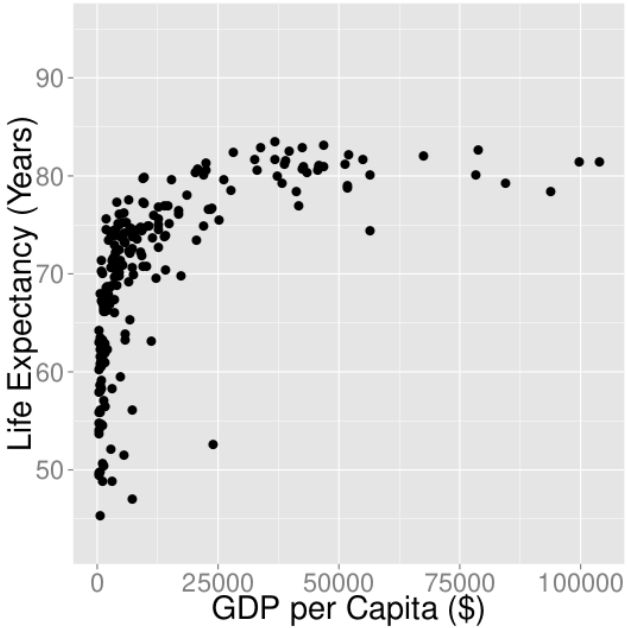
Super-ajuste (Overfitting) e Seleção de Modelos

INFORMAÇÃO,

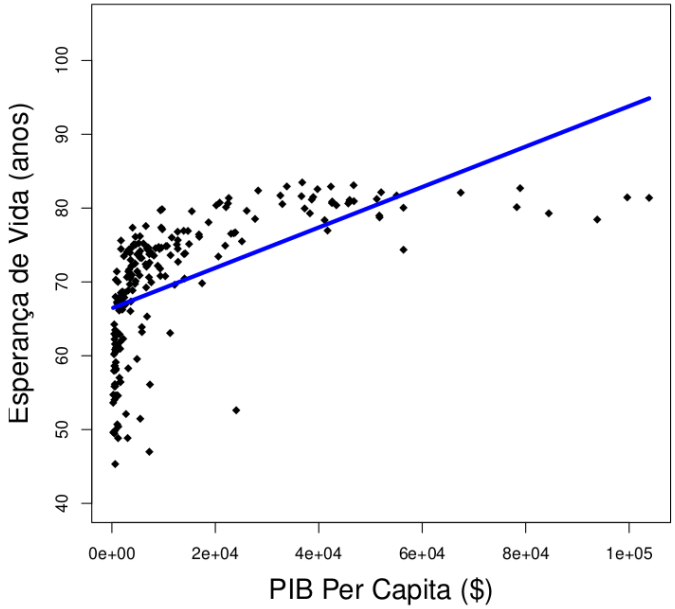
TECNOLOGIA

& INOVAÇÃO

REGRESSÃO LINEAR



REGRESSÃO LINEAR



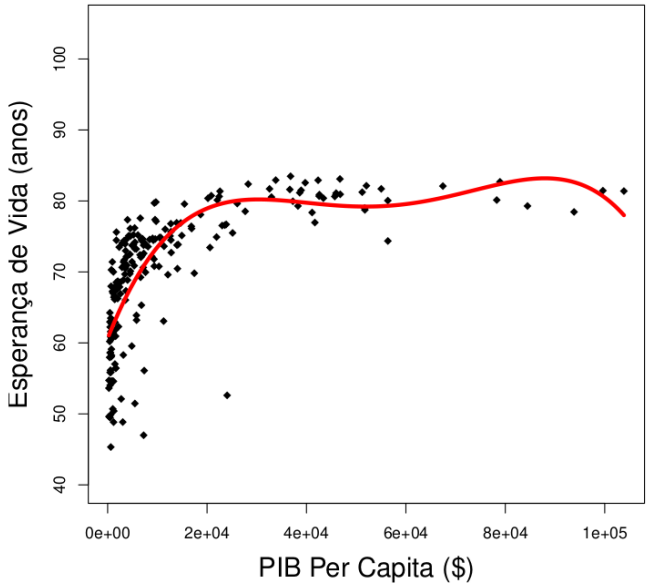
$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

