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

Uma análise de regressão logística usando componentes principais

Gustavo Ramos de Goes

Trabalho de Conclusão de Curso

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

Uma análise de regressão logística usando componentes principais

Gustavo Ramos de Goes
Orientadora: Teresa Cristina Martins Dias

Trabalho de Conclusão de Curso apresentado como parte dos requisitos para obtenção do título de Bacharel em Estatística.

São Carlos
Junho de 2024

Federal University of São Carlos
Exact and Technology Sciences Center
Department of Statistics

A logistic regression analysis using principal components

Gustavo Ramos de Goes
Advisor: Teresa Cristina Martins Dias

Bachelors dissertation submitted to the Department of Statistics, Federal University of São Carlos - DEs-UFSCar, in partial fulfillment of the requirements for the degree of Bachelor in Statistics.

São Carlos
June 20th, 2024

Gustavo de Ramos Goes

Uma análise de regressão logística usando componentes principais

Este exemplar corresponde à redação final do trabalho de conclusão de curso devidamente corrigido e defendido por Gustavo Ramos de Goes e aprovado pela banca examinadora.

Aprovado em 25 de janeiro de 2024

Banca Examinadora:

- Profa. Dra. Teresa Cristina Martins Dias (Orientadora)
- Prof. Dr. Márcio Luis Lanfredi Viola
- Profa. Dra. Andressa Cerqueira

Resumo

Neste Trabalho de Conclusão de Curso (TCC) estudamos, no contexto modelos de regressão logística, o caso em que o número de covariáveis é grande e as covariáveis são correlacionadas. Nesta situação, usamos a técnica análise de componentes principais a fim de reduzir o número de covariáveis (dimensão) envolvidas no conjunto de dados. Apresentamos dois exemplos de regressão logística e um de análise de componentes principais, além de uma comparação entre uma aplicação com as duas metodologias, aplicadas conjuntamente, e uma aplicação usando, apenas, a regressão logística.

Palavras-chave: *análise de componentes principais, redução de dimensionalidade, regressão logística.*

Abstract

In this Bachelors dissertation we studied, in the context of logistic regression models, the case in which the number of covariables is large and the covariables are correlated. In this situation, we use the principal component analysis technique in order to reduce the number of the variables (dimension) involved in the data set. We present two examples of logistic regression and one of principal component analysis, as well as a comparison between an application with both methodologies, applied jointly, and an application using only logistic regression.

Keywords: *dimensionality reduction, logist regression, principal component analysis.*

Lista de Figuras

Lista de Tabelas

Sumário

Capítulo 1

Introdução

Para a análise de dados envolvendo variáveis binárias ou dicotômicas, a regressão logística é um modelo muito popular para a obtenção da estimativa das probabilidades de interesse. Apesar de não ser o mais indicado é comum que esta seja dicotomizada a fim de utilizar a regressão logística para o calculo da probabilidade de sucesso. Existem outros modelos que possibilitam a modelagem de dados binários, como por exemplo, árvore de classificação, mas a regressão logística se destaca é frequentemente preferida devido às suas propriedades únicas e interpretabilidade direta, de acordo com Hosmer Jr *et al.* (2013). A regressão logística utiliza a função *logit* para modelar a relação entre as variáveis independentes e a probabilidade de um evento ocorrer. Essa função *logit* transforma as probabilidades lineares em uma escala que varia de menos infinito a mais infinito, sendo especialmente útil para casos em que a relação entre as variáveis preditoras e a resposta não é linear. Além disso, a regressão logística oferece medidas como razão de chances, que são úteis na interpretação dos efeitos das variáveis independentes sobre a probabilidade do evento de interesse.

Mesmo com as qualidades citadas a respeito da regressão logística, podemos encontrar dificuldades na interpretação e no entendimento do modelo quando trabalhamos com muitas variáveis, por exemplo, um modelo de regressão logística no qual há um número alto de coeficientes traria dificuldades na visualização e interpretação do mesmo. Essa dificuldade causada pela alta dimensionalidade não pode ser resolvido retirando variáveis do conjunto de dados, uma vez que essa atitude causaria perda de informação.

Muitas vezes trabalhar com um banco de dados com alta dimensionalidade pode ser um problema. A partir disso, desenvolveram-se técnicas para lidar com problemas de alta dimensionalidade, entre elas, a Análise de Componentes Principais (ACP). Essa técnica

foi inicialmente escrita por Pearson (1901), posteriormente por Hotteling (1933) apresentou métodos computacionais com o objetivo de analisar as estruturas de correlação das variáveis envolvidas no estudo. Em linhas gerais, a ACP tem como objetivo, além da redução da dimensão dos dados, a identificação de padrões e estrutura, compactação de informação e facilidade na visualização dos dados.

1.1 Objetivo

Com o intuito de superar as limitações decorrentes da alta dimensionalidade e da complexidade na interpretação de modelos de regressão logística, surge a necessidade de explorar abordagens inovadoras. A união entre a regressão logística e a Análise de Componentes Principais oferece uma solução promissora. A ACP, ao reduzir a dimensão dos dados, poderia ser empregada como um meio para pré-processar as variáveis originais, mitigando os desafios enfrentados pela interpretação e visualização em modelos com um grande número de coeficientes. Ao empregar ambas as ferramentas conjuntamente, é possível alcançar uma representação mais compacta e informativa dos dados, preservando ao mesmo tempo as informações relevantes, o que pode resultar em modelos com interpretações mais elucidativas. Neste contexto, o objetivo neste trabalho é investigar a relação entre a regressão logística e a análise de componentes principais como uma abordagem integrada para o tratamento de conjuntos de dados com variáveis correlacionadas e de alta dimensionalidade. Dessa forma, o começamos o trabalho desenvolvendo a base teórica tanto da regressão logística para variável resposta binária, com a função de verossimilhança e a interpretação dos coeficientes, como também da ACP, com a construção, escolha e interpretação dos componentes principais. Após estruturar a teoria foram desenvolvidos três exemplos, sendo dois de regressão logística e um ACP. Na segunda parte do trabalho, fizemos uma aplicação utilizando componentes principais como covariáveis de um modelo de regressão logística e outra sem a utilização dos componentes afim de analisar o desempenho na classificação e a interpretabilidade dos coeficientes estimados.

Capítulo 2

Modelos de regressão logística e componentes principais

Neste capítulo abordamos a metodologia aplicada ao longo deste trabalho, especificamente, modelos de regressão logística e análise de componentes principais. Na Seção 2.1 apresentamos a construção do modelo de rgressão logística, estimação e interpretação dos coeficientes associados às covariáveis e na Seção 2.2 apresentamos a análise de componentes principais, isto é, obtenção das componentes e interpretação.

2.1 Modelos de regressão logística

Um modelo de regressão logística é uma técnica usada com o intuito de analisar e modelar a relação entre uma variável binária (ou dicotômica), que possui apenas dois resultados possíveis (sucesso ou fracasso), e uma ou mais variáveis independentes. Nesse modelo, calcula-se os coeficientes que indicam como as variáveis independentes influenciam a *log-odds* da variável dependente, permitindo assim estimar a probabilidade de sucesso em relação aos valores das variáveis independentes. O resultado do modelo é uma curva logística que descreve a relação entre as variáveis independentes e a probabilidade do evento de interesse. Essa técnica é bastantete utilizada em estudos epidemiológicos, previsões de sucesso/falha e em problemas onde a variável dependente é categórica e envolve dois resultados distintos.

Para ilustrar a aplicação prática da regressão logística, focamos em seu uso na construção de modelos de crédito. Construir modelos de crédito envolve avaliar a probabilidade de um mutuário inadimplir com seus pagamentos com base em vários fatores, como

renda, histórico de crédito e informações demográficas. A regressão logística é uma ferramenta valiosa para estimar a probabilidade de inadimplência e classificar os mutuários em diferentes categorias de risco.

Nas próximas seções exploramos os fundamentos teóricos da regressão logística baseados no livro Myers *et al.* (2012), o processo de modelagem e interpretação dos resultados. Além disso, discutimos brevemente a seleção de variáveis e avaliação do desempenho do modelo. Começamos examinando os fundamentos matemáticos da regressão logística. Em seguida, nos aprofundamos no conceito da função logística, que é o núcleo dessa técnica estatística. Por fim, discutimos métodos de estimativa de parâmetros, medidas de ajuste do modelo e a interpretação do modelo resultante.

2.1.1 Modelos de regressão logística para variável resposta binária

Considere a situação em que a variável resposta em um problema de regressão assume apenas dois valores possíveis, 0 ou 1. Estas podem ser atribuições arbitrárias resultantes da observação de uma resposta qualitativa. Por exemplo, a resposta pode ser o resultado de uma classificação de clientes em um banco de dados como adimplentes (sucesso) ou inadimplentes (falha).

Suponha que o modelo de regressão linear múltiplo possa ser ajustado, isto é,

$$Y_i = \mathbf{x}_i^{'}\boldsymbol{\beta} + \varepsilon_i,\ i = 1, 2, \ldots, n. \tag{2.1}$$

em que:

- $\mathbf{Y}$ é o vetor coluna dos valores da variável dependente para todas as observações de dimensão $n \times 1$
- $\mathbf{x_i}$ é a matriz das variáveis independentes, em que cada linha contém os valores das variáveis para uma observação com dimensão $n \times (p + 1)$
- $\boldsymbol{\beta}$ é o vetor coluna formado pelos coeficientes de regressão associada cada variável independente e do intercepto com dimensão $(p + 1) \times 1$
- $\boldsymbol{\varepsilon_i}$ é o vetor coluna dos termos de erro para cada observação com dimensão $n \times 1$.

Assumimos que a variável resposta é uma váriavel aleatória Bernoulli com distribuição

de probabilidade dada por:

$$Y_i = \begin{cases} 1, & \text{com probabilidade } \pi_i \\ 0, & \text{com probabilidade } 1 - \pi_i. \end{cases}$$

Agora, desde que $E(\varepsilon_i) = 0$, o valor esperado da variável resposta é

$$E(Y_i) = 1(\pi_i) + 0(1 - \pi_i) = \pi_i.$$

Usando $E(\varepsilon_i) = 0$,

$$E(Y_i) = \mathbf{x}_i'\boldsymbol{\beta} = \pi_i \tag{2.2}$$

Isso significa que a resposta esperada dada pelo preditor linear é a probabilidade de que a variável de resposta assuma o valor 1. Existem alguns problemas com o modelo de regressão formado pela Equação (2.1). Primeiramente, observe que, se a resposta for binária, os termos de erro ε_i podem assumir apenas dois valores, a saber,

$$\begin{aligned} \varepsilon_i &= 1 - \mathbf{x}_i'\boldsymbol{\beta} \quad \text{quando } y_i = 1, \\ \varepsilon_i &= -\mathbf{x}_i'\boldsymbol{\beta} \quad \text{quando } y_i = 0. \end{aligned}$$

Consequentemente, os erros neste modelo não podem ser normais. Em segundo lugar, a variância do erro não é constante, pois

$$\sigma_{y_i}^2 = E\{Y_i - E(Y_i)\}^2 = (1 - \pi_i)^2 \pi_i + (0 - \pi_i)^2 (1 - \pi_i) = \pi_i (1 - \pi_i)$$

Note que a última expressão nada mais é que:

$$\sigma_{Y_i}^2 = E(Y_i)[1 - E(Y_i)]$$

pela Equação 2.2. Isso indica que a variância das observações (que é a mesma que a variância dos erros) é uma função da média. Finalmente, há uma restrição na função de resposta, porque

$$0 \leq E(Y_i) = \pi_i \leq 1, \ i = 1, \ldots n.$$

Essa restrição pode causar sérios problemas com a escolha de uma função de resposta linear, como inicialmente assumimos na Equação (2.1). Seria possível ajustar um modelo aos dados de forma que os valores previstos da resposta estão fora do intervalo $(0, 1)$.

Geralmente, quando a variável de resposta é binária, há evidências empíricas consideráveis indicando que a forma da função de resposta deve ser não linear. Uma função monotonicamente crescente (ou decrescente) em forma de S (ou em forma de S inversa) geralmente é empregada. Essa função é chamada de função de resposta logística e tem a forma

$$E(Y_i) = \frac{\exp\left(\mathbf{x_i}'\boldsymbol{\beta}\right)}{1+\exp\left(\mathbf{x_i}'\boldsymbol{\beta}\right)} = \frac{1}{1+\exp\left(-\mathbf{x_i}'\boldsymbol{\beta}\right)}.$$

A função de resposta logística pode ser facilmente linearizada. Essa abordagem define a parte estrutural do modelo em termos de uma função da média da função de resposta. Seja Y uma variável aleatória tal que $Y \sim Bernoulli(\pi)$ e a probabilidade de sucesso, $\pi(x)$, dada por:

$$\eta = \mathbf{x}'\boldsymbol{\beta} \tag{2.3}$$

o preditor linear em que η é definido pela transformação

$$\eta = \ln(\frac{\pi}{1-\pi}) \tag{2.4}$$

Esta transformação é chamada de transformação logito que é a função de ligação que relaciona π com η. A transformação logit é uma abordagem muito popular para modelar dados de Bernoulli ou binomial. A transformação mapeia π, que é limitado entre 0 e 1, para a reta numérica real. A razão

$$\frac{\pi}{1-\pi} \tag{2.5}$$

é a de razão das probabilidades de ocorrer sucesso e fracasso, chamada de razão de odds. Aplicando a transformação logito na razão (2.5) temos que

$$g(x) = \ln\left(\frac{\pi(x)}{1-\pi(x)}\right),$$

com

$$g(x) = \beta_0 + \beta_1 x_1 + ... + \beta_p x_p.$$

Então,

$$\ln\left(\frac{\pi(x)}{1-\pi(x)}\right) = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p$$

$$\frac{\pi}{1-\pi} = \exp\left\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\right\}$$

$$\pi(x) = \exp\left\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\right\}\left(1-\pi(x)\right)$$

$$\pi(x) = \exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\} - \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\pi(x)$$

$$\pi(x) + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}\pi(x) = \exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\}$$

$$\pi(x)(1 + \exp\{\beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p\}) = \exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\}$$

$$\pi(x) = \frac{\exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\}}{1 + \exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_p x_p\}}. \tag{2.6}$$

O modelo em (2.6) é chamado de modelo de regressão logística. O interesse é estimar o vetor parâmetros $\boldsymbol{\beta}$ e consequentemente o vetor de probabilidades $\boldsymbol{\pi}$. Para isso, o método da máxima verossimilhança é usado.

2.1.2 Função de verossimilhança

A função de verossimilhança é uma maneira usada para obter os estimadores dos parâmetros envolvidos no modelo. A função de verossimilhança é escrita em função dos parâmetros (quantidades a serem estimadas), condicionada aos valores registrados (variáveis resposta e covariáveis) (Mood *et al.*, 1974).

