Métodos Quantitativos IV - Lista 7 (resolução)

Departamento de Ciência Política da FFLCH-USP

Luiz Henrique da Silva Batista (Número USP: 12687228)

2023-12-19

Esta lista compreende questões sobre regressão linear múltipla. **Os alunos devem entregar um arquivo PDF contendo as respostas e o script para replicação.**

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

Carregando pacotes

```
library(tidyverse)
library(janitor)
library(stargazer)
library(vroom)
library(knitr)
library(car)
```

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/ene

# Selecione as seguintes variáveis:
## NU_NOTA_CH
## NU_NOTA_MT
## NU_NOTA_CN
## NU_NOTA_LC
## TP_SEXO
```

```r
## TP_COR_RACA
## TP_ESCOLA

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

# Importando os dados
df <- vroom("datasets/microdados_enem_2022/DADOS/MICRODADOS_ENEM_2022.csv",
            delim = ";") |>
  # Limpa o nome das colunas
  janitor::clean_names() |>
  # Seleciona somente as colunas de interesse
  select(nu_nota_mt, nu_nota_ch, nu_nota_cn, nu_nota_lc,
         tp_sexo, tp_cor_raca, tp_escola)
```

Limpe os dados para a análise. Para resolver as questões, utilize uma amostra aleatória com 10.000 observações. Lembre-se de utilizar a função `set.seed()`.

Verifique como as variáveis estão codificadas e faça os ajustes necessários para a análise.

```r
set.seed(123)

# Verificando a codificação das variaveis
# glimpse(df)

dados <- df |>
  # Recodificando os valores das variáveis
  mutate(
    tp_cor_raca = case_when(
     # Codifica como NA a categoria 0 ("não declarada")
      tp_cor_raca == 0 ~ NA,
      tp_cor_raca == 1 ~ "branca",
      tp_cor_raca == 2 ~ "negra",
      tp_cor_raca == 3 ~ "negra",
      tp_cor_raca == 4 ~ "amarela",
      tp_cor_raca == 5 ~ "indígena",
      # Codifica como NA a categoria 6 ("Não dispõe da informação")
      tp_cor_raca == 6 ~ NA
      ),
    tp_escola = case_when(
      # Codifica como NA a categoria 1 ("não respondeu")
      tp_escola == 1 ~ NA,
      tp_escola == 2 ~ "pública",
      tp_escola == 3 ~ "privada"
      ),
```

```
    # Transforma as variáveis tp_sexo, tp_cor_raca e tp_escola em factor
    across(.cols = tp_sexo:tp_escola, .fns = as.factor)
    ) |>
  # Remove valores ausentes
  drop_na() |>
  # Seleciona somente os candidatos que obtiveram notas maiores que 0 em
  # todas as provas
  filter(if_all(
    .cols = contains("nota"),
    .fns = ~ . > 0
  )) |>
  # Seleciona 10 mil casos aleatorios
  sample_n(size = 10000)

# Verificando novamente a codificação das variaveis
# glimpse(dados)
```

No dataset original, as variáveis `tp_cor_raca` e `tp_escola` foram codificadas como numéricas (cada número fazendo referência a uma categoria). Assim, o R interpretou essas variáveis como numéricas, ao invés de categóricas, que é o que queremos. Várias eram as formas de transformar para formato `chr`, mas optamos por fazer referência direta às categorias. Assim, no caso da variável `tp_escola`, o número 2 se tornou "pública", que era a codificação indicada no dicionário.

Também transformamos as variáveis `tp_sexo`, `tp_cor_raca` e `tp_escola` em *factor*. Isso foi necessário para, posteriormente, alterarmos as categorias de referência ao rodar o modelo de regressão.

Além dessas transformações, removemos todos os `NA` do conjunto de dados e selecionamos somente os candidatos com nota maior que 0 (zero) em todas as provas. Assumimos que os candidatos com nota igual à 0 (zero) são aqueles que faltaram no dia da prova. Desse modo, optamos por rodar um modelo de regressão somente com os aqueles que compareceram nos dois dias de prova. Por fim, feito esses ajustes, selecionamos uma amostra de 10 mil casos aleatórios.

Exercício 2

Utilize as demais variáveis selecionadas para prever a nota de matemática por meio de uma regressão linear múltipla.

