Fundamentos de Aprendizagem Estatística

Tiago Mendonça dos Santos

tiagoms.com
tiagomendonca
tiagoms1@insper.edu.br

Prática (1)

A informação sistemática de X sobre Y será dada pela função de regressão não linear definida por:

$$g(x) = 45 \cdot \tanh\left(\frac{x}{1,9} - 7\right) + 57.$$

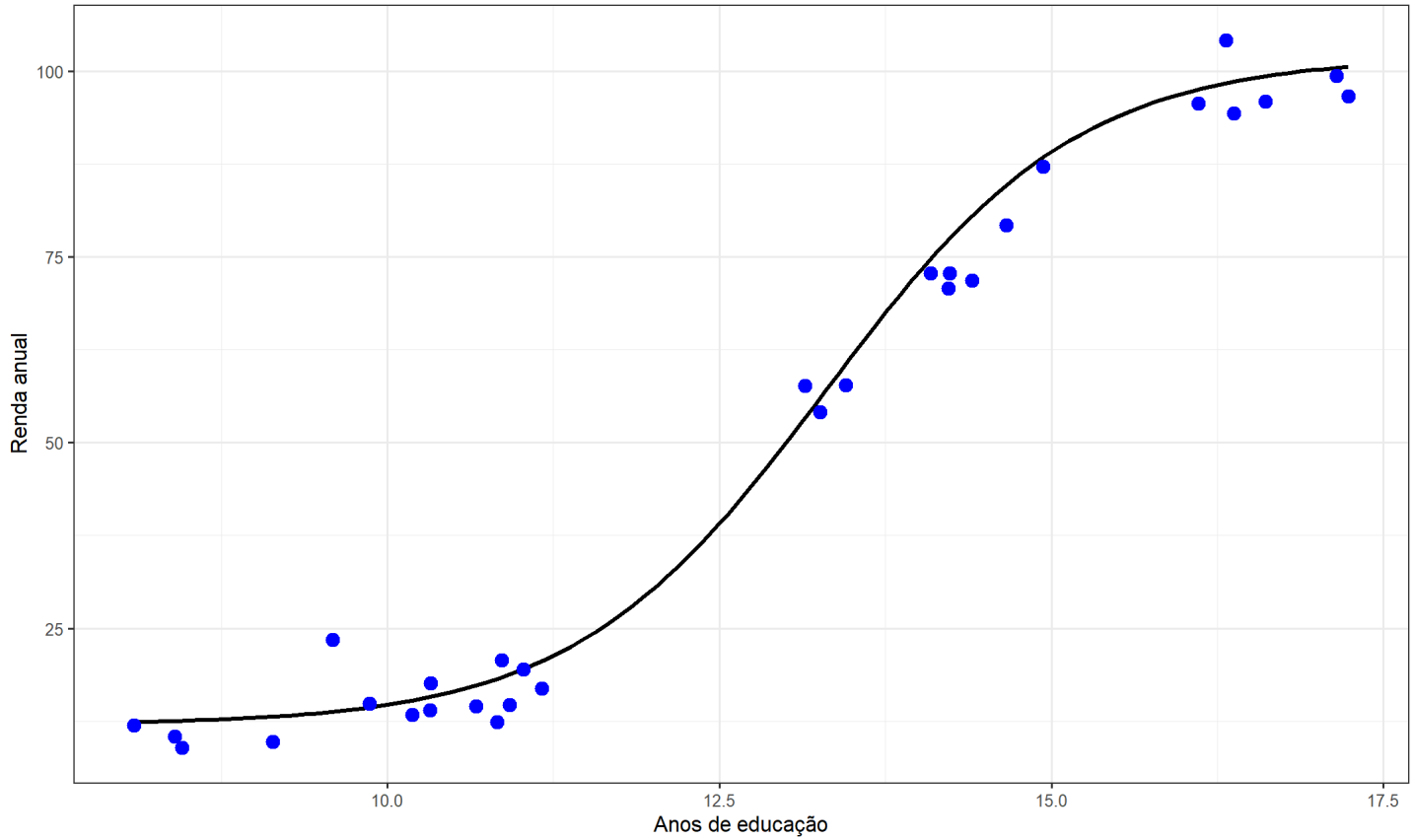
Para a preditora X utilizaremos uma distribuição uniforme entre 8 e 18 anos de estudo.

O erro aleatório terá distribuição normal com esperança 0 e desvio padrão igual a 4.

Note que estamos em uma situação particular na qual conhecemos $g(\cdot)$.

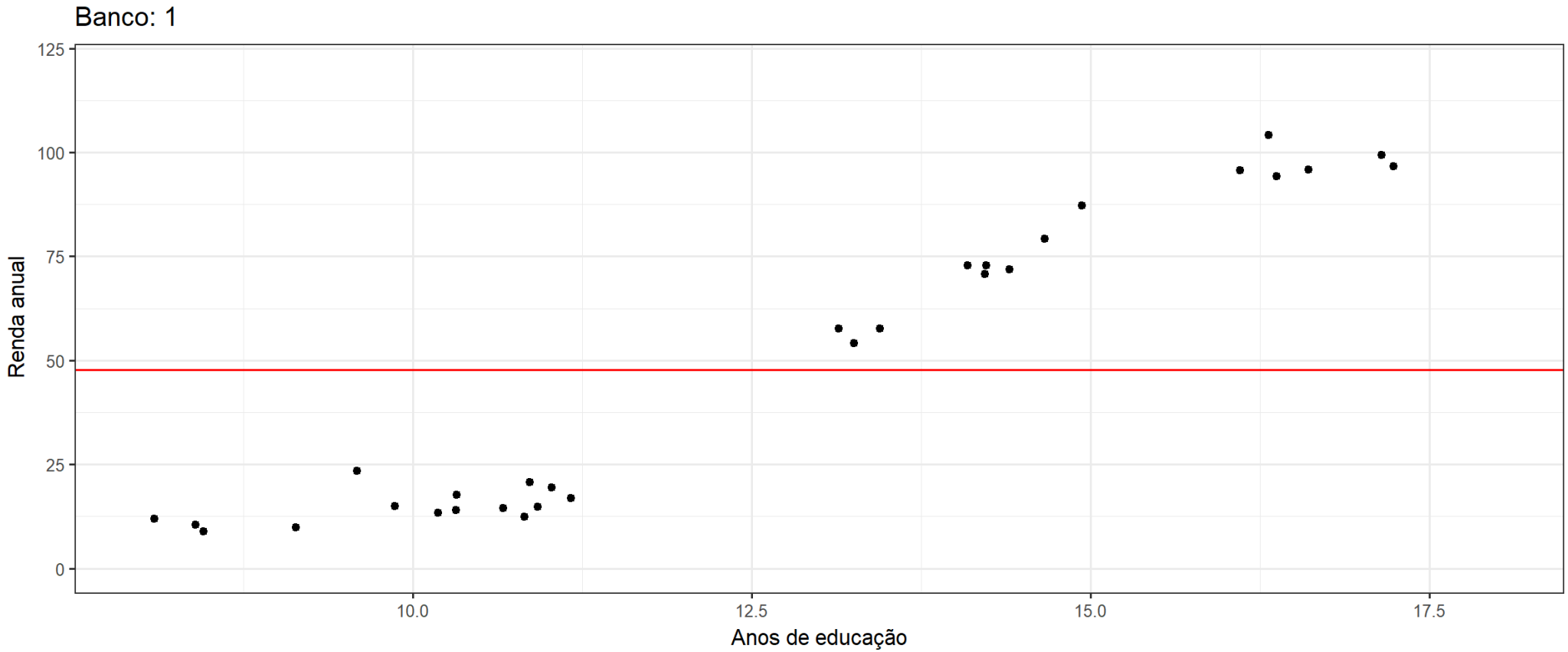
Prática (2)

Na figura a seguir temos o gráfico da função de regressão g e de 30 pares (x_i, y_i) de dados de treinamento gerados pelo modelo aditivo especificado.



Viés

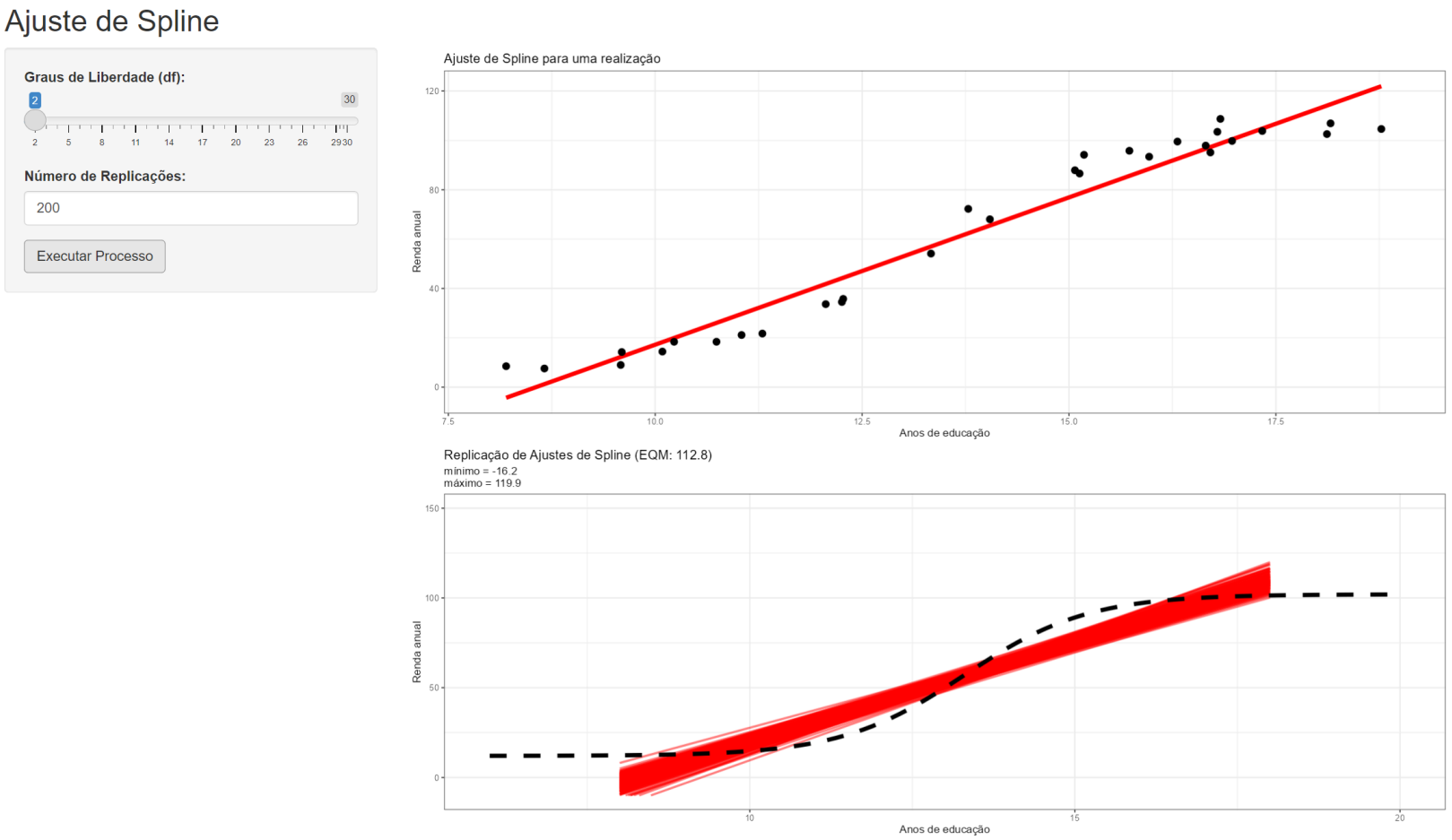
Considerando a geração de 5.000 modelos, para $\hat{g}(x) = \frac{\sum_{i=1}^{30} Y_i}{30}$ temos



Lembre-se que $\text{Viés}[\hat{g}(x)] = \text{E}(\hat{g}(x)) - g(x)$.

