Tera

Estatística e Modelagem de Dados

Aula 19:

Algoritmos de Classificação - 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 - SAS Institute Brasil

  - Consultora de Pré Vendas - SAS Institute Brasil

  - Professora do curso SAS Academy for Data Science

crisrodrigues_27@hotmail.com

Índice

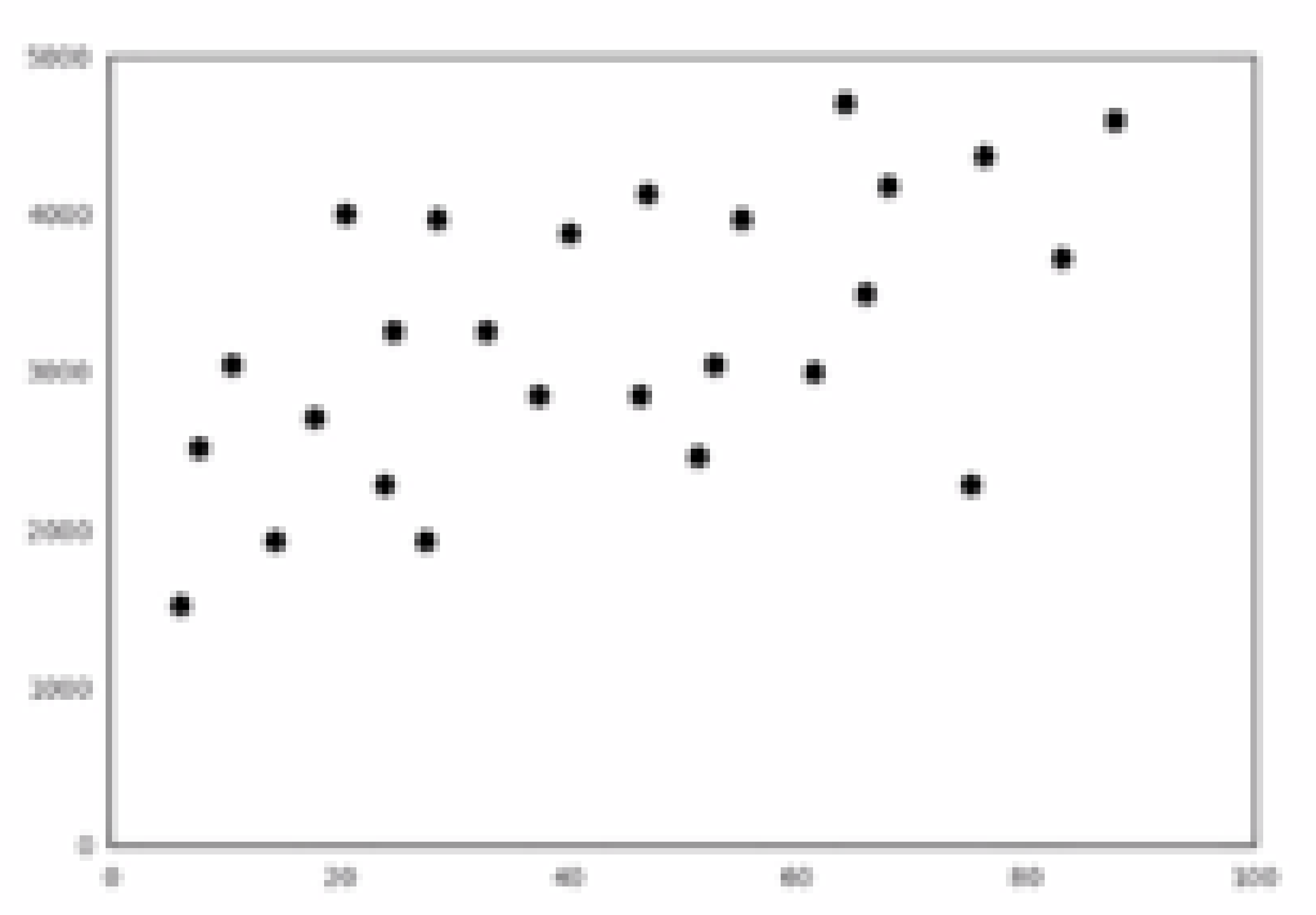
- Revisão Regressão Linear
- Motivação
- Forma Funcional do Modelo de Regressão Logística
- Aplicações
- Superfície de Ajuste e Interpretação
- Odds Ratio
- Ponto de Corte
- Tratamento das variáveis
- Seleção de Variáveis
- Matriz de Confusão
- Curva ROC

Revisão: Regressão Linear Simples

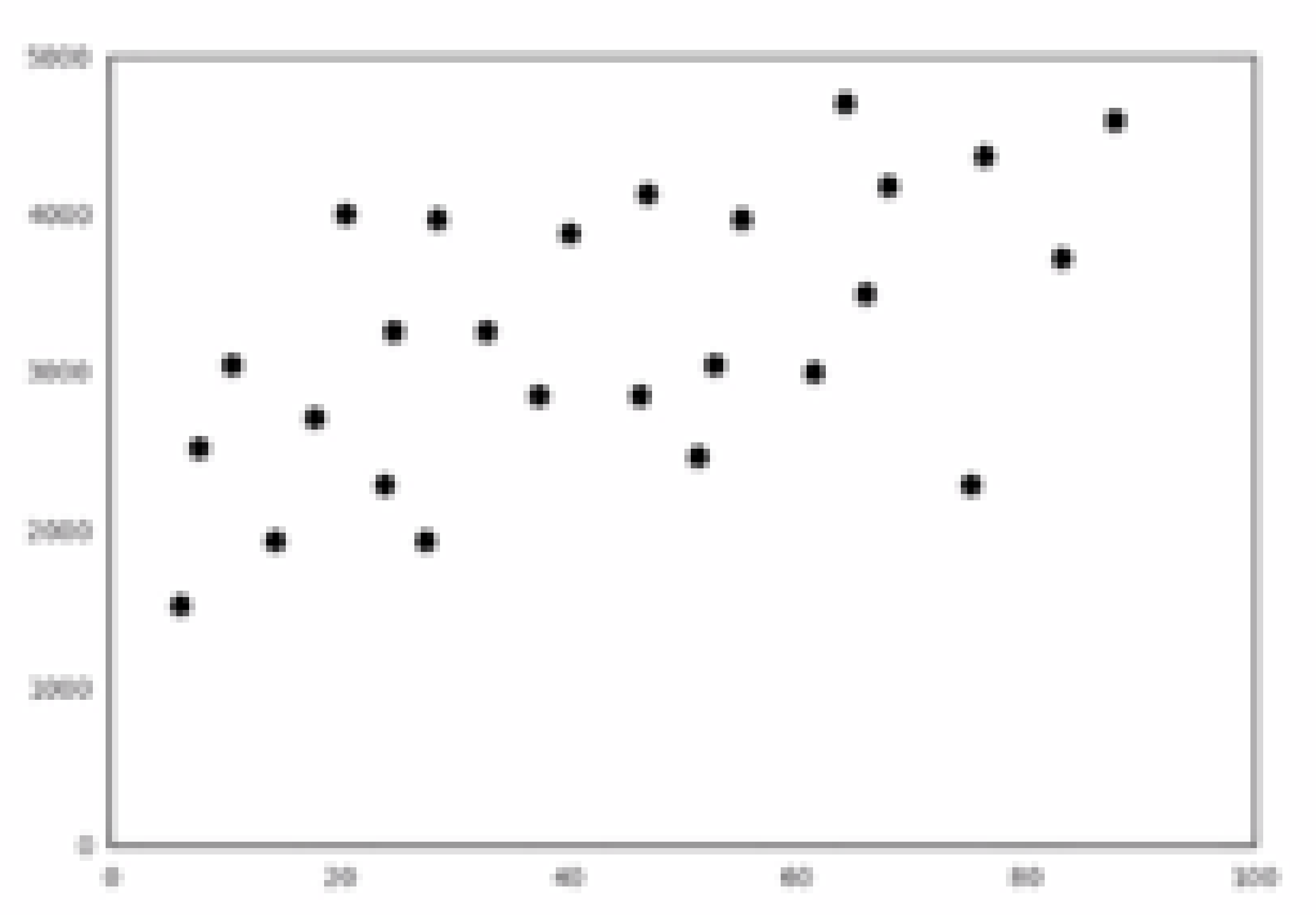
- $y = \beta_0 + \beta_1 x + \mathrm{e}$
  - $y = target$
  - $x = variável\ preditora$ contínua
  - $\beta_0, \beta_1 = parâmetros\ do\ modelo$

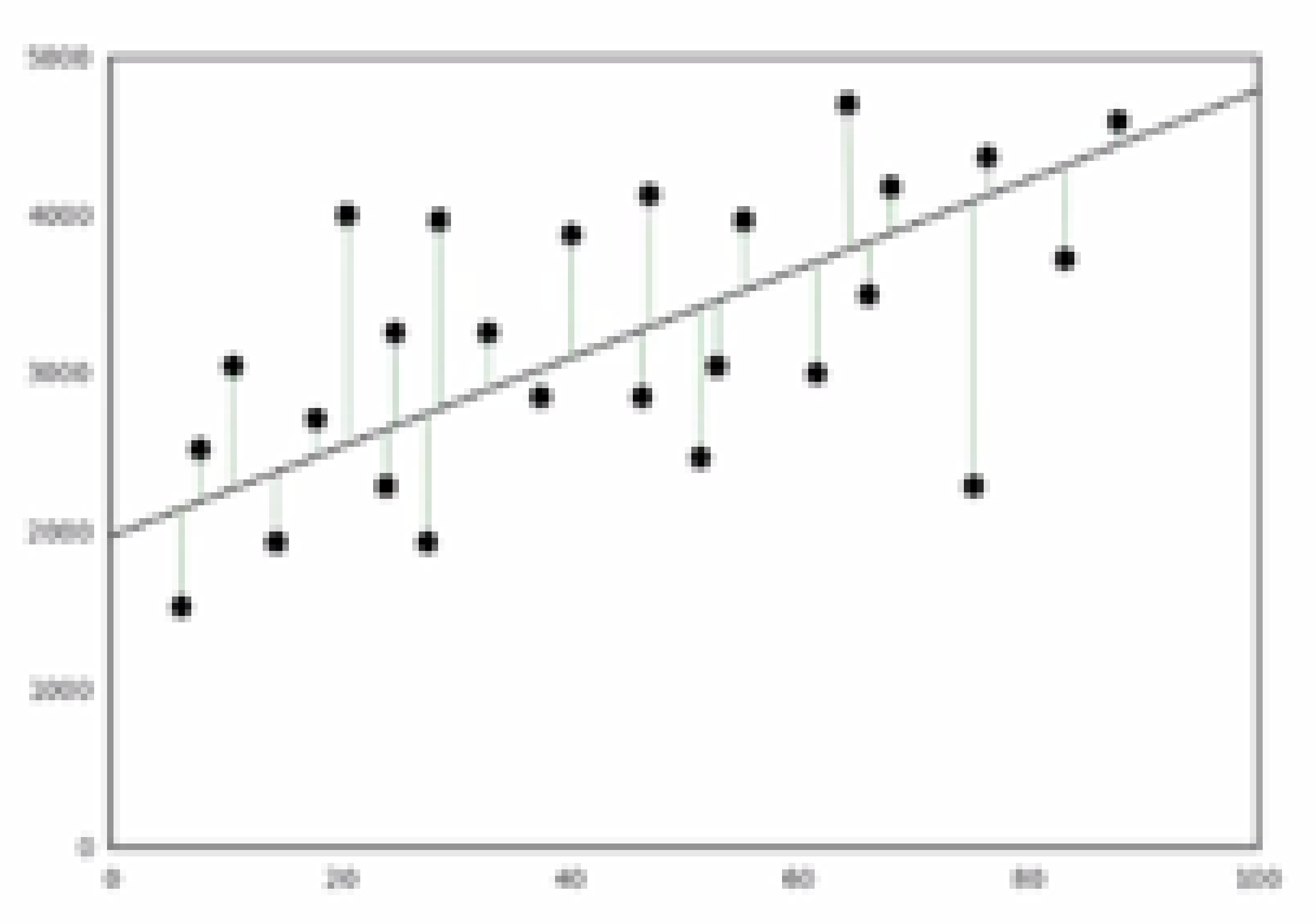
- Como escolher β_0 e β_1?

