Enap

Aulas 9 - PNAD

Marcelo Prudente e Rafael Giacomin

03 de abril de 2018

Enap

Baixando pacotes necessários

Porque PNADs com R?

Enap

- A forma mais fácil de analisar os dados da PNAD e PNADC é por meio do pacote *lowdown*.
- É a forma atualmente indicada pelo **IBGE** para a análise desses dados
- Acesse o site **Analyze Survey Data for Free**

Instalando o pacote lodown

Enap

- Primeiro passo, instalar lodown.

```
# instalar e utilizar pacote httr
install.packages(httr);  library(httr)

# reorganizar configuração para baixar pacote
set_config(config(ssl_verifypeer = 0L))

# instalar e utilizar rcpp
install.packages("Rcpp"); library(Rcpp)

# devtools - para baixar do github
library(devtools)

# instalar lodown
install_github( "ajdamico/lodown" , dependencies = TRUE )
```

Instalando os pacotes survey e srvyr

Enap

- Vamos utilizar dois pacotes: *survey* e *srvyr*

```
# instalar pacote survey
install.packages("survey")
library(survey)

# instalar e utilizar pacote srvyr
install.packages("srvyr")
library("srvyr")
```

BAIXAR DADOS DA PNAD

Criando catálogo dos arguivos a serem baixados

Enap

- O primeiro passo para acessar as PNADs é criar um catálogo com as observações. Essa aborgagem permite escolher o período a ser baixado das PNADS.

```
# utilizar pacote lodown
library(lodown)

# Pesquisa Nacional de Amostra de Domicílios
pnad_cat <-
    get_catalog( "pnad" ,
        output_dir = file.path( path.expand( "~" ) , "PNAD" )
```

Criando catálogo dos arguivos a serem baixados

Enap

- Ainda, é possível fazer o mesmo especificando a pasta em que se deseja armazenar os arquivos.

```
# Cria um catálogo com todas os dados disponíveis
meu.path <- "C:/PNADC"
pnad_cat <-
    get_catalog( "pnad" ,
        output_dir = meu.path )
```

Baixando os dados

Enap

- Depois de criado o catálogo, vizualize.

```
View(pnad_cat)
```

- Por ser um data.frame você pode extrair um subconjunto do período que deseja analisar:

```
# Tirar um subconjunto dos dados
pnadc_cat <- pnad_cat%>%
  filter(year == 2015)
```

- Depois, baixar:

```
lodown( "pnad" , pnad_cat )
```

Ler o arquivo da PNAD

Enap

```
# Diretório dos arquivos
setwd(meu.path)
# Listar arquivos
list.files()
# readr
library(readr)
# baixar pnad
pnadc_df <- read_rds("2015 main")
```

PNAD - pontos importantes

Preliminares (1)

Enap

- O banco de dados das PNADs exigem a criação de variáveis derivadas para a análise.
- O ideal é conhecer bem o dicionário de variáveis (ver arquivo na pasta dicionarios_pnad)
- Nesse processo será possível revisar alguns comandos já apresentados.

Preliminares (2)

Enap

- No site *asdfree* você notará que os autores utilizam diversas formas para manipular os dados.
  - Em geral, comandos de base do sistema.
- No entanto, a manipulação pode ser executada sem problemas com os comandos do *dplyr*.

PNAD: lendo os arquivos

Enap

```
# Depois de baixados
pnad_cat <- pnad_cat %>% filter(year == 2015)
# Ler variáveis
pnad <- readRDS(pnad_cat$output_filename)

# Estrutura do banco
str(pnad)
```

PNAD - manipulação dos dados

Para que manipular dados da PNAD?

Enap

- Como em qualquer base, os dados podem ser manipulados para fornecerem mais informações.
- Para a PNAD, faremos o mesmo.
- Aqui, vamos utilizar os nossos conhecimentos de *dplyr*.

Transformar em numéricas

Enap

- É provável que nem todas variáveis da PNAD baixada tenham a classe que devem originalmente estar: numéricas.
- Peça a estrutura dos dados:

```
str(pnad)
```

- Execute a transformação

```
# transformar as variáveis em numericas
pnad <- pnad %>%
  mutate_if(is.character, as.numeric)
```

Recodificar algumas variáveis PNAD

Enap

- Vamos recodificar algumas variáveis da PNAD

```
source("C:/curso_r_enap/funcoes/age_cat.R")
pnad <- pnad %>%
  mutate(
         # cria faixa de idades
         fx_idade = age.cat(v8005, upper = 80, by = 5) ,
         # criar uma variável para determinar quem são os adol
         adolescentes = as.numeric( v8005 > 12 & v8005 < 20 )
         # se o indivíduo trabalha antes dos 13 anos
         trab_antes_treze = as.numeric( v9892 < 13 ))
```

Exemplo de manipulação

Exemplo: criar variável situação da família

Enap

- Outra questão que a PNAD pode responder é o tipo de família predominante no Brasil.
- Porém, será necessário criar novas variávies para encontrar os objetos desejados,
- De acordo com as variáveis do banco, é possível encontrar seis tipos de família.

