Seleção de modelos

Thomas Peron

thomas.peron@usp.br
Sala 3-250 B

Problema: Explicar Y em termos de X_0, X_1, X_2, ..., X_{p-1}.

Quais covariáveis escolher?

$$Y = \beta_0 + \beta_1 X_1 + \varepsilon$$

$$Y = \beta_0 + \beta_1 X_1 + \beta_3 X_3 + \varepsilon$$

Seleção de variáveis

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_4 X_4 + \varepsilon$$

$$\vdots$$

$$Y = \beta_0 + \beta_1 X_1^2 + \beta_2 X_2^2 + \beta_{1,2} X_1 X_2 + \varepsilon$$

Seleção de modelos

$$\vdots$$

Cada escolha muda a matriz $\mathbb{X}$

X_1	X_2	X_3	$\cdots$	X_{p-1}	Y
80	32	70		13	69
3	80	7		54	38
29	90	10		61	58
86	22	48		32	35
98	70	31		90	8
83	86	7		0	75
20	12	70		77	18
18	76	39		21	21
73	62	71	$\cdots$	33	0
10	69	26		1	76
91	92	75		42	37
42	48	27		89	25
20	80	29		18	22
0	20	18		90	75
97	39	64		58	39

Generalizações e otimismo do modelo

Mínimos quadrados:

$$\vec{\hat{\beta}} = \arg\min_{\vec{\beta}} (\vec{Y} - \mathbb{X}\vec{\beta})^T (\vec{Y} - \mathbb{X}\vec{\beta})$$

Erro associado a um ponto novo $(X_1, X_2, ..., X_{p-1})$:

$$\mathbb{E}\left[\left(Y - \left(\hat{\beta}_0 + \sum_{j=1}^{p-1} X_j \hat{\beta}_j\right)\right)^2\right] \qquad (\hat{\beta}_0 \text{ e } \hat{\beta}_j \text{ estimados com conj. de treinamento})$$

Há um estimador?

Treino e teste

$\vec{Y}' \equiv$ dados novos (teste, *out-of-sample*).

Valor esperado do erro do conjunto de teste:

$$\mathbb{E}\left[\frac{1}{n}(\vec{Y}' - \mathbb{X}\vec{\hat{\beta}})^T(\vec{Y}' - \mathbb{X}\vec{\hat{\beta}})\right] \qquad (\vec{\hat{\beta}} \text{ estimado com conj. treinamento})$$

Ou

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{Y}_i)^2\right]$$

Lembrando que

$$\text{MSE}_{\text{treino}} = \frac{1}{n-p}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 = \frac{SSE}{n-p}$$

Treino e teste

$Y' \equiv$ dados novos

$\hat{Y} \equiv$ modelo ajustado (treino)

$$\begin{aligned}\mathbb{E}\left[(Y_i' - \hat{Y}_i)^2\right] &= \mathrm{Var}[Y_i' - \hat{Y}_i] + (\mathbb{E}[Y_i - \hat{Y}_i])^2 \\ &= \mathrm{Var}[Y_i'] + \mathrm{Var}[\hat{Y}_i] - 2\mathrm{cov}[Y_i', \hat{Y}_i] + (\mathbb{E}[Y_i] - \mathbb{E}[\hat{Y}_i])^2\end{aligned}$$

Mas $\mathbb{E}[Y_i'] = \mathbb{E}[Y_i]$ e $\mathrm{Var}[Y_i'] = \mathrm{Var}[Y_i]$:

$$\begin{aligned}\mathbb{E}\left[(Y_i' - \hat{Y}_i)^2\right] &= \mathrm{Var}[Y_i] + \mathrm{Var}[\hat{Y}_i] + (\mathbb{E}[Y_i] - \mathbb{E}[\hat{Y}_i])^2 \\ &= \mathbb{E}\left[(Y_i - \hat{Y}_i)^2\right] + 2\mathrm{cov}[Y_i, \hat{Y}_i]\end{aligned}$$

Tomando a média sobre todos dados:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{Y}_i)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right] + \frac{2}{n}\sum_{i=1}^{n}\mathrm{cov}[Y_i, \hat{Y}_i]$$

$\sigma^2 H_{ii}$ (Primeiras listas)

Treino e teste

$Y' \equiv$ dados novos

$\hat{Y} \equiv$ modelo ajustado (treino)

Logo,

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{Y}_i)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right] + \frac{2}{n}\sigma^2 \mathbf{tr}\mathbf{H}$$

Mas $\mathbf{tr}\mathbf{H} = p + 1$:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{Y}_i)^2\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2\right] + \frac{2}{n}\sigma^2(p+1)$$

Fazemos então:

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i' - \hat{Y}_i)^2\right] \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \frac{2}{n}\sigma^2(p+1)$$

$$\boxed{\text{MSE}_{\text{teste}} \simeq \text{MSE}_{\text{treino}} + \frac{2}{n}\sigma^2(p+1)}$$

Minimizar $\text{MSE}_{\text{treino}}$ ignora $\text{MSE}_{\text{teste}}$.

