DCP098
Fundamentos para Avaliação Quantitativa de Políticas Públicas

Regressão múltipla e controle estatístico

Aula 09
02 de maio de 2022

Ana Paula Karruz

Como é que chama o nome disso?

Definição	**Exemplo**
O **estimando** (*estimand*) é a quantidade de interesse cujo valor verdadeiro desejamos conhecer; aka: **parâmetro**, **coeficiente**	β_1
O **estimador** (*estimator*) é um método para estimar o estimando	$\hat{\beta}_1 = \frac{\sum_{i=1}^{N}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{N}(X_i - \bar{X})^2}$
A **estimativa** (*estimate*) é um valor numérico para o estimando que resulta do uso de um estimador particular; aka: **coeficiente estimado**	$\hat{\beta}_1$

Regressão simples

$Y = \beta_0 + \beta_1 X_1 + \varepsilon$	$\beta_1 = \Delta Y / \Delta X_1$

- Tudo que desconhecemos é incorporado ao ε (erro aleatório)
- Por premissa, ε tem média zero e não se correlaciona com X_1; assim, o efeito médio das variáveis omitidas é capturado pelo β_0
- Porque é difícil manter tudo o mais constante na prática e porque não controlamos por outros fatores, é como se só enxergássemos $\Delta X \rightarrow \Delta Y$, sem **considerar um possível ΔZ**
- Portanto, a regressão simples em si **não nos permite verificar se há causalidade**

Regressão multivariada é uma regressão com duas ou mais variáveis explicativas

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, ..., X_k$ são mantidos constantes)

Interpretação detalhada adiante

- Cada X é uma variável independente diferente
- k é o número total de variáveis independentes

*Muitas vezes, **uma única variável ou** talvez um **subconjunto de variáveis são de interesse primário**. Referimo-nos às **outras variáveis independentes** como **variáveis de controle**, pois são incluídas para controlar os fatores que podem afetar a variável dependente e serem correlacionados com as variáveis independentes de interesse primário. **Variáveis de controle e grupos de controle são diferentes**: uma variável de controle é uma variável adicional que incluímos em um modelo, enquanto um grupo de controle é o grupo ao qual comparamos o grupo de tratamento em um experimento.*

Bailey (2016: 203)

Regressão multivariada é uma regressão com duas ou mais variáveis explicativas

$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$	$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, ..., X_k$ são mantidos constantes)

Como calcular os parâmetros?

- Assim como na regressão bivariada, MQO encontra os βhats que minimizam a soma dos quadrados dos resíduos

$$\hat{\epsilon}_i^2 \quad = \quad (Y_i - (\hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}))^2$$

- As “fórmulas vão ficando desajeitadas”; para k = 2:

$$\hat{\beta}_1 = \frac{(\sum yx_1)(\sum x_2^2) - (\sum yx_2)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{\beta}_2 = \frac{(\sum yx_2)(\sum x_1^2) - (\sum yx_1)(\sum x_1 x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1 x_2)^2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}_1 - \hat{\beta}_2 \overline{X}_2$$

Melhor não calcular isto a mão ☺

where lower case variables indicate deviations from the mean, as in $y = Y_i - \overline{Y}$; $x_1 = X_{1i} - \overline{X}_1$; and $x_2 = X_{2i} - \overline{X}_2$.

Fonte: Studenmund (2006: 43).

Regressão multivariada é uma regressão com duas ou mais variáveis explicativas

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_k X_k + \varepsilon$$

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, \ldots, X_k$ são mantidos constantes)

Como interpretar $\beta_1$hat (ou qualquer dos outros coeficientes de inclinação)?

- Um aumento de uma unidade no respectivo X está associado a uma variação estimada de $\beta_1$hat em Y, **mantendo-se constante(s) a(s) outra(s) variável(eis) explicativa(s) incluída(s) do modelo**
  - Usualmente, substitui-se o trecho final (negritado) por **ceteris paribus ou cæteris paribus**, que significa “tudo o mais constante” em latim

Regressão múltipla

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, ..., X_k$ são mantidos constantes)

- A regressão múltipla fortalece nossa capacidade de apurar causalidade; com ela, podemos incluir tantos controles quanto desejarmos

MQO multivariado combate a endogeneidade "puxando" variáveis do termo de erro para a equação estimada.
Bailey (2016:207)

- Todavia, a regressão múltipla **também é limitada como instrumento para apurar não espuriedade/ causalidade**: sempre pode existir um outro fator determinante de Y que esteja correlacionado com algum X incluído na equação e sobre o qual (esse fator omitido) não possuímos ciência ou dados

Um variável independente é exógena se sua variação não é relacionada a fatores embutidos no ε

Exogeneidade é o oposto de endogeneidade

“exo” = externo; variável está fora do modelo no sentido de que não se correlacionada com outros fatores que influenciam Y

Exogeneidade: **corr(X,ε) = 0**

“endo” = interno; variável está dentro do modelo no sentido de que se correlaciona com outros fatores que influenciam Y

Endogeneidade: **corr(X,ε) ≠ 0**

Lembrete
Ordem das variáveis não altera correlação: corr(X,ε) = corr(ε,X)

*Estatisticamente falando, destacamos esse grande desafio ao dizer que a variável donut é endógena. **Uma variável independente é endógena se as mudanças nela estiverem relacionadas a fatores no termo de erro. […] A endogeneidade está em toda parte; é endêmica.***

Bailey (2016: 14-15)

Regressão múltipla

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_k X_k + \varepsilon$$

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, ..., X_k$ são mantidos constantes)

Como pessoas práticas, reconhecemos que é improvável que possamos observar todas as fontes possíveis de endogeneidade que, se não incluídas na equação, comporão o termo de erro. Mas se pudermos medir mais variáveis e extrair mais fatores do termo de erro, nossas estimativas normalmente se tornarão menos tendenciosas e serão distribuídas mais próximas do valor real.

