Tutorial de Predição Conforme

Helton Graziadei
DEs/UFSCar
helton@ufscar.br

Laboratório de Estatística Aplicada

6 de Dezembro, 2025

Motivação

- Imagine que você trabalhe em uma empresa que opera uma plataforma de anúncios de imóveis online.
- **Objetivo:** estimar o preço de venda do imóvel.
- Implementa-se um modelo de regressão com localização, área, idade, vaga de garagem etc.

Motivação

- O modelo devolve **predições pontuais** de preço.

Motivação

- O modelo devolve **predições pontuais** de preço.

Pergunta: isto é suficiente?

Motivação

A empresa necessita responder perguntas como:

- Se o proprietário anunciar por R$ 520 mil, há risco de estar supervalorizando em relação ao mercado?
- Até quanto pode pedir sem ficar fora da realidade?

Precisamos não só do valor médio, mas de **intervalos de preço** (quantificar a incerteza).

Motivação

- **Quantificação de incerteza**: caracterizar a incerteza das predições.
- Foco: distribuição de $Y \mid \mathbf{x}$ e/ou intervalos de predição.
- Meta: capturar diferentes fontes de incerteza + limitação de informação e obter resultados **calibrados**.

Motivação

Queremos intervalos com alta probabilidade de conter Y_{n+1}.

- Regressão linear (ingênuo):

$$\hat{y}_{n+1} \pm t_{\alpha/2}\, s \sqrt{1 + \frac{1}{n} + \frac{(x_{n+1} - \overline{x})^2}{S_{xx}}}$$

- Válido sob:
  1. Erros i.i.d's com distribuição Normal.
  2. Linearidade.
  3. Homoscedasticidade.

Questão: como obter intervalos de predição sem essas suposições ou resultados assintóticos?

Um pouco de teoria

Definição

A sequência $V_1, V_2, \ldots, V_n$ *de variáveis aleatórias é permutável se, para todo* $\pi = (\pi_1, \ldots, \pi_n) : \{1, \ldots, n\} \to \{1, \ldots, n\}$, *vale:*

$$(V_1, \ldots, V_n) \overset{d}{\sim} (V_{\pi_1}, \ldots, V_{\pi_n}).$$

Permutabilidade é uma condição mais fraca do que i.i.d. (**Exemplo**).

Um pouco de teoria

Exemplo

Considere realizações de $n = 6$ variáveis aleatórias:

V_1	V_2	V_3	V_4	V_5	V_6
4	6	8	4	3	8

Então $\sum_{i=1}^{n} \mathbb{I}(V_i \leq V_{(k)}) \geq k$.

Um pouco de teoria

Resultado

Se $V_1, \ldots, V_n$ são permutáveis, então para todo $i, k = 1, \ldots, n$:

$$P(V_i \leq V_{(k)}) \geq \frac{k}{n}.$$

Se as variáveis V_i têm valores distintos (q.c.):

$$P(V_i \leq V_{(k)}) = \frac{k}{n}.$$

Um pouco de teoria

Duas versões principais:

1. *Full conformal prediction*
2. *Split conformal prediction*

Focaremos em **split conformal**.

Um pouco de teoria

- Modelo pré-treinado $\hat{\mu} : \mathbb{R}^p \to \mathbb{R}$.
- Amostra de calibração $(X_1, Y_1), \dots, (X_n, Y_n)$ de variáveis permutáveis.
- Nível de descobertura $0 < \alpha < 1$.

Objetivo: intervalo para Y_{n+1} usando $\hat{\mu}(X_{n+1})$, com garantias probabilísticas.

