



DCP098

Fundamentos para Avaliação Quantitativa de Políticas Públicas

Correlação, equação da reta e equação de regressão

Aula 05
13 de abril de 2022

Ana Paula Karruz

Aterrissando: o que as limitações da correlação têm a ver com a nossa prática de avaliação?



Correlation is not going to cut it!

Abordagem analítica

Questão

Correlação

Regressão

1 Há uma associação entre valores observados de X e Y?



2 Qual é a direção 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?



Have we met before?

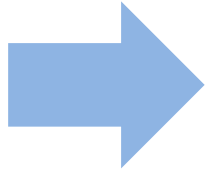
6 Quanto Y varia quando X varia, **mantendo-se constantes as demais influências sobre Y?**



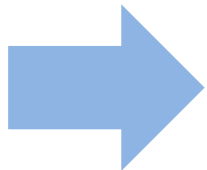
Equação da reta

$$y = ax + b$$

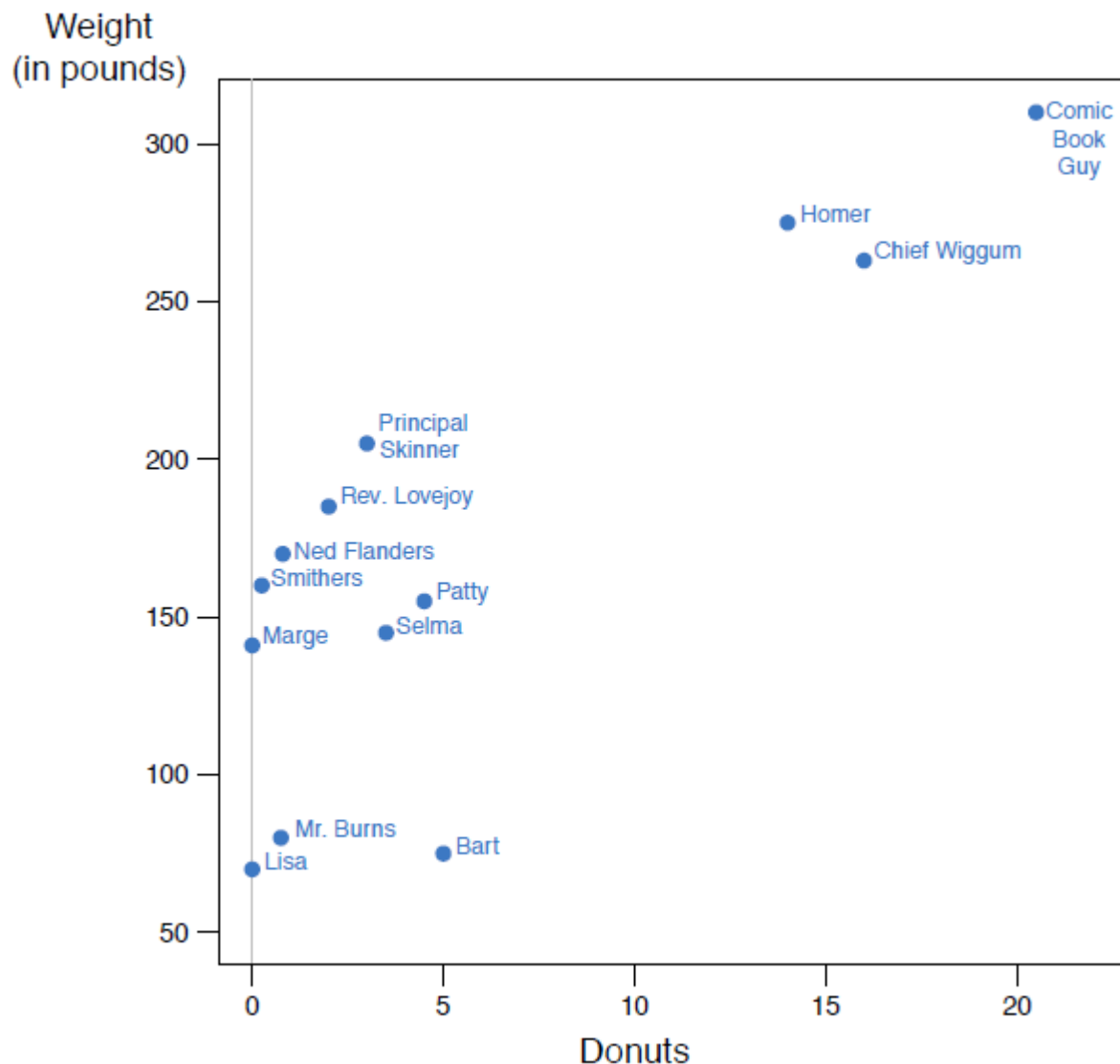




Regressão é sobre identificar a linha que melhor descreva nossos dados. Joia!



