André Ribeiro Pinheiro da Silva

Modelando a probabilidade de ocorrência de eventos raros

Niterói - RJ, Brasil

10 de maio 2021

Universidade Federal Fluminense

André Ribeiro Pinheiro da Silva

Modelando a probabilidade de ocorrência de eventos raros

Trabalho de Conclusão de Curso

Monografia apresentada para obtenção do grau de Bacharel em Estatística pela Universidade Federal Fluminense.

Orientador(a): Prof. Dr. Jony Arrais Pinto Junior

Niterói - RJ, Brasil

10 de maio 2021

Universidade Federal Fluminense

André Ribeiro Pinheiro da Silva

Modelando a probabilidade de ocorrência de eventos raros

Monografia de Projeto Final de Graduação sob o título *"Modelando a probabilidade de ocorrência de eventos raros"*, defendida por André Ribeiro Pinheiro da Silva e aprovada em 10 de maio 2021, na cidade de Niterói, no Estado do Rio de Janeiro, pela banca examinadora constituída pelos professores:

Prof. Dr. Jony Arrais Pinto Junior
Departamento de Estatística – UFF

Prof. Dra. Márcia Marques de Carvalho
Departamento de Estatística – UFF

Prof. Dra. Jessica Quintanilha Kubrusly
Departamento de Estatística – UFF

Niterói, 10 de maio 2021

Ficha catalográfica automática - SDC/BIME
Gerada com informações fornecidas pelo autor

S586m Silva, André Ribeiro Pinheiro da
Modelando a probabilidade de ocorrência de eventos raros / André Ribeiro Pinheiro da Silva ; Jony Arrais Pinto Junior, orientador. Niterói, 2021.
77 f. : il.

Trabalho de Conclusão de Curso (Graduação em Estatística)-Universidade Federal Fluminense, Instituto de Matemática e Estatística, Niterói, 2021.

1. Modelo de Regressão Logística. 2. Evento Raro. 3. Abordagem de Firth. 4. Produção intelectual. I. Pinto Junior, Jony Arrais, orientador. II. Universidade Federal Fluminense. Instituto de Matemática e Estatística. III. Título.

CDD -

Bibliotecário responsável: Debora do Nascimento - CRB7/6368

Resumo

O modelo de regressão logística, surgiu na primeira metade do século XX, e é um dos mais populares para descrever a relação existente entre uma variável resposta binária e um conjunto de variáveis explicativas. Entretanto, é conhecido na literatura que este modelo apresenta problemas quando se trata da modelagem de um evento raro ou quando se trabalha com amostras pequenas. Um evento é considerado raro se a variável aleatória binária possui um número de ocorrências do evento de interesse (sucesso) consideravelmente mais baixo que o número de ocorrências de não interesse (fracassos). O desbalanceamento entre essas duas categorias, sucessos e fracassos, faz com que o modelo de regressão logística subestime a probabilidade de ocorrência do evento de interesse. Na literatura existem diversas alternativas apontadas para tentar solucionar este problema. A mais utilizada é o uso da abordagem de Firth à regressão logística. O objetivo deste trabalho é aplicar dois métodos de regressão logística para dados com cenários de eventos raros da área médica e financeira, buscando fazer uma comparação entre os métodos. A aplicação feita para a base médica busca compreender o impacto de fatores de risco, por exemplo, frequência cardíaca e colesterol, em doenças coronarianas. Já a aplicação feita para a base financeira, busca reconhecer transações fraudulentas com cartão de crédito por meio de variáveis explicativas resultantes de uma Análise de Componentes Principais (ACP) e outras, como por exemplo, o valor da transação e o tempo da primeira transação realizada. Os métodos de Regressão Logística usual e Regressão Logística de Firth (ou Abordagem de Firth) foram aplicados aos dois problemas e seus resultados comparados. Os dois métodos apresentaram resultados semelhantes, com uma pequena vantagem para a abordagem de Firth.

Palavras-chave: Modelo de Regressão Logística. Evento Raro. Abordagem de Firth.

Dedicatória

Dedico este trabalho a todos que acreditaram e apostaram em mim nessa jornada.

Agradecimentos

Primeiramente gostaria de agradecer aos meus pais, Sérgio e Berenice, que em nenhum momento desacreditaram de mim e sempre estiveram presentes para me dar apoio, todo suporte necessário e carinho.

Ao meu grande amigo, o qual tenho o prazer de chamar de irmão, Raphael Godoi que sempre esteve na torcida, incentivando e apoiando.

Aos amigos que fiz dentro da faculdade. Esses foram bem importantes para mim durante a jornada dentro da UFF.

A meu orientador professor Jony Arrais pelas orientações dadas ao logo do processo de construção deste trabalho.

Sumário

Lista de Tabelas

1 Introdução

1.1 Motivação

Nas mais diversas situações do cotidiano é comum se ter o interesse em investigar se existe o relacionamento entre uma variável de desfecho, usualmente denominada variável resposta, e uma ou mais variáveis explicativas. Pode-se citar, por exemplo, o desejo de entender o relacionamento da escolaridade e do sexo (variáveis explicativas) na renda (variável resposta). Em muitas dessas situações, a variável resposta assume apenas dois resultados, um que será chamado de sucesso e o outro de fracasso. Suponha que um gerente do banco esteja interessado na variável desfecho: "o cliente pagou o empréstimo solicitado?". Neste cenário, pagar o empréstimo seria considerado fracasso e não pagar o empréstimo seria considerado sucesso. Definindo sucesso e fracasso desta maneira, pode-se identificar fatores que aumentam a probabilidade de um cliente ser um mal pagador. Logo, entende-se uma variável binária, ou dicotômica, como sendo uma variável aleatória que assume um de dois valores possíveis (CLAYTON; HILLS, 2013).

O modelo de regressão logística usual, que é um caso particular dos Modelos Lineares Generalizados (MLG), é um dos modelos mais usado nesse cenário onde têm-se variáveis binárias. Tal como outros modelos estatísticos de regressão, o modelo de regressão logística apareceu para responder problemas para os quais o modelo de regressão linear clássico não podia ser aplicado (TURKMAN; SILVA, 2000), destacando-se, desde logo, os problemas com variável resposta binária (HOSMER; LEMESHOW, 2000). A literatura tem apresentado este modelo, como o mais importante para dados de resposta categórica (AGRESTI, 2002). Geralmente usado quando se está interessado em saber a relação entre uma variável resposta (dependente) e uma ou mais variáveis explicativas (independentes).

Entretanto, o modelo de regressão logística, citado anteriormente, usualmente não apresenta bons resultados se o número de sucesso for demasiadamente pequeno (HEINZE; SCHEMPER, 2002). Quando os dados estão extremamente desbalanceados, por exemplo, se os dados contiverem uma enorme proporção de fracasso, estes sucessos podem ser vistos

como raros (PAAL, 2014). Pode-se citar como exemplo de evento raro, situações cotidianas do mercado financeiro, entre as quais inclui-se o interesse em determinar a probabilidade de que cada cliente venha a cometer uma possível ação fraudulenta, sendo que essa proporção de clientes fraudadores é extremamente pequena. Um outro exemplo, este na área da saúde, é o desejo em determinar a probabilidade de que uma determinada pessoa venha a apresentar alguma infecção epidemiológica que atinge apenas uma ínfima parcela da população. Mesmo atingindo pequenas parcelas, o interesse em investigar esses eventos está presente no dia a dia de diversos pesquisadores e é de suma importância. Deste modo, é necessário usar ferramentas apropriadas para tal.

Ao longo do tempo, os eventos raros têm se mostrado bastante difíceis de se explicar e prever (KING G.; ZENG, 2001b). Existem duas teorias principais. Uma relacionada aos métodos estatísticos e outra associada a coleta de dados. Na próxima seção estas teorias serão discutidas.

Na literatura, existem muitas maneiras propostas para que sejam contornados os problemas acima apresentados. Ao longo do presente trabalho será abordada uma possibilidade de solução deste problema, por meio do modelo Regressão Logística de Firth (ou Abordagem de Firth).

A regressão logística de Firth (ou Abordagem de Firth) foi proposta por Firth (1993). O modelo está baseado na penalização da função de log-verossimilhança pela priori de Jeffrey's, se o parâmetro objetivo for o parâmetro canônico de uma família exponencial, como é o caso dos modelos lineares generalizados (MLG) e da maioria das funções de ligação, como por exemplo, logit. Este método propõe um novo estimador com viés mais baixo do que o do estimador de máxima verossimilhança.

1.2 Revisão Bibliográfica

O modelo de regressão logística, que é pertencente a classe dos Modelos Lineares Generalizados (MLG), é um dos mais utilizados para a modelagem de variáveis binárias. O modelo teve a sua primeira aparição em um bioensaio realizado por Berkson no ano de 1944 (MCCULLAGH; NELDER, 1989).

A literatura já apresentou algumas situações nos quais o modelo de regressão logística pode não apresentar bons resultados. A primeira delas diz respeito ao número de sucesso, isto é, se o mesmo for demasiadamente pequeno (HEINZE; SCHEMPER, 2002). Este cenário é chamado de dados desbalanceados, por exemplo, se os dados contiverem uma

enorme proporção de fracasso, estes sucessos podem ser vistos como raros (PAAL, 2014). Ao longo do tempo, cenários com eventos raros têm se mostrado difíceis de se modelar e se prever (KING G.; ZENG, 2001b). Duas linhas são apresentadas na literatura, a primeira linha acredita que ao se aplicar o modelo de regressão logística convencional em um cenário com eventos raros são gerados vieses nos coeficientes de regressão estimados. Esses vieses gerados são sempre na mesma direção, segundo King G.; Zeng (2001b), devido as probabilidades geradas serem sempre pequenas. Para a segunda linha de pensamento, o problema encontra-se nas escolhas feitas pelos pesquisadores que, muitas vezes, possuem recursos fixos e limitados que impactam na escolha das variáveis. E, consequentemente, optam por escolher conjuntos de dados muito grandes com poucas variáveis explicativas, e muitas vezes mal mensuradas (KING G.; ZENG, 2001a).

Na literatura, existem muitas maneiras propostas para que sejam contornados os problemas acima apresentados. Dentre elas, pode-se citar o uso do modelo de regressão logística condicional em estudos de caso-controle; o modelo de regressão logística com correção do viés, usando correção a priori e pesos; regressão binária com função de ligação potência e reversa de potência. As devidas explicações e definições dos respectivos modelos citados podem ser encontrados nos trabalhos de Santos (2017) e Huayanay (2019).

Ao longo do presente trabalho será abordada uma possibilidade de solução deste problema, por meio do modelo regressão logística de Firth (ou Abordagem de Firth).

1.3 Objetivos

O objetivo do trabalho é comparar o modelo de regressão logística com o modelo de regressão logística de Firth (ou Abordagem de Firth) na modelagem de dados binários, considerando que o sucesso é um evento raro. Sendo assim, definem-se os seguintes objetivos específicos:

- Estudar o modelo de regressão logística usual;
- Estudar a alternativa do modelo de regressão logística de Firth (ou Abordagem de Firth) para a modelagem de dados binários, com eventos raros;
- Aplicar em conjuntos de dados reais.

1.4 Organização

A organização do trabalho é descrita a seguir. No Capítulo 2, são os materiais e métodos utilizados para elaboração do trabalho são explicados. No Capítulo 3, os resultados da aplicação dos métodos nas bases de dados reais são apresentados. No Capítulo 4, se encontra a conclusão dos resultados apresentados no trabalho.

2 Materiais e Métodos

Este capítulo está estruturado da seguinte forma. Na primeira seção será apresentado o conceito de dados com eventos raros. Nas seções seguintes, serão apresentados algumas definições importantes para definir os modelos utilizados para modelar a probabilidade de eventos raros, como a definição de modelos lineares generalizados, com ênfase em regressão logística. Em seguida, os conceitos sobre a regressão logística de Firth serão apresentados.

2.1 Dados com eventos raros

Neste capítulo será introduzido um breve conceito sobre dados com eventos raros para uma melhor compreensão.

De acordo com (KING G.; ZENG, 2001a), os dados com eventos raros correspondem a variáveis dependentes binárias que, tendo em conta as suas dimensões, têm um alto valor de *não-eventos* do que *eventos de interesse*, isto é, na variável dependente teremos um número elevado de 0's do que de 1's, onde 1 são os *eventos de interesse* e 0 os *não-eventos*. Existem vários exemplos de situações nas quais deparam-se com dados com eventos raros, dentro deles podemos citar: *guerras*, *fraude no cartão de crédito* e *infecções epidemiológicas*.

Após um breve conceito sobre eventos raros, é vulgar na literatura sobre este tema ler-se, como definição, que um evento é considerado raro quando o evento de interesse ocorre em menos de 10% das observações. Nos textos que foram utilizados como referência e base para esse trabalho, é possível observar que usualmente um evento raro é definido quando o evento de interesse ocorre entre 10% e 15% das observações. Porém, neste presente trabalho, o evento será considerado raro quando o evento de interesse ocorrer em menos de 16% das observações. Ou seja, o evento será raro quando o número de *eventos de interesse* estiver entre 0% até 16%.

2.2 Modelo de Regressão linear

O modelo de regressão linear é um modelo estatístico que considera uma variável resposta Y_i ligada a um vetor de variáveis explicativas $\boldsymbol{X}_i = (1, x_{i1}, \ldots, x_{ip})$. Este modelo têm dois componentes. Um deles é determinístico, representado por uma função de $\boldsymbol{X}_i$ que indica a informação de Y_i obtida a partir do conhecimentos das variáveis explicativas. O segundo, é um componente aleatório $\boldsymbol{\epsilon}$, denominado erro aleatório, que representa outros fatores de Y_i que não são explicados por $\boldsymbol{X}_i$.

Um modelo de regressão linear pode ser representado da seguinte forma:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \ldots + \beta_p x_{ip} + \epsilon_i, i = 1, \ldots, n; \tag{2.1}$$

em que Y_i é a i-ésima observação da variável resposta, x_{ij} é a i-ésima observação da j-ésima variável explicativa, $\beta_j, j = 0, 1, \ldots, p$ são os efeitos associados as variáveis explicativas. ϵ_i é o erro aleatório do modelo associado a i-ésima observação. O modelo apresentado em (2.1), pode ser escrito da forma matricial. Deste modo, o termo $\boldsymbol{X\beta}$ representa o componente determinístico do modelo e $\boldsymbol{\epsilon}$ o componente aleatório. O modelo é representado da seguinte forma:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{\epsilon}, \tag{2.2}$$

$$\text{em que, } \boldsymbol{X} = \begin{bmatrix} 1 & x_{11} & \ldots & x_{1p} \\ . & . & \ldots & . \\ . & . & \ldots & . \\ . & . & \ldots & . \\ 1 & x_{n1} & \ldots & x_{np} \end{bmatrix}, \quad \boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T, \quad \boldsymbol{\epsilon} = (\epsilon_1, \ldots, \epsilon_n)^T.$$

No modelo de regressão linear, algumas hipóteses devem ser satisfeitas para que o modelo seja válido. A relação da variável resposta com cada variável explicativa deve ser do tipo linear. O erro aleatório do modelo tem média zero e variância σ^2 constante para todo i, além disso os erros devem ser não correlacionados. A última hipótese é que o erro tem distribuição normal, ou seja, $\boldsymbol{\epsilon} \sim N(0, \sigma^2 \boldsymbol{I_n})$, $\boldsymbol{I_n}$ é a matriz identidade de ordem n.

