DCP098
Fundamentos para Avaliação Quantitativa de Políticas Públicas

Grau de ajuste (*goodness of fit*)

Aula 17
30 de maio de 2022

Ana Paula Karruz

Agenda para esta aula e as seguintes

1. **Grau de ajuste**
2. Regressão multivariada: precisão
3. Multicolinearidade
4. Heteroscedasticidade

Grau de ajuste (goodness of fit): A regressão está bem ajustada aos dados?

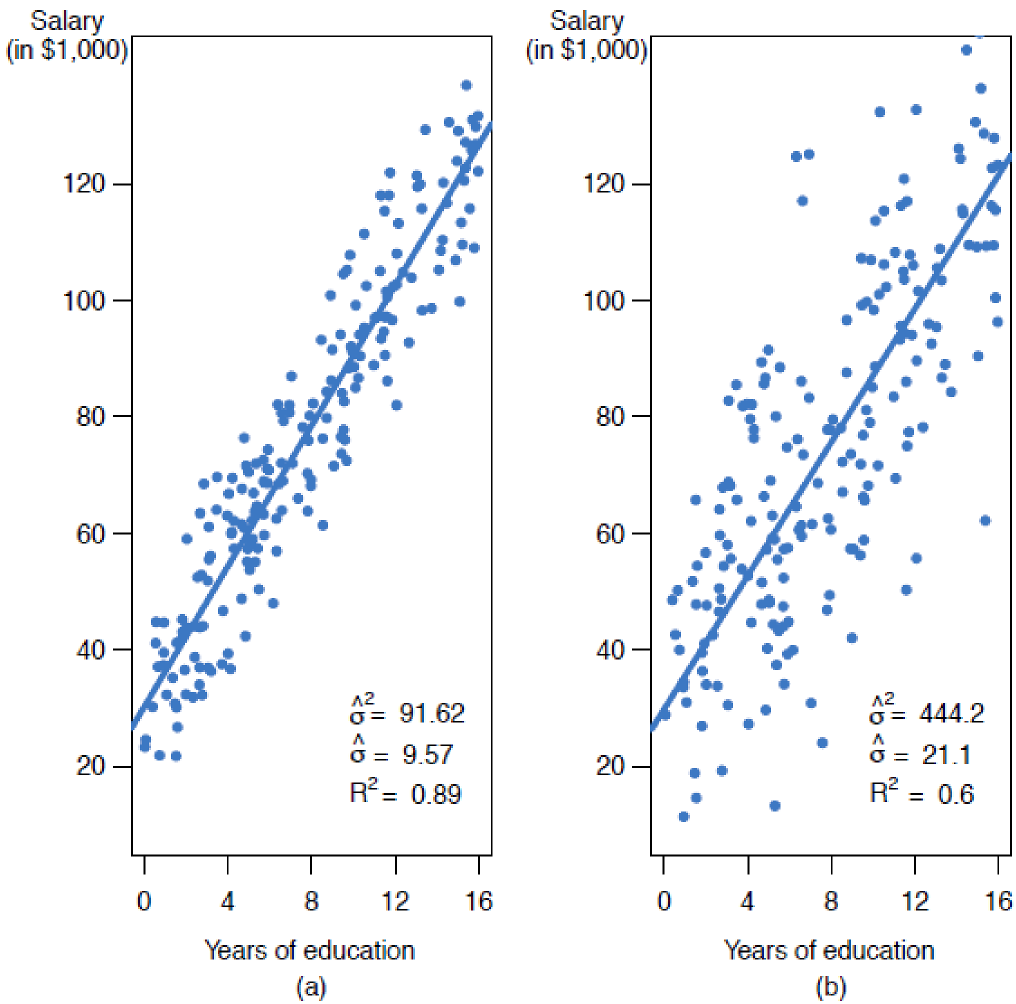


FIGURE 3.9: Plots with Different Goodness of Fit

Fonte: Bailey (2016: 110).

Grau de ajuste (goodness of fit):
A regressão está bem ajustada aos dados?

- Esta pergunta **não tem resposta direta**, e isso **não é um problema**: tipicamente, nossa preocupação principal é estimar o efeito de X sobre Y de maneira acurada (i.e., sem viés); portanto, nosso foco não é fazer a melhor previsão de Y

- Podemos analisar o grau de ajuste **em regressões simples ou múltiplas**; na múltipla, interessa também **saber se a adição de variáveis está melhorando o ajuste**

- Bailey (2016: 106-111) propõe três formas de avaliar o ajuste da regressão:

  – Via **gráfico de dispersão e plotagem da linha de regressão**: útil para detectar outliers, porém análise é subjetiva

  – Via **erro padrão da regressão (σhat)**: grosseiramente, corresponde à distância média entre os valores observados e previstos de Y; no R, o erro padrão da regressão é denominado residual standard error

  – Via **coeficiente de determinação, aka R^2**: corresponde à proporção da variação de Y em torno de sua média (Ybar) que é "explicada" pelo modelo

Abordagem mais usada

Grau de ajuste (goodness of fit):
A regressão está bem ajustada aos dados?

