Análise de Dados Categorizados

Prática 3: Regressão Logística Binária

Prof. Dr. Márcio Augusto

Contents

1 Modelo de Regressão Logística Simples

$$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = \alpha + \beta x. \tag{1}$$

$$\pi(x) = \frac{e^{\alpha+\beta x}}{1+e^{\alpha+\beta x}}. \tag{2}$$

4.1.2 Odds Ratio

$$\frac{\pi(x)}{1-\pi(x)} = \exp(\alpha+\beta x) = e^{\alpha}(e^{\beta})^{x}.$$

Exercício 1: Regreção Logística Simples

O arquivo `dcc.txt` apresenta os dados de um estudo sobre doença coronária cardíaca. Os dados são originado de Hosmer e Lemeschow (2000), e trata-se de uma amostra de 100 pessoas que foram submetidos a um eletrocardiograma (ecg). A variável dependente (resposta) é a ocorrência ou não (1 ou 0) de doença coronária cardíaca (cdc) e as covariáveis são:

- Idade, em anos
- Sexo (0 se feminio e 1 se masculino)
- ecg (0 se ecg $< 0,1$, 1 se $0,1 \leq$ ecg $< 0,2$ e 2 se ecg $\geq 0,2$).

Vamos ajustar um modelo de regressão logística para verificar o efeito da variável idade na ocorrência de doença coronária cardíaca.

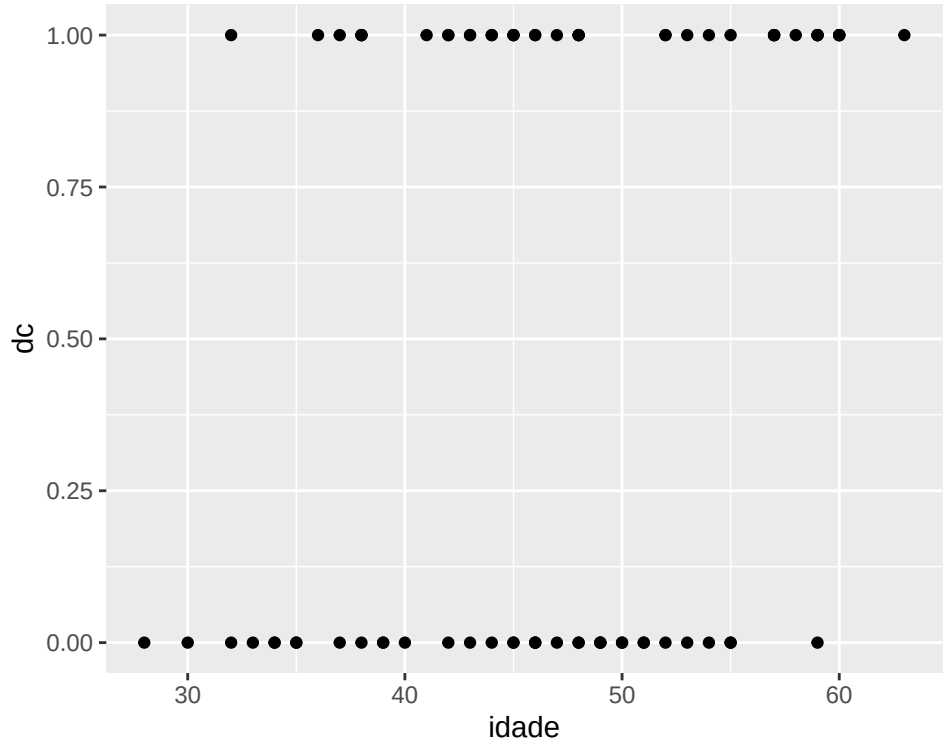
```
##Pacotes necessários
library(ggplot2)
library(mfx)
library(ResourceSelection)
library(performance)
library(caret)
library(pROC)
library(hnp)
```

```
# lendo o banco de dados
dados<-read.table("dcc.txt",h=T)
summary(dados)
```

```
##        dc              sexo              ecg              idade
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.0000   Min.   :28.00
##  1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:0.0000   1st Qu.:41.25
##  Median :1.0000   Median :1.0000   Median :1.0000   Median :46.50
##  Mean   :0.5256   Mean   :0.5769   Mean   :0.7436   Mean   :46.90
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:53.75
##  Max.   :1.0000   Max.   :1.0000   Max.   :2.0000   Max.   :63.00
```

