Cecília Satie Shiraiwa

Tópicos em Regressão para Riscos Relativos

Brasília

2019

Cecília Satie Shiraiwa

Tópicos em Regressão para Riscos Relativos

Dissertação apresentada como requisito para obtenção do título de Mestre em Estatística pelo Programa de Pós-Graduação em Estatística da Universidade de Brasília

Universidade de Brasília

Instituto de Ciências Exatas

Departamento de Estatística

Orientador: Prof. Dr. Bernardo Borba de Andrade

Brasília

2019

Agradecimentos

Sou grata:

Ao meu estimado orientador, professor Bernardo, pela confiança e pelos valiosos ensinamentos;

Ao professor Nakano, pelas contribuições na qualificação e por aceitar novamente fazer parte da banca;

Ao professor Manoel pela disposição em vir de Campina Grande para fazer parte da banca;

Ao coordenador do Programa de Pós-Graduação em Estatística, professor Raul, pelo apoio;

Aos professores do departamento por compartilharem tão generosamente seus conhecimentos;

Aos colegas de curso pelo companheirismo;

À minha família e amigos por compreenderem minha ausência nas horas de dedicação aos estudos;

Aos meus pais que despertaram em mim, desde cedo, a curiosidade e o desejo pelo conhecimento e com sua dedicação me permitiram trilhar meu caminho;

A Deus pelas oportunidades que me foram concedidas.

Não é sobre chegar ao topo do mundo e saber que venceu
É sobre escalar e sentir que o caminho te fortaleceu.
Trem Bala - Ana Vilela

Resumo

O modelo de regressão logística é, provavelmente, o modelo para dados binários mais popular. Algumas justificativas para sua ampla utilização são a simplicidade computacional e o uso da razão de chances em estudos de caso controle. Entretanto, sua interpretação não é tão trivial, comumente levando à supervalorização do efeito das covariáveis. O modelo log-binomial oferece uma interpretação dos resultados de forma mais intuitiva, por meio do risco relativo. O fato do modelo log-binomial apresentar restrições no espaço paramétrico torna o método para seu ajuste mais complexo. Entretanto, já existem soluções para lidar com essas restrições. Neste trabalho serão abordadas as vantagens do modelo log-binomial, além das soluções para lidar com a restrição no espaço paramétrico e comentaremos o processo de estimação considerando as abordagens clássica e bayesiana. Por fim, estudaremos, via simulação e reamostragem, o comportamento dos testes de Vuong e de Cox para escolha do melhor modelo, na comparação da regressão log-binomial com a regressão logística.

Palavras-chave: Risco relativo. Regressão log-binomial. Dados binários. Comparação de modelos. Simulação. Análise bayesiana.

Abstract

The logistic regression model is probably the most popular model for binary data. Some justifications for its wide use are computational simplicity and the use of odds ratio in case control studies. However, its interpretation is not so trivial, usually leading to overvaluation of the covariates effects. Log-binomial model offers more intuitive results interpretation, by relative risk. The fact that log-binomial model presents constraints in the parametric space makes fit method more complex. However, there are solutions available to deal with these constraints. In this work we will discuss the advantages of the log-binomial model, as well as the solutions to deal with the constraint in the parametric space, and we will discuss the estimation process considering the classical and Bayesian approaches. Finally, we will study Vuong and Cox tests to choose the best model between log-binomial regression and logistic regression by simulation and resampling.

Keywords: Relative Risk. Log-binomial regression. Binary data. Model comparison. Simulation. Bayesian analysis.

Lista de tabelas

Lista de abreviaturas e siglas

EM: *Expectation Maximization*

EMV: Estimador de Máxima Verossimilhança

MLG: Modelos Lineares Generalizados

MQO: Método de Mínimos Quadrados Ordinários

MQP: Método de Mínimos Quadrados Ponderados

MQPI: Método de Mínimos Quadrados Ponderados Iterativamente

PGD: Processo Gerador de Dados

Sumário

1 Introdução

Modelos de regressão para dados dicotômicos são amplamente utilizados em diversas áreas, como na medicina, na economia, na engenharia, etc. Frequentemente estamos interessados em saber quais aspectos estão relacionados com um determinado evento de interesse, por exemplo, quais características dos pacientes que podem estar relacionadas com o desenvolvimento de determinada doença; quais as característica dos clientes que têm mais propensão a realizar uma compra; quais características dos clientes de uma instituição financeira podem indicar que ele não honrará seus compromissos.

Por vezes, mesmo respostas que são intervalares são transformadas em binárias pela facilidade de lidar com os dados e possibilidade de padronização. Pode ser definido um valor de corte da variável intervalar, de tal forma que valores acima desse corte são considerados de uma classe e valores abaixo, de outra classe. Um exemplo disso é a definição de ativo problemático dada pelo Banco Central e que todas as instituições financeiras devem atender, são considerados ativos problemáticos, contratos que ultrapassam 90 dias em atraso.

O modelo de regressão logística é o mais conhecido e utilizado em virtude de possuir soluções computacionais há mais tempo e por se acreditar que seus parâmetros são de fácil interpretação.

A dificuldade no ajuste do modelo log-binomial situa-se na necessidade de impor uma restrição no seu espaço paramétrico. Entretanto, a capacidade de processamento dos computadores atuais tornaram viáveis algoritmos de otimização para solução de variados problemas. Além disso, a interpretação dos parâmetros da regressão logística, por meio da razão de chances, é simples quanto à direção da covariável (diretamente ou inversamente proporcional à resposta), porém a interpretação quanto à magnitude da chance não é trivial.

Uma alternativa ao modelo logístico é o modelo log-binomial, cujos parâmetros são mais facilmente interpretados, por meio do risco relativo. Modelos de regressão para riscos relativos são amplamente utilizado em áreas como epidemiologia (CHAO et al., 2010), medicina (LI et al., 2009), nutrição (CHATZI et al., 2012), farmacologia (LIU et al., 2015), psicologia (GLOVER et al., 2011), ciências sociais (JUN; ACEVEDO-GARCIA, 2007), ecologia (WILLIAMSON; GASTON, 2005).

Apesar da maior facilidade de interpretação dos parâmetros e da popularidade em áreas como a ciências sociais e epidemiologia, o modelo log-binomial não é muito explorado em outros segmentos como economia e finanças. Nesses campos, o modelo logístico é mais comumente aplicado, possivelmente porque no passado o modelo log-binomial apresentava

dificuldade de ajuste.

Mesmo nas áreas em que o modelo log-binomial é habitual, muitas vezes são utilizados métodos indiretos de estimação, entretanto, já existem soluções que permitem obter estimativas robustas diretamente pelo método de máxima verossimilhança.

Este trabalho tem o intuito de evidenciar que atualmente os problemas de estimação do modelo log-binomial foram solucionados, estudar testes de comparação de modelos não encaixados, mais especificamente na comparação de funções de ligação e descrever o estado da arte em termos de inferência bayesiana.

Nos próximos capítulos serão abordados assuntos relacionados à regressão log-binomial. No Capítulo 2 serão detalhadas as medidas de razão de chances e risco relativo, bem como sua relação com os modelos logístico e log-binomial respectivamente, além de esclarecer porque a primeira medida é mais inteligível que a segunda. Ainda no Capítulo 2 serão descritos os problemas que podem ocorrer na estimação dos parâmetros e como podem ser solucionados. No Capítulo 3 serão estudados testes de comparação de modelos não aninhados por meio de cenários simulados e reamostragem de dados reais presentes na literatura. O Capítulo 4 tratará da análise bayesiana para o modelo log-binomial.

2 O Modelo Log-Binomial e Avanços Recentes

Modelos para dados dicotômicos são amplamente utilizados em diversas áreas do conhecimento. O modelo logístico é o mais conhecido, entretanto, o entendimento dos parâmetros, em especial sua magnitude não é trivial, como será visto neste capítulo.

Uma opção de modelo para resposta binária é o modelo log-binomial, cujos parâmetros são facilmente interpretados por meio do risco relativo.

2.1 Risco Relativo vs Razão de Chances

As medidas de associação mais utilizadas em áreas biomédicas são o risco relativo e a razão de chances, as quais revisamos a seguir. Na sequência, os respectivos modelos de regressão (log-binomial e logístico) são apresentados.

2.1.1 Razão de Chances

Chance (C) é a razão entre a probabilidade p de ocorrer um evento de interesse e a probabilidade $1-p$ de não ocorrer, ou seja,

$$\mathrm{C} = \frac{p}{1-p}.$$

Se temos dois grupos de indivíduos, então, podemos calcular a Razão de Chances (RC), que é a razão entre a chance no grupo 1 e a chance no grupo 2,

$$\mathrm{RC} = \frac{\frac{p_1}{1-p_1}}{\frac{p_2}{1-p_2}},$$

em que p_1 é a probabilidade de ocorrência de um determinado evento de interesse no grupo 1 e p_2 é a probabilidade de ocorrência no grupo 2.

A Razão de Chances pode assumir valores de zero a infinito. Caso seja maior que um, pertencer ao grupo 1 é um fator de risco; se menor que um, pertencer ao grupo 1 é um fator de proteção (em geral, o evento de interesse se refere a algo negativo, como doença, inadimplência, acidente, etc.); se igual a um, a chance nos dois grupos é a mesma.

No contexto do risco de crédito, por exemplo, temos interesse em identificar clientes que possam vir a descumprir suas obrigações de pagamento. É comum usarmos o termo "mau" para clientes que chegam à inadimplência e "bom" para os demais. Vamos supor que

Tabela 1 – Exemplo - Risco Relativo e Razão de Chances

	Mau	Bom	Total	p
Proprietário de imóveis	50	150	200	0,25
Não proprietário de imóveis	100	100	200	0,50
Total	150	250	400	-

observamos as seguintes contagens para os grupos da Tabela 1, neste exemplo hipotético, somente o total foi fixado.

No caso dos clientes que possuem imóveis, a chance de ser mau é dada por:

$$\mathrm{C}_{ComImo} = \frac{0,25}{1-0,25} = 0,33.$$

No caso dos clientes que não possuem imóveis, a chance de ser mau é dada por:

$$\mathrm{C}_{SemImo} = \frac{0,50}{1-0,50} = 1.$$

A razão de chances é:

$$\mathrm{RC} = \frac{0,33}{1} = 0,33.$$

Como a razão de chances é menor que um, possuir imóveis é um fator de proteção, ou seja, existe menor chance dos clientes que possuem imóveis de serem maus do que aqueles que não possuem imóveis.

Chance tem um interpretação distinta da probabilidade. No exemplo acima, a probabilidade de um cliente proprietário de imóveis ser um mau pagador é 0,5, já a chance de um cliente não proprietário de imóveis ser um mau pagador é 1. Chance não é uma medida usada rotineiramente e sua gradação pode confundir usuários menos familiarizados com seu cálculo.

2.1.2 Risco Relativo

Risco (R) é uma medida mais comum a todas as áreas e é equivalente à probabilidade.

$$\mathrm{R} \equiv p.$$

Se temos dois grupos de indivíduos, então podemos calcular o Risco Relativo (RR) que é a razão entre o risco no grupo 1 e o risco no grupo 2.

$$\mathrm{RR} = \frac{p_1}{p_2}.$$

Como exemplo, podemos calcular os riscos da Tabela 1 da seguinte forma:

$$\mathrm{R}_{ComImo} = 0,25,$$

$$\mathrm{R}_{SemImo} = 0,50,$$
$$\mathrm{RR} = \frac{0,25}{0,50} = 0,50.$$

A probabilidade (ou risco) de um cliente proprietário de imóveis ser mau pagador é metade da probabilidade de um clientes não proprietário de imóveis ser mau pagador.

Assim como a Razão de Chances, o Risco Relativo pode assumir valores de zero a infinito e, caso seja maior que um, pertencer ao grupo 1 é um fator de risco; se menor que um, pertencer ao grupo 1 é um fator de proteção; se igual a um, o risco nos dois grupos é o mesmo. Entretanto, a sua gradação é mais facilmente interpretada, uma vez que o valor do risco equivale ao valor da probabilidade do evento.

Vale ressaltar que quando o total nas colunas é fixo, como nos ensaios de caso-controle, não existe sentido no risco relativo, já a razão de chances pode ser utilizada.

Quando o evento é raro, a razão de chances é aproximadamente igual ao risco relativo. Entretanto, quando o evento é comum, interpretar a razão de chances como se fosse o risco relativo pode conduzir a uma associação exagerada. Por exemplo, se $p_1 = 0,01$ e $p_2 = 0,02$, então a razão de chances é $0,4949$ e o risco relativo $0,5000$, já se $p_1 = 0,9$ e $p_2 = 0,8$, a razão de chances é $2,2500$ e o risco relativo $1,1250$.

2.1.3 Regressão Logística e Razão de Chances

O modelo logístico é um modelo linear generalizado (MLG), com distribuição Bernoulli e função de ligação $g(\cdot)$ dada pelo logaritmo neperiano da chance, conhecido como Logit. Seja $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ik})$ um vetor de k covariáveis mais um intercepto, $y_i \sim Bernoulli(p_i)$, $i = 1, 2, ..., n$, de tal forma que $P(y_i = 1|\mathbf{x}) = p_i$, então, o modelo de regressão logístico é dado por

$$g(p_i) = \mathrm{logit}(p_i) = \log(\mathrm{Chance}) = \log\left(\frac{p_i}{1-p_i}\right) = \mathbf{x}_i'\boldsymbol{\beta},$$

em que $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ é o vetor de coeficientes de regressão. Consequentemente,

$$p_i = \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{1+\exp(\mathbf{x}_i'\boldsymbol{\beta})} = \frac{1}{1+\exp(-\mathbf{x}_i'\boldsymbol{\beta})}.$$

Os coeficientes da regressão logística podem ser interpretados por meio da razão de chances. Vamos supor, sem perda de generalidade, um modelo com duas variáveis explicativas, x_1 e x_2. O índice i foi omitido para simplificar a notação. Se queremos calcular a razão de chances entre aqueles que apresentam determinado valor de x_1 e aqueles que apresentam uma unidade a mais, $x_1 + 1$, mantendo as demais variáveis constantes, temos:

$$\log(\mathrm{Chance}) = \mathbf{x}'\boldsymbol{\beta} = \beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2,$$

$$\mathrm{Chance} = \exp(\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2),$$

$$\text{RC} = \frac{\exp(\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2)}{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)} = \exp(\beta_1).$$

Portanto, se $\beta_1 < 0$, então $0 < \text{RC} < 1$ e um aumento em x_1 é um fator de proteção. Se $\beta_1 > 0$, então $\text{RC} > 1$ e um aumento em x_1 é um fator de risco. Se $\beta_1 = 0$, então $\text{RC} = 1$ e a chance é a mesma para qualquer valor de x_1.

2.1.4 Regressão Log-Binomial e Risco Relativo

O modelo log-binomial é um MLG, com distribuição Bernoulli e função de ligação $g(.)$ dada pelo logaritmo neperiano. Seja $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ik})$ um vetor de k covariáveis mais um intercepto, $y_i \sim Bernoulli(p_i)$, $i = 1, 2, ..., n$, de tal forma que $P(y_i = 1|\mathbf{x}) = p_i$, então, o modelo de regressão log-binomial é dado por

$$g(p_i) = \log(p_i) = \mathbf{x}_i'\boldsymbol{\beta},$$

$$p_i = \exp(\mathbf{x}_i'\boldsymbol{\beta}), (\mathbf{x}_i'\boldsymbol{\beta}) \leq 0.$$

Observe que no modelo log-binomial, $p_i = \exp(\mathbf{x}_i'\boldsymbol{\beta})$ e a imagem da função exponencial pode apresentar valores de zero a infinito, mas só temos interesse em valores de p_i entre zero e um. Portanto, para que $p_i \in (0, 1)$ precisamos garantir que $(\mathbf{x}_i'\boldsymbol{\beta}) \leq 0$.

Os coeficientes da regressão log-binomial $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, ..., \beta_k)$ podem ser interpretados por meio do risco relativo. Vamos supor, sem perda de generalidade, um modelo com duas variáveis explicativas, x_1 e x_2. O índice i foi omitido para simplificar a notação. Se queremos entender o efeito de somar uma unidade em x_1, mantendo as demais variáveis constantes, podemos calcular o risco daqueles que apresentam determinado valor $x_1 + 1$ relativo àqueles que apresentam x_1 da seguinte forma:

$$\text{Risco} \equiv p = \exp(\mathbf{x}'\boldsymbol{\beta}),$$

$$\text{RR} = \frac{\exp(\beta_0 + \beta_1(x_1 + 1) + \beta_2 x_2)}{\exp(\beta_0 + \beta_1 x_1 + \beta_2 x_2)} = \exp(\beta_1).$$

A interpretação quanto a ser um fator de risco ou proteção, é semelhante à da regressão logística. Se $\beta_1 < 0$, então $0 < \text{RR} < 1$ e um aumento em x_1 é um fator de proteção. Se $\beta_1 > 0$, então $\text{RR} > 1$ e um aumento em x_1 é um fator de risco. Se $\beta_1 = 0$, então $\text{RR} = 1$ e o risco é o mesmo para qualquer valor de x_1.

Entretanto, entender a magnitude do risco ou da proteção é bem mais acessível na regressão log-binomial. As Tabelas 2 e 3 mostram exemplos em que $\beta_1 = 0,6931$ e, portanto, $\exp(\beta_1) = 2$. Seja a um valor qualquer pertencente ao domínio de x_1, vamos denotar $P_1 = P(y_i = 1|x_1 = a)$ e $P_2 = P(y_i = 1|x_1 = a + 1)$. Portanto, a chance de $y_i = 1$ quando $x_1 = a$ é $C_1 = P_1/(1 - P_1)$ e a chance de $y_i = 1$ quando $x_1 = a + 1$ é $C_2 = P_2/(1 - P_2)$. O risco de $y_i = 1$ quando $x_1 = a$ é $R_1 \equiv P_1$ e o risco de $y_i = 1$ quando

$x_1 = a + 1$ é $R_2 \equiv P_2$. Assim a razão de chances é dada por RC $= C_2/C_1$ e o risco relativo por RR $= R_2/R_1$. Na Tabela 2, C_2 é o dobro de C_1 e para que isso ocorra é preciso que $P_2 = 2P_1/(1 + P_1)$. Na Tabela 3, R_2 é o dobro de R_1 e para que isso ocorra, a relação entre P_1 e $P2$ é bem mais simples, $P_2 = 2P_1$. Observe que se $P_1 > 0,5$, o risco relativo nunca poderá ser igual a 2, pois P_2 não pode ser maior que 1, já a razão de chances pode apresentar qualquer valor positivo.

Tabela 2 – Exemplos com RC = 2

P_1	P_2	C_1	C_2	RR	β_1^{logit}	β_1^{log}
0,0100	0,0198	0,0101	0,0202	1,9802	0,6931	0,6832
0,0500	0,0952	0,0526	0,1053	1,9048	0,6931	0,6444
0,1000	0,1818	0,1111	0,2222	1,8182	0,6931	0,5978
0,1500	0,2609	0,1765	0,3529	1,7391	0,6931	0,5534
0,2000	0,3333	0,2500	0,5000	1,6667	0,6931	0,5108
0,2500	0,4000	0,3333	0,6667	1,6000	0,6931	0,4700
0,3000	0,4615	0,4286	0,8571	1,5385	0,6931	0,4308
0,3500	0,5185	0,5385	1,0769	1,4815	0,6931	0,3930
0,4000	0,5714	0,6667	1,3333	1,4286	0,6931	0,3567
0,4500	0,6207	0,8182	1,6364	1,3793	0,6931	0,3216

Nota: β_1^{logit} é o valor de β_1 se o modelo de regressão for o logístico e β_1^{log} é o valor de β_1 se o modelo regressão for log-binomial.

Tabela 3 – Exemplos com RR = 2

P_1	P_2	C_1	C_2	RC	β_1^{logit}	β_1^{log}
0,0100	0,0200	0,0101	0,0204	2,0204	0,7033	0,6931
0,0500	0,1000	0,0526	0,1111	2,1111	0,7472	0,6931
0,1000	0,2000	0,1111	0,2500	2,2500	0,8109	0,6931
0,1500	0,3000	0,1765	0,4286	2,4286	0,8873	0,6931
0,2000	0,4000	0,2500	0,6667	2,6667	0,9808	0,6931
0,2500	0,5000	0,3333	1,0000	3,0000	1,0986	0,6931
0,3000	0,6000	0,4286	1,5000	3,5000	1,2528	0,6931
0,3500	0,7000	0,5385	2,3333	4,3333	1,4663	0,6931
0,4000	0,8000	0,6667	4,0000	6,0000	1,7918	0,6931
0,4500	0,9000	0,8182	9,0000	11,0000	2,3979	0,6931

Nota: β_1^{logit} é o valor de β_1 se o modelo de regressão for o logístico e β_1^{log} é o valor de β_1 se o modelo regressão for log-binomial.

Se x_1 é intervalar, dobrar o risco pelo acréscimo de uma única unidade em seu valor parece exagerado e incomum. Entretanto, se considerarmos covariáveis binárias de presença/ausência de uma determinada característica, valores altos de RR podem ser mais frequentes. Por exemplo, se x_1 for dicotômica e

$$x_1 = \begin{cases} 1, & \text{se cliente já atrasou pagamentos no passado} \\ 0, & \text{se cliente nunca atrasou pagamentos} \end{cases}$$

então, $\exp(\beta_1)$ é a probabilidade de se observar o evento de interesse (mau pagador) dos clientes que já atrasaram pagamentos, relativa à probabilidade dos que nunca atrasaram pagamentos, mantendo-se as demais varáveis constantes.