Um pouco de teoria

Escores de conformidade:

$$R_i = \left| Y_i - \hat{\mu}(X_i) \right|, \quad i = 1, \ldots, n.$$

- Como a sequência é permutável, os escores também são.
- Ordene:
$$R_{(1)} \leq R_{(2)} \leq \cdots \leq R_{(n)}.$$
- Escolha
$$\hat{r}_\alpha = R_{(\lceil (1-\alpha)(n+1) \rceil)}.$$

Um pouco de teoria

Conjunto conforme:

$$C^{(\alpha)}(X_{n+1}) = \{y \in \mathbb{R} : R_{n+1} \leq \hat{r}_\alpha\}.$$

Cobertura:

$$P\big(Y_{n+1} \in C^{(\alpha)}(X_{n+1})\big) = P(R_{n+1} \leq \hat{r}_\alpha) \geq \frac{\lceil (1-\alpha)(n+1) \rceil}{n+1} \geq 1-\alpha.$$

Se os escores são distintos (sem empates):

$$\boxed{1 - \alpha \leq P\big(Y_{n+1} \in C^{(\alpha)}(X_{n+1})\big) < 1 - \alpha + \frac{1}{n+1}}$$

Um pouco de teoria

Considere $R_i = |Y_i - \hat{\mu}(X_i)|$, $i = 1, \ldots, n+1$, então:

$$
\begin{aligned}
P\big(Y_{n+1} \in C^{(\alpha)}(X_{n+1})\big) &= P(R_{n+1} \leq \hat{r}_\alpha) = P\big(|Y_{n+1} - \hat{\mu}(X_{n+1})| \leq \hat{r}_\alpha\big) \\
&= P\big(\hat{\mu}(X_{n+1}) - \hat{r}_\alpha \leq Y_{n+1} \leq \hat{\mu}(X_{n+1}) + \hat{r}_\alpha\big) \\
&\geq 1 - \alpha
\end{aligned}
$$

Logo, o intervalo de predição conforme de nível α é:

$$
C^{(\alpha)}(x_{n+1}) = [\hat{\mu}(X_{n+1}) - \hat{r}_\alpha \,;\, \hat{\mu}(X_{n+1}) + \hat{r}_\alpha]
$$

Um pouco de teoria

Revisando:

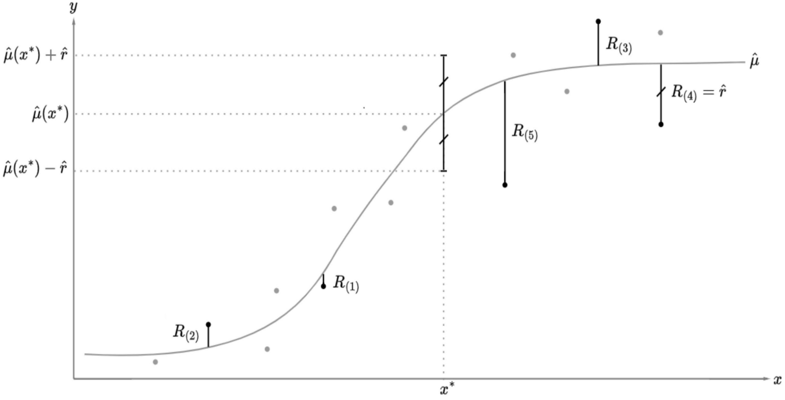
1. Obtenha $R_i = |Y_i - \hat{\mu}(X_i)|$, $i = 1, \dots, n$.
2. Ordene: $R_{(1)} \leq R_{(2)} \leq \cdots \leq R_{(n)}$.
3. Calcule $\hat{r}_\alpha = R_{(\lceil (1-\alpha)(n+1) \rceil)}$.
4. Intervalo conforme:

$$C^{(\alpha)}(x_{n+1}) = [\, \hat{\mu}(x_{n+1}) - \hat{r}_\alpha \, , \, \hat{\mu}(x_{n+1}) + \hat{r}_\alpha \,].$$

Um pouco de teoria

Supondo $n = 5$ e $\alpha = 0.4$, segue que

$$r_{0.4} = R_{\lceil 0.6 \times 6 \rceil} = R_{(4)}$$



Um pouco de teoria

- **Validade em amostra finita:** cobertura marginal $\approx 1 - \alpha$, sem teoria assintótica.
- **Agnóstico ao modelo:** vale para qualquer preditor $\hat{\mu}$, desde que os dados sejam permutáveis.
- **Papel da permutabilidade:** a única hipótese é de permutabilidade entre calibração e teste.

