A3 Modelización Predictiva

Francisco Javier Melchor González

7/12/2020

Contents

#1. Datos y Estadística descriptiva ##1.1 Lectura de datos

```
house_filepath <- "../Data/house.csv"
house <- read.csv(file=house_filepath, header=TRUE, sep=";", na.strings=c(""," ","NA"))
head(house)
```

```
##   price resid_area air_qual room_num   age dist1 dist2 dist3 dist4 teachers
## 1     5      48.10    0.693    5.453 100.0  1.57  1.26  1.79  1.34     19.8
## 2    12      48.10    0.614    5.304  97.3  2.28  1.99  2.41  1.73     19.8
## 3    14      51.89    0.624    6.174  93.6  1.86  1.54  1.87  1.18     18.8
## 4    18      51.89    0.624    6.431  98.8  1.96  1.61  1.92  1.77     18.8
## 5    19      35.19    0.515    5.985  45.4  4.89  4.64  5.05  4.67     19.8
## 6    20      35.96    0.499    5.841  61.4  3.39  3.28  3.62  3.22     20.8
##   poor_prop airport n_hos_beds n_hot_rooms waterbody rainfall bus_ter
## 1     30.59      NO       9.30      13.040      Lake       26     YES
## 2     24.91      NO       9.34      15.096      Lake       39     YES
## 3     24.16      NO       5.68      10.112      Lake       28     YES
## 4     15.39      NO       8.16      14.144      None       41     YES
## 5      9.74      NO       6.38      11.152      Lake       28     YES
## 6     11.41      NO       7.50      15.160      None       39     YES
##        parks Sold
## 1 0.06525315    0
## 2 0.06192155    0
## 3 0.05697699    0
## 4 0.05636501    0
## 5 0.04769962    0
## 6 0.04535682    0
```

```
str(house)
```

```
## 'data.frame':    506 obs. of  19 variables:
##  $ price      : num  5 12 14 18 19 20 20 20 21 21 ...
##  $ resid_area : num  48.1 48.1 51.9 51.9 35.2 ...
##  $ air_qual   : num  0.693 0.614 0.624 0.624 0.515 0.499 0.437 0.489 0.538 0.544 ...
##  $ room_num   : num  5.45 5.3 6.17 6.43 5.99 ...
##  $ age        : num  100 97.3 93.6 98.8 45.4 61.4 74.5 100 87.3 58.8 ...
##  $ dist1      : num  1.57 2.28 1.86 1.96 4.89 3.39 4.33 3.95 4.53 4.07 ...
##  $ dist2      : num  1.26 1.99 1.54 1.61 4.64 3.28 3.72 3.86 3.94 3.86 ...
##  $ dist3      : num  1.79 2.41 1.87 1.92 5.05 3.62 4.26 4.14 4.36 4.24 ...
##  $ dist4      : num  1.34 1.73 1.18 1.77 4.67 3.22 3.9 3.55 4.13 3.84 ...
##  $ teachers   : num  19.8 19.8 18.8 18.8 19.8 20.8 21.3 21.4 19 21.6 ...
##  $ poor_prop  : num  30.59 24.91 24.16 15.39 9.74 ...
##  $ airport    : chr  "NO" "NO" "NO" "NO" ...
##  $ n_hos_beds : num  9.3 9.34 5.68 8.16 6.38 ...
##  $ n_hot_rooms: num  13 15.1 10.1 14.1 11.2 ...
##  $ waterbody  : chr  "Lake" "Lake" "Lake" "None" ...
##  $ rainfall   : int  26 39 28 41 28 39 22 60 50 36 ...
##  $ bus_ter    : chr  "YES" "YES" "YES" "YES" ...
##  $ parks      : num  0.0653 0.0619 0.057 0.0564 0.0477 ...
##  $ Sold       : int  0 0 0 0 0 0 0 0 0 0 ...
```

```
house$airport <- as.factor(house$airport)
house$waterbody <- as.factor(house$waterbody)
house$bus_ter <- as.factor(house$bus_ter)
house$Sold <- as.factor(house$Sold)

str(house)
```

```
## 'data.frame':    506 obs. of  19 variables:
##  $ price      : num  5 12 14 18 19 20 20 20 21 21 ...
##  $ resid_area : num  48.1 48.1 51.9 51.9 35.2 ...
##  $ air_qual   : num  0.693 0.614 0.624 0.624 0.515 0.499 0.437 0.489 0.538 0.544 ...
##  $ room_num   : num  5.45 5.3 6.17 6.43 5.99 ...
##  $ age        : num  100 97.3 93.6 98.8 45.4 61.4 74.5 100 87.3 58.8 ...
##  $ dist1      : num  1.57 2.28 1.86 1.96 4.89 3.39 4.33 3.95 4.53 4.07 ...
##  $ dist2      : num  1.26 1.99 1.54 1.61 4.64 3.28 3.72 3.86 3.94 3.86 ...
##  $ dist3      : num  1.79 2.41 1.87 1.92 5.05 3.62 4.26 4.14 4.36 4.24 ...
##  $ dist4      : num  1.34 1.73 1.18 1.77 4.67 3.22 3.9 3.55 4.13 3.84 ...
##  $ teachers   : num  19.8 19.8 18.8 18.8 19.8 20.8 21.3 21.4 19 21.6 ...
##  $ poor_prop  : num  30.59 24.91 24.16 15.39 9.74 ...
##  $ airport    : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ n_hos_beds : num  9.3 9.34 5.68 8.16 6.38 ...
##  $ n_hot_rooms: num  13 15.1 10.1 14.1 11.2 ...
##  $ waterbody  : Factor w/ 4 levels "Lake","Lake and River",..: 1 1 1 3 1 3 3 3 4 2 ...
##  $ rainfall   : int  26 39 28 41 28 39 22 60 50 36 ...
##  $ bus_ter    : Factor w/ 1 level "YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ parks      : num  0.0653 0.0619 0.057 0.0564 0.0477 ...
##  $ Sold       : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