Devido a essa relação, o modelo log-binomial também é chamado de modelo de regressão de risco relativo.

2.2 Estimação Frequentista do Modelo Log-Binomial

Seja $\mathbf{x}_i = (1, x_{i1}, x_{i2}, ..., x_{ik})$ um vetor de k covariáveis mais um intercepto, $y_i \sim Bernoulli(p_i)$, $i = 1, 2, ..., n$, de tal forma que $P(y_i = 1|\mathbf{x}) = p_i$ e $\log(p_i) = \mathbf{x}_i'\boldsymbol{\beta}$, a função de verossimilhança do modelo log-binomial é dada por

$$L(p_i) = \prod_{i=1}^{n} p_i^{y_i}(1 - p_i)^{1-y_i},$$

que equivale a

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} \exp(\mathbf{x}_i'\boldsymbol{\beta})^{y_i}(1 - \exp(\mathbf{x}_i'\boldsymbol{\beta}))^{1-y_i}.$$

Observe que ocorre uma redução de dimensionalidade, não é necessário estimar n parâmetros p_i, $i = 1, 2, ..., n$, mas somente $k + 1$ parâmetros β_j, $j = 0, 1, 2, ..., k$.

A log-verossimilhança é dada por

$$\ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} y_i \log(\exp(\mathbf{x}_i'\boldsymbol{\beta})) + \sum_{i=1}^{n}(1 - y_i)\log(1 - \exp(\mathbf{x}_i'\boldsymbol{\beta})),\ (\mathbf{x}_i'\boldsymbol{\beta}) \leq 0.$$

Em geral, os softwares de análise estatística utilizam o método de mínimos quadrados ponderados iterativamente para estimação de modelos lineares generalizados por máxima verossimilhança. Entretanto, esse método pode apresentar baixa taxa de convergência no caso do modelo log-binomial pois os algoritmos usuais não tratam restrições durante o processo de estimação dos parâmetros. No modelo log-binomial o espaço paramétrico é restrito sendo necessário garantir que $\mathbf{x}_i'\boldsymbol{\beta} \leq 0$. Se o algoritmo não considerar essa condição, pode não ser capaz de chegar a valores dentro do espaço paramétrico ou se tornar tão lento que seria impraticável sua utilização. Mais detalhes sobre os motivos que levam à falta de convergência, são descritos nas próximas seções.

Note que o termo $\log(0)$, decorrente de $\mathbf{x}_i'\boldsymbol{\beta} = 0$, só correria quando $y_i = 0$ e o termo $0\log(0) = 0$.

Andrade e Andrade (2017) propõem um método de estimação por programação não linear, por meio de barreira adaptativa logarítmica que não apresenta essa limitação. Andrade (2018), disponibiliza uma ferramenta de aplicação desse método por meio do pacote `lbreg` implementado na linguagem `R`.

2.3 Problemas de Estimação

Nesta seção discutiremos cinco problemas que podem ocorrer na estimação dos parâmetros: separação completa, inicialização fora da fronteira, falta de convergência devido à restrição no espaço paramétrico, repulsão e estimador de máxima verossimilhança (EMV) na fronteira.

Nos ateremos a estimação pelo método da máxima verossimilhança, discussões sobre métodos de quasi-verossimilhança para o modelo log-binomial podem ser encontradas em Marschner (2015) e Andrade e Andrade (2017).

2.3.1 Separação

O problema de separação completa, também chamado de verossimilhança monotônica, ocorre no modelo logístico quando não é possível determinar o valor de algum coeficiente β_j que maximiza a função de verossimilhança, pois o coeficiente diverge para $-\infty$ ou $+\infty$. Na prática, ocorre quando uma variável ou uma combinação linear de variáveis é perfeitamente preditiva.

Uma forma intuitiva de entender o problema de separação no modelo logístico, é por meio da Razão de Chances.

Tabela 4 – Tabela de Contingência

	Y=1	Y=0
X=1	a	b
X=0	c	d

A partir da Tabela 4, podemos calcular a razão da chance de $X = 1$ em relação a $X = 0$ é

$$\text{RC} = \frac{P(Y=1|X=1)/(1-P(Y=1|X=1))}{P(Y=1|X=0)/(1-P(Y=1|X=0))} = \frac{P(Y=1|X=1)/P(Y=0|X=1)}{P(Y=1|X=0)/P(Y=0|X=0)},$$

$$\text{RC} = \frac{\frac{a}{a+b}/\frac{b}{a+b}}{\frac{c}{c+d}/\frac{d}{c+d}} = \frac{a/b}{c/d} = \frac{ad}{bc},$$

portanto, se $b = 0$ e/ou $c = 0$, RC é indeterminada.

Como a escolha da característica que representa $X = 1$ ou $X = 0$ é arbitrária, devemos calcular, então, o caso inverso, isto é, a razão da chance de $X = 0$ em relação a $X = 1$,

$$\text{RC} = \frac{P(Y=1|X=0)/(1-P(Y=1|X=0))}{P(Y=1|X=1)/(1-P(Y=1|X=1))} = \frac{P(Y=1|X=0)/P(Y=0|X=0)}{P(Y=1|X=1)/P(Y=0|X=1)},$$

$$\text{RC} = \frac{\frac{c}{c+d}/\frac{d}{c+d}}{\frac{a}{a+b}/\frac{b}{a+b}} = \frac{c/d}{a/b} = \frac{bc}{ad},$$

portanto, se $a = 0$ e/ou $d = 0$, RC é indeterminada.

Assim, se qualquer uma das caselas a, b, c ou d for igual a zero, a razão de chances não pode ser determinada.

Para simplificação denotamos X como uma variável dicotômica, mas poderia ser uma variável contínua. Por exemplo, poderíamos substituir $X = 1$ e $X = 0$ por $X \leq 10$ e $X > 10$.

Um método comumente utilizado para permitir estimação por máxima verossimilhança em um modelo logístico que apresenta separação completa, é conhecido como Máxima Verossimilhança Penalizada (FIRTH, 1993; HEINZE; SCHEMPER, 2002). Consiste em aplicar uma penalização na função de verossimilhança. Para modelos da família exponencial com ligação canônica é sugerida a seguinte função de verossimilhança penalizada

$$L^*(\boldsymbol{\beta}) = L(\boldsymbol{\beta})|\det(I(\boldsymbol{\beta}))|^{\frac{1}{2}},$$

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

Como será visto na seção 4.2.3 a função de penalização $|\det(I(\boldsymbol{\beta}))|^{\frac{1}{2}}$ equivale a priori de Jeffreys, portanto, as metodologias são equivalentes.

2.3.2 Inicialização

O espaço paramétrico do modelo log-binomial é dado por $\Theta = \{\boldsymbol{\beta} : p_i \in (0,1)\} = \{\boldsymbol{\beta} : \mathbf{x}_i'\boldsymbol{\beta} \leq 0, \forall i = 1, ..., n\}$, portanto,

$$\beta_0 \leq -\sum_{j=1}^{k} \beta_j x_{ij}, \ \forall i. \tag{2.1}$$

Para garantir que o algoritmo inicie dentro do espaço paramétrico, podemos utilizar um dos dois procedimentos descritos a seguir.

Sejam $\beta_j^{(0)}$ os valores iniciais para os parâmetros β_j, $j = 0, 1, ..., k$. A opção mais simples para garantir $\mathbf{x}'\boldsymbol{\beta} \leq 0$ é atribuir um valor negativo para $\beta_0^{(0)}$ e zero para os demais $\beta_j^{(0)}$, $j > 0$.

Andrade e Andrade (2017) sugerem o uso das estimativas do modelo Poisson com função de ligação logarítmica como valores iniciais para β_j, $j > 0$ e uma função dessas estimativas como valor inicial para β_0.

Sejam $\beta_j^{(Poi)}$ os valores das estimativas dos parâmetros β_j, $j = 0, 1, ..., k$, em um modelo de Poisson com função de ligação logarítmica. Estimavas iniciais $\beta_j^{(0)}$ dentro da fronteira do modelo log-binomial são dadas por

$$\beta_j^{(0)} = \beta_j^{(Poi)}, \text{ para } j > 0, \text{ e}$$

$$\beta_0^{(0)} = \left[\min_i \left(-\sum_{j=1}^{k} \beta_j^{(Poi)} x_{ij}\right)\right] - \epsilon,$$

isso garante (2.1). O valor

$$\epsilon = \frac{1}{10}\left|\min_i \left(-\sum_{j=1}^{k} \beta_j^{(Poi)} x_{ij}\right)\right|,$$

é sugerido como referência, mas ϵ pode ser qualquer valor estritamente positivo.

Mesmo utilizando alguma estratégia de inicializaçao, não há garantia de que as iterações permanecerão dentro do espaço paramétrico. A seguir são descritas algumas formas de garantir que as iterações permaneçam em Θ.

2.3.3 Permanência no Espaço Paramétrico

A estimativa de máxima verossimilhança dos parâmetros do modelo de regressão log-binomial é dada pelo vetor $\hat{\boldsymbol{\beta}}$ pertencente ao espaço paramétrico $\Theta = \{\boldsymbol{\beta} : p_i(\boldsymbol{\beta}) \in (0,1)\}$ que maximiza a função de log-verossimilhança,

$$\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\Theta}{\operatorname{argmax}}\ \ell(\boldsymbol{\beta}).$$

O método computacional mais comumente utilizado para resolver essa maximização, é o método de mínimos quadrados ponderados iterativamente (MQPI), que é equivalente ao algoritmo escore de Fisher, esse assunto será detalhado na seção 2.3.3.2.

2.3.3.1 Newton-Raphson

O método iterativo de Newton-Raphson resolve equações não lineares, assim, pode ser empregado para determinar o valor de $\hat{\boldsymbol{\beta}}$ que maximiza $\ell(\boldsymbol{\beta})$. Seja

$$\mathbf{u} = \left(\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_2}, ..., \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_k}\right)'$$

o vetor transposto de derivadas parciais de primeira ordem de $\ell(\boldsymbol{\beta})$, $\mathbf{H}$ a matriz de derivadas parciais de segunda ordem, chamada de matriz Hessiana, formada pelos termos $h_{ab} = \partial^2 \ell(\boldsymbol{\beta})/\partial\beta_a\beta_b$ e $t = 0, 1, 2, ...$ o número da iteração. A função $\ell(\boldsymbol{\beta})$ é aproximada pelo polinômio de segunda ordem da expansão de Taylor,

$$\ell(\boldsymbol{\beta}) \approx \ell\left(\boldsymbol{\beta}^{(t)}\right) + \mathbf{u}^{(t)}\left(\beta - \beta^{(t)}\right) + \frac{1}{2}\left(\beta - \beta^{(t)}\right)' \mathbf{H}^{(t)}\left(\beta - \beta^{(t)}\right),$$

derivando e igualando a zero,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \approx \mathbf{u}^{(t)} + \mathbf{H}^{(t)}\left(\beta - \beta^{(t)}\right) = 0.$$

Assim, o processo iterativo atualiza a estimativa de $\boldsymbol{\beta}$ por

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} - \left(\mathbf{H}^{(t)}\right)^{-1} \mathbf{u}^{(t)}.$$

Marschner (2015) afirma que o método escore de Fisher (2.3.3.2) é mais estável que o de Newton-Rapshon sobretudo quando as iterações estão afastadas da estimativa de máxima verossimilhança. O método de Newton-Rapshon não apresenta problema de repulsão (2.3.4), mas pode atenuar o problema de saída do espaço paramétrico. O autor sugere, então, uma abordagem na qual o algoritmo de Fisher seria usado nas primeiras iterações e então substituído pelo método de Newthon-Rapshon.

2.3.3.2 Score de Fisher e Mínimos Quadrados Ponderados Iterativamente

O método score de Fisher é uma adaptação do método de Newton, a diferença em relação ao método de Newton-Raphson é a forma como utiliza a matriz Hessiana $\mathbf{H}$. O método score de Fisher utiliza os valores esperados dos termos da matriz Hessiana, que formam a chamada matriz de informação esperada, ou simplesmente matriz de informação $\mathcal{I}_e = E(-\mathbf{H})$. O método de Newton-Raphson utiliza os valores da própria matriz Hessiana que formam a chamada matriz de informação observada $\mathcal{I}_o = -(\mathbf{H})$. O processo iterativo método score de Fisher é dado por

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + \left(\boldsymbol{\mathcal{I}}_e^{(t)}\right)^{-1} \mathbf{u}^{(t)}, \tag{2.2}$$

em que $\mathcal{I}_e^{(t)}$ é formada pelos elementos $-E(\partial^2 \ell(\boldsymbol{\beta})/\partial\beta_a\beta_b)$, $a, b = 0, 1, ..., k$.

Nos MLGs $\boldsymbol{\mathcal{I}}_e = \mathbf{X}'\mathbf{W}\mathbf{X}$, em que $\mathbf{W}$ é uma matriz diagonal formada pelos elementos $w_i = (\partial\mu_i/\partial\eta_i)^2/var(y_i)$, $\mu_i = E(y_i)$ e $\eta_i = \beta_0 + \sum_{j=1}^{k} x_{ij}\beta_j$.

O método score de Fisher é equivalente ao MQPI que por sua vez são sucessivos ciclos do método de mínimos quadrados ponderado (MQP). O MQP é usado para estimar $\boldsymbol{\beta}$ em modelos lineares $\mathbf{z} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$. O estimador de mínimos quadrados é dado por

$$\left(\mathbf{X}'\mathbf{W}^{-1}\mathbf{X}\right)^{-1} \mathbf{X}'\mathbf{W}^{-1}\mathbf{z},$$

em que $\mathbf{W}^{-1} = var(\boldsymbol{\epsilon})$. O método é chamado de ponderado, pois dá pesos w_i maiores ou menores para diferentes observações. No método de mínimos quadrados ordinários (MQO) o peso é o mesmo para todos os registros e o estimador de mínimos quadrados é dado por $(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{z}$.

Utilizando resultados de MLGs descritos mais detalhadamente em Agresti (2015), temos

$$\mathbf{u} = \mathbf{X}'\mathbf{W}\mathbf{D}^{-1}(\mathbf{y} - \boldsymbol{\mu}),$$

em que $\mathbf{D} = diag\{\partial\mu_i/\partial\eta_i\}$ e $\mathbf{W} = diag\{(\partial\mu_i/\partial\eta_i)^2/var(y_i)\}$. Assim, multiplicando ambos os lados de (2.2) por $\boldsymbol{\mathcal{I}}_e = \mathbf{X}'\mathbf{W}\mathbf{X}$, temos que

$$(\mathbf{X}'\mathbf{W}^{(t)}\mathbf{X})\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}'\mathbf{W}^{(t)}\mathbf{X})\boldsymbol{\beta}^{(t)} + \mathbf{X}'\mathbf{W}^{(t)}(\mathbf{D}^{(t)})^{-1}(\mathbf{y} - \boldsymbol{\mu}^{(t)})$$

$$(\mathbf{X}'\mathbf{W}^{(t)}\mathbf{X})\boldsymbol{\beta}^{(t+1)} = \mathbf{X}'\mathbf{W}^{(t)}[\mathbf{X}\boldsymbol{\beta}^{(t)} + (\mathbf{D}^{(t)})^{-1}(\mathbf{y} - \boldsymbol{\mu}^{(t)})],$$

então,

$$\boldsymbol{\beta}^{(t+1)} = (\mathbf{X}'\mathbf{W}^{(t)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(t)}\mathbf{z}^{(t)},$$

em que $\mathbf{z}^{(t)}$ tem elementos

$$z_i^{(t)} = \eta_i^{(t)} + \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}(y_i - \mu^{(t)}).$$

A correspondência $\mathbf{z}^{(t)} = \mathbf{X}\boldsymbol{\beta}^{(t)} + (\mathbf{D}^{(t)})^{-1}(\mathbf{y} - \boldsymbol{\mu}^{(t)})$ pode ser vista como um modelo linear

$$\mathbf{z}^{(t)} = \mathbf{X}\boldsymbol{\beta}^{(t)} + \boldsymbol{\epsilon}^{(t)},$$

em que $\boldsymbol{\epsilon}^{(t)} = (\mathbf{D}^{(t)})^{-1}(\mathbf{y} - \boldsymbol{\mu}^{(t)})$. Assim, o vetor $\mathbf{z}^{(t)}$ é uma aproximação linearizada da função de ligação $g(\cdot)$ avaliada em y_i

$$g(y_i) \approx g(\mu_i^{(t)}) + (y_i - \mu^{(t)})g'(\mu_i^{(t)}) = \eta_i^{(t)} + \frac{\partial \eta_i^{(t)}}{\partial \mu_i^{(t)}}(y_i - \mu^{(t)}) = z_i^{(t)}.$$

Assim, a cada ciclo do MQPI, realizamos um ajuste de MQP no qual obtemos $\boldsymbol{\beta}^{(t+1)}$ a partir de $\mathbf{z}^{(t)}$ e X usando os pesos $\mathbf{W}^{(t)}$ (inversa da covariância). A partir de $\boldsymbol{\beta}^{(t+1)}$ temos uma nova aproximação de $\boldsymbol{\eta}^{(t+1)} = \mathbf{X}\boldsymbol{\beta}^{(t+1)}$ e de $\mathbf{z}^{(t+1)}$ pra serem usadas no próximo ciclo.

Assim, obter um EMV de um MLG por MQPI equivale a obter sucessivamente o estimador de mínimos quadrados de um modelo linear correlato. O MQPI foi desenvolvido para permitir ajustar MLGs com as ferramentas computacionais disponíveis na época que lidavam somente com modelos lineares.

2.3.3.3 Fracionamento do Passo

O fracionamento do passo (*Step Halving*) é um método que visa fazer com que os valores estimados nas iterações do algoritmo de escore de Fisher retornem para o interior do espaço paramétrico, caso tenham saído.

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + 0.5^h \mathcal{I}_e(\boldsymbol{\beta}^{(t)})^{-1}\mathbf{u}^{(t)},\ h = 0, 1, 2, ...$$

em que $h = \min\{h \in \mathbb{N} : \boldsymbol{\beta}_{c+1} \in \Theta\}$

Esse método está presente na função `glm` do `R`. Entretanto, o algoritmo se torna lento quando se aproxima da fronteira, uma estratégia para obter convergência pode ser aumentar a opção `maxit` que define o número de iterações máximas que a função executará. Assim, o algoritmo terá mais ciclos para reduzir o tamanho dos passos aumentando a possibilidade de retornar ao interior do espaço paramétrico. Porém, aumentar o número de iterações nem sempre será suficiente para resolver o problema, outras estratégias são descritas a seguir.

2.3.3.4 Algoritmo EM

Marschner e Gillett (2011) propuseram um método computacional especificamente desenvolvido para o modelo log-binomial que apresenta taxas de convergência maiores que o método de escore de Fisher. O método utiliza um algoritmo EM (*Expectation Maximization*) aplicado a um modelo de variáveis latentes. Os autores apresentam a metodologia com covariáveis discretas estritamente positivas e depois estendem para variáveis contínuas que são tratadas como discretas, com a justificativa de que qualquer medida das covariáveis precisa ser representada por um número finito de casas decimais e portanto pode ser reescalonada para um número inteiro.

- **Covariáveis categóricas**

Sejam n respostas independentes $Y_i = \{Y_1, ..., Y_n\}$, $Y_i \sim Binomial(n_i, p_i)$, cada uma com k covariáveis e um termo constante, o vetor de covariáveis do registro i é dado por $\mathbf{x}_i = (x_{i0}, ..., x_{ik})$, em que $x_{i0} = 1$. Considere que as covariáveis x_j, $j = 0, ..., k$, sejam categóricas, apresentem k_j níveis e $\{x_j^{(1)}, x_j^{(2)}, ..., x_j^{(k_j)}\}$ sejam os valores possíveis para x_j. O espaço de covariáveis $\mathcal{X}$ é o produto cartesiano dos j conjuntos de valores possíveis das covariáveis $x_j = \{x_j^{(1)}, x_j^{(2)}, ..., x_j^{(k_j)}\}$.