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: 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 “verdadeiro” (i.e., efeito geral na população de interesse) de X sobre Y (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:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \varepsilon_i$$

$$= \underbrace{E(Y | X)_i}_{\text{Parte sistemática}} + \underbrace{\varepsilon_i}_{\text{Parte estocástica}}$$

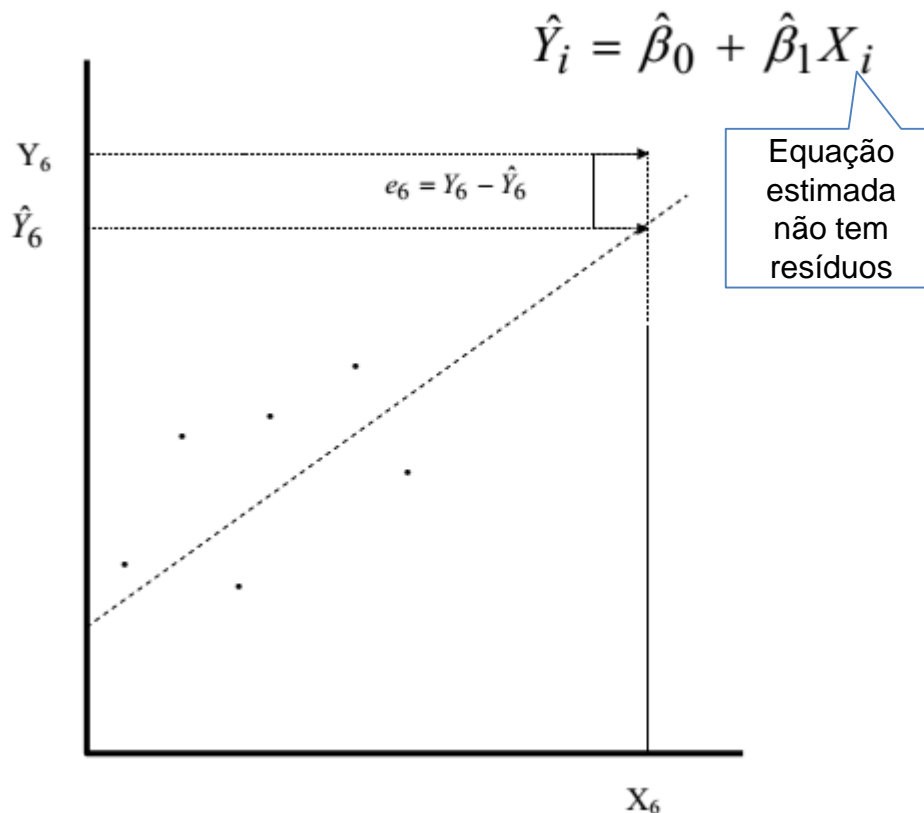
- Desafio:** não observamos os valores dos coeficientes populacionais; nós os estimamos a partir dos dados observados

- Função de regressão estimada:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + e_i$$

