Econometria I (2022) — FEA-RP/USP

Professor: Daniel Domingues dos Santos — Monitor: Fábio Hideki Nishida

---

Lista Prática 3

Exercício 1 (Wooldridge, 2006, Exercício 2.1)**.** *Neste exercício, usaremos a base de dados de Papke (1995), que possui informações sobre a participação e contribuição em planos previdência privada de empresas nos EUA, chamada de 401(k):*

```r
data(k401k, package="wooldridge")
```

- `prate`*: é o percentual de trabalhadores contribuindo ativamente à previdência privada.*
- `mrate`*: é a taxa de "generosidade" da empresa, isto é, a razão de quanto a empresa contribui para a previdência privada de seu funcionário sobre o quanto este funcionário contribuiu com seu próprio salário. Por exemplo, se a empresa auxilia com \$0,50 para cada \$1,00 de contribuição do funcionário, então a taxa de generosidade* `mrate` $= 0,50$.

Queremos saber a relação entre a taxa de participação de funcionários (`prate`*) e a a taxa de generosidade da empresa (*`mrate`*).*

a) *Encontre as médias de taxa de participação (*`prate`*) e de taxa de generosidade (*`mrate`*)*

b) *Estime "na mão" (sem usar a função* `lm()`*) o seguinte modelo de regressão simples:*

$$\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 mrate$$

Para isto, use as fórmulas dos estimadores de MQO:

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x} \qquad e \qquad \hat{\beta}_1 = \frac{Cov(x,y)}{Var(x)}$$

Além de reportar as estimativas de $\hat{\beta}_0$ e $\hat{\beta}_1$, também informe o número de observações .

c) *Interprete o intercepto e o coeficiente de* `mrate`*.*

d) *Encontre o valor ajustado/predito de* `prate` *quando* `mrate` $= 3,5$*. É uma previsão razoável? Explique.*

Resposta:

a)

```r
mean(k401k$prate)
mean(k401k$mrate)
```

```
[1] 87.36291
[1] 0.7315124
```

b) Usando equações (2.17) e (2.19) de Wooldridge (2006), temos:

```
b1hat = cov(k401k$prate, k401k$mrate) / var(k401k$mrate)
b0hat = mean(k401k$prate) - mean(k401k$mrate)*b1hat

b1hat
b0hat
nrow(k401k)
```

```
[1] 5.861079
[1] 83.07546
[1] 1534
```

c) Quando $mrate = 0$ (taxa de generosidade é nula), a taxa de participação é, em média, de 83,1%. A cada incremento na contribuição empresarial correspondente a 100% da contribuição do trabalhador, aumenta-se aproximadamente 5,9% a participação no programa de previdência privada.

d) O valor predito de participação $\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1.3,5 = 103,59$ não faz sentido na realidade, pois o percentual de participação não pode ser maior do que 100%. Isso mostra que, em casos em que a variável dependente é limitada, o modelo de MQO pode resultar previsões estranhas em valores extremos.

```
prate_hat_35 = b0hat + b1hat*3.5
prate_hat_35
```

```
[1] 103.5892
```

□

Exercício 2 (Wooldridge, 2019, Exercício C2.5)**.** *Neste exercício, usaremos a base de dados com informações de empresas na indústria química.*

```
data(rdchem, package="wooldridge")
```

- `rd`: *é o gasto anual em pesquisa e desenvolvimento (P&D) em milhões de dólares*
- `sales`: *é a venda anual da empresa em milhões de dólares*

Queremos saber o quanto as vendas (`sales`*) afetam os gastos com P&D.*

a) Escreva um modelo empírico (não é a equação estimada) que implique uma elasticidade constate entre `rd` *e* `sales`*. Qual é o parâmetro da elasticidade?*

b) Estime o modelo proposto usando a base de dados `rdchem`*. Qual é a elasticidade estimada entre* `rd` *e* `sales`*? Explique, em palavras, o que essa elasticidade estimada significa.*

Resposta:

a) Modelo empírico:

$$\log(rd) = \beta_0 + \beta_1 \log(sales) + u,$$

em que a elasticidade é dada por β_1.

b) Ver tabela 2.3 da seção 2.4 de Wooldridge (2006). A elasticidade é de $\hat{\beta}_1 = 1,076$. A cada aumento de 1% nas vendas, há um incremento médio de 1,076% nos gastos com P&D.

```r
lm(log(rdchem$rd) ~ log(rdchem$sales))
```

```
Call:
lm(formula = log(rdchem$rd) ~ log(rdchem$sales))

Coefficients:
      (Intercept)  log(rdchem$sales)
           -4.105              1.076
```

