workshop5

juan cepeda- edward romero- carlos bejarano

2023-04-20

#MODELOI MEJORADO

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.2 --
## v ggplot2 3.4.1     v purrr   1.0.1
## v tibble  3.1.8     v dplyr   1.1.0
## v tidyr   1.3.0     v stringr 1.5.0
## v readr   2.1.3     v forcats 1.0.0
## -- Conflicts ------------------------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
## get parent script folder
folder <- dirname(rstudioapi::getSourceEditorContext()$path)
parentFolder <- dirname(folder)

insurance <- read.csv(("C:/Users/USUARIO/Downloads/macine_learning/datasets/insurance.csv")
                      , stringsAsFactors = TRUE)

#obtenemos la variable de edad al cuadrado
insurance$age2 <- insurance$age^2
#otraa forma de hacer para tener la edad al cuadrado
#insurancem<-mutate(insurance,age2
#              (age^2) )#nuevo dataset que contiene la variable edad al cuadrado

# volvemos la variable BMI en binario para saber si la persona es o no obesa
#despues de 30 cuenta como una persona obesa
insurance$bmi30 <- ifelse(insurance$bmi >= 30, 1, 0)
#clasificamos si una persona es fumadora
insurance$smokeryes<-ifelse(insurance=='yes', 1, 0)
#insurance$bmi30*smoker<-insurance$bmi30 + smokeryes + bmi30:smokeryes
#agregamos interaccion de Bmi30 y esmokeryes


colnames(insurance)
```

```
##  [1] "age"       "sex"       "bmi"       "children"  "smoker"    "region"
##  [7] "charges"   "age2"      "bmi30"     "smokeryes"
```

```
summary(insurance)
```

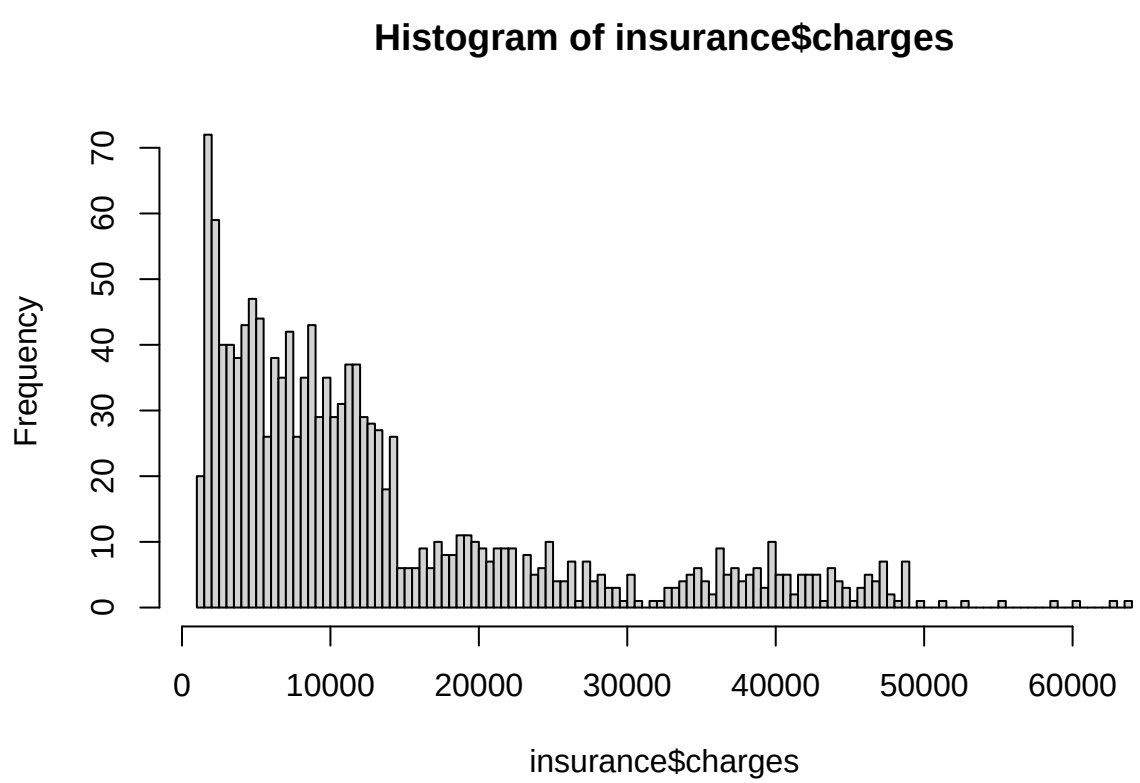
```
##       age            sex           bmi           children     smoker
##  Min.   :18.00   female:662   Min.   :15.96   Min.   :0.000   no :1064
##  1st Qu.:27.00   male  :676   1st Qu.:26.30   1st Qu.:0.000   yes: 274
```

