# Lista 1

Diogo Wolff Surdi
Econometria I
FGV

16 de Março de 2020

## Questão 1

$$
\begin{aligned}
Cov(X,Y) &= \mathbb{E}[(X-\mathbb{E}[X])(Y-\mathbb{E}[Y])] \\
&= \mathbb{E}[XY - X\mathbb{E}[Y] - \mathbb{E}[X]Y + \mathbb{E}[X]\mathbb{E}[Y]] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] - \mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\
&= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
\end{aligned}
$$

Se $\mathbb{E}[X] = 0$ ou $\mathbb{E}[Y] = 0$, então o segundo termo se anula, logo $Cov(X,Y) = \mathbb{E}[XY]$.

## Questão 2

$$
\begin{aligned}
\mathbb{E}[X] &= \int_x x f_x dx \\
&= \int_x x \left( \int_y f_{X,Y=y} dy \right) dx \\
&= \int_x x \left( \int_y f_{X|Y=y} f_{Y=y} dy \right) dx \\
&= \int_x \int_y x f_{X|Y=y} f_{Y=y} dy dx \\
&= \int_y \int_x x f_{X|Y=y} f_{Y=y} dx dy \\
&= \int_y \left( \underbrace{\int_x x f_{X|Y=y} dx}_{=\mathbb{E}[X|Y]} \right) f_{Y=y} dy \\
&= \mathbb{E}[\mathbb{E}[X|Y]]
\end{aligned}
$$

# Questão 3

As propriedades do MQO são linearidade nos parâmetros, aleatoriedade da amostra, esperança condicional do erro nula ($\mathbb{E}[u|x_1, x_2, ..., x_n] = 0$), e não-existência de colinearidade perfeita.

## (a)

Sabemos que $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x}$ e $\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{\sum_{i=1}^n (x_i - \bar{x})^2}$. Começando pelo segundo item, podemos realizar as tranformações:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})y_i}{SQT_x} = \frac{\sum_{i=1}^n (x_i - \bar{x})(\beta_0 + \beta_1 x_i + u_i)}{SQT_x}$$

O numerador pode ser transformado em:

$$\begin{aligned}
&\sum_{i=1}^n (x_i - \bar{x})\beta_0 + \sum_{i=1}^n (x_i - \bar{x})\beta_1 x_i + \sum_{i=1}^n (x_i - \bar{x})u_i \\
&= \beta_0 \sum_{i=1}^n (x_i - \bar{x}) + \beta_1 \sum_{i=1}^n (x_i - \bar{x})x_i + \sum_{i=1}^n (x_i - \bar{x})u_i
\end{aligned}$$

Note que $\sum_{i=1}^n (x_i - \bar{x}) = 0$ e $\sum_{i=1}^n (x_i - \bar{x})x_i = \sum_{i=1}^n (x_i - \bar{x})^2 = SQT_x$, logo podemos reescrever a equação como:

$$\hat{\beta}_1 = \beta_1 SQT_x + \sum_{i=1}^n (x_i - \bar{x})u_i = \beta_1 + \frac{1}{SQT_x}\sum_{i=1}^n (x_i - \bar{x})u_i$$

Com isso, basta tomar a esperança condicional em relação à amostra para encontrar o resultado:

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1|x_1, ..., x_n] &= \beta_1 + \mathbb{E}[\frac{1}{SQT_x}\sum_{i=1}^n (x_i - \bar{x})u_i|x_1, ..., x_n] \\
&= \beta_1 + \frac{1}{SQT_x}\mathbb{E}[\sum_{i=1}^n (x_i - \bar{x})u_i|x_1, ..., x_n] \\
&= \beta_1 + \frac{1}{SQT_x}\sum_{i=1}^n (x_i - \bar{x})\mathbb{E}[u_i|x_1, ..., x_n] \\
&= \beta_1 + \frac{1}{SQT_x}\sum_{i=1}^n (x_i - \bar{x}) \times 0 \\
&= \beta_1
\end{aligned}$$

Sabendo que vale para $\hat{\beta}_1$, a prova para $\hat{\beta}_0$ é trivial. Temos que $\bar{y} = \beta_0 + \beta_1\bar{x} + \bar{u}$, logo $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1\bar{x} = \beta_0 + (\beta_1 - \hat{\beta}_1)\bar{x} + \bar{u}$. Com isso:

$$\begin{aligned}\mathbb{E}[\hat{\beta}_0|x_1,...,x_n] &= \beta_0 + \mathbb{E}[(\beta_1 - \hat{\beta}_1)\bar{x}|x_1,...,x_n)] + \mathbb{E}[\bar{u}|x_1,...,x_n] \\ &= \beta_0 + \bar{x}\mathbb{E}[(\beta_1 - \hat{\beta}_1)|x_1,...,x_n] \\ &= \beta_0\end{aligned}$$

# Questão 4

Seja $\mathbb{E}[u] = \alpha_0$. Então, basta-se tomar $y = (\beta_0+\alpha_0)+\beta_1 x+u-\alpha_0$, gerando novos $\beta_0^* = \beta_0+\alpha_0$ e $u^* = u - \alpha_0$. Note que $\mathbb{E}[u^*] = 0$, logo encontramos o resultado esperado.

# Questão 5

## (a)

O peso com 0 cigarros é 199.77, enquanto que o peso para 20 cigarros é 189.49. Essa difirença indica que mães que fumam cigarros tendem a ter filhos com peso menor ao nascer.

## (b)

Apesar da correlação ser não-nula, podem existir outros fatores correlacionados com a queda no peso dos bebês. O caso pode ser de que mães que fumam durante a gravidez desconhecem os riscos que isso gera, e assim também podem incorrer em outros comportamentos maléficos para a saúde do bebê, sendo que a perda de peso pode ser causada por esses efeitos ulteriores.

