# Gabarito - Lab Aula 7

```
library(tidyverse)
library(olsrr)
library(lmtest)
library(stargazer)

dados <- read.csv("Aula 7 - Lab 4 - Base moodle.csv")
```

**Uma das hipóteses do modelo MQO é a de que a relação entre as variáveis é linear. Isto acarreta um obstáculo: e se a relação entre x e y existir, mas não for linear?**

**Usando o arquivo que está disponibilizado no Moodle para esta aula ("Base_Lista_Aula_7"), faça o seguinte:**

**Considere que esta base indica duas variáveis dependentes (Y1 e Y2), duas explicativas (X1 e X2) e dois controles (C1 e C2).**

**Comecemos com a variável dependente Y1.**

**1. Faça como na semana passada: Rode cinco modelos introduzindo uma variável dependente por vez – primeiro x1, depois x2 - e depois as dependentes juntas x1 e x2, depois x1, x2 e c1 e c2. O que acontece com os coeficientes destas variáveis entre os modelos? Discuta;**

```
reg_x1 <- dados %>% lm(y1 ~ x1, data=.)
reg_x2 <- dados %>% lm(y1 ~ x2, data=.)
reg_x1x2 <- dados %>% lm(y1 ~ x1 + x2, data=.)
reg_completa <- dados %>% lm(y1 ~ x1 + x2 + c1 + c2, data=.)

stargazer(reg_x1, reg_x2, reg_x1x2, reg_completa,
          style="ajps",
          column.labels=c("x1", "x2", "x1 + x2", "x1 + x2 + c1 + c2"),
          omit.stat=c("f"),
          header=FALSE)
```

*Resultado das regressões está na Tabela 1.*

*Os coeficientes dos modelos ficam estáveis quanto ao tamanho do efeito e significância estatística. A maior mudança que ocorre com a inclusão das variáveis de controle c1 e c2 está no coeficiente em x2, que sobe de -0,11 para -0,06, e acaba perdendo a significância estatística com um p-valor superior a 0,1.*

**2. Realize os testes de multicolinearidade e de heteroscedasticidade. Discuta os resultados e os corrija se for o caso;**

```
bptest(reg_completa)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_completa
## BP = 123.79, df = 4, p-value < 2.2e-16
```

Tabela 1:

| | y1 | | | |
|---|---|---|---|---|
| | **x1** | **x2** | **x1 + x2** | **x1 + x2 + c1 + c2** |
| | **Model 1** | **Model 2** | **Model 3** | **Model 4** |
| x1 | 6.635*** | | 6.635*** | 6.687*** |
| | (0.171) | | (0.170) | (0.170) |
| x2 | | −0.110 | −0.110* | −0.061 |
| | | (0.085) | (0.064) | (0.067) |
| c1 | | | | −1.854*** |
| | | | | (0.262) |
| c2 | | | | 0.012 |
| | | | | (0.092) |
| Constant | 18.048*** | 32.197*** | 18.926*** | 24.041*** |
| | (0.426) | (0.758) | (0.666) | (1.018) |
| N | 2000 | 2000 | 2000 | 2000 |
| R-squared | 0.431 | 0.001 | 0.432 | 0.447 |
| Adj. R-squared | 0.431 | 0.0003 | 0.431 | 0.445 |
| Residual Std. Error | 11.437 (df = 1998) | 15.157 (df = 1998) | 11.432 (df = 1997) | 11.289 (df = 1995) |