```
##  Median :39.00                Median :30.40   Median :1.000
##  Mean   :39.21                Mean   :30.66   Mean   :1.095
##  3rd Qu.:51.00                3rd Qu.:34.69   3rd Qu.:2.000
##  Max.   :64.00                Max.   :53.13   Max.   :5.000
##        region       charges           age2          bmi30
##  northeast:324   Min.   : 1122   Min.   : 324   Min.   :0.0000
##  northwest:325   1st Qu.: 4740   1st Qu.: 729   1st Qu.:0.0000
##  southeast:364   Median : 9382   Median :1521   Median :1.0000
##  southwest:325   Mean   :13270   Mean   :1734   Mean   :0.5284
##                  3rd Qu.:16640   3rd Qu.:2601   3rd Qu.:1.0000
##                  Max.   :63770   Max.   :4096   Max.   :1.0000
##     smokeryes.age        smokeryes.sex        smokeryes.bmi     smokeryes.children    smokeryes.smok
##  Min.   :0            Min.   :0            Min.   :0            Min.   :0            Min.   :0.00000
##  1st Qu.:0            1st Qu.:0            1st Qu.:0            1st Qu.:0            1st Qu.:0.00000
##  Median :0            Median :0            Median :0            Median :0            Median :0.00000
##  Mean   :0            Mean   :0            Mean   :0            Mean   :0            Mean   :0.20478
##  3rd Qu.:0            3rd Qu.:0            3rd Qu.:0            3rd Qu.:0            3rd Qu.:0.00000
##  Max.   :0            Max.   :0            Max.   :0            Max.   :0            Max.   :1.00000
```

```
summary(insurance$charges)
```

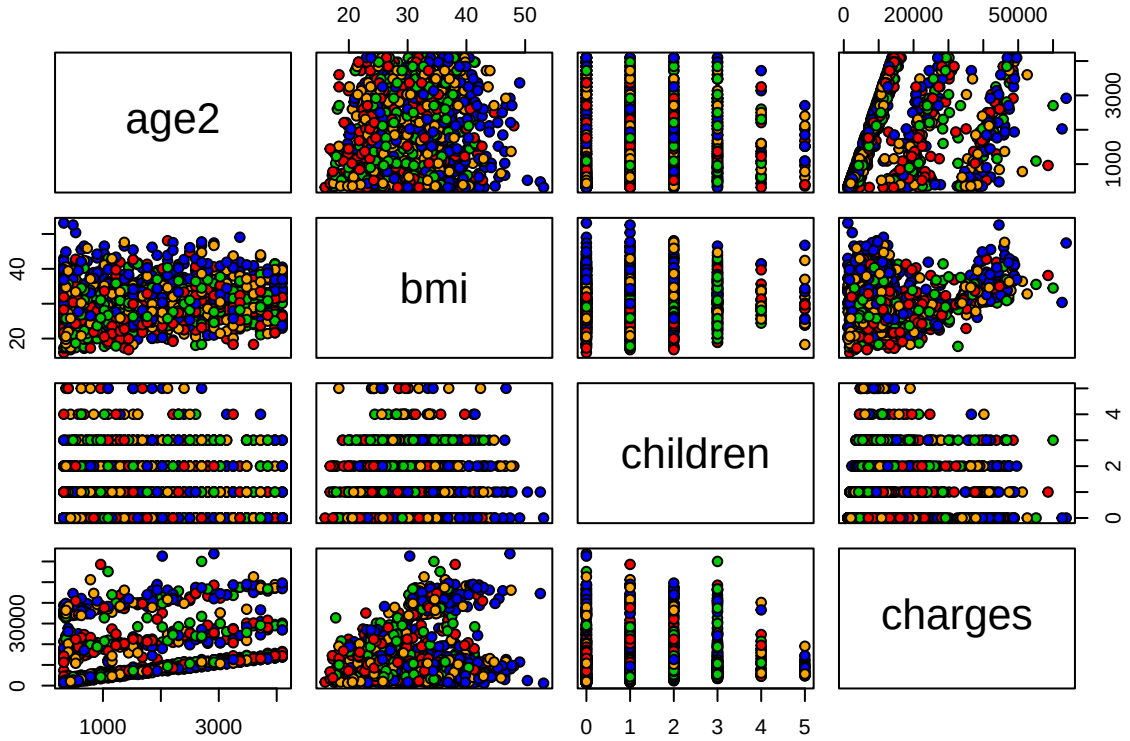
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##    1122    4740    9382   13270   16640   63770
```

```
hist(insurance$charges, breaks=100)
```




```
pairs(insurance[c("age2", "bmi", "children", "charges")]
```

```
      ,pch=21, bg=c("red","green3","blue", "orange")[unclass(insurance$region)])
```




```
if(!require(psych))
  install.packages("psych")
```

```
## Loading required package: psych
```

```
## Warning: package 'psych' was built under R version 4.2.3
```

