Regressão Linear!

Aula Introdutória

Cícero Azevedo
Evandro Vianna
Gabriel Orlando
Geovanne Mansano
Lucca Barberato

Tópicos

01
Aprendizado Supervisionado

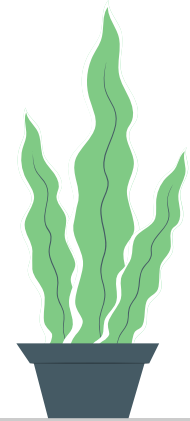
02
Regressão Linear

03
Classificação vs Regressão

04
Predição vs Inferência

05
Análise de resultados

06
Problemas com a regressão



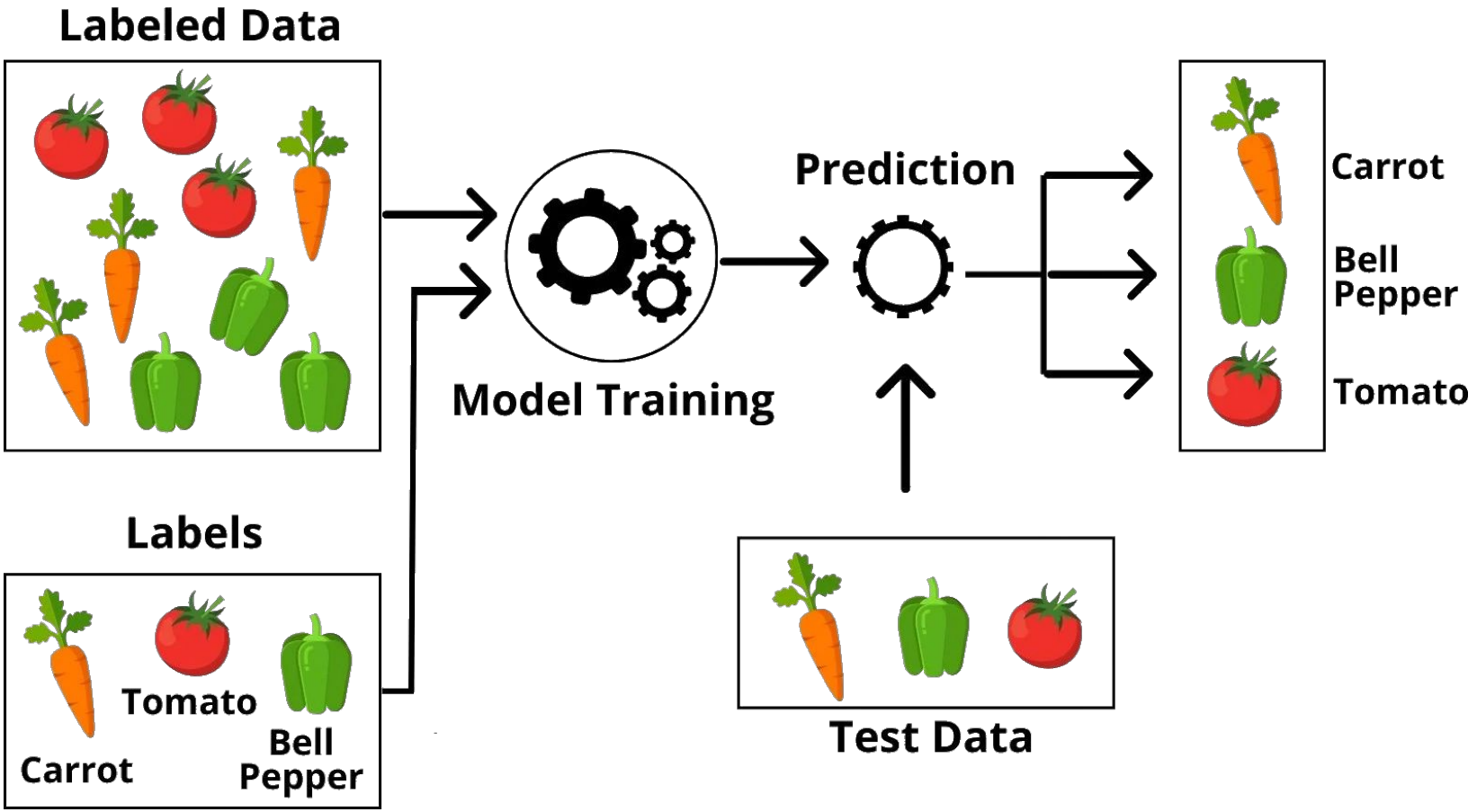
01

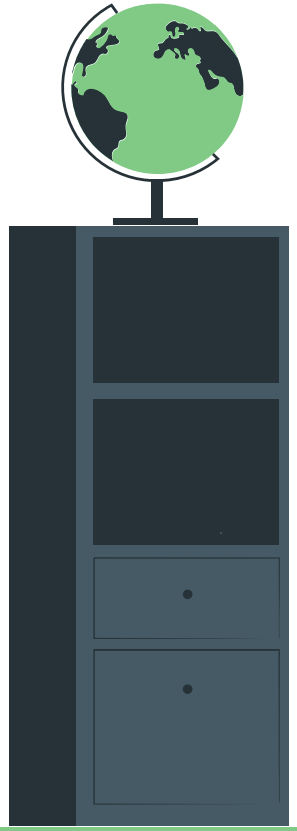
Aprendizado Supervisionado

Breve descrição sobre este tipo de tarefa

Aprendizado Supervisionado

- Dados para treinamento são rotulados, ou seja, incluem a saída correta
- Modelos tentam prever a saída (nesse caso valores numéricos continuos) e se adaptam, visando minimizar o erro
- Busca criar um modelo que consiga aprender a relação entre as entradas e a saída
- Ao final o modelo deve ser capaz de receber dados não vistos e prever a saída correta


