Lego 2 – 2ª Lista de Exercícios

Felipe Lamarca

2025-10-27

Parte 1: Conceitual

Questão 1: algumas explicações curtas

a. *O conceito de homocedasticidade.* Trata-se da ideia de que a variabilidade das observações em diferentes grupos é constante – i.e., a variância é idêntica para todas as observações.

b. *O conceito de heterocedasticidade.* Ao contrário da homocedasticidade, a heterocedasticidade indica que a variabilidade das observações em diferentes grupos é distinta – i.e., a variância não é constante.

c. *A ideia de que pode haver correlação entre os erros das observações.* É a noção de que os erros das observações podem estar relacionados entre si. De fato, quando utilizamos, por exemplo, amostragem por conglomerados, ou quando entrevistamos indivíduos dentro de um mesmo domicílio, é intuitivo imaginar que as respostas estejam correlacionadas, assim como os erros. Parte da variação, nesse caso, tem a ver com fatores não observados, como características específicas do conglomerado ou do domicílio.

d. *Pressuposto de Homocedasticidade (ML.5).* Trata-se, no contexto de regressão, do pressuposto de que a matriz de variância-covariância dos erros é composta tão somente de uma diagonal com valores (variâncias) constantes. Embora não seja um pressuposto necessário (e às vezes sequer desejável), ele facilita as contas, conforme veremos no exercício 4.

e. *Pressuposto de Normalidade dos Erros (ML.6).* Trata-se da suposição de que os erros seguem uma distribuição normal com média zero e variância constante (isto é, inclui a suposição de homocedasticidade). É uma suposição auxiliar, sobretudo quando precisamos fazer inferências em amostras pequenas e não há garantia, pelo Teorema Central do Limite, de que $\widehat{\beta}$ segue distribuição normal.

f. *Erros robustos à heterocedasticidade, de Huber-White.* É uma flexibilização do ML.5. A hipótese nesse caso é que a matriz de variância-covariância dos erros é uma diagonal de valores potencialmente distintos entre si e com zeros ao redor. Mais especificamente, cada observação pode ter um desvio-padrão arbitrário qualquer – e, nesse caso, o próprio resíduo ao quadrado é uma boa aproximação.

g. *Erros clusterizados.* Semelhante ao que vimos em 1c, em que há covariância entre as observações. Suponha, novamente, uma amostragem por conglomerados. Nesse caso, há correlação espacial entre as observações e, portanto, os erros estão correlacionados.

h. *A diferença entre colinearidade perfeita e multicolinearidade.* Colinearidade perfeita, de saída, impossibilita o ajuste de modelos de regressão, porque torna a matriz X singular. Em particular, isso ocorre quando uma coluna da matriz X é idêntica a outra coluna, ou é uma combinação linear de uma ou mais colunas. A multicolinearidade, por sua vez, significa apenas que duas ou mais variáveis explicativas estão correlacionadas, mas não de forma perfeita.

i. *Fator de inflação de variância (ou Variance Inflation Factor – VIF).* Quando as variáveis explicativas estão correlacionadas entre si, fica mais difícil delimitar o efeito de cada covariável e, portanto, os erros-padrão são maiores. O VIF mede quanto a variância de um coeficiente de regressão é inflada devido à multicolinearidade.

Questão 2: Sobre o uso de diferentes tipos de estimadores de variância das estimativas da regressão

a. *Quando devemos usar erros homocedásticos (ou "clássicos")?* Devemos utilizar erros homocedásticos quando a variância dos erros em diferentes níveis da variável resposta é constante.

b. *Quando devemos usar erros heterocedásticos (Huber-White)?* Usamos erros heterocedásticos quando a variância dos erros em diferentes níveis da variável resposta é sistematicamente distinta.

c. *Por que, às vezes, a diferença entre eles e os erros clássicos podem ser um indicativo de má especificação funcional?* De fato, conforme King and Roberts (2015), a diferença entre os dois tipos de erro pode estar associada, na verdade, a um erro na forma funcional – dependendo da maneira como traçamos a curva, reta ou qualquer forma que seja, a variância poderia ser constante. Então, quando calculamos os erros de Huber-White e os homocedásticos e a diferença é muito significativa, provavelmente deveríamos dar um passo através e avaliar se estamos acertando a forma funcional.

