

Machine Learning

Aula 3 - Parte I

Prof. Thiago A. N. De Andrade

Universidade Federal de Santa Maria
Departamento de Estatística

2025-08-19

Aviso aos estudantes

- Este é um material novo e atualizado, elaborado especialmente para nosso curso **Machine Learning - UFSM 2025.2**. Entretanto, **não se configura em conteúdo original**. É apenas uma compilação resumida de conteúdos presentes nas referências citadas. Em resumo: é indispensável consultar as referências indicadas.
- As imagens não são autorais e os respectivos créditos são reservados aos autores.
- Este material foi integralmente produzido em R Markdown, utilizando o pacote `xaringan`, que possibilita a criação de apresentações **ninja**.

Supervised Learning

- Regressão linear;
- Regularização *Ridge* (penalidade L2);
- Regularização *Lasso* (penalidade L1);
- Regularização *Elastic Net* (penalidades L1 e L2);

Modelo de Regressão Linear Simples

Em regressão linear simples, pretende-se explicar o comportamento de uma quantidade por meio de uma variável explicativa:

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i \quad i = 1, \dots, n,$$

em que

- y_i é o valor da variável resposta na i -ésima observação;
- x_i é o valor da variável explicativa na i -ésima observação;
- β_0 e β_1 são parâmetros desconhecidos a serem estimados;
- ϵ_i é um termo de erro aleatório.

Suposições:

1. O modelo está corretamente especificado (a relação entre x_i e y_i é linear, e não há variáveis omitidas relevantes).
2. $E(\epsilon_i) = 0, \forall i.$
3. $\text{Var}(\epsilon_i) = \sigma^2, 0 < \sigma^2 < \infty, \forall i.$
4. $\text{Cov}[(\epsilon_i, \epsilon_j)] = 0, \forall i \neq j.$
5. x_i assume pelo menos dois valores.
6. Os $\epsilon_i, \forall i$, têm distribuição normal (necessária para testes de hipóteses e intervalos de confiança)

