Renda per capita Metropolitana - Porto Alegre - RS

Pós-Graduação Lato Senso - Big Data, Data Science e Data Analytics

Luciano Teixeira

08/julho/2018

Contents

1 Introdução da Análise

O arquivo utilizado, se refere aos dados municipais do Atlas do desenvolvimento humano no Brasil referentes aos Censos de 1991, 2000 e 2010 em http://www.atlasbrasil.org.br/2013/pt/download/.

Foram escolhidas 5 variáveis explicativas para a renda per capita dos municípios.

* IDHM: Índice de Desenvolvimento Humano Municipal
* ESPVIDA: Esperança de vida ao nascer
* GINI: Índice de Gini
* PESOURB: População residente na área urbana
* T_FBSUPER: Taxa de frequência bruta ao ensino superior

A amostra será demonstrada por meio de uma análise descritiva des variáveis explicaivas em relação à evolução da renda per capita dos municípios da região metropolitana de Porto Alegre sobre os anos de 1991, 2000 e 2010.

Como método de análise, será utilizado regressão linear múltipla onde a VR é a renda per capita e as variáveis explicativas são as 5 escolhidas no passo 2.

2 Inicializando Bibliotecas

Como primeiro passo, serão carregadas a seguintes bibliotecas. Caso estas não se encontrem instaladas, é necessário que esta instalação seja eetuada.

```
library(readr)
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(readxl)
library(ggplot2)
library(stringi)
library(stringr)
library(car)
```

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

```
##
## Attaching package: 'car'
```

```
## The following object is masked from 'package:dplyr':
##
##     recode
```

3 Importando Dados Brutos

```
dadosbrutos <- read_excel("atlas2013_municipios.xlsx")
```

4 Especificando os Dados

Comandos Encadeados podem demonstrar um principio de Machine Learning, segregando cidades, Estado e Região. No caso deste modelo, oi delimitado a Região Metropolitana de Porto Alegre, podendo ser aplicado em qualquer estado, macro região ou micro região, com pequenos ajustes.

Este encadeameno de funçoes, substiui uma série de passos, utilizados anteriormente para chegar à um resultado muito mais enchuto, levando em consideração proficionais de analise de dados com poucos recursos em questão de equipamenos, como por exemplos computadores de pequeno porte, pouca memória e processador limitado.

4.1 Dados de 2000

```
dadosrs <-
  filter(
    select(
      subset.data.frame(dadosbrutos, UF == 43),
      ANO,
      UF,
      MUNICIPIO,
      RDPC,
      IDHM,
      ESPVIDA,
      GINI,
      PESOURB,
      T_FBSUPER
    ),
    MUNICIPIO %in% c("VIAMÃO","TRIUNFO","TAQUARA","LEOPOLDO","SÃO JERÔNIMO",
                     "SAPUCAIA DO SUL","SAPIRANGA","SANTO ANTÔNIO DA PATRULHA",
                     "PORTÃO","PORTO ALEGRE","PAROBÉ","HAMBURGO","NOVA SANTA RITA",
                     "NOVA HARTZ","MONTENEGRO","IVOTI","GUAÍBA","GRAVATAÍ","GLORINHA",
                     "ESTÂNCIA VELHA","ESTEIO","ELDORADO DO SUL","DOIS IRMÃOS",
                     "CHARQUEADAS","CAPELA DE SANTANA","CANOAS","CAMPO BOM",
                     "CACHOEIRINHA","ARROIO DOS RATOS","ARARICÁ","ALVORADA"),
    ANO == 2000
  )
```

4.2 Listando os Dados de 2000

```
head(dadosrs)
```

```
## # A tibble: 6 x 9
##     ANO    UF MUNICIPIO         RDPC  IDHM ESPVIDA  GINI PESOURB T_FBSUPER
##   <dbl> <dbl> <chr>            <dbl> <dbl>   <dbl> <dbl>   <dbl>     <dbl>
## 1  2000    43 ALVORADA          429. 0.582    73.3  0.44  183365      6.71
## 2  2000    43 ARARICÁ           422. 0.565    71.3  0.42    3493      3.76
## 3  2000    43 ARROIO DOS RATOS  444. 0.61     71.4  0.5    12528     15.0
## 4  2000    43 CACHOEIRINHA      629. 0.672    74.2  0.47  107564     15.7
## 5  2000    43 CAMPO BOM         726. 0.669    74.1  0.48   51838     18.1
## 6  2000    43 CANOAS            704. 0.665    74.1  0.52  306093     22.8
```

Total de 29 registros.

