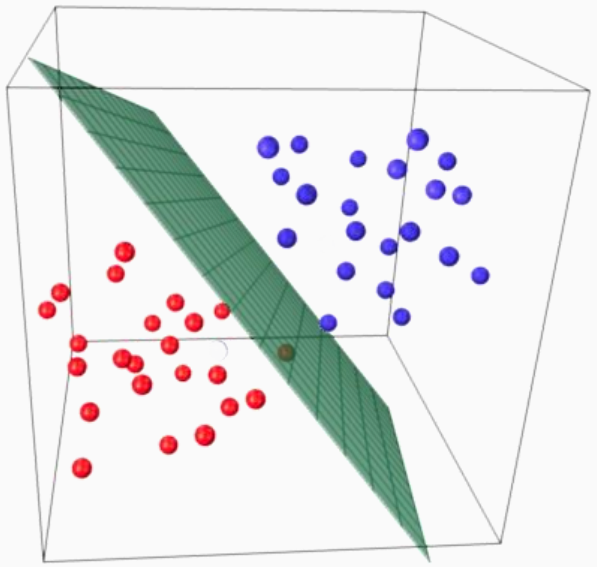
Modelo de Regressão Logística

Alexandre Cristovão Maiorano

alexandre.maiorano@usp.br

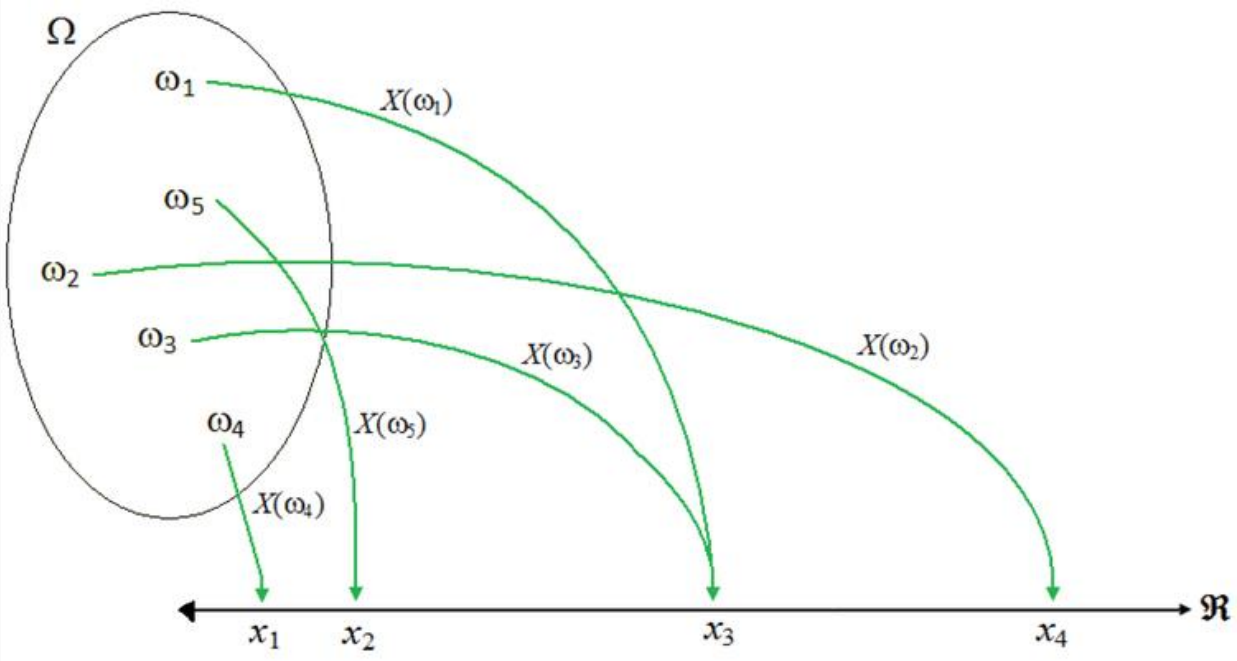
ICMC/USP



Variável aleatória

- Variável aleatória X é uma função que leva do espaço real para o espaço amostral

$$X : \Omega \rightarrow \Re$$



Variáveis quantitativas

- Contínua: assume qualquer valor numérico em um determinado intervalo ou coleção de intervalos
  - Altura
  - Peso
  - Duração de uma ligação telefônica
  - Tempo até chegar um novo cliente

- Discreta: assume valores que podem ser contados
  - Número de filhos
  - Número de gols
  - Número de cartões de crédito
  - Número de carros que passam pelo pedágio

Variáveis qualitativas

- Nominais: não existe ordenação entre as categorias
  - Sexo
  - Cor do olho
  - Doente/não doente
  - Raça/cor

