Lab08 - Previsão com Regressão Linear

Machine Learning com o R - Análise Macro

Thalles Quinaglia Liduares

2022-08-15

Upload database

```
setwd("C:\\Program Files\\R\\Dados\\ML")

data<-read.csv("insurance.csv", stringsAsFactors = T)

attach(data)
```

Upload packages

```
library(GGally)
library(ggpubr)
library(olsrr)
```

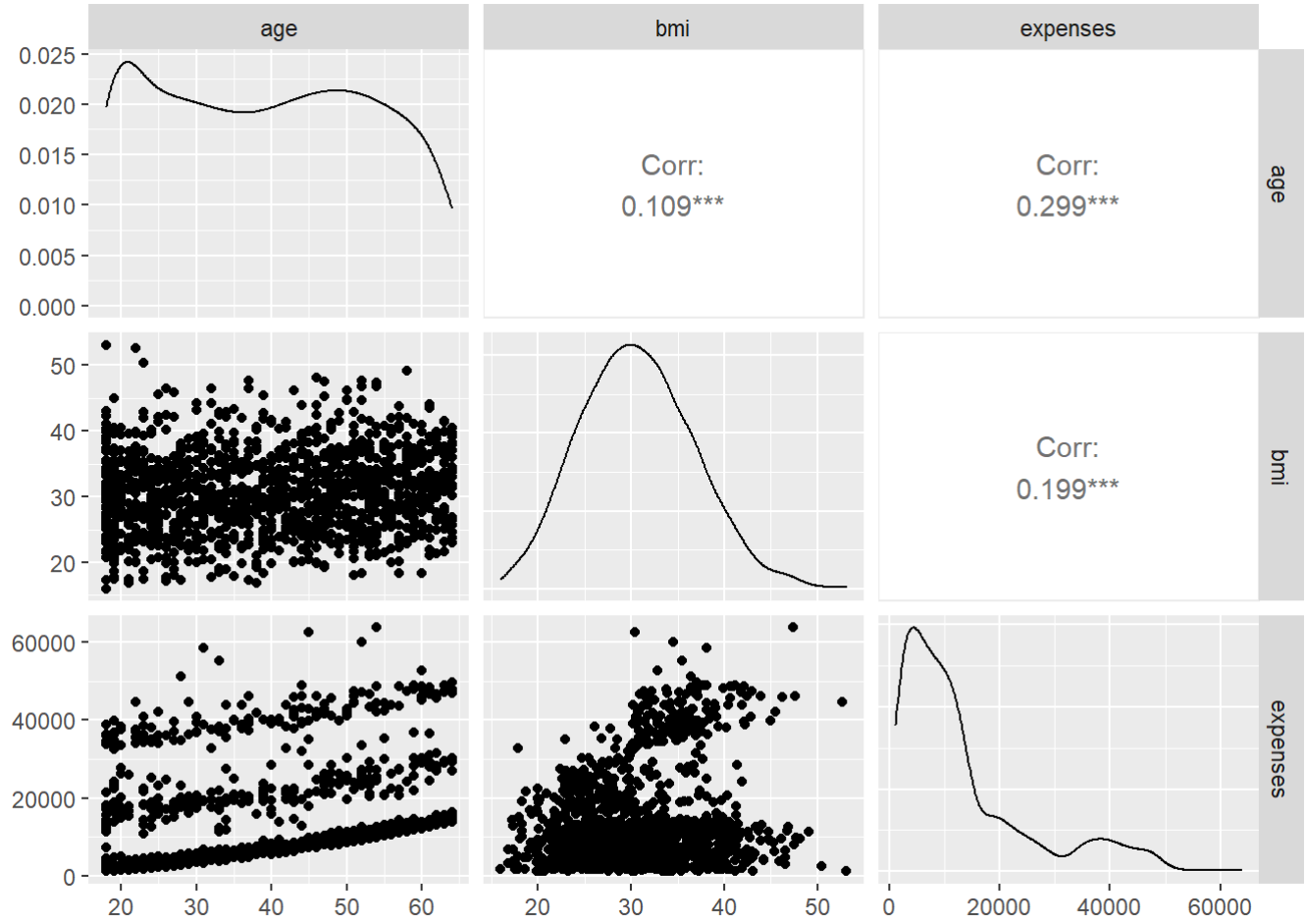
Análise dos dados

```
summary(data)
```

```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:662   Min.   :16.00   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
##  Median :39.00                Median :30.40   Median :1.000
##  Mean   :39.21                Mean   :30.67   Mean   :1.095
##  3rd Qu.:51.00                3rd Qu.:34.70   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.10   Max.   :5.000
##        region       expenses
##  northeast:324   Min.   : 1122
##  northwest:325   1st Qu.: 4740
##  southeast:364   Median : 9382
##  southwest:325   Mean   :13270
##                  3rd Qu.:16640
##                  Max.   :63770
```

