Correlação e Regressão

PRI5003 - Aula 10

Instituto de Relações Internacionais - Universidade de São Paulo

14 de Junho de 2018

Outline

O que vimos até agora

Nossas últimas aulas têm se concentrado em verificar a associação de variáveis:

- Aula 8: associação entre variáveis quantitativas e categóricas (testes de diferença de médias e de proporções)
- Aula 9: associação entre variáveis categóricas (qui-quadrado)
- Aula 10: associação entre variáveis quantitativas (correlação e regressão)

Introdução

Em ciências sociais, é frequente termos interesse em saber se duas variáveis quantitativas estão associadas:

- Quando um país reduz barreiras comerciais, aumenta sua taxa de crescimento?
- Candidatos que investem mais em campanha se elegem com mais frequência?
- Países que produzem mais riqueza têm menor taxa de mortalidade infantil?

Introdução

Definimos que duas variáveis estão associadas quando, ao mudar o valor de uma delas, encontramos variaçao também na segunda. Podemos, então, pensar na correlação como covariância.

O termo correlação significa relação em dois sentidos (co + relação), e é usado em estatística para designar a força que mantém unidos dois conjuntos de valores. A verificação da existência e do grau de relação entre as variáveis é o objeto de estudo da correlação.

Correlação

Relembrando a fórmula da variância

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

É o mesmo que

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

Correlação

Para encontrarmos a covariância entre duas variáveis, incluímos o segundo fenômeno de interesse

Covariância

$$\text{cov(x, y)} = \sum_{i=1}^{n} \frac{(x_i - \bar{x}) * (y_i - \bar{y})}{N}$$

Correlação

O índice de covariância que calculamos tem um problema: não está padronizado. Assim, não podemos comparar covariâncias entre variáveis com medidas muito desiguais.

Para normalizarmos a covariância, utilizamos o *coeficiente r de Pearson*. Basta dividir a covariância pelo desvio-padrão das duas variáveis.

r de Pearson

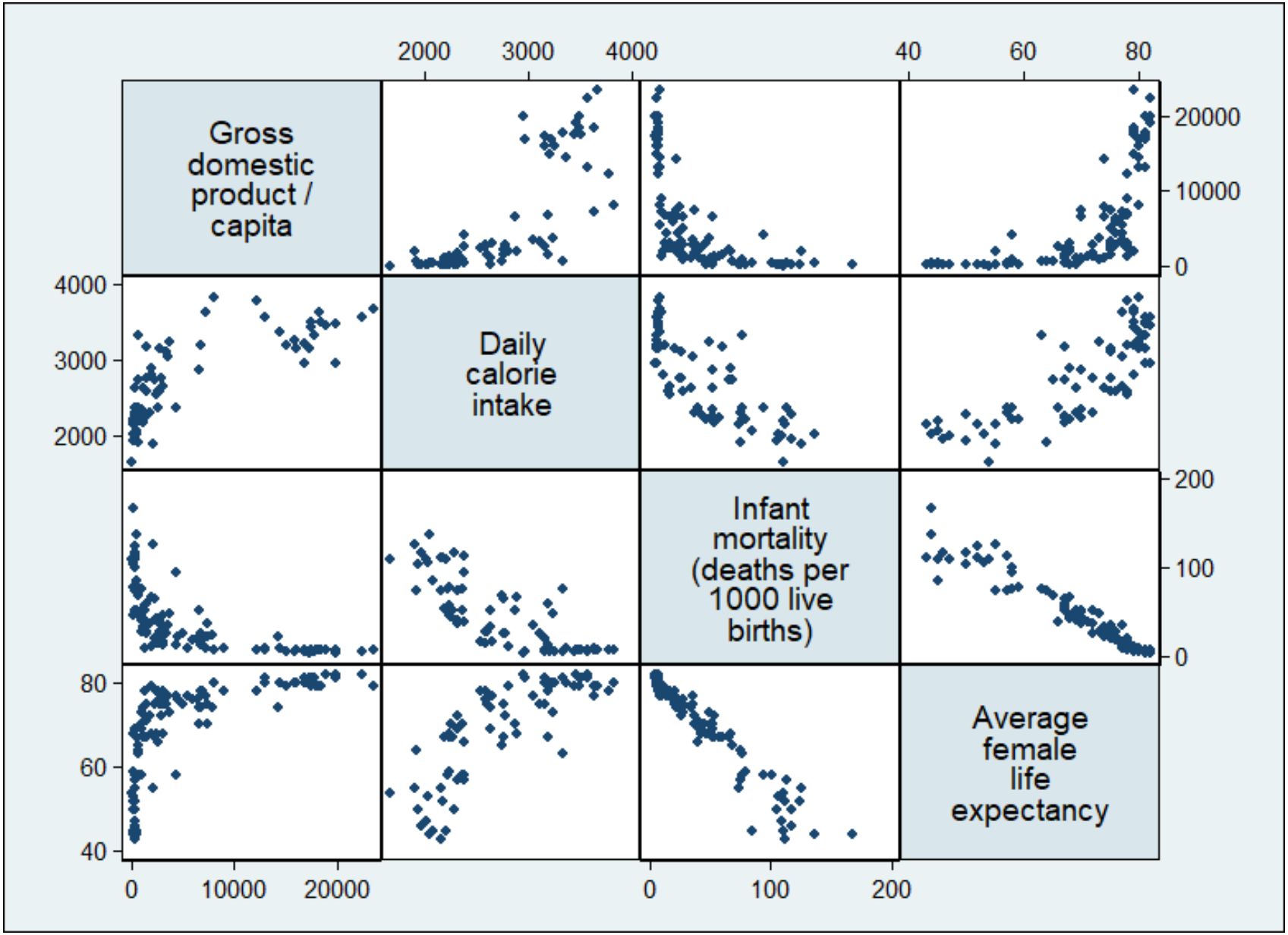
$$r = \sum_{i=1}^{n} \frac{cov(x, y)}{\sigma_x \sigma_y}$$

