

**Tera**

# Módulo 4:

## Estatística e Modelagem de Dados

### Aula 17: Regressão Logística





# Instrutora

## Cristiane Rodrigues

- **Bacharel em Matemática – UNESP Rio Claro.**
- **Mestre em Estatística – USP Piracicaba**
- **Experiências Profissionais:**
  - Modelagem de Credito para PF e PJ – Banco Bradesco
  - Experiência com Segmentação e Análise de Series temporais – Atento
  - Consultora Analítica no SAS Institute Brasil
  - Professora do curso SAS Academy for Data Science



# Índice

- Revisão Regressão Logística
- Motivação
- Forma Funcional do Modelo de Regressão Logística
- Aplicações
- Superfície de Ajuste e Interpretação
- Odds Ratio
- Ponto de Corte e Estimação
- Tratamento das variáveis
- Seleção de Variáveis
- Matriz de Confusão
- Curva ROC



# Revisão: Supervised X Unsupervised Learning

- **Cenário 1:** Você é uma criança e vê diferentes tipos de animais, seu pai lhe diz que esse animal é um cão ... depois dele, dando algumas dicas, você vê um novo tipo de cachorro, que você nunca viu antes, mas você o identifica como um cão e não como gato ou macaco ou batata.

Este cenário é um exemplo de classificação supervisionada, onde você tem alguém para orientá-lo e aprender conceitos, de modo que, quando uma nova amostra chega ao seu caminho, mesmo que você não tenha visto antes, você ainda pode identificá-la.

- **Cenário2:** Você vai fazer uma viagem para um novo país, o qual você não conhece muito sobre sua comida, cultura, idioma, etc. No entanto, a partir do primeiro dia, você começa a caminhar por lá, aprendendo o que comer e o que não comer, encontrar um caminho para a praia ou para o hotel, etc.

Este cenário é um exemplo de classificação não supervisionada, onde você tem muitas informações, mas não sabe o que fazer inicialmente. Uma distinção importante é que, não há ninguém para guiá-lo e você tem que encontrar uma saída por conta própria. Então, com base em alguns critérios, você começa a gerar essas informações em grupos que fazem sentido para você.



# Revisão: Supervised Learning

- Usam dados com marcação da variável resposta
- Variáveis preditoras + variável target
- Objetivo: prever a variável target, usando as variáveis preditoras
  - Regressão: variável target é contínua
  - Classificação: variável target é composta de categorias
- Nomenclaturas para as variáveis
  - Independentes = Preditoras = Características = Input
  - Target = Dependente = Resposta



# Revisão: Regressão Linear Simples

- $y = \beta_0 + \beta_1 x + e$ 
  - $y = target$
  - $x = variável\ preditora$  contínua
  - $\beta_0, \beta_1 = parâmetros\ do\ modelo$

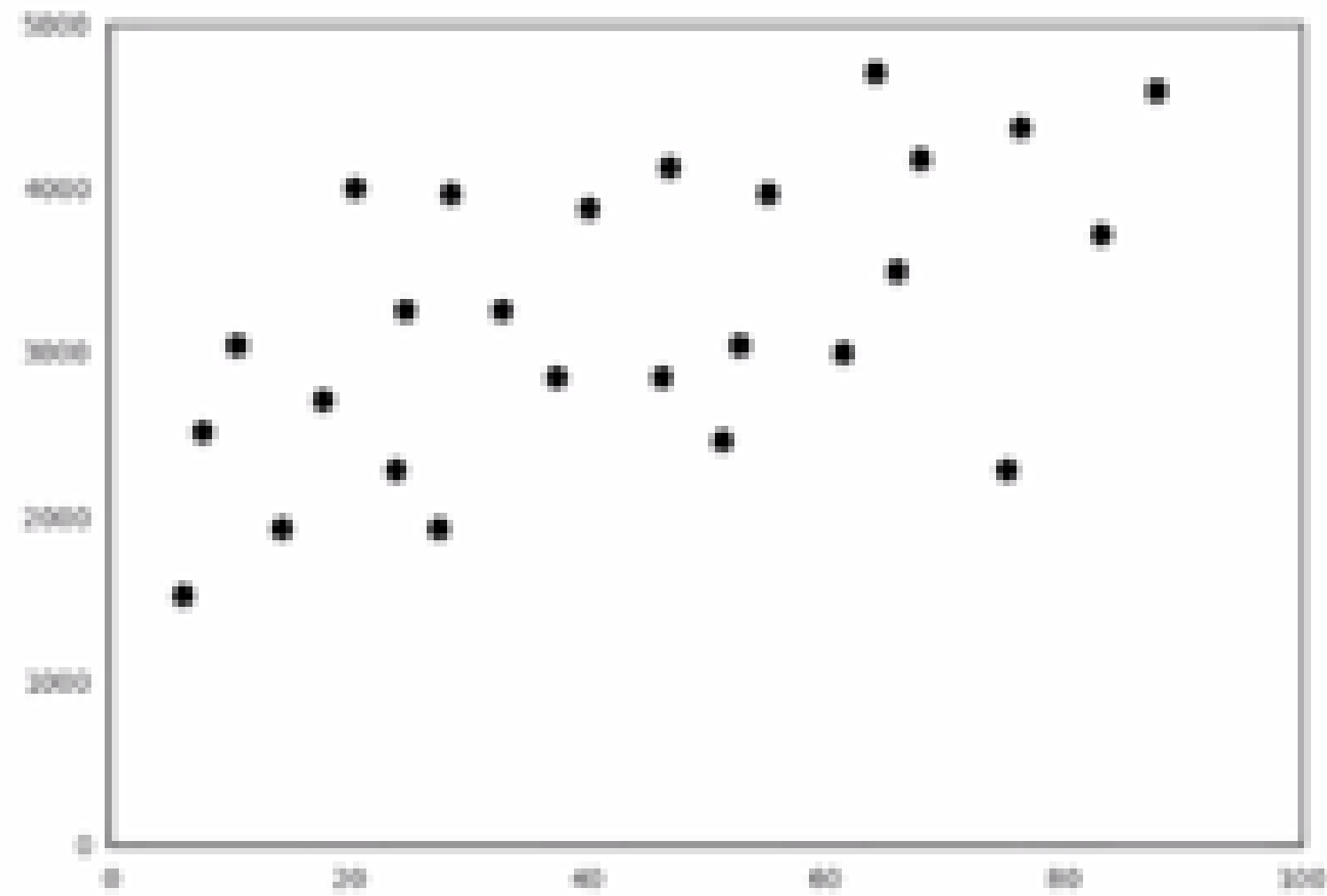
- Como escolher  $\beta_0$  e  $\beta_1$ ?

Os valores de  $\beta_0$  e  $\beta_1$  podem ser estimados pelo método dos mínimos quadrados, minimizando a soma dos erros quadráticos

$$\sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$$

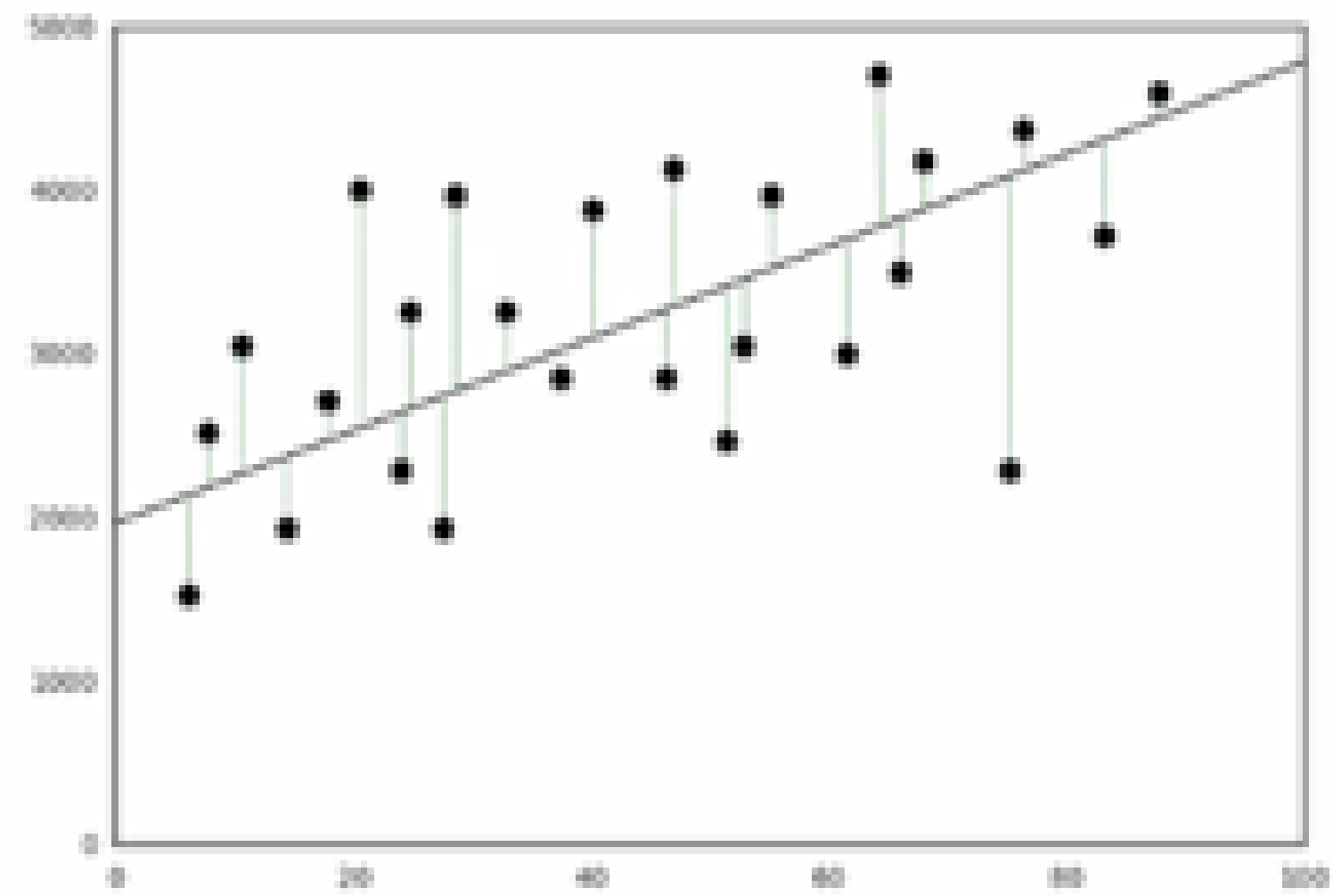
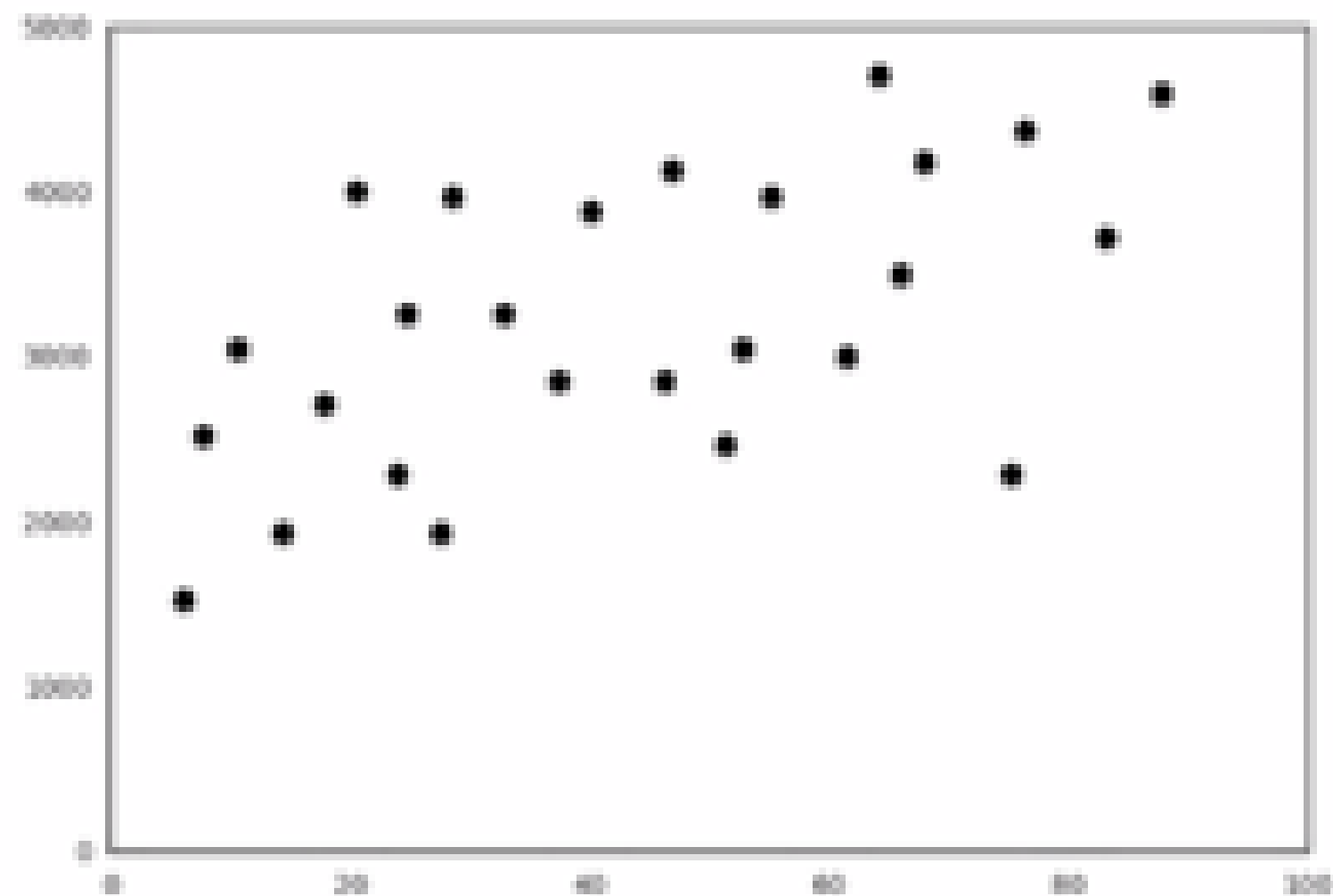


# Revisão: Regressão Linear Simples

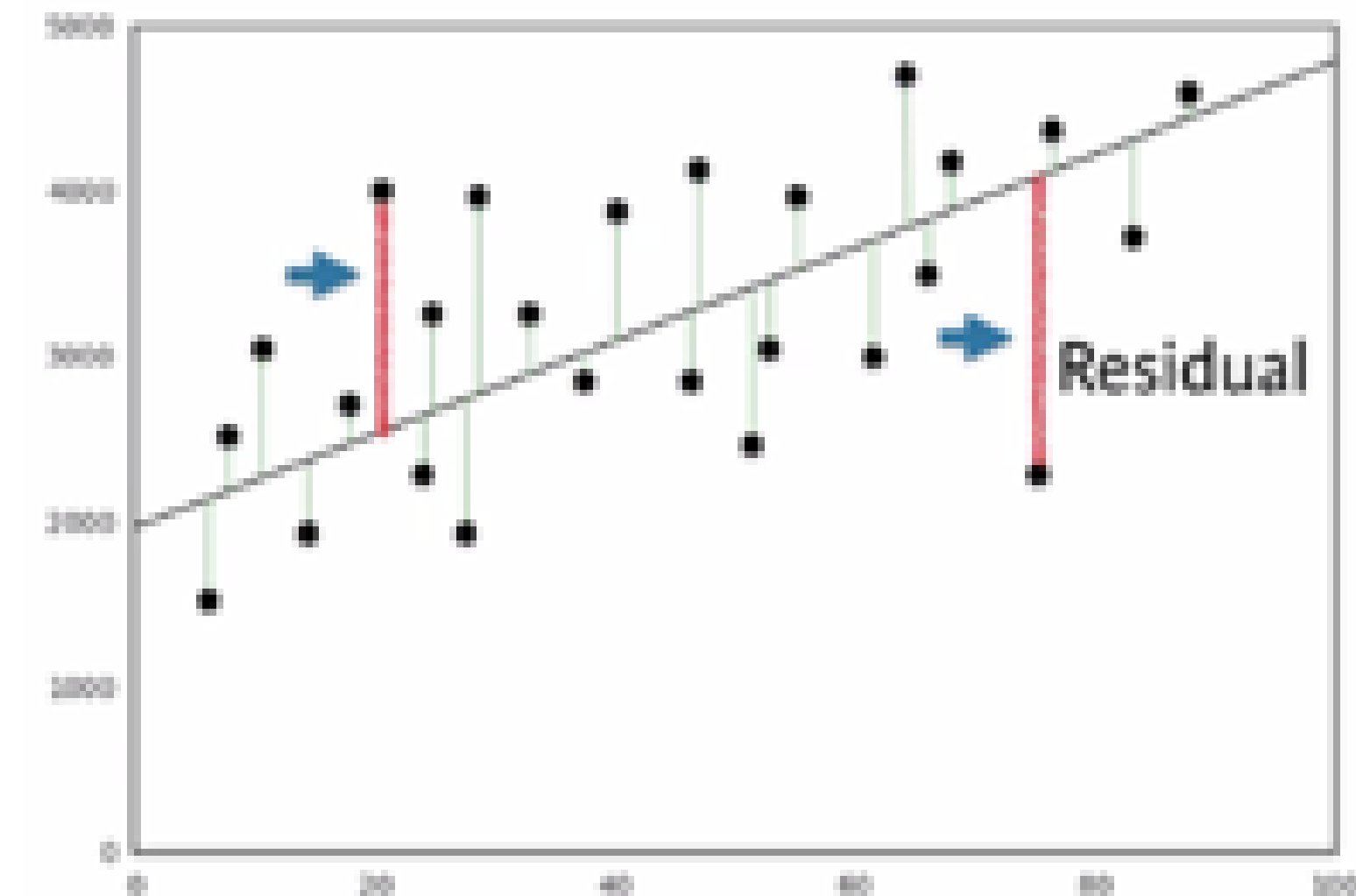
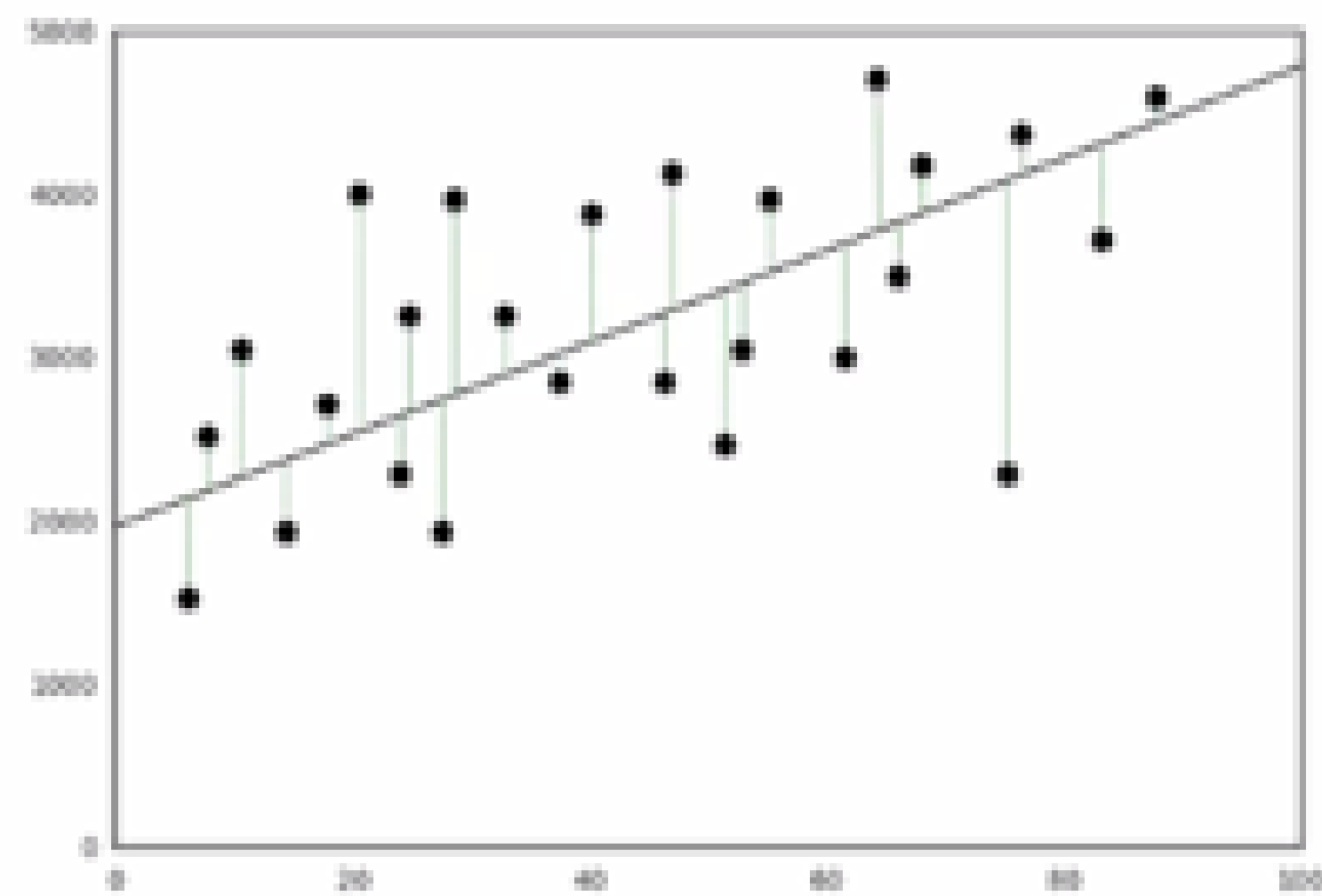
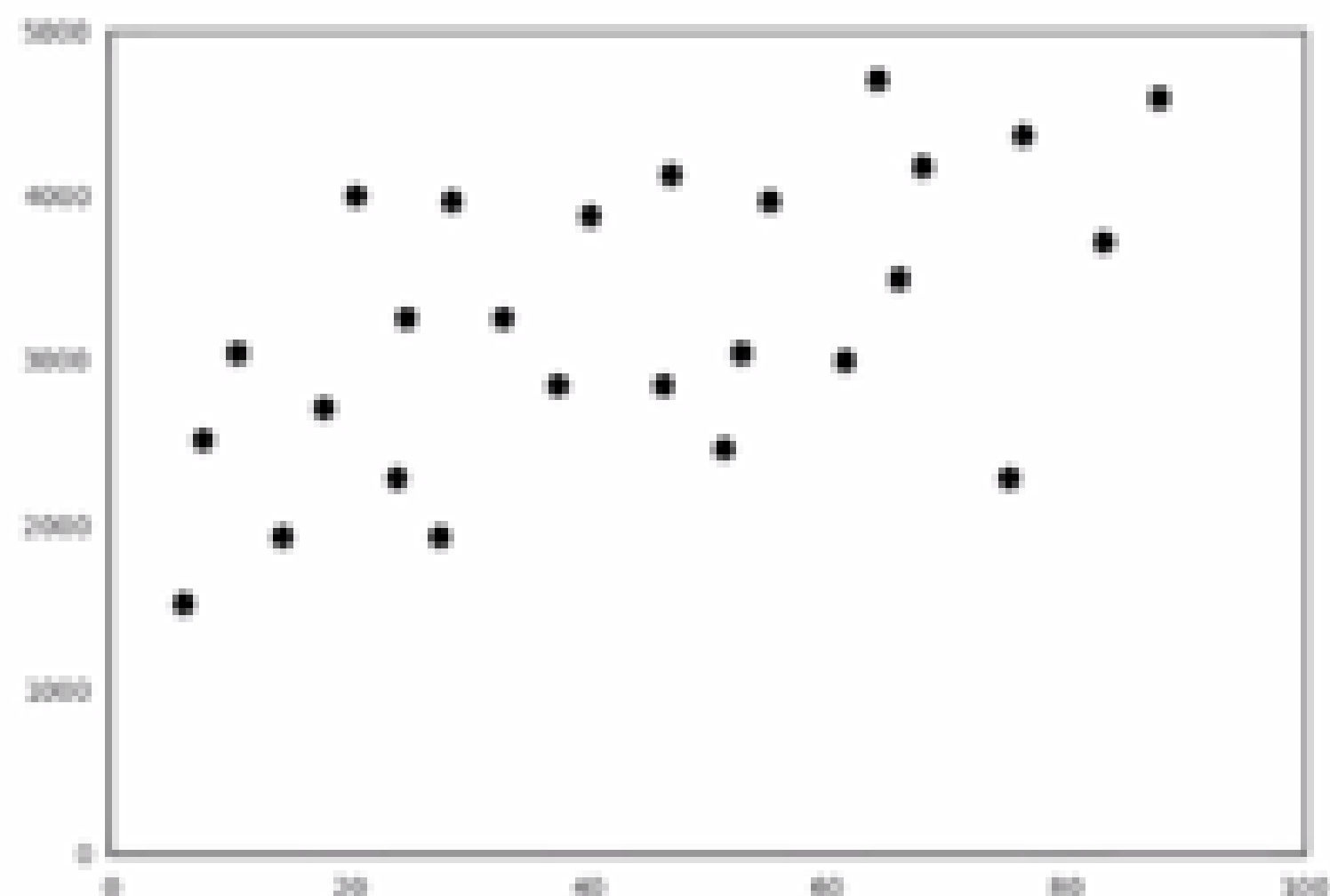