```
##1.2 Descriptiva y visualizaci?n
colSums(is.na(house))
```

```
##       price  resid_area    air_qual    room_num         age       dist1
##           0           0           0           0           0           0
##       dist2       dist3       dist4    teachers   poor_prop     airport
##           0           0           0           0           0           0
##  n_hos_beds n_hot_rooms   waterbody    rainfall     bus_ter       parks
##           8           0           0           0           0           0
##        Sold
##           0
```

```
factors = unlist(lapply(house, is.factor))
which(factors, arr.ind = TRUE)
```

```
##   airport waterbody   bus_ter      Sold
##        12        15        17        19
```

```
levels(house$airport)
```

```
## [1] "NO"  "YES"
```

```
levels(house$waterbody)
```

```
## [1] "Lake"           "Lake and River" "None"           "River"
```

```
levels(house$bus_ter)
```

```
## [1] "YES"
```

```
levels(house$Sold)
```

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

```
par(mfrow=c(2,2))

counts <- table(house$waterbody)
barplot(counts, main="Distribuci?n de tipos de fuente natural de agua dulce
        que hay en la ciudad", xlab="N?mero de fuentes por cada categor?a",
        col = rainbow (length(levels(house$waterbody))))

colorForPieCharts = rainbow(length(levels(house$airport)) +
                              length(levels(house$bus_ter)) +
                              length(levels(house$Sold)))

levels(house$airport)
```

```
## [1] "NO"  "YES"
```

```
mytableAirport <- table(house$airport)
pctAirport <- round(mytableAirport/sum(mytableAirport)*100)
lblsAirport <- paste(names(mytableAirport), "\n", pctAirport, sep="")
lblsAirport <- paste (lblsAirport, '%', sep="")
pie(mytableAirport, labels = lblsAirport, col=colorForPieCharts[1:2],
    main="Pie Chart of Airport\n")


