Relatório 1 - Regressão

Flavio Margarito Martins de Barros

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)
require(Hmisc)
```

```
## Lendo o banco de dados
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")
```

```
## Sumario dos dados
d <- Hmisc::describe(dados)
```

dados

9 Variables 1030 Observations

cimento

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	280	1	281.2	118.5	143.7	153.5	192.4	272.9	350.0	425.0	480.0

lowest : 102.0 108.3 116.0 122.6 132.0, highest: 522.0 525.0 528.0 531.3 540.0

escoria

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	187	0.907	73.9	91.71	0.0	0.0	0.0	22.0	142.9	192.0	236.0

lowest : 0.00 0.02 11.00 13.61 15.00, highest: 290.20 305.30 316.10 342.10 359.40

cinza

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	163	0.834	54.19	67.08	0.0	0.0	0.0	0.0	118.3	141.1	167.0

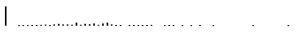
lowest : 0.00 24.46 24.51 24.52 59.00, highest: 194.00 194.90 195.00 200.00 200.10

agua

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	205	0.998	181.6	23.82	146.1	154.6	164.9	185.0	192.0	203.5	228.0

lowest : 121.75 126.60 127.00 127.30 137.80, highest: 228.00 236.70 237.00 246.90 247.00

super_plastificante



n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	155	0.95	6.203	6.426	0.00	0.00	0.00	6.35	10.16	12.21	16.05

```
lowest :  0.00  1.72  1.90  2.00  2.20, highest: 22.00 22.10 23.40 28.20 32.20
```

agregador_grosso

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	284	1	972.9	88.55	842.0	852.1	932.0	968.0	1029.4	1076.5	1104.0

```
lowest :  801.0  801.1  801.4  811.0  814.0, highest: 1124.4 1125.0 1130.0 1134.3 1145.0
```

agregador_fino

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	304	1	773.6	89.87	613.0	664.1	730.9	779.5	824.0	880.8	898.1

```
lowest : 594.0 605.0 611.8 612.0 613.0, highest: 925.7 942.0 943.1 945.0 992.6
```

idade

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	14	0.925	45.66	50.89	3	3	7	28	56	100	180

```
lowest :   1   3   7  14  28, highest: 120 180 270 360 365
```

Value	1	3	7	14	28	56	90	91	100	120	180	270	360	365
Frequency	2	134	126	62	425	91	54	22	52	3	26	13	6	14
Proportion	0.002	0.130	0.122	0.060	0.413	0.088	0.052	0.021	0.050	0.003	0.025	0.013	0.006	0.014

forca_compressiva

n	missing	distinct	Info	Mean	Gmd	.05	.10	.25	.50	.75	.90	.95
1030	0	938	1	35.82	18.92	10.96	14.20	23.71	34.44	46.14	58.82	66.80

```
lowest :  2.331808  3.319827  4.565021  4.782206  4.827711
highest: 79.400056 79.986111 80.199848 81.751169 82.599225
```

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]
categorias <- treino[,Ind_categoricas]
```

Redução de dimensionalidade

Estrutura de correlações

Como são todas variáveis numéricas inicialmente veremos na matriz de correlação se há algumas relação mais forte entre pares de variáveis. Se houver poderemos escolher somente uma das variáveis pois adiconar outra variável fortemente correlacionanda não adicionaria novas informações e traria dificuldades no processo de estimação em virtude de possível multicolinearidade.

```
## Adicionando pacote corrplot
require(corrplot)
```

```
## Carregando pacotes exigidos: corrplot
```

```
## corrplot 0.92 loaded
```