σ^2?

Estatística C *(Mallow's C statistic)*

$Y' \equiv$ dados novos

$\hat{Y} \equiv$ modelo ajustado (treino)

$$C_p = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \frac{2}{n}\hat{\sigma}^2(p+1)$$

MSE(Modelo completo)

MSE$(X_1, X_2, ..., X_{P-1})$

$$C_p = \text{MSE} + (\text{penalidade})$$

Diferença entre modelos:

$$\Delta C_p = \text{MSE}_1 - \text{MSE}_2 + \frac{2}{n}\hat{\sigma}^2(p_1 - p_2)$$

Versão alternativa:

$$C_p = \frac{n\text{MSE}}{\hat{\sigma}^2} - n + 2p$$

O problema com R^2 e sua versão ajustada

Coeficiente de determinação:

$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} \qquad \text{SSE}_{p+1} < \text{SSE}_p$$

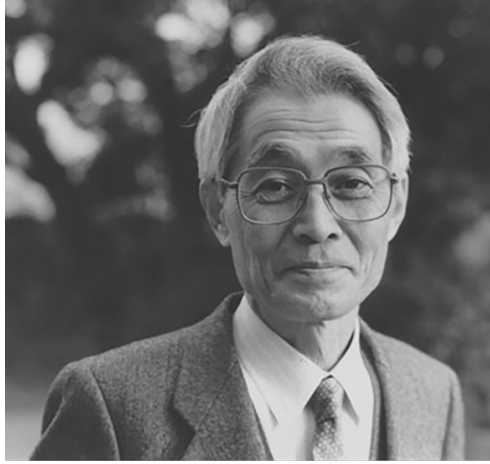
Coeficiente de determinação ajustado:

$$R^2_{\text{adj}} = 1 - \frac{\text{MSE}}{\frac{\text{SST}}{n-1}}$$

Equivalente a minimizar $\text{MSE}_{\text{treino}}$.

Subestima $\text{MSE}_{\text{teste}}$.

Akaike Information Criterion (AIC)



Hirotugu Akaike
(1927-2009)

Coeficiente para um modelo S:

$$\mathrm{AIC}(S) = L_S - \dim(S)$$

$L_S \equiv$ log da verossimilhança do modelo S.

$\dim(S) \equiv$ número de parâmetros ajustáveis.

Para modelos Gaussianos lineares:

$$L_S = -\frac{n}{2}(1 + \log 2\pi) - \frac{n}{2}\log \mathrm{MSE}$$

Comparando modelos de p_1 e p_2 covariáveis:

$$\frac{-2\hat{\sigma}^2}{n}\Delta\mathrm{AIC} \approx \Delta\mathrm{MSE} + \frac{2}{n}\hat{\sigma}^2(p_1 - p_2) = \Delta C_p$$

Resultados semelhantes ao critério C_p

Leave-one-out cross-validation score (LOOCV)

Definição:

$$\text{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left[Y_i - \hat{Y}_i^{(-i)}\right]^2, \qquad \hat{Y}_i^{(-i)} \text{ modelo ajustado sem } i$$

Definição alternativa:

(Identidade matricial de Woodbury & Sherman-Morrison formula)

$$\text{LOOCV} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{Y_i - \hat{Y}_i}{1 - H_{ii}}\right)^2$$

(a.k.a PRESS: "predictive residual sum of squares")

No limite $n \to \infty$:

$$\text{LOOCV} \approx \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2(1 - 2H_{ii}) \approx \text{MSE} + 2\sigma^2(p+1)$$

Métodos passo-a-passo (*stepwise*)

Inicie com o modelo completo $(X_1, X_2, ..., X_{P-1})$.

1. Elimine os coeficientes não significantes.
2. Escolha seu método favorito (C_p, AIC, LOOCV, ...),
   e considere a removação de um coeficiente.
   Escolha o modelo com o melhor valor do critério adotado.

Métodos passo a passo são *greedy*.

Populares, mas não muito eficientes.

