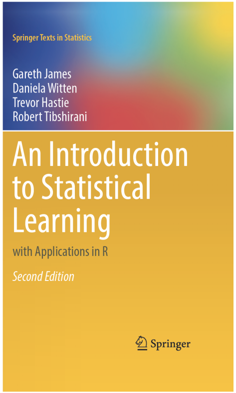
Aula 3: Regressão Linear e Revisão da Semana 1

Machine Learning

Paulo Orenstein

Verão, 2025
IMPA



Capítulo 2: Noções gerais de aprendizado

Créditos de figuras e slides: James, Witten, Hastie e Tibshirani

Setup de machine learning

- Dados de treino: $\{(x_i, y_i)\}_{i=1}^{n_{\text{tr}}}$ iid, onde $x_i \in \mathcal{X}$ e $y_i \in \mathcal{Y}$
- Função preditiva: $\hat{f}_{\text{Tr}} : \mathcal{X} \to \mathcal{Y}$
- Função-perda: $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$
  - Regressão: $\mathcal{Y} = \mathbb{R}$ e $L(y, \hat{f}(x)) = (y - \hat{f}(x))^2$, com $f(x) = \mathbb{E}[Y|X = x]$
  - Classificação: $\mathcal{Y} = \{0, 1\}$ e $L(y, \hat{f}(x)) = \mathbb{I}_{[y \neq \hat{f}(x)]}$, com $f(x) = \operatorname{argmax}_{j \in \{0,1\}} \mathbb{P}[Y = j|X = x]$
- Objetivo: minimizar risco $\hat{f}^* = \operatorname{argmin}_{\hat{f} \in \mathcal{F}} \mathbb{E}[L(Y, \hat{f}(X))]$, onde $\mathcal{F}$ precisa ser escolhido
- Na prática, usamos o risco empírico, ou seja, trocamos valor esperado por média:

$$\hat{f}^* = \operatorname*{argmin}_{\hat{f} \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} L(y_i, \hat{f}(x_i)) = \operatorname{argmin}_{\hat{f} \in \mathcal{F}} \hat{L}(\{y_i, \hat{f}(x_i)\}_{i=1}^{n})$$

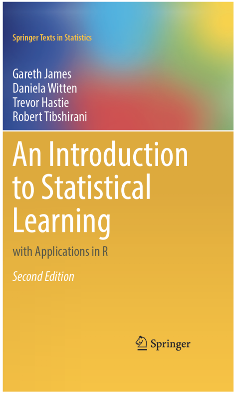
- Três componentes básicos de ML: (i) classe de modelos $\mathcal{F}$; (ii) função-perda $\hat{L}$; (iii) otimizador
- Exemplos: regressão linear e kNN

Testando o sucesso de um modelo

- No caso de regressão, queremos saber se $\mathrm{MSE} = \mathbb{E}[(Y - \hat{f}(X))^2]$ é pequeno
- Usar os dados de treino com $\widehat{\mathrm{MSE}}_{\mathrm{Tr}} = \frac{1}{n_{\mathrm{tr}}} \sum_{i \in \mathrm{Tr}} (y_i - \hat{f}(x_i))^2$, por si só, não é boa ideia
- Usamos um conjunto de teste: $\widehat{\mathrm{MSE}}_{\mathrm{Te}} = \frac{1}{n_{\mathrm{te}}} \sum_{i \in \mathrm{Te}} (y_i - \hat{f}(x_i))^2$
- Uma maneira de decompor esses valores é através do trade-off viés-variância: para novo (x_*, y_*),

$$\mathbb{E}[(y_* - \hat{f}(x_*))^2] = \mathbb{E}\left[\left(y_* - \mathbb{E}[\hat{f}(x_*)] + \mathbb{E}[\hat{f}(x_*)] - \hat{f}(x_*)\right)^2\right] = \left(\mathrm{bias}(\hat{f}(x_*))\right)^2 + \mathbb{V}[\hat{f}(x_*)] + \mathbb{V}[\varepsilon],$$

  onde $\mathrm{bias}(\hat{f}(x_*)) = \mathbb{E}[\hat{f}(x_*)] - f(x_*)$, e o erro pode ser redutível ou irredutível
- Viés: erro introduzido por aproximar f com alguma $\hat{f}$ que pode não ser capaz de capturar as nuances de f; métodos lineares estimando f não-lineares têm viés alto
- Variância: quanto nossas estimativas mudam se os dados de treino mudarem; métodos que colam nos dados de treino têm alta variância