Labeled Data
Labels
Carrot
Tomato
Bell Pepper
Model Training
Prediction
Test Data
Carrot
Bell Pepper
Tomato



02

Regressão Linear

O que é Regressão Linear? Quais os tipos? Fórmulas?

$$Y = \beta_0 + \beta_1 X + e$$

Variável reposta | Intecepto | Coeficiente angular | Variável explicativa | Erro

Tipos de Regressão Linear

Simples

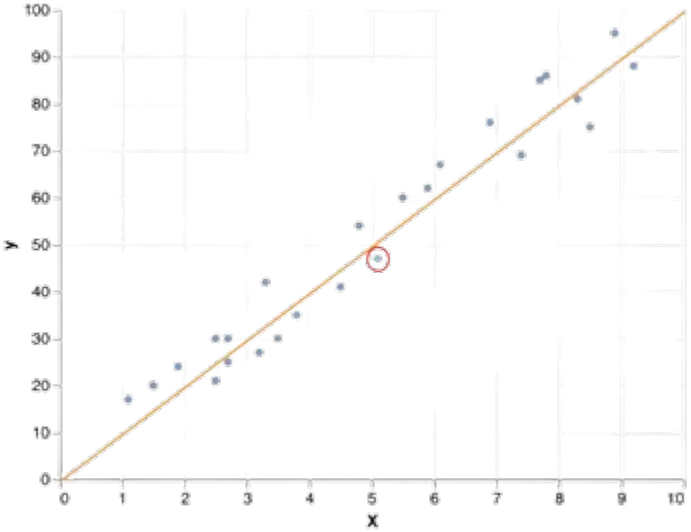
- Apenas 1 variável dependente e 1 variável independente
- Modela a relação entre as 2 variáveis com uma reta
- Busca encontrar os parâmetros a e b da equação da reta y = ax + b

Múltipla

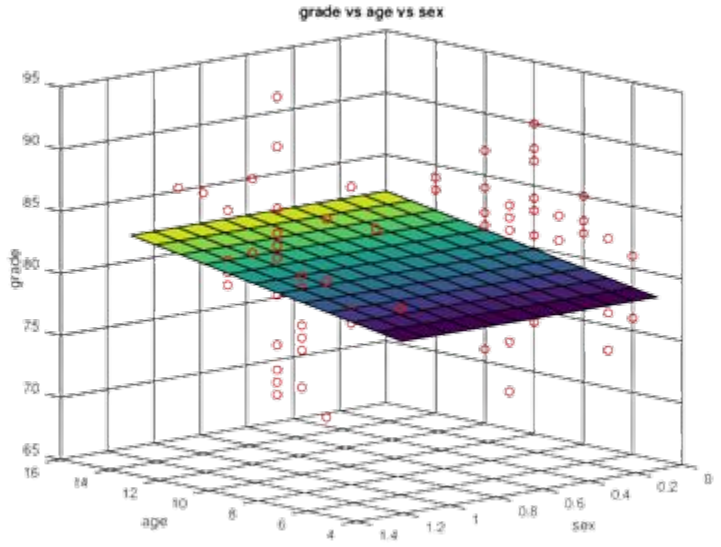
- Mais de uma variável dependente e uma independente
- Modela a relação entre as variáveis através de planos/hiperplanos
- Busca encontrar os parâmetros a, b, c, ... da equação do hiperplano (depende da dimensão)

