lista3_gabarito

Gabriel_Mardegan

2024-09-30

Métodos quantitativos - regressão linear

Lista 3 - propriedades da esperança [gabarito]

monitor: Gabriel Rodrigues Mardegan

1. Importando a base de dados PNADc

A PNAD Contínua (Pesquisa Nacional por Amostra de Domicílios Contínua) é uma pesquisa realizada pelo IBGE (Instituto Brasileiro de Geografia e Estatística) que tem como objetivo principal fornecer informações sobre o mercado de trabalho, assim como características socioeconômicas da população brasileira, de forma contínua e trimestral. A pesquisa investiga temas como emprego, desemprego, renda, migração, educação, e outros aspectos, permitindo análises detalhadas sobre a evolução das condições de vida e trabalho no Brasil ao longo do tempo. Sua unidade de investigação é o domicílio (por isso o nome). Como o próprio nome já indica, trata-se de uma AMOSTRA, e por isso seus dados devem ser tratados enquanto tal. No link a seguir, vocês podem conferir a planilha do dicionário das variáveis utilizadas na PNADc de 2022:

link

2. Limpando a base de dados

```
#trata-se de um extenso banco de dados
#mais de 470mil observações para 423 variáveis
#para esta lista, vamos selecionar 6 variáveis:
#ano, trimestre, UF, sexo, renda e horas trabalhadas

#ao atribuir uma operação a um objeto de mesmo nome do dataframe anterior,
#você o substitui/atualiza em seu environment
dados_pnad <-
  dados_pnad %>%
  select(Ano, Trimestre, UF, V2007, VD4020, VD4035)
```

```r
#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

head(dados_pnad)
```

```
## # A tibble: 6 x 6
##   Ano   Trimestre UF       sexo   renda horas_trabalhadas
##   <chr> <chr>     <fct>    <fct>  <dbl>             <dbl>
## 1 2023  4         Rondônia Homem   5000                40
## 2 2023  4         Rondônia Mulher    NA                NA
## 3 2023  4         Rondônia Mulher  3500                36
## 4 2023  4         Rondônia Mulher    NA                NA
## 5 2023  4         Rondônia Mulher    NA                NA
## 6 2023  4         Rondônia Homem     NA                NA
```

3. Estatística descritiva dos dados

3.1) A renda média

renda_media
2828.871

3.2) A variância da renda

renda_variancia
19570671

3.3) A renda média dos homens e das mulheres

sexo	renda_media
Homem	3036.147
Mulher	2546.456

3.4) A renda média em cada estado brasileiro

UF	renda_media
Rondônia	2669.754
Acre	2406.400
Amazonas	2109.684
Roraima	2568.822
Pará	2187.960
Amapá	2806.506
Tocantins	2659.720
Maranhão	1510.355
Piauí	1946.789
Ceará	1733.271
Rio Grande do Norte	2360.079
Paraíba	2297.151
Pernambuco	1985.519
Alagoas	1925.597
Sergipe	1916.972
Bahia	1808.199
Minas Gerais	2783.375
Espírito Santo	2945.184
Rio de Janeiro	3481.701
São Paulo	3613.323
Paraná	3249.759
Santa Catarina	3410.939
Rio Grande do Sul	3471.415
Mato Grosso do Sul	3184.045
Mato Grosso	3355.650
Goiás	3179.700
Distrito Federal	5061.156

3.5) A covariância entre a renda e o número de horas trabalhadas

cov_renda_horas
9186.645

4.Linearidade da esperança

renda_horas_media_1
5771.216

media_renda	media_horas	renda_horas_media_2
2828.871	37.61353	5770.583

renda_horas_media_1	media_renda	media_horas	renda_horas_media_2
5771.216	2828.871	37.61353	5770.583

Os dois lados da equação chegaram no mesmo resultado (não precisamente iguais, devido a arrendondamento). Os cálculos do exercício demonstraram a **propriedade da linearidade da esperança matemática**. A esperança de uma combinação linear de variáveis aleatórias é igual à combinação linear das esperanças dessas variáveis. De maneira simplificada, a esperança da soma (das variáveis), é igual à soma das esperanças. Entretanto, é importante fazer uma ressalva: escolhemos duas variáveis que possuem grandezas que não são comparáveis (renda e horas trabalhadas), o que dificulta a interpretação dos resultados em termos práticos.

