Regressão para pesquisas sociais

Dia 1 - Exercício 1: correlação

Thiago Cordeiro Almeida

September 14, 2025

0..1 Introdução

Vamos trabalhar neste documento com dados oriundos do site Correlações Espúrias.

Em especial, vamos explorar a seguinte questão: **Qual a relação entre o consumo per capita de margarina nos EUA e a taxa de divórcio no estado de Maine (EUA)??**

Seguindo as etapas que destacamos nos slides, começaremos reproduzindo uma exploração dos dados e, posteriormente, entraremos em detalhes sobre a relação entre as variáveis e, por fim, testando a correlação entre elas.

0..2 Importação dos dados

Ao longo dessas análises, usaremos a gramática de programação `tidyverse`. Caso você ainda não esteja familiarizado/a com essa forma de programação, sugiro a leitura do material da Curso-R que deixei nas referências do curso, há um capítulo sobre isso. Eles, com certeza, te convencerão sobre a potencialidade dessa forma de programção.

Dentro do `tidyverse`, há distintas bibliotecas. Usaremos duas chamadas `dplyr` – para manipulação dos dados – e `ggplot2` – para gráficos.

```
# ajustes gerais
options(scipen = 99999)

## pacotes necessarios
# instalando pacote de gerenciador de pacotes, pacman
ifelse(!require(pacman),install.packages("pacman"),require(pacman))
```

```
[1] TRUE
```

```r
p_load(tidyverse) # importando pacote que usaremos, tidyverse

# importando dados
base <- read_csv2("dados_correlacao.csv")
anos = seq(from = 2000,to = 2009,by = 1) # anos
label_margarina = "Consumo per capita de margarina nos EUA" # rótulo da variavel
label_divorcio = "Taxa de divórcio no estado de Maine (EUA)" # rótulo da variavel
```

Agora que criamos as variáveis com os dados, vamos criar uma base

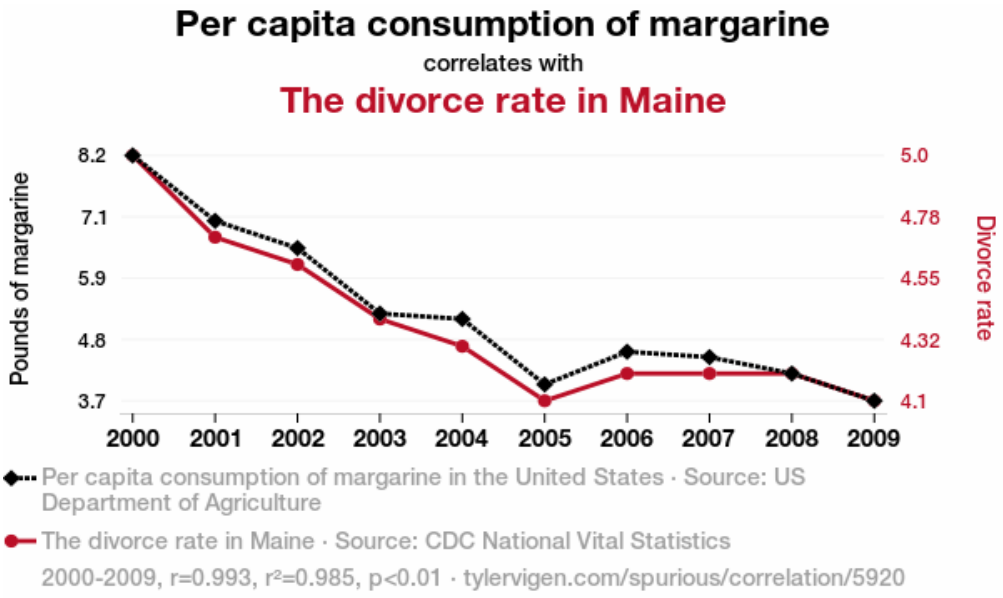
```r
## criando base de dados com essas informações
# obs: vamos criar uma base de classe 'tibble'
# por ser mais facil de trabalhar com pipelines

# base <- tibble(
#   margarina,
#   divorcio,
#   anos
# )

base <- as.tibble(base)
```

