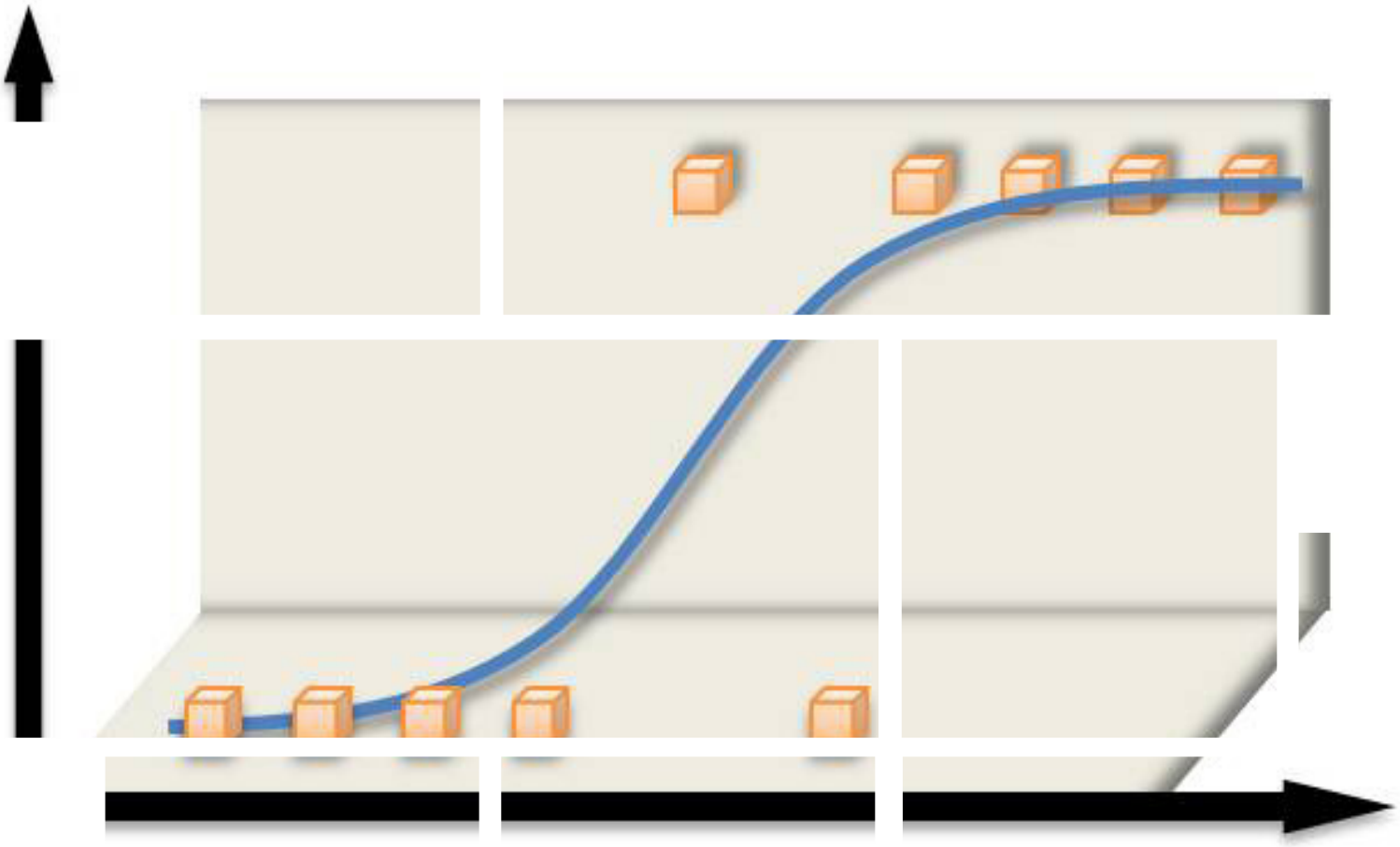
ISSN 1806-9193
Dezembro/2022

DOCUMENTOS

529

Aplicação da Regressão Logística em Dados Experimentais Utilizando o *Software* R



Embrapa

ISSN 1806-9193
Dezembro/2022

Empresa Brasileira de Pesquisa Agropecuária
Embrapa Clima Temperado
Ministério da Agricultura, Pecuária e Abastecimento

DOCUMENTOS 529

Aplicação da Regressão Logística em Dados Experimentais Utilizando o *Software* R

Ricardo Alexandre Valgas

Embrapa Clima Temperado
Pelotas, RS
2022

Embrapa Clima Temperado
BR-392, km 78, Caixa Postal 403
CEP 96010-971, Pelotas, RS
Fone: (53) 3275-8100
www.embrapa.br/clima-temperado
www.embrapa.br/fale-conosco

Revisão de texto
Bárbara Chevallier Cosenza

Normalização bibliográfica
Marilaine Schaun Pelufê

Editoração eletrônica
Nathália Santos Fick (46.431.873/0001-50)

Foto de capa
Ricardo Valgas

1ª edição
Publicação digital: PDF

Dados Internacionais de Catalogação na Publicação (CIP)
Embrapa Clima Temperado

V169a Valgas, Ricardo Alexandre
Aplicação da Regressão Logística em dados experimentais utilizando o software R / Ricardo Alexandre Valgas. - Pelotas: Embrapa Clima Temperado, 2022.
29 p. (Documentos / Embrapa Clima Temperado, ISSN 1806-9193 ; 529).

1. Estatística. 2. Estatística agrícola. 3. Análise estatística. 4. Método estatístico. I. Valgas, Ricardo Alexandre. II. Série.

CDD 519.5

Marilaine Schaun Pelufê (CRB10/1274)

Autores

Ricardo Alexandre Valgas
Estatístico, mestre em Métodos Numéricos em Engenharia, pesquisador da Embrapa Clima Temperado, Pelotas, RS

Apresentação

A utilização de técnicas de estatística experimental vem sendo amplamente utilizada nos diversos campos da ciência. Há alguns anos, o conceito de experimentação era algo relativamente novo, porém passou a ganhar espaço cada vez maior nos setores produtivos, devido aos bons resultados obtidos na avaliação de experimentos de pesquisas, área em que seus princípios e prática já estavam mais difundidos.

Com o avanço da informática, a análise estatística foi se tornando uma ferramenta cada vez mais comum para comprovar as hipóteses formuladas nos projetos de pesquisas. Nessa linha, o programa estatístico **R** foi ocupando lugar de destaque pelo seu uso crescente no meio agronômico. Atualmente, é um dos programas de análise de dados mais difundido no meio científico.

Buscando motivar o aprendizado e a aplicação da regressão logística no contexto da pesquisa agropecuária, este trabalho foi elaborado no **Rstudio**, uma das interfaces disponíveis do *software* R. Portanto, com esta publicação, espera-se possibilitar a qualificação técnica de pessoas com interesse nessa área, em especial pesquisadores e estudantes.

Roberto Pedroso de Oliveira
Chefe-Geral
Embrapa Clima Temperado

Sumário

Regressão logística: introdução

A técnica de regressão logística foi desenvolvida por volta da década de 1960, como proposta para realizar predições ou explicar determinados fenômenos nos quais a variável de interesse tivesse apenas dois resultados possíveis. O estudo pioneiro da aplicação da técnica foi entitulado *Framingham Heart Study*, realizado em cooperação com a Universidade de Boston, que tinha como principal objetivo identificar e modelar os principais fatores que desencadeavam doenças cardiovasculares em um grupo de 5.209 pessoas variando entre 30 e 60 anos de idade (Fávero et al., 2009). Vários fatores de risco foram modelados e identificados pela técnica de regressão logística, entre eles tabagismo, sedentarismo, obesidade, diabetes, hipertensão arterial e altas taxas de colesterol.

Quando se fala em modelos de regressão logística, remete-se aos conceitos básicos de modelagem e mineração de dados. Minerar dados nada mais é do que utilizar técnicas para prever padrões em um conjunto de dados que não poderiam ser observados de forma simples.

Então, a mineração de dados inclui, entre suas várias tarefas, prever o valor de um atributo com base nos valores de outros atributos. Para isso, são construídos modelos relacionados com a variável resposta ou dependente, em função de variáveis independentes ou regressoras.

A regressão logística é uma das técnicas que faz uma modelagem de previsão, a partir de uma variável de interesse do tipo categórica (geralmente binária). A partir do modelo ajustado, é possível calcular a probabilidade de um evento ocorrer, para uma observação aleatória.

