Universidade Estadual da Paraíba
Centro de Ciências e Tecnologia
Departamento de Estatística

Marcia de Lourdes Alves Teotônio

Uso da regressão logística para avaliar o efeito de antissépticos em cirurgia hospitalar

Campina Grande
Outubro de 2016

Marcia de Lourdes Alves Teotônio

Uso da regressão logística para avaliar o efeito de antissépticos em cirurgia hospitalar

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

Orientadora:
Profa. Dra. Ana Patricia Bastos Peixoto

Campina Grande
Outubro de 2016

T314u Teotônio, Marcia de Lourdes Alves.
Uso da regressão logística para avaliar o efeito de antissépticos em cirurgia hospitalar [manuscrito] / Marcia de Lourdes Alves Teotônio. - 2016.
29 p. : il.

Digitado.
Trabalho de Conclusão de Curso (Graduação em Estatística) - Universidade Estadual da Paraíba, Centro de Ciências e Tecnologia, 2016.
"Orientação: Profa. Dra. Ana Patricia Bastos Peixoto, Departamento de Estatística".

1. Regressão logística. 2. Teste de Wald. 3. Teste da razão de verossimilhança. 4. Estatística. I. Título.

21. ed. CDD 519.2

Marcia de Lourdes Alves Teotônio

Uso da regressão logística para avaliar o efeito de antissépticos em cirurgia hospitalar

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

Aprovado em: 31 / 10 /2016

Banca Examinadora:

Profa. Dra. Ana Patricia Bastos Peixoto
Orientadora

Prof. Dr. Gustavo Henrique Esteves
Universidade Estadual da Paraíba

Profa. Ms. Maria das Vitórias Alexandre Serafim
Universidade Estadual da Paraíba

Dedicatória

Aos meus pais Antônio Alves Correia Filho e Luzia Pereira Correia (in memoria).

Dedico

Dedicatória

Ao meu esposo, Hélio Teotônio Alves.
Aos meus filhos Marcélio, Hélio Filho e Luzirene

Dedico

Agradecimentos

Primeiramente agradeço a nosso Mestre Maior “DEUS”, pois sem Ele, não teria alcançado meu objetivo nessa fase da minha vida, Ele que sempre me deu forças para não desistir desse meu sonho e me iluminou nas dificuldades que apareciam, fazendo com que não desistisse dessa conquista.

Aos meus pais, Antônio Alves Correia Filho e Luzia Pereira Correia (in memoria), que com o vosso amor e dedicação souberam me criar apesar das dificuldades, sei que se minha mãe estivesse entre nós estaria orgulhosa da minha conquista.

Ao meu esposo, Hélio Teotônio Alves, que sempre me acompanhou nesta trajetória, muitas das vezes me trazia na Universidade e ficava me esperando até o final das aulas, pensei em desistir, mas ele com seu amor e dedicação sempre me incentivou a continuar e graças a ele estou realizando um sonho que sempre almejei.

Aos meus filhos, Marcélio, Hélio Filho e Luzirene que muitas das vezes me ajudaram em algumas disciplinas, me incentivando a não desistir.

Aos amigos, Sônia, Abraão, Damião, Arnete, Claúdio, Vanessa, Deyse, Aline, que sempre estavam prontos a me ajudar nas dificuldades que apareciam, espero que sejam ”amigos para sempre”, que esse nosso vínculo se perpetue por anos e seja infinito.

A minha orientadora, Ana Patrícia, que sempre esteve pronta para tirar minhas dúvidas, prestativa e orientando com paciência e dedicação.

E por fim, agradeço a todos que me ajudaram direta e indiretamente nesta jornada para realizar este trabalho.

Deus, abençoe a todos, muita PAZ.

"Diante de qualquer obstáculo, reflete no bem, porque no curso de todas as circunstâncias, por trás dos contratempos da vida, a Bondade de Deus jaz oculta".

Emmanuel

Resumo

A Regressão Logística é útil para descrever relacionamentos entre uma variável categórica , do tipo binária, e um conjunto de variáveis explicativas, permitindo modelar a probabilidade do sucesso de um evento de interesse. A interpretação e o uso desses modelos depende frequentemente da estimativa pontual para os parâmetros de interesse. Para o modelo ser validado deve-se analisar cuidadosamente através de técnicas adequadas que verifique as pressuposições do modelo. Neste trabalho o pesquisador está interessado em saber o efeito do antisséptico antes e após a utilização do mesmo em pacientes que sobreviveram, após a amputação dos membros superiores e inferiores. Para tanto, utilizou-se da regressão logística para inferir sobre os dados observados. Após a formulação e o ajuste do modelo, foi realizada à análise de diagnóstico. Na análise de diagnóstico, foi utilizado dois tipos de resíduos para avaliar a qualidade do ajuste que são, Os resíduos de Pearson e os de Deviance. Na análise desse experimento utilizou-se o Teste da Verossimilhança e o Teste de Wald, para estimar os parâmetros do modelo e verificar se são estatisticamente significativo.

Palavras-chaves: Envelope simulado, Resíduos, Teste da Razão de Verossimilhança, Teste de Wald.

