Métodos Quantitativos IV - Lista 6 (resolução)

Departamento de Ciência Política da FFLCH-USP

Luiz Henrique da Silva Batista (Número USP: 12687228)

2023-11-30

Essa a lista compreende questões sobre a verificação dos pressupostos do modelo de regressão linear.

```
# Material de apoio para esta lista:
# https://jonnyphillips.github.io/Analise_de_Dados_2022/
```

Carregando pacotes

```
library(tidyverse)
library(janitor)
library(stargazer)
library(arrow)
library(dbplyr, warn.conflicts = FALSE)
library(duckdb)
library(tictoc)
```

Exercício 1

Nesta lista, utilizaremos os microdados do ENEM relativos a 2022. Faça o download dos dados do ENEM de 2022 e do dicionário de dados. Leia o dicionário de dados.

```
# Link: https://www.gov.br/inep/pt-br/acesso-a-informacao/dados-abertos/
# microdados/enem
# Selecione as variáveis NU_NOTA_CH e NU_NOTA_MT
```

Importando a base

Para importar a base, vamos transformar nosso dataset original em formato parquet, frequentemente usado ao se lidar com Big Data. A recomendação está no livro *R for Data Science*, escrito por Hadley, Mine Çetinkaya-Rundel and Garrett Grolemund.

```r
# Importando a base usando open_dataset

# setwd(dir = "C:\\sua\\estrutura\\de\\pastas\\no\\windows")
# setwd(dir = "C:/sua/estrutura/de/pastas/no/linux/ou/mac")

tic()
df <- open_dataset(
  # Defina aqui o caminho do seu arquivo MICRODADOS_ENEM_2022.csv
  sources = "datasets/microdados_enem_2022/DADOS/MICRODADOS_ENEM_2022.csv",
  format = "csv",
  delim = ";"
)
toc()
```

```
## 0.08 sec elapsed
```

```r
# Definindo caminho onde nossa base será salva
pq_path <- "datasets/microdados_enem_2022_pq"
# No meu caso, é na pasta "microdados_enem_2022_pq" do meu projeto R

# Transformando nosso df .csv em formato parquet (só é necessário fazer isso uma
# vez, por isso comentamos essa parte do código)
# tic()
# df |>
#   write_dataset(path = pq_path, format = "parquet")
# toc()

# Importando nossa base em formato parquet
tic()
enem_pq <- open_dataset(pq_path)
toc()
```

```
## 0.14 sec elapsed
```

As funções `tic()` e `toc()` funcionam como um relógio. Usamos somente para demonstrar a velocidade de leitura dos dados.

Importante: conforme estamos usando caminhos relativos para importar as bases de dados, lembre-se de indicar o caminho do seu PC. No início do código acima, a funçao `setwd()` define a estrutura de pastas no qual seu arquivo R Markdown está localizado, isto é, onde está no seu PC.

Limpe os dados para a análise.

```r
# Selecionando as colunas de interesse
dados <- enem_pq |>
  # Seleciona somente as colunas de interesse
  select(NU_INSCRICAO, NU_NOTA_CH, NU_NOTA_MT) |>
  # Apresenta (não converte!) os dados em formato data.frame
  collect() |>
  # Limpando os nomes das colunas
  janitor::clean_names() |>
  # Remove valores ausentes (só é possível depois de dar o collect)
  drop_na()
```

Exercício 2

Utilize a nota do ENEM em ciências humanas para prever a nota em matemática por meio de uma regressão linear simples. Interprete os resultados.

```r
reg <- lm(nu_nota_mt ~ nu_nota_ch, data = dados)

stargazer::stargazer(reg,
                     type = "latex",
                     style = "ajps",
                     title = "Regressão linear",
                     single.row = T,
                     keep.stat = c("n"),
                     dep.var.labels = "Nota em matemática",
                     header = F,
                     dep.var.caption= "")
```

Table 1: Regressão linear

	Nota em matemática
nu_nota_ch	0.905*** (0.001)
Constant	62.975*** (0.399)
N	2344823