Em outras palavras, o modelo de regressão logístico permite:

1) modelar a probabilidade de um dos eventos da variável resposta ocorrer em função das variáveis independentes;

2) estimar a probabilidade de um dos eventos da variável resposta ocorrer, para uma observação selecionada, contra a probabilidade desse evento não ocorrer. Isso é chamado de *odds ratio* ou razão de chance, valor muito explorado nesse tipo de modelagem;

3) prever a contribuição de cada variável regressora no modelo estimado.

Embora o modelo de regressão logística seja não linear, é possível linearizá-lo fazendo uma transformação na variável resposta chamada *logit* (ou logito), a qual é determinada na função de ligação no momento do ajuste. Com a linearização da variável resposta, obtém-se uma interpretação mais simples e direta das estimativas dos parâmetros, considerando-se a *odds ratio*.

Regressão logística e modelos lineares generalizados

A característica básica e necessária para que a regressão logística binária possa ser aplicada é: “explorar um conjunto de dados no qual a variável de interesse seja binária, ou seja, com apenas dois resultados possíveis”. Por exemplo: a planta está doente ou não, ocorre ou não o parasitismo, o animal prefere ou não prefere a nova dieta, o solo está apto ou não para o plantio, entre outras situações análogas. Em todas elas, a regressão logística é a metodologia mais robusta para interpretar a variável binária (dependente) em função de um conjunto de variáveis preditoras (independentes).

Voltando aos modelos de regressão clássicos, na regressão linear simples, o método de mínimos quadrados ordinários (MQO) é o mais utilizado para promover a melhor estimativa não viesada dos parâmetros populacionais. Ou seja, os parâmetros estimados por MQO apresentam a menor variação entre todas as estimativas possíveis.

É claro que, na sua aplicação, alguns pressupostos devem ser atendidos, como a normalidade dos resíduos, a homocedasticidade (igualdade de variâncias) e a linearidade dos dados. No entanto, quando se trata de uma variável resposta do tipo binária, geralmente esses pressupostos não são atendidos, e as estimativas por mínimos quadrados já não se torna a mais apropriada. Nesse contexto, a regressão logística torna-se a mais adequada para produzir estimativas mais consistentes dos parâmetros do modelo.

Inicialmente, é preciso relembrar as características dos modelos de regressão de forma ampla. Dado o modelo clássico de regressão linear simples:

$$Y = \beta_0 + \beta_1 X + \epsilon$$

Sabe-se que Y corresponde à variável a ser estimada (dependente), X representa a variável informada (independente), β_0 o intercepto (valor de Y quando X assume zero) e β_1 o coeficiente de regressão, o qual representa a variação de Y quando se aumenta uma unidade em X. Já o parâmetro estocástico representa o erro do modelo.

Assim, pode-se dizer que a regressão logística é um caso particular dos modelos lineares generalizados (GLM) nas situações em que a variável resposta é dicotômica, ou seja, 0 ou 1. Nesse caso, o espaço paramétrico da probabilidade predito pelo modelo deve ser o mesmo intervalo de Y.

Isso torna a interpretação da função logística mais fácil, pois tem-se o valor estimado da probabilidade diretamente: à medida que os valores de X aumentam, a probabilidade se aproxima de 1, mas quando os valores de X diminuem, a probabilidade se aproxima de 0.

A regressão logística utiliza a curva logística para apresentar a relação entre a variável dependente e as variáveis independentes. Nessa curva (ou função), os valores preditos permanecem no intervalo entre 0 e 1, sendo definida pelos coeficientes do modelo que são estimados.

Regressão logística binária

Considere a situação em que o modelo de regressão logística será representado por uma variável resposta Y, com dois resultados possíveis, assumindo uma distribuição de Bernoulli em que Y = 1 representa o sucesso (evento de interesse) e Y = 0 o fracasso, e terá apenas uma variável regressora independente X. Sendo *p* a probabilidade de sucesso, a probabilidade condicional *P(Y|X)* é dada por:

$$P(Y = 1|X) = p$$

$$P(Y = 0|X) = 1 - p$$

Nesse caso, representa a probabilidade de Y ocorrer condicionado ao valor de X, o que é algo definido na hipótese inicial do modelo. Portanto, a esperança matemática da variável de Bernoulli é dada por $E(Y|X)=1p+0(1-p)=p$.

Nesse momento, admite-se, por hipótese, que a variável binária Y possa ser modelada pela definição clássica apresentada anteriormente: $Y=\beta_0+\beta_1 X+\epsilon$. Assim, a esperança matemática é dada pela esperança da parte determinística, uma vez que se assume a hipótese de que $E(\epsilon)=0$:

$$E(Y|X) = E(\beta_0 + \beta_1 X) + E(\epsilon) = \beta_0 + \beta_1 X$$

resultando, portanto, na equação:

$$p = \beta_0 + \beta_1 X$$

Pela expressão acima, ocorre que o modelo, no formato apresentado, não é apropriado nos casos em que a variável X é contínua, porque, assim como X, *p* assumiria qualquer valor real, o que não é possível quando se trata de uma probabilidade, a qual deve ocorrer no intervalo $0 \leq p \leq 1$.

Portanto, o uso matemático e empírico de uma função flexível, de fácil interpretação, contínua e diferenciável, como a função logística, embasam e justificam sua aplicação.

Regressão logística binária múltipla

No contexto em que existam *k* variáveis preditoras $X_1, X_2, \ldots, X_k$ para estimar uma variável binária Y, tem-se um modelo de regressão logística binomial múltipla. O objetivo permanece o mesmo: estimar a probabilidade desconhecida *p*, mas agora através da combinação linear de *k* variáveis independentes.

Então, a única variável X no modelo logístico simples passa a ser representada por um vetor de variáveis preditoras, assim como *β* passa a ser representado por um vetor da forma:

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \ldots \\ \beta_k \end{bmatrix} X = \begin{bmatrix} 1 \\ X_1 \\ \ldots \\ X_k \end{bmatrix}$$

Mas como se faz a conexão das variáveis independentes com a variável dependente Y? Através de uma função, a qual combina linearmente as variáveis, e que pode retornar qualquer valor de uma distribuição de probabilidades de Bernoulli (domínio de 0 a 1). Essa razão de probabilidades é chamada de chance ou *odds* (em inglês), dada por:

$$odds = \frac{p}{1-p}$$

onde seu logaritmo natural (ou logit) é dado pela expressão a seguir, e a sua forma é apresentada na Figura 1:

$$\ln(odds) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

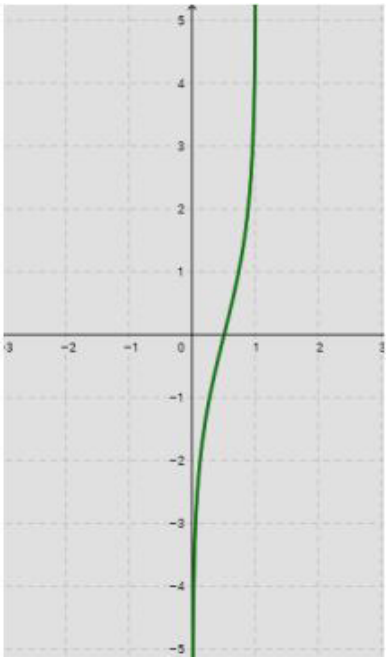


Figura 1. Gráfico da função *logit.*
Fonte: Gonzalez (2018).

Para obter a função de resposta do modelo logístico, calcula-se a inversa da função *logit* e assim obter as probabilidades adequadamente no eixo Y:

$$logit^{-1}(Z) = \frac{1}{1+e^{-(z)}} = \frac{e^{z}}{1+e^{(z)}} = \frac{e^{(\beta' X)}}{1+e^{(\beta' X)}}$$

sendo:

$$Z = \beta' X = \beta_0 + \beta_1 X_1 + \cdots + \beta_k X_k$$

a combinação linear das variáveis preditoras e seus coeficientes.

Segundo Gonzalez (2018), os gráficos da inversa da função *logit* (Figura 2) e da função *logit* são basicamente os mesmos, havendo apenas uma rotação de 90 graus e a troca das coordenadas x e y, tornando a função inversa, com domínio entre 0 e 1 no eixo Y.