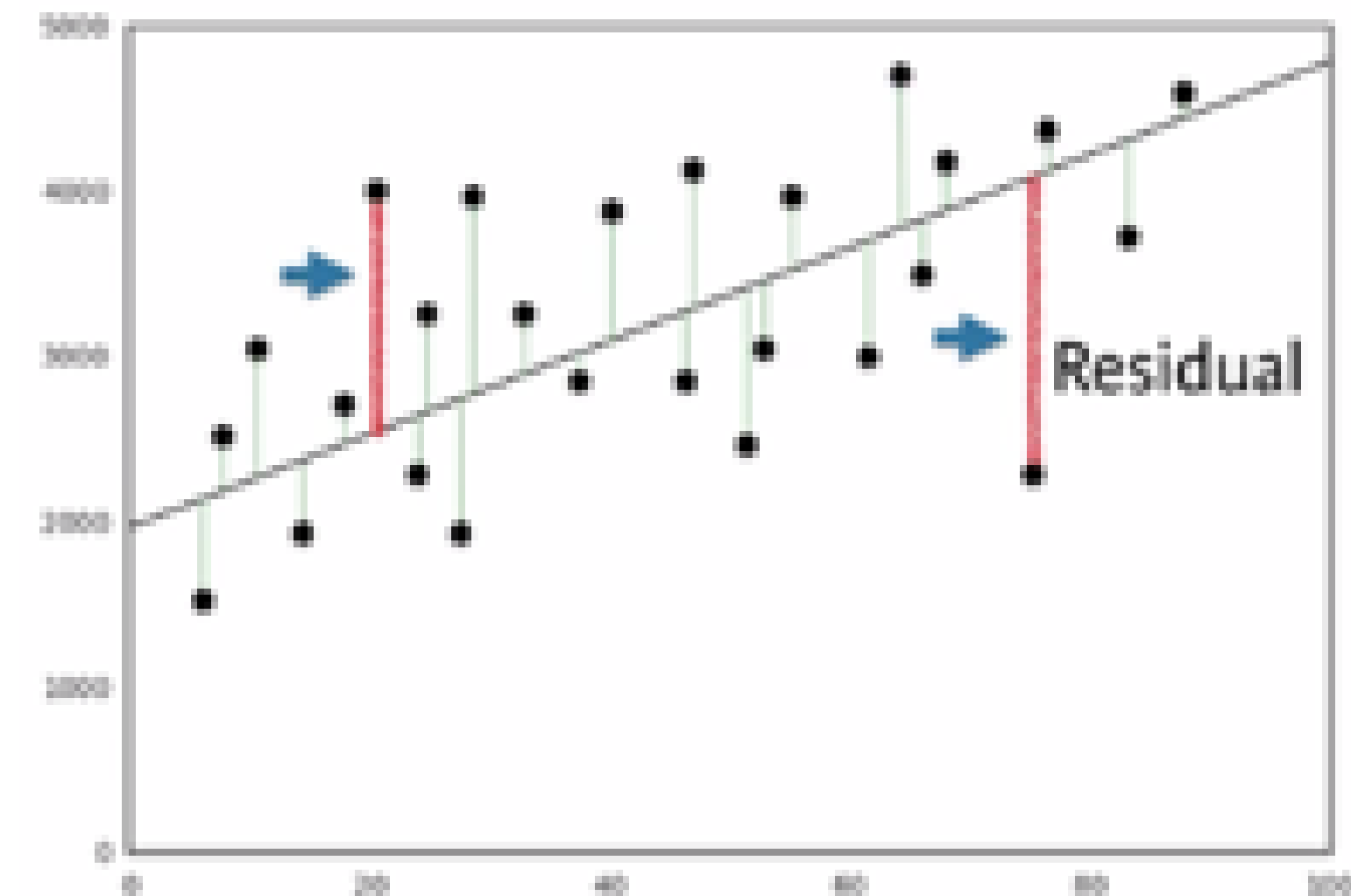
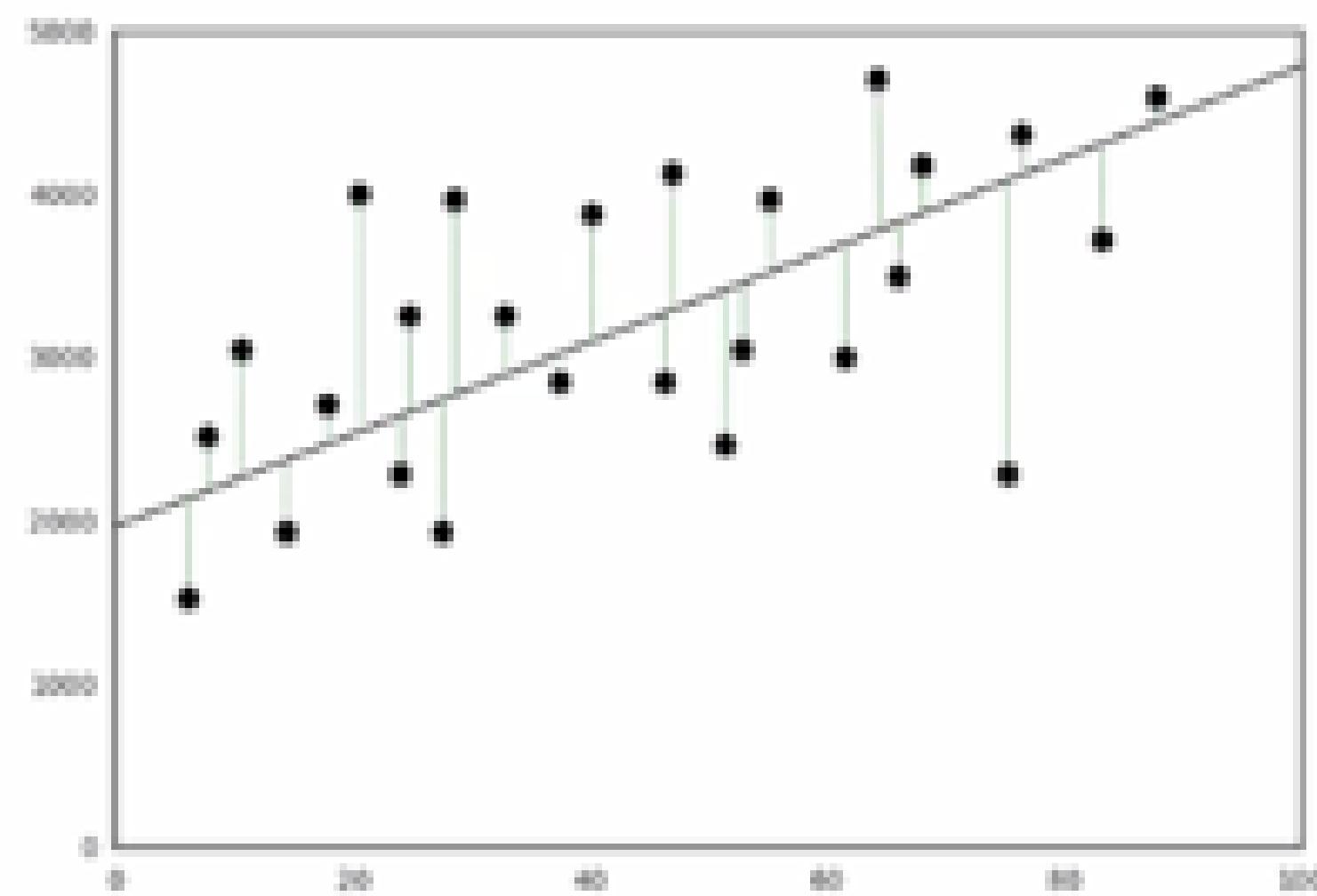
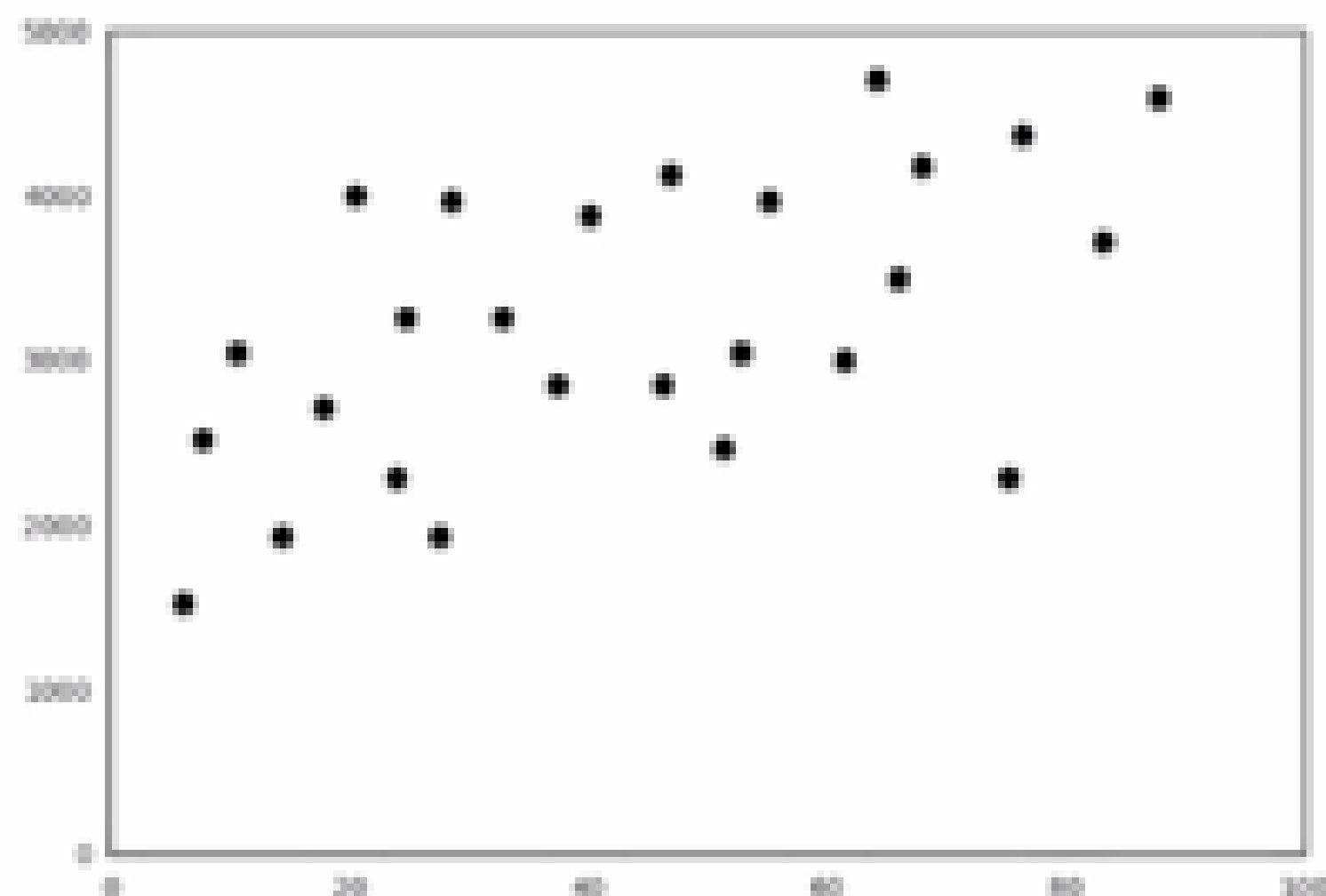
# Revisão: Regressão Linear Simples