Analise de correlação das variaveis numéricas

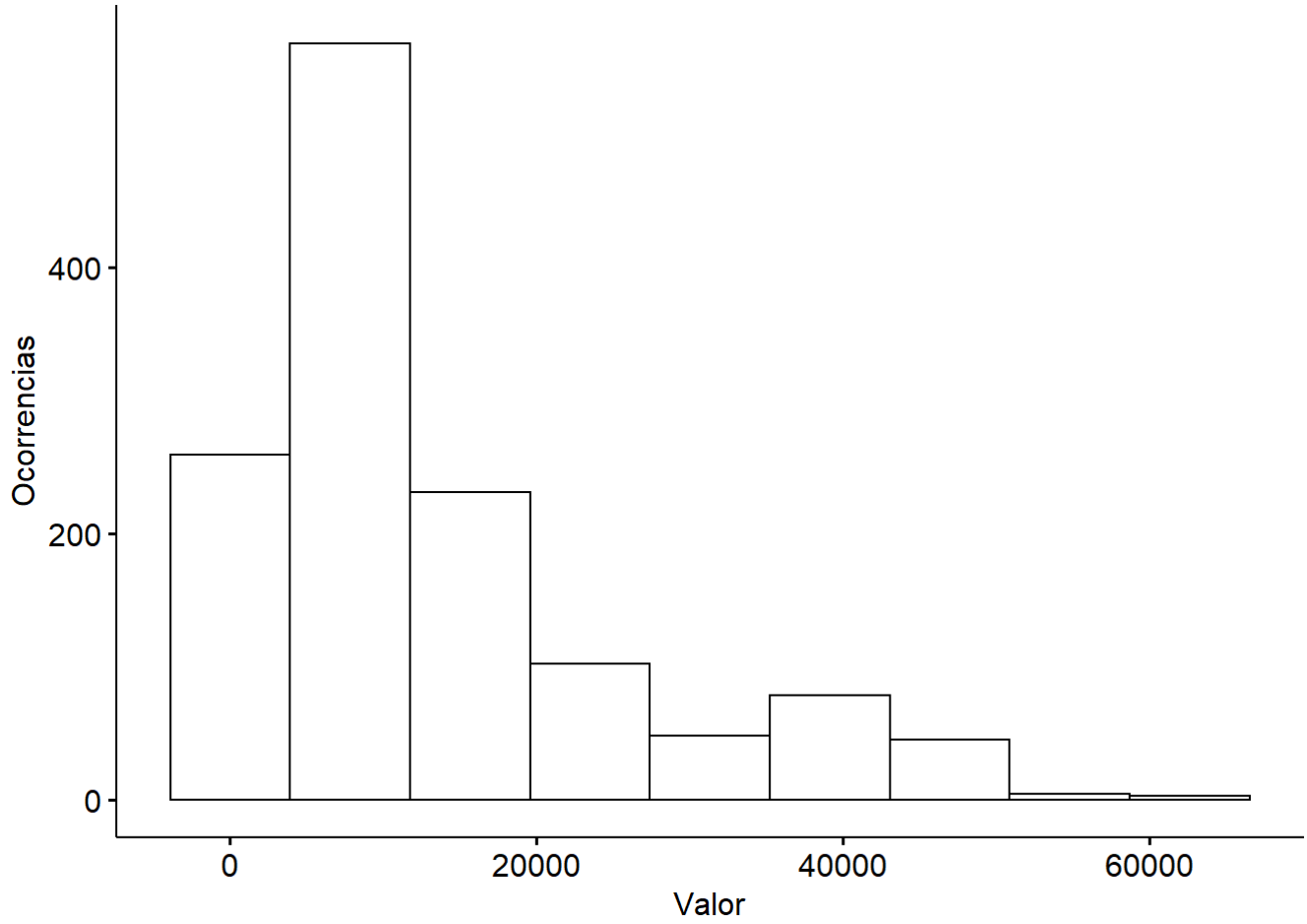
```
data %>%
GGally::ggpairs(c(1,3,7))
```



