GABARITO - LISTA 1

A NATUREZA DA ECONOMETRIA E DOS DADOS ECONÔMICOS

Mateus Cardoso

25/04/2021

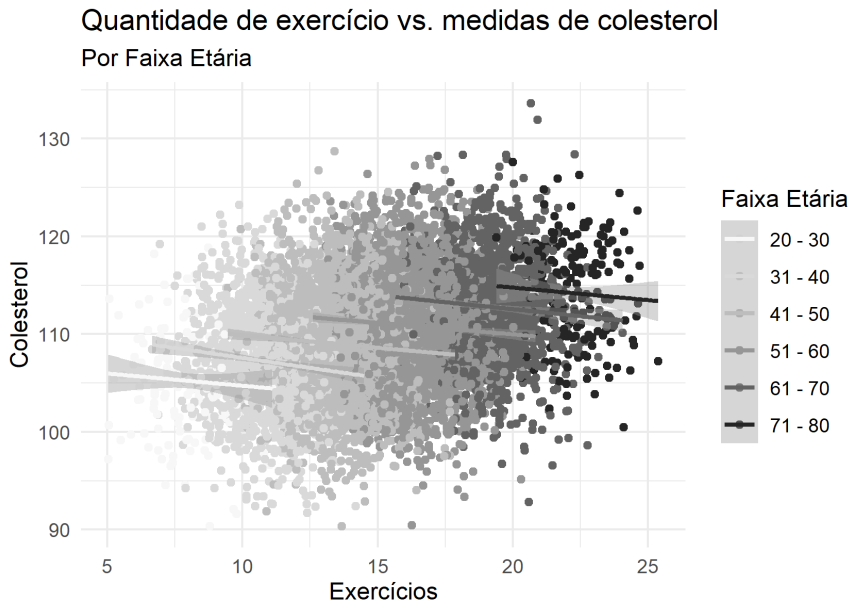
1) O experimento a ser projetado deve ter um ou mais grupos de tratamento e um grupo de controle; por exemplo, um tratamento poderia ser estudar por quatro horas, e o controle seria não estar estudando (sem tratamento). Os alunos seriam atribuídos aleatoriamente aos grupos de tratamento e controle, e o efeito causal das horas de estudo no desempenho seria estimado comparando as notas médias do meio do semestre de cada um dos grupos de tratamento às do grupo de controle. O maior obstáculo seria garantir que os alunos nos diferentes grupos de tratamento passem o número correto de horas estudando. Como você pode ter certeza de que os alunos do grupo de controle não estudem, já que isso pode comprometer suas notas? Como você pode ter certeza de que todos os alunos no grupo de tratamento realmente estudam por quatro horas?

2) Este experimento precisa dos mesmos ingredientes do experimento da pergunta anterior: grupos de tratamento e de controle, atribuição aleatória e um procedimento para analisar o dados experimentais resultantes. Aqui, existem dois níveis de tratamento: não usar cinto de segurança (grupo de controle) e usar o cinto de segurança (grupo tratado). Esses tratamentos devem ser aplicado durante um período de tempo especificado, como por exemplo no ano seguinte. O efeito do uso do cinto de segurança sobre fatalidades no trânsito pode ser estimado como a diferença entre as taxas de mortalidade no grupo de controle e no grupo de tratamento. Uma complicação para este estudo é garantir que os participantes sigam o tratamento (usar ou não o cinto de segurança). Porém, acima de tudo, este estudo levanta sérias preocupações éticas porque instrui os participantes a se envolverem em práticas perigosas conhecidas comportamento (não usar cinto de segurança).

3) Não faz sentido colocar a questão em termos de causalidade. Os economistas presumiriam que os alunos escolhem uma combinação de *estudo* e *trabalho* (e outras atividades, como assistir às aulas, lazer e dormir) com base no comportamento racional, maximizando a utilidade sujeita à restrição de que hajam somente 168 horas por semana. Podemos então usar métodos estatísticos para medir a associação entre *estudo* e *trabalho*, incluindo análise de regressão. Mas não estaríamos afirmando que uma variável "causa" a outra. Ambas são variáveis de escolha do aluno.

4) A conclusão errônea de que há uma relação positiva entre quantidade de exercícios e níveis de colesterol decorre do problema de tentar extrapolar resultados estatísticos como causais.

Neste caso, a quantidade de exercícios e o nível de colesterol estão sendo impactados por uma terceira variável: a idade. Se fizermos o mesmo gráfico, porém, levando em conta a idade dos participantes, veremos que a conclusão pode mudar.



Agora, a conclusão retirada do gráfico é que exercícios na verdade reduzem o colesterol.

Neste exemplo, a idade está agindo como uma variável confusora (*confounding variable*). Isto significa que a idade da pessoa impacta ao mesmo tempo a quantidade de exercícios e o nível de colesterol no corpo. Mais importante que a simples correlação entre variáveis é o modelo de estrutura causal que explique de que forma estas variáveis se relacionam.