No ajuste de um modelo de regressão logística assumimos que a variável resposta segue uma distribuição Bernoulli, com probabilidade de sucesso π. Sendo assim, considerando uma amostra aleatória com n observações temos que :

$$Y_i \sim \text{Bernoulli}(\pi_i), i = 1, 2, \ldots, n,$$

em que

$$E(Y_i) = \pi_i, i = 1, 2, \ldots, n. \tag{2.7}$$

Assim, as respostas individuais (ou seja, o y_i) são modeladas quando assumimos que sua distribuição é Bernoulli. Uma vez que cada observação segue uma distribuição de Bernoulli, a distribuição de probabilidade para a observação i-ésima é dada por

$$f(y_i) = \pi_i^{y_i}(1 - \pi_i)^{1-y_i}, i = 1, 2, \ldots, n$$

e a função de verossimilhança é

$$L(\boldsymbol{\pi}|\boldsymbol{y}) = \prod_{i=1}^{n}\left[\pi_i^{y_i}(1 - \pi_i)^{1-y_i}\right].$$

Aplicando a função logarítmica e usando a expressão (2.6), obtemos

$$\begin{aligned}
l(\boldsymbol{\beta}|\boldsymbol{y},\boldsymbol{x}) &= \sum_{i=1}^{n}\left(y_i \ln\left(\pi_i\right) + (1-y_i)\ln\left(1-\pi_i\right)\right) \\
&= \sum_{i=1}^{n}\left(y_i \ln\left(\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right) + (1-y_i)\ln\left(1-\frac{e^{x_i'\beta}}{1+e^{x_i'\beta}}\right)\right) \\
&= \sum_{i=1}^{n}\left(y_i\left(x_i'\beta - \ln\left(1+e^{x_i'\beta}\right)\right) + (1-y_i)\ln\left[\frac{\left(1+e^{x_i'\beta}\right)}{1+e^{x_i'\beta}}\right]\right) \\
&= \sum_{i=1}^{n}\left(y_i x_i'\beta - y_i \ln\left(1+e^{x_i'\beta}\right) + \ln\left[\frac{\left(1+e^{x_i'\beta}\right)-e^{x_i'\beta}}{1+e^{x_i'\beta}}\right]\right. \\
&\quad \left. -y_i \ln\left[\frac{\left(1+e^{x_i'\beta}\right)-e^{x_i'\beta}}{1+e^{x_i'\beta}}\right]\right) \\
&= \sum_{i=1}^{n}\left(y_i x_i'\beta - y_i \ln\left(1+e^{x_i'\beta}\right) - \ln\left(1+e^{x_i'\beta}\right) + y_i \ln\left(1+e^{x_i'\beta}\right)\right) \\
&= \sum_{i=1}^{n}\left(y_i x_i'\beta - \ln\left(1+e^{x_i'\beta}\right)\right). \qquad (2.8)
\end{aligned}$$

O estimador de máxima verossimilhança $\hat{\boldsymbol{\beta}}$ é encontrado derivando a expressão (2.8), para cada um dos parâmetros em $\boldsymbol{\beta}$, e igualando a zero, ou seja,

$$\frac{\partial l(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n}\left(y_i x_i' - \frac{e^{x_i'\boldsymbol{\beta}}}{(1+e^{x_i'\boldsymbol{\beta}})}x_i'\right) = 0, \text{ para } j = 0, 1, \ldots, p.$$

2.1.3 Interpretação dos coeficientes e probabilidade

A interpretação de coeficientes e probabilidades em modelos de regressão é crucial para entender como as variáveis independentes afetam a variável dependente. Os coeficientes indicam a magnitude e direção da influência de cada preditor, enquanto as probabilidades fornecem as chances de ocorrência de um evento.

Dentro da regressão logística, os coeficientes estimados são extremamente importantes na interpretação dos resultados, pois eles nos dão informações sobre sua direção (maior ou menor probabilidade de estar associado a categoria de interesse) por meio do seu sinal e sobre sua magnitude (o quanto a coeficiente influência na variável resposta).

Destacando a relação entre coeficientes e probabilidades, no nosso modelo de interesse (regressão logística) os coeficientes estimados são expressos em termos de *log-odds* que nada mais é do que a função logarítima de uma *odd* (chance) acontecer. Porém, para fins de facilitar a interpretação, transformamos os log-*odd* em probabilidades e chances:

- Probabilidade: é usada a função exponencial de forma a transformar o *log-odds* em probabilidades. Por exemplo: seja nossa categoria de interesse um cliente ser classificado como adimplente, essa probabilidade é dada por: $P(Y = 1|\mathbf{X}) = \frac{\exp(logit)}{1+\exp(logit)}$. Ou seja, esse valor está contido no intervalo [0,1].

- Coeficientes: Na regressão logística, os coeficientes estimados indicam como a mudança em uma unidade em uma variável independente afeta o logaritmo das *odds* do evento de interesse. Se o coeficiente for positivo, espera-se que o logaritmo das *odds* (e, portanto, as *odds*) aumente com o aumento da variável independente. Se for negativo, espera-se uma diminuição nas *odds*. A interpretação específica depende da escala e natureza das variáveis envolvidas.

2.2 Análise de componentes principais

Conforme Johnson e Wichern (2002), a análise de componentes principais (ACP) é um poderoso método estatístico destinado a compreender a estrutura subjacente de um conjunto de variáveis. Seu foco principal é revelar as relações e padrões nos dados, transformando as variáveis originais em um novo conjunto de componentes ortogonais. Os objetivos fundamentais da ACP são duplos: simplificação dos dados e maior interpretabilidade.

Normalmente, um conjunto de dados consiste em um número grande variáveis, e para capturar completamente a variabilidade total, seria necessário utilizar todas as suas variáveis. No entanto, a ACP pode reduzir efetivamente essa complexidade identificando um subconjunto menor de componentes-chave, frequentemente, denominados "componentes principais". Notavelmente, esses componentes principais podem reter uma quantidade significativa das informações originais, permitindo que eles substituam as variáveis iniciais enquanto mantém características essenciais. Como resultado, um grande conjunto de dados com medições em várias variáveis pode ser representado de forma eficiente por um conjunto reduzido de medições em alguns componentes principais.

Além de suas capacidades de redução de dados, a ACP tem a qualidade de revelar relações previamente não percebidos nos dados. Esse recurso permite que os pesquisadores façam interpretações novas que talvez não fossem evidentes por meio de análises convencionais.

A ACP serve como uma ferramenta valiosa em um contexto analítico mais amplo,

frequentemente atuando como uma etapa intermediária em investigações mais extensas. Por exemplo, os componentes principais derivados podem ser usados como entradas para análises de regressão ou análises de agrupamento.

Por fim, vale a pena ressaltar que em regressão, tratamos $\boldsymbol{X}$ como covariável, enquanto que na ACP o tratamos como variável.

2.2.1 Componentes Principais de uma população

Ao usar a técnica ACP o objetivo está em encontrar combinações lineares particulares das p variáveis aleatórias $X_1, X_2, \ldots, X_p$. Geometricamente, essas combinações lineares representam a seleção de um novo sistema de coordenadas, obtido por meio da rotação do sistema original com $X_1, X_2, \ldots, X_p$ como os eixos de coordenadas. Os novos eixos representam as direções de maior variabilidade, permitindo uma descrição mais simples e concisa da estrutura de covariância.

Os componentes principais dependem exclusivamente da matriz de covariância $\boldsymbol{\Sigma}$ (ou da matriz de correlação $\boldsymbol{\rho}$) das p variáveis $X_1, X_2, \ldots, X_p$, o uso dessa técnica não requer a suposição de distribuição normal multivariada dos dados. Entretanto, os componentes principais derivados de populações multivariadas normais possuem interpretações úteis em relação aos elipsoides de densidade constante. Além disso, quando a população é multivariada normal, é possível realizar inferências a partir dos componentes amostrais.

Seja o vetor aleatório $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ com matriz de covariância $\boldsymbol{\Sigma}$ e autovalores $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$, sendo $\mathbf{A}$ a matriz dos coeficientes da ACP resultante da decomposição espectral (de acordo com Johnson e Wichern (2002)) da matriz $\boldsymbol{\Sigma}$, dada por:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1p} \\ a_{21} & a_{22} & \cdots & a_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ a_{p1} & a_{p2} & \cdots & a_{pp} \end{bmatrix}.$$

Considere as combinações lineares:

$$\begin{aligned}
Y_1 &= \mathbf{a}_1^{'}\mathbf{X} = a_{11}X_1 + a_{12}X_2 + \cdots + a_{1p}X_p \\
Y_2 &= \mathbf{a}_2^{'}\mathbf{X} = a_{21}X_1 + a_{22}X_2 + \cdots + a_{2p}X_p \\
&\vdots \\
Y_p &= \mathbf{a}_p^{'}\mathbf{X} = a_{p1}X_1 + a_{p2}X_2 + \cdots + a_{pp}X_p.
\end{aligned}$$

Obtemos:

$$\begin{aligned}
\text{Var}\,(Y_i) &= \mathbf{a}_i^{'}\mathbf{\Sigma}\mathbf{a}_i \qquad i &= 1, 2, \ldots, pe \\
\text{Cov}\,(Y_i, Y_k) &= \mathbf{a}_i^{'}\mathbf{\Sigma}\mathbf{a}_k \quad i, k &= 1, 2, \ldots, p.
\end{aligned}$$

O primeiro componente principal é a combinação linear com a maior variância. É possível aumentar $\text{Var}\,(Y_1)$ multiplicando $\mathbf{a}_1$ por uma constante. Para eliminar essa indeterminação, é conveniente restringir $\mathbf{a}_1$ a ter comprimento unitário:

Primeiro componente principal = combinação linear $\mathbf{a}_1^{'}\mathbf{X}$ que maximiza

$$\text{Var}\left(\mathbf{a}_1^{'}\mathbf{X}\right) \text{ sujeito a } \mathbf{a}_1^{'}\mathbf{a}_1 = 1.$$

Segundo componente principal = Combinação linear $\mathbf{a}_2^{'}\mathbf{X}$ que maximiza

$$\begin{aligned}
&\text{Var}\left(\mathbf{a}_2^{'}\mathbf{X}\right) \text{ sujeito a } \mathbf{a}_2^{'}\mathbf{a}_2 = 1 \text{ e} \\
&\text{Cov}\left(\mathbf{a}_1^{'}\mathbf{X}, \mathbf{a}_2^{'}\mathbf{X}\right) = 0.
\end{aligned}$$

***i*-ésimo componente principal** = Combinação linear $\mathbf{a}_i^{'}\mathbf{X}$ que maximiza

$$\begin{aligned}
&\text{Var}\left(\mathbf{a}_i^{'}\mathbf{X}\right) \text{ sujeito a } \mathbf{a}_j^{'}\mathbf{a}_i = 1 \text{ e} \\
&\text{Cov}\left(\mathbf{a}_i^{'}\mathbf{X}, \mathbf{a}_k^{'}\mathbf{X}\right) = 0 \text{ para } k < i.
\end{aligned}$$

Seja $\mathbf{\Sigma}$ a matriz com pares autovalor-autovetor $(\lambda_1, \mathbf{e}_1)$, $(\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$, em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Então, o i-ésimo componente principal é determinado por:

$$Y_i = e_i^{'}\mathbf{X} = e_{i1}X_1 + e_{i2}X_2 + \cdots + e_{ip}X_p, \quad i = 1, 2, \ldots, p.$$

Assim, temos:

$$\begin{aligned} \text{Var}\,(Y_i) &= \mathbf{e}_i'\Sigma\mathbf{e}_i = \lambda_i \qquad i = 1, 2, \ldots, p, \\ \text{Cov}\,(Y_i, Y_k) &= \mathbf{e}_i'\Sigma\mathbf{e}_k = 0 \qquad i \neq k, \end{aligned}$$

em que

$$\mathbf{\Sigma} = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \lambda_p \end{bmatrix}$$

Se alguns λ_i forem iguais, as escolhas dos vetores de coeficientes correspondentes, $\mathbf{e}_i$, e, portanto, Y_i, não são únicas. Os componentes principais são não correlacionados e suas variâncias são iguais aos autovalores de $\mathbf{\Sigma}$.

Suponha que $\mathbf{X}' = [X_1, X_2, \ldots, X_p]$ tenha a matriz de covariância $\mathbf{\Sigma}$, com pares autovalor-autovetor $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda_p, \mathbf{e}_p)$, em que $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$. Sejam $Y_1 = \mathbf{e}_1'\mathbf{X}, Y_2 = \mathbf{e}_2'\mathbf{X}, \ldots, Y_p = \mathbf{e}_p'\mathbf{X}$ os componentes principais. Então,

$$\sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} = \sum_{i=1}^{p} \text{Var}\,(X_i) = \lambda_1 + \lambda_2 + \cdots + \lambda_p = \sum_{i=1}^{p} \text{Var}\,(Y_i)\,.$$

Pelo resultado acima temos que:

$$\begin{aligned} \text{Variância total da população} &= \sigma_{11} + \sigma_{22} + \cdots + \sigma_{pp} \\ &= \lambda_1 + \lambda_2 + \cdots + \lambda_p, \end{aligned}$$

e consequentemente, a proporção da variância total devida (explicada) pelo k-ésimo componente principal é o k-ésimo termo dividido pela da variância total da população:

$$\left(\begin{array}{c} \text{Proporção da variância total} \\ \text{da população devida} \\ \text{ao } k\text{ -ésimo} \\ \text{componente principal.} \end{array} \right) = \frac{\lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p} \quad , k = 1, 2, \ldots, p.$$

Se a maior parte da variabilidade total da população, por exemplo, entre 80% a 90%, pode ser atribuída aos primeiros um, dois ou três componentes principais, então esses componentes podem substituir as variáveis originais sem causar grande perda de informação.

Cada componente do vetor de coeficientes $\mathbf{e}_i' = [e_{i1}, \ldots, e_{ik}, \ldots, e_{ip}]$ também é impor-

tante e merece ser examinado. A magnitude de e_{ik} indica a relevância da k-ésima variável para o i-ésimo componente principal, independentemente das outras variáveis. Em outras palavras, e_{ik} está diretamente relacionado ao coeficiente de correlação entre Y_i e X_k.

Se obtivermos os componentes principais $Y_1 = \mathbf{e}_1^{'}\mathbf{X}, Y_2 = \mathbf{e}_2^{'}\mathbf{X}, \ldots, Y_p = \mathbf{e}_p^{'}\mathbf{X}$ a partir da matriz de covariância $\mathbf{\Sigma}$, então os coeficientes de correlação entre esses componentes Y_i e as variáveis originais X_k são expressos por:

$$\rho_{Y_i,X_k} = \frac{e_{ik}\sqrt{\lambda_i}}{\sqrt{\sigma_{kk}}} \quad i,k = 1,2,\ldots,p$$

Neste contexto, $(\lambda_1, \mathbf{e}_1), (\lambda_2, \mathbf{e}_2), \ldots, (\lambda p, \mathbf{e}_p)$ representam os pares de autovalor-autovetor associados à matriz $\mathbf{\Sigma}$.

Embora as correlações das variáveis com os componentes principais possam ser úteis para interpretar os componentes, elas refletem apenas a contribuição individual de cada variável X em cada componente Y, isoladamente. Isso significa que as correlações não fornecem uma medida da importância de uma variável X para um componente Y considerando as outras variáveis X presentes. Por essa razão, alguns estatísticos recomendam que apenas os coeficientes e_{ik} sejam utilizados para interpretar os componentes. Os coeficientes e as correlações podem resultar em classificações diferentes da importância das variáveis para um componente específico, mas, na prática, essas classificações geralmente são bastante semelhantes. Normalmente, variáveis com coeficientes relativamente grandes (em valor absoluto) tendem a ter correlações relativamente altas, fazendo com que as duas medidas de importância, a abordagem multivariada e a univariada, levem a resultados parecidos. Recomenda-se examinar tanto os coeficientes quanto as correlações para obter uma interpretação mais completa dos componentes principais.