4.3 Sumário do Dados de 2000

```
summary(dadosrs)
```

```
##       ANO             UF      MUNICIPIO             RDPC
##  Min.   :2000   Min.   :43   Length:29          Min.   : 376.3
##  1st Qu.:2000   1st Qu.:43   Class :character   1st Qu.: 504.5
##  Median :2000   Median :43   Mode  :character   Median : 575.1
##  Mean   :2000   Mean   :43                      Mean   : 600.1
##  3rd Qu.:2000   3rd Qu.:43                      3rd Qu.: 670.7
##  Max.   :2000   Max.   :43                      Max.   :1399.5
##       IDHM           ESPVIDA           GINI           PESOURB
##  Min.   :0.5340   Min.   :70.20   Min.   :0.3700   Min.   :   1285
##  1st Qu.:0.6090   1st Qu.:72.06   1st Qu.:0.4500   1st Qu.:  13785
##  Median :0.6280   Median :73.45   Median :0.4800   Median :  34367
##  Mean   :0.6339   Mean   :73.26   Mean   :0.4862   Mean   : 107818
##  3rd Qu.:0.6680   3rd Qu.:74.25   3rd Qu.:0.5200   3rd Qu.:  91956
##  Max.   :0.7440   Max.   :76.11   Max.   :0.6100   Max.   :1320739
##    T_FBSUPER
##  Min.   : 3.76
##  1st Qu.:11.16
##  Median :14.25
##  Mean   :15.55
##  3rd Qu.:18.08
##  Max.   :42.01
```

4.4 Dados de 2010

```
dadosrs_2010 <-
  filter(
    select(
      subset.data.frame(dadosbrutos, UF == 43),
      ANO,
      UF,
      MUNICIPIO,
      RDPC,
      IDHM,
      ESPVIDA,
      GINI,
      PESOURB,
      T_FBSUPER
    ),
    MUNICIPIO %in% c("VIAMÃO","TRIUNFO","TAQUARA","LEOPOLDO","SÃO JERÔNIMO",
                     "SAPUCAIA DO SUL","SAPIRANGA","SANTO ANTÔNIO DA PATRULHA",
                     "PORTÃO","PORTO ALEGRE","PAROBÉ","HAMBURGO","NOVA SANTA RITA",
                     "NOVA HARTZ","MONTENEGRO","IVOTI","GUAÍBA","GRAVATAÍ","GLORINHA",
                     "ESTÂNCIA VELHA","ESTEIO","ELDORADO DO SUL","DOIS IRMÃOS",
                     "CHARQUEADAS","CAPELA DE SANTANA","CANOAS","CAMPO BOM",
                     "CACHOEIRINHA","ARROIO DOS RATOS","ARARICÁ","ALVORADA"),
    ANO == 2010
  )
```

4.5 Listando os Dados de 2010

```
head(dadosrs_2010)
```

```
## # A tibble: 6 x 9
##     ANO    UF MUNICIPIO         RDPC  IDHM ESPVIDA  GINI PESOURB T_FBSUPER
##   <dbl> <dbl> <chr>            <dbl> <dbl>   <dbl> <dbl>   <dbl>     <dbl>
## 1  2010    43 ALVORADA          600. 0.699    77.4  0.43  195673      17.2
## 2  2010    43 ARARICÁ           610. 0.679    74.4  0.35    3996      15.4
## 3  2010    43 ARROIO DOS RATOS  624. 0.698    75.0  0.46   12956      27.2
## 4  2010    43 CACHOEIRINHA      844. 0.757    76.4  0.44  118278      37.7
## 5  2010    43 CAMPO BOM         880. 0.745    76.1  0.43   57338      29.0
## 6  2010    43 CANOAS            952. 0.75     76.8  0.51  323827      42
```

Total de 29 registros.