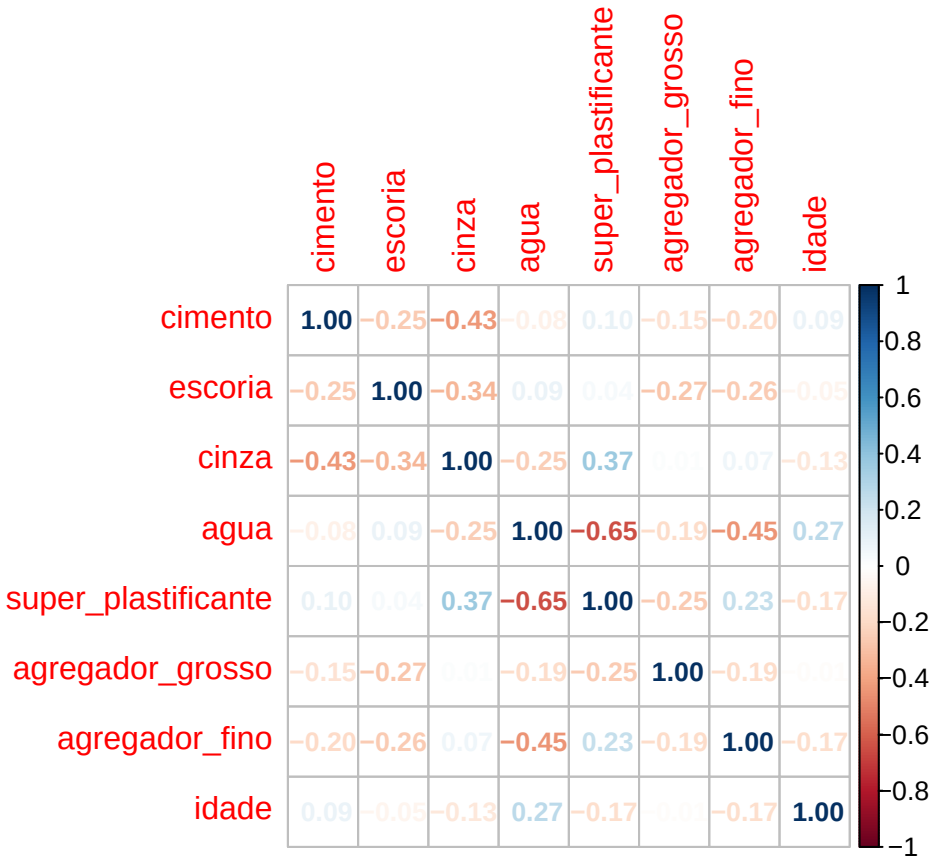
```
require(GGally)
```

```
## Carregando pacotes exigidos: GGally
```

```
## Carregando pacotes exigidos: ggplot2
```

```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

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