# Revisão: Regressão Linear Simples



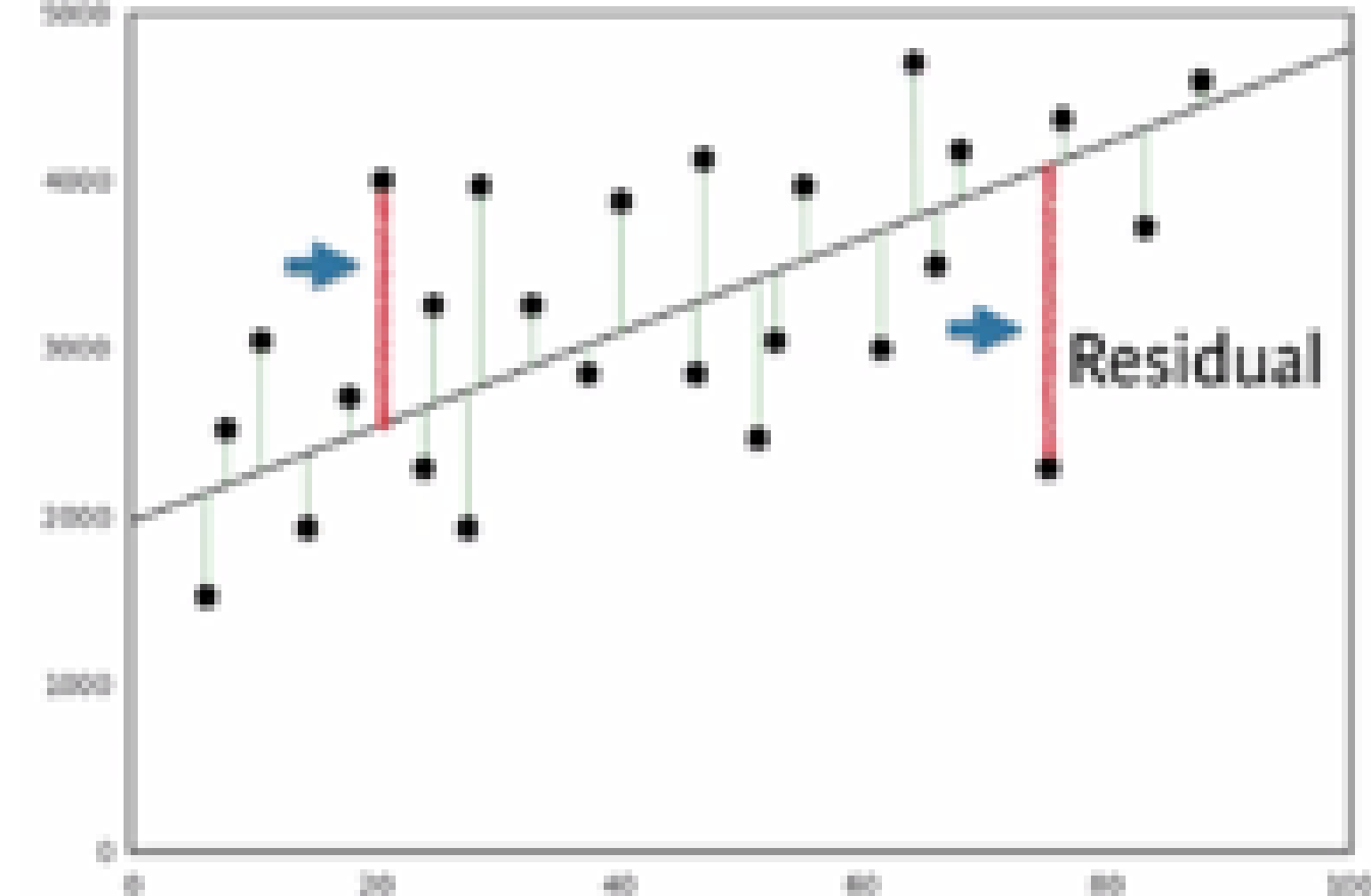
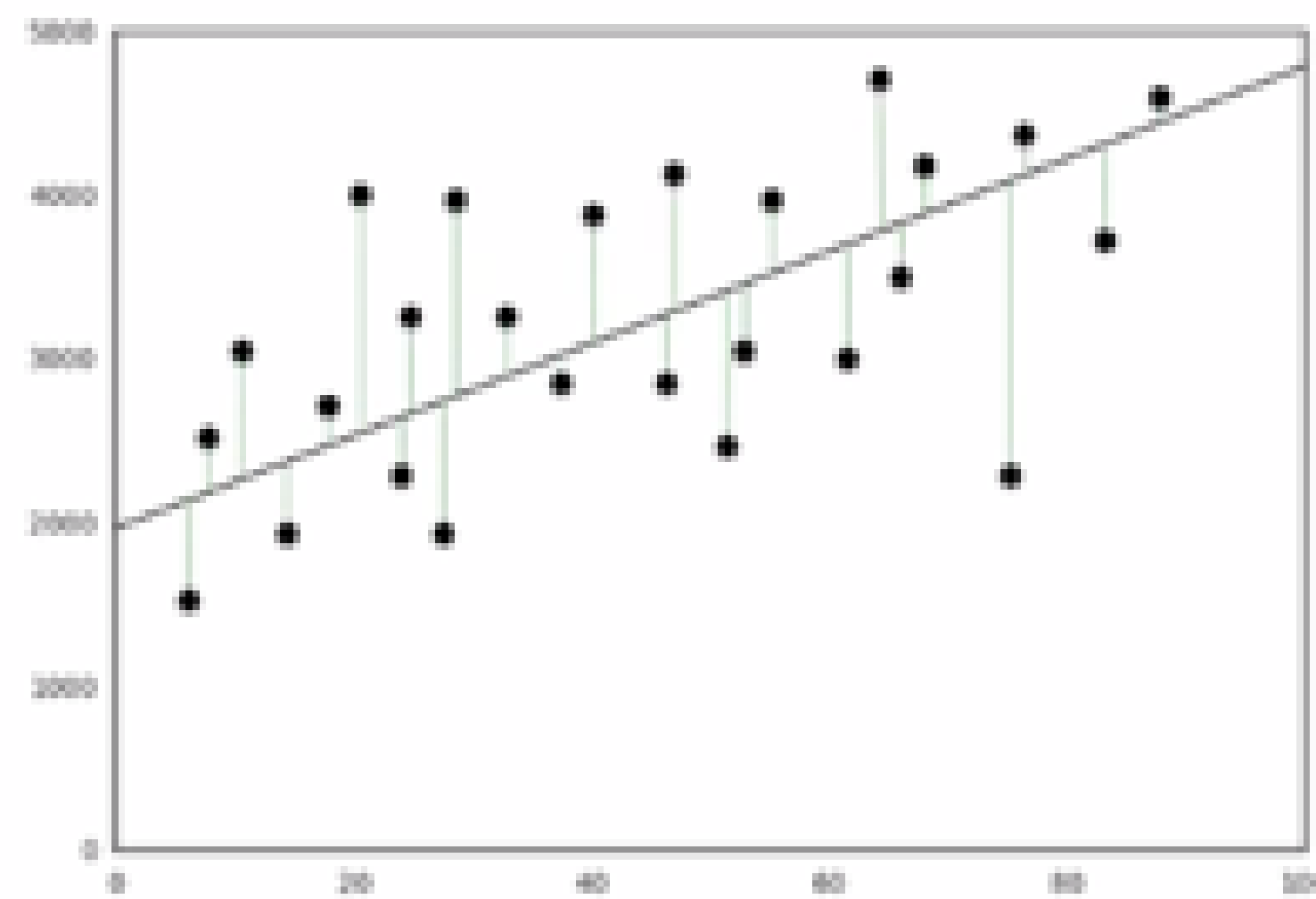
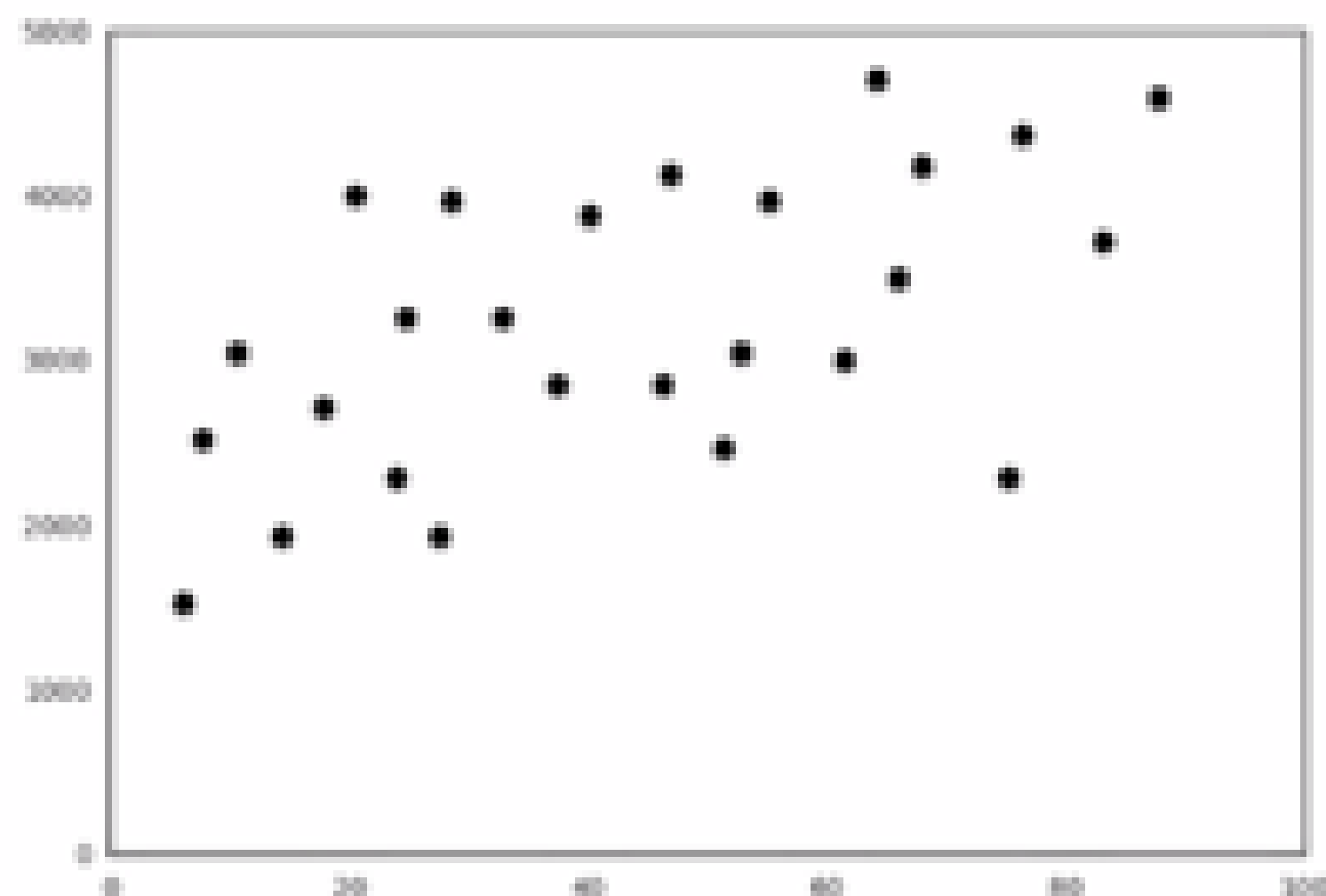
# Revisão: Regressão Linear Simples



- Na regressão linear o objetivo é escolher a reta que minimiza a função de erro, ou seja, que diminui a distância entre o ajuste e os dados



# Revisão: Regressão Linear Simples



- Na regressão linear o objetivo é escolher a reta que minimiza a função de erro, ou seja, que diminui a distância entre o ajuste e os dados
- Na regressão linear múltipla temos a inserção de mais variáveis preditoras e podemos escrever o modelo da seguinte forma:

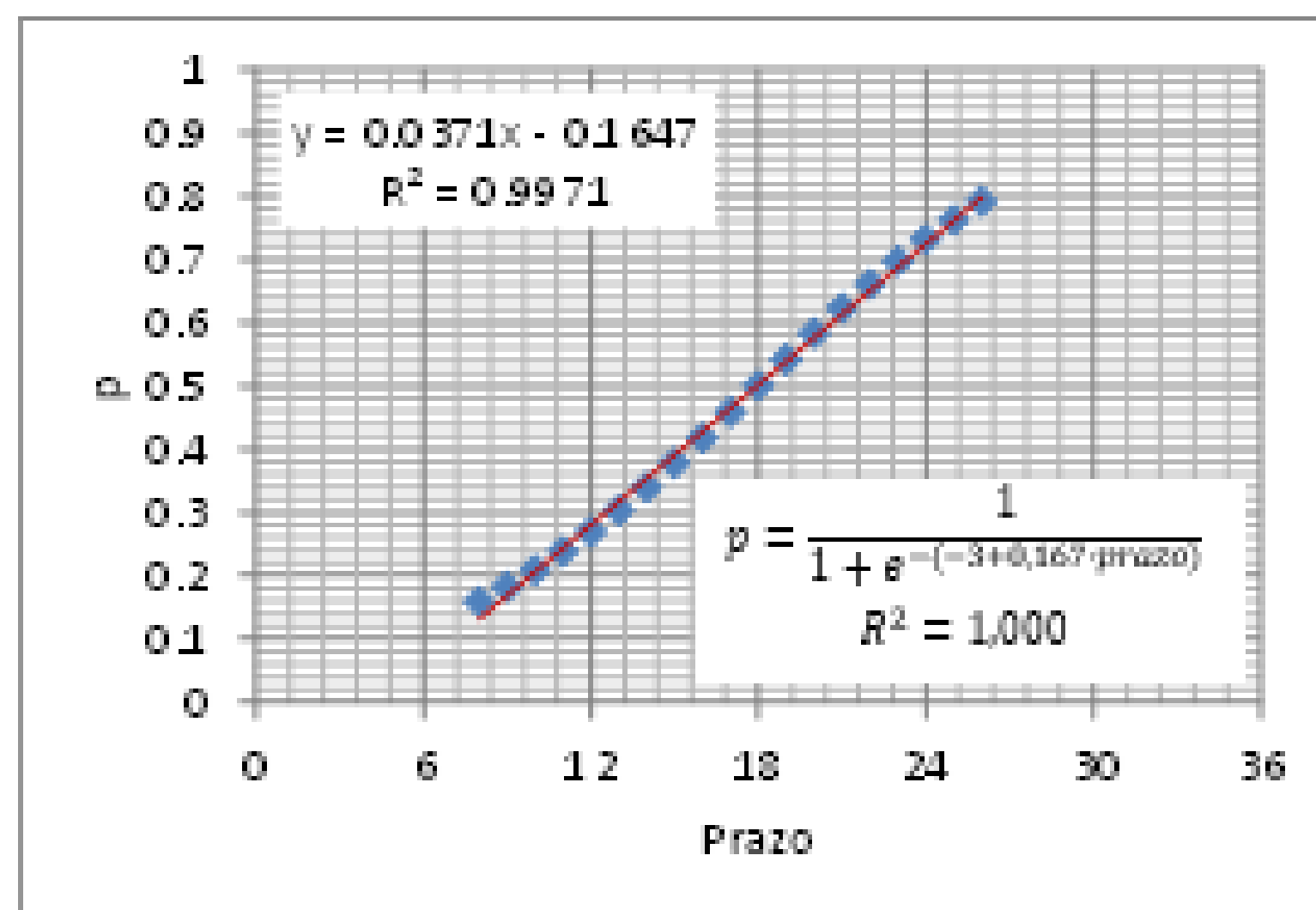
$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k$$



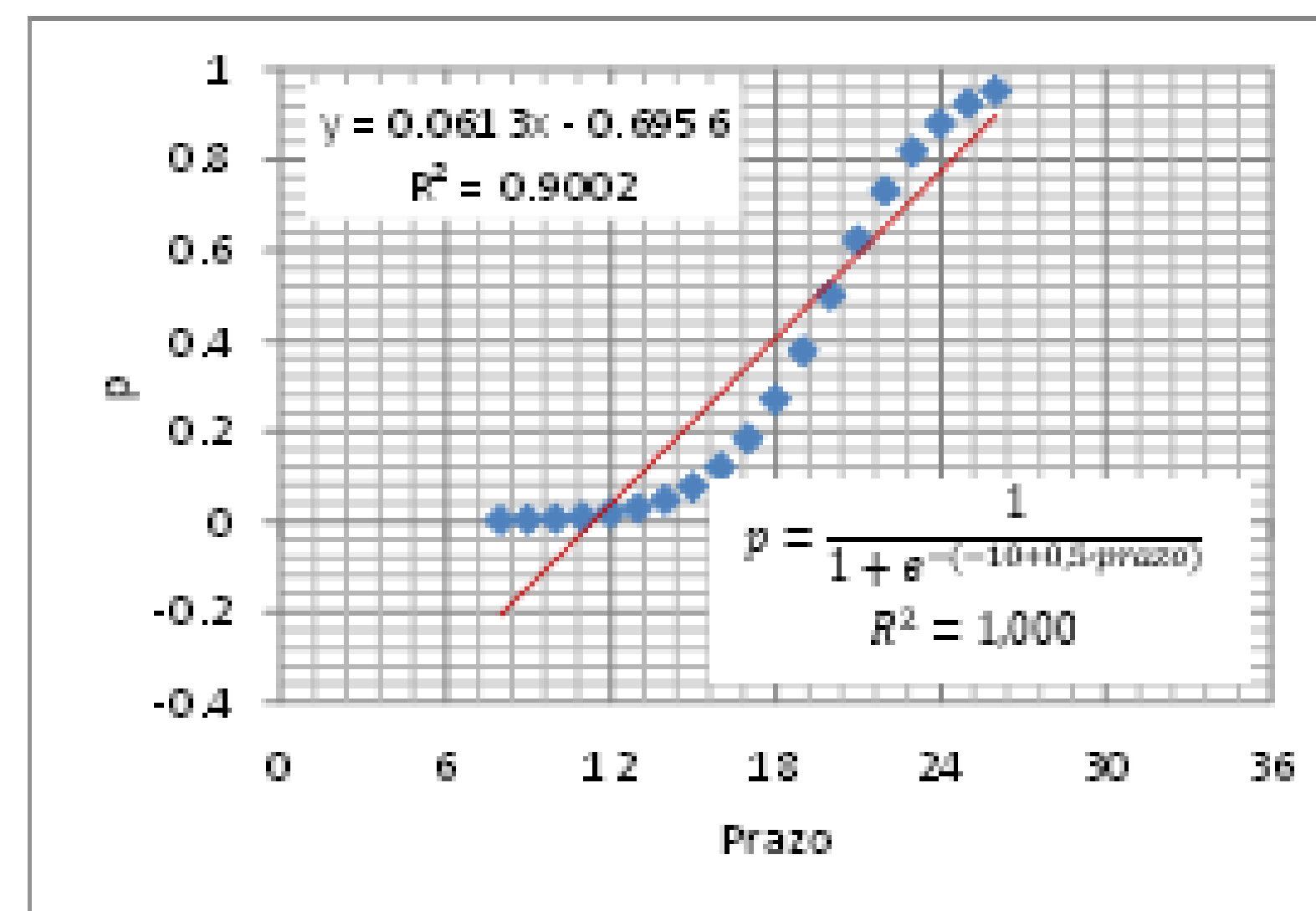
# Comparação

Ajustando modelos de regressão linear para dois tipos de dados diferentes

- Probabilidade variando entre 0,15 e 0,85 substituição
- Probabilidade menor que 0,15 ou maior que 0,85



A equação linear **é suficiente** para modelar bem os dados



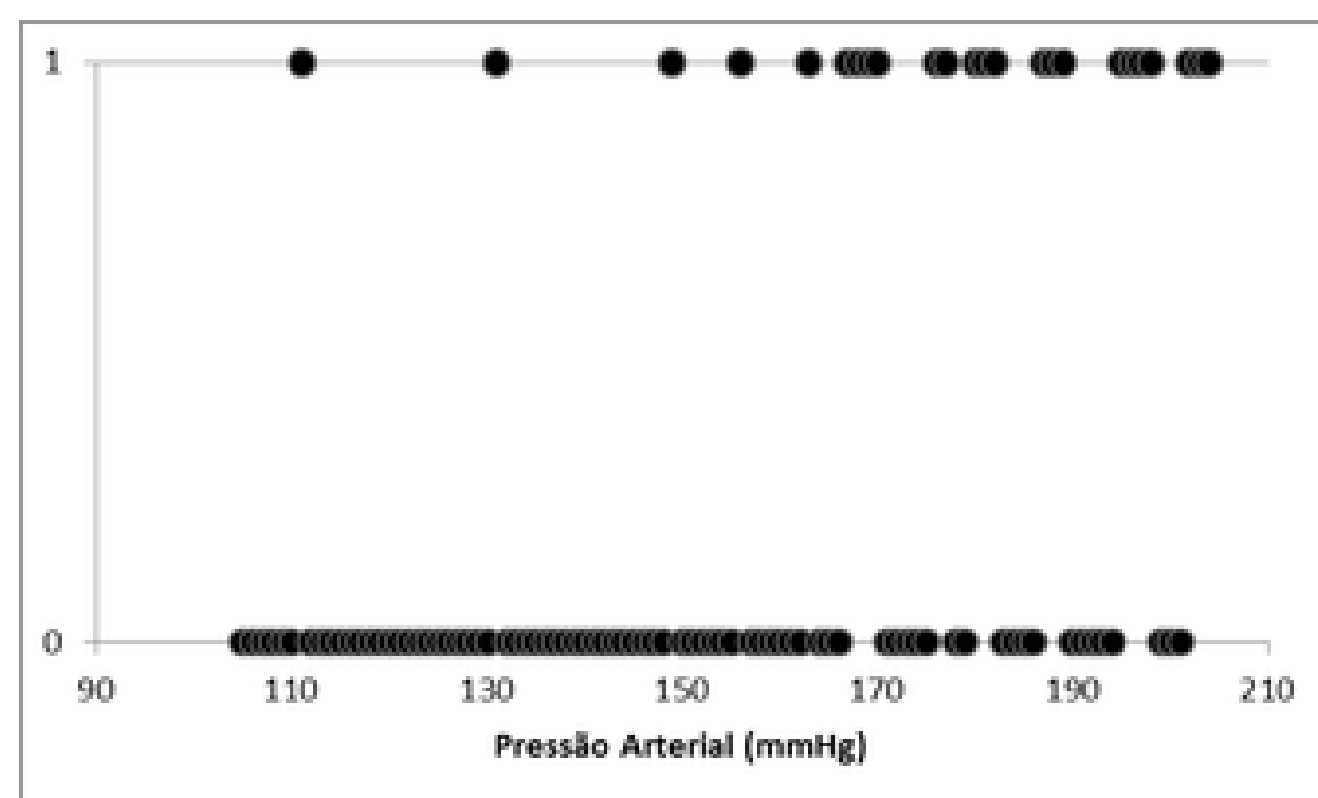
A equação linear **não é suficiente** para modelar bem os dados



# Motivação

- As equações apresentadas no tópico anterior são equações do tipo linear.
- Nem sempre as variáveis se comportam como uma reta, portanto nem sempre uma equação linear será uma equação adequada para descrever o comportamento de uma variável em relação à outra. Isso é especialmente verdade quando temos uma variável binária: 0 ou 1.