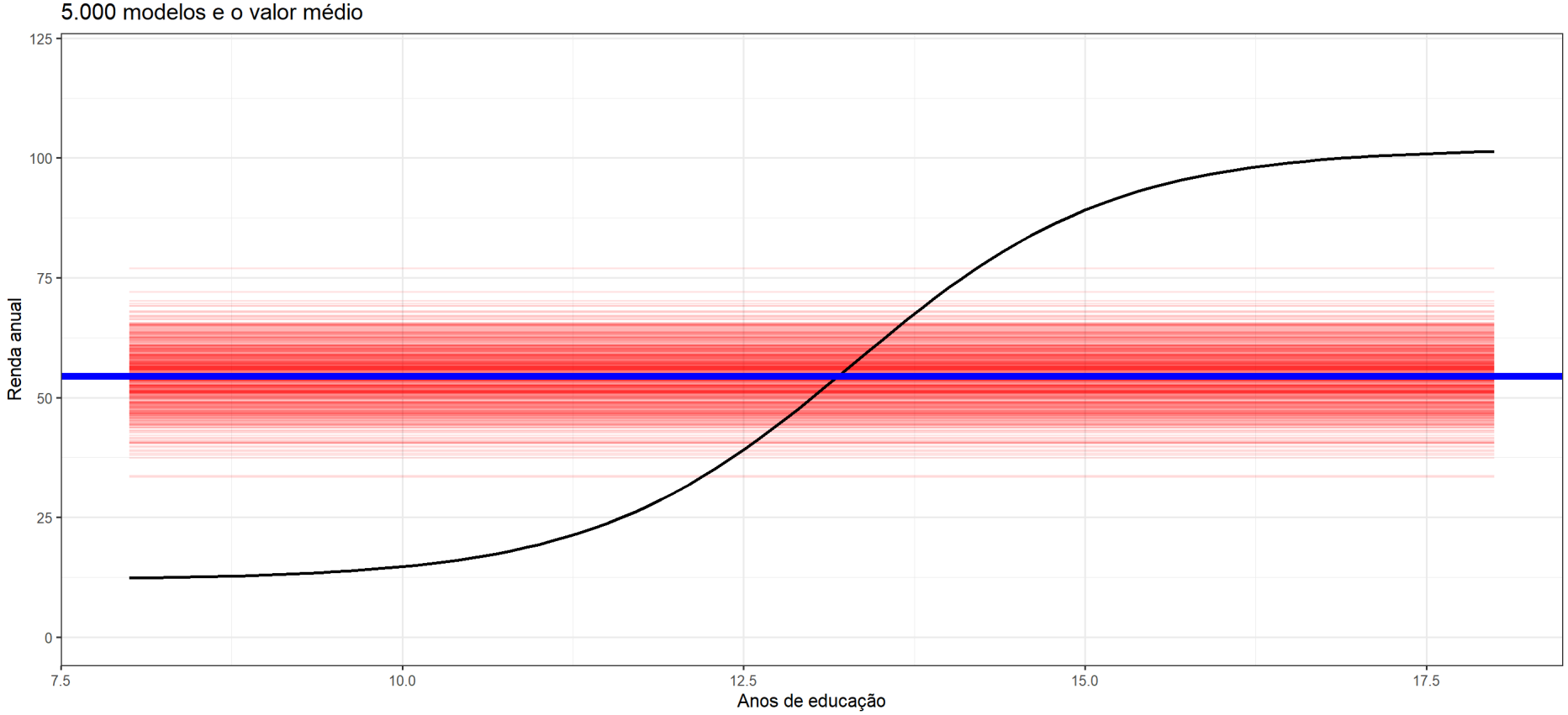
Simulação

Execute algumas simulações em tiagomendonca.github.io/exe_01.



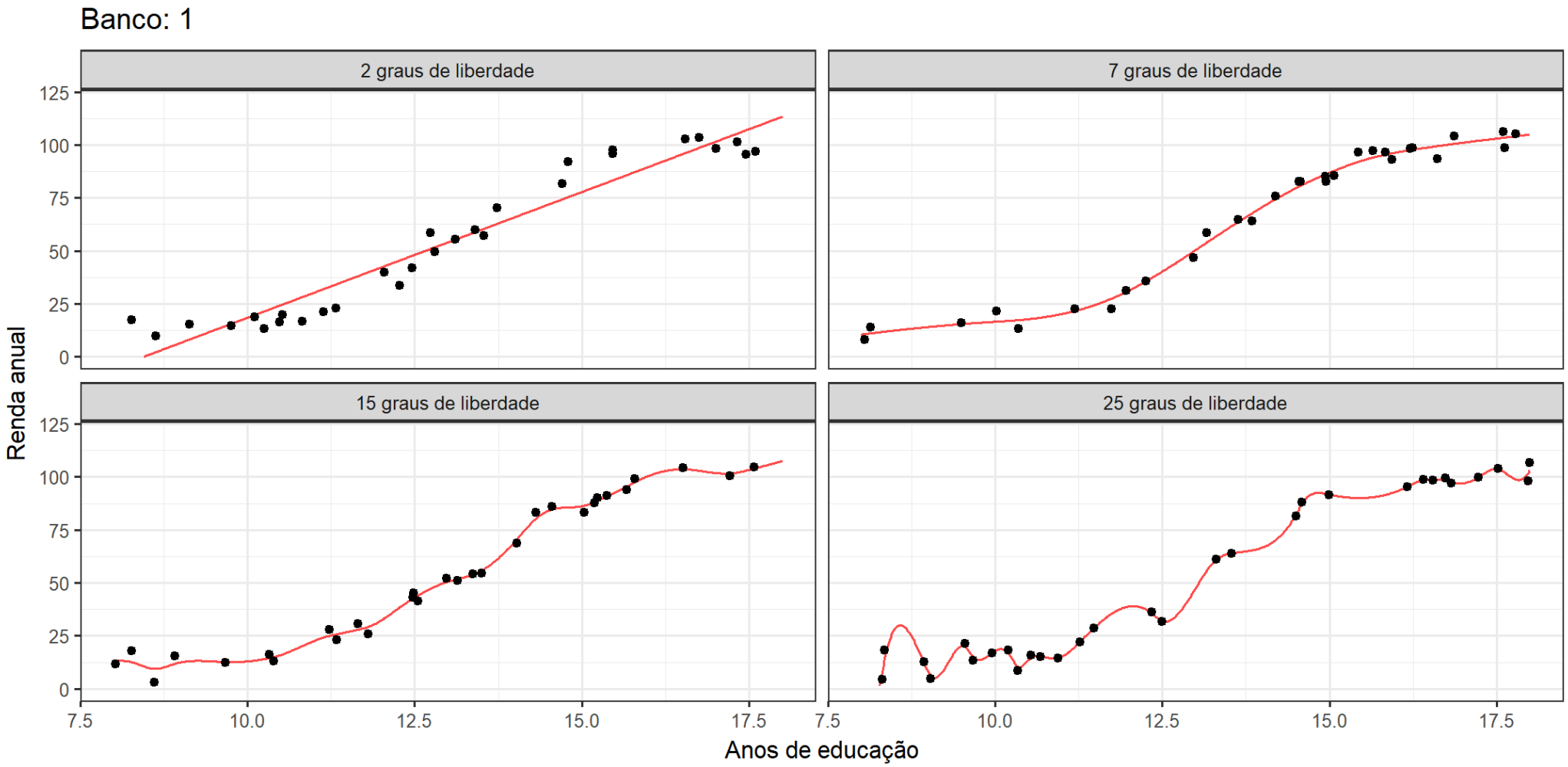
Viés

Considerando a geração de 5.000 modelos, para $\hat{g}(x) = \frac{\sum_{i=1}^{30} Y_i}{30}$ temos

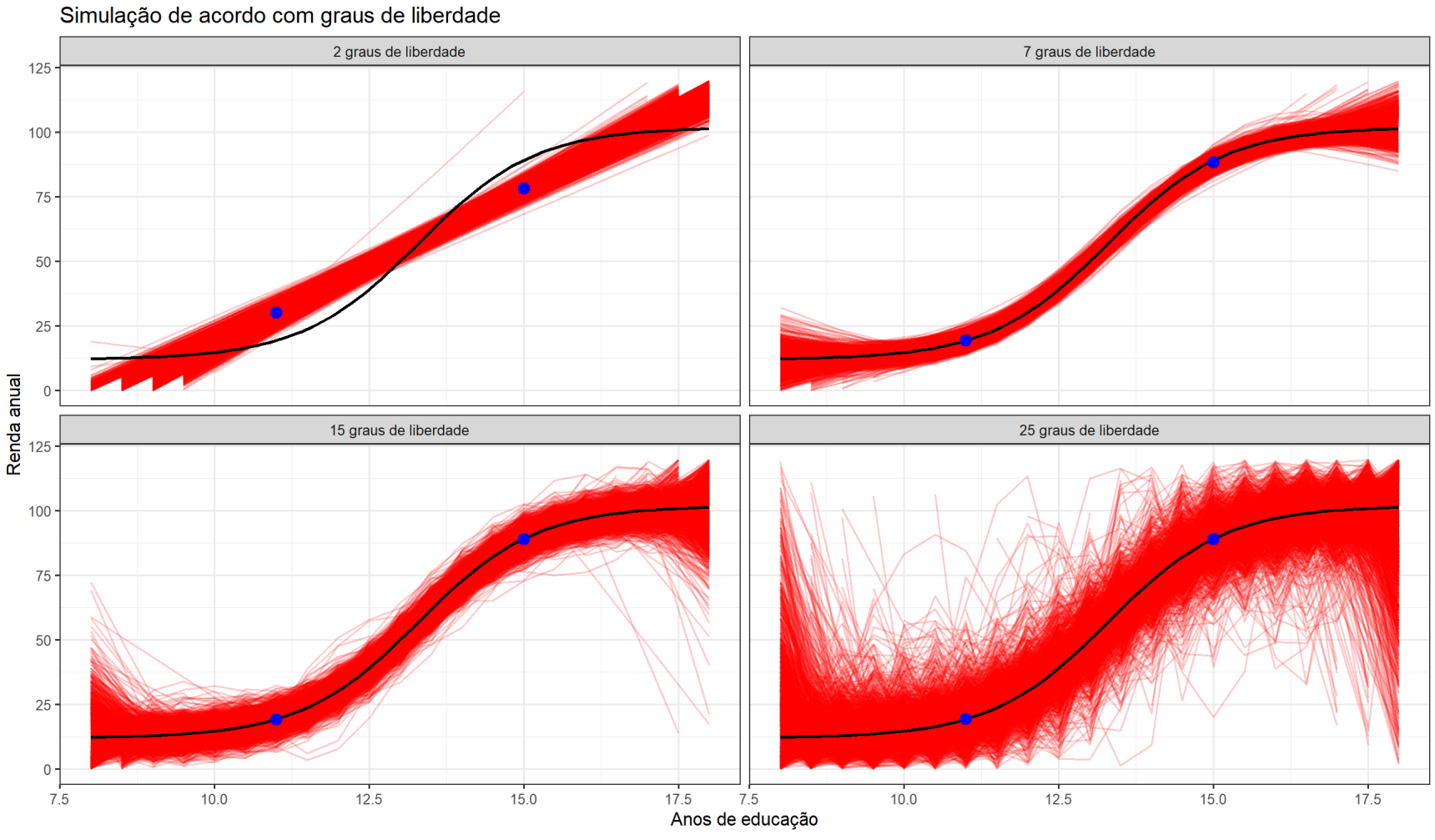


Lembre-se que $\text{Viés}[\hat{g}(x)] = \text{E}(\hat{g}(x)) - g(x)$.

