Machine Learning e políticas públicas

Prof. Rafael B. Barbosa
Universidade Federal do Ceará
Fortaleza/2020

Sumário

Metodos de shirinkage

Com o surgimento de novas bases de dados, métodos de estimação tradicionais (MQO/MV) tornam-se inadequados.
Seja o seguinte modelo linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

Então:

- Se $n \geq p \Longrightarrow$MQO é factível
- Se $p \gg n \Longrightarrow$ MQO é factível, porém não existe uma única solução para os β's

Com o surgimento de novas bases de dados, métodos de estimação tradicionais (MQO/MV) tornam-se inadequados.
Seja o seguinte modelo linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

Então:

- Se $n \geq p \Longrightarrow$MQO é factível
- Se $p \gg n \Longrightarrow$ MQO é factível, porém não existe uma única solução para os β's

Com o surgimento de novas bases de dados, métodos de estimação tradicionais (MQO/MV) tornam-se inadequados.
Seja o seguinte modelo linear:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \epsilon_i$$

Então:

- Se $n \geq p \Longrightarrow$MQO é factível
- Se $p \gg n \Longrightarrow$ MQO é factível, porém não existe uma única solução para os β's

Métodos de shirinkage utillizam o princípio dos mínimos quadrados para obter o melhor modelo dentre p variáveis.

Todavia, o princípio dos minimos quadrados é aplicado restringindo (ou regularizando) o procedimento de otimização.

Suponha o seguinte modelo linear com p regressores:
$y_i = \beta_0 + \beta_1 x_{i1} + ... + \beta_{ip} x_{ip} + \epsilon_i$

O princípio dos minimos quadrados restritos é dado por:

$$min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

sujeito a $g(\beta_j)$

Escrevendo este problema em termos do metodo de otimização de lagrange temos:

$$\mathcal{L} = \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda\, g(\beta_j)$$

A variável λ é chamada de parâmetro de afinação (*tunning parameter*) e $\lambda\, g(\beta_j)$ é chamado de penalidade de shrinkage.

A introdução da penalidade de shirinkage faz com que os regressores estimados reduzam seus valores caso não houvesse penalidade. Variáveis mais irrelevantes a previsão de y_i são mais penalizadas, recebendo valores menores.

A diferença entre os métodos de shrinkage reside na escolha da função g, isto é, na forma como se penaliza os $\beta's$.

Método de Ridge

O método de Ridge utiliza uma penalidade

$$g(\beta) = \sum_{j=1}^{p} \beta_j^2 = \beta_1^2 + ... + \beta_p^2.$$

O problema de minimização é:

$$min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 + \lambda \sum_{j=1}^{p} \beta_j^2$$

O parâmetro de afinação serve para dar mais ou menos peso a restrição

1. Se $\lambda = 0$, então a minimização é a mesma de MQO
2. Se $\lambda \to \infty$, então os parâmetros $\beta \to 0$.
3. Ou seja, para cada valor de λ teremos um diferente valor de β_{Ridge}
4. Portanto, a escolha do λ mais adequado é fundamental.

O parâmetro de afinação serve para dar mais ou menos peso a restrição

1. Se $\lambda = 0$, então a minimização é a mesma de MQO
2. Se $\lambda \to \infty$, então os parâmetros $\beta \to 0$.
3. Ou seja, para cada valor de λ teremos um diferente valor de β_{Ridge}
4. Portanto, a escolha do λ mais adequado é fundamental.

O parâmetro de afinação serve para dar mais ou menos peso a restrição

1. Se $\lambda = 0$, então a minimização é a mesma de MQO
2. Se $\lambda \to \infty$, então os parâmetros $\beta \to 0$.
3. Ou seja, para cada valor de λ teremos um diferente valor de β_{Ridge}
4. Portanto, a escolha do λ mais adequado é fundamental.

O parâmetro de afinação serve para dar mais ou menos peso a restrição

1. Se $\lambda = 0$, então a minimização é a mesma de MQO
2. Se $\lambda \to \infty$, então os parâmetros $\beta \to 0$.
3. Ou seja, para cada valor de λ teremos um diferente valor de β_{Ridge}
4. Portanto, a escolha do λ mais adequado é fundamental.

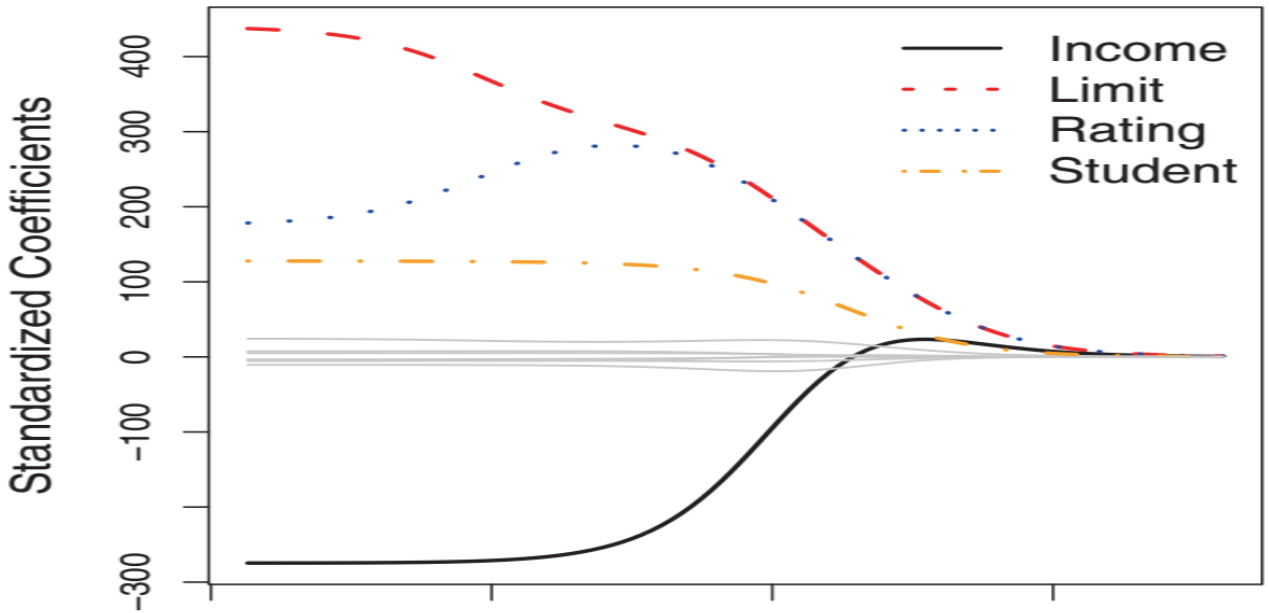
O parâmetro de afinação serve para dar mais ou menos peso a restrição

1. Se $\lambda = 0$, então a minimização é a mesma de MQO
2. Se $\lambda \to \infty$, então os parâmetros $\beta \to 0$.
3. Ou seja, para cada valor de λ teremos um diferente valor de β_{Ridge}
4. Portanto, a escolha do λ mais adequado é fundamental.

Observação: A redução dos parametros estimados foi aplicada apenas em $\beta = (\beta_1, ..., \beta_p)$ e não em β_0, pois este não mensura o impacto de variável nenhuma.

Exemplo: Usando a base de dados Credit, foi regredido um modelo linear simples para cada uma das onze variaveis contra balance, usando estimador de Ridge: $balance_i = \beta_0 + \sum_{j=1}^{11} \beta_j x_{ij} + \epsilon_i$.

Primeiramente, vejamos o que ocorre a medida que aumentamos o valor do parâmetro de afinação.

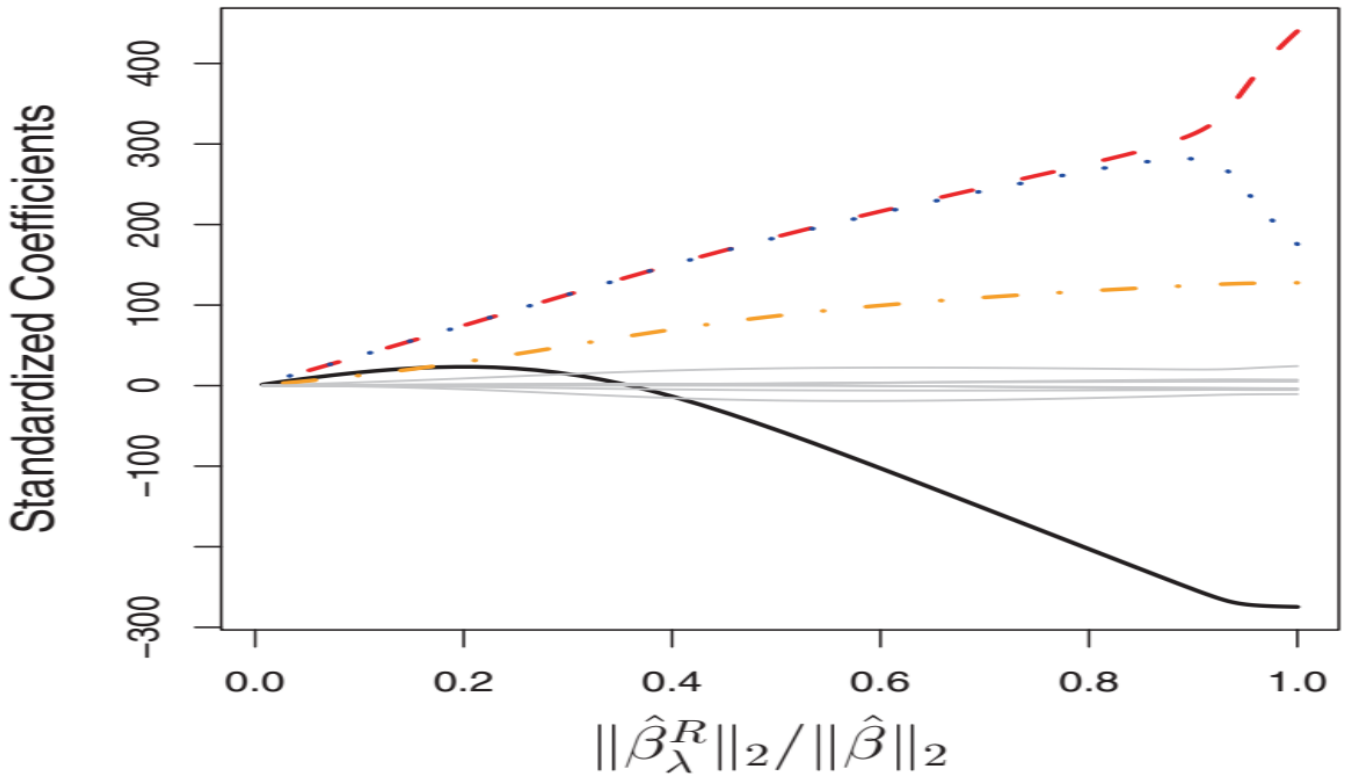


1. À medida que λ aumenta
2. Se $\lambda \to \infty$, então os parâmetros $\beta \to 0$.

1. À medida que λ aumenta
2. Se $\lambda \to \infty$, então os parâmetros $\beta \to 0$.

1. À medida que λ aumenta
2. Se $\lambda \to \infty$, então os parâmetros $\beta \to 0$.

Para comparar o que acontece quando a restrição é colocada no processo de minimização, vamos comparar o mesmo modelo com a estimativa de MQO.



Método do Lasso

Método do lasso

O método de Ridge possui a desvantagem de incluir ao final da estimação todas as p variáveis do modelo original. Isto é, o método de Ridge não realiza regularização.

Metodos do lasso

O método do lasso utiliza uma penalidade do tipo: $\| \beta \|_1 = \sum | \beta |$. Ou seja, o problema de otimização será:

$$min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2$$

Sujeito a

$$\sum_{j=1}^{p} | \beta_j | \leq s$$

Método do Lasso

Ou em termos do método de Lagrange:

$$\mathcal{L} = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 - \lambda \sum_{j=1}^{p} \mid \beta_j \mid$$

Método do lasso

O método do lasso restinge a soma dos quadrados dos resíduos (SQR) a uma restrição normada do tipo L_1.

Por sua vez, o método de Ridge aplica uma restrição normada do tipo L_2, dada por: $\| \beta \|_2 = \sum \beta_j^2$

Ao contrarío do método do Ridge, o método do Lasso realiza uma regularização das variáveis:

1. O lasso força alguns $\beta's$ estimados a receber valor zero.
2. O lasso realiza portanto: seleção de variáveis ou esparçamento ou regularização.
3. Significando que: dependendo do parâmetro de afinamento λ um número menor de variáveis restará no modelo como sendo as mais relevantes para a previsão.

Ao contrarío do método do Ridge, o método do Lasso realiza uma regularização das variáveis:

1. O lasso força alguns $\beta's$ estimados a receber valor zero.
2. O lasso realiza portanto: seleção de variáveis ou esparçamento ou regularização.
3. Significando que: dependendo do parâmetro de afinamento λ um número menor de variáveis restará no modelo como sendo as mais relevantes para a previsão.

Ao contrarío do método do Ridge, o método do Lasso realiza uma regularização das variáveis:

1. O lasso força alguns $\beta's$ estimados a receber valor zero.
2. O lasso realiza portanto: seleção de variáveis ou esparçamento ou regularização.
3. Significando que: dependendo do parâmetro de afinamento λ um número menor de variáveis restará no modelo como sendo as mais relevantes para a previsão.

Ao contrarío do método do Ridge, o método do Lasso realiza uma regularização das variáveis:

1. O lasso força alguns $\beta's$ estimados a receber valor zero.
2. O lasso realiza portanto: seleção de variáveis ou esparçamento ou regularização.
3. Significando que: dependendo do parâmetro de afinamento λ um número menor de variáveis restará no modelo como sendo as mais relevantes para a previsão.

Podemos comparar os três métodos vistos até aqui: Ridge e Lasso.
RIDGE:

$$min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ sujeito a } \sum_{j=1}^{p} \mid \beta_j^2 \mid \leq s$$

LASSO:

$$min \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \right)^2 \text{ sujeito a } \sum_{j=1}^{p} \mid \beta_j \mid \leq s$$

Observe que:

1. O parâmetro s mensura o grau que a restrição terá sobre os parâmetros
2. Se s muito grande, então a restrição é mais larga e a soma dos parâmetros pode ser maior
3. Se s muito pequeno, então a restrição é mais apertada e a soma dos parâmetros deve ser menor.
4. O valor do s possui correlação com o parâmetro de afinação.