Abstract

The Logistic Regression is useful to describe relationships between a categorical variable (the binary type) and a set of explanatory variables, allowing to model the probability of success of an event of interest. The interpretation and use of these models often depends on the point estimate for the parameters of interest. For the model to be validated should be analyzed carefully by appropriate techniques to verify the assumptions of the model. The data used in this study the researcher is interested to know the effect of an antiseptic before and after its use in patients who survived after amputation of the upper and lower limbs. After the formulation and model adjustment was performed the analysis of diagnosis. In diagnostic analysis, we used two types of waste to evaluate the quality of fit that are residues of Pearson and the Deviance. In analyzing this experiment we used the test of likelihood and Wald test to estimate the model parameters and check whether they are statistically significant.

Key-words: Simple regression Logistics, Diagnostic Analysis, Likelihood Ratio Test, Wald test.

Lista de Figuras

Lista de Tabelas

Sumário

1 Introdução

A finalidade do antisséptico é impedir a propagação de micro organismos em tecidos vivos com o uso de substâncias químicas usadas como bactericidas ou bacteriostáticos. Tudo o que for utilizado para degradar ou inibir a propagação de micro organismo na superfície da pele ou mucosas é antisséptico. Essas substâncias são utilizadas para desinfectar ferimentos, evitando ou reduzindo o risco de infecção por ação de germes. No Estados Unidos os antissépticos encontram-se sob o controle da Food and Drug Administration (FDA) no que concerne à sua eficácia e no uso clínico, no Brasil é controlado pela Agência Nacional de Vigilância Sanitária (ANVISA).

Com o intuito de modelar fenômenos de inúmeras naturezas, a regressão logística, tem sido uma das principais ferramentas estatísticas empregadas atualmente, sendo largamente utilizada em diversos tipos de problemas. Paula (2002) explica que mesmo quando a resposta não é originalmente binária, alguns pesquisadores têm dicotomizado a variável resposta de modo que a probabilidade de sucesso possa ser modelada por intermédio da regressão logística. Tudo isso deve-se, principalmente, à facilidade de interpretação dos parâmetros de um modelo logístico e também pela possibilidade do uso desse tipo de metodologia em análise com objetivo de discriminação.

De acordo com Moral (2013), diversos testes foram desenvolvidos para verificar as pressuposições de modelos de regressão, bem como para verificar a qualidade do ajuste. Atkinson (1985) propõe a adição de um envelope simulado, de modo que, sob o modelo correto, é provável que os pontos estejam dentro do envelope. O objetivo não é desenvolver uma região de aceitação ou rejeição de observações, mas sim um guia para que tipo de formato seja esperado, caso o modelo esteja correto. Em análises envolvendo modelos lineares generalizados, o gráfico meio-normal com envelope simulado é útil para a verificação da qualidade do ajuste.

Diante dos fatos apresentados, este trabalho apresenta um estudo dos efeitos do uso de antissépicos no tratamento de pacientes que passaram por cirurgia, para verificar a eficácia do mesmo, para tanto, utilizou-se a regressão logística para verificar se o uso de antisséptico tem sido eficaz e para ser possível formar uma estimativa razoável de sua influência sobre a salubridade de um hospital.

2 Fundamentação Teórica

O conteúdo desta seção relata,de modo geral, os aspectos relacionados ao modelos de regressão logística.

2.1 Regressão logística

A função logística surgiu em 1845, ligada a problemas de crescimento demográfico, problemas em que, até os dias de hoje, essa função é utilizada (LIMA, 2002). Na década de 30, essa metodologia passou a ser aplicada no âmbito da biologia, e posteriormente nas áreas relacionadas a problemas econômicos e sociais. Paula (2002), mostra que, apesar do modelo de regressão logística ser conhecido desde os anos 1950, foi devido a trabalhos do estatístico David Cox, na década de 1970, que essa técnica se tornou bastante popular entre os usuários de Estatística.

Esse tipo de regressão é uma ferramenta analítica, baseada nos princípios da regressão múltipla, que tanto trabalha com variáveis métricas quanto com variáveis não métricas ou categóricas e, procura predizer a relação entre uma e mais variáveis conhecidas, procurando explicar determinada situação e sua dependência entre as variáveis. A variável dependente pode ser qualitativa e assumir apenas um entre dois resultados possíveis.

Para Corrar, Paulo e Dias Filho (2007), a regressão logística busca explicar ou valorizar uma variável, a mesma diferencia-se da regressão múltipla em função de outras variáveis conhecidas ou observadas de natureza qualitativa, os autores mostram que a regressão logística tem sua atuação não somente na áreas econômica como também na área médica, como por exemplo os fatores que caracterizam um grupo de indivíduos doentes em relação a indivíduos sãos, onde o principal objetivo da mesma é explicar determinado evento. Esse conjunto de variáveis categóricas existente no estudo podem ser totalmente significativas ou não ter influência nenhuma.

De acordo com Berry, Demerit e Esarey (2010), para que haja confiança na probabilidade do evento é necessário que exista uma interação consistente entre as variáveis categóricas fazendo com que diminua a incidência de resíduos que por ventura possa prejudicar o resultado. Os resíduos podem afetar o resultado por não terem significância

estatística ou simplesmente não afetarem diretamente no evento.