Bailey (2016:205-206)

- **Desenhos de pesquisa experimentais ou quase-experimentais** oferecem maior validade na apuração de relações causais

Se corr(ε,X) ≠ 0, então X é considerado uma variável endógena

Endogeneidade faz MQO produzir um estimador enviesado do verdadeiro β

Endogeneidade é uma preocupação constante em nossas vidas

Ilustração: como a regressão múltipla descontamina βhat do efeito das demais variáveis explicativas

Vide APÊNDICE: Sobre descontaminação de βhat

Controle estatístico: “mantendo-se constante(s) a(s) outra(s) variável(eis) explicativa(s) do modelo”

Em essência, a regressão multivariada calcula o βhat “líquido”, “descontaminado” do efeito de outras variáveis explicativas incluídas no modelo

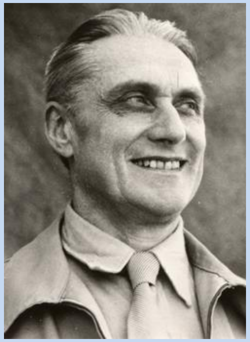
*[Ao interpretar estimativas de regressão multivariada] O que eles [estatísticos] realmente querem dizer é mais como **“descontar”** o efeito de outras variáveis no modelo. A lógica por trás de dizer que outros fatores são constantes é que, **uma vez que tenhamos calculado os efeitos de outras variáveis, é como se os valores dessas variáveis fossem iguais para cada observação**.*

…

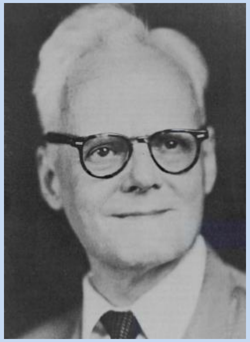
*Portanto, quando alguém diz algo como “**mantendo X_2 constante**, o efeito estimado de um aumento de uma unidade em X_1 é $\beta_1$hat”, o que se quer dizer é que, **considerando o efeito de X_2, estima-se que o efeito de X_1 seja $\beta_1$hat**.*

Bailey (2016: 197-198)

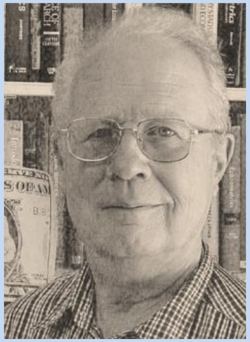
Teorema de Frisch-Waugh-Lovell (FWL)



Ragnar Frisch
(Norueguês, 1895-1973)



Frederick V. Waugh
(Americano, 1898-1974)



Michael C. Lovell
(Americano, 1930-)

- Em uma regressão multivariada, $\beta_1$hat pode ser obtido seguindo os passos abaixo:
  - Regresse X_1 em X_2, ..., X_k
  - Compute os resíduos (r_1) obtidos por essa regressão
  - Regresse Y em r_1

Regredimos Y especificamente na parte de X_1 que não se correlaciona com as demais variáveis explicativas

- O mesmo vale para todos os coeficientes de inclinação

Ilustrando o Teorema de Frisch-Waugh-Lovell (FWL)

```
> dados <-read_dta('auto.dta')

> summary(dados[,c("price", "mpg", "trunk", "foreign")])
     price            mpg            trunk          foreign      
 Min.   : 3291   Min.   :12.00   Min.   : 5.00   Min.   :0.0000  
 1st Qu.: 4220   1st Qu.:18.00   1st Qu.:10.25   1st Qu.:0.0000  
 Median : 5006   Median :20.00   Median :14.00   Median :0.0000  
 Mean   : 6165   Mean   :21.30   Mean   :13.76   Mean   :0.2973  
 3rd Qu.: 6332   3rd Qu.:24.75   3rd Qu.:16.75   3rd Qu.:1.0000  
 Max.   :15906   Max.   :41.00   Max.   :23.00   Max.   :1.0000  
```

Ilustrando o Teorema de Frisch-Waugh-Lovell (FWL)

	Dependent variable:			
	price			
	(1)	(2)	(3)	(4)
aux_reg_mpg_res	-261.989***			
	(69.595)			
aux_reg_trunk_res		83.646		
		(100.977)		
aux_reg_foreign_res			1,887.461**	
			(804.226)	
mpg				-261.989***
				(64.913)
trunk				83.646
				(86.501)
foreign				1,887.461***
				(711.416)
Constant	6,165.257***	6,165.257***	6,165.257***	10,033.080***
	(315.581)	(343.611)	(332.751)	(2,256.685)
Observations	74	74	74	74
R2	0.164	0.009	0.071	0.293
Adjusted R2	0.153	-0.004	0.058	0.263
Residual Std. Error	2,714.734 (df = 72)	2,955.856 (df = 72)	2,862.436 (df = 72)	2,532.103 (df = 70)
F Statistic	14.172*** (df = 1; 72)	0.686 (df = 1; 72)	5.508** (df = 1; 72)	9.683*** (df = 3; 70)
Note:				*p<0.1; **p<0.05; ***p<0.01

Nestes três modelos, X é um resíduo e, portanto, Xbar = 0; como a reta de regressão simples passa pela coordenada (Xbar, Ybar), nestes três casos o intercepto = Ybar

Algumas outras propriedades de MQO

Estas propriedades **não** dependem de atendimento às premissas

Regressão simples	Regressão múltipla
▪ Reta estimada passa pela coordenada (Xbar, Ybar)	▪ Reta estimada passa pela coordenada (Xbar, Ybar), desde que demais variáveis explicativas encontrem-se em seus valores médios
▪ Soma dos resíduos = 0, desde que haja um termo de intercepto na equação – Como a soma dos resíduos = 0 a média dos resíduos também é nula – Na prática, soma dos resíduos ≅ 0, por causa de arredondamentos	▪ Idem
▪ Média de Yhat = Ybar	▪ Idem

Atenção: a soma dos quadrados dos resíduos não é zero. Para que esta soma fosse zero, todos os resíduos teriam de ser nulos (já que o quadrado de um resíduo é sempre não negativo).

Aprofun-damento

Fonte: Hoffmann (2016:49-50).

