FIAP

MBA

MBA EM DATA SCIENCE & AI

STATISTICS WITH R

AULA 6

Regressão Logística

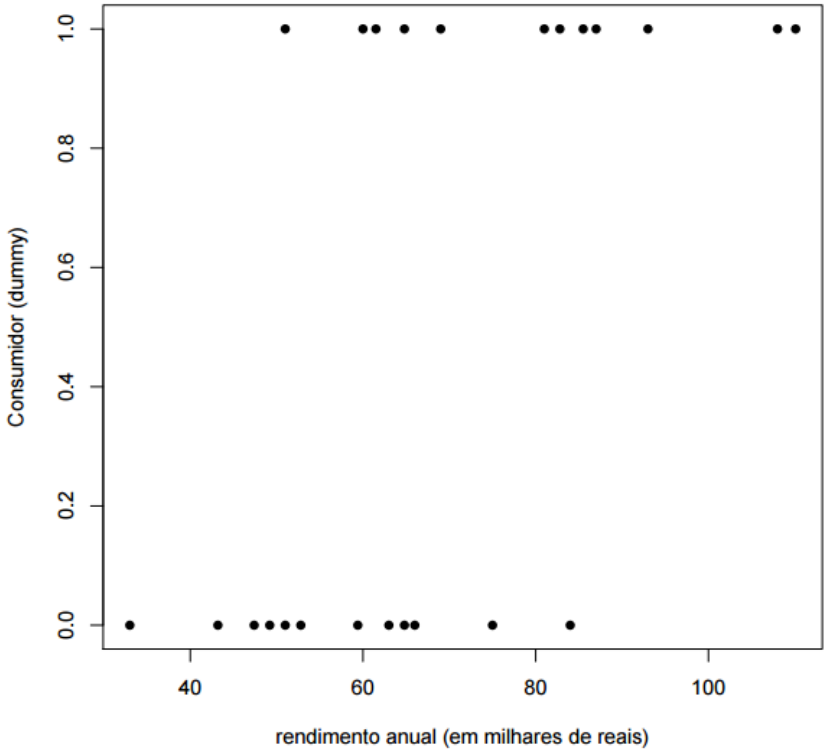
Um modelo simples e poderoso para predizer a probabilidade de ocorrência de um evento dicotômico.

REGRESSÃO LOGÍSTICA

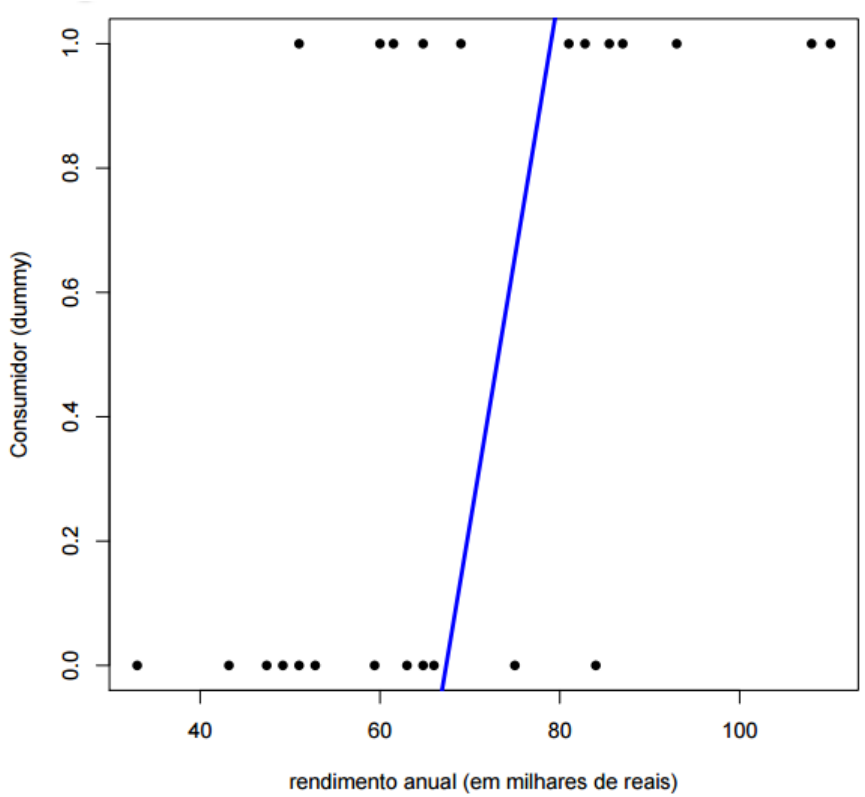
Consumidor de trator

- 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?

