Introdução à Bioestatística

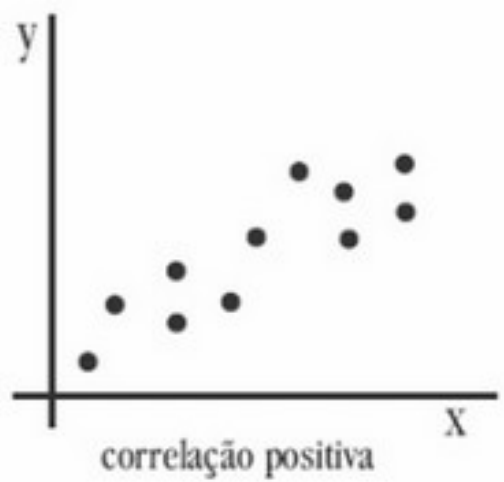
Marcelo Goulart Correia

Instituto Nacional de Cardiologia

May 25, 2015

- Correlação –> Associação entre duas variáveis
- Verificar se duas variáveis mudam em conjunto
- Existir correlação não significa que uma variável seja causa ou consequência da outra

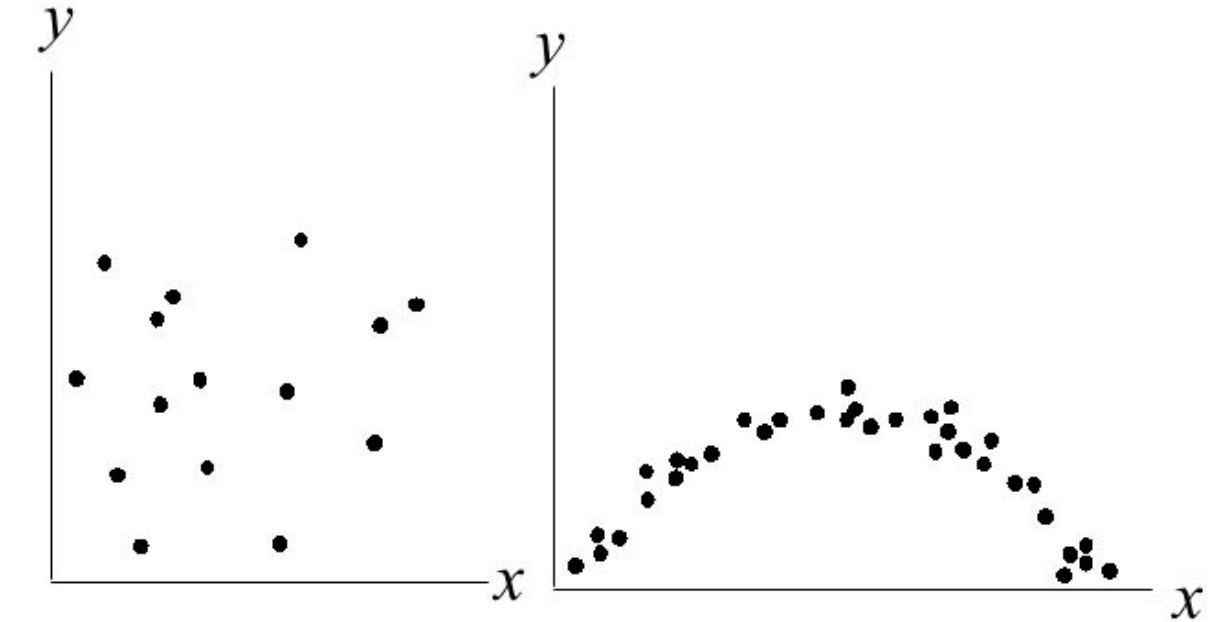
- Existem dois tipos de correlação
  - Correlação linear –> A dispersão das correlações segue uma forma linear
  - Correlação não-linear –> A dispersão das correlações não segue uma forma linear



correlação positiva

correlação negativa

Não correlação



(g) Nenhuma Correlação (h) Correlação Não linear

- Teste de correlação de Pearson
  - Estatística utilizada para calcular o grau de relacionamento entre duas variáveis com distribuição normal

$$r = \frac{N \sum_{i=1}^{k} xy - \sum_{i=1}^{k} x \sum_{i=1}^{k} y}{\sqrt{[N(\sum_{i=1}^{k} x)^2 - \sum_{i=1}^{k} (x^2)] * [N(\sum_{i=1}^{k} y)^2 \sum_{i=1}^{k} (y^2)]}} \quad (1)$$