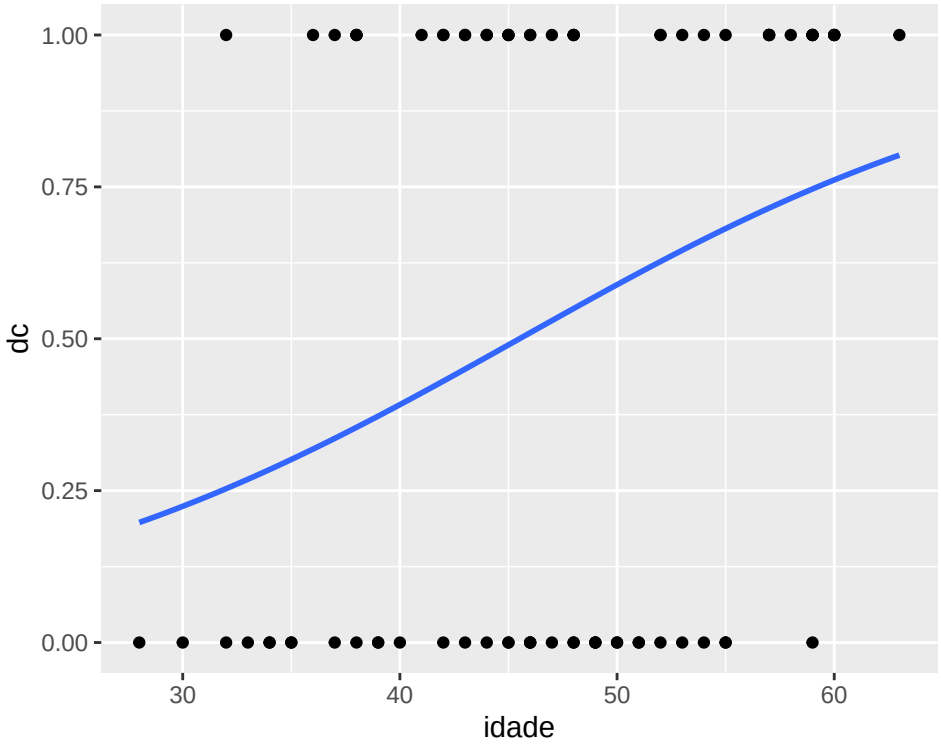
Observa-se na figura abaixo a dispersão dos "casos" e dos "nao-casos" de doença coronária cardíaca relacionando-se com a variável idade.

```
ggplot(dados, aes(x=idade, y=dc)) +
  geom_point()
```




```
ggplot(dados, aes(x=idade, y=dc)) +
  geom_point() +
  stat_smooth(method="glm", method.args=list(family="binomial"), se=FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



Vamos ajustar um modelo de regressão logística com a variável dependente dcc e a variável indepedente idade.

```
m1<-glm(dc~ idade, family = binomial(link="logit"), data = dados)
summary(m1)
```

Ajuste do Modelo

```
##
## Call:
## glm(formula = dc ~ idade, family = binomial(link = "logit"),
##     data = dados)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.64310    1.41838  -2.568  0.01021 *
## idade        0.08006    0.02993   2.675  0.00748 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 107.926  on 77  degrees of freedom
## Residual deviance:  99.902  on 76  degrees of freedom
## AIC: 103.9
##
## Number of Fisher Scoring iterations: 4
```

- Se observa o intercepto com o valor de -3,64, sendo que para a análise aqui proposta da relação entre

'cd' e 'idade' não obtém-se um significado prático para este resultado.

- No entanto, a variável de interesse é idade, que no modelo de regressão obteve o coeficiente de 0,08.
- Pelo fato de ser positivo informa que quando a idade se eleva, elevam-se as chances de ocorrência de doença coronária cardíaca.
- De igual forma, nota-se que há significância estatística (p=0,007) na utilização da variável idade para o modelo, mostrando que possui importância ao modelo de regressão proposto.

```
anova(m1, test = "Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: dc
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                     77    107.926
## idade  1   8.0236        76     99.902 0.004617 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Análise dos resíduos

```
dev<-residuals(m1, type='deviance');
#dev
QL<-sum(dev^2);
p1<-1-pchisq(QL,1);
cbind(QL,p1)
```