5. Esperança condicional

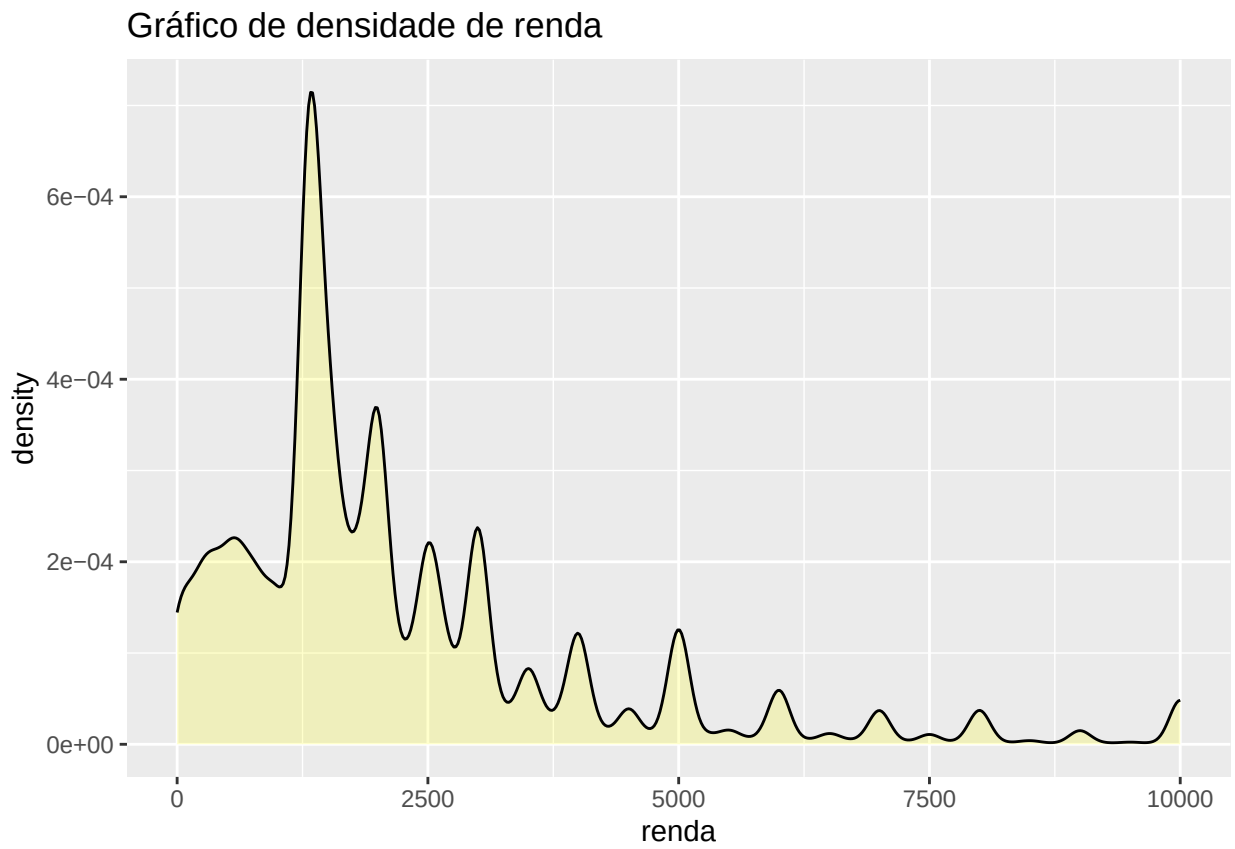
5.1) No ano de 2023, com os dados do PNAD referentes ao quarto semestre, considerando os entrevistados que trabalharam entre 10 e 20 horas semanais, a renda foi de 1348.17 reais (perceba, através do arquivo script, que utilizei aqui um código "in-line" para escrever o resultado da operação diretamente em meu texto).

