PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

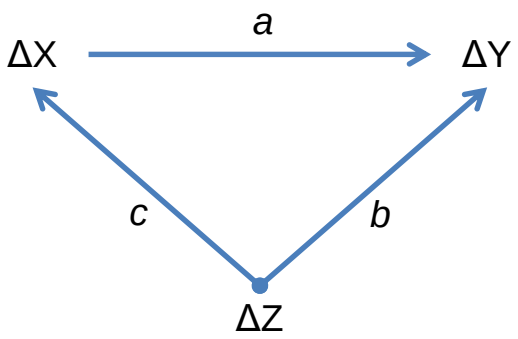
Regressão bivariada
Noções de regressão multivariada

Aula 2
14 de setembro de 2022

Ana Paula Karruz

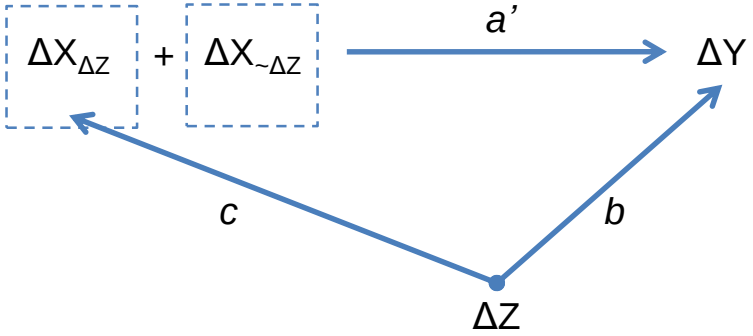
Ilustrando uma associação espúria: como evitar essa cilada?

a é associação espúria; ***a'*** pode não ser espúria



Associação *a* é espúria (i.e., não é uma medida válida da associação causal entre X e Y) porque reporta uma combinação de dois efeitos: efeito de X sobre Y e efeito de Z sobre Y.

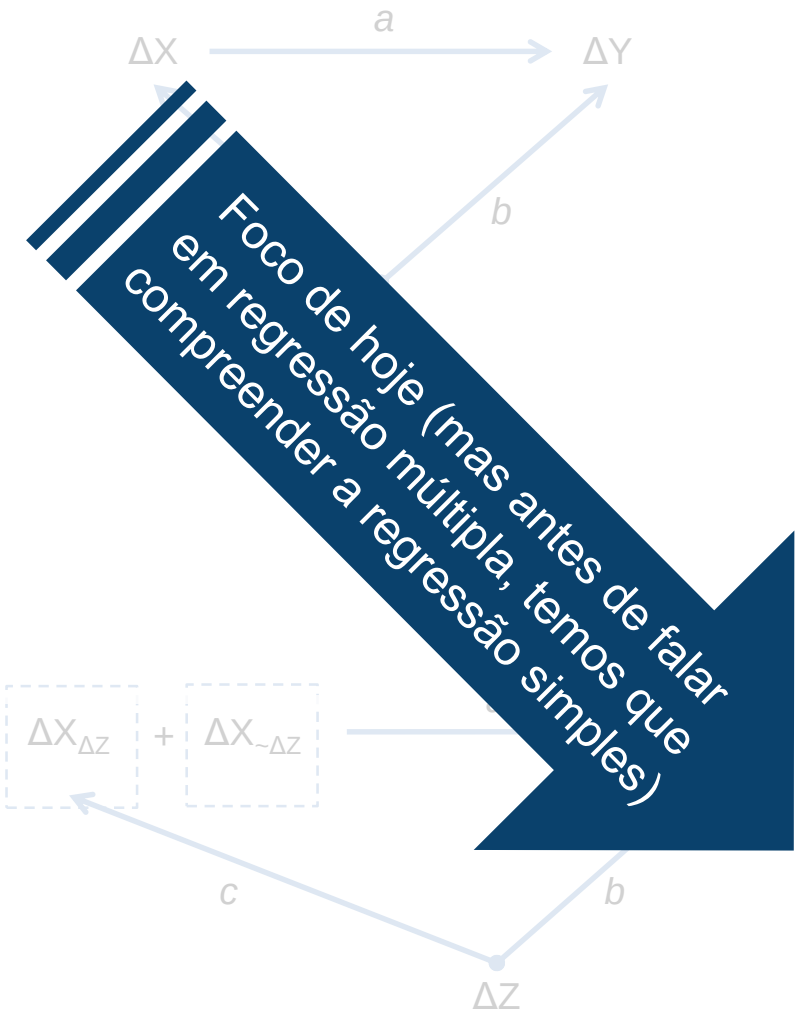
- Para que *a* fosse não espúria, para cada potencial fonte de espuriedade Z, as associações b e c não poderiam existir concomitante (ou seja, pelo menos uma delas teria de ser nula)
- **Atribuição aleatória ao tratamento** (desenho experimental) e **análise por estrato de Z eliminam variações de Z;** se Z não varia, não se associa com X ou Y na amostra
- Análise por estrato de Z é uma solução inferior à atribuição aleatória ao tratamento; estratificação deve ocorrer para todo potencial Z, e pode render estratos com poucas observações



- Se associação ***a'*** for não nula para todas as potenciais fontes de espuriedade Z, então a proposição de que variações em X causam variações em Y é válida.
- **O controle estatístico** (i.e., regressão múltipla em que se estima o efeito de X sobre Y controlando-se por Z) e o uso de **desenho quase experimental** (e.g., variável instrumental) são formas de se **identificar a associação *a'***
- Porém, ambas são soluções inferiores à atribuição aleatória ao tratamento

Ilustrando uma associação espúria: como evitar essa cilada?

a é associação espúria; ***a'*** pode não ser espúria



Associação *a* é espúria (i.e., não é uma medida válida da associação causal entre X e Y) porque reporta uma combinação de dois efeitos: efeito de X sobre Y e efeito de Z sobre Y.

- Para que *a* fosse não espúria, para cada potencial fonte de espuriedade Z, as associações b e c não poderiam existir concomitante (ou seja, pelo menos uma delas teria de ser nula)
- **Atribuição aleatória ao tratamento** (desenho experimental) e **análise por estrato de Z eliminam variações de Z;** se Z não varia, não se associa com X ou Y na amostra
- Análise por estrato de Z é uma solução inferior à atribuição aleatória ao tratamento; estratificação deve ocorrer para todo potencial Z, e pode render estratos com poucas observações

- Se associação *a'* for não nula para todas as potenciais fontes de espuriedade Z, então a proposição de que variações em X causam variações em Y é válida.
- **O controle estatístico** (i.e., regressão múltipla em que se estima o efeito de X sobre Y controlando-se por Z) e o uso de **desenho quase experimental** (e.g., variável instrumental) são formas de se **identificar a associação *a'***
- Porém, ambas são soluções inferiores à atribuição aleatória ao tratamento

Responde questão

Correlation is not going to cut it!

Questão	Abordagem analítica: Correlação	Abordagem analítica: Regressão
1 Há uma associação entre valores observados de X e Y?	✓	✓
2 Qual é a direção (sinal) dessa associação?	✓	✓
3 Qual é a magnitude (força) dessa associação?	✓	✓
4 Qual o valor estimado de Y para um dado X?		✓
5 Quanto Y varia quando X varia?		✓
6 Quanto Y varia quando X varia, **mantendo-se constantes as demais influências sobre Y**?		✓

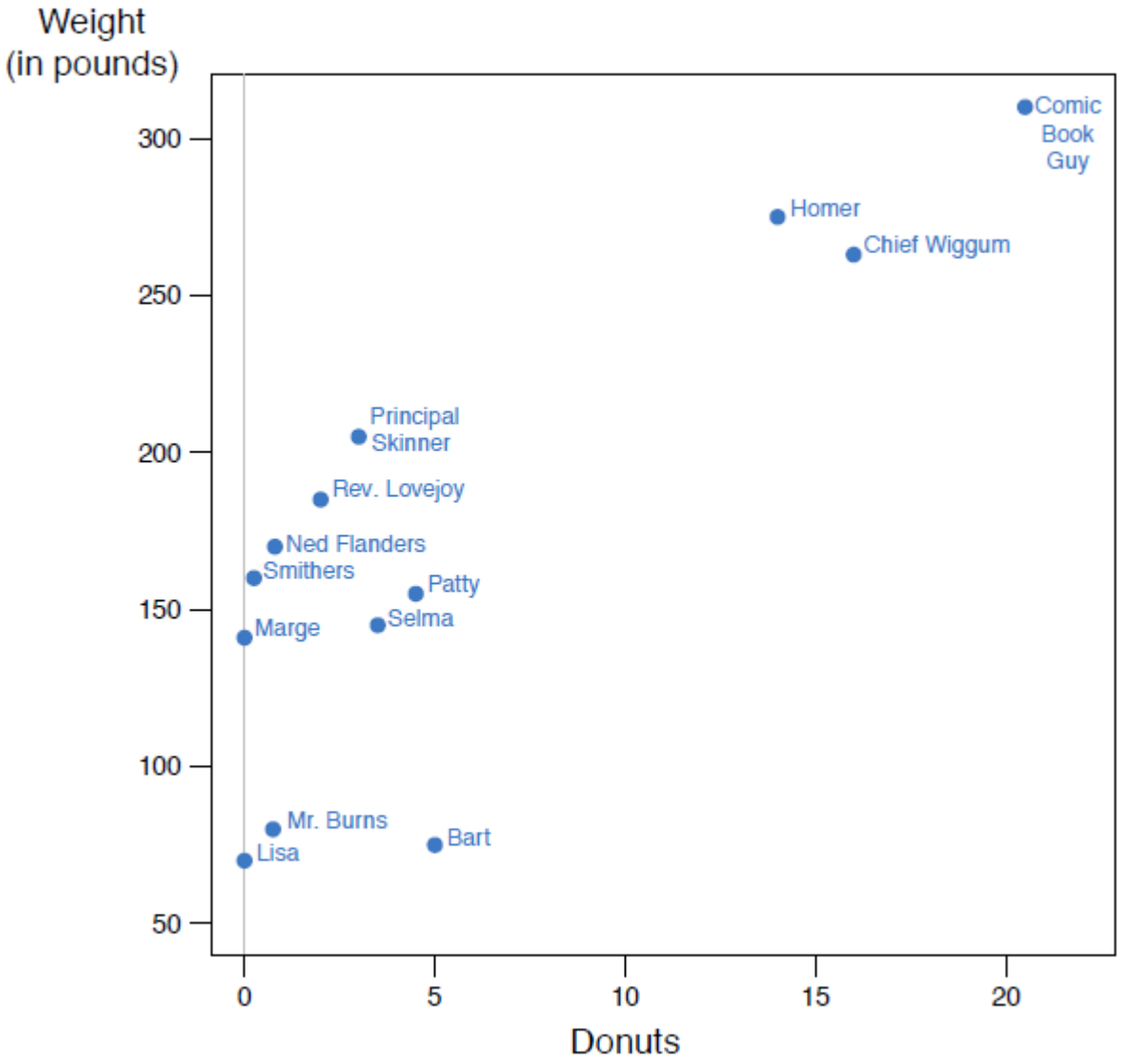
Have we met before?

