

Controle estatístico

Avaliação de Políticas Públicas B

12, 24 e 26 de março de 2025

Leitura básica:

CHEIN, Luciana. **Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas**. Brasília: Escola Nacional de Administração Pública, 2019, pp. 09-45, Disponível em: <http://repositorio.enap.gov.br/handle/1/4788>

Leitura complementar:

Demais obras disponibilizadas na pasta “Referências sobre regressão linear”, disponível em https://drive.google.com/drive/folders/1MndWD_X3Vg9hVGv7-QlyGtYExgp2XAua?usp=sharing


Como podemos separar o efeito de X sobre Y do efeito de Z sobre Y?










- **Correlação** envolve apenas duas variáveis (não dá conta de considerar Z na análise)
- **Controle estatístico (i.e., regressão múltipla)** pode acomodar numerosas variáveis na análise (dependendo do tamanho da amostra), mas não é a estratégia mais potente para se reduzir o risco de espuriedade
 - ➡ Não conhecemos a lista completa de variáveis do tipo Z que podem ser fonte de espuriedade, e frequentemente não temos dados para representar essas variáveis
- Pesquisas com **desenho experimental e quase experimental**, se pudermos garantir o atendimento de certas premissas, são mais potentes em afastar espuriedades que o controle estatístico

- **Frequentemente, as avaliações de impacto combinam desenhos experimentais ou quase experimentais com controle estatístico**

Nosso foco até o final
do Módulo I.

Controle estatístico (via análise de regressão) responde questões que correlação não dá conta de responder

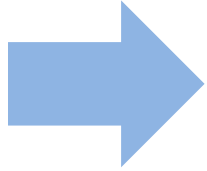
 Responde questão

Questão	Correlação	Controle estatístico
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?		
<i>Have we met before?</i>		
6 Quanto Y varia quando X varia, mantendo-se constantes as demais influências sobre Y?		