2.2.2 Resumindo a Variação da Amostra por componentes Principais

Agora temos o quadro necessário para estudar o problema de resumir a variação em n medições em p variáveis com algumas combinações lineares criteriosamente escolhidas.

Suponha que os dados $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n$ representam n observações independentes de alguma população p-dimensional com vetor médio $\boldsymbol{\mu}$ e matriz de covariância $\mathbf{\Sigma}$. Por esses dados, calcula-se o vetor médio da amostra $\overline{\mathbf{x}}$, a matriz de covariância da amostra $\mathbf{S}$ e a matriz de correlação da amostra $\mathbf{R}$.

Nosso objetivo nesta seção é construir combinações lineares não correlacionadas das características medidas que representam grande parte da variação na amostra. As combinações não correlacionadas com as maiores variâncias são chamadas de componentes principais da amostra.

Lembrando que os n valores de qualquer combinação linear

$$\mathbf{a}_1^{'}\mathbf{x} = a_{11}x_{j1} + a_{12}x_{j2} + \cdots + a_{1p}x_{jp}, \quad j = 1, 2, \ldots, n,$$

os componentes principais da amostra têm média amostral $\mathbf{a}_1^{'}\overline{\mathbf{x}}$ e variância amostral $\mathbf{a}_1^{'}\mathbf{S}\mathbf{a}_1$. Além disso, as combinações lineares $\left(\mathbf{a}_1^{'}\mathbf{x}_j, \mathbf{a}_2^{'}\mathbf{x}_j\right)$, de duas combinações lineares distintas, têm covariância amostral $\mathbf{a}_1^{'}\mathbf{S}\mathbf{a}_2$. Os componentes principais da amostra são obtidos como as combinações lineares que possuem a maior variância amostral. Para garantir coerência com as quantidades populacionais, normalmente normalizamos os vetores de coeficientes $\mathbf{a}_i$ de forma que $\mathbf{a}_i^{'}\mathbf{a}_i = 1$.

Mais especificamente, para obter o primeiro componente principal da amostra, procuramos a combinação linear $\mathbf{a}_1^{'}\mathbf{x}_j$ que maximize a variância amostral sujeita à restrição $\mathbf{a}_1^{'}\mathbf{a}_1 = 1$.

Da mesma forma, para o segundo componente principal da amostra, buscamos a combinação linear $\mathbf{a}_2^{'}\mathbf{x}_j$ que maximize a variância amostral, mas agora adicionamos a restrição adicional de que a covariância amostral entre as combinações $\left(\mathbf{a}_1^{'}\mathbf{x}_j, \mathbf{a}_2^{'}\mathbf{x}_j\right)$ deve ser igual a zero.

Esse processo é repetido para os passos subsequentes, em que procuramos as combinações lineares que maximizam a variância amostral, levando em consideração as restrições das covariâncias amostrais entre as combinações anteriores.

Em resumo, os componentes principais da amostra são obtidos como as combinações lineares que capturam a maior variação dos dados amostrais, sendo essas combinações mutuamente não correlacionadas. Isso nos permite resumir a variação dos dados em um espaço de dimensão menor, destacando as características mais significativas dos dados originais.

$$\frac{\mathbf{a}_1^{'}\mathbf{S}\mathbf{a}_1}{\mathbf{a}_1^{'}\mathbf{a}_1}.$$

O máximo é o maior autovalor $\hat{\lambda}_1$ alcançado por $\mathbf{a}_1 =$ autovetor $\hat{\mathbf{e}}_1$ de $\mathbf{S}$. Escolhas sucessivas de $\mathbf{a}_i$ maximizam sujeito a $0 = \mathbf{a}_i^{'}\mathbf{S}\hat{\mathbf{e}}_k = \mathbf{a}_i^{'}\hat{\lambda}k\hat{\mathbf{e}}_k$, ou seja, $\mathbf{a}_i$ é perpendicular a

$\hat{\mathbf{e}}_k$. Assim, obtemos os seguintes resultados sobre os componentes principais da amostra:

Se $\boldsymbol{S} = \{s_{ik}\}$ são os elementos da matriz de covariância amostral $p \times p$ com pares de autovalor-autovetor $\left(\hat{\lambda}_1, \hat{\mathbf{e}}_1\right), \left(\hat{\lambda}_2, \hat{\mathbf{e}}_2\right) \ldots, \left(\hat{\lambda}_p, \hat{\mathbf{e}}_p\right)$, o i-ésimo componente principal da amostra é dado por:

$$\hat{y}_i = \hat{\mathbf{e}}_{\mathrm{i}}^{'}\mathbf{x} = \hat{e}_{i1}x_1 + \hat{e}_{i2}x_2 + \cdots + \hat{e}_{ip}x_p, \quad i = 1, 2, \ldots, p$$

em que $\hat{\lambda}_1 \geq \hat{\lambda}_2 \geq \cdots \geq \hat{\lambda}_p \geq 0$ e $\mathbf{x}$ representa um valor observado de $X_1, X_2, \ldots, X_p$. Então,

$$\text{Variância amostral } (\hat{y}_k) = \hat{\lambda}_k, \quad k = 1, 2, \ldots, p$$

$$\text{Covariância amostral } (\hat{y}_i, \hat{y}_k) = 0, \quad i \neq k$$

$$\text{Variância amostral total } = \sum_{i=1}^{p} s_{ii} = \hat{\lambda}_1 + \hat{\lambda}_2 + \cdots + \hat{\lambda}_p$$

$$r_{\hat{y}_l, x_k} = \frac{\hat{e}_{ik}\sqrt{\hat{\lambda}_i}}{\sqrt{s_{kk}}}, \quad i, k = 1, 2, \ldots, p.$$

Denotamos os componentes principais da amostra por $\hat{y}_1, \hat{y}_2, \ldots, \hat{y}_p$, independentemente de terem sido obtidos a partir de S ou $\mathbf{R}$. Os componentes construídos a partir de S e $\mathbf{R}$ não são os mesmos. Em geral, ficará claro pelo contexto qual matriz está sendo utilizada, e a notação única $\hat{y}_i$ é conveniente. Também é conveniente rotular os vetores de coeficientes do componente por $\hat{\mathbf{e}}_i$ e as variâncias do componente por $\hat{\lambda}_i$ para ambas as situações.

As observações $\mathbf{x}_j$ são frequentemente "centralizadas" subtraindo-se $\overline{\mathbf{x}}$. Isso não tem efeito na matriz de covariância amostral S e fornece o i-ésimo componente principal da amostra igual a:

$$\hat{y}_i = \hat{\mathbf{e}}_i^{'}(\mathbf{x} - \overline{\mathbf{x}}), \quad i = 1, 2, \ldots, p$$

Para qualquer observação do vetor $\mathbf{x}$, se considerarmos os valores do i-ésimo componente

$$\hat{y}_{ji} = \hat{\mathbf{e}}_i^{'}\left(\mathbf{x}_j - \overline{\mathbf{x}}\right), \quad j = 1, 2, \ldots, n$$

gerado ao substituir cada observação $\mathbf{x}_j$ no lugar do arbitrário $\mathbf{x}$, então

$$\overline{\hat{y}}_i = \frac{1}{n}\sum_{j=1}^{n} \hat{\mathbf{e}}_i' \left(\mathbf{x}_j - \overline{\mathbf{x}}\right) = \frac{1}{n}\hat{\mathbf{e}}_i' \left(\sum_{j=1}^{n} \left(\mathbf{x}_j - \overline{\mathbf{x}}\right)\right) = \frac{1}{n}\hat{\mathbf{e}}_i'\mathbf{0} = 0$$

Isso significa que a média amostral de cada componente principal é zero. As variâncias amostrais ainda são dadas pelos $\hat{\lambda}_i$.

2.2.3 Escolha do número de componentes Principais

A análise de componentes principais pode ser feita utilizando o critério dos autovalores maiores do que 1 ou por meio do gráfico *scree plot*. O *scree plot* é uma ferramenta visual que auxilia na determinação do número de componentes principais, identificando o ponto em que a curva se estabiliza, formando um cotovelo.

Além disso, de acordo com Johnson e Wichern (2002), é recomendado que a variância total acumulada seja pelo menos 80%. Para ilustrar a escolha do número de componentes, foi realizado um exemplo no Capítulo 3.

2.2.4 Interpretação dos componentes Principais

A interpretação dos resultados da análise de componentes principais (ACP) desempenha um papel fundamental na compreensão das relações subjacentes entre as variáveis originais e os componentes gerados. Esta subseção aborda a abordagem e os passos essenciais para interpretar os componentes principais de um conjunto de dados. As seguintes interpretações são realizadas:

- **Explorando a Variância Explicada:** Antes de iniciar a interpretação, é prudente examinar a proporção da variância explicada por cada componente principal. Isso pode ser realizado traçando um gráfico da proporção acumulada da variância explicada em relação ao número de componentes. A decisão de quantos componentes reter deve ser baseada em um equilíbrio entre a redução da dimensionalidade e a retenção da informação crítica. Geralmente, uma proporção acumulada de 0,85 a 0,95 pode indicar uma escolha razoável de componentes a serem mantidos;

- **Selecionando componentes e Interpretando Cargas:** Uma vez determinado o número de componentes a serem retidos, é hora de interpretar as cargas dos componentes principais. As cargas representam as contribuições relativas das variáveis

originais na formação de cada componente. Valores positivos e negativos nas cargas indicam a direção do impacto da variável sobre o componente. Para interpretar as cargas:

  - **Cargas Elevadas:** Variáveis com cargas próximas a 1 ou -1 em um componente têm uma forte relação com esse componente. Uma carga próxima a 1 indica uma relação positiva, enquanto uma carga próxima a -1 indica uma relação negativa. Essas variáveis podem ser consideradas como aquelas que mais influenciam o componente;
  - **Cargas Próximas a 0:** Cargas próximas a 0 indicam que a variável tem pouca influência na formação desse componente. Isso pode significar que a variável tem pouca relação com o padrão representado pelo componente;

- **Análise Contextual e Domínio do Problema:** A interpretação das cargas dos componentes principais não é uma tarefa trivial e muitas vezes requer conhecimento de domínio. Por exemplo, se a análise de componentes principais for aplicada a dados de mercado financeiro, as cargas podem representar setores econômicos, indicadores financeiros ou outras métricas relevantes. O contexto é essencial para dar significado às relações identificadas pelas cargas.

2.3 Alguns passos do ajusto modelo de regressão logística

Nesta seção explicamos os passos feitos na aplicação da regressão logística, assim como a demonstração de algumas métricas utilizadas.

2.3.1 Separação dos dados e validação cruzada

A separação dos dados em conjuntos de treinamento e teste é uma etapa fundamental ao realizar uma análise de regressão logística (ou qualquer outra técnica de modelagem preditiva), de acordo com Izbicki e dos Santos (2020). O erro quadrático médio, ou risco observado, é uma medida comum de avaliação da qualidade de um modelo de previsão ou ajuste a um conjunto de dados e tende a subestimar o verdadeiro risco de um modelo.

Quando utilizado na seleção de modelos, ele pode resultar em *overfitting*, ou seja, um ajuste excessivamente preciso aos dados de treinamento. Sua fórmula é dada por:

$$\mathrm{EQM}(g) := \frac{1}{n}\sum_{i=1}^{n}\left(Y_i - g\left(\mathbf{X}_i\right)\right)^2.$$

Isto ocorre pois g foi escolhida de modo a ajustar bem $(\mathbf{X}_1, Y_1), \ldots, (\mathbf{X}_n, Y_n)$.

Uma opção para solucionar esse problema é dividir os dados em dois conjuntos: Conjunto Teste e Conjunto Treino. Podemos usar, por exemplo, 70% dos dados para realizar o treinamento e o restante para o teste. Utilizamos o conjunto de dados de treinamento apenas para estimar os parâmetros do parâmetro ou função, geralmente representada como g (por exemplo, estimar os coeficientes em uma regressão linear). Por outro lado, usamos o conjunto de dados de validação exclusivamente para avaliar a qualidade do modelo.

É uma prática sólida escolher aleatoriamente as amostras que constituem os conjuntos de treinamento e validação. Isso é feito usando um gerador de números aleatórios para determinar quais amostras serão usadas para treinamento e quais para validação. Essa abordagem ajuda a evitar questões decorrentes de um banco de dados previamente ordenado com base em alguma covariável (por exemplo, um observador pode ter organizado as entradas com base em uma variável específica).

O ato de separar os dados em duas partes e usar uma delas para calcular a estimativa do risco é conhecido como *data splitting* (divisão dos dados). Uma versão aprimorada desse método é a validação cruzada, que utiliza todo o conjunto de dados de amostra. Dentro da validação cruzada, utilizamos o método *Holdout*, que particiona os dados em dois conjuntos, um para treinamento e outro para validação, de acordo com uma proporção específica estabelecida antecipadamente, como 80-20 (onde 80% dos dados são destinados ao treinamento e 20% para validação) ou 70-30. Durante esse processo, o algoritmo é treinado com o subconjunto destinado ao treinamento, e os dados remanescentes são utilizados para fazer previsões.

A separação dos dados em treinamento e teste possui vários benefícios, entre eles:

1. *Avaliar o desempenho do modelo:* A divisão dos dados permite que o modelo seja treinado em um conjunto de dados (conjunto de treinamento) e, posteriormente, testado em outro conjunto de dados independente (conjunto de teste). Isso nos ajuda a avaliar a capacidade de generalização do modelo, ou seja, sua capacidade de fazer previsões precisas em dados não utilizados durante o treinamento;

2. *Prevenir overfitting:* O *overfitting* ocorre quando o modelo se ajusta excessivamente aos dados de treinamento, capturando padrões e ruídos específicos desse conjunto, mas não é capaz de generalizar para novos dados. A separação em conjuntos de treinamento e teste ajuda a identificar se o modelo está superajustado aos dados de treinamento, permitindo ajustes para melhorar a generalização;

3. *Avaliar o desempenho em dados novos:* Ao testar o modelo em um conjunto de teste independente, podemos obter uma estimativa mais realista do desempenho do modelo em dados futuros. Isso é especialmente relevante quando aplicamos o modelo em cenários do mundo real, onde a qualidade das previsões é de interesse primordial;

4. *Validação do modelo:* A separação dos dados permite que usemos diferentes métricas de avaliação, como acurácia, precisão, *recall*, *F1-score*, entre outras, para avaliar o desempenho do modelo de forma mais abrangente e identificar suas fraquezas e pontos fortes;

5. *Melhorar o processo de modelagem:* Ao testar vários modelos ou ajustar parâmetros, a separação dos dados permite comparar e selecionar o modelo mais adequado para a tarefa específica, com base em seu desempenho em dados de teste.

