# Mineração de Dados

# Aula 3: Mais sobre Regressão Linear e Seleção de Modelos

Rafael Izbicki

**Resumo** Vimos que o objetivo de um problema de predição é, com base em um conjunto de treinamento, estimar $g(x)$ tal que, dadas as covariáveis de uma nova observação, $x$, $g(x)$ seja próxima à resposta desta nova observação, $y$.

Tal objetivo pode ser traduzido como encontrar $g$ tal que $R(g) = \mathbb{E}[(g(X) - Y)^2)]$ seja baixo.

A solução para tal problema é dada por $g(x) = \mathbb{E}[Y|x]$, ou seja, a melhor solução é dada pela função de regressão de $Y$ em $x$. Como tal função é desconhecida, podemos estimá-la.

Uma maneira de se fazer isso é através de uma regressão linear. Note que não estamos assumindo que a verdadeira regressão é linear, estamos apenas usando tal método para criar um preditor de $Y$.

Estudamos também o problema de seleção de modelos: Seja $\mathbb{G}$ um conjunto de funções $g(x)$. Ex:

$$\begin{aligned}\mathbb{G} = \{&g(x) = \widehat{\beta}_0,\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x,\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2, \ \ldots\\ &g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x + \widehat{\beta}_2 x^2 + \ldots + \widehat{\beta}_{30} x^{30}\}\end{aligned}$$

**Como escolher $g \in \mathbb{G}$ com menor risco $R(g)$?**

**Como escolher $g \in \mathbb{G}$ com menor risco $R(g)$?**

Para resolver tal problema é necessário estimar $R(g)$. O erro quadrático médio amostral $EQM(g) := n^{-1} \sum_{i=1}^{n} (Y_i - g(X_i))^2$ é um estimador extremamente viesado $R(g)$: em geral ele subestima tal quantidade. Selecionar modelos com base no $EQM$ leva ao overfitting.

Vimos duas alternativas para estimar $R(g)$:

- penalização: $R(g) \approx EQM(g) + \mathcal{P}(g)$
- validação cruzada: dividir a amostra em treinamento vs validação; estimar $R(g)$ usando somente o conjunto de validação

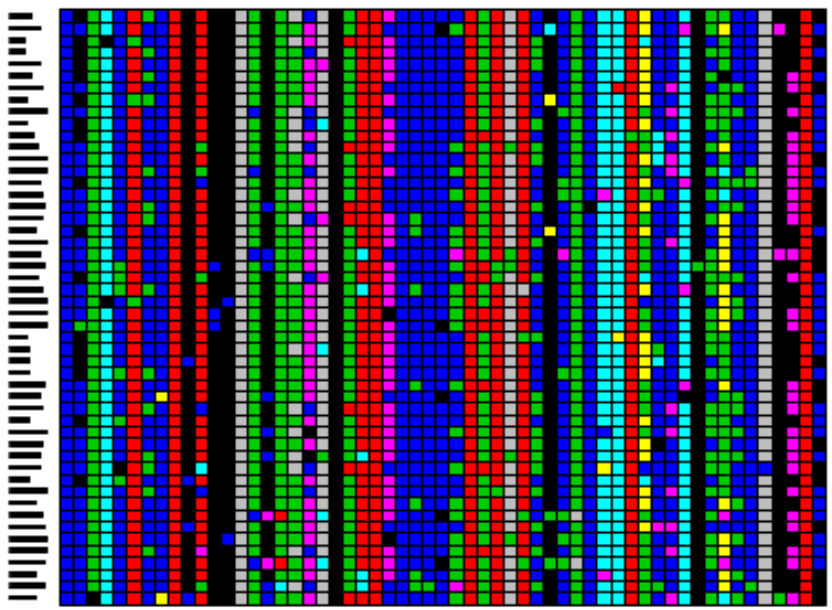
**Nesta Aula:** Vamos também estudar como a metodologia de penalização é usada no caso de regressão linear múltipla, e como ela motiva uma alternativa ao estimador de mínimos quadrados usada em regressão moderna.

