# REGRESSÃO LOGÍSTICA BINÁRIA

## Relação entre as redes sociais e a geração de lixo

Como as redes sociais podem afetar o gerenciamento de resíduos gerados em casa?

Relatório de Análise Estatística by LdeSV

## Contents



# METODOLOGIA DA ANÁLISE ESTATÍSTICA

Esta é uma análise estatística de cunho exploratório em que foi realizada uma regressão logística binária sequida por análise de deviância e seleção do modelo (completo *versus* média geral).

O modelo completo é o que possui $n$ parâmetros (em particular, $n = 10$); já o modelo médio é aquele que possui apenas o intercepto (que representa a média geral). Tomou-se como variável resposta $y$ a questão 1 sobre a geração de resíduo doméstico e sem relação com uso de rede social, donde 1 correspondeu à "Sim" e 0 à "Não". Já as variáveis preditoras $x$ são as questões sobre gênero, escolaridade e as questões de 2 a 5 (as quais que relacionaram o uso de rede social).

Então, para testar o efeito de alguma variável preditora sobre a chance de obter a resposta, aplicou-se o teste da razão da verossimilhança (*Likelyhood Ratio Test*, LTR). A regra de decisão do LTR foi: se o teste for significativo ao nível de 5% de probabilidade de erro ($p \leq 5$), então se conclui que existe o efeito de alguma das variáveis preditoras. Por fim, se obteve a razão de possibilidades (*odds ratio*) a partir dos coeficentes do modelo ajustado aos dados.

Note que, se utiliza o conceito de variáveis latentes, que por definição não são diretamente observadas, mas são inferidas através de um modelo matemático-estatístico e da mensuração de variáveis observáveis. Em geral, tais variáveis não podem ser acessadas diretamente, mas possuem manifestações no mundo real (e.g., personalidade: extroversão ou introversão).

# PREÂMBULO da/para ANÁLISE DE DADOS

```
## Formatacao numerica
options(scipen = 6, digits = 8) #saidas com notação científica
options(OutDec = ",") #saidas com separador decimal escolhido (, ou .)

## Instalando pacote(s)
library(readr)
```

## Importando os dados

```
dados <- read_csv("dados.csv")
head(dados) #leitura das linhas iniciais do conjunto de dados coletado via questionário
```

```
## # A tibble: 6 x 7
##   genero    escolaridade        y segue aplica compartilha aprende_aplica
##   <chr>     <chr>           <dbl> <chr> <chr>  <chr>       <chr>
## 1 Feminino  Ensino superior     1 Sim   Sim    Sim         Sim
## 2 Feminino  Ensino médio       1 Sim   Sim    Sim         Não
## 3 Feminino  Ensino superior     1 Sim   Sim    Sim         Sim
## 4 Feminino  Ensino superior     1 Sim   Sim    Sim         Sim
## 5 Masculino Ensino superior     1 Sim   Não    Não         Não
## 6 Masculino Ensino superior     1 Sim   Sim    Sim         Sim
```

em que: $y$ = variável resposta (questão 1, sem relação com rede social) em que 1=Sim e 0=Não; $x$ = variáveis explicativas ou preditoras (questões de 2 a 5, que relaciona o uso de rede social).

## Pré-processamento dos dados

Passo necessário para implementação da análise no R.

```
str(dados) #função que exibe de forma compacta a estrutura de um objeto R arbitrário
```

```
## spec_tbl_df [267 x 7] (S3: spec_tbl_df/tbl_df/tbl/data.frame)
##  $ genero        : chr [1:267] "Feminino" "Feminino" "Feminino" "Feminino" ...
##  $ escolaridade  : chr [1:267] "Ensino superior" "Ensino médio" "Ensino superior" "Ensino superior" ...
##  $ y             : num [1:267] 1 1 1 1 1 1 0 0 1 1 ...
##  $ segue         : chr [1:267] "Sim" "Sim" "Sim" "Sim" ...
##  $ aplica        : chr [1:267] "Sim" "Sim" "Sim" "Sim" ...
##  $ compartilha   : chr [1:267] "Sim" "Sim" "Sim" "Sim" ...
##  $ aprende_aplica: chr [1:267] "Sim" "Não" "Sim" "Sim" ...
##  - attr(*, "spec")=
##   .. cols(
##   ..   genero = col_character(),
##   ..   escolaridade = col_character(),
##   ..   y = col_double(),
##   ..   segue = col_character(),
##   ..   aplica = col_character(),
##   ..   compartilha = col_character(),
##   ..   aprende_aplica = col_character()
##   .. )
##  - attr(*, "problems")=<externalptr>
```

```
dados$genero=as.factor(dados$genero) # transformando a coluna "genero" em fator
dados$escolaridade=as.factor(dados$escolaridade) # transformando a coluna "escolaridade" em fator
```