2.2 Modelo de regressão logística simples

A regressão linear simples tem o objetivo de descrever as relações entre a variável resposta (Y) e a variável explicativa (X). Já na regressão logística, a variável resposta (Y) é dicotômica, isto é, atribui-se o valor 1 para o acontecimento de interesse (sucesso) e o valor 0 ao acontecimento complementar (fracasso) com probabilidade $\pi_i = P(Y = 1|X = x_i)$ e $1 - \pi_i = P(Y = 0|X = x_i)$. Para descrever a média condicional de Y dado X com distribuição logística, é utilizada a notação π_i (HOSMER; LEMESHOW, 1985).

Considera-se uma série de eventos binários, em que $(Y_1, Y_2, ..., Y_n)$ são variáveis aleatórias independentes com distribuição Bernoulli, com probabilidade de sucesso (π_i), isto é, $Y \sim Ber(\pi_i)$ e denota-se $\mathbf{x_i^T}$ a i-ésima linha da matriz (X), em que $i = 1, 2, ..., n$.

A probabilidade de sucesso do modelo logístico simples é definida como

$$\pi_i = \pi(x_i) = P(Y = 1|X = x_i) = \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}, \tag{2.1}$$

e a probabilidade de fracasso

$$1 - \pi_i = 1 - \pi(x_i) = P(Y = 0|X = x_i) = \frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}, \tag{2.2}$$

em que, $\beta = (\beta_0, \beta_1)^T$ é o vetor de parâmetros desconhecidos.

Uma diferença importante entre o modelo de regressão logística e o modelo de regressão linear pode ser notada, quando diz respeito a natureza da relação entre a variável resposta e as variáveis independentes. Em qualquer problema de regressão, a quantidade a ser modelada é o valor médio da variável resposta dados os valores das variáveis independentes. Essa quantidade é chamada de média condicional, denotada por $E(Y|X = x_i)$ em que Y é a variável resposta e x_i os valores das variáveis independentes. Na regressão linear tem-se $-\infty < E(Y|X = x_i) < \infty$ e na regressão logística, devido a natureza da variável resposta $0 \leq E(Y|X = x_i) \leq 1$.

Outra diferença importante entre um modelo de regressão linear e o modelo de regressão logístico refere-se a distribuição condicional da variável resposta. No modelo de regressão linear assume-se que uma observação da variável resposta pode ser expressa por $Y_i = E(Y|X = x_i) + \varepsilon_i$, em que ε_i é chamado de erro, com distribuição normal, média zero e variância constante. Isto não ocorre, quando a resposta é dicotômica. O valor da

variável resposta dado x_i, é expresso por $Y_i = \pi_i + \varepsilon_i$ como a quantidade ε_i que pode assumir somente um de dois possíveis valores, isto é, $\varepsilon_i = 1 - \pi_i$ para $Y_i = 1$ ou $\varepsilon_i = -\pi_i$ para $Y_1 = 0$, segue que ε_i tem distribuição com média zero e variância $\pi_i(1 - \pi_i)$ (HOSMER; LEMESHOW, 1985).

A transformação de π_i, é interpretada como o logaritmo da razão das chances entre π_i e $1-\pi_i$. Esta transformação é bastante empregada em estudos toxicológicos, epidemiológicos e de outras áreas, sendo definida como:

Transformação logito da razão de chances

$$g(x_i) = \ln\left[\frac{\pi_i}{1-\pi_i}\right] = \beta_0 + \beta_1 x_i. \tag{2.3}$$

2.2.1 Estimação dos parâmetros do modelo

Supondo que x_i, y_i sejam uma amostra independente com n pares de observações, y_i representa o valor da variável resposta dicotômica e x_i é o valor da variável independente da i-ésima observação em que $i = 1, 2, ..., n$. Para o ajuste do modelo de regressão logística simples, segundo a expressão (2.1), é necessário estimar os parâmetros desconhecidos β_0 e β_1. O método de máxima verossimilhança é utilizado para estimar esses parâmetros. A função de distribuição de Y_i para o modelo de regressão logística simples com $Y_i \sim Ber(\pi_i)$.

Como as observações são independentes, a função de distribuição de probabilidade conjunta de $y_1, y_2, ..., y_n$ será

$$\prod_{i=1}^{n} f(y_i, \pi_i) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}, \quad y_i \in \{0, 1\}.$$

Então a função de verossimilhança é dada por

$$L(\beta) = \prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}, \quad \beta \in \mathbb{R}.$$

O princípio da máxima verossimilhança é estimar o valor de β que maximiza $L(\beta)$.

