Gabarito Aula 6 - Lab 3

Suponha que uma pesquisadora tenha encontrado os seguintes resultados para o seu modelo de regressão conforme a tabela a seguir:

	(1)	(2)
Variáveis	y	y
Explicativa_1	1.564***	1.354
	(0.346)	(1.227)
Explicativa_2	0.906**	1.045
	(0.416)	(0.279)
Explicativa_3		0.000722
		(0.0143)
Controle_1		0.843
		(0.627)
Controle_2		0.408
		(0.325)
Constante	-1.975***	-3.119
	(0.395)	(0.382)
Observações	616	616
R2 ajustado	0.24	0.89

Erros padrão em parêntesis
*** p<0.01, ** p<0.05, * p<0.1

Observe os coeficientes obtidos, os valores dos erros-padrão, das significâncias e as medidas de ajuste do modelo.

1) O que muda com a introdução das variáveis explicativas e o controle no modelo?

Com a introdução da variável explicativa 3 e os controles, os coeficientes estimados para as variáveis explicativas 1 e 2 perdem significância estatística devido ao aumento do valor do erro padrão, ainda que os valores dos coeficientes permaneceram próximos. O modelo 2 apresentou melhor ajuste, com um R^2 indicando que o conjunto de variáveis do modelo 2 explica 89% da variação da variável dependente.

2) Os resultados apresentados no modelo 2 são convincentes? Por que? Discuta. (Dica: considere em sua resposta as medidas de significância estatística e a de ajuste do modelo)

Contudo, o fato de não haver uma nova variável "roubando" o efeito no segundo modelo, levanta um sinal de alerta quanto às premissas do modelo linear. Sendo assim, os resultados do modelo 2 não são convincentes e precisamos investigar melhor.

A situação ocorrida no resultado apresentado acima se manifesta de outras formas também.

3) Qual o nome da ocorrência detectada no modelo 2 acima?

A ocorrência detectada acima é a multicolinearidade, em que há correlação entre variáveis independentes em um modelo de regressão.

Usando o arquivo que está disponibilizado no Moodle para esta aula ("Base_Lista_Aula_6"), faça o seguinte: Considere que esta base indica duas variáveis dependentes (Y1 e Y2), duas explicativas (X1) e dois controles (C1 e C2). Comecemos com a variável dependente Y1.

**4) Rode quatro modelos introduzindo uma variável independente por vez – primeiro x1, depois x2 - e depois as dependentes juntas x1 e x2, depois x1, x2 e c1 apenas. O que acontece com os coeficientes destas variáveis entre os modelos? Discuta;

```
library(tidyverse)
library(stargazer)
library(lmtest)
library(olsrr)
dados <- read.csv("Aula 6 - Lab 3 - Base moodle.csv")

reg1 <- dados %>% lm(y1 ~ x1, data=.)
reg2 <- dados %>% lm(y1 ~ x2, data=.)
reg3 <- dados %>% lm(y1 ~ x1 + x2, data=.)
reg4 <- dados %>% lm(y1 ~ x1 + x2 + c1, data=.)

stargazer(reg1, reg2, reg3, reg4,
          style="ajps",
          column.labels=c("x1", "x2", "x1 + x2", "x1 + x2 + c1"),
          omit.stat=c("f"),
          header=FALSE)
```

Tabela 1:

	y1			
	x1	**x2**	**x1 + x2**	**x1 + x2 + c1**
	Model 1	**Model 2**	**Model 3**	**Model 4**
x1	12.275***		12.275***	8.251*
	(4.036)		(4.012)	(4.373)
x2		−3.621***	−3.621***	−3.872***
		(1.007)	(1.003)	(1.007)
c1				5.030**
				(2.197)
Constant	−11.095	42.420***	17.871	12.841
	(8.321)	(8.306)	(11.524)	(11.708)
N	1000	1000	1000	1000
R-squared	0.009	0.013	0.022	0.027
Adj. R-squared	0.008	0.012	0.020	0.024
Residual Std. Error	63.790 (df = 998)	63.674 (df = 998)	63.409 (df = 997)	63.275 (df = 996)

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

Resultado na tabela 1

O coeficiente em x1 permanece estável e significativo nos modelos 1 e 3, perdendo efeito com a inclusão da variável de controle c1 e deixando de ser estatisticamente significante ao nível de 99,9%, passando para 90,0%.

Já x2 apresenta estabilidade entre todos os modelos em que aparece em termos de tamanho do efeito e significância estatística.

Por fim, o modelo 4 introduz a variável c1, com efeito positivo de 5.03 sobre y1 e estatisticamente significativo ao nível de 95%.

