UNIVERSIDADE FEDERAL DE PELOTAS - UFPEL
PROGRAMA DE PÓS GRADUAÇÃO EM ORGANIZAÇÕES E MERCADO - PPGOM

Aplicação de Econometria Básica no R

Sérgio Alves Daneris

sergiodanerisalves@gmail.com

SOBRE MIM

Formação Acadêmica:

- Graduação em Economia - (UNIFRAN)
- Mestre em Economia Aplicada - (PPGOM - UFPEL)
- Doutorando em Economia Aplicada - (PPGOM - UFPEL)
- Graduação (Em Andamento) em Física - (UNIFRAN)

Grupo de Pesquisa:

- GAPPS - Grupo de Avaliação de Políticas Públicas (PPGOM- UFPEL)

Linhas de Pesquisa:

- Economia da Saúde
- Economia do Meio Ambiente
- Energy Economics

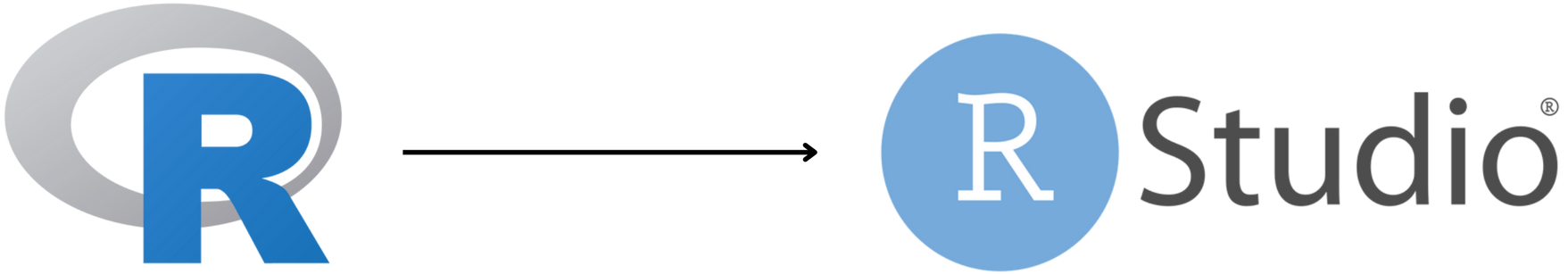
Repositório de Materiais da Aula:

- https://github.com/sergiodaneris

INDICE

INTRODUÇÃO AO SOFTWARE R E IDE RSTUDIO

- R é uma **linguagem e ambiente** para computação estatística e gráficos, sendo assim um sistema de **armazenamento e manipulação** eficiente de dados.
- O software oferece uma **ampla variedade de técnicas estatísticas**: modelagem linear e não-linear, testes estatísticos clássicos, séries temporais, classificação, agrupamentos (clustering), entre outras.
- O RStudio é uma IDE (Ambiente de Desenvolvimento Integrado) que oferece uma **interface intuitiva** e um **conjunto de ferramentas** que facilitam o uso do software R.



PROCESSO DE INSTALAÇÃO

Links para download

- Software R (Windows) → https://cran-r.c3sl.ufpr.br/bin/windows/base/
- Software R (Linux) → https://cran-r.c3sl.ufpr.br/bin/linux/
- Software R (MacOS) → https://cran-r.c3sl.ufpr.br/bin/macosx/
- R Studio → https://posit.co/download/rstudio-desktop/

3

FORMATOS DE DATAFRAMES

Formatos de dataframes suportados no R:

- CSV (Comma-Separated Values): é um formato de **texto simples**, onde os valores das colunas são **separados por vírgulas (ou ponto e vírgula).**
- XLS / XLSX (Microsoft Excel Spreadsheet): é o formato padrão das **planilhas do Microsoft Excel**.
- DTA (Stata Data File): é o formato nativo do software **Stata**.
- SAV: Arquivos do **SPSS (software estatístico).**
- RDS: Formato **nativo do R.**
- JSON: Arquivo em formato **texto estruturado (usado em APIs).**
- ODS: Formato de planilhas do **LibreOffice.**

4

PACOTES DE LEITURA DE DADOS

Instalação e Carregamento de Pacotes no R

1- install.packages(‘pacote’)

2- library(‘pacote’)

Leitura de base de dados:

read.csv: Nativo do R. **(Não necessita de pacote)**
Ex: dados <- read.csv("dados.csv")

readxl: Lê arquivos **Excel (.xls e .xlsx)**
Ex: dados <- read_excel("dados.xlsx", sheet = "Planilha1")

haven: Lê arquivos de softwares estatísticos como **Stata, SPSS e SAS.**
Ex: dados <- read_dta("arquivo.dta")

jsonlite: Lê e escreve arquivos **JSON.**
Ex: dados <- fromJSON("dados.json")

BASE DE DADOS 1

Base de dados utilizada como exemplo: Cattaneo 2

A base foi construída para analisar o efeito do cuidado pré-natal adequado sobre o peso do bebê ao nascer.