O modelo log-binomial é especificado pela função de ligação logarítmica

$$p_i = P(\mathbf{x}_i; \boldsymbol{\beta}) = \exp\left[\sum_{j=0}^{k} \beta_j(x_{ij})\right] = \prod_{j=0}^{k} \exp\left[\beta_j(x_{ij})\right],$$

o valor de β_j depende do nível da covariável x_j apresentado pelo registro i,

$$\beta_j(x_{ij}) = \begin{cases} \beta_j^{(1)}, \text{ se } x_{ij} = x_j^{(1)} \\ \beta_j^{(2)}, \text{ se } x_{ij} = x_j^{(2)} \\ ... \\ \beta_j^{(k_j)}, \text{ se } x_{ij} = x_j^{(k_j)} \end{cases}. \tag{2.3}$$

Definimos novos parâmetros θ_j, tais que,

$$p_i = P(\mathbf{x}_i; \boldsymbol{\theta}) = \prod_{j=0}^{k} \left[\theta_j(x_{ij})\right]. \tag{2.4}$$

De (2.3) e (2.4) temos que $\theta_j(x_{ij}) = \exp\left[\beta_j(x_{ij})\right]$. Escolhemos um nível de referência $x_j^{(r_j)}$ para cada covariável e estipulamos $\theta_j\left(x_j^{(r_j)}\right) = 1$, que equivale a $\beta_j\left(x_j^{(r_j)}\right) = 0$. Assim, $\theta_j\left(x_j^{(1)}\right)$ pode ser interpretado como o risco relativo do nível $x_j^{(1)}$ em relação ao nível de referência $x_j^{(r_j)}$. A escolha do vetor de referência $\mathbf{r} = \left(x_0^{(r_0)}, x_1^{(r_1)}, ..., x_k^{(r_k)}\right)$ é arbitrária e não afeta o resultado final. O espaço paramétrico de $\boldsymbol{\theta} = (\theta_0, \theta_1, ..., \theta_k)$ é

$$\Theta = \{\boldsymbol{\theta} : p_i \in (0, 1), \forall \mathbf{x} \in \mathcal{X}\},$$

e a log-verossimilhaça é dada por

$$\ell(\boldsymbol{\theta}) = \sum_{i=1}^{n} Y_i \log(P(\mathbf{x}_i;\boldsymbol{\theta})) + \sum_{i=1}^{n}(n_i - Y_i)\log(1 - P(\mathbf{x}_i;\boldsymbol{\theta})).$$

Como Y_i tem distribuição Binomial, pode ser expressa como uma soma de ensaios de Bernoulli,

$$Y_i = \prod_{m=1}^{n_i} Y_i^{[m]}.$$

Marschner e Gillett (2011) propõem representar cada resposta binária observada $Y_i^{[m]}$ como um produto de variáveis binárias latentes não observadas $Z_{ij}^{[m]}$,

$$Y_i^{[m]} = \prod_{j=0}^{k} Z_{ij}^{[m]}. \tag{2.5}$$

Como $Z_{ij}^{[m]}$ são variáveis binárias, $Y_i^{[m]}$ será igual a 1 somente se todos $Z_{ij}^{[m]}$ s'ão iguais a 1,

$$Y_i^{[m]} = 1 \iff Z_{ij}^{[m]} = 1,\ \forall j \in \{0, 1, ..., k\},$$

assim,

$$p_i = P(Y_i^{[m]} = 1) = P(Z_{i0}^{[m]} = 1, Z_{i1}^{[m]} = 1, ..., Z_{ij}^{[m]} = 1),$$

usando (2.4),

$$P(Z_{ij}^{[m]} = 1) = \theta_j(x_{ij}) \text{ e } P(Z_{ij}^{[m]} = z) = [\theta_j(x_{ij})]^z[1 - \theta_j(x_{ij})]^{1-z},\ z \in \{0, 1\}.$$

O modelo de variáveis latentes é definido em um espaço paramétrico restrito, no qual o risco $P(\mathbf{x};\boldsymbol{\theta})$ associado a qualquer $\mathbf{x} \in \mathcal{X}$ não excede o risco $P(\mathbf{r};\boldsymbol{\theta})$ associado ao vetor de valores de referência $\mathbf{r}$.

$$\Theta(\mathbf{r}) = \{\boldsymbol{\theta} : 0 \leq P(\mathbf{x};\boldsymbol{\theta}) \leq P(\mathbf{r};\boldsymbol{\theta}),\ \forall \mathbf{x} \in \mathcal{X}\} \subset \Theta.$$

As variáveis latentes apresentam distribuição Binomial,

$$Z_{ij}^{[m]} \sim Binomial(n_i, \theta_j(x_{ij})),$$

e a função de log verossimilhança para $\boldsymbol{\theta} \in \Theta(\mathbf{r})$ é dada por

$$\ell(\boldsymbol{\theta}|Z_{ij}) = \sum_{i=1}^{n}\sum_{j=0}^{k} Z_{ij}\log[\theta_j(x_{ij})] + (n_i - Z_{ij})\log[1 - \theta_j(x_{ij})].$$

Podemos maximizar $\ell(\boldsymbol{\theta})$ para $\Theta(\mathbf{r})$ utilizando um algoritmo EM.

Passo E

$$\mathcal{Q}\left(\boldsymbol{\theta}|\hat{\boldsymbol{\theta}}^{(t)}\right) = E\left[\ell(\boldsymbol{\theta}|Z_{ij})\middle|Y_i;\hat{\boldsymbol{\theta}}^{(t)}\right] = \ell\left(\boldsymbol{\theta}|\hat{Z}_{ij}^{(t)}\right),$$

$$\hat{Z}_{ij}^{(t)} = E\left[Z_{ij}|Y_i;\hat{\boldsymbol{\theta}}^{(t)}\right] = Y_i + (m_i - Y_i)e_{ij}\hat{\boldsymbol{\theta}}^{(t)},$$

$$e_{ij}(\boldsymbol{\theta}) = E[Z_{ij}^{[m]}|Y_i^{[m]} = 0] = \frac{\theta_j(x_{ij}) - P(\mathbf{x}_i;\boldsymbol{\theta})}{1 - P(\mathbf{x}_i;\boldsymbol{\theta})}.$$

Passo M

$$\hat{\theta}_j^{(t+1)}(x) = \frac{\sum_{i\in I_{jx}} \hat{Z}_{ij}^{(t)}}{\sum_{i\in I_{jx}} N_i},$$

em que, $I_{jx} = i : x_{ij} = x$ é o conjunto de observações que a covariável $j = x$.

Uma vez que

$$\Theta = \bigcup_{\mathbf{r}\in\mathcal{X}} \Theta(\mathbf{r}),$$

se realizarmos as maximizações para todos os escolhas possíveis de $\mathbf{r}$, ao menos um dos resultados será o máximo no espaço paramétrico completo Θ.

- **Covariáveis contínuas**

Seja $Y_i \sim Binomial(n_i, p_i)$, suponha um modelo log-binomial com k covariáveis e um termo constante, o vetor de covariáveis é dado por $\mathbf{x}_i = (x_{i0}, ..., x_{ik})$, com $x_{i0} = 1$. Seja $x_j^{(0)} = min(x_{ij})$ e $x_j^{(1)} = max(x_{ij})$, o espaço das covariáveis é estipulado como o produto cartesiano dos intervalos entre o mínimo das covariáveis, $\mathcal{X} = \prod_{j=0}^{k}(x_j^{(0)}, x_j^{(1)})$. Definimos novos parâmetros θ_j, tais que,

$$p_i = P(\mathbf{x}_i;\boldsymbol{\theta}) = e^{\left(\sum_{j=0}^{k}\beta_j x_{ij}\right)} = \prod_{j=0}^{k} e^{(\beta_j x_{ij})} = \prod_{j=0}^{k}\left(e^{\beta_j}\right)^{x_{ij}} = \prod_{j=0}^{k}(\theta_j)^{x_{ij}}, \tag{2.6}$$

portanto, $\theta_j = \exp(\beta_j)$ pode ser interpretado como o risco relativo associado ao acréscimo de uma unidade na covariável j. O espaço paramétrico de $\boldsymbol{\theta} = (\theta_0, ..., \theta_k)$ é dado por $\Theta = \{\boldsymbol{\theta} : p_i \in (0, 1), \forall \mathbf{x} \in \mathcal{X}\}$.

Para definir as variáveis latentes, que serão utilizadas no algoritmo EM, é utilizada uma reparametrização do modelo. Para cada covariável, é escolhido um valor de referência r_j, que pode ser seu valor mínimo $x_j^{(0)}$ ou máximo $x_j^{(1)}$ observados. Assim, existem 2^k escolhas possíveis para $\mathbf{r} = (r_0, ..., r_k)$, esse espaço de valores é denotado por $\mathcal{P}$. Diferentes escolhas de $\mathbf{r}$ implicam em diferentes reparametrizações.

Para $j = 1, ..., k$, define-se:

$$u_{i0} = 1,\ u_{ij} = (-1)^{s_j}(x_{ij} - r_j),\ \alpha_0 = \beta_0 + \sum_{j=1}^{k}\beta_j r_j,\ \beta_j = (-1)^{s_j}\alpha_j,$$

em que, $s_j = 0$ se $r_j = x_j^{(0)}$ e $s_j = 1$ se $r_j = x_j^{(1)}$.

Definindo novos parâmetros $\lambda_j = \exp(\alpha_j)$, como em (2.2), temos:

$$p_i = P(\mathbf{x}_i;\boldsymbol{\theta}) = P(\mathbf{u}_i;\boldsymbol{\lambda}) = e^{\left(\sum_{j=0}^{k}\alpha_j u_{ij}\right)} = \prod_{j=0}^{k}(\lambda_j)^{u_{ij}}.$$

A vantagem de usar u_{ij} ao invés de x_{ij} é que $u_{ij} \geq 0$, o que permite ver λ_j como probabilidades no modelo latente.

Assumindo, sem perda de generalidade, que u_{ij} é um inteiro, $Y_i^{[m]}$ podem ser vistos como um produto de variáveis binárias latentes independentes, como em (2.5),

$$Y_i^{[m]} = \prod_{j=0}^{k} \prod_{u=1}^{u_{ij}} Z_{iju}^{[m]}, \text{ em que } P\left(Z_{iju}^{[m]} = z\right) = \lambda_j^z (1-\lambda_j)^{1-z}, z \in \{0,1\}.$$

As variáveis latentes apresentam distribuição Binomial,

$$Z_{ij} = \sum_{k=1}^{n_i} \sum_{u=1}^{u_{ij}} Z_{iju}^{[m]} \sim Binomial(n_i u_{ij}, \lambda_j).$$

Podemos maximizar a função de máxima verossimilhança, para $\Theta(\mathbf{r})$, utilizando um algoritmo EM.

Passo E

$$\hat{Z}_{ij}^{(t)} = Y_i + (n_i - Y_i)\frac{\hat{\lambda}_j^{(t)} - P(\mathbf{u}_i; \hat{\boldsymbol{\lambda}})}{1 - P(\mathbf{u}_i; \hat{\boldsymbol{\lambda}})}.$$

Passo M

$$\hat{\lambda}_j^{(t+1)} = \frac{\sum_{i=1}^{n} u_{ij} \hat{Z}_{ij}^{(t)}}{\sum_{i=1}^{n} u_{ij}}.$$

Semelhante ao caso das variáveis categóricas, é preciso realizar a maximização para cada uma das 2^k possíveis escolhas de $\mathbf{r} \in \mathcal{P}$.

A implementação do algoritmo EM para ajuste de modelo log-binomial está disponível no pacote `logbin` (DONOGHOE; MARSCHNER, 2018). Os autores oferecem duas opções de solução. Quando é utilizada a opção `method="cem"`, o algoritmo EM é executado para todas as combinações de valores possíveis para o vetor $\mathbf{r}$, conforme descrito acima. Entretanto o processamento pode se tornar muito prolongado quando existem muitas covariáveis. Para evitar esse inconveniente, Donoghoe e Marschner (2016) propõe um método que requer a execução do algoritmo EM somente uma vez, para isso, são inseridos parâmetros extras no modelo, um para cada covariável, esse método é dado pela opção `method="em"`.

2.3.3.5 Barreira Adaptativa

Os métodos de otimização visam obter os valores em que uma função $f(\mathbf{x})$ atinge seu mínimo. Existem métodos de otimização com restrição e sem restrição. As restrições podem ser de desigualdade $h(\mathbf{x}) \leq 0$ ou de igualdade $g(\mathbf{x}) = 0$.

Um problema de otimização genérico pode ser expresso da seguinte forma: seja $\mathbf{x} \in \mathbb{R}^d$, $f : \mathbb{R}^d \to \mathbb{R}$, $h_i : \mathbb{R}^d \to \mathbb{R}$ e $g_j : \mathbb{R}^d \to \mathbb{R}$, encontrar

$$\min_{\mathbf{x}} \{f(\mathbf{x})\}$$
$$\text{sujeito a } h_i(\mathbf{x}) \leq 0 \text{ e } g_j(\mathbf{x}) = 0.$$

em que $f(\mathbf{x})$ é a função objetivo, $h_i(\mathbf{x})$, $i = 1, 2, ..., r$ são as restrições de desigualdade e $g_j(\mathbf{x})$, $j = 1, 2, ..., l$ são as restrições de igualdade.

O método de barreira trata o problema de otimização com restrições convertendo-o em um problemas mais amplo de otimização sem restrição, para isso, inclui as restrições como um termo de penalização. Assim, dado um problema de otimização com r restrições de desigualdade,

$$\begin{aligned} &\min_{\mathbf{x}} \{f(\mathbf{x})\} \\ &\text{sujeito a } h_i(\mathbf{x}) \leq 0, \end{aligned} \tag{2.7}$$

em que $i = 1, 2, ..., r$, podemos reescreve-lo como um problema de otimização sem restrição:

$$\min_{\mathbf{x}} \{f(\mathbf{x}) + penalização\},$$

isto é,

$$\min_{\mathbf{x}} \left\{ f(\mathbf{x}) + \delta \sum_{i=1}^{r} \psi(h_i(\mathbf{x})) \right\},$$

em que δ é chamado de parâmetro de barreira, $\sum_{i=1}^{r} \psi(h_i(\mathbf{x}))$ é a função barreira e $f(\mathbf{x}) + \delta \sum_{i=1}^{r} \psi(h_i(\mathbf{x}))$ é a função auxiliar. Quando $\delta \to 0$ e $\sum_{i=1}^{r} \psi(h_i(\mathbf{x})) \to \infty$ a solução do problema de barreira converge para a solução ótima do problema (2.7).

Existem diferentes formas para a penalização, a mais comum é a logarítmica que facilita as derivações, assim, no método de barreira logarítmica,

$$\sum_{i=1}^{r} \psi(h_i(\mathbf{x})) = -\sum_{i=1}^{r} \log(-h_i(\mathbf{x})).$$

No processo iterativo o parâmetro de barreira δ deve convergir para zero. Uma forma simples de atualizar δ a cada iteração é multiplicá-lo por um valor $\gamma \in (0, 1)$, entretanto, se δ decresce muito lentamente, um número muito grande de iterações pode ser necessário e se δ decresce muito rapidamente, pode não ser possível encontrar uma solução sem violar as restrições. Assim, a melhor forma de atualização de δ depende de $f(\mathbf{x})$ e da magnitude de $\mathbf{x}$. Existem diferentes formas de atualização de δ, as que levam em conta o progresso do algoritmo são chamados adaptativos.

Lange (1994) propõe um método de barreira logarítmica adaptativa em que o parâmetro de barreira δ é mantido constante durante o processo, diferentemente dos métodos clássicos de barreira, em que δ decresce gradualmente para zero. Para isso, realiza uma alteração na função barreira que garante, para $f(\mathbf{x})$ convexa, que a função decresça a

cada nova iteração, isto é, $f\left(\mathbf{x}^{(t+1)}\right) \leq f\left(\mathbf{x}^{(t)}\right)$, o que traz maior estabilidade ao algoritmo. Seja o problema de otimização de uma função convexa,

$$\min_{\mathbf{x}} \{f(\mathbf{x})\}$$
$$\text{sujeito a } x_i \leq 0,\ i = 1, 2, ..., r,$$

nos métodos clássicos de barreira logarítmica, a função auxiliar é

$$f(\mathbf{x}) - \delta^{(t)} \sum_{i=1}^{r} \log(-x_i),$$

pelo método proposto, a função auxiliar passa a ser

$$f(\mathbf{x}) - \delta \sum_{i=1}^{r} x_i^{(t)} \log(-x_i),$$

portanto, o valor da próxima iteração $\mathbf{x}^{(t+1)}$ é produzido utilizando os valores de $\mathbf{x}^{(t)}$ na iteração atual.

De forma mais geral, se tivermos como restrição de desigualdade funções de $\mathbf{x}$,

$$\min_{\mathbf{x}} \{f(\mathbf{x})\}$$
$$\text{sujeito a } h_i(\mathbf{x}) \leq 0,\ i = 1, 2, ..., r,$$

então, a função auxiliar passar a ser,

$$f(\mathbf{x}) + \delta \sum_{i=1}^{r} \left\{ h_i\left(\mathbf{x}^{(t)}\right) \log\left(-h_i(\mathbf{x})\right) + h_i(\mathbf{x}) \right\}$$

Obter o EMV para um modelo log-binomial pode ser visto como um problema de otimização com restrições de desigualdade, em que $f(\mathbf{x}) = -\ell(\boldsymbol{\beta})$, $d = k + 1$, $h_i(\mathbf{x}) = \beta_0 + \sum_{j=1}^{k} \beta_j x_{ij} = \mathbf{x}_i'\boldsymbol{\beta} \leq 0$, $i = 1, 2, ..., n$. Ou seja,

$$\min_{\boldsymbol{\beta}} \{-\ell(\boldsymbol{\beta})\}$$
$$\text{sujeito a } \mathbf{x}_i'\boldsymbol{\beta} \leq 0,\ i = 1, 2, ..., n,$$

que pode ser resolvido pelo problema de otimização sem restrições

$$\min_{\boldsymbol{\beta}} \left\{ -\ell(\boldsymbol{\beta}) + \delta \sum_{i=1}^{n} \left[\mathbf{x}_i'\boldsymbol{\beta}^{(t)} \log(-\mathbf{x}_i'\boldsymbol{\beta}) + \mathbf{x}_i'\boldsymbol{\beta} \right] \right\}.$$

Esse método para ajuste do modelo log-binomial foi implementado por Andrade e Andrade (2017) e está disponível no pacote `lbreg` (ANDRADE, 2018) do `R`.

2.3.4 Repulsão

Marschner (2015) esclarece que, mesmo que a estimativa de máxima verossimilhança seja um ponto estacionário dentro do espaço paramétrico, pode haver um ponto de repulsão

que fará com que o algoritmo de escoragem de Fisher seja incapaz de convergir e apresente oscilação periódica ou aperiódica .

Seja um processo iterativo

$$\boldsymbol{\beta}^{(t+1)} = F(\boldsymbol{\beta}^{(t)}),$$

no qual a função de iteração F atualiza a estimativa atual $\boldsymbol{\beta}^{(t)} \in \mathbb{R}^p$ para o novo valor $\boldsymbol{\beta}^{(t+1)}$, a matriz de convergência desse processo é dada pelos elementos

$$G_{ij}(\boldsymbol{\beta}) = \frac{\partial F_i(\boldsymbol{\beta})}{\partial \beta_j}.$$

Um ponto estacionário $\hat{\boldsymbol{\beta}}$ é um ponto de atração do processo iterativo, se o raio espectral $\rho(G(\hat{\boldsymbol{\beta}}))$ é menor que um. Por outro lado, se $\rho(G(\hat{\boldsymbol{\beta}})) > 1$, então $\hat{\boldsymbol{\beta}}$ é um ponto de repulsão e o algoritmo só conseguirá convergir se já for inicializado com o valor $\hat{\boldsymbol{\beta}}$.

Derivando, em relação a estimativa de máxima verossimilhança $\hat{\boldsymbol{\beta}}$, a função F do processo de escoragem de Fisher, caracterizada em (2.2), temos a matriz de convergência

$$G(\boldsymbol{\beta}) = \mathcal{I}_e(\boldsymbol{\beta})^{-1}\{\mathcal{I}_e(\boldsymbol{\beta}) - \mathcal{I}_o(\boldsymbol{\beta})\}, \tag{2.8}$$

em que $\mathcal{I}_e$ é a matriz de informação esperada e $\mathcal{I}_o$ é a matriz de informação observada.

Em qualquer MLG canônico, a matriz de informação esperada coincide com a matriz de informação observada, portanto, $\rho(G(\hat{\boldsymbol{\beta}})) = 0$ e a estimativa de máxima verossimilhança nunca será um ponto de repulsão. No modelo logístico,

$$\mathcal{I}_o = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{(1+\exp(\mathbf{x}_i'\boldsymbol{\beta}))^2},$$

como $\mathcal{I}_o$ não depende de y_i, então $\mathcal{I}_e = \mathcal{I}_o$. Por outro lado, no modelo log-binomial

$$\mathcal{I}_o(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2} = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})(1-y_i)}{(1-\exp(\mathbf{x}_i'\boldsymbol{\beta}))^2} \text{ e}$$

$$\mathcal{I}_e(\boldsymbol{\beta}) = E\left(-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\right) = \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i' \frac{\exp(\mathbf{x}_i'\boldsymbol{\beta})}{(1-\exp(\mathbf{x}_i'\boldsymbol{\beta}))},$$

o que pode ocasionar o problema repulsão.

2.3.4.1 Fracionamento do Passo e Log-verossimilhança Crescente

Para tratar a questão da fronteira e da repulsão ao mesmo tempo, Marschner (2015) acrescenta uma condição de parada da redução dos passos no método fracionamento do passo, dessa forma,

$$\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)} + 0.5^h \mathcal{I}_e(\boldsymbol{\beta}^{(t)})^{-1}\mathbf{u}^{(t)},\ h = 0, 1, 2, ...$$

em que, $h = \min\{h \in \mathbb{N} : \boldsymbol{\beta}_{c+1} \in \Theta \text{ e } \ell(\boldsymbol{\beta}^{(t+1)}) \geq \ell(\boldsymbol{\beta}^{(t)})\}$. Esse método está implementado no pacote `glm2` (MARSCHNER, 2011) do `R`.

2.3.5 EMV na Fronteira

Quando o EMV está no interior do espaço paramétrico, sua distribuição é assintoticamente normal e podemos construir intervalos de confiança para $\hat{\boldsymbol{\beta}}$ utilizando o fato de que

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \mathcal{I}_e^{-1}(\boldsymbol{\beta}_0)), \tag{2.9}$$

em que $\boldsymbol{\beta}_0$ é o verdadeiro parâmetro e $\mathcal{I}_e^{-1}(\boldsymbol{\beta}_0)$ é a matriz de informação esperada.