Complexidade

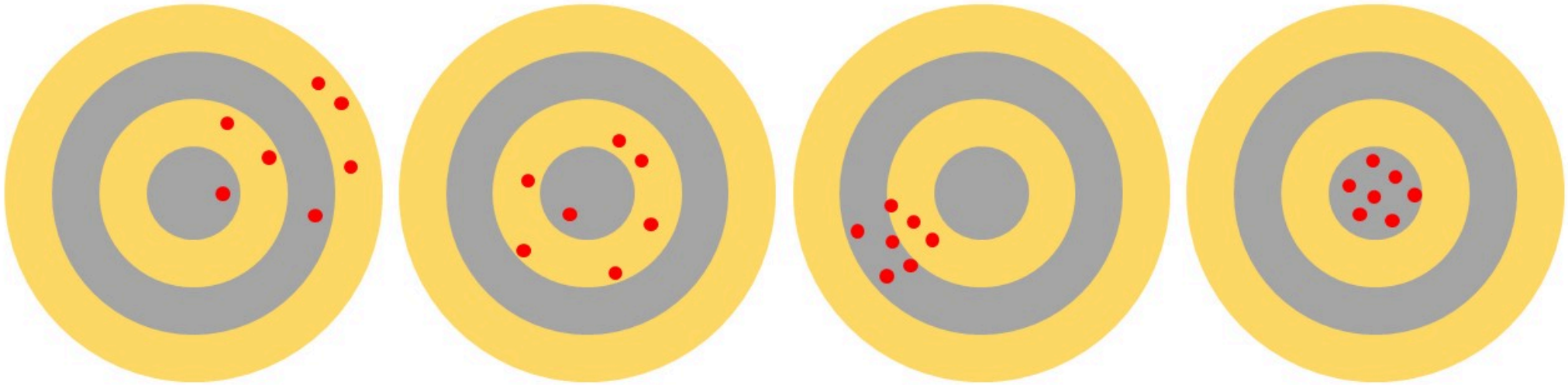


Complexidade

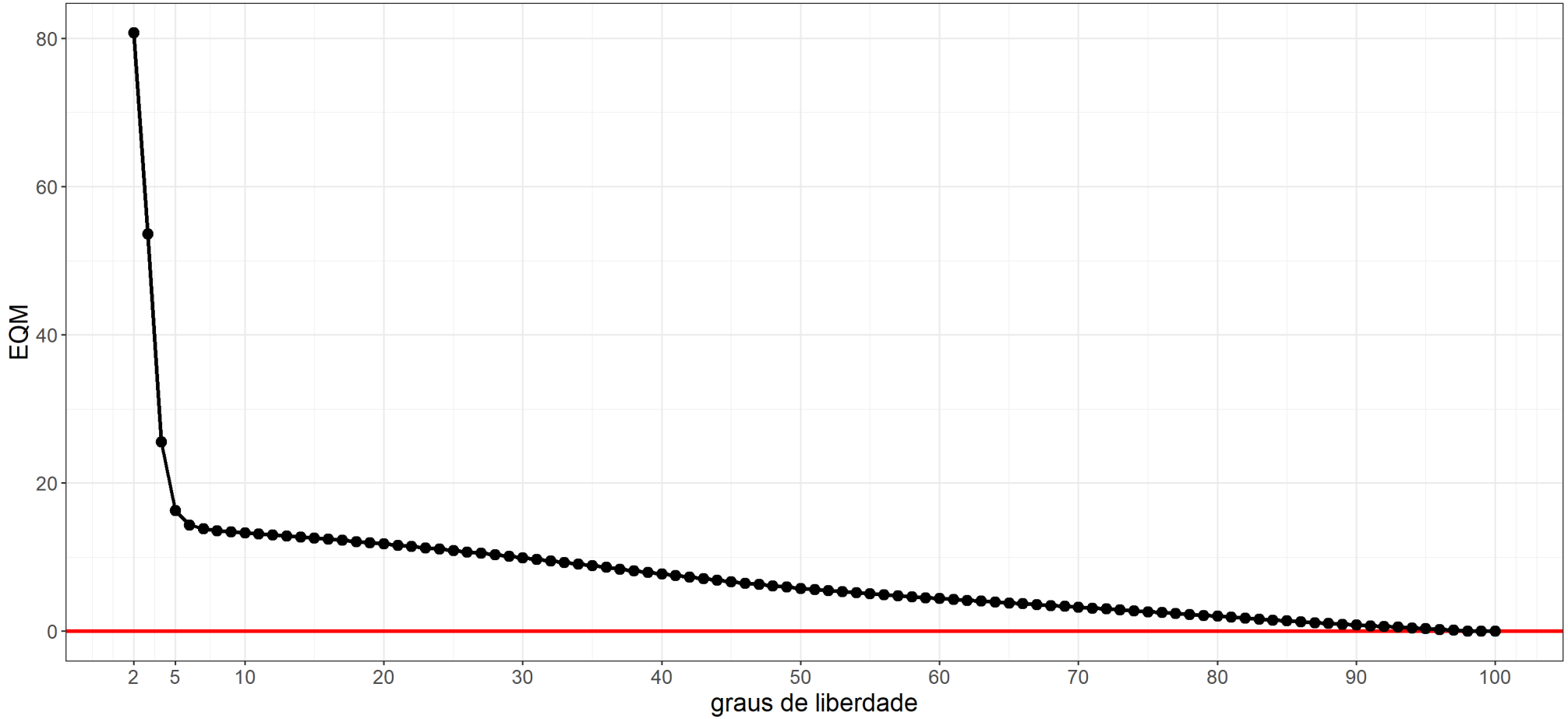


Revisão (2)

Como esses esquemas se relacionam, em termos de viés e variância, com o que vimos no modelo anterior?

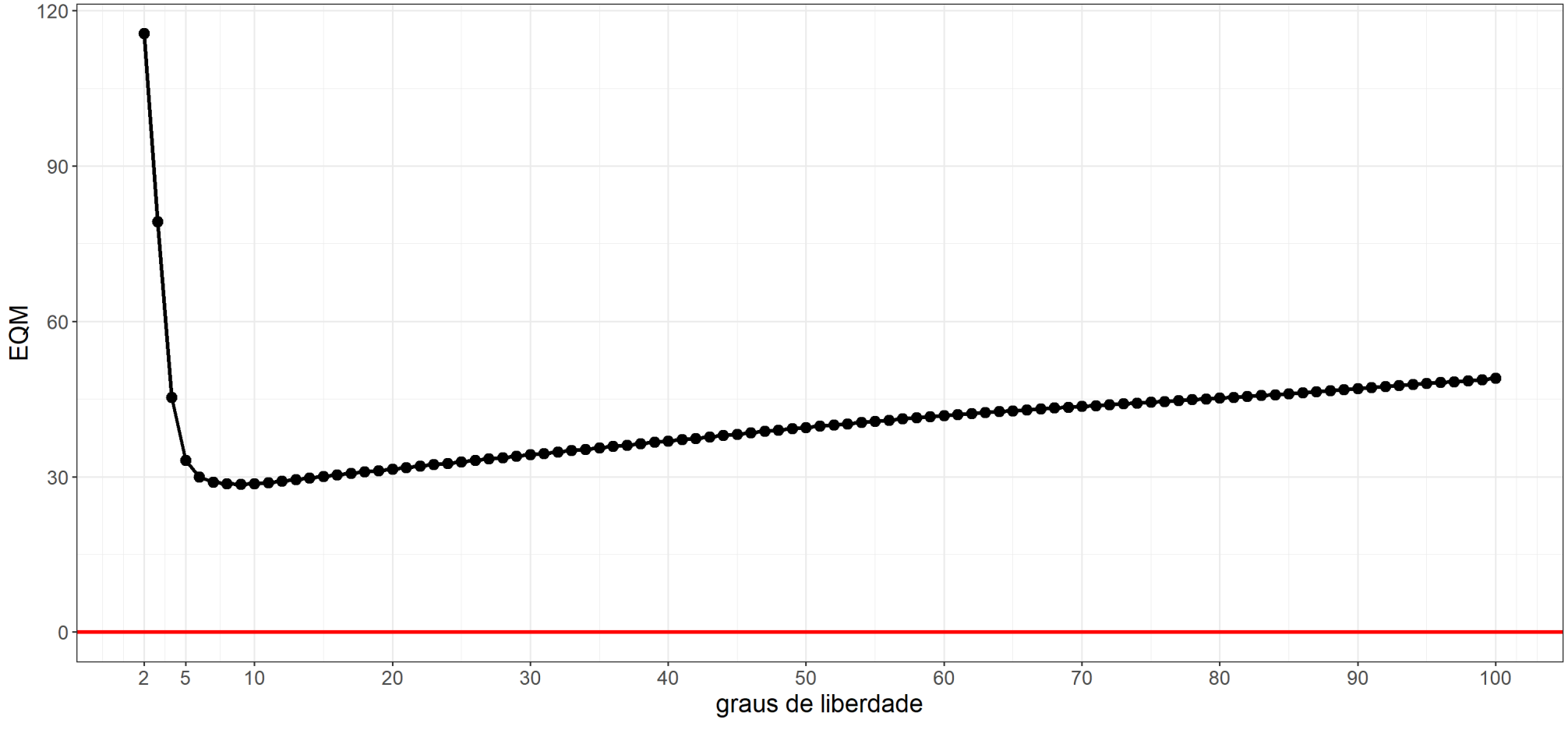


Avaliação do erro

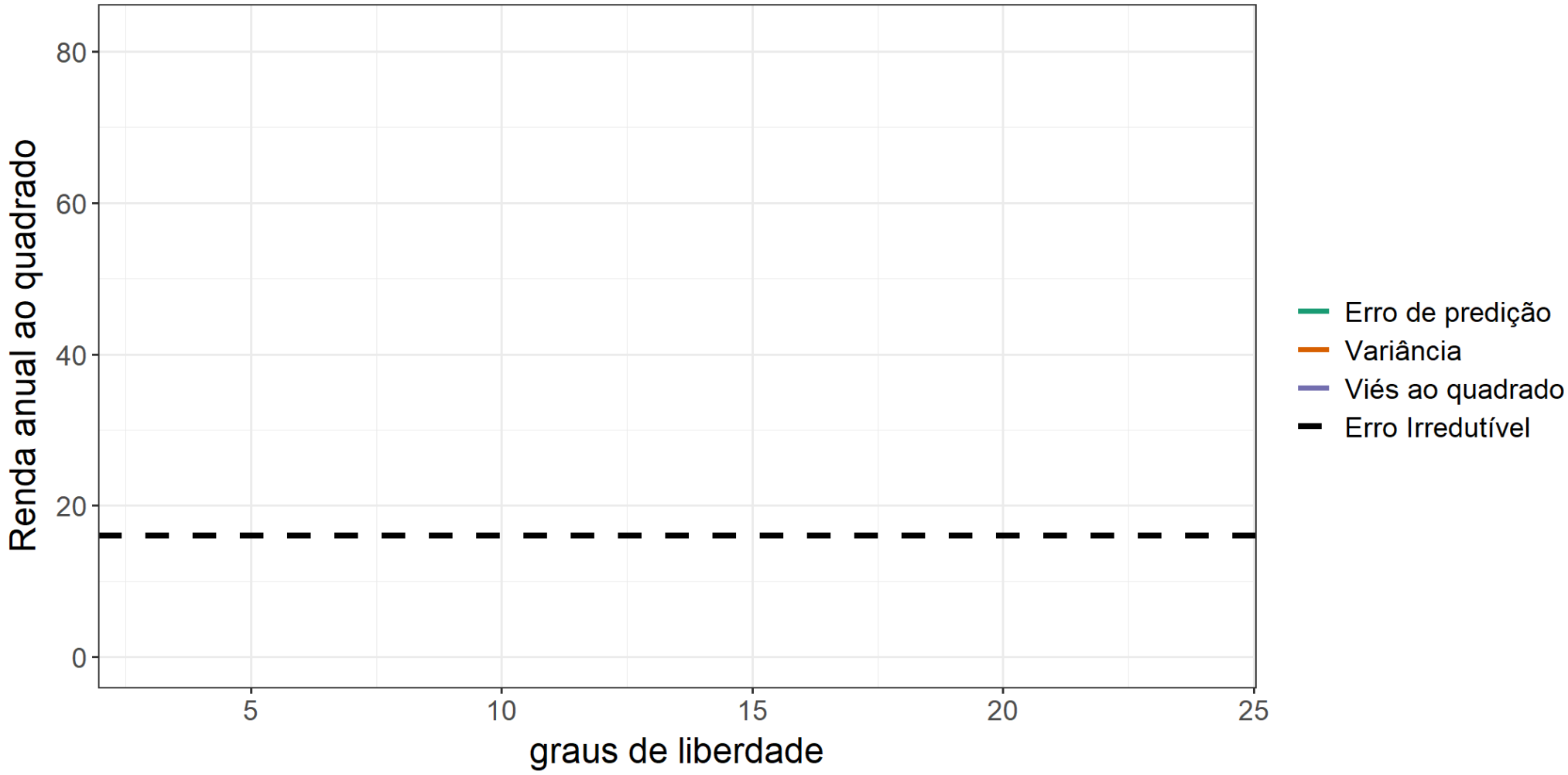


Avaliação do erro

Agora considere as mesmas variações de modelo, mas calculando o EQM a partir de um conjunto independente do utilizado para a modelagem.



