Análise de Mediação

Introdução, Método do Produto e Definição Formal

José Luiz Padilha – DEST/UFPR

09 de maio, 2025

Conteúdo

1. Introdução
2. Abordagem de Baron e Kenny
3. Aplicação 1: Dados simulados
4. Um pouco de teoria
5. Aplicação 2: Dados reais
6. Bibliografia

Introdução

Introdução

Na análise de mediação, o objetivo é entender como uma variável mediadora M pode explicar a relação entre uma variável independente T e uma variável dependente Y.

Análise de mediação: investigar mecanismos causais examinando o papel de variáveis intermediárias que (supostamente) estão caminho causal entre as variáveis tratamento (exposição) e resposta.

Um mediador deve, portanto, ser uma variável pós tratamento que ocorre antes da resposta ser realizada.

Exemplo 1

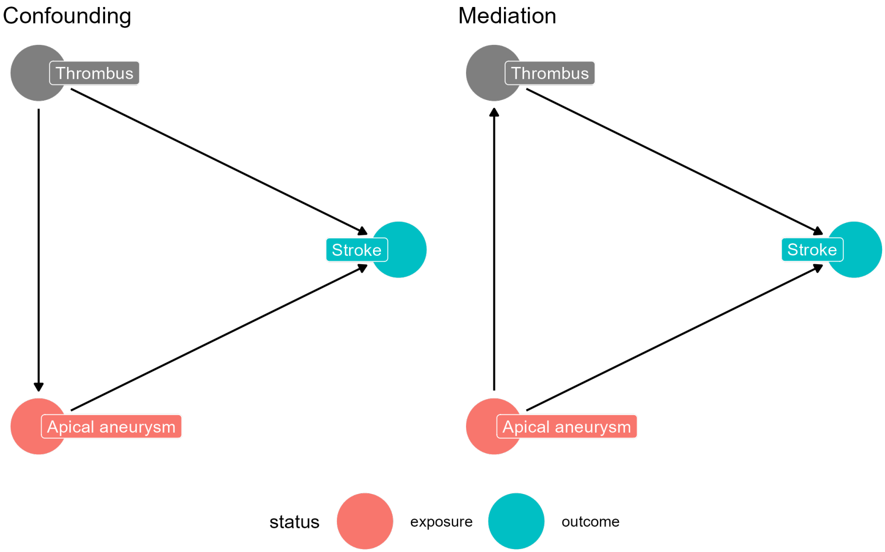
Considere um estudo em que se busca determinar os fatores de risco associados ao acidente vascular cerebral (AVC) cardioembólico em pacientes com cardiopatia chagásica crônica (CCC).

Diversas características clínicas, eletrocardiográficas e ecocardiográficas foram definidas na literatura como fatores de risco para AVC na Doença de Chagas.

Em particular, o foco está no efeito do aneurisma apical do ventrículo esquerdo, alteração da morfologia cardíaca típica da CCC, sobre o risco de AVC.

O aneurisma apical é um fator confundidor ou de mediação? A figura a seguir ilustra, de forma resumida, as duas situações.

Exemplo 1: confundidor ou mediador?



Exemplo 2

Considere um outro estudo que busca avaliar o efeito do sexo sobre o controle de anticoagulação oral.

A população de interesse faz uso de varfarina, um anticoagulante oral, terapia indicada para pacientes com doença cardiovascular e elevado risco de tromboembolismo.

A variável resposta é chamada TTR (tempo em que o paciente permanece dentro da faixa terapêutica) e varia de 0% a 100%.

Fatores como idade, tabagismo, etilismo, comorbidades, e uso de outros medicamentos, podem ser considerados fatores de confundimento ou mediadores?

Exemplo 3

Suponha que estamos interessados na eficácia clínica de um novo produto farmacêutico que está entrando no mercado.

Suponha que os pacientes são aleatorizados em dois grupos (Grupo A e Grupo B), e o principal desfecho de interesse é a expectativa de vida.

Suponha ainda que esses pacientes tiveram sua sensação de bem-estar medida após a atribuição do tratamento.

Em tese, essa sensação de bem-estar poderia ser afetada pela atribuição do tratamento. Acreditamos ainda que o bem-estar tenha impacto na expectativa de vida.

Neste caso, a sensação de bem-estar não é uma covariável baseline, mas um mediador, pois é medida após a atribuição ao tratamento.

Abordagem de Baron e Kenny

Abordagem de Baron e Kenny

Considere uma amostra de tamanho n e que para cada unidade (indivíduo) observamos (T_i, M_i, X_i, Y_i).

T_i é uma variável binária, M_i é a variável mediadora, X_i é um vetor de covaráveis observadas pré exposição e Y_i é a resposta de interesse.

