

# Relatório 1 - Regressão

Flavio Margarito Martins de Barros Gabriel Tupinamba da Cunha Leandro Gustavo Leite Machado

14/05/2022

## Conjunto de dados

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

```
## Carregando os pacotes
```

```
require(readxl)
require(corrplot)
require(psych)
require(kableExtra)
require(caret)
require(GGally)
```

```
## Lendo o banco de dados
```

```
## Fonte: https://archive.ics.uci.edu/ml/datasets/Concrete+Compressive+Strength
dados <- read_excel(path = "Concrete_Data.xls", sheet = 1)
```

```
## Trocando os nomes das variáveis para o português
```

```
colnames(dados) <-
c(
  "cimento",
  "escoria",
  "cinza",
  "agua",
  "super_plastificante",
  "agregador_grosso",
  "agregador_fino",
  "idade",
  "forca_compressiva"
)
```

```
## Descrição básica dos dados
```

## Cimento	-- numérica -- kg / m3	-- Variável explicativa
## Escoria	-- numérica -- kg / m3	-- Variável explicativa
## Cinza	-- numérica -- kg / m3	-- Variável explicativa
## Agua	-- numérica -- kg / m3	-- Variável explicativa
## Super plastificante	-- numérica -- kg / m3	-- Variável explicativa
## Agregadro grosso	-- numérica -- kg / m3	-- Variável explicativa

```
## Agregador fino      -- numérica -- kg / m3      -- Variável explicativa
## Idade               -- numérica -- Dias (1~365) -- Variável explicativa
## Força compressiva   -- numérica -- MPa          -- Target
```

```
## Sumario dos dados
describe(dados)
```

```
##          vars      n   mean      sd median trimmed   mad    min    max
## cimento          1 1030 281.17 104.51 272.90  273.47 117.72 102.00  540.0
## escoria          2 1030  73.90  86.28  22.00   62.43  32.62   0.00  359.4
## cinza            3 1030  54.19  64.00   0.00   46.85   0.00   0.00  200.1
## agua            4 1030 181.57  21.36 185.00  181.19  19.27 121.75  247.0
## super_plastificante 5 1030   6.20   5.97   6.35   5.56   7.87   0.00   32.2
## agregador_grosso   6 1030 972.92  77.75 968.00  973.49  68.64 801.00 1145.0
## agregador_fino     7 1030 773.58  80.18 779.51  776.41  67.44 594.00  992.6
## idade            8 1030  45.66  63.17  28.00   32.53  31.13   1.00  365.0
## forca_compressiva  9 1030  35.82  16.71  34.44   34.96  16.20   2.33   82.6
##
##          range skew kurtosis   se
## cimento    438.00  0.51   -0.53  3.26
## escoria    359.40  0.80   -0.52  2.69
## cinza      200.10  0.54   -1.33  1.99
## agua       125.25  0.07    0.11  0.67
## super_plastificante 32.20  0.91    1.39  0.19
## agregador_grosso  344.00 -0.04   -0.61  2.42
## agregador_fino   398.60 -0.25   -0.11  2.50
## idade        364.00  3.26   12.07  1.97
## forca_compressiva  80.27  0.42   -0.32  0.52
```

```
sum(is.na(data.frame(dados)))
```

```
## [1] 0
```

*# Como podemos notar, temos 1030 observações, 8 variáveis explicativas e nossa variável de interesse (f*  
*# Pela descrição básica dos dados, não temos nenhum dado que parece fugir dos valores esperados (por ex*