Histograma dos gastos médicos

```
ht <- gghistogram(data, x = "expenses", bins=9, xlab ="Valor", ylab="Ocorrencias")

ggarrange(ht)
```



Particionando a amostra entre treino e teste

```
set.seed(1608)

part_data<-floor(0.80*nrow(data))

treino_data <-sample(seq_len(nrow(data)), size = part_data)

treino<-data[treino_data, ]

teste<-data[-treino_data,]
```

Modelo com todas variaveis do dataset

```
lm1<-lm(expenses~.,treino)

summary(lm1)
```

```
##
## Call:
## lm(formula = expenses ~ ., data = treino)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -10798.1  -2945.8   -965.3   1517.9  30262.5
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     -12280.17    1098.71 -11.177  < 2e-16 ***
## age                246.63      13.41  18.394  < 2e-16 ***
## sexmale           -340.10     374.91  -0.907 0.364533
## bmi                361.28      32.23  11.210  < 2e-16 ***
## children           525.12     152.01   3.454 0.000573 ***
## smokeryes        23765.07     463.55  51.268  < 2e-16 ***
## regionnorthwest     87.14     543.77   0.160 0.872711
## regionsoutheast   -895.74     535.82  -1.672 0.094873 .
## regionsouthwest   -920.69     541.71  -1.700 0.089498 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6074 on 1061 degrees of freedom
## Multiple R-squared:  0.7492, Adjusted R-squared:  0.7473
## F-statistic: 396.2 on 8 and 1061 DF,  p-value: < 2.2e-16
```

Previsão e acurácia do modelo

```
teste$prev1<-predict(lm1, newdata=teste)

teste$prev_acc<-round(teste$prev1/teste$expenses,2)

teste$prev_acc<-teste$prev_acc-1

teste$prev_acc<-abs(teste$prev_acc)

summary(teste$prev_acc)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##  0.0000  0.1700  0.3150  0.4944  0.6100  7.6900
```

Como há um valor minimo igual a 0, houve pelo menos um caso onde o modelo previu com 100% de acurácia o valor dos gastos médicos.

Porcentagem do erro médio absoluto

```
erro_médio<-round(mean(abs(lm1$residuals)/treino$expenses)*100,2)

erro_médio
```

```
## [1] 42.65
```

Teste de normalidade dos resíduos

```
ols_test_normality(lm1)
```

```
## Warning in ks.test.default(y, "pnorm", mean(y), sd(y)): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9044         0.0000
## Kolmogorov-Smirnov        0.157          0.0000
## Cramer-von Mises         104.2405        0.0000
## Anderson-Darling         32.6613         0.0000
## -----------------------------------------------
```

Com base no teste de Kolmogorov-Smirnov, rejeita-se a hipótese nula de que os residuos seguem uma distribuição normal. Portanto, a inferência estatistisca do modelo fica prejudicada.

Versão 2: Sem inclusão da variável `region`

```
lm2<-lm(expenses~.-region,treino)
```

```
ols_test_normality(lm2)
```

```
## Warning in ks.test.default(y, "pnorm", mean(y), sd(y)): ties should not be
## present for the Kolmogorov-Smirnov test
```

```
## -----------------------------------------------
##        Test             Statistic       pvalue
## -----------------------------------------------
## Shapiro-Wilk              0.9048         0.0000
## Kolmogorov-Smirnov        0.1585         0.0000
## Cramer-von Mises         104.2405        0.0000
## Anderson-Darling         32.4498         0.0000
## -----------------------------------------------
```

Os resíduos permanecem não seguindo uma distribuição normal. Portanto, recomenda-se cautela na interpretação dos resultados obtidos.