levels(house$bus_ter)
```

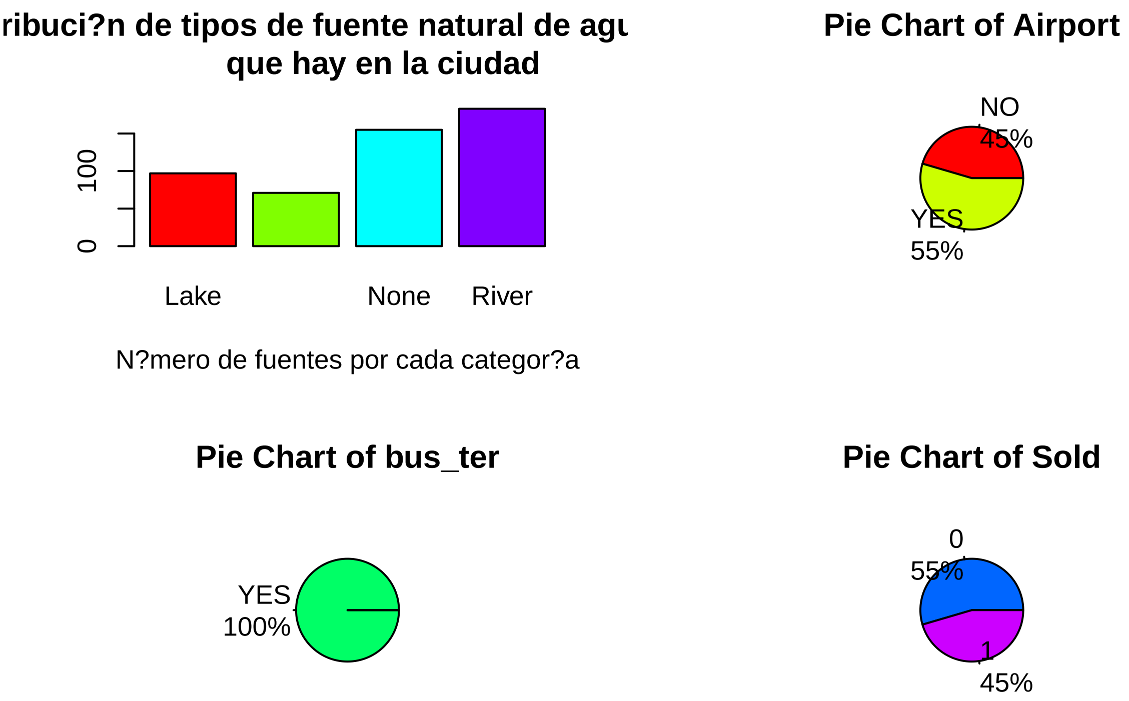
```
## [1] "YES"
```

```
mytableBus_ter <- table(house$bus_ter)
pctBus_ter <- round(mytableBus_ter/sum(mytableBus_ter)*100)
lblsBus_ter <- paste(names(mytableBus_ter), "\n", pctBus_ter, sep="")
lblsBus_ter <- paste (lblsBus_ter, '%', sep="")
pie(mytableBus_ter, labels = lblsBus_ter, col=colorForPieCharts[3:3],
    main="Pie Chart of bus_ter\n")

levels(house$Sold)
```

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

```
mytableSold <- table(house$Sold)
pctSold <- round(mytableSold/sum(mytableSold)*100)
lblsSold <- paste(names(mytableSold), "\n", pctSold, sep="")
```

```r
lblsSold <- paste (lblsSold, '%', sep="")
pie(mytableSold, labels = lblsSold, col=colorForPieCharts[4:5],
    main="Pie Chart of Sold\n")
```