Capítulo 3: Métodos lineares para regressão

Créditos de figuras e slides: James, Witten, Hastie e Tibshirani

Regressão linear múltipla

- Com vetores $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{X} \in \mathbb{R}^{n\times(p+1)}$, vamos estimar $\boldsymbol{\beta} \in \mathbb{R}^{p\times 1}$ via $\hat{\boldsymbol{\beta}} = \operatorname{argmin}_{\tilde{\boldsymbol{\beta}}}(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})^T(\mathbf{y} - \mathbf{X}\tilde{\boldsymbol{\beta}})$:

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$$

- Supondo os erros independentes com $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$ e $\mathbb{V}[\boldsymbol{\varepsilon}] = \sigma^2\mathbf{I}$,

$$\mathbb{E}[\hat{\boldsymbol{\beta}}] = \boldsymbol{\beta}, \qquad \mathbb{V}[\hat{\boldsymbol{\beta}}] = (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2$$

- Supondo $\boldsymbol{\varepsilon} \sim N(\mathbf{0}, \sigma^2\mathbf{I})$, vale que $\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, (\mathbf{X}^T\mathbf{X})^{-1}\sigma^2)$

- Para fazer testes de hipótese precisamos estimar σ^2; usamos $\hat{\sigma}^2 = \frac{1}{n-(p+1)}\sum_{i=1}^n (y_i - \hat{y}_i)^2$

- Testes de hipótese:
  - Testar $H_0 : \beta_j = 0$: usamos um teste t
  - Testar $H_0 : \beta_{k+1} = \beta_{k+2} = \cdots = \beta_p = 0$: usamos um teste F

- Filosofia: assumir H_0, e usar uma estatística de distribuição conhecida para ver se o valor observado é extremo demais; a probabilidade de ser tão extremo ou mais é o p-valor

Questões fundamentais

- Como saber qual subconjunto de previsores é o melhor para estimar Y?
  - Não dá pra testar todos: forward, backward ou mixed selection

- Como saber se algum previsor $X_1, \dots, X_p$ é de fato significativo para prever Y?
  - Teste t ou teste F

- Como saber se o fit do modelo é bom o suficiente?
  - $\text{RSE} = \sqrt{\frac{1}{n-(p+1)}\text{RSS}}$ ou, se possível, $\widehat{\text{MSE}}_{\text{Te}} = \frac{1}{n_{\text{te}}} \sum_{i \in \text{Te}} (y_i - \hat{f}(x_i))^2$

- Como prever o valor de Y? É possível construir intervalos de confiança?
  - $\hat{y}_{n+1} = x_{n+1}^T \hat{\boldsymbol{\beta}}$; existem intervalos para esse valor

Extensões

- Previsores qualitativos binários: incluir $X_i = \mathbb{I}_{[i \text{ na classe } 1]}$

$$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{se } i\text{-ésima instância está na classe } 1 \\ \beta_0 + \varepsilon_i, & \text{caso contrário} \end{cases}$$

- Previsores qualitativos em geral: incluir $X_{i1} = \mathbb{I}_{[i \text{ na classe } 1]}$ e $X_{i2} = \mathbb{I}_{[i \text{ na classe } 2]}$, etc

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} \varepsilon_i = \begin{cases} \beta_0 + \beta_1 + \varepsilon_i, & \text{se } i\text{-ésima pessoa é da classe } 1 \\ \beta_0 + \beta_2 + \varepsilon_i, & \text{se } i\text{-ésima pessoa é da classe } 2 \\ \beta_0 + \varepsilon_i, & \text{se } i\text{-ésima pessoa é da classe } 0 \end{cases}$$

- Interações: $y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \times X_2) + \varepsilon$
  - Com interações, um aumento em X_1 tem efeito em y que depende do valor de X_2
  - Princípio da hierarquia: ao incluir interação, incluir também efeitos principais

- Não-linearidades: $y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \cdots + \beta_p X^p + \varepsilon$ (regressão polinomial)

Capítulo 3

- Regressão linear simples: previsão e inferência
- Regressão linear múltipla: previsão e inferência
- Extensões: previsores qualitativos, interações, não-linearidades
- Problemas: correlação nos erros, heterocedasticidade, pontos de alavanca e colinearidade

Problemas

1. Correlação entre os erros
2. Variância dos erros não-constante (heterocedasticidade)
3. Outliers
4. Pontos de alavanca
5. Colinearidade