□

Exercício 3. *Neste exercício, usaremos as funções* `runif()` *e* `rnorm()` *para gerar números aleatórios com distribuições uniforme e normal, respectivamente.*

a) *Gere o vetor* `x` *com 10.000 números aleatórios a partir de uma distribuição uniforme no intervalo [0, 10]. Qual é a média e o desvio padrão de x?*

b) *Gere o vetor* `z` *com 10.000 números aleatórios usando $z = 2x + \tilde{u}, \quad \tilde{u} \sim N(0, 4^2)$. Qual é a média e o desvio padrão de z? Qual é a correlação entre x e z?*

c) *Gere o vetor $\tilde{\varepsilon}$ (*`e_til`*) com 10.000 números aleatórios a partir de uma distribuição $N(0, 6^2)$. Qual é a média e o desvio padrão de $\tilde{\varepsilon}$? Qual é a correlação entre $\tilde{u}$ e cada uma das demais variáveis x e z? Além disso, verifique a correlação entre x e a soma $(\tilde{\varepsilon} + z)$.*

d) *Gere o vetor* `y`*, considerando o seguinte modelo real:*

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 z + \tilde{\varepsilon}, \tag{2.1}$$

em que $\tilde{\beta}_0 = 10$, $\tilde{\beta}_1 = 2$ e $\tilde{\beta}_2 = 3$. Agora, estime por MQO o seguinte modelo empírico:

$$y = \beta_0 + \beta_1 x + \varepsilon. \tag{2.2}$$

A estimação conseguiu recuperar $\hat{\beta}_0 \approx \tilde{\beta}_0$ e $\hat{\beta}_1 \approx \tilde{\beta}_1$? Explique.

e) *Obtenha os resíduos de MQO, $\hat{\varepsilon}$, e verifique se valem as seguintes condições amostrais (sujeitas a algum erro de arredondamento):*

$$\frac{1}{N}\sum_{i=1}^{N} \hat{\varepsilon}_i = 0 \qquad e \qquad \frac{1}{N}\sum_{i=1}^{N} x_i \hat{\varepsilon}_i = 0$$

f) *O que os resultados do item (e) dizem sobre as condições de momento populacionais $E(\varepsilon) = 0$ e $E(x\varepsilon) = 0$?*

g) *Denote como <u>Caso I</u> o modelo real visto até agora com $\tilde{\beta}_2 = 3$ e $z = 2x + \tilde{u}$, em que $\tilde{u} \sim N(0, 4^2)$. Agora, gere novamente observações z e y, e estime por OLS o modelo empírico (2.2) para cada um dos seguintes modelos reais:*

- *<u>Caso II</u>:* $\tilde{\beta}_2 = -3$ *e* $z = 2x + \tilde{u}, \quad \tilde{u} \sim N(0, 4^2)$
- *<u>Caso III</u>:* $\tilde{\beta}_2 = 3$ *e* $z = -2x + \tilde{u}, \quad \tilde{u} \sim N(0, 4^2)$
- *<u>Caso IV</u>:* $\tilde{\beta}_2 = -3$ *e* $z = -2x + \tilde{u}, \quad \tilde{u} \sim N(0, 4^2)$

Considerando os sinais do parâmetro da variável omitida z, $\tilde{\beta}_2$, e da sua covariância com x, $cov(x, z)$, em quais casos a estimativa do parâmetro de x é sobre-estimada ($\hat{\beta}_1 > \tilde{\beta}_1$)? E em quais é sub-estimada ($\hat{\beta}_1 < \tilde{\beta}_1$)?

<u>Resposta</u>:

a) $\bar{x} \approx 5$ e $dp(x) \approx 2,9$

```
N = 10000
x = runif(N, 0, 10)
mean(x)
sd(x)
```

```
[1] 5.048937
[1] 2.886117
```

b) $\bar{z} \approx 10,1$, $dp(z) \approx 7$ e $corr(x, z) \approx 0,82$

```
z = 2*x + rnorm(N, 0, 4)
mean(z)
sd(z)
cor(x, z)
```

```
[1] 10.12043
[1] 6.990625
[1] 0.819311
```

c) $\bar{\tilde{\varepsilon}} \approx 0$, $dp(\tilde{\varepsilon}) \approx 6$, $corr(x, \tilde{\varepsilon}) \approx 0$, $corr(z, \tilde{\varepsilon}) \approx 0$ e $corr(x, \tilde{\varepsilon} + z) \approx 0,63$

```
e_til = rnorm(N, 0, 6)
mean(e_til)
sd(e_til)
cor(x, e_til)
cor(z, e_til)
cor(x, e_til + z)
```

```
[1] 0.04037083
[1] 5.95413
[1] 0.008887763
[1] 0.009301469
[1] 0.6326533
```

d) $\bar{y} \approx 50$ e $dp(y) \approx 26$

```
y = 10 + 2*x + 3*z  + e_til
mean(y)
sd(y)
```

```
[1] 50.52809
[1] 26.62365
```

