GABARITO - LISTA 2

MODELO DE REGRESSÃO LINEAR SIMPLES

Mateus Cardoso

04/05/2021

1) Hipótese RLS.1 (Linear em parâmetros): No modelo populacional, a variável dependente, y, está relacionada com a variável independente, x, e com o erro (ou perturbação), u, como $y = \beta_0 + \beta_1 x + u$, em que β_0 e β_1, são os parâmetros do intercepto e da inclinação populacionais, respectivamente.

Hipótese RLS.2 (Amostragem aleatória): Temos uma amostra aleatória de tamanho n, $\{(x_i, y_i) : i = 1, 2, ..., n)\}$, seguindo o modelo populacional na Hipótese RLS.1.

Hipótese RLS.3 (Variação amostral na variável explicativa): Os resultados amostrais em x, a saber $\{x_i,\ i = 1, ..., n)\}$, não são todos de mesmo valor.

Hipótese RLS.4 (Média condicional zero): O erro u tem zero como valor esperado, quaisquer que sejam os valores das variáveis. Em outras palavras, $E(u|x) = 0$.

Hipótese RLS.5 (Homoscedasticidade): O erro u tem a mesma variância quaisquer que sejam os valores das variáveis explicativas. Em outras palavras: $Var(u|x) = \sigma^2$.

2) O R^2 é uma medida de ajuste do modelo de regressão linear. Representa o quanto da variação total é explicada pela regressão. Seu valor varia de 0 a 1, sendo que um R^2 de 0 significa que não há nenhuma relação linear entre Y e X e um R^2 de 1 significa uma relação linear perfeita.

Vale reforçar que o coeficiente de determinação mede apenas a qualidade de ajuste do modelo, não implicando em qualquer relação causal entre as variáveis. O R^2 demonstra somente o quanto da variação em Y é explicada pelas variáveis inseridas em nosso modelo.[1]

3) a) Para estimar os parâmetros, precisamos primeiro dos valores de $(x_i - \bar{x})(y_i - \bar{y})$ e $(x_i - \bar{x})^2$.

Antes, calculamos as médias de x e y, chegando a $\bar{x} = 3$ e $\bar{y} = 5$.

[1] Uma boa introdução sobre causalidade pode ser encontrada no primeiro capítulo do livro Guia Brasileiro de Análise de Dados. Na lista de material complementar podem ser encontrados alguns livros sobre inferência causal. Sugiro Cunningham (2021) por possuir versão online gratuita e exemplos em R, Python e Stata. Huntington-Klein (2021) também possui versão online gratuita e foca na intuição por trás das fórmulas, podendo ser melhor para quem prefere um tratamento menos matemático. Em português, há os livros Avaliação Econômica de Projetos Sociais, do Itaú Social; e Avaliação de Impacto na Prática, do Banco Mundial. Estes, porém, possuem maior enfoque na avaliação de impacto de políticas públicas.

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
0	3	-3	-2	6	9
1	2	-2	-3	6	4
1	3	-2	-2	4	4
2	5	-1	0	0	1
3	4	0	-1	0	0
3	4	0	-1	0	0
4	7	1	2	2	1
5	6	2	1	2	4
5	7	2	2	4	4
6	9	3	4	12	9

Tendo os valores, podemos partir para os somatórios

$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum(x - \bar{x})^2$
36	36

Substituindo os valores na fórmula, temos $\hat{\beta} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{36}{36} = 1$.

Para descobrir $\hat{\alpha}$, basta substituir o valor de $\hat{\beta}$ na equação da reta. $\bar{y} = \hat{\alpha} + \hat{\beta}\bar{x} \rightarrow 5 = \hat{\alpha} + 1 * 3 \rightarrow \hat{\alpha} = 2$.

Portanto, chegamos às estimativas dos parâmetros $\hat{\alpha} = 2$ e $\hat{\beta} = 1$.

b) Precisamos calcular a SST e a SSE.