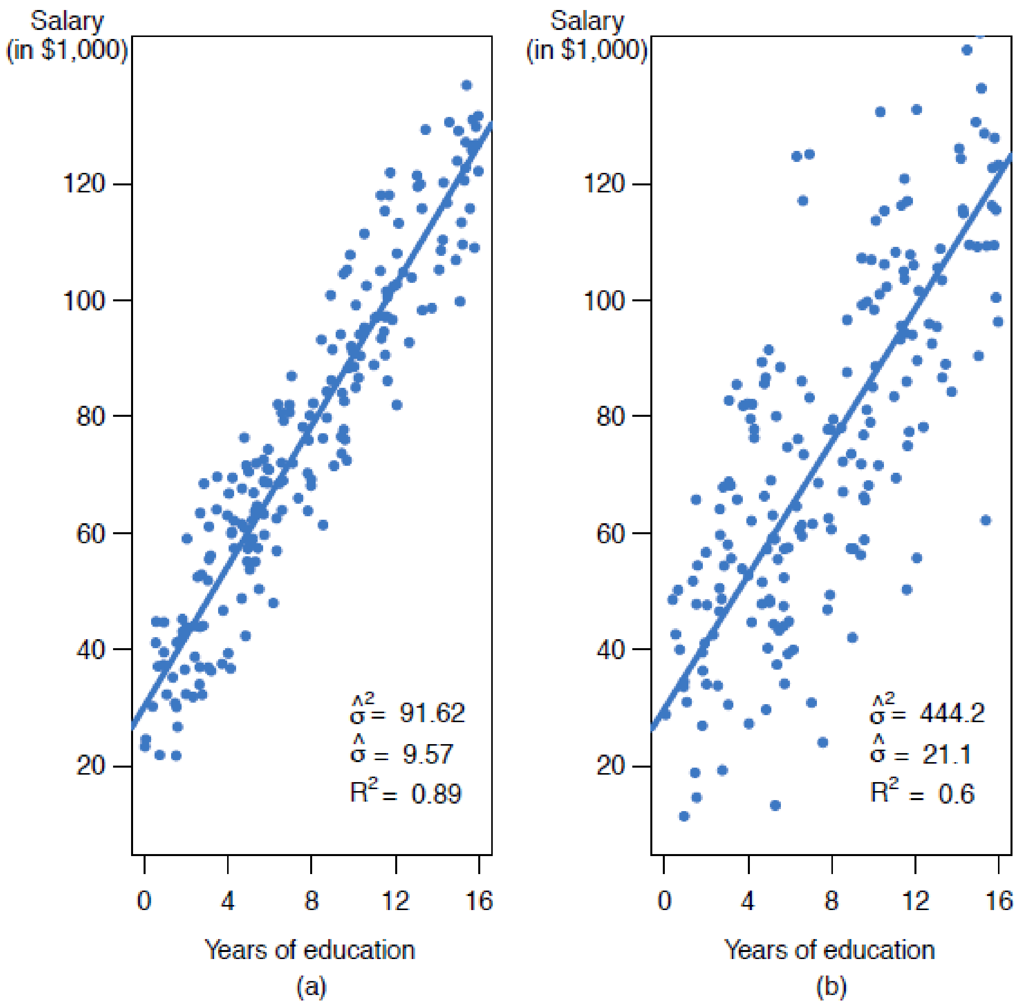
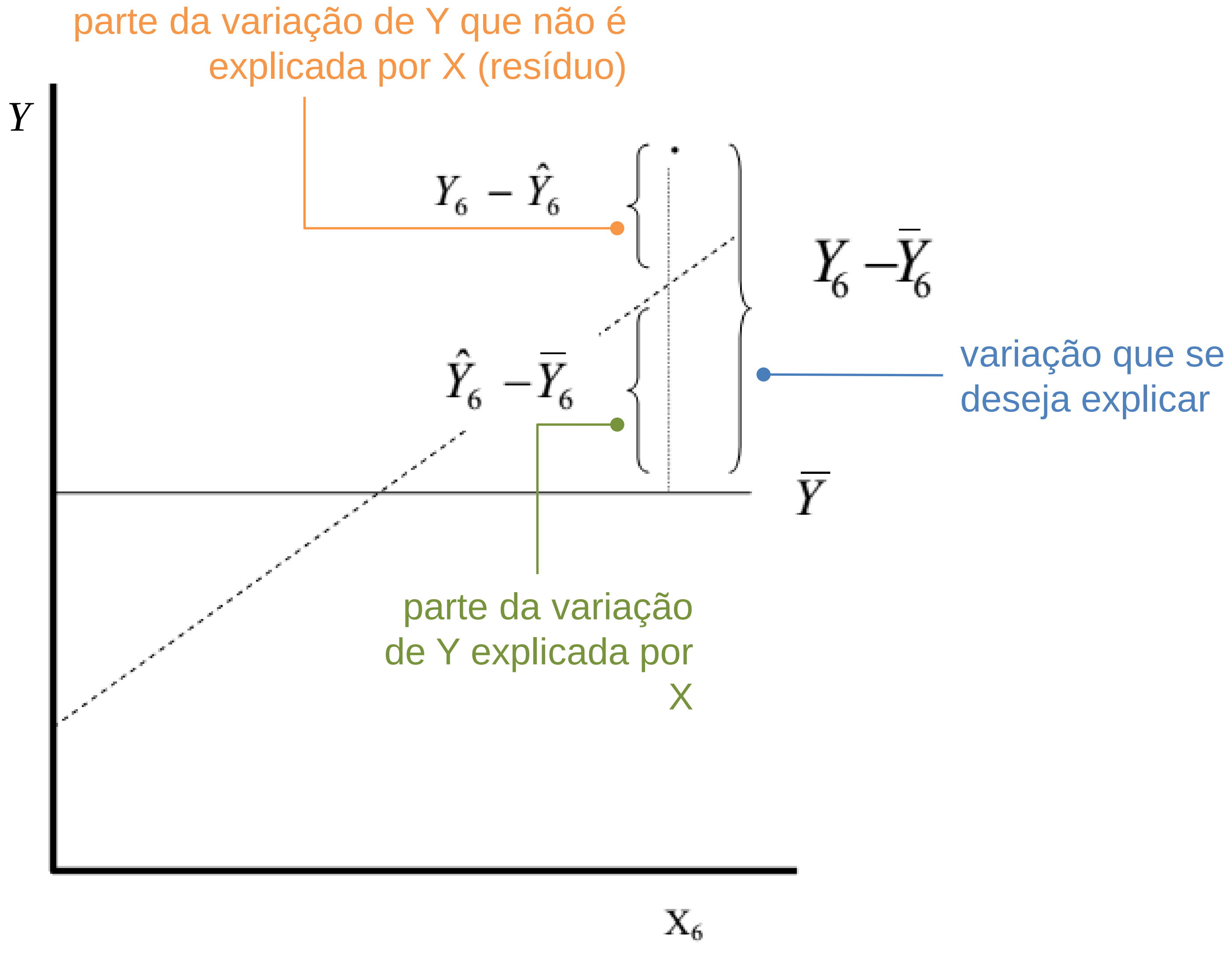