REGRESSÃO POLINOMIAL

REGRESSÃO LINEAR

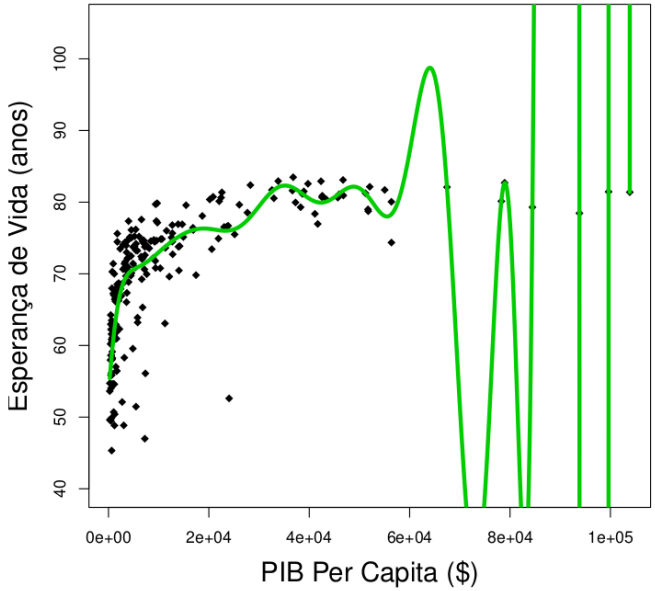
$$g(x) = \overbrace{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \widehat{\beta}_3 x_3 + \widehat{\beta}_4 x_4}^{\widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \widehat{\beta}_3 x^3 + \widehat{\beta}_4 x^4}$$

REGRESSÃO LINEAR



$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \widehat{\beta}_3 x^3 + \widehat{\beta}_4 x^4$$

REGRESSÃO LINEAR



$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \ldots + \widehat{\beta}_{50} x^{50}$$