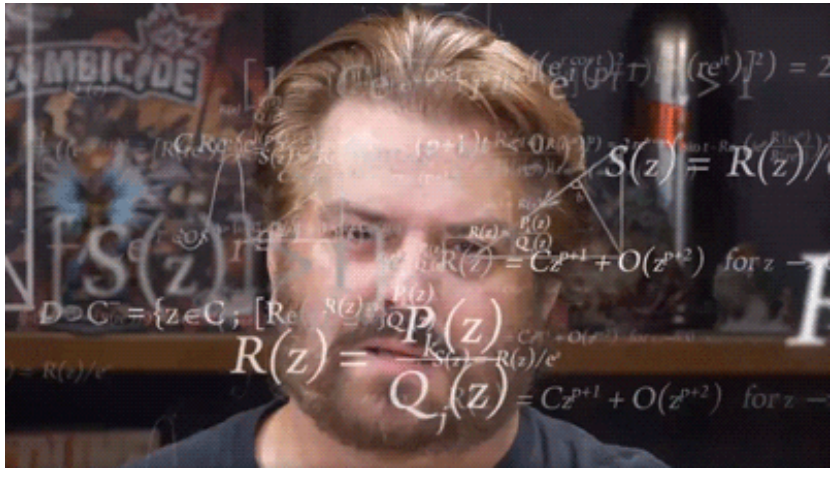
Perde-ganha viés-variância (1)



Erro redutível e irredutível

Lembre que podemos decompor o erro de predição da seguinte forma:

$$\begin{aligned}\mathrm{E}[(Y - \hat{g}(\mathbf{X}))^2] &= \int \left(\mathrm{Viés}^2[\hat{g}(\mathbf{x})] + \mathrm{Var}[\hat{g}(\mathbf{x})]\right) dF_X(\mathbf{x}) + \sigma^2 \\ &= \int \mathrm{EQM}[\hat{g}(\mathbf{x})]\, dF_X(\mathbf{x}) + \sigma^2.\end{aligned}$$



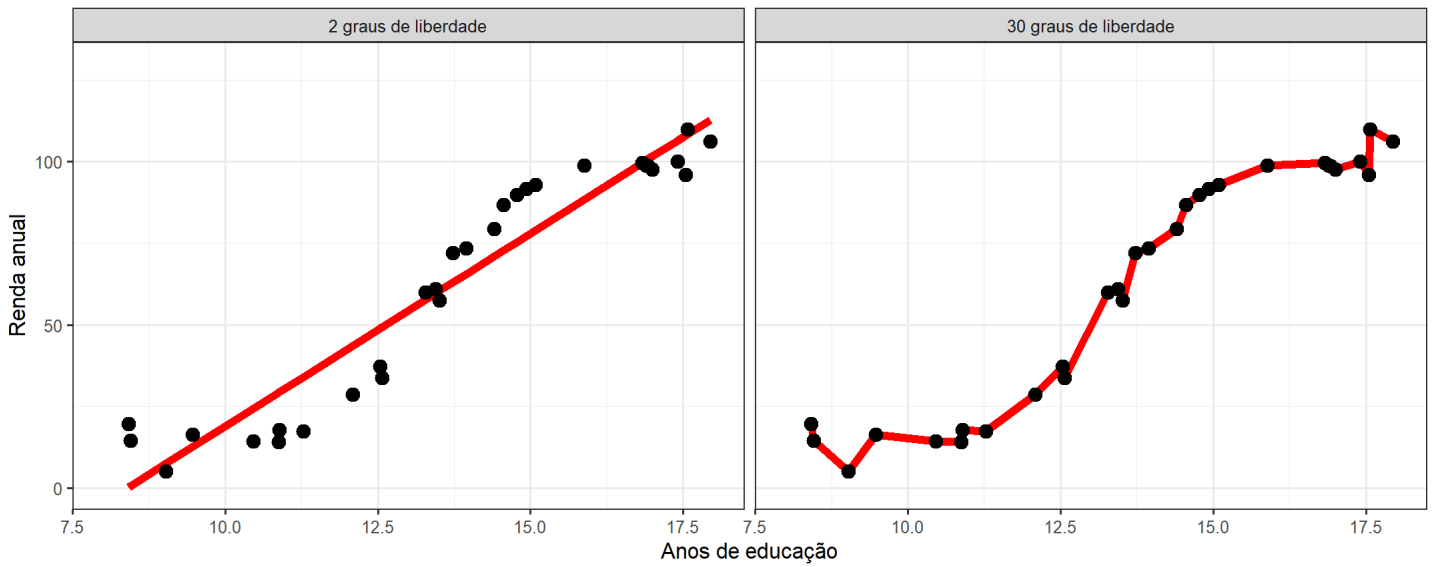
A interpretação deste resultado é que podemos reduzir $\mathrm{EQM}[\hat{g}(\mathbf{x})]$ escolhendo o método de aprendizagem que produz $\hat{g}$, mas não controlamos o tamanho de σ^2.

Erro quadrático médio de treinamento (1)

- Na prática não conhecemos g!!
- Como estimar o erro de predição esperado?

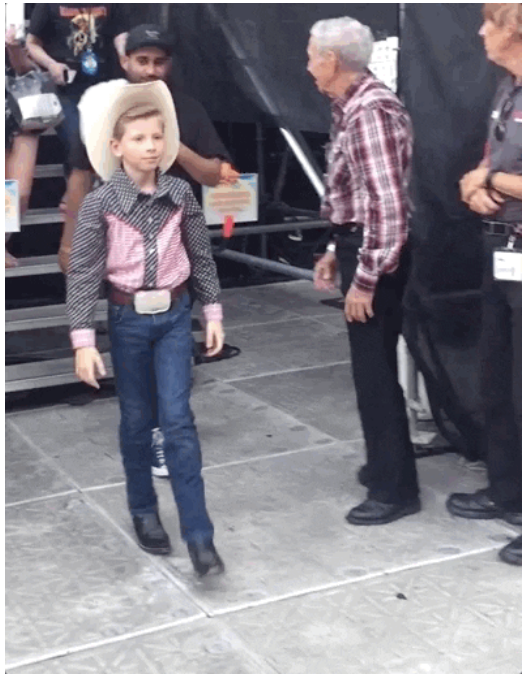
$$\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{g}(x_i))^2$$

Problema?



Como estimar o erro de predição?

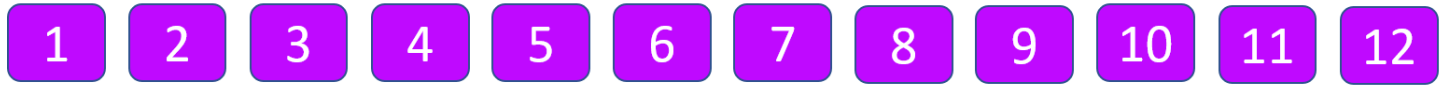
Já vimos que não podemos ajustar um modelo e fazer a avaliação preditiva utilizando o mesmo conjunto de dados.



No entanto, podemos utilizar algumas abordagens como treinamento/teste ou validação cruzada.

Validation Set Approach

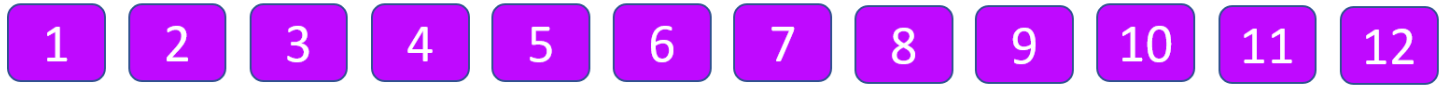
Divisão em dois conjuntos: **treinamento** e **validação**.



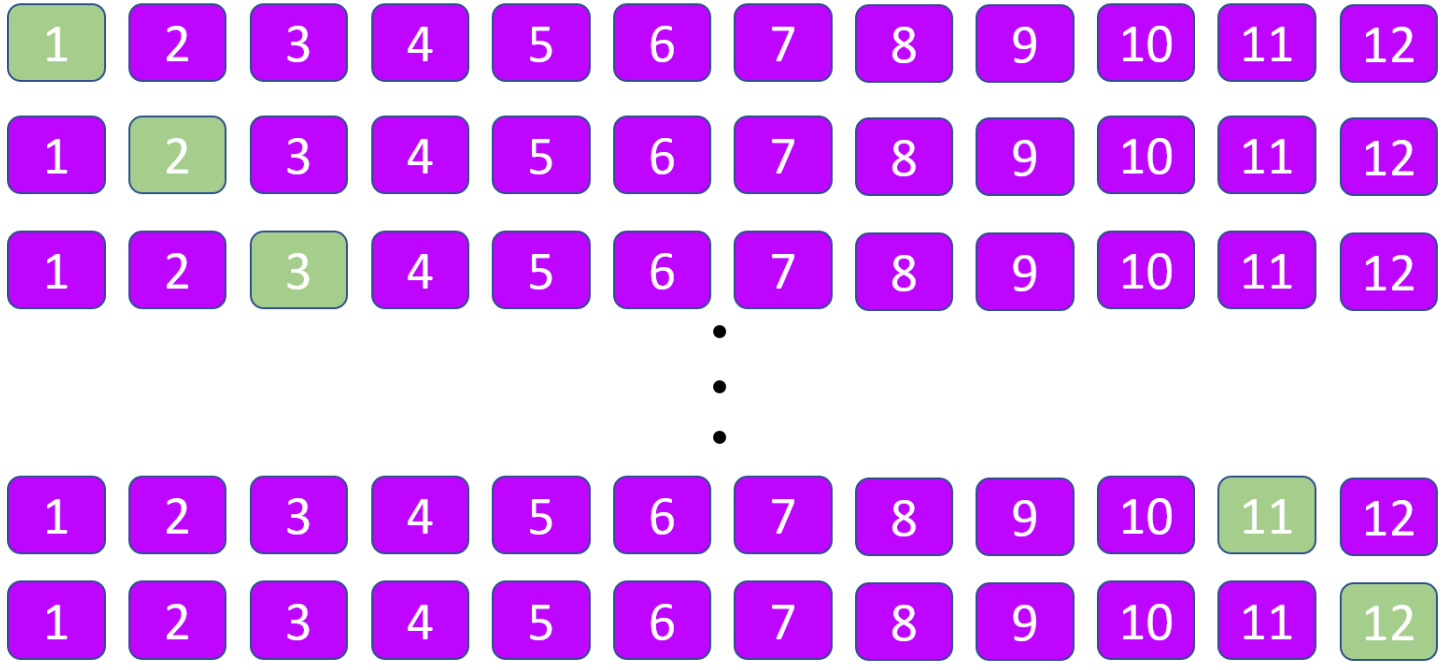
Por exemplo, considerando 75% das observações para treinamento e 25% para validação, seria possível ter a seguinte divisão.

Leave-One-Out Cross-Validation

Considere o seguinte conjunto:

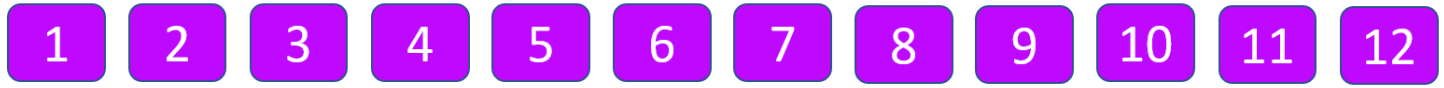


Nessa abordagem iremos considerar uma observação a cada vez.