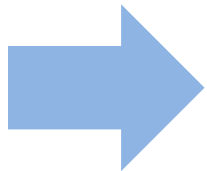
Equação da reta

$$y = ax + b$$

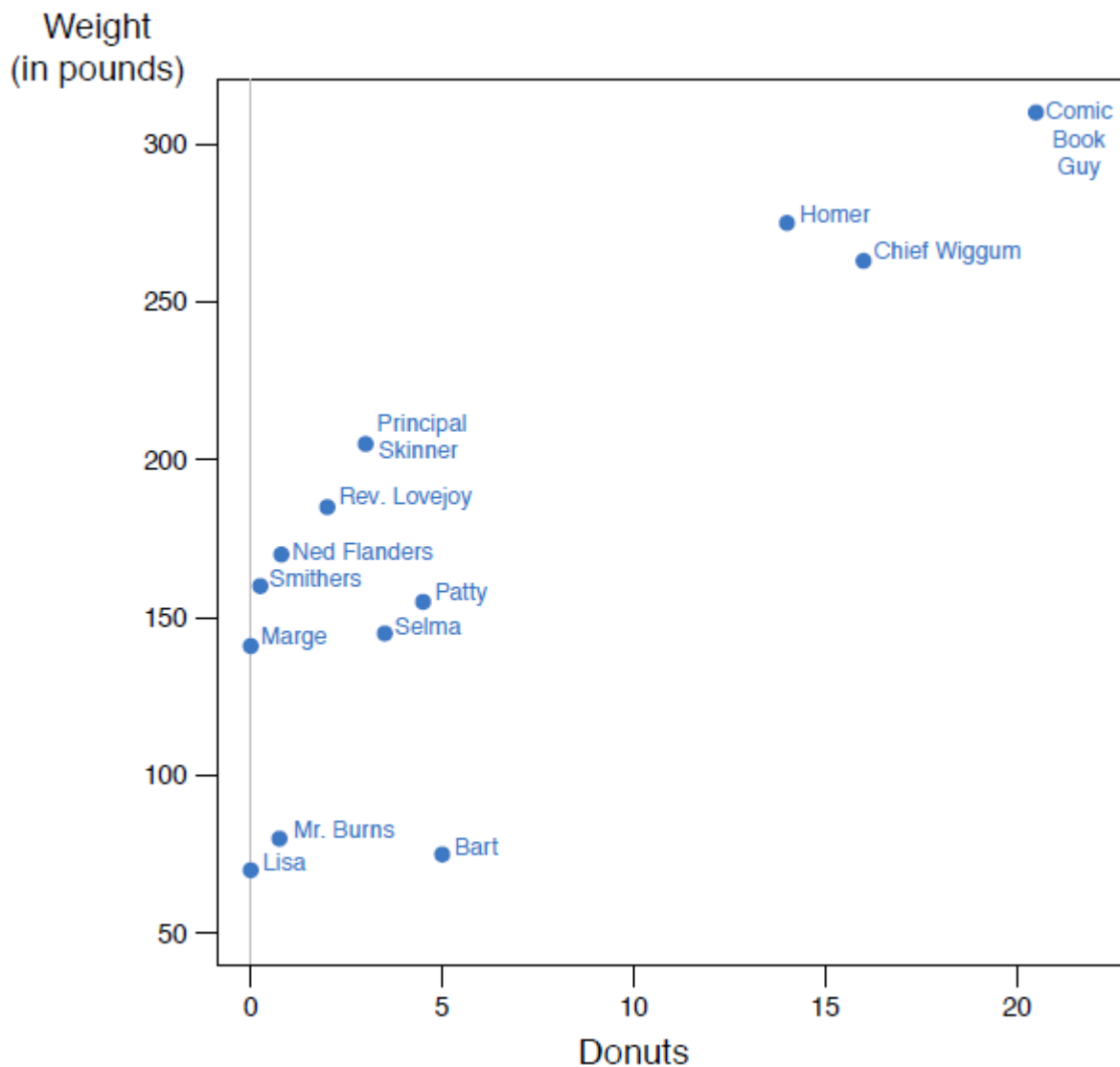




Regressão é sobre identificar a linha que melhor descreve nossos dados. Joia!



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



→ 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?

Fonte: Bailey (2016, p. 6).

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)
- 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:

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

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

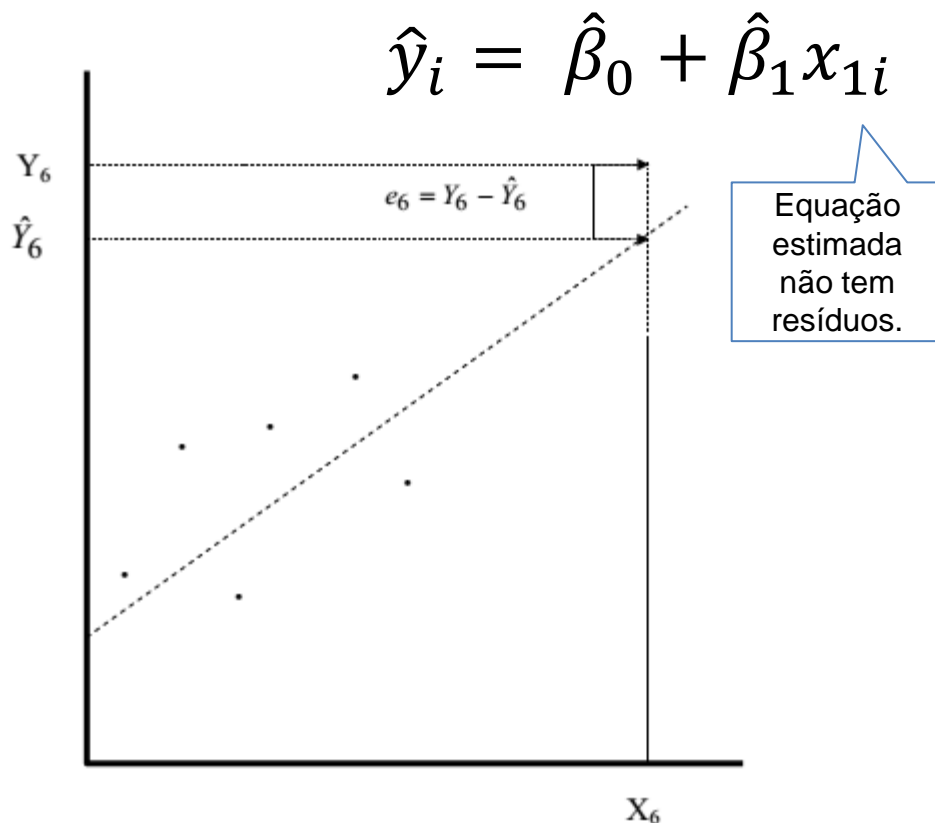
- Desafio:** não observamos os valores dos coeficientes populacionais; nós os estimamos a partir dos dados observados

- Função de regressão estimada:

