (2006), a análise de resíduos e diagnóstico é utilizada para detectar problemas como:

- Presença de observações discrepantes (pontos aberrantes);
- Inadequação das pressuposições para os erros aleatórios ou a médias;
- Colinearidade entre as colunas da matriz do modelo;
- Forma funcional do modelo inadequada;
- Presença de observações influentes.

Basicamente as estatísticas de influência definem quanto a eliminação de uma observação em particular pode influenciar no ajuste do modelo. A seguir, as medidas geral-

mente utilizadas para os resíduos e diagnósticos serão abordadas.

3.4.1 *Diagonal da matriz H*

Os elementos da matriz **H** são utilizados para detectar pontos extremos no espaço designado. Tais pontos exercem um papel importante no ajuste final dos parâmetros de um modelo estatístico, ou seja, sua eliminação pode implicar mudanças substanciais dentro de uma análise estatística.

No modelo de regressão linear, a matriz **H** é definida por:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T,$$

conhecida como matriz de projeção da solução de mínimos quadrados ou matriz *hat*.

Como nos modelos de regressão logística, $\text{Var}(\epsilon_i) = \pi(\mathbf{x}_i)(1 - \pi(\mathbf{x}_i))$ não é constante, a matriz de projeção para o modelo logístico fica definida como:

$$\mathbf{H} = \mathbf{Q}^{1/2}\mathbf{X}(\mathbf{X}^T\mathbf{Q}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Q}^{1/2},$$

em que a utilização dos elementos da diagonal principal de **H** é utilizada para detectar a presença de pontos de alavanca. Hosmer e Lemeshow (1989) mostram, contudo, que o uso da diagnonal principal da matriz de projeção **H** deve ser feito com algum cuidado em regressão logística e que as interpretações são diferentes daquelas no caso do modelo linear. Portanto, a diagonal da matriz $\widehat{\mathbf{H}}$ é dada por:

$$\widehat{h}_{ii} = \widehat{\pi}(\mathbf{x}_i)(1 - \widehat{\pi}(\mathbf{x}_i))\mathbf{x}_i^T[\mathbf{I}(\widehat{\boldsymbol{\beta}})]^{-1}\mathbf{x}_i;\ i = 1, 2, ..., n.$$

3.4.2 *Resíduo de Pearson*

O resíduo de Pearson auxilia na classificação de uma observação que pode ser considerada como outlier. O resíduo para cada elemento amostral é definido como a diferença entre os valores observados e os valores preditos e é definido por

$$r_i = y_i - \widehat{\pi}(\mathbf{x}_i).$$

Devido ao efeito da escala de medição, esse tipo de resíduo não é útil para detectar outliers. Desta forma, Souza (2006) sugere que é necessário transformá-lo para eliminar o efeito das variáveis resposta e preditora. No modelo logístico, o resíduo de Pearson

transformado é definido por:

$$(rp)_i = \frac{y_i - \widehat{\pi}(\mathbf{x}_i)}{\sqrt{\widehat{\pi}(\mathbf{x}_i)(1-\widehat{\pi}(\mathbf{x}_i))}},\ i = 1, 2, ..., n,$$

sendo que, no caso desses valores serem pequenos, há indicação de que o modelo está bem ajustado. Os resíduos de Pearson são componentes da estatística χ^2 de Pearson.

3.4.3 *Resíduo de Deviance*

Segundo Souza (2006), os resíduos de deviance são componentes da deviance, sendo utilizados para detectar os erros no ajuste do modelo e medem a discrepância entre o modelo saturado e o modelo restrito em relação as observações y_i. O resultado da deviance é uma estatística de bondade de ajuste, para cada indivíduo ($i = 1, 2, ..., n$) baseada no logaritmo da função de verossimilhança, definida por:

$$d_i = \begin{cases} -\sqrt{-2\ln(1-\widehat{\pi}(\mathbf{x}_i))} & \text{, se } y_i = 0 \\ \pm\sqrt{2\left[y_i \ln\left(\frac{y_i}{\widehat{\pi}(\mathbf{x}_i)}\right) + (-y_i)\ln\left(\frac{1-y_i}{1-\widehat{\pi}(\mathbf{x}_i)}\right)\right]} & \text{, se } 0 < y_i < 1 \\ \sqrt{-2\ln(\widehat{\pi}(\mathbf{x}_i))} & \text{, se } y_i = 1 \end{cases}$$

3.4.4 *C e CBar*

São medidas de diagnóstico baseadas no intervalo de confiança, que fornecem medidas de influência das observações individuais sob $\boldsymbol{\beta}$, e têm a mesma ideia da Distância de Cook na regressão linear. Essa medida por ser escrita como:

$$C_i = \frac{(rp)_i^2 h_{ii}}{(1-h_{ii})^2},\ i = 1, 2, ..., n.$$

Christensen (1997) define uma nova medida $\overline{C}_i$, também conhecida por *CBar*, em termos da medida C_i, que é definida como:

$$\overline{C}_i = \frac{(rp)_i^2 h_{ii}}{(1-h_{ii})},\ i = 1, 2, ..., n.$$

3.4.5 *DIFCHISQ e DIFDEV*

A DIFCHISQ é útil para detectar as observações mal ajustadas, isto é, as observações que contribuam pesadamente na diferença entre os dados e os valores preditos. Fazendo

uso das aproximações lineares e a estatística χ^2 de Pearson, Souza (2006) a define como:

$$DIFCHISQ_i = \frac{\overline{C}_i}{h_{ii}} = \frac{(rp)_i^2}{1 - h_{ii}},\ i = 1, 2, ..., n.$$

De forma similar, a DIFDEV é utilizada para detectar observações que são influentes na estimação do ajuste do modelo logístico (SOUZA, 2006). Baseada no resíduo da deviance, ela é definida por:

$$DIFDEV_i = d_i^2 + \overline{C}_i = d_i^2 + \frac{(rp)_i^2}{h_{ii}(1 - h_{ii})},\ i = 1, 2, ..., n.$$

As estatísticas de diagnóstico apresentadas em toda esta seção são conceitualmente interessantes, pois permitem identificar as observações que contribuem para um mal ajuste do modelo e que também tenham grande influência nas estimativas dos parâmetros. Depois de identificadas, pode-se decidir sobre a sua permanência ou não na análise. Desta forma, os gráficos de diagnóstico são de grande utilidade para detectar pontos influentes no modelo logístico.

4 TESTES DE DIAGNÓSTICO

O testes de diagnósticos servem para rastrear um fator de risco de determinado experimento. Esses testes, que serão descritos a seguir, auxiliam no ajuste do modelo de regressão logística.

4.1 Resultados falso-positivos e falso-negativos

Tomando como exemplo um estudo de avaliação de determinada doença, há o interesse de saber se as pessoas testadas têm ou não a doença. O resultado do teste pode ser positivo (prevendo que a pessoa está doente) ou negativo (prevendo que a pessoa não está doente), podendo ou não coincidir com a verdadeira situação da pessoa. Desta forma, têm-se as seguintes definições:

- **Falso-positivo**: ocorre quando o teste dá positivo, mas o indivíduo não está doente. É também conhecido como **erro tipo I**.
- **Falso-negativo**: ocorre quando o teste dá negativo, mas o indivíduo está doente. É também conhecido como **erro tipo II**.

4.2 Sensibilidade e especificidade

Sensibilidade e especificidade são duas medidas importantes do funcionamento de um teste. A **sensibilidade** (pode também ser chamada de taxa de verdadeiro positivo) refere-se à capacidade de um teste para detectar uma doença quando ela está presente, ou seja, é a probabilidade de um resultado positivo de um teste, dado que o indivíduo realmente esteja doente. O teste será altamente sensível se a probabilidade for alta e o teste não será sensível se ele falhar na detecção da doença em indivíduos doentes. A taxa com que isso ocorre é chamada de **taxa de erro falso-negativo** (TFN).

A **especificidade** (também chamada de taxa de verdadeiro negativo) é capacidade de um teste indicar ausência de doença quando ela não está presente, ou seja, é a probabilidade de um resultado negativo de um teste, dado que o indivíduo não está doente. Quando essa probabilidade é alta, o teste é altamente específico e, se o teste não é específico, indicará falsamente a presença de doença em indivíduos não-doentes. A taxa com que isso ocorre é chamada de **taxa de erro falso-positivo** (TFP). A especificidade e a taxa de erro falso-positivo somam 1(100%).

O teste ideal, com 100% de sensibilidade e especificidade, raramente existe na prática, pois a tentativa de melhorar a sensibilidade frequentemente tem o efeito de diminuir a especificidade.

Para calcular essas medidas, os dados referentes aos indivíduos estudados e os resultados dos testes podem ser colocados em uma tabela 2x2, sugerida por Jekel et al (2006), e estes são mostrados na Tabela 3.

Tabela 3 - Tabela padrão 2x2 comparando os resultados do teste e a verdadeira condição da doença nos indivíduos testados.

Verdadeira condição do teste	Resultado		Total
	Doente	Não-Doente	
Positivo	a	b	$a+b$
Negativo	c	d	$c+d$
Total	$a+c$	$b+d$	$a+b+c+d$

Fonte: Jekel et al (2006).

A interpretação das células é a seguinte:

a = indivíduos com um resultado de teste verdadeiro-positivo.

b = indivíduos com um resultado de teste falso-positivo.

c = indivíduos com um resultado de teste falso-negativo.

d = indivíduos com um resultado de teste verdadeiro-negativo.

$a+b$ = todos os indivíduos com um resultado de teste positivo.

$c+d$ = todos os indivíduos com um resultado de teste negativo.

$a+c$ = todos os indivíduos com a doença.

$b+d$ = todos os indivíduos sem a doença.

$a+b+c+d$ = todos os indivíduos estudados.

As fórmulas são as seguintes:

$a/(a+c)$ = sensibilidade.

$d/(b+d)$ = especificidade.

$b/(b+d)$ = taxa de erro falso-positivo.

$c/(a+c)$ = taxa de erro falso-negativos.
$a/(a+b)$ = valor preditivo positivo.
$d/(c+d)$ = valor preditivo negativo.
$[a/(a+c)]/[b/(b+d)] = (a/b)/[(a+c)/(b+d)]$ = razão de probabilidades positiva (RPb+).
$[c/(a+c)]/[d/(b+d)] = (c/d)/[(a+c)/(b+d)]$ = razão de probabilidades negativa (RPb−).
$(a+c)/(a+b+c+d)$ = prevalência.

4.3 Valor preditivo

Apesar da grande utilidade, a sensibilidade e especificidade não respondem duas questões questões clínicas importantes:

- Se o resultado do teste de um paciente é positivo, qual é a probabilidade dele ter a doença que está sob investigação?
- Se o resultado é negativo, qual a probabilidade do paciente não ter a doença?

Para tais questões, que são influenciadas pela sensibilidade, pela especificidade e pela prevalência, Jekel et al (2006) diz que elas podem ser respondidas fazendo-se uma análise horizontal na Tabela 3, na qual a fórmula $a/(a+b)$ é usada para calcular o **valor preditivo positivo**. No estudo de uma determinada população, essa medida indica a probabilidade do indivíduo ter a doença, dado que o resultado do teste é positivo. De maneira análoga, a fórmula $d/(c+d)$ é utilizada para calcular o **valor preditivo negativo**, que indica a probabilidade do indivíduo não ter a doença, dado que o resultado do teste é negativo.

A **prevalência** é o número total de pessoas doentes dividido pelo número de pessoas estudadas, então, de acordo com a Tabela 3, é $(a+c)/(a+b+c+d)$.

