Universidade de Brasília
Instituto de Ciências Exatas
Departamento de Estatística

# Um Escore de Risco para Classificação de Transações Suspeitas de Lavagem de Dinheiro via Regressão Ordinal

**Maria Clara Vieira Borba**

Orientador: Prof.º Dr. Eduardo Yoshio Nakano

Brasília, 2017

# Um Escore de Risco para Classificação de Transações Suspeitas de Lavagem de Dinheiro via Regressão Ordinal

Versão definitiva da dissertação apresentada ao Departamento de Estatística do Instituto de Ciências Exatas da Universidade de Brasília durante a defesa realizada em 05/12/2017, como requisito parcial para a obtenção do título de Mestre em Estatística.

Banca Examinadora:

---

Prof. Eduardo Yoshio Nakano - EST/UnB (Orientador)

---

Prof. Juliana Betini Fachini - EST/UnB

---

Prof. Marcelo Angelo Cirillo - DEX/UFLA

*Eu sei que não sou nada e que talvez nunca tenha tudo. Aparte isso, eu tenho em mim todos os sonhos do mundo.*
*<Fernando Pessoa>*

# Resumo

O objetivo deste trabalho foi apresentar um escore de risco para classificação de transações financeiras suspeitas de lavagem de dinheiro. Esse escore é obtido a partir dos resultados de um modelo de regressão logística ordinal cumulativo. O escore proposto foi ilustrado em dados reais visando classificar os associados quanto o seu risco de ter realizado movimentações com indícios do crime de lavagem de dinheiro, sendo que o mesmo se mostrou uma ferramenta útil para ranquear e classificar esses associados.

**Palavras-chave:** regressão logística, lavagem de dinheiro, modelos cumulativos, risco.

# Abstract

The purpose of this paper is to present a risk score to classify suspected money laundering transactions. This score is obtained from the results of a cumulative ordinal logistic regression model. The proposed score was illustrated using real data in order to classify the associates as to their risk of having carried out movements with indications of the crime of money laundering, and this proved to be an useful tool to rank and classify these associates.

**Keywords:** logistic regression, money laundering, cumulative models, risk.

# Índice

# Lista de Figuras

# Lista de Tabelas

# Capítulo 1

# Introdução

Em diversos estudos existe o interesse em avaliar a influência de fatores sobre uma resposta dicotômica ou politômica. Essa análise pode ser realizada por meio de modelos de Regressão Logística. Essa metodologia é uma das mais importantes para dados com variável resposta categórica e é uma técnica estatística amplamente utilizada em diversos campos do conhecimento, tais como: medicina, em estudos epidemiológicos para avaliar fatores de riscos na incidência de doenças; ciências sociais, para explicar as intenções de voto em atos eleitorais; ou econometria, na predição de grupos de risco para a obtenção de crédito. Essa técnica tem como grande diferencial a praticidade de interpretação dos resultados e apresenta ainda qualidade ao ser capaz, não somente de mostrar qual combinação de variáveis explicativas é melhor, como também estimar sua significância na variável resposta.

Ela se tornou uma ferramenta popular para aplicação em negócios. Por exemplo, nos dias atuais, uma das atividades muito em voga na mídia em geral que assola a comunidade internacional é o crime de lavagem de dinheiro. Nesse contexto, a Regressão Logística pode ser utilizada com a intenção de descrever as características (variáveis independentes ou covariáveis) de movimentações com indícios do crime, bem como ser utilizada como uma técnica de mineração dos dados de uma empresa, em que não é viável a análise do perfil de movimentação de cada cliente, individualmente.

De acordo com Agresti [4], sob o ponto de vista matemático, a regressão logística dicotômica é razoavelmente flexível e fácil de ser utilizada, além de permitir uma interpretação dos resultados bastante rica e direta. Ainda, escores de risco podem ser obtidos a partir dos resultados de uma análise de regressão logística.

O número de trabalhos na literatura com o uso da regressão logística dicotômica foi impulsionado após trabalhos de Agresti [3], Cox [7] e Hosmer e Lemeshow [9], entre

outros. Entretanto, essa abundância se contrasta com a escassez de trabalhos que consideram variáveis respostas politômicas. Essa escassez é maior ainda quando a variável resposta politômica apresenta natureza ordinal e pode ser atribuída à maior complexidade do modelo e reduzir opções de modelagem oferecidas em *softwares* estatísticos comerciais [1].

Neste contexto, este trabalho teve como objetivo a abordagem da técnica de Regressão Logística Ordinal, utilizando Modelos Cumulativos, bem como a apresentação de um escore de risco, calculado a partir dos resultados obtidos pelo modelo de regressão logística ordinal. A metodologia desse trabalho foi aplicada em dados reais para construir um escore de risco de lavagem de dinheiro. Todas as análises deste trabalho foram realizadas pelo *software* livre R [13].

# Capítulo 2

# Revisão Bibliográfica

## 2.1 Lavagem de Dinheiro

Lavar dinheiro significa transformar recursos de origem ilícita, adquirido por meio de atividades criminosas, em algo supostamente lícito, dissimulando sua fonte por meio de complexas manobras financeiras para realizar atividades legais e ilegais, para que se utilizem esses recursos sem atrair a atenção à atividade principal ou às pessoas envolvidas na geração dos lucros.

O crime de lavagem de dinheiro pode resultar em consequências econômicas, sociais e de segurança devastadoras, afetando principalmente países em desenvolvimento, mercados emergentes frágeis e países com sistemas financeiros frágeis. Alguns efeitos são: aumento do crime e da corrupção, por países considerados paraísos fiscais serem mais vulneráveis; e enfraquecimento das instituições financeiras, considerando o risco de reputação, riscos operacionais, riscos legais e riscos de concentração [2].

Em muitos casos, a lavagem de dinheiro envolve uma série complexa de operações que costumam ser difíceis de separar. Contudo, em geral, é possível apontar três fases no processo de lavagem de dinheiro:

1. Colocação: introdução dos recursos obtidos de forma ilegal no sistema financeiro;
2. Ocultação: conversão do produto do crime em qualquer outra forma e criação de camadas complexas de operações financeiras para dificultar o rastreamento para fins de auditoria, bem como ocultar a fonte e titularidade dos recursos;
3. Integração: reinserção dos recursos ilícitos na economia através de operações comerciais aparentemente legítimas ou pessoais regulares.

Em cooperativas de crédito, por serem de pequeno porte, o crime de lavagem de dinheiro em comparação aos bancos e demais instituições financeiras, é mais difícil de ser cometido, por ser mais fácil identificar qualquer tipo de atividade suspeita quando há um volume menor de operações. Apesar disso, são vulneráveis, já que quanto mais serviços financeiros forem oferecidos por uma cooperativa de crédito, maior é o potencial de risco de lavagem de dinheiro. Em cooperativas, o número de clientes que oferecem meios para que possíveis criminosos possam ocultar seus recursos ilegais tende a aumentar. Além disso, há a realização de altos níveis de operações em dinheiro, o que aumenta o risco de ocorrência do crime [2].

## 2.2 Regressão Logística Clássica

Considerando a Regressão Logística Clássica, a variável resposta assume somente dois valores qualitativos, comumente denotados por "sucesso"e "fracasso", sendo representada por uma variável indicadora binária de valores 0 e 1 [6].

Sendo assim, tem-se $Y$, uma variável binária, como descrito acima. Sua distribuição é a de Bernoulli, sendo especificada pelas probabilidades $P(Y = 1) = \pi$ de sucesso e $P(Y = 0) = (1 - \pi)$ de fracasso. Sua média é $E(Y) = \pi$. Suponha $\mathbf{X}$ um vetor de variáveis explicativas $(X_1, X_2, ..., X_k)$, correspondendo às variáveis independentes do estudo, em que as probabilidades do modelo mudam linearmente em $\mathbf{x}$. Agora, $\pi = P(Y = 1|\mathbf{x})$ denota a probabilidade de sucesso para os valores específicos das variáveis explicativas.

Considerando que em modelos de regressão linear, $\mu = E(Y)$ é uma função linear de $\mathbf{X}$, em que $0 < E(Y|x) < 1$ quando se tem uma resposta binária, um modelo análogo seria

$$E(Y_i|\mathbf{x}) = \pi_i = \theta_0 + \gamma_1 x_{i1} + \ldots + \gamma_k x_{ik} = \mathbf{x}_i'\gamma$$

em que $\gamma' = [\theta_0, \gamma_1, \ldots, \gamma_k]$ é o vetor de coeficientes de regressão a serem estimados e $\mathbf{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ik})$ é o vetor de variáveis explicativas do $i$-ésimo indivíduo da amostra, $i = 1, \ldots, n$.

Contudo, ao considerar um modelo de regressão linear cuja variável resposta é binária, surgem os seguintes problemas [6]:

1. **Não normalidade dos erros**

   Cada erro $\epsilon_i$ pode assumir somente dois valores, o que indica que os erros associados ao modelo de regressão logística não seguem uma distribuição Normal.

Considerando $\epsilon_i = Y_i - (\theta_0 + \gamma_1 x_{i1} + \ldots + \gamma_k x_{ik})$, tem-se que:

$$\epsilon_i = 1 - (\theta_0 + \gamma_1 x_{i1} + \ldots + \gamma_k x_{ik}), \quad \text{se } Y_i = 1 \text{ e}$$
$$\epsilon_i = -\theta_0 - \gamma_1 x_{i1} - \ldots - \gamma_k x_{ik}, \quad \text{se } Y_i = 0$$

Assim, o modelo de regressão linear já não se torna apropriado, por supor normalidade nos erros.

2. **Heterocedasticidade**

   Ainda levando em consideração os erros $\epsilon_i$, quando a resposta é uma variável indicadora, eles não possuem variâncias iguais, já que:

$$Var(Y_i) = \pi_i(1 - \pi_i) = E(Y_i)[1 - E(Y_i)]$$

   Como $\epsilon_i = (Y_i - \pi_i)$ e $\pi_i$ é uma constante, tem-se que $Var(Y_i) = Var(\epsilon_i)$. Assim,

$$\begin{aligned} Var(\epsilon_i) &= E(Y_i)[1 - E(Y_i)] \\ &= (\theta_0 + \gamma_1 x_{i1} + \ldots + \gamma_k x_{ik})(1 - \theta_0 - \gamma_1 x_{i1} - \ldots - \gamma_k x_{ik}), \end{aligned}$$

   que depende de cada $x_i$, fazendo com que a variância dos erros difiram para cada nível de $\mathbf{x}$.

3. **Restrição no Modelo**

   Como o que está sendo modelado são probabilidades, existe a seguinte restrição para a resposta média do modelo:

$$0 \leq E(Y_i) = \pi_i \leq 1. \tag{2.1}$$

   Funções de resposta lineares, como a do modelo de regressão linear podem não atender satisfatoriamente a essa restrição, o que seria claramente uma impropriedade matemática.

Diante das dificuldades apontadas, principalmente em (2.1), caso se queira modelar as probabilidades, não seria uma escolha adequada utilizar o modelo de regressão linear. Frequentemente, então, é necessário transformar as variáveis explicativas, tornando a resposta esperada, a ser estimada pelo modelo, em uma função não linear.

A transformação mais comumente aplicada utiliza a função exponencial para garantir valores compreendidos no intervalo (0,1), resultando na função de resposta

logística [5]:

$$E(Y_i|\mathbf{x}) = \pi_i = \frac{\exp(\theta_0 + \gamma_1 x_{i1} + \ldots + \gamma_k x_{ik})}{1 + \exp(\theta_0 + \gamma_1 x_{i1} + \ldots + \gamma_k x_{ik})}$$
$$= \frac{\exp(\mathbf{x}_i'\gamma)}{1 + \exp(\mathbf{x}_i'\gamma)}, \quad i = 1, \ldots, n. \tag{2.2}$$

Visto que $\mathbf{x}_i'\gamma = \theta_0 + \gamma_1 x_{i1} + \ldots + \gamma_k x_{ik}$ pode assumir qualquer valor real enquanto $\pi_i$ está restrito ao intervalo (0,1), o objetivo da transformação foi atingido.

A transformação de $\pi_i$ utilizada para obtenção da forma aditiva é chamada de transformação logito e é definida da seguinte forma [4]:

$$logit(\pi_i) = log\left(\frac{\pi_i}{1 - \pi_i}\right) = \mathbf{x}_i'\gamma, \quad i = 1, \ldots, n. \tag{2.3}$$

Essa transformação em (2.3) é contínua, linear nos seus parâmetros e pode assumir qualquer valor real. Dessa forma, várias propriedades assumidas por um modelo de regressão linear são satisfeitas.

O fato da interpretação dos coeficientes do modelo ser simples e útil, foi um dos fatores que tornam a regressão logística mais interessante e que contribuiu para sua popularização [10].

A interpretação é baseada na razão de chances. O termo $\left(\frac{\pi}{1-\pi}\right)$ é conhecido como chance de sucesso. Logo, mantidas constantes todas as demais variáveis explicativas do modelo, ao se aumentar uma unidade em $x_k$, a chance de sucesso estimado anteriormente a este incremento é $e^{\gamma_k}$ [5].

A estimação dos parâmetros do modelo é feita pelo método de Máxima Verossimilhança. Relembrando que cada uma das $n$ respostas da amostra é uma variável de Bernoulli independente, as distribuições de probabilidade podem ser representadas como:

$$P_{Y_i}(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, \quad y_i = 0, 1 \quad \text{e} \quad i = 1, \ldots, n.$$

Considerando que as $n$ observações $Y_i$ são independentes, a função de probabilidade conjunta é dada por:

$$P(y_1, \ldots, y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \pi_i^{y_i}(1 - \pi_i)^{1-y_i}$$

Substituindo $\pi_i$ por (2.2), obtemos a expressão da função de verossimilhança:

$$L(\gamma|Y,X) = \prod_{i=1}^{n} \left( \frac{\exp(\mathbf{x_i'}\gamma)}{1+\exp(\mathbf{x_i'}\gamma)} \right)^{y_i} \left( \frac{1}{1+\exp(\mathbf{x_i'}\gamma)} \right)^{1-y_i}, \tag{2.4}$$

em que $\gamma' = [\theta_0, \gamma_1, \ldots, \gamma_k]$ é o vetor de parâmetros, $Y' = [y_1, \ldots, y_n]$ é o vetor das respostas e $\mathbf{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ é o vetor de covariáveis do $i$-ésimo indivíduo da amostra.