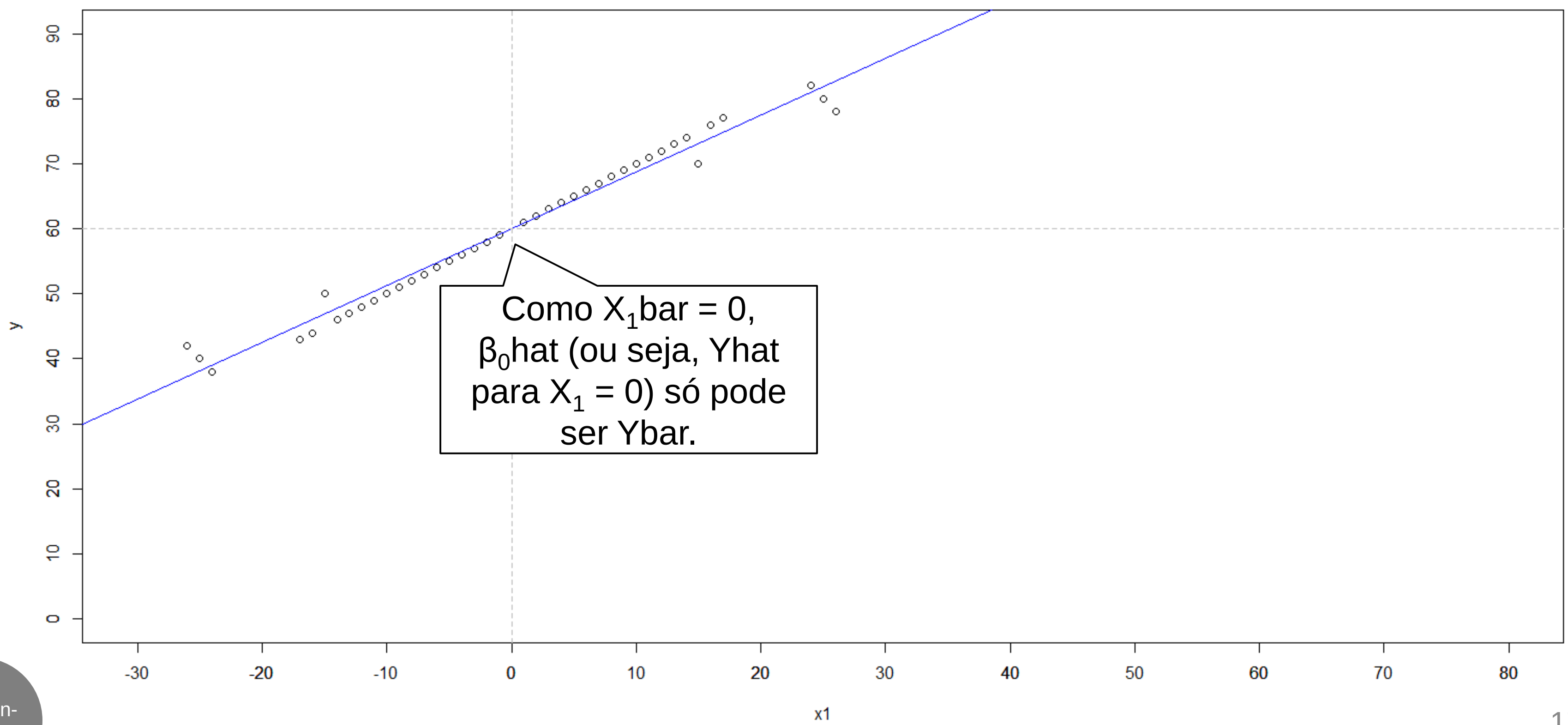
Na regressão simples, a reta estimada passa, necessariamente, por (Xbar, Ybar)

```
> summary(dados)
       y              x1               x2       
 Min.   :38   Min.   :-26.00   Min.   :-20.0  
 1st Qu.:50   1st Qu.:-10.25   1st Qu.:  9.5  
 Median :60   Median :  0.00   Median : 30.0  
 Mean   :60   Mean   :  0.00   Mean   : 30.0  
 3rd Qu.:70   3rd Qu.: 10.25   3rd Qu.: 50.5  
 Max.   :82   Max.   : 26.00   Max.   : 80.0  
```

```
> reg1 = lm(y ~ x1, data = dados)
> summary(reg1)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 60.00000    0.28381  211.41   <2e-16 ***
x1           0.87548    0.02097   41.74   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

38 degrees of freedom
```



Aprofun-damento

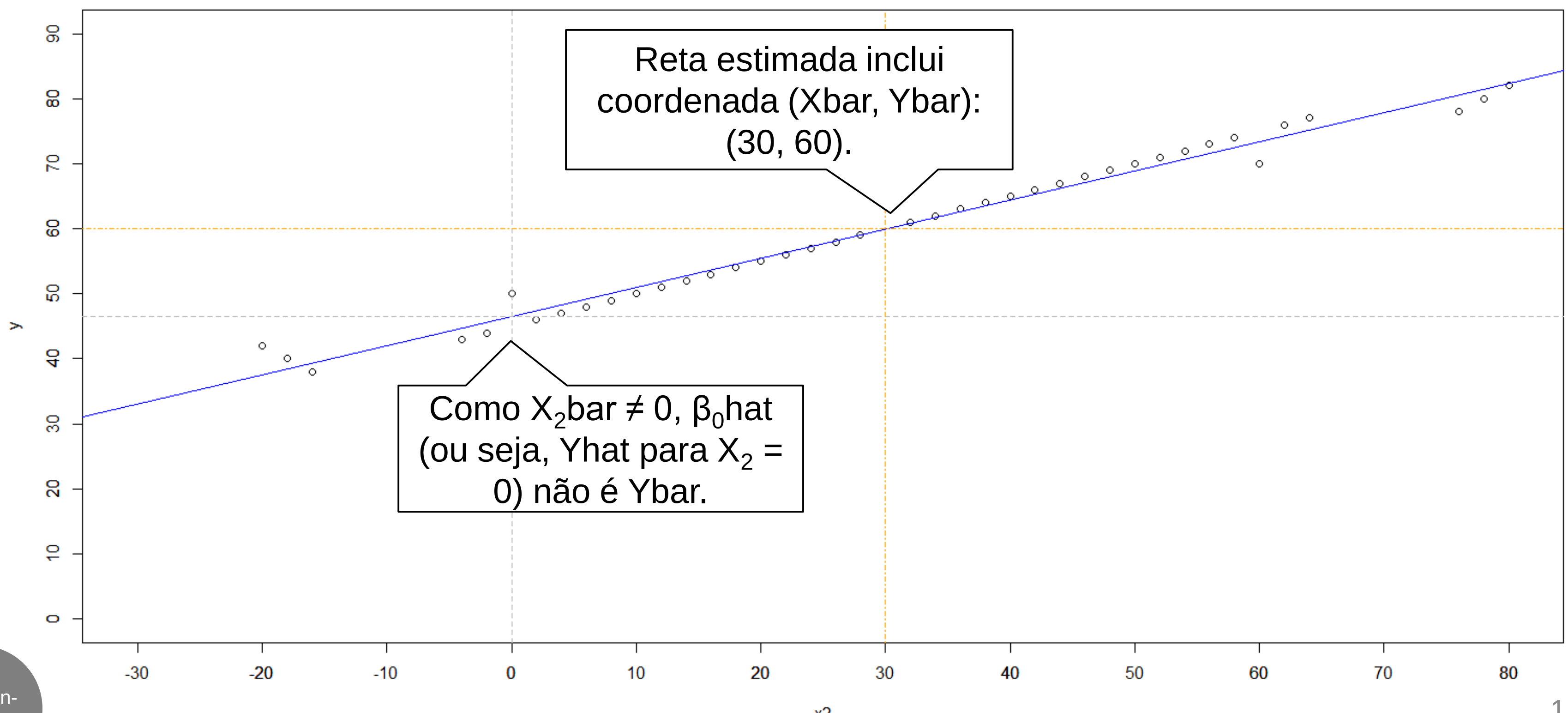
Na regressão simples, a reta estimada passa, necessariamente, por (Xbar, Ybar)

```
> summary(dados)
       y             x1                x2      
 Min.   :38   Min.   :-26.00   Min.   :-20.0  
 1st Qu.:50   1st Qu.:-10.25   1st Qu.:  9.5  
 Median :60   Median :  0.00   Median : 30.0  
 Mean   :60   Mean   :  0.00   Mean   : 30.0  
 3rd Qu.:70   3rd Qu.: 10.25   3rd Qu.: 50.5  
 Max.   :82   Max.   : 26.00   Max.   : 80.0  
```

```
> reg2 = lm(y ~ x2, data = dados)
> summary(reg2)

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 46.557610   0.367842  126.57   <2e-16 ***
x2           0.448080   0.009187   48.77   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

38 degrees of freedom
```