## Preparação dos dados

```
## Separando o conjunto de dados em treino e teste
set.seed(2)
inTrain <- createDataPartition(dados$forca_compressiva, p = 7/10)[[1]]
treino <- dados[inTrain,]
teste <- dados[-inTrain,]

## Mantendo casos completos em treino e teste
treino <- treino[complete.cases(treino),]
teste <- teste[complete.cases(teste),]

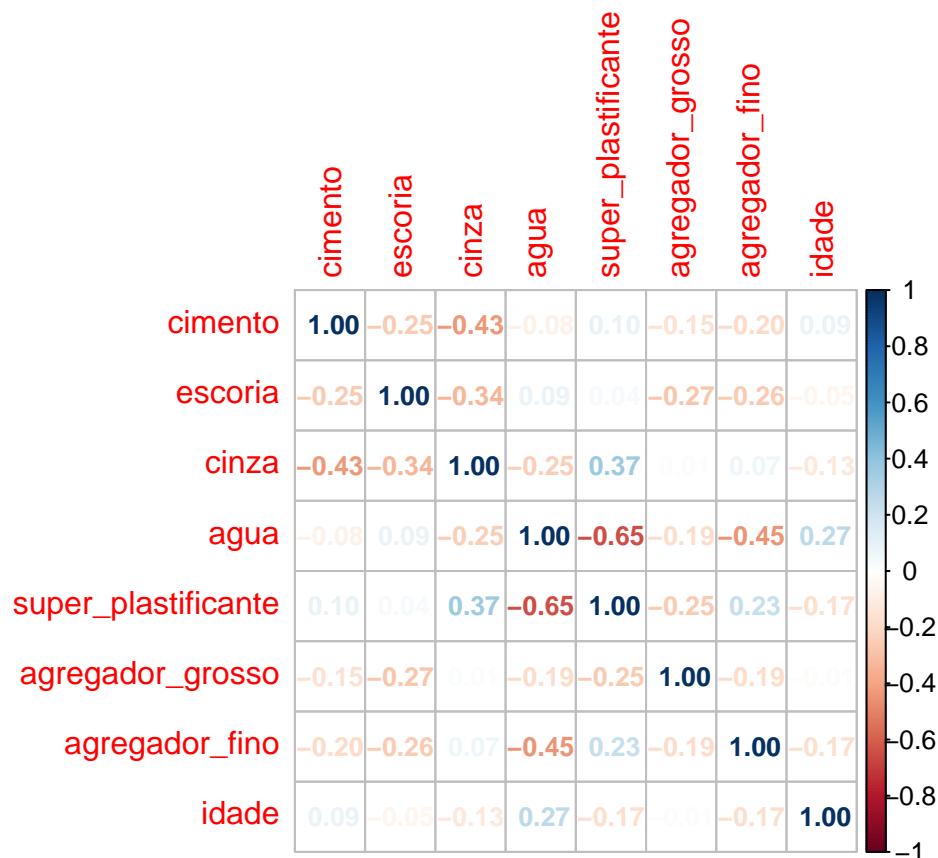
## Separando a variavel resposta, categóricas e numericas
resposta <- treino$forca_compressiva
resposta_teste <- teste$forca_compressiva
```

```
## Removendo a variável resposta
treino <- treino[,-ncol(treino)]
teste <- teste[,-ncol(teste)]

## Retendo as numéricas
Ind_numericas <- colnames(treino)[sapply(treino, is.numeric)]
Ind_categoricas <- colnames(treino)[sapply(treino, function(x) !is.numeric(x))]
numericas <- treino[,Ind_numericas]
categoricas <- treino[,Ind_categoricas]
```

## Redução de dimensionalidade

```
## Analisando as correlações
M <- cor(numericas, use = 'complete.obs')
corrplot(M, method='number', diag = T, number.cex = 0.8)
```



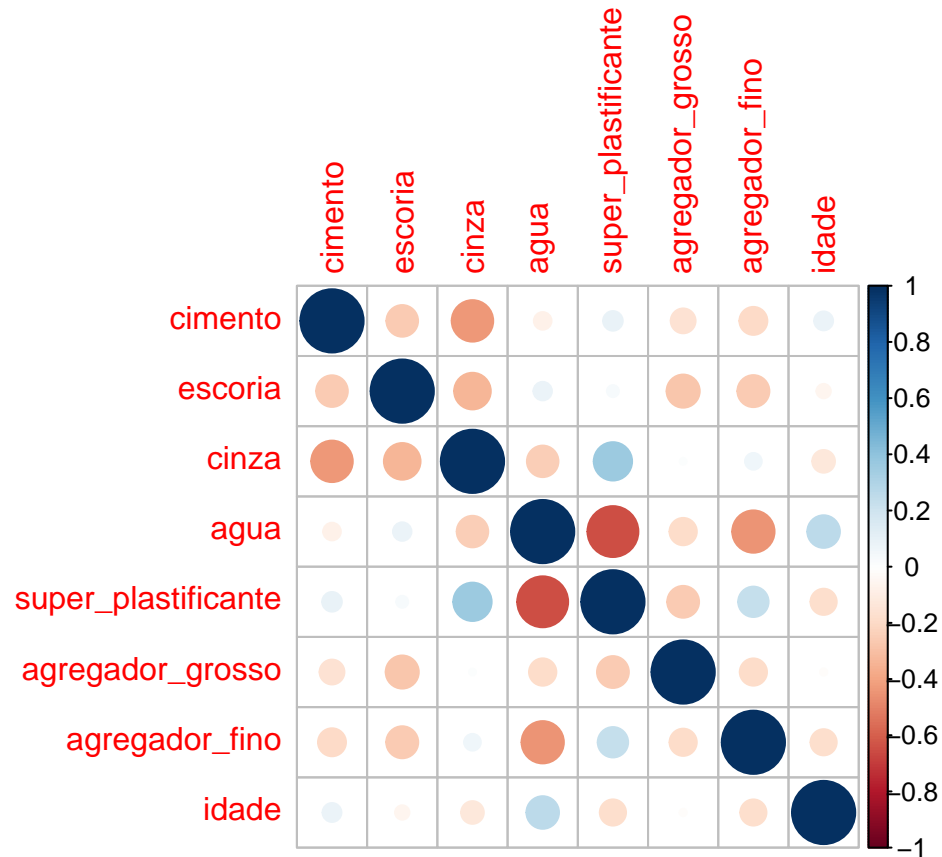
```
summary(M[upper.tri(M)])
```

```
##      Min.  1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.64810 -0.25116 -0.16258 -0.11522  0.04417  0.36742
```