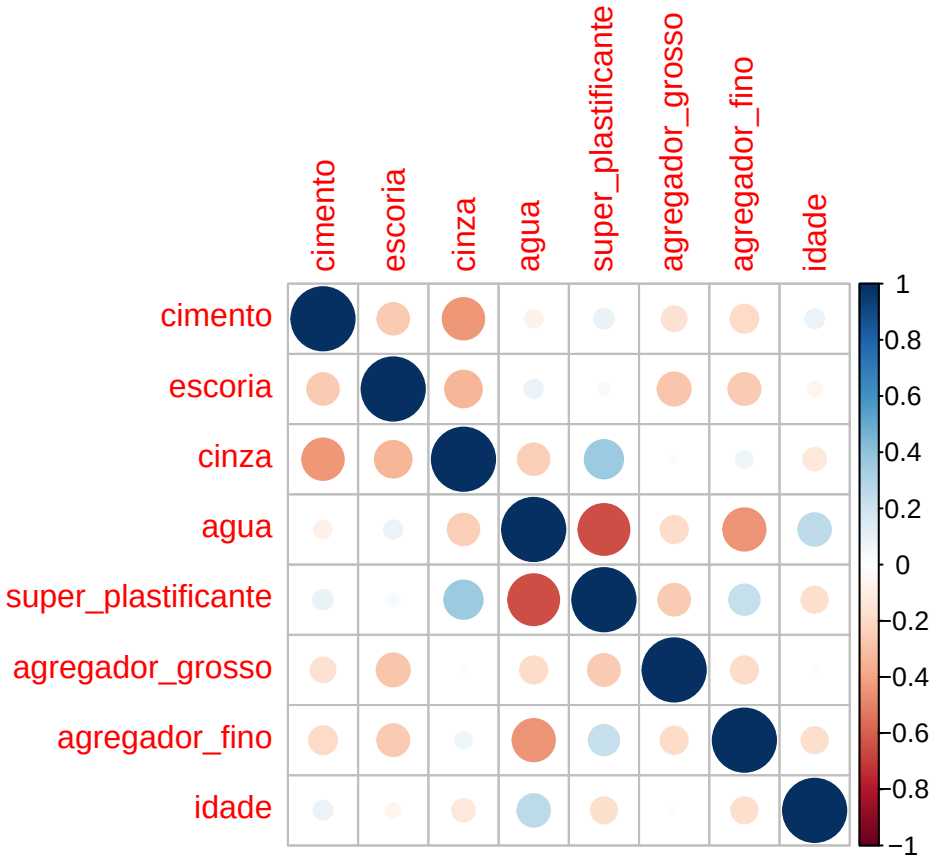
```
summary(M[upper.tri(M)])
```

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

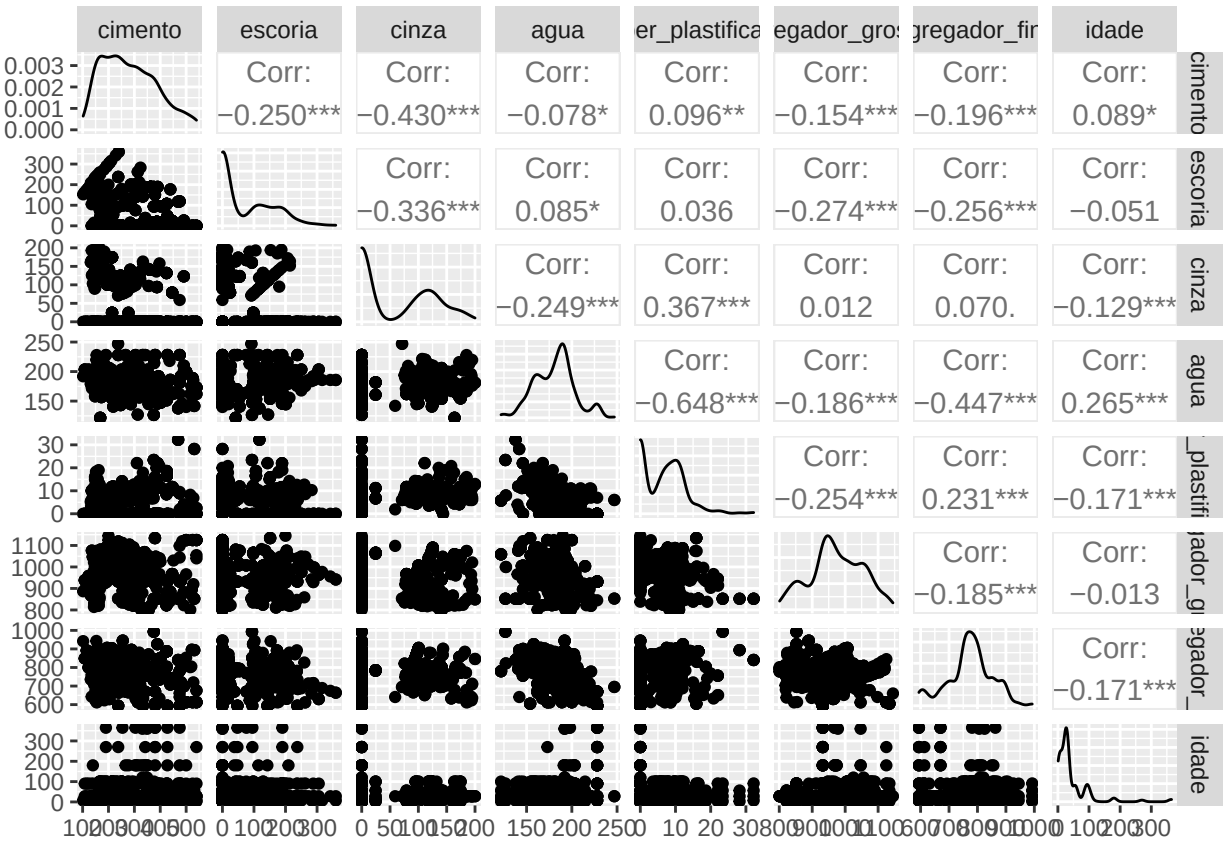
```
## Imprimindo as correlações na forma de circulos
M <- cor(numericas, use = 'complete.obs')
summary(M[upper.tri(M)])
```

