FIAP

MBA EM DATA SCIENCE & AI

APPLIED STATISTICS

AULA 6

Teste de Diagnósticos

Curva RoC

Regressão Logística

TESTE DE DIAGNÓSTICOS

Teste de Diagnósticos

Digamos que exista um teste para detectar uma certa doença

- Padrão-Ouro indica quem são os verdadeiros doentes.
- Teste indica qual foi o diagnóstico pelo teste.

	Padrão - Ouro	
Teste	D -	D +
T -		
T+		

Teste de Diagnósticos

Verdadeiro Positivo (VP): T+ em D+, ou seja, quantos foram diagnosticados corretamente com a Doença.

	Padrão - Ouro	
Teste	D -	D +
T -		
T+		

Teste de Diagnósticos

Verdadeiro Positivo (VP): T+ em D+, ou seja, quantos foram diagnosticados corretamente com a Doença.

	Padrão - Ouro	
Teste	D -	D +
T -		
T+		VP

Teste de Diagnósticos

Verdadeiro Negativo (VN): T- em D-, ou seja, quantos foram diagnosticados corretamente sem a Doença.

	Padrão - Ouro	
Teste	D -	D +
T -	VN	
T+		VP

Teste de Diagnósticos

Falso Positivo (FP): T+ em D-, ou seja, quantos foram diagnosticados incorretamente com a Doença.

	Padrão - Ouro	
Teste	D -	D +
T -	VN	
T+	FP	VP

Teste de Diagnósticos

Falso Negativo (FN): T- em D+, ou seja, quantos foram diagnosticados incorretamente sem a Doença.

	Padrão - Ouro	
Teste	D -	D +
T -	VN	FN
T+	FP	VP

Sensibilidade (Recall)

Dos pacientes doentes, quantos foram corretamente identificados?

S = VP / (VP + FN)

	Padrão - Ouro	
Teste	D -	D +
T -	VN	FN
T+	FP	VP

Valor Preditivo Positivo (Precision)

Dos pacientes classificados como doentes, quantos foram corretamente identificados?

$$VPP = VP / (VP + FP)$$

	Padrão - Ouro	
Teste	D -	D +
T -	VN	FN
T+	FP	VP

Especificidade

Dos pacientes não doentes, quantos foram corretamente identificados?

$$E = VN / (VN + FP)$$

	Padrão - Ouro	
Teste	D -	D +
T -	VN	FN
T+	FP	VP

Valor Preditivo Negativo

Dos pacientes classificados como não doentes, quantos foram corretamente identificados?

VPN = VN / (VN + FN)

	Padrão - Ouro	
Teste	D -	D +
T -	VN	FN
T+	FP	VP

Exemplo

Um teste ergométrico foi realizado com o objetivo de detectar doença coronariana (Wiener, 1979 em Soares e Siqueira). O diagnóstico preciso de doença coronariana foi determinado por angioplastia (padrão ouro).

Exemplo

Um teste ergométrico foi realizado com o objetivo de detectar doença coronariana (Wiener, 1979 em Soares e Siqueira). O diagnóstico preciso de doença coronariana foi determinado por angioplastia (padrão ouro).

Exemplo

Um teste ergométrico foi realizado com o objetivo de detectar doença coronariana (Wiener, 1979 em Soares e Siqueira). O diagnóstico preciso de doença coronariana foi determinado por angioplastia (padrão ouro).

	Padrão-Ouro	
Teste	D -	D +
T -	327	208
T+	115	815

Exemplo

Um teste ergométrico foi realizado com o objetivo de detectar doença coronariana (Wiener, 1979 em Soares e Siqueira). O diagnóstico preciso de doença coronariana foi determinado por angioplastia (padrão ouro).

	Padrão-Ouro	
Teste	D -	D +
T -	327	208
T+	115	815

S = 815 / (815 + 208) = 0,797

Exemplo

Um teste ergométrico foi realizado com o objetivo de detectar doença coronariana (Wiener, 1979 em Soares e Siqueira). O diagnóstico preciso de doença coronariana foi determinado por angioplastia (padrão ouro).

	Padrão-Ouro	
Teste	D -	D +
T -	327	208
T+	115	815

E = 327 / (327 + 115)

E = 0,740

Exemplo

Um teste ergométrico foi realizado com o objetivo de detectar doença coronariana (Wiener, 1979 em Soares e Siqueira). O diagnóstico preciso de doença coronariana foi determinado por angioplastia (padrão ouro).

