Inferência Estatística (GCC1625)
Introdução ao Pareamento por Escore de Propensão

Prof. Eduardo Bezerra

2025.1

1. Introdução

Um **estudo observacional** é um tipo de investigação em que o pesquisador **não interfere na atribuição dos tratamentos ou exposições**. Em vez disso, ele apenas observa e coleta dados sobre o que ocorre naturalmente. Diferente de **experimentos controlados** (como ensaios clínicos randomizados), nos estudos observacionais não há randomização, o que pode introduzir *viés de seleção*.

Viés de seleção ocorre quando os grupos comparados em um estudo diferem sistematicamente em características que também afetam o desfecho, comprometendo a validade da inferência causal. Em outras palavras: é um erro que surge porque *quem recebe o tratamento é diferente de quem não recebe*, por razões que influenciam os resultados.

Estudos observacionais são comuns em contextos onde a manipulação direta seria antiética, impraticável ou muito cara (como em epidemiologia, economia ou ciências sociais). Os exemplos a seguir ilustram como o pesquisador trabalha com dados do "mundo real", assumindo o desafio de controlar vieses com métodos estatísticos.

- Epidemiologia
  - **Objetivo**: Avaliar se fumantes têm maior risco de desenvolver câncer de pulmão.
  - **Observação**: Os pesquisadores acompanham grupos de fumantes e não fumantes ao longo do tempo, sem interferir no hábito de fumar.
- Economia
  - **Objetivo**: Estimar o efeito da conclusão do ensino superior sobre a renda.
  - **Observação**: Compara-se a renda de indivíduos com e sem diploma, sem que o pesquisador escolha quem estuda ou não.
- Ciências Sociais
  - **Objetivo**: Investigar se filhos de pais divorciados apresentam maior taxa de evasão escolar.
  - **Observação**: Coletam-se dados de famílias com diferentes estruturas, sem intervenção nos arranjos familiares.

Em estudos observacionais, o objetivo muitas vezes é estimar o efeito causal de uma intervenção (ou tratamento) sobre um desfecho de interesse. No entanto, ao contrário de experimentos randomizados, o pesquisador não controla a alocação do tratamento. Isso gera o problema do viés de seleção, pois os grupos tratado e controle podem diferir sistematicamente em variáveis que também influenciam o desfecho.

A técnica de **pareamento por escore de propensão** (*propensity score matching*, PSM) é usada para estimar efeitos causais em estudos observacionais, e seu principal objetivo é justamente mitigar o viés de seleção decorrente de diferenças nas covariáveis observadas entre os grupos tratado e controle.

Ao emparelhar indivíduos com probabilidades similares de receber o tratamento, o PSM busca criar grupos comparáveis, como se o tratamento tivesse sido alocado aleatoriamente dentro desses pares ou estratos. Assim, ele aproxima o cenário observacional de

um experimento controlado, sob a suposição de ausência de confundidores não observados.[1]

O pareamento por escore de propensão é amplamente utilizado em Economia, Medicina, Ciências Sociais e Avaliação de Políticas Públicas. Trata-se de uma ferramenta essencial para a análise causal em contextos onde a randomização não é viável.

O objetivo deste documento é explorar:

- O conceito de escore de propensão e suas hipóteses fundamentais;
- Como estimar escores de propensão via regressão logística;
- Técnicas de pareamento usando esses escores;
- Avaliação do balanceamento e estimativas do efeito causal;
- Um exemplo com dados sintéticos e código em Python.

Exemplo

Suponha que queremos avaliar o efeito de um curso de capacitação profissional sobre o salário mensal dos participantes. Se compararmos diretamente o salário médio de quem fez o curso com quem não fez, corremos o risco de capturar diferenças que não se devem ao curso, mas sim a características como idade, escolaridade ou experiência prévia.

Alternativas para contornar o viés de seleção

Para tentar ajustar essas diferenças, uma alternativa poderosa é o uso de técnicas de **pareamento**, que consistem em encontrar, para cada indivíduo tratado, um ou mais indivíduos não tratados com características similares. A ideia é formar pares comparáveis, como se a alocação tivesse sido aleatória dentro de subconjuntos homogêneos.