5) Agora rode um quinto modelo completo com as duas variáveis explicativas e as duas variáveis de controle. O que acontece com os coeficientes agora? Você detecta alguma alteração importante? Quais seriam os motivos?

```
reg5 <- dados %>% lm(y1 ~ x1 + x2 + c1 + c2, data=.)
stargazer(reg4, reg5,
          style="ajps",
          column.labels=c("x1 + x2 + c1", "x1 + x2 + c1 + c2"),
          omit.stat=c("f"),
          header=FALSE)
```

Tabela 2:

	y1	
	x1 + x2 + c1	x1 + x2 + c1 + c2
	Model 1	Model 2
x1	8.251*	−4.461
	(4.373)	(14.342)
x2	−3.872***	−2.934**
	(1.007)	(1.425)
c1	5.030**	6.064**
	(2.197)	(2.462)
c2		2.201
		(2.365)
Constant	12.841	36.456
	(11.708)	(27.946)
N	1000	1000
R-squared	0.027	0.028
Adj. R-squared	0.024	0.024
Residual Std. Error	63.275 (df = 996)	63.279 (df = 995)

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

Resultado na tabela 2.

Com a inclusão de c2, a alteração mais importante a ser observada é a mudança do sinal do coeficiente em x1, que passou de positivo em todos os modelos rodados até então para um valor negativo, invertendo o efeito desta sobre y1. Além disso, este coeficiente perde significância estatística, com o aumento do erro padrão.

A possível causa para esse comportamento é a presença de multicolinearidade.

6) Procure na lista de comandos do R para a aula de hoje um que permita checar se a suposta causa do comportamento observado considerado no item anterior.

Podemos usar o teste VIF (variance inflation factor) para testar a multicolinearidade do modelo. Para tanto, usamos a função `ols_vif_tol` do pacote `olsrr`:

```
ols_vif_tol(reg4)
```

```
##   Variables Tolerance      VIF
## 1        x1 0.8383838 1.192771
## 2        x2 0.9880952 1.012048
## 3        c1 0.8300000 1.204819
```

```
ols_vif_tol(reg5)
```

```
##   Variables  Tolerance       VIF
## 1        x1 0.07794811 12.829047
## 2        x2 0.49328359  2.027231
## 3        c1 0.66100000  1.512859
## 4        c2 0.07963855 12.556732
```

Nestes testes, valores próximos a 1 indicam que não há multicolinearidade entre variáveis independentes de um modelo, enquanto valores maiores constituem evidência de multicolinearidade entre elas. Sendo assim, nota-se que há evidências muito maiores de multicolinearidade no modelo em que adicionamos a segunda variável de controle c2, que é justamente a que apresenta o maior fator encontrado no teste realizado, junto com x1.

Há um outro teste necessário a ser feito com o intuito de se certificar de que não há violação às hipóteses básicas do modelo MQO.

7) Rode um modelo multivariado Y = f (X2, C1, C2), primeiro apenas com a única variável explicativa x2 e depois, incluindo os controles. Escreva a nova equação estimada;

```
reg6 <- dados %>% lm(y1 ~ x2, data=.)
reg7 <- dados %>% lm(y1 ~ x2 + c1 + c2, data=.)
stargazer(reg6, reg7,
          style="ajps",
          column.labels=c("x2", "x1 + x2 + c1 + c2"),
          omit.stat=c("f"),
          header=FALSE)
```

Tabela 3:

	y1	
	x2	x1 + x2 + c1 + c2
	Model 1	Model 2
x2	-3.621^{***}	-3.228^{***}
	(1.007)	(1.065)
c1		5.651^{***}
		(2.073)
c2		1.500^{**}
		(0.721)
Constant	42.420^{***}	28.328^{***}
	(8.306)	(9.897)
N	1000	1000
R-squared	0.013	0.028
Adj. R-squared	0.012	0.025
Residual Std. Error	63.674 (df = 998)	63.250 (df = 996)

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

Resultados na tabela 3.

