Atividade Avaliativa

Análise de dados categorizados

Ana Maria Alves da Silva

2025-06-14

Instruções

- O desenvolvimento desta atividade deve ser realizada de forma individual ou em dupla.
- Deve-se completar o arquivo Rmd enviado na atividade.
- É necessário devolver o arquivo em Rmd e em pdf.
- Valor da atividade: 10 pontos.

Descrição da atividade

Um estudo desenvolvido pela Profa. Denise Gonçalves, do Departamento de Otorrinolaringologia da UFMG, teve como interesse a ocorrência de manifestações otorrinolaringológicas em pacientes HIV positivos. Neste estudo, 112 pacientes foram acompanhados no período de março de 1993 a fevereiro de 1195, sendo 91 HIV positivo e 21 HIV negativo. A classificação quanto à infecção pelo HIV seguiu os critérios Do *Center for Disease Control* (CDC, 1987), sendo ela: HIV soronegativo (não possui o HIV), HIV soropositivo assintomático (possui o vírus mas não desenvolveu o quadro clínico de AIDS), com ARC (*Aids Related Complex*: apresenta baixa imunidade e outros indicadores clínicos que antecedem o quadro clínico de AIDS), ou AIDS (apresenta infecções oportunistas que definem AIDS).

As Covariáveis medidas no estudo são:

- **id** : Idade do Paciente (medida em anos);
- **sex**: Sexo do Paciente (0 se Masculino e 1 se Feminino)
- **grp**: Grupo de Risco (1 se HIV Soronegativo, 2 se HIV Soropositivo Assintomático, 3 se ARC e 4 se AIDS)
- **ats**: Atividade Sexual (1 se Homossexual, 2 se Bissexual e 3 se Heterossexual)
- **ud**: Uso de Droga Injetável (1 se Sim e 2 se Não)
- **ac**: Uso de Cocaína por Aspiração (1 se Sim e 2 se Não)

Questão 1 Faça a leitura do conjunto de dados *aids.txt* e formate as variáveis **sex**, **grp**, **ats**, **ud** e **ac** para fator.

Solução:

Primeiro vamos

```
library(readr)
# Caminho corrigido usando barra normal
dados <- read.table("C:/Users/anama/OneDrive/Documents/especializacao/modulo_14/atividade_avaliativa/ai
                    header = TRUE,
                    sep = "",  # ajuste conforme o separador real
                    stringsAsFactors = FALSE)

# Checagem
print(dim(dados))
```

```
## [1] 133  12
```

```
print(is.data.frame(dados))
```

```
## [1] TRUE
```

Vamos verificar se há valores ausentes antes de transformar em fator.

```
colSums(is.na(dados))
```

```
##  pac   id  sex  grp   ti   tf cens  cd4  cd8  ats   ud   ac
##    0    2    0    0    0    0    0   46   46   26   26   26
```