X	Y	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$\hat{y}$	$(\hat{y} - \bar{y})$	$(\hat{y} - \bar{y})^2$
0	3	-2	4	2	-3	9
1	2	-3	9	3	-2	4
1	3	-2	4	3	-2	4
2	5	0	0	4	-1	1
3	4	-1	1	5	0	0
3	4	-1	1	5	0	0
4	7	2	4	6	1	1
5	6	1	1	7	2	4
5	7	2	4	7	2	4
6	9	4	16	8	3	9

Basta somar os valores de $(y_i - \bar{y})^2$ e $(\hat{y_i} - \bar{y})^2$ para obter SST e SSE, respectivamente.

$\sum(y_i - \bar{y})^2$ (SST)	$\sum(\hat{y_i} - \bar{y})^2$ (SSE)
44	36

Resta substituir os valores na fórmula: $R^2 = \frac{SSE}{SST} = \frac{36}{44} = 0,818$. Ou seja, aproximadamente 82% da variação total em y é explicada por fatores que estão incluídos em nosso modelo.

c) Para calcular a variância do erro, precisamos primeiro calcular os valores de $\hat{u}_i^2$.

Y	$\hat{y}$	$\hat{u}$	$\hat{u^2}$
3	2	1	1
2	3	-1	1
3	3	0	0
5	4	1	1
4	5	-1	1
4	5	-1	1
7	6	1	1
6	7	-1	1
7	7	0	0
9	8	1	1

A fórmula para obter uma estimativa não viesada da variância do erro é $\frac{\sum \hat{u}_i^2}{n-2}$. Subsituindo os valores, temos $\frac{8}{10-2} = 1$. Portanto, a variância do erro é $\sigma^2 = 1$.

d) Aproveitando a tabela feita no exercício 1.b, podemos caclular a média dos valores observados a média dos valores estimados. Chegamos a:

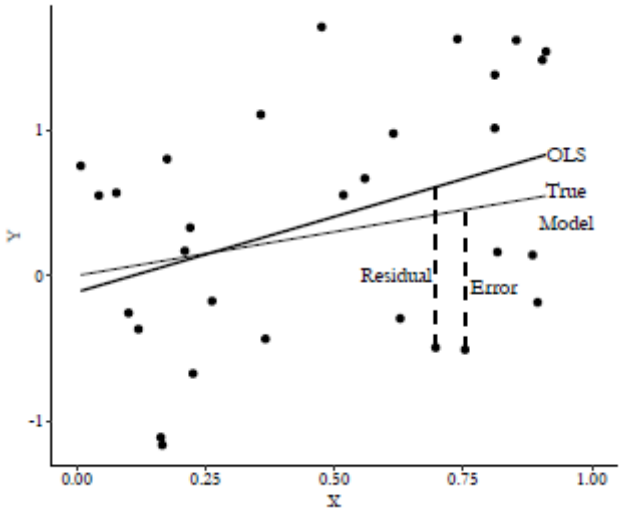
$\bar{y}$	$\bar{\hat{y}}$
5	5

A média dos valores estimados ($\bar{\hat{y}}$) é igual à média dos valores da amostra ($\bar{y}$) porque a soma dos resíduos é igual a 0. Se $y_i = \hat{y_i} + \hat{u_i}$, então $\frac{\sum y_i}{n} = \frac{\sum \hat{y_i}}{n}$, dado que $\sum \hat{u_i} = 0$.

4) O resíduo ($\hat{u}_i$) é a diferença entre o y estimado ($\hat{y}_i$) e o y observado y_i. Portanto, é facilmente calculado com qualquer amosra de dados.

O erro (u_i) é a diferença entre o y observado e o y que teríamos estimado caso tivessemos dados populacionais. Contém tudo que causa y mas não está incluído em nosso modelo.