Use com cautela, ou não use.

Única justificativa: 2^{p-1} modelos possíveis.

Resumo

C_p e AIC em vez de R^2_{adj} ou apenas $\text{MSE}_{\text{treino}}$.

LOOCV também é estimador não viesado para $\text{MSE}_{\text{teste}}$.

Resultados mostram que:

> $\text{MSE}_{\text{teste}}$ do modelo selecionado por LOOCV é o mais próximo do modelo ideal.

Para $n \to \infty$, C_p, AIC e LOOCV se tornam similares.

C_p e AIC são visto como aproximações rápidas para LOOCV.

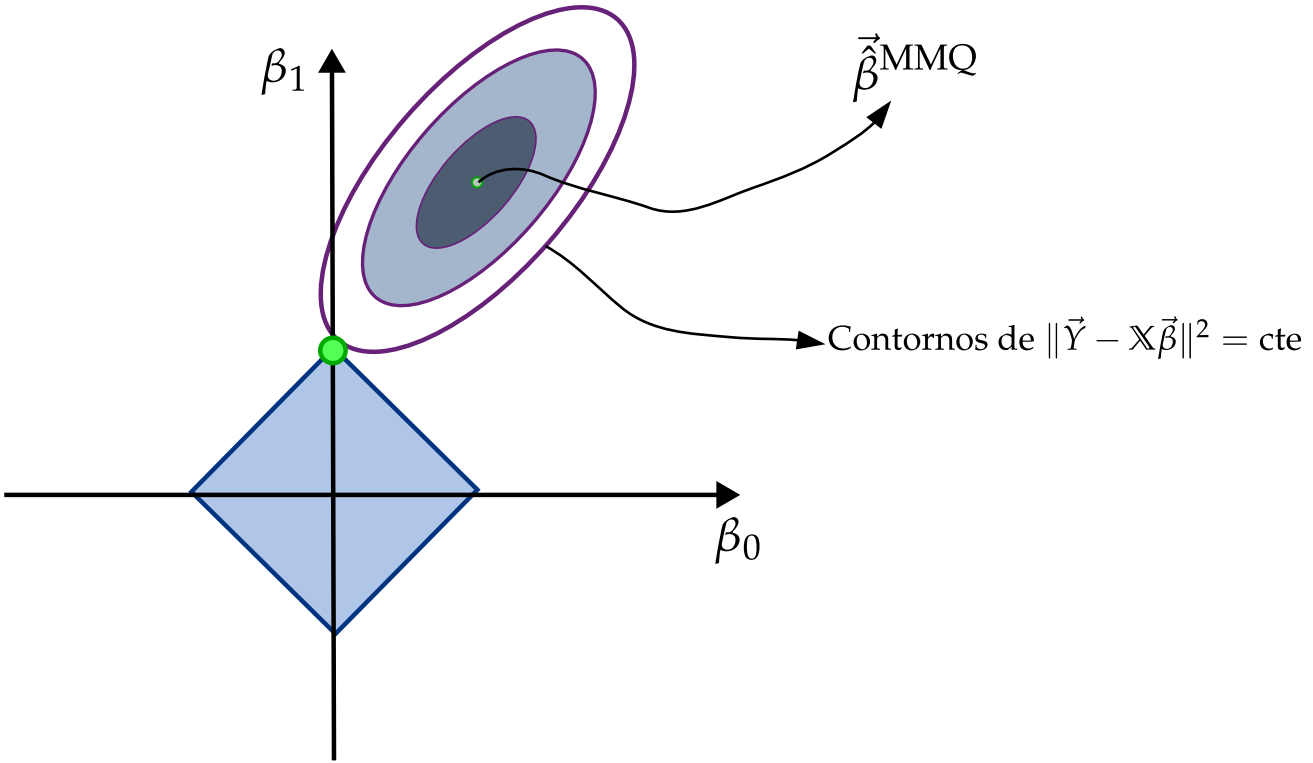
Há teoremas que dizem que:

> Para $n \to \infty$, se o modelo verdadeiro estiver entre os que estiverem sendo comparados, LOOCV, C_p e AIC levarão a modelos com mais covariáveis do que o necessário.