Não vamos excluir os dados ausentes, pois isso reduz a base sem necessidade no entanto, quando necessário iremos omiti-los da análise. Vale ressaltar também que alguns testes estatísticos permintem omitir os valores ausentes na hora de aplicar o teste.. Agora, vamos transforma as variáveis em fator.

```
dados$sex <- factor(dados$sex, levels = c(0, 1),
                    labels = c("Masculino", "Feminino"))
dados$grp <- factor(dados$grp,
                    levels = c(1, 2, 3, 4),
                    labels = c("HIV-", "HIV+ Assintomático", "ARC", "AIDS"))
dados$ats <- factor(dados$ats, levels = c(1, 2, 3),
                    labels = c("Homossexual", "Bissexual", "Heterossexual"))
dados$ud <- factor(dados$ud, levels = c(1, 2), labels = c("Sim", "Não"))
dados$ac <- factor(dados$ac, levels = c(1, 2), labels = c("Sim", "Não"))

# Visualização do início do dataset formatado
head(dados)
```

```
##   pac id       sex               grp ti  tf cens cd4 cd8           ats   ud
## 1   1 31 Masculino              AIDS  0   0    1  NA  NA Heterossexual  Não
## 2   2 22  Feminino HIV+ Assintomático  0 378    0 132 715 Heterossexual  Não
## 3   3 32 Masculino              AIDS  0  84    1  75 315 Heterossexual  Não
## 4   4 36 Masculino HIV+ Assintomático  0 109    0  NA  NA Heterossexual  Não
## 5   5 34 Masculino HIV+ Assintomático  0 134    1  NA  NA          <NA> <NA>
## 6   6 29 Masculino HIV+ Assintomático  0 338    0  NA  NA   Homossexual  Não
```

```
##     ac
## 1  Não
## 2  Não
## 3  Não
## 4  Não
## 5 <NA>
## 6  Não
```

```r
# Verificando as transformações
str(dados)
```

```
## 'data.frame':    133 obs. of  12 variables:
##  $ pac : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ id  : int  31 22 32 36 34 29 29 22 38 32 ...
##  $ sex : Factor w/ 2 levels "Masculino","Feminino": 1 2 1 1 1 1 2 1 1 2 ...
##  $ grp : Factor w/ 4 levels "HIV-","HIV+ Assintomático",..: 4 2 4 2 2 2 3 4 4 1 ...
##  $ ti  : num  0 0 0 0 0 0 0 0 0 0 ...
##  $ tf  : num  0 378 84 109 134 338 311 0 182 77 ...
##  $ cens: int  1 0 1 0 1 0 0 0 1 1 ...
##  $ cd4 : num  NA 132 75 NA NA NA 73 58 NA NA ...
##  $ cd8 : int  NA 715 315 NA NA NA 590 775 NA NA ...
##  $ ats : Factor w/ 3 levels "Homossexual",..: 3 3 3 3 NA 1 3 2 1 3 ...
##  $ ud  : Factor w/ 2 levels "Sim","Não": 2 2 2 2 NA 2 2 2 2 2 ...
##  $ ac  : Factor w/ 2 levels "Sim","Não": 2 2 2 2 NA 2 2 2 2 2 ...
```

Questão 2 Faça as seguintes análises:

(i) Construa a tabela de contigência entre as variáveis **sex** e **ud**.

Solução:

Nesse caos, vamos omitir os valores ausentes.

```r
dados_sub <- na.omit(dados[, c("sex", "ud")])
tabela_sex_ud <- table(dados_sub$sex, dados_sub$ud)
# Tabela de contingência com totais e proporções
addmargins(tabela_sex_ud)
```

```
##
##             Sim Não Sum
##   Masculino   8  65  73
##   Feminino    2  32  34
##   Sum        10  97 107
```

```r
# Proporções por linha
prop.table(tabela_sex_ud, margin = 1)
```

```
##
##                    Sim        Não
##   Masculino 0.10958904 0.89041096
##   Feminino  0.05882353 0.94117647
```

```r
# Proporções por coluna
prop.table(tabela_sex_ud, margin = 2)
```

```
##
##                    Sim       Não
##   Masculino 0.8000000 0.6701031
##   Feminino  0.2000000 0.3298969
```

(ii) Verifique se existe associação entre as variáveis **sex** e **ud**.

Solução:

```r
# Teste Qui-quadrado sem correção
teste_qui2 <- chisq.test(tabela_sex_ud, correct = FALSE)

# Frequências esperadas
freq_esperadas <- teste_qui2$expected

# Verificar se alguma frequência esperada < 5
frequencia_baixa <- any(freq_esperadas < 5)

# Se necessário, aplica o teste exato de Fisher
if (frequencia_baixa) {
  teste_final <- fisher.test(tabela_sex_ud)
  metodo <- "Teste exato de Fisher"
  valor_p <- teste_final$p.value
  or <- round(teste_final$estimate, 3)
  ci <- paste0("IC 95% = [",
               round(teste_final$conf.int[1], 3), "; ",
               round(teste_final$conf.int[2], 3), "]")
} else {
  teste_final <- teste_qui2
  metodo <- "Teste Qui-quadrado de Pearson"
  valor_p <- teste_final$p.value
  or <- NA
  ci <- NA
}

# Interpretação automatizada
cat("Método utilizado:", metodo, "\n")
```

```
## Método utilizado: Teste exato de Fisher
```

```r
cat("Valor-p:", round(valor_p, 4), "\n")
```

```
## Valor-p: 0.4978
```

```r
if (valor_p < 0.05) {
  cat("Conclusão: Há evidência estatística de associação entre sexo
      e uso de droga injetável.\n")
} else {
  cat("Conclusão: Não há evidência estatística de associação entre sexo
      e uso de droga injetável.\n")
}
```

```
## Conclusão: Não há evidência estatística de associação entre sexo
##       e uso de droga injetável.
```

```r
if (!is.na(or)) {
  cat("Razão de chances estimada (OR):", or, "\n")
  cat(ci, "\n")
}
```

```
## Razão de chances estimada (OR): 1.958
## IC 95% = [0.361; 19.977]
```