Coeficiente de Correlação

O coeficiente de correlação linear *r* varia de -1 a $+1$ e sua interpretação dependerá do valor numérico e do sinal:

- $r = 1 \longrightarrow$ correlação positiva perfeita
- $0 < r < 1 \longrightarrow$ correlação positiva
- $r = 0 \longrightarrow$ não há correlaçao (variáveis independentes)
- $-1 < r < 0 \longrightarrow$ correlação negativa
- $r = -1 \longrightarrow$ correlação negativa perfeita

Lembram da prova?



Teste da Correlação

Fazendo inferências sobre a associação entre variáveis

Quando estamos trabalhando com uma amostra, é interessantes testar a associação entre as variáveis na população. Vamos calcular se a correlação linear ρ é estatisticamente significante.

H_0: Não há correlaçao populacional ($\rho = 0$)
H_a: Há correlação populacional ($\rho \neq 0$)

Estatística-teste

$$t = r\sqrt{\frac{n-2}{1-r^2}}$$

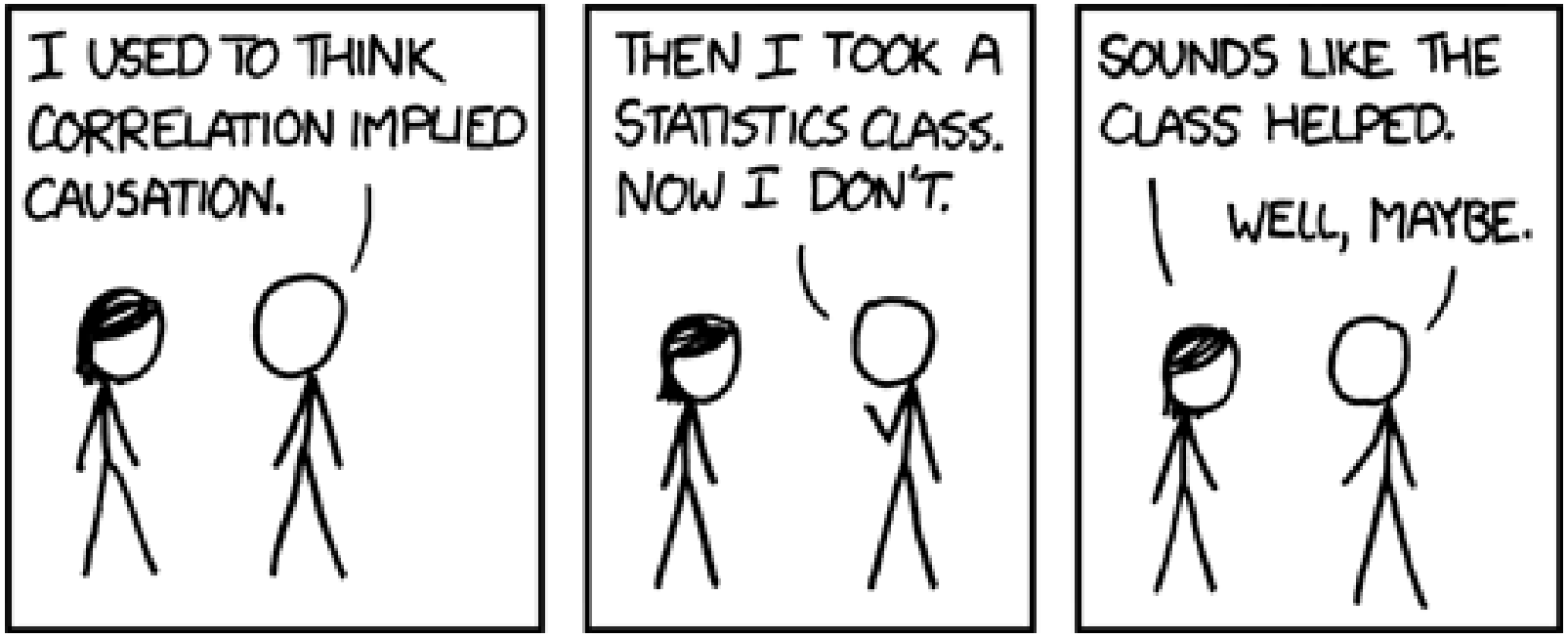
Correlação e causalidade

Quando encontramos associação entre duas variáveis, não conseguimos distinguir se:

1. X influencia Y
2. Y influencia X
3. X e Y se influenciam mutuamente
4. Z influencia X e Y (variável omitida)

A regressão dá o primeiro passo para corrigir isso.

