

# Regressão Multivariada

PRI5003 - Lab 10

Instituto de Relações Internacionais - Universidade de São Paulo

22 de Junho de 2017

# Outline

Associação e causalidade (de novo)

Forma funcional

Exemplo no Stata

Interpretação dos coeficientes

Análise do uso da regressão em um paper

# Associação e causalidade

- ▶ Nas últimas 3 aulas vimos como avaliar a associação entre duas variáveis
- ▶ Porém, como temos enfatizado, **associação não é causalidade**
- ▶ A regressão multivariada é um avanço no sentido de achar causalidade
- ▶ Veremos hoje que como ela ajuda, mas veremos também que ela não resolve completamente o problema

### 3 condições para a causalidade

Para concluir que

$$X \longrightarrow Y$$

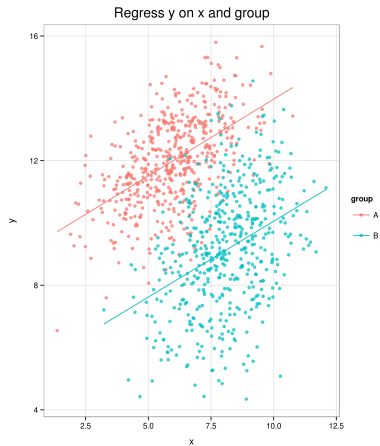
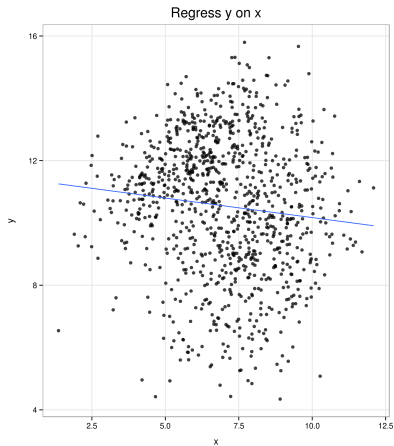
precisamos estabelecer três condições:

1. Ordem temporal:  $Y$  não pode acontecer antes de  $X$ . É a condição de mais fácil verificação.
2. Associação:  $X$  e  $Y$  devem variar mutuamente. Avaliada com as ferramentas que vimos até agora no curso.
3. **Eliminação de alternativas**: outras explicações plausíveis para  $Y$  são descartadas. É a condição mais difícil de verificar.

## Como eliminar explicações alternativas?


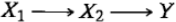



- ▶ Desenhos experimentais permitem o controle do processo de geração dos dados, por isso são considerados o padrão de ouro de estudos causais. O modelo mais usado nas ciências sociais é o de *potential outcomes*, desenvolvido por Donald Rubin.
- ▶ Quando não podemos fazer um experimento, buscamos eliminar explicações alternativas por meio de **variáveis de controle**. A regressão multivariada abre espaço para incluirmos esses controles no modelo.
- ▶ A ideia é observar o efeito de  $X$  sobre  $Y$  quando a influência de explicações alternativas é eliminada.

# O que acontece quando adicionamos um controle?

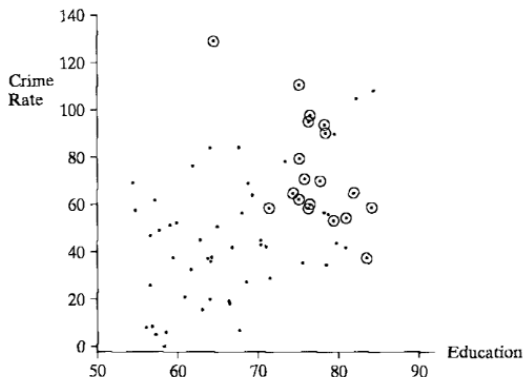


# O que acontece quando adicionamos um controle?

TABLE 10.5: Some Three-Variable Relationships

Graph	Name of Relationship	What Happens after Controlling for $X_2$
	Spurious $X_1 Y$ association	Association between $X_1$ and $Y$ disappears.
	Chain relationship; $X_2$ intervenes; $X_1$ indirectly causes $Y$	Association between $X_1$ and $Y$ disappears.
	Interaction	Association between $X_1$ and $Y$ varies according to level of $X_2$ .
	Multiple causes	Association between $X_1$ and $Y$ does not change.
	Both direct and indirect effects of $X_1$ on $Y$	Association between $X_1$ and $Y$ changes, but does not disappear.

## Um exemplo prático



**FIGURE 11.4:** Scatterplot Relating Crime Rate and Education. The circled points are the counties highest on urbanization. A regression line fitting the circled points has negative slope, even though the regression line passing through *all* the points has positive slope (Simpson's paradox).