```
##      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 as maiores correlações foram de $0,65$, não podemos afirmar que há pares de variáveis redundantes. Portanto optamos por não retirar nenhuma variável nessa etapa.

Análise de redundância

Na análise de redundância utilizamos regressões de cada variável tendo as outras como suas preditoras, inclusive com componentes não lineares via *splines* cúbicos. Essa análise é superior ao correlograma no sentido de que considera não somente as relações lineares dois a dois, mas também a capacidade das preditoras fornecerem informações sobre as outras preditoras de forma conjunta.

```
redun(~ ., r2 = .8, type = "adjusted", data = numericas)
```

```
##
## Redundancy Analysis
##
## redun(formula = ~., data = numericas, r2 = 0.8, type = "adjusted")
##
## n: 722   p: 8    nk: 3
##
## Number of NAs:    0
##
## Transformation of target variables forced to be linear
##
## R-squared cutoff: 0.8    Type: adjusted
##
## R^2 with which each variable can be predicted from all other variables:
##
##             cimento             escoria               cinza                agua
##               0.876               0.863               0.874               0.857
```

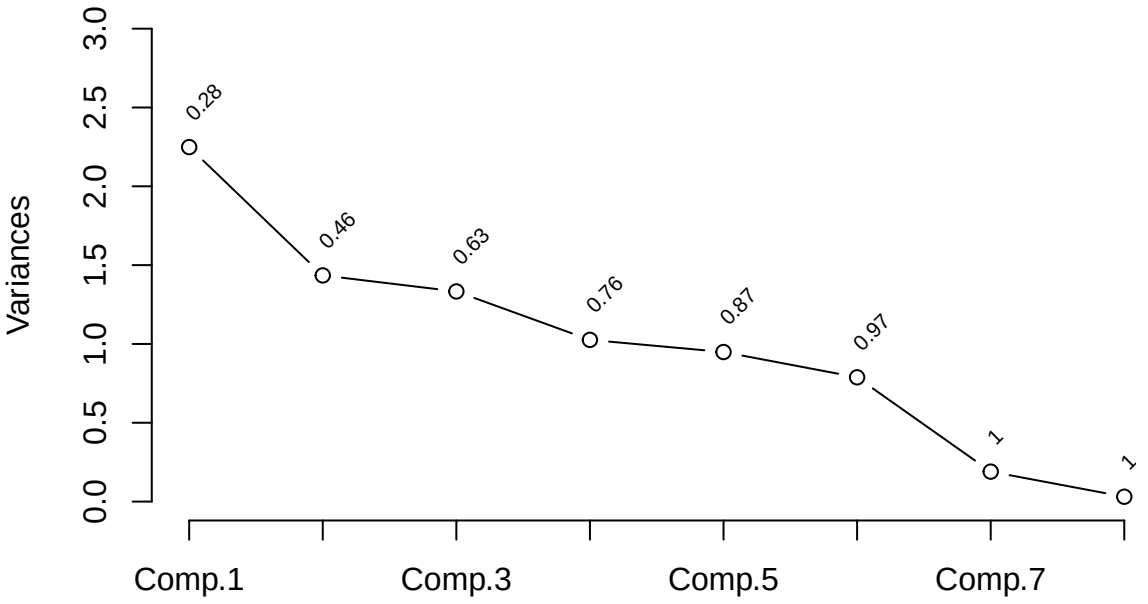
```
## super_plastificante    agregador_grosso      agregador_fino               idade
##               0.669               0.814               0.854               0.161
##
## Rendundant variables:
##
## cimento
##
## Predicted from variables:
##
## escoria cinza agua super_plastificante agregador_grosso agregador_fino idade
##
##   Variable Deleted   R^2 R^2 after later deletions
## 1          cimento 0.876
```

Como resultado dessa análise as variáveis cimento, considerando como critério um $R^2 $ água e cinza podem ser facilmente preditas a partir das outras, portanto serão excluídas.

Estrutura das variáveis com PCA

```
# Calculando o PCA
prin.raw <- princomp (~ ., cor = TRUE , data = numericas)
plot (prin.raw, type = 'lines' , main = ' ' , ylim = c (0 ,3))

# Adicionando a variância cumulativa explicada
addscree <- function (x , npcs = min (10 , length (x$sdev)) ,
                      plotv = FALSE ,
                      col =1 , offset = .8 , adj =0 , pr = FALSE) {
                        vars <- x$sdev^2
                        cumv <- cumsum(vars)/sum(vars)
                        if (pr) print(cumv)
                        text (1: npcs , vars [1: npcs ] + offset*par ('cxy')[2] ,
                        as.character(round(cumv [1: npcs ], 2)),
                        srt =45 , adj = adj , cex = .65 , xpd = NA , col = col)
                        if ( plotv ) lines (1: npcs , vars [1: npcs ], type = ' b ' , col = col )
}
addscree (prin.raw)
```



Não parece haver uma estrutura onde as primeiras componente dominam as outras. Portanto com base nessa análise ainda não teríamos indicação de eliminar variáveis.

Modelagem