Correlação e causalidade

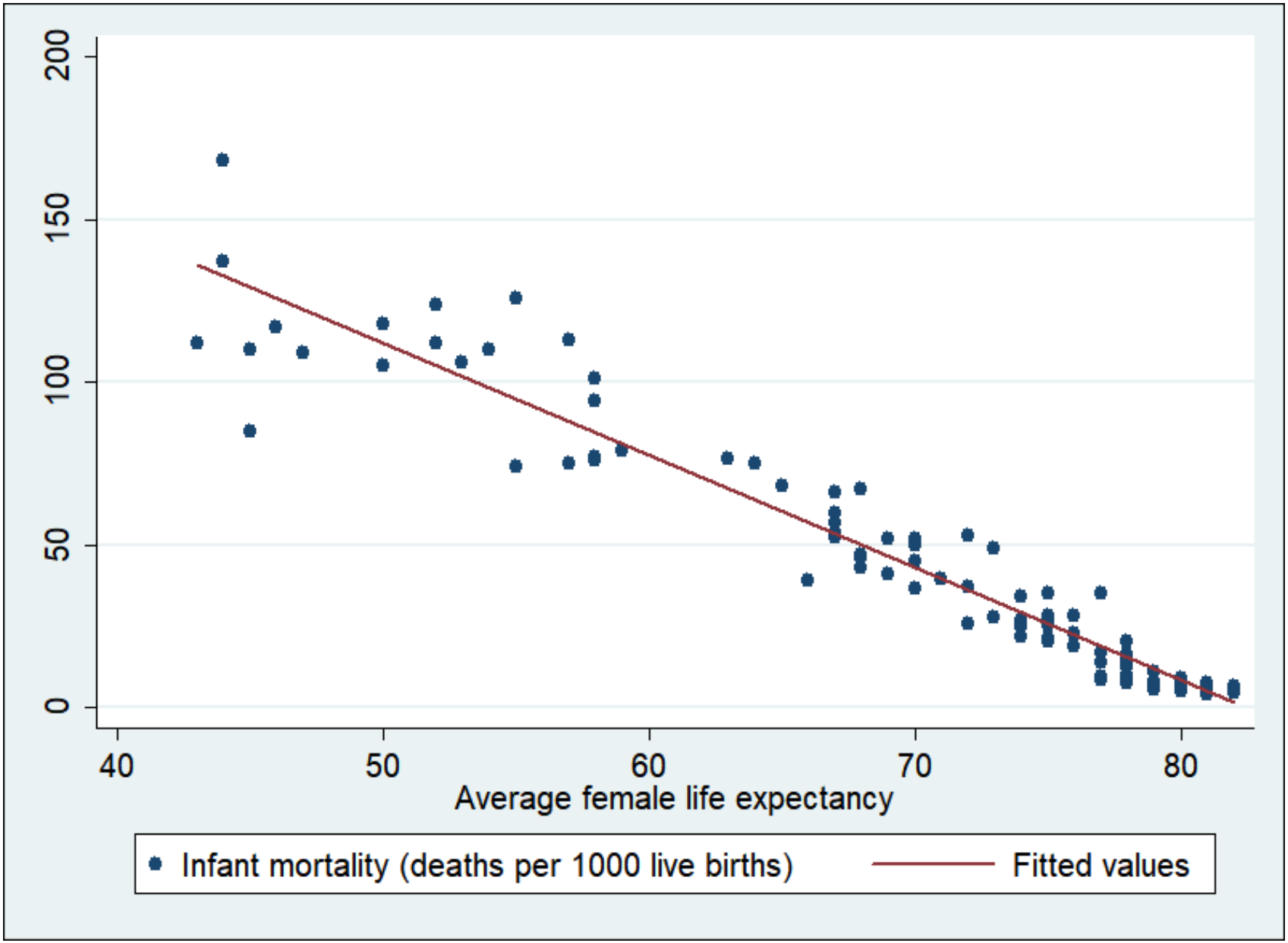


Regressão linear

Partindo de um gráfico de dispersão, o primeiro passo da regressão é ajustar uma reta que passe o mais próximo possível de todos os pontos. A função que gera essa reta tem o nome de *equação de regressão*.

A partir dessa reta, podemos fazer previsões para y a partir de valores específicos de x

Exemplo


200
150
100
50
0
40
50
60
70
80
Average female life expectancy
