Fernanda Nanci Scacabarozi

Modelagem de eventos raros: um estudo comparativo.

São Carlos, janeiro de 2012

Universidade Federal de São Carlos
Centro de Ciências Exatas e de Tecnologia
Departamento de Estatística

Fernanda Nanci Scacabarozi

Modelagem de eventos raros: um estudo comparativo

Dissertação apresentada ao Departamento de Estatística da Universidade Federal de São Carlos-Des-UFSCar, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Orientador: Carlos Alberto Ribeiro Diniz

São Carlos, janeiro de 2012

Ficha catalográfica elaborada pelo DePT da Biblioteca Comunitária da UFSCar

S277me

Scacabarozi, Fernanda Nanci.
Modelagem de eventos raros : um estudo comparativo / Fernanda Nanci Scacabarozi. -- São Carlos : UFSCar, 2012.
116 f.

Dissertação (Mestrado) -- Universidade Federal de São Carlos, 2012.

1. Probabilidades. 2. Modelo logito. 3. Modelo logito limitado. 4. Modelo logito generalizado. 5. Modelo logito com resposta de origem. 6. Estimadores KZ. I. Título.

CDD: 519.2 (20ª)

Fernanda Nanci Scacabarozi

MODELAGEM DE EVENTOS RAROS

Dissertação apresentada à Universidade Federal de São Carlos, como parte dos requisitos para obtenção do título de Mestre em Estatística.

Aprovada em 16 de janeiro de 2012.

BANCA EXAMINADORA

Presidente _____
Prof. Dr. Carlos Alberto Ribeiro Diniz (DEs-UFSCar/Orientador)

1º Examinador _____
Prof. Dr. Francisco Louzada Neto (ICMC-USP)

2º Examinador _____
Prof. Dr. Jorge Luis Bazán Guzmán (PUC-Peru)

Resumo

Em algumas situações, nas mais diversas áreas do conhecimento, a variável resposta de interesse possui distribuição dicotômica extremamente desbalanceada. No mercado financeiro é comum o interesse em determinar a probabilidade de que cada cliente venha a cometer uma ação fraudulenta, sendo que a proporção de clientes fraudadores é extremamente pequena. Na área da saúde existe o interesse em determinar a probabilidade de que uma determinada pessoa venha a apresentar alguma infecção epidemiológica que atinge apenas uma diminuta parcela da população. No entanto, existem estudos que revelam que o modelo de regressão logística usual, amplamente utilizado na modelagem de dados binários, não produz bons resultados quando este é construído utilizando bases de dados extremamente desbalanceadas. Na literatura, encontramos algumas propostas para o ajuste de modelos que levam em conta esta característica, tal como os estimadores KZ sugeridos por King e Zeng (2001) para o modelo de regressão logística aplicado em bases de dados com eventos raros. Neste trabalho apresentamos esta metodologia e um estudo de simulação para verificar a qualidade destes estimadores. Outras propostas encontradas na literatura são o modelo logito limitado sugerido por Cramer (2004) que limita superiormente a probabilidade de sucesso e o modelo logito generalizado sugerido por Stukel (1988) que apresenta dois parâmetros de forma e funciona melhor que o modelo logito usual nas situações em que a curva de probabilidade não é simétrica em torno do ponto $\frac{1}{2}$. Neste trabalho apresentamos algumas simulações para verificar as vantagens do usos destes modelos.

Palavras-chave: modelo logito, modelo logito limitado, modelo logito generaizado, modelo logito com resposta de origem, estimadores KZ, medidas preditivas.

Abstract

In some situations, in various areas of knowledge, the response variable of interest has dichotomous distribution extremely unbalanced. In the financial market is the common interest in determining the probability that each customer will commit a fraudulent action, and the proportion of customers fraudsters is extremely small. In health there is interest in determining the probability that a particular person will present some epidemiological infection that affects only a small fraction of the population. However, there are studies that show that the usual logistic regression model, widely used in the modeling of binary data, does not produce good results when it is built using databases extremely unbalanced. In the literature, we find some proposals for adjusting models them that take into account this characteristic, such as KZ estimators suggested by King and Zeng (2001) for the logistic regression model applied to databases with events rare. We present this methodology and a simulation study to verify the quality of these estimators. Other proposals in the literature are limited logit model suggested by Cramer (2004) that upper limit to the probability of success and the generalized logit model suggested by Stukel (1988) which has two shape parameters and works better than the usual logit model in situations that the probability curve is not symmetrical around the point $\frac{1}{2}$. In this paper we present some simulations to verify the advantages of the use of these models.

Palavras-chave: model logit model limited, generalized logit model, logit model with response of origin, KZ estimators, measures forecasts.

Sumário

Lista de Figuras

Lista de Tabelas

Capítulo 1

Introdução

Em muitas situações práticas é comum o interesse em descrever o relacionamento entre uma variável resposta e uma ou mais variáveis explicativas. Quando a variável resposta é contínua o método dos mínimos quadrados usual é utilizado na estimação dos parâmetros. No entanto, muitas vezes a variável resposta de interesse possui distribuição discreta dicotômica e nestes casos os estimadores de mínimos quadrados não possuem as propriedades usuais. Alguns modelos para a análise de dados binários foram discutidos por Cox (1970), destacando o modelo de regressão logística pelas suas características. Uma delas é a flexibilidade, do ponto de vista matemático, da função logística. Outra propriedade importante do modelo é a interpretação dos seus coeficientes que é muito útil, principalmente em problemas biológicos.

Assim, a regressão logística é um modelo probabilístico de regressão não linear que se encaixa nas situações em que as variáveis resposta são discretas e os erros não são normalmente distribuídos.

No entanto, existem diversas situações em que a variável resposta de interesse possui distribuição dicotômica extremamente desbalanceada. No mercado financeiro

é comum o interesse em determinar a probabilidade de que cada cliente venha a cometer uma ação fraudulenta, sendo que a proporção de clientes fraudadores é extremamente pequena. Na área da saúde existe o interesse em determinar a probabilidade de que uma determinada pessoa venha a apresentar alguma infecção epidemiológica que atinge apenas uma diminuta parcela da população. Existem alguns estudos que revelam que o modelo de regressão logística usual subestima a probabilidade do evento de interesse quando este é construído utilizando bases de dados extremamente desbalanceadas (King e Zeng, 2001). Além disso, os estimadores de máxima verossimilhança dos parâmetros do modelo de regressão logística são viciados nestes casos.

Uma técnica de amostragem muito utilizada na construção das bases de dados para o ajuste modelo logito usual na situação de desbalanceamento são as amostras *state-dependent*. Neste trabalho apresentamos uma breve discussão sobre este procedimento e o Método de Correção a Priori que permite manter as propriedades dos estimadores de máxima verossimilhança quando o modelo logito é ajustado em amostras construídas por meio desta técnica. Além disso, realizamos um estudo de simulação para verificar o poder preditivo dos modelos ajustados considerando as amostras *state-dependent*.

Sabemos que os parâmetros do modelo de regressão logística são assintoticamente não viciados, no entanto, de acordo com King e Zeng(2001), este vício persiste mesmo quando as amostras são grandes. McCullagh e Nelder (1989) sugeriram um estimador para o vício de qualquer modelo linear generalizado que foi adaptado por King e Zeng (2001) para o uso concomitante com as amostras *state-dependent* para o modelo de regressão logística. Neste trabalho apresentamos esta correção efetuada nos estimadores de máxima verossimilhança bem como algumas simulações para verificar o impacto causado no vício destas estimativas.

Apresentamos também algumas correções realizadas na probabilidade de sucesso estimada por meio do modelo de regressão logística que foram sugeridas por King e Zeng (2001). Tais correções permitem diminuir o vício e o erro quadrático médio destas probabilidades estimadas. Para verificar a eficiência desta metodologia foram realizadas algumas simulações para averiguar a vantagem das mesmas no poder preditivo do modelo de regressão logística. Os estimadores dos parâmetros e também da probabilidade de evento do modelo logito que levam em conta as correções sugeridas por King e Zeng (2001) são chamados de estimadores KZ.

Existem modelos encontrados na literatura desenvolvidos especialmente para a situação de dados binários desbalanceados. Um deles é o modelo logito generalizado sugerido por Stukel(1988). Este modelo possui dois parâmetros de forma e funciona melhor do que o modelo logito usual em situações em que a curva de probabilidade esperada é assimétrica. Outro modelo encontrado é o logito limitado sugerido por Cramer (2004). Este permite que seja estabelecido um limite superior para a probabilidade de sucesso. Neste trabalho apresentamos uma breve discussão sobre as características destes modelos assim como um estudo de simulação para verificar as qualidades dos mesmos quando comparados ao modelo logito usual.

Muitas vezes, a variável resposta é originalmente fruto de uma distribuição discreta ou contínua, ou seja, ela tem uma distribuição original que não a de Bernoulli e, por alguma razão esta variável foi dicotomizada através de um ponto de corte C arbitrário. O modelo de regressão logística pode agregar a informação sobre a distribuição da variável de origem no ajuste do modelo logito usual. Dessa forma, o modelo pode ter a variável resposta, por exemplo, pertencente a família exponencial no contexto dos modelos lineares generalizados com função de ligação composta. Esta metodologia foi apresentada por Suissa e Blais (1995) considerando dados reais de estudos clínicos e também dados simulados com distribuição original log-normal, Paula

e Diniz(2011) estenderam os resultados para algumas distribuições de origem pertencentes a família exponencial. Dependendo do ponto de corte utilizado a variável resposta pode apresentar um desbalanceamento muito acentuado. Neste trabalho apresentamos o desenvolvimento de modelos de regressão logística com resposta de origem normal, exponencial e log-normal e realizamos um estudo comparativo utilizando dados simulados para verificar as vantagens do uso do modelo logito que agrega a resposta de origem comparado ao modelo logito usual na situação de desbalanceamento extremo.

O trabalho está organizado como segue:

No Capítulo 2 apresentamos o modelo logito usual (Hosmer e Lemeshow(1989)) e a metodologia proposta por King e Zeng(2001) para a obtenção de estimadores para os parâmetros deste modelo na situação de eventos raros. Neste capítulo encontramos também a técnica de amostragem *state-dependent*, muito utilizado no mercado financeiro na construção de amostras para o ajuste do modelo de regressão logística na situação de desbalanceamento.

No Capítulo 3 encontramos as principais características do modelo logito limitado proposto por Cramer(2004).

No Capítulo 4 apresentamos o modelo logito generalizado proposto por Stukel (1988) e as técnicas utilizadas na estimação dos parâmetros do mesmo.

No Capítulo 5 apresentamos o modelo logito com resposta de origem considerando três ditribuições da variável resposta: normal, exponencial e lognormal.

No Capítulo 6 é destinado a exposição das técnicas utilizadas na comparação dos modelos estudados. Neste Capítulo apresentamos uma breve discussão sobre as medidas preditivas e de qualidade de ajuste utilizadas neste trabalho.

No Capítulo 7 encontramos um estudo de simulações realizado com o intuito de comparar a capacidade preditiva e a qualidade de ajuste dos modelos de classificação

estudados na situação de eventos raros.

No Capítulo 8 apresentamos algumas simulações efetuadas com o intuito de comparar o desempenho do modelo logito usual com o modelo logito com resposta de origem na qualidade das estimativas produzidas destes modelos.

No Capítulo 9 encontramos uma análise de dados reais de fraude bancária em que ajustamos os modelos de classificação estudados e efetuamos comparações através das medidas preditivas e de qualidade de ajuste.

No Capítulo 10 é destinado a exposição das conclusões finais.

Capítulo 2

O Modelo Logito Usual

2.1 Introdução

A Regressão Logística é um modelo probabilístico de regressão não linear que se encaixa em situações em que a variável resposta é discreta e os erros não são normalmente distribuídos. A variável resposta Y apresenta dois possíveis resultados (sucesso e fracasso), geralmente o sucesso é alguma característica de interesse.

Em problemas de regressão, a quantidade de interesse é a esperança condicional da variável resposta Y dado uma covariável X, $E(Y|X = x)$. Neste caso, assumimos que esta média pode ser expressa como uma equação linear em x, tal como:

$$E(Y|x) = \beta_0 + \beta_1 \times x. \tag{2.1}$$

Considerando que o valor da covariável varia entre $-\infty$ e ∞, o valor desta média também varia entre $-\infty$ e ∞.

Quando a variável resposta segue distribuição de Bernoulli, sua média condicional deve estar entre 0 e 1, ou seja, $0 \leq E(Y|X = x) \leq 1$, aproximando-se de 0 e de 1

gradualmente em forma de S.

Considerando uma amostra de n elementos de uma variável resposta com distribuição de Bernoulli e um conjunto de $p-1$ covariáveis associadas a esta variável, sendo y_1, y_2, $\cdots$, y_n os valores observados da variável resposta para todos os indivíduos da amostra e x_{i1}, x_{i2}, $\cdots$, x_{ip-1} as covariáveis observadas para o i-ésimo indivíduo, de acordo com o modelo de regressão logística, a probabilidade de sucesso para o i-ésimo indivíduo é dada por:

$$\pi\left(\boldsymbol{x_i}\right) = \frac{e^{\eta_i}}{1+e^{\eta_i}}, \tag{2.2}$$

com $\eta_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_{p-1} x_{ip-1}$. Aplicando a transformação logito em $\pi\left(\boldsymbol{x}_i\right)$ temos o preditor linear $\eta_i = \boldsymbol{x_i'\beta}$ e $\boldsymbol{x_i} = (x_{i1}, x_{i2}, \cdots, x_{ip-1})$ e $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})$:

$$logito\left(\pi\left(\boldsymbol{x}_i\right)\right) = log\left(\frac{\pi\left(\boldsymbol{x}_i\right)}{1+\pi\left(\boldsymbol{x}_i\right)}\right) = \boldsymbol{x_i'\beta}. \tag{2.3}$$

2.2 Estimação

O método mais comum de estimação dos parâmetros do modelo logito usual é o método de máxima verossimilhança. Considerando um conjunto de variáveis aleatórias $Y_1, Y_2, \cdots, Y_n$ com distribuição de Bernoulli e $\boldsymbol{X_i}$ um vetor contendo $p-1$ covariáveis, de forma que $Y|X_i$ siga distribuição de Bernoulli com probabilidade de sucesso $\pi\left(\boldsymbol{x_i}\right)$. Assim, a distribuição de probabilidade de $Y_i|\boldsymbol{X_i}$ pode ser escrita como:

$$P\left(Y_i = y_i|\boldsymbol{x_i}\right) = \left(\pi\left(\boldsymbol{x_i}\right)\right)^{y_i}\left(1-\pi\left(\boldsymbol{x_i}\right)\right)^{1-y_i}, \tag{2.4}$$

com $y_i = 0, 1$ e $i = 1, \cdots, n$.

Assumindo a independência entre as observações podemos escrever a função

de verossimilhança do modelo como $L(\boldsymbol{\beta}|y) = \prod_n^{i=1} \pi(\boldsymbol{x_i})^{y_i} (1 - \pi(\boldsymbol{x_i}))^{1-y_i}$, substituindo os valores de $\pi(\boldsymbol{x_i})$ e aplicando o logaritmo temos:

$$ln(L(\boldsymbol{\beta}|y)) = \sum_{Y_i=1} ln(\pi(\boldsymbol{x_i})) + \sum_{Y_i=0} ln(1 - \pi(\boldsymbol{x_i})). \tag{2.5}$$

O valor de $\boldsymbol{\beta}$ que maximiza (2.5) é o estimador de máxima verossimilhança denotado por $\hat{\boldsymbol{\beta}}$. Este estimador é obtido por meio de métodos numéricos de otimização.

Sabemos que $\hat{\boldsymbol{\beta}}$ é consistente e assintoticamente eficiente e possui matriz de variâncias e covariâncias $V\left(\hat{\boldsymbol{\beta}}\right)$ dada por:

$$V\left(\hat{\boldsymbol{\beta}}\right) = \left[\sum_{i=1}^{n} \pi(\boldsymbol{x_i})(1 - \pi(\boldsymbol{x_i}))\boldsymbol{x_i}\boldsymbol{x_i}'\right]^{-1}. \tag{2.6}$$

Analisando a expressão (2.6) notamos que os eventos raros afetam a estimação de $V\left(\hat{\boldsymbol{\beta}}\right)$ já que neste tipo de aplicação $\pi(\boldsymbol{x_i}) = P(Y_i = 1|\boldsymbol{x_i})$ é muito baixa para todas as observações. No entanto, caso o modelo possua uma covariável que possua uma associação bastante significativa com a variável resposta a estimativa de $\pi(\boldsymbol{x_i})$ para $Y_i = 1$ será geralmente maior do que para $Y_i = 0$. Assim, o fator $\pi(\boldsymbol{x_i})(1 - \pi(\boldsymbol{x_i}))$ será maior para $Y_i = 1$ do que para $Y_i = 0$, e consequentemente, quanto maior a quantidade de eventos na amostra menor será a variâncias dos estimadores dos parâmetros do modelo.

2.3 Interpretação dos coeficientes do modelo

Sabemos que a interpretação de qualquer modelo de regressão exige a possibilidade de extrair informações práticas dos coeficientes estimados. No caso do modelo de regressão logística é fundamental o conhecimento do impacto causado por cada

variável na determinação da probabilidade do evento de interesse.

Uma medida muito útil na interpretação dos coeficientes do modelo de regressão logística é a *odds ratio*. Na presença de apenas uma covariável categórica x, com dois níveis $x = 0$ e $x = 1$, a *odds ratio* é dada por:

$$\Psi = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))}. \tag{2.7}$$

Como $\pi(1) = e^{\beta_0+\beta_1}/1 + e^{\beta_0+\beta_1}$, $\pi(0) = e^{\beta_0}/1 + e^{\beta_0}$, $1 - \pi(1) = 1/1 + e^{\beta_0+\beta_1}$ e $1 - \pi(0) = 1/1 + e^{\beta_0}$ temos que:

$$\Psi = \frac{\left(\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1}{1+e^{\beta_0}}\right)}{\left(\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1}{1+e^{\beta_0+\beta_1}}\right)} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}. \tag{2.8}$$

A *odds ratio* é uma medida de associação largamente utilizada especialmente na pesquisa epidemiológica e pode ser interpretada como a propensão que o indivíduo possui de assumir o evento de interesse quando $x = 1$, comparada com $x = 0$. Por exemplo, se y denota a presença de câncer de pulmão e x é uma variável indicadora que denota se o indivíduo é fumante ($x = 1$) ou não (x=0). Se $\hat{\Psi} = 2$ podemos dizer que o câncer de pulmão é duas vezes mais provável em fumantes.

2.4 Amostras *state-dependent*

Uma estratégia muito comum utilizada na construção de amostras para o ajuste de modelos de regressão logística na situação de dados desbalanceados é selecionar uma amostra contendo todos os eventos presentes na base de dados original e selecionar, via amostragem aleatória simples sem reposição, um número de não eventos igual ou superior ao número de eventos, no entanto este número deve sempre

ser menor do que a quantidade de observações representando não evento presentes na amostra. Estas amostras, denominadas *state-dependent*, são muito utilizadas principalmente no mercado financeiro, no entanto, para validar as inferências realizadas para os parâmetros obtidos por meio destas amostras algumas adaptações são necessárias. Neste trabalho utilizamos o Método de Correção a Priori.

2.4.1 Método de Correção a Priori

A técnica de correção a priori envolve o cálculo dos estimadores de máxima verossimilhança dos parâmetros do modelo de regressão logística e a correção destas estimativas com base na informação a priori da fração de eventos na população τ (prevalência populacional, ou seja, a proporção de eventos na população) e a fração de eventos observados na amostra $\bar{y}$ (prevalência amostral, ou seja, a proporção de eventos na amostra).

No modelo de regressão logística, os estimadores de máxima verossimilhança $\hat{\beta}_i$, $i = 1, \cdots, p-1$, são estimadores consistentes e eficientes dos β_i considerando o planejamento amostral em questão. No entanto, para que $\hat{\beta}_0$ seja consistente e eficiente ele deve ser corrigido de acordo com a seguinte expressão:

$$\hat{\beta}_0 - ln\left[\left(\frac{1-\tau}{\tau}\right)\left(\frac{\bar{y}}{1-\bar{y}}\right)\right]. \tag{2.9}$$

A maior vantagem da técnica de correção a priori é a facilidade de uso, já que os parâmetros do modelo de regressão logística podem ser estimados da forma usual e apenas o intercepto deve ser corrigido.

2.5 Os estimadores KZ para o Modelo de Regressão Logística

Segundo King e Zeng (2001) o estimador $\hat{\boldsymbol{\beta}}$ é viciado, na situação de eventos raros, para $\boldsymbol{\beta}$ mesmo quando o tamanho da amostra é grande . Além disso, mesmo que $\hat{\boldsymbol{\beta}}$ seja corrigido pelo vício estimado, $P\left(Y=1|\hat{\boldsymbol{\beta}}, \boldsymbol{x_i}\right)$ é viciado para $\pi(\boldsymbol{x_i})$. Nesta seção, discutimos métodos para a correção destes estimadores.

2.5.1 Correção nos parâmetros

Sabemos que o estimador de máxima verossimilhança para $\boldsymbol{\beta}$, $\hat{\boldsymbol{\beta}}$, é assintoticamente não viciado. No entanto, quando os dados em questão são extremamente desbalanceados este vício persiste mesmo quando o tamanho da amostra é grande.

Segundo McCullagh e Nelder (1989) o vício do estimador do vetor de parâmetros de qualquer modelo linear generalizado pode ser estimado como:

$$vicio\left(\hat{\boldsymbol{\beta}}\right) = \left(X^{'}WX\right)^{-1} X^{'}W\xi, \tag{2.10}$$

sendo $X'WX$ é a matriz de informação de Fisher, $\xi_i = -0,5\mu_i^{''}/\mu_i^{'}Q_{ii}$, μ_i é a inversa da função de ligação que relaciona $\mu_i = E\left(Y_i\right)$ ao preditor linear $\eta_i = x_i^{'}\beta$, $\mu_i^{'}$ e $\mu_i^{''}$ são as derivadas de primeira e segunda ordem de μ_i com relação a η_i e Q_{ii} é o i-ésimo elemento da diagonal principal de $X\left(X^{'}W^{'}X\right)X^{'}$.

Temos que $\mu_i^{'} = exp\left(\eta_i\right)$ $1+exp\left(\eta_i\right)$ e $\mu_i^{''} = exp\left(\eta_i\right)/\left(1-exp\left(\eta_i\right)\right)/(1+exp\left(\eta_i\right))^3$, logo $\mu_i^{''}/\mu_i^{'} = 1-exp\left(\eta_i\right)/1+exp\left(\eta_i\right)$. Assim,

$$\xi_i = -0,5\left(\frac{1-exp\left(\eta_i\right)}{\left(1+exp\left(\eta_i\right)\right)}\right)Q_{ii}. \tag{2.11}$$

O cálculo deste vício pode ser adaptado quando utilizamos amostras state-dependent considerando $P(Y_i = y_i) = \pi_i^{\omega_1 y_i}(1-\pi_i)^{\omega_0(1-y_i)}$, onde $\omega_1 = \frac{\tau}{\bar{y}}$ e $\omega_0 = \frac{1-\tau}{1-\bar{y}}$.

Assim, $\mu_i = E(Y_i) = \left(\frac{1}{1+exp(-\eta_i)}\right)^{\omega_1} \equiv \pi_i^{\omega_1}$,
$\mu_i^{'} = \omega_1 \pi_i^{\omega_1}(1-\pi_i)$, $\mu_i^{''} = \omega_1 \pi_i^{\omega_1}(1-\pi_i)[\omega_1 - (1-\omega_1)\pi_i]$ e $\xi_i = 0,5 Q_{ii}[(1-\omega_1)\pi_i - \omega_1]$.

Logo a matriz de informação de Fisher do modelo é dada por:

$$-E\left(\frac{\partial^2 L_\omega(\boldsymbol{\beta}|y)}{\partial\beta_j \partial\beta_k}\right) = \sum_{i=1}^{n} \pi_i(1-\pi_i) x_j \omega_i x_k^{'} = \left[X^{'} W_\omega X\right]_{j,k}, \tag{2.12}$$

com $W_\omega = diag[\pi_i(1-\pi_i)\omega_i]$.

Assim, o estimador corrigido pelo vício é dado por $\tilde{\boldsymbol{\beta}} = \hat{\boldsymbol{\beta}} - vicio\left(\hat{\boldsymbol{\beta}}\right)$. Segundo McCullagh e Nelder (1989) a matriz de variâncias e covariâncias de $\tilde{\boldsymbol{\beta}}$ é aproximadamente $\left(\frac{n}{n+p-1}\right)^2 V\left(\hat{\boldsymbol{\beta}}\right)$. Como $\left(\frac{n}{n+p-1}\right)^2 < 1$ temos que $V\left(\tilde{\boldsymbol{\beta}}\right) < V\left(\hat{\boldsymbol{\beta}}\right)$, ou seja, a diminuição no vício dos estimadores do modelo causa uma diminuição na variância dos mesmos.

2.5.2 Correção nas probabilidades estimadas

De acordo com os resultados apresentados na seção anterior $\tilde{\boldsymbol{\beta}}$ é menos viciado do $\hat{\boldsymbol{\beta}}$ para $\boldsymbol{\beta}$, além disso $V\left(\tilde{\boldsymbol{\beta}}\right) < V\left(\hat{\boldsymbol{\beta}}\right)$. Assim, $\tilde{\pi}(\boldsymbol{x_i})$ é preferível a $\hat{\pi}(\boldsymbol{x_i})$.

No entanto, segundo Geisser (1993) e King et.al. (2001) este estimador não é ótimo porque não leva em conta a incerteza a respeito de $\boldsymbol{\beta}$. Ignorar esta incerteza na estimação pode gerar estimativas viesadas da probabilidade de evento.

Uma maneira de levar em contar a incerteza na estimação do modelo é escrever $\pi(\boldsymbol{x_i})$ como:

$$P(Y_i = 1) = \int P(Y_i = 1|\boldsymbol{\beta^*}) P(\boldsymbol{\beta^*}) d\boldsymbol{\beta^*}, \tag{2.13}$$

com $P(\cdot)$ representando a incerteza com relação à $\boldsymbol{\beta}$. Sob o ponto de vista Bayesiano

usamos a densidade a posteriori de $\boldsymbol{\beta} \sim Normal\left[\boldsymbol{\beta}|\tilde{\boldsymbol{\beta}}, V\left(\tilde{\boldsymbol{\beta}}\right)\right]$. Uma forma aproximada de escrever (2.13) é expandir em série de Taylor a expressão $\pi(\boldsymbol{x_0}) = \frac{e^{\boldsymbol{x_0'\beta}}}{1+e^{\boldsymbol{x_0'\beta}}}$ em torno de $\tilde{\boldsymbol{\beta}}$ até a segunda ordem. Assim,

$$P\left(Y_0 = 1\right) \approx \tilde{\pi}(\boldsymbol{x_0}) + \left[\frac{\partial \pi(\boldsymbol{x_0})}{\partial \boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)$$

$$+\frac{1}{2}\left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)\left[\frac{\partial^2 \pi(\boldsymbol{x_0})}{\partial \beta' \partial \boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} \left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right), \tag{2.14}$$

sendo que $\left[\frac{\partial \pi(\boldsymbol{x_0})}{\partial \boldsymbol{\beta}}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = \tilde{\pi}(\boldsymbol{x_0})\left(1 - \tilde{\pi}(\boldsymbol{x_0})\right)\boldsymbol{x_0'}\left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right)$,
$\left[\frac{\partial \pi(\boldsymbol{x_0})}{\partial \boldsymbol{\beta}\partial \boldsymbol{\beta}'}\right]_{\boldsymbol{\beta}=\tilde{\boldsymbol{\beta}}} = (0,5 - \tilde{\pi}(\boldsymbol{x_0}))\,\tilde{\pi}(\boldsymbol{x_0})\left(1 - \tilde{\pi}(\boldsymbol{x_0})\right)\boldsymbol{x_0'} D \boldsymbol{x_0}$ e D é uma matriz de ordem k x k sendo que o elemento k, j da matriz de D é igual a $\left(\beta_k - \tilde{\beta}_k\right)\left(\beta_j - \tilde{\beta}_j\right)$. Sob a perspectiva Bayesiana, $\pi(\boldsymbol{x_0})$ e $\boldsymbol{\beta}$ são variáveis aleatórias, mas por outro lado, $\tilde{\pi}(\boldsymbol{x_0})$ e $\tilde{\boldsymbol{\beta}}$ são função dos dados. Tomando a esperança da expressão (2.10) temos:

$$P\left(Y_0 = 1\right) = E\left(\frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}}\right) \approx \tilde{\pi}(\boldsymbol{x_0}) +$$

$$\tilde{\pi}(\boldsymbol{x_0})\left(1 - \tilde{\pi}(\boldsymbol{x_0})\right)\boldsymbol{x_0'}b + \left(0.5 + \tilde{\pi}(\boldsymbol{x_0})\right)\left(\tilde{\pi}(\boldsymbol{x_0}) - \tilde{\pi}^2(\boldsymbol{x_0})\right)\boldsymbol{x_0'}\left[V\left(\tilde{\boldsymbol{\beta}}\right) + bb'\right]\boldsymbol{x_0'}, \tag{2.15}$$

onde $b = E\left(\boldsymbol{\beta} - \tilde{\boldsymbol{\beta}}\right) \approx 0$, assim podemos escrever $\pi(\boldsymbol{x_i})$ como:

$$\pi_i = P\left(Y_i = 1\right) = \pi(\tilde{\boldsymbol{x_i}}) + C_i. \tag{2.16}$$

A constante C_i é dada por $C_i = (0.5 - \tilde{\pi}(\boldsymbol{x_i}))\,\tilde{\pi}(\boldsymbol{x_i})\left(1 - \tilde{\pi}(\boldsymbol{x_i})\right)\boldsymbol{x_i'} V\left(\tilde{\boldsymbol{\beta}}\right)\boldsymbol{x_i}$, representando o fator de correção.