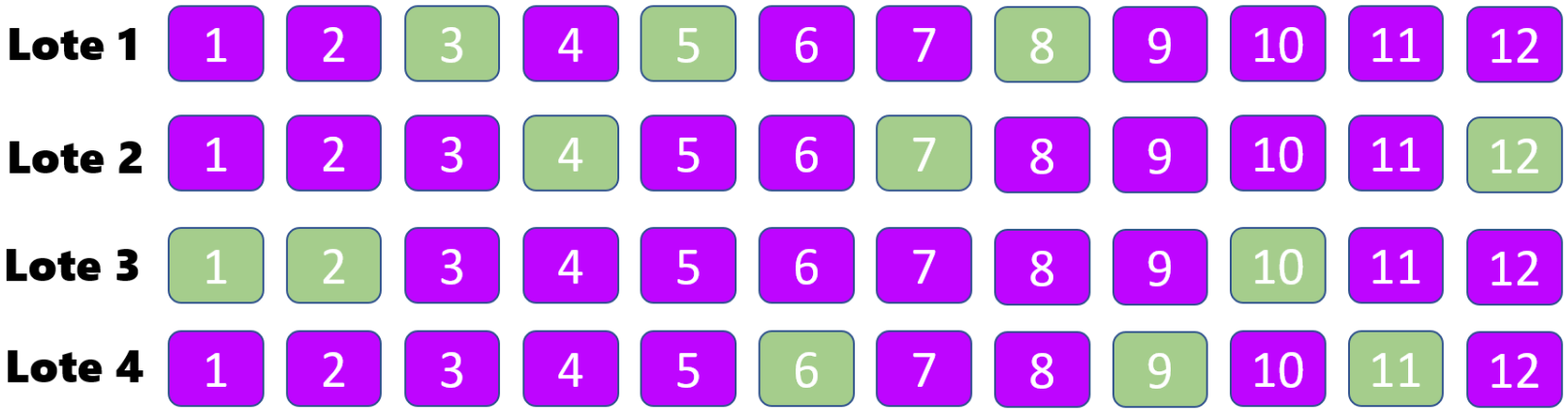


k-fold Cross-Validation

Considere o seguinte conjunto:



Vamos separar esses dados em 4 lotes.



Relação entre os dois procedimentos

- LOOCV é um caso particular da validação cruzada para $k = n$.
- LOOCV é um procedimento com mais custo computacional.
- O procedimento LOOCV, em geral, apresenta uma variância maior do que a validação cruzada em k lotes. Uma vez que

$$\text{Var}(\text{CV}_{(n)}) = \frac{1}{n^2}\sum_{i=1}^{n}\text{Var}(\text{EQM}_i) + \frac{2}{n^2}\sum_{1\leq i\leq j\leq n}\text{Cov}(\text{EQM}_i, \text{EQM}_j).$$

A prática indica que as escolhas $k = 5$ e $k = 10$ são um bom compromisso entre viés, variância e custo computacional.

Exemplo - Treino/Teste

Neste exemplo vamos considerar a abordagem **treino/teste** para estimar o erro de previsão. Considere o seguinte conjunto de dados e separação:

id	valor	grupo
1	5	Treino
2	10	Treino
3	0	Teste
4	20	Treino
5	5	Treino
6	30	Teste

Qual seria a previsão/regressor/estimativa? Vamos considerar a média amostral.

$$\bar{y} = \frac{y_1 + y_2 + y_4 + y_5}{4} = 10$$

Agora vamos calcular uma medida de desempenho.

$$\text{EQM} = \frac{\sum(y_i - \bar{y})^2}{n} = \frac{(y_3 - 10)^2 + (y_6 - 10)^2}{2} = 250$$

obs: é possível considerar diferentes métricas como, por exemplo, perda média absoluta.

Exemplo - Validação Cruzada em 3 lotes

Neste exemplo vamos considerar a abordagem de **validação cruzada em 3 lotes** para estimar o erro de previsão. Considere o seguinte conjunto de dados e lotes:

id	valor	lote
1	5	1
2	10	2
3	0	1
4	20	3
5	5	2
6	30	3

Para o lote 1, temos:

$$\bar{y}_1 = 16.25 \quad \text{e} \quad \text{EQM}_1 = \frac{(5-16.25)^2 + (0-16.25)^2}{2} = 195.3$$

Complete o procedimento com os dois lotes restantes.

$$\text{EQM} = \frac{1}{3}\sum_{i=1}^{3} \text{EQM}_i = 221.9$$

Note que, para cada lote, o procedimento funciona como a abordagem treinamento/teste.

Exemplo - Treino/Teste - Classificação

Neste exemplo vamos considerar a abordagem **treino/teste** para estimar o erro de previsão. Considere o seguinte conjunto de dados e separação:

id	resposta	grupo
1	sim	Treino
2	sim	Treino
3	não	Teste
4	não	Treino
5	sim	Treino
6	sim	Teste

Qual seria a previsão/regressor/estimativa? Vamos considerar a classe mais frequente como previsão. Nesse caso, $\hat{y} =$ "sim".

Agora vamos calcular uma medida de desempenho.

$$\text{Erro} = \frac{\sum \mathbb{I}(y_i \neq \hat{y})}{n} = \frac{\mathbb{I}(y_3 \neq \text{sim}) + \mathbb{I}(y_6 \neq \text{sim})^2}{2} = \frac{1+0}{2} = 0.5$$

Exemplo - Validação Cruzada em 2 lotes - Classificação

Neste exemplo vamos considerar a abordagem de **validação cruzada em 2 lotes** para estimar o erro de previsão. Considere o seguinte conjunto de dados e lotes:

id	resposta	lote
1	sim	1
2	sim	2
3	não	2
4	não	1
5	sim	2
6	sim	1

Para o lote 1, temos:

$$\hat{y}_1 = \text{sim} \quad \text{e} \quad \text{Erro}_1 = \frac{\mathbb{I}(y_1 \neq \text{sim}) + \mathbb{I}(y_4 \neq \text{sim}) + \mathbb{I}(y_6 \neq \text{sim})}{2} = \frac{1}{3}$$

Complete o procedimento para o segundo lote.

$$\text{Erro} = \frac{1}{2}\sum_{i=1}^{2} \text{Erro}_i = \frac{1}{2}\left(\frac{1}{3} + \frac{1}{3}\right) = \frac{1}{3}$$

Note que, para cada lote, o procedimento funciona como a abordagem treinamento/teste.

Regressão KNN

Regressão kNN

A regressão de k vizinhos mais próximos (KNN) é baseado na ideia de que objetos semelhantes tendem a compartilhar características semelhantes.

Esse método pode ser resumido da seguinte forma:

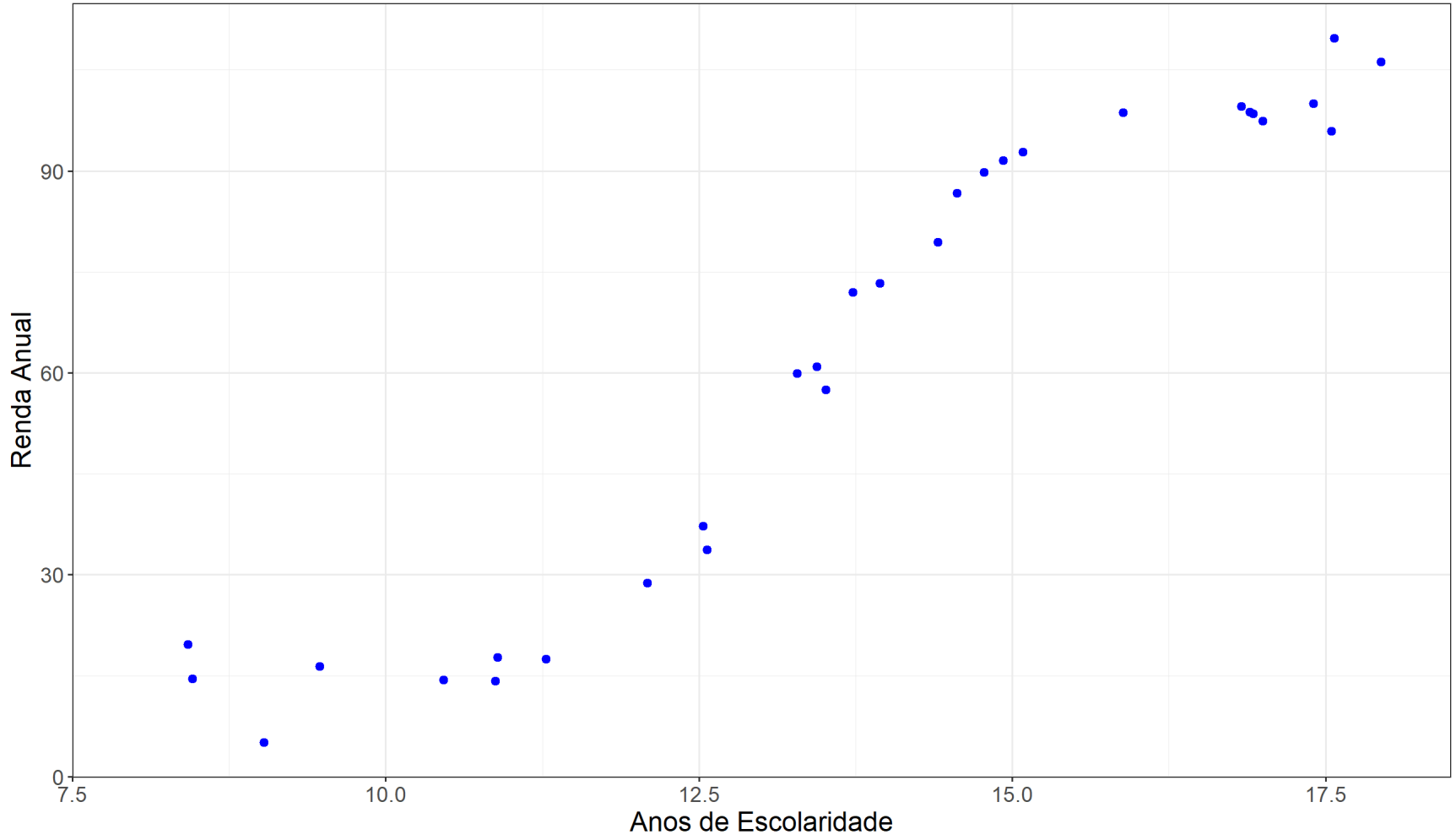
- encontre os k vizinhos mais próximos (e.g. distância euclidiana) de uma observação
- Calcule a previsão a partir da média (pode ser ponderada) da resposta dos k vizinhos

Note que k representa um hiperparâmetro de forma que:

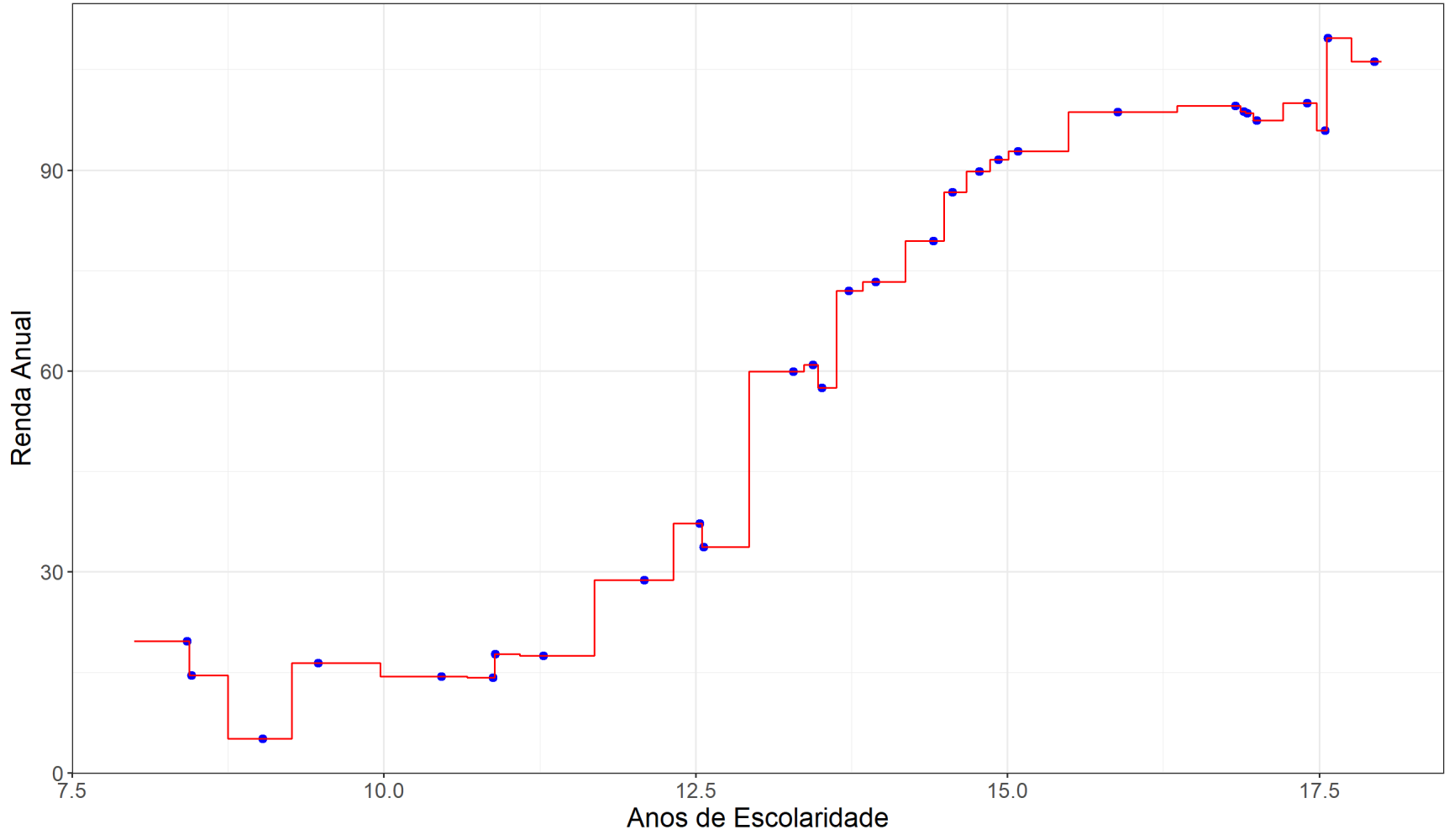
- valores grandes dessa quantidade tornam a regressão pouco flexível e
- valores reduzidos levam a uma regressão com pouca suvidade.