FIGURE 3.9: Plots with Different Goodness of Fit

Fonte: Bailey (2016: 110).



parte da variação de Y que não é
explicada por X (resíduo)
Y
$Y_6 - \hat{Y}_6$
$Y_6 - \bar{Y}_6$
$\hat{Y}_6 - \bar{Y}_6$
variação que se
deseja explicar
$\bar{Y}$
parte da variação
de Y explicada por
X
X_6

Soma dos quadrados

- Soma dos quadrados total (SQT) é uma medida da variação amostral total em Y_i; mede a dispersão de Y_i em torno de sua média (Ybar):

$$SQT = \sum_{i=1}^{n} (y_i - \bar{y})^2$$

- Soma dos quadrados explicada (SQE) mede a variação amostral de Yhat em torno de Ybar:

$$SQE = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$$

- Soma dos quadrados dos resíduos (SQR) mede a variação amostral dos resíduos:

$$SQR = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

- Variação total em Y_i é a soma da variação explicada e da variação não explicada:

$$SQT = SQE + SQR$$

R^2 é a proporção da variação de Y em torno de Ybar que é explicada pela regressão

- R^2 é a razão entre a variação explicada (SQE) e a variação total (SQT):

$$R^2 = \frac{SQE}{SQT} = \frac{\sum(\hat{Y}_i - \overline{Y})^2}{\sum(Y_i - \overline{Y})^2}$$

$$0 \leq R^2 \leq 1$$

- R^2 pode também ser definido como o **quadrado da correlação entre Y e Yhat***; **intuição:** se o modelo explica bem os dados, valores observados e previstos serão altamente correlacionados

- Na **regressão simples, R^2** corresponde ao **quadrado da correlação entre Y e X_1***

* Interpretação válida para modelos com termo de intercepto.