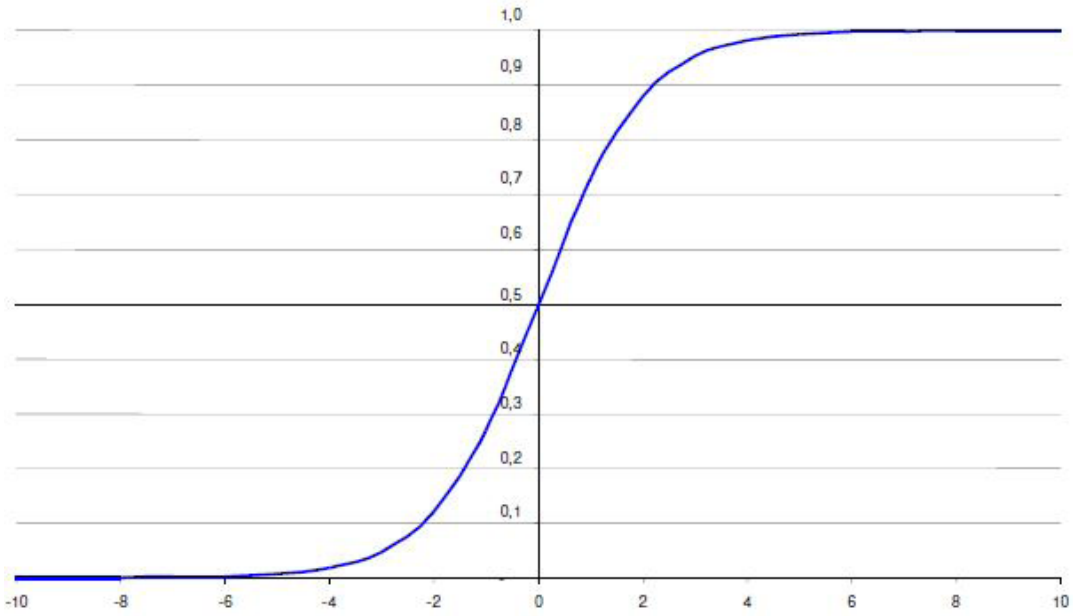


Figura 2. Gráfico da inversa da função *logit*.
Fonte: Gonzalez (2018).

Como visto anteriormente, o objetivo do modelo logístico é estimar o valor de *p*. Para isso parte-se da função *logit* (considerando uma única variável preditora X), resultando na equação:

$$ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k = \beta_0 + \sum \beta_k X_k$$

Os cálculos matemáticos são similares ao caso que contempla uma única variável regressora. Utiliza-se o antilogaritmo na equação acima para isolar *p*:

$$\frac{p}{1-p} = \mathrm{e}^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}$$

Assim, obtendo o modelo de regressão logístico múltiplo:

$$\hat{p} = \frac{e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}{1+e^{(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}} = \frac{1}{1+e^{-(\beta_0+\beta_1 X_1+\beta_2 X_2+\cdots+\beta_k X_k)}}$$

A equação acima é a equação de regressão desejada e utilizada para estimar o valor da probabilidade $\hat{p}$ com base nos coeficientes (estimados pelo método da máxima verossimilhança) e os valores apresentados da variável X.

Estimação dos coeficientes do modelo de regressão

Nessa etapa do processo de ajuste do modelo de regressão logística, é interessante saber como os parâmetros β são estimados. Utiliza-se um método estatístico que, a partir de dados amostrais, busca estimar os parâmetros de uma função, a função de verossimilhança, que apresentam a maior probabilidade dos dados da amostra ocorrerem. Em outras palavras, a estimação por máxima verossimilhança fornece os valores de $\hat{\beta}$ do modelo logístico que permite identificar a contribuição de cada variável regressora para que o evento de interesse ocorra.

Sendo $(X_1, X_2, \ldots, X_k, Y)$ uma amostra aleatória com k variáveis preditoras e $p_i = P(Y_i = 1 \mid X_i)$ e $1 - p_i = P(Y_i = 0 \mid X_i)$ as probabilidades teóricas. Para cada observação $y_i \in \{0,1\}$, tem-se a probabilidade condicional:

$$P(Y_i|X_i) = p_i^{y_i}(1 - p_i)^{1-y_i}$$

Assumindo independência amostral, a função de verossimilhança é dada pelo produto:

$$L(\beta) = \prod_{i=1}^{k} p_i^{y_i} (1 - p_i)^{1-y_i}$$

Daqui em diante, o procedimento é todo matemático. Passando o logaritmo natural, tem-se a função log-verossimilhança:

$$l(\beta) = \ln[L(\beta)] = \sum_{i=1}^{n} [y_i \ln p_i + (1 - y_i)\ln(1 - p_i)]$$

Para encontrar o valor do parâmetro β que maximiza ln[L(β)],, basta obter as derivadas parciais:

$$\frac{\partial[L(\beta)]}{\partial\beta_0} = \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} \hat{p}_i = 0$$

$$\frac{\partial[L(\beta)]}{\partial\beta_j} = \sum_{i=1}^{n} x_{ij}\, y_i - \sum_{i=1}^{n} x_{ij}\, \hat{p}_i = 0, j \in 1, \ldots, k$$

Logo, para encontrar a solução das equações acima, é preciso considerar que a primeira derivada do vetor de estimativas satisfaz $\partial L(\hat{\beta}) = 0$. Ainda, essas equações também são não lineares em seus parâmetros, exigindo, portanto, um método interativo para solucioná-las.

Tal procedimento requer um apoio computacional para que os cálculos possam ser feitos de forma rápida e segura, o que não era possível há algumas décadas. Hoje, aplicar o método interativo de Newton-Raphson, por exemplo, é fácil e acessível, permitindo que diversas simulações sejam executadas até se encontrar o conjunto de parâmetros que produzem a maior log-verossimilhança.

Exemplos de aplicação da regressão logística

Para começar a entender a relação existente entre os modelos lineares e logístico, considere a simulação da infestação de uma doença em plantas com a seguinte codificação: 1 para planta infestada e 0 para não infestada, de acordo com a quantidade de horas de exposição a um determinado contaminante.

Nesse caso, a variável dependente é infestação e variável independente é o tempo de exposição. Na Figura 3, pode-se observar que os únicos valores da variável resposta são 0 e 1. Também, à medida que o tempo de exposição aumenta, a quantidade de plantas infestadas também aumenta. Fica notável que não se pode utilizar uma regressão linear simples para modelar a variável resposta, porque certamente os pressupostos do modelo serão violados.

```
x<-c(seq(0,24,1))
y<-c(0,0,0,0,0,1,0,1,0,1,1,1,0,1,1,1,1,1,1,1,0,1,1,1,1)
plot(x,y,xlab=(“Tempo de exposição”),ylab=(“Manifestação da doença”),
main=(“”))
```

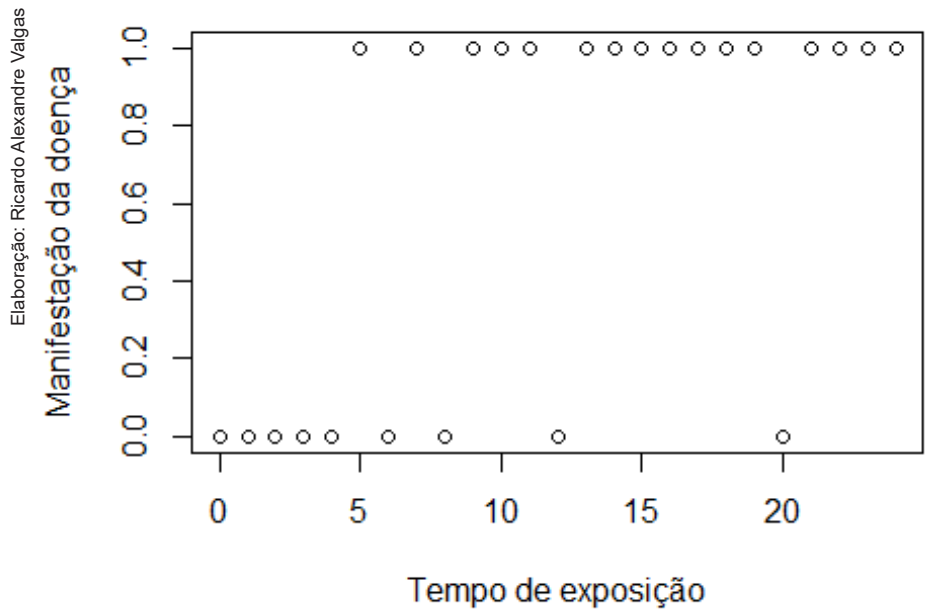


Figura 3. Simulação de dados da infestação de uma doença em plantas.