Aprofun-damento

As propriedades mais bacanas de MQO

Veja, propriedade é diferente de premissa

Se as premissas de MQO forem atendidas,
MQO é ***BLUE****!*

- ***Best*** *(variância mínima, i.e., máxima precisão)*
- ***Linear***
- ***Unbiased*** *(livre de vieses)*
- ***Estimator***

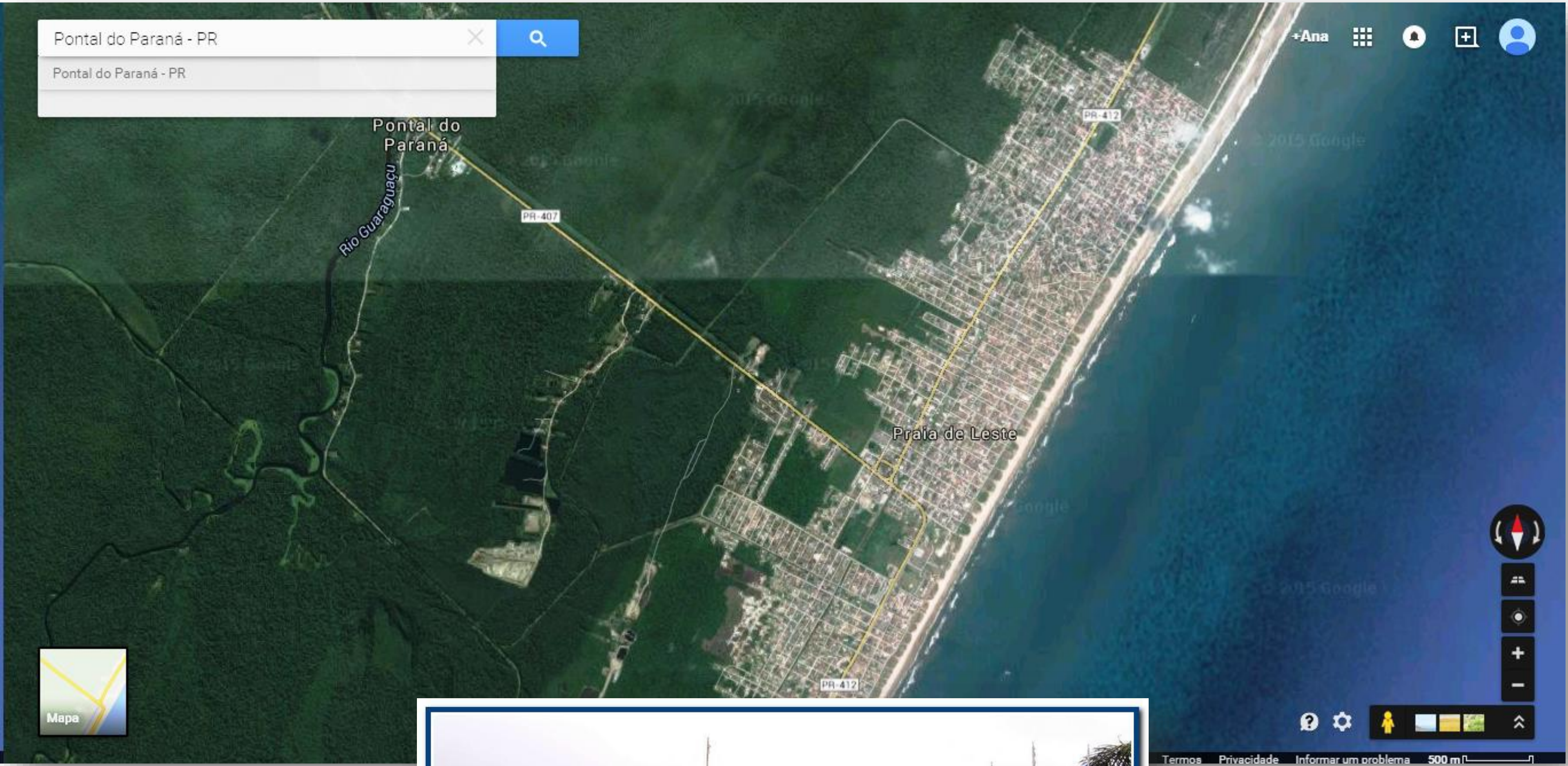
- Se as premissas de MQO estiverem atendidas, MQO é “melhor” em relação às alternativas, quais sejam: adaptações de MQO (e.g., Mínimos Quadrados Generalizados) e estimadores de Máxima Verossimilhança (um algoritmo iterativo que permite não linearidades nos βs)

- As propriedades BLUE do MQO são provadas pelo teorema de Gauss-Markov

Aprofun-
damento

APÊNDICE: Sobre descontaminação de βhat

Ilustração: Sorvete dos afogados



Reunião pública de prestação de contas do Prefeito e Secretários, 31/mar/2015

– **Secretário de Desenvolvimento Econômico:** *Nossa economia está bastante aquecida, principalmente em função do desempenho do setor de alimentos. Em particular, a venda de sorvetes cresceu enormemente no último trimestre, especialmente como sobremesa do almoço*.