5.2) No ano de 2023, com os dados do PNAD referentes ao quarto semestre, considerando os entrevistados que trabalharam 20 horas semanais ou menos, a renda foi de 1562.97 reais. É curioso que, ao abarcar entrevistados que trabalharam menos de 10 horas semanais, a média de rendimentos tenha aumentado em relação ao exercício anterior. Esse resultado suscita uma investigação para compreender quais fatores influiram nessa diferença entre quem trabalha a partir de 10 horas semanais, e quem trabalha menos de 10 horas semanais.

6. Probabilidade condicional

```
dados_pnad_6 <- dados_pnad %>%
  filter(renda <= 10000)
#o número de observações caiu mais da metade, de 473mil para 193mil
```

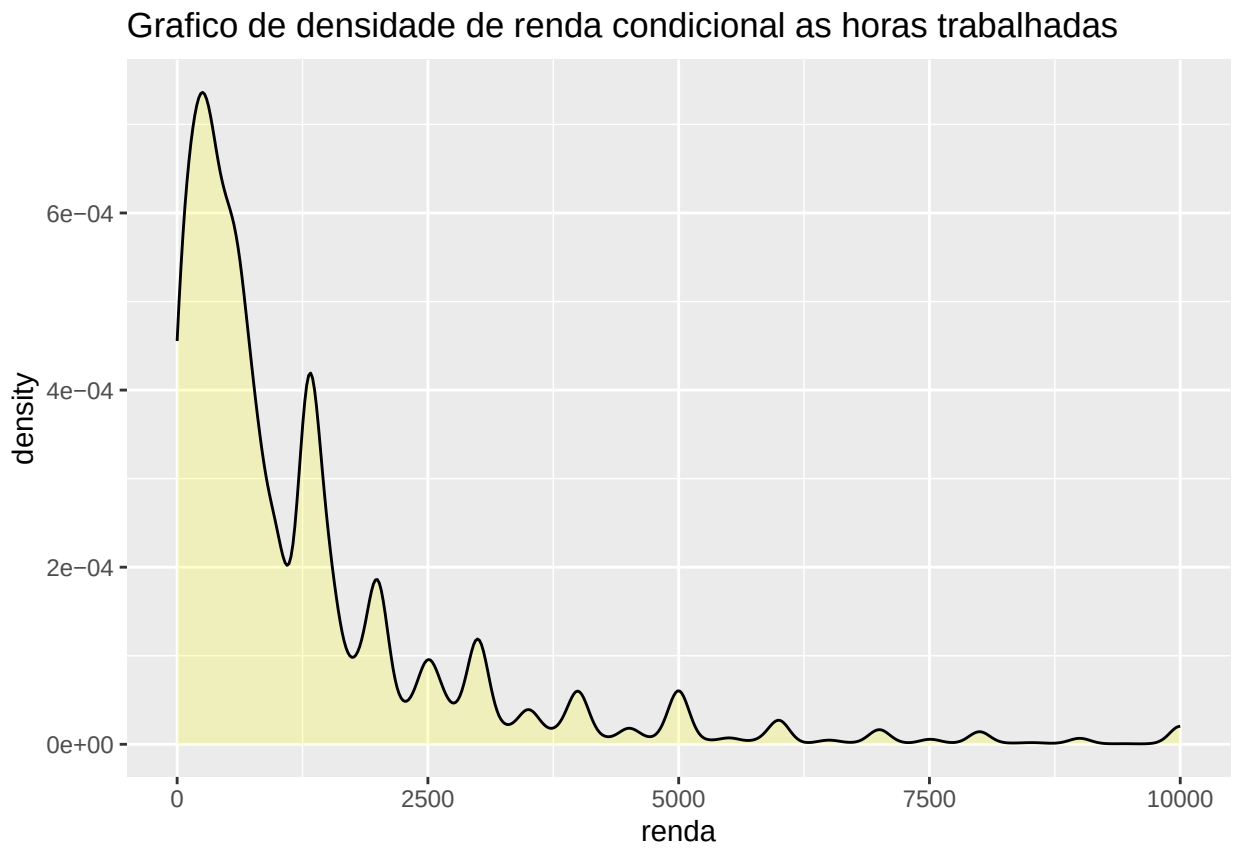
6.1) Gráfico de densidade



Percebe-se que a maioria dos respondentes afirmou ter rendimento próximo de 1250 reais no quarto semestre de 2023. Para subsidiar a análise, o salário mínimo em 2023 foi de R$1320,00 mensais. Na faixa de rendimentos entre R$5000,00 e R$10000, há uma minoria de respondentes. Esse gráfico está em consonância com as análises socioeconômicas acerca da grande parcela da população que ganha até 2 salários mínimos.

6.2) A probabilidade de um respondente escolhido aleatoriamente na amostra ter rendimento maior que 1000 e menor que 2000 reais é de 33.13%.

6.3) Gráfico de densidade da renda condicional ao número de horas trabalhadas



Ao comparar o gráfico criado para a questão 6.3 com o gráfico gerado na questão 6.1, percebe-se que o número de respondentes que tem rendimento próximo a um salário mínimo caiu em relação à densidade de respondentes que possuem rendimento próximo a zero. Uma hipótese para esse resultado, é que, ao filtrar entrevistados que trabalharam no máximo vinte horas semanais, boa parte daqueles registrados com 1sm ficaram de fora dos dados, e boa parte dos respondentes restantes tenham trabalhos informais, com menor rendimento. Essa hipótese suscita novas investigações de nossos dados, como por exemplo, verificar as variáveis referentes aos tipos de ocupação, com foco na proporção entre trabalhadores formais vs. informais.