4.6 Sumário do Dados de 2010

```
summary(dadosrs_2010)
```

```
##       ANO             UF      MUNICIPIO             RDPC
##  Min.   :2010   Min.   :43   Length:29          Min.   : 533.9
##  1st Qu.:2010   1st Qu.:43   Class :character   1st Qu.: 687.0
##  Median :2010   Median :43   Mode  :character   Median : 733.3
##  Mean   :2010   Mean   :43                      Mean   : 789.8
##  3rd Qu.:2010   3rd Qu.:43                      3rd Qu.: 871.4
##  Max.   :2010   Max.   :43                      Max.   :1758.3
##       IDHM          ESPVIDA           GINI           PESOURB
##  Min.   :0.661   Min.   :73.93   Min.   :0.3400   Min.   :   2067
##  1st Qu.:0.711   1st Qu.:75.57   1st Qu.:0.4200   1st Qu.:  18062
##  Median :0.726   Median :76.37   Median :0.4400   Median :  41484
##  Mean   :0.727   Mean   :76.21   Mean   :0.4431   Mean   : 117138
##  3rd Qu.:0.747   3rd Qu.:76.95   3rd Qu.:0.4700   3rd Qu.:  93064
##  Max.   :0.805   Max.   :78.23   Max.   :0.6000   Max.   :1409351
##    T_FBSUPER
##  Min.   :15.42
##  1st Qu.:24.41
##  Median :28.99
##  Mean   :30.84
##  3rd Qu.:35.09
##  Max.   :64.55
```

5 Segregando os Dados em Relação a Variavel Y = RDPC

```
Y_RDPC_2010 <- c(dadosrs_2010$RDPC)
```

6 Inserindo os dados de 2010 na base de 2000

```
dadosrs<- data.frame(dadosrs,Y_RDPC_2010)
```

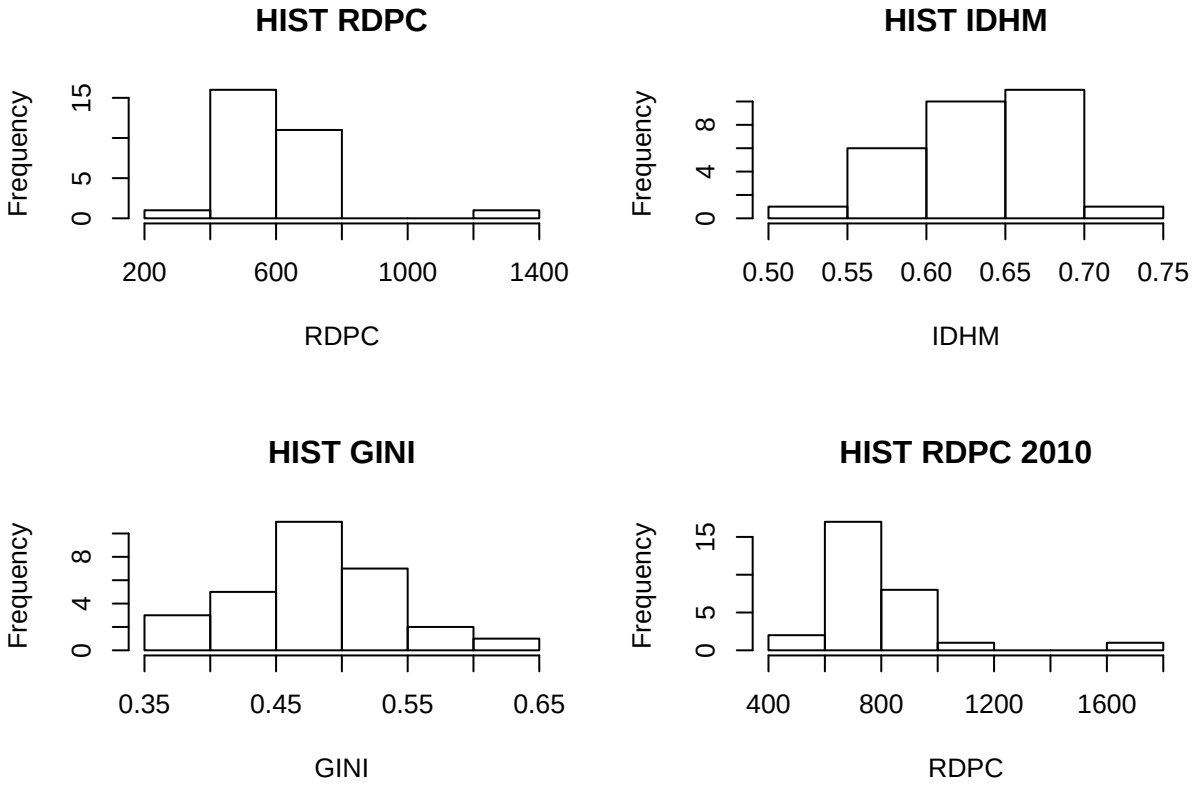
7 Sumário de variáveis

```
summary(dadosrs)
```

```
##       ANO             UF      MUNICIPIO             RDPC
##  Min.   :2000   Min.   :43   Length:29          Min.   : 376.3
##  1st Qu.:2000   1st Qu.:43   Class :character   1st Qu.: 504.5
##  Median :2000   Median :43   Mode  :character   Median : 575.1
##  Mean   :2000   Mean   :43                      Mean   : 600.1
##  3rd Qu.:2000   3rd Qu.:43                      3rd Qu.: 670.7
##  Max.   :2000   Max.   :43                      Max.   :1399.5
##       IDHM           ESPVIDA           GINI           PESOURB
##  Min.   :0.5340   Min.   :70.20   Min.   :0.3700   Min.   :   1285
##  1st Qu.:0.6090   1st Qu.:72.06   1st Qu.:0.4500   1st Qu.:  13785
##  Median :0.6280   Median :73.45   Median :0.4800   Median :  34367
##  Mean   :0.6339   Mean   :73.26   Mean   :0.4862   Mean   : 107818
##  3rd Qu.:0.6680   3rd Qu.:74.25   3rd Qu.:0.5200   3rd Qu.:  91956
##  Max.   :0.7440   Max.   :76.11   Max.   :0.6100   Max.   :1320739
##    T_FBSUPER      Y_RDPC_2010
##  Min.   : 3.76   Min.   : 533.9
##  1st Qu.:11.16   1st Qu.: 687.0
##  Median :14.25   Median : 733.3
##  Mean   :15.55   Mean   : 789.8
##  3rd Qu.:18.08   3rd Qu.: 871.4
##  Max.   :42.01   Max.   :1758.3
```