- Ordinais: existe ordenação entre as categorias
  - Nível de escolaridade
  - Estágio da doença
  - Escala de dor
  - Nível de satisfação

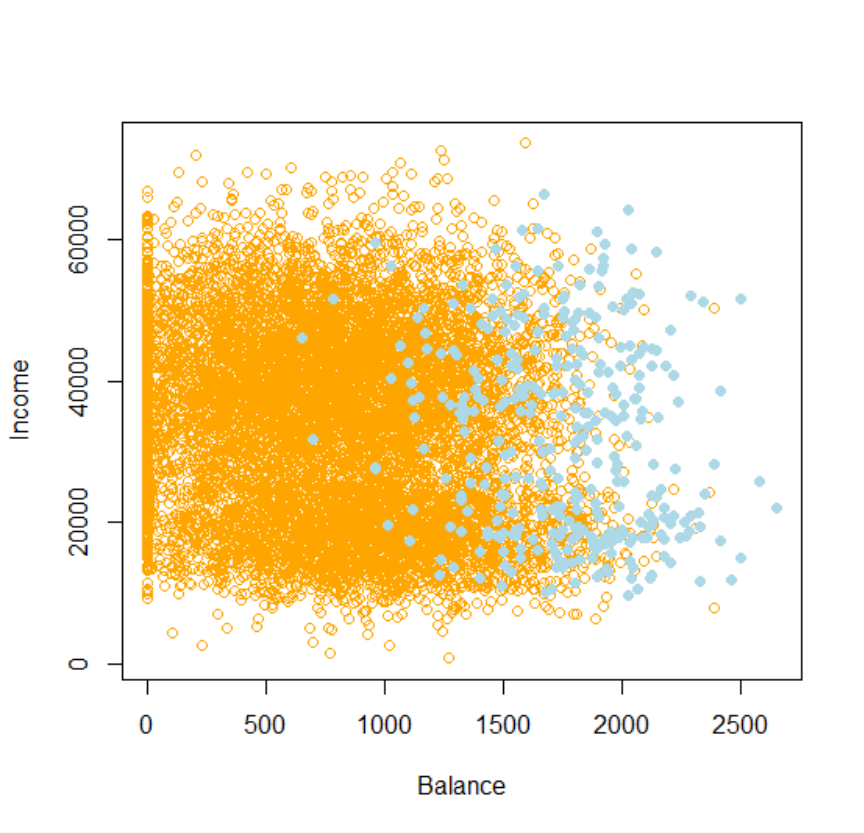
Classificação

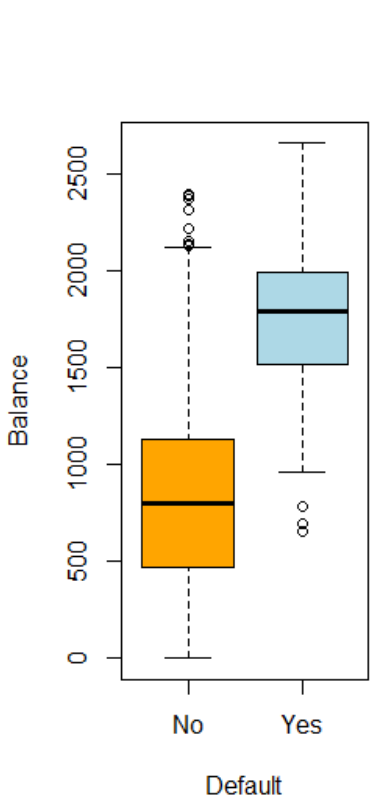
- Variáveis qualitativas podem assumir valores em um conjunto $\mathcal{C}$, tal como
  - Sexo $\in$ {masculino, feminino}
  - Cor dos olhos $\in$ {preto, castanho, azul, verde}
  - Raça/cor $\in$ {branco, pardo, preto, indígena, amarelo}
  - Doença $\in$ {sim, não}

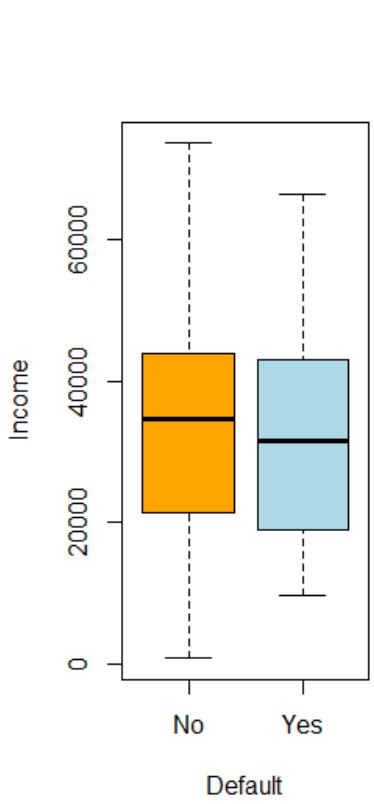
- A partir de um vetor X e uma resposta qualitativa Y que assume valores no conjunto $\mathcal{C}$, queremos construir uma função $C(X)$ que leva em conta valores de X e prevê o valor de Y, isto é, $C(X) \in \mathcal{C}$

