PROGRAMA DE
PÓS-GRADUAÇÃO EM
CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

Aprofundamento em regressão multivariada

Aula 5
5 de outubro de 2022

Ana Paula Karruz

Agenda para esta aula e a próxima

1. Grau de ajuste (não é tema exclusivo da regressão múltipla)
2. Regressão multivariada: precisão
3. Multicolinearidade
4. Heteroscedasticidade (não é exclusivo da regressão múltipla)
5. Viés de variável omitida (não é exclusivo da regressão múltipla)
6. Variável irrelevante (estudaremos da perspectiva da regressão múltipla)

Agenda para esta aula

1. **Grau de ajuste**

2. Regressão multivariada: precisão

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

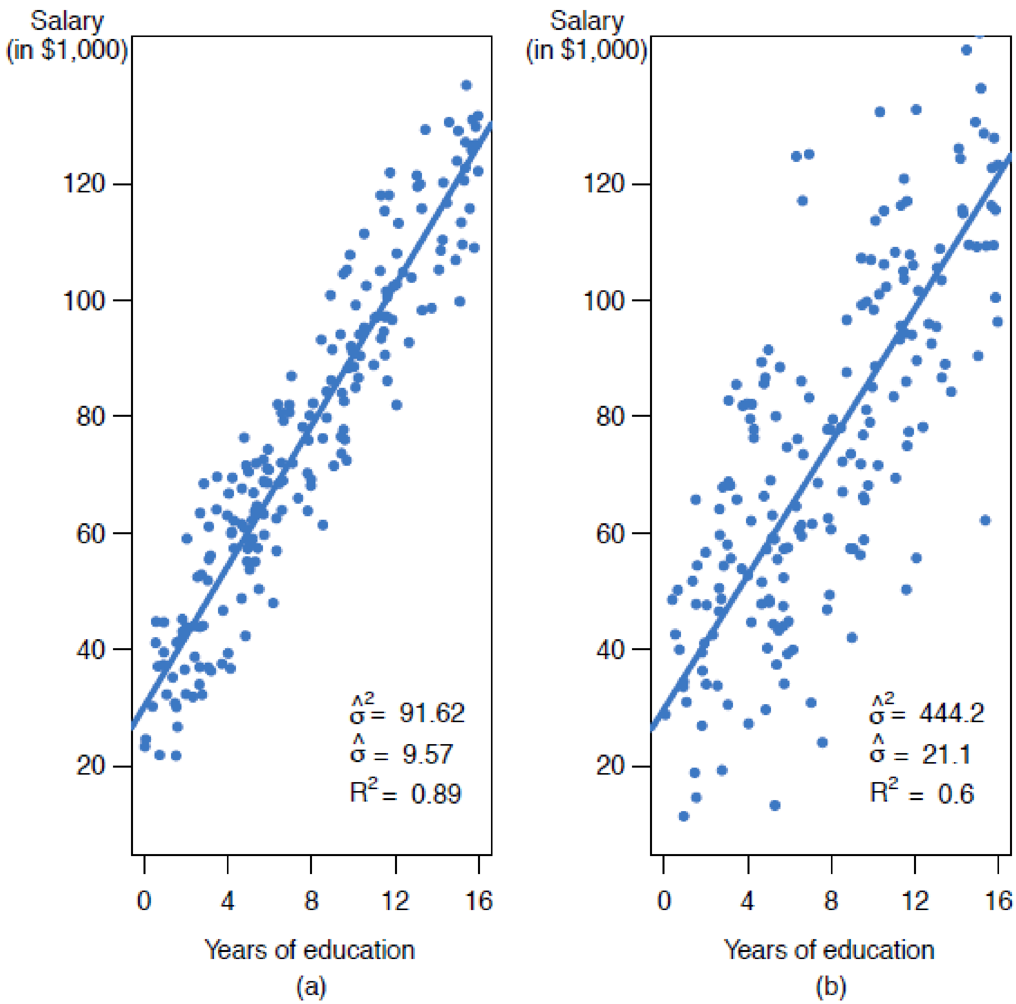


FIGURE 3.9: Plots with Different Goodness of Fit

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 para se 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?

