Monitoria de Econometria II

Jean Haendell

Carregando as bibliotecas

Uma boa prática no uso do R é a de sempre carregar as bibliotecas que serão utilizadas no início do script. Para a ap1, essas instalações não serão mandatórias.

Não obstante, para fins didáticos, irei carregar a biblioteca `wooldridge`, que contém diversas bases de dados. Caso você queira replicar os exemplos, basta: 1) instalar a referida biblioteca utilizando o comando `install.packages("wooldridge")`; 2) carregar a biblioteca utilizando o seguinte código:

```
library(wooldridge)
```

Caso você utilize o Rstudio no mesmo computador novamente, provavelmente não será necessário instalar a biblioteca, bastando apenas carregá-la com o `library()`.

Regressão Linear no R - MQO

Para fazer uma regressão linear no R, basta utilizar a função `lm()` com seus respectivos argumentos. Para um modelo com duas variáveis independentes, temos, por exemplo:

```
lm(data = base_de_dados, variavel_dependente ~ variavelindependente1 + variavelindepe
ndente2)
```

Obs.: `lm()` vem de *linear model*.

Nesse caso, foi considerado um modelo com duas variáveis independentes. `data = base_de_dados` indica qual base de dados estamos utilizando.

Exemplo : Fazendo a regressão do salário em relação à educação e experiência

Nesse exemplo, faremos a regressão do salário em educação e experiência, utilizando a base de dados “wage1”, presente na biblioteca wooldridge.

Podemos visualizar essa base de dados, bem como alguns valores, da seguinte maneira:

```
head(wage1)
```

```
##   wage educ exper tenure nonwhite female married numdep smsa northcen south
## 1 3.10   11     2      0        0      1       0      2    1         0     0
## 2 3.24   12    22      2        0      1       1      3    1         0     0
## 3 3.00   11     2      0        0      0       0      2    0         0     0
## 4 6.00    8    44     28        0      0       1      0    1         0     0
## 5 5.30   12     7      2        0      0       1      1    0         0     0
## 6 8.75   16     9      8        0      0       1      0    1         0     0
##   west construc ndurman trcommpu trade services profserv profocc clerocc
## 1    1        0       0        0     0        0        0       0       0
## 2    1        0       0        0     0        1        0       0       0
## 3    1        0       0        0     1        0        0       0       0
## 4    1        0       0        0     0        0        0       0       1
## 5    1        0       0        0     0        0        0       0       0
## 6    1        0       0        0     0        0        1       1       0
##   servocc    lwage expersq tenursq
## 1       0 1.131402       4       0
## 2       1 1.175573     484       4
## 3       0 1.098612       4       0
## 4       0 1.791759    1936     784
## 5       0 1.667707      49       4
## 6       0 2.169054      81      64
```

A funcão `head()` serve para termos nocão de quais variáveis estão presentes nessa base de dados, bem como ter conhecimento dos valores que essas variáveis recebem.

Nesse exemplo, estamos interessados nas variáveis salário, educação e experiência, que nessa base de dados estão como `lwage` (nesse caso, `lwage` significa $\log(wage)$. Isso é feito para que possamos ter uma resposta em percentual), `educ` e `exper` .

Isto é, o nosso modelo econométrico é:

$$\log(wage) = \beta_0 + \beta_1 educ + \beta_2 exper$$

Rodando esse modelo e salvando-o em um objeto chamado `modelo1` , temos

```
modelo1 <- lm(data = wage1, lwage ~ educ + exper)
```

ATENÇÃO:

tanto `<-` quanto `=` servem para atribuírmos valores a novos objetos (variáveis). Nesse caso, basicamente estamos criando uma nova variável (`modelo1`) e atribuindo o resultado do modelo estimado por `lm(data = wage1, lwage ~ educ + exper)` a ela.