- Geralmente estamos interessados em estimar a probabilidade de X pertencer a cada categoria de $\mathcal{C}$

Classificação







Classificação

- Suponha que queremos criar uma regra de classificação para a variável Default (mau pagador), que codificamos como

$$Y = \begin{cases} 0, & \text{se não} \\ 1, & \text{se sim} \end{cases}$$

- Esse é um caso de variável qualitativa binária que podemos utilizar o modelo de regressão logística para estimar

$$E(Y|X = x) = P(Y = 1|X = x)$$

Modelo de regressão logística

- Consideramos $p(X) = P(Y = 1|X = x)$ e a variável Balance (saldo devedor) para prever a Default (mau pagador). O modelo logístico é dado pela expressão

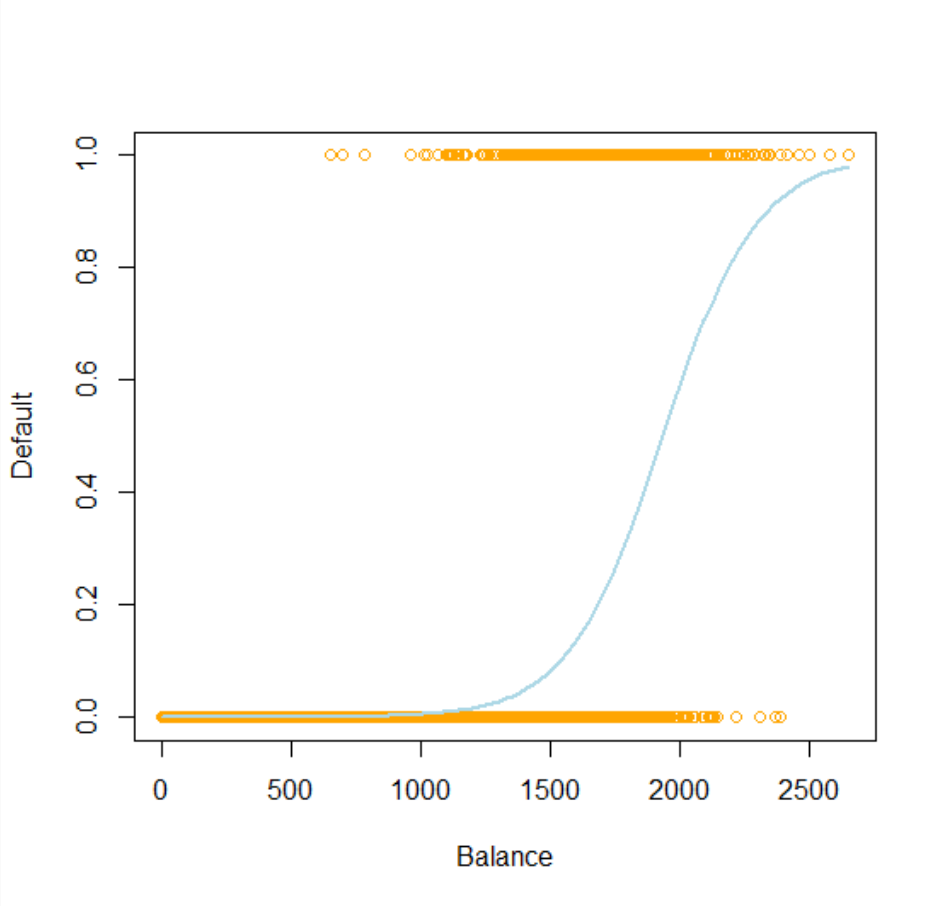
$$p(X) = \frac{e^{\beta_0 + \beta_1 X}}{1 + e^{\beta_0 + \beta_1 X}}$$

- Nesse caso, não importa os valores de β_0, β_1 ou X, o valor de $p(X)$ sempre estará entre 0 e 1.
- Rearranjando os termos, podemos obter

$$\log\left(\frac{p(X)}{1 - p(X)}\right) = \beta_0 + \beta_1 X$$

- Essa transformação é chamada de *log odds* ou logito de $p(X)$

Modelo de regressão logística



	Estimativa	Erro Padrão	Estatística Teste	P-Valor
Intercepto	-10.6513	0.3612	-29.4900	<0.0001
Balance	0.0055	0.0002	24.9500	<0.0001

Predição

- Qual a probabilidade de uma pessoa com saldo devedor de \$1500.00 não pagar o banco $(Y = 1)$?