Note que, os valores de $\boldsymbol{X}$ são conhecidos, pois foram observados, além disso, $\boldsymbol{\beta}$, embora desconhecido, é constante. Assim, a única variável aleatório no modelo são os

erros aleatórios. Portanto, dado que $\boldsymbol{X}$ foi observado, $\boldsymbol{Y}$ é uma combinação linear de variáveis aleatórias com distribuição normal, logo, $\boldsymbol{Y}$ também tem distribuição normal,

$$\boldsymbol{Y}|\boldsymbol{X}, \boldsymbol{\beta}, \sigma^2 \sim N\left(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2 \boldsymbol{I_n}\right). \tag{2.3}$$

Se alguma das hipóteses não são atendidas, o modelo não é válido. Por exemplo, se $\boldsymbol{Y}$ é um vetor composto por uma variáveis binárias, não é plausível assumir que sua distribuição seja normal, de forma que o uso do modelo de regressão linear não é adequado. Nesses casos, é preciso utilizar modelos que permitam que $\boldsymbol{Y}$ tenha diferentes distribuições. Uma classe de modelos que permite esta possibilidade, é classe dos modelos lineares generalizados.

2.3 Modelos Lineares Generalizados

Os modelos lineares generalizados foram apresentados por Nelder e Wedderburn (1972). Essa classe de modelos é uma generalização dos modelos de regressão linear, flexibilizando uma das principais restrições dos modelos lineares de que a variável resposta $\boldsymbol{Y}$ deve seguir uma distribuição normal. $\boldsymbol{Y}$ pode assumir alguma distribuição que pertença à família exponencial. Segue abaixo, a definição de família exponencial no caso uniparamétrico, que será o caso abordado neste trabalho.

Definição 2.1 *Seja $\boldsymbol{Y}$ uma variável aleatória com f.p ou f.d.p dada por $f(\boldsymbol{y}\,|\boldsymbol{\theta})$, $\boldsymbol{Y}$ pertence à família exponencial se:*

$$f(\boldsymbol{y}\,|\boldsymbol{\theta}) = a(\boldsymbol{\theta})\, b(\boldsymbol{y}) \exp\left(c(\boldsymbol{\theta})\, d(\boldsymbol{y})\right), \boldsymbol{y} \in A \tag{2.4}$$

em que, $\boldsymbol{\theta}$ é o parâmetro desconhecido da distribuição de $\boldsymbol{Y}$, as funções $a(\cdot)$, $b(\cdot)$, $c(\cdot)$, $d(\cdot)$ são bem definidas e A é um conjunto que não depende do parâmetro desconhecido θ.

Se $\boldsymbol{Y}$ segue uma distribuição que pertence à família exponencial uniparamétrica, então para explicar $\boldsymbol{Y}$, é preciso conhecer o seu parâmetro que, por enquanto, será denotado por $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_n)^T$, um para cada observação. As notações usadas para definir estes modelos serão os mesmos de modelos lineares.

Na classe de modelos lineares generalizados o vetor de parâmetros $\boldsymbol{\theta}$ é relacionado ao preditor linear $\boldsymbol{X_i}\boldsymbol{\beta}$ por meio de uma função de ligação g. A seguir, a forma geral de um

modelo linear generalizado:

$$g\left(E(Y_i)\right) = \boldsymbol{X_i\beta} = \beta_0 + \beta_1 x_{i1} + \ldots + \beta_p x_{ip}, i = 1, \ldots, n. \tag{2.5}$$

Observe que, se $Y_i \sim N(\theta_i, \sigma^2)$ e $g(E(Y_i)) = g(\theta) = \theta$, sendo g a função de identidade, o modelo linear especificado na seção anterior é obtido. Neste trabalho, foi utilizado apenas um dos casos da classe dos modelos lineares generalizados, o modelo de regressão logística, que deve ser usado quando a variável resposta é binária ou dicotômica. A seguir o modelo será apresentado de forma mais detalhada.

2.4 Modelos de Regressão Logística

O modelo de regressão logística é um caso particular da classe de modelos lineares generalizados. Ela tem como objetivo produzir, a partir de um conjunto de observações, um modelo que permita a predição de valores tomados por uma variável categórica, frequentemente binária, a partir de uma série de variáveis explicativas contínuas e/ou binárias.

Para uma definição completa do modelo, suponha-se que foram observados n unidades experimentais. De cada unidade experimental, foi observado uma variável reposta Z_i e p variáveis explicativas, como por exemplo sexo, estado civil, etc. A variável Z_i assume somente dois resultados, aqui será chamado de sucesso um dos resultados e fracasso o seu complementar. Desta forma, Z_i é definido da seguinte forma:

$$Z_i = \begin{cases} 1 \text{ , se a } i\text{-ésima unidade experimental é sucesso} \\ 0 \text{ , caso contrário} \end{cases}, i = 1, \ldots, n.$$

Considere-se que foi observado uma amostra aleatória simples de tamanho n, isto é, $Z_1, Z_2, \ldots, Z_n$. Sem perda de generalidade, suponha que foram observados $p = 2$ variáveis explicativas, sexo e estado civil e que o conjunto de todas os possíveis resultados observados foi

$$V = \{SF, SM, CF, CM, V_iF, V_iM\},$$

em que S é solteiro, C é casado, V_i é viúvo, F é feminino e M é masculino. V representa o conjunto com todas as configurações observadas. Supondo p covariáveis observadas, pode-se dizer que a cardinalidade de V é N e $N = 6$. Desta forma, pode-se dividir a nossa amostra em J estratos e definir em cada estrato a variável $Z_l{}^j$, $j = 1, \ldots, N$ e $l = 1, \ldots, n_j$,

$n_1 + n_2 + \ldots + n_N = n$ da seguinte forma:

$$Z_l{}^j = \begin{cases} 1 \text{ , se o indivíduo } l \text{ do estrato } j \text{ é sucesso} \\ 0 \text{ , caso contrário} \end{cases}$$

em que $P(Z_l{}^j = 1) = P_j$ e $P(Z_l{}^j = 0) = 1 - P_j$. Deste modo, será definida a variável

$$Y_j = \sum_{l=1}^{n_j} Z_l{}^j \sim Binomial(n_j, P_j) \text{ , j = 1, ..., N.}$$

Para estimar a probabilidade P_j, usa-se o modelo de regressão logística, definido a seguir:

$$\log\left(\frac{P_j}{1-P_j}\right) = \boldsymbol{X}_i\boldsymbol{\beta}^T, \tag{2.6}$$

em que $\log\left(\frac{P_j}{1-P_j}\right)$ é a função de ligação do modelo, $\boldsymbol{X}_i$ é a matriz de ordem $n \times p$ do modelo e $\boldsymbol{\beta}$ é o vetor dos parâmetros desconhecidos de dimensão $p \times 1$, cujos elementos são os coeficientes de regressão.

Em uma abordagem frequentista, o método de máxima verossimilhança é comumente utilizado para obter-se as estimativas dos parâmetros desconhecidos. A seguir o método é detalhado.

2.5 Função de Verossimilhança

Considerando $\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)^T$ um vetor aleatório e $f(y;\theta)$ a função de probabilidade conjunta de Y, com $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_p)^T$. A função de verossimilhança, $L(\theta; y)$, algebricamente é igual a função de probabilidade $f(y;\theta)$. Lembrando que a diferença entre elas reside no que é considerado conhecido e desconhecido. Por exemplo, na função de verossimilhança os valores do vetor $\boldsymbol{Y}$ são conhecidos e os valores do vetor $\boldsymbol{\theta}$ desconhecidos, isto é, L é função de θ, já na função de probabilidade ocorre o contrário (DOBSON; BARNETT, 2018).

A função de verossimilhança de forma geral é definida por:

$$\begin{aligned} L(\theta; y) &= L(\theta; Y_1, Y_2, \ldots, Y_n) \\ &= f(Y_1, Y_2, \ldots, Y_n; \theta) \\ &= f(Y_1; \theta) f(Y_2; \theta) \ldots f(Y_n; \theta) \\ &= \prod_{i=1}^{n} f(Y_i; \theta) \end{aligned} \tag{2.7}$$

onde os Y_i's são independentes e identicamente distribuídos (i.i.d's).

Sendo assim, ao aplicar a função de verossimilhança para a estimação do vetor de parâmetros $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$, que é o problema de interesse deste trabalho, têm-se o interesse de determinar o valor dos parâmetros que maximiza a probabilidade de se obter o cojunto de dados observados.

Com base na função (2.7), obtem-se a seguinte equação:

$$log(1 - p_i) = -log[1 + exp(\boldsymbol{X}_i \boldsymbol{\beta}^T)] \tag{2.8}$$

Dessa forma, pode-se aplicar a função logarítmica em cima desta função de verossimilhança, obtendo assim a função de log-verossimlhança dada por:

$$\begin{aligned} l(\boldsymbol{\beta}; Y) &= \left(\sum_{j=1}^{N} Y_j \log \left(\frac{P_j}{1 - P_j} \right) + n_j log(1 - P_j) + log \binom{n_j}{Y_j} \right) \\ &= \sum_{j=1}^{N} \left(y_j \boldsymbol{X}_j \boldsymbol{\beta}^T - n_j log[1 + exp(\boldsymbol{X}_j \boldsymbol{\beta}^T)] + log \binom{n_j}{Y_j} \right). \end{aligned} \tag{2.9}$$

Considera-se a função log-verossimilhança para efeitos de contas pois, é considerada mais simples de ser trabalhar do que a função de verossimilhança.

O objetivo então, é determinar o vetor $\boldsymbol{\beta}$ para qual a função de verossimilhança é máxima. O primeiro passo consiste em derivar a função de log-verossimilhança para cada β_k, $k = 1, \ldots, p$, do vetor $\boldsymbol{\beta}$, e igualar a derivada a zero. Como há $p + 1$ parâmetros então, serão obtidos $p + 1$ que correspondem a um sistema de equações.

Os valores dos parâmetros β's que são solução desse sistema de equações são chamados de estimadores de máxima verossimilhança e são denotados por $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$.

As equações do sistema não são lineares nos parâmetros. Então, para se chegar a uma solução tem-se que recorrer aos métodos conhecidos como métodos numéricos interativos.

Um método interativo para a obtenção da solução do sistema é o método iterativo dos mínimos quadrados ponderados.

2.6 Teste de Significância dos Parâmetros

O teste de Wald, permite que seja avaliado se algum ou todos os coeficientes são não nulos, usando para tal uma estatística de teste, em geral denominada de estatística de Wald. Essa estatística compara as estimativas de máxima verossimilhança dos parâmentros β_k, $k = 0, 1, 2, \ldots, p$ com seu erro padrão. A expressão da estatística univariada é da seguinte forma:

$$W_k = \frac{\hat{\boldsymbol{\beta}}_k}{\hat{EP}(\hat{\boldsymbol{\beta}}_k)} \tag{2.10}$$

onde $k = 0, 1, 2, \ldots, p$ e EP é o erro padrão.

A estatística de Wald segue uma distribuição normal padrão, $W_k \sim N(0, 1)$, se for considerada a hipótese nula $H_0 : \boldsymbol{\beta}_k = 0$ (HOSMER; LEMESHOW, 2000). Se $k = 0$ então é necessário calcular a estatística de teste para o coeficiente independente β_0 e será testado se o coeficiente é nulo ou não nulo. Porém, se $k = 1, 2, \ldots, p$ então será avaliado se alguma das variáveis X_j não deveria constar no modelo de regressão, isto é, se aquela variável não está relacionada com o desfecho. Agresti (2002), define a expressão da estatística de Wald multivariada da seguinte forma:

$$W_k = (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*)^T [cov(\hat{\boldsymbol{\beta}})]^{-1} (\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^*), \tag{2.11}$$

onde $cov(\hat{\boldsymbol{\beta}})$ é o número de parâmetros não redundantes em $\boldsymbol{\beta}$ e considerando, $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$, $\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \ldots, \hat{\beta}_p)^T$ e $\boldsymbol{\beta}^* = ({\beta_0}^*, {\beta_1}^*, \ldots, {\beta_p}^*)^T$.

O objetivo então, por exemplo, é testar as seguintes hipóteses:

$$\boldsymbol{H_0} : \boldsymbol{\beta}_1 = {\boldsymbol{\beta}_1}^*$$
$$\boldsymbol{H_1} : \boldsymbol{\beta}_1 \neq {\boldsymbol{\beta}_1}^*$$

A estatística anterior segue uma distribuição χ^2 com $p + 1$ graus de liberdade. Dessa forma, para um nível de significância α, será considerado a hipótese alternativa, H_1, como verdadeira caso consiga-se rejeitar a hipótese nula, H_0, isto é, se for verificado que $W > \chi^2_{1-\alpha}(p + 1)$.

2.7 Interpretação dos parâmetros - Razão de Chances

A função de ligação logit é a mais utilizada no ajuste do modelo Binomial por ser capaz de fornecer uma interpretação conveniente dos parâmetros.

A chance de ocorrer o evento, dada a ligação logit é dado por:

$$\left(\frac{P_j}{1-p_j}\right) = exp(\boldsymbol{X}_j\boldsymbol{\beta}) = exp(\beta_0 + \beta_1 x_{j_1} + \beta_2 x_{j_2} + \ldots + \beta_p x_{j_p}) \tag{2.12}$$

Por exemplo, se uma variável independente contínua x_1, for acrescida de uma unidade, mantendo as outras variáveis independentes do modelo fixas, a chance do evento fica:

$$\begin{aligned}\left(\frac{P_j^*}{1-P_j^*}\right) &= exp(\beta_0 + \beta_1(x_{j_1}+1) + \beta_2 x_{j_2} + \ldots + \beta_p x_{j_p}) \\ &= exp(\beta_0 + \beta_1 x_{j_1} + \beta_2 x_{j_2} + \ldots + \beta_p x_{j_p} + \beta_1) \\ &= exp(\beta_0 + \beta_1 x_{j_1} + \beta_2 x_{j_2} + \ldots + \beta_p x_{j_p}) exp(\beta_1) \\ &= \left(\frac{P_j}{1-p_j}\right) exp(\beta_1)\end{aligned} \tag{2.13}$$

Assim, a razão de chances (odds ratio) de (x_1+1) em relação a x_1, $OR(x_1+1, x_1)$ é dada por

$$OR(x_1+1, x_1) = \frac{\frac{P_j^*}{1-P_j^*}}{\frac{P_j}{1-p_j}} = exp(\beta_1),$$

ou seja, a chance do evento ocorrer entre os indivíduos que diferem na variável x_1 em 1 unidade é $exp(\beta_1)$. Neste caso, a estimativa da razão de chances é $\hat{OR}(x_1+1, x_1) = exp(\hat{\beta}_1)$. De forma geral, a estimativa da razão de chances com o acréscimo de c unidades, ou seja, substituindo $(x_{j_1}+1)$ por $(x_{j_1}+c)$ em (2.17), é dada por

$$\hat{OR}(x_1+c, x_1) = exp(c\hat{\beta}_1). \tag{2.14}$$

Por (2.14), temos que $log(\hat{OR}(x_1+1, x_1)) = c\hat{\beta}_1$. Assim, um intervalo de $100(1-\alpha)\%$ de confiança aproximado para as estimativas da razão de chances é obtido ao calcular inicialmente um intervalo de confiança para β e, então transformar seus limites, ou seja,

$$IC(100(1-\alpha)\%, OR(x_1+1, x_1)) = exp[c\hat{\beta}_1 \pm z_{1-\frac{\alpha}{2}} c\hat{EP}(\beta_1)], \tag{2.15}$$

em que $c\hat{E}P(\beta_1)$ é a estimativa do erro padrão de $\hat{\beta}_1$, obtida da raiz quadrada do segundo termo da diagonal principal de $I(\hat{\beta}^{-1})$, em que $I(\hat{\beta}^{-1})$ é a inversa da matriz de informação de Fisher estimada.

É muito comum que no modelo exista pelo menos uma variável independente que seja categórica. Nesses casos, variáveis explicativas auxiliares são utilizadas.

2.8 Modelo de Regressão Logística de Firth

O método proposto por Firth (1993) surgiu pelo fato dos métodos utilizados para correção naquela época serem, de acordo com suas palavras, apenas "corretivos"e pouco "preventivos", uma vez que os estimadores de máxima verossimilhança eram primeiro calculados e só depois corrigidos. Dessa forma, o método de Firth está baseado na penalização da função de log-verossimilhança pela priori de Jeffrey's, se o parâmetro objetivo for o parâmetro canônico de uma família exponencial, como é o caso do modelos lineares generalizados (MLG) e da maioria das funções de ligação, como por exemplo, a logit.