As estimativas de máxima verossimilhança dos parâmetros são os valores que maximizam a função (2.4). Não há formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos para tal fim, como por exemplo o de Newton-Raphson [10].

## 2.3 Regressão Logística Multi-Categórica

A Regressão Logística é mais frequentemente utilizada para modelar a relação entre uma variável resposta dicotômica e um conjunto de variáveis independentes. Em muitos casos, porém, a variável resposta é categórica, possuindo mais de dois níveis.

Uma variável resposta é definida como sendo do tipo categórica quando ela possui uma escala de medida formada por um conjunto de categorias. As categorias que a variável dependente assumem podem ser de natureza nominal ou ordinal. Uma variável ordinal ou intervalar é quantitativa, porque cada nível sobre sua escala pode ser comparado a um outro nível em termos de magnitude maior ou menor de certa característica. Esse tipo de variável ordinal é de uma natureza bem diferente das variáveis qualitativas, que são medidas em uma escala nominal. Os níveis dessas variáveis diferem em qualidade, não em quantidade. Então, a ordem de apresentações das categorias de uma variável nominal não é relevante [6].

Nesse contexto, enquadra-se uma abordagem que utiliza uma generalização da regressão logística vista na Seção anterior, chamada de Regressão Logística Multi-categórica [10]. Assim, discute-se essa regressão que estuda a relação entre um conjunto de covariáveis $X_1, X_2, ..., X_k$ e uma variável resposta $Y$, esta podendo ter ou não uma ordenação natural. Quando há uma ordem natural entre as categorias, fala-se de uma Regressão Logística Ordinal, caso contrário, quando a variável resposta é puramente qualitativa, de uma Regressão Logística Nominal [12].

O objetivo deste capítulo é descrever parte dos modelos de regressão para variáveis resposta categóricas nominais e ordinais.

### 2.3.1 Regressão Logística Nominal Clássica

Seja $R$ o número de categorias de uma variável resposta $Y$, sendo $[\pi_1, \pi_2, \ldots, \pi_R]$ as respectivas probabilidades, que satisfazem a condição $\sum_{r=1}^{R} \pi_r = 1$. Considerando a existência de $n$ observações independentes, a probabilidade de todas as formas que essas $n$ observações podem se associar às $R$ categorias pode ser especificada por uma distribuição de probabilidades multinomial [4].

A $i$-ésima observação pode ser escrita a partir de $R$ variáveis respostas binárias, $Y_{i1}, \ldots, Y_{iR}$, em que:

$$Y_{ir} = \begin{cases} 1, & \text{se a } i\text{-ésima resposta está na categoria } r \\ 0, & \text{caso contrário} \end{cases},$$

em que $i = 1, 2, \ldots, n$ e $r = 1, 2, \ldots, R$.

Pelo fato da $i$-ésima variável resposta estar associada a somente uma categoria, deve-se considerar que $\sum_{r=1}^{R} Y_{ir} = 1$, sendo $\pi_{ir}$ a probabilidade da categoria $r$ ser selecionada para a $i$-ésima resposta, isto é, $\pi_{ir} = P(Y_{ir} = 1)$.

Para os casos da Seção 2.2, em que se considera a Regressão Logística Binária, tem-se que $R = 2$. Ao denotar $Y_i = 1$ se a $i$-ésima resposta for da Categoria 1, por exemplo, e $Y_i = 0$ se a $i$-ésima resposta for da Categoria 2. Então:

$$\pi_{i1} = \pi_i \qquad \text{e} \qquad \pi_{i2} = 1 - \pi_i, \quad i = 1, 2, \ldots, n.$$

Nesse caso, o *Logit* de $\pi_i$ é modelado utilizando um preditor linear e, como há somente 2 categorias na Regressão Logística Binária, o *Logit* de fato compara a probabilidade da resposta ser da Categoria 1 com a probabilidade da resposta ser da Categoria 2:

$$logit(\pi_i) = \log\left[\frac{\pi_i}{1 - \pi_i}\right] = \log\left[\frac{\pi_{i1}}{\pi_{i2}}\right] = logit(\pi_{i12}) = \mathbf{x}_i' \gamma_{12}.$$

Note que o subscrito "12" foi adotado para enfatizar que o preditor linear está modelando o logaritmo da razão das probabilidades para as Categorias 1 e 2 [10].

Generalizando para $R$ categorias, tem-se $R(R-1)/2$ pares de categorias para comparação e, consequentemente, $R(R-1)/2$ preditores lineares. Portanto, os modelos logísticos multi-categóricos nominais utilizam simultaneamente todos os pares de categorias, especificando a chance da resposta estar em uma categoria em comparação com outra (lembrando que não há ordenação nas categorias). No entanto, não é necessário desenvolver todos os $R(R-1)/2$ modelos logísticos, já que, na prática, uma

categoria é escolhida como base e, então, todas as demais categorias são comparadas a ela. A escolha da categoria base, também chamada de categoria (nível) de referência, é arbitrária [6].

Por exemplo, ao utilizar a categoria $R$ como base para análise, considera-se as $R-1$ comparações a essa categoria. O *Logit* para a $r$-ésima comparação é [10]:

$$logit(\pi_{irR}) = log\left[\frac{\pi_{ir}}{\pi_{iR}}\right] = \mathbf{x}_i'\gamma_{rR}, \quad r = 1, 2, \ldots, R-1 \quad \text{e} \quad i = 1, \ldots, n.$$

Como todas as comparações são feitas com a categoria $R$, pode-se reescrever a equação acima por:

$$logit(\pi_{ir}) = log\left[\frac{\pi_{ir}}{\pi_{iR}}\right] = \mathbf{x}_i'\gamma_{r}, \quad r = 1, 2, \ldots, R-1 \quad \text{e} \quad i = 1, \ldots, n. \tag{2.5}$$

Torna-se suficiente levar em consideração apenas $R-1$ *Logits*, pelo fato de qualquer outro *Logit* poder ser obtido através deles [10]. Em geral, para comparar as categorias $l$ e $m$:

$$log\left[\frac{\pi_{il}}{\pi_{im}}\right] = \mathbf{x}_i'(\gamma_l - \gamma_m), \quad i = 1, 2, \ldots n.$$

Os coeficientes do Modelo Nominal podem ser interpretados da mesma maneira dos coeficientes da Regressão Logística Clássica, por meio da razão de chances. Deve-se ficar atento apenas em relação aos parâmetros que estão sendo analisados e suas categorias [5].

De acordo com as expressões em (2.5), obtém-se as $R-1$ expressões diretas para as probabilidades de cada categoria em termo dos $R-1$ preditores lineares $\mathbf{x}'\gamma_r$. As expressões resultantes são [10]:

$$\pi_{ir} = \frac{\exp(\mathbf{x}_i'\gamma_r)}{1 + \sum_{l=1}^{R-1}\exp(\mathbf{x}_i'\gamma_l)}, \quad r = 1, 2, \ldots, R-1 \quad \text{e} \quad i = 1, \ldots, n. \tag{2.6}$$

Para que possa ser feita a estimação simultânea dos $R-1$ vetores de parâmetros $\gamma_1, \gamma_2, \ldots, \gamma_{R-1}$ é preciso obter a função de verossimilhança dos dados e aplicar o método de máxima verossimilhança. Assim, para $n$ observações independentes e $R$ categorias, a função de probabilidade conjunta é dada por [10]:

$$P(y_1, \ldots, y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n}\left[\prod_{r=1}^{R}\pi_{ir}^{y_{ir}}\right]. \tag{2.7}$$

Visto que $\pi_{iR} = 1 - \sum_{r=1}^{R-1} \pi_{ir}$ e $y_{iR} = 1 - \sum_{r=1}^{R-1} y_{ir}$, tem-se que:

$$P(y_1, \ldots, y_n) = \prod_{i=1}^{n} P_{Y_i}(y_i) = \prod_{i=1}^{n} \left[ \left( \prod_{r=1}^{R-1} \pi_{ir}^{y_{ir}} \right) \left( 1 - \sum_{r=1}^{R-1} \pi_{ir} \right)^{1 - \sum_{r=1}^{R-1} y_{ir}} \right]$$

Usando como referência a igualdade (2.6) e considerando a categoria $R$ como base, obtém-se a expressão da função de verossimilhança desejada [10]:

$$L(\gamma_1, \gamma_2, \ldots, \gamma_{R-1} | Y, X) =$$

$$= \prod_{i=1}^{n} \left\{ \left[ \prod_{r=1}^{R-1} \left( \frac{\exp(\mathbf{x}_i' \gamma_r)}{1 + \sum_{l=1}^{R-1} \exp(\mathbf{x}_i' \gamma_l)} \right)^{y_{ir}} \right] \left[ \left( \frac{1}{1 + \sum_{l=1}^{R-1} \exp(\mathbf{x}_i' \gamma_l)} \right)^{1 - \sum_{r=1}^{R-1} y_{ir}} \right] \right\}, \quad (2.8)$$

em que $\gamma_l' = [\theta_l, \gamma_{l1}, \gamma_{l2} \ldots, \gamma_{lk}]$, $\quad l = 1, 2, \ldots, R-1$ são os vetores dos parâmetros; $\mathbf{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ é o vetor de covariáveis do $i$-ésimo indivíduo da amostra; e $y_i = (y_{i1}, y_{i2}, \ldots, y_{ik})$ é o vetor de respostas do $i$-ésimo indivíduo da amostra. Aqui, $y_{ir}$ é igual a 1 se a resposta do $i$-ésimo indivíduo pertence à categoria $r$, $r = 1, 2, \ldots, R-1$.

As estimativas de máxima verossimilhança de $\gamma$ são os valores que maximizam (2.8). Não existem formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos como por exemplo o de Newton-Raphson [10].

### 2.3.2 Regressão Logística Ordinal Clássica

Variáveis categóricas ordinais são importantes em muitas áreas de estudo, principalmente em situações em que medidas exatas não são possíveis. Caracterizam-se por apresentar uma ordenação entre seus possíveis valores e estão muito presentes em Ciências Sociais, em particular, para medir atitudes e opiniões sobre vários assuntos, assim como *status* de diversos tipos. Também costumam ocorrer em campos como *marketing* e em disciplinas médicas e de saúde pública.

As variáveis ordinais podem ser originadas de formas diferentes: variáveis contínuas agrupadas e variáveis categóricas naturalmente ordenadas. A primeira é formada pela categorização de uma variável contínua. Como exemplo, temos a variável número de anos de estudo, que pode ser medida de forma ordinal por meio da categorização 0-8, 9-12, 13-16, 17 ou mais anos.

A segunda forma consiste na avaliação de uma informação não quantificável, casos em que uma medida precisa nem sempre é possível, associada a níveis de uma escala ordinal, realizando a coleção de categorias naturalmente ordenadas. Como ilustração, temos o risco de ocorrência de um crime, que pode ser classificado como "alto", "médio"ou "baixo".

Uma variável categórica é referida como ordinal ao invés de intervalar quando há uma ordem clara das categorias, mas as distâncias absolutas entre elas são desconhecidas. Assim, para muitas variáveis categóricas ordinais, é sensato imaginar a existência de uma variável contínua subjacente. Para se aproximar da escala subjacente, é frequentemente útil associar um conjunto "razoável"de scores às categorias.

#### 2.3.2.1 Modelos Cumulativos

Ao se realizar um estudo em que os dados são categóricos, o tamanho da amostra é frequentemente crítico. Assim, é de suma importância fazer uso de toda a informação disponível, sendo necessário levar em consideração a ordem das categorias, o que é feito pelos modelos de regressão ordinal. Apesar das técnicas de regressão logística nominal também possuírem a capacidade de analisar respostas ordinais, levar em conta a ordem das categorias resulta em um modelo de mais fácil interpretação [10].

Assume-se que a variável observada $Y$ é uma categorização da variável contínua latente $U$. No caso de variáveis respostas contínuas agrupadas, a variável latente pode ser considerada a variável subjacente não observada. Já no caso de variáveis categóricas naturalmente ordenadas, assume-se que a variável observada $Y$ é uma categorização da variável contínua latente $U$, uma avaliação sobre uma escala contínua subjacente, utilizada apenas para facilitar a interpretação e a construção do modelo.

Para um dado vetor $\mathbf{x}$ de variáveis explicativas, a abordagem do limite das categorias sugere que a variável observada $Y$, com $Y \in \{1, \ldots, R\}$, e a variável latente $U$ estão conectadas por:

$$Y = r \Leftrightarrow \theta_{r-1} < U < \theta_r, \quad r = 1, \ldots, R \tag{2.9}$$

em que $(-\infty = \theta_0 < \theta_1 < \ldots < \theta_R = \infty)$. Isso significa que $Y$ é uma versão categorizada de $U$, determinada pelos pontos $\theta_1, \ldots, \theta_{R-1}$. Além disso, assume-se que a variável latente $U$ é determinada pelas variáveis explicativas de forma linear:

$$U = -\mathbf{x}'\gamma + \epsilon \tag{2.10}$$

em que $\gamma = (\gamma_1, \ldots, \gamma_k)$ é o vetor de parâmetros e $\epsilon$ é uma variável aleatória com

distribuição acumulada $F$.