4.4 Razão de probabilidades

Conforme visto em Jekel et al (2006), a **razão de probabilidade positiva** (RPb+) é a razão entre a sensibilidade de um teste e a taxa de erro falso-positivo de um teste, ou seja

$$\text{RPb+} = \frac{\text{sensibilidade}}{\text{TFP}} = \frac{a/a+c}{b/b+d}.$$

A RPb+ é a razão entre aquilo que se deseja (teste sensível) e algo não desejado (TFP), então quanto maior for o valor de RPb+, melhor será o teste. Para ser um bom teste, a

razão deve ser muito maior do que 1.

De forma semelhante, a **razão de probabilidade negativa** (RPb−) é a razão entre a taxa de erro falso-negativo e a especificidade, ou seja

$$\text{RPb}- = \frac{\text{TFN}}{\text{especificidade}} = \frac{c/(a+c)}{d/(b+d)}.$$

Nesse caso, RPb− é a razão entre algo não desejável (TFN) e algo desejável (especificidade), portanto quanto menor (mais próximo de 0), melhor será o teste.

Em resumo, deseja-se RPb+ grande e RPb− pequeno.

4.5 Curvas de característica operatória do receptor

Sabe-se que qualquer forma de diagnóstico tem uma certa imprecisão. O teste ideal seria aquele que fosse altamente sensível e específico ao mesmo tempo. Porém, na realidade, isso não é possível. Geralmente são encontrados testes com alta sensibilidade e pouca especificidade ou testes com baixa sensibilidade e alta especificidade. Muitos testes estão baseados em uma medida clínica que pode assumir uma série de valores; nesse caso, há um compromisso inerente entre a sensibilidade e especificidade.

A curva de Característica Operatória do Receptor (ou curva ROC - *Receiver Operating Characteristic*) ilustra a relação entre sensibilidade e especificidade, e pode ser utilizada para decidir um bom ponto de corte. Uma curva ROC é um gráfico de linha que plota a sensibilidade do teste em função da probabilidade de um resultado falso-positivo (1-especificidade) para uma série de diferentes pontos de corte.

Segundo Jekel et al (2006), as curvas ROC estão sendo cada vez mais vistas na literatura médica. Diz-se que o nome surgiu na Inglaterra, durante a segunda batalha da Grã-Bretanha, quando o desempenho dos operadores receptores de radar era avaliado por meio do seguinte fundamento: um verdadeiro-positivo consistia em um aviso prévio correto de que os aviões alemães estavam vindo sobre o canal Inglês; um falso-positivo quando um operador receptor mandava um alarme, mas nenhum avião inimigo aparecia e um falso-negativo ocorria quando os aviões alemães apareciam sem aviso prévio dos operadores de radar.

Quando um teste de diagnóstico existente é avaliado, esse tipo de gráfico pode ser usado como auxílio da avaliação da utilidade do teste e para determinar o ponto de corte mais apropriado. Quanto mais perto a linha está no canto superior esquerdo do gráfico,

mais preciso é o teste. Além disso, o ponto que se encontra mais próximo desse canto é normalmente escolhido como o corte que maximiza simultaneamente tanto a sensibilidade como a especificidade (PAGANO & GAUVREAU, 2004).

5 APLICAÇÃO

5.1 Descrição do problema

Como aplicação para modelos de regressão logística, será ajustado um modelo para predizer a ocorrência de pré-eclampsia (PE), na qual é uma doença que ocorre apenas durante a gravidez, principalmente no últimos três meses de gestação, sendo considerada a maior causa de morbimortalidade materno-fetal. O mais comum é que apareça depois da 37ª semana, mas, na realidade, pode acontecer em qualquer época da segunda metade da gravidez, incluindo durante o parto ou depois (geralmente nas primeiras 48 horas). Também é possível ter sintomas de pré-eclâmpsia antes de 20 semanas, mas somente em casos mais raros. Apesar de inúmeros estudos, a etiologia da PE permanece desconhecida.

Paula, L. G. (2010) diz que diversos fatores foram identificados como capazes de aumentar o risco de pré-eclâmpsia, sendo esta primariamente uma doença da primeira gestação, ocorrendo em 2 a 7% de nulíparas saudáveis. Outros fatores de risco incluem gestações múltiplas, pré-eclâmpsia em gestação prévia, obesidade, história familiar de pré-eclâmpsia e eclâmpsia, hipertensão crônica, diabetes, idade materna superior a 35 anos, dentre outros fatores. É possível que também exista predisposição genética.

Basicamente, essa doença seria causada pela presença da placenta e pela reação materna à placentação. Acredita-se que a placentação inadequada não seria a causa, mas forte fator predispondente (PAULA, L. G. 2010).

Classificação da pré-eclampsia: precoce e tardia.

Segundo Brandão et al (2010), a PE é classificada de acordo com a gravidade das manifestações da doença em leve ou grave. No entanto, estudos mais recentes têm sugerido uma nova classificação baseada na época de início das manifestações clínicas. Esses estudos propõem essa classificação em:

1. Precoce: para pacientes que apresentam o início da sintomatologia antes das 34 sem-

anas de gravidez associando-se principalmente à remodelação placentária incorreta, com evidências de lesões isquêmicas ao exame da placenta.

2. Tardia: para pacientes nas quais os sintomas iniciam-se após as 34 semanas. Por sua vez, está mais associada a fatores constitucionais maternos, com índice de massa corporal (IMC) aumentado.

5.2 Metodologia utilizada

Os dados desta aplicação foram levantados por Alves (2012), onde o mesmo realizou um estudo em pacientes consecutivas atendidas no Laboratório de Medicina Materno-Fetal UECE/HGF, que se encontra no Hospital Geral de Fortaleza (HGF), em Fortaleza/Ceará. A coleta dos dados foi no período de agosto de 2009 até fevereiro de 2011, tendo uma duração de 19 meses. Houve um recrutamento inicial de 550 pacientes, mas ao utilizar alguns critérios de exclusão, a amostra final foi de 487 pacientes.

A base de dados envolve variáveis dicotômicas como presença de diabetes, ocorrência anterior de PE, incisura na artérias uterinas, dentre outras; bem como variáveis contínuas sendo representadas pelas medidas de pressão em artérias uterinas e oftálmicas, índice de massa corporal, etc. Desta forma, com base na metodologia apresentada no Capítulo 3, o interesse é predizer $P(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$, em que Y é a variável de interesse (ou resposta), na qual representa a ocorrência (sucesso) ou não de PE, e $\mathbf{X} = (X_1, ..., X_p)$ é o conjunto de covariáveis preditoras para sua ocorrência, sendo $\mathbf{x} = (x_1, ..., x_p)$ um valor particular.

Na análise descritiva dos dados, foi construída uma tabela resumo (Tabela 6) contendo o valor mínimo, o 1º quartil, a mediana, a média, o 3º quartil e o valor máximo de cada variável contínua (ou numérica) utilizada na aplicação. Para as variáveis categóricas foram construídas tabelas de contigênca 2x2 (Apêndice). Os resultados dos devidos testes de associação Qui-Quadrado e Exato de Fisher destas variáveis com a ocorrência de PE encontram-se na Tabela 7.

A análise gráfica de todas as variáveis do estudo está no Apêndice, onde cada variável contínua foi representada por um boxplot de acordo com os níveis da ocorrência de pré-eclampsia. Para cada variável categórica foi construído o gráfico de barras emparelhado de acordo com os níveis da ocorrência de PE e os níveis de fatores da respectiva variável.

Em seguida foi ajustado o modelo de regressão logística e, para a escolha do modelo final, fez-se o uso de técnicas de seleção de variáveis utilizando os métodos *backward*,

forward e *stepwise*. Com base em critérios de seleção de modelos, tais como o AIC, o BIC, o $-2LL$ e a curva ROC, os modelos dados nas três técnicas de seleção de variáveis são comparados para que, no fim, seja escolhido dentre eles um modelo que possa predizer se determinada paciente venha a sofrer ou não de pré-eclampsia. Depois de selecionado o modelo, prossegue-se com as estimativas dos parâmetros, os testes de hipóteses e a devida análise de diagnóstico para a validação do mesmo.

5.3 Análise descritiva

5.3.1 Apresentação dos dados

O banco de dados utilizado para a análise possui 23 variáveis explicativas, nas quais 8 são categóricas e 15 são numéricas. Ele é composto de 487 observações, dividida em 31 pacientes com pré-eclampsia e 456 pacientes sem pré-eclampsia. A descrição de cada variável categórica está na Tabela 4, enquanto que a descrição das variáveis numéricas está na Tabela 5.

Tabela 4 - Descrição das variáveis categóricas.

Variável	**Rótulo**	**Classes**	**Código**
paridade	Paridade	0	Primeiro filho
		1	Tem pelo menos um filho
pe_ant	Ocorrência anterior de PE	0	Não teve
		1	Teve
familar_pe	Caso de PE na famíla	0	Não tem caso
		1	Tem caso
prematuro_ant	Parto prematuro anteriormente	0	Não teve
		1	Teve
hiperten_cronica	Hipertensão crônica	0	Não tem
		1	Tem
diabetes	Diabetes	0	Não tem
		1	Tem
uterD_Incis	Artéria uterina direita com incisura	0	Sem incisura
		1	Com incisura
uterE_incis	Artéria uterina esquerda com incisura	0	Sem incisura
		1	Com incisura
PE	Pré-eclampsia	0	Não tem pré-eclampsia
		1	Tem pré-eclampsia

Tabela 5 - Descrição das variáveis numéricas.

Variável	**Rótulo**
idade	Idade
imc	Índide de massa corporal
uterD_IR	Artéria uterina direita - Índice de resistência
uterD_AB	Artéria uterina direita - Relação AB
uterE_IR	Artéria uterina esquerda - Índice de resistência
uterE_AB	Artéria uterina esquerda - Relação AB
Media_IP	Média do indíce pulsatilidade das artérias uterinas
oftIR	Pressão arterial oftálmica - Índice de resistência
oftIP	Pressão arterial oftálmica - Índice de pulsatilidade
oftAB	Pressão arterial oftálmica - Relação AB
oftPS	Pressão arterial oftálmica - Pico sistólico
oftPD1	Pressão arterial oftálmica - Pico diastólico
Razao_pico	Razão de pico
menor_IP	Menor índice de pulsatilidade
PAM	Pressão arterial média

A Tabela 6 mostra algumas estatísticas descritivas das variáveis numéricas do banco de dados. Enquanto que na Tabela 7 encontram-se os resultados dos níveis descritivos dos testes de associação de cada variável categórica com a ocorrência de PE.

Tabela 6 - Resumo descritivo das variáveis numéricas.

Variável	**Mínimo**	**1º Quartil**	**Mediana**	**Média**	**3º Quartil**	**Máximo**
idade	13,00	21,00	26,00	26,27	31,00	45,00
imc	0,19	22,02	24,64	25,28	27,66	51,44
uterD_IR	0,37	0,60	0,71	0,70	0,80	1,25
uterD_AB	0,78	2,48	3,38	3,97	4,96	13,00
uterE_IR	0,38	0,63	0,72	0,73	0,80	9,59
uterE_AB	0,64	2,63	3,53	4,16	4,75	23,00
Media_IP	0,58	1,17	1,51	1,55	1,83	6,27
oftIR	0,39	0,76	0,80	0,81	0,85	2,60
oftIP	0,31	1,69	1,99	2,06	2,41	5,93
oftAB	0,98	4,07	5,00	5,63	6,32	22,00
oftPS	14,37	28,63	35,07	36,56	42,22	91,43
oftPD1	3,86	15,74	20,04	21,35	25,03	65,13
PAM	53,33	73,33	76,67	80,19	86,67	126,67
Razao_pico	0,21	0,51	0,58	0,59	0,66	1,12
menor_IP	0,36	0,92	1,17	3,55	1,58	1113,00

De acordo com os resultados da Tabela 7, somente as variáveis "pe_ant"(ocorrência anterior de PE), "familar_pe"(histórico de PE na família) e "prematuro_ant"(parto prematuro anteriormente) têm indícios de associação estatística com a ocorrência de PE. Tanto no teste de Qui-Quadrado como no Teste Exato de Fisher, testou-se a hipótese

nula (H_0) de que não há relação entre a variável em questão e a ocorrência de PE e, considerando um nível de 5%, rejeitou-se H_0, mostrando assim indícios de associação entre a covariável e a variável resposta.

