DCP098
Fundamentos para Avaliação Quantitativa de Políticas Públicas

Regressão simples: como interpretar e calcular coeficientes

Aula 06
18 de abril de 2022

Ana Paula Karruz

Interpretação dos coeficientes estimados

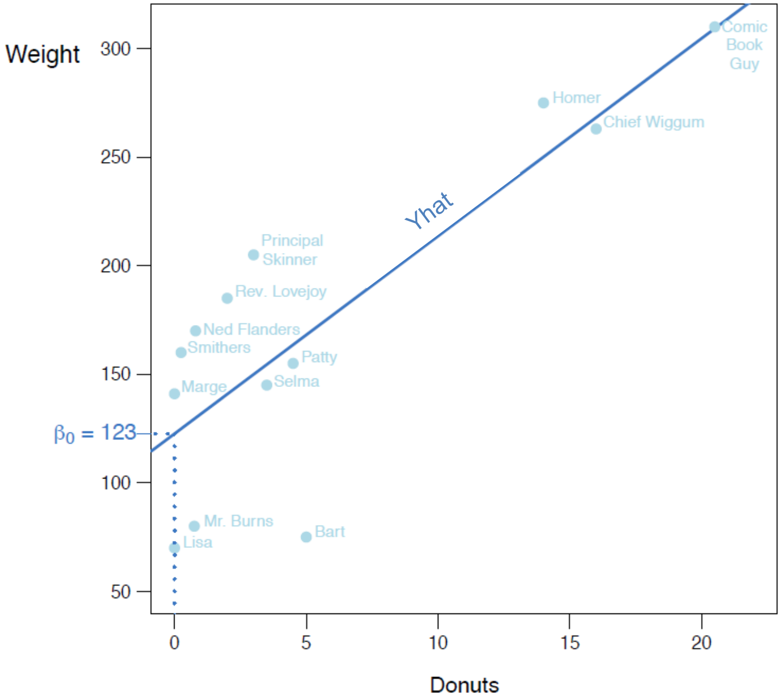


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

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:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 * X_i$$

$$\hat{Y}_i = 121.6 + 9.2 * X_i$$

onde Y = peso em libras, X = 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 está associado a um aumento 9,2 pounds no peso

ceteris paribus?