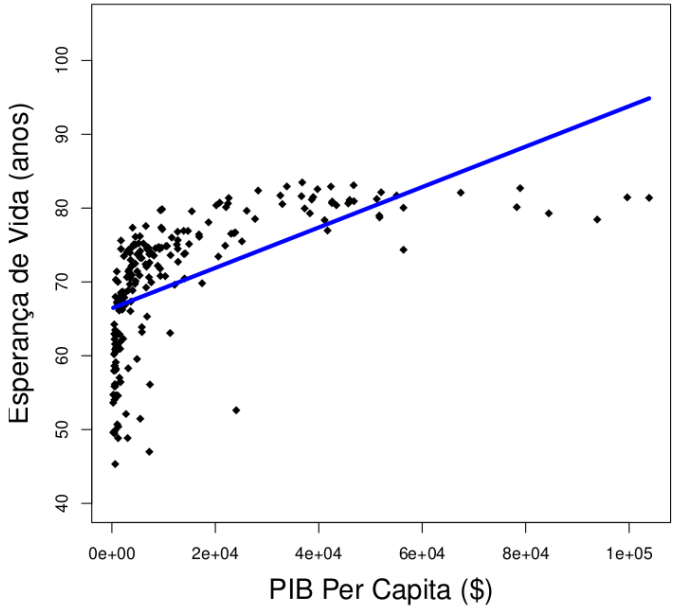
BIAS-VARIANCE TRADEOFF

Bias-Variance Tradeoff

Risco = Variabilidade Intrínseca + Variância + Viés

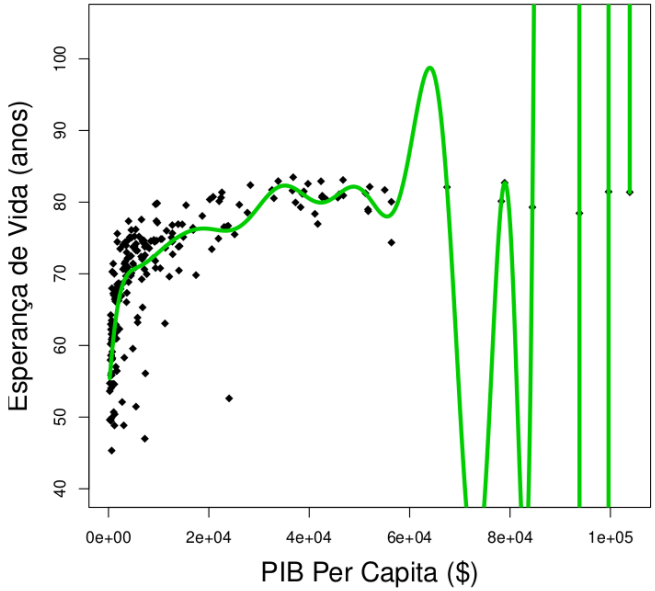
Shiny

VIÉS ALTO



$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x$$

VARIÂNCIA ALTA



$$g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \ldots + \widehat{\beta}_{50} x^{50}$$

SELEÇÃO DE MODELOS

Encontrar a melhor função de predição em

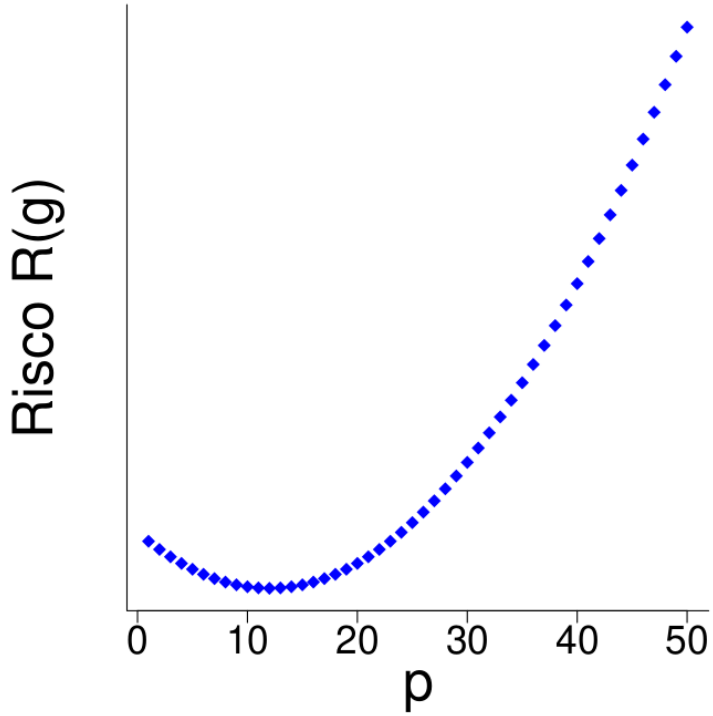
$$\mathbb{G} = \left\{ g(x) : g(x) = \beta_0 + \sum_{i=1}^{p} \beta_i x^i, \text{ para } p \in \{1, 2, \ldots, 50\} \right\}$$

Qual o melhor p?

$p = 50$: super-ajuste $\Rightarrow$ baixo poder preditivo.

$p = 1$: sub-ajuste $\Rightarrow$ baixo poder preditivo.

SELEÇÃO DE MODELOS



COMO ESCOLHER g?