Tabela 7 - Níveis descritivos dos testes de associação entre as variáveis categóricas e PE.

Variável	**Teste**	
	Qui-Quadrado	**Exato de Fisher**
pe_ant	< 0,001*	-
familar_pe	0,0026*	-
prematuro_ant	0,0078*	-
hiperten_cronica	-	0,1423
diabetes	-	0,6136
uterD_Incis	0,6354	-
uterE_incis	0,1414	-
paridade	0,3236	-

*Teste significativo a 5%.

5.4 Ajuste do modelo

5.4.1 *Seleção do modelo*

Em posse das covariáveis e da variável resposta (PE), e com o uso de técnicas de seleção de variáveis descritas na Seção 2.6, foram selecionados modelos sem interação e com iteração de ordem 2. Os modelos sem interação tiveram resultados iguais nas três técnicas de seleção de variáveis e, ao fazer a seleção de modelos com interação, não houve subconjunto de variáveis com iterações significativas.

A Tabela 8 mostra os resultados dos testes de nulidade global da Razão de Verossimilhança, Score e Wald para os coeficientes do modelo sem interação. Cada teste têm como hipótese nula $H_0 : \boldsymbol{\beta} = \mathbf{0}$ e é considerado significativo se o seu valor-p for menor que um nível de significância α especificado (nesta aplicação será considerado $\alpha = 0,05$). Para o modelo sem interação, que é o mesmo nos três métodos de seleção, os testes de nulidade global foram significativos, bem como as variáveis selecionadas também foram significativas. Então, este modelo será candidato a um modelo final.

Tabela 8 - Testes de nulidade global dos coeficientes para o modelo sem interação.

Teste	$\boldsymbol{\chi^2}$	**Valor-p**
Razão de verossimilhança	38,1964	<0,001
Score	47,5955	<0,001
Wald	32,5982	<0,001

Na Tabela 9 encontram-se os resultados dos testes de nulidade global para o modelo com interação de ordem 2 para os três métodos de seleção de variáveis.

Tabela 9 - Testes de nulidade global dos coeficientes para o modelo com interação.

Teste	**Stepwise**		**Backward**		**Forward**	
	$\boldsymbol{\chi^2}$	**Valor-p**	$\boldsymbol{\chi^2}$	**Valor-p**	$\boldsymbol{\chi^2}$	**Valor-p**
Razão de verossimilhança	34,4801	<0,001	155,7953	<0,001	45,202	<0,001
Score	43,1953	<0,001	123,2021	<0,001	52,6799	<0,001
Wald	29,8099	<0,001	29,5081	0,9836	35,8776	<0,001

Com os resultados apresentados na Tabela 9, têm-se os seguintes comentários:

- No método *stepwise*, os testes de nulidade global da razão de verossimilhança, Score e Wald para testar a hipótese nula $\boldsymbol{\beta} = \mathbf{0}$ foram significativos, além de que para cada coeficiente o teste Wald também foi significativo. Desta forma, o modelo gerado neste método é um candidato para ser o modelo final.

- No método *backward*, somente o teste de nulidade global Wald não foi significativo, além de que alguns coeficientes do modelo gerado também não foram. Com base nisso, foi reajustado um novo modelo somente com as variáveis que foram significativas, mas o teste de nulidade global Wald continuou não sendo significativo e alguns coeficientes também não. A partir disso, mais dois modelos foram reajustados, porém o teste de nulidade global Wald não foi significativo, apesar de todas as variáveis reajustadas serem significativas. Então, optou-se que este modelo seria descartado para ser um candidato a modelo final.

- No método *forward*, todos os testes de nulidade global foram significativos, porém algumas variáveis não foram significativas. A partir disso, foi reajustado um novo modelo somente com as variáveis significativas, onde todos os testes de nulidade global foram significativos e todas as variáveis selecionadas também foram consideradas significativas para o modelo reajustado. Sendo assim, este último modelo gerado também é um candidato a modelo final.

A Tabela 10 mostra os resultados dos critérios de seleção dos três modelos candidatos a modelo final. Os modelos candidatos são:

Modelo 1 : $\text{logit}[\pi(\mathbf{x})] = \beta_0 + \beta_1 \text{familar\_pe} + \beta_2 \text{imc} + \beta_3 \text{oftPD1} + \beta_4 \text{pe\_ant} + \beta_5 \text{paridade}$

Modelo 2 : $\text{logit}[\pi(\mathbf{x})] = \beta_0 + \beta_1 \text{familar\_pe} + \beta_2 \text{imc} + \beta_3 \text{pe\_ant} + \beta_4 \text{paridade}$

Modelo 3 : $\text{logit}[\pi(\mathbf{x})] = \beta_0 + \beta_1 \text{familar\_pe} + \beta_2 \text{imc} + \beta_3 \text{oftPD1}$

em que o **modelo 1** vem dos três métodos de seleção de variáveis sem interação; o **modelo 2** vem do método *stepwise* com interação; e o **modelo 3** vem do método *forward* com iteração após um reajuste.

Tabela 10 - Critérios de seleção de modelos.

Critério	**Modelo**		
	Modelo 1	**Modelo 2**	**Modelo 3**
AIC	204,552	206,269	217,896
BIC	229,682	227,21	234,649
-2LL	192,552	196,269	209,896

Dos critérios apresentados na Tabela 10, tem-se que o **modelo 1** apresenta o menor AIC e o menor -2LL. Para facilitar ainda mais a comparação desses modelos, as Figuras 2, 3 e 4 apresentam as curvas ROC dos modelos 1, 2 e 3, respectivamente. Dentre elas, o **modelo 1** apresenta uma maior área sob a curva (0,8277), sendo assim considerado o melhor modelo para predizer a ocorrência de PE.

Figura 2 - Curva ROC do modelo 1.

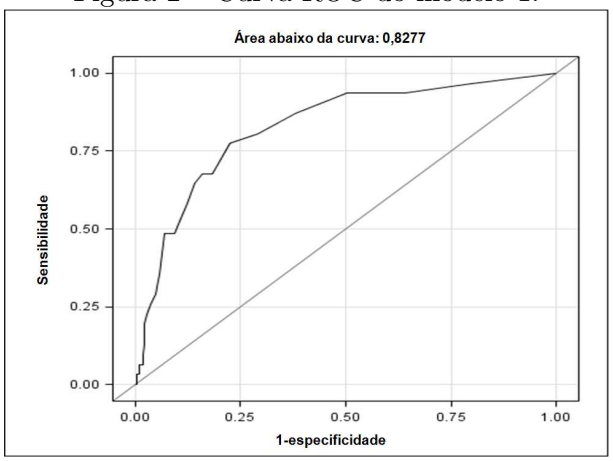


Figura 3 - Curva ROC do modelo 2.

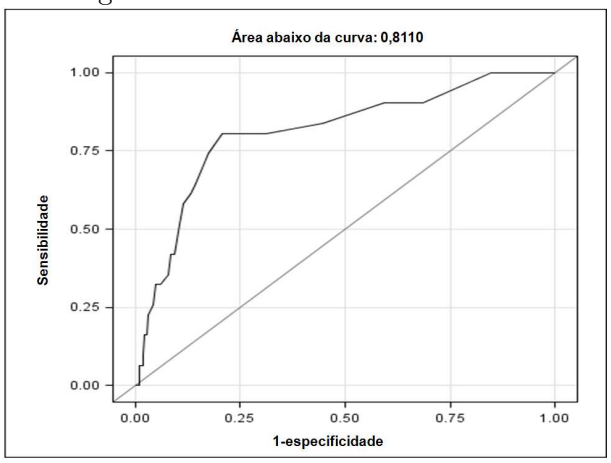
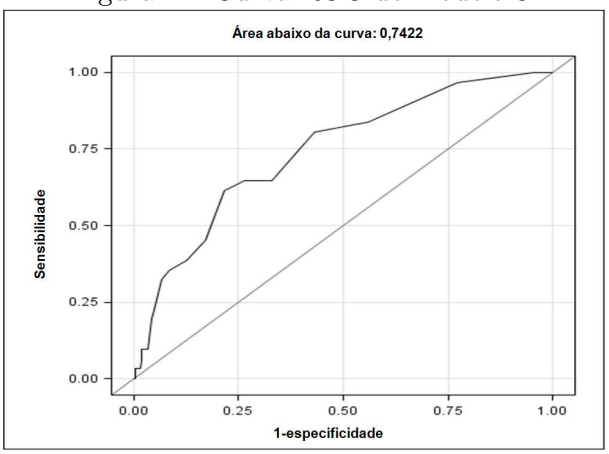


Figura 4 - Curva ROC do modelo 3.



5.4.2 *Interpretação dos resultados*

O modelo ajustado é constituído de três variáveis categóricas (familar_pe, pe_ant e paridade) e duas variáveis contínuas (imc e oftPD1). A Tabela 11 mostra as estatísticas de qualidade do ajuste e, de acordo com os seus resultados, o modelo é considerado bom, além de que os testes de nulidade global dos coeficientes foi significativo, conforme foram mostrados na Tabela 8.

Tabela 11 - Estatísticas de qualidade do ajuste.

Critério	**Valor**	**G.l.**	**Valor/G.l.**	Pr > χ^2
Deviance	192,5523	481	0,4003	1
Pearson	427,8899	481	0,8896	0,9606

Na Tabela 12 encontra-se a partição do teste de Hosmer e Lemeshow e os resultados dos seus testes de bondade do ajuste encontram-se na Tabela 13, na qual mostra que a hipótese nula de que não existem diferenças significativas entre os valores preditos

e observados é aceita, concluindo assim que o modelo representa a realidade de forma satisfatória.

Tabela 12 - Partição do teste de Hosmer e Lemeshow.

Grupo	**Total**	**PE=1**		**PE=0**	
		Observado	**Esperado**	**Observado**	**Esperado**
1	49	0	0,42	49	48,58
2	49	1	0,63	48	48,37
3	49	1	0,91	48	48,09
4	49	0	1,27	49	47,73
5	49	1	1,64	48	47,36
6	49	1	2,03	48	46,97
7	49	2	2,56	47	46,44
8	49	4	3,44	45	45,56
9	49	6	5,69	43	43,31
10	46	15	12,42	31	33,58

Tabela 13 - Resultados do teste de Hosmer e Lemeshow.

$\boldsymbol{\chi^2}$	**G.l.**	Pr > $\boldsymbol{\chi^2}$
3,7452	8	0,8793

Os testes globais também foram significativos, como já foi mostrado anteriormente na Tabela 8. Na Tabela 14 está a análise das estimativas de máxima verossimilhança.

Tabela 14 - Estimativas de máxima verossimilhança.