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0 + \hat{\beta}_1 X}}{1 + e^{\hat{\beta}_0 + \hat{\beta}_1 X}} = \frac{e^{-10.6513 + 0.0055 * 1500}}{1 + e^{-10.6513 + 0.0055 * 1500}} = 8.31\%$$

- E saldo de \$2500.00?

$$\hat{p}(X) = \frac{e^{-10.6513 + 0.0055 \times 2500}}{1 + e^{-10.6513 + 0.0055 \times 2500}} = 95.68\%$$

Razão de chances

- Odds: compara a probabilidade sucesso ($Y = 1$) com a probabilidade de fracasso ($Y = 0$)

$$g(X) = \frac{p(X)}{1 - p(X)} = e^{\beta_0 + \beta_1 X}$$

- Nosso objetivo é determinar a razão entre odds. Fazendo a razão de odds (OR) para dois valores distintos

$$OR = \frac{g(X_{j+1})}{g(X_j)} = \frac{e^{\beta_0 + \beta_1 X_{j+1}}}{e^{\beta_0 + \beta_1 X_j}}$$

Razão de chances

- Temos que

$$\log(OR)$$
$$= \log\left(\frac{g(X_{j+1})}{g(X_j)}\right) = \log\left(g(X_{j+1})\right) - \log\left(g(X_{j+1})\right)$$
$$= \beta_0 + \beta_1 X_{j+1} - \beta_0 - \beta_1 X_j = \beta_1 \left(X_{j+1} - X_j\right)$$

- Tomando $X_{j+1} - X_j = 1$ (uma unidade) teremos

$$OR = e(\beta_1)$$

Razão de chances

- Para uma variável categórica levamos em conta as chamadas variáveis dummy (indicadora)
- Supondo uma variável que indica o sexo de uma pessoa {masculino, feminino}, teremos a seguinte variável

$$X = \begin{cases} 0, & \text{se feminino} \\ 1, & \text{se masculino} \end{cases}$$

- A razão de odds será igual a

$$OR = e\{\beta_1 \ (X_{\text{masc}} - X_{\text{fem}})\} = e\{\beta_1 \ (1 - 0)\} = e(\beta_1 \ )$$

Razão de chances

- Características
  - Se $\beta_1 > 0$ então $OR > 1 \Longrightarrow p(X_{j+1}) > p(X_j)$
  - Se $\beta_1 < 0$ então $OR < 1 \Longrightarrow p(X_{j+1}) < p(X_j)$
- Para $\hat{\beta}_1 = 0.0055$, temos

		I.C.	
	OR	2.5%	97.5%
Balance	1.0055	1.0051	1.006

Modelo de regressão logística

- Com p variáveis, teremos

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 \ + \beta_1 \ X_1 + \cdots + \beta_p \ X_p$$

$$p(X) = \frac{e^{\beta_0 \ + \beta_1 \ X_1 + \cdots + \beta_p \ X_p}}{1 + e^{\beta_0 \ + \beta_1 \ X_1 + \cdots + \beta_p \ X_p}}$$

	Estimativa	OR	Erro Padrão	Estatística Teste	P-Valor
Intercepto	-10.8690	-	0.4923	-22.0800	< 0.0001
Balance	0.0057	1.0057	0.0002	24.7380	< 0.0001
Income	0.0000	1.0000	0.0000	0.3700	0.7115
Student [Yes]	-0.6468	0.5237	0.2363	-2.7380	0.0062

Predição

- Qual a probabilidade de uma pessoa com saldo devedor de $1000.00, salário $5000.00 e estudante não pagar o banco?

$$\hat{p}(X) = \frac{e^{-10.8690+0.0057\times1000+0.000\times5000-0.6468\times1}}{1+e^{-10.8690+0.0057\times1000+0.000\times5000-0.6468\times1}} = 0.03\%$$

- E se o saldo for $2000?

$$\hat{p}(X) = \frac{e^{-10.8690+0.0057\times2000+0.000\times5000-0.6468\times1}}{1+e^{-10.8690+0.0057\times2000+0.000\times5000-0.6468\times1}} = 47.11\%$$

Como avaliar o modelo?