A seguir verificaremos uma forma de estimar o valor ótimo do hiperparâmetro k com uso da validação cruzada.

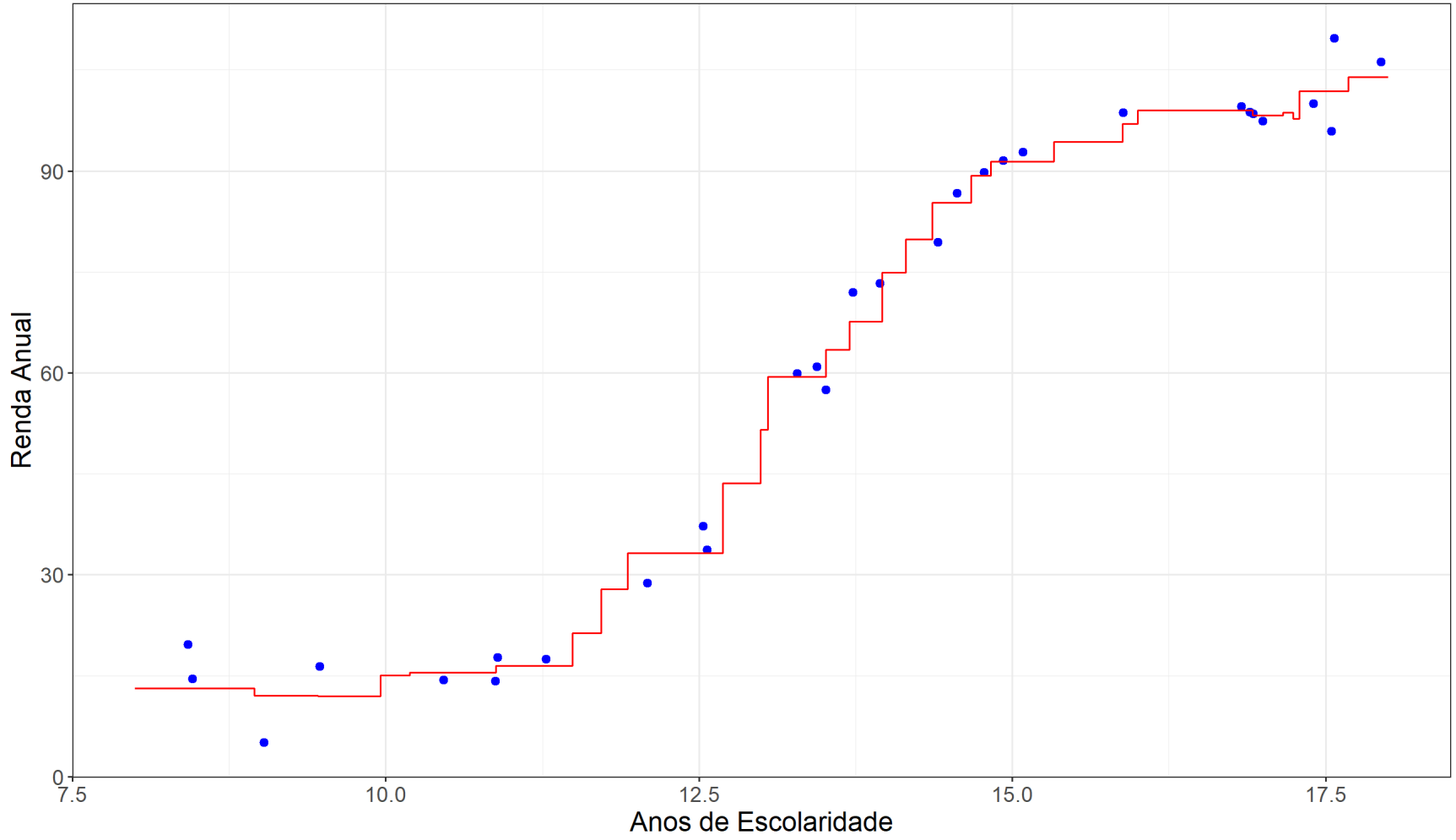
Regressão K-NN (não confundir com k da VC)



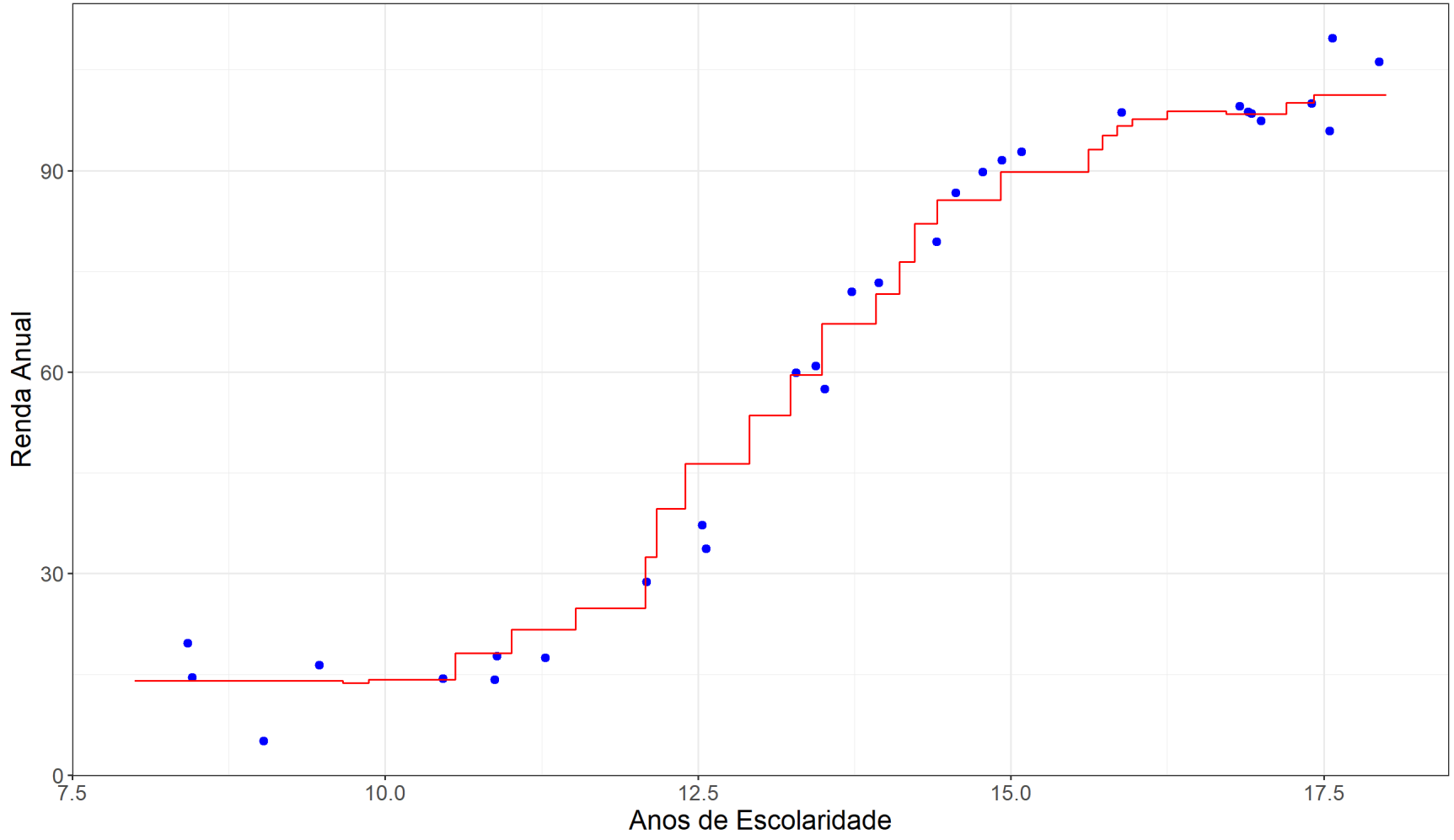
Regressão 1-NN



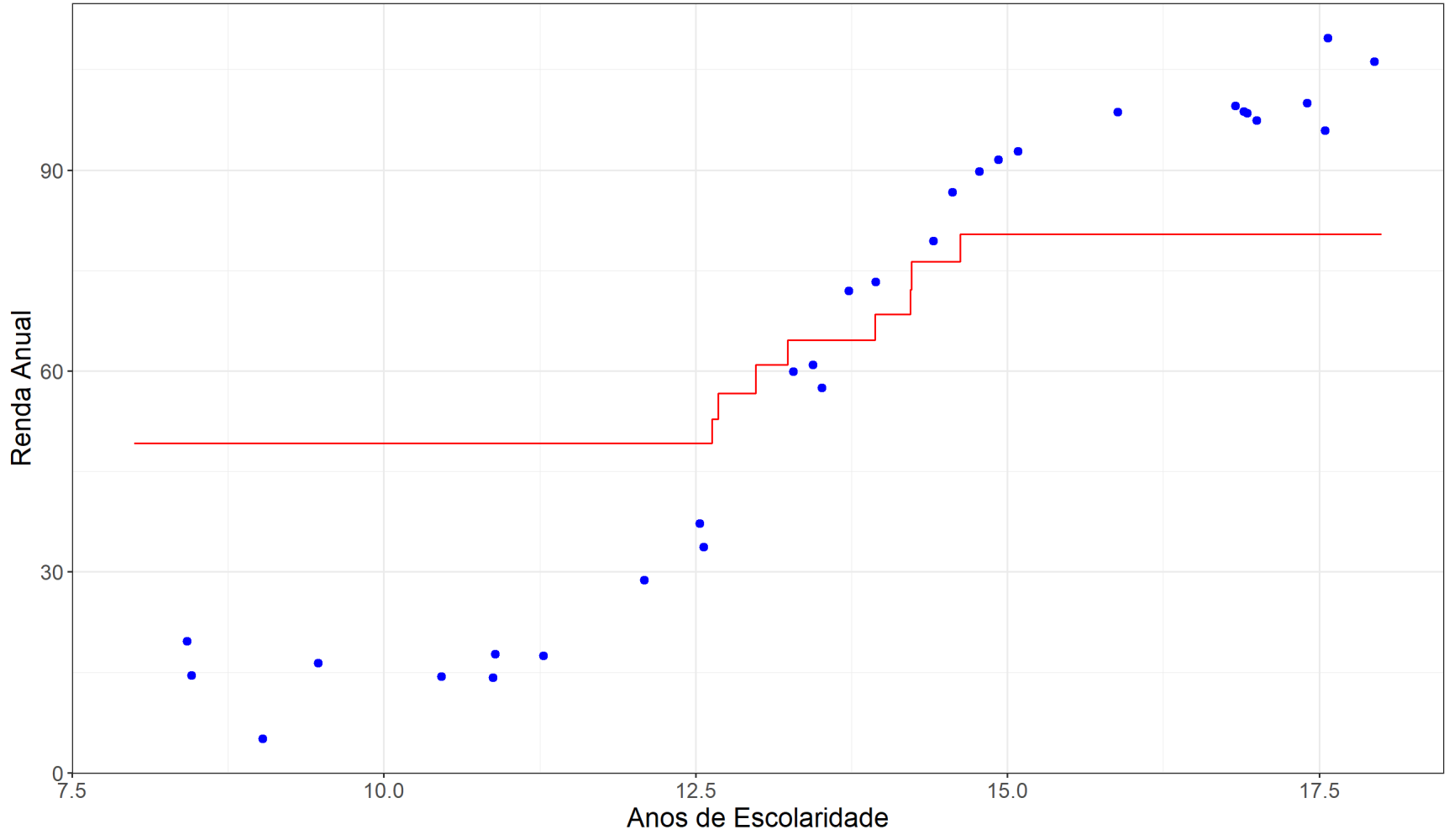
Regressão 3-NN



Regressão 6-NN



Regressão 22-NN



Simulação

Vamos retomar o exemplo da simulação e definir os graus de liberdade com a utilização da validação cruzada.