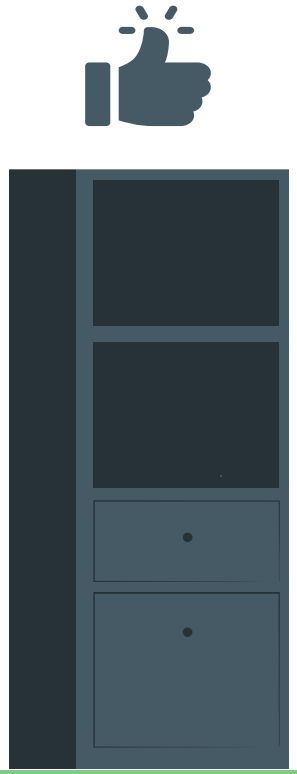
Tipos de Regressão Linear

SIMPLES



MÚLTIPLA





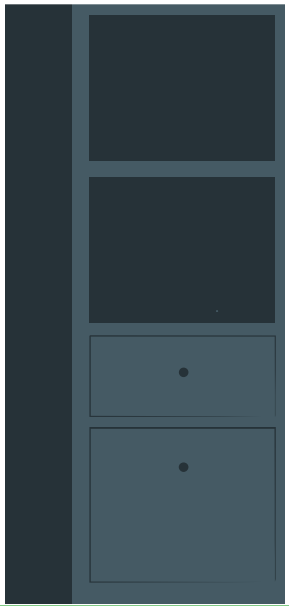
03

Classificação vs Regressão

Diferenças entre a tarefa de classificação e regressão

Regressão	Classificação
Valor numérico continuo	Prever a categoria de uma observação dada
Encontrar uma relação entre variáveis de entrada e saída	Estimar um “classificador” que gera como saída a classificação qualitativa de um dado
Prever o preço de uma casa	Classificar e-mails como spam ou não spam

04



Predição vs Inferência

Diferença entre as principais abordagens que podem ser realizadas com a regressão

Predição	Inferência
Usar um modelo treinado para estimar ou prever valores futuros com base em dados passados	Proposições sobre um universo, a partir de dados de uma amostra
É feita com base em identificação de padrões, que definem correlação entre fatos	Processo de extrair informações, insights ou conhecimentos para atacar um problema
O modelo preditivo ajuda em uma tomada de decisão mais eficiente	Pode entender como as variáveis estão relacionadas

Exemplo

Suponha que temos um conjunto de dados contendo informações sobre o número de horas de estudo semanal de alunos e suas respectivas notas finais em uma disciplina.

Horas de Estudo (x)	**Nota Final (y)**
2	62
3	65
4	68
5	70
6	73
7	76

Exemplo de Inferência

Nesse caso, queremos usar a regressão linear para inferir a relação *entre as horas de estudo e o desempenho acadêmico dos alunos.*

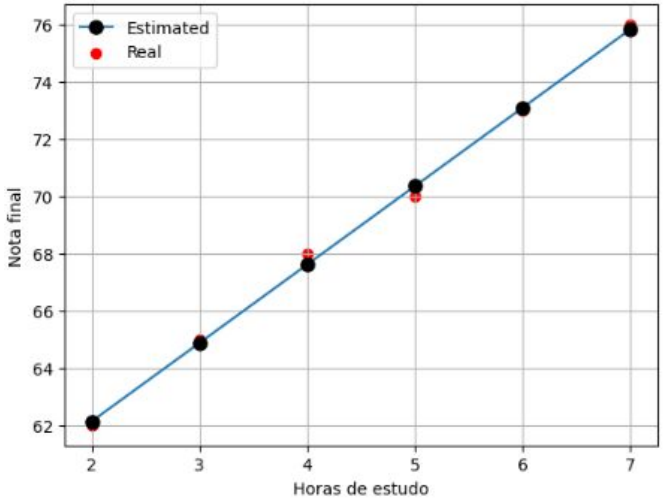
Após aplicar a regressão linear, encontramos a equação da reta: y = 2.74x + 56.65

- x representa o número de horas de estudo
- y representa a nota final

Exemplo de Inferência

Nesse caso, queremos usar a regressão linear para inferir a relação *entre as horas de estudo e o desempenho acadêmico dos alunos.*