Equação da reta

$$y = ax + b$$

Regressão é sobre identificar a linha que melhor descreva nossos dados. Joia!

Todavia, a realidade dos processos sociais não é muito “alinhada”.



Fonte: Bailey (2016: 6).

Linha reta não descreverá perfeitamente os dados.

Portanto, nosso modelo não preverá exatamente os valores de Y.

Nosso modelo linear de weight em função de donuts é uma simplificação da realidade.

Como incorporar ao modelo nossas incertezas sobre Y?

Modelo de regressão linear simples

- Também chamado de modelo de regressão linear de duas variáveis ou modelo de regressão linear bivariada

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

Esta é a equação de regressão **populacional**. Como ela se compara com a **equação da reta**? E com a equação de regressão **estimada**?

- Terminologia

Y	X
Variável dependente	Variável independente
Variável explicada	Variável explicativa
Variável prevista	Variável previsora (ou preditora)
Regressando	Regressor
Variável de resposta	
	Variável de interesse (foco da análise causal)
	Covariável (se regressão múltipla)
	Variável de controle (covariável que não é foco da análise causal)

Equação de regressão populacional

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

- A estimação de Y (via estimação de β_0 e β_1) é executada a partir de **dados empíricos amostrais**

- Todavia, o que se deseja estimar são os coeficientes mais gerais que descrevem a **relação entre X e Y no espaço teórico de todas as amostras possíveis**

- Em outras palavras, buscamos **generalizar** a linha de regressão **para além da amostra em questão**, pois estamos interessados no efeito de X sobre Y na população de interesse (e não no efeito particular observado na nossa amostra, especificamente)

- Para tanto, o modelo de regressão é fundado na existência de uma **linha de regressão "teórica" ou "populacional"**, que nunca será observada, a qual desejamos estimar a partir da nossa amostra

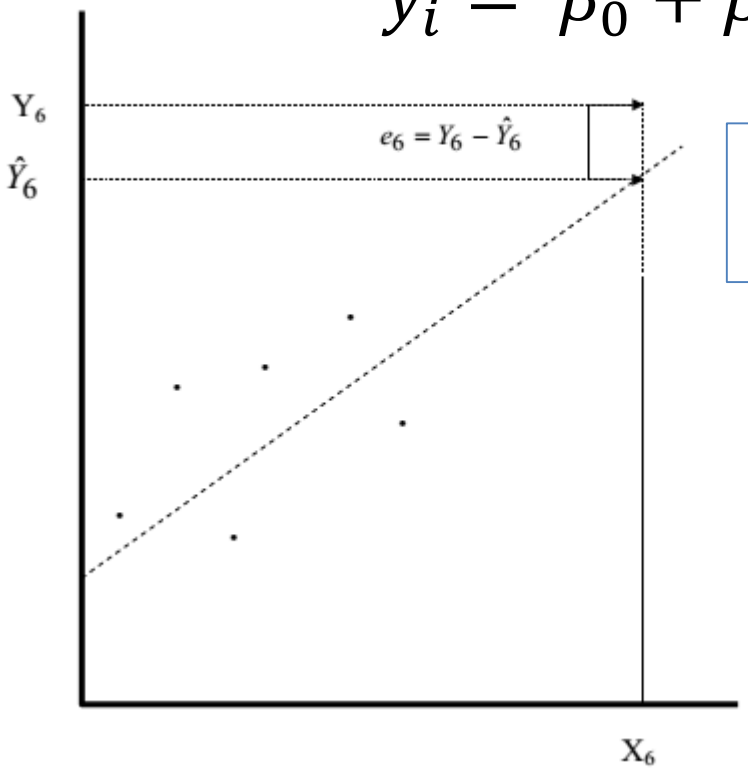
Função de regressão: populacional x estimada

- Função de regressão populacional:

$$y_i = \beta_0 + \beta_1 x_{1i} + \varepsilon_i$$

$$y_i = \underbrace{E(y \mid x_1)_i}_{\text{Parte sistemática}} + \underbrace{\varepsilon_i}_{\text{Parte estocástica}}$$

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$



Equação estimada não tem resíduos

- **Desafio:** não observamos os valores dos coeficientes populacionais; nós os estimamos a partir dos dados observados

- **Função de regressão estimada:**

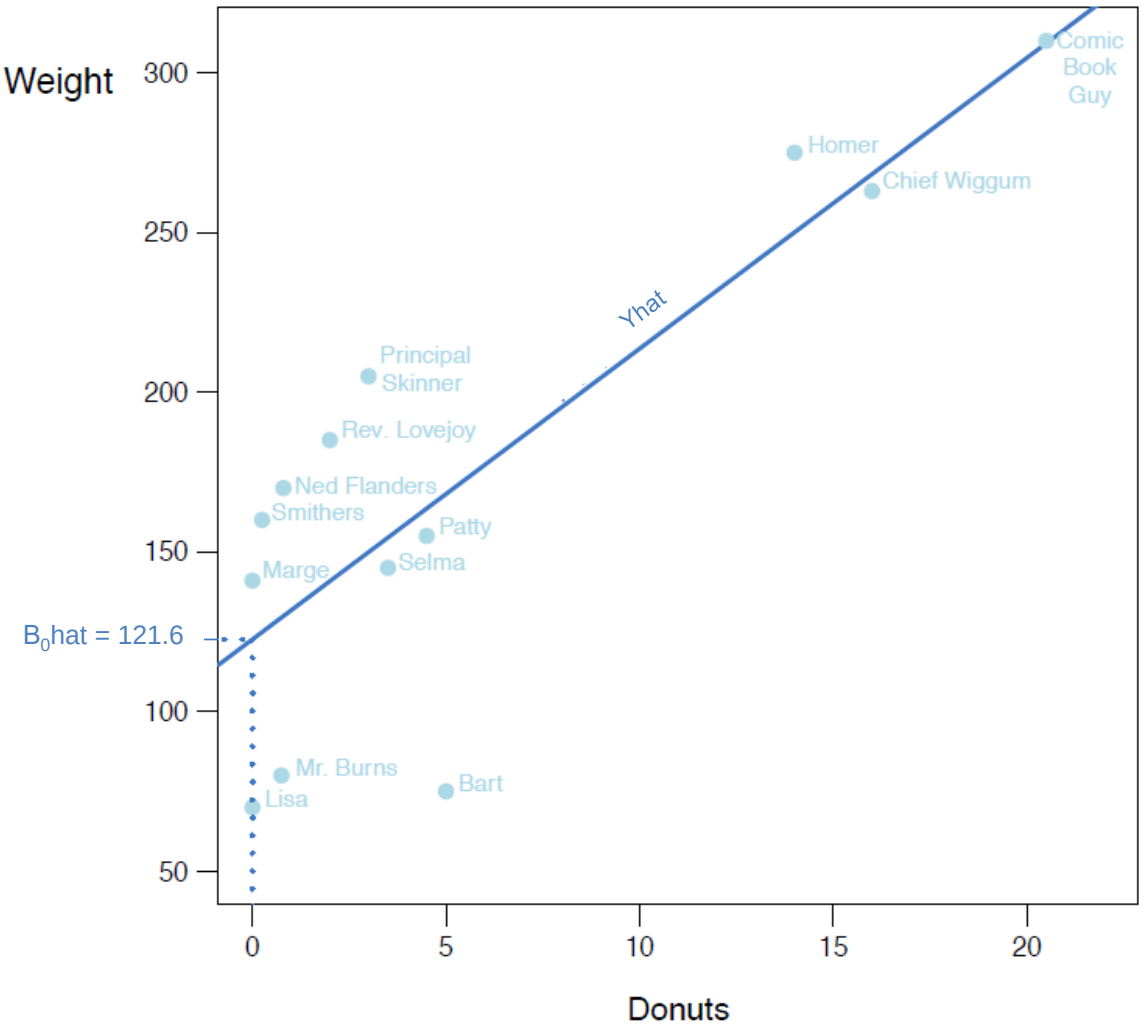
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + resíduo_i$$

$$y_i = \hat{y}_i + resíduo_i$$

$$resíduo_i = y_i - \hat{y}_i$$

1 libra = 0,454 quilograma
1 quilograma = 2,205 libras

Reta estimada



$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

$$\hat{y}_i = 121{,}6 + 9{,}2 x_{1i}$$

Voltaremos a esta regressão estimada

FIGURE 1.3: Regression Line for Weight and Donuts in Springfield
Fonte: Adaptado de Bailey (2016, p 9).