Leitura da base de dados:

```
library('haven')
library('readxl')

cattaneo2CSV ← read.csv('C:/Users/sergi/Desktop/Faculdade/Semana Academica Economia/Curso Sergio/base de dados/cattaneo2.csv')

cattaneo2XLS ← read_excel('C:/Users/sergi/Desktop/Faculdade/Semana Academica Economia/Curso Sergio/base de dados/cattaneo2.xlsx')

cattaneo2DTA ← read_dta('C:/Users/sergi/Desktop/Faculdade/Semana Academica Economia/Curso Sergio/base de dados/cattaneo2.dta')
```

BASE DE DADOS 1

Comandos para explorar, entender e resumir uma base de dados:

head() <- Mostra as primeiras linhas do conjunto de dados.

Ex: head(dados)

dim() <- Mostra as dimensões da base: número de linhas e colunas.

Ex : dim(dados)

nrow() ← Mostra quantas linhas existem.

Ex: nrow(dados)

ncol() <- Mostra quantas colunas existem.

Ex: ncol(dados)

str() <- Mostra a estrutura da base (tipo de cada variável, valores iniciais, etc).

Ex: str(dados)

BASE DE DADOS 1

Comandos para explorar, entender e resumir uma base de dados:

summary() <- Mostra um resumo estatístico (média, mínimo, máximo, mediana, etc.) de cada variável.

Ex: summary(dados)

class() <- Mostra o tipo de objeto (ex: data.frame, tibble, etc.).

Ex: class(dados)

unique() <- Lista os valores únicos de uma variável.

Ex: unique(dados$estado)

is.na() + sum() <- Conta quantos valores faltantes (NA) existem.

Ex: sum(is.na(dados))

BASE DE DADOS 1

```
head(cattaneo2CSV)
```

```
##   bweight mmarried mhisp fhisp foreign alcohol deadkids mage medu fage fedu
## 1    3459        1     0     0       0       0        0   24   14   28   16
## 2    3260        0     0     0       1       0        0   20   10    0    0
## 3    3572        1     0     0       1       0        0   22    9   30    9
## 4    2948        1     0     0       0       0        0   26   12   30   12
## 5    2410        1     0     0       0       0        0   20   12   21   14
## 6    3147        0     0     0       0       0        0   27   12   40   12
##   nprenatal monthslb order msmoke mbsmoke mrace frace prenatal birthmonth
## 1        10       30     2      0       0     1     1        1          12
## 2         6       42     3      0       0     0     0        1           7
## 3        10       17     3      0       0     1     1        1           3
## 4        10       34     2      0       0     1     1        1           1
## 5        12        0     1      0       0     1     1        1           3
## 6         9        0     1      0       0     1     1        1           4
##   lbweight fbaby prenatal1
## 1        0     0         1
## 2        0     0         1
## 3        0     0         1
## 4        0     0         1
## 5        1     1         1
## 6        0     1         1
```