Analisando o fator de correção da expressão acima nota-se que ele será maior à medida que o número de zeros na amostra diminui, pois este é diretamente propor-

cional a $V(\tilde{\beta})$.

O estimador da probalidade de sucesso $\pi(\boldsymbol{x_i})^* = \tilde{\pi}(\boldsymbol{x_i}) + C_i$ é chamado de estimador KZ1. Estudos de simulação revelam que este estimador possui o quadrado médio do erro menor do que estimador usual.

Segundo King e Zeng(2001) o fator de correção C_i é um estimador do vício da probabilidade de sucesso. Assim, um estimador aproximadamente não viciado para a probabilidade é dado por: $\pi(\boldsymbol{x_i})^{**} = \tilde{\pi}(\boldsymbol{x_i}) - C_i$. Este estimador é chamado de estimador KZ2.

Capítulo 3

O Modelo Logito Limitado

3.1 Introdução

O modelo logito limitado provém de uma modificação do modelo logito usual. Essa modificação é dada pelo acréscimo de um parâmetro que quantifica um limite superior para a probabilidade de sucesso. Assim, a probabilidade de sucesso dada as covariáveis é dada por:

$$\pi\left(\boldsymbol{x_i}\right) = \omega \frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}},\ 0 < \omega < 1. \tag{3.1}$$

Este modelo foi proposto por Cramer (2004). Nesse trabalho Cramer aplicou o modelo de regressão logística usual, o modelo complementar log-log e o modelo logito limitado em uma base de dados de uma instituição financeira holandesa. Os dados em questão apresentavam baixa incidência do evento de interesse. O teste de Hosmer e Lemeshow indicou que o modelo logito limitado foi o mais adequado na base de dados analisada. Segundo Cramer, o parâmetro ω tem a capacidade de absorver o impacto de possíveis covariáveis significativas excluídas da base de dados.

O modelo logito limitado também foi utilizado por Moraes e Diniz (2008) em dados reais de fraude bancária. De acordo com os resultados obtidos o modelo logito limitado apresentou uma performance superior ao modelo logito usual de acordo com as estatísticas que medem a qualidade do ajuste: AIC(Critério de Informação de Akaike), SC(Critério de Schwartz) e KS(Estatística de Kolmogorov-Smirnov).

Neste trabalho, temos como objetivo comparar o modelo logito limitado com o modelo logito usual tanto na qualidade do ajuste quanto na capacidade preditiva utilizando dados simulados e dados reais.

3.2 Estimação

Como a variável resposta $Y_i \sim Bernoulli(\pi(\boldsymbol{x_i}))$ as probabilidades de sucesso e fracasso são dadas por $P(Y_i = 1|\boldsymbol{x_i}) = \pi(\boldsymbol{x_i})$ e $P(Y_i = 1|\boldsymbol{x_i}) = 1 - \pi(\boldsymbol{x_i})$, respectivamente. Assim, o logaritmo da função de verossimilhança é dado por:

$$l(\beta, \omega) = \sum_{i=1}^{n} \left(y_i ln \left(\omega \left(\frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} \right) \right) + (1 - y_i) ln \left(1 - \omega \left(\frac{e^{\boldsymbol{x_i'\beta}}}{1 + e^{\boldsymbol{x_i'\beta}}} \right) \right) \right) I_{(0,1)}(\omega). \tag{3.2}$$

Os estimadores de Máxima Verossimilhança são obtidos maximizando a expressão 3.2. As derivadas da função de verossimilhança com relação aos parâmetros $\beta_0, \beta_1, \cdots, \beta_{p-1}$ e ω são dadas por:

$$\sum_{i=1}^{n} \omega(y_i - \pi(\boldsymbol{x_i})), \tag{3.3}$$

$$\sum_{k=1}^{p-1} \sum_{i=1}^{n} x_{ij} \omega(y_i - \pi(\boldsymbol{x_i})), \tag{3.4}$$

e

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

Notamos que estas equações são não-lineares nos parâmetros, fato que impossibilita a solução explícita do sistema de equações. Assim, precisamos recorrer a algum método de otimização para encontrar as estimativas de máxima verossimilhança dos parâmetros em questão. No entanto, devido as características da função, sua maximização nem sempre é possível utilizando os procedimentos usuais de otimização numérica. Uma alternativa é considerar a reparametrização $\theta = log\left(\frac{\omega}{1-\omega}\right)$, neste caso a função de verossimilhança pode ser escrita como:

$$l\left(\beta,\omega\right)=\sum_{i=1}^{n}\left(y_i ln\left(\left(\frac{e^{\theta}}{1+e^{\theta}}\right)\left(\frac{1}{1+exp(-x_i'\beta)}\right)\right)\right.$$

$$\left.+(1-y_i)ln\left(1-\left(\frac{e^{\theta}}{1+e^{\theta}}\right)\left(\frac{1}{1+exp(-x_i'\beta)}\right)\right)\right, \tag{3.6}$$

com $-\infty<\theta<\infty$. Para maximizar 3.6 utilizamos o algoritmo BFGS, implementado no software R, proposto simultaneamente e independentemente por Broyden(1970), Fletcher(1970), Godfarb(1970) e Shano(1970).

3.3 Interpretação dos coeficientes do modelo

Como já ressaltamos no capítulo anterior, é fundamental o conhecimento do impacto causado por cada covariável na determinação da probabilidade do evento de interesse.

Na presença de apenas uma covariável categórica x, com dois níveis $x=0$ e $x=1$, temos que $\pi(1)=\omega\frac{e^{\beta_0}}{1+e^{\beta_0}}$, $\pi(0)=\omega\frac{e^{\beta_0}}{1+e^{\beta_0}}$, $1-\pi(1)=\frac{1+e^{\beta_0+\beta_1}(1-\omega)}{1+e^{\beta_0+\beta_1}}$ e $1-\pi(0)=$

$\frac{1+e^{\beta_0}(1-\omega)}{1+e^{\beta_0}}$. Assim, a *razão das chances* é dada por:

$$\Psi = \frac{\left(\omega\frac{e^{\beta_0+\beta_1}}{1+e^{\beta_0+\beta_1}}\right)\left(\frac{1+e^{\beta_0+\beta_1}}{1+(1-\omega)e^{\beta_0+\beta_1}}\right)}{\left(\omega\frac{e^{\beta_0}}{1+e^{\beta_0}}\right)\left(\frac{1+e^{\beta_0}}{1+e^{\beta_0}(1-\omega)}\right)} = \frac{e^{\beta_1}+(1-\omega)e^{\beta_0+\beta_1}}{1+(1-\omega)e^{\beta_0+\beta_1}} \tag{3.7}$$

A propensão que um indivíduo possui de assumir o evento de interesse quando $x = 1$, comparada com $x = 0$ é dada por $\frac{e^{\beta_1}+(1-\omega)e^{\beta_0+\beta_1}}{1+(1-\omega)e^{\beta_0+\beta_1}}$.

3.4 O Método BFGS

O Método BFGS (Broyden, Fletcher, Goldfarb e Shano) é uma técnica de otimização que utiliza um esquema iterativo para buscar um ponto ótimo. O processo de otimização parte de um valor inicial θ_0 e na iteração t verificamos se o ponto θ_t encontrado é ou não o ponto ótimo. Caso este não seja o ponto ótimo, calculamos um vetor direcional Δ_t e realizamos uma otimização secundária, conhecida como "busca em linha", para encontrar o tamanho do passo ótimo λ_t. De forma que, $\theta_{t+1} = \theta_t + \lambda_t\Delta_t$, onde será realizada uma nova busca pelo ponto ótimo.

O vetor direcional Δ_t é tomado como $\Delta_t = \omega_t g_t$, em que g_t é o gradiente (vetor de primeiras derivadas) no passo t e ω_t é uma matriz positiva-definida calculada no passo t.

O Método BFGS, assim como o Método de Newton-Raphson, são casos particulares do método gradiente. O Método de Newton-Raphson utiliza $\omega_t = -H^{-1}$, sendo H a matriz hessiana, entretando quando o valor do ponto inicial θ_0 não está próximo do ponto ótimo a matriz $-H^{-1}$ pode não ser positiva definida dificultando o uso do método.

No método BFGS, uma estimativa de $-H^{-1}$ é construída iterativamente. Para tanto, gera-se uma sequência de matrizes $\omega_{t+1} = \omega_t + E_t$. A matriz ω_0 é a matriz

identidade e E_t é também uma matriz positiva definida, pois em cada passo do processo iterativo ω_{t+1} é a soma de duas matrizes positiva definida.

A matriz E_t é dada por:

$$E_t = \frac{\delta_t \delta_t}{\delta_t' \gamma_t} + \frac{\omega_t \gamma_t \gamma_t' \omega_t}{\gamma_t' \omega_t \gamma_t} - \nu_t dt \tag{3.8}$$

com $\delta_t = \lambda_t \Delta_t = \theta_{t+1} - \theta_t$, $\gamma_t = g(\theta_{t+1}) - g(\theta_t)$, $\nu_t = \gamma_t' \omega_t \gamma_t$ e $d_t = \left(\frac{1}{\gamma_t \delta_t}\right) \gamma_t - \left(\frac{1}{\gamma_t' \omega_t \gamma_t}\right) \omega_t \gamma_t$.

Capítulo 4

O Modelo Logito Generalizado

4.1 Introdução

O modelo de regressão logística usual é largamente utilizado para modelar a dependência de dados binários e variáveis explicativas. Este sucesso deve-se a sua vasta aplicabilidade, simplicidade da sua fórmula e sua fácil interpretação. Este modelo funciona bem em muitas situações, contudo ele tem como suposição a simetria no ponto $\frac{1}{2}$ da curva de probabilidade esperada $\pi(x)$ e que sua forma seja a da função de distribuição acumulada da distribuição logística. Segundo Stukel(1988) nas situações em que as caudas da distribuição de $\pi(x)$ são mais pesadas, o modelo logito usual não funciona bem.

Na Figura 4.1 encontram-se os gráficos da curva de probabilidade $\pi(x)$ considerando as prevalâncias amostrais de 1%, 15%, 30% e 50%. De acordo com estes gráficos, na situação de baixa prevalência, a suposição de simetria na curva $\pi(x)$ no ponto $\frac{1}{2}$ não é verificada. Este fato indica que o modelo logito usual não é adequado para modelar dados com desbalanceamento acentuado.

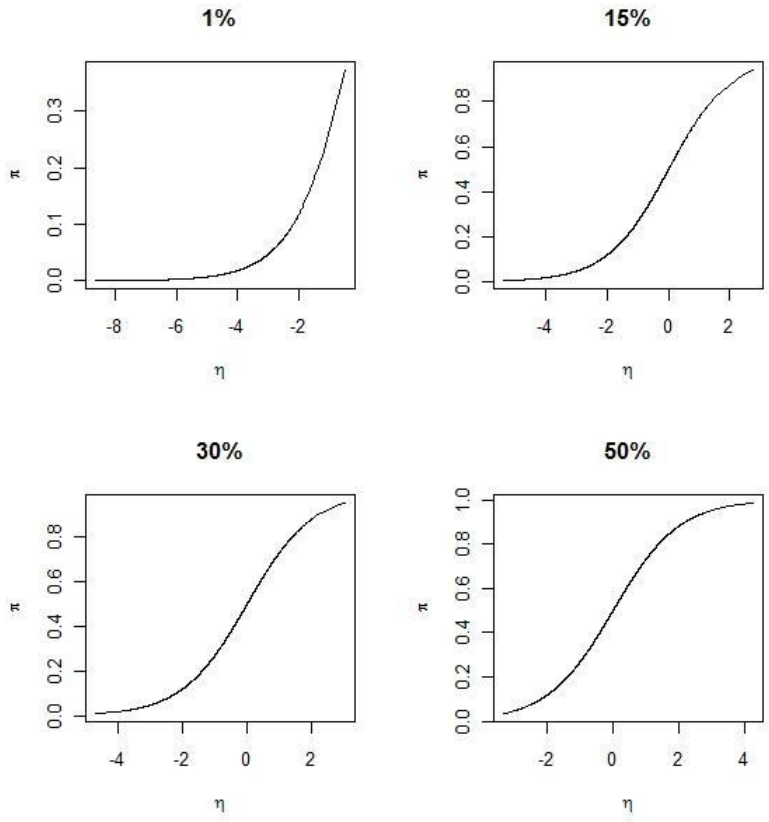

Figura 4.1: Curvas de probalidade considerando diferentes prevalências.

Muitos autores apresentaram propostas de modelos que generalizam o modelo logito padrão. Prentice (1976) sugeriu uma ligação bi-paramétrica, utilizando a função de distribuição acumulada da transformação $log\left(F_{2m_1,2m_2}\right)$. A família de distribuições $log(F)$ contém a distribuição logística($m_1 = m_2 = 1$), a Gaussiana, as distribuições do Mínimo e Máximo Extremo, a Exponencial, a distribuição de Laplace e a Exponencial Refletida. Este modelo é eficaz em muitas situações devido a sua flexibilidade, no entanto, computacionalmente este apresenta dificuldades já que as curvas de probabilidades estimadas devem ser calculadas através da soma de séries infinitas.

Pregibon (1980) definiu uma família de funções de ligação que inclui a ligação logito como um caso especial. A curva de probabilidade esperada é a solução implícita da equação $\left(\pi^{\lambda_1-\lambda_2}-1\right)/\left(\lambda_1-\lambda_2\right)-\left(\left(1-\pi\right)^{\lambda_1+\lambda_2}-1\right)/\left(\lambda_1+\lambda_2\right)=\eta$. O parâmetro λ_1 controla as caudas da distribuição e λ_2 determina a simetria da curva de probabilidade π. Aranda-Ordaz(1981) sugeriram dois modelos uniparamétricos, um deles simétrico e o outro assimétrico, como alternativa ao modelo logito padrão. O modelo simétrico é dado pela transformação $2(\pi^{\delta_1}-(1-\pi)^{\delta_1})/\delta_1(\pi^{\delta_1}+(1-\pi)^{\delta_1})=\eta$, quando $\delta_1 \to 0$ no limite é o modelo logito. Já o modelo assimétrico é dado por $log(((1-\pi)^{-\delta_2}-1)/\delta_2)=\eta$, quando $\delta_2=1$ este é o modelo logito e se $\delta_2=0$ o modelo é o complementar log-log.

4.2 O Modelo Logito Generalizado

A forma geral do modelo logito generalizado, proposto por Stukel (1988), é dada por:

$$\pi_\alpha(\boldsymbol{x_i})=\frac{e^{h_\alpha(\eta)}}{1+e^{h_\alpha(\eta)}}, \tag{4.1}$$

sendo que,

$$log\left(\frac{\pi_\alpha(\boldsymbol{x_i})}{1-\pi_\alpha(\boldsymbol{x_i})}\right)=h_\alpha(\eta), \tag{4.2}$$

com $h_\alpha(\eta)$ são funções não-lineares estritamente crescentes indexadas por dois parâmetros de forma α_1 e α_2.

Para $\eta \geq 0$, $(\pi \geq \frac{1}{2})$ $h_\alpha(\eta)$ é dada por:

$$h_\alpha = \begin{cases} \alpha_1^{-1}(e^{(\alpha_1|\eta|)} - 1), \ \alpha_1 > 0, \\ \eta, \ \alpha_1 = 0, \\ -\alpha_1^{-1} log\left(1 - \alpha_1 |\eta|\right), \ \alpha_1 < 0, \end{cases} \tag{4.3}$$

e para $\eta \leq 0$ $(\pi \leq \frac{1}{2})$,

$$h_\alpha = \begin{cases} \alpha_2^{-1}(e^{(\alpha_2|\eta|)} - 1), \ \alpha_2 > 0, \\ \eta, \ \alpha_2 = 0, \\ -\alpha_2^{-1} log\left(1 - \alpha_2 |\eta|\right), \ \alpha_2 < 0. \end{cases} \tag{4.4}$$

Para $\alpha_1 = \alpha_2 = 0$, o modelo resultante é o logito usual.

As funções h aumentam mais rapidamente ou mais lentamente que a curva do modelo logito usual como podemos ver na Figura 4.2. Os parâmetros α_1 e α_2 determinam o comportamento das caudas. Se $\alpha_1 = \alpha_2$, a curva de probabilidade correspondente é simétrica.

4.3 Estimação

Os estimadores de Máxima Verossimilhança de $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ podem ser obtidos utilizando o algoritmo delta, sugerido por Jorgensen(1984). Este algoritmo é equivalente a usar o procedimento de Mínimos Quadrados Ponderados para o ajuste dos parâmetros de modelos lineares generalizados, porém, neste caso a matriz do modelo é atualizada depois de cada iteração. No caso do modelo logito generalizado a matriz do modelo é a matriz usual $\boldsymbol{X}$ acrescida de duas colunas contendo as variáveis,

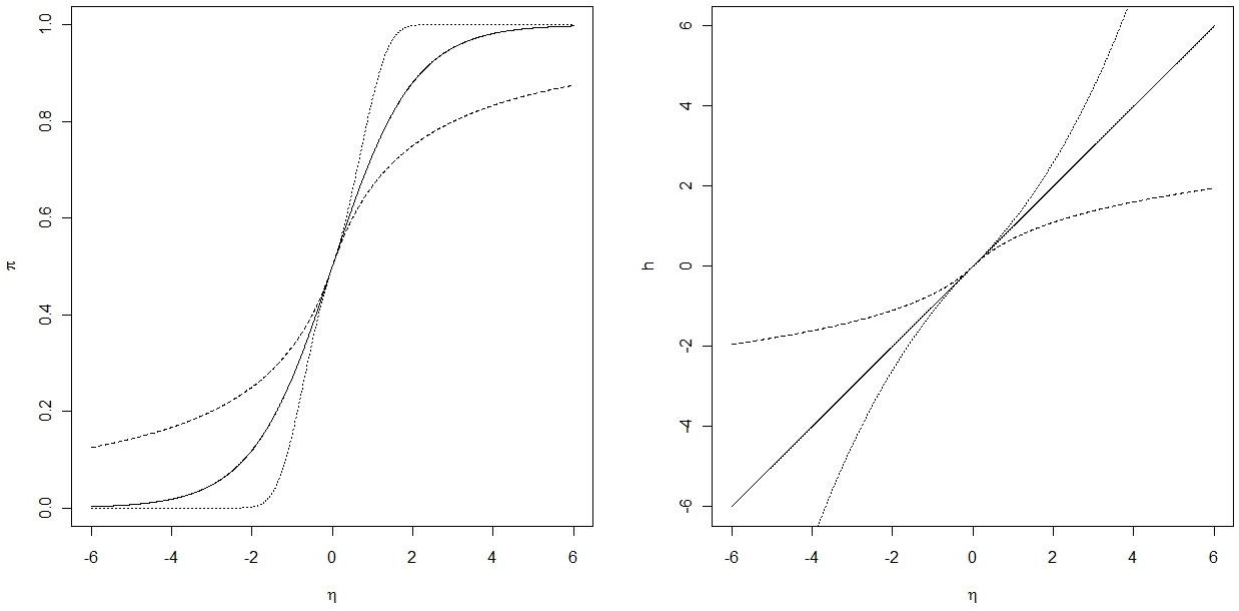

Figura 4.2: Gráfico de h e π, a linha sólida representa o modelo logito usual, a linha tracejada corresponde ao modelo logito generalizado com $\boldsymbol{\alpha} = (-1, -1)$ e a linha pontilhada corresponde ao modelo logito generalizado com $\boldsymbol{\alpha} = (0.25, 0.25)$.

$(z_{1,t+1}, z_{2,t+1} = -\frac{\partial g(\pi)}{\partial \alpha_1}, -\frac{\partial g(\pi)}{\partial \alpha_2})|_{\hat{\boldsymbol{\beta}}, \hat{\boldsymbol{\alpha}}_t}$ com:

$$z_{i,t+1} = \alpha_i^{-2}(\alpha_i|\eta| - 1 + exp(-\alpha_i|\eta|)sgn(\eta)),\ \alpha_i > 0,$$

$$z_{i,t+1} = \frac{1}{2}\eta^2 sgn(\eta),\ \alpha_i = 0,$$

e

$$z_{i,t+1} = \alpha_i^{-2}(\alpha_i|\eta| + (1 - \alpha_i|\eta|)log(1 - \alpha_i|\eta|))sgn(\eta), \alpha_i < 0.$$

sendo que $\alpha_i = \hat{\alpha_{i,t}}$, $\eta = \hat{\eta}_t = \boldsymbol{x\beta_t}$ e $(\hat{\boldsymbol{\beta}}_{\boldsymbol{t}}, \hat{\boldsymbol{\alpha}}_{\boldsymbol{t}})$ é a estimativa de $(\boldsymbol{\beta}, \boldsymbol{\alpha})$ na t-ésima iteração. Os elementos de $\boldsymbol{z}$ correspondem aos parâmetros de forma e devem ser atualizados a cada iteração.

Uma maneira alternativa de estimar os parâmetros do modelo logito generalizado consiste em estimar o vetor de parâmetros $\boldsymbol{\beta}$ considerando vários valores de $\boldsymbol{\alpha}$ e escolher como estimativa o conjunto de valores que maximize a verossimilhança (Stukel(1985)). Neste trabalho utilizamos este método para estimar os parâmetros do modelo logito generalizado.

Capítulo 5

Modelo Logito com resposta de origem

5.1 Introdução

Em muitas situações práticas possuímos uma variável resposta binária com distribuição original pertencente a alguma classe de distribuições. Ou seja, a variável resposta possui alguma distribuição original que não é a de Bernoulli e, por alguma razão, foi dicotomizada através de um ponto de corte C arbitrário. Assim, podemos adicionar características da distribuição original da variável resposta no modelo de regressão logística usual. Esta metodologia foi proposta, inicialmente, por Suissa(1995) e ampliada por Suissa e Blais (1995) em uma estrutura de modelos lineares generalizados com função de ligação composta para ajustar modelos de regressão logística com resposta log-normal. Paula e Diniz(2011) estenderam esta técnica para outras variáveis resposta pertencentes a família exponencial.

Neste trabalho, apresentamos a construção e o desenvolvimento dos modelos de

regressão logística para os casos de variável resposta com distribuição normal, exponencial e log-normal.

5.2 Modelo Normal

Seja $R_1, R_2, \cdots, R_n$ variáveis aleatórias independentes seguindo distribuição $N(\mu_i, \sigma^2)$, para $i = 1, \cdots, n$. Considerando C um ponto de corte arbitrário e $Y_1, Y_2, \cdots, Y_n$ tal que $Y_i = 1$ se $R_i > C$ e $Y_i = 0$ se $R_i \leq C$, $i = 1, \cdots, n$. Dessa forma, $P(Y_i = 1) = P(R_i > C) = \pi_i$ e $P(Y_i = 0) = P(R_i \leq C) = 1 - \pi_i$, $i = 1, \cdots, n$. Assim, $Y_i \sim Bernoulli(\pi_i)$.

Na presença de $p - 1$ covariáveis relacionadas com a variáveis resposta, a probabilidade de sucesso pode ser escrita através do modelo de regressão logística na forma:

$$\pi(\boldsymbol{x_i}) = E(Y_i) = P(Y_i) = 1 = \frac{exp(\boldsymbol{x_i'\beta})}{1 + exp(\boldsymbol{x_i'\beta})} = g^{-1}(\boldsymbol{x_i'\beta}), i = 1, \cdots, n, \quad (5.1)$$

com $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$, o vetor de parâmetros associado as covariáveis do modelo. Logo,

$$\pi(\boldsymbol{x_i}) = P(Y_i > C) = P\left[Z_i > \frac{C - \mu_i}{\sigma}\right] = P\left[Z_i < \frac{\mu_i - C}{\sigma}\right] = \phi\left(\frac{\mu_i - C}{\sigma}\right), \quad (5.2)$$

sendo Z_i uma variável aleatória com distribuição normal padrão e distribuição acumulada ϕ. Das equações 5.1 e 5.2 temos que:

$$\pi(\boldsymbol{x_i}) = \phi\left(\frac{\mu_i - C}{\sigma}\right) = g^{-1}(\boldsymbol{x_i'\beta}), i = 1, \cdots, n, \quad (5.3)$$

ou ainda,

$$g\left(\pi(\boldsymbol{x_i})\right) = g\left[\phi\left(\frac{\mu_i - C}{\sigma}\right)\right] = \boldsymbol{x_i'\beta} = \eta_i, i = 1, \cdots, n, \tag{5.4}$$

com $g\left[\phi\left(\cdot\right)\right]$ uma função de ligação composta que origina o preditor linear $\boldsymbol{x_i'\beta}$. Tomando $\gamma_i = (\mu_i - C)/\sigma$ e assumindo σ conhecido, pode-se dizer que este modelo faz parte da classe dos modelos lineares generalizados, cujo componente aleatório é o conjunto de variáveis independentes com distribuição $N\left(\gamma_i, 1\right)$ e o componente sistemático dado pela função de ligação composta $g\left[\phi\left(\cdot\right)\right]$ e pelo preditor linear $\eta_i = \boldsymbol{x_i'\beta}, i = 1, \cdots, n$.

A partir de 5.3 podemos escrever μ_i como:

$$\mu_i = \sigma\phi^{-1}\left[g^{-1}\left(\boldsymbol{x_i'\beta}\right)\right] + C, i = 1, \cdots, n. \tag{5.5}$$

Logo, a função de verossimilhança pode ser escrita como:

$$L\left(\boldsymbol{\beta}, \boldsymbol{\sigma^2}; \boldsymbol{r}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(r_i - \sigma\phi^{-1}\left[g^{-1}\left(\boldsymbol{x_i'\beta}\right)\right] - C\right)^2\right\}. \tag{5.6}$$

E o logaritmo da função de verossimilhança é dado por:

$$lnL\left(\boldsymbol{\beta}, \sigma; \boldsymbol{r}\right) = l\left(\beta, \sigma; \boldsymbol{r}\right) = -\frac{n}{2}ln\left(2\pi\sigma^2\right) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left(r_i - \sigma\phi^{-1}\left[g^{-1}\left(\boldsymbol{x_i'\beta}\right)\right] - C\right)^2. \tag{5.7}$$

5.3 Modelo Exponencial

Seja $R_1, R_2, \cdots, R_n$ variáveis aleatórias independentes seguindo distribuição $Exponencial\,(\theta_i)$, $i = 1, \cdots, n$, isto é,

$$f\,(r_i) = \theta_i exp\,(-\theta_i r_i)\,, \theta_i > 0, i = 1, \cdots, n. \tag{5.8}$$

Considerando C um ponto de corte arbitrário e $Y_1, Y_2, \cdots, Y_n$ tal que $Y_i = 1$ se $R_i > C$ e $Y_i = 0$ se $R_i \leq C$, $i = 1, \cdots, n$. Dessa forma, $P\,(Y_i = 1) = P\,(R_i > C) = \pi_i$ e $P\,(Y_i = 0) = P\,(R_i \leq C) = 1 - \pi_i$, $i = 1, \cdots, n$. Assim, $Y_i \sim Bernoulli\,(\pi_i)$.

Na presença de $p-1$ covariáveis relacionadas com a variáveis resposta, a probabilidade de sucesso pode ser escrita através do modelo de regressão logística como:

$$\pi(\boldsymbol{x_i}) = E\,(Y_i) = P\,(Y_i = 1) = \frac{exp\,(\boldsymbol{x_i'\beta})}{1 + exp\,(\boldsymbol{x_i'\beta})} = g^{-1}\,(\boldsymbol{x_i'\beta})\,, i = 1, \cdots, n, \tag{5.9}$$

com $\boldsymbol{\beta} = (\beta_0, \beta_1, \cdots, \beta_{p-1})'$, o vetor de parâmetros associado as covariáveis do modelo. Dessa forma,

$$P\,(R_i > C) = exp\,(-\theta_i C)\,, i = 1, \cdots, n. \tag{5.10}$$

Das equações 5.9 e 5.10 temos:

$$exp\,(-\theta_i C) = g^{-1}\,(\boldsymbol{x_i'\beta})\,, \tag{5.11}$$

isolando θ_i, temos:

$$\theta_i = -\frac{-ln\,[g^{-1}\,(\boldsymbol{x_i'\beta})]}{C}. \tag{5.12}$$

Portanto,

$$g\,[exp\,(-\theta_i C)] = \boldsymbol{x_i'\beta}. \tag{5.13}$$

sendo $g\left[exp\left(\cdot\right)\right]$ a função de ligação que origina o preditor linear $\boldsymbol{x_i'\beta}$.