d. *Quando devemos usar erros clusterizados?* Os erros clusterizados devem ser considerados quando há correlação entre as observações. Em particular, quando estamos tratando de *survey*, erros clusterizados normalmente são importantes quando a amostragem não segue o paradigma SRS (*Simple Random Sampling*) – isto é, quase sempre –, porque o próprio processo de amostragem introduz correlação espacial entre as observações (por exemplo, quando entrevistamos os indivíduos de um mesmo domicílio ou conglomerado).

Questão 3: derivando a matriz de variância-covariância de

Nesta questão, vamos derivar a fórmula mais geral da matriz de variância-covariância de $\hat{\beta}$, ou seja,

$$\text{Var}(\hat{\beta}) = (X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1}.$$

De saída, precisamos tomar a fórmula do estimador de $\hat{\beta}$, qual seja:

$$\hat{\beta} = (X^\top X)^{-1} X^\top y.$$

Sabemos que $y = X\beta + \vec{\epsilon}$. Substituindo na equação anterior, chegamos ao seguinte:

$$\begin{aligned}\hat{\beta} &= (X^\top X)^{-1} X^\top (X\beta + \vec{\epsilon}) \\ &= (X^\top X)^{-1} X^\top X\beta + (X^\top X)^{-1} X^\top \vec{\epsilon}.\end{aligned}$$

Apliquemos o operador de variância condicional dado X, isto é, dado o conjunto de dados efetivamente observado. Com isso, teremos:

$$\mathrm{Var}(\hat{\beta}|X) = \mathrm{Var}((X^\top X)^{-1} X^\top X\beta|X) + \mathrm{Var}((X^\top X)^{-1} X^\top \vec{\epsilon}|X).$$

Como β é fixo, não aleatório – afinal, representa os coeficientes verdadeiros na população –, a primeira variância é zero. Portanto, nos limitamos somente ao segundo termo, que pode ser simplificado mantendo o operador da variância aplicado apenas aos erros. Sendo $\Omega = \mathrm{Var}(\vec{\epsilon}|X)$, teremos:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}|X) &= \cancel{\mathrm{Var}((X^\top X)^{-1} X^\top X\beta|X} + \mathrm{Var}((X^\top X)^{-1} X^\top \vec{\epsilon}|X) \\ &= (X^\top X)^{-1} X^\top \mathrm{Var}(\vec{\epsilon}|X)((X^\top X)^{-1} X^\top)^\top \\ &= (X^\top X)^{-1} X^\top \mathrm{Var}(\vec{\epsilon}|X) X (X^\top X)^{-1} \\ &= (X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1}.\end{aligned}$$

Com isso, chegamos à expressão desejada.

Questão 4: a matriz de variância-covariância de $\hat{\beta}$ sob homocedasticidade

Sob homocedasticidade, a variância dos erros é constante. Matematicamente, isso significa que $\Omega = \sigma^2 I$, ou seja, a variância dos erros é a matriz identidade multiplicada por um único σ^2. Isso simplifica bastante as contas; veja:

$$\begin{aligned}\mathrm{Var}(\hat{\beta}) &= (X^\top X)^{-1} X^\top \Omega X (X^\top X)^{-1} \\ &= (X^\top X)^{-1} X^\top (\sigma^2 I) X (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1} X^\top X (X^\top X)^{-1} \\ &= \sigma^2 \cancel{(X^\top X)^{-1} X^\top X} (X^\top X)^{-1} \\ &= \sigma^2 (X^\top X)^{-1}.\end{aligned}$$

Parte 2: Prática

Na parte prática da lista, utilizaremos os dados do BH Area Survey, Pesquisa sobre a Região Metropolitana de Belo Horizonte, que foi a campo em 2002, sob coordenação da importantíssima socióloga Neuma Aguiar, então docente do Departamento de Sociologia da UFRM. Trata-se de um survey que utiliza amostragem probabilística em várias etapas e, portanto, é complexa.