O valor médio do tempo de exposição é facilmente calculado resultando em 9,76 horas e pode ser utilizado como um valor de referência para estabelecer o comportamento dos dados. Plantas com horas de exposição acima da média apresentam uma maior incidência da doença do que plantas expostas a um número de horas menor que a média. Em termos de probabilidade, a chance de se observar uma planta infestada aumenta para uma exposição acima da média, enquanto que o inverso ocorre para valores abaixo da média.

É nesse sentido que a regressão logística busca determinar a probabilidade de ocorrência do evento de interesse, ou seja, quando Y assume o valor 1, e também estabelece a correlação existente entre as variáveis X e Y.

Tendo em mente essa lógica, juntamente com a parte teórica descrita anteriormente, é possível determinar quais são as etapas que se deve seguir para poder utilizar o modelo de regressão logística em uma análise de dados.

A seguir serão apresentados três exemplos da aplicação da regressão logística (simulações 1, 2 e 3). Nos dois primeiros, o modelo não é bem ajustado por motivos que podem ocorrer com mais frequência. No último exemplo, o modelo logístico é bem ajustado e o método é explorado para mostrar sua aplicação prática.

Exemplo prático 1: presença de multicolinearidade

Até aqui foram descritas as características teóricas sobre o modelo de regressão logística. Mesmo que esse não seja o principal objetivo do trabalho, não seria adequado abordar diretamente sua aplicação, pois o conhecimento teórico da técnica auxilia a implementação computacional, além de facilitar o entendimento dos cálculos e estimativas obtidas ao longo da sua aplicação.

A partir de agora, o programa utilizado será o **R** (R Core Team, 2018) através do *Rstudio* e todo *script* será apresentado para mostrar os procedimentos aplicados.

Considere um conjunto fictício de dados de 18 empresas ligadas ao agronegócio, no qual são apresentadas as variáveis: faturamento (milhões de R$), número de empregados, número de projetos agrícolas vigentes e número de projetos sociais. Após uma auditoria externa, as empresas foram classificadas quanto à necessidade de contratação de seguro: 1 para necessita e 0 para não necessita.

Planilha de dados 1 - Dados simulados do exemplo 1.

```
# Lendo o arquivo de dados
dados<-read.table("dados_ex1.txt",header=TRUE)
dados
```

```
##    Empresa Faturamento Empregados Projetos Social Seguro
## 1        1         289      15186      297      3      0
## 2        2         275      14711      200      2      0
## 3        3         258      13945      184      2      0
## 4        4         199      10263      203      1      0
## 5        5         170       9756      122      1      0
## 6        6         149       7929      165      1      0
## 7        7         132       6541      102      1      0
## 8        8         125       7175      154      1      0
## 9        9         106       5034      104      1      0
## 10      10         105       4987      107      1      0
## 11      11         101       4329      111      1      0
## 12      12         100       4581       98      0      1
## 13      13          99       3946       75      1      1
## 14      14          97       3048       74      0      1
## 15      15          84       2489       63      0      1
## 16      16          50       1945       48      0      1
## 17      17          19        486       26      0      1
## 18      18          16        311       20      0      1
```

Observe que a variável de interesse **Seguro** está na última coluna do arquivo e apresenta os valores 0 e 1 para cada empresa. Aqui cabe algumas considerações importantes:

1) o modelo logístico é obtido a partir das informações de todas as variáveis, tanto as regressoras quanto a variável resposta. Isso significa que o conjunto de dados deve ser completo, com dados para todas as variáveis.

2) o modelo logístico não será bem ajustado, caso a variável de interesse esteja desbalanceada, isto é, se houver uma discrepância muito grande entre as respostas. O ideal é que haja um equilíbrio entre as duas classificações, isto é, no caso desse exemplo, o ideal seria que o número de empresas que precisam de seguro seja similar ao número de empresas que não precisam de seguro.

Iniciando a análise exploratória dos dados, observa-se a natureza de cada variável através da função *str*, permitindo verificar se os dados foram digitados corretamente. Também são apresentadas algumas estatísticas das variáveis regressoras:

```
# Nomeando as variáveis

attach(dados)

# Informações de cada variável

str(dados)
```

```
## 'data.frame':    18 obs. of  6 variables:
##  $ Empresa    : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Faturamento: int  289 275 258 199 170 149 132 125 106 105 ...
##  $ Empregados : int  15186 14711 13945 10263 9756 7929 6541 7175 5034 4987 ...
##  $ Projetos   : int  297 200 184 203 122 165 102 154 104 107 ...
##  $ Social     : int  3 2 2 1 1 1 1 1 1 1 ...
##  $ Seguro     : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
# Estatísticas descritivas

summary(dados[,2:5])
```

```
##   Faturamento      Empregados       Projetos          Social
##  Min.   : 16.0   Min.   :  311   Min.   : 20.00   Min.   :0.0000
##  1st Qu.: 97.5   1st Qu.: 3272   1st Qu.: 74.25   1st Qu.:0.0000
##  Median :105.5   Median : 5010   Median :105.50   Median :1.0000
##  Mean   :131.9   Mean   : 6481   Mean   :119.61   Mean   :0.8889
##  3rd Qu.:164.8   3rd Qu.: 9299   3rd Qu.:162.25   3rd Qu.:1.0000
##  Max.   :289.0   Max.   :15186   Max.   :297.00   Max.   :3.0000
```

A matriz de correlação entre as variáveis é dada por:

```
require(corrplot)

# Matriz de correlação

corrplot(cor(dados[,2:5]), method = "number")
```

Na Figura 4, pode-se notar que existe uma correlação muito alta entre as variáveis. Isso chama a atenção e pode indicar a presença de multicolinearidade, uma vez que existem correlações acima de 90%.

Figura 4. Matriz de correlação entre as variáveis do exemplo 1.

Para verificar isso, calcula-se o fator de inflação da variância, conhecido como *VIF*, para cada variável. Valores de *VIF* entre 1 e 5 indicam que existe correlação entre as variáveis preditoras, mas uma correlação aceitável. Entre 5 e 10 indicam uma alta correlação, o que pode invalidar o modelo ajustado. Já acima de 10 indicam que os coeficientes do modelo não são estimados adequadamente, devido à presença de multicolinearidade.

Primeiramente, obtém-se a equação do modelo de regressão linear considerando todas as variáveis regressoras. Para isso, é preciso que a variável de interesse seja do tipo FATOR, o que pode ser definido aplicando a função *factor*:

```
y <- factor(Seguro) ~ Faturamento + Empregados + Projetos + Social
```

Em seguida, parte-se para a estimação do modelo logístico pelo método dos mínimos quadrados, com a função *lrm* do pacote **rms**:

```
require(rms)

# Ajustando o modelo de regressão logístico

modelo <- lrm(y,data=dados)
modelo
```

```
## Logistic Regression Model
##
##  lrm(formula = y, data = dados)
##
##                       Model Likelihood        Discrimination    Rank Discrim.
##                             Ratio Test               Indexes          Indexes
##  Obs         18    LR chi2      24.05    R2           1.000    C       1.000
##   0          11    d.f.             4    g           32.963    Dxy     1.000
##   1           7    Pr(> chi2) <0.0001    gr    2.068803e+14    gamma   1.000
##  max |deriv|  3                          gp           0.503    tau-a   0.503
##                                         Brier        0.000
##
##             Coef     S.E.     Wald Z Pr(>|Z|)
```

```
##  Intercept     4.1548 147.2459  0.03  0.9775
##  Faturamento   1.1434   6.0171  0.19  0.8493
##  Empregados   -0.0171   0.0896 -0.19  0.8491
##  Projetos     -0.3187   1.9006 -0.17  0.8668
##  Social      -18.4531 120.8117 -0.15  0.8786
##
```

Pode-se verificar na saída acima os valores estimados dos coeficientes do modelo logístico encontrado, assim como o erro padrão das estimativas, coeficiente de Wald (para verificar a qualidade de ajuste do modelo) e o *p-valor* de cada variável regressora (para o teste de hipótese de significância das estimativas), além de mais algumas estatísticas sobre o modelo ajustado.