A função de verossimilhança para o modelo logístico com resposta exponencial é dada por:

$$L\left(\boldsymbol{\beta};\boldsymbol{r}\right)=\prod_{i=1}^{n}\left\{-\frac{ln\left[g^{-1}\left(\boldsymbol{x_i'\beta}\right)\right]\left[g^{-1}\left(\boldsymbol{x_i'\beta}\right)\right]^{\frac{r_i}{C}}}{C}\right\}. \tag{5.14}$$

Aplicando o logaritmo em 5.14 temos:

$$l\left(\boldsymbol{\beta};\boldsymbol{r}\right)=\sum_{i=1}^{n}ln\left\{-ln\left[g^{-1}\left(\boldsymbol{x_i'\beta}\right)\right]\right\}+\frac{1}{C}\sum_{i=1}^{n}r_i ln\left[g^{-1}\left(\boldsymbol{x_i'\beta}\right)\right]-nln\left(C\right). \tag{5.15}$$

5.4 Modelo Log-Normal

Seja $R_1, R_2, \cdots, R_n$ variáveis aleatórias independentes seguindo distribuição $LN\left(\mu_i,\sigma^2\right)$, para $i=1,\cdots,n$. Ou seja, $ln\left(R_1\right), ln\left(R_2\right), \cdots, ln\left(R_n\right)$ seguem distribuição normal com média μ_i e variância σ^2. Considerando C um ponto de corte arbitrário e $Y_1, Y_2, \cdots, Y_n$ tal que $Y_i=1$ se $R_i>C$ e $Y_i=0$ se $R_i\leq C$, $i=1,\cdots,n$. Dessa forma, $P\left(Y_i=1\right)=P\left(R_i>C\right)=\pi_i$ e $P\left(Y_i=0\right)=P\left(R_i\leq C\right)=1-\pi_i$, $i=1,\cdots,n$. Assim, $Y_i\sim Bernoulli\left(\pi_i\right)$.

Devido a sua relação com a distribuição $Normal$, os resultados para a distribuição $Log-Normal$ podem ser obtidos utilizando os resultados presentes na seção 5.2. Basta apenas substituir a constante C por $log(C)$ e a variável resposta R_i por $log(R_i)$, para $i=1,\cdots,n$. Dessa forma, a probabilidade de sucesso $\pi(\boldsymbol{x_i})$ é dada por:

$$\pi(\boldsymbol{x_i})=P\left[Z_i<\frac{\mu_i-log(C)}{\sigma}\right]=\phi\left[\frac{\mu_i-log(C)}{\sigma}\right], i=1,\cdots,n, \tag{5.16}$$

sendo Z_i uma variável aleatória com distribuição normal padrão e distribuição acumulada ϕ. Logo, de 5.16:

$$\mu_i = \sigma\phi^{-1}\left[g^{-1}\left(\boldsymbol{x_i'}\boldsymbol{\beta}\right)\right] + log(C). \tag{5.17}$$

Considerando 5.17, a função de verossimilhança pode ser escrita como:

$$L\left(\boldsymbol{\beta}, \sigma; \boldsymbol{r}\right) = \left(2\pi\sigma^2\right)^{-\frac{n}{2}} exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}\left[log(r_i) - \mu_i\right]^2\right\}, \tag{5.18}$$

com $\mu_i = \sigma\phi^{-1}\left[g^{-1}(\boldsymbol{x_i'}\boldsymbol{\beta})\right] + log(C)$, $\imath = 1, \cdots, n$. O logaritmo da função de verossimilhança pode ser escrito como:

$$lnL(\boldsymbol{\beta}, \sigma; \boldsymbol{r}) = -\frac{n}{2}log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}\left\{log(r_i) - \sigma\phi^{-1}\left[g^{-1}(\boldsymbol{x_i'}\boldsymbol{\beta})\right] - log(C)\right\}^2. \tag{5.19}$$

Capítulo 6

Comparação dos Modelos

6.1 Introdução

Neste Capítulo apresentamos as técnicas utilizadas na comparação dos modelos estudados. Para verificar a qualidade dos ajustes foram utilizadas as medidas AIC, BIC e -2log(Verossimilhança). Na análise do poder preditivo dos modelos utilizamos as medidas sensibilidade, especificidade, valor preditivo positivo, valor preditivo negativo, acurácia e coeficiente de correlação de Mathews. Além disso, utilizamos os testes não paramétricos de Mann-Whitney e Kruskal-Wallis para verificar a igualdade de médias destas medidas obtidas por meio dos modelos analisados.

6.2 Qualidade do Ajuste

Um modelo é uma representação simplificada de alguma situação da vida real destinado a ilustrar alguns aspectos do problema sem se ater a todos os detalhes. Muitas vezes mais de um modelo pode descrever o mesmo fenômeno, assim existe a

necessidade de obter medidas baseadas em pricípios científicos que permitam escolher qual deles é mais adequado para cada situação.

Uma maneira de comparar dois ou mais modelos é através do valor da medida -2log(verossimilhança). Quanto menor esta medida melhor é o modelo. No entanto, tal método não fornece uma verdadeira comparação pois não leva em conta a quantidade de parâmetros de cada modelo.

Akaike (1974) utilizou a Informação de Kullback-Leibler para verificar se um dado modelo é adequado. Seja p o número de parâmetros do modelo, a medida AIC é dada por:

$$AIC = -2log(L(\hat{\theta})) + 2p. \tag{6.1}$$

O Critério de Informação Bayesiano (BIC) proposto por Schwartz (1978) leva em conta o número de parâmetros do modelo p e o tamanho da amostra n e é dado por:

$$BIC = -2log(L(\hat{\theta})) + plog(n). \tag{6.2}$$

6.3 Curva ROC

A curva ROC (receiver-operating characteristic) foi desenvolvida no início dos anos 50 com o objetivo de quantificar a habilidadade dos radares em distinguir um sinal de ruído (Zweig e Campbell, 1993). Na década seguinte, a curva ROC foi utilizada em psicologia experimental, e na década de 70 esta técnica se disseminou em vários ramos da pesquisa biomédica.

Esta metodologia é uma forma gráfica de se avaliar a qualidade de um modelo de classificação. Além disso, ela pode ser utilizada na determinação de um ponto

de corte P_c utilizado como valor limite para classificar uma nova observação como evento ou não-evento.

O ponto de corte P_c é um número entre 0 e 1, e para cada valor temos um modelo de classificação diferente e, dessa forma, podemos calcular as probabilidades condicionais de que uma observação que apresente a ocorrência do evento ser classificada corretamente pelo modelo e de uma observação que represente não evento ser classificada como tal, estas medidas são denominadas respectivamente, sensibilidade(SE) especificidade(ES).

Para cada ponto de corte P_c calculamos as medidas sensibilidade e especificidade. Quanto maior o ponto de corte maior será a especificidade e menor a sensibilidade do modelo e em contrapartida, quanto menor o ponto de corte, maior a sensibilidade e menor a especificidade. Assim, a curva ROC é construída tendo no seu eixo horizontal os valores de (1-especificidade) e no seu eixo vertical os valores da sensibilidade.

Sabemos que a escolha de um ponto de corte ótimo depende do objetivo de estudo do modelo. De acordo com Shäfer (1989) uma seleção de pontos de corte ótimos pode ser obtida através da combinação linear máxima entre a sensibilidade e a especificidade. Neste trabalho, utilizamos a curva ROC para escolher o ponto de corte de cada modelo estudado, sendo este a probabilidade do evento de interesse em que curva mais se aproxima do canto superior esquerdo.

6.4 Medidas Preditivas

Após construir um modelo de clasificação devemos avaliar a capacidade do mesmo em distinguir entre as observações em que há ou não a ocorrência do evento de interesse (Guirado, 2010). Esta avaliação é realizada por meio da comparação das previsões do modelo com a verdadeira classificação da observação.

Seja D uma variável que indica a classificação original da resposta de interesse, podendo portanto assumir dois valores 0 e 1 e T, uma variável que corresponde a classificação do modelo obtida por meio do ponto de corte adotado, pondendo assumir também dois valores, 0 e 1. Para cada observação calculamos a probabilidade de ocorrência do evento de interesse p_i, com $0 < p_i < 1$. Escolhido o ponto de corte P_c, temos que $T = 1$ se $p_i > P_c$ e $T = 0$ se $p_i < P_c$.

Para cada ponto de corte P_c podemos construir a matriz de confusão apresentada na Tabelas 6.1.

Tabela 6.1: Matriz de Confusão.

	Real		
Resultado do Modelo	Positivo(D+)	Negativo(D-)	Total
Positivo(T+)	a (VP)	b (FP)	a+b
Negativo(T-)	c (FN)	d (VN)	c+d
Total	a+c	b+d)	a+b+c+d

Analisando a matriz de confusão temos:

- a representa o número de observações representando evento classificadas como tal, ou seja, os verdadeiros positivos (VP);

- b representa o número de observações representando evento classificadas incorretamente, ou seja, o número de resultados falso-positivos(FP);

- c é o número de observações representando não evento classificadas como evento, ou seja, o número de resultados falso-negativos(FN);

- d é o número de observações representando não evento classificadas corretamente, ou seja, os resultados verdadeiro negativo (VN);

- $a + c$ é o total de observações representando evento;

- $b + d$ total de observações representando não evento;
- $a + b$ total de observações classificadas como evento;
- $c + d$ total de observações classificadas como não evento;

Neste trabalho utilizamos como medidas preditivas: a sensibilidade, a especificidade, o valor preditivo positivo, o valor preditivo negativo, a acurácia e o coeficiente de Mathews. Todas estas medidas são calculadas utilizando os valores das medidas explicitadas na Matriz de Confusão.

Outra medida de extrema importância, principalmente quando tratamos de eventos raros, é a *Prevalência* que pode ser definida como a proporção de observações propensas a característica de interesse. Ou seja, a probabilidade de uma observação apresentar a característica de interesse antes do modelo ser ajustado e pode ser estimada por:

$$p = P(D+) = \frac{a + c}{a + b + c + d}. \tag{6.3}$$

6.4.1 Sensibilidade(SE)

Sensibilidade (SEN) é definida como a probabilidade de que o modelo de classificação resulte em um resultado positivo, dado que a observação correspondente é relativa a um evento. Ou seja, a sensibilidade corresponde à proporção de eventos que são classificados como tal, e é dada por:

$$SEN = P(T + |D+) = \frac{VP}{VP + FN} = \frac{a}{a + c}. \tag{6.4}$$

6.4.2 Especificidade(ES)

Especificidade (ES) é a probabilidade de que uma observação que representa um não-evento seja classificada como tal. Ou seja, é proporção de não-eventos classificada corretamente, e é dada por:

$$ES = P(T - |D-) = \frac{VN}{VN + FP} = \frac{d}{b + d}. \tag{6.5}$$

Um modelo muito sensível raramente deixará de diagnosticar a característica de interesse e um modelo muito específico dificilmente classificará como evento uma observação que livre da característica de interesse.

6.4.3 Acurácia

A Acurácia (ACC) é definida como a proporção de acertos de um modelo de classificação, ou seja, o percertual VP e VN dentre todos os possíveis resultados, e é definida por:

$$ACC = \frac{VP + VN}{VP + FP + VN + FN}. \tag{6.6}$$

Esta medida também pode ser vista como uma média ponderada da sensibilidade e da especificidade em relação ao número de observações que apresentam ou não a característica de interesse de uma determinada população. É importante ressaltar que a acurácia não é uma medida que deve ser analisada isoladamente na escolha de um modelo pois ela é influenciada pela sensibilidade, especificidade e prevalência. Além disso, dois modelos com sensibilidade e especificidade muito diferentes podem produzir valores semelhantes de acurácia se forem aplicados a populações com prevalências muito diferentes.

Para ilustrar o efeito da prevalência na acurácia de um modelo podemos supor

uma população que apresente 5% de seus integrantes com a característica de interesse, então se um modelo classificar todos os indivíduos como não portadores da característica temos um percentual de acerto de 95%, ou seja, a acurácia é alta e o modelo é pouco informativo.

6.4.4 Valor Preditivo Positivo(VPP) e Valor Preditivo Negativo (VPN)

O Valor Preditivo Positivo (VPP) de um modelo é a proporção de observações representando o evento de interesse dentre aqueles indivíduos que o modelo identificou como evento. Já o Valor Preditivo Negativo (VPN) é a proporção de individuos que representam não evento dentre aqueles identificados como não evento pelo modelo. O VPP e o VPN são definidos respectivamente como:

$$VPP = P(D + |T+) = \frac{VP}{VP + FP} = \frac{a}{a+b}, \tag{6.7}$$

e

$$VPP = P(D - |T-) = \frac{VN}{VN + FN} = \frac{d}{c+d}. \tag{6.8}$$

Estas medidas devem ser interpretadas com cautela, pois sofrem a influência da prevalência populacional.

Caso a estimativa da sensibilidade e da especificidade sejam confiáveis, o valor preditivo positivo (VPP) pode ser estimado via Teorema de Bayes, utilizando uma estimativa da prevalência (Linnet, 1988):

$$VPP = \frac{Sensibilidade \times Prevalencia}{Sensibilidade \times Prevalencia + (1 - Especificidade) \times (1 - Prevalencia)}. \tag{6.9}$$

Da mesma forma, o valor preditivo negativo (VPN) pode ser estimado por:

$$VPN = \frac{Especificidade \times (1 - Prevalencia)}{Especificidade \times (1 - Prevalencia) + Sensibilidade \times Prevalencia}. \tag{6.10}$$

6.4.5 Coeficiente de Correlação de Mathews (MCC)

O coeficiente de correlação proposto por Matthews (Mathews, 1975) é uma medida de desempenho que pode ser utilizada no caso de prevalências extremas. É uma adaptação do Coeficiente de Correlação de Pearson e mede o quanto as variáveis T e D tendem a apresentar o mesmo sinal de magnitude após serem padronizadas, Baldi(2000).

O (MCC) retorna um valor entre -1 e +1. Um valor de 1 representa uma previsão perfeita, um acordo total, o valor 0 representa uma previsão completamente aleatória e -1 uma previsão inversa, ou seja, total desacordo. O MCC utiliza as 4 medidas apresentadas na matriz de confusão (VP, VN, FP e FN) e é dado por:

$$MPP = \frac{VP * VN - FP * FN}{\sqrt{(VP + FP)\,(VP + FN)\,(VN + FP)\,(VN + FN)}}. \tag{6.11}$$

Capítulo 7

Simulações

Neste Capítulo apresentamos um estudo de simulações realizado com o intuito de comparar a qualidade de ajuste e a capacidade preditiva dos modelos de regressão logística com estimadores usuais, KZ1 e KZ2, modelo logito limitado e generalizado.

Apresentamos um estudo comparativo do vício dos estimadores de máxima verossimilhança usual e KZ para o modelo de regressão logística considerando diferentes prevalências. Adotamos a prevalência de 1% na geração dos dados artificiais.

As medidas preditivas utilizadas são a sensibilidade, a especificidade, o valor preditivo positivo, o valor preditivo negativo, a acurácia e o coeficiente de correlação de Mathews. A qualidade de ajuste é verificada de acordo com o AIC, BIC e -2log(verossimilhança).

Incialmente utilizamos o modelo logito usual na geração dos dados artificiais, simulamos uma covariável com distribuição normal e fixamos valores para os parâmetros β_0 e β_1. Foram geradas 200 amostras de tamanho 20000 que foram divididas em amostras treinamento com 70% dos dados utilizada no desenvolvimento dos modelos e amostra teste com 30% da amostra original utilizada na verificação da capacidade

preditiva de cada modelo. Com estes dados ajustamos os modelos logito usual com os estimadores de máxima varossimilhança convencionais e estimadores KZ e também os modelos logito limitado e logito generalizado. O processo foi repetido utilizando os modelos logito limitado e logito generalizado na geração dos dados.

7.1 Comparação do vício entre os estimadores usuais e KZ para o modelo de regressão logística

Nesta seção analisamos o desempenho dos estimadores usuais e KZ para o modelo de regressão logística no que diz respeito ao vício amostral.

Foram geradas 1000 amostras de tamanho 1000 com duas variáveis, a variável resposta com distribuição de Bernoulli e uma covariável com distribuição normal. O valor dos parâmetros do modelo foram estabelecidos em $\beta_0 = -4,5$ e $\beta_1 = 1$, e a variável resposta foi gerada seguindo distribuição de Bernoulli com média $\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$. A prevalência populacional foi tomada como a média das prevalências amostrais. A simulação foi repetida considerando as prevalências de 1%, 5% e 10%, obtidas mudando os valores dos parâmetros da covariável.

De cada amostra foram retiradas sucessivamente 10%, 30%, 50%, 70% e 90% das observações representando não evento. O modelo de regressão logística com estimadores usuais e estimadores corrigidos pelo vício foram ajustados na amostra completa e nas sub-amostras obtidas.

Nas Figuras 7.1 e 7.2 encontramos o vício dos parâmetros β_0 e β_1 considerando a prevalência de 1%, respectivamente. Já nas Figuras 7.3 e 7.4 encontramos o vício dos parâmetros β_0 e β_1 considerando a prevalência de 5%, respectivamente. E, nas Figuras 7.5 e 7.6 encontramos o vício dos parâmetros β_0 e β_1 considerando a prevalência

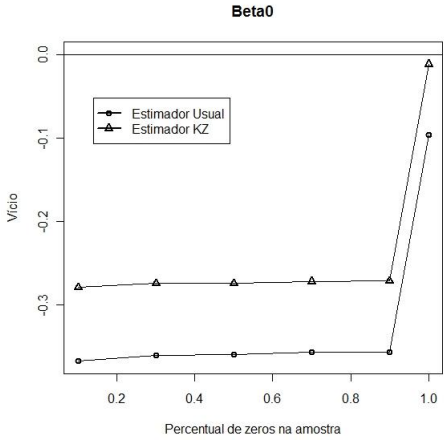

Figura 7.1: Vício de β_0 com prevalência de 1%.

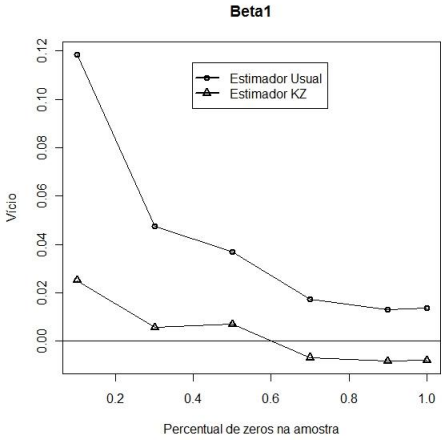

Figura 7.2: Vício de β_1 com prevalência de 1%.

de 10%, respectivamente.

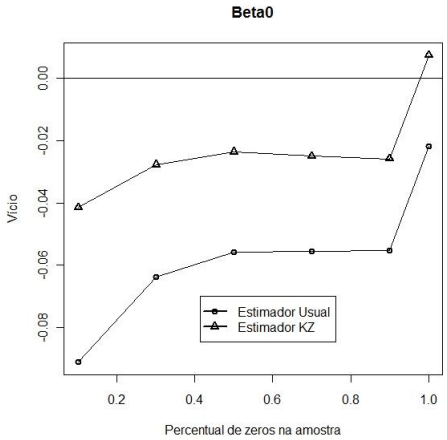

Figura 7.3: Vício de β_0 com prevalência de 5%.

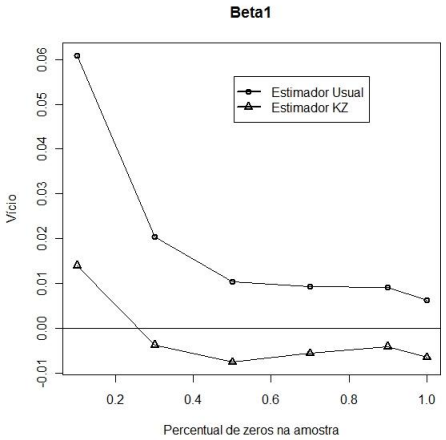

Figura 7.4: Vício de β_1 com prevalência de 5%.

Notamos que o vício dos estimadores KZ é sempre inferior ao vício dos estimadores usuais de máxima verossimilhança. Além disso, a medida que retiramos observações representando não evento das amostras o vício de ambos os estimadores aumenta.

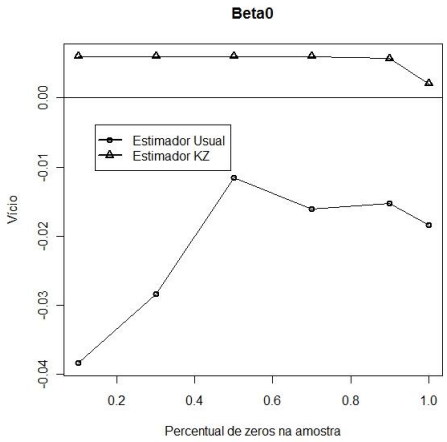

Figura 7.5: Vício de β_0 com prevalência de 10%.

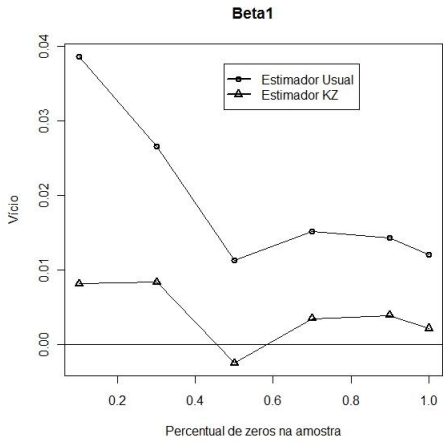

Figura 7.6: Vício de β_1 com prevalência de 10%.

7.2 Modelo Logito Usual

Nesta seção apresentamos os resultados encontrados através dos dados gerados utilizando o modelo logito usual.

Inicialmente estabelecemos os valores dos parâmetros em $\beta_0 = 1$ e $\beta_1 = -5$, geramos 200 amostras de tamanho 20000 simulando uma covariável x com distribuição normal com média 2,5 e variância 1 e a variável resposta foi gerada seguindo distribuição de Bernoulli com média $\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$. Cada amostra foi separada em amostra treinamento com 70% dos dados utilizada no desenvolvimento dos modelos e amostra teste com 30% dos dados utilizada na verificação da capacidade preditiva dos modelos.

Em cada amostra ajustamos os modelos logito usual com estimadores convencionais e estimadores KZ e os modelos logito limitado e logito generalizado. Na Figura 7.7 encontramos um gráfico de dispersão da medida AIC do modelo logito usual e logito limitado, na Figura 7.8 apresentamos o gráfico de dispersão do BIC

do modelo logito usual e logito limitado e na Figura 7.9 encontramos o gráfico de dispersão da estatística -2log(verossimilhança) dos modelos logito usual e logito limitado na situação em que os dados foram gerados através do modelo logito usual. Observamos que de acordo com as medidas de qualidade de ajuste analisadas o desempenho dos modelos logito usual e logito limitado é bastante parecido. Este fato é reflexo das estimativas do parâmetro ω do modelo logito limitado que ficaram muito próximas de 1 em todas as amostras simuladas.

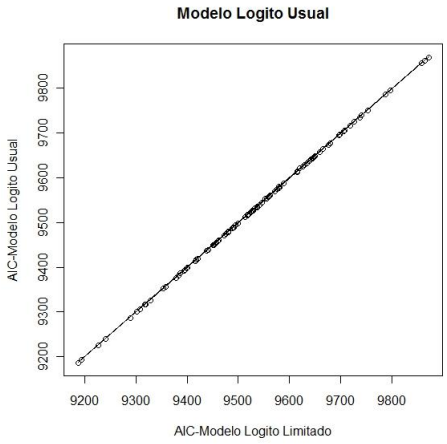

Figura 7.7: AIC - Modelo logito usual x Modelo logito limitado.

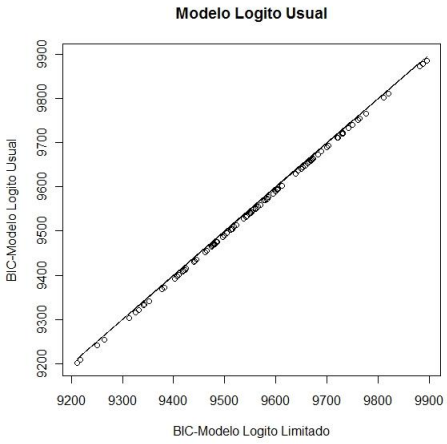

Figura 7.8: BIC - Modelo logito usual x Modelo logito limitado.

Na Figura 7.10 encontramos um gráfico de dispersão do AIC do modelo logito usual e logito generalizado, na Figura 7.11 apresentamos o gráfico de dispersão do BIC do modelo logito usual e logito generalizado e na Figura 7.12 encontramos o gráfico de dispersão da estatística -2log(verossimilhança) dos modelos logito usual e logito generalizado na situação em que as amostras foram geradas através do modelo logito usual. Observamos que de acordo com as medidas de qualidade de ajuste analisadas o desempenho dos modelos logito usual é bastante superior ao desempenho

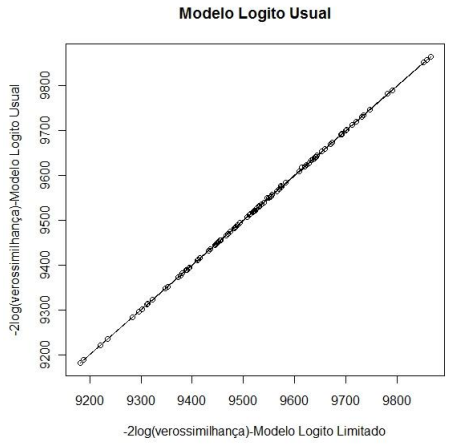

Figura 7.9: -2log(verossimilhança) - Modelo logito usual x Modelo logito limitado.

do modelo logito generalizado.

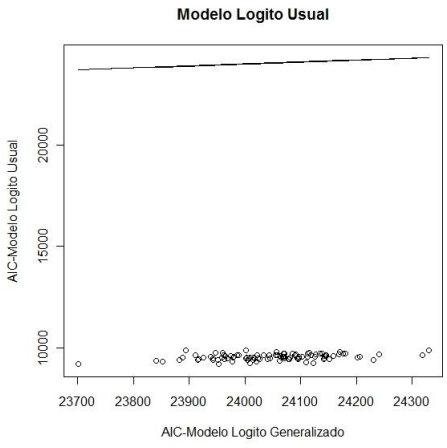

Figura 7.10: AIC - Modelo logito usual x Modelo logito generalizado.

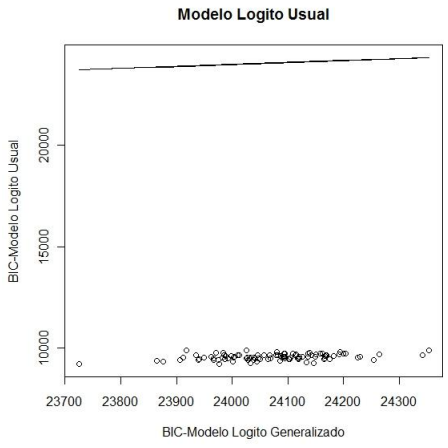

Figura 7.11: BIC - Modelo logito usual x Modelo logito generalizado.

Na Figura 7.13 encontramos um gráfico de dispersão do AIC do modelo logito generalizado e logito limitado, na Figura 7.14 apresentamos o gráfico de dispersão do BIC do modelo logito generalizado e logito limitado e na Figura 7.15 encontramos

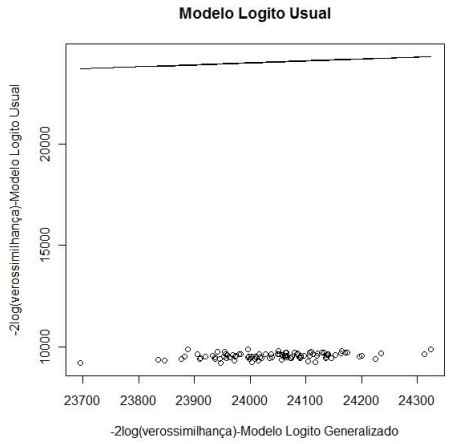

Figura 7.12: -2log(verossimilhança) - Modelo logito usual x Modelo logito generalizado.

o gráfico de dispersão da estatística -2log(verossimilhança) dos modelos logito generalizado e logito limitado na situação em que as amostras foram geradas através do modelo logito usual. Observamos que, de acordo com as medidas de qualidade de ajuste analisadas, o desempenho dos modelos logito limitado é bastante superior ao desempenho do modelo logito generalizado.

Com o intuito de analisar o desempenho preditivo dos modelos de classificação estudados, construímos boxplots das medidas preditivas para cada modelo.

Na Figura 7.16 encontramos os boxplots para a sensibilidade dos modelos ajustados na situação em que os dados foram gerados de acordo com o modelo logito usual. Notamos que a mediana é bastante próxima para todos os modelos, no entanto a variabilidade da medida obtida no modelo logito generalizado é superior a dos outros modelos. Todos os modelos apresentaram pontos atípicos, no entanto, o modelo logito generalizado foi o que apresentou uma maior dispersão.