- Teste de correlação de Spearman
  - Estatística utilizada para calcular o grau de relacionamento entre duas variáveis ordinais ou numéricas não-paramétricas
  - Utiliza postos para seu cálculo

$$\rho = 1 - \frac{6 \sum_{i=1}^{k} d_i^2}{n(n^2 - 1)} \tag{2}$$

- Correlação fraca –> 0 a 0,30
- Correlação moderada –> 0,30 a 0,70
- Correlação forte –> 0,70 a 1

- Exemplo
  - Deseja-se investigar se existe relação entre peso e altura de um determinado grupo de pacientes

Elemento	Altura	Peso
1	182	95
2	175	72
3	170	64
4	180	95
5	183	79
6	160	72
7	156	64
8	157	62
9	168	67
10	175	75
11	176	75
12	190	96

- r = 0,8257 (Valor crítico = 0,576 com GL = 10)
- p = 0,0009
- Existe correlação significativa entre as medidas para um nível de significância de 5%

- Exemplo
  - Deseja-se investigar se existe relação entre as notas e o QI dos alunos

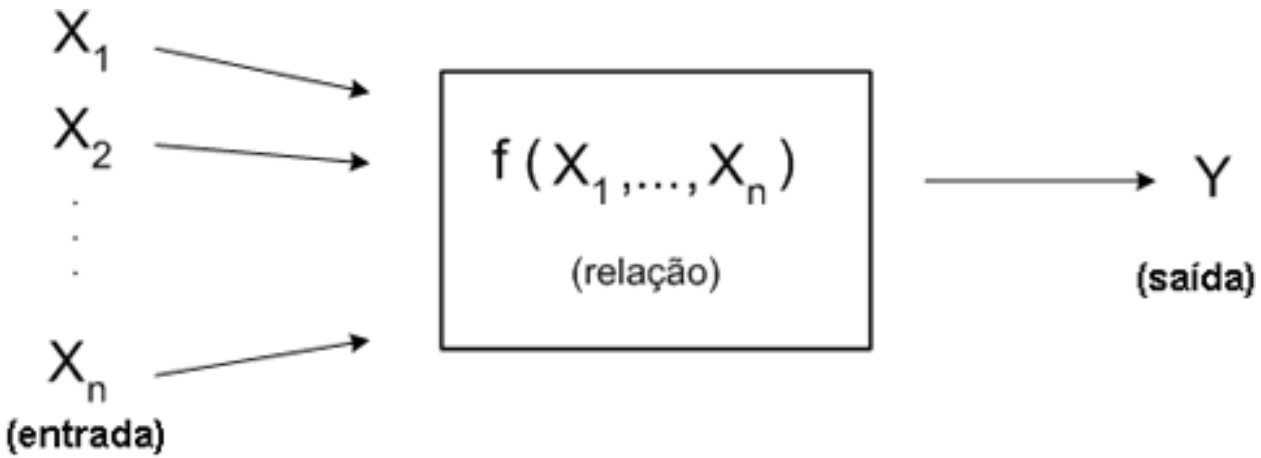
Elemento	Nota	QI
1	8	127
2	9,5	149
3	10	150
4	9,1	135
5	6,5	122
6	9	129
7	9,5	142
8	5,2	100
9	9,1	136
10	9,3	139

- r = 0,9939 (Valor crítico = 0,648)
- p = 0,000000005
- Existe correlação significativa entre as medidas para um nível de significância de 5%

- Exercício

- Correlação –> O quão bem x e y variam em conjunto
- Regressão –> Encontrar a reta que melhor prevê y em função de x

- Verificar se duas ou mais variáveis são relacionadas
- Utiliza-se de modelagem matemática
- De forma ilustrada um modelo funciona da seguinte maneira:



