Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

Comparação dos Modelos Lineares Generalizados Logístico e Log-Binomial

Rodrigo Ferrari Lucas Lassance

Trabalho apresentado ao Departamento de Estatística da Universidade de Brasília para a obtenção do título de Bacharel em Estatística.

Brasília
2015

Rodrigo Ferrari Lucas Lassance

Comparação dos Modelos Lineares Generalizados Logístico e Log-Binomial

Trabalho apresentado ao Departamento de Estatística da Universidade de Brasília para a obtenção do título de Bacharel em Estatística.

Orientador: Prof. Dr. **Bernardo Borba de Andrade**

Brasília
2015

Agradecimentos

AGRADECIMENTOS

Agradeço primeiramente à minha família por ter me dado a oportunidade de chegar onde cheguei. Agradeço em especial à minha mãe, que me ajudou com minhas dificuldades sempre que pôde e que sempre me incentivou, e ao meu pai Marcelo, que se deu ao trabalho de revisar o texto e me ajudar a melhorá-lo.

Agradeço também aos meus colegas que conheci na UnB durante o curso, em especial ao Mateus e à Geiziane por me darem companhia e apoio sempre que possível.

Agradeço à minha namorada Beatriz por ter lido o texto e me ajudado assistindo e dando sugestões para minha apresentação.

Por último, agradeço ao departamento de estatística por ter me proporcionado a oportunidade de me tornar um estatístico. Em especial, ao professor doutor Bernardo por me orientar e aos professores doutores George e Joanlise por fazerem parte da minha banca.

Resumo

Este trabalho se destinou a buscar formas de comparar os modelos de regressão Logística e Log-Binomial, estabelecendo qual seria o mais adequado para um determinado conjunto de dados. Consequentemente buscou-se estabelecer qual das medidas, razão de chances ou risco relativo, seria mais adequada para o mesmo conjunto. Para isso, foram avaliadas algumas medidas por meio de simulações e em dados reais (especialmente por meio de estatísticas de diagnóstico), constatando-se que as mais indicadas para esse tipo tarefa foram a *deviance* e a raiz do erro quadrático médio.

Também foi avaliada a afirmação em Blizzard e Hosmer [2] de que a estatística de Hosmer-Lemeshow segue uma qui-quadrado com oito graus de liberdade quando os dados são provenientes e uma Log-Binomial e os dados são divididos em dez grupos. A afirmação se mostrou duvidosa e o teste teve baixo poder nos casos analisados, que consistiram no uso da função *link* incorreta e algumas transformações da covariável contínua.

Sumário

Lista de Figuras

Lista de Tabelas

Introdução

No meio médico existe um debate acerca do uso das medidas de associação razão de chances e risco relativo, havendo diferentes razões para se definir a preferência de um em detrimento do outro. Além disso, são recorrentes os casos de erros de interpretação sobre a razão de chances, enquanto a do risco relativo é mais intuitiva, tanto para o profissional da área quanto para o público leigo. A razão de chances é uma medida que pode ser obtida diretamente pela Regressão Logística (RL), enquanto o risco relativo é obtido diretamente pela Regressão Log-Binomial (RLB).

A intenção do presente trabalho é buscar estabelecer um critério puramente estatístico para a escolha do modelo a ser utilizado e, consequentemente, observar qual medida, razão de chances ou risco relativo, seria a mais indicada para diferentes situações.

A RL e a RLB surgiram na Estatística com o intuito de lidar com variáveis de resposta dicotômica, mas de maneiras diferenciadas. É comum o uso desse tipo de variável quando se está interessado em um dado "sucesso", que seria uma observação, ou um conjunto de observações específicas agrupadas em uma categoria, em relação a uma "falha", que seria simplesmente o que não é dado como sucesso. Em várias áreas do conhecimento, entre elas a médica e a das engenharias, há um interesse em certas medidas provenientes desse tipo de variável.

O risco relativo é muitas vezes calculado de forma indireta e imprecisa. Como exemplo de casos assim há a aproximação via regressão log-Poisson e outros, tal como revisado em Andrade [1]. A RLB, por outro lado, fornece estimativas adequadas e de forma direta, mas é um modelo que possui restrições em seus parâmetros, o que por muito tempo foi um agente desestimulador de seu uso. Ainda assim, já foram elaboradas maneiras de contornar esse problema e fornecer um fácil acesso a recursos computacionais que permitem a estimação dos parâmetros. Isso inclusive para um público com pouco conhecimento na área de programação, o que serve como estímulo ao aumento do uso da RLB nos tempos mais recentes.

Este trabalho trata do uso de métodos de comparação entre a RL e a RLB. Tais comparações são feitas de forma a determinar, com base exclusivamente em princípios estatísticos, qual dos modelos seria o mais adequado para os dados. Tal escolha definiria qual das medidas seria a mais válida para um determinado banco de dados, o risco relativo

ou a razão de chances. A pesquisa durou dez meses, abrangendo o período de março até dezembro.

É importante ressaltar o fato de que não existe na literatura atual nenhum estudo com foco na escolha exclusiva entre os dois modelos com base em critérios estatísticos. Considerou-se realizar o método comparação de *links* de Pregibon [9], que fornece famílias de *links* que permitem uma certa comparação a partir do uso de *deviances*. Essa alternativa não se mostrou interessante, dado que a intenção é comparar dois modelos de forma exclusiva, mas o método fornece uma comparação entre uma família de modelos, tendo esses dois como casos específicos. Além disso, a comparação seria confusa, visto que seria preciso ser obtida a distribuição de probabilidade dessa família para uma comparação mais válida por *deviance*.

Neste trabalho, foram utilizadas duas formas de comparação entre os modelos:

1. Medidas de ajuste, como *deviance*, erro quadrático médio e erro de classificação;

2. Diagnóstico dos resíduos.

Inicialmente realizou-se uma comparação direta de *deviances*, observando se o fato de uma *deviance* ser maior que a outra fornece um critério válido para diferenciar qual seria o modelo mais apropriado. Nesse caso, o modelo mais indicado seria aquele que possuísse menor *deviance*.

Em adição, comparou-se a raiz dos erros quadráticos médios (REQM) dos dois modelos. O modelo com menores valores seria mais adequado para a situação. Foram comparadas também proporções do erro de classificação (MEC), calculada pelo número de vezes em que a observação predita foi diferente da observação real e dividindo tal número pelo total de observações. O uso da REQM mostrou-se mais eficiente na classificação correta do modelo que a MEC, que não apresentou grandes diferenças no valor, mesmo quando o modelo era incorreto.

A análise de diagnóstico para uma variável foi realizada com base em Blizzard e Hosmer [2] e Hosmer e Lemeshow [5]. Para uma variável explicativa, foram separados quatro cenários, dois para a RL e dois para a RLB, e tomadas algumas medidas para se realizar o diagnóstico. Esse pareceu fornecer interpretações concordantes ao que se esperaria tomando-se como base as medidas de ajuste. Para duas variáveis explicativas, a análise de diagnóstico foi realizada para apenas um cenário, em que as medidas também demonstraram validez.

Blizzard e Hosmer [2] afirmaram que o teste de Hosmer-Lemeshow possui distribuição χ^2_8 quando o número de grupos é oito para o modelo RLB, mas conforme algumas das análises deste texto essa afirmação se mostrou questionável. Além disso, o poder do teste se mostrou insatisfatório, especialmente quando ele é usado para determinar qual o modelo seria mais adequado aos dados.

As medidas de associação mencionadas foram comparadas com base em dados simulados. As simulações do modelo RLB foram geradas a partir dos cenários apresentados no

artigo de Blizzard e Hosmer [2]. As simulações do modelo RL, por sua vez, foram geradas a partir de sugestões do professor orientador.

Ambas as simulações foram geradas com uma variável explicativa, que é contínua e uniforme. Posteriormente, foram realizadas simulações com duas variáveis explicativas, uma contínua e uma dicotômica. Finalmente, foram calculadas as medidas selecionadas em conjuntos de dados reais.

Capítulo 1

Metodologia

Para a realização das análises estatística foi utilizado o*software* R (versão 3.2.0). Todas as funções utilizadas no R referentes ao modelo Log-Binomial foram obtidas do artigo de Andrade [1].

As análises iniciais se deram através de dados simulados. Este trabalho começou com a tentativa do uso do método de comparação de *links* estabelecido por Pregibon [9]. A partir de uma família de *links* (em que a escolha de certos valores fornece um *link* específico, como por exemplo o da regressão Logística), o uso dos dados forneceria estimativas para esses valores, determinando então qual seria o *link* a ser utilizado.

Não se optou por seguir tal abordagem devido a dois aspectos. Primeiramente, o método mencionado compara um modelo inicial, neste caso RL ou RLB, à uma família de *links*. Mas o interesse deste trabalho é a comparação de apenas esses dois modelos, de forma exclusiva. O segundo aspecto envolve o problema de se determinar qual modelo seria significativamente melhor do que o modelo inicial, pois ainda não foi analisada de forma suficiente uma família de *links* que fornecesse essa informação. Visto que não foi encontrada uma família de *links* que contemplasse ambos os modelos e que possuísse uma distribuição de probabilidade já descoberta, optou-se por não se utilizar tal abordagem.

Então, optou-se pela comparação direta entre *deviances*. A ideia consiste em se obter o valor da *deviance* de um conjunto de dados para ambos os modelos e observar qual das *deviances* obtidas é menor. Assim, o modelo escolhido é aquele que possuir a menor *deviance*.

Outra forma de comparação foi através da raiz do erro quadrático médio, seguindo uma lógica semelhante à comparação por *deviances*. Seria obtida a raiz do erro quadrático médio de ambos os modelos e observado em qual modelo essa é menor, sendo então classificado como o modelo correto.

Utilizou-se ainda a proporção (ou média) do erro de classificação. Obtendo as probabilidades estimadas dos modelos, elas seriam utilizadas para determinar valores preditos, que seriam comparados com os valores verdadeiros. A proporção seria a razão entre o número de vezes em que os valores preditos não foram iguais aos valores verdadeiros e o número total de termos. A proporção que fosse menor indicaria qual modelo é mais

adequado.

Para a análise de diagnóstico, foram utilizadas medidas baseadas no resíduo de Pearson e no resíduo *deviance*, além de uma aproximação da estatística $\Delta\beta$. O teste de Hosmer-Lemeshow foi aplicado com certa ressalva por Hosmer [4] ter mostrado que o teste tem poder consideravelmente baixo para a RL e por análises feitas neste trabalho.

Simulações foram implementadas tanto para dados que têm sua origem no modelo Log-Binomial quanto no modelo Logístico. Supõe-se o desconhecimento desta origem e procura-se identificá-las por meio das medidas de comparação descritas acima.

As simulações para o modelo RLB foram obtidas com base no artigo de Blizzard e Hosmer [2], com a ideia de que as curvas da função logística e da RLB se diferenciam mais conforme o aumento da probabilidade condicional. Sendo assim, para o caso de uma variável explicativa foram simulados dados do modelo RLB que fornecem uma probabilidade condicional mais alta (acima de 90%) e outros que fornecessem probabilidade condicional menor (em torno de 50%). Similarmente, com base em sugestões do professor orientador, foram simulados dados para o modelo RL com parâmetros que permitissem uma boa diferenciação entre as curvas. No caso de duas variáveis explicativas para a RLB, todos os cenários apresentaram probabilidade condicional em torno de 90%.

Capítulo 2

Revisão de Literatura

Este capítulo descreve as diferentes ferramentas estatísticas utilizadas ao longo do trabalho. O termo n é associado ao tamanho da amostra obtida. Quando se trata de uma variável aleatória, letras em maiúsculo se referem à variável, enquanto as mesmas letras em minúsculo se referem às observações desta variável. Ressalta-se mais uma vez que não há literatura sobre a comparação de *links*.

2.1 Modelos Lineares Generalizados

Paula [8] apresenta uma definição de modelos lineares generalizados, e sua notação será seguida.

Considera-se $\mathbf{Y} = (Y_1, Y_2, \cdots, Y_n)$ um vetor de variáveis aleatórias com função de densidade de probabilidade pertencente à família exponencial, ou seja,

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

em que ϕ é uma constante, θ_i é uma função que depende apenas dos parâmetros em uma dada observação, $b(\theta_i)$ é uma função de θ_i e $c(y_i, \theta_i)$ é uma função de y_i e θ_i.

Além disso, um modelo linear generalizado possui uma parte sistemática, um preditor linear e uma função de ligação (ou *link*). Tais elementos são descritos a seguir:

- Parte sistemática: $g(\mu_i) = \eta_i$;
- Preditor linear: $\eta_i = \mathbf{x}_i'\boldsymbol{\beta}$, sendo $\boldsymbol{\beta} = (\beta_1\ \beta_2\ \cdots\ \beta_p)'$, $p < n$ o vetor coluna de parâmetros e $\mathbf{x}_i = (x_{i,1}\ x_{i,2}\ \cdots\ x_{i,p})'$ os valores das variáveis explicativas, também conhecidas como covariáveis;
- Função de ligação: $g(\cdot)$.

Pode-se mostrar que a distribuição de Bernoulli, que é a distribuição assumida para a variável resposta, pertence à família exponencial. Sua função distribuição de probabilidade

é dada por:

$$f(y_i; p) = p^{y_i}(1-p)^{1-y_i} = \exp\left(y_i \log\left(\frac{p}{1-p}\right) + \log(1-p)\right).$$

Portanto, $\phi = 1$, $\theta_i = \log\left(\frac{p}{1-p}\right)$, $b(\theta_i) = \log(1+e^{\theta_i})$ e $c(y_i, \phi) = 0$.

Para os modelos RL e RLB, as especificações da parte sistemática, do preditor linear e da função de ligação constam em suas respectivas seções.

2.2 Modelo Logístico

O modelo Logístico surge como uma forma de se trabalhar com variáveis dicotômicas sob o ponto de vista da regressão. Hosmer e Lemeshow [5] explicitam a importância do uso da RL ao trabalhar com variáveis resposta dicotômicas, além de uma boa exploração de suas características e formas de análise. O conteúdo desta seção se baseia no que foi expresso no trabalho.

Em contraste com a regressão linear tradicional, em que se estima o próprio valor da variável resposta, o modelo RL fornece a probabilidade estimada da ocorrência de um evento quando considerado em função de covariáveis, o que permite uma exploração mais profunda de um problema em questão. $\pi(\mathbf{x}_i)$ representa a probabilidade da ocorrência daquilo que se está considerando como sucesso (e ao qual normalmente é atribuído o valor 1) na observação i, condicionada às covariáveis $\mathbf{x}_i = (x_{i,0}\ x_{i,1}\ \cdots\ x_{i,p})'$, sendo que $x_{i,0} = 1, \forall i$.

Sendo $\boldsymbol{\beta} = (\beta_0\ \beta_1\ \cdots\ \beta_p)'$ o vetor coluna dos p parâmetros do modelo, $\pi(\mathbf{x}_i)$ pode ser obtido por meio da função logística

$$\pi(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{1 + e^{\mathbf{x}_i'\boldsymbol{\beta}}}.$$

A sua função *link* é dada pelo logito e permite a obtenção de uma função linear, ou seja,

$$\text{logito}(\pi(\mathbf{x}_i)) = \log\left(\frac{\pi(\mathbf{x}_i)}{1-\pi(\mathbf{x}_i)}\right) = \log\left(\frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}/(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})}{1-e^{\mathbf{x}_i'\boldsymbol{\beta}}/(1+e^{\mathbf{x}_i'\boldsymbol{\beta}})}\right) = \log(e^{\mathbf{x}_i'\boldsymbol{\beta}}) = \mathbf{x}_i'\boldsymbol{\beta}.$$

Sendo Y_i a variável resposta de interesse na observação i, então é possível se concluir que a variável Y_i possui distribuição de Bernoulli com parâmetro $\pi(\mathbf{x}_i)$, pois a variável resposta é apenas uma (a observação i) e possui dois valores possíveis. E sendo $f_{Y_i}(y_i)$ a função de distribuição de probabilidade da Bernoulli, então $f_{Y_i}(y_i) = \pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{1-y_i}$.

2.2.1 Estimação dos parâmetros por máxima verossimilhança

A função do logaritmo da verossimilhança da variável resposta, para uma amostra de n observações, será dada por:

$$L(\pi(\mathbf{x}_i)|y_i) = \prod_{i=1}^{n} f_{Y_i}(y_i) = \prod_{i=1}^{n} \pi(\mathbf{x}_i)^{y_i}(1-\pi(\mathbf{x}_i))^{1-y_i}.$$

Consequentemente, o logaritmo da função é expresso por:

$$\log(L(\pi(\mathbf{x}_i)|y_i)) = l(\pi(\mathbf{x}_i)|y_i) = \sum_{i=1}^{n}[y_i \log(\pi(\mathbf{x}_i)) + (1-y_i)\log(1-\pi(\mathbf{x}_i))]. \qquad (2.1)$$

A forma mais comum de estimação dos parâmetros se dá por meio da maximização da verossimilhança. São selecionados os valores para o vetor de parâmetros $\boldsymbol{\beta}$ que fornecem o maior valor possível para a equação (2.1). Os parâmetros são obtidos derivando a função de log-verossimilhança em função de cada um dos parâmetros e igualando a 0. Sendo j um valor inteiro,

$$\frac{\partial l(\pi(\mathbf{x}_i)|y_i)}{\partial \beta_j} = \begin{cases} \sum_{i=1}^{n}(y_i - \pi(\mathbf{x}_i)) & : j = 0 \\ \sum_{i=1}^{n} x_{ij}(y_i - \pi(\mathbf{x}_i)) & : 1 \leq j \leq p \end{cases}.$$

Em termos vetoriais,

$$\frac{\partial l(\pi(\mathbf{x}_i)|y_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i(y_i - \pi(\mathbf{x}_i)).$$

Obtidas as derivadas da função, os parâmetros são obtidos numericamente.