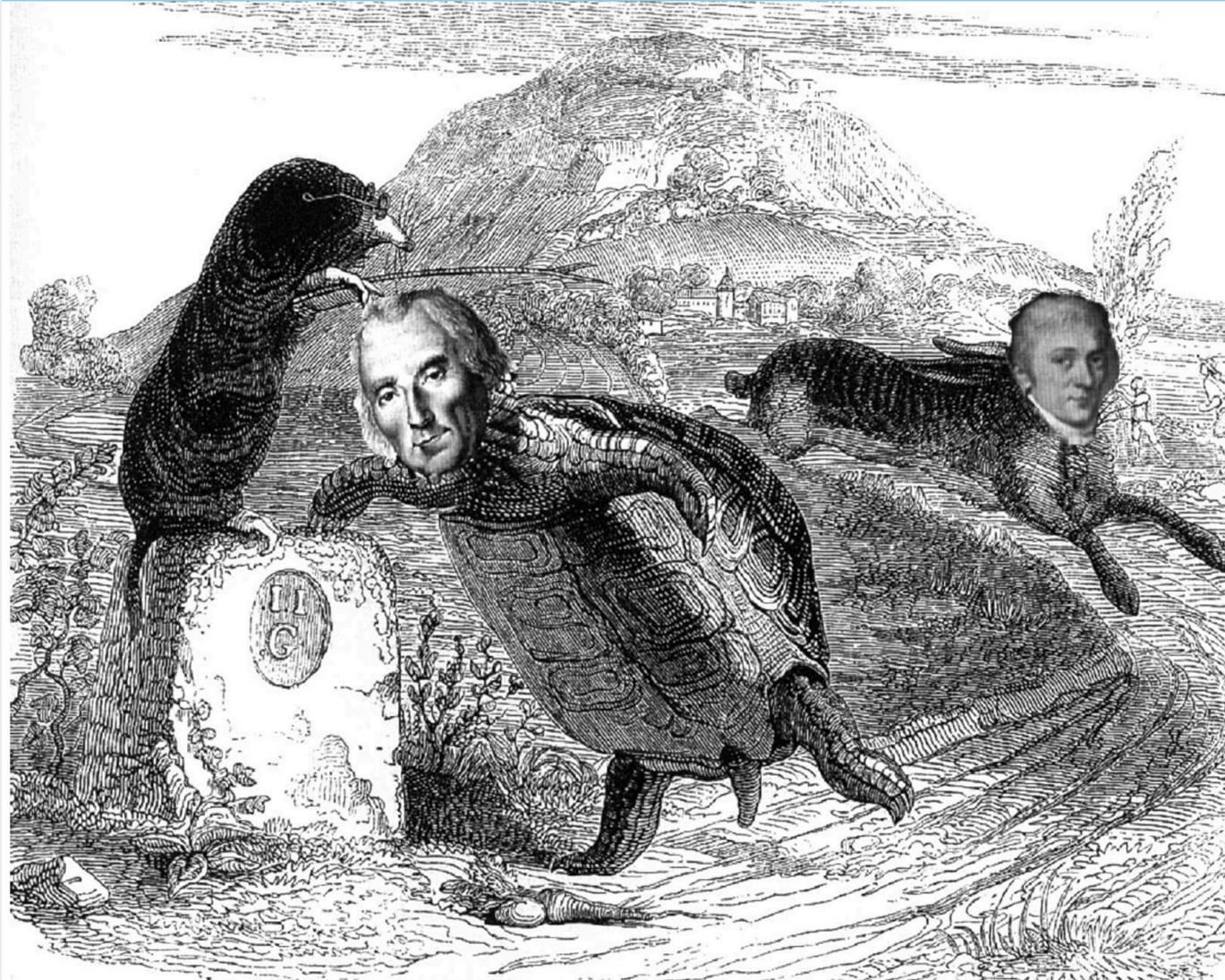
Objeto de desejo: Estimar β_0 e β_1 com base nos dados $(x_1, y_1), \dots, (x_n, y_n)$.

- **Pergunta:** Como estimar os parâmetros β_0 e β_1 do modelo?
- **Resposta:** Que tal minimizar alguma quantidade que indique o agregado dos erros?

Sugestões:

1. Minimize $\sum_{i=1}^n \epsilon_i$. Ruim: erros positivos e negativos se cancelam.
2. Minimize $\sum_{i=1}^n |\epsilon_i|$. Proposta de Laplace: regressão L1.
3. Minimize $\sum_{i=1}^n \epsilon_i^2$. Proposta de Gauss: regressão L2. Método de mínimos quadrados ordinários.

A LEBRE GAUSSIANA E A TARTARUGA LAPLACIANA



Queremos encontrar a reta que erre menos:

$$\begin{pmatrix} \hat{\beta}_0 \\ \hat{\beta}_1 \end{pmatrix} = \underset{\begin{pmatrix} \beta_0 \\ \beta_1 \end{pmatrix} \in \mathcal{R}^2}{\text{Arg Min}} \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2.$$

Equivalentemente:

Equivalentemente, a função objetivo pode ser expressa em termos da raiz do erro quadrático médio (RMSE, Root Mean Square Error):

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2}.$$

- No treinamento (otimização), prefere-se usar o MSE (Mean Squared Error) por razões computacionais (a raiz dificulta a derivação).

Atenção ao contexto!

Função de custo (MSE):

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Usada durante o ajuste do modelo.
- Divisor: n .

Estimativa da variância residual:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Usada em inferência estatística.
- Divisor: graus de liberdade.