R^2 alto não é condição necessária, nem suficiente, para que uma análise de regressão seja útil

- Seleção de variáveis explicativas com base no tamanho do R^2 **pode levar a modelos absurdos**

- Podemos ter um modelo **carregado de endogeneidade com R^2 alto**

- R^2 pequeno indica que não incluímos no modelo fatores importantes para a formação de Y; mas isso não significa que fatores em ε estejam correlacionados com uma ou mais das variáveis explicativas – em outras palavras, R^2 pequeno **nada diz sobre acurácia (ou falta dela)**

- Não há um valor mínimo de R^2 para que a regressão seja crível; é **comum experimentos gerarem R^2 baixo**

R^2 alto não é condição necessária, nem suficiente, para que uma análise de regressão seja útil

Pode haver todos os tipos de razões para o R^2 ser baixo – o mundo pode ser tão confuso que σhat^2 seja alto, por exemplo – mas o modelo pode, no entanto, fornecer informações valiosas.

σhat^2 = variância da regressão

Bailey (2016: 109)

Uma preocupação razoável poderia ser que devemos ser cautelosos com os resultados do OLS quando o ajuste do modelo parece muito ruim. Não é assim que funciona. Os coeficientes nos dão as melhores estimativas com base nos dados. Os erros padrão dos coeficientes incorporam o ajuste ruim (via the σhat^2).

Bailey (2016: 116-117)

Sobre focar no R2: *Não é assim que avaliamos os modelos [...] avaliamos a força das relações estimadas com base em estimativas de coeficientes e erros padrão, não olhando diretamente para R^2.*

Bailey (2016: 205)

Grau de ajuste e seleção de variáveis explicativas

- O R^2 **nunca diminui** quando outra variável independente é adicionada à regressão – o R^2 pode manter-se constante, mas normalmente aumenta

- Isso ocorre porque a **soma do quadrado dos resíduos nunca aumenta** quando variáveis explicativas são acrescentadas ao modelo

*Em outras palavras, **toda vez que adicionamos uma variável** a um modelo, **não pioramos o ajuste** e, **na prática, melhoramos o ajuste pelo menos um pouco**, mesmo que a variável adicionada não afete verdadeiramente a variável dependente. Apenas por acaso, estimar um coeficiente diferente de zero nessa variável normalmente melhorará o ajuste de algumas observações. Portanto, o **R^2 mantém ou aumenta à medida que adicionamos variáveis**.*

Bailey (2016: 231)

- Essa característica faz de R^2 uma **estatística fraca para decidir se devemos incluir mais** variáveis no modelo