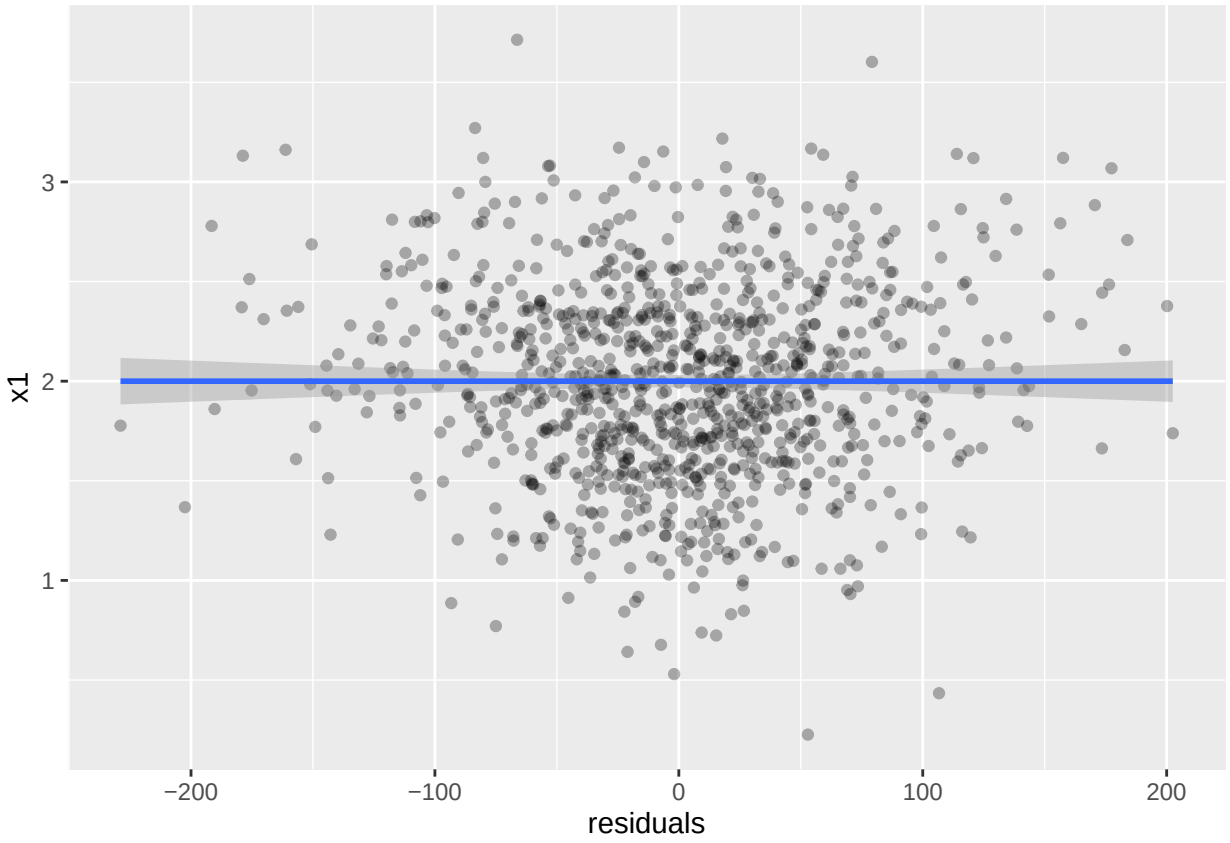
$\hat{y} = 28.33 + -3.62 * x2 + \epsilon$ $\hat{y} = 28.33 + -3.23 * x2 + 5.54 * c1 + 1.50 * c2 + \epsilon$

8) Construa um gráfico de dispersão que relacione a variável explicativa com os resíduos. O que é possível notar neste gráfico a respeito da relação entre ambas as variáveis?

```
dados$residuals <- reg5$residuals

ggplot(dados, aes(residuals, x1))+
  geom_point(alpha=0.3)+
  stat_smooth(method="lm")
```

```
## `geom_smooth()` using formula 'y ~ x'
```



Com este gráfico, vemos que a relação entre a variável independente de interesse x1 não está correlacionada com os resíduos do modelo, indicando que não há evidência de heterocedasticidade, visto que caso houvesse, veríamos um padrão claro positivo ou negativo entre os valores do resíduos e os valores de x1 (algo do tipo: resíduos maiores estão atrelados a valores maiores de x1).

9) Encontre um outro teste entre os comandos para a aula de hoje que permita testar se esta outra hipótese foi ou não violada. Reporte os resultados. Qual sua conclusão?

Para testar a heterocedasticidade também podemos usar a função `bptest` *do pacote* `lmtest`:

```
bptest(reg4)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg4
## BP = 116.85, df = 3, p-value < 2.2e-16
```

```
bptest(reg5)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  reg5
## BP = 116.73, df = 4, p-value < 2.2e-16
```

Nesta função, aplica-se um teste de hipótese no qual a hipótese nula corresponde aos resíduos estarem distribuídos com variância igual, e a hipótese alternativa aos resíduos estarem distribuídos com variância desigual. Com isso, na verdade podemos observar a heterocedasticidade em ambos os modelos.