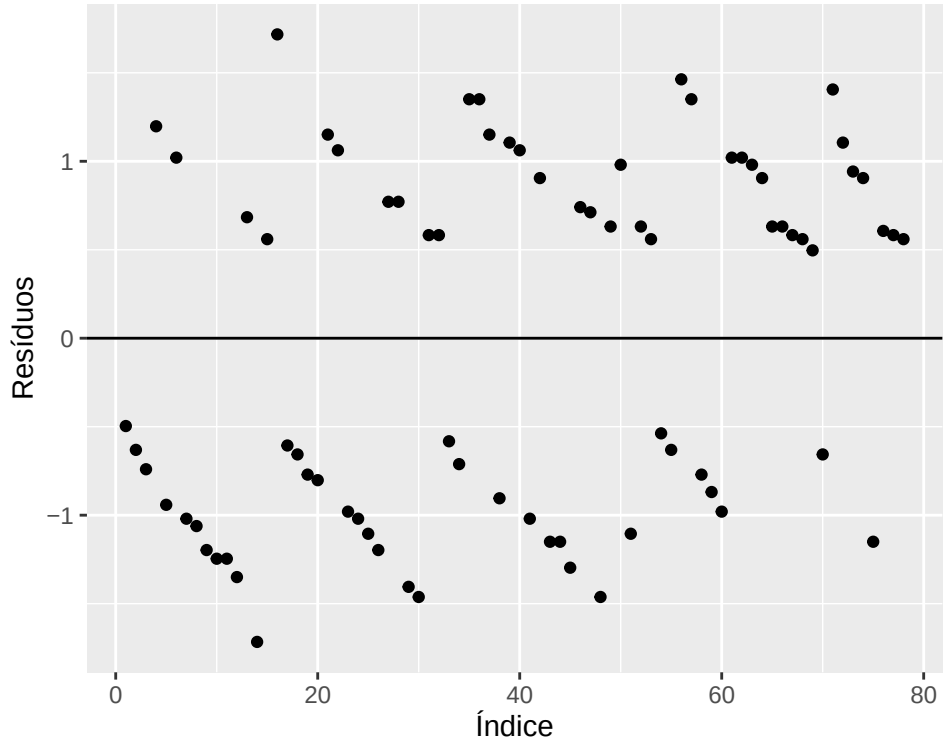
```
##            QL p1
## [1,] 99.90212  0
```

```
rpears<-residuals(m1, type='pearson');
#rpears
QP<-sum(rpears^2);
p2<-1-pchisq(QP,1);
cbind(QP,p2)
```

```
##            QP p2
## [1,] 76.73771  0
```

```
## grafico dos resíduos de pearson
resp <- data.frame(indice = 1:nrow(dados),
                   residuos = residuals(m1, type = "pearson"))
ggplot(resp, aes(x = indice, y = residuos)) +
  geom_point() +
```

```
  geom_hline(yintercept = 0) +
  labs(x = "Índice", y = "Resíduos")
```

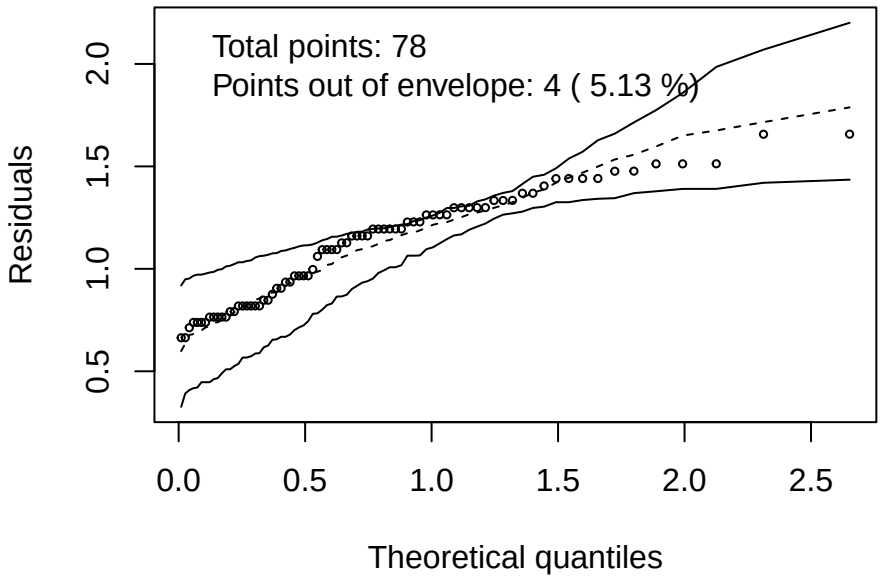



```
#Envelope Simulado

g <- hnp(m1, print.on=TRUE, plot=FALSE)
```

```
## Binomial model
```

```
plot(g)
```



O teste Hosmer e Lemeshow O teste de Hosmer e Lemeshow é utilizado para demonstrar a qualidade do ajuste do modelo, ou seja, se o modelo pode explicar os dados observados.