Após aplicar a regressão linear, encontramos a equação da reta: **y = 2.74x + 56.65**

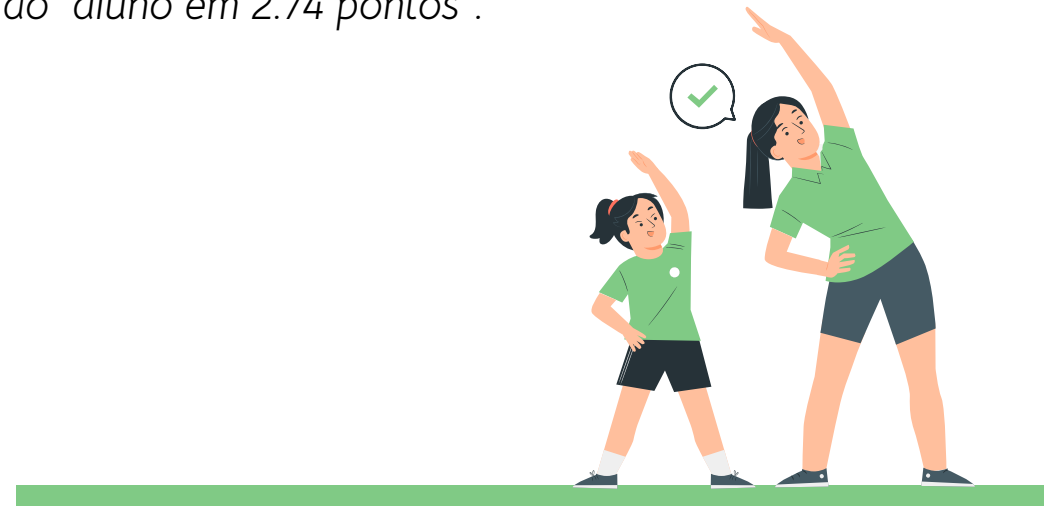


- x representa o número de horas de estudo
- y representa a nota final

Exemplo de Inferência

Assim, podemos fazer inferências como:

"Para a nossa população, espera-se que, em média, cada hora adicional de estudo aumente a nota final do aluno em 2.74 pontos".



Exemplo de Predição

Agora, consideremos um novo semestre em que temos as seguintes informações dos alunos:

- Suas respectivas horas de estudo.
- Mas não temos as notas finais.

Aqui, o objetivo é usar a regressão linear para fazer predições e estimar as notas finais com base nas horas de estudo.

Exemplo de Predição

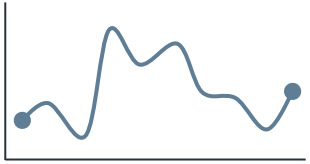
Suponha que, com base nos dados anteriores, obtivemos a mesma equação da reta:

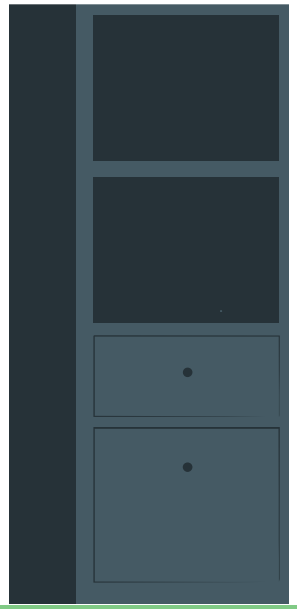
- $y = 2.74x + 56.65$.

Agora, se tivermos um aluno que estudou 5 horas por semana, podemos usar a regressão linear para prever sua nota final:

- Substituindo x por 5 na equação, obtemos **$y = (2.74 * 5) + 56.65$**

O que nos dá uma estimativa de que esse aluno pode obter uma nota final de 70.35.