Um pouco de teoria

- **Eficiência dos intervalos:** modelos melhores produzem intervalos, em média, mais curtos.
- **Split conformal:** parte da amostra só para calibração (*trade-off* ajuste vs. calibração).
- **Cobertura marginal, não condicional:** garantia em média na população, não para cada x fixo.

Um pouco de teoria

- Escore de resíduo absoluto gera intervalos de comprimento fixo.
- Solução: escore normalizado (localmente ponderado):

$$R_i = \frac{|y_i - \hat{y}_i|}{\hat{\sigma}_i}.$$

- Obtenha $\hat{\sigma}_i$ ajustando um modelo para os resíduos absolutos $|y_i - \hat{y}_i|$.

Diagnóstico

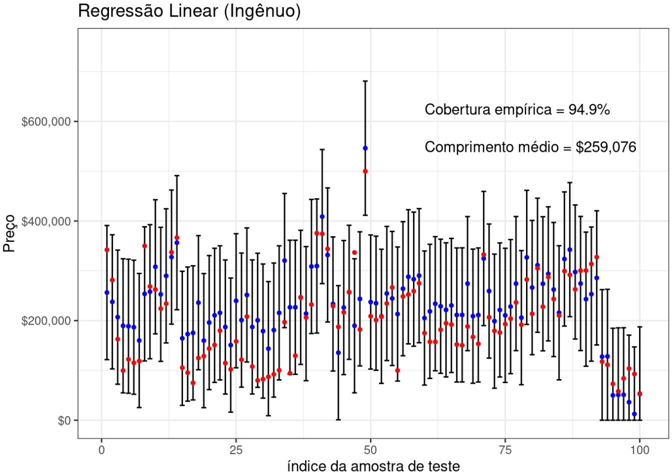
Focaremos em duas verificações:

- Cobertura empírica no conjunto de teste próxima de $1 - \alpha$.
- Comparar comprimentos dos intervalos entre modelos, buscando intervalos curtos que se adaptem à variabilidade local.

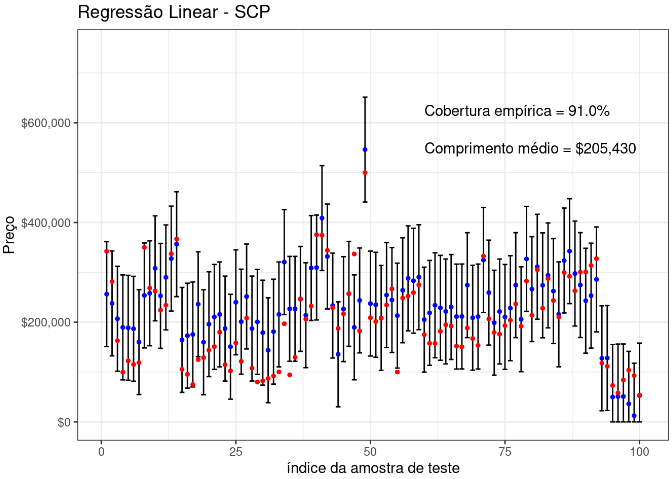
Lab em R

Hora de colocar a mão na massa :)