```
# Modelo
reg <- lm(nu_nota_mt ~ nu_nota_ch + nu_nota_cn + nu_nota_lc +
            tp_sexo + tp_cor_raca + tp_escola, data = dados)
```

Escreva a equação da regressão em notação escalar e em notação matricial.

Notação escalar

$$Nota\ em\ matemática = \alpha + \beta_1 \cdot X_{nota\ CH} + \beta_2 \cdot X_{nota\ CN} + \beta_3 \cdot X_{nota\ LC} + \beta_4 \cdot X_{sexo} + \beta_5 \cdot X_{raça} + \beta_6 \cdot X_{escola}$$

Notação matricial

$$\begin{bmatrix} \hat{y}_1 \\ \hat{y}_2 \\ . \\ . \\ . \\ \hat{y}_n \end{bmatrix} = \begin{bmatrix} 1 & nu\_nota\_ch_1 & nu\_nota\_cn_1 & nu\_nota\_lc_1 & sexo_1 & raca_1 & escola_1 \\ 1 & nu\_nota\_ch_2 & nu\_nota\_cn_2 & nu\_nota\_lc_2 & sexo_2 & raca_2 & escola_2 \\ 1 & . & . & . & . & . & . \\ 1 & . & . & . & . & . & . \\ 1 & . & . & . & . & . & . \\ 1 & nu\_nota\_ch_n & nu\_nota\_cn_n & nu\_nota\_lc_n & sexo_n & raca_n & escola_n \end{bmatrix} = \begin{bmatrix} \alpha \\ \beta_1 \\ \beta_2 \\ \beta_3 \\ \beta_4 \\ \beta_5 \\ \beta_6 \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ . \\ . \\ . \\ \varepsilon_N \end{bmatrix}$$

Apresente os resultados da regressão em uma tabela (stargazer).

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

Table 1: Regressão múltipla

	Nota em matemática
nu_nota_ch	0.414*** (0.017)
nu_nota_cn	0.383*** (0.015)
nu_nota_lc	0.310*** (0.016)
tp_sexoM	29.295*** (1.632)
tp_cor_racabranca	9.186 (6.121)
tp_cor_racaindígena	1.841 (13.367)
tp_cor_racanegra	−4.536 (6.092)
tp_escolapública	−25.708*** (2.169)
Constant	−21.126** (9.805)
N	10000

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

Exercício 3

Interprete o coeficiente estimado para cada uma das variáveis. Para as variáveis categóricas, qual é a categoria de referência? Explique o significado do coeficiente de cada categoria. Mude a categoria de referência e verifique se sua interpretação se altera.

R: Vamos fazer uma interpretação em partes, começando pelo intercepto, passando pelas variáveis numéricas e terminando com as variáveis categóricas.

Intercepto

Sabemos que o intercepto é o valor estimado quando todas as nossas variáveis preditoras são 0 (zero). Assim, nosso intercepto, aqui, significa o valor estimado para a nota em matemática quando a nota em ciências humanas, ciências da natureza e linguagens é 0 (zero), assim como quando a raça é amarela, o sexo é masculino e a escola é privada, que são as categorias de referência das variáveis categóricas No caso, nosso intercepto, considerando essas condições, é de -21.126, resultado nada factível.

Variáveis numéricas

Cada coeficiente estimado pelo modelo representa o efeito marginal preditivo *médio* sobre a nossa variável dependente, *mantendo todas as demais variáveis preditoras constantes*. Com isso em mente, nosso modelo de regressão estima que a nota de matemática aumenta cerca de:

- 0.414 pontos para cada 1 (um) ponto a mais obtido na prova de ciências humanas;
- 0.383 pontos para cada 1 (um) ponto a mais obtido na prova de ciências da natureza;
- 0.31 pontos para cada 1 (um) ponto a mais obtido na prova de linguagens.

Variáveis categóricas

Quando interpretamos as variáveis categóricas, é preciso ter em mente, assim como fizemos na interpretação das variáveis numéricas, que estamos falando do efeito marginal preditivo *médio*, *mantendo todas as demais variáveis preditoras constantes* e que estamos comparando com as respectivas categorias de referências.

Desse modo, no caso da variável *sexo*, nosso modelo estima que a nota em matemática:

- é cerca de 29.295 maior para candidatos do sexo masculino em comparação com candidatas do sexo feminino.

No caso da variável *raça*, não tivemos coeficientes estatisticamente significativos, considerando que a categoria de referência é *amarela*.

Por fim, para a variável *tipo de escola*, o modelo estima que a nota em matemática:

- diminui cerca de 25.708 em relação a candidatos de escolas privadas (categoria de referência).

Alterando a categoria de referência

Vamos alterar a categoria de referência das variáveis *sexo*, *raça* e *escola* e verificar os resultados.