É vantajoso se estimar por verossimilhança principalmente por duas características muito importantes:

- Propriedade de invariância: Sendo $\hat{\boldsymbol{\beta}}$ o estimador de máxima verossimilhança de $\boldsymbol{\beta}$, então, para uma função de $\boldsymbol{\beta}$, $g(\boldsymbol{\beta})$, o estimador de máxima verossimilhança de $g(\boldsymbol{\beta})$ é justamente $g(\hat{\boldsymbol{\beta}})$. Ou seja, $\hat{\pi}(\mathbf{x}_i) = \frac{e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}{1+e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}}$ é o estimador de máxima verossimilhança de $\pi(\mathbf{x}_i)$;

- Distribuição assintótica: Os estimadores de máxima verossimilhança possuem a propriedade de que quando o número de observações na amostra tende a infinito e sob certas condições de regularidade (especificadas em Casella e Berger [3]), seguem a distribuição normal, $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, I_F^{-1}(\boldsymbol{\beta}))$. $I_F(\boldsymbol{\beta})$ representa a matriz de informação de Fisher, aproximada pela matriz de informação de Fisher observada ($I_F(\hat{\boldsymbol{\beta}})$), dada por:

$$I_F(\hat{\boldsymbol{\beta}}) = -\left(\frac{\partial^2 l(\pi(\mathbf{x}_i)|y_i)}{\partial\boldsymbol{\beta}\partial\boldsymbol{\beta}}\right)\Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = \sum_{i=1}^{n} \mathbf{x}_i\mathbf{x}_i'\hat{\pi}(\mathbf{x}_i)(1-\hat{\pi}(\mathbf{x}_i)).$$

Ou seja, deriva-se duas vezes a função de log-verossimilhança e então os valores de $\boldsymbol{\beta}$ são substituídos por seus estimadores de máxima verossimilhança, $\hat{\boldsymbol{\beta}}$.

2.2.2 Interpretação dos parâmetros

Antes de se falarsobre a interpretação dos parâmetros, atenta-se para a necessidade de que unidade de mudança será definida para uma variável independente (covariável), pois o tamanho dos valores fornecidos por cada variável pode diferir consideravelmente. A partir disso, deve-se então definir dois valores da covariável para comparação e utilizá-los na equação do logito para a estimação da razão de chances.

2.2.3 Razão de Chances

Considerando duas observações diferentes, $\mathbf{x}_i$ e $\mathbf{x}_j$, a razão de chances (RC) é definida como:

$$RC = \frac{\pi(\mathbf{x}_i)/(1-\pi(\mathbf{x}_i))}{\pi(\mathbf{x}_j)/(1-\pi(\mathbf{x}_j))}.$$

Tal medida indica o quão maior ou menor é a chance de sucesso do evento no numerador com relação ao evento no denominador, sendo a chance a exponencial do logito.

Como exemplo, caso houvesse uma variável explicativa x_1 e ela fosse dicotômica, então seria

$$RC = \frac{\pi(1)/(1-\pi(1))}{\pi(0)/(1-\pi(0))} = \frac{e^{\beta_0+\beta_1}}{e^{\beta_0}} = e^{\beta_1}.$$

O mesmo valeria caso houvesse várias variáveis explicativas sendo mantidas constantes, e apenas a variável explicativa dicotômica tendo seu valor alterado. Num caso mais geral, para uma variável contínua cujos valores selecionados foram a e b, a RC seria $RC = e^{(a-b)\beta_1}$.

Reitera-se aqui que a interpretação da razão de chances precisa ser em função de chances, não de probabilidades. Existem casos em que é possível uma interpretação aproximada em função da probabilidade, mas isto será melhor explicado na subseção referente ao risco relativo.

Pela propriedade de invariância dos estimadores de máxima verossimilhança, o estimador de máxima verossimilhança de RC é dado por $\widehat{RC} = \frac{\hat{\pi}(\mathbf{x}_i)/(1-\hat{\pi}(\mathbf{x}_i))}{\hat{\pi}(\mathbf{x}_j)/(1-\hat{\pi}(\mathbf{x}_j))}$.

2.3 Modelo Log-Binomial

A RLB é utilizada para o mesmo tipo de variável resposta que a RL. As informações presentes nessa parte foram obtidas principalmente de Blizzard e Hosmer [2], além de Andrade [1].

Mais uma vez, o modelo busca a estimação da probabilidade de sucesso de um determinado evento com base nas covariáveis analisadas. A notação utilizada será semelhante

à notação para a RL, visto que há muitos aspectos similares. A probabilidade de sucesso $\pi(\mathbf{x}_i)$ assume uma nova forma, dada por

$$\pi(\mathbf{x}_i) = e^{\mathbf{x}_i'\boldsymbol{\beta}}.$$

Consequentemente, a função *link* é mais simples que o caso do modelo RL, bastando obter-se o logaritmo da probabilidade, de modo que

$$\log(\pi(\mathbf{x}_i)) = \log(e^{\mathbf{x}_i'\boldsymbol{\beta}}) = \mathbf{x}_i'\boldsymbol{\beta}.$$

2.3.1 Restrição do espaço paramétrico

É importante notar que, com a probabilidade de sucesso $\pi(\mathbf{x}_i)$ definida desta forma, estimativas do vetor de parâmetros $\boldsymbol{\beta}$ não podem mais assumir qualquer valor, pois dado x_i, $\boldsymbol{\beta}$ precisa satisfazer a relação

$$\pi(\mathbf{x}_i) \leq 1 \Rightarrow \log(e^{\mathbf{x}_i'\boldsymbol{\beta}}) \leq \log(1) \Rightarrow \mathbf{x}_i'\boldsymbol{\beta} \leq 0,$$

ou seja, impõe-se aos parâmetros uma restrição linear.

2.3.2 Estimação por máxima verossimilhança

A equação da função de verossimilhança é exatamente a mesma do caso da RL, bastando substituir o valor da probabilidade $\pi(\mathbf{x}_i)$. A diferença se dá no processo de derivação em função dos parâmetros como a seguir:

$$\frac{\partial l(\pi(\mathbf{x}_i)|y_i)}{\partial \boldsymbol{\beta}} = \sum_{i=1}^{n} \mathbf{x}_i \frac{y_i - \pi(\mathbf{x}_i)}{1 - \pi(\mathbf{x}_i)}.$$

Pode-se obter também a distribuição assintótica dos estimadores de máxima verossimilhança usando a inversa da matriz informação de Fisher, mas somente se $bm\beta$ estiver no interior do espaço paramétrico (Andrade [1]). Sendo $\hat{\pi}(\mathbf{x}_i) = e^{\mathbf{x}_i'\hat{\boldsymbol{\beta}}}$, a matriz de informação de Fisher observada é expressa por:

$$I_F(\hat{\boldsymbol{\beta}}) = -\left(\frac{\partial^2 l(\pi(\mathbf{x}_i)|y_i)}{\partial \boldsymbol{\beta} \partial \boldsymbol{\beta}} \right) \Bigg|_{\boldsymbol{\beta}=\hat{\boldsymbol{\beta}}} = -\sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \frac{\hat{\pi}(\mathbf{x}_i)(y_i - 1)}{(1 - \hat{\pi}(\mathbf{x}_i))^2}.$$

O problema da estimação dos parâmetros, no caso do modelo RLB, surge devido à restrição paramétrica. Não é possível se realizar a mesma estimação por máxima verossimilhança que foi realizada numericamente no caso do modelo RL Portanto, é necessário se recorrer a algum outro meio.

2.3.3 Estimação dos parâmetros

Andrade [1] descreve o uso do algoritmo conhecido como barreira adaptativa para a estimação de parâmetros do modelo RLB. Esse algoritmo impede que as estimativas dos parâmetros saiam do espaço paramétrico durante as iterações do algoritmo, fornecendo estimativas mais próximas dos verdadeiros valores quando estes estão próximos da região de restrição.

2.3.4 Interpretação dos parâmetros

Mais uma vez, chama-se a atenção para a necessidade de se observar qual unidade de mudança será usada de uma das variáveis independentes. No caso da RLB, o risco relativo é estimado diretamente, enquanto que para a RL o RC é estimada diretamente.

2.3.5 Risco Relativo

Sendo $\mathbf{x}_i$ e $\mathbf{x}_j$ referentes a duas observações diferentes de um mesmo conjunto de variáveis, o risco relativo (RR) é definido como:

$$RR = \frac{\pi(\mathbf{x}_i)}{\pi(\mathbf{x}_j)}.$$

Como se pode observar, $RR \simeq RC$ nos casos em que $(1 - \pi(\mathbf{x}_i))/(1 - \pi(\mathbf{x}_i)) \simeq 1$, ou seja, a razão de chances e o risco relativo são próximos quando a probabilidade de sucesso é pequena. Isso permite que a razão de chances seja interpretada de forma aproximada ao risco relativo, mas apenas neste caso.

O risco relativo é interpretado em função de probabilidades. Tal medida fornece a informação de quão mais ou menos provável é o sucesso de uma observação em detrimento da outra. Isso permite uma interpretação muito mais intuitiva da medida, ao contrário da RC, que é interpretada em função das chances.

A RLB fornece uma maneira simples de obtenção desta medida:

$$RR = \frac{\pi(\mathbf{x}_i)}{\pi(\mathbf{x}_j)} = \frac{e^{\mathbf{x}_i'\boldsymbol{\beta}}}{e^{\mathbf{x}_j'\boldsymbol{\beta}}} = e^{\mathbf{x}_i'\boldsymbol{\beta} - \mathbf{x}_j'\boldsymbol{\beta}} = e^{(\mathbf{x}_i' - \mathbf{x}_j')\boldsymbol{\beta}}.$$

Como exemplo, havendo apenas uma variável explicativa x_1, dicotômica, o RR é dado por

$$RR = \frac{\pi(1)}{\pi(0)} = e^{\beta_1}. \tag{2.2}$$

Caso x_1 fosse uma variável contínua, escolhendo-se valores quaisquer a e b, ter-se-ia $RR = e^{(a-b)\beta_1}$. Comumente, considera-se que $a - b = 1$, obtendo a expressão (2.2). Porém, ressalta-se que, dependendo da variável que se está analisando, a alteração em uma

unidade pode não fornecer uma informação que seja interessante, fazendo-se necessário o uso de uma diferença maior ou menor.

Mais uma vez, pela propriedade de invariância, o estimador de máxima verossimilhança de RR é dado por $\widehat{RR} = e^{(\mathbf{x}_i' - \mathbf{x}_j')\hat{\boldsymbol{\beta}}}$.

2.4 Medidas de Ajuste e Diagnóstico para Variável Resposta Dicotômica

A seguir, seguem algumas das medidas propostas por Hosmer e Lemeshow [5] com o intuito de avaliar a qualidade do ajuste de um dos modelos com relação aos dados binários.

Ressalta-se que, em muitas das medidas, é importante a separação dos diferentes padrões de covariável, que seriam os casos em que todas as covariáveis possuem o mesmo valor, mas não necessariamente valor igual para a variável resposta. Assim, aos padrões iguais é designada a notação $\mathbf{x}_j$, em que j vai de 1 a J, sendo J o número de padrões distintos. Os padrões de covariável devem ser levados em consideração especialmente quando J é muito menor do que n.

Sendo m_j o número de observações em que $\mathbf{x} = \mathbf{x}_j$, existem duas teorias diferentes que tratam do problema de diagnóstico para variável resposta dicotômica, que são:

m-assintótica: Teoria em que se mantém o número dos padrões de covariável constantes enquanto se aumenta o tamanho da amostra;

n-assintótica: Teoria em que o número dos padrões de covariável aumenta conforme aumenta o tamanho da amostra.

2.4.1 Teste de Hosmer-Lemeshow

O teste de Hosmer-Lemeshow, também conhecido como teste dos decis-de-risco, se trata de um teste que compara a frequência da ocorrência das observações com as estimações teóricas, obtidas através do modelo ajustado. Consiste em realizar um ordenamento das observações e dos valores ajustados em função dos valores ajustados. As observações ordenadas são então separadas em g grupos, sendo que comumente $g = 10$.

Também é possível se realizar a separação dos grupos de outro modo (separando em um grupo as observações cujas probabilidades estimadas estão entre 0 e 0.1, em outro grupo as que estão entre 0.1 e 0.2 e assim sucessivamente até o grupo com probabilidades estimadas entre 0.9 e 1), mas há o risco de um determinado grupo conter poucas observações e esse não foi o critério utilizado no trabalho.

A estatística do teste é

$$\hat{C} = \sum_{j=0}^{1} \sum_{k=1}^{g} \frac{(o_{jk} - \hat{e}_{jk})^2}{\hat{e}_{jk}},$$

sendo o_{1k} o número de sucessos no k-ésimo grupo, o_{0k} o número de fracassos, $\hat{e}_{1k}$ a soma dos valores ajustados para o k-ésimo grupo e $\hat{e}_{0k}$ a soma de valor ajustado subtraindo 1.

A hipótese nula analisada é a de que o modelo se ajusta bem aos dados. Para a RL, $\hat{C} \sim \chi^2_{g-2}$ quando o modelo correto é ajustado (resultado obtido por simulação). Para a RLB, Blizzard e Hosmer [2] afirmam que a distribuição é a mesma.

2.4.2 Matriz "Chapéu" e Valores de *Leverage*

A matriz chapéu para modelos de variável resposta dicotômica é obtida através de uma aproximação linear para os valores ajustados. Sendo $\mathbf{V}$ a matriz diagonal J X J de termo geral

$$v_j = m_j\hat{\pi}(\mathbf{x}_j)[1 - \hat{\pi}(\mathbf{x}_j)];\ j = 1, 2, \cdots, J$$

quando os dados são da RL e

$$v_j = \frac{\hat{\pi}(\mathbf{x}_j)}{1 - \hat{\pi}(\mathbf{x}_j)};\ j = 1, 2, \cdots, J$$

quando os dados são da RLB. A matriz chapéu, denotada por $\mathbf{H}$, é dada por

$$\mathbf{H} = \mathbf{VX(X'VX)}^{-1}\mathbf{X'}.$$

Desse modo, os valores da diagonal, conhecidos como valores *leverage* e denotados por h_j, possuem limite superior de $1/m_j$, o que permite uma noção de distância. Ainda assim, tal noção só é possível no intervalo de probabilidades estimadas entre 0,1 e 0,9.

2.4.3 Medidas Baseadas no Resíduo de Pearson

O resíduo de Pearson é uma medida simples se para avaliar a diferença entre os valores ajustados e os preditos. Sua fórmula é dada por

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

Como forma de se resumir a informação contida nos resíduos, tem-se a estatística resumo baseada em Pearson, chamada de qui-quadrado de Pearson, dada por

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

Sob m-assintótica, é sabido que $\mathbf{X}^2 \sim \chi^2_{J-(p+1)}$.

Hosmer e Lemeshow [5] mostram que o resíduo de Pearson não possui variância 1 e

sugerem como se obter o resíduo de Pearson padronizado para o padrão $\mathbf{x}_j$, expresso por

$$r_{Sj} = \frac{r_j}{\sqrt{1-h_j}}.$$

Uma outra medida, que representa o decrescimento no valor do qui-quadrado de Pearson quando se retira o padrão $\mathbf{x}_j$, é

$$\Delta\mathbf{X}_j^2 = \frac{r_j^2}{1-h_j} = r_{sj}^2.$$

Tal medida é responsável por indicar um ajuste ruim do padrão da covariável $\mathbf{x}_j$.

2.4.4 Medidas Baseadas no Resíduo Deviance

O resíduo *deviance* é dado por

$$d_j = d(y_j, \hat{\pi}_j) = \sqrt{2m_j|\log(\hat{\pi}_j)|},$$

quando $y_j = m_j$ e

$$d_j = -\sqrt{2m_j|\log(1-\hat{\pi}_j)|},$$

quando $y_j = 0$. Seu propósito é o mesmo do resíduo de Pearson.

A estatística resumo baseada na *deviance* é dada por

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

Para o caso da *deviance*, uma medida equivalente ao $\Delta\mathbf{X}^2$ é $\Delta\mathbf{D}_j$, dada por

$$\Delta\mathbf{D}_j = d_j^2 + \frac{r_j^2 h_j}{1-h_j}.$$

A mesma distribuição de $\Delta\mathbf{X}_j^2$ se aplica para $\Delta\mathbf{D}_j$ sob m-assintótica.

2.4.5 Diferença Padronizada Entre β e $\beta_{\text{-j}}$

Outra medida, criada com o intuito de se observar a influência de um determinado padrão de covariável na estimação dos parâmetros do modelo, é a diferença padronizada entre $\boldsymbol{\beta}$ e $\boldsymbol{\beta}_{-j}$, sendo $\boldsymbol{\beta}_{-j}$ os valores estimados dos betas quando o padrão $\mathbf{x}_j$ é retirado. Tal medida é denotada por $\Delta\hat{\beta}_j$ e, segundo Hosmer e Lemeshow [5], seu valor pode ser aproximado por

$$\Delta\hat{\beta}_j = \frac{r_j^2 h_j}{(1-h_j)^2} = \frac{r_{Sj}^2 h_j}{1-h_j}.$$

2.5 *Deviance*

A *deviance*, denotada por D, é definida por

$$D = -2\log(L(\boldsymbol{\beta}|Y)) = -2l(\boldsymbol{\beta}|Y).$$

2.5.1 Teste de Razão de Verossimilhança - TRV

A aplicação mais marcante dessa medida se dá pelo Teste da Razão de Verossimilhança (TRV), que permite se analisar a significância de um ou mais parâmetros dentro de um modelo através da comparação de *deviances* de modelos que sejam aninhados. Se a diferença entre o número de parâmetros dos dois modelos é k, então a estatística do teste é dada por

$$TRV = D(modelo\ reduzido) - D(modelo\ mais\ completo) \sim \chi^2_k,$$

e a estatística do teste segue uma distribuição qui-quadrado com k graus de liberdade.