### Definindo a linha de base ou níveis de referência

Primeiro tem-se que escolher os níveis referência da análise exploratória através da função relevel.

```
# Para a variável resposta (dependente) genero (Feminino, Masculino, Outro) foi escolhido o nivel "Femi
dados$genero <- relevel(dados$genero, ref = "Feminino")

# Para a variável resposta (dependente) escolaridade foi escolhido o nivel "Ensino fundamental incomple
dados$escolaridade <- relevel(dados$escolaridade, ref = "Ensino fundamental incompleto")

# Para a variável preditora (independente) $y$ o nivel "Sim" (1) é o valor referência.
```

## ANÁLISE DE DADOS - Ajuste do modelo

Tomou-se a questão 1 como váriável resposta (dependente) $y$ e as questões sobre gênero, escolaridade e hábitos nas redes sociais (i.e: questões 2, 3, 4 e 5) como variáveis preditoras (independentes) $x$. Então foram ajustados dois modelos:

(i) SEM as variáveis preditoras, em que considera-se a média geral.

(ii) COM todas as variáveis preditoras.

Estes dois modelos foram comparados pelo teste da razão de verossimilhança.

**Regra de decisão**: Se o teste for significativo ao nível de 5% de probabilidade de erro ($p \leq 5$), então se conclui que existe o efeito de alguma das variáveis preditoras.

Modelo SEM as variáveis preditoras

```
modelo_0 = glm(y ~ 1, data = dados, family = binomial("logit"))
```

Modelo COM as variáveis preditoras

```
modelo = glm(y ~ ., data = dados, family = binomial("logit"))
summary(modelo)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial("logit"), data = dados)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1,93258  -0,47768   0,60432   0,79156   1,46625
##
## Coefficients:
##                                          Estimate  Std. Error z value
## (Intercept)                              14,125591 1029,121608  0,0137
## generoMasculino                          -0,681139    0,315064 -2,1619
## generoOutro                              15,265068 1455,397892  0,0105
## escolaridadeEnsino fundamental completo  -14,175207 1029,121908 -0,0138
## escolaridadeEnsino médio                 -13,958317 1029,121558 -0,0136
## escolaridadeEnsino superior              -13,972077 1029,121498 -0,0136
## segueSim                                   1,108363    0,354338  3,1280
## aplicaSim                                  0,350617    0,401343  0,8736
## compartilhaSim                            -0,018502    0,375809 -0,0492
## aprende_aplicaSim                          0,073322    0,312815  0,2344
##                                         Pr(>|z|)
## (Intercept)                              0,98905
## generoMasculino                          0,03063 *
```

```
## generoOutro                                0,99163
## escolaridadeEnsino fundamental completo  0,98901
## escolaridadeEnsino médio                   0,98918
## escolaridadeEnsino superior              0,98917
## segueSim                                   0,00176 **
## aplicaSim                                  0,38233
## compartilhaSim                             0,96073
## aprende_aplicaSim                          0,81468
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 300,835  on 266  degrees of freedom
## Residual deviance: 278,948  on 257  degrees of freedom
## AIC: 298,948
##
## Number of Fisher Scoring iterations: 14
```

# RESULTADOS

## ANÁLISE DE DADOS - ANAVA e teste LRT

Análise de Variância (ANAVA) seguida pelo *Likelyhood Ratio Test* (LRT) ou teste da razão da verossimilhança.

```
anova(modelo_0, modelo, test = "LRT") # Likelyhood Ratio Test (teste da razão da verossimilhança)
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ 1
## Model 2: y ~ genero + escolaridade + segue + aplica + compartilha + aprende_aplica
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       266    300,835
## 2       257    278,948  9  21,8872 0,0092435 **
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

### Risco relativo ou “chance”

```
exp(coef(modelo)) # calculo dos coeficientes = risco relativo ou "chance" (*odds*)
```

```
##                              (Intercept)                           generoMasculino
##                            1,3635347e+06                             5,0604049e-01
##                             generoOutro escolaridadeEnsino fundamental completo
##                            4,2612279e+06                             6,9788806e-07
##                 escolaridadeEnsino médio             escolaridadeEnsino superior
##                            8,6692143e-07                             8,5507511e-07
##                                segueSim                                 aplicaSim
##                            3,0293941e+00                             1,4199427e+00
##                           compartilhaSim                          aprende_aplicaSim
##                            9,8166829e-01                             1,0760766e+00
```

```
# Tabela de frequência das variáveis preditoras significativas na análise
with(dados, table(dados$genero,dados$segue))
```