2.3.2 Ajuste do modelo - *stepwise*

A técnica conhecida *stepwise* é um método de ajuste de modelos de regressão que automatiza a seleção das variáveis que servirão como preditores (Izbicki e dos Santos, 2020). A cada etapa, o algoritmo adiciona ou remove uma variável do conjunto de variáveis explicativas com base em um critério específico predefinido. Geralmente, esses critérios podem ser testes estatísticos, como:

- AIC (critério de informação de Akaike, do inglês *Information Criterion*)

$$-2l\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 \mid \mathbf{X}, \boldsymbol{y}\right) + 2d,$$

em que $l\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 \mid \mathbf{X}, \boldsymbol{y}\right)$ é o logaritmo da verossimilhança calculado com base nas estimativas de verossimilhança dos parâmetros e d é o número de parâmetros no modelo;

- BIC (Bayesian Information Criterion)

$$-2l\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 \mid \mathbf{X}, \boldsymbol{y}\right) + d \log n,$$

em que $l\left(\hat{\boldsymbol{\beta}}, \hat{\sigma}^2 \mid \mathbf{X}, \boldsymbol{y}\right)$ é o logaritmo da verossimilhança calculado com base nas estimativas de verossimilhança dos parâmetros e d é o número de parâmetros no modelo; ou

- o valor-p do teste de significância para o coeficiente de regressão.

Este procedimento automatizado visa simplificar o processo de seleção de variáveis e pode ser aplicado sequencialmente para determinar quais variáveis fornecem o melhor ajuste ao modelo de regressão.

Stepwise Forward

O método *Stepwise Forward* é uma técnica comum de seleção de variáveis em modelos de regressão. Inicialmente, o modelo começa sem nenhuma variável explicativa e, em cada etapa, uma variável é adicionada ao modelo com base em sua contribuição para a explicação da variância no resultado. A seleção das variáveis ocorre de forma iterativa e leva em consideração critérios estatísticos, como o valor-p (*p-value*) e medidas de ajuste do modelo.

Para ajustar o modelo de regressão logística utilizando o método *Stepwise Forward*, seguimos os seguintes passos:

1. *Inicialização do modelo:* Iniciamos o modelo sem nenhuma variável explicativa;

2. *Adição de Variáveis:* Iterativamente, adicionamos a variável que proporciona o maior ganho de ajuste estatisticamente significativo ao modelo. Isso é feito considerando critérios como o valor-p associado ao coeficiente estimado;

3. *Critérios de parada:* O processo de adição de variáveis continua até que não haja mais variáveis com significância estatística suficiente para melhorar o modelo ou até que sejam atingidos critérios de parada predefinidos, como um valor máximo de variáveis a serem incluídas;

4. *Avaliação do modelo:* Após o ajuste do modelo, avaliamos sua qualidade usando métricas de ajuste, como o AIC (*Akaike Information Criterion*) ou o BIC (*Bayesian*

Information Criterion). Também verificamos a significância estatística e interpretabilidade dos coeficientes estimados;

5. *Validação cruzada:* Para evitar overfitting, realizamos validação cruzada do modelo utilizando conjuntos de treinamento e teste separados. Isso nos permite estimar a capacidade de generalização do modelo em dados não utilizados no ajuste.

Stepwise Backward

O método *Stepwise Backward* é outra abordagem comum de seleção de variáveis em modelos de regressão, mas difere do *Stepwise Forward* na direção em que as variáveis são adicionadas ou removidas do modelo. No *Backward Stepwise*, começamos com um modelo que inclui todas as variáveis explicativas disponíveis e, em cada etapa, consideramos a remoção de uma variável por vez com base em certos critérios estabelecidos.

O processo de ajustar um modelo de regressão logística usando o método *Backward Stepwise* é conduzido da seguinte forma:

1. Inicialização do Modelo: Começamos com um modelo que inclui todas as variáveis explicativas disponíveis;

2. Remoção de Variáveis: Iterativamente, consideramos a remoção de uma variável do modelo. Avaliamos o impacto da remoção da variável na qualidade do modelo, usando critérios como o valor-p associado ao coeficiente estimado daquela variável;

3. Critérios de Parada: O processo de remoção de variáveis continua até que não haja mais variáveis que possam ser removidas sem prejudicar significativamente a qualidade do modelo, ou até que critérios de parada pré-definidos sejam alcançados, como a remoção de um número máximo de variáveis;

4. Avaliação do Modelo: Após a etapa de remoção das variáveis, avaliamos a qualidade do modelo resultante. Isso inclui a análise de métricas de ajuste, como o AIC (*Akaike Information Criterion*) ou o BIC (*Bayesian Information Criterion*), bem como a análise da significância estatística e interpretabilidade dos coeficientes estimados nas variáveis remanescentes;

5. Validação Cruzada: Assim como no método *Stepwise Forward*, é recomendável realizar a validação cruzada do modelo ajustado usando conjuntos de treinamento e

teste separados para verificar a capacidade de generalização do modelo e evitar o *overfitting*.

Em resumo, enquanto o *Stepwise Forward* começa sem variáveis e adiciona uma por uma, o *Backward Stepwise* começa com todas as variáveis e remove uma por uma, ambos com o objetivo de encontrar um modelo final que forneça um bom ajuste aos dados.

2.3.3 Curva ROC

A Curva ROC do inglês é uma representação visual valiosa usada para avaliar o desempenho de modelos de classificação, especialmente, em situações nas quais lidamos com dois resultados possíveis, como "positivo"e "negativo". O gráfico mostra como a taxa de acertos para as situações verdadeiramente positivas (TPR) se relaciona com a taxa de erros para os casos que foram erroneamente classificados como positivos (FPR), enquanto ajustamos o ponto de corte do modelo.

Neste gráfico é possível observar o comportamnento do à medida que mudamos o limite de decisão. Quanto mais perto do canto superior esquerda do gráfico, melhor o modelo, pois significa que está classificando corretamente muitos casos positivos (alta TPR) e cometendo poucos erros ao considerar negativos como positivos (baixo FPR). Podemos resumir o desempenho geral do modelo com uma única métrica chamada Área Sob a Curva (AUC), ajudando-nos a comparar diferentes modelos e escolher aquele que se ajusta melhor ao nosso problema de classificação.

A AUC é uma métrica amplamente utilizada para avaliar o desempenho de classificadores em problemas de classificação binária. Em diversos estudos, a AUC tem sido empregada para construir modelos de aprendizado otimizados e também para comparar diferentes algoritmos de aprendizado. Ao contrário de métricas que dependem de um limiar ou probabilidade específica, a AUC reflete o desempenho geral do classificador em ordenar corretamente os exemplos.

Para o caso de problemas de classificação binária, o valor da AUC pode ser calculado da seguinte maneira:

$$AUC = \frac{S_p - n_p\left(n_n + 1\right)/2}{n_p n_n}$$

em que S_p é a soma de todos os exemplos positivos corretamente classificados, e n_p e n_n representam o número de exemplos positivos e negativos, respectivamente. A AUC foi comprovada tanto teoricamente quanto empiricamente como uma métrica superior à

acurácia para avaliar o desempenho do classificador e identificar soluções ótimas durante o treinamento do modelo.

2.3.4 Medidas de avaliação

No campo da classificação de dados, diferentes tipos de informações, como dados comerciais, textos, DNAs e imagens, podem ser classificados. A classificação de dados pode ser realizada por meio de duas abordagens principais: classificação binária e multiclasse.

De acordo com Hossin e Sulaiman (2015), os principais objetivos das métricas de avaliação são avaliar a capacidade de generalização do classificador treinado, selecionar o melhor classificador entre diferentes opções e discriminar e selecionar a melhor solução durante o treinamento da classificação. Para isso, a escolha adequada da medida é crucial.

Embora a precisão seja comumente utilizada para discriminar a melhor solução, ela apresenta algumas limitações, especialmente em cenários de distribuição desigual de classes.

2.3.5 Tipos de Medidas Discriminadoras Baseadas em Limiar

As medidas de avaliação usadas em problemas são empregadas em duas etapas: treinamento, para otimizar o algoritmo de classificação, e teste, para medir a eficácia do classificador produzido com dados não vistos anteriormente. Nosso foco é explorar como tais medidas podem discriminar e selecionar a melhor solução para construir classificadores. As medidas discriminadoras baseadas em limiar são usados em problemas de classificação binária, permitindo avaliar quão bem a solução ótima é discriminada durante o treinamento do classificador. Usando tais medidas, uma tabela de confundimento é construída, na qual é possível observar a quantidade de observações classificadas corretamente e incorretamente para cada classe prevista e real.

Na Tabela 2.3.1 as quantidades VP (verdadeiros positivos) e VN (verdadeiros negativos) representam o número de resultados positivos e negativas corretamente classificadas, respectivamente. Além disso, FP (falsos positivos) e FN (falsos negativos) representam o número de resultados negativos e positivas classificadas incorretamente, respectivamente.

Com base nos resultados na Tabela 2.3.1, várias medidas comumente utilizadas podem ser calculadas para avaliar o desempenho do classificador sob diferentes perspectivas. Essas medidas são essenciais para determinar a eficácia da solução encontrada pelo clas-

sificador durante o treinamento.

Tabela 2.3.1: Tabela de Confundimento para Classificação Binária.

	Positivos reais	Negativos reais
Positivos preditos	Verdadeiro positivo (VP)	Falso negativo (FN)
Negativos preditos	Falso positivo (FP)	Verdadeiro Negativo (VN)

A partir da Tabela 2.3.1, podemos obter algumas medidas bastante utilizadas, tais como:

- Acurácia: ACC = (VP + VN) / (VP + FP + VN + FN) (proporção de previsões corretas em relação ao número total de casos).
- Taxa de erro: ER = (FP + FN) / (VP + FP + VN + FN) (proporção de previsões incorretas em relação ao número total de casos).
- Sensibilidade/*Recall*: S = VP/(VP + FN) (proporção de verdadeiros positivos (TP) em relação ao total de casos reais positivos).
- Especificidade: E = VN/(VN + FP) (proporção de verdadeiros negativos (TN) em relação ao total de casos reais negativos).
- Valor preditivo positivo/Precisão: VPP = VP/(VP + FP) (é a probabilidade de que um caso classificado como positivo seja realmente positivo).
- Valor preditivo negativo: VPN = VN/(VN + FN) (probabilidade de que um casp-classificado como negativo seja realmente negativo).
- Estatística F1: F1 = $\frac{2}{1/S+1/VPP}$ (a média harmônica entre S e VPP).

Capítulo 3

Aplicação e Exemplos

Neste capítulo acrescentamos aplicações com as metodologias estudadas anteriormente. Primeiramente, aplicamos a regressão logística em um conjunto de dados do Titanic, dimensionalidade baixa, e em seguida em um banco de dados que visa detectar *spam* em *e-mails*, dimensionalidade alta, por fim, aplicamos análise de componentes principais em dados relacionados a socioeconomia de Madison.

3.1 Regressão logística aplicada em dados com baixa dimensionalidade

Neste exemplo optamos por um banco de dados mais simples com o intuito de exibir a interpretabilidade e o entendimento do modelo para tal caso, cujos dados estão disponíveis no *software* R Core Team (2023).

3.1.1 Descrição dos dados:

O naufrágio do Titanic é um dos mais trágicos naufrágios da história. Em 15 de abril de 1912, durante sua viagem inaugural, o RMS Titanic, considerado impossível de afundar, afundou após colidir com um *iceberg*. Infelizmente, não havia botes salva-vidas suficientes para todos a bordo, resultando na morte de 1502 dos 2224 passageiros ou tripulantes. Dentre os passageiros, alguns grupos de pessoas tinham maior probabilidade de sobreviver do que outros. Para investigar um pouco mais desse caso, construímos um modelo de predição usando regressão logística usando as seguintes variáveis de uma amostra de 889 passageiros ou tripulantes.

- **Y**: Indica se a pessoa sobreviveu (1) ou não (0);
- X_1: Classe em que a pessoa viajou (1=1^a; 2=2^a; 3=3^a);
- X_2: Gênero da pessoa (1=masculino ou 0=feminino);
- X_3: Idade da pessoa em anos;
- X_4: Número de irmãos e cônjuges (irmão, irmã, meio-irmão ou meia-irmã, marido ou esposa) do passageiro a bordo do Titanic;
- X_5: Número de pais e filhos (mãe, pai, filho, filha, enteado ou enteada) do passageiro a bordo do Titanic;
- X_6: Porto de embarque (0 = Cherbourg; 1 = Queenstown; 2 = Southampton);

3.1.2 Análise descritiva

Começamos a análise descritiva pelo sexo dos passageiros, a proporção do sexo feminino é de 35%, enquanto que 65% são homens . Em relação aos locais de embarcação: Cherbourg, Queenstown ou Southampton, temos que a porcentagem de embarcações em cada cidade é: 20%, 9% e 61%, respectivamente. Assim como nas duas variáveis analisadas sobre o gênero, não há um equilíbrio entre as proporções de todas as possibilidades, mas isso se dá devido a terceira classe que conta com 54% dos passageiros, mais que a metade, possuindo mais que o dobro de passageiros que a primeira classe (24%) e a segunda (22%).

A variável IC demonstra que a grande maioria dos viajantes possuem no máximo um irmão ou cônjuge a bordo, pois 92% dos passageiros possuem no máximo um desses dois graus de parentesco. O mesmo ocorre com o número de pais/filhos a bordo, em que 89% passageiros possuem no máximo um pai/filho a bordo.

Outra informação importante a se notar é a idade, que possui uma amplitude considerável que vai de 0 até 80 anos, sendo que sua média é de 29 anos e sua mediana de 28 anos.

Analisando os sobreviventes, temos que dos 889 passageiros a bordo apenas 35% sobreviveram, dos quais 74% eram mulheres e 15% eram crianças (passageiros com idades menores ou iguais a 11 anos). Além disso, 51% das crianças sobreviventes eram do sexo feminino e 49% do sexo masculino. Feita as análises de todas as variáveis e a análise

conjunta de algumas variáveis de interesse, analisamos uma matriz de correlação que é uma tabela que mostra as correlações entre várias variáveis em um conjunto de dados. Ela é uma ferramenta útil para identificar padrões e associações lineares entre as variáveis, ajudando na análise exploratória de dados e na seleção de variáveis para modelos de regressão.

Tabela 3.1.1: Correlação entre as variáveis utilizadas no exemplo.

	X_1	X_3	X_4	X_5
X_1	1,00	-0,38	0,06	0,02
X_3	-0,38	1,00	-0,19	-0.13
X_4	0,06	-0,19	1,00	0,37
X_5	0,02	-0,13	0,37	1,00

Observando os resultados da Tabela 3.1.1 concluímos que não há variáveis correlacionadas. Dessa forma, podemos começar a ajustar o modelo.

3.1.3 Separação os dados

Então, após a realização da análise descritiva, separamos os dados em treinamento e teste considerando os 916 primeiros termos para treinamento e o restante para teste, uma vez que já estão distribuidos de forma aleatória já que a coleta dos dados não possui ordem específica, sendo suas proporções 70% e 30%, respectivamente, de forma que cada grupo resultante tenha a proporção de sobreviventes e mortos parecidas com os dados completos. Por fim, chegamos as seguintes proporções:

Tabela 3.1.2: Tabela da proporção de sobreviventes em cada conjunto de dados.