Significado de ε

$$Y = \beta_0 + \beta_1 X_1 + \boldsymbol{\varepsilon}$$

- Na análise de regressão simples, todos os fatores (além do X) que afetam Y são tratados como **não observados**
- **ε, chamado de erro aleatório ou termo estocástico**, representa todos esses fatores
- **O que o erro capta?** Everything we haven't accounted for in our model!
  - **Aleatoriedade intrínseca ao comportamento**. "Springfield residents are much too complicated for donuts to explain them completely (except, apparently, Comic Book Guy)." (Bailey, 2016: 10)
  - **Variáveis omitidas** (e.g., sex, height, other eating habits, exercise patterns, genetics)

Há fatores concretos em ε:
fatores omitidos que afetam sistematicamente o Y; neste sentido, a equação populacional que norteia a estimação é uma simplificação

Coeficientes: inclinação e intercepto

$$Y = \boldsymbol{\beta_0} + \boldsymbol{\beta_1} X_1 + \varepsilon$$

β_1

Tipicamente, estamos muito interessados em β_1, pois esse coeficiente caracteriza a relação entre X e Y (variação esperada em Y quando X aumenta em uma unidade).

β_0

Em geral, não damos muita atenção ao β_0. Apesar de esse coeficiente ser importante para ajustar a reta de regressão, normalmente não é o foco da pesquisa determinar o valor de Y quando X = 0.

Se β_0 estiver ausente, assume-se que $\beta_0 = 0$ e que, portanto, a reta de regressão atravessa a origem

Significado do intercepto (β_0)

- O intercepto **muitas vezes não tem significado real** porque é o valor previsto da variável dependente quando todas as variáveis independentes na regressão assumem o valor 0

  – **Frequentemente, esse é um cenário que não faz sentido**, porque cai fora do intervalo de dados aceitáveis – por exemplo, na equação que prevê **salário como uma função da idade**, não temos indivíduos para quem idade = 0

- O intercepto faz um trabalho de **"coleta de lixo"**. O efeito médio de todas as **variáveis omitidas** no modelo é atribuído ao erro (ε). Como, por definição (premissa), o **valor esperado do erro é 0**, qualquer desvio em relação a esse valor é forçado na estimação da constante (e/ou dos parâmetros de inclinação, no caso de viés de variável omitida, como veremos adiante)

Sempre adicione β_0 a sua regressão. O trabalho de "coleta de lixo" é necessário.

Falando de premissas...
Premissas de MQO quanto ao erro
(apenas aquelas que nos interessam neste momento)

1. Valor esperado do erro é zero:

$$\mathbf{E(\varepsilon) = 0}$$

2. X e ε não são sistematicamente relacionados (isto é, o valor esperado do erro não depende de X):

Premissa da média condicional zero

$$\mathbf{E(\varepsilon|X) = 0}$$

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

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).

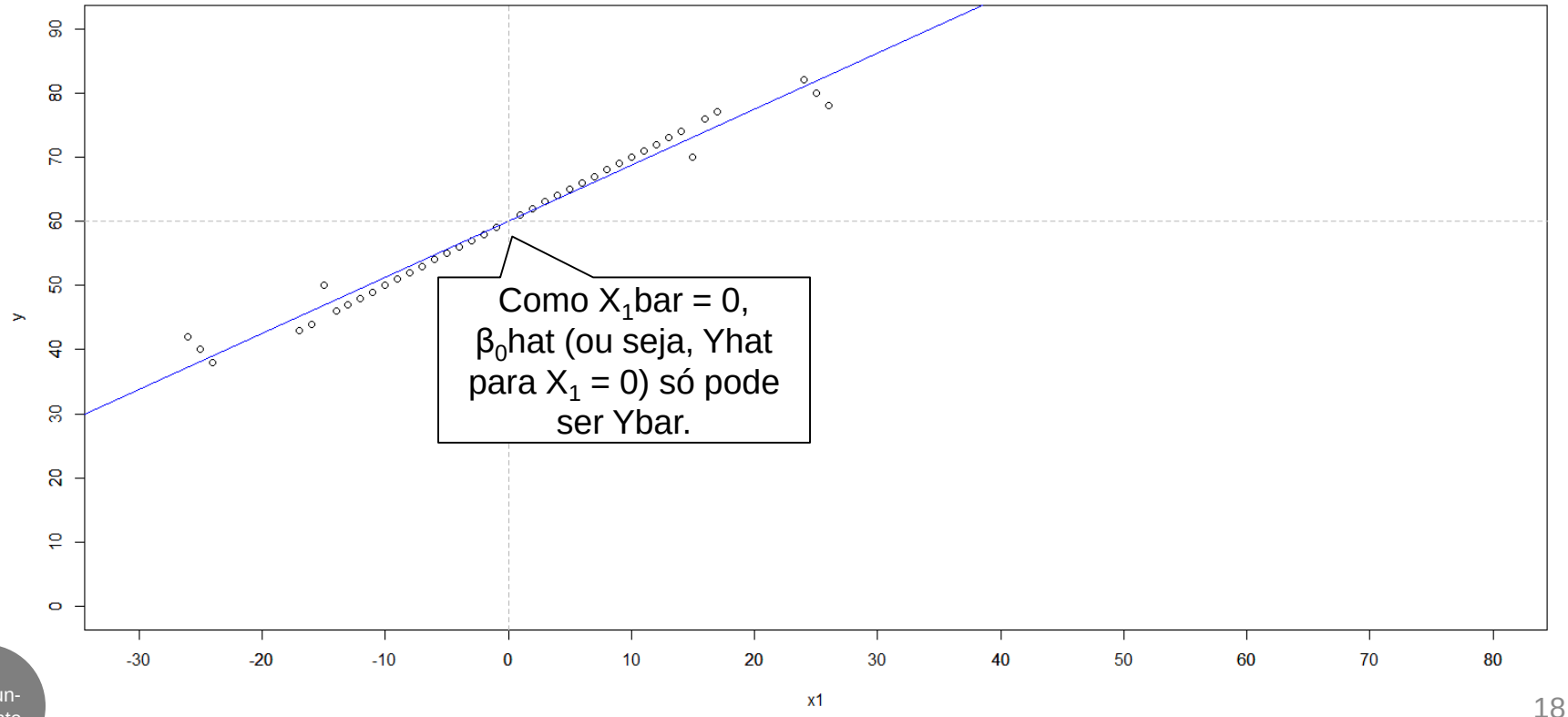
Na regressão simples, a reta estimada passa, necessariamente, por (Xbar, Ybar)

```
> 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

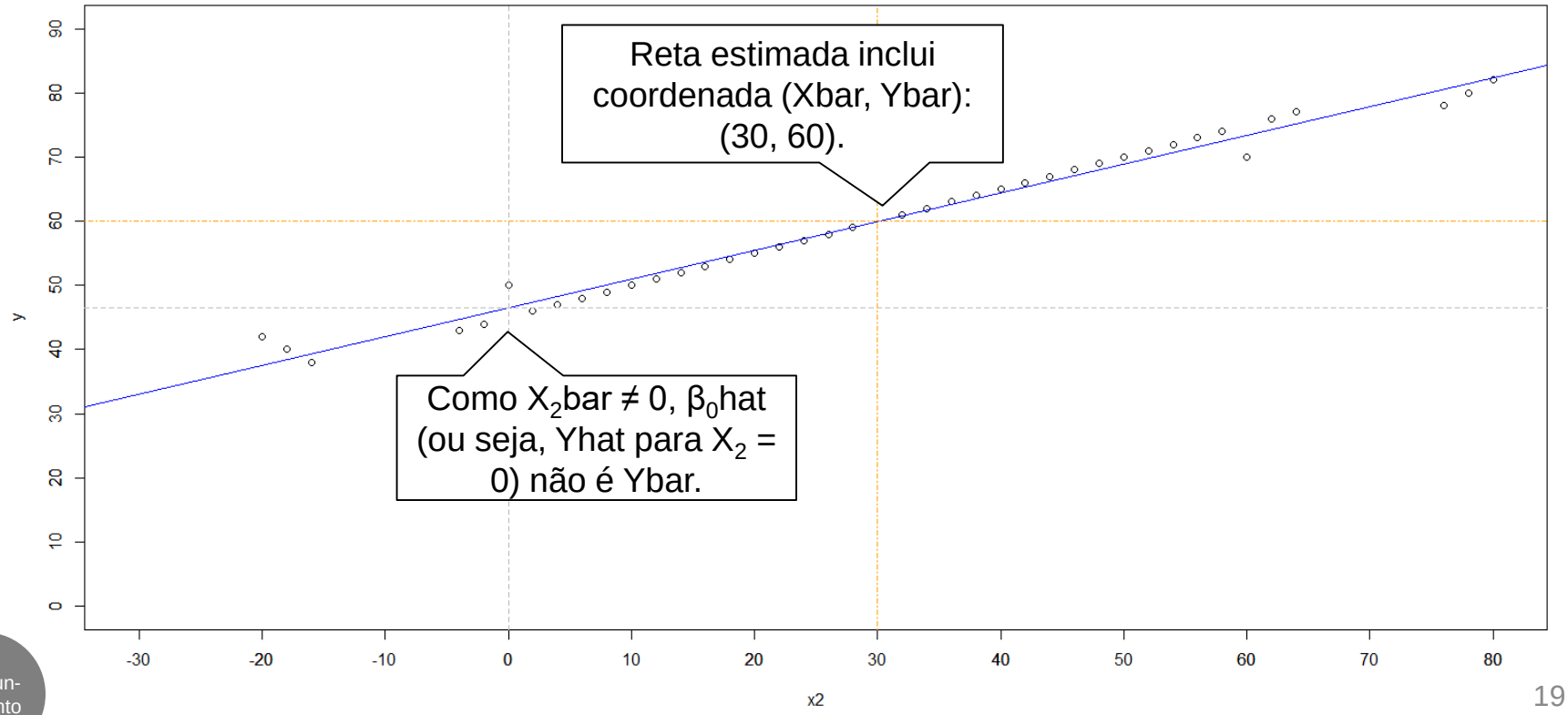
Na regressão simples, a reta estimada passa, necessariamente, por (Xbar, Ybar)