Infant mortality (deaths per 1000 live births)
Fitted values

Regressão linear

Equação da reta que você viu na escola

$$y = ax + b$$

Notação do modelo populacional

$$y = \alpha + \beta x + \epsilon$$

- y é a variável dependente;
- α é o intercepto;
- β é a inclinação da reta;
- x é a variável independente.

Ilustração

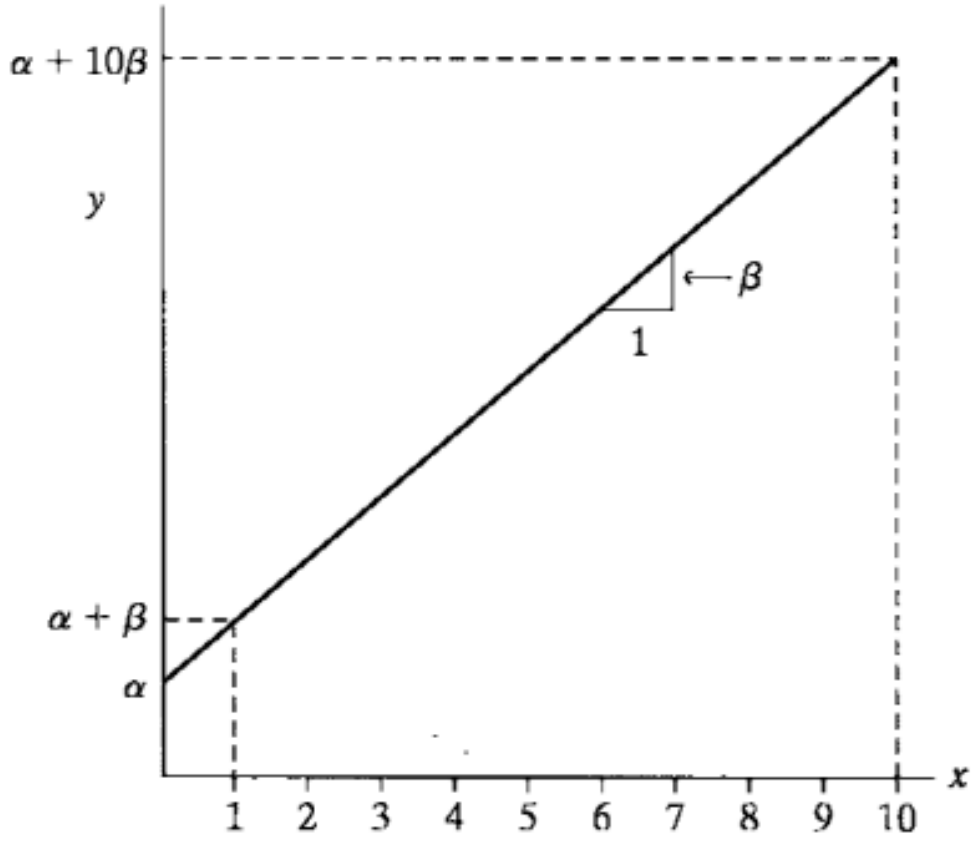


FIGURE 9.2: Graph of the Straight Line $y = \alpha + \beta x$. The y-intercept is α and the slope is β.