***p < .01; **p < .05; *p < .1

Formalmente, o modelo pode ser representado pela seguinte expressão:

$$Nota\ em\ matemática = \alpha + \beta \cdot X_{Nota\ em\ ciências\ humanas}$$

Nosso modelo de regressão estima que a nota de matemática aumenta, em média, cerca de 0.905 pontos para cada 1 (um) ponto a mais obtido na prova de ciências humanas. Para

ajudar na interpretação, isso significa que para cada aumento de 10 pontos na prova de ciências humanas esperamos que a nota em matemática aumente 9.05 ($0.905 \cdot 10$). Além disso, quando a nota em ciências humanas é zero temos uma nota de matemática esperada de 62.975.

Exercício 3

Calcule os intervalos de confiança dos coeficientes.

Intervalo de confiança de 95% para α	Intervalo de confiança de 95% para β
(62.176; 63.774)	(0.904; 0.907)

Por que o R utiliza a distribuição t (e não a distribuição normal) para as inferências?

R: Utilizar a distribuição normal para calcular o erro padrão implica supor que sabemos o desvio padrão populacional (o que nem sempre é o caso) e que nossos dados estão normalmente distribuídos conforme indica o Teorema do Limite Central (68-95-99). A distribuição t-student, por outro lado, pode ser utilizada para "*qualquer* tamanho de amostra aleatória" (Agresti e Finlay, 2012, p. 142). O cálculo do erro padrão, assim, não é feito com desvio padrão populacional, mas com o desvio padrão amostral. Como consequência, a distribuição t-student é mais dispersa do que a normal. Se quisermos, por exemplo, calcular um intervalo com 95% de confiança usando a distribuição normal deveríamos fazer $\bar{y} + -1,96 \cdot \frac{\sigma}{\sqrt{n}}$, mas usando a dsitribuição t deveríamos fazer $\bar{y} + -2,048 \cdot \frac{s}{\sqrt{n}}$.