A hipótese nula H_0 deste teste é a de que as proporções observadas e esperadas são as mesmas ao longo da amostra.

```
# primeiro método, usando o pacote ResourceSelection
h <- hoslem.test(dados$dc,fitted(m1),g=10)
h
```

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  dados$dc, fitted(m1)
## X-squared = 15.244, df = 8, p-value = 0.05458
```

```
# Segundo método, usando o pacote
performance_hosmer(m1, n_bins = 10)
```

```
## # Hosmer-Lemeshow Goodness-of-Fit Test
##
##   Chi-squared: 15.244
##            df:  8
##       p-value:  0.055
```

```
## Summary: model seems to fit well.
```

Estimando a Razão de Chances O modelo estimado fica expresso por

$$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -3,6431 + 0,08 \times \text{idade}. \tag{3}$$

e a odds estimada é

$$\frac{\pi(x)}{1-\pi(x)} = \exp(-3,6431 + 0,08 \times \text{idade})$$

O modelo de regressão logística, porém, traz os resultados dos estimadores na forma logarítma, ou seja, o log das chances da variável idade no modelo é 0,08. No entanto, para uma interpretação mais enriquecida da relação da idade com o dc é necessária a transformação deste coeficiente, ou seja, que seja efetuada a exponenciação da(s) variavel(eis) da regressão. Assim, obtém-se a razão das chances (OR - Odds Ratio em inglês) para as variáveis independentes.

```
coef(m1) ## valores dos coeficientes
```

```
## (Intercept)       idade
## -3.64310033  0.08005666
```

```
confint(m1) ## intervalo de confiança para os coeficientes
```

```
##                   2.5 %     97.5 %
## (Intercept) -6.59302767 -0.9783453
## idade        0.02390432  0.1423646
```

```
exp(coef(m1)) ## razões de chances
```

```
## (Intercept)       idade
##  0.02617108  1.08334844
```

```
exp(cbind(OR=coef(m1), confint(m1)))
```

```
##                      OR       2.5 %    97.5 %
## (Intercept) 0.02617108 0.001369886 0.3759326
## idade       1.08334844 1.024192318 1.1529970
```

Usando o pacote `mfx`

```
### Usando o pacote mfx
#require(mfx)
logitor(dc ~ idade,data = dados)
```

```
## Call:
## logitor(formula = dc ~ idade, data = dados)
##
## Odds Ratio:
##       OddsRatio Std. Err.      z    P>|z|
## idade  1.083348  0.032424 2.6749 0.007476 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- A razão de chances da variável idade foi de 1,0833. Portanto, para cada variação de um ano nada idade, aumentam-se 8,33% as chances da ocorrência de doença coronária cardíaca.

Predição o modelo é utilizado para construção da predição de todos os valores das idades de todos os indivíduos desta amostra.

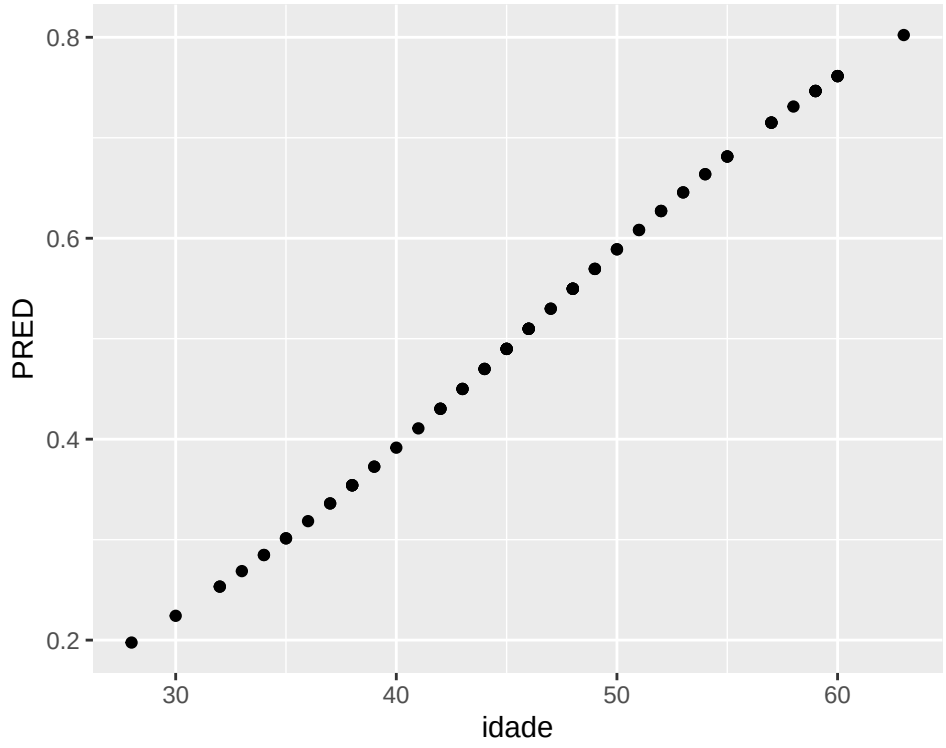
Para isto, será criada um novo objeto contendo somente a variável dependente do modelo (idade) e em sequida, é criada nova coluna constando os valores preditos.