Vamos começar importando as bibliotecas necessárias e os dados do survey. Além disso, vamos aplicar algumas manipulações disponibilizadas no próprio enunciado da lista.

```r
# importa as libs
library(tidyverse)
library(stargazer)
library(foreign)
library(fixest)
library(gt)
library(modelsummary)
library(survey)

# remove o uso de e^
options(scipen = 999)
options(modelsummary_factory_latex = "gt")

# leitura do banco de dados
p <- read.spss(file = "dados/BH Survey 2002/PRMBH_2002_DadosRecodificados.sav",
               use.value.labels = T,
               to.data.frame = T,
               reencode ="latin1")

# manipulacoes disponibilizadas no enunciado
p <- p %>%
  mutate(
    idade = as.numeric(E2),
    mulher = ifelse(E1 %in% "Mulher", 1, 0),
    brancos = case_when(
      RC4A %in% "Branco (a)" ~ 1,
      RC4A %in% "Amarelo (a)" ~ 1,
      RC4B %in% "Branco (a)" ~ 1,
      RC4B %in% "Amarelo (a)" ~ 1,
      RC4B %in% "Claro(a)" ~ 1,
      TRUE ~ 0),
    anosEstudo = ifelse(ANOSEST %in% "17 anos ou mais", "17", as.character(ANOSEST)),
    anosEstudo = as.numeric(anosEstudo),
    religiao = case_when(
      R4 %in% "Católico (a) não praticante" ~ "1. Catolico nao-praticante",
      R4 %in% "Católico (a) praticante" ~ "2. Catolico praticante",
      R4 %in% "Evangélico(a)" ~ "3. Evangelico",
      R4 %in% "Não tem religião?" ~ "5. Sem Religiao",
      TRUE ~ "4. Outra"
      ),
    classes_rendaDom = case_when(
      RENDA %in% "Até 1 salário mínimo (SM) (até R$ 200,00)" ~ "1. Até 1SM",
      RENDA %in% "Mais de 1 a 3 SM" ~ "2. 1 a 3 SM",
      RENDA %in% "Mais de 3 a 5 SM" ~ "3. 3 a 5 SM",
      RENDA %in% "Mais de 5 a 10 SM" ~ "4. 5 a 10 SM",
      RENDA %in% "Mais de 10 a 20 SM" ~ "5. 10 a 20 SM",
      RENDA %in% "Mais de 20 SM" ~ "6. 20+ SM",
```

```
      TRUE ~ NA_character_
      ),
    aborto = V9,
    aborto = ifelse(V9 %in% "Nunca aceitável", 1, aborto),
    aborto = ifelse(V9 %in% "Sempre aceitável", 10, aborto),
    aborto = ifelse(V9 %in% c("NS", "NR", "NA"), NA_character_, aborto),
    aborto = as.numeric(aborto)
    ) %>%
  select(
    idade,
    mulher,
    brancos,
    anosEstudo,
    religiao,
    classes_rendaDom,
    aborto,
    PESOFIM,
    F3
    )
```

As manipulações fazem o seguinte:

- A idade está medida em anos.
- "Mulher" é uma variável *dummy* indicadora do sexo feminino.
- "Brancos" é uma variável *dummy* indicadora da raça autodeclarada, "Branca" ou "Amarela" são 1.
- anosEstudo é a variável de anos de escolarização formal.
- A variável de religião agrupa categorias pequenas, com poucos casos.
- A renda domiciliar total está medida de forma categórica, em classes/faixas.