Considerando-se o teste no qual a hipótese nula seja que os efeitos das estimativas é igual a zero, todas as variáveis apresentaram um *p-valor* não significativo, isto é, o modelo ajustado contém variáveis que não conseguem explicar a variável resposta.

Como a matriz de correlações acusou haver uma correlação alta entre as variáveis, é possível que a presença de multicolinearidade possa estar influenciando na qualidade do ajuste do modelo.

Executando a função *VIF* pode-se verificar que os valores obtidos estão bem acima do valor 10, que é considerado o limite para a presença de multicolinearidade. Logo, esse pressuposto não é atendido.

```
vif(modelo)
```

```
## Faturamento  Empregados    Projetos      Social
##   46.947099   34.699595    2.959404    2.804075
```

Uma alternativa para contornar esse problema seria retirar do modelo as variáveis que apresentam multicolinearidade e fazer o ajuste de um novo modelo contendo as demais variáveis regressoras, pois essas atendem esse pressuposto e podem promover um ajuste melhor. No entanto, todas as variáveis regressoras são não significativas no ajuste do modelo proposto.

Com isso, chega-se à conclusão de que os parâmetros do modelo não representam adequadamente as melhores estimativas. Na prática, não é possível estimar a probabilidade de uma empresa contratar um seguro considerando as variáveis inseridas no modelo logístico.

Exemplo prático 2: modelo logístico não significativo

Nesse outro exemplo, apresenta-se a situação na qual o modelo ajustado, embora atenda os pressupostos necessários, não apresenta variáveis significativas.

Considere-se o conjunto de dados simulados contendo 11 genótipos de batata e 3 variáveis: número médio de dias até o início da tuberização, rendimento médio por planta (kg) e a classificação dos genótipos em 1 para precoce e 0 para não precoce.

Planilha de dados 2 - Dados simulados do exemplo 2.

```
dados<-read.table("dados_ex2.txt",header=TRUE)
dados
```

```
##       Genotipo  Maturacao  Rendimento Precocidade
## 1         1       39.63       8.94       1
## 2         2       38.23       9.86       1
## 3         3       37.27       8.49       0
```

```
## 4          4       36.73       9.36       0
## 5          5       35.87       6.96       1
## 6          6       34.00       7.20       1
## 7          7       33.33       4.20       1
## 8          8       33.21       7.04       1
## 9          9       32.50       5.00       0
## 10        10       30.67       5.40       0
## 11        11       29.67       3.90       0
```

A variável de interesse é denominada **Precocidade** e será a variável dependente no modelo logístico. As variáveis regressoras são **Maturação** e **Rendimento**.

Antes de iniciar o ajuste do modelo, a variável resposta será transformada em fator e será feita a conferência da categoria de referência que está sendo considerada pelo **R**:

```
attach(dados)

# Transformando a variável em fator
Precocidade<-as.factor(Precocidade)


# Verificando a categoria de referência
levels(Precocidade)

## [1] "0" "1"
```

A categoria de referência é aquela classificada como "0", portanto as probabilidades estimadas serão em relação às cultivares não precoces.

O ajuste do modelo logístico também pode ser feito usando a função *glm* indicando a família *binomial*. Nessa situação, o *software* entende que a função de ligação a ser utilizada é a **logito**, por *default*, por isso sua inclusão é opcional.

```
# Verificando o nome das variáveis do conjunto de dados
colnames(dados)

## [1] "Genotipo"    "Maturacao" "Rendimento"  "Precocidade"

# Ajustando o modelo de regressão logístico
ajuste <- glm(Precocidade~Tuberizacao+Rendimento,

family=binomial(link='logit'),data=dados)
```

Antes de detalhar as estimativas encontradas, deve-se verificar a presença de multicolinearidade. Neste exemplo, espera-se que esse pressuposto seja atendido corretamente:
vif(ajuste)

```
## Maturacao  Rendimento
##    4.488771    4.488771
```

O valor apresentado é inferior a 5, o que representa ausência de multicolinearidade. O valor é repetido, pois são apenas duas variáveis regressoras. Quando o modelo contar com três ou mais, esse valor provavelmente será diferente para cada variável.

Em seguida, usa-se a função *summary* para visualizar os parâmetros do modelo. O resultado é similar ao exemplo anterior, quando foi utilizada função *lrm*.

```
summary(ajuste)
```

```
##
## Call:
## glm(formula = Precocidade ~ Maturacao + Rendimento, family = binomial(link = "log-
it"),
##     data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.6362  -0.9411   0.4365   0.8203   1.5104
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -18.1989    13.9291  -1.307    0.191
## Tuberizacao   0.6485     0.5262   1.232    0.218
## Rendimento   -0.5813     0.7281  -0.798    0.425
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 15.158  on 10  degrees of freedom
## Residual deviance: 12.730  on  8  degrees of freedom
## AIC: 18.73
##
## Number of Fisher Scoring iterations: 3
```

A função executada apresenta os valores dos coeficientes estimados, o erro-padrão, a estatística padronizada *z* do teste de significância dos parâmetros e os respectivos *p-valor*. A hipótese nula do teste é que as estimativas dos coeficientes do modelo sejam iguais a zero, isto é, o modelo é não significativo para um *p-valor* acima de 0,05.

Verificando as variáveis regressoras do modelo, observa-se que as duas apresentam um *p-valor* maior do que 0,05 (0,218 e 0,425 respectivamente), indicando que a hipótese nula do teste não deve ser rejeitada. Logo, pode-se inferir que os efeitos das variáveis independentes no modelo não é significativo.

Outra forma de avaliar a significância do modelo é utilizar a função *ANOVA* com o teste *Chisq*:

```
anova(ajuste, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Precocidade
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                           10     15.158
## Maturacao    1  1.74033         9     13.418   0.1871
## Rendimento   1  0.68827         8     12.730   0.4068
```

Nesse teste, a hipótese nula é a mesma que a anterior, no entanto a estatística de teste é obtida com base na função *Deviance* ao se acrescentar uma a uma as variáveis regressoras. Embora os testes tenham abordagens diferentes, espera-se que o resultado final seja parecido com os valores obtidos no teste anterior. Conforme o último resultado, as variáveis regressoras permanecem não significativas no modelo, em nível de 95% de confiança.

Na prática, isso representa dizer que o modelo logístico, embora ajustado, não é adequado para classificar a variável resposta, ou seja, uma cultivar em relação à precocidade. Um modelo mais eficiente poderá ser obtido se, principalmente, a variável Rendimento for substituída por outras que possam agregar qualidade ao ajuste do modelo, e assim proporcionar uma classificação válida quanto à precocidade.

Exemplo prático 3: modelo logístico bem ajustado

Por fim, no último exemplo prático deste trabalho, apresenta-se um exercício simulado no qual o modelo logístico apresenta-se completamente adequado para fazer a classificação de uma amostragem de rochas.

O conjunto de dados a seguir representa uma amostra de tamanho 100 contendo 3 variáveis simuladas (*var1*, *var2* e *var3*), que representam atributos geoquímicos das rochas da região Sul do Brasil. A primeira variável é numérica e fornece um valor percentual de bases, sendo gerada a partir de uma distribuição normal com média 40 e desvio padrão 15. A segunda variável também é numérica,considera a presença de um determinado metal (em mg/kg), e foi gerada a partir de uma distribuição normal com média 10 e desvio padrão 4. A terceira variável é binária e simula o nível de outros minerais não desejáveis (em mg/kg): 0 foi atribuído para valores baixos e 1 para valores altos.

A variável de interesse foi codificada para representar a possibilidade de uso na agricultura: 1 para rochas que atendem os requisitos mínimos e podem ser utilizadas e 0 para as rochas que foram reprovadas e não apresentam potencial de uso na agricultura.

```
dados<-read.table("dados_ex3.txt",header=TRUE)
attach(dados)

# Transformando a variável em fator
Uso<-as.factor(Uso)

# Gerando o arquivo de dados
dados <- data.frame(var1,var2,var3,Uso)
str(dados)

## 'data.frame':    100 obs. of  4 variables:
##  $ var1  : int  47 45 61 29 30 16 41 37 17 57 ...
##  $ var2  : int  11 10 14 10 13 7 11 13 9 8 ...
##  $ var3  : int  0 1 1 1 1 1 0 0 1 0 ...
##  $ Uso: Factor w/ 2 levels "0","1": 2 2 2 1 2 1 1 1 1 1 ...

# Verificando a categoria de referência para o modelo logístico
levels(Uso)

## [1] "0" "1"
```

A variável de interesse denominada **Uso** apresenta somente os valores 0 e 1 e foi transformada em um fator. A categoria de referência para os cálculos das probabilidades do modelo ajustado é a categoria "0".