(iii) Determine a razão de chances (Odds ratio) entre as variáveis **sex** e **ud**. Interprete os resultados.

Solução:

```r
library(epitools)
resultado_or <- oddsratio(tabela_sex_ud, method = "wald")
resultado_or$measure  # mostra estimativa e IC 95%
```

```
##            odds ratio with 95% C.I.
##             estimate     lower    upper
##   Masculino 1.000000        NA       NA
##   Feminino  1.969231 0.3951015 9.814869
```

```r
resultado_or$p.value  # valor-p dos testes
```

```
##            two-sided
##             midp.exact fisher.exact chi.square
##   Masculino         NA           NA         NA
##   Feminino   0.4378603    0.4978387  0.4009123
```

A razão de chances de uma mulher usar droga injetável em comparação a um homem é 1.97, ou seja, quase o dobro. No entanto, o intervalo de confiança é muito amplo e inclui o valor 1, o que indica alta incerteza. O valor-p = 0.4978 (teste exato de Fisher) é maior que 0.05, portanto, não há evidência estatística de associação significativa entre sex e ud.

Questão 3 Faça as seguintes análises:

(i) Construa a tabela de contigência entre as variáveis **ud** e **grp**.

Solução:

Vamos omitir os valores ausentes.

```r
# Remover valores ausentes apenas nas variáveis ud e grp
dados_sub3 <- na.omit(dados[, c("ud", "grp")])

# Tabela de contingência entre uso de droga injetável e grupo de risco
tabela_ud_grp <- table(dados_sub3$ud, dados_sub3$grp)
tabela_ud_grp
```

```
##
##       HIV- HIV+ Assintomático ARC AIDS
##   Sim    1                  3   1    5
##   Não   16                 35  18   28
```

(ii) Verifique se existe associação entre as variáveis **ud** e **grp**.

Solução:

```r
# Teste Qui-quadrado sem correção
teste_qui_ud_grp <- chisq.test(tabela_ud_grp, correct = FALSE)

# Verificar frequências esperadas
freq_esperadas <- teste_qui_ud_grp$expected
frequencia_baixa <- any(freq_esperadas < 5)

# Aplicar teste apropriado
if (frequencia_baixa) {
  teste_final <- fisher.test(tabela_ud_grp)
  metodo <- "Teste exato de Fisher"
  valor_p <- teste_final$p.value
} else {
  teste_final <- teste_qui_ud_grp
  metodo <- "Teste Qui-quadrado de Pearson"
  valor_p <- teste_final$p.value
}

# Interpretação automatizada
cat("Método utilizado:", metodo, "\n")
```

```
## Método utilizado: Teste exato de Fisher
```

```r
cat("Valor-p:", round(valor_p, 4), "\n")
```

```
## Valor-p: 0.6919
```

```r
if (valor_p < 0.05) {
  cat("Conclusão: Há evidência estatística de associação entre uso de droga
      injetável e grupo de risco.\n")
} else {
  cat("Conclusão: Não há evidência estatística de associação entre uso de droga
      injetável e grupo de risco.\n")
}
```

```
## Conclusão: Não há evidência estatística de associação entre uso de droga
##       injetável e grupo de risco.
```

(iii) Determine as razão de chances (Odds ratio) entre as variáveis **sex** e **ud**. Interprete os resultados.