– **Secretário de Saúde Pública:** *Minhas notícias são menos animadoras. O último trimestre foi marcado por um aumento substancial de mortes por afogamento, concentradas no fim da tarde*.

– **Presidente da Câmara Municipal:** *Visualizo oportunidade de alteração das nossas normas. É preciso banir o consumo de sorvete, antes que mais e mais dos nossos munícipes sejam vitimizados pela praga do afogamento!*

Proposição causal do Presidente da Câmara Municipal

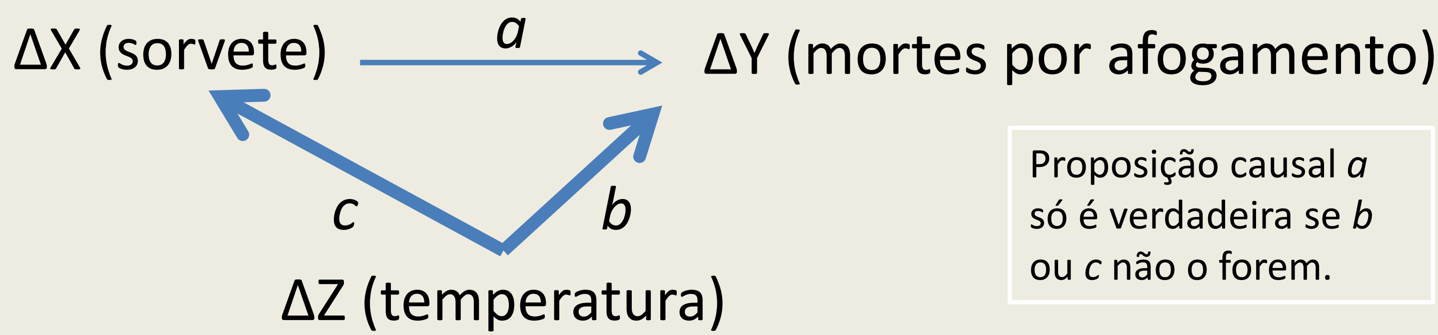
O aumento no consumo de sorvete é a causa da elevação das mortes por afogamento.

ΔX (sorvete) $\xrightarrow{a}$ ΔY (mortes por afogamento)

Como checar se a proposição causal *a* é verdadeira?

Critério	Atendido?	Evidência
Anterioridade	✓	▪ Consumo de sorvete está concentrado na sobremesa do almoço, enquanto afogamentos concentram-se no final da tarde
Correlação	✓	▪ Existe associação empírica entre consumo de sorvete e afogamentos (consumo de sorvete cresceu no último trimestre, assim como casos de afogamento); correlação > 0
Não-espuriedade	?	**Próximo passo:** Checar se existe algum outro fator que poderia explicar Δ afogamento e que se correlacione com consumo de sorvete (p.ex., Δ temperatura)

Temperatura atende aos 3 critérios, sugerindo que a proposição causal *a* é espúria



Proposição causal *a* só é verdadeira se *b* ou *c* não o forem.

Checando se as proposições causais *b* e *c* são verdadeiras:

Critério	Atendido?	Evidência
Anterioridade	✓	▪ Podemos admitir que o consumo de sorvete e as mortes por afogamento intensificaram-se alguns dias após o aumento da temperatura média diária
Correlação	✓	▪ Existe associação empírica entre temperatura e mortes por afogamento (temperatura elevou-se no último trimestre em relação ao trimestre anterior, assim como as fatalidades por afogamento); correlação > 0 ▪ Existe associação empírica entre consumo de sorvete e temperatura; correlação > 0
Não-espuriedade	✓	▪ Além de alta plausibilidade de que ΔZ esteja causando tanto ΔY quanto ΔX, não temos indicação de que qualquer coisa tenha mudado nas praias de Pontal: sinalização, número de salva-vidas por banhista, etc.

Muito incomodado, o Presidente da Câmara intervém:

– *Veja, entendo que o verão realmente leve a mais afogamentos e mortes por afogamento. Mas não estou convencido de que sorvetes sejam seguros para o povo de Pontal. Acredito que possam ser responsáveis por uma porção considerável das mortes, enquanto outras tantas, admito, são resultado lamentável do verão.*

Agora, o presidente está argumentando que a existência de *b* e *c* não exclui a possibilidade de existir uma relação causal parcial entre X e Y. E ele tem razão.

Analisando a proposição de causalidade parcial entre consumo de sorvete e mortes

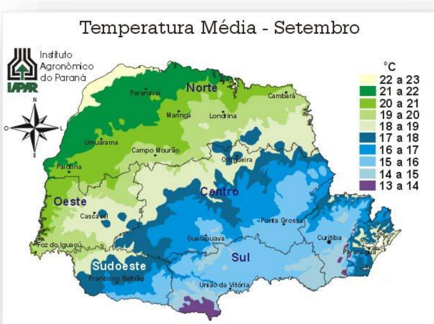
Se sorvete tem algum **efeito independente** em mortes por afogamento (ou seja, um efeito que não seja puramente fruto do aumento da temperatura), então:

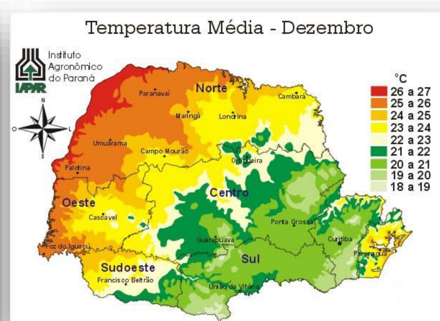
ΔX (sorvete) $\xrightarrow{a'}$ ΔY (mortes por afogamento)