Já na Figura 7.17 apresentamos os boxplots da medida espeficidade para cada um dos modelos estudados quando os dados foram gerados através do modelo logito

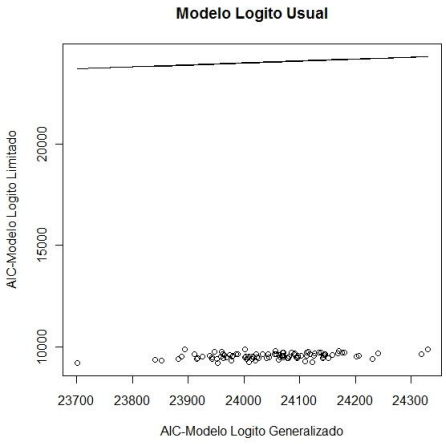

Figura 7.13: AIC - Modelo logito usual x Modelo logito generalizado.

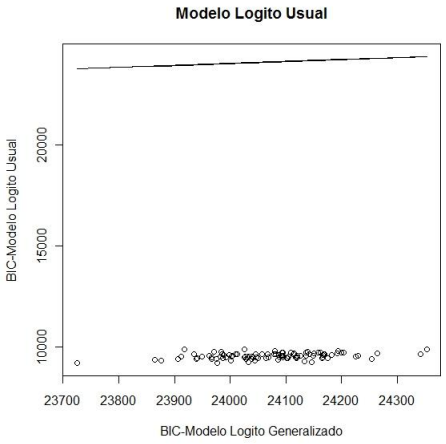

Figura 7.14: BIC - Modelo logito usual x Modelo logito generalizado.

usual. Observamos que a mediana para o modelo logito generalizado é inferior a mediana desta medida para os outros modelos, no entanto, a variabilidade da especificidade para este modelo é inferior a dos demais modelos.

Os boxplots do valor preditivo positivo obtidos através dos dados gerados através do modelo logito usual encontram-se na Figura 7.18. A mediana desta medida para o modelo logito generalizado é inferior a mediana dos demais modelos no entanto, a variabilidade desta medida para o modelo logito generalizado também é inferior a varibialidade desta medida para os demais modelos. Todos os modelos apresentaram ponto atípicos no entanto, a amplitude do valor preditivo positivo calculada com base no modelo logito generalizado foi a menor.

Na Figura 7.19 apresentamos os boxplots para o valor preditivo negativo para dados gerados através do modelo logito usual. A mediana desta medida para todos os modelos é bastante próxima, no entanto a variabilidade para o modelo logito generalizado é superior a variabilidade dos outros modelos.

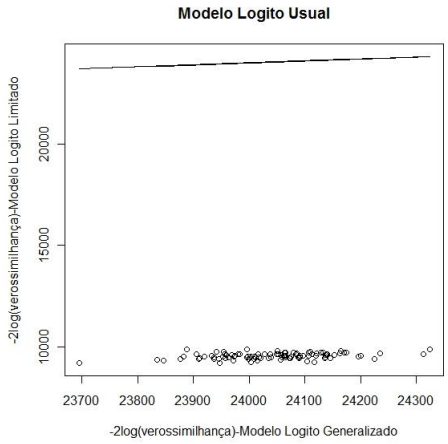

Figura 7.15: -2log(verossimilhança) - Modelo logito usual x Modelo logito generalizado.

Na Figura 7.21 apresentamos os boxplots para o coeficiente de correlação de mathews considerando os dados gerados através do modelo logito usual. A mediana desta medida é inferior para o modelo logito generalizado, enquanto que os demais modelos apresentam um desempenho bastante parecido.

Na Tabela 7.1 encontramos intervalos de confiança empíricos de 90%, 95% e 99% das medidas preditivas dos modelos de classificação ansalisados considerando dados gerados através do modelo logito usual.

7.3 Modelo Logito Limitado

Nesta seção apresentamos os resultados encontrados através dos dados gerados utilizando o modelo logito limitado.

Inicialmente estabelecemos os valores dos parâmetros em $\beta_0 = 1$, $\beta_1 = 5$ e $\omega = 0,01$, geramos 200 amostras de tamanho 20000 simulando uma covariável x com distribuição normal com média 2,5 e variância 1 e a variável resposta foi gerada

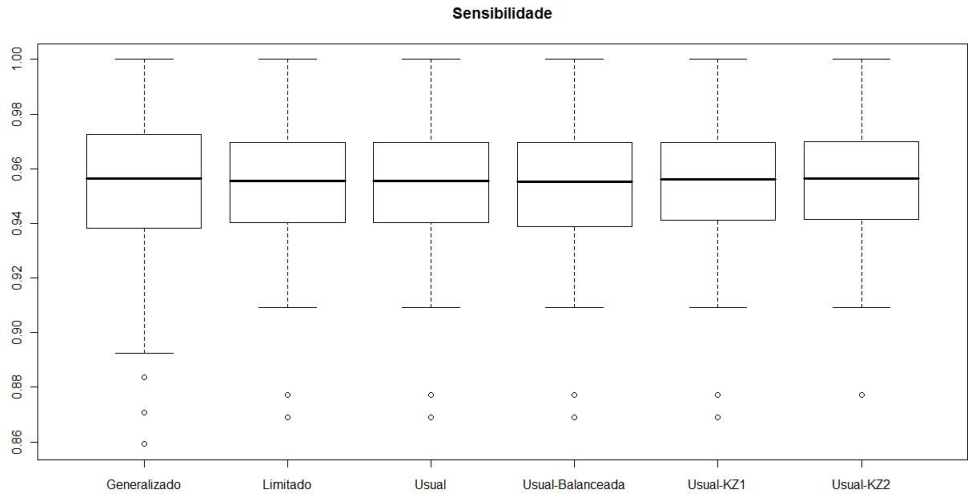

Figura 7.16: Modelo Logito Usual - Sensibilidade.

seguindo distribuição de Bernoulli com média $\pi(x_i) = \omega \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$. Cada amostra foi separada em amostra treinamento do 70% dos dados utilizada no desenvolvimento dos modelos e amostra teste com 30% dos dados utilizada na verificação da capacidade preditiva dos modelos.

Observamos que a covariável apresenta baixa associação com a variável resposta devido a presença do parâmetro ω limitando superiormente a probabilidade de sucesso, este fato é refletido nas medidas preditivas dos modelos de classificação estudados.

Em cada amostra, ajustamos os modelos logito usual com estimadores convencionais e estimadores KZ e os modelos logito limitado e logito generalizado. Na Figura 7.22 encontramos um gráfico de dispersão do AIC do modelo logito usual e logito limitado, na Figura 7.23 apresentamos o gráfico de dispersão do BIC do modelo logito usual e logito limitado e na Figura 7.24 encontramos o gráfico de dispersão da estatística -2log(verossimilhança) dos modelos logito usual e logito limitado. Estas medidas foram calculadas considerando dados gerados através do modelo logito

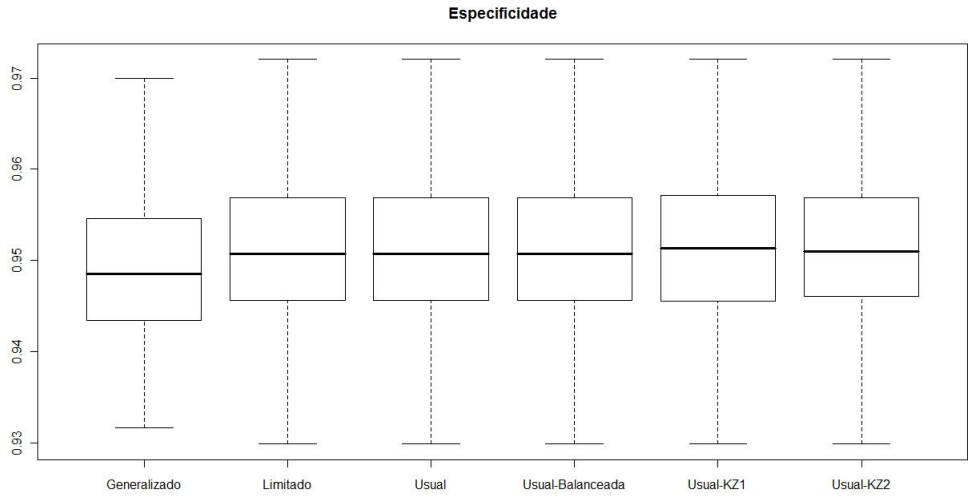

Figura 7.17: Modelo Logito Usual - Especificidade.

limitado.

Na Figura 7.22 encontramos um gráfico de dispersão do AIC do modelo logito usual e logito limitado, na Figura 7.23 apresentamos o gráfico de dispersão do BIC do modelo logito usual e logito limitado e na Figura 7.24 encontramos o gráfico de dispersão da estatística -2log(verossimilhança) dos modelos logito usual e logito limitado. Estas medidas foram calculadas considerando dados gerados através do modelo logito limitado.

Observamos através das medidas de qualidade de ajuste que o modelo logito limitado foi ligeiramente superior ao modelo logito usual em todas as amostras analisadas, considerando a situação em que os dados são gerados de acordo com o modelo logito limitado.

Na Figura 7.28 encontramos um gráfico de dispersão do AIC do modelo logito generalizado e logito limitado, na Figura 7.29 apresentamos o gráfico de dispersão do BIC do modelo logito generalizado e logito limitado e na Figura 7.30 encontramos o

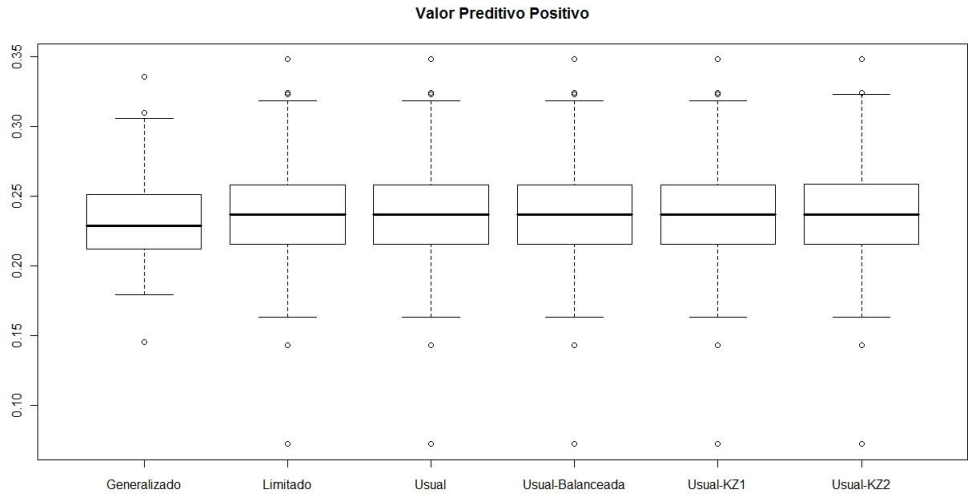

Figura 7.18: Modelo Logito Usual - Valor Preditivo Positivo.

gráfico de dispersão da estatística -2log(verossimilhança) dos modelos logito generalizado e logito limitado. Estas medidas foram calculadas considerando dados gerados através do modelo logito limitado.

Notamos que o modelo logito limitado possui um ajuste mais adequado que o modelo logito generalizado em todas as amostras analisadas na situação em que o mecanismo de geração de dados utiliza o modelo logito limitado.

Na Figura 7.31 encontramos um gráfico de dispersão do AIC do modelo logito generalizado e logito limitado, na Figura 7.32 apresentamos o gráfico de dispersão do BIC do modelo logito generalizado e logito limitado e na Figura 7.33 encontramos o gráfico de dispersão da estatística -2log(verossimilhança) dos modelos logito generalizado e logito limitado.

De acordo com as medidas de qualidade de ajuste analisadas observamos que o desempenho dos modelos logito usual e logito generalizado é bastante similar na situação em os dados são gerados de acordo com o modelo logito limitado.

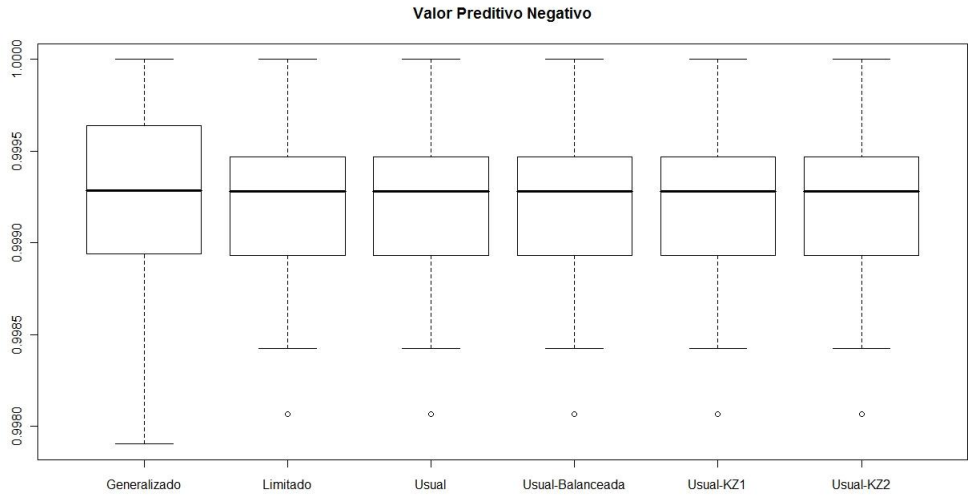

Figura 7.19: Modelo Logito Usual - Valor Preditivo Negativo.

Na Tabela 7.2 encontramos os intervalos empíricos para as medidas preditivas obtidas por meio de cada um dos modelos estudados, na situação em que os dados foram gerados por meio do modelo logito generalizado.

Na Figura 7.31 encontramos os boxplots da medida sensibilidade para os diversos modelos ajustados nos dados gerados por meio do modelo logito limitado. Notamos que o modelo logito generalizado e o modelo logito usual construído em amostras balanceadas apresentaram medianas próximas e menores do que a mediana desta medida calculada por meio dos outros modelos.

Os boxplots para a medida especificidade encontram-se na Figura 7.32 na situação que os dados foram gerados através do modelo logito limitado. Notamos que a dispersão desta medida para o modelo logito usual ajustado em amostras balanceadas com estimadores convencionais e estimadores KZ é superior a dispersão desta mesma medida obtida por meio dos demais modelos de classificação.

Na Figura 7.33 apresentamos os boxplots da medida valor preditivo positivo

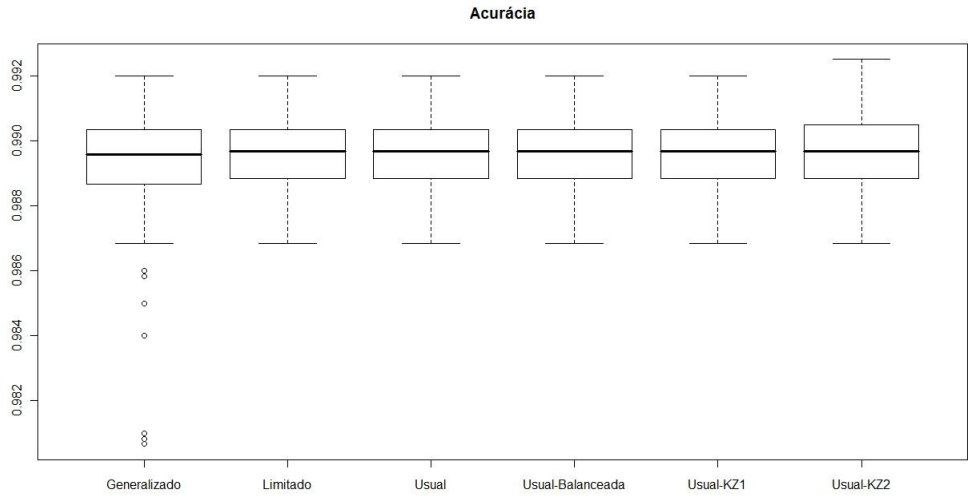

Figura 7.20: Modelo Logito Usual - Acurácia.

obtida através dos diversos modelos de classificação estudados para os dados gerados através do modelo logito limitado. O pior desempenho observado considerando esta medida foi o do modelo logito generalizado.

Na Figura 7.34 apresentamos os boxplots da medida valor preditivo negativo para os dados gerados através do modelo logito limitado. O modelo logito generalizado foi o que apresentou o melhor desempenho preditivo considerando apenas esta medida.

Na Figura 7.35 encontramos os boxplots da acurácia. Todos os modelo analisados apresentaram desempenho similar.

Na Figura 7.36 apresentamos os boxplots do Coeficiente de Correlação de Mathews. Notamos que para todos os modelos esta medida está próxima de zero indicando uma aleatoriedade na previsão de evento, este fato é reflexo da falta de associação entre a variável resposta e covariável.

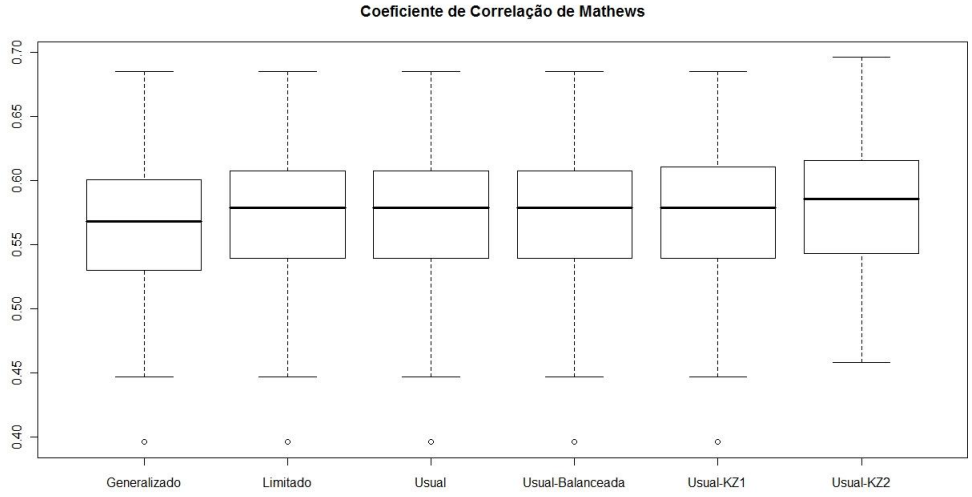

Figura 7.21: Modelo Logito Usual - Coeficiente de Correlação de Mathews.

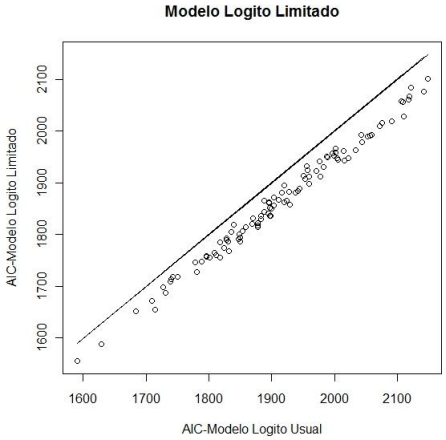

Figura 7.22: AIC - Modelo logito usual x Modelo logito limitado.

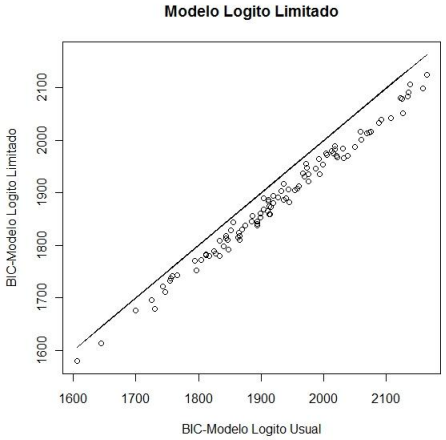

Figura 7.23: BIC - Modelo logito usual x Modelo logito limitado.

7.4 Modelo logito generalizado

Nesta seção apresentamos os resultados encontrados através dos dados gerados utilizando o modelo logito generalizado.

Tabela 7.1: Medidas Preditivas - Modelo Logito Usual.

	Modelos	90%	95%	99%
SEN	Usual	0.912 0.989	0.909 0.989	0.873 1.000
	Usual-Balanceado	0.911 0.989	0.902 1.000	0.876 1.000
	Usual-KZ1	0.908 0.990	0.899 1.000	0.874 1.000
	Usual-KZ2	0.908 0.990	0.905 1.000	0.874 1.000
	Limitado	0.912 0.989	0.909 0.989	0.8730 1.000
	Generalizado	0.894 0.990	0.887 1.000	0.864 1.000
ESP	Usual	0.939 0.966	0.936 0.967	0.930 0.970
	Usual-Balanceado	0.928 0.963	0.924 0.965	0.921 0.967
	Usual-KZ1	0.928 0.964	0.923 0.967	0.919 0.967
	Usual-KZ2	0.927 0.964	0.925 0.967	0.921 0.967
	Limitado	0.939 0.966	0.936 0.967	0.930 0.970
	Generalizado	0.939 0.962	0.936 0.964	0.932 0.967
VPP	Usual	0.182 0.293	0.169 0.320	0.107 0.335
	Usual-Balanceado	0.182 0.293	0.169 0.320	0.107 0.335
	Usual-KZ1	0.182 0.291	0.169 0.320	0.107 0.335
	Usual-KZ2	0.183 0.318	0.169 0.323	0.107 0.347
	Limitado	0.182 0.293	0.169 0.320	0.107 0.335
	Generalizado	0.184 0.283	0.180 0.302	0.162 0.322
VPN	Usual	0.998 0.999	0.998 0.999	0.998 1.000
	Usual-Balanceado	0.998 0.999	0.998 0.999	0.998 1.000
	Usual-KZ1	0.998 0.999	0.998 0.999	0.998 1.000
	Usual-KZ2	0.998 0.999	0.998 0.999	0.998 1.000
	Limitado	0.998 0.999	0.998 0.999	0.998 1.000
	Generalizado	0.998 0.999	0.998 1.000	0.997 1.000
ACC	Usual	0.987 0.991	0.987 0.991	0.986 0.991
	Usual-Balanceado	0.987 0.991	0.987 0.991	0.986 0.991
	Usual-KZ1	0.987 0.991	0.987 0.991	0.986 0.991
	Usual-KZ2	0.988 0.992	0.987 0.992	0.987 0.992
	Limitado	0.987 0.991	0.987 0.991	0.986 0.991
	Generalizado	0.985 0.991	0.980 0.991	0.980 0.991
MCC	Usual	0.477 0.655	0.477 0.655	0.421 0.681
	Usual-Balanceado	0.477 0.655	0.459 0.668	0.421 0.681
	Usual-KZ1	0.471 0.655	0.454 0.668	0.421 0.681
	Usual-KZ2	0.498 0.669	0.474 0.683	0.459 0.693
	Limitado	0.477 0.655	0.459 0.668	0.421 0.681
	Generalizado	0.471 0.652	0.459 0.665	0.421 0.681

Inicialmente estabelecemos os valores dos parâmetros em $\beta_0 = 4$, $\beta_1 = 2$ e $\alpha_1 = 3$, geramos 200 amostras de tamanho 20000 simulando uma covariável x com distribuição normal com média 2,5 e variância 1 e a variável resposta foi gerada seguindo distribuição de Bernoulli com média $\pi(x_i) = \frac{exp(h)}{1+exp(h)}$ com $h = -\frac{1}{\alpha_1}(exp(\alpha_1 |\eta|) - 1)$. Cada amostra foi separada em amostra treinamento com 70% dos dados utilizada no desenvolvimento dos modelos e amostra teste com 30% dos dados utilizada na

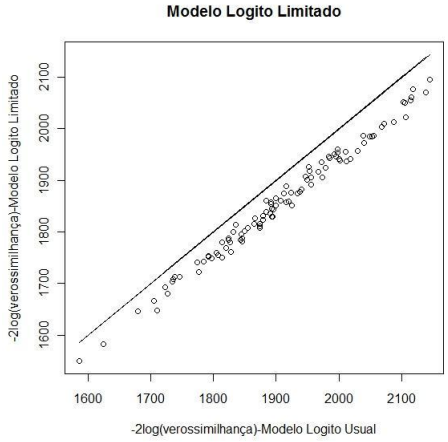

Figura 7.24: -2log(verossimilhança) - Modelo logito usual x Modelo logito limitado.

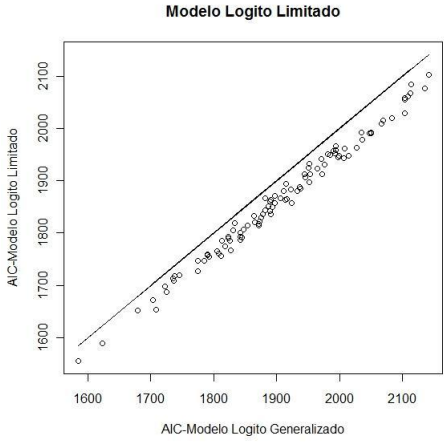

Figura 7.25: AIC - Modelo logito generalizado x Modelo logito limitado.

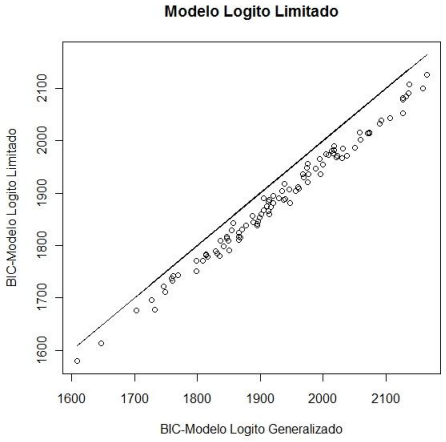

Figura 7.26: BIC - Modelo logito generalizado x Modelo logito limitado.

verificação da capacidade preditiva dos modelos.

Em cada amostra ajustamos os modelos logito usual com estimadores convencionais e estimadores KZ e os modelos logito limitado e logito generalizado. Na Figura 7.37 encontramos um gráfico de dispersão do AIC do modelo logito genera-

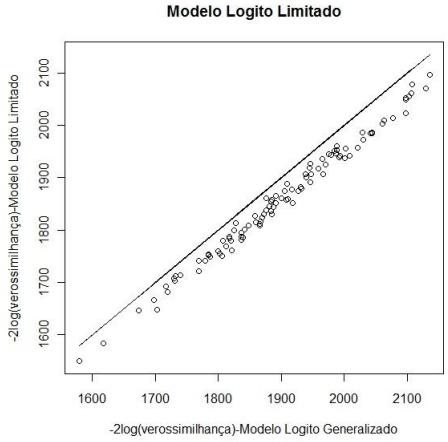

Figura 7.27: -2log(verossimilhança) - Modelo logito generalizado x Modelo logito limitado.

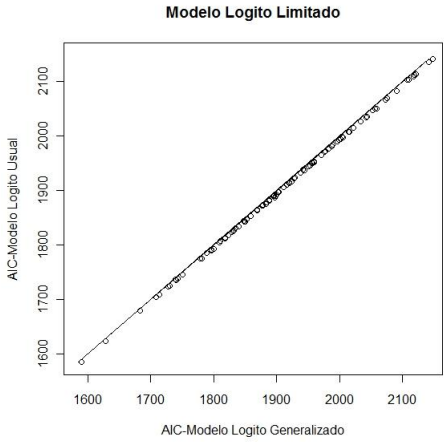

Figura 7.28: AIC - Modelo logito generalizado x Modelo logito usual.

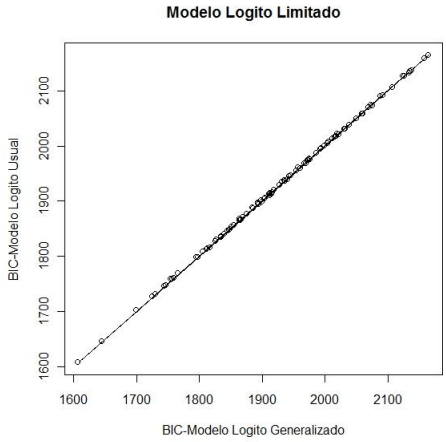

Figura 7.29: BIC - Modelo logito generalizado x Modelo logito usual.

lizado e logito usual, na Figura 7.38 apresentamos o gráfico de dispersão do BIC do modelo logito generalizado e logito usual e na Figura 7.39 encontramos o gráfico de dispersão da estatística -2log(verossimilhança) dos modelos logito generalizado e

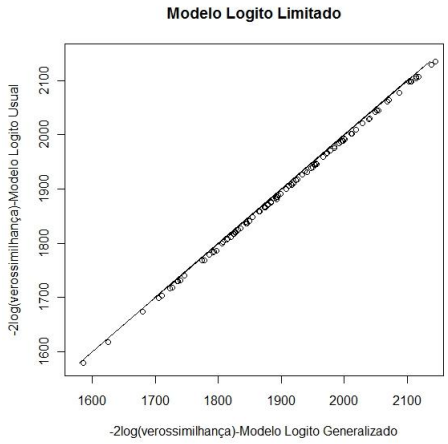

Figura 7.30: -2log(verossimilhança) - Modelo logito generalizado x Modelo logito usual.

logito usual. Estas medidas foram calculadas na situação em que os dados foram gerados através do modelo logito generalizado.

De acordo com as medidas de qualidade de ajuste apresentadas o modelo logito generalizado apresentou-se superior ao modelo logito usual em todas as amostras analisadas quando estas foram geradas através do modelo logito generalizado.

Na Figura 7.40 apresentamos o gráfico de dispersão da medida AIC para os modelos logito generalizado e logito limitado ajustados na amostras geradas através do modelo logito generalizado. Na Figura 7.42 encontramos o gráfico de dispersão da medida BIC dos modelos logito generalizado e logito limitado e na Figura 7.43 apresentamos o gráfico de dispersão da medida -2log(verossimilhança) para tais modelos.