Por exemplo: queremos saber os valores de pressão arterial entre pessoas que tiveram ou não um AVC. Se classificarmos “presença de AVC” igual a 1 e “ausência de AVC” igual a 0, teremos um gráfico tipo o abaixo, o qual não parece se ajustar bem com uma reta



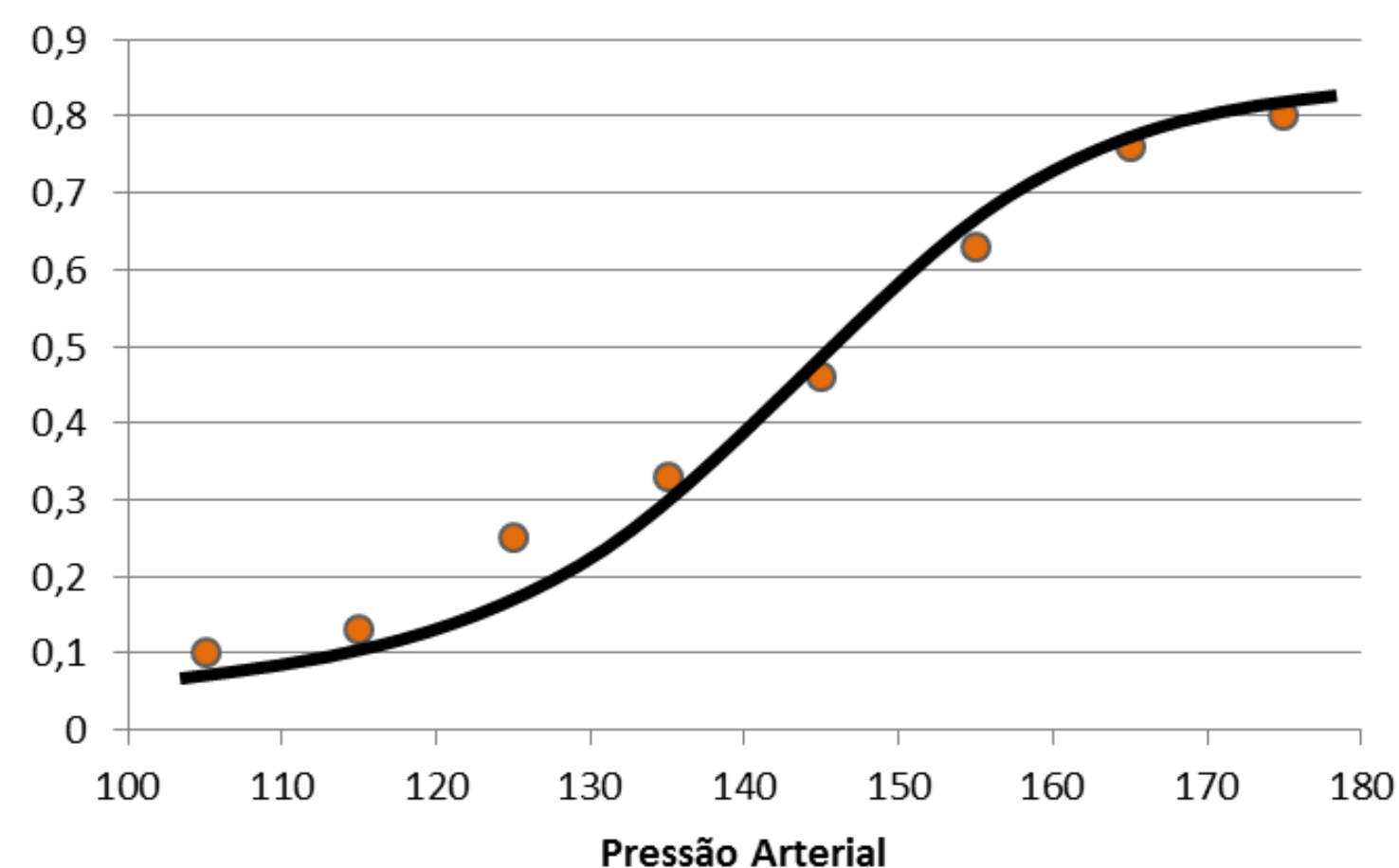
- só tem dois valores: 0 ou 1
- os pontos estão mais concentrados próximos:
  - ao valor 0, em que os valores de pressão arterial são mais baixos
  - ao valor 1, em que os valores de pressão arterial são mais altos
- significa que: provavelmente à medida que aumenta a pressão arterial, aumenta a incidência de AVC.

Mas em quanto?



# Forma Funcional

- Quando transformamos uma variável com valores 1 e 0 em proporções, acontece um fenômeno que o gráfico fica mais ou menos assim:



- Alguns estatísticos perceberam que essa curva poderia ser escrita em forma de função, porém ela não é linear, mas sim bem mais complexa, e pode ser descrita assim:

$$p_i = \frac{1}{1 + e^{-n}},$$

em que  $p_i$  é a proporção de eventos para cada  $x_i$  e

$$n = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$



# Função de ligação logit

- Essa probabilidade em forma de S é muito difícil de interpretar pois o y aumenta em velocidades diferentes ao longo do eixo x.
- A ideia é tornar a equação uma reta novamente para ficar mais fácil de interpretar o resultado
- Para fazer isso, vamos utilizar a transformação Logit, a qual é composta por duas transformações
  1. Transformar o p em uma chance:  $\frac{p}{1-p}$
  2. Aplicando o logaritmo a chance

$$\begin{aligned} \text{logit}(p_i) &= \ln\left(\frac{p_i}{1-p_i}\right) = \ln\left(\frac{\frac{1}{1+e^{-n}}}{1-\frac{1}{1+e^{-n}}}\right) = \ln\left(\frac{\frac{1}{1+e^{-n}}}{\frac{e^{-n}}{1+e^{-n}}}\right) = \ln\left(\frac{1}{e^{-n}}\right) = \ln(e^n) = n \\ &= \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki} \end{aligned}$$

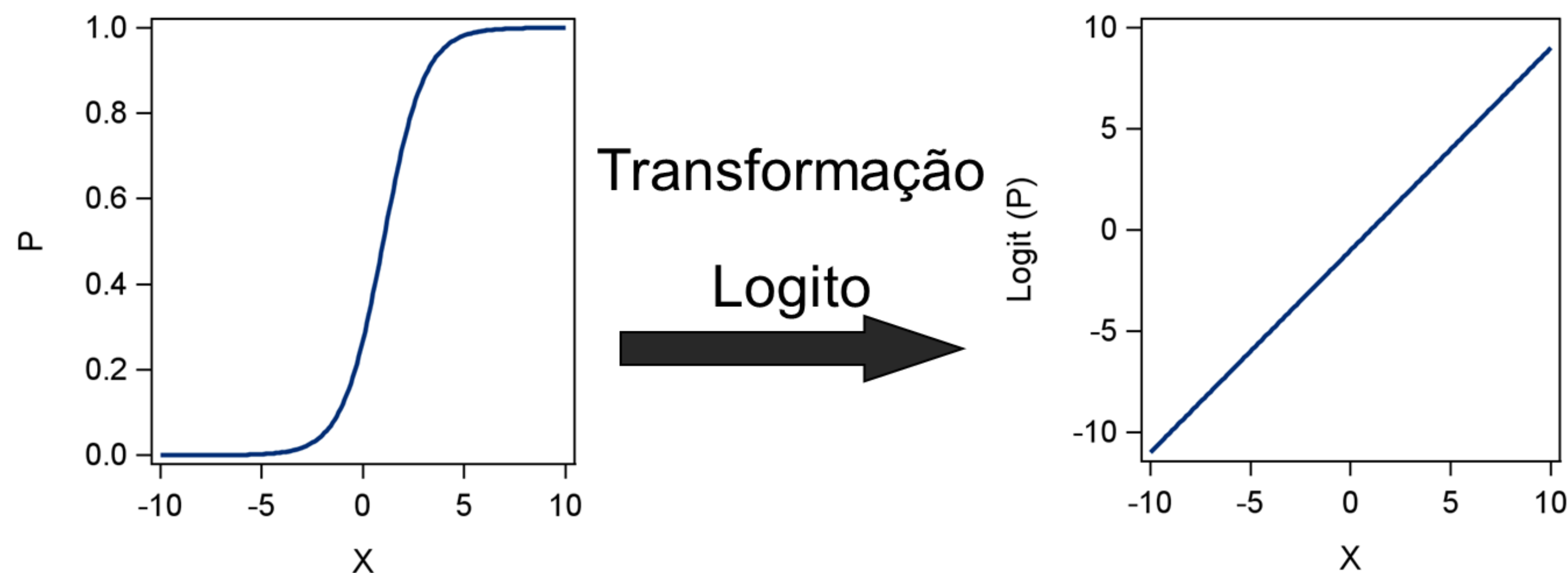
Desta forma voltamos para uma relação linear entre o logit de  $p_i$  e as variáveis input.



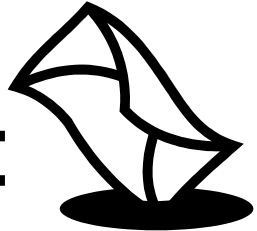





# Transformação

- Usando a transformação Logito podemos sair de um problema não linear e voltar para a modelagem de um problemas linear.



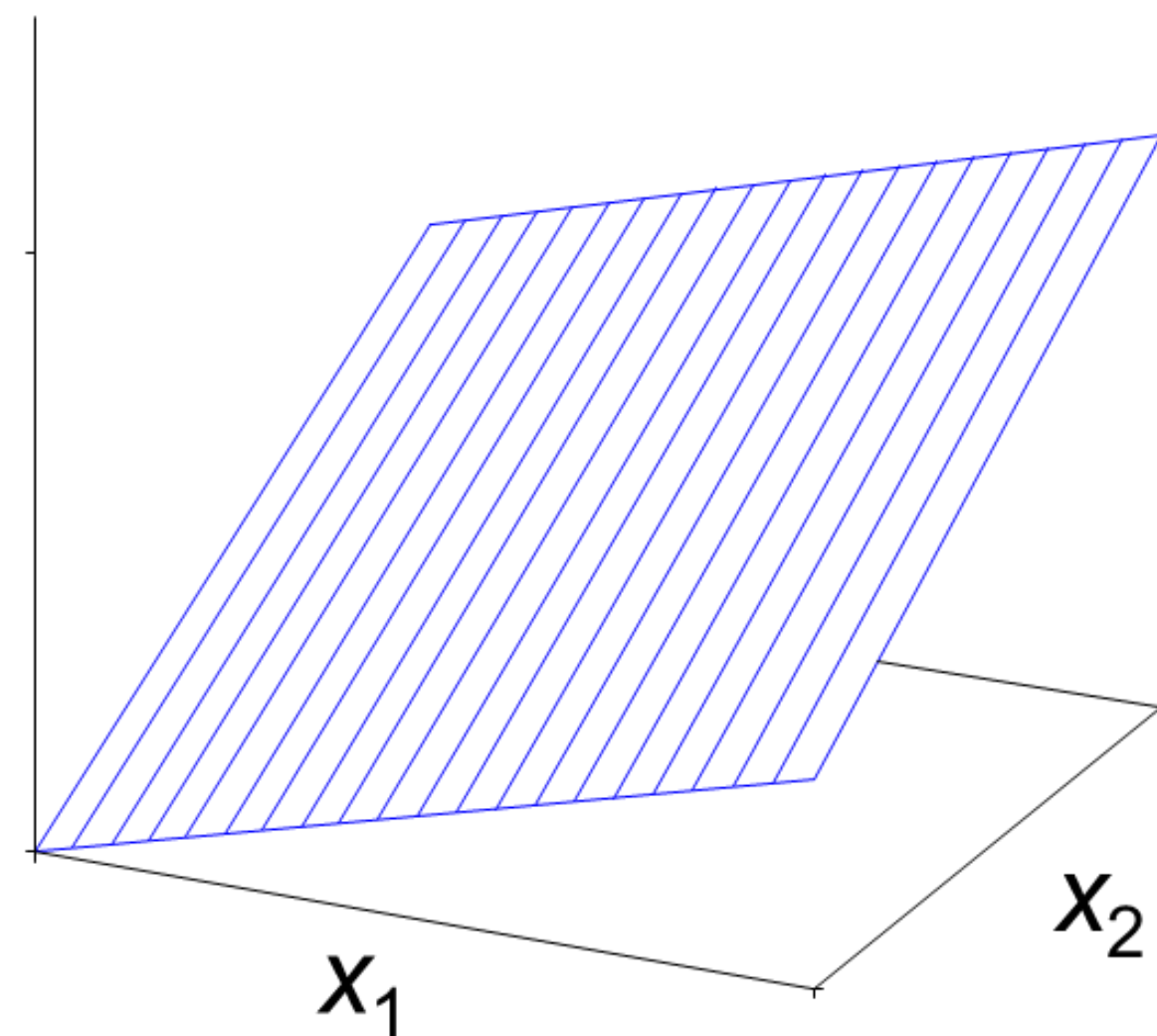
# Aplicações da Regressão Logística - Classificação

- Marketing:   
Objetivo: Encontrar segmentos de clientes mais prováveis a aderir a uma promoção  
Target: Se o cliente aderiu ou não a alguma promoção passada  
Inputs: Histórico de compras, Localidade, Salário,...
- RH – Pedido de demissão de funcionários:   
Objetivo: Verificar a probabilidade do funcionário deixar a empresa  
Target: Se o funcionário saiu ou não da empresa no mês anterior  
Inputs: Tempo de serviço, nível de satisfação, salário, cargo,...
- Credit Scoring:   
Objetivo: Verificar a probabilidade do cliente entrar em default  
Target: Se o funcionário entrou ou não e default nos últimos 90 dias  
Inputs: Saldo médio em cc, se recebe em conta, saldo máximo, quantidade de meses em risco
- Detecção de Fraude:   
Objetivo: Verificar fraude ou abuso em novas transações ou solicitações  
Target: Se o cliente cometeu ou não fraude na transação com cartão de crédito pela internet  
Inputs: Valor médio de pagamento por sessão, número de sessões abertas, ...



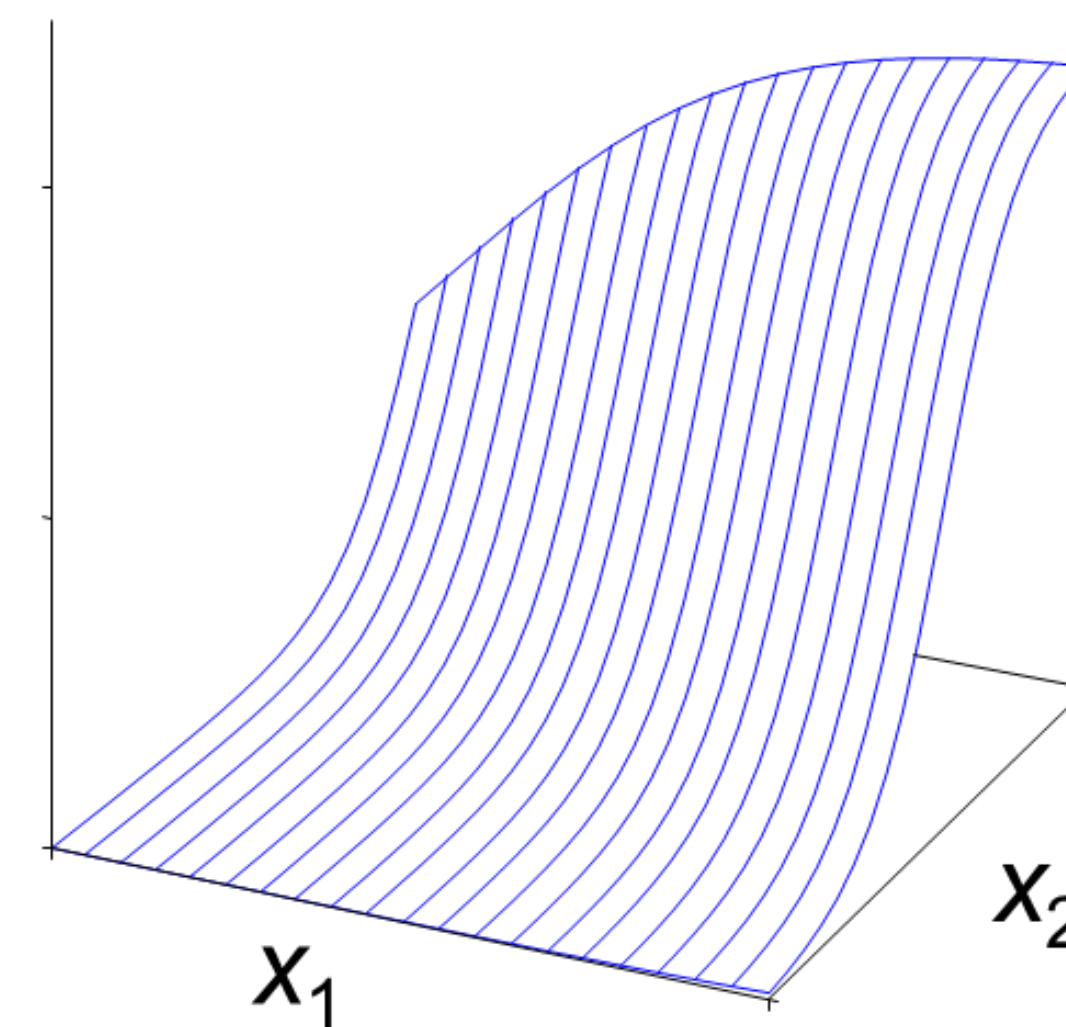
# Superfície de Ajuste e Interpretação

$$\text{logit}(p_i) = \beta_0 + \beta_1 x_1 + \beta_2 x_2$$



$$\text{logit}(p) \in (0, \infty)$$

$$p = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2)}}$$



$$(p) \in (0, 1)$$

Interpretação:  
Mudar uma unidade  
em  $x_2$

$\beta_2$  muda na logit

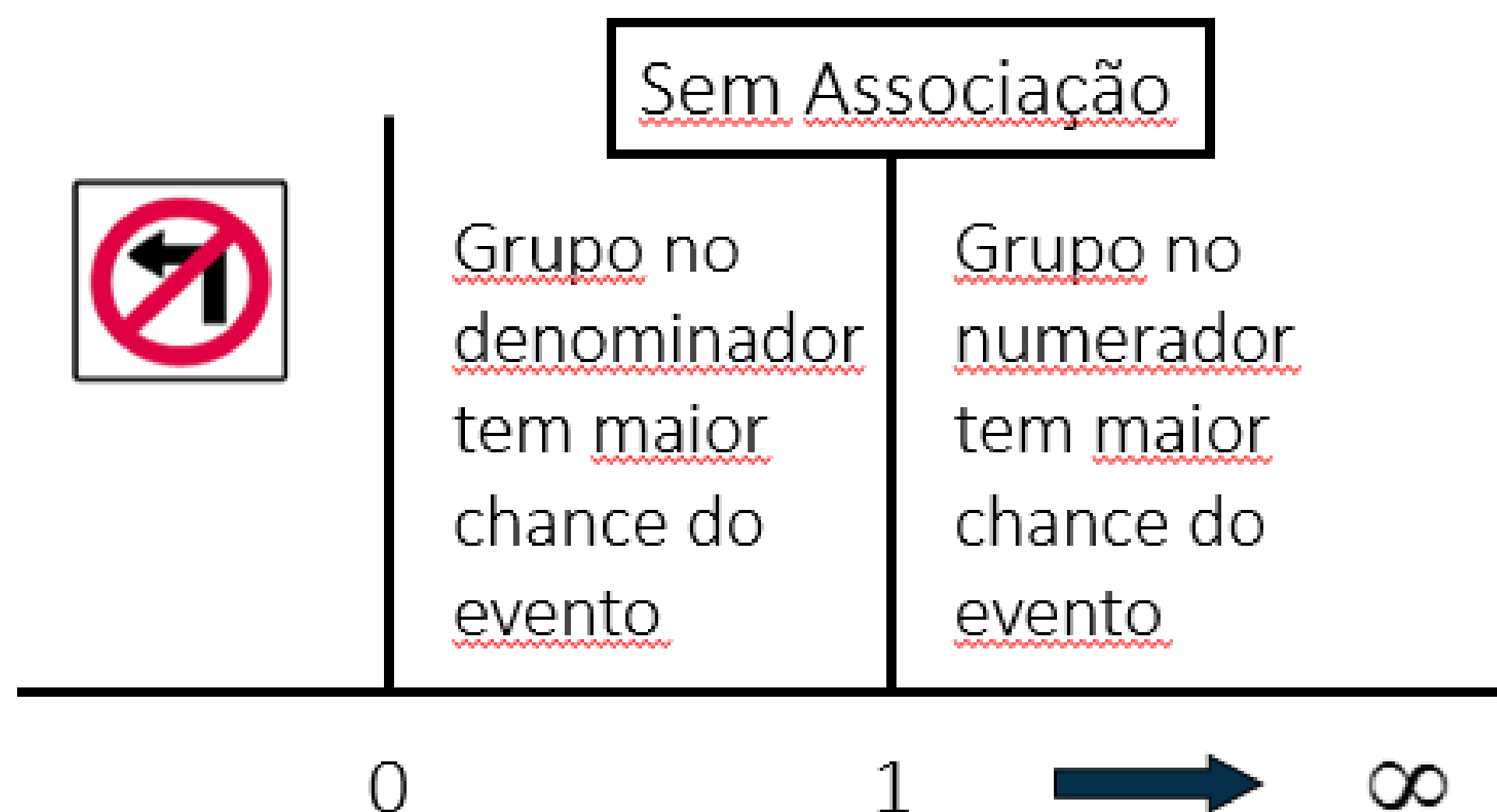
$100(\exp(\beta_2)-1)\%$  muda na odds



# Odds Ratio

- $Odds = \frac{p}{1-p} = e^n$ , chance do evento ocorrer. Em que  $n = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_k x_{ki}$ .
- $Odds_{Ratio} = \frac{Odds_{grupo_A}}{Odds_{grupo_B}} = \frac{e^n_{grupo_A}}{e^n_{grupo_B}}$ , chance do evento ocorrer se for do Grupo\_A com relação ao Grupo\_B

**Odds Ratio  $\in (0, \infty)$**



- **Odds Ratio = 1**  $\rightarrow \frac{Odds_{Grupo1}}{Odds_{Grupo2}} = 1 \rightarrow p_1 = p_2$ , ou seja, não há associação entre a variável preditora e a resposta
- **Odds Ratio > 1**  $\rightarrow \frac{Odds_{Grupo1}}{Odds_{Grupo2}} > 1 \rightarrow Odds_{Grupo1} > Odds_{Grupo2}$ , ou seja, o grupo no numerador tem maior chance do evento ocorrer que o grupo no denominador
- **Odds Ratio < 1**  $\rightarrow \frac{Odds_{Grupo1}}{Odds_{Grupo2}} < 1 \rightarrow Odds_{Grupo1} < Odds_{Grupo2}$ , ou seja, o grupo no numerador tem menor chance do evento ocorrer que o grupo no denominador