Este método propõe um novo estimador com viés mais baixo do que o do estimador de máxima verossimilhança. Esse novo estimador é obtido, nos modelos da família exponencial, através da resolução da seguinte equação:

$$U^*(\boldsymbol{\beta}) = U(\boldsymbol{\beta}) + \left(\frac{1}{2}\right) tr[I^{-1}(\boldsymbol{\beta})\frac{d}{d\boldsymbol{\beta}}I(\boldsymbol{\beta})] \tag{2.16}$$

em que tr = traço e $U^*(\boldsymbol{\beta})$ é a função de score modificada de modo que a solução resulte um estimador com o viés menor do que o obtido pela método de verossimilhança.

Nesse caso, para o modelo de regressão logística a equação (2.16) pode ser escrita da seguinte forma:

$$U^*(\boldsymbol{\beta}_j) = \sum_{i=1}^{n}(Y_i - \mu_i)x_{ij} + \sum_{i=1}^{n} h_i(0.5 - \mu_i)x_{ij} \text{ ; j = 0,1,. . . ,p} \tag{2.17}$$

h_i é o i-ésimo elemento da diagonal de $\boldsymbol{H} = \boldsymbol{W}^{\left(\frac{1}{2}\right)}\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{W}\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{W}^{\left(\frac{1}{2}\right)}$ e $\boldsymbol{W} = diag[\hat{\mu}_i(1-\hat{\mu}_i)]$. A solução dessa equação pode ser obtida por meio de métodos numéricos, como exemplo o método de Newton Raphson.

2.8.1 Distribuição a priori de Jeffreys

Nesta seção, será abordado um pouco melhor a definição da distribuição a priori de Jeffreys. Antes serão definidas algumas equações importantes.

Seguindo então uma notação idêntica a utilizada por McCullagh e Nelder (1989), considere:

$$U_r(\boldsymbol{\theta}) = \frac{\partial}{\partial \theta_r} \text{ e } U_r s(\boldsymbol{\theta}) = \frac{\partial^2}{\partial^2 \theta_{rs}} \tag{2.18}$$

onde $\boldsymbol{\theta} = (\theta_1, \theta_2), \ldots, \theta_p)^T$ é um parâmetro vetor.

Os cumulantes conjuntos nulos são dados por:

$$K_{r,s} = n^{-1}E[U_r U_s],\ K_{r,s,t} = n^{-1}E[U_r U_s U_t],\ K_{r,s\,t} = n^{-1}[U_r U_{st}] \tag{2.19}$$

e assim sucessivamente. Aos cumulantes nulos, verificam-se as seguintes igualdades:

$$K_{rs} + K_{r,s} = 0 \text{ e } K_{rst} + K_{r,s\,t} + K_{s,r\,t} + K_{t,r\,s} + K_{r,s,t} = 0 \tag{2.20}$$

Será considerada uma modificação bastante geral na função dos scores da forma

$$U_r{}^*(\boldsymbol{\theta}) = U_r(\boldsymbol{\theta}) + A_r(\boldsymbol{\theta}) \tag{2.21}$$

onde $A_r(\theta)$ está sempre dependendo dos dados. Sendo assim, se for suposto que $\hat{\boldsymbol{\theta}}$ e $\boldsymbol{\theta}^*$ são tais que satisfazem, respectivamente

Usando um argumento fechado utilizado por McCullagh e Nelder (1989), baseado na expansão de $U_r{}^*(\boldsymbol{\theta}^*)$ sobre o valor de $\boldsymbol{\theta}$, o viés de $\boldsymbol{\theta}^*$, tem-se que

$$E(\boldsymbol{\theta}^* - \boldsymbol{\theta})^r = n^{-1}K^{r,s}[-K^{t,u}(K_{s,t,u} + \frac{K_{s,t\,u}}{2} + \alpha_s) + O(n^{-\frac{3}{2}})]$$

onde $K^{r,s}$ denota a matriz inversa de informação de Fisher $K_{r,s}$ e α_s denota a expectativa nula de A_s.

Na expressão anterior, o termo

$$-n^{-1}K^{r,s}[-K^{t,u}(K_{s,t,u} + \frac{K_{s,t\,u}}{2} = n^{-1}{b_1}^r(\boldsymbol{\theta}),$$

corresponde ao viés de primeira ordem de $\hat{\boldsymbol{\theta}}$ (FIRTH, 1993).

Assim, para a remoção do termo de primeira ordem do viés utiliza-se A_r, caso se verifique que

$$\alpha_r = -K_{r,s}{b_1}^s + O(n^{-\frac{1}{2}})].$$

Em uma notação matricial o vetor A deve ser tal que

$$E(A) = \frac{-i(\boldsymbol{\theta})b_1(\boldsymbol{\theta})}{n} + O(n^{-\frac{1}{2}}).$$

Existem dois candidatos mais ou menos óbvios para a escolha de A. À semelhança do que foi proposto por Firth (1993), teremos:

$$A^{(E)} = \frac{-i(\boldsymbol{\theta})b_1(\boldsymbol{\theta})}{n} \tag{2.22}$$

$$A^{(O)} = \frac{-I(\boldsymbol{\theta})b_1(\boldsymbol{\theta})}{n}, \tag{2.23}$$

usando, respectivamente, a informação esperada e a informação observada.

Sendo assim, após definidas as equações acima é possível especificar priori de Jeffreys.

A utilização da distribuição a priori de Jeffreys como função de penalização constitui uma restrição à família exponencial.

Se $\boldsymbol{\theta}$ é um parâmetro pertencente a um modelo de família exponencial, então temos que $K_{r,s_t} = n^{-1}E(U_rU_{st}) = 0$, para todo r, s e t. Lembrando que K_{r,s_t} é um cumulante conjunto nulo. Onde em sua equação temos as derivadas da função de log-verossimilhança, (2.22) e (2.23), definidas por McCullagh (MCCULLAGH, 1987). Deste modo, o r-ésimo elemento de $A^{(E)}(\boldsymbol{\theta})$, ou equivalentemente $A^{(O)}(\boldsymbol{\theta})$, é, considerando (2.20), dado por:

$$a_r = -\frac{nK_{r,\,s}b_1^s}{n} = \frac{K_{r,\,s}K^{r,t}K^{u,v}K_{t,\,u,v}}{2} = \frac{K^{u,v}K_{r,\,u,v}}{2} = -\frac{K^{u,v}K_{r,\,u,v}}{2}. \tag{2.24}$$

Utilizando a notação matricial, é possível escrever a expressão anterior da seguinte forma:

$$a_r = \frac{1}{2}tr[i^{-1}(\frac{\partial i}{\partial \boldsymbol{\theta}_r})] = \frac{\partial i}{\partial \boldsymbol{\theta}_r}[\frac{1}{2}log|i(\boldsymbol{\theta})|]. \tag{2.25}$$

A solução de $U_r^* = U_r + a_r = 0$ localiza-se em um ponto estacionário de

$$l^*(\boldsymbol{\theta}) = l(\boldsymbol{\theta}) + \frac{1}{2}log|i(\boldsymbol{\theta})|, \tag{2.26}$$

ou equivalentemente, da função de verossimilhança penalizada:

$$L^*(\boldsymbol{\theta}) = L(\boldsymbol{\theta})|i(\boldsymbol{\theta})|^{\frac{1}{2}}. \tag{2.27}$$

A função de penalização concluída é $|i(\boldsymbol{\theta})|^{\frac{1}{2}}$, que corresponde à expressão da distribuição a priori de Jeffreys (que é invariante por reparametrizações do parâmetro $\boldsymbol{\theta}$), introduzida por H. Jeffreys em 1946. Mais detalhes sobre o passo a passo e de como chega-se a essa conclusão, podem ser encontrados em Firth (1993).

2.9 Medidas de Comparação entre os Modelos

Neste trabalho, serão considerados dois modelos para dados binários. Após realizados os ajustes dos modelos de regressão, naturalmente tem-se o questionamento sobre qual dos ajustes é o melhor. A seguir serão apresentados alguns procedimentos de obtenção dessas medidas para que as comparações sejam feitas entre os modelos.

2.9.1 Qualidade do Ajuste

Muitas vezes mais de um modelo pode descrever o mesmo fenômeno, dessa forma, existe a necessidade de obter medidas baseadas em princípios científicos que permitam escolher qual deles é mais adequado para cada situação.

Uma dessas maneiras de comparar dois ou mais modelos é atráves do valor da medida -2log(L($\hat{\theta}$)). Quanto menor esta medida, melhor é o modelo.

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

$$AIC = -2log(L(\hat{\theta})) + 2p \tag{2.28}$$

onde $L(\hat{\theta})$ é a função de verossimilhança e p o número de parâmetros.

Outro método utilizado para saber se o modelo é adequado, é o Critério de Informação Bayesiano (BIC) proposto por Schwarz (1978) que leva em conta o número de parâmetros do modelo p e o tamanho da amostra n. A medida BIC é dada por:

$$BIC = -2log(L(\hat{\theta})) + plog(n) \tag{2.29}$$

2.9.2 Medidas Desempenho

À obtenção das medidas de desempenho é um procedimento importante para a avaliação do poder de predição do modelo. Isto é feito, em geral, por meio da sensibilidade, especificidade, acurácia, valores preditivos positivos e valores preditivos negativos. Para efeitos de representação da equação de cada medida, considere $\hat{Y} = 1$ se um indivíduo selecionado ao acaso da população em estudo for classificado como evento e $\hat{Y} = 0$ se classificado como não-evento. Dessa forma, as equações são definidas da seguinte forma:

- Sensibilidade: probabilidade de classificação correta do envento:

$$SE = P(\hat{Y} = 1|Y = 1) = \frac{P(\hat{Y} = 1; Y = 1)}{P(Y = 1)}.$$

- Especificidade: probabilidade de classificação correta do não-evento:

$$ES = P(\hat{Y} = 0|Y = 0) = \frac{P(\hat{Y} = 0; Y = 0)}{P(Y = 0)}.$$

- Acurácia: probabilidade de classificação correta:

$$ACC = P(Y = 1; \hat{Y} = 1) + P(Y = 0; \hat{Y} = 0).$$

- Verdadeiro Preditivo Positivo: probabilidade do indivíduo ser evento, dado que foi classificado como evento:

$$VPP = P(Y = 1|\hat{Y} = 1) = \frac{P(Y = 1; \hat{Y} = 1)}{P(\hat{Y} = 1)}.$$

- Verdadeiro Preditivo Negativo: probabilidade do indivíduo ser não-evento, dado que foi classificado como não-evento:

$$VPN = P(Y = 0|\hat{Y} = 0) = \frac{P(Y = 0; \hat{Y} = 0)}{P(\hat{Y} = 0)}.$$

A Tabela 1 apresenta os possíveis cenários ao se realizar uma classificação. Segue abaixo as definições referentes a cada componente da tabela:

- verdadeiro positivo (VP): número de eventos classificados corretamente como eventos.
- verdadeiro negativo (VN): número de não-eventos classificados corretamente como não-eventos.

- falso positivo (FP): número de não-eventos classificados incorretamente como eventos.
- falso negativo (FN): número de eventos classificados incorretamente como não-eventos.

Tabela 1: Tabela de classificação do modelo

	Observado	
Predito	0	1
0	VN	FN
1	FP	VP

Através dos possíveis resultados na classificação dos indivíduos, é possível estimar as probabilidades das medidas de desempenho. Logo, é possível obter as equações de estimativa da sensibilidade, especificidade, acurácia, verdadeiros positivos e verdadeiros negativos. São elas:

- $$\hat{SE} = \frac{VP}{VP + FN}$$
- $$\hat{ES} = \frac{VN}{VN + FP}$$
- $$\hat{ACC} = \frac{VP + VN}{n}$$
- $$\hat{VPP} = \frac{VP}{VP + FP}$$
- $$\hat{VPN} = \frac{VN}{VN + FN}$$

Vale ressaltar que, neste trabalho as estimativas das medidas de desempenho serão apresentadas em porcentagem.

2.9.3 Curva AUC

A curva AUC é derivada da curva ROC, então inicialmente será apresentado a curva ROC, que significa "Receiver Operating Characteristic".

A curva ROC mostra o quão bom o modelo criado pode distinguir entre dois resultados (já que é utilizado para classificação). Essas duas coisas podem ser fracasso ou sucesso, ou positivo e negativo. Os melhores modelos conseguem distinguir com precisão o binômio.

A ROC possui dois parâmetros:

- Taxa de verdadeiro positivo, que é dado por,

$$\frac{\text{verdadeiros positivos}}{\text{verdadeiros positivos} + \text{falsos negativos}}$$

- Taxa de falso positivo, que é dado por,

$$\frac{\text{falsos positivos}}{\text{falsos positivos} + \text{verdadeiros negativos}}$$

Uma curva ROC delineia "Taxa de verdadeiro positivo vs. Taxa de falso positivo" em diferentes limiares de classificação.

Assim, na tentativa de simplificar a análise da ROC, a AUC ("area under the ROC curve") nada mais é que uma maneira de resumir a curva ROC em um único valor, agregando todos os limiares da ROC, calculando a "área sob a curva".

O valor do AUC varia de 0 até 1 e o limiar entre a classe é $0,5$. Ou seja, acima desse limite, o algoritmo classifica em uma classe e abaixo na outra classe. Quanto maior o AUC, melhor.

O interessante do AUC é que a métrica é invariante em escala, uma vez que trabalha com precisão das classificações ao invés de seus valores absolutos. Além disso, também mede a qualidade das previsões do modelo, independentemente do limiar de classificação.

Na representação gráfica que se segue encontram-se as representações das curvas ROC para os dois modelos de regressão, regressão logística usual e regressão logística de Firth, aplicados na base de estudos da área médica e considerando para tal o melhor modelo. Essa base e os respectivos resultados obtidos através das aplicações dos modelos, serão vistos mais a frente e com mais detalhes.