	Padrão-Ouro	
Teste	D -	D +
T -	327	208
T+	115	815

VPP = 815 / (815 + 115)

VPP = 0,876

Exemplo

Um teste ergométrico foi realizado com o objetivo de detectar doença coronariana (Wiener, 1979 em Soares e Siqueira). O diagnóstico preciso de doença coronariana foi determinado por angioplastia (padrão ouro).

	Padrão-Ouro	
Teste	D -	D +
T -	327	208
T+	115	815

VPN = 327 / (327 + 208)

VPN = 0,611

Curva RoC

- A Curva ROC (**Receiver Operating Characteristic Curve)** surgiu durante a segunda Guerra Mundial para distinguir:

  - Sinal de avião inimigo;
  - Ruído irrelevante (pássaros, etc)

É utilizada para:

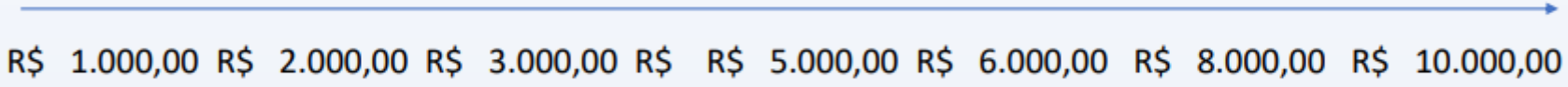
  - Achar um ponto de corte de uma variável continua
  - Avaliar a capacidade dessa variável de classificar uma variável dicotômica

- Usando **sensibilidade** e **especificidade**.

Curva RoC

Como funciona?

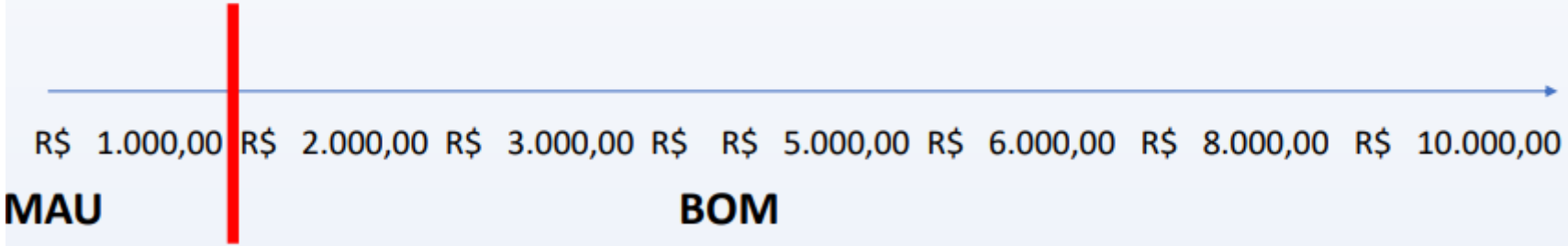
Considere a variável **Salário** abaixo:



Curva RoC

Como funciona?

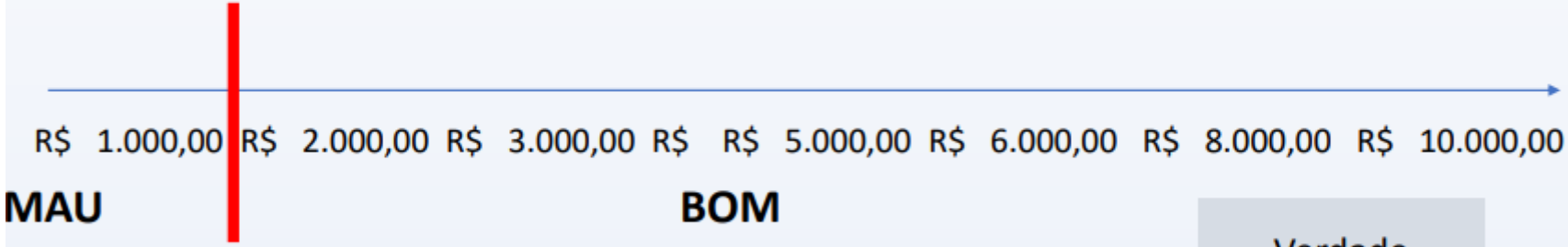
Considere a variável **Salário** abaixo:



Curva RoC

Como funciona?