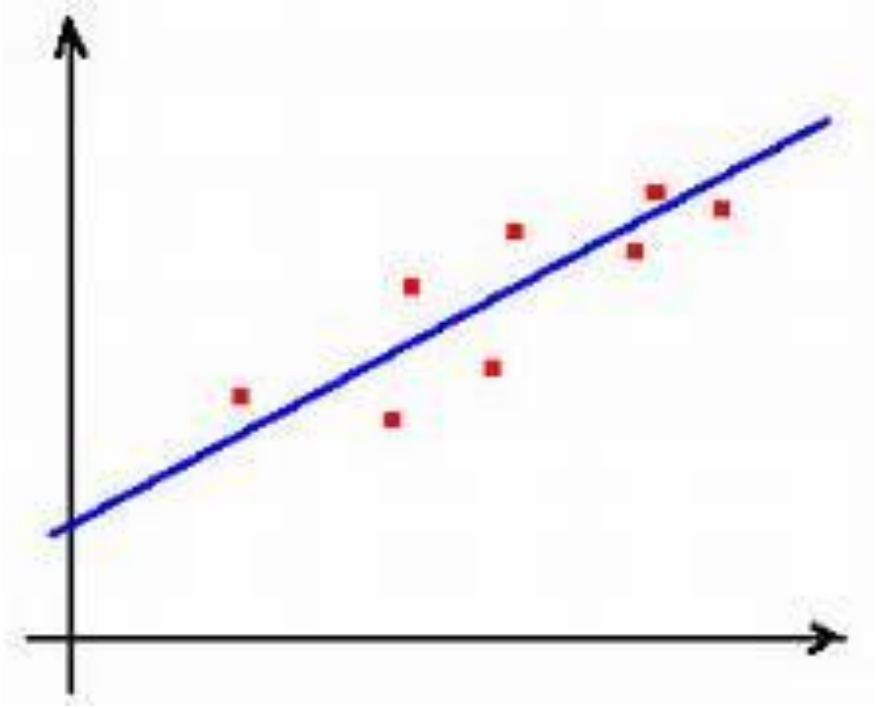
Consumidor de trator



Ajuste por uma reta??

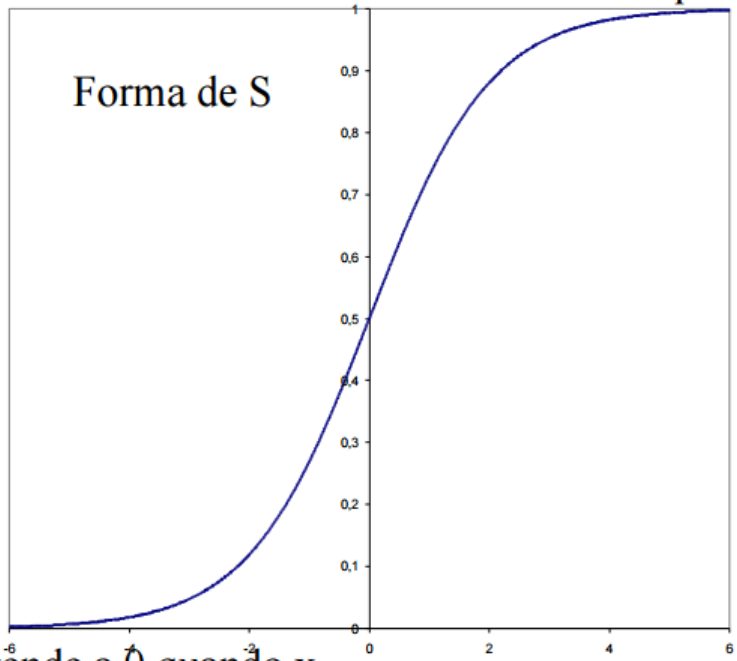


Não parece razoável!!!



Modelo Logístico

tende a 1 quando x aumenta



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 Software R

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

```
modelo <- glm(consumidor~rendimento + tamanho.lote + criacao.gado, data = dados, family = 'binomial')
```

Indica que o modelo é logístico

Akaike Information Criterion (AIC)

- O critério de informação de Akaike (AIC) é um método matemático para avaliar o quão bem um modelo se ajusta aos dados a partir dos quais foi gerado. Em estatística, o AIC é usado para comparar diferentes modelos possíveis e determinar qual é o melhor para os dados.

$$AIC = 2K - 2\ln(L)$$

K: o número de variáveis
L: A função de verossimilhança do modelo

Para o modelo logístico a função L é uma função descrita da seguinte maneira.

$$L(\beta_0, \beta_1) = p_i^{y_i}(1-p_i)^{1-y_i}$$

Método StepWise

- Incremento de variáveis a cada rodada.
- A cada rodada é medido o AIC e 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

AIC no R

step(modelo)

MASS::stepAIC(modelo)

```
direction = c("both", "backward", "forward")
```

Variance Inflation Factor (VIF)

- A multicolinearidade na análise de regressão ocorre quando duas ou mais variáveis preditoras são altamente correlacionadas entre si, de modo que não fornecem informações únicas ou independentes no modelo de regressão.

