Lista 3: Esperança

FLP 0468/FLS 6183

Prof. Manoel Galdino
Gabriel Mardegan
Pedro Reis

Para entregar até: 29 de setembro de 2024

Exercícios

Esta terceira lista tem como objetivo continuar a revisão de estatística básica, com foco nas propriedades da esperança condicional. Note que ainda não entramos em exercícios relacionados diretamente aos conceitos de regressão linear, mas a lógica por detrás dos modelos de regressão já está presente.

Como base de dados, vamos utilizar a Pesquisa Nacional por Amostra de Domícilios Contínua (PNADc). Antes de iniciar os exercícios, procure pesquisar na internet o que é a PNAD e quais dados ela levanta.

1 Baixando o banco de dados PNADc (2023, 4o semestre)

O pacote "PNADcIBGE" permite importar bases de dados da PNAD diretamente para o environment do R. Primeiro, instale e ative o pacote. Depois, importe os dados do último trimestre de 2023 e as variáveis de interesse para esta lista, conforme o código no chunk abaixo:

```
#instale o pacote da PNAD
#install.packages("PNADcIBGE")

#carrege o pacote
library(PNADcIBGE)

#importe os dados de interesse através da função "get_pnadc"
dados_pnad <- get_pnadc(
    year = 2023,
    quarter = 4,
    design = FALSE,
    savedir = tempdir()
)
```

O banco de dados contém aproximadamente 473mil observações para 423 variáveis. O download pode demorar algum tempo.

Vamos entender cada linha do nosso código:

- Baixamos dados referentes a 2023 ("year = 2023") do 4o semestre ("quarter = 4").
- Por razões didáticas, selecionamos `design = FALSE` para ignorar o plano amostral. Optamos por importar apenas os microdados que serão utilizados nos exercícios da lista. Oportunamente, em sua pesquisa, é altamente recomendado deixar a função "design" como `TRUE`.
- "savedir" define o caminho do diretório onde o arquivo da base de dados será salvo. neste caso, num diretório temporário (default da função). É possível utilizar essa função para salvar os arquivos da base de dados em alguma pasta de seu computador, o que dispensará a necessidade de baixar os dados a todo momento.
- O seguinte link explica algumas das funções do pacote: https://cran.r-project.org/web/packages/PNADcIBGE/PNADcIBGE.pdf Muitas das informações parecem complicadas, mas é interessante explorar para começar a se familiarizar com documentos de apoio de banco de dados.

2 Limpando a base de dados

```r
#vamos selecionar apenas as variáveis úteis para a lista:
library(tidyverse)
library(tidylog)

dados_pnad <-
    dados_pnad %>%
    select(Ano, Trimestre, UF, V2007, VD4020, VD4035)

#renomeando as variáveis:
dados_pnad <-
    dados_pnad %>%
    rename(
        sexo = V2007,
        renda = VD4020,
        horas_trabalhadas = VD4035
    )

# utilize as funções "glimpse()", "View()" ou "head()"
# para inspecionar seu banco de dados
```

3 Estatística descritiva dos dados

Calcule:

1. A renda média;

2. A variância da renda;

3. A renda média dos homens e das mulheres;

4. A renda média em cada estado brasileiro;

5. A covariância entre a renda e o número de horas trabalhadas.

Nota: perceba que muitos valores do banco de dados aparecem como `NA`, que significa "not available" (não disponível). Ou seja, são valores ausentes, ou informações que faltam em nossa base, seja porque não foram coletados, sejam porque não existem. De toda forma, a presença de NAs na base de dados interfere nos cálculos que desejamos fazer. Para isso, é importante especificar "na.rm = TRUE" nas funções utilizadas, removendo os valores indisponíveis durante a operação. A função `cov()`, a ser utilizada no último item não possui o argumento "na.rm". Em seu lugar, use o argumento `use = "complete.obs"`, o que faz com que a covariância seja calculada usando apenas as observações completas.

Dica: a função `kable()` do pacote "knitr" dentro de seu pipe (%>%) gera uma tabela intuitiva para apresentação dos resultados.

4 Linearidade da esperança

Vamos aproveitar que temos uma base de dados concreta para articular a prática do R com conceitos trabalhados em aula, sobretudo o conceito de esperança matemática. Temos a seguinte equação:

$$\mathbb{E}[aX + bY] = a \times \mathbb{E}[X] + b \times \mathbb{E}[Y] \tag{1}$$

1. Considerando que `X = "renda"`, `Y = "horas_trabalhadas`, e "a" e "b" são constantes que multiplicam as variáveis (a = 2, b = 3), exemplifique a veracidade da equação utilizando os dados da PNAD.