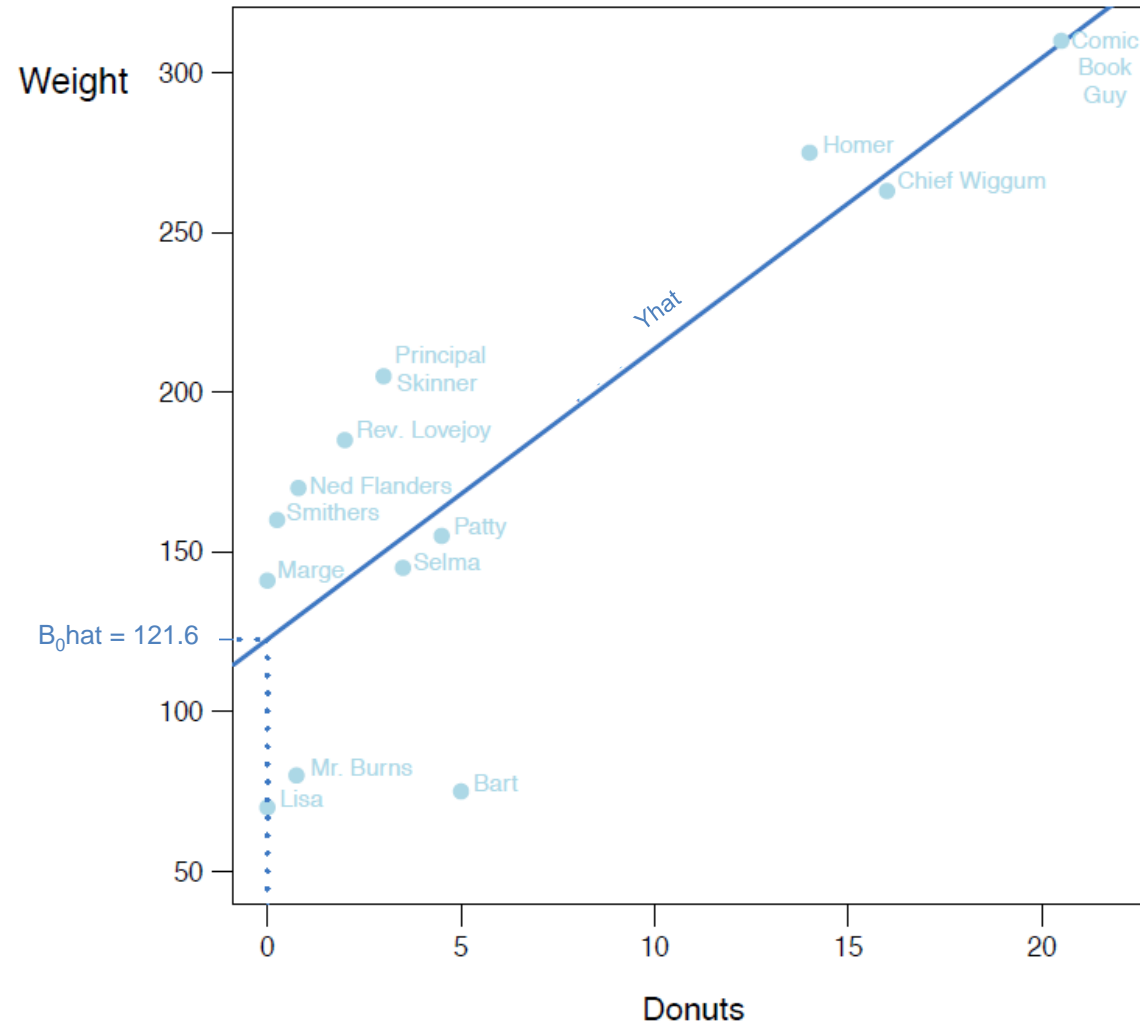
$$y_i = \hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \text{resíduo}_i$$

$$y_i = \hat{y}_i + \text{resíduo}_i$$

$$\text{resíduo}_i = y_i - \hat{y}_i$$



Reta estimada



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

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

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

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

- 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?** Tudo aquilo que não incluímos no nosso modelo!
 - **Aleatoriedade intrínseca ao comportamento.** “Os residentes de Springfield são complicados demais para que apenas o consumo de donuts possa explicá-los completamente (exceto, aparentemente, o Comic Book Guy).” (Bailey, 2016, p. 10)
 - **Variáveis omitidas** (e.g., sexo, altura, outros hábitos alimentares, prática de exercícios físicos, genética)

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

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

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.