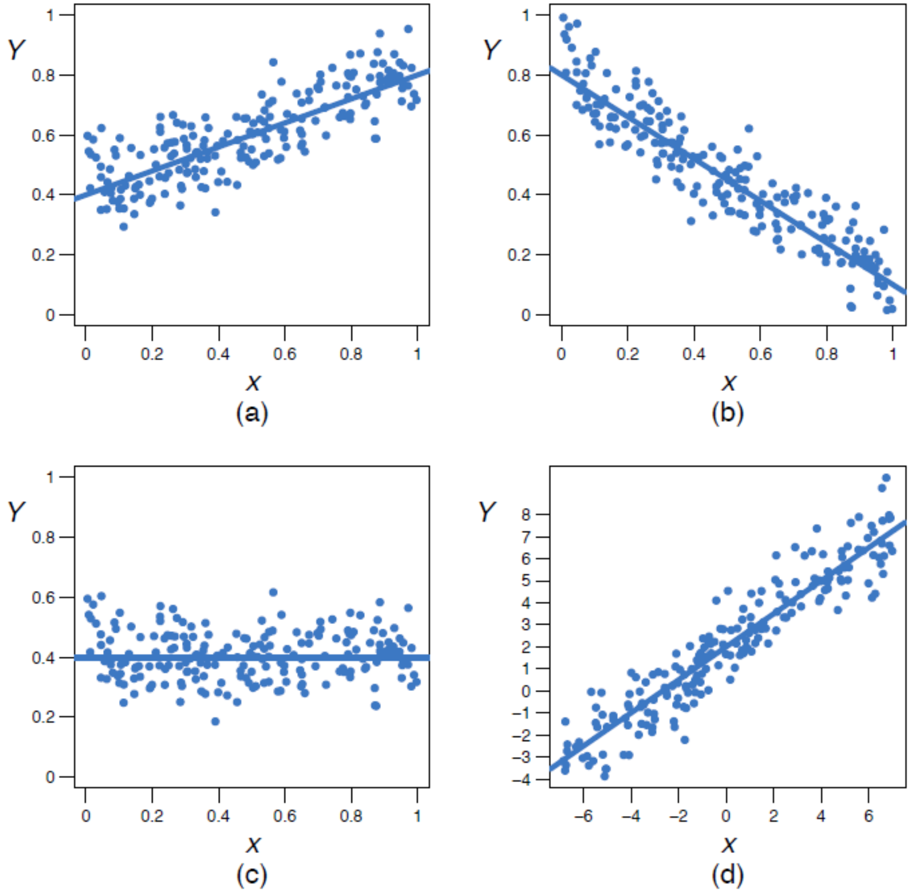


FIGURE 1.4: Examples of Lines Generated by Core Statistical Model

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?

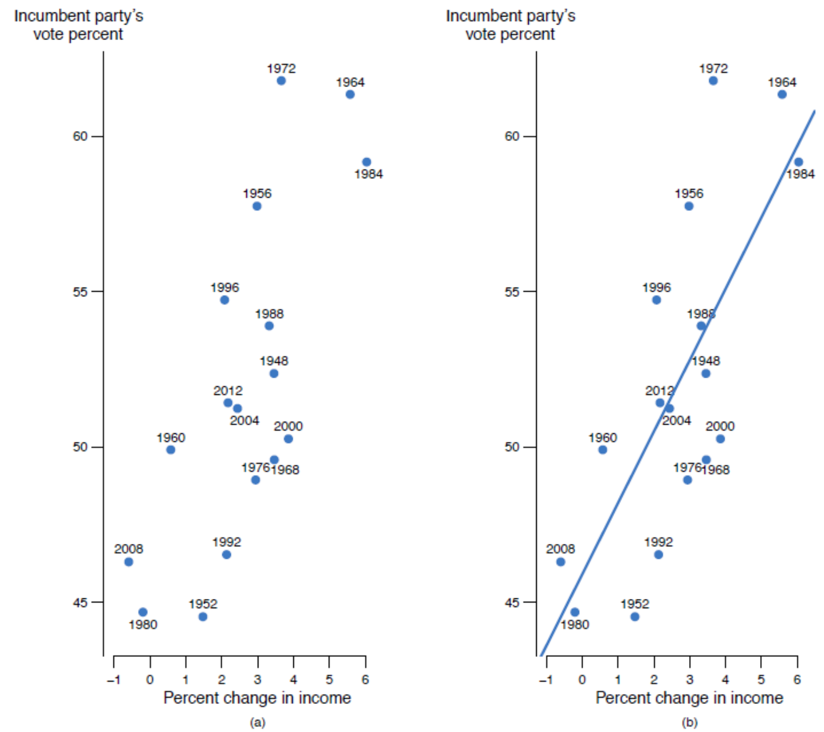


FIGURE 3.1: Relationship Between Income Growth and Vote for the Incumbent President's Party, 1948-

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 caracterize 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?