05

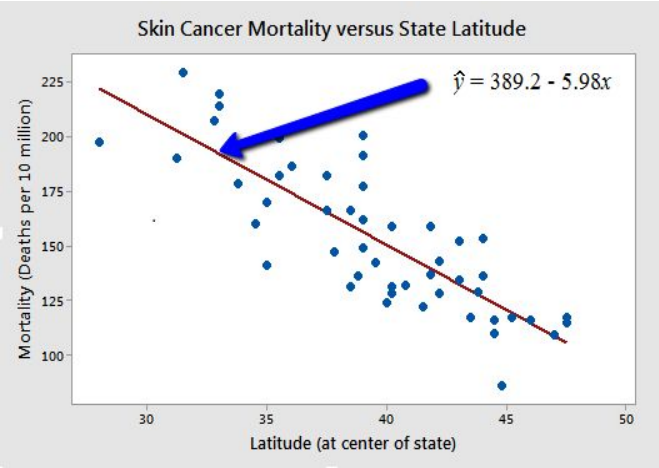
Análise dos resultados

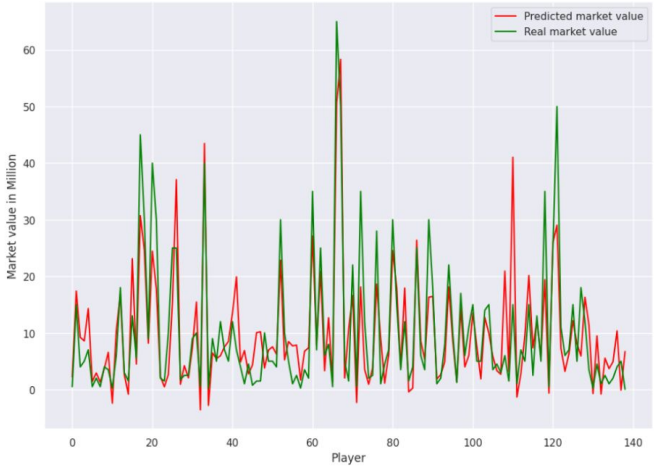
Principais métodos e estratégias para analisar os resultados da regressão linear

Análise dos resultados

- Técnicas diferentes devem ser empregadas para a Inferência e Predição

- **Inferência**: Análise de Coeficientes, Testes de hipóteses, Multicolinearidade, etc...
- **Predição**: Cross-Validation, R-2 Score, LASSO, etc...





Teste de hipóteses

- Testar a validade de uma hipótese

- Na regressão: Testar se nossos coeficientes têm influência na variável resposta
- Utilizamos o Teste-T, Teste-Z ou Teste-F

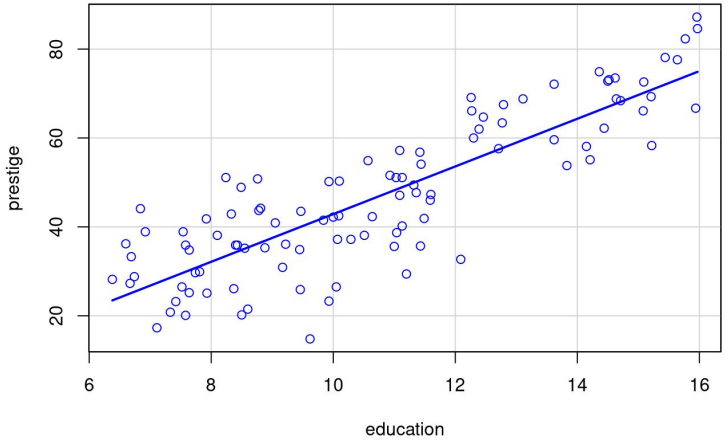
$$IGC_i = \beta_0 + \beta_1 fastfood_i + u_i$$

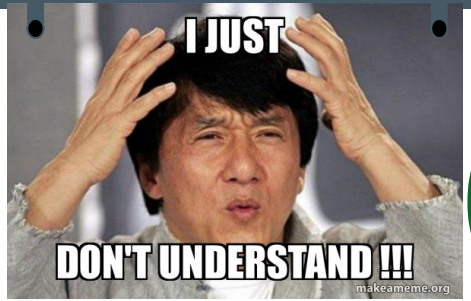
$$\begin{cases} H_0: \beta_1 = 0 \longrightarrow \text{o consumo de } \textit{fastfood} \text{ não é relevante para explicar o IGC de um indivíduo} \\ H_A: \beta_1 \neq 0 \longrightarrow \text{o consumo de } \textit{fastfood} \text{ é relevante para explicar o IGC de um indivíduo} \end{cases}$$

Verificação do Erro

- Não utilizamos a acurácia/precisão/recall
- Predição de valores contínuos
- Utiliza-se o erro do nosso modelo

- **Métricas**: RMSE, MAE, R2, MRE, etc...