# Mais sobre Seleção de Modelos

$Y =$ HIV resistance

$X_j$ = amino acid in position $j$ of the virus.

$$Y = \beta_0 + \beta_1 X_1 + \cdots + \beta_{100} X_{100} + \epsilon$$



Suponhamos que há $p$ covariáveis $(x_1, \ldots, x_p)$ e que desejamos fazer uma regressão linear.

Podemos buscar o melhor modelo dentre

$$
\begin{aligned}
\mathbb{G} = \{ & g(x) = \widehat{\beta}_0, \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1, \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_2, \\
& \ldots \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_p, \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_1 x_2, \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_1 x_3, \\
& \ldots \\
& g(x) = \widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \widehat{\beta}_2 x_2 + \ldots + \widehat{\beta}_p x_p \}
\end{aligned}
$$

Porque não usar todas as covariáveis?

(1) Várias variáveis não importam (ou importam pouco)

(2) Overfitting

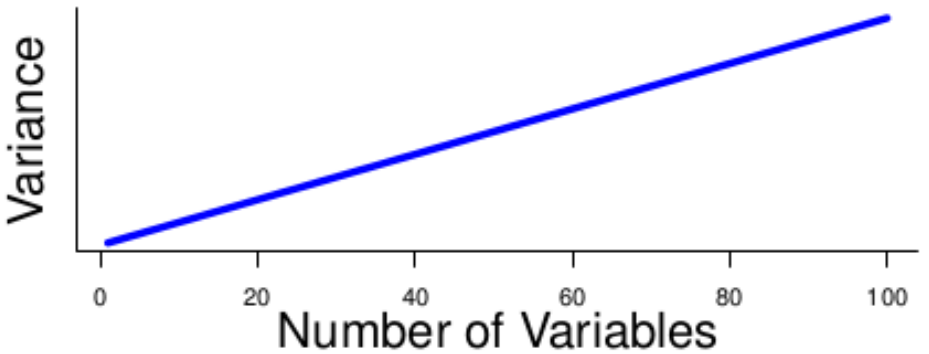
(3) Muitos coeficientes para se estimar: tarefa difícil de se fazer com uma amostra pequena

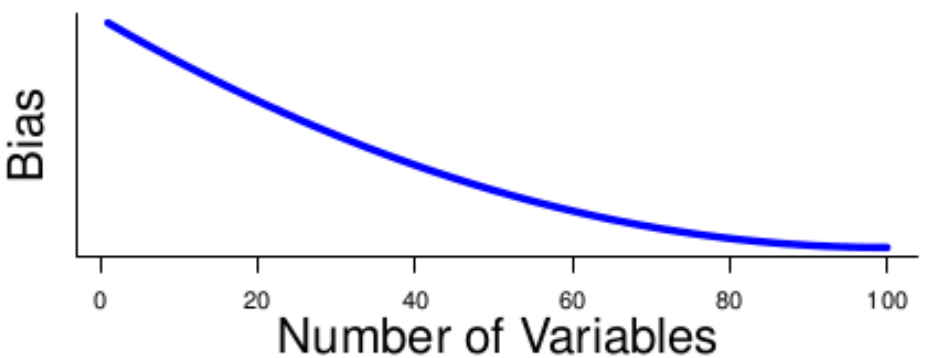
risco preditivo $=$ viés$^2$ $+$ variância $+$ erro inevitável

Viés: erro por omitir variáveis importantes

Variância: error por ter que estimar muitos parâmetros (tarefa difícil)

# The Bias-Variance Tradeoff



Queremos então selecionar um dentre todos os modelos disponíveis

Há $2^p$ (porquê?) modelos!!

Se $p = 30$, são 1.073.741.824 modelos!!

Se $p = 100$, são mais modelos que átomos no universo!!

Apesar de, em teoria, podermos estimar o risco de cada modelo usando penalização (AIC) ou validação cruzada, há tantos modelos que computacionalmente isso é impossível (levaria muito tempo).