Diante do exposto e levando em consideração que a modelagem da probabilidade de ocorrência de uma das classes de uma variável dependente é feita em termos de probabilidades acumuladas, por existir uma relação de ordem entre elas [4], a distribuição de probabilidades da variável observada $Y$ é dada por:

$$\begin{aligned} P(Y \leq r|\mathbf{x}) &= F(\theta_r + \mathbf{x}'\gamma) \\ &= P(Y = 1|\mathbf{x}) + P(Y = 2|\mathbf{x}) + \ldots + P(Y = r|\mathbf{x}), \quad r = 1, \ldots, R \end{aligned} \tag{2.11}$$

As probabilidades acumuladas refletem a ordenação natural $0 \leq P(Y \leq 1|\mathbf{x}) \leq P(Y \leq 2|\mathbf{x}) \leq \ldots \leq P(Y \leq R-1|\mathbf{x}) \leq 1$ e, por isso, o modelo é denominado Modelo Cumulativo com função de distribuição $F$. Modelos para probabilidades cumulativas não utilizam a última categoria, $P(Y \leq R|\mathbf{x})$ visto que ela é necessariamente igual a 1 (informação referente à última classe redundante) [12].

Para verificar (2.11), utiliza-se a equação 2.9 e obtém-se:

$$\begin{aligned} P(Y \leq r|\mathbf{x}) =& P(Y = 1|\mathbf{x}) + P(Y = 2|\mathbf{x}) + \ldots + P(Y = r|\mathbf{x}) \\ =& P(\theta_0 < U < \theta_1) + P(\theta_1 < U < \theta_2) + \ldots + P(\theta_{r-1} < U < \theta_r) \\ =& F_U(\theta_r) - F_U(\theta_0). \end{aligned}$$

Como $\theta_0 = -\inf$, temos $F_U(\theta_0) = 0$ e, sendo $U$ determinada por (2.10), $F_U(\theta_r) = P(\mathbf{x}'\gamma + \epsilon \leq \theta_r) = P(\epsilon \leq \theta_r + \mathbf{x}'\gamma) = F(\theta_r + \mathbf{x}'\gamma)$, em que $F$ é a função de distribuição acumulada da variável aleatória $\epsilon$.

O inverso da função $F$ ($F^{-1}$) é designada função de ligação (*Link*) por fazer a associação linear entre a parte aleatória do modelo, $P(Y \leq r)$ e a parte sistemática ($\mathbf{x}'\gamma$). Ou seja:

$$Link(P(Y \leq r)) = \theta_r + \mathbf{x}'\gamma$$

Várias são as opções para se usar como função de ligação, cuja utilização no modelo ordinal é recomendável de acordo com o tipo de distribuição de probabilidades que as classes da variável dependente apresentam. Esta escolha deve ser feita com cuidado, pois uma escolha inapropriada pode comprometer a significância do modelo e sua capacidade preditiva [6]. As cinco principais funções de ligação estão descritas na Tabela 2.1:

Tabela 2.1: Principais Funções de Ligação ($F^{-1}$)

| Nome | Função de Ligação |
|---|---|
| *Logit* | $\log\left[\frac{P(Y\leq r)}{P(Y>r)}\right]$ |
| *Complemento Log-log* | $\log\{-\log[1-P(Y\leq r)]\}$ |
| *Log-log Negativo* | $-\log\{-\log[P(Y\leq r)]\}$ |
| *Cauchit* | $\tan(\pi(P(Y\leq r)-0,5))$ |
| *Probit* | $\phi^{-1}(P(Y\leq r))$, onde $\phi$ é a f.d.p. da N(0,1) |

Na prática, a função de ligação *Logit* é a mais utilizada, devido a sua interpretação interessante dos coeficientes do modelo e da sua matemática simples. Essa será a abordagem explorada neste texto.

Existem algumas metodologias de análise para quando a variável resposta em estudo possui mais de dois níveis e natureza ordenada, dentre elas os modelos: Odds Proporcionais, Odds Proporcionais Parciais, Razão Contínua e Estereótipos. Esses pertecem à família logística, que contém uma hierarquia de modelos tanto para dados ordenados quanto nominais. Existe um maior destaque para o modelo de chances proporcionais que assume como pressuposição básica que os coeficientes, para cada variável explicativa do modelo, são os mesmos para todos os logitos. Este será o modelo considerado neste trabalho.

#### 2.3.2.2 Modelo Logístico Cumulativo de Chances Proporcionais

O modelo é proposto através de uma analogia com a regressão logística usual, de forma que o *Logit* das probabilidades cumulativas são [10]:

$$\begin{aligned} Logit[P(Y_i \leq r|\mathbf{x})] =& \log\left[\frac{P(Y_i \leq r|\mathbf{x})}{1-P(Y_i \leq r|\mathbf{x})}\right] \\ =& \theta_r - \gamma_1 x_{i1} - \ldots - \gamma_k x_{ik}, \quad r = 1, \ldots, R-1 \end{aligned} \tag{2.12}$$

Consequentemente,

$$P(Y_i \leq r|\mathbf{x}) = \frac{\exp(\theta_r + \mathbf{x}'\gamma)}{1+\exp(\theta_r + \mathbf{x}'\gamma)}, \quad r = 1, \ldots, R-1 \tag{2.13}$$

O modelo ordinal atrás definido permite estimar o logaritmo da probabilidade da variável dependente tomar os valores de classes inferiores ou iguais a $r$, comparativamente com a probabilidade de tomar os valores das classes superiores a $r$.

Note que os coeficientes de regressão $\gamma' = [\gamma_1, \ldots, \gamma_k]$ em (2.12) não apresentam índice $r$, obrigando o modelo a pressupor que os efeitos das variáveis independentes sobre o $P(Y_i \leq r)$ é igual para todas as classes [10]. Assim, a resposta observada em cada classe apenas se encontra deslocada para a direita ou para a esquerda, em função de $\theta_r$. Isso resulta em um modelo mais parcimonioso. Para uma $\gamma_p > 0$, um aumento em algum $X_p$ resulta na diminuição da probabilidade da variável dependente tomar valores de ordem inferiores ou iguais a $r$ (mantendo as demais variáveis explicativas constantes), ou seja, quando $X_p$ aumenta, $Y$ diminui.

A interpretação do modelo pode usar as razões de chance para as probabilidades cumulativas e seus complementos [12]. Para dois valores $x_1$ e $x_2$ de uma das variáveis explicativas $X_k$ do estudo, a razão de chances comparando as probabilidades cumulativas, para todas as classes da variável dependente, é dada por (mantendo as demais variáveis explicativas constantes):

$$\text{razão de chances} = \frac{P(Y \leq r|X_k = x_2)/P(Y > r|X_k = x_2)}{P(Y \leq r|X_k = x_1)/P(Y > r|X_k = x_1)} \tag{2.14}$$

O log dessa razão de chances é a diferença entre os logitos cumulativos para esses dois valores de $X_k$. Isso é igual a $\gamma_k(x_2 - x_1)$. Se $x_2 - x_1 = 1$, a chance da variável resposta assumir valores menores para qualquer categoria é multiplicado por $e^{-\gamma_k}$ para cada unidade acrescida em $X_k$.

Os parâmetros do modelo $\theta_1, \ldots, \theta_{R-1}$ e $\boldsymbol{\gamma}$ são estimados simultaneamente pelo método de Máxima Verossimilhança. Para isso, é necessário a obtenção da função de verossimilhança para os dados, lembrando que o modelo pressupõe que as curvas de probabilidade das $R-1$ classes da variável dependente são iguais para todas as classes e são calculadas de forma cumulativa.

A partir de (2.7), para $n$ observações independentes e $R$ categorias, a função de verossimilhança é dada por [5]:

$$L(\theta, \gamma|X, Y) = \prod_{i=1}^{n} \left[ \prod_{r=1}^{R} [P(Y_i \leq r|\mathbf{x}) - P(Y_i \leq r-1|\mathbf{x})]^{y_{ir}} \right]$$

Substituindo $P(Y_i \leq R|\mathbf{x}) = 1$, $P(Y_i \leq 0|\mathbf{x}) = 0$ e $P(Y_i \leq r|\mathbf{x})$, $r = 1, \ldots, R-1$, por (2.13) encontra-se a expressão desejada da função de verossimilhança, em termos

de $\theta_1, \ldots, \theta_{R-1}$ e $\gamma$, é dada por:

$$L(\theta_1, \ldots, \theta_{R-1}, \gamma | X, Y) = \prod_{i=1}^{n} \left[ \left( \frac{\exp(\theta_1 + x_i'\gamma)}{1 + \exp(\theta_1 + x_i'\gamma)} \right)^{y_{i1}} \right] \times$$
$$\times \prod_{i=1}^{n} \left[ \left( \prod_{r=2}^{R-1} \left( \frac{\exp(\theta_r + x_i'\gamma)}{1 + \exp(\theta_r + x_i'\gamma)} - \frac{\exp(\theta_{r-1} + x_i'\gamma)}{1 + \exp(\theta_{r-1} + x_i'\gamma)} \right)^{y_{ir}} \right) \right] \times$$
$$\times \prod_{i=1}^{n} \left[ \left( \frac{1}{1 + \exp(\theta_{R-1} + x_i'\gamma)} \right)^{y_{iR}} \right], \qquad (2.15)$$

em que $\theta_1, \ldots, \theta_{R-1}$ e $\gamma' = [\gamma_1, \gamma_2, \ldots, \gamma_k]$ são os parâmetros a serem estimados; $\mathbf{x}_i' = [x_{i1}, x_{i2}, \ldots, x_{ik}]$ é o vetor de covariáveis do $i$-ésimo indivíduo da amostra e; $y_i = (y_{i1}, y_{i2}, \ldots, y_{iR})$ é o vetor de respostas do $i$-ésimo indivíduo da amostra. Aqui, $y_{ir} = 1$ indica que a resposta do $i$-ésimo indivíduo pertence à categoria $r$, $r = 1, 2, \ldots, R$.

As estimativas de máxima verossimilhança são os valores dos parâmetros que maximizam 2.15. Não existem formas analíticas fechadas para o cálculo dessas estimativas, devendo ser utilizado algoritmos numéricos, como por exemplo o de Newton-Raphson [10].

O Modelo Logístico Cumulativo apresentado em tem sido também chamado de Modelo de Chances Proporcionais, devido a uma propriedade especial: se duas populações caracterizadas por variáveis explicativas $x_1$ e $x_2$ são consideradas, a razão de chances acumuladas para as duas populações é dada por:

$$\frac{P(Y \leq r|x_1)/P(Y > r|x_1)}{P(Y \leq r|x_2)/P(Y > r|x_2)} = \exp\{(x_1 - x_2)'\gamma\},$$

que não depende de $r$.

# Capítulo 3

# Escore de Risco

O objetivo deste capítulo é apresentar um escore de risco ou índice de gravidade para os associados, para que seja possível determinar em qual grupo de risco cada associado se encontra, com o auxílio do vetor de probabilidades **P**, resultante da aplicação da regressão logística multicategórica ordinal. Esse índice de gravidade é definido a partir da distância (euclidiana) entre o ponto **P** e a melhor situação possível (onde o grupo de menor gravidade é classificado com probabilidade 1)[11].

## 3.1 Índice para dois grupos $(k = 2)$

Nesse caso, supõe-se que exite o problema de classificação entre dois grupos, $G_0$ e $G_1$, em que $G_1$ é o grupo de maior gravidade. Seja **P** o vetor de probabilidades resultante da aplicação de uma regressão logística binária com diversas variáveis explicativas.

Indicando agora o resultado da regressão logística como $(p_0, p_1) = (1 - p_1, p_1)$, note que $p_1$ é proporcional a distância euclidiana entre $\mathbf{i} = (1; 0)$ e $\mathbf{P} = (p_0; p_1)$, isto é, a razão entre distância de $\mathbf{P} = (p_0, p_1)$ à melhor situação possível, $\mathbf{i} = (1; 0)$ e a distância da pior situação possível, $(0; 1)$, à melhor situação possível, $\mathbf{i} = (1; 0)$, ou seja,

$$\frac{D\{(1;0);(p_0;p_1)\}}{D\{(1;0);(0;1)\}} = \frac{\sqrt{(1-p_0)^2+p_1^2}}{\sqrt{2}} = \frac{p_1\sqrt{2}}{\sqrt{2}} = p_1,$$

em que $D\{(a;b);(c;d)\} = [(a-c)^2 + (b-d)^2]^{1/2}$.

Isto significa que $D\{(1;0);(p_0;p_1)\}$ corresponde ao percurso completo entre o ponto observado e a melhor situação possível, passando por todos os pontos melhores

do que $\mathbf{P} = (p_0, p_1)$ [11].

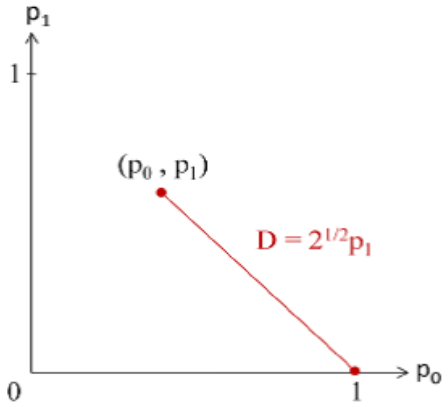


Figura 3.1: Distância percorrida pelo Índice de Gravidade para $k = 2$ grupos.

Assim, no caso de $k = 2$ grupos, o índice de gravidade denotado por $IG$, é definido por [11]:

$$IG_2(\mathbf{P}) = \frac{D\{(1;0);(p_0;p_1)\}}{D\{(1;0);(0;1)\}} = p_1.$$

## 3.2 Índice para três grupos $(k = 3)$

Considere o caso em que se tem três grupos ordenados por gravidade, de $G_0$ a $G_2$: $G_0$ é o grupo de menor gravidade, $G_1$ é o de gravidade intermediária e $G_2$ o de maior gravidade.

Analogamente ao caso de dois grupos, considere agora os vetores $\mathbf{P} = (p_0; p_1; p_2)$ e $\mathbf{i} = (1; 0; 0)$, que representam, respectivamente, o vetor do resultado da regressão logística ordinal multivariada em uma unidade amostral genérica e o vetor da melhor situação possível.

Inicialmente, o objetivo é calcular a distância percorrida entre os pontos $\mathbf{P}$ e $\mathbf{i}$, de forma a contemplar todos os pontos que sejam melhores que $\mathbf{P}$ e piores que $\mathbf{i}$.