Exercício 4

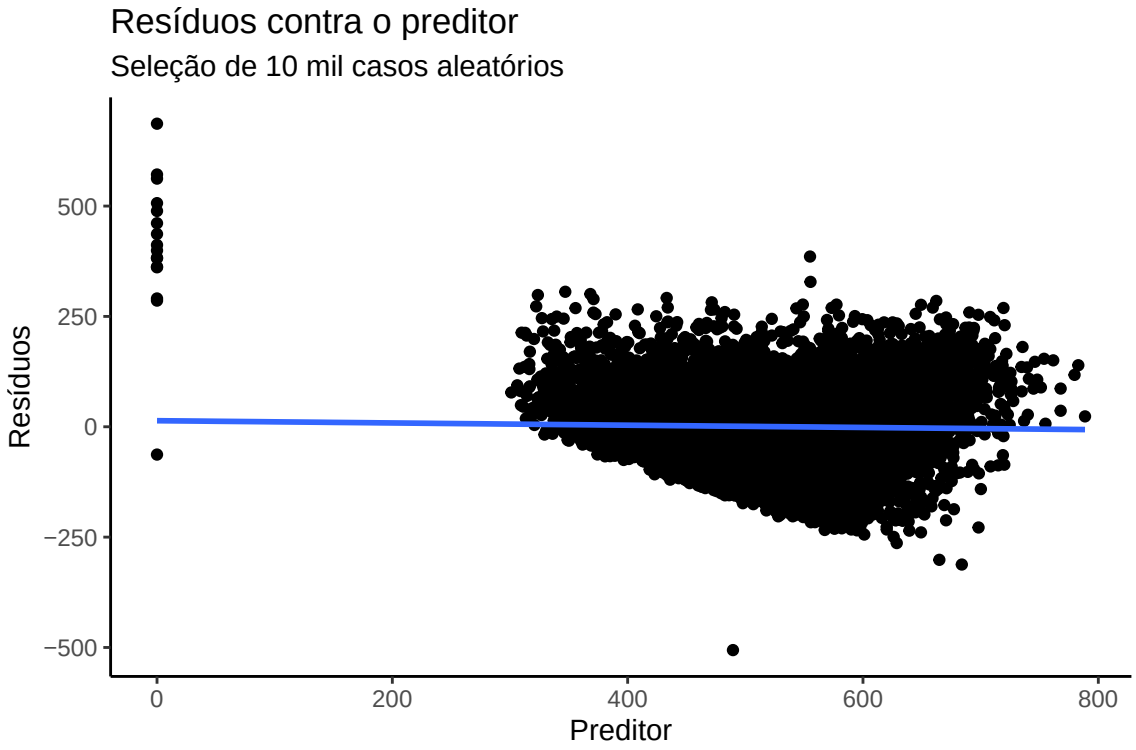
Utilize os resíduos para testar graficamente o seguinte: $E[\hat{e}|X = x] = 0$. Explique a sua conclusão.

```
residuos <- data.frame(residuos = residuals(reg),
                       preditor = dados$nu_nota_ch)

residuos |>
  # Seleciona 10 mil casos aleatórios
  sample_n(size = 10000) |>
  ggplot(aes(x = preditor, y = residuos)) +
  geom_point() +
  geom_smooth(method="lm", se=F) +
  theme_classic() +
  labs(
    title = "Resíduos contra o preditor",
    subtitle = "Seleção de 10 mil casos aleatórios",
```

```
    x = "Preditor",
    y = "Resíduos"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



R: O gráfico acima demonstra que $E[\hat{e}|X = x] = 0$. Conforme os resíduos representam a diferença entre a previsão do modelo de regressão $\hat{y}_i$ e o valor observado y_i (Galdino, 2023), uma média igual a 0 (zero), condicionada a um dado conjunto de valores aleatórios $X = x$, significa que nosso modelo não está "tendencioso", atribuindo mais (ou menos) peso a determinados valores. Em outras palavras, $E[\hat{e}|X = x] = 0$ deveria significar que não estamos percebendo nenhum padrão claro em uma dada distribuição aleatória dos nossos dados.

No entanto, embora tenhamos $E[\hat{e}|X = x] = 0$ no gráfico que plotamos, é possível identificar que os resíduos tendem a diminuir sistematicamente à medida que os valores do preditor se aproximam de 600. Além disso, a distribuição dos pontos parece apresentar algum padrão, não sendo totalmente aleatória a distribuição ao longo dos valores do nosso preditor.

Qual é a motivação desse teste? Em outras palavras, qual pressuposto do modelo linear gostaríamos de satistazer?

R: O objetivo desse teste é satisfazer o pressuposto de que os resíduos devem ter esperança igual a zero, ou seja, que $E[\hat{e}|X = x] = 0$.

Qual é a consequência da violação desse pressuposto?

R: Segundo Dalson (2019), violar esse pressuposto acarreta em viés no intercepto, o que prejudicaria a capacidade preditiva do modelo.

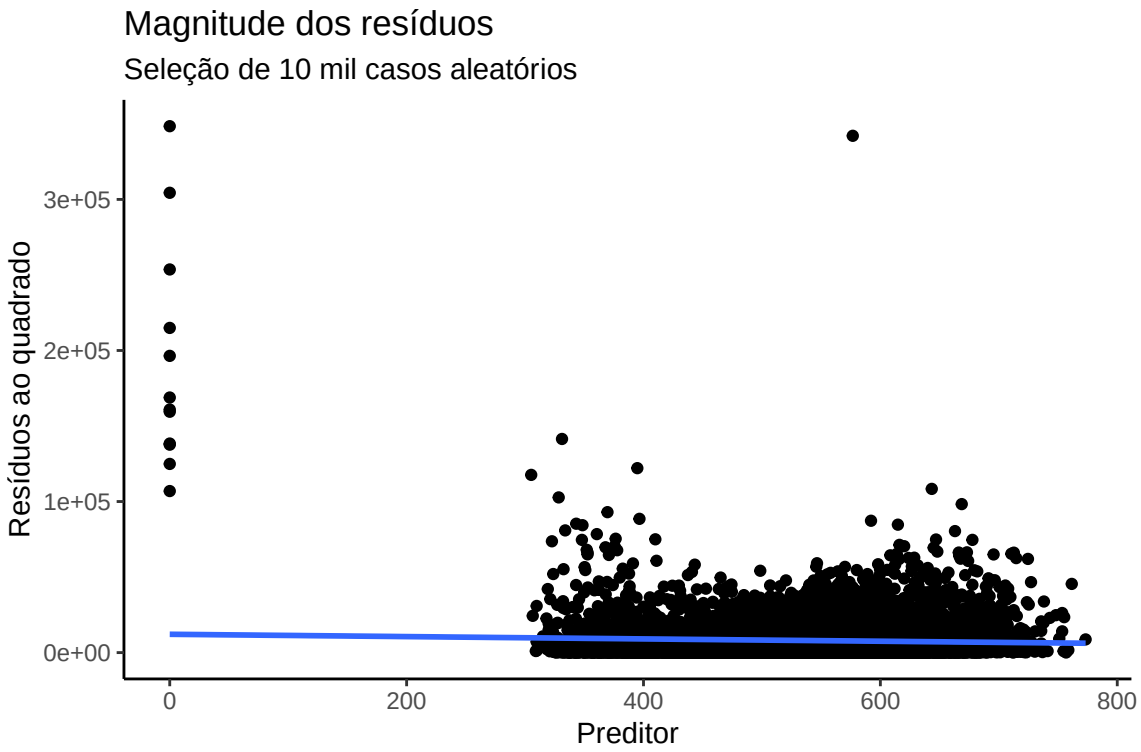
Exercício 5

Com base nos resíduos, utilize um gráfico para testar a hipótese de homocedasticidade dos erros. Explique a sua conclusão.