```
# Alterando as categorias de referência
dados2 <- dados |>
  mutate(tp_sexo = fct_relevel(tp_sexo, "M"),
         tp_cor_raca = fct_relevel(tp_cor_raca, "branca"),
         tp_escola = fct_relevel(tp_escola, "privada"))

# Criando novo modelo de regressão
reg2 <- lm(nu_nota_mt ~ nu_nota_ch + nu_nota_cn + nu_nota_lc +
             tp_sexo + tp_cor_raca + tp_escola, data = dados2)

stargazer::stargazer(reg2,
                     type = "latex",
                     style = "ajps",
                     title = "Regressão múltipla",
                     single.row = T,
                     keep.stat = c("n"),
                     dep.var.labels = "Nota em matemática",
                     header = F,
                     dep.var.caption= "")
```

É possível perceber que os resultados são diferentes e isso acontece porque alteramos as categorias de referência. No primeiro modelo, as categorias de referência eram: ter obtido nota 0 (zero) em todas as demais provas, ser do sexo feminino, raça amarela e ter estudado em escola privada.

No segundo modelo, as categorias de referência são: ter obtido nota 0 (zero) em todas as demais provas, ser do sexo masculino, raça branca e ter estudado em escola privada.

Assim, a primeira diferença que podemos notar reside nos interceptos. No primeiro modelo, o intercepto era -21.126 e neste segundo modelo temos um intercepto de 17.355. Como foi dito, o intecepto é o valor quando todas as outras variáveis apresentam o valor 0 (zero).

O intecepto nem sempre é interpretável, como já sabemos.

Table 2: Regressão múltipla

	Nota em matemática
nu_nota_ch	0.414*** (0.017)
nu_nota_cn	0.383*** (0.015)
nu_nota_lc	0.310*** (0.016)
tp_sexoF	−29.295*** (1.632)
tp_cor_racaamarela	−9.186 (6.121)
tp_cor_racaindígena	−7.345 (12.033)
tp_cor_racanegra	−13.722*** (1.688)
tp_escolapública	−25.708*** (2.169)
Constant	17.355** (8.210)
N	10000

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

Em relação aos demais coeficientes, vemos que os valores estimados para as variáveis numéricas são os mesmos, mas temos mudanças nas variáveis categóricas.

Para a variável *sexo*, temos um coeficiente com sinal invertido agora:

- No modelo 2, a nota em matemática (mantendo tudo o mais constante) é cerca de 29.295 menor para candidatas do sexo feminino em comparação com candidatos do sexo masculino.

Para a variável *raça*, no modelo 1 os coeficientes para as categorias *branca*,*indígena* e *negra* não foram estatisticamente significativos quando comparados à categoria de referência *amarela*. No modelo 2, mudamos a categoria de referência para *branca* e o coeficiente estimado para *negra* deu estatisticamente significante:

- A nota de matemática é cerca de 13.722 menor em relação à *branca* (mantendo tudo o mais constante).

Exercício 4

Discuta as diferenças entre o modelo da lista 6 (com apenas uma variável independente) e o modelo desta lista (múltiplas variáveis independentes).

R: Na lista 6, rodamos um modelo de regressão lienar simples, ou seja, com apenas uma variável (nota obtida na prova de ciências humanas). No modelo desta lista, rodamos um modelo de regressão linear múltipla, ou seja, com mais de uma variável (nota obtida nas provas de ciências humanas, de ciências da natureza, de linguagens, sexo e raça do candidato e tipo de escola que estudou no ensino médio).

Desse modo, uma diferença fundamental que surge entre os modelos está na interpretação que fazemos dos coeficientes. Nos dois modelos, o intecepto representa o valor estimado quando todas as outras variáveis assume o valor 0. No entanto, em um modelo regressão linear múltipla, é preciso saber qual a categoria de referência (que será o valor 0) das variáveis qualitativas.

Além disso, ma regressão múltipla quando interpretamos um coeficiente qualquer, temos que ter em mente que estamos mantendo constante o efeito de todas as outras variáveis. Isso significa que, por exemplo, ao interpretar o impacto da nota obtida em ciências da natureza, estamos verificando o efeito exclusivo da nota obtido em ciências da natureza, isolando o efeito das demais variáveis (raça, sexo, etc.).

Exercício 5

Analise a plausibilidade da satisfação dos pressupostos do modelo linear. Analise a performance preditiva do modelo.

R: Abaixo, optamos por checar os pressupostos com base no segundo modelo.

Gráfico dos resíduos pelos valores ajustados

Aqui, os pontos deeriam se espalhar aleatoriamente ao redor da linha linha horizontal, sem seguir um padrão sistemático, sugerindo que a condição de esperança condicional zero dos resíduos está sendo atendida. No entanto, o gráfico abaixo mostra que existe um padrão na distribuição dos pontos.