Como dito, o modelo estimado foi salvo em `modelo1` . Agora, precisamos verificar os resultados. Para isso, podemos utilizar a função `summary()` . Em suma, essa função resume os resultados do modelo estimado e nos retorna o seguinte:

```
summary(modelo1)
```

```
##
## Call:
## lm(formula = lwage ~ educ + exper, data = wage1)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.05800 -0.30136 -0.04539  0.30601  1.44425
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 0.216854   0.108595   1.997   0.0464 *
## educ        0.097936   0.007622  12.848  < 2e-16 ***
## exper       0.010347   0.001555   6.653 7.24e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4614 on 523 degrees of freedom
## Multiple R-squared:  0.2493, Adjusted R-squared:  0.2465
## F-statistic: 86.86 on 2 and 523 DF,  p-value: < 2.2e-16
```

Interpretação dos resultados

Agora, vamos interpretar os resultados da estimação anterior.

1. *Interpretando os efeitos marginais*: O aumento de um ano de educação leva, em média, a um aumento de $9,79\%$ do salário, mantendo a experiência fixa. (A interpretação de $exper$ fica a cargo do estudante.)

2. *Realizando o teste de hipóteses*: Vamos realizar o teste de hipóteses para a variável $educ$. Nossa hipótese nula é de que um ano a mais de educação não tem efeito no salário. Nossa hipótese alternativa, nesse caso, será a de que a educação tem efeito no salário (embora seja convencional considerar como hipótese alternativa que esse efeito é positivo, aqui considerarei também a possibilidade de ser negativo). O teste de hipóteses é, então, construído da seguinte maneira:

$$\mathrm{H}_0 : \beta_1 = 0$$

$$\mathrm{H}_1 : \beta_1 \neq 0$$

Olhando para o resultado da regressão, vemos que o valor t em educ é de $12,489$. Se $|t| > 2$, então a variável é estatisticamente significante ao nível de 5%. Logo, $educ$ é estatisticamente significante. (a significância de $exper$ fica a cargo do estudante).

Atenção: Considerar uma variável estatisticamente significante se |t| > 2 é uma regra de bolso com base em um teste bicaudal com um nível de significância de 5%, em que o t crítico é próximo de 2. Mais explicações serão dadas na monitoria.

3. *Coeficiente de determinação - R^2*: O R^2 observado é de 0.2493, o que significa que $24,93\%$ da variação na variável dependente é causada pela variação nas variáveis independentes.

4. *Significância Global*: Para analisar a significância global de um modelo, olharemos para o $\mathrm{p\text{-}value}$ do modelo, que se encontra na última linha do nosso `summary(modelo1)`. O p-valor indica o menor nível de significância em que uma variável (ou um modelo, nesse caso) é estatisticamente significante. Quanto menor o nível de significância, mais difícil é para uma variável (ou modelo) ser estatisticamente

significante. Por isso, se observarmos no resultado que $p - value < 0.05$, sabemos que o modelo também será significante ao nível de 5%, que é o mais utilizado. Por isso, nesse caso, rejeitamos a hipótese nula de que o modelo **não** é estatisticamente significante.

Em suma: o teste de hipóteses se dá por:

$$\mathrm{H}_0 : \beta_j = 0, j = 1, 2$$

$$\mathrm{H}_1 : \beta_j \neq 0, \text{para qualquer } j = 1, 2$$

.

Observamos, nesse modelo, um p-valor de $2.2\text{e-}16$, um número beeeem inferior a 0.05. Ou seja, o modelo é globalmente significante.

$$\text{p-value} = 2.2\text{e-}16 < 0.05 \implies \text{rejeita a hipótese nula.}$$

Vamos a um exemplo da lista?

Questão COCA-

Nessa questão, precisamos carregar a base de dados `coca` , que não pertence à biblioteca do Wooldridge. No meu caso, especificarei o caminho do arquivo no computador que estou utilizando. Entretanto, para carregar, basta baixar a base no sigaa, ir em “file”, no canto superior esquerdo, e depois em “open file”.