	Cônjuge	Filhos	Sexo	Soma
Casal sem filhos	1	0	6	7
Casal com filhos	1	7	6	14
Mulher sozinha	0	0	4	4
Mãe com filhos	0	7	4	11
Homem sozinho	0	0	2	2
Pai com filhos	0	7	2	9

Exemplo: situação da família: passo 1

Enap

- É necessário saber como é a família em cada domicílio. A variável central é a **v0401 - condição na unidade domiciliar**.

```
# Recodificar variáveis
pnad <- pnad %>%
  # identifica o domicilio
  mutate(domicilio = factor(paste0(v0101, v0102, v0103))) %>%
  group_by(domicilio) %>%
  # identifica quem tem conjuge no domicilio
  mutate(tem_conjuge =as.numeric(any(v0401 == 2)))%>%
  group_by(domicilio, tem_conjuge)%>%
  # identifica quem tem filhos
  mutate(tem_filhos = ifelse(any(v0401 == 3), 7, 0))
```

Exemplo: situação da família: passo 2

Enap

- Para identificar os 6 tipos de família, é necessário recodificar um pouco mais.
- Agora, atribuimos que o sexo de todos os conjuges como **6**, pois para esse grupo a variável sexo não importa.

```
pnad <- pnad %>% group_by(domicilio)%>%
  mutate(sexo_conjuge= ifelse(tem_conjuge ==1, 6,
                              v0302),
         sit_fam = (sexo_conjuge + tem_filhos + tem_conjuge))
```

Exemplo: situação da família: passo 3

Enap

- Por fim, vamos renomar as variáveis

```
pnad$sit_fam <- recode(pnad$sit_fam,
                       "2" = "Homem Sozinho",
                       "4" = "Mulher Sozinha",
                       "9" = "Pai com filhos",
                       "11" = "Mãe com filhos",
                       "14" = "Casal com filhos",
                       "7" =  "Casal sem filhos")
```

Outras manipulações

Enap

- Ao longo da análise da PNAD promoveremos outras manipulações.
- Atente para a lógica subjacente à criação de novas variáveis. Ela irá se repetir em todo o processo.

Preparando o survey para análise

Survey? (1)

Enap

- Os grandes *surveys*, a exemplo das PNADs, se diferenciam filosoficamente e substantivamente da aborgagem tradicional das amostras aleatórias.
- Em poucas palavras, a amostragem aleatória não produz estimadores corretos para *surveys* grandes e complexos.
- Em pesquisas domiciliares, se a estratificação envolvesse aleatoriamente indivíduos, seria necessário visitar milhões de lugares para uma pesquisa.

Survey? (2)

Enap

- Por isso, as questões logísticas (custos!) levam a conglomerar as amostras - concentrar geograficamente as entrevistas é financeiramente mais efetivo.
- Nas PNADs há uma amostragem por conclomerados:
  - A unidade de seleção pode ser o município, que congrega setores censitários, que contém domicílios.
  - Seleciona-se uma amostra dos municípios, depois uma amostra dos setores censitários, em seguida dos domicílios.

Analisando surveys

Enap

- Por conta do seu desenho, a análise de *surveys* não pode ser feita diretamente.
- Deve ser feita a "correção"" das unidades primárias amostrais e dos estratos de amostragem.

Analisando surveys: pacotes

Enap

- A análise de dados estruturados em *surveys* está implementada pelo pacote *survey*.

```
library(survey)
```

- Para a PNAD, precisamos adequar a amostra **estratificando** e **pós-estratificando**.

Pré-estratificação da PNAD

Enap

- Nos grandes *surveys*, a amostra aleatória não é tão utilizada, pois outros desenhos dão maior precisão a menor custo.
- Na PNAD, adota-se um plano amostral estratificado, conglomerado com dois ou três estágios de seleção dependendo do estrato.
- Nos conglomerados, o primeiro estágio é chamado Unidade Primária de Amostragem (UPA).
  - Nos municípios representativos há dois conglomerados. O primeiro estágio é o setor censitário. O segundo, os domicílios.
  - Nos municípios **não representativos**, há um conglomerado (município), subdividido em um segundo conglomerado (setor censitários), subdivido em um terceiro conglomerado (domicílios.)