Considere a variável **Salário** abaixo:



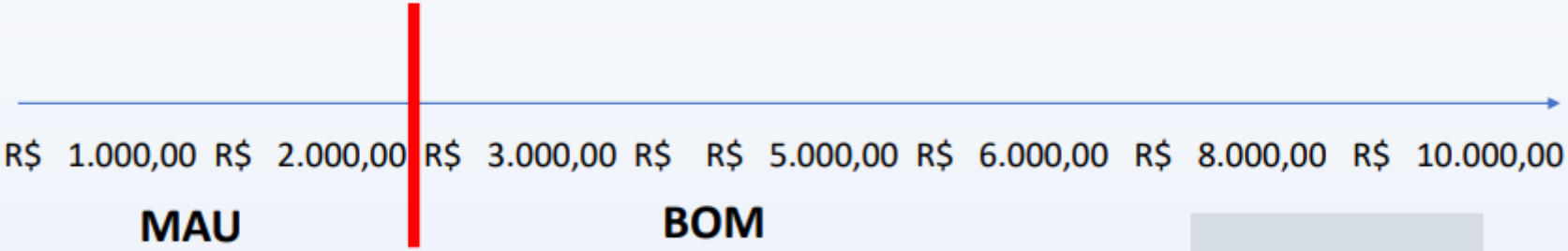
Calculamos sensibilidade e especificidade

	Verdade	
Classif.	MAU	BOM
MAU	VN	FN
BOM	FP	VP

Curva RoC

Como funciona?

Considere a variável **Salário** abaixo:



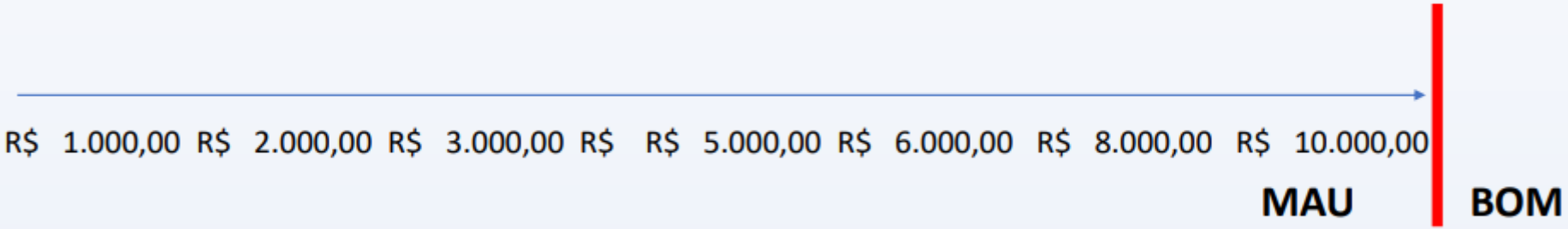
Calculamos sensibilidade e especificidade

	Verdade	
Classif.	MAU	BOM
MAU	VN	FN
BOM	FP	VP

Curva RoC

Como funciona?

Considere a variável **Salário** abaixo:



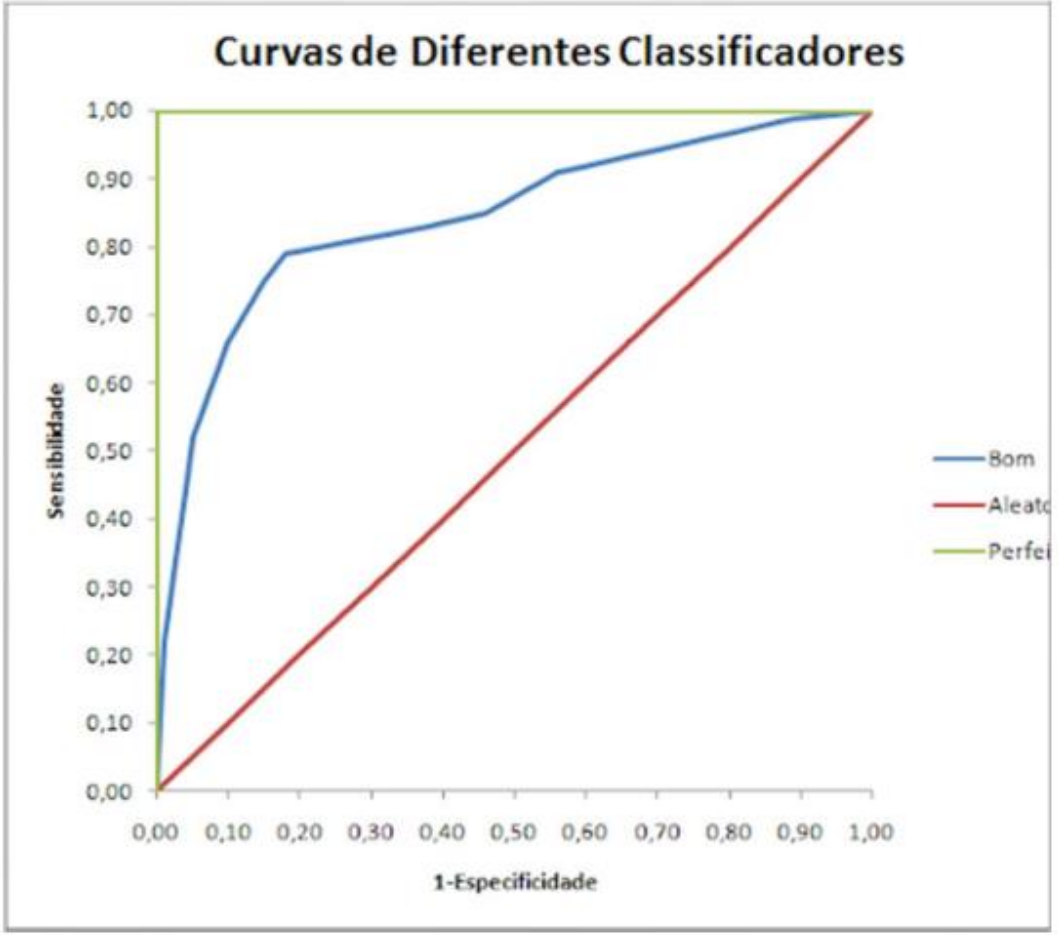
O processo termina quando varremos todas as possibilidades