As medidas de qualidade de ajuste apresentadas indicam que o modelo logito generalizado apresentaram um desempenho superior ao modelo logito limitado nas amostras geradas através do modelo logito generalizado.

Nas Figura 7.43, 7.44 e 7.45 encontramos o gráfico de dispersão das medidas AIC, BIC e -2log(verossimilhança) dos modelos logito usual e logito limitado na situação

Tabela 7.2: Medidas Preditivas - Modelo Logito Limitado.

	Modelos	90%	95%	99%
SEN	Usual	0.408 0.619	0.378 0.648	0.357 0.704
	Usual-Balanceado	0.351 0.639	0.326 0.670	0.301 0.703
	Usual-KZ1	0.391 0.617	0.373 0.637	0.331 0.740
	Usual-KZ2	0.390 0.636	0.368 0.659	0.337 0.688
	Limitado	0.408 0.619	0.378 0.648	0.357 0.704
	Generalizado	0.336 0.611	0.319 0.622	0.304 0.655
ESP	Usual	0.432 0.583	0.419 0.592	0.393 0.614
	Usual-Balanceado	0.408 0.616	0.385 0.620	0.363 0.656
	Usual-KZ1	0.395 0.603	0.377 0.622	0.359 0.640
	Usual-KZ2	0.390 0.593	0.368 0.620	0.332 0.637
	Limitado	0.432 0.583	0.419 0.592	0.393 0.614
	Generalizado	0.428 0.605	0.414 0.633	0.410 0.648
VPP	Usual	0.016 0.022	0.015 0.023	0.014 0.024
	Usual-Balanceado	0.016 0.022	0.015 0.023	0.012 0.024
	Usual-KZ1	0.016 0.023	0.015 0.023	0.014 0.024
	Usual-KZ2	0.016 0.022	0.015 0.023	0.014 0.024
	Limitado	0.016 0.022	0.015 0.023	0.014 0.024
	Generalizado	0.010 0.018	0.009 0.019	0.008 0.020
VPN	Usual	0.980 0.985	0.979 0.986	0.978 0.987
	Usual-Balanceado	0.980 0.985	0.979 0.986	0.978 0.987
	Usual-KZ1	0.980 0.985	0.979 0.986	0.978 0.987
	Usual-KZ2	0.980 0.985	0.979 0.986	0.978 0.987
	Limitado	0.980 0.985	0.979 0.986	0.978 0.987
	Generalizado	0.981 0.988	0.981 0.989	0.980 0.989
ACC	Usual	0.433 0.580	0.423 0.589	0.398 0.610
	Usual-Balanceado	0.433 0.580	0.422 0.589	0.394 0.610
	Usual-KZ1	0.433 0.582	0.423 0.589	0.398 0.610
	Usual-KZ2	0.433 0.580	0.423 0.589	0.398 0.610
	Limitado	0.433 0.580	0.423 0.589	0.398 0.610
	Generalizado	0.431 0.601	0.418 0.629	0.412 0.643
MCC	Usual	-0.013 0.0287	-0.017 0.034	-0.025 0.037
	Usual-Balanceado	-0.013 0.028	-0.017 0.034	-0.025 0.037
	Usual-KZ1	-0.015 0.028	-0.018 0.034	-0.025 0.037
	Usual-KZ2	-0.013 0.028	-0.017 0.034	-0.025 0.037
	Limitado	-0.013 0.028	-0.017 0.034	-0.025 0.037
	Generalizado	-0.022 0.022	-0.025 0.024	-0.031 0.028

em que os dados foram gerados através do modelo logito generalizado, respectivamente.

As medidas de qualidade de ajuste apresentadas indicam que o modelo logito limitado obteve um desempenho superior ao modelo logito usual quando as amostras foram geradas através do modelo logito generalizado.

Na Figura 7.46 encontramos os boxplots da sensibilidade para cada modelo de

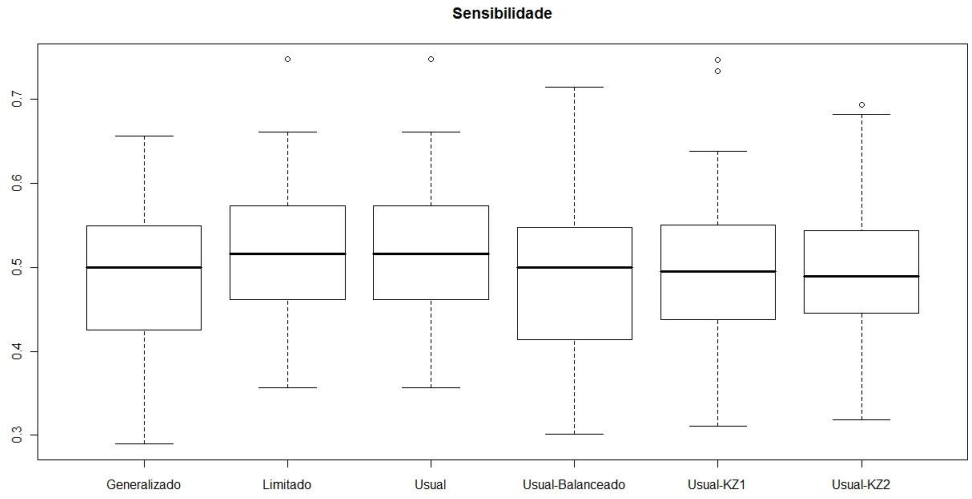

Figura 7.31: Modelo Logito Limitado - Sensibilidade.

classificação analisado para os dados gerados de acordo com o modelo logito generalizado.

Notamos que a mediana desta medida está muito próxima para todos os modelos, no entanto, o modelo logito usual foi o que apresentou uma maior quantidade de pontos atípicos e consequentemente uma maior dipersão.

Na Figura 7.47 apresentamos os boxplots para especificidade calculada através dos modelos estudados para os dados gerados através do modelo logito generalizado. O modelo logito generalizado foi o que apresentou o melhor desempenho, enquanto que o modelo logito usual apresentou o pior desempenho.

Os boxplots para o valor preditivo positivo são encontrados na Figura 7.48 para os dados gerados através do modelo logito generalizado. O modelo logito generalizado foi o que apresentou o melhor desempenho considerando esta medida, em contrapartida, o modelo logito usual construído através de amostras completas apresentou o pior desempenho.

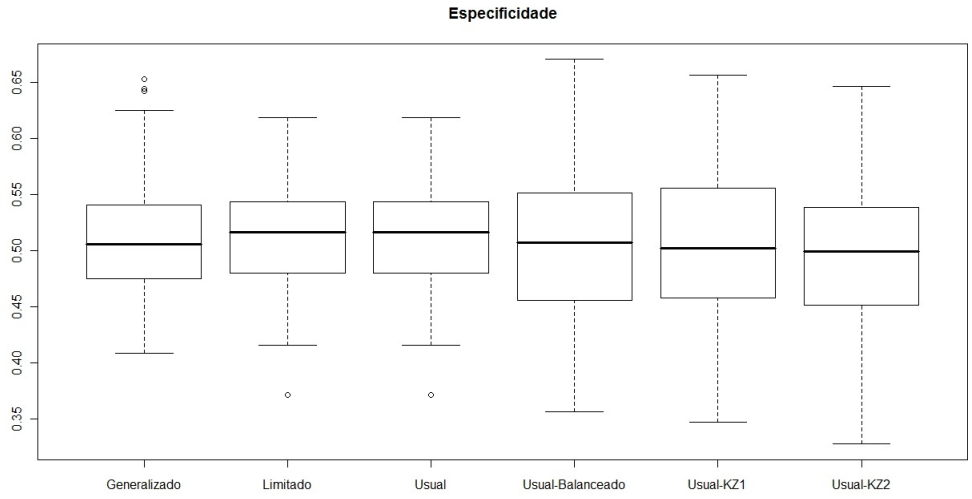

Figura 7.32: Modelo Logito Limitado - Especificidade.

Na Figura 7.49 estão os boxplots do valor preditivo negativo para os diversos modelos analisados para os dados gerados através do modelo logito generalizado. O modelo logito usual apresentou diversos pontos atípicos e uma maior dispersão quando comparado aos outros modelos.

Na Figura 7.50 apresentamos os boxplots para a acurácia nos diversos modelos para os dados gerados através do modelo logito generalizado. O modelo gerador dos dados foi o que apresentou o melhor desempenho enquanto que o modelo logito usual, ajustado em amostras completas, foi o pior modelo considerando esta medida.

Os boxplots para o coeficiente de correlação de mathews para os dados gerados de acordo com o modelo logito generalizado estão na Figura 7.51. O modelo logito apresentou o melhor desempenho considerando esta medida enquanto que o modelo logito usual foi o que apresentou o pior desempenho.

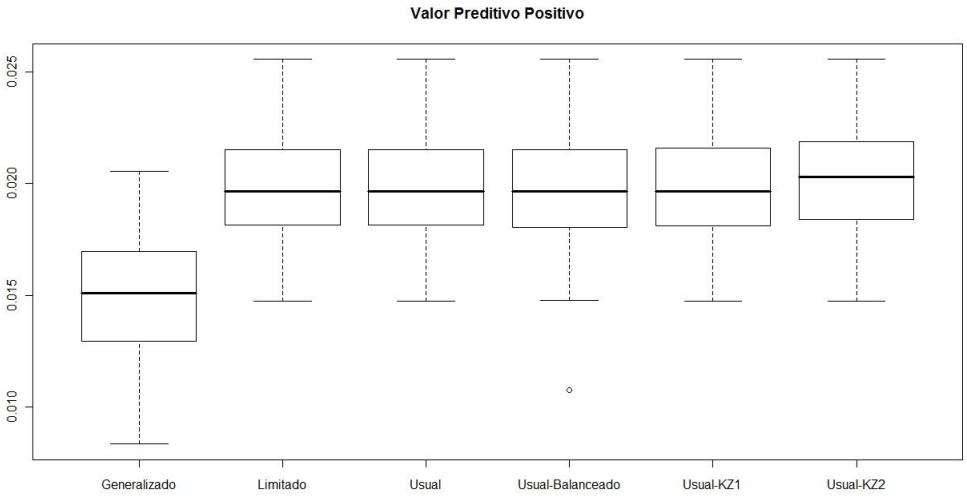

Figura 7.33: Modelo Logito Limitado - Valor Preditivo Positivo.

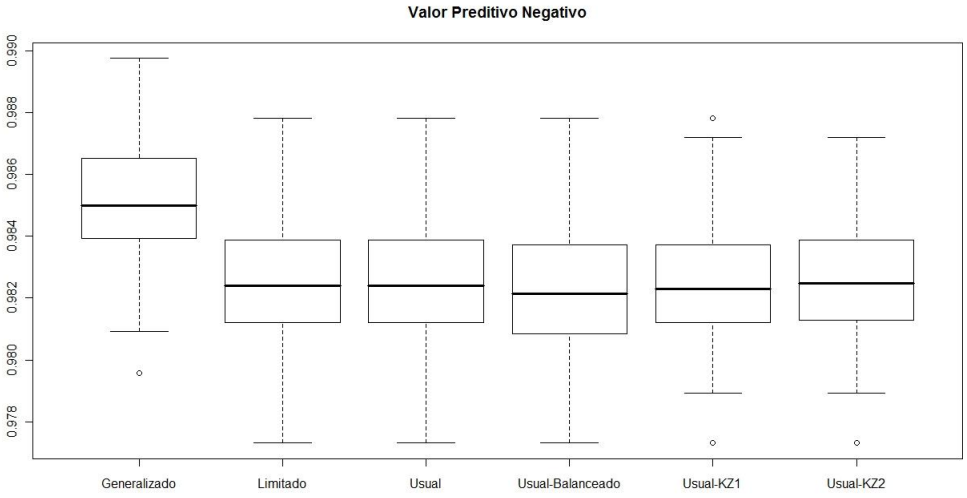

Figura 7.34: Modelo Logito Limitado - Valor Preditivo Negativo.

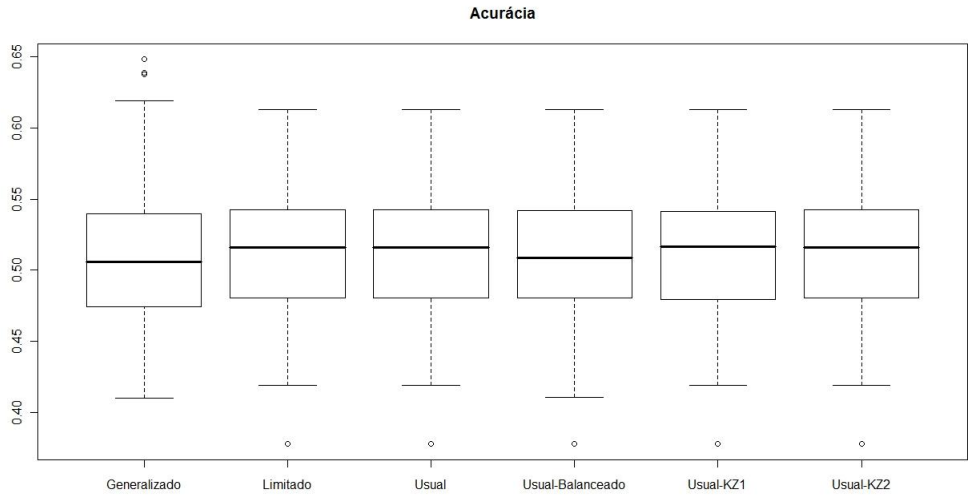

Figura 7.35: Modelo Logito Limitado - Acurácia.

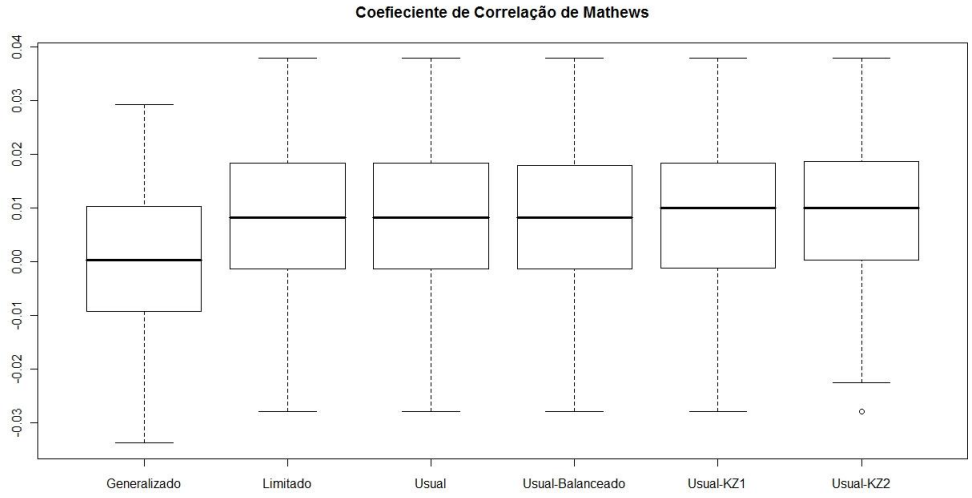

Figura 7.36: Modelo Logito Limitado - Coeficiente de Correlação de Mathews.

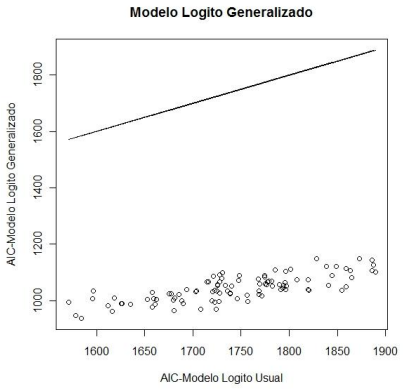

Figura 7.37: AIC - Modelo logito generalizado x Modelo logito usual.

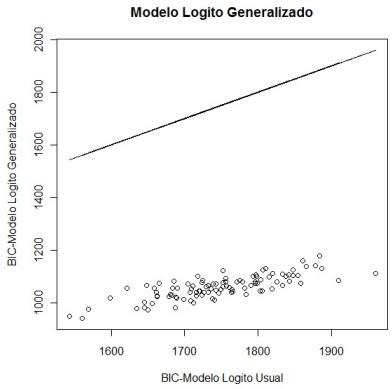

Figura 7.38: BIC - Modelo logito generalizado x Modelo logito usual.

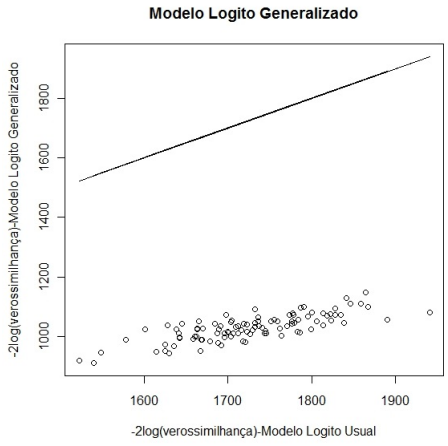

Figura 7.39: -2log(verossimilhança) - Modelo logito generalizado x Modelo logito usual.

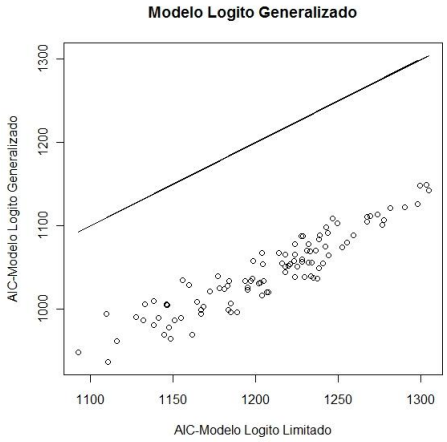

Figura 7.40: AIC - Modelo logito generalizado x Modelo logito limitado.

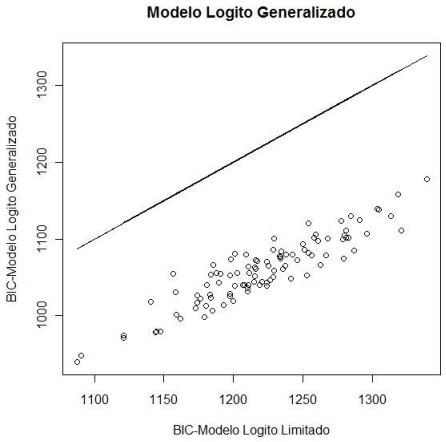

Figura 7.41: BIC - Modelo logito generalizado x Modelo logito limitado.

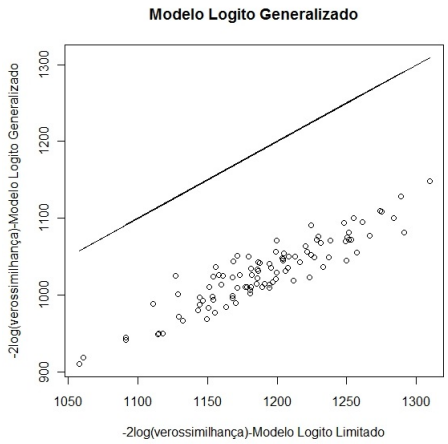

Figura 7.42: -2log(verossimilhança) - Modelo logito generalizado x Modelo logito limitado.

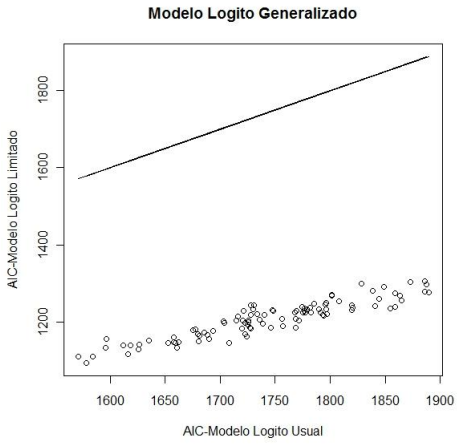

Figura 7.43: AIC - Modelo logito usual x Modelo logito limitado.

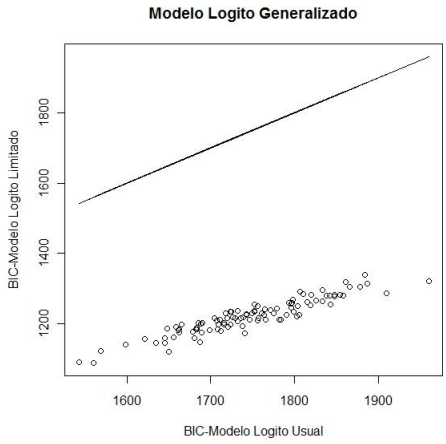

Figura 7.44: BIC - Modelo logito usual x Modelo logito limitado.

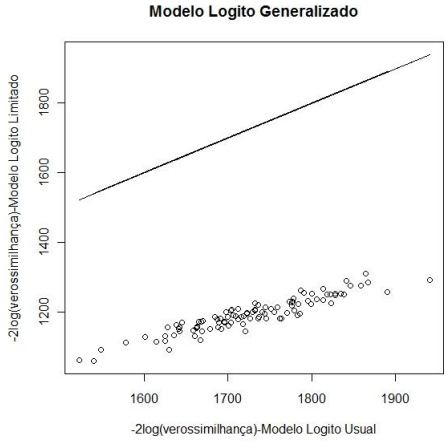

Figura 7.45: -2log(verossimilhança) - Modelo logito usual x Modelo logito limitado.

Tabela 7.3: Medidas Preditivas-Modelo Logito Generalizado.

	Modelos	90%	95%	99%
SEN	Usual	0.784 1.000	0.688 1.000	0.507 1.000
	Usual-Balanceado	0.946 1.000	0.929 1.000	0.908 1.000
	Usual-KZ1	0.946 1.000	0.929 1.000	0.908 1.000
	Usual-KZ2	0.946 1.000	0.929 1.000	0.908 1.000
	Limitado	0.946 1.000	0.929 1.000	0.908 1.000
	Generalizado	0.945 1.000	0.936 1.000	0.898 1.000
ESP	Usual	0.901 0.974	0.893 0.977	0.880 0.980
	Usual-Balanceado	0.954 0.966	0.953 0.966	0.951 0.968
	Usual-KZ1	0.953 0.966	0.952 0.966	0.951 0.968
	Usual-KZ2	0.953 0.966	0.952 0.966	0.951 0.968
	Limitado	0.952 0.966	0.951 0.966	0.951 0.968
	Generalizado	0.962 0.973	0.961 0.974	0.959 0.975
VPP	Usual	0.122 0.301	0.116 0.310	0.098 0.329
	Usual-Balanceado	0.219 0.285	0.206 0.291	0.200 0.307
	Usual-KZ1	0.210 0.281	0.206 0.291	0.200 0.302
	Usual-KZ2	0.218 0.281	0.212 0.291	0.204 0.297
	Limitado	0.217 0.281	0.210 0.290	0.203 0.295
	Generalizado	0.254 0.336	0.250 0.345	0.238 0.358
VPN	Usual	0.997 1.000	0.995 1.000	0.993 1.000
	Usual-Balanceado	0.999 1.000	0.999 1.000	0.998 1.000
	Usual-KZ1	0.999 1.000	0.999 1.000	0.998 1.000
	Usual-KZ2	0.999 1.000	0.999 1.000	0.998 1.000
	Limitado	0.999 1.000	0.999 1.000	0.998 1.000
	Generalizado	0.999 1.000	0.999 1.000	0.998 1.000
ACC	Usual	0.902 0.972	0.894 0.973	0.882 0.975
	Usual-Balanceado	0.953 0.966	0.953 0.966	0.951 0.968
	Usual-KZ1	0.952 0.966	0.953 0.966	0.951 0.968
	Usual-KZ2	0.954 0.966	0.953 0.966	0.951 0.968
	Limitado	0.953 0.966	0.953 0.966	0.951 0.968
	Generalizado	0.962 0.973	0.961 0.974	0.960 0.975
MCC	Usual	0.327 0.501	0.314 0.514	0.289 0.526
	Usual-Balanceado	0.449 0.519	0.446 0.524	0.439 0.531
	Usual-KZ1	0.449 0.519	0.446 0.524	0.439 0.531
	Usual-KZ2	0.449 0.519	0.446 0.524	0.434 0.531
	Limitado	0.449 0.519	0.446 0.524	0.434 0.531
	Generalizado	0.487 0.565	0.485 0.574	0.472 0.581

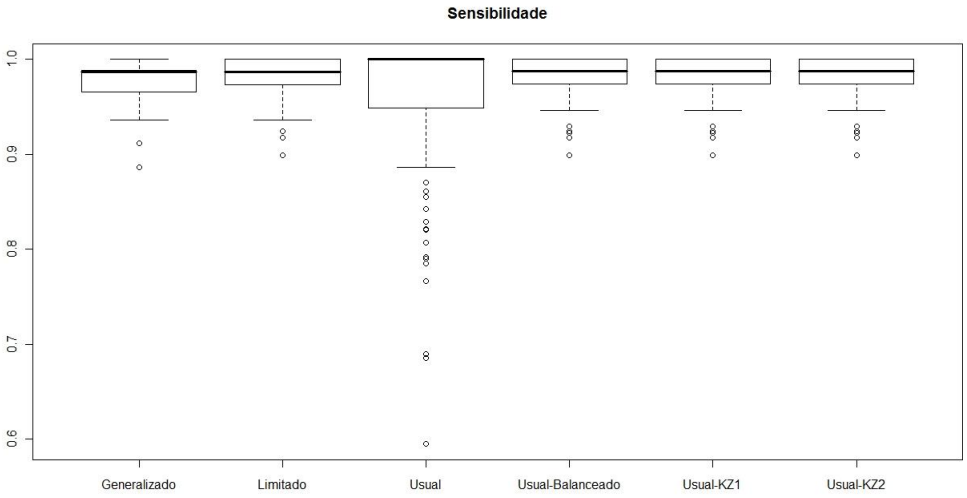

Figura 7.46: Modelo Logito Limitado - Sensibilidade.

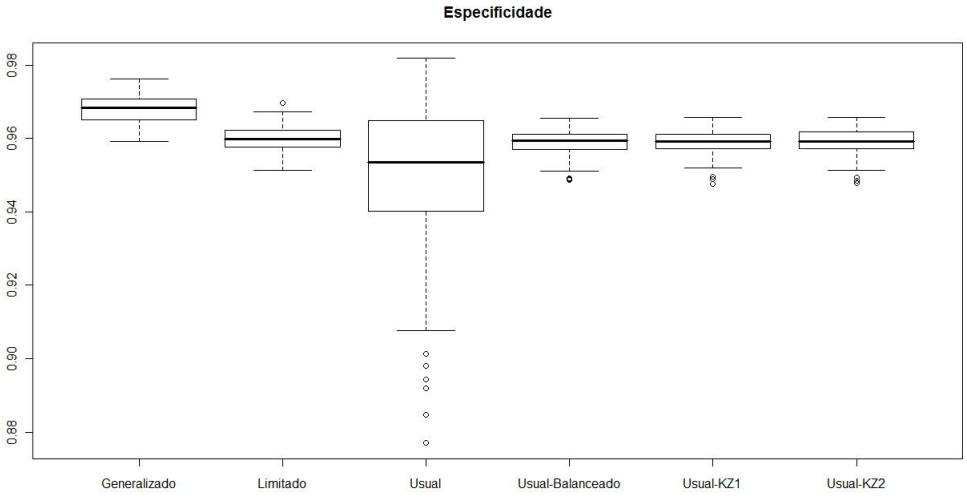

Figura 7.47: Modelo Logito Limitado - Especificidade.

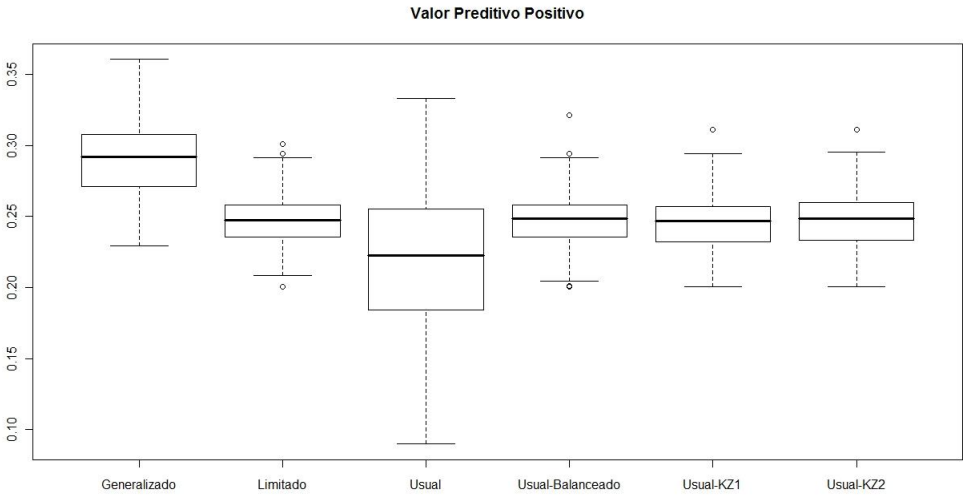

Figura 7.48: Modelo Logito Limitado - Valor Preditivo Positivo.