# Odds Ratio em um Modelo de Regressão Logística

- Considere o seguinte modelo de regressão logística estimado

$$\text{logit}(p) = -.7567 + .4373*(\text{sexo})$$

em que feminino é codificado com 1 e masculino com 0

- Razão de chances estimada (Femino para Masculino) é:

$$\text{odds ratio} = \frac{\text{odds feminino}}{\text{odds masculino}} = \frac{e^{n1}}{e^{no}} = \frac{e^{-0.7567+0.4373*(1)}}{e^{-0.7567+0.4373*(0)}} = \frac{e^{-0.7567+0.4373*(1)}}{e^{-0.7567}} = e^{0.4373} = 1.55$$

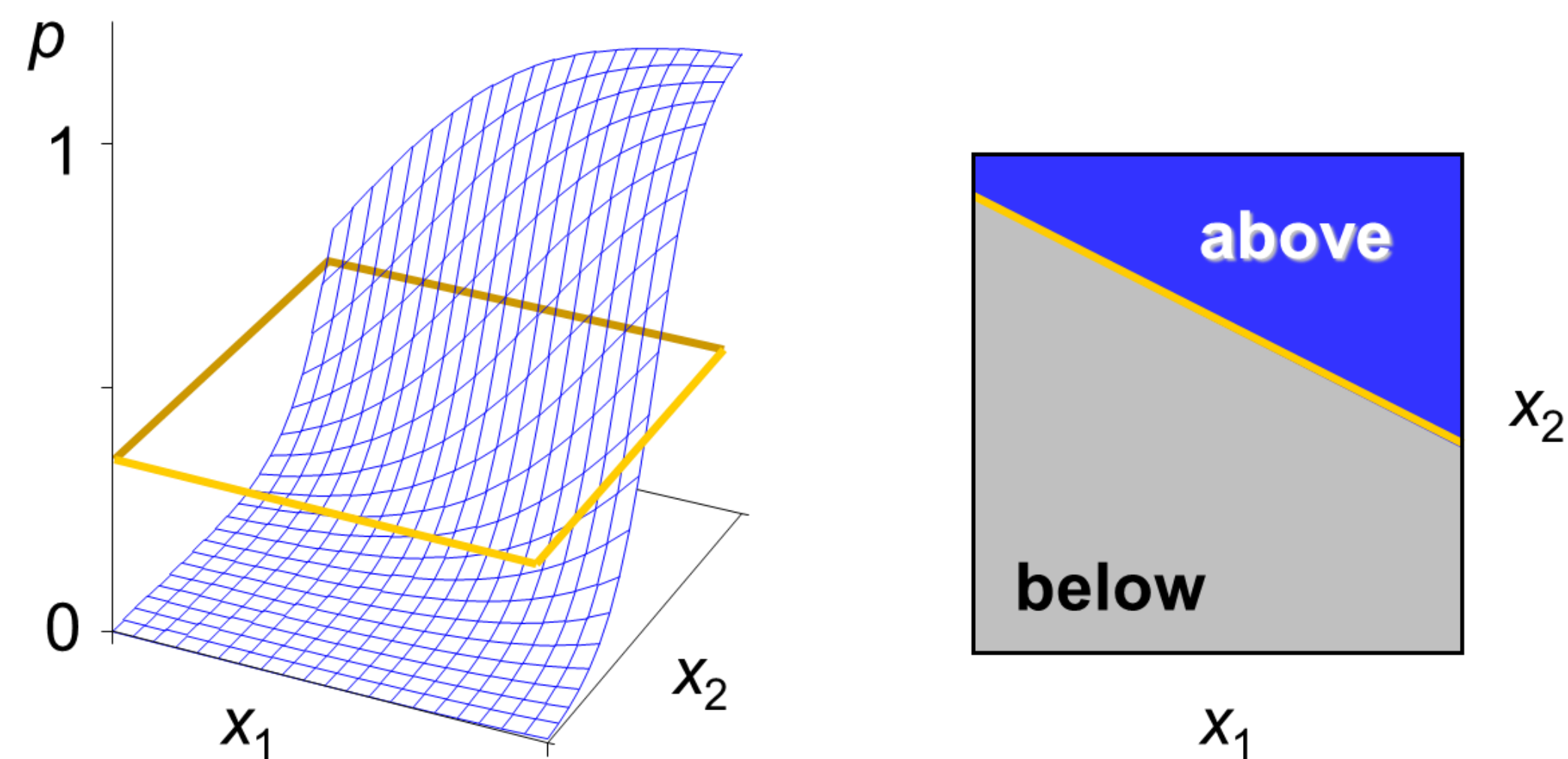
- Interpretação: A chance de ocorrer o sexo feminino é 1,55 vezes a chance de ocorrer o sexo masculino





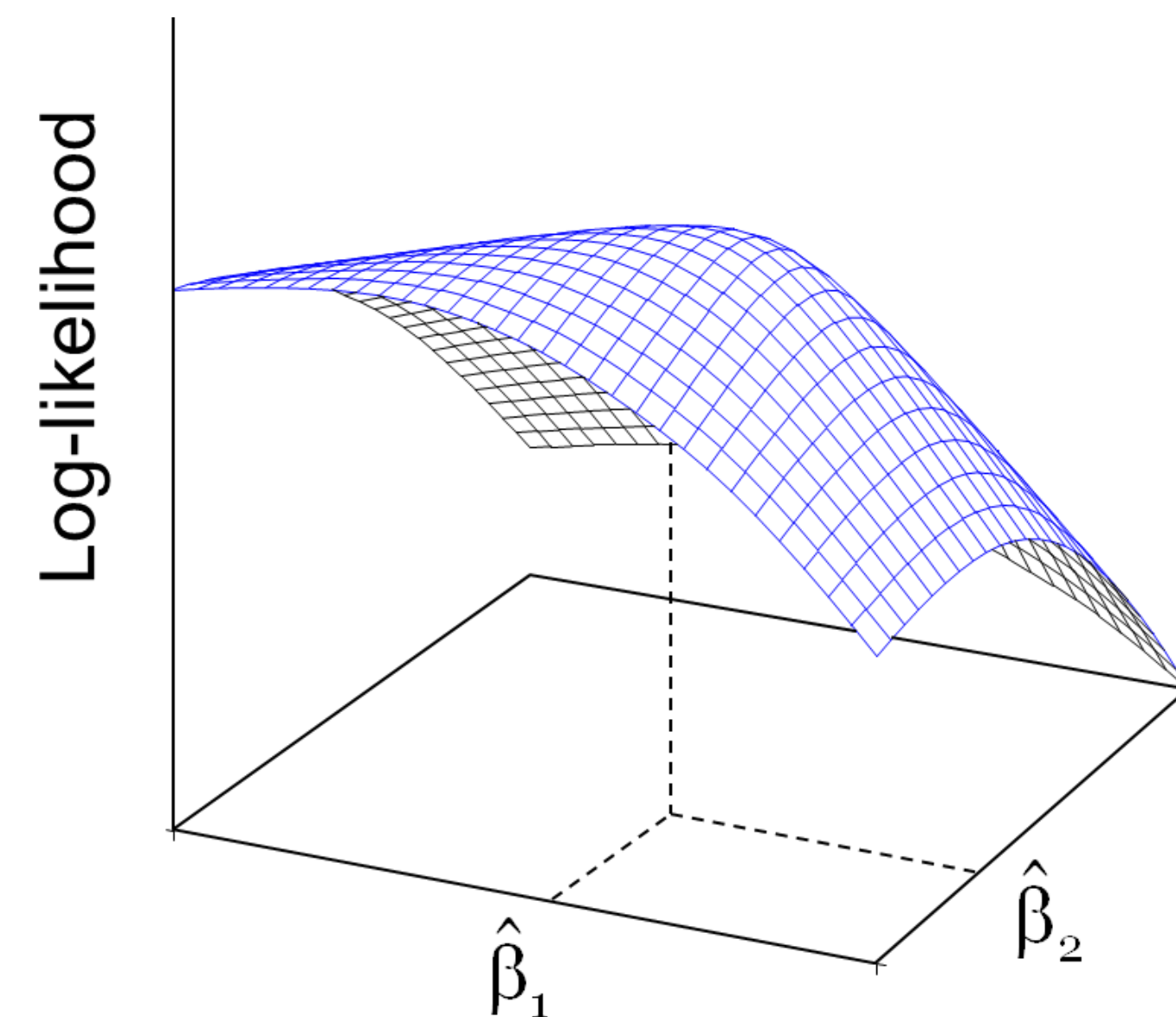
# Discriminação e Estimação

- Discriminação



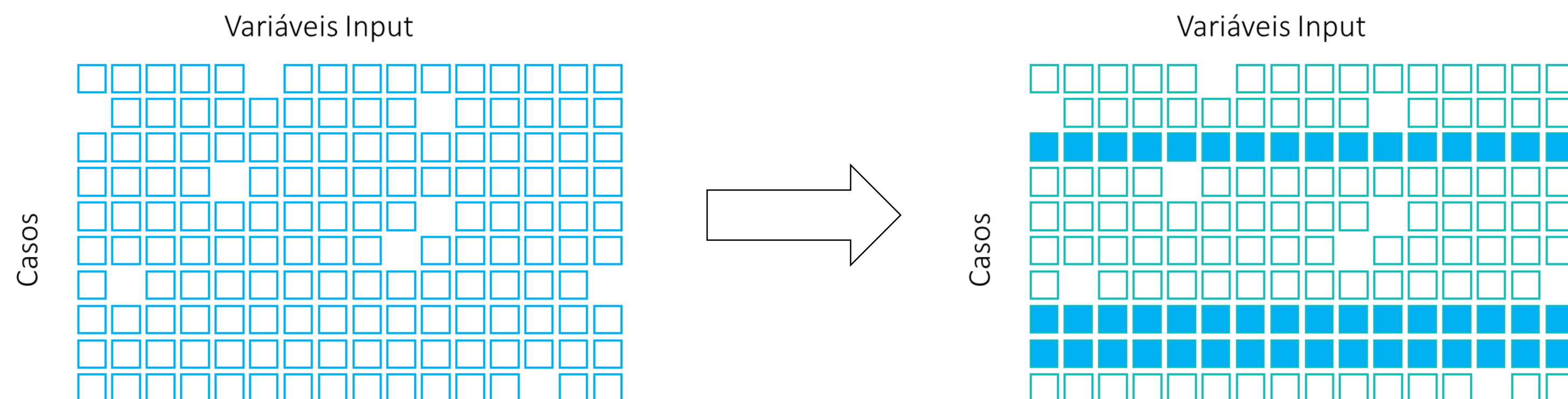
Ponto de Corte

- Estimação



# Tratamento das variáveis - Missing

## 1. Complete case analysis



## 2. Imputação + Variáveis indicadoras de missing

Dado Incompleto	Dado Completo	Indicadora de Missing
34	34	0
63	63	0
.	30	1
22	22	0
26	26	0
54	54	0
18	18	0
.	30	1
47	49	0
20	20	0

Mediana= 30

### Métodos para substituir o missing na imputação:

- Média, Mediana, Moda, Zero, Criação de nova categoria, regressão,...

### Criando a indicadora de missing:

- Indicadora de Missing =  $\begin{cases} 1, & \text{se a obs é missing} \\ 0, & \text{caso contrario} \end{cases}$



# Tratamento das variáveis - Categóricas

Problemas causados por variável input categórica

1. Variáveis com muitos níveis: se expandir em dummies
  - aumenta a dimensão
  - produzirá inputs redundantes e irrelevantes

2. Quase completa separação: Quando um nível da categoria tem taxa de evento target igual a 0 ou a 100% das observações
  - afeta a convergência dos parâmetros
  - Pode levar a escolha errada das variáveis

Criando dummies da variável classe

	classe		D_A	D_B	D_C
1	B	1	0	1	0
2	A	2	1	0	0
3	B	3	0	1	0
4	C	4	0	0	1
5	A	5	1	0	0
6	A	6	1	0	0
7	C	7	0	0	1

	0	1	D <sub>A</sub>	D <sub>B</sub>	D <sub>C</sub>	logit
A	28	7	1	0	0	-1.39
B	16	0	0	1	0	-Infinity
C	94	11	0	0	1	-2.14
D	23	21	0	0	0	-0.08





# Tratamento das variáveis - Categóricas

Soluções para os problemas causados por variável input categórica

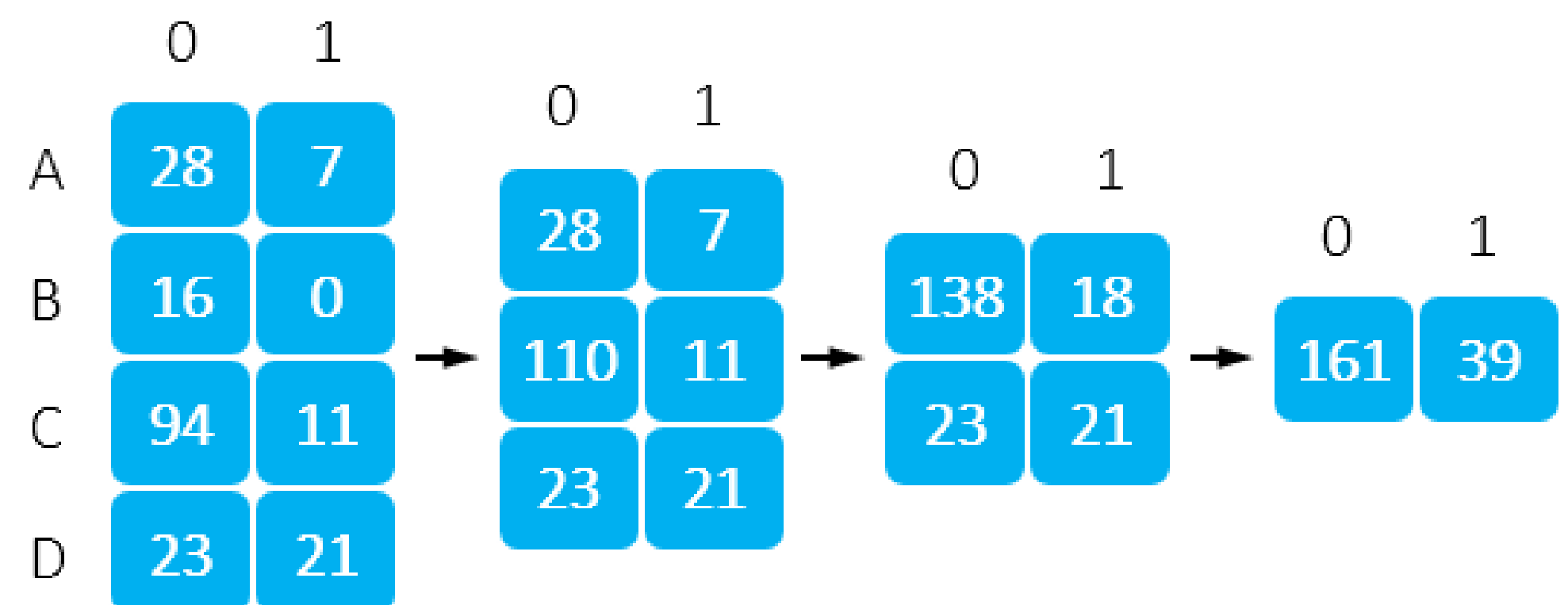
1. Thresholding: Juntar categorias baseado no número de observações

Categoria	Número de observações
A	1562
B	970
C	223
D	111
E	85
F	23
G	17
H	12
I	5

Crie uma nova categoria "Outros" e então crie dummies para cada uma das seis categorias

2. Clusterização: Juntar as categorias das variáveis considerando

- menor redução da estatística de  $\chi^2$
- Taxas de respostas semelhante
- número de observações na categoria



Merged:	B and C	A and BC	ABC and D	
$\chi^2 =$	31.7	30.7	28.6	0
	100%	97%	90%	0%

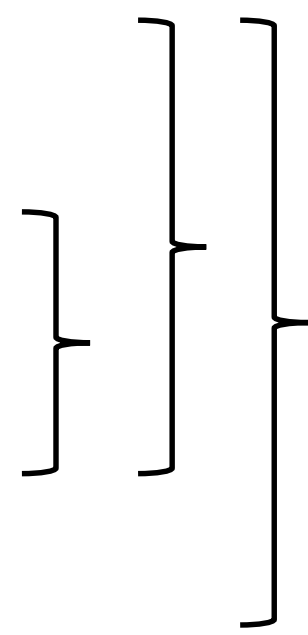


# Tratamento das variáveis - Categóricas

Soluções para os problemas causados por variável input categórica

3. Weight of Evidence (WOE): substitui o valor da categoria pelo  $\log(\text{odds})$  do evento

	0	1	p	WOE
A	28	7	0,200	0,25
B	16	0	0	0
C	94	11	0,105	0,117
D	23	21	0,477	0,912



# Estudo de Caso

## Ajustando um modelo de Regressão Logística no Python

**Fonte da dados:** `kaggle`

**Link:** <https://www.kaggle.com/kost13/us-income-logistic-regression/data>

**Resumo:** Dados do Censo Adulto Americano referentes a renda para fatores sociais como Idade, Educação, raça, etc.

**Objetivo:** Ajustar um modelo de regressão logística, em uma base de treinamento, para uma resposta binária, fazer a previsão desta resposta e avaliar a qualidade de ajuste do modelo em uma base de teste.



# Estudo de Caso

## Ajustando um modelo de Regressão Logística no Python

### Parte 1 : Tratando as Variáveis do modelo

- Missing
- Variáveis categóricas



# Tratamento das variáveis - Redundância

Redundância: Variáveis input altamente correlacionadas

## 1. Problemas das variáveis redundantes:

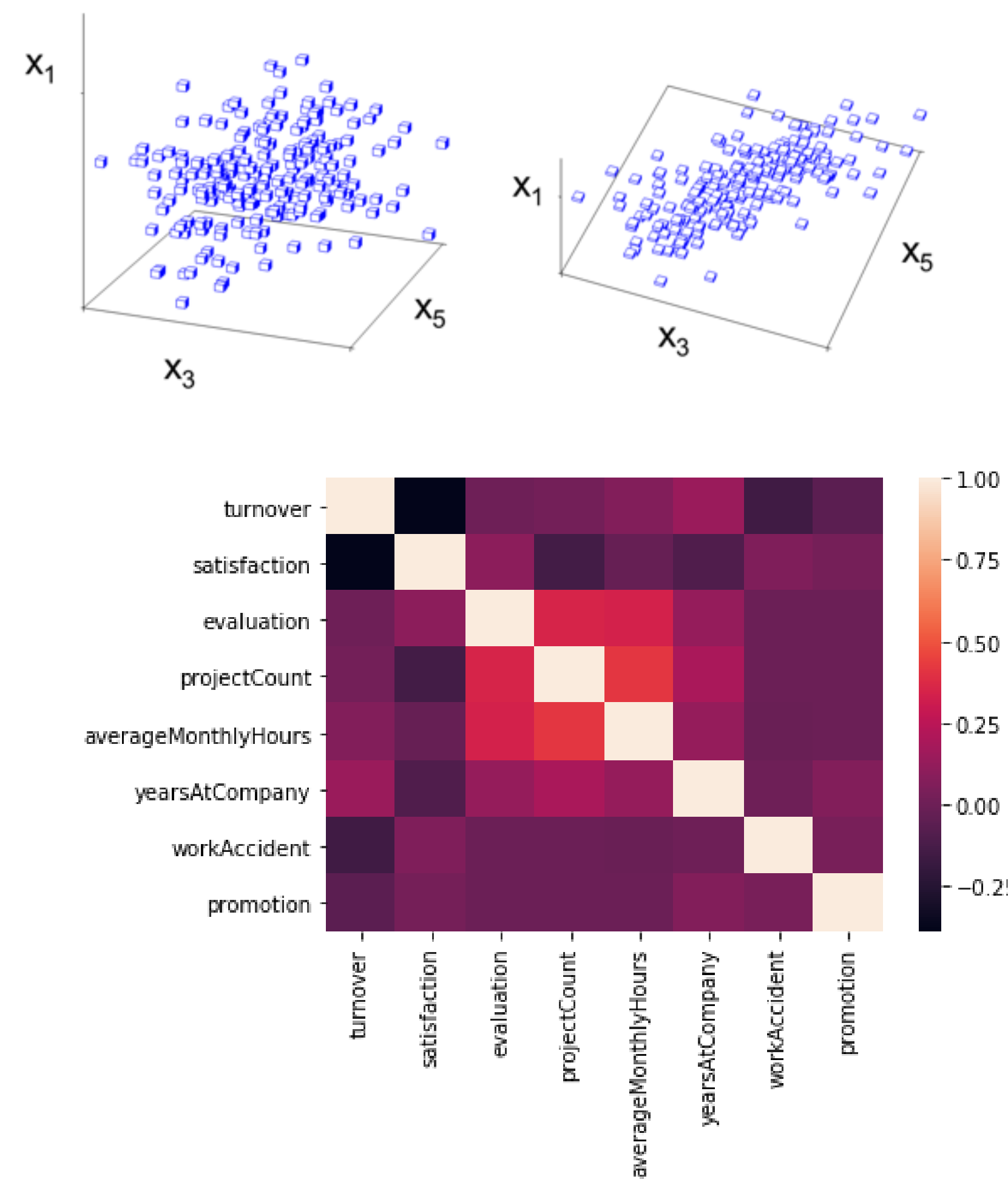
- desestabiliza a estimação dos parâmetros
- aumenta o risco de overfitting
- pode confundir a interpretação
- aumenta o tempo computacional para a estimação dos parâmetros
- aumenta o custo da coleção dos dados

## 2. Como verificar se as variáveis são redundantes:

- Matriz de correlação entre as variáveis input

## 3. Como resolver o problema da redundância

- Excluir da análise as variáveis que são altamente correlacionadas entre si e destas a que tem menor correlação com a variável resposta



# Tratamento das variáveis - Irrelevância

Irrelevância: Variáveis inputs pouco correlacionadas com a variável resposta

1. Problemas das variáveis irrelevantes:
  - ao utilizar algum método de seleção de variáveis pode-se selecionar a variável incorreta
2. Como verificar se as variáveis são irrelevantes:
  - Matriz de correlação das variáveis input com a variável target
3. Como resolver o problema da redundância
  - Excluir da análise as variáveis que tem baixa correlação com a variável resposta, mas antes verificar se a interação entre as variáveis com baixa correlação aumenta o poder de predição do modelo.