Aplicando o logaritmo, a expressão é definida como

$$\begin{aligned} l(\beta) = \ln[L(\beta)] &= \ln\left[\prod_{i=1}^{n} \pi_i^{y_i} (1-\pi_i)^{1-y_i}\right] \\ &= \sum_{i=1}^{n} [y_i \ln(\pi_i) + (1-y_i)\ln(1-\pi_i)] \\ &= \sum_{i=1}^{n} [y_i \ln(\pi_i) + \ln(1-\pi_i) - y_i \ln(1-\pi_i)] \\ &= \sum_{i=1}^{n} \left[y_i \ln\left(\tfrac{\pi_i}{1-\pi_i}\right) + \ln(1-\pi_i)\right]. \end{aligned}$$

Substituindo as equações (2.2) e (2.3), têm-se

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

Para encontrar o valor de β que maximiza $l(\beta)$, deriva-se $l(\beta)$ em a relação cada parâmetro (β_0, β_1), obtendo-se duas equações

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

que igualadas a zero, geram o seguinte sistema de equações

$$\sum_{i=1}^{n} (y_i - \pi_i) = 0, \tag{2.4}$$

$$\sum_{i=1}^{n} x_i (y_i - \pi_i) = 0, \tag{2.5}$$

em que $\pi_i = \frac{\exp(\beta_0+\beta_1 x_i)}{1+\exp(\beta_0+\beta_1 x_i)}$

Como as equações (2.4) e (2.5) não são lineares em β_0 e β_1, são necessários métodos iterativos para resolução, estes disponíveis em vários programas computacionais.

2.2.2 Interpretação dos parâmetros

Segundo Agresti (2002) na regressão logística a interpretação dos parâmetros é obtida quando a variável independente é dicotômica, ou seja, compara-se a probabilidade do evento ocorrer com a probabilidade do evento não ocorrer. Defini-se a chance do evento ocorrer como a probabilidade do evento ocorrer dividido pela probabilidade do evento não

ocorrer, conhecida como razão de chances (Odds Ratio). O logaritmo da razão de chances é dada por

$$g(x_i) = \ln\left[\frac{\pi_i}{1-\pi_i}\right].$$

Se a variável X for codificada como 0 ou 1, então a chance da resposta quando $x = 1$ é definida como $\pi(1)/[1-\pi(1)]$ e quando $x = 0$ é definida como $\pi(0)/[1-\pi(0)]$

Então a razão de chances, pode ser denotada por OR, é definida por

$$\begin{aligned} OR = \frac{\pi(1)/[1-\pi(1)]}{\pi(0)/[1-\pi(0)]} &= \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{(\beta_0+\beta_1)}}\right)\Big/\left(\frac{1}{1+e^{(\beta_0+\beta_1)}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{(\beta_0)}}\right)\Big/\left(\frac{1}{1+e^{(\beta_0)}}\right)} \\ &= \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} \\ &= e^{\beta_1}. \end{aligned}.$$

Logo, o logaritmo da razão de chances é dado por

$$\ln(OR) = \ln\left[e^{\beta_1}\right] = \beta_1.$$

A razão de chances é um parâmetro de grande interesse no modelo de regressão logística devido a sua fácil interpretação. A distribuição assimétrica de β_1 é devida ao fato do seu limite tender a zero (PAULA, 2004). As inferências são frequentemente baseadas na distribuição do $\ln(OR) = \beta_1$, o qual tende a seguir uma distribuição normal, mesmo para pequenas amostras. Assim sendo, a razão de chances é definida como a chance de ocorrência de um evento entre indivíduos que tem um fator de risco, comparados com indivíduos não expostos, sujeitos ao evento.

2.2.3 Teste de Wald

O teste de Wald é uma técnica estatística utilizada para verificar se o parâmetro é estatisticamente significativo. Segundo Agresti (2002), o teste de Wald é uma das inúmeras maneiras de testar os parâmetros associados com um grupo de variáveis explicativas. O teste de Wald é obtido por comparação entre a estimativa de máxima verossimilhança do parâmetro $(\hat{\beta}_1)$ e a estimativa de seu erro padrão, sob a hipótese $H_0 : \beta_1 = 0$, assim, a estatística do teste de Wald para a regressão logística é

$$W_j = \frac{\hat{\beta}_1}{DP(\hat{\beta}_1)}. \tag{2.6}$$

Esse teste segue uma distribuição assintoticamente normal, portanto quando o valor-p

é menor que o nível de significância, conclui-se que os parâmetros β_1 são significativos no modelo.

2.2.4 Teste da razão de verossimilhança

O teste da razão de verossimilhança é obtido por meio da comparação de valores observados da variável resposta com os valores preditos obtidos dos modelos com e sem a variável em questão. A comparação dos valores observados com os valores preditos é baseado no logaritmo da verossimilhança. Para entender melhor essa comparação, é útil pensar em um valor observado da variável resposta também como sendo um valor predito resultante de um modelo saturado. Um modelo saturado é aquele que contém tantos parâmetros quanto observações. A função de verossimilhança é baseada na seguinte expressão

$$D = -2\ln\left[\frac{\text{verossimilhança do modelo ajustado}}{\text{verossimilhança do modelo saturado}}\right]. \tag{2.7}$$

Com o propósito de assegurar a significância de uma variável independente, compara-se o valor da estatística D com e sem a variável na equação. A mudança em D devido a inclusão da variável no modelo é obtida da seguinte maneira:

$$G = D(\text{modelo sem a variavel}) - D(\text{modelo com a variavel})$$

Pode-se então escrever a estatística G como sendo:

$$D = -2\ln\left(L_s\right) + 2\ln\left(L_c\right),$$

em que, L_s é a verossimilhança do modelo sem a covariável e L_c e a verossimilhança do modelo com a covariável, onde irá testar $H_0 : \beta_i = 0$ vs $H_0 : \beta_i \neq 0$. Sob hipótese nula, a estatística G tem distribuição qui-quadrado com 1 grau de liberdade