BASE DE DADOS 1

```
str(cattaneo2CSV)
```

```
## 'data.frame':    4642 obs. of  23 variables:
##  $ bweight   : int  3459 3260 3572 2948 2410 3147 3799 3629 2835 3880 ...
##  $ mmarried  : int  1 0 1 1 1 0 1 1 1 1 ...
##  $ mhisp     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ fhisp     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ foreign   : int  0 1 1 0 0 0 0 0 0 0 ...
##  $ alcohol   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ deadkids  : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mage      : int  24 20 22 26 20 27 27 24 21 30 ...
##  $ medu      : int  14 10 9 12 12 12 12 12 12 15 ...
##  $ fage      : int  28 0 30 30 21 40 29 33 24 33 ...
##  $ fedu      : int  16 0 9 12 14 12 14 12 9 15 ...
##  $ nprenatal : int  10 6 10 10 12 9 16 11 20 9 ...
##  $ monthslb  : int  30 42 17 34 0 0 29 0 0 27 ...
##  $ order     : int  2 3 3 2 1 1 3 1 1 2 ...
##  $ msmoke    : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mbsmoke   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ mrace     : int  1 0 1 1 1 1 1 1 1 1 ...
##  $ frace     : int  1 0 1 1 1 1 1 1 1 1 ...
##  $ prenatal  : int  1 1 1 1 1 1 1 1 1 1 ...
##  $ birthmonth: int  12 7 3 1 3 4 12 6 6 12 ...
##  $ lbweight  : int  0 0 0 0 1 0 0 0 0 0 ...
##  $ fbaby     : int  0 0 0 0 1 1 0 1 1 0 ...
##  $ prenatal1 : int  1 1 1 1 1 1 1 1 1 1 ...
```

```
summary(cattaneo2CSV)
```

```
##     bweight        mmarried          mhisp             fhisp
##  Min.   : 340   Min.   :0.0000   Min.   :0.00000   Min.   :0.00000
##  1st Qu.:3033   1st Qu.:0.0000   1st Qu.:0.00000   1st Qu.:0.00000
##  Median :3390   Median :1.0000   Median :0.00000   Median :0.00000
##  Mean   :3362   Mean   :0.6997   Mean   :0.03404   Mean   :0.03705
##  3rd Qu.:3725   3rd Qu.:1.0000   3rd Qu.:0.00000   3rd Qu.:0.00000
##  Max.   :5500   Max.   :1.0000   Max.   :1.00000   Max.   :1.00000
##     foreign           alcohol           deadkids           mage
##  Min.   :0.00000   Min.   :0.00000   Min.   :0.0000   Min.   :13.0
##  1st Qu.:0.00000   1st Qu.:0.00000   1st Qu.:0.0000   1st Qu.:22.0
##  Median :0.00000   Median :0.00000   Median :0.0000   Median :26.0
##  Mean   :0.05343   Mean   :0.03231   Mean   :0.2594   Mean   :26.5
##  3rd Qu.:0.00000   3rd Qu.:0.00000   3rd Qu.:1.0000   3rd Qu.:30.0
##  Max.   :1.00000   Max.   :1.00000   Max.   :1.0000   Max.   :45.0
##       medu             fage             fedu         nprenatal
##  Min.   : 0.00   Min.   : 0.00   Min.   : 0.00   Min.   : 0.00
##  1st Qu.:12.00   1st Qu.:24.00   1st Qu.:12.00   1st Qu.: 9.00
##  Median :12.00   Median :28.00   Median :12.00   Median :11.00
##  Mean   :12.69   Mean   :27.27   Mean   :12.31   Mean   :10.76
##  3rd Qu.:14.00   3rd Qu.:33.00   3rd Qu.:14.00   3rd Qu.:13.00
##  Max.   :17.00   Max.   :60.00   Max.   :17.00   Max.   :40.00
##     monthslb          order            msmoke          mbsmoke
##  Min.   :  0.00   Min.   : 0.000   Min.   :0.0000   Min.   :0.0000
##  1st Qu.:  0.00   1st Qu.: 1.000   1st Qu.:0.0000   1st Qu.:0.0000
##  Median : 13.00   Median : 2.000   Median :0.0000   Median :0.0000
##  Mean   : 23.07   Mean   : 1.892   Mean   :0.3996   Mean   :0.1861
##  3rd Qu.: 35.00   3rd Qu.: 2.000   3rd Qu.:0.0000   3rd Qu.:0.0000
##  Max.   :272.00   Max.   :12.000   Max.   :3.0000   Max.   :1.0000
##      mrace            frace           prenatal       birthmonth
##  Min.   :0.0000   Min.   :0.0000   Min.   :0.000   Min.   : 1.00
##  1st Qu.:1.0000   1st Qu.:1.0000   1st Qu.:1.000   1st Qu.: 4.00
##  Median :1.0000   Median :1.0000   Median :1.000   Median : 7.00
##  Mean   :0.8406   Mean   :0.8137   Mean   :1.202   Mean   : 6.54
##  3rd Qu.:1.0000   3rd Qu.:1.0000   3rd Qu.:1.000   3rd Qu.: 9.00
##  Max.   :1.0000   Max.   :1.0000   Max.   :3.000   Max.   :12.00
##     lbweight           fbaby          prenatal1
##  Min.   :0.00000   Min.   :0.000   Min.   :0.0000
##  1st Qu.:0.00000   1st Qu.:0.000   1st Qu.:1.0000
##  Median :0.00000   Median :0.000   Median :1.0000
##  Mean   :0.06032   Mean   :0.438   Mean   :0.8014
##  3rd Qu.:0.00000   3rd Qu.:1.000   3rd Qu.:1.0000
##  Max.   :1.00000   Max.   :1.000   Max.   :1.0000
```