Revisão: Regressão Linear Simples

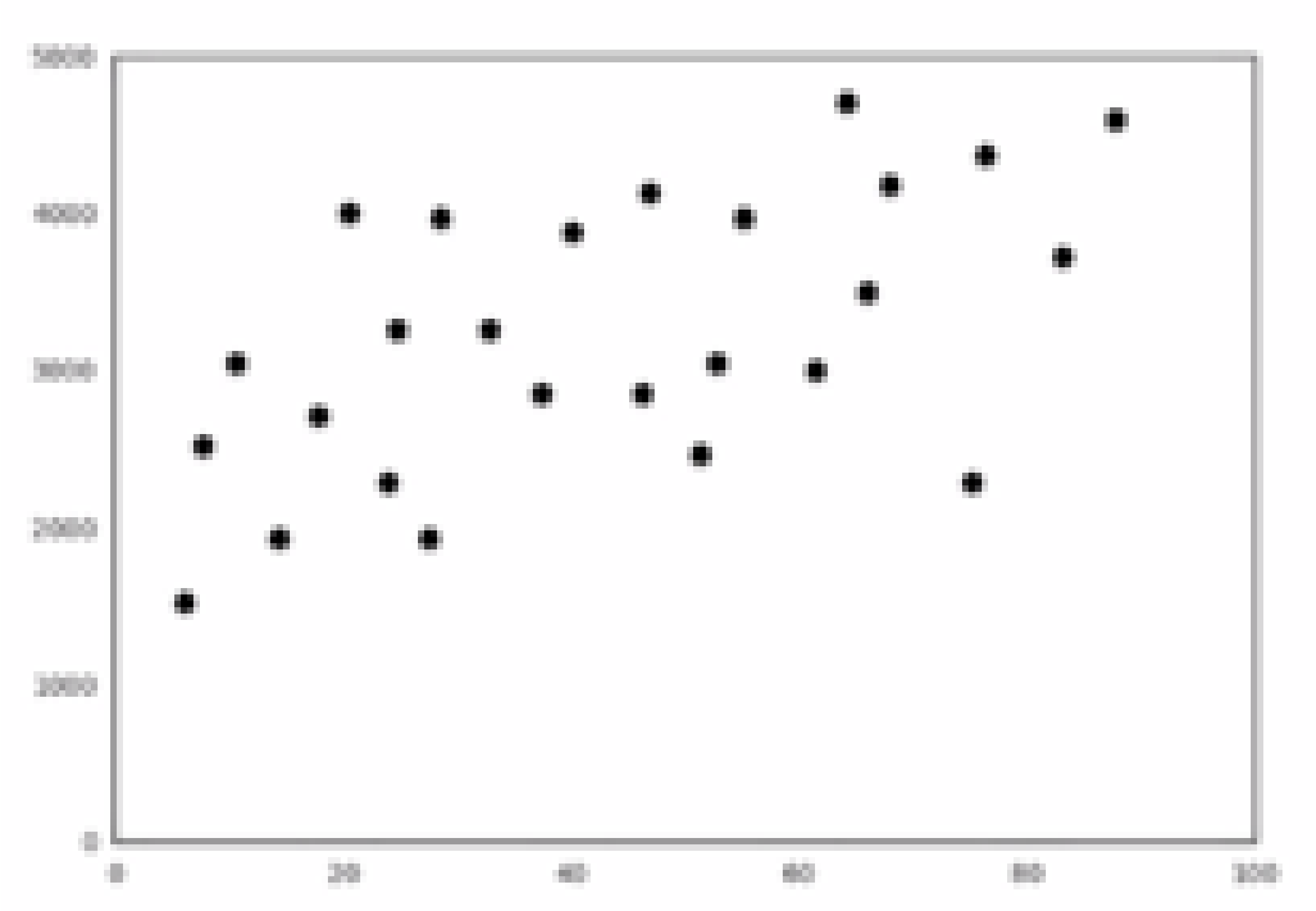


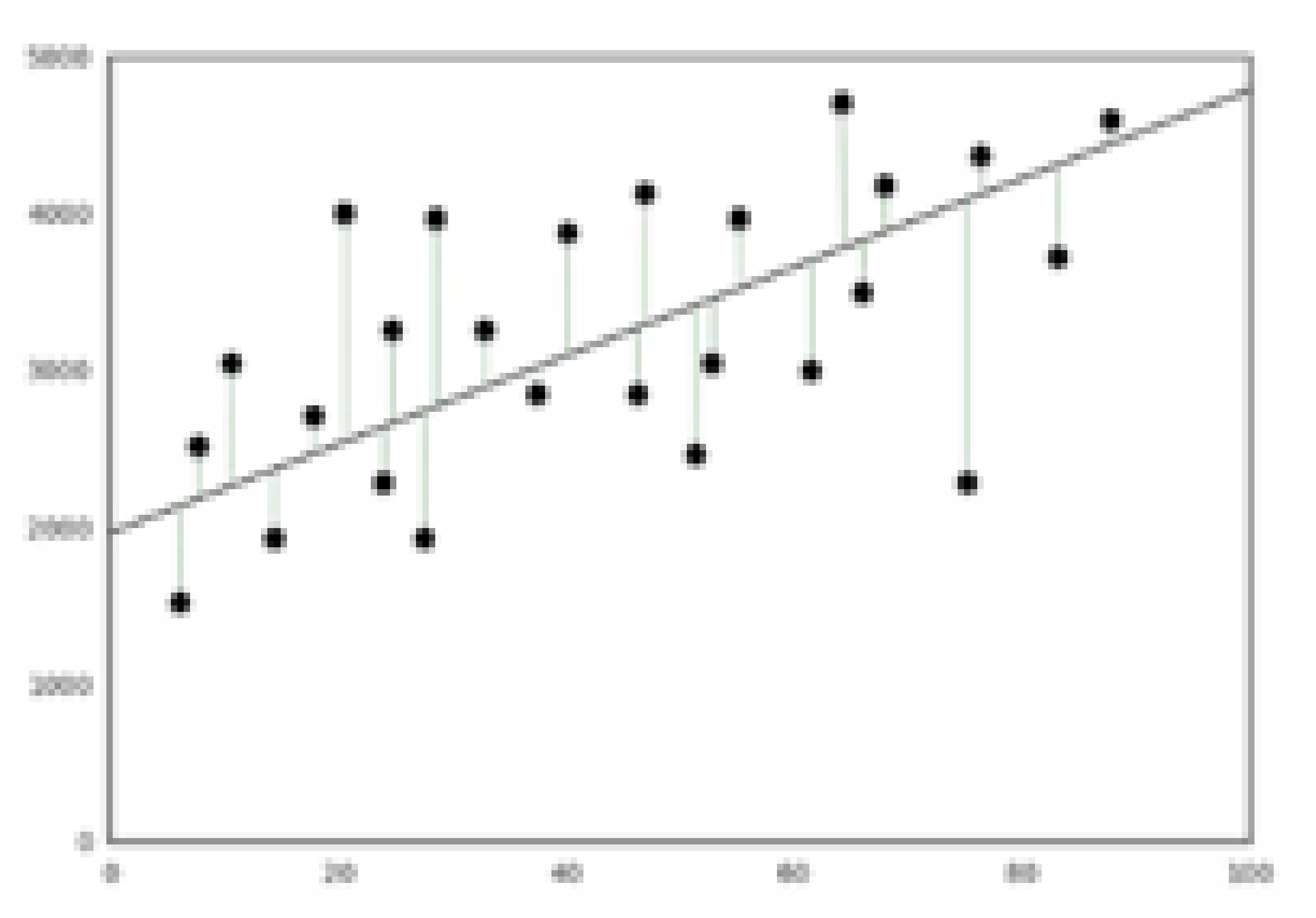
Revisão: Regressão Linear Simples

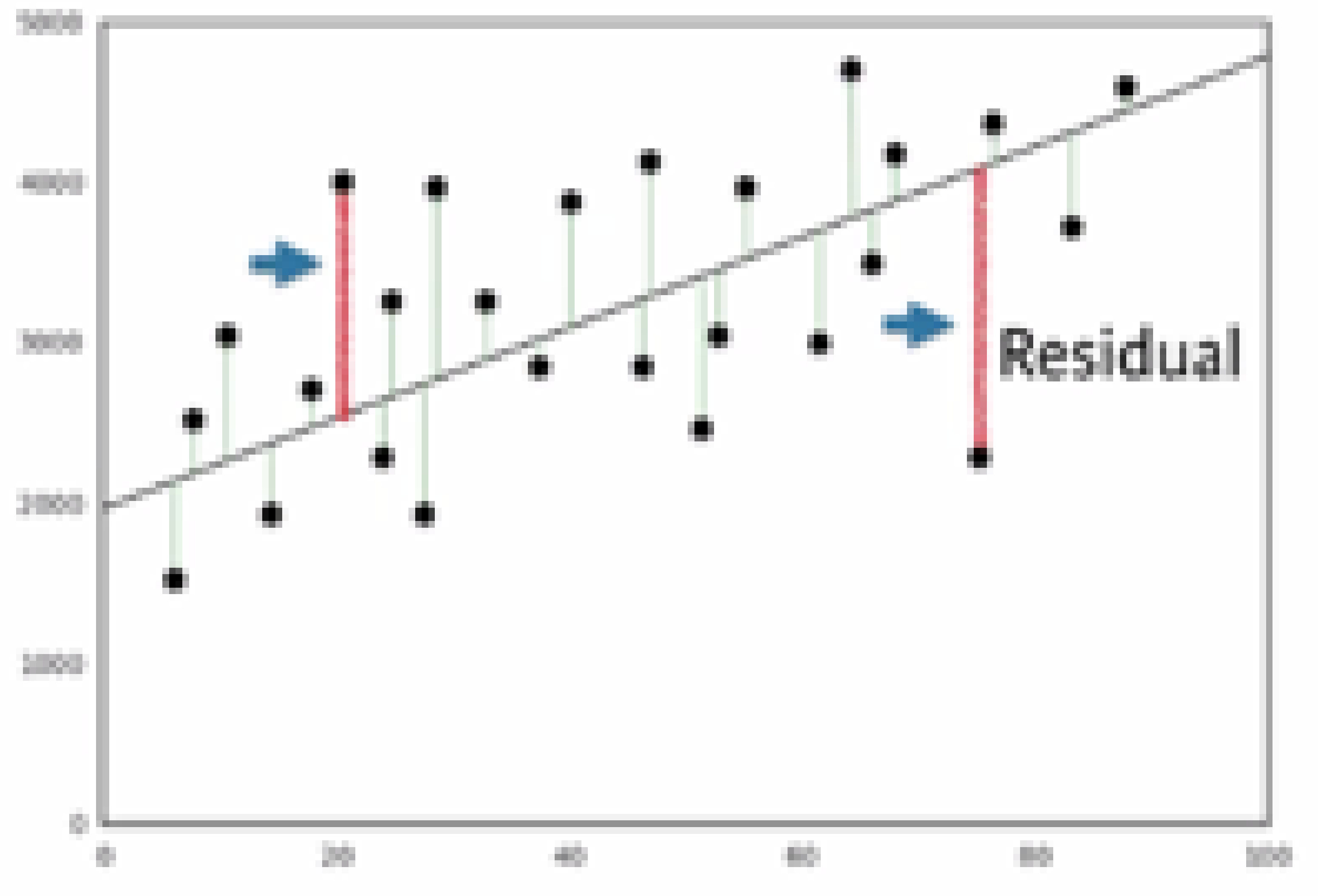




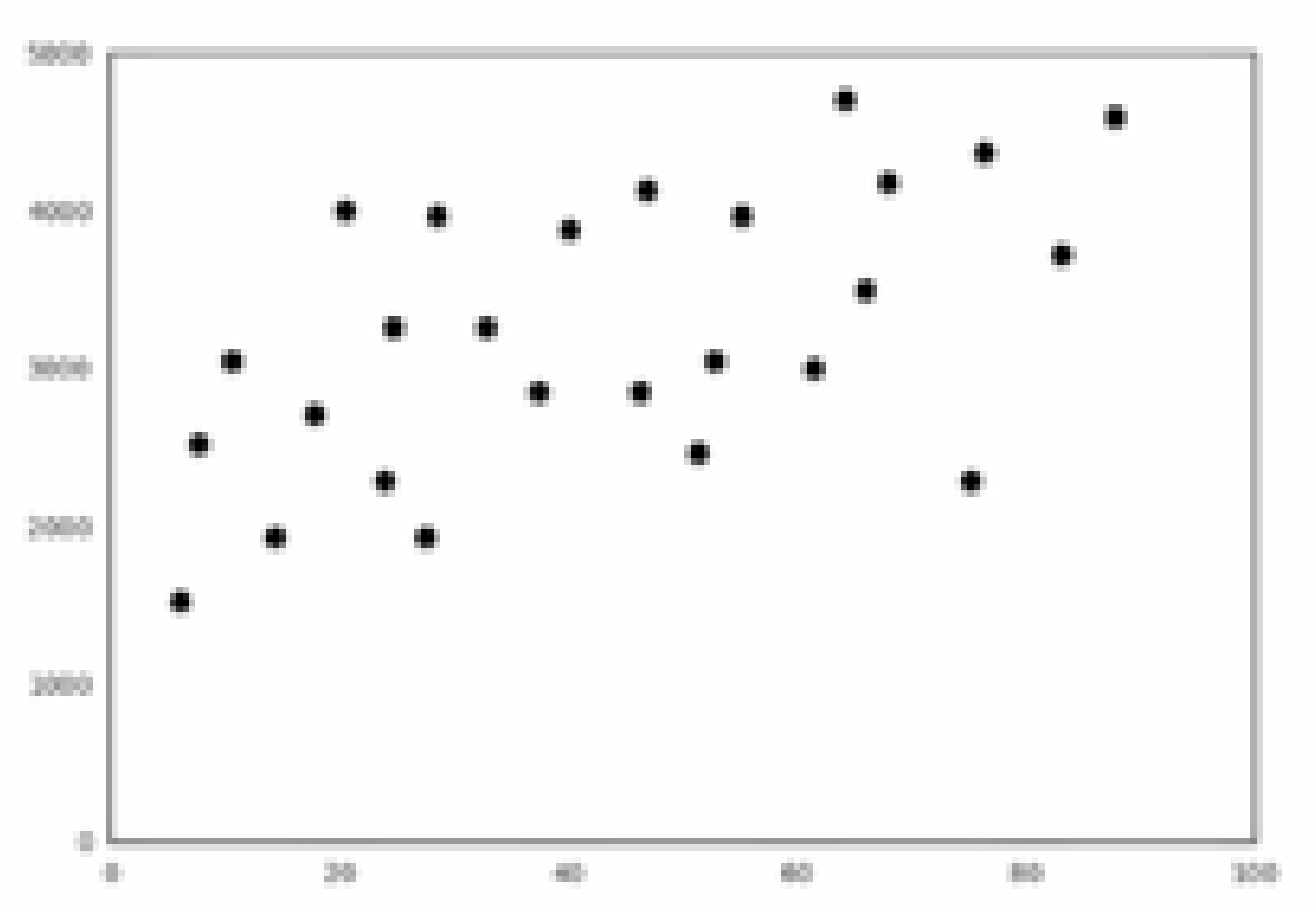
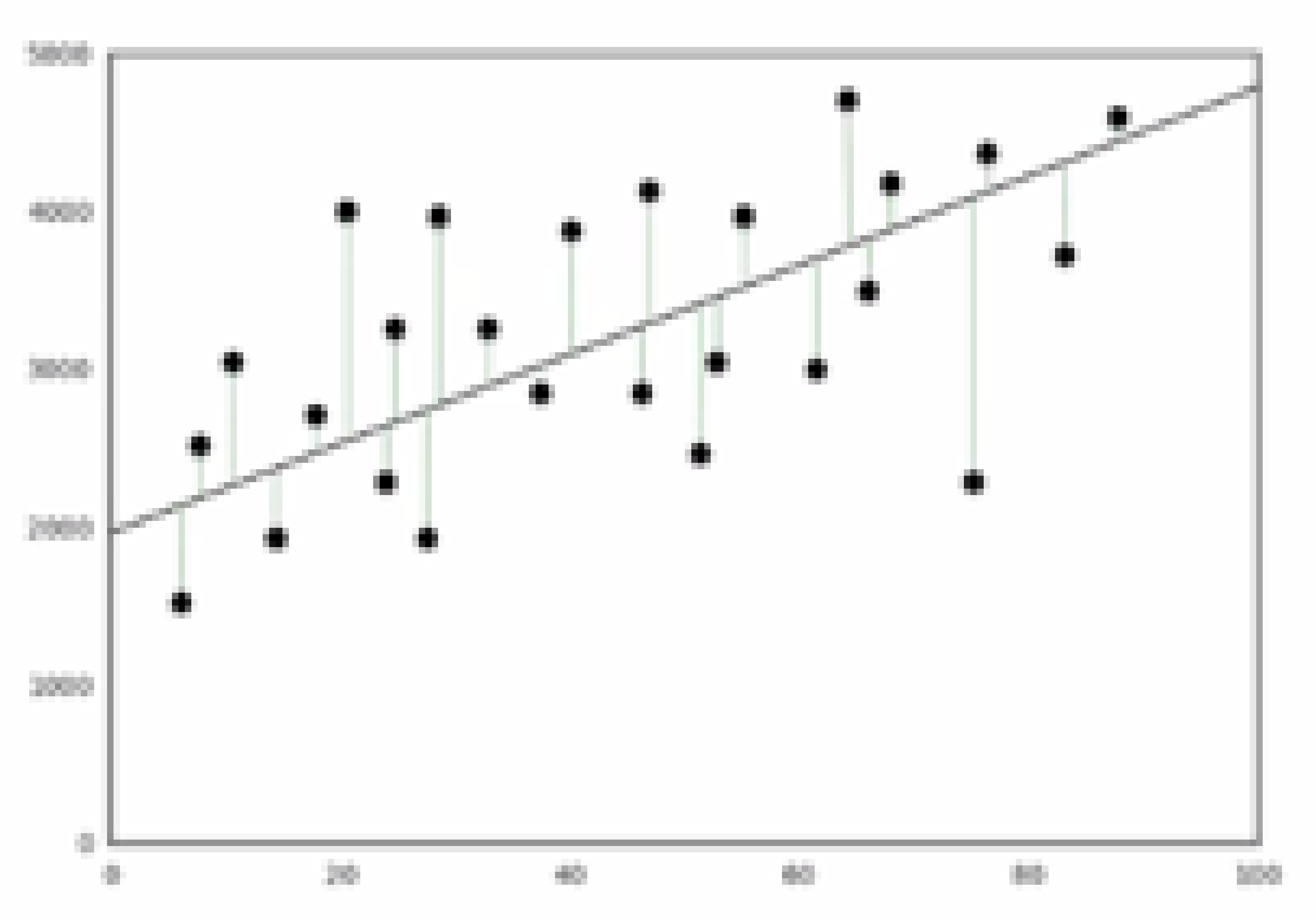
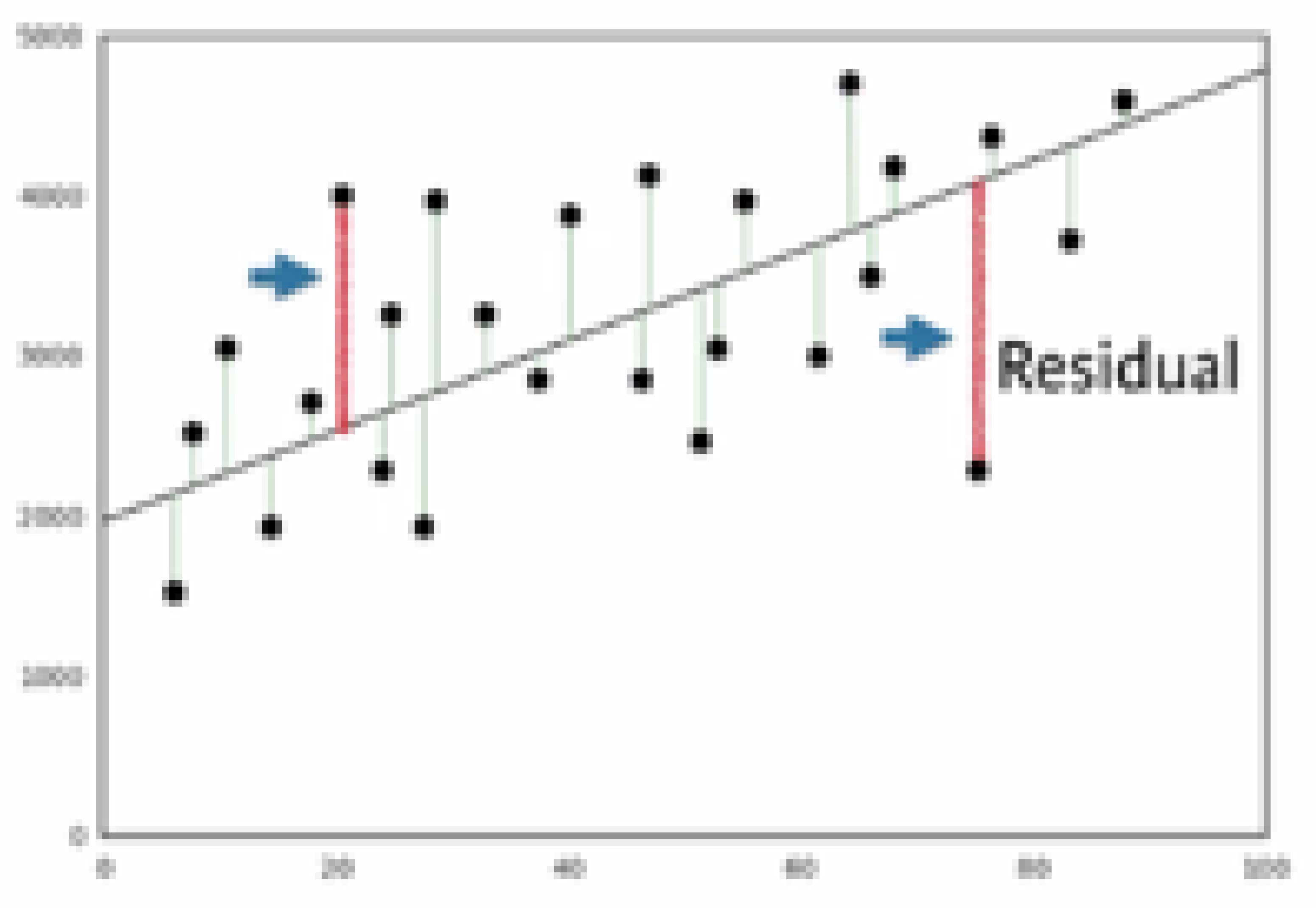
Revisão: Regressão Linear Simples





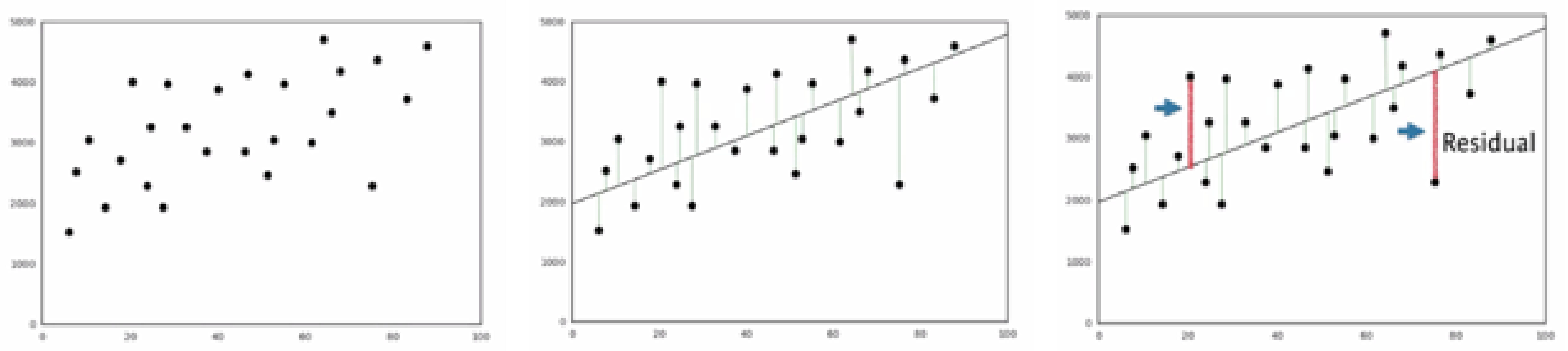


Revisão: Regressão Linear Simples

- 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

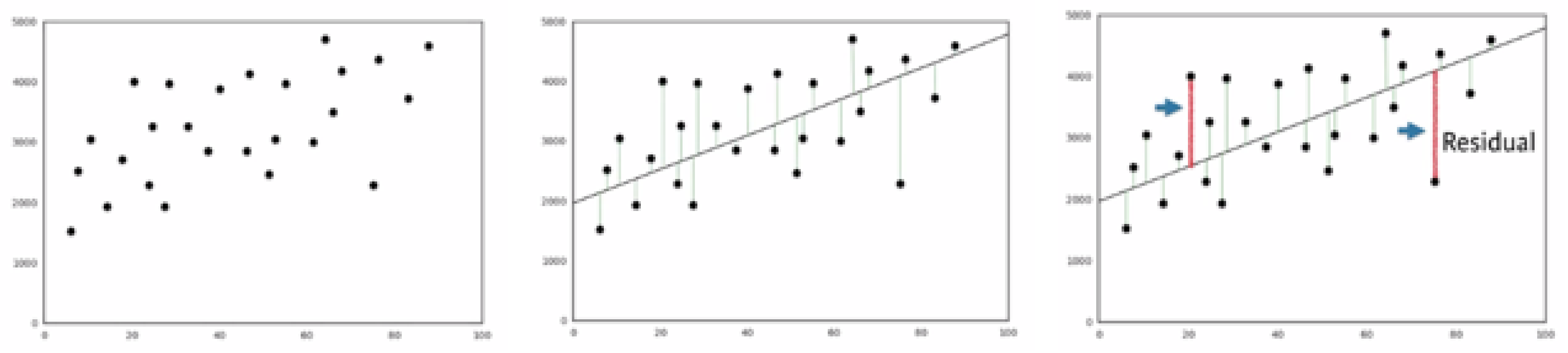


- Na regressão linear o objetivo é escolher a reta que minimiza a função de erro, ou seja, que diminui a distância entre o ajuste e os dados