Interpretação dos coeficientes estimados

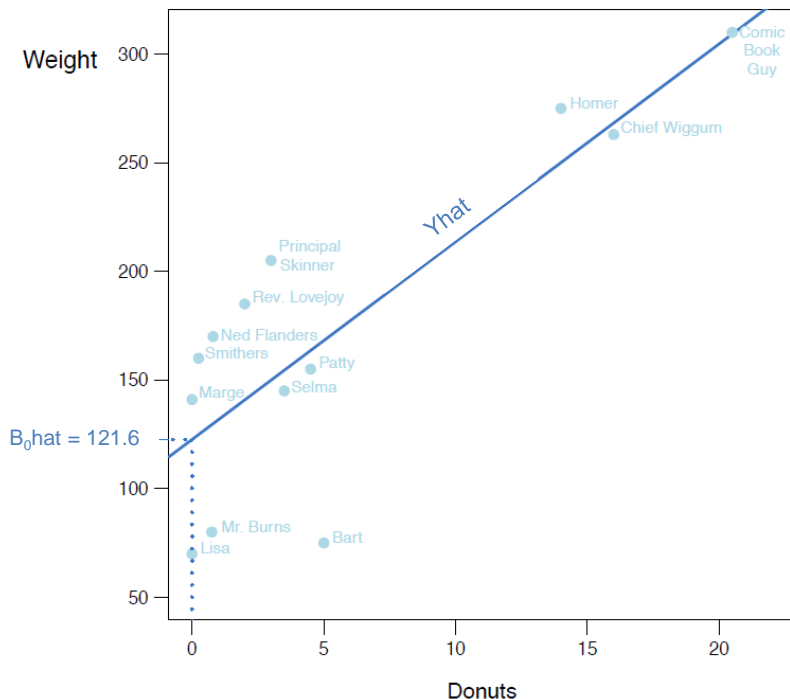


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

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

Ceteris paribus? Regressões são modelos causais, mas nem sempre oferecem estimativas causais válidas.

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

- Estimação da linha de regressão fornece os seguintes resultados:

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

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

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

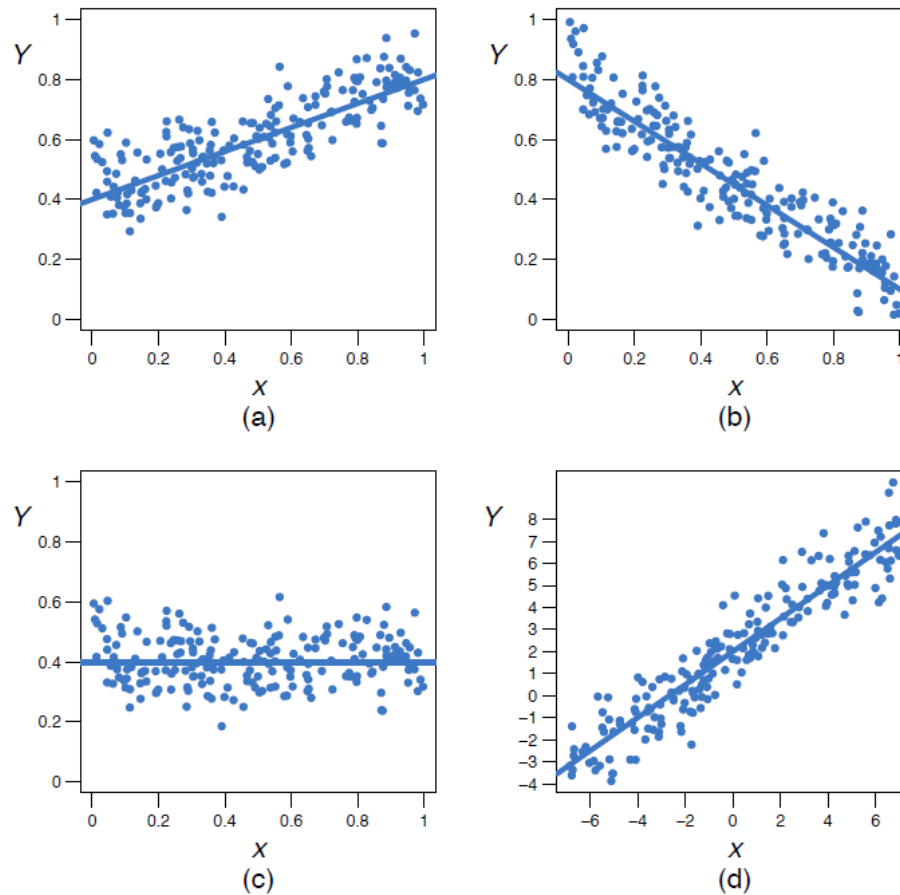
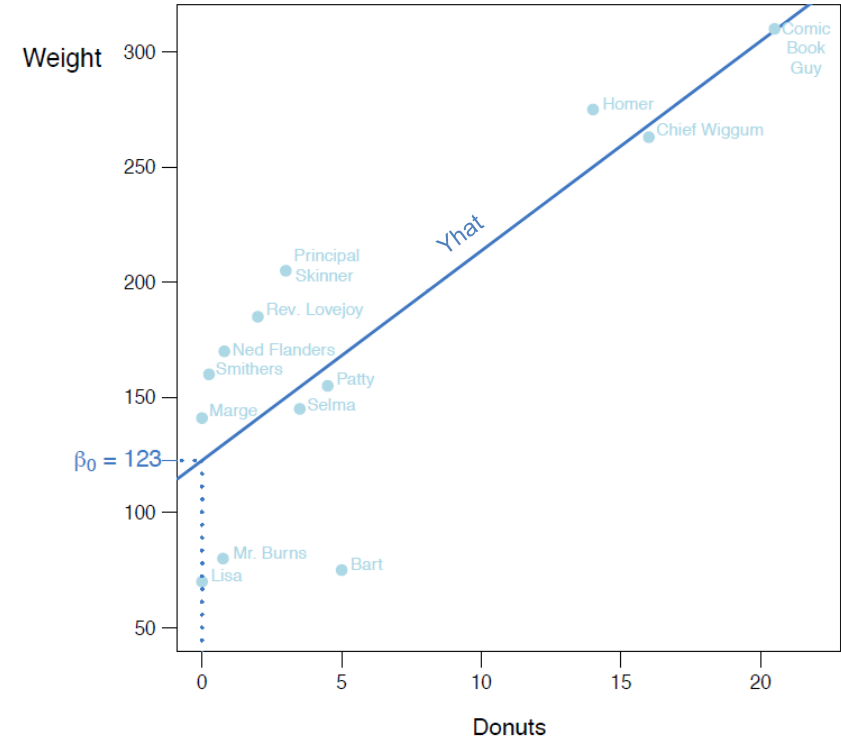
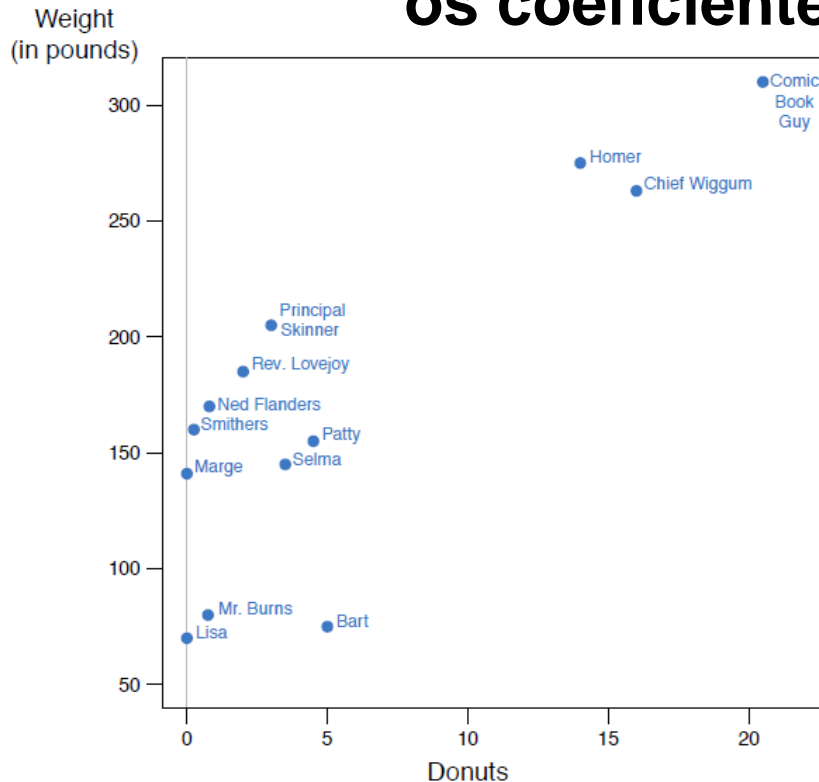