Solução:

```r
resultado_or_2 <- oddsratio(tabela_ud_grp, method = "wald")
resultado_or_2$measure  # mostra estimativa e IC 95%
```

```
##      odds ratio with 95% C.I.
##        estimate      lower    upper
##   Sim 1.0000000         NA       NA
##   Não 0.7291667 0.07029608 7.563494
```

```r
resultado_or_2$p.value  # valor-p dos testes
```

```
##      two-sided
##       midp.exact fisher.exact chi.square
##   Sim         NA           NA         NA
##   Não   0.853358    0.6918851  0.5678957
```

A razão de chances estimada para os indivíduos que não usam droga injetável, comparados aos que usam, foi de 0.729, o que sugere uma menor chance de estar em um grupo de risco mais grave. No entanto, o intervalo de confiança inclui 1, indicando alta incerteza na estimativa. Além disso, o valor-p = 0.6919 (do teste exato de Fisher) é muito maior que 0.05, o que indica ausência de evidência estatística de associação entre o uso de droga injetável e o grupo de risco.

Questão 4 Utilizando as covariáveis **id**, **sex**, **grp**, **ats**, **ud** e **ac**, e o modelo de regressão logística, responda:

(i) Faça o ajuste do modelo de regressão logística, considerando a variável **cens** como variável resposta e todos os efeitos principais de todas as covariáveis.

Solução:

```r
# Selecionar apenas as colunas relevantes e remover valores ausentes
dados_modelo <- na.omit(dados[, c("cens", "id", "sex", "grp", "ats", "ud", "ac")])

# Ajuste do modelo de regressão logística
```

```r
modelo_logistico <- glm(cens ~ id + sex + grp + ats + ud + ac,
                        data = dados_modelo,
                        family = binomial())

# Resumo do modelo
summary(modelo_logistico)
```

```
##
## Call:
## glm(formula = cens ~ id + sex + grp + ats + ud + ac, family = binomial(),
##     data = dados_modelo)
##
## Coefficients:
##                        Estimate Std. Error z value Pr(>|z|)
## (Intercept)           -3.20401    1.77719  -1.803  0.07141 .
## id                    -0.02767    0.03233  -0.856  0.39208
## sexFeminino            0.40218    0.74875   0.537  0.59117
## grpHIV+ Assintomático  0.42569    1.20143   0.354  0.72310
## grpARC                 1.97936    1.19506   1.656  0.09766 .
## grpAIDS                3.15824    1.16000   2.723  0.00648 **
## atsBissexual          -0.03785    0.76586  -0.049  0.96058
## atsHeterossexual      -0.39087    0.77950  -0.501  0.61606
## udNão                  0.51389    1.44827   0.355  0.72272
## acNão                  0.81411    1.79776   0.453  0.65066
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 115.264  on 104  degrees of freedom
## Residual deviance:  92.667  on  95  degrees of freedom
## AIC: 112.67
##
## Number of Fisher Scoring iterations: 5
```

(ii) Avalie a qualidade do modelo ajustado.

Solução:

1. Vamos Comparar a Null Deviance com a Residual Deviance:

- Null deviance: 115.264
- Residual deviance: 92.667

a diferença entre as duas é (TRV): 22.597 com 9 Graus de liberdade da diferença, pois. $104 - 95 = 9$ Vamos testar essa diferença com um Teste da Razão de Verossimilhança (TRV):