06

Problemas da Regressão

Como lidar e identificar os principais problemas que envolvem regressão linear para as diferentes tarefas

Problemas da Regressão

Alguns dos principais problemas que serão retratados são:

- **Multicolinearidade:** alta correlação existente entre duas ou mais variáveis independentes.
- **Causalidade e correlação:** falsa relação entre variáveis.
- **Overfitting (sobreajuste):** treinou até demais.

Outros problemas possíveis mas que não serão tratados nessa aula:

- **Underfitting (sub ajuste):** pouca informação no dataset
- **Dados desbalanceados:** modelo pode ser tendencioso em direção à classe dominante
- **Sobreajuste de hiperparâmetros:** são atributos que controlam o treinamento do modelo de machine learning

Multicolinearidade

É uma situação em que duas ou mais variáveis independentes estão altamente correlacionadas. Podendo causar vários problemas em uma análise de regressão, como:

- Coeficientes de regressão não confiáveis
- Erros padrão inflados (erros padrão heteroscedásticos)
  - a variância dos erros não é a mesma para todos as variáveis independentes
- Testes de hipóteses não confiáveis

Multicolinearidade

Se você não tiver certeza se há multicolinearidade em seus dados, existem algumas maneiras de verificar:

- Inflação da variância (VIF), sendo uma medida da multicolinearidade de cada variável independente.

- Gráfico de dispersão das variáveis independentes. Se você ver muitos pontos agrupados juntos no gráfico, isso pode ser um sinal de multicolinearidade.

Tratamento da Multicolinearidade

Existem algumas maneiras de lidar com a multicolinearidade:

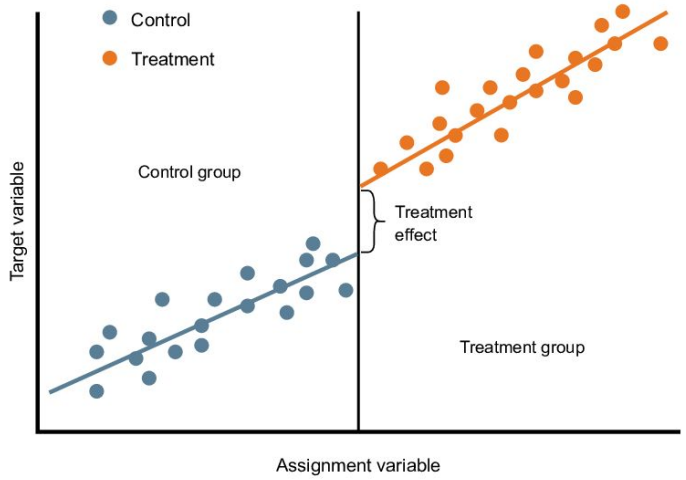
- Uma maneira é remover uma das variáveis independentes que estão altamente correlacionadas.

- Utilizar um método de regressão que seja menos sensível à multicolinearidade, como a regressão por mínimos quadrados generalizados.