Baron e Kenny (1986) propuseram uma abordagem para análise de mediação, que é baseada no seguinte sistema de equações lineares (linear structural equation model - LSEM).

$$Y_i = \alpha_1 + \beta_1 T_i + \varepsilon_{1i} \tag{1}$$
$$M_i = \alpha_2 + \beta_2 T_i + \varepsilon_{2i} \tag{2}$$
$$Y_i = \alpha_3 + \beta_3 T_i + \gamma M_i + \varepsilon_{3i} \tag{3}$$

Abordagem de Baron e Kenny

- Covariáveis pré exposição podem ser incluídas no modelo.

- Baron e Kenny (1986) sugeriram que a existência de efeitos de mediação podem ser testados ao ajustar separadamente as três regressões lineares e testar as hipóteses nulas (1) $\beta_1 = 0$, (2) $\beta_2 = 0$, e (3) $\beta_3 = 0$.
  - Se todas as hipóteses nulas são rejeitadas, então $\beta_2 \times \gamma$ ("método do produto dos coeficientes") pode ser interpretado como um efeito de mediação (efeito indireto de T em Y).
  - Efeito direto de T em Y: β_3.

- Equação (1) é redundante e desnecessária.

Aplicação 1: Dados simulados

Aplicação 1[1]

Suponha que o comprimento da sépala influencia a chance de uma flor ser polinizada por uma abelha.

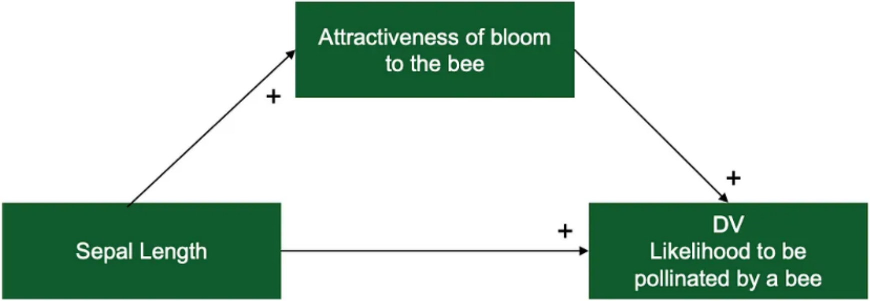
No entanto, este efeito não seria um efeito direto, mas sim mediado pelo grau de atratividade da flor para a abelha.

A hipótese é que existe um efeito indireto do comprimento da sépala na chance de a flor ser polinizada que ocorre através da atratividade da flor.

Serão usados os dados do conjunto `iris`.

[1] Exercício baseado em https://medium.com/data-science/doing-and-reporting-your-first-mediation-analysis-in-r-2fe423b92171

Aplicação 1


Attractiveness of bloom to the bee
+
+
+
Sepal Length
DV
Likelihood to be pollinated by a bee

Aplicação 1

```r
data(iris)
set.seed(1234)
head(iris)
```

```
  Sepal.Length Sepal.Width Petal.Length Petal.Width Species
1          5.1         3.5          1.4         0.2  setosa
2          4.9         3.0          1.4         0.2  setosa
3          4.7         3.2          1.3         0.2  setosa
4          4.6         3.1          1.5         0.2  setosa
5          5.0         3.6          1.4         0.2  setosa
6          5.4         3.9          1.7         0.4  setosa
```

```r
df <- data.frame(Sepal.Length = sample(iris$Sepal.Length, 1000,
                                       replace = TRUE))
```

Como apenas o comprimento da sépala está disponível, as duas outras variáveis serão simuladas. Para estabilidade das estimativas, usaremos $n = 1000$.

Aplicação 1

Primeiro, geramos o mediador (atratividade), em função do comprimento da sépala.

```
df$mediator <- df$Sepal.Length*0.5 + rnorm(nrow(df))
```

O ruído aleatório pode envolver outros atributos específicos da flor, como a cor, forma, cheiro, etc.

Agora, geramos a variável dependente (chance de polinização) em função do mediador. Neste caso, sem efeito direto do comprimento da sépala na chance de polinização.

```
df$dv <- df$mediator*0.8 + rnorm(nrow(df))
```

O ruído aleatório pode envolver atributos como localização da flor, condições de clima, solo, etc.

Aplicação 1

No sistema de equações

$$
\begin{align}
Y_i &= \alpha_1 + \beta_1 T_i + \varepsilon_{1i} \tag{1}\\
M_i &= \alpha_2 + \beta_2 T_i + \varepsilon_{2i} \tag{2}\\
Y_i &= \alpha_3 + \beta_3 T_i + \gamma M_i + \varepsilon_{3i} \tag{3}
\end{align}
$$

fizemos $\alpha_2 = 0$, $\beta_2 = 0.50$ em (2), e $\alpha_3 = 0$, $\beta_3 = 0$, e $\gamma = 0.80$.