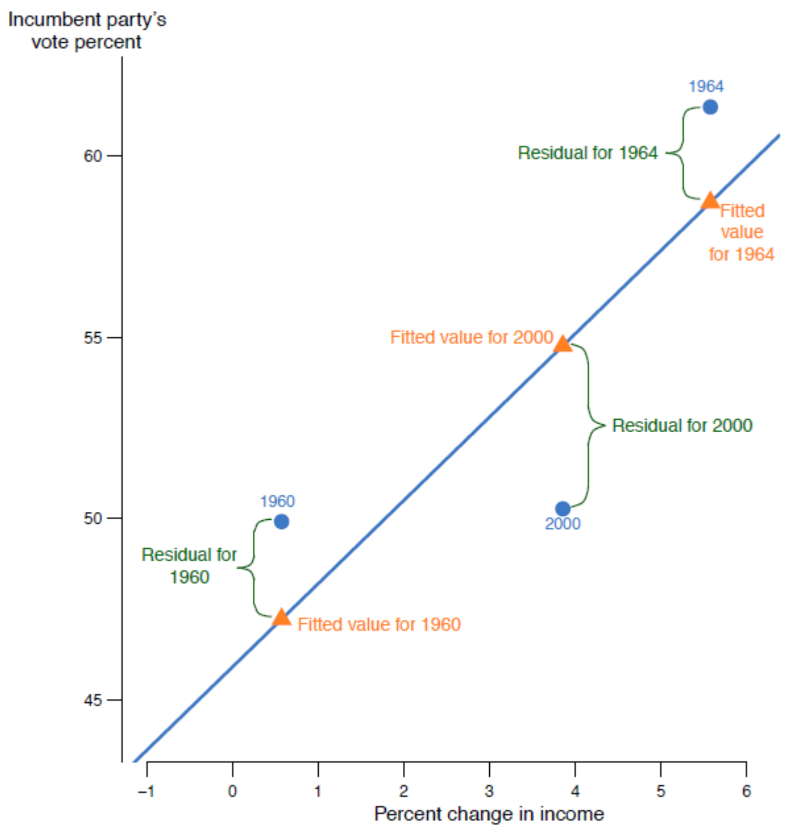


FIGURE 3.3: Fitted Values and Residuals for Observations in Table 3.1

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):

  $$\text{Resíduo}_i = Y_i - (\beta_0\text{hat} + \beta_1\text{hat}X_i)$$

  Há várias representações de resíduo, entre elas:

  $$e_i,\ \hat{e}_i,\ r_i,\ \varepsilon_i\text{hat},\ Y_i - Y_i\text{hat}$$

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

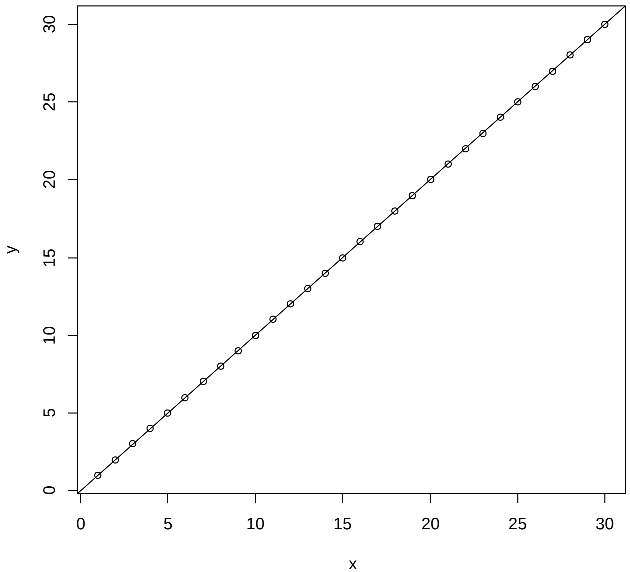
$$\sum_{i=1}^{N} \hat{e}_i^2 = \sum \left(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_1\right)^2$$

Especificamente, MQO identifica a linha que minimiza a soma do <u>quadrado</u> dos resíduos

- Valores elevados ao quadrado são sempre positivos

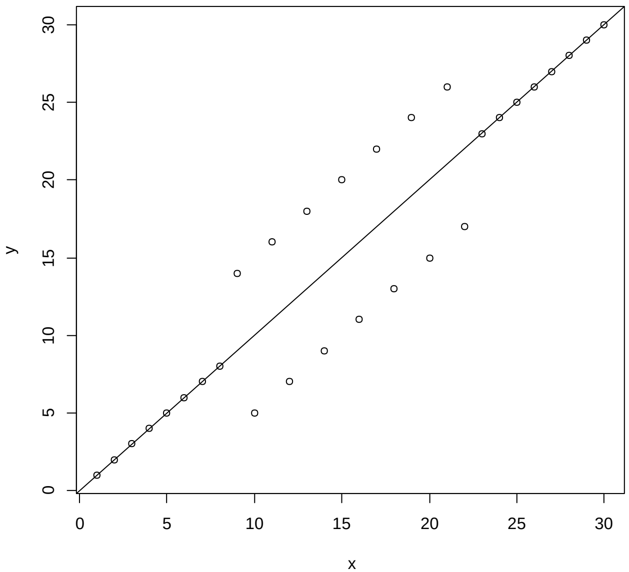
- Resíduos são elevados ao quadrado para que erros de estimação "para cima" e "para baixo" não se anulem

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)

- Conceitualmente, numerador é uma medida de quanto os valores dos **pares ordenados (X_i, Y_i)** são **relacionados entre si**

- 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 oculto "Minimização da soma dos quadrados dos resíduos")
- Na regressão simples, essa fórmula implica que a **reta estimada passará**, necessariamente, **pela coordenada ($\bar{X}$, $\bar{Y}$)**.