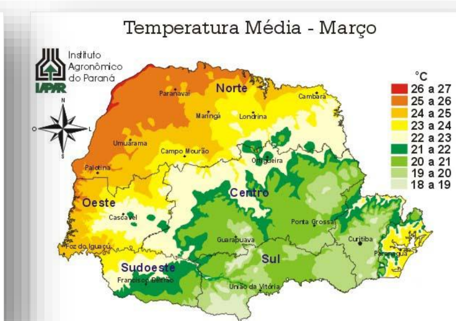
ΔZ (temperatura) = 0
(i.e., sempre quente ou sempre frio)

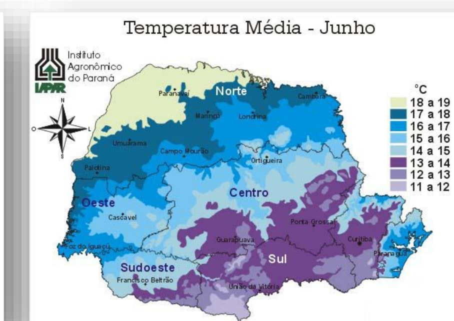
> Proposição causal *a’* só é verdadeira se existir correlação não nula entre X e Y num cenário de temperatura constante.

No Paraná, onde as estações do ano são bem definidas, não é possível observar ΔX e ΔY a valores constantes de Z em períodos mais longos que 3 meses.

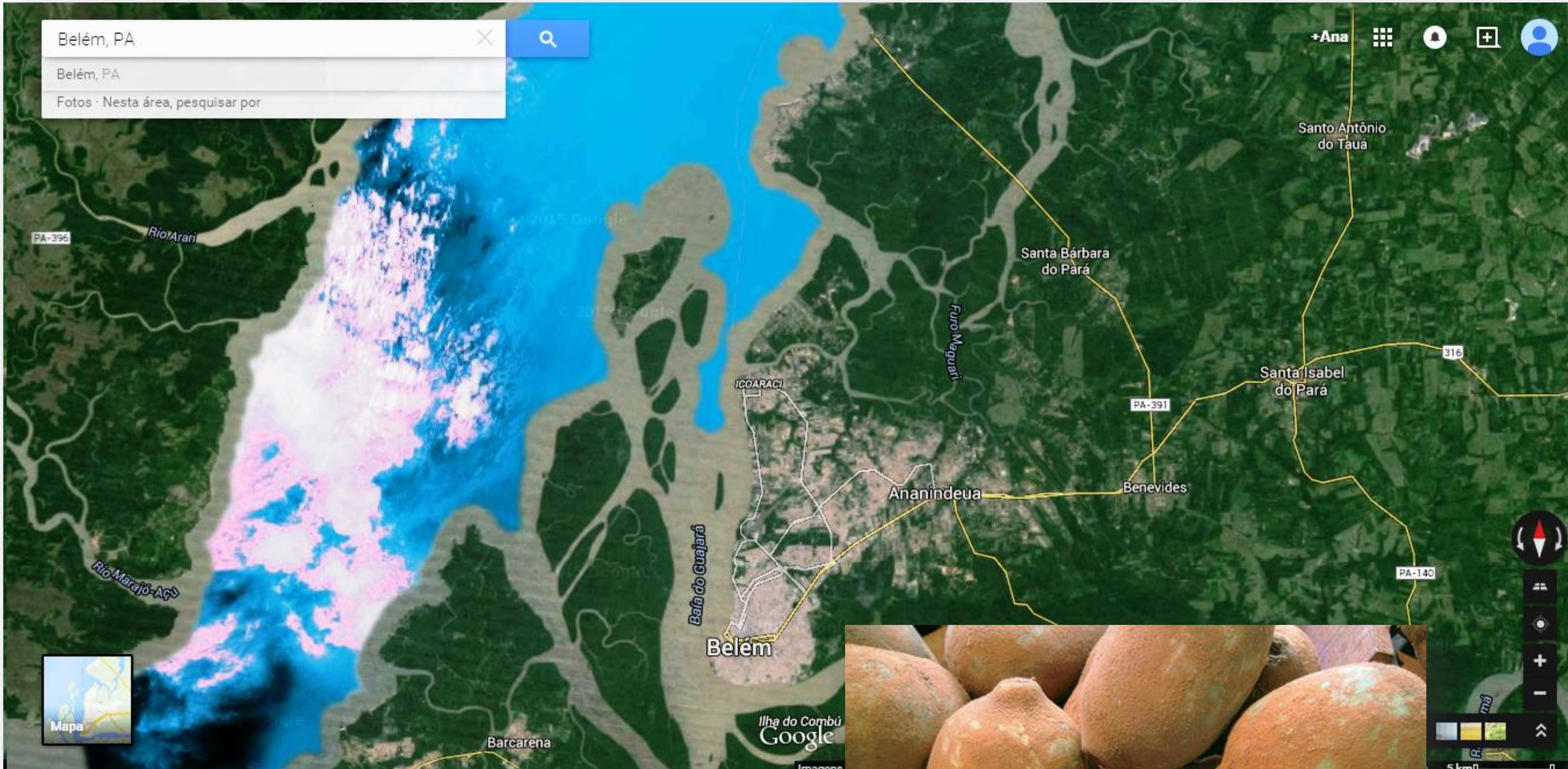








“Cidades das Mangueiras” e do cupuaçu, Belém é sempre quente



Será que existe uma **correlação** entre consumo de sorvete e mortes por afogamento em **Belém**, onde é **sempre quente**?