Outra confusão comum é associar com "variância amostral"
 $\hat{S}^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$.

$$\begin{aligned}
S(\beta_0, \beta_1) &= \sum_{i=1}^n \epsilon_i^2 = \sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2 \\
&= \sum_{i=1}^n y_i^2 + n\beta_0^2 + \beta_1^2 \sum_{i=1}^n x_i^2 + 2\beta_0\beta_1 \sum_{i=1}^n x_i \\
&\quad - 2\beta_0 \sum_{i=1}^n y_i - 2\beta_1 \sum_{i=1}^n x_i y_i.
\end{aligned}$$

$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_0} = 0$$
$$\frac{\partial S(\beta_0, \beta_1)}{\partial \beta_1} = 0$$

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}$$
$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Uma distinção importante:

- **Métrica**: Usada para avaliar o desempenho dos modelos na tarefa de prever o y .
- **Função de custo** (ou função objetivo): usada para estimar os parâmetros do modelo.

Modelo de Regressão Linear Múltipla (MRLM)

Intuições iniciais

- Cada variável explicativa X_j tem um efeito **linear e aditivo** sobre a variável resposta Y .
- O objetivo é quantificar o efeito de cada X_j sobre Y , **constantes os demais preditores**.

- O modelo permite responder perguntas como:
 - "Qual é o efeito de X_1 sobre Y **com as demais variáveis mantidas constantes?**"
 - "Se aumentarmos X_2 em uma unidade, quanto esperamos que Y mude em média?"
- A linearidade significa que o efeito de cada variável é:
 - **Constante:** não depende do valor de X .
 - **Independente:** o efeito de uma variável não muda com o valor de outra (a menos que sejam incluídas interações).

Definição do modelo

Forma não matricial

O **MRLM** estende o modelo simples para incluir múltiplas variáveis explicativas.

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i$$

- Y_i : variável resposta
- X_{ji} : j-ésima variável explicativa para a i-ésima observação
- β_0 : intercepto
- β_j : coeficiente da j-ésima variável explicativa
- ε_i : erro aleatório associado, com $E(\varepsilon_i) = 0$ e $Var(\varepsilon_i) = \sigma^2$

Forma matricial

A forma matricial do MRLM é:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

em que:

- \mathbf{Y} é um vetor $n \times 1$ da variável resposta.
- \mathbf{X} é a matriz $n \times (p + 1)$ de variáveis explicativas, incluindo a coluna de 1's para o intercepto.
- $\boldsymbol{\beta}$ é um vetor $(p + 1) \times 1$ dos parâmetros desconhecidos.
- $\boldsymbol{\varepsilon}$ é um vetor $n \times 1$ dos termos de erro.

Forma matricial explícita

$$\mathbf{Y} = \begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix}, \quad \mathbf{X} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix}$$

$$\boldsymbol{\beta} = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix}, \quad \boldsymbol{\varepsilon} = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Usando os componentes matriciais, o modelo pode ser expresso como

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{p1} \\ 1 & X_{12} & X_{22} & \cdots & X_{p2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{pn} \end{bmatrix} \cdot \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}.$$

Ou seja, para cada observação $i = 1, 2, \dots, n$:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_p X_{pi} + \varepsilon_i.$$

Suposições do modelo

1. Especificação correta do modelo
2. Média dos erros igual a zero: $E(\varepsilon_i) = 0, \forall i, i = 1, \dots, n;$
3. Homoscedasticidade: $Var(\varepsilon_i) = \sigma^2, \forall i, i = 1, \dots, n;$
4. Independência dos erros: $Cov(\varepsilon_i, \varepsilon_j) = 0, \forall i \neq j;$
5. Ausência de multicolinearidade perfeita: As variáveis explicativas não são combinações lineares exatas entre si.
6. Normalidade dos erros: (necessária para inferência) $\varepsilon_i \sim N(0, \sigma^2), \forall i, i = 1, \dots, n;$

Decorre da suposição segundo a qual $E(\varepsilon) = 0$, que o modelo a ser estimado é:

$$E(\mathbf{Y}|\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}.$$

Ou seja, assim como no modelo de regressão linear simples, o que explicamos é o comportamento da **média de \mathbf{Y} condicional à matriz de regressores \mathbf{X}** .