2.5.2 Critério de Informação de Akaike - AIC

A *deviance* também é utilizada no cálculo do Critério de Informação de Akaike (AIC), dado por:

$$AIC = 2p + D.$$

O AIC é utilizado para indicar qual o modelo que melhor se ajusta a um determinado conjunto de dados, qual pode ser a melhor função de distribuição de probabilidade para esse mesmo conjunto. Quanto menor for o valor do AIC, melhor o modelo estaria se adequando aos dados.

Existem outras medidas que realizam uma função similar, como o AIC corrigido e o BIC (critério de informação bayesiano). A diferença entre esses critérios é simplesmente o termo que se soma à *deviance*, que além do número de parâmetros também leva em conta o tamanho da amostra.

Vale notar que, neste trabalho, uma comparação entre os valores do AIC, de AICc ou de BIC são equivalentes a uma comparação entre *deviances*. Como o interesse é comparar a mesma amostra sendo modelada pela RL e pela RLB, mantendo-se as mesmas covariáveis, tanto o tamanho da amostra quanto o número de parâmetros são iguais. Usando o AIC de exemplo,

$$\begin{aligned} AIC(\text{RL}) - AIC(\text{RLB}) &= 2p + D(\text{RL}) - (2p + D(\text{RLB})) \\ &= D(\text{RL}) - D(\text{RLB}) \end{aligned} \tag{2.3}$$

Como mencionado anteriormente, essa medida se baseia nos dados amostrais. Por-

tanto, diferentes amostras da mesma população podem levar a diferentes conclusões sobre qual o modelo correto.

2.6 Raiz do Erro Quadrático Médio

O erro quadrático médio (EQM) é uma ferramenta comumente utilizada na análise residual. Sua fórmula é dada por:

$$EQM = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\pi}(\mathbf{x}_i))^2.$$

James et al. [6] fazem menção à essa medida como meio de mensurar a qualidade de um determinado ajuste aos dados. No caso em questão, o ajuste que está sendo feito é o da probabilidade estimada $\hat{\pi}(\mathbf{x}_i)$ aos dados observados y_i. Caso os valores preditos sejam próximos do valor observado, o erro quadrático médio será pequeno.

Normalmente, para se testar o ajuste, parte dos dados observados não são usados para a estimação do modelo. Isso se dá pelo fato de que o modelo é estimado com base na amostra, portanto é o melhor modelo para determinada amostra. Para amostras diferentes, modelos diferentes surgirão, acarretando que mesmo que o EQM da amostra seja pequeno, o ajuste pode ser ruim.

A raiz do erro quadrático médio (REQM) é obtida-se tomando a raiz quadrada do EQM. A raiz REQM é utilizada por permitir uma interpretação mais fácil.

2.7 Proporção do Erro de Classificação

O erro de classificação (EC) consiste em se obter o número de vezes em que os dados são erroneamente classificados quando se compara com os valores estimados pelo modelo (Hosmer e Lemeshow [5]).

Sendo $\hat{y}_i$ o valor predito para a observação, para se classificar uma observação predita como assumindo valor 0 ou 1, é necessário escolher um valor c tal que:

$$\hat{y}_i = \begin{cases} 0 & : \hat{\pi}(\mathbf{x}_i) < c \\ 1 & : \hat{\pi}(\mathbf{x}_i) \geq c \end{cases}.$$

O valor mais comumente assumido é $c = 0.50$.

Hosmer e Lemeshow [5] mencionam que utilizar a classificação dessa maneira para o caso do modelo RL não é muito adequado pelo fato de se limitar probabilidades a apenas dois valores fechados (no caso, 0 e 1). Ainda assim, pode ser uma medida apropriada quando o interesse maior está na classificação.

2.8 Simulações

As simulações realizadas foram feitas com base em Blizzard e Hosmer [2] e sugestões do professor orientador. A estimação dos parâmetros e de outros aspectos da RLB foram feitos com os códigos presentes em Andrade [1].

Foram realizadas simulações com uma e duas variáveis explicativas. Para uma variável, essa segue uma distribuição uniforme dentro do intervalo $[-6, a]$, sendo que o valor de a é escolhido dependendo do que se está interessado em realizar. Portanto, $X_i \sim U[-6, a]$ e $f(x_i) = \frac{1}{a+6}$. Para duas variáveis, a primeira é $D \sim$ Bernoulli(p) e a segunda é $U \sim$ Uniforme$(-6 + 2d, 2 + 2d)$, sendo d um dado valor de D.

Blizzard e Hosmer [2] tomam por hipótese para as simulações que as curvas dos modelos RL e RLB são bastante próximas em probabilidades estimadas inferiores a 0.5 e que apenas em valores superiores passam a divergir.

2.8.1 Simulações para o modelo Log-Binomial

Blizzard e Hosmer [2] apresentam um meio de simular dados da RLB que fornece probabilidades condicionais altas ($P(Y_i = 1|x = a) = \pi(a)$), ou mesmo probabilidades mais próximas de 0.5. No caso,

$$\pi(x_i) = e^{\beta_0 + \beta_1 x_i}.$$

A partir de informações sobre a probabilidade marginal de sucesso ($P(Y_i = 1)$) e da probabilidade condicional, elaboraram oito cenários, quatro dos quais serão utilizados neste trabalho. Em quatro cenários, as probabilidades de sucesso condicionadas a a foram valores superiores a 0.90, enquanto na outra metade ficaram próximas de 0.50.

As simulações provenientes do modelo RLB neste trabalho seguiram quatro passos:

1. Definir um valor para β_0, β_1 e a;
2. Simular X de uma Uniforme no intervalo $[-6, a]$;
3. Obter o valor das probabilidades $\pi(x_i) = e^{\beta_0 + \beta_1 x_i}$;
4. Simular Y_i com distribuição Bernoulli com probabilidade de sucesso $\pi(x_i)$.

Com os dados simulados, ajusta-se os modelos RL e RLBe então obtém-se as medidas de qualidade de ajuste mencionadas para ambos.

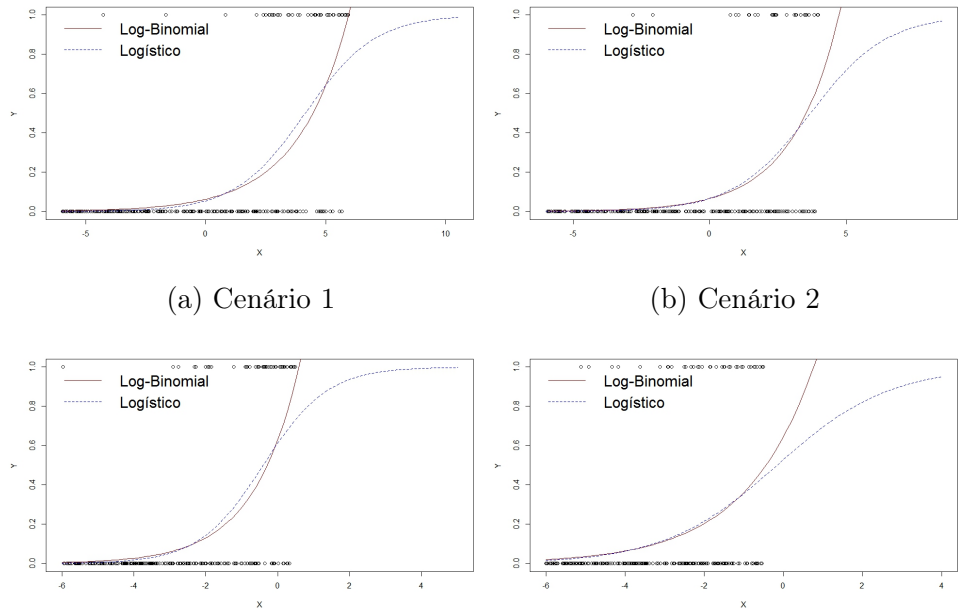
Observou-se que existe uma relação entre o valor máximo a da uniforme e os valores dos parâmetros β_0 e β_1. Quando $\beta_0 + \beta_1 a \simeq 0 \Rightarrow \pi(a) \simeq 1$, é possível mostrar que β_0 e β_1 estão na fronteira do espaço paramétrico, que pode acarretar em problemas de estimação. Blizzard e Hosmer [2] mencionam tal problema, buscando contorná-lo sem grande sucesso. Andrade [1] apresenta solução para esse problema por funções que constam em seu artigo.

A Tabela 2.1 apresenta os parâmetros e probabilidades dos cenários 1, 2, 7 e 8 de Blizzard e Hosmer [2], que foram utilizados neste trabalho.Aqui serão renumerados para, 1, 2, 3 e 4 respectivamente.

Tabela 2.1: Cenários da RLB e seus respectivos parâmetros e probabilidades aproximadas

Cenário	β_0	β_1	a	$P(Y=1)$	$P(Y=1\|X=a)$
1	-2.30259	0.38376	6	0.22	1.00
2	-2.30259	0.38376	4	0.12	0.46
3	-0.35667	0.70808	0.5	0.22	1.00
4	-0.35667	0.70808	-0.5	0.12	0.49

A Figura 2.1 apresenta ajustes para os modelos RL e RLB para uma simulação com $n = 250$ para cada um dos cenários usados. O ajuste nos cenários 1 e 3 pela RLB são superiores aos fornecidos pela RL, enquanto que nos cenários 2 e 4 há pouca diferença entre os modelos adotados.



(a) Cenário 1

(b) Cenário 2

(c) Cenário 3

(d) Cenário 4

Figura 2.1: Ajuste dos modelos Logístico e Log-Binomial para os cenários de dados da RLB

Para o caso de duas variáveis explicativas, foram utilizados os quatro cenários apresentados em Blizzard e Hosmer [2]. Os passos para a realização da simulação mudam ligeiramente, envolvendo os seguintes passos:

1. Definir um valor para β_D, β_U e p;

2. Simular D de uma Bernoulli com probabilidade p;

3. Simular U de uma Uniforme no intervalo $[-6+2d, 2+2d]$;

4. Obter o valor das probabilidades $\pi(x_i) = e^{\beta_0+\beta_D d_i+\beta_U u_i}$;

5. Simular Y_i com distribuição Bernoulli com probabilidade de sucesso $\pi(x_i)$.

A Tabela 2.2 apresenta os diferentes cenários. Sendo β_D o parâmetro referente à variável dicotômica e β_U o parâmetro referente à variável uniforme.

Tabela 2.2: Cenários da RLB com duas variáveis explicativas

Cenário	Intercepto	β_D	β_U	p	$P(Y=1)$	$P(Y=1\|\mathbf{D}=d, U=2+2d)$
15	ln(0.3)	ln(1.5)	0.18	0.2	0.28	0.92
16	ln(0.3)	ln(1.5)	0.18	0.5	0.36	0.92
17	ln(0.3)	ln(2.0)	0.10	0.2	0.33	0.90
18	ln(0.3)	ln(2.0)	0.10	0.5	0.43	0.90

2.8.2 Simulações para o modelo Logístico

As simulações do modelo RL foram elaboradas com o objetivo de se obter curvas ajustadas que fossem consideravelmente diferentes das da RLB com um ajuste superior da RL para os dados. Mais uma vez, tratou-se de se utilizar apenas uma variável explicativa, ou seja,

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

Desta vez, as simulações foram realizadas com base em valores dos parâmetros β_0, β_1 e valor máximo da uniforme a sugeridos pelo professor orientador. Foram sugeridas dez possibilidades e foram escolhidas seis delas que permitiam uma melhor diferenciação entre as curvas da RL e da RLB.

Os passos para se realizar a simulação foram semelhantes aos do caso anterior, alterando apenas o passo da determinação do valor das probabilidades.

1. Definir um valor para β_0, β_1 e a;

2. Simular X de uma Uniforme no intervalo $[-6, a]$;

3. Obter o valor das probabilidades $\pi(x_i) = \frac{e^{\beta_0+\beta_1 x_i}}{1+e^{\beta_0+\beta_1 x_i}}$;

4. Simular Y_i com distribuição Bernoulli com probabilidade de sucesso $\pi(x_i)$.

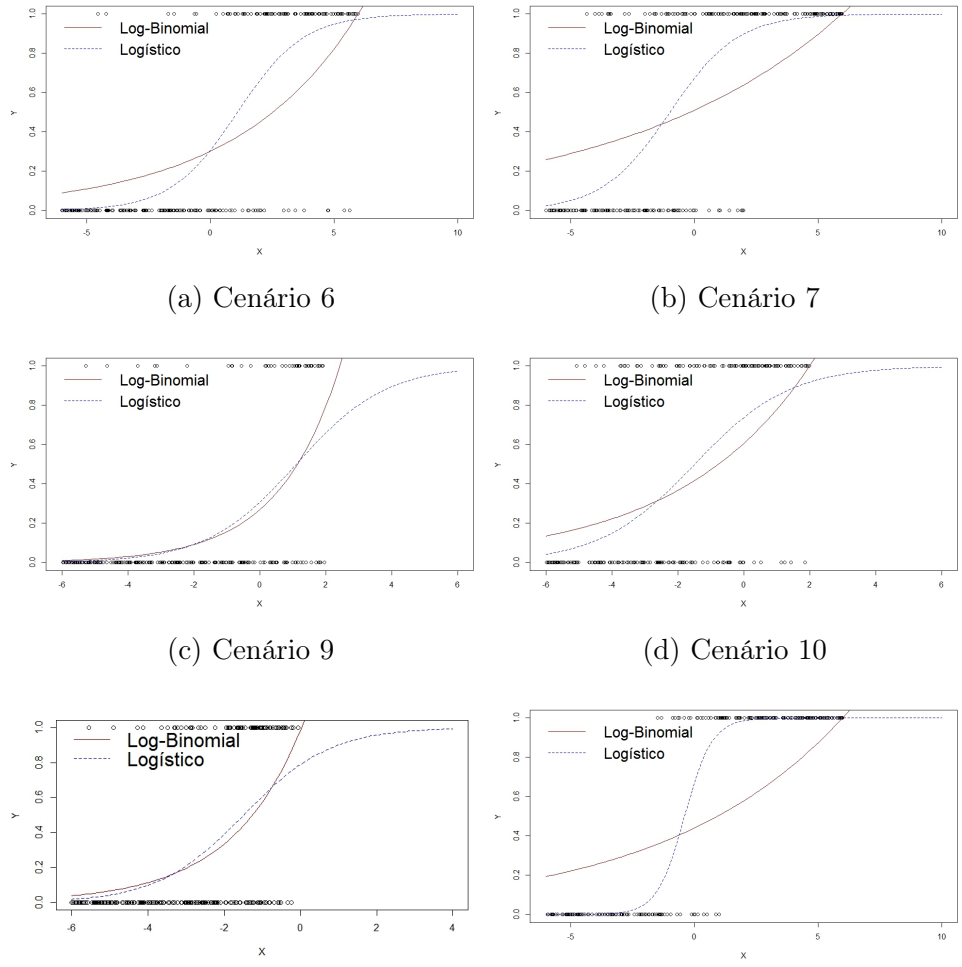
A Tabela 2.3 apresenta os parâmetros e probabilidades para os 10 cenários propostos. Os cenários que apresentaram maiores probabilidades $P(Y=1|X=a)$, mostrando maior diferença entre os 2 modelos (RL e RLB), e portanto foram selecionados foram os cenários 6, 7, 9, 10, 13 e 14.

Nos cenários 6, 7, 10 e 14 as curvas dos dois modelos estavam bem diferenciadas. Nas configurações 9 e 13, havia alguma diferenciação entre as curvas. A seguir, consta a imagem de uma simulação.

Tabela 2.3: Cenários e seus respectivos parâmetros e probabilidades

C.	β_0	β_1	a	$P(Y=1)$	$P(Y=1\|X=a)$	C.	β_0	β_1	a	$P(Y=1)$	$P(Y=1\|X=a)$
5	-1	0.25	6	0.30	0.62	6	-1	0.75	6	0.39	0.97
7	1	0.75	6	0.61	1.00	8	-1	0.25	2	0.20	0.38
9	-1	0.75	2	0.16	0.62	10	1	0.75	2	0.42	0.92
11	-1	0.25	0	0.16	0.27	12	-1	0.75	0	0.07	0.27
13	1	0.75	0	0.29	0.73	14	1	1.5	6	0.56	1.00

A Figura 2.2 apresenta o ajuste para os 6 melhores cenários.