No meu caso, o caminho do arquivo é o seguinte:

```
load("/Users/emilia.franca/Downloads/coca.RData")
```

Com a base de dados carregada, podemos ver as variáveis que estão presentes nela usando nossa conhecida função `head()` :

```
head(coca)
```

```
##       PRICE QUAL      QUANT TREND
## 1 57.50000 62.5 1000.00000     1
## 2 77.16049 62.5  453.60000     1
## 3 77.60141 19.0   28.35000     1
## 4 84.65608 19.0   14.17500     1
## 5 77.60141 19.0    7.08750     1
## 6 91.71076 19.0    3.54375     1
```

O modelo dado pela questão 1 é construído da seguinte maneira:

$$price = \beta_1 + \beta_2 quant + \beta_3 qual + \beta_4 trend + e$$

Trazendo para a linguagem do R, temos

```
modelococa = lm(data = coca, PRICE ~ QUANT + QUAL + TREND)
```

Após ter definido e salvo o modelo, podemos ver os resultados da estimação:

```
summary(modelococa)
```

```
##
## Call:
## lm(formula = PRICE ~ QUANT + QUAL + TREND, data = coca)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -43.479 -12.014  -3.743  13.969  43.753
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 90.84669    8.58025  10.588 1.39e-14 ***
## QUANT       -0.05997    0.01018  -5.892 2.85e-07 ***
## QUAL         0.11621    0.20326   0.572   0.5700
## TREND       -2.35458    1.38612  -1.699   0.0954 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 20.06 on 52 degrees of freedom
## Multiple R-squared:  0.5097, Adjusted R-squared:  0.4814
## F-statistic: 18.02 on 3 and 52 DF,  p-value: 3.806e-08
```

Podemos, então, começar a responder a questão.

a) Estime o modelo acima e interprete os resultados.

- O R^2 do modelo é de 0.5097, o que significa que 50,97% da variação no preço é explicada pela variação nas variáveis independentes.

- Para analisar a significância dos coeficientes, podemos construir um teste de hipóteses como

$$\mathrm{H}_0 : \beta_j = 0, j = 2, 3, 4$$

$$\mathrm{H}_1 : \beta_j \neq 0,\ j = 2, 3, 4$$

  Nesse caso, rejeitamos a hipótese nula ao nível de 5% se $|t_{\hat{\beta}_j}| > 2$.

1. **QUANT**: o valor de t é de -5,892. Em módulo, esse valor é maior que 2. Portanto, QUANT é estatisticamente significante ao nível de 5%.
2. **QUAL**: o valor de t é de 0,572. Em módulo, esse valor é menor que 2. Portanto, falhamos em rejeitar a hipótese nula ao nível de 5%.
3. **TREND**: o valor de t é de -1,699. Em módulo, esse valor é menor que 2. Portanto, falhamos em rejeitar a hipótese nula ao nível de 5%.

- Para testar a significância global do modelo, construímos o seguinte teste de hipóteses:

$$\mathrm{H}_0 : \beta_j = 0, j = 2, 3, 4$$

$$\mathrm{H}_1 : \beta_j \neq 0, \text{para qualquer } j = 2, 3, 4$$

  Diremos que o modelo é globalmente significante se o p-valor global for inferior a 0.05. Observamos

que o p-valor deste modelo é igual a 3.806e-08. Este valor é menor que 0.05, portanto, o modelo é globalmente significante.

- Passemos agora para a análise dos efeitos marginais das variáveis explicativas do modelo:

1. **QUANT**: Um aumento de uma unidade na quantidade leva, em média, a uma diminuição de 0,06 no preço, tudo o mais constante.
2. **QUAL**: Um aumento de uma unidade na qualidade leva, em média, a um aumento de 0,12 no preço, tudo o mais constante.
3. **TREND**: Um aumento de uma unidade no tempo leva, em média, a uma diminuição de 2,35 no preço, tudo o mais constante.

- Para analisar a coerência teórica, lançaremos mão de nossos conhecimentos econômicos.