Os valores de β_0 e β_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



- 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

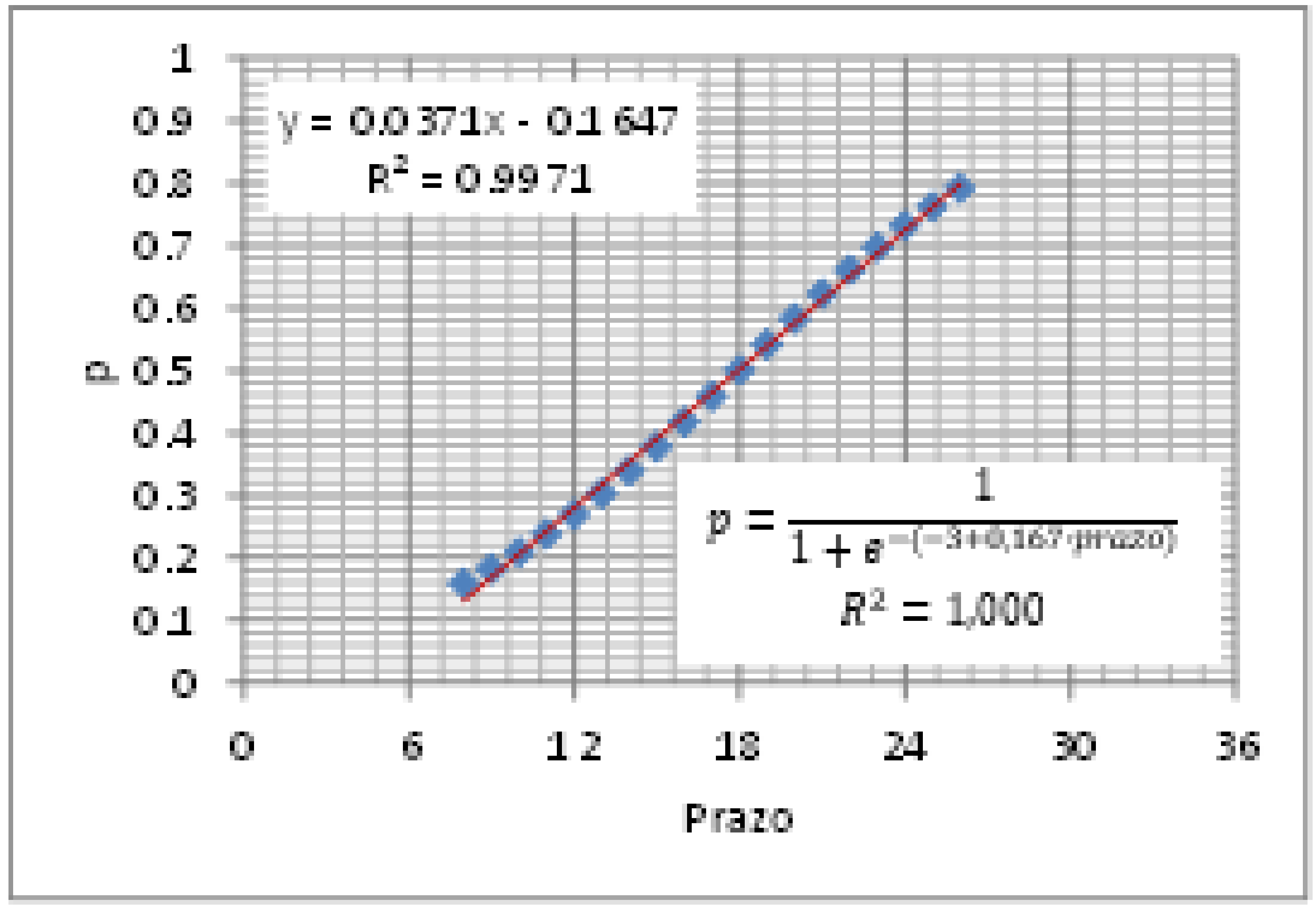
- Na regressão linear múltipla temos a inserção de mais variáveis preditoras e podemos escrever o modelo da seguinte forma:

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

Comparação

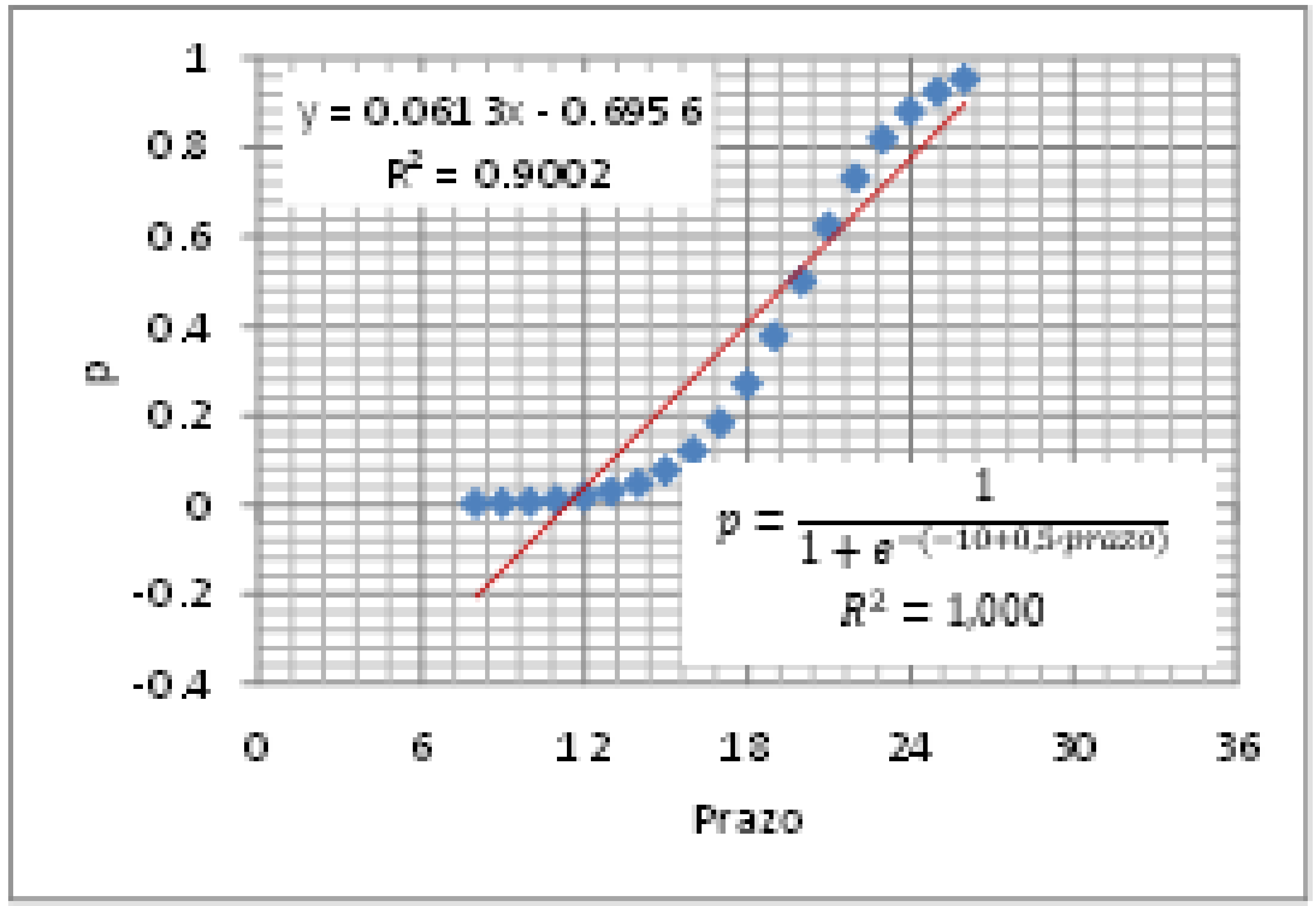
Ajustando modelos de regressão linear para dois tipos de dados diferentes

- Probabilidade variando entre 0,15 e 0,85 substituição



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

- Probabilidade menor que 0,15 ou maior que 0,85



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.

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.

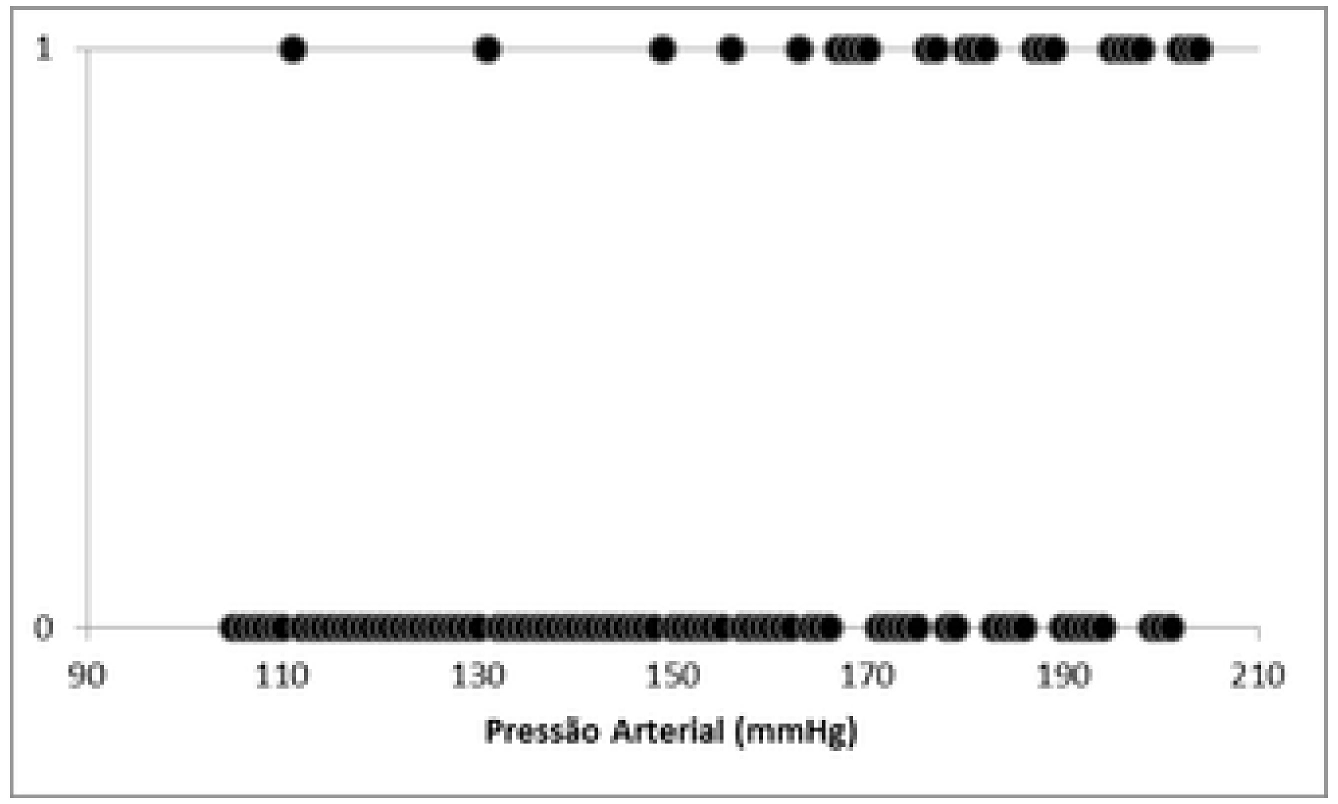
Por exemplo:

Queremos saber os valores de pressão arterial entre pessoas que tiveram ou não um AVC.

Vamos classificar

- "presença de AVC" igual a 1
- "ausência de AVC" igual a 0

teremos um gráfico tipo o abaixo, o qual não parece se ajustar bem com uma reta

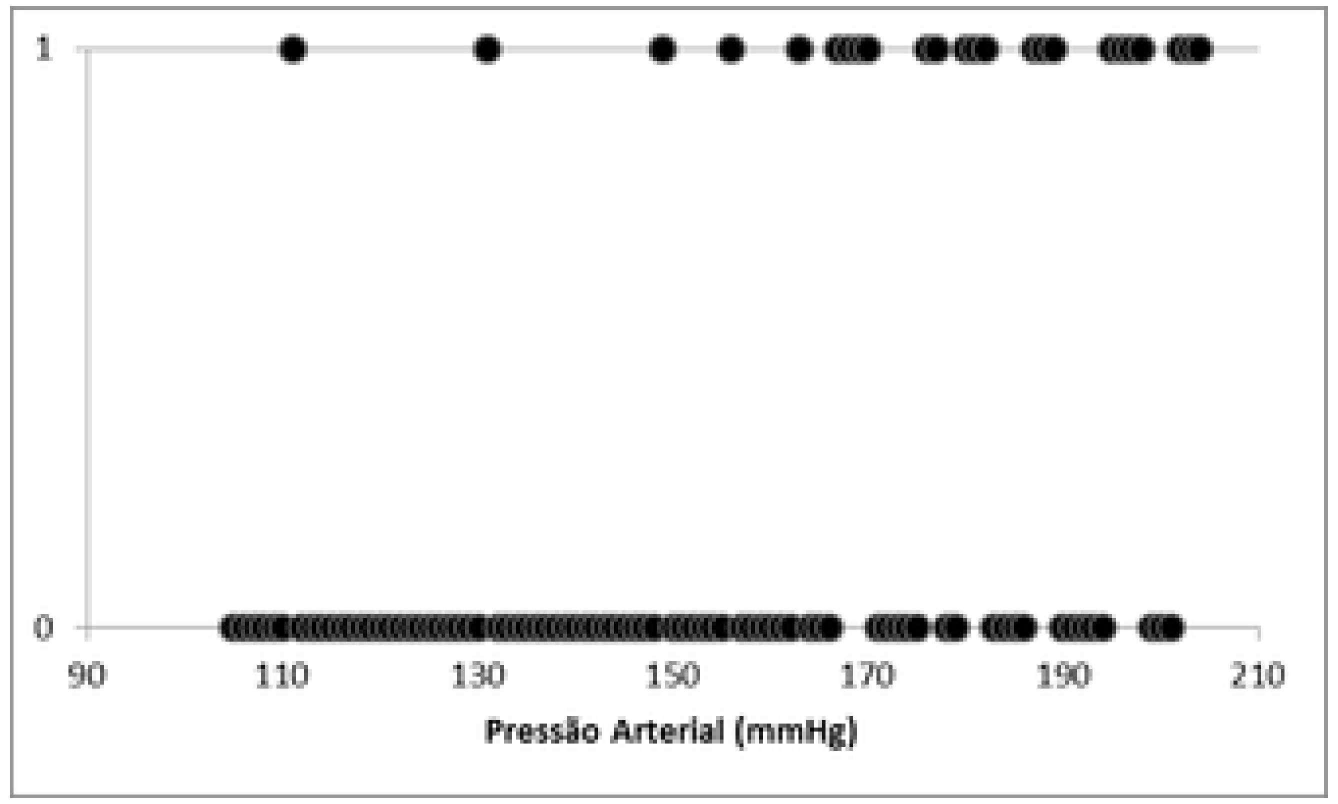


Motivação

- 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.
Vamos classificar
- "presença de AVC" igual a 1
- "ausência de AVC" igual a 0

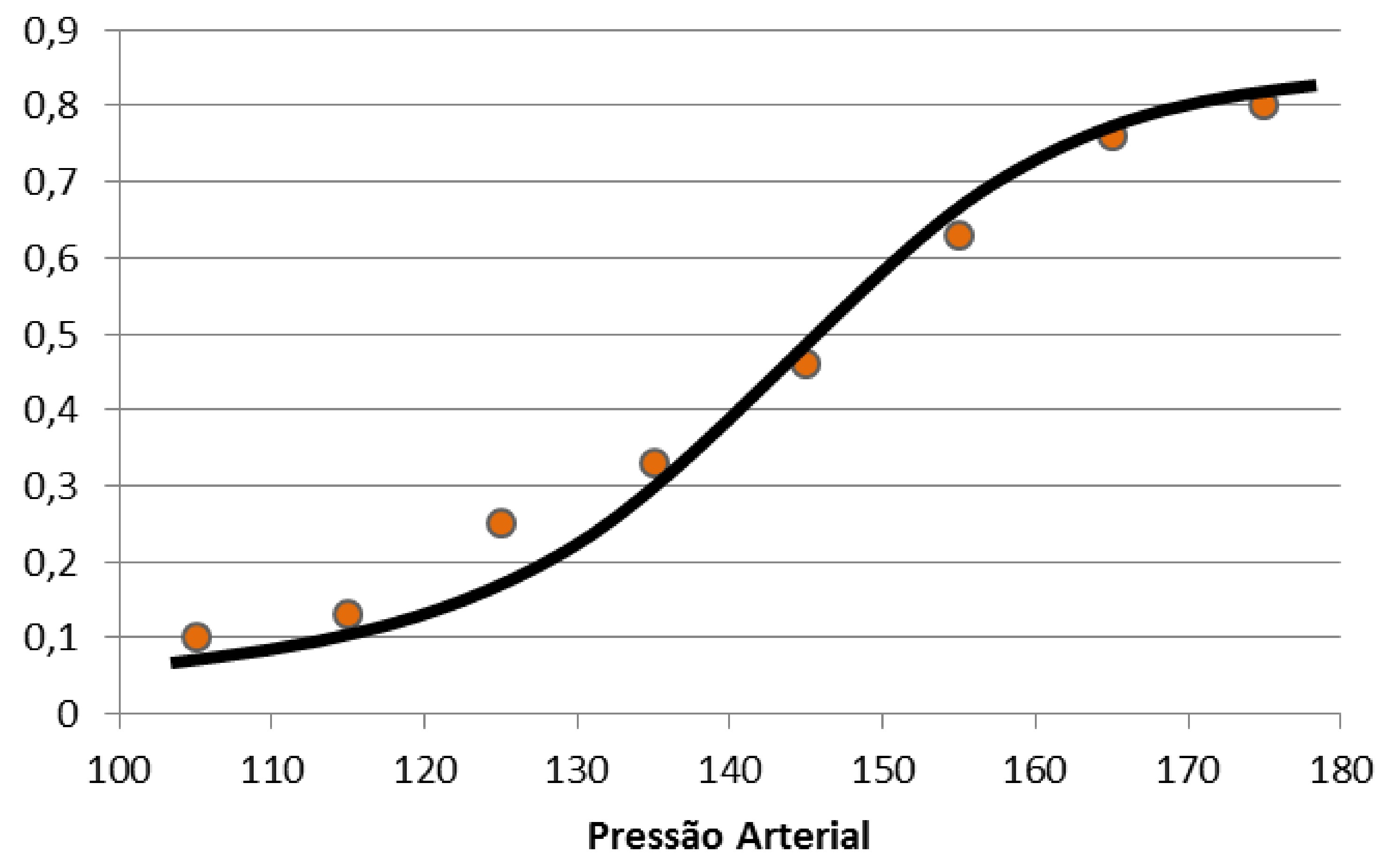
teremos um gráfico tipo o abaixo, o qual não parece se ajustar bem com uma reta