(a) Cenário 6

(b) Cenário 7

(c) Cenário 9

(d) Cenário 10

(e) Cenário 13

(f) Cenário 14

Figura 2.2: Ajuste dos modelos Logístico e Log-Binomial para os cenários de dados da RL

Capítulo 3

Resultados

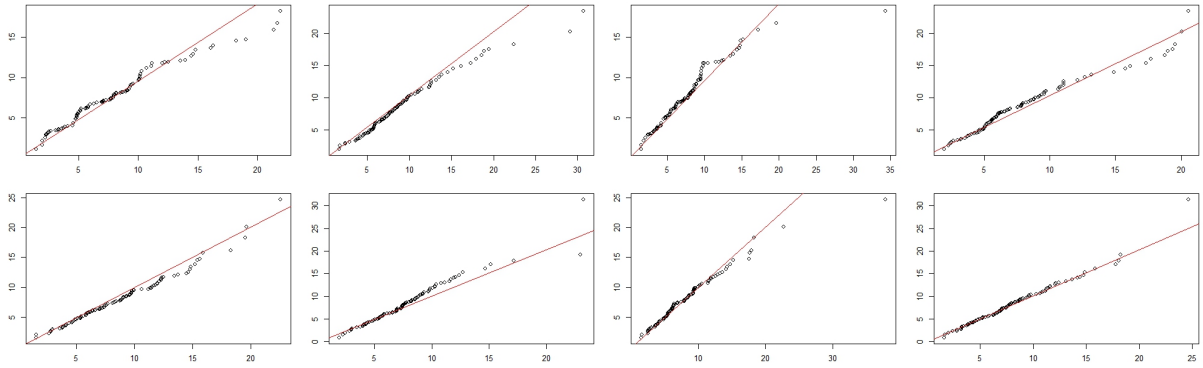
3.1 Teste de Hosmer-Lemeshow

Para a avaliação da validade do teste de Hosmer-Lemeshow para o caso em que os dados são de uma RLB, foi realizada uma análise gráfica e outra do poder do teste. Em Hosmer et al [4] já há a menção de que o teste para a RL não apresenta poder muito alto, além de ser sensível à forma de se dividir os grupos. Assim, espera-se que os mesmos problemas ocorram também ao tentar aplicar o teste para dados de uma RLB. Neste trabalho, os valores para a estatística $\hat{C}$ e os respectivos p-valores foram todos obtidos por meio do pacote "*ResourceSelection*" do R.

3.1.1 Análise Gráfica

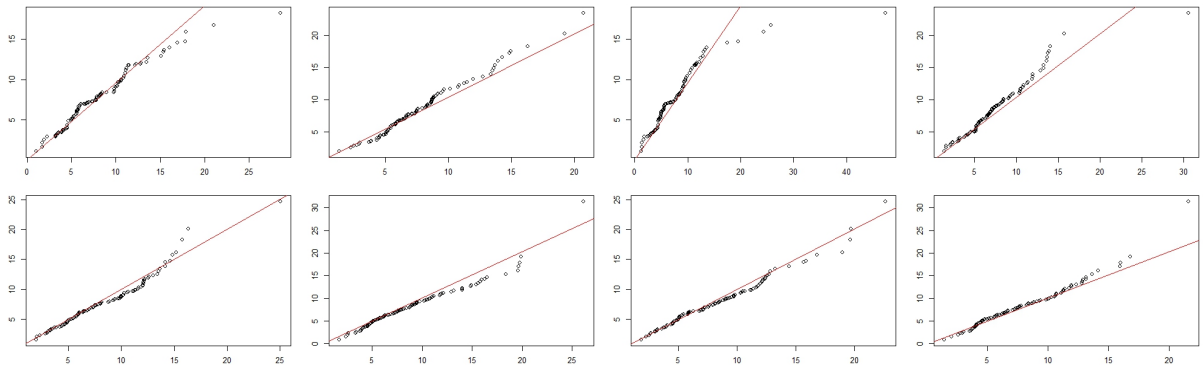
Sob a hipótese nula de que o modelo se ajusta aos dados, Blizzard e Hosmer [2] afirmam que a estatística do teste de Hosmer-Lemeshow possui distribuição χ^2_8 quando $g = 10$ mesmo para o caso em que o modelo verdadeiro é a RLB. Um *QQ-plot* de uma χ^2_8 para a estatística de teste na simulação foi utilizado para observar se os dados se comportam da forma esperada. Para cada um dos oito Cenários (1, 2, 3, 4, 15, 16, 17, 18) e quatro tamanhos de amostras ($n = 100, 250, 500, 1000$) da RLB, obteve-se a estatística $\hat{C}$ mil vezes e seus valores foram usados no *QQ-plot* (Figura 3.1). Observa-se que nos Cenários com uma variável explicativa (1, 2, 3 e 4), embora em alguns dos casos o ajuste pareça adequado, há outros que escapam de forma consistente do valor esperado, tornando duvidosa a ideia de que a estatística para a RLB possua distribuição χ^2_8 quando $g = 10$.

Já nos Cenários com duas variáveis explicativas (15, 16, 17 e 18), de forma geral não houve muitos casos em que os valores da estatística de teste se afastaram consideravelmente do *QQ-plot*, contrastando com o caso univariado que teve mais casos. Não foi possível se concluir sobre a validade da estatística. Para se observar melhor a possível validade da estatística julgou-se necessário realizar uma análise do poder da mesma, dessa vez para todos os Cenários, de forma a haver uma possibilidade de comparação entre o teste aplicado à RL e à RLB.



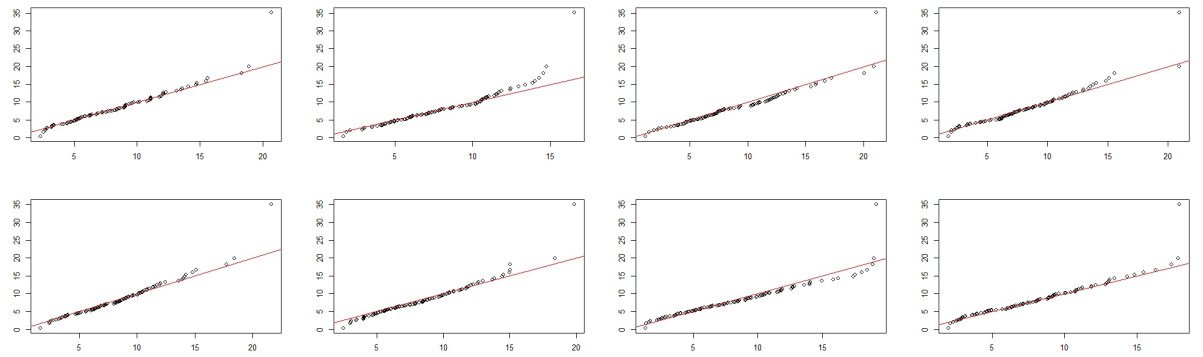
(a) Cenário 1

(b) Cenário 2



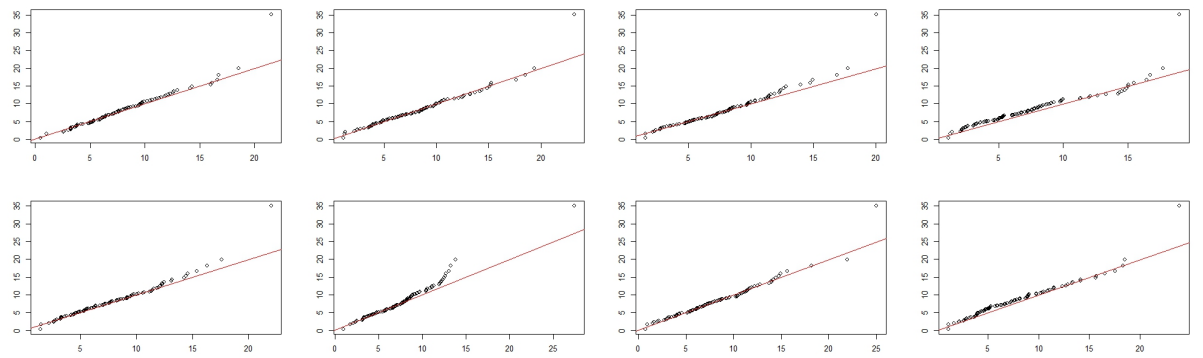
(c) Cenário 3

(d) Cenário 4



(e) Cenário 15

(f) Cenário 16



(g) Cenário 17

(h) Cenário 18

Figura 3.1: *QQ-plot* Para os quatro Cenários da RLB com uma variável, com $n = 100$ (canto superior esquerdo), $n = 250$ (canto superior direito), $n = 500$ (canto inferior esquerdo), $n = 1000$ (canto inferior direito)

3.1.2 Análise do Poder do Teste

A seguir, para se ter uma ideia mais precisa acerca da qualidade do teste como um todo, foram realizadas simulações específicas para esse fim. Hosmer et al. [4] analisam o poder do teste em uma série de situações para a RL e comentam sobre algumas das desvantagens de utilizar o teste. Em especial, a estatística parece ser sensível à forma como se dá o agrupamento, além de fornecer baixo poder em várias circunstâncias. Além disso, explicitam-se as situações possíveis em que se haveria um mau ajuste do modelo, que se resumem em três e podem acontecer simultaneamente:

1. A função *link* considerada é incorreta;

2. O preditor linear utilizado é incorreto;

3. A variância não é a de uma distribuição Bernoulli.

Com base nessas constatações, buscou-se realizar algumas análises de situações específicas em que o primeiro ou o segundo item acima ocorrem, para então se obter um poder do teste simulado. Para essa situação em específico, foram realizadas mil simulações para diferentes circunstâncias, mantendo-se o nível de $\alpha = 0.05$. Tais circunstâncias são válidas tanto para os Cenários de uma variável explicativa quanto de duas, e foram:

- Utilizar a função *link* incorreta (RL caso os dados sejam da RLB e RLB caso os dados sejam da RL);

- Utilizar a variável explicativa contínua exponencializada;

- Utilizar a variável explicativa contínua ao quadrado;

- Utilizar o inverso da variável explicativa contínua;

- Utilizar o módulo da variável explicativa contínua.

Para se ter uma ideia inicial, nenhum dos três pontos que podem levar a um mau ajuste foi utilizado, ou seja, foram mantidas as condições corretas que deveriam levar à conclusão de que o modelo está bem ajustado. Tais informações constam na Tabela 3.1. Pode-se observar que a taxa de rejeição empírica se manteve próxima da teórica. Além disso, com exceção do Cenário 14, todos os cenários apresentaram média e desvio padrão com uma certa proximidade entre si. O Cenário 14 difere dos outros, com desvio padrão da estatística bastante superior aos demais casos. Isso se deu porque é o Cenário que melhor se ajusta à RL, chegando a apresentar estatísticas de valor quase oitenta vezes superior ao valor crítico que define a rejeição em 5%.

Tabela 3.1: Medidas da estatística de Hosmer-Lemeshow para mil simulações ($\alpha = 0.05$)

Cenário	Média	Desvio padrão	Taxa de rejeição empírica
1	8.39	4.20	0.06
2	7.87	3.76	0.04
3	8.01	3.70	0.04
4	8.16	4.17	0.06
6	8.18	4.88	0.07
7	7.78	4.53	0.05
9	7.56	4.26	0.05
10	8.10	3.97	0.05
13	7.96	3.89	0.04
14	8.33	43.56	0.06
15	8.26	4.03	0.06
16	7.91	3.85	0.04
17	7.81	3.91	0.04
18	7.58	3.58	0.04

A análise do poder do teste consta na Tabela 3.2. Cada valor se refere ao poder do teste em um determinado cenário sob uma determinada especificação incorreta do modelo. Os elementos de cada coluna representam, respectivamente:

- ***Link:*** situação em que o modelo incorreto é ajustado aos dados;
- **Exponencial:** situação em que se usou a variável explicativa contínua exponencializada para o ajuste do modelo;
- **Quadrado:** situação em que se usou a variável explicativa contínua ao quadrado para o ajuste do modelo;
- **Inverso:** situação em que se usou o inverso da variável explicativa contínua para o ajuste do modelo;
- **Módulo:** situação em que se usou o módulo da variável explicativa contínua para o ajuste do modelo.

Sendo $\mathbf{x}$ a covaríavel simulada contínua, foram realizadas mil simulações de dados com $n = 250$ para cada cenário do modelo correto (por exemplo, RLB no cenário 1 e RL no cenário 6). Em seguida, ajusta-se um modelo incorretamente especificado, cuja especificação é dada a seguir. O erro de especificação pode ser:

1) ***Link***

a) Ajustar os dados simulados pelo modelo incorreto (RLB quando provém de uma RL e RL quando provém de uma RLB);

b) Obter a estatística de Hosmer-Lemeshow;

c) Tomar a média das vezes em que, com $\alpha = 0.05$, a hipótese nula do teste não foi rejeitada.

2) Exponencial

a) Tomar $\mathbf{e} = \exp(\mathbf{x})$;

b) Ajustar pelo modelo correto do cenário a variável resposta original e **e** (e outras, se houver) como covariável;

c) Obter a estatística de Hosmer-Lemeshow;

d) Tomar a média das vezes em que, com $\alpha = 0.05$, a hipótese nula do teste não foi rejeitada.

3) Quadrado

a) Tomar $\mathbf{q} = \mathbf{x}^2$;

b) Ajustar pelo modelo correto do cenário a variável resposta original e **q** (e outras, se houver) como covariável;

c) Obter a estatística de Hosmer-Lemeshow;

d) Tomar a média das vezes em que, com $\alpha = 0.05$, a hipótese nula do teste não foi rejeitada.

4) Inverso

a) Tomar $\mathbf{i} = 1/\mathbf{x}$;

b) Ajustar pelo modelo correto do cenário a variável resposta original e **i** (e outras, se houver) como covariável;

c) Obter a estatística de Hosmer-Lemeshow;

d) Tomar a média das vezes em que, com $\alpha = 0.05$, a hipótese nula do teste não foi rejeitada.

5) Módulo

2) Tomar $\mathbf{m} = |\mathbf{x}|$;

3) Ajustar pelo modelo correto do cenário a variável resposta original e **m** (e outras, se houver) como covariável;

4) Obter a estatística de Hosmer-Lemeshow;

5) Tomar a média das vezes em que, com $\alpha = 0.05$, a hipótese nula do teste não foi rejeitada.

A Tabela 3.2 segue abaixo. Como era de se esperar (pois essa função não altera drasticamente a forma em que as observações se distribuem), para a função módulo o teste em geral não teve um poder elevado. Em contrapartida, o poder em alguns dos casos para a função inversa foi elevado, mas parece ser inferior quando há duas variáveis explicativas, o que pode sugerir que à medida que o número de covariáveis aumenta, mais difícil é a identificação de um mau ajuste. A função quadrática foi a que apresentou menor poder entre os casos em que o preditor linear é incorreto, ficando abaixo de 50% em todos os casos com exceção do Cenário 3. Em geral, para o caso da exponencial o poder do teste foi bom, sendo menos elevado nos dois últimos Cenários. Por último, para o caso da *link*, o teste parece ser pouco poderoso para realmente realizar uma discriminação, exceto quando se tratam de casos mais acentuados, como o do Cenário 14.

Tabela 3.2: Poder estimado do teste de Hosmer-Lemeshow para diferentes circunstâncias em mil simulações

Cenário	*Link*	Exponencial	Quadrado	Inverso	Módulo
1	0.28	1.00	0.07	1.00	0.06
2	0.06	0.56	0.47	0.97	0.50
3	0.25	0.96	0.66	1.00	0.06
4	0.07	0.26	0.23	0.40	0.06
6	0.99	1.00	0.09	1.00	0.06
7	1.00	0.98	0.08	1.00	0.05
9	0.03	0.73	0.23	1.00	0.41
10	0.58	0.99	0.41	1.00	0.23
13	0.08	0.77	0.26	0.98	0.04
14	1.00	0.70	0.17	1.00	0.11
15	0.06	0.52	0.11	0.58	0.14
16	0.11	0.58	0.17	0.77	0.20
17	0.05	0.21	0.06	0.25	0.07
18	0.08	0.25	0.06	0.40	0.07

Para o caso de duas variáveis explicativas, o teste se mostrou pouco poderoso para quase qualquer circunstância analisada. Mais uma vez, isso pode ser um indicativo de que o poder do teste se reduz à medida que o número de covariáveis aumenta. Além disso, o teste parece ter pouca sensibilidade para transformações de ordem quadrática e

modular, o que sugere que pode haver mais transformações das covariáveis que fornecem testes menos poderosos, tornando mais duvidosas as conclusões que se pode chegar por meio do teste para casos reais.

Todas essas observações tornam a validade do teste ainda mais duvidosa e mostram que ele claramente não deve ser usado para indicar qual o modelo mais adequado aos dados, independente de qual o modelo verdadeiro.

3.2 Análise das Medidas de Ajuste

3.2.1 Uma Variável Explicativa

Como foi mencionado anteriormente, para todas as medidas de comparação utilizadas há o problema de a amostra influenciar nos resultados, podendo favorecer o modelo incorreto. Sendo assim, a maneira que se pensou para se reduzir esse problema foi simular os mesmos dados mil vezes (ou seja, mantendo os mesmos valores de β_0, β_1 e a para essas simulações) e obter as respectivas medidas em cada um deles.

Foram tomadas amostras de tamanho 100, 250, 500 e 1000. Como o comportamento frente ao aumento da amostra serviu apenas para aumentar ligeiramente os resultados desejados, manteve-se a amostra fixa em 250 nos resultados presentes nesta seção.

As siglas em cada uma das tabelas representam o seguinte:

- Dev: Proporção das simulações em que a RLB foi a com menor *deviance*, ou seja, a porcentagem de vezes em que a RLB foi selecionada como o modelo correto segundo o critério *deviance*;
- REQM: Proporção das simulações em que o modelo RLB foi aquele com menor raiz do erro quadrático médio, ou seja, a porcentagem de vezes em que foi selecionado como o modelo correto;
- Dev e REQM: Indica a proporção de vezes em que tanto a raiz do erro quadrático médio quanto a *deviance* indicaram ao mesmo tempo que o modelo RLB era o correto;
- MEC-LB: Proporção média do erro de classificação quando se está ajustando pelo modelo RLB;
- MEC-L: Proporção média do erro de classificação quando se está ajustando pelo modelo RL;
- P(Y=1|X=0): A probabilidade de sucesso condicionada ao valor de X sendo 0, ou seja, o valor de $\pi(0)$;
- P(Y=1): A probabilidade marginal de sucesso.