Proporção de sobreviventes					
Conjunto Completo		Conjunto de treino		Conjunto de teste	
Sobreviventes	Não sobreviventes	Sobreviventes	Não sobreviventes	Sobreviventes	Não sobreviventes
0,38	0,62	0,39	0,61	0,36	0,64

3.1.4 Ajuste do Modelo

Ajustamos modelo de regressão logística utilizando o método *Stepwise Forward* com as variáveis de treino, pois trata-se de um conjunto de dados com poucas variáveis. O objetivo desse método é selecionar as variáveis mais relevantes para o modelo, avançando

iterativamente ao adicionar ou remover variáveis com base em critérios específicos, como a significância estatística ou o poder preditivo.

Após aplicar o método *Stepwise Forward* ao conjunto de dados, obtivemos um modelo de regressão logística final com as variáveis selecionadas de acordo com os critérios mencionados. Os resultados incluem os coeficientes estimados, os valores-p associados e outras métricas relevantes para a interpretação e validação do modelo.

Tabela 3.1.3: Coeficientes do modelo ajustado.

Variável	**Estimativa**	**Erro Padrão**	**z valor**	**Pr(>|z|)**
(Intercepto)	3,911	0,473	8,596	$< 2 \times 10^{-16}$
$X_{1,2}$	-0,872	0,297	-3,092	0,002
$X_{1,3}$	-2,175	0,298	-7,222	$5,13 \times 10^{-13}$
$X_{2,1}$	-2,677	0,201	-13,531	$< 2 \times 10^{-16}$
X_3	-0,031	0,008	-4,903	$9,43 \times 10^{-7}$
X_4	-0,248	0,109	-2,947	0,003
X_5	-0,089	0,119	-0,785	0,432
$X_{6,1}$	-0,076	0,381	-0,148	0,883
$X_{6,2}$	-0,465	0,240	-1,813	0,070

Além das informações contidas na tabela 3.1.3, o AIC (*Akaike Information Criterion*) é 805,44. O AIC é uma medida de qualidade do ajuste do modelo e é usado para comparar diferentes modelos. Quanto menor o valor do AIC, melhor o ajuste do modelo aos dados, chegando ao modelo:

$$\hat{\pi}(\mathbf{x_i}) = \frac{\exp\left(3,911 - 0,872X_{1,2} - 2,175X_{1,3} - 2,677X_{2,1} - 0,031X_3 - 0,248X_4\right)}{1 + \exp\left(3,911 - 0,872X_{1,2} - 2,175X_{1,3} - 2,677X_{2,1} - 0,031X_3 - 0,248X_4\right)} \tag{3.1}$$

em que, por exemplo, $X_{1,2}$ representa a variável X_1 quando a classe é igual a dois e a categoria de referência é 1.

Para interpretar cada coeficiente estimado, assumimos que todos os outros se mantém constantes. Dessa forma:

- O coeficiente estimado de $X_{1,2}$ (-0,872): Esse coeficiente indica a mudança esperada no *log-odds* (logaritmo da razão de chances) da sobrevivência quando uma pessoa está na segunda classe em comparação com a primeira classe. Um valor negativo sugere que estar na segunda classe está associado a uma menor probabilidade de sobrevivência em relação a primeira classe;

- O coeficiente estimado de $X_{1,3}$ (-2,175): Da mesma forma, esse coeficiente indica a mudança esperada no *log-odds* da sobrevivência quando uma pessoa está na ter-

ceira classe em comparação com a primeira classe. Novamente, um valor negativo sugere que estar na terceira classe está associado a uma menor probabilidade de sobrevivência em relação à primeira classe.

- O coeficiente de $X_{2,1}$ (-2,677): Esse coeficiente estimado indica a mudança esperada no *log-odds* da sobrevivência quando uma pessoa é do sexo masculino em comparação com o sexo feminino. Um valor negativo sugere que ser do sexo masculino está associado a uma menor probabilidade de sobrevivência em relação ao sexo feminino.

- O coeficiente estimado de X_3 (-0,031): Esse coeficiente indica a mudança esperada no *log-odds* da sobrevivência para cada aumento de uma unidade na idade da pessoa. Um valor negativo sugere que o aumento da idade está associado a uma diminuição na probabilidade de sobrevivência a cada ano de vida.

- O coeficiente estimado de X_4 (-0,248): Esse coeficiente indica a mudança esperada no *log-odds* da sobrevivência para cada aumento de uma unidade no número de irmãos/cônjuges a bordo. Um valor negativo sugere que um maior número de irmãos/cônjuges a bordo está associado a uma menor probabilidade de sobrevivência.

3.1.5 Curva ROC

Utilizamos essa ferramenta para descobrir o ponto de corte ideal do nosso problema e chegamos ao gráfico na Figura 3.1:

Sendo o ponto de corte igual a 0,87, quando a probabilidade de sobrevivência estimada para um indivíduo é maior ou igual a 0,87, o modelo classificará essa pessoa como "sobrevivente"(1). Por outro lado, se a probabilidade estimada for menor do que 0,87, o modelo classificará a pessoa como "não sobrevivente"(0). Para melhor entendimento, na Tabela 3.1.4 são mostrados os resultados.

Tabela 3.1.4: Comparação entre predição e resultados reais dos sobreviventes.

Indivíduo	670	671	672	673	674
Predição	0,09	0,06	0,90	0,72	0,37
Classificação via curva ROC	0	0	1	0	0
Classificação real	0	0	1	1	0

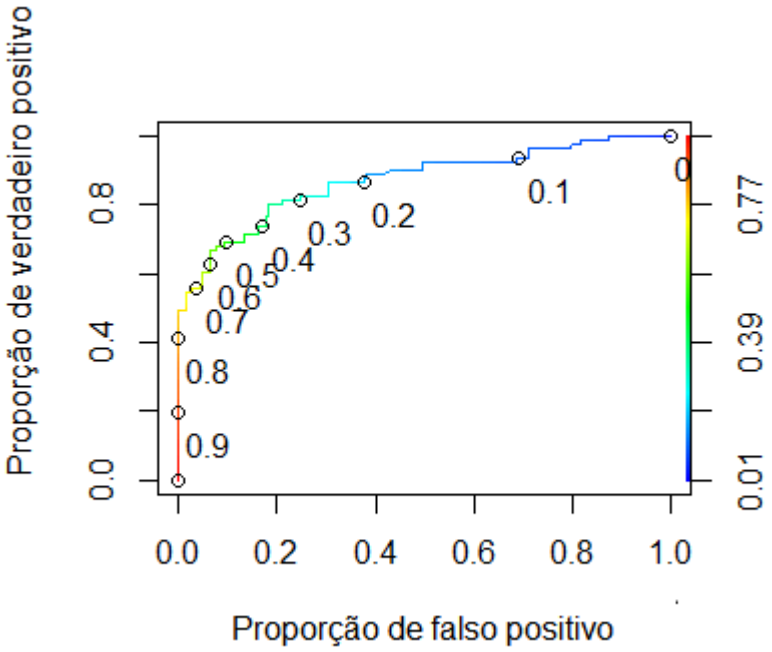

Figura 3.1: Curva ROC para o modelo ajustado.

3.1.6 Tabela de confundimento

Considerando o exemplo do Titanic, construímos a Tabela 3.2.4.

Tabela 3.1.5: Tabela de confusão.

	Referência	
Predição	**0**	**1**
0	$\frac{141}{222}$	$\frac{60}{222}$
1	0	$\frac{21}{222}$

Com base na Tabela 3.2.4, temos:

- Verdadeiros Negativos (VN): Na célula correspondente à previsão 0 e à classe verdadeira 0, temos um valor de 141. Isso significa que o modelo corretamente previu que 141 pessoas não sobreviveram;

- Falsos Positivos (FP): Na célula correspondente à previsão 1 e à classe verdadeira 0, temos um valor de 0. Isso indica que o modelo não cometeu nenhum erro do tipo I, ou seja, não previu erroneamente que alguém sobreviveu quando na verdade não sobreviveu;

- Falsos Negativos (FN): Na célula correspondente à previsão 0 e à classe verdadeira 1, temos um valor de 60. Isso significa que o modelo cometeu 60 erros do tipo II, ou seja, falhou em prever corretamente que 60 pessoas sobreviveram;

- Verdadeiros Positivos (VP): Na célula correspondente à previsão 1 e à classe verdadeira 1, temos um valor de 21. Isso indica que o modelo corretamente previu que 21 pessoas sobreviveram;

3.2 Regressão Logística aplicada nos dados com alta dimensionalidade

Neste exemplo, procuramos um banco de dados com mais variáveis que o exemplo anterior com a finalidade de exibir a diferença entre os modelos;

3.2.1 Descrição dos dados

Este conjunto de dados chamado de*spam database* é amplamente reconhecido e possui uma variável resposta binária. Cada linha representa um *e-mail*, que é classificado como *spam* ou não *spam*. O conjunto de dados contém 4601 *e-mails* com 48 variáveis que registram a frequência de ocorrência de palavras específicas nos *e-mails*, seis variáveis que registram a porcentagem de vezes que um caractere específico aparece nos *e-mails*, e mais três variáveis que registram o comprimento das sequências de letras maiúsculas.

3.2.2 Variáveis

Algumas das variáveis do conjunto de dados são:

- Y = *is.spam*: Indica se o *e-mail* é considerado como spam (0 = não, 1 = sim).

- $\mathbf{X_1}$= *word.freq.make*: Porcentagem de vezes que a palavra *make* apareceu *e-mail*.

- $\mathbf{X_2}$=*word.freq.address*: Porcentagem de vezes que a palavra *address* apareceu no *e-mail*.

- $\mathbf{X_{57}}$=*capital.run.length.total*: Número total de letras maiúsculas no *e-mail*.

3.2.3 Análise descritiva

Na análise descritiva das 58 variáveis notamos algumas semelhanças no que diz respeito ao formato dos dados. A última coluna do conjunto de dados `spambase.data` indica se o *e-mail* foi considerado como *spam* (1) ou não (0), ou seja, *e-mails* comerciais não solicitados. A maioria das variáveis indica se uma palavra ou caractere específico aparece frequentemente no *e-mail*. As variáveis de *run-length* (55 a 57) registram o comprimento de sequências de letras maiúsculas consecutivas. Aqui estão as definições das variáveis:

1. 48 variáveis contínuas [0,100] do tipo `word_freq_WORD`, que representam a porcentagem de palavras no *e-mail* que correspondem a *WORD*, ou seja,

$$100 \times \frac{\text{(número de vezes que WORD aparece no } e\text{-}mail)}{\text{(número total de palavras no } e\text{-}mail).}$$

   Uma "palavra" neste caso é qualquer sequência de caracteres alfanuméricos delimitada por caracteres não alfanuméricos ou o fim da *string*;

2. Seis variáveis contínuos [0,100] do tipo `char_freq_CHAR`, que representam a porcentagem de caracteres no *e-mail* que correspondem a *CHAR*, ou seja,

$$100 \times \frac{\text{(número de ocorrências de CHAR)}}{\text{(número total de caracteres no } e\text{-}mail)}$$

3. Uma variável inteira [1,...] do tipo `capital_run_length_average`, que representa o comprimento médio de sequências ininterruptas de letras maiúsculas;

4. Uma variável inteira [1,...] do tipo `capital_run_length_longest`, que representa o comprimento da sequência ininterrupta mais longa de letras maiúsculas;

5. Uma Vvriável inteira [1,...] do tipo `capital_run_length_total`, que representa a soma dos comprimentos das sequências ininterruptas de letras maiúsculas, ou seja, o número total de letras maiúsculas no *e-mail*;

6. Uma variável nominal {0,1} do tipo *spam*, que indica se o *e-mail* foi considerado *spam* (1) ou não (0), ou seja, *e-mails* comerciais não solicitados.

3.2.4 Separação dos dados

Conforme explicado na Seção 3.1, após a realização da análise descritiva, separamos os dados em treinamento e teste por meio de um sorteio aleatório com proporções 70% e 30%, respectivamente, de forma que cada grupo resultante tenha a proporção de *spam* e não *spam* parecidas com os dados completos. Dessa forma, obtemos as proporções mostradas na Tabela 3.2.1.

Tabela 3.2.1: Tabela da proporção de *spam* em cada conjunto de dados.

Proporção de *spam*					
Conjunto Completo		Conjunto de treino		Conjunto de teste	
spam	Não *spam*	*spam*	Não *spam*	*spam*	Não *spam*
0,39	0,61	0,40	0,60	0,40	0,60

3.2.5 Ajuste do Modelo

Nesta seção, descrevemos o processo de ajuste do modelo de regressão logística utilizando o método *Stepwise Backward* nos dados de treinamento, com o objetivo de detectar e prever *e-mails* de *spam* em nosso conjunto de dados. Devido à natureza complexa e extensa do conjunto, optamos por essa abordagem para selecionar as variáveis mais relevantes e construir um modelo mais eficiente.

Inicialmente, nosso conjunto de dados contém uma ampla gama de variáveis que medem a frequência de palavras, caracteres e outros padrões em cada *e-mail*, além de informações adicionais sobre os *e-mails* considerados *spam* ou não. Ao ajustar o modelo, todas as variáveis disponíveis foram inicialmente incluídas.

O método *Stepwise Backward* foi aplicado para realizar uma seleção sistemática das variáveis mais importantes para a predição de *e-mails* de *spam*. A cada iteração, uma variável foi removida do modelo se sua exclusão resultasse em um impacto mínimo na qualidade do ajuste, avaliado pelo critério estatístico AIC (Akaike Information Criterion), resultando nos seguintes coeficientes apresentados na Tabela 3.2.2.

Tabela 3.2.2: Coeficientes do Modelo de Regressão Logística.