Variance Inflation Factor (VIF)

- Um valor de 1 indica que não há correlação entre uma determinada variável preditora e quaisquer outras variáveis preditoras no modelo.
- Um valor entre 1 e 5 indica correlação moderada entre uma determinada variável preditora e outras variáveis preditoras no modelo, mas isso geralmente não é grave o suficiente para exigir atenção.
- Um valor maior que 5 indica correlação potencialmente grave entre uma determinada variável preditora e outras variáveis preditoras no modelo. Nesse caso, as estimativas de coeficiente e os valores-p na saída da regressão provavelmente não são confiáveis.

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.

Voltando no Exemplo

```
Call:
glm(formula = consumidor ~ rendimento + tamanho.lote + criacao.gado,
    family = "binomial", data = dados)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.53313  -0.25098  -0.00622   0.42648   2.24423

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  -30.02783   13.46137  -2.231   0.0257 *
rendimento     0.12994    0.05938   2.188   0.0286 *
tamanho.lote   1.07593    0.53481   2.012   0.0442 *
criacao.gado   1.94320    1.64534   1.181   0.2376
---
Signif. codes:
0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Essa variável não foi significativa de acordo com o nível de significância de 5%

Podemos utilizar o método Stepwise

```
Call:
glm(formula = consumidor ~ rendimento + tamanho.lote, family = "binomial",
    data = dados)

Deviance Residuals:
     Min        1Q    Median        3Q       Max
-1.74044  -0.29685   0.00439   0.44750   1.86821

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)  -25.9382    11.4871  -2.258   0.0239 *
rendimento     0.1109     0.0543   2.042   0.0412 *
tamanho.lote   0.9638     0.4628   2.083   0.0373 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 33.271  on 23  degrees of freedom
Residual deviance: 15.323  on 21  degrees of freedom
AIC: 21.323
```

Todas são significativas a 5%

Interpretando o modelo

	beta <dbl>	OR <dbl>	IC2.5 <dbl>	IC97.5 <dbl>	p.value <dbl>
(Intercept)	-25.938	0.000	0.000	0.033	0.024
rendimento	0.111	1.117	1.004	1.243	0.041
tamanho.lote	0.964	2.622	1.058	6.494	0.037

- A chance de uma pessoa tornar-se consumidor de trator aumenta em 11,7% (1,117 – 1) a cada aumento de unidade de **rendimento** (ou seja, a cada aumento de 1 mil).

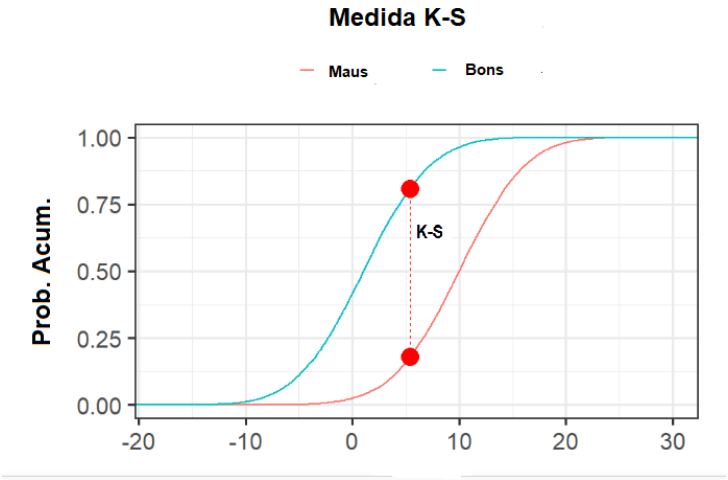
- A chance de uma pessoa tornar-se consumidor de trator aumenta em 162,2% (2,622 – 1) a cada aumento de unidade **de tamanho do lote**.

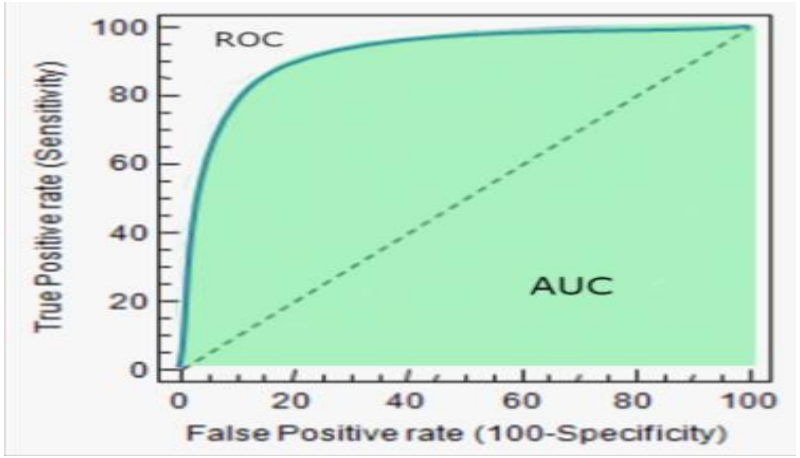
Verificando o VIF do modelo final

VAR <chr>	VIF <dbl>
rendimento	1.524088
tamanho.lote	1.524088

Todas são menores que 5, portando podem ficar no modelo final

Performance do modelo





- Medida K-S é a maior distância entre a curva de prob. Acumulada dos bons e maus.

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

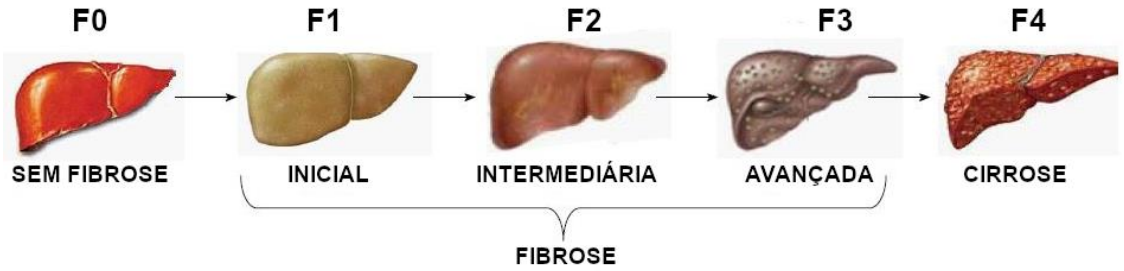
Performance do modelo