## Estudo de Caso

# Ajustando um modelo de Regressão Logística no Python

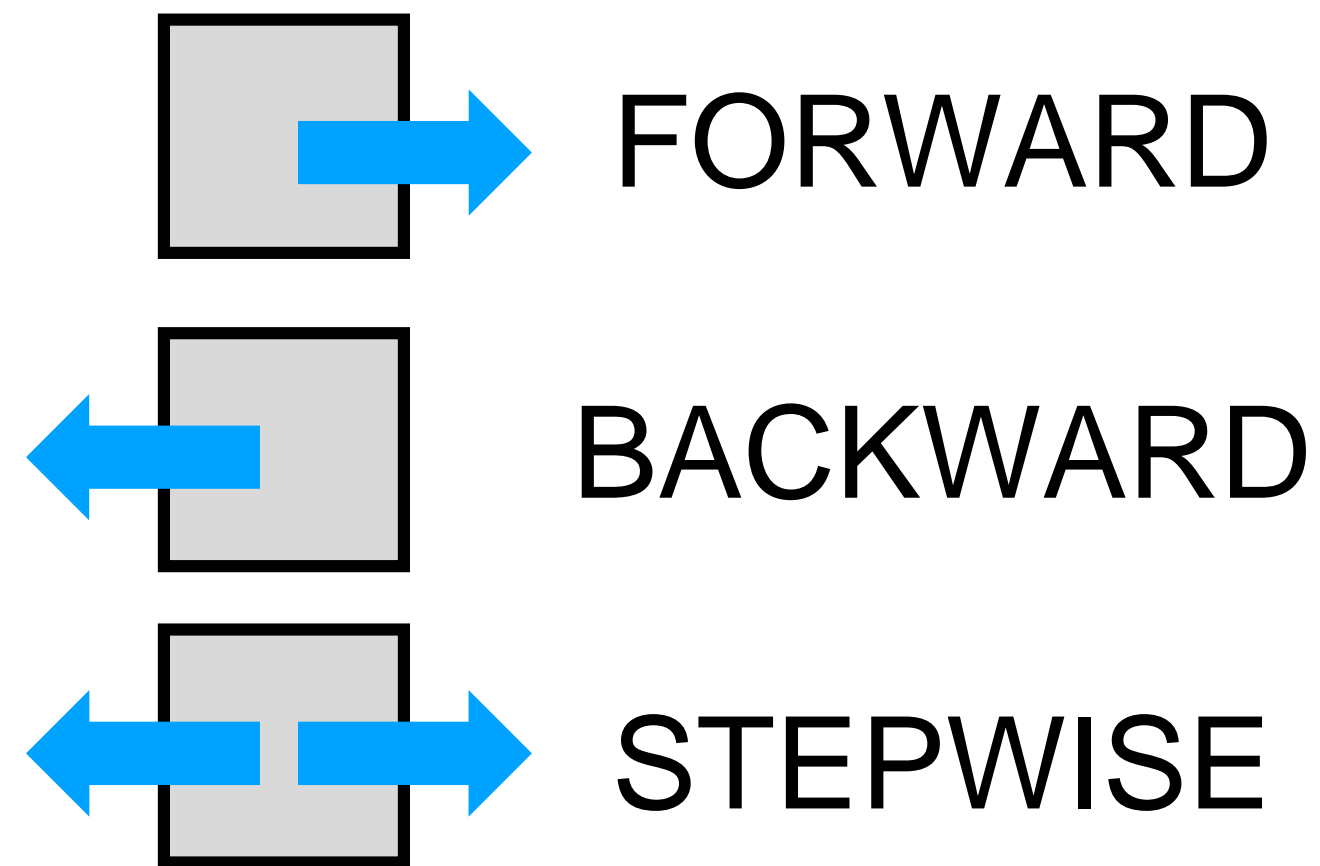
### Parte 2 : Correlação

- Entre as variáveis input
- Entre cada input e a target



# Seleção de Variáveis

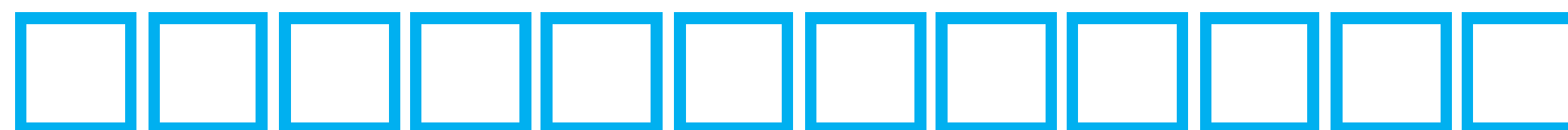
- Para diminuir a dimensão com conjunto de dados e assim facilitar a análise, podemos utilizar métodos de seleção de variáveis que testam todos os possíveis modelos e retornam o que melhor ficou ajustado.
  - Dependendo do número de variáveis estes métodos se tornam muito caros computacionalmente
- Métodos sequenciais



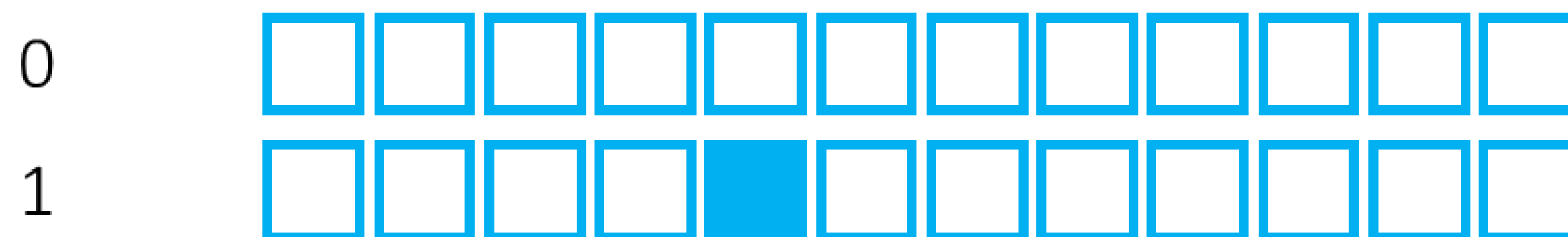


# Seleção de Variáveis - Forward

0



# Seleção de Variáveis - Forward

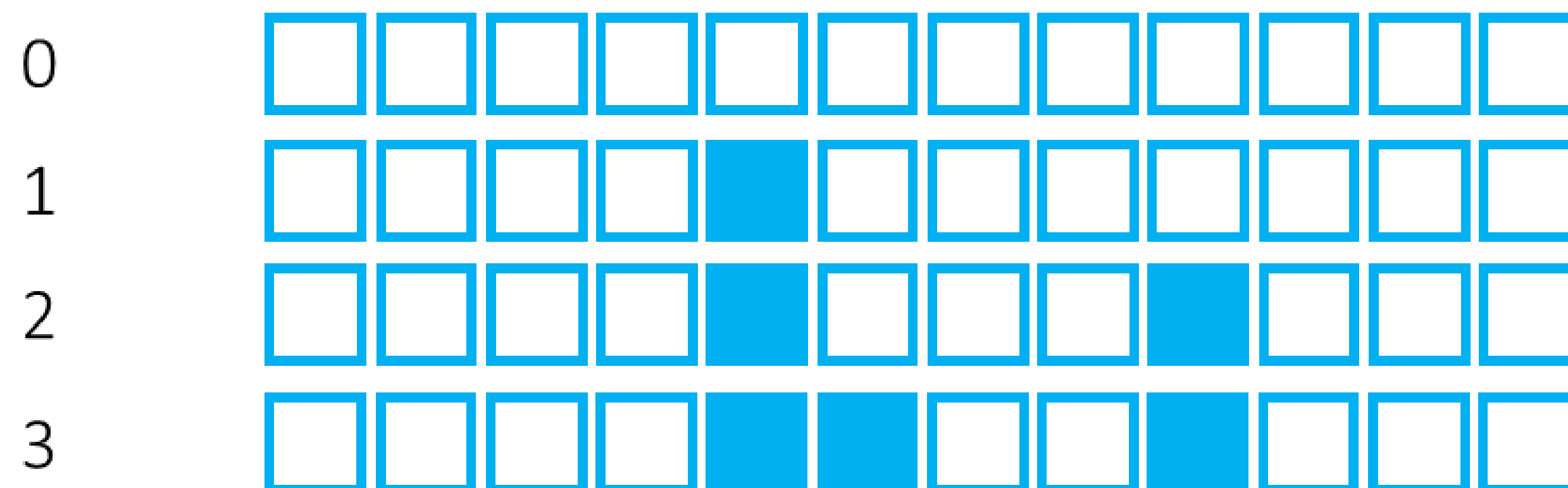


# Seleção de Variáveis - Forward

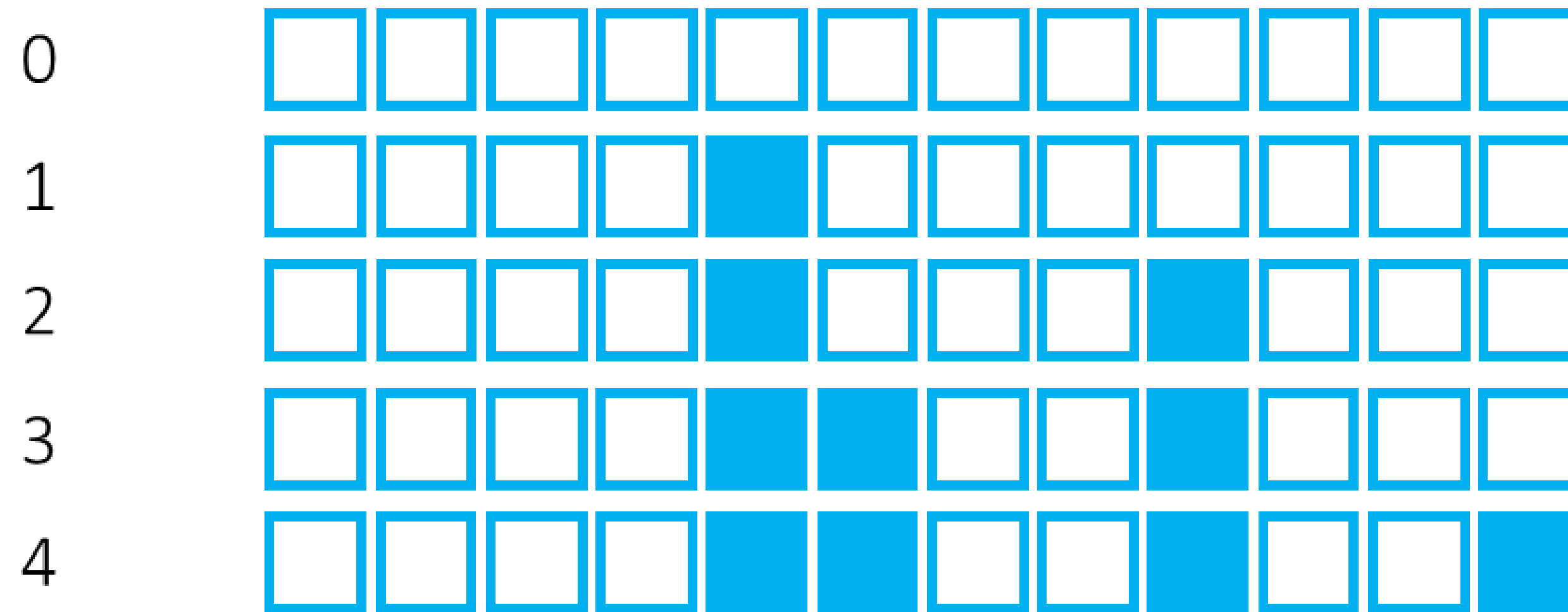
0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



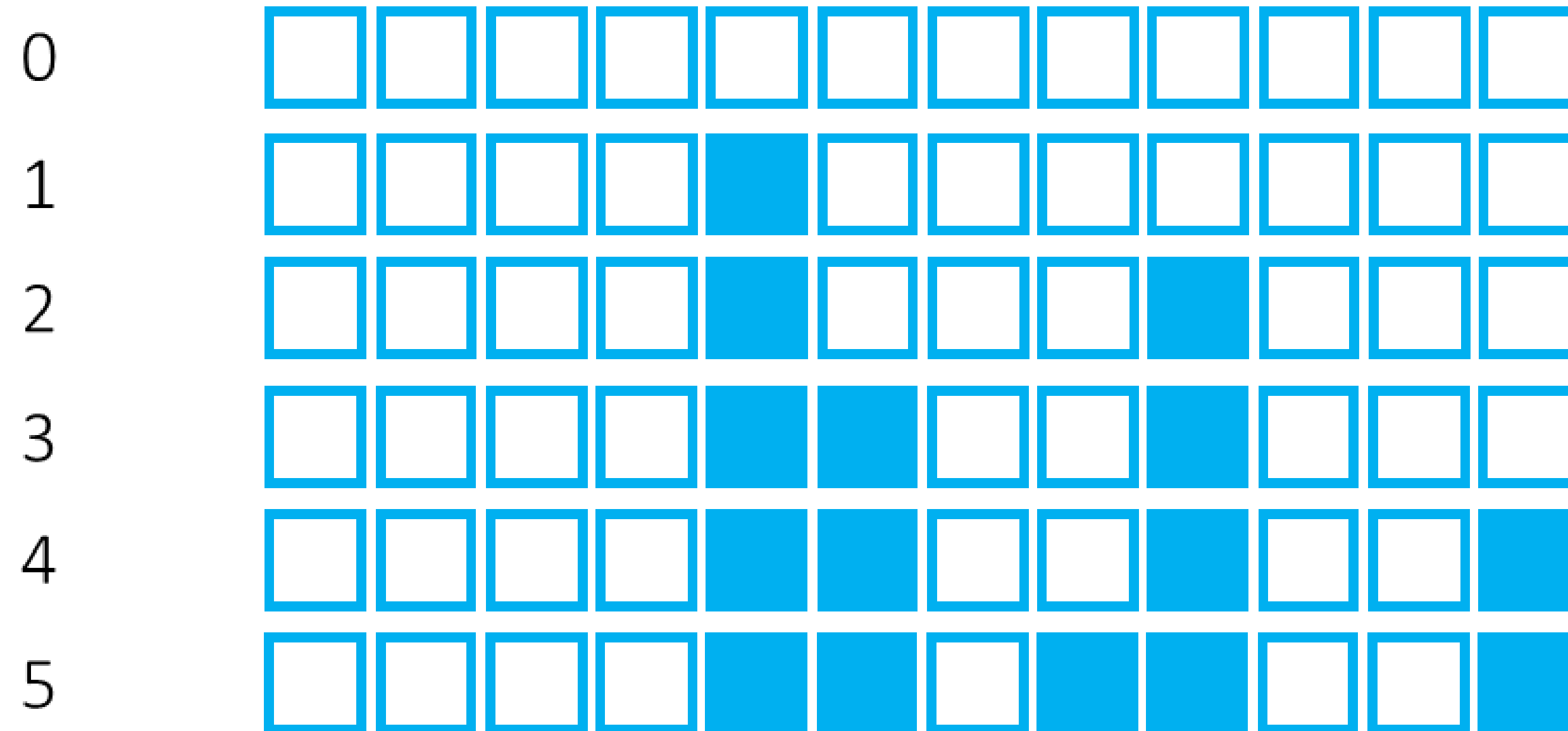
# Seleção de Variáveis - Forward



# Seleção de Variáveis - Forward



# Seleção de Variáveis - Forward



# Seleção de Variáveis - Forward

0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Stop	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