Quando o EMV está na fronteira para alguma observação, isto é, quando $\mathbf{x}_i'\boldsymbol{\beta} = 0$ e, portanto, $\hat{p}_i = 1$, a relação (2.9) não é válida. Andrade e Andrade (2017) propõe uma forma mais geral que permite construir intervalos de confiança para $\hat{\boldsymbol{\beta}}$ válidos inclusive nesses casos, a relação (2.9) passa a ser

$$\sqrt{n}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0) \xrightarrow{D} N(\mathbf{0}, \boldsymbol{\Sigma}),$$

em que, $\boldsymbol{\Sigma} = \mathcal{I}_e^{-1} - \mathcal{I}_e^{-1}\mathbf{A}'(\mathbf{A}\mathcal{I}_e^{-1}\mathbf{A})^{-1}\mathbf{A}\mathcal{I}_e^{-1}$ e $\mathbf{A}\hat{\boldsymbol{\beta}} = \mathbf{0}$. Esse solução está implementada na função `lbreg` do `R`.

2.3.6 Exemplos

Nos exemplos a seguir, foram utilizados os conjuntos de dados do pacote `lbreg`.

2.3.6.1 Conjunto de dados `Birth`

O conjunto de dados `Birth` refere-se ao peso de 900 bebês ao nascer. A variável resposta `lowbw` é dada de forma binária, 1 em caso de nascimento com baixo peso e 0 caso contrário. As três variáveis explicativas são categóricas e referentes aos hábitos e condições da mãe. A variável `alc` expressa a frequência de consumo de álcool, possui três classes, 1=baixa, 2=moderada ou 3=alta; a variável `smo` possui duas classes, 1=fumante ou 2=não fumante e a variável `soc` indica a classe social à qual a mãe pertence e possui três classes, 1=baixa, 2=média ou 3=alta. O percentual observado de nascimentos com baixo peso é de 10,89%.

Nesse conjunto de dados, não ocorreu nenhum problema de estimação descritos nas subseções 2.3.2, 2.3.3 e 2.3.4. Todos os pacotes testados: `glm`, `glm2`, `logbin` com `method="cem"`, `logbin` com `method="em"` e `lbreg`, foram capazes de ajustar o modelo log-binomial e produziram estimativas semelhantes, divergindo somente a partir da quarta casa decimal. Como os resultados são semelhantes, somente o resultado do pacote `lbreg` é apresentado na Tabela 5.

O algoritmo da função `logbin` com `method="cem"` não convergiu dentro das 10.000 iterações máximas, valor padrão da função. Aumentar a opção `maxit` foi suficiente para alcançar a convergência que ocorreu na iteração 12.441.

A categoria 2=moderada da variável `alc`, frequência de consumo de álcool, não apresentou risco maior que a categoria 1=baixa, a um nível de significa de 0,10. Da mesma forma, as categorias 2=média e 3=alta da variável `soc`, classe social, não apresentaram risco maior que a categoria 1=baixa, caso o interesse seja de predição, essas categorias podem ser retiradas do modelo.

Tabela 5 – Estimativas do pacote `lbreg` para o conjunto de dados `Birth`

	Estimativa	Erro Padrão	estatística z	p-valor	RR	IC RR Inf	IC RR Sup
Intecepto	-2,764	0,203	-13,609	< 2.2e-16			
alc2	0,175	0,274	0,638	0,524	1,191	0,696	2,039
alc3	0,680	0,215	3,160	0,002	1,974	1,295	3,010
smo2	0,500	0,201	2,484	0,013	1,648	1,111	2,445
soc2	0,293	0,233	1,255	0,209	1,340	0,849	2,116
soc3	0,300	0,243	1,231	0,218	1,349	0,837	2,174

2.3.6.2 Conjunto de dados `Heart`

O conjunto de dados `Heart` é referente a 16.949 pacientes que tiveram ataque cardíaco. A variável resposta `Heart` é dada de forma binária, 1 em caso de morte dentro de 30 dias após o ataque cardíaco e 0 caso contrário. As quatro variáveis explicativas são categóricas. A variável `age` apresenta três classes, 1=menos de 65 anos, 2=de 65 a 75 anos ou 3=acima de 75 anos; a variável `severity` apresenta três classes em ordem crescente de gravidade; a variável `region` apresenta três classes, 1=países ocidentais, 2=américa latina ou 3=leste europeu e a variável `onset` refere-se à demora até o atendimento do paciente, apresenta três classes, 1=menos de 2 horas, 2=de 2 a 4 horas ou 3=acima de 4 horas. O percentual observado de mortes é de 6,17%.

A função `glm` não foi capaz de inicializar com valores dentro do espaço paramétrico. Conforme descrito na seção 2.3.2, uma opção para valores iniciais é um vetor com valor negativo para β_0 e zero para os demais coeficientes, outra opção é utilizar uma função dos valores de um modelo de poisson. Em ambas as situações o algoritmo não foi capaz de convergir, mesmo aumentando o número máximo de iterações para 100.000. Por meio da opção `trace=1` é possível verificar que a *deviance* fica oscilando entre os valores 6955,164, 6941,807, 6944,300, 6943,054, 6957,479, 6941,698, 6943,137 e 6942,831, o que indica que se trata de um problema de repulsão, questão tratada na seção 2.3.4. Dessa forma, aumentar o argumento `maxit` não resolve o problema de convergência, pois o algoritmo acaba entrando em um ciclo de repetições e não é capaz de sair. É possível notar que mesmo partindo de valores iniciais distintos, depois de algumas iterações, entra no mesmo ciclo de valores. Observando que a variação das *deviances* é da ordem da quarta casa decimal, podemos alterar o argumento de tolerância de `epsilon=1e-8` para `epsilon=1e-4`, assim o algoritmo consegue declarar convergência, entretanto, essa não é uma solução para o problema de

repulsão, mas uma forma de reduzir o rigor sobre os resultados. Existem recursos mais eficientes como a utilização dos pacotes `glm2`, `logbin` e `lbreg`.

A função `glm2` também não foi capaz de inicializar dentro do espaço paramétrico, pois utiliza o mesmo procedimento do `glm`. Foi preciso fornecer valores iniciais, como as duas opções dadas na seção 2.3.2. O algoritmo foi capaz de convergir dentro do número máximo de iterações padrão da função, `maxit=25`, a convergência tendo ocorrido na 14ª iteração.

No caso das funções `logbin` e `lbreg` não foi necessário fornecer valores iniciais e os algoritmos convergiram usando os valores padrão para os argumentos da função.

As estimativas dos coeficientes das funções `glm2`, `logbin` com `method="cem"`, `logbin` com `method="em"` e `lbreg` foram semelhantes, divergindo somente a partir da quarta casa decimal. Na Tabela 6 são apresentados os valores obtidos com a função `lbreg`.

Assim como no exemplo anterior, caso o interesse seja de predição, algumas variáveis podem ser retiradas do modelo.

Tabela 6 – Estimativas do pacote `lbreg` para o conjunto de dados `Heart`

	Estimativa	Erro Padrão	estatística z	p-valor	RR	IC RR Inf	IC RR Sup
Intercepto	-4,028	0,089	-45,333	< 2.2e-16			
age2	1,104	0,089	12,346	< 2.2e-16	3,017	2,532	3,596
age3	1,927	0,093	20,742	< 2.2e-16	6,871	5,727	8,244
severity2	0,703	0,070	10,056	< 2.2e-16	2,020	1,762	2,317
severity3	1,376	0,088	15,646	< 2.2e-16	3,961	3,334	4,706
region2	0,076	0,181	0,420	0,675	1,079	0,757	1,537
region3	0,483	0,087	5,528	3,2e-8	1,620	1,365	1,923
onset2	0,059	0,069	0,854	0,393	1,061	0,926	1,215
onset3	0,172	0,079	2,174	0,030	1,188	1,017	1,387

2.3.6.3 Conjunto de dados PCS

O conjunto de dados `PCS` é um estudo de câncer de próstata e compreende 360 pacientes. A variável resposta `tumor` é dada de forma binária, 1 em caso de penetração do tumor na capsula da próstata e 0 caso contrário. Apresenta sete variáveis explicativas das quais três são categóricas e quatro intervalares. A variável `race` possui duas classes, 1=branco ou 2=negro; a variável `dpros` refere-se ao resultado do exame retal digital e possui 4 classes; a variável `dcaps` refere-se à detecção de comprometimento da capsula no exame retal, possui dois níveis, 1=não ou 2=sim; a variável `age` refere-se à idade do paciente; a variável `psa` refere-se à concentração do antígeno psa no sangue, em ng/ml; a variável `vol` refere-se ao volume do tumor, medido por ultrassonografia, em cm^3 e a variável `gleason` refere-se a uma pontuação dada por meio de biopsia, possui valores de 0 a 10, sendo 10 o pior prognóstico. O percentual observado de penetração da capsula é de 40,26%.

Assim como no exemplo anterior, as funções `glm` e `glm2` não foram capazes de inicializar dentro do espaço paramétrico, fornecendo os valores iniciais descritos na seção 2.3.2, o algoritmo convergiu, sendo necessário somente aumentar o valor do argumento `maxit`. Entretanto, observamos que os valores estimados diferem quando usamos valores iniciais diferentes. Apesar disso, não há um aviso de que os resultados podem estar imprecisos, as únicas mensagens de alerta referem-se ao fato de ser necessário realizar o fracionamento do passo devido à saída do espaço paramétrico.

Tabela 7 – Estimativas do pacote `glm` para o conjunto de dados `PCS` com valores de inicialização distintos

	glm - $0^{(1)}$		glm - $poi^{(2)}$	
	Estimativa	Erro Padrão	Estimativa	Erro Padrão
Intercepto	-2,960	0,500	-3,058	0,511
age	-0,005	0,007	-0,005	0,007
race2	-0,002	0,074	-0,007	0,075
dpros2	0,501	0,210	0,520	0,216
dpros3	0,694	0,204	0,722	0,210
dpros4	0,563	0,215	0,585	0,221
dcaps2	0,079	0,090	0,075	0,091
psa	0,002	0,001	0,002	0,001
vol	-0,007	0,002	-0,007	0,002
gleason	0,293	0,037	0,304	0,038
Residual Deviance	406.73		406.86	
Iterações	68		49	

Nota: (1) inicializado com β_0 negativo e demais coeficientes iguais a zero; (2) inicializado uma função dos valores do modelo de poisson.

A função `glm2` trouxe o mesmo resultado com o mesmo número de iterações para alcançar a convergência, então a modificação no Step Having, descrita na seção 2.3.4.1 não foi necessária, as *deviances* a cada iteração foram sempre decrescentes.

A função `logbin` retorna estimativas para os coeficientes, mas não calcula os valores da variância, assim, não traz o erro padrão, a estatística z e o p-valor, informando que a estimava de máxima verossimilhança está na fronteira e que não é possível utilizar a matriz de covariância assintótica.

Conforme descrito na seção 2.3.5, a função `lbreg` é capaz de trazer uma solução mesmo nos casos em que o EMV está na fronteira, na Tabela 8 são apresentados os valores obtidos com a função `lbreg`. Assim como nos exemplos anteriores, caso o interesse seja de predição, algumas variáveis podem ser retiradas do modelo.

2.3.6.4 Roteiro

Ao consultar artigos recentes das áreas da saúde e ciências sociais percebemos que os métodos alternativos de estimação ainda são muito utilizados. Entretanto, avanços

Tabela 8 – Estimativas do pacote `lbreg` para o conjunto de dados `PCS`

	Estimativa	Erro Padrão	estatística z	p-valor	RR	IC RR Inf	IC RR Sup
Intercepto	-2,997	0,584	-5,131	2,9e-07			
age	-0,005	0,008	-0,654	0,513	0,995	0,979	1,011
race2	-0,007	0,073	-0,101	0,920	0,993	0,860	1,146
dpros2	0,528	0,230	2,299	0,022	1,695	1,081	2,658
dpros3	0,726	0,219	3,308	0,001	2,067	1,344	3,178
dpros4	0,591	0,237	2,497	0,013	1,807	1,136	2,874
dcaps2	0,069	0,097	0,707	0,480	1,071	0,885	1,296
psa	0,002	0,001	1,896	0,058	1,002	1,000	1,004
vol	-0,007	0,002	-3,434	0,001	0,993	0,989	0,997
gleason	0,297	0,040	7,352	2,0e-13	1,346	1,243	1,456

na capacidade de processamento dos computadores e desenvolvimento de novos métodos e algoritmos permitem obter estimativas robustas diretamente pelo método de máxima verossimilhança.

Como exemplo de métodos alternativos, podemos citar o método COPY (DEDDENS; PETERSEN; LEI, 2003) que utiliza dados aumentados, o método que aplica a regressão logística em dados modificados (SCHOUTEN et al., 1993), o método de quasi-verossimilhança que emprega a regressão de poisson com função de ligação log (ZOU, 2004) e o método que obtém as estimativas de risco relativo pelo redimensionamento das estimativas de razão de chances dadas pela regressão logística (SANTOS et al., 2008).

Para obter estimavas de máxima verossimilhança sem necessidade de recorrer a métodos indiretos, podem ser seguidas as seguintes etapas:

1. Caso o algoritmo de mínimos quadrados ponderados não seja capaz de inicializar, fornecer valores dentro do espaço paramétrico como descrito na seção 2.3.2.

2. Caso o algoritmo não alcance convergência, aumentar o número de iterações máximas que a função realiza, argumento `maxit` na função `glm`.

3. Se mesmo com o aumento de iterações máximas não ocorrer convergência, utilizar pacotes específicos para regressão log-binomial.

Na prática, havendo disponibilidade, é conveniente sempre utilizar os pacotes desenvolvidos para estimação de risco relativo, já que como visto no exemplo do conjunto de dados `PCS` quando a os valores estão próximos à fronteira, a função `glm` pode trazer resultados imprecisos sem dar nenhuma advertência.

3 Testes de Hipótese para Seleção de Modelos não Aninhados

A estatística do teste da razão de verossimilhança para comparação de modelos apresenta distribuição qui-quadrado somente nos casos em que os modelos são aninhados. Neste capítulo veremos duas opções de testes para modelos não aninhados, inclusive com função de ligação diferentes, o que, em princípio, nos permitirá comparar o ajuste do modelo log-binomial com o logístico.

Cabe ressaltar que existe distinção entre medidas de ajuste e testes de hipótese. Medidas como AIC e BIC auxiliam na decisão da escolha do modelo, uma vez que podemos escolher aquele que apresentar menor valor, entretanto não testam se de fato existe diferença significativa entre os modelos.

3.1 Teste de Cox

Sejam as hipóteses H_f, sob a qual os dados seguem o modelo F_θ e possuem densidade $f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})$ e H_g, sob a qual os dados seguem o modelo G_γ e possuem densidade $g(\mathbf{y}|\mathbf{x}, \boldsymbol{\gamma})$,

$$\mathrm{H}_f : f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}),$$
$$\mathrm{H}_g : g(\mathbf{y}|\mathbf{z}, \boldsymbol{\gamma}),$$

em que,

- $\mathbf{y}$: vetor de valores observados de dimensão $n \times 1$;
- $\mathbf{x}$: vetor de covariáveis do modelo F_θ de dimensão $n \times p$;
- $\boldsymbol{\theta}$: vetor de parâmetros desconhecidos do modelo F_θ de dimensão $p \times 1$;
- $\mathbf{z}$: vetor de covariáveis do modelo G_γ de dimensão $n \times q$;
- $\boldsymbol{\gamma}$: vetor de parâmetros desconhecidos do modelo G_γ de dimensão $q \times 1$.

A estatística do teste de Cox é dada por

$$\mathrm{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) = \bar{d} - \hat{\mathrm{E}}_f[\bar{d}], \tag{3.1}$$

em que

$$\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i,$$

$$d_i = \log \frac{f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{z}_i, \hat{\boldsymbol{\gamma}})},$$

$\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\gamma}}$ são os estimadores de máxima verossimilhança de $\boldsymbol{\theta}$ e $\boldsymbol{\gamma}$ sob H_f e H_g respectivamente e $\hat{\mathrm{E}}_f$ é a esperança sob H_f.

Se G_γ é aninhado a F_θ então, $\mathrm{E}_f[\bar{d}] = 0$, nesse caso, pode ser usado o teste de razão de verossimilhança habitual, no qual $2\bar{d}$ tem distribuição qui-quadrado.

Assintoticamente, a variância de $n^{1/2}\,\mathrm{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$, sob H_f, é dada por

$$v_f^2(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*) = V_f\left[\log \frac{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0)}{g(\mathbf{y}|\mathbf{z}, \boldsymbol{\gamma}_*)}\right] - \Psi'(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*)F(\boldsymbol{\theta}_0)^{-1}\Psi(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*),$$

em que $V_f(\cdot)$ é a variância sob $f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0)$, $\boldsymbol{\theta}_0$ é o verdadeiro valor de $\boldsymbol{\theta}$ sob H_f e $\boldsymbol{\gamma}_*$ é o valor de $\boldsymbol{\gamma}$ sob H_f,

$$\Psi(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*) = E_f\left[\frac{\partial \log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}}\left(\log \frac{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0)}{g(\mathbf{y}|\mathbf{z}, \boldsymbol{\gamma}_*)}\right)\right],$$

$$F(\boldsymbol{\theta}_0) = -E_f\left[\frac{\partial^2 \log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0 \partial \boldsymbol{\theta}_0'}\right] = E_f\left[\frac{\partial \log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0}\frac{\partial \log f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0)}{\partial \boldsymbol{\theta}_0'}\right].$$

A estatística do teste de Cox padronizada é dada por

$$\mathrm{T}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) = n^{1/2}\,\mathrm{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})/\hat{v}_f(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) \sim N(0, 1),$$

em que $\hat{v}_f(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$ é um estimador consistente de $v_f(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*)$.

3.1.1 Estimativa do denominador do teste

Pesaran e Pesaran (1993) demonstram que a variância $v_f(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*)$ pode ser consistentemente estimada por

$$\hat{v}_f(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) = \frac{1}{n}\mathbf{d}'\left\{\mathbf{I}_n - \mathbf{R}(\hat{\boldsymbol{\theta}})\left[\mathbf{R}'(\hat{\boldsymbol{\theta}})\mathbf{R}(\hat{\boldsymbol{\theta}})\right]^{-1}\mathbf{R}'(\hat{\boldsymbol{\theta}})\right\}\mathbf{d},$$

em que $\mathbf{d}' = (d_1, d_2, ..., d_N)$ e $\mathbf{R}(\hat{\boldsymbol{\theta}})$ é a matriz $n \times (p+1)$

$$\mathbf{R}(\hat{\boldsymbol{\theta}}) = \begin{bmatrix} 1 & \frac{\partial \log f(y_1|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial \log f(y_1|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_p} \\ 1 & \frac{\partial \log f(y_2|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial \log f(y_2|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_p} \\ \cdots & \cdots & \cdots & \\ 1 & \frac{\partial \log f(y_n|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_1} & \cdots & \frac{\partial \log f(y_n|\mathbf{x}, \boldsymbol{\theta})}{\partial \theta_p} \end{bmatrix}.$$

Assim, v_f^2 pode ser estimado ajustando uma regressão em que d_i é a variável resposta e as covariáveis são $\frac{\partial \log f(y_i|\mathbf{x},\boldsymbol{\theta})}{\partial \theta_1}$, $\frac{\partial \log f(y_i|\mathbf{x},\boldsymbol{\theta})}{\partial \theta_2}$, ... , $\frac{\partial \log f(y_i|\mathbf{x},\boldsymbol{\theta})}{\partial \theta_p}$ e um intercepto. O erro padrão estimado dessa regressão pode ser usado como estimativa de v_f^2.

3.1.2 Simulação do numerador do teste

3.1.2.1 Simulação de $\hat{\boldsymbol{\gamma}}_*$

Uma forma de estimar $\mathrm{E}_f[\bar{d}]$ é pela medida de proximidade de Kullback-Leibler (PESARAN; PESARAN, 1993) de H_f em relação a H_g

$$\hat{\mathrm{E}}_f[\bar{d}] = C(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*) = \int \log \frac{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0)}{g(\mathbf{y}|\mathbf{z}, \boldsymbol{\gamma}_*)} f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_0) dy. \tag{3.2}$$

Para estimar a medida de Kullback-Leibler, sob H_f,

$$\widehat{C(\boldsymbol{\theta}_0, \boldsymbol{\gamma}_*)} = C(\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\gamma}}_*),$$

precisamos de uma estimativa consistente de $\boldsymbol{\gamma}_*$ que pode ser obtida por simulação.