```
> 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

Ilustração:
Comando para estimar regressão em R

```
Call:
lm(formula = dados$Weight..pounds. ~ dados$Donuts.per.week)

Residuals:
    Min      1Q  Median      3Q     Max
-92.731 -13.508   3.916  36.081  55.716

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            121.613     16.593   7.329 1.49e-05 ***
dados$Donuts.per.week    9.224      1.959   4.707 0.000643 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 45.81 on 11 degrees of freedom
Multiple R-squared:  0.6683,  Adjusted R-squared:  0.6381
F-statistic: 22.16 on 1 and 11 DF,  p-value: 0.0006426
```

Ilustração:
Comando para estimar regressão em R

```
Call:
lm(formula = dados$Weight..kilograms ~ dados$Donuts.per.week)

Residuals:
    Min      1Q  Median      3Q     Max
-42.062  -6.127   1.776  16.366  25.272

Coefficients:
                      Estimate Std. Error t value Pr(>|t|)
(Intercept)            55.1629     7.5265   7.329 1.49e-05 ***
dados$Donuts.per.week   4.1837     0.8888   4.707 0.000643 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 20.78 on 11 degrees of freedom
Multiple R-squared:  0.6683,  Adjusted R-squared:  0.6381
F-statistic: 22.16 on 1 and 11 DF,  p-value: 0.0006426
```

Interpretação dos coeficientes estimados

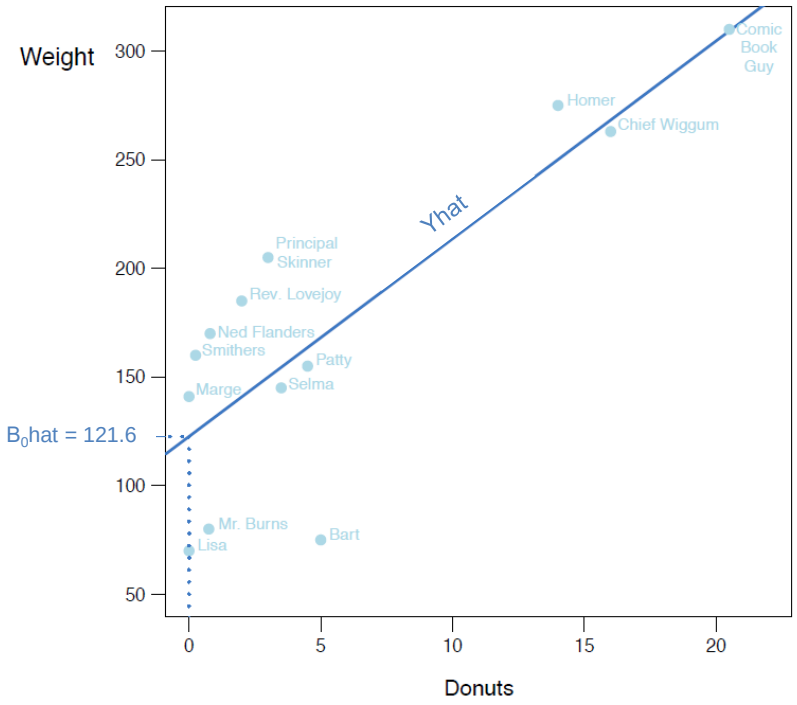


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

Fonte: Adaptado de Bailey (2016, p 9).

1 libra = 0,454 quilograma
1 quilograma = 2,205 libras

- Estimação da linha de regressão fornece os seguintes resultados:

$$\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i}$$

$$\hat{y}_i = 121{,}6 + 9{,}2 x_{1i}$$

onde Y = peso em libras, X_1 = donuts por semana, e i é o subscrito que indexa indivíduos

- **Intercepto (constante):** Estima-se que indivíduos que não consomem donuts pesem 121,6 pounds, em média

- **Coeficiente de inclinação:** Estima-se que o consumo de um donut adicional por semana esteja associado a um aumento 9,2 pounds no peso

ceteris paribus?

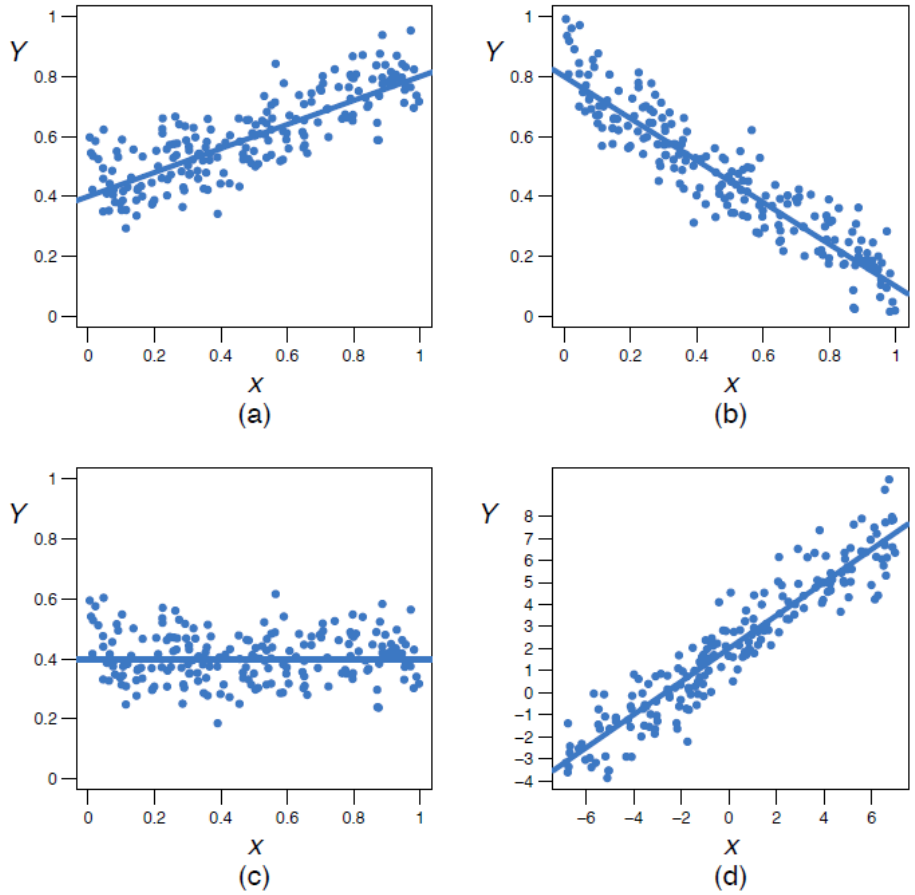


FIGURE 1.4: Examples of Lines Generated by Core Statistical Model

Discussion Questions

For each of the panels in Figure 1.4, determine whether β_0 and β_1 are greater than, equal to, or less than zero. (Be careful with β_0 in panel (d)!)

Fonte: Bailey (2016: 12).

Como são calculados os coeficientes de regressão?

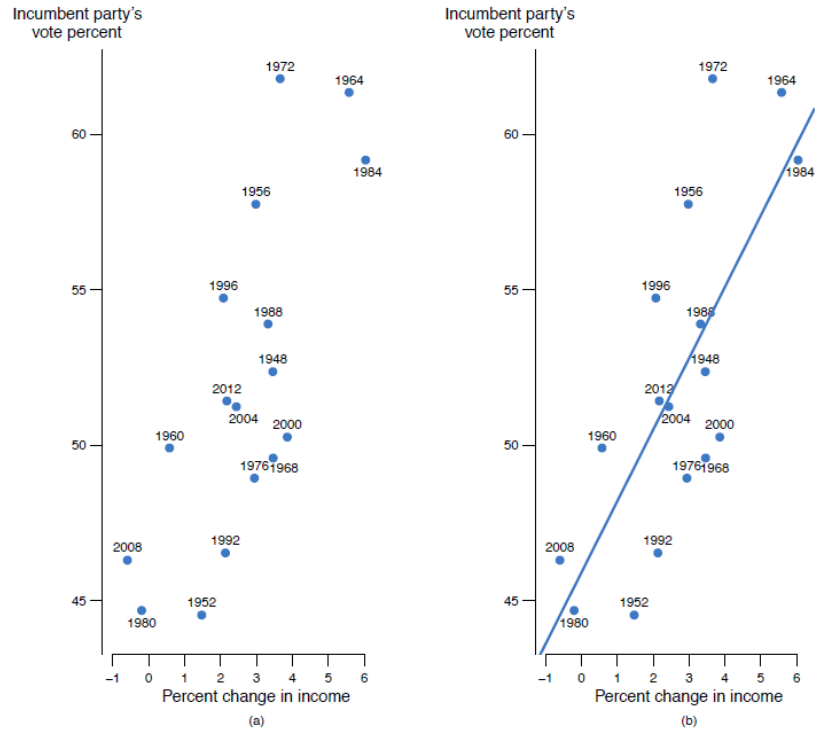


FIGURE 3.1: Relationship Between Income Growth and Vote for the Incumbent President's Party, 1948-

Fonte: Bailey (2016: 66).