Vamos ver duas soluções para resolver esse problema: (i) stepwise regression (ii) lasso

## Notação

$S$: subconjunto de covariáveis. Ex:

$$S = \{x_1, x_4\}, \quad S = \{x_{12}\}, \quad S = \{\}$$

$g_S$: modelo ajustado utilizando-se as covariáveis $S$ (i.e., mínimos quadrados utilizando-se as covariáveis $S$; note que na realidade $g_S = g_{\beta,S}$)

O objetivo dos métodos de seleção de modelos vistos nas aulas anteriores é encontrar a solução para

$$\arg\min_S \widehat{R}(g_S),$$

onde $\widehat{R}(g_S)$ é uma estimativa do risco de $g_S$ (vimos como fazer isso usando AIC e validação cruzada).

Como há $2^p$ subconjuntos $S$, é muito demorado achar este mínimo.

Alternativa 1: usamos uma *heurística* para achar o melhor $S$ (ou ao menos algo próximo ao melhor).

Exemplo: forward stepwise selection:

1. Para $j = 1, \ldots, p$, ajuste a regressão de $Y$ na $j$-ésima variável $X_j$. Seja $\widehat{R}(g_j)$ o risco estimado desta função (usando AIC ou validação cruzada). Defina $\widehat{j} = argmin_j \widehat{R}(g_j)$

2. Para cada $j \in S^c$, ajuste a regressão de $Y = \beta_j X_j + \sum_{s \in S} \beta_s X_s + \epsilon$, e seja $\widehat{R}(g_j)$ o risco estimado desta função (usando AIC ou validação cruzada). Defina $\widehat{j} = argmin_{j \in S^c} \widehat{R}(g_j)$ e atualize $S \longleftarrow S \cup \widehat{j}$

3. Repita o passo anterior até que todas as variáveis estejam em $S$ ou até que não seja possível mais ajustar a regressão

4. Selecione o modelo com menor risco estimado.

Se $p = 20$, são 1.048.576 submodelos, mas apenas 211 (=1+p(p+1)/2)) são contemplados no forward stepwise .

Existem outras abordagens, e.g., backwards stepwise, forward-backwards stepwise etc.

Estas técnicas são mais rápidas que simplesmente estimar o risco de todos os subconjuntos $S$, mas ainda assim podem ser lentas se $p$ e grande.

Além disso, não há garantias de que o melhor modelo será encontrado.

## Segunda abordagem: Penalização esperta.

Se usamos o critério AIC para estimar o risco, estamos buscando

$$\arg\min_S \widehat{R}(g_S) = \arg\min_S EQM(g_S) + K_1 \times p_S,$$

ou, mais precisamente

$$\arg\min_{\beta,S} \widehat{R}(g_{\beta,S}) = \arg\min_{\beta,S} EQM(g_{\beta,S}) + K_1 \times p_{\beta,S},$$

em que $p_{\beta,S} = \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)$ é o número de covariáveis em $S$ e $K_1$ é uma constante que não depende de $p$ (Aula 2).

Pode-se mostrar que esse problema pode ser reescrito como

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2 \text{ sujeito a } \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) \leq K_2,$$

em que $K_2$ é uma constante que depende de $K_1$.

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2 \text{ sujeito a } \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) \leq K_2,$$

**Interpretação**. $K_2$ pequeno: soluções com muitos zeros; $K_2$ grande: soluções com poucos zeros

Como vimos, é difícil resolver esse problemas pois temos fazer uma busca $2^p$ modelos.

Uma abordagem para resolver isso é mudar a forma de se fazer a penalização.

# O método do Lasso

$$\arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2 \text{ sujeito a } \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) \leq K_2,$$

O termo $\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)$ mede o quão *esparsa* é nossa solução.

Ex:

$\beta = (1, 5, 6, -1, -4, 0, 0, 0, 0, 0, 0, 0, 0)$ tem $\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) = 5$, e

$\beta = (1, 5, 6, -1, -4, 1, 4, 1, 1, 5, 8, 12, 1)$ tem $\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0) = 13$.

Outra maneira de se fazer isso é usando a norma L1: $\sum_{j=1}^{p} |\beta_j|$

Ex:

$\beta = (1, 5, 6, -1, -4, 0, 0, 0, 0, 0, 0, 0, 0)$ tem $\sum_{j=1}^{p} |\beta_j| = 17$, e

$\beta = (1, 5, 6, -1, -4, 1, 4, 1, 1, 5, 8, 12, 1)$ tem $\sum_{j=1}^{p} |\beta_j| = 50$.