Gerar R simulações de $\mathbf{y}_i = (y_1, y_2, ..., y_n)$, assumindo o modelo e os coeficientes de F_θ, assim, cada simulação gera um vetor $\mathbf{y}_j = (y_{1j}, y_{2j}, ..., y_{nj})$, $j = 1, 2, ..., R$. Para cada simulação, ajustar um modelo G_γ, obtendo um vetor de estimativas $\hat{\boldsymbol{\gamma}}_j = (\hat{\gamma}_{1j}, \hat{\gamma}_{2j}, ..., \hat{\gamma}_{qj})$. Uma estimativa para $\boldsymbol{\gamma}_*$ pode ser obtida pela média dos valores de $\hat{\boldsymbol{\gamma}}_j$ das simulações:

$$\hat{\boldsymbol{\gamma}}_*^S = \frac{1}{R} \sum_{j=1}^{R} \hat{\boldsymbol{\gamma}}_j.$$

3.1.2.2 Simulação de $\mathrm{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}})$

Sejam $L_f(\mathbf{y}, \hat{\boldsymbol{\theta}}) = \frac{1}{n} \sum_{i=1}^{n} \log f(y_i|\mathbf{x}, \hat{\boldsymbol{\theta}})$ e $L_g(\mathbf{y}, \hat{\boldsymbol{\gamma}}) = \frac{1}{n} \sum_{i=1}^{n} \log f(y_i|\mathbf{z}, \hat{\boldsymbol{\gamma}})$, de (3.1) e (3.2) a estatística do teste de Cox pode ser reescrita como

$$\mathrm{N}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) = L_f(\mathbf{y}, \hat{\boldsymbol{\theta}}) - L_g(\mathbf{y}, \hat{\boldsymbol{\gamma}}) - C(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}_*). \tag{3.3}$$

O terceiro termo da equação (3.3) pode ser estimado por simulação utilizando o mesmo processo descrito na seção 3.1.2.1.

$$\widehat{C(\hat{\boldsymbol{\theta}}_0, \hat{\boldsymbol{\gamma}}_*)} = C(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}_*^S) = \frac{1}{R} \sum_{j=1}^{R} \left[L_f(\mathbf{y}_j, \hat{\boldsymbol{\theta}}) - L_g(\mathbf{y}_j, \hat{\boldsymbol{\gamma}}_*^S) \right]$$

No caso do teste do modelo log-binomial contra o modelo logístico e vice e versa, $C(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}_*^S)$ pode ser calculada analiticamente e não é necessária essa segunda simulação.

Se combinarmos o teste de Cox de H_f contra H_g e de H_g contra H_f, que chamaremos de T e $\tilde{\mathrm{T}}$ respectivamente e, se z_α é o valor crítico do teste, podemos ter quatro situações:

1. Se $|\,\mathrm{T}\,| < z_\alpha$ e $|\tilde{\mathrm{T}}| \geq z_\alpha$, então H_g é rejeitada em favor de H_f;
2. Se $|\,\mathrm{T}\,| \geq z_\alpha$ e $|\tilde{\mathrm{T}}| < z_\alpha$, então H_f é rejeitada em favor de H_g;
3. Se $|\,\mathrm{T}\,| \geq z_\alpha$ e $|\tilde{\mathrm{T}}| \geq z_\alpha$, então H_g e H_f são rejeitadas;
4. Se $|\,\mathrm{T}\,| < z_\alpha$ e $|\tilde{\mathrm{T}}| < z_\alpha$, então H_g e H_f não são rejeitadas.

3.2 Teste de Vuong

$$\mathrm{H}_0 : \mathrm{E}_0 \left[\log \frac{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})}{g(\mathbf{y}|\mathbf{x}, \boldsymbol{\gamma})} \right] = 0,$$

em que E_0 é a esperança da verdadeira densidade $h(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})$.

Diferentemente do teste de Cox, que pressupões que F_θ é o verdadeiro modelo, o teste de Vuong é mais genérico, permitindo que o verdadeiro modelo H_β seja desconhecido e diferente de F_θ e G_γ,

Equivalente a testar

$$\mathrm{H}_0 : \mathrm{E}_h \left[\log \frac{h(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})}{g(\mathbf{y}|\mathbf{x}, \boldsymbol{\gamma})} \right] - \mathrm{E}_h \left[\log \frac{h(\mathbf{y}|\mathbf{x}, \boldsymbol{\beta})}{f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta})} \right] = 0,$$

Razão de verossimilhança de F_θ contra G_γ.

$$\mathrm{LR}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) = \ell(\hat{\boldsymbol{\theta}}) - \ell(\hat{\boldsymbol{\gamma}}) = \sum_{i=1}^{n} \log \frac{f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\gamma}})}$$

O teste de Vuong apresenta três tipos de comparação de modelos

1. Modelos aninhados: $G_\gamma \subset F_\theta$;
2. Modelos estritamente não aninhados: $G_\gamma \cap F_\theta = \varnothing$;
3. Modelos sobrepostos: $G_\gamma \cap F_\theta \neq \varnothing$, $G_\gamma \subsetneqq F_\theta$ e $F_\theta \subsetneqq G_\gamma$.

Vamos nos ater ao segundo caso, pois nosso interesse é testar o modelo log-binomial contra o logístico.

Como f e g são supostamente incorretos, utilizaremos a notação $\boldsymbol{\theta}_*$ e $\boldsymbol{\gamma}_*$ como pseudo valores de $\boldsymbol{\theta}$ e $\boldsymbol{\gamma}$. Cameron e Trivedi (2005) demonstram que mesmo com as funções mal especificadas, os estimadores de máxima verossimilhança $\hat{\boldsymbol{\theta}}$ e $\hat{\boldsymbol{\gamma}}$ convergem para os pseudos parâmetros $\boldsymbol{\theta}_*$ e $\boldsymbol{\gamma}_*$ respectivamente.

Modelos estritamente não aninhados Se $f(\mathbf{y}|\mathbf{x}, \boldsymbol{\theta}_*) \neq g(\mathbf{y}|\mathbf{x}, \boldsymbol{\gamma}_*)$, sob H_0

$$n^{-1/2}\,\mathrm{LR}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) \xrightarrow{d} N(0, \omega_*^2),$$

em que

$$\omega_*^2 = \mathrm{Var}_0 \left[\log \frac{f(\mathbf{y}|\mathbf{x}, \hat{\boldsymbol{\theta}}_*)}{g(\mathbf{y}|\mathbf{x}, \hat{\boldsymbol{\gamma}}_*)} \right],$$

em que Var_0 é a variância da verdadeira distribuição h.

Um estimador consistente de ω_*^2 é

$$\hat{\omega}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \right)^2 - \left(\frac{1}{n} \sum_{i=1}^{n} \log \frac{f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \right)^2.$$

Assim, a estatística do teste é dada por

$$\mathrm{V} = n^{-1/2} \, \mathrm{LR}(\hat{\boldsymbol{\theta}}, \hat{\boldsymbol{\gamma}}) / \hat{\omega} \xrightarrow{d} N(0, 1).$$

Assintoticamente, a estimativa da variância se reduz ao primeiro termo

$$\tilde{\omega}^2 = \frac{1}{n} \sum_{i=1}^{n} \left(\log \frac{f(y_i|\mathbf{x}_i, \hat{\boldsymbol{\theta}})}{g(y_i|\mathbf{x}_i, \hat{\boldsymbol{\gamma}})} \right)^2$$

Seja z_α o valor crítico do teste, podemos ter três situações:

1. Se $\mathrm{V} > z_\alpha$, então H_g é rejeitada em favor de $\mathrm{H}_f : \mathrm{E}_0[\log(f/g)] > 0$;
2. Se $\mathrm{V} < -z_\alpha$, então H_f é rejeitada em favor de $\mathrm{H}_g : \mathrm{E}_0[\log(f/g)] < 0$;
3. Se $|\,\mathrm{V}\,| < z_\alpha$, então a discriminação entre os dois modelos não é possível.

3.3 Simulação de Cenários

Para comparar os testes descritos neste Capítulo, foram feitas simulações dos cenários propostos por Blizzard e Hosmer (2006), que serão designados por BH1 a BH12 e os cenários propostos por Savu, Liu e Yasui (2010), que serão designados por S1 a S8. Além disso, foram gerados mais 12 cenários baseados em BH1 a BH12, com alteração da função resposta do processo gerador de dados (PGD) de $\log^{-1}$ para logit^{-1} que foram designados por BH1-logit a BH12-logit. Foram geradas 100 simulações de cada cenário, cada simulação com $n = 500$ registros e a estimativa de $\hat{\boldsymbol{\gamma}}_*$, para o teste de Cox, foi calculada por meio de $R = 200$ repetições.

Os cenários BH1 a BH8 apresentam uma variável explicativa x e a resposta Y é gerada de tal forma que

$$P(Y = 1|x) = \log^{-1}(\beta_0 + \beta_1 x) = \exp(\beta_0 + \beta_1 x),$$

com $x \sim Uniforme(-6, a)$ e valores para β_0, β_1 e a conforme Tabela 9. No caso de BH1-logit a BH8-logit, substituímos $\log^{-1}$ por logit^{-1}, assim,

$$P(Y = 1|x) = \text{logit}^{-1}(\beta_0 + \beta_1 x) = \left(\frac{\exp(\beta_0 + \beta_1 x)}{1 + \exp(\beta_0 + \beta_1 x)}\right).$$

Tabela 9 – Especificações dos Cenários BH1 a BH8

Cenário	β_0	β_1	a
BH1	-2,30259	0,38376	6
BH2	-2,30259	0,38376	4
BH3	-1,20397	0,56687	2
BH4	-1,20397	0,56687	1
BH5	-0,69315	0,65200	1
BH6	-0,69315	0,65200	0
BH7	-0,35667	0,70808	0,5
BH8	-0,35667	0,70808	-0,5

Os cenários BH9 a BH12 apresentam duas variáveis explicativa u e d e a resposta Y é gerada de tal forma que

$$P(Y = 1|u, d) = \log^{-1}(\beta_0 + \beta_d d + \beta_u u) = \exp(\beta_0 + \beta_d d + \beta_u u),$$

com $d \sim Bernoulli(\pi)$, $u|d \sim Uniforme(-6 + 2d, 2 + 2d)$ e valores para β_0, β_d, β_u e π conforme Tabela 10. No caso de BH9-logit a BH12-logit, substituimos $\log^{-1}$ por logit^{-1}, assim,

$$P(Y = 1|u, d) = \text{logit}^{-1}(\beta_0 + \beta_d d + \beta_u u) = \left(\frac{\exp(\beta_0 + \beta_d d + \beta_u u)}{1 + \exp(\beta_0 + \beta_d d + \beta_u u)}\right).$$

Tabela 10 – Especificações dos Cenários BH9 a BH12

Cenário	β_0	β_d	β_u	π
BH9	log(0,3)	log(1,5)	0,18	0,2
BH10	log(0,3)	log(1,5)	0,18	0,5
BH11	log(0,3)	log(2,0)	0,10	0,2
BH12	log(0,3)	log(2,0)	0,10	0,5

Os cenários S1 a S8 apresentam quatro variáveis explicativa x_1, x_2, x_3 e E, tais que $x_1 \sim Bernoulli(0, 5)$, $x_2 \sim Multinomial(0, 3, 0, 3, 0, 4)$, $x_3 \sim Uniforme(-1, 2)$ e

$$P(E = 1|\mathbf{x}) = \text{logit}^{-1}(-1 + x_1 - x_{2(2)} + x_{2(3)} + x_3).$$

Se $E = 0$, então, a resposta Y é gerada de tal forma que

$$P(y = 1|E = 0, \mathbf{x}) = g(\gamma_0 + h(\mathbf{x})). \tag{3.4}$$

Se $E = 1$, então, a resposta Y é gerada de tal forma que

$$P(y = 1|E = 1, \mathbf{x}) = 3g(\gamma_0 + h(\mathbf{x})) \tag{3.5}$$

Os valores de γ_0 e as funções g e h são descritas na Tabela 11. As funções g são as inversas das principais funções de ligação utilizadas para respostas binárias.

Tabela 11 – Especificações dos Cenários S1 a S8

Cenário	γ_0	g	h
S1	-2,10	$\exp(x)$	$-(x_1 + x_{2(2)} + x_{2(3)} + x3)$
S2	-1,90	$1 - \exp(-\exp(x))$	$-(x_1 + x_{2(2)} + x_{2(3)} + x3)$
S3	-1,70	logit^{-1}	$-(x_1 + x_{2(2)} + x_{2(3)} + x3)$
S4	-1,48	probit^{-1}	$-(x_1 + x_{2(2)} + x_{2(3)} + x3)$
S5	-1,10	$\exp(x)$	$-max(x_1 + x_{2(2)} + x_{2(3)} + x3, 0)$
S6	-0,90	$1 - \exp(-\exp(x))$	$-max(x_1 + x_{2(2)} + x_{2(3)} + x3, 0)$
S7	-0,70	logit^{-1}	$-max(x_1 + x_{2(2)} + x_{2(3)} + x3, 0)$
S8	-0,48	probit^{-1}	$-max(x_1 + x_{2(2)} + x_{2(3)} + x3, 0)$

Os cenários BH1 a BH12 e S1 a S8 foram elaborados para testar convergência de algoritmos e metodologias para estimação de modelos log-binomiais. Ao aplicar os mesmos cenários nos testes de seleção de modelos não aninhados, observamos um fato curioso. Os PGDs utilizam as inversas das principais funções de ligação para dados binários, log, log-log complementar, logit e probit, entretanto, o modelo ajustado com a função resposta correspondente não necessariamente traz o menor AIC. O PGD garante que os dados seguirão determinado modelo, entretanto, não garante que esse será o que apresenta melhor ajuste.

Foram geradas 100 simulações de cada um dos 32 cenários descritos acima, nos quais foram ajustados MLGs com resposta binária e função de ligação log, logit, probit e log-log complementar e, então, verificado qual apresentava o menor AIC. Os resultados são apresentados no Anexo A. Podemos observar que, em alguns casos, como em BH1, BH3, BH5, BH7, BH10 e BH12, existe predominância do MLG correspondente ao PDG. Entretanto, em outros casos como em BH2, BH4, BH6, BH8, BH6-logit, BH8-logit, BH10-logit e BH11-logit essa predominância é menos acentuada. Por outro lado, pode acontecer de a maioria dos ajustes feitos por outra função de ligação ser melhor, como em BH12-logit, BH7-logit e BH1-logit. No caso dos cenários S1 a S8, isso ocorre com mais frequência, possivelmente porque, como visto em (3.4) e (3.5), o PGD utiliza o valor gerado pela função resposta do MLG somente nos casos em que $E = 0$, quando $E = 1$, esse valor é multiplicado por 3 o que pode estar gerando dados com distribuição distinta.

Se fazem necessários estudos mais detalhados sobre o processo de criação de cenários de forma a produzir situações em que exista predominância de melhor ajuste do modelo gerador dos dados. Isso permitirá testar com mais precisão o poder dos testes de COX e

VUONG. Durante o processo de simulação podem estar sendo gerados outliers que fazem com que a função geradora não corresponda ao melhor ajuste.

Foram realizados os testes de COX e VUONG nas simulações, para escolha entre o modelo log-binomial ou logístico, os resultados são apresentados no Anexo B. A quantidade de simulações em que cada modelo apresentou menor AIC está nas duas últimas colunas. Estes resultados diferem dos apresentados Anexo A, pois neste caso estão sendo confrontados somente dois modelos. Esses valores nos dão uma noção se o cenário gera dados que seguem um modelo em específico ou se não existe predominância. De fato, quando existe predomínio de menor AIC no modelo log-binomial, os testes apresentam maior número de rejeição do modelo logístico em favor do modelo log-binomial e vice e versa. Observamos, ainda, que o teste de Cox, em geral, rejeita mais a hipótese nula do que o modelo de Vuong.

Além dos testes de COX e VUONG, uma terceira sugestão de investigação para escolha do modelo que se ajusta melhor aos dados é por meio de *bootstrap*. Como já existem soluções para os problemas de ajuste do modelo log-binomial como descrito na seção 2.3, podemos ajustar os modelo log-binomial e logístico para n reamostragens dos dados e verificar se algum modelo tem predominância de ajuste melhor.

3.4 Medidas de diagnóstico de acurácia

As medidas de diagnóstico de acurácia indicam a capacidade de predição do modelo, isto é, a capacidade de diferenciar os indivíduos que apresentam a característica de interesse (positivos) dos que não apresentam (negativos). Por meio dos modelos log-binomial e logístico, obtemos as estimativas $\hat{p}_i$, para determinar os valores preditos $\hat{y}_i \in \{0, 1\}$, uma regra usual é

$$\hat{y}_i = \begin{cases} 0, \text{ se } \hat{p}_i \leq 0,5 \\ 1, \text{ se } \hat{p}_i > 0,5 \end{cases}$$

em que 0,5 pode ser chamado de ponto de corte. Outros valores de ponto de corte podem ser escolhidos dependendo do quão grave ou dispendioso é cometer os erros tipo I (falso positivo) e tipo II (falso negativo) no contexto em estudo. Quanto maior o ponto de corte, maior será o erro tipo II e quanto menor o ponto de corte maior será o erro tipo I.

Após a definição do ponto de corte, pode ser elaborada uma tabela de contingência dos valores observados versus valores preditos, também chamada de matriz de confusão. Os elementos da matriz de confusão, são apresentados na Tabela 12.

A partir da matriz de confusão, podemos calcular as medidas de acurácia, sensibilidade e especificidade.

A acurácia é proporção de predições corretas, sem considerar se são positivos ou

Tabela 12 – Matriz de Confusão

	$y_i = 0$	$y_i = 1$
$\hat{y}_i = 0$	VN (verdadeiro negativo)	FN (falso negativo)
$\hat{y}_i = 1$	FP (falso positivo)	VP (verdadeiro positivo)

Nota: y_i são os valores observados e $\hat{y}_i$ são os valores preditos

negativos,

$$\text{Acurácia} = \frac{\text{VP} + \text{VN}}{\text{VP} + \text{FP} + \text{VN} + \text{FN}}.$$

A sensibilidade é proporção de verdadeiros positivos e mede a capacidade do modelo de classificar um indivíduo como positivo ($\hat{y}_i = 1$), dado que realmente é positivo ($y_i = 1$),

$$\text{Sensibilidade} = \frac{\text{VP}}{\text{VP} + \text{FN}}.$$

A especificidade é proporção de verdadeiros negativos e mede a capacidade do modelo de classificar um indivíduo como negativo ($\hat{y}_i = 0$), dado que realmente é negativo ($y_i = 0$),

$$\text{Especificidade} = \frac{\text{VN}}{\text{VN} + \text{FP}}.$$

A curva ROC é construída por meio de um gráfico cujo eixo das ordenadas é dado pela Sensibilidade e o eixo das abscissas por 1-Especificidade para diferentes valores de ponto de corte. Quanto maior a área sob a curva ROC (AUROC), maior a capacidade do modelo de diferenciar os positivos dos negativos.

3.5 Exemplos

Como exemplos de aplicação, foram utilizados os conjuntos de dados `Birth`, `Heart` e `PCS` do pacote `lbreg` descritos na seção 2.3.6 e uma das simulações do cenário BH1. Para a realização dos testes de Cox e Vuong, foi desenvolvida a função `cv.test` em `R`, constante no Anexo C.2. Essa função será incluída na próxima versão do pacote `lbreg`. No cálculo da área sob a curva ROC, foi utilizado o pacote `pROC` (ROBIN et al., 2011).

3.5.0.1 Conjunto de dados `Birth`

No teste de Cox - log vs logit para o conjunto de dados `Birth`, a estatística do teste foi 0,43, portanto, não rejeitamos o modelo log-binomial em favor do logístico. No teste de Cox - logit vs log, a estatística do teste foi -0,90, portanto não rejeitamos o modelo logístico em favor do log-binomial. Se combinarmos os dois testes de Cox, como descrito na seção 3.1, chegamos à decisão de não rejeitar nenhum dos dois modelos. No teste de Vuong, a estatística do teste foi 0,64, portanto, concluímos que não há diferença entre os dois ajustes.

Foram retiradas 1000 amostras com reposição de tamanho 900 do conjunto de dados `Birth`, para as quais foram ajustados os modelos log-binomial e logístico. Dessas 1000 amostras, 648 apresentaram AIC menor no modelo log-binomial e 352 no logístico, portanto não existe predominância de nenhum modelo, o que corrobora com os resultados dos testes de Cox e Vuong.

A Tabela 13 exibe as medidas de diagnóstico de acurácia. Não foi possível utilizar o ponto de corte 0,5 pois o valor máximo de $\hat{p}_i$ foi 0,28 e 0,29 no ajuste dos modelos log-binomial e logístico respectivamente. O conjunto de dados `Birth` apresenta somente três covariáveis categóricas, duas delas com três classes e uma com duas classes, portanto, só existem 18 valores possíveis para o vetor $\mathbf{x}_i = (1, x_{i1}, x_{i2}, x_{i3})$. As medidas de acurácia, sensibilidade e especificidade foram calculadas utilizando o ponto de corte 0,14 que é o valor do terceiro quartil dos valores estimados $\hat{p}_i$ em ambos os modelos.

3.5.0.2 Conjunto de dados `Heart`

No teste de Cox - log vs logit para o conjunto de dados `Heart`, a estatística do teste foi -4,64, portanto, rejeitamos o modelo log-binomial em favor do logístico. No teste de Cox - logit vs log, a estatística do teste foi 2,84, portanto rejeitamos o modelo logístico em favor do log-binomial. Assim, combinando os dois testes de Cox, como descrito na seção 3.1, chegamos à decisão de rejeitar os dois modelos. No teste de Vuong, a estatística do teste foi -3,98, portanto, rejeitamos o modelo log-binomial em favor do logístico.

Foram retiradas 1000 amostras com reposição de tamanho 16949 do conjunto de dados `Heart`, para as quais foram ajustados os modelos log-binomial e logístico. Em todas as 1000 amostras o menor AIC foi do modelo logístico.

A Tabela 13 exibe as medidas de diagnóstico de acurácia, foi considerado o ponto de corte 0,5 para calcular as medidas de acurácia, sensibilidade e especificidade. Apesar do modelo logístico ter sido escolhido pelo testes de Vuong e ter apresentado menor AIC em todas as reamostragens, para predição, não parece haver grande vantagem.