- 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:



- Algum estatístico percebeu 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:

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

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

$$\eta = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

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

Função de ligação logit

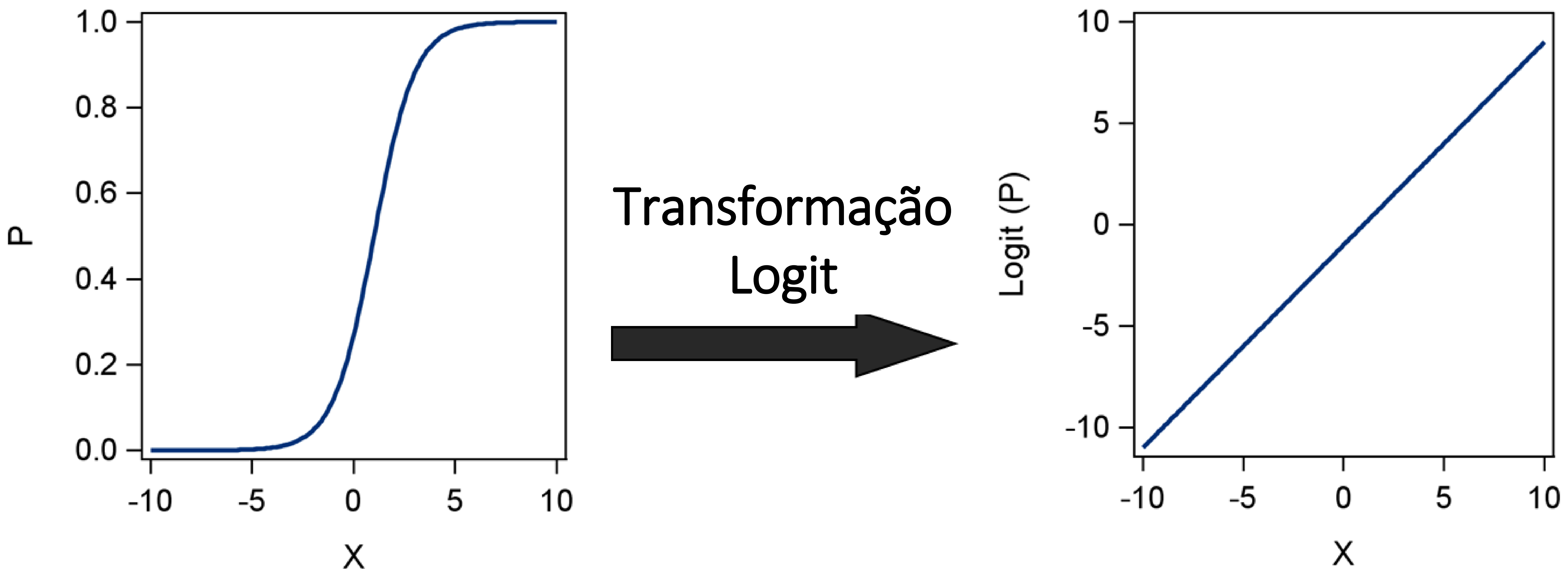
- Essa probabilidade em forma de S é muito difícil de interpretar pois o y aumenta em velocidades diferentes ao longo do eixo x.
- A ideia é tornar a equação uma reta novamente para ficar mais fácil de interpretar o resultado
- Para fazer isso, utilizamos a transformação Logit, a qual é composta por duas transformações
  1. Transformação ODDS
  2. Transformação Logaritmica

$$logit(p_i) = \eta = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$$

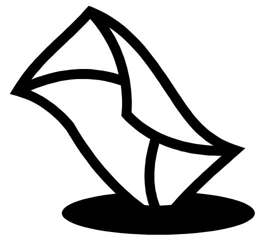
Desta forma voltamos para uma relação linear entre o logit de p_i e as variáveis input.

Transformação

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



Aplicações

- <u>Marketing</u>: 

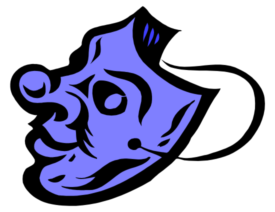
Objetivo: Encontrar segmentos de clientes mais prováveis a aderir a uma promoção
Target: Se o cliente aderiu ou não a alguma promoção passada
Inputs: Histórico de compras, Localidade, Salário,...

- <u>RH</u> – 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,...

- <u>Credit Scoring</u>: 

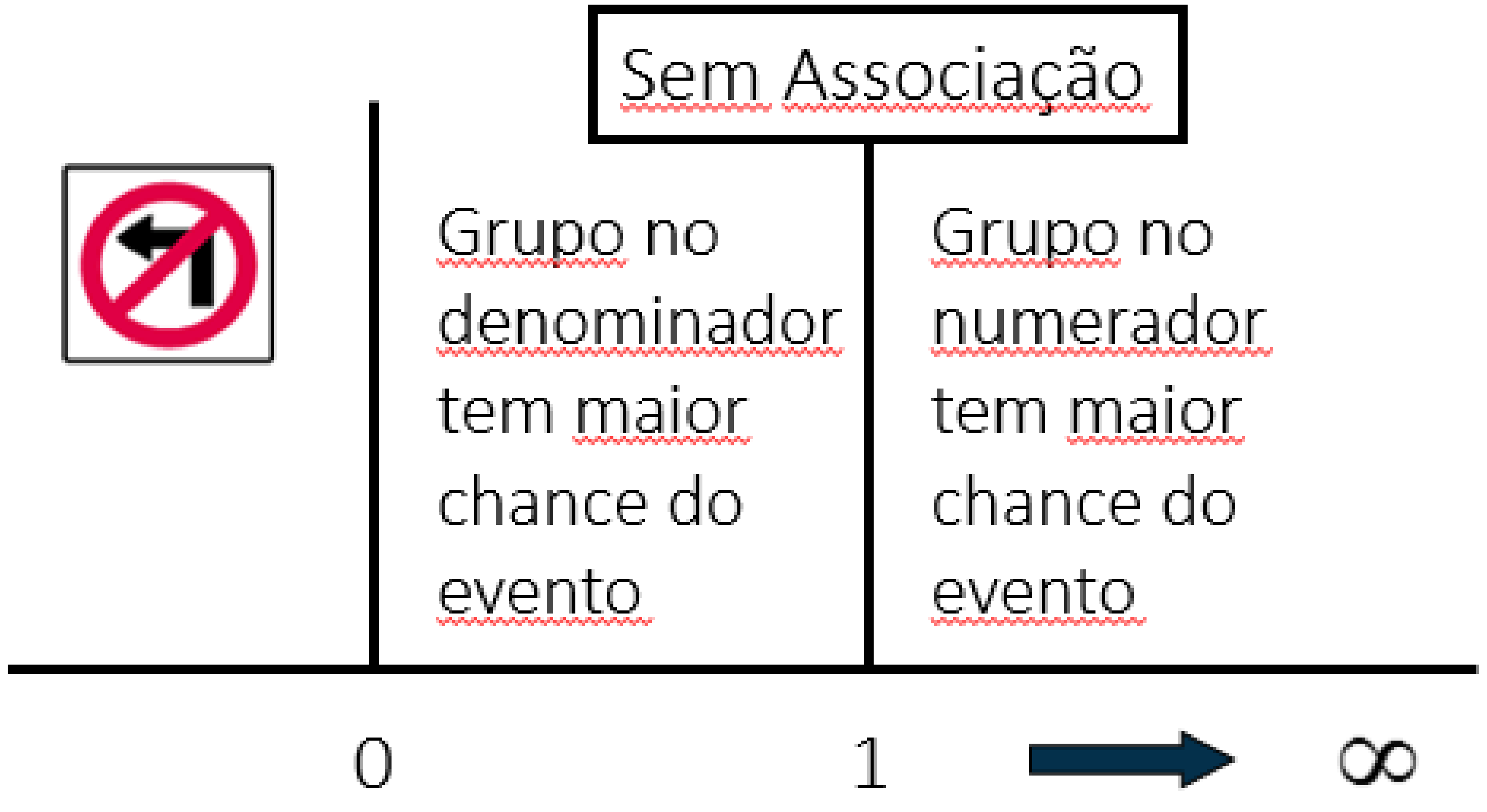
Objetivo: Verificar a probabilidade do cliente entrar em default
Target: Se o funcionário entrou ou não e default nos últimos 90 dias
Inputs: Saldo médio em conta, se recebe em conta, saldo máximo, quantidade de meses em risco

- <u>Detecção de Fraude</u>: 

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, ...

Interpretação dos coeficientes - Odds Ratio

- $\boldsymbol{Odds} = \frac{p}{1-p} = e^{n}$, chance do evento ocorrer. Em que $n = \beta_o + \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_k x_{ki}$.
- $\boldsymbol{Odds_{Ratio}} = \frac{Odds_{grupo_A}}{Odds_{grupo_B}} = \frac{e^{n}_{grupo\_A}}{e^{n}_{grupo\_B}}$, chance do evento ocorrer se for do Grupo_A com relação ao Grupo_B



Odds Ratio ϵ $(\mathbf{0}, \infty)$

- **Odds Ratio = 1** → $\frac{Odds\ Grupo1}{Odds\ Grupo\ 2} = 1$ → p1=p2, ou seja, não há associação entre a variável preditora e a resposta
- **Odds Ratio >** 1 → $\frac{Odds\ Grupo1}{Odds\ Grupo\ 2} > 1$ → $Odds\ Grupo1 > Odds\ Grupo2$, ou seja, o grupo no numerador tem maior chance do evento ocorrer que o grupo no denominador
- **Odds Ratio < 1** → $\frac{Odds\ Grupo1}{Odds\ Grupo\ 2} < 1$ → $Odds\ Grupo1 < Odds\ Grupo2$, 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

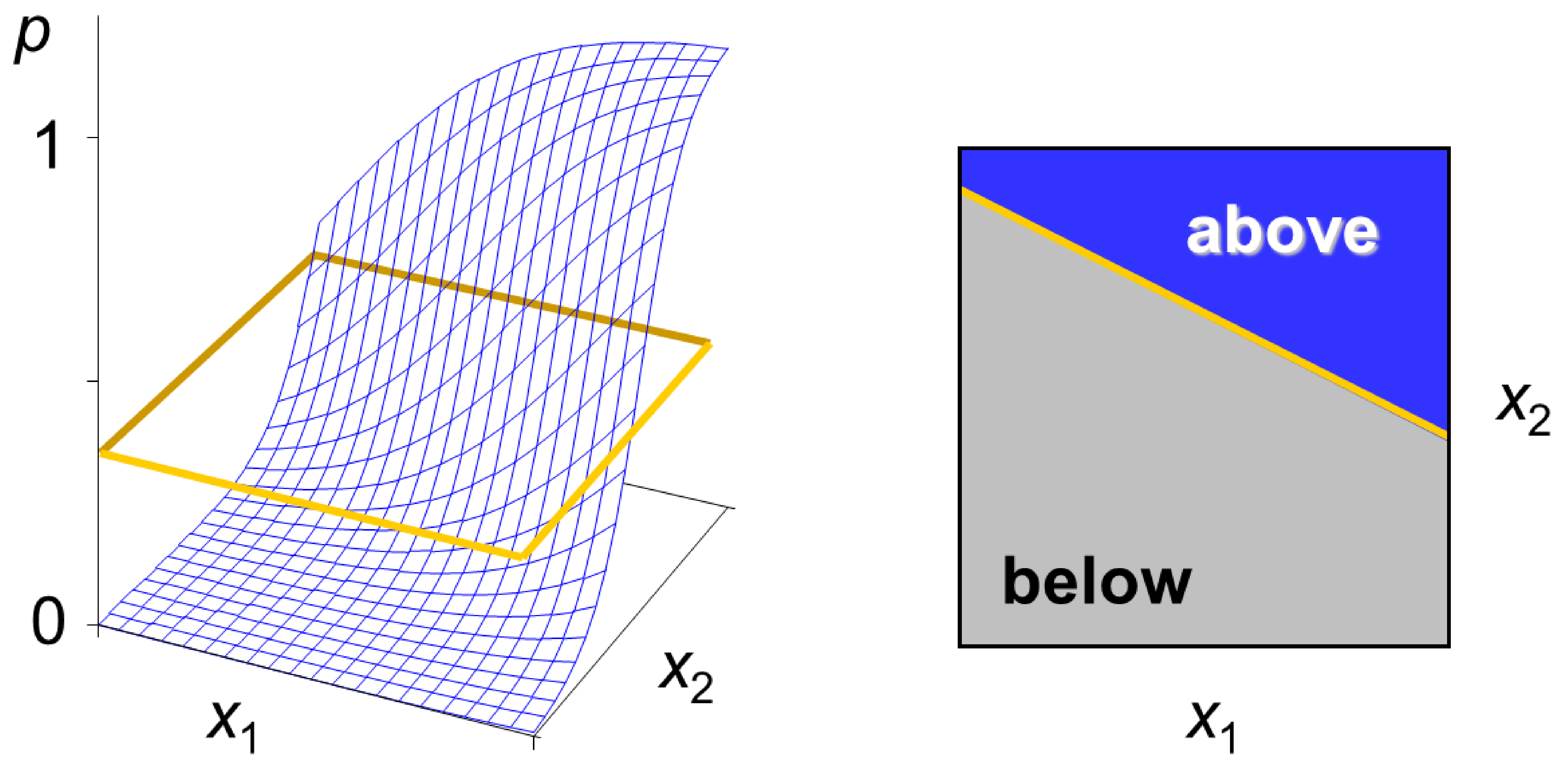
- Razão de chances estimada (Femino para Masculino) é:

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

O que isso que dizer?

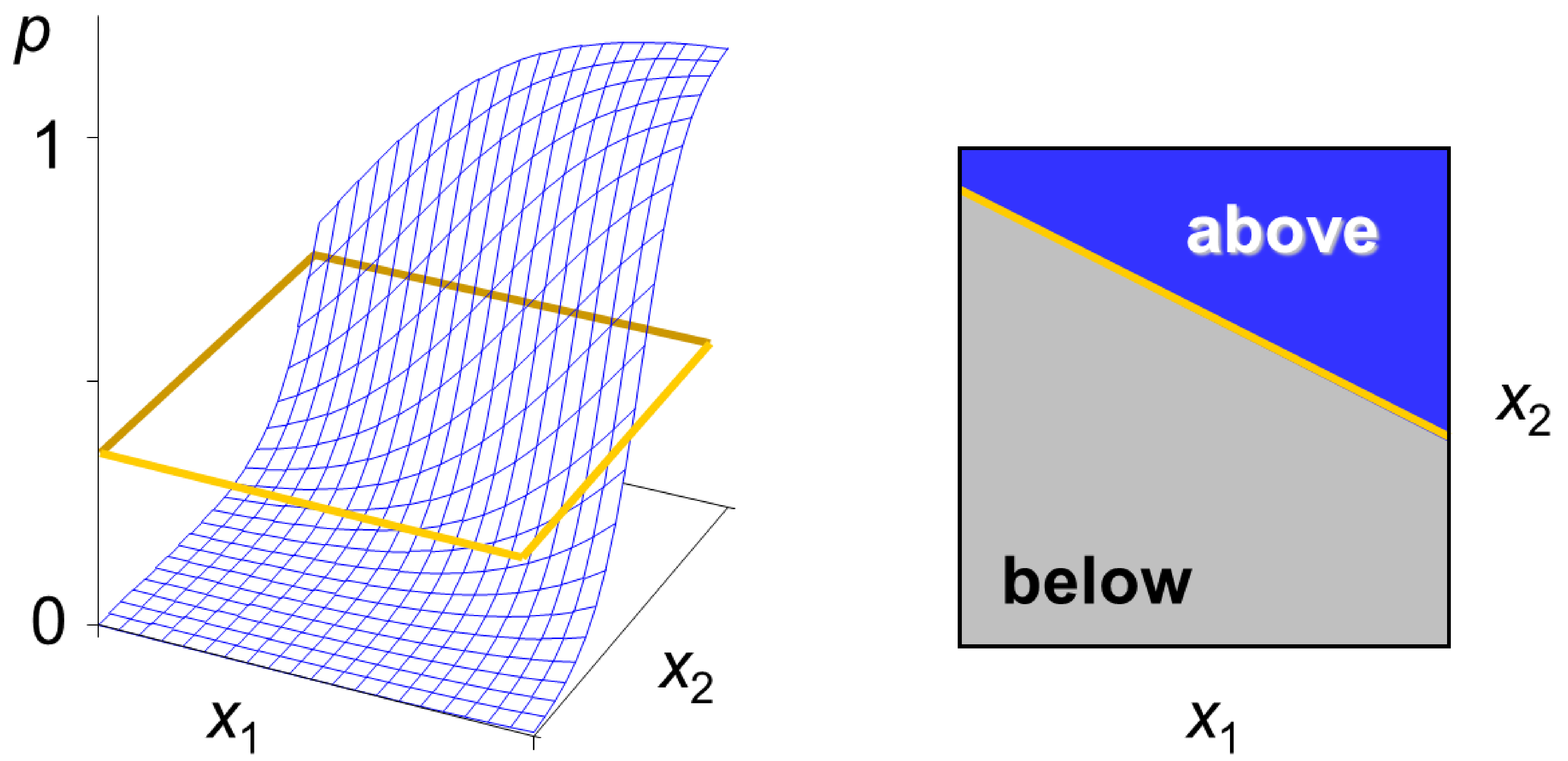
Interpretação: A chance de ocorer o sexo feminino é 1,55 vezes a chance de ocorrer o sexo masculino

Discriminação – Ponto de corte



Como definir o ponto de corte ???

Discriminação – Ponto de corte



Como definir o ponto de corte ???