Nota: The income change variable refers to the last year of incumbent party's term.

- A partir de dados como os do painel (a), estimamos a **linha que melhor caracteriza a relação** entre as duas variáveis

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

$$\text{Incumbent party vote share}_i = \beta_0 + \beta_1 \text{Income change}_i + \varepsilon_i$$

ε_i = all other factors affecting elections (e.g., wars, scandals)

- Um **algoritmo (regra de cálculo)** poderoso para estimação dos parâmetros (β_0 e β_1) é **Mínimos Quadrados Ordinários (MQO, OLS em inglês)**

Como MQO "encontra" a linha que melhor descreve os dados?

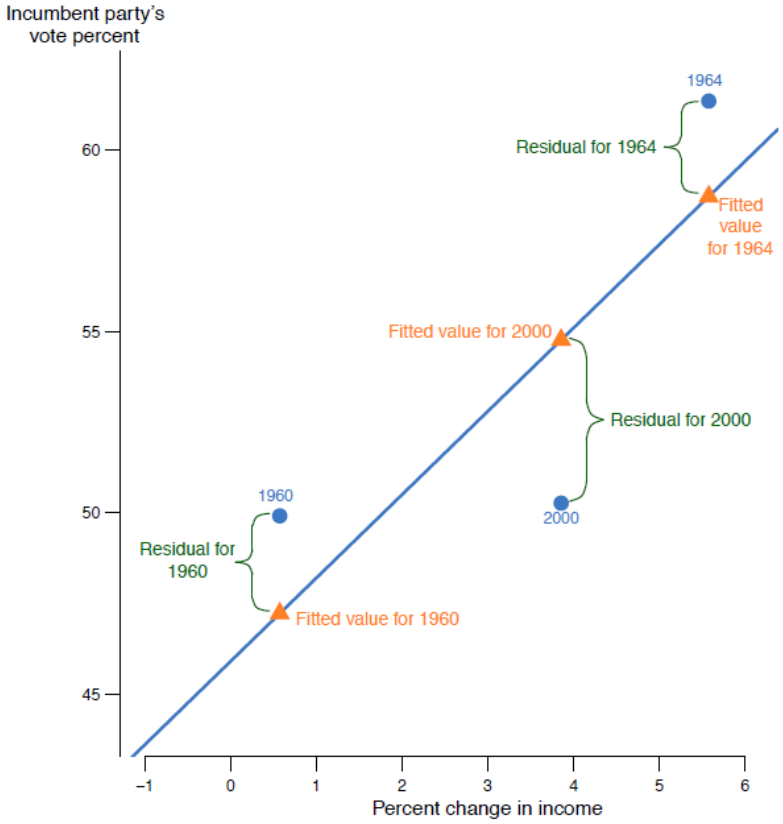


FIGURE 3.3: Fitted Values and Residuals for Observations in Table 3.1

Fonte: Bailey (2016: 76).

- MQO identifica a **linha que minimiza a soma da distância entre cada valor observado de Y_i e a linha** (linha informa o valor estimado de Y_i, aka $Y_i$hat ou **fitted value**)

- Essa **distância** corresponde ao **resíduo**; o resíduo é a **manifestação empírica do erro aleatório "verdadeiro"** (o ε_i do modelo populacional):

$$resíduo_i = y_i - \left(\hat{\beta}_0 + \hat{\beta}_1 x_{1i}\right)$$

Há várias representações de resíduo, entre elas:

$$e_i,\ \hat{e}_i,\ r_i,\ \hat{\varepsilon}_i,\ (y_i - \hat{y}_i)$$

- Especificamente, MQO minimiza a seguinte expressão, em que $\hat{e}_i$ = resíduo:

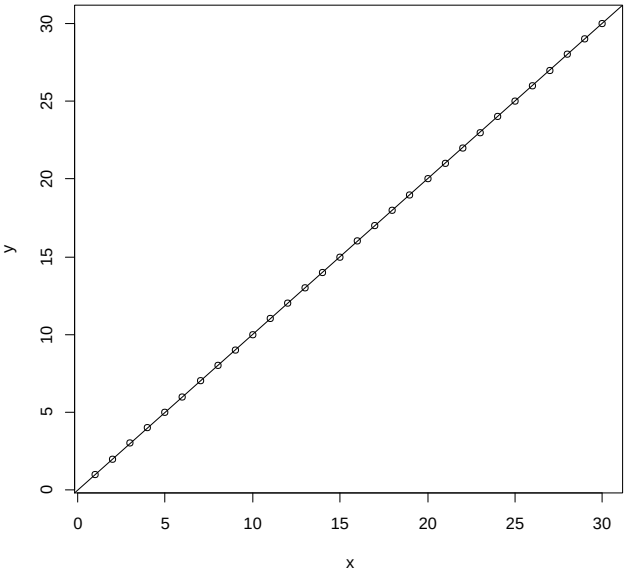
$$\sum_{i=1}^{n} \hat{e}_i^2 = \sum \left(y_i - \hat{\beta}_0 - \hat{\beta}_1 x_{1i}\right)^2$$

Especificamente, MQO identifica a linha que minimiza a soma do quadrado dos resíduos

- Valores elevados ao quadrado são sempre positivos

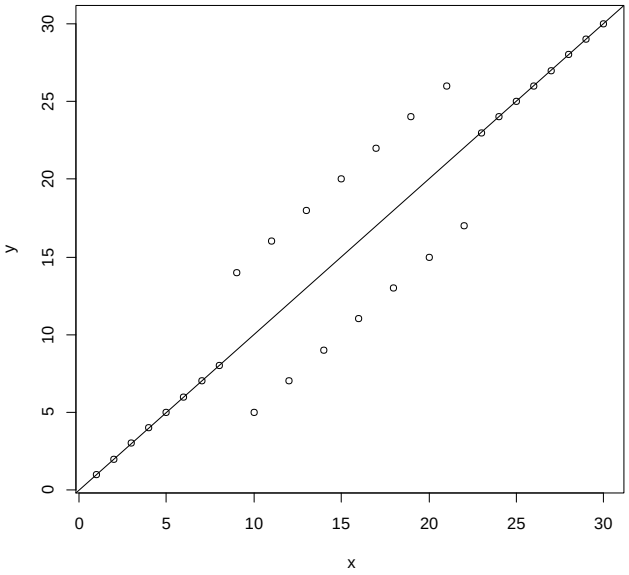
- Resíduos são elevados ao quadrado para que erros de estimação “para cima” e “para baixo” não se anulem

Cenário 1



Soma dos resíduos = 0
Soma dos resíduos^2 = 0

Cenário 2



Soma dos resíduos = 0
Soma dos resíduos^2 = 350

- Outra possibilidade seria minimizar a média do módulo do resíduo (i.e., o desvio médio), mas essa ideia não se convencionou (Gorard, S. Revisiting a 90-year-old debate: the advantages of the mean deviation, *British Journal of Educational Studies*, v. 53, n. 4, December 2005, p. 417-430. https://doi.org/10.1111/j.1467-8527.2005.00304.x

Minimização da soma do quadrado dos resíduos

- Queremos identificar a combinação de coeficientes estimados (β_0 e β_1) que minimiza a soma dos quadrados dos resíduos; aqui, usaremos a letra *a* para designar o β_0, a letra *b* para designar o β_1, *e* a letra e para designar o resíduo:

$$Z = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} [Y_i - (a + bX_i)]^2$$

- No ponto em que a soma do quadrado dos resíduos (Z) for mínima, a derivada parcial de Z em relação a *a* e a derivada parcial de Z em relação a *b* serão nulas; assim, a minimização passa por obter essas derivadas e igualá-las a zero:

$$\frac{\partial Z}{\partial a} = -2 \sum [Y_i - (a + bX_i)] = 0$$

$$\frac{\partial Z}{\partial b} = 2 \sum [Y_i - (a + bX_i)](-X_i) = 0$$

- Os valores de *a* e de *b* que minimizam Z (ou seja, que fazem ambas as derivadas nulas) atendem ao seguinte sistema de equações normais:

$$\begin{cases} na + b \sum X_i = \sum Y_i \\ a \sum X_i + b \sum X_i^2 = \sum X_i Y_i \end{cases}$$

- Na prática, determinamos *b* em primeiro lugar:

$$b = \frac{n \sum XY - (\sum X)(\sum Y)}{n \sum X^2 - (\sum X)^2} = \frac{\sum (X - \overline{X})(Y - \overline{Y})}{\sum (X - \overline{X})^2}$$

$$a = \frac{\sum Y}{n} - b \frac{\sum X}{n} = \overline{Y} - b\overline{X}$$

Fonte: Hoffmann (2016)

Aprofun-damento

Essa minimização produz equações para as estimativas dos coeficientes de inclinação e intercepto

Coeficiente: inclinação

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Fonte: https://weeklyblabber.wordpress.com/2018/01/12/deja-vu-evoking-memories/

- **Numerador** é chamado de **soma dos produtos cruzados** (mesmo numerador do coeficiente de correlação de Pearson)
- Conceitualmente, numerador é uma medida de quanto os valores dos **pares ordenados (x_i, y_i)** são **associados**
- Denominador **"padroniza" o numerador**, fazendo com que $\beta_1$hat seja a variação em Y esperada quando X varia em **uma unidade**

Essa minimização produz equações para as estimativas dos coeficientes de inclinação e intercepto

Coeficiente: Intercepto

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

- A estimativa do intercepto corresponde à **diferença entre a média de Y (i.e., $\bar{y}$), de um lado, e $\beta_1$hat vezes a média de X (i.e., $\bar{x}$), de outro**
- Esta fórmula é obtida a partir da **solução de um sistema de equações com incógnitas $\beta_0$hat e $\beta_1$hat** (detalhes no slide de aprofundamento “Minimização da soma dos quadrados dos resíduos”)
- Na regressão simples, essa fórmula implica que a **reta estimada passará**, necessariamente, **pela coordenada ($\bar{x}$, $\bar{y}$)**.

