GABARITO - LISTA 4

MODELO DE REGRESSÃO LINEAR MÚLTIPLA: INFERÊNCIA

Mateus Cardoso

30/06/2021

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

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

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

A hipótese RLM.1 descreve o relacionamento populacional que esperamos estimar, e explicitamente especifica β_j − os efeitos populacionais ceteris paribus da x_j sobre y − como os parâmetros de interesse.

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.

Essa hipótese de amostragem aleatória significa que possuímos dados que podem ser usados para estimarmos a β_j, e que os dados foram selecionados como representativos da população descrita 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.

Sempre que temos uma amostra de dados, precisamos saber se podemos usar os dados para calcularmos as estimativas MQO, a $\hat{\beta}_j$. Essa é a função da Hipótese RLM.3: se tivermos variação amostral em cada variável independente e nenhum relacionamento linear exato entre as variáveis independentes, poderemos calcular a $\hat{\beta}_j$.

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

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

Como já discutimos, presumindo que as não observáveis são, na média, não relacionadas com as variáveis explicativas, é vital para se derivar a primeira propriedade estatística de cada

estimador MQO: sua ausência de viés no parâmetro populacional correspondente. É claro, todas as hipóteses anteriores são usadas para demonstrar a ausência de viés.

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

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

Comparada com a Hipótese RLM.4, a hipótese de homoscedasticidade é de importância secundária; particularmente, a Hipótese RLM.5 não tem influência na ausência de viés das $\hat{\beta}_j$. Ainda assim, a homoscedasticidade tem duas implicações importantes: (1) Podemos derivar fórmulas das variâncias amostrais cujos componentes são fáceis de serem caracterizados; (2) Podemos concluir, sob as hipóteses RLM.1 até a RLM.5 de Gauss-Markov, que os estimadores MQO têm a menor variância entre todos os estimadores lineares, não viesados.

Hipótese RLM.6 (Normalidade): O erro populacional u é independente das variáveis explicativas $x_1, x_2, ..., x_k$, e é normalmente distribuído com média zero e variância σ^2 : $u \sim Normal(0, \sigma^2)$.

Adicionamos a Hipótese RLM.6 à distribuição amostral exata das estatísticas t e estatísticas F, de forma que possamos realizar testes de hipóteses. No próximo capítulo, veremos que a RLM.6 pode ser eliminada se tivermos uma amostra de tamanho razoavelmente grande. A Hipótese RLM.6 realmente implica uma propriedade de eficiência mais forte dos MQO: os estimadores MQO têm a menor variância entre todos os estimadores não viesados; o grupo de comparação não estará mais restrito a estimadores lineares na $(y_i : i = 1, 2, ..., n)$.

2) (i) e (iii) geralmente fazem com que a estatística t não tenha uma distribuição t sob H_0.

A homoscedasticidade é uma das premissas do MRLC. Uma variável importante omitida viola a Hipótese MLR.3.

As Hipóteses do MRLC não contêm nenhuma afirmação acerca das correlações amostrais entre as variáveis independentes, exceto para descartar o caso em que a correlação é igual a um.

3) (i) $H_0 : \beta_3 = 0$ e $H_1 : \beta_3 \neq 0$.

(ii) Tudo mais constante, uma população maior aumenta a demanda por moradias para aluguel, o que deve aumentar as rendas. A demanda por habitação geral é maior quando a renda média é maior, elevando o custo da habitação, incluindo as taxas de aluguel.

(iii) A interpretação de um modelo Log-Log (Elasticidade) se dá da seguinte forma: quando x aumenta em 1%, espera-se uma variação de $1 * \beta_k\%$ em y. Então, como temos $\beta_1 = 0,066$, quando *pop* aumentar em 1%, esperamos uma variação de 0,066% em *rent*. Uma afirmação correta para o postulado é: "um aumento de 10% na população aumenta o valor do aluguel($rent$) em $0,066 * 10 = 0,66\%$."

(iv) Com GL (Graus de Liberdade)[1] $= 64 - 3 - 1 = 60$, o valor crítico de 1% para um teste bicaudal é 2,660[2]. A estatística t é $\frac{\hat{\beta}_3 - \beta_{H_0}}{ep(\hat{\beta}_3)} = \frac{0{,}0056 - 0}{0{,}0017} = 3,29$, o que está bem acima do valor crítico. Então, β_3 é estatisticamente diferente de zero ao nível de 1%, ou seja, rejeita-se a hipótese nula.

4) (i) Para a aproximação normal padronizada, temos que o valor crítico de t é igual a $1,96$ para o nível de confiança de 95%. Temos que o valor instervalo de confiança para $hsGPA$ é $0,412 \pm 1,96 * 0,094 = 0,412 \pm 0,184$. Portanto, $IC(hsGPA, 95\%) = [0,228; 0,596]$.