- Taxa de eventos da população original
- Regras de Négócio
- Ponto de máximo da Curva ROC

Tratamento das variáveis

1. Missing

Solução:

- Complete case analysis
- Imputação
- Variáveis indicadoras de missing

2. Categoricas

- **Variáveis com muitos níveis**
  - Aumento na dimensão ao criar dummys para cada nível
  - Produção de inputs redundantes e irrelevantes
- **Thresholding**: Juntar categorias baseado no número de observações
- **Clusterização**: Juntar as categorias baseado na taxa de resposta

Tratamento das variáveis

3. Redundância

Variáveis input altamente correlacionadas

Problemas:

- 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

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

4. Irrelevância

Variáveis inputs pouco correlacionadas com a variável resposta

Problema: Pode afetar a escolha das variáveis no momento de seleção

Solução: 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

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

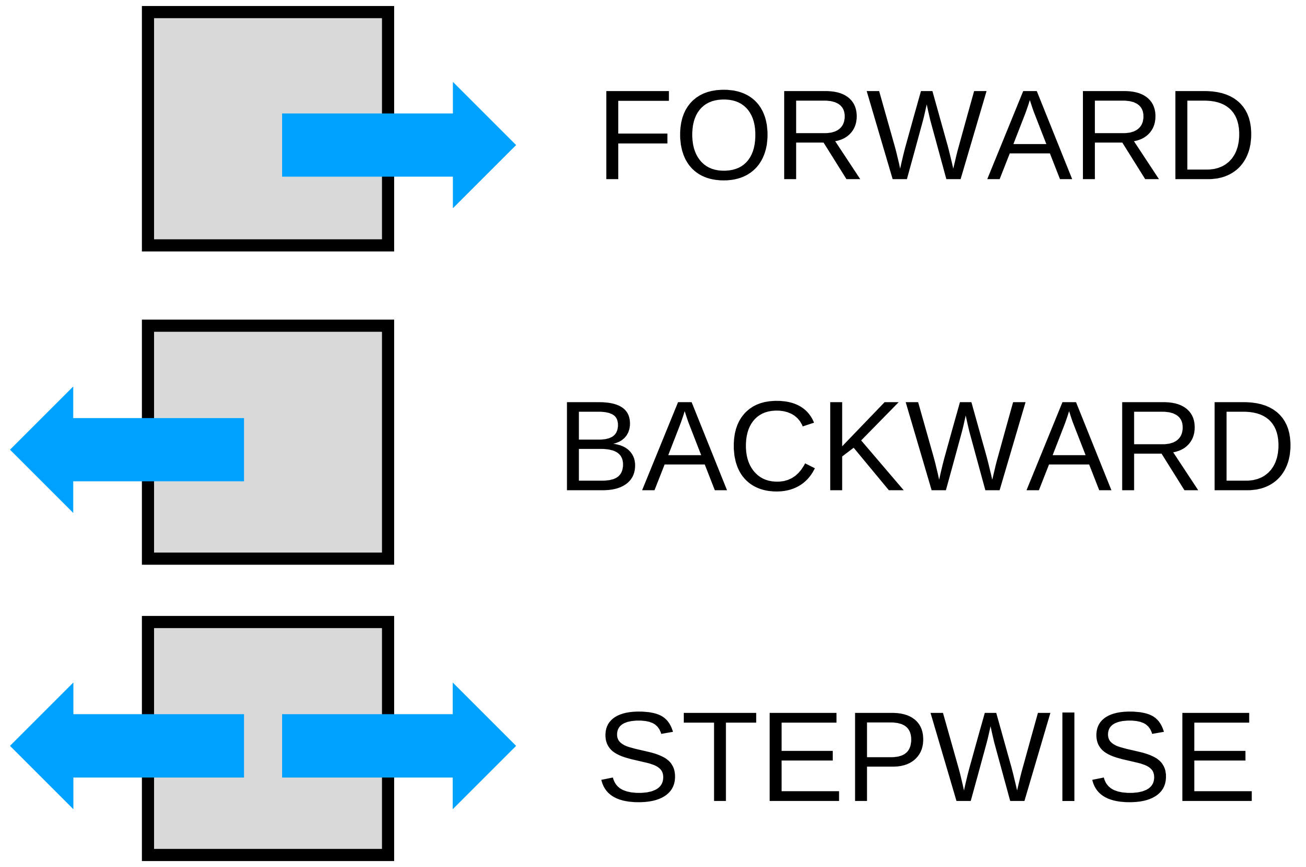
Parte 1 : Tratando as Variáveis do modelo