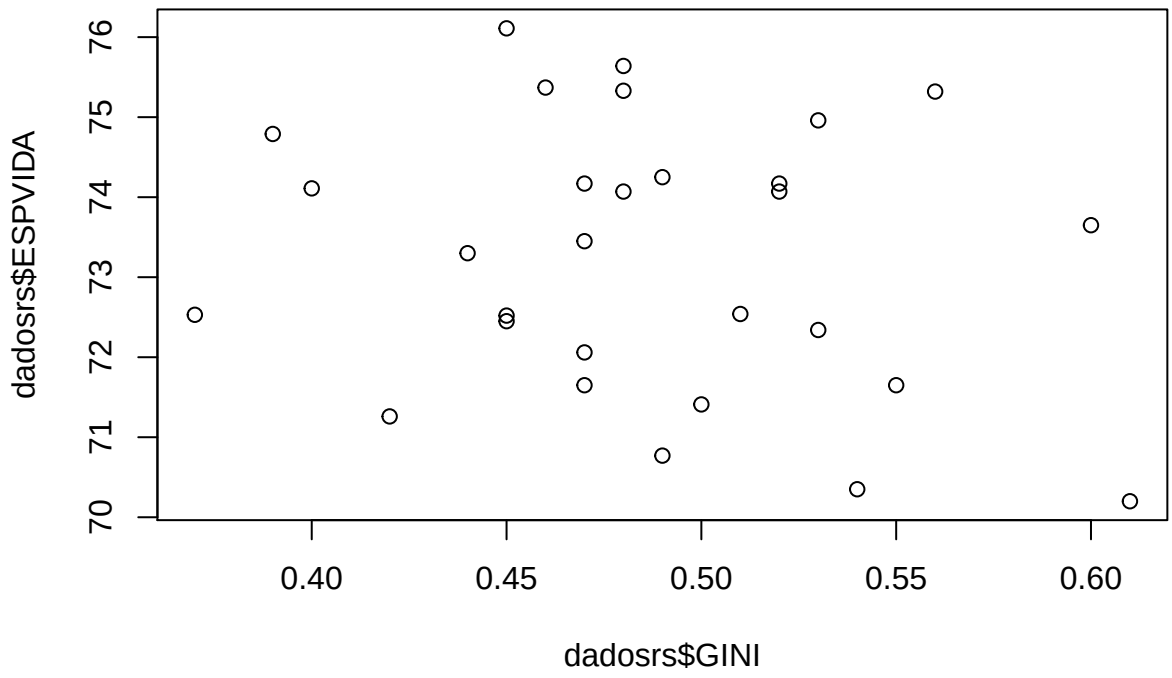
8 Histogramas

```
par(mfrow = c(2,2))

hist(dadosrs$RDPC, xlab = "RDPC", main = "HIST RDPC")
hist(dadosrs$IDHM, xlab = "IDHM", main = "HIST IDHM")
hist(dadosrs$GINI, xlab = "GINI", main = "HIST GINI")
hist(dadosrs$Y_RDPC_2010, xlab = "RDPC", main = "HIST RDPC 2010")
```