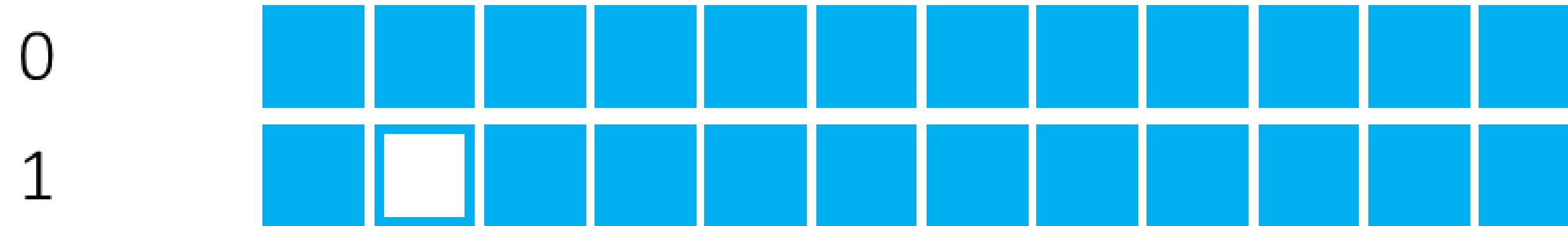
# Seleção de Variáveis - Backward

0

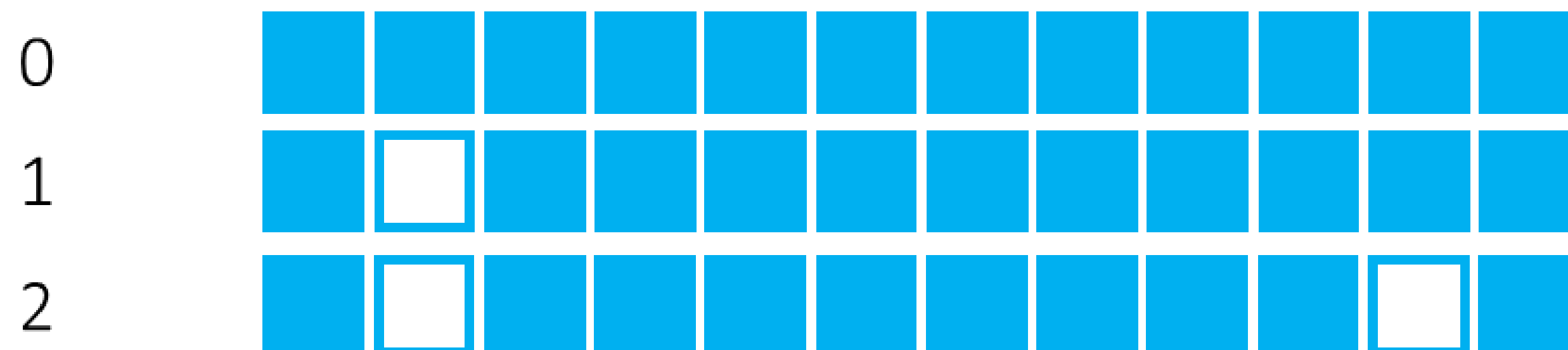




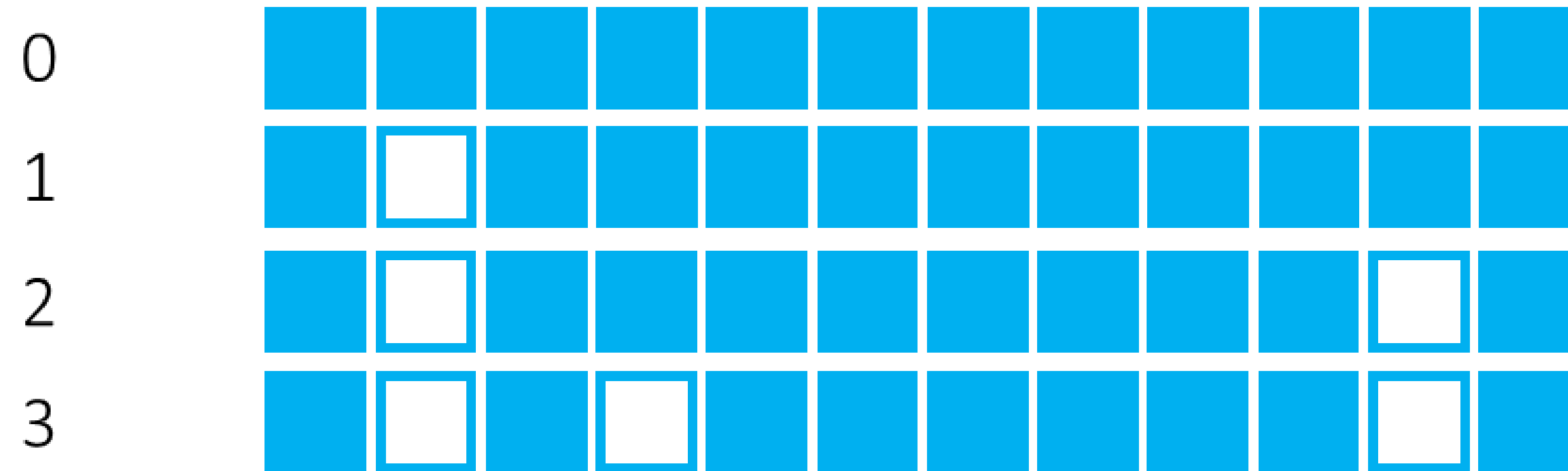
# Seleção de Variáveis - Backward



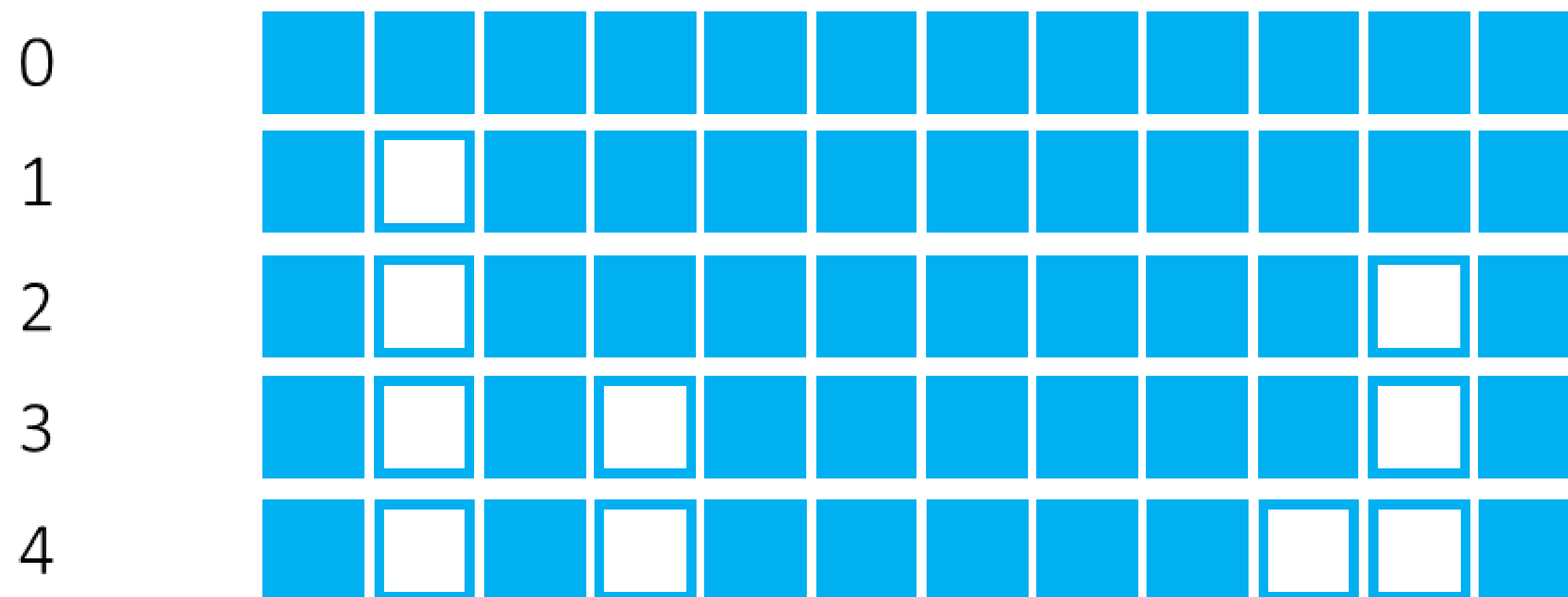
# Seleção de Variáveis - Backward



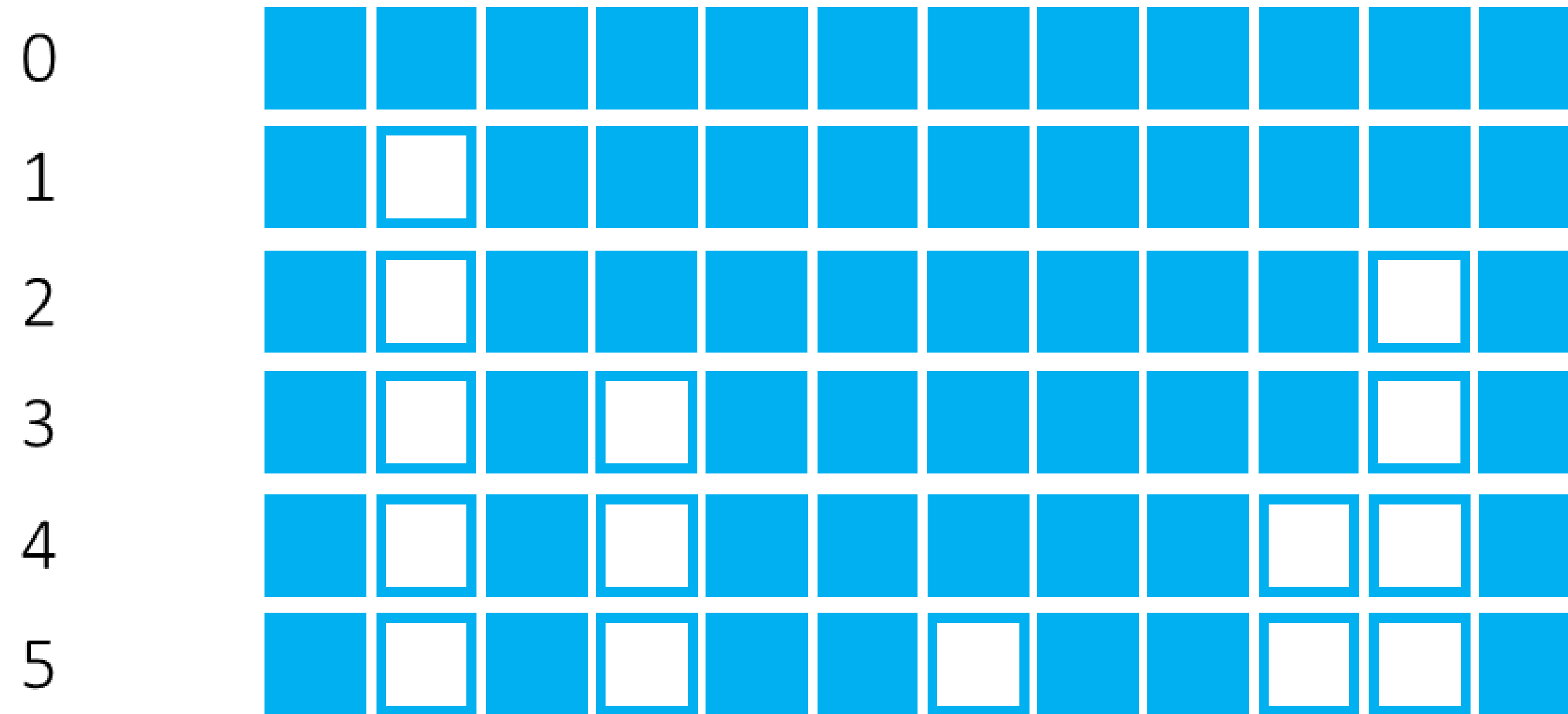
# Seleção de Variáveis - Backward



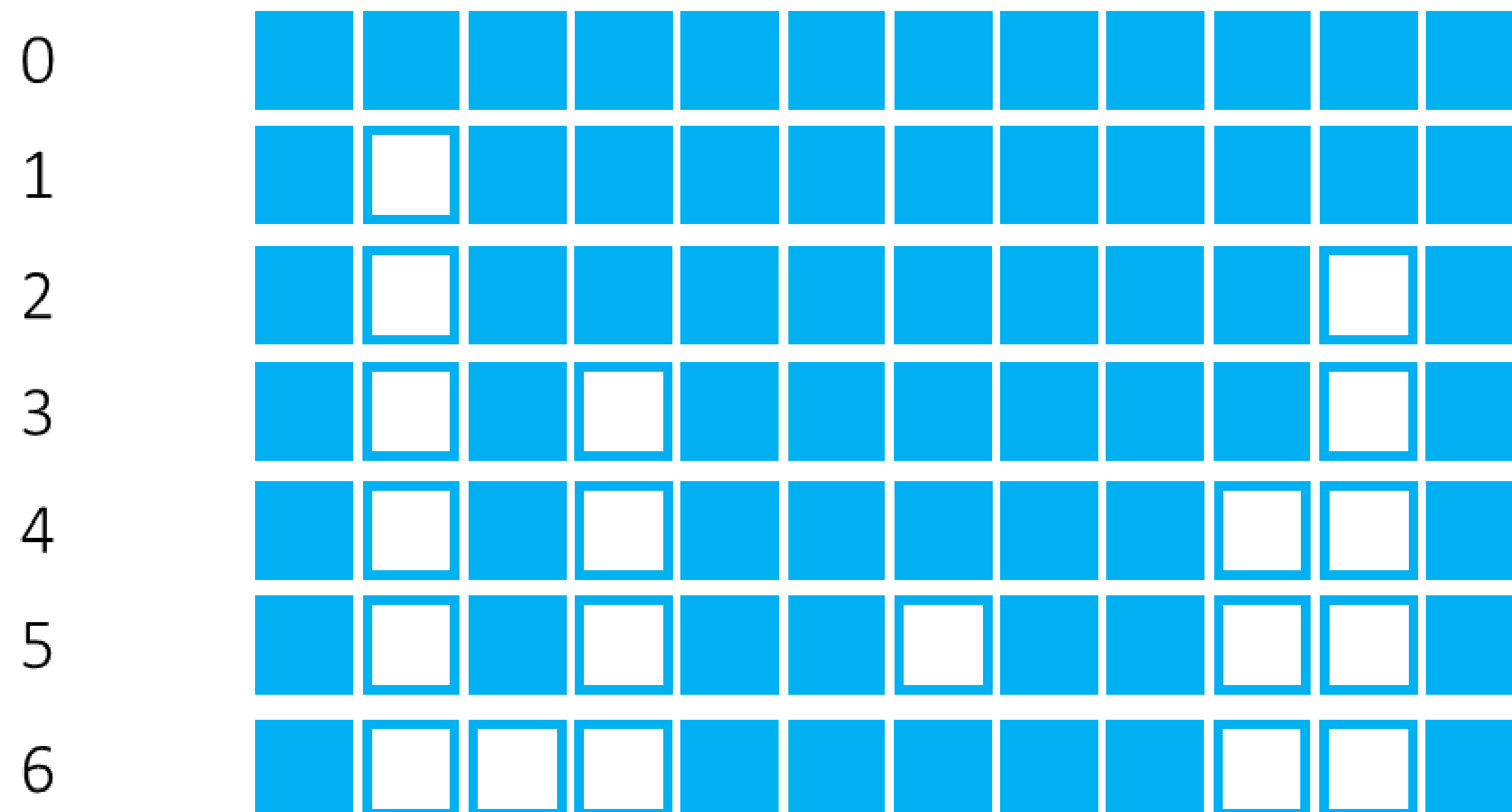
# Seleção de Variáveis - Backward



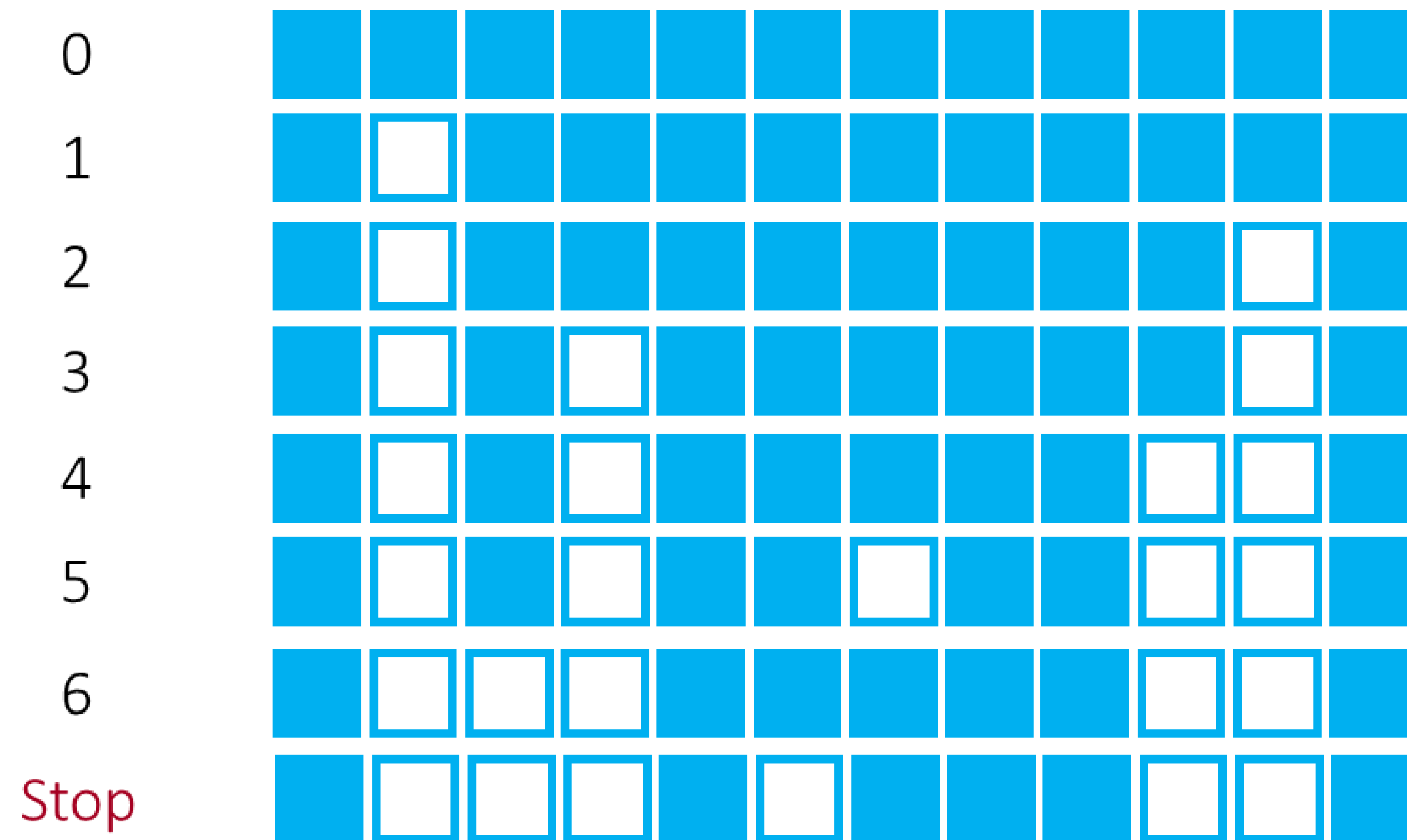
# Seleção de Variáveis - Backward



# Seleção de Variáveis - Backward

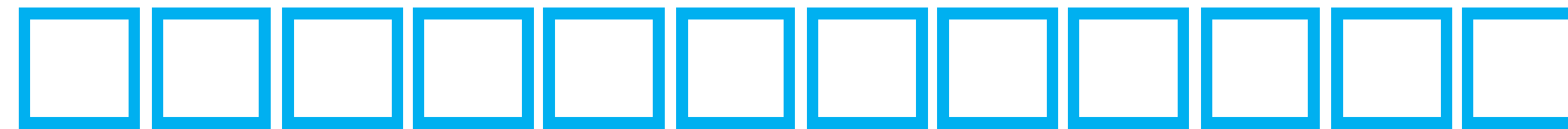


# Seleção de Variáveis - Backward



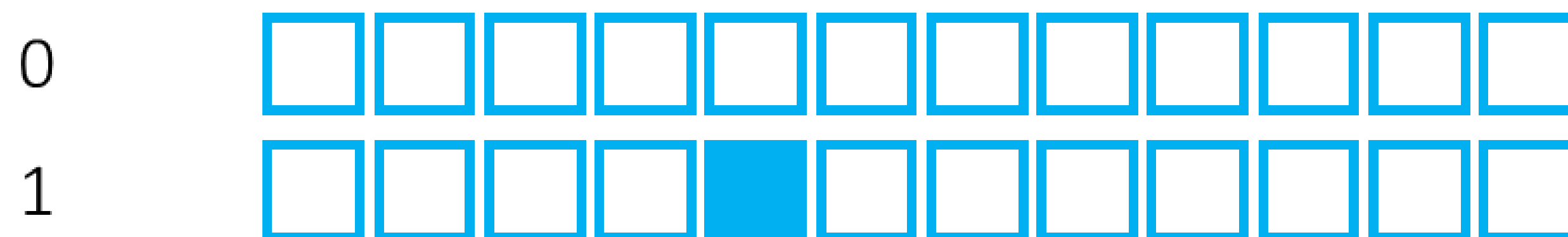
# Seleção de Variáveis - Stepwise

0





# Seleção de Variáveis - Stepwise

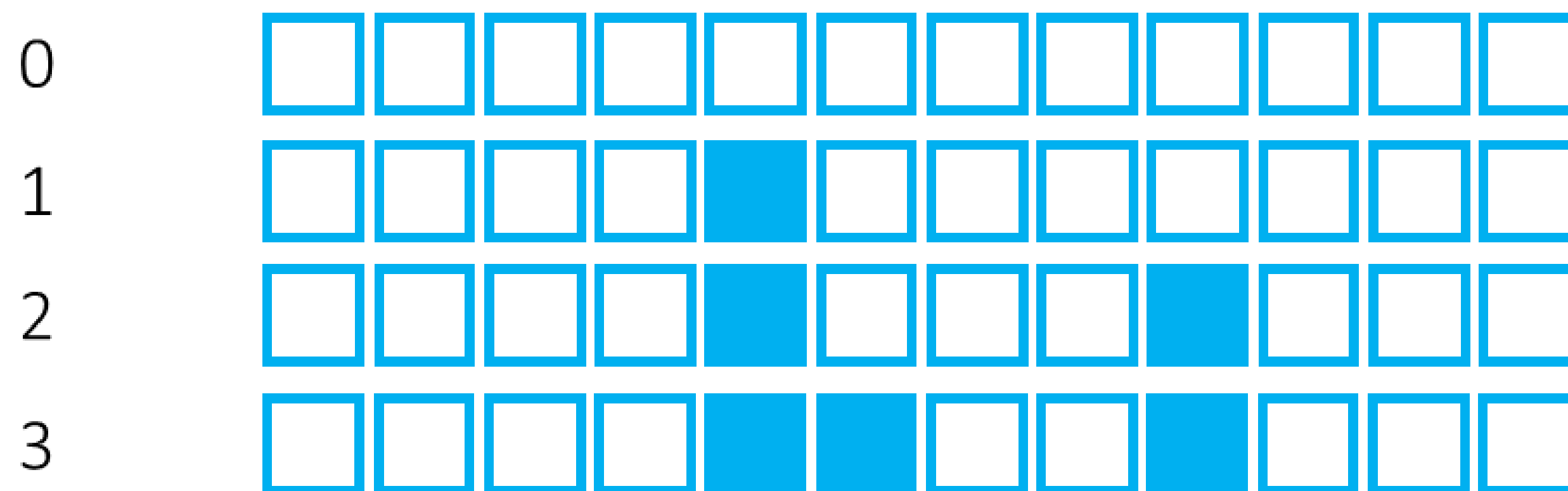


# Seleção de Variáveis - Stepwise

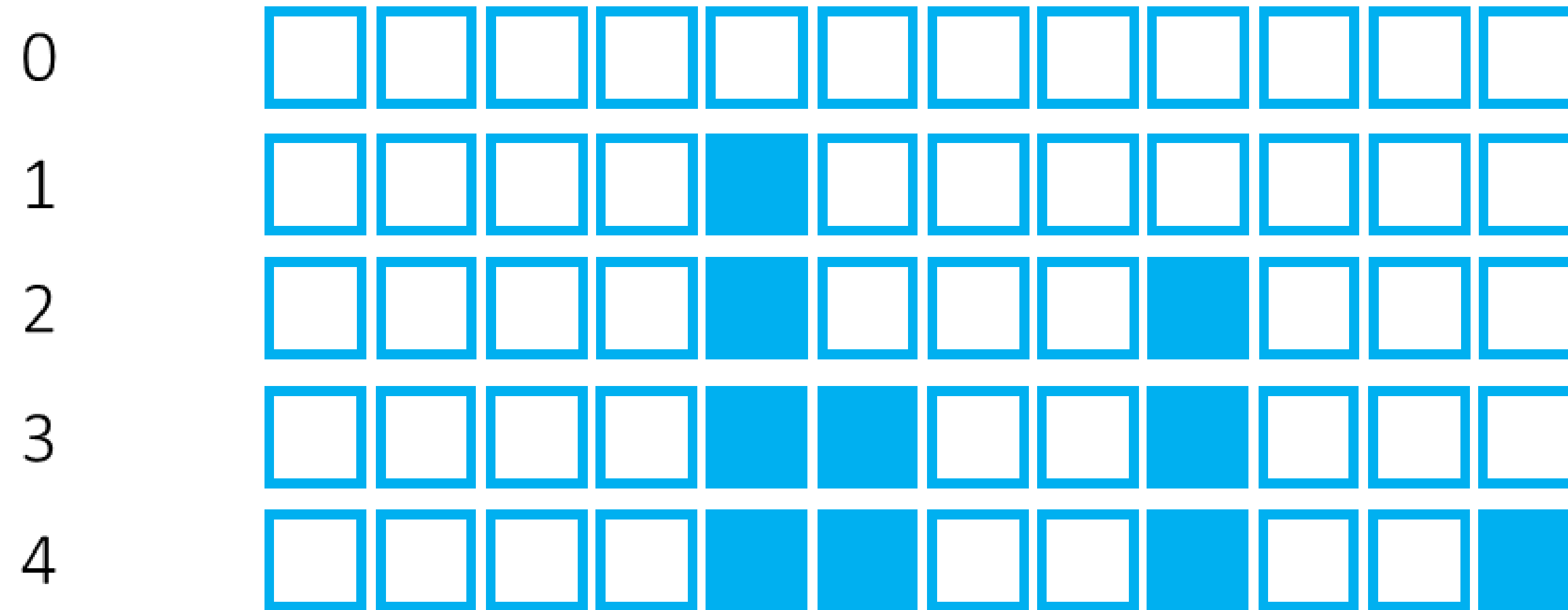
0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



