Renda per capita Metropolitana - Porto Alegre - RS

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

Luciano Teixeira

01/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("NOVO HAMBURGO", "SÃO LEOPOLDO", "SAPUCAIA DO SUL",
                     "ESTEIO", "CANOAS", "PORTO ALEGRE", "GUAÍBA"),
    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 CANOAS         704. 0.665    74.1  0.52  306093      22.8
## 2  2000    43 ESTEIO         729. 0.693    74.2  0.49   79961      28.6
## 3  2000    43 GUAÍBA         575. 0.654    72.5  0.51   91956      14.2
## 4  2000    43 NOVO HAMBURGO  770. 0.671    74.4  0.54  231989      24.7
## 5  2000    43 PORTO ALEGRE  1400. 0.744    73.6  0.6  1320739      42.0
## 6  2000    43 SÃO LEOPOLDO   729. 0.656    73.4  0.54  192895      28.5
```

Total de 7 registros.

4.3 Sumário do Dados de 2000

```
summary(dadosrs)
```

```
##       ANO             UF      MUNICIPIO             RDPC
##  Min.   :2000   Min.   :43   Length:7           Min.   : 539.2
##  1st Qu.:2000   1st Qu.:43   Class :character   1st Qu.: 639.3
##  Median :2000   Median :43   Mode  :character   Median : 728.6
##  Mean   :2000   Mean   :43                      Mean   : 777.9
##  3rd Qu.:2000   3rd Qu.:43                      3rd Qu.: 749.7
##  Max.   :2000   Max.   :43                      Max.   :1399.5
##       IDHM           ESPVIDA           GINI           PESOURB
##  Min.   :0.6330   Min.   :72.45   Min.   :0.4500   Min.   :  79961
##  1st Qu.:0.6550   1st Qu.:72.94   1st Qu.:0.5000   1st Qu.: 106885
##  Median :0.6650   Median :73.65   Median :0.5200   Median : 192895
##  Mean   :0.6737   Mean   :73.53   Mean   :0.5214   Mean   : 335064
##  3rd Qu.:0.6820   3rd Qu.:74.16   3rd Qu.:0.5400   3rd Qu.: 269041
##  Max.   :0.7440   Max.   :74.38   Max.   :0.6000   Max.   :1320739
##    T_FBSUPER
##  Min.   :12.69
##  1st Qu.:18.52
##  Median :24.68
##  Mean   :24.79
##  3rd Qu.:28.55
##  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("NOVO HAMBURGO", "SÃO LEOPOLDO", "SAPUCAIA DO SUL",
                     "ESTEIO", "CANOAS", "PORTO ALEGRE", "GUAÍBA"),
    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 CANOAS         952. 0.75     76.8  0.51  323827      42
## 2  2010    43 ESTEIO         956. 0.754    75.6  0.48   80643      44.0
## 3  2010    43 GUAÍBA         759. 0.73     75.0  0.47   93064      36.4
## 4  2010    43 NOVO HAMBURGO 1012. 0.747    76.1  0.53  234798      38.4
## 5  2010    43 PORTO ALEGRE  1758. 0.805    76.4  0.6  1409351      64.6
## 6  2010    43 SÃO LEOPOLDO   940. 0.739    76.6  0.53  213238      40.7
```

Total de 7 registros.

4.6 Sumário do Dados de 2010