2.2.5 Intervalo de confiança para os parâmetros

As estimativas do intervalo de confiança para os parâmetros é a mesma teoria estatística que usamos para os testes de significância do modelo. Em particular, um intervalo de confiança para a inclinação e intercepto são baseados em seus respectivos testes de Wald. O intervalo de confiança de β_i é dado por

$$IC(\beta_1;(1-\alpha) = [\beta_1 - Z_{1-\alpha/2}DP(\beta_1); \beta_1 + Z_{1-\alpha/2}DP(\beta_1)],$$

já para o intercepto β_0, tem-se:

$$IC(\beta_0;(1-\alpha) = [\widehat{\beta}_0 - Z_{1-\alpha/2}DP(\widehat{\beta}_0); \widehat{\beta}_0 + Z_{1-\alpha/2}DP(\widehat{\beta}_0)],$$

em que, $Z_{1-\alpha/2}$, é o ponto normal padrão correspondente ao quantil de $100(1-\alpha)$ de probabilidade.

2.2.6 Intervalo de confiança para logito

O modelo de regressão logística tem como parte linear a função logito, tem-se como estimador

$$\hat{g}(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x.$$

O estimador da logito requer a obtenção da variância da soma através do estimador da variância, isto é

$$v\hat{a}r[\hat{g}(x)] = v\hat{a}r(\widehat{\beta}_0) + x^2 v\hat{a}r(\widehat{\beta}_1) + 2\ x\ c\hat{o}v(\widehat{\beta}_0, \widehat{\beta}_1). \tag{2.8}$$

A logito tem como intervalo de confiança

$$IC(g(x), 1-\alpha) = [\hat{g}(x) - Z_{1-\alpha/2}DP(\hat{g}(x));\ \hat{g}(x) + Z_{1-\alpha/2}DP(\hat{g}(x))],$$

em que $DP(\hat{g}(x))$, é a raiz quadrada da expressão (2.8) e $Z_{1-\alpha/2}$ é o ponto da normal padrão, de $100(1-\alpha)$ % de confiança.

2.2.7 Intervalo de confiança para os valores ajustados

O estimador dos valores ajustados fornece o estimador do logito e seu intervalo de confiança. Logo, o intervalo de confiança dos valores ajustados é dado por

$$IC(\pi, 1-\alpha) = \Big[\frac{e^{\hat{g}(x) - Z_{1-\alpha/2}DP\hat{g}(x)}}{1 + e^{\hat{g}(x) - Z_{1-\alpha/2}DP\hat{g}(x)}}; \frac{e^{\hat{g}(x) - Z_{1-\alpha/2}DP\hat{g}(x)}}{1 - e^{\hat{g}(x) - Z_{1-\alpha/2}DP\hat{g}(x)}}\Big]$$

2.2.8 Intervalo de confiança para a Odds Ratio

Sabendo-se que os limites do intervalo de confiança para β_1 é $\beta_I = \widehat{\beta}_1 - Z_{1-\alpha/2}DP(\hat{\beta}_1)$ e $\beta_S = \widehat{\beta}_1 + Z_{1-\alpha/2}DP(\hat{\beta}_1)$ Odds Ratio tem seu intervalo de confiança desta forma

$$IC(\text{Odds Ratio}; 1-\alpha) = [e^{\beta_I}; e^{\beta_S}]$$

2.3 Bondade do ajuste

A ideia de se verificar a bondade de ajuste refere-se ao teste estatístico aplicado na obtenção do modelo, visando-se aferir se este é o mais indicado.

2.3.1 Estatística de *Pearson*

As medidas da qualidade do ajuste são funções dos resíduos definidos como a diferença entre o valor observado e os valores ajustados $(y - \hat{y})$ na regressão linear, enquanto que na regressão logística existem diferentes formas de calcular essa diferença. Pode-se ver que os valores ajustados na regressão logística são calculados para cada covariável e depende das estimativas de probabilidade para cada covariável, vamos denotar os valores ajustados para a j-ésima covariável como $\hat{y}_j$ em que:

$$\hat{y}_j = m_j\hat{\pi}_j,$$

em que m_j é o número de observações na covariável j e $\hat{\pi}_j$ é a probabilidade ajustada dos indivíduos em j. Sendo assim a medida de Pearson para a diferença entre o observado e predito é:

$$r(y_j, \hat{\pi}_j) = \frac{(y_j - m_j\hat{\pi}_j)}{\sqrt{m_j\hat{\pi}_j(1-\hat{\pi}_j)}},$$

em que y_j o número de indivíduos em j com $y = 1$. Logo, a estatística qui-quadrado de Pearson é dada por:

$$\chi^2 = \sum_{j=1}^{J} r(y_j, \hat{\pi}_j)^2.$$

Portanto, pode-se dizer que χ^2 se aproxima, assintoticamente, $\chi^2_{J-(p+1)}$,p o número de covariáveis do modelo ajustado e J é o número de covariáveis. Porém, em alguns casos essa aproximação é ruim.