Curva ROC

- Plotando Sensibilidade x

(1 – Especificidade) obtemos a curva.

- A Área debaixo da curva nos dá uma medida de qualidade do teste ou classificador.



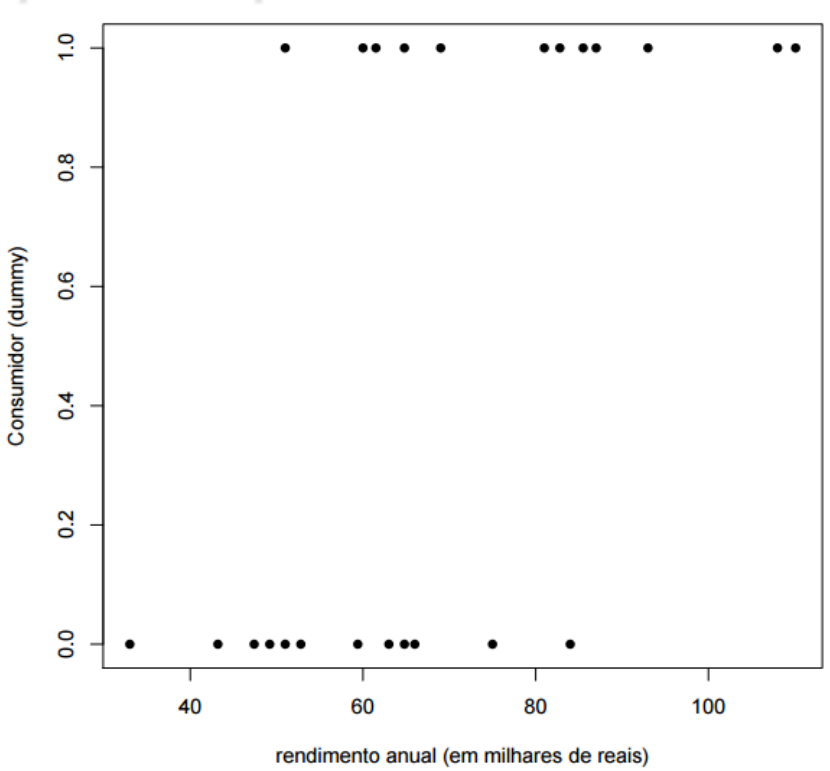
Um modelo simples e poderoso para predizer a probabilidade de ocorrência de um evento dicotômico.