FIGURE 1.4: Examples of Lines Generated by Core Statistical Model

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)!)

Como são calculados os coeficientes de regressão?



Fonte: Adaptado de Bailey (2016, p. 6, 9).

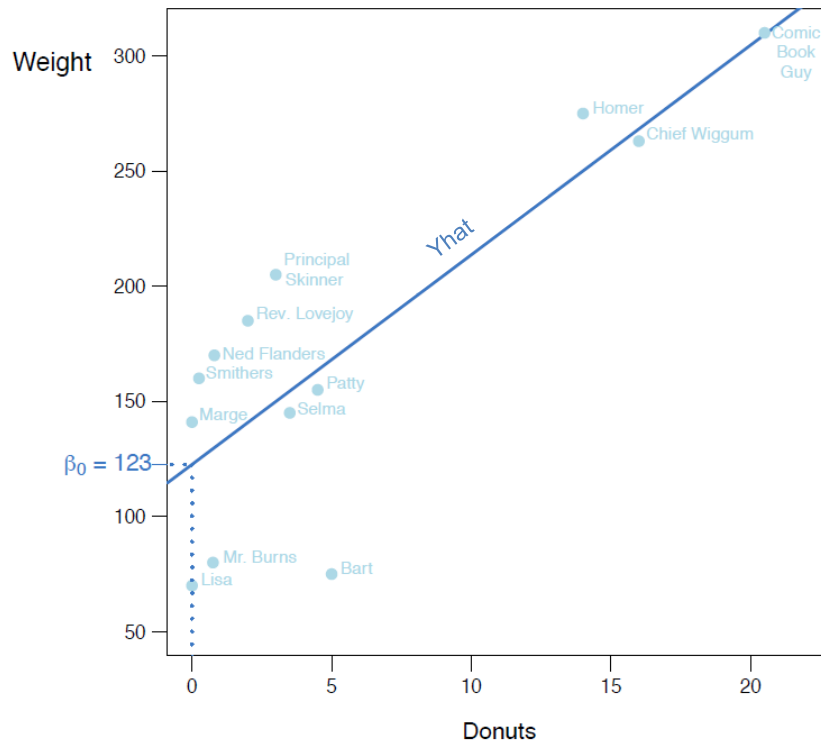
- A partir de dados como os do painel à esquerda, estimamos a **linha que melhor caracteriza a relação** entre as duas variáveis