1. **QUANT**: Faz sentido de que o aumento da quantidade ofertada leve a uma diminuição dos preços, tudo o mais constante. Logo, é coerente. Outra explicação pode ser a fornecida no item b)
2. **QUAL**: Faz sentido que o aumento da qualidade da cocaína eleve seu preço, tudo o mais constante. Logo, é coerente.
3. **TREND**: A interpretação aqui é mais dúbia: o passar dos anos faz com que o preço diminua, tudo o mais constante. Faz-se necessário conhecer mais do fenômeno para chegar a uma conclusão, mas vale lembrar que essa variável não é significante para explicar preço nesse modelo. Mais informações no item d)

b) **Argumenta-se que quanto maior o número de vendas, maior a possibilidade de se ser preso; assim, os vendedores aceitam preços mais baixos se puderem vender quantidades maiores. Estabeleça H_0 e H_A apropriadas para testar essa hipótese**

- Nesse caso, queremos realizar o teste de hipóteses para a variável QUANT. Como o coeficiente que a acompanha no modelo é o β_2, o teste de hipóteses se dá pelo seguinte:

$$\mathrm{H}_0 : \beta_2 = 0$$

$$\mathrm{H}_1 : \beta_2 \neq 0$$

O teste se dá da mesma forma que no primeiro item. Lá, vimos que a quantidade é estatisticamente significante, porque seu valor t, em módulo, é maior que 2. Portanto, rejeitamos a hipótese nula, que na prática dizia que a quantidade não tem impacto no preço.

c) **Teste a hipótese de que a qualidade da cocaína não tem qualquer influência no preço contra a alternativa de que paga-se um prêmio pela cocaína de melhor qualidade.**

- Nesse caso, queremos realizar o teste de hipóteses para a variável QUAL. Como o coeficiente que a acompanha no modelo é o β_3, o teste de hipóteses se dá pelo seguinte:

$$\mathrm{H}_0 : \beta_3 = 0$$

$$\mathrm{H}_1 : \beta_3 \neq 0$$

Como vimos, essa variável não é estatisticamente significante. Portanto, falhamos em rejeitar a hipótese nula, que na pratica diz que a qualidade da cocaína não influencia no preço.

Questão Sleep

```
load("/Users/emilia.franca/Downloads/sleep75.RData")
head(sleep75)
```

```
##   age black case clerical construc educ earns74 gdhlth inlf leis1 leis2 leis3
## 1  32     0    1        0        0   12       0      0    1  3529  3479  3479
## 2  31     0    2        0        0   14    9500      1    1  2140  2140  2140
## 3  44     0    3        0        0   17   42500      1    1  4595  4505  4227
## 4  30     0    4        0        0   12   42500      1    1  3211  3211  3211
## 5  64     0    5        0        0   14    2500      1    1  4052  4007  4007
## 6  41     0    6        0        0   12       0      1    1  4812  4797  4797
##   smsa  lhrwage   lothinc male marr prot rlxall selfe sleep slpnaps south
## 1    0 1.955861 10.075380    1    1    1   3163     0  3113    3163     0
## 2    0 0.357674  0.000000    1    0    1   2920     1  2920    2920     1
## 3    1 3.021887  0.000000    1    1    0   3038     1  2670    2760     0
## 4    0 2.263844  0.000000    0    1    1   3083     1  3083    3083     0
## 5    0 1.011601  9.328213    1    1    1   3493     0  3448    3493     0
## 6    0 2.957511 10.657280    1    1    1   4078     0  4063    4078     0
##   spsepay spwrk75 totwrk union worknrm workscnd exper yngkid yrsmarr    hrwage
## 1       0       0   3438     0    3438        0    14      0      13  7.070004
## 2       0       0   5020     0    5020        0    11      0       0  1.429999
## 3   20000       1   2815     0    2815        0    21      0       0 20.529997
## 4    5000       1   3786     0    3786        0    12      0      12  9.619998
## 5    2400       1   2580     0    2580        0    44      0      33  2.750000
## 6       0       0   1205     0       0     1205    23      0      23 19.249998
##   agesq
## 1  1024
## 2   961
## 3  1936
## 4   900
## 5  4096
## 6  1681
```

O modelo a ser estimado nessa questão é

$$\text{dormir} = \beta_0 + \beta_1 \text{trabtot} + u$$

. Ou seja, queremos saber se o número de horas trabalhadas influencia no sono. Vamos aos itens.

a) Reporte seus resultados na forma de equação, juntamente com o número de observações e R^2. O que o intercepto significa nessa equação?