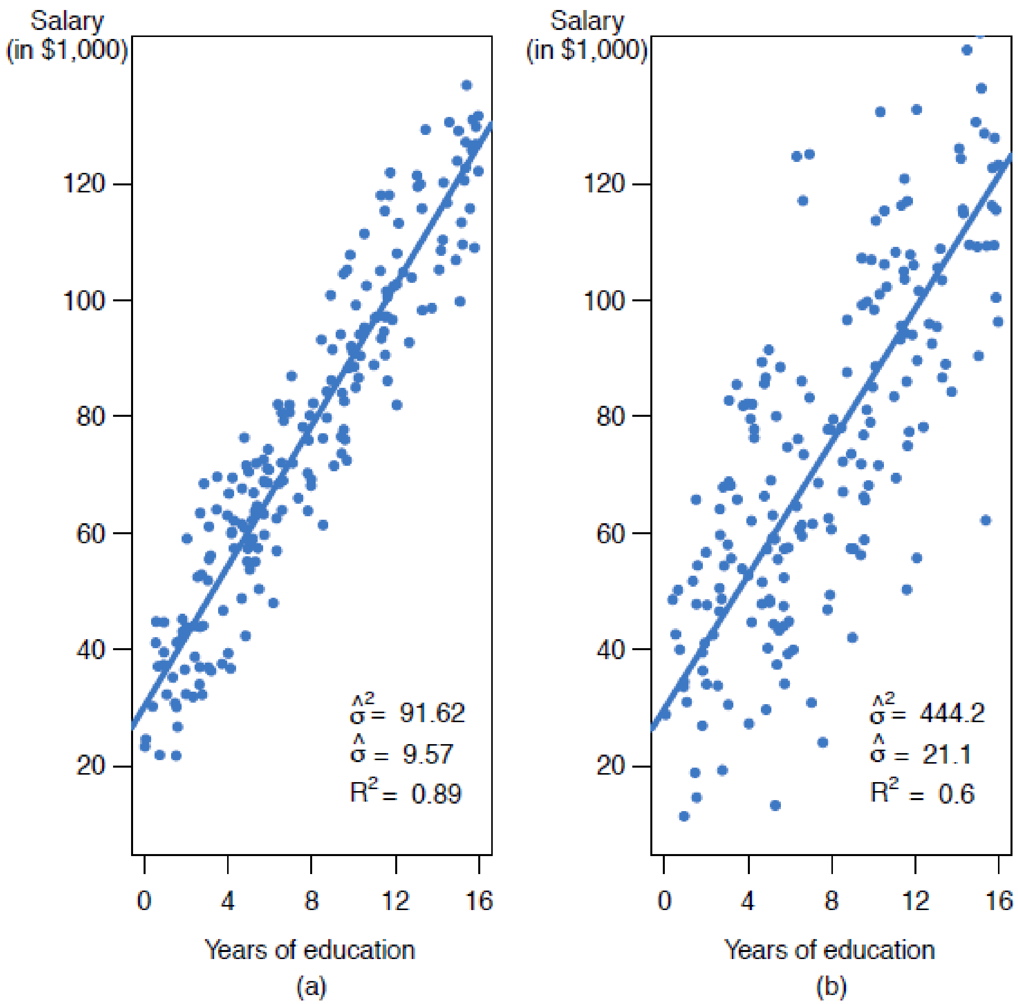