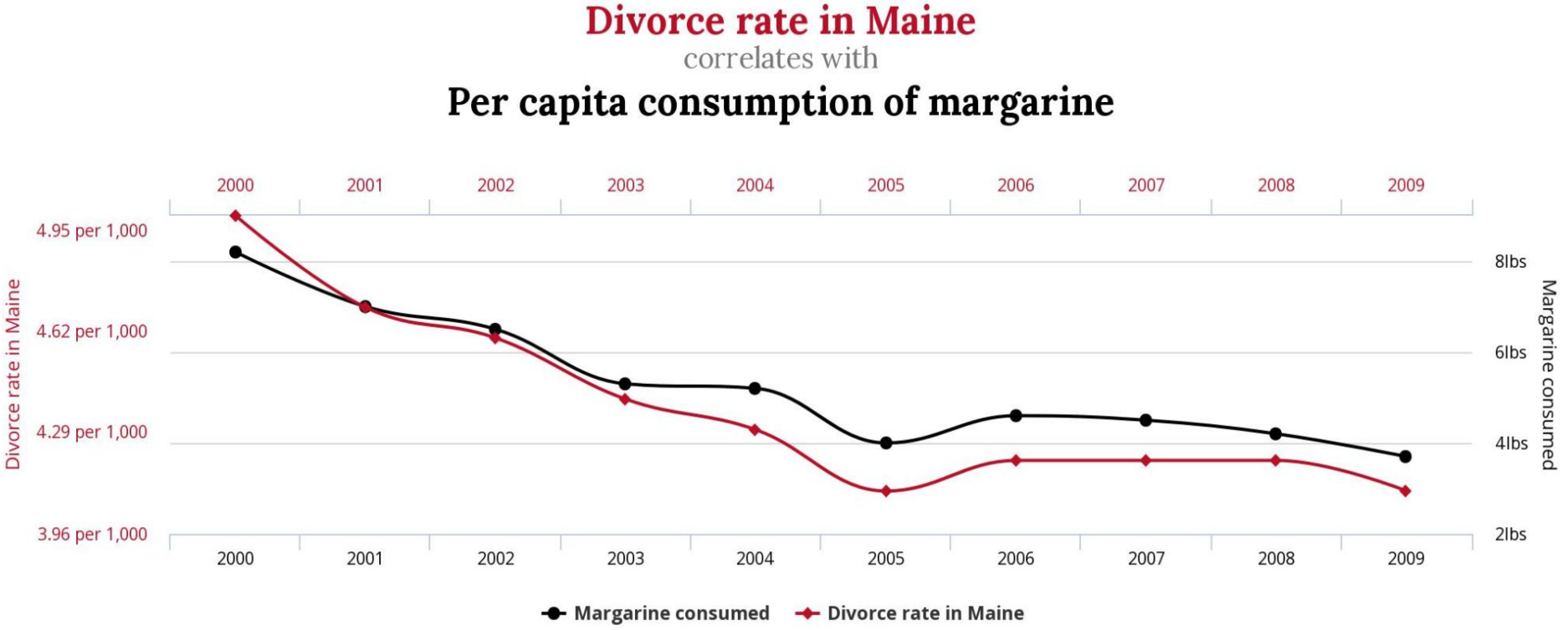
Causalidade

- Regressão linear por si só não captura causa do evento
- Devemos nos atentar em falsas causalidades que podem ser geradas
- Para verificar causalidade podemos usar o RDD

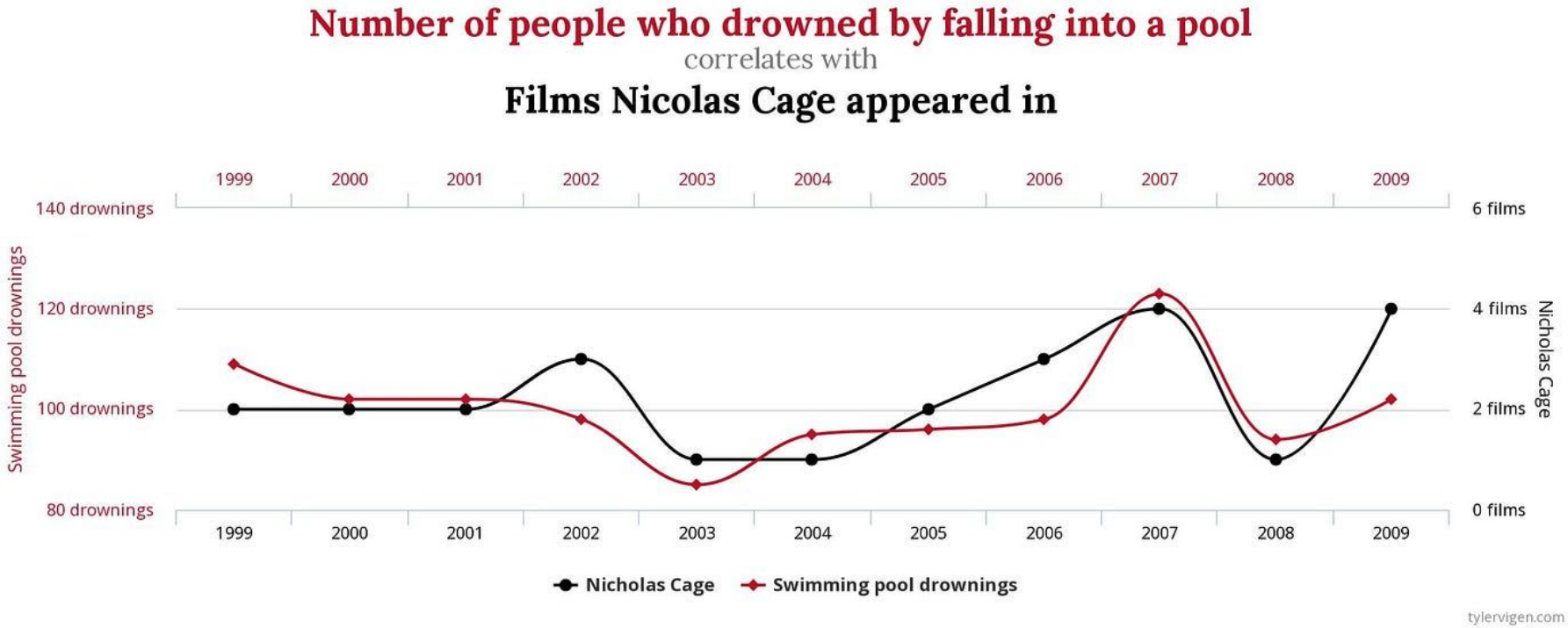
- Verifica o efeito causal de uma intervenção em uma variável de interesse
- Analisamos se de fato há a relação de causa