REGRESSÃO LOGÍSTICA

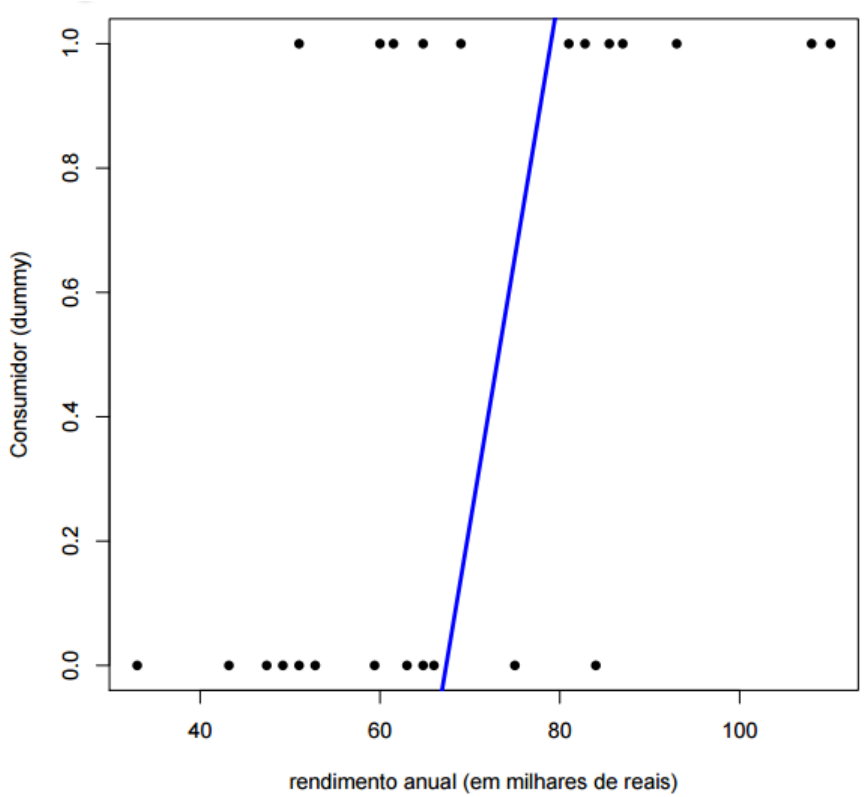
Consumidor de trator

- Por meio de um modelo gostaríamos de poder classificar se uma pessoa é ou não um provável consumidor de trator.
- Qual seria um possível modelo para esse problema?

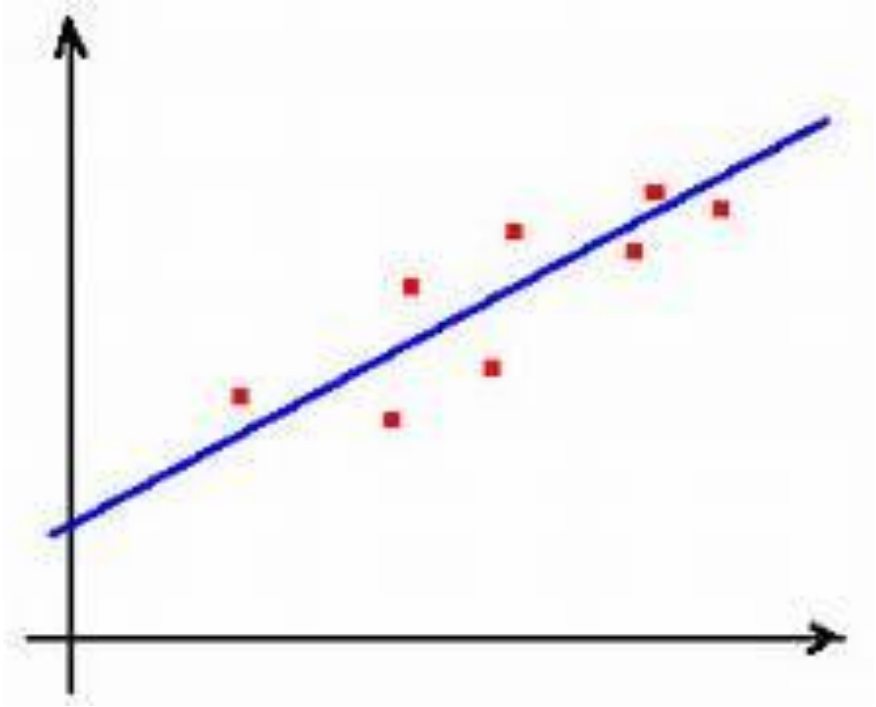
Consumidor de trator



Ajuste por uma reta??