Problema 1: correlação nos erros

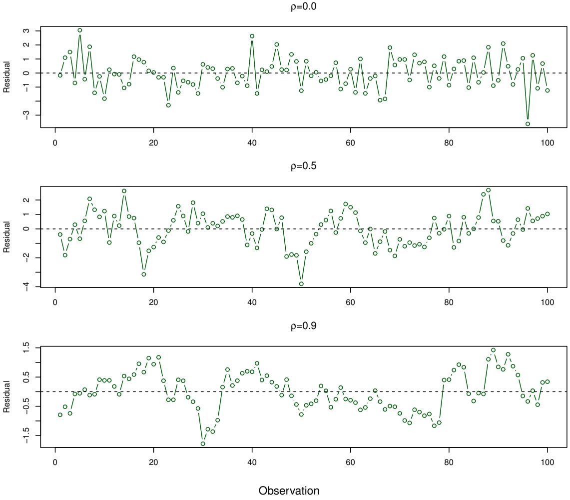
- Assumimos que $y_i = f(x_i) + \varepsilon_i$ para $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$

- Se as hipóteses de independência ou Normalidade falham, nossos procedimentos inferenciais ficam inválidos (*e.g.*, erros padrão, intervalos de confiança, testes de hipóteses)

- Exemplo: se há correlação entre os erros ε_i (ou seja, não são iid)
  - séries temporais (*e.g.*, finanças)
  - séries espaciais (*e.g.*, meteorologia)

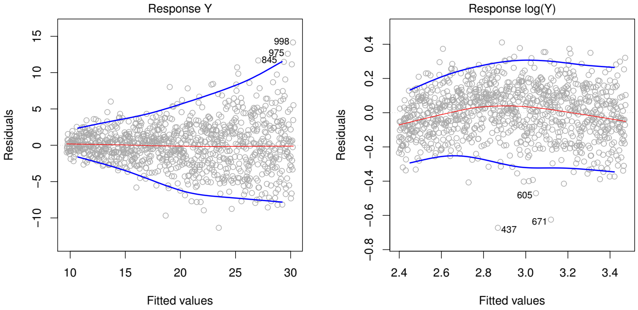
- Há métodos específicos para lidar com correlação nos erros (*e.g.*, processos gaussianos)

Problema 1: correlação nos erros



Problema 2: heterocedasticidade

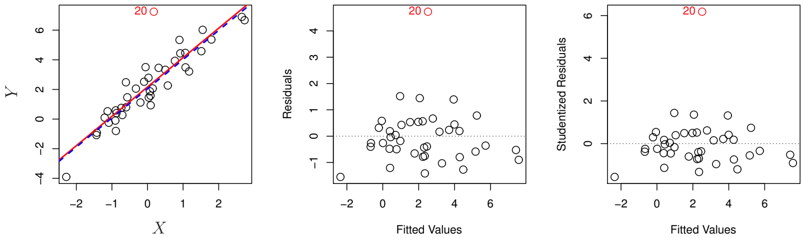
- Mesmo se os erros forem independentes e Normais, pode ser que $\mathbb{V}[\varepsilon_i] = \sigma_i^2$ (ou seja, variável)
- Um diagnóstico é visualizar resíduos vs previsões:



- Aqui, resíduos têm variância maior quanto maior a previsão. Solução: transformações
- Outra solução: atribuir pesos para as observações (regressão linear com pesos)

Problema 3: outliers

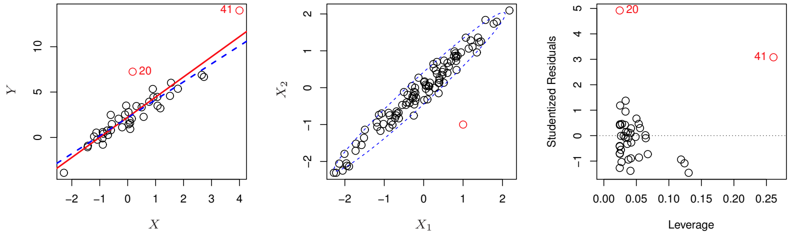
- Um outlier são pontos com erros enormes, em geral por questões estruturais



- Não costumam afetar o fit, mas afetam nossas medidas de avaliação (*e.g.*, MSE)
- Soluções:
  - Se é um erro estrutural, basta remover o ponto (cuidado!)
  - Se não, o modelo talvez precise ser mais complexo ou fazer uso de outros previsores

Problema 4: pontos de alavanca

- Se outliers são valores poucos usuais de y, pontos de alavanca são x_i pouco usuais
- Pontos de alavanca têm efeito extremo no fit