- Matriz de confusão
- Medidas de desempenho
- Curva ROC
- Estatística de Kolmogorov-Smirnov (KS)

Ponto de corte

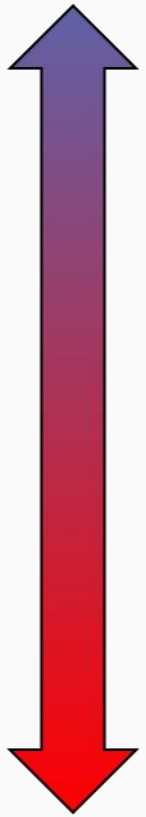
- Um ponto de corte c é um ponto cujo valor é considerado como "divisor" da probabilidade estimada pelo modelo
- Considerando $Y = 1$, temos

$$\hat{p}(X) \geq c \Longrightarrow \hat{Y} = 1$$

$$\hat{p}(X) < c \Longrightarrow \hat{Y} = 0$$

Ponto de corte

ESCORE	Evento	Frequência Total
0 - 100	0.92	91.92
101 - 200	13.72	3.28
201 - 300	19.54	1.74
301 - 400	37.63	0.93
401 - 500	42.65	0.68
501 - 600	58.54	0.41
601 - 700	68.57	0.35
701 - 800	77.14	0.35
801 - 900	90.91	0.22
901 - 1000	83.33	0.12
TOTAL	3.33	100.00



$\longrightarrow \quad c = 0.0312$

Matriz de confusão

		Condição Predita			
		Positivo	Negativo		
Condição Verdadeira	Positivo (T)	Verdadeiro Positivo (TP)	Falso Negativo (FN) (Erro tipo II)	Sensibilidade (SEN)	$= \frac{TP}{TP + FN}$
	Negativo (N)	Falso Positivo (FP) (Erro tipo I)	Verdadeiro Negativo (TN)	Especificidade (SPE)	$= \frac{TN}{FP + TN}$
		Valor Preditivo Positivo (PPV) $= \frac{TP}{TP + FP}$	Valor Preditivo Negativo (NPV) $= \frac{TN}{TN + FN}$	Acurácia (ACC)	$= \frac{TP + TN}{P + N}$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

Matriz de confusão

		Condição Predita	
		Positivo	Negativo
Condição Verdadeira	Positivo (T)	301 (TP)	32 (FN)
	Negativo (N)	1342 (FP)	8325 (TN)

Sensibilidade (SEN) $= \frac{301}{301 + 32} = 90.39$

Especificidade (SPE) $= \frac{8325}{1342 + 8325} = 86.12$

Acurácia (ACC) $= \frac{301 + 8325}{333 + 9667} = 86.26$

Valor Preditivo Positivo (PPV) $= \frac{301}{301 + 1342} = 18.32$

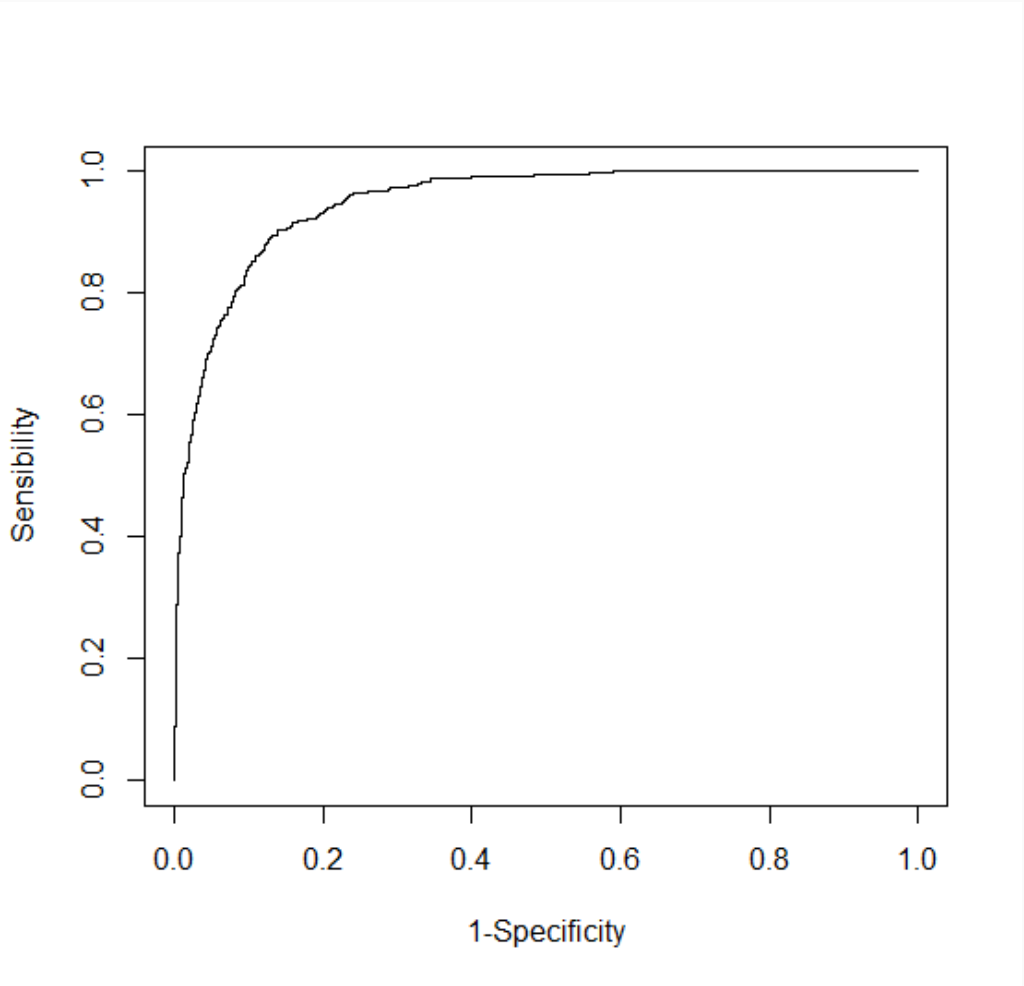
Valor Preditivo Negativo (NPV) $= \frac{8325}{8325 + 32} = 99.62$