Realizando-se uma análise exploratória do conjunto de dados, é possível fazer o histograma das variáveis 1 e 2, que são contínuas (Figura 5). A linha vermelha representa a média dos dados:

```
# Dividindo a janela gráfica em uma linha e duas colunas
par(mfrow=c(1,2))
hist(var1, border="black", xlab="Variável 1", ylab="Frequência", main="")

# Adicionando uma linha vertical vermelha representando a média dos dados
abline(v=mean(var1), col="red", lwd=2)

hist(var2, border="black", xlab="Variável 2", ylab="Frequência", main="")
abline(v=mean(var2), col="red", lwd=2)
```

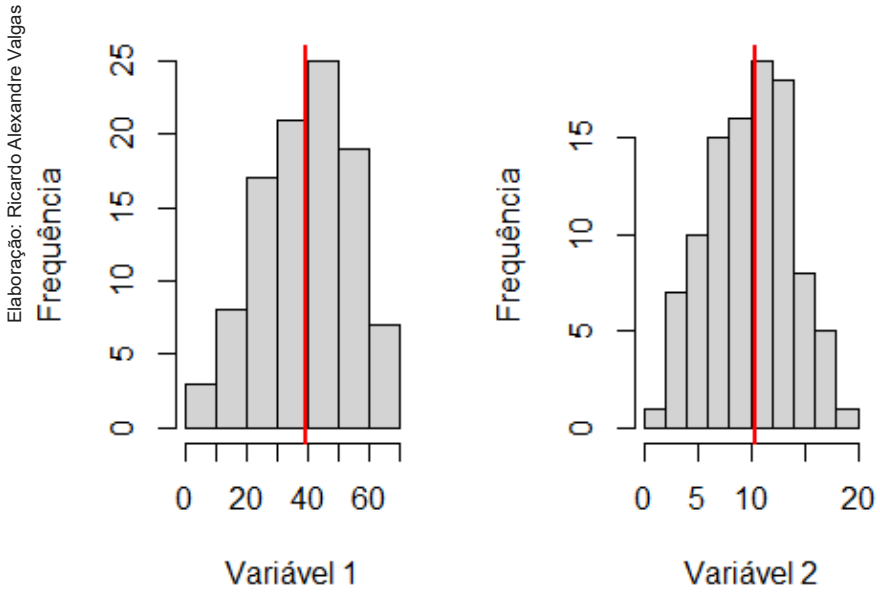


Figura 5. Histograma das variáveis 1 (% de bases) e 2 (mg/kg de metal). Dados simulados.

A variável 3 apresenta duas categorias e sua frequência é apresentada na Figura 6 a seguir:

```
barplot(data.frame(table(var3))[,2], names.arg = c(0, 1), col="gray",xlab="Variável 3",
ylab="Frequência", main ="")
```

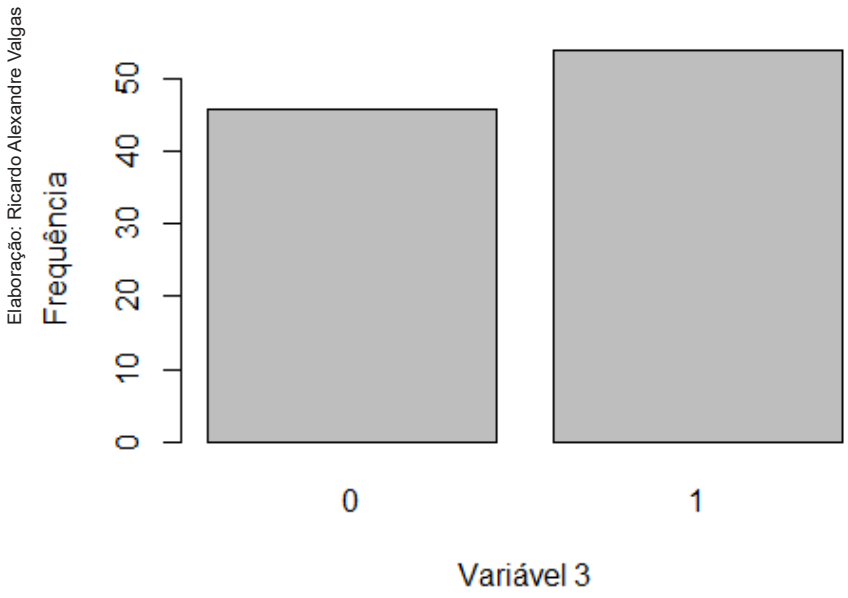


Figura 6. Dados da variável 3 (nível de alguns minerais) (dados simulados).

O modelo de regressão logística para as três variáveis independentes (saturado) será ajustado usando a função *glm* considerando a família *binomial* e a função de ligação *logit*:

```
# Carregando alguns pacotes
library(visreg)

# Ajustando o modelo saturado
mod_sat<-glm(Uso ~ var1 + var2 + factor(var3), family=binomial(link='logit'), data=da-
dos)


summary(mod_sat)

##
## Call:
## glm(formula = Uso ~ var1 + var2 + factor(var3), family = binomial(link = "logit"),
##     data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.8717  -0.6536  -0.1604   0.6811   2.0803
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -6.08849    1.42953  -4.259 2.05e-05 ***
## var1           0.10684    0.02382   4.486 7.25e-06 ***
## var2           0.03674    0.06591   0.557 0.577260
## factor(var3)1  2.25370    0.57954   3.889 0.000101 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.27  on 99  degrees of freedom
## Residual deviance:  91.90  on 96  degrees of freedom
## AIC: 99.9
##
## Number of Fisher Scoring iterations: 5
```

O resultado obtido fornece os coeficientes estimados (*Estimate*) do modelo na escala *logit*, o erro padrão das estimativas (*Std. Error*), o valor calculado *z* da estatística padronizada (*z value*) usada no teste de hipótese de que as estimativas dos coeficientes do modelo são significativamente diferentes de zero, e o *p-valor* (*Pr(>|z|)*) associado para cada valor de *z*.

Embora as estimativas dos coeficientes não tenham uma interpretação direta, como no caso dos modelos de regressão linear, é possível examinar o sinal das estimativas (positivo ou negativo) e comparar sua direção em relação à variável resposta: as três variáveis regressoras possuem um efeito positivo sobre a probabilidade das rochas não atenderem os requisitos (categoria de referência **0**), pois apresentam sinal positivo.

Olhando para a significância de cada variável no modelo, nota-se que as variáveis 1 e 3 apresentam *** e, portanto, um valor significativo ($p < 0,05$), enquanto a variável 2 não ($p = 0,577260$). Isso indica que o modelo logístico completo, contendo todas as variáveis independentes, pode ser simplificado retirando-se a variável não significativa e mantendo as demais.

O ajuste do modelo logístico considerando apenas as variáveis significativas 1 e 3 é dado a seguir:

```
# Ajuste do modelo logístico com as duas variáveis significativas
mod_sig <- glm(Uso ~ var1 + factor(var3), family=binomial(link='logit'), data=dados)


summary(mod_sig)

##
## Call:
## glm(formula = Uso ~ var1 + factor(var3), family = binomial(link = "logit"),
##     data = dados)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.7699  -0.6885  -0.1577   0.6927   2.0923
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -5.65669    1.17732  -4.805 1.55e-06 ***
## var1           0.10631    0.02376   4.474 7.69e-06 ***
## factor(var3)1  2.20467    0.57024   3.866 0.000111 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 138.269  on 99  degrees of freedom
## Residual deviance:  92.213  on 97  degrees of freedom
## AIC: 98.213
##
## Number of Fisher Scoring iterations: 5
```

Matematicamente, a expressão do modelo é dada por:

$$\ln\left(\frac{prob_{Uso}}{1-prob_{Uso}}\right) = -5{,}65669 + 0{,}10631var1 + 2{,}20467var3$$

Comparando dois modelos ajustados

Observe-se que as duas variáveis regressoras permanecem significativas no novo modelo. Então surge a questão: qual dos dois modelos apresenta o melhor ajuste: o modelo anterior, completo, com todas as variáveis regressoras (significativas ou não), ou o modelo simplificado, contendo apenas as variáveis significativas?