Lasso Regression

Para soluções esparsas:

$$\vec{\hat{\beta}} = \underset{\vec{\beta} \in \mathbb{R}^P}{\arg\min} \left\| \vec{Y} - \mathbb{X}\vec{\beta} \right\|^2 + \lambda \left\| \vec{\beta} \right\|_1$$



Penalizações alternativas

Minimizar:

$$\vec{\hat{\beta}} = \underset{\vec{\beta} \in \mathbb{R}^P}{\arg\min} \left\| \vec{Y} - \mathbb{X}\vec{\beta} \right\|^2 + \lambda \left\| \vec{\beta} \right\|_0$$

Se $\lambda = 2\hat{\sigma}^2$: AIC

Se $\lambda = \hat{\sigma}^2 \log n$: BIC

$$\left\| \vec{\beta} \right\|_0 = \sum_{j=0}^{p} \mathbb{1}\{b_j = 0\}$$

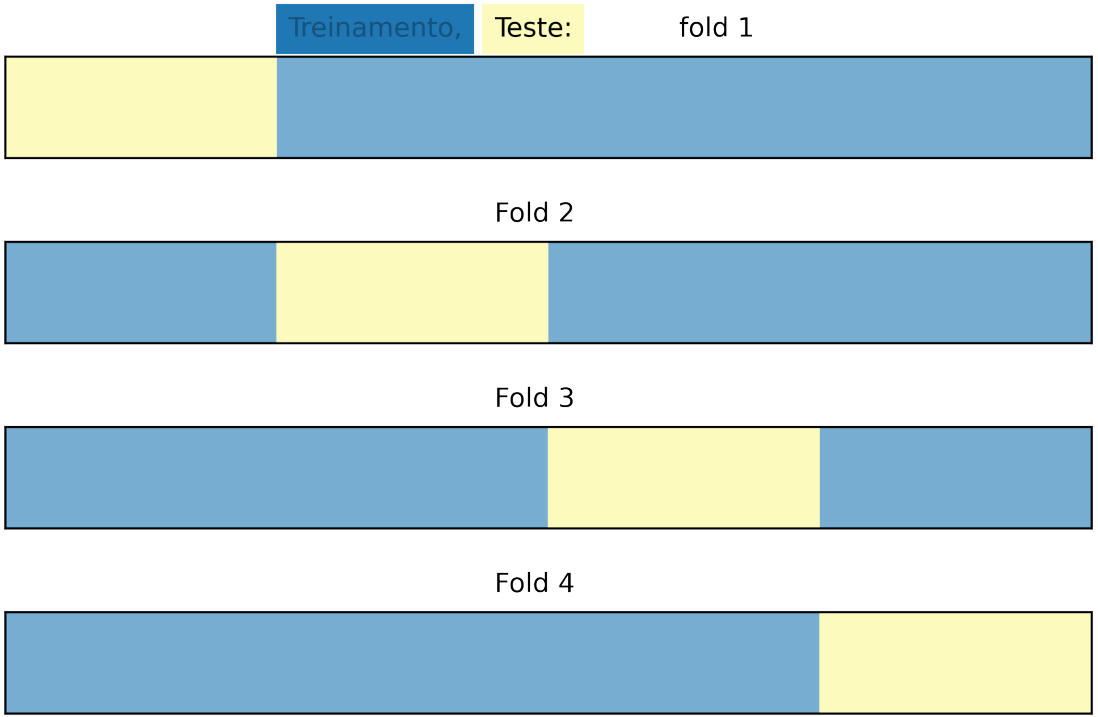
k-Fold Cross-Validation

1. Divida os dados em *k* partes.

2. Para cada parte (*fold*):

    - Para cada *fold*: divida em teste e treino.
    - Ajuste o modelo com o conjunto de treino
    - Calcule o MSE do conjunto de teste.

3. Calcule a média de MSE (teste) sobre todas as *folds.*




```
from sklearn.model_selection import KFold

KFold(n_splits=n_folds, shuffle=False)
```

