GABARITO - LISTA 3

MODELO DE REGRESSÃO LINEAR MÚLTIPLA: ESTIMAÇÃO

Mateus Cardoso

25/06/2021

1) Hipótese RLM.1 (Linear em parâmetros): O modelo na população pode ser escrito da seguinte forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_k x_k + u$$

em que β_0, β_1, ..., β_k são os parâmetros desconhecidos (constantes) de interesse e u é um erro aleatório não observável ou termo de perturbação.

Hipótese RLM.2 (Amostragem aleatória): Temos uma amostra aleatória de n observações, $\{(x_{i1}, x_{i2}, ..., x_{ik}, y_i) : i = 1, 2, ..., n\}$, seguindo o modelo populacional na Hipótese RLM.1.

Hipótese RLM.3 (Colinearidade imperfeita): Na amostra (e, portanto, na população), nenhuma das variáveis independentes é constante e não existem relacionamentos lineares exatos entre as variáveis independentes.

Hipótese RLM.4 (Média condicional zero): O erro u tem zero como valor esperado dados quaisquer valores das variáveis independentes. Em outras palavras:

$$E(u|x_1, x_2, ..., x_k) = 0.$$

Hipótese RLM.5 (Homoscedasticidade): O erro u tem a mesma variância dados quaisquer valores das variáveis explicativas. Em outras palavras:

$$Var(u|x_1, ..., x_k) = \sigma^2.$$

2) Sob as primeiras quatro hipóteses de Gauss-Markov (RLM.1 a RLM.4), os estimadores de MQO são não viesados. Isso implica que a inclusão de uma variável irrelevante em um modelo não tem nenhum efeito sobre a inexistência de viés dos estimadores de intercepto e de inclinação. De outro lado, omitir uma variável importante faz que MQO seja viesado. Em muitas circunstâncias, a direção do viés pode ser determinada.

3) O Teorema de Gauss-Markov nos diz que sob as Hipóteses RLM.1 a RLM.5, os estimadores de MQO $\hat{\beta}_0$, $\hat{\beta}_1$, ..., $\hat{\beta}_k$ são os melhores estimadores lineares não viesados de β_0, β_1, ..., β_k, respectivamente.

A importância do Teorema de Gauss-Markov é que, quando um conjunto padrão de hipóteses se mantém, não necessitamos procurar por estimadores não viesados alternativos, pois nenhum será melhor que MQO. Equivalentemente, se somos apresentados a um estimador que é tanto linear quanto não viesado, então sabemos que a variância desse estimador é pelo menos tão grande quanto a variância de MQO; nenhum cálculo adicional é necessário para demonstrar isso.

4) Apenas (ii), omitir uma variável importante, pode causar viés, e isso só é verdade quando o a variável omitida é correlacionada com as variáveis explicativas incluídas.

A hipótese de homoscedasticidade, RLM.5, não desempenha nenhum papel em mostrar que os estimadores MQO são não viesados (Homoscedasticidade foi usada para obter as fórmulas de variância usuais para $\hat{\beta}_k$).

Além disso, o grau de colinearidade entre as variáveis explicativas da amostra, mesmo que se reflita em uma correlação tão alta quanto $0,95$, não afeta as suposições de Gauss-Markov. Somente se houver uma relação linear perfeita entre duas ou mais variáveis explicativas haverá violação de RLM.3.

5) (i) Se os adultos trocam o sono pelo trabalho, mais trabalho implica menos sono (tudo mais constante), então $\beta_1 < 0$.

(ii) Os sinais de β_2 e β_3 não são tão óbvios. Alguém poderia argumentar que pessoas mais educadas gostam de tirar mais proveito da vida e, assim, tudo mais constante, dormem menos ($\beta_2 < 0$). A relação entre sono e idade é mais complicada do que este modelo sugere.

(iii) Como *totwrk* está em minutos, devemos converter cinco horas em minutos: $\Delta totwrk = 5 * 60 = 300$. Então, prevê-se que o *sleep* caia em $0,148 * 300 = 44,4$ minutos. Em uma semana, 45 minutos a menos de sono não é uma mudança considerável.

(iv) Mais educação implica menos tempo previsto para dormir, mas o efeito é muito pequeno. Se presumimos que a diferença entre a faculdade e o ensino médio é de quatro anos, o graduado dorme cerca de 45 minutos a menos por semana, tudo mais constante.