```
summary(dadosrs_2010)
```

```
##       ANO             UF      MUNICIPIO             RDPC
##  Min.   :2010   Min.   :43   Length:7           Min.   : 733.3
##  1st Qu.:2010   1st Qu.:43   Class :character   1st Qu.: 849.4
##  Median :2010   Median :43   Mode  :character   Median : 952.1
##  Mean   :2010   Mean   :43                      Mean   :1015.7
##  3rd Qu.:2010   3rd Qu.:43                      3rd Qu.: 983.7
##  Max.   :2010   Max.   :43                      Max.   :1758.3
##       IDHM           ESPVIDA           GINI           PESOURB
##  Min.   :0.7260   Min.   :74.99   Min.   :0.4400   Min.   :  80643
##  1st Qu.:0.7345   1st Qu.:75.61   1st Qu.:0.4750   1st Qu.: 111767
##  Median :0.7470   Median :76.11   Median :0.5100   Median : 213238
##  Mean   :0.7501   Mean   :76.03   Mean   :0.5086   Mean   : 355056
##  3rd Qu.:0.7520   3rd Qu.:76.53   3rd Qu.:0.5300   3rd Qu.: 279313
##  Max.   :0.8050   Max.   :76.83   Max.   :0.6000   Max.   :1409351
##    T_FBSUPER
##  Min.   :35.09
##  1st Qu.:37.39
##  Median :40.70
##  Mean   :43.02
##  3rd Qu.:43.02
##  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)
```

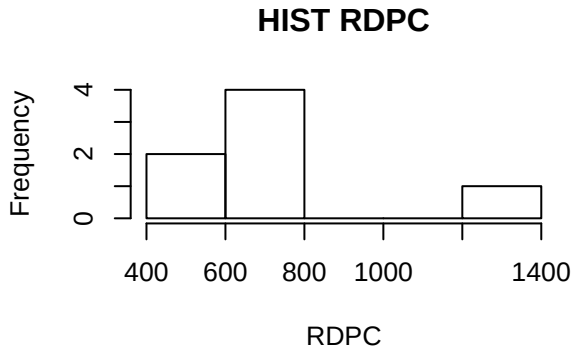
7 Sumário de variáveis

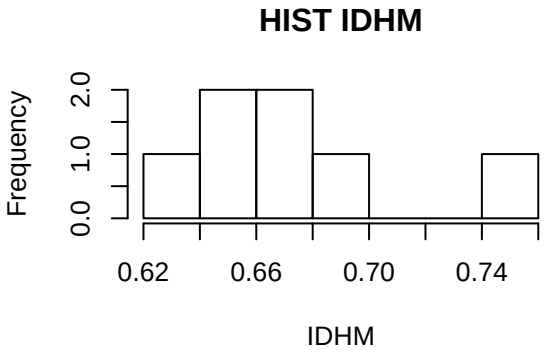
```
summary(dadosrs)
```

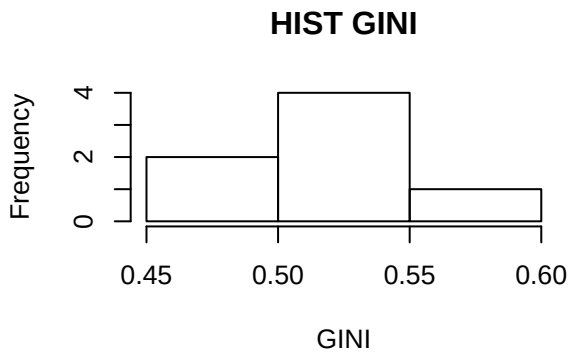
```
##       ANO             UF      MUNICIPIO             RDPC
##  Min.   :2000   Min.   :43   Length:7           Min.   : 539.2
##  1st Qu.:2000   1st Qu.:43   Class :character   1st Qu.: 639.3
##  Median :2000   Median :43   Mode  :character   Median : 728.6
##  Mean   :2000   Mean   :43                      Mean   : 777.9
##  3rd Qu.:2000   3rd Qu.:43                      3rd Qu.: 749.7
##  Max.   :2000   Max.   :43                      Max.   :1399.5
##       IDHM           ESPVIDA           GINI           PESOURB
##  Min.   :0.6330   Min.   :72.45   Min.   :0.4500   Min.   :  79961
##  1st Qu.:0.6550   1st Qu.:72.94   1st Qu.:0.5000   1st Qu.: 106885
##  Median :0.6650   Median :73.65   Median :0.5200   Median : 192895
##  Mean   :0.6737   Mean   :73.53   Mean   :0.5214   Mean   : 335064
##  3rd Qu.:0.6820   3rd Qu.:74.16   3rd Qu.:0.5400   3rd Qu.: 269041
##  Max.   :0.7440   Max.   :74.38   Max.   :0.6000   Max.   :1320739
##    T_FBSUPER      Y_RDPC_2010
##  Min.   :12.69   Min.   : 733.3
##  1st Qu.:18.52   1st Qu.: 849.4
##  Median :24.68   Median : 952.1
##  Mean   :24.79   Mean   :1015.7
##  3rd Qu.:28.55   3rd Qu.: 983.7
##  Max.   :42.01   Max.   :1758.3
```

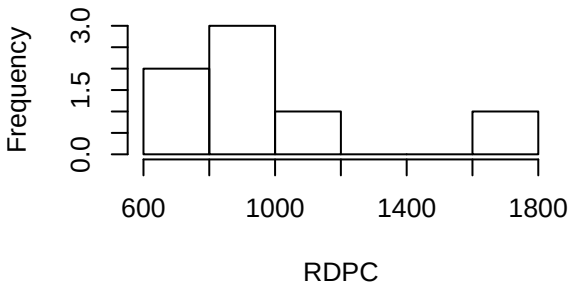
8 Histogramas

