Exercício 3 - Solução

Microeconometria - IBM0288

Prof. Raphael Gouvea

2026-04-09

1 Questão 3 - AP1 2025.02

a. Interprete o valor do coeficiente da variável Educ. O valor do coeficiente é estatísticamente diferente de zero?

O coeficiente estimado $\hat{\beta}_1 = 0{,}142$ indica que, em média, cada ano adicional de educação está associado a um aumento de aproximadamente **14,2%** no salário ($100 \times 0{,}142 = 14{,}2\%$), mantendo os demais fatores constantes. O coeficiente é significativo a 5% uma vez que o seu valor em módulo é mais que o dobro do desvio-padrão e, portanto, o intervalo de confiança em torno do valor estimado não inclui o zero.

> **Nota de interpretação:** Como a variável dependente é $\ln W$ e Educ é medida em anos (variável contínua), o coeficiente representa uma semielasticidade: uma variação de 1 unidade em Educ implica variação percentual de $100 \times \hat{\beta}_1$ em W.

b. Explique por que o modelo sofre de viés de variável omitida.

O modelo sofre de viés de variável omitida porque *habilidade* (A_i) satisfaz simultaneamente as duas condições necessárias para gerar este tipo de viés:

1. **Afeta a variável dependente:** pessoas mais hábeis tendem a ter salários mais altos, logo $\text{Corr}(A_i, \ln W_i) \neq 0$.
2. **Está correlacionada com o regressor incluído:** pessoas mais hábeis tendem a acumular mais anos de educação, logo $\text{Corr}(A_i, \text{Educ}_i) \neq 0$.

Como A_i não é observada, ela fica absorvida no erro u_i. Isso viola a hipótese de exogeneidade estrita ($\mathbb{E}[u_i \mid \text{Educ}_i] = 0$), de modo que o estimador de MQO é **viesado e inconsistente** para β_1.

c. Qual o sinal esperado do viés de variável omitida? Explique sua resposta.

O sinal do viés de variável omitida é dado por:

$$\text{Viés} = \tilde{\beta}_A \cdot \delta_{A,\text{Educ}}$$

onde $\tilde{\beta}_A$ é o coeficiente de *habilidade* em uma regressão de $\ln W$ sobre Educ e A, e $\delta_{A,\text{Educ}}$ é o coeficiente de *habilidade* em uma regressão auxiliar de A sobre Educ (que capta o sinal de $\text{Corr}(A_i, \text{Educ}_i)$).

Assim, assumindo que $\tilde{\beta}_A > 0$ (habilidade aumenta o salário); e $\delta_{A,\text{Educ}} > 0$ (habilidade está positivamente correlacionada com educação), temos que:

$$\text{Viés} = (+) \times (+) > 0$$

O estimador de MQO **superestima** o verdadeiro efeito causal de educação sobre salários. Parte do coeficiente estimado $\hat{\beta}_1$ captura o prêmio salarial associado à habilidade, e não apenas o retorno da educação em si.

d. Qual problema o pesquisador pretende evitar ao usar erros-padrão robustos conforme descrito no título da tabela?

O uso de **erros-padrão robustos à heterocedasticidade** visa evitar inferências equivocadas quando a variância do erro não é constante ao longo dos valores do regressor, isto é, quando $\text{Var}(u_i \mid \text{Educ}_i) \neq \sigma^2$.

Sob heterocedasticidade, os erros-padrão convencionais de MQO são inconsistentes: podem subestimar ou superestimar a variabilidade dos estimadores, tornando os testes t e F inválidos. Os erros-padrão robustos fornecem estimativas consistentes da matriz de covariância dos estimadores mesmo na presença de heterocedasticidade de forma desconhecida,sem exigir nenhuma suposição sobre a estrutura de $\text{Var}(u_i \mid X_i)$.

> **Importante:** o uso de erros-padrão robustos não corrige o viés de variável omitida; ele apenas garante inferência válida sobre os parâmetros estimados.

e. Algum tempo depois de estimar a regressão acima, o pesquisador encontra uma base de dados em painel em que cada indivíduo é observado por um período de 20 anos. Utilizando essa nova base de dados, seria possível eliminar o viés de variável omitida causado pelo fato de que habilidade é não observada? Seria necessário fazer algum ajuste para os erros-padrão do modelo a ser estimado? Justifique.

Sob a hipótese de que habilidade não varia no tempo, seria possível eliminar o viés de variável omitida utilizando um painel em que cada indivíduo i é observado por $T = 20$ anos ao se estimar um modelo de **efeitos fixos**:

$$\ln W_{it} = \alpha_i + \beta_1 \, \text{Educ}_{it} + u_{it}$$

O efeito fixo individual α_i absorve *todas* as características não observadas constantes no tempo — incluindo habilidade (A_i), desde que ela não varie ao longo dos 20 anos (a hipótese de identificação utilizada). A estimação do modelo de efeitos fixos elimina α_i, removendo, portanto, o viés causado por A_i.