- Os modelos mais utilizados são:
  - Regressão linear –> Variável resposta é numérica
  - Regressão logística –> Variável resposta binária, nominal ou ordinal

- Principais objetivos de uma análise de regressão:
  - Predição
  - Seleção de variáveis
  - Estimação de parâmetros
  - Inferência

- O modelo de regressão linear é representado pela fórmula

$$\boxed{Y = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + ... + \hat{\beta}_n x_n + \varepsilon_i} \quad (3)$$

- $\hat{\beta}_0$ é o intercepto (ou constante) do modelo
- $\hat{\beta}_1, ..., \hat{\beta}_n$ são os coeficientes da regressão
- $x_1, ..., x_n$ são os valores das variáveis explicativas
- ε_i são os erros aleatórios independentes
- Os parâmetros são estimados através do método de mínimos quadrados

- Pressupostos para regressão linear:
  - Linearidade
  - Independência estatística dos erros
  - Homocedasticidade
  - Normalidade

$$
\begin{cases}
n\widehat{\beta}_0 + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i2} + \ldots + \widehat{\beta}_p \sum_{i=1}^{n} x_{ip} = \sum_{i=1}^{n} Y_i \\
\widehat{\beta}_0 \sum_{i=1}^{n} x_{i1} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{i1}^2 + \widehat{\beta}_2 \sum_{i=1}^{n} x_{i1}x_{i2} + \ldots + \widehat{\beta}_p \sum_{i=1}^{n} x_{i1}x_{ip} = \sum_{i=1}^{n} x_{i1}Y_i \\
\vdots \\
\widehat{\beta}_0 \sum_{i=1}^{n} x_{ip} + \widehat{\beta}_1 \sum_{i=1}^{n} x_{ip}x_{i1} + \widehat{\beta}_2 \sum_{i=1}^{n} x_{ip}x_{i2} + \ldots + \widehat{\beta}_p \sum_{i=1}^{n}{}^{2}_{ip} = \sum_{i=1}^{n} x_{ip}Y_i.
\end{cases}
$$

Figure : Método dos mínimos quadrados

- Exemplo
  - Investigar a relação entre o tempo que o indivíduo leva para reagir a um certo estimulo e algumas de suas características tais como sexo, idade e acuidade visual

ID	Tempo	Sexo	Idade	Acuidade	ID	Tempo	Sexo	Idade	Acuidade
1	96	M	20	90	11	109	M	30	90
2	92	F	20	100	12	100	F	30	80
3	106	M	20	80	13	112	F	35	90
4	100	F	20	90	14	105	F	35	80
5	98	F	25	100	15	118	M	35	70
6	104	M	25	90	16	108	M	35	90
7	110	M	25	80	17	113	F	40	90
8	101	F	25	90	18	112	F	40	90
9	116	F	30	70	19	127	M	40	60
10	106	M	30	90	20	117	M	40	80

- Exemplo - Regressão linear simples
- Y:Tempo de reação e X:Idade
  - $n = 20, \Sigma y_i = 2150, \Sigma x_i = 600, \Sigma x_i y_i = 65400, \Sigma xi^2 = 19000$
  - $\hat{\beta}_1 = \frac{65400 + 20 * 30 * 107,5}{19000 - 20 * 30^2} = 0,90$
  - $\hat{\beta}_0 = 107,5 - 0,90 * 30 = 80,50$
  - $Y = 80,5 + 0,90 * x_i$
  - x_i = Idade

```
Call:
lm(formula = Tempo ~ Idade, data = RegLin)

Residuals:
    Min      1Q  Median      3Q     Max
-7.500  -4.125  -0.750   2.625  10.500

Coefficients:
            Estimate Std. Error t value Pr(>|t|)
(Intercept)  80.5000     5.4510  14.768 1.67e-11 ***
Idade         0.9000     0.1769   5.089 7.66e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 5.593 on 18 degrees of freedom
Multiple R-squared:  0.5899,    Adjusted R-squared:  0.5672
F-statistic:  25.9 on 1 and 18 DF,  p-value: 7.662e-05
```