## (c)

$$125 = 199.77 - 0.514x \quad \longrightarrow \quad x \approx 145$$

## (d)

Como há muitas mulheres não-fumantes no grupo, há grande quantidade de variação que não é resultante de fumar, mas que afeta a regressão pois está sendo contabilizada. Com isso, há um forte viés de omissão presente na regressão, explicando o valor exorbitante encontrado no item anterior.

# Questão 6

As hipóteses são versões n-dimensoniais das propriedades do MQO, somadas à hipótese de homoscedasticidade. O teorema é importante pois afirma que o método dos mínimos quadrados gera o estimador não-enviesado de menor variância.

# Questão 7

## (a)

O sinal de $\beta_0$ deve ser positivo, pois casas tem valor positivo logo a média deve o ser para qualquer valor das outras variáveis. Como a poluição diminui a qualidade de vida das pessoas, espera-se que $\beta_1$ possua valor negativo.

## (b)

É razoável supor que há menos poluição em bairros nobres, pois pessoas ricas evitariam áreas com essa característica, e moradores de tais localidades possuem maior renda, podendo construir casas com mais quartos. Com isso, tais fatores são negativamente correlacionados. Considerando agora a modelagem sem a segunda variável independente, temos que observar o efeito da omissão dela na estimação.

Sabemos que $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)y_i}{\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2}$. Sendo o modelo verdadeiro $y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + u_i$, ao substituir no denominador de $\hat{\beta}$ encontramos:

$$\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)y_i = \beta_1 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2 + \beta_2 \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2} + \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)u_i$$

Voltando para $\beta_1$, isto é, dividindo a equação por $\sum_{i=1}^{n}(x_{i1}-\bar{x}_1)^2$, e tirando o valor esperado, encontramos que:

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2}}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2}$$

Pois $\mathbb{E}[u_i] = 0$. Note que, $\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2} = \sum_{i=1}^{n}(x_{i1} - \bar{x}_1)(x_{i2} - \bar{x}_2) = (n-1)Cov(x_1, x_2)$, logo o sinal da correlação entre os dois é o mesmo do termo da equação. Como visto anteriormente, tal valor será então negativo. Como casas com mais quartos tendem a ser mais caras, podemos supor que $\beta_2>0$. Com isso, o valor esperado será:

$$\mathbb{E}[\hat{\beta}_1] = \beta_1 + \beta_2 \frac{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)x_{i2}}{\sum_{i=1}^{n}(x_{i1} - \bar{x}_1)^2} < \beta_1$$

Com isso, podemos afirmar que a regressão *subestima* o valor de $\beta_1$ (isso é, a regressão sobrevaloriza o impacto da poluição).

## (c)

Sim, a relação é como esperado. É válido afirmar que $-0.781$ está mais próximo do que $-1.0043$ da elasticidade verdadeira, pois há mais variáveis explicativas não-redundantes, logo a estimação deve estar mais próxima do modelo real.

# Questão 8

## (a)

```
library(foreign)
dados <- read.dta("C:/Users/dwsur/OneDrive/Area de Trabalho/
docs/livros do curso/5 periodo/econometria 1/lista 1/
econometria_2020-master/ceosal2.dta")
mean(dados[,'salary'])
mean(dados[,'ceoten'])
```

Os valores encontrados foram salário médio de 865.9 e permanência média de 7.955.

## (b)

Gero o vetor de dados da coluna que se encaixam na especificação, e encontro o tamanho dele. No caso, é 5.

```
ano0 <- which(dados$ceoten==0)
length(ano0)
```

Para encontrar a maior permanência é mais simples. No caso, é 37.

```
max(dados$ceoten)
```

## (c)

A função log(*salary*) já está computada em uma coluna dos dados.

```
modelo <- lm(lsalary ~ ceoten, dados)
modelo
```

```
Call:
lm(formula = lsalary ~ ceoten, data = dados)

Coefficients:
(Intercept)       ceoten
6.505498     0.009724
```

Para cada ano de permanência no cargo, o salário aumenta cerca de 0.97%.

# Questão 9

## (a)

```
regressao <- lm(price ~
        sqrft + bdrms,
        dados)
```

```
regressao

Call:
lm(formula = price ~ sqrft + bdrms, data = dados)

Coefficients:
(Intercept)        sqrft        bdrms
   -19.3150       0.1284      15.1982
```

A função é então $price = -19.3150 + 0.1284 sqrft + 15.1982 bdrms + u$.

## (b)

O aumento estimado com um quarto a mais no preço é de 15.19 milhares de dólares.

## (c)

O aumento estimado seria de $15.19 + 140 \times 0.1284 \approx 33.17$ milhares de dólares.

## (d)

```
summary(regressao)

Call:
lm(formula = price ~ sqrft + bdrms, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-127.627  -42.876   -7.051   32.589  229.003

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept) -19.31500   31.04662  -0.622    0.536
sqrft         0.12844    0.01382   9.291 1.39e-14 ***
bdrms        15.19819    9.48352   1.603    0.113
---

Residual standard error: 63.04 on 85 degrees of freedom
Multiple R-squared:  0.6319,    Adjusted R-squared:  0.6233
F-statistic: 72.96 on 2 and 85 DF,  p-value: < 2.2e-16
```

O $R^2$ é 0.6319, logo as variáveis explicam cerca de 63.2% da variação.

## (e)

$$-19.3150 + 0.1284 \times 2438 + 15.1982 \times 4 = 354.517$$

## (f)

O resíduo é $\hat{u} = y - \hat{y} = 300 - 354.517 = -54.517$, logo, o valor sugere que ele pagou pouco pela casa.