MI628 / ME920 - Inferência Causal

– Lista 1 –

Carlos Trucíos

Instruções

- A resolução da lista será discutida no dia 12/03 em sala de aula (a participação será avaliada). Os **alunos** apresentarão a solução e farão a discussão pertinente de cada um dos exercícios.
- Os exercícios computacionais deverão ser resolvidos com antecedencia de forma que seja possível ver o código, gráficos, tabelas e outros resultados obtidos (sugestão: Github / Colab / Posit Cloud).
- Os exercícios são para ambas as turmas, exceto quando o contrário seja explicitado.

Exercícios

1. Para o caso de tabelas de contigência, já foram apresentadas as definições de RD, RR e OR. Mostre que:

   a. $Z \perp\!\!\!\perp Y$, $RD = 0$, $RR = 1$ e $OR = 1$ são afirmações equivalentes.
   b. Se todos os p_{zy} são positivos, então $RD > 0$ é equivalente a $RR > 1$ e também a $OR > 1$.
   c. $OR \approx RR$ se $Pr(Y = 1|Z = 1)$ e $Pr(Y = 1|Z = 0)$ são pequenos.

2. Seja

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho YZ \\ \rho_{XZ} & \rho YZ & 1 \end{pmatrix}\right).$$

   a. [**MI628**] Mostre que $\rho_{YZ|X} = \dfrac{\rho_{YZ} - \rho_{XY}\rho_{ZX}}{\sqrt{1-\rho^2_{XY}} \times \sqrt{1-\rho^2_{ZX}}}$.
   b. De um exemplo numérico (simule um caso) em que $\rho_{YZ} > 0$ mas $\rho_{YZ|X} < 0$ [Paradoxo de Simpson para um vetor Normal trivariado].

1. Para o caso de tabelas de contigência, já foram apresentadas as definições de RD, RR e OR. Mostre que:

   a. $Z \perp\!\!\!\perp Y$, $RD = 0$, $RR = 1$ e $OR = 1$ são afirmações equivalentes.
   b. Se todos os p_{zy} são positivos, então $RD > 0$ é equivalente a $RR > 1$ e também a $OR > 1$.
   c. $OR \approx RR$ se $Pr(Y = 1|Z = 1)$ e $Pr(Y = 1|Z = 0)$ são pequenos.

1) a) $RD = 0 \Rightarrow p_{11}/(p_{11}+p_{10}) - p_{01}/(p_{01}+p_{00}) = 0 \Rightarrow p_{11}/(p_{11}+p_{10}) = p_{01}/(p_{01}+p_{00})$

$\Rightarrow \dfrac{p_{11}/(p_{11}+p_{10})}{p_{01}/(p_{01}+p_{00})} = 1 \Rightarrow RR = 1$

$RR = 1 \Rightarrow \dfrac{p_{11}/(p_{11}+p_{10})}{p_{01}/(p_{01}+p_{00})} = 1 \Rightarrow \dfrac{p_{11}+p_{10}}{p_{11}} = \dfrac{p_{01}+p_{00}}{p_{01}}$

$\Rightarrow \cancel{1} + \dfrac{p_{10}}{p_{11}} = \cancel{1} + \dfrac{p_{00}}{p_{01}} \Rightarrow \dfrac{p_{11}p_{00}}{p_{10}p_{01}} = 1 \Rightarrow OR = 1$

$Z \perp\!\!\!\perp Y \Rightarrow P(Z=z, Y=y) = P(Z=z)P(Y=y)$

$\Rightarrow p_{11} = (p_{11}+p_{10})(p_{11}+p_{01}) = p_{11}^2 + p_{11}(p_{10}+p_{01}) + p_{10}p_{01}$

$\Rightarrow p_{11} + p_{10} + p_{01} + \dfrac{p_{10}p_{01}}{p_{11}} = 1$

$\Rightarrow \dfrac{p_{10}p_{01}}{p_{11}} = 1 - (p_{10}+p_{01}+p_{11}) = p_{00}$

$\Rightarrow \dfrac{p_{11}p_{00}}{p_{10}p_{01}} = 1 \Rightarrow OR = 1$

1) b) $RD = \dfrac{p_{11}}{p_{11}+p_{10}} - \dfrac{p_{01}}{p_{01}+p_{00}} > 0$

$\Rightarrow \dfrac{p_{11}}{p_{11}+p_{10}} > \dfrac{p_{01}}{p_{01}+p_{00}}$

$\Rightarrow \dfrac{p_{11}}{p_{11}+p_{10}} \Big/ \dfrac{p_{01}}{p_{01}+p_{00}} > 1 \Rightarrow RR > 1$