Pré-estratificação

Enap

- Para as PNADS até 2015, a pré-estratificação deve assumir a seguinte forma:

```
prestratified_design <-
    svydesign(
        # upa
        id = ~ v4618 ,
        # estrato
        strata = ~ v4617 ,
        data = pnad,
        # v4610 - inv fracao amostral
        weights = ~ pre_wgt ,
        nest = TRUE
    )

rm( pnad ) ; gc()
```

Pós-estratificação: ideia geral (1)

Enap

- Algumas vezes, um *survey* sobreamostra grupos de população - idade, gênero, outros.
- Em outros termos, a *pós-estratificação* é uma forma de calibrar os dados.
- As técnicas de pós-estratificação são utilizadas para ajustar os pesos amostrais e melhorar a eficiência dos estimadores.
- Assim, ajustes são feitos nos pesos amostrais de modo que a população estimada total condiza com a população total conhecida.

Pós-estratificação: ideia geral (2)

Enap

- Para a PNAD, especificamente, Ruiz e Silva (2014) apontam que:

  O método adotado para a calibração dos pesos utiliza informações auxiliares provenientes das projeções da população para cada Unidade de Federação (UF) segundo o tipo de área (região metropolitana - RM - e não metropolitana).

- Com isso, para a PNAD 2015 são analisados as 27 UFs e nove Regiões Metropolitanas (Belém, Fortaleza, Recife, Salvador, Belo Horizonte, Rio de Janeiro, São Paulo, Curitiba e Porto Alegre).

Pós-estratificação: computar

Enap

- Pós-estratificar:

```
pop_types <-
    data.frame(
        v4609 = unique( pnad$v4609 ) ,
        Freq = unique( pnad$v4609 )
    )
pnad_design <-
    postStratify(
        design = prestratified_design ,
        strata = ~ v4609 ,
        population = pop_types
    )

rm( prestratified_design ) ; gc()
```

PNAD_DESIGN

Enap

- Finalmente, temos o arquivo pronto para a análise, que chamamos de *pnad_design*.
- Qual a classe do *pnad_design*??

```
class(pnad_design)
# transformar em tbl_svy
pnad_design <- as_survey(pnad_design)
# classe
class(pnad_design)
```

- A análise desse objeto será feita de forma distinta, conforme explicitado a seguir.

Análise da PNAD

pacote svryr

Enap

- Para a análise dos surveys, utilizaremos o pacote **srvyr** que utiliza a sintaxe do *dplyr* para executar as análises.
- Há cinco comandos principais para a análise:
  - **survey_mean()**
  - **survey_ratio()**
  - **survey_total()**
  - **survey_quantile()**
  - **survey_median()**

Análises: totais survey_total()

Enap

- Ajuste para o cálulo da variância

```
# opção para ajustar o cálculo da variância
options( survey.lonely.psu = "adjust" )
```

- **Algumas estimativas**:

```
# população estimada do Brasil em 2015
pnad_design %>%
  summarize(pop_brasil = survey_total(one))
# população estimadad do Brasil por regiao em 2015
pnad_design %>%
  group_by(region)%>%
  summarize(pop_brasil = survey_total(one))
```

- Tente extrair a população por Estado

Análises: média survey_mean()

Enap

```
# rendimento por sexo
pnad_design %>%
  group_by(v0302)%>%
  summarize(rendimento = survey_mean(v4718, na.rm = TRUE))

# rendimento por faixa etária
pnad_design %>%
  group_by(fx_idade)%>%
  summarize(rendimento = survey_mean(v4718, na.rm = TRUE))
```

- Qual o rendimento médio por sexo e raça?

Análises mediana survey_median()

Enap

```
# rendimento por sexo
pnad_design %>%
  group_by(v0302)%>%
  summarize(rendimento = survey_median(v4718, na.rm = TRUE))

# rendimento por faixa etária
pnad_design %>%
  filter(v8005>10)%>%
  group_by(fx_idade)%>%
  summarize(rendimento = survey_median(v4718, na.rm = TRUE))
```

- Compare a média e a mediana dos salários em uma mesma tabela

Análises dos quantis survey_quantile()

Enap

```
# distribuição da renda por grupos
pnad_design %>%
  summarize(rendimento =
              survey_quantile(v4718, c(0.25, 0.5, 0.75),
                              na.rm = TRUE, covmat = TRUE))

# rendimento por faixa etária
pnad_design %>%
  summarize(rendimento =
              survey_quantile(v4718, seq(0.1, 1, 0.1),
                              na.rm = TRUE))
```

Exercícios

Enap

- Resolver exercícios.