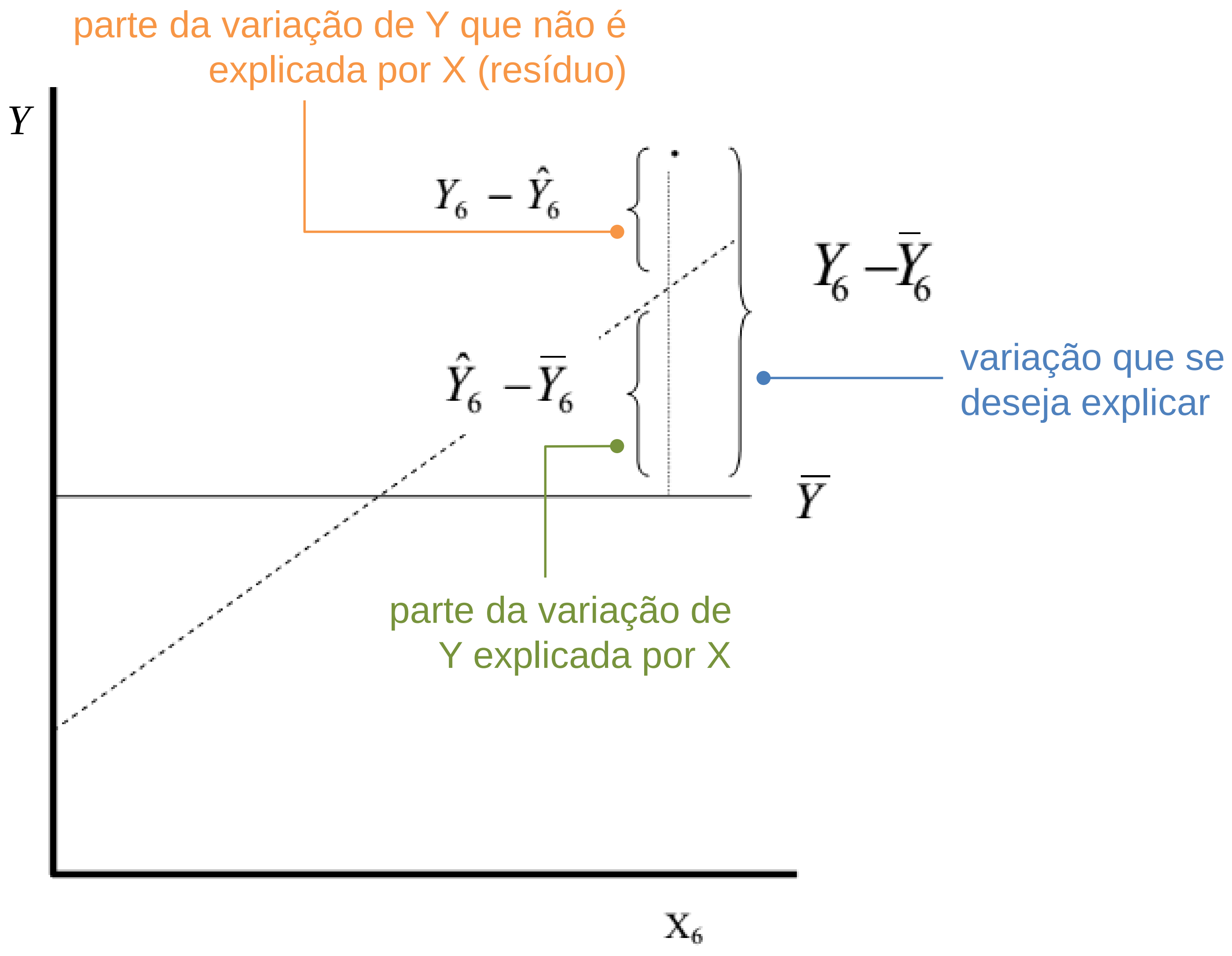