## Um exemplo prático

- ▶ O gráfico mostra que os distritos com maior taxa de educação são aqueles com maior criminalidade.
- ▶ Os pontos com um círculo são aqueles com maior taxa de urbanização.
- ▶ Se traçarmos uma reta sobre **todos** os pontos, ela terá inclinação positiva
- ▶ Porém, se a reta for traçada somente entre os pontos com maior urbanização, a inclinação será negativa
- ▶ Portanto, quando controlamos por urbanização, o efeito da educação nas taxas de criminalidade é negativo.
- ▶ A imagem ilustra também o paradoxo de Simpson (que tecnicamente não é um paradoxo, nem foi descoberto por Simpson).

# Equação

## Equação da regressão múltipla

$$y = \alpha + \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_k x_k + \epsilon$$

## No exemplo acima

$$crime = \alpha + \beta_1 * educ + \beta_2 * urban + \epsilon$$

O termo de erro é calculado da mesma maneira que vimos na regressão bivariada. A única diferença é que  $\hat{y}$  é previsto por diversas variáveis, em vez de uma só.

## Relembrando o termo de erro

**Resíduo** é a diferença entre a previsão feita para  $y$  pela reta ajustada e o valor observado para  $y$ .

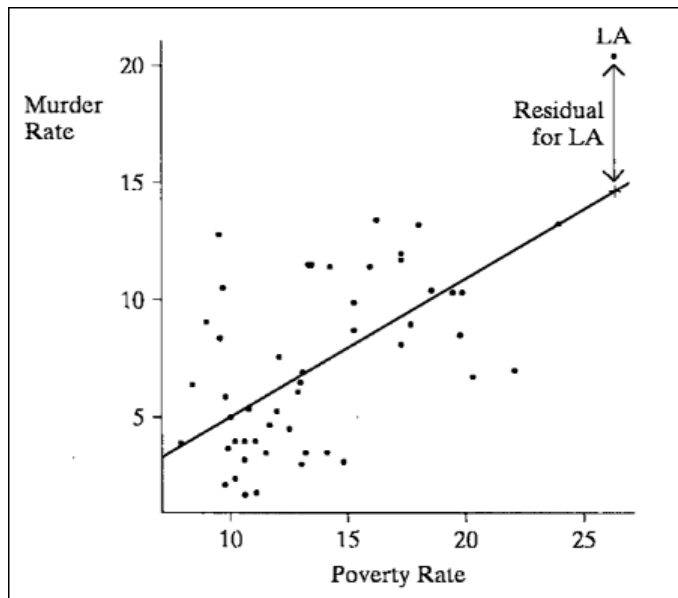
Em notação, resíduo =  $y - \hat{y}$ .

Para sabermos o desvio geral das observações em relação à previsão, calculamos a **soma dos quadrados dos erros**:

$$SQE = \sum (y - \hat{y})^2$$

Como a função minimiza o quadrado dos erros entre a reta e os pontos observados, ela é chamada *linha dos mínimos quadrados*

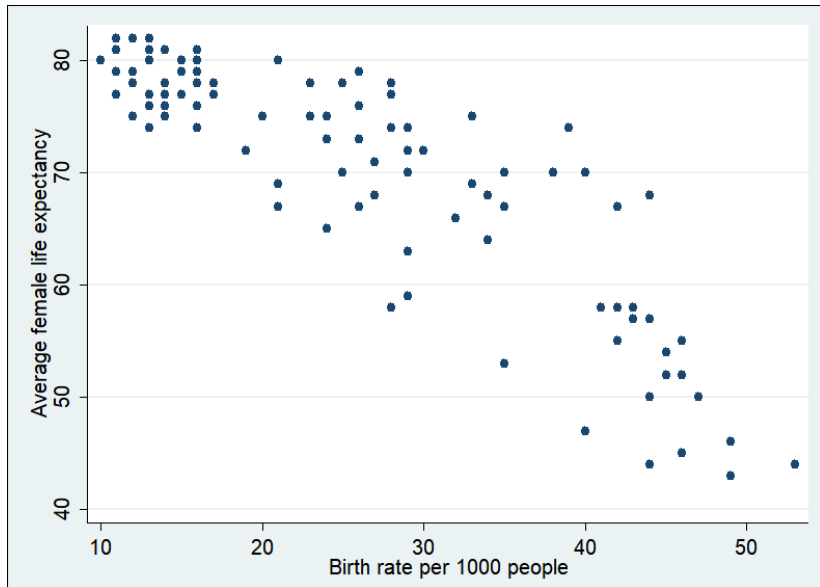
# Ilustração



# No Stata

- ▶ No banco de dados `world95.dta`, observamos uma associação forte e negativa entre a expectativa de vida feminina e a taxa de natalidade.
- ▶ A regressão bivariada entre ambas as variáveis apresenta efeito significativo.
- ▶ Essa associação faz sentido? Vamos observar.