Observe que:

1. O parâmetro s mensura o grau que a restrição terá sobre os parâmetros
2. Se s muito grande, então a restrição é mais larga e a soma dos parâmetros pode ser maior
3. Se s muito pequeno, então a restrição é mais apertada e a soma dos parâmetros deve ser menor.
4. O valor do s possui correlação com o parâmetro de afinação.

Observe que:

1. O parâmetro s mensura o grau que a restrição terá sobre os parâmetros
2. Se s muito grande, então a restrição é mais larga e a soma dos parâmetros pode ser maior
3. Se s muito pequeno, então a restrição é mais apertada e a soma dos parâmetros deve ser menor.
4. O valor do s possui correlação com o parâmetro de afinação.

Observe que:

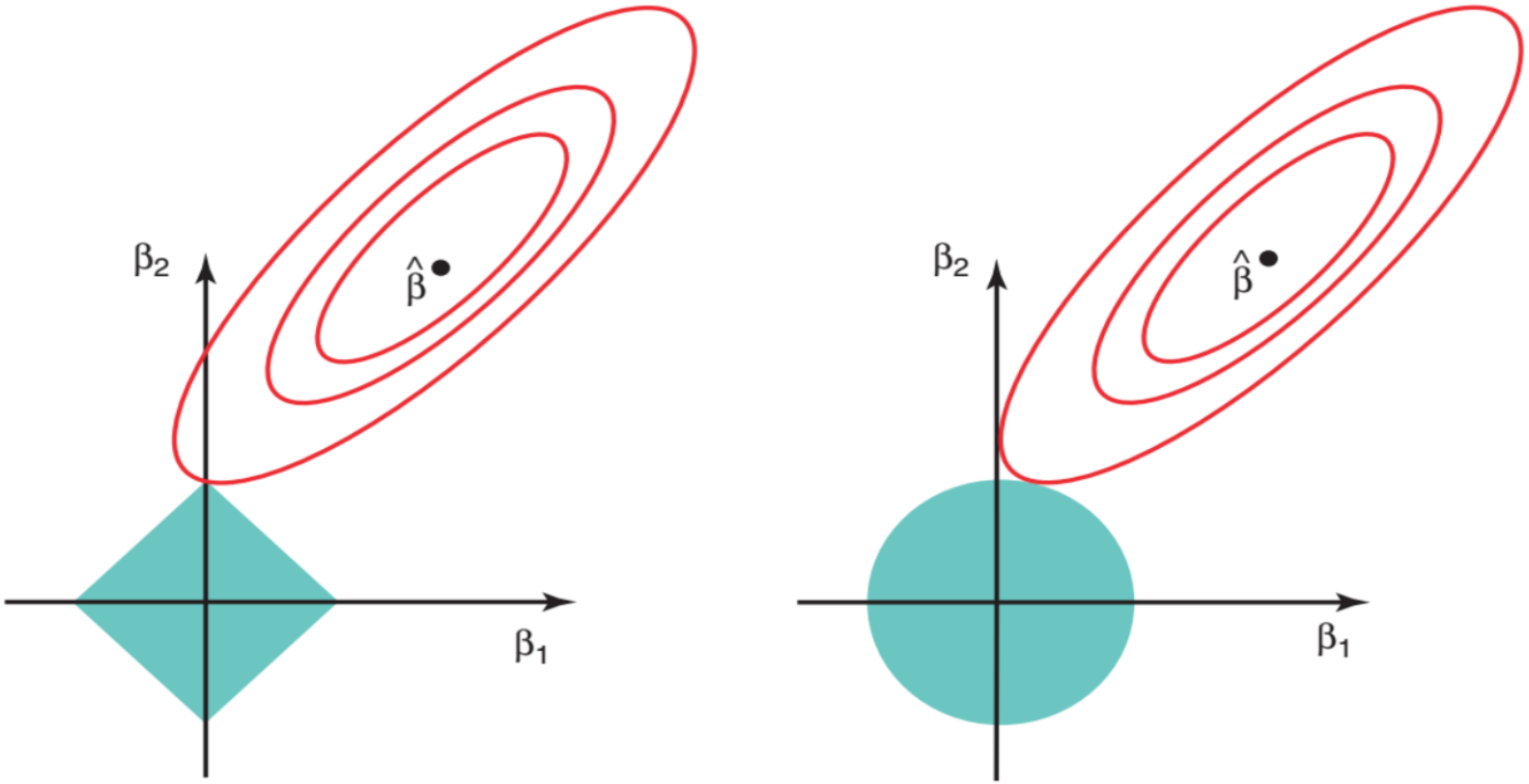
1. O parâmetro s mensura o grau que a restrição terá sobre os parâmetros
2. Se s muito grande, então a restrição é mais larga e a soma dos parâmetros pode ser maior
3. Se s muito pequeno, então a restrição é mais apertada e a soma dos parâmetros deve ser menor.
4. O valor do s possui correlação com o parâmetro de afinação.

Observe que:

1. O parâmetro s mensura o grau que a restrição terá sobre os parâmetros
2. Se s muito grande, então a restrição é mais larga e a soma dos parâmetros pode ser maior
3. Se s muito pequeno, então a restrição é mais apertada e a soma dos parâmetros deve ser menor.
4. O valor do s possui correlação com o parâmetro de afinação.

Por que o Lasso realiza seleção de variáveis e o Ridge não:

Figura: Otimização do Lasso e do Ridge



Observe que:

1. Se $s\,(\lambda = 0)$ for muito grande, então poderá englobar $\hat{\beta}$ e com isso o ótimo será neste ponto.
2. Sendo s da forma como representado na figura, note que: a restrição do lasso permite solução ótima de canto, isto é, um dos $\beta's$ será igual a zero.
3. Por sua vez, a restrição do Ridge permite que o ótimo seja obtido de forma que $\beta's$ sejam diferentes de zero (o conjunto restrição é côncavo)
4. Por fim, note que em ambos os casos, as restrições promovem a redução do valor dos parâmetros em relação ao $\hat{\beta}$.

Observe que:

1. Se $s\,(\lambda = 0)$ for muito grande, então poderá englobar $\hat{\beta}$ e com isso o ótimo será neste ponto.
2. Sendo s da forma como representado na figura, note que: a restrição do lasso permite solução ótima de canto, isto é, um dos $\beta's$ será igual a zero.
3. Por sua vez, a restrição do Ridge permite que o ótimo seja obtido de forma que $\beta's$ sejam diferentes de zero (o conjunto restrição é côncavo)
4. Por fim, note que em ambos os casos, as restrições promovem a redução do valor dos parâmetros em relação ao $\hat{\beta}$.

Observe que:

1. Se $s\,(\lambda = 0)$ for muito grande, então poderá englobar $\hat{\beta}$ e com isso o ótimo será neste ponto.
2. Sendo s da forma como representado na figura, note que: a restrição do lasso permite solução ótima de canto, isto é, um dos $\beta's$ será igual a zero.
3. Por sua vez, a restrição do Ridge permite que o ótimo seja obtido de forma que $\beta's$ sejam diferentes de zero (o conjunto restrição é côncavo)
4. Por fim, note que em ambos os casos, as restrições promovem a redução do valor dos parâmetros em relação ao $\hat{\beta}$.

Observe que:

1. Se $s\,(\lambda = 0)$ for muito grande, então poderá englobar $\hat{\beta}$ e com isso o ótimo será neste ponto.
2. Sendo s da forma como representado na figura, note que: a restrição do lasso permite solução ótima de canto, isto é, um dos $\beta's$ será igual a zero.
3. Por sua vez, a restrição do Ridge permite que o ótimo seja obtido de forma que $\beta's$ sejam diferentes de zero (o conjunto restrição é côncavo)
4. Por fim, note que em ambos os casos, as restrições promovem a redução do valor dos parâmetros em relação ao $\hat{\beta}$.

Observe que:

1. Se $s\,(\lambda = 0)$ for muito grande, então poderá englobar $\hat{\beta}$ e com isso o ótimo será neste ponto.
2. Sendo s da forma como representado na figura, note que: a restrição do lasso permite solução ótima de canto, isto é, um dos $\beta's$ será igual a zero.
3. Por sua vez, a restrição do Ridge permite que o ótimo seja obtido de forma que $\beta's$ sejam diferentes de zero (o conjunto restrição é côncavo)
4. Por fim, note que em ambos os casos, as restrições promovem a redução do valor dos parâmetros em relação ao $\hat{\beta}$.

Aplicação do Lasso na avaliação de políticas públicas

Veremos como aplicar os métodos de regularização para melhorar a identificação do efeito causal em problemas com dados transversais ou em painel.

1. Regularização para melhorar seleção sobre observáveis
2. Regularização aplicada a variáveis instrumentais
3. Inferência Causal e Regularização
4. Qual método de *shrinkage* aplicar? Exemplo de quando Ridge é mais apropriado que o LASSO.

Veremos como aplicar os métodos de regularização para melhorar a identificação do efeito causal em problemas com dados transversais ou em painel.

1. Regularização para melhorar seleção sobre observáveis
2. Regularização aplicada a variáveis instrumentais
3. Inferência Causal e Regularização
4. Qual método de *shrinkage* aplicar? Exemplo de quando Ridge é mais apropriado que o LASSO.

Veremos como aplicar os métodos de regularização para melhorar a identificação do efeito causal em problemas com dados transversais ou em painel.

1. Regularização para melhorar seleção sobre observáveis
2. Regularização aplicada a variáveis instrumentais
3. Inferência Causal e Regularização
4. Qual método de *shrinkage* aplicar? Exemplo de quando Ridge é mais apropriado que o LASSO.

Veremos como aplicar os métodos de regularização para melhorar a identificação do efeito causal em problemas com dados transversais ou em painel.

1. Regularização para melhorar seleção sobre observáveis
2. Regularização aplicada a variáveis instrumentais
3. Inferência Causal e Regularização
4. Qual método de *shrinkage* aplicar? Exemplo de quando Ridge é mais apropriado que o LASSO.

Veremos como aplicar os métodos de regularização para melhorar a identificação do efeito causal em problemas com dados transversais ou em painel.

1. Regularização para melhorar seleção sobre observáveis
2. Regularização aplicada a variáveis instrumentais
3. Inferência Causal e Regularização
4. Qual método de *shrinkage* aplicar? Exemplo de quando Ridge é mais apropriado que o LASSO.

Considere o seguinte problema estrutural:

$$y_i = d_i\alpha_0 + g(w_i) + e_i \tag{1}$$

Em que:
y_i é uma variável de resultado associada a algum tratamento (binário) d_i.
$g(w_i)$ é um conjunto de variáveis de controle, que possuem forma funcional g
e_i é o erro, tal que: $E[e_i|\, d_i,\, w_i] = 0$

Para que $\hat{\alpha}_0$ represente o efeito causal de d_i sobre y_i é preciso garantir que: y_i é indenpendente de d_i dado os controles $g(w_i)$, isto é: $y_i \perp d_i | g(w_i)$.

O conjunto de variáveis $g(w_i)$ refere-se as variáveis de controle observadas que garantem que a independência entre d_i e y_i seja válida.

Exemplo:
Problema de analisar o efeito do treinamento de professores quando estes são submetidos a uma seleção observada.

A idependência somente é garantida quando é incluido como controle os critérios de seleção.
Em outras aplicações, a função $g(w_i)$ pode ser desconhecida e a sua não inclusão pode não garantir a hipótese de independência condicionada.

Assuma que $g(w_i)$ possua alta-dimensão e seja aproximadamente linear. Isto é:

$$g(w_i) = \sum_{j=1}^{p} \beta_j x_{i,j} + r_{p,i}$$

Em que:
$x_{i,j} = P(w_i)$ é um conjunto de variáveis de controle fruto de transformações de P de w_i.
$r_{p,i}$ é o erro de aproximação de $g(w_i)$ por $\sum_{j=1}^{p} \beta_j x_{i,j}$.
Por fim, $p > n$, o que caracteriza a dimensão elevada.

ABORDAGEM TRADICIONAL:

A abordagem tradicional em economia para a escolha de x_i consiste em selecionar as variáveis de acordo com a experiência e o conhecimento prévio do problema.

Ex. Estimação da equação Minceriana.

$$log\ sal_i = \beta_0 + \beta educ_i + g(w_i) + e_i$$

Quais variáveis incluir em $g(w_i)$?

- Experiência e experiência ao quadrado
- Gênero
- Região onde vive
- Raça
- etc

Quais variáveis incluir em $g(w_i)$?

- Experiência e experiência ao quadrado
- Gênero
- Região onde vive
- Raça
- etc

Quais variáveis incluir em $g(w_i)$?

- Experiência e experiência ao quadrado
- Gênero
- Região onde vive
- Raça
- etc

Quais variáveis incluir em $g(w_i)$?

- Experiência e experiência ao quadrado
- Gênero
- Região onde vive
- Raça
- etc

Quais variáveis incluir em $g(w_i)$?

- Experiência e experiência ao quadrado
- Gênero
- Região onde vive
- Raça
- etc

Quais variáveis incluir em $g(w_i)$?

- Experiência e experiência ao quadrado
- Gênero
- Região onde vive
- Raça
- etc

Digamos que tenha sido escolhidas $k < p$ variáveis:

$$log\ sal_i = \beta_0 + \beta educ_i + \sum_{j=1}^{k} \beta_j x_{i,j} + r_{k,i} + e_i$$

Chame: $u_i = r_{k,i} + e_i$

É possível que $E[educ_i,\ u_i] = E[educ_i,\ r_{k,j}] \neq 0$ e a estimação de β não seja causal.

Ou seja, mesmo quando a variável de tratamento é exógena, a inclusão não adequada de variáveis de controle pode gerar estimativas viesas.

Esse é um caso de VOV causado pelo fato de que a seleção das variáveis observáveis é inadequada.

Belloni, Chernozhukov e Hansen (2014) tratam deste problema e propõem o método do **duplo pós-lasso** para contorná-lo.