A hipótese crucial é que **habilidade seja constante ao longo do tempo** para cada indivíduo. Se A_i variar (por exemplo, por acumulação de capital humano), o efeito fixo não resolverá o problema do viés.

Em relação aos erros-padrão, seria necessário utilizar erros-padrão clusterizados no nível do indivíduo, pois as observações de um mesmo indivíduo ao longo do tempo tendem a ser **correlacionadas entre si**. Ignorar essa estrutura torna os erros-padrão convencionais inconsistentes.

2 Reprodução Tabela 7.1

Antes de iniciar a atividade, vamos chamar os pacotes que serão utilizados. Nesta resolução são utilizados os seguintes pacotes do R. [1]

```
library(tidyverse) # Conjunto de pacotes e funções para realizar análise de dados no
library(readxl) # Para fazer leitura e também exportar arquivos em formato Excel.
library(gt) # Para elaboração de tabela com qualidade publicação
library(here) # Para facilitar trabalhar com caminhos de pastas
library(sandwich)   # vcovHC() para erros-padrão robustos
library(lmtest)     # coeftest()
```

Conforme descrito na atividade, o primeiro passo é ler os dados no R. Vamos salvar os dados no objetivo `caschool`.

```
# Carregar os dados
# Ajuste o caminho conforme a localização do arquivo na sua máquina
caschool <- read_excel(here::here("labs","SW_Datasets","caschool.xlsx"))
```

Agora, vamos estimar as cinco especificações da tabela 7.1.

```
#Especificação 1
espec_1 <- lm(testscr ~ str, data = caschool)

#Especificação 2
espec_2 <- lm(testscr ~ str + el_pct, data = caschool)

#Especificação 3
espec_3 <- lm(testscr ~ str + el_pct + meal_pct, data = caschool)

#Especificação 4
espec_4 <- lm(testscr ~ str + el_pct + calw_pct, data = caschool)

#Especificação 5
espec_5 <- lm(testscr ~ str + el_pct + meal_pct+ calw_pct, data = caschool)
```

[1] Como não quero reportar no arquivo pdf as mensagens que o R reproduz após a execução do comando `library`, inseri opções específicas para esse bloco de código. Para maiores detalhes, veja a página de referência sobre opções de execução.

Abaixo, segue o código e a Tabela Tabela 1 com os resulados gerados a partir dos modelos estimados. [2]

Utilizando os resultados da Tabela Tabela 1, seguem respostas para as perguntas realizadas em cada item. É importante ter clareza de que respostas diferentes também podem ser consideradas corretas, desde que a justificativa dada seja consistente com os conceitos econométricos estudados.

a. A especificação (3) poderia ser escolhida como especificação mais apropriada, uma vez que além da variável de interesse controla por características dos alunos que poderiam gerar viés de variável omitida caso não fossem incluídas no modelo. As especificações (4) e (5) apresentam especificações alternativas para diferentes formas de se controlar pelas condições socioeconômicas dos alunos.

b. O resultado reportado na coluna (3) indica que, ao reduzir a razão aluno-professor em um aluno por professor, a média na nota do teste aumenta aproximanete em 1 ponto, mantendo-se constante as características dos alunos. Note que o resultado é robusto, ou seja, o coeficiente é realitivamente simliar para as demais especificações com controles.

c. Caso responda sim, o aluno deve justificar porque o modelo não sofre de viés de variável omitida, ou seja, que não há nenhuma variável não incluída no modelo que seja correlacionada com a razão aluno-professor e as notas médias do teste.

Caso responda não, o aluno deve apresentar outros mecanismos omitidos que estejam correlacionados com a razão aluno-professor e as notas médias do teste.

d. Eu pediria ao secretário para me apresentar por quais motivos ele acreditava que os resultados do estudo feito para Califórnia poderiam ser extrapolados para população do município dado as grandes diferenças socioeconômicas, culturais e institucionais entre o estado americano e nosso país. Caso a explicação não fosse convincente, eu chegaria a conclusão que o argumento teria um problema de validade externa.

```
# ------------------------------------------------------------------------------
# 1. MODELOS EM LISTA
# ------------------------------------------------------------------------------

modelos <- list(espec_1, espec_2, espec_3, espec_4, espec_5)

# ------------------------------------------------------------------------------
# 2. FUNÇÃO AUXILIAR: extrai coef + EP robusto HC1
# ------------------------------------------------------------------------------

extrair <- function(modelo, var) {
  ct <- coeftest(modelo, vcov. = vcovHC(modelo, type = "HC1"))
  if (var %in% rownames(ct)) {
    list(
      coef = formatC(ct[var, "Estimate"],   digits = 3, format = "f"),
```