O escore de propensão

Introduzido por [Rosenbaum and Rubin, 1983], o **escore de propensão** é definido como a probabilidade de um indivíduo receber o tratamento condicionalmente às covariáveis observadas:

$$e(x) = \Pr(D = 1 \mid X = x)$$

Onde:

- $D \in \{0, 1\}$ é a variável indicadora de tratamento,
- X é o vetor de covariáveis observadas.

A vantagem do escore de propensão é que, sob certas condições, ele permite balancear os grupos tratado e controle em uma única dimensão escalar, mesmo quando há muitas covariáveis.

[1] Confundidores não observados são variáveis que afetam tanto o tratamento quanto o desfecho, mas que não foram medidas ou incluídas na análise. Por não serem controladas, essas variáveis podem distorcer a estimativa do efeito causal, mesmo após o uso de técnicas como o PSM.

2. Fundamentos Teóricos

O escore de propensão foi introduzido por [Rosenbaum and Rubin, 1983] como uma ferramenta para reduzir o viés de seleção em estudos observacionais. A ideia central é que, sob certas condições, o pareamento com base no escore de propensão permite estimar efeitos causais mesmo na ausência de um experimento aleatorizado.

Definição

Dado um vetor de covariáveis X, o escore de propensão é definido como:

$$e(x) = \mathbb{P}(D = 1 \mid X = x)$$

onde $D \in \{0, 1\}$ é a variável indicadora de tratamento. Em outras palavras, $e(x)$ é a probabilidade condicional de um indivíduo receber o tratamento, dado seu perfil de covariáveis.

Motivação

Em estudos observacionais, os grupos tratado e controle geralmente diferem em várias covariáveis. Ao invés de ajustar diretamente para todas elas, o escore de propensão oferece uma forma de colapsar essas informações em uma única métrica escalar, mantendo o potencial de controle do viés de confusão.

Hipóteses fundamentais

Para que o pareamento por escore de propensão seja válido para inferência causal, duas hipóteses principais devem ser satisfeitas:

1. **Ignorabilidade condicional (ou unconfoundedness)**:

   $$(Y(0), Y(1)) \perp D \mid X$$

   Essa hipótese afirma que, dado X, a alocação do tratamento é independente dos resultados potenciais.

2. **Suporte comum (ou overlap)**:

   $$0 < \Pr(D = 1 \mid X) < 1$$

   Isso garante que, para qualquer combinação de covariáveis, há indivíduos tanto no grupo tratado quanto no grupo controle.

Teorema de Rosenbaum & Rubin (1983)

Se as duas hipóteses acima forem verdadeiras, então:

$$(Y(0), Y(1)) \perp D \mid e(X)$$

Ou seja, condicionar sobre o escore de propensão é suficiente para tornar o tratamento independente dos resultados potenciais. Assim, podemos emparelhar indivíduos com valores similares de $e(X)$, como se o tratamento tivesse sido randomizado dentro desses estratos.

Vantagens do escore de propensão

- Reduz a dimensionalidade do problema (de p covariáveis para 1 escalar);
- Permite diversas estratégias: pareamento, estratificação, ponderação (IPW), ou covariável na regressão;
- Facilita a visualização do balanceamento entre grupos.

Métodos de estimativa

Na prática, o escore de propensão é estimado via modelos preditivos, usualmente por meio de um modelo de regressão logística:

$$\log\left(\frac{e(x)}{1-e(x)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

Modelos mais flexíveis como árvores de decisão, random forests ou redes neurais também podem ser utilizados, mas exigem cuidados adicionais com overfitting e interpretabilidade.

3. Exemplo em Python

O exemplo apresentado nesta seção ilustra como implementar o pareamento por escore de propensão em Python, estimando o efeito do tratamento e avaliando o balanceamento. A metodologia pode ser ampliada para conjuntos de dados maiores, com validação do pareamento e análise de sensibilidade.

Considere uma amostra hipotética de 6 indivíduos. O objetivo é avaliar o efeito de um curso de capacitação (variável `Tratamento`) sobre o salário mensal (`Salário`). As covariáveis observadas são `Idade` e `Escolaridade`.