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$
$$\text{Weight}_i = \beta_0 + \beta_1 \text{Donuts per week}_i + \varepsilon_i$$

ε_i = todos os outros fatores que afetam Weight (e.g., prática de exercícios, predisposição genética, aleatoriedades)

- 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?



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

- 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 \hat{Y}_i ou **fitted value**) – na verdade, minimiza a soma dos quadrados dessa distância
- Essa **distância** corresponde ao **resíduo**; o resíduo é a **manifestação empírica do erro aleatório “verdadeiro”** (o ε_i do modelo populacional):

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

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:

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

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:

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

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

Fonte: Hoffmann (2016)

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

Regressão simples: coeficiente de inclinação estimado

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



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 $\hat{\beta}_1$ 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

Regressão simples: intercepto estimado

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

- A estimativa do intercepto corresponde à **diferença entre a média de Y (i.e., \bar{y}), de um lado, e $\hat{\beta}_1$ 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 $\hat{\beta}_0$ e $\hat{\beta}_1$** (detalhes no slide de aprofundamento “Minimização da soma dos quadrados dos resíduos”)
- Na regressão simples, essa fórmula implica que a **reta estimada passará, necessariamente, por (\bar{x}, \bar{y})** . $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x}$

Alerta:

Não confunda $\hat{\beta}_0$ com a média amostral de Y para $X = 0$

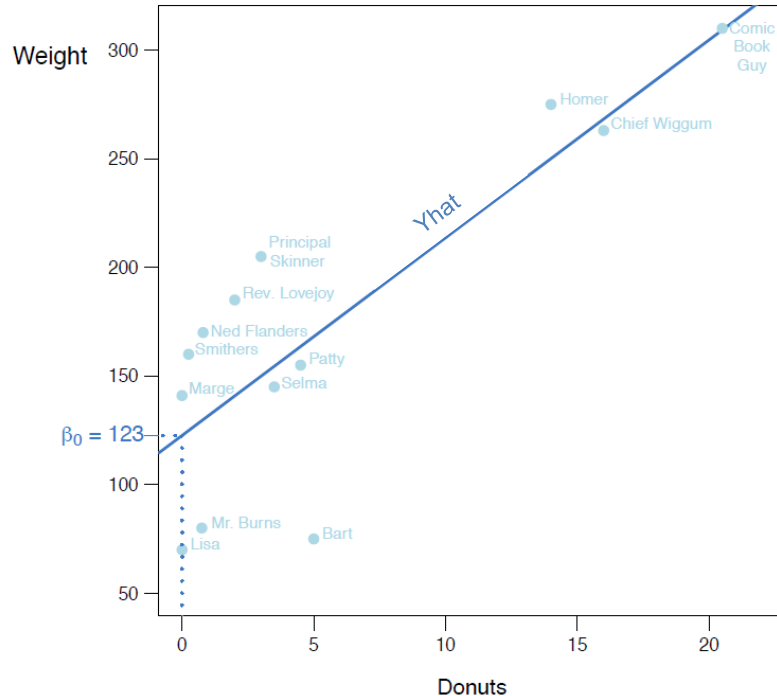


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

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

- $\hat{\beta}_0$ é o valor estimado do peso para indivíduos que não consumiram donuts na última semana: 121,6 libras
- Note que $\hat{\beta}_0$ não é o peso médio dos indivíduos que não consumiram donuts na última semana (apenas Marge e Lisa):
 $\bar{Y} = 171,8$
 $(\bar{Y}|X=0) = 105,5$
 $(\bar{Y}|X>0) = 183,9$
- 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 prevalente é aquele das pessoas que consumiram donut

Como é que chama o nome disso?