```
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")
```



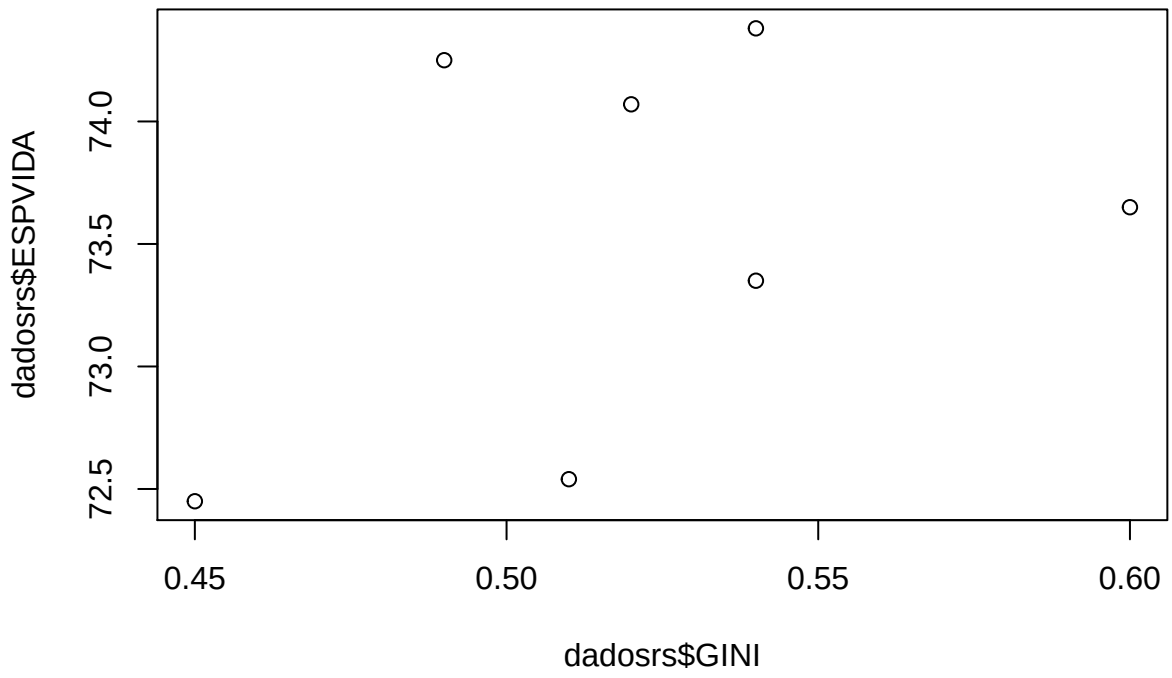






9 Análise de correlação linear entre duas variáveis quantitativas

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




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

```
## [1] 0.3936927
```

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:
##       1       2       3       4       5       6       7
## -33.460  -2.321  -6.362  30.554   2.277  -3.052  12.363
##
## Coefficients:
```

```
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -5.607e+02  2.620e+03  -0.214    0.866
## IDHM         1.596e+03  1.617e+03   0.987    0.504
## ESPVIDA     -2.961e+00  3.871e+01  -0.076    0.951
## GINI         3.462e+02  8.496e+02   0.407    0.754
## PESOURB      3.609e-04  1.196e-04   3.018    0.204
## T_FBSUPER    7.231e+00  5.828e+00   1.241    0.432
##
## Residual standard error: 47.61 on 1 degrees of freedom
## Multiple R-squared:  0.9954, Adjusted R-squared:  0.9725
## F-statistic: 43.47 on 5 and 1 DF,  p-value: 0.1146
```

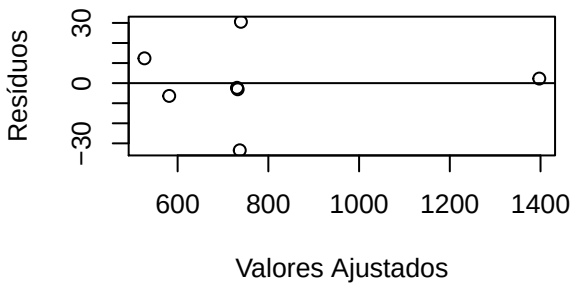
12 Intervalos de confiança para os coeficientes da equação.