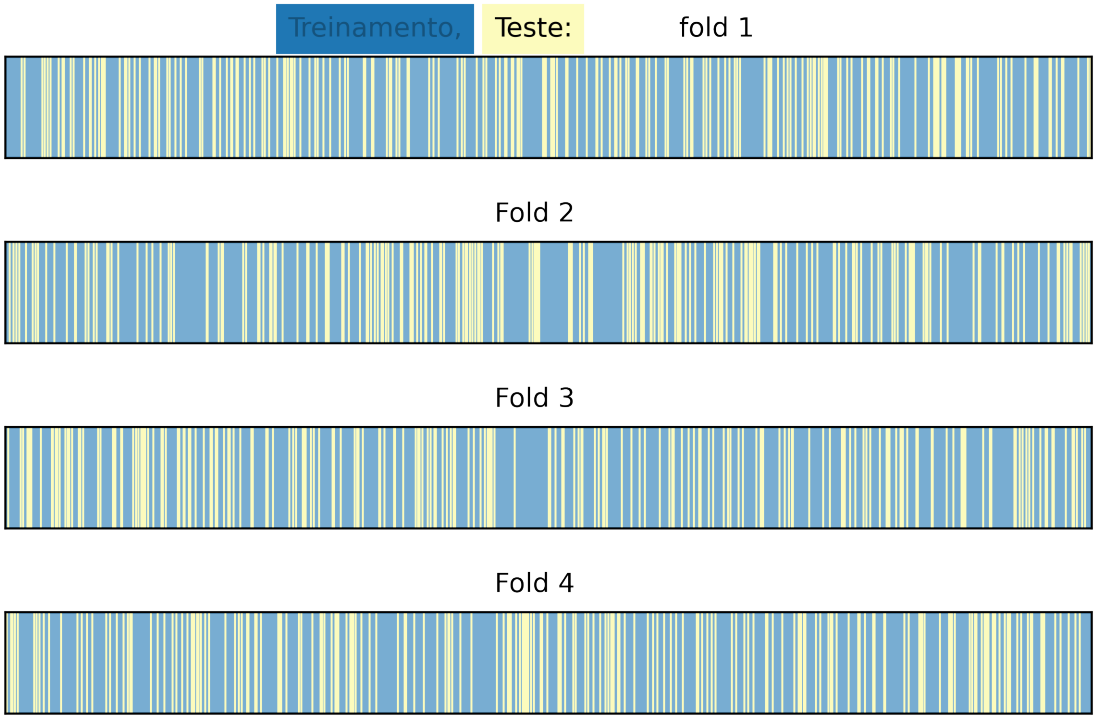
k-Fold Cross-Validation

1. Divida os dados em k partes.

2. Para cada parte (*fold*):

    - Para cada *fold*: divida em teste e treino.
    - Ajuste o modelo com o conjunto de treino
    - Calcule o MSE do conjunto de teste.

3. Calcule a média de MSE (teste) sobre todas as *folds.*




```
from sklearn.model_selection import KFold

KFold(n_splits=n_folds, shuffle=True)
```

Exemplo – *Surgical dataset* (Cap. 9, Kutner)

Explicar Y em termos de $X_1, X_2, ..., X_8$ ($n = 54$ amostras).

p	SSE_p	R^2_p	$R^2_{\text{adj},p}$	C_p	AIC_p	LOOCV_p
1	12.808	0	0	240.452	-75.703	13.296
2	7.332	0.428	0.417	117.409	-103.827	8.025
3	4.312	0.663	0.65	50.472	-130.483	5.065
4	2.843	0.778	0.765	18.914	-150.985	3.469
5	2.179	0.83	0.816	5.751	-163351	2.738
6	2.082	0.837	0.821	5.541	-163.805	2.739
7	2.005	0.843	0.823	5.787	-163.834	2.772
8	1.972	0.846	0.823	7.029	-162.736	2.809
9	1.971	0.846	0.819	9	-160.771	2.931

Exemplo – *Surgical dataset* (Cap. 9, Kutner)

Explicar Y em termos de $X_1, X_2, ..., X_8$ ($n = 54$ amostras).

p	SSE_p	R^2_p	$R^2_{\text{adj},p}$	C_p	AIC_p	LOOCV_p
1	12.808	0	0	240.452	-75.703	13.296
2	7.332	0.428	0.417	117.409	-103.827	8.025
3	4.312	0.663	0.65	50.472	-130.483	5.065
4	2.843	0.778	0.765	18.914	-150.985	3.469
5	2.179	0.83	0.816	5.751	-163351	2.738
6	2.082	0.837	0.821	5.541	-163.805	2.739
7	2.005	0.843	0.823	5.787	-163.834	2.772
8	1.972	0.846	0.823	7.029	-162.736	2.809
9	1.971	0.846	0.819	9	-160.771	2.931

Referências

Stanley H. Chan, Introduction to Probability for Data Science (2021)
https://probability4datascience.com/

Claeskens, Gerda and Nils Lid Hjort (2008). Model Selection and Model Averaging.
Cambridge, England: Cambridge University Press.

Dirk Kroese et al., Data Science and Machine Learning (2022) [Seção 5.3.2].
https://people.smp.uq.edu.au/DirkKroese/DSML/

Rafael Izbicki, Aprendizado de Máquina Estatístico (2020).
http://www.rizbicki.ufscar.br/AME.pdf

Kutner, Montgomery...