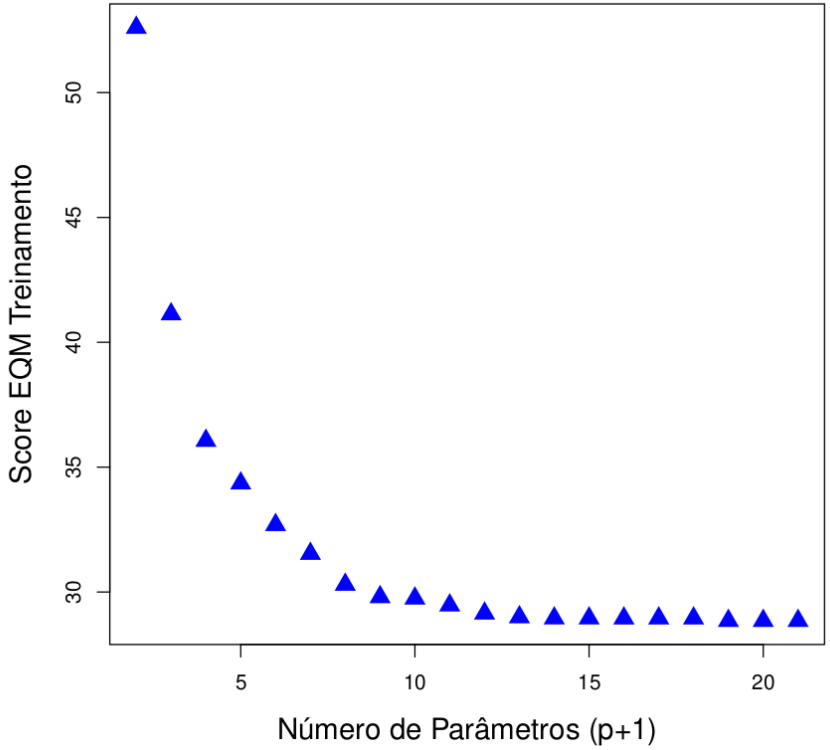
Estimação do Risco

Como estimar o risco $R(g) = \mathbb{E}\left[(Y - g(\mathbf{X}))^2\right]$?

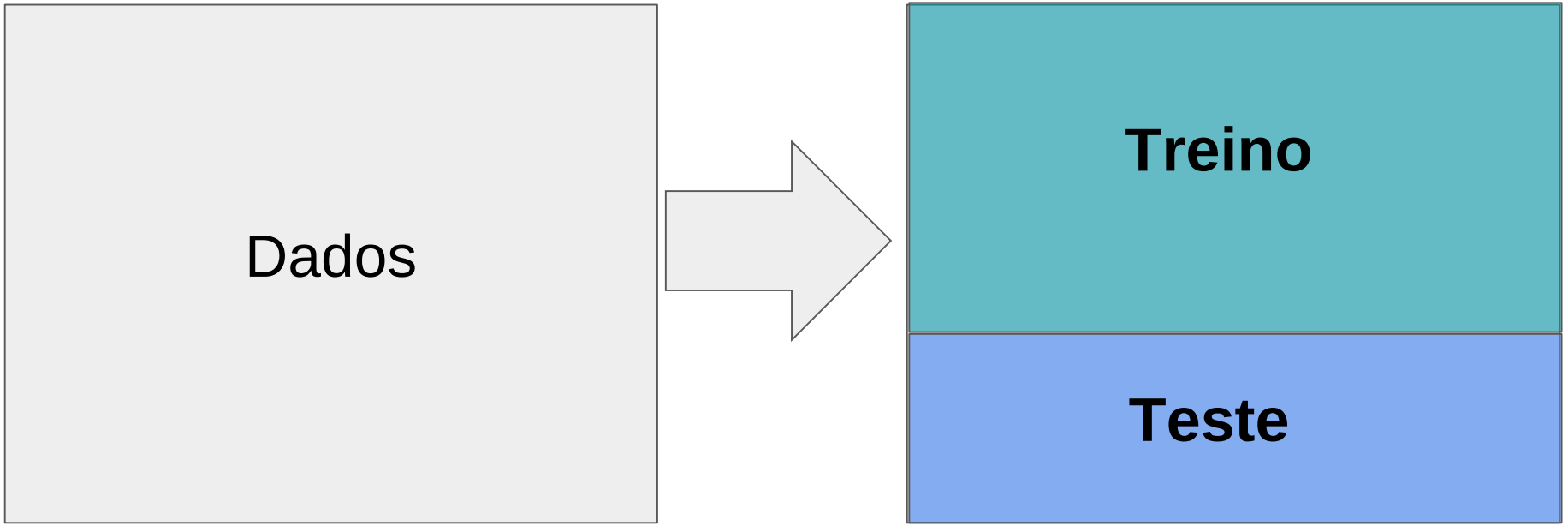
$$\frac{1}{n}\sum_{i=1}^{n}(Y_i - g(\mathbf{X}_i))^2 := EQM(g)\ ?$$