Parâmetro	**G.l.**	**Estimativa**	**Erro padrão**	$\boldsymbol{\chi^2}$ **Wald**	**Pr** > $\boldsymbol{\chi^2}$
Intercepto	1	-5,1874	1,1107	21,8107	<0,0001
familar_pe 1	1	0,5238	0,2234	5,4970	0,0190
imc	1	0,0930	0,0356	6,8157	0,0090
oftPD1	1	0,0406	0,0203	4,0167	0,0451
pe_ant 1	1	1,0486	0,3032	11,9609	0,0005
paridade 0	1	0,7921	0,2669	8,8071	0,0030

Com base na Tabela 14, o modelo ajustado é

$$\begin{aligned}\text{logit}[\pi(\mathbf{x})] = -5,1874 + 0,5238\text{familar\_pe} + 0,093\text{imc} \\ + 0,0406\text{oftPD1} + 1,0486\text{pe\_ant} + 0,7921\text{paridade} \quad (5.1)\end{aligned}$$

Logo, o logaritmo natural da razão de chances de uma paciente ter PE é explicado pela combinação linear das variáveis familar_pe, imc, oftPD1, pe_ant e paridade. No caso das variáveis categóricas, o modelo foi ajustado tendo como base o valor relacionado ao

fator de risco. Desta forma, deseja-se verificar se mulheres com histórico familiar de PE na família (valor = 1) tem um risco maior de ter PE, sendo que o mesmo raciocínio vale para mulheres que já tiveram PE anteriormente (valor = 1). No caso da variável paridadade, o valor = 0 foi utilizado para verificar se mulheres que estão na primeira gravidez tem um maior risco de ter PE.

Percebe-se que no modelo (5.1) todos os coeficientes estimados são positivos, assim todos eles contribuem para o aumento da probabilidade da ocorrência de pré-eclampsia, logo, todas as variáveis do modelo são fatores de risco. Na Tabela 15 encontram-se as estimativas pontuais e os respectivos intervalos de confiança das razões de chances para cada coeficiente do modelo, com 95% de confiança. Como o valor 1 não está contido em nenhum dos intervalos calculados, a chance de PE difere para cada nível de fator nas variáveis categóricas, assim como a chance difere para cada unidade de aumento nas variáveis contínuas.

Através da Tabela 15 é possível fazer as seguintes interpretações para os coeficientes do modelo logístico:

- Pacientes com histórico familiar de pré-eclâmpsia são 2,85 mais propícias a terem PE do que pacientes que não têm esse histórico.
- Para cada unidade de aumento no IMC da paciente, a chance de PE aumenta em 9,7%.
- Mulheres têm seu risco aumentado em 4,1% para cada unidade de aumento do pico diastólico da pressão arterial oftálmica.
- Pacientes que tiveram pré-eclâmpsia anteriormente são 8,14 vezes mais propícias a desenvolverem PE do que mulheres que não tiverem PE anteriormente.
- Mulheres que nunca tiveram filhos têm aproximadamente 4,9 vezes mais chances de terem PE do que mulheres que já têm um filho ou mais.

Tabela 15 - Estimativas das razões de chances dos coeficientes.

Efeito	**Estimativa pontual**	**Intervalo de confiança**	
familar_pe 1 vs 0	2,851	1,187	6,844
imc	1,097	1,023	1,177
oftPD1	1,041	1,001	1,084
pe_ant 1 vs 0	8,143	2,481	26,724
paridade 0 vs 1	4,876	1,713	13,882

5.4.3 *Avaliação do modelo*

Para cada perfil de paciente é obtida a probabilidade dela vir a ter pré-eclâmpsia, porém somente essa probabilidade não diz se ela vai ter PE ou não. As probabilidades (ou escores do modelo) para cada paciente encontram-se na Tabela 26 (Apêndice).

É preciso um critério para escolha de um ponto de corte, que está compreendido entre 0 e 1. O ponto de corte é o valor a partir do qual uma paciente possa ter pré-eclâmpsia, em que o ideal é estabelecer esse ponto com base em algum critério baseado nas taxas de detecção ou má classificação do modelo, no qual deseja-se que o modelo ajustado seja bastante sensível e bastante específico. Desta forma, o critério utilizado nesta aplicação foi a maximização da soma da sensibilidade com a especificidade.

Com base na Tabela 28 (Apêndice), tem-se que o modelo tem a soma de sua sensibilidade e especificidade maximizada quando o ponto de corte é igual a 0,06. Então, diz-se que uma paciente terá PE quando o seu perfil produzir uma probabilidade de PE superior a 0,06. Outra ferramenta utilizada para avaliar a qualidade de ajuste do modelo é a curva ROC. A curva ROC para o modelo em questão encontra-se na Figura 2, na qual mostra o ganho em sensibilidade à medida que a taxa de falso-positivo (1-especificidade) aumenta e tem uma área sob a curva de 0,8277, sugerindo que o modelo é bastante eficiente em discriminar pacientes que terão ou não pré-eclâmpsia.

A Tabela 16 mostra a tabela de classificação do modelo com base no ponto de corte igual a 0,06. Os demais resultados dos testes de diagnóstico do modelo encontram-se na Tabela 17.

Tabela 16 - Tabela de classificação do modelo.

Previsão do modelo	**Situação real**		**Total**
	PE = 1	PE = 0	
PE = 1	23	119	142
PE = 0	8	337	345
Total	31	456	487

Tabela 17 - Testes de diagnóstico do modelo.

Indicadores	**Valor**
Capacidade de acerto total	73,9%
Sensibilidade	74,2%
Especificidade	73,9%
Taxa de falso-positivo	26,1%
Taxa de falso-negativo	25,8%
Valor preditivo positivo	83,8%
Valor preditivo negativo	97,7%
Razão de probabilidade positiva	2,8
Razão de probabilidade negativa	0,3
Prevalência	6,4%

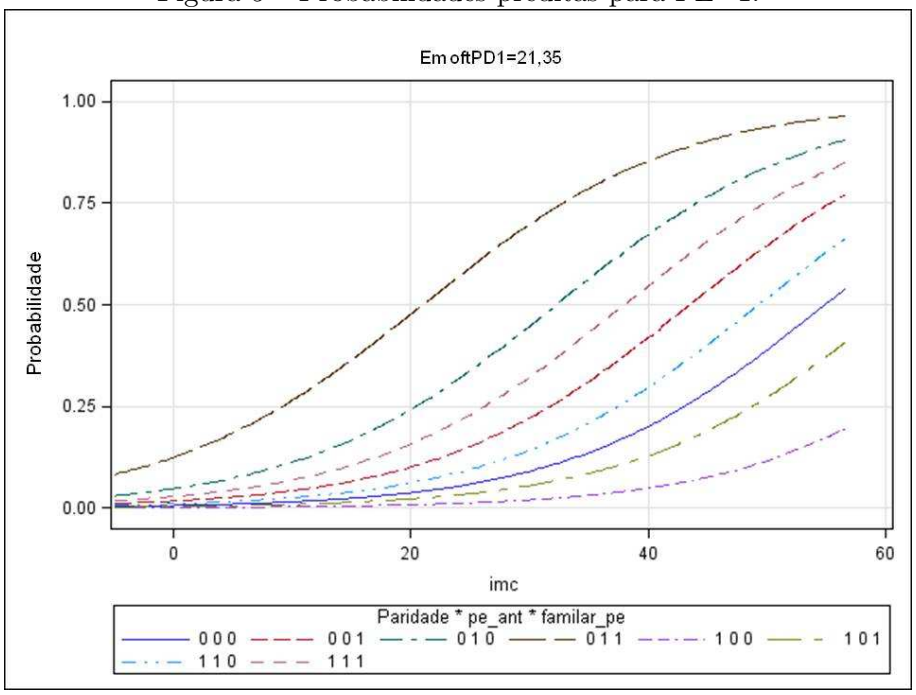
A partir dos resultados apresentados na Tabela 17, seguem as interpretações de cada teste de diagnóstico:

- A capacidade de acerto total foi de 73,9%, indicando que o modelo classificou corretamente 73,9% das 487 pacientes estudadas.
- 74,2% das pacientes previstas pelo modelo com PE, realmente têm a doença. Desta maneira, o modelo tem uma sensibilidade de 74,2%.
- 73,9% das pacientes previstas pelo modelo sem PE, realmente não têm a doença. Assim, o modelo tem uma especificidade de 73,9%.
- 26,1% das pacientes previstas pelo modelo com PE, não têm a doença, ou seja, foram classificadas incorretamente como doentes pelo modelo ajustado.
- 25,8% das pacientes previstas pelo modelo sem PE, são doentes, ou seja, foram classificadas incorretamente como não-doentes pelo modelo ajustado.
- A probabilidade da paciente ter PE, dado que o modelo a classificou como doente é de 83,8%.
- A probabilidade da paciente não ter PE, dado que o modelo a classificou como não-doente é de 97,7%.
- A razão entre a sensibilidade do modelo ajustado e sua taxa de erro falso-positivo é 2,8.
- A razão entre a taxa de erro falso-negativo do modelo ajustado e sua especificidade é 0,3.

- A prevalência de pacientes com PE dentre as mulheres estudadas é de 6,4%.

Na Figura 5 estão as probabilidades preditas para a ocorrência de pré-eclampsia para cada perfil de paciente, em que no eixo x estão os valores do imc e no eixo y está a probabilidade de ocorrência de PE. Nota-se que uma paciente que ainda não teve filhos, já teve PE anteriormente e tem casos de PE na família tem um risco maior de vir a sofrer de PE à medida que o imc aumenta. Enquanto que mulheres que já têm pelo menos um filho, não teve PE anteriormente e nem têm casos de PE na família têm um risco bem menor de ter a doença.

Figura 5 - Probabilidades preditas para PE=1.



5.4.4 *Análise de resíduos e diagnósticos*

Nos gráficos mostrados nas Figuras 6 e 7, o eixo vertical representa o valor do diagnóstico, enquanto que o eixo horizontal representa as pacientes do estudo. Esses gráficos mostram que os resíduos estão relativamente próximos de zero, sugerindo um bom ajuste para o modelo. Vale destacar as observações #54 e #264 nos gráficos **Resíduo de Pearson**, **DIFCHISQ** e **DIFDEV** como possíveis outliers, enquanto que as observações #87 e #333 destacam-se nos gráficos **C** e **Cbar**, e a observação #169 destaca-se no gráfico **Diagonal da Matriz H** também como possível *outlier*.

Descrição dos possíveis outliers.

A paciente #54 tem um imc no valor de 30,36, que segundo a OMS (Organização Mundial da Saúde), ela está com obesidade classe I. A paciente #87 também tem obesidade classe I ($\text{imc} = 30,08$) e também tem um valor de oftPD1 igual a 65,13, que é considerado bem elevado, segundo as estatísticas descritivas apresentadas anteriormente. Já a paciente #264 está com excesso de peso ($\text{imc} = 25,03$) e o seu valor da oftPD1 igual a 18,61 não é considerado um outlier. É importante ressaltar que todas estas pacientes são casos de PE, apesar delas não possuírem os demais fatores de risco, isto é, ambas já têm pelo menos um filho ($\text{paridade} = 1$), não têm casos de PE na família ($\text{familar\_pe} = 0$) e não sofreram de PE em gestações anteriores ($\text{pe\_ant} = 0$).

As pacientes #169 e #333 não são casos de PE, porém a paciente #169 tem o fator de risco $\text{familar\_pe} = 1$ e um elevado valor da oftPD1 (57,25), enquanto que a paciente #333 tem os fatores de risco $\text{familar\_pe} = 1$, $\text{pe\_ant} = 1$ e um grande valor do imc (41,81), sendo classificado como obesidade classe III.

A retirada dessas observações não traria grandes benefícios para o modelo, já que algumas delas são casos de PE, o que já são poucos casos e essa retirada iria diminuir ainda mais a prevalência de mulheres doentes no estudo.