A distância euclidiana entre $\mathbf{P}$ e $\mathbf{P_1} = (p_0; p_1 + p_2; 0)$, que contempla todos os pontos melhores que $\mathbf{P}$ e piores que $\mathbf{P_1}$ (Figura 3.2), é dada por:

$$D_1 = D\{(p_0; p_1 + p_2; 0); (p_0; p_1; p_2)\} = \sqrt{p_2^2 + p_2^2} = p_2\sqrt{2}.$$

Adicionando a distância euclidiana entre $\mathbf{P_1}$ e $\mathbf{i} = (1;0;0)$, contemplando todos os pontos piores que $\mathbf{i}$ e melhores do que $\mathbf{P_1}$ (Figura 3.2),

$$\begin{aligned} D_0 &= D\{(1;0;0);(p_0;p_1+p_2;0)\} \\ &= \sqrt{(1-p_0)^2+(p_1+p_2)^2} = \sqrt{(p_1+p_2)^2+(p_1+p_2)^2} = \sqrt{2}(p_1+p_2) \end{aligned}$$

Desta forma, a distância total entre $\mathbf{P}$ e $\mathbf{i}$ será igual a:

$$D = D_0 + D_1 = \sqrt{2}(p_1 + 2p_2). \tag{3.1}$$

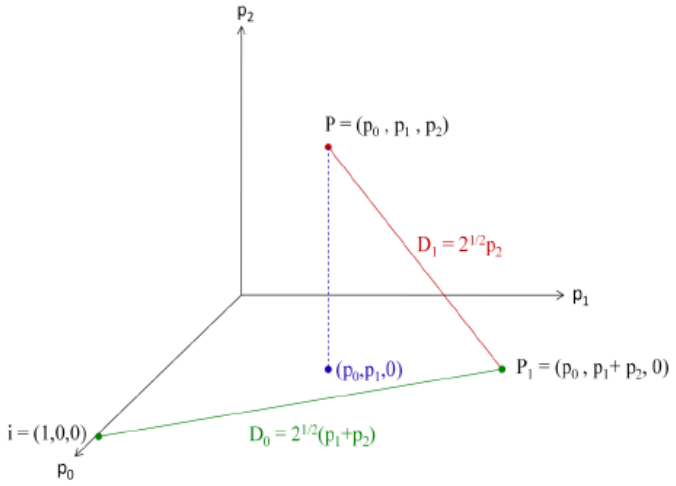


Figura 3.2: Distância percorrida pelo Índice de Gravidade para $k = 3$ grupos.

Considerando este caso de $k = 3$ grupos, o valor máximo possível atingido por (3.1) é $2\sqrt{2}$, a distância percorrida total entre (1;0;0) e (0;0;1). Dessa forma, pode-se considerar como índice de gravidade a função $IG_3$, definida por [11]:

$$\begin{aligned} IG_3(\mathbf{P}) &= \frac{D}{D\{(0;0;1);(0;1;0)\} + D\{(1;0;0);(0;1;0)\}} \\ &= \frac{\sqrt{2}(p_1+2p_2)}{2\sqrt{2}} = \frac{p_1+2p_2}{2}. \end{aligned} \tag{3.2}$$

## 3.3 Índice para $k$ grupos ($k \geq 2$)

Aqui é apresentada uma extensão dos resultados para um caso genérico quando temos $k$ ($k \geq 2$) grupos ordenados para classificação, $G_0$ a $G_{(k-1)}$. Aqui, $G_0$ é o

grupo de menor gravidade. Os demais grupos são definidos ordenadamente segundo o aumento da gravidade. Finalmente, $G_{(k-1)}$ é o grupo de maior gravidade.

Como mostrado anteriormente, o objetivo inicial é calcular a distância percorrida entre os pontos $\mathbf{P} = (p_0, p_1, ..., p_{k-1})$ e $\mathbf{i} = (1, 0, ..., 0)$, de forma a contemplar todos os pontos que sejam melhores que $\mathbf{P}$ e piores que $\mathbf{i}$. Para isso, serão calculadas:

**i.** a distância entre $\mathbf{P}$ e $\mathbf{P}_{(k-2)}$

$$\begin{aligned} D_{(k-2)} &= D\{(p_0; p_1; \ldots; p_{(k-2)} + p_{(k-1)}; 0); (p_0; p_1; \ldots; p_{(k-2)}; p_{(k-1)})\} \\ &= \sqrt{p_{(k-1)}^2 + p_{(k-1)}^2} = \sqrt{2} p_{(k-1)} \end{aligned}$$

**ii.** a distância entre $\mathbf{P}_{(k-2)}$ e $\mathbf{P}_{(k-3)}$

$$\begin{aligned} D_{(k-3)} &= D\{(p_0; p_1; \ldots; p_{(k-3)} + p_{(k-2)} + p_{(k-1)}; 0; 0); (p_0; p_1; \ldots; p_{(k-3)}; p_{(k-2)} + p_{(k-1)}; 0)\} \\ &= \sqrt{(p_{(k-2)} + p_{(k-1)})^2 + (p_{(k-2)} + p_{(k-1)})^2} = \sqrt{2}(p_{(k-2)} + p_{(k-1)}) \end{aligned}$$

**iii.** e assim por diante, até a distância entre $\mathbf{P}_1$ e $\mathbf{i}$

$$\begin{aligned} D_0 &= D\{(1; 0; \ldots; 0); (p_0; p_1 + p_2 + \ldots + p_{(k-2)} + p_{(k-1)}; 0; \ldots; 0)\} \\ &= \sqrt{(1 - p_0)^2 + \left(\sum_{i=1}^{k-1} p_i\right)^2} = \sqrt{\left(\sum_{i=1}^{k-1} p_i\right)^2 + \left(\sum_{i=1}^{k-1} p_i\right)^2} = \sqrt{2} \sum_{i=1}^{k-1} p_i \\ &= \sqrt{2}(p_1 + p_2 + \ldots + p_{(k-1)}) \end{aligned}$$

Lembrando que $p_0 + \sum_{i=1}^{k-1} p_i = 1$.

Logo, a distância total considerada é dada por:

$$D = \sum_{r=0}^{k-2} D_r = \sqrt{2} \sum_{j=1}^{k-1} j p_j = \sqrt{2}(p_1 + 2p_2 + \ldots + (k-1)p_{k-1})$$

Considerando o caso de $k$ grupos ($k \geq 2$), o valor máximo possível atingido por $D_r$ é $\sqrt{2}$, $r = 0, 1, \ldots, k-2$. Assim, a distância total entre $(1; 0; 0; \ldots; 0)$ e $(0; 0; \ldots; 0; 1)$ é dada por:

$$\max\{D\} = \sum_{r=0}^{k-2} \max\{D_r\} = \sum_{r=0}^{k-2} \sqrt{2} = (k-1)\sqrt{2}.$$

Logo, para o caso geral de $k$ ($k \geq 2$) grupos, o índice de gravidade, denotado por $IG$, é definido por [11]:

$$IG_k(\mathbf{P}) = \frac{D}{\max\{D\}}$$
$$= \frac{\sqrt{2}(p_1 + 2p_2 + \ldots + (k-1)p_{(k-1)})}{(k-1)\sqrt{2}} = \frac{p_1 + 2p_2 + \ldots + (k-1)p_{(k-1)}}{(k-1)} \quad (3.3)$$

**Nota 1**: O índice agrega pesos maiores para os grupos de maior gravidade, dando peso zero ao grupo $G_0$, de menor gravidade. O índice é dividido pela distância máxima possível simplesmente para padronizá-lo numa escala 0 a 1, que pode também ser representado por 0 a 100%.

**Nota 2**: Ao invés de um índice de gravidade (IG), poderia haver o interesse em definir um Índice de Qualidade (IQ). O índice de qualidade é o inverso do IG e ele é definido a partir da distância (euclidiana) entre o ponto $\mathbf{P}$, resultante da aplicação de uma regressão logística, e a pior situação possível. Com um desenvolvimento análogo ao índice de gravidade definido em (3.3), o Índice de Qualidade (IQ) é definido por:

$$IQ_k(\mathbf{P}) = \frac{(k-1)p_0 + (k-2)p_1 + \ldots + 2p_{(k-3)} + p_{(k-2)}}{k-1}$$

Note que:

$$IQ_k(\mathbf{P}) = 1 - IG_k(\mathbf{P})$$

Ou seja, se o interesse é trabalhar com um índice de qualidade (gravidade), ao invés de um índice de gravidade (qualidade), basta subtrair esse índice da unidade.

# Capítulo 4

# Base de Dados

## 4.1 Introdução

Para se prevenir contra a lavagem de dinheiro dentro de uma empresa, mais especificamente uma cooperativa de crédito, há a necessidade de analisar o perfil das movimentações realizadas pelos seus associados/clientes, com o intuito de verificar qualquer tipo de atipicidade, como movimentações acima da capacidade financeira dos mesmos. Apesar de uma cooperativa de crédito ser de pequeno porte ao se comparar com bancos e demais instituições financeiras, ainda sim torna-se inviável monitorar todas as transações realizadas por cada associado.

Dentro da cooperativa de crédito exite uma base que contém todas as movimentações realizadas por meio das contas correntes de pessoas físicas, consolidadas por mês, mas essa base é considerada muito grande quando se sabe que os analistas responsáveis pela verificação da existência de incompatibilidade entre transação e renda do associado analisarão individualmente cada caso, conferindo extratos, justificativas, origem e destino de recursos, entre outros. Logo, é preciso criar uma amostra, com o mínimo de perda de casos atípicos e com suspeita de lavagem de dinheiro.

Ao final da análise individual de todos os associados selecionados, as ocorrências que apresentarem movimentações realmente atípicas, em que não houver justificativa para as expressivas movimentações e possuírem indícios de lavagem de dinheiro, devem ser comunicadas ao COAF (Conselho de Controle de Atividades Financeiras), órgão que tem por finalidade disciplinar, aplicar penas administrativas, receber, examinar e identificar as ocorrências suspeitas de atividades ilícitas previstas na lei, sem prejuízo de outros órgãos ou entidades.

Como aqui o objetivo é analisar associados do Sicoob Confederação que movi-

mentam acima da sua capacidade financeira, é válido ressaltar que valores movimentados no mês analisado inferiores a renda do associado ou valores muito baixos não foram considerados, visto que uma pessoa física, quando não considerado eventos excepcionais, pode movimentar no mês no máximo o valor da sua renda e, no ano, em torno de 13 vezes a renda, considerando o recebimento do $13^{o}$ salário.

## 4.2 Descrição do Estudo

A base de dados utilizada foi disponibilizada pelo Sicoob Confederação, referente à todas as movimentações financeiras realizadas à crédito por meio das Contas Correntes de pessoas físicas associadas à instituição, entre Fevereiro/2016 a Janeiro/2017. A amostra consistia em 82.531 associados, cujas movimentações consolidadas são classificadas com risco "baixo", "médio" e "alto", cujas definições foram feitas de acordo com os seguintes critérios:

**(a)** BAIXO (Resposta=1): associado classificado como baixo risco, ou seja, que não apresenta movimentações suspeitas de lavagem de dinheiro e não deve ser encaminhado para análise aprofundada;

**(b)** MÉDIO (Resposta=2): associado classificado como médio risco, apresentando movimentações atípicas em relação a sua capacidade financeira e, consequentemente foi gerada uma ocorrência para ser realizada uma análise aprofundada, mas suas movimentações são justificadas;

**(c)** ALTO (Resposta=3): associado classificado como alto risco, que possui movimentações atípicas em relação a sua capacidade financeira, foi gerada uma ocorrência para ser realizada uma análise aprofundada e, devido a incompatibilidades e falta de justificativas para as atipicidades, o associado é comunicado ao COAF.

Uma observação importante a se fazer é que, devido a grande quantidade de observações na base de dados, não é viável a realização de testes de hipóteses, já que, em todos eles, a hipótese nula seria rejeitada. Com isso, as decisões para chegar em resultados foram baseadas, de forma geral, em técnicas gráficas.

Para o estudo, apenas 6 variáveis se mostraram relevantes para explicar a variável resposta. A classificação das variáveis preditoras consideradas na análise é dada a seguir, sendo que o significado de cada uma foi omitido diante do sigilo na forma de identificar transações ilícitas, suspeitas de lavagem de dinheiro.

- $X_1$: variável quantitativa contínua;
- $X_2$: variável quantitativa contínua;
- $X_3$: variável quantitativa contínua;
- $X_4$: variável quantitativa contínua;
- $X_5$: variável qualitativa nominal (binária);
- $X_6$: variável qualitativa nominal (binária).

As variáveis $X_1$ e $X_2$ têm grande importância por representarem os valores movimentados ao longo do ano, não só considerando o mês de referência, como as variáveis $X_3$ e $X_4$, já que assim variações bruscas podem ser identificadas. Contudo, essas duas variáveis poderiam apresentar valores negativos. Nesse caso, todos os valores negativos foram substituídos por 1, como sendo o menor valor positivo observado e podendo ser tratados como tendo o mesmo risco, ao desconsiderar o valor das outras variáveis. Já as variáveis $X_5$ e $X_6$ são indicadoras da presença de características fundamentais para a identificação do perfil cadastrado do associado.

Vale ressaltar que em todas as variáveis quantitativas do estudo, $X_1$ a $X_4$, foi realizada uma transformação logarítmica, para suavizar seus valores discrepantes, descritos nas Tabelas 4.1, 4.2, 4.3 e 4.4, respectivamente.