(ii) Não, pois o valor $0,4$ está dentro do intervalo de confiança de 95%.

(iii) Sim, pois 1 está fora do intervalo de confiança de 95%.

Exercícios no R

1) (i) Tratando-se de um modelo Nível-Log, a interpretação do coeficiente β_1 é (*ceteris paribus*): quando x aumenta em 1%, espera-se uma variação de $\frac{\beta_1}{100}$ unidades em y. Neste caso, se as despesas do Candidato A subirem em 1%, espera-se que os votos do Candidato A aumentem em 0,01 pontos percentuais.

(ii) A hipótese nula é $H_0 : \beta_2 = -\beta_1$. De forma equivalente, podemos reescrever $H_0 : \beta_1 + \beta_2 = 0$.

(iii) Primeiro, vamos abrir o pacote tidyverse e salvar a base VOTE1 do Wooldridge em nosso environment do R.

```
library(tidyverse)

votos <- wooldridge::vote1
```

Agora, vamos visualizar nossa base de dados:

```
votos %>% glimpse()
```

[1] $GL = n - k - 1$, onde n é o tamanho da amostra e k é o número de parâmetros de **INCLINAÇÃO**(Ou seja, o núemro de variáveis explicativas usadas no modelo).

[2] Na tabela t de Student, encontre o ponto de encontro entre a coluna de 99% para o teste bicaudal e a linha de 60 graus de liberdade.

```
## Rows: 173
## Columns: 10
## $ state    <chr> "AL", "AK", "AZ", "AZ", "AR", "AR", "CA", "CA", "CA", "CA", "~
## $ district <int> 7, 1, 2, 3, 3, 4, 2, 3, 5, 6, 7, 11, 12, 16, 19, 23, 24, 27, ~
## $ democA   <int> 1, 0, 1, 0, 0, 1, 0, 1, 1, 1, 1, 1, 0, 1, 0, 1, 1, 1, 1, 1, 1~
## $ voteA    <int> 68, 62, 73, 69, 75, 69, 59, 71, 76, 73, 68, 71, 52, 79, 50, 6~
## $ expendA  <dbl> 328.296, 626.377, 99.607, 319.690, 159.221, 570.155, 696.748,~
## $ expendB  <dbl> 8.737, 402.477, 3.065, 26.281, 60.054, 21.393, 193.915, 7.695~
## $ prtystrA <int> 41, 60, 55, 64, 66, 46, 58, 49, 71, 64, 53, 58, 49, 54, 54, 5~
## $ lexpendA <dbl> 5.793916, 6.439952, 4.601233, 5.767352, 5.070293, 6.345908, 6~
## $ lexpendB <dbl> 2.167567, 5.997638, 1.120048, 3.268846, 4.095244, 3.063064, 5~
## $ shareA   <dbl> 97.40767, 60.88104, 97.01476, 92.40370, 72.61247, 96.38355, 7~
```

Podemos estimar nosso modelo:

```
mod <- lm(voteA ~ lexpendA + lexpendB + prtystrA, data = votos)

summary(mod)
```

```
##
## Call:
## lm(formula = voteA ~ lexpendA + lexpendB + prtystrA, data = votos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3968  -5.4174  -0.8679   4.9551  26.0660
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.07893    3.92631   11.48   <2e-16 ***
## lexpendA     6.08332    0.38215   15.92   <2e-16 ***
## lexpendB    -6.61542    0.37882  -17.46   <2e-16 ***
## prtystrA     0.15196    0.06202    2.45   0.0153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.712 on 169 degrees of freedom
## Multiple R-squared:  0.7926, Adjusted R-squared:  0.7889
## F-statistic: 215.2 on 3 and 169 DF,  p-value: < 2.2e-16
```

Nossa equação estimada é

$$\widehat{voteA} = 45,079 + 6,083 lexpendA - 6,615 lexpendB + 0,152 prtystrA$$

$$(3,926) \qquad (0,382) \qquad (0,379) \qquad (0,062)$$

$$n = 173, \quad R^2 = 0,793.$$

Tanto os gastos do Candidato A quanto o Candidato B afetam os resultados, pois as estimativas são significantes com $p - valor$ próximo de zero (***).

Mesmo que os coeficientes de $log(expendA)$ e $log(expendB)$ possuam magnitudes similares e sinal oposto, não podemos testar a hipótese nula formulada anteriormente pois não temos o erro padão de $\hat{\beta}_1 + \hat{\beta}_2$.

