Regressão para pesquisas sociais

Dia 2 - Exercício: RLS e RLM

Thiago Cordeiro Almeida

September 17, 2025

1. Introdução

Vamos trabalhar neste documento com dados do Censo Demográfico de 2010 e Cadastro Único.

Estamos interessados em responder à seguinte pergunta: **Quais os fatores associados ao acesso ao esgotamento sanitário de populações vulneráveis no Brasil?**

Vamos seguir as etapas descritas nos slides para cada modelo a ser estimado.

2. Importação dos dados

Ao longo dessas análises, usaremos a gramática de programação `tidyverse`. Caso você ainda não esteja familiarizado/a com essa forma de programação, sugiro a leitura do material da Curso-R que deixei nas referências do curso, há um capítulo sobre isso. Eles, com certeza, te convencerão sobre a potencialidade dessa forma de programção.

Dentro do `tidyverse`, há distintas bibliotecas. Usaremos duas chamadas `dplyr` – para manipulação dos dados – e `ggplot2` – para gráficos.

```
# ajustes gerais
options(scipen = 9999999)
rm(list = ls())

## pacotes necessarios
# instalando pacote de gerenciador de pacotes, pacman
ifelse(!require(pacman),install.packages("pacman"),require(pacman))
```

```
[1] TRUE
```

```r
p_load(tidyverse, here, skimr) # importando pacote que usaremos, tidyverse

# importando dados
# setwd() # CONFIGURE O SEU DIRETORIO DE TRABALHO

# o comando 'here()' faz com que trabalhemos onde esta o nosso codigo ou projeto

diretorio <- file.path(here(),"dia 2","pratica")
# diretorio <- file.path(here())

dados <- read_csv2(file.path(diretorio,"dados_acesso_esgoto.csv"))
```

3. Exploratória da base de dados

Vamos fazer uma breve exploração dos dados, para compreender como a nossa base de dados está organizada.

A começar, vamos usar o pacote que o Guilherme nos recomendou na última aula, chamado `skimr`[1].

```r
dados |>
  skim()
```

Table 1: Data summary

Name	dados
Number of rows	10410
Number of columns	9
Column type frequency:	
character	3
numeric	6
Group variables	None

Variable type: character

skim_variable	n_missing	complete_rate	min	max	empty	n_unique	whitespace
regiao	0	1	3	12	0	5	0
pop_faixas	0	1	9	19	0	6	0
municipio_rural	0	1	5	6	0	2	0

Variable type: numeric

[1]Para mais informações sobre ele, você pode acessar esse texto da Curso-R.

skim_variable	n_missing	complete_rate	mean	sd	p0	p25	p50	p75	p100	hist
ano	0	1	2016.51	1.11	2015.00	2016.00	2017.00	2017.00	2018	
em	0	1	32.15	12.89	0.00	22.73	31.15	40.54	100	
acesso_esgoto	0	1	34.09	32.37	0.03	3.37	23.99	60.63	100	
pbf	0	1	64.08	18.10	0.00	49.81	67.53	79.20	100	
desocupados	0	1	78.57	9.73	16.79	72.52	78.50	84.89	100	
informalidade	0	1	58.12	20.96	0.00	45.60	60.00	72.77	100	

Com base no que este código nos retorna, podemos ter uma visão geral da distribuição das variáveis, segundo algumas informações principais, como medidas de tendência central e dispersão, para variáveis categorizadas como contínuas, valores mínimos e máximos, número de casos ausentes, etc.

4. Modelos

5. Modelo 1 - PBF

O nosso modelo 1 consiste em incluir somente PBF como variável explicativa. Pode ser descrito em sua forma populacional como:

$$AcessoEsgoto = \beta_0 + \beta_1 \cdot PBF + \epsilon$$

Para estimar, basta rodar:

```
modelo1 <- lm(
  acesso_esgoto ~ pbf,
  data = dados
)
```

Para obter seus resultados, temos:

```
summary(modelo1)
```

```

Call:
lm(formula = acesso_esgoto ~ pbf, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-76.830 -20.134  -7.137  23.807  83.405

Coefficients:
            Estimate Std. Error t value            Pr(>|t|)
```

```
(Intercept) 88.06360    1.02972   85.52 <0.0000000000000002 ***
pbf         -0.84226    0.01546  -54.47 <0.0000000000000002 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.56 on 10408 degrees of freedom
Multiple R-squared:  0.2218,    Adjusted R-squared:  0.2217
F-statistic:  2966 on 1 and 10408 DF,  p-value: < 0.00000000000000022
```

Podemos observar que um aumento em 1% do percentual de famílias que recebem bolsa família reduziria, em média, em 0.84% o percentual de famílias com acesso à rede de esgoto.