Assim como $\sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0)$, $\sum_{j=1}^{p} |\beta_j|$ também quantifica quantos coeficientes "grandes" temos. Vetores com muitos coeficientes grandes tem $\sum_{j=1}^{p} |\beta_j|$ alto; vetores com poucos coeficientes grandes tem $\sum_{j=1}^{p} |\beta_j|$ baixo.

No lasso, ao invés de buscarmos

$$\arg\min_{\beta} \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{i,1}-\ldots-\beta_p x_{i,p})^2 \text{ sujeito a } \sum_{j=1}^{p}\mathbb{I}(\beta_j \neq 0) \leq K_2,$$

buscamos por

$$\arg\min_{\beta} \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{i,1}-\ldots-\beta_p x_{i,p})^2 \text{ sujeito a } \sum_{j=1}^{p}|\beta_j| \leq K_2,$$

**Interpretação**. $K_2$ pequeno: soluções com muitos zeros; $K_2$ grande: soluções com poucos zeros

**Observações**: (i) a solução do lasso é fácil de ser encontrada, (ii) ela em geral possui muitos zeros. A prova dessas afirmações vai além do escopo deste curso.

O problema do lasso pode ser reescrito como

$$\arg\min_{\beta} \sum_{i=1}^{n}(y_i-\beta_0-\beta_1 x_{i,1}-\ldots-\beta_p x_{i,p})^2 + \lambda\sum_{j=1}^{p}|\beta_j|$$

Como escolher $\lambda$? $\lambda$ alto: penalização alta; $\lambda$ baixo: penalização baixa

Validação cruzada!

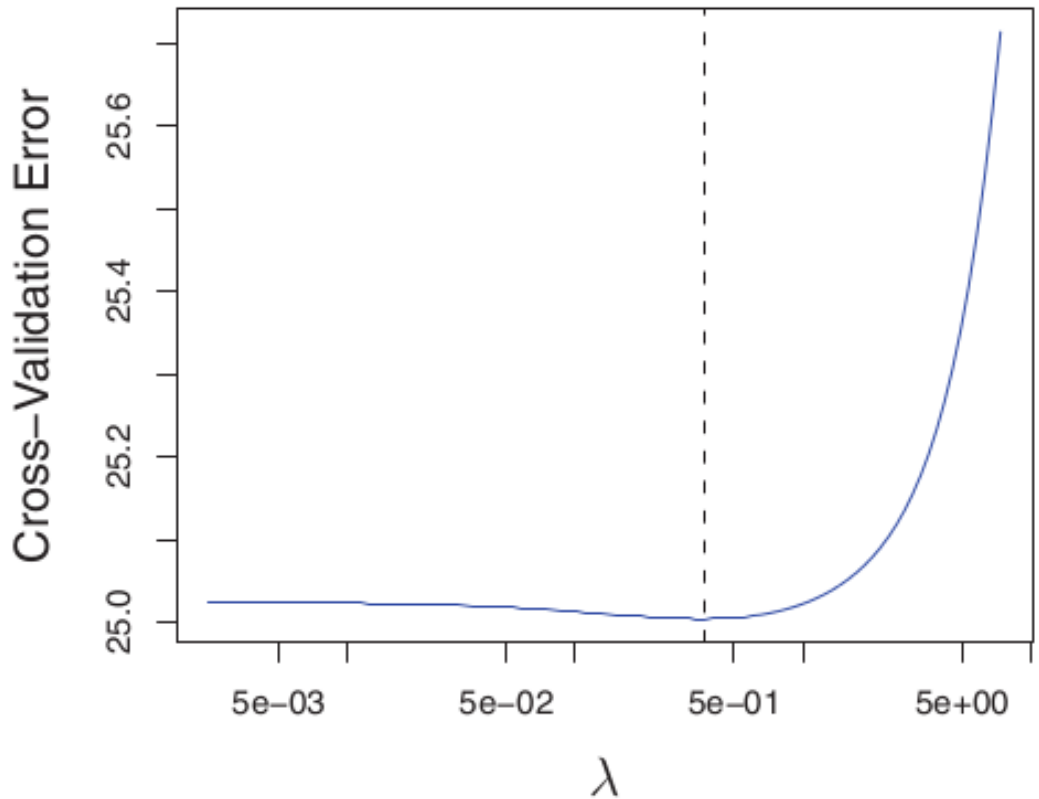
Para cada $\lambda$, buscamos

$$\beta^{\lambda} \equiv \arg\min_{\beta} \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_{i,1} - \ldots - \beta_p x_{i,p})^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Em seguida, buscamos

$$\arg\min_{\beta^{\lambda}} \widehat{R}(\beta^{\lambda})$$

$\lambda = 0$....