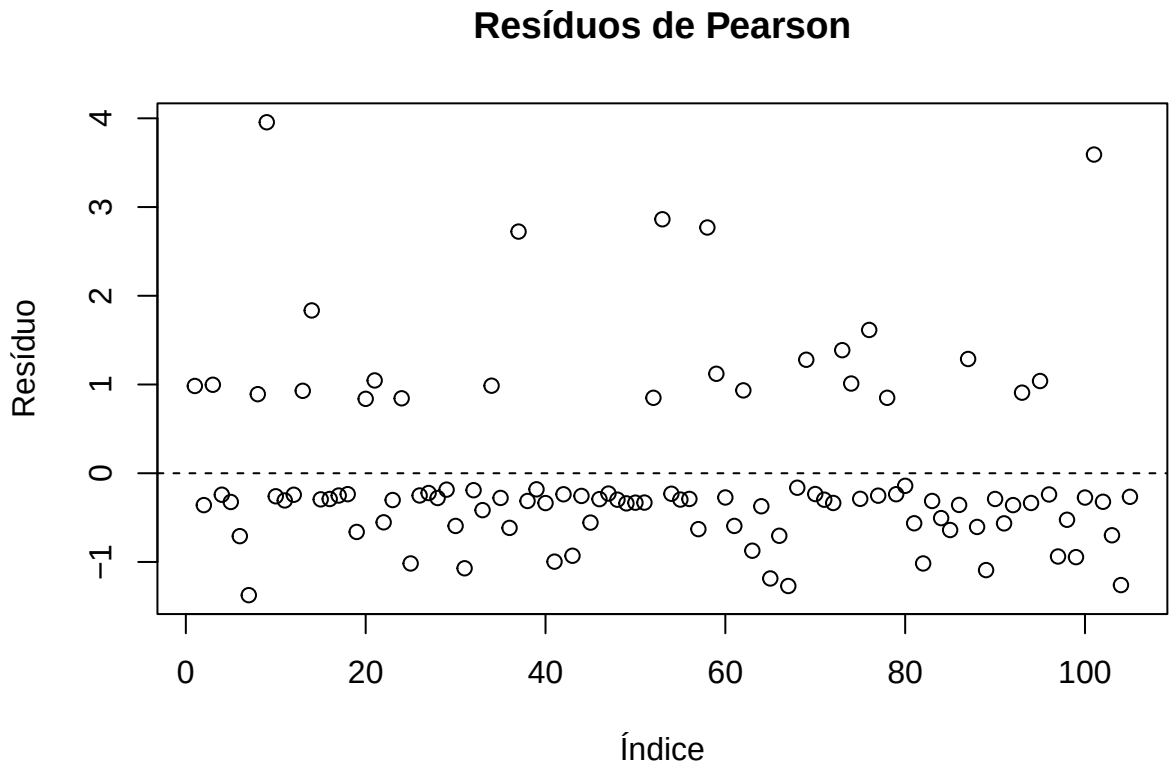
```r
pchisq(22.597, df = 9, lower.tail = FALSE)
```

```
## [1] 0.00716772
```

Como o valor-p é menor que 0.05, o modelo ajustado é significativamente melhor que o modelo nulo. Note também que a AIC do modelo é 112.67. Valores de AIC menores indicam modelos melhores (quando comparando com outros modelos). Sozinho, o AIC indica apenas que o modelo é parcimonioso, mas deve ser comparado com modelos alternativos.

Vamos verificar os resíduos:

```
# Resíduos de Pearson
plot(residuals(modelo_logistico, type = "pearson"),
     main = "Resíduos de Pearson", ylab = "Resíduo", xlab = "Índice")
abline(h = 0, lty = 2)
```



O gráfico dos resíduos de Pearson mostra que a maioria das observações se ajusta bem ao modelo. No entanto, algumas observações têm resíduos elevados, acima de 2, indicando que o modelo pode não estar ajustando perfeitamente todos os casos. Apesar disso, não há padrão sistemático nos resíduos, o que sugere que o modelo está, em geral, bem especificado.

```
library(ResourceSelection)
hoslem.test(dados_modelo$cens, fitted(modelo_logistico), g = 10)
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  dados_modelo$cens, fitted(modelo_logistico)
## X-squared = 5.201, df = 8, p-value = 0.7359
```