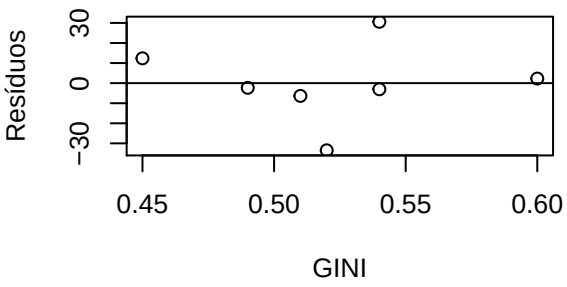
```
confint(reg)
```

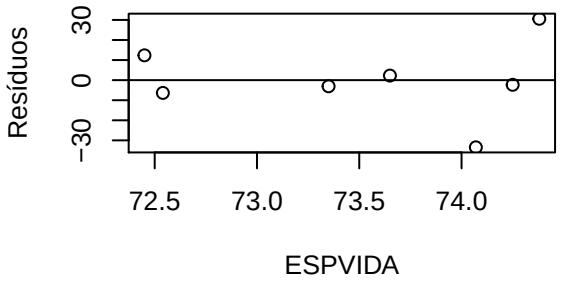
```
##                     2.5 %       97.5 %
## (Intercept) -3.385014e+04 3.272879e+04
## IDHM        -1.894629e+04 2.213917e+04
## ESPVIDA     -4.947954e+02 4.888738e+02
## GINI        -1.044888e+04 1.114127e+04
## PESOURB     -1.158877e-03 1.880740e-03
## T_FBSUPER   -6.681589e+01 8.127878e+01
```

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 = 0.94644, p-value = 0.6972
```

15 Outliers

Não foi detetado nenhum outliers.

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