- Resíduos = $Y_{Real} - Y_{Predito}$
- Coeficientes:
  - Estimativa = Betas da regressão (Coeficientes)
  - Erro padrão dos Betas
  - Valor de t = Estatística para determinar a contribuição de uma dada variável na variável resposta
  - Valor de p = Determina o nível de significância da estatística t
- Erro padrão dos resíduos
- R^2
  - Múltiplo = Representa a proporção da variabilidade de Y explicada pelas variáveis regressoras
  - Ajustado = É a medida anterior ajustada pelo grau de liberdade
- Estatística F = Verifica se há relação linear entre a variável resposta (Y) e as variáveis explicatívas (X_n)

- Estatística t $= \frac{\hat{\beta}_j}{\sqrt{\sigma^2 \hat{C}_{jj}}} \sim t_{(n-p-1)}$
- R^2 múltiplo = SQR/SQT
- $R^2_\alpha = 1 - \frac{n-1}{n-p}(1 - R^2)$
- Estatística F

Fonte de variação	Graus de liberdade	Soma dos quadrados	Quase variâncias	F0
Regressão	p	SQR	QMR = SQR / p	F0 = QMR / QME
Erro (Resíduo)	N-p	SQE	QME = SQE / n-p-1	
Total	n-1	SQT		

- Exemplo
  - Os coeficientes contribuem para a variável resposta
  - As medidas de R^2 mostram que existem correlação moderada na linha ajustada pelo modelo
  - Existe relação linear entre as variáveis analisadas

- Exemplo - Regressão linear múltipla
- Y:Tempo de reação e X:Idade, Sexo e Acuidade

```
Call:
lm(formula = Tempo ~ Sexo + Idade + Acuidade, data = RegLin)

Residuals:
    Min      1Q  Median      3Q     Max
-8.1113 -0.9892  0.8090  2.1108  4.5067

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) 119.84744   10.41492  11.507 3.77e-09 ***
Sexo[T.M]     2.79153    1.67600   1.666 0.115249
Idade         0.67922    0.12297   5.523 4.63e-05 ***
Acuidade     -0.40141    0.09376  -4.281 0.000573 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.53 on 16 degrees of freedom
Multiple R-squared:  0.8548,    Adjusted R-squared:  0.8275
F-statistic: 31.39 on 3 and 16 DF,  p-value: 6.165e-07
```

- Exemplo
  - A variável Sexo não contríbui para a varável resposta
  - As medidas de R^2 mostram que existem correlação forte na linha ajustada pelo modelo
  - Existe relação linear entre as variáveis analisadas
  - $Y = 119,85 + 2,79 * Idade + 0,68 * Idade - 0,40 * Acuidade$

- Exercício

- O modelo de regressão logística para variável resposta binária é representado pela fórmula

$$p = \frac{e^{\alpha+\hat{\beta}_1 x_1+...+\hat{\beta}_k x_k}}{1 + e^{\alpha+\hat{\beta}_1 x_1+...+\hat{\beta}_k x_k}} \quad (4)$$

- Transformação logit
- α é o intercepto (ou constante) do modelo
- $\hat{\beta}_1, ..., \hat{\beta}_n$ são os coeficientes da regressão
- $x_1, ..., x_n$ são os valores das variáveis explicativas
- Os parâmetros são estimados através do método de máxima verossimilhança