Assim, pode ser plotado um gráfico completo com todas as probabilidades desta base de dados:

```
# Filtrando a idade dos indivíduos


# Criando campo de predição para cada idade dos indivíduos
dados$PRED=predict(m1, newdata=dados, type="response")

# Plotando a probabilidade predita pelo modelo
require(ggplot2)
ggplot(dados, aes(x=idade, y=PRED)) +
  geom_point()
```



Poder preditivo do modelo

```
dados$pdata <- as.factor(
    ifelse(
        predict(m1,
                newdata = dados,
                type = "response")
        >0.5,"1","0"))

confusionMatrix(dados$pdata, as.factor(dados$dc), positive="1")
```

```
## Confusion Matrix and Statistics
```

```
##
##           Reference
## Prediction  0  1
##          0 18 16
##          1 19 25
##
##                Accuracy : 0.5513
##                  95% CI : (0.4344, 0.6641)
##     No Information Rate : 0.5256
##     P-Value [Acc > NIR] : 0.3677
##
##                   Kappa : 0.0966
##
##  Mcnemar's Test P-Value : 0.7353
##
##             Sensitivity : 0.6098
##             Specificity : 0.4865
##          Pos Pred Value : 0.5682
##          Neg Pred Value : 0.5294
##              Prevalence : 0.5256
##          Detection Rate : 0.3205
##    Detection Prevalence : 0.5641
##       Balanced Accuracy : 0.5481
##
##        'Positive' Class : 1
##
```

A matriz de confusão exibe as seguintes informações:

- True Positives (TP): 18, o número de não doentes previstos corretamente (classe 0)
- True Negatives (TN): 25, o número de doentes previstos corretamente (classe 1)
- False Negatives (FN): 19, o número de não doentes incorretamente previstos como doentes (classe 1)
- False Positives (FP): 16, o número de doentes incorretamente previstos como não doentes (classe 0)

Alem dessa informaões, são apresentadas as seguintes estatísticas:

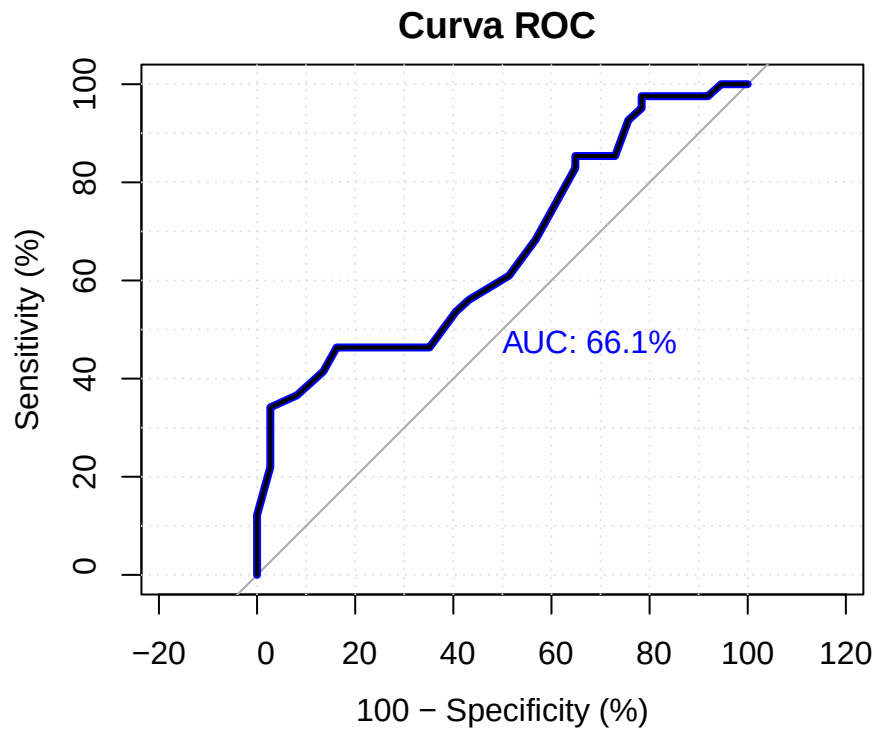
- Accuracy: 0,5513 (55,13%), a proporção de previsões corretas (tanto verdadeiros positivos quanto verdadeiros negativos) entre o número total de casos.
- No Information Rate (NIR): 0.5256, a precisão que poderia ser obtida prevendo sempre a classe majoritária (classe 1 neste caso).
- Mcnemar's Test P-Value: 0.7353 , o valor de p para um teste estatístico que compara o número de falsos positivos e falsos negativos. Um valor de p alto (tipicamente maior que 0,05) indica que não há diferença significativa entre o número de falsos positivos e falsos negativos.