Algoritmo Duplo Pós-Lasso

1. Aplicar o procedimento de LASSO para selecionar variáveis que prevejam o comportamento de y_i, digamos S_1
2. Aplicar o procedimento do LASSO para selecionar variáveis que prevejam o comportamento de d_i, digamos S_2
3. Estimar por MQO o modelo em (1) incluindo como variáveis de controle no lugar de $g(w_i)$ as variáveis selecionadas S_1 e S_2.

Belloni, Chernozhukov e Hansen (2014) tratam deste problema e propõem o método do **duplo pós-lasso** para contorná-lo.

Algoritmo Duplo Pós-Lasso

1. Aplicar o procedimento de LASSO para selecionar variáveis que prevejam o comportamento de y_i, digamos S_1
2. Aplicar o procedimento do LASSO para selecionar variáveis que prevejam o comportamento de d_i, digamos S_2
3. Estimar por MQO o modelo em (1) incluindo como variáveis de controle no lugar de $g(w_i)$ as variáveis selecionadas S_1 e S_2.

Belloni, Chernozhukov e Hansen (2014) tratam deste problema e propõem o método do **duplo pós-lasso** para contorná-lo.

Algoritmo Duplo Pós-Lasso

1. Aplicar o procedimento de LASSO para selecionar variáveis que prevejam o comportamento de y_i, digamos S_1
2. Aplicar o procedimento do LASSO para selecionar variáveis que prevejam o comportamento de d_i, digamos S_2
3. Estimar por MQO o modelo em (1) incluindo como variáveis de controle no lugar de $g(w_i)$ as variáveis selecionadas S_1 e S_2.

Belloni, Chernozhukov e Hansen (2014) tratam deste problema e propõem o método do **duplo pós-lasso** para contorná-lo.

Algoritmo Duplo Pós-Lasso

1. Aplicar o procedimento de LASSO para selecionar variáveis que prevejam o comportamento de y_i, digamos S_1
2. Aplicar o procedimento do LASSO para selecionar variáveis que prevejam o comportamento de d_i, digamos S_2
3. Estimar por MQO o modelo em (1) incluindo como variáveis de controle no lugar de $g(w_i)$ as variáveis selecionadas S_1 e S_2.

A hipótese da validade do algoritmo do duplo pós-lasso $g(w_i)$ decorre da possibilidade de aplicar procedimentos de regularização (LASSO) de forma a selecionar p variáveis de controle de forma que o erro de aproximação seja não correlacionado com a variável de tratamento.

Sendo $\hat{g}(w_i)$ o conjunto de variáveis selecionado e $\hat{r}_{p,i}$ o erro de aproximação estimado, então:

$$E[d_i, \hat{r}_{p,i}] = 0$$

Assim, é dito que $g(w_i)$ é aproximadamente esparço.

Características do método do duplo pós-lasso

1. É um método que dirigido por dados (*data-driven*)
2. Compreende as estimações paramétricas e não-paramétricas
3. Simples de estimar

Características do método do duplo pós-lasso

1. É um método que dirigido por dados (*data-driven*)
2. Compreende as estimações paramétricas e não-paramétricas
3. Simples de estimar

Características do método do duplo pós-lasso

1. É um método que dirigido por dados (*data-driven*)
2. Compreende as estimações paramétricas e não-paramétricas
3. Simples de estimar

Características do método do duplo pós-lasso

1. É um método que dirigido por dados (*data-driven*)
2. Compreende as estimações paramétricas e não-paramétricas
3. Simples de estimar

Mas por que DUPLO pós-Lasso?

Suponha a equação:

$$y_i = \alpha_0 d_i + \sum_{j=1}^{p} \beta_j x_{i,j} + r_{k,i} + e_i$$

O procedimento do Lasso poderia ser aplicado apenas nesta equação, selecionando de $\sum_{j=1}^{p} \beta_j x_{i,j}$ as variáveis mais relevantes para prever y_i.

Por que isto não é suficiente para que: $E[d_i, \hat{r}_{p,i}] = 0$?

Dito de outra forma, por que precisamos aplicar o procedimento do Lasso sobre:

$$d_i = \sum_{j=1}^{p} \beta_j x_{i,j} + r_{k,i} + u_i$$

Exemplo de aplicação em Econ-Tech

Altualmente, muitas empresas de tecnologia estão contratando economistas para analisar o efeito causal de seus produtos.
Ex:
Suponha que uma empresa de e-commerce deseja saber qual o efeito no número de vendas de uma promoção especial.
O problema da empresa é:

$$y_i = \alpha_0 d_i + \sum_{j=1}^{p} \beta_j x_{i,j} + e_i$$

Em que:
y_i número de vendas dos produtos durante a campanha. Apenas alguns produtos receberam desconto.
d_i indivíduos que receberam o cupom de desconto;

Exemplo de aplicação em Econ-Tech

Mesmo assumindo que d_i seja distribuído aleatoriamente, por que o simples cálculo da equação anterior pode ser viesado?

Exemplo de aplicação em Econ-Tech

Por que a compra do produto pode estar relacionada a quem recebeu o desconto e ao tipo de produto com desconto.

Neste caso, existem inúmeras características dos produtos que podem ser utilizadas como controle.
Por exemplo:
- Número de pesquisas por produtos
- Número de clicks secundários
- Clicks que leram especificações
- Tempo de leitura da pagina
- Características de quem clicou.

Exemplo de aplicação em Econ-Tech

Neste caso, utilizar o método do duplo pós-lasso é uma alternativa interessante.

- O número de variáveis de controle pode ser muito maior do que o número de indivíduos
- Não há uma clara intuição de quais são os melhores controles
- Possívelmente tais controles são não lineares: interseccionados com as variáveis características dos clientes.

Legalização do aborto afeta a criminalidade futura?

- Donohue e Levitt (2001) afirmaram que a legalização do aborto na década de 1980 em alguns estados dos EUA reduziu a criminalidade em 1990.
- Testaram o seguinte modelo usando dif-dif:
- $y_{cit} = \alpha_c a_{cit} + w_{it}^{'}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit}$
- Em que: y_{cit} indice de criminalidade $c = (violento, propriedade, homicidio)$ no estado i e no ano t, a_{cit} uma medida da taxa de abordo relevante para o tipo de crime c (baseada na idade)
- w_{it} conjunto de variáveis de controle (variantes no tempo) ao nível estadual; δ_{cit}, γ_{ct} são efeitos transversais e temporais específicos do estado, respectivamente.

- Donohue e Levitt (2001) afirmaram que a legalização do aborto na década de 1980 em alguns estados dos EUA reduziu a criminalidade em 1990.
- Testaram o seguinte modelo usando dif-dif:
- $y_{cit} = \alpha_c a_{cit} + w_{it}^{'}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit}$
- Em que: y_{cit} indice de criminalidade $c = (violento, propriedade, homicidio)$ no estado i e no ano t, a_{cit} uma medida da taxa de abordo relevante para o tipo de crime c (baseada na idade)
- w_{it} conjunto de variáveis de controle (variantes no tempo) ao nível estadual; δ_{cit}, γ_{ct} são efeitos transversais e temporais específicos do estado, respectivamente.

- Donohue e Levitt (2001) afirmaram que a legalização do aborto na década de 1980 em alguns estados dos EUA reduziu a criminalidade em 1990.
- Testaram o seguinte modelo usando dif-dif:
- $y_{cit} = \alpha_c a_{cit} + w_{it}^{'}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit}$
- Em que: y_{cit} indice de criminalidade $c = (violento, propriedade, homicidio)$ no estado i e no ano t, a_{cit} uma medida da taxa de abordo relevante para o tipo de crime c (baseada na idade)
- w_{it} conjunto de variáveis de controle (variantes no tempo) ao nível estadual; δ_{cit}, γ_{ct} são efeitos transversais e temporais específicos do estado, respectivamente.

- Donohue e Levitt (2001) afirmaram que a legalização do aborto na década de 1980 em alguns estados dos EUA reduziu a criminalidade em 1990.
- Testaram o seguinte modelo usando dif-dif:
- $y_{cit} = \alpha_c a_{cit} + w_{it}^{'}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit}$
- Em que: y_{cit} indice de criminalidade $c = (violento, propriedade, homicidio)$ no estado i e no ano t, a_{cit} uma medida da taxa de abordo relevante para o tipo de crime c (baseada na idade)
- w_{it} conjunto de variáveis de controle (variantes no tempo) ao nível estadual; δ_{cit}, γ_{ct} são efeitos transversais e temporais específicos do estado, respectivamente.

- Donohue e Levitt (2001) afirmaram que a legalização do aborto na década de 1980 em alguns estados dos EUA reduziu a criminalidade em 1990.
- Testaram o seguinte modelo usando dif-dif:
- $y_{cit} = \alpha_c a_{cit} + w_{it}^{'}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit}$
- Em que: y_{cit} indice de criminalidade $c = (violento, propriedade, homicidio)$ no estado i e no ano t, a_{cit} uma medida da taxa de abordo relevante para o tipo de crime c (baseada na idade)
- w_{it} conjunto de variáveis de controle (variantes no tempo) ao nível estadual; δ_{cit}, γ_{ct} são efeitos transversais e temporais específicos do estado, respectivamente.

- Donohue e Levitt (2001) afirmaram que a legalização do aborto na década de 1980 em alguns estados dos EUA reduziu a criminalidade em 1990.
- Testaram o seguinte modelo usando dif-dif:
- $y_{cit} = \alpha_c a_{cit} + w_{it}^{'}\beta_c + \delta_{ci} + \gamma_{ct} + \epsilon_{cit}$
- Em que: y_{cit} indice de criminalidade $c = (violento, propriedade, homicidio)$ no estado i e no ano t, a_{cit} uma medida da taxa de abordo relevante para o tipo de crime c (baseada na idade)
- w_{it} conjunto de variáveis de controle (variantes no tempo) ao nível estadual; δ_{cit}, γ_{ct} são efeitos transversais e temporais específicos do estado, respectivamente.

- w_{it} inclui as seguintes variáveis: log do número de prisioneiros defasados, log do número de policiais defasados, taxa de desemprego, renda per capita, taxa de pobreza, Programa sobre jovens (0 a 15 anos) chamado AFDC; dummy para o estado com lei de desarmamento, consumo de cerveja per capita.
- A abordagem de Donohue e Levitt (2001) funcionará apenas se as variáveis que capturam os componentes específicos dos estados e do tempo sejam representadas por poucas variáveis.
- Uma alternativa é a inclusão de um conjunto de variáveis específicas para o estados mas com tendência linear ou não linear.

- w_{it} inclui as seguintes variáveis: log do número de prisioneiros defasados, log do número de policiais defasados, taxa de desemprego, renda per capita, taxa de pobreza, Programa sobre jovens (0 a 15 anos) chamado AFDC; dummy para o estado com lei de desarmamento, consumo de cerveja per capita.
- A abordagem de Donohue e Levitt (2001) funcionará apenas se as variáveis que capturam os componentes específicos dos estados e do tempo sejam representadas por poucas variáveis.
- Uma alternativa é a inclusão de um conjunto de variáveis específicas para o estados mas com tendência linear ou não linear.

- w_{it} inclui as seguintes variáveis: log do número de prisioneiros defasados, log do número de policiais defasados, taxa de desemprego, renda per capita, taxa de pobreza, Programa sobre jovens (0 a 15 anos) chamado AFDC; dummy para o estado com lei de desarmamento, consumo de cerveja per capita.
- A abordagem de Donohue e Levitt (2001) funcionará apenas se as variáveis que capturam os componentes específicos dos estados e do tempo sejam representadas por poucas variáveis.
- Uma alternativa é a inclusão de um conjunto de variáveis específicas para o estados mas com tendência linear ou não linear.

- w_{it} inclui as seguintes variáveis: log do número de prisioneiros defasados, log do número de policiais defasados, taxa de desemprego, renda per capita, taxa de pobreza, Programa sobre jovens (0 a 15 anos) chamado AFDC; dummy para o estado com lei de desarmamento, consumo de cerveja per capita.
- A abordagem de Donohue e Levitt (2001) funcionará apenas se as variáveis que capturam os componentes específicos dos estados e do tempo sejam representadas por poucas variáveis.
- Uma alternativa é a inclusão de um conjunto de variáveis específicas para o estados mas com tendência linear ou não linear.

- CBH então propuseram incluir um conjunto de variáveis não lineares e não linerares interados com características específicas do estado.
- Foram incluidas 284 novas variáveis: variáveis em nível, primeira diferença, nível inicial, diferença inicial, média estadual das varíaveis time-state, quadrados das varíaveis, interações de todas as variáveis citadas e tendências cúbicas.
- $$\begin{cases} \Delta y_{cit} = \alpha_c \Delta a_{cit} + z'_{cit}\beta_c + \bar{\gamma_{ct}} + \Delta\epsilon_{cit} \\ \Delta a_{cit} = z'_{cit}\Pi_c + \bar{k}_{ct} + \Delta v_{vit} \end{cases}$$
- Em que: $\Delta y_{cit} = y_{cit} - y_{cit-1}$ e $\bar{\gamma_{ct}}$, $\bar{k}_{ct}$ são efeitos temporais.

- CBH então propuseram incluir um conjunto de variáveis não lineares e não linerares interados com características específicas do estado.
- Foram incluidas 284 novas variáveis: variáveis em nível, primeira diferença, nível inicial, diferença inicial, média estadual das varíaveis time-state, quadrados das varíaveis, interações de todas as variáveis citadas e tendências cúbicas.
- $$\begin{cases} \Delta y_{cit} = \alpha_c \Delta a_{cit} + z'_{cit}\beta_c + \bar{\gamma_{ct}} + \Delta\epsilon_{cit} \\ \Delta a_{cit} = z'_{cit}\Pi_c + \bar{k}_{ct} + \Delta v_{vit} \end{cases}$$
- Em que: $\Delta y_{cit} = y_{cit} - y_{cit-1}$ e $\bar{\gamma_{ct}}$, $\bar{k}_{ct}$ são efeitos temporais.