# Associação



# Regressão bivariada

. reg lifeexpf birth_rt						
Source	SS	df	MS	Number of obs = 102		
Model	8844.93696	1	8844.93696	F( 1, 100) = 298.37		
Residual	2964.43559	100	29.6443559	Prob > F = 0.0000		
Total	11809.3725	101	116.924481	R-squared = 0.7490		
				Adj R-squared = 0.7465		
				Root MSE = 5.4447		
lifeexpf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
birth_rt	-.7555723	.0437421	-17.27	0.000	-.8423554	-.6687892
_cons	89.72942	1.26713	70.81	0.000	87.21547	92.24337

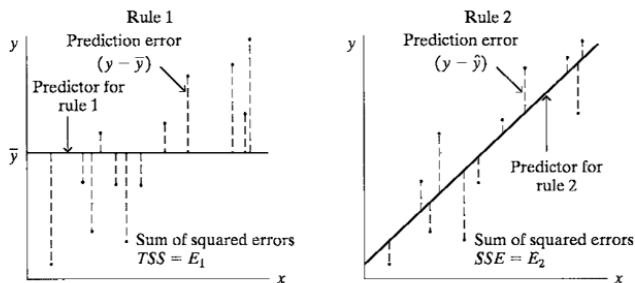
- ▶ Quais são as hipóteses sendo testadas?
- ▶ Interprete os coeficientes
- ▶ Avalie o ajuste do modelo

# Regressão multivariada

. reg lifeexpf birth_rt gdp_cap lit_fema urban fertilty						
Source	SS	df	MS	Number of obs = 80		
Model	7678.17442	5	1535.63488	F( 5, 74) = 65.85		
Residual	1725.77558	74	23.3212916	Prob > F = 0.0000		
Total	9403.95	79	119.037342	R-squared = 0.8165		
				Adj R-squared = 0.8041		
				Root MSE = 4.8292		
lifeexpf	Coef.	Std. Err.	t	P> t	[95% Conf. Interval]	
birth_rt	-.3955945	.2222784	-1.78	0.079	-.8384939	.0473049
gdp_cap	.0000631	.0001647	0.38	0.703	-.0002652	.0003914
lit_fema	.1022896	.0384592	2.66	0.010	.025658	.1789212
urban	.1418301	.0320536	4.42	0.000	.0779619	.2056983
fertilty	.1617463	1.309272	0.12	0.902	-2.447035	2.770528
_cons	63.98663	5.150286	12.42	0.000	53.72447	74.2488

- ▶ Quais são as hipóteses sendo testadas?
- ▶ Interprete os coeficientes
- ▶ Avalie o ajuste do modelo





**FIGURE 9.13:** Graphical Representation of Rule 1 and Total Sum of Squares  $E_1 = TSS = \sum (y - \bar{y})^2$ , Rule 2 and Residual Sum of Squares  $E_2 = SSE = \sum (y - \hat{y})^2$

O  $R^2$  compara quão boa é a reta ajustada em relação a uma previsão que só leva em consideração a média de  $y$ , ignorando  $x$

# Regressão multivariada

- ▶ Além das estatísticas a que já estamos habituados, podemos analisar mais duas: o teste F e o  $R^2$  ajustado
- ▶ O teste F avalia se o conjunto das suas VIs impacta a VD.
- ▶ O  $R^2$  ajustado é uma medida de ajuste que penaliza a inclusão de VIs não explicativas

## **Why the Move to Free Trade? Democracy and Trade Policy in the Developing Countries**

Helen V. Milner with Keiko Kubota

**Abstract** Rising international trade flows are a primary component of globalization. The liberalization of trade policy in many developing countries has helped foster the growth of these flows. Preceding and concurrent with this move to free trade, there has been a global movement toward democracy. We argue that these two trends are related: democratization of the political system reduces the ability of governments to use trade barriers as a strategy for building political support. Political leaders in labor-rich countries may prefer lower trade barriers as democracy increases. Empirical evidence supports our claim about the developing countries from 1970–99. **Regime change toward democracy is associated with trade liberalization, controlling for many factors.** Conventional explanations of economic reform, such as economic crises and external pressures, seem less salient. Democratization may have fostered globalization in this period.