FIGURE 3.9: Plots with Different Goodness of Fit

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): mede a variação amostral total de Y; mede a dispersão de Y 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:

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

- A variação total de Y (SQT) corresponde à soma da variação explicada (SQE) e da variação não explicada (SQR):

$$SQT = SQE + SQR$$

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):

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

$$0 \le R^2 \le 1$$

- 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**

  – Nota: 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 **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

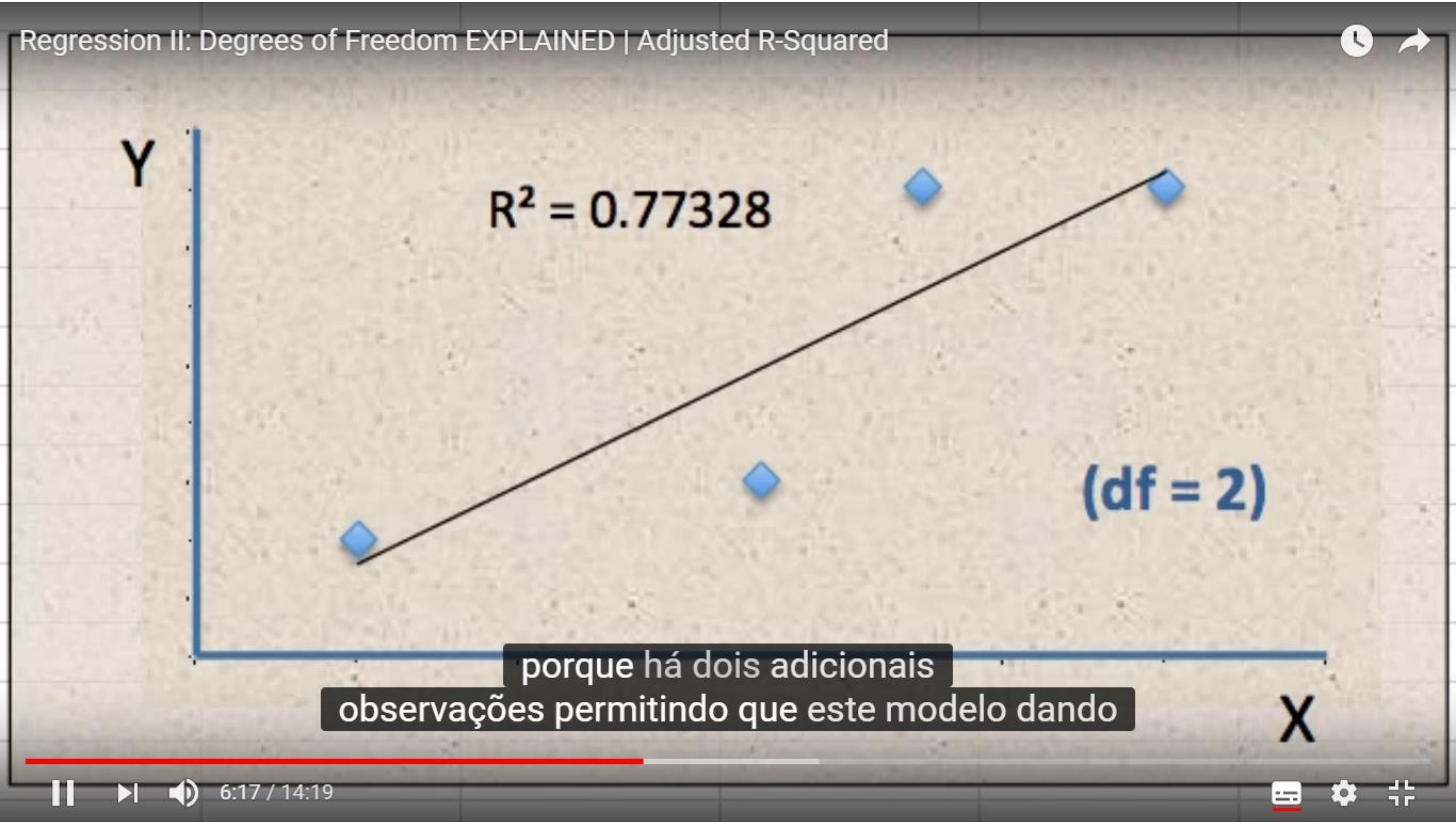
- O R^2 ajustado corrige o R^2 para refletir a perda de graus de liberdade que ocorre quando incluímos variáveis na regressão

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

- 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 $\leq R^2$
- 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

Interpretação dos coeficientes estimados

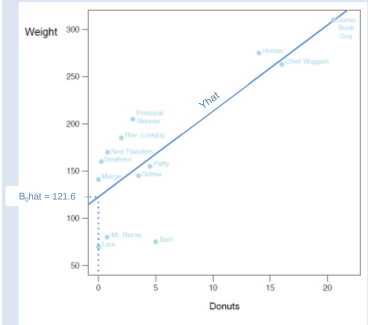


FIGURE 1.3: Regression Line for Weight and Donuts in Springfield

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:

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

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

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

```
> 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
```

Agenda para esta aula

1. Grau de ajuste

2. **Regressão multivariada: precisão**

Dois desafios à análise estatística inferencial: aleatoriedade e endogeneidade

Fontes de incerteza quanto ao efeito estimado de X sobre Y

Sampling randomness: amostras de diferentes tamanhos geram coeficientes estimados diferentes; amostras diferentes de um mesmo tamanho também geram coeficientes estimados diferentes; na estatística frequentista, coeficiente populacional é fixo)

Modeled randomness: aleatoriedade e complexidade na formação de Y redundam em variáveis omitidas; nota: aqui não estamos falando de variáveis omitidas correlacionadas com X

Variáveis omitidas correlacionadas com X: existência dessas variáveis implica espuriedade

Aleatoriedade (compromete a **precisão**)

Como a regressão múltipla aumenta a precisão das estimativas?

Endogeneidade (compromete a **acurácia**)

$$\hat{\beta}_j$$

(qualquer coeficiente de inclinação estimado)

Numa regressão <u>bivariada</u>, a variância de $\beta_1$hat* é dada por:

[Obs.: A variância de $\beta_0$hat é dada por outra fórmula.]