- CBH então propuseram incluir um conjunto de variáveis não lineares e não linerares interados com características específicas do estado.
- Foram incluidas 284 novas variáveis: variáveis em nível, primeira diferença, nível inicial, diferença inicial, média estadual das varíaveis time-state, quadrados das varíaveis, interações de todas as variáveis citadas e tendências cúbicas.
- $$\begin{cases} \Delta y_{cit} = \alpha_c \Delta a_{cit} + z'_{cit}\beta_c + \bar{\gamma_{ct}} + \Delta\epsilon_{cit} \\ \Delta a_{cit} = z'_{cit}\Pi_c + \bar{k}_{ct} + \Delta v_{vit} \end{cases}$$
- Em que: $\Delta y_{cit} = y_{cit} - y_{cit-1}$ e $\bar{\gamma_{ct}}$, $\bar{k}_{ct}$ são efeitos temporais.

- CBH então propuseram incluir um conjunto de variáveis não lineares e não linerares interados com características específicas do estado.
- Foram incluidas 284 novas variáveis: variáveis em nível, primeira diferença, nível inicial, diferença inicial, média estadual das varíaveis time-state, quadrados das varíaveis, interações de todas as variáveis citadas e tendências cúbicas.
- $$\begin{cases} \Delta y_{cit} = \alpha_c \Delta a_{cit} + z'_{cit}\beta_c + \bar{\gamma_{ct}} + \Delta\epsilon_{cit} \\ \Delta a_{cit} = z'_{cit}\Pi_c + \bar{k}_{ct} + \Delta v_{vit} \end{cases}$$
- Em que: $\Delta y_{cit} = y_{cit} - y_{cit-1}$ e $\bar{\gamma_{ct}}$, $\bar{k}_{ct}$ são efeitos temporais.

- CBH então propuseram incluir um conjunto de variáveis não lineares e não linerares interados com características específicas do estado.
- Foram incluidas 284 novas variáveis: variáveis em nível, primeira diferença, nível inicial, diferença inicial, média estadual das varíaveis time-state, quadrados das varíaveis, interações de todas as variáveis citadas e tendências cúbicas.
- $$\begin{cases} \Delta y_{cit} = \alpha_c \Delta a_{cit} + z_{cit}^{'} \beta_c + \bar{\gamma_{ct}} + \Delta \epsilon_{cit} \\ \Delta a_{cit} = z_{cit}^{'} \Pi_c + \bar{k}_{ct} + \Delta v_{vit} \end{cases}$$
- Em que: $\Delta y_{cit} = y_{cit} - y_{cit-1}$ e $\bar{\gamma_{ct}}$, $\bar{k}_{ct}$ são efeitos temporais.

Effect of Abortion on Crime

	Type of crime					
	Violent		*Property*		*Murder*	
Estimator	*Effect*	*Std. error*	*Effect*	*Std. error*	*Effect*	*Std. error*
First-difference	−.157	.034	−.106	.021	−.218	.068
All controls	.071	.284	−.161	.106	−1.327	.932
Double selection	−.171	.117	−.061	.057	−.189	.177

- O subconjunto selecionado inclui tendência não linear relacionada aos estados.
- Logo, tais tendências não lineares são bons previsores para o tratamento.
- Resultado: Qnd controles temporais não lineares são incluídos, abortos não são significativos para explicar certos tipos de crime.

- O subconjunto selecionado inclui tendência não linear relacionada aos estados.
- Logo, tais tendências não lineares são bons previsores para o tratamento.
- Resultado: Qnd controles temporais não lineares são incluídos, abortos não são significativos para explicar certos tipos de crime.

- O subconjunto selecionado inclui tendência não linear relacionada aos estados.
- Logo, tais tendências não lineares são bons previsores para o tratamento.
- Resultado: Qnd controles temporais não lineares são incluídos, abortos não são significativos para explicar certos tipos de crime.

- O subconjunto selecionado inclui tendência não linear relacionada aos estados.
- Logo, tais tendências não lineares são bons previsores para o tratamento.
- Resultado: Qnd controles temporais não lineares são incluídos, abortos não são significativos para explicar certos tipos de crime.

Método de Árvores

Métodos de árvores

- Métodos de árvores realizam uma divisão do espaço previsor, seja por segmentação ou por estratificação, em um número de regiões que são utilizadas para fazer a previsão da variável alvo.
- Usa-se a média ou a moda na amostra de treino para acessar a informação de cada região.
- Nesta aula analisaremos três métodos principais:árvores de regressão (e classificação) e suas extensões: bagging e floresta aleatória
- Métodos de bagging podem ser aplicados em outras técnicas de ML e fazem parte dos métodos de agregação de previsões (*ensemble models*)

Métodos de árvores

- Métodos de árvores realizam uma divisão do espaço previsor, seja por segmentação ou por estratificação, em um número de regiões que são utilizadas para fazer a previsão da variável alvo.
- Usa-se a média ou a moda na amostra de treino para acessar a informação de cada região.
- Nesta aula analisaremos três métodos principais:árvores de regressão (e classificação) e suas extensões: bagging e floresta aleatória
- Métodos de bagging podem ser aplicados em outras técnicas de ML e fazem parte dos métodos de agregação de previsões (*ensemble models*)

Métodos de árvores

- Métodos de árvores realizam uma divisão do espaço previsor, seja por segmentação ou por estratificação, em um número de regiões que são utilizadas para fazer a previsão da variável alvo.
- Usa-se a média ou a moda na amostra de treino para acessar a informação de cada região.
- Nesta aula analisaremos três métodos principais:árvores de regressão (e classificação) e suas extensões: bagging e floresta aleatória
- Métodos de bagging podem ser aplicados em outras técnicas de ML e fazem parte dos métodos de agregação de previsões (*ensemble models*)

Métodos de árvores

- Métodos de árvores realizam uma divisão do espaço previsor, seja por segmentação ou por estratificação, em um número de regiões que são utilizadas para fazer a previsão da variável alvo.
- Usa-se a média ou a moda na amostra de treino para acessar a informação de cada região.
- Nesta aula analisaremos três métodos principais:árvores de regressão (e classificação) e suas extensões: bagging e floresta aleatória
- Métodos de bagging podem ser aplicados em outras técnicas de ML e fazem parte dos métodos de agregação de previsões (*ensemble models*)

Métodos de árvores

- Métodos de árvores realizam uma divisão do espaço previsor, seja por segmentação ou por estratificação, em um número de regiões que são utilizadas para fazer a previsão da variável alvo.
- Usa-se a média ou a moda na amostra de treino para acessar a informação de cada região.
- Nesta aula analisaremos três métodos principais:árvores de regressão (e classificação) e suas extensões: bagging e floresta aleatória
- Métodos de bagging podem ser aplicados em outras técnicas de ML e fazem parte dos métodos de agregação de previsões (*ensemble models*)

INTRODUÇÃO A ÁRVORES DE REGRESSÃO

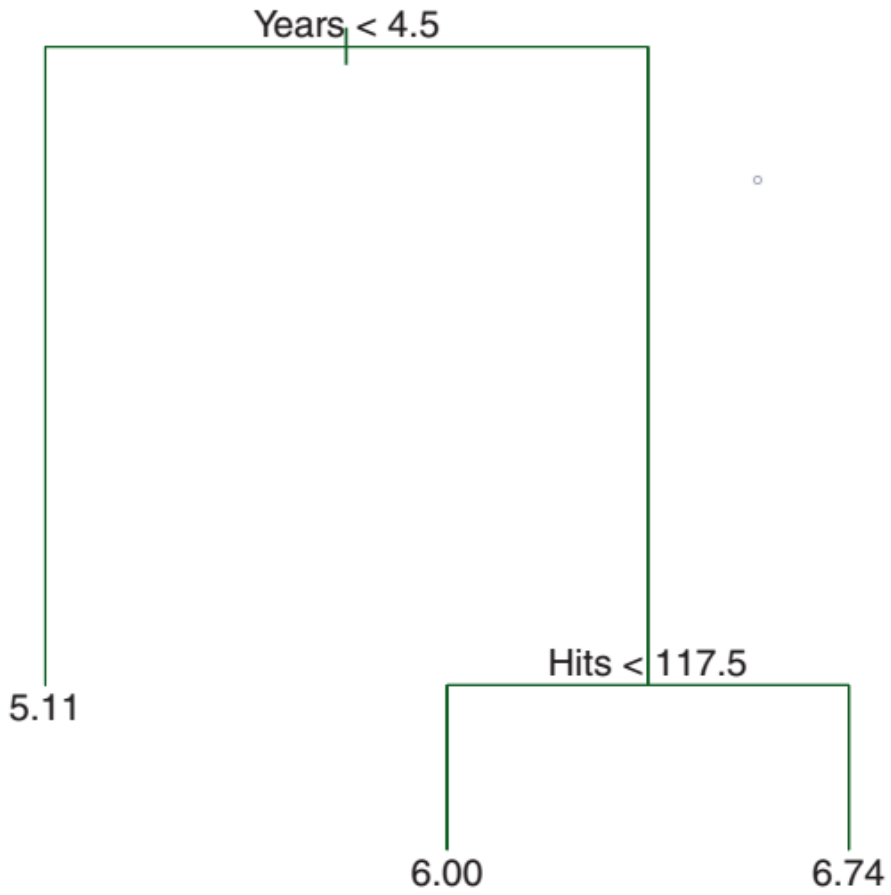
Introdução a árvore de regressão

Métodos de árvores de decisão podem ser aplicados a regressão ou a classificação.
Vamos iniciar com métodos de regressão.

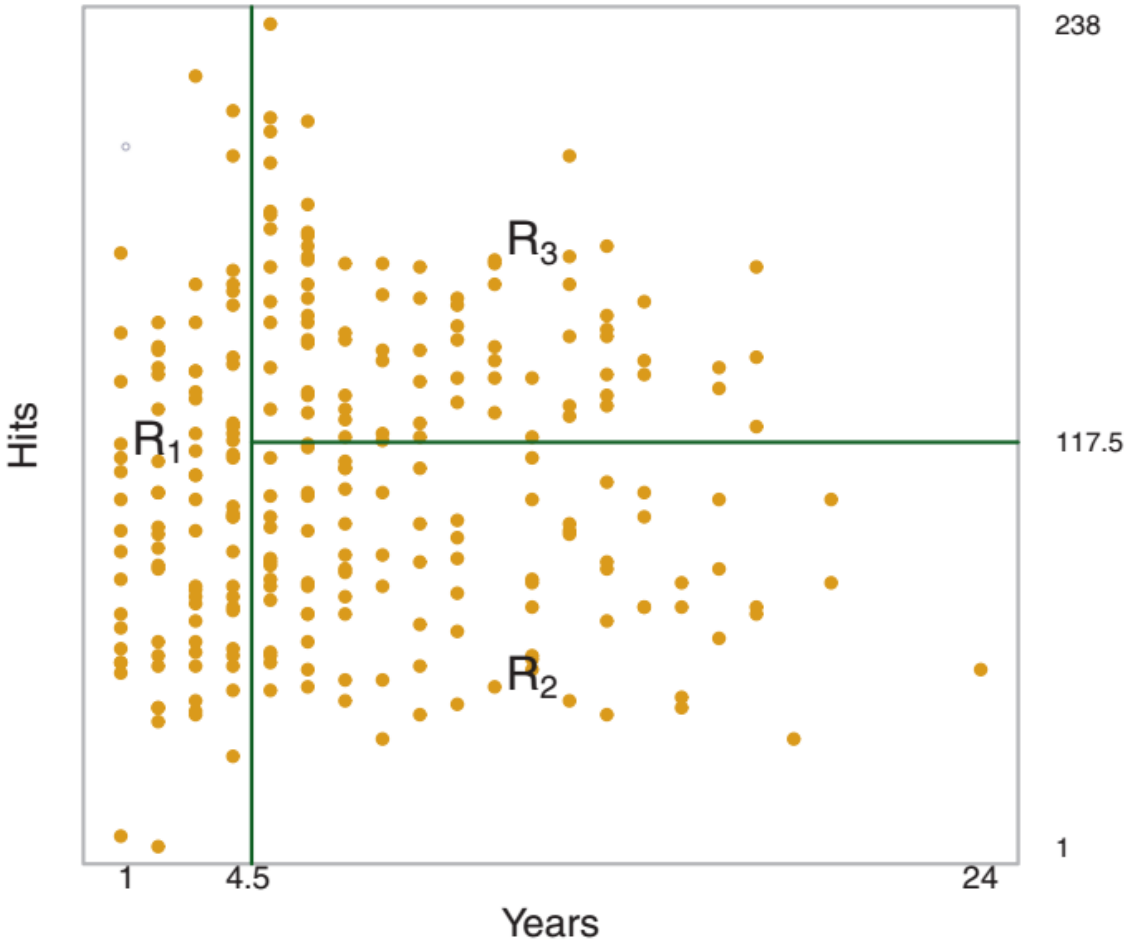
Ex. Prevendo o salário dos jogadores de Baseball usando árvore de regressão (exemplo: James et al (2013), pg. 303)

A ideia é prever o salário dos jogadores de baseball baseado em duas variáveis: anos de experiência (Years) e número de rebatidas no ano anterior (Hits)
A figura abaixo apresenta um exemplo de árvore em que o espaço previsor é dividido de acordo com as características das variáveis.

Introdução a árvore de regressão



Para ver como o espaço previsor é subdividido, observe a figura abaixo:



Introdução a árvore de regressão

A média do log salário dos jogadores que possuem menos de 4.5 anos de experiência é 5.107, o que implica em um salario medio de $\$1000 \times e^{5.107} = \$165.174$ dólares.

Para jogadores com mais de 4.5 anos de experiência, a região de previsão é subdividida pelo número de batidas no ano anterior.