Tabela 3.3: Proporção em % aproximada na qual cada medida de ajuste favorece a RLB a partir de simulações do modelo Log-Binomial ($n = 250$)

Cenário	Dev	REQM	Dev e REQM	MEC-LB	MEC-L	P(Y=1|X=0)	P(Y=1)
1	84	81	78	15	15	10	21
2	54	56	49	12	12	10	12
3	83	81	78	15	15	70	21
4	52	56	47	12	12	0	12

Vale ressaltar que, no Cenário 4, a probabilidade condicional foi 0 pelo fato de que X segue um uniforme no intervalo [-6, -0.5] nesse caso. Portanto, como é um valor fora do intervalo de X, não é possível que ele seja obtido.

Como era de se esperar, a *deviance* e a raiz do erro quadrático médio selecionaram o modelo RLB de forma conjunta na maior parte das vezes. Além disso, como também era de se esperar, as duas medidas indicaram o modelo RLB mais vezes quando as probabilidades de sucesso condicionadas a a eram altas, e indicaram menos quando as probabilidades eram mais próximas de 0.5, ou seja, quando as duas curvas são bastante semelhantes. Esse é um indicativo de que as medidas possuem a capacidade de determinar qual o melhor modelo quando um se ajusta consideravelmente melhor que o outro.

A média do erro de classificação, por sua vez, foi muito próxima em todos os casos analisados, sugerindo que não deve ser boa ferramenta de comparação. Isso faz sentido quando se considera o argumento em Hosmer e Lemeshow [5] de que se está limitando a probabilidade a apenas dois valores, acarretando em uma considerável perda de informação.

Esses resultados podem ser vistos também sob uma ótica probabilística. Como exemplo, a probabilidade estimada de se cometer o erro tipo I (afirmar que os dados são da RL quando na verdade são da RLB) quando se utiliza o critério da *deviance* é de 16% no primeiro Cenário (pois 84% é a proporção de vezes em que o critério *deviance* indicou corretamente o modelo). Ao se utilizar apenas a REQM ainda no mesmo cenário, essa probabildiade seria de 81%. Pelos valores obtidos da Tabela 3.4, quanto melhor for o ajuste da RLB, menor será a probabilidade de erro tipo I ao serem utilizadas as medidas *deviance* e REQM.

Tabela 3.4: Proporção em % na qual cada medida de ajuste favorece a RLB a partir de simulações do modelo Logístico ($n = 250$)

	Dev	REQM	Dev e REQM	MEC-LB	MEC-L	P(Y=1|X=0)	P(Y=1)
Cenário 6	0	0	0	18	15	27	39
Cenário 7	0	0	0	15	15	73	61
Cenário 9	26	29	22	15	15	27	16
Cenário 10	1	1	1	22	21	73	42
Cenário 13	21	23	17	23	23	73	29
Cenário 14	0	0	0	11	8	73	56

Avaliou-se todos os cenários do modelo RL em que as curvas eram consideravelmente diferentes. Então, seria esperado que em todos os casos haja um bom favorecimento da RL.

Mais uma vez, a *deviance* e a raiz do erro quadrático médio serviram para realizar uma boa comparação entre os dois modelos. Nos Cenários 6, 7 e 14, nenhuma dessas medidas favorecerem o modelo RLB. Nas configurações 9 e 13 (em que, mesmo havendo diferença, não era uma diferença tão grande quanto nas outras configurações) as duas medidas também serviram para fazer a diferenciação, mas com uma capacidade menor. Além disso, na maior parte dos casos, a *deviance* e a raiz do erro quadrático médio, conjuntamente, forneceram a mesma conclusão.

Para a média do erro de classificação, observa-se que na maior parte dos casos não há muita diferença, por mais que haja configurações em que as curvas são extremamente diferentes uma da outra.

Pela ótica probabilística há uma alteração na interpretação com relação ao primeiro caso. Para a RL, os valores obtidos fornecem a probabilidade estimada de se cometer o erro tipo II (não rejeitar que os dados são de uma RLB quando na realidade eles são da RL). Para exemplificar, a probabilidade de se afirmar erroneamente que os dados do Cenário 9 são provenientes da RLB é de aproximadamente 21% ao se utilizar o critério *deviance* e 29% pelo critério REQM.

Conclui-se então que, ao menos com base nesses dois conjuntos de simulações, a *deviance* e a raiz do erro quadrático médio são boas medidas para a realização das comparações, enquanto a média do erro de classificação não avalia bem.

3.2.2 Duas Variáveis Explicativas

As siglas utilizadas para o caso de uma variável explicativa se mantêm para o caso de duas variáveis. Manteve-se também o tamanho da amostra de 250 pois observou-se que, conforme a amostra cresce, mais precisa é a medida para determinar o modelo verdadeiro. Os resultados para os quatro Cenários constam na tabela abaixo. Embora os resultados para a média do erro de classificação não tenham sido informativos para uma variável explicativa, a medida foi observada também para duas variáveis.

Tabela 3.5: Proporção em % na qual cada medida de ajuste favorece a RLB a partir de simulações do modelo Log-Binomial ($n = 250$)

Cenário	Dev	REQM	Dev.e.REQM	MECLB	MECL	Marginal	Condicional
15	71	70	68	25	25	28	92
16	79	76	74	28	28	36	92
17	71	69	68	27	27	33	90
18	74	72	71	30	30	43	90

Mais uma vez, observou-se que a média do erro de classificação não foi capaz de discriminar bem entre os modelos, pois forneceu resultados muito próximos independentemente

do modelo utilizado. A *deviance* e a raiz do erro quadrático médio, por sua vez, forneceram resultados melhores e, mais uma vez, concordantes um com o outro. A interpretação probabilística desses casos é a mesma do caso univariado da RLB, ou seja, pode ser visto como uma estimativa da probabilidade de erro tipo I.

3.3 Análise de Diagnóstico

Para a realização da análise de diagnóstico, foram escolhidos quatro Cenários com uma variável explicativa (Cenários 1, 2, 6 e 14) e um cenário com duas (Cenário 18). Como o teste de Hosmer-Lemeshow pode não ser válido para os Cenários da RLB, ele será utilizado apenas nos sexto e décimo quarto Cenários. O tamanho da amostra para todos os casos foi de $n = 250$.

3.3.1 Cenário 1

A estatística qui-quadrado de Pearson e a estatística resumo da *deviance* forneceram, respectivamente, os valores 181.076 e 159.3652. Como a covariável é contínua, conforme o tamanho da amostra aumenta o número de padrões também aumentará, caracterizando assim um caso de n-assintótica. Sendo assim, recomenda-se a utilização de valores das estatísticas de forma comparativa aos outros casos.

Pelo gráfico dos valores *leverage* na Figura 3.2 há uma indicação de que as observações são próximas de zero, com exceção da observação para a qual se atribuiu probabilidade 1. Como foi comentado, a observação com *leverage* 1 não pode ser interpretada como distante da média, então não se sabe dizer ao certo o que significaria tal valor. Levando essas questões em consideração, pode-se dizer que o Cenário 1 não parece apresentar observações muito influentes.

Foram destacadas as partes em que os valores ajustados assumem valores 0.1 e 0.9 pois fora do intervalo explicitado a medida pode não dar uma noção de distância, sendo portanto não confiável.

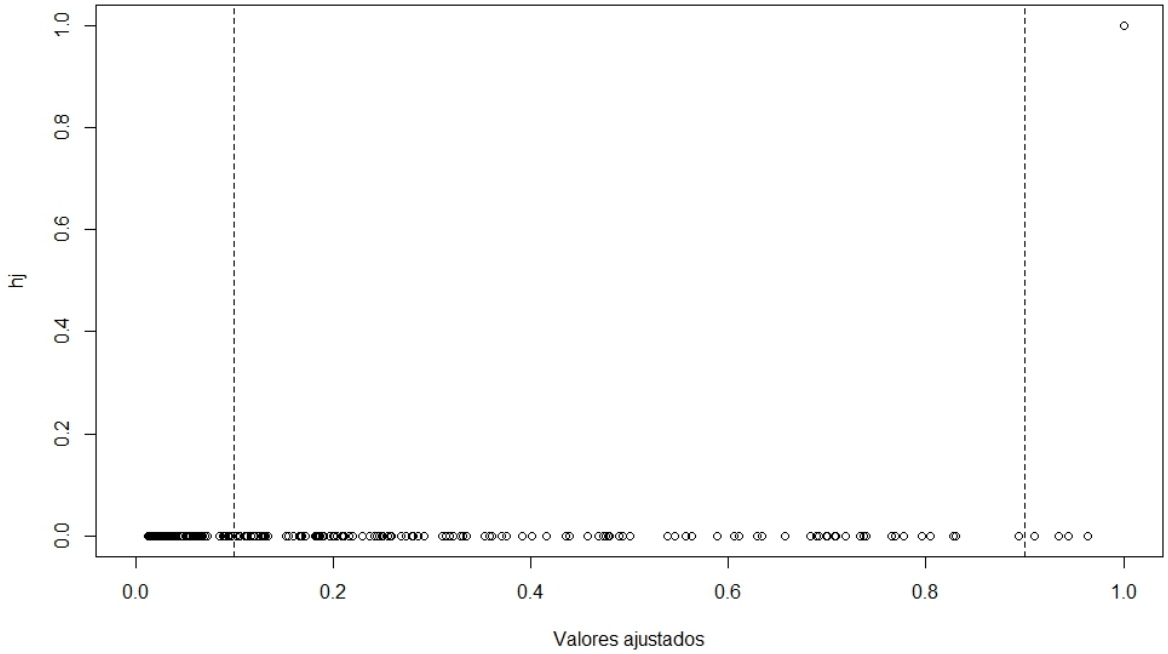


Figura 3.2: Valores *leverage* e valores ajustados (Cenário 1).

A Figura 3.3 possui uma aparência similar à da Figura 3.2. Isso se dá pela equação utilizada para se obter os valores de $\Delta\beta_j$. Ainda assim, pode-se observar que há uma observação cuja influência nos valores dos coeficientes é alta, tratando-se da mesma observação da Figura 3.2. Para considerar ou não essa observação como discrepante, serão utilizadas outras medidas de diagnóstico.

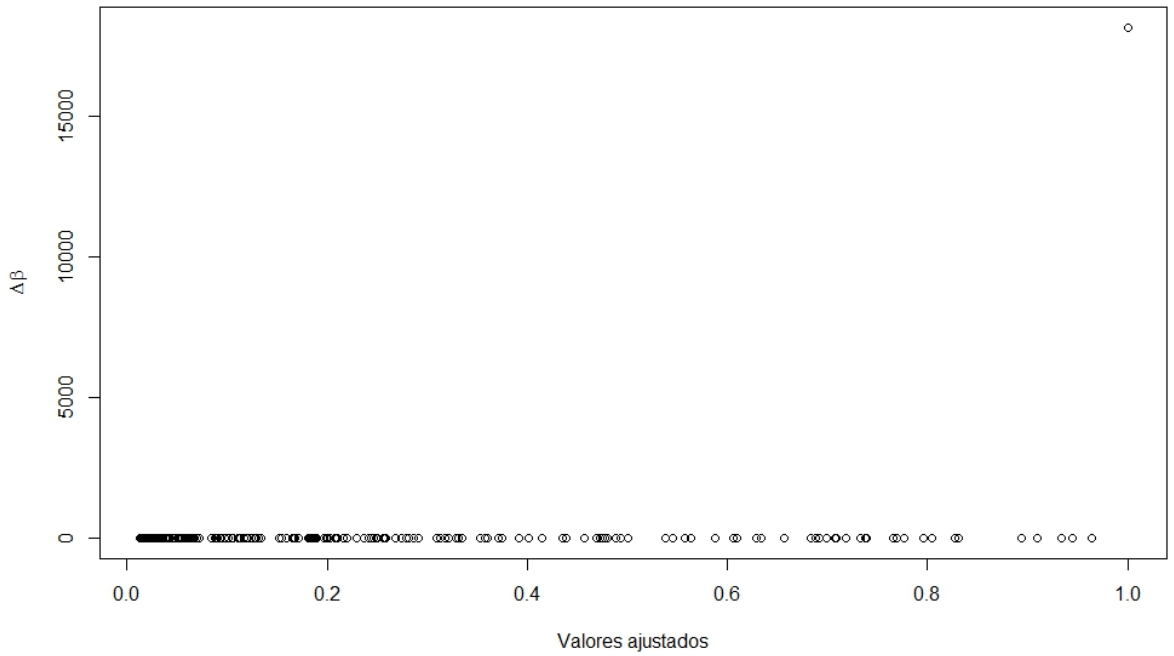
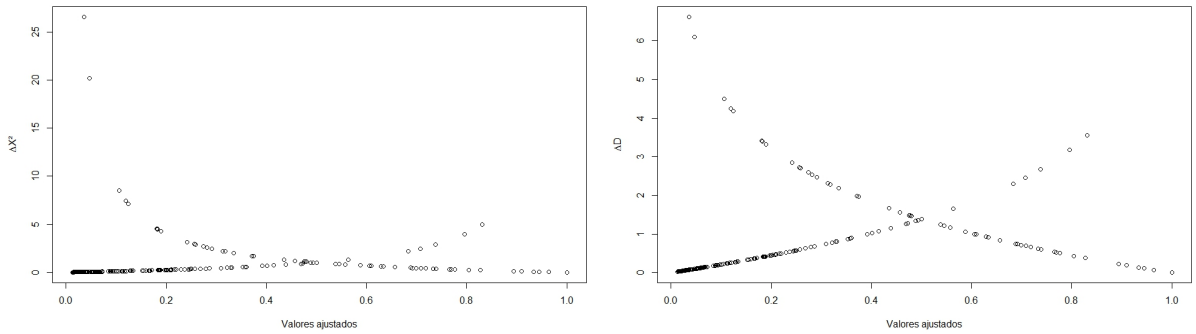


Figura 3.3: $\Delta\beta_j$ e valores ajustados (Cenário 1).

Com relação às medidas $\Delta\mathbf{X}_j^2$ e $\Delta\mathbf{D}_j$, em ambas as imagens há dois valores que se sobressaem em relação aos outros, mas nenhum deles é a observação que se destacou nas Figuras anteriores. Foi possível observar que esses dois valores são os mesmos para os dois gráficos, podendo então ser considerado que esses dois padrões de covariável se ajustam mal ao modelo. Ainda assim, não podem ser consideradas observações discrepantes, pois os outros gráficos do Cenário 1 não trazem destaque algum para essas observações.



(a) $\Delta\mathbf{X}_j^2$ (b) $\Delta\mathbf{D}_j$

Figura 3.4: Estatísticas contra valores ajustados (Cenário 1).

3.3.2 Cenário 2

Os valores obtidos para $\mathbf{X}^2$ e D foram, respectivamente, 243.4363 e 175.1721, ou seja, tomando apenas essas medidas resumo como base se concluiria que o ajuste é inferior ao do primeiro Cenário. Na subseção anterior nada foi comentado a respeito dessas estatísticas por não haver outro Cenário em que elas pudessem ser olhadas comparativamente.

Ao olhar para a Figura 3.5, fica aparente que os valores ajustados não ultrapassam o valor de 0.60, sendo que seu valor máximo nesse caso foi de aproximadamente 0.56. Esse é outro aspecto que serve como indicativo de que o ajuste não está tão adequado. Na Figura 2.1b, fica claro que a curva ajustada da RLB não chega ao valor 1 fora da nuvem de dados.

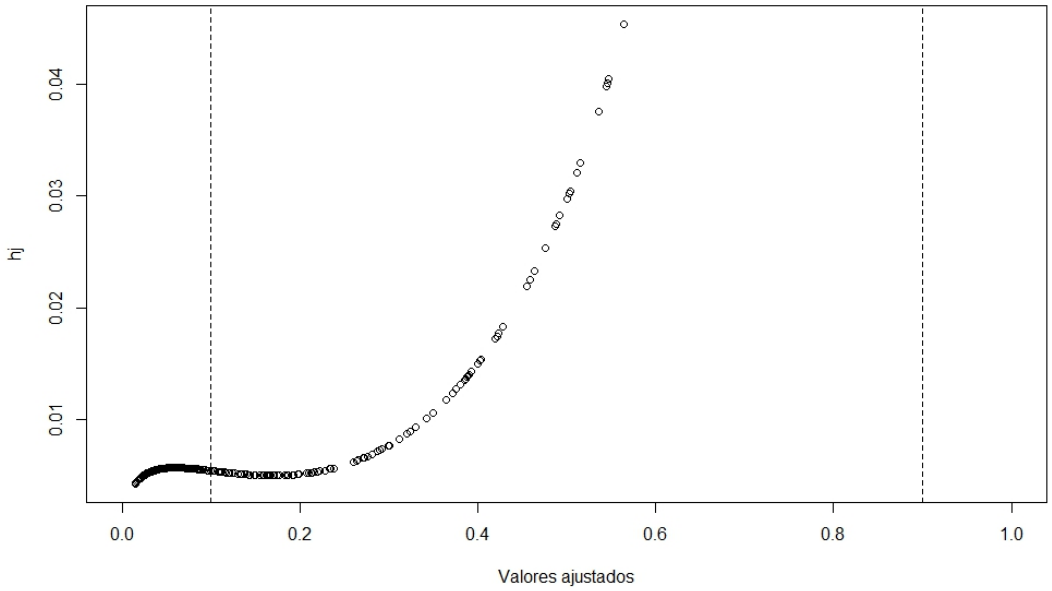


Figura 3.5: Valores *leverage* e valores ajustados (Cenário 2).