Como temos apenas uma amostra finita, o melhor ajuste que obtivermos não será exatamente o mais adequado para toda a população. Então, tudo o que podemos realmente ver será o resíduo. A diferença entre o erro e o resíduo pode ser visualizada no gráfico abaixo, retirado de Huntington-Klein (2021):



5) a) Renda, idade e estrutura familia (como número de irmãos, por exemplo) são apenas algumas possibilidades. A princípio, cada um destes fatores pode ser correlacionado com os anos de educação. (Renda e a educação provavelmente estão positivamente correlacionados; idade e educação podem estar negativamente correlacionados porque as mulheres mais jovens possuem, em média, mais escolaridade; e número de irmãos e a educação provavelmente estão negativamente correlacionados.)

b) Não se os fatores listados na parte (i) estiverem correlacionados com *educ*. Já que estamos mantendo estes fatores fixos, eles fazem parte do erro (u). Mas se u está correlacionado com *educ*, então $E(u|educ) \neq 0$, e assim a hipótese RLS.4 (Média condicional zero) falha.

6) a) Primeiro, calculamos as médias $\bar{x} = 170$ e $\bar{y} = 111$, então montamos a tabela com os valores necessários:

X	Y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
80	70	-90	-41	3690	8100
100	65	-70	-46	3220	4900
120	90	-50	-21	1050	2500
140	95	-30	-16	480	900
160	110	-10	-1	10	100
180	115	10	4	40	100
200	120	30	9	270	900
220	140	50	29	1450	2500
240	155	70	44	3080	4900
260	150	90	39	3510	8100

Somando os valores, temos:

$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum(x - \bar{x})^2$
16800	33000

Substituindo os valores na fórmula, temos $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{16800}{33000} = 0,509$.

Para descobrir $\hat{\beta}_0$, basta substituir o valor de $\hat{\beta}_1$ na equação da reta. $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \rightarrow 111 = \hat{\beta}_0 + 0,509 * 170 \rightarrow \hat{\beta}_0 = 24,47$.

Portanto, chegamos às estimativas dos parâmetros $\hat{\beta}_0 = 24,47$ e $\hat{\beta} = 0,509$.

b) Precisamos calcular a SST e a SSE.

X	Y	$(y_i - \bar{y})$	$(y_i - \bar{y})^2$	$\hat{y}$	$(\hat{y} - \bar{y})$	$(\hat{y} - \bar{y})^2$
80	70	-41	1681	65.19	-45.81	2098.5561
100	65	-46	2116	75.37	-35.63	1269.4969
120	90	-21	441	85.55	-25.45	647.7025
140	95	-16	256	95.73	-15.27	233.1729
160	110	-1	1	105.91	-5.09	25.9081
180	115	4	16	116.09	5.09	25.9081
200	120	9	81	126.27	15.27	233.1729
220	140	29	841	136.45	25.45	647.7025
240	155	44	1936	146.63	35.63	1269.4969
260	150	39	1521	156.81	45.81	2098.5561

Basta somar os valores de $(y_i - \bar{y})^2$ e $(\hat{y}_i - \bar{y})^2$ para obter SST e SSE, respectivamente.

$\sum(y_i - \bar{y})^2$ (SST)	$\sum(\hat{y}_i - \bar{y})^2$ (SSE)
8890	8549.673

Resta substituir os valores na fórmula: $R^2 = \frac{SSE}{SST} = \frac{8549.673}{8890} = 0,962$.

c)Para estimar a elasticidade-renda do consumo, precisamos de um modelo log-log. Ou seja, devemos transformar o X em $\ln(X)$ e Y em $\ln(Y)$. $\ln(X)$ medirá a variação percentual da renda e $\ln(Y)$ medirá a variação percentual do consumo. Nosso modelo passará a ser, portanto, $\ln(y_i) = \beta_0 + \beta_1 \ln(x_i) + u$.