Introdução a árvore de regressão

Ou seja, essas duas variáveis permitem escrever três regiões distintas.

$$R_1 = \{X | \ Years < 4.5\}$$

$$R_2 = \{X | \ Years \geq 4.5, \ Hits > 117.5\}$$

$$R_3 = \{X | \ Years \geq 4.5, \ Hits \leq 117.5\}$$

Introdução a árvore de regressão

Os salários médios de cada grupo de jogadores são respectivamente:

$$
\begin{aligned}
R_1: &\quad \$1000 \times e^{5.107} = \$165.174 \\
R_2: &\quad \$1000 \times e^{5.999} = \$402.834 \\
R_3: &\quad \$1000 \times e^{6.740} = \$845.346
\end{aligned}
$$

Introdução a árvore de regressão

Nomeclatura

1. As regioes R_1, R_2 e R_3 são chamadas folhas (ou nós terminais)
2. Os pontos nos quais o espaço previsor é dividido são chamados de nós internos ($Years = 4.5$ e $Hits = 117.5$)
3. Os segmentos de linha que seguem ate os nós terminais sao chamados de ramos (*branches*)

Introdução a árvore de regressão

Nomeclatura

1. As regioes R_1, R_2 e R_3 são chamadas folhas (ou nós terminais)
2. Os pontos nos quais o espaço previsor é dividido são chamados de nós internos ($Years = 4.5$ e $Hits = 117.5$)
3. Os segmentos de linha que seguem ate os nós terminais sao chamados de ramos (*branches*)

Introdução a árvore de regressão

Nomeclatura

1. As regioes R_1, R_2 e R_3 são chamadas folhas (ou nós terminais)
2. Os pontos nos quais o espaço previsor é dividido são chamados de nós internos ($Years = 4.5$ e $Hits = 117.5$)
3. Os segmentos de linha que seguem ate os nós terminais sao chamados de ramos (*branches*)

Introdução a árvore de regressão

Nomeclatura

1. As regioes R_1, R_2 e R_3 são chamadas folhas (ou nós terminais)
2. Os pontos nos quais o espaço previsor é dividido são chamados de nós internos ($Years = 4.5$ e $Hits = 117.5$)
3. Os segmentos de linha que seguem ate os nós terminais sao chamados de ramos (*branches*)

Introdução a árvore de regressão

A interpretação da árvore de regressão da figura 1 é a seguinte:

- Years é a variavel mais importante para determinar o salário, de modo que jogadores com menos experiência ganham menos que jogadores com mais experiência.
- No conjunto composto por jogadores com menos experiência, o número de rebatidas (Hits) não é relevante para o salário médio.
- No conjunto formado por jogadores mais experientes ($Years > 4.5$) o número de rebatidas do ano anterior é relevante para determinação do salário
  - Jogadores com mais rebatidas ($Hits > 117.5$) ganham em média mais que o dobro de salário que os jogadores com um número baixo de rebatidas ($845.346 = $402.834 = $442.512)

Introdução a árvore de regressão

A interpretação da árvore de regressão da figura 1 é a seguinte:

- Years é a variavel mais importante para determinar o salário, de modo que jogadores com menos experiência ganham menos que jogadores com mais experiência.
- No conjunto composto por jogadores com menos experiência, o número de rebatidas (Hits) não é relevante para o salário médio.
- No conjunto formado por jogadores mais experientes ($Years > 4.5$) o número de rebatidas do ano anterior é relevante para determinação do salário
  - Jogadores com mais rebatidas ($Hits > 117.5$) ganham em média mais que o dobro de salário que os jogadores com um número baixo de rebatidas ($845.346 = $402.834 = $442.512)

Introdução a árvore de regressão

A interpretação da árvore de regressão da figura 1 é a seguinte:

- Years é a variavel mais importante para determinar o salário, de modo que jogadores com menos experiência ganham menos que jogadores com mais experiência.
- No conjunto composto por jogadores com menos experiência, o número de rebatidas (Hits) não é relevante para o salário médio.
- No conjunto formado por jogadores mais experientes ($Years > 4.5$) o número de rebatidas do ano anterior é relevante para determinação do salário
  - Jogadores com mais rebatidas ($Hits > 117.5$) ganham em média mais que o dobro de salário que os jogadores com um número baixo de rebatidas ($845.346 = $402.834 = $442.512)

Introdução a árvore de regressão

A interpretação da árvore de regressão da figura 1 é a seguinte:

- Years é a variavel mais importante para determinar o salário, de modo que jogadores com menos experiência ganham menos que jogadores com mais experiência.
- No conjunto composto por jogadores com menos experiência, o número de rebatidas (Hits) não é relevante para o salário médio.
- No conjunto formado por jogadores mais experientes ($Years > 4.5$) o número de rebatidas do ano anterior é relevante para determinação do salário
  - Jogadores com mais rebatidas ($Hits > 117.5$) ganham em média mais que o dobro de salário que os jogadores com um número baixo de rebatidas ($\$845.346 - \$402.834 = \$442.512$)

Introdução a árvore de regressão

A interpretação da árvore de regressão da figura 1 é a seguinte:

- Years é a variavel mais importante para determinar o salário, de modo que jogadores com menos experiência ganham menos que jogadores com mais experiência.
- No conjunto composto por jogadores com menos experiência, o número de rebatidas (Hits) não é relevante para o salário médio.
- No conjunto formado por jogadores mais experientes ($Years > 4.5$) o número de rebatidas do ano anterior é relevante para determinação do salário
  - Jogadores com mais rebatidas ($Hits > 117.5$) ganham em média mais que o dobro de salário que os jogadores com um número baixo de rebatidas ($\$845.346 - \$402.834 = \$442.512$)

Introdução a árvore de regressão

Características do método de árvore de decisão

- Facil interpretação
- Representável por um gráfico simples
- Uma vez encontrados os nós internos, o cálculo subsequente é simples.

Introdução a árvore de regressão

Características do método de árvore de decisão

- Facil interpretação
- Representável por um gráfico simples
- Uma vez encontrados os nós internos, o cálculo subsequente é simples.

Introdução a árvore de regressão

Características do método de árvore de decisão

- Facil interpretação
- Representável por um gráfico simples
- Uma vez encontrados os nós internos, o cálculo subsequente é simples.

Introdução a árvore de regressão

Características do método de árvore de decisão

- Facil interpretação
- Representável por um gráfico simples
- Uma vez encontrados os nós internos, o cálculo subsequente é simples.

Introdução a árvore de regressão

Etapas para construção da árvore de decisão

1. Dividir o espaço previsor formado a partir das variáveis $X_1, ..., X_p$ em J regiões distintas e não sobrepostas, $R_1, ..., R_J$.
2. Para cada região R_j a mesma previsão é atribuída.

Introdução a árvore de regressão

Etapas para construção da árvore de decisão

1. Dividir o espaço previsor formado a partir das variáveis $X_1, ..., X_p$ em J regiões distintas e não sobrepostas, $R_1, ..., R_J$.
2. Para cada região R_j a mesma previsão é atribuída.

Introdução a árvore de regressão

Etapas para construção da árvore de decisão

1. Dividir o espaço previsor formado a partir das variáveis $X_1, ..., X_p$ em J regiões distintas e não sobrepostas, $R_1, ..., R_J$.
2. Para cada região R_j a mesma previsão é atribuída.

Introdução a árvore de regressão

DETERMINAÇÃO DAS REGIÕES

O objetivo de realizar qualquer separação do espaço previsor de forma a minimizar a seguinte expressão:

$$\sum_{j=1}^{J}\sum_{i\in R_j}(y_i - \hat{y}_{R_j})^2$$

Entretanto, tal abordagem é computacionalmente infactível quando existem muitas observações (n) e muitas variáveis (p). Por isso, adota-se a abordagem conhecida como divisão binária recursiva

DIVISÃO BINÁRIA RECURSIVA

Trata-se de uma abordagem *top-down* que permite dividir o espaço de previsão. O método inicia fazendo a melhor divisão de uma única variável.

Posteriormente, subdivide as regiões obtidas em outras a partir das novas variáveis.

Seja $X_1, ..., X_p$ o conjunto de variáveis. Uma variável é escolhida, digamos X_j.

Um ponto de corte s é escolhido de forma a reduzir a SQR.

Duas regiões são obtidas: $R_1(j, s) = \{X|X_j < s\}$ e $R_2(j, s) = \{X|X_j \geq s\}$.
Assim, s é obtido de forma a minimizar:

$$\sum_{i:\, x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:\, x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2$$

Em que: $\hat{y}_{R_j}$ representa a média da resposta do grupo j.

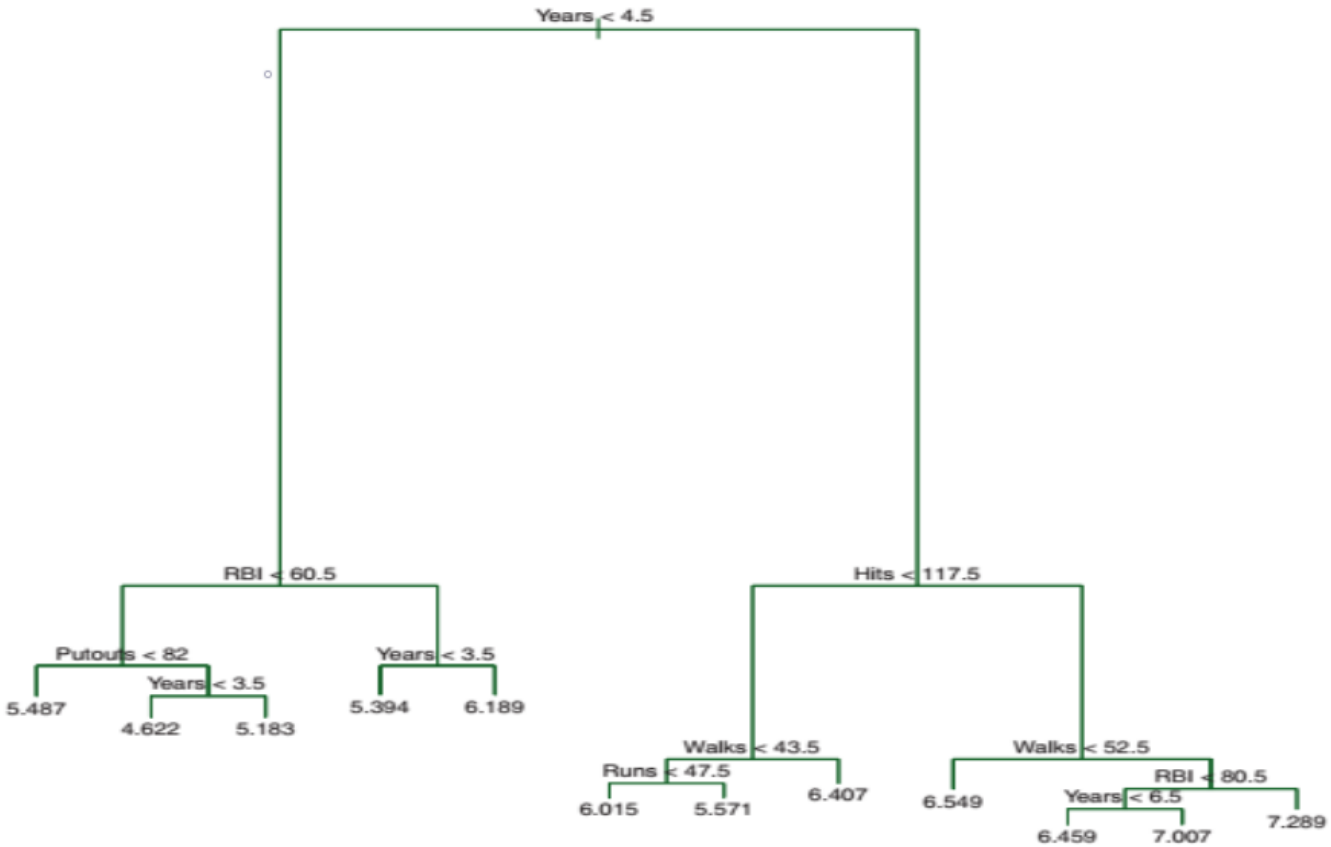
Este procedimento é realizado para todas as variáveis e a variável inicial escolhida é aquela que minimiza a expressão acima.

Após o primeiro ponto de corte, o procedimento repete-se para as subregiões e repete-se até um critério de parada ser alcançado.

Depois que todas as regiões são construídas, pode-se calcular a previsão para cada região.

Exemplo

Usando a base de dados Hitters a arvore é dada por:



Em que:
RBI = run batted in
Runs = Corridas

- **VANTAGENS**

1. Árvores são simples de explicar para não especialistas
2. Árvores podem ser dispostas graficamente gerando uma forma amigável de visualização
3. Variáveis qualitativas são mais fácieis de serem tratadas. Não requerem a criação de variáveis dummies.

- **DESVANTAGENS**

1. Árvores não possuem o desempenho preditivo semelhante aos modelos como os métodos de regularização e modelos lineares.

- **VANTAGENS**

1. Árvores são simples de explicar para não especialistas
2. Árvores podem ser dispostas graficamente gerando uma forma amigável de visualização
3. Variáveis qualitativas são mais fácieis de serem tratadas. Não requerem a criação de variáveis dummies.

- **DESVANTAGENS**

1. Árvores não possuem o desempenho preditivo semelhante aos modelos como os métodos de regularização e modelos lineares.

- **VANTAGENS**

1. Árvores são simples de explicar para não especialistas
2. Árvores podem ser dispostas graficamente gerando uma forma amigável de visualização
3. Variáveis qualitativas são mais fácieis de serem tratadas. Não requerem a criação de variáveis dummies.

- **DESVANTAGENS**

1. Árvores não possuem o desempenho preditivo semelhante aos modelos como os métodos de regularização e modelos lineares.

- **VANTAGENS**

1. Árvores são simples de explicar para não especialistas
2. Árvores podem ser dispostas graficamente gerando uma forma amigável de visualização
3. Variáveis qualitativas são mais fácieis de serem tratadas. Não requerem a criação de variáveis dummies.

- **DESVANTAGENS**

1. Árvores não possuem o desempenho preditivo semelhante aos modelos como os métodos de regularização e modelos lineares.

- **VANTAGENS**

1. Árvores são simples de explicar para não especialistas
2. Árvores podem ser dispostas graficamente gerando uma forma amigável de visualização
3. Variáveis qualitativas são mais fácieis de serem tratadas. Não requerem a criação de variáveis dummies.

- **DESVANTAGENS**

1. Árvores não possuem o desempenho preditivo semelhante aos modelos como os métodos de regularização e modelos lineares.

- **VANTAGENS**

1. Árvores são simples de explicar para não especialistas
2. Árvores podem ser dispostas graficamente gerando uma forma amigável de visualização
3. Variáveis qualitativas são mais fácieis de serem tratadas. Não requerem a criação de variáveis dummies.

- **DESVANTAGENS**

1. Árvores não possuem o desempenho preditivo semelhante aos modelos como os métodos de regularização e modelos lineares.