- Eles podem ser medidos pelo seu valor de influência:

$$h_{ii} = \frac{\partial \hat{y}_i}{\partial y_i} = \mathbf{H}_{ii} = (\mathbf{X}(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T)_{ii} \in [1/n, 1]$$

Problema 5: colinearidade

- Problema: há uma relação linear entre quaisquer dois previsores
- Por exemplo, se **`balance`** $\approx$ **`limit`** $+$ **`rating`**, mas **`rating`** $= 80 + 0.06 \times$ **`limit`**
- Nesse caso, $\mathbf{X}$ não tem posto cheio e $\mathbf{X}^T\mathbf{X}$ não é inversível (lembre que $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$)
- Mais concretamente, isso quer dizer que os parâmetros não são identificáveis
  - Se $X_1 = X_2$, então $(\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2)$ resulta nas mesmas previsões que $(\hat{\beta}_0, \hat{\beta}_1 + 100, \hat{\beta}_2 - 100)$. Dessa forma, a variância dos estimadores vai para infinito
  - Mesmo com casos menos extremos, a variância dos estimadores pode ficar muito alta
- Muitas vezes, colinearidade é descuido (*e.g.*, incluindo um indicador para **`brasileiro`** e outro para **`não-brasileiro`**)

Problema 5: colinearidade

- Com 2 previsores, é possível usar a correlação para diagnosticar colinearidade
- Quando há p previsores, é mais difícil: correlações dois-a-dois não revelam multicolinearidade
- Diagnóstico: fator de inflação da variância:

$$\text{VIF}(\hat{\beta}_j) = \frac{1}{1 - R^2_{*j}},$$

  onde R^2_{*j} é o valor da estatística R^2 ao regredir X_j em X_{-j}.
- Isso mede o grau de colinearidade de X_j por X_{-j}
- Matematicamente, o VIF consegue isolar a influencia das variáveis X_{-j} na variância de $\hat{\beta}_j$:

$$\widehat{\mathbb{V}}[\hat{\beta}_j] = \hat{\sigma}^2(\mathbf{X}^T\mathbf{X})^{-1}_{jj} = \hat{\sigma}^2(X_j^T X_j - \hat{\boldsymbol{\beta}}_{*j}^T(\mathbf{X}_{-j}^T\mathbf{X}_{-j})\hat{\boldsymbol{\beta}}_{*j})^{-1} = \frac{\hat{\sigma}^2}{(n-1)\widehat{\mathbb{V}}[X_j]} \cdot \frac{1}{1 - R^2_{*j}}$$

Questões computacionais

- Calcular $\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}$ é perigoso (a inversão é numericamente instável e lenta)
- Em geral, a inversão pode ser feita via decomposição QR ou SVD
- QR: se $\mathbf{X} = \mathbf{QR}$, com $\mathbf{Q}^T\mathbf{Q} = \mathbf{I}$ e $\mathbf{R}$ triangular superior, pela condição de primeira ordem,

$$
\begin{aligned}
(\mathbf{X}^T\mathbf{X})\hat{\boldsymbol{\beta}} &= \mathbf{X}^T\mathbf{y} \\
(\mathbf{R}^T\mathbf{R})\hat{\boldsymbol{\beta}} &= \mathbf{R}^T\mathbf{Q}^T\mathbf{y} \\
\mathbf{R}\hat{\boldsymbol{\beta}} &= \mathbf{Q}^T\mathbf{y}
\end{aligned}
$$

  e, como $\mathbf{R}$ é diagonal superior, é possível resolver usando substituição reversa
- SVD: se $\mathbf{X} = \mathbf{UDV}^T$, $\mathbf{X}^T\mathbf{X} = \mathbf{VD}^2\mathbf{V}^T$, e

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y} = \mathbf{VD}^{-2}\mathbf{V}^T\mathbf{VDU}^T\mathbf{y} = \mathbf{VD}^{-1}\mathbf{U}^T\mathbf{y}$$

Perguntas para revisão

- Que resultados deixam de valer se há correlação nos erros?
- Como visualizar a presença de correlação nos erros?
- O que é heterocedasticidade? Por que é um problema? Como diagnosticar? Como resolver?
- O que são outliers e pontos de alavanca? Por que são um problema? Como diagnosticar? Como resolver?
- O que é colinearidade? Por que é um problema? Como diagnosticar? Como resolver?
- Por que inverter $X^T X$ não é boa ideia para obter $\hat{\beta}$? O que fazer ao invés?