Definição

O **estimando** (*estimand*) é a quantidade de interesse cujo valor verdadeiro desejamos conhecer; também conhecido como: **parâmetro, coeficiente**

O **estimador** (*estimator*) é um método para estimar o estimando (e.g., MQO)

A **estimativa** (*estimate*) é um valor numérico para o estimando que resulta do uso de um estimador particular; também conhecido como: **parâmetro estimado, coeficiente estimado**

Exemplo

$$\beta_1$$

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

$$\hat{\beta}_1$$

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

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

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, \dots, X_k \text{ são mantidos constantes})$

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

Interpretação detalhada adiante.

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, p. 203)

Regressão multivariada é uma regressão com duas ou mais variáveis explicativas

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

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, \dots, X_k \text{ são mantidos constantes})$

Como calcular os parâmetros?

- Assim como na regressão bivariada, MQO encontra os $\hat{\beta}$ hats que minimizam a soma dos quadrados dos resíduos

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

- 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 à 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 $x_2 = X_{2i} - \bar{X}_2$.

Regressão multivariada é uma regressão com duas ou mais variáveis explicativas

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

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, \dots, X_k \text{ são mantidos constantes})$

Como interpretar $\hat{\beta}_1$ (ou qualquer dos outros coeficientes de inclinação)?

- Um aumento de uma unidade no respectivo X está associado a uma variação estimada de $\hat{\beta}_1$ em Y , **mantendo-se constante(s) a(s) outra(s) variável(eis) explicativa(s) incluída(s) do modelo**
 - Usualmente, substitui-se o trecho final (negrito) por **ceteris paribus** ou **cæteris paribus**, que significa “tudo o mais constante” em latim

Vide [apêndice](#) sobre como a regressão múltipla “mantém” demais covariáveis constantes (teorema de Frisch-Waugh-Lovell).

Exemplo/ Atividade

Três modelos de regressão foram estimados para se calcular o custo de imóveis residenciais de certa região da cidade. Para tanto, foram utilizadas as seguintes variáveis:

HOUSEPRICE = Preço do imóvel, em USD

SQFT = Área construída do imóvel, em square feet


BEDRMS = Número de quartos no imóvel

BATHS = Número de banheiros no imóvel

Baseando-se nos resultados das três regressões reportadas a seguir, e desconsiderando questões de significância estatística:

1. Escreva a equação de regressão estimada para cada modelo. Arredonde os valores para o inteiro mais próximo.
2. Interprete os coeficientes de SQFT, BEDRMS e BATHS do terceiro modelo.
3. O que explica os coeficientes negativos de BEDRMS e BATHS no modelo 3?

Modelo 1



Model with Y = HOUSEPRICE (USD) | Estimated coefficient for SQFT: [[138.75031952]]

Estimated intercept: [52350.90728647]

R2: 0.820521867058809

OLS Regression Results

Dep. Variable:

HOUSEPRICE

R-squared:

0.821

Model:

OLS

Adj. R-squared:

0.806

Method:

Least Squares

F-statistic:

54.86

Date:

Wed, 06 Nov 2024

Prob (F-statistic):

8.20e-06

Time:

22:45:28

Log-Likelihood:

-166.79

No. Observations:

14

AIC:

337.6

Df Residuals:

12

BIC:

338.9

Df Model:

1

Covariance Type:

nonrobust

coef

std err

t

P>|t|

[0.025

0.975]

const

5.235e+04

3.73e+04

1.404

0.186

-2.89e+04

1.34e+05

SQFT

138.7503

18.733

7.407

0.000

97.935

179.566

Modelo 2

Model with Y = HOUSEPRICE (USD) | Estimated coefficient for SQFT, BEDRMS: [[148.31382509 -23910.61412805]]
Estimated intercept: [121178.68271482]
R2: 0.8346727773986125

OLS Regression Results										
Dep. Variable:	HOUSEPRICE	R-squared:	0.835							
Model:	OLS	Adj. R-squared:	0.805							
Method:	Least Squares	F-statistic:	27.77							
Date:	Wed, 06 Nov 2024	Prob (F-statistic):	5.02e-05							
Time:	22:49:16	Log-Likelihood:	-166.22							
No. Observations:	14	AIC:	338.4							
Df Residuals:	11	BIC:	340.4							
Df Model:	2									
Covariance Type:	nonrobust									
	coef	std err	t	P> t	[0.025	0.975]				
const	1.212e+05	8.02e+04	1.511	0.159	-5.53e+04	2.98e+05				
SQFT	148.3138	21.208	6.993	0.000	101.635	194.992				
BEDRMS	-2.391e+04	2.46e+04	-0.970	0.353	-7.81e+04	3.03e+04				

Modelo 3

Model with Y = HOUSEPRICE (USD) | Estimated coefficient for SQFT, BEDRMS, BATHS: [[154.79989497 -21587.51917702 -12192.75743263]]
Estimated intercept: [129061.63452086]
R2: 0.8359763585711256