O propósito do exercício será o de tentar compreender a posição individual quanto ao tema do aborto. Utilizaremos a variável `V9` (aqui recodificada como `aborto`), que indica o nível de aceitação do indivíduo em relação ao aborto numa escala discreta de 1 a 10, na qual 1 significa "Nunca aceitável" e 10 indica "Sempre aceitável". Em particular, nosso objetivo é entender *o que determina patamares maiores de aceitação.*

Questão 5: Estatísticas descritivas

De saída, vamos apresentar estatísticas descritivas univariadas de todas as variáveis a serem utilizadas. Em particular, utilizaremos o pacote `survey` para incorporar o desenho amostral e facilitar o cálculo das estatisticas descritivas utilizando pesos amostrais.

```
# cria o desenho amostral
desenho <- svydesign(ids = ~1, weights = ~PESOFIM, data = p)

desc <- lapply(c("anosEstudo", "idade", "aborto"), function(v) {
  f <- as.formula(paste0("~", v))
  media <- svymean(f, desenho, na.rm = TRUE)
```

```r
  mediana <- svyquantile(f, desenho, quantiles = 0.5, na.rm = TRUE, ci = FALSE)
  var_ <- svyvar(f, desenho, na.rm = TRUE)
  dp <- sqrt(as.numeric(var_))
  n_validos <- sum(!is.na(p[[v]]))

  data.frame(
    variavel = v,
    media = as.numeric(media),
    mediana = as.numeric(mediana),
    desvio_padrao = dp,
    n_validos = n_validos,
    prop_validos = (n_validos / nrow(p)) * 100
  )
}) %>% bind_rows()

# frequencias das variaveis categoricas

# Mulheres
tab_mulher <- svytable(~ factor(mulher, exclude = NULL), desenho)
mulher_freq <- as.data.frame(tab_mulher)
names(mulher_freq) <- c("categoria", "n")
mulher_freq <- mulher_freq %>%
  mutate(prop = n / sum(n)) %>%
  arrange(desc(n), categoria)

# Brancos
tab_brancos <- svytable(~ factor(brancos, exclude = NULL), desenho)
brancos_freq <- as.data.frame(tab_brancos)
names(brancos_freq) <- c("categoria", "n")
brancos_freq <- brancos_freq %>%
  mutate(prop = n / sum(n)) %>%
  arrange(desc(n), categoria)

# Religião
tab_religiao <- svytable(~ factor(religiao, exclude = NULL), desenho)
religiao_freq <- as.data.frame(tab_religiao)
names(religiao_freq) <- c("categoria", "n")
religiao_freq <- religiao_freq %>%
  mutate(prop = n / sum(n)) %>%
  arrange(desc(n), categoria)

# Classes de renda do domicílio
tab_renda <- svytable(~ factor(classes_rendaDom, exclude = NULL), desenho)
classes_rendaDom_freq <- as.data.frame(tab_renda)
names(classes_rendaDom_freq) <- c("categoria", "n")
classes_rendaDom_freq <- classes_rendaDom_freq %>%
  mutate(prop = n / sum(n)) %>%
```

Table 1: Estatísticas descritivas das variáveis contínuas

Variável	Média	Mediana	Desvio-padrão	N válidos	% válidos
anosEstudo	7.324405	7	4.632490	1028	99.90282
idade	22.940924	20	16.360699	1029	100.00000
aborto	3.038374	1	2.957872	1021	99.22255

Table 2: Estatísticas descritivas das variáveis categóricas

(a) Sexo

Categoria	n	%
1	536.1992	52.1%
0	492.8009	47.9%

(b) Cor / raça

Categoria	n	%
0	626.9731	60.9%
1	402.0270	39.1%

(c) Religião

Categoria	n	%
2. Catolico praticante	446.35236	43.4%
3. Evangelico	235.48107	22.9%
1. Catolico nao-praticante	210.16105	20.4%
4. Outra	83.39586	8.1%
5. Sem Religiao	53.60974	5.2%

(d) Faixas de renda domiciliar

Categoria	n	%
2. 1 a 3 SM	302.86579	29.4%
4. 5 a 10 SM	228.32328	22.2%
3. 3 a 5 SM	202.83508	19.7%
5. 10 a 20 SM	118.96537	11.6%
6. 20+ SM	68.07135	6.6%
1. Até 1SM	65.97261	6.4%
NA	41.96660	4.1%

```
  arrange(desc(n), categoria)
```

```
desc %>%
  gt() %>%
  cols_label(
    variavel = "Variável", media = "Média", mediana = "Mediana",
    desvio_padrao = "Desvio-padrão", n_validos = "N válidos", prop_validos = "% válidos"
  )
```

Por ora não precisamos de recodificações adicionais, e nenhuma variável requer transformação logarítmica. Além disso, há algumas informações relevantes: em primeiro lugar, a aceitação do aborto era baixa, pelo menos no período em que os dados foram coletados. Além disso, a proporção de religiosos é alta: aproximadamente 65% da população era católica praticante ou evangélica; e a idade média era de 23 anos.