3.5.0.3 Conjunto de dados `PCS`

No teste de Cox - log vs logit para o conjunto de dados `PCS`, a estatística do teste foi -4.83, portanto, rejeitamos o modelo log-binomial em favor do logístico. No teste de Cox - logit vs log, a estatística do teste foi -0.06, portanto não rejeitamos o modelo logístico em favor do log-binomial. Portanto, combinando os dois testes de Cox, como descrito na seção 3.1, chegamos à decisão de rejeitar o modelo log-binomial em favor do logístico e a mesma decisão pode ser tomada pelo resultado do teste de Vuong, cuja a estatística de teste foi -3,98.

Os três registros que apresentavam valor nulo foram desconsiderados e foram

retiradas 1000 amostras com reposição de tamanho 377 do conjunto de dados `Heart`. Foram ajustados os modelos log-binomial e logístico. Em todas as 1000 amostras o modelo logístico apresentou o menor AIC.

A Tabela 13 exibe as medidas de diagnóstico de acurácia, foi considerado o ponto de corte 0,5 para calcular as medidas de acurácia, sensibilidade e especificidade. Os valores confirmam os resultados do testes já que o modelo logístico apresenta maior capacidade de predição.

3.5.0.4 Conjunto de dados simulado do cenário BH1

Afim de confrontar os resultados obtidos nos conjuntos de dados anteriores, foi selecionado um conjunto de dados em que o modelo log-binomial foi escolhido nos testes. Em uma das simulações com 500 registros do cenário BH1 a estatística do teste de Cox - log vs logit foi 0,60 e a estatística do teste de Cox - logit vs log foi -3,43. Combinando o dois testes, chegamos à decisão de rejeitar o modelo logístico em favor do log-binomial. Nessa mesma simulação, a estatística do teste de Vuong foi 2,06, portanto também rejeita o modelo o modelo logístico em favor do log-binomial. Realizando o bootsrap como descrito acima, 964 amostras tiveram o melhor ajuste no modelo log-binomial e somente 36 no logístico.

A Tabela 13 exibe as medidas de diagnóstico de acurácia, foi considerado o ponto de corte 0,5 para calcular as medidas de acurácia, sensibilidade e especificidade. Apesar do modelo log-binomial ter sido escolhido pelo testes de Vuong e ter apresentado menor AIC na maioria das reamostragens, para predição, não parece haver grande vantagem.

Tabela 13 – Medidas de diagnóstico de acurácia dos conjuntos de dados `Birth`, `Heart`, `PCS` e uma das simulações do cenário BH1

	`Birth`		`Heart`		`PCS`		BH1	
	log	logit	log	logit	log	logit	log	logit
Acurácia	0,748	0,719	0,938	0,938	0,695	0,751	0,878	0,870
Senstividade	0,337	0,398	0,021	0,030	0,344	0,623	0,539	0,560
Especificidade	0,798	0,758	0,998	0,998	0,929	0,836	0,954	0,939
AUROC	0,624	0,624	0,758	0,759	0,801	0,826	0,880	0,880

4 Estimação Bayesiana

Na abordagem Clássica os parâmetros são vistos como valores fixos. A inferência frequentista se baseia na possibilidade de repetibilidade dos experimentos e muitas vezes recorre a resultados assintóticos. O nível de confiança de um intervalo de confiança calculado para uma amostra de tamanho n é a probabilidade do intervalo conter o verdadeiro parâmetro, se fossem retiradas todas as amostras possíveis de tamanho n.

Na abordagem Bayesiana, os parâmetros possuem uma distribuição de probabilidade. A distribuição inicial dos parâmetros é chamada de priori, por meio do Teorema de Bayes, essa distribuição é atualizadas com a informação contida nos dados, então, tem-se a distribuição a posteriori dos parâmetros. O nível de confiança do intervalo de credibilidade tem uma interpretação direta e representa a probabilidade do verdadeiro parâmetro pertencer ao intervalo.

A cultura de dados é uma necessidade a qualquer empresa na atualidade, entretanto, ainda não é uma realidade, em especial nas empresas de menor porte. Na prática cotidiana, muitas vezes não temos disponível a informação de forma estruturada em base de dados acessíveis. Apesar da escassez de dados consistentes, em geral, os envolvidos na atividade têm um bom conhecimento do que pode influenciar na predição, já que estão habituados a tomar suas decisões com base nas suas percepções. A informação é um bem valioso para as empresas e não pode ser desperdiçada. A inferência bayesiana permite converter o conhecimento de um especialista em informação quantitativa para ser usada de forma objetiva.

Gelman et al. (2014) afirmam que a ideia de informação a priori já está implicitamente presente na regressão clássica, por exemplo, quando não incluímos no estudo variáveis que sabemos não ter valor preditivo.

A abordagem bayesiana oferece uma variedade de metodologias, uma vez que, além da necessidade de definir o modelo que será usado para a variável resposta, ainda abre espaço para a discussão da distribuição inicial dos parâmetros.

Diversos autores discutem e propõem diferentes tipos de distribuição a priori. Para nortear o presente trabalho foram selecionados alguns artigos e livros que serão usados como referência e estão elencados na Tabela 14.

4.1 Análise Bayesiana

Em modelos bayesianos, atualizamos a informação que possuímos a priori a respeito de β, resumida na densidade $\pi(\beta)$, com a informação dos dados observados $X = x$ por

meio do teorema de Bayes.

A informação contida nos dados é representada pela verossimilhança $L(\beta; x)$. A distribuição de β após a atualização é chamada de distribuição a posteriori e dada por:

$$\pi(\beta|x) = \frac{\pi(\beta)L(\beta; x)}{\int \pi(\beta)L(\beta; x)d\beta} \propto \pi(\beta)L(\beta|x).$$

DeGroot (2005) apresenta um exemplo ilustrativo de como a Inferência Bayesiana difere da Inferência Clássica. Suponha que são amostrados independentemente itens para verificar se são defeituosos de diferentes formas:

1. Uma amostra aleatória de n itens é selecionada, n é inteiro positivo fixo;
2. Itens são selecionados, um a um, até que exatamente y itens defeituosos tenham sido obtidos, y é inteiro positivo fixo;
3. Itens são selecionados ao acaso, um a um, até que o inspetor seja chamado para resolver um outro problema;
4. Itens são selecionados ao acaso até que o inspetor acredite que já possui informação suficiente sobre θ.

Não importa como foi retirada a amostra, se foram observados n itens $(x_1, ..., x_n)$, dos quais y eram defeituosos,

$$L(\theta; x) \propto \theta^y(1-\theta)^{n-y}.$$

Portanto, a distribuição a posteriori será idêntica e, consequentemente, a inferência a respeito de θ também será a mesma. O mesmo não ocorre na modelagem clássica, na qual haveria 4 modelos distintos.

No caso dos Modelos Binomiais, a verossimilhança é dada por:

$$L(\boldsymbol{\beta}; x) = \prod_{i=1}^{n} \exp(x'\beta)^{y_i}(1 - \exp(x'\beta))^{1-y_i} I_\Theta(\boldsymbol{\beta}).$$

O termo $I_\Theta(\boldsymbol{\beta})$ é a função indicadora do espaço paramétrico de $\boldsymbol{\beta}$, $I_\Theta(\boldsymbol{\beta}) = 1$ se $\boldsymbol{\beta} \in \Theta$ e $I_\Theta(\boldsymbol{\beta}) = 0$ se $\boldsymbol{\beta} \notin \Theta$. No modelo log-binomial, $\Theta = \{\boldsymbol{\beta} : \boldsymbol{x'\beta} \leq 0\}$, no modelo logístico com intercepto e k covariáveis $\Theta = \{\mathbb{R}^{k+1}\}$.

A distribuição a priori $\pi(\beta)$ pode assumir diversos formatos dependente do conhecimento prévio que se tem do problema, esse assunto é abordado na próxima seção.

A abordagem bayesiana do modelo log-binomial é pouco discutida na literatura. Efetuando-se uma busca pelos termos log-binomial e *bayesian*, encontramos somente 11

artigos na base *Web of Science*. Retirando-se os artigos que não discutem o modelo log-binomial bayesiano, mas somente realizam uma aplicação em dados reais, sem entrar em detalhes do método estatístico, restam os seguintes artigos de Chu e Cole (2010), Zhou, Sivaganesan e Longla (2014), Torman e Camey (2015), Salmerón, Cano e Chirlaque (2015), Pedroza et al. (2016) e Janani et al. (2017).

O primeiro artigo a tratar da abordagem bayesiana para o modelo log-binomial é o de Chu e Cole. Os autores utilizam uma distribuição a priori normal, $\beta_j \sim N(0, 10^2)$. A estimação é feita por MCMC por meio do WINBUGS.

4.2 Prioris

As distribuições a priori podem ser caracterizadas por diferentes aspectos.

1. Quanto ao nível de informação, a distribuição a priori pode ser:
   - não informativa, assume completa ignorância a respeito dos parâmetros;
   - informativa, assume algum grau de conhecimento a respeito dos parâmetros.
2. Quanto a aplicação, a distribuição a priori pode ser:
   - específica para determinado problema;
   - genérica, podendo ser usada como um escolha padrão, também chamada de *reference prior*.
3. Quanto a depender ou não da amostra, a distribuição a priori pode ser:
   - subjetiva, quando não depende dos dados que compõem a verossimilhança;
   - objetiva, quando depende de alguma informação oriunda dos dados que compõem a verossimilhança.
4. Quanto a integrabilidade, a distribuição a priori pode ser:
   - própria se é integrável positiva;
   - imprópria se não é integrável.

Se a priori é própria a posteriori será própria também. Se a priori for imprópria, é preciso garantir que a posteriori seja própria para permitir realizar inferências a respeito dos parâmetros.

4.2.1 Priori uniforme

Podemos chamar de priori uniforme, aquela em que todos os valores possíveis de β apresentam igual chance de ocorrer, isto é, segue distribuição uniforme.

$$\pi(\boldsymbol{\beta}) \propto k.$$

Se o intervalo de variação de β é ilimitado, então a priori é imprópria, pois sua integral não é igual a 1.

Kass e Wasserman (1996) afirmam que Bayes em 1763 utilizou uma distribuição a priori constante para estimar o parâmetro de uma distribuição binomial. Já Datta e Ghosh (1996) declaram que a priori uniforme foi introduzida por Laplace em 1812.

Alguns autores chamam a distribuição a priori uniforme de "flat", ao passo que outros utilizam o termo para qualquer priori não informativa. A fim de evitar confusão não utilizaremos a expressão "flat".

Uma distribuição é dita própria quando é integrável. Quando a distribuição a priori é própria a distribuição a posteriori é sempre própria. Quando a distribuição a priori não é própria, é preciso verificar se a posteriori é própria.

No caso do modelo log-binomial a priori uniforme é imprópria, Zhou, Sivaganesan e Longla (2014) demonstram que para a existência da distribuição a posteriori de $(\beta_0, \beta_1, ..., \beta_k)$, basta que o espaço vetorial abrangido por S tenha posto maior ou igual a $k+1$, sendo S o conjunto de todos os vetores para os quais $Y_i = 1$, ou seja,

$$S = \{(1, x_{1i}, ..., x_{ki}) : Y_i = 1, 1 \leq i \leq n\} \subset \mathbb{R}^{k+1}.$$

O que equivale a dizer que para a existência da posteriori, basta existir pelo menos $k+1$ registros não idênticos com $Y = 1$.

4.2.2 Priori Própria

Zhou, Sivaganesan e Longla (2014) propõem a construção de uma priori não informativa própria para modelos log-binomiais, conforme descrito abaixo.

Inicialmente, deve-ser centralizar as covariáveis em zero (valor mínimo negativo e o máximo positivo). Em seguida, atribuir como distribuição marginal a priori para β_0 uma distribuição exponencial padrão,

$$\pi(\beta_0) \propto \exp(\beta_0), \text{ para } \beta_0 < 0,$$

dessa forma, com β_0 fixado, o espaço paramétrico de $(\beta_1, ..., \beta_k)$ passa a ser

$$\Theta_{\beta_0} = \{(\beta_1, ..., \beta_k) : \beta_1 x_{1i} + ... + \beta_k x_{ki} \leq -\beta_0, i = 1, ..., n\}.$$

Quando Θ_{β_0} é limitado, é possível atribuir como distribuição condicional para $(\beta_1, ..., \beta_k)$ uma distribuição uniforme,

$$\pi(\beta_1, ..., \beta_k|\beta_0) \propto \frac{1}{|\beta_0|^k},$$

Portanto,

$$\pi(\boldsymbol{\beta}) = \pi(\beta_1, ..., \beta_k|\beta_0)\pi(\beta_0) \propto \frac{\exp(\beta_0)}{|\beta_0|^k}.$$

Quando Θ_{β_0} não é limitado, os autores sugerem a utilização de prioris independentes Cauchy, como descrito na seção 4.2.5.

4.2.3 Priori de Jeffreys

A priori de Jefrreys foi a primeira das chamadas prioris de referência, que são prioris escolhidas a partir de uma regra formal e podem também ser chamadas de *conventional prior*, *default prior* ou *generic prior* (KASS; WASSERMAN, 1996).

Jeffreys acreditava que uma priori não informativa deveria ser invariante para reparametrizações de β o que não ocorre com a priori uniforme, então propôs uma priori invariante para transformações monótonas que considera mais provável os valores de β com maior Informação de Fisher.

Conforme a regra de Jeffreys (1946), a priori é proporcional à raiz quadrada do determinante da matriz de informação de Fisher.

$$\pi(\boldsymbol{\beta}) \propto |\det(I(\boldsymbol{\beta}))|^{\frac{1}{2}}.$$

Essa priori pode ser própria ou imprópria. Ibrahim e Laud (1991) mostram uma condição suficiente para a posteriori ser própria e uma condição necessária e suficiente para a priori de Jeffreys ser própria, no caso geral da família exponencial com dispersão: sejam $y_1, ..., y_n$ respostas independentes com densidade f tal que

$$f(y_i; \theta_i, \phi) = \exp\left\{\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right\}$$

com suporte livre de parâmetros tal que

$$\theta_i = g(x_i'\beta)$$

Ibrahim e Laud (1991) mostram que se

$$\int_S \exp\left\{\phi^{-1}w(yr - b(r))\right\}\left(\frac{\partial^2 b(r)}{\partial r^2}\right)^{\frac{1}{2}} dr < \infty,$$

em que S é o espaço paramétrico do parâmetro canônico θ, então a distribuição a posteriori de β é própria.

Além disso, se e somente se

$$\int_S \left(\frac{\partial^2 b(r)}{\partial r^2} \right)^{\frac{1}{2}} dr < \infty,$$

então a priori de Jeffrey é própria.

4.2.4 G-prior

Zellner propôs as chamadas *g-priors* para modelos lineares, atribuindo uma distribuição normal multivariada para $\boldsymbol{\beta}$ e uma priori de Jeffreys para σ^2 .

$$\boldsymbol{\beta}|\sigma^2 \sim N_k(\tilde{\boldsymbol{\beta}}, g\sigma^2(X'X)^{-1}),$$
$$\pi(\sigma^2) \propto \sigma^{-2}.$$

Assim, a escolha da priori se limita à definição das médias $\widetilde{\beta}$ da normal multivariada e de uma constante g.

Essa constante pode ser interpretada como uma medida inversamente proporcional a quantidade de informação disponível na priori em relação à amostra. Se $g = 1$ o peso da priori é o mesmo que da amostra, se $g = 2$ a priori tem o mesmo peso que 50% da amostra, se $g = 100$ a priori tem o mesmo peso que 1% da amostra e se $g = n$ (n é o tamanho da amostra) a priori tem o mesmo peso que uma única observação da amostra (MARIN; ROBERT, 2014).

O fato da priori depender de X não constitui um problema, uma vez que todo o modelo é condicionado em X. Já no caso em que X é composto por variáveis defasadas, como em séries temporais, não é recomendado o uso das *g-priors* (MARIN; ROBERT, 2014).

4.2.5 Priori t-student fracamente informativa

Gelman et al. (2008) afirmam que antes de sua proposta de metodologia, toda a literatura só oferecia duas opções extremas de distribuição a priori: distribuições totalmente informativas, usando informações específicas para o problema ou distribuições não informativas em geral motivadas pelo princípio da invariância. Então propuseram uma opção intermediária, uma distribuição informativa, mas que possa ser usada em uma gama grande de aplicações.

O trabalho de Gelman et al. (2008) se relaciona com a metodologia das Prioris de Médias Condicionais de Bedrick, Christensen e Johnson (1996) (seção 4.2.6) ao determinar

a distribuição a priori com base nos possíveis efeitos que alterações nos parâmetros podem causar na resposta. Por outro lado, se diferencia das Prioris de Médias Condicionais ao propor uma metodologia genérica que pode ser usada em variadas situações evitando a necessidade de refazer o estudo dos efeitos das alterações dos parâmetros a cada novo estudo.

A fim de se padronizar o método para que possa ser usado em variadas situações, é necessário escalonar as variáveis não binárias para que apresentem média 0 e desvio padrão igual a 0,5 e as variáveis binárias para que tenham média 0 e amplitude 1. Por exemplo, uma variável binária que apresenta 10% de 1 e 90% de 0, passaria a apresentar os valores -0.1 e 0.9, respectivamente.

A distribuição t-student apresenta somente o parâmetro de forma, dado pelos graus de liberdade ν, e tem função densidade de probabilidade dada por:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{x^2}{\nu}\right)^{-\left(\frac{\nu+1}{2}\right)}, x \in \mathbb{R}$$

A distribuição t-student localização-escala é expressa por meio de três parâmetros, localização μ, escala σ e forma ν, e tem função densidade de probabilidade dada por:

$$f(x) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sigma\sqrt{\nu\pi}\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{(x-\mu)^2}{\nu\sigma^2}\right)^{-\left(\frac{\nu+1}{2}\right)}.$$

Gelman et al. (2008) sugerem como distribuição a priori para os coeficientes, uma t-student localização-escala com média (localização) 0, escala σ e graus de liberdade (forma) ν. No caso da regressão logística, recomendam a distribuição a priori Cauchy com parâmetro de localização 0 e escala 2.5 para cada coeficiente e uma Cauchy com parâmetro de localização 0 e escala 10 para o intercepto. A distribuição Cauchy coincide com a distribuição t-student com um grau de liberdade.

Gelman et al. (2008) propõem um método de estimação dos parâmetros com distribuição a priori t-student para modelos lineares generalizados incorporando um algoritmo EM ao método de mínimos quadrados ponderados. A distribuição t-student é expressa como uma mistura de distribuições normais com variância desconhecida, o algoritmo EM é incluído no algoritmo de mínimos quadrados ponderados iterativos para estimar essa variância. O método está implementada em `R` por meio da função `bayesglm` do pacote `arm`.

4.2.6 Prioris de Médias Condicionais

Em muitas situações, o maior conhecedor do problema que se quer descrever em forma de modelos estatísticos, tem pouco ou nenhum conhecimento de distribuições de

probabilidade. Dessa forma, definir uma distribuição a priori para os parâmetro se torna um desafio.

Uma forma acessível de se tratar o conhecimento prévio é inquirir o especialista a respeito da sua opinião das valores da variável resposta que seriam observados caso fossem considerados determinados valores das covariáveis. A partir desse conhecimento empírico, podemos chegar a uma priori para os parâmetros. Com esse intuito, Bedrick, Christensen e Johnson (1996) propuseram distribuições a priori informativas para modelos lineares generalizados chamadas Priori de Médias Condicionais (CMP) que podem ser aplicados a MLGs.

No caso dos modelos de resposta binomial, em que $y_i|\boldsymbol{x}_i \sim binomial(n_i, p_i)$ e $p_i = E(y_i|\boldsymbol{x}_i)$, a priori de $\boldsymbol{\beta}$ é induzida por $\tilde{\boldsymbol{p}} = (\tilde{p_1}, ..., \tilde{p_k})'$, $\tilde{p_i} = E(\tilde{y_i}|\tilde{\boldsymbol{x}}_i)$ é a probabilidade de sucesso pressuposta pelo especialista para um certo vetor de covariáveis $\tilde{\boldsymbol{x}}_i$. A distribuição Beta é uma escolha comum para distribuição a priori de probabilidades, com base no conhecimento do especialista sobre os valores prováveis para os $\tilde{p_i}$ define-se a_{1i} e a_{2i}, tais que $\tilde{p_i} \sim beta(a_{1i}, a_{2i})$, assim,

$$\pi_0(\tilde{\boldsymbol{p}}) \propto \prod_{i=1}^{k} \tilde{p_i}^{a_{1i}-1}(1-\tilde{p_i})^{a_{2i}-1}.$$

Seja $\tilde{\boldsymbol{X}}$ a matriz $k \times k$ formada por $\tilde{\boldsymbol{x}}_i{}'$ como suas linhas, $g(.)$ a função de ligação, $r(.) = g^{-1}(.)$ a função resposta, $R(.)$ uma transformação que aplica $r(.)$ a cada elemento de um vetor e $\pi_0(\tilde{\boldsymbol{p}})$ a distribui a priori em $\tilde{\boldsymbol{p}}$, então, usando a técnica de mudança de variável:

$$\pi(\boldsymbol{\beta}) = \pi_0(R(\tilde{\boldsymbol{X}}\boldsymbol{\beta}))\left|\frac{\partial R(\tilde{\boldsymbol{X}}\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\right|.$$

Assim priori em $\boldsymbol{\beta}$ é dada por:

$$\pi(\boldsymbol{\beta}) \propto \prod_{i=1}^{k} r(\tilde{x}_i'\beta)^{a_{1i}-1}[1-r(\tilde{x}_i'\beta)]^{a_{2i}-1}\dot{r}(\tilde{x}_i'\beta),$$

em que, $\dot{r}(.)$ é a primeira derivada de $r(.)$. Bedrick, Christensen e Johnson (1997) dão exemplos de como definir a_{1i} e a_{2i}.