Recordatório

Erro padrão de $\beta_j$hat, o **se($\beta_j$hat)** = Raiz quadrada da var($\beta_j$hat)

$$var(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{n * var(X)}$$

Variância da regressão = (Residual standard error)^2. Obtenha o residual standard error no output da regressão

$\hat{\sigma}^2$

Variância da regressão

- Mede quão bem o modelo explica a variação de Y
- Seu cálculo baseia-se nos resíduos
- Aqui, k = número de variáveis explicativas
- É também uma estimativa da variância de ε
- **Intuição:** média do quadrado da distância entre valores observados e previstos de Y

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

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n}\hat{\epsilon}_i^2}{n - k - 1}$$

n

Tamanho da amostra

- **Intuição:** mais dados implicam menor variância, pois a chance de o acaso nos levar às caudas da distribuição de $\beta_1$hat é menor em amostras maiores (i.e., menor sampling randomness)

$var(X)$

Variância X (amostral)

- Quanto mais X variar, mais precisa será a distribuição de $\beta_1$hat
- **Intuição:** se X varia pouco, não temos muita informação para estimar o efeito da variação de X sobre a variação de Y

$$var(X) = s^2$$

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n - 1}$$

* Fórmula de var($\beta_1$hat) é mais complicada quando erros são correlacionados ou heteroscedásticos, mas as intuições sobre variância da regressão, tamanho da amostra e var(X) se aplicam. Voltaremos a esse ponto em aulas futuras.

Na regressão múltipla, var(βhat) é influenciada também pela correlação entre as variáveis explicativas

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

$\beta_1 = \Delta Y / \Delta(X_1 \mid X_2, ..., X_k$ são mantidos constantes$)$

- Para erros homoscedásticos e independentes entre si*, sendo X_j uma variável explicativa:

$$var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{n * var(X_j) * (1 - R_j^2)}$$

Observe o numerador: regressão múltipla tende a gerar estimativas mais precisas

- **$(1 – R_j^2)$** is the new kid on the block
- Wait... Hold on... We already know his face! Or, at least, the face of a close relative of his:
  – **R_j^2** é o **R^2** (coeficiente de determinação) de uma **regressão auxiliar** em que **X_j** é a variável explicada; todas as outras variáveis explicativas do modelo principal assumem o papel de variáveis explicativas da regressão auxiliar; **cada X_j produz uma regressão auxiliar diferente**

NEW KIDS ON THE BLOCK
Boy band americana formada em 1986; separou-se em 1994, retornando em 2008.

Regressões auxiliares não são modelos causais; servem apenas para apurar o grau de correlação entre as variáveis explicativas do modelo principal.

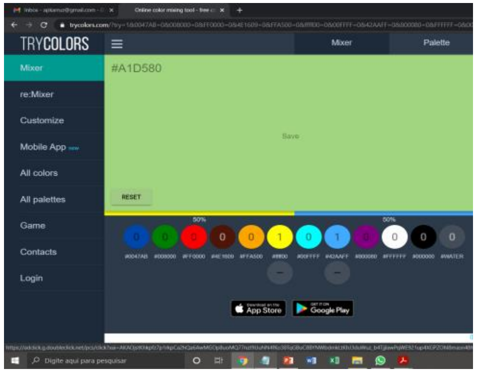
* Fórmula de var(βhat) é mais complicada quando erros são correlacionados ou heteroscedásticos, mas as intuições sobre variância da regressão, tamanho da amostra, var(X) e multicolineridade se aplicam.

var($\beta_j$hat) é maior quando X_j é bastante correlacionado com uma ou mais das outras variáveis explicativas

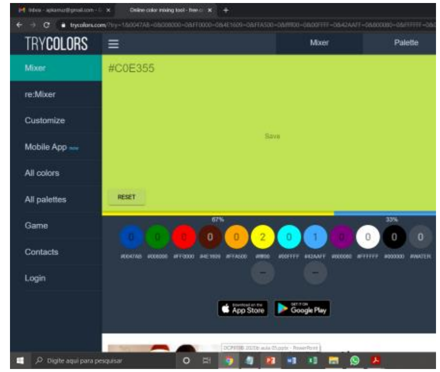
$$var(\hat{\beta}_j) = \frac{\hat{\sigma}^2}{n * var(X_j) * (1 - R_j^2)}$$