Levaria ao super-ajuste

Estimação do Risco



Data Splitting



Data Splitting

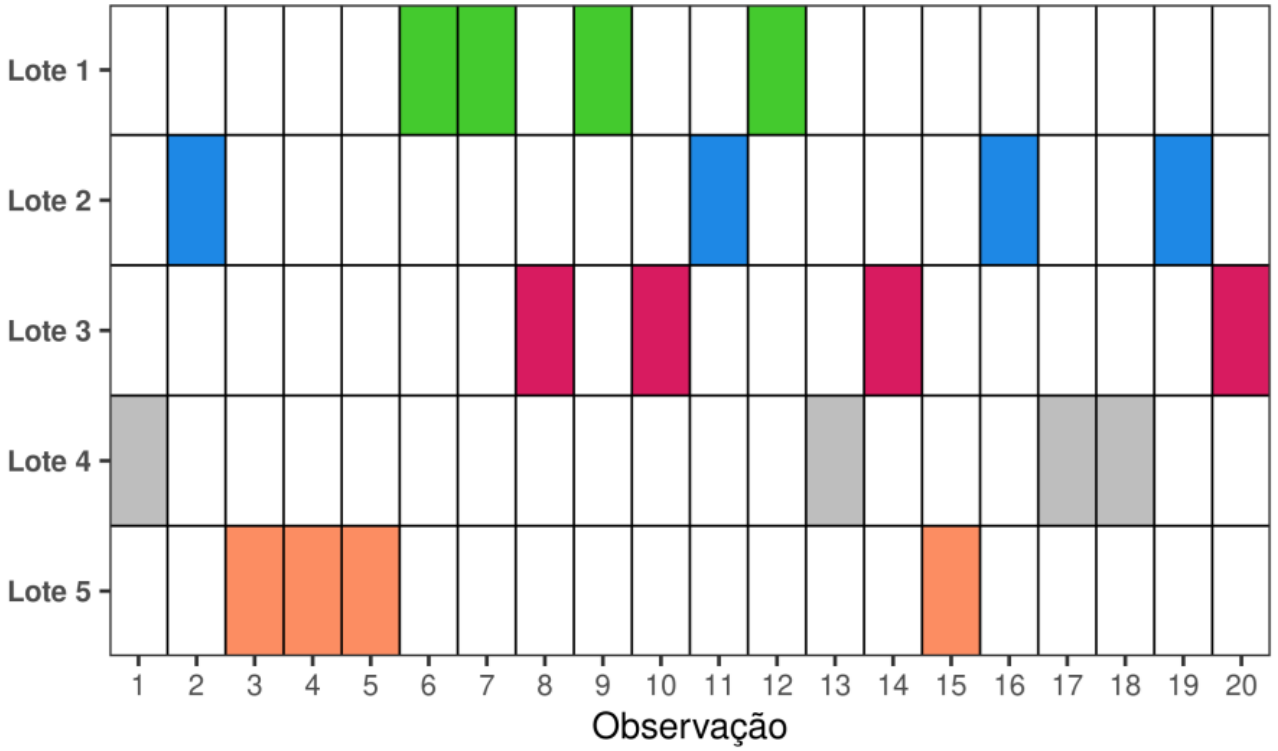
$$\overbrace{(\mathbf{X}_1, Y_1), (\mathbf{X}_2, Y_2), \ldots, (\mathbf{X}_s, Y_s)}^{\text{Treinamento (e.g., 70\%)}}, \overbrace{(\mathbf{X}_{s+1}, Y_{s+1}), \ldots, (\mathbf{X}_n, Y_n)}^{\text{Validação (e.g., 30\%)}}$$