Alerta:
Não confunda $\beta_0$hat com a média amostral de Y para X = 0

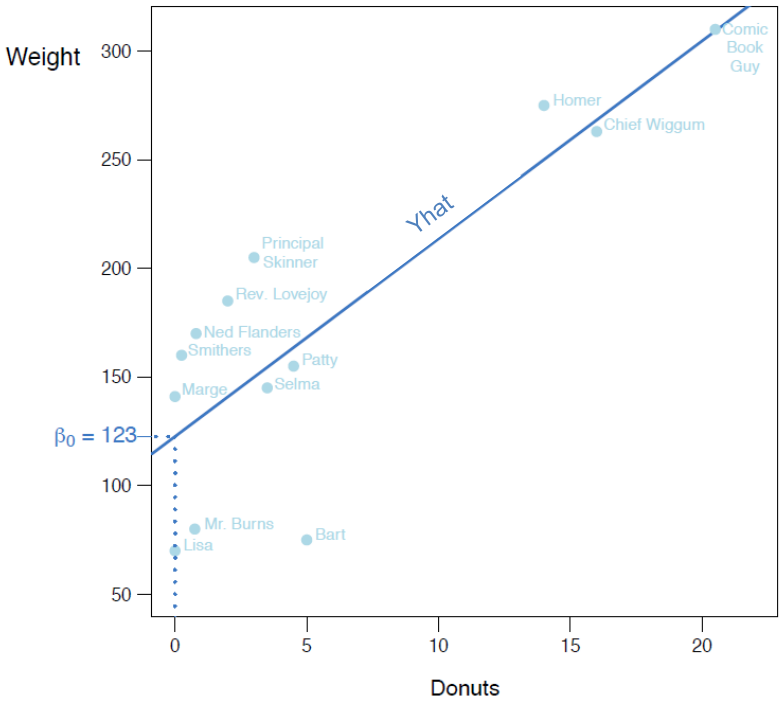


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

Fonte: Adaptado de Bailey (2016, p. 9).

- $\beta_0$hat é o valor estimado do peso para indivíduos que não consumiram donuts na última semana: 121,6 libras

- Note que $\beta_0$hat não é o peso médio dos indivíduos que não consumiram donuts na última semana (apenas Marge e Lisa):
  Ybar = 171,8
  (Ybar|X=0) = 105,5
  (Ybar|X>0) = 183,9

- Modelo sobrestima o peso médio das pessoas que não consumiram donuts, pois há poucos casos nessa condição; no processo de "encaixe" da reta, o padrão prevalecente é aquele das pessoas que consumiram donut

Aprofun-damento

Voltando ao exemplo: Bivariate OLS and presidential elections

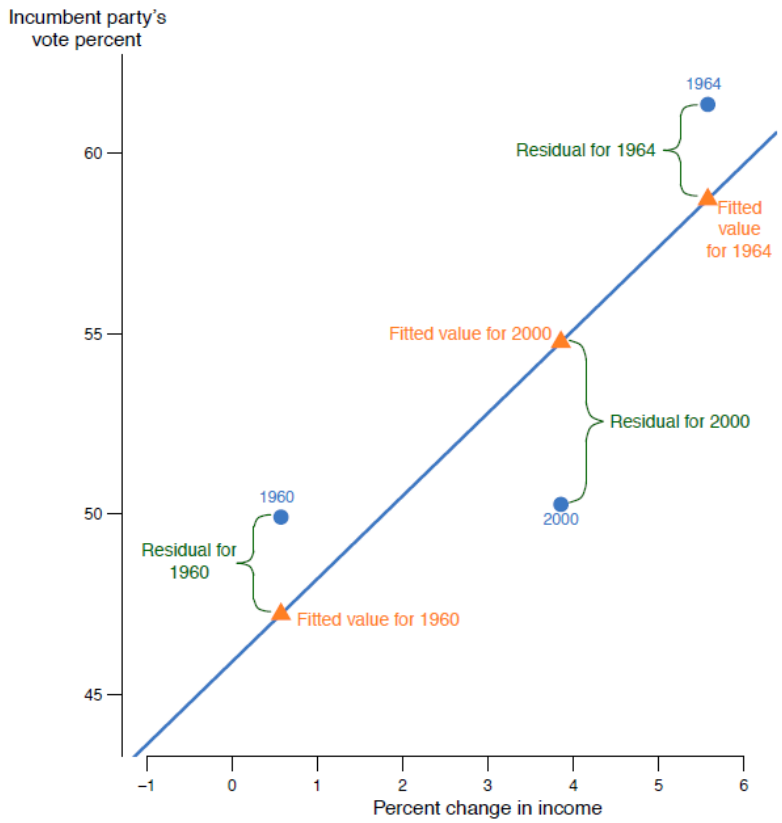


FIGURE 3.3: Fitted Values and Residuals for Observations in Table 3.1

Fonte: Bailey (2016: 76).

$$\widehat{Incumbent\ party\ vote\ share}_i = \hat{\beta}_0 + \hat{\beta}_1 Income\ change_i$$

$$= 45.9 + 2.3 \times Income\ change_i$$

Table 3.1: Selected Observations from Election and Income Data

Year	Income change (X)	Incumbent party vote share (Y)	Fitted value ($\hat{Y}$)	Residual ($\hat{\epsilon}$)
1960	0.58	49.9	47.2	2.7
1964	5.58	61.3	58.7	2.6
2000	3.85	50.2	54.7	-4.5

Fonte: Bailey (2016: 77).

Como interpretar coeficientes estimados, valores previstos (Yhat) e resíduos?

Estima-se que o percentual de votos recebidos pelo partido do incumbente no cenário de estagnação econômica (i.e., variação percentual da renda = 0) seja 45,9%.

Estima-se que o aumento de um ponto percentual na variação percentual da renda (e.g., passar de um crescimento de 3% para 4%) esteja associado com um aumento de 2,3 pontos percentuais no percentual de votos recebidos pelo partido do incumbente.

Sobre a diferença entre ponto percentual e variação percentual, vide: https://educacao.uol.com.br/disciplinas/matematica/ponto-percentual-nao-confunda-com-porcentagem.htm

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 ΔX ➔ ΔY, 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 é 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, ao mesmo tempo, podem estar 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

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

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

Como calcular os parâmetros?

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

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

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

$$\hat{\beta}_1 = \frac{(\sum yx_1)(\sum x_2^2) - (\sum yx_2)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

$$\hat{\beta}_2 = \frac{(\sum yx_2)(\sum x_1^2) - (\sum yx_1)(\sum x_1x_2)}{(\sum x_1^2)(\sum x_2^2) - (\sum x_1x_2)^2}$$

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

Melhor não calcular isto a mão ☺

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

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 endógena se for correlacionada com 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)

- Em comparação com a simples aplicação de regressão múltipla em dados observacionais, **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

Voltaremos a este ponto

Endogeneidade é uma preocupação constante em nossas vidas

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

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

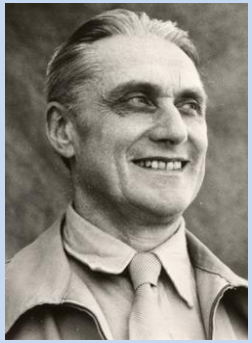
*[Ao interpretar estimativas de regressão multivariada] O que eles [estatísticos] realmente querem dizer é mais como **“descontar”** o efeito de outras variáveis no modelo. A lógica por trás de dizer que outros fatores são constantes é que, **uma vez que tenhamos calculado os efeitos de outras variáveis,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

…

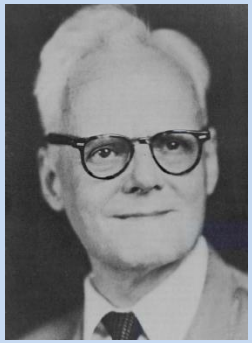
*Portanto, quando alguém diz algo como “**mantendo X_2 constante**, o efeito estimado de um aumento de uma unidade em X_1 é $\beta_1$hat”, o que se quer dizer é que, **considerando o efeito de X_2, estima-se que o efeito de X_1 seja $\beta_1$hat**.*

Bailey (2016: 197-198)

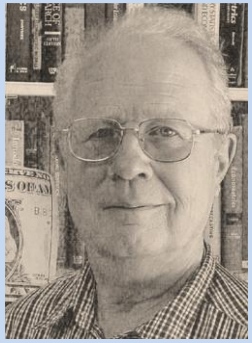
Teorema de Frisch-Waugh-Lovell (FWL)