4.2.7 Exemplos

Como exemplos de aplicação, foram realizados ajustes de modelos com priori uniforme (4.2.1) e com priori t-student fracamente informativa (4.2.5) nos conjuntos de dados do pacote `lbreg` descritos na seção 2.3.6.

As estimativas dos parâmetros de regressão do modelo log-binomial com priori uniforme foram realizadas utilizando o algoritmo de Metropolis-Hasting por meio da função `MCMCmetrop1R` do pacote `MCMCpack` (MARTIN; QUINN; PARK, 2011).

Tabela 14 – Referência de artigos

Autor	**Ano**	**Priori**	**Modelo**
Jeffreys (1946)	1946	Jeffreys	Geral
Zellner (1983)	1983	G-prior	Linear
Ibrahim e Laud (1991)	1991	Jeffreys	MLG
Berger e Bernardo (1992)	1992	Reference Prior	Geral
Bedrick, Christensen e Johnson (1996)	1996	Conditional Means Prior	MLG
Bedrick, Christensen e Johnson (1997)	1997	Conditional Means Prior	Logistico
Datta e Mukerjee (2004)	2004	Matching Prior	Geral
Genkin, Lewis e Madigan (2007)	2007	Laplace	Logístico
Marin e Robert (2014)	2007	G-prior	MLG
Gelman et al. (2008)	2008	t-student	MLG
Chu e Cole (2010)	2010	$N(0, 10^2)$	Log-Binomial
Zhou, Sivaganesan e Longla (2014)	2014	$\exp(\beta_0)/\lvert\beta_0\rvert^k$	Log-Binomial
Hanson et al. (2014)	2014	G-prior	Logístico

As estimativas dos parâmetros de regressão do modelo log-binomial com priori t-student fracamente informativa foram realizadas por meio da função `bayesglm` do pacote `arm`.

4.2.7.1 Conjunto de dados `Birth`

No conjunto de dados `Birth` que possui 900 registros, apesar de termos utilizado prioris difusas (uniforme e t-student), essas prioris parecem estar influenciando na estimativa dos parâmetros, uma vez que os resultados de todas as metodologias aplicadas no Capítulo 2 geraram resultados semelhantes divergindo somente a partir da quarta casa decimal e os resultados dos modelos bayesianos divergem já a partir da segunda casa decimal. Os valores estimados utilizando as prioris uniforme e t-student, são apresentados na Tabela 15 e na Tabela 16 respectivamente.

Tabela 15 – Estimativas para o conjunto de dados `Birth` - priori uniforme

	Estimativa	Desvio Padrão	Erro Padrão	Erro Padrão de Monte Carlo
Intecepto	-2,795	0,207	0,001	0,007
alc2	0,165	0,277	0,002	0,009
alc3	0,670	0,219	0,002	0,007
smo2	0,488	0,200	0,001	0,006
soc2	0,297	0,237	0,002	0,008
soc3	0,308	0,244	0,002	0,007

Tabela 16 – Estimativas para o conjunto de dados `Birth` - priori t-student

	Estimativa	Erro Padrão
Intecepto	-2,752	0,201
alc2	0,166	0,271
alc3	0,669	0,214
smo2	0,497	0,201
soc2	0,284	0,231
soc3	0,291	0,241

4.2.7.2 Conjuntos de dados `Heart` e `PCS`

Mesmo em um conjunto de dados maior como o `Heart` com 16.949 registros, o resultado do modelo bayesiano com priori uniforme divergiu já a partir da segunda casa decimal, como pode ser verificado na Tabela 17. O mesmo ocorreu com o conjunto de dados `PCS` que apresenta 380 registros dos quais três foram excluídos por apresentar valores nulos, como pode ser confirmado pela Tabela 18.

Tabela 17 – Estimativas para o conjunto de dados `Heart` - priori uniforme

	Estimativa	Desvio Padrão	Erro Padrão	Erro Padrão de Monte Carlo
Intercepto	-4,064	0,089	4,5e-04	2,4e-03
age2	1,105	0,091	4,5e-04	2,5e-03
age3	1,957	0,094	4,7e-04	2,6e-03
severity2	0,693	0,069	3,5e-04	2,0e-03
severity3	1,489	0,083	4,2e-04	2,4e-03
region2	0,070	0,177	8,9e-04	4,8e-03
region3	0,726	0,092	4,6e-04	2,9e-03
onset2	0,067	0,066	3,3e-04	1,8e-03
onset3	0,186	0,076	3,8e-04	2,0e-03

Tabela 18 – Estimativas para o conjunto de dados `PCS` - priori uniforme

	Estimativa	Desvio Padrão	Erro Padrão	Erro Padrão de Monte Carlo
Intercepto	-2,946	0,209	1,0e-03	3,7e-03
age	-0,006	0,001	4,5e-06	8,6e-05
race2	-0,007	0,044	2,2e-04	2,4e-02
dpros2	0,564	0,209	1,0e-03	3,7e-03
dpros3	0,682	0,211	1,1e-03	4,1e-03
dpros4	0,574	0,210	1,1e-03	3,7e-03
dcaps2	0,030	0,030	1,5e-04	1,5e-02
psa	0,002	0,001	4,4e-06	7,8e-05
vol	-0,007	0,003	1,4e-05	3,8e-04
gleason	0,294	0,005	2,6e-05	1,9e-03

Nos conjunto de dados `Heart` e `PCS`, o algoritmo da função `bayesglm` não convergiu mesmo fornecendo valores inicias próximos ao EMV (estimados por meio do pacote `lbreg`). Para o ajuste do modelo com priori uniforme por Metropolis-Hasting foi possível inserir as restrições do espaço paramétrico juntamente com a função de verossimilhança. Uma

inspeção mais detalhada da função `bayesglm` se faz necessária para compreender como inserir restrições em seu algoritmo.

5 Considerações finais

O modelo log-binomial estima o risco relativo e o modelo logístico estima a razão de chances. O risco relativo é uma medida de associação mais naturalmente assimilada do que a razão de chances, entretanto, as dificuldades computacionais impediram por muito tempo que esse modelo fosse amplamente utilizado.

A dificuldade no ajuste do modelo log-binomial situa-se na necessidade de impor uma restrição no seu espaço paramétrico. Mesmo oferecendo valores iniciais dentro do espaço paramétrico, o algoritmo de otimização pode acabar saindo do espaço paramétrico e não ser capaz de retornar ou ficar oscilando em valores próximos à fronteira se tornando lento e não alcançar convergência. Sobre esse tema, destaca-se na literatura a contribuição de Andrade (2018), que desenvolveu o pacote `lbreg` para ajuste do modelo log-binomial por programação não-linear. Apesar de outros autores também terem implementado algoritmos para lidar com as restrições nos parâmetros, somente Andrade e Andrade (2017) apresentam uma solução para os casos em que o EMV está na fronteira. Utilizando-se desse estudo é possível adaptar outros algoritmos de otimização para o ajuste do modelo log-binomial.

Portanto, atualmente, já existem soluções computacionais e metodológicas para lidar com as restrições do modelo log-binomial e não existe mais a necessidade deste ser preterido diante do modelo logístico.

A solução das questões de convergência no ajuste do modelo log-binomial, nos permitiu aplicá-lo de forma iterativa em diversas simulações de cenários e reamostragens de dados.

Para verificar o melhor ajuste entre o modelo de regressão log-binomial e o modelo de regressão logística foram aplicados os testes de COX e de Vuong. Diante dos resultados das simulações, os testes parecem ter pouco poder o que abre caminho para pesquisa de outras metodologias e testes de hipótese para escolha entre esses dois modelos. A falta de poder dos testes pode estar relacionada ao fato da resposta ser binária, assim, ainda existe ampla possibilidade de simulações com outros modelos de regressão utilizando a função `cv.test` constante no Anexo C.2.

A comparação entre os resultados dos testes de Cox e de Vuong e as medidas de diagnóstico de acurácia nos mostraram que um ajuste melhor aos dados, não necessariamente trará um grande ganho na capacidade de predição do modelo. Além disso, se o intuito é somente de predição, uma mudança no ponto de corte tem uma influência muito maior na especificidade e na sensibilidade do que uma mudança de modelo.

Com a popularização dos algoritmos de aprendizado de máquina, ficou clara a necessidade de se distinguir modelos preditivos de modelos puramente inferenciais. A aplicação de aprendizado de máquina, em geral, é feita em situações em que a única preocupação é a predição, uma vez que as relações entre as variáveis pode se tornar tão complexa que fica impossível interpretar como cada uma influencia a resposta.

Por outro lado, existem situações em que interpretar os parâmetros é tão ou mais importante que a predição. É neste contexto que o modelo log-binomial ganha grande importância.

A abordagem bayesiana do modelo log-binomial ainda é pouco discutida na literatura. Chu e Cole (2010) foram os primeiros a abordar o tema e utilizaram uma distribuição a priori normal $\beta_j \sim N(0, 10^2)$. Zhou, Sivaganesan e Longla (2014) propõem uma priori própria não informativa. Torman e Camey (2015) estendem o uso da abordagem bayesiana do modelo log-binomial para respostas politônicas. Salmerón, Cano e Chirlaque (2015), propõem uma reparametrização e um amostrador de Gibss específico que possibilitam a redução de erro de Monte Carlo. Janani et al. (2017) fazem uma comparação entre os métodos frequentista e bayesiano para ajuste do modelo log-binomial. Como foi visto na seção 4.2.7, mesmo na abordagem bayesiana, algoritmos desenvolvidos para MLGs nem sempre são capazes de convergir no caso do modelo log-binomial, o qual não utiliza a função de ligação canônica. Ainda existe um vasto campo de pesquisa no desenvolvimento teórico e implementação computacional na abordagem bayesiana do modelo log-binomial.

Alguns dos códigos em `R` que foram utilizados neste trabalho estão no Apêndice C. Não foram inseridos todos os códigos devido à extensão, mas a partir dos exemplos disponíveis é possível fazer adaptações para reproduzir os demais resultado.

Referências

AGRESTI, A. *Foundations of linear and generalized linear models*. [S.l.]: John Wiley & Sons, 2015. Citado na página 21.

ANDRADE, B. B. de. *Log-Binomial Regression with Constrained Optimization*. [S.l.], 2018. Disponível em: <https://CRAN.R-project.org/package=lbreg>. Citado 3 vezes nas páginas 17, 28 e 57.

ANDRADE, B. B. de; ANDRADE, J. M. de L. Some results for maximum likelihood estimation of adjusted relative risks. *Communications in Statistics - Theory and Methods*, Taylor & Francis, p. 1–20, 2017. Citado 6 vezes nas páginas 17, 18, 19, 28, 30 e 57.

BEDRICK, E. J.; CHRISTENSEN, R.; JOHNSON, W. A new perspective on priors for generalized linear models. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 91, n. 436, p. 1450–1460, 1996. Citado 3 vezes nas páginas 51, 53 e 54.

BEDRICK, E. J.; CHRISTENSEN, R.; JOHNSON, W. Bayesian binomial regression: Predicting survival at a trauma center. *The American Statistician*, Taylor & Francis, v. 51, n. 3, p. 211–218, 1997. Citado 2 vezes nas páginas 53 e 54.

BERGER, J. O.; BERNARDO, J. M. On the development of the reference prior method. *Bayesian statistics*, v. 4, p. 35–60, 1992. Citado na página 54.

BLIZZARD, L.; HOSMER, W. Parameter estimation and goodness-of-fit in log binomial regression. *Biometrical Journal*, Wiley Online Library, v. 48, n. 1, p. 5–22, 2006. Citado na página 39.

CAMERON, A. C.; TRIVEDI, P. K. *Microeconometrics: methods and applications*. [S.l.]: Cambridge university press, 2005. Citado na página 38.

CHAO, C. et al. Correlates for human papillomavirus vaccination of adolescent girls and young women in a managed care organization. *American journal of epidemiology*, Oxford University Press, v. 171, n. 3, p. 357–367, 2010. Citado na página 10.

CHATZI, L. et al. Mediterranean diet adherence during pregnancy and fetal growth: Inma (spain) and rhea (greece) mother–child cohort studies. *British Journal of Nutrition*, Cambridge University Press, v. 107, n. 1, p. 135–145, 2012. Citado na página 10.

CHU, H.; COLE, S. R. Estimation of risk ratios in cohort studies with common outcomes a bayesian approach. *Epidemiology*, LWW, v. 21, n. 6, p. 855–862, 2010. Citado 3 vezes nas páginas 48, 54 e 58.

DATTA, G. S.; GHOSH, M. On the invariance of noninformative priors. *The Annals of Statistics*, Institute of Mathematical Statistics, v. 24, n. 1, p. 141–159, 1996. Citado na página 49.

DATTA, G. S.; MUKERJEE, R. Matching priors for prediction. In: *Probability Matching Priors: Higher Order Asymptotics*. [S.l.]: Springer, 2004. p. 99–115. Citado na página 54.

DEDDENS, J. A.; PETERSEN, M. R.; LEI, X. Estimation of prevalence ratios when proc genmod does not converge. In: SAS INSTITUTE INC. CARY, NC. *Proceedings of the 28th annual SAS users group international conference*. [S.l.], 2003. v. 30, p. 270–28. Citado na página 34.

DEGROOT, M. H. *Optimal statistical decisions*. [S.l.]: John Wiley & Sons, 2005. v. 82. Citado na página 47.

DONOGHOE, M. W.; MARSCHNER, I. C. Fast stable relative risk regression using an overparameterised em algorithm. In: DUPUY, J.-F.; JOSSE, J. (Ed.). *Proceedings of the 31st International Workshop on Statistical Modelling*. [S.l.]: Statistical Modelling Society, 2016. v. 1, p. 93–98. Citado na página 26.

DONOGHOE, M. W.; MARSCHNER, I. C. logbin: An r package for relative risk regression using the log-binomial model. *Journal of Statistical Software*, v. 86, p. 1–22, 2018. Citado na página 26.

FIRTH, D. Bias reduction of maximum likelihood estimates. *Biometrika*, Oxford University Press, v. 80, n. 1, p. 27–38, 1993. Citado na página 19.

GELMAN, A. et al. *Bayesian data analysis*. [S.l.]: CRC press, 2014. Citado 2 vezes nas páginas 46 e 64.

GELMAN, A. et al. A weakly informative default prior distribution for logistic and other regression models. *The Annals of Applied Statistics*, Institute of Mathematical Statistics, v. 2, n. 4, p. 1360–1383, 2008. Citado 3 vezes nas páginas 51, 52 e 54.

GENKIN, A.; LEWIS, D. D.; MADIGAN, D. Large-scale bayesian logistic regression for text categorization. *Technometrics*, Taylor & Francis, v. 49, n. 3, p. 291–304, 2007. Citado na página 54.

GLOVER, M. et al. Driving kids to smoke? children's reported exposure to smoke in cars and early smoking initiation. *Addictive behaviors*, Elsevier, v. 36, n. 11, p. 1027–1031, 2011. Citado na página 10.

HANSON, T. E. et al. Informative g-priors for logistic regression. *Bayesian Analysis*, International Society for Bayesian Analysis, v. 9, n. 3, p. 597–612, 2014. Citado na página 54.

HEINZE, G.; SCHEMPER, M. A solution to the problem of separation in logistic regression. *Statistics in medicine*, Wiley Online Library, v. 21, n. 16, p. 2409–2419, 2002. Citado na página 19.

IBRAHIM, J. G.; LAUD, P. W. On bayesian analysis of generalized linear models using jeffreys's prior. *Journal of the American Statistical Association*, Taylor & Francis, v. 86, n. 416, p. 981–986, 1991. Citado 2 vezes nas páginas 50 e 54.

JANANI, L. et al. Comparison between bayesian approach and frequentist methods for estimating relative risk in randomized controlled trials: a simulation study. *Journal of Statistical Computation and Simulation*, Taylor & Francis, v. 87, n. 4, p. 640–651, 2017. Citado 2 vezes nas páginas 48 e 58.

JEFFREYS, H. An invariant form for the prior probability in estimation problems. *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, JSTOR, p. 453–461, 1946. Citado 2 vezes nas páginas 50 e 54.

JUN, H.-J.; ACEVEDO-GARCIA, D. The effect of single motherhood on smoking by socioeconomic status and race/ethnicity. *Social science & medicine*, Elsevier, v. 65, n. 4, p. 653–666, 2007. Citado na página 10.

KASS, R. E.; WASSERMAN, L. The selection of prior distributions by formal rules. *Journal of the American Statistical Association*, Taylor & Francis Group, v. 91, n. 435, p. 1343–1370, 1996. Citado 2 vezes nas páginas 49 e 50.

LANGE, K. An adaptive barrier method for convex programming. *Methods and Applications of Analysis*, International Press of Boston, v. 1, n. 4, p. 392–402, 1994. Citado na página 27.

LI, C. et al. Prevalence of pre-diabetes and its association with clustering of cardiometabolic risk factors and hyperinsulinemia among us adolescents: National health and nutrition examination survey 2005–2006. *Diabetes care*, Am Diabetes Assoc, v. 32, n. 2, p. 342–347, 2009. Citado na página 10.

LIU, X. et al. Exposure to bisphenol-a and reproductive hormones among male adults. *Environmental toxicology and pharmacology*, Elsevier, v. 39, n. 2, p. 934–941, 2015. Citado na página 10.

MARIN, J.-M.; ROBERT, C. P. *Bayesian essentials with R*. [S.l.]: Springer, 2014. v. 48. Citado 2 vezes nas páginas 51 e 54.

MARSCHNER, I. C. glm2: Fitting generalized linear models with convergence problems. *The R Journal*, v. 3, p. 12–15, 2011. Citado na página 29.

MARSCHNER, I. C. Relative risk regression for binary outcomes: Methods and recommendations. *Australian & New Zealand Journal of Statistics*, Wiley Online Library, v. 57, n. 4, p. 437–462, 2015. Citado 4 vezes nas páginas 18, 21, 28 e 29.

MARSCHNER, I. C.; GILLETT, A. C. Relative risk regression: reliable and flexible methods for log-binomial models. *Biostatistics*, Oxford University Press, v. 13, n. 1, p. 179–192, 2011. Citado 2 vezes nas páginas 23 e 24.

MARTIN, A. D.; QUINN, K. M.; PARK, J. H. Mcmcpack: Markov chain monte carlo in r. Foundation for Open Access Statistics, 2011. Citado na página 53.

PEDROZA, C. et al. Performance of models for estimating absolute risk difference in multicenter trials with binary outcome. *BMC medical research methodology*, BioMed Central, v. 16, n. 1, p. 113, 2016. Citado na página 48.

PESARAN, M. H.; PESARAN, B. A simulation approach to the problem of computing cox's statistic for testing nonnested models. *Journal of Econometrics*, Elsevier, v. 57, n. 1-3, p. 377–392, 1993. Citado 2 vezes nas páginas 36 e 37.

ROBIN, X. et al. proc: an open-source package for r and s+ to analyze and compare roc curves. *BMC bioinformatics*, BioMed Central, v. 12, n. 1, p. 77, 2011. Citado na página 43.

SALMERÓN, D.; CANO, J. A.; CHIRLAQUE, M. D. Reducing monte carlo error in the bayesian estimation of risk ratios using log-binomial regression models. *Statistics in medicine*, Wiley Online Library, v. 34, n. 19, p. 2755–2767, 2015. Citado 2 vezes nas páginas 48 e 58.

SANTOS, C. A. S. et al. Estimating adjusted prevalence ratio in clustered cross-sectional epidemiological data. *BMC Medical Research Methodology*, BioMed Central, v. 8, n. 1, p. 80, 2008. Citado na página 34.

SAVU, A.; LIU, Q.; YASUI, Y. Estimation of relative risk and prevalence ratio. *Statistics in medicine*, Wiley Online Library, v. 29, n. 22, p. 2269–2281, 2010. Citado na página 39.

SCHOUTEN, E. G. et al. Risk ratio and rate ratio estimation in case-cohort designs: Hypertension and cardiovascular mortality. *Statistics in medicine*, Wiley Online Library, v. 12, n. 18, p. 1733–1745, 1993. Citado na página 34.

TORMAN, V. B. L.; CAMEY, S. A. Bayesian models as a unified approach to estimate relative risk (or prevalence ratio) in binary and polytomous outcomes. *Emerging themes in epidemiology*, BioMed Central, v. 12, n. 1, p. 8, 2015. Citado 2 vezes nas páginas 48 e 58.

WILLIAMSON, M.; GASTON, K. J. The lognormal distribution is not an appropriate null hypothesis for the species–abundance distribution. *Journal of Animal Ecology*, Wiley Online Library, v. 74, n. 3, p. 409–422, 2005. Citado na página 10.

ZELLNER, A. Applications of bayesian analysis in econometrics. *Journal of the Royal Statistical Society. Series D (The Statistician)*, JSTOR, v. 32, n. 1/2, p. 23–34, 1983. Citado na página 54.

ZHOU, R.; SIVAGANESAN, S.; LONGLA, M. An objective bayesian estimation of parameters in a log-binomial model. *Journal of Statistical Planning and Inference*, Elsevier, v. 146, p. 113–121, 2014. Citado 4 vezes nas páginas 48, 49, 54 e 58.

ZOU, G. A modified poisson regression approach to prospective studies with binary data. *American journal of epidemiology*, Oxford University Press, v. 159, n. 7, p. 702–706, 2004. Citado na página 34.

APÊNDICE A – Algoritmo EM

Para entender o algoritmo EM, vamos rever alguma relações de probabilidade condicional, nos atendo ao caso de variáveis discretas que é o interesse deste trabalho.