X	Y	ln(X)	ln(Y)	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(x_i - \bar{x})(y_i - \bar{y})$	$(x_i - \bar{x})^2$
80	70	4.382027	4.248495	-0.6897461	-0.4220224	0.2910883	0.4757497
100	65	4.605170	4.174387	-0.4666025	-0.4961303	0.2314957	0.2177179
120	90	4.787492	4.499810	-0.2842810	-0.1707079	0.0485290	0.0808157
140	95	4.941642	4.553877	-0.1301303	-0.1166407	0.0151785	0.0169339
160	110	5.075174	4.700480	0.0034011	0.0299628	0.0001019	0.0000116
180	115	5.192957	4.744932	0.1211841	0.0744145	0.0090179	0.0146856
200	120	5.298317	4.787492	0.2265447	0.1169741	0.0264999	0.0513225
220	140	5.393628	4.941642	0.3218548	0.2711248	0.0872628	0.1035905
240	155	5.480639	5.043425	0.4088662	0.3729075	0.1524693	0.1671716
260	150	5.560682	5.010635	0.4889089	0.3401177	0.1662866	0.2390319

Somando os valores, temos:

$\sum(x_i - \bar{x})(y_i - \bar{y})$	$\sum(x - \bar{x})^2$
1.02793	1.367031

Substituindo os valores na fórmula, temos $\hat{\beta}_1 = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sum(x_i - \bar{x})^2} = \frac{1,02793}{1,367031} = 0,752$.

Para descobrir $\hat{\beta}_0$, basta substituir o valor de $\hat{\beta}_1$ na equação da reta. $\bar{y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{x} \rightarrow 4,6705 = \hat{\beta}_0 + 0,752 * 5,0717 \rightarrow \hat{\beta}_0 = 0,857$.

Assim, chegamos às estimativas dos parâmetros $\hat{\beta}_0 = 0,857$ e $\hat{\beta} = 0,752$.

Na interpretação dos coeficientes, temos que:

- $\hat{\beta}_0$: quando $\ln(X) = 0$, o valor esperado de Y é $e^{\hat{\beta}_0}$. Com $\hat{\beta}_0 = 0,857$, temos portanto que quando $\ln(X) = 0$, o valor esperado de Y é $e^{0,857} = US$2,36.
- $\hat{\beta}_1$: quando X aumenta em 1%, esperamos um aumento de $\hat{\beta}_1$% em Y. Portanto, com $\hat{\beta} = 0,752$, quando quando X aumenta em 1%, esperamos um aumento de $0,752$% em Y.

Portanto, diante de uma variação de 1% na renda semanal, há uma variação de 0,752% no consumo semanal.

Exercícios no R

1) Inicialmente, criamos um tibble com os valores requisitados. Para definir o desvio-padrão, basta inserir o argumento "sd =" dentro da função rnorm().

```
set.seed(1) #garante que somente uma amostra aleatória seja gerada,
            #assim, garante reproducibilidade caso clique de novo

n<-10000  #definindo n = 10000

dados<-tibble(x = rnorm(n, sd = 9), #dist. normal com n = 10000 e sd = 9
              u = rnorm(n, sd = 36), #dist. normal com n = 10000 e sd = 9
              y = 3 + 2*x + u, #modelo dado no problema
              yhat = predict(lm(y ~ x)), #predict() nos dá o valor do y estimado
              uhat = residuals(lm(y ~ x))) #resíduos

head(dados) #visualizar a base
```

```
## # A tibble: 6 x 5
##       x     u      y   yhat  uhat
##   <dbl> <dbl>  <dbl>  <dbl> <dbl>
## 1 -5.64 -29.0 -37.2   -8.53 -28.7
## 2  1.65 -38.0 -31.7    6.19 -37.9
## 3 -7.52 -37.3 -49.3  -12.3  -37.0
## 4 14.4  -42.7 -11.0   31.8  -42.8
## 5  2.97 -18.0  -9.08   8.84 -17.9
## 6 -7.38 -18.9 -30.7  -12.1  -18.6
```

a) Com a sunção sum(), podemos checar se $\sum \hat{u}_i = 0$.

```
round(sum(dados$uhat), 5) #Arredondamos para 5 casas decimais para evitar notação científica
```

```
## [1] 0
```

Devido a arredondamentos feitos pelo R, a função sum() não retornará exatamente 0, mas sim um valor estatisticamente igual a 0.

b)

```
mean(dados$yhat)
```

```
## [1] 2.731491
```

```
mean(dados$y)
```

```
## [1] 2.731491
```

Portanto, as médias do y observado e do estimado são iguais.