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  
```

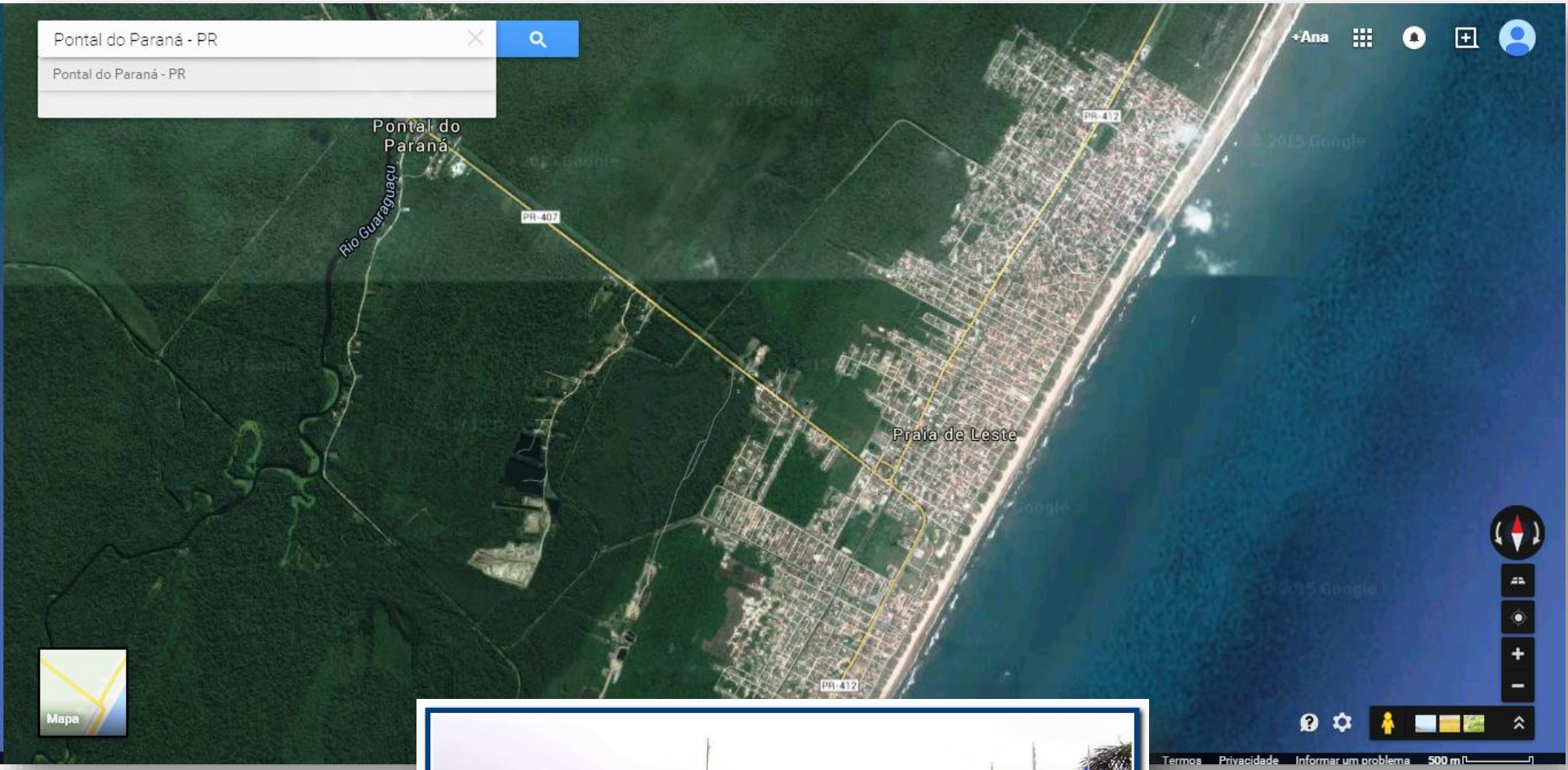
Ilustrando o Teorema de Frisch-Waugh-Lovell (FWL)

	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

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) —*a*→ ΔY (mortes por afogamento)

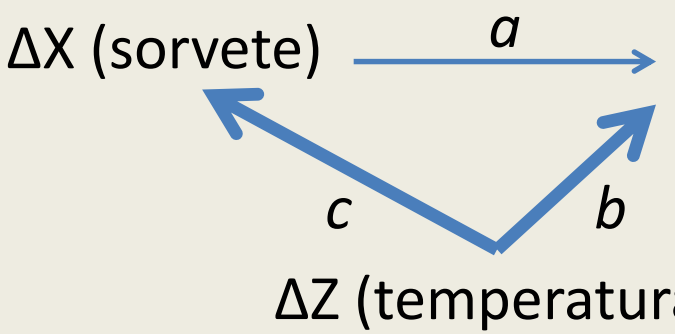
Como checar se a proposição causal *a* é verdadeira?

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:

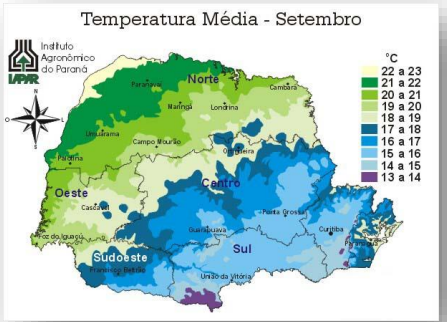
$$\Delta X \text{ (sorvete)} \xrightarrow{\alpha'} \Delta Y \text{ (mortes por afogamento)}$$

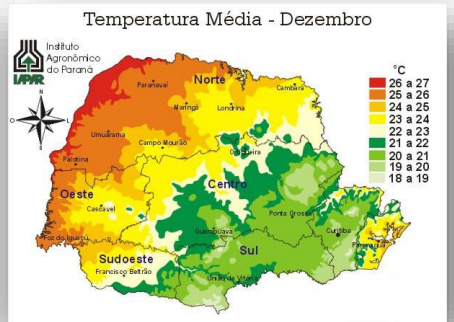
$$\Delta Z \text{ (temperatura)} = 0$$

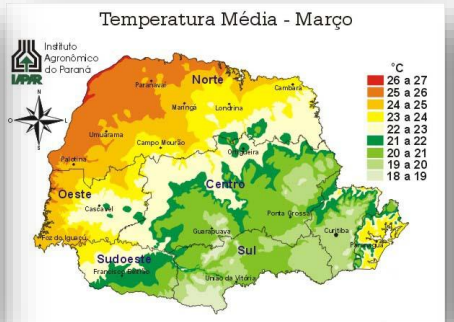
(i.e., sempre quente ou sempre frio)

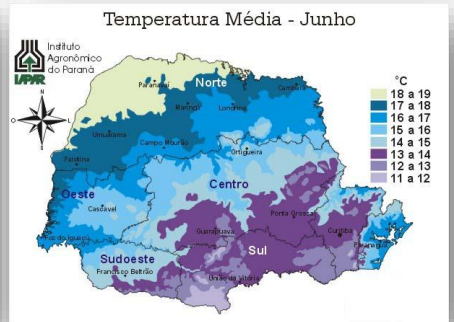
Proposição causal *α'* só é verdadeira se existir correlação não nula entre X e Y num cenário de temperatura constante.

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.

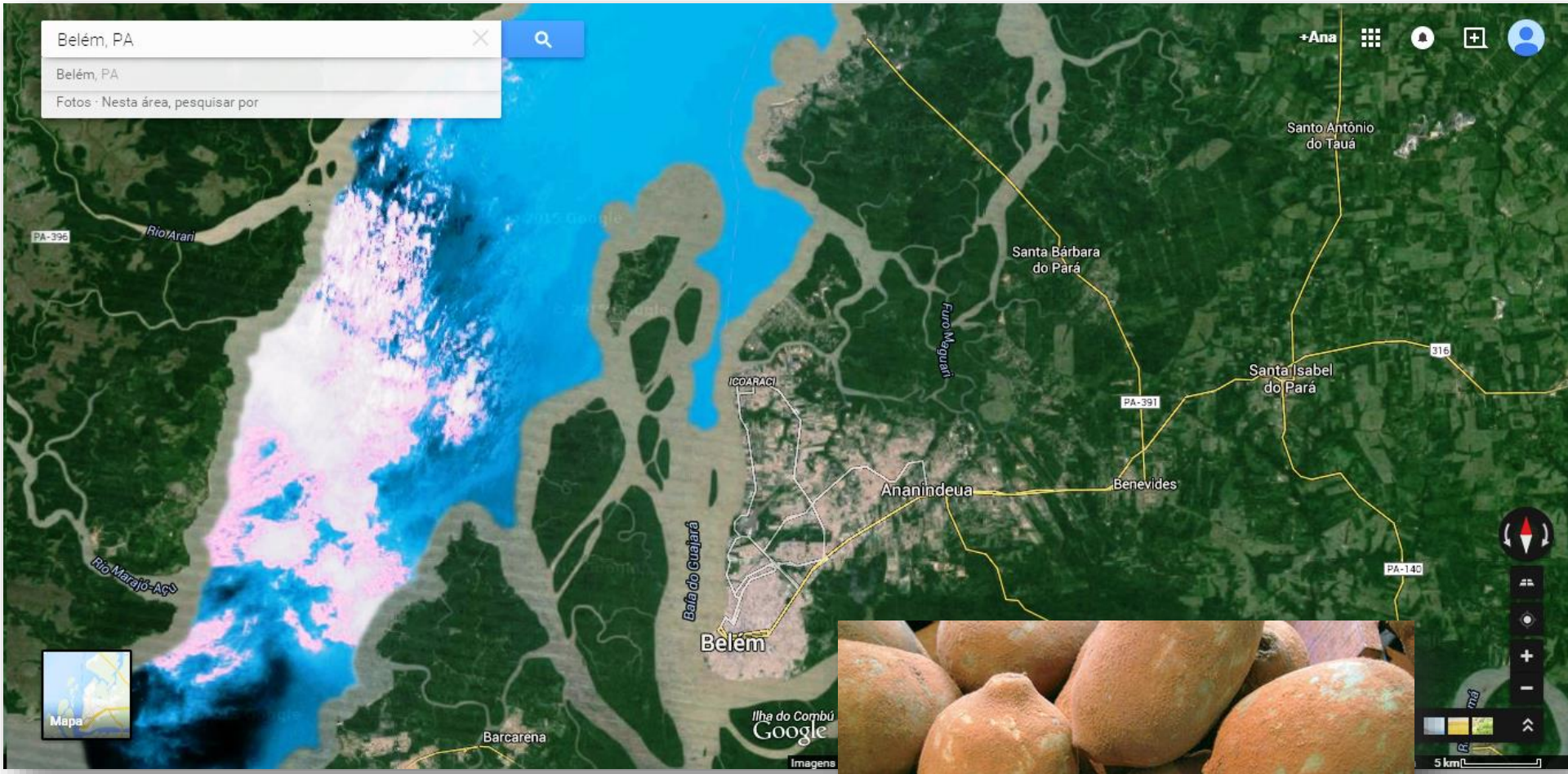








“Cidades das Mangueiras” e do cupuaçu, Belém é sempre quente



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

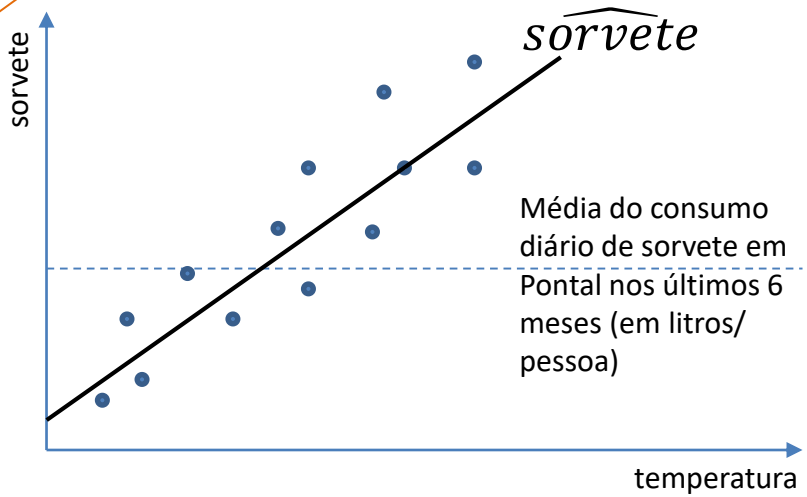
- É plausível imaginar que **além da temperatura, outros fatores** possam influenciar consumo de sorvete, p.ex.: lançamentos, mudança na preferência dos consumidores, disponibilidade de outras sobremesas

- 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!