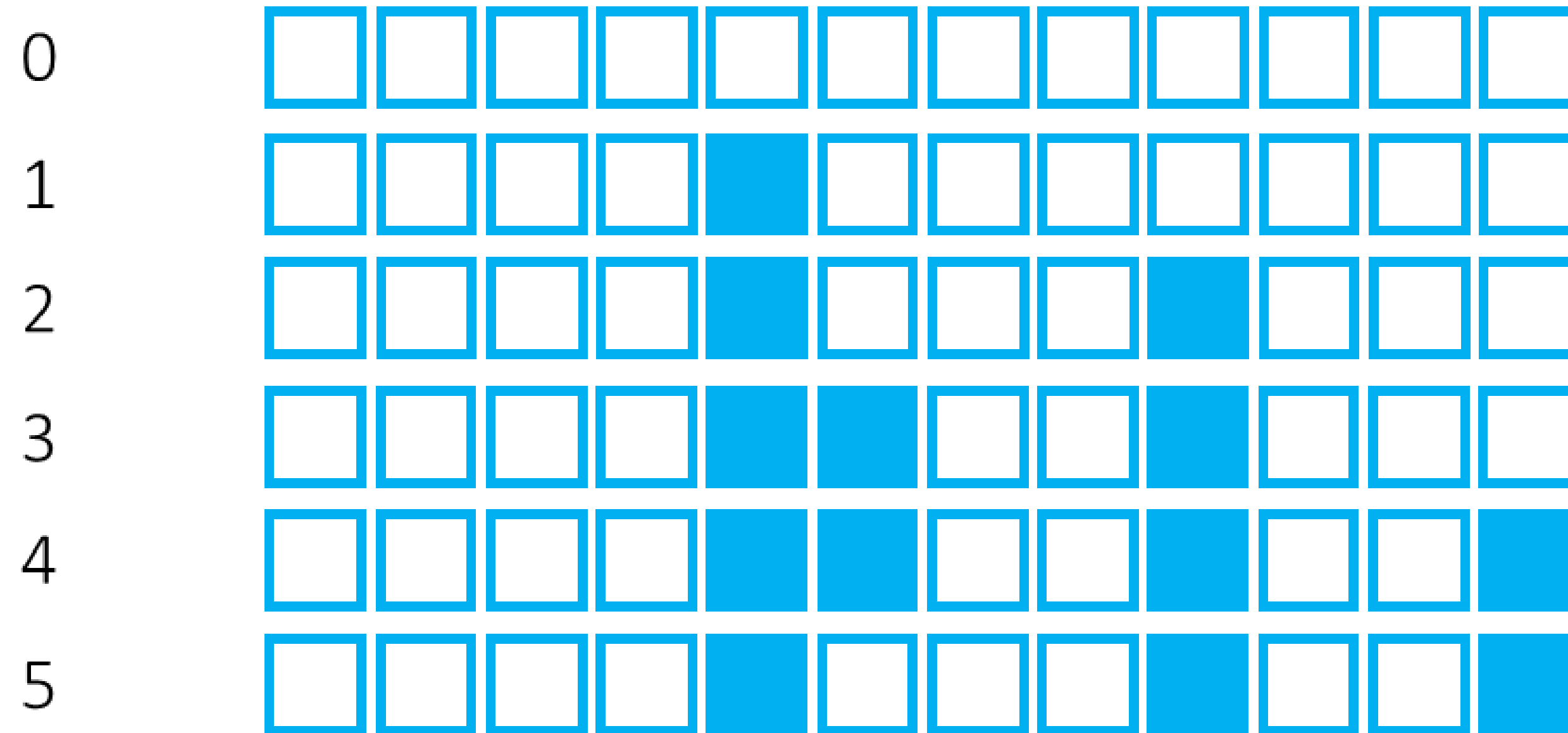
# Seleção de Variáveis - Stepwise



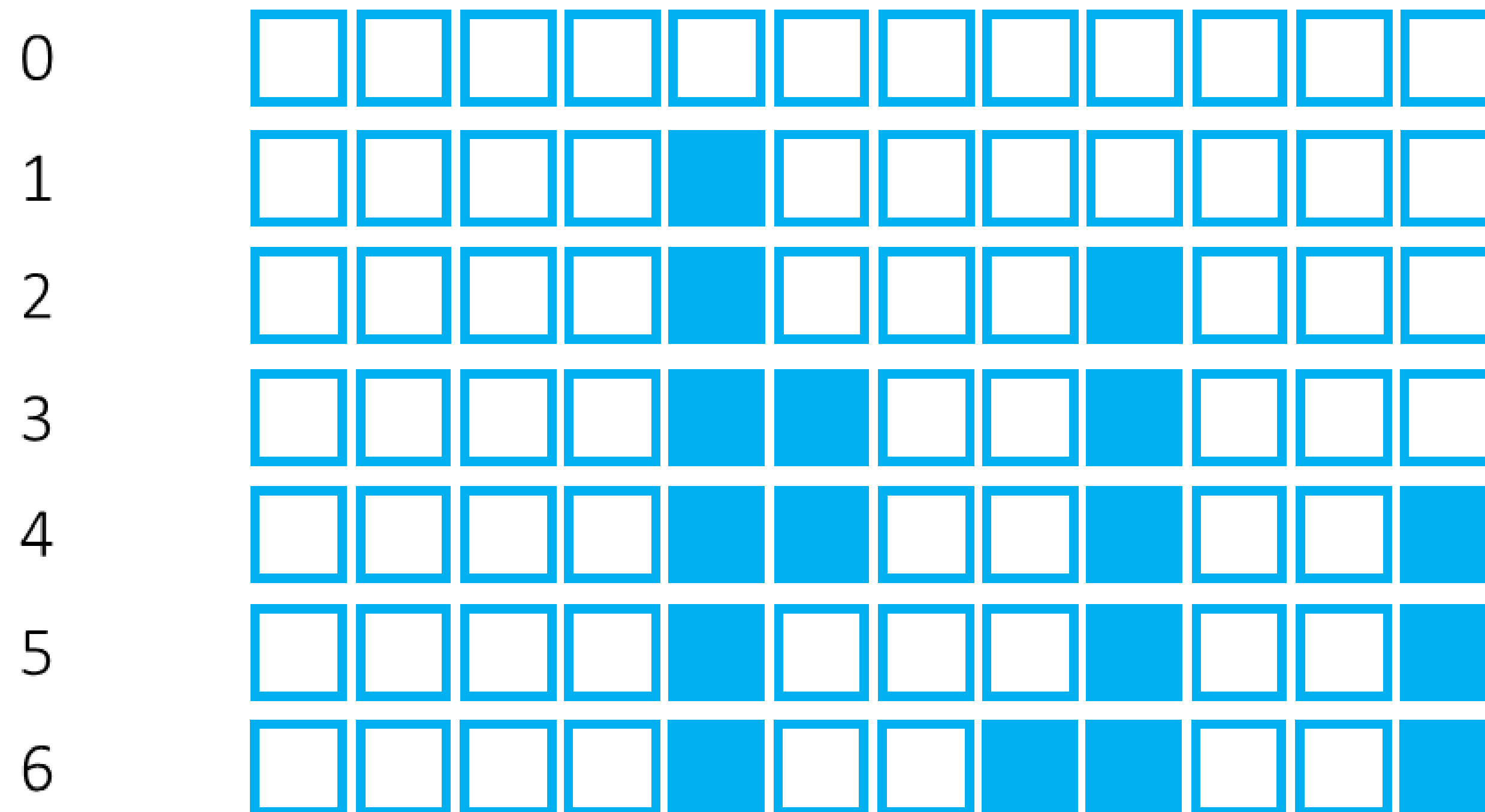
# Seleção de Variáveis - Stepwise



# Seleção de Variáveis - Stepwise



# Seleção de Variáveis - Stepwise



# Seleção de Variáveis - Stepwise

0	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
1	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
2	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
3	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
4	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
5	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
6	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>
Stop	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>	<input checked="" type="checkbox"/>	<input type="checkbox"/>	<input checked="" type="checkbox"/>



## Estudo de Caso

# Ajustando um modelo de Regressão Logística no Python

## Parte\_3 : Método de seleção de variáveis – Forward





# Ajuste do Modelo – Matriz de Confusão

		Classificação Predita		
		0	1	
Classificação Real	0	TN	FP	Actual Negative
	1	FN	TP	Actual Positive
		Predicted Negative	Predicted Positive	

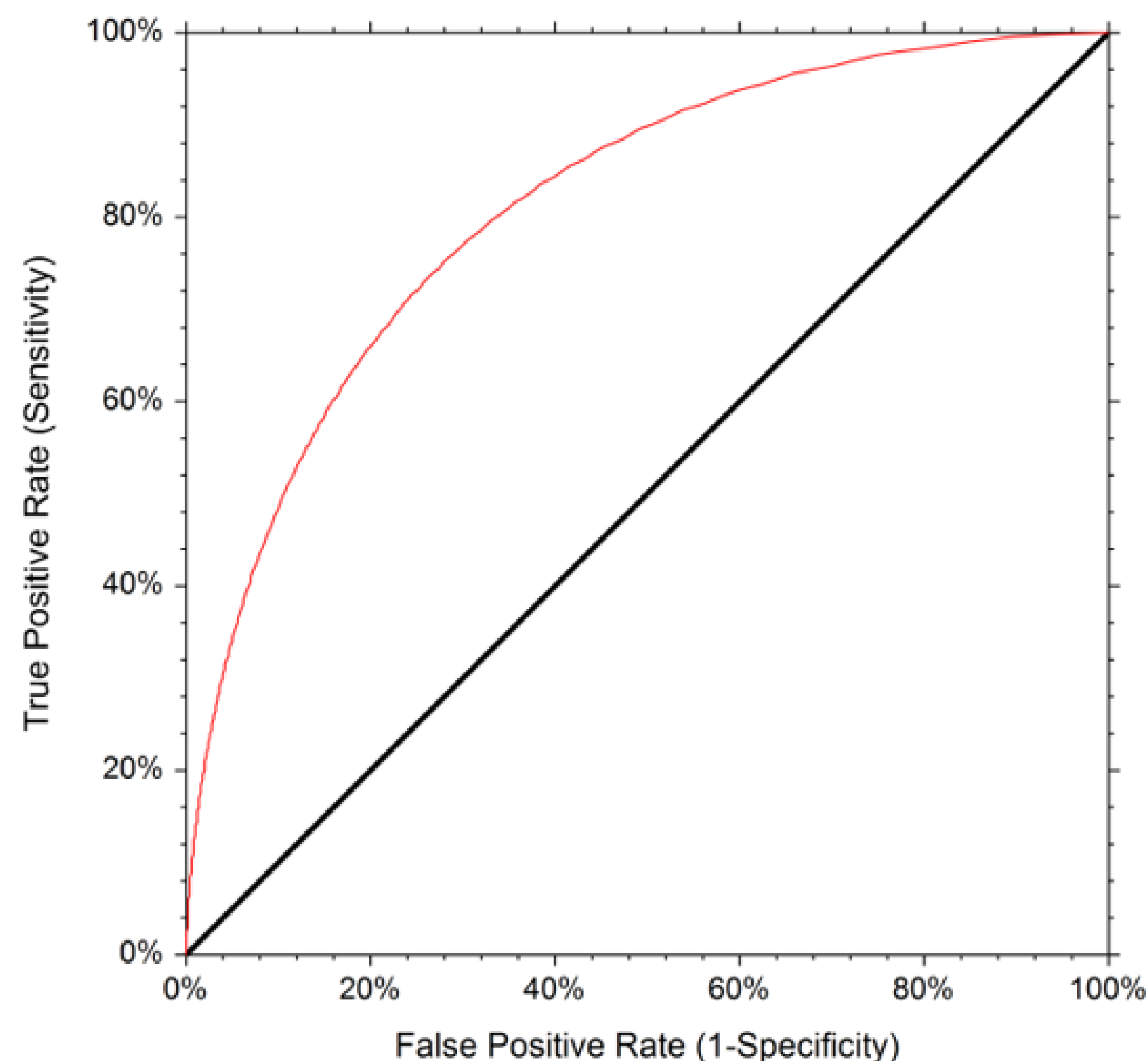
TN: True Negative  
 TP: True Positive  
 FN: False Negative  
 FP: False Positive

- Métricas** para avaliar a qualidade do ajuste do modelo
  - **Missclassification** =  $\frac{FP+FN}{Total\ de\ casos}$
  - **Acurácia** =  $\frac{TP+TN}{Total\ de\ casos}$
  - **Precision** =  $P = \frac{TP}{TP+FP}$ 
    - Altos valores de precision estão relacionados a baixa taxa de FP
  - **Recall** =  $R = \frac{TP}{TP+FN}$ 
    - Altos valores de recall estão relacionados a baixa taxa de FN
- Conclusões:**
  - Alto recall e Baixo precision -> prejudica o cliente, pois o cliente era bom (0) e foi classificado como ruim (1).
  - Baixo recall e Alto precision -> beneficia o cliente, pois o cliente era ruim (1) e foi classificado como bom (0).
  - Altos valores de precision e recall são indicativos de um modelo bem ajustado



# Ajuste do Modelo – Curva ROC

A curva ROC, mede, fração a fração, quantos 1's foram capturados (taxa de true positive) vs quantos 0's foram capturados (taxa de false positive).



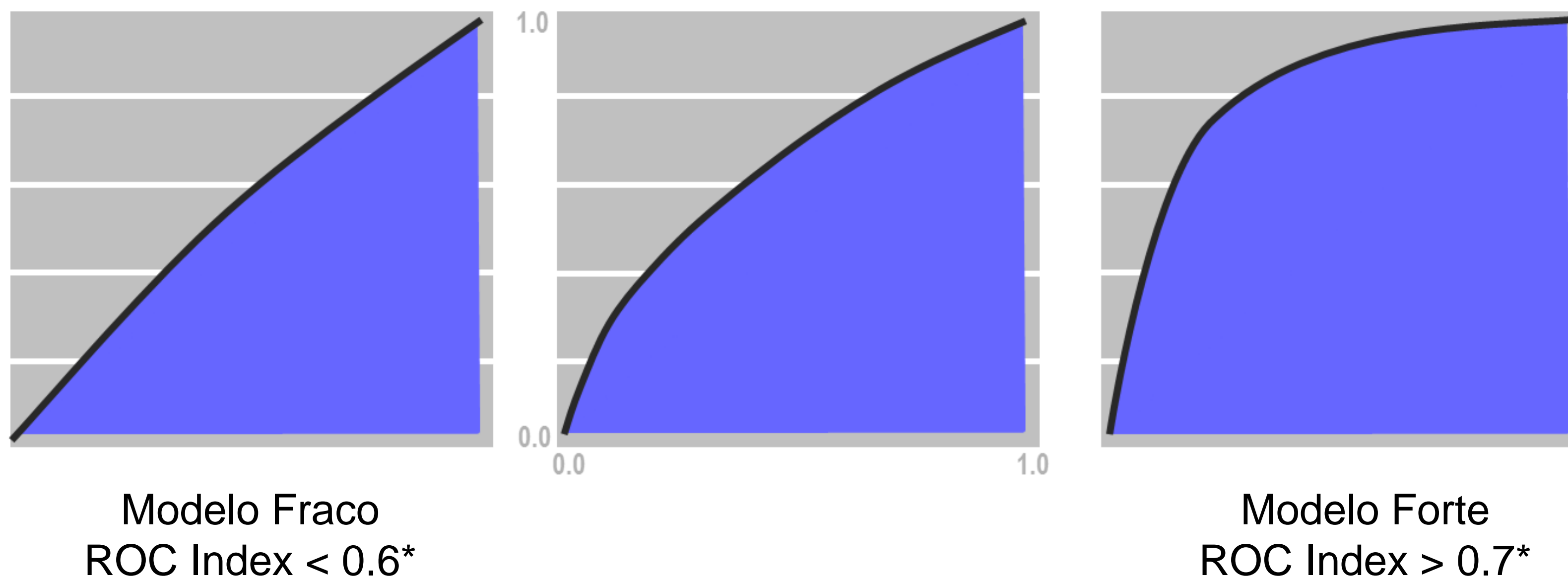
- **Métricas**

- *Sensibilidade = Recall* =  $\frac{TP}{TP+FN}$
- *Especificidade* =  $\frac{TN}{TN+FP}$



# Ajuste do Modelo – Curva ROC

Quanto maior a área sob a curva melhor é o modelo ajustado



\* Regras de bolso sempre são perigosas, o modelo ideal depende sempre do problema modelado.



# Estudo de Caso

## Ajustando um modelo de Regressão Logística no Python

**Parte\_4.1: Ajustar um modelo de regressão Logística na base de treinamento usando sklearn**

**Parte\_4.2: Validar o modelo na base de teste usando: AUC, precision, recall**



# Estudo de Caso

## Ajustando um modelo de Regressão Logística no Python

**Parte\_5.1: Ajustar um modelo de regressão Logística na base de treinamento usando statsmodel**

**Parte\_5.2: Validar o modelo na base de teste usando: AUC, precision, recall**



## Desafio

# Ajustar um modelo de Regressão Logística no Python

1. Tratar as Variáveis da base de dados: Missing e Categoricals
2. Verificar a correlação entre as variáveis
3. Dividir a base em treinamento e teste
4. Seleção de variáveis
4. Ajustar um modelo de regressão Logística
5. Prever na base de teste
6. Avaliar a qualidade do ajuste do modelo: acurácia, precision, recall



# DÚVIDAS?!



## Referências

1. <https://ebmacademy.wordpress.com/2015/08/17/o-fantasma-da-regressao-logistica/>
2. <https://www.kaggle.com/kost13/us-income-logistic-regression>
3. [http://planspace.org/20150423-forward selection with statsmodels/](http://planspace.org/20150423-forward_selection_with_statsmodels/)





# Obrigada

***Cristiane Rodrigues***

*[crisrodrigues\\_27@hotmail.com](mailto:crisrodrigues_27@hotmail.com)*