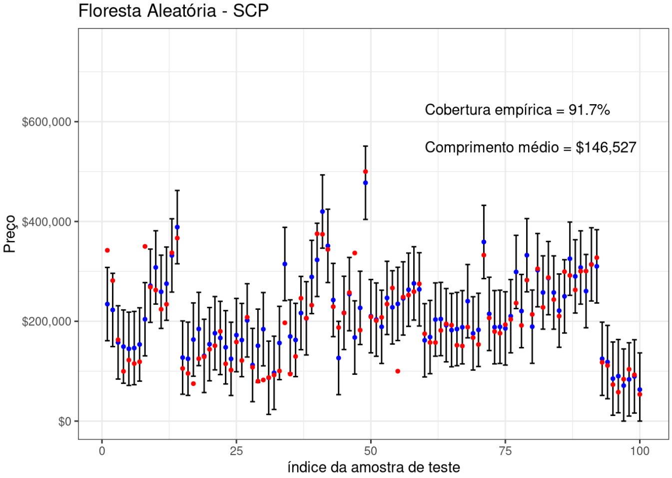
Dados California Housing



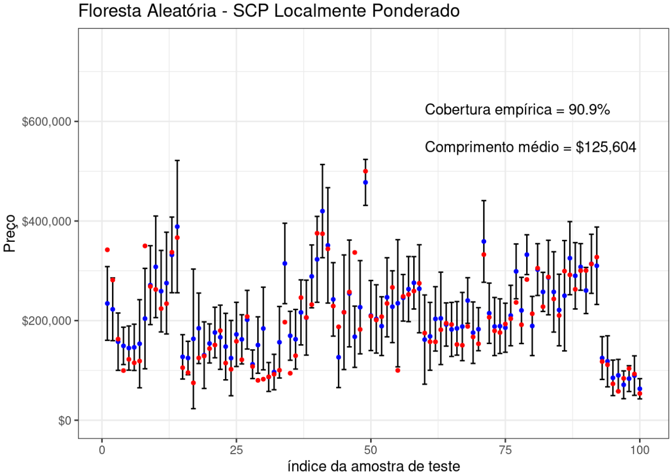
Dados California Housing



Dados California Housing



Dados California Housing



Conclusões

- Problema motivador: precificar imóveis via **intervalos plausíveis** de preço.
- Métodos tradicionais: intervalos sob suposições fortes (linearidade, normalidade, homoscedasticidade).
- Predição conforme: intervalos com **cobertura em amostra finita** $\approx 1 - \alpha$, sob permutabilidade.

Conclusões

- **Agnóstica ao modelo**: vale para qualquer $\hat{\mu}$; eficiência reflete a qualidade preditiva.
- Extensões práticas: escores normalizados e diagnóstico via cobertura empírica e largura dos intervalos.
- Predição conforme como ferramenta flexível de **quantificação de incerteza** em modelos estatísticos e de de aprendizado de máquina.

Exercício 1

1. Considere os seguintes valores para os escores de conformidade, baseados no resíduo absoluto, no conjunto de calibração ($n = 7$):

R_1	R_2	R_3	R_4	R_5	R_6	R_7
2.3	1.7	3.1	0.9	2.8	1.2	2.0

Suponha que, para um novo ponto de teste x_8, $\hat{\mu}(x_8) = 12.3$.

(a) Fixe $\alpha = 0.4$, ordene os escores $R_1, \ldots, R_7$ e determine

$$\hat{r}_\alpha = R_{(\lceil (1-\alpha)(n+1) \rceil)}.$$

(b) Encontre o intervalo de predição conforme $C^{(0.4)}(x_8)$.

(c) Fixe $\alpha = 0.3$ e encontre o intervalo de predição conforme. Comente a diferença entre este intervalo e o intervalo obtido no item (b).

Exercício 2

Use o conjunto `Boston` do pacote `MASS`, dividindo-o em 40% treino, 30% calibração e 30% teste.

1. Ajuste uma regressão linear usando a amostra de treino (`medv` é a variável resposta).
2. Na amostra de teste, para cada observação:
   (a) construa o **intervalo de predição clássico** da regressão linear com nível $1 - \alpha = 0{,}9$;
   (b) construa o **intervalo de predição conforme** com o mesmo nível, usando a amostra de calibração e o escore $R_i = |Y_i - \hat{\mu}(X_i)|$.
3. Calcule as coberturas empíricas e comprimentos médios para os dois tipos de intervalos.

Perguntas:

(a) Qual abordagem ficou mais próxima da cobertura nominal?

(b) Qual produziu intervalos mais curtos?

(c) Quais vantagens você observa na predição conforme em relação à abordagem clássica?