***p < .01; **p < .05; *p < .1

```
ols_vif_tol(reg_completa)
```

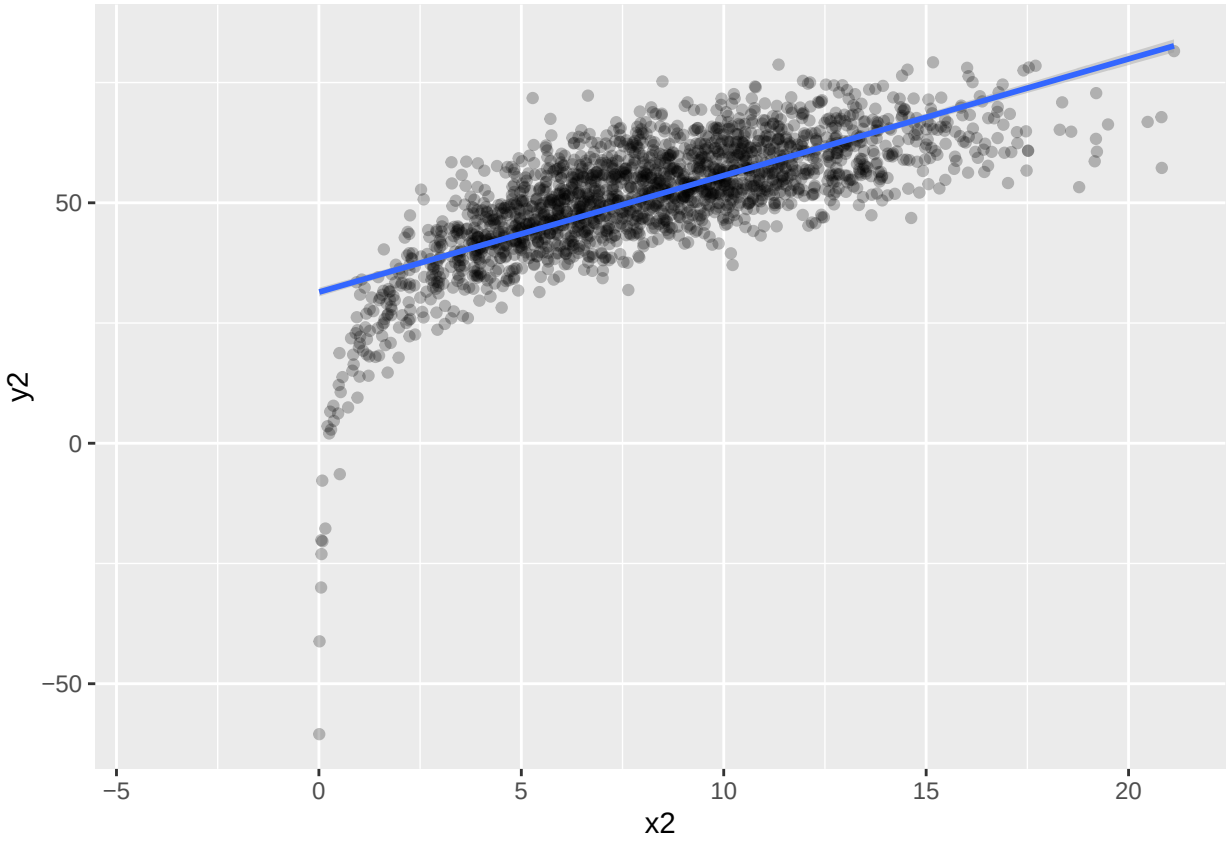
```
##   Variables Tolerance      VIF
## 1        x1 0.9844387 1.015807
## 2        x2 0.8817110 1.134158
## 3        c1 0.9275600 1.078097
## 4        c2 0.8446014 1.183990
```

*O resultado da função* `bptest` *indica que há heterocedasticidade no modelo, ou seja, os resíduos não variam da mesma maneira ao longo da amostra dos dados. Levantando um alerta importante sobre as premissas do modelo linear. Já o teste para multicolinearidade VIF, não traz evidências de correlação serial entre as variáveis explicativas do modelo.*

**A pergunta que você deve se fazer é: este modelo é bom o suficiente? Como aqui não fizemos nenhuma análise descritiva das variáveis, temos apenas o resultado do modelo de regressão para julgar a adequabilidade do próprio modelo.**

**3. Construa um gráfico de dispersão entre a variável X2 e Y2. Analise este gráfico em termos da linearidade da relação;**