Figura 6 - Diagnósticos de influência das observações.

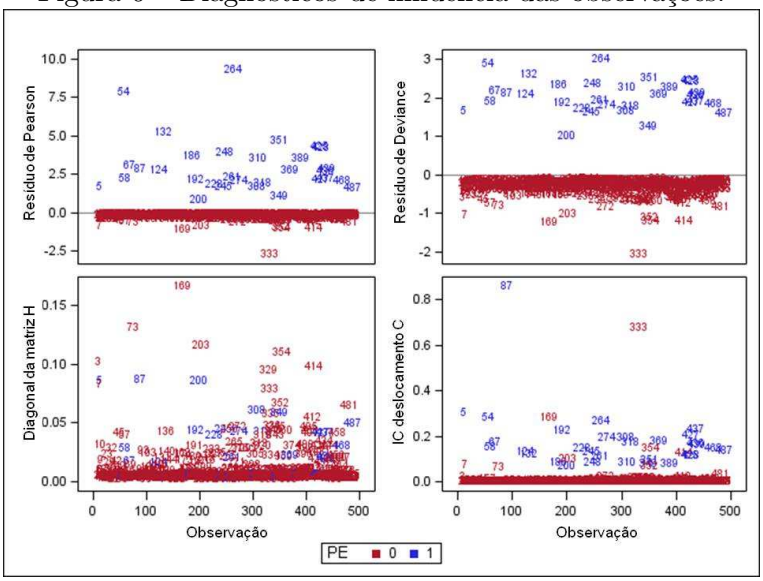
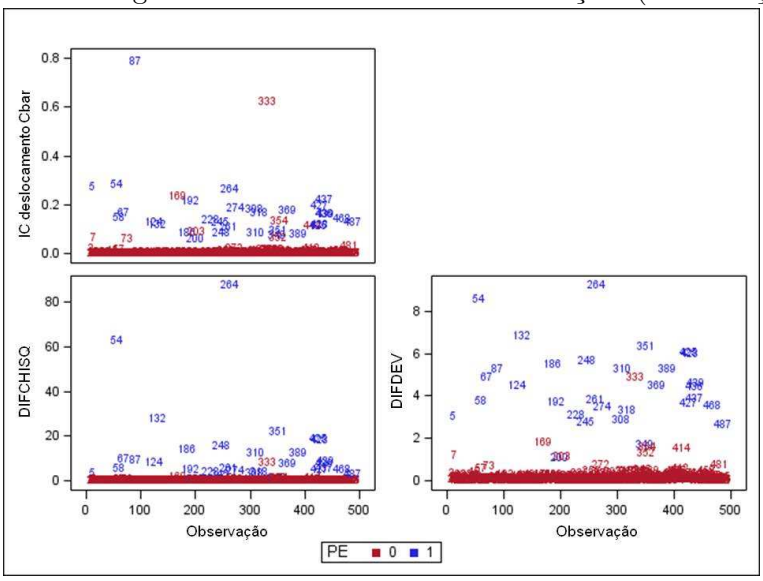


Figura 7 - Diagnósticos de influência das observações (continuação).



Os gráficos para os **DFBetas** mostrados nas Figuras 8 e 9 indicam que a observação #333 mostrou-se influente nas estimativas dos parâmetros de familar_pe1 e imc. Para as estimativas dos parâmetros de oftPD1 destacam-se as observações #5, #87 e #169. Já para os parâmetros pe_ant1 e paridade0, as observações #54, #87, #264, #333 e #354 mostraram-se influentes.

Figura 8 - Diagnósticos de influências das estimativas dos parâmetros.

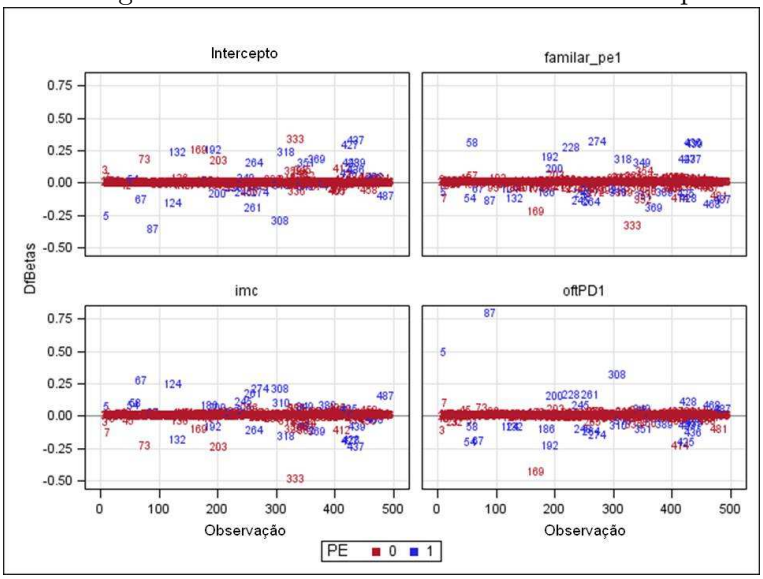
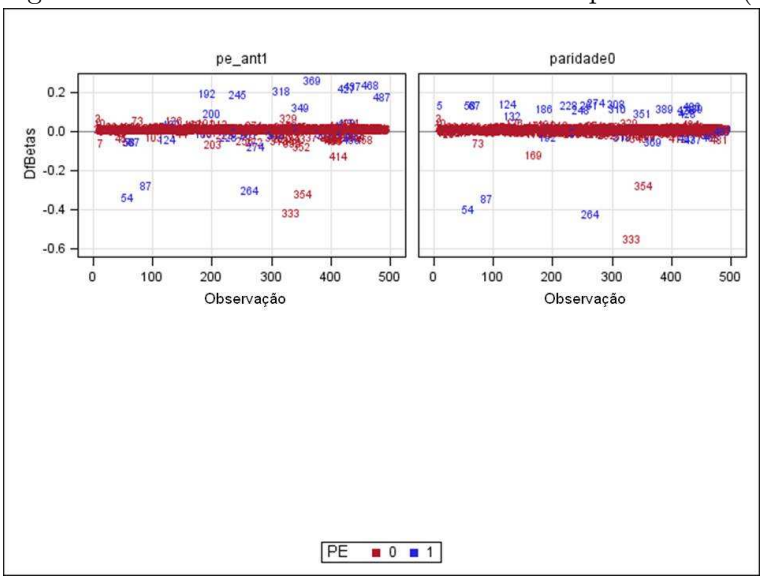
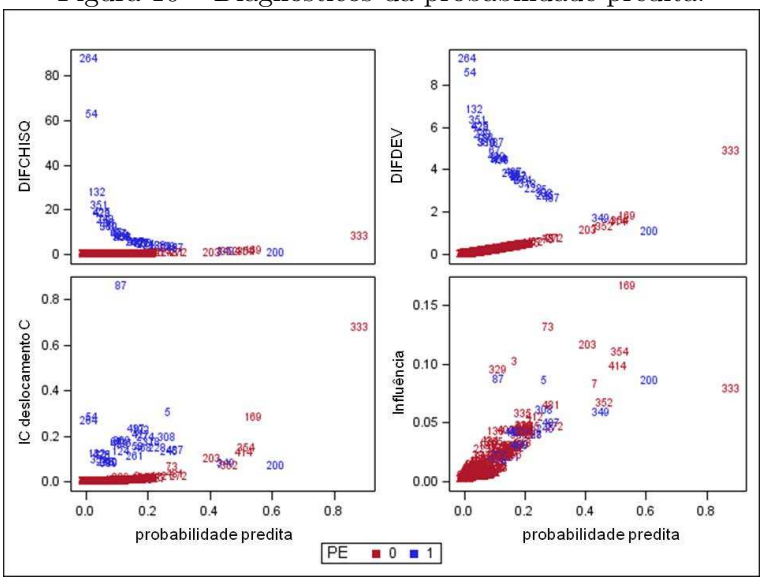


Figura 9 - Diagnósticos de influências das estimativas dos parâmetros (continuação).



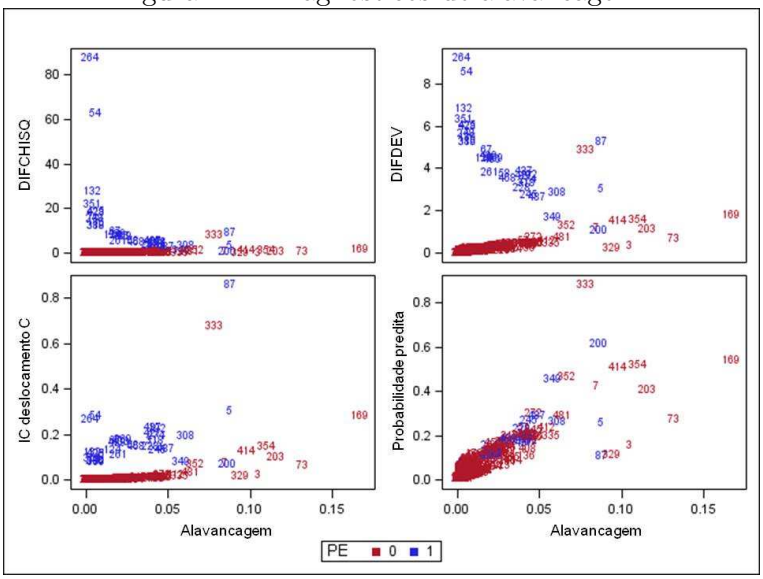
Nos gráficos de diagnósticos contra as probabilidades previstas (Figura 10) as observações #54, #264 e #333 se destacam nos gráficos **DIFCHISQ** e **DIFDEV**. Enquanto que no gráfico **IC deslocamento C** destacam-se as observações #87 e #333. No gráfico de **Influência** as observações #169 e #333 destacam-se como outliers.

Figura 10 - Diagnósticos da probabilidade predita.



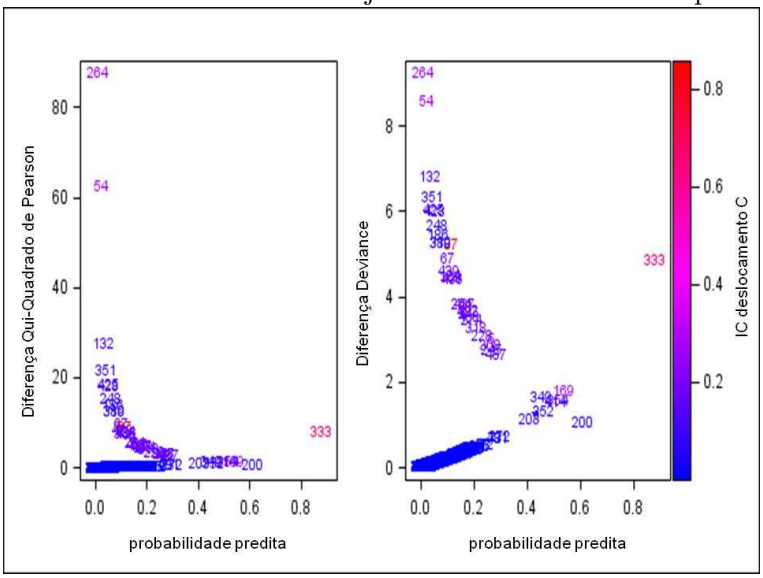
As mesmas observações citadas nos gráficos da Figura 10 também mostraram-se influentes nos gráficos de diagnósticos contra a alavancagem mostrados na Figura 11.

Figura 11 - Diagnósticos de alavancagem.