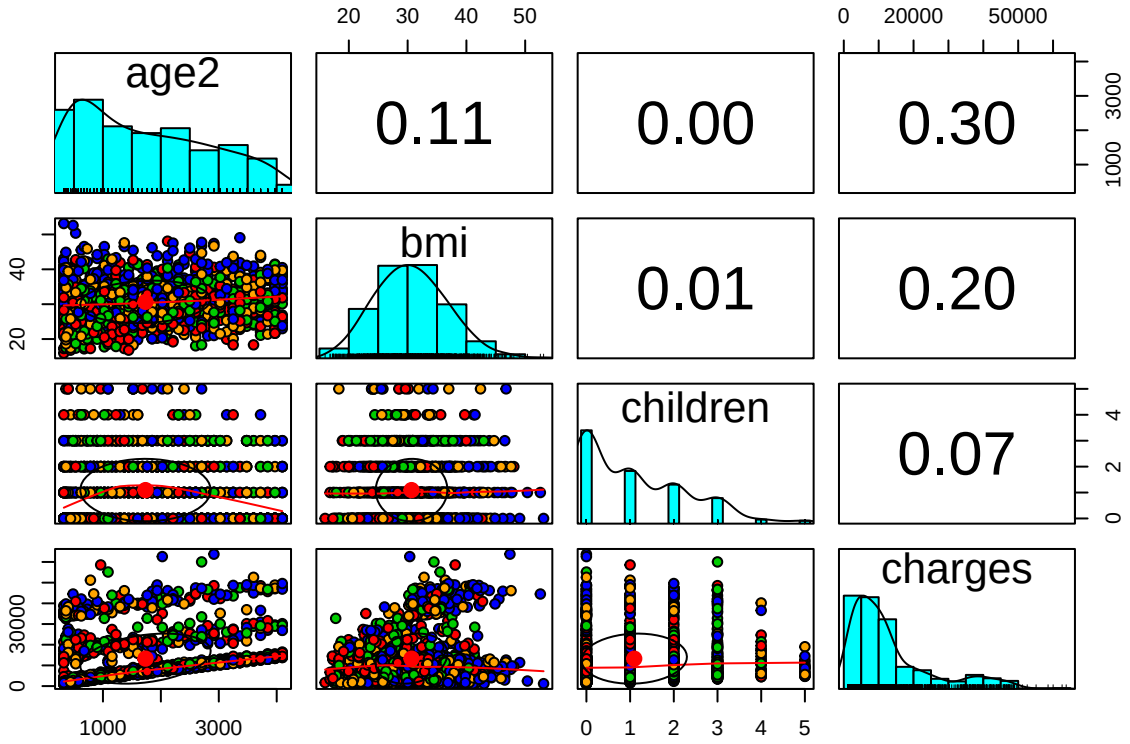
```
##
## Attaching package: 'psych'
##
## The following objects are masked from 'package:ggplot2':
##
##     %+%, alpha
```

```
library(psych)
pairs.panels(insurance[c("age2",

                         "bmi"

                         , "children", "charges")]

             ,pch=21, bg=c("red","green3","blue", "orange")[unclass(insurance$region)])
```




```
#predictors
#set.seed(1)
predictors <- colnames(insurance)[-7]#todas la variables predictores menos la variable objetivo charges
sample.index <- sample(1:nrow(insurance)#recorre todos los datos(1338) y le asigna un numero
                       ,nrow(insurance)*0.7#sacamos el  70% de todos los numeros
                       ,replace = F)#no se reeemplazan los datos

train.data <- insurance[sample.index,c(predictors,"charges"),drop=F]#filas de sample.index,todas las co
test.data <- insurance[-sample.index,c(predictors,"charges"),drop=F]#todas las muestras que sean el sam



# as we are using all variables, we can write ~ .
#añadimos a la formula el efecto de interaccion
ins_model <- lm(charges ~ bmi30 + smokeryes + bmi30:smokeryes, data = train.data)#creamos modelo predic

#tenemos al final un modelo de prediccion un poco mas acertado con un 78% con el modelo de #smokeryes y
#bmi tiene un costo por cambio de unidad de 1151 y esmokeryes de 13619 cuesta mas un #ciudadano que fum

#ins_model <- lm(charges ~ bmi30 + smokeryes + bmi30:smokeryes, data = train.data)
ins_model
```

```
##
## Call:
## lm(formula = charges ~ bmi30 + smokeryes + bmi30:smokeryes, data = train.data)
##
```

```
## Coefficients:
##             (Intercept)                   bmi30             smokeryesage
##                  7858.8                   993.9                      NA
##            smokeryessex            smokeryesbmi        smokeryeschildren
##                      NA                      NA                      NA
##         smokeryessmoker         smokeryesregion         smokeryescharges
##                 13557.3                      NA                      NA
##           smokeryesage2          smokeryesbmi30       bmi30:smokeryesage
##                      NA                      NA                      NA
##      bmi30:smokeryessex      bmi30:smokeryesbmi  bmi30:smokeryeschildren
##                      NA                      NA                      NA
##   bmi30:smokeryessmoker   bmi30:smokeryesregion   bmi30:smokeryescharges
##                 19136.2                      NA                      NA
##     bmi30:smokeryesage2     bmi30:smokeryesbmi30
##                      NA                      NA
```