$$\mathrm{MCC} = \frac{301 \times 8325 - 1342 \times 32}{\sqrt{(301 + 1342) \times (301 + 32) \times (8325 + 1342) \times (8325 + 32)}} = 37.05$$

Curva ROC

- Ferramenta gráfica para visualizar a capacidade de predição do modelo de acordo com diversos pontos de corte obtidos elas probabilidades estimadas
- Variando os pontos de corte, também variamos as medidas de desempenho
- A curva ROC é obtida a partir de duas medidas:
  - 1-Especificidade: proporção de negativos classificados como positivos
  - Sensibilidade: proporção de positivos classificados como positivos
- Avaliar a área sob a curva!

Curva ROC



$$AUC = 94.96$$

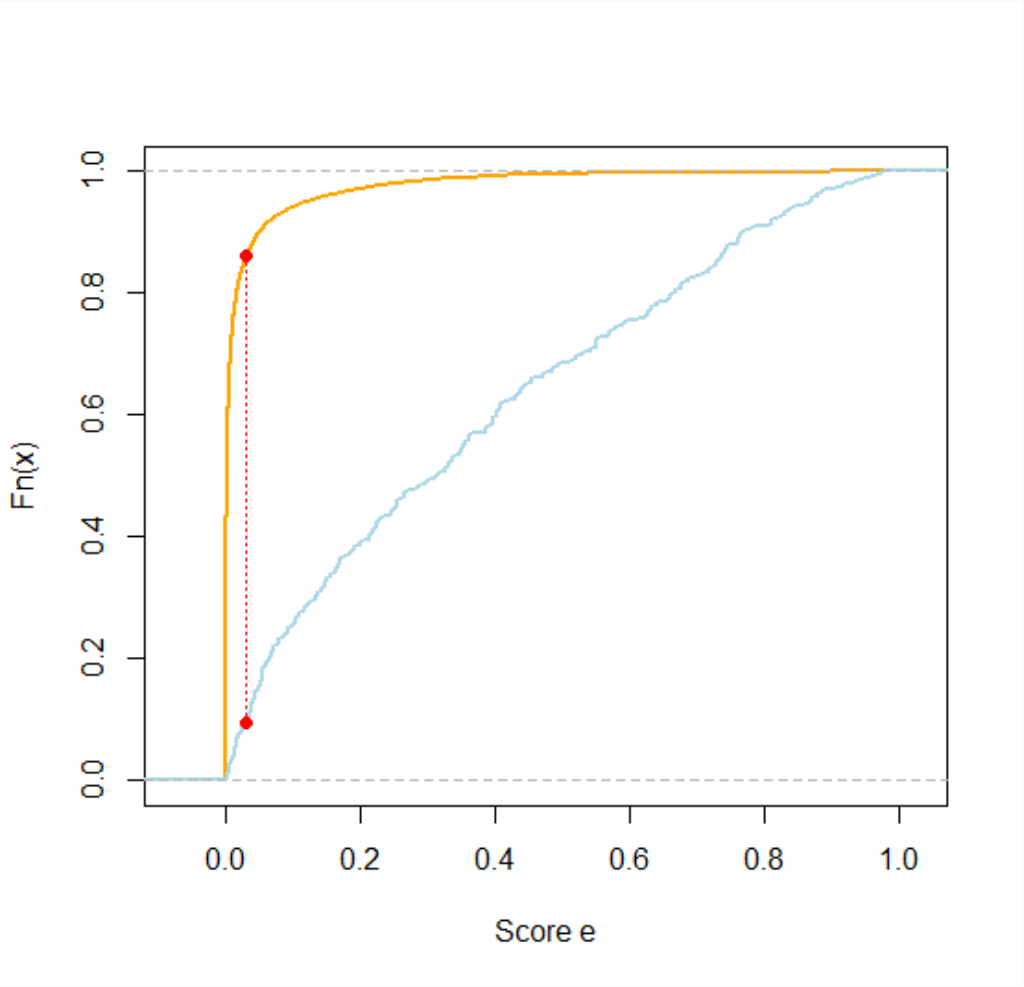
Estatística de Kolmogorov-Smirnov (KS)