Tabela 4.1: Medidas resumo da variável $X_1$ por categoria de resposta $Y$.

| $\exp\{X_1\}$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| Mínimo | 1 | 11.620 | 1 | 1 |
| Média | 191.500 | 763.900 | 1.040.000 | 215.500 |
| Máximo | 44.300.000 | 17.320.000 | 17.190.000 | 44.300.000 |
| Mediana | 100.500 | 500.600 | 622.300 | 105.400 |

Tabela 4.2: Medidas resumo da variável $X_2$ por categoria de resposta $Y$.

| $\exp\{X_2\}$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| Mínimo | 1 | 1 | 1 | 1 |
| Média | 382.000 | 1.254.000 | 1.763.000 | 367.600 |
| Máximo | 54.360.000 | 28.270.000 | 33.170.000 | 54.360.000 |
| Mediana | 173.200 | 822.300 | 1.044.000 | 181.100 |

Tabela 4.3: Medidas resumo da variável $X_3$ por categoria de resposta $Y$.

| $\exp\{X_3\}$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| Mínimo | 1 | 2,09 | 1,58 | 1 |
| Média | 680,70 | 7.705 | 9.590 | 956,7 |
| Máximo | 826.900 | 3.213.000 | 1.839.000 | 3.213.000 |
| Mediana | 8,05 | 49,98 | 67,22 | 8,46 |

Tabela 4.4: Medidas resumo da variável $X_4$ por categoria de resposta $Y$.

| $\exp\{X_4\}$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| Mínimo | 8,22 | 6.528 | 3.873 | 8,22 |
| Média | 56.100 | 237.700 | 293.500 | 63.330 |
| Máximo | 12.480.000 | 9.114.000 | 7.335.000 | 12.480.000 |
| Mediana | 25.610 | 130.900 | 161.400 | 26.800 |

## 4.3 Análise Descritiva dos Dados

De acordo com o que foi dito na Seção 4.2, diante da grande amplitude entre os valores mínimos e máximos das variáveis $X_1$ a $X_4$, foi realizada uma transformação logarítmica para reduzir as discrepâncias entre as observações. Assim, durante o estudo, as variáveis transformadas foram consideradas como: $\log(X_1)$, $\log(X_2)$, $\log(X_3)$ e $\log(X_4)$.

Sendo $Y$ a variável resposta considerada no modelo, a Tabela 4.5 apresenta seus valores absolutos e relativos. De acordo com ela, a maioria da amostra é representada por associados classificados como baixo risco, enquanto o grupo mais grave consiste em apenas 1,25% da base.

Tabela 4.5: Valores absolutos e relativos da variável resposta $Y$

| Resposta | Frequência | % |
|---|---|---|
| 1 | 79.564 | 96,40 |
| 2 | 1.937 | 2,35 |
| 3 | 1.030 | 1,25 |

As Tabelas 4.6, 4.7, 4.8 e 4.9 apresentam as medidas resumo da transformação das variáveis utilizadas no estudo. Usando como referência essas tabelas e os box-plots das Figuras 4.1, 4.2, 4.3 e 4.4, é possível verificar que existe uma melhor distribuição dos dados ao se fazer o logaritmo das mesmas.

Tabela 4.6: Medidas resumo da variável contínua $\log(X_1)$ por categoria de resposta $Y$.

| $\log(X_1)$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| Mínimo | 0,00 | 9,36 | 0,00 | 0,00 |
| Média | 10,34 | 13,16 | 13,39 | 10,45 |
| Máximo | 17,61 | 16,67 | 16,66 | 17,61 |
| Mediana | 11,52 | 13,12 | 13,34 | 11,57 |

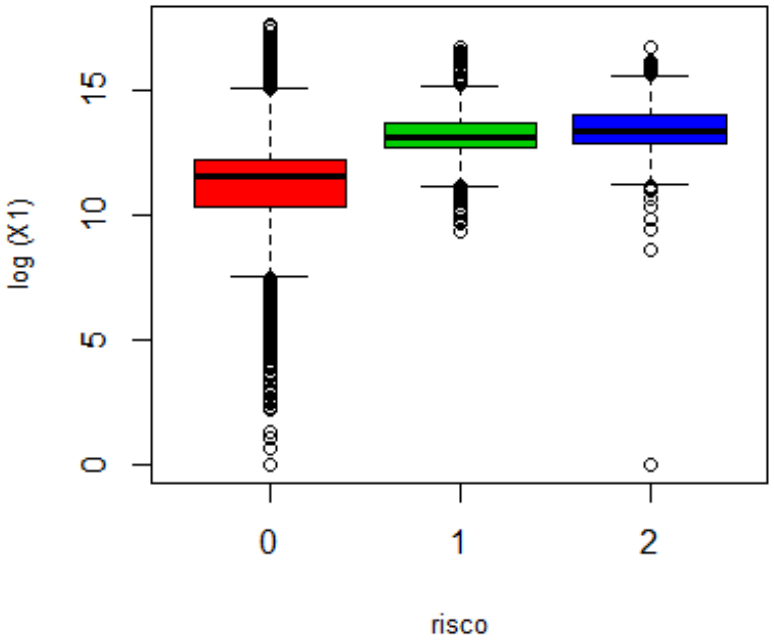


Figura 4.1: Boxplot da variável $\log(X_1)$ por categoria de resposta.

De acordo com as informações acima, é possível perceber que a média dos valores das categorias de médio e alto risco, da variável $\log(X_1)$, estão mais próximas do que a média da categoria de baixo risco, o que indica que essa variável é capaz de discriminar melhor os associados que não apresentam movimentações com indícios de lavagem de dinheiro.

Tabela 4.7: Medidas resumo da variável contínua $\log(X_2)$ por categoria de resposta $Y$.

| $\log(X_2)$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| Mínimo | 0,00 | 0,00 | 0,00 | 0,00 |
| Média | 10,11 | 13,60 | 13,81 | 10,24 |
| Máximo | 17,81 | 17,16 | 14,43 | 12,83 |
| Mediana | 12,06 | 13,62 | 13,86 | 12,11 |

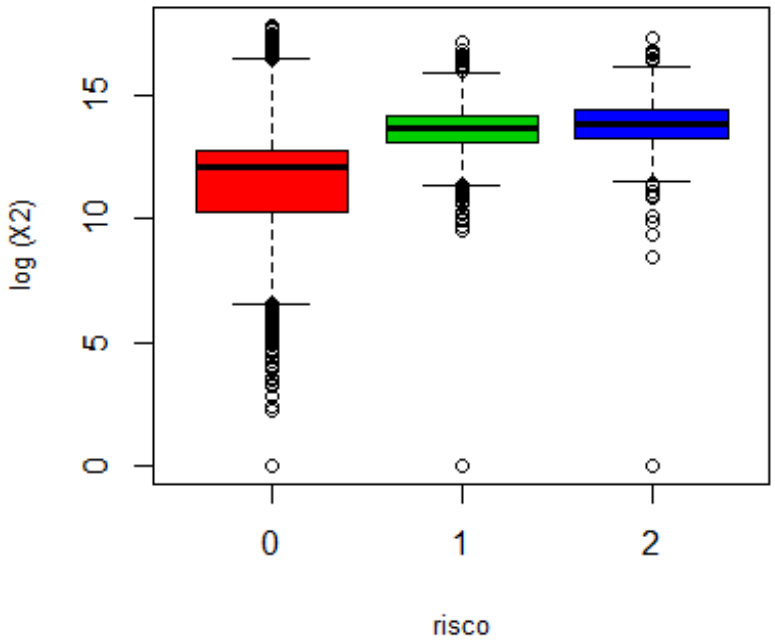


Figura 4.2: Boxplot da variável $\log(X_2)$ por categoria de resposta.

A variável transformada $\log(X_2)$ tem como valor mínimo, em cada uma das categorias de $Y$. Isso se deve, em grande parte, pela transformação de valores negativos em zero, por motivo de tratamento dos dados para melhorar o ajuste do modelo. Da mesma forma do $\log(X_1)$, a variável é capaz de discriminar melhor os associados classificados como baixo risco daqueles como médio e alto. Curiosamente, o maior valor da variável em questão no grupo em que $Y = 1$ é superior aos máximos das categorias $Y = 2$ e $Y = 3$.

Tabela 4.8: Medidas resumo da variável contínua $\log(X_3)$ por categoria de resposta $Y$.

| $\log(X_3)$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| Mínimo | 0,00 | 0,73 | 0,45 | 0,00 |
| Média | 2,27 | 4,11 | 4,37 | 2,34 |
| Máximo | 13,63 | 14,98 | 14,42 | 14,98 |
| Mediana | 2,08 | 3,91 | 4,20 | 2,13 |

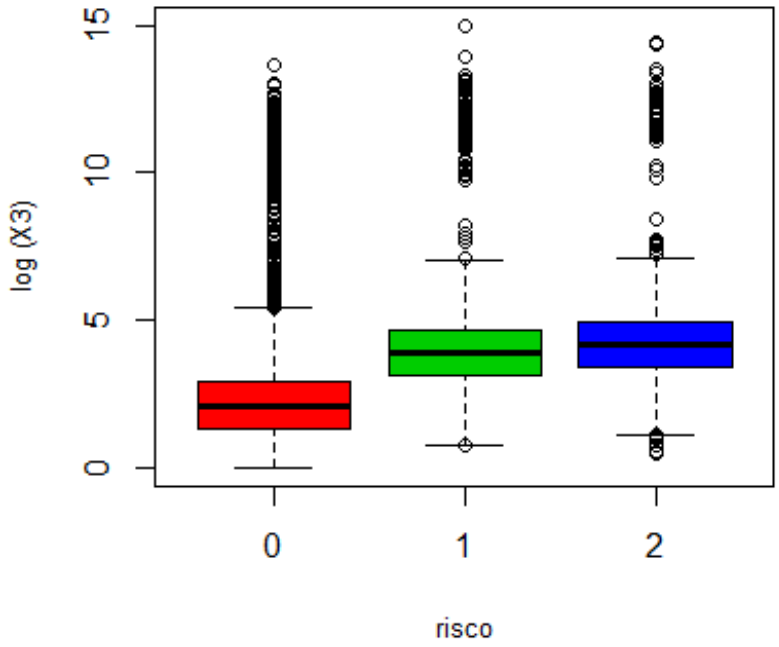


Figura 4.3: Boxplot da variável $\log(X_3)$ por categoria de resposta.

Ao comparar os valores obtidos das médias da variável $\log(X_3)$ e as demais, é possível perceber que esses são menores, característica que levou à uma maior importância na utilização dessa variável no estudo, sendo capaz também de discriminar os associados de acordo com seu nível de risco.

Tabela 4.9: Medidas resumo da variável contínua $\log(X_4)$ por categoria de resposta $Y$.

| $\log(X_4)$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| Mínimo | 2,10 | 8,78 | 8,26 | 2,10 |
| Média | 10,25 | 11,86 | 12,08 | 10,31 |
| Máximo | 16,34 | 16,03 | 15,81 | 16,34 |
| Mediana | 10,15 | 11,78 | 11,99 | 10,20 |

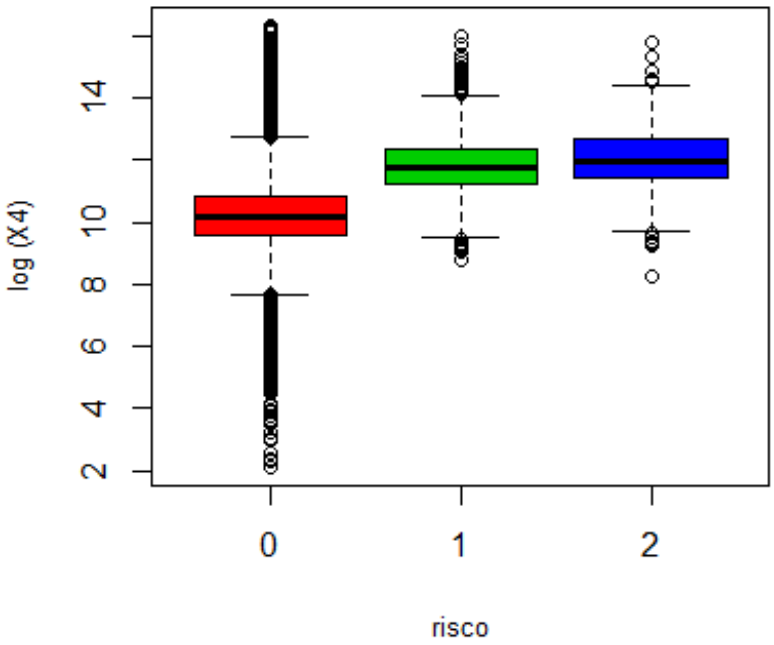


Figura 4.4: Boxplot da variável $\log(X_4)$ por categoria de resposta.

Considerando a Tabela 4.9 e Figura 4.4, existe uma grande discrepância entre os valores mínimos das categorias de médio e alto risco, com a de baixo risco, grande indicativo de que o $\log(X_4)$ é capaz de discriminar esse último grupo. Apesar do maior valor do grupo em que $Y = 3$ ser inferior aos demais, as médias apresentadas são coerentes com a gravidade dos níveis da resposta $Y$.

A seguir, é mostrada nas Tabelas 4.10 e 4.11 as frequências relativas às covariáveis qualitativas que permaneceram na base.

Tabela 4.10: Distribuição conjunta da variável binária $X_5$ e $Y$.

| $X_5$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| 0 | 78.993 (99%) | 1.898 (98%) | 1.008 (98%) | 81.899 (99%) |
| 1 | 571 (1%) | 39 (2%) | 22 (2%) | 632 (1%) |
| Total | 79.564 (100%) | 1.937 (100%) | 1.030 (100%) | 82.531 (100%) |

Tabela 4.11: Distribuição conjunta da variável binária $X_6$ e $Y$.

| $X_6$ | Baixo Risco | Médio Risco | Alto Risco | Total |
|---|---|---|---|---|
| 0 | 55.712 (70%) | 1.388 (72%) | 826 (80%) | 57.926 (70%) |
| 1 | 23.852 (30%) | 549 (28%) | 204 (20%) | 24.605 (30%) |
| Total | 79.564 (100%) | 1.937 (100%) | 1.030 (100%) | 82.531 (100%) |

Através da Tabela 4.10, percebe-se que, na amostra, associados com a presença da característica representada por $\log(X_5)$ possuem maior probabilidade de serem classificados como médio ou alto risco. De forma análoga, associados com a característica da variável indicadora $\log(X_6)$ apresentam maior probabilidade de serem classificadas como baixo risco.