Voltaremos a este ponto

Alerta:

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

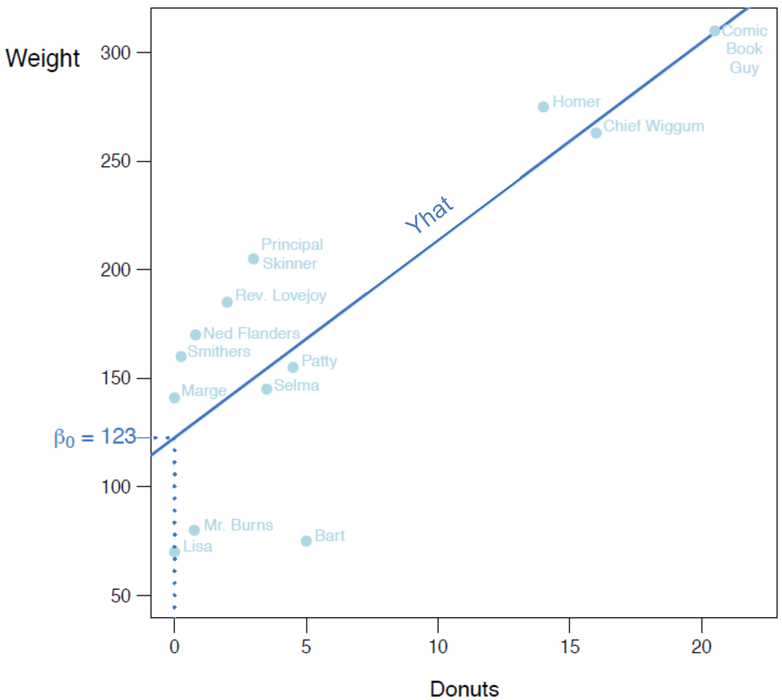


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

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

Voltando ao exemplo:
Bivariate OLS and presidential elections

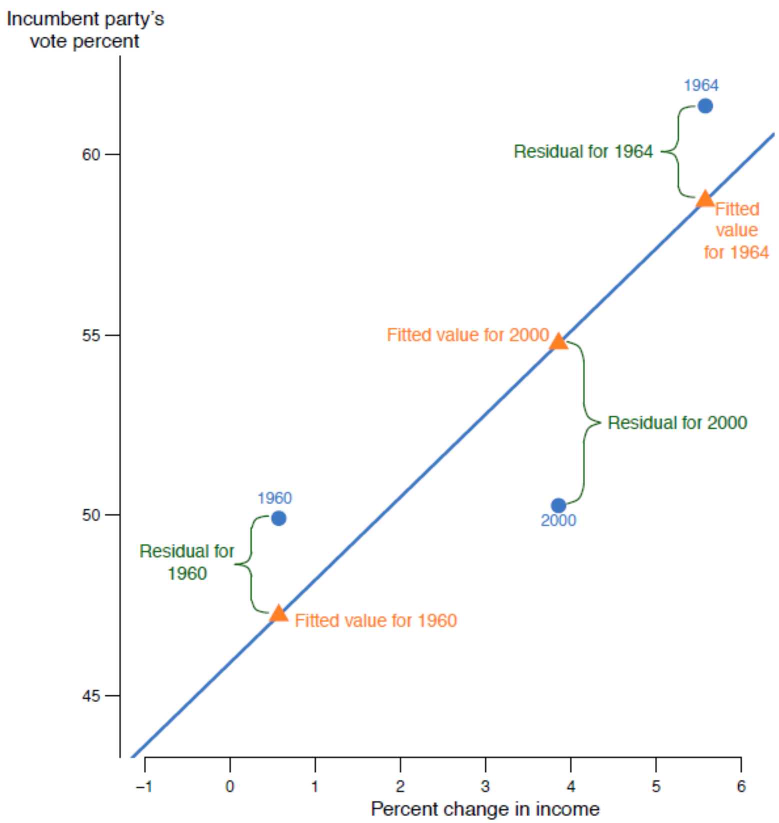


FIGURE 3.3: Fitted Values and Residuals for Observations in Table 3.1

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

DCP098
Fundamentos para Avaliação Quantitativa de Políticas Públicas

Regressão simples: como interpretar e calcular coeficientes

Aula 06
18 de abril de 2022

Ana Paula Karruz