2. O que o exercício demonstrou?

5 Esperança condicional

Agora trabalharemos explicitamente com a esperança condicional. Note que essa lógica estava implícita nas questões anteriores ("quero saber a média da renda de acordo com determinada característica da população"). Assuma novamente duas variáveis aleatórias, X e Y, tais que X = renda e Y = horas trabalhadas.

Calcule e interprete o resultado em sua resposta:

1. $\mathbb{E}[X \mid 10 \leq Y \leq 20]$

2. $\mathbb{E}[X \mid Y \leq 20]$

6 Probabilidade condicional

Para os itens desta questão, remova todas as observações cuja renda seja superior a 10.000 reias.

1. Apresente um gráfico de densidade da variável renda. Interprete.
2. Qual é a probabilidade de que, ao retirarmos aleatoriamente uma observação (um indivíduo/ respondente) dessa base de dados, sua renda seja estritamente maior do que 1000 e estritamente menor que 2000 reais? Apenas para propósitos didáticos, ignore o erro amostral e trate a sua base de dados como uma população (não faça isso em sua pesquisa).
3. Apresente um gráfico de densidade da renda dado que as horas trabalhadas (Y) sejam menores ou iguais a 20.
4. Calcule e interprete o resultado: $\mathbb{P}(1000 < X < 2000 \,|\, Y \leq 20)$

7 Questão extra/optativa [pós-graduação]

Utilize os dados da PNAD para praticar as propriedades da esperança abaixo.

1. Mostre que se:
$$\frac{\sum_{i=1}^{n} X_i}{n} = \overline{X} \tag{2}$$
Então:
$$\mathbb{E}\left[7 \times X_i - 7 \times \overline{X}\right] = 0 \tag{3}$$
2. Exemplifique a veracidade da equação:
$$\mathbb{E}\left[(X_i - \mathbb{E}[X_i])^2\right] = \mathbb{E}\left[X_i^2\right] - (\mathbb{E}[X_i])^2 \tag{4}$$

Nota final: Excetuando a questão 8 (de crédito extra), esta lista tem os mesmos exercícios para as turmas da graduação e da pós-graduação. Como foi avisado pelo moodle, os alunos da graduação podem enviar o script com as respostas dos exercícios em formato ".R". Para os alunos da pós-graduação, incentiva-se que busquem organizar suas listas no formato RMarkdown, enviando um artigo em formato ".Rmd" (com o script para replicação) e outro em PDF.

Como mencionado na lista 1, o RMarkdown permite gerar documentos em PDF a partir do script do R. Assim, é possível criar relatórios limpos, organizados e, principalmente, que contenham os produtos gerados a partir das nossas operações com as diversas ferramentas do software. Entretanto, é importante que esse arquivo em PDF contenha apenas aquilo que se quer apresentar ao leitor, dispensando ou ocultando códigos desnecessários, eventuais mensagens ou erros indicados pelo programa.

Para isso, o RMarkdown nos permite configurar os chunks de maneira que o relatório final contenha apenas aquilo que nos desejamos apresentar ao público em geral. Alguns comandos possíveis são:

- `echo` (`= TRUE` ou `= FALSE`) para apresentar o código do chunk no documento.

- `results` (`= "markup"` ou `= "hide"`) para apresentar os resultados da execução do código.
- `error` (`= TRUE` ou `= FALSE`) para apresentação dos erros.
- `warning` (`= TRUE` ou `= FALSE`) para apresentação de avisos.
- `message` (`= TRUE` ou `= FALSE`) para apresentação de mensagens.

Essas especificações devem ser escritas após a letra r no cabeçalho do chunk, como no exemplo a seguir:

````
```{r exemplo, echo=TRUE, error=FALSE, warning=FALSE, message=FALSE}
2^3  #exponenciação
```
````

Ao compilarmos o relatório com "Knit", a única coisa desse chunk que será impressa no documento final é o resultado do cálculo (8), nada do código ou o pensamento do R. O padrão (default) do programa leva à inclusão de todos os cinco tipos de conteúdo no relatório final.

Perceba que no exemplo nomeamos o chunk como "exemplo". Em longos scripts, isso pode ser uma boa estratégia para organizar os chunks, e ajudar nas etapas na escrita, assim como utilizamos nomes em capítulos de teses e dissertações.