- **VANTAGENS**

1. Árvores são simples de explicar para não especialistas
2. Árvores podem ser dispostas graficamente gerando uma forma amigável de visualização
3. Variáveis qualitativas são mais fácieis de serem tratadas. Não requerem a criação de variáveis dummies.

- **DESVANTAGENS**

1. Árvores não possuem o desempenho preditivo semelhante aos modelos como os métodos de regularização e modelos lineares.

Árvores para classificação

Se o objetivo é criar árvores para a classificação, e não para a regressão, necessita-se apenas modificar a medida do custo-complexidade da árvore.

Em essência o algoritmo de construção da árvore continua o mesmo.

Seja o nó m representando a região R_m, com N_m observações.

$$\hat{p}_{m,k} = \frac{1}{N_m} \sum_{X_i \in R_m} I(y_i = k)$$

Em que: k representa o caso de interesse e $\hat{p}_{m,k}$ é a proporção estimada de casos de sucesso em cada nó.

Existem três principais medidas de erro de classificação que geralmente são usadas para a construção de árvores:

- Erro de classificação: $\frac{1}{N_m} \sum_{X_i \in R_m} I(y_i \neq k) = 1 - \hat{p}_{m,k}$
- Índice de Gini: $\sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$
- Entropia cruzada (ou *deviance*) $-\sum_{k=1}^{K} \hat{p}_{m,k} \times log\, \hat{p}_{m,k}$

Existem três principais medidas de erro de classificação que geralmente são usadas para a construção de árvores:

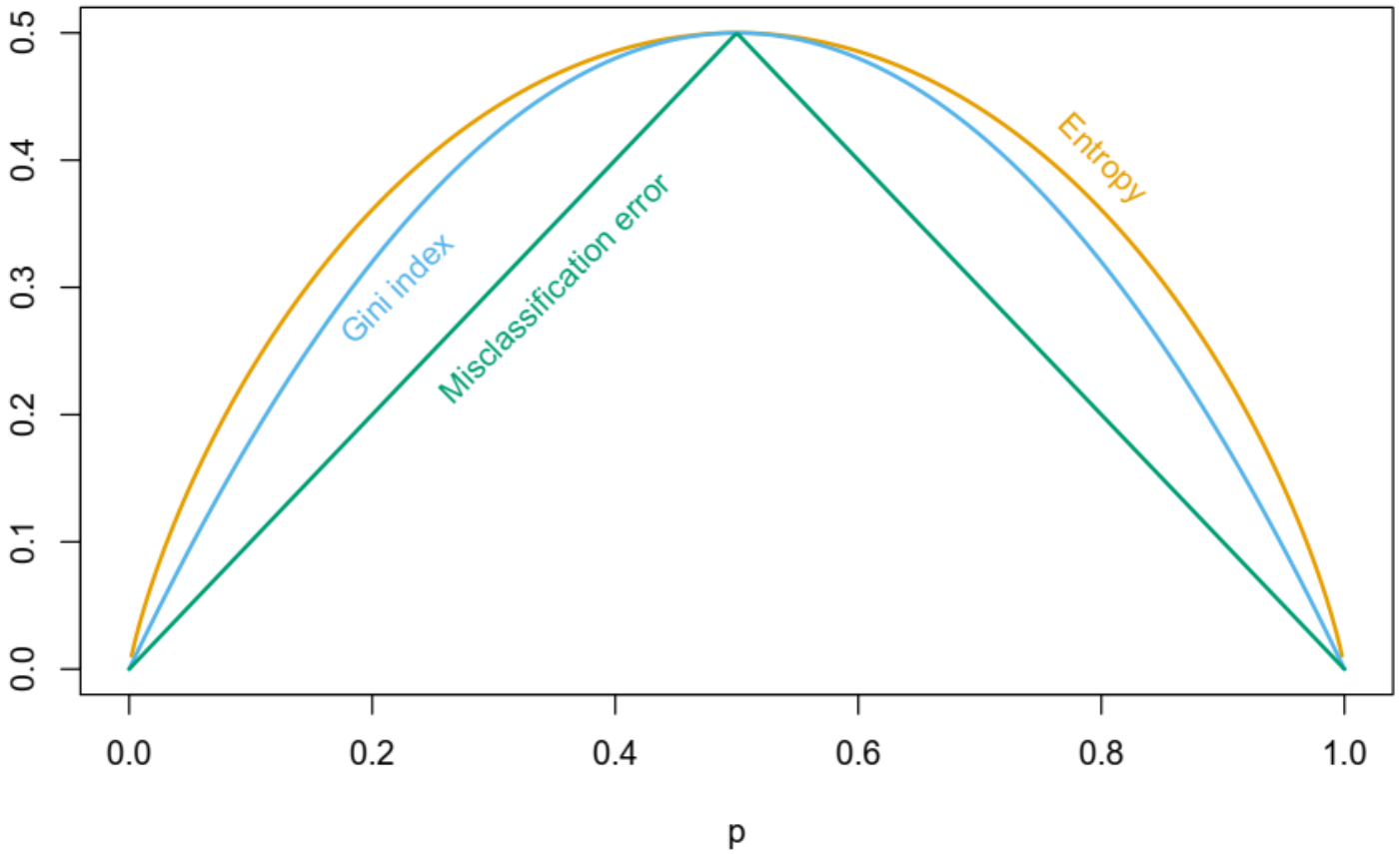
- Erro de classificação: $\frac{1}{N_m} \sum_{X_i \in R_m} I(y_i \neq k) = 1 - \hat{p}_{m,k}$
- Índice de Gini: $\sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$
- Entropia cruzada (ou *deviance*) $-\sum_{k=1}^{K} \hat{p}_{m,k} \times log\ \hat{p}_{m,k}$

Existem três principais medidas de erro de classificação que geralmente são usadas para a construção de árvores:

- Erro de classificação: $\frac{1}{N_m} \sum_{X_i \in R_m} I(y_i \neq k) = 1 - \hat{p}_{m,k}$
- Índice de Gini: $\sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$
- Entropia cruzada (ou *deviance*) $-\sum_{k=1}^{K} \hat{p}_{m,k} \times log\ \hat{p}_{m,k}$

Existem três principais medidas de erro de classificação que geralmente são usadas para a construção de árvores:

- Erro de classificação: $\frac{1}{N_m} \sum_{X_i \in R_m} I(y_i \neq k) = 1 - \hat{p}_{m,k}$
- Índice de Gini: $\sum_{k=1}^{K} \hat{p}_{m,k}(1 - \hat{p}_{m,k})$
- Entropia cruzada (ou *deviance*) $-\sum_{k=1}^{K} \hat{p}_{m,k} \times log\, \hat{p}_{m,k}$


0.0
0.1
0.2
0.3
0.4
0.5
0.0
0.2
0.4
0.6
0.8
1.0
p
Gini index
Misclassification error
Entropy

BAGGING

O termo bagging é uma abreviação para a agregação via bootstrap.

Em tese, a agregação de previsões tem perfomance melhor que previsões individuais.

A ideia de que a agregação melhora a previsão decorre da distribuição amostral da média.

Considere que Z_1, Z_2, ..., Z_n seja um conjunto de variaveis aleatórias independentes com mesma variância σ^2.

Então, $(1/n)\sum_{i=1}^{n} Z_i$ possui variância σ^2/n.

Portanto, se $n \to \infty$, então a variância de $(1/n)\sum_{i=1}^{n} Z_i$ tende para zero.

Assim, o procedimento de agregação consiste em fazer várias previsões usando potenciais amostras de treino: $\hat{f}^1(x)$, $\hat{f}^2(x)$,...,$\hat{f}^B(x)$ e posteriormente tirar a média dessas previsões.

$$\hat{f}_{avg}(x) = \left(\frac{1}{B}\right)\sum_{i=1}^{B}\hat{f}^i(x)$$

Normalmente utiliza-se dois procedimentos distintos para fazer a agregação:

1. Usando um conjunto de dados $X = \{x_1, ..., x_p\}$ aplicá-se diferentes previsores $f^i(X)$ e realiza-se a agregadação, podendo ser ponderada ou média simples.
2. Usando um conjunto de dados $X = \{x_1, ..., x_p\}$aplicá-se um mesmo previsor $f(X)$ para diferentes sub-amostras de X, $f(X \mid S)$ e realizá-se a agregação.

Normalmente utiliza-se dois procedimentos distintos para fazer a agregação:

1. Usando um conjunto de dados $X = \{x_1, ..., x_p\}$ aplicá-se diferentes previsores $f^i(X)$ e realiza-se a agregadação, podendo ser ponderada ou média simples.
2. Usando um conjunto de dados $X = \{x_1, ..., x_p\}$aplicá-se um mesmo previsor $f(X)$ para diferentes sub-amostras de X, $f(X \mid S)$ e realizá-se a agregação.

Normalmente utiliza-se dois procedimentos distintos para fazer a agregação:

1. Usando um conjunto de dados $X = \{x_1, ..., x_p\}$ aplicá-se diferentes previsores $f^i(X)$ e realiza-se a agregadação, podendo ser ponderada ou média simples.
2. Usando um conjunto de dados $X = \{x_1, ..., x_p\}$aplicá-se um mesmo previsor $f(X)$ para diferentes sub-amostras de X, $f(X \mid S)$ e realizá-se a agregação.

O problema do último procedimento de agregação é que não temos acesso a várias amostras de treino aleatórias da mesma distribuição populacional.

O procedimento de bagging simula as amostras de treino utilizando o algoritmo de bootstrap.

O procedimento de bagging pode ser utilizado em diferentes modelos: modelos de regressão, modelos de classificação, modelos de árvores etc.

É possível mostrar que todo procedimento de bagging utiliza apenas 2/3 das observações.
Isto é, existirão 1/3 de observações que não são utilizadas na agregação bagging. Essas observações deixadas de fora são chamadas de *out-of-bag* (OOB) e podem ser utilizadas para fazer testes de valildação.

ALGORITMO DO BAGGING

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j estime o modelo de previsão (Podendo ser regressão linear ou outro método qualquer como árvores), $\hat{f}_j(x_{ij})$, $j = 1, ..., B$
- $\hat{f}_{bag} = \frac{1}{B}\sum_{j=1}^{B} \hat{f}_j(x_{ij})$

ALGORITMO DO BAGGING

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j estime o modelo de previsão (Podendo ser regressão linear ou outro método qualquer como árvores), $\hat{f}_j(x_{ij})$, $j = 1, ..., B$
- $\hat{f}_{bag} = \frac{1}{B}\sum_{j=1}^{B} \hat{f}_j(x_{ij})$

ALGORITMO DO BAGGING

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j estime o modelo de previsão (Podendo ser regressão linear ou outro método qualquer como árvores), $\hat{f}_j(x_{ij})$, $j = 1, ..., B$
- $\hat{f}_{bag} = \frac{1}{B}\sum_{j=1}^{B} \hat{f}_j(x_{ij})$

ALGORITMO DO BAGGING

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j estime o modelo de previsão (Podendo ser regressão linear ou outro método qualquer como árvores), $\hat{f}_j(x_{ij})$, $j = 1, ..., B$
- $\hat{f}_{bag} = \frac{1}{B}\sum_{j=1}^{B} \hat{f}_j(x_{ij})$

ALGORITMO DO BAGGING

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j estime o modelo de previsão (Podendo ser regressão linear ou outro método qualquer como árvores), $\hat{f}_j(x_{ij})$, $j = 1, ..., B$
- $\hat{f}_{bag} = \frac{1}{B}\sum_{j=1}^{B} \hat{f}_j(x_{ij})$

Para entendermos por que métodos de agregação tendem a melhorar a previsão, considere a medida de erro quadrático médio.

Seja (y_i, x_i) um conjunto de observações independentes e identicamente distribuídas em $\mathcal{O}$.

Considere $f_{ag}(x) = E_{\mathcal{O}}\hat{f}(x^*)$ uma agregação obtida do procedimento de bootstrap amostrado sobre $\mathcal{O}$.

Observe que o procedimento de bootstrap é obtido da população $\mathcal{O}$ e não sobre a amostra de treino. Isso é importante por que na prática não é realizado desta forma.

$$E_{\mathcal{O}}[Y - \hat{f}(x)]^2 = E_{\mathcal{O}}[Y - f_{ag}(x) + f_{ag}(x) - \hat{f}(x)]^2$$

$$= E_{\mathcal{O}}[Y - f_{ag}(x)]^2 + E_{\mathcal{O}}[f_{ag}(x) - \hat{f}(x)]^2$$

$$\geq E_{\mathcal{O}}[Y - f_{ag}(x)]^2$$

Ou seja,
Caso o procedimento de bootstrap seja adequado, bagging nunca aumentará o erro quadrático médio.

FLORESTA ALETÓRIA

Floresta aleátoria (Random Forest) consiste num método para melhorar o desempenho das árvores de regressão, proposto por Breiman (2001).

O problema principal do método de bagging é que as árvores construídas a partir de um procedimento de bootstrap são fortemente correlacionadas entre si.

Seja B iid variáveis aleatórias, com variância individual σ^2. Portanto, a média das variâncias será $\frac{1}{B}\sigma^2$.

Se as variáveis forem apenas identicamente distribuídas, mas não independentes, então existe possibilidade de correlação entre as B variáveis aleatórias. Sendo tal correlação igual a ρ, logo a média das variáveis será:

$$\rho\sigma^2 + \frac{1}{B}\sigma^2$$

O procedimento de bagging reduz a variância da previsão ao se fazer $B \to \infty$. Mas o primeiro termo não é alterado pelo bagging.

A essência do Random Forest é melhorar a previsão da agregação ao se reduzir a correlação entre as árvores.