Questão 6: Ajuste de alguns modelos

Agora, ajustaremos algumas regressões múltiplas para modelar a aceitação do aborto, controlando pelas covariáveis já indicadas nas estatísticas descritivas. Mais especificamente, ajusto três modelos, cada

um assumindo diferentes hipóteses sobre a distribuição dos erros para o cálculo da variância dos $\hat{\beta}$'s: homocedasticidade, heterocedasticidade e erros clusterizados. Os resultados estão indicados na Table 3.

```
# remove os NANs
p_flt <- p %>% drop_na()

# formula da regressao
formula <- aborto ~ idade + mulher + brancos + anosEstudo + religiao +
  classes_rendaDom

# modelo base para os betas (de fato, o beta nao muda -- os erros sim!)
mod_homocedastico <- feols(formula, data = p_flt, weights = ~PESOFIM, vcov = "iid")
mod_heterocedastico <- feols(formula, data = p_flt, weights = ~PESOFIM, vcov = "hetero")
mod_clusterizado <- feols(formula, data = p_flt, weights = ~PESOFIM, vcov = cluster ~ F3)

tab <- modelsummary(
  list(
    "Homocedástico" = mod_homocedastico,
    "Heterocedástico" = mod_heterocedastico,
    "Erros clusterizados" = mod_clusterizado
  ),
  stars = TRUE,
  fmt = 3,
  output = "gt"
)

tab
```

Neste caso, é razoável afirmar que o erro-padrão mais adequado é aquele calculado *sob a hipótese de erros clusterizados*. De fato, quando consideramos o processo de amostragem complexa, em etapas, há correlação entre as observações – e esta hipótese lida com isso da forma mais adequada.

Questão 7: Uma interpretação do modelo com erros clusterizados

Os resultados indicam, de saída, que idade não apresenta um efeito estatisticamente significativo. Aliás, trata-se de uma variável cujo efeito provavelmente se confunde, pelo menos em parte, com os anos de estudo. O sexo, por sua vez, também não possui efeito estatisticamente significativo – com exceção do caso em que impus a hipótese de homocedasticidade aos erros, o que, de fato, subestima a variância do β.

Anos de estudo têm associação forte e estatisticamente significativa com a aceitação do aborto. Em particular, quando aumentamos os anos de estudo em uma unidade, a aceitação do aborto aumenta em 0.142, tudo mais constante. Ao mesmo tempo, a religião também parece ter um efeito negativo sobre a aceitação do aborto – um resultado esperado, naturalmente. Em relação a um católico não-praticante, um evangélico tem, em média e mantido os demais fatores constantes, 0.832 pontos a menos na aceitação do aborto.