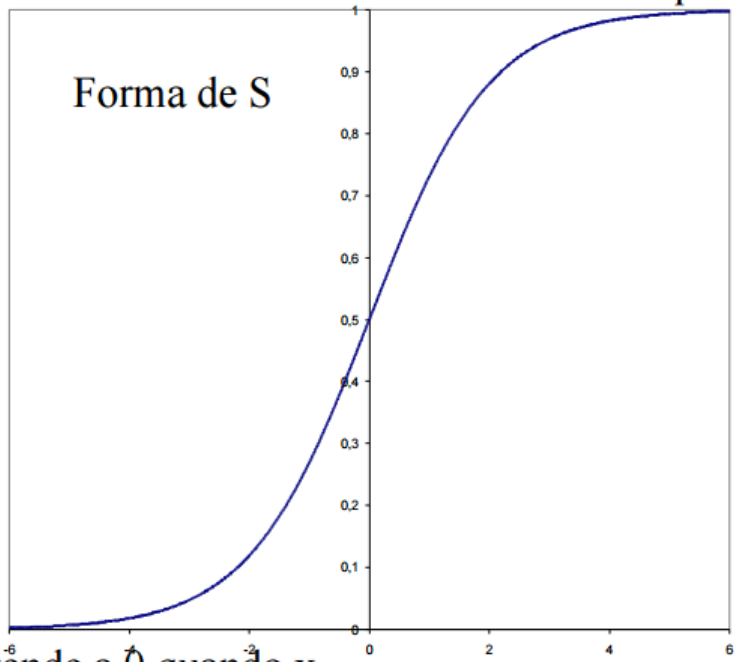
Não parece razoável!!!



Modelo Logístico

tende a 1 quando x aumenta

Forma de S



tende a 0 quando x diminui

Caso crescente

$\beta_1 > 0$

Modelo logístico

Y = variável dependente dicotômica

$X_1, X_2, \ldots X_p$ = variáveis independentes

Objetivo: encontrar uma relação funcional entre $P(Y = 1)$ e $X_1, X_2, \ldots X_p$ (regressão pela média).

Chance do evento de interesse

Modelar o logaritmo neperiano (ln) da chance de ocorrência do evento de interesse:

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)=\beta_0+\beta_1 x$$

$$P(Y=1)=\frac{e^{\beta_0+\beta_1 x}}{1+e^{\beta_0+\beta_1 x}}=\frac{1}{1+e^{-(\beta_0+\beta_1 x)}}$$

Exemplo: Consumidor de trator

FIAP MBA+

Y: 1=consumidor, 0=não consumidor

x_1: rendimento anual (em milhares de reais)

x_2: tamanho do lote (em hectares)

x_3: 1=se há criação de gado, 0=caso contrário

MODELO

$$P(\text{consumidor}) = \frac{e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}{1+e^{\beta_0+\beta_1 x_1+\beta_2 x_2+\beta_3 x_3}}$$

No Python

```
modelo = sm.GLM(y, X_sm, family =
sm.families.Binomial()).fit()
```

Indica que o modelo é logístico pertence a uma classe de modelos mais genéricos (glm ou mlg em português.)

Indica que o modelo é logístico

Saída do modelo

	coef	std err	z	P>|z|	[0.025	0.975]
const	-30.0278	13.463	-2.230	0.026	-56.415	-3.641
rendimento	0.1299	0.059	2.188	0.029	0.014	0.246
tamanho.lote	1.0759	0.535	2.012	0.044	0.028	2.124
criacao.gado	1.9432	1.645	1.181	0.238	-1.282	5.168

Coeficientes do Modelo

Saída do modelo

	coef	std err	z	P>|z|	[0.025	0.975]
const	-30.0278	13.463	-2.230	0.026	-56.415	-3.641
rendimento	0.1299	0.059	2.188	0.029	0.014	0.246
tamanho.lote	1.0759	0.535	2.012	0.044	0.028	2.124
criacao.gado	1.9432	1.645	1.181	0.238	-1.282	5.168

Erro padrão

Saída do modelo

	coef	std err	z	P>|z|	[0.025	0.975]
const	-30.0278	13.463	-2.230	0.026	-56.415	-3.641
rendimento	0.1299	0.059	2.188	0.029	0.014	0.246
tamanho.lote	1.0759	0.535	2.012	0.044	0.028	2.124
criacao.gado	1.9432	1.645	1.181	0.238	-1.282	5.168

P-valor

Saída do modelo

	coef	std err	z	P>|z|	[0.025	0.975]
const	-30.0278	13.463	-2.230	0.026	-56.415	-3.641
rendimento	0.1299	0.059	2.188	0.029	0.014	0.246
tamanho.lote	1.0759	0.535	2.012	0.044	0.028	2.124
criacao.gado	1.9432	1.645	1.181	0.238	-1.282	5.168

IC95%

Como interpretar os Coeficientes?