9 Análise de correlação linear entre duas variáveis quantitativas

```
plot(dadosrs$GINI,dadosrs$ESPVIDA)
```




```
cor(dadosrs$GINI,dadosrs$ESPVIDA)
```

```
## [1] -0.1923844
```

10 Aplicação da Regressao Multipla

```
reg <- lm(RDPC ~ IDHM + ESPVIDA + GINI + PESOURB + T_FBSUPER, data = dadosrs)
```

11 Teste a significância global do modelo de regressão.

```
summary(reg)
```

```
##
## Call:
## lm(formula = RDPC ~ IDHM + ESPVIDA + GINI + PESOURB + T_FBSUPER,
##     data = dadosrs)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -100.383  -42.143   -0.363   38.871   91.328
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.686e+02  5.908e+02  -0.285  0.77798
## IDHM         1.004e+03  4.913e+02   2.043  0.05267 .
## ESPVIDA     -1.204e+00  8.384e+00  -0.144  0.88709
## GINI         4.911e+01  2.530e+02   0.194  0.84778
## PESOURB      3.012e-04  5.657e-05   5.324  2.1e-05 ***
## T_FBSUPER    1.056e+01  3.138e+00   3.366  0.00267 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 56.03 on 23 degrees of freedom
## Multiple R-squared:  0.9255, Adjusted R-squared:  0.9093
## F-statistic: 57.16 on 5 and 23 DF,  p-value: 3.287e-12
```

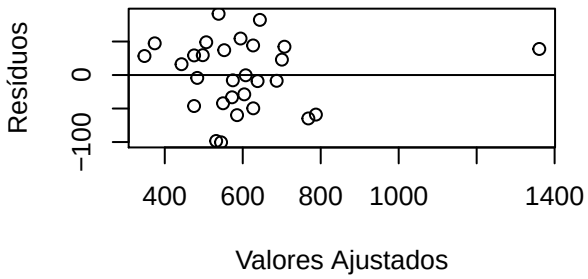
12 Intervalos de confiança para os coeficientes da equação.

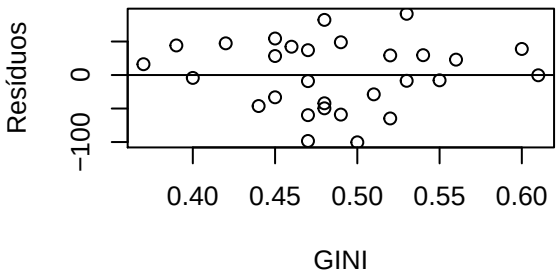
```
confint(reg)
```

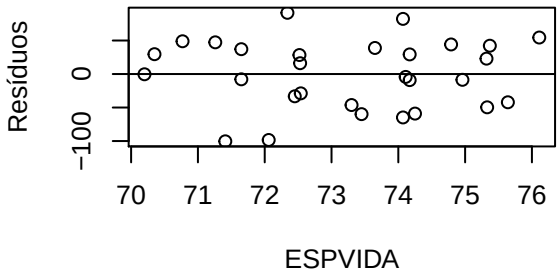
```
##                     2.5 %       97.5 %
## (Intercept) -1.390756e+03 1.053655e+03
## IDHM        -1.256329e+01 2.020167e+03
## ESPVIDA     -1.854803e+01 1.614057e+01
## GINI        -4.742425e+02 5.724709e+02
## PESOURB      1.841587e-04 4.182144e-04
## T_FBSUPER    4.071895e+00 1.705401e+01
```

13 Distribuição dos Resíduos

```
par(mfrow = c(2,2))

plot(fitted(reg),residuals(reg),xlab="Valores Ajustados",ylab="Resíduos")
abline(h=0)

plot(dadosrs$GINI,residuals(reg),xlab="GINI",ylab="Resíduos")
abline(h=0)

plot(dadosrs$ESPVIDA,residuals(reg),xlab="ESPVIDA",ylab="Resíduos")
abline(h=0)
```







14 Teste de Shapiro

```
shapiro.test(reg$residuals)
```

```
##
##  Shapiro-Wilk normality test
##
## data:  reg$residuals
## W = 0.96435, p-value = 0.4185
```