[2] Para essa tabela, eu pedi auxílio do Claude. Veja o prompt inicial no anexo. Depois, revisei o código e fiz alguns ajustes para corrigir erros e detalhes de formatação.

```
      se   = formatC(ct[var, "Std. Error"], digits = 3, format = "f")
    )
  } else {
    list(coef = "", se = "")
  }
}

# ------------------------------------------------------------------------------
# 3. CONSTRUÇÃO DO DATA FRAME
# ------------------------------------------------------------------------------

# Variáveis na ordem desejada e seus rótulos
variaveis <- c("str", "el_pct", "meal_pct", "calw_pct")

label_map <- c(
  str      = "Razão Aluno-Professor",
  el_pct   = "% Alunos Aprendendo Inglês",
  meal_pct = "% Alunos c/ Lanche Subsidiado",
  calw_pct = "% Alunos c/ Assist. por Renda"
)

# Monta as linhas de coeficientes explicitamente como data.frame
linhas_list <- vector("list", length(variaveis) * 2)
idx <- 1

for (var in variaveis) {
  coefs <- sapply(modelos, function(m) extrair(m, var)$coef)
  ses   <- sapply(modelos, function(m) {
    se <- extrair(m, var)$se
    if (se == "") "" else paste0("(", se, ")")
  })

  # Linha do coeficiente
  linhas_list[[idx]] <- data.frame(
    variavel = label_map[[var]],
    M1 = coefs[1], M2 = coefs[2], M3 = coefs[3],
    M4 = coefs[4], M5 = coefs[5],
    stringsAsFactors = FALSE
  )
  idx <- idx + 1

  # Linha do erro-padrão (rótulo vazio)
  linhas_list[[idx]] <- data.frame(
    variavel = "",
    M1 = ses[1], M2 = ses[2], M3 = ses[3],
    M4 = ses[4], M5 = ses[5],
    stringsAsFactors = FALSE
  )
```

```
  idx <- idx + 1
}

tabela_coef <- do.call(rbind, linhas_list)

# Estatísticas de ajuste
ser   <- sapply(modelos, function(m) formatC(summary(m)$sigma,           digits = 3,
r2adj <- sapply(modelos, function(m) formatC(summary(m)$adj.r.squared,   digits = 3,
n_obs <- sapply(modelos, function(m) as.character(nobs(m)))

tabela_estat <- data.frame(
  variavel = c("SER", "R² Ajustado", "n"),
  M1 = c(ser[1], r2adj[1], n_obs[1]),
  M2 = c(ser[2], r2adj[2], n_obs[2]),
  M3 = c(ser[3], r2adj[3], n_obs[3]),
  M4 = c(ser[4], r2adj[4], n_obs[4]),
  M5 = c(ser[5], r2adj[5], n_obs[5]),
  stringsAsFactors = FALSE
)

# Tabela completa
tabela_completa <- rbind(tabela_coef, tabela_estat)

# Índices de linha para os grupos
idx_str      <- 1:2                              # str: coef + EP
idx_controle <- 3:nrow(tabela_coef)              # demais regressores
idx_estat    <- (nrow(tabela_coef) + 1):nrow(tabela_completa)

# Índices das linhas de EP (para itálico) - linhas pares dentro de tabela_coef
idx_ep <- seq(2, nrow(tabela_coef), by = 2)

# ------------------------------------------------------------------------------
# 4. TABELA COM gt
# ------------------------------------------------------------------------------

tab <- tabela_completa |>
  gt(rowname_col = "variavel") |>

  # Cabeçalho
  tab_header(
    title    = "Determinantes do Desempenho Escolar na Califórnia",
    subtitle = "Variável dependente: nota média nos testes (testscr)"
  ) |>

  # Rótulos das colunas
  cols_label(
    M1 = "(1)", M2 = "(2)", M3 = "(3)", M4 = "(4)", M5 = "(5)"
  ) |>
```

```r
# Grupos de linhas
tab_row_group(label = "Estatísticas de Ajuste",  rows = idx_estat)    |>
tab_row_group(label = "Variáveis de Controle",   rows = idx_controle) |>
tab_row_group(label = "Regressor de Interesse",  rows = idx_str)      |>

# Alinhamento
cols_align(align = "center", columns = c(M1, M2, M3, M4, M5)) |>
cols_align(align = "left",   columns = variavel) |>

# Itálico + cor cinza nas linhas de EP
tab_style(
  style     = cell_text(style = "italic", color = "grey40", size = "small"),
  locations = cells_stub(rows = idx_ep)
) |>
tab_style(
  style     = cell_text(style = "italic", color = "grey40", size = "small"),
  locations = cells_body(rows = idx_ep)
) |>

# Negrito nos cabeçalhos de coluna
tab_style(
  style     = list(cell_text(weight = "bold"), cell_fill(color = "#f2f2f2")),
  locations = cells_column_labels()
) |>

# Notas de rodapé
tab_footnote(
  footnote  = "Erros-padrão robustos HC1 entre parênteses.",
  locations = cells_column_labels(columns = M1)
) |>
tab_footnote(
  footnote  = "SER = desvio-padrão dos resíduos (Root MSE).",
  locations = cells_stub(rows = which(tabela_completa$variavel == "SER"))
) |>

# Opções gerais
tab_options(
  table.font.names               = "Arial",
  table.font.size                = 12,
  heading.title.font.size        = 14,
  heading.title.font.weight      = "bold",
  row_group.font.weight          = "bold",
  row_group.background.color     = "#e8edf2",
  column_labels.background.color = "#d0d8e4",
  table.border.top.style         = "solid",
  table.border.top.width         = px(2),
  table.border.bottom.style      = "solid",
```