Além disso, os valores *leverage* demonstram uma tendência crescente durante praticamente todo o processo, o que faz sentido considerando que assim que a curva ajustada começa a subir, ela se afasta cada vez mais dos valores. É válido notar que há poucas observações de sucesso e observações de fracasso em uma frequência relevante durante todo o intervalo da covariável, o que deve também influenciar no prejuízo do ajuste.

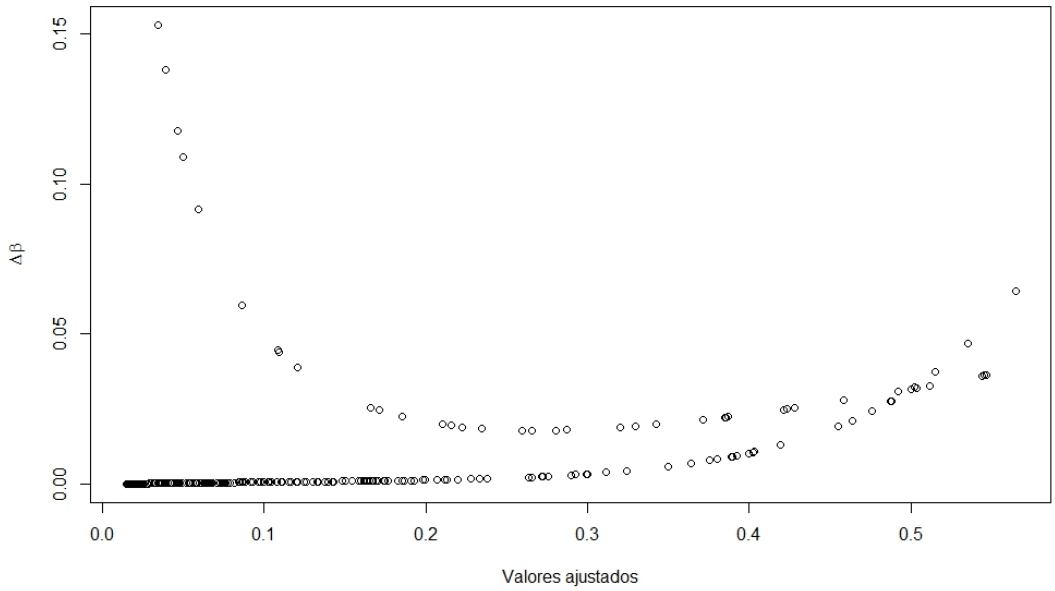
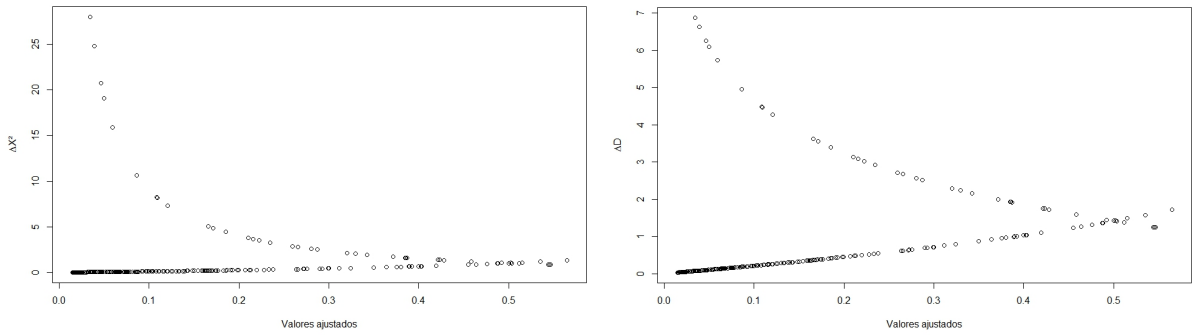


Figura 3.6: $\Delta\beta_j$ e valores ajustados (Cenário 2).

Para os valores de $\Delta\beta_j$, há uma presença maior de observações que influenciam no valor estimado dos coeficientes, apresentando-se praticamente na região entre 0 e 0.1. Em contrapartida, conforme o valor ajustado aumenta a estatística se reduz, dando indícios de crescimento após 0.5. As observações com maiores valores de $\Delta\beta_j$ são justamente aquelas em que os valores ajustados eram baixos (entre 0 e 0.1), mas os valores observados foram um. Como a maior concentração de zeros se dá justamente no início, não surpreende que esses valores sejam influentes. Por outro lado, conforme os valores ajustados aumentam, os $\Delta\beta_j$ se mantêm mais baixos, possivelmente pelo fato de haver uma maior presença de uns, mas mesmo assim não deixar de seguir com uma presença muito maior de zeros.



(a) $\Delta\mathbf{X}_j^2$ (b) $\Delta\mathbf{D}_j$

Figura 3.7: Estatísticas contra valores ajustados (Cenário 2).

Ao contrário do primeiro Cenário, os gráficos da Figura 3.7 possuem um formato consideravelmente similar, com muitas observações de valor mais alto enquanto há um

outro conjunto de valor inferior, sendo cinco das de valor alto as mesmas indicadas na Figura 3.6. Isso acaba sendo uma forte indicação de um ajuste ruim para vários padrões de covariável, o que só serve para indicar mais que o ajuste utilizado não foi bom, por mais que os dados tenham sido simulados de uma RLB.

3.3.3 Cenário 6

Pelo teste de Wald, o p-valor obtido para a hipótese do parâmetro da variável explicativa ser zero foi de aproximadamente zero, rejeitando a hipótese. Sendo assim, considera-se que a presença da covariável adiciona capacidade explicativa ao modelo.

Ao se utilizar o teste de Hosmer-Lemeshow para dez grupos, o p-valor obtido foi de 0.06132. Tal valor deixa a decisão muito ao critério do valor crítico que for selecionado, o que sugere um possível mau ajustamento do modelo aos dados.

A estatística qui-quadrado de Pearson e a estatística resumo da *deviance* forneceram, respectivamente, os valores 338.0097 e 157.9698. Como se trata de um outro modelo, é melhor que tais medidas sejam comparadas apenas com o Cenário seguinte.

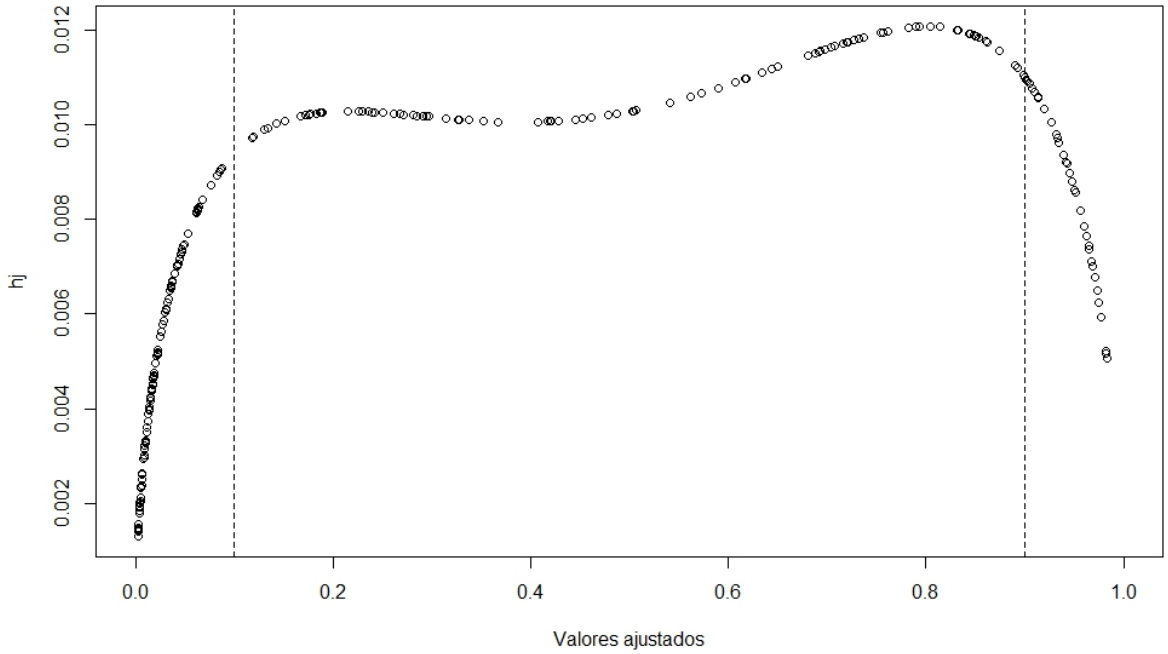


Figura 3.8: Valores *leverage* e valores ajustados (Cenário 6).

Ao contrário dos Cenários referentes à RLB, a imagem com os valores *leverage* do Cenário 6 mostra-os sem uma tendência crescente aparente. Eles possuem certa constância na maior parte do gráfico, com leve subida em torno de 0.5, decrescendo em torno de 0.8. Pelo que parece, nenhum valor se distancia muito daquele dado pelo ajuste. Junto ao teste de Hosmer-Lemeshow conclui-se que mesmo que o ajuste do modelo não seja adequado

(apesar de ser o modelo verdadeiro), nenhum ponto parece se distanciar muito da média.

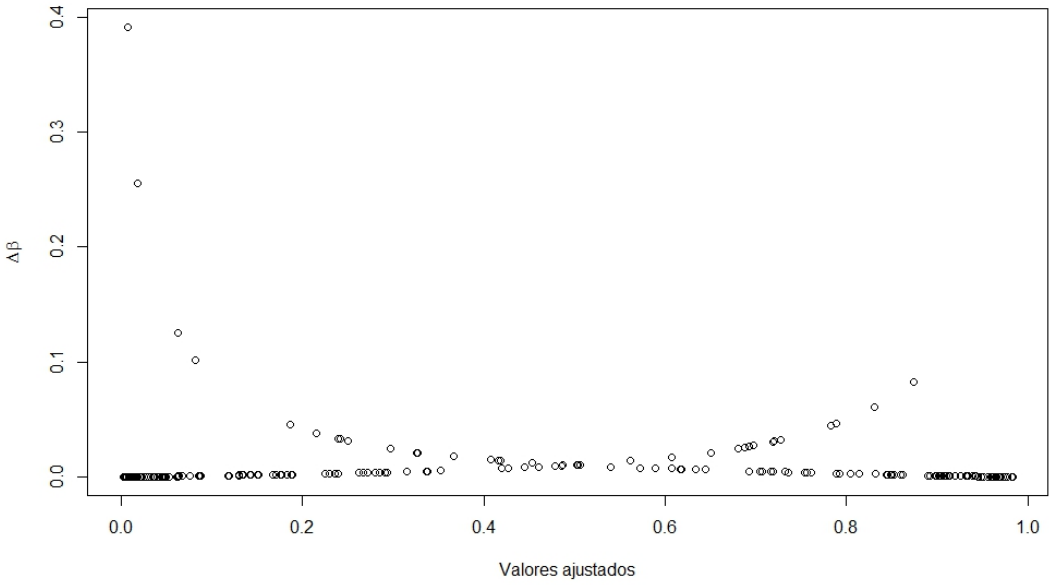
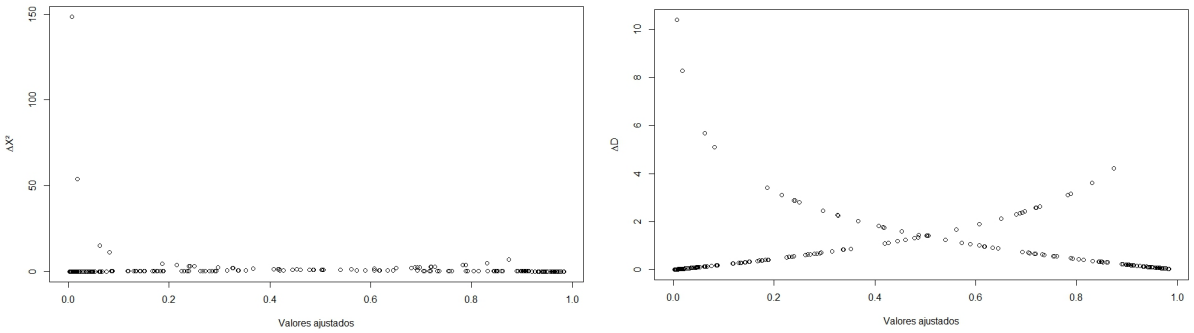


Figura 3.9: $\Delta\beta_j$ e valores ajustados (Cenário 6).

Assim como nos casos do modelo RLB, os maiores valores do gráfico referente ao $\Delta\beta_j$ estão ao lado esquerdo. Mais uma vez, trataram-se de observações às quais o modelo atribuiu valores consideravelmente próximos de zero, quando na realidade a observação foi um, em ambos os casos. No resto do gráfico, os valores permanecem quase todos abaixo de 0.1 e não apresentam tanta variação. O distanciamento de duas das observações em comparação às outras sugere que podem se tratar de observações discrepantes.

A Figura 3.10b apresenta também duas observações que parecem se sobressair com relação às demais, que são as mesmas comentadas na Figura 3.9. Portanto, pode-se concluir que se tratam de dois valores discrepantes, o que fortalece a suspeita de ajuste inadequado.



(a) $\Delta\mathbf{X}_j^2$ (b) $\Delta\mathbf{D}_j$

Figura 3.10: Estatísticas contra valores ajustados (Cenário 6).

3.3.4 Cenário 14

Para esse Cenário, assim como no anterior, rejeitou-se a hipótese de que o parâmetro da variável explicativa fosse zero com um p-valor de aproximadamente dois vezes dez elevado a menos onze. Além disso, o teste de Hosmer-Lemeshow forneceu um p-valor de 0.8509, favorecendo a hipótese de que o ajuste do modelo é adequado para os dados em análise. Os valores de $\mathbf{X}^2$ e D foram, respectivamente, 117.5006 e 104.6275, menores do que os do Cenário 6.

As observações da Figura 3.11 apresentam alguma constância, mas dessa vez com uma tendência decrescente em quase toda a região analisada, o que pode ser sinal de um bom ajuste, pois a distância da média não só parece pequena como também decresce conforme o tempo passa.

Como a curva desse Cenário possui um rápido crescimento, além de espaços nos extremos em que há a presença de apenas uma das duas possíveis observações e uma parte mais prolongada de uns, faz sentido que os valores *leverage* decresçam conforme os valores ajustados aumentam.

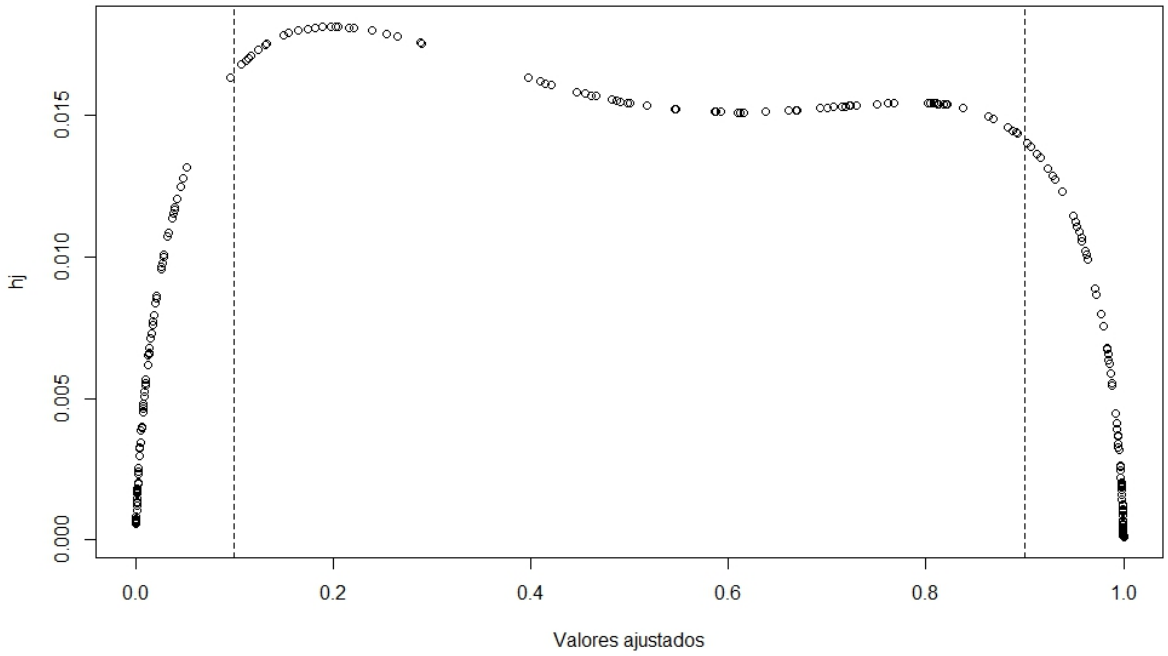


Figura 3.11: Valores *leverage* e valores ajustados (Cenário 14).

Deve-se tomar um certo cuidado com a interpretação da Figura 3.12. Por mais que os valores pareçam mais dispersos do que em relação ao Cenário anterior, isso não é verdade pelo fato de que o eixo da estatística em questão está menor, indo até 0.3.

O formato do gráfico é idêntico ao anterior, mas com apenas uma observação que chama a atenção, pelas mesmas razões já comentadas anteriormente para outras configurações

de Cenário. Do lado direito, a observação máxima está ligeiramente maior do que no Cenário anterior, mas não se julgou ser uma alteração relevante.