Dados foram apurados. Em Belém, correlação (sorvete, morte por afogamento) = 0

Não se dando por vencido, o Presidente da Câmara esbraveja:

- *Mas nossos sorvetes são diferentes dos de Belém! Os deles são de cupuaçu, os nosso de pinhão. Não dá para comparar!*

Percebendo a tensão no ar, técnicos presentes reúnem-se para planejar uma análise que pudesse ajudar a entender se realmente existe um potencial mortífero no sorvete de pinhão

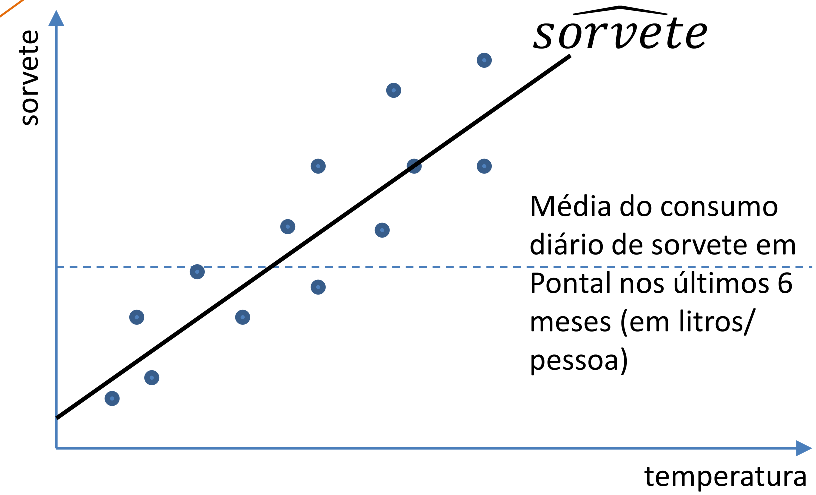
- É plausível imaginar que **além da temperatura, outros fatores** possam influenciar consumo de sorvete, p.ex.: lançamentos, mudança na preferência dos consumidores, disponibilidade de outras sobremesas

- Se assim for, ao tentar prever consumo de sorvete **meramente em função da temperatura**, acertaríamos algumas vezes mas **erraríamos feio** em outras tantas

- A parte da variação de **X que não pode ser explicada** por **Z** seria representada pelos **resíduos**, i.e., as distâncias entre nossas previsões e o que realmente aconteceu

“Encaixando” a reta de regressão nos dados e calculando os resíduos

Isto é uma regressão simples!

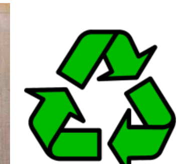


$$\widehat{sorvete} = \widehat{\beta_0} + \widehat{\beta_1} temperatura$$

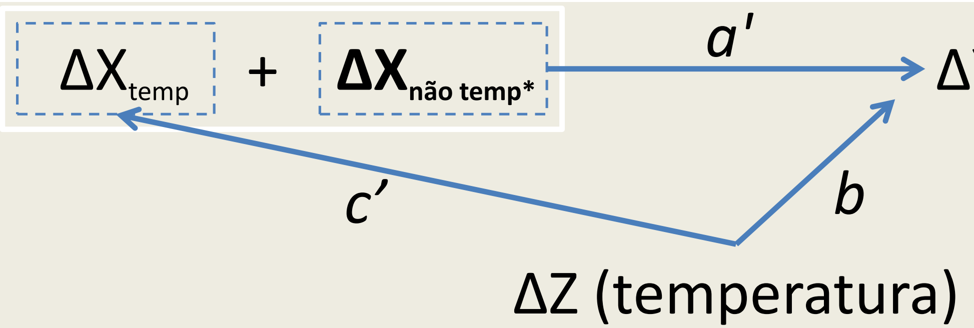
A distância entre cada ponto observado e a reta estimada representa o resíduo, ou seja, a parte da **variação de sorvete que não é explicada pela temperatura**:

$$resíduo = sorvete - \widehat{sorvete}$$

Técnicos agora correm para **checar se a variação residual de sorvete** poderia justificar **mortes por afogamento**.



A variação residual de sorvete explica as mortes por afogamento?



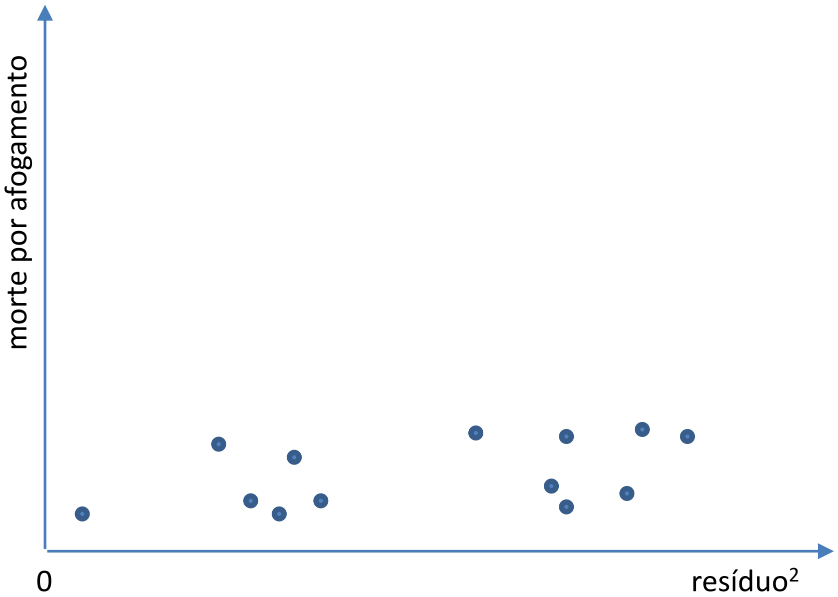
Proposição causal *a'* só é verdadeira **se o resíduo da regressão** de sorvete em temperatura **explicar** a variação de **mortes por afogamento.**

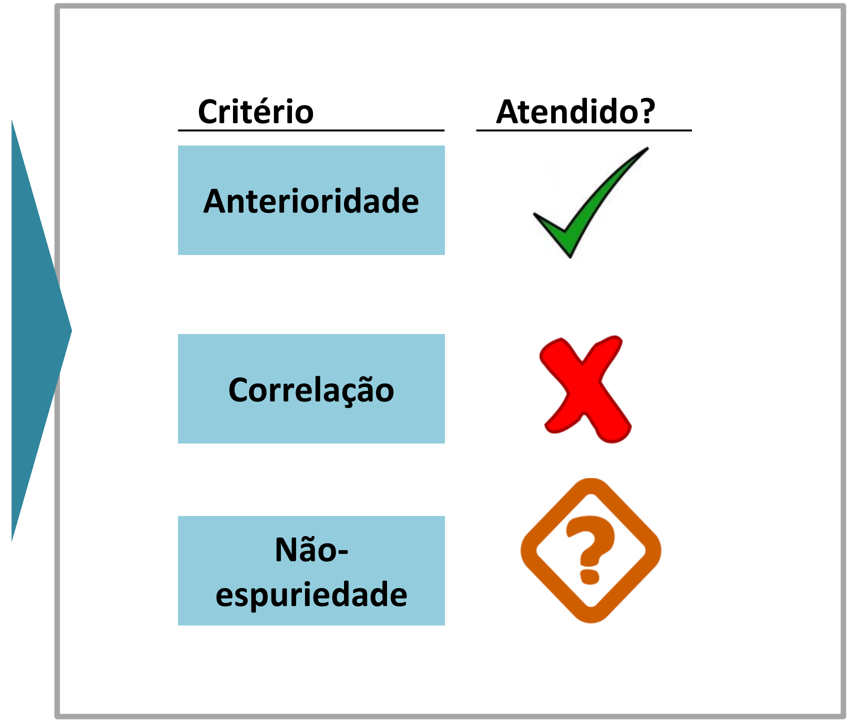
Como checar se a proposição causal *a'* é verdadeira?