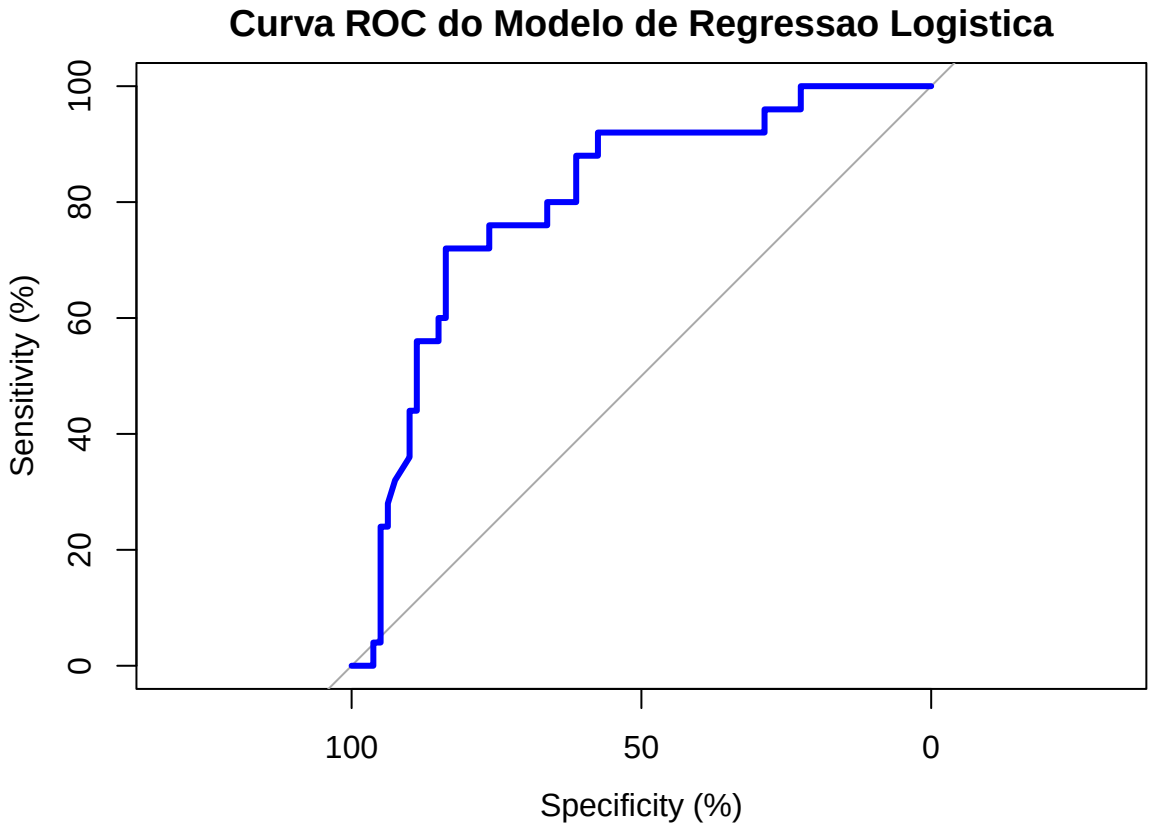
Como o valor-p do teste de Hosmer e Lemeshow foi 0.7359, que é maior que 0.05, não há evidência de falta de ajuste do modelo. Portanto, o modelo de regressão logística apresenta um bom ajuste aos dados, segundo esse critério.

(iii) Avalie a predição do modelo ajustado utilizando a curva ROC.

Solução:

```
library(pROC)
# Curva ROC e AUC
roc_modelo <- roc(dados_modelo$cens, fitted(modelo_logistico), percent = TRUE)

# Plotar curva ROC
plot(roc_modelo, main = "Curva ROC do Modelo de Regressao Logistica",
     col = "blue", lwd = 3)
```




```
auc(roc_modelo)
```

```
## Area under the curve: 80.22%
```

A curva ROC gerada a partir do modelo ajustado mostra a capacidade de discriminar corretamente entre os indivíduos com `cens = 1` e `cens = 0`. A **AUC (Área sob a Curva)** foi de **80.2%**, o que indica que o modelo possui muito boa capacidade preditiva e poder discriminativo.

(iv) interprete os resultados do modelo ajustado.

Solução:

Com base em todos os critérios — ajuste global (TRV e AIC), qualidade de ajuste (Hosmer e Lemeshow), comportamento dos resíduos e curva ROC —, podemos concluir que o modelo de regressão logística ajustado apresenta bom desempenho global e preditivo. Ele é estatisticamente significativo, bem ajustado aos dados e possui capacidade razoável de discriminação entre os indivíduos com cens = 0 e cens = 1.