(iv) Para tanto, precisaremos de um parâmetro $\theta_1 = \beta_1 + \beta_2$. Adicionando na equação original e rearranjando, temos

$$\widehat{voteA} = \beta_0 + \theta_1 lexpendA + \beta_2[lexpendB - lexpendA] + \beta_3 prtystrA + u.$$

Podemos estimar nosso novo modelo, depois de criar a variável de $lexpendB - lexpendA$:

```
votos %>%
  mutate(lexpB_lexpA = lexpendB-lexpendA) -> votos

mod2 <- lm(voteA ~ lexpendA + lexpB_lexpA + prtystrA, data = votos)

summary(mod2)
```

```
##
## Call:
## lm(formula = voteA ~ lexpendA + lexpB_lexpA + prtystrA, data = votos)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -20.3968  -5.4174  -0.8679   4.9551  26.0660
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) 45.07893    3.92631  11.481   <2e-16 ***
## lexpendA    -0.53210    0.53309  -0.998   0.3196
## lexpB_lexpA -6.61542    0.37882 -17.463   <2e-16 ***
## prtystrA     0.15196    0.06202   2.450   0.0153 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 7.712 on 169 degrees of freedom
## Multiple R-squared:  0.7926, Adjusted R-squared:  0.7889
## F-statistic: 215.2 on 3 and 169 DF,  p-value: < 2.2e-16
```

De nossa equação estimada, temos $\hat{\theta}_1 = -0,532$ e $ep(\hat{\theta}_1) = 0,533$. A estatística t para a hipótese formulada em (ii) é $\frac{-0,532}{0,533} \approx -1$. Dessa forma, não é possível rejeitar H_0.

2) De início, vamos armazenar a base 401KSUBS no R.

```
ksubs<- wooldridge::k401ksubs
```

(i) Para encontrar quantas residências com apenas uma pessoa existem no conjunto de dados, basta aplicar um filtro e contar.

```
ksubs %>%
  filter(fsize == 1) %>%
  count()
```

```
##      n
## 1 2017
```

Temos 2017 residências com apenas uma pessoa no conjunto de dados.

(ii) Vamos estimar o modelo, após salvar a base com os filtros necessários.

```
ksubs %>%
  filter(fsize == 1) -> ksubs2

mod3 <- lm(nettfa ~ inc + age, data = ksubs2)

summary(mod3)
```

```
##
## Call:
## lm(formula = nettfa ~ inc + age, data = ksubs2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -179.95  -14.16   -3.42    6.03 1113.94
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -43.03981    4.08039 -10.548   <2e-16 ***
## inc           0.79932    0.05973  13.382   <2e-16 ***
## age           0.84266    0.09202   9.158   <2e-16 ***
## ---
```

```
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 44.68 on 2014 degrees of freedom
## Multiple R-squared:  0.1193, Adjusted R-squared:  0.1185
## F-statistic: 136.5 on 2 and 2014 DF,  p-value: < 2.2e-16
```

O coeficiente em *inc* indica que mais um dólar em renda (mantendo *age* fixo) resulta em cerca de 80 centavos a mais no $nettfa$ previsto; nenhuma surpresa nisso. O coeficiente de idade significa que, mantendo a renda constante, se uma pessoa ficar um ano mais velha, prevê-se que a sua $nettfa$ aumente cerca de \$843 (lembrando que $nettfa$ é medida em milhares de dólares). Novamente, nenhuma surpresa.

(iii) O intercepto nos indica a $nettfa$ prevista quando $age = 0$ e $inc = 0$, portanto, não é interessante para a análise.

(iv) A estatística t é $\frac{0,843-1}{0,092} \approx -1,71$. Contra a hipótese alternativa $H_1 : \beta_2 < 1$, ao nível de significância de 1% temos o p-valor, que é a probabilidade de $|T| > 2,567$ como aproximadamente 0,044 (encontramos este valor com o comando `pnorm(-1.71)`). Então, podemos rejeitar a hipótese nula a 5%, mas não a 1% (contra a hipótese unicaudal).

(v) Fazendo a regressão simples:

```
mod4 <- lm(nettfa ~ inc, data = ksubs2)

summary(mod4)
```

```
##
## Call:
## lm(formula = nettfa ~ inc, data = ksubs2)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -185.12  -12.85   -4.85    1.78 1112.66
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.5709     2.0607   -5.13 3.18e-07 ***
## inc           0.8207     0.0609   13.48  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 45.59 on 2015 degrees of freedom
## Multiple R-squared:  0.08267,    Adjusted R-squared:  0.08222
## F-statistic: 181.6 on 1 and 2015 DF,  p-value: < 2.2e-16
```

O coeficiente de inclinação em inc na regressão simples é de cerca de 0,821, o que não é muito diferente de 0,799 obtido na parte (ii). Acontece que a correlação entre *inc* e *age* na amostra de pessoas solteiras é apenas cerca de 0,039,

```
cor(ksubs2$age, ksubs2$inc)
```

```
## [1] 0.03905864
```

o que ajuda a explicar por que o simples e as estimativas de regressão múltipla não são muito diferentes.