15 Outliers

Não foi detetado nenhum outliers.

```
which(rstudent(reg) > 2)
```

```
## 22
## 22
```

16 Multicolinearidade

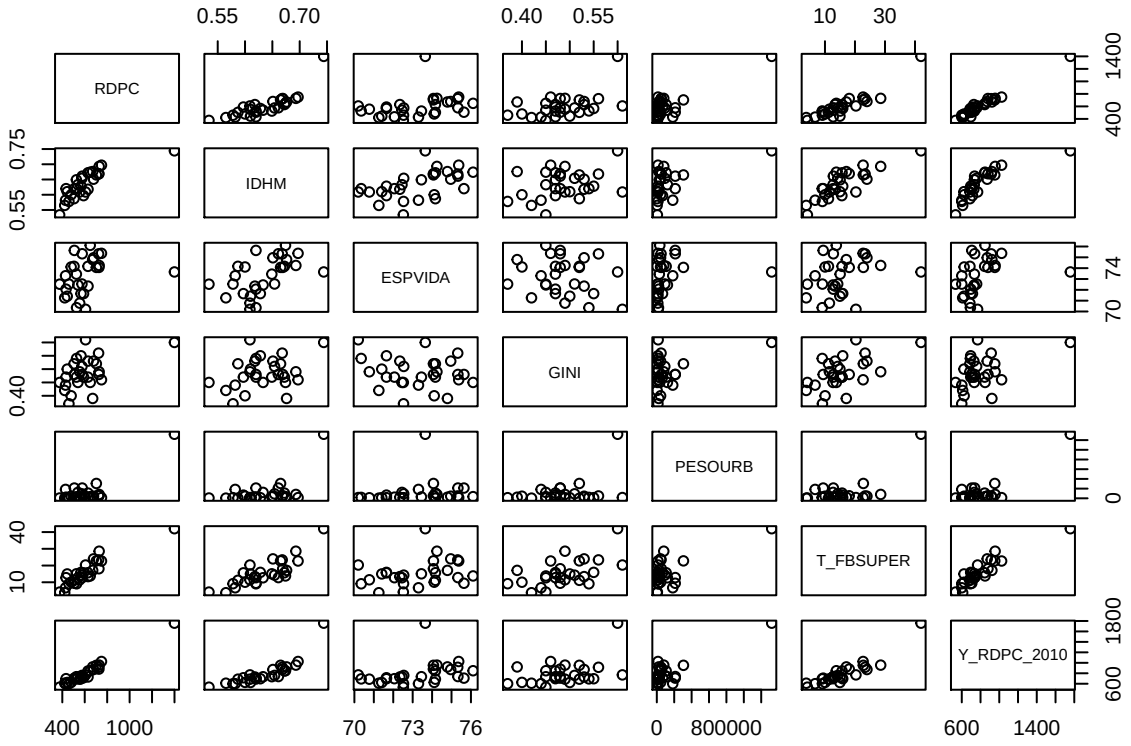
Multicolinearidade consiste em um problema comum em regressões, no qual as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas.

16.1 Multicolinearidade 01

```
attach(dadosrs)
```

```
## The following object is masked _by_ .GlobalEnv:
##
##     Y_RDPC_2010
```

```
pairs(dadosrs[,c(4:10)])
```




```
round(cor(dadosrs[,c(4:10)]),3)
```

```
##                RDPC  IDHM ESPVIDA   GINI PESOURB T_FBSUPER Y_RDPC_2010
## RDPC          1.000 0.812   0.305  0.492   0.811     0.906       0.983
## IDHM          0.812 1.000   0.527  0.322   0.512     0.824       0.833
## ESPVIDA       0.305 0.527   1.000 -0.192   0.157     0.287       0.345
## GINI          0.492 0.322  -0.192  1.000   0.371     0.562       0.439
## PESOURB       0.811 0.512   0.157  0.371   1.000     0.645       0.826
## T_FBSUPER     0.906 0.824   0.287  0.562   0.645     1.000       0.905
## Y_RDPC_2010 0.983 0.833   0.345  0.439   0.826     0.905       1.000
```

16.2 Multicolinearidade 02

Foram detetados valores superiores a 5, que indicam associação muito fraca entre variáveis explicativas.

```
vif(reg)
```

```
##      IDHM   ESPVIDA      GINI   PESOURB T_FBSUPER
##  4.411788  1.747208  1.836354  1.716155  5.346944
```