- Missing
- Variáveis categóricas
- Redundância/Irrelevância

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

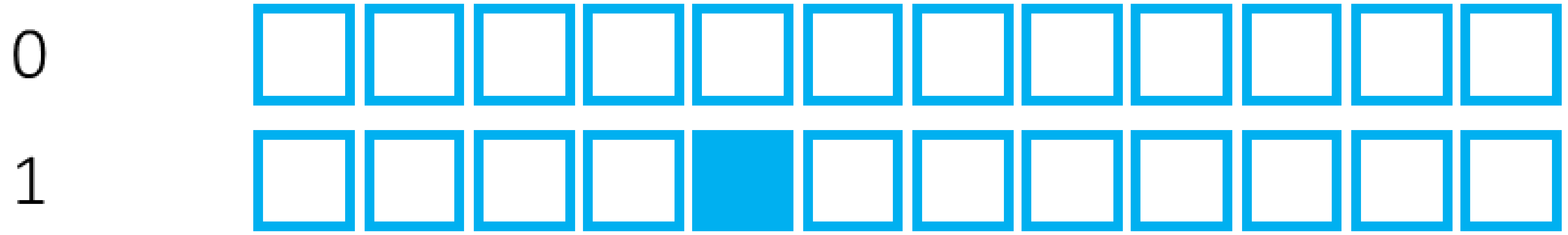
- Métodos sequenciais



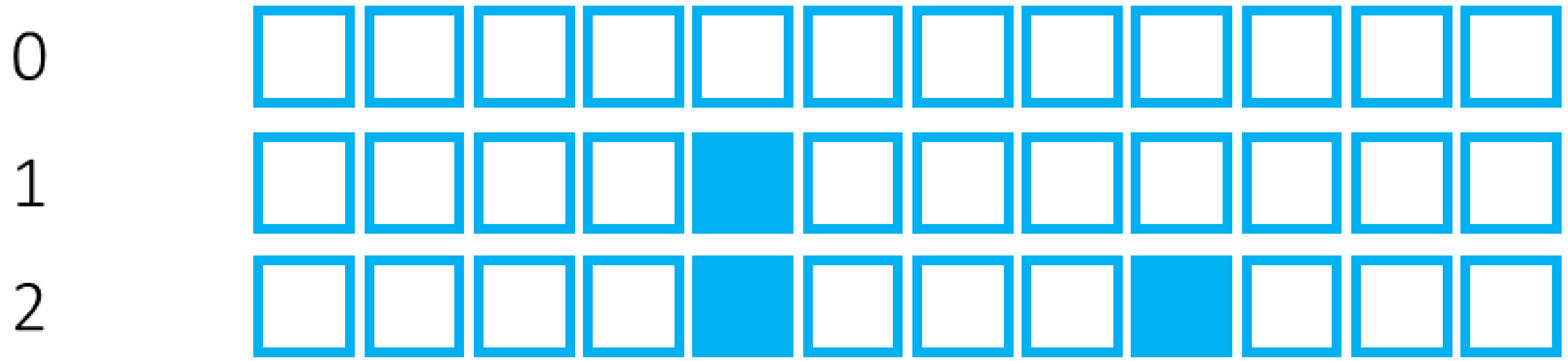
Seleção de Variáveis - Forward

0 □□□□□□□□□□□□

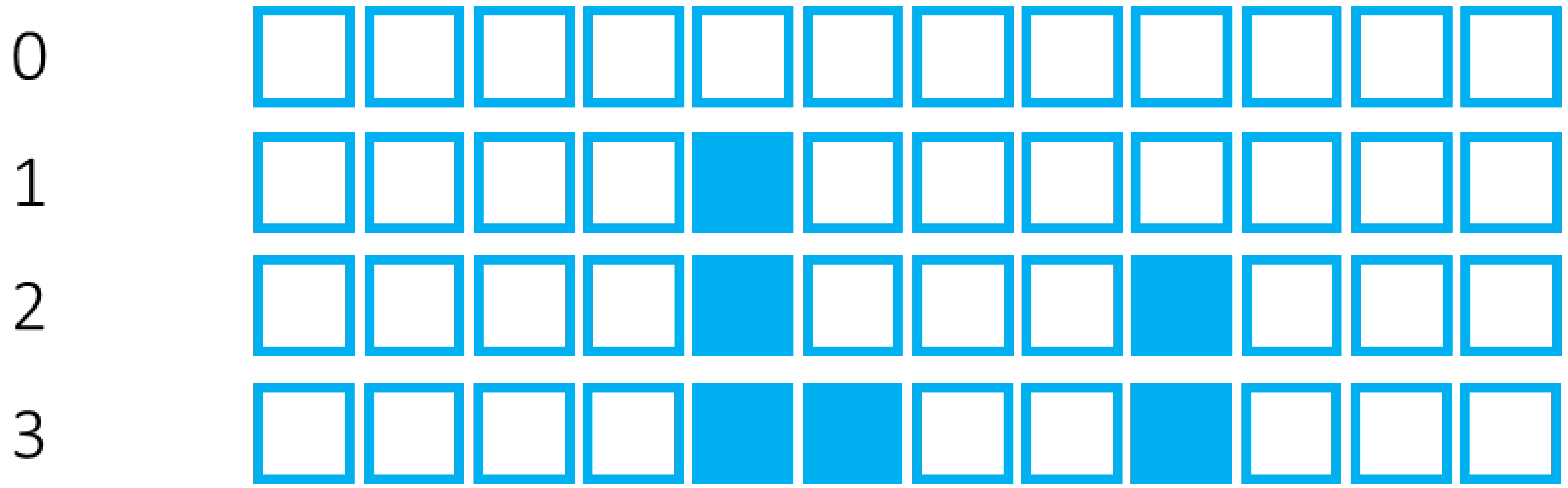
Seleção de Variáveis - Forward



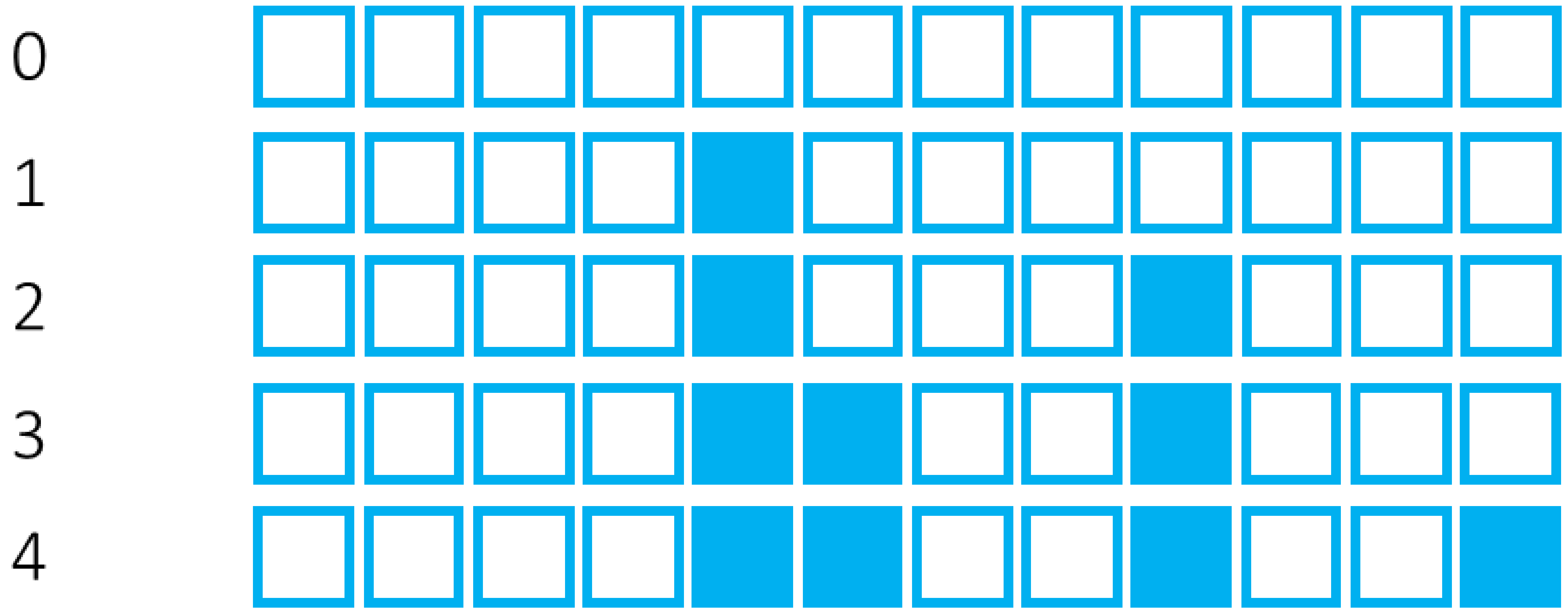
Seleção de Variáveis - Forward



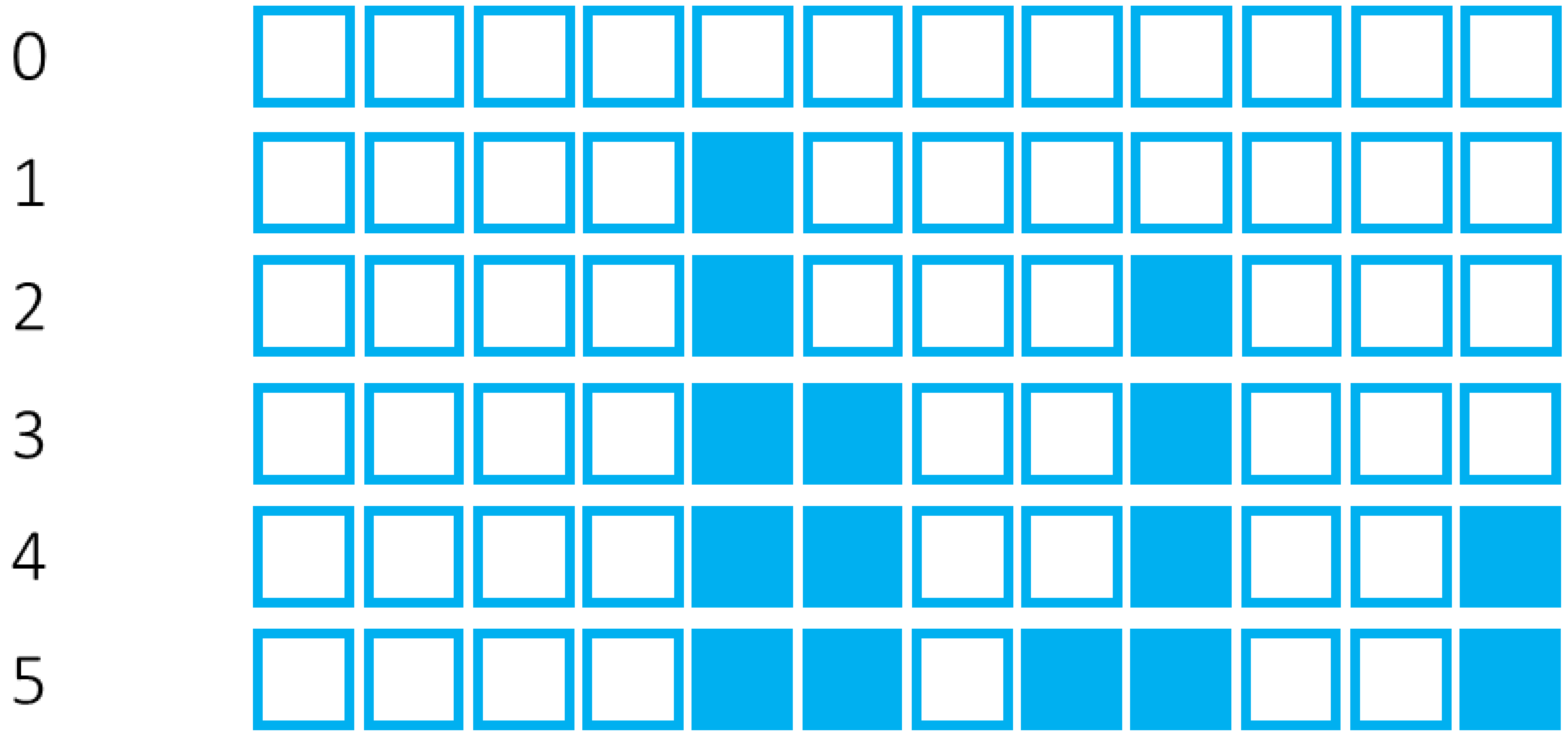
Seleção de Variáveis - Forward



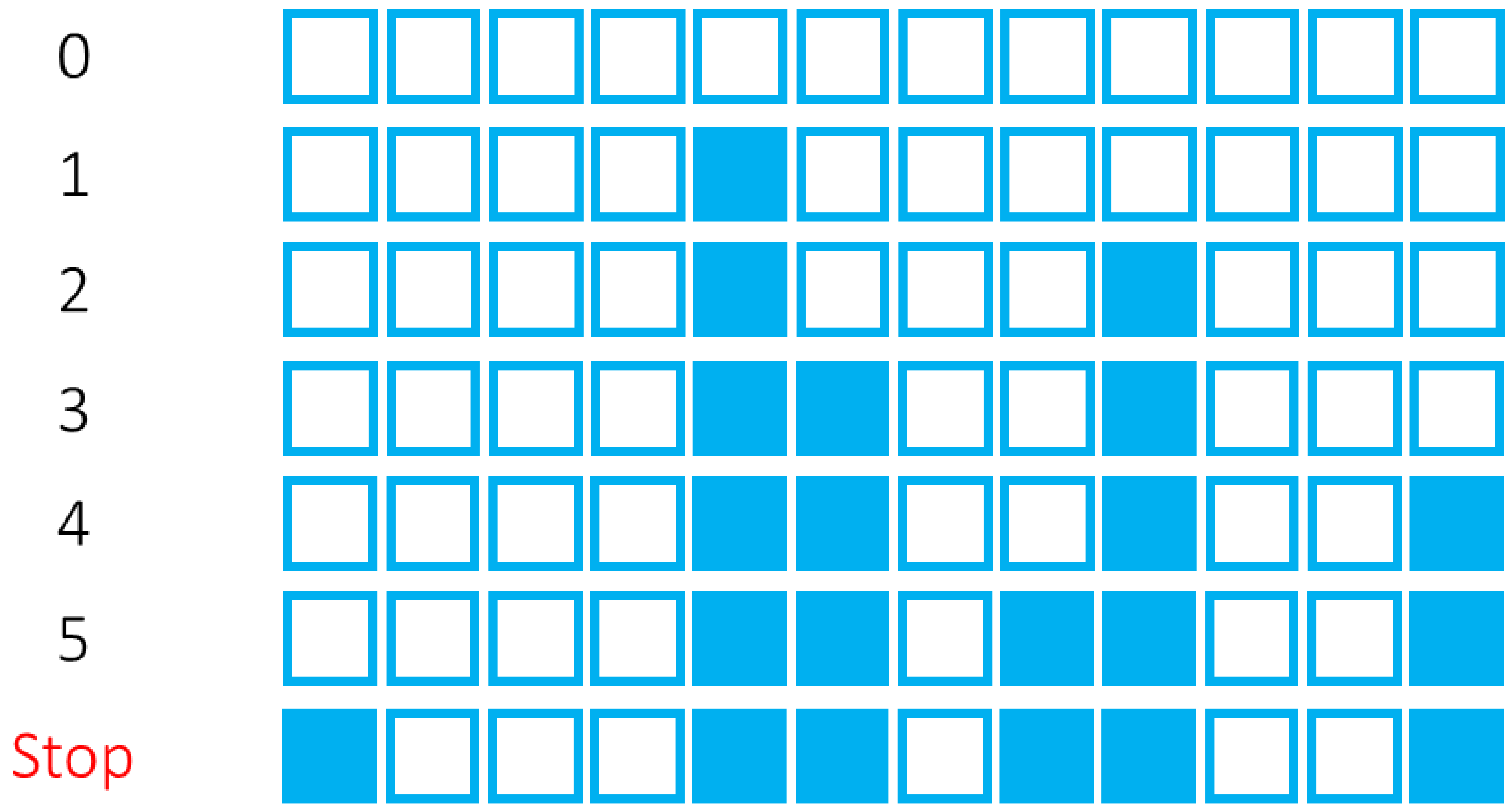
Seleção de Variáveis - Forward



Seleção de Variáveis - Forward



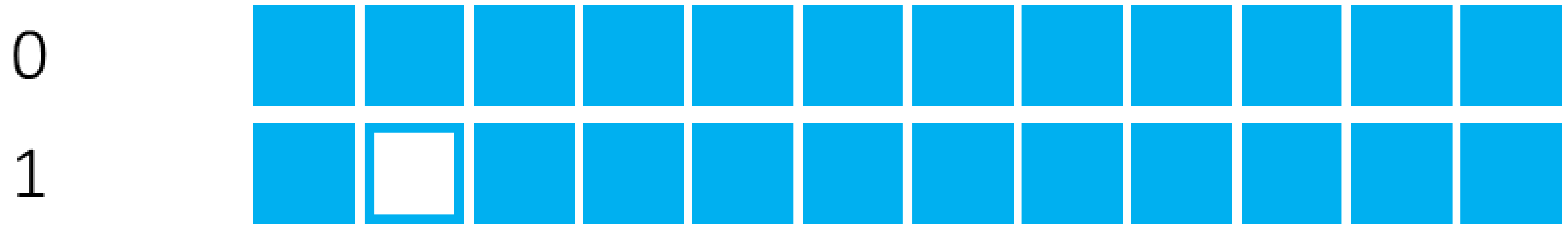
Seleção de Variáveis - Forward



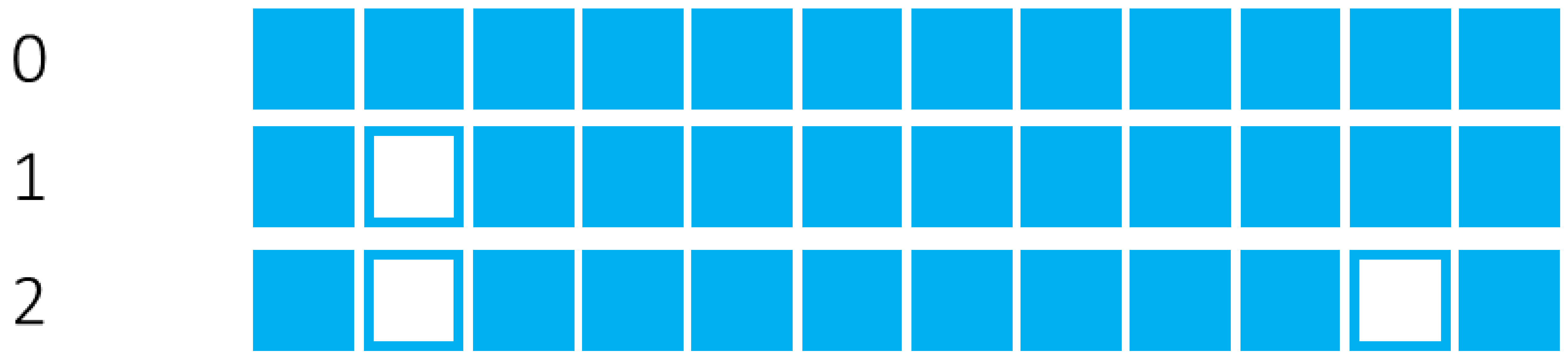
Seleção de Variáveis - Backward

0

Seleção de Variáveis - Backward

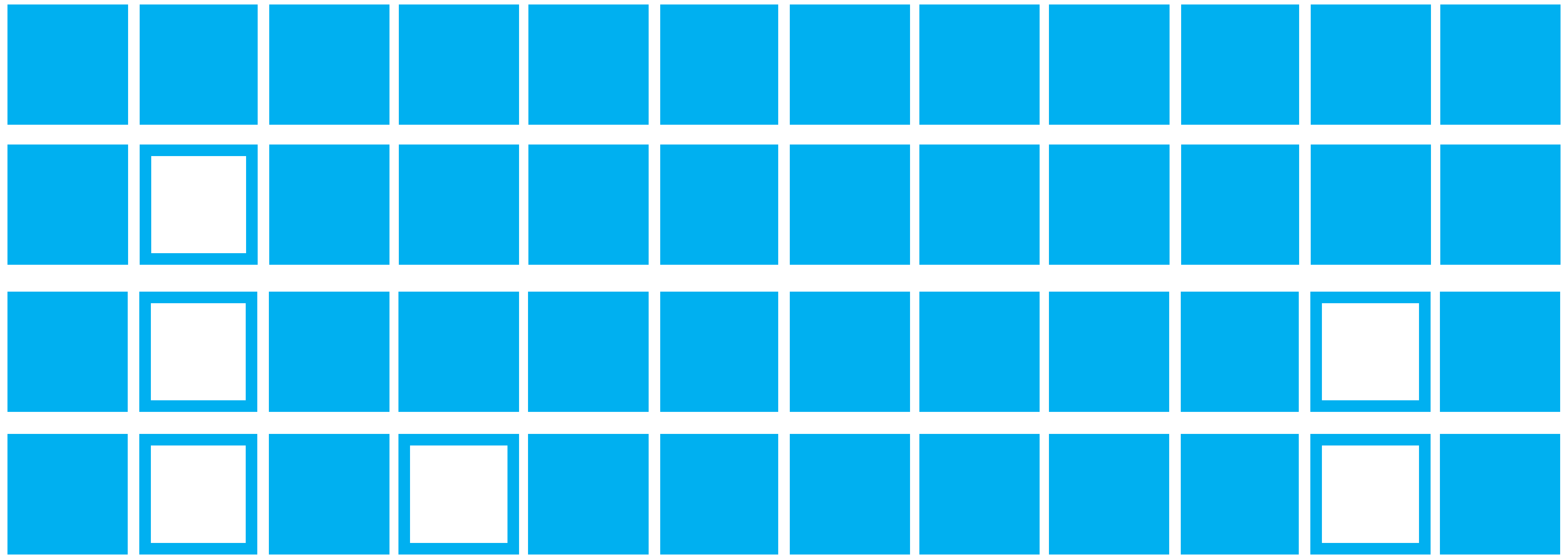


Seleção de Variáveis - Backward

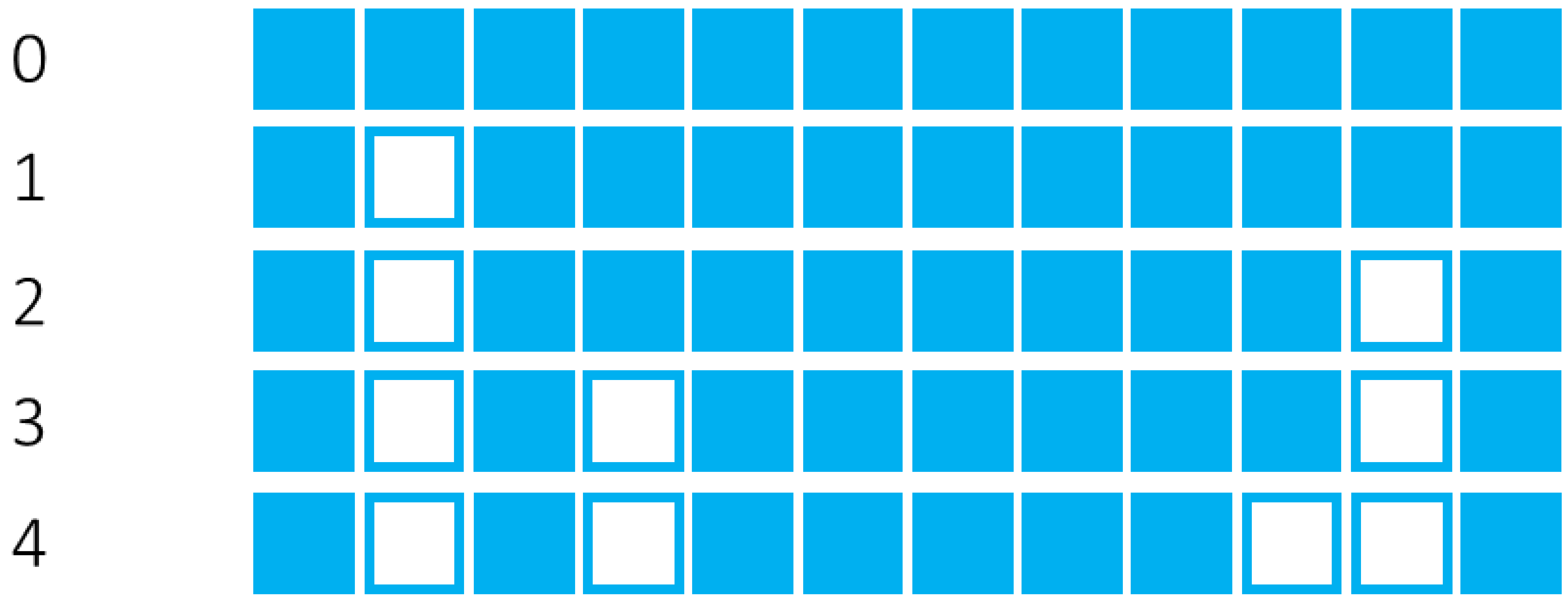


Seleção de Variáveis - Backward

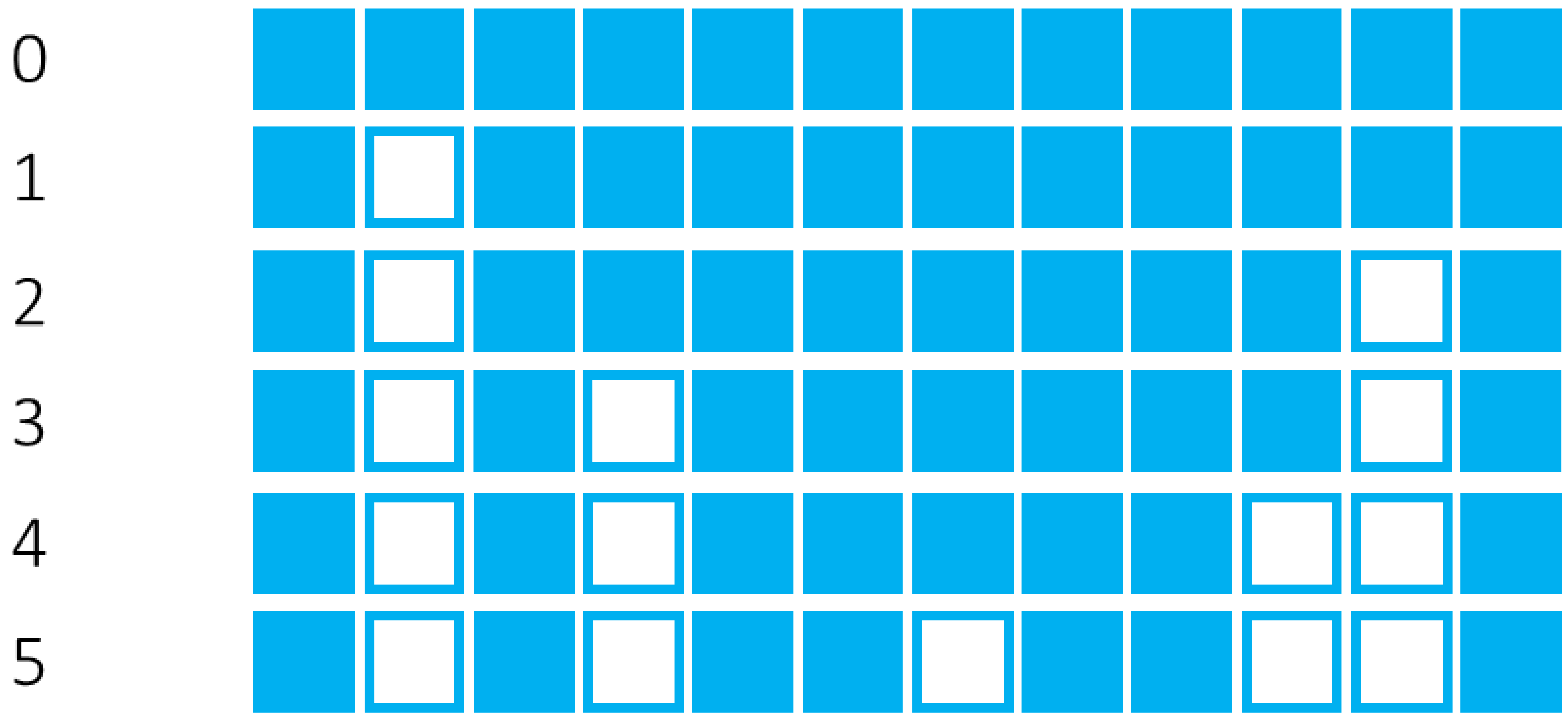
0
1
2
3



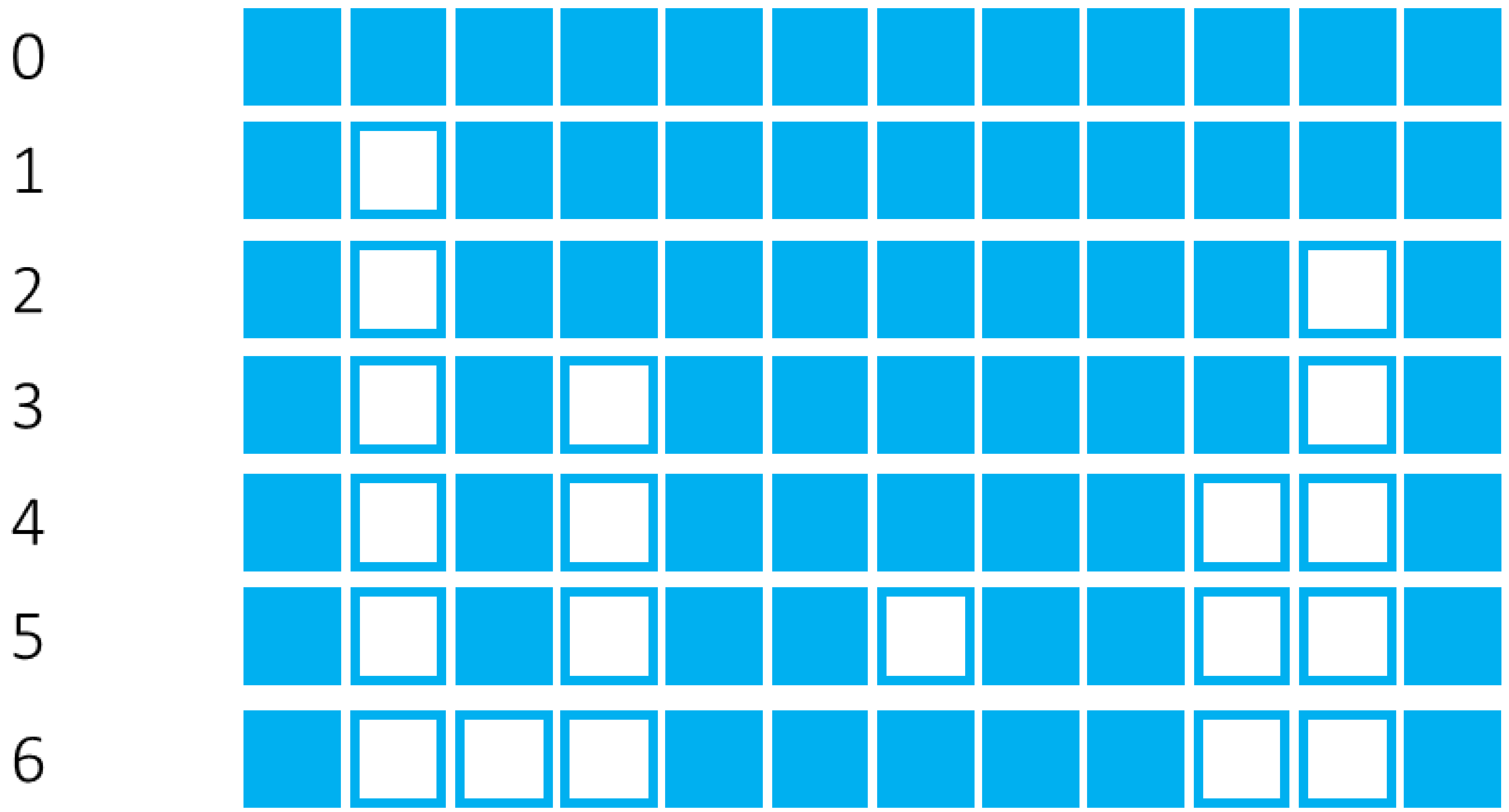
Seleção de Variáveis - Backward



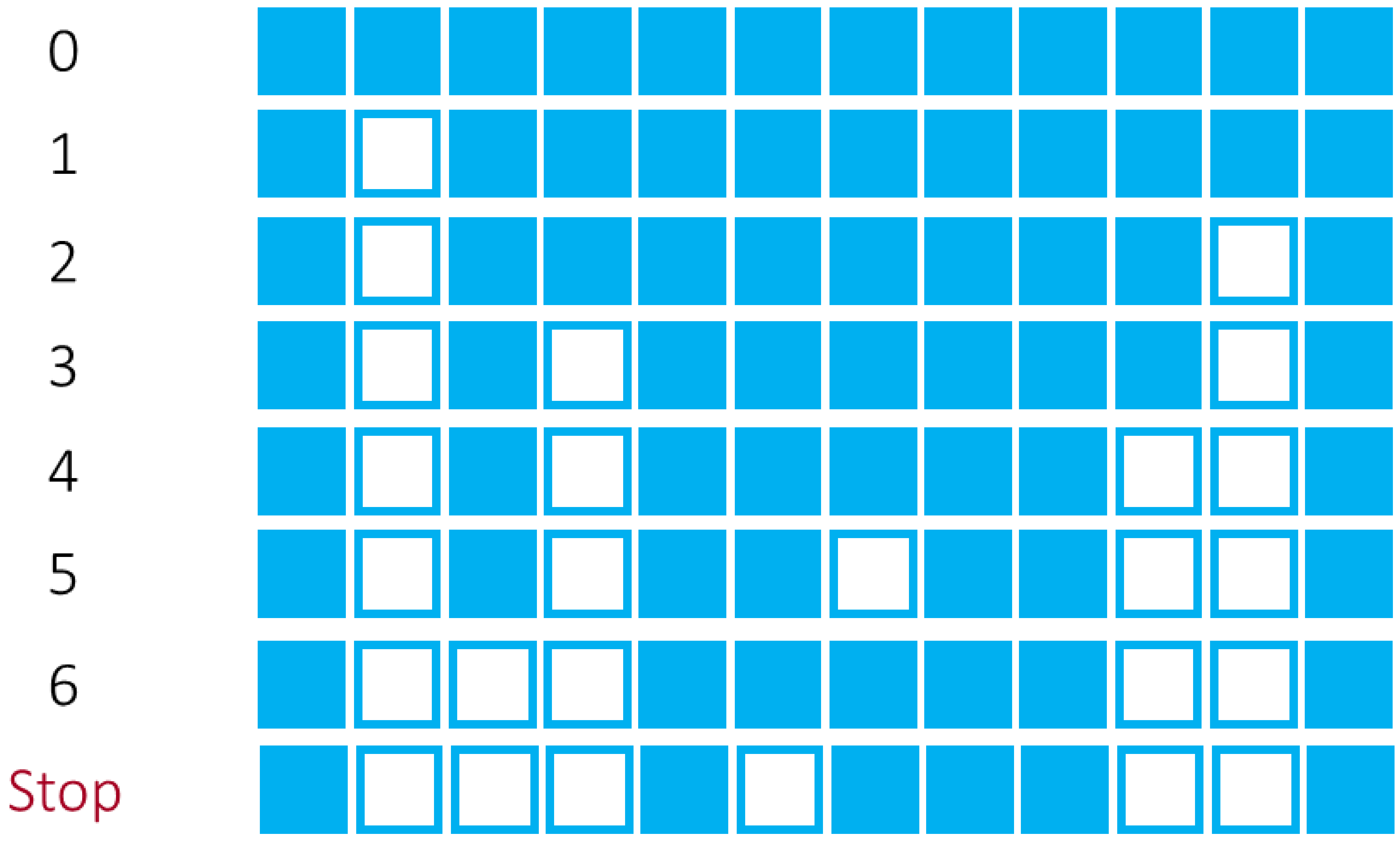
Seleção de Variáveis - Backward



Seleção de Variáveis - Backward



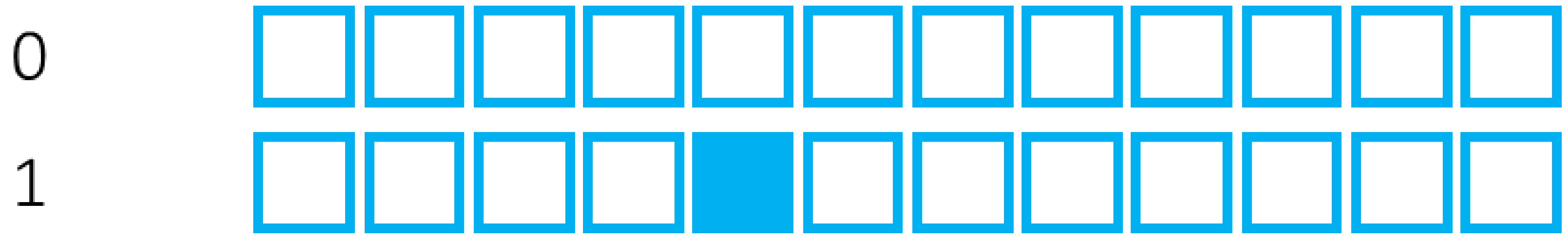
Seleção de Variáveis - Backward



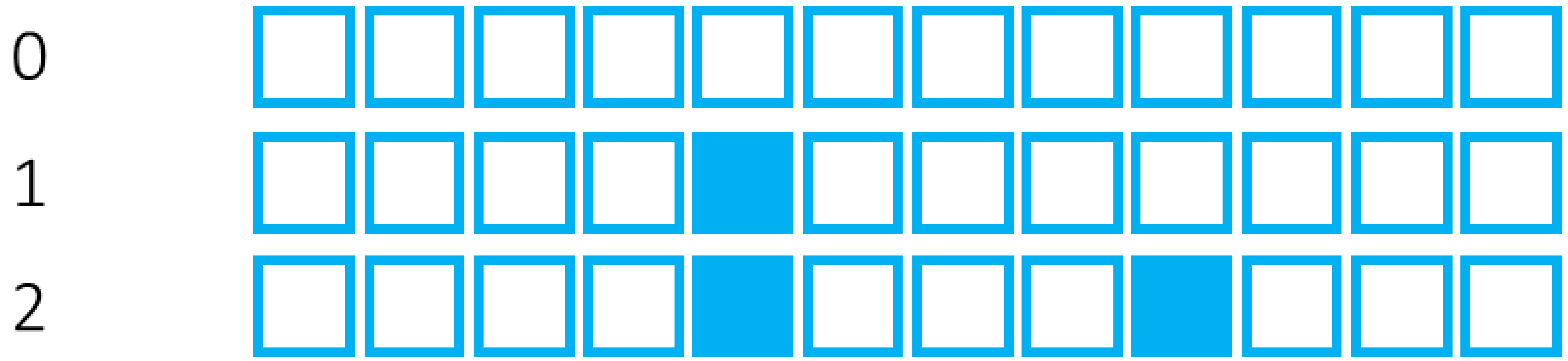
Seleção de Variáveis - Stepwise

0 □□□□□□□□□□□□

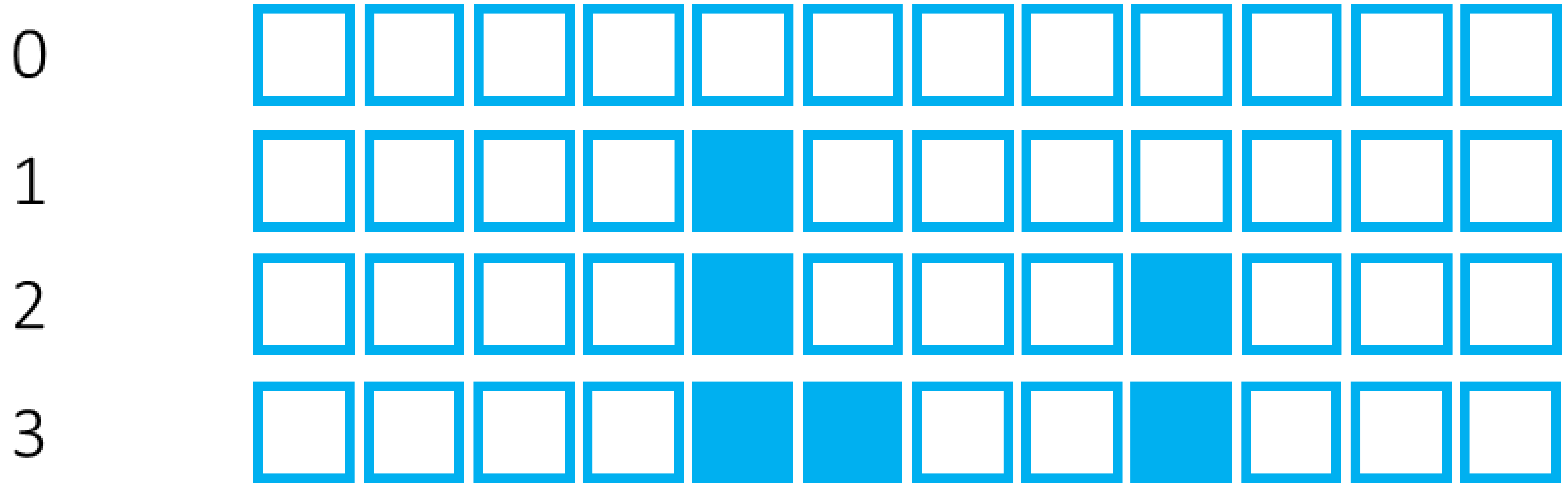
Seleção de Variáveis - Stepwise



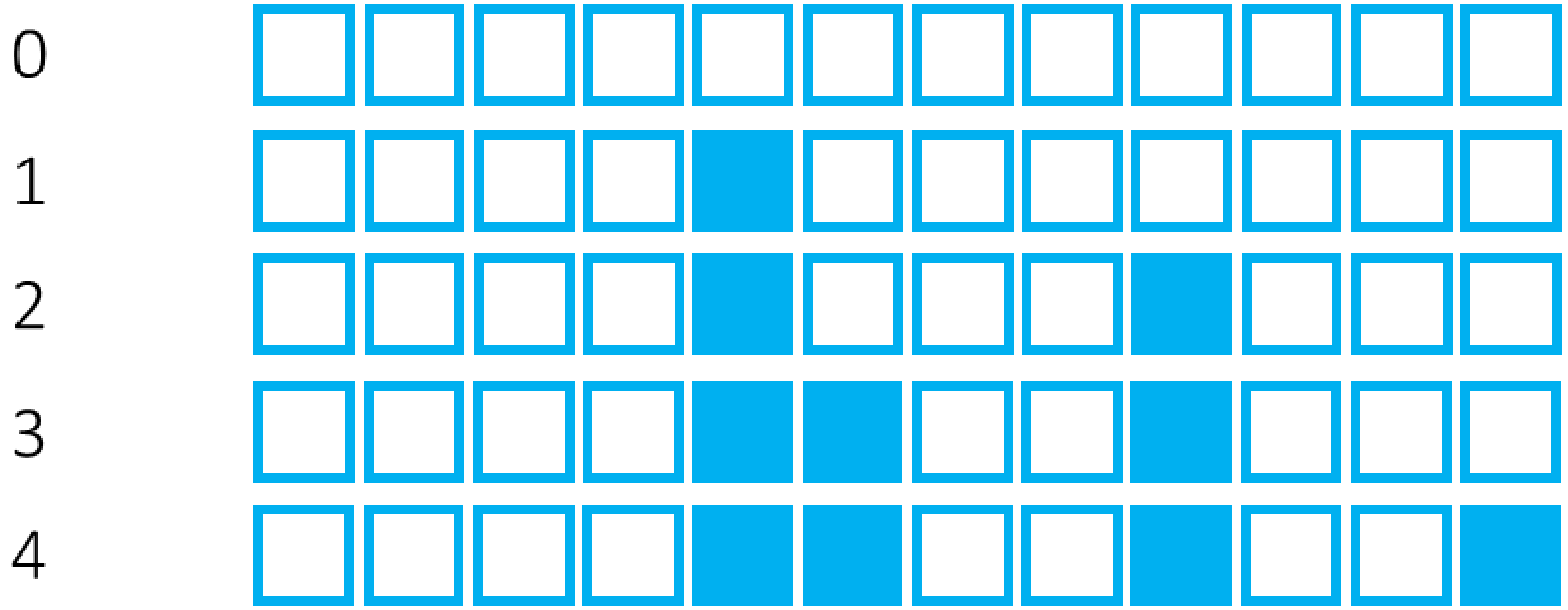
Seleção de Variáveis - Stepwise



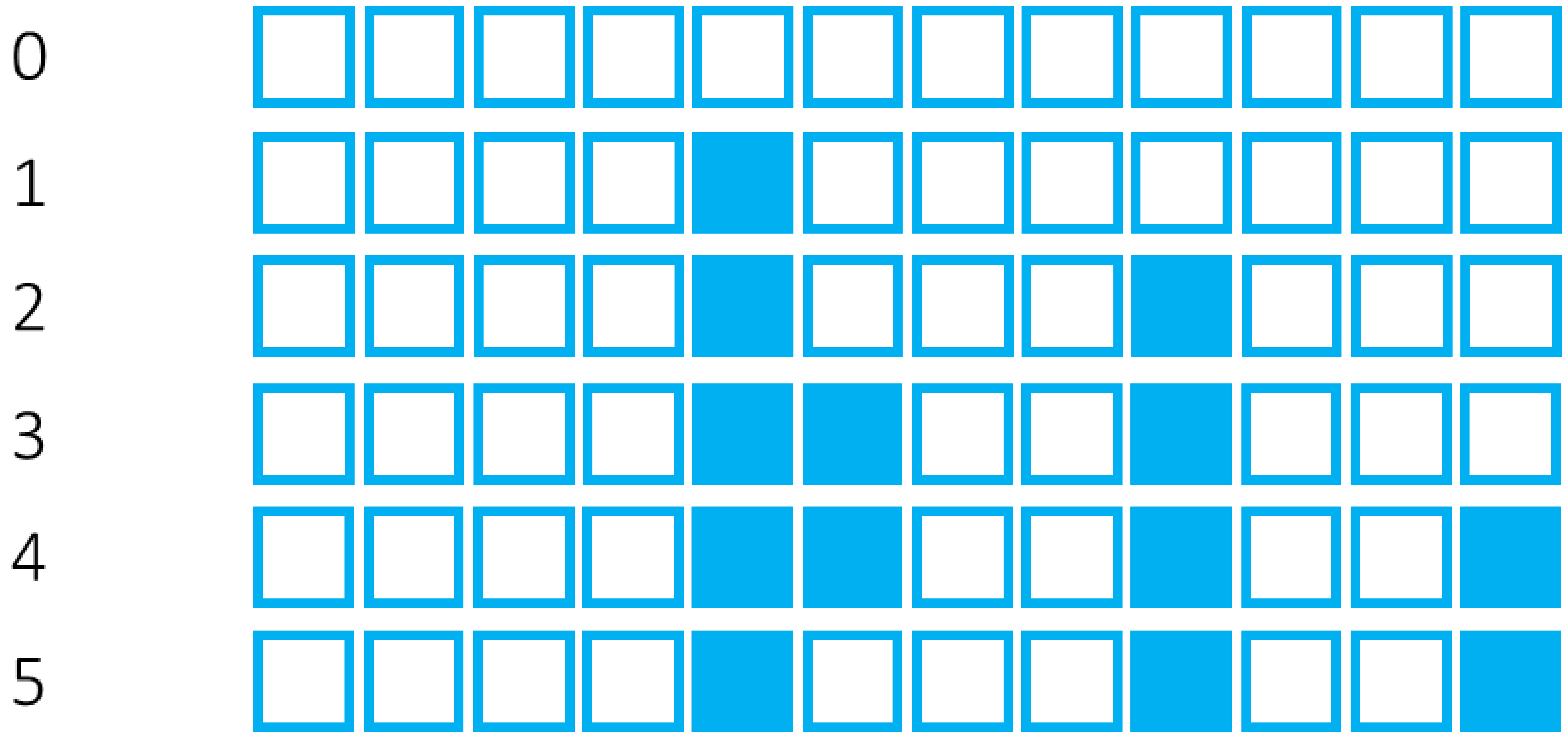
Seleção de Variáveis - Stepwise



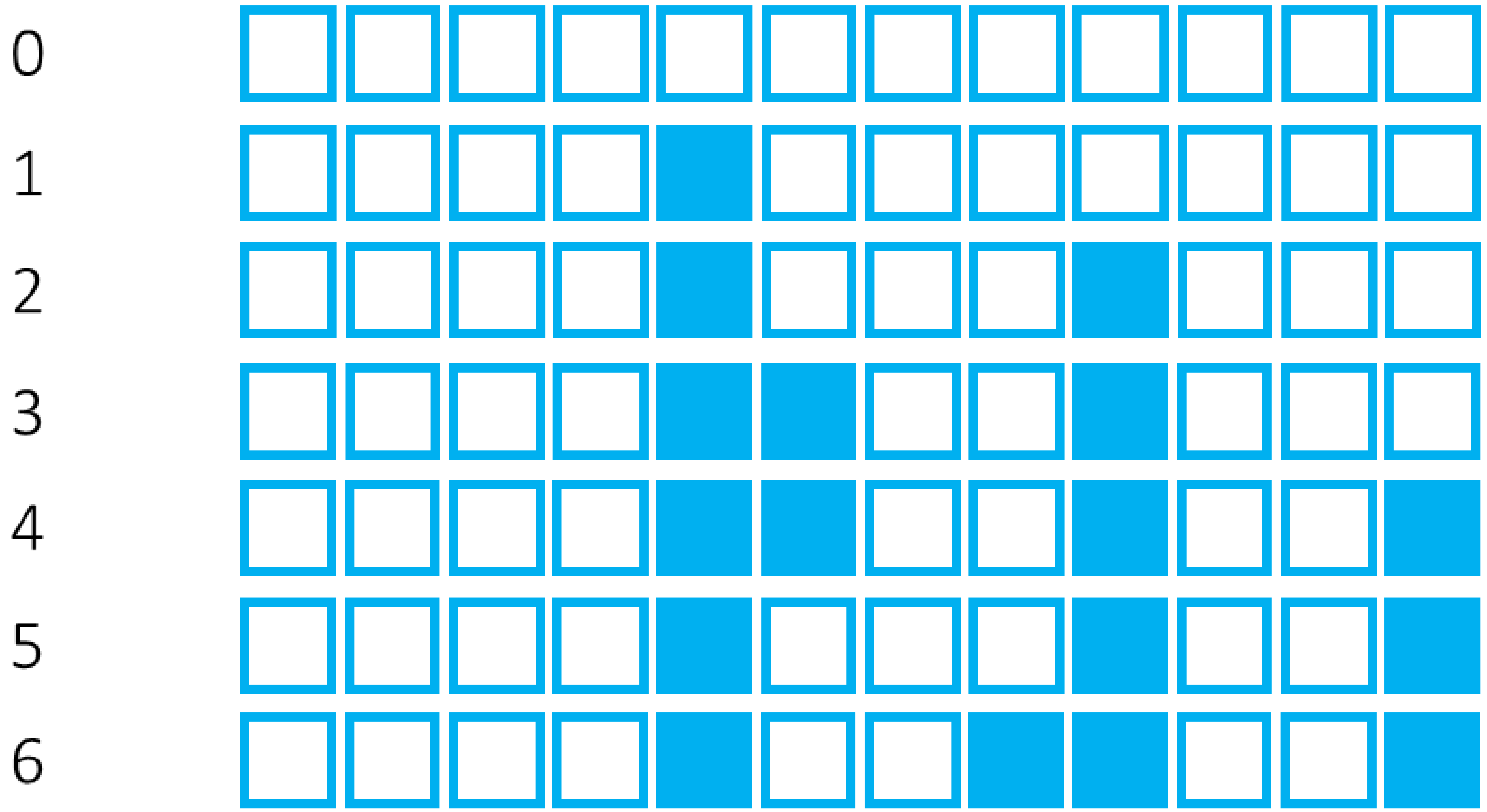
Seleção de Variáveis - Stepwise