PACOTES DPLYR E LUBRIDATE

Dplyr: O dplyr é uma gramática de manipulação de dados que permite realizar operações em data frames de forma simples e encadeada

Comandos:

group_by() ← Agrupa os dados por uma ou mais variáveis, **Ex: group_by(dados, sexo).**

filter() ← Filtra linhas com base em condições, **Ex: filter(dados, idade > 18).**

mutate() ← Cria ou modifica variáveis, **Ex: mutate(dados, idade_ao_quadrado = idade^2).**

summarise() ← Resume dados (geralmente com médias, somas etc.), **Ex: summarise(dados, media_idade = mean(idade)).**

inner_join() ← Mantém apenas as observações correspondentes nas duas bases.
Ex: inner_join(base1, base2, by = "id")

PACOTES DPLYR E LUBRIDATE

left_join() ← Mantém todas as linhas da base da esquerda, completando com dados da direita quando houver correspondência.
Ex: left_join(base1, base2, by = "id")

right_join() ← Mantém todas as linhas da base da direita.
Ex: right_join(base1, base2, by = "id")

full_join() ← Mantém todas as linhas de ambas as bases, mesmo sem correspondência.
Ex: full_join(base1, base2, by = "id")

semi_join() ← Mantém apenas as linhas da base da esquerda que têm correspondência na base da direita (sem adicionar colunas).
Ex: semi_join(base1, base2, by = "id")

anti_join() ← Mantém as linhas da base da esquerda que não têm correspondência na base da direita.
Ex: anti_join(base1, base2, by = "id")

PACOTES DPLYR E LUBRIDATE

Lubridate: Facilita o trabalho com datas e horários no R, permitindo extrair o ano, mês ou dia de uma data, converter strings em objetos de data reconhecidos pelo R, além de trabalhar com fusos horários, intervalos e durações.

Comandos:

ymd() <- Converte uma string no formato ano-mês-dia para data / Cria Variável

de tempo

Ex: dados$tempo <- ymd(paste(dados$ano, dados$mes, "01", sep = "-"))

year() <- Extrai o ano de uma data

Ex: year(ymd("2025-11-08")) → 2025

month() <- Extrai o mês

Ex: month(ymd("2025-11-08")) → 11

PACOTES DPLYR E LUBRIDATE

Lubridate + Dplyr: Após converter um valor no formato ano-mês-dia para data com o comando, ymd() do pacote Lubridate, podemos criar junto com o pacote Dplyr duas colunas ano e mês extraindo os valores da coluna data.

dados <- dados %>%

mutate(

ano = year(data),

mes = month(data))

dados <- dados %>% ← Altera o dataframe (manda o resultado da etapa anterior para a próxima).

mutate ← Cria ou modifica colunas dentro do dataframe.

ano = year(data) ← Extrai o ano da coluna data.

mes = month(data) ← Extrai o mês da coluna data.

BASES DE DADOS 2

Bases de dados utilizada como exemplo:
1) Pessoas que não completaram o ensino médio
2) Pessoas que não completaram o ensino fundamental

As bases de dados apresentam as médias de pessoas que **não concluíram o ensino fundamental e o ensino médio**, com informações por **estado, município e para o Brasil** como um todo. Os dados têm como fonte a **Pesquisa Nacional por Amostra de Domicílios (PNAD)** e foram coletados e disponibilizados pelo **Instituto de Pesquisa Econômica Aplicada (IPEA).**

Leitura da base de dados:

```
PESSOASFUNDAMENT <- read.csv('C:/Users/sergi/Desktop/Faculdade/Semana Academica Economia/Curso Sergio/base de dados/PESSOASFUNDAMENT.csv')
PESSOASENSMEDIO <- read.csv('C:/Users/sergi/Desktop/Faculdade/Semana Academica Economia/Curso Sergio/base de dados/PESSOASENSMEDIO.csv')
```

BASES DE DADOS 2

```
head(PESSOASFUNDAMENT)
```

```
##                  code       date value  uname tcode
## 1 PNADCA_PCT25SFUUF 2016-01-01  40.5 Brazil     0
## 2 PNADCA_PCT25SFUUF 2017-01-01  39.6 Brazil     0
## 3 PNADCA_PCT25SFUUF 2018-01-01  38.5 Brazil     0
## 4 PNADCA_PCT25SFUUF 2019-01-01  37.2 Brazil     0
## 5 PNADCA_PCT25SFUUF 2022-01-01  34.0 Brazil     0
## 6 PNADCA_PCT25SFUUF 2023-01-01  33.1 Brazil     0
```