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

ALGORITMO DO RANDOM FOREST

- Seja uma amostra de treino $(Y,\ X) = \{(y_1,\ x_1), ..., (y_n,\ x_n)\}$ com Y a variável de interesse e X o preditor
- Usando bootstrap, construa B sub-conjuntos de $(Y,\ X)$, digamos $(Y_j,\ X_j)$, $j = 1, ..., B$
- Para cada sub-conjunto j construa uma árvore até alcançar um nó terminal adequado, seguindo a cada nó o seguinte procedimento.
  1. Selecione m variáveis aleatoriamente das p variáveis disponíveis.
  2. Das m variáveis, escolha a melhor e construa um nó.
  3. Repita o procedimento até alcançar o nó terminal
- Chame as árvores obtidas de $\{T_j\}_{j=1}^{B}$
- Realiza a previsão por agregação: $\hat{f}_{rf} = \frac{1}{B}\sum^{B} T_i(x)$

Aplicação do método de árvores na avaliação de políticas públicas

Um problema interessante em avaliação de políticas públicas é identificar até que ponto o efeito tratamento depende das variáveis de controle.
Em essência este é um problema do tipo CATE:

$$\tau(x) = E[Y_i(1) - Y_i(0) | X_i = x]$$

$$\tau = E[\tau(x)]$$

$$\tau(x) = E[\tau_i | X_i = x] = E[Y_i(1) - Y_i(0) | X_i = x]$$

$$= E[Y_i(1) | X_i = x] - E[Y_i(0) | X_i = x] \qquad \textit{linearidade}$$

$$= E[Y_i(1) | W_i = 1, X_i = x] - E[Y_i(0) | W_i = 0, X_i = x] \qquad \textit{ignorabilidade}$$

$$= \mu(1, x) - \mu(0, x)$$

Será que para diferentes grupos de indivíduos o efeito tratamento é diferente?
Isto é:

$$\tau(x) = \mu(1, x) - \mu(0, x)$$

Problemas empíricos tradicionais em economia não enfrentam dificuldade para estimar pois $\mu(W, x)$ pode ser estimado sob a hipótese do ignorabilidade. Entretanto, duas situações podem dificultar o cálculo do efeito heterogêneo.

1. O problema da maldição da dimensionalidade (*Curse of Dimensionality*) que prejudica o desempenho dos tradicionais.
2. O cálculo da heterogeneidade em dados experimentais não é direta.

Athey e Imbens (2016) e Athey e Wager (2017) optam por tratar deste problema por meio de métodos de árvores, o que eles chamam de causal tree ou causal forest.

Considere o seguinte setup:

- É observado a seguinte tripla $\{(W_i, Y_i, X_i)\}_{i=1}^{N}$
- em que: W_i representa se o indivíduo i é tratado,
- Y_i é uma variável binária indicando o resultado do tratamento $(Y_i(1), Y_i(0))$
- X_i é um conjunto de variáveis de controle para o indivíduo i
- O conjunto de dados é do tipo experimental, isto é, a hipótese do ignorabilidade é satisfeita

Considere o seguinte setup:

- É observado a seguinte tripla $\{(W_i, Y_i, X_i)\}_{i=1}^{N}$
- em que: W_i representa se o indivíduo i é tratado,
- Y_i é uma variável binária indicando o resultado do tratamento $(Y_i(1), Y_i(0))$
- X_i é um conjunto de variáveis de controle para o indivíduo i
- O conjunto de dados é do tipo experimental, isto é, a hipótese do ignorabilidade é satisfeita

Considere o seguinte setup:

- É observado a seguinte tripla $\{(W_i, Y_i, X_i)\}_{i=1}^{N}$
- em que: W_i representa se o indivíduo i é tratado,
- Y_i é uma variável binária indicando o resultado do tratamento $(Y_i(1), Y_i(0))$
- X_i é um conjunto de variáveis de controle para o indivíduo i
- O conjunto de dados é do tipo experimental, isto é, a hipótese do ignorabilidade é satisfeita

Considere o seguinte setup:

- É observado a seguinte tripla $\{(W_i, Y_i, X_i)\}_{i=1}^{N}$
- em que: W_i representa se o indivíduo i é tratado,
- Y_i é uma variável binária indicando o resultado do tratamento $(Y_i(1), Y_i(0))$
- X_i é um conjunto de variáveis de controle para o indivíduo i
- O conjunto de dados é do tipo experimental, isto é, a hipótese do ignorabilidade é satisfeita

Considere o seguinte setup:

- É observado a seguinte tripla $\{(W_i, Y_i, X_i)\}_{i=1}^{N}$
- em que: W_i representa se o indivíduo i é tratado,
- Y_i é uma variável binária indicando o resultado do tratamento $(Y_i(1), Y_i(0))$
- X_i é um conjunto de variáveis de controle para o indivíduo i
- O conjunto de dados é do tipo experimental, isto é, a hipótese do ignorabilidade é satisfeita

Considere o seguinte setup:

- É observado a seguinte tripla $\{(W_i, Y_i, X_i)\}_{i=1}^{N}$
- em que: W_i representa se o indivíduo i é tratado,
- Y_i é uma variável binária indicando o resultado do tratamento $(Y_i(1), Y_i(0))$
- X_i é um conjunto de variáveis de controle para o indivíduo i
- O conjunto de dados é do tipo experimental, isto é, a hipótese do ignorabilidade é satisfeita

O objetivo é estimar o CATE

$$\tau(x) = E[\tau_i | X_i = x] = E[Y_i(1) - Y_i(0) | X_i = x]$$

O problema de aplicar diretamente o método de árvores é que tal método não foi desenvolvido para calcular o efeito causal.

Então, a separação do espaço de variáveis é obtida aumentando o viés e reduzindo ao máximo a variância.

Todavia, na análise de efeito tratamento o foco é justamente o contrário: o importante é ter viés mínimo.

Athey e Imbens (2016) propuseram uma forma de aplicação dos métodos de árvores que reduz o viés da separação do espaço das variáveis.

Tal método foi chamado de estimação "honesta" do espaço particionado.

O método consiste nos seguintes passos:

1. Separa-se a amostra em três conjunto diferentes sem interseção entre si: conjunto de treino, conjunto de teste e conjunto de estimação
2. A árvore é construida utilizando métodos tradicionais de ML: treina uma árvore no conjunto de treino e verifica seu poder de previsão no conjunto de teste
3. Por fim, o conjunto de estimação utilizado para estimar o efeito tratamento. Quando aplicado à árvore anteriormente particionada tem-se o efeito heterogêneo.

O método consiste nos seguintes passos:

1. Separa-se a amostra em três conjunto diferentes sem interseção entre si: conjunto de treino, conjunto de teste e conjunto de estimação
2. A árvore é construida utilizando métodos tradicionais de ML: treina uma árvore no conjunto de treino e verifica seu poder de previsão no conjunto de teste
3. Por fim, o conjunto de estimação utilizado para estimar o efeito tratamento. Quando aplicado à árvore anteriormente particionada tem-se o efeito heterogêneo.

O método consiste nos seguintes passos:

1. Separa-se a amostra em três conjunto diferentes sem interseção entre si: conjunto de treino, conjunto de teste e conjunto de estimação
2. A árvore é construida utilizando métodos tradicionais de ML: treina uma árvore no conjunto de treino e verifica seu poder de previsão no conjunto de teste
3. Por fim, o conjunto de estimação utilizado para estimar o efeito tratamento. Quando aplicado à árvore anteriormente particionada tem-se o efeito heterogêneo.

O método consiste nos seguintes passos:

1. Separa-se a amostra em três conjunto diferentes sem interseção entre si: conjunto de treino, conjunto de teste e conjunto de estimação
2. A árvore é construida utilizando métodos tradicionais de ML: treina uma árvore no conjunto de treino e verifica seu poder de previsão no conjunto de teste
3. Por fim, o conjunto de estimação utilizado para estimar o efeito tratamento. Quando aplicado à árvore anteriormente particionada tem-se o efeito heterogêneo.

A intuição deste procedimento é que uma vez que o conjunto de estimação não é utilizado para determinar a complexidade da árvore, então, o cálculo do efeito tratamento não será afetado em termos de redução de viés.

Este procedimento foi estendido por Athey e Wager (2017) por meio de florestas aleatórias. A diferença essêncial é que ao invés da construção de uma única árvore é construído uma família de árvores e a árvore resultante é obtida ao se combinar todas as árvores da família.

Artigo: Do Planning Prompts Increase Educational Success? Evidence from Randomized Controlled Trials in MOOCs (Andor et al (2019))

Motivação

- Estudantes que fazem MOOCs tem elevada taxas de evasão
- Uma explicação: custo de oportunidade baixo e loss-aversion

Artigo: Do Planning Prompts Increase Educational Success? Evidence from Randomized Controlled Trials in MOOCs (Andor et al (2019))

Motivação

- Estudantes que fazem MOOCs tem elevada taxas de evasão
- Uma explicação: custo de oportunidade baixo e loss-aversion

Artigo: Do Planning Prompts Increase Educational Success? Evidence from Randomized Controlled Trials in MOOCs (Andor et al (2019))

Motivação

- Estudantes que fazem MOOCs tem elevada taxas de evasão
- Uma explicação: custo de oportunidade baixo e loss-aversion

Artigo: Do Planning Prompts Increase Educational Success? Evidence from Randomized Controlled Trials in MOOCs (Andor et al (2019))

Experimento

- Elevar a participação dos estudante solicitando que eles determinem uma data para realizar a próxima aula (prompt) e os informando sobre essa data e a importância de ter metas (email).
- Experimento randomizado realizado nas plataformas: openHPI e openSAP

Artigo: Do Planning Prompts Increase Educational Success? Evidence from Randomized Controlled Trials in MOOCs (Andor et al (2019))

Experimento

- Elevar a participação dos estudante solicitando que eles determinem uma data para realizar a próxima aula (prompt) e os informando sobre essa data e a importância de ter metas (email).
- Experimento randomizado realizado nas plataformas: openHPI e openSAP

Artigo: Do Planning Prompts Increase Educational Success? Evidence from Randomized Controlled Trials in MOOCs (Andor et al (2019))

Experimento

- Elevar a participação dos estudante solicitando que eles determinem uma data para realizar a próxima aula (prompt) e os informando sobre essa data e a importância de ter metas (email).
- Experimento randomizado realizado nas plataformas: openHPI e openSAP