```
# Gráfico Magnitude dos Resíduos contra o Preditor
residuos_sq <- data.frame(residuos_sq = residuals(reg)^2,
                          preditor = dados$nu_nota_ch)

residuos_sq |>
  # Seleciona 10 mil casos aleatórios
  sample_n(size = 10000) |>
  ggplot(aes(x=preditor, y = residuos_sq)) +
  geom_point() +
  geom_smooth(method="lm", se=F) +
  theme_classic() +
  labs(
    title = "Magnitude dos resíduos",
    subtitle = "Seleção de 10 mil casos aleatórios",
    x = "Preditor",
    y = "Resíduos ao quadrado"
  )
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



R: Aqui, estamos verificando a relação entre os resíduos (ao quadrado) e a nossa variável independente. Como o gráfico acima, queremos visualizar o tamanho desses resíduos em relação aos valores da variável independente. Se o tamanho dos resíduos variar ao longo dos valores da v. independente, não estamos satisfazendo o pressuposto de homocedasticidade. Mas este não é o caso dos nossos dados, que satisfazem o pressuposto de homocedasticidade. Isso pode ser verificado a partir da reta plotada no gráfico, que é horizontal.

Qual é a consequência da violação desse pressuposto?

R: Se nossos resíduos variarem significativamente (para mais ou para menos) quando os valores da nossa v. independente aumentam ou diminuem, isso colocaria em xeque nossos intervalos de confiança e os testes de significância.

Exercício 6

Com base nos resíduos, utilize um gráfico para testar a hipótese de normalidade dos erros. Explique a sua conclusão.