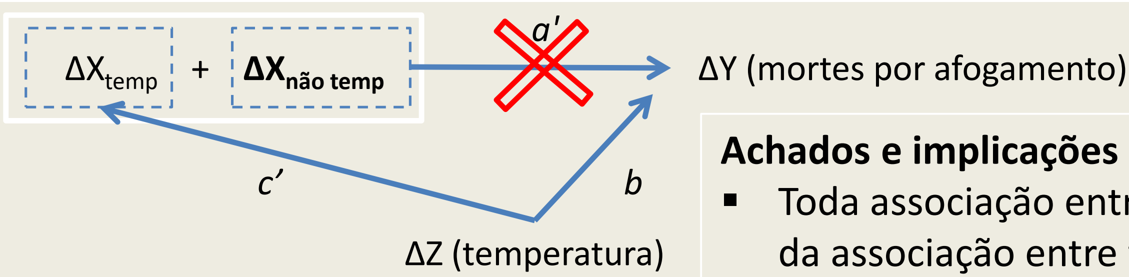
Critério	Atendido?	Evidência
Anterioridade	✓	▪ Consumo de sorvete está concentrado na sobremesa do almoço, enquanto afogamentos concentram-se no final da tarde
Correlação	?	Vamos trabalhar com 2 cenários para a correlação entre o resíduo de sorvete e as mortes por afogamento
Não-espuriedade	?	

* $\Delta X_{não\ temp}$ = ΔResíduo da regressão de sorvetes x temperatura.

Cenário 1: Correlação (resíduo², morte por afogamento) ≅ 0



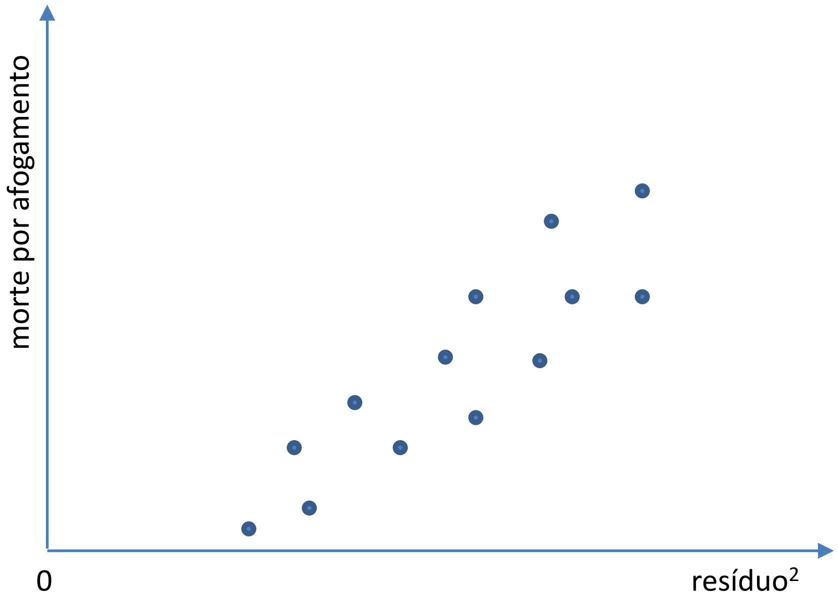




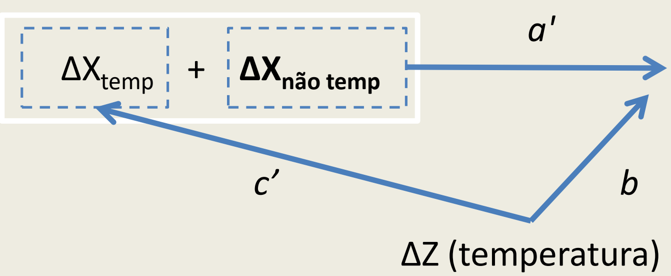
Achados e implicações práticas:

- Toda associação entre sorvete e mortes por afogamento vem da associação entre temperatura e consumo de sorvete; **não há causalidade parcial, ou seja, não existe um efeito independente de sorvete sobre mortes por afogamento**
- O Presidente da Câmara tirou conclusões precipitadas; **não será preciso proibir sorvete em Pontal**

Cenário 2: Correlação (resíduo², morte por afogamento) > 0



Critério	Atendido?
Anterioridade	
Correlação	
Não-espuriedade	



Achados e implicações práticas:

- **Mesmo após "controlar"** por temperatura – i.e., descontaminando consumo de sorvete de sua correlação com temperatura – a análise **não é capaz de derrubar a assertiva do Presidente da Câmara**
- **Medida cautelar** deve suspender consumo de sorvete até que a investigação seja concluída – i.e., que se **determine se há espuriedade** na relação entre o **resíduo** de sorvete e as **mortes** por afogamento.

O que acabou de acontecer?

- Nos 2 cenários, **não mantivemos a temperatura constante de fato** (nem conseguiríamos, dado o perfil climático local); não pudemos realmente observar a variação de X na ausência de variação de Z
- Todavia, conseguimos **descontaminar ΔX de ΔZ** de outra forma: separando a **parte de ΔX correlacionada com ΔZ da parte de ΔX que é independente de ΔZ**, e focando nesta última
- Porque chegamos ao mesmo efeito (descontaminação de ΔX), consideramos que ***a'* é relação (parcial) entre ΔX e ΔY mantendo a temperatura "constante" ou "controlando" pela temperatura**

É para isto que serve a regressão múltipla!

DCP098
Fundamentos para Avaliação Quantitativa de Políticas Públicas

Regressão múltipla e controle estatístico

Aula 09
02 de maio de 2022

Ana Paula Karruz