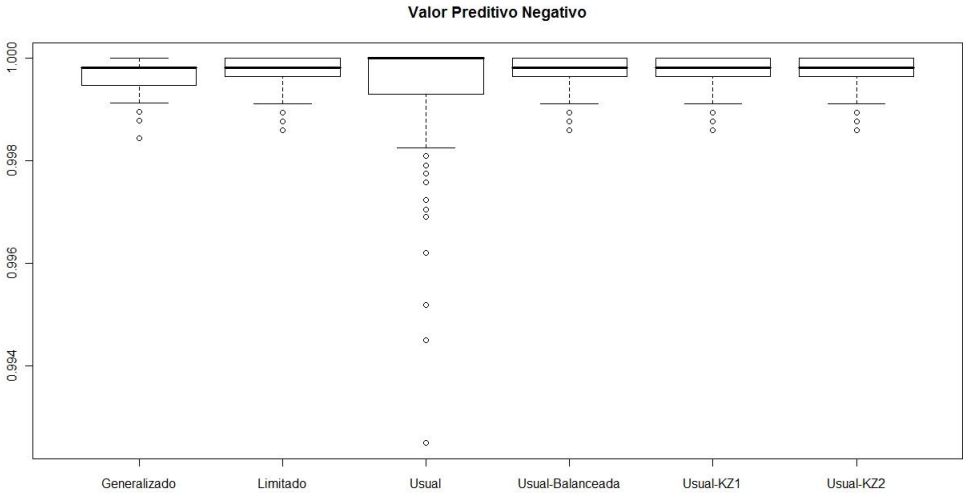

Figura 7.49: Modelo Logito Limitado - Valor Preditivo Negativo.

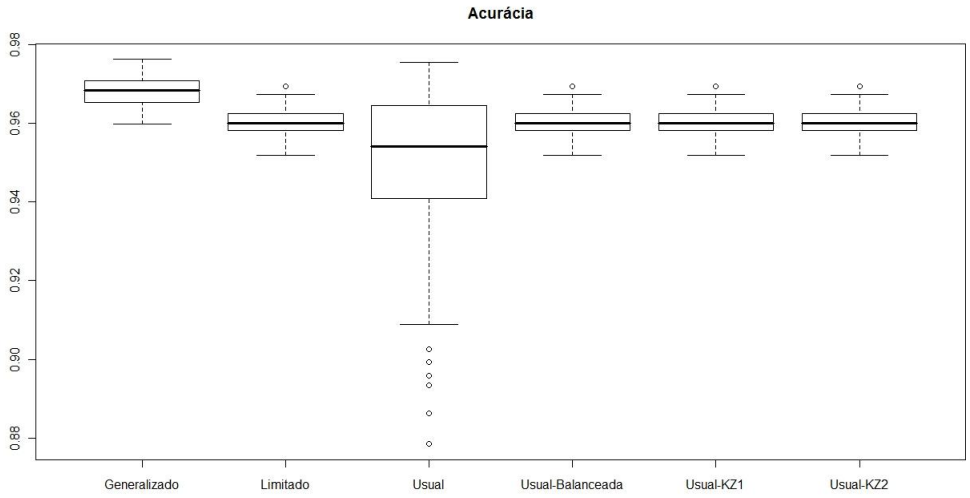

Figura 7.50: Modelo Logito Limitado - Acurácia.

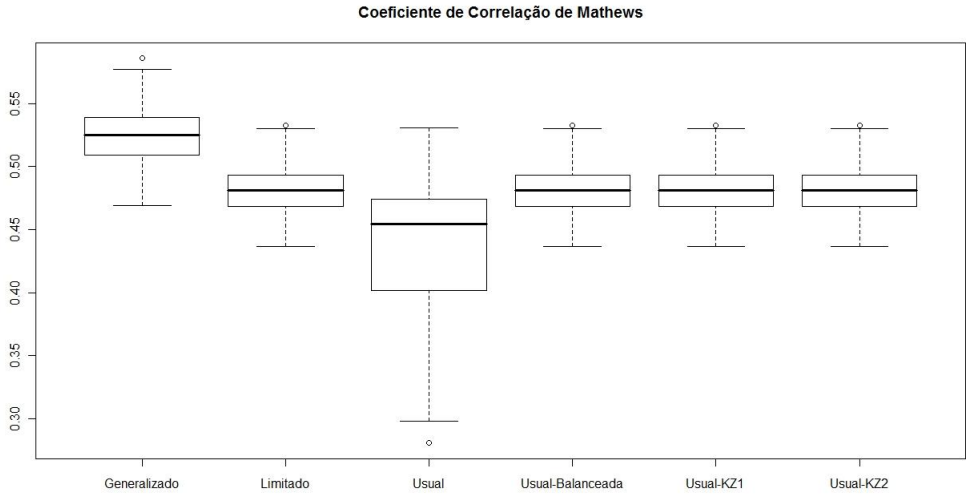

Figura 7.51: Modelo Logito Limitado - Coeficiente de Correlação de Mathews.

Capítulo 8

Análise do Modelo Logito com resposta de origem

Neste Capítulo apresentamos um estudo de simulações realizado com o intuito de comparar a qualidade de ajuste dos modelos logito com resposta de origem e usual considerando diferentes prevalências e tamanhos amostrais. O objetivo central é analisar o impacto da prevalência e do tamanho amostral na qualidade do ajuste de ambos os modelos.

As distribuições normal, exponencial e lognormal foram utilizadas como distribuições de origem. Para cada distribuição consideramos três tamanhos amostrais n=100, n=500 e n=5000 e para cada tamanho amostral consideramos duas prevalências.

Utilizamos o vício, o erro quadrático médio e o erro absoluto médio para analisar a qualidade das estimativas de ambos os modelos. Além disso, calculamos intervalos de confiança assintóticos para os parâmetros e analisamos a probabilidade de cobertura e a amplitude média dos intervalos construídos através dos modelos estudados. Para verificar a convergência das estimativas de ambos os modelos calculamos a

razão das mesmas e construímos intervalos de confiança empíricos para estas razões. Analisamos também as estimativas da razão das chances de cada modelo através de intervalos de confiança empíricos.

8.1 Distribuição de origem Normal

Nesta seção analisamos o desempenho do modelo logito com resposta de origem normal através de dados artificiais. Na geração dos dados utilizamos três variáveis explicativas com distribuição de Bernoulli, X_{i1}, X_{i2} e X_{i3}. A geração de dados foi repetida para três tamanhos amostrais, n=100, n=500 e n=5000. Foram geradas 1000 amostras de tamanho n com variável resposta $R_i \sim N(\mu_i, \sigma^2)$, com $\sigma = 100$ e $\mu_i = \sigma\phi^{-1}\left[g^{-1}\left(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i1} + \beta_3 xi3\right)\right] + C$, $i = 1, \cdots, n$. Os valores atribuídos para o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ para a geração de μ_i foram, $\beta_0 = -4.5$, $\beta_1 = 1.0$, $\beta_2 = 1.2$ e $\beta_3 = 2.0$. O ponto de corte considerado foi $C = 110$.

Nas amostras de tamanho 100, geramos covariáveis $X_{i1} \sim Bernoulli(0, 2)$, $X_{i2} \sim Bernoulli(0, 2)$ e $X_{i2} \sim Bernoulli(0, 2)$ para obter a prevalência de 0,1. Para obter a prevalência de 0,2 simulamos as covariáveis $X_{i1} \sim Bernoulli(0, 5)$, $X_{i2} \sim Bernoulli(0, 5)$ e $X_{i2} \sim Bernoulli(0, 5)$. Nas amostras de tamanho 500 geramos covariáveis $X_{i1} \sim Bernoulli(0, 3)$, $X_{i2} \sim Bernoulli(0, 3)$ e $X_{i2} \sim Bernoulli(0, 3)$ para obter a prevalência de 0,05 e para alcançar a prevalência de 0,1 simulamos as covariáveis $X_{i1} \sim Bernoulli(0, 5)$, $X_{i2} \sim Bernoulli(0, 5)$ e $X_{i2} \sim Bernoulli(0, 5)$. Nas amostras de tamanho 5000 analisamos as prevalências de 0,01 e de 0,1. Na primeira situação geramos covariáveis $X_{i1} \sim Bernoulli(0, 1)$, $X_{i2} \sim Bernoulli(0, 1)$ e $X_{i2} \sim Bernoulli(0, 1)$ e para obter a prevalência de 0,1 geramos covariáveis $X_{i1} \sim Bernoulli(0, 4)$, $X_{i2} \sim Bernoulli(0, 4)$ e $X_{i2} \sim Bernoulli(0, 4)$.

É importante ressaltar que as prevalências foram escolhidas de acordo com o

tamanho amostral, isso porque as covariáveis geradas possuem distribuição de Bernoulli e como sabemos no ajuste de um modelo de regressão logística é necessário no mínimo uma observação representando evento e uma representando não evento para cada combinação das classes das covariáveis.

Tabela 8.1: Qualidade do ajuste- Distribuição de origem Normal - n=100.

		Modelo logístico usual				Modelo resposta de origem			
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,1	β_0	-4,503	80,560	4,824	-9,003	-0,336	0,547	0,572	-4,836
	β_1	0,354	6,004	0,937	1,354	0,031	0,206	0,360	1,031
	β_2	0,709	12,864	1,283	1,909	0,050	0,199	0,350	1,250
	β_3	3,095	55,515	3,668	5,095	0,058	0,248	0,395	2,058
p=0,2	β_0	-4,505	81,576	4,977	-9,005	-0,277	0,652	0,620	-4,777
	β_1	0,579	9,180	1,033	1,579	0,061	0,205	0,350	1,061
	β_2	0,617	9,633	1,085	1,817	0,048	0,210	0,362	1,248
	β_3	3,156	54,278	3,704	5,156	0,060	0,259	0,399	2,060

Na Tabela 8.1 encontramos o vício amostral, o erro quadrático médio (EQM), o erro absoluto médio (EAM) e a média das estimativas dos parâmetros para as amostras de tamanho n=100 e prevalência p=0,1 e p=0,2. Notamos que o vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com resposta de origem são inferiores as mesmas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual. Ou seja, há evidências de que o modelo logito com resposta de origem produziu um ajuste com qualidade superior ao modelo logito usual.

Na Tabela 8.2 encontramos intervalos empíricos para a razão entre as estimativas do modelo logito limitado e do modelo logito com resposta de origem considerando amostras de tamanho 100 e a prevalência de 0,1 e 0,2. Observamos que os modelos apresentam estimativas divergentes já que a amplitude dos intervalos é elevada.

Nas Tabela 8.3 e 8.4 encontramos os intervalos empíricos para a razão das chances dos modelos logito usual e logito com resposta de origem, respectivamente, para as

Tabela 8.2: Intervalos de Confiança Empiricos da razão das estimativas - Distribuição de origem Normal - n=100.

	Parâmetros	90%	95%	99%
p=0,1	β_0	0.776 4.787	0.748 5.000	0.655 6.722
	β_1	-0.314 3.040	-0.932 5.789	-6.249 14.609
	β_2	0.003 3.031	-0.265 9.801	-2.253 14.093
	β_3	0.508 8.820	0.379 9.827	0.041 12.003
p=0,2	β_0	0.748 4.820	0.711 5.386	0.635 7.487
	β_1	0.113 2.608	-0.152 9.218	-2.100 15.061
	β_2	0.286 2.547	0.062 8.676	-0.669 14.266
	β_3	0.521 8.753	0.391 10.112	0.204 12.265

três covariáveis.

Sabemos que para X_1 o valor real da razão das chances é de $exp(\beta_1) = 2,718$, já para X_2 a razão das chances é de $exp(\beta_2) = 3.320$ e para X_3 temos que a razão das chances é de $exp(\beta_3) = 7.389$. Notamos que os intervalos empíricos para as estimativas da razão das chances do modelo logito usual possuem amplitude bastante elevada indicando que as estimativas deste modelo são muito imprecisas. Já o modelo logito com resposta de origem apresenta intervalos empíricos com amplitude menor indicando uma maior precisão nas estimativas.

Tabela 8.3: Intervalos de Confiança Empíricos da razão das chances - Modelo Logito Usual - Distribuição de origem Normal - n=100.

	Parâmetros	90%	95%	99%
p=0,1	β_1	0.818 14.134	0.623 18.103	3.479e-01 2.761e+08
	β_2	1.001 19.441	7.966e-01 2.110e+08	4.671e-01 4.614e+08
	β_3	2.465 2.583e+08	1.925 3.094e+08	1.071 5.893e+08
p=0,2	β_1	1.035 12.862	8.628e-01 6.730e+07	5.804e-01 1.393e+08
	β_2	1.253 15.817	1.038 9.362e+07	7.474e-01 1.996e+08
	β_3	2.406 1.419e+08	1.972 1.625e+08	1.281 3.113e+08

Nas Tabelas 8.5 e 8.6 encontramos a probabilidade de cobertura e a amplitude média dos intervalos de confiança assintóticos para os parâmetros dos modelos logito usual e logito com resposta de origem normal para amostras de tamanho 100 a

Tabela 8.4: Intervalos de Confiança Empíricos da razão das chances - Modelo logito com resposta de origem - Distribuição de origem Normal-n=100.

	Parâmetros	90%	95%	99%
p=0,1	β_1	1.361 5.920	1.210 7.005	0.934 10.521
	β_2	1.685 7.459	1.497 8.558	1.163 12.108
	β_3	3.641 18.834	3.259 23.708	2.586 30.708
p=0,2	β_1	1.437 6.302	1.276 7.125	0.873 11.280
	β_2	1.702 7.705	1.506 9.057	1.140 12.864
	β_3	3.626 18.494	3.013 23.568	2.315 30.796

prevalências de 0,1 e 0,2. Observamos que os intervalos de confiança de ambos os modelos atingem o nível de confiança estabelecido, no entanto, a amplitude média dos intervalos de confiança assintóticos dos parâmetros do modelo logito com resposta de origem é inferior a amplitude média dos intervalos de confiança dos parâmetros do modelo logito usual.

Tabela 8.5: Probabilidade de Cobertura - Distribuição de origem Normal - n=100.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,1	β_0	0.943	0.965	0.988	0.903	0.952	0.987
	β_1	0.926	0.973	0.998	0.916	0.957	0.990
	β_2	0.924	0.968	0.998	0.908	0.956	0.991
	β_3	0.968	0.994	0.990	0.909	0.958	0.989
p=0,2	β_0	0.905	0.945	0.982	0.907	0.956	0.994
	β_1	0.916	0.964	0.991	0.888	0.955	0.992
	β_2	0.891	0.952	0.987	0.909	0.956	0.994
	β_3	0.976	0.953	0.992	0.908	0.958	0.990

Na Tabela 8.7 encontramos o vício amostral, o erro quadrático médio (EQM), o erro absoluto médio (EAM) e a média das estimativas dos parâmetros nas amostras de tamanho n=500 e prevalência p=0,05 e p=0,10. Notamos que vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com resposta de origem são inferiores a estas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual.

Tabela 8.6: Amplitude Média - Distribuição de origem Normal - n=100.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,1	β_0	1358.629	1623.727	2129.071	2.007	2.399	3.145
	β_1	142.3367	170.109	223.052	1.432	1.712	2.245
	β_2	55.737	66.612	87.3441	1.652	1.975	2.589
	β_3	1233.317	1473.964	1932.698	1.596	1.908	2.502
p=0,2	β_0	462.286	552.489	724.436	2.124	2.538	3.327
	β_1	25.408	30.365	39.816	1.363	1.629	2.136
	β_2	7.431	8.881	11.644	1.319	1.576	2.067
	β_3	432.832	517.288	678.280	1.548	1.850	2.426

Tabela 8.7: Qualidade do ajuste - Distribuição de origem Normal - n=500.

		Modelo logístico usual				Modelo resposta de origem			
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,05	β_0	-0,708	8,859	1,052	-5,208	-0,038	0,084	0,231	-4,538
	β_1	0,072	0,584	0,438	1,072	0,011	0,046	0,172	1,011
	β_2	-0,111	0,882	0,452	-1,311	-0,002	0,048	0,176	-1,202
	β_3	0,630	8,452	0,989	2,630	0,022	0,061	0,198	2,022
p=0,10	β_0	-0,067	3,221	0,704	-4,567	0,013	0,010	0,263	-4,367
	β_1	0,089	0,215	0,351	1,089	0,014	0,045	0,169	1,014
	β_2	-0,276	0,194	0,340	-1,276	-0,227	0,049	0,176	-1,227
	β_3	0,301	3,326	0,628	2,301	0,030	0,060	0,195	2,030

Na Tabela 8.8 encontramos intervalos de confiança empíricos das razão entre as estimativas do modelo do logito usual e do modelo logito com resposta de origem. Notamos que a amplitude destes intervalos diminui a medida que aumentamos o tamanho a amostral e a prevalência.

Nas Tabelas 8.9 e 8.11 encontramos intervalos de confiança empíricos para a razão da chances dos modelos logito usual e logito com resposta de origem normal, respectivamente considerando amostras de tamanho 500. A precisão dos intervalos para a razão das chances do modelo logito com resposta de origem apresentam amplitude inferior quando comparamos a amplitude do modelo logito usual.

Nas Tabelas 8.11 e 8.12 encontramos a probabilidade de cobertura e a amplitude média dos intervalos de confiança assintóticos dos modelos logito usual e logito com

Tabela 8.8: Intervalos de Confiança Empiricos da razão das estimativas - Distribuição de origem Normal - n=500.

	Medida	90%	95%	99%
p=0,05	β_0	0.841 1.353	0.810 4.266	0.760 4.669
	β_1	0.290 1.855	0.147 2.060	-0.157 2.491
	β_2	0.290 1.855	0.147 2.060	-0.157 2.491
	β_3	0.639 1.757	0.585 8.011	0.496 8.911
p=0,10	β_0	0.803 1.280	0.780 1.338	0.735 4.324
	β_1	0.470 1.817	0.344 2.020	0.081 2.468
	β_2	0.585 1.590	0.514 1.667	0.353 2.018
	β_3	0.686 1.584	0.629 1.791	0.551 7.978

Tabela 8.9: Intervalos de Confiança Empiricos da razão das chances - Distribuição de origem Normal - n=500.

	Medida	90%	95%	99%
p=0,05	β_1	1.288 7.365	1.116 9.673	0.856 14.946
	β_2	0.110 0.614	0.072 0.706	0.038 0.877
	β_3	3.349 31.556	3.0243e+00 7.518e+07	2.592 1.076e+08
p=0,1	β_1	1.514 6.427	1.353 7.985	1.076 11.948
	β_2	0.123 0.527	0.102 0.568	0.071 0.697
	β_3	3.744 31.450	3.349 36.630	2.809 5.309e+07

Tabela 8.10: Intervalos de Confiança Empiricos da razão das chances - Modelo logito com resposta de origem - Distribuição de origem Normal - n=500.

	Medida	90%	95%	90%
p=0,05	β_1	1.941 3.913	1.846 4.303	1.618 4.765
	β_2	0.206 0.425	0.196 0.451	0.173 0.535
	β_3	5.06374 11.292	4.669 12.596	4.157 14.853
p=0,1	β_1	1.942 3.854	1.813 4.169	1.601 5.000
	β_2	0.201 0.415	0.185 0.443	0.163 0.472
	β_3	5.134 11.395	4.808 12.283	4.275 15.130

resposta de origem normal para amostras de tamanho 500. Os intervalos assintóticos para os parâmetros de ambos os modelos atingem os níveis críticos de confiança estabelecidos, no entanto, a amplitude média dos intervalos de confiança dos modelos logito com resposta de origem é inferior, ou seja, as estimativas produzidas por este modelo são mais precisas.

Tabela 8.11: Probabilidade de Cobertura - Distribuição de origem Normal - n=500.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,05	β_0	0.888	0.952	0.992	0.897	0.948	0.990
	β_1	0.904	0.952	0.992	0.883	0.945	0.995
	β_2	0.876	0.945	0.990	0.895	0.946	0.988
	β_3	0.902	0.962	0.998	0.909	0.955	0.990
p=0,10	β_0	0.911	0.958	0.993	0.901	0.950	0.992
	β_1	0.890	0.946	0.992	0.884	0.945	0.990
	β_2	0.891	0.950	0.989	0.889	0.949	0.987
	β_3	0.899	0.959	0.990	0.893	0.952	0.990

Tabela 8.12: Amplitude Média - Distribuição de origem Normal - n=500.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,05	β_0	1.570	1.877	2.461	0.835	0.997	1.308
	β_1	1.312	1.568	2.056	0.698	0.835	1.094
	β_2	1.312	1.568	2.056	0.700	0.837	1.097
	β_3	1.414	1.690	2.216	0.727	0.868	1.138
p=0,10	β_0	1.621	1.938	2.541	0.890	1.064	1.395
	β_1	1.101	1.316	1.725	0.623	0.745	0.977
	β_2	1.045	1.248	1.637	0.648	0.775	1.016
	β_3	1.384	1.654	2.169	0.689	0.824	1.080

Na Tabela 8.13 encontramos o vício amostral, o erro quadrático médio (EQM), e erro absoluto médio (EAM) e a média das estimativas dos parâmetros nas amostras de tamanho n=5000 com prevalências de p=0,01 e p=0,10. Notamos que o vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com resposta de origem são inferiores a estas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual.

Como era esperado a medida que aumentamos o tamanho amostral o vício, o EQM e o EAM diminuem em ambos os modelos.

Na Tabela 8.14 encontramos intervalos de confiança empíricos para a razão entre as estimativas do modelo logito usual e do modelo logito com resposta de origem normal para amostras de tamanho 5000.

Tabela 8.13: Qualidade do ajuste - Distribuição de origem Normal-n=5000.

		Modelo logístico usual			Modelo resposta de origem				
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,01	β_0	-0,053	0,020	0,110	-4,553	-0,016	0,005	0,057	-4,516
	β_1	-0,025	0,136	0,283	0,974	0,003	0,024	0,124	1,003
	β_2	-0,008	0,069	0,206	1,191	-0,004	0,011	0,086	1,195
	β_3	-0,010	0,070	0,210	1,989	-0,006	0,018	0,108	1,993
p=0,1	β_0	-0,018	0,023	0,121	-4,518	-0,009	0,006	0,065	-4,509
	β_1	0,006	0,011	0,085	1,006	0,007	0,0035	0,0469	1,007
	β_2	0,001	0,012	0,087	1,201	0,002	0,004	0,054	1,202
	β_3	0,015	0,019	0,109	2,015	0,005	0,004	0,053	2,005

Notamos que a medida que aumentamos o tamanho amostral as estimativas de ambos os modelos convergem.

Tabela 8.14: Intervalos de Confiança Empiricos da razão das estimativas - Distribuição de origem Normal - n=5000.

	Medida	90%	95%	99%
p=0,01	β_0	0.966 1.056	0.960 1.063	0.945 1.078
	β_1	0.339 1.481	0.172 1.596	-0.153 1.751
	β_2	0.664 1.324	0.573 1.392	0.391 1.451
	β_3	0.810 1.189	0.776 1.224	0.678 1.282
p=0,10	β_0	0.953 1.049	0.945 1.057	0.928 1.074
	β_1	0.860 1.142	0.831 1.169	0.766 1.217
	β_2	0.885 1.110	0.869 1.142	0.831 1.156
	β_3	0.915 1.100	0.894 1.113	0.845 1.145

Na Tabelas 8.15 e8.16 apresentamos os intervalos de confiança empíricos para a razão das chances dos modelos logito usual e logito com resposta de origem normal para amostras de tamanho 5000, respectivamente. A amplitude dos intervalos para o modelo logito com resposta de origem é inferior a amplitude dos intervalos para o modelo logito usual indicando que a precisão das estimativas produzidas pelo modelo logito com resposta de origem é superior. Além disso, comparando os resultados obtidos através dos diversos tamanhos amostrais estudados notamos que em amostras maiores a precisão das estimativas é superior.

Tabela 8.15: Intervalos de Confiança Empiricos da razão das chances - Modelo logito usual - Distribuição de origem Normal - n=5000.

	Medida	90%	95%	99%
p=0,01	β_1	1.319 4.597	1.156 5.171	0.880 5.912
	β_2	2.167 5.069	1.916 5.484	1.514 6.200
	β_3	4.742 11.218	4.288 12.152	3.511 13.892
p=0,10	β_1	2.311 3.295	2.239 3.397	2.113 3.587
	β_2	5.856 9.068	5.575 9.481	5.193 10.168
	β_3	6.022 9.375	5.781 9.736	5.185 10.790

Tabela 8.16: Intervalos de Confiança Empiricos da razão das chances - Modelo logito com resposta de origem - Distribuição de origem Normal-n=5000.

	Medida	90%	95%	99%
p=0,01	β_1	2.089 3.504	2.015 3.691	1.811 4.133
	β_2	2.752 3.947	2.674 4.069	2.497 4.342
	β_3	5.856 9.068	5.575 9.481	5.193 10.168
p=0,10	β_1	2.492 3.024	2.451 3.072	2.350 3.187
	β_2	2.973 3.736	2.904 3.823	2.815 3.96
	β_3	6.682 8.311	6.512 8.493	6.232 8.864

Na Tabela 8.17 apresentamos a probabilidade de cobertura para os intervalos de confiança empíricos para os parâmetros do modelo logito usual e do modelo logito com resposta de origem normal para amostras de tamanho 5000. Ambos os intervalos atingem o nível nominal de confiança porém, a amplitude média dos intervalos de confiança para o modelo logito com resposta de origem é inferior.

Tabela 8.17: Probabilidade de Cobertura - Distribuição de origem Normal - n=5000.

	Modelo logístico usual				Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,01	β_0	0.917	0.962	0.996	0.916	0.958	0.990
	β_1	0.893	0.949	0.992	0.899	0.947	0.991
	β_2	0.894	0.954	0.995	0.900	0.947	0.990
	β_3	0.899	0.963	0.994	0.911	0.952	0.987
p=0,10	β_0	0.893	0.943	0.989	0.916	0.958	0.990
	β_1	0.910	0.958	0.992	0.913	0.962	0.990
	β_2	0.884	0.945	0.988	0.908	0.948	0.988
	β_3	0.900	0.941	0.986	0.922	0.971	0.994

Tabela 8.18: Amplitude Média - Distribuição de origem Normal - n=5000.

	Modelo logístico usual				Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,01	β_0	0.472	0.564	0.740	0.240	0.287	0.376
	β_1	1.203	1.438	1.885	0.493	0.589	0.773
	β_2	1.113	1.330	1.744	0.483	0.577	0.757
	β_3	0.869	1.038	1.361	0.447	0.534	0.700
p=0,10	β_0	0.471	0.562	0.737	0.276	0.330	0.433
	β_1	0.346	0.413	0.542	0.197	0.236	0.309
	β_2	0.334	0.400	0.524	0.198	0.237	0.311
	β_3	0.375	0.448	0.588	0.211	0.252	0.331

8.2 Distribuição de origem Exponencial

Nesta seção analisamos o desempenho do modelo logístico com resposta de origem exponencial através de dados artificiais. Na geração dos dados utilizamos três variáveis explicativas com distribuição de Bernoulli, X_{i1}, X_{i2} e X_{i3}. A geração de dados foi repetida para três tamanhos amostrais, n=100, n=500 e n=5000. Foram geradas 1000 amostras de tamanho n com variável resposta $R_i \sim Exp(\lambda_i)$, com $\lambda_i = -\frac{log\left[g^{-1}(\beta_0+beta_1x_{i1}+\beta_2x_{i2}+\beta_2x_{i3})\right]}{C}$, $i = 1, \cdots, n$. Os valores atribuídos para o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ para a geração de μ_i foram, $\beta_0 = -5$, $\beta_1 = 2.0$, $\beta_2 = 1.0$ e $\beta_3 = 0.2$. O ponto de corte considerado foi $C = 40000$.

Nas amostras de tamanho 100, geramos covariáveis $X_{i1} \sim Bernoulli(0,5)$, $X_{i2} \sim Bernoulli(0,5)$ e $X_{i2} \sim Bernoulli(0,5)$ para obter a prevalência de 0,1. Para obter a prevalência de 0,2 simulamos as covariáveis $X_{i1} \sim Bernoulli(0,5)$, $X_{i2} \sim Bernoulli(0,5)$ e $X_{i2} \sim Bernoulli(0,5)$. Nas amostras de tamanho 500 geramos covariáveis $X_{i1} \sim Bernoulli(0,3)$, $X_{i2} \sim Bernoulli(0,3)$ e $X_{i2} \sim Bernoulli(0,3)$ para obter a prevalência de 0,05 e para alcançar a prevalência de 0,1 simulamos as covariáveis $X_{i1} \sim Bernoulli(0,5)$, $X_{i2} \sim Bernoulli(0,5)$ e $X_{i2} \sim Bernoulli(0,5)$. Nas amostras de tamanho 5000 analisamos as prevalências de 0,01 e de 0,1. Na

primeira situação geramos covariáveis $X_{i1} \sim Bernoulli(0,1)$, $X_{i2} \sim Bernoulli(0,1)$ e $X_{i2} \sim Bernoulli(0,1)$ e para obter a prevalência de 0,1 geramos covariáveis $X_{i1} \sim Bernoulli(0,4)$, $X_{i2} \sim Bernoulli(0,4)$ e $X_{i2} \sim Bernoulli(0,3)$.

Na Tabela 8.19 encontramos o vício amostral, o erro quadrático médio (EQM), e erro absoluto médio (EAM) e a média das estimativas dos parâmetros nas amostras de tamanho n=100 e prevalência p=0,10 e p=0,20. Notamos que o vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com resposta de origem são inferiores a estas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual.