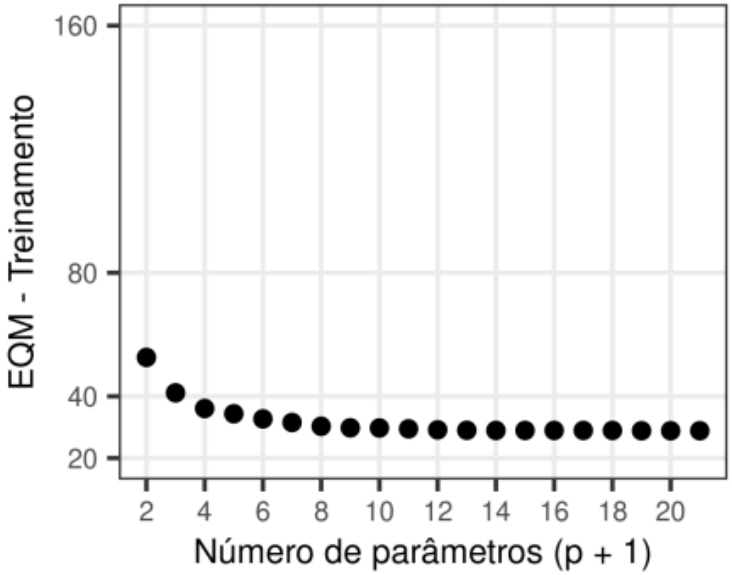
Treinamento: estimar g

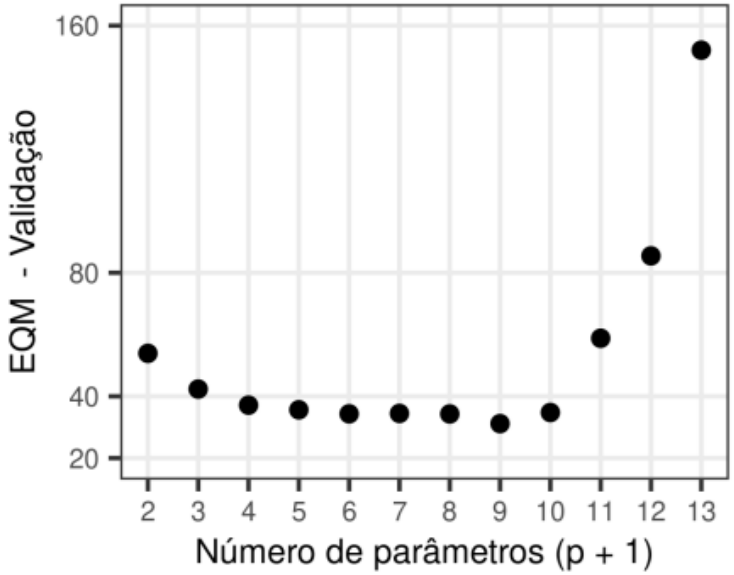
Validação: estimar $R(g)$

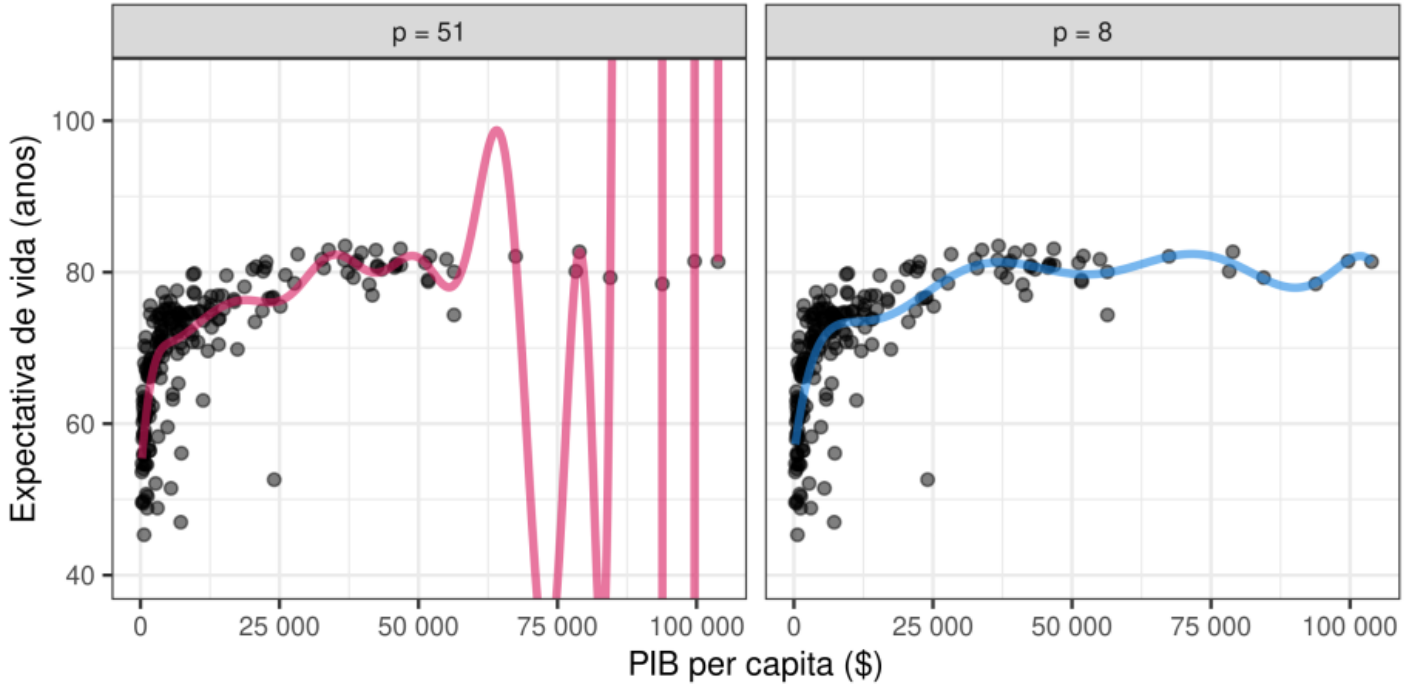
$$R(g) \approx \frac{1}{n-s} \sum_{i=s+1}^{n} (Y_i - g(\mathbf{X}_i))^2$$

Validação Cruzada




EQM - Treinamento
160
80
40
20
2 4 6 8 10 12 14 16 18 20
Número de parâmetros (p + 1)


EQM - Validação
160
80
40
20
2 3 4 5 6 7 8 9 10 11 12 13
Número de parâmetros (p + 1)


p = 51
p = 8
Expectativa de vida (anos)
100
80
60
40
0
25 000
50 000
75 000
100 000
PIB per capita ($)

Tarefa - IML 1.1

Precificação de notebooks

Precificação (*pricing*) de bens é um importante instrumento para a gestão estratégica e operacional de um negócio. O objetivo é entender o quanto um produto vale no mercado e quanto os consumidores estão dispostos a pagar por ele. Nesta tarefa, iremos elaborar um modelo de precificação de notebooks.

Características do banco de dados:

- 2160 observações
- 14 colunas (13 features e 1 variável resposta)
- Dados de característica do notebook
- Preço do notebook

Na sua tarefa você vai ter que criar um modelo de regressão para explicar o preço de um veículo de acordo com as suas características.

Você tem total liberdade para selecionar os modelos que deseja testar, criar novas covariáveis, remover covariáveis e selecionar técnicas de comparação entre os modelos que você escolher.

Pontos importantes!

Análise descritiva

- A análise descritiva faz parte do processo de modelagem
- Entenda os dados
- Identifiquem o que é variável resposta e o que é variável explicativa
- Identifique missing, outliers e outros problemas de dados

Modelagem

- Entendam qual técnica de aprendizado faz mais sentido para o dados
- Testem diferentes modelos
- Avaliem os modelos por alguma medida de comparação
- Selecionem variáveis de forma inteligente
- Conclua qual modelo fornece o melhor ajuste

Estrutura

- Utilize a metodologia CRISP-DM
- Motive o desenvolvimento da análise (*business case*)
- Siga uma ordem lógica no desenvolvimento
- Faça comentários explicando seu raciocínio (o porquê de utilizar determinada função, alguma característica importante dos dados, motivação para próximos passos)
- Explicite a interpretação do modelo
- Retire comentários desnecessários ou que não impactem diretamente o código ou a interpretação
- Façam a atividade utilizando o jupyter notebook