(v) Não surpreendentemente, as três variáveis explicativas explicam apenas cerca de 11,3% do variação em *sleep*. Um fator importante no termo de erro é a saúde. Outro é estado civil, e se a pessoa tem filhos. Saúde, estado civil, e o número e as idades das crianças geralmente seriam correlacionados com o *totwrk*.

6) (i) Não. Por definição, $study + sleep + work + leisure = 168$. Portanto, se mudarmos o tempo de *study*, devemos mudar pelo menos uma das outras categorias para que a soma ainda seja 168.

(ii) Da parte (i), podemos escrever, por exemplo, *study* como uma função linear perfeita das outras variáveis independentes: $study = 168 - sleep - work - leisure$. Isso vale para todas as observações, então RLM.3 é violado.

(iii) Simplesmente elimine uma das variáveis independentes, digamos *leisure*:

$$GPA = \beta_0 + \beta_1 study + \beta_2 sleep + \beta_3 work + u.$$

Agora, por exemplo, β_1 é interpretado como a mudança no GPA quando *study* aumenta em uma hora, enquanto *sleep*, *work* e u são mantidos fixos. Se estamos mantendo *sleep* e *work* fixos, mas aumentamos *study* em uma hora, então devemos reduzir *leisure* em uma hora. Os outros parâmetros de inclinação têm uma interpretação semelhante.

7) (i) Dado que x_1 é altamente correlacionado com x_2 e x_3, e que essas últimas variáveis têm grandes efeitos parciais em y, os coeficientes de regressão simples e múltipla em x_1 podem diferir bastante. Para mais detalhes, ler seção 3-3b do Wooldridge.

(ii) Neste caso, esperamos que $\tilde{\beta}_1$ e $\hat{\beta}_1$ sejam semelhantes (sujeito, é claro, ao que queremos dizer com "Quase não correlacionado"). O grau de correlação entre x_2 e x_3 não afeta diretamente a estimativa de regressão múltipla em x_1 se x_1 é essencialmente não correlacionado com x_2 e x_3.

(iii) Neste caso, estamos (desnecessariamente) introduzindo multicolinearidade na regressão: x_2 e x_3 têm pequenos efeitos parciais em y e ainda são altamente correlacionados com x_1. Adicionar x_2 e x_3 aumenta o erro padrão do coeficiente de x_1 substancialmente, então $\text{ep}(\hat{\beta}_1)$ tende a ser maior que $\text{ep}(\tilde{\beta}_1)$.

Exercícios no R

1) Primeiramente, devemos chamar os pacotes wooldridge e tidyverse e importar a base "ceosal2" para nosso Environment.

```
library(tidyverse)
library(wooldridge)

ceo <- ceosal2
```

Agora, vamos visualizar a base para ver que variáveis ela possui e como estas se apresentam.

```
ceo %>% glimpse()
```

```
## Rows: 177
## Columns: 15
## $ salary   <int> 1161, 600, 379, 651, 497, 1067, 945, 1261, 503, 1094, 601, 35~
## $ age      <int> 49, 43, 51, 55, 44, 64, 59, 63, 47, 64, 54, 66, 72, 51, 63, 4~
## $ college  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 0, 1, 1, 1~
## $ grad     <int> 1, 1, 1, 0, 1, 1, 0, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 0, 1, 0, 1~
## $ comten   <int> 9, 10, 9, 22, 8, 7, 35, 32, 4, 39, 26, 39, 37, 25, 21, 7, 38,~
## $ ceoten   <int> 2, 10, 3, 22, 6, 7, 10, 8, 4, 5, 7, 8, 37, 1, 11, 7, 4, 12, 2~
## $ sales    <dbl> 6200, 283, 169, 1100, 351, 19000, 536, 4800, 610, 2900, 1200,~
## $ profits  <int> 966, 48, 40, -54, 28, 614, 24, 191, 7, 230, 34, 8, 35, 234, 9~
## $ mktval   <dbl> 23200, 1100, 1100, 1000, 387, 3900, 623, 2100, 454, 3900, 533~
## $ lsalary  <dbl> 7.057037, 6.396930, 5.937536, 6.478509, 6.208590, 6.972606, 6~
## $ lsales   <dbl> 8.732305, 5.645447, 5.129899, 7.003066, 5.860786, 9.852194, 6~
## $ lmktval  <dbl> 10.051908, 7.003066, 7.003066, 6.907755, 5.958425, 8.268732, ~
## $ comtensq <int> 81, 100, 81, 484, 64, 49, 1225, 1024, 16, 1521, 676, 1521, 13~
## $ ceotensq <int> 4, 100, 9, 484, 36, 49, 100, 64, 16, 25, 49, 64, 1369, 1, 121~
## $ profmarg <dbl> 15.580646, 16.961130, 23.668638, -4.909091, 7.977208, 3.23157~
```