Primeiro, precisamos passar o modelo para a linguagem do R e estimá-lo.

```
modelosleep1= lm(data = sleep75, sleep ~ totwrk)
summary(modelosleep1)
```

```
##
## Call:
## lm(formula = sleep ~ totwrk, data = sleep75)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2429.94  -240.25     4.91   250.53  1339.72
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) 3586.37695   38.91243  92.165   <2e-16 ***
## totwrk        -0.15075    0.01674  -9.005   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 421.1 on 704 degrees of freedom
## Multiple R-squared:  0.1033, Adjusted R-squared:  0.102
## F-statistic: 81.09 on 1 and 704 DF,  p-value: < 2.2e-16
```

Agora, podemos reportar os resultados:

- Nessa estimação, verificamos que um aumento de 1 minuto de trabalho durante a semana leva a uma diminuição de 0,15 minutos dormidos durante a semana.
- Reportando os resultados em forma de equação, temos

$$\hat{\text{dormir}} = 3586 - 0,15 trabtot$$

- O número de observações é de 706 (degrees of freedom + números de parâmetros do modelo).
- O intercepto significa o total de minutos dormidos na semana caso o número de minutos trabalhados na semana seja igual a zero. Isto é, se $totwrk = 0$, então o número de minutos dormidos na semana será aproximadamente 3586, o que dá em torno de 60 horas.

b) **Se trabtot aumenta em duas horas, em quanto tempo se estima que dormir irá cair? Você acha que isso é um efeito grande?**

Se houver um aumento de duas horas em trabtot, temos um aumento de 120 minutos, o que provoca uma diminuição de $120 \cdot 0,15$ em sleep. Ou seja, cai em 18 o número de minutos dormidos na semana. Podemos dizer que este é um efeito pequeno para um aumento de 2 horas de trabalho.

Questão Tuna

```
load("/Users/emilia.franca/Downloads/tuna.RData")
head(tuna)
```

```
##   sal1 apr1 apr2 apr3 disp dispad
## 1 6439 0.66 0.82 0.79    1      0
## 2 3329 0.62 0.80 0.59    1      0
## 3 3415 0.62 0.77 0.63    0      0
## 4 2909 0.62 0.66 0.81    0      0
## 5 2598 0.63 0.65 0.81    0      0
## 6 3773 0.69 0.63 0.80    0      0
```

a) Estime, por mínimos quadrados, o modelo log-linear

$$\ln(SAL1) = \beta_1 + \beta_2 \text{APR1} + \beta_3 \text{APR2} + \beta_4 \text{APR3} + \beta_5 \text{DISP} + \beta_6 \text{DispAd} + e$$

Passando o modelo para a linguagem do R, temos

```
modelotuna= lm(data = tuna, log(sal1) ~ apr1 + apr2 + apr3 + disp + dispad)
summary(modelotuna)
```

```
##
## Call:
## lm(formula = log(sal1) ~ apr1 + apr2 + apr3 + disp + dispad,
##     data = tuna)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.70001 -0.21573 -0.03785  0.26241  0.74457
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.9848     0.6464  13.900  < 2e-16 ***
## apr1         -3.7463     0.5765  -6.498 5.17e-08 ***
## apr2          1.1495     0.4486   2.562 0.013742 *
## apr3          1.2880     0.6053   2.128 0.038739 *
## disp          0.4237     0.1052   4.028 0.000209 ***
## dispad        1.4313     0.1562   9.165 6.04e-12 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.3397 on 46 degrees of freedom
## Multiple R-squared:  0.8428, Adjusted R-squared:  0.8257
## F-statistic: 49.33 on 5 and 46 DF,  p-value: < 2.2e-16
```