Tabela 1

Determinantes do Desempenho Escolar na Califórnia

Variável dependente: nota média nos testes (testscr)

	(1)[1]	(2)	(3)	(4)	(5)
Regressor de Interesse					
Razão Aluno-Professor	-2.280	-1.101	-0.998	-1.308	-1.014
	(0.519)	*(0.433)*	*(0.270)*	*(0.339)*	*(0.269)*
Variáveis de Controle					
% Alunos Aprendendo Inglês		-0.650	-0.122	-0.488	-0.130
		(0.031)	*(0.033)*	*(0.030)*	*(0.036)*
% Alunos c/ Lanche Subsidiado			-0.547		-0.529
			(0.024)		*(0.038)*
% Alunos c/ Assist. por Renda				-0.790	-0.048
				(0.068)	*(0.059)*
Estatísticas de Ajuste					
SER[2]	18.581	14.464	9.080	11.654	9.084
R^2 Ajustado	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

[1]Erros-padrão robustos HC1 entre parênteses.
[2]SER = desvio-padrão dos resíduos (Root MSE).

```
    table.border.bottom.width      = px(2),
    row_group.border.top.style     = "solid",
    row_group.border.top.width     = px(1),
    source_notes.font.size         = 10
  )

# ------------------------------------------------------------------------------
# 5. EXIBIÇÃO / EXPORTAÇÃO
# ------------------------------------------------------------------------------

tab
```

3 Anexo - Prompt

Abaixo segue o prompt solicitado ao Claude para elaboração da tabela em formato publicação. O primeiro código gerado pelo Claude veio com problemas, de modo que a apresentação do resultado estava gerando mais de 5 colunas e diversos valores NA. Após a apresentação inicial, foi necessário fazer ajustes no código.

Prompt:

Claude, eu tenho 5 modelos de regressão estimados em R conforme reportado abaixo. Eu preciso preparar uma tabela de resultados formatada para publicação. Para produzir a tabela, eu gostaria que fosse utilizado o pacote gt em R. A tabela deve conter: 1. Os cinco modelos apresentados em colunas; 2. Os coeficientes devem ser apresentados em duas partes. Na primeira, no topo, o regressor de interesse razão aluno-professor (variável str). Na segunda parte, logo abaixo, serão apresentados os controles: % de alunos aprendendo inglês (el_pct); % alunos com lanche subsídiado (meal_pct); % alunos que qualificam para assistência por critério de renda (calw_pct). 3. A terceira parte da tabela deve apresentar as seguintes estatísticas descritivas: SER, R2-ajustado e n. 4. Por fim, é importante que os erros-padrão sejam reportados entre parenteses abaixo dos coeficientes e seja reportado o desvio-padrão robusto (HC1).

Observação: eu não preciso que você rode os modelos, apenas apresente o código para execução.

Código das especificações:

```
#Especificação 1
espec_1 <- lm(testscr ~ str, data = caschool)

#Especificação 2
espec_2 <- lm(testscr ~ str + el_pct, data = caschool)

#Especificação 3
espec_3 <- lm(testscr ~ str + el_pct + meal_pct, data = caschool)

#Especificação 4
espec_4 <- lm(testscr ~ str + el_pct + calw_pct, data = caschool)

#Especificação 5
espec_5 <- lm(testscr ~ str + el_pct + meal_pct+ calw_pct, data = caschool)
```