2) (i) Primeiro, criaremos um dataframe para armazenar os dados do ano de 1996.

```
library(wooldridge)
```

```
## Warning: package 'wooldridge' was built under R version 4.0.5
```

```
countymurders_1996<-countymurders %>%
filter(year == 1996)
```

Agora, podemos contar quantos condados tiveram zero assassinatos no ano de 1996.

```
countymurders_1996 %>% filter(murders == 0) %>% count()
```

```
##      n
## 1 1051
```

Contaremos agora quantos condados tiveram pelo menos uma execução.

```
countymurders_1996 %>% filter(execs >= 1) %>% count()
```

```
##    n
## 1 31
```

Por fim, verificamos qual o maior número de execuções:

```
max(countymurders_1996$execs)
```

```
## [1] 3
```

(ii) Estimaremos a regressão com o comando lm(), que significa linear model. Não esqueça que nossa análise se limita apenas ao ano de 1996.

```
reg<- lm(murders ~ execs, data = countymurders_1996)
```

Com o comando summary(), podemos ver todas as informações solicitadas.

```
summary(reg)
```

```
##
## Call:
## lm(formula = murders ~ execs, data = countymurders_1996)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -149.12   -5.46   -4.46   -2.46 1338.99
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)   5.4572     0.8348   6.537      0.0000000000779 ***
## execs        58.5555     5.8333  10.038 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 38.89 on 2195 degrees of freedom
## Multiple R-squared:  0.04389,    Adjusted R-squared:  0.04346
## F-statistic: 100.8 on 1 and 2195 DF,  p-value: < 0.00000000000000022
```

Vemos que temos os coeficientes $\hat{\beta}_0 = 5,457$ e $\hat{\beta}_1 = 58,555$, e nosso modelo possui R^2 de $0,044$. A função summary() no entanto, não nos fornece o tamanho da amostra. Podemos utilizar o pacote modelsummary para obter uma tabela melhor.

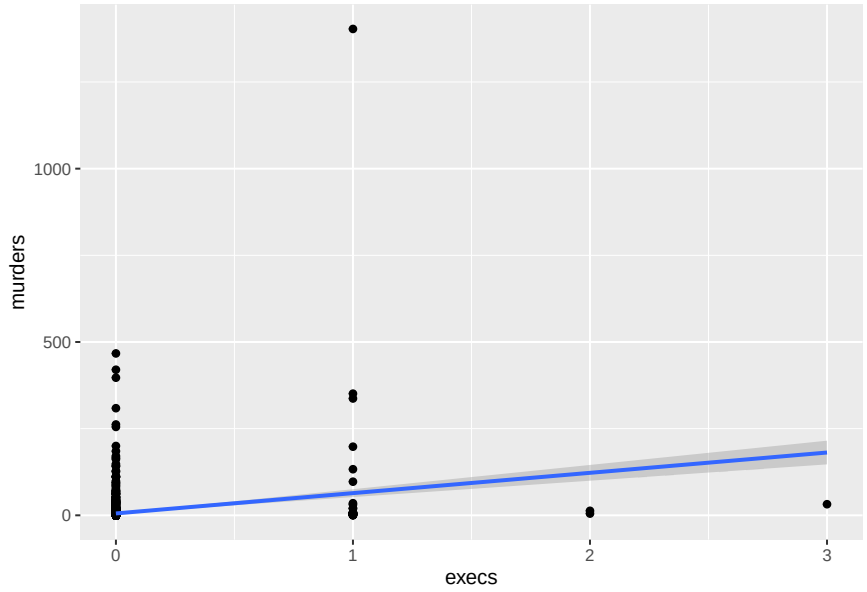
```
library(modelsummary)

modelsummary(reg, estimate = c("{estimate} ({std.error}){stars}"))%>%
  kable_styling(full_width = F,
                latex_options = "HOLD_position")
```

	Model 1
(Intercept)	5.457 (0.835)***
	(0.835)
execs	58.555 (5.833)***
	(5.833)
Num.Obs.	2197
R2	0.044
R2 Adj.	0.043
AIC	22323.8
BIC	22340.9
Log.Lik.	-11158.905
F	100.766

Vemos que o tamanho da amostra é n = 2197. Os argumentos além de `modelsummary(reg)` são apenas para melhorar a forma da tabela no documento.