b) Discuta e interprete as estimativas de β_2, β_3 e $\beta_4$$

- β_2: o valor estimado de β_2 indica que o aumento de 1 unidade no preço da marca 1 faz com que as vendas da marca 1 caiam em 374,63%. Na prática, toda a venda da marca 1 é eliminada.
- β_3: o valor estimado de β_3 indica que o aumento de 1 unidade no preço da marca 2 faz com que as vendas da marca 1 aumentem em 114,95%. Na prática, a venda da marca 1 mais que dobra.
- β_4: o valor estimado de β_4 indica que o aumento de 1 unidade no preço da marca 3 faz com que as vendas da marca 1 aumentem em 128,80%. Na prática, a venda da marca 1 mais que dobra.

c) Os sinais e as grandezas relativas de β_5 e β_6 são coerentes com a lógica econômica?

O sinal positivo tanto na estimativa do coeficiente em disp e dispad faz sentido: um aumento na propaganda leva a aumentos nas vendas. Além disso, faz sentido que dispad seja maior que disp, pois a primeira envolve propaganda em duas frentes, e a segunda em apenas uma.

d) No nível $\alpha = 0,05$ de significância...

- O primeiro teste proposto é

$$\text{H}_0 : \beta_5 = 0$$

$$H_1 : \beta_5 \neq 0$$

O valor t de disp é maior que 2. Logo, rejeitamos a hipótese nula.

- O segundo teste é

$$H_0 : \beta_6 = 0$$

$$H_1 : \beta_6 \neq 0$$

O valor t de dispad é maior que 2. Logo, rejeitamos a hipótese nula.

- O terceiro teste é

$$H_0 : \beta_5 = 0, \beta_6 = 0$$

$$H_1 : \beta_5 \text{ ou } \beta_6 \neq 0$$

Neste caso, podemos proceder da seguinte maneira:

```
# Ajustando o modelo completo
modelo_completo <- lm(data = tuna, log(sal1) ~ apr1 + apr2 + apr3 + disp + dispad)

# Ajustando o modelo reduzido
modelo_reduzido <- lm(data = tuna, log(sal1) ~ apr1 + apr2 + apr3)

# Realizando o teste de hipóteses
resultado <- anova(modelo_reduzido, modelo_completo)
print(resultado)
```

```
## Analysis of Variance Table
##
## Model 1: log(sal1) ~ apr1 + apr2 + apr3
## Model 2: log(sal1) ~ apr1 + apr2 + apr3 + disp + dispad
##   Res.Df     RSS Df Sum of Sq      F    Pr(>F)
## 1     48 15.0023
## 2     46  5.3073  2     9.695 42.015 4.172e-11 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nesse teste, se o p-valor for menor que 0.05, rejeitamos a hipótese nula, o que acontece nesse caso, pois o p-valor é de 4.172e-11

- O quarto teste é

$$H_0 : \beta_6 \leq \beta_5$$

$$H_1 : \beta_6 > \beta_5$$

Nesse caso, faremos

```
# Carregue o pacote necessário
library(car)
```

```
## Loading required package: carData
```

```
# Realize o teste de hipótese
test <- linearHypothesis(modelotuna, "dispad - disp = 0")
print(test)
```

```
## Linear hypothesis test
##
## Hypothesis:
## - disp  + dispad = 0
##
## Model 1: restricted model
## Model 2: log(sal1) ~ apr1 + apr2 + apr3 + disp + dispad
##
##   Res.Df     RSS Df Sum of Sq      F   Pr(>F)
## 1     47 10.7332
## 2     46  5.3073  1    5.4259 47.029 1.49e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Nesse caso, se o p-valor for menor que 0,05, rejeitamos a hipótese nula.