Exercícios no R

1) Primeiro, devemos instalar o pacote que possui as bases de dados do Wooldridge. Fazemos isto com o comando `install.packages("wooldridge")`. Para abrir o pacote, utilizamos `library(wooldridge)`. Iremos utilizar também o pacote tidyverse.

```
library(wooldridge)
library(tidyverse)
```

Primeiro, precisamos checar a base de dados, quais são suas colunas, etc. Podemos fazer isto com o comando `glimpse()`.

```
glimpse(wage1)
```

```
## Rows: 526
## Columns: 24
## $ wage     <dbl> 3.10, 3.24, 3.00, 6.00, 5.30, 8.75, 11.25, 5.00, 3.60, 18.18,~
## $ educ     <int> 11, 12, 11, 8, 12, 16, 18, 12, 12, 17, 16, 13, 12, 12, 12, 16~
## $ exper    <int> 2, 22, 2, 44, 7, 9, 15, 5, 26, 22, 8, 3, 15, 18, 31, 14, 10, ~
## $ tenure   <int> 0, 2, 0, 28, 2, 8, 7, 3, 4, 21, 2, 0, 0, 3, 15, 0, 0, 10, 0, ~
## $ nonwhite <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ female   <int> 1, 1, 0, 0, 0, 0, 0, 1, 1, 0, 1, 1, 0, 0, 0, 0, 1, 1, 1, 1, 1~
## $ married  <int> 0, 1, 0, 1, 1, 1, 0, 0, 0, 1, 0, 0, 1, 0, 1, 1, 1, 0, 1, 1, 0~
## $ numdep   <int> 2, 3, 2, 0, 1, 0, 0, 0, 2, 0, 0, 0, 2, 0, 1, 1, 0, 0, 3, 0, 0~
## $ smsa     <int> 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ northcen <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ south    <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ west     <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ construc <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ ndurman  <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ trcommpu <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ trade    <int> 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ services <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ profserv <int> 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 0, 1, 0, 0, 0, 1, 0, 1, 0, 1, 1~
## $ profocc  <int> 0, 0, 0, 0, 0, 1, 1, 1, 1, 1, 1, 0, 0, 0, 1, 1, 0, 1, 0, 0, 1~
## $ clerocc  <int> 0, 0, 0, 1, 0, 0, 0, 0, 0, 0, 0, 1, 0, 0, 0, 0, 1, 0, 0, 0, 0~
## $ servocc  <int> 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0~
## $ lwage    <dbl> 1.1314021, 1.1755733, 1.0986123, 1.7917595, 1.6677068, 2.1690~
## $ expersq  <int> 4, 484, 4, 1936, 49, 81, 225, 25, 676, 484, 64, 9, 225, 324, ~
## $ tenursq  <int> 0, 4, 0, 784, 4, 64, 49, 9, 16, 441, 4, 0, 0, 9, 225, 0, 0, 1~
```

a) Calculamos a média com o comando `mean()`.

```
mean(wage1$educ)
```

```
## [1] 12.56274
```

Portanto, temos que a média de anos de educação é de 12,56 anos.

Para encontrar os menores e maiores valores, utilizamos as funções `min()` e `max()`.

```
min(wage1$educ)
```

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

```
max(wage1$educ)
```

```
## [1] 18
```

O menor valor para anos de estudo é de 0 e o maior é de 18.

Poderíamos também ter calculado todas estas informações apenas com o comando `summary()`.

```
summary(wage1$educ)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    0.00   12.00   12.00   12.56   14.00   18.00
```

b)

```
mean(wage1$wage)
```

```
## [1] 5.896103
```

O salário-hora médio é de aproximadamente US$5,90.

c) Nesta base de dados, *female* é uma variável binária que possui valor 1 para mulher e 0 para homem. Para contar, utilizamos o comando `count()`.

```
wage1 %>% count(female)
```

```
##   female   n
## 1      0 274
## 2      1 252
```

Temos em nossa amostra 274 homens e 252 mulheres.