6. Modelo 2 - Informalidade

O nosso modelo 2 consiste em incluir a Informalidade, para além do PBF como variável explicativa. Pode ser descrito em sua forma populacional como:

$$AcessoEsgoto = \beta_0 + \beta_1 \cdot PBF + \beta_2 \cdot Informalidade + \epsilon$$

Para estimar, basta rodar:

```r
modelo2 <- lm(
  acesso_esgoto ~ pbf + informalidade,
  data = dados
)
```

Para obter seus resultados, temos:

```r
summary(modelo2)
```

```
Call:
lm(formula = acesso_esgoto ~ pbf + informalidade, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-76.908 -20.132  -7.205  23.797  84.304

Coefficients:
              Estimate Std. Error t value             Pr(>|t|)
(Intercept)   90.58242    1.19783  75.622 < 0.0000000000000002 ***
pbf           -0.83111    0.01569 -52.973 < 0.0000000000000002 ***
informalidade -0.05564    0.01355  -4.107            0.0000403 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.54 on 10407 degrees of freedom
Multiple R-squared:  0.2231,    Adjusted R-squared:  0.2229
F-statistic:  1494 on 2 and 10407 DF,  p-value: < 0.00000000000000022
```

Neste caso, o que obtemos é uma relação negativa em que, aumentando a informalidade entre os jovens em 1%, haveria uma redução do percentual de famílias com acesso ao esgotamento sanitário em 0.05%. Para o PBF, parece manter o mesmo sentido e intensidade.

7. Modelo 3 - Ensino Médio

O nosso modelo 3 consiste em incluir somente frequência ao Ensino Médio, para além de PBF, como variável explicativa. Pode ser descrito em sua forma populacional como:

$$AcessoEsgoto = \beta_0 + \beta_1 \cdot PBF + \beta_2 \cdot EnsinoMedio + \epsilon$$

Para estimar, basta rodar:

```r
modelo3 <- lm(
  acesso_esgoto ~ pbf + em,
  data = dados
)
```

Para obter seus resultados, temos:

```r
summary(modelo3)
```

```
Call:
lm(formula = acesso_esgoto ~ pbf + em, data = dados)

Residuals:
   Min     1Q Median     3Q    Max
-72.79 -20.27  -6.87  23.68  81.82

Coefficients:
            Estimate Std. Error t value             Pr(>|t|)
(Intercept) 82.23602    1.29138  63.681 < 0.0000000000000002 ***
pbf         -0.83249    0.01548 -53.779 < 0.0000000000000002 ***
em           0.16178    0.02173   7.444    0.000000000000105 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
Residual standard error: 28.48 on 10407 degrees of freedom
Multiple R-squared:  0.2259,    Adjusted R-squared:  0.2258
F-statistic:  1519 on 2 and 10407 DF,  p-value: < 0.00000000000000022
```

Neste caso, o que obtemos é uma relação positiva em que, aumentando a frequência média dos jovens ao Ensino Médio em 1%, haveria um aumento do percentual de famílias com acesso ao esgotamento sanitário em 0.16%. Por outro lado, a associação do PBF mantém a mesma.

8. Modelo 4 - PBF, Informalidade e Ensino Médio

O nosso modelo 4 consiste em incluir todas as variáveis explicativas do acesso ao esgotamento sanitário. Pode ser descrito em sua forma populacional como:

$$AcessoEsgoto = \beta_0 + \beta_1 \cdot PBF + \beta_2 \cdot Informalidade + \beta_3 \cdot EnsinoMedio + \epsilon$$

Para estimar, basta rodar:

```r
# modelo 4 - todos juntos

modelo4 <- lm(
  acesso_esgoto ~ pbf + informalidade + em,
  data = dados
)
```

Para obter seus resultados, temos:

```r
summary(modelo4)
```

```
Call:
lm(formula = acesso_esgoto ~ pbf + informalidade + em, data = dados)

Residuals:
    Min      1Q  Median      3Q     Max
-73.676 -20.199  -6.888  23.703  81.620

Coefficients:
             Estimate Std. Error t value             Pr(>|t|)
(Intercept)  84.75562    1.42782  59.360 < 0.0000000000000002 ***
pbf          -0.82132    0.01570 -52.303 < 0.0000000000000002 ***
informalidade -0.05569    0.01351  -4.122   0.0000377921116428 ***
em            0.16183    0.02172   7.452   0.0000000000000991 ***
---
```

```
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 28.46 on 10406 degrees of freedom
Multiple R-squared:  0.2272,    Adjusted R-squared:  0.227
F-statistic:  1020 on 3 and 10406 DF,  p-value: < 0.00000000000000022
```

Os resultados parecem se manter semelhantes aos anteriores!