2.3.2 Estatística de *Deviance*

De acordo com Souza (2006), existem muitas estatísticas para medir esta discrepância, das quais a mais utilizada está baseada na função de verossimilhança, proposta por Nelder e Wedderburn (1972), com o nome de Deviance. Os autores comparam o valor da função de verossimilhança para o modelo proposto com $p+1$ parâmetros ($L(\hat{\beta}_0, ..., \hat{\beta}_p)$) ao seu valor no modelo saturado ($L(y_1, ..., y_n)$). Para esta comparação é conveniente utilizar menos duas vezes o logaritmo do quociente destes máximos. Assim, a Deviance é definida como:

$$D = -2 \ln \left[\frac{L\left(\hat{\beta}_0, \ldots, \hat{\beta}_p\right)}{L\left(y_1, \ldots, y_n\right)} \right],$$

equação na qual verifica-se a utilização de um teste de razão de verossimilhança generalizado. No modelo de regressão logística, considerando o modelo com as proporções estimadas $\hat{\pi}_i$, temos que a *deviance* pode ser escrita como:

$$D = 2 \sum_{i=1}^{n} \left[y_1 \ln \left(\frac{y_i}{\hat{\pi}_i} \right) + (1 - y_i) \ln \left(\frac{1 - y_i}{1 - \hat{\pi}_i} \right) \right].$$

A estatística baseada no resíduo da deviance é dada por

$$D = \sum_{j=1}^{J} d(y_j, \hat{\pi}_j)^2.$$

A distribuição assintótica da deviance é dada por Collet (1991) é uma $\chi^2_{(n-p)}$, em que p é o número de parâmetros do modelo.

2.4 Diagnóstico do modelo

Na regressão seja ela simples, múltipla, logística, é preciso proceder com a análise dos resíduos para a validação da qualidade do modelo estimado, só assim pode-se avaliar as "distâncias"entre os valores observados e os valores estimados. Há várias medidas de modo a detectar se existem diferenças significativas entre os valores observados e os valores estimados. A análise de resíduos e diagnóstico é utilizada para detectar problemas, tais como:

i) presença de observações discrepantes (pontos aberrantes);

ii) inadequação das pressuposições para os erros aleatórios ou para as médias;

iii) colinearidade entre as colunas da matriz do modelo;

iv) forma funcional do modelo inadequada;

v) presença de observações influentes.

2.4.1 Gráficos meio-normais com envelope simulados

De acordo com Moral (2013) em análises envolvendo modelos lineares generalizados, o gráfico meio-normal com envelope simulado é bastante útil para a verificação da qualidade do ajuste, principalmente quando se trata de análises utilizando modelos que incorporam superdispersão. Essa técnica, relativamente simples, para verificação da adequação do ajuste de um determinado modelo a um conjunto de observações, proposta por Atkinson (1985), consiste em se plotar os valores absolutos ordenados de uma determinada medida de diagnóstico (diferentes tipos de resíduos, distância de Cook, medida de alavanca, etc.) versus as estatísticas de ordem esperadas da distribuição meio-normal, calculadas por

$$\Phi^{-1}\left(\frac{i+n-\frac{1}{8}}{2n+\frac{1}{2}}\right), \qquad i=1,2,...,n,$$

sendo n o tamanho da amostra e adicionando-se a esse gráfico um envelope simulado. Geralmente, utilizam-se os resíduos de Pearson ou componentes da deviance escalonados como medida de diagnóstico para a verificação da qualidade de ajuste de um modelo linear generalizado.

Para obtenção do envelope simulado, segue-se os seguintes passos (MORAL, 2013)

i) Obter $d_{(i)}$, os valores de uma quantidade diagnóstica em valor absoluto e em ordem crescente;

ii) Simular 99 amostras do modelo ajustado com o mesmo valor para as varíaveis explanatórias;

iii) Fazer o ajuste do modelo para as 99 amostras e, para cada ajuste, obter quantidade diagnóstica de interesse, $d^*_{j(i)}$, $j=1,...,99$ e $i=1,...,n$ em valor absoluto e em ordem crescente;

iv) Para cada i, computar os percentis 5%, 50% e 95%;

v) Fazer o gráfico desses percentis e dos d, observados, contra as estatísticas de ordem da distribuição meio-normal.

3 Aplicação

Os dados utilizados na elaboração dos resultados da análise estatística foram obtidos no endereço *http* : *//www.stat.ufl.edu/ winner/datasets.html*, através do arquivo *http* : *//www.stat.ufl.edu/ winner/data/lister.dat*, maiores informações sobre a coleta dos dados encontram-se em Lister (1870), no seu artigo *Effects of the Antiseptic System of Treatment Upon the Salubrity of a Surgical Hospital*, no artigo o autor verificou a sobrevivência dos pacientes que tiveram os membros superiores e inferiores amputados antes e após a descoberta e utilização de antisséptico. A pesquisa foi realizada entre os anos de 1864 e 1869 com um total de 75 pacientes, cujas variáveis estudadas foram: Antisséptico (1= uso, 0=não uso), Membros (1=inferiores, 2=superiores), Diagnóstico (1=recuperação, 0=morte).