- Assim, exatos 40.0% ($\beta_2 \times \gamma$) da informação do "comprimento da sépala" é transmitida para a variável resposta (chance de polinização). O efeito mediado é igual ao efeito total esperado.
- O efeito direto (β_3) vale zero, pois não há efeito direto do comprimento da sépala na chance de polinização.

Aplicação 1

```r
fit.totaleffect <- lm(dv ~ Sepal.Length, df)
printCoefmat(coef(summary(fit.totaleffect)))
```

```
             Estimate Std. Error t value  Pr(>|t|)
(Intercept)  0.204231   0.283506  0.7204    0.4715
Sepal.Length 0.370826   0.047696  7.7747 1.877e-14 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

O efeito total estimado ($\hat{\beta}_1 = 0.371$) é significativo e bastante próximo do valor real ($\beta_2\gamma + \beta_3 = 0.50 \times 0.80 + 0 = 0.40$).

Aplicação 1

```r
fit.mediator <- lm(mediator ~ Sepal.Length, df)
printCoefmat(coef(summary(fit.mediator)))
```

```
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  0.099888   0.219881  0.4543   0.6497
Sepal.Length 0.482243   0.036992 13.0363   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

O efeito do tratamento no mediador também é significativo e próximo do valor real ($\beta_2 = 0.50$).

Aplicação 1

```r
fit.dv <- lm(dv ~ Sepal.Length + mediator, df)
printCoefmat(coef(summary(fit.dv)))
```

```
                Estimate Std. Error t value Pr(>|t|)
 (Intercept)  0.12758757 0.22797920  0.5596   0.5758
 Sepal.Length 0.00080675 0.04148779  0.0194   0.9845
 mediator     0.76728699 0.03281686 23.3809   <2e-16 ***
 ---
 Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' '
```

Note a significância do efeito do mediador, mas não do tratamento. Este resultado indica uma mediação total.

Se o efeito do tratamento ainda fosse significativo, teríamos uma mediação incompleta ou parcial. Ou seja, haveria um efeito direto do tratamento na resposta.

Aplicação 1

Utilizaremos o pacote `mediation` para fazer a análise de mediação. Entraremos com os modelos estimados e o pacote os utiliza para realizar a análise de mediação, com intervalos de confiança bootstrap e valores-p.

Os principais quantidades de interesse são:

- `ACME` (*average causal mediation effects*). É o efeito indireto do tratamento (comprimento da sépala) sobre a resposta (chance de polinização) que passa pelo mediador (atratividade da flor).
- `ADE` (*average direct effects*). É o efeito direto do tratamento na resposta.
- `Total Effect`. É o efeito total (direto + indireto) do tratamento na resposta
- `Prop. Mediated`. Descreve a proporção do efeito do tratamento na resposta que passa pelo mediador.

Aplicação 1

```
library(mediation)
results <- mediate(fit.mediator, fit.dv, treat = 'Sepal.Length',
                   mediator = 'mediator', boot = T)
summary(results)
```

```

Causal Mediation Analysis

Nonparametric Bootstrap Confidence Intervals with the Percentile M

               Estimate 95% CI Lower 95% CI Upper p-value
ACME           0.370019     0.309181         0.43  <2e-16 ***
ADE            0.000807    -0.083465         0.09    0.97
Total Effect   0.370826     0.281742         0.47  <2e-16 ***
Prop. Mediated 0.997824     0.800069         1.27  <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Sample Size Used: 1000
```

Aplicação 1

Principais resultados:

- `ACME` (efeito médio causal de mediação) é significativo e igual a 0.37. É o efeito indireto do comprimento da sépala na chance de polinização que passa pelo mediador (atratividade da flor).

- `ADE` (efeito médio direto) é igual a 0.001 e não significativo. É o efeito direto do comprimento da sépala na chance de polinização. O efeito é nulo.

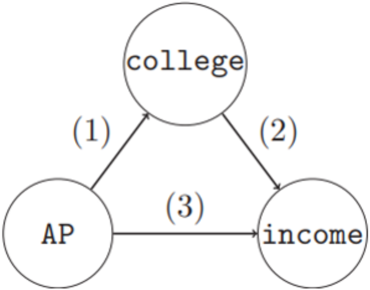
- `Total Effect` é igual a 0.371 e significativo. Soma dos efeitos diretos e indiretos.