A Figura 12 mostra os diagnósticos de exclusão contra as probabilidades e as cores das observações previstas de acordo com o intervalo de confiança de deslocamento do diagnóstico, no qual destacam-se as observações #54, #264 e #333.

Figura 12 - Influência no modelo ajustado e estimativas dos parâmetros.



6 CONCLUSÕES

O modelo logístico é frequentemente utilizado em situações em que a variável resposta é de natureza binária (sucesso e fracasso), na qual seus valores são expressos em termos de probabilidade. Para entender melhor o contexto de regressão logística, foi feita uma abordagem dos modelos lineares generalizados no Capítulo 2, para que em seguida, a teoria do modelo logístico (Capítulo 3) fosse mostrada.

Os testes de diagnósticos são essenciais para uma avaliação do modelo ajustado, pois a partir deles pode-se ter conhecimento da sua capacidade de ajuste, sensibilidade, especificidade e Curva ROC do modelo. Tais resultados normalmente são utilizados como critério para o ponto de corte na classificação da variável de interesse. Logo, a partir de um determinado valor de probabilidade, a variável resposta é classificada como sucesso.

O principal objetivo desta aplicação foi ajustar um modelo logístico capaz de prever se uma mulher grávida pode vir a ter pré-eclâmpsia. O modelo foi adquirido a partir de um estudo com 487 mulheres grávidas, das quais 31 desenvolveram a doença. Apesar de poucos casos em relação ao total, conseguiu-se um bom ajuste para o modelo.

Depois de efetuada a seleção das 23 variáveis explicativas através do método *stepwise* (escolhido como o melhor modelo entre os métodos *backward* e *forward*), cinco delas foram selecionadas, das quais três são categóricas (familar_pe, pe_ant e paridade) e duas são contínuas (imc e oftPD1). O modelo logístico ajustado atende os pré-requisitos de qualidade do ajuste, tais como o teste de Hosmer e Lemeshow e os testes de nulidade global dos coeficientes. Com base nos testes de diagnóstico, o modelo tem aproximadamente 83% de eficiência em discriminar pacientes que terão ou não pré-eclâmpsia, sensibilidade de 74,2% e especificidade de 73,9%, sendo essas taxas consideradas boas. O critério para a escolha do ponto de corte foi a maximização da soma da sensibilidade com a especificidade, sendo esse ponto igual a 0,06, ou seja, diz-se que uma paciente terá PE quando o seu perfil produzir uma probabilidade de PE superior a 0,06.

A análise de resíduos e diagnósticos mostrou alguns pontos influentes, porém optou-se

por não removê-los, já que a maioria deles são casos de pré-eclâmpsia, que já são poucos na amostra, além de que valores extremos de algumas variáveis também contribuíram para a alavancagem desses pontos.

Algumas variáveis que surgem com frequência na literatura como sendo associadas com PE não se mostraram significativas no modelo, por exemplo, hipertensão crônica e diabetes. Isso se deve ao fato de os modelos serem estudos independentes, provindos de populações diferentes, com peculiaridades culturais e geográficas distintas. Um estudo de meta-análise mais aprofundado poderia considerar esses aspectos das amostras, possivelmente envolvendo dados adicionais sobre hábitos das pacientes.

REFERÊNCIAS

AGRESTI, A. **Categorical data analysis.** 2nd ed. Gainesville : University of Florida, 2002. 710 p.

AKAIKE, H. A new look at statistical model identification. **IEEE Transaction on Automatic Control**, v. AC-19, p.716-723, dec. 1974.

ALLISON, P. D. **Logistic regression using the SAS system:** theory and application. Cary, N.C.: SAS Institute, 1999. 304 p.

ALVES, J. A G. **Predição de Pré-eclâmpsia através da associação de fatores maternos à avaliação tríplice vascular no primeiro trimestre de gestação**. 2012. 146f. Tese (Doutorado em Saúde Coletiva) - Doutorado em Saúde Coletiva em Associação AMPLA de IES UECE/UFC/UNIFOR, Universidade de Estadual do Ceará, Universidade Federal do Ceará, Fortaleza, 2012.

ANDERSEN, E. B. **Introduction to the statistical analysis of categorical data**. New York: Springer, 1996.

BRANDÃO, A. H. F. et al. Dilatação fluxo-mediada da artéria braquial como método de avaliação da função endotelial na pré-eclâmpsia e em gestantes normotensas. **Rev Med Minas Gerais**, Belo Horizonte, v. 21, p. 9-13, jan. 2011.

BRANDÃO, A. H. F. et al. Predição de pré-eclâmpsia: a realidade atual e as direções futuras. **FEMINA**, Belo Horizonte, v. 38, p. 488-491, set. 2010.

CHRISTENSEN, R. **Log-linear models & logistic regression**. New York: Springer-Verlang, 1997, 500p.

CORDEIRO, G.M. **Modelos Lineares Generalizados**. Campinas: VII SINAPE, 1986. 286p.

CORRAR, L. J.; PAULO, E.; DIAS FILHO, J. M. **Análise multivariada para os cursos de administração, ciências contábeis e economia**. São Paulo: Fundação Instituto de Pesquisas Contábeis, Atuariais e Financeiras, 2009. 344p

COSTA, S. C. **Regressão logística aplicada na identificação de fatores de risco**

para doenças animais domésticos. 1997. 104f. Dissertação (Mestrado em Agronomia) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 1997.

COX, D. R.; SNELL, E.J. **Analysis of binary data**. London: Chapman &Hall, 1989. 236p.

DEMÉTRIO, C. G. B. **Modelos lineares generalizados em experimentação agronômica**. Piracicaba: ESALQ-USP, 2002. 121p.

DRAPER, N.R.; SMITH, R. **Applied regression analysis**. 3rd. ed. New York: John Wiley, 1996. 706p.

FIGUEIRA, C. V. **Modelos de regressão logística**. 2006. 149f. Dissertação (Mestrado em Matemática) - Instituto de Matemática, Universidade Federal do Rio Grande do Sul, Porto Alegre, 2006.

GUJARATI, D. N. **Econometria básica**. 4. ed. Editora Elsevier, 2006. 812p.

HOFFMANN, R. **Análise de regressão:** uma introdução à econometria. 4. ed. São Paulo: HUCITEC/EDUSP, 2006. 379p.

HOSMER, D.; LEMESHOW, S. **Applied logistic regression**. New York: John Wiley, 1989. 307p.

JEKEL, J. F.; KATZ, D. L.; ELMORE, J. G. **Epidemiologia, bioestatística e medicina preventiva**. 2. ed. Porto Alegre: Artmed, 2006. 432p.

KELINBAUM, D. G.; KLEIN, M. **Logistic regression, a self-learning text**. 3rd. ed. New York: Springer, 2010. 701p.

LOBATO JÚNIOR, D. **Influência local em modelos de regressão**. 2005. 106f. Dissertação (Mestrado em Matemática) - Centro de Ciências e Tecnologia, Universidade Federal de Campina Grande, Campina Grande, 2005.

MCCULLAGH, P.; NELDER, J. **Generalized linear models**. 2nd. ed. London: Chapman & Hall, 1989. 532p.

NELDER, J. A.; WEDDERBURN, R. W. M. Generalized Linear Models. **Journal of the Royal Statistical Society Series A**, v. 135, p. 370-384, 1972.

MELCA, L. A. V. **Estudo do índice de resistência da artéria oftálmica como preditor dopplerfluxométrico da pré-eclâmpsia**. 2007. 70f. Dissertação (Mestrado em Saúde da Mulher) - Faculdade de Medicina, Universidade Federal de Minas Gerais,

Belo Horizonte, 2007.

PAGANO, M.; GAUVREAU, K. **Princípios de bioestatística.** 2. ed. São Paulo: Cengage Learning, 2004. 506p.

PAULA, G. A. **Modelos de regressão com apoio computacional**. São Paulo: IME-USP, 2010. 403p.

PAULA, L. G. **Eclâmpsia e pré-eclâmpsia:** estudo comparativo e experiência no Hospital São Lucas da PUCRS. 2010. 90f. Tese (Doutorado em Medicina e Ciências da Saúde) - Faculdade de Medicina, Pontifícia Universidade Católica do Rio Grande do Sul, Porto Alegre, 2010.

PINO, F. A. Modelos de decisão binários: uma revisão. **Rev. de Economia Agrícola**, São Paulo, v. 54, p.43-57, jan./jun. 2007.

PLASENCIA, W. et al. Uterine artery Doppler at 11 + 0 to 13 + 6 weeks and 21 + 0 to 24 + 6 weeks in the prediction of pre-eclampsia. **Ultrasound Obstet Gynecol**, London, v.32, p.138-146, abr. 2008.

QUEIROZ, N. M. O. B. **Regressão logística:** uma estimativa bayesiana aplicada na identificação de fatores de risco para HIV, em doadores de sangue. 2004. 95f. Dissertação (Mestrado em Biometria) - Departamento de Física e Matemática, Universidade Federal Rural de Pernambuco, Recife, 2004.

SCHWARZ, G. Estimating the dimensional of a model. **Annals of Statistics**, Hayward, v.6, n.2, p.461-464, mar. 1978.

SILVA, G. L. **Modelos logísticos para dados binários**. 1992. 118f. Dissertação (Mestrado em Estatística) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo, 1992.

SILVEIRA, D. M. I. et al. **Gestação de alto risco:** manual técnico. 3. ed. Brasília: Ministério da Saúde, 2000. 164p.

SOUZA, E. C. **Análise de influência local no modelo de regressão logística**. 2006. 101f. Dissertação (Mestrado em Agronomia) - Escola Superior de Agricultura Luiz de Queiroz, Universidade de São Paulo, Piracicaba, 2006.

STOKES, M. E.; CHARLES S. D.; GARY G. K. **Categorical data analysis using the SAS System**. 2nd. ed. USA: SAS Institute Inc., 2000. 626p.

TURKMAN, M.; SILVA, G. **Modelos lineares generalizados:** da teoria à prática.

Lisboa: UL/UTL, 2000, 151p.

VENABLES, W. N.; RIPLEY, B. D. **Modern applied statistics with S**. 4th. ed. New York: Springer, 2002. 495p.

APÊNDICE

Tabelas de contigência 2x2

Tabela 18 - PE anterior.

PE anterior	PE		Total
	0	1	
0	423	21	444
1	33	10	43
Total	456	31	487

Tabela 19 - Familiar PE.

Familar PE	PE		Total
	0	1	
0	388	20	408
1	68	11	79
Total	456	31	487

Tabela 20 - Prematuro anterior.

Prematuro anterior	PE		Total
	0	1	
0	434	26	460
1	22	5	27
Total	456	31	487

Tabela 21 - Hipertensão crônica.

Hipertensão crônica	PE		Total
	0	1	
0	438	28	466
1	18	3	21
Total	456	31	487

Tabela 22 - Diabetes.

Diabetes	PE		Total
	0	1	
0	441	31	472
1	15	0	15
Total	456	31	487

Tabela 23 - Incisura - artéria uterina direita.

Incisura - artéria uterina direita	PE		Total
	0	1	
0	226	14	240
1	230	17	247
Total	456	31	487

Tabela 24 - Incisura - artéria uterina esquerda.

Incisura - artéria uterina esquerda	PE		Total
	0	1	
0	224	11	235
1	232	20	252
Total	456	31	487

Tabela 25 - Paridade.

Paridade	PE		Total
	0	1	
0	223	18	241
1	233	13	246
Total	456	31	487

Gráficos das variáveis categóricas

Figura 13 - PE anterior, familiar PE e parto prematuro anteriormente.