$RR = \dfrac{p_{11}}{p_{11}+p_{10}} \Big/ \dfrac{p_{01}}{p_{01}+p_{00}} > 1$

$\Rightarrow \dfrac{p_{01}+p_{00}}{p_{01}} > \dfrac{p_{11}+p_{10}}{p_{11}}$

$\Rightarrow \cancel{1} + p_{00}/p_{01} > \cancel{1} + p_{10}/p_{11} \Rightarrow \dfrac{p_{11}p_{00}}{p_{10}p_{01}} > 0 \Rightarrow OR > 1$

1) c) $RR = \frac{P_{11}}{P_{11}+P_{10}} \Big/ \frac{P_{01}}{P_{01}+P_{00}} \approx \frac{P_{11}}{P_{10}} \Big/ \frac{P_{01}}{P_{00}} = \frac{P_{11}P_{00}}{P_{10}P_{01}} = OR$

2. Seja

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho_{XY} & \rho_{XZ} \\ \rho_{XY} & 1 & \rho YZ \\ \rho_{XZ} & \rho YZ & 1 \end{pmatrix}\right).$$

a. **[MI628]** Mostre que $\rho_{YZ|X} = \dfrac{\rho_{YZ} - \rho_{XY}\rho_{ZX}}{\sqrt{1-\rho_{XY}^2} \times \sqrt{1-\rho_{ZX}^2}}$.

b. De um exemplo numérico (simule um caso) em que $\rho_{YZ} > 0$ mas $\rho_{YZ|X} < 0$ [Paradoxo de Simpson para um vetor Normal trivariado].

2) a)

Dado $\begin{pmatrix} Y_1 \\ Y_2 \end{pmatrix} \sim N\left(\begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right)$

Temos $Y_1 | Y_2 = y_2 \sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(y_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}\right)$

Logo $Y,Z|X \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix} + \begin{pmatrix} \rho_{xy} \\ \rho_{xz} \end{pmatrix} \cdot 1^{-1} \cdot (x-0), \begin{pmatrix} 1 & \rho_{yz} \\ \rho_{yz} & 1 \end{pmatrix} - \begin{pmatrix} \rho_{xy} \\ \rho_{xz} \end{pmatrix} \cdot 1^{-1} \cdot \begin{pmatrix} \rho_{xy} & \rho_{xz} \end{pmatrix}\right)$

$\sim N\left(\begin{pmatrix} \rho_{xy} \\ \rho_{xz} \end{pmatrix} \cdot x, \begin{pmatrix} 1-\rho_{xy}^2 & \rho_{yz} - \rho_{xy}\rho_{xz} \\ \rho_{yz} - \rho_{xy}\rho_{xz} & 1-\rho_{xz}^2 \end{pmatrix}\right)$

Temos $\rho_{xy} = \dfrac{cov(X,Y)}{\sqrt{var(X)\,var(Y)}}$

Logo $\rho_{yz|x} = \dfrac{\rho_{yz} - \rho_{xy}\rho_{xz}}{\sqrt{(1-\rho_{xy}^2)(1-\rho_{xz}^2)}}$

2) b)

$\rho_{yz} = 0{,}68$

$\rho_{xy} = \rho_{xz} = 0{,}9$ $\Rightarrow$ $\rho_{yz|x} = -0{,}68$

3. O *dataset* `lalonde` do pacote `Matching` contém informação do tratamento (`treat`) [1: tratamento, 0: controle], do salario em 1978 (`re78`), bem como de outras 10 covariáveis (ou seja, existem $2^{10} = 1024$ possíveis subconjuntos de covariáveis). Ajuste as 1024 regressões com todos os possíveis subconjuntos de covariáveis e reporte o coeficiente associado ao tratamento.

   a. Quantas vezes o tratamento foi positivo e significativo?
   b. Quantas vezes o tratamento foi negativo e significativo?
   c. Quantas vezes o tratamento não foi significativo?

4. Assuma que os resultados potenciais são gerador por

$$Y(0) \sim N(0,1), \quad \tau = -0.5 + Y(0), \quad e Y(1) = Y(0) + \tau.$$

   O mecanismo de atribuição de tratamento é de tal forma que $Z = 1$ se $\tau \geq 0$ e $Z = 0$ se $\tau < 0$. O resultado observado é dado por $Y = ZY(1) + (1 - Z)Y(0)$.

   a. Calcule $\mathbb{E}(Y|Z = 1) - \mathbb{E}(Y|Z = 0)$.
   b. Simule um caso (apenas uma replicação) para verificar o resultado obtido no item anterior.
   c. **Dica:** se $X \sim N(\mu, \sigma)$, $\mathbb{E}(X|a < X < b) = \mu - \sigma \dfrac{\phi(\frac{b-\mu}{\sigma}) - \phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$.