Podemos também fazer um gráfico:

```
countymurders_1996 %>%
  ggplot(aes(x = execs, y = murders))+
  geom_point()+
  geom_smooth(method = "lm")
```



(iii) Podemos interpretar nosso coeficiente de inclinação $\hat{\beta}_1$ da forma: quando o número de execuções cresce em uma unidade, o número de assassinatos cresce em 58,55 unidades. Portanto, a equação estimada na verdade não sugere um efeito dissuasor, mas um efeito positivo do número de execuções sobre o número de assassinatos.

Essa relação no entanto, carece de maior análise. Isso será discutido no item (v).

(iv) O menor número de assassinatos que pode ser previsto pela equação estimada é o coeficiente $\hat{\beta}_0$. Ele indica que, quando a variável execuções toma valor igual a zero, o número de assassinatos é 5,457.

Se formos olhar para os dados, essa estimativa não se sustenta. Por exemplo, se selecionarmos entre os condados que tiveram número de execuções igual a zero os 10 condados com maiores números de assassinatos, veremos que os valores fogem muito à nossa estimativa de β_0.

```
countymurders_1996 %>%
  filter(execs == 0) %>%
  select(murders, execs) %>%
  top_n(10, murders) %>%
  arrange(-murders)
```

```
##    murders execs
## 1      467     0
## 2      420     0
## 3      397     0
## 4      309     0
## 5      262     0
## 6      255     0
```

```
## 7      200     0
## 8      185     0
## 9      171     0
## 10     166     0
```

Quanto ao resíduo, este é igual a $y_i - \hat{y}_i$. No caso de um condado com $y_i = 0$ e $x_i = 0$, o resíduo tomará o valor de $\hat{u}_i = -\hat{\beta}_0$, que neste caso é $\hat{u}_i = -5,457$.

(v) Pelo valor do R^2 se vê que *execs* não explica muito da variação em *murders*. Uma análise de regressão simples não é adequada para determinar se a pena capital tem um efeito dissuasor sobre os assassinatos pois muitos outros fatores podem também afetar o número de assassinatos. Por exemplo a renda per capita do condado, a taxa de desemprego, o % de pessoas que já haviam sido presas anteriormente, o nível educacional, entre outros.

Provavelmente há também uma relação no sentido contrário entre as variáveis, com o número de assassinatos influenciando o número de execuções. Pode haver também a presença de uma variável confusora, que afeta tanto *murders* quanto *execs* simultâneamente.

Para exemplificar, podemos fazer uma regressão levando em consideração outros fatores além de *execs*.

```
reg2<- lm(murders ~ execs + density + popul + rpcpersinc, data = countymurders_1996)

summary(reg2)
```

```
##
## Call:
## lm(formula = murders ~ execs + density + popul + rpcpersinc,
##     data = countymurders_1996)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -196.30   -2.78    0.62    4.10  340.73
##
## Coefficients:
##                 Estimate   Std. Error t value             Pr(>|t|)
## (Intercept) 23.489929162  1.855381994  12.660 < 0.0000000000000002 ***
## execs       13.456380412  2.849095051   4.723  0.00000247049732958 ***
## density      0.002050700  0.000259578   7.900  0.00000000000000437 ***
## popul        0.000121531  0.000001552  78.286 < 0.0000000000000002 ***
## rpcpersinc  -0.002356557  0.000148631 -15.855 < 0.0000000000000002 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 18.63 on 2192 degrees of freedom
## Multiple R-squared:  0.781,  Adjusted R-squared:  0.7806
## F-statistic:  1954 on 4 and 2192 DF,  p-value: < 0.00000000000000022
```

Podemos ver que nosso $\hat{\beta}_1$ passou de $58,55$ para $13,456$, e o R^2 passou de $0,044$ para $0,781$.