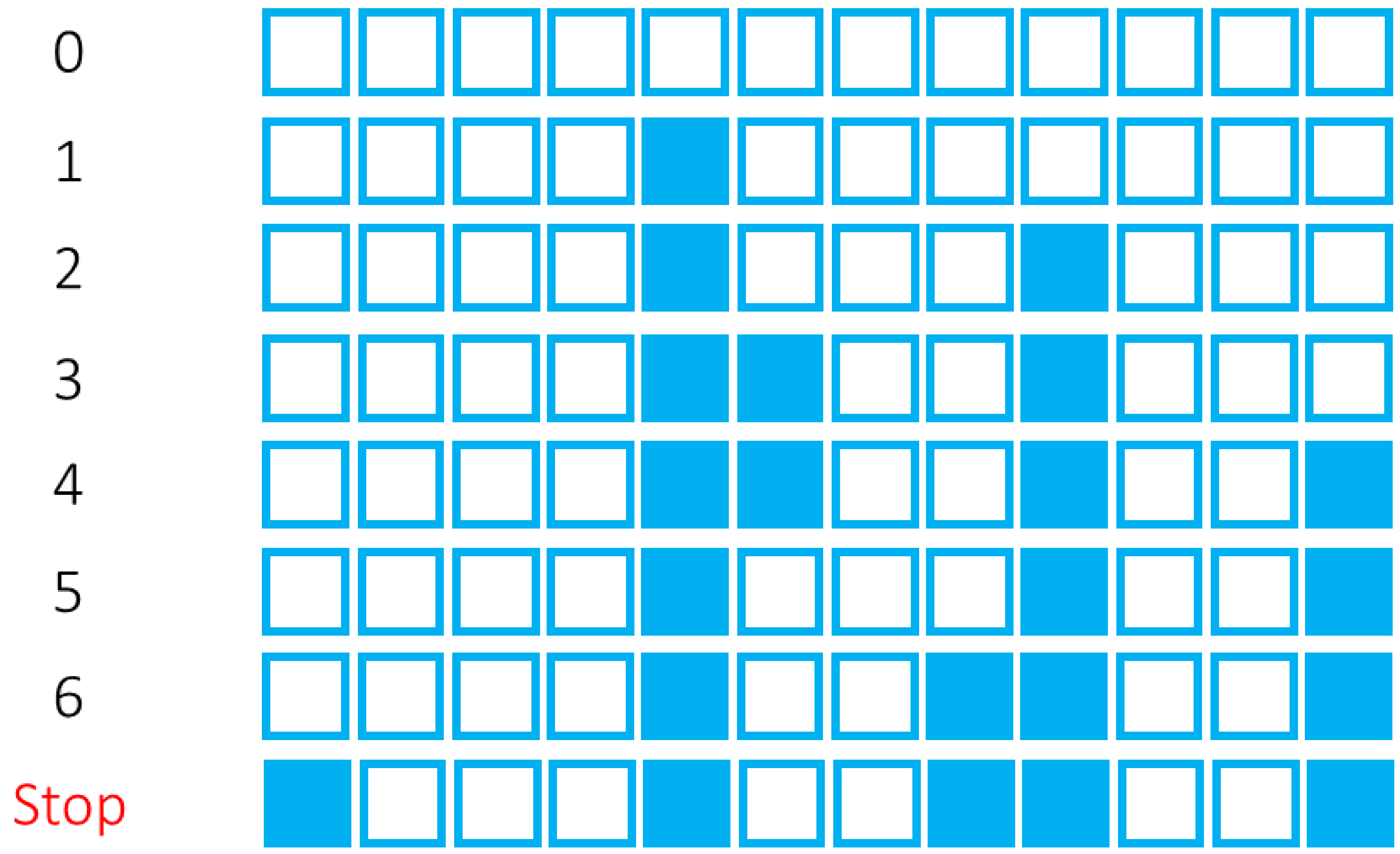
Seleção de Variáveis - Stepwise



Seleção de Variáveis - Stepwise



Seleção de Variáveis - Stepwise



Estudo de Caso
Ajustando um modelo de Regressão Logística no Python

Parte_2 : Seleção de variáveis – Forward

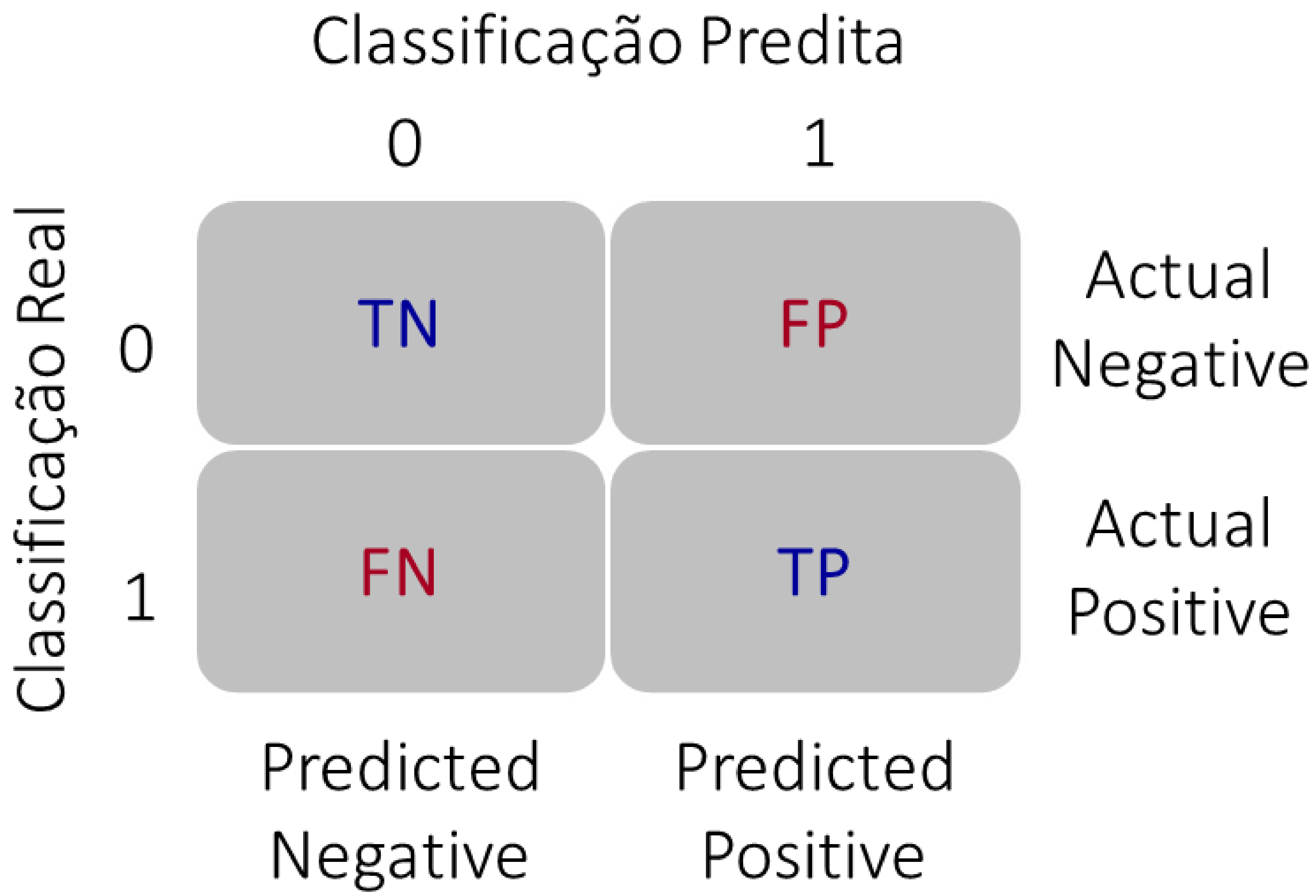
Ajuste do Modelo – Matriz de Confusão

Classificação Predita

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

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

Ajuste do Modelo – Matriz de Confusão



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

- ***Métricas*** para avaliar a qualidade do ajuste do modelo
  - **Missclassification** $= \frac{FP+FN}{Total\ de\ casos}$
  - **Acurácia** $= \frac{TP+TN}{Total\ de\ casos}$

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 Positve
FN: False Negative
FP: False Positive

- ***<u>Métricas</u>*** para avaliar a qualidade do ajuste do modelo

  - $\textbf{Missclassification} = \frac{FP+FN}{Total\ de\ casos}$

  - $\textbf{Acurácia} = \frac{TP+TN}{Total\ de\ casos}$

  - $\textbf{Precision} = \mathrm{P} = \frac{TP}{TP+FP}$

    - Altos valores de precision estão relacionados a baixa taxa de FP

  - $\textbf{Recall} = \mathrm{R} = \frac{TP}{TP+FN}$

    - Altos valores de recall estão relacionados a baixa taxa de FN

Ajuste do Modelo – Matriz de Confusão

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

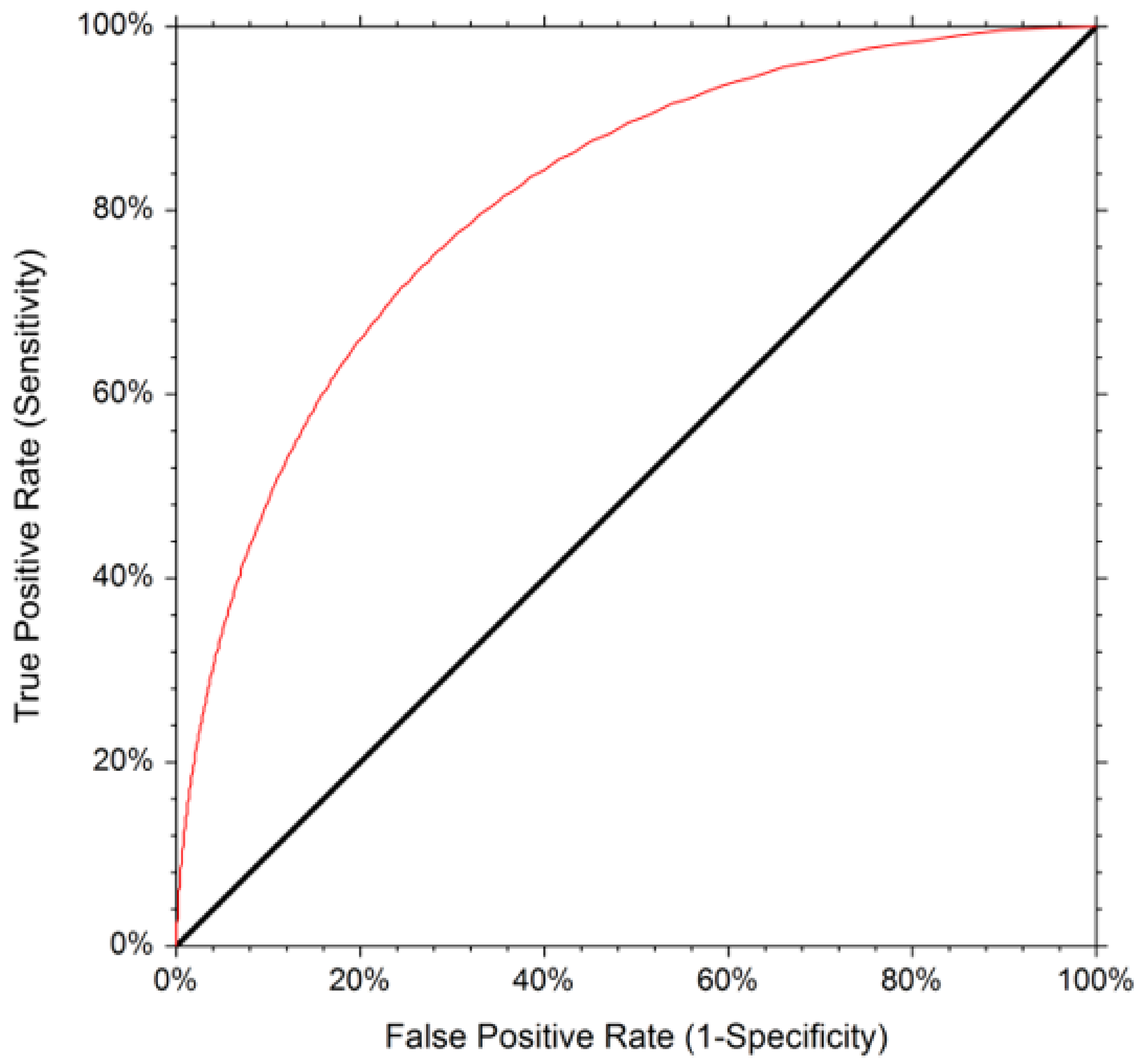
TN: True Negative
TP: True Positve
FN: False Negative
FP: False Positive

- ***Métricas*** para avaliar a qualidade do ajuste do modelo
  - **Missclassification** $= \dfrac{FP+FN}{Total\ de\ casos}$