5. De um exemplo numérico (simule um caso) em que $\tau = n^{-1} \sum_{i=1}^{n} (Y_i(1) - Y_i(0)) > 0$ mas a proporção de vezes que $Y_i(1) > Y_i(0)$ é menor do que 0.5. Ou seja, o efeito causal médio é positivo mas o tratamento beneficia menos da metade das unidades em análise.

4. Assuma que os resultados potenciais são gerador por

$$Y(0) \sim N(0,1), \quad \tau = -0.5 + Y(0), \quad eY(1) = Y(0) + \tau.$$

O mecanismo de atribuição de tratamento é de tal forma que $Z = 1$ se $\tau \geq 0$ e $Z = 0$ se $\tau < 0$. O resultado observado é dado por $Y = ZY(1) + (1 - Z)Y(0)$.

a. Calcule $\mathbb{E}(Y|Z = 1) - \mathbb{E}(Y|Z = 0)$.

b. Simule um caso (apenas uma replicação) para verificar o resultado obtido no item anterior.

c. **Dica:** se $X \sim N(\mu, \sigma)$, $\mathbb{E}(X|a < X < b) = \mu - \sigma \dfrac{\phi(\frac{b-\mu}{\sigma}) - \phi(\frac{a-\mu}{\sigma})}{\Phi(\frac{b-\mu}{\sigma}) - \Phi(\frac{a-\mu}{\sigma})}$.

4) a)

Se $Z_i = 0 \Rightarrow \tau_i < 0 \Rightarrow -0,5 + Y_i(0) < 0 \Rightarrow Y_i(0) < 0,5$

$$\Rightarrow E(Y|Z=0) = E(Y(0) \mid Y(0) < 0,5)$$

$$= 0 - 1 \times \frac{\phi(0,5) - \phi(-\infty)}{\Phi(0,5) - \Phi(-\infty)} = -\frac{\phi(0,5)}{\Phi(0,5)} \approx -0,509$$

Dado $Y_i(1) = Y_i(0) + \tau_i = 2 \times Y_i(0) - 0,5 \Rightarrow Y(1) \sim N(-0,5, 2)$

Se $Z_i = 1 \Rightarrow \tau_i \geq 0 \Rightarrow Y(0) \geq 0,5 \Rightarrow Y(1) \geq 2 . 0,5 - 0,5 = 0,5$

$$\Rightarrow E(Y|Z=1) = E(Y(1) \mid Y(1) \geq 0,5)$$

$$= -0,5 - 2 \times \frac{\phi(\infty) - \phi\left(\frac{0,5+0,5}{2}\right)}{\Phi(\infty) - \Phi\left(\frac{0,5+0,5}{2}\right)}$$

$$= -0,5 + 2 \times \frac{\phi(0,5)}{1 - \Phi(0,5)} \approx 1,782$$

Entao $E(Y|Z=1) - E(Y|Z=0) \approx 2,291$

5. De um exemplo numérico (simule um caso) em que $\tau = n^{-1}\sum_{i=1}^{n}(Y_i(1) - Y_i(0)) > 0$ mas a proporção de vezes que $Y_i(1) > Y_i(0)$ é menor do que 0.5. Ou seja, o efeito causal médio é positivo mas o tratamento beneficia menos da metade das unidades em análise.

$$Y(1) \sim B(p_1) \rightarrow \begin{cases} Z_i = 1 & \text{se } Y_i(1) > Y_i(0) \\ Z = 0, & \text{c.c} \end{cases}$$
$$Y(0) \sim B(p_0)$$

$$\Rightarrow \tau = E[Y \mid Z=1] - E[Y \mid Z=0] = p_1 - p_0$$

$$\Rightarrow Pr(Y(1) > Y(0)) = Pr(Y(1) = 1) \cdot Pr(Y(0) = 0)$$
$$= p_1 \times (1 - p_0)$$

In [49]:

```
n_samples = 10000

p_1 = 0.5
y_1 = bernoulli.rvs(p_1, size=n_samples)
p_0 = 0.3
y_0 = bernoulli.rvs(p_0, size=n_samples)

tau = sum(y_1 - y_0) / n_samples
print(f'tau = {tau}')

y_1_gt_y_0 = np.sum(y_1>y_0) / n_samples
print(f'%[Y(1)>Y(0)] = {y_1_gt_y_0 *100}%')
```

```
tau = 0.195
%[Y(1)>Y(0)] = 34.69%
```

In [50]:

```
print("Exact")
tau_Exact = p_1 - p_0
print(f'tau = {tau}')
print(f'%[Y(1)>Y(0)] = {p_1 * (1-p_0) * 100}%')
```

```
Exact
tau = 0.195
%[Y(1)>Y(0)] = 35.0%
```