```r
str(house)
```

```
## 'data.frame':    506 obs. of  19 variables:
##  $ price       : num  5 12 14 18 19 20 20 20 21 21 ...
##  $ resid_area  : num  48.1 48.1 51.9 51.9 35.2 ...
##  $ air_qual    : num  0.693 0.614 0.624 0.624 0.515 0.499 0.437 0.489 0.538 0.544 ...
##  $ room_num    : num  5.45 5.3 6.17 6.43 5.99 ...
##  $ age         : num  100 97.3 93.6 98.8 45.4 61.4 74.5 100 87.3 58.8 ...
##  $ dist1       : num  1.57 2.28 1.86 1.96 4.89 3.39 4.33 3.95 4.53 4.07 ...
##  $ dist2       : num  1.26 1.99 1.54 1.61 4.64 3.28 3.72 3.86 3.94 3.86 ...
##  $ dist3       : num  1.79 2.41 1.87 1.92 5.05 3.62 4.26 4.14 4.36 4.24 ...
##  $ dist4       : num  1.34 1.73 1.18 1.77 4.67 3.22 3.9 3.55 4.13 3.84 ...
##  $ teachers    : num  19.8 19.8 18.8 18.8 19.8 20.8 21.3 21.4 19 21.6 ...
##  $ poor_prop   : num  30.59 24.91 24.16 15.39 9.74 ...
##  $ airport     : Factor w/ 2 levels "NO","YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ n_hos_beds  : num  9.3 9.34 5.68 8.16 6.38 ...
##  $ n_hot_rooms: num  13 15.1 10.1 14.1 11.2 ...
##  $ waterbody   : Factor w/ 4 levels "Lake","Lake and River",..: 1 1 1 3 1 3 3 3 4 2 ...
##  $ rainfall    : int  26 39 28 41 28 39 22 60 50 36 ...
##  $ bus_ter     : Factor w/ 1 level "YES": 1 1 1 1 1 1 1 1 1 1 ...
##  $ parks       : num  0.0653 0.0619 0.057 0.0564 0.0477 ...
##  $ Sold        : Factor w/ 2 levels "0","1": 1 1 1 1 1 1 1 1 1 1 ...
```

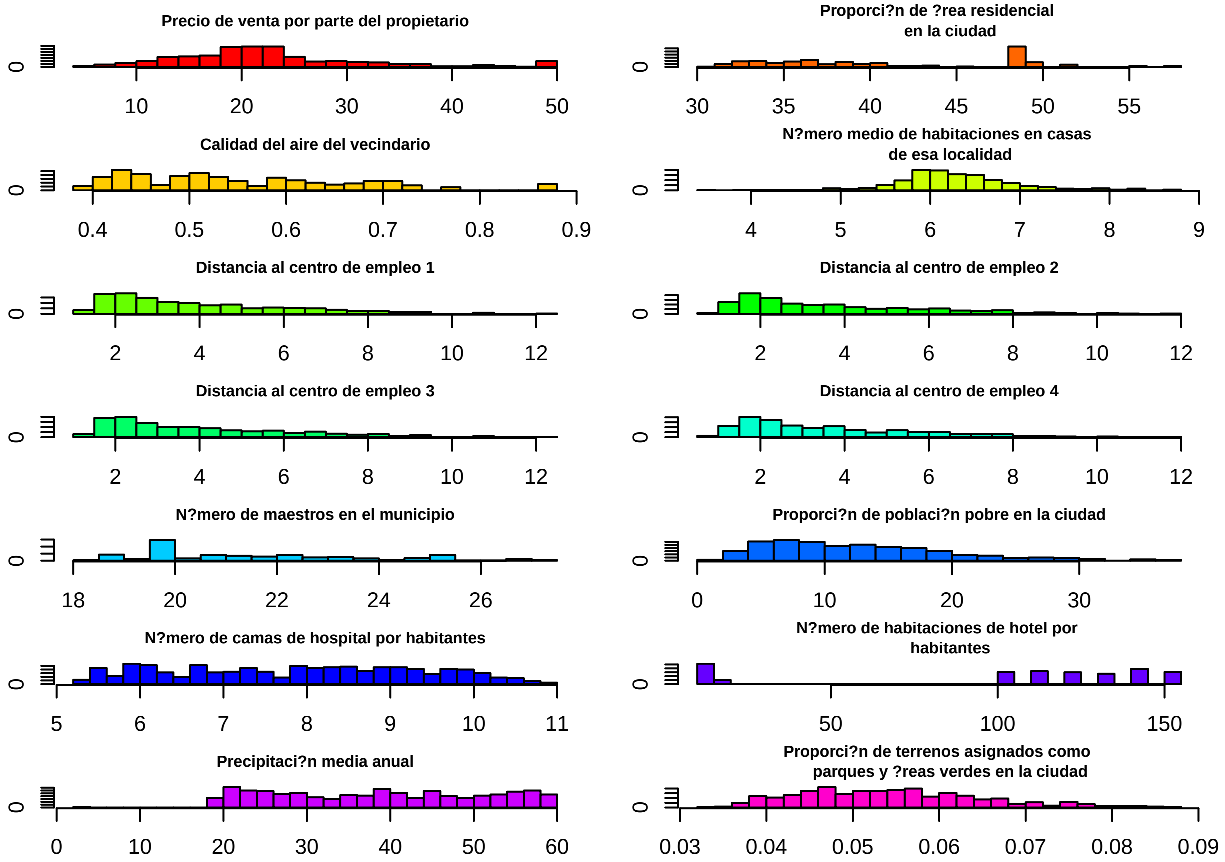
```r
numeric = unlist(lapply(house, is.numeric))
which(numeric, arr.ind = TRUE)
```