```
score = predict(modelo, dados, type = "response")

pred <- ROCR::prediction(score, dados[[names(modelo$model)[[1]]]])
perf <- ROCR::performance(pred, "tpr", "fpr")

ks <-  max(attr(perf, 'y.values')[[1]] - attr(perf, 'x.values')[[1]])
auc <- ROCR::performance(pred, measure = "auc")
auc <- auc@y.values[[1]]
```

KS <dbl>	AUC <dbl>
0.75	0.92

Exercício

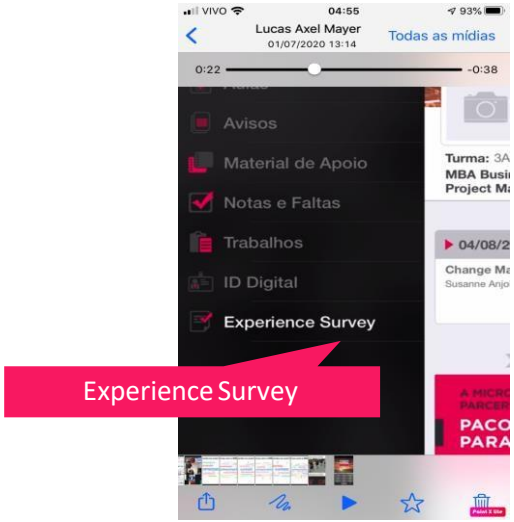
- Utilize a data set “base fibrose” para modelar o conjunto “F0F1” x “F2F3F4”, onde Fi é o grau da doença Hepática Fibrose.

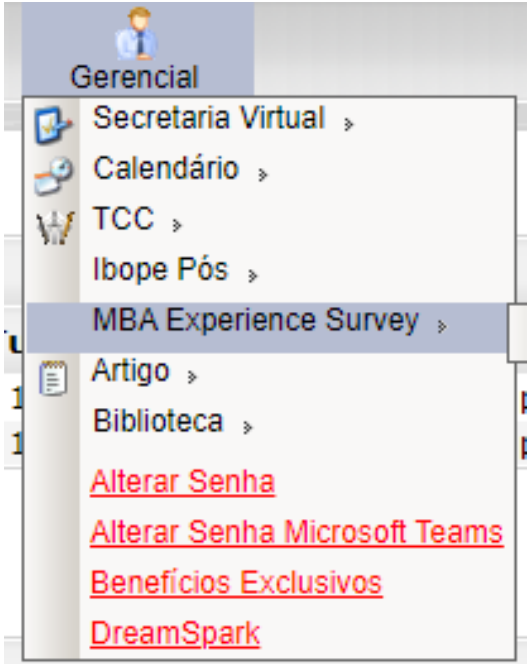


O que você achou da aula de hoje?

Pelo aplicativo da FIAP

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





OBRIGADO

in/lafphd

FIAP MBA+


FIAP