# Capítulo 5

# Resultados

## 5.1 Modelo de Regressão Logística Ordinal

O Modelo Logístico Ordinal pode ser utilizado quando a variável resposta possui mais de duas categorias, levando em consideração a ordenação que pode existir entre elas. Deseja-se ajustar esse modelo para que seja possível definir o risco de determinado associado, levando em consideração as informações existentes (variáveis $X_1$ a $X_6$).

As estimativas de máxima verossimilhança de $\gamma$, assim como erro padrão para o modelo logístico ordinal, são apresentadas na Tabela 5.1 e o modelo (2.13) pode ser escrito como:

$$Logit[P(Y \leq 1|\mathbf{x})] = \theta_1 - 0,58X_1 - 0,31X_2 - 0,36X_3 - 0,58X_4 - 2,14X_5 + 0,43X_6$$
$$Logit[P(Y \leq 2|\mathbf{x})] = \theta_2 - 0,58X_1 - 0,31X_2 - 0,36X_3 - 0,58X_4 - 2,14X_5 + 0,43X_6$$

Tabela 5.1: Estimativas dos parâmetros do modelo Logístico Ordinal

| | Estimativas | EP |
|---|---|---|
| $\theta_1$ | 21,9532 | 0,2784 |
| $\theta_2$ | 23,2773 | 0,2838 |
| $X_1$ | 0,5870 | 0,059971 |
| $X_2$ | 0,3130 | 0,045674 |
| $X_3$ | 0,3616 | 0,009519 |
| $X_4$ | 0,5808 | 0,029173 |
| $X_5$ | 2,1453 | 0,176457 |
| $X_6$ | -0,4386 | 0,052149 |

Então,

$$P(Y \leq 1) = \frac{\exp(21,95 - 0,58X_1 - 0,31X_2 - 0,36X_3 - 0,58X_4 - 2,14X_5 + 0,43X_6)}{1 + \exp(21,95 - 0,58X_1 - 0,31X_2 - 0,36X_3 - 0,58X_4 - 2,14X_5 + 0,43X_6)}$$

e

$$P(Y \leq 2) = \frac{\exp(23,27 - 0,58X_1 - 0,31X_2 - 0,36X_3 - 0,58X_4 - 2,14X_5 + 0,43X_6)}{1 + \exp(23,27 - 0,58X_1 - 0,31X_2 - 0,36X_3 - 0,58X_4 - 2,14X_5 + 0,43X_6)}$$

Assim, quanto maior o valor das variáveis $X_1$, $X_2$, $X_3$ ou $X_4$, maior será o risco relativo de um associado ser classificado como "Alto Risco". Da mesma forma, a presença de $X_5$ tende a tornar o risco do associado pior, dado que o risco relativo de se tornar baixo é mais reduzido do que o de se tornar médio. Por outro lado, a presença de $X_6$ parece aumentar mais o risco relativo de ser um associado de "Baixo Risco".

Considerando um exemplo em que o associado possui as movimentações ($X_1$ a $X_4$) e classificações ($X_5$ e $X_6$) representadas na Tabela 5.2, é possível ajustar o modelo de Regressão Logística Ordinal para determinar o risco deste associado na empresa.

Tabela 5.2: Variáveis de um determinado associado.

| Variável | $X_i$ | $\log(X_i)$ |
|---|---|---|
| *variável 1* | 894.295,07 | 13,70 |
| *variável 2* | 1.279.227,63 | 14,06 |
| *variável 3* | 73,36 | 4,29 |
| *variável 4* | 377.491,34 | 12,84 |
| *variável 5* | 1 | n/a |
| *variável 6* | 0 | n/a |

Utilizando como referência o resultado do modelo encontrado na Seção 5.1, tem-se que:

$$\begin{aligned} Logit[P(Y \leq 1|\mathbf{x})] &= 21,95 - 0,58 \times 13,70 - 0,31 \times 14,06 - 0,36 \times 4,29 - \\ &\quad - 0,58 \times 12,84 - 2,14 \times 1 + 0,43 \times 0 \\ &= -1,4862 \qquad \text{e} \end{aligned}$$

$$\begin{aligned} Logit[P(Y \leq 2|\mathbf{x})] &= 23,27 - 0,58 \times 13,70 - 0,31 \times 14,06 - 0,36 \times 4,29 - \\ &\quad - 0,58 \times 12,84 - 2,14 \times 1 + 0,43 \times 0 \\ &= -0,1662 \end{aligned}$$

Os valores acima encontrados implicam em:

$$P(Y \leq 1) = \frac{\exp(-1,4862)}{1+\exp(-1,4862)} = 0,1845$$

$$P(Y \leq 2) = \frac{\exp(-0,1662)}{1+\exp(-0,1662)} = 0,4585$$

Sabendo que as probabilidades acima são acumuladas, então, $P(Y = j) = P(Y \leq j) - P(Y \leq j-1)$, $j = 2,3$. Assim, a probabilidade desse associado ser identificado como sendo "baixo risco" é $P(Y = 1) = P(Y \leq 1) = 0,1845$.

De forma análoga, a probabilidade do mesmo associado ser classificado como baixo risco ou médio risco é $P(Y \leq 2) = P(Y = 1) + P(Y = 2) = 0,4585$. Isso significa que a probabilidade do associado ser identificado como "médio risco" é:

$$P(Y = 2) = P(Y \leq 2) - P(Y \leq 1) = 0,4585 - 0,1845 = 0,2740$$

Agora, sabendo que $P(Y \leq 3) = 1$, já que o associado só pode estar inserido em uma das três categorias em análise, a probabilidade do associado ser classificado como alto risco é $P(Y = 3) = 1 - P(Y \leq 2) = 0,5415$.

Tabela 5.3: Probabilidade de classificação do associado por risco.

| Risco | Probabilidade |
|---|---|
| Baixo | 18,45% |
| Médio | 27,40% |
| Alto | 54,15% |

Como $P(Y = 3) > \max\{P(Y = 2), P(Y = 3)\}$, o modelo de regressão logística ordinal classificará esse indivíduo como alto risco.

Realizando esse mesmo cálculo para todos os indivíduos da amostra, a Tabela 5.4 apresenta a matriz de confusão com as classificações reais e preditas segundo o modelo de Regressão Logística Ordinal.

Tabela 5.4: Matriz de confusão baseada na maior probabilidade.

**Valores Predito e Observado**

| **Observado** | **Predito** | | | **Total** |
|---|---|---|---|---|
| | Baixo | Médio | Alto | |
| Baixo | 79.367 | 0 | 197 | 79.564 |
| Médio | 1.818 | 0 | 119 | 1.937 |
| Alto | 915 | 0 | 115 | 1.030 |
| **Total** | 82.100 | 0 | 431 | 82.531 |

**Taxa de Acerto Geral**

| **Observado** | **Predito** | | | **Total** |
|---|---|---|---|---|
| | Baixo | Médio | Alto | |
| Baixo | 96,17% | 0% | 0,24% | 96,40% |
| Médio | 2,20% | 0% | 0,14% | 2,35% |
| Alto | 1,11% | 0% | 0,14% | 1,25% |
| **Total** | 99,47% | 0% | 0,52% | 100% |

Observando a Tabela 5.4, percebe-se que o modelo apresentou um bom ajuste dos dados ao considerar a predição dos associados classificados como baixo risco, tendo um acerto de $(^{79.367}/_{79.564}) = 99,75\%$. Apesar desse resultado, o modelo não foi capaz de prever de forma eficaz os associados que deveriam ser classificados como médio ou alto risco. Isso acontece pelo fato de que o ponto de corte determinado leva em consideração que as probabilidades de um associado ser classificado com cada nível de $Y$ são as mesmas e, como de forma geral a $P(Y = 1) = 96,40\%$, conforme indicado na Tabela 4.5, o modelo não apresenta um bom ajuste.

As classificações apresentadas pela Tabela 5.4 foram feitas considerando a maior probabilidade estimada. Contudo, como é possível ver, o número de associados em cada faixa de risco não é balanceado. Desta forma, uma regra de classificação com base na maior probabilidade estimada não trará bons resultados. Assim, uma alternativa é classificar os associados segundo a prevalência de cada classe observada na amostra.

Por se tratar de três classes, optou-se inicialmente por classificar os associados de maior risco (já que é a principal classe de interesse no estudo), classificando posteriormente as demais classes. Essa classificação foi realizada com base na prevalência de cada classe, apresentada pela Tabela 4.5. Segue abaixo o algorítimo de classificação:

1. Ao calcular a probabilidade do indivíduo ser classificado como alto risco, se $p_3 > 0,01248$, é possível predizer que $Y = 3$, por ter probabilidade superior à probabilidade de prevalência da categoria;

2. Caso contrário, calcular a probabilidade do indivíduo ser classificado em cada uma das outras duas categorias. Se

$$\frac{p_2}{p_1 + p_2} > \frac{0,02347}{0,96405 + 0,02347} = 0,02376,$$

   então o indivíduo pode ser classificado como médio risco, em que $Y = 2$;

3. Ao não se enquadrar em nenhum dos itens acima, automaticamente o indivíduo será classificado com resposta $Y = 1$, ou seja, baixo risco.

As matrizes de confusão apresentadas na Tabela 5.5 mostram que o percentual total de acerto foi de 79,71% indicando um bom ajuste do Modelo de Regressão Logística Ordinal. O percentual de associados "alto risco" erroneamente classificados como baixo ou médio risco é de 9,22%. Consequentemente, na predição do grupo mais grave, o acerto do modelo é de 90,78%, o que considera essa técnica capaz de discriminar principalmente os casos que devem ser comunicados ao COAF após a análise de suas informações.

Tabela 5.5: Matriz de confusão baseada na prevalência.

**Valores Predito e Observado**

| **Observado** | **Predito** | | | **Total** |
|---|---|---|---|---|
| | Baixo | Médio | Alto | |
| Baixo | 64.752 | 3.886 | 10.926 | 79.564 |
| Médio | 179 | 102 | 1.656 | 1.937 |
| Alto | 58 | 37 | 935 | 1.030 |
| **Total** | 64.989 | 4.025 | 13.517 | 82.531 |

**Taxa de Acerto Geral**

| **Observado** | **Predito** | | | **Total** |
|---|---|---|---|---|
| | Baixo | Médio | Alto | |
| Baixo | 78,46% | 4,71% | 13,23% | 96,40% |
| Médio | 0,22% | 0,12% | 2,01% | 2,35% |
| Alto | 0,07% | 0,05% | 1,13% | 1,25% |
| **Total** | 78,75% | 4,88% | 16,37% | 100% |

Diante dessas informações, é possível encontrar o escore de risco deste associado. Assim, considerando $p_0 = P(Y = 1)$, $p_1 = P(Y = 2)$ e $p_2 = P(Y = 3)$ na fórmula (3.2), o escore é dado por:

$$Escore = \frac{p_1 + 2p_2}{2} = 0,6785. \tag{5.1}$$

A Tabela 5.6 apresenta o resumo dos valores gerais dos escores segundo o nível de risco dos associados. Note que, como esperado, os associados de maior risco apresentaram, em média, maior escore de risco.

Tabela 5.6: Resumo dos valores dos escores segundo o risco.

| **Medidas** | **Risco** | | |
|---|---|---|---|
| | Baixo | Médio | Alto |
| Mínimo | 0,0000000 | 0,0000719 | 0,0000002 |
| 1º Quartil | 0,0009326 | 0,0424200 | 0,0650400 |
| Mediana | 0,0048170 | 0,0932400 | 0,1417000 |
| Média | 0,0185200 | 0,1583000 | 0,2174000 |
| 3º Quartil | 0,0142700 | 0,2061000 | 0,3014000 |
| Máximo | 0,9766000 | 0,9917000 | 0,9921000 |

Dado o limite máximo de M associados a serem enviados para as cooperativas para a análise individual, uma aplicação direta do índice de risco é simplesmente alertar os M associados com os maiores escores de risco. Esse procedimento traz como vantagem tornar o processo objetivo (não tendo mais necessidade da análise subjetiva), mas existe a desvantagem de não ser possível controlar o erro $\epsilon$.

Agora, considerando M=10.000, o ponto de corte é definido como $corte = 0.04154211$, o que implica que $\epsilon = 0,0864$. Isto é, estima-se que 8,64% dos associados classificados como alto risco não seriam comunicados ao COAF.

### 5.1.1 Definição do Ponto de Corte do Escore de Risco

O escore proposto se mostrou capaz de ordenar (ranquear) os associados segundo o seu risco. Visto que o erro de não alertar um associado classificado como alto risco é muito mais grave do que alertar aquele classificado como baixo risco, a proposta é definir um ponto de corte que permita discriminar este último. A situação ideal é a definição de um ponto de corte em que a ocorrência de falsos negativos fosse controlada.

O ponto de corte é definido como o $\epsilon$-ésimo quantil dos escores de risco dos associados de alto risco. Esse corte garante que apenas $\epsilon \times 100\%$ dos casos que foram comunicados ao COAF não sejam selecionados.

Nesse exemplo, considerando $\epsilon = 0,05$ (corte em que se estima que 5% dos associados que seriam comunicados não serão alertados), o ponto de corte é dado por $corte = 0,01916968$. Assim, todos os associados cujo escore de risco é menor do que 0,01916968 seriam automaticamente filtrados e considerados com movimentações sem indícios de lavagem de dinheiro. Esse corte corresponde a uma redução de $77,58\%$ do número de associados encaminhados para análise subjetiva. Considerando a amostra inicial de 82.531 associados, essa redução corresponde a 64.028, restando apenas 18.503 associados para análise subjetiva. A Tabela 5.7 apresenta o ponto de corte do escore de risco e seu respectivo percentual de redução segundo o percentual de erro fixado.