Vale ressaltar aqui que seus respectivos valores de AUC são:

- Para Figura 1 o valor do AUC $= 0,7525138$
- Para Figura 2 o valor do AUC $= 0,7526155$

Figura 1: Representação gráfica da curva ROC para o modelo de regressão logística usual

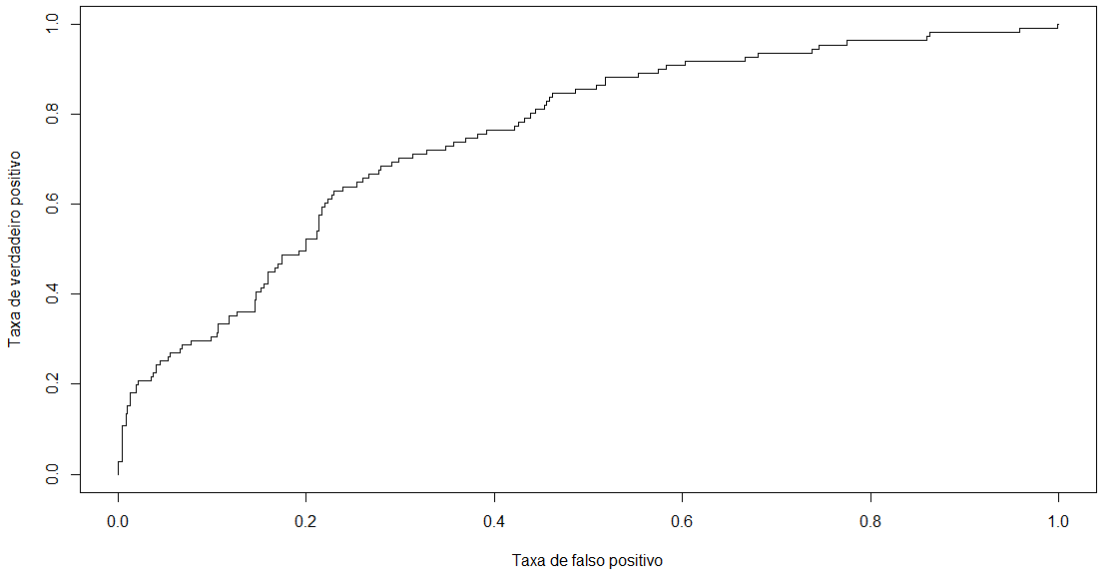
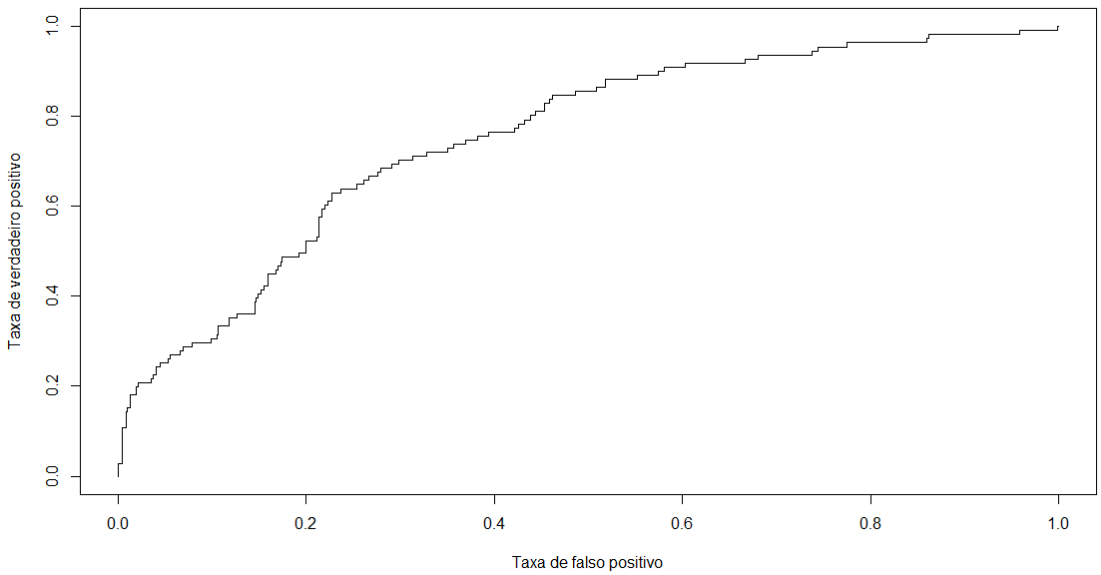


Figura 2: Representação gráfica da curva ROC para o modelo de regressão logística de Firth



3 Análise dos Resultados

Neste capítulo, será feita a aplicação dos dois modelos descritos anteriormente, modelo de Regressão Logística usual e o de Regressão Logística de Firth, a um conjunto de dados sobre a área médica e um conjuntos de dados da área financeira.

3.1 Resultados encontrados para a base da área médica

3.1.1 Descrição dos dados

Afim de aplicar os métodos estudados em dados reais da área médica, serão considerados os dados do estudo *Framinghanm Heart Study* (CARROLL et al., 1984). O interesse do estudo é verificar o impacto de fatores de risco em doenças coronarianas na população americana. As doenças coronarianas se desenvolvem por conta do depósito de substâncias gordurosas nas paredes das artérias, restringindo assim o fluxo sanguíneo e são as causas de ataque cardíaco.

Nesta base tem-se a variável resposta que indica a presença ou ausência da doença coronariana em um período de 10 anos, chamada de *TenYearCHD* e as variáveis preditoras (ou fatores de risco) que são:

- homem: se o paciente é homem ou não (1, se sim e 0 para não);
- idade: Idade, em anos, no momento do exame médico;
- educação: Uma variável categórica da escolaridade dos participantes, com os níveis: algum ensino médio, ensino médio / GED, alguma faculdade / escola profissional, faculdade;
- fumante atual: Tabagismo atual no momento dos exames;
- cigarros por dia (CPD): Número de cigarros fumados por dia;

- BPmeds: Uso de medicamento anti-hipertensivo no exame;
- AVC prevalente (AVC): se o paciente teve ou não um AVC (1, se teve e 0 para não teve);
- hipertenso: se o paciente era hipertenso ou não (1, se sim e 0 para não);
- diabetes: se o paciente tinha diabetes ou não (1, se sim e 0 para não);
- colesterol total: Colesterol total (mg / dL);
- pressão arterial sistólica (PAS): Pressão arterial sistólica (mmHg);
- pressão arterial diastólica (PAD): Pressão arterial diastólica (mmHg);
- IMC: Índice de massa corporal;
- frequência cardíaca: Frequência cardíaca (batimentos / minuto);
- glicose: Nível de glicose no sangue (mg / dL).

Os modelos estudados foram utilizados para verificar quais fatores de risco influenciam na probabilidade de surgimento de doenças coronarianas na população. A variável resposta, Y_i, indica a presença ou ausência de doenças coronarianas em um período de 10 anos (chamamos de TenYearCHD) para a i-ésimo pessoa e os fatores de risco (variáveis preditoras).

O estudo consiste na observação de 4.240 pessoas. Porém, só será feita análise em cima de 3.658 observações, devido a 582 observações, dessas 4.240, apresentarem alguns valores ausentes, e assim sendo retiradas da base de dados.

Ao serem feitas as primeiras análises, conforme a Tabela 2, pode-se perceber que o número de presença é bem menor do que a quantidade de ausências. Dessa forma, a proporção de sucessos (ou a proporção de uns) é de $15,2\%$ em uma população de 31 a 65 anos de idade, configurando assim, um caso de dados desbalanceados. Vale ressaltar que a Tabela 2 foi obtida através da base de dados completa.

Tabela 2: Frequência de pessoas segundo presença/ausência de doenças coronarianas

Ausência	Presença
3.101	557

Tabela 3: Análise das variáveis quantitativas

Variável	Média da variável na Amostra Completa	Média da variável entre os indivíduos com a doença coronariana
idade	49,55	54,27
CPD	9,02	10,48
colesterol total	236,84	246,35
PAS	132,37	143,98
PAD	82,91	87,15
IMC	25,78	26,56
frequência cardíaca	75,73	76,31
glicose	81,85	88,73

Nas Tabelas 3 e 4, encontram-se as análises descritivas referentes as variáveis explicativas da base de dados completa.

Analisando a Tabela 3, é possível verificar a coerência entre os valores relativo às duas amostras, completa e a que tem presença de doença coronariana. Em todas as variáveis quantitativas analisadas na Tabela, chama a atenção o fato de que na amostra com a presença da doença coronariana os valores são consistentemente superiores, o que era esperado do ponto de vista clínico. Tal fato, é um indicador de consistência da base de dados.

Na Tabela 4, da mesma forma que na Tabela 3, o resultado de todas as variáveis também apontam para a consistência da base de dados. Em todas as variáveis de anamnese discriminados na tabela, os seus respectivos resultados são coerentes com o que se espera do ponto de vista clínico. Alguns valores valem destacar:

- A incidência da doença coronariana entre os homens é mais de 50% superior ao manifestado nas mulheres;
- A incidência da doença coronariana entre os que possuem educação superior completa $(56,74\%)$ é substancialmente elevada. É provável que tal valor decorra da maior expectativa de vida entre os indivíduos desse segmento social, o que faz sentido, considerando a alta incidência entre os idosos;
- A incidência da doença coronariana entre os indivíduos portadores de diabetes é mais de 140% superior aos indivíduos não portadores de diabetes.

O trabalho titulado *"Avaliação do Risco de Doença Coronariana em Adultos e Idosos no Município de Lagêado do Tabocal/BA"*,Mascarenhas C. H. M; REIS (2009), aponta

Tabela 4: Análise das variáveis qualitativas

Variável	Fatores	Percentual do Fator na Amostra completa	Percentual do Fator entre os indivíduos com doença coronariana
homem	SIM	44,36%	18,91%
	NÃO	55,63%	12,28%
educação	Ensino médio incompleto	21,06%	19,07%
	Ensino médio completo/GED	15,2%	23,8%
	Faculdade incompleta/escola profissional	8,4%	37%
	Faculdade completa	5,84%	56,74%
fumante atual	SIM	48,91%	15,93%
	NÃO	51,1%	14,55%
BPMeds	SIM	3,03%	33,33%
	NÃO	96,96%	14,66%
AVC	SIM	0,57%	38,1%
	NÃO	99,43%	15,1%
hipertenso	SIM	31,16%	24,91%
	NÃO	68,83%	10,84%
diabetes	SIM	2,7%	35,35%
	NÃO	97,3%	14,67%

resultados bastante alinhados aos valores encontrados nas Tabelas 3 e 4 e suas respectivas análises.

Sendo assim, dando seguimento as análises, realizou-se uma descrição mais detalhada a respeito do modelo. Primeiramente, a base de dados será dividida, de forma aleatória, em duas bases (ou grupos) chamadas "Treino"e "Teste". A primeira base, Treino, costuma representar cerca de 70% da totalidade dos dados e é utilizada para o aprendizado e a criação do modelo. Já a segunda base, Teste, costuma representar cerca de 30% da totalidade dos dados e é utilizada para realizar as previsões do modelo, permitindo assim que o desempenho real seja verificado.

Neste trabalho, ressalta-se que a porcentagem utilizada para a base Treino foi de 80% e a para base Teste foi de 20%. Vale também uma análise a respeito dos respectivos sucessos e fracassos referentes a cada base. Esses resultados são apresentados nas tabelas que se seguem.

Tabela 5: Frequência de pessoas segundo presença/ausência de doenças coronarianas - Base Treino

Ausência	Presença
2.481	446

Tabela 6: Frequência de pessoas segundo presença/ausência de doenças coronarianas - Base Teste

Ausência	Presença
620	111

Dessa forma, a proporção de sucessos para a base Treino é de $15,2\%$ em uma população de 31 a 65 anos e a proporção de sucessos para a base Teste também é de $15,2\%$ em uma população de 31 a 65 anos. Uma observação que vale a pena ser destacada, é que já era esperado que essas proporções de sucessos, obtidas para a base Treino e Teste, fossem iguais a obtida pela base completa. Ficando claro que existe uma fidelidade estatística entre as três bases. Ou seja, esse fato ocorre pela ideia de que ao se dividir a base completa, preservam-se nas duas bases resultantes dessa divisão, Treino e Teste, uma proporção de sucessos e fracassos que resultam no final a mesma proporção de sucessos obtida pela base completa.

3.1.2 Ajustando o modelo de regressão logística usual

A análise dos dados feita acima e as que serão realizadas ainda nesta seção, foram feitas utilizando o programa **R**. Dessa forma, inicialmente é encontrado o modelo de regressão logística usual utilizando para o ajustamento todas as variáveis da base de Treino. O modelo foi encontrado utilizando a função *glm*. Após ajustar o modelo de interesse, é frequente querer saber algumas informações a respeito. Essas informações são as estimativas pontuais, os erros padrão referentes as estimativas pontuais, os valores observados da estatística de Wald e os p-valores do teste de Wald. No **Anexo A**, é possível ver essas informações que foram obtidas pela função "$summary()$".

Ao analisar os resultados, é possível verificar que várias das variáveis não se mostram significativas para o modelo. Dessa forma, fazendo uma seleção de variáveis objetivando minimizar o AIC, é possível encontrar o modelo que melhor se ajusta aos dados. Para esse caso da regressão logística usual, é possível utilizar uma função chamada "$stepAIC()$" que realiza essa seleção de variáveis e permite que sejam avaliados os vários modelos obtidos afim de selecionar aquele que melhor se ajusta.

É possível ver os valores do AIC obtidos através da saída da função para os vários modelos na tabela a seguir. No **Anexo B**, estará a saída mais detalhada da função "*stepAIC*()".

Tabela 7: Valores dos AIC dos modelos avaliados com a regressão logística usual

Modelo	AIC
Modelo1	2.250,27
Modelo2	2.248,27
Modelo3	2.246,49
Modelo4	2.244,75
Modelo5	2.243,08
Modelo6	2.241,71
Modelo7	2.240,9
Modelo8	**2.240,69**

O critério *AIC*, foi variando de modelo para modelo, sendo como tal, uma medida para avaliar qual destes é o melhor. Dessa forma, levando em conta os valores apresentados na Tabela 7, que foram obtidos através da função "*stepAIC*()" aplicada a base de Treino, o melhor modelo segundo este critério é o *modelo*8.

Tabela 8: Ajuste do modelo final pelo método usual para os dados da base médica

Variável	Fatores	Estimativa	Razão de Chances	Erro Padrão	p-valor
homem		0,4793	1,6149	0,1204	<0,001
idade		0,0626	1,0646	0,0073	<0,001
educação	Ensino médio completo/GED	-0,3164	0,7287	0,1369	0,0208
	Faculdade incompleta/escola profissional	-0,3992	0,6708	0,1711	0,0196
	Faculdade completa	-0,2573	0,7731	0,1847	0,1635
CPD		0,0201	1,0203	0,0046	<0,001
hipertenso		0,2541	1,2893	0,1519	0,0943
colesterol total		0,0024	1,0024	0,0012	0,0524
PAS		0,0120	1,0120	0,0032	0,0002
glicose		0,0071	1,0072	0,0020	0,0003

Ao analisar as medidas obtidas do modelo final na Tabela 8, pode-se chegar a conclusão de que apesar das variáveis "*hipertenso*" e *"colesterol total"*, possuírem um p-valor acima do $\alpha = 0,05$ que foi estabelecido, elas acabam também se tornando variáveis significativas para

o modelo, pois a presença delas faz com que o "*AIC*" seja o menor possível. Ressalta-se que o *modelo*8 foi obtido através do uso da função "*glm*" na base de Treino. Consequentemente, como a Tabela 8 é o ajuste do modelo final pelo método usual, ela também vem da base de Treino.

Dando seguimento às análises, interpreta-se a razão de chances (RC) obtida para cada variável. Essa razão de chances, é a chance de doença (do evento "desenvolver a doença coronariana") entre indivíduos expostos dividido pela chance de doença entre os não-expostos. Abaixo segue essa interpretação.

- homem: RC $= 1,6149$, então isso quer dizer que a chance de um homem desenvolver a doença é de $61,49\%$ maior do que a chance de uma mulher;
- idade: RC $= 1,0646$, então isso quer dizer que para um certo indivíduo a cada ano de vida a mais, a chance dele ter a doença aumenta em $6,46\%$;
- educação ensino médio completo/GED: RC $= 0,7287$, isso é o mesmo que obter $\frac{1}{0,7287}$, ou seja, o indivíduo com ensino médio incompleto tem a chance de $37,23\%$ de desenvolver a doença quando comparado com um indivíduo com ensino médio completo/GED;
- faculdade incompleta/escola profissional: RC $= 0,6708$, isso é o mesmo que obter $\frac{1}{0,6708}$, ou seja, o indivíduo com ensino médio incompleto tem a chance de $49,07\%$ de desenvolver a doença quando comparado com um indivíduo com faculdade incompleta/escola profissional;
- educação faculdade completa: RC $= 0,7731$, isso é o mesmo que obter $\frac{1}{0,7731}$, ou seja, o indivíduo com ensino médio incompleto tem a chance de $29,34\%$ de desenvolver a doença quando comparado com um indivíduo com faculdade completa;
- CPD: RC $= 1,0203$, então isso quer dizer que um indivíduo aumenta em $2,03\%$ a chance de ter a doença a cada um cigarro a mais que ele fume por dia. Se ele passar a fumar 10 cigarros a mais, a chance dele ter a doença aumenta em $22,26\%$, uma vez que $\exp(10\beta) = 1,222624$;
- hipertenso: RC $= 1,2893$, então isso quer dizer que um indivíduo com hipertensão tem $28,93\%$ a mais de chance de ter a doença quando comparado a outro que não tem hipertensão, considerando os demais fatores de risco não variando de um indivíduo para o outro;

- colesterol total: RC $= 1,0024$, então se aumentar no indivíduo 1 mg/dL no colesterol total ele tem $0,24\%$ a mais de chance de desenvolver a doença em relação a outro indivíduo que no teve aumento no colesterol total;

- PAS: RC $= 1,0120$, então se aumentar no indivíduo 1 mmHg na pressão arterial sistólica (PAS) ele tem $1,20\%$ a mais de chance de desenvolver a doença em relação a outro indivíduo que no teve aumento na pressão arterial sistólica (PAS);

- glicose: RC $= 1,0072$, então se aumentar no indivíduo 1 mg/dL na taxa de glicose ele tem $0,72\%$ a mais de chance de desenvolver a doença em relação a outro indivíduo que não teve aumento na taxa de glicose.