```
##       price  resid_area    air_qual    room_num         age       dist1
##           1           2           3           4           5           6
##       dist2       dist3       dist4    teachers   poor_prop  n_hos_beds
##           7           8           9          10          11          13
## n_hot_rooms    rainfall       parks
##          14          16          18
```

```r
length(which(numeric, arr.ind = TRUE))
```

```
## [1] 15
```

```r
colorForHistograms = rainbow(length(which(numeric, arr.ind = TRUE)))

par(mfrow=c(7,2),mar=c(2,2,2,2))

hist(house$price, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[1],main="Precio de venta por parte del propietario",cex.main=0.8, cex.lab=0

hist(house$resid_area, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[2],main="Proporci?n de ?rea residencial
     en la ciudad",cex.main=0.8, cex.lab=0.8)

hist(house$air_qual, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[3],main="Calidad del aire del vecindario"
     ,cex.main=0.8, cex.lab=0.8)

hist(house$room_num, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[4],main="N?mero medio de habitaciones en casas
     de esa localidad", cex.main=0.8, cex.lab=0.8)

hist(house$dist1, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[5],main="Distancia al centro de empleo 1",
     cex.main=0.8, cex.lab=0.8)

hist(house$dist2, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[6],main="Distancia al centro de empleo 2",
     cex.main=0.8, cex.lab=0.8)

hist(house$dist3, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[7],main="Distancia al centro de empleo 3",
     cex.main=0.8, cex.lab=0.8)

hist(house$dist4, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[8],main="Distancia al centro de empleo 4",
     cex.main=0.8, cex.lab=0.8)

hist(house$teachers, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[9],main="N?mero de maestros en el municipio",
     cex.main=0.8, cex.lab=0.8)

hist(house$poor_prop, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[10],main="Proporci?n de poblaci?n pobre en la ciudad",
     cex.main=0.8, cex.lab=0.8)

hist(house$n_hos_beds, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[11],main="N?mero de camas de hospital por habitantes",
     cex.main=0.8, cex.lab=0.8)

hist(house$n_hot_rooms, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[12],main="N?mero de habitaciones de hotel por
     habitantes", cex.main=0.8, cex.lab=0.8)

hist(house$rainfall, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[13],main="Precipitaci?n media anual",
     cex.main=0.8, cex.lab=0.8)

hist(house$parks, breaks=sqrt(dim(house)[1]),
     col=colorForHistograms[14],main="Proporci?n de terrenos asignados como
     parques y ?reas verdes en la ciudad",
     cex.main=0.8, cex.lab=0.8)
```



2. Modelo de regresión lineal

2.1 Modelo de RLS

2.1.1 Calcular

Enunciado: *Estimar por mínimos cuadrados ordinarios dos modelos lineales que expliquen la variable price, uno en función de la variable teachers y otro en función de la variable poor_prop.*

```r
get_cov_muestral<- function(x,y){
    mean_x = mean(x)
    mean_y = mean(y)
    sum = 0
    for (i in 1:length(x)){
        sum = sum + ((x[i] - mean_x)*(y[i] - mean_y))
    }
    return (sum/(length(x) - 1))
}

get_var_muestral <- function(x){
    mean_x = mean(x)
    sum = 0
    for (i in 1:length(x)){
        sum = sum + ((x[i]-mean_x)^2)
    }
    return (sum/(length(x) - 1))
}

get_b1 <- function(x,y){
    Sxy = get_cov_muestral(x,y)
    S2x = get_var_muestral(x)

    return (Sxy/S2x)
}

get_b0 <- function(x,y){
    mean_y = mean(y)
    b1 = get_b1(x,y)
    mean_x = mean(x)

    return(mean_y - (b1*mean_x))
}
```

- teachers: 18.8585379 + 0.1192217x
- poor_prop: 25.6331722 -0.5761549x