- Probabilidade condicional
$$p(x|z) = \frac{p(x,z)}{p(z)}; \tag{A.1}$$

- Probabilidade conjunta
$$p(x,z) = p(z|x)p(x) = p(x|z)p(z);$$

- Probabilidade marginal
$$p(x) = \sum_Z p(x,z) = \sum_Z p(x|z)p(z) = \sum_Z p(z|x)p(x); \tag{A.2}$$

- Probabilidade marginal como esperança
$$p(x) = \sum_Z p(x|z)p(z) = E_Z[p(x|z)];$$

- Função de verossimilhança como probabilidade condicional
$$L(\mathbf{x};\boldsymbol{\beta}) = p(\mathbf{x}|\boldsymbol{\beta}) = \sum_Z p(\mathbf{x},\mathbf{z}|\boldsymbol{\beta}). \tag{A.3}$$

O algoritmo EM utiliza uma sequencia de problemas de maximização mais fáceis que levam a resposta do problema original, se utilizando do fato de que a distribuição marginal de uma variável aleatória X equivale a soma da distribuição conjunta de X e Z para todos os valores de Z, conforme (A.2). Assim pode ser incorporada uma variável latente Z, que simplifique o problema de maximização. Com a inclusão de Z o modelo passa ser chamado de modelo completo, ou ainda, como Z é uma variável auxiliar e não observável, também é chamado de *missing data model*.

A variável latente tem distribuição $Z \sim p(\mathbf{z}|\mathbf{x};\boldsymbol{\beta})$, $\mathbf{z} = z_1, ..., z_m$ são chamados de dados aumentados, e a verossimilhança do modelo completo é denotada por $L(\mathbf{z},\mathbf{x};\boldsymbol{\beta})$.

A verossimilhança original e a verossimilhança do modelo completo apresentam a seguinte relação:

$$L(\mathbf{x};\boldsymbol{\beta}) = E_Z[L(\mathbf{z},\mathbf{x};\boldsymbol{\beta})] = \sum_Z L(\mathbf{z},\mathbf{x};\boldsymbol{\beta})p(\mathbf{z}|\mathbf{x};\boldsymbol{\beta}).$$

Mas como usualmente em inferência, é preferível trabalhar com a log-verossimilhança, assim, utilizando (A.3), (A.1) e aplicando log, temos que

$$\log L(\mathbf{x};\boldsymbol{\beta}) = \log p(\mathbf{x}|\boldsymbol{\beta}) = \log p(\mathbf{z},\mathbf{x}|\boldsymbol{\beta}) - \log p(\mathbf{z}|\mathbf{x};\boldsymbol{\beta}).$$

Como o lado esquerdo da equação não depende de $\mathbf{z}$, se tirarmos a esperança de ambos os lados,

$$\log L(\mathbf{x};\boldsymbol{\beta}) = E_Z[\log p(\mathbf{z},\mathbf{x}|\boldsymbol{\beta})] - E_Z[\log p(\mathbf{z}|\mathbf{x};\boldsymbol{\beta})] =$$

$$E_Z[\log L(\mathbf{z},\mathbf{x};\boldsymbol{\beta})] - E_Z[\log p(\mathbf{z}|\mathbf{x};\boldsymbol{\beta})].$$

O termo $E_Z[\log p(\mathbf{z}|\mathbf{x};\boldsymbol{\beta})]$ não é necessário para a maximização de $\log L(\mathbf{x};\boldsymbol{\beta})$ essa demonstração pode ser feita pelo Teorema de Jensen (GELMAN et al., 2014). O termo $E_Z[\log L(\mathbf{z},\mathbf{x};\boldsymbol{\beta})]$ é, então, denotado por $\mathcal{Q}(\boldsymbol{\beta}|\boldsymbol{\beta}^{(0)})$ e pode ser visto como a esperança da log-verossimilhança dos dados completos.

O algoritmo EM segue uma sequencia iterativa de dois passos, um de cálculo de esperança (E) e outro de maximização (M).

1. Atribuir um valor inicial para $\hat{\boldsymbol{\beta}}^{(t)} = \hat{\boldsymbol{\beta}}^{(0)}$.
2. Passo E: Calcular a esperança

$$\mathcal{Q}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(t)}) = E_Z[\log L(\mathbf{z},\mathbf{x};\boldsymbol{\beta})].$$

3. Passo M: Maximizar $\mathcal{Q}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(t)})$ em $\boldsymbol{\beta}$

$$\hat{\boldsymbol{\beta}}^{(t+1)} = \underset{\boldsymbol{\beta}}{\operatorname{argmax}}\ \mathcal{Q}(\boldsymbol{\beta}|\hat{\boldsymbol{\beta}}^{(t)}).$$

4. Repetir 2 e 3 até um ponto estacionário, isto é, $\boldsymbol{\beta}^{(t+1)} = \boldsymbol{\beta}^{(t)}$.

ANEXO A – Comparação das simulações por AIC

Tabela 19 – Quantidade de simulações em que o modelo apresentou maior AIC

Cenários	PGD	log	logit	probit	cloglog
BH1	log	86	1	1	12
BH2	log	48	14	27	11
BH3	log	77	2	0	21
BH4	log	57	11	19	13
BH5	log	76	3	4	17
BH6	log	45	14	29	12
BH7	log	76	0	1	23
BH8	log	50	9	31	10
BH9	log	69	3	8	20
BH10	log	74	2	2	22
BH11	log	64	11	7	18
BH12	log	74	1	8	17
BH1-logit	logit	28	29	0	43
BH2-logit	logit	48	52	0	0
BH3-logit	logit	28	55	0	17
BH4-logit	logit	51	49	0	0
BH5-logit	logit	25	46	1	28
BH6-logit	logit	38	62	0	0
BH7-logit	logit	28	39	0	33
BH8-logit	logit	44	56	0	0
BH9-logit	logit	53	47	0	0
BH10-logit	logit	39	57	0	4
BH11-logit	logit	42	58	0	0
BH12-logit	logit	39	45	0	16
S1	log	44	9	40	7
S2	cloglog	39	10	44	7
S3	logit	47	7	42	4
S4	probit	15	9	72	4
S5	log	6	41	46	7
S6	cloglog	5	28	62	5
S7	logit	2	25	69	4
S8	probit	0	4	96	0

ANEXO B – Testes de COX e VUONG em cenários simulados

Tabela 20 – Teste de COX - log-binomial vs logística, teste de COX - logística vs log-binomial, teste de COX combinado, teste de VUONG e menor AIC de 100 simulações dos cenários BH1 a BH12 e S1 a S8

Cenário	Média de $\hat{p}$	COX1 - log vs logit		COX2 - logit vs log		COX - combinado				VUONG			Menor AIC	
		Rejeita log	Não Rejeita log	Rejeita logit	Não Rejeita logit	Rejeita logit	Rejeita log	Rejeita ambas	Não rejeita ambas	Rejeita logit	Rejeita log	Não há diferença	log	logit
BH1	0.215	2	98	82	18	80	0	2	18	38	0	62	94	6
BH2	0.121	4	96	4	96	2	2	2	94	1	0	99	57	43
BH3	0.207	4	96	50	50	46	0	4	50	21	0	79	90	10
BH4	0.131	3	97	1	99	1	3	0	96	0	3	97	62	38
BH5	0.208	9	91	58	42	57	8	1	34	25	0	75	87	13
BH6	0.124	3	97	2	98	2	3	0	95	1	0	99	52	48
BH7	0.216	6	94	80	20	75	1	5	19	39	0	61	91	9
BH8	0.122	3	97	4	96	3	2	1	94	2	2	96	55	45
BH9	0.277	10	90	37	63	29	2	8	61	15	0	85	78	22
BH10	0.360	4	96	46	54	42	0	4	54	16	0	84	86	14
BH11	0.327	5	95	21	79	17	1	4	78	7	0	93	70	30
BH12	0.434	10	90	30	70	23	3	7	67	15	0	85	83	17
S1	0.067	37	63	20	80	5	22	15	58	0	0	100	47	53
S2	0.077	32	68	12	88	5	25	7	63	0	0	100	44	56
S3	0.090	18	82	5	95	2	15	3	80	0	0	100	53	47
S4	0.023	98	2	90	10	0	8	90	2	0	0	100	20	80
S5	0.168	30	70	3	97	0	27	3	70	0	11	89	9	91
S6	0.185	44	56	3	97	0	41	3	56	0	19	81	5	95
S7	0.200	42	58	8	92	0	34	8	58	0	23	77	4	96
S8	0.105	97	3	28	72	0	69	28	3	0	82	18	0	100

Tabela 21 – Teste de COX - log-binomial vs logística, teste de COX - logística vs log-binomial, teste de COX combinado, teste de VUONG e menor AIC de 100 simulações dos cenários BH1-logit a BH12-logit

Cenário	Média de $\hat{p}$	COX1 - log vs logit		COX2 - logit vs log		COX - combinado				VUONG			Menor AIC	
		Rejeita log	Não Rejeita log	Rejeita logit	Não Rejeita logit	Rejeita logit	Rejeita log	Rejeita ambas	Não rejeita ambas	Rejeita logit	Rejeita log	Não há diferença	log	logit
BH1-logit	0.150	21	79	7	93	2	16	5	77	0	7	93	37	63
BH2-logit	0.094	4	96	3	97	0	1	3	96	0	0	100	48	52
BH3-logit	0.142	18	82	7	93	1	12	6	81	0	9	91	29	71
BH4-logit	0.105	8	92	2	98	0	6	2	92	0	5	95	51	49
BH5-logit	0.146	13	87	10	90	1	4	9	86	0	11	89	28	72
BH6-logit	0.101	9	91	3	97	0	6	3	91	0	4	96	38	62
BH7-logit	0.148	9	91	3	97	1	7	2	90	1	4	95	34	66
BH8-logit	0.100	11	89	3	97	0	8	3	89	0	4	96	44	56
BH9-logit	0.202	7	93	6	94	1	2	5	92	0	0	100	53	47
BH10-logit	0.246	6	94	3	97	0	3	3	94	0	1	99	42	58
BH11-logit	0.235	4	96	4	96	1	1	3	95	0	0	100	42	58
BH12-logit	0.289	4	96	2	98	0	2	2	96	0	0	100	48	52

ANEXO C – Códigos em R

C.1 Códigos do Capítulo 2

```
data(Heart)
#Valores iniciais
mod.poi<- glm(Heart ~ age + severity + region + onset
              ,data = Heart
              ,family = poisson(link = 'log'))
b0 <- mod.poi$coefficients
x <- model.matrix(mod.poi)
mX <- as.matrix(-x[, -1])
b0[1] <- min(mX %*% b0[-1]) -.1*abs(min(mX %*% b0[-1]))

Heart.glm <- glm (Heart ~ age + severity + region + onset
                  ,family=binomial(link='log')
                  ,data = Heart
                  ,start=b0
                  ,maxit=100000)

Heart.glm2 <- glm2 (Heart ~ age + severity + region + onset
                   ,family=binomial(link='log')
                   ,data = Heart
                   ,start=b0)

Heart.logbin.cem <- logbin (Heart ~ age + severity + region + onset
                            ,data = Heart
                            ,method='em')

Heart.lbreg <- lbreg(Heart ~ age + severity + region + onset
                     ,data=Heart)
```

C.2 Códigos do Capítulo 3

```
cv.test <- function(null, alt, nsim, alpha=0.05 ){
    mf <- null
    mg <- alt
    if(class(null)[1]=="lbreg"){
        ff <- mf$formula
        lf <- mf$loglik
        modframe <- model.frame(formula=ff)
        terms <- attr(modframe, "terms")
        X <- model.matrix(terms, modframe)
```

```
    Y <- as.matrix(  model.response(modframe) )
    mf$family <- binomial(link="log")
    dff <- mf$df
}else{
    X <- model.matrix(mf)
    Y <- as.matrix( model.response(model.frame(mf)) )
    lf <- logLik(mf)
    dff <- mf$df.resid
}
if(class(alt)[1]=="lbreg"){
    lg <- mg$loglik
    mg$family <- binomial(link="log")
    dfg <- mg$df
}else{
    X <- model.matrix(mg)
    Y <- as.matrix( model.response(model.frame(mg)) )
    lg <- logLik(mg)
    dfg <- mg$df.resid
}
pf <- fitted(mf)
pg <- fitted(mg)
d <- Y*(log(pf)-log(pg)) + (1-Y)*(log(1-pf)-log(1-pg))
eta <- X%*%coef(mf)
ef <- switch(mf$family$link,
    logit = Y-pf,
    log = (Y-pf)/(1-pf),
    probit = dnorm(eta)*(Y-pf)/(pf*(1-pf)),
    cloglog = -exp(-eta)*(Y-pf)/(1-pf),
    cauchit =  dcauchy(eta)*(Y-pf)/(pf*(1-pf)) )
Z <- sweep(as.matrix(X[,-1]), 1, ef, "*")
aux <- lm.fit(y=d, x=cbind(1,Z))
v <- sqrt( mean(resid(aux)^2) )
LRdif <- mean(d)

# MC estimate of gama under f
gama <- matrix(NA, nrow=nsim, ncol=length(coef(mf)))
N <- nrow(X)
if(mg$family$link=="log"){
    for(i in 1:nsim){
        ystar <- as.matrix(rbinom(N, size = 1, prob = pf)
        b0 <- coef( mg )
        mX <- as.matrix(-X[,-1])
        b0[1] <-  min( mX %*% b0[-1] ) - 1
        gama[i,] <- coef(lbreg:::lbreg.fit(y=ystar, x=X, start.beta=b0, tol=.9999, delta=1))
    }
}else{
    for(i in 1:nsim){
```

```
            ystar <- rbinom(N, size = 1, prob = pf )
            gama[i,] <- coef(glm.fit(y=ystar,x=X,family=mg$family))
        }
    }
    ghat <- colMeans(gama)
    pg.star <- mg$family$linkinv( X%*%ghat )
    KL <- mean(pf*(log(pf)-log(pg.star)) + (1-pf)*(log(1-pf)-log(1-pg.star)))
    cox.stat <- sqrt(N)*(LRdif - KL)/v
    LR <- c(lf, lg)/N
    df.resid <- c(dff,dfg)
    names(LR) <- names(df.resid) <- c(paste("null.",mf$family$link,sep="")
                                      ,paste("alt.",mg$family$link,sep=""))
    vuong.stat <- sqrt(N)*LRdif/sd(d)
    return(list(LR=LR, df.resid=df.resid,
                KL=KL, cox=cox.stat,
                vuong = vuong.stat,
                zcrit=qnorm(1-alpha/2)) )
}

sim_S <- function (scenario       =NULL
                   ,nsim_scenario  =10
                   ,n              =500
                   ,nsim_gamma     =200) {

  sim <-  data.frame(iteration      = integer(nsim_scenario)
                     ,iteration_seed = integer(nsim_scenario)
                     ,mean_y         = integer(nsim_scenario)
                     ,LR_log         = numeric(nsim_scenario)
                     ,LR_logit       = numeric(nsim_scenario)
                     ,COX1           = numeric(nsim_scenario)
                     ,COX2           = numeric(nsim_scenario)
                     ,VUONG          = numeric(nsim_scenario))

  seed_ini <- sample(1:2^15, 1)
  set.seed(seed_ini)
  n_tot= n*nsim_scenario
  x1  <- rbinom(n_tot, size=1, prob=0.5 )
  x2  <- rmultinom(n_tot, size=1, prob=c(0.3, 0.3, 0.4))
  x22 <- x2[2,]
  x23 <- x2[3,]
  x3  <- runif (n_tot, min=-1, max=2)
  x   <- cbind(x1,x22,x23,x3)
  pE <- exp(-1+x1-x22+x23+x3)/(1+exp(-1+x1-x22+x23+x3))
  E  <- rbinom(n_tot, size=1, prob=pE)
  gamma0 <- c(-2.1, -1.9, -1.7, -1.48, -1.1, -0.9, -0.7, -0.48)
  if (scenario %in% 1:4) {h = quote(-(x1+x22+x23+x3))}
    else if (scenario %in% 5:8) {h = quote(-pmax(x1+x22+x23+x3,0))}
```

```
  if (scenario %in% c(1,5)) {g = quote(exp(gamma0[scenario] + eval(h)))}
    else if (scenario %in% c(2,6)) {g = quote(1-exp(-exp(gamma0[scenario] + eval(h))))}
    else if (scenario %in% c(3,7)) {g = quote(exp(gamma0[scenario] + eval(h))
                                              /(1+exp(gamma0[scenario] + eval(h))))}
    else if (scenario %in% c(4,8)) {g = quote(pnorm(gamma0[scenario] + eval(h)))}
  py <- ifelse(E == 0, eval(g)
       ,ifelse(E == 1, 3*eval(g)
       ,NA))
  py <- ifelse(py>1,1,py)
  y  <- rbinom(n_tot, size=1, prob=py)
  sim$iteration_seed <- sample(1:2^15, nsim_scenario)
  for(i in 1:nsim_scenario) {
    ini=(n*i)-(n-1)
    end=(n*i)
    m_lbreg <- lbreg(y[ini:end] ~ x[ini:end,])
    m_logistic <- glm(y[ini:end] ~ x[ini:end,], family=binomial)
    set.seed(sim$iteration_seed[i])
    t1 <- cv.test(m_lbreg   , m_logistic, nsim= nsim_gamma)
    t2 <- cv.test(m_logistic, m_lbreg   , nsim= nsim_gamma)
    sim$iteration[i]      = i
    sim$mean_y[i]         = mean(y[ini:end])
    sim$LR_log[i]         = t1$LR[1]
    sim$LR_logit[i]       = t1$LR[2]
    sim$COX1[i]           = t1$cox
    sim$COX2[i]           = t2$cox
    sim$VUONG[i]          = t1$vuong
  }

  Result_COX <-  ifelse(abs(sim[,"COX1"])< sim[,"zcrit"] & abs(sim[,"COX2"])>=sim[,"zcrit"], 1
                ,ifelse(abs(sim[,"COX1"])>=sim[,"zcrit"] & abs(sim[,"COX2"])< sim[,"zcrit"], 2
                ,ifelse(abs(sim[,"COX1"])>=sim[,"zcrit"] & abs(sim[,"COX2"])>=sim[,"zcrit"], 3
                ,ifelse(abs(sim[,"COX1"])< sim[,"zcrit"] & abs(sim[,"COX2"])< sim[,"zcrit"], 4
                ,NA))))

  Result_VUONG <-  ifelse(    sim[,"VUONG"] >  sim[,"zcrit"], 1
                  ,ifelse(    sim[,"VUONG"] < -sim[,"zcrit"], 2
                  ,ifelse(abs(sim[,"VUONG"])<  sim[,"zcrit"], 3
                  ,NA)))

  sim <- cbind(sim, Result_COX, Result_VUONG)

  return( list(seed_ini = seed_ini, simulation=sim, y=y, x=x))
}

S1 <- sim_S(scenario=1, nsim_scenario=100, n=500, nsim_gamma=200)

reamostragem <- function(tam_dados= 100
```

```
                        ,tam_amostra=100
                        ,n_sim=100
                        ,formula= NULL
                        ,data=NULL){

  result <- data.frame(iteration         = integer(n_sim)
                       ,aic_log          = numeric(n_sim)
                       ,aic_logit        = numeric(n_sim)
                       ,stringsAsFactors =FALSE)

  seed <- sample(1:2^15, 1)
  set.seed(seed)
  size_tot = tam_amostra*n_sim
  linhas=sample.int(n = tam_dados, size = size_tot , replace = TRUE)

  for(i in 1: n_sim) {
    result$iteration[i] = i
    ini=(tam_amostra*i)-(tam_amostra-1)
    end=(tam_amostra*i)
    dados=data[linhas[ini:end],]
    m_log <- lbreg (formula
                    ,data = dados)
    result$aic_log[i] = -2 * m_log$loglik + 2*length(m_log$coefficients)
    m_logit <- glm (formula
                    ,data = dados
                    ,family=binomial)
    result$aic_logit[i] = m_logit$aic
  }
  return( list(seed = seed, result = result))
}

boot.Birth <- reamostragem (tam_dados = 900
                            ,tam_amostra =900
                            ,n_sim = 1000
                            ,formula = lowbw ~ alc + smo + soc
                            ,data = Birth)
```

C.3 Códigos do Capítulo 4

```
attach(Birth)
yvector = lowbw
x1 = alc
x11=ifelse(x1==1,1,0)
x12=ifelse(x1==2,1,0)
x13=ifelse(x1==3,1,0)
x2 = smo
```

```
x21=ifelse(x2==1,1,0)
x22=ifelse(x2==2,1,0)
x3 = soc
x31=ifelse(x3==1,1,0)
x32=ifelse(x3==2,1,0)
x33=ifelse(x3==3,1,0)
Xdata = cbind (1,x12,x13,x22,x32,x33)
detach(Birth)

logfun <- function(beta, y, X){
  eta <- X %*% beta
  p <- ifelse(eta<=0, exp(eta), 1e-8)
  sum( y * log(p) + (1-y)*log(1-p) )
}

post.samp.log.Birth <- MCMCmetrop1R(fun=logfun
                                    ,X=Xdata
                                    ,y=yvector
                                    ,logfun=T
                                    ,theta.init= c(-1,0,0,0,0,0)
                                    ,burnin=500
                                    ,mcmc= 20000
                                    ,seed = 1234)

Birth.bayesglm <- bayesglm (formula = lowbw ~ alc + smo + soc
                            ,family = binomial(link='log')
                            ,data = Birth)
```