O que é o resíduo?

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

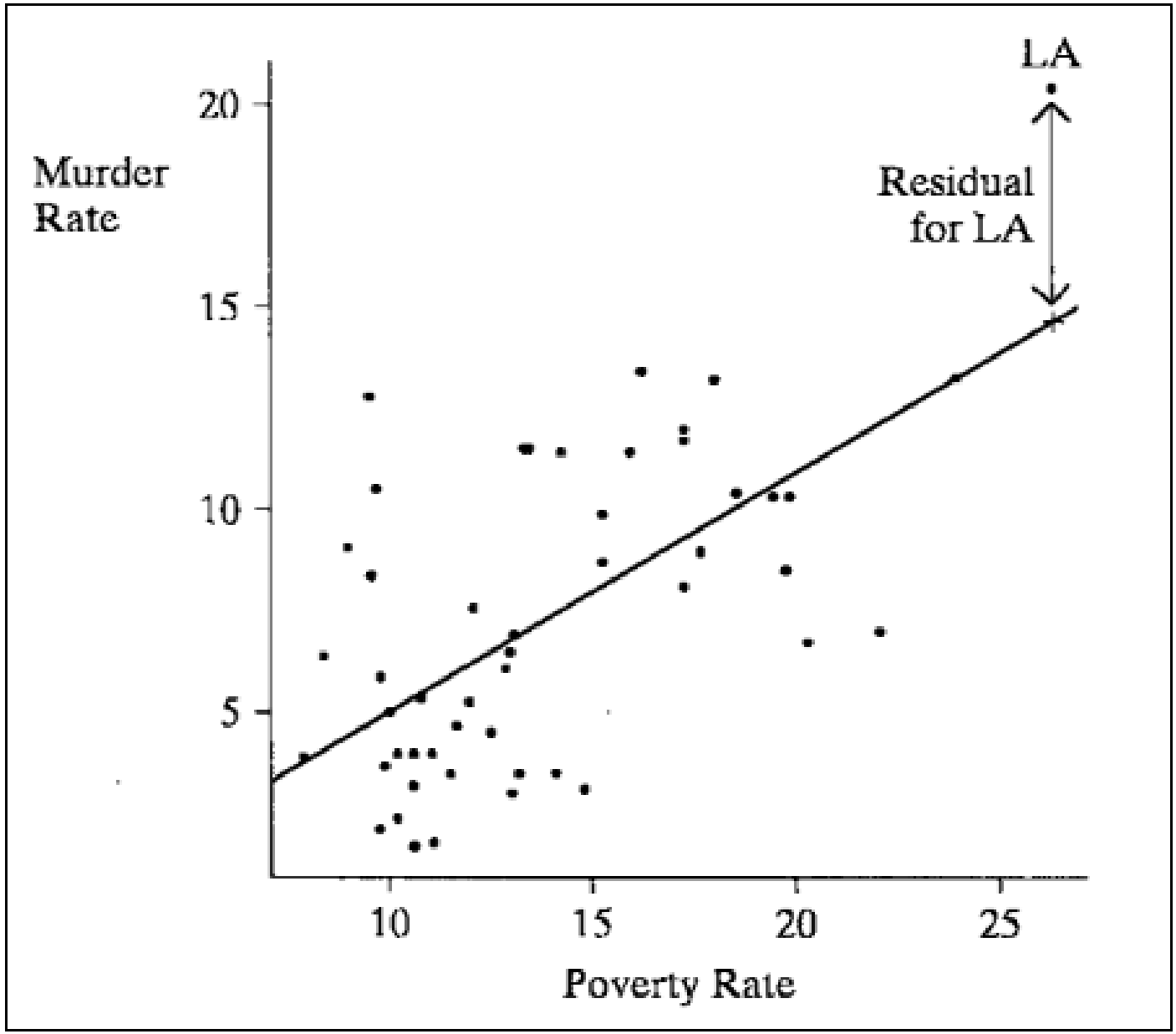
Em notação, resíduo$= y - \hat{y}$.