Outras métricas apresentadas:

- Sensitivity (Taxa de recall ou verdadeiro positivo): 0.6098 , a proporção de casos positivos reais (doentes) que foram identificados corretamente pelo modelo.
- Specificity: 0.4865, a proporção de casos negativos reais (não doentes) que foram identificados corretamente pelo modelo.
- Positive Predictive Value (PPV): 0.5682, a proporção de previsões positivas (doentes previstos) que foram realmente positivas (doentes verdadeiros).

- Negative Predictive Value (NPV): 0.5294, a proporção de previsões negativas (não doentes previstos) que foram realmente negativas (verdadeiros não doentes).
- Prevalence: 0.5256, a proporção de casos verdadeiramente positivos (doentes) no conjunto de dados.
- Detection Rate: 0.3205, a proporção de casos verdadeiramente positivos que foram detectados corretamente pelo modelo.
- Detection Prevalence: 0.5641, a proporção de casos previstos como positivos (doentes) pelo modelo.
- Balanced Accuracy: 0.5481, a média de sensibilidade e especificidade, fornecendo uma avaliação equilibrada do desempenho do modelo em ambas as classes.

```
#Curva ROC
roc1 = roc(as.factor(dados$dc), fitted(m1),
           percent=TRUE,
           plot=TRUE, smoothed = TRUE, grid=TRUE,
           print.auc=TRUE, show.thres=TRUE, col="blue",lwd =4,
           legacy.axes=TRUE,   main="Curva ROC " )
roc2 <- roc(as.factor(dados$dc), fitted(m1),
             plot=TRUE, add=TRUE, percent=roc1$percent)
```



2 Modelo de Regressão Logística Multipla

Vamos ajustar um modelo de regressão logística multipla para investigar a associação do sexo, da idade e do ecg com a doença coronária cardíaca.

Modelo 1 Iniciamos ajustando um modelo mais completo, com os efeitos principais e com as interações.

```
ajust1<-glm(dc ~sexo+ecg+idade+sexo*ecg+sexo*idade+ecg*idade+sexo*ecg*idade,
            family=binomial(link="logit"), data=dados)
```

```r
summary(ajust1)
```

```
##
## Call:
## glm(formula = dc ~ sexo + ecg + idade + sexo * ecg + sexo * idade +
##     ecg * idade + sexo * ecg * idade, family = binomial(link = "logit"),
##     data = dados)
##
## Coefficients:
##                 Estimate Std. Error z value Pr(>|z|)
## (Intercept)    -3.994176   3.718019  -1.074    0.283
## sexo            1.241693   4.598831   0.270    0.787
## ecg            -1.137157   4.388678  -0.259    0.796
## idade           0.060987   0.074962   0.814    0.416
## sexo:ecg       -1.257617   5.474010  -0.230    0.818
## sexo:idade     -0.002877   0.095008  -0.030    0.976
## ecg:idade       0.038508   0.087603   0.440    0.660
## sexo:ecg:idade  0.038087   0.115802   0.329    0.742
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 107.926  on 77  degrees of freedom
## Residual deviance:  85.414  on 70  degrees of freedom
## AIC: 101.41
##
## Number of Fisher Scoring iterations: 5
```

```r
anova(ajust1, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: dc
##
## Terms added sequentially (first to last)
##
##
##                Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                             77    107.926
## sexo            1   6.0903        76    101.835 0.013592 *
## ecg             1   6.7558        75     95.080 0.009344 **
## idade           1   8.2684        74     86.811 0.004034 **
## sexo:ecg        1   0.0351        73     86.776 0.851411
## sexo:idade      1   0.0329        72     86.743 0.856049
## ecg:idade       1   1.2211        71     85.522 0.269146
## sexo:ecg:idade  1   0.1084        70     85.414 0.741972
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

- A saída anterior, mostra as *deviances* e suas diferenças, que corresponde ao TRV, associados aos modelos sequenciais ajustados aos dados.
- É possível concluir pela não significância do efeito da interação tripla, visto que o TRV = 0,1084 (valor p = 0,7420 , g.l = 1).

- A mesma conclusão é válida para os efeitos das três interações duplas.
- Assim, o modelo selecionado é aquele com os efeitos principais de sexo, ecg e idade.

Modelo 2 Ajustando o modelo com apenas efeitos principais de sexo, ecg e idade.