OLS Regression Results										
Dep. Variable:	HOUSEPRICE	R-squared:	0.836							
Model:	OLS	Adj. R-squared:	0.787							
Method:	Least Squares	F-statistic:	16.99							
Date:	Wed, 06 Nov 2024	Prob (F-statistic):	0.000299							
Time:	22:50:31	Log-Likelihood:	-166.16							
No. Observations:	14	AIC:	340.3							
Df Residuals:	10	BIC:	342.9							
Df Model:	3									
Covariance Type:	nonrobust									
	coef	std err	t	P> t	[0.025	0.975]				
const	1.291e+05	8.83e+04	1.462	0.175	-6.77e+04	3.26e+05				
SQFT	154.7999	31.940	4.847	0.001	83.632	225.968				
BEDRMS	-2.159e+04	2.7e+04	-0.799	0.443	-8.18e+04	3.86e+04				
BATHS	-1.219e+04	4.33e+04	-0.282	0.784	-1.09e+05	8.42e+04				

Regressão múltipla

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

$$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, \dots, X_k \text{ são mantidos constantes})$$

- 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, p. 207)

- Todavia, a regressão múltipla **também é limitada como instrumento para apurar 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 correlaciona com outros fatores que influenciam Y

Exogeneidade: $\text{corr}(X, \varepsilon) = 0$



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

Endogeneidade: $\text{corr}(X, \varepsilon) \neq 0$

Lembrete

Ordem das variáveis não altera correlação: $\text{corr}(X, \varepsilon) = \text{corr}(\varepsilon, X)$.

*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, p. 14-15)

Regressão múltipla

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

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, \dots, X_k \text{ são mantidos constantes})$

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, p. 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

apêndice: teorema de FWL

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



Em essência, a regressão multivariada calcula o $\hat{\beta}$ “líquido”, “descontaminado” do efeito de outras variáveis explicativas incluídas no modelo

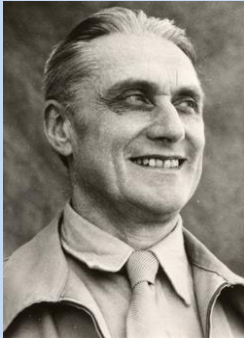
*[Ao interpretar estimativas de regressão multivariada] O que eles [estatísticos] realmente querem dizer é mais como “descontar” o efeito de outras variáveis no modelo. A lógica por trás de dizer que outros fatores são constantes é que, **uma vez que tenhamos calculado os efeitos de outras variáveis, 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 $\hat{\beta}_1$ ”, o que se quer dizer é que, **considerando o efeito de X_2 , estima-se que o efeito de X_1 seja $\hat{\beta}_1$.**

Bailey (2016, p. 197-198)

Teorema de Frisch-Waugh-Lovell (FWL)



Ragnar Frisch
(Norueguês, 1895-1973)



Frederick V. Waugh
(Americano, 1898-1974)



Michael C. Lovell
(Americano, 1930-2018)

- Em uma regressão multivariada, $\hat{\beta}_1$ pode ser obtido seguindo os passos abaixo:
 - Regresse X_1 em X_2, \dots, X_k
 - Compute os resíduos (r_1) obtidos por essa regressão
 - Regresse Y em r_1
- O mesmo vale para todos os coeficientes de inclinação

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



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:					
	(1)	(2)	price	(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*** (315.581)	6,165.257*** (343.611)	6,165.257*** (332.751)		10,033.080*** (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:					

Nestes três modelos, X é um resíduo e, portanto, $X_{bar} = 0$; como a reta de regressão simples passa pela coordenada (X_{bar}, Y_{bar}) , nestes três casos o intercepto = Y_{bar}

*p<0.1; **p<0.05; ***p<0.01

Nestes três modelos, X é um resíduo e, portanto, $\bar{X} = 0$; como a reta de regressão simples passa pela coordenada (\bar{X}, \bar{Y}) , nestes três casos o intercepto = \bar{Y}

Controle estatístico

Avaliação de Políticas Públicas B

12, 24 e 26 de março de 2025

Leitura básica:

CHEIN, Luciana. **Introdução aos modelos de regressão linear: um passo inicial para compreensão da econometria como uma ferramenta de avaliação de políticas públicas**. Brasília: Escola Nacional de Administração Pública, 2019, pp. 09-45, Disponível em: <http://repositorio.enap.gov.br/handle/1/4788>

Leitura complementar:

Demais obras disponibilizadas na pasta “Referências sobre regressão linear”, disponível em https://drive.google.com/drive/folders/1MndWD_X3Vg9hVGv7-QlyGtYExgp2XAua?usp=sharing