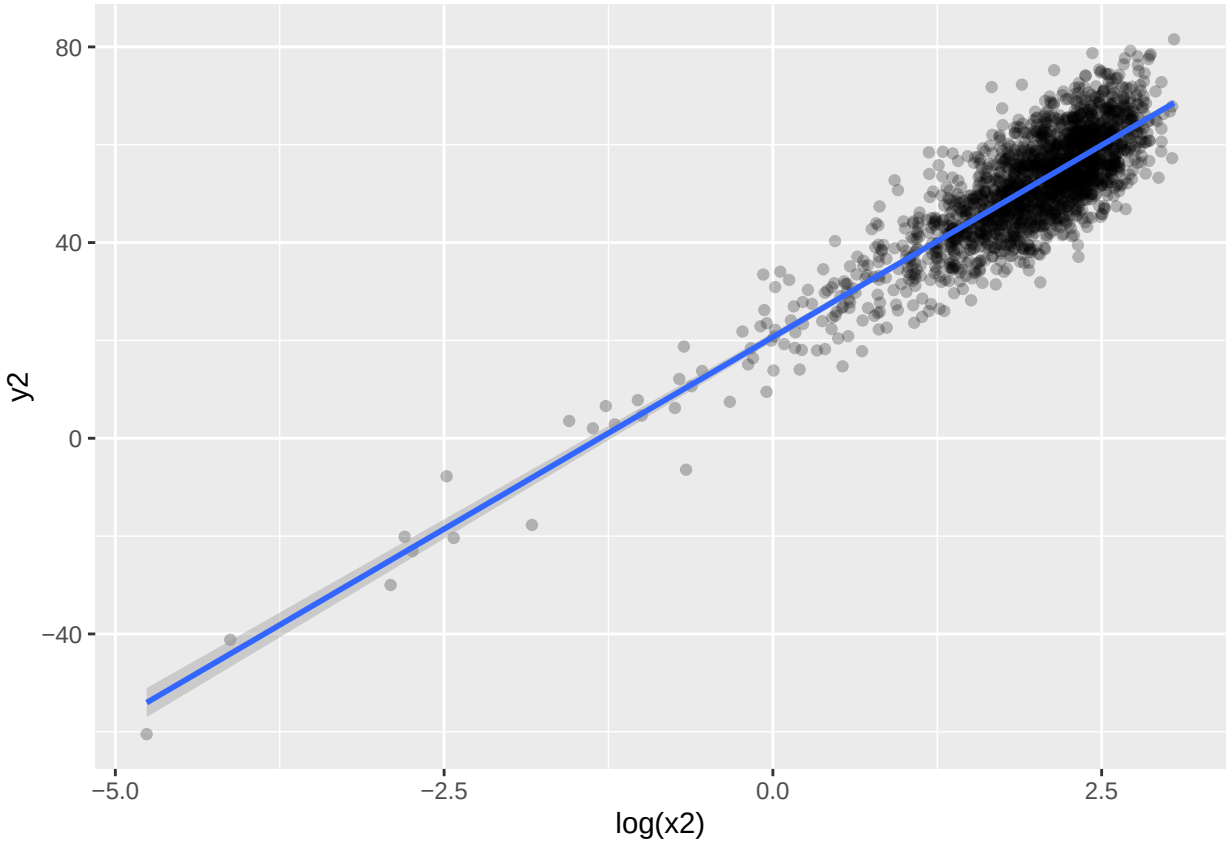
```
ggplot(dados, aes(x2, y2))+
  geom_point(alpha=0.25)+
  stat_smooth(method="lm")
```




*A relação entre as variáveis não é linear, aparenta ser uma distribuição logaritmica.*

**4. Há alguma transformação de variável que possa ser realizada para adequar melhor o modelo? Explique;**

```
ggplot(dados, aes(log(x2), y2))+
  geom_point(alpha=0.25)+
  stat_smooth(method="lm")
```




*Podemos transformar a variável, podemos calcular o logaritmo natural de x2.*

**a. Caso sua resposta seja "sim", faça a transformação da variável e reporte os resultados, comparando-o com o modelo sem transformação;**

```
reg_completa_log <- dados %>% lm(y1 ~ x1 + log(x2) + c1 + c2, data=.)
```

```
## Warning in log(x2): NaNs produzidos
```

```
stargazer(reg_completa,
          reg_completa_log,
          style="ajps",
          column.labels=c("Sem transformação", "log(x2)"),
          omit.stat=c("f"),
          header=FALSE)
```