Para responder essa pergunta, utilizam-se informações que servem de parâmetro para avaliar a bondade (ou a qualidade) do ajuste de modelos: a *deviance* (em português, desvio), a *Null deviance* que corresponde ao desvio do modelo nulo, sem nenhuma variável regressora, e a *Residual deviance*, que é o desvio residual do modelo que foi ajustado.

Primeiro em relação ao modelo nulo: comparando-se os dois valores da *deviance*, verifica-se que o desvio residual (*residual deviance*) foi menor (92,213), quando comparado ao modelo nulo (*null deviance*), que foi de 138,269. O valor menor do desvio residual indica que o modelo ajustado apresenta-se melhor ao incluir somente as variáveis regressoras 1 e 3.

No entanto, antes de se fazer a comparação em relação ao modelo completo, é necessário saber que a comparação entre dois modelos ajustados quaisquer é feita pela comparação da informação de Akaike (*AIC*), a qual também se baseia na *deviance*. Quanto menor for o valor de *AIC*, melhor o ajuste. Portanto, ao se comparar dois valores de *AIC*, o modelo que apresentar o menor valor será o que proporciona o melhor ajuste. Por si só, a informação de Akaike não possui uma interpretação direta.

Então, comparando-se o modelo saturado em relação ao modelo simplificado, tem-se que o primeiro fornece um *AIC* de 99,9 e o segundo 98,213. Como o valor diminuiu, admite-se que o modelo contendo somente as duas variáveis regressoras significativas apresenta um ajuste melhor que o modelo com todas as variáveis.

Entre os pressupostos do modelo, é preciso verificar a presença de multicolinearidade. Os valores obtidos são todos satisfatórios conforme segue:

```
require(rms)
```

```
vif(mod_sig)
```

```
##          var1 factor(var3)1
##      1.208724      1.208724
```

Estimação da razão de chances

Com o pressuposto de multicolinearidade atendido, o modelo ajustado pode ser explorado mais a fundo. Sabendo-se que nos modelos de regressão logística os resultados dos estimadores estão na forma logaritmo, pode-se obter uma melhor interpretação da relação entre as variáveis dependente e independentes efetuando-se a exponenciação das variáveis de regressão, e assim obter a *odds ratio* (razão de chances) do modelo.

Para isso, utiliza-se a função *summ* do pacote **jtools**. O resultado da razão de chances encontrado foi 1,11 para a variável 1 e 9,07 para a variável 3. Isso significa que, para cada variação de uma unidade nas variáveis 1 e 3, as chances aumentam em 1,11 vezes [ou (1,11 - 1) * 100 = 11%] e 9,07 vezes [ou (9,07 - 1) * 100 = 807%], respectivamente, a chance da variável **Uso** pertencer a categoria 0. Na prática, isso significa que a variável 3, quando classificada como 1 (presença de outros minerais indesejáveis), tem um peso bem maior para classificar a variável resposta.

```
require(jtools)
```

```
summ(mod_sig, exp = T)
```

```
## MODEL INFO:
## Observations: 100
## Dependent Variable: Uso
## Type: Generalized linear model
##   Family: binomial
##   Link function: logit
##
## MODEL FIT:
## <U+03C7>²(2) = 46.06, p = 0.00
## Pseudo-R² (Cragg-Uhler) = 0.49
## Pseudo-R² (McFadden) = 0.33
## AIC = 98.21, BIC = 106.03
##
## Standard errors: MLE
## ------------------------------------------------------------
##                       exp(Est.)   2.5%   97.5%   z val.      p
## ------------------- ----------- ------ ------- -------- ------
## (Intercept)                0.00   0.00    0.04    -4.80   0.00
## var1                       1.11   1.06    1.17     4.47   0.00
## factor(var3)1              9.07   2.97   27.72     3.87   0.00
## ------------------------------------------------------------
```

Construindo um intervalo de confiança de 95% para cada razão de chances, tem-se o resultado a seguir. Na coluna 2,5% está o limite inferior e na coluna 97,5% o limite superior:

```
exp(cbind(OR = coef(mod_sig), confint(mod_sig)))
```

```
##                         OR        2.5 %     97.5 %
## (Intercept)   0.003494062 0.0002657397  0.0280239
## var1          1.112169332 1.0657154370  1.1706966
## factor(var3)1 9.067241478 3.1602982703 30.2802989
```

Pseudo - R^2

No modelo de regressão logístico não há uma estatística resumo que forneça a variação na variável resposta explicada pelo modelo, como ocorre nos modelos lineares com o coeficiente de determinação R^2 . Há uma medida similar, o *Pseudo - R^2*, que fornece uma ideia do poder preditivo/explicativo do modelo.

A interpretação do *Pseudo - R^2* é simples: quanto mais próximo de zero, menor será a diferença entre o modelo nulo e o modelo estimado. Quanto mais próximo de um, maior será essa diferença, indicando que as variáveis regressoras presentes no modelo estimado não contribuem para a explicação da variável dependente.

Contribuição dos parâmetros do modelo

Como o modelo logístico simplificado está bem ajustado, é possível identificar qual a contribuição de cada variável regressora no modelo. Através da *ANDEVA* (*Analysis of Deviance*), ou de uma análise da função *deviance*, realiza-se uma comparação entre os desvios do modelo saturado e dos modelos aninhados com um número menor de parâmetros. Para isso, executa-se a função *anova* informando o teste qui-quadrado:

```
anova(mod_sig, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: Uso
##
## Terms added sequentially (first to last)
##
##
##              Df Deviance Resid. Df Resid. Dev  Pr(>Chi)
## NULL                            99    138.269
## var1          1   27.561        98    110.708 1.522e-07 ***
## factor(var3)  1   18.496        97     92.213 1.703e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

O teste realizado compara os valores dos desvios dos modelos aninhados acrescentando de forma sequencial um parâmetro (variável) por vez, a partir do modelo nulo. A coluna *Deviance*, nesse caso, apresenta a redução do desvio residual em relação ao modelo anterior (com uma regressora a menos), seguido pela coluna *Pr(>Chi)* contendo o *p-valor* para o teste qui-quadrado. A conclusão obtida é de que as variáveis 1 e 3 contribuem de forma significativa com a função de verossimilhança do modelo.

Predição de probabilidades

Após verificar a qualidade do modelo e constatar que o mesmo foi bem ajustado, pode-se verificar como o modelo prevê uma determinada amostra. Para os valores: variável 1 = 45 e variável 3 = 1 (valores altos de minerais não desejáveis), o modelo de regressão logístico ajustado fornece uma probabilidade de 79,11% da rocha não ser adequada para o uso na agricultura.

```
pred1=data.frame(var1 = 45, var3=factor(1))
pred1$prob=predict(mod_sig, newdata=pred1, type="response")
pred1
```

```
##   var1 var3      prob
## 1   45    1 0.7911776
```

Alterando somente a categoria da variável 3, tem-se: variável 1 = 45 e variável 3 = 0; a probabilidade da rocha ser rejeitada diminui para 29,47%. Isso faz todo sentido, uma vez que, ao classificar a variável 3 com valores baixos de minerais não desejáveis, a probabilidade da rocha ser descartada cai significativamente.

```
pred2=data.frame(var1 = 45, var3=factor(0))
pred2$prob=predict(mod_sig, newdata=pred2, type="response")
pred2
```

```
##   var1 var3      prob
## 1   45    0 0.2947074
```

Graficamente, é possível explorar a predição do modelo em relação aos valores da variável 3 (a mais significativa). Isso é feito criando-se uma tabela com os valores médios da variável 1 e ordenando-os com os dois valores da variável 3. Feito isso, aplica-se o modelo logístico ajustado, obtém-se o erro padrão das estimativas e calcula-se os limites do intervalo de 95% de confiança (Figura 7).