3.1.3 Ajustando o modelo de regressão logística de Firth

As análises dos dados para o caso do modelo de regressão logística de Firth também foram obtidas utilizando o programa **R**. Inicialmente, é encontrado o modelo utilizando para o ajustamento todas as variáveis da base de Treino. O modelo foi encontrado utilizando a função *brglm*. Após ajustar o modelo de interesse, é frequente querer saber algumas informações a respeito dele. Essas informações são: as estimativas pontuais, os erros padrão referentes as estimativas pontuais, os valores observados da estatística de Wald e os p-valores do teste de Wald. No **Anexo C**, é possível ver essas informações que foram obtidas pela função "*summary*()".

Ao analisar os resultados é possível verificar que várias das variáveis não se mostram significativas para o modelo. Dessa forma, fazendo uma seleção de variáveis objetivando minimizar o AIC, é possível encontrar o modelo que melhor se ajusta aos dados. Para o caso da regressão logística de Firth, não é possível utilizar a função "*stepAIC*(). Então, será feito um passo a passo de retirada de variáveis do modelo, seguindo o mesmo princípio da função, até que se encontre o modelo ideal e com o valor do AIC mais baixo.

É possível ver os valores do AIC obtidos através da saída da função para os vários modelos na tabela a seguir. No **Anexo D**, estará a saída mais detalhada desse passo a passo feito.

O critério AIC, foi variando de modelo para modelo, sendo como tal, uma medida para avaliar qual destes é o melhor. Dessa forma, levando em conta os valores apresentados na Tabela 9, que foi obtida através da base Treino, o melhor modelo segundo este critério é o *brglm model8*.

Tabela 9: Valores dos AIC dos modelos avaliados com a regressão logística de Firth

Modelo	AIC
brglm model1	2.250,3
brglm model2	2.248,3
brglm model3	2.246,6
brglm model4	2.244,9
brglm model5	2.243,1
brglm model6	2.241,8
brglm model7	2.240,9
brglm model8	**2.240,7**

Tabela 10: Ajuste do modelo final do método de Firth para os dados da base médica

Variável	Fatores	Estimativa	Razão de Chances	Erro Padrão	p-valor
homem		0,4773	1,6117	0,1200	<0,001
idade		0,0623	1,0643	0,0072	<0,001
educação	Ensino médio completo/GED	-0,3136	0,7308	0,1365	0,0216
	Faculdade incompleta/escola profissional	-0,3931	0,6749	0,1704	0,0210
	Faculdade completa	-0,2500	0,7787	0,1839	0,1739
CPD		0,0200	1,0202	0,0046	<0,001
hipertenso		0,2544	1,2897	0,1515	0,0931
colesterol total		0,0023	1,0024	0,0012	0,0520
PAS		0,0119	1,0120	0,0032	0,0002
glicose		0,0071	1,0071	0,0020	0,0004

Ao analisar as medidas obtidas do modelo final na Tabela 10, pode-se chegar a conclusão de que apesar das variáveis "*hipertenso*" e *"colesterol total"*, possuírem um p-valor acima do $\alpha = 0,05$ que foi estabelecido, elas acabam também se tornando variáveis significativas para o modelo pois a presença delas faz com que o "AIC" seja o menor possível. Vale ressaltar que o *brglm model8* foi obtido através do uso da função $brglm$ na base de Treino. Consequentemente, como a Tabela 10 é o ajuste do modelo final de Firth, ela também é obtida através da base de Treino.

Dando seguimento às análises, interpreta-se a razão de chances (RC) obtida para cada variável. Essa razão de chances, é a chance de doença (do evento "desenvolver a doença coronariana") entre indivíduos expostos dividido pela chance de doença entre os

não-expostos. Abaixo segue essa interpretação.

- homem: RC $= 1,6117$, então isso quer dizer que a chance de um homem desenvolver a doença é de $61,17\%$ maior do que a chance de uma mulher;
- idade: RC $= 1,0643$, então isso quer dizer que para um certo indivíduo a cada ano de vida a mais, a chance dele ter a doença aumenta em $6,43\%$ quando comparado ao ano anterior;
- educação ensino médio completo/GED: RC $= 0,7308$, isso é o mesmo que obter $\frac{1}{0,7308}$, ou seja, o indivíduo com ensino médio incompleto tem a chance de $36,83\%$ de desenvolver a doença quando comparado com um indivíduo com ensino médio completo/GED;
- faculdade incompleta/escola profissional: RC $= 0,6749$, isso é o mesmo que obter $\frac{1}{0,6749}$, ou seja, o indivíduo com ensino médio incompleto tem a chance de $48,17\%$ de desenvolver a doença quando comparado com um indivíduo com faculdade incompleta/escola profissional;
- educação faculdade completa: RC $= 0,7787$, isso é o mesmo que obter $\frac{1}{0,7787}$, ou seja, o indivíduo com ensino médio incompleto tem a chance de $28,41\%$ de desenvolver a doença quando comparado com um indivíduo com faculdade completa;
- CPD: RC $= 1,0202$, então isso quer dizer que um indivíduo aumenta em $2,02\%$ a chance de ter a doença a cada um cigarro a mais que ele fume por dia. Se ele passar a fumar 10 cigarros a mais, a chance dele ter a doença aumenta em $22,14\%$, uma vez que $\exp(10\beta) = 1,221402$;
- hipertenso: RC $= 1,2897$, então isso quer dizer que um indivíduo com hipertensão tem $28,97\%$ a mais de chance de ter a doença quando comparado a outro que não tem hipertensão, considerando os demais fatores de risco não variando de um indivíduo para o outro;
- colesterol total: RC $= 1,0024$, então se aumentar no indivíduo 1 mg/dL no colesterol total ele tem $0,24\%$ a mais de chance de desenvolver a doença em relação a outro indivíduo que no teve aumento no colesterol total;
- PAS: RC $= 1,0120$, então se aumentar no indivíduo 1 mmHg na pressão arterial sistólica (PAS) ele tem $1,20\%$ a mais de chance de desenvolver a doença em relação a outro indivíduo que no teve aumento na pressão arterial sistólica (PAS);

- glicose: RC $= 1,0071$, então se aumentar no indivíduo 1 mg/dL na taxa de glicose ele tem $0,71\%$ a mais de chance de desenvolver a doença em relação a outro indivíduo que no teve aumento na taxa de glicose.

3.1.4 Comparando os dois ajustes

Após serem obtidos os melhores modelos para cada regressão logística, em questão, será feita a comparação entre eles através das seguintes medidas de desempenho:

- sensibilidade;
- especificidade;
- verdadeiro preditivo positivo (VPP);
- verdedeiro preditivo negativo (VPN);
- acurácia;
- AUC.

Essa comparação é feita para saber qual dos dois métodos desempenha o melhor papel neste cenário de eventos raros com a base de dados *Framinghanm Heart Study* (CARROLL et al., 1984). Na Tabela 11, encontra-se a matrize de confusão, que foram obtidas através da base de Teste, respectivas de cada modelo. Na Tabela 12, encontram-se os valores de cada medida de desempenho, obtidas também através da base de Teste, referente a cada modelo. Já na Tabela 13, encontra-se a comparação entre as razões de chance de cada variável do modelo que foram obtidas através da base de Treino.

Tabela 11: Matriz de confusão dos modelos

	Falso	Verdadeiro
0	617	3
1	102	9

Tabela 12: Comparação dos modelos pelas medidas de desempenho

Modelo	Sensibilidade	Especificidade	VPP	VPN	Acurácia	AUC
modelo8	0,75	0,85	0,08	0,995	0,856	0,7525138
brglm model8	0,75	0,85	0,08	0,995	0,856	0,7526155

Tabela 13: Comparação dos modelos pela razão de chances e erro padrão

Variável	Fatores	RC (modelo8)	Erro Padrão (modelo8)	RC (brglm model8)	Erro Padrão (brglm model8)
homem		1,6149	0,1204	1,6117	0,1200
idade		1,0646	0,0073	1,0643	0,0072
educação	ensino médio/GED	0,7287	0,1369	0,7308	0,1365
	alguma faculdade/escola profissional	0,6708	0,1711	0,6749	0,1704
	faculdade	0,7731	0,1847	0,7787	0,1839
CPD		1,0203	0,0046	1,0202	0,0046
hipertenso		1,2893	0,1519	1,2897	0,1515
colesterol total		1,0024	0,0012	1,0024	0,0012
PAS		1,0120	0,0032	1,0120	0,0032
glicose		1,0072	0,0020	1,0071	0,0020

Ao analisar a Tabela 12, destaca-se o baixo valor do VPP. Esse baixo valor indica a conclusão de que ao testar um tipo de indivíduo seriam necessários realizar 100 testes para diagnosticar a doença coronariana em apenas $0,08$ indivíduos, representando assim, um gasto muito grande de recursos e uma relação custo-benefício muito baixa. Em contra partida, por se tratar de um cenário com dados desbalanceados, o valor de VPN é bastante alto, uma vez que, os modelos são capazes de classificar, de forma mais precisa, *verdadeiros negativos* do que os *verdadeiros positivos*.

Feitas as devidas análises, é possível perceber que apesar de serem modelos obtidos por métodos diferentes, um pelo modelo de regressão logística usual e o outro pelo modelo de regressão logística de Firth, pouco se alterou na questão dos valores obtidos pelas medidas. Basicamente, todas as medidas permaneceram as mesmas de um modelo para o outro. Destaca-se que ambos os modelos, tanto o escolhido pela regressão logística usual quanto o escolhido pela regressão logística de Firth, possuem as mesmas variáveis explicativas em sua composição.

Dessa forma, apesar das medidas estarem bem parecidas, deve-se escolher para esse caso com a base de dados médica, o modelo proveniente do método da regressão logística de Firth. Essa escolha decorresse do pequeno aumento que ocorre em sua medida AUC em relação ao modelo obtido pelo método de regressão logística usual.

3.2 Resultados encontrados para a base da área financeira

3.2.1 Descrição dos dados

Afim de aplicar os métodos estudados em dados reais da área financeira, serão considerados os dados obtidos do site kaggle (https://www.kaggle.com/mlg-ulb/creditcardfraud). O interesse do estudo é reconhecer transações fraudulentas com cartão de crédito

Nesta base contém apenas variáveis de entrada numéricas que são o resultado de uma transformação ACP (Análise de Componentes Principais). Infelizmente, devido a questões de confidencialidade, não têm-se os recursos originais e mais informações básicas sobre os dados. Os recursos $V1, V2, \ldots, V28$ são os componentes principais obtidos com o ACP, os únicos recursos que não foram transformados com o ACP são 'Time' e 'Amount'. A variável 'Time' contém os segundos decorridos entre cada transação e a primeira transação no conjunto de dados. A variável 'Amount' é o valor da transação. A variável resposta 'Class', assume o valor 1 em caso de fraude e 0 em caso contrário.

Os modelos foram utilizados para verificar quais fatores de risco influenciam na probabilidade de fraude no cartão de crédito. A variável resposta, Y_i, indica se ocorreu fraude ou não (chamamos de Class) para a i-ésima observação e os fatores de risco (variáveis preditoras). O estudo consiste na observação de 284.807 transações.

Ao serem feitas as primeiras análises, conforme a Tabela 14, pode-se perceber que o número de presença é bem menor do que a quantidade de ausências. Dessa forma, a proporção de sucessos (ou a proporção de uns) é de $0,173\%$, configurando assim, um caso de dados desbalanceados. Vale ressaltar que a Tabela 14 foi obtida através da base de dados completa.

Tabela 14: Frequência dos dados segundo presença/ausência de fraude

Ausência	Presença
284.315	492

Dando seguimento às análises, foi realizada uma descrição mais detalhada a respeito do modelo. Primeiramente, a base de dados será dividida, de forma aleatória, em duas bases (ou grupos) chamadas "Treino"e "Teste". A primeira base, Treino, costuma representar cerca de 70% da totalidade dos dados e é utilizada para o aprendizado e a criação do modelo. Já a segunda base, Teste, costuma representar cerca de 30% da totalidade dos dados e é

utilizada para realizar as previsões do modelo, permitindo assim que o desempenho real seja verificado.

Neste trabalho, evidencia-se que a porcentagem utilizada para a base Treino foi de 80% e a para base Teste foi de 20%. Vale também uma análise a respeito dos respectivos sucessos e fracassos referentes a cada base. Esses resultados são apresentados nas tabelas que se seguem.

Tabela 15: Frequência de pessoas segundo presença/ausência de fraude - Base Treino

Ausência	Presença
227.452	394

Tabela 16: Frequência de pessoas segundo presença/ausência de fraude - Base Teste

Ausência	Presença
56.863	98

Dessa forma, a proporção de sucessos para a base Treino é de $0,173\%$ em uma população de 31 a 65 anos e a proporção de sucessos para a base Teste também é de $0,173\%$ em uma população de 31 a 65 anos. Uma observação que vale a pena ser destacada, é que já era esperado que essas proporções de sucessos, obtidas para a base Treino e Teste, fossem iguais a obtida pela base completa. Ficando claro que existe uma fidelidade estatística entre as três bases. Ou seja, esse fato ocorre pela ideia de ao se dividir a base completa, preservar nas duas bases resultantes dessa divisão, Treino e Teste, uma proporção de sucessos e fracassos que resultasse no final a mesma proporção de sucessos obtida pela base completa.

3.2.2 Ajustando o modelo de regressão logística usual

A análise dos dados feita acima e as que serão realizadas ainda nesta seção, foram feitas utilizando o programa **R**. Dessa forma, inicialmente é encontrado o modelo de regressão logística usual utilizando para o ajustamento todas as variáveis da base de Treino. O modelo foi encontrado utilizando a função *glm*. Após ajustar o modelo de interesse, é frequente querer saber algumas informações a respeito. Essas informações são as estimativas pontuais, os erros padrão referentes as estimativas pontuais, os valores observados da estatística de Wald e os p-valores do teste de Wald. No **Anexo E**, é possível ver essas informações que foram obtidas pela função "*summary*()".

Ao analisar os resultados é possível verificar que várias das variáveis não se mostram significativas para o modelo. Dessa forma, fazendo uma seleção de variáveis objetivando minimizar o AIC, é possível encontrar o modelo que melhor se ajusta aos dados. Para o caso desta base de dados, não é possível utilizar a função "$stepAIC()$ por uma limitação da mesma decorrente do tamanho da base. Então, será feito um passo a passo de retirada de variáveis do modelo, seguindo o mesmo princípio da função, até que se encontre o modelo ideal e com o valor do AIC mais baixo. Assim, foi utilizado um critério que retirou da análise todas as variáveis cujo p-valor foi maior que que $\alpha = 0,20$.

É possível ver os valores do AIC obtidos através do passo a passo feito para os vários modelos na tabela a seguir. No **Anexo F**, estará a saída mais detalhada desse passo a passo de retirada de variáveis que foi feito.

Tabela 17: Valores dos AIC dos modelos avaliados com a regressão logística usual.

Modelo	AIC
modelo1	1.813,6
modelo2	**1.803,8**

O critério AIC, foi variando de modelo para modelo, sendo como tal, uma medida para avaliar qual destes é o melhor. Dessa forma, levando em conta os valores apresentados na Tabela 17, que foi obtida através da base de Treino, o melhor modelo segundo este critério é o $modelo2$.