*Resultados na tabela 2.*

**b. Interprete os parâmetros estimados. O que a transformação muda em sua análise inicial sem a transformação?;**

*Os coeficientes em x1, c1 e c2 permanecem estáveis e com o mesmo nível de significância estatística, apresentando uma pequena melhora no $R^2$ ajustado do modelo, que passa de 0,445 para 0,450.*

**c. Não se esqueça de avaliar a multicolinearidade e a heteroscedasticidade.**

Tabela 2:

| | y1 | |
|---|---|---|
| | **Sem transformação** | **log(x2)** |
| | **Model 1** | **Model 2** |
| x1 | 6.687*** | 6.719*** |
| | (0.170) | (0.171) |
| x2 | −0.061 | |
| | (0.067) | |
| log(x2) | | −0.152 |
| | | (0.396) |
| c1 | −1.854*** | −1.925*** |
| | (0.262) | (0.264) |
| c2 | 0.012 | 0.070 |
| | (0.092) | (0.091) |
| Constant | 24.041*** | 24.272*** |
| | (1.018) | (1.138) |
| N | 2000 | 1954 |
| R-squared | 0.447 | 0.451 |
| Adj. R-squared | 0.445 | 0.450 |
| Residual Std. Error | 11.289 (df = 1995) | 11.237 (df = 1949) |

***p < .01; **p < .05; *p < .1

```
bptest(reg_completa_log)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_completa_log
## BP = 123.95, df = 4, p-value < 2.2e-16
```

```
ols_vif_tol(reg_completa_log)
```

```
##   Variables Tolerance      VIF
## 1        x1 0.9827860 1.017516
## 2   log(x2) 0.9136637 1.094495
## 3        c1 0.9324063 1.072494
## 4        c2 0.8745796 1.143406
```

*Testando para heterocedasticidade, também encontramos um p-valor próximo a 0 que, portanto, nos faz rejeitar a hipótese nula do modelo em que a variância dos resíduos está distribuída de maneira igual. Já o teste VIF indica que não há multicolinearidade entre as variáveis independentes, com os valores próximos a 1.*

**Repita o exercício acima, agora utilizando a variável Y2 como variável resposta em seu modelo.**

**5. Se for necessária alguma transformação, neste caso, de que forma ela influenciará a sua interpretação dos betas? Explique.**

```
reg_completa_log_y2 <- dados %>% lm(y2 ~ x1 + log(x2) + c1 + c2, data=.)
```

```
## Warning in log(x2): NaNs produzidos
```

```
stargazer(reg_completa_log,
          reg_completa_log_y2,
          style="ajps",
          column.labels=c("log(X2)", "log(x2)"),
          omit.stat=c("f"),
          header=FALSE)
```

Tabela 3:

| | **y1** | **y2** |
|---|---|---|
| | **log(X2)** | **log(x2)** |
| | **Model 1** | **Model 2** |
| x1 | 6.719*** | −0.062 |
| | (0.171) | (0.082) |
| log(x2) | −0.152 | 15.095*** |
| | (0.396) | (0.191) |
| c1 | −1.925*** | 3.942*** |
| | (0.264) | (0.127) |
| c2 | 0.070 | −0.026 |
| | (0.091) | (0.044) |
| Constant | 24.272*** | 10.029*** |
| | (1.138) | (0.547) |
| N | 1954 | 1954 |
| R-squared | 0.451 | 0.812 |
| Adj. R-squared | 0.450 | 0.812 |
| Residual Std. Error (df = 1949) | 11.237 | 5.404 |

***p < .01; **p < .05; *p < .1

```
bptest(reg_completa_log_y2)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg_completa_log_y2
## BP = 16.431, df = 4, p-value = 0.002493
```

```
ols_vif_tol(reg_completa_log_y2)
```

```
##   Variables Tolerance      VIF
## 1        x1 0.9827860 1.017516
## 2   log(x2) 0.9136637 1.094495
## 3        c1 0.9324063 1.072494
## 4        c2 0.8745796 1.143406
```

*O principal ponto de atenção quando transformamos alguma variável para um modelo é a interpretação deste coeficiente. Com o logaritmo natural de x2, agora não estamos mais pensando em um aumento de 1 unidade em x2 gerando um efeito $\beta$ em y2. Agora estamos lidando com escalas logaritmicas.*