MANIPULAÇÃO BASE DE DADOS 2

```
PESSOASFUNDAMENT2 <- PESSOASFUNDAMENT %>%
  filter(year(date) == 2024)
```

```
head(PESSOASFUNDAMENT2)
```

```
##                 code       date value   uname tcode
## 1 PNADCA_PCT25SFUUF 2024-01-01  31.7  Brazil     0
## 2 PNADCA_PCT25SFUUF 2024-01-01  34.6 Regions     1
## 3 PNADCA_PCT25SFUUF 2024-01-01  41.2 Regions     2
## 4 PNADCA_PCT25SFUUF 2024-01-01  26.5 Regions     3
## 5 PNADCA_PCT25SFUUF 2024-01-01  31.0 Regions     4
## 6 PNADCA_PCT25SFUUF 2024-01-01  27.8 Regions     5
```

```
PESSOASFUNDAMENT3 <- PESSOASFUNDAMENT %>% filter(value > 30.0)

head(PESSOASFUNDAMENT3)
```

```
##                 code       date value  uname tcode
## 1 PNADCA_PCT25SFUUF 2016-01-01  40.5 Brazil     0
## 2 PNADCA_PCT25SFUUF 2017-01-01  39.6 Brazil     0
## 3 PNADCA_PCT25SFUUF 2018-01-01  38.5 Brazil     0
## 4 PNADCA_PCT25SFUUF 2019-01-01  37.2 Brazil     0
## 5 PNADCA_PCT25SFUUF 2022-01-01  34.0 Brazil     0
## 6 PNADCA_PCT25SFUUF 2023-01-01  33.1 Brazil     0
```

MANIPULAÇÃO BASE DE DADOS 2

```r
PESSOASFUNDAMENT4 <- PESSOASFUNDAMENT %>%
  mutate(
    date = ymd(date),       # converte a coluna para data
    ano = year(date),       # extrai o ano
    mes = month(date)       # extrai o mês
  )

head(PESSOASFUNDAMENT4)
```

```
##                  code       date value  uname tcode  ano mes
## 1 PNADCA_PCT25SFUUF 2016-01-01  40.5 Brazil     0 2016   1
## 2 PNADCA_PCT25SFUUF 2017-01-01  39.6 Brazil     0 2017   1
## 3 PNADCA_PCT25SFUUF 2018-01-01  38.5 Brazil     0 2018   1
## 4 PNADCA_PCT25SFUUF 2019-01-01  37.2 Brazil     0 2019   1
## 5 PNADCA_PCT25SFUUF 2022-01-01  34.0 Brazil     0 2022   1
## 6 PNADCA_PCT25SFUUF 2023-01-01  33.1 Brazil     0 2023   1
```

MANIPULAÇÃO BASE DE DADOS 2

```
PESSOASFUNDAMENT5 <- PESSOASFUNDAMENT4 %>%
  group_by(uname, ano, mes) %>%
  summarise(media = mean(value, na.rm = TRUE))
```

```
## `summarise()` has grouped output by 'uname', 'ano'. You can override using the
## `.groups` argument.
```

```
head(PESSOASFUNDAMENT5)
```

```
## # A tibble: 6 x 4
## # Groups:   uname, ano [6]
##   uname    ano   mes media
##   <chr>  <dbl> <dbl> <dbl>
## 1 Brazil  2016     1  40.5
## 2 Brazil  2017     1  39.6
## 3 Brazil  2018     1  38.5
## 4 Brazil  2019     1  37.2
## 5 Brazil  2022     1  34
```

REFERÊNCIAS:

- Wickham H, François R, Henry L, Müller K, Vaughan D (2025). dplyr: A Grammar of Data Manipulation. R package version 1.1.4, https://dplyr.tidyverse.org.

- Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, Grolemund G, Hayes A, Henry L, Hester J, Kuhn M, Pedersen TL, Miller E, Bache SM, Müller K, Ooms J, Robinson D, Seidel DP, Spinu V, Takahashi K, Vaughan D, Wilke C, Woo K, Yutani H (2019). “Welcome to the tidyverse.” Journal of Open Source Software, 4(43), 1686.

- R Core Team (2021). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria.

- Wickham H, Miller E, Smith D (2025). haven: Import and Export 'SPSS', 'Stata' and 'SAS' Files