Agora, estimando o modelo empírico (1.2), temos

```
fit = lm(y ~ x)
fit
```

```
(Intercept)            x
     10.430        7.942
```

Como $\hat{\beta}_0 \approx 10,4 \approx 10 = \tilde{\beta}_0$ e $\hat{\beta}_1 \approx 7,9 > 2 = \tilde{\beta}_1$, a estimação conseguiu recuperar apenas $\tilde{\beta}_0$ do modelo real, enquanto $\hat{\beta}_1$ é viesado (sobre-estimado) e não recuperou $\tilde{\beta}_1$.

Isto se dá pelo viés de variável omitida. Como o modelo empírico (2.2) não incluiu z como covariada, então ela entrou dentro de $\varepsilon = \tilde{\varepsilon} + z$ e, portanto, não é válida a hipótese $E(x\varepsilon) = 0$, o que compromete as estimativas da regressão por MQO. De fato, vimos no item (b) que $corr(x, \tilde{\varepsilon} + z) \approx 0,63 \neq 0$.

e) $\sum_{i=1}^{n} \hat{\varepsilon}_i \approx 0, \quad \sum_{i=1}^{n} x_i \hat{\varepsilon}_i \approx 0$

```
sum(fit$residual)
sum(fit$residual * x)
```

```
[1] 2.039452e-12
[1] 5.794154e-11
```

f) O que se pode dizer é que o estimador de MQO faz com que $\sum_{i=1}^{n} \hat{\varepsilon}_i$ e $\sum_{i=1}^{n} x_i \hat{\varepsilon}_i$ sejam sempre iguais a zero. Isto ocorre, pois o MQO escolhe $\hat{\beta}_0$ e $\hat{\beta}_1$ que resultem em resíduos $\hat{\varepsilon}$ $(= y - \hat{y})$ que satisfaçam essas contrapartidas amostrais. No entanto, isto não quer dizer que as hipóteses $E(\varepsilon) = 0$ e $E(x\varepsilon) = 0$ sejam verdadeiras. Se $E(x\varepsilon) \neq 0$, então a estimação de $\hat{\beta}_1$ será viesada.

g) II $\hat{\beta}_1 < \tilde{\beta}_1 = 2$, quando $\tilde{\beta}_2 < 0$ e $cov(x, z) > 0$

```
z = rnorm(10000, 2*x, 4)
y = 10 + 2*x - 3*z + u
cor(x, z)
lm(y ~ x)
```

```
[1] 0.8245678

(Intercept)            x
     10.338       -4.067
```

III $\hat{\beta}_1 < \tilde{\beta}_1 = 2$, quando $\tilde{\beta}_2 > 0$ e $cov(x, z) < 0$

```
z = rnorm(10000, -2*x, 4)
y = 10 + 2*x + 3*z + e_til
cor(x, z)
lm(y ~ x)
```

```
[1] -0.8263252

(Intercept)            x
      10.14        -4.04
```

IV $\hat{\beta}_1 > \tilde{\beta}_1 = 2$, quando $\tilde{\beta}_2 < 0$ e $cov(x, z) < 0$

```
z = rnorm(10000, -2*x, 4)
y = 10 + 2*x - 3*z + e_til
cor(x, z)
lm(y ~ x)
```

```
[1] -0.819522

(Intercept)            x
     10.166        7.997
```

Portanto, assumindo os sinais de $\tilde{\beta}_2$ e $cov(x, z)$, conseguimos ao menos analisar que, em relação ao valor verdadeiro $\tilde{\beta}_1$, a estimativa

- $\hat{\beta}_1$ é sobre-estimada ($\hat{\beta}_1 > \tilde{\beta}_1$), se $\tilde{\beta}_2.cov(x, z) > 0$, e
- $\hat{\beta}_1$ é sub-estimada ($\hat{\beta}_1 < \tilde{\beta}_1$), se $\tilde{\beta}_2.cov(x, z) < 0$.

$\square$