0..3 Exploração dos dados

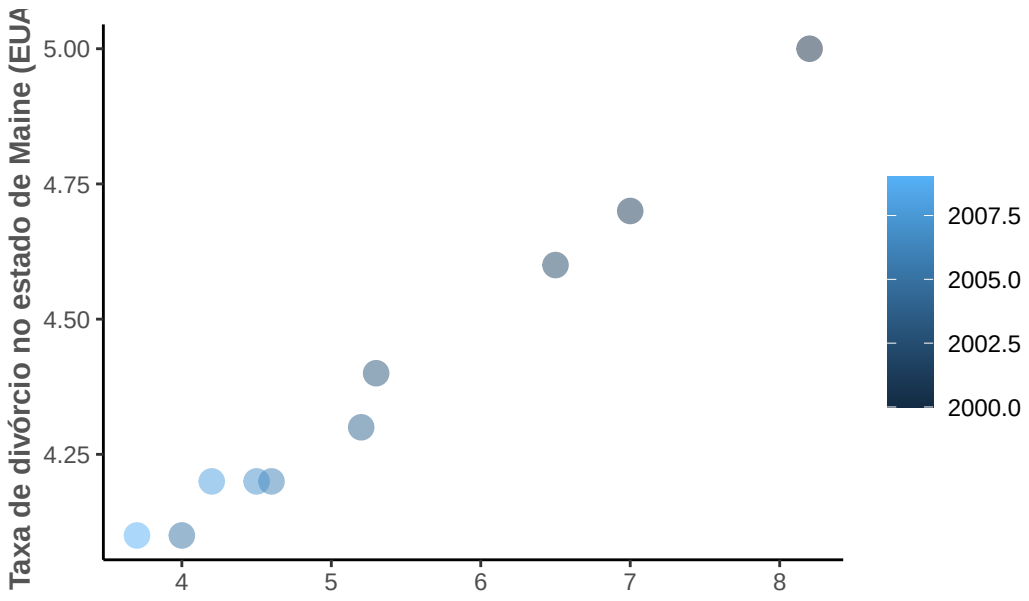
0..3.1 Distribuição ao longo do tempo para ambas as variáveis



Bom, parece que elas se distribuem de modo similar ao longo dos anos. Vamos relacionar ambas para termos maior certeza disso?

0..3.2 Distribuição de ambas as variáveis: gráfico de dispersão

```
base |>
  ggplot() +
  aes(x = margarina, y = divorcio, color = anos) +
  geom_point( size = 4, alpha = .5) +
  labs(
    x = label_margarina,
    y = label_divorcio,
    caption = "Fonte: https://www.tylervigen.com/spurious/correlation/5920"
  ) +
  theme_classic() +
  theme(
    legend.title = element_blank(),
    axis.title.y = element_text(color = "grey33",face = "bold"),
    axis.title.x = element_text(color = "red4",face = "bold")
  )
```



Aparentemente, temos um bom sinal de que descobrimos algo! elas não só aparentam ter uma relação muito semelhante ao longo do tempo, como também apresentam mudanças que segue uma mesma distribuição ao longo dos anos!!!

Para confirmar nossa exploração dos dados, vamos fazer o teste de correlação estatística, utilizando a correlação de Pearson.

0..3.3 Análise de correlação entre as duas variáveis

```r
cor.test(base$margarina, base$divorcio) # teste de correlação, default 'Pearson'
```

```
    Pearson's product-moment correlation

data:  base$margarina and base$divorcio
t = 23.055, df = 8, p-value = 0.0000000133
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.9676666 0.9983038
sample estimates:
      cor
0.9925585
```

Uau! Há uma relação quase completa entre elas, com um nível de significância que nos permite dizer que, estatisticamente, elas apresentam uma relação positiva.

Como explicar isso? Façam suas hipóteses!