Table 3: Modelos por abordagem para calcular a variância dos betas

	Homocedástico	Heterocedástico	Erros clusterizados
(Intercept)	1.323**	1.323**	1.323**
	(0.439)	(0.415)	(0.455)
idade	0.006	0.006	0.006
	(0.006)	(0.008)	(0.008)
mulher	-0.362*	-0.362	-0.362
	(0.179)	(0.227)	(0.235)
brancos	0.356+	0.356	0.356
	(0.188)	(0.240)	(0.222)
anosEstudo	0.142***	0.142***	0.142***
	(0.026)	(0.035)	(0.034)
religiao2. Catolico praticante	-0.202	-0.202	-0.202
	(0.237)	(0.313)	(0.284)
religiao3. Evangelico	-0.832**	-0.832**	-0.832**
	(0.270)	(0.294)	(0.302)
religiao4. Outra	-0.386	-0.386	-0.386
	(0.367)	(0.418)	(0.459)
religiao5. Sem Religiao	0.578	0.578	0.578
	(0.425)	(0.462)	(0.531)
classes_rendaDom2. 1 a 3 SM	0.476	0.476+	0.476
	(0.377)	(0.275)	(0.308)
classes_rendaDom3. 3 a 5 SM	0.525	0.525+	0.525+
	(0.397)	(0.316)	(0.312)
classes_rendaDom4. 5 a 10 SM	1.203**	1.203**	1.203**
	(0.412)	(0.406)	(0.404)
classes_rendaDom5. 10 a 20 SM	1.388**	1.388**	1.388**
	(0.457)	(0.458)	(0.436)
classes_rendaDom6. 20+ SM	2.542***	2.542***	2.542***
	(0.548)	(0.593)	(0.635)
Num.Obs.	963	963	963
R2	0.185	0.185	0.185
R2 Adj.	0.174	0.174	0.174
AIC	4633.6	4633.6	4633.6
BIC	4701.8	4701.8	4701.8
RMSE	2.64	2.64	2.64
Std.Errors	IID	Heteroskedasticity-robust	by: F3

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Questão 8: Outra especificação funcional

Já observamos que a religião é estatisticamente significativa no caso do grupo dos evangélicos. Agora, vamos verificar o que faz mais sentido: manter uma variável com todas as categorias de religião, ou criar uma variável binária indicando se o indivíduo é evangélico ou não.

Primeiro, vamos criar variáveis binárias para cada categoria de religião. Incluiremos todas elas no modelo 1 – com exceção de "católicos não-praticantes", que será nossa categoria de referência, algo necessário para garantir a invertibilidade de $X^\top X$. Depois, no modelo 2, incluo apenas a variável binária de evangélicos. Os resultados estão indicados na Table 4.

```
# binarias de cada religiao
p_flt <- p_flt %>%
  mutate(
    catolico_nao_praticante = as.integer(religiao == "1. Catolico nao-praticante"),
    catolico_praticante = as.integer(religiao == "2. Catolico praticante"),
    evangelico = as.integer(religiao == "3. Evangelico"),
    outra_religiao = as.integer(religiao == "4. Outra"),
    sem_religiao = as.integer(religiao == "5. Sem Religiao")
    )

# modelo 1 -- religiao com catolico nao praticante como referencia
formula1 <- aborto ~ idade + mulher + brancos + anosEstudo + classes_rendaDom +
  catolico_praticante + evangelico + outra_religiao + sem_religiao

mod1 <- feols(formula1, data = p_flt, weights = ~PESOFIM, vcov = ~F3)

# modelo 2 -- apenas evangelico
formula2 <- aborto ~ idade + mulher + brancos + anosEstudo + classes_rendaDom +
  evangelico

mod2 <- feols(formula2, data = p_flt, weights = ~PESOFIM, vcov = ~F3)

# tabelinha com gt
tab <- modelsummary(
  list("Todas as religiões" = mod1, "Evangélicos" = mod2),
  stars = TRUE,
  fmt = 3,
  output = "gt"
)

tab
```

Na prática, não há grandes diferenças entre os dois modelos do ponto de vista das medidas de ajuste. No entanto, o modelo 2 parece ter estimado com mais confiança (i.e., um menor desvio-padrão, comparativamente) o efeito de ser evangélico sobre a aceitação do aborto; além disso, o efeito de cada uma das demais religiões não é estatisticamente significativo. Por esses motivos, seguiremos com o modelo que inclui apenas a variável binária que indica se o indivíduo é ou não evangélico.