$$Y_i = Y_i\text{hat} + \underbrace{e_i}_{\text{Resíduo}}$$

$$e_i = Y_i - Y_i\text{hat}$$



Reta estimada

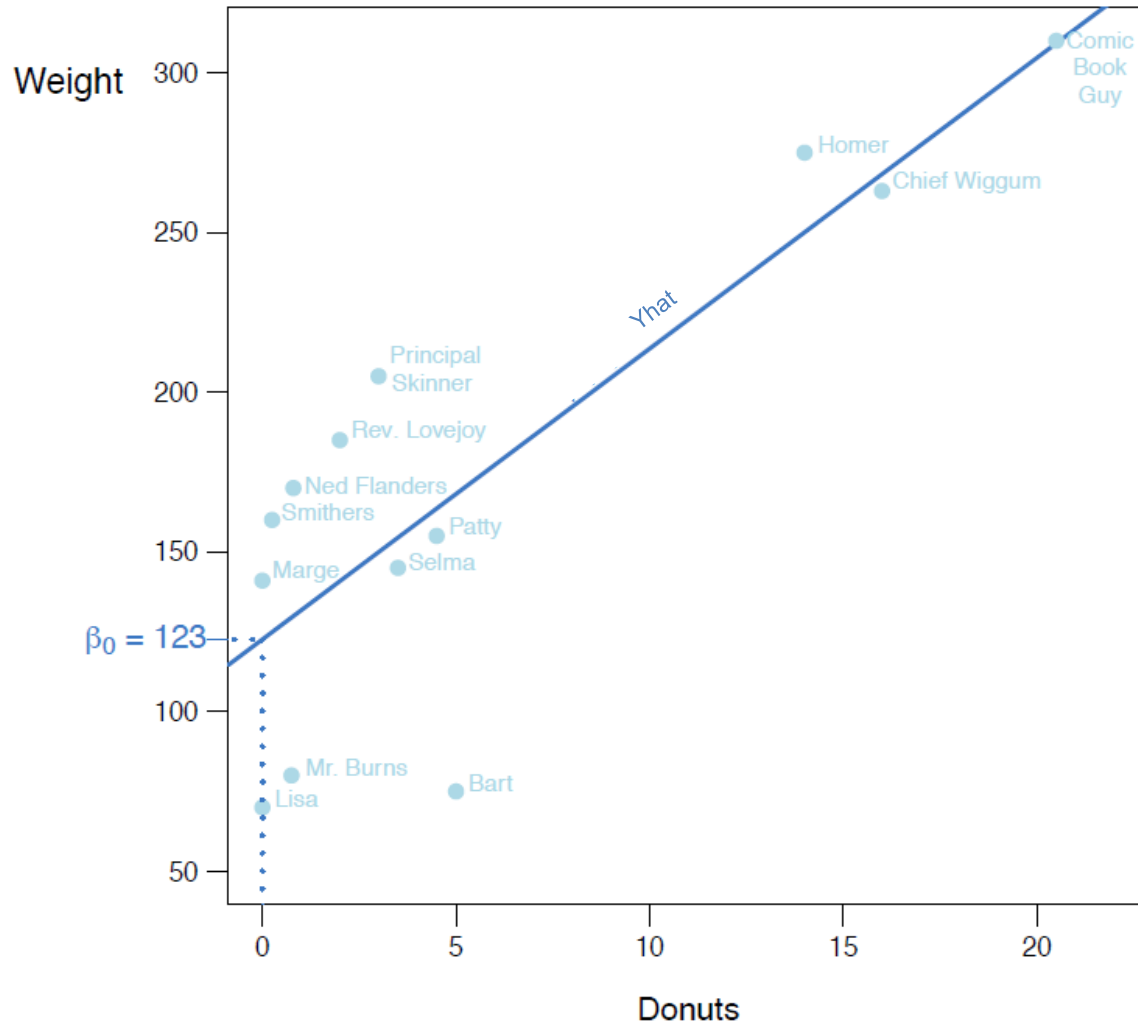


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield
Fonte: Adaptado de Bailey (2016, p 9).

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?** 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 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. 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:

$$E(\varepsilon) = 0$$

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

Premissa da média condicional zero

$$E(\varepsilon|X) = 0$$

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 correlaciona com outros fatores que influenciam Y



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

*Speaking statistically, we highlight this major statistical challenge by saying that the donut variable is endogenous. **An independent variable is endogenous if changes in it are related to factors in the error term.***

[...]

Endogeneity is everywhere; it's endemic.

Bailey (2016: 14-15)

Premissas de MQO (lista estendida)

O que eventualmente ficou de fora não é correlacionado com as variáveis incluídas – se for, teremos o problema do viés de variável omitida

- O modelo de regressão é **linear nos parâmetros** (i.e., coeficientes são adicionados e com expoente = 1)

- ϵ , o termo de erro aleatório, tem **média populacional = 0**
- Não há correlação entre cada uma das variáveis explicativas e o termo de erro – esta é a premissa de **exogeneidade: correlação $(X_j, \epsilon) = 0$**

Premissa da média condicional zero é resultado destas duas premissas

- O termo de erro tem variância constante (i.e., é **não há heteroscedasticidade**)
- Os erros não são correlacionados entre si (i.e., **não existe autocorrelação** serial ou espacial)
- As variáveis explicativas não são uma função linear das outras (i.e., **não há multicolinearidade**)
- O **erro** tem distribuição **normal**: $\epsilon \sim N(0, \sigma^2)$

Premissa requerida para teste de hipótese e intervalo de confiança, não para estimação MQO

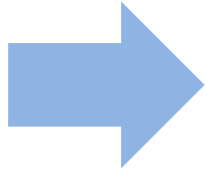


Ilustração:

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



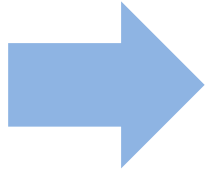


Ilustração:

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





DCP098

Fundamentos para Avaliação Quantitativa de Políticas Públicas

Correlação, equação da reta e equação de regressão

Aula 05
13 de abril de 2022

Ana Paula Karruz