Variável	**Estimativa**	**Erro Padrão**	**Valor Z**	**Valor P**
Intercepto	-2,100e+00	1,835e-01	-11,442	$<$ 2e-16
X_1	-5,480e-01	2,812e-01	-1,949	0,051317
X_2	-1,647e-01	8,716e-02	-1,890	0,058813
X_3	2,353e-01	1,413e-01	1,666	0,095763
X_4	1,717e+00	1,703e+00	1,008	0,313343
X_5	5,774e-01	1,212e-01	4,765	1,89e-06
...	...	...	...	...
X_{56}	3,723e-01	5,326e-02	6,990	2,75e-12
X_{57}	8,192e-04	2,370e-04	3,456	0,000548

Além das informações contidas na Tabela 3.2.2, o AIC (Akaike Information Criterion) é 1266,5. Portanto, chegamos ao modelo final:

$$\begin{aligned} A = &- 2,100 - 0,548X_1 - 0,165X_2 + 0,235X_3 \\ &+ 1,717X_4 + 0,577X_5 \\ &+ 1,159X_6 + 2,338X_7 + 0,7328X_8 \\ &+ 1,028X_9 - 0,2446X_{12} \\ &+ 0,425X_{14} + 1,277X_{15} + 0,8961X_{16} \\ &+ 1,306X_{17} + 0,2746X_{18} \\ &+ 1,764X_{20} + 0,2151X_{21} + 2,102X_{23} + 1,727X_{24} - 2,189X_{25} \\ &- 0,851X_{26} - 11,38X_{27} + 0,5145X_{28} - 2,056X_{29} - 1,030X_{33} \\ &- 2,018X_{35} + 1,161X_{36} - 0,844X_{39} - 2,214X_{42} \\ &- 1,730X_{43} - 1,889X_{44} - 0,8207X_{45} - 1,703X_{46} \\ &- 2,973X_{47} - 2,939X_{48} - 1,455X_{49} + 2,777X_{52} \\ &+ 3,344X_{53} + 3,161X_{54} + 0,3723X_{55} + 0,0008192X_{57} \end{aligned}$$

cuja probabilidade estimada de cada *e-mail* ser *spam* ou não é dada por:

$$\hat{\pi}(\mathbf{x_i}) = \frac{\exp(A)}{1 + \exp(A)} \tag{3.2}$$

A interpretação de alguns dos coeficientes estimados são:

- X_1 = **word_freq_make** (-0,548): Esse coeficiente indica a mudança esperada no *log-odds* da probabilidade de um *e-mail* ser classificado como spam quando a frequência da palavra *make* aumenta em uma unidade, mantendo todas as outras variáveis constantes. Um valor negativo sugere que um aumento na frequência da palavra *make* está associado a uma diminuição na probabilidade de ser *spam*

- X_2=**word_freq_address** (-0,165): Similarmente, esse coeficiente indica a mudança esperada no *log-odds* da probabilidade de ser *spam* quando a frequência da palavra *address* aumenta em uma unidade, mantendo todas as outras variáveis constantes. Um valor negativo sugere que um aumento na frequência da palavra *address* está associado a uma diminuição na probabilidade de ser *spam*

- X_3 = **word_freq_all** (0,235): Esse coeficiente indica a mudança esperada no *log-odds* da probabilidade de ser *spam* quando a frequência da palavra *all* aumenta em uma unidade, mantendo todas as outras variáveis constantes. Um valor positivo sugere que um aumento na frequência da palavra *all* está associado a um aumento na probabilidade de ser *spam*.

É importante lembrar que essas interpretações são feitas mantendo todas as outras variáveis constantes. Valores positivos nos coeficientes indicam uma relação positiva com a probabilidade de ser *spam*, enquanto valores negativos indicam uma relação negativa. Coeficientes maiores em magnitude têm um impacto mais significativo nas chances de ser *spam*. A constante (-2,100) é o valor de referência quando todas as frequências de palavras e caracteres são zero.

3.2.6 Curva ROC

Utilizamos essa ferramenta para descobrir o ponto de corte do nosso problema (Figura 3.2).

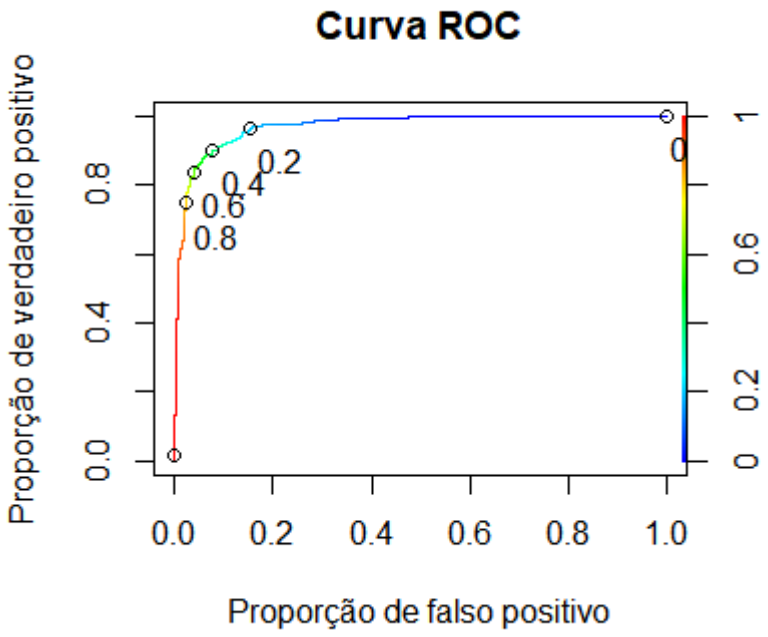

Figura 3.2: Curva Roc para o modelo ajustado.

Sendo o ponto de corte igual a 0,96, quando a probabilidade de *spam* estimado para um *e-mail* é maior ou igual a 0,96, o modelo classificará esse *e-mail* como *spam* (1). Por outro lado, se a probabilidade estimada for menor do que 0,96, o modelo classificará o *e-mail* como não *spam* (0). Para melhor entendimento, mostramos na Tabela 3.2.3 a indicação da observação (*e-mail*), ovalor da predição a partir do modelo ajustado nos dados de teste, a classificação da curva ROC a partir do ponto de corte e a real classificação do *e-mail*.

Tabela 3.2.3: Comparação entre classificação real e via curva ROC.

Observação	11	14	15	18	20
Predição	0,77	0,81	0,98	0,99	0,99
Classificação via curva ROC	0	0	1	1	1
Classificação real	1	1	1	1	1

De acordo com a Tabela 3.2.3, as observações 11 e 14 classificaram um *e-mail* como *spam* quando na verdade não eram. Já as observações 15, 18 e 20 acertaram suas predições.

3.2.7 Tabela de confundimento

Na Tabela 3.2.4 observamos a proporção de erros e acertos do modelo ajustado.

Tabela 3.2.4: Tabela de Confundimento.

	Referência	
Predição	**0**	**1**
0	$\frac{800}{1331}$	$\frac{242}{1331}$
1	$\frac{12}{1331}$	$\frac{277}{1331}$

- **Verdadeiros Negativos (VN)**: O modelo classificou corretamente 800 *e-mails* como não *spam* (classe 0);

- **Falsos Positivos (FP)**: O modelo erroneamente classificou 242 *e-mails* como *spam* (classe 1), quando na verdade são não *spam* (classe 0);

- **Falsos Negativos (FN)**: O modelo erroneamente classificou 12 *e-mails* como não *spam* (classe 0), quando na verdade são *spam* (classe 1);

- **Verdadeiros Positivos (VP)**: O modelo classificou corretamente 277 *e-mails* como *spam* (classe 1).

3.3 Aplicação de ACP

Para ilustrar a aplicação da ACP, usamos informações socioeconômicas em 14 regiões censitárias de Madison, Wisconsin, USA. Cinco variáveis foram observadas, cujos valores são mostrados na Tabela 3.2.8

X_1 = população total ($\times$1000);

X_2 = tempo mediano de estudo (em anos);

X_3 = total de empregados ($\times$1000);

X_4 = total de empregados no setor da saúde ($\times$1000);

X_5 = valor mediano da residência ($\times US\$10.000$).

Tabela 3.3.1: Dados Socioeconômicos de Madison.

População	Escolaridade	Empregados	Empregados na Saúde	Valor da Residência
5,935	14,2	2,265	2,27	2,91
1,523	13,1	0,597	0,75	2,62
2.599	12.7	1,237	1.11	1.72
4,009	15,2	1,649	0,81	3,02
4,687	14,7	2,312	2,50	2,22
8,044	15,6	3,641	4,51	2,36
2,766	13,3	1,244	1,03	1,97
6,538	17,0	2,618	2,39	1,85
6,451	12,9	3,147	5,52	2,01
3,314	12,2	1,606	2,18	1,82
3,777	13,0	2,119	2,83	1,80
1,530	13,8	0,798	0,84	4,25
2,768	13,6	1,336	1,75	2,64
6,585	14,9	2,763	1,91	3,17

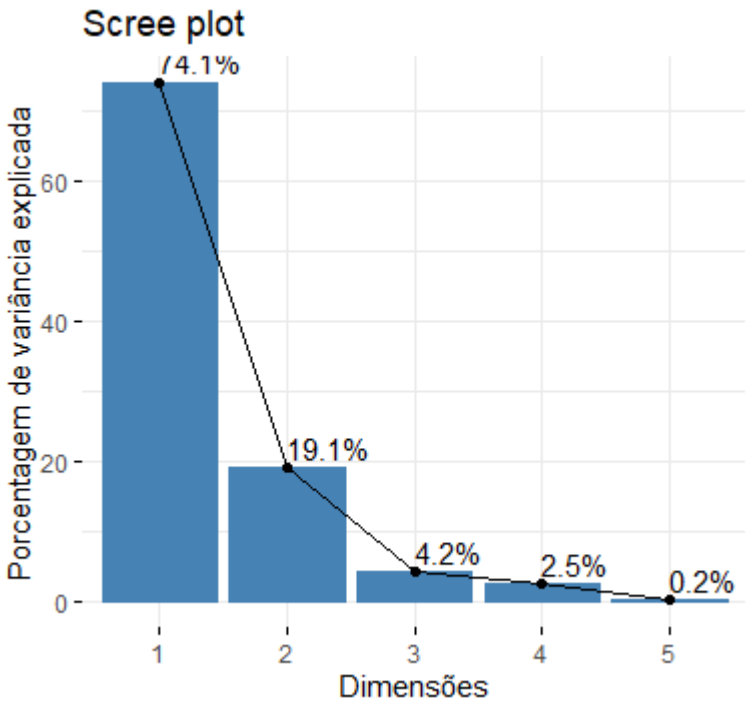

Figura 3.3: Gráfico do "Cotovelo" para os componentes dos dados.

Tabela 3.3.2: Medidas dos Componentes Principais.

Medida	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
Autovalores	3,029	1,291	0,573	0,095	0,012
Proporção de variação	0,6058	0,2582	0,1145	0,0191	0,0024
Variação acumulada	0,6058	0,8640	0,9785	0,9976	1,0000

Tabela 3.3.3: Cargas dos Componentes Principais.

Variável	Comp.1	Comp.2	Comp.3	Comp.4	Comp.5
População	$\mathbf{0,5584}$	0,1314	0,0079	0,5506	0,6065
Escolaridade	0,3133	$\mathbf{0,6289}$	$-0,5490$	$-0,4527$	$-0,0066$
Empregados	$\mathbf{0,5683}$	0,0043	0,1173	0,2681	$-0,7690$
Empregados na saúde	$\mathbf{0,4866}$	$-0,3096$	0,4549	$-0,6480$	0,2013
Valor da residência	$-0,1743$	$\mathbf{0,7010}$	0,6912	0,0151	$-0,0142$

No exemplo apresentado, com base no *Scree plot* mostrado na Figura 3.3, a sugestão é escolher 2 componentes principais, o que é razoável considerando o número de variáveis presentes e a proporção das variâncias explicadas somam 0,86. Além dos autovalores dos dois primeiros componentes serem maiores do que 1, a proporção de variância explicada desses dois componentes juntos é maior que 0.8. Além disso, observamos pela Tabela 3.3.3 que no Componente 1 as variáveis com maiores cargas são: População, Empregados e Empregados na saúde, enquanto que no Componente 2 as maiores cargas são protagonizadas pela Escolaridade e pelo Valor da Residência. Dessa forma, podemos nomear os dois componentes da seguinte forma:

- **Componente 1 = V1 =** População e cidadãos com emprego na cidade de Madison.
- **Componente 2 = V2 =** Poder de aquisição.

No entanto, a escolha do número de componentes principais não possui uma resposta definitiva e depende de diversos fatores. É importante considerar a proporção da variância total explicada, os tamanhos relativos dos autovalores e as interpretações dos componentes na análise específica. Além disso, componentes com autovalores pequenos podem indicar dependências lineares inesperadas nos dados, requerendo investigações adicionais.

O *scree plot* é uma valiosa ferramenta para determinar o número apropriado de componentes, mas é essencial combinar essa análise visual com o conhecimento do domínio e dos objetivos da pesquisa. A decisão pode envolver *trade-offs* entre a simplicidade do modelo e a retenção de informações importantes dos dados.

Capítulo 4

Aplicação

Com o intuito de unir as duas técnicas estatísticas estudadas ao longo do trabalho, utilizamos o banco de dados *Connectionist Bench (Sonar, Mines vs. Rocks)* disponibilizada em Sejnowski e Gorman que contém 111 observações obtidas ao refletir sinais de sonar em um cilindro de metal em vários ângulos e sob diferentes condições e 97 observações obtidas de rochas sob condições semelhantes. O sonar é um sistema acústico que utiliza ondas sonoras para detecção, localização e identificação de objetos subaquáticos ou submarinos. No contexto deste conjunto de dados, sinais de sonar foram enviados em direção a um cilindro de metal e rochas, e os padrões resultantes foram registrados.

A aplicação se divide em três etapas após a análise descritiva, sendo a primeira a aplicação da regressão logística, a segunda a união da ACP com a própria regressão logística e, por fim, comparativos e conclusões.

4.1 Análise dos dados

Cada padrão é um conjunto de 60 números na faixa de 0,0 a 1,0. Cada número representa a energia dentro de uma banda de frequência específica, integrada ao longo de um determinado período de tempo. A abertura de integração para frequências mais altas ocorre mais tarde no tempo, uma vez que essas frequências são transmitidas mais tarde durante o *chirp* (sinal acústico ou de radar cuja frequência varia continuamente ao longo do tempo).

O rótulo associado a cada registro contém a letra "R" se o objeto for uma rocha e "M" se for uma mina (cilindro de metal). Os números nos rótulos estão em ordem crescente de ângulo de visão, mas não codificam o ângulo diretamente, como mostrado na Tabela

4.1.1.

Tabela 4.1.1: Tabela do conjunto de dados *Connectionist Bench*.

ID	**Tipo**	**V1**	**V2**	...	**V59**	**V60**
1	R	0,020	0,037		0,009	0,003
2	R	0,045	0,052	...	0,005	0,004
...	...	...	...	...	...	...
207	M	0,030	0,035	...	0,003	0,004
208	M	0,052	0,043	...	0,007	0,003

Dessa forma, as 60 energias diferentes (V1, V2, ..., V60) foram utilizadas como covariáveis e o tipo ("R"ou "M") como variável resposta.

Com o intuito de analisar a proporção da variável resposta, plotamos o gráfico de barras representado na Figura 4.1.

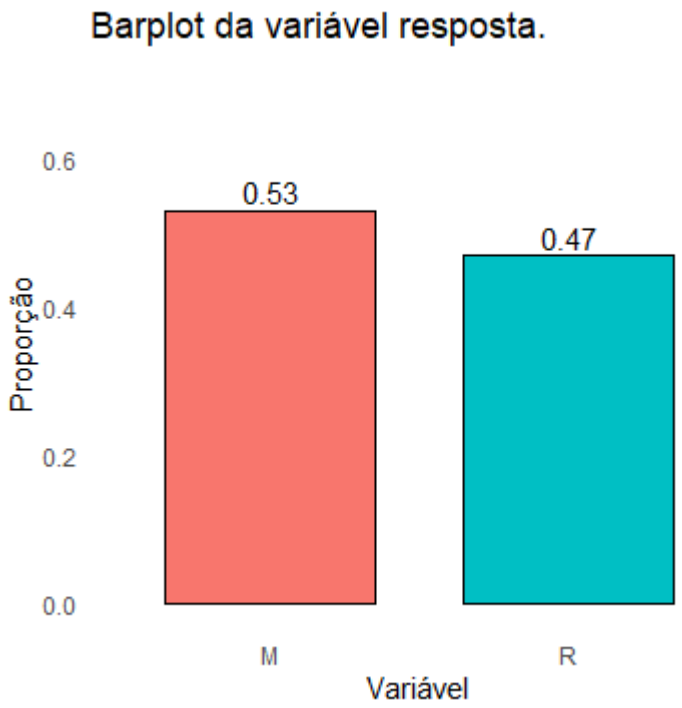

Figura 4.1: Gráfico de barras das variáveis "rocha" e "metal".