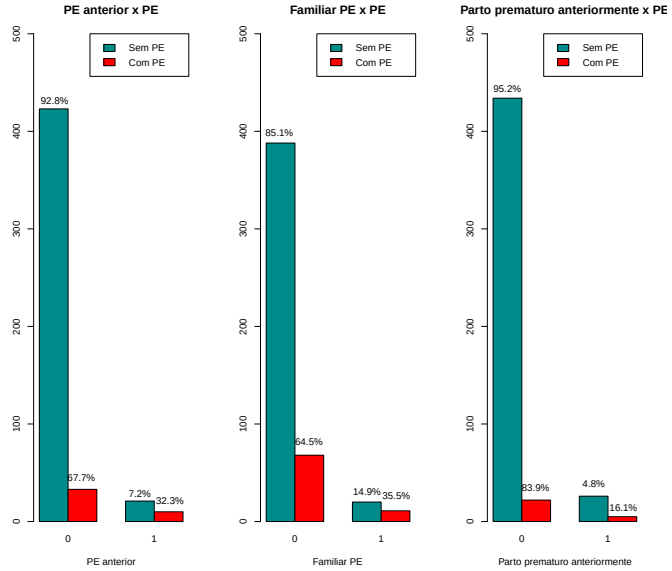


Figura 14 - Hipertensão crônica, diabetes e artéria uterina direita com incisura.

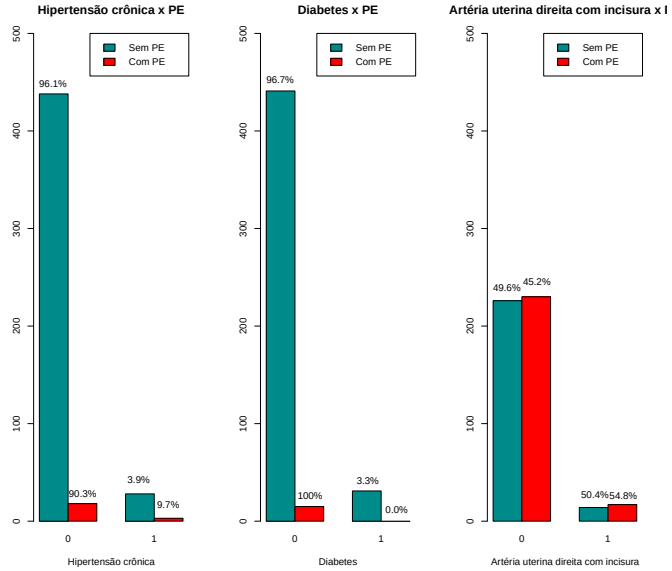
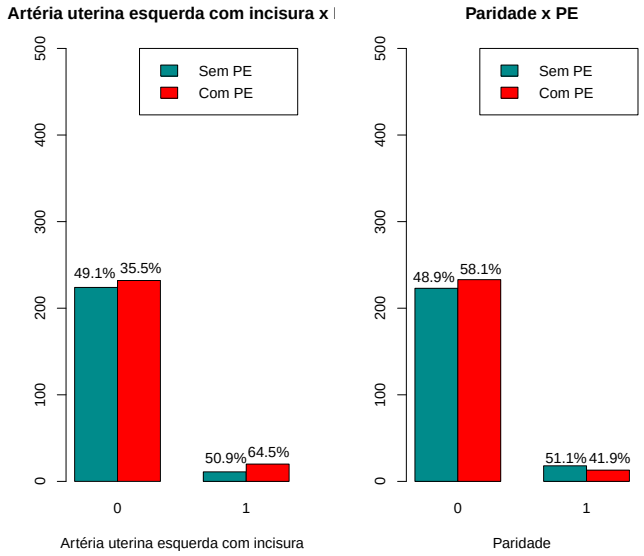


Figura 15 - Artéria uterina esquerda com incisura e paridade.



Gráficos das variáveis contínuas

Figura 16 - Idade, imc, pressão uterina direita-IR e AB.

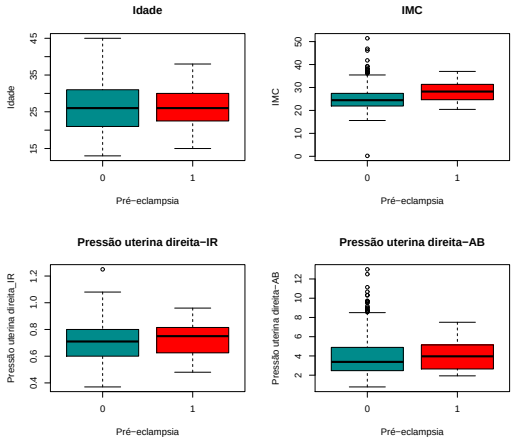


Figura 17 - Pressão uterina esquerda-(IR e AB), média-IP e menor-IP.

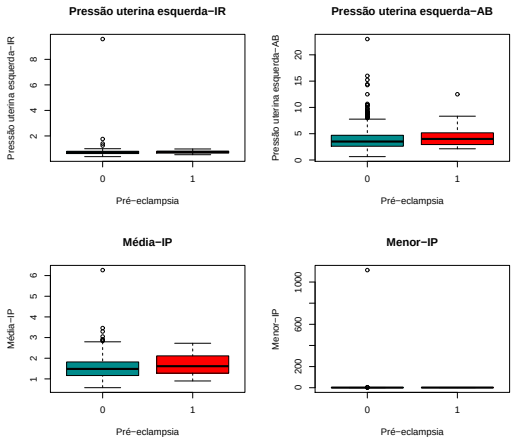


Figura 18 - Pressão oftálmica-(IR,IP, AB e PS).

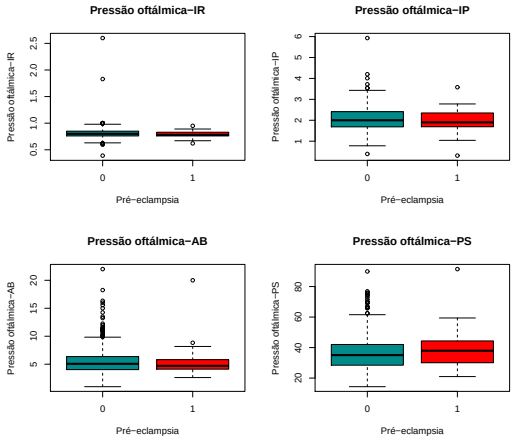
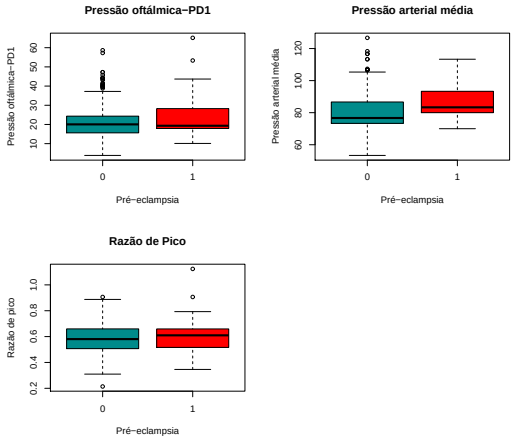


Figura 19 - Pressão oftálmica-PD1, pressão arterial média e razão de pico.



Tabelas

Tabela 26 - Escores do modelo logístico para PE.

Paciente	PE	Escores(PE=1)	Paciente	PE	Escores(PE=1)	Paciente	PE	Escores(PE=1)
1	0	0,023	83	0	0,037	165	0	0,017
2	0	0,021	84	0	0,022	166	0	0,011
3	0	0,154	85	0	0,028	167	0	0,075
4	0	0,017	86	0	0,022	168	0	0,119
5	1	0,255	87	1	0,108	169	0	0,539
6	0	0,023	88	0	0,061	170	0	0,040
7	0	0,420	89	0	0,009	171	0	0,134
8	0	0,032	90	0	0,046	172	0	0,041
9	0	0,073	91	0	0,035	173	0	0,057
10	0	0,086	92	0	0,049	174	0	0,025
11	0	0,011	93	0	0,109	175	0	0,018
12	0	0,047	94	0	0,038	176	0	0,026
13	0	0,105	95	0	0,071	177	0	0,051
14	0	0,055	96	0	0,007	178	0	0,116
15	0	0,047	97	0	0,020	179	0	0,065
16	0	0,013	98	0	0,032	180	0	0,046
17	0	0,011	99	0	0,052	181	0	0,036
18	0	0,030	100	0	0,053	182	0	0,018
19	0	0,012	101	0	0,057	183	0	0,016
20	0	0,017	102	0	0,043	184	0	0,044
21	0	0,056	103	0	0,138	185	0	0,015
22	0	0,020	104	0	0,030	186	1	0,068
23	0	0,129	105	0	0,023	187	0	0,021
24	0	0,071	106	0	0,029	188	0	0,070
25	0	0,014	107	0	0,040	189	0	0,133
26	0	0,017	108	0	0,036	190	0	0,031
27	0	0,014	109	0	0,014	191	0	0,079
28	0	0,011	110	0	0,030	192	1	0,177
29	0	0,041	111	0	0,017	193	0	0,038
30	0	0,031	112	0	0,040	194	0	0,035
31	0	0,010	113	0	0,031	195	0	0,034
32	0	0,120	114	0	0,036	196	0	0,022
33	0	0,073	115	0	0,034	197	0	0,045
34	0	0,012	116	0	0,043	198	0	0,041
35	0	0,007	117	0	0,051	199	0	0,064
36	0	0,045	118	0	0,011	200	1	0,611
37	0	0,050	119	0	0,045	201	0	0,024
38	0	0,035	120	0	0,051	202	0	0,025
39	0	0,008	121	0	0,033	203	0	0,408
40	0	0,018	122	0	0,038	204	0	0,018
41	0	0,053	123	0	0,009	205	0	0,008
42	0	0,049	124	1	0,114	206	0	0,035
43	0	0,032	125	0	0,013	207	0	0,097
44	0	0,055	126	0	0,011	208	0	0,034
45	0	0,185	127	0	0,031	209	0	0,031
46	0	0,016	128	0	0,017	210	0	0,012
47	0	0,094	129	0	0,028	211	0	0,048
48	0	0,079	130	0	0,017	212	0	0,065
49	0	0,037	131	0	0,115	213	0	0,067
50	0	0,025	132	1	0,035	214	0	0,018
51	0	0,069	133	0	0,009	215	0	0,095
52	0	0,023	134	0	0,021	216	0	0,008
53	0	0,008	135	0	0,058	217	0	0,006
54	1	0,016	136	0	0,106	218	0	0,027
55	0	0,027	137	0	0,012	219	0	0,012
56	0	0,048	138	0	0,066	220	0	0,030
57	0	0,233	139	0	0,088	221	0	0,013
58	1	0,164	140	0	0,130	222	0	0,047
59	0	0,042	141	0	0,008	223	0	0,062
60	0	0,012	142	0	0,072	224	0	0,009
61	0	0,010	143	0	0,044	225	0	0,059
62	0	0,037	144	0	0,063	226	0	0,009
63	0	0,063	145	0	0,010	227	0	0,117
64	0	0,036	146	0	0,056	228	1	0,230
65	0	0,010	147	0	0,022	229	0	0,019
66	0	0,034	148	0	0,031	230	0	0,049
67	1	0,095	149	0	0,014	231	0	0,012
68	0	0,039	150	0	0,013	232	0	0,022
69	0	0,047	151	0	0,029	233	0	0,084
70	0	0,068	152	0	0,040	234	0	0,034
71	0	0,017	153	0	0,011	235	0	0,146
72	0	0,037	154	0	0,018	236	0	0,030
73	0	0,273	155	0	0,015	237	0	0,008
74	0	0,009	156	0	0,007	238	0	0,013
75	0	0,058	157	0	0,007	239	0	0,022
76	0	0,015	158	0	0,019	240	0	0,008
77	0	0,033	159	0	0,012	241	0	0,007
78	0	0,047	160	0	0,006	242	0	0,051
79	0	0,010	161	0	0,061	243	0	0,022
80	0	0,042	162	0	0,023	244	0	0,005
81	0	0,026	163	0	0,015	245	1	0,268
82	0	0,074	164	0	0,036	246	0	0,044

Tabela 27 - Escores do modelo logístico para PE (continuação).