$$Y = g(x) + \epsilon = 45 \cdot \tanh\left(\frac{x}{1,9} - 7\right) + 57 + \epsilon,$$

em que $\epsilon \sim N(0, 4^2)$.

```
n_obs <- 100

set.seed(123)

dados <- tibble(x = sort(runif(n = n_obs, min = 8, max = 18)),
                y = 45*tanh(x/1.9 - 7) + 57 + rnorm(n = n_obs, mean = 0, sd = 4))
```

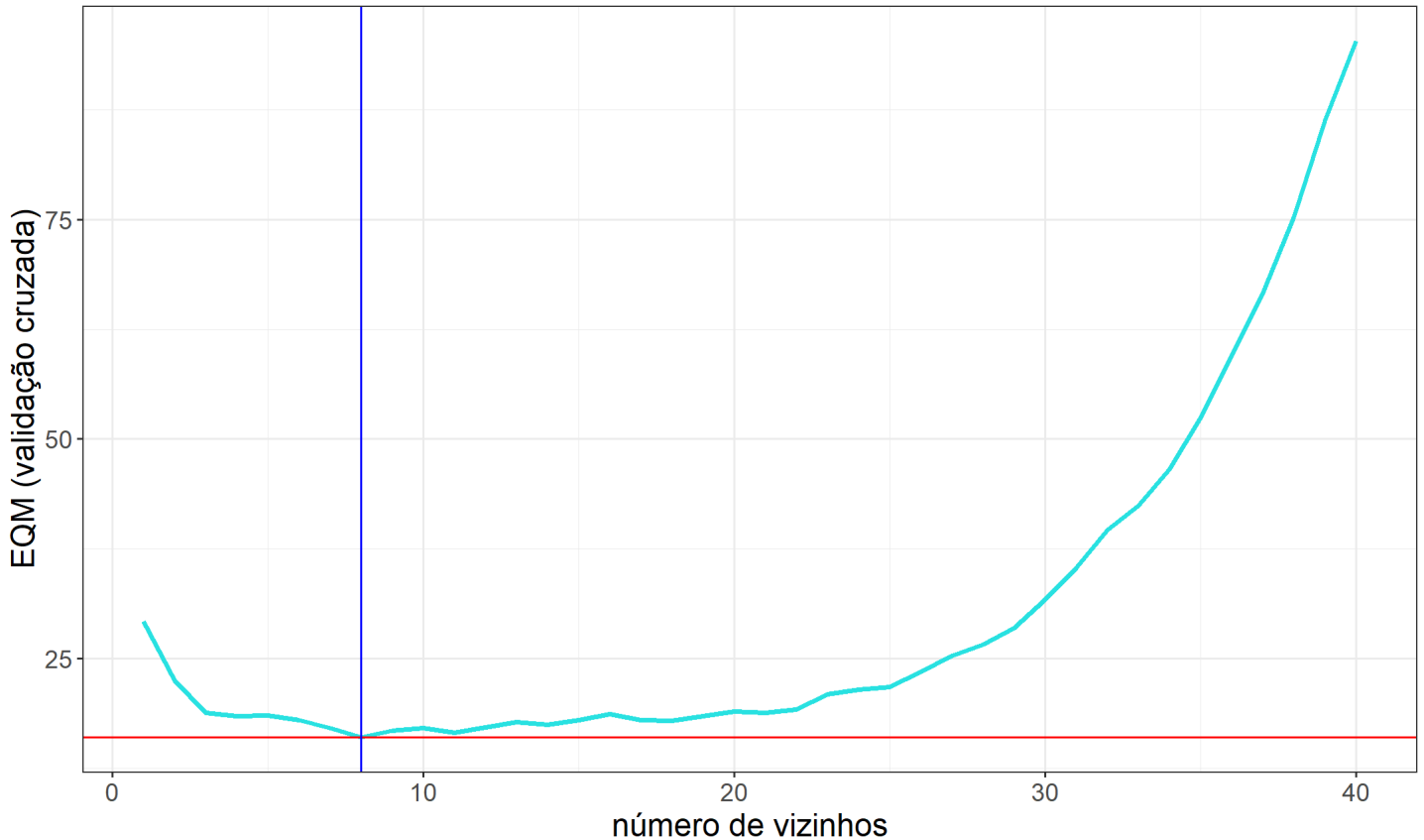
Regressão KNN

Considere os parâmetros definidos a seguir e identifique o número ótimo de vizinhos para a regressão a partir do procedimento de validação cruzada.

- **n_sample**: número de observações do conjunto de treinamento (100)
- **folds**: número de lotes da validação cruzada (5-*fold*)
- **n_vizinhos**: de 1 a 40

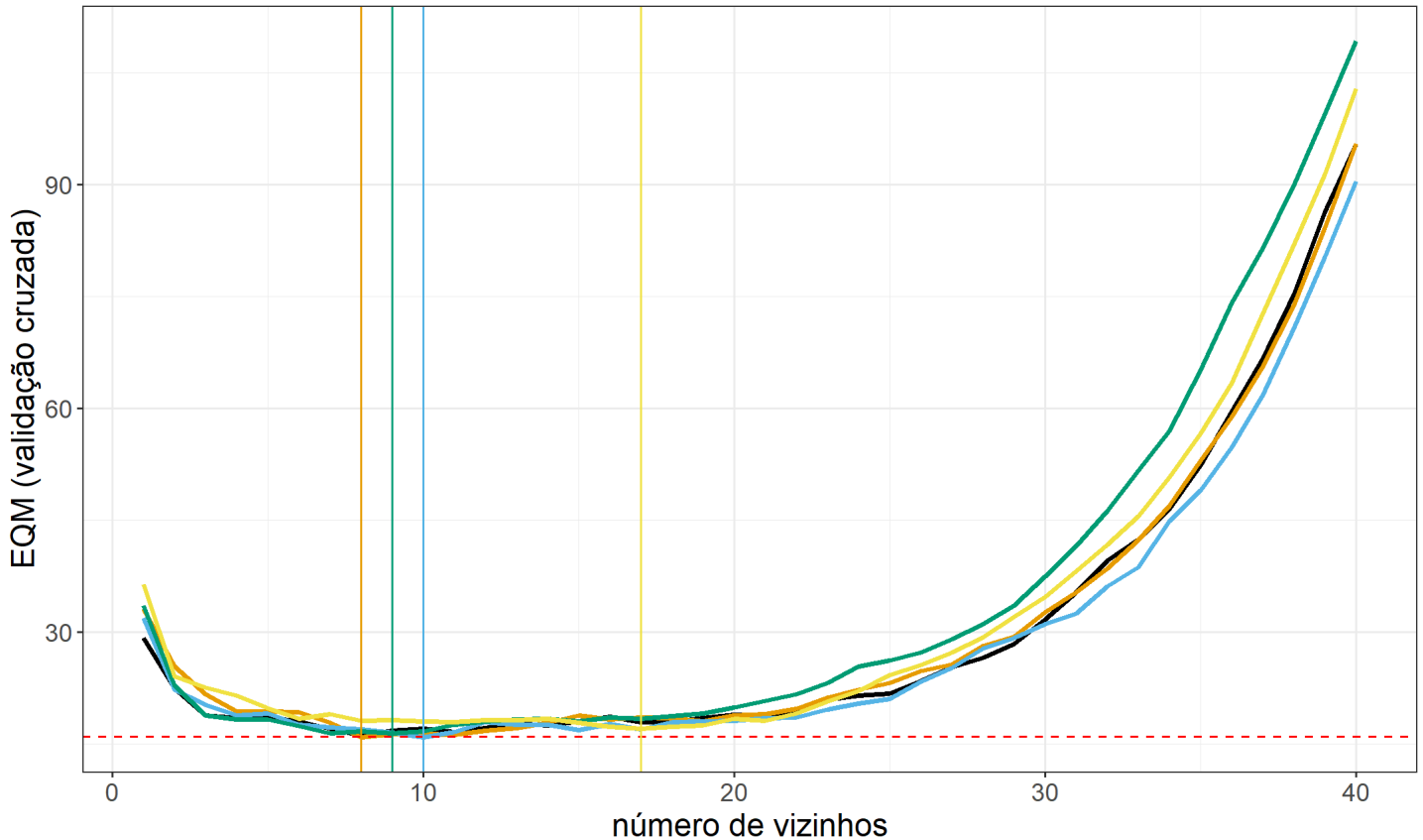
EQM - Validação Cruzada 5 Lotes (*CV 5-folds*)