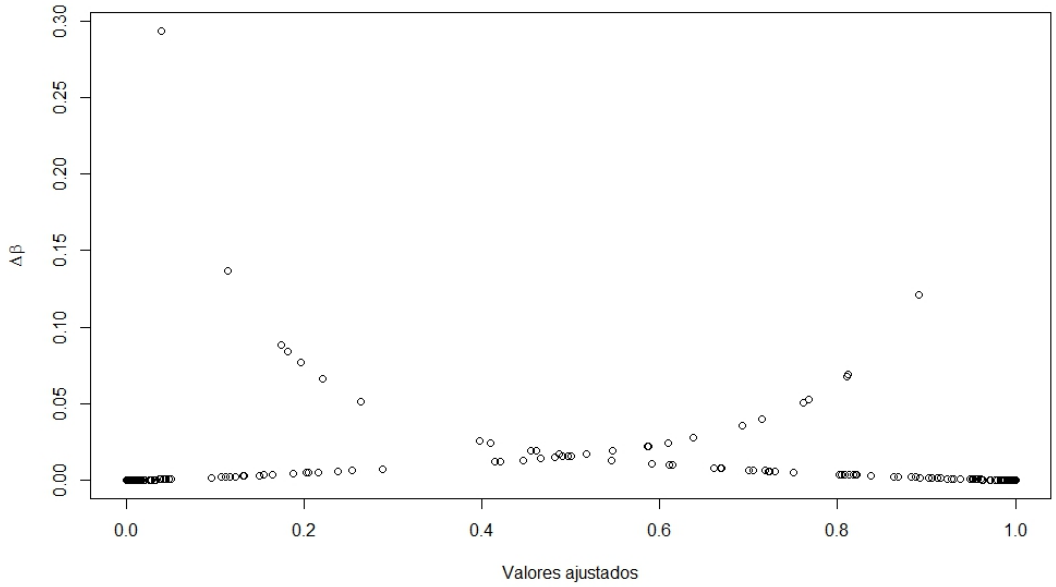
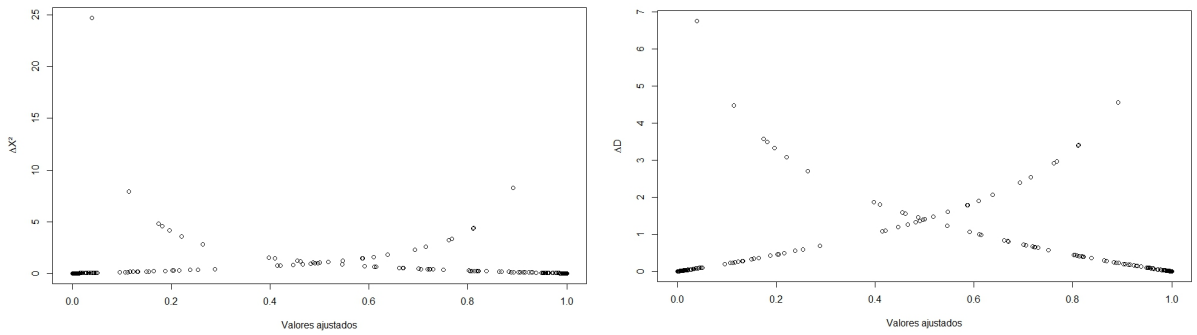


Figura 3.12: $\Delta\beta_j$ e valores ajustados (Cenário 14).

Ambos os gráficos da Figura 3.13 parecem fornecer uma ideia semelhante. A imagem à esquerda mostra que os valores estão seguindo um padrão sem grandes variações, excetuando-se a mesma observação comentada para a Figura 3.12.

A imagem à direita, por sua vez, parece apresentar um comportamento mais acentuado do que no Cenário anterior, mas isso se dá mais uma vez por conta dos limites do eixo de $\mathbf{\Delta D}_j$, que está menor. A mesma observação que parece se sobressair é a comentada anteriormente para os outros casos. Sendo assim, essa é a única observação que realmente parece fugir do padrão, embora apenas ligeiramente.



(a) $\mathbf{\Delta X}_j^2$ (b) $\mathbf{\Delta D}_j$

Figura 3.13: Estatísticas contra valores ajustados (Cenário 14).

3.3.5 Cenário 18

O Cenário 18 foi escolhido em detrimento dos outros por dar a impressão de ser o mais informativo entre eles na questão da análise de diagnóstico. Os Cenários 15 e 17 forneceram menores valores *leverage* e não deram indício de ajuste inadequado, enquanto o Cenário 16 é similar ao 18 graficamente. O tamanho da amostra também foi de 250.

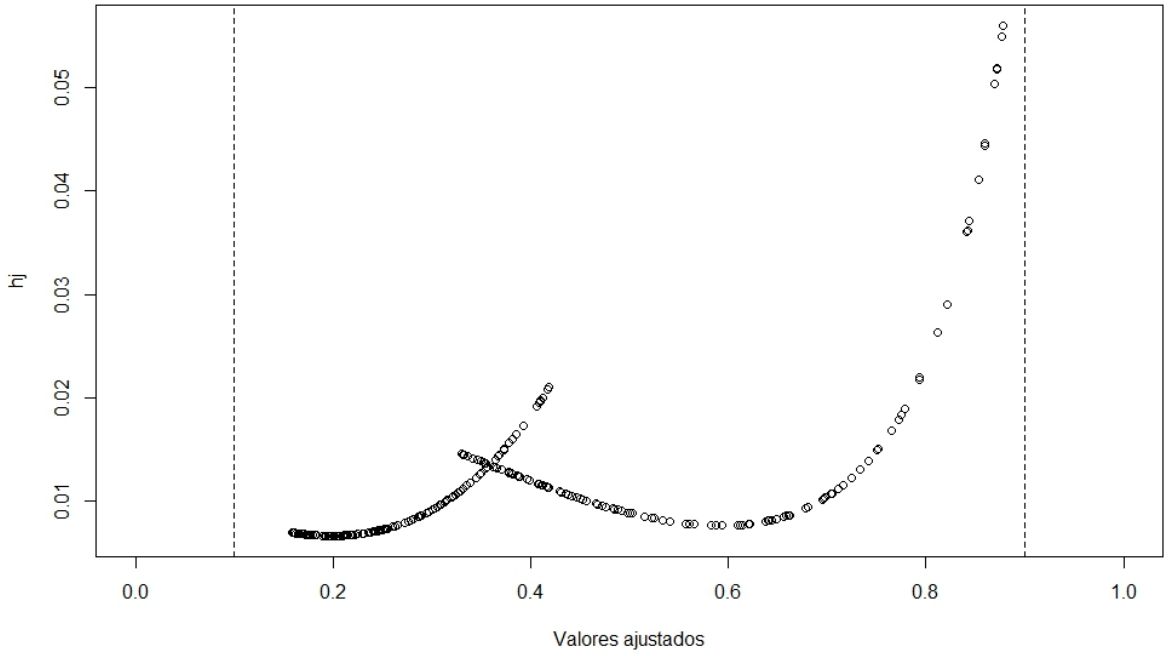


Figura 3.14: Valores *leverage* e valores ajustados (Cenário 18).

Para o gráfico dos valores *leverage*, observam-se duas tendências. Na primeira, que tem fim pouco depois de 0.4, há um crescimento, mas com as observações ainda pequenas quando comparado ao valor máximo possível, 1. A segunda tendência, começando antes de 0.4, cresce de modo mais prolongado, mas cessa antes de 0.06, que ainda é um valor consideravelmente baixo. Nenhuma probabilidade estimada foi próxima de 0 ou 1, indicando a possibilidade de um ajuste inadequado. De qualquer forma, nenhuma das observações parece provocar um grande desvio da média.

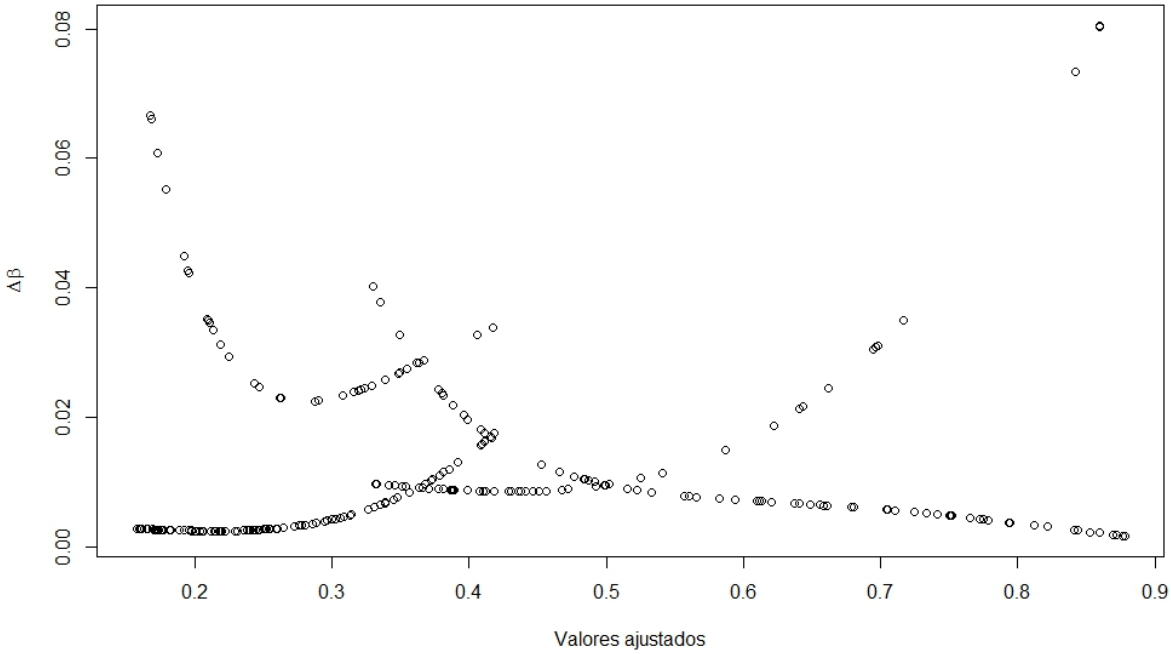
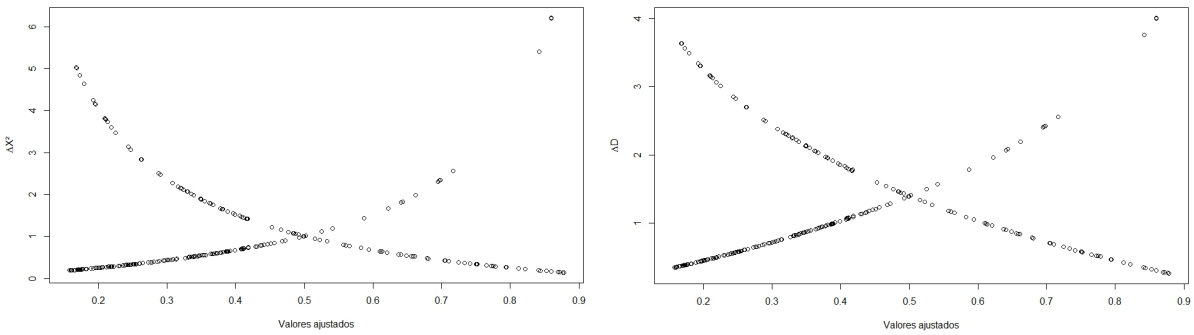


Figura 3.15: $\Delta\beta_j$ e valores ajustados (Cenário 18).

A Figura 3.15 apresenta mais tendências diferentes do que nos outros Cenários. Três observações no canto direito da Figura se sobressaem, isso ocorreu pois o modelo atribuiu probabilidade de sucesso próxima de 1, quando as observações eram 0. Em decorrência disso, são observações com uma influência mais forte na determinação dos valores dos coeficientes.



(a) $\Delta\mathbf{X}_j^2$ (b) $\Delta\mathbf{D}_j$

Figura 3.16: Estatísticas contra valores ajustados (Cenário 18).

Os dois últimos gráficos destacam as mesmas três observações da Figura 3.15, embora apresentem também valores próximos a esses no outro extremo. Sendo assim, não fica claro se essas observações se tratam de *outliers*.

3.4 Análise de dados reais

Os bancos de dados que serão estudados nessa seção foram fornecidos pelo professor orientador, e são os mesmos que constam em Andrade [1].

3.4.1 Banco de dados "Vaso"

As informações sobre esse banco de dados podem ser encontradas na página da *web* support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#statug_logistic_sect057.htm (acessado em 15/10/2015). Para a RLB, o modelo a ser ajustado seria

$$P(Constriction = 1|Volume, Rate) = \exp[\beta_{10} + \beta_{11}Volume + \beta_{12}Rate],$$

Enquanto para a RL seria

$$P(Constriction = 1|Volume, Rate) = \frac{\exp[\beta_{20} + \beta_{21}Volume + \beta_{22}Rate]}{1 + \exp[\beta_{20} + \beta_{21}Volume + \beta_{22}Rate]}.$$

A amostra é de tamanho $n = 39$. A seguir, foram obtidas algumas medidas para que fosse realizada a comparação entre os dois modelos. Alguns dos resultados foram aproximados por não acarretar em comprometimento da interpretação.

Tabela 3.6: Medidas de Ajuste para o banco de dados "Vaso" ($n = 39$).

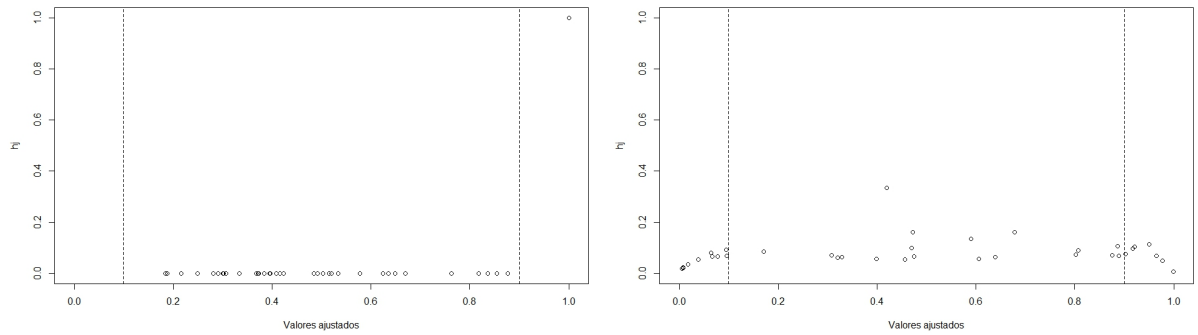
Modelo	*Deviance*	REQM	EC	$\mathbf{X}^2$	$\hat{C}$ (Hosmer-Lemeshow)	p-valor do teste
RLB	37.56	0.40	8	27.01	4.89	0.7795
RL	29.77	0.34	4	39.01	17.81	0.0227

Com esses resultados é possível observar que tanto a *deviance* quanto a REQM sugerem que o modelo mais adequado é a RL. Como comentado anteriormente, o EC e a estatística de Hosmer-Lemeshow podem não ser confiáveis, então são analisados com ressalvas. Tanto o EC quanto $\hat{C}$ privilegiaram a RLB, o segundo indicando inclusive pelo p-valor que a RLB se ajusta bem aos dados, enquanto a RL não.

A RC (por meio da RL) obtida quando se varia uma unidade da variável *Volume* foi de 48.52846, enquanto variando uma unidade da variável *Rate* foi de 14.14156. Ambos indicam que com o aumento dessas variáveis, há também um aumento relevante da chance de ocorrência delas. Já para o RR (por meio da RLB), os valores obtidos para variação de uma unidade nas variáveis foram, respectivamente, 1.597674 e 1.705619. Isso indica que as covariáveis estão sendo mais informativas no caso da RL do que no da RLB e portanto a RL estaria melhor ajustada.

A Figura 3.17 apresenta valores *leverage* próximos de zero para quase todas as observações da RLB, enquanto para a RL os valores são mais altos, apesar de não passarem de 0.2 para quase todas as observações. Há duas observações para as quais foi atribuído

valor 1 na RLB, e ambas possuem *leverage* 1. Com base apenas nisso, a RLB aparenta possuir um melhor ajuste.

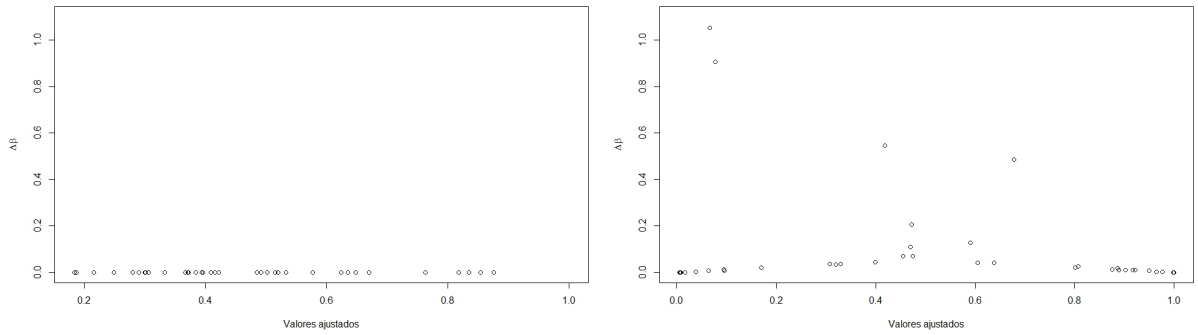


(a) Log-Binomial (b) Logística

Figura 3.17: Estatística *leverage* do banco de dados "Vaso" para ambos os modelos.

A Figura 3.18 foi limitada para a RLB, pois havia duas observações cuja estatística era muito alta, comprometendo a visibilidade do gráfico e a possibilidade de comparação. Essas duas observações foram as mesmas que se destacaram para a RLB na Figura 3.17.