```
ajust2<-glm(dc~sexo+ecg+idade, family=binomial(link="logit"), data=dados)
 summary(ajust2)
```

```
##
## Call:
## glm(formula = dc ~ sexo + ecg + idade, family = binomial(link = "logit"),
##     data = dados)
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.64176    1.80614  -3.124  0.00179 **
## sexo         1.35643    0.54645   2.482  0.01306 *
## ecg          0.87320    0.38433   2.272  0.02309 *
## idade        0.09285    0.03509   2.646  0.00815 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 107.926  on 77  degrees of freedom
## Residual deviance:  86.811  on 74  degrees of freedom
## AIC: 94.811
##
## Number of Fisher Scoring iterations: 4
```

```
 anova(ajust2, test="Chisq")
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: dc
##
## Terms added sequentially (first to last)
##
##
##       Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                     77    107.926
## sexo   1   6.0903        76    101.835 0.013592 *
## ecg    1   6.7558        75     95.080 0.009344 **
## idade  1   8.2684        74     86.811 0.004034 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

$$\text{logit}[\pi(x)] = \log\left[\frac{\pi(x)}{1-\pi(x)}\right] = -5,6417 + 1,3564x_{i1} + 0,8732x_{i2} + 0,0928 \times \text{idade}. \tag{4}$$

e a odds estimada é

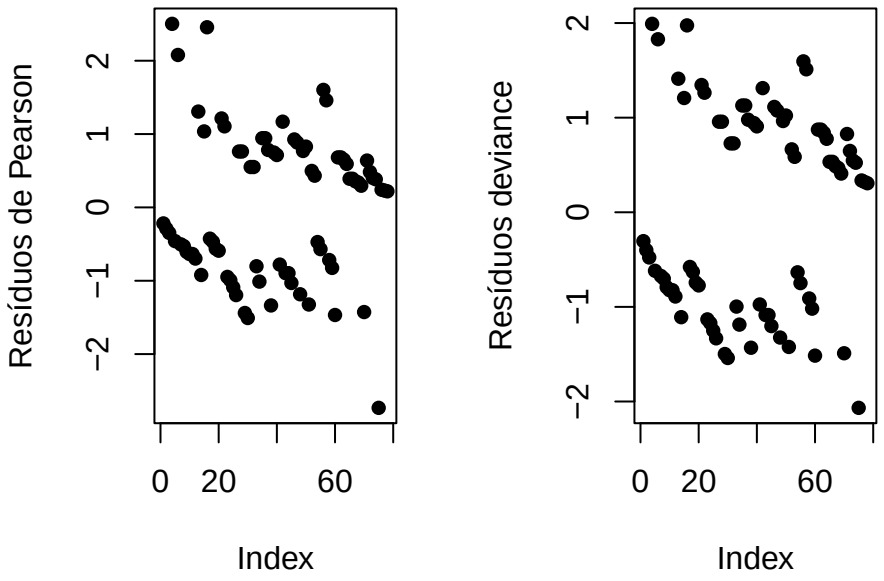
$$\frac{\pi(x)}{1-\pi(x)} = \exp(-5,6417 + 1,3564x_{i1} + 0,8732x_{i2} + 0,0928 \times \text{idade})$$

em que:

- $x_{i1} = 0$ se sexo feminino e 1 se masculino;
- $x_{i2=0}$ se ecg $< 0,1$ e 1 se $0,1 \leq$ ecg $< 0,2$ e 2 se ecg $\geq 0,2$

Avaliando a qualidade do ajuste do modelo

```
dev<-residuals(ajust2, type='deviance');
#dev
rpears<-residuals(ajust2, type='pearson');
#rpears
par(mfrow=c(1,2))
plot(rpears, pch=16, ylab="Resíduos de Pearson")
plot(dev, pch=16, ylab="Resíduos deviance")
```




```
hoslem.test(dados$dc,fitted(ajust2),g=10)
```