Cross−Validation Error
25.0
25.2
25.4
25.6
5e−03
5e−02
5e−01
5e+00
λ

**Comentários adicionais:**

Uma grande vantagem do lasso: funciona quando $p > n$

Muitas vezes o lasso funciona melhor se as covariáveis são normalizadas para ter média 0 e variância 1

Note que ao fazer o lasso, estamos aumentando o viés do estimador dado por mínimos quadrados (pois estamos "jogando fora" algumas covariáveis), por outro lado, estamos diminuindo a variância dele.

risco preditivo $=$ viés$^2$ $+$ variância $+$ erro inevitável

# Resumindo

Em muitos problemas reais temos várias covariáveis.

Um modelo de regressão linear simples que usa todas elas pode ter performance preditiva muito ruim, pois com ele é necessário estimar muitos coeficientes, além de que muitas vezes certas covariáveis influenciam muito pouco na variável resposta.

Devemos então selecionar qual o melhor subconjunto de covariáveis para ser utilizado.

Como visto nas aulas anteriores, podemos, para cada modelo, estimar o risco $R(g)$, e buscar aquele com menor risco.

Quando há muitas covariáveis, esse procedimento leva tempo demais.

# Resumindo

Vimos duas alternativas para contornar esse problema.

Método 1: Usamos uma heurística para buscar o melhor modelo. Aqui, estudamos o forward stepwise, no qual começamos avaliando o risco do modelo que só tem intercepto, e então passamos a tentar inserir apenar uma covariável por vez. Buscamos sempre a variável que minimiza o risco estimado.

# Resumindo

Método 2: Mudamos a penalização usada. Aqui vimos o lasso, que ao invés de buscar

$$\arg\min_{\beta,S} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1 - \ldots - \beta_p x_p)^2 + \lambda \sum_{j=1}^{p} \mathbb{I}(\beta_j \neq 0),$$

buscamos por

$$\arg\min_{\beta} \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_1 - \ldots - \beta_p x_p)^2 + \lambda \sum_{j=1}^{p} |\beta_j|$$

Rápido de ser resolvido; funciona para $p > n$; também induz vários 0's na solução (interpretabilidade, fácil de ser usado na prático)

$\lambda$ em geral é escolhido por validação cruzada.

# Aplicação

Dados simulados: $Y = X_1 + X_2 + X_3 + X_4 + X_5 + \epsilon$.

Além de $x_1, \ldots, x_5$, observamos mais 15 variáveis não relacionadas a $y$.

# Resultados:

## Todos os subconjuntos:

Tempo: 1 hora e 20 minutos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

## Forward stepwise:

Tempo: 0.46 segundos; $\widehat{R}(g^*) = 0.30$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5, x_{10}, x_{12}, x_{13}, x_{19}, x_{20}$

## Lasso:

Tempo: 0.09 segundos; $\widehat{R}(g^*) = 0.25$

Covariáveis selecionadas: $x_1, x_2, x_3, x_4, x_5$

# No R

Todos os subconjuntos: `library (leaps)`

```
modelo = regsubsets("y ~ x", data=dados)
```

Forward stepwise:

```
modelo = regsubsets("y ~ x", data=dados,
nvmax=20 ,method="forward")
```

Lasso: `library(glmnet)`

```
lasso.mod = glmnet(x,  y, alpha =1)
cv.out = cv.glmnet(x,  y, alpha =1)
plot(cv.out)
bestlam = cv.out$lambda.min
```

Mais detalhes: ISLR, capítulos 6.5.1, 6.5.2 e 6.6.2