Paciente	PE	Escores(PE=1)	Paciente	PE	Escores(PE=1)	Paciente	PE	Escores(PE=1)
247	0	0,025	328	0	0,016	408	0	0,141
248	1	0,062	329	0	0,112	409	0	0,052
249	0	0,010	330	0	0,008	410	0	0,015
250	0	0,021	331	0	0,033	411	0	0,077
251	0	0,041	332	0	0,058	412	0	0,235
252	0	0,009	333	0	0,880	413	0	0,093
253	0	0,052	334	0	0,057	414	0	0,507
254	0	0,073	335	0	0,195	415	0	0,106
255	0	0,108	336	0	0,200	416	0	0,023
256	0	0,197	337	0	0,024	417	0	0,146
257	0	0,049	338	0	0,021	418	0	0,038
258	0	0,037	339	0	0,009	419	0	0,031
259	0	0,009	340	0	0,179	420	0	0,106
260	0	0,032	341	0	0,018	421	0	0,040
261	1	0,157	342	0	0,031	422	0	0,046
262	0	0,017	343	0	0,183	423	0	0,025
263	0	0,009	344	0	0,035	424	0	0,042
264	1	0,011	345	0	0,222	425	1	0,052
265	0	0,100	346	0	0,032	426	0	0,147
266	0	0,021	347	0	0,020	427	1	0,179
267	0	0,010	348	0	0,103	428	1	0,053
268	0	0,016	349	1	0,452	429	0	0,049
269	0	0,032	350	0	0,009	430	0	0,026
270	0	0,122	351	1	0,044	431	0	0,116
271	0	0,055	352	0	0,463	432	0	0,045
272	0	0,300	353	0	0,036	433	0	0,059
273	0	0,017	354	0	0,515	434	0	0,085
274	1	0,194	355	0	0,064	435	0	0,038
275	0	0,037	356	0	0,023	436	1	0,120
276	0	0,157	357	0	0,055	437	1	0,164
277	0	0,007	358	0	0,019	438	0	0,014
278	0	0,022	359	0	0,133	439	1	0,108
279	0	0,072	360	0	0,202	440	0	0,010
280	0	0,032	361	0	0,053	441	0	0,081
281	0	0,026	362	0	0,027	442	0	0,035
282	0	0,013	363	0	0,069	443	0	0,046
283	0	0,021	364	0	0,018	444	0	0,152
284	0	0,038	365	0	0,011	445	0	0,024
285	0	0,020	366	0	0,011	446	0	0,011
286	0	0,016	367	0	0,013	447	0	0,036
287	0	0,064	368	0	0,006	448	0	0,033
288	0	0,056	369	1	0,117	449	0	0,085
289	0	0,009	370	0	0,038	450	0	0,018
290	0	0,038	371	0	0,011	451	0	0,011
291	0	0,011	372	0	0,075	452	0	0,051
292	0	0,172	373	0	0,035	453	0	0,007
293	0	0,088	374	0	0,107	454	0	0,015
294	0	0,014	375	0	0,023	455	0	0,046
295	0	0,013	376	0	0,011	456	0	0,164
296	0	0,113	377	0	0,018	457	0	0,045
297	0	0,017	378	0	0,037	458	0	0,204
298	0	0,034	379	0	0,051	459	0	0,014
299	0	0,010	380	0	0,053	460	0	0,127
300	0	0,047	381	0	0,080	461	0	0,067
301	0	0,008	382	0	0,091	462	0	0,029
302	0	0,052	383	0	0,014	463	0	0,063
303	0	0,072	384	0	0,029	464	0	0,060
304	0	0,032	385	0	0,013	465	0	0,044
305	0	0,135	386	0	0,060	466	0	0,012
306	0	0,046	387	0	0,041	467	0	0,116
307	0	0,024	388	0	0,027	468	1	0,184
308	1	0,263	389	1	0,075	469	0	0,008
309	0	0,013	390	0	0,026	470	0	0,022
310	1	0,076	391	0	0,063	471	0	0,030
311	0	0,026	392	0	0,074	472	0	0,026
312	0	0,035	393	0	0,018	473	0	0,014
313	0	0,195	394	0	0,031	474	0	0,050
314	0	0,055	395	0	0,024	475	0	0,010
315	0	0,013	396	0	0,141	476	0	0,028
316	0	0,048	397	0	0,095	477	0	0,064
317	0	0,026	398	0	0,161	478	0	0,017
318	1	0,210	399	0	0,034	479	0	0,013
319	0	0,200	400	0	0,188	480	0	0,036
320	0	0,169	401	0	0,029	481	0	0,287
321	0	0,080	402	0	0,036	482	0	0,013
322	0	0,040	403	0	0,043	483	0	0,055
323	0	0,060	404	0	0,012	484	0	0,009
324	0	0,030	405	0	0,194	485	0	0,072
325	0	0,011	406	0	0,034	486	0	0,018
326	0	0,027	407	0	0,193	487	1	0,287
327	0	0,046						

Tabela 28 - Tabela de classificação geral do modelo.

Nível de probabilidade	Correto		Incorreto		Correto	Porcentagens	
	PE=1	PE=0	PE=1	PE=0		Sensibilidade	Especificidade
0,000	31	0	456	0	6,4	100	0
0,010	30	38	418	1	14	96,8	8,3
0,020	29	131	325	2	32,9	93,5	28,7
0,030	29	191	265	2	45,2	93,5	41,9
0,040	28	259	197	3	58,9	90,3	56,8
0,050	24	304	152	7	67,4	77,4	66,7
0,060	23	337	119	8	73,9	74,2	73,9
0,070	20	359	97	11	77,8	64,5	78,7
0,080	20	375	81	11	81,1	64,5	82,2
0,090	19	385	71	12	83	61,3	84,4
0,100	17	391	65	14	83,8	54,8	85,7
0,110	15	396	60	16	84,4	48,4	86,8
0,120	15	406	50	16	86,4	48,4	89
0,130	15	410	46	16	87,3	48,4	89,9
0,140	14	417	39	17	88,5	45,2	91,4
0,150	10	422	34	21	88,7	32,3	92,5
0,160	9	424	32	22	88,9	29	93
0,170	8	427	29	23	89,3	25,8	93,6
0,180	7	430	26	24	89,7	22,6	94,3
0,190	7	431	25	24	89,9	22,6	94,5
0,200	5	434	22	26	90,1	16,1	95,2
0,210	5	440	16	26	91,4	16,1	96,5
0,220	4	443	13	27	91,8	12,9	97,1
0,230	4	443	13	27	91,8	12,9	97,1
0,240	3	444	12	28	91,8	9,7	97,4
0,250	3	446	10	28	92,2	9,7	97,8
0,260	2	446	10	29	92	6,5	97,8
0,270	2	446	10	29	92	6,5	97,8
0,280	2	446	10	29	92	6,5	97,8
0,290	2	446	10	29	92	6,5	97,8
0,300	2	446	10	29	92	6,5	97,8
0,310	2	447	9	29	92,2	6,5	98
0,320	2	449	7	29	92,6	6,5	98,5
0,330	2	449	7	29	92,6	6,5	98,5
0,340	2	449	7	29	92,6	6,5	98,5
0,350	2	449	7	29	92,6	6,5	98,5
0,360	2	449	7	29	92,6	6,5	98,5
0,370	2	449	7	29	92,6	6,5	98,5
0,380	2	449	7	29	92,6	6,5	98,5
0,390	2	449	7	29	92,6	6,5	98,5
0,400	2	449	7	29	92,6	6,5	98,5
0,410	2	449	7	29	92,6	6,5	98,5
0,420	1	449	7	30	92,4	3,2	98,5
0,430	1	449	7	30	92,4	3,2	98,5
0,440	1	449	7	30	92,4	3,2	98,5
0,450	1	449	7	30	92,4	3,2	98,5
0,460	1	450	6	30	92,6	3,2	98,7
0,470	1	451	5	30	92,8	3,2	98,9
0,480	1	451	5	30	92,8	3,2	98,9
0,490	1	451	5	30	92,8	3,2	98,9
0,500	1	452	4	30	93	3,2	99,1
0,510	1	452	4	30	93	3,2	99,1
0,520	1	452	4	30	93	3,2	99,1
0,530	1	452	4	30	93	3,2	99,1
0,540	1	452	4	30	93	3,2	99,1
0,550	1	452	4	30	93	3,2	99,1
0,560	1	452	4	30	93	3,2	99,1
0,570	1	453	3	30	93,2	3,2	99,3
0,580	0	454	2	31	93,2	0	99,6
0,590	0	454	2	31	93,2	0	99,6
0,600	0	454	2	31	93,2	0	99,6
0,610	0	454	2	31	93,2	0	99,6
0,620	0	454	2	31	93,2	0	99,6
0,630	0	454	2	31	93,2	0	99,6
0,640	0	454	2	31	93,2	0	99,6
0,650	0	455	1	31	93,4	0	99,8
0,660	0	455	1	31	93,4	0	99,8
0,670	0	455	1	31	93,4	0	99,8
0,680	0	455	1	31	93,4	0	99,8
0,690	0	455	1	31	93,4	0	99,8
0,700	0	455	1	31	93,4	0	99,8
0,710	0	455	1	31	93,4	0	99,8
0,720	0	455	1	31	93,4	0	99,8
0,730	0	455	1	31	93,4	0	99,8

Tabela 29 - Tabela de classificação geral do modelo (continuação).

Nível de probabilidade	Correto		Incorreto		Correto	Porcentagens	
	PE=1	PE=0	PE=1	PE=0		Sensibilidade	Especificidade
0,740	0	455	1	31	93,4	0	99,8
0,750	0	455	1	31	93,4	0	99,8
0,760	0	455	1	31	93,4	0	99,8
0,770	0	455	1	31	93,4	0	99,8
0,780	0	455	1	31	93,4	0	99,8
0,790	0	455	1	31	93,4	0	99,8
0,800	0	455	1	31	93,4	0	99,8
0,810	0	455	1	31	93,4	0	99,8
0,820	0	455	1	31	93,4	0	99,8
0,830	0	455	1	31	93,4	0	99,8
0,840	0	455	1	31	93,4	0	99,8
0,850	0	455	1	31	93,4	0	99,8
0,860	0	455	1	31	93,4	0	99,8
0,870	0	455	1	31	93,4	0	99,8
0,880	0	455	1	31	93,4	0	99,8
0,890	0	455	1	31	93,4	0	99,8
0,900	0	455	1	31	93,4	0	99,8
0,910	0	455	1	31	93,4	0	99,8
0,920	0	455	1	31	93,4	0	99,8
0,930	0	455	1	31	93,4	0	99,8
0,940	0	456	0	31	93,6	0	100
0,950	0	456	0	31	93,6	0	100
0,960	0	456	0	31	93,6	0	100
0,970	0	456	0	31	93,6	0	100
0,980	0	456	0	31	93,6	0	100
0,990	0	456	0	31	93,6	0	100
1,000	0	456	0	31	93,6	0	100