Table 4: Modelos com diferentes covariáveis de religião

	Todas as religiões	Evangélicos
(Intercept)	1.323**	1.281**
	(0.455)	(0.387)
idade	0.006	0.004
	(0.008)	(0.008)
mulher	-0.362	-0.423+
	(0.235)	(0.233)
brancos	0.356	0.390+
	(0.222)	(0.228)
anosEstudo	0.142***	0.138***
	(0.034)	(0.033)
classes_rendaDom2. 1 a 3 SM	0.476	0.508+
	(0.308)	(0.306)
classes_rendaDom3. 3 a 5 SM	0.525+	0.534+
	(0.312)	(0.316)
classes_rendaDom4. 5 a 10 SM	1.203**	1.205**
	(0.404)	(0.398)
classes_rendaDom5. 10 a 20 SM	1.388**	1.403**
	(0.436)	(0.431)
classes_rendaDom6. 20+ SM	2.542***	2.534***
	(0.635)	(0.624)
catolico_praticante	-0.202	
	(0.284)	
evangelico	-0.832**	-0.717***
	(0.302)	(0.205)
outra_religiao	-0.386	
	(0.459)	
sem_religiao	0.578	
	(0.531)	
Num.Obs.	963	963
R2	0.185	0.181
R2 Adj.	0.174	0.173
AIC	4633.6	4639.0
BIC	4701.8	4692.5
RMSE	2.64	2.66
Std.Errors	by: F3	by: F3

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

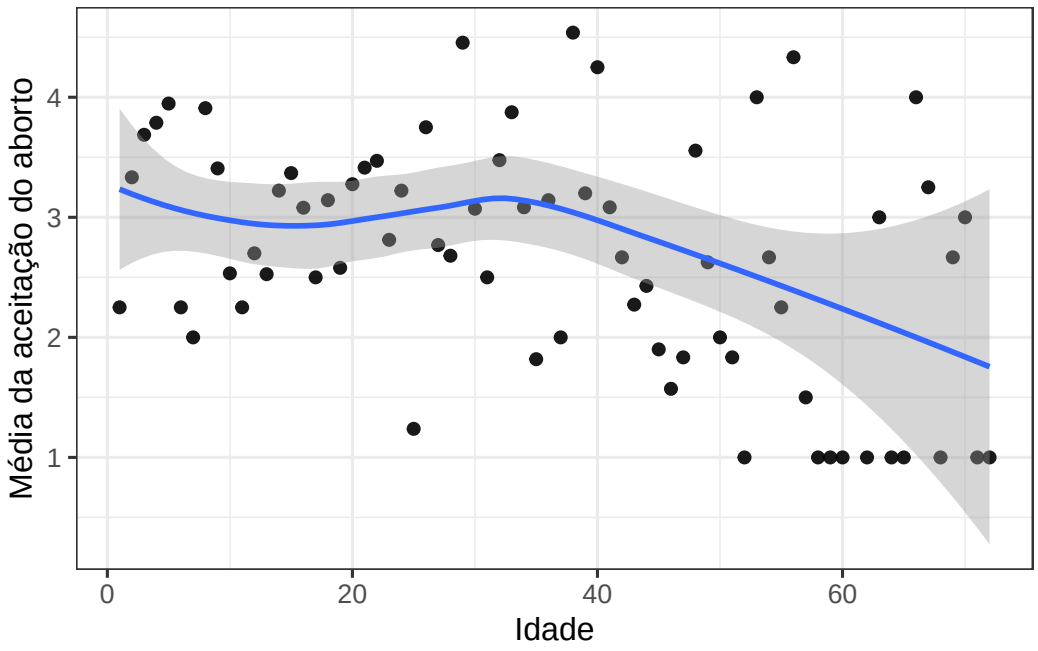
Questão 9: Especificações funcionais com polinômios

Agora, vamos verificar se há necessidade de incluirmos algum termo polinomial associado à idade. Vejamos a Figure 1, que mostra a distribuição da média de aceitação do aborto por idade.

```
# Média por idade
mean_by_age <- p_flt %>%
  group_by(idade) %>%
  summarise(media_aborto = mean(aborto, na.rm = TRUE), .groups = "drop")

ggplot() +
  geom_point(data = mean_by_age,
             aes(x = idade, y = media_aborto),
             size = 1.8, alpha = 0.9) +
  geom_smooth(data = p_flt,
              aes(x = idade, y = aborto),
              se = TRUE) +
  labs(
    x = "Idade",
    y = "Média da aceitação do aborto",
  ) +
  theme_bw(base_size = 12)
```