Tabela de dados

ID	Tratamento (D)	Idade	Escolaridade	Salário (R$)
1	1	25	16	3200
2	1	30	18	3500
3	1	28	14	3000
4	0	27	14	2800
5	0	35	18	3400
6	0	24	16	3100

Etapa 1 – Estimar escore de propensão

Assuma que utilizamos regressão logística com covariáveis `Idade` e `Escolaridade`, resultando nos seguintes escores de propensão estimados:

ID	Escore de propensão
1	0.65
2	0.80
3	0.50
4	0.48
5	0.75
6	0.60

Etapa 2 – Pareamento 1:1 com vizinho mais próximo

Emparelhe cada tratado com um controle cujo escore de propensão seja o mais próximo:

Tratado	Controle Pareado	Diferença de Escore
1 (0.65)	6 (0.60)	0.05
2 (0.80)	5 (0.75)	0.05
3 (0.50)	4 (0.48)	0.02

Etapa 3 – Comparar desfechos pareados

Par	Salário Tratado	Salário Controle
1 vs 6	3200	3100
2 vs 5	3500	3400
3 vs 4	3000	2800

Diferença média:

$$\frac{(3200-3100)+(3500-3400)+(3000-2800)}{3} = \frac{100+100+200}{3} = 133{,}33$$

Conclusão do exemplo

A diferença média dos salários após o pareamento é de R$ 133,33. Esse valor pode ser interpretado como uma estimativa do **efeito médio do tratamento sobre os tratados (ATT)** sob as hipóteses do modelo de escore de propensão.

Este exemplo destaca como o pareamento contribui para a construção de grupos comparáveis, minimizando o viés de seleção com base nas covariáveis observadas.

4. Implementação em Python

Nesta seção, implementamos o exemplo numérico da Seção 3 utilizando a linguagem Python. Utilizamos a biblioteca `scikit-learn` para ajustar o modelo de escore de propensão e a técnica de `vizinho mais próximo` para realizar o pareamento.

1. Preparação dos dados

Listing 1: Criação do conjunto de dados

```
import pandas as pd

data = pd.DataFrame({
    'D': [1, 1, 1, 0, 0, 0],
    'Idade': [25, 30, 28, 27, 35, 24],
    'Escolaridade': [16, 18, 14, 14, 18, 16],
    'Salario': [3200, 3500, 3000, 2800, 3400, 3100]
})
```

2. Estimação do escore de propensão

Listing 2: Ajuste de regressão logística

```
from sklearn.linear_model import LogisticRegression

X = data[['Idade', 'Escolaridade']]
y = data['D']

model = LogisticRegression()
model.fit(X, y)

# Adiciona coluna com o escore de propensão estimado
data['propensity_score'] = model.predict_proba(X)[:, 1]
```

Nota: Utilizamos a regressão logística para estimar a probabilidade de receber tratamento, condicionalmente às covariáveis.

3. Separação entre tratados e controles

Listing 3: Separação dos grupos

```
treated = data[data['D'] == 1]
control = data[data['D'] == 0]
```

4. Pareamento 1:1 com vizinho mais próximo

Listing 4: Pareamento baseado no escore

```
from sklearn.neighbors import NearestNeighbors

nn = NearestNeighbors(n_neighbors=1)
nn.fit(control[['propensity_score']])

# Identifica os índices dos controles mais próximos
distances, indices = nn.kneighbors(treated[['propensity_score']])
```

```
matched_control = control.iloc[indices.flatten()].copy()
matched_control.index = treated.index  # para alinhamento
```

5. Comparação dos salários

Listing 5: Cálculo do efeito médio do tratamento

```
matched_data = pd.concat([treated, matched_control])
print("Média por grupo após pareamento:")
print(matched_data.groupby('D')['Salario'].mean())
```

Saída esperada:

```
Média por grupo após pareamento:
D
0    3100.0
1    3233.33
Name: Salario, dtype: float64
```

ATT estimado: R$ 133,33

6. Visualização dos escores

Listing 6: Gráfico de densidade dos escores

```
import seaborn as sns
import matplotlib.pyplot as plt

plt.figure(figsize=(8,4))
sns.kdeplot(data=data[data.D == 1]['propensity_score'], label='Tratado')
sns.kdeplot(data=data[data.D == 0]['propensity_score'], label='Controle')
plt.title("Distribuição dos Escores de Propensão")
plt.xlabel("Escore")
plt.legend()
plt.show()
```

Este gráfico permite avaliar se os grupos têm sobreposição de escores de propensão (condição de *overlap*).

References

[Rosenbaum and Rubin, 1983] Rosenbaum, P. R. and Rubin, D. B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika*, 70(1):41–55.