Tabela 5.7: Valor do escore de corte e % de redução da amostra baseados na % do erro.

| %Erro | Corte | Redução | %Erro | Corte | Redução |
|---|---|---|---|---|---|
| 1 | 0,002714 | 0,369098 | 16 | 0,041502 | 0,878724 |
| 2 | 0,007991 | 0,600187 | 17 | 0,043573 | 0,883389 |
| 3 | 0,012807 | 0,703227 | 18 | 0,045450 | 0,887715 |
| 4 | 0,016276 | 0,747719 | 19 | 0,047738 | 0,892234 |
| 5 | 0,019170 | 0,775818 | 20 | 0,050519 | 0,897348 |
| 6 | 0,021918 | 0,796731 | 21 | 0,052770 | 0,901201 |
| 7 | 0,023551 | 0,808072 | 22 | 0,055338 | 0,905163 |
| 8 | 0,026566 | 0,825157 | 23 | 0,058973 | 0,910276 |
| 9 | 0,028800 | 0,835710 | 24 | 0,061354 | 0,913778 |
| 10 | 0,031086 | 0,845064 | 25 | 0,065040 | 0,917898 |
| 11 | 0,032327 | 0,849778 | 26 | 0,067390 | 0,920563 |
| 12 | 0,034646 | 0,858005 | 27 | 0,069068 | 0,922393 |
| 13 | 0,036013 | 0,862694 | 28 | 0,071801 | 0,925022 |
| 14 | 0,037789 | 0,867953 | 29 | 0,073169 | 0,926210 |
| 15 | 0,039272 | 0,872714 | 30 | 0,075627 | 0,928609 |

### 5.1.2 Cálculo dos Intervalos de Confiança das estimativas do erro e percentual de redução da amostra

Tendo em vista que as estimativas do erro (erro de não alertar um associado de alto risco) e dos percentuais de redução da amostra estão sujeitas a desvios amostrais, esses desvios foram estimados por meio da técnica de reamostragem Bootstrap não paramétrica [8]. Essa técnica visa obter $B$ reamostragens, selecionadas de forma independente e com reposição da amostra original.

Desta forma, para cada reamostragem (que tem o mesmo tamanho da amostra original), calcula-se o ponto de corte, estimativa do erro e do percentual da redução da amostra, resultando em $B$ estimativas do erro e percentual da redução. Assim, limites inferiores e superiores de um intervalo $(1 - \alpha) \times 100\%$ de confiança são dados, respectivamente, pelos quantis $\alpha/2$ e $1 - {}^{\alpha}/_{2}$ dos $B$ resultados obtidos. Neste trabalho considerou-se $B = 1000$ reamostragens.

A Tabela 5.8 apresenta os intervalos de 95% de confiança do percentual da redução da amostra ao fixar erro e a Tabela 5.9 apresenta os intervalos de 95% de confiança do erro ao se fixar o percentual de redução da amostra.

Tabela 5.8: Intervalos de Confiança para os cortes dos escores e suas respectivas reduções na amosta segundo o erro.

| **%Erro** | **Corte** | **Redução** | |
|---|---|---|---|
| | | Pontual | IC(95%) |
| 1 | 0,002714 | 0,369098 | (0,265639 ; 0,569915) |
| 2 | 0,007991 | 0,600187 | (0,465576 ; 0,705492) |
| 3 | 0,012807 | 0,703227 | (0,616869 ; 0,753679) |
| 4 | 0,016276 | 0,747719 | (0,704509 ; 0,784189) |
| 5 | 0,019170 | 0,775818 | (0,733383 ; 0,803965) |
| 6 | 0,021918 | 0,796731 | (0,764714 ; 0,816112) |
| 7 | 0,023551 | 0,808072 | (0,787651 ; 0,832852) |
| 8 | 0,026566 | 0,825157 | (0,802049 ; 0,841337) |
| 9 | 0,028800 | 0,835710 | (0,813216 ; 0,850411) |
| 10 | 0,031086 | 0,845064 | (0,826884 ; 0,857642) |
| 11 | 0,032327 | 0,849778 | (0,836215 ; 0,863196) |
| 12 | 0,034646 | 0,858005 | (0,845305 ; 0,867215) |
| 13 | 0,036013 | 0,862694 | (0,851108 ; 0,873769) |
| 14 | 0,037789 | 0,867953 | (0,856211 ; 0,879077) |
| 15 | 0,039272 | 0,872714 | (0,861384 ; 0,884953) |
| 16 | 0,041502 | 0,878724 | (0,865707 ; 0,889998) |
| 17 | 0,043573 | 0,883389 | (0,872367 ; 0,893902) |
| 18 | 0,045450 | 0,887715 | (0,876694 ; 0,899263) |
| 19 | 0,047738 | 0,892234 | (0,881957 ; 0,902854) |
| 20 | 0,050519 | 0,897348 | (0,886840 ; 0,907078) |
| 21 | 0,052770 | 0,901201 | (0,891212 ; 0,912362) |
| 22 | 0,055338 | 0,905163 | (0,895674 ; 0,916020) |
| 23 | 0,058973 | 0,910276 | (0,899042 ; 0,919545) |
| 24 | 0,061354 | 0,913778 | (0,903247 ; 0,922140) |
| 25 | 0,065040 | 0,917898 | (0,906845 ; 0,924356) |
| 26 | 0,067390 | 0,920563 | (0,912222 ; 0,925774) |
| 27 | 0,069068 | 0,922393 | (0,914889 ; 0,928451) |
| 28 | 0,071801 | 0,925022 | (0,918090 ; 0,931080) |
| 29 | 0,073169 | 0,926210 | (0,920804 ; 0,933310) |
| 30 | 0,075627 | 0,928609 | (0,922973 ; 0,934800) |

Tabela 5.9: Cortes e erros pontuais e seus IC(95%) segundo a redução da amostra.

| **%Redução** | **Corte** | **Erro** | |
|---|---|---|---|
| | | Pontual | IC(95%) |
| 10 | 0,000004 | 0,000971 | (0,000000 ; 0,003012) |
| 20 | 0,000493 | 0,002913 | (0,000912 ; 0,007744) |
| 30 | 0,001677 | 0,007767 | (0,002857 ; 0,013385) |
| 40 | 0,003222 | 0,011650 | (0,005664 ; 0,017017) |
| 50 | 0,005213 | 0,015534 | (0,008661 ; 0,024225) |
| 60 | 0,007985 | 0,020388 | (0,010846 ; 0,029150) |
| 70 | 0,012610 | 0,028155 | (0,019788 ; 0,039424) |
| 71 | 0,013241 | 0,031068 | (0,021073 ; 0,042614) |
| 72 | 0,013948 | 0,034951 | (0,023345 ; 0,045980) |
| 73 | 0,014698 | 0,036893 | (0,024900 ; 0,048151) |
| 74 | 0,015551 | 0,039806 | (0,027356 ; 0,050939) |
| 75 | 0,016485 | 0,041748 | (0,030268 ; 0,055020) |
| 76 | 0,017458 | 0,043689 | (0,032107 ; 0,058825) |
| 77 | 0,018494 | 0,047573 | (0,035271 ; 0,062265) |
| 78 | 0,019708 | 0,052427 | (0,039632 ; 0,064901) |
| 79 | 0,021005 | 0,056311 | (0,042547 ; 0,070858) |
| 80 | 0,022367 | 0,063107 | (0,048971 ; 0,078657) |
| 81 | 0,023864 | 0,073786 | (0,056678 ; 0,087851) |
| 82 | 0,025634 | 0,076699 | (0,062434 ; 0,094596) |
| 83 | 0,027603 | 0,084466 | (0,068825 ; 0,102342) |
| 84 | 0,029763 | 0,096117 | (0,079158 ; 0,115005) |
| 85 | 0,032392 | 0,110680 | (0,091574 ; 0,129447) |
| 86 | 0,035306 | 0,123301 | (0,105871 ; 0,147130) |
| 87 | 0,038386 | 0,142718 | (0,122602 ; 0,166507) |
| 88 | 0,042129 | 0,162136 | (0,139410 ; 0,185448) |
| 89 | 0,046622 | 0,186408 | (0,162065 ; 0,209526) |
| 90 | 0,052030 | 0,207767 | (0,183432 ; 0,233881) |
| 91 | 0,058824 | 0,229126 | (0,203278 ; 0,256801) |
| 92 | 0,066839 | 0,257282 | (0,233753 ; 0,284592) |
| 93 | 0,076906 | 0,302913 | (0,276779 ; 0,329626) |
| 94 | 0,088714 | 0,348544 | (0,315569 ; 0,375483) |
| 95 | 0,107087 | 0,402913 | (0,373594 ; 0,434837) |
| 96 | 0,131247 | 0,474757 | (0,445631 ; 0,503498) |
| 97 | 0,169282 | 0,562136 | (0,534813 ; 0,593137) |
| 98 | 0,231819 | 0,663107 | (0,636614 ; 0,690942) |
| 99 | 0,364810 | 0,800000 | (0,776913 ; 0,821434) |

As estimativas pontuais dos erros e percentuais de redução da amostra, com seus respectivos intervalos de confiança, segundo o erro e o percentual de redução são apresentados na Figura 5.1.

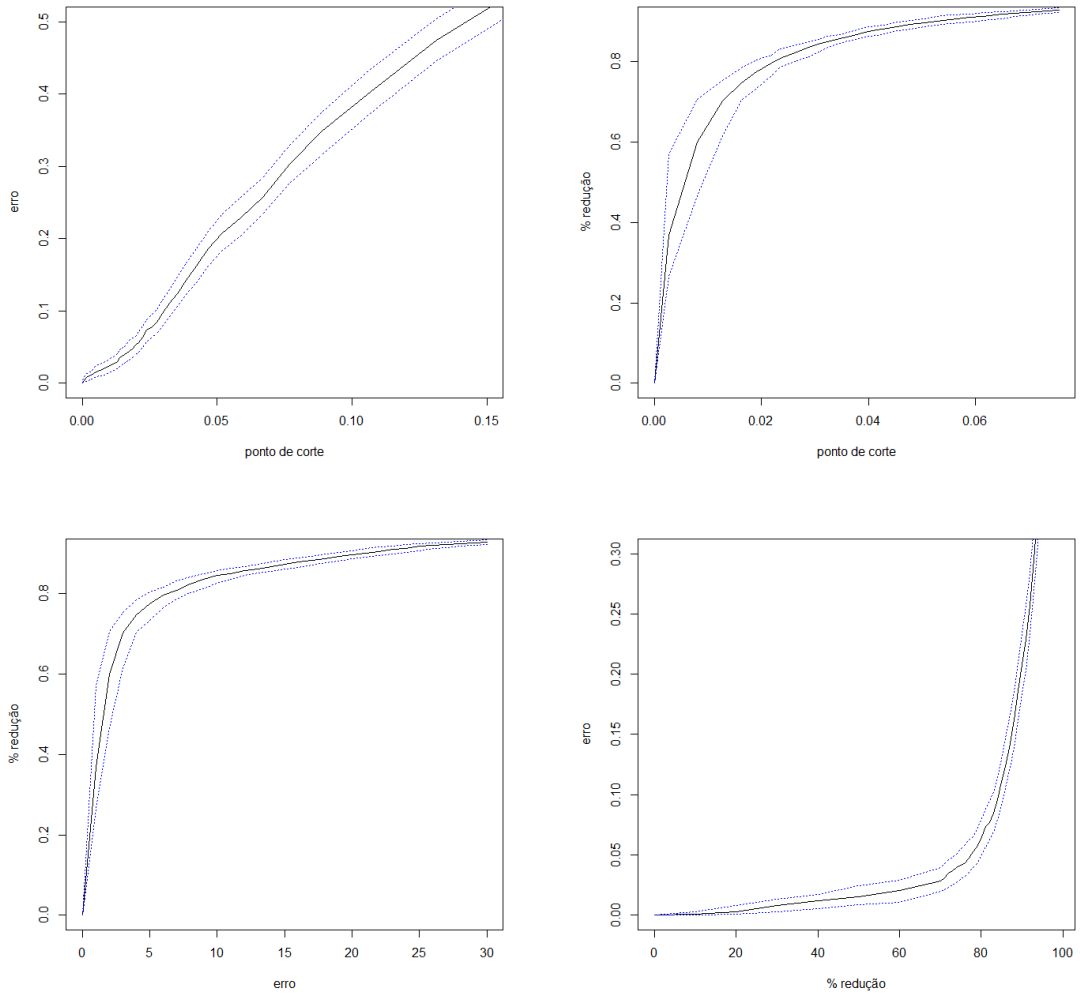


Figura 5.1: Variação dos erros, cortes e percentuais de redução da amostra.

### 5.1.3 Aplicação Prática do Escore de Risco

Na prática, o escore de risco pode ser aplicado de duas formas distintas:

1. Se o interesse é controlar o percentual de erro, o escore pode ser utilizado para definir o ponto de corte ótimo, de forma que o erro de classificação não ultrapasse o valor estabelecido.

   Considere um exemplo, assumindo um erro de classificação de 5%. Veja na Tabela

5.9 que, com 95% de confiança, uma redução de 73% da amostra resultaria em um erro máximo de 4,815% e esse erro foi consequência do ponto de corte 0,014698. Esse resultado sugere que, ao adotar o ponto de corte 0,014698, com 95% de confiança, o erro máximo cometido será menor do que 5%.

Segundo os resultados apresentados na Tabela 5.9, o ponto de corte que resultaria em um erro máximo de 5% pode ser obtido por meio de uma interpolação simples entre os pontos 0,014698 (que apresenta um erro máximo de 4,815%) e o ponto 0,015551 (que apresenta um erro máximo de 5,094%), resultando em um corte no ponto 0,015264. Em outras palavras, ao classificar como baixo risco aqueles associados com escore de risco menor do que 0,015264, com 95% de confiança, o percentual máximo de classificações incorretas será de 5%.

É possível notar pela Tabela 5.8, fixando o erro em 5%, o percentual de redução da amostra será de 77,58% (IC: 73,34% - 80,39%). No exemplo desse trabalho, há um total de 82.531 associados. Com o corte definido acima, espera-se uma redução mínima de 73,34% (limite inferior do IC), que equivale a $82.531 \times 73,34\% = 60.528$ associados classificados como baixo risco. Neste caso, ainda restariam 22.003 associados para serem avaliados pela análise subjetiva, caso a cota de associados que poderão ser alertados nesse mês seja menor do que esse valor.