De acordo com a Figura 4.1, a proporção da variável resposta se encontra bem distribuída, ou seja, os dados estão balanceados.

Além de analisar a variável resposta, construimos *box-plots* para todas as energias conforme é mostrado na Figura 4.2.

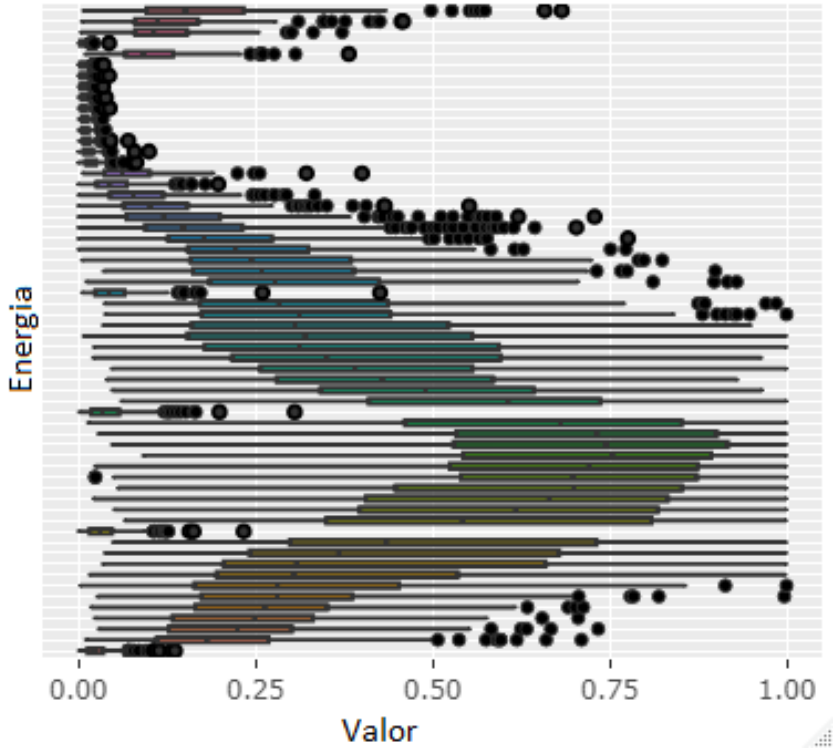

Figura 4.2: *Boxplots* para todas as covariáveis.

Pelo gráfico de *boxplots* das covariáveis ordenadas de forma que o primeiro *boxplot* é a energia V1 e o último a energia V60, há um indicativo com base na amostra observada de que cada covariável tem uma forte associação com seus vizinhos, uma vez que a amplitude das caixas e suas médias apresentam uma certa tendência. Para identificar se realmente há essa correlação, construimos a Figura 4.3.

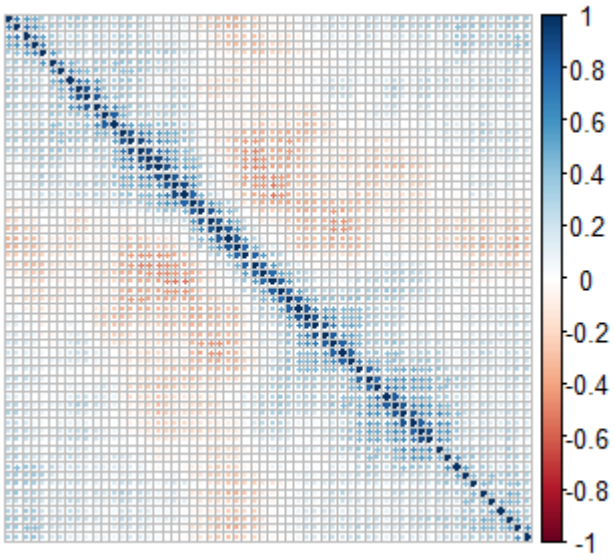

Figura 4.3: Gráfico de correlações das covariáveis.

Pelo gráfico de correlação (4.3) das energias, conseguimos confirmar o indicativo de que cada energia possui uma correlação forte com seus vizinhos devido a intensidade das cores dos blocos próximos a diagonal principal.

Com a análise descritiva feita, seguimos para as aplicações da regressão logística e em seguida da ACP juntamente com a regressão logística.

4.2 Separação dos dados

Para iniciar o ajuste separamos os dados em dois conjuntos, um de treino e outro de teste, sendo as proporções iguais a 80% e 20%, respectivamente. Calculando as frequência relativas da variável resposta em cada conjunto obtemos as tabelas 4.2.1 e 4.2.2 que nos mostram uma proprorção semelhante ao banco de dados completo.

Tabela 4.2.1: Tabela da proporção da variável resposta para o conjunto de treino.

Metal	**Rocha**
0,531	0,469

Tabela 4.2.2: Tabela da proporção da variável resposta para o conjunto de teste.

Metal	**Rocha**
0,535	0,465

4.3 Regressão logística

Antes de iniciar o ajuste do modelo é necessário lidar com o fato das variáveis serem altamente correlacionadas com seus vizinhos próximos, pois a multicolinearidade pode causar instabilidade nos coeficientes, redução na eficiência do modelo, problema na seleção de variáveis, entre outros. A solução proposta para esse problema foi selecionar variáveis quatro a quatro a partir de V4, ou seja, nessa seção trabalhamos com o conjunto de dados contento 15 variáveis: V1, V4, V8, V12, V16, V20, V24, V28, V32, V36, V40, V44, V48, V52, V56 e V60, conforme a Figura 4.4.

De acordo com a Figura 4.4, é possível notar que a correlação entre as variáveis é aceitável para iniciar o ajuste do modelo, uma vez que 0,5 é a maior correlação presente.

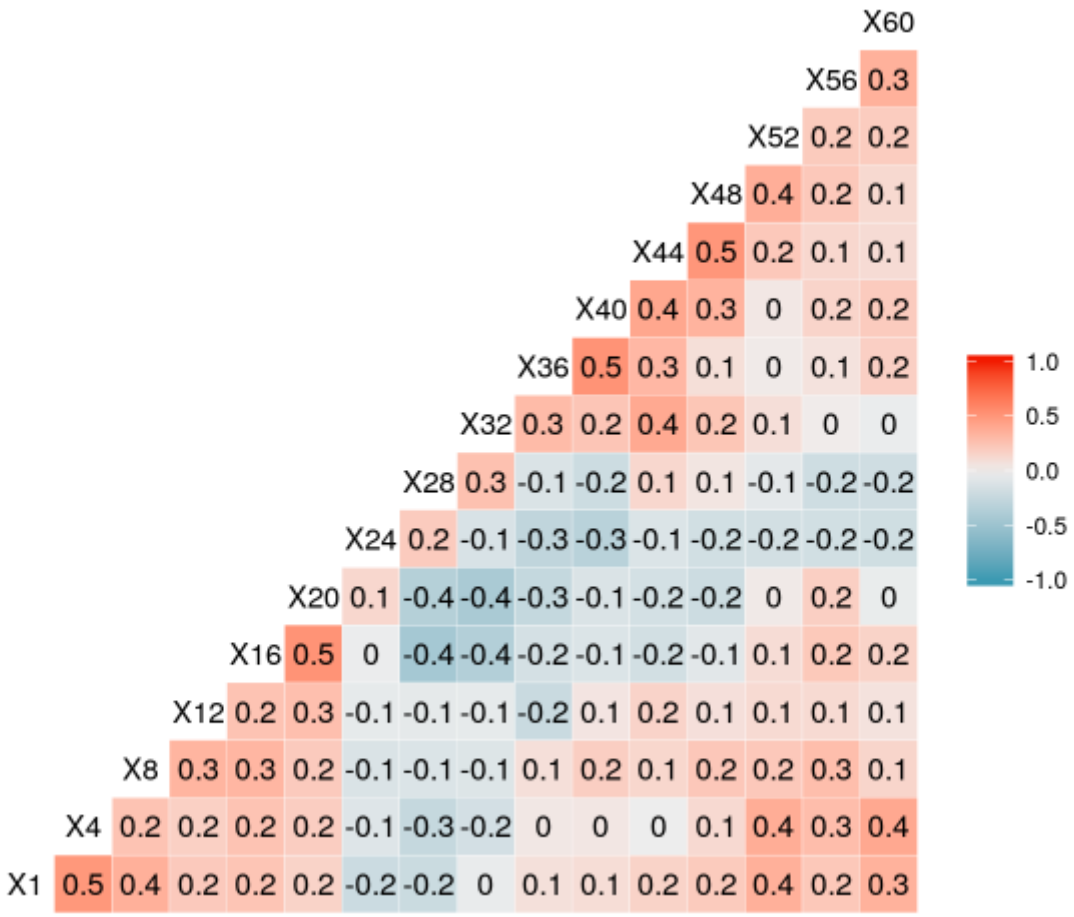

Figura 4.4: Gráfico de correlação das variáveis selecionadas.

4.3.1 Ajuste do modelo

Com os dados separados, foi feito o ajuste do modelo no conjunto de teste no *software* R e os resultados presentes na Tabela 4.3.1

indicam que com um nível se significância de 5% as covariáveis que são relevantes para o modelo além do intercepto são as energias associadas a V1, V4, V12, V16, V20, V36, V44 e V48. Portanto, o modelo final com a variável resposta"Rocha"como classe positiva é dado por:

$$\hat{\pi}(\mathbf{v_i}) = \frac{\exp\left(7,97 - -40,55V_1 - 20,95V_4 - 6,61V_{12} + 3,57V_{16} - 4,32V_{20} + 5,24V_{36}, -11,68V_{44} - 22,58V_{48}\right)}{1 + \exp\left(7,97 - -40,55V_1 - 20,95V_4 - 6,61V_{12} + 3,57V_{16} - 4,32V_{20} + 5,24V_{36}, -11,68V_{44} - 22,58V_{48}\right).} \quad (4.1)$$

Assim, podemos interpretar os coeficientes estimados da seguinte forma:

- **Intercepto 7,97:** representando a estimativa da razão de chances da nova observação ser classificada como "Rocha"quando todas as covariáveis são zero.

Tabela 4.3.1: Tabela com os coeficientes estimados e p-valor das covariáveis do ajuste do modelo no conjunto de treino.

Coeficientes	**Estimativa**	**p-valor**
Intercepto	**7,97**	**0,001**
V1	**-40,55**	**0,008**
V4	**-20,95**	**0,044**
V8	0,62	0,866
V12	**-6,61**	**0,004**
V16	**3,57**	**0,021**
V20	**-4,32**	**0,006**
V24	-2,13	0,074
V28	-1,24	0,323
V32	1,66	0,293
V36	**5,24**	**0,001**
V40	3,24	0,131
V44	**-11,68**	**0,001**
V48	**-22,68**	**0,001**
V52	-64,68	0,060
V56	95,83	0,084
V60	-32,23	0,623

- $-40,55V_1$**:** Mantendo as outras variáveis constantes, o aumento em uma unidade de V_1 diminui a razão de chances, em média, da nova observação ser classificada como "Rocha"em 40,55.

- $-20,95,55V_4$**:** Mantendo as outras variáveis constantes, o aumento em uma unidade de V_4 diminui a razão de chances, em média, da nova observação ser classificada como "Rocha"em 20,95.

A interpretação dos demais coeficientes estimados seguem de forma análoga.

4.3.2 Classificação e avaliação do modelo

Para efetuar a classificação do modelo ajustado usamos o conjunto teste e um ponto de corte igual a 0,5 e para realizar a avaliação da capacidade de classificação usamos uma matriz de confusão. Seguindo dessa forma, chegamos aos respectivos resultados:

Tabela 4.3.2: Tabela de confundimento para o modelo ajustado de regressão logística.

	Referência	
Predição	**M**	**R**
M	$\frac{15}{41}$	$\frac{6}{41}$
R	$\frac{7}{41}$	$\frac{13}{41}$

Através da Tabela 4.3.2 conseguimos fazer as seguintes conclusões a respeito do modelo final:

- **Acurácia:** Cerca de 68,29% das classificações feitas pelo modelo estão corretas em relação a ambas as classes.
- **Sensibilidade:** O modelo é capaz de identificar corretamente cerca de 71,43% das amostras de "Rocha".
- **Especificidade:** O modelo acerta em torno de 65% das vezes quando se trata de classificar corretamente as amostras de "Metal".
- **VPP:** Cerca de 68,18% das amostras classificadas como "Rocha"pelo modelo são realmente "Rocha".
- **VPN:** Cerca de 68,42% das amostras classificadas como "Metal"pelo modelo são realmente "Metal".

4.4 ACP e regressão logística

Nesta seção, vamos unir as duas técnicas estatísticas estudadas ao longo do trabalho. Primeiramente, aplicando ACP nos dados e em seguida usando os componentes resultantes como variáveis.

4.4.1 Análise de componentes principais

Para a escolha do número de componentes principais, inicialmente foi construido um gráfico de cotovelo (Figura 4.5) no *software* R.

Pela Figura 4.5, percebe-se que há a presença do cotovelo no componente três. No entanto, para termos mais *insights* a respeito da escolha do número de componentes

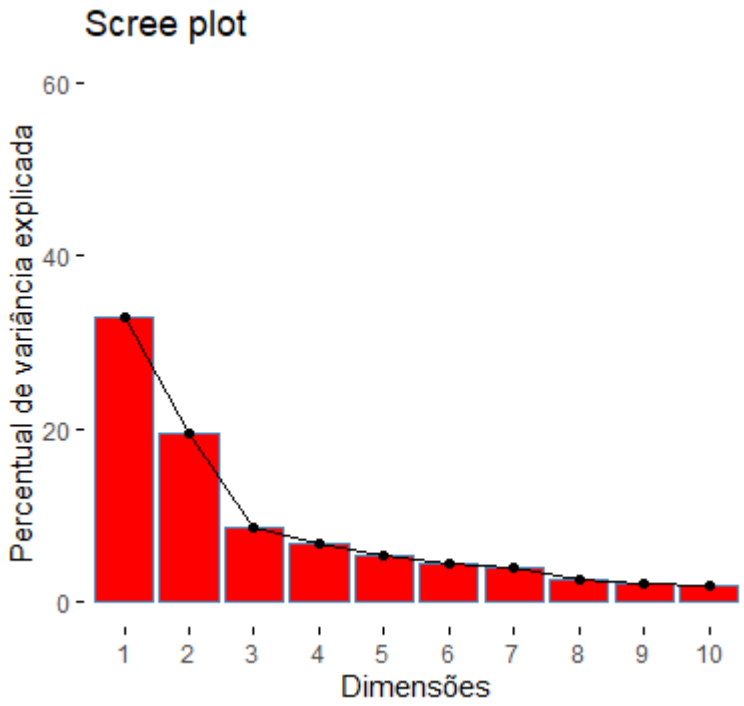

Figura 4.5: Gráfico do cotovelo dos componentes do banco de dados de treino.

também analisamos o percentual de variabildiade explicada acumulada por cada um até o componente 10.