Esses ***R_j^2*** *nos dizem o* ***quanto as outras variáveis independentes explicam X_j****. Se as outras variáveis explicarem X_j muito bem, o R_j^2 será alto e – aqui está a sacada principal – o denominador será menor. Observe que o denominador da fórmula da var($\beta_j$hat) é (1 – R_j^2). Lembre-se de que qualquer R^2 está entre 0 e 1, então,* ***à medida que R_j^2 fica maior, 1 – R_j^2 diminui*** *o que por sua vez* ***faz var($\beta_j$hat) aumentar****. A intuição é que se a variável* ***X_j*** *for* ***praticamente indistinguível*** *das outras variáveis explicativas, fica mais* ***difícil dizer quanto X_j afeta Y*** *e teremos, portanto, uma maior* ***var($\beta_j$hat)****.*

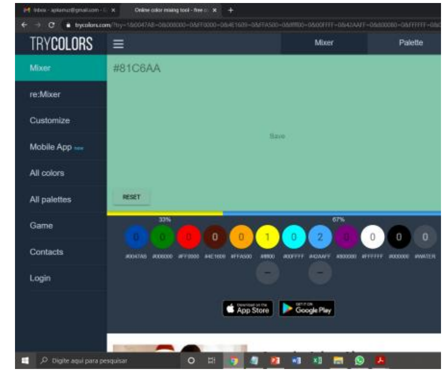
Bailey (2016: 227)



$Yhat = B_0hat + B_1hatAmarelo + B_2hatAzulRoyal$

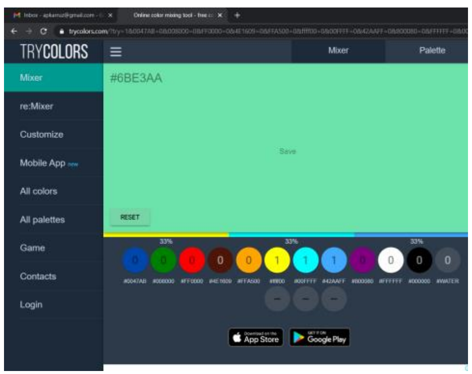


Aumenta Amarelo

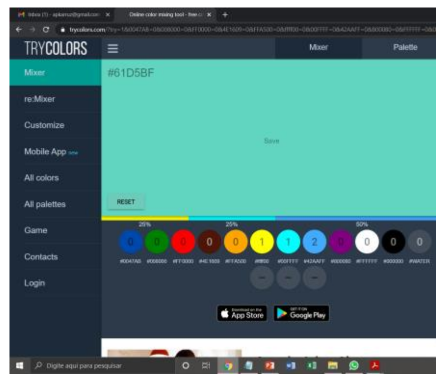


Aumenta AzulRoyal

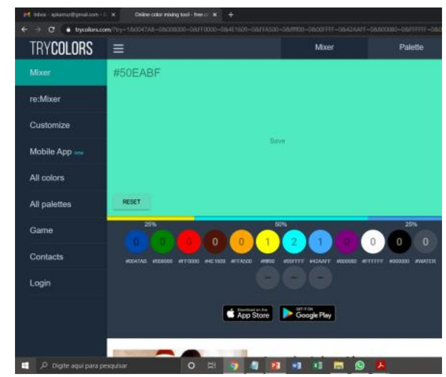
- Neste primeiro modelo, as duas cores usadas para prever Y (Y = tons de verde) são bastante distintas; o efeito da intensificação do Amarelo é bem diferente do efeito da intensificação do AzulRoyal



$Yhat = B_0hat + B_1hatAmarelo + B_2hatAzulRoyal + B_3hatAzulClaro$



Aumenta AzulRoyal



Aumenta AzulClaro

- No segundo modelo, duas das três cores usadas para prever Y são muito semelhantes entre si; tal semelhança traz imprecisão para a estimativa do efeito de cada uma dessas cores na composição de Y
- Ao adicionarmos AzulClaro, estamos demandando que o modelo estime mais um parâmetro, porém a informação “nova” trazida pela variável adicional é pouca: mantendo-se constante AzulRoyal, sobra pouca variação em AzulClaro (e vice-versa)

https://trycolors.com/

PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIA POLÍTICA | UFMG

DCP919B
Metodologia II

Aprofundamento em regressão multivariada

Aula 5
5 de outubro de 2022

Ana Paula Karruz