Com exceção das duas observações mencionadas, os valores da RLB são menores que os da RL. Apesar disso, como mencionado, os dois valores para a RLB são muito elevados, indicando forte influência sob o modelo. Sendo assim, há a suspeita de que elas possam ser discrepantes, e a conclusão sobre qual modelo é o mais adequado por essa estatística não é evidente.

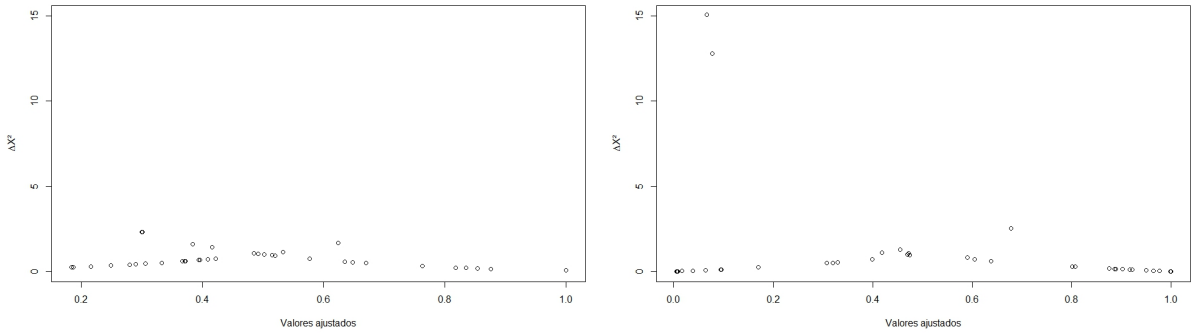


(a) Log-Binomial (b) Logística

Figura 3.18: $\Delta\beta_j$ (exceto duas observações da RLB) do banco de dados "Vaso" para ambos os modelos.

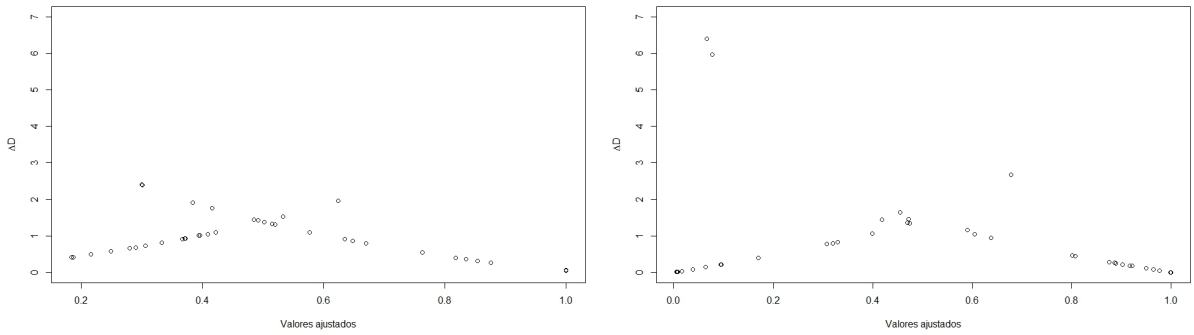
A Figura 3.19 não deixa tão claro qual modelo está fornecendo resultados melhores. Como comparar as observações de forma pareada não faz sentido, optou-se então por ordenar as estatísticas e observar a dispersão delas. Constatou-se que em mais de 92% dos casos as observações da RL foram inferiores às da RLB, o que é indicativo de um melhor ajuste. O mesmo ocorre na Figura 3.20. Nas Figuras 3.18b, 3.19b e 3.20b, as duas

observações que se sobressaem são as mesmas em todos. Apesar disso, como a Figura 3.17b não destaca tais observações, elas não serão consideradas *outliers*. Analogamente, vale o mesmo para as observações destacadas nas Figuras 3.17a e 3.18a.



(a) Log-Binomial (b) Logística

Figura 3.19: ΔX_j^2 do banco de dados "Vaso" para ambos os modelos.



(a) Log-Binomial (b) Logística

Figura 3.20: ΔD_j do banco de dados "Vaso" para ambos os modelos.

Com todas essas informações, parece válido considerar que a RL se ajusta melhor a esse banco de dados.

3.4.2 Banco de dados "Death"

O banco de dados "Death" também pode obtido através do SAS. Ele possui as variáveis *death* (variável resposta dicotômica), *blackd* (etnia do réu dicotomizada), *whitvic* (etnia da vítima dicotomizada), uma medida indicando a culpabilidade da vítima chamada *culp* e duas medidas diferentes para indicar a seriedade do crime, chamadas *serious* e *serious2*. O modelo a ser ajustado segue uma lógica semelhante ao do caso do banco "Vaso".

Mais uma vez, foram obtidas as medidas de ajuste para ambos os modelos.

Apenas a estatística $\mathbf{X}^2$ favoreceu a RLB nas comparações. Ainda assim, vale observar que os REQM tiveram valores próximos entre si, podendo estar indicando equivalência

Tabela 3.7: Medidas de Ajuste para o banco de dados "Death" ($n = 147$).

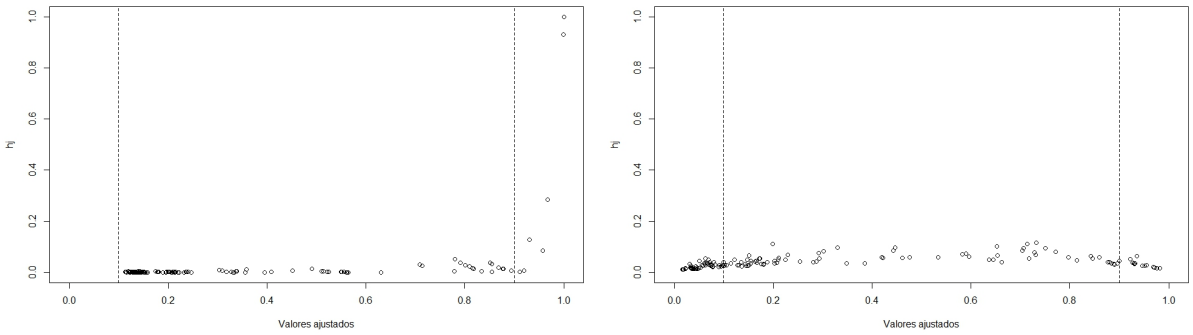
Modelo	*Deviance*	REQM	EC	$\mathbf{X}^2$	$\hat{C}$ (Hosmer-Lemeshow)	p-valor do teste
RLB	123	0.36	26	120	14	0.08
RL	110	0.34	24	135	5	0.73

entre os modelos. A estatística de Hosmer-Lemeshow, por sua vez, indicou um bom ajuste da RL, enquanto forneceu um p-valor para a RLB baixo o suficiente para que a decisão de considerar um modelo bem ajustado dependa do valor crítico a ser estabelecido.

Pela Tabela 3.8, observa-se que as medidas parecem concordar entre si na questão de se há ou não diferença nas probabilidades e chances quando se altera em uma unidade cada uma das variáveis em separado. Ainda assim, a RC fornece valores mais altos, o que pode servir de indício de que as variáveis estão sendo mais informativas para esse modelo.

Tabela 3.8: RR e RC obtidos com alteração em uma unidade de cada variável.

Medida	*blackd*	*whitvic*	*serious*	*culp*	*serious2*
RR	1.18	1.11	0.98	1.61	1.02
RC	5.07	2.44	0.86	3.62	1.66

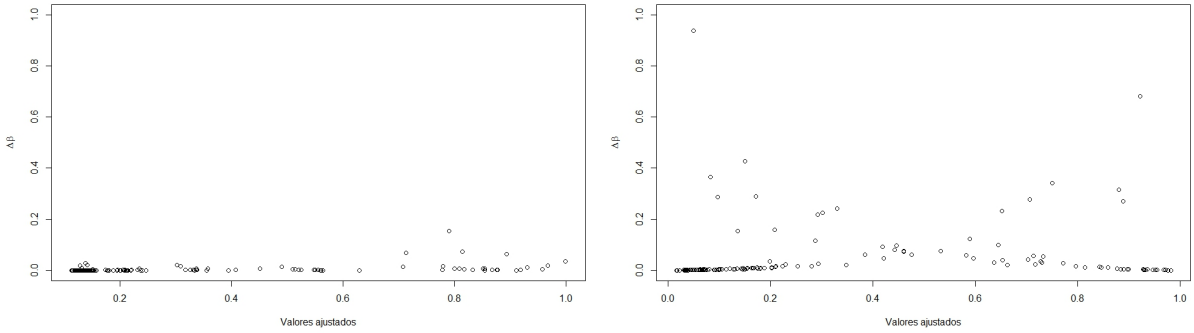


(a) Log-Binomial (b) Logística

Figura 3.21: Estatística *leverage* do banco de dados "Death" para ambos os modelos.

O gráfico acima mostra, mais uma vez, os valores *leverage* da RLB são inferiores aos da RL, apesar de terem crescimento vertiginoso após 0.9. Como não se pode interpretar isso como afastamento da média, a conclusão a se tomar sobre qual modelo escolher não fica clara. Além disso, observa-se que não probabilidades estimadas para a RLB antes de 0.1, o que pode ser um sinal de ajuste inadequado.

A Figura 3.22 foi limitada para a RLB por possuir uma observação de valor próximo de 7000, pois apresentar essa observação no gráfico poderia comprometer a comparação entre os modelos. De modo geral, a RLB apresenta observações com menor influência sobre os coeficientes do que a RL. Apesar disso, os valores de $\Delta\beta_j$ da RL não são tão altos, pois não ultrapassam 1. Considerando essa informações, inclusive a observação de



(a) Log-Binomial (b) Logística

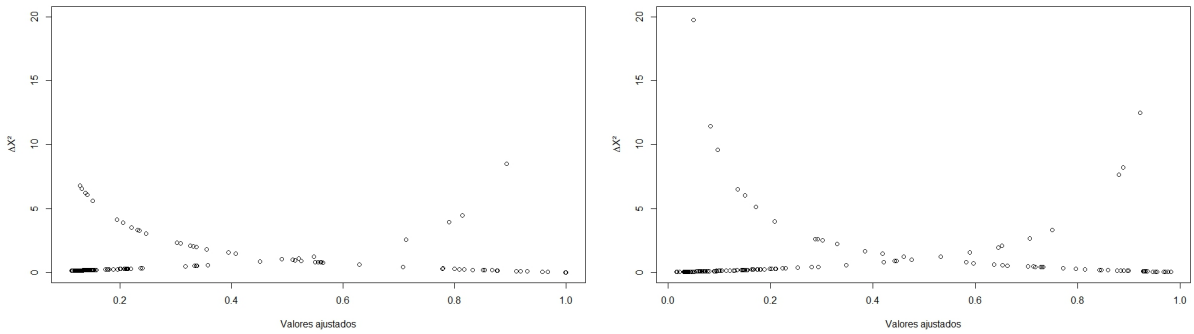
Figura 3.22: $\Delta\beta_j$ (limitado para a RLB) do banco de dados "Death" para ambos os modelos.

valor elevado da RLB, ainda não fica claro qual modelo estaria melhor ajustado para os dados.

A Figura 3.23 revela que a estrutura de dispersão da estatística é parecida nos dois casos, mas que a RL possui observações distribuídas num intervalo maior, que pode ser um indicativo de pior ajuste.

As duas observações mais extremas da Figura 3.23b são as mesmas observações indicadas pela Figura 3.22. Mas, ao menos para a observação extrema cujo valor ajustado é próximo de 1, não parece haver o indicativo de ser observação discrepante.

Realizando um ordenamento das estatísticas e comparando entre as ordenadas do outro modelo, observou-se quem em mais de 86% dos casos a RL forneceu valores menores do que a RLB.

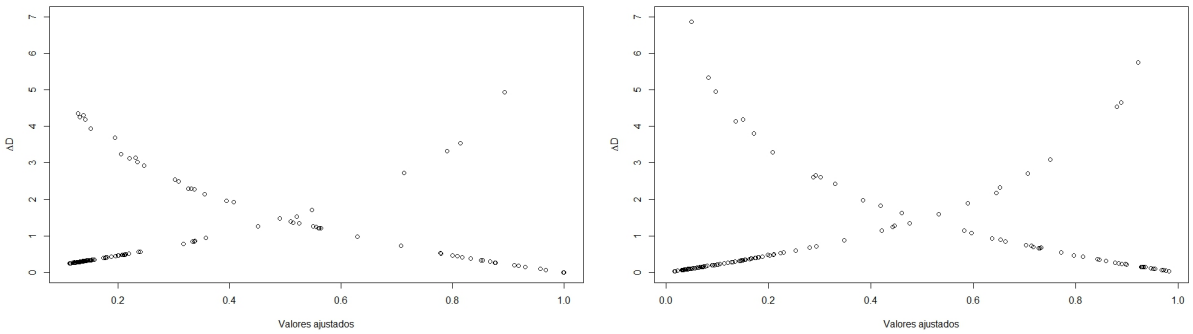


(a) Log-Binomial (b) Logística

Figura 3.23: ΔX_j^2 do banco de dados "Death" para ambos os modelos.

Na Figura 3.24, mais uma vez a RL apresenta observações mais extremas do que a RLB. Além disso, foi observado que a observação mais extrema da RL na Figura 3.24b é a mesma comentada para as Figuras 3.22b e 3.23b. Ainda assim, parece estar próxima demais das outras observações para se afirmar que se trata de uma observação discrepante.

Realizando o ordenamento das estatísticas como no caso anterior, constatou-se que em mais de 87% dos casos as da RLB foram superiores às da RL. Com base em todas essas informações, não se julgou que existem observações discrepantes em nenhum dos casos.



(a) Log-Binomial (b) Logística

Figura 3.24: ΔD_j do banco de dados "Death" para ambos os modelos.

A partir de todas essas informações, a conclusão ainda não é tão clara, mas parece haver uma ligeira preferência pelo uso da RL.

Capítulo 4

Considerações Finais

O presente trabalho buscou avaliar a capacidade de determinadas medidas estabelecerem qual modelo deveria ser utilizado entre a RL e a RLB. Pelo que foi apresentado, observou-se que o EC não possui uma boa capacidade de comparação, enquanto o *deviance* e o REQM apresentaram alguma capacidade, além de fornecerem conclusões concordantes na maioria dos casos avaliados. Assim, parece válido utilizar essas medidas para se obter alguma referência quanto a qual seria o modelo mais adequado.

Outro aspecto que foi avaliado foi a afirmação em Blizzard e Hosmer [2] de que a estatística de Hosmer-Lemeshow teria distribuição χ^2_8 quando o número de grupos é dez. Essa afirmação se mostrou duvidosa, e o poder do teste foi baixo em casos de importância, especialmente o de diferenciação entre os *links*. Também se levantou a suspeita de que a estatística de Hosmer-Lemeshow se torna menos poderosa conforme aumenta o número de covariáveis. Com base nessas questões, sugere-se que não se utilize essa medida para tomar decisões.

Uma característica que se manteve nos casos reais avaliados foi o de que a estatística *leverage* e as de $\Delta\beta_j$ fornecem valores menores na RLB durante quase todo o intervalo de probabilidades estimadas, tendo um aumento quando elas se aproximam de 1. As outras medidas de diagnóstico em geral apresentam valores maiores na RLB, mas as da RL apresentam maior valor máximo. Para se ter uma ideia mais clara disso é necessário um número maior de casos para avaliação. Em decorrência disso, muitas vezes houve dificuldade em se tomar uma decisão acerca do ajuste com base apenas nas medidas de diagnóstico.

Por último, o autor acredita que analisar bancos de dados em que é válido o uso de m-assintótica pode ajudar a revelar mais informações sobre a comparação, além de permitir, de acordo com Hosmer e Lemeshow [5], que sejam feitos outros testes de hipótese.

Referências Bibliográficas

[1] ANDRADE, B. B. Estimation of Log-Binomial Regression via Nonlinear Programming in R. **Biomed** (artigo submetido). 1, 5, 10, 11, 12, 18, 43

[2] BLIZZARD, L.; HOSMER, D. W. Parameter Estimation and Goodness-of-Fit in Log Binomial Regression. **Biometrical Journal** vol. 48, 2006, p. 5-22. i, 2, 3, 6, 10, 14, 18, 19, 23, 49

[3] CASELLA, G.; BERGER, R. L. **Inferência Estatística**. 2 ed, São Paulo, Cengage Learning, 2010. 9

[4] HOSMER, D. W. et. al. A Comparison of Goodness-of-fit Tests for The Logistic Regression Model **Statistics in Medicine**, vol. 16, 1997, p. 965-980. 6, 23, 25

[5] HOSMER, D. W.; LEMESHOW, S. **Applied Logistic Regression**. 3 ed, New Jersey, John Wiley and Sons, 2013. 2, 8, 13, 14, 15, 17, 30, 49

[6] JAMES, G. et al. **An Introduction to Statistical Learning**. New York, Springer, 2014. 17

[7] KOENKER, R. Parametric Links for Binary Response. **R News**, vol. 6/4, oct. 2006.

[8] PAULA, G. A. **Modelos de Regressão com Apoio Computacional**. 2013. Disponível em: https://www.ime.usp.br/∼giapaula/texto_2013.pdf. Acesso em: 31/05/2015. 7

[9] PREGIBON, D. Goodness of Link Tests for Generalized Linear Models. **Journal of the Royal Statistical Society**, Series C, vol. 29, n. 1, p. 15-23, 1980. 2, 5

[10] PRENTICE, R.L. A Generalization of the Probit and Logit Methods for Dose Response Curves. **Biometrics**, vol. 32, n. 4, p. 761-768, 1976.