- Baseada no teste não paramétrico de Kolmogorov-Smirnov e testa se duas funções associadas a duas populações são idênticas ou não
- A estatística KS mede o quanto as funções de distribuições empíricas dos escores dos grupos positivos ($Y = 1$) e negativos ($Y = 0$)

$$KS = \text{máx}|F_P(e) - F_N(e)|$$

- Em que $F_P(e)$ e $F_N(e)$ são as proporções de positivos e negativos com escore menor ou igual a e

Estatística de Kolmogorov-Smirnov (KS)



$$KS = 0.7651$$

Como selecionar variáveis?

- Forward
  - Começa do modelo nulo e adiciona variável a cada passo
- Backward
  - Começa do modelo completo e remove variável a cada passo
- Stepwise
  - Combinação dos dois anteriores e cada passo testa se variáveis devem ser incluídas ou removidas
- Subset
  - Testa todas as combinações possíveis das variáveis

Validação do modelo

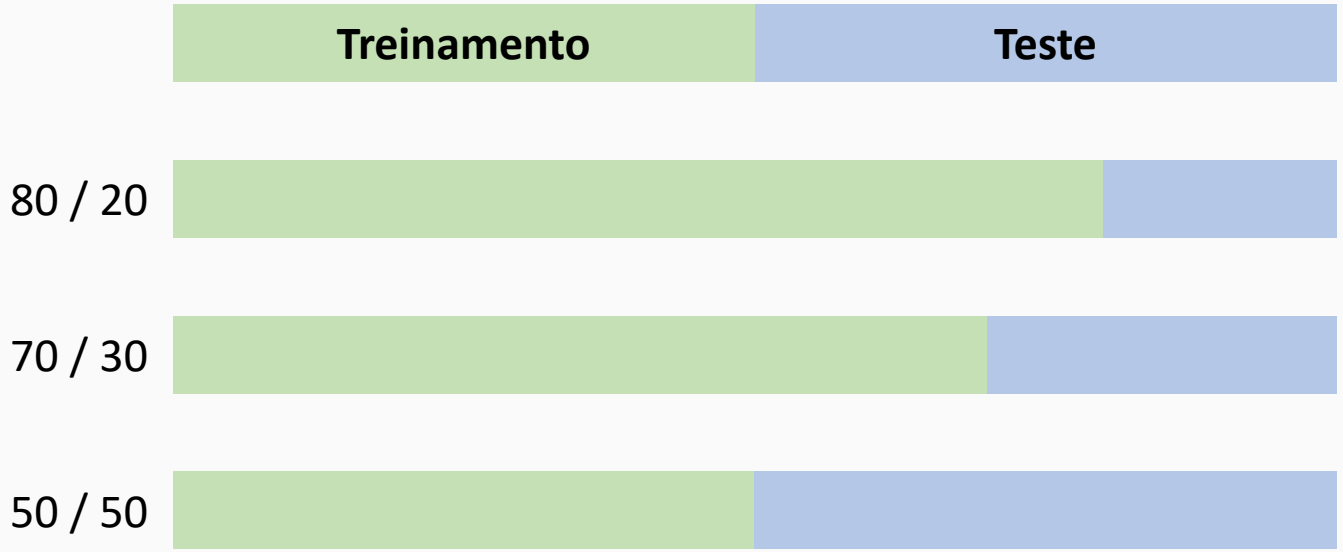
- Dividimos a amostra aleatoriamente em duas partes
  - Treinamento
  - Teste

- O modelo é ajustado com os dados da amostra treinamento e usado para prever as respostas da amostra teste

- Objetivo: evitar superestimação do modelo

Validação do modelo

- Podemos separar a amostra de várias formas



K-fold cross-validation

- Dividimos a amostra em K partes iguais.
- Separamos uma parte k e ajustamos o modelo nas outras $K - 1$ partes conjuntamente.
- Fazemos esse procedimento para todas as partes $k = 1,2, \dots, K$ e combinamos os resultados