Isso significa que $\mathbf{X}\boldsymbol{\beta}$ representa a **parte sistemática**, enquanto ε captura as variações aleatórias não explicadas pelo modelo.

Função de Custo: Erro Quadrático Médio (MSE)

Queremos encontrar os coeficientes que **minimizam o erro médio entre os valores observados e os preditos**:

$$J(\boldsymbol{\beta}) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$$

- Essa é a **função de custo** utilizada para estimar os parâmetros β .
- Também chamada de **Erro Quadrático Médio** (*Mean Squared Error, MSE*).
- No treinamento (otimização), prefere-se usar o EQM por razões computacionais ao invés da Raiz do Erro Quadrático Médio (a raiz dificulta a derivação).

Atenção ao contexto!

Função de custo (MSE):

$$J(\beta) = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Usada durante o ajuste do modelo.
- Divisor: n .

Estimativa da variância residual:

$$\hat{\sigma}^2 = \frac{1}{n - p - 1} \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

- Usada em inferência estatística.
- Divisor: graus de liberdade.

Modelo de Regressão Ridge

| O que significa?

Introduzimos um hiperparâmetro λ que **controla** a **magnitude** dos coeficientes (shrinkage).

| Como usar?

Adicionamos uma **penalização L2** à função de custo do modelo linear:
 $\lambda \sum_{j=1}^p \beta_j^2$ (sem penalizar o intercepto).

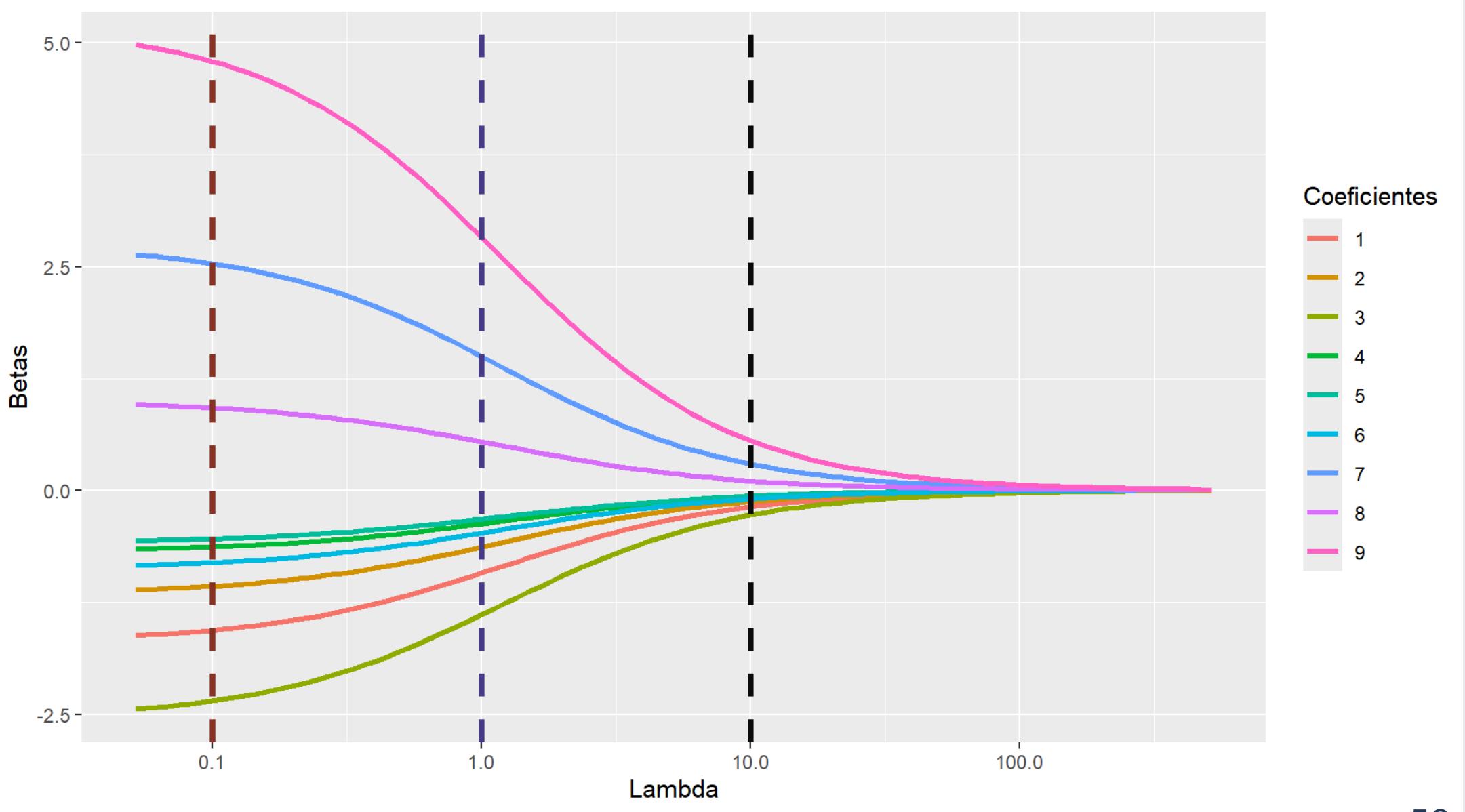
| Seleção de variáveis?