Na Tabela 18, são apresentadas as variáveis que compõe o $modelo2$ e algumas de suas medidas. Vale ressaltar que o $modelo2$ foi obtido através do uso da função *"glm"* na base de Treino.

Ao analisar as medidas obtidas do modelo final na Tabela 18, pode-se chegar a conclusão também de que todas essas variáveis são significativas para o modelo. Uma vez que, todos os p-valores estão abaixo do $\alpha = 0,20$ que foi estabelecido.

Dando seguimento às análises, interpreta-se a razão de chances (RC) obtida para as variáveis "Time"e "Amount". Uma vez que, as demais variáveis que compõe o modelo foram obtidas através de ACP, elas perdem sua capacidade de interpretação do modelo. Essa razão de chances, é a chance de fruade (do evento "ocorreu uma fraude de cartão de crédito") entre indivíduos que foram expostos a esse caso dividido pela chance de indivíduos não-expostos. Abaixo segue essa interpretação.

- Time: RC = 1, então essa variável não está relacionada com a questão da fraude de

Tabela 18: Variáveis do modelo2

Variável	Estimativa	Razão de Chances	Erro Padrão	p-valor
Time	-4.717e-06	1	1.940e-06	0.015047
V1	1.271e-01	1.135	4.130e-02	0.002087
V4	7.280e-01	2.071	6.989e-02	<0,001
V5	1.212e-01	1.128	3.529e-02	0.000593
V7	-2.052e-01	0.814	6.311e-02	0.001145
V8	-1.421e-01	0.87	2.418e-02	<0,001
V9	-1.801e-01	0.835	9.298e-02	0.052741
V10	-9.136e-01	0.401	9.206e-02	<0,001
V13	-3.205e-01	0.726	8.612e-02	0.000198
V14	-5.428e-01	0.581	5.309e-02	<0,001
V20	-4.587e-01	0.632	8.102e-02	<0,001
V21	4.415e-01	1.555	5.620e-02	<0,001
V22	7.376e-01	2.09	1.316e-01	<0,001
V23	-1.219e-01	0.885	6.249e-02	<0,001
V27	-8.654e-01	0.421	1.207e-01	<0,001
V28	-3.484e-01	0.706	1.122e-01	0.001909
Amount	9.550e-04	1.001	3.417e-04	0.005185

cartão de crédito;

- Amount: RC = 1.128, então se aumentar em 1 unidade o Amount o tem-se $12,8\%$ a mais chance de ocorrer a fraude no cartão de crédito em relação a não se ter esse aumento no Amount.

3.2.3 Ajustando o modelo de regressão logística de Firth

As análises dos dados para o caso do modelo de regressão logística de Firth também foram obtidas utilizando o programa **R**. Assim, inicialmente é encontrado o modelo utilizando para o ajustamento todas as variáveis da base de Treino. O modelo foi encontrado utilizando a função *brglm*. Após ajustar o modelo de interesse, é frequente querer saber algumas informações a respeito. Essas informações são as estimativas pontuais, os erros padrão referentes as estimativas pontuais, os valores observados da estatística de Wald e os p-valores do teste de Wald. No **Anexo G**, é possível ver essas informações que foram obtidas pela função "$summary()$".

Ao analisar os resultados é possível verificar que várias das variáveis não se mostram significativas para o modelo. Dessa forma, fazendo uma seleção de variáveis objetivando minimizar o AIC, é possível encontrar o modelo que melhor se ajusta aos dados. Para o caso desta base de dados, não é possível utilizar a função "$stepAIC()$ por uma limitação

da mesma decorrente do tamanho da base. Então, será feito um passo a passo de retirada de variáveis do modelo, seguindo o mesmo princípio da função, até que se encontre o modelo ideal e com o valor do AIC mais baixo. Portanto, foi utilizado um critério que retirou da análise todas as variáveis cujo p-valor foi maior que que $\alpha = 0,20$.

É possível ver os valores do AIC obtidos através do passo a passo feito para os vários modelos na tabela a seguir. No **Anexo H**, estará a saída mais detalhada desse passo a passo de retirada de variáveis que foi feito.

Tabela 19: Valores dos AIC dos modelos avaliados com a regressão logística de Firth.

Modelo	AIC
brglm model1	1.816,5
brglm model2	**1.805,8**

O critério AIC, foi variando de modelo para modelo, sendo como tal, uma medida para avaliar qual destes é o melhor. Dessa forma, levando em conta os valores apresentados na Tabela 19, que foi obtida através da base de Treino, o melhor modelo segundo este critério é o *"brglm model2"*.

Na Tabela 20, são apresentadas as variáveis que compõe o *"brglm model2"* e algumas de suas medidas. Vale ressaltar que o $modelo2$ foi obtido através do uso da função *"brglm"* na base de Treino.

Ao analisar as medidas obtidas do modelo final na Tabela 16, pode-se chegar a conclusão também de que todas essas variáveis são significativas para o modelo. Uma vez que, todos os p-valores estão abaixo do $\alpha = 0,20$ que foi estabelecido.

Dando seguimento às análises, interpreta-se a razão de chances (RC) obtida para as variáveis "Time"e "Amount". Uma vez que, as demais variáveis que compõe o modelo foram obtidas através de ACP, elas perdem sua capacidade de interpretação do modelo. Essa razão de chances, é a chance de fraude (do evento "ocorreu uma fraude de cartão de crédito") entre indivíduos que foram expostos a esse caso dividido pela chance de indivíduos não-expostos. Abaixo segue essa interpretação.

- Time: RC = 1, então essa variável não está relacionada com a questão da fraude de cartão de crédito;
- Amount: RC = 1.001, então se aumentar em 1 unidade o Amount o tem-se $0,1\%$ a mais chance de ocorrer a fraude no cartão de crédito em relação a não se ter esse aumento no Amount..

Tabela 20: Variáveis do brglm model2

Variável	Estimativa	Razão de Chances	Erro Padrão	p-valor
Time	-4.600e-06	1	1.898e-06	0.015339
V1	1.151e-01	1.222	3.677e-02	0.001743
V4	7.301e-01	2.075	6.563e-02	<0,001
V5	1.300e-01	1.138	2.940e-02	<0,001
V7	-2.093e-01	0.811	5.104e-02	<0,001
V8	-1.484e-01	0.862	2.118e-02	<0,001
V9	-1.706e-01	0.843	8.766e-02	0.051618
V10	-9.059e-01	0.404	8.671e-02	<0,001
V13	-3.145e-01	0.73	8.470e-02	0.000205
V14	-5.373e-01	0.584	4.989e-02	<0,001
V20	-4.766e-01	0.621	7.161e-02	<0,001
V21	4.286e-01	1.535	5.294e-02	<0,001
V22	7.261e-01	2.07	1.285e-01	<0,001
V23	-1.281e-01	0.879	4.695e-02	0.006366
V27	-8.606e-01	0.423	1.135e-01	<0,001
V28	-3.093e-01	0.734	7.886e-02	<0,001
Amount	1.000e-03	1.001	2.689e-04	0.000200

3.2.4 Comparando os dois ajustes

Após serem obtidos os melhores modelos para cada regressão logística, em questão, será feita a comparação entre eles através das seguintes medidas de desempenho:

- sensibilidade;
- especificidade;
- verdadeiro preditivo positivo (VPP);
- verdadeiro preditivo negativo (VPN);
- acurácia;
- AUC.

Essa comparação é feita para saber qual dos dois métodos desempenha o melhor papel neste cenário de eventos raros com a base de dados do kaggle (https://www.kaggle.com/mlg-ulb/creditcardfraud). Na Tabela 21, encontra-se a matriz de confusão, obtida através da base de Teste, respectiva ao cada modelo. Na Tabela 22, encontram-se os valores de cada medida de desempenho, obtidas também através da base de Teste, referente a cada modelo.

Já na Tabela 23, encontra-se a comparação entre as razões de chance de cada variável do modelo.

Tabela 21: Matriz de confusão dos modelos

	Falso	Verdadeiro
0	56.844	19
1	30	68

Tabela 22: Comparação dos modelos pelas medidads de desempenho

Modelo	Sensibilidade	Especificidade	VPP	VPN	Acurácia	AUC
modelo2	0,782	0,994	0,694	0,9996	0,9991	0,9744238
brglm model2	0,782	0,994	0,694	0,9996	0,9991	0,9750403

Feitas as devidas análises, é possível perceber que apesar de serem modelos obtidos por métodos diferentes, um pelo modelo de regressão logística usual e o outro pelo modelo de regressão logística de Firth, pouco se alterou na questão dos valores obtidos pelas medidas. Basicamente, todas as medidas permaneceram as mesmas de um modelo para o outro. Vale ressaltar que ambos os modelos, tanto o escolhido pela regressão logística usual quanto o escolhido pela regressão logística de Firth, possuem as mesmas variáveis explicativas em sua composição.

Portanto, apesar das medidas estarem bem parecidas, deve-se escolher para esse caso com a base de dados financeira, o modelo proveniente do método da regressão logística de Firth. Essa escolha decorresse ao pequeno aumento que ocorre em sua medida AUC em relação ao modelo obtido pelo método de regressão logística usal.

Tabela 23: Comparação dos modelos pela razão de chances e erro padrão

Variável	RC (modelo2)	Erro Padrão (modelo2)	RC (brglm model2)	Erro Padrão (brglm model2)
Time	1	0,00000194	1	0,000001898
V1	1,135	0,0413	1,222	0,03677
V4	2,071	0,06989	2,075	0,06563
V5	1,128	0,03529	1,138	0,02940
V7	0,814	0,06311	0,811	0,05104
V8	0,87	0,02418	0,862	0,02118
V9	0,835	0,09298	0,843	0,09766
V10	0,401	0,09206	0,404	0,08671
V13	0,726	0,08612	0,73	0,08470
V14	0,581	0,05209	0,584	0,04989
V20	0,632	0,08102	0,621	0,07161
V21	1,555	0,05620	1,535	0,05294
V22	2,09	0,1316	2,07	0,1285
V23	0,885	0,06249	0,879	0,04695
V27	0,421	0,1207	0,423	0,1135
V28	0,706	0,1122	0,734	0,07886
Amount	1,001	0,0003417	1,011	0,0002689

4 Conclusões

O foco principal deste trabalho foi o estudo da regressão logística no contexto de eventos raros. A ideia era estudar a metodologia para modelagem de dados binários e investigar uma alternativa de modelagem no cenário com a presença de eventos raros.

Neste sentido foram estudados e comparados dois modelos de regressão logística. Os dois modelos abordados foram: o modelo de regressão logística usual e o modelo de regressão logística de Firth. Essa comparação foi realizada utilizando-se dois conjuntos de dados reais, um da área média e outro da área financeira.

No que diz respeito à comparação entre os modelos, era de se esperar que o modelo de regressão logística usual apresentasse resultados piores quando comparados com a abordagem de Firth na presença de eventos raros. Isso porque, como King G.; Zeng (2001b) já tinham relatado em seu texto, o modelo de regressão logística, pode subestimar drasticamente a probabilidade de eventos raros.

Neste trabalho, foi observado uma equivalência entre os dois modelos estudados. Apesar de se considerar dois cenários com porcentagens diferentes de eventos raros, pouca diferença foi observada entre as duas abordagens.

A principal diferença observada entre as medidas analisadas diz respeito a área sob a curva (AUC) dos modelos. Investigações futuras e mais detalhadas precisam ser feitas.

Algumas considerações merecem destaque. Por exemplo, as acurácias pelos dois modelos em ambos os cenários foram acima de 80%. Entretando, o VPP na base da área da saúde foi extremamente baixo.

Logo, e ainda que os resultados não tenham sido os melhores, os objetivos propostos para este presente trabalho foram cumpridos, uma vez que os objetivos inicialmente listados foram abordados.

4.1 Trabalhos futuros

Como trabalhos futuros, entende-se que algumas linhas podem ser seguidas para esse assunto aqui abordado. São elas:

- realizar um estudo mais aprofundado a respeito da regressão logística de Firth;
- realizar um estudo simulado para à comparação de vários modelos de regressão logística;
- aplicar a metodologia estudada em outros dados da área médica no Brasil;
- aplicar a metodologia estudada na base médica fazendo a ACP para saber se assim o VPP melhora.

Referências

AGRESTI, A. Categorical data analysis (2nd ed). New Jersey: John Wiley and Sons, 2002.

AKAIKE, H. A new look at the statistical model identification. IEEE Transactions on Automatic Control, v. 19, p. 716–723, 1974.

CARROLL, R. J. et al. On errors-in-variables for binary regression models. Biometrika, v. 71, p. 19–26, 1984.

CLAYTON, D.; HILLS, M. Statistical models in epidemiology. Oxford, p. 384, 2013.

DOBSON, A. J.; BARNETT, A. G. An introduction to generalized linear models (3rd ed). CHAPMAN and HALL/CRC, 2018.

FIRTH, D. Bias reduction of maximum likelihood estimates. Biometrika, v. 80, n. 1, p. 27–38, 1993.

HEINZE, G.; SCHEMPER, M. A solution to the problem of separation in logistic regression. Section of Clinical Biometrics; Department of Medical Computer Sciences; University of Vienna, p. 01–11, 2002.

HOSMER, D.; LEMESHOW, S. Applied logistic regression. Wiley Series in Probability and Statistics, p. 01–10, 2000.

HUAYANAY, A. Modelos de regressÃo para resposta binÁria na presenÇa de dados desbalanceados. Universidade Federal de São Carlos, p. 33–50, 2019.

KING G.; ZENG, L. Explaining rare events in international relations. International Organization, v. 55, n. 3, p. 693–715, 2001.

KING G.; ZENG, L. Logistic regression in rare events data. Cambridge University Press, v. 9, n. 2, p. 137–163, 2001.

MASCARENHAS C. H. M; REIS, L. A. S. M. S. Avaliação do risco de doença coronariana em adultos e idosos no município de lagêado do tabocal/ba. Arq. Ciêne. Saúde Unipar, Umuarama, 2009.

MCCULLAGH, P. Tensor methods in statistics. Chapman and Hall, 1987.

MCCULLAGH, P.; NELDER, J. A. Generalized linear models (2nd ed). Chapman and Hall/CRC, v. 2, 1989.

NELDER, J. A.; WEDDERBURN, R. W. Generalized linear models. *Journal of the Royal Statistical Society: Series A (General)*, Wiley Online Library, v. 135, n. 3, p. 370–384, 1972.

PAAL, B. V. d. A comparison of different methods for modelling rare events data. Ghent University, p. 75, 2014.

SANTOS, R. dos. Regressão logística em dados com eventos raros. Faculdade de Ciências Universidade do Porto, p. 27–38, 2017.

SCHWARZ, G. Estimating the dimensional of a model. Annals of Statistics, v. 6, p. 461–464, 1978.

TURKMAN, M. A.; SILVA, G. Modelos lineares generalizados - da teoria à prática. UL and UTL: Lisboa, 2000.