Número de vizinhos = 8

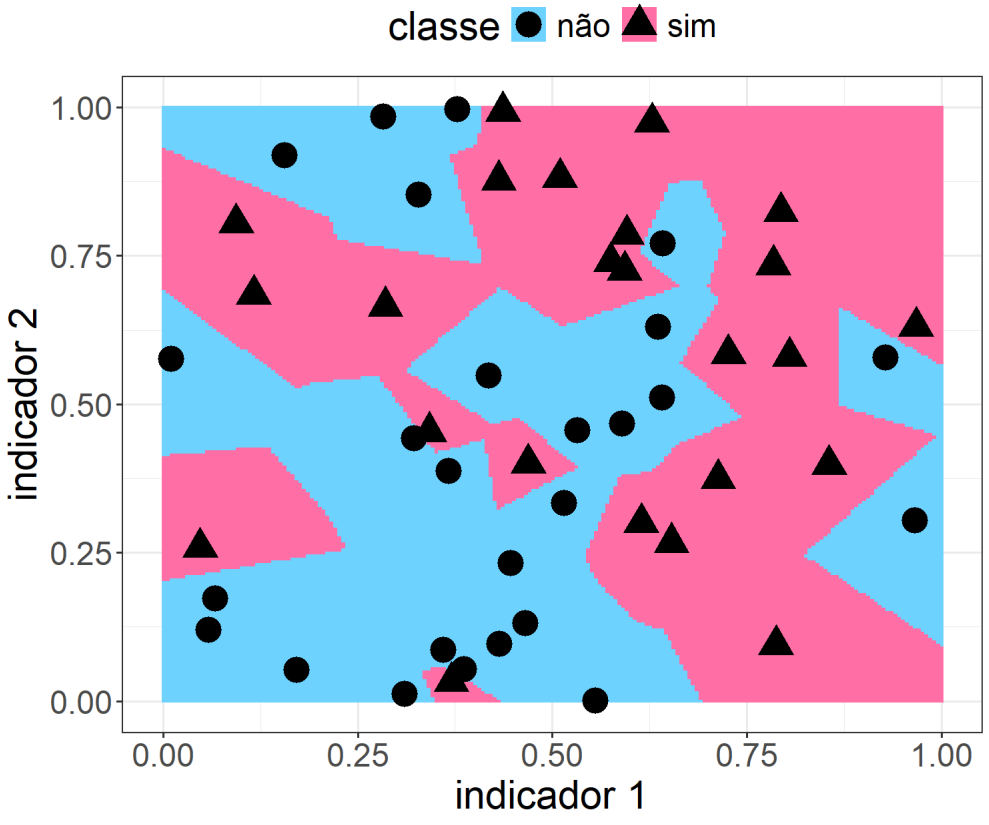


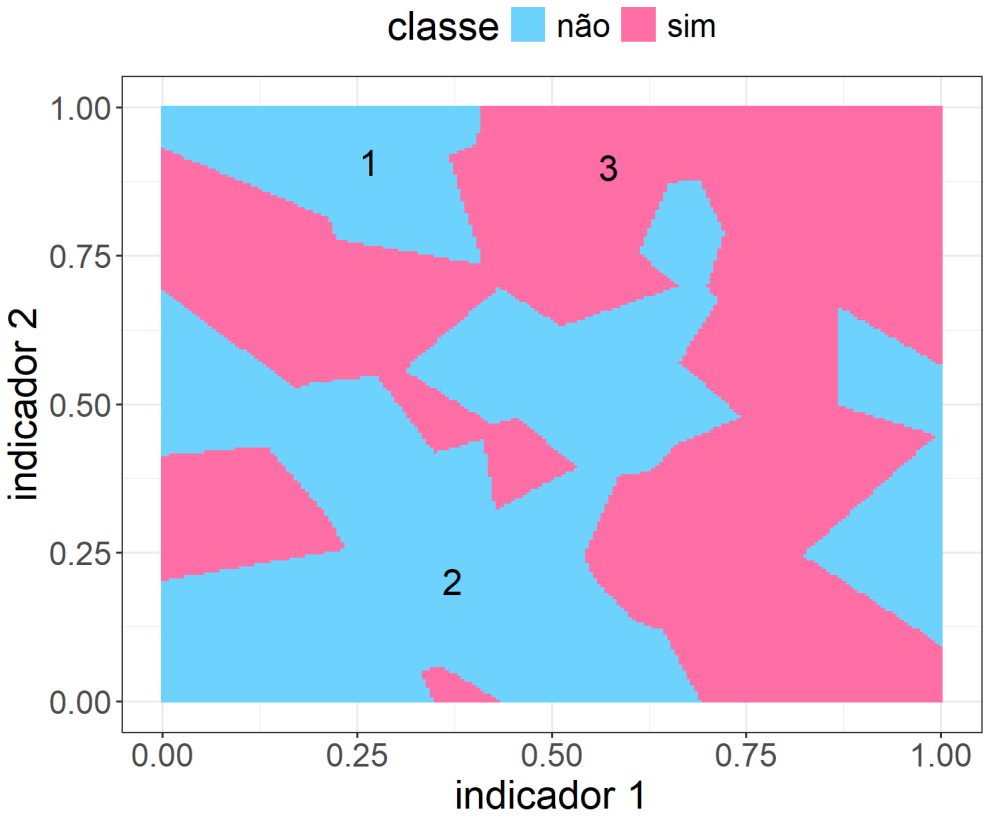
EQM - Validação Cruzada 5 Lotes (*CV 5-folds*)

Número de vizinhos = 8, 8, 10, 9 e 17



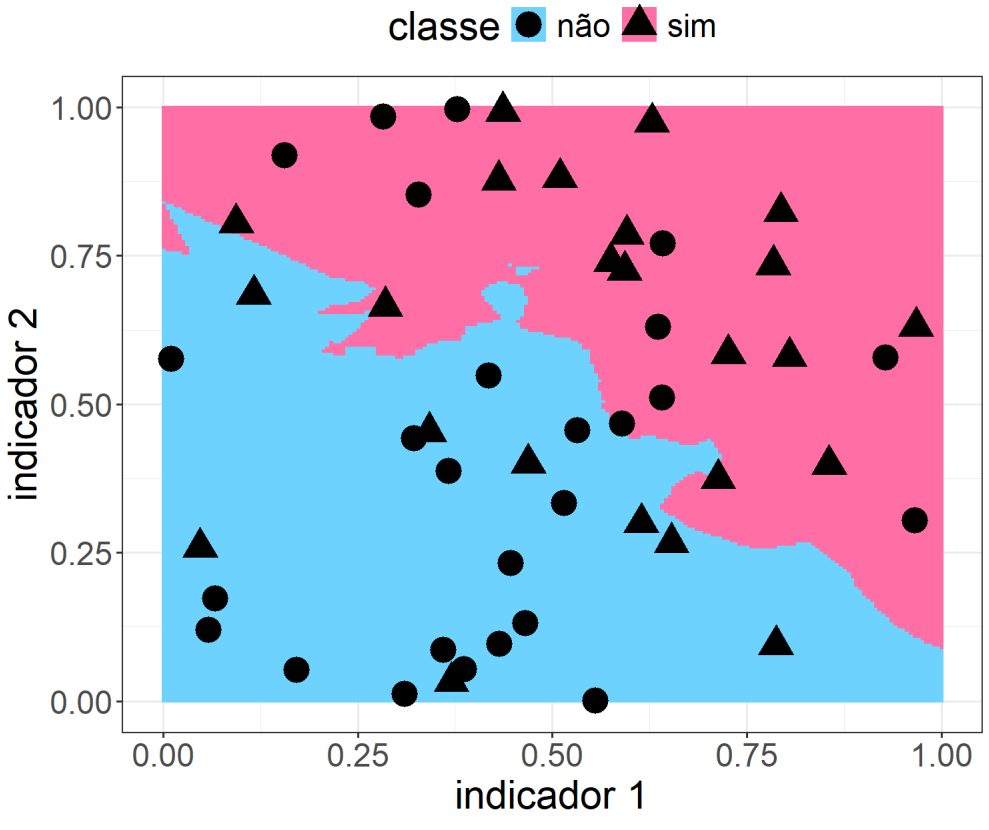
1NN - Classificação

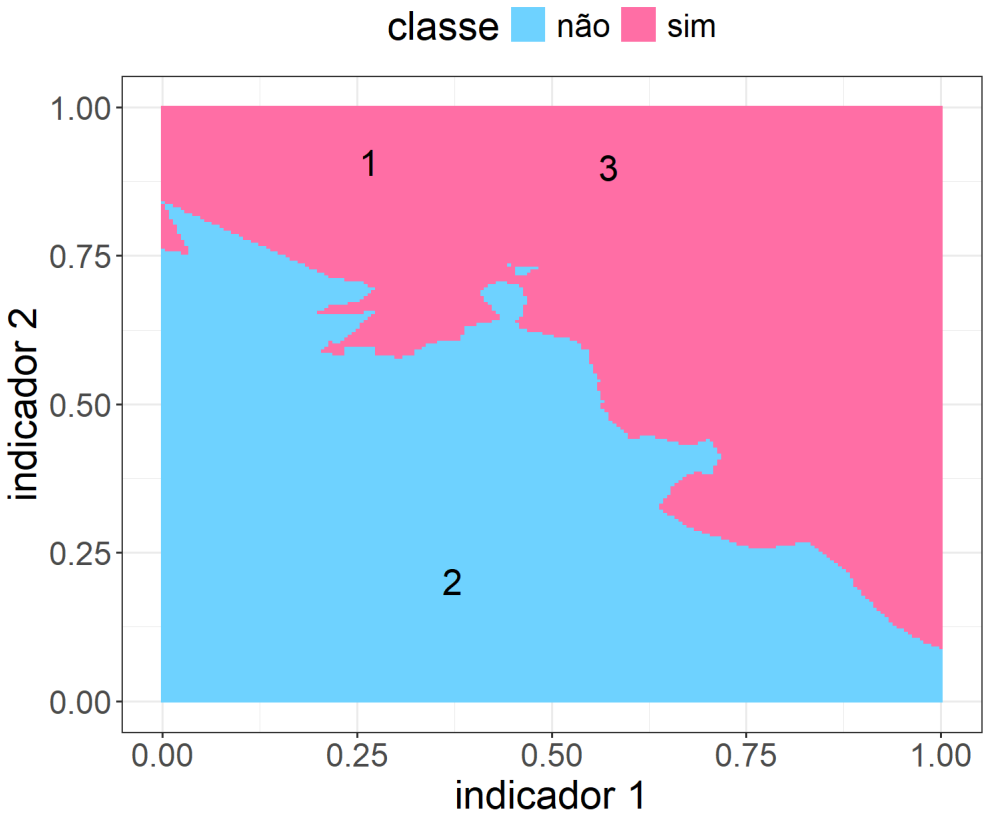




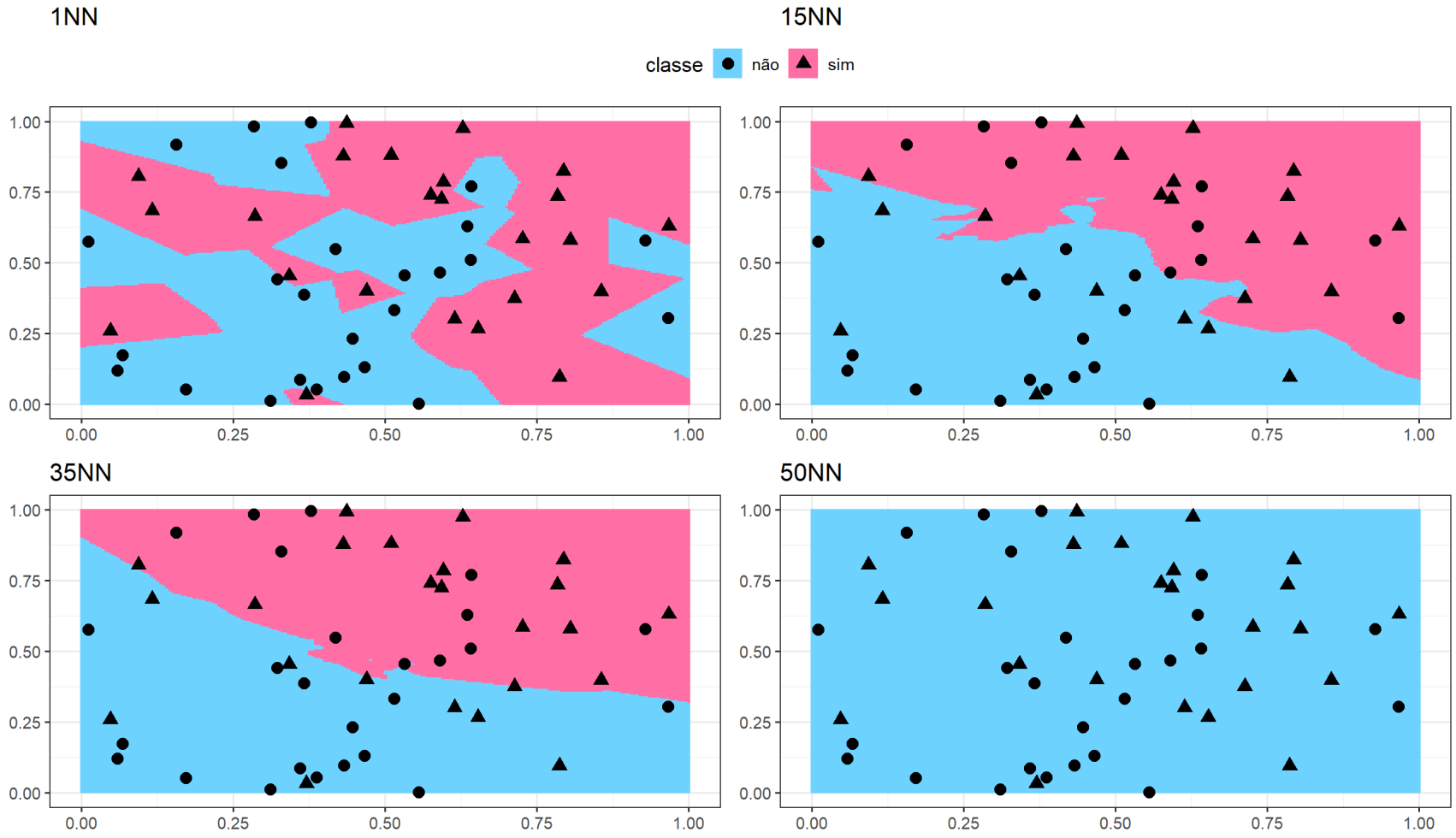
Se a observação 1 é, de fato, "não", a observação 2 é "não" e a observação 3 é "sim", qual o erro estimado?

10NN - Classificação





KNN - Classificação



Obrigado!

tiagoms.com

tiagomendonca

tiagoms1@insper.edu.br