

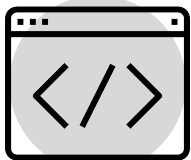
# Profissão: Cientista de Dados



# BOAS PRÁTICAS



# Regressão II



- **Diferencie previsão de explicação**
- **Faça inferência sobre os parâmetros**
- **Domine variáveis qualitativas**
- **Diferencie qualidade de complexidade**
- **Selecione modelos**
- **Regularização**



# Diferencie previsão de explicação

- Na previsão, o foco deve ser na redução de erro. No entanto, na explicação, além de explicar a variabilidade, também se busca medir o erro de previsão, fazer um intervalo de confiança ou um teste de hipóteses.
- A regressão é um excelente algoritmo de previsão, mas existem algoritmos melhores derivados da escola de Machine Learning, como árvore de decisão.



# Faça inferência sobre os parâmetros



- Entenda que os parâmetros de regressão são funções dos dados e, portanto, também são variáveis aleatórias: Se conhecermos a lei geradora do parâmetro, podemos fazer inferências sobre ele.
- Conheça as condições sob as quais a distribuição do parâmetro de regressão é conhecida: os erros devem ter uma distribuição normal, devem ser independentes e a variância deve ser homogênea.
- Ao realizar testes de hipóteses sobre o parâmetro de regressão, lembre-se de que a hipótese nula é que o parâmetro é igual a zero, o que significa que a variável não é relevante para o modelo.
- Saiba como calcular a estatística T e o valor p, e como interpretar esses resultados: Se o valor p for pequeno, rejeitamos a hipótese nula, enquanto se for grande, não rejeitamos.

# Domine variáveis qualitativas

- Dependendo do contexto, uma variável quantitativa e discreta pode ser tratada como contínua ou qualitativa no modelo. Entenda a natureza da variável antes de decidir como tratá-la.
- Use funções para criar matrizes de dados: A matriz X deve ser composta por variáveis independentes, enquanto a matriz Y deve ser a variável resposta. Isso ajuda a organizar seus dados e facilita a análise.



# Diferencie qualidade de complexidade

- Pondere a complexidade do modelo com sua precisão: Quando há muitas variáveis, é importante ponderar a complexidade do modelo com sua precisão. Um modelo mais complexo não é necessariamente melhor.
- Adicionar variáveis adicionais pode complicar o modelo e aumentar o risco de overfitting. Portanto, é importante ser criterioso ao adicionar variáveis ao modelo.



# Selecione modelos

- Não se baseie apenas no valor  $p$  para determinar a relevância de uma variável. Considere também outros indicadores de qualidade do modelo, como o critério de Akaike (AIC).
- Ao implementar o algoritmo stepwise, execute-o até que nenhuma alteração seja feita no modelo. Isso indica que o modelo está finalizado.





# Regularização

- Utilize a regularização para reduzir a complexidade do modelo. Isso pode ajudar a evitar o overfitting e melhorar a generalização do modelo para dados não vistos.
- Lembre-se de que a construção de um modelo é uma mistura de arte e ciência. Há muitas decisões a serem tomadas, como a escolha das variáveis, a parametrização e possíveis ajustes no modelo.



# Bons estudos!