Para sabermos o desvio geral das obsevações 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


LA
Residual
for LA
Murder
Rate
20
15
10
5
10
15
20
25
Poverty Rate

E como calular os coeficientes?

Inclinação

$$\hat{\beta} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Intercepto

$$\hat{\alpha} = \bar{y} - \hat{\beta}\bar{x}$$

Como ler a saída do Stata

```
. reg VOTE GROWTH

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  1,    32) =   15.70
       Model |  385.31241      1  385.312461            Prob > F      =  0.0004
    Residual |  785.539343     32  24.5481045           R-squared     =  0.3291
-------------+------------------------------           Adj R-squared =  0.3081
       Total |  1170.8518      33  35.4803577           Root MSE      =  4.9546

------------------------------------------------------------------------------
       VOTE |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
     GROWTH |   .6249078   .1577315     3.96   0.000     .3036193    .941963
      _cons |   51.50816   .8569026    60.11   0.000     49.76271    53.25361
------------------------------------------------------------------------------
```

1. Quais são a VD e a VI?
2. Identifique α e β
3. Onde está o SQE?
4. Quais são as hipóteses sendo testadas?

Teste de hipóteses na regressão

- $H_0 : \beta = 0$
- $n - k$ são os graus de liberdade. n é o número de observações no banco, k é o número de variáveis
- Cuidado com bancos de dados com muitas variáveis e poucas observações!
- $n > k$
- O teste de hipóteses para α segue a mesma lógica

Estatística-teste

$$t_{n-k} = \frac{\hat{\beta} - \beta_0}{ep(\hat{\beta})}$$

Root Mean-Squared Error

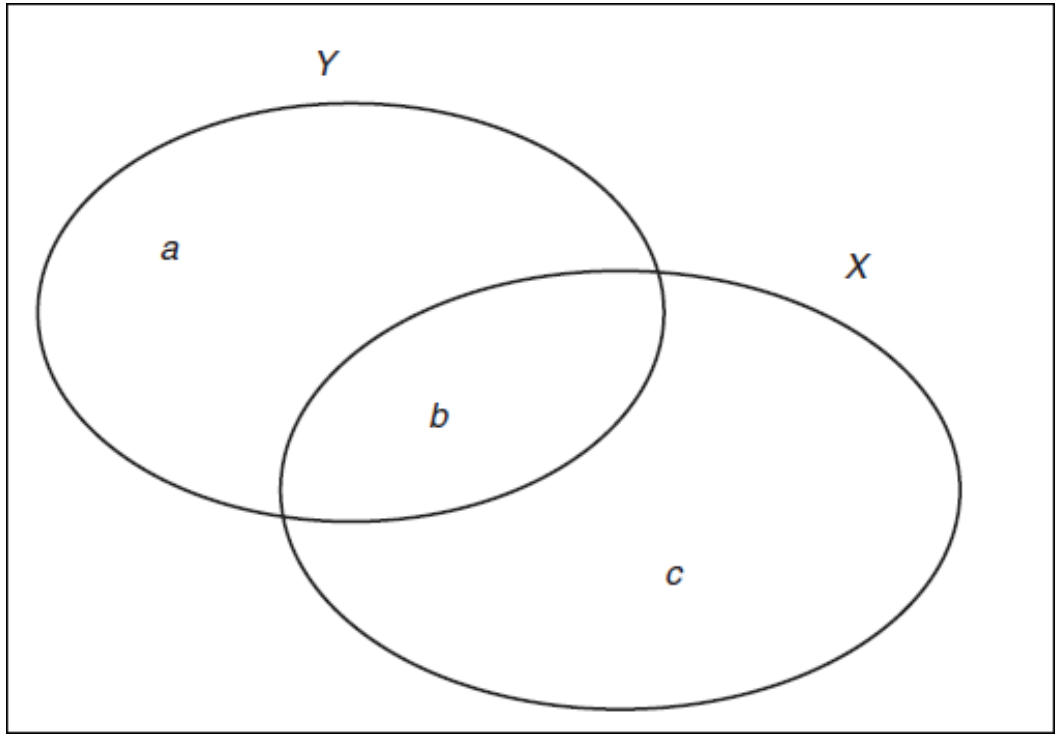
Root MSE

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

- O nome parece complicado
- A fórmula parece complicada
- O conceito é simples: é a distância média entre os pontos e a reta de ajuste
- Quanto menor o valor, melhor o ajuste

R^2

A imagem abaixo ilustra a intuição do r^2



- a é a variação residual de y
- b é a variação conjunta de x e y
- Quanto maior a área de b e menor a área de a, maior será o R^2. Em consequência, melhor será o ajuste (mas não é tão simples).