- `Prop. Mediated` é igual a 0.998 e significativo. Proporção do efeito na resposta que passa pelo mediador. No caso, temos mediação total, pois o efeito direto é nulo.

Um pouco de teoria

Um pouco de teoria

Considere um estudo de um programa que incentiva alunos do ensino médio a participarem de determinado programa. Temos, como hipótese, a seguinte relação causal.



Um pouco de teoria

Em termos das variáveis (T_i, M_i, X_i, Y_i):

- T: *advanced placement (AP) course*[2],
- M: *college attendance*, indicando se o aluno frequentou o ensino superior.
- Y: log do salário, por exemplo, após 10 anos.
- X: *ses*, covariável pré tratamento que indica o nível socioeconômico do aluno.

A relação causal entre T e Y é hipoteticamente mediada por M: fazer o curso AP incentivaria ou ajudaria os alunos a frequentar a faculdade.

[2] curso oferecido em escolas dos Estados Unidos e permitem que os alunos do ensino médio estudem em um nível universitário e, em alguns casos, obtenham créditos universitários.

Um pouco de teoria

M é uma função do tratamento, assumindo os valores $M(T = t)$.

A resposta potencial pode ser escrita como função de T e M, ou seja, $Y(T = t, M = m)$, ou $Y(t, m)$.

- Por exemplo, $Y_i(T_i = 1, M_i = M_i(1)) = Y_i(T_i = 1)$ é o log do salário de um aluno que participa do curso AP e tem o nível de ensino resultante [de ter participado].
- Alternativamente, $Y_i(T_i = 0, M_i = M_i(0)) = Y_i(T_i = 0)$ é o log do salário do indivíduo que não participa do programa e tem o nível de ensino resultante.

A formulação $M_i = M_i(t)$ permite avaliar respostas estritamente contrafactuais. Por exemplo $Y_i(T_i = 1, M_i = M_i(0))$ representa a log renda do indivíduo se ele participa do programa, mas com o nível de formação superior que ele teria sem ter participado do programa.

Um pouco de teoria

Suponha T binário. O efeito total do tratamento é dado por

$$\tau_i = Y_i(1, M_i(1)) - Y_i(0, M_i(0)).$$

O efeito causal de mediação é:

$$\delta_i(0) = Y_i(T_i = 0, M_i = M_i(1))) - Y_i(T_i = 0, M_i = M_i(0))),$$

ou

$$\delta_i(1) = Y_i(T_i = 1, M_i = M_i(1))) - Y_i(T_i = 1, M_i = M_i(0))).$$

É a diferença na resposta potencial quando o mediador assume o valor que teria na condição de tratamento em oposição à condição de controle, enquanto o próprio tratamento é mantido constante.

O ACME (*average causal mediation effect*) é dado por $E(\delta)$.

Um pouco de teoria

Efeitos diretos, são também de dois tipos:

$$\zeta_i(0) = Y_i(T_i = 1, M_i = M_i(0)) - Y_i(T_i = 0, M_i = M_i(0)),$$

ou

$$\zeta_i(1) = Y_i(T_i = 1, M_i = M_i(1)) - Y_i(T_i = 0, M_i = M_i(1)).$$

Representa a diferença na resposta potencial quando o tratamento é alterado, mas o mediador é mantido constante no valor que seria obtido se o tratamento fosse igual a 0 ou 1.

O ADE (*average direct effect*) é dado por $E(\zeta)$.

Um pouco de teoria

A soma das duas quantidades é igual ao efeito total do tratamento (τ).

Por exemplo,

$$
\begin{aligned}
\zeta(1) + \delta(0) =& Y(T = 1, M = M(1))) - Y(T = 0, M = M(1)))+ \\
& Y(T = 0, M = M(1))) - Y(T = 0, M = M(0))) \\
=& Y(T = 1, M = M(1))) - Y(T = 0, M = M(0))) \\
=& Y(T = 1) - Y(T = 0)
\end{aligned}
$$

De forma similar para $\zeta(0) + \delta(1)$.

A análise de mediação decompõe o efeito tratamento T sobre Y em uma porção dependente de M e uma dependente de outros mecanismos.

Aplicação 2: Dados reais

Aplicação 2

Relação aneurisma e AVC.

Bibliografia

Principais Referências Usadas

Pearl, J., "Causality. Models, Reasoning and Inference", *Cambridge University Press*, 2009.

Reis, R. C. P. "MAT02010 - Tópicos Avançados em Estatística II. Introdução a Inferência Causal". *UFRGS*, 2019.

Brumback, B. A., "Fundamentals of Causal Inference With R", *CRC Press*, 2022.

Sales, A. C. "mediation Package in R. Journal of Educational and Behavioral Statistics", 2017