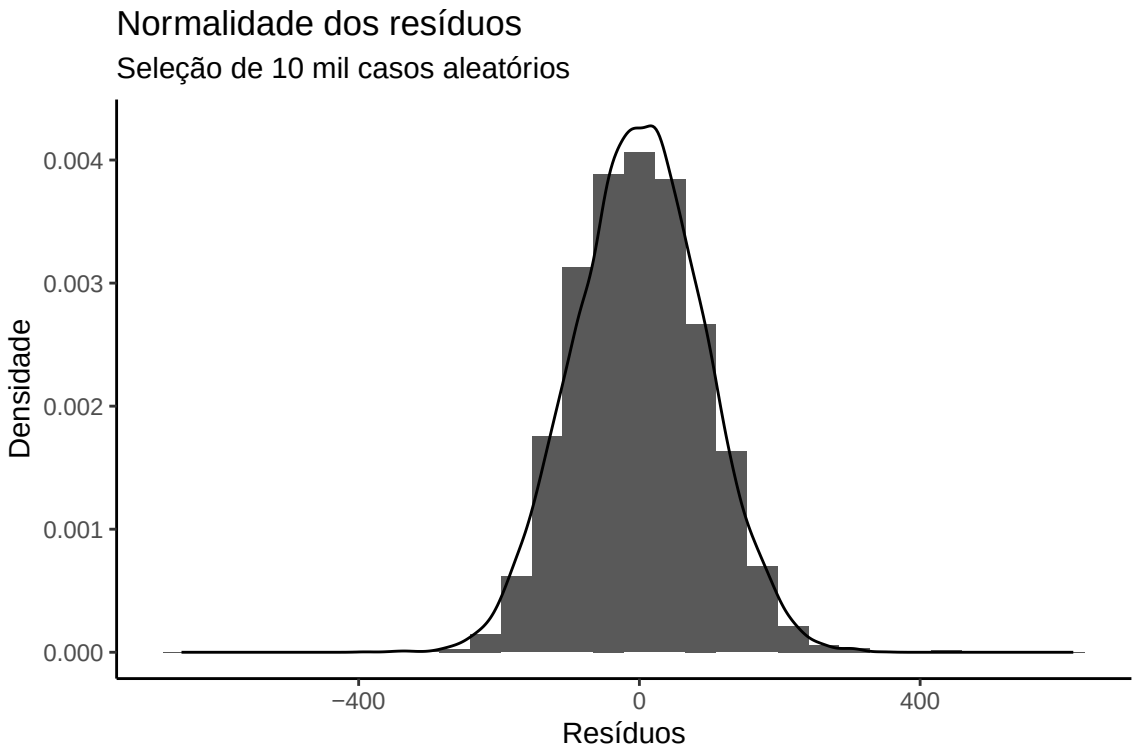
```
# Testando a normalidade
normalidade <- data.frame(residuos = residuals(reg),
                          preditor = dados$nu_nota_ch,
                          density_points = rnorm(
                            n = length(residuals(reg)),
                            mean = 0,
```

```
                            sd = sd(residuals(reg))
                            )
                          )

# print(sd(residuals(reg)))

normalidade |>
  # Seleciona 10 mil casos aleatórios
  sample_n(size = 10000) |>
  ggplot(aes(x = residuos)) +
  geom_histogram(aes(y = ..density..)) +
  geom_density(aes(x = density_points), colour = "black") +
  theme_classic() +
  labs(
    title = "Normalidade dos resíduos",
    subtitle = "Seleção de 10 mil casos aleatórios",
    x = "Resíduos",
    y = "Densidade"
    )
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



R: O gráfico compara um histograma dos resíduos do nosso modelo com uma curva de densidade teórica. Essa curva representa a distribuição dos resíduos se fossem distribuídos de

forma normal, com média zero e desvio padrão igual ao dos nossos resíduos. Essa comparação visual ajuda a entender se os resíduos se aproximam de uma distribuição normal. Como conclusão, podemos dizer que o pressuposto de normalidade dos resíduos pode ser satisfeita.

Qual é a consequência da violação desse pressuposto?

R: Violar esse pressuposto coloca em dúvida a confiança que podemos ter na construção dos nossos intervalos de confiança para os parâmetros e, portanto, também coloca em dúvida a confiança na inferência.

Exercício 7

Apresente seus resultados em um arquivo PDF. Garanta que seu arquivo esteja limpo, contendo as respostas, os gráficos e as tabelas, mas não eventuais mensagens e erros. O arquivo PDF pode ser gerado diretamente a partir do R por meio do RMarkdown ou do RSweave. Para os alunos de graduação, isso é recomendado, mas não obrigatório. Adicionalmente, forneça o script para replicação.