5 Anexos

5.1 Anexo A - Sumário do modelo 1 (glm) base médica

```
> summary(modelo1)

Call:
glm(formula = TenYearCHD ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.7896  -0.5977  -0.4251  -0.2856   2.8279

Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.891310   0.789357  -9.997  < 2e-16 ***
maleTRUE           0.481721   0.122395   3.936 8.29e-05 ***
age                0.061612   0.007532   8.180 2.83e-16 ***
education2        -0.304740   0.138188  -2.205  0.02744 *
education3        -0.380501   0.172581  -2.205  0.02747 *
education4        -0.248696   0.186288  -1.335  0.18187
currentSmokerTRUE  0.107326   0.173633   0.618  0.53650
cigsPerDay         0.017873   0.006930   2.579  0.00991 **
BPMedsTRUE         0.302236   0.259205   1.166  0.24361
prevalentStrokeTRUE 0.571425  0.522648   1.093  0.27425
prevalentHypTRUE   0.239382   0.156012   1.534  0.12494
diabetesTRUE      -0.003528   0.351650  -0.010  0.99200
totChol            0.002400   0.001243   1.931  0.05344 .
sysBP              0.012897   0.004293   3.004  0.00266 **
```

```
diaBP            −0.003364   0.007204  −0.467  0.64049
BMI               0.008459   0.014237   0.594  0.55241
heartRate        −0.003715   0.004662  −0.797  0.42553
glucose           0.007173   0.002644   2.713  0.00667 **
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2498.5  on 2926  degrees of freedom
Residual deviance: 2214.3  on 2909  degrees of freedom
AIC: 2250.3

Number of Fisher Scoring iterations: 5
```

5.2 Anexo B - Saída da função stepAIC(modelo1) base médica

```
> stepAIC(modelo1)
Start:  AIC=2250.27
TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + diabetes + totChol +
    sysBP + diaBP + BMI + heartRate + glucose

                 Df Deviance    AIC
- diabetes        1   2214.3 2248.3
- diaBP           1   2214.5 2248.5
- BMI             1   2214.6 2248.6
- currentSmoker   1   2214.7 2248.7
- heartRate       1   2214.9 2248.9
- prevalentStroke 1   2215.4 2249.4
- BPMeds          1   2215.6 2249.6
<none>                2214.3 2250.3
- prevalentHyp    1   2216.6 2250.6
```

```
- totChol          1    2218.0 2252.0
- education       3    2222.2 2252.2
- cigsPerDay      1    2220.8 2254.8
- glucose         1    2221.9 2255.9
- sysBP           1    2223.3 2257.3
- male            1    2229.8 2263.8
- age             1    2283.3 2317.3

Step:  AIC=2248.27
TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + totChol + sysBP +
    diaBP + BMI + heartRate + glucose

                 Df Deviance    AIC
- diaBP           1    2214.5 2246.5
- BMI             1    2214.6 2246.6
- currentSmoker   1    2214.7 2246.7
- heartRate       1    2214.9 2246.9
- prevalentStroke 1    2215.4 2247.4
- BPMeds          1    2215.6 2247.6
<none>                 2214.3 2248.3
- prevalentHyp    1    2216.6 2248.6
- totChol         1    2218.0 2250.0
- education       3    2222.2 2250.2
- cigsPerDay      1    2220.8 2252.8
- sysBP           1    2223.3 2255.3
- glucose         1    2226.6 2258.6
- male            1    2229.8 2261.8
- age             1    2283.3 2315.3

Step:  AIC=2246.49
TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + totChol + sysBP +
    BMI + heartRate + glucose
```

```
                 Df Deviance    AIC
- BMI             1   2214.8 2244.8
- currentSmoker   1   2214.9 2244.9
- heartRate       1   2215.2 2245.2
- prevalentStroke 1   2215.6 2245.6
- BPMeds          1   2215.8 2245.8
<none>                2214.5 2246.5
- prevalentHyp    1   2216.7 2246.7
- totChol         1   2218.2 2248.2
- education      3   2222.7 2248.7
- cigsPerDay      1   2221.0 2251.0
- sysBP           1   2226.6 2256.6
- glucose         1   2227.1 2257.1
- male            1   2229.8 2259.8
- age             1   2287.7 2317.7

Step:  AIC=2244.75
TenYearCHD ~ male + age + education + currentSmoker + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + totChol + sysBP +
    heartRate + glucose

                 Df Deviance    AIC
- currentSmoker   1   2215.1 2243.1
- heartRate       1   2215.4 2243.4
- prevalentStroke 1   2215.9 2243.9
- BPMeds          1   2216.1 2244.1
<none>                2214.8 2244.8
- prevalentHyp    1   2217.1 2245.1
- totChol         1   2218.5 2246.5
- education      3   2223.5 2247.5
- cigsPerDay      1   2221.4 2249.4
- glucose         1   2227.6 2255.6
- sysBP           1   2227.7 2255.7
- male            1   2230.2 2258.2
- age             1   2287.8 2315.8
```

```
Step:  AIC=2243.08
TenYearCHD ~ male + age + education + cigsPerDay + BPMeds + prevalentStro
    prevalentHyp + totChol + sysBP + heartRate + glucose

                 Df Deviance    AIC
- heartRate       1   2215.7 2241.7
- prevalentStroke 1   2216.2 2242.2
- BPMeds          1   2216.5 2242.5
<none>                2215.1 2243.1
- prevalentHyp    1   2217.4 2243.4
- totChol         1   2218.8 2244.8
- education      3   2223.8 2245.8
- glucose         1   2227.9 2253.9
- sysBP           1   2227.9 2253.9
- male            1   2230.6 2256.6
- cigsPerDay      1   2234.1 2260.1
- age             1   2287.8 2313.8

Step:  AIC=2241.71
TenYearCHD ~ male + age + education + cigsPerDay + BPMeds + prevalentStro
    prevalentHyp + totChol + sysBP + glucose

                 Df Deviance    AIC
- prevalentStroke 1   2216.9 2240.9
- BPMeds          1   2217.2 2241.2
<none>                2215.7 2241.7
- prevalentHyp    1   2217.9 2241.9
- totChol         1   2219.3 2243.3
- education      3   2224.4 2244.4
- glucose         1   2228.1 2252.1
- sysBP           1   2228.1 2252.1
- male            1   2232.1 2256.1
- cigsPerDay      1   2234.1 2258.1
- age             1   2289.9 2313.9
```

```
Step:  AIC=2240.9
TenYearCHD ~ male + age + education + cigsPerDay + BPMeds + prevalentHyp
    totChol + sysBP + glucose

               Df Deviance    AIC
- BPMeds        1   2218.7 2240.7
<none>              2216.9 2240.9
- prevalentHyp  1   2219.2 2241.2
- totChol       1   2220.5 2242.5
- education     3   2225.8 2243.8
- sysBP         1   2229.3 2251.3
- glucose       1   2229.4 2251.4
- male          1   2233.2 2255.2
- cigsPerDay    1   2235.1 2257.1
- age           1   2291.6 2313.6

Step:  AIC=2240.69
TenYearCHD ~ male + age + education + cigsPerDay + prevalentHyp +
    totChol + sysBP + glucose

               Df Deviance    AIC
<none>              2218.7 2240.7
- prevalentHyp  1   2221.5 2241.5
- totChol       1   2222.4 2242.4
- education     3   2227.6 2243.6
- glucose       1   2231.4 2251.4
- sysBP         1   2232.1 2252.1
- male          1   2234.6 2254.6
- cigsPerDay    1   2236.7 2256.7
- age           1   2294.4 2314.4

Call:  glm(formula = TenYearCHD ~ male + age + education + cigsPerDay +
    prevalentHyp + totChol + sysBP + glucose, family = binomial,
    data = train)
```

```
Coefficients:
     (Intercept)          maleTRUE               age         education2
education3
      −8.119562          0.479334          0.062663          −0.316434
−0.399265
       education4       cigsPerDay  prevalentHypTRUE           totChol
sysBP
       −0.257338          0.020128          0.254144          0.002401
0.012008
         glucose
        0.007186

Degrees of Freedom: 2926 Total (i.e. Null);  2916 Residual
Null Deviance:      2499
Residual Deviance: 2219         AIC: 2241
```

5.3 Anexo C - Sumário brglm base médica

```
> summary(brglm_model)

Call:
brglm(formula = TenYearCHD ~ ., family = "binomial", data = train)


Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       −7.858832   0.791302  −9.932  < 2e−16 ***
maleTRUE           0.504092   0.120995   4.166  3.1e−05 ***
age                0.062803   0.007412   8.474  < 2e−16 ***
education         −0.121224   0.055917  −2.168  0.03016 *
currentSmokerTRUE  0.104432   0.173064   0.603  0.54622
cigsPerDay         0.017803   0.006919   2.573  0.01009 *
BPMedsTRUE         0.308524   0.258122   1.195  0.23198
```

```
prevalentStrokeTRUE   0.583261   0.518808   1.124  0.26091
prevalentHypTRUE      0.239078   0.155199   1.540  0.12345
diabetesTRUE          0.037451   0.349180   0.107  0.91459
totChol               0.002332   0.001235   1.889  0.05889 .
sysBP                 0.012794   0.004273   2.994  0.00275 **
diaBP                -0.003380   0.007170  -0.471  0.63737
BMI                   0.010398   0.014157   0.734  0.46266
heartRate            -0.003937   0.004643  -0.848  0.39649
glucose               0.006817   0.002632   2.590  0.00959 **
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2428.9  on 2926  degrees of freedom
Residual deviance: 2217.4  on 2911  degrees of freedom
Penalized deviance: 2100.083
AIC:  2249.4
```

5.4 Anexo D - Passo a passo da função 'stepAIC()' para o modelo brglm base médica

```
> brglm_model1=brglm(TenYearCHD ~ male + age + education + currentSmoker
                     BPMeds + prevalentStroke + prevalentHyp + totChol +
                     diaBP + BMI + heartRate + glucose, data = train,
                   family = binomial)
> summary(brglm_model1)

Call:
brglm(formula = TenYearCHD ~ male + age + education + currentSmoker +
    cigsPerDay + BPMeds + prevalentStroke + prevalentHyp + totChol +
    sysBP + diaBP + BMI + heartRate + glucose, family = binomial,
    data = train)
```

```
Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)        -7.873102   0.780100 -10.092  < 2e-16 ***
maleTRUE            0.504441   0.120979   4.170 3.05e-05 ***
age                 0.062809   0.007413   8.473  < 2e-16 ***
education          -0.121349   0.055929  -2.170 0.030030 *
currentSmokerTRUE   0.104216   0.173038   0.602 0.546994
cigsPerDay          0.017807   0.006920   2.573 0.010080 *
BPMedsTRUE          0.309632   0.258040   1.200 0.230163
prevalentStrokeTRUE 0.582962   0.518688   1.124 0.261048
prevalentHypTRUE    0.239234   0.155155   1.542 0.123097
totChol             0.002334   0.001234   1.891 0.058607 .
sysBP               0.012805   0.004273   2.997 0.002729 **
diaBP              -0.003393   0.007168  -0.473 0.635962
BMI                 0.010448   0.014152   0.738 0.460336
heartRate          -0.003959   0.004643  -0.853 0.393891
glucose             0.006988   0.002039   3.428 0.000609 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2431.9  on 2926  degrees of freedom
Residual deviance: 2217.4  on 2912  degrees of freedom
Penalized deviance: 2102.189
AIC:  2247.4
```

```
> brglm_model2=brglm(TenYearCHD ~ male + age + education + currentSmoker
+                    BPMeds + prevalentStroke + prevalentHyp + totChol
+                    BMI + heartRate + glucose, data = train, family=bi
> summary(brglm_model2)
```

```
Call:
brglm(formula = TenYearCHD ~ male + age + education + currentSmoker +
    cigsPerDay + BPMeds + prevalentStroke + prevalentHyp + totChol +
    sysBP + BMI + heartRate + glucose, family = binomial, data = train)


Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.980151   0.748889 -10.656  < 2e-16 ***
maleTRUE           0.498038   0.120201   4.143 3.42e-05 ***
age                0.063541   0.007268   8.742  < 2e-16 ***
education         -0.123527   0.055759  -2.215 0.026734 *
currentSmokerTRUE  0.107079   0.172944   0.619 0.535817
cigsPerDay         0.017769   0.006922   2.567 0.010256 *
BPMedsTRUE         0.310317   0.257890   1.203 0.228863
prevalentStrokeTRUE 0.578362  0.518393   1.116 0.264558
prevalentHypTRUE   0.229083   0.153685   1.491 0.136065
totChol            0.002342   0.001235   1.897 0.057839 .
sysBP              0.011539   0.003335   3.460 0.000540 ***
BMI                0.009100   0.013853   0.657 0.511257
heartRate         -0.004064   0.004640  -0.876 0.381205
glucose            0.007060   0.002033   3.473 0.000514 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2435.5  on 2926  degrees of freedom
Residual deviance: 2217.6  on 2913  degrees of freedom
Penalized deviance: 2112.285
AIC:  2245.6


> brglm_model3=brglm(TenYearCHD ~ male + age + education + cigsPerDay + B
```

```
+                       prevalentStroke + prevalentHyp + totChol + sysBP +
+                       heartRate + glucose, data = train, family=binomial
> summary(brglm_model3)

Call:
brglm(formula = TenYearCHD ~ male + age + education + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + totChol + sysBP +
    BMI + heartRate + glucose, family = binomial, data = train)


Coefficients:
                    Estimate Std. Error z value Pr(>|z|)
(Intercept)       -7.919855   0.741568 -10.680  < 2e-16 ***
maleTRUE           0.498543   0.120289   4.145 3.40e-05 ***
age                0.063260   0.007252   8.723  < 2e-16 ***
education         -0.123319   0.055763  -2.211 0.027003 *
cigsPerDay         0.020881   0.004744   4.401 1.08e-05 ***
BPMedsTRUE         0.313940   0.257650   1.218 0.223043
prevalentStrokeTRUE 0.575517   0.518056   1.111 0.266604
prevalentHypTRUE   0.228149   0.153645   1.485 0.137567
totChol            0.002332   0.001235   1.889 0.058950 .
sysBP              0.011567   0.003335   3.469 0.000523 ***
BMI                0.008009   0.013752   0.582 0.560305
heartRate         -0.003994   0.004639  -0.861 0.389193
glucose            0.007042   0.002033   3.463 0.000534 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2439.5  on 2926  degrees of freedom
Residual deviance: 2218.0  on 2914  degrees of freedom
Penalized deviance: 2116.162
AIC:  2244
```

```
> brglm_model4=brglm(TenYearCHD ~ male + age + education + cigsPerDay + B
+                    prevalentStroke + prevalentHyp + totChol + sysBP +
+                    glucose, data = train, family=binomial)
> summary(brglm_model4)

Call:
brglm(formula = TenYearCHD ~ male + age + education + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + totChol + sysBP +
    heartRate + glucose, family = binomial, data = train)


Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)      -7.751801   0.679392 -11.410  < 2e-16 ***
maleTRUE          0.502205   0.120188   4.178 2.93e-05 ***
age               0.063044   0.007242   8.706  < 2e-16 ***
education        -0.127163   0.055405  -2.295 0.021724 *
cigsPerDay        0.020653   0.004731   4.365 1.27e-05 ***
BPMedsTRUE        0.318506   0.257566   1.237 0.216235
prevalentStrokeTRUE 0.579497   0.516558   1.122 0.261929
prevalentHypTRUE  0.237064   0.152930   1.550 0.121107
totChol           0.002354   0.001234   1.907 0.056463 .
sysBP             0.011865   0.003297   3.599 0.000320 ***
heartRate        -0.003944   0.004637  -0.850 0.395056
glucose           0.007119   0.002034   3.500 0.000466 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2443.3  on 2926  degrees of freedom
Residual deviance: 2218.3  on 2915  degrees of freedom
Penalized deviance: 2125.062
```

```
AIC:  2242.3
```

```
> brglm_model5=brglm(TenYearCHD ~ male + age + education + cigsPerDay + B
+                      prevalentStroke + prevalentHyp + totChol + sysBP +
+                    data = train, family=binomial)
> summary(brglm_model5)

Call:
brglm(formula = TenYearCHD ~ male + age + education + cigsPerDay +
    BPMeds + prevalentStroke + prevalentHyp + totChol + sysBP +
    glucose, family = binomial, data = train)


Coefficients:
                   Estimate Std. Error z value Pr(>|z|)
(Intercept)       -8.018424   0.606889 -13.212  < 2e-16 ***
maleTRUE           0.514312   0.119403   4.307 1.65e-05 ***
age                0.063527   0.007222   8.797  < 2e-16 ***
education         -0.125446   0.055411  -2.264 0.023578 *
cigsPerDay         0.020137   0.004690   4.293 1.76e-05 ***
BPMedsTRUE         0.329305   0.257212   1.280 0.200444
prevalentStrokeTRUE 0.594335   0.516417   1.151 0.249781
prevalentHypTRUE   0.228225   0.152559   1.496 0.134661
totChol            0.002307   0.001232   1.872 0.061229 .
sysBP              0.011582   0.003280   3.531 0.000414 ***
glucose            0.006988   0.002029   3.445 0.000571 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2447.3  on 2926  degrees of freedom
Residual deviance: 2219.1  on 2916  degrees of freedom
```

```
Penalized deviance: 2136.556
AIC:  2241.1
```

```
> brglm_model6=brglm(TenYearCHD ~ male + age + education + cigsPerDay + B
+                          prevalentHyp + totChol + sysBP + glucose,
+                        data = train, family=binomial)
> summary(brglm_model6)

Call:
brglm(formula = TenYearCHD ~ male + age + education + cigsPerDay +
    BPMeds + prevalentHyp + totChol + sysBP + glucose, family = binomial,
    data = train)


Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.021227   0.606848 -13.218  < 2e-16 ***
maleTRUE         0.514340   0.119352   4.309 1.64e-05 ***
age              0.063667   0.007217   8.822  < 2e-16 ***
education       -0.127678   0.055372  -2.306 0.021120 *
cigsPerDay       0.020025   0.004689   4.271 1.95e-05 ***
BPMedsTRUE       0.358588   0.255460   1.404 0.160409
prevalentHypTRUE 0.234010   0.152373   1.536 0.124595
totChol          0.002307   0.001232   1.872 0.061138 .
sysBP            0.011577   0.003278   3.532 0.000413 ***
glucose          0.007026   0.002030   3.460 0.000540 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2449.6  on 2926  degrees of freedom
Residual deviance: 2220.3  on 2917  degrees of freedom
```

```
Penalized deviance: 2139.058
AIC:  2240.3
```

```
> brglm_model7=brglm(TenYearCHD ~ male + age + education + cigsPerDay + p
+                    totChol + sysBP+ glucose, data = train, family=bin
> summary(brglm_model7)

Call:
brglm(formula = TenYearCHD ~ male + age + education + cigsPerDay +
    prevalentHyp + totChol + sysBP + glucose, family = binomial,
    data = train)


Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.097226   0.604179 -13.402  < 2e-16 ***
maleTRUE         0.506614   0.119090   4.254 2.10e-05 ***
age              0.064036   0.007212   8.880  < 2e-16 ***
education       -0.128092   0.055365  -2.314 0.020690 *
cigsPerDay       0.019920   0.004689   4.248 2.15e-05 ***
prevalentHypTRUE 0.254710   0.151277   1.684 0.092233 .
totChol          0.002344   0.001231   1.904 0.056902 .
sysBP            0.011998   0.003268   3.671 0.000241 ***
glucose          0.007052   0.002022   3.487 0.000488 ***
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2453.0  on 2926  degrees of freedom
Residual deviance: 2222.1  on 2918  degrees of freedom
Penalized deviance: 2143.636
AIC:  2240.1
```