2)

```
library(wooldridge)
library(tidyverse)
```

Primeiro, devemos observar a base de dados.

```
glimpse(countymurders)
```

```
## Rows: 37,349
## Columns: 20
## $ arrests     <int> 2, 3, 2, 7, 3, 1, 1, 2, 0, 5, 0, 1, 5, 3, 4, 5, 8, 4, 9, 8~
## $ countyid    <int> 1001, 1001, 1001, 1001, 1001, 1001, 1001, 1001, 1001, 1001~
## $ density     <dbl> 54.05000, 53.66000, 53.75000, 53.78000, 53.91000, 54.11000~
## $ popul       <int> 32216, 31984, 32036, 32056, 32128, 32248, 32888, 33264, 33~
## $ perc1019    <dbl> 20.63000, 20.19000, 19.66000, 19.10000, 18.54000, 18.06000~
## $ perc2029    <dbl> 15.28000, 15.55000, 15.73000, 15.88000, 15.92000, 15.87000~
## $ percblack   <dbl> 22.33000, 22.07000, 21.80000, 21.53000, 21.26000, 20.96000~
## $ percmale    <dbl> 40.25000, 40.36000, 40.42000, 40.47000, 40.51000, 40.45000~
## $ rpcincmaint <dbl> 167.670, 167.990, 166.630, 176.530, 166.250, 153.120, 151.~
## $ rpcpersinc  <dbl> 8780.80, 8232.80, 8327.61, 8545.55, 8965.16, 9254.02, 9885~
## $ rpcunemins  <dbl> 29.160, 43.920, 71.410, 72.220, 40.360, 44.540, 38.350, 35~
## $ year        <int> 1980, 1981, 1982, 1983, 1984, 1985, 1986, 1987, 1988, 1989~
## $ murders     <int> 2, 1, 3, 7, 2, 2, 4, 1, 0, 3, 1, 1, 1, 1, 1, 5, 7, 4, 6, 7~
## $ murdrate    <dbl> 0.6208096, 0.3126563, 0.9364465, 2.1836790, 0.6225100, 0.6~
## $ arrestrate  <dbl> 0.6208095, 0.9379690, 0.6242977, 2.1836790, 0.9337650, 0.3~
## $ statefips   <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1~
## $ countyfips  <int> 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 3, 3, 3~
## $ execs       <int> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
## $ lpopul      <dbl> 10.38022, 10.37299, 10.37462, 10.37524, 10.37748, 10.38121~
## $ execrate    <dbl> 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0~
```

a) Criaremos o objeto condados para armazenar a contagem. Antes, utilizaremos o comando `filter()` para criar um dataframe com apenas o ano de 1996.

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

condados<- countymurders_1996 %>%
  distinct(countyid) %>%
  count()

condados
```

```
##      n
## 1 2197
```

Para descobrir quantos condados tiveram 0 assassinatos, utilizaremos mais um filtro.

```
condados_sem_assassinatos<-countymurders_1996 %>%
  filter(murders == 0) %>%
  distinct(countyid) %>%
  count()
```

```
condados_sem_assassinatos
```

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

Para descobrir qual é a porcentagem de condados que tiveram zero execuções, criaremos o objeto zero_execucao, e então dividiremos pelo total de condados.

```
zero_execucao<- countymurders_1996 %>%
  filter(execs == 0) %>%
  distinct(countyid) %>%
  count()

zero_execucao
```

```
##      n
## 1 2166
```

São 2166 condados com zero execuções.

```
zero_execucao/condados
```

```
##           n
## 1 0.9858898
```

98,59% dos condados tiveram 0 execuções.

b) Utilizaremos a função `max()` para encontrar o maior número e `mean()` para encontrar a média.

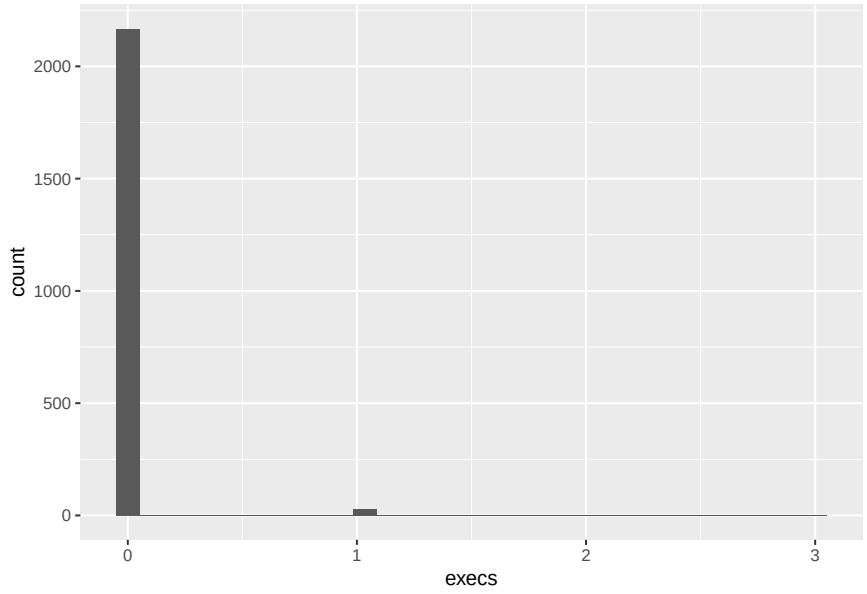
```
countymurders_1996 %>%
  summarise(max_murders = max(murders),
            max_execs = max(execs),
            mean_execs = mean(execs))
```

```
##   max_murders max_execs mean_execs
## 1        1403         3 0.01593081
```

O maior número de assassinatos é 1403. O número mais alto de execuções é 3.

A média de execuções é de 0,0159. Para descobrir o motivo do valor ser pequeno, podemos fazer um histograma para olhar a distribuição da frequência dos valores.

```
ggplot(countymurders_1996, aes(x = execs))+
  geom_histogram()
```



Logo vemos que o motivo de a média ser tão baixa é a quantidade de condados onde o número de execuções é igual a zero.

c) Para encontrar o coeficiente de correlação, utilizamos a função `cor()`.

```
cor(countymurders_1996$murders, countymurders_1996$execs)
```

```
## [1] 0.2095042
```

Há uma baixa correlação entre o número de assassinatos e o número de execuções, no valor de 0,209.