```
summary(ins_model)
```

```
##
## Call:
## lm(formula = charges ~ bmi30 + smokeryes + bmi30:smokeryes, data = train.data)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -19402  -4390  -1068   2961  28058
##
## Coefficients: (16 not defined because of singularities)
##                          Estimate Std. Error t value Pr(>|t|)
## (Intercept)               7858.8      319.9  24.563   <2e-16 ***
## bmi30                      993.9      442.5   2.246   0.0249 *
## smokeryesage                  NA         NA      NA       NA
## smokeryessex                  NA         NA      NA       NA
## smokeryesbmi                  NA         NA      NA       NA
## smokeryeschildren             NA         NA      NA       NA
## smokeryessmoker          13557.3      681.8  19.885   <2e-16 ***
## smokeryesregion               NA         NA      NA       NA
## smokeryescharges              NA         NA      NA       NA
## smokeryesage2                 NA         NA      NA       NA
## smokeryesbmi30                NA         NA      NA       NA
## bmi30:smokeryesage            NA         NA      NA       NA
## bmi30:smokeryessex            NA         NA      NA       NA
## bmi30:smokeryesbmi            NA         NA      NA       NA
## bmi30:smokeryeschildren       NA         NA      NA       NA
## bmi30:smokeryessmoker    19136.2      937.2  20.419   <2e-16 ***
## bmi30:smokeryesregion         NA         NA      NA       NA
## bmi30:smokeryescharges        NA         NA      NA       NA
## bmi30:smokeryesage2           NA         NA      NA       NA
## bmi30:smokeryesbmi30          NA         NA      NA       NA
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5960 on 932 degrees of freedom
## Multiple R-squared:  0.7729, Adjusted R-squared:  0.7722
## F-statistic:  1057 on 3 and 932 DF,  p-value: < 2.2e-16
```

```r
# al aumentar la edad al cuadrado no predice bien los cargos por cambio de unidad de edad en mis coefic

prediction <- predict(ins_model, test.data)
```

```
## Warning in predict.lm(ins_model, test.data): prediction from a rank-deficient
## fit may be misleading
```

```r
#calculate RMSE
RMSE.df = data.frame(predicted = prediction, actual = test.data$charges,

                     SE = sqrt((prediction - test.data$charges)^2))

head(RMSE.df)
```

```
##    predicted    actual         SE
## 5   7858.771  3866.855  3991.9162
## 7   8852.652  8240.590   612.0619
## 10  7858.771 28923.137 21064.3655
## 12 21416.120 27808.725  6392.6046
## 15 41546.251 39611.758  1934.4929
## 23  8852.652  1137.011  7715.6405
```

```r
sum(RMSE.df$SE)/nrow(RMSE.df)
```

```
## [1] 4272.787
```