## Milner e Kubota (2005), *Why Move to Free Trade?*

- ▶ Trata-se de um exemplo típico de paper observacional em um *journal* de ponta.
- ▶ VI principal e VD expostas de maneira clara.
- ▶ Separação entre VI principal da pesquisa e variáveis de controle.
- ▶ Controle de explicações alternativas.
- ▶ Cuidado em não falar de causalidade.

# Milner e Kubota (2005), *Why Move to Free Trade?*

## Estatística descritiva

**TABLE 1.** *Summary statistics*

<i>Variable</i>	<i>Observations</i>	<i>Mean</i>	<i>Standard deviation</i>	<i>Minimum</i>	<i>Maximum</i>
TARIFF	907	20.54	15.06	0	102.2
SW OPEN	2790	0.31	0.46	0	1
DATE	5370	1984	8.66	1970	1999
REGIME	3367	-2.07	6.95	-10	10
DEM	4187	0.30	0.46	0	1
DICTATOR	4213	4.74	2.81	1	8
SGL PARTY	5370	0.20	0.40	0	1
MILITARY	5370	0.11	0.32	0	1
PERSONAL	5370	0.17	0.37	0	1
GDP PC	3691	2885.51	4645.60	0	44164.5
LN POP	4880	15.11	2.00	10.57	20.95
EC CRISIS	3403	0.06	0.24	0	1
BP CRISIS	2636	0.59	0.49	0	1
OFFICE	3009	8.43	8.12	0	44
IMF	4008	0.15	0.35	0	1
GATT	4672	0.48	0.50	0	1
FDI	3076	1.90	5.29	-27.24	184.56
US HEG	5370	0.27	0.02	0.24	0.31
AV TARIFF	5370	14.91	11.53	0	30.52

# Milner e Kubota, *Why Move to Free Trade?*

Equação estimada

The basic equation estimating the relationship between democracy and trade policy is:

$$\begin{aligned} \text{tradepolicy}_{i,t} = & \beta_0 + \beta_1 \text{REGIME}_{i,t-1} + \beta_2 \text{IMF}_{i,t-1} + \beta_4 \text{OFFICE}_{i,t-1} \\ & + \beta_5 \text{GDPPC}_{i,t-1} + \beta_5 \text{LNPOP}_{i,t-1} + \beta_6 \text{ECCRISIS}_{i,t-1} \\ & + \beta_7 \text{BPCRISIS}_{i,t-1} + \beta_8 \text{AVOPEN}_{t-1} + u_i + \varepsilon_{i,t} \end{aligned}$$

- ▶ VD = nível de barreira tarifária
- ▶ IMF = assinatura de acordo com o FMI
- ▶ OFFICE = anos de governo
- ▶ EC e BP CRISIS = crise econômica (PIB e Balança de Pagamentos)
- ▶ AVOPEN = Abertura de mercado dos países mais industrializados

# Milner e Kubota, *Why Move to Free Trade?*

Tabela de regressão

<i>Dependent variable</i>	<i>Tariff rates</i>					
	(1)	(2)	(3)	(4)	(5)	(6)
POLITY	-0.264*** (0.096)	-0.247** (0.096)	-0.262*** (0.101)	-0.262*** (0.096)	-0.251*** (0.096)	-0.249*** (0.096)
GDP PC	0.000** (0.000)	0.001*** (0.000)	0.001*** (0.000)	0.000** (0.000)	0.000*** (0.000)	0.000*** (0.000)
LN POP	36.24*** (5.106)	32.50*** (5.433)	34.99*** (6.222)	36.37*** (5.162)	36.61*** (4.976)	36.72*** (5.084)
EC CRISIS		-0.777 (0.670)				
BP CRISIS			0.709 (0.672)			
IMF				0.248 (0.375)		
US HEG					21.515 (15.769)	
FIVE OPEN						-1.646 (1.523)
Constant	2,781*** (203.9)	2,762*** (194.9)	2,821*** (239.2)	2,798*** (209.3)	2,830*** (195.7)	2,581*** (304.3)
<i>Observations</i>	774	765	738	765	774	734
<i>Countries</i>	101	100	98	101	101	101
<i>R</i> <sup>2</sup>	0.79	0.79	0.79	0.79	0.79	0.80
<i>Wald chi</i> <sup>2</sup>	3724	4996	1312	1454	635	767
<i>Prob &gt; chi</i> <sup>2</sup>	0.00	0.00	0.00	0.00	0.00	0.00