```
ggplot() +
  aes(x = reg$fitted.values, y = resid(reg2)) +
  labs(
    title = "Resíduos vs. Valores Ajustados",
    subtitle = "Amostra de 10 mil casos aleatórios",
    x = "Valores Ajustados",
    y = "Resíduos") +
  geom_point() +
  geom_smooth(method="lm", se=F) +
  theme_classic()
```

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

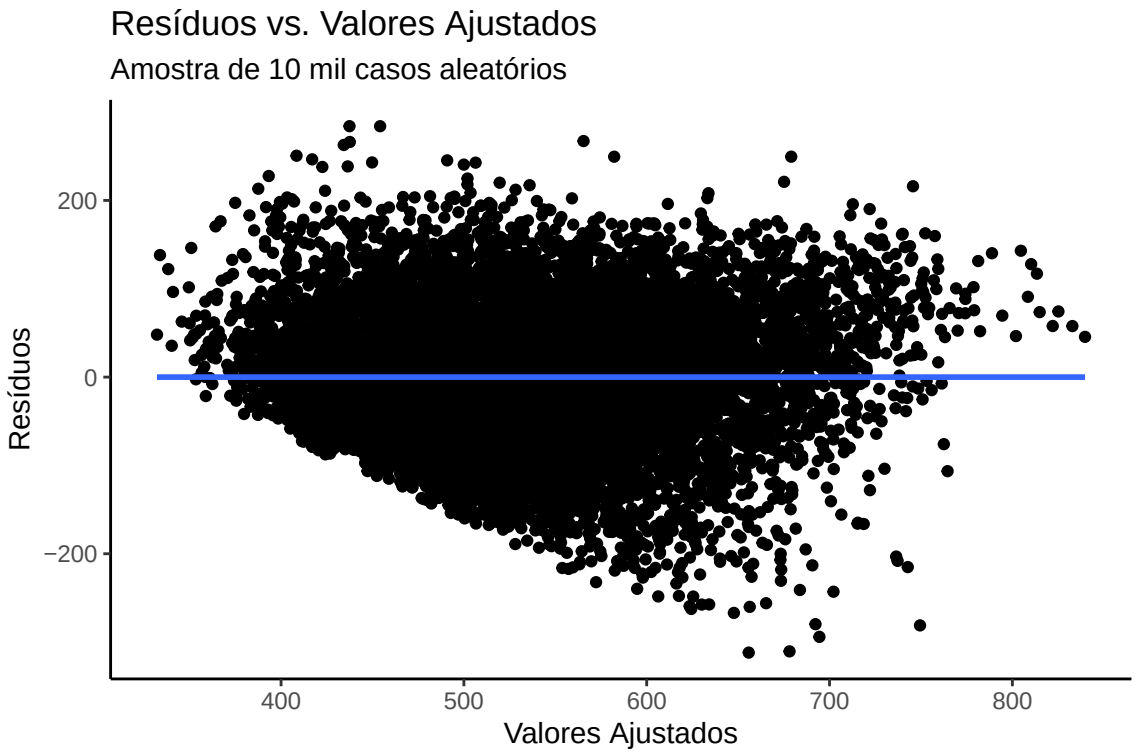
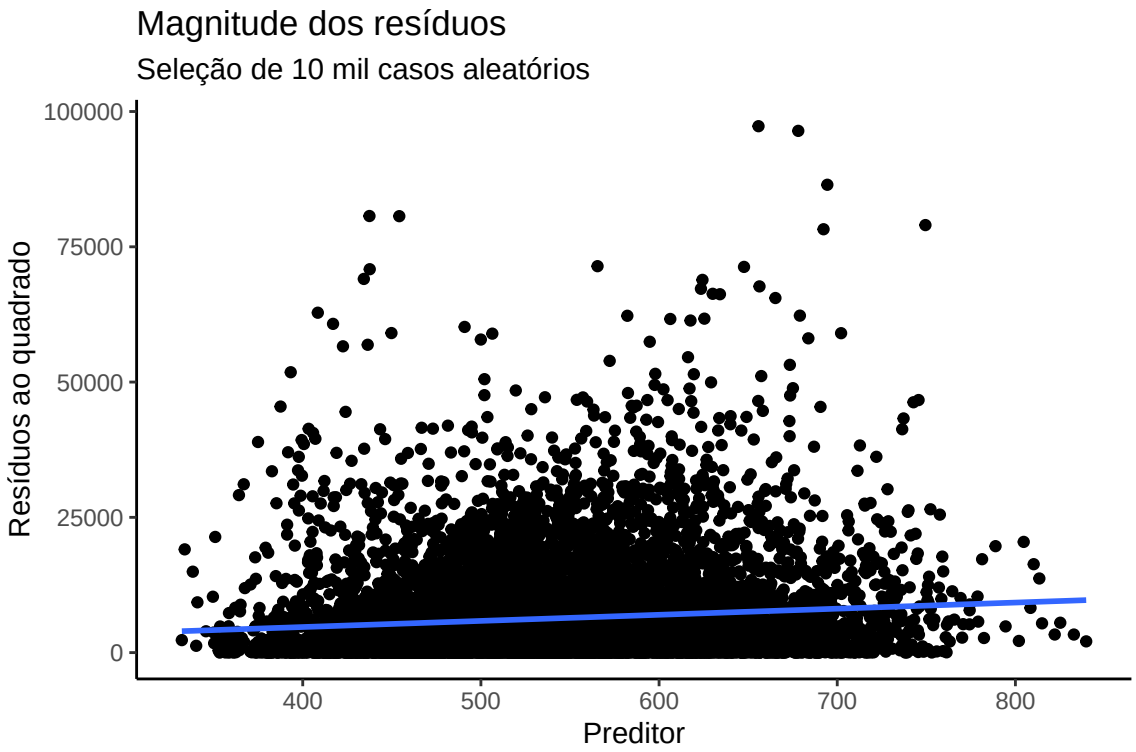


Gráfico Magnitude dos Resíduos contra o Preditor

Aqui, estamos testando a homocedasticidade dos resíduos. Para satisfazer esse pressupsoto, a reta plotada deve ser aproximadamente horizontal, indicando que não devemos ter sistematicamente mais pontos acima ou abaixo da reta. No gráfico abaixo, podemos perceber que a reta não é exatamente horizontal. Nesse caso, o recomedável seria construir inervalos de confiança com erro padrão robusto.