Teste HOsmer-Lemeshow

```
##
##  Hosmer and Lemeshow goodness of fit (GOF) test
##
## data:  dados$dc, fitted(ajust2)
## X-squared = 4.4215, df = 8, p-value = 0.8172
```

```
performance_hosmer(ajust2, n_bins = 10)
```

```
## # Hosmer-Lemeshow Goodness-of-Fit Test
##
##   Chi-squared: 4.421
##            df: 8
##       p-value: 0.817
```

```
## Summary: model seems to fit well.
```

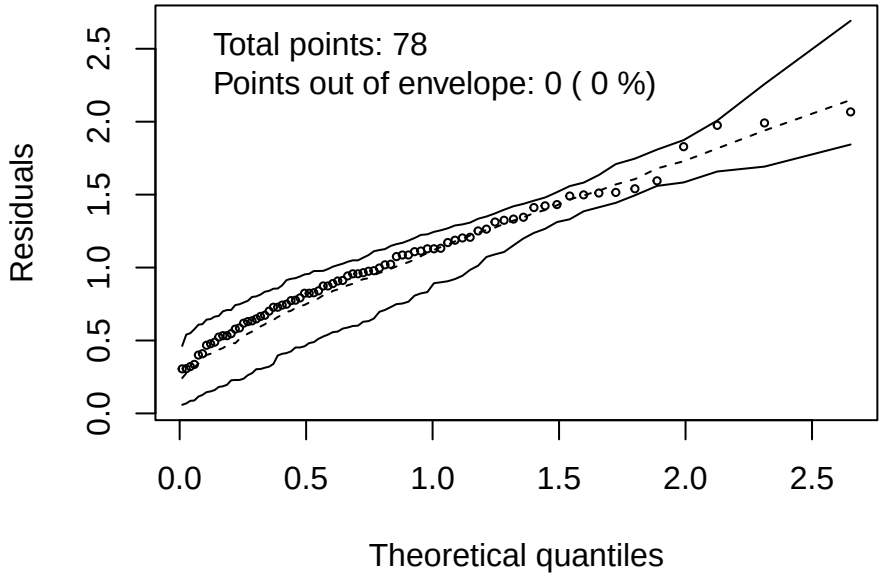
Portanto o teste indica evidências a favor do modelo ajustado.

```
#Envelope Simulado

g <- hnp(ajust2, print.on=TRUE, plot=FALSE)
```

```
## Binomial model
```

```
plot(g)
```



Interpretação

A partir das expressões das probabilidades e chances associadas ao modelo, tem-se que a razão de chances entre pacientes dos sexos masculino e feminino fica :

$$\frac{\exp(\beta_0+\beta_1)}{\exp(\beta_0)} = \exp(\beta_1) \qquad \text{ou} \qquad \frac{\exp(\beta_0+\beta_1+\beta_2)}{\exp(\beta_0+\beta_2)} = \exp(\beta_1)$$

Assim, a estimativa para a razão de chances entre pacientes homens e mulheres, ajustada para ECG, resulta em $\hat{OR} = \exp(\hat{\beta}_1) = \exp(1,3564) = 3,88$ o que indica que os homens apresentam, aproximadamente, 4 vezes mais chance de doença coronária do que às mulheres.

De modo similar, tem-se que a razão de chances entre pacientes com ECG alto e ECG baixo resulta em

$$\frac{\exp(\beta_0+\beta_2)}{\exp(\beta_0)}=\exp(\beta_2) \qquad \text{ou} \qquad \frac{\exp(\beta_0+\beta_1+\beta_2)}{\exp(\beta_0+\beta_2)}=\exp(\beta_2)$$

Logo, a estimativa para a razão de chances entre pacientes com ECG $\geq 0,1$ e com ECG $< 0,1$, ajustada para o sexo é $\hat{OR}=\exp(\hat{\beta}_2)=\exp(0,8732)=2,39$. Portanto, a chance de doença coronária dos pacientes com ECG $\geq 0,1$ foi de 2,39 vezes a dos pacientes com ECG $< 0,1$.

Curva ROC

```
#Curva ROC
roc1 = roc(as.factor(dados$dc), fitted(ajust2),
           percent=TRUE,
           plot=TRUE, smoothed = TRUE, grid=TRUE,
           print.auc=TRUE, show.thres=TRUE, col="blue",lwd =4,
           legacy.axes=TRUE,   main="Curva ROC " )
roc2 <- roc(as.factor(dados$dc), fitted(ajust2),
             plot=TRUE, add=TRUE, percent=roc1$percent)
```