Podemos ver que já possuimos as variáveis *lsalary*, *lsales* e *lmktval*, que correspondem, respectivamente, ao logarítimo natural das variáveis *salary*, *sales* e *mktval*. Estas variáveis serão necessárias para a estimação do modelo de elasticidade em (i).

(i) Para estimar o modelo, utilizaremos o comando lm().

```
elast <- lm(lsalary ~ lsales + lmktval, data = ceo)

summary(elast)
```

```
##
## Call:
## lm(formula = lsalary ~ lsales + lmktval, data = ceo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.28060 -0.31137 -0.01269  0.30645  1.91210
##
## Coefficients:
##             Estimate Std. Error t value             Pr(>|t|)
## (Intercept)  4.62092    0.25441  18.163 < 0.0000000000000002 ***
## lsales       0.16213    0.03967   4.087            0.0000667 ***
## lmktval      0.10671    0.05012   2.129               0.0347 *
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5103 on 174 degrees of freedom
## Multiple R-squared:  0.2991, Adjusted R-squared:  0.2911
## F-statistic: 37.13 on 2 and 174 DF,  p-value: 0.00000000000003727
```

Escrevendo nossos resultados em forma de equação, temos

$$\widehat{ln(salary)} = 4,621 + 0,162ln(sales) + 0,107ln(mktval)$$

(ii) Se voltarmos ao início deste exercício, onde fizemos a visualização da base de dados, já podemos ver o porquê de não poder usar a variável $profits$ na forma logarítimica: existem valores negativos. Teremos então de adicionar esta variável no formato de nível.

```
modelo2 <- lm(lsalary ~ lsales + lmktval + profits, data = ceo)

summary(modelo2)
```

```
##
## Call:
## lm(formula = lsalary ~ lsales + lmktval + profits, data = ceo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.27002 -0.31026 -0.01027  0.31043  1.91489
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 4.68692438 0.37972940  12.343 < 0.0000000000000002 ***
## lsales      0.16136826 0.03991005   4.043            0.0000792 ***
## lmktval     0.09752857 0.06368863   1.531                0.128
## profits     0.00003566 0.00015196   0.235                0.815
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5117 on 173 degrees of freedom
## Multiple R-squared:  0.2993, Adjusted R-squared:  0.2872
## F-statistic: 24.64 on 3 and 173 DF,  p-value: 0.000000000000253
```

Podemos escrever nossa equação agora como

$$\widehat{ln(salary)} = 4,687 + 0,161ln(sales) + 0,098ln(mktval) + 0,000036profits$$

O coeficiente de lucros é muito pequeno. Aqui, os lucros são medidos em milhões, portanto, se $profits$ aumentam em $1 bilhão, o que significa $\Delta profits = 1.000$ − uma grande mudança − o salário previsto aumenta cerca de 3,6%, apenas. No entanto, devemos lembrar de que estamos mantendo as vendas ($lsales$) e o valor de mercado ($lmktval$) fixos.

Essas variáveis explicam apenas 29,9% dos salários dos CEOs. Devemos observar que o R-quadrado continua o mesmo que no modelo estimado em (i), portanto, não perderíamos nada ao retirar a variável $profits$ do modelo. Devemos notar também que as estimativas dos parâmetros de $lmktval$ e $profits$ não possuem significância estatística.

Respondendo à pergunta, o modelo não explica a maior parte da variação dos salários dos CEOs.