- $K = 5$

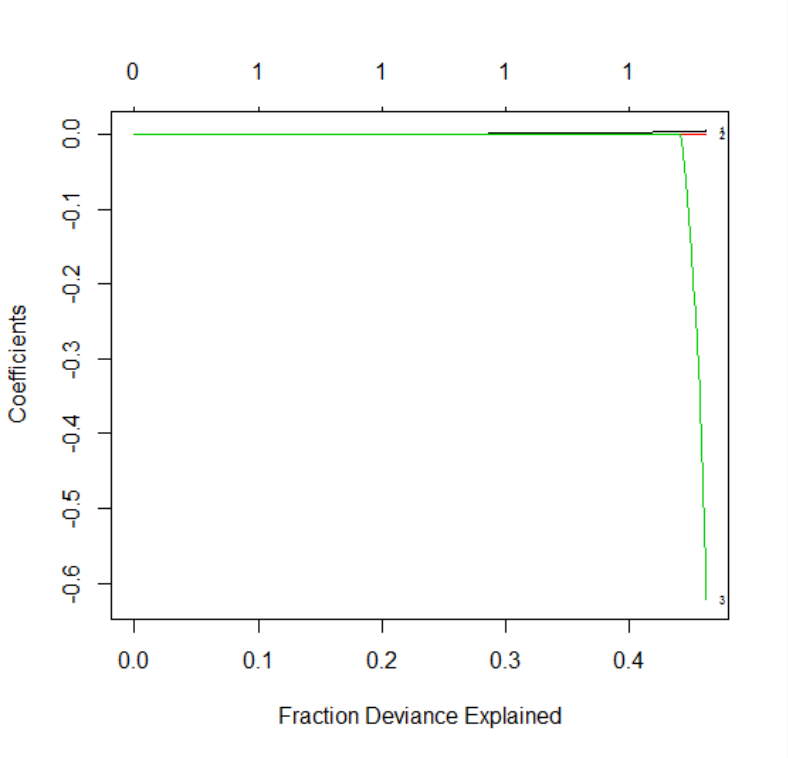
1	2	3	4	5
Teste	Treinamento	Treinamento	Treinamento	Treinamento

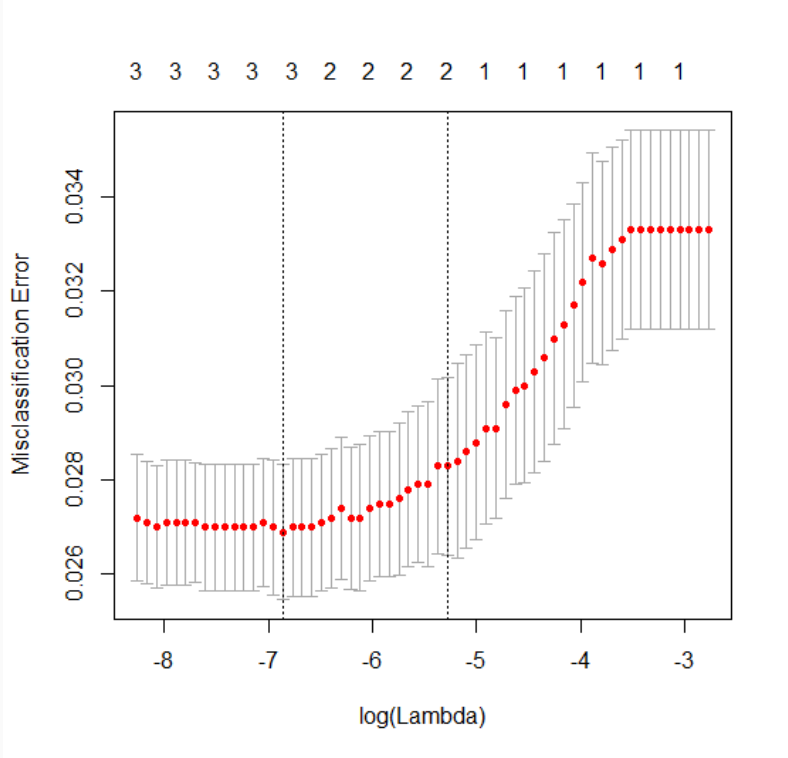
Métodos alternativos - Lasso

- O Lasso faz os coeficientes tenderem a zero
- Método utilizado para selecionar variáveis
- Utiliza penalização dos parâmetros
- Minimiza a quantidade

$$\sum_{i=1}^{n}\left(Y_i - \beta_0 - \sum_{j=1}^{p}\beta_j X_{ij}\right)^2 + \lambda\sum_{j=1}^{p}|\beta_j| = \mathrm{RSS} + \lambda\sum_{j=1}^{p}|\beta_j|$$

Métodos alternativos - Lasso





λ=0.001

	Logística	Lasso
Intercepto	-10.8690	-10.3595
Balance	0.0057	0.0054
Income	0.0000	0.0000
Student [Yes]	-0.6468	-0.5475

Métodos alternativos - Outros

- Modelo Binário com outras funções de ligação
  - Probit
  - Complementary log-log
- Árvore de Decisão
- Redes Neurais
- Redes Bayesianas
- Support Vector Machine
- Análise Discriminante

Obrigado!