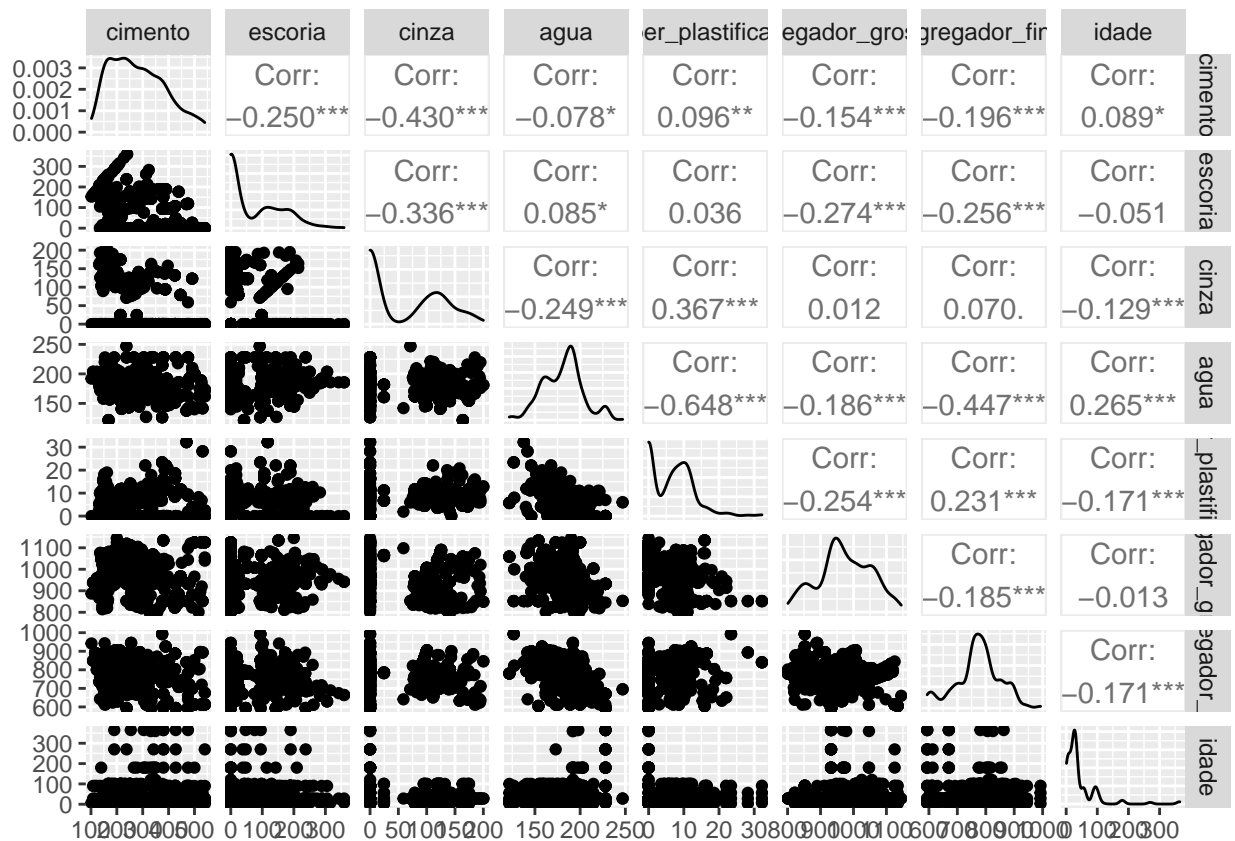
```
## Imprimindo as correlações na forma de círculos
M <- cor(numericas, use = 'complete.obs')
summary(M[upper.tri(M)])
```

```
##      Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## -0.64810 -0.25116 -0.16258 -0.11522  0.04417  0.36742
```

```
corrplot(M, method='circle')
```



```
## Visualizando as correlações
ggpairs(numericas)
```



### Como podemos notar, não temos uma correlação muita alta (pensando em módulo) entre as covariáveis.   
 ### Dependendo do objetivo da análise, se queremos acertar o valor da força compressiva de uma certa b

## Modelagem