```
# Criação da tabela
predvar3=with(dados,data.frame(var1=mean(var1),var3=factor(0:1)))

predvar3=cbind(predvar3,predict(mod_sig,
                                newdata=predvar3,
                                type="response",
                                se.fit=TRUE))
# Renomeando as variáveis
names(predvar3)[names(predvar3)=='fit']="prob"
names(predvar3)[names(predvar3)=='se.fit']="se.prob"

# Estimando os intervalos de confiança
predvar3$LL=predvar3$prob-1.96*predvar3$se.prob
predvar3$UL=predvar3$prob+1.96*predvar3$se.prob

require(ggplot2)
ggplot(predvar3, aes(x=var3,y=prob))+
  geom_errorbar(aes(ymin=LL, ymax=UL), width=0.2,lty=1,lwd=1,col="red")+
  geom_point(shape=18, size=5, fill="black")+
  scale_x_discrete(limits=c("0","1"))+
  labs(title="", x="Variável 3",y="Pr(y=0)")
```

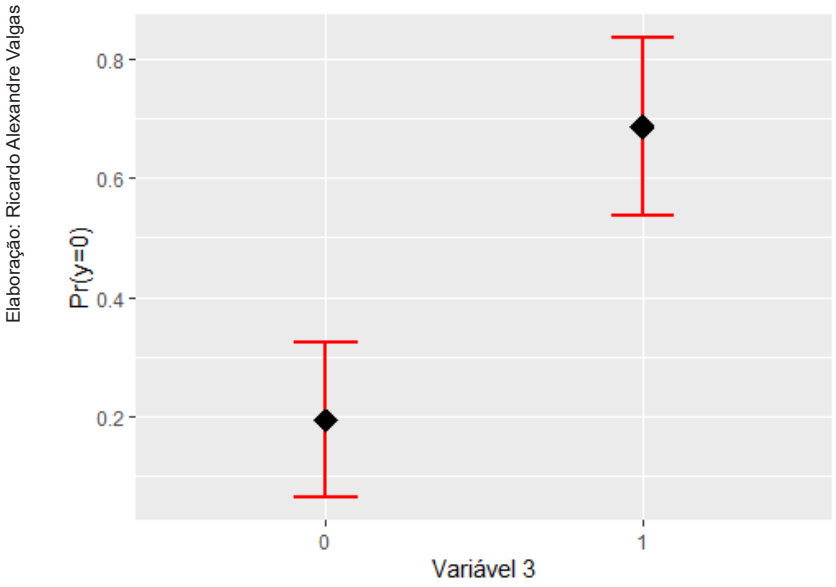


Figura 7. Probabilidades preditas pelo modelo logístico ajustado em relação à variável 3. Dados simulados.

Análise de resíduos

Para iniciar a análise de resíduos, é necessário obter os resíduos de Pearson, e em seguida calcular a estatística de Pearson.

```
# Calculando os resíduos de Pearson
resP<-data.frame(indice=1:length(var1),residuos=residuals(mod_sig,type="pearson"))
```

Visualmente, pode-se fazer o gráfico dos resíduos de Pearson (Figura 8): o desejável é que os pontos estejam concentrados entre -2 e 2, indicando um bom comportamento. Para os dados do exemplo 3, apenas dois pontos tiveram valor em módulo maior que 2, o que é tolerável pela grande quantidade de dados.

```
ggplot(resP,aes(x=sample(indice),y=residuos))+geom_point()+geom_hline(yintercept=0)+
  labs(x="Amostra",y="Resíduos")
```

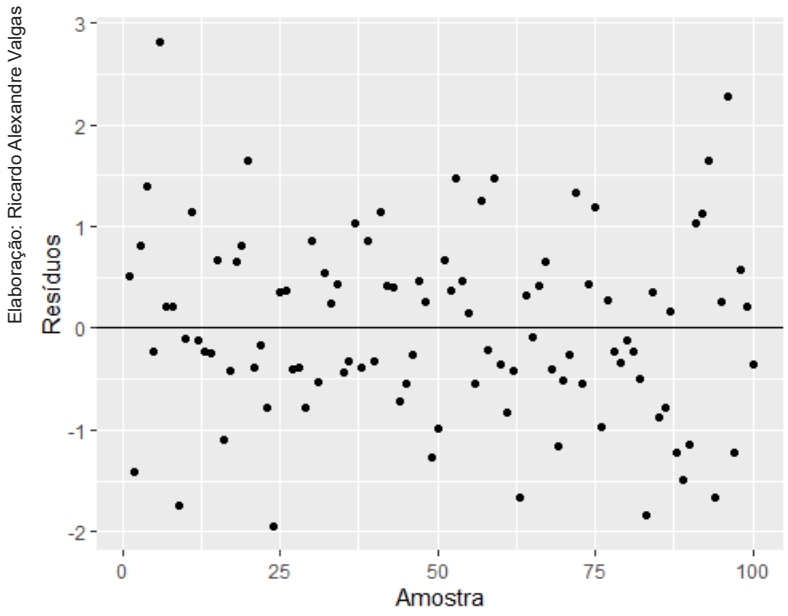


Figura 8. Gráfico dos resíduos de Pearson do modelo logístico ajustado (dados simulados).

O teste associado à estatística qui-quadrado de Pearson para o teste de resíduos tem como hipótese nula que o modelo está bem ajustado, logo espera-se que o teste não seja rejeitado. O *p-valor* da estatística do teste é dada por:

```
pchisq(sum(resP$residuos^2), df = mod_sig$df.residual, lower.tail = F)
```

```
## [1] 0.8662746
```

Como o *p-valor* encontrado é maior que 0,05, a hipótese nula não é rejeitada, em nível de 95% de confiança. Portanto, pode-se garantir que o modelo escolhido está bem ajustado.

Considerações finais

A metodologia de regressão logística apresentada neste trabalho aprofunda-se nos conceitos, na teoria e, principalmente, na prática, de forma a proporcionar ao leitor o entendimento necessário para sua aplicação.

Demonstrou-se que a regressão logística requer que uma quantidade menor de pressupostos sejam atendidos, em comparação com a metodologia de modelos lineares. Assim, torna-se facilmente aplicada em experimentos nas áreas da agronomia, entomologia, zootecnia, entre outros. Para isso, basta delinear um experimento ou a coleta de dados buscando modelar uma variável resposta dicotômica, em função de algumas variáveis regressoras que apresentem uma correlação razoável entre elas.

Disponibilizou-se ao leitor deste trabalho toda a sequência de uso do *software* **R** para processar uma análise de dados do início ao fim, obtendo os modelos de regressão logísticos e explorando seus resultados. Diferentes exemplos foram demonstrados, com variadas abordagens e contextos, de forma a maximizar o entendimento das situações em que os modelos de regressão logísticos podem ou não apresentar um bom resultado.

Referências

FAVERO, L. P.; BELFIORE, P.; SILVA, F. L. da; CHAN, B. L. **Análise de dados**: modelagem multivariada para tomada de decisões. Rio de Janeiro: Elsevier, 2009. p. 440-467.

GONZALEZ, L. de A. **Regressão Logística e suas Aplicações**. 2018. Monografia - UFMA.

R Core Team. **R**: A language and environment for statistical computing. Vienna: R Foundation for Statistical Computing, 2018. Disponível em: https://www.R-project.org/.

Literatura recomendada

ARAUJO, G. L. D. de. **Métodos de estimação em regressão logística com efeito aleatório: aplicação em germinação de sementes**. 2012. Dissertação (Mestrado – Universidade Federal de Viçosa, Viçosa.

FERNANDES, A. A. T.; FIGUEIREDO FILHO, D. B.; ROCHA, E. C. da; NASCIMENTO, W. da S. Leia este artigo se você quiser aprender regressão logística. **Revista de Sociologia e Política,** 2020. Disponível em: https://www.scielo.br/j/rsocp/a/RWjPthhKDYbFQYydbDr3MgH/?lang=pt. Acesso em: 5 abr. 2022.

LOESCH, C.; HOELTGEBAUM, M. **Métodos Estatísticos Multivariados**. São Paulo: Saraiva, 2012. 288 p.

MINGOTI, S. A. **Análise de dados através de métodos de estatística multivariada**: uma abordagem aplicada. Belo Horizonte: Editora UFMG, 2007. 297 p.

SMOLSKI, F. M. S.; BATTISTI, I. E.; CHASSOT, T.; REIS, D. I.; KASZUBOWSKI, E.; RIEGER, D. S. Capacitação em análise estatística de dados utilizando o software livre R. **Revista Ciência em Extensão**, v.14, n. 3, p. 123-134, 2018. Disponível em: https://ojs.unesp.br/index.php/revista_proex/article/viewFile/1823/2073 Acesso em: 05 out. 2022.

Embrapa

Clima Temperado

MINISTÉRIO DA
AGRICULTURA, PECUÁRIA
E ABASTECIMENTO

PÁTRIA AMADA
BRASIL
GOVERNO FEDERAL

CGPE 017891