2. Se o interesse é automatizar completamente o processo, o escore de risco pode ser utilizado simplesmente como um instrumento para ranquear os associados segundo seu risco e alertar aqueles com maiores riscos. O número de alertados será definido segundo a cota mensal de alertas disponível. Esse procedimento tem a vantagem de tirar a dependência da análise subjetiva, mas tem a desvantagem de não controlar o erro de classificação, pois o mesmo depende do percentual de redução da amostra.

   Como exemplo, considere que do total de 82.531 associados, o limite de alertas do mês correspondente seja de 8.253 alertas, isto é, 10% do número total de associados (ou 90% de redução da amostra). Neste caso, serão alertados os 8.253 associados com maior escore de risco, resultando em um ponto de corte do escore de risco igual a 0,05203. Essa classificação automática resultaria numa estimativa de erro de classificação de 20,77% (IC: 18,34% - 23,39%) (Tabela 5.9).

   Considere agora que o limite de alertas do mês correspondente seja de 16.000 alertas, isto é, aproximadamente 20% do número total de associados (ou 80% de redução da amostra). Neste caso, a estimativa do erro de classificação será de 6,31% (IC: 4,90% - 7,87%) (Tabela 5.9).

# Capítulo 6

# Conclusão

Os resultados obtidos sugerem que o modelo de Regressão Logístico Ordinal é adequado para ajustar os dados sobre a classificação do risco dos associados cujas movimentações podem apresentar indícios de lavagem de dinheiro, através das 6 variáveis selecionadas. Como esperado, diante de uma amostra de mais de 80.000 observações, qualquer técnica de seleção de variáveis acusaria significância de qualquer variável considerada no modelo. Assim, a decisão de incluir ou retirar determinada covariável no modelo foi feita com base em técnicas gráficas e considerando a importância subjetiva dessa covariável, ao invés da significância estatística.

Utilizando somente o modelo de regressão logística ordinal, foi possível classificar os associados de acordo com o risco de apresentarem transações suspeitas de lavagem de dinheiro. Como visto, o bom ajuste do modelo foi devido à elevada concordância entre a predição do modelo e as respostas observadas (taxa geral de acerto de 79,71%).

Na aplicação considerada neste trabalho, o erro de classificação mais grave é o de não alertar um associado (classificá-lo como sem indícios de lavagem de dinheiro) que deveria ser comunicado ao COAF, sendo esse o falso negativo. A taxa de falso negativo obtida pelo modelo de regressão logística ordinal (considerando a classificação baseada na prevalência das respostas) foi de 9,22%.

Apesar de ser possível controlar a taxa de falso negativo do modelo de regressão logística ordinal modificando os pontos de corte das classificações, a determinação desses pontos é complicada e pode se tornar confusa por se ter uma resposta tricotômica. Considere, por exemplo, um associado que é classificado como alto risco se $p_3 > 0,1$ (que é diferente de $1/3$ e também da prevalência). Caso $p_3 < 0,1$ (ou $p_1 + p_2 > 0,9$), resta classificar esse associado como baixo ou médio risco. Logo, a definição dessa classificação não segue um regra clara, ao contrário da prevalência. Além disso, o modelo

de regressão logística, por si só, falha em ranquear os associados quanto ao seu risco. Por apresentar três categorias, o uso da probabilidade $p_3$ ou seu respectivo preditor linear não pode ser considerado um escore de risco. Isso porque dois associados com o mesmo valor de $p_3$ terão o mesmo risco, mesmo apresentando valores diferentes de $p_1$ e $p_2$.

O escore de risco apresentado neste trabalho se mostrou útil para ranquear os associados quanto ao seu risco de ter realizado movimentações com indícios de lavagem de dinheiro. Por se tratar de um único escore (e não um vetor de resultados, como na regressão logística ordinal), a definição de um ponto de corte de classificação se torna simples. A consequência direta dessa propriedade é que os percentuais de classificação incorreta (e/ou correta) podem ser facilmente controladas. Na prática, esse escore pode ser utilizado de duas formas distintas: 1) se o interesse é controlar a taxa de falso negativo, esse escore permite reduzir o número de associados submetidos à análise subjetiva. Neste caso, controla-se a taxa de falso negativo, mas não se controla o número de associados classificados com risco de apresentarem indícios de lavagem de dinheiro; 2) se o interesse é automatizar o processo e extinguir o passo da análise subjetiva, o escore de risco pode ser utilizado como um mecanismo para ranquear os associados quanto ao seu risco e, simplesmente, alertar os M associados que apresentaram os maiores escores. Neste caso, não há controle da taxa de falso negativo, que dependerá de M, o número de associados que serão alertados no mês corrente.Dessa forma, o bom desempenho do processo totalmente automatizado depende de um valor alto para o limite de alertas.

# Referências Bibliográficas


[1] ABREU, M.N.S.; SIQUEIRA, A.L.; CAIAFFA, W.T. *Regressão Logística Ordinal em estudos epidemiológicos*. Rev. Saúde Pública. v.43, n.1, p.183-194, 2009.

[2] ACAMS. *Guia de Estudo: Exame de Certificação CAMS*. 5 ed. ACAMS, 2015. ISBN: 978-0-9777495-4-6

[3] AGRESTI, A. *Categorical Data Analysis*. 1 ed. John Wiley & Sons, 1990.

[4] AGRESTI, A. *An Introduction Categorical Data Analysis*. 2 ed. John Wiley & Sons, 2007. ISBN: 978-0-471-22618-5.

[5] AGRESTI, A. *Categorical Data Analysis*. 3 ed. John Wiley & Sons, 2012. ISBN: 978-0-470-46363-5

[6] CELLA, L.O.G. *Regressão Ordinal Bayesiana*. 93 f. Dissertação (Mestrado em Estatística) - Instituto de Ciências Exatas, Universidade de Brasília, Brasília. 2013.

[7] COX, D.R. *The Analysis of Binary Data*. Methuen, London, 1970.

[8] DAVISON, A.C., HINKLEY, D.V. *Bootstrap Methods and their Application*. Cambridge University Press, Cambridge. 1997.

[9] HOSMER, D.W; LEMESHOW, S. *Applied Logistic Regression*. John Wiley & Sons, New York, 2002.

[10] KUTNER, M.H., NACHTSHEIM, C.J., NETER, J. e LI, W. *Applied Linear Statistical Models*. 5 ed. McGraw-Hill/Irwin, 2004. ISBN: 0-256-02547-9.

[11] NAKANO, E.Y. *Soluções bayesianas para alguns problemas clássicos com dados discretos*. 111 f. Tese (Doutorado em Ciências) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo. 2010.

[12] OKURA, R.I.S. *Modelos de Regressão para Variáveis Categóricas Ordinais com aplicações ao problema de classificação*. 110 f. Dissertação (Mestrado em Ciências) - Instituto de Matemática e Estatística, Universidade de São Paulo, São Paulo. 2008.

[13] R Core Team (2017). R: *A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN $3-900051-07-0$, $< http://www.R-project.org >$.

# Apêndice A

# Script do Software R

```
#pacotes utilizados
library("ggplot2")
library(sm)
attach(mtcars)
require(MASS)

#base de dados original
dados.ori<-read.csv2(file.choose())


### Variáveis do estudo ###

#resposta
y<-as.factor(dados.ori$RESPOSTA)


x1[which(x1<1)]<-1
x1<-log(x1)

x2[which(x2<1)]<-1
x2<-log(x2)

n<-length(y)
```

```
#modelo de regressão logística
est<-1
est<-polr(y~x1+x2+x3+x4+x5+x6)
summary(est)
dados<-cbind(y,x1,x2,x3,x4,x5,x6)

###############################################
### predição dos valores da amostra #########
###############################################

probab<-predict(est,x1,type="probs")
predicao<-predict(est,x1,type="class")
escore.risco<-(probab[,2]+2*probab[,3])/2
dados<-cbind(dados,probab,escore.risco,predicao)
dados<-as.data.frame(dados)

table(y,predicao) #### baseado na maior probab


#####################################################
### predição com cortes baseados na prevalência ###
#####################################################

n<-length(dados$y)
cortes<-table(dados$y)/n
pred2<-rep(0,n)
for(i in 1:length(dados$y)){
if (probab[i,3]>=cortes[3]) pred2[i]<-2
else if ((probab[i,2]/(probab[i,1]+probab[i,2])) >=
cortes[2]/(cortes[1]+cortes[2])) pred2[i]<-1
}
table(y,pred2)/n ### baseado na prevalencia

#####################################
#### corte considerando erro de 5% ##
#####################################
```

```
corte<-quantile((dados[which(dados[,1]==3),6]),.05)
corte*100

#### percentual de redução da amostra

mean(dados[,6]<corte)
dados<-dados[which(dados[,6]<corte),]
corte

#### redução x corte
reducao<-numeric()
corte<-numeric()
for (i in 1:30){
corte[i]<-quantile((dados[which(dados[,1]==3),6]),i/100)
reducao[i]<-mean(dados[,6]<corte[i])
}
reducao

plot(seq(1,30)/100,reducao*100,type="l",ylim=c(0,100),
xlab="erro",ylab="% redução")
abline(v=0.05,col=2)

plot(seq(1,30)/100,corte,type="l",ylim=c(0,.5),
xlab="erro",ylab="% redução")

plot(density(dados.ori$SCORE_ANOMALIA[-which(y==2)]),xlim=c(0,.005),col=1)
points(density(dados.ori$SCORE_ANOMALIA[which(y==2)]),col=2,type="l")


###### Cortes ######

dados.cortes<-read.csv2(file.choose())

########## ANÁLISE DESCRITIVA ###########

par(mfrow=c(1,1))
hist(x1, probability = T, main = "X1", breaks = 20, xlab='x1',
```

```
ylab='Frequência',col = 'blue',xlim = c(0,20))
hist(x2, probability = T, main = "X2", breaks = 20, xlab='x2',
ylab='Frequência', col = 'blue',xlim = c(0,20))
hist(x3, probability = T, main = "X3", breaks = 20, xlab='x3',
ylab='Frequência', col = 'blue',xlim = c(0,20))
hist(x4, probability = T, main = "X4", breaks = 20, xlab='x4',
ylab='Frequência', col = 'blue',xlim = c(0,20))


#BOXPLOT
par(mfrow=c(2,2))
colfill<-c(2:(2+length(levels(cyl.f))))
box1 <- boxplot(x1~y, xlab="risco", ylab="log (X1)",
col=colfill, cex.lab=0.8)
box2 <- boxplot(x2~y, xlab="risco", ylab="log (X2)",
col=colfill, cex.lab=0.8)
box3 <- boxplot(x3~y, xlab="risco", ylab="log (X3)",
col=colfill, cex.lab=0.8)
box4 <- boxplot(x4~y, xlab="risco", ylab="log (X4)",
col=colfill, cex.lab=0.8)

x11<-summary(x1[which(y==0)])
x12<-summary(x1[which(y==1)])
x13<-summary(x1[which(y==2)])
x14<-summary(x1)
cbind(x11,x12,x13,x14)

x21<-summary(x2[which(y==0)])
x22<-summary(x2[which(y==1)])
x23<-summary(x2[which(y==2)])
x24<-summary(x2)
cbind(x21,x22,x23,x24)

x31<-summary(x3[which(y==0)])
x32<-summary(x3[which(y==1)])
x33<-summary(x3[which(y==2)])
x34<-summary(x3)
```

```
cbind(x31,x32,x33,x34)

x41<-summary(x4[which(y==0)])
x42<-summary(x4[which(y==1)])
x43<-summary(x4[which(y==2)])
x44<-summary(x4)
cbind(x41,x42,x43,x44)

## exponenciais
x11<-summary(exp(x1)[which(y==0)])
x12<-summary(exp(x1)[which(y==1)])
x13<-summary(exp(x1)[which(y==2)])
x14<-summary(exp(x1))
cbind(x11,x12,x13,x14)

x21<-summary(exp(x2)[which(y==0)])
x22<-summary(exp(x2)[which(y==1)])
x23<-summary(exp(x2)[which(y==2)])
x24<-summary(exp(x2))
cbind(x21,x22,x23,x24)

x31<-summary(exp(x3)[which(y==0)])
x32<-summary(exp(x3)[which(y==1)])
x33<-summary(exp(x3)[which(y==2)])
x34<-summary(exp(x3))
cbind(x31,x32,x33,x34)

x41<-summary(exp(x4)[which(y==0)])
x42<-summary(exp(x4)[which(y==1)])
x43<-summary(exp(x4)[which(y==2)])
x44<-summary(exp(x4))
cbind(x41,x42,x43,x44)


## FAZENDO O EXEMPLO

x1_exemplo <- x1[65957]
```

```
x2_exemplo <- x2[65957]
x3_exemplo <- x3[65957]
x4_exemplo <- x4[65957]
x5_exemplo <- x5[65957]
x6_exemplo <- x6[65957]

exp_x1_exemplo <- exp(x1_exemplo)
exp_x2_exemplo <- exp(x2_exemplo)
exp_x3_exemplo <- exp(x3_exemplo)
exp_x4_exemplo <- exp(x4_exemplo)

exemplo_var_ori <- cbind(exp_x1_exemplo,exp_x2_exemplo,exp_x3_exemplo,
exp_x4_exemplo,x5_exemplo,x6_exemplo)
exemplo_var_log <- cbind(x1_exemplo,x2_exemplo,x3_exemplo,
x4_exemplo,x5_exemplo,x6_exemplo)

logit1 <- est$zeta[1] - sum(exemplo_var_log*est$coefficients)
logit2 <- est$zeta[2] - sum(exemplo_var_log*est$coefficients)

probab1 <- exp(logit1)/(1+exp(logit1))
probab2 <- exp(logit2)/(1+exp(logit2)) - probab1
probab3 <- 1 - probab2 - probab1

escore_exemplo <- (probab2 + 2*probab3)/2
```