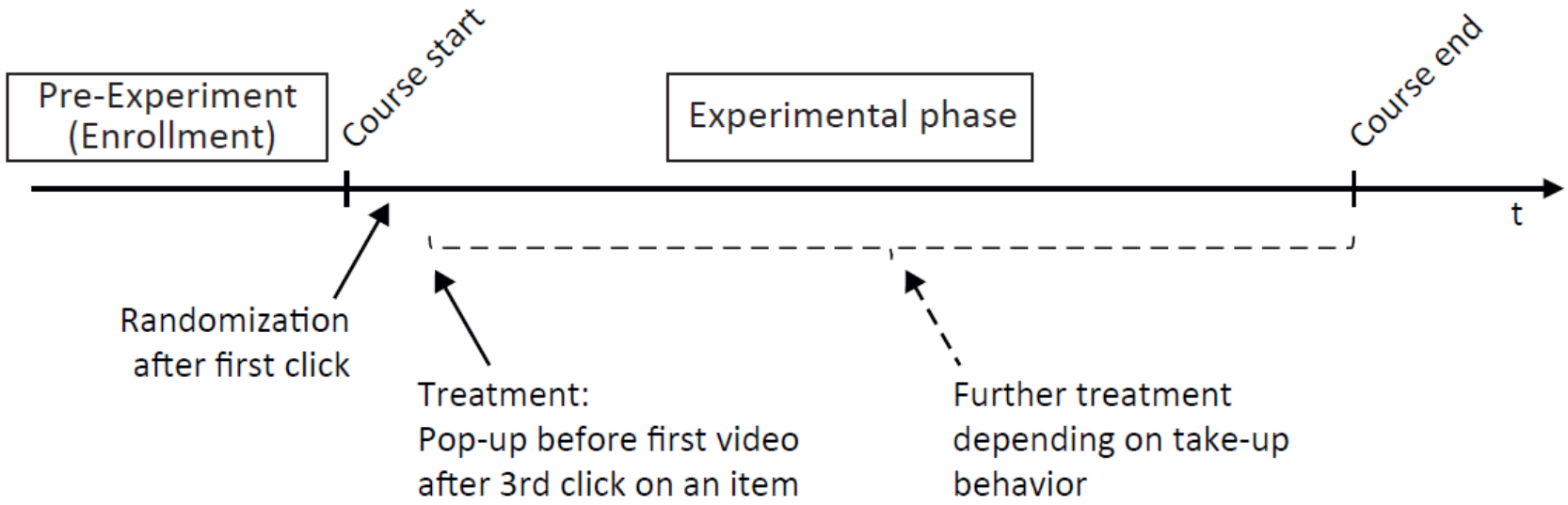


Figure 1: Timeline of the experiment

Figure 2: Treatment group pop-up

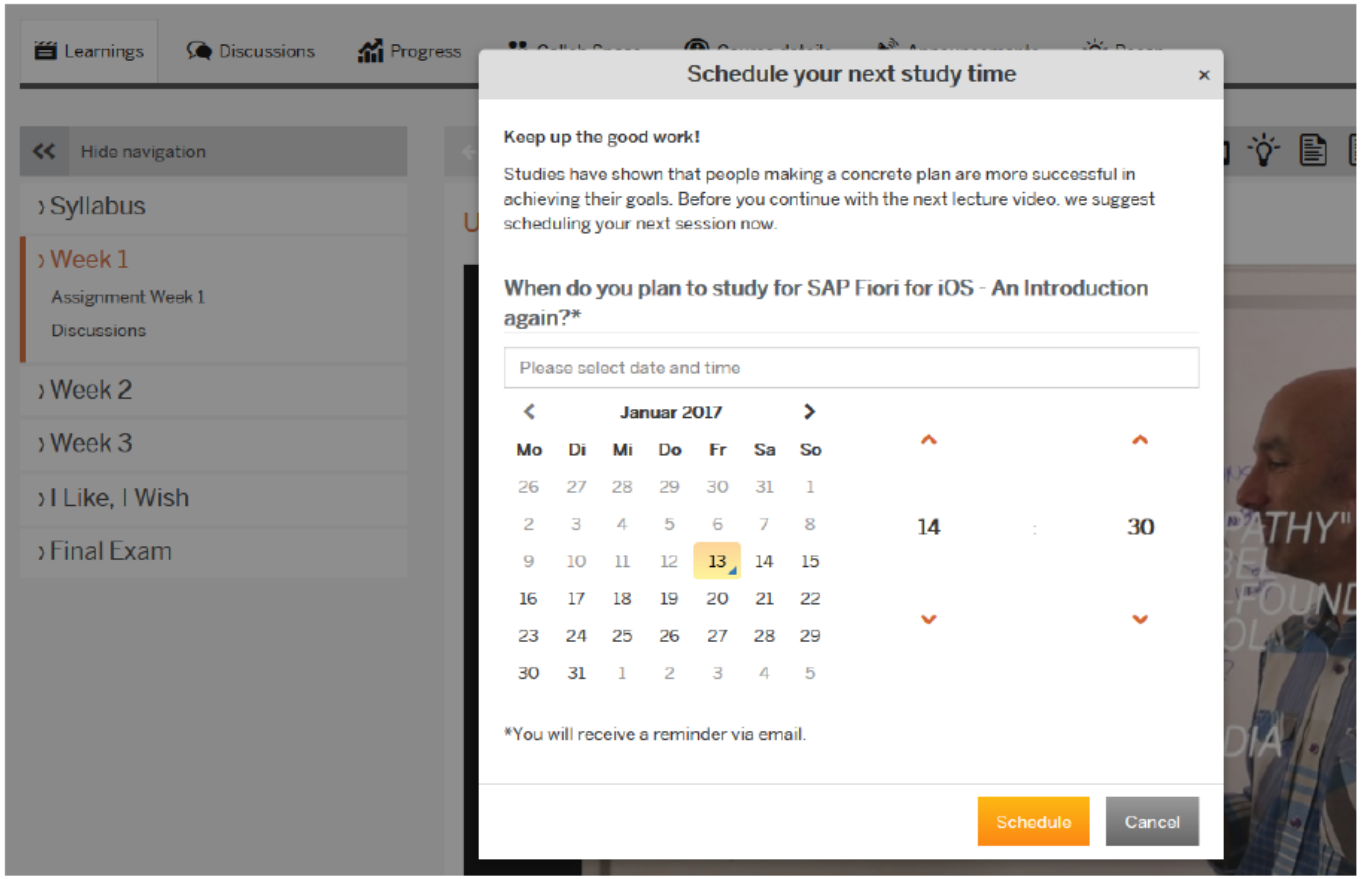
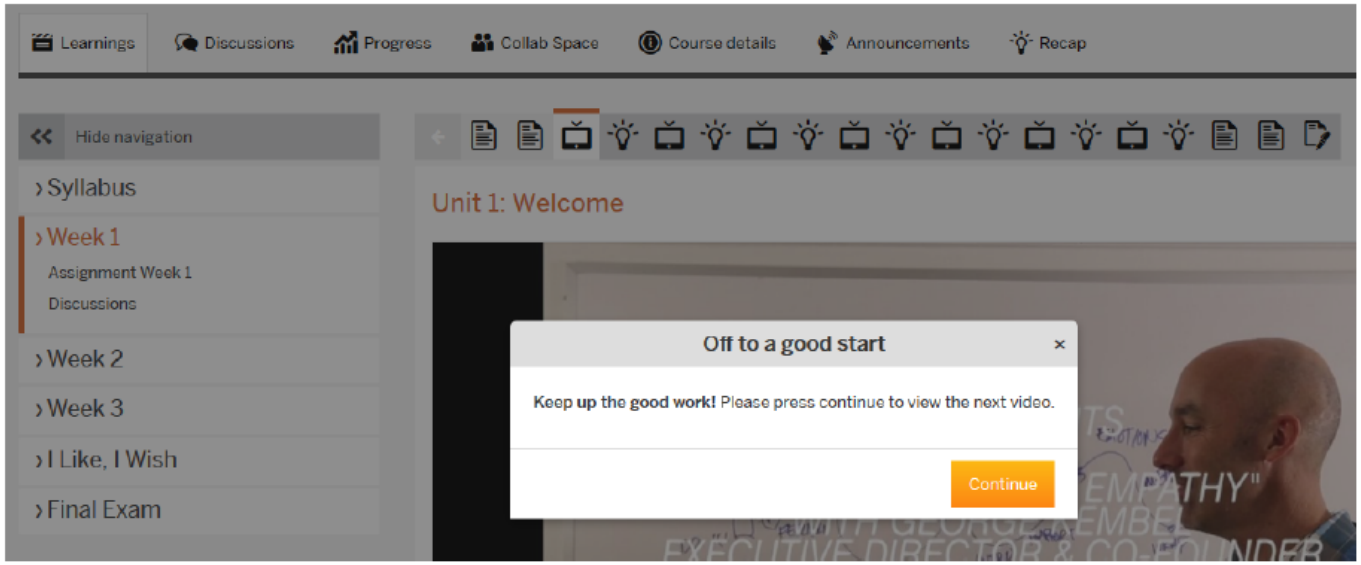


Figure 3: Control group pop-up



	(1) Certificate	(2) Sessions	(3) Duration	(4) Points	(5) Quizzes	(6) Videos	(7) Satisfaction	(8) Stress
Treatment	-0.005	0.066	0.002	-2.075	-0.558	0.078	0.001	-0.032
	(0.007)	(0.418)	(0.187)	(1.428)	(0.449)	(2.057)	(0.017)	(0.020)
Constant	0.299***	14.507***	5.290***	84.422***	19.135***	47.135***	0.503***	0.522***
	(0.007)	(0.320)	(0.135)	(1.783)	(0.377)	(1.994)	(0.019)	(0.020)
CI effect sizes	[-7%, 3%]	[-4%, 6%]	[-4%,6%]	[-8%,1%]	[-6%,2%]	[-8%,9%]	[-3%,3%]	[-7%,7%]
N	15574	15574	15574	15574	15574	15574	2679	2553

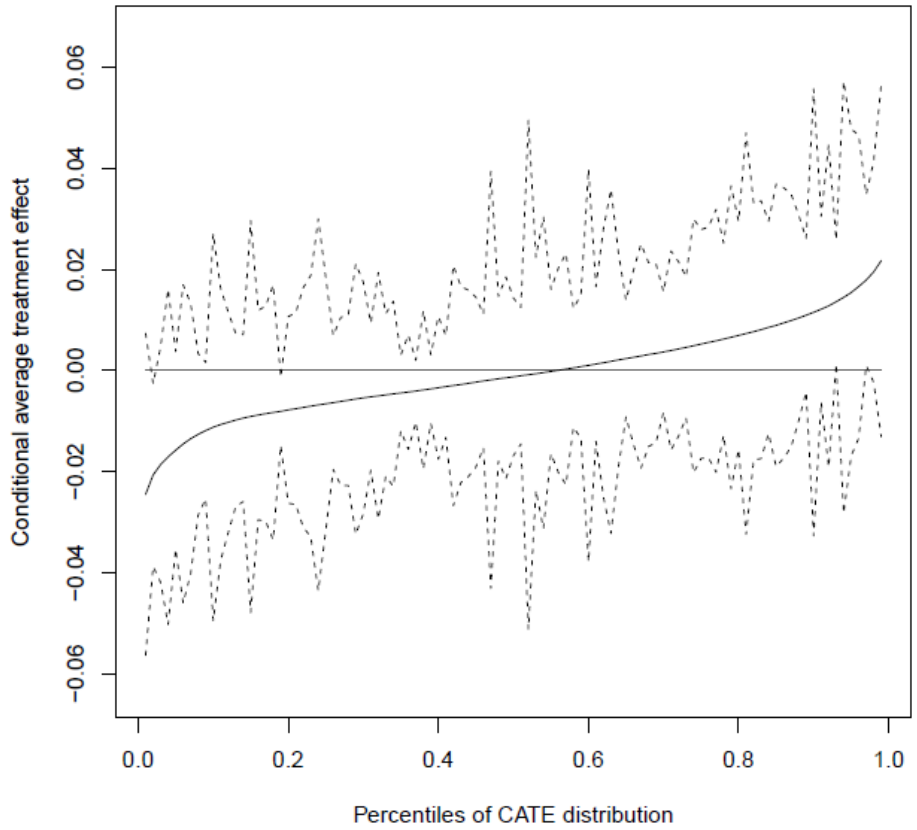
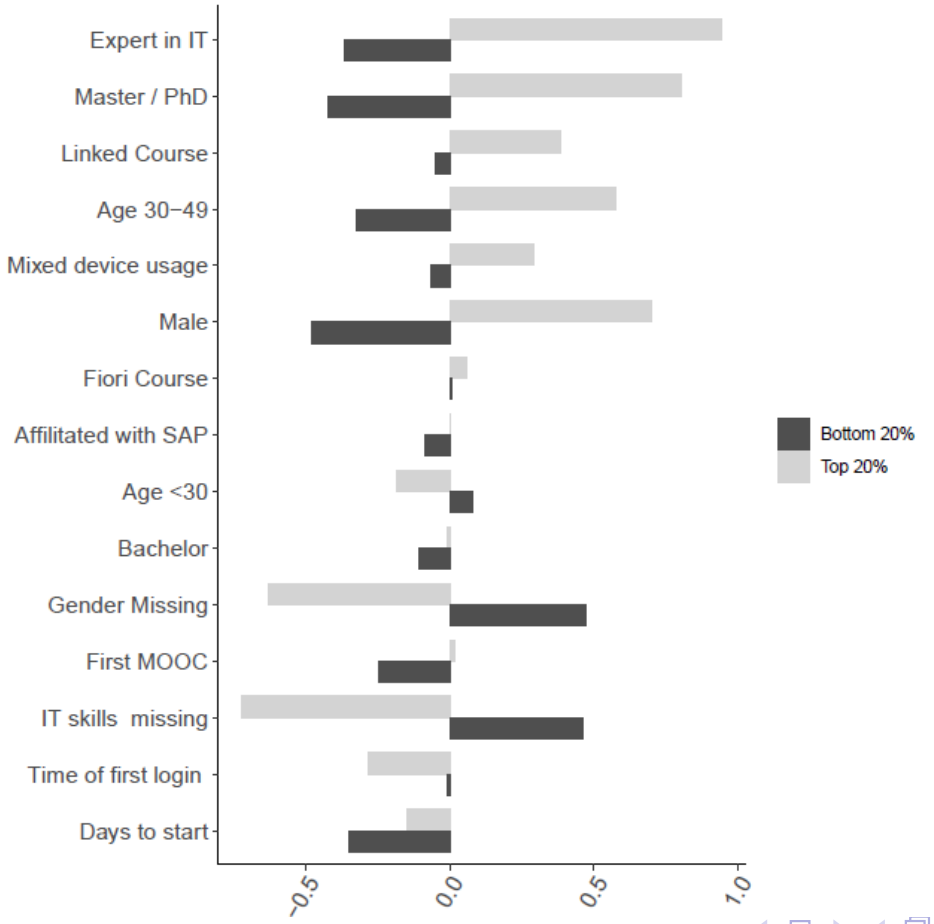

Conditional average treatment effect
0.06
0.04
0.02
0.00
−0.02
−0.04
−0.06
0.0
0.2
0.4
0.6
0.8
1.0
Percentiles of CATE distribution

Figure 6: Characteristics of the bottom and top of the CATE distribution



Encerramento do mini curso...

Muito ainda para estudar...

- Métodos de deep learning
- Aplicações em vários métodos de avaliação de políticas públicas, desde de RCT até controle sintético, RDD etc
  - Uma boa revisão desta literatura: The Impact of Machine Learning on Economics (Athey (2019))
- Apesar de ML não ajudar diretamente na identificação causal, pode contribuir indiretamente para a análise das estimativas causais.
  - Aplicável em situações com muitos dados, muitas informações e pouca estrutura para guiar análises

Muito ainda para estudar...

- Métodos de deep learning
- Aplicações em vários métodos de avaliação de políticas públicas, desde de RCT até controle sintético, RDD etc
  - Uma boa revisão desta literatura: The Impact of Machine Learning on Economics (Athey (2019))
- Apesar de ML não ajudar diretamente na identificação causal, pode contribuir indiretamente para a análise das estimativas causais.
  - Aplicável em situações com muitos dados, muitas informações e pouca estrutura para guiar análises

Muito ainda para estudar...

- Métodos de deep learning
- Aplicações em vários métodos de avaliação de políticas públicas, desde de RCT até controle sintético, RDD etc
  - Uma boa revisão desta literatura: The Impact of Machine Learning on Economics (Athey (2019))
- Apesar de ML não ajudar diretamente na identificação causal, pode contribuir indiretamente para a análise das estimativas causais.
  - Aplicável em situações com muitos dados, muitas informações e pouca estrutura para guiar análises

Muito ainda para estudar...

- Métodos de deep learning
- Aplicações em vários métodos de avaliação de políticas públicas, desde de RCT até controle sintético, RDD etc
  - Uma boa revisão desta literatura: The Impact of Machine Learning on Economics (Athey (2019))
- Apesar de ML não ajudar diretamente na identificação causal, pode contribuir indiretamente para a análise das estimativas causais.
  - Aplicável em situações com muitos dados, muitas informações e pouca estrutura para guiar análises

Muito ainda para estudar...

- Métodos de deep learning
- Aplicações em vários métodos de avaliação de políticas públicas, desde de RCT até controle sintético, RDD etc
  - Uma boa revisão desta literatura: The Impact of Machine Learning on Economics (Athey (2019))
- Apesar de ML não ajudar diretamente na identificação causal, pode contribuir indiretamente para a análise das estimativas causais.
  - Aplicável em situações com muitos dados, muitas informações e pouca estrutura para guiar análises

Muito ainda para estudar...

- Métodos de deep learning
- Aplicações em vários métodos de avaliação de políticas públicas, desde de RCT até controle sintético, RDD etc
  - Uma boa revisão desta literatura: The Impact of Machine Learning on Economics (Athey (2019))
- Apesar de ML não ajudar diretamente na identificação causal, pode contribuir indiretamente para a análise das estimativas causais.
  - Aplicável em situações com muitos dados, muitas informações e pouca estrutura para guiar análises

- Prof. Rafael B. Barbosa
- Email: rafael.barbosa@ufc.br
- educLAB: https://educlab.com.br/

- Prof. Rafael B. Barbosa
- Email: rafael.barbosa@ufc.br
- educLAB: https://educlab.com.br/

- Prof. Rafael B. Barbosa
- Email: rafael.barbosa@ufc.br
- educLAB: https://educlab.com.br/

- Prof. Rafael B. Barbosa
- Email: rafael.barbosa@ufc.br
- educLAB: https://educlab.com.br/

- Prof. Rafael B. Barbosa
- Email: rafael.barbosa@ufc.br
- educLAB: https://educlab.com.br/

OBRIGADO!