- Exemplo
- Analisar o efeito de variáveis como o valor da nota do exame de aptidão, a média do valor da nota escolar e o prestígio da escola, na aprovação em um vestibular
- Y:Passou no vestibular? e X:Nota do exame de aptidão (NEA), Média Escolar e Prestígio

```
Call:
glm(formula = aprovado ~ NEA + media + prestigio, family = "binomial",
    data = mydata)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-1.6268  -0.8662  -0.6388   1.1490   2.0790

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     -3.989979   1.139951  -3.500 0.000465 ***
NEA              0.002264   0.001094   2.070 0.038465 *
media            0.804038   0.331819   2.423 0.015388 *
prestigio[T.2]  -0.675443   0.316490  -2.134 0.032829 *
prestigio[T.3]  -1.340204   0.345306  -3.881 0.000104 ***
prestigio[T.4]  -1.551464   0.417832  -3.713 0.000205 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 499.98  on 399  degrees of freedom
Residual deviance: 458.52  on 394  degrees of freedom
AIC: 470.52

Number of Fisher Scoring iterations: 4
```

```
Analysis of Deviance Table

Model 1: aprovado ~ NEA + media + prestigio
Model 2: aprovado ~ 1

  Resid. Df Resid. Dev Df Deviance  Pr(>Chi)
1       394     458.52
2       399     499.98 -5  -41.459 7.578e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

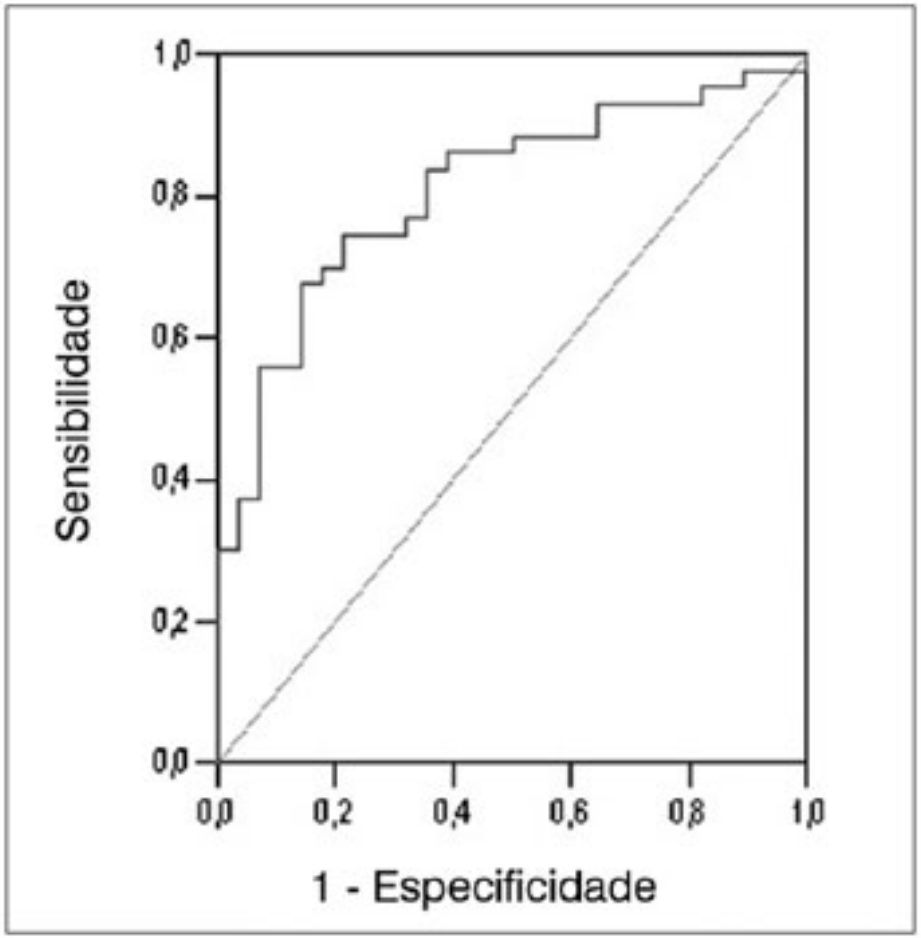
	OR	2.5 %	97.5 %
(Intercept)	0.0185001	0.001889165	0.1665354
NEA	1.0022670	1.000137602	1.0044457
media	2.2345448	1.173858216	4.3238349
prestigio[T.2]	0.5089310	0.272289674	0.9448343
prestigio[T.3]	0.2617923	0.131641717	0.5115181
prestigio[T.4]	0.2119375	0.090715546	0.4706961

- Resíduos da desviância = Calcular a discordância entre a resposta real e a calculada através da função de verossimilhança