6.4) Considerando respondentes que trabalharam até 20 horas semanais, a probabilidade de ser alguém com faixa de renda maior que 1000 e menor que 2000 reais, é de 18.4%.

7. Novamente as propriedades da esperança (questão extra/optativa)

7.1)

```
dados_pnad71 <- dados_pnad %>%
  mutate(renda1 = renda*7,
         renda2 = mean(renda, na.rm = TRUE)*7)

dados_pnad71 %>% summarise(esperanca71 = mean(renda1-renda2, na.rm = TRUE)) %>% kable()
```

esperanca71
0

A média das diferenças entre os valores das observações individuais Xi e a média de X é zero, e essa relação se mantém quando multiplicamos ambos os valores por uma mesma constante (7), conforme aponta a equação.

7.2)

```
## # A tibble: 1 x 2
##      renda1    renda2
##       <dbl>     <dbl>
## 1 19570573. 19570573.
```

```r
#resolvendo de maneira desagregada (passo a passo)

#primeira parte da equação
dados_pnad_73 <- dados_pnad %>%
  mutate(renda1 = (renda - mean(renda, na.rm = TRUE))^2) #dentro colchetes
#a esperança da primeira parte
dados_pnad_73 <- dados_pnad_73 %>%
  mutate(renda1 = mean(renda1, na.rm = TRUE))

#segunda parte da equação, primeira esperança
dados_pnad_73 <- dados_pnad_73 %>%
  mutate(renda2 = mean(renda^2, na.rm = TRUE))

#segunda parte da equação, segunda esperança
dados_pnad_73 <- dados_pnad_73 %>%
  mutate(renda3 = mean(renda, na.rm = TRUE)^2)

#resolvendo a segunda parte da equação
dados_pnad_73 <- dados_pnad_73 %>%
  mutate(renda4 = renda2 - renda3)

#verificando a equação
dados_pnad_73 %>%
  summarise(final = mean(renda1 - renda4, na.rm = TRUE))%>%
  kable()
```

final
0

Note que a primeira parte da equação é a definição de variância: subtrai-se a média da variável de cada

valor individualmente, e eleva-se ao quadrado antes de calcular a média. A segunda parte da equação é uma maneira alternativa de calcular a variância.

Créditos: Esta lista foi elaborada a partir da lista do ano anterior (2023), confeccionada por Davi Ferreira Veronese, a quem agradeço a gentileza.