Exemplo


Divorce rate in Maine
correlates with
Per capita consumption of margarine
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
4.95 per 1,000
4.62 per 1,000
4.29 per 1,000
3.96 per 1,000
Divorce rate in Maine
8lbs
6lbs
4lbs
2lbs
Margarine consumed
Margarine consumed
Divorce rate in Maine
tylervigen.com

Exemplo

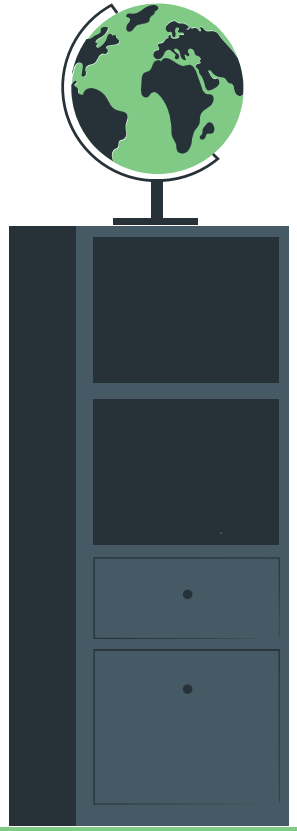

Number of people who drowned by falling into a pool
correlates with
Films Nicolas Cage appeared in
1999
2000
2001
2002
2003
2004
2005
2006
2007
2008
2009
140 drownings
120 drownings
100 drownings
80 drownings
Swimming pool drownings
6 films
4 films
2 films
0 films
Nicholas Cage
Nicholas Cage
Swimming pool drownings
tylervigen.com

Overfitting

- Modelo se ajusta muito bem aos dados de treino, mas não consegue generalizar, ou seja, não prevê bem dados nunca vistos

- Possíveis causas:
  - Modelo muito complexo
  - Poucos dados
  - Variáveis irrelevantes
  - Dados pouco representativos
  - Dados desbalanceados

- Como evitar?
  - Avaliar desempenho com validação-cruzada
  - Oversampling/Undersampling
  - Normalização e padronização
  - Seleção de atributos (ex: LASSO)



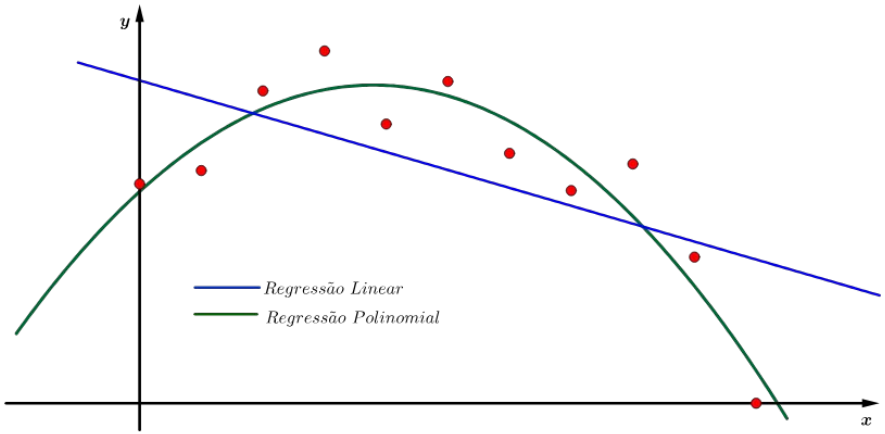
Tópicos adicionais

Estes tópicos não serão abordados durante esta oferta, mas podem ser estudados futuramente

Tópicos adicionais

Para as próximas ofertas podemos pensar nos seguintes tópicos:

- Outros tipos de regressão (Ex: Logarítimica, Polinomial, etc...)
- Combinação de modelos



Vamos para a apresentação prática!