Tabela 8.19: Qualidade do ajuste - Distribuição de origem Exponencial - n=100.

		Modelo logístico usual				Modelo resposta de origem			
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,10	β_0	-3,600	75,891	4,883	-8,600	0,085	0,662	0,648	-4,914
	β_1	3,647	64,894	4,225	5,647	0,232	0,591	0,588	2,232
	β_2	0,566	10,207	10,207	1,566	0,077	0,395	0,489	1,077
	β_3	0,016	1,032	0,654	0,216	0,005	0,347	0,456	0,205
p=0,20	β_0	-4,342	87,183	5,921	-9,342	0,067	1,298	0,877	-4,932
	β_1	4,522	75,257	5,034	6,522	0,412	1,024	0,732	2,412
	β_2	0,710	11,822	1,182	1,710	0,125	0,422	0,502	1,125
	β_3	0,133	1,785	0,551	0,333	0,071	0,277	0,411	0,271

Na Tabela 8.20 temos os intervalos de confiança empíricos das razão das estimativas dos modelo logito usual e do modelo logito com resposta de origem exponencial para amostras de tamanho 100. Observamos que estes intervalos apresentam amplitude elevada indicando que para este tamanho amostral as estimativas são divergentes.

Nas Tabelas 8.21 e 8.22 apresentamos os intervalos de confiança empíricos da razão das chances dos modelos logito usual e logito com resposta de origem exponencial para amostras de tamanho 100. Sabemos que para X_1 o valor real da razão das chances é de $exp(\beta_1) = 7,389$, já para X_2 a razão das chances é de $exp(\beta_2) = 2,718$

Tabela 8.20: Intervalos de Confiança Empiricos da razão das estimativas - Distribuição de origem Exponencial - n=100.

	Medida	90%	95%	99%
p=0,10	β_0	0.679 4.573	0.647 4.912	0.557 6.380
	β_1	0.471 8.358	0.360 9.571	0.113 11.0788
	β_2	-0.202 3.474	-0.637 8.374	-3.855 13.732
	β_3	-6.789 5.696	-13.973 8.674	-53.990 39.919
p=0,20	β_0	0.653 4.451	0.613 4.879	0.525 6.311
	β_1	0.504 7.637	0.453 8.844	0.289 10.907
	β_2	0.504 7.637	0.453 8.844	0.289 10.907
	β_3	-2.626 5.806	-9.324 12.909	-68.938 30.141

e para X_3 temos quea razão das chances é de $exp(\beta_3) = 1,221$. Observamos que a amplitude dos intervalos empíricos para o modelo logito usual possuem amplitude bastante elevada quando comparados aos intervalos empíricos do modelo logito com resposta de origem exponencial.

Tabela 8.21: Intervalos de Confiança Empiricos da razão das chances - Modelo logito usual - Distribuição de origem Exponencial - n=100.

	Medida	99%	95%	90%
p=0,10	β_1	2.119 2.318e+08	1.689 2.555e+08	1.155 4.360e+08
	β_2	0.846 12.567	6.862e-01 1.308e+08	4.253e-01 4.070e+08
	β_3	0.353 4.979	0.279 6.712	0.142 13.057
p=0,20	β_1	2.515 1.261e+08	1.967 1.374e+08	1.387 1.615e+08
	β_2	1.068 13.517	0.873 7.687e+07	0.656 1.391422e+08
	β_3	0.504 3.722	0.431 4.750	0.302 7.715

Tabela 8.22: Intervalos de Confiança Empiricos da razão das chances - Distribuição de origem Exponencial - n=100.

	Medida	90%	95%	99%
p=0,10	β_1	3.108 33.926	2.580 51.557	1.903 94.672
	β_2	1.154 8.501	0.946 11.383	0.650 19.047
	β_3	0.478 3.274	0.368 4.089	0.221 5.950
p=0,20	β_1	3.298 59.722	2.753 116.984	1.908 362.389
	β_2	1.218 9.885	1.060 13.104	0.856 23.324
	β_3	0.603 2.986	0.508 3.894	0.419 6.144

Nas Tabelas 8.23 e 8.24 encontramos a probabilidade de cobertura e a amplitude média dos intervalos de confiança assintóticos para os parâmetros dos modelos logito usual e logito com resposta de origem exponencial para amostras de tamanho 100. Ambos os intervalos atingem o nível de confiança nominal no entanto, a amplitude média dos intervalos para os parâmetros do modelo logito com resposta de origem é inferior a amplitude média dos parâmetros do modelo logito usual.

Tabela 8.23: Probabilidade de Cobertura - Distribuição de origem Exponencial - n=100.

	Modelo logístico usual						
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,10	β_0	0.873	0.928	0.981	0.892	0.946	0.986
	β_1	0.950	0.979	0.996	0.894	0.953	0.988
	β_2	0.956	0.980	0.999	0.897	0.957	0.998
	β_3	0.903	0.970	1,000	0.913	0.956	0.989
p=0,20	β_0	0.818	0.884	0.954	0.865	0.918	0.977
	β_1	0.958	0.976	0.992	0.889	0.948	0.988
	β_2	0.947	0.973	0.995	0.904	0.961	0.996
	β_3	0.900	0.976	0.997	0.931	0.965	0.995

Tabela 8.24: Amplitude Média - Distribuição de origem Exponencial - n=100.

	Modelo logístico usual				Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,10	β_0	2036.767	2434.185	3191.763	2.601	3.109	4.076
	β_1	1760.008	2103.424	2758.061	4.076	2.294	2.741
	β_2	344.343	411.532	539.611	2.039	2.436	3.195
	β_3	15.283	18.265	23.950	1.931	2.308	3.026
p=0,20	β_0	1768.530	2113.610	2771.417	3.311	3.957	5.189
	β_1	1557.282	1861.142	2440.375	2.671	3.193	4.187
	β_2	239.606	286.358	375.480	2.045	2.444	3.205
	β_3	38.931	46.528	61.009	1.729	2.066	2.710

Na Tabela 8.25 encontramos o vício amostral, o erro quadrático médio (EQM), o erro absoluto médio (EAM) e a média das estimativas dos parâmetros nas amostras de tamanho n=500 e prevalência p=0,05 e p=0,10. Notamos que o vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com

resposta de origem são inferiores a estas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual.

Tabela 8.25: Qualidade do ajuste - Distribuição de origem Exponencial - n=500.

		Modelo logístico usual			Modelo resposta de origem				
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,05	β_0	-1,738	22,877	1,982	-6,738	-0,067	0,170	0,328	-5,067
	β_1	1,355	21,120	1,765	3,355	-0,050	0,112	0,267	1,949
	β_2	0,116	0,883	0,461	1,116	-0,030	8,884E-05	0,239	0,969
	β_3	0,044	0,201	0,346	0,244	0,0098	0,079	0,221	0,209
p=0,10	β_0	-2,530	39,085	3,004	-7,530	-0,103	0,350	0,468	-5,103
	β_1	2,231	34,408	2,664	4,231	0,038	0,181	0,334	2,038
	β_2	0,253	2,931	0,580	1,253	0,034	0,126	0,282	1,034
	β_3	0,036	0,1641	0,310	0,236	0,021	0,094	0,240	0,221

Os intervalos de confiança empíricos para a razão das estimativas dos modelos logito usual e logito com resposta de origem exponencial para amostras de tamanho 500 são apresentados na Tabela 8.26. A amplitude dos intervalos é inferior aquela observada para as amostras de tamanho 100 indicando que a medida que aumentamos o tamanho amostral as estimativas de ambos os modelos convergem.

Tabela 8.26: Intervalos de Confiança Empiricos da razão das estimativas - Distribuição de origem Exponencial - n=500.

	Medida	90%	95%	99%
p=0,05	β_0	0.8817 4.002	0.847 4.202	0.801 4.583
	β_1	0.627 8.034	0.557 9.215	0.431 10.253
	β_2	0.400 1.908	0.240 2.147	-0.214 3.186
	β_3	-3.281 8.070	-9.039 14.974	-55.707 46.835
p=0,10	β_0	0.819 3.969	0.787 4.143	0.710 4.463
	β_1	0.638 7.938	0.559 8.610	0.433 10.275
	β_2	0.478 1.974	0.324 2.283	-0.078 10.589
	β_3	-3.021 5.072	-7.695 7.616	-36.691 19.996

Os intervalos empíricos para a razão das chances dos modelo logito usual e logito com resposta de origem exponencial para amostras de tamanho 500 estão nas Tabela 8.27 e 8.28. Observamos que a precisão das estimativas do modelo logito com resposta

de origem é superior quando comparada ao modelo logito usual.

Tabela 8.27: Intervalos de Confiança Empiricos da razão das chances - Distribuição de origem Exponencial - n=500.

	Medida	90%	95%	99%
p=0,05	β_1	3.025 7.660e+07	2.558 8.549e+07	2.006 9.614e+07
	β_2	1.308 8.607	1.175 9.905	0.915 18.757
	β_3	0.638 2.683	0.563 3.205	0.414 4.978
p=0,1	β_1	3.134 4.246e+07	2.649 4.493e+07	1.956 4.966e+07
	β_2	1.392 8.285	1.195 13.733	9.645e-01 3.991e+07
	β_3	0.700 2.486	0.602 2.975	0.520 4.564

Tabela 8.28: Intervalos de Confiança Empiricos da razão das chances - Distribuição de origem Exponencial - n=500.

	Medida	90%	95%	99%
p=0,05	β_1	4.162 12.484	3.912 13.668	3.328 17.133
	β_2	1.609 4.402	1.502 4.751	1.262 6.082
	β_3	0.726 2.183	0.593 2.833	
p=0,1	β_1	4.088 16.613	3.691 19.857	2.876 26.191
	β_2	1.611 5.145	1.459 5.763	1.271 8.368
	β_3	0.769 2.034	0.705 2.278	0.557 3.002

Nas Tabelas 8.29 e 8.30 apresentamos a probabilidade de cobertura e a amplitude média dos intervalos de confiança empíricos dos parâmetros do modelo logito usual e do modelo logito com resposta de origem exponencial para amostras de tamanho 500. Ambos os intervalos atingem o nível de confiança nominal no entanto, a amplitude média dos intervalos de confiança para os parâmetros do modelo logito com resposta de origem é inferior.

Na Tabela 8.31 encontramos o vício amostral, o erro quadrático médio (EQM), e erro absoluto médio (EAM) e a média das estimativas dos parâmetros nas amostras de tamanho n=5000 e prevalência p=0,01 e p=0,10. Notamos que o vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com

Tabela 8.29: Probabilidade de Cobertura - Distribuição de origem Exponencial - n=500.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,05	β_0	0.949	0.979	0.994	0.887	0.942	0.990
	β_1	0.950	0.972	0.990	0.911	0.959	0.992
	β_2	0.938	0.969	0.993	0.904	0.951	0.994
	β_3	0.900	0.960	0.996	0.911	0.952	0.987
p=0,10	β_0	0.949	0.987	0.998	0.904	0.953	0.993
	β_1	0.930	0.958	0.991	0.915	0.956	0.994
	β_2	0.903	0.963	0.994	0.901	0.958	0.990
	β_3	0,904	0.961	0.994	0.907	0.955	0.988

Tabela 8.30: Amplitude Média - Distribuição de origem Exponencial - n=500.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,05	β_0	460.077	549.848	720.975	1.885	2.253	2.955
	β_1	433.819	518.467	679.826	1.422	1.699	2.228
	β_2	31.593	37.757	49.508	1.167	1.395	1.829
	β_3	1.344	1.606	2.106	0.995	1.189	1.559
p=0,10	β_0	329.414	393.690	516.217	1.356	1.621	2.126
	β_1	321.059	383.705	503.123	1.129	1.349	1.769
	β_2	9.609	11.485	15.059	1.001	1.197	1.570
	β_3	1.417	1.694	2.221	0.937	1.120	1.469

resposta de origem são inferiores a estas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual.

Tabela 8.31: Qualidade do ajuste - Distribuição de origem Exponencial - n=5000.

		Modelo logístico usual				Modelo resposta de origem			
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,01	β_0	-0,155	0,053	0,182	-5,155	-0,020	0,006	0,063	-5,020
	β_1	-0,003	0,081	0,225	1,996	-0,054	0,026	0,130	1,945
	β_2	-0,028	0,133	0,285	0,971	-0,020	0,032	0,141	0,979
	β_3	-0,095	0,497	0,382	0,104	-0,007	0,042	0,168	0,192
p=0,10	β_0	-0,043	0,090	0,237	-5,043	-0,007	0,031	0,142	-5,007
	β_1	0,031	0,062	0,196	2,031	0,005	0,017	0,105	2,005
	β_2	0,002	0,025	0,129	1,002	-0,002	0,012	0,088	0,997
	β_3	0,006	0,645	0,794	0,206	0,002	0,009	0,076	0,202

Os intervalos de confiança empíricos para a razão das estimativas do modelo

logito usual e do modelo logito com resposta de origem exponencial para amostras de tamanho 5000. Observamos que para a prevalência p=0,10 as estimativas dos parâmetros são convergentes com excessão das estimativas do parâmetro β_3 cujos intervalos empíricos apresentaram uma amplitude maior.

Tabela 8.32: Intervalos de Confiança Empíricos da razão das estimativas - Distribuição de origem Exponencial - n=5000.

	Medida	90%	95%	99%
p=0,01	β_0	0.977 1.082	0.970 1.096	0.956 1.115
	β_1	0.819 1.211	0.778 1.243	0.713 1.310
	β_2	0.379 1.452	0.191 1.599	-0.073 1.863
	β_3	-6.988 9.467	-13.022 17.958	-47.257 64.539
p=0,10	β_0	0.933 1.093	0.922 1.111	0.897 1.137
	β_1	0.855 1.192	0.830 1.237	0.793 1.319
	β_2	0.821 1.196	0.789 1.231	0.741 1.318
	β_3	0.189 1.942	-0.404 2.619	-5.631 5.837

Os intervalos de confiança empíricos para a razão das chances dos modelos logito usual e logito com resposta de origem exponencial para amostras de tamanho 5000 estão nas Tabelas 8.33 e 8.34. Notamos que as estimativas do modelo logito com resposta de origem são mais precisas quando comparadas as estimativas para o modelo logito usual. Além disso, as estimativas considerando a amplitude de 0,10 são mais precisas do que as estimativas para a prevalência de 0,01.

Tabela 8.33: Intervalos de Confiança Empiricos da razão das chances - Modelo logito usual - Distribuição de origem Exponencial - n=5000.

	Medida	90%	95%	99%
p=0,01	β_1	4.530 11.670	4.165 12.969	3.554 15.648
	β_2	1.404 4.522	1.185 4.989	0.941 5.964
	β_3	0.485 2.152	0.356 2.318	0.168 2.729
p=0,10	β_1	5.260 11.965	4.937 12.673	4.460 15.054
	β_2	2.113 3.539	2.009 3.759	1.858 4.154
	β_3	1.014 1.512	0.974 1.579	0.909 1.660

Tabela 8.34: Intervalos de Confiança Empiricos da razão das chances - Distribuição de origem Exponencial - n=5000.

	Medida	90%	95%	99%
p=0,01	β_1	5.497 9.103	5.261 9.578	4.744 10.269
	β_2	1.940 3.621	1.845 3.787	1.693 4.175
	β_3	0.855 1.686	0.796 1.777	0.732 1.882
p=0,10	β_1	6.043 9.257	5.885 9.739	5.353 10.352
	β_2	2.269 3.268	2.210 3.369	2.061 3.581
	β_3	1.049 1.443	1.019 1.472	0.970 1.551

Nas Tabelas 8.35 e 8.36 encontram-se a probabilidade de cobertura e a amplitude média dos intervalos de confiança assintóticos dos parâmetros dos modelo logito usual e logito com resposta de origem exponencial para amostras de tamanho 5000. Ambos os modelos atingem o nível nominal no entanto, a amplitude média dos intervalos para o modelo logito com resposta de origem é inferior a amplitude média dos intervalos assintóticos para o modelo logito usual.

Tabela 8.35: Probabilidade de Cobertura - Distribuição de origem Exponencial - n=5000.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,01	β_0	0.879	0.947	0.995	0.907	0.954	0.989
	β_1	0.889	0.948	0.990	0.894	0.956	0.990
	β_2	0.892	0.947	0.989	0.891	0.953	0.991
	β_3	0,899	0.967	0.996	0.910	0.961	0.997
p=0,10	β_0	0.908	0.957	0.991	0.898	0.952	0.989
	β_1	0.910	0.964	0.993	0.919	0.959	0.992
	β_2	0.900	0.950	0.990	0.957	0.957	0.989
	β_3	0.901	0.956	0.992	0.895	0.951	0.992

8.3 Distribuição de origem Lognormal

Nesta seção analisamos o desempenho do modelo logístico com resposta de origem lognormal através de dados artificiais. Na geração dos dados utilizamos três variáveis

Tabela 8.36: Amplitude Média - Distribuição de origem Exponencial - n=5000.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,01	β_0	0.665	0.795	1.043	0.265	0.317	0.416
	β_1	0.916	1.095	1.435	0.520	0.621	0.815
	β_2	1.112	1.329	1.742	0.603	0.721	0.945
	β_3	3.960	4.732	6.205	0.688	0.822	1.079
p=0,10	β_0	0.989	1.182	1.550	0.584	0.698	0.916
	β_1	0.806	0.964	1.264	0.529	0.694	0.362
	β_2	0.527	0.630	0.826	0.362	0.433	0.568
	β_3	0.412	0.493	0.646	0.310	0.371	0.487

explicativas com distribuição de Bernoulli, X_{i1}, X_{i2} e X_{i3}. A geração de dados foi repetida para três tamanhos amostrais, n=100, n=500 e n=5000. Foram geradas 1000 amostras de tamanho n com variável resposta $R_i \sim LN(\mu_i, \sigma^2)$, com $\mu_i = \sigma\phi^{-1}\left[g^{-1}(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3})\right] + log(C)]$, $i = 1, \cdots, n$. Os valores atribuídos para o vetor de parâmetros $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3)'$ para a geração de μ_i foram, $\beta_0 = -7$, $\beta_1 = 1.0$, $\beta_2 = 2.0$ e $\beta_3 = 5.0$ e $\sigma = 1$. O ponto de corte considerado foi $C = 10$.

Nas amostras de tamanho 100, geramos covariáveis $X_{i1} \sim Bernoulli(0, 5)$, $X_{i2} \sim Bernoulli(0, 5)$ e $X_{i2} \sim Bernoulli(0, 5)$ para obter a prevalência de 0,1. Para obter a prevalência de 0,2 simulamos as covariáveis $X_{i1} \sim Bernoulli(0, 5)$, $X_{i2} \sim Bernoulli(0, 5)$ e $X_{i2} \sim Bernoulli(0, 5)$. Nas amostras de tamanho 500 geramos covariáveis $X_{i1} \sim Bernoulli(0, 3)$, $X_{i2} \sim Bernoulli(0, 3)$ e $X_{i2} \sim Bernoulli(0, 3)$ para obter a prevalência de 0,05 e para alcançar a prevalência de 0,1 simulamos as covariáveis $X_{i1} \sim Bernoulli(0, 5)$, $X_{i2} \sim Bernoulli(0, 5)$ e $X_{i2} \sim Bernoulli(0, 5)$. Nas amostras de tamanho 5000 analisamos as prevalências de 0,01 e de 0,1. Na primeira situação geramos covariáveis $X_{i1} \sim Bernoulli(0, 1)$, $X_{i2} \sim Bernoulli(0, 1)$ e $X_{i2} \sim Bernoulli(0, 1)$ e para obter a prevalência de 0,1 geramos covariáveis $X_{i1} \sim Bernoulli(0, 4)$, $X_{i2} \sim Bernoulli(0, 4)$ e $X_{i2} \sim Bernoulli(0, 4)$.

Na Tabela 8.32 encontramos o vício amostral, o erro quadrático médio (EQM), o

erro absoluto médio (EAM) e a média das estimativas dos parâmetros nas amostras de tamanho n=100 e prevalência p=0,10 e p=0,20. Notamos que o vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com resposta de origem exponencial são inferiores a estas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual.

Tabela 8.37: Qualidade do ajuste - Distribuição de origem Lognormal - n=100.

		Modelo logístico usual				Modelo resposta de origem			
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,10	β_0	-15,371	320,061	15,652	-22,371	-0,570	1,240	0,852	-7,570
	β_1	0,524	8,683	1,096	1,524	0,018	0,220	0,372	1,018
	β_2	1,710	32,846	2,249	3,710	0,032	0,269	0,412	2,032
	β_3	12,509	195,630	13,060	17,509	0,058	0,531	0,565	5,058
p=0,20	β_0	-12,514	210,188	12,889	-19,514	-0,371	1,073	0,798	-7,371
	β_1	0,098	0,876	0,583	1,098	0,034	0,184	0,348	1,034
	β_2	0,557	7,770	0,998	2,557	0,047	0,240	0,386	2,046
	β_3	11,475	175,402	12,071	16,475	0,060	0,531	0,576	5,060

Na Tabela 8.33 apresentamos os intervalos de confiança empíricos para a razão das estimativas dos modelos logito usual e logito com resposta de origem lognormal para amostras de tamanho 100.

Tabela 8.38: Intervalos de Confiança Empiricos da razão das estimativas - Distribuição de origem Lognormal - n=100.

	Medida	90%	95%	99%
p=0,10	β_0	0.823 5.040	0.758 5.596	0.663 7.869
	β_1	-0.421 3.276	-1.309 9.179	-5.762 18.877
	β_2	0.504 8.031	0.422 9.597	0.136 11.707
	β_3	0.642 4.913	0.585 5.093	0.507 5.911
p=0,20	β_0	0.823 3.624	0.775 3.900	0.718 5.481
	β_1	-0.023 2.069	-0.547 2.420	-3.931 4.737
	β_2	0.618 1.705	0.520 2.033	0.374 8.826
	β_3	0.690 4.704	0.637 4.901	0.560 5.348

Sabemos que para X_1 o valor real da razão das chances é de $exp(\beta_1) = 2,718$, já para X_2 a razão das chances é de $exp(\beta_2) = 7,389$ e para X_3 temos quea razão

das chances é de $exp(\beta_3) = 148,413$. Os intervalos de confiança empíricos para a razão das chances dos modelo logito usual e logito com resposta de origem lognormal para amostras de tamanho 100 apresentados nas Tabelas 8.39 e 8.40 indicam que as estimtivas do modelo logito com resposta de origem são mais precisas.

Tabela 8.39: Intervalos de Confiança Empiricos da razão das chances - Modelo logito usual - Dsitribuição de origem Lognormal - n=100.

	Medida	90%	95%	99%
p=0,10	β_1	0.770 20.132	0.544 41.179	2.896e-01 1.005e+09
	β_2	2.199 7.074e+08	1.866 1.191e+09	1.220 4.960e+09
	β_3	2.020e+01 2.106e+09	1.499e+01 3.422e+09	9.193 4.912e+14
p=0,20	β_1	1.001 9.219	0.847 13.079	0.595 25.676
	β_2	2.674 44.997	2.254 65.523	1.807 1.467e+09
	β_3	2.749e+01 1.148e+09	2.174e+01 1.384e+09	1.297e+01 2.372e+09

Tabela 8.40: Intervalos de Confiança Empiricos da razão das chances - Modelo logito com resposta de origem - Distribuição de origem Lognormal - n=100.

	Medida	90%	95%	99%
p=0,10	β_1	1.347 6.073	1.215 7.199	0.829 10.228
	β_2	2.199 7.074e+08	1.866 1.191e+09	1.220 4.960e+09
	β_3	53.162 525.355	44.052 865.100	33.693 1365.209
p=0,20	β_1	1.402 5.670	1.230 6.477	0.993 8.194
	β_2	3.595 18.001	3.145 21.366	2.509 31.662
	β_3	51.983 589.962	45.575 753.0350	36.223 1281.020

Nas Tabelas 8.41 e 8.42 apresentamos a probabilidade de cobertura e a amplitude média dos intervalos de confiança assintóticos para os parâmetros dos modelos logito usual e logito com resposta de origem normal para amostras de tamanho 100. Os intervalos para ambos os modelos atingem o nível de confiança nominal e a amplitude média dos intervalos para o modelo com resposta de origem é inferior a amplitude dos intervalos para o modelo logito usual.

Na Tabela 8.43 encontramos o vício amostral, o erro quadrático médio (EQM), e erro absoluto médio (EAM) e a média das estimativas dos parâmetros nas amostras

Tabela 8.41: Probabilidade de Cobertura - Distribuição de origem Lognormal - n=100.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,10	β_0	0.981	0.991	0.998	0.889	0.948	0.989
	β_1	0.925	0.972	1,000	0.912	0.958	0.990
	β_2	0.936	0.981	0.997	0.899	0.953	0.991
	β_3	0,910	0.965	0.988	0.908	0.949	0.991
p=0,20	β_0	0.986	0.995	0.999	0.902	0.945	0.990
	β_1	0.897	0.959	0.992	0.901	0.955	0.988
	β_2	0.934	0.978	0.996	0.899	0.950	0.993
	β_3	0,900	0.977	0.992	0.898	0.955	0.988

Tabela 8.42: Amplitude Média - Distribuição de origem Lognormal - n=100.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,10	β_0	7590.144	9071.147	11894.31	3.104	3.710	4.865
	β_1	283.0997	338.338	443.637	1.543	1.844	2.418
	β_2	1076.024	1285.980	1686.208	1.638	1.958	2.568
	β_3	6955.821	8313.054	10900.28	2.405	2.874	3.769
p=0,20	β_0	5105.962	6102.247	8001.416	3.076	3.677	4.821
	β_1	14.997	17.924	23.502	1.432	1.711	2.244
	β_2	172.958	206.705	271.037	1.533	1.832	2.403
	β_3	5011.561	5989.427	7853.483	2.352	2.811	3.686

de tamanho n=100 e prevalência p=0,05 e p=0,10. Notamos que o vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com resposta de origem são inferiores a estas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual.

Na Tabela 8.52 apresentamos os intervalos empíricos da razão das estimativas dos modelos logito usual e do modelo logito com resposta de origem lognormal para amostras de tamanho 500. Observamos que a amplitude dos intervalos para a prevalência de 0.10 é menor do que a prevalência considerando a amostra de pravalência 0,10.

Nas Tabelas 8.53 e 8.54 encontram-se os intervalos de confiança empíricos para a razão das chances dos modelos logito usual e logito com resposta de origem lognormal

Tabela 8.43: Qualidade do ajuste - Distribuição de origem Lognormal - n=500.

		Modelo logístico usual				Modelo resposta de origem			
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,05	β_0	-4,766	77,247	5,241	-11,766	-0,050	0,145	0,304	-7,050
	β_1	0,0088	0,400	0,494	1,008	0,010	0,079	0,224	1,010
	β_2	0,099	0,353	0,461	2,099	0,027	0,083	0,230	2,027
	β_3	4,728	77,228	5,209	9,728	0,039	0,122	0,284	5,039
p=0,10	β_0	-2,977	48,569	3,563	-9,977	-0,030	0,152	0,308	-7,030
	β_1	0,036	0,174	0,333	1,036	0,018	0,054	0,186	1,018
	β_2	0,058	0,165	0,317	2,058	0,019	0,056	0,189	2,019
	β_3	3,038	49,031	3,538	8,038	0,072	0,120	0,274	5,072

Tabela 8.44: Intervalos de Confiança Empiricos da razão das estimativas - Distribuição de origem Lognormal - n=500.

	Medida	90%	95%	99%
p=0,05	β_0	0.839 3.319	0.821 3.422	0.788 3.594
	β_1	-0.019 1.961	-0.241 2.129	-1.089 2.966
	β_2	0.628 1.484	0.561 1.593	0.416 1.769
	β_3	0.787 4.295	0.756 4.428	0.706 4.629
p=0,10	β_0	0.841 3.236	0.827 3.315	0.776 3.437
	β_1	0.428 1.633	0.323 1.753	0.114 2.139
	β_2	0.760 1.300	0.719 1.387	0.628 1.518
	β_3	0.804 4.106	0.773 4.200	0.731 4.383

para amostras de tamanho 500. Os intervalos indicam que as estimativas da razão das chances do modelo logito com resposta de origem são mais precisas que as estimativas obtidas por meio do modelo logito usual.

Tabela 8.45: Intervalos de Confiança Empiricos da razão das chances - Modelo logito Usual - Distribuição de origem Lognormal - n=500.

	Medida	90%	95%	99%
p=0,05	β_1	0.985 7.553	0.839 10.031	0.564 16.388
	β_2	3.411 22.344	2.849 26.005	2.096 41.956
	β_3	5.133e+01 1.318e+09	4.186e+01 1.957e+09	3.070e+01 5.587e+09
p=0,10	β_1	1.466 5.623	1.277 6.690	1.079 8.562
	β_2	4.159 15.291	3.775 17.260	3.018 27.124
	β_3	5.179e+01 1.509e+09	4.406e+01 1.739e+09	3.189e+01 2.050e+09

Nas Tabelas 8.47 e 8.48 encontram-se a probabilidade de cobertura e a amplitude

Tabela 8.46: Intervalos de Confiança Empiricos da razão das chances - Modelo logito com reposta de origem - Distribuição de origem Lognormal - n=500.