Figure 1: Média por idade com suavização não-paramétrica



De fato, parece que o efeito da idade não é exatamente linear. Com isso em mente, vamos ajustar 3 modelos, sendo o primeiro sem termo polinômial, e os outros dois incluindo um termo quadrático e um cúbico da idade, respectivamente. Os resultados estão discriminados na Table 5.

```r
# modelo "normal"
fml_lin <- aborto ~ idade + mulher + brancos + anosEstudo + classes_rendaDom +
  evangelico

# quadratico
fml_quad <- aborto ~ idade + I(idade^2) + mulher + brancos + anosEstudo +
  classes_rendaDom + evangelico

# quadratico + cubico
fml_cub <- aborto ~ idade + I(idade^2) + I(idade^3) + mulher + brancos +
  anosEstudo + classes_rendaDom + evangelico

# ajuste dos modelos
m_lin  <- feols(fml_lin,  data = p_flt, weights = ~PESOFIM, vcov = ~F3)
m_quad <- feols(fml_quad, data = p_flt, weights = ~PESOFIM, vcov = ~F3)
m_cub  <- feols(fml_cub,  data = p_flt, weights = ~PESOFIM, vcov = ~F3)

# apresenta a tabela
tab <- modelsummary(
  list("Linear" = m_lin, "Quadrático" = m_quad, "Cúbico" = m_cub),
  stars = TRUE,
  fmt = 3,
  output = "gt"
)

tab
```

É esperado, naturalmente, que os termos polinomiais não sejam estatisticamente significativos – mesmo porque eles são altamente correlacionados com o termo de idade simples. De todo modo, a inclusão dos termos polinomiais adiciona uma complexidade ao modelo que, na prática, não parece se justificar. Ao contrário: não há ganhos significativos, por exemplo, em R^2 ou RMSE, enquanto há um incremento (marginal) no AIC e no BIC. Por isso, o modelo que não inclui termos polinomiais parece o mais adequado.

Referências

King, Gary, and Margaret E. Roberts. 2015. "How Robust Standard Errors Expose Methodological Problems They Do Not Fix, and What to Do about It." *Political Analysis* 23 (2): 159–79. https://doi.org/10.1093/pan/mpu015.

Table 5: Modelos com termos polinomiais da idade

	Linear	Quadrático	Cúbico
(Intercept)	1.281**	1.132*	1.356*
	(0.387)	(0.496)	(0.528)
idade	0.004	0.023	-0.021
	(0.008)	(0.023)	(0.055)
mulher	-0.423+	-0.427+	-0.423+
	(0.233)	(0.232)	(0.232)
brancos	0.390+	0.399+	0.405+
	(0.228)	(0.226)	(0.226)
anosEstudo	0.138***	0.138***	0.138***
	(0.033)	(0.032)	(0.033)
classes_rendaDom2. 1 a 3 SM	0.508+	0.476	0.458
	(0.306)	(0.302)	(0.301)
classes_rendaDom3. 3 a 5 SM	0.534+	0.520	0.513
	(0.316)	(0.320)	(0.320)
classes_rendaDom4. 5 a 10 SM	1.205**	1.190**	1.174**
	(0.398)	(0.399)	(0.394)
classes_rendaDom5. 10 a 20 SM	1.403**	1.379**	1.374**
	(0.431)	(0.430)	(0.432)
classes_rendaDom6. 20+ SM	2.534***	2.491***	2.494***
	(0.624)	(0.632)	(0.635)
evangelico	-0.717***	-0.723***	-0.715***
	(0.205)	(0.204)	(0.204)
I(I(idade^2))		0.000	0.001
		(0.000)	(0.002)
I(I(idade^3))			0.000
			(0.000)
Num.Obs.	963	963	963
R2	0.181	0.182	0.184
R2 Adj.	0.173	0.173	0.173
AIC	4639.0	4640.2	4642.3
BIC	4692.5	4698.7	4705.6
RMSE	2.66	2.66	2.66
Std.Errors	by: F3	by: F3	by: F3

+ $p < 0.1$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$