```
ggplot() +
  aes(x = reg$fitted.values, y = resid(reg2)^2) +
  labs(
    title = "Magnitude dos resíduos",
    subtitle = "Seleção de 10 mil casos aleatórios",
    x = "Preditor",
    y = "Resíduos ao quadrado"
  ) +
  geom_point() +
  geom_smooth(method="lm", se=F) +
  theme_classic()
```

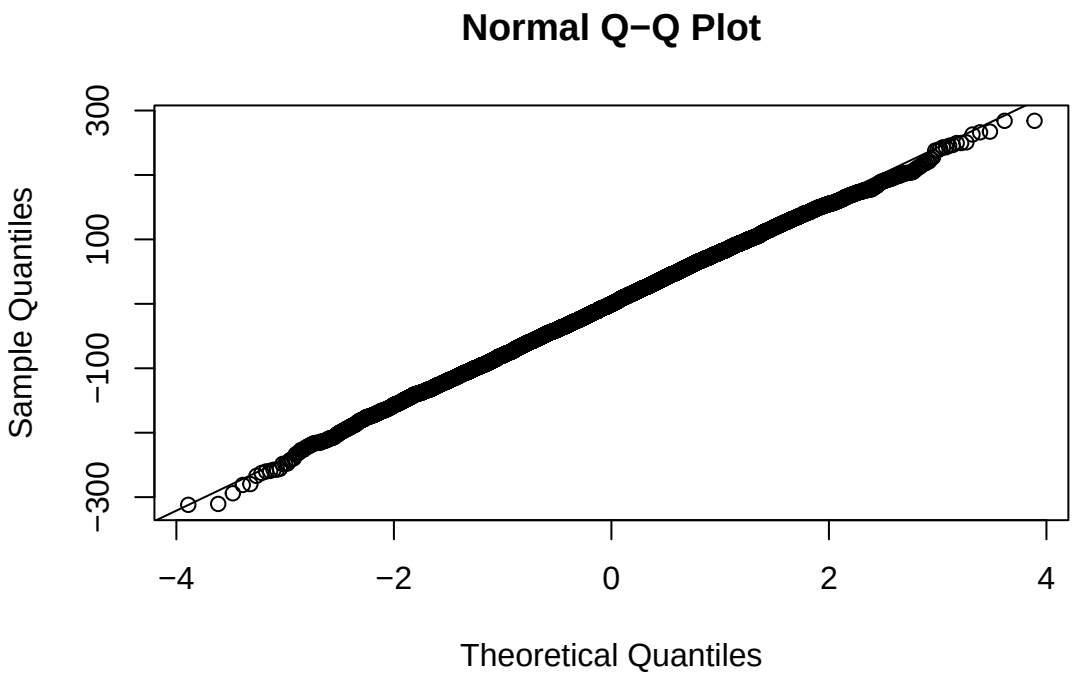
```
## `geom_smooth()` using formula = 'y ~ x'
```



Normalidade dos erros

O modelo satisfaz o pressuposto de normalidade dos erros

```
qqnorm(residuals(reg2))
qqline(residuals(reg2))
```



Multicolinearidade

```
vif(reg2) |> knitr::kable()
```

	GVIF	Df	GVIF^(1/(2*Df))
nu_nota_ch	2.506498	1	1.583192
nu_nota_cn	1.734200	1	1.316890
nu_nota_lc	2.390848	1	1.546237
tp_sexo	1.029600	1	1.014692
tp_cor_raca	1.111985	3	1.017848
tp_escola	1.224296	1	1.106479

A tabela VIF acima indica a presença de multicolinearidade entre as variáveis independentes. Existe uma problema de multicolinearidade quando este vif estiver acima de 10.

- As variáveis `nu_nota_ch` e `nu_nota_lc` apresentam uma influência moderada da multicolinearidade em suas estimativas.
- As demais variáveis (`nu_nota_cn`, `tp_sexo`, `tp_cor_raca`, `tp_escola`) têm baixa influência da multicolinearidade em seus coeficientes estimados.
- No geral, a multicolinearidade parece ser um problema moderado, pois a maioria das variáveis tem valores GVIF próximos de 1, indicando baixa influência da multicolinearidade nas estimativas do modelo.

Exercício 6

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.