O ajuste dos modelos foi feito por meio do software gratuito R (R CORE TEAM, 2013), para gerar os gráficos meio-normais com envelope simulado para o modelo, foi utilizada a função `hnp()`, que gera objetos com a classe `hnp`. Os resultados obtidos a partir do uso do modelo logístico informam que por meio do modelo de regressão logística, avaliou-se os efeitos causados no tratamento durante o uso do antisséptico. Pelo que se pode observar na Figura 1 que houve efeito do uso se antisséptico no tratamentos dos pacientes.

O ajuste do modelo logístico aos dados mostrou que há evidências de efeito do antisséptico e que o intercepto não foi significativo ao nível de 5% de significância, ou seja, os pacientes que submeteram-se ao tratamento após terem seus membros amputados melhoraram, isto é, o uso do antisséptico surtiu efeito (Tabela 1).

Tabela 1: Análise de variância para o resultado em função do uso de antisséptico no tratamento de pacientes que tiveram seus membros amputados.

	Estimativas	Erro Padrão	Valor t	Pr(>|t|)
Intercepto	0,1719	0,3393	0,506	0,6125
Antisséptico	1,5628	0,5578	2,802	$0,0051^{**}$

Como observa-se Tabela 2, na análise de desvios para o uso de antisséptico o resultado foi significativo $\Pr(> \chi^2)$=0,0032, logo pode-se dizer que o efeito do antisséptico foi eficaz para os pacientes que foram submetidos ao tratamento. Há uma forte evidência de dife-

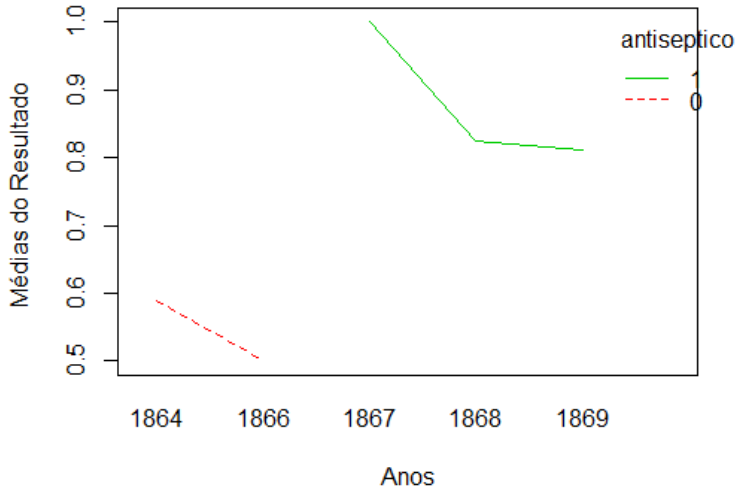

Figura 1: Interação entre variáveis anos e antisséptico para o resultado do uso de antissépticos

rença nos antissépticos usados antes e depois de ter os membros amputados (a estatística qui-quadrado tem um valor resultante dado por $\chi^2 = 8,69$ em 1 grau de liberdade, sendo o p-valor aproximadamente nulo). O desvio nulo do modelo foi de 90,77 com 74 graus de liberdade e o desvio residual foi de 82,08 com 73 graus de liberdade.

Tabela 2: Análise de desvios para o resultado em função do uso de antissépticos no tratamento de pacientes que tiveram seus membros amputados.

	Df	Deviance	Resid. Df	Resid. Dev	Pr(>Chi)
Modelo Nulo			74	90,77	
antisséptico	1	8,69	73	82,08	$0,0032^{**}$

A razão de chances entre do uso de antisséptico é 4,77, ou seja, o paciente que usar o antisséptico tem 4 vezes a chance de recuperar-se e não ser acometido a morte (Tabela 3).

Tabela 3: Intervalo de confiança para a Razão de Chances para os parâmetros do modelo de regressão logística.

	OR	2,5 %	97,5 %
(Intercepto)	1,19	0,61	2,34
antisséptico(1)	4,77	1,67	15,24

Na Figura 2 apresenta-se o gráfico do envelope simulado e não observa-se indícios de que a distribuição utilizada seja inadequada. O envelope é bastante útil para verificação

da qualidade do ajuste, logo pode-se observar que o modelo é adequado.

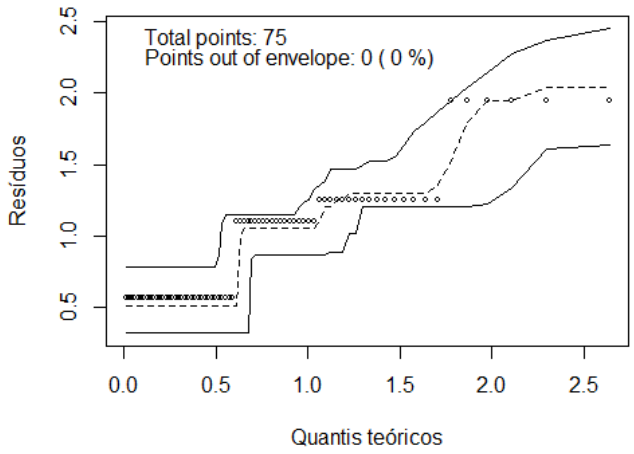

Figura 2: Gráficos meio-normais com envelope simulado para os modelos binomiais

Na Figura (3), estão os gráficos dos resíduos do modelo. Percebe-se que o modelo, também por esta análise, está bem ajustado, pela proximidade dos pontos ordenados a reta no gráfico do quantil do normal padrão e os resíduos *Deviance* ordenados. No gráfico da distância de Cook, não há presença de ponto discrepante.