  - **Acurácia** $= \dfrac{TP+TN}{Total\ de\ casos}$
  - **Precision** $= \mathrm{P} = \dfrac{TP}{TP+FP}$
    - Altos valores de precision estão relacionados a baixa taxa de FP
  - **Recall** $= \mathrm{R} = \dfrac{TP}{TP+FN}$
    - 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).

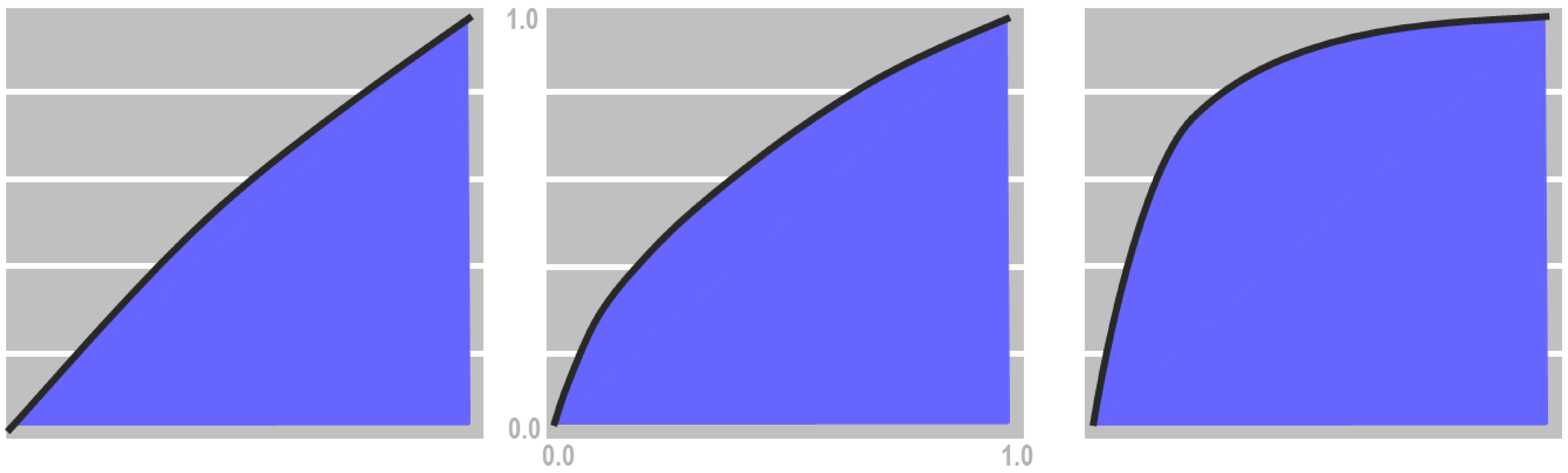


- ***Métricas***

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

Ajuste do Modelo – Curva ROC

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

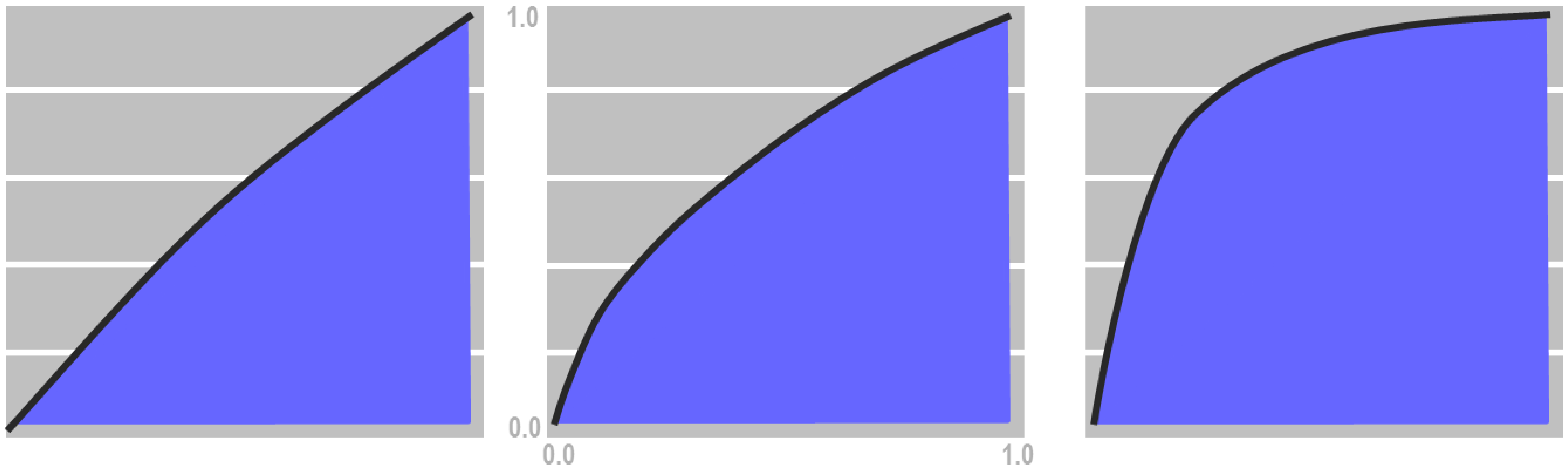


Modelo Fraco
ROC Index < 0.6*

Modelo Forte
ROC Index > 0.7*

Ajuste do Modelo – Curva ROC

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



Modelo Fraco
ROC Index < 0.6*

Modelo Forte
ROC Index > 0.7*

* 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_3

- **Ajustar um modelo de regressão Logística na base de treinamento usando sklearn**
- **Validar o modelo na base de teste usando: AUC, precision e recall**

Estudo de Caso
Ajustando um modelo de Regressão Logística no Python

Parte_4

- **Ajustar um modelo de regressão Logística na base de treinamento usando statsmodel**
- **Validar o modelo na base de teste usando: AUC, precision e recall**

Desafio
Ajustar um modelo de Regressão Logística no Python

1. Tratar as Variáveis: Missing, Categoricas, redundância e irrelevância
2. Dividir a base em treinamento e teste
3. Seleção de variáveis
4. Ajustar um modelo de regressão Logística
5. Prever na base de teste
6. Avaliar a qualidade do ajuste do modelo: AUC, precision e 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/

Obrigada

Cristiane Rodrigues
crisrodrigues_27@hotmail.com