- Coeficientes:
  - Estimativa = Betas da regressão (Coeficientes)
  - Erro padrão dos Betas
  - Valor de z = Estatística para determinar a contribuição de uma dada variável na variável resposta (Estatística de Wald)
  - Valor de p = Determina o nível de significância da estatística t
- Erro padrão dos resíduos
- Desviância:
  - Nula = Desviância do modelo com apenas a constante
  - Residual = Estatística teste para qualidade de ajuste do modelo
- AIC = Medida de qualidade do modelo
- Escore de iterações de Fisher = Número de iterações necessárias para o ajuste

- Exemplo
  - Os coeficientes contribuem para a variável resposta
  - Para cada aumento de unidade na média escolar, a razão de chance do aluno ser aprovado num vestibular é de 2,23
  - Escolas com ranking inferiores aumentam a chance de não aprovação em vestibulares

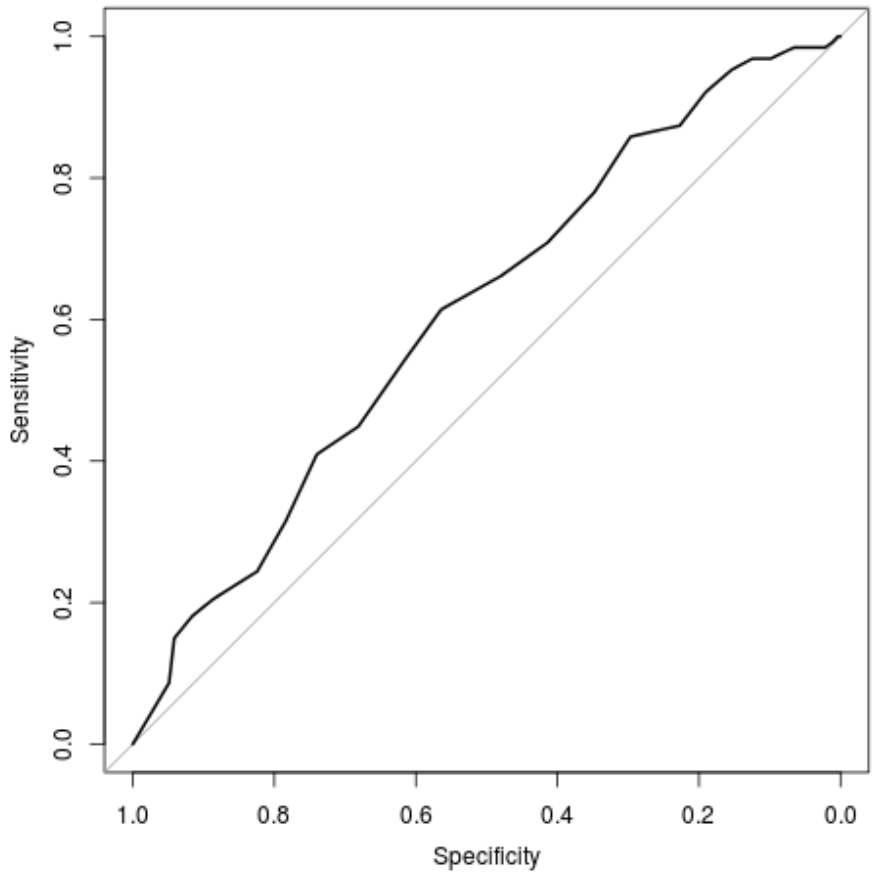
- Exercício

- Utilizado para a regressão logística com resposta binária
- Determina o ponto de corte para definir quando cada valor será assumido como resposta
- Através da determinação ótima da maior especificidade e sensibilidade


Sensibilidade
1,0
0,8
0,6
0,4
0,2
0,0
0,0
0,2
0,4
0,6
0,8
1,0
1 - Especificidade

- Sensibilidade = $\#VP/(\#VP + \#FN)$
- Especificidade = $\#VN/(\#VN + \#FP)$
- Acurácia = $(\#VP + \#VN)/n$
- Área sub a Curva:
  - 1 - 0,90 –> Excelente
  - 0,90 - 0,80 –> Bom
  - 0,80 - 0,70 –> Razoável
  - 0,70 - 0,60 –> Ruim
  - 0,60 - 0,50 –> Fraca

- Exemplo



- Exemplo

```
Call:
roc.formula(formula = admit ~ prob, data = mydata)

Data: prob in 273 controls (admit 0) < 127 cases (admit 1).
Area under the curve: 0.6094
```

- Exercício