Ridge não realiza seleção de variáveis: em geral **não zera** coeficientes; apenas os **encolhe**.

A nova função de custo passa a ser

$$\text{MSE}_{Ridge} = \text{MSE} + \lambda \sum_{j=1}^p \beta_j^2$$

- Quanto maior λ , **maior a penalização e menores os β_j .**
- No limite, coeficientes de variáveis pouco contributivas podem ficar **muito pequenos, mas não chegam a zero.**
- **Existe** um λ ótimo (escolhido por validação cruzada) que minimiza o erro de predição.
- Penalização aplicada aos coeficientes β_1, \dots, β_p (**intercepto não penalizado**).



Modelo de Regressão Lasso

| O que significa?

Introduzimos um hiperparâmetro λ que controla a magnitude dos coeficientes
e pode induzir esparsidade.

| Como usar?

Adicionamos uma **penalização L1**: $\lambda \sum_{j=1}^p |\beta_j|$ (sem penalizar o intercepto).

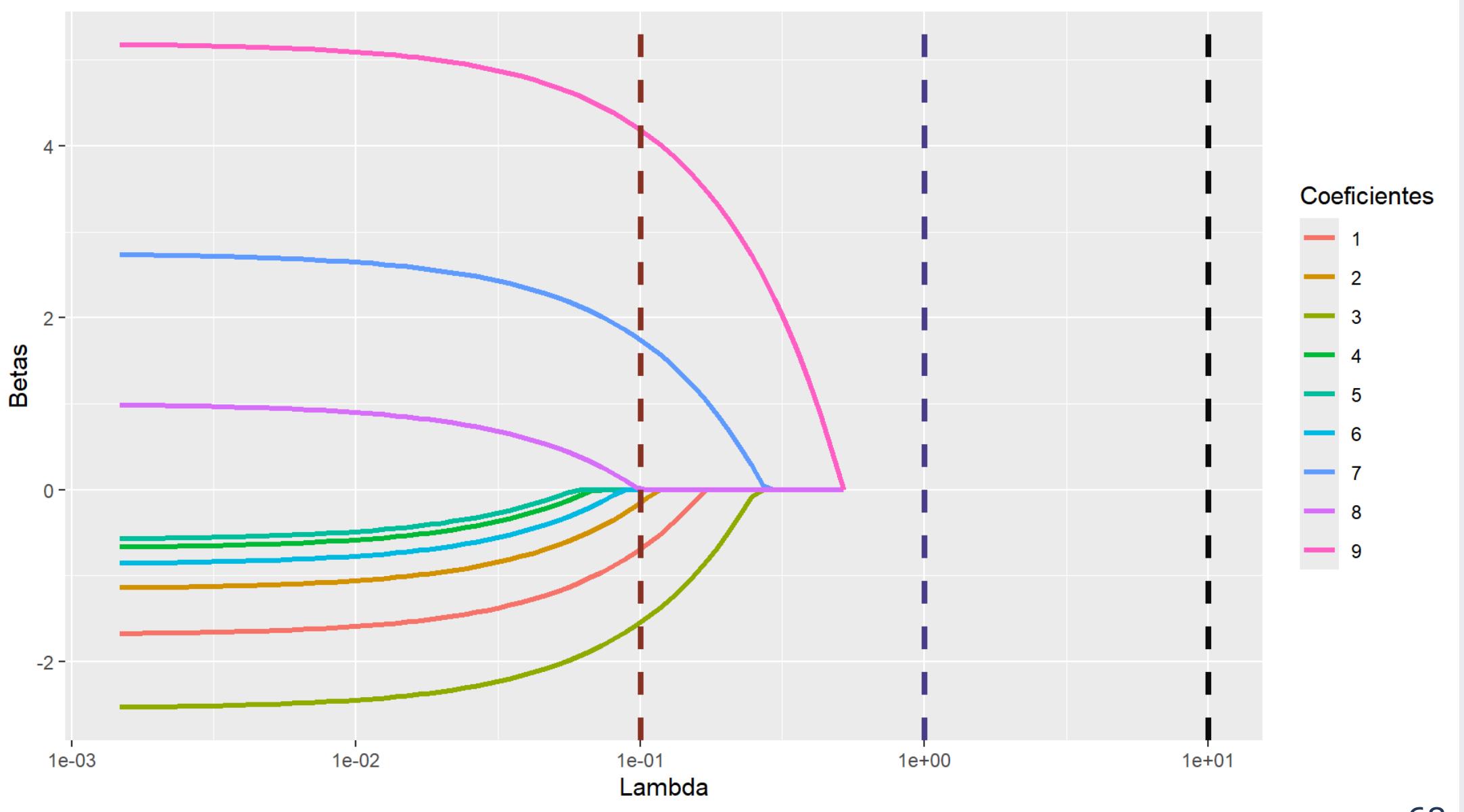
| Seleção de variáveis?