5.5 Anexo E - Sumário do modelo 1 (glm) base financeira

```
> summary(modelo1)

Call:
glm(formula = Class ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-5.1778  -0.0287  -0.0191  -0.0121   4.6359

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.391e+00  2.798e-01 -29.986  < 2e-16 ***
Time        -4.173e-06  2.527e-06  -1.651 0.098638 .
V1           1.249e-01  4.855e-02   2.572 0.010116 *
V2          -2.820e-03  6.422e-02  -0.044 0.964973
V3           1.748e-02  5.964e-02   0.293 0.769415
V4           7.422e-01  8.466e-02   8.767  < 2e-16 ***
V5           1.003e-01  7.458e-02   1.344 0.178880
V6          -1.057e-01  8.205e-02  -1.288 0.197841
V7          -1.403e-01  7.504e-02  -1.870 0.061436 .
V8          -1.834e-01  3.427e-02  -5.352 8.71e-08 ***
V9          -2.523e-01  1.252e-01  -2.015 0.043907 *
V10         -8.888e-01  1.109e-01  -8.016 1.09e-15 ***
V11          6.432e-03  9.101e-02   0.071 0.943659
V12          2.839e-02  9.468e-02   0.300 0.764337
V13         -3.336e-01  8.959e-02  -3.724 0.000196 ***
V14         -5.329e-01  6.776e-02  -7.864 3.71e-15 ***
V15         -9.570e-02  9.676e-02  -0.989 0.322621
V16         -1.556e-01  1.435e-01  -1.084 0.278361
V17          2.165e-03  7.857e-02   0.028 0.978016
V18         -2.609e-02  1.466e-01  -0.178 0.858767
V19          1.029e-01  1.099e-01   0.936 0.349133
```

```
V20          -4.664e-01   9.292e-02   -5.020  5.17e-07 ***
V21           3.579e-01   6.706e-02    5.338  9.41e-08 ***
V22           5.286e-01   1.500e-01    3.524  0.000424 ***
V23          -9.486e-02   6.456e-02   -1.469  0.141752
V24           1.295e-01   1.671e-01    0.775  0.438389
V25          -4.448e-02   1.501e-01   -0.296  0.766884
V26           2.369e-02   2.136e-01    0.111  0.911695
V27          -8.533e-01   1.408e-01   -6.060  1.36e-09 ***
V28          -3.242e-01   1.060e-01   -3.060  0.002211 **
Amount        8.938e-04   3.965e-04    2.254  0.024175 *
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5799.1  on 227845  degrees of freedom
Residual deviance: 1751.6  on 227815  degrees of freedom
AIC: 1813.6

Number of Fisher Scoring iterations: 12
```

5.6 Anexo F - Passo a passo da função 'stepAIC()' para o modelo glm base financeira

```
> modelo1 = glm(Class ~ ., data = train, family=binomial)
Warning message:
glm.fit: probabilidades ajustadas numericamente 0 ou 1 ocorreu
> summary(modelo1)

Call:
glm(formula = Class ~ ., family = binomial, data = train)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
```

```
−5.1778  −0.0287  −0.0191  −0.0121   4.6359

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) −8.391e+00  2.798e−01 −29.986  < 2e−16 ***
Time        −4.173e−06  2.527e−06  −1.651 0.098638 .
V1           1.249e−01  4.855e−02   2.572 0.010116 *
V2          −2.820e−03  6.422e−02  −0.044 0.964973
V3           1.748e−02  5.964e−02   0.293 0.769415
V4           7.422e−01  8.466e−02   8.767  < 2e−16 ***
V5           1.003e−01  7.458e−02   1.344 0.178880
V6          −1.057e−01  8.205e−02  −1.288 0.197841
V7          −1.403e−01  7.504e−02  −1.870 0.061436 .
V8          −1.834e−01  3.427e−02  −5.352 8.71e−08 ***
V9          −2.523e−01  1.252e−01  −2.015 0.043907 *
V10         −8.888e−01  1.109e−01  −8.016 1.09e−15 ***
V11          6.432e−03  9.101e−02   0.071 0.943659
V12          2.839e−02  9.468e−02   0.300 0.764337
V13         −3.336e−01  8.959e−02  −3.724 0.000196 ***
V14         −5.329e−01  6.776e−02  −7.864 3.71e−15 ***
V15         −9.570e−02  9.676e−02  −0.989 0.322621
V16         −1.556e−01  1.435e−01  −1.084 0.278361
V17          2.165e−03  7.857e−02   0.028 0.978016
V18         −2.609e−02  1.466e−01  −0.178 0.858767
V19          1.029e−01  1.099e−01   0.936 0.349133
V20         −4.664e−01  9.292e−02  −5.020 5.17e−07 ***
V21          3.579e−01  6.706e−02   5.338 9.41e−08 ***
V22          5.286e−01  1.500e−01   3.524 0.000424 ***
V23         −9.486e−02  6.456e−02  −1.469 0.141752
V24          1.295e−01  1.671e−01   0.775 0.438389
V25         −4.448e−02  1.501e−01  −0.296 0.766884
V26          2.369e−02  2.136e−01   0.111 0.911695
V27         −8.533e−01  1.408e−01  −6.060 1.36e−09 ***
V28         −3.242e−01  1.060e−01  −3.060 0.002211 **
Amount       8.938e−04  3.965e−04   2.254 0.024175 *
```

Signif. **codes**: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

(Dispersion parameter **for binomial family** taken to be 1)

Null **deviance**: 5799.1 **on** 227845 degrees of freedom
Residual **deviance**: 1751.6 **on** 227815 degrees of freedom
AIC: 1813.6

Number of Fisher Scoring iterations: 12

5.7 Anexo G - Sumário do brglm base financeira

> **summary**(brglm_**model**)

Call:
brglm(**formula** = Class ~ ., **family** = "binomial", **data** = train)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−8.278e+00	2.653e−01	−31.208	< 2e−16	***
Time	−4.082e−06	2.427e−06	−1.682	0.092560	.
V1	1.199e−01	4.592e−02	2.611	0.009039	**
V2	−2.774e−02	5.369e−02	−0.517	0.605340	
V3	1.165e−02	5.630e−02	0.207	0.836075	
V4	7.156e−01	7.648e−02	9.357	< 2e−16	***
V5	9.050e−02	6.587e−02	1.374	0.169467	
V6	−9.042e−02	7.577e−02	−1.193	0.232722	
V7	−1.593e−01	6.547e−02	−2.434	0.014941	*
V8	−1.884e−01	3.064e−02	−6.148	7.86e−10	***
V9	−2.829e−01	1.139e−01	−2.483	0.013011	*
V10	−8.632e−01	1.003e−01	−8.608	< 2e−16	***
V11	5.036e−03	8.701e−02	0.058	0.953843	

```
V12          5.823e-03  8.902e-02   0.065 0.947843
V13         -3.177e-01  8.606e-02  -3.692 0.000223 ***
V14         -5.260e-01  6.386e-02  -8.238  < 2e-16 ***
V15         -9.839e-02  9.231e-02  -1.066 0.286462
V16         -1.615e-01  1.301e-01  -1.242 0.214369
V17          2.017e-02  7.392e-02   0.273 0.784992
V18         -1.791e-02  1.336e-01  -0.134 0.893416
V19          9.378e-02  1.024e-01   0.916 0.359823
V20         -4.826e-01  8.248e-02  -5.851 4.88e-09 ***
V21          3.492e-01  6.090e-02   5.734 9.83e-09 ***
V22          5.189e-01  1.415e-01   3.668 0.000245 ***
V23         -1.007e-01  5.366e-02  -1.876 0.060663 .
V24          1.363e-01  1.591e-01   0.856 0.391847
V25         -4.998e-02  1.429e-01  -0.350 0.726527
V26          1.717e-02  2.031e-01   0.085 0.932643
V27         -8.506e-01  1.303e-01  -6.529 6.64e-11 ***
V28         -3.001e-01  8.103e-02  -3.703 0.000213 ***
Amount       8.762e-04  2.781e-04   3.151 0.001629 **
---
Signif. codes: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5768.6  on 227845  degrees of freedom
Residual deviance: 1754.5  on 227815  degrees of freedom
Penalized deviance: 1546.376
AIC:  1816.5
```

5.8 Anexo H - Passo a passo da função 'stepAIC()' para o modelo brglm base financeira

```
> brglm_model = brglm(formula = Class ~ . , family = "binomial", data = t
> summary(brglm_model)
```

Call:
brglm(**formula** = Class ~ ., **family** = "binomial", **data** = train)

Coefficients:

	Estimate	Std. Error	z value	Pr(>|z|)	
(Intercept)	−8.278e+00	2.653e−01	−31.208	< 2e−16	***
Time	−4.082e−06	2.427e−06	−1.682	0.092560	.
V1	1.199e−01	4.592e−02	2.611	0.009039	**
V2	−2.774e−02	5.369e−02	−0.517	0.605340	
V3	1.165e−02	5.630e−02	0.207	0.836075	
V4	7.156e−01	7.648e−02	9.357	< 2e−16	***
V5	9.050e−02	6.587e−02	1.374	0.169467	
V6	−9.042e−02	7.577e−02	−1.193	0.232722	
V7	−1.593e−01	6.547e−02	−2.434	0.014941	*
V8	−1.884e−01	3.064e−02	−6.148	7.86e−10	***
V9	−2.829e−01	1.139e−01	−2.483	0.013011	*
V10	−8.632e−01	1.003e−01	−8.608	< 2e−16	***
V11	5.036e−03	8.701e−02	0.058	0.953843	
V12	5.823e−03	8.902e−02	0.065	0.947843	
V13	−3.177e−01	8.606e−02	−3.692	0.000223	***
V14	−5.260e−01	6.386e−02	−8.238	< 2e−16	***
V15	−9.839e−02	9.231e−02	−1.066	0.286462	
V16	−1.615e−01	1.301e−01	−1.242	0.214369	
V17	2.017e−02	7.392e−02	0.273	0.784992	
V18	−1.791e−02	1.336e−01	−0.134	0.893416	
V19	9.378e−02	1.024e−01	0.916	0.359823	
V20	−4.826e−01	8.248e−02	−5.851	4.88e−09	***
V21	3.492e−01	6.090e−02	5.734	9.83e−09	***
V22	5.189e−01	1.415e−01	3.668	0.000245	***
V23	−1.007e−01	5.366e−02	−1.876	0.060663	.
V24	1.363e−01	1.591e−01	0.856	0.391847	
V25	−4.998e−02	1.429e−01	−0.350	0.726527	
V26	1.717e−02	2.031e−01	0.085	0.932643	
V27	−8.506e−01	1.303e−01	−6.529	6.64e−11	***

```
V28          -3.001e-01   8.103e-02   -3.703 0.000213 ***
Amount        8.762e-04   2.781e-04    3.151 0.001629 **
---
Signif. codes:  0  ***  0.001  **  0.01  *  0.05  .  0.1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5768.6  on 227845  degrees of freedom
Residual deviance: 1754.5  on 227815  degrees of freedom
Penalized deviance: 1546.376
AIC:  1816.5

> # Primeiro modelo:
> brglm_model1=brglm(Class ~ Time + V1 + V4 + V5 + V7 + V8 + V9 + V10 + V
+                        V20 + V21 + V22 + V23 + V27 + V28 +
+                        Amount, data = train, family = binomial)
> summary(brglm_model1)

Call:
brglm(formula = Class ~ Time + V1 + V4 + V5 + V7 + V8 + V9 +
    V10 + V13 + V14 + V20 + V21 + V22 + V23 + V27 + V28 + Amount,
    family = binomial, data = train)


Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept) -8.391e+00  2.189e-01 -38.335  < 2e-16 ***
Time        -4.600e-06  1.898e-06  -2.424 0.015339 *
V1           1.151e-01  3.677e-02   3.131 0.001743 **
V4           7.301e-01  6.563e-02  11.125  < 2e-16 ***
V5           1.300e-01  2.940e-02   4.423 9.76e-06 ***
V7          -2.093e-01  5.104e-02  -4.100 4.12e-05 ***
V8          -1.484e-01  2.118e-02  -7.006 2.45e-12 ***
V9          -1.706e-01  8.766e-02  -1.946 0.051618 .
V10         -9.059e-01  8.671e-02 -10.448  < 2e-16 ***
```

V13	−3.145e−01	8.470e−02	−3.713	0.000205 ***
V14	−5.373e−01	4.989e−02	−10.770	< 2e−16 ***
V20	−4.766e−01	7.161e−02	−6.655	2.83e−11 ***
V21	4.286e−01	5.294e−02	8.096	5.67e−16 ***
V22	7.261e−01	1.285e−01	5.649	1.61e−08 ***
V23	−1.281e−01	4.695e−02	−2.728	0.006366 **
V27	−8.606e−01	1.135e−01	−7.581	3.43e−14 ***
V28	−3.093e−01	7.886e−02	−3.922	8.79e−05 ***
Amount	1.000e−03	2.689e−04	3.719	0.000200 ***

Signif. **codes**: 0 *** 0.001 ** 0.01 * 0.05 . 0.1

(Dispersion parameter **for binomial family** taken to be 1)

Null **deviance**: 5793.4 **on** 227845 degrees of freedom
Residual **deviance**: 1769.8 **on** 227828 degrees of freedom
Penalized **deviance**: 1624.941
AIC: 1805.8