	coef	std err	z	P>|z|	[0.025	0.975]
const	-30.0278	13.463	-2.230	0.026	-56.415	-3.641
rendimento	0.1299	0.059	2.188	0.029	0.014	0.246
tamanho.lote	1.0759	0.535	2.012	0.044	0.028	2.124
criacao.gado	1.9432	1.645	1.181	0.238	-1.282	5.168

Como interpretar os Coeficientes?

	coef	std err	z	P>|z|	[0.025	0.975]
const	-30.0278	13.463	-2.230	0.026	-56.415	-3.641
rendimento	0.1299	0.059	2.188	0.029	0.014	0.246
tamanho.lote	1.0759	0.535	2.012	0.044	0.028	2.124
criacao.gado	1.9432	1.645	1.181	0.238	-1.282	5.168

Equação usa o logaritmo, isso não dá uma interpretação direta

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)=\beta_0+\beta_1 x$$

Como interpretar os Coeficientes?

	coef	std err	z	P>|z|	[0.025	0.975]
const	-30.0278	13.463	-2.230	0.026	-56.415	-3.641
rendimento	0.1299	0.059	2.188	0.029	0.014	0.246
tamanho.lote	1.0759	0.535	2.012	0.044	0.028	2.124
criacao.gado	1.9432	1.645	1.181	0.238	-1.282	5.168

Equação usa o logaritmo, isso não dá uma interpretação direta

$$\ln\left(\frac{P(Y=1)}{P(Y=0)}\right)=\beta_0+\beta_1 x$$

Vamos precisar do conceito de ODDs Ratio (OR)

Chance (Odds) x Probabilidade

Probabilidade

- **Ex**: Se tenho uma moeda na qual a cada 10 jogadas, obtenho cara 8 vezes e coroa 2 vezes, qual a **probabilidade** de dar cara?

$$\frac{8}{10} = 80\%$$

A cada 10 jogadas, tenho 8 caras

Chance (Odds)

- **Ex**: Se tenho uma moeda na qual a cada 10 jogadas, obtenho cara 8 vezes e coroa 2 vezes, qual a **chance** de dar cara?

$$\frac{8}{(10-8)} = 4$$

A cada Coroa, tenho 4 Caras

- Relação entre Chance e Probabilidade:

$$\text{Chance} = \frac{\text{Prob}}{(1-\text{Prob})}$$

Quanto maior a probabilidade, maior a chance. Ou vice versa.

Em português, costumamos mencionar “probabilidade” e “chance” como sinônimos, mas são numericamente distintos.

Odds Ratio (OR) – Razão de Chances

Agora um exemplo de **Odds Ratio (OR) .** Considere os dados abaixo:

	Doença X	Sem Doença X	Total
Consome Fritura	400	100	500
Não consome fritura	200	300	500

OR

$$\frac{\text{**Chance** de ter a doença consumindo fritura}}{\text{**Chance** de ter a doença Não consumindo fritura}} = \frac{400 / 100}{200/300} = \frac{4}{0{,}66} = 6$$

A **Chance** da pessoa que consome fritura ter a doença é **6x** maior do que o da pessoa que não consome.

Odds Ratio (OR) – Modelo logístico

Para obter o **Odds Ratio (OR) ,**
basta utilizar a exponencial do expoentes betas.

Interpretando o modelo

	coef					
const	-30.0278					
rendimento	0.1299					
tamanho.lote	1.0759	0.555	2.912	0.044	0.028	2.124
criacao.gado	1.9432	1.645	1.181	0.238	-1.282	5.168

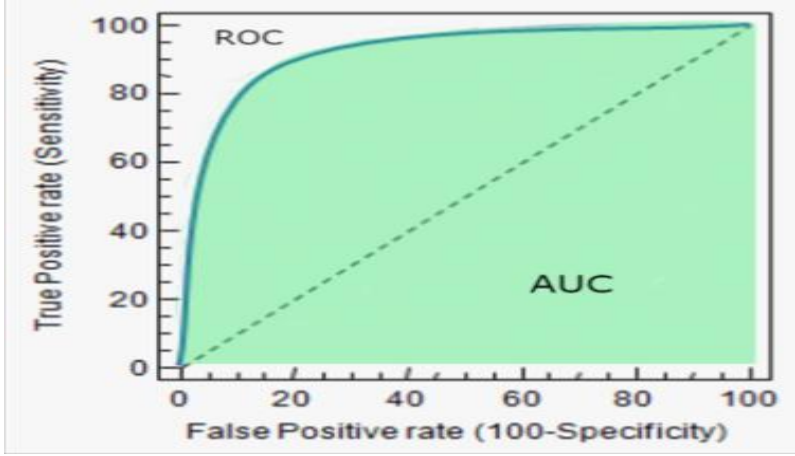
```
rendimento    1.138
tamanho.lote 2.932
criacao.gado 6.981
```

- A chance de uma pessoa tornar-se consumidor de trator aumenta em 13,8% (1,138 – 1) a cada aumento de unidade de **rendimento** (ou seja, a cada aumento de 1 mil).

