PLN para a Ciência Política e Políticas Públicas Públicas

Professora: Lorena Barberia

Semana 6

Universidade de São Paulo

Tópicos da Aula

1. Inferência
2. Aprendizado Estatístico com Regressão Linear
3. Predição

Universidade de São Paulo

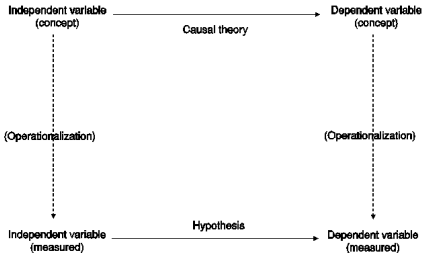
Inferência

- Nosso objetivo é selecionar uma amostra para investigar a relação entre as variáveis explicativas (X) e as variáveis dependentes (Y).
- Dado certos pressupostos, pensamos que nossas estimativas nos permitem afirmar o que esperamos encontrar na população (e o contrafactual).
- Usamos dados observacionais para estimar um modelo (paramétrico ou não-paramétrico) e comparamos o valor observado versus o valor predito, ou seja o ajuste dos dados ao modelo.

Universidade de São Paulo

Inferência

From theory to hypothesis



Universidade de São Paulo

Inferência

- Podemos pensar em 2 diferentes tipos de planos de amostragem:
  - Os planos de **amostragem probabilística** atribuem a cada membro da população uma probabilidade conhecida diferente de zero de inclusão na amostra. Seleção aleatória
  - Os planos de **amostragem dependentes no modelo** assumem que as variáveis de interesse na pesquisa seguem uma distribuição de probabilidade conhecida e que a escolha dos elementos da amostra maximiza a precisão da estimativa para as estatísticas de interesse.

Universidade de São Paulo

Inferência

- 2 diferentes tipos de métodos de inferência:
  - Modelos **não paramétricos**, e.g. $y \sim N(xB + \epsilon, \sigma^2)$ não há suposições relativas à natureza da população da qual uma amostra é extraída ou métodos baseados na distribuição de probabilidade para a seleção da amostra.
  - Modelos **paramétricos**, suposições relativas à distribuição específica da população e parâmetros ou métodos que consideram a distribuição de probabilidade para a variável aleatória de interesse.

Universidade de São Paulo

Métodos Paramétricos: Vantagens

- **Simplicidade**: Os métodos paramétricos são geralmente mais simples de entender e implementar em comparação com métodos não paramétricos.
- **Eficiência Computacional**: Eles geralmente requerem menos recursos computacionais.
- **Interpretabilidade**: Modelos paramétricos muitas vezes produzem resultados mais interpretais, permitindo entender a relação entre as variáveis.
- **"Econômicos"**: É mais adequando quando só temos um pequeno número de observações.

Universidade de São Paulo

Métodos Paramétricos - Desvantagens

- **Restrições de Modelagem**: Eles podem não ser capazes de capturar relações complexas nos dados, tornando-os menos adequados para problemas não lineares.
- **Vieses de Modelo**: Os modelos paramétricos podem introduzir viés se a forma funcional escolhida não corresponder à verdadeira relação entre as variáveis.
- **Limitações na Flexibilidade**: São menos flexíveis para acomodar variações nos dados, o que pode resultar em um desempenho inferior em alguns casos.

Universidade de São Paulo

Métodos Não Paramétricos

- Eles não fazem suposições específicas sobre a forma funcional da distribuição subjacente dos dados.
- Em vez disso, esses métodos são mais flexíveis e se adaptam melhor a relações complexas entre variáveis.
- Os métodos não paramétricos são ideais quando não temos informações prévias claras sobre a estrutura dos dados ou quando as relações são altamente não lineares.

Universidade de São Paulo

Métodos Não Paramétricos - Vantagens

- **Flexibilidade** - Ideais para capturar relações complexas e não lineares entre variáveis.
- **Ajuste Simples** - Adequados para uma gama diversificada de problemas.
- **Menos Viés** - Métodos não paramétricos tendem a ter menos viés e podem se adaptar melhor a uma variedade de situações do mundo real.

Universidade de São Paulo

Métodos Não Paramétricos - Desvantagens

- **Maior Demanda Computacional**;
- **Interpretação Complexa** - A flexibilidade dos métodos não paramétricos pode tornar os modelos difíceis de interpretar;
- **Maior Risco de Overfitting** - Como esses métodos podem se ajustar demais aos dados, há um risco maior de overfitting, especialmente quando o conjunto de dados é pequeno.

Universidade de São Paulo

Aprendizado Estatístico com Regressão Linear

$$Y = f(X) + \varepsilon$$

Universidade de São Paulo

Regressão Linear

- Reescrevendo a função utilizando um modelo de regressão linear, teríamos:

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \varepsilon$$

Universidade de São Paulo

Conceitos básicos: O contrafactual

$$Y = \beta_0 + \beta_1 * X_1 + \beta_2 * X_2 + \varepsilon$$

Universidade de São Paulo

Conceitos básicos: Mean Square Error (MSE)

1. Se g é a estimativa do parâmetro γ:
2. $MSE = V(g) + E(g - \gamma)^2 = Varianca + \text{Viés}^2$
3. Se comparamos modelos para estimar g, preferimos o modelo que produz estimativas com menor viés e menor variância, porém muitas vezes precisamos comparar podemos preferir um estimador com viés por ele ter menor variação.
4. RMSE é a raiz quadrada de MSE, ou seja $\sqrt{MSE}$.

Universidade de São Paulo

Conceitos básicos: Mean Square Error (MSE)

$$
\begin{aligned}
E(Y - \hat{Y})^2 &= E[f(X) + \varepsilon - \hat{f}(X)]^2 \\
&= [f(X) - \hat{f}(X)]^2 + Var(\varepsilon)
\end{aligned}
\tag{1}
$$

Universidade de São Paulo

Conceitos básicos: Testes de hipóteses e Erro Tipo I e II

1. Erro Tipo 1 - Rejeitar a hipótese quando é verdadeira.
2. Erro Tipo 2 - Não rejeitar a hipótese quando ela é falsa.

	Valor Verdeiro	
Amostra	Verdeiro	Falso
Hipótese Rejeitada	Erro Tipo 1 Falso positivo α = probabilidade de rejeitar quando verdadeiro	Decisão correta (Verdadeiro Negativo) $1 - \beta$
Hipótese Rejeitada	Decisão correta (Verdadeiro Falso) 1-α = *probabilidadede*	Erro Tipo 2 Falso negativo β

Universidade de São Paulo

Corpus

A construção de um corpus envolve uma estrategia de amostragem e de desenho de pesquisa.

O Problema em Aprendizado com Supervisão

- Y = classificação (e.g. posicionamento, sentimento, etc.)
- X = vetor de características preditoras
- *Training Data Set* = $(x_1, y_1, ...x_N, y_N)$
- Coleta e Avaliação de outras Amostras para avaliar se o *training* data set corretamente classifica outras amostras
- Statistical Framework: $y = f(x) + \epsilon$

Universidade de São Paulo

Dúvidas?

Universidade de São Paulo