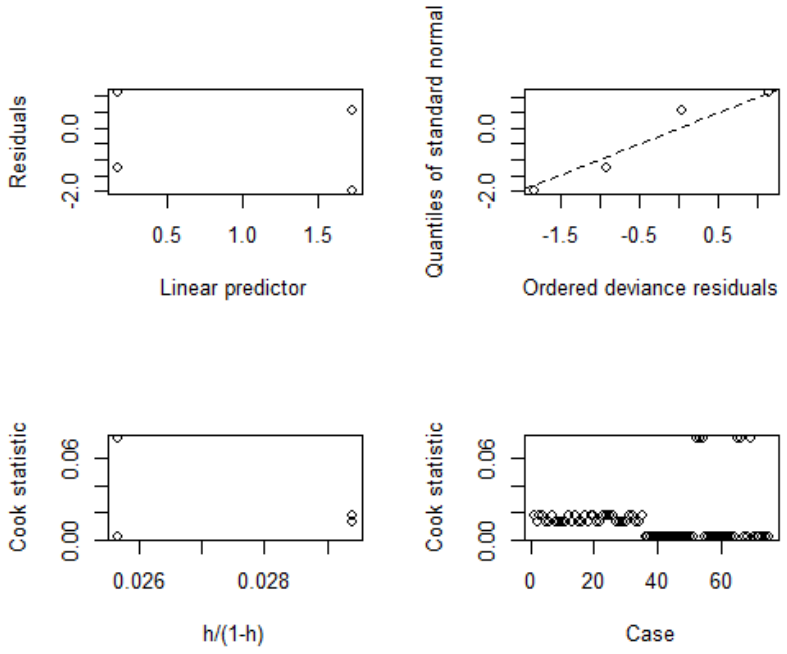

Figura 3: Gráficos de diagnóstico dos dados ajustados ao modelo binomial para o resultado do uso de antissépticos

4 Conclusão

Através do modelo de regressão logística, foi possível avaliar os efeitos causados no tratamento durante o uso do antisséptico, observou-se que após o ajuste do modelo, o antisséptico foi o que melhor respondeu ao tratamento. Foi utilizada as técnicas de diagnósticos e análises de resíduos, onde chegou-se a conclusão por meio do gráfico de envelope que o modelo ajustado foi adequado. Logo, a partir das análises e do modelo ajustado pode-se afirmar que houve efeito do uso dos antissépticos no tratamento dos pacientes.

5 Referências Bibliográficas

AGRESTI, A. **Categorical data analysis**. Hoboken: John Wiley & Sons, 2002. 710 p.

ATKINSON, A. C. **Plots, transformations, and regression: An introduction to graphical methods of diagnostic regression analysis.** Oxford: Clarendon, 1985. 282 p.

BERRY, WILLYAM D.; DEMERITT, JACQUELINE H.R.; ESAREY, JUSTIN. Testing for interaction in binary logit and probit models: is a product term essential. **American Journal of Political Science**. v. 54, n.1, p. 248-266, 2010.

COLLETT, D. **Modelling Binary Data**. Chapman and Hall, London, 1991.

CORRAR, L.J; PAULO, E; DIAS FILHO, J. M. A**nálise Multivariada: para cursos de administração, ciências contábeis e economia**. São Paulo, Atlas/FIPECAFI, 2007

HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. John Wiley, New York, 1989.

LIMA, J. **A análise econômico-financeira de empresas sob a ótica da estatística multivariada**. 2002. 192 f. Dissertação (Mestrado em Métodos Numéricos em Engenharia), Curso de Pós-graduação em Tecnologia e Ciências Exatas, Universidade Federal do Paraná, Curitiba, 2002.

LISTER. J. Effects of the antiseptic system of treatment upon the salubrity of a surgical hospital. **Lancet** 1:4-6 and 40-42, 1870.

MORAL, R. A. **Modelagem estatística e ecológica de relações tróficas em pragas e inimigos naturais**. Dissertação (Mestrado) - Escola Superior de Agricultura ?Luiz de Queiroz?, Piracicaba, 2013, 173 p.

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

PAULA, G. A. **Modelos de Regressão com apoio computacional**. São Paulo: IME/USP. 2002.

Paula, G. A. **Modelos de regressão: Com apoio computacional.** Instituto de matemática e estatística - Universidade de São Paulo. ed. IME-USP.2004.

R CORE TEAM. R: A language an environment for statistical computing. R foundation for statistical computing, Vienna, Áustria, disponível em: $< http : //www.R - projeto.org >$, acesso em:10 jul. 2016.

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