	Medida	90%	95%	99%
p=0,05	β_1	1.729 4.370	1.592 4.775	1.355 5.725
	β_2	4.695 12.076	4.333 13.329	3.723 16.799
	β_3	87.411 276.608	80.711 307.584	66.383 365.080
p=0,10	β_1	1.870 4.040	1.738 4.379	1.547 5.029
	β_2	5.284 11.179	4.874 11.960	4.216 14.180
	β_3	92.624 286.421	84.107 324.509	65.911 400.060

média dos intervalos de confiança assintóticos dos para os parâmetros dos modelos logito usual e logito com resposta de origem. Os intervalos de ambos os modelos atingem o nível nominal de confiança estabelecido no entanto, os intervalos para o modelo logito com resposta de origem são mais precisos pois apresentam menor amplitude.

Tabela 8.47: Probabilidade de Cobertura - Distribuição de origem Lognormal - n=500.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,05	β_0	0.956	0.979	0.996	0.899	0.950	0.987
	β_1	0.902	0.959	0.993	0.912	0.957	0.993
	β_2	0.899	0.959	0.997	0.902	0.954	0.992
	β_3	0.905	0.975	0.991	0.896	0.953	0.992
p=0,10	β_0	0.934	0.959	0.986	0.896	0.938	0.987
	β_1	0.914	0.953	0.997	0.901	0.947	0.992
	β_2	0.915	0.962	0.991	0.908	0.958	0.991
	β_3	0,912	0.954	0.985	0.889	0.949	0.983

Na Tabela 8.49 encontramos o vício amostral, o erro quadrático médio (EQM), e erro absoluto médio (EAM) e a média das estimativas dos parâmetros nas amostras de tamanho n=5000 e prevalência p=0,01 e p=0,10. Notamos que o vício, o erro quadrático médio e o erro absoluto médio das estimativas do modelo logito com resposta de origem são inferiores a estas medidas calculadas utilizando as estimativas produzidas pelo modelo logito usual.

Tabela 8.48: Amplitude Média - Distribuição de origem Lognormal - n=500.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,05	β_0	1445.068	1727.033	2264.528	1.245	1.488	1.952
	β_1	1.969	2.353	3.086	0.950	1.135	1.489
	β_2	1.870	2.236	2.931	0.952	1.138	1.492
	β_3	1444.938	1726.877	2264.324	1.140	1.362	1.786
p=0,10	β_0	780.328	932.587	1222.831	1.492	1.492	1.957
	β_1	1.363	1.629	2.136	0.767	0.917	1.202
	β_2	1.339	1.600	2.098	0.790	0.945	1.239
	β_3	780.159	932.385	1222.566	1.072	1.281	1.680

Tabela 8.49: Qualidade do ajuste - Distribuição de origem Lognormal - n=5000.

		Modelo logístico usual				Modelo resposta de origem			
	Parâmetros	Vício	EQM	EAM	Estimativas	Vício	EQM	EAM	Estimativas
p=0,01	β_0	-0,146	0,460	0,351	-7,146	-0,011	0,013	0,093	-7,011
	β_1	-0,022	0,113	0,265	0,977	-0,0004	0,016	0,101	0,999
	β_2	-0,0003	0,094	0,241	1,999	-0,0005	0,016	0,101	1,999
	β_3	0,104	0,468	0,357	5,104	-0,008	0,0146	0,096	4,991
p=0,10	β_0	-0,046	0,100	0,249	-7,046	-0,004	0,015	0,101	-7,004
	β_1	-0,001	0,013	0,092	0,998	-0,002	0,004	0,055	0,997
	β_2	0,001	0,014	0,095	2,001	0,003	0,004	0,055	2,003
	β_3	0,043	0,088	0,233	5,043	0,001	0,010	0,083	5,001

Os intervalos de confiança empíricos da razão das estimativas dos modelos logito usual e logito com resposta de origem lognormal para amostras de tamanho 5000 encontrados na Tabela 8.43 indicam que as estimativas de ambos os modelos convergem. Além disso, a amplitude destes intervalos considerando a prevalência de 0,10 é inferior a amplitude apresentada pelos intervalos considerando a prevalência de 0,01.

Os intervalos empíricos para a razão das chances dos modelos logito usual e logito com resposta de origem lognormal para amostras de tamanho 5000 encontrados na Tabela 8.50 indicam uma precisão superior nas estimativas obtidas através do modelo logito com resposta de origem. Além disso, quando comparamos a precisão dos resultados considerando as duas prevalências analisadas observamos que a amplitude

Tabela 8.50: Intervalos de Confiança Empiricos da razão das estimativas - Distribuição de origem Lognormal - n=5000.

	Medida	90%	95%	99%
p=0,01	β_0	0.932 1.126	0.919 1.159	0.894 1.254
	β_1	0.402 1.480	0.302 1.563	0.077 1.786
	β_2	0.761 1.238	0.724 1.288	0.617 1.367
	β_3	0.900 1.174	0.883 1.216	0.847 1.356
p=0,10	β_0	0.944 1.072	0.932 1.085	0.921 1.125
	β_1	0.844 1.157	0.818 1.192	0.780 1.240
	β_2	0.922 1.076	0.908 1.089	0.879 1.089
	β_3	0.930 1.097	0.920 1.117	0.891 1.169

dos intervalos construídos através de amostras com prevalência de 0,10 é inferior a amplitude dos intervalos obtidos considerando amostras com prevalância de 0,01.

Tabela 8.51: Intervalos de Confiança Empiricos da razão das chances - Modelo logito usual - Distribuição Lognormal - n=5000.

	Medida	90%	95%	99%
p=0,01	β_1	1.457 4.469	1.306 4.940	1.062 6.093
	β_2	4.397 12.062	3.973 13.435	3.336 16.684
	β_3	87.12 369.905	81.527 437.423	62.517 904.604
p=0,10	β_1	2.234 3.276	2.159 3.431	2.053 3.712
	β_2	6.059 8.966	5.886 9.274	5.574 10.018
	β_3	101.215 255.177	94.817 288.825	82.262 402.277

Tabela 8.52: Intervalos de Confiança Empiricos da razão das chances - Modelo logito com resposta de origem - Distribuição de origem Lognormal - n=5000.

	Medida	90%	95%	99%
p=0,01	β_1	2.207 3.329	2.130 3.473	2.009 3.890
	β_2	6.034 9.192	5.818 9.528	5.428 10.209
	β_3	120.810 180.553	117.774 187.391	110.959 199.106
p=0,10	β_1	2.433 3.037	2.362 3.123	2.300 3.265
	β_2	6.636 8.323	6.482 8.496	6.168 8.739
	β_3	124.913 176.059	121.539 180.823	115.152 192.856

Nas Tabelas 8.53 e 8.54 apresentamos a probabilidade de cobertura e a amplitude média dos intervalos de confiança assintóticos dos parâmetros dos modelos logito

usual e logito com resposta de origem lognormal para amostras de tamanho 5000. O nível de confiança nominal é observado nos intervalos de ambos os modelos contudo, os intervalos para os parâmetros do modelo logito com resposta de origem são mais precisos.

Tabela 8.53: Probabilidade de Cobertura - Dsitribuição de origem Lognormal - n=5000.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,01	β_0	0.917	0.975	0.995	0.908	0.954	0.992
	β_1	0.898	0.952	0.993	0.921	0.961	0.992
	β_2	0.900	0.947	0.990	0.899	0.952	0.995
	β_3	0.905	0.970	0.992	0.910	0.967	0.993
p=0,10	β_0	0.914	0.961	0.992	0.901	0.948	0.989
	β_1	0.899	0.954	0.994	0.899	0.953	0.987
	β_2	0.900	0.944	0.993	0.899	0.946	0.983
	β_3	0.900	0.960	0.994	0.901	0.948	0.987

Tabela 8.54: Amplitude Média - Distribuição de origem Lognormal - n=5000.

		Modelo logístico usual			Modelo resposta de origem		
	Parâmetros	90%	95%	99%	90%	95%	99%
p=0,01	β_0	3.670	4.387	5.752	0.388	0.464	0.608
	β_1	1.094	1.308	1.715	0.432	0.517	0.678
	β_2	0.990	1.183	1.551	0.417	0.498	0.653
	β_3	3.662	4.376	5.739	0.412	0.492	0.645
p=0,10	β_0	0.969	1.159	1.519	0.395	0.472	0.619
	β_1	0.387	0.463	0.607	0.226	0.270	0.354
	β_2	0.384	0.459	0.602	0.236	0.282	0.370
	β_3	0.908	1.085	1.423	0.330	0.395	0.518

Capítulo 9

Análise de dados reais

Neste Capítulo analisamos um conjunto de dados reais de uma instituição financeira cuja variável resposta representa fraude em cartão de crédito. Como trata-se de dados reais, as covariáveis são tratadas com nomes fictícios. Os dados originais possuem 172452 observações das quais apenas 2234 representam fraude, cerca de 1,30%.

A base de dados possui dez variáveis explicativas, além da variável resposta que indica fraude.

As variáveis explicativas foram categorizadas. Inicialmente, foram divididas em dez classes, ou seja, os decis. Após análises bivariadas chegamos a categorização final utilizada no ajuste dos modelos. Aplicamos a técnica de seleção de variáveis *stepwise* e esta técnica indicou cinco covariáveis que devem permanecer no modelo final.

Inicialmente dividimos a amostra original em amostra treinamento, com 70% dos dados, amostra em que os modelos foram ajustados, e amostra teste com 30% dos dados, utilizada para calcular as medidas preditivas referente a cada modelo.

Na Tabela 9.1 encontramos as estimativas dos parâmetros do modelo de regressão

logística usual e os testes individuais de Wald.

Tabela 9.1: Parâmetros estimados modelo logito usual.

Variáveis	GL	Estimativas	Erro Padrão	Teste de Wald	Valor p
Intercepto	1	-2.677	0.159	280.6489	0.0001
X_1	1	0.588	0.034	290.583	0.0001
X_2	1	0.500	0.062	65.021	0.0001
X_2	1	0.215	0.064	11.307	0.0008
X_2	1	-0.068	0.067	1.052	0.304
X_2	1	-0.336	0.064	27.249	.0001
X_3	1	0.522	0.087	36.013	.0001
X_4	1	-0.411	0.146	7.916	0.004
X_4	1	0.445	0.275	2.616	0.105
X_5	1	-0.720	0.130	30.625	.0001
X_5	1	-0.233	0.085	7.560	0.006
X_5	1	0.094	0.069	1.853	0.173
X_5	1	0.278	0.070	15.788	.0001
X_5	1	0.161	0.110	2.134	0.144
X_5	1	0.449	0.093	23.300	.0001

Todas as variáveis apresentadas na Tabela 9.1 são siginificativas de acordo com o Teste de Wald.

Na Tabela 9.2 apresentamos as estimativas dos parâmetros do modelo logito limitado juntamente com o Teste de Wald que indica que todas as variáveis apresentadas são significativas no modelo assim como o parâmetro ω.

Na Tabela 9.3 apresentamos as estimativas dos parâmetros do modelo logito generalizado juntamente com o Teste de Wald para verificar a significância destas estimativas.

Na Tabela 9.4 encontramos as medidas AIC, BIC e -2log(verossimilhança) para os três modelos ajustados. Da acordo com ests medidas o modelo logito limitado apresentou o melhor desempenho seguido pelo modelo logito usual e pelo modelo logito generalizado.

Na Tabela 9.5 apresentamos as medidas preditivas para os modelos logito usual, logito limitado, logito generalizado e logito usual construídos em amostras balance-

Tabela 9.2: Parâmetros estimados modelo logito limitado.

Variáveis	GL	Estimativas	Erro Padrão	Teste de Wald	Valor p
w	1	0.234	0.089	2.611	0.009
Intercepto	1	-0.770	0.686	-1.121	0.261
X_1	1	0.704	0.077	9.116	¡0.001
X_2	1	0.602	0.091	6.546	¡0.001
X_2	1	0.240	0.078	3.083	0.0020
X_2	1	-0.082	0.078	-1.058	0.289
X_2	1	-0.401	0.080	-4.964	¡0.0001
X_3	1	0.677	0.138	4.891	¡0.001
X_4	1	-0.553	0.265	-2.086	0.036
X_4	1	0.707	0.516	1.370	0.170
X_5	1	-0.795	0.146	-5.437	¡0.001
X_5	1	-0.270	0.097	-2.773	0.005
X_5	1	0.099	0.080	1.232	0.217
X_5	1	0.323	0.086	3.749	0.0001
X_5	1	0.149	0.129	1.155	0.247
X_5	1	0.528	0.122	4.305	0.00001

Tabela 9.3: Parâmetros estimados modelo logito generalizado.

Variáveis	GL	Estimativas	Erro Padrão	Teste de Wald	Valor p
α_1	1	1.02			
Intercepto	1	-1.266	0.050	-25.106	¡0.001
X_1	1	0.140	0.008	16.233	¡0.001
X_2	1	0.118	0.015	7.875	¡0.001
X_2	1	0.046	0.015	3.031	0.002
X_2	1	-0.016	0.015	-1.116	0.264
X_2	1	-0.079	0.013	-5.728	¡0.001
X_3	1	0.131	0.023	5.564	¡0.001
X_4	1	-0.103	0.046	-2.255	0.024
X_4	1	0.136	0.089	1.514	0.129
X_5	1	-0.147	0.025	-5.816	¡0.001
X_5	1	-0.052	0.018	-2.881	0.003
X_5	1	0.017	0.015	1.101	0.270
X_5	1	0.060	0.016	3.717	0.0002
X_5	1	0.025	0.025	1.007	0.313
X_5	1	0.104	0.023	4.478	¡0.001

Tabela 9.4: Medidas de qualidade do ajuste.

Modelo	AIC	BIC	-2*log(verossimilhança)
Logito Usual	8726.026	8854.676	8696.815
Logito Limitado	8725.026	8819.315	8693.026
Logito Generalizado	8729.12	8823.409	8697.120

adas com estimadores KZ1 e KZ2.

Tabela 9.5: Medidas preditivas.

Modelo	SEN	ESP	VPP	VPN	ACC	MCC
Logito Usual	0.632	0.683	0.052	0.985	0.682	0.109
Logito Usual-Balanceado	0.622	0.673	0.051	0.985	0.662	0.107
Logito Limitado	0.632	0.681	0.052	0.985	0.680	0.108
Logito Generalizado	0.713	0.616	0.049	0.987	0.618	0.109
Usual KZ1	0.701	0.627	0.049	0.986	0.629	0.109
Usual KZ2	0.703	0.674	0.053	0.985	0.674	0.113

Da Tabela 9.5 notamos que o modelo logito usual com estimadores KZ2 construído em amostras balanceadas apresentou um desempenho preditivo ligeiramente superior aos demais modelos. O Coeficiente de Correlação de Mathews está bastante próximo para todos os modelos.O modelo logito generalizado apresentou a maior sensibilidade seguido do modelo logito usual aplicado em amostras balanceadas com estimadores KZ2.

Dos resultados apresentados podemos concluir que o desempenho preditivo dos modelos de classificação estudados foi bastante parecido considerando os dados reais, no entanto, o modelo logito usual com estimadores KZ foi o que apresentou medidas indicando um poder predito mais efetivo.

Capítulo 10

Conclusões

Existem diversas situações em que a variável resposta de interesse possui distribuição dicotômica extremamente desbalanceada. Alguns estudos revelam que o modelo de regressão logística usual subestima a probabilidade de sucesso quando é construído utilizando bases de dados extremamente desbalanceadas (King e Zeng, 2001). Além disso, os estimadores de máxima verossimilhança dos parâmetros do modelo de regressão logística são viciados neste caso.

Uma técnica de amostragem muito utilizada na construção das bases de dados para o ajuste modelo logito usual na situação de desbalanceamento são as amostras *state-dependent*. Neste trabalho apresentamos uma breve discussão sobre este procedimento e o Método de Correção a priori que permite manter as propriedades usuais dos estimadores de máxima verossimilhança quando o modelo logito é ajustado em amostras construídas por meio desta técnica. Além disso, realizamos um estudo de simulações para verificar o poder preditivo dos modelos ajustados considerando as amostras *state-dependent*.

Sabemos que os parâmetros do modelo de regressão logística são assintoticamente

não viciados, no entanto, de acordo com King e Zeng(2001) este vício persiste mesmo quando as amostras são grandes. McCullagh e Nelder (1989) sugeriram um estimador para o vício de qualquer modelo linear generalizado que foi adaptado por King e Zeng (2001) para o uso concomitante com as amostras *state-dependent* para o modelo de regressão logística. Neste trabalho apresentamos esta correção efetuada nos estimadores de máxima verossimilhança bem como algumas simulações para verificar o impacto causado no vício destas estimativas.

Apresentamos também algumas correções realizadas na probabilidade de sucesso estimada por meio do modelo de regressão logística que foram sugeridas por King e Zeng (2001). Tais correções permitem diminuir o vício e o erro quadrático médio destas probabilidades. Para verificar a eficiência desta metodologia foram realizadas algumas simulações para averiguar a vantagem das mesmas no poder preditivo do modelo de regressão logística.

Existem modelos encontrados na literatura desenvolvidos especialmente para a situação de dados binários desbalanceados. Um deles é o modelo logito generalizado sugerido por Stukel(1988). Este modelo possui dois parâmetros de forma e funciona melhor do que o modelo logito usual em situações em que a curva de probabilidade esperada é assimétrica. Outro modelo encontrado é o logito limitado sugerido por Cramer (2004). Este permite que seja estabelecido um limite superior para a probabilidade de sucesso. Neste trabalho apresentamos uma breve discussão sobre as características destes modelos assim como um estudo de simulação para verificar as qualidades dos mesmos quando comparados ao modelo logito usual.

Muitas vezes, a variável resposta é originalmente fruto de uma distribuição discreta ou contínua, ou seja, ela tem uma distribuição original que não a de Bernoulli e, por alguma razão esta variável foi dicotomizada através de um ponto de corte C dessa forma, o modelo pode ter a variável resposta, por exemplo, pertencente a

família exponencial no contexto dos modelos lineares generalizados com função de ligação composta. Esta metodologia foi apresentada por Suissa&Blais (1995) considerando dados reais de estudos clínicos e também dados simulados com distribuição original log-normal, Paula & Diniz(2011) estenderam os resultados para qualquer distribuição de origem pertencente a família exponencial. Dependendo do ponto de corte utilizado a variável resposta pode apresentar um desbalanceamento muito acentuado. Neste trabalho apresentamos o desenvolvimento de modelos de regressão logística com resposta de origem normal, exponencial e log-normal e realizamos um estudo comparativo utilizando dados simulados para verificar as vantagens do uso do modelo logito que agrega a resposta de origem comparado ao modelo logito usual na situação de desbalanceamento extremo.

Algumas simulações foram realizadas para comparar as propostas encontradas na literatura para o ajuste de dados binários extremamente desbalanceados com o modelo logito usual amplamente utilizado na modelagem de dados com variável reposta dicotômica. Os resultados encontrados revelam que a correção proposta por McCullagh e Nelder (1989) e adaptada por King e Zeng (2001) para os estimadores do modelo de regressão logística permite diminuir o vício, ou seja, o estimador corrigido pelo vício estimado é preferível ao estimador usual.

Apresentamos um estudo de simulações realizado para verificar o poder preditivo de cada um dos modelos de classificação apresentados. Inicialmente geramos dados de acordo com o modelo logito usual utilizando apenas uma covariável, e ajustamos nestes dados o modelo logito usual utilizando as amostras completas e o modelo logito usual utilizando amostras balanceadas com estimadores usuais, estimadores KZ1 e KZ2. Ajustamos ainda nestes dados os modelos logito limitado e logito generalizado. As medidas de qualidade de ajuste indicam que o modelo logito generalizado foi o que apresentou os piores resultados. Nestes dados, a capacidade preditiva dos

modelo logito usual e limitado apresentaram o melhor desempenho sendo que esta foi bastante parecida, já o modelo logito generalizado é o que apresentou o pior desempenho preditivo. No entanto devemos ressaltar, que o parâmetros ω do modelo logito limitado foi estimado em valores muito próximos de 1, logo devemos a esse fato a igualdade da capacidade preditiva entre os modelos logito usual e limitado.

O processo foi repetido para amostras com uma covariável geradas através do modelo logito limitado. Nesta situação a covariável apresenta uma baixa associação com a variável resposta, assim todas as medidas preditivas indicaram baixa qualidade para todos os modelos. O modelo que apresentou o melhor desempenho preditivo foi o modelo logito limitado, e o que apresentou a pior performance preditiva foi o modelo logito generalizado.

Na situação em que as amostras foram geradas através do modelo logito generalizado tal modelo foi o que apresentou o melhor desempenho preditivo enquanto que o modelo logito usual ajustado em amostras completas apresentou o pior desempenho.

Notamos que o desempenho preditivo de todos os modelos está extremamente relacionado com a dependência entre a variável resposta e as covariáveis. Assim, os modelos ajustados com os dados gerados através dos modelos logito usual e generalizado apresentaram desempenho superior ao desempenho dos modelos gerados através do modelo logito limitado em decorrência da baixa associação entre a variável resposta e a covariável.

De acordo com os resultados obtidos não existe um modelo que melhor em todas as situações, no entanto, sabemos que a escolha do mesmo está diretamente relacionada com a dependência entre a variável resposta e as covariáveis. Em trabalhos futuros pretendemos desenvolver estudos para identificar qual modelo é melhor para cada situação. Além disso, é de grande interesse realizar estudos para verificar o comportamento destes modelos de classificação considerando diversos tamanhos

amostrais e também diferentes covariáveis.

De acordo com os resultados apresentados no Capítulo 8 notamos que o modelo logito com resposta de origem apresenta um desempenho bastante superior ao desempenho do modelo logito na estimação dos parâmetros dos modelos principalmento quando as amostras em questão são pequenas. A medida que aumentamos o tamanho das amostras a diferença da qualidade de ajuste entre os dois modelos diminui para todas as distribuições de origem estudadas.

No Capítulo 9 apresentamos uma análise de dados reais relativos a fraude bancária provenientes de uma instituição financeira. O modelo logito usual com estimadores KZ2 construídos em amostras balanceadas apresentou um desempenho preditivo ligeiramente superior aos demais modelos.

Referências Bibliográficas

[1] Cox,D. R. 1970."The analysis of Binary Data". Methuen, London.

[2] King, G. e Zeng, L. 2001. "Logistic Regression in Rare Events Data". Cambridge,MA: Harvard University.

[3] McCullagh, P. e J. A. Nelder, 1989. "Generalized Linear Models", 2nd ed. New York: Chapman and Hall.

[4] Stukel, T. A., 1988. "Generalized Logistic Models", Journal of Statitical Association, vol. 83, 426-431.

[5] Cramer, J. S., 2004. "Scoring bank loans that may go wrong", Statistica Neerlandica, vol. 58, 365-380.

[6] Suissa, S. e Blais, L., 1995. "Binary Regression with continous outcomes", Statistics in Medicine, vol. 14, 247-255.

[7] Diniz, C. A. R. e Paula, Marcelo,2011. "Regressão Logística binária com resposta pertencente a família exponencial", Des-UFSCar.

[8] Hosmer, W. e Lemeshow, S., 1989. "Applied Logistic Regression", John Wiley, New York.

[9] Geisser, S. 1993. "Predictive Inference: An Introduction", New York: Chapman and Hall.

[11] King, G. e Zeng, L. 2000a. "Inference in Case-Control Studies with Limited Auxilliary Information"(in press). (Preprint at http://Gking.harvard.edu.).

[12] Moraes, D. 2008. "Modelagem de fraude em cartaão de crédito", Dissertação de Mestrado, Des-UFSCar.

[13] Broyden, C. G. 1970. "The convergence of a class of double-rank minimization algorithms", partes I e II, Inst. Math. Appl., Malden, vol6, 76-90 e 222-231.

[14] Fletcher, R. 1970. "A new approach to variable metric algorithms", Comput. J., Oxford, vol 13, 317-322.

[14] Goldfarb, D. 1970. "A family of variable metric methods derived by variational means", Math Comp., Boston, vol 26, 23-26.

[15] Shano, D. F. 1970. "Conditioning of quasi-Newton methods for functions minimization", Math Comp., Boston, vol 24, 647-657.

[16] Prentice, R. L. 1976, "Generalization of the Probit and Logit Methods for Dose Response Curves", Biometrics, vol 32, 761-768.

[17] Pregibon, D. 1980, "Goodness of Link Tests for Generalized Linear Models", Applied Statitics, vol 29, 15-24 .

[18] Aranda-Ordaz, F. J. 1981, "On Two Families of Transformations to Additivity for Binary Response Data", Biometrika, vol 68, 357-363.

[19] Jorgensen, B. 1984, "The Delta Algorithm and GLIM", International Statistical Review, vol 52, 283-300.

[20] Stukel, T. A. 1985, "Implementation of an Algorithm for Fitting a Class of Generalized Logistic Models", Generalized Linear Models Conference Proceedings, Spring-Verlag, 160-167.

[21] Akaike, H. 1974, "A new look at the statistical model identification", IEEE Transactions on Automatic Control, Boston, vol 19, 716-723.

[22] Schwarz, G. 1978, "Estimating the dimensional of a model", Annals of Statistics, vol 6, 461-464.

[23] Zwlig, M. H. e Campbell, G. 1993, "Receiver-operating characteristic (ROC) plots", clin. Chem., vol 29, 561-577.

[24] Guirado, L. 2010. "Comparação do desempenho de Modelos Lineares Generalizados (MLG) e Modelos Aditivos Generalizados (MAG) na predição de dados Financeiros em credit score", 2010", Dissertação de Mestrado, Des-UFSCar.

[25] Linnet, K. e Brandt, E. 1986, "Assessing diagnostic tests once an optimal cutoff points has been selected", Clin. Chem.,vol 32, 1341-1346.

[26] Mathews, B. W. 1975, "Comparison of the predicted and observed secondary structure of t4 phage lysozyme", Biochim. Biophys. Acta., vol 405, 442-451.

[27] Baldi, P.; Brunak, S.; Chauvin, Y.; Andersen, C. A. F.; Nielsen, H. 2000, "Assessing the accuracy of prediction algorithms for classification: an overview", Bioinformatics 2000, 412-424.

[28] Kruskal, W. e Wallis, W. A. 1952, "Use of ranks in one-criterion variance analysis", Journal of American Statistical Association, 583-621.

[29] Mann, H. B. e Whitney, D. R. 1947, "On a Test of wheter one of Two Random Variables in stochastically larger than the other", Annals of Mathematical Statistics, vol 18, 50-60.

[30] Breiman, L. 1998, "Arcing classifiers", The Annals of Statistics, vol 26, 801-849.

Apêndice A

O Método BFGS

Neste Apêndice apresentamos o código em R utilizado no ajuste no modelo logito generalizados para amostras com seis covariáveis.

```
###Ajuste do Modelo Logito Generalizado



L1=function(theta){

beta0=theta[1]
beta1=theta[2]
beta2=theta[3]
beta3=theta[4]
beta4=theta[5]
beta5=theta[6]
beta6=theta[7]

eta=beta0+beta1*dados[,2]+beta2*dados[,3]+beta3*dados[,4]
+beta4*dados[,5]+beta5*dados[,6]+beta6*dados[,7]

h=(1/k)*log(1-k*abs(eta))
pi=exp(h)/(1+exp(h))
vero=sum(log((dbinom(dados[,1], 1, pi ))))
```

```
return(-vero)
}


#####


L2=function(theta){

beta0=theta[1]
beta1=theta[2]
beta2=theta[3]
beta3=theta[4]
beta4=theta[5]
beta5=theta[6]
beta6=theta[7]

eta=beta0+beta1*dados[,2]+beta2*dados[,3]+beta3*dados[,4]
+beta4*dados[,5]+beta5*dados[,6]+beta6*dados[,7]

h=-(1/k)*(exp(k*abs(eta))-1)
pi=exp(h)/(1+exp(h))
vero=sum(log((dbinom(dados[,1], 1, pi ))))
```

```
return(-vero)
}

ajuste <- glm(dados[,1] ~ dados[,2] + dados[,3] + dados[,4]
+dados[,5]+dados[,6]+dados[,7],
                family = binomial)

val_in=ajuste$coefficients

alpha1=seq(-3,-0.1,by=0.1)
m1=numeric(length(alpha1))
estimativa1=matrix(0,300,8)

alpha2=seq(0.01,3,by=0.01)
m2=numeric(length(alpha2))
estimativa2=matrix(0,300,8)



for(i in 1:length(alpha1)){

k=alpha1[i]

ajuste1=optim(val_in,L1,method="BFGS")

est=numeric(8)
```

```
est[1]=k

est[2:8]=ajuste1$par

m1[i]=-ajuste1$value

estimativa1[i,]=est

}

for(i in 1:length(alpha2)){

k=alpha2[i]

ajuste2=optim(val_in,L2,method="BFGS")

est=numeric(8)

est[1]=k

est[2:8]=ajuste2$par

m2[i]=-ajuste2$value

estimativa2[i,]=ajuste2$par
```

```
}



m=numeric(length(m1)+length(m2))
m[1:length(m1)]=m1
m[(length(m1)+1):length(m)]=m2
estimativa=rbind(estimativa1,estimativa2)
ind=which(m==max(m))
est=estimativa(ind,)
```