- A chance de uma pessoa tornar-se consumidor de trator aumenta em 193,2% (2,932 – 1) a cada aumento de unidade **de tamanho do lote**.

- A chance de uma pessoa tornar-se consumidor é de 6,91 vezes a mais se ela for **criador de gado.**

Performance do modelo



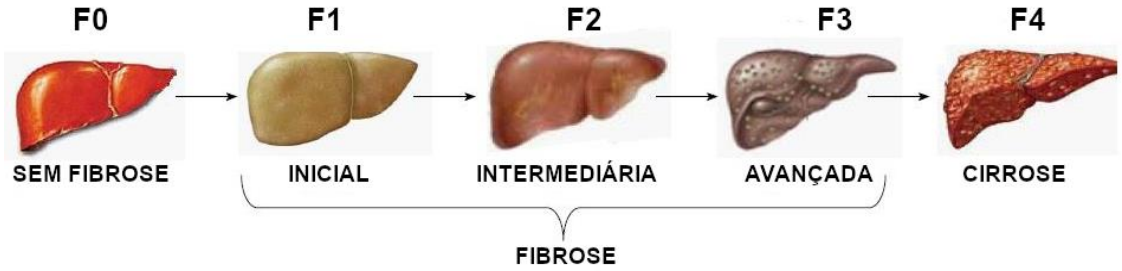
- Medida AUC (Area Under Curve) é a área formada pela curva de sensibilidade x 1 - especificidade)

Método StepWise

- Incremento de variáveis a cada rodada.
- A cada rodada é medido a acurácia do modelo e é verificado o se a o modelo piorou ou melhorou.
  - Se o modelo piorar, a variável é retirada
  - Se o modelo melhorar, a variável é inclusa

Exercício

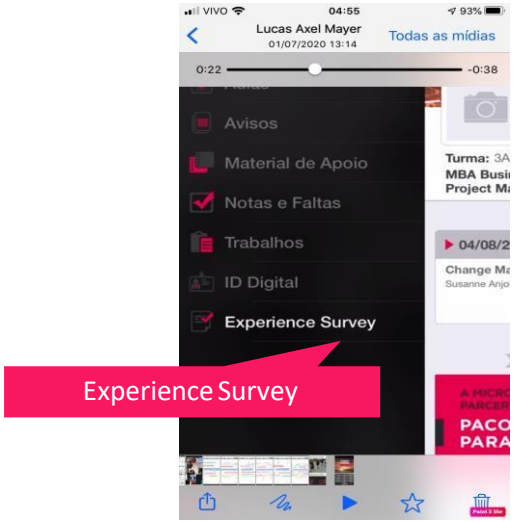
- Utilize a data set “base fibrose” para modelar o conjunto “F0F1” x “F2F3F4”, onde Fi é o grau da doença Hepática Fibrose.

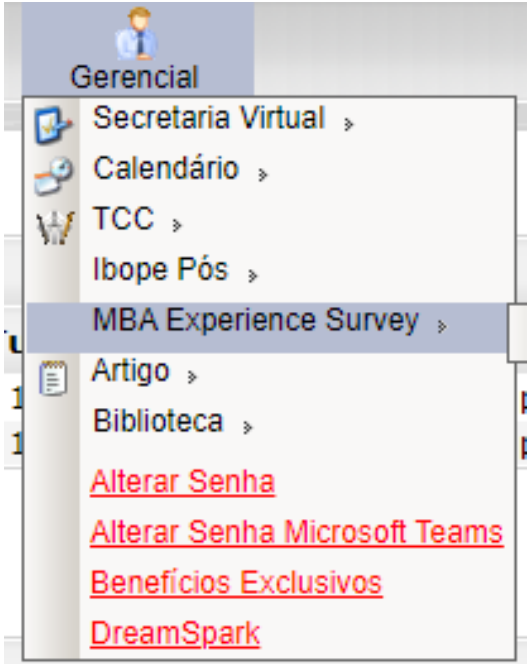


O que você achou da aula de hoje?

Pelo aplicativo da FIAP

(Entrar no FIAPP, e no menu clicar em Experience Survey)





OBRIGADO

in/lafphd

FIAP MBA+


FIAP