

### 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 *401k*:

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

- *prate*: é o percentual de trabalhadores contribuindo ativamente à previdência privada.
- *mrte*: 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  $mrte = 0,50$ .

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

- Encontre as médias de taxa de participação (*prate*) e de taxa de generosidade (*mrte*)
- Estime “na mão” (sem usar a função `lm()`) o seguinte modelo de regressão simples:

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

Para isto, use as fórmulas dos estimadores de MQO:

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

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

- Interprete o intercepto e o coeficiente de *mrte*.
- Encontre o valor ajustado/predito de *prate* quando  $mrte = 3,5$ . É uma previsão razoável? Explique.

Resposta:

```
a) mean(k401k$prate)
2 mean(k401k$mrte)
```

```
1 [1] 87.36291
2 [1] 0.7315124
```

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

```
1 b1hat = cov(k401k$prate, k401k$mrate) / var(k401k$mrate)
2 b0hat = mean(k401k$prate) - mean(k401k$mrate)*b1hat
3
4 b1hat
5 b0hat
6 nrow(k401k)

1 [1] 5.861079
2 [1] 83.07546
3 [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  $\widehat{prate} = \hat{\beta}_0 + \hat{\beta}_1 \cdot 3,5 = 103,59$  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.

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

1 [1] 103.5892
```

□

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

```
1 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.*

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

- Modelo empírico:

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

em que a elasticidade é dada por  $\beta_1$ .

- 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.

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

1 Call:
2 lm(formula = log(rdchem$rd) ~ log(rdchem$sales))
3
4 Coefficients:
5      (Intercept)      log(rdchem$sales)
6      -4.105              1.076
```

□

**Exercício 3** (Wooldridge, 2019, Exercício C2.8 - Modificado). *Neste exercício, usaremos as funções `runif()` e `rnorm()` para gerar números aleatórios com distribuições uniforme e normal, respectivamente.*

- Gere o vetor  $\mathbf{x}$  com 10.000 números aleatórios a partir de uma distribuição uniforme no intervalo  $[0, 10]$ . Qual é a média e o desvio padrão amostrais de  $x$ ?
- Gere o vetor  $\mathbf{z}$  com 10.000 números aleatórios a partir de uma distribuição  $N(2x, 4^2)$ . Qual é a média e o desvio padrão amostrais de  $z$ ? Qual é a correlação entre  $x$  e  $z$ ?
- Gere o vetor  $\tilde{u}$  (`u_til`) com 10.000 números aleatórios a partir de uma distribuição  $N(0, 6^2)$ . Qual é a média e o desvio padrão amostrais de  $\tilde{u}$ ? 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 de variáveis  $\tilde{u} + z$ .
- Gere o vetor  $\mathbf{y}$ , considerando o seguinte modelo real:

$$y = \tilde{\beta}_0 + \tilde{\beta}_1 x + \tilde{\beta}_2 z + \tilde{u}, \quad (3.1)$$

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

$$y = \beta_0 + \beta_1 x + u. \quad (3.2)$$

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

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

$$\sum_{i=1}^n \hat{u}_i = 0 \quad e \quad \sum_{i=1}^n x_i \hat{u}_i = 0 \quad (\text{Wooldridge, 2006, 2.60})$$

- O que os resultados do item (e) dizem sobre as hipóteses  $E(u) = 0$  e  $E(xu) = 0$ ?
- Considere que o modelo real visto até agora é o Caso I: em que  $\tilde{\beta}_2 = 3$  e  $z \sim N(2x, 4^2)$ . Agora, gere novamente  $z$ , calcule  $y$  e estime o modelo empírico (3.2) para cada um dos seguintes casos:

- Caso II:  $\tilde{\beta}_2 = -3$  e  $z \sim N(2x, 4^2)$
- Caso III:  $\tilde{\beta}_2 = 3$  e  $z \sim N(-2x, 4^2)$
- Caso IV:  $\tilde{\beta}_2 = -3$  e  $z \sim N(-2x, 4^2)$

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 ( $\hat{\beta}_1 < \tilde{\beta}_1$ )?

Resposta:

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

```
1 x = runif(10000, 0, 10)
2 mean(x)
3 sd(x)
```

```
1 [1] 5.048937
2 [1] 2.886117
```

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

```
1 z = rnorm(10000, 2*x, 4)
2 mean(z)
3 sd(z)
4 cor(x, z)
```

```
1 [1] 10.12043
2 [1] 6.990625
3 [1] 0.819311
```

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

```
1 u_til = rnorm(10000, 0, 6)
2 mean(u_til)
3 sd(u_til)
4 cor(x, u_til)
5 cor(z, u_til)
6 cor(x, u_til + z)
```

```
1 [1] 0.04037083
2 [1] 5.95413
3 [1] 0.008887763
4 [1] 0.009301469
5 [1] 0.6326533
```

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

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

```
1 [1] 50.52809
2 [1] 26.62365
```

Agora, estimando o modelo empírico (3.2), temos

```

1 fit = lm(y ~ x)
2 fit

1 Call:
2 lm(formula = y ~ x)
3
4 Coefficients:
5 (Intercept)          x
6      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 (3.2) não incluiu  $z$  como covariada, então ela entrou dentro de  $u = \tilde{u} + z$  e, portanto, não é válida a hipótese  $E(xu) = 0$ , o que compromete as estimativas da regressão por MQO. De fato, vimos no item (b) que  $\text{corr}(x, \tilde{u} + z) \approx 0,63 \neq 0$ .

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

```

1 sum(fit$residual)
2 sum(fit$residual * x)

1 [1] 2.039452e-12
2 [1] 5.794154e-11

```

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

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

```

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

1 [1] 0.8245678
2
3 Coefficients:
4 (Intercept)          x
5      10.338        -4.067

```

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

```

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

1 [1] -0.8263252
2
3 Coefficients:
4 (Intercept)          x
5      10.14         -4.04

```

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

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

```
1 [1] -0.819522
2
3 Coefficients:
4 (Intercept)          x
5      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  $\tilde{\beta}_2 \cdot cov(x, z) > 0$ , e
- $\hat{\beta}_1$  é sub-estimada ( $\hat{\beta}_1 < \tilde{\beta}_1$ ), se  $\tilde{\beta}_2 \cdot cov(x, z) < 0$ .

□