Pela Tabela 4.1.1, percebe-se que os três primeiros componentes explicam cerca de 60% da variabilidade total dos dados, mas para dar continuidade ao trabalho, optamos por escolher um número de componentes que explique pelo menos 80%. Dessa forma, o ajuste do modelo foi feito com os sete primeiros componentes principais que representam $81,3\%$ da variabilidade total dos dados.

Tabela 4.4.1: Tabela do percentual de variabildiade explicada acumulada para os 10 primeiros componentes.

Componente	**Variabilidade explicada acumulada**
CP1	0,328
CP2	0,522
CP3	0,609
CP4	0,675
CP5	0,728
CP6	0,773
CP7	0,813
CP8	0,838
CP9	0,859
CP10	0,879

A escolha do compoenentes é realizada usando a cargas fatorias (Tabela 4.4.2).

Tabela 4.4.2: Cargas fatorias das 30 primeiras variáveis para os sete primeiros componentes principais.

	CP1	CP2	CP3	CP4	CP5	CP6	CP7
V1	0,00	0,02	-0,01	0,01	-0,01	0,00	-0,00
V2	0,00	0,02	-0,02	0,02	-0,01	-0,01	-0,01
V3	0,01	0,02	-0,02	0,02	-0,02	-0,03	-0,01
V4	0,01	0,03	-0,01	0,01	-0,02	-0,02	-0,01
V5	0,02	0,03	-0,01	0,03	-0,03	-0,03	-0,02
V6	0,02	0,02	-0,00	0,05	0,00	-0,04	-0,03
V7	0,02	0,02	-0,01	0,05	-0,00	-0,04	-0,05
V8	0,01	0,04	-0,02	0,09	-0,04	0,00	-0,09
V9	0,01	0,04	-0,03	0,16	-0,04	0,10	-0,14
V10	0,02	0,05	-0,03	0,18	-0,07	0,10	-0,17
V11	0,05	0,05	-0,05	0,19	-0,06	0,06	-0,14
V12	0,05	0,04	-0,07	0,18	-0,06	0,02	-0,08
V13	0,08	0,05	-0,09	0,20	-0,03	-0,05	-0,07
V14	0,13	0,09	-0,11	0,18	0,02	-0,13	-0,11
V15	0,17	0,14	-0,13	0,17	0,11	-0,21	-0,12
V16	0,21	0,19	-0,07	0,18	0,20	-0,25	-0,11
V17	0,25	0,19	0,04	0,17	0,30	-0,26	-0,05
V18	0,27	0,17	0,10	0,08	0,26	-0,08	0,03
V19	0,28	0,14	0,10	0,09	0,12	0,18	0,11
V20	0,27	0,11	0,12	0,10	-0,01	0,43	0,09
V21	0,27	0,05	0,19	0,10	-0,09	0,40	0,08
V22	0,26	-0,08	0,28	0,09	-0,10	0,14	0,17
V23	0,20	-0,18	0,30	0,02	-0,07	-0,11	0,22
V24	0,13	-0,24	0,28	0,00	-0,03	-0,22	0,10
V25	0,08	-0,29	0,29	0,05	-0,03	-0,24	-0,04
V26	0,03	-0,32	0,23	0,10	-0,06	-0,10	-0,21
V27	-0,02	-0,34	0,13	0,17	0,03	0,02	-0,32
V28	-0,11	-0,29	-0,00	0,24	0,15	0,16	-0,24
V29	-0,16	-0,23	-0,10	0,22	0,27	0,21	-0,03
V30	-0,18	-0,10	-0,08	0,27	0,28	0,17	0,10

Analisando carga por carga em cada componente, é possível associar as 60 energias mais relevantes em cada componente e, se for necessário, nomear de acordo com essas variáveis:

- **CP1:** V18, V19.
- **CP2:** V1, V4, V26, V27, V28.
- **CP3:** V22, V23, V24, V25, V35, V36.
- **CP4:** V6, V9, V10, V11, V12, V13, V14, V31, V45.
- **CP5:** V29, V30, V40, V41, V42, V43, V46.

Tabela 4.4.3: Cargas fatorias das 30 últimas variáveis para os sete primeiros componentes principais.

	CP1	CP2	CP3	CP4	CP5	CP6	CP7
V31	-0,18	-0,04	-0,04	0,26	0,25	-0,06	0,25
V32	-0,18	0,00	0,08	0,21	0,15	-0,06	0,30
V33	-0,19	0,05	0,13	0,06	0,11	0,04	0,27
V34	-0,19	0,13	0,21	-0,02	0,17	0,08	0,19
V35	-0,20	0,19	0,32	-0,04	0,20	-0,03	0,00
V36	-0,19	0,21	0,37	-0,07	0,12	-0,08	-0,21
V37	-0,15	0,21	0,31	-0,03	0,06	-0,02	-0,29
V38	-0,14	0,21	0,18	0,04	-0,05	0,13	-0,22
V39	-0,11	0,18	0,08	0,08	-0,18	0,18	-0,15
V40	-0,10	0,15	0,09	0,07	-0,21	0,02	-0,07
V41	-0,11	0,12	0,05	0,12	-0,22	-0,16	0,12
V42	-0,10	0,08	0,02	0,18	-0,25	-0,18	0,15
V43	-0,08	0,04	0,03	0,20	-0,22	-0,09	0,07
V44	-0,08	0,03	0,07	0,23	-0,17	-0,03	0,03
V45	-0,08	0,05	0,02	0,29	-0,22	-0,00	0,05
V46	-0,07	0,03	-0,02	0,22	-0,19	-0,04	0,11
V47	-0,04	0,01	-0,04	0,13	-0,11	-0,07	0,10
V48	-0,03	0,01	-0,03	0,09	-0,08	-0,04	0,03
V49	-0,01	0,01	-0,02	0,05	-0,04	-0,02	0,01
V50	-0,00	0,01	-0,00	0,01	-0,01	-0,00	0,01
V51	-0,00	0,00	-0,01	0,01	-0,01	-0,01	0,01
V52	-0,00	0,00	-0,00	0,01	-0,01	-0,00	0,01
V53	-0,00	0,00	-0,00	0,00	-0,00	-0,00	0,00
V54	0,00	0,00	0,00	0,00	-0,00	-0,00	0,00
V55	0,00	0,01	-0,00	0,00	-0,00	-0,00	0,00
V56	0,00	0,00	0,00	0,00	-0,00	-0,00	0,00
V57	0,00	0,00	-0,00	0,00	-0,00	-0,00	-0,00
V58	0,00	0,00	-0,00	0,00	-0,00	-0,00	0,00
V59	-0,00	0,00	-0,00	0,00	-0,00	-0,00	0,00
V60	-0,00	0,00	-0,00	0,00	-0,00	-0,00	-0,00

- **CP6:** V3, V15, V16, V17, V20, V21, V32, V33, V44.

- **CP7:** V8, V37, V38.

As cargas dos componentes nos ajudam a interpretar melhor como cada variável o influencia. Podemos interpretar, por exemplo, as qautro primeiras cargas mais influentes do primeiro componente principal (CP1):

- **V17 (0,25):** A variável V17 tem uma carga positiva significativa no CP1, indicando que valores mais altos de V17 estão associados a pontuações mais altas no CP1.

- **V18 (0,27):** V18 contribui positivamente para o CP1, e valores mais altos dessa variável estão associados a pontuações mais altas no CP1.
- **V19 (0,28):** V19 também contribui positivamente para o CP1, e valores mais altos de V19 estão associados a pontuações mais altas no CP1.
- **V20 (0,27):** A variável V20 tem uma carga positiva no CP1, indicando que valores mais altos de V20 estão associados a pontuações mais altas no CP1.

4.4.2 Ajuste do modelo de regressão logística utilizando ACP

Assim como na seção anterior, utilizamos o comando "glm" para ajustar o modelo. Os resultados dos coeficientes estimados e do p-valor associado a cada um estão na Tabela 4.4.4.

Tabela 4.4.4: Coeficientes estimados de cada componente e seus p-valores associados.

Coeficiente	**Estimativa**	p**-valor**
Intercepto	-0,233	0,245
CP1	**-0,757**	**0,008**
CP2	0,453	0,179
CP3	1,126	0,026
CP4	**-3,850**	**6e-07**
CP5	**4,119**	**2e06**
CP6	**-1,600**	**0,028**
CP7	-0,347	0,646

Os valores presentes na Tabela 4.4.4 indicam que com um nível de significância de 5% as covariáveis CP1, CP3, CP4, CP5 e CP6 são relevantes para o modelo ajustado. Portanto, o modelo final com a variável resposta "Rocha"como classe positiva é dado por:

$$\hat{\pi}(\mathbf{v_i}) = \frac{\exp\left(-0,757\text{CP1} + 1,126\text{CP3} - 3,850\text{CP4} + 4,119\text{CP5} - 1,600\text{CP6}\right)}{1 + \exp\left(-0,757\text{CP1} + 1,126\text{CP3} - 3,850\text{CP4} + 4,119\text{CP5} - 1,600\text{CP6}\right)}. \tag{4.2}$$

Assim, podemos interpretar os coeficientes estimados de forma análoga ao primeiro componente da seguinte maneira:

- **CP1 -0,757:** Mantendo as outras componentes constantes, o aumento em uma unidade da **CP1** diminui a média da resposta em -0,757.

No entanto, CP1 é composto pelas cargas apresentadas nas tabelas 4.4.2 e 4.4.3, portanto para o primeiro compenente aumentar em uma unidade é necessário que, dada uma nova observação V

$$\text{v3} \times 0,01 + \text{v4} \times 0,01 + \text{v5} \times 0,02 + ... + \text{v49} \times 0,01 = 1. \tag{4.3}$$

ou seja, para CP1 aumentar em uma unidade é necessário que a combinação linear entre as cargas do primeiro componente e a nova observação seja igual a 1.

4.4.3 Classificação e avaliação do modelo.

Para realizar a classificação do modelo ajustado usamos o conjunto de teste e um ponto de corte igual a 0,5 e para efetetuar a availação da capacidade do modelo usamos uma matriz de confusão. Seguindo assim, chegamos a matriz dada por:

Tabela 4.4.5: Tabela de confundimento para o modelo ajustado de regressão logística com componentes principais.

	Referência	
Predição	**M**	**R**
M	$\frac{16}{41}$	$\frac{6}{41}$
R	$\frac{5}{41}$	$\frac{14}{41}$

Através da Tabela 4.4.5 conseguimos fazer as seguintes conclusões a respeito do modelo final:

- **Acurácia:** Cerca de 73,17% das classificações feitas pelo modelo estão corretas em relação a ambas as classes.
- **Sensibilidade:** O modelo é capaz de identificar corretamente cerca de 70% das amostras de "Rocha".
- **Especificidade:** O modelo acerta em torno de 76,19% das vezes ao classificar corretamente as amostras de "Metal".
- **Valor Preditivo Positivo (VPP):** Cerca de 73,68% das amostras classificadas como "Rocha"pelo modelo são realmente "Rocha".
- **Valor Preditivo Negativo (VPN):** Cerca de 72,73% das amostras classificadas como "Metal"pelo modelo são realmente "Metal".

Capítulo 5

Conclusão e considerações finais

Na parte inicial do trabalho estudamos as bases teóricas de duas metodologias estatísticas, sendo elas a regressão logística e a análise de componentes princiapis, além de exemplificar suas aplicações em bancos de dados distintos. A partir desses estudos, foi realizada a união dos dois métodos com o intuito de investigar tanto a interpretabilidade como a avaliação de um modelo de regressão logística com componentes principais ocupando o papel de covariável. Para tanto, selecionamos um banco de dados adequado, com alta dimensionalidade e correlação entre as variáveis, para primeiramente aplicar a regressão logística com as variáveis originais como covariáveis, posteriormente aplicar a regressão logística com os componentes principais como covariáveis e, por fim, efetuar conclusões e considerações finais a respeito do uso de componentes principais como uma forma de reduzir a dimensionalidade em uma aplicação de regressão logística em relação ao uso do banco de dados original.

No que diz respeito a aplicação regressão logística sem o uso de componentes principais, obtivemos acerto em 28 das 41 observações reservadas como teste, ou seja, $68,29\%$, além da interpretação dos coeficientes do modelo ajustado ser simples para o leitor.

Já a aplicação regressão logística com o uso de componentes apresentou um melhor desempenho, acertando 30 das 41 observações reservadas como teste, ou seja, $73,17\%$. Em relação a questão da interpretabilidade, há uma dificuldade maior em interpretar os coeficientes estimados, uma vez que cada componente principal pode carregar em si cargas de todas as variáveis do banco de dados.

É importante ressaltar que para a aplicação da análise de componentes principais para a redução de dimensionalidade em um banco de dados é necessário que este tenha variáveis correlacionadas, dessa forma selecionamos um *dataset* que atende a exigência.

No entanto, quando utilizamos bancos de dados com variáveis altamente correlacionadas para uma análise de regressão há a paresença de multicolinearidade, em que uma das maneiras de resolvar essa questão é eliminando algumas variáveis correlacionadas (como foi feito na aplicação), podendo resultar em um banco de dados com dimensão menor que o original. Outra observação importante é em relação a acurácia dos modelos, o modelo de regressão logística com componentes principais como covariáveis obteve uma porcentagem de acerto em quase 5% das classificações em relação ao outro modelo, no entanto, considerando o tamanho do banco de dados essa diferença não é muito expressiva, uma vez que o primeiro modelo acertou 38 classificações e o segundo 40.

Portanto, concluimos que para o banco de dados utilizado a união das duas metodologias estatísticas estudas no trabalho é válida. Em relação a regressão logística usual, a utilização de componentes principais em uma análise de regressão logística pode levar a um bom desempenho de classificação, mas há uma certa dificuldade em relação a interpretação dos coeficientes estimados do modelo final. No entanto, esse estudo requer mais aprofundamento, podendo ser aplicado em mais banco de dados e ser comparado com outras téncnicas de redução de dimensionalidade para obter conclusões mais fundamentadas a respeito da união das duas metodologias estudadas no trabalho.

Referências Bibliográficas

Hosmer Jr, D. W.; Lemeshow, S.; Sturdivant, R. X. *Applied logistic regression*, volume 398. John Wiley & Sons, 2013.

Hossin, M.; Sulaiman, M. N. A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process (IJDKP)*, v. 5(2), p. 1, 2015.

Izbicki, R.; dos Santos, T. M. *Aprendizado de máquina: uma abordagem estatística*. São Carlos, SP: Rafael Izbicki, 2020.

Johnson, R. A.; Wichern, D. W. *Applied multivariate statistical analysis*. Prentice Hall, 6^a ed., 2002.

Mood, A. M.; Boes, D. C.; Graybill, F. A. *Introduction to the Theory of Statistics*. McGraw-Hill, 3^a ed., 1974.

Myers, R. H.; Montgomery, D. C.; Vining, G. G.; Robinson, T. J. *Generalized linear models: with applications in engineering and the sciences*. John Wiley & Sons, 2012.

R Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2023. URL https://www.R-project.org/.

Sejnowski, T.; Gorman, R. Connectionist Bench (Sonar, Mines vs. Rocks). UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5T01Q, jhgbd.