```
## named integer(0)
```

16 Multicolinearidade

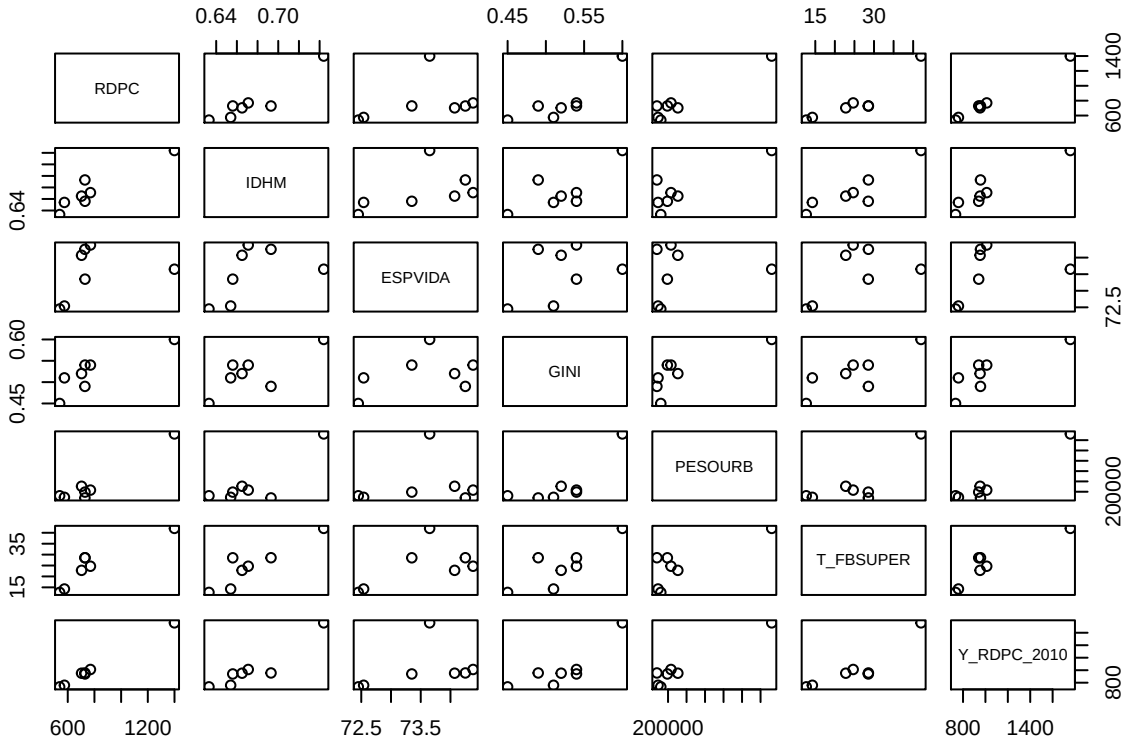
Multicolinearidade consiste em um problema comum em regressões, no qual as variáveis independentes possuem relações lineares exatas ou aproximadamente exatas.

16.1 Multicolinearidade 01

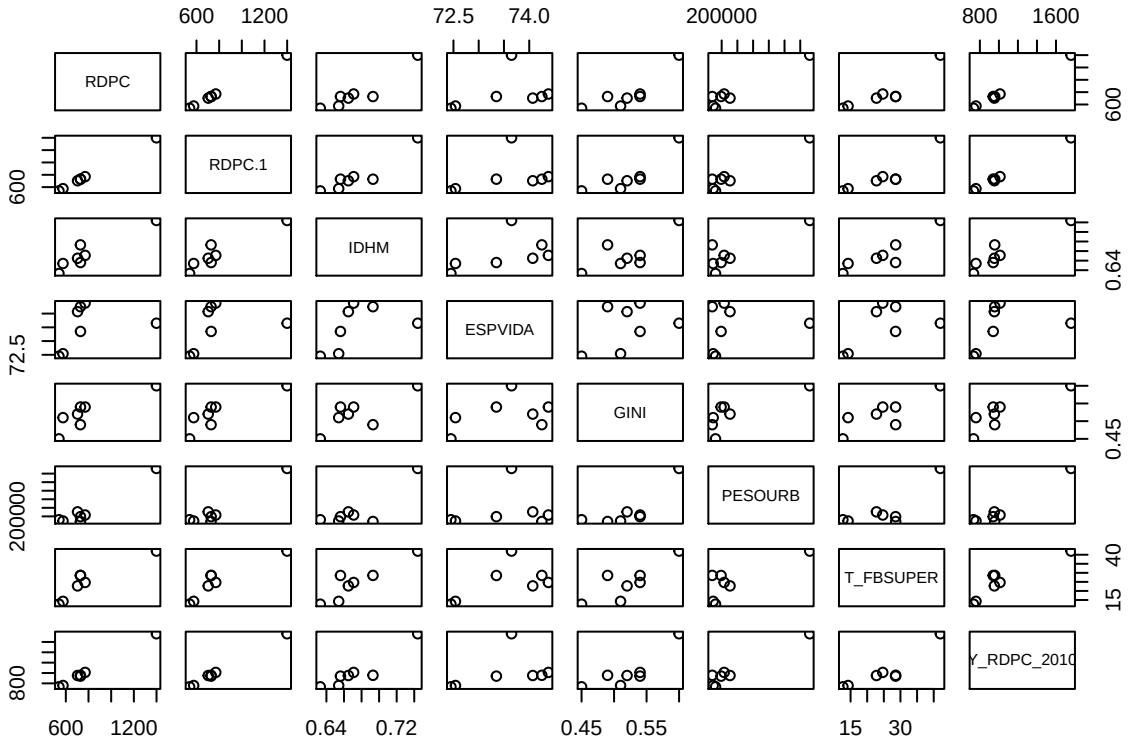
```
attach(dadosrs)
```

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

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




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




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

```
##              RDPC  IDHM ESPVIDA  GINI PESOURB T_FBSUPER Y_RDPC_2010
## RDPC        1.000 0.936   0.338 0.855   0.965     0.911       0.999
## IDHM        0.936 1.000   0.482 0.763   0.848     0.907       0.936
## ESPVIDA     0.338 0.482   1.000 0.394   0.157     0.566       0.357
## GINI        0.855 0.763   0.394 1.000   0.799     0.826       0.850
## PESOURB     0.965 0.848   0.157 0.799   1.000     0.789       0.967
## T_FBSUPER   0.911 0.907   0.566 0.826   0.789     1.000       0.908
## Y_RDPC_2010 0.999 0.936   0.357 0.850   0.967     0.908       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
##  8.944670  2.463163  4.167646  7.404158  8.787186
```