R^2 versus R^2 ajustado

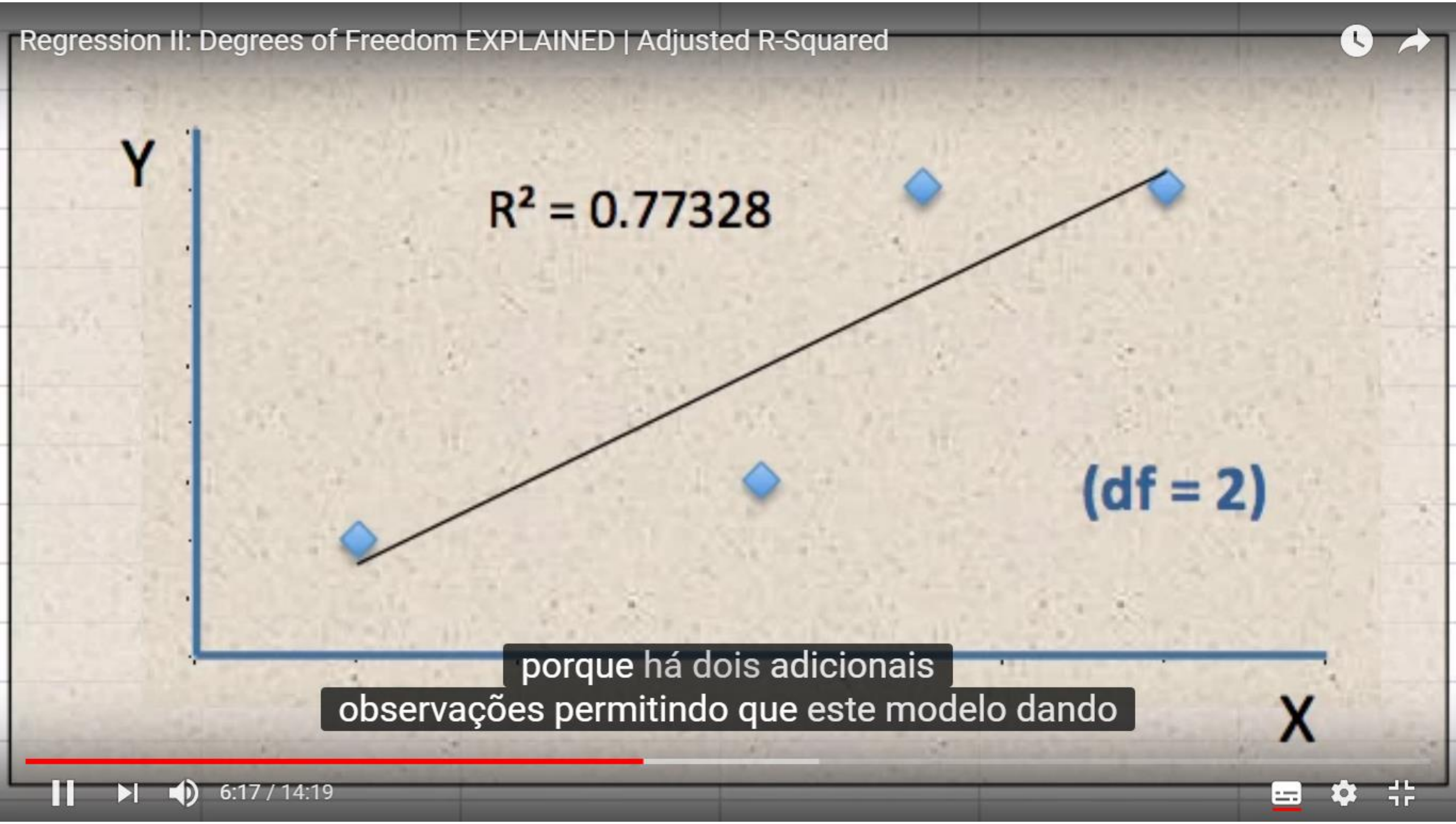
- O R^2 ajustado corrige o R^2 para refletir a perda de graus de liberdade que ocorre quando incluímos variáveis na regressão

$$\overline{R}^2 = 1 - (1 - R^2)\frac{n-1}{n-k-1}$$

- Graus de liberdade = n – k – 1, onde:
  - n = tamanho da amostra
  - k = número de variáveis independentes
  - k + 1 = número de parâmetros estimados (inclui intercepto)

Ao estimarmos mais parâmetros com os mesmos dados, estamos reduzindo o número de observações livres

R^2 ajustado e graus de liberdade



https://youtu.be/4otEcA3gjLk

R^2 versus R^2 ajustado

- **R^2 ajustado não possui uma interpretação direta**

- R^2 ajustado ≤ R^2

- R^2 ajustado pode assumir valores **negativos**; Exemplo de Wooldridge (2008: 191): $R^2 = 0{,}10$; $n = 51$; $k = 10$; R^2 ajustado = - 0,125

- Quando o tamanho da **amostra** for muito **grande**, **não haverá diferença** material entre o R^2 e o R^2 ajustado; neste caso, **melhor guiar-se pelo R^2**, pois o **R^2 ajustado não possui uma interpretação direta**

- Para sabermos qual a contribuição individual de cada variável adicionada, é preciso observar o R^2 ajustado em **modelos aninhados que difiram unicamente quanto a essa variável** (presente em um modelo, ausente no outro)

Exemplo: R^2 e R^2 ajustado

Two models of estimated weight:

Variable	Model 1	Model 2
Height	6.38	6.36
Mail Box #		0.02
Intercept	103.4	102.3
N	20	20
R^2	0.74	0.75
Adjusted R^2	0.73	0.72

R2 e R2 ajustado no R

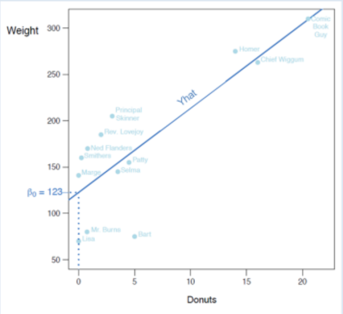


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

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?

```
> summary(reg.pounds)

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
```

DCP098
Fundamentos para Avaliação Quantitativa de Políticas Públicas

Grau de ajuste (*goodness of fit*)

Aula 17
30 de maio de 2022

Ana Paula Karruz