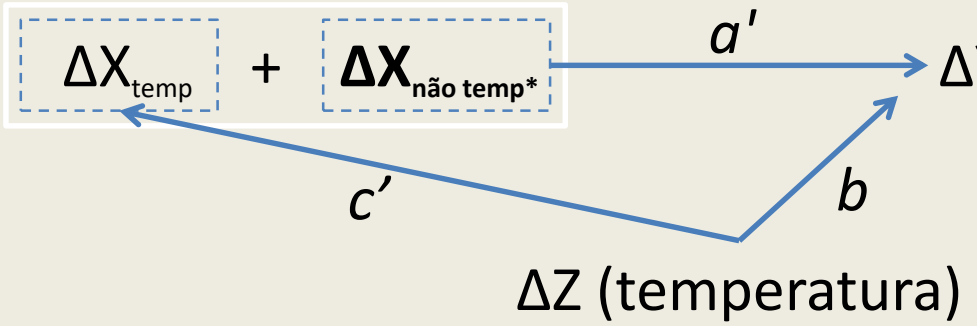
$$\widehat{sorvete} = \widehat{\beta_0} + \widehat{\beta_1} temperatura$$

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?



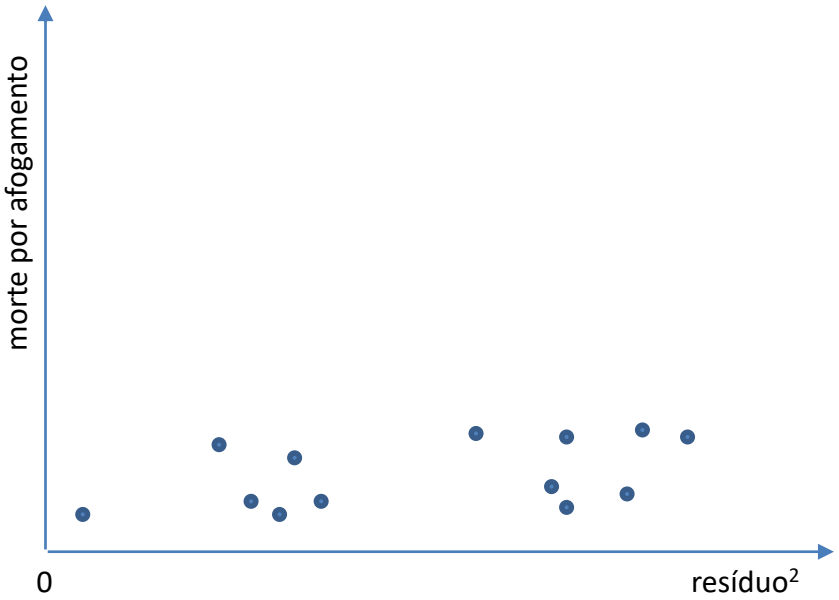
Proposição causal *a'* só é verdadeira **se o resíduo da regressão** de sorvete em temperatura **explicar** a variação de **mortes por afogamento.**

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

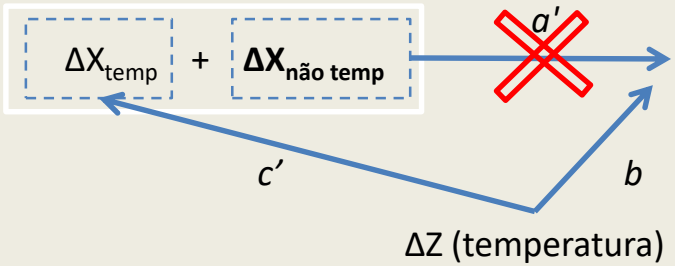
Critério	Atendido?	Evidência
Anterioridade	✓	▪ Consumo de sorvete está concentrado na sobremesa do almoço, enquanto afogamentos concentram-se no final da tarde
Correlação	?	Vamos trabalhar com 2 cenários para a correlação entre o resíduo de sorvete e as mortes por afogamento
Não-espuriedade	?	

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

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



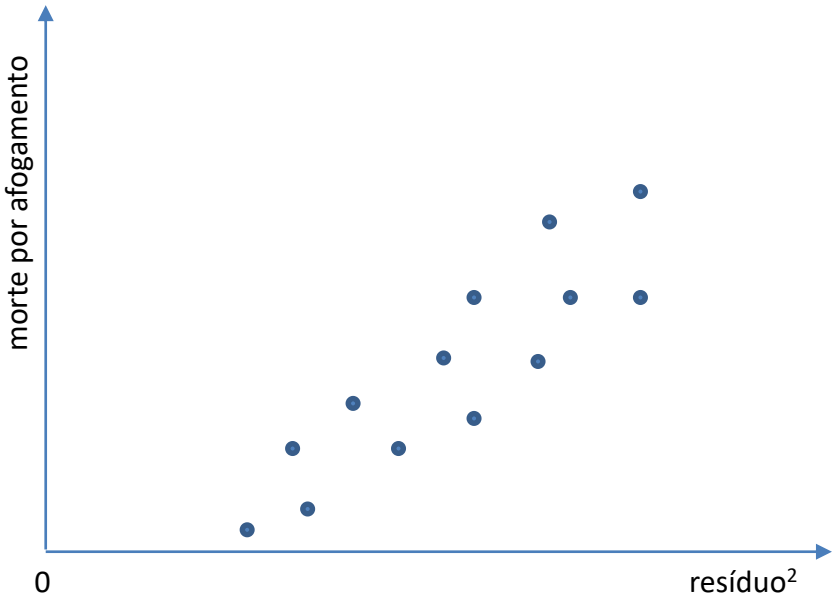
Critério	Atendido?
Anterioridade	✓
Correlação	✗
Não-espuriedade	?



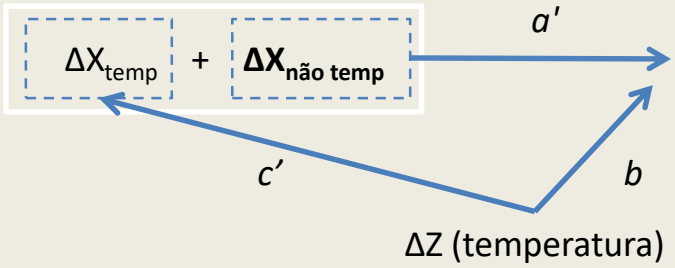
Achados e implicações práticas:

- 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 ***α'* é relação (parcial) entre ΔX e ΔY mantendo a temperatura "constante" ou "controlando" pela temperatura**

É para isto que serve a regressão múltipla!

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

Regressão bivariada
Noções de regressão multivariada

Aula 2
14 de setembro de 2022

Ana Paula Karruz