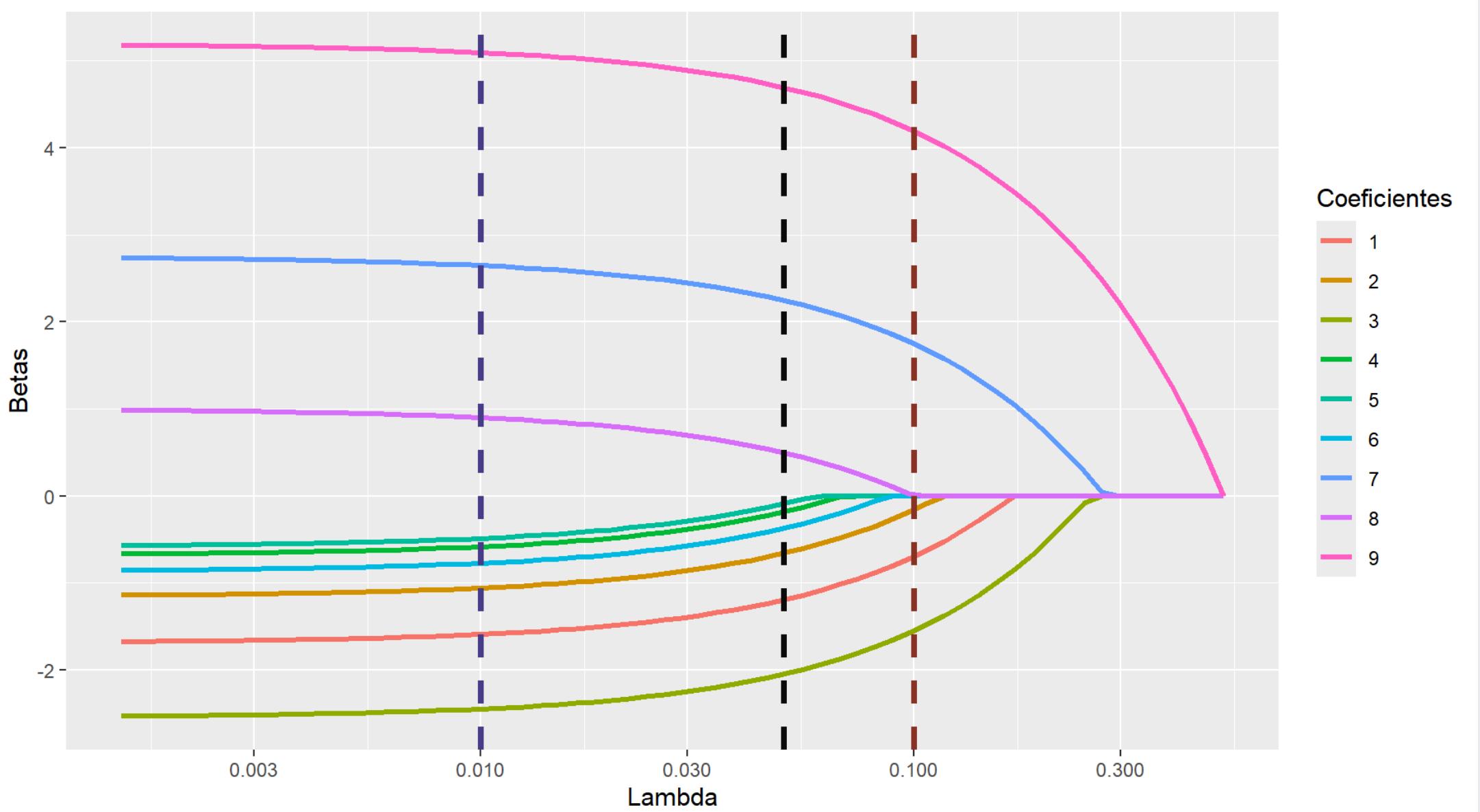
Lasso pode zerar coeficientes, funcionando como seleção de variáveis
(dependente de λ).

A nova função de custo passa a ser

$$\text{MSE}_{Lasso} = \text{MSE} + \lambda \sum_{j=1}^p |\beta_j|$$

- Quanto maior λ , maior a penalização e menor a magnitude dos β_j 's.
- **Diferente do Ridge**, o Lasso pode **zerar** coeficientes de variáveis pouco relevantes.
- **Existe** um λ ótimo (via validação cruzada) que equilibra viés-variância.
- Penalização aplicada aos coeficientes β_1, \dots, β_p (**intercepto não penalizado**).





Referências

- **Regularization and Variable Selection via the Elastic Net**
- **Regression Shrinkage and Selection via the Lasso**
- **Ridge Regression: Biased Estimation for Nonorthogonal Problems**
- **A Survey of Cross-Validation Procedures for Model Selection**

- **The Predictive Sample Reuse Method with Applications**
- **Estimating the Error Rate of a Prediction Rule: Improvement on Cross-Validation**
- **A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation**
- **Some Studies in Machine Learning Using the Game of Checkers**

- Computing Machinery and Intelligence
- Machine Learning: an introduction
- An Introduction to Statistical Learning
- Aprendizado de máquina: uma abordagem estatística
- Materiais Curso R

Não deixe de entrar em contato comigo para tirar suas dúvidas:
thiagoan.andrade@gmail.com

Estamos no  @thiagoan.andrade para networking e socializações

Obrigado!

Thanks!