(iii) Vamos estimar o modelo com a variável *ceoten*, que mede o tempo que o CEO está na empresa (*tenure*).

```
modelo3 <- lm(lsalary ~ lsales + lmktval + profits + ceoten, data = ceo)

summary(modelo3)
```

```
##
## Call:
## lm(formula = lsalary ~ lsales + lmktval + profits + ceoten, data = ceo)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.48792 -0.29369  0.00827  0.29951  1.85524
##
## Coefficients:
##               Estimate Std. Error t value             Pr(>|t|)
## (Intercept) 4.55778026 0.38025481  11.986 < 0.0000000000000002 ***
## lsales      0.16223394 0.03948256   4.109            0.0000614 ***
## lmktval     0.10175975 0.06303296   1.614               0.1083
## profits     0.00002905 0.00015035   0.193               0.8470
## ceoten      0.01168467 0.00534202   2.187               0.0301 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5062 on 172 degrees of freedom
## Multiple R-squared:  0.3183, Adjusted R-squared:  0.3024
## F-statistic: 20.08 on 4 and 172 DF,  p-value: 0.0000000000001387
```

Temos a equação

$$\widehat{ln(salary)} = 4,558 + 0,162ln(sales) + 0,102ln(mktval) + 0,000029profits + 0,012ceoten$$

O retorno percentual estimado para mais um ano de permanência do CEO, tudo mais constante, é cerca de 1,2%.

(iv) Para encontrar a correlação, utilizamos o comando cor().

```
cor(ceo$lmktval, ceo$profits)
```

```
## [1] 0.7768976
```

A correlação amostral entre *lmktval* e $profits$ é de cerca de 0,78 (78%), o que é consideravelmente alto. Como sabemos, isso não causa viés nos estimadores MQO, embora possa fazer com que suas variâncias sejam grandes. Dada a correlação bastante substancial entre o valor de mercado da empresa e seus lucros, não é surpreendente que a variável $profits$ não acrescente nada para explicar os salários dos CEOs. Também, os lucros são uma medida de curto prazo de como a empresa está, enquanto seu valor de mercado é baseado no passado, presente, e futuro da lucratividade da empresa.

2) Primeiramente vamos salvar a base de dados solicitada em nosso Environment.

```
peso_bebes<- bwght
```

(i) Provavelmente $\beta_2 > 0$, dado que mais renda tipicamente implica em melhor nutrição da mãe e melhores cuidados pré-natais.

(ii) Por um lado, um aumento na renda geralmente aumenta o consumo de um bem, e *cigs* e $faminc$ podem ser positivamente correlacionados. Por outro lado, a renda familiar também é maior para famílias com mais educação e educação e tabagismo tendem a ser negativamente correlacionados.

Veremos agora qual das suposições está correta para esta amostra

```
cor(peso_bebes$cigs, peso_bebes$faminc)
```

```
## [1] -0.1730449
```

A correlação de amostra entre *cigs* e $faminc$ é de cerca de $-0,173$, indicando uma correlação negativa entre as variáveis.

(iii)

```r
sem_faminc <- lm(bwght ~ cigs, data = peso_bebes)
```

A função summary do R não nos retorna o tamanho da amostra. Portanto utilizaremos a função modelsummary, do pacote modelsummary.

```r
library(modelsummary)

modelsummary(sem_faminc) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position")
```

	Model 1
(Intercept)	119.772
	(0.572)
cigs	-0.514
	(0.090)
Num.Obs.	1388
R2	0.023
R2 Adj.	0.022
AIC	12276.9
BIC	12292.6
Log.Lik.	-6135.457
F	32.235

Temos então a equação

$$\widehat{bwght} = 119,772 - 0,514cigs, \quad n = 1.388, \quad R^2 = 0,227.$$

Agora, o modelo que considera a renda da família:

```r
com_faminc <- lm(bwght ~ cigs + faminc, data = peso_bebes)

modelsummary(com_faminc) %>%
  kableExtra::kable_styling(latex_options = "HOLD_position")
```

	Model 1
(Intercept)	116.974
	(1.049)
cigs	-0.463
	(0.092)
faminc	0.093
	(0.029)
Num.Obs.	1388
R2	0.030
R2 Adj.	0.028
AIC	12268.8
BIC	12289.8
Log.Lik.	-6130.414
F	21.274

O segundo modelo nos dá a equação

$$\widehat{bwght} = 116,974 - 0,463cigs + 0,093faminc, \quad n = 1.388, \quad R^2 = 0,030.$$

O efeito do tabagismo é ligeiramente menor quando a renda da família é adicionada à regressão, mas a diferença não é grande. Isso se deve ao fato de que cigs e faminc não são muito correlacionados, e o coeficiente de faminc é pequeno (A variável $faminc$ é medida em milhares, então \$10.000 a mais na renda de 1988 aumenta o peso previsto ao nascer em apenas 0,93 onças, que correspondem a 26,365 gramas).