Activitat 3: Models predictius

Marc Cervera Rosell

2024-05-24

```
setRepositories(ind=2) # Per descarregar paquets de CRAN
```

1. Regressió lineal

1.1 Preparació de les dades

```
tryCatch({
  data <- read.csv("Casas.csv", header = TRUE)
  print("Fitxer llegit correctament")
}, error = function(e){
  cat("ERROR en el moment de llegir el fitxer:",conditionMessage(e), "\n")
})
```

```
## [1] "Fitxer llegit correctament"
```

```
any(is.na(data))
```

```
## [1] FALSE
```

El fitxer no conté valors NA

Canvi de peus quadrats a metres quadrats i de dòlars a euros:

```
for (i in seq_along(data$sqft_living)) {
  data$sqft_living[i] <- data$sqft_living[i] * 0.0929
  data$sqft_lot[i] <- data$sqft_lot[i] * 0.0929
  data$sqft_living15[i] <- data$sqft_living15[i] * 0.0929
  data$sqft_lot15[i] <- data$sqft_lot15[i] * 0.0929
  data$sqft_basement[i] <- data$sqft_basement[i] * 0.0929
  data$price[i] <- data$price[i] * 0.93
}
```

La funció *seq_along()* ens permet iterar sobre la columna indicada i, atès que, totes les columnes tenen la mateixa longitud no hi ha problema a posar com a argument una columna o una altra.

```
# Canvi de noms de les columnes afectades per les conversions monetàries i
# per les conversions del sistema imperial al sistema mètric
names(data)[names(data) == "price"] <- "price_eur"
names(data)[names(data) == "sqft_living"] <- "m2_living"
names(data)[names(data) == "sqft_lot"] <- "m2_lot"
names(data)[names(data) == "sqft_living15"] <- "m2_living15"
names(data)[names(data) == "sqft_lot15"] <- "m2_lot15"
names(data)[names(data) == "sqft_basement"] <- "m2_basement"
```

```
columns <- names(data)
type <- sapply(data, class)
for (i in seq_along(columns)) {
  cat("La columna", columns[i], "és de tipus", type[i], "\n")
}
```

```
## La columna date és de tipus character
## La columna price_eur és de tipus numeric
## La columna bedrooms és de tipus integer
## La columna bathrooms és de tipus numeric
## La columna m2_living és de tipus numeric
## La columna m2_lot és de tipus numeric
## La columna floors és de tipus numeric
## La columna waterfront és de tipus integer
## La columna view és de tipus integer
## La columna condition és de tipus integer
## La columna m2_basement és de tipus numeric
## La columna yr_built és de tipus integer
## La columna yr_renovated és de tipus integer
## La columna zipcode és de tipus integer
## La columna m2_living15 és de tipus numeric
## La columna m2_lot15 és de tipus numeric
```

Després d'analitzar els tipus de les variables del fitxer es determina realitzar un canvi de tipus de les següents variables:

- *bathrooms* -> Actualment és de tipus numèric, i no es canviarà atès que si s'observa **l'apartat de discusions del dataset**, es podrà veure el significat dels decimals. Per veure la web del dataset, clicar sobre el text en negreta.
- *floors* -> Actualment és de tipus numèric. En aquest cas, no s'ha trobat cap explicació per als decimals d'aquesta variable, per tant es decideix fer el canvi a *integer* sota la lògica de que no podem tenir mitja planta o 0.64 plantes.

```
data <- transform(data,
                  floors = as.integer(floors))
```

```
columns <- names(data)
type <- sapply(data, class)
for (i in seq_along(columns)) {
  cat("La columna", columns[i], "és de tipus", type[i], "\n")
}
```

```
## La columna date és de tipus character
## La columna price_eur és de tipus numeric
## La columna bedrooms és de tipus integer
## La columna bathrooms és de tipus numeric
## La columna m2_living és de tipus numeric
## La columna m2_lot és de tipus numeric
## La columna floors és de tipus integer
## La columna waterfront és de tipus integer
## La columna view és de tipus integer
## La columna condition és de tipus integer
## La columna m2_basement és de tipus numeric
## La columna yr_built és de tipus integer
## La columna yr_renovated és de tipus integer
```

```
## La columna zipcode és de tipus integer
## La columna m2_living15 és de tipus numeric
## La columna m2_lot15 és de tipus numeric
```

Com s'observa, després d'aplicar la funció *transform()* s'han modificat els tipus.

1.2 Estudi de correlació lineal

Considerant que s'han d'excloure dues de les variables del fitxer en el moment del càlcul de la correlació lineal, cal seleccionar, primer, aquelles columnes que sí que s'usaran en el càlcul.

```
subset_estudi_correlacio <- data[, c("price_eur", "bedrooms", "bathrooms", "m2_living",
                                     "m2_lot", "floors", "waterfront", "view",
                                     "condition", "m2_basement", "yr_built",
                                     "yr_renovated", "m2_living15", "m2_lot15")]
```

```
matriu_correlacio <- cor(subset_estudi_correlacio)
indexs <- which(matriu_correlacio > 0.2, arr.ind = TRUE)
indexs_ordenats <- indexs[order(matriu_correlacio[indexs], decