Modelos Computacionais em Economia

Marcleiton Morais

Universidade Federal do Tocantins (UFT)

24 de junho de 2021

Tipos de dados

Os dados podem ser:

- **Primários**:
  - **Coletados diretamente** por um entrevistador em campo, por um pesquisador em laboratório ou através do disparo de questionário por email ou similar (ex: IBGE, pesquisa experimental etc);
- **Secundários:**
  - Dados coletados por instituições públicas ou privadas os quais são disponibilizados na internet, e que são fruto de coleta direta ou mesmo **gerados automaticamente** a partir do uso de um sistema qualquer (ex: Uber etc);
  - Também podem ser fruto de **alimentação descentralizada** ou de **autodeclaração** (ex: DATASUS, Censo de Portugal etc).

Tipos de dados

Os dados podem ser:

- **Quantitativos**: são variáveis com valores reais ($\Re$) e que, geralmente, formam uma amostra aleatória da população (hipótese 1), uma vez que a distribuição normal tem suporte no $\Re$. O modelo selecionado para sua representação deve ser aproximadamente o populacional (hipótese 2).
  Exemplos de violação:
  - **Variáveis não negativas**: salários de trabalhadores e preços de casas são não-negativos e, portanto, não podem ser normalmente distribuídos de um modo estrito. Dados de duração (desemprego, secunda prisão);
  - **Variáveis não negativas com frequentes zeros**: variável contínua positiva coexiste com uma grupo de observações discretas com valor zero.
    Ex: despesas familiares com um bem de consumo em um dado período de tempo. Comprou ou não? Quanto? Microeconomicamente falando, solução de canto ou interior? Ver [Wooldridge, 2002] ("corner solution models").

Tipos de dados

Os dados podem ser:

Exemplos de violação:

- **Variáveis truncadas**: são variáveis em que todas as observações acima ou abaixo de um valor de corte (threshold) são excluídas. Essa variável tem distribuição truncada na amostra (mesmo sendo aleatória) diferente da população.
  Ex: notas de cortes para admissão de alunos em programas de pós-graduação.
- **Variáveis censuradas**: uma variável é censurada se parte do seu domínio, por exemplo a a reta real, é considerado parcialmente, apenas um intervalo ao invés dos valores atuais são observados nos dados.
  Ex: Contribuição social, Plano de saúde etc. Valores proporcionais ao ganho até um teto, depois permanecem constantes.
  Quanto tempo leva mulheres de 15 anos de uma amostra para ter o primeiro parto? Observamos apenas as que tiveram em dado período, não de todas.
  Salários são observados apenas para quem trabalha. Quanto receberia quem não trabalha?

Tipos de dados

Os dados podem ser:

Exemplos de violação:

- **Variáveis de contagem**: são variáveis fruto da resposta em relação a frequência de um evento. As respostas tem a forma de inteiros não negativos 0,1,2,... ou 0,1,2,...,n.
  Ex: número de patentes registradas por uma empresa, número de vítimas de acidente aéreo em um dado ano, números de transações na bolsa em um dia o número de notas de cortes para admissão de alunos em programas de pós-graduação.

O **modelo de regressão linear** tende a ser inapropriado para esses casos.

Tipos de dados

Os dados podem ser:

- **Qualitativos (Categóricos):** são dados discretos.
  - **Binário**: a variável assume dois resultados possíveis e indica a presença ou ausência de certa propriedade.
    Ex: Sexo: (Feminino,Masculino) (0,1); Trabalho em tempo integral: (Não,Sim) (0,1).
  - **Multinomial**: a variável assume três ou mais resultados possíveis e indica a qualidade de um objeto usando um conjunto mutualmente excludentes, exaustivas e não ordenadas categorias.
    Ex: *Empregado*? (tempo integral, tempo parcial, desempregado, fora da PEA). *Portfólio*: (renda fixa apenas, renda fixa e variável, renda variável epenas, NDO)
  - **Ordenado**: a variável assume três ou mais resultados possíveis e indica a qualidade de um objeto usando um conjunto de características mutuamente excludente, exaustivo e ordenada, mas as diferenças entre as categorias não são definidas.
    Ex: Escala de satisfação: (completamente satisfeito, um pouco satisfeito, neutro, um pouco insatisfeito, completamente insatisfeito).

Tipos de dados

- **Séries temporais:** é um conjunto de observações dos valores que uma variável assume em diferentes momentos do tempo (diariamente, semanalmente, mensalmente, trimestralmente, anualmente etc);
- **Dados em corte transversal:** são dados em que uma ou mais variáveis foram coletadas no mesmo ponto do tempo (censo demográfico feito a cada dez anos);

Unidade	$X(t_1)$	$Y(t_1)$	$X(t_2)$	$Y(t_2)$
1				
2				
$\vdots$				
n				

Tipos de dados

- **Definindo microdados**:
  O aspecto conceitual usado para caracterizar ***microdados*** é, principalmente, a dimensão *cross-sectional*, implicando que o modelo básico de amostragem é caracterizado por independência entre as observações. Esses dados fornecem informações sobre unidades individuais.
  Características:
  - Coss-sectional;
  - Observacional;
  - Escala de medida não contínua.

Unidade	$X(t_1)$	$Y(t_1)$	$X(t_2)$	$Y(t_2)$
1				
2				
$\vdots$				
n				

Tipos de dados

- **Dados combinados:** há elementos tanto de séries temporais quanto de corte transversal;
- **Dados em painel, longitudinais ou de micropainel:** são um tipo especial de dados combinados nos quais a mesma unidade em corte transversal é pesquisada ao longo do tempo. Uma mesma unidade é entrevistada para verificar se houve alguma alteração nas diversas variáveis desde o último levantamento.
  - **Painel balanceado:** o número de observações for o mesmo para todas as unidades;
  - **Painel desbalanceado:** o número de observações não for o mesmo para todas as unidades.

Unidade	$X(t_1)$	$Y(t_1)$	$X(t_2)$	$Y(t_2)$
1				
2				
$\vdots$				
n				

Bases de Dados em Economia

Nacionais

Acesso livre:

- Instituto Brasileiro de Geografia e Estatística: ▸ IBGE ▸ SIDRA;
- Ministério do Trabalho: ▸ RAIS/CAGED;
- Ministério da Saúde: ▸ DATASUS;
- Instituto Nacional de Estudos e Pesquisas Educacionais Anísio Teixeira: ▸ INEP;
- Instituto de Pesquisa Econômica Aplicada: ▸ IPEADATA;
- Secretaria Especial de Comércio Exterior e Assuntos Internacionais. ▸ MDIC;

Bases de Dados em Economia

Internacionais

- Acesso restrito:
  - Business Source Complete: ▸ EBSCO;
  - Academic Search Premier: ▸ EBSCO;
  - Science Direct: ▸ Sci;
  - Web of Science: ▸ CAPES;
  - Journal Storage: ▸ JSTOR;
  - Scopus: ▸ Scopus;
  - Economatica: ▸ Economatica;
  - S&P Global: ▸ SPGlobal.

Bases de Dados em Economia

Internacionais

- Acesso livre:
  - Comdinheiro: ▸ Comdinheiro;
  - The World Bank: ▸ DataBank;
  - The Enterprise Surveys (ES) (World Bank): ▸ Surveys;
  - Organização das Nações Unidas para Alimentação e Agricultura: ▸ FAO;
  - Organização para a Cooperação e Desenvolvimento Econômico: ▸ OCDE;
  - Fundo Monetário Internacional: ▸ FMI.

Análise de dados do IPEADATA

```
library(tidyverse)

install.packages("remotes")

install.packages("ipeadatar")

## ipeadatar <IPEA>
# sidrar <SIDRA, IBGE>
# rbcb <BCB> remotes::install_github('wilsonfreitas/rbc
# wbstats <Banco Mundial>

library(ipeadatar)
# Lista as series disponpiveis
series_ipea <- ipeadatar::available_series()
#Indice de producao fisica de alimentos (media 1989(10
prod_fisica <- ipeadatar::ipeadata("ABIA12_ALIM12")
```

Análise de dados do IPEADATA

```
# Paises disponiveis
paises <- ipeadatar::available_countries()

# Territorios disponiveis
terr <- ipeadatar::available_territories()

# Metadados
metadados <- ipeadatar::metadata("ABIA12_ALIM12")

# PIB Real
pib_real <- ipeadatar::ipeadata("PAN4_PIBPMG4")
# desemprego
desemprego <- ipeadatar::ipeadata("Dese_Pnad")
# taxa de desocupacao
tx_deso <- ipeadatar::ipeadata("PNADC_TXDES_UF")
```

Análise de dados do SIDRA

```
# Sidra
library(sidrar)

library(sidrar)
ipca_15 <- sidrar::get_sidra(3065,
period = "202101-202105")

sidrar::info_sidra(6442)
pms <- sidrar::get_sidra(
6442,
period = "201701-202104",
variable = 8676,
geo = "State",
geo.filter = list("State" = 17),
header = TRUE,
format = 4
)
```

Análise de dados do rbcb (BCB)

```
remotes::install_github("wilsonfreitas/rbcb")
library(rbcb)
rbcb::search_series("pib")
divida_pib <- rbcb::get_series(4536,
start_date = "2020-01-01", end_date = "2020-12-01")
anual_expec <- rbcb::get_annual_market_expectations(
"PIB Total", end_date = "2020-06-01")
mensaL_expec <- rbcb::get_monthly_market_expectations(
"IGP-DI", start_date="2020-01-01",
end_date = "2020-06-01")
```

Análise de dados do rbcb (World Bank)

```
library(wbstats)

wb_paises <- wbstats::wb_countries()
niveis_renda <- wbstats::wb_income_levels()
wb_indic <- wbstats::wb_indicators(lang = "en",
 include_archive = FALSE)
gdp <- wbstats::wb_data(indicator = "NY.GDP.MKTP.CD",
startdate = 2015, enddate = 2016)
br <- wb_data(indicator = "1.0.HCount.1.90usd",
 country = "BR")
```

Salvar em csv

```
readr::write_excel_csv(br, "br_HCount.csv")
```

Salvar em xlsx

```
writexl::write_xlsx(br, "br_xlsx.xlsx")
```

Análise de dados usando o BD+

Definição dos pacotes:

```
library(tidyverse)
library(basedosdados)
```

Criação de um perfil no Google BigQuery:

- Criar e fazer login: Link.
- Passos seguintes:
  - Abrir o console;
  - Criar um novo projeto;
  - Abrir o projeto;
  - Copiar ID do projeto em "informações do projeto".

Defina o seu projeto no Google Cloud a partir do R:

```
set_billing_id("ID do projeto")
```

Análise de dados usando o BD+

Utilizando o Query para selecionar uma base de dados: Após selecionar a base de dados no Base dos Dados (Link), o caminho para definir o query estará em "mais informações".

```
# Para carregar o dado direto no R
querya <- "SELECT * FROM `basedosdados.br_sp_gov_ssp.ocorrencias_registradas`"

queryb <- "SELECT * FROM `...` LIMIT 100"

queryc <- "SELECT * FROM `...` WHERE ano BETWEEN 2012 AND 2019"

df <- read_sql(querya)
```

Leitura dos de segurança pública (SP):

```
data <- read_sql(query)
grande_sp <- data %>%
filter(regiao_ssp == "Grande Sao Paulo (exclui a Capital)") %>%
mutate(id_municipio = as.numeric(id_municipio)) %>%
mutate(furto_de_veiculo = as.numeric(furto_de_veiculo)) %>%
mutate(ano = as.character(ano))
```

Análise de dados usando o BD+

Roubo de veiculos por municipios - Osasco, Barueri, Jandira:

```
grande_sp_furtos <- grande_sp %>%
select(ano, mes, id_municipio, regiao_ssp,
roubo_de_veiculo, furto_de_veiculo) %>%
filter(id_municipio %in% c(3525003, 3505708, 3534401))
```

Gráfico:

```
a <- ggplot(grande_sp_furtos, aes(ano, furto_de_veicul
geom_bar(stat = "identity") +
ggtitle("Furtos de veiculos")
a
```

Análise de dados usando o BD+

Indices de educação:

```
query2 <- "SELECT * FROM 'basedosdados.br_sp_seduc_idesp.escola'
WHERE ano BETWEEN 2010 AND 2019"
data2 <- read_sql(query2)
```

Filtrando para o municipio de Barueri:

```
idesp_barueri <- data2 %>%
filter(id_municipio == 3556453)
```

Query do fluxo escolar:

```
query3 <- "SELECT * FROM 'basedosdados.br_sp_seduc_fluxo_escolar
LIMIT 100"

data3 <- read_sql(query3)
```

Análise de dados usando o BD+

Futebol:

```
query <- "SELECT * FROM
'basedosdados.mundo_transfermarkt_competicoes.brasileirao_serie_
WHERE ano_campeonato = 2020"
```

Leitura:

```
data <- read_sql(query)
soma <- data %>%
select(idade_media_titular_man, time_man, rodada) %>%
mutate(idade_media_titular_man = as.numeric(idade_media_titular_
%>%
filter(time_man == "Santos FC")
writexl::write_xlsx(soma, "soma.xlsx")
df<-read_xlsx("soma.xlsx")
ggplot(df, aes(x= rodada, y= idade_media_titular_man))+
geom_line() +
geom_point() +
labs(x= "Rodada", y= "Idade media como Mandante") +
ggtitle("Santos FC - Idade media como mandante/2020") +
scale_y_continuous(limits = c(23,28))
```

Análise de dados usando o BETS

Séries temporais BACEN e FGV/IBRE:

- Versão CRAN:

```
install.packages("BETS")
library(BETS)
```

- Versão DEV:

```
devtools::install_github("nmecsys/BETS")
```

Exemplo de como buscar banco de dados:

```
BETSsearch(description,
src,
periodicity,
unit,
code, view = TRUE,
lang = "en")
igpm <- BETSsearch(description = "IGP-M", lang = "pt")
dados_igpm <- BETSget(code = 189)
case <- as_tibble(dados_igpm)
```

Análise de dados Nativos - RDS,CSV/XLS/XLSX

Carregando o pacote *readxl*:

```
library(readxl)
```

Lendo um arquivo do diretório - POF:

```
domicilio <- readRDS("Domicilio.rds") %>%
janitor::clean_names()

morador <- readRDS("Morador.rds") %>%
janitor::clean_names()
```

Unindo dois data frames

```
result <- full_join(domicilio, morador)
```

Lendo arquivos XLSX/XLS

```
ex <- read_xlsx("mymsa.xlsx") %>%
janitor::clean_names() %>%
remove_empty()
```

Análise de dados Nativos - RDS,CSV/XLS/XLSX

Limpando os nomes da coluna com a função (tabyl):

```
tabyl(ex, meat_colour) %>%
adorn_pct_formatting(digits = 0, affix_sign = TRUE)

query <- "SELECT * FROM
'basedosdados.br_ibge_pam.municipio_lavouras_permanentes'
WHERE ano BETWEEN 2010 AND 2019"
dados <- read_sql(query)

dados_to <- dados %>%
filter(sigla_uf == "TO")
saveRDS(dados_to, "pam-2010-19")
```

Análise de dados do Yahoo Finance

Biblioteca *BatchGetSymbols*:

```
install.packages("BatchGetSymbols")
library(BatchGetSymbols)

acao <- c("COGN3.SA")
bg <- "2020-08-08"
lst <- Sys.Date()
bench <- "^BVSP"
freq

dados <- BatchGetSymbols(tickers = acao,bench.ticker = bench,
first.date = bg,last.date = lst)

ibov <- dados$df.tickers
ibov_ajust <- ibov %>%

janitor::clean_names()

ggplot(data = ibov_ajust, mapping = aes(ref_date, price_close)) +
geom_line(color = "#006600") + labs(y = "Preco", title = , subtitle, caption) +
theme_fivethirtyeight()
```

Análise de microdados da PNAD Contínua

Tipos de microdados:

- Trimestral, que contém a parte básica investigada pela pesquisa, contendo variáveis conjunturais de mercado de trabalho referentes a um trimestre civil;
- Anual, que contém temas estruturais específicos investigados na pesquisa para um ano civil.

Periodicidade:

- Mensal - Conjunto restrito de indicadores relacionados à força de trabalho e somente para o nível geográfico de Brasil;
- Trimestral - Conjunto de indicadores relacionados à força de trabalho para todos os níveis de divulgação da pesquisa;
- Anual - Demais temas permanentes da pesquisa e indicadores complementares à força de trabalho; e
- Variável - Outros temas ou tópicos dos temas permanentes a serem pesquisados com maior periodicidade ou ocasionalmente.

Análise de microdados da PNAD Contínua

Temas e tópicos suplementares (trimestral/visitas):

- Educação (2o trimestre); e
- Acesso à televisão e à Internet e posse de telefone móvel celular para uso pessoal (4o trimestre)
- Habitação (1a visita);
- Características gerais dos moradores (1a visita);
- Informações adicionais da força de trabalho (1a visita);
- Outras formas de trabalho (afazeres domésticos, cuidados de pessoas, produção para o próprio consumo e trabalho voluntário) (5a visita);
- Trabalho de crianças e adolescentes (5a visita); e Rendimentos de outras fontes (1a e 5a visitas).

Análise de microdados da PNAD Contínua

Pacotes PNADcIBGE e survey

- **PNADcIBGE**: foi desenvolvido para facilitar o download, importação e análise dos dados amostrais da Pesquisa Nacional por Amostra de Domicílios Contínua realizada pelo Instituto Brasileiro de Geografia e Estatística - IBGE.
- **Survey**: através do objeto criado com este pacote *PNADcIBGE*, é possível utilizar o pacote *survey* para realizar análises considerando o efeito do esquema de seleção utilizado no plano amostral complexo da pesquisa e calcular corretamente as medidas de erro das estimativas, considerando o estimador de pós-estratificação utilizado na pesquisa.

Análise de microdados da PNAD Contínua

A instalação do pacote

```
install.packages("PNADcIBGE")
library(PNADcIBGE)
```

Importação *online*

```
help("get_pnadc")
```

- **Trimestrais**:

```
#Condicao em relacao a forca de trabalho (VD4001)
#e Condicao de ocupacao (VD4002) do 3o trimestre/2017.
dadosPNADc <- get_pnadc(year = 2017, quarter = 3,
vars=c("VD4001","VD4002"))
```

Análise de microdados da PNAD Contínua

Clase do objeto (default)

```
dadosPNADc
class(dadosPNADc)
```

Clase do objeto (Data-frame)

```
dadosPNADc_brutos <- get_pnadc(year = 2017, quarter = 3,
vars = c("VD4001", "VD4002"), design = FALSE)

dadosPNADc_brutos
class("dadosPNADc_brutos")
```

Excluindo labels

```
dadosPNADc_brutos_sem <- get_pnadc(year = 2017, quarter = 3,
vars = c("VD4001", "VD4002"), design = FALSE, labels = FALSE)
```

Análise de microdados da PNAD Contínua

Carregar microdados no diretório temporário:

```
pnadc_dat<-
get_pnadc(2017, quarter = 3, interview = NULL,
vars = c("VD4001", "VD4002"),labels = F, #Quais e como ver as va
design = F, #informa o objeto para analise
(base de microdados ou objeto com desenho amostral)
savedir = tempdir() #informa diretorio para salvar arquivo
        )
```

Análise de microdados da PNAD Contínua

- **Microdados anuais**:

1a Visita 2006

- Temas:

  -Características Adicionais do Mercado de Trabalho;

  -Características Gerais dos Moradores;

  -Características Gerais dos Domicílios.

  -Rendimentos de Outras Fontes

```
dadosPNADc_anual <- get_pnadc(year = 2016, interview = 1)
dadosPNADc_anual
```

Análise de microdados da PNAD Contínua

Importação *offline*: três funções

- read_pnadc: Para a leitura do arquivo .txt dos microdados;
- pnadc_labeller: Opcional. Coloca os rótulos dos níveis nas variáveis categóricas;
- pnadc_design: Cria o objeto do plano amostral para a análise com o pacote *survey*.

Exemplo de leitura para o 3º trimestre de 2017:

```
#Leitura
dados_pnadc <- read_pnadc("PNADC_032017.txt",
"Input_PNADC_trimestral.txt")
#Labels
\dados_pnadc <- pnadc_labeller(dados_pnadc,
"dicionario_das_variaveis_PNAD_Continua_microdados.xls")
#Objeto do plano amostral
dados_pnadc <- pnadc_design(dados_pnadc)
```

A importação offline é feita da mesma forma tanto para microdados mensais quanto anuais.

Análise de microdados da PNAD Contínua

Definindo o design manualmente:

```
#Baixando os dados
dadosPNADC <- get_pnadc(year = 2017, quarter = 3,
vars = c("VD4001", "VD4002", "V2009", "VD4015",
"VD4016"), design = F)

# Adiciona coluna de 1's ao arquivo de microdados
dadosPNADC$one <- 1

# Conta o numero de pessoas da amostra
sum(dadosPNADC$one)
```

Análise de microdados da PNAD Contínua

Carregando os pacotes:

```
# Opcao para permitir estimar variancia quando houver
apenas 1 observacao na UPA
options( survey.lonely.psu = "adjust" )
# Troca . por , nos outputs
options(OutDec=",")
# Instala e Carrega pacotes necessarios
# Pacote para facilitar manuseio dos dados
#install.packages("tidyverse")
library(tidyverse)
library(survey)
# Pacote para analisar dados amostrais
#install.packages("srvyr")
library(srvyr)
```

Análise de microdados da PNAD Contínua

Cálculo de estimativas considerando a estrutura do plano amostral complexo:

```
#############################################################
# Declara Plano Amostral - Usa pacote survey
#Declara estrutura do plano amostral complexo
pnadc_plano <- svydesign(
ids = ~ UPA ,              # Declara a unidade amostral mais granular
#UPA - Unidade Primaria de Amostragem
strata = ~ Estrato , # Declara a variavel que contem os estratos
weights = ~ V1027 ,  # Declara variavel com pesos
data = dadosPNADC ,  # Declara base de microdados
nest = TRUE              # Declara que os estratos podem conter
identificacoes identicas para UPA's distintas
)
# Sumariza o objeto sobre a estrutura do plano amostral
summary(pnadc_plano)
```

Análise de microdados da PNAD Contínua

Calibração dos pesos com base nas estimativas de população produzidas pelo IBGE:

```
#############################################################
# Tabela com frequencias populacionais (estimativas IBGE para
calibracao)
df_pos <- data.frame( posest = unique( pnadc$posest ) ,
Freq = unique( pnadc$V1029 ) )
df_pos

#############################################################
# Calibra pesos
pnadc_calib <- postStratify( pnadc_plano , ~ posest , df_pos )
# Outras opcoes: funcoes rake() e calibrate()

# Obtem fatores de calibracao dos pesos da amostra
pnadc_fatores <- weights(pnadc_calib) / weights(pnadc_plano)
boxplot(pnadc_fatores, horizontal = TRUE,
xlab="Fatores de calibracao")
```

Análise de microdados da PNAD Contínua

Estimação de totais populacionais

A função *svytotal*:

- O nome da variável que se deseja calcular o total, precedido por um ;
- O nome do objeto do plano amostral;
- A opção na.rm = T, que remove as observações onde a variável é não-aplicável.

Análise de microdados da PNAD Contínua

Estimativas da população

```
#############################################################
# Validar estimativas populacionais
svytotal( ~ one , pnadc_calib )        #estimativa populacional
```

Estimativa divulgada pelo IBGE: ▸ Link.
Efeito do Plano Amostral (EPA) considerando o Rendimento Habitual do Trabalho Principal

```
#############################################################
# Avalia Efeito do Plano Amostral
# Compara estimativas do RHTP com e sem plano
(+ calculo do EPA de Kish)
round(svymean( ~ VD4016 , subset(pnadc_plano, V2009 >= 14)
  , na.rm = TRUE ),2)

round(svymean( ~ VD4016 , subset(pnadc_calib, V2009 >= 14)
  , na.rm = TRUE, deff="replace"),2)
```

Análise de microdados da PNAD Contínua

Salva objeto final

```
# Salva objeto final
saveRDS(pnadc_calib,"pnadc_calib_20173")

# Limpa objetos da memoria
rm(pnadc_plano,df_pos,dadosPNADC)

##############################################################
# Carrega dados da PNADC
pnadc_calib <- readRDS(file="pnadc_calib_20173")

# Modifica objeto para permitir sintaxe tipo tidyverse
pnadc_calib <- as_survey_design(pnadc_calib)
```

Análise com pacote *survey*

Pacote survey

O pacote survey (▸ Link) é um pacote criado especificamente para análise e modelagem de dados amostrais complexos.

PNAD Contínua do 3º trimestre de 2017

UF	Unidade da Federação
UPA	Unidade Primária de Amostragem
Estrato	Variável que contém os estratos
V1027	Peso do domicílio e das pessoas
V1029	Projeção da população
posest	Domínios de projeção
V2007	Sexo
V2009	Idade do morador na data de referência
V2010	Cor ou raça
V3007	Já concluiu algum outro curso de graduação?

Análise com pacote *survey*

Pacote survey

O pacote survey (▸ Link) é um pacote criado especificamente para análise e modelagem de dados amostrais complexos.

PNAD Contínua do 3º trimestre de 2017	
VD3004	Nível de instrução mais elevado alcançado (pessoas de 5 anos ou mais de idade)
VD4001	Condição em relação à força de trabalho na semana de referência para pessoas de 14 anos ou mais de idade
VD4002	Condição de ocupação na semana de referência para pessoas de 14 anos ou mais de idade
VD4015	Tipo de remuneração
VD4016	Rendimento mensal habitual do trabalho principal (>= 14 anos)
VD4020	Rendimento mensal efetivo de todos os trabalhos (>= 14 anos)
VD4035	Horas efetivamente trabalhadas na semana de referência em todos os trabalhos (>= 14 anos)

Análise de microdados da PNAD Contínua

Importando os dados

```
variaveis_selecionadas <- c("UF", "V1029", "V1027", "Estrato",
"UPA", "posest", "V2007", "V2009", "V2010", "V3007", "VD3004",
"VD4001", "VD4002", "VD4015", "VD4016", "VD4020", "VD4035")

pnadc <- get_pnadc(year = 2017, quarter = 3,
vars = variaveis_selecionadas, design = T)
```

Análise de microdados da PNAD Contínua

Estimando estatísticas de interesse

```
# Prepara variaveis para calculo de estimativas
pnadc <-  update(pnadc,
idade5 = factor( 1 + findInterval( V2009 , seq( 5 , 60 ,5))),
nivel_renda = factor( 1 + findInterval( VD4020 , seq( 5 , 60 , 5
sexo = as.numeric( V2007 == 1 ) ,
pia = as.numeric( V2009 >= 14 ) ,
analfabeto = 1*(V3001==2),
ocupado = ifelse( pia == 1 , as.numeric( VD4002 %in% 1 ) ,NA),
desocup30 = ifelse( pia == 1 , as.numeric( VD4002 %in% 2 ) ,NA),
pea_c = as.numeric( ocupado == 1 | desocup30 == 1 ),
# (rendimento habitual do trabalho principal)
VD4016n = ifelse( pia %in% 1 & VD4015 %in% 1 , VD4016 , NA ) ,
# (rendimento efetivo do todos os trabalhos)
VD4020n = ifelse( pia %in% 1 & VD4015 %in% 1 , VD4020 , NA ) ,
#indicador de nivel superior
VD3001n = 1*( VD3004 == 7)
)
```

Análise de microdados da PNAD Contínua

Estimando Totais

- **Variáveis Numéricas**:

```
totalrenda <- svytotal(~VD4020, pnadc, na.rm = T)
totalrenda
```

Coeficientes de variação:

```
cv(totalrenda)
```

Intervalos de confiança:

```
confint(totalrenda) #intervalo de confianca de 95% (padrao)
confint(totalrenda, level= .99) #intervalo de confianca de 99%
```

Análise de microdados da PNAD Contínua

Estimando Totais

- **Variáveis Categóricas**:

```
totalsexo <- svytotal(~V2007, pnadc, na.rm = T)
totalsexo
totalsexoraca <- svytotal(~V2007 + V2010, pnadc, na.rm
totalsexoraca
```

- **Cruzamento de duas ou mais variáveis**:

```
totalsexoEraca <- svytotal(~ interaction(V2007, V2010)
pnadc, na.rm = T)
ftable(totalsexoEraca)
```

Análise de microdados da PNAD Contínua

Estimando Totais

- **Estimando Médias**:

```
mediarenda <- svymean(~VD4020, pnadc, na.rm = T)
mediarenda
```

Coeficiente de variação:

```
cv(mediarenda)
```

Intervalo de confiança:

```
confint(mediarenda)
```

Análise de microdados da PNAD Contínua

Estimando Totais

- **Estimando Proporções**:

```
propsexo <- svymean(~V2007, pnadc, na.rm = T)
propsexo
#mais de uma variavel
propsexoraca <- svymean(~V2007 + V2010, pnadc,
 na.rm = T)
propsexoraca
#Cruzamento de variaveis
propsexoEraca <- svymean(~ interaction(V2007, V2010),
pnadc, na.rm = T)
ftable(propsexoEraca)
```

- **Estimando a proporção de um cruzamento**:

```
propsexoEraca <- svymean(~ interaction(V2007, V2010),
pnadc, na.rm = T)
ftable(propsexoEraca)
```

Análise de microdados da PNAD Contínua

Estimando Razões

A taxa de desocupação é a razão entre o total de pessoas desocupadas pelo total de pessoas na força de trabalho.

```
txdesocup <- svyratio(~VD4002 == "Pessoas desocupadas",~VD4001
 == "Pessoas na forca de trabalho", pnadc, na.rm = T)
txdesocup
```

Cálculos de coeficiente de variação e intervalos de confiança:

```
cv(txdesocup)

confint(txdesocup)
```

Análise de microdados da PNAD Contínua

Estimando Medianas e Quantis

Usando a função *svyquantile*:

```
medianarenda <- svyquantile(~VD4020, pnadc,
quantiles = .5, na.rm = T)
medianarenda
```

O erro padrão (*ci = TRUE*)

```
medianarenda <- svyquantile(~VD4020, pnadc,
quantiles = .5, na.rm = T, ci = TRUE)
medianarenda
```

Então o erro padrão e coeficiente de variação são:

```
SE(medianarenda)
cv(medianarenda)
```

Vários quantis simultaneamente:

```
quantisrenda <- svyquantile(~VD4020, pnadc,
quantiles = c(.1,.25,.5,.75,.9), na.rm = T)
quantisrenda
```

Análise de microdados da PNAD Contínua

Estimação para um Domínio

Condicionais com igualdade e desigualdade:

```
mediarendaM <- svymean(~VD4020, subset(pnadc,
V2007 == "Mulher")   , na.rm = T)
mediarendaM

txdesocup25 <- svyratio(~VD4002 =="Pessoas desocupadas
~VD4001 == "Pessoas na forca de trabalho",
subset(pnadc, V2009>=25) , na.rm = T)
txdesocup25
```

Análise de microdados da PNAD Contínua

Estimação para um Domínio

Múltiplas condições:

```
nivelinstrHP30 <- svymean(~VD3001, subset(pnadc,
V2007 == "Homem" & V2010 == "Parda" & V2009 > 30),
na.rm = T)
nivelinstrHP30
```

Diversas análises para um mesmo domínio:

```
dadosPNADc_mulheres <- subset(pnadc,
V2007 == "Mulher")
dadosPNADc_mulheres
```

Análise de microdados da PNAD Contínua

Estimação para Vários Domínios (*svyby*)

Argumentos:

- A variável da qual se deseja calcular a quantidade;
- A variável que define os domínios;
- O objeto do plano amostral;
- A função utilizada para calcular a quantidade de interesse (*svytotal,svymean,svyratio,svyquantile,...*)

Frequência relativa de homens e mulheres em cada nível de instrução:

```
freqSexoInstr <- svyby(~V2007, ~VD3001, pnadc,
svymean, na.rm = T)
freqSexoInstr
```

O inverso:

```
freqInstrSexo <- svyby(~VD3001, ~V2007, pnadc,
svymean, na.rm = T)
freqInstrSexo
```

Análise de microdados da PNAD Contínua

Estimação para Vários Domínios (*svyby*)

Renda média efetiva por unidade da federação:

```
mediaRendaUF <- svyby(~VD4020, ~UF, pnadc,
svymean, na.rm = T)
mediaRendaUF
```

Intervalo de confiança:

```
confint(mediaRendaUF)
```

Cruzamentos de variáveis categóricas (*interaction*):

```
txdesocupSexoRaca <- svyby(~VD4002 == "Pessoas desocupadas",
~interaction(V2007,V2010), pnadc, svyratio,
denominator = ~VD4001 == "Pessoas na forca de trabalho",
na.rm = T, vartype = "cv")socupadas",
txdesocupSexoRaca
```

Análise de microdados da PNAD Contínua

Gráficos para Dados Amostrais

Histograma:

```
svyhist(~ as.numeric(VD4035), pnadc, main = "Histograma",
 xlab = "Numero de Horas Trabalhadas")

svyhist(~ as.numeric(VD4035), pnadc, freq = TRUE,
main = "Histograma", xlab = "Numero de Horas Trabalhadas")
```

Boxplot:

```
#Sem grupo
svyboxplot(VD4035 ~ 1, pnadc, main = "Boxplot do
 Numero de Horas Trabalhadas")
#Com grupo
svyboxplot(VD4035 ~ V2007, pnadc, main = "Boxplot
do Numero de Horas Trabalhadas por Sexo")
#Outilers
svyboxplot(VD4035 ~ V2007, pnadc, main = "Boxplot
do Numero de Horas Trabalhadas por Sexo", all.outliers = TRUE)
```

Gráficos de Dispersão

Representação do pesos através do argumento *style*:

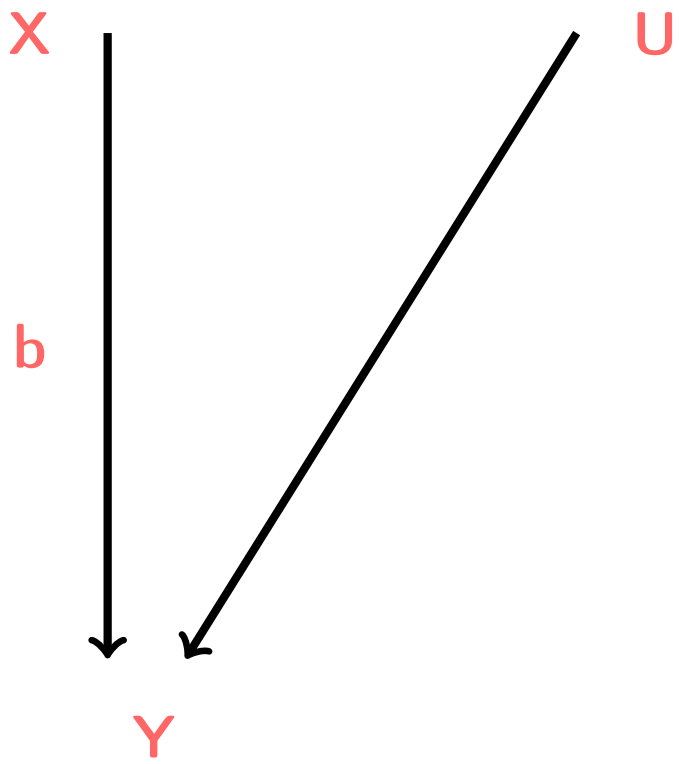
```
svyplot(VD4020 ~ VD4035, pnadc, style = "bubble",
xlab = "Horas efetivamente trabalhadas",
ylab = "Rendimento efetivo")

svyplot(VD4020 ~ VD4035, pnadc, style = "transparent",
xlab = "Horas efetivamente trabalhadas",
 ylab = "Rendimento efetivo")
```

Ordinary least squares (OLS)

OLS is the work-horse model of microeconometrics. It is quite simple to estimate. It is straightforward to understand. It presents reasonable results in a wide variety of circumstances.

Estimating the Causal Effect (year of schooling (X) and person's income (Y))



Estimating the Causal Effect

A Linear Causal Model:

Individual i earns income y_i determined by their education level x_i and unobserved characteristics v_i.

$$y_i = a + bx_i + v_i$$

where a and b are the parameters that determine how much icome individual i earns and how much of that is determined by their leval of education.

Our goal is to estimate these parameters from the data we have.

Estimating the Causal Effect

Simulation of the Causal Effect

Sumulated data

- Linear relationship between x and y with an intercept of 2 and a slope of 3.
- Unobserved characteristics is ditributed **standard normal** **(**$v_i \sim \mathcal{N}(0,1)$ **)**.

We want to estimate the value of b, which has a true value of 3.

```
# Create a simulated data set
set.seed(123456789)
# use to get the exact same answer each time the code is run.
# you need to set the seed each time you want to get the
# same answer.
N <- 100
# Set N to 100, to represent the number of observations.
a <- 2
b <- 3 # model parameters of interest
# Note the use of <- to mean "assign".
x <- runif(N)
```

Estimating the Causal Effect

Simulation of the Causal Effect

Sumulated data

- Linear relationship between x and y with an intercept of 2 and a slope of 3.
- Unobserved characteristics is ditributed **standard normal** ($v_i \sim \mathcal{N}(0,1)$).

We want to estimate the value of b, which has a true value of 3.

```
# create a vector where the observed characteristic, x,
# is drawn from a uniform distribution.
u <- rnorm(N)
# create a vector where the unobserved characteristic,
# v is drawn from a standard normal distribution.
y <- a + b*x + v # create a vector y
# * allows a single number to be multiplied through
# the whole vector
# + allows a single number to be added to the whole vector
# or for two vectors of the same length to be added together.
```

Estimating the Causal Effect

Averaging to Estimate the Causal Effect

Plot of x and y with the true relationship represented by the line.

```
mean(y[x > 0.95]) - mean(y[x < 0.05])
plot(x, y) # creates a simple plot
abline(a = 2, b = 3) # adds a linear function to the plot.
# a - intercept, b - slope.

#mean takes an average
#the logical expression inside the square brackets
#creates an index for the elements of y where the logical
#expression in x holds.
```

By taking the difference in the average of Y calculated at two different values of X, we can determine how X affects the average value of Y . In essence, this is what OLS does.

Estimating the Causal Effect

Assumptions of the OLS Model

Unobserved characteristics enter independently and additively:

- **Independence:** states that conditional on observed characteristics (the $X's$), the unobserved characteristic (the U) has independent effects on the outcome of interest (Y).
  Our estimated model does not allow students from wealthy families to be more likely to go to college and get a good job due to their family background.
- **Additive:** states that unobserved characteristics enter the model additively.
  Attending college increases everyone's income by the same amount

Matrix Algebra of the OLS Model

Standard Algebra of the OLS Model

Consider

$$y_i = a + bx_i + v_i$$

and let $a = 2$. So

$$b = \frac{y_i - 2 - v_i}{x_i} \tag{1}$$

This highlights two problems:

- **First:** the observed terms ($\{y_i, x_i\}$) are different for each person i, but Equation 1 states that b is exactly the same for each person.
- **Second:** second problem is that the unobserved term (v_i) is unobserved.

"kill two birds with one stone"

We can determine b by averaging:

$$\frac{1}{N}\sum^{N} y_i = \frac{1}{N}\sum^{N}(a + bx_i + v_i)$$

References

Wooldridge, J. M. (2002).
Econometric analysis of cross section and panel data mit press.
Cambridge, MA, 108.

Obrigado!