R^2

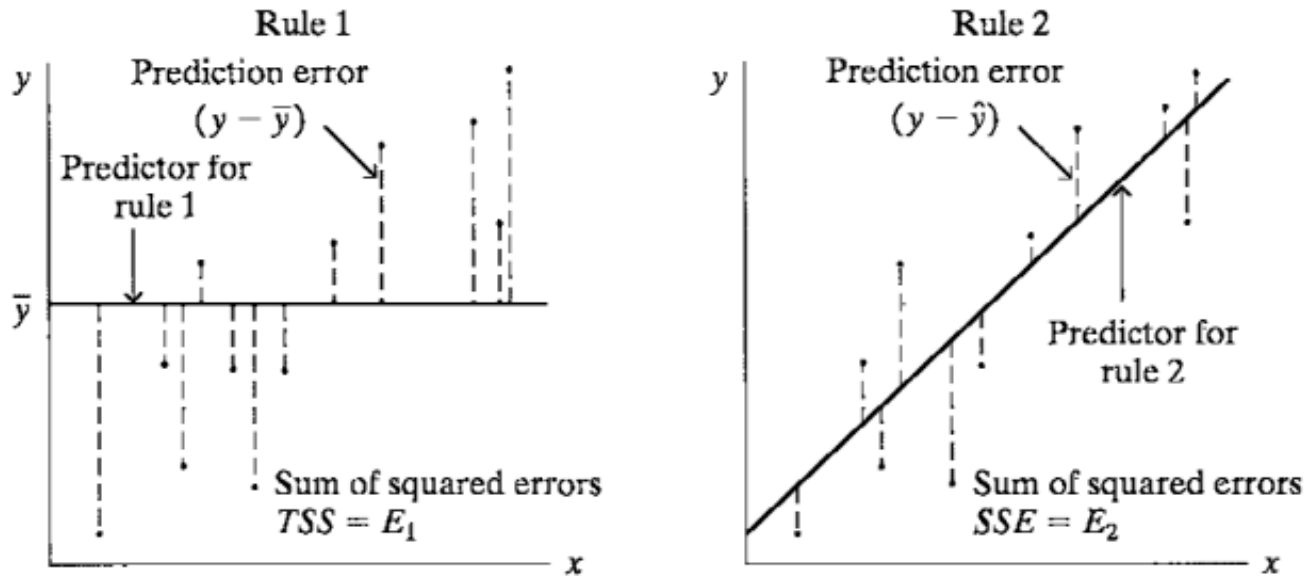


FIGURE 9.13: Graphical Representation of Rule 1 and Total Sum of Squares $E_1 = TSS = \Sigma(y - \bar{y})^2$, Rule 2 and Residual Sum of Squares $E_2 = SSE = \Sigma(y - \hat{y})^2$

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

R^2 - Coeficiente de determinação

$$r^2 = \frac{SQT - SQE}{SQT}$$

Soma dos quadrados totais

$$SQT = \sum (y - \bar{y})^2$$

Soma dos quadrados dos erros

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

Voltando ao Stata

```
. reg VOTE GROWTH

      Source |       SS       df       MS              Number of obs =      34
-------------+------------------------------           F(  1,    32) =   15.70
       Model |  385.31241      1  385.312461            Prob > F      =  0.0004
    Residual |  785.539343     32  24.5481045           R-squared     =  0.3291
-------------+------------------------------           Adj R-squared =  0.3081
       Total |  1170.8518     33  35.4803577           Root MSE      =  4.9546

------------------------------------------------------------------------------
       VOTE |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
------------+----------------------------------------------------------------
     GROWTH |   .6249078   .1577315     3.96   0.000     .3036193    .941963
      _cons |   51.50816   .8569026    60.11   0.000     49.76271    53.25361
------------------------------------------------------------------------------
```

- Se SQT e SQE forem próximos, o numerador será baixo e, consequentemente, o ajuste será baixo
- Quanto mais distante o SQE estiver do SQT, maior será o numerador, e maior será o ajuste