```
##
##             Não Sim
##   Feminino   33 158
##   Masculino  18  57
##   Outro       1   0
```

### Conclusão

Em particular, podemos inferir que:

- o gênero influência, sendo que homens tem 0,506 vezes mais chances que mulheres de responder sim na questão 1 (y); e
- seguir conteúdos na rede social influência, sendo que quem respondeu sim tem 3,029 vezes mais chances de responder sim na questão 1 (y) do que quem respondeu não (em seguir conteúdos em rede social).

## Adendo

Como para gênero se obteve-se os níveis feminio, masculino e outro como resposta; pode-se repetir a análise para obter a *odds ratio* tomando como referência outro nível (você deve ter notado que na análise anterior tomou-se o gênero feminino como referência ;)

- Dados:

```
dados <- read_csv("dados.csv")
```

```
dados$genero=as.factor(dados$genero) # transformando a coluna "genero" em fator
dados$escolaridade=as.factor(dados$escolaridade) # transformando a coluna "escolaridade" em fator
```

- Definindo outro nível de referência (gênero masculino):

```
# Para a variável resposta (dependente) genero (Feminino, Masculino, Outro) foi escolhido o nivel "Femi
dados$genero <- relevel(dados$genero, ref = "Masculino")

# Para a variável resposta (dependente) escolaridade foi escolhido o nivel "Ensino fundamental incomple
dados$escolaridade <- relevel(dados$escolaridade, ref = "Ensino fundamental incompleto")

# Para a variável preditora (independente) $y$ o nivel "Sim" (1) é o valor referência.
```

- Modelos:

```
modelo_0 = glm(y ~ 1, data = dados, family = binomial("logit"))
```

```
modelo = glm(y ~ ., data = dados, family = binomial("logit"))
summary(modelo)
```

```
##
## Call:
## glm(formula = y ~ ., family = binomial("logit"), data = dados)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1,93258  -0,47768   0,60432   0,79156   1,46625
##
## Coefficients:
##                                      Estimate  Std. Error z value
## (Intercept)                         13,444452 1029,121637  0,0131
## generoFeminino                       0,681139    0,315064  2,1619
```

```
## generoOutro                                15,946207 1455,397893  0,0110
## escolaridadeEnsino fundamental completo  -14,175207 1029,121911 -0,0138
## escolaridadeEnsino médio                 -13,958318 1029,121560 -0,0136
## escolaridadeEnsino superior              -13,972077 1029,121501 -0,0136
## segueSim                                   1,108363    0,354338  3,1280
## aplicaSim                                  0,350617    0,401343  0,8736
## compartilhaSim                            -0,018502    0,375809 -0,0492
## aprende_aplicaSim                          0,073322    0,312815  0,2344
##                                           Pr(>|z|)
## (Intercept)                               0,98958
## generoFeminino                            0,03063 *
## generoOutro                               0,99126
## escolaridadeEnsino fundamental completo   0,98901
## escolaridadeEnsino médio                  0,98918
## escolaridadeEnsino superior               0,98917
## segueSim                                  0,00176 **
## aplicaSim                                 0,38233
## compartilhaSim                            0,96073
## aprende_aplicaSim                         0,81468
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 300,835  on 266  degrees of freedom
## Residual deviance: 278,948  on 257  degrees of freedom
## AIC: 298,948
##
## Number of Fisher Scoring iterations: 14
```

```
anova(modelo_0, modelo, test = "LRT") # Likelyhood Ratio Test (teste da razão da verossimilhança)
```

```
## Analysis of Deviance Table
##
## Model 1: y ~ 1
## Model 2: y ~ genero + escolaridade + segue + aplica + compartilha + aprende_aplica
##   Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
## 1       266    300,835
## 2       257    278,948  9  21,8872 0,0092435 **
## ---
## Signif. codes:  0 '***' 0,001 '**' 0,01 '*' 0,05 '.' 0,1 ' ' 1
```

```
exp(coef(modelo)) # calculo dos coeficientes = risco relativo ou "chance" (*odds*)
```

```
##                             (Intercept)                           generoFeminino
##                           6,9000376e+05                            1,9761265e+00
##                             generoOutro escolaridadeEnsino fundamental completo
##                           8,4207253e+06                            6,9788804e-07
##                  escolaridadeEnsino médio                escolaridadeEnsino superior
##                           8,6692140e-07                            8,5507508e-07
##                                segueSim                                aplicaSim
##                           3,0293941e+00                            1,4199427e+00
##                           compartilhaSim                          aprende_aplicaSim
##                           9,8166829e-01                            1,0760766e+00
```

- Conclusão: Acrescenta-se apenas que mulheres tem 1,976 vezes mais chances que homens de responder

sim na questão 1 (y).

# REFERÊNCIAS

R Core Team. R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria, 2021. URL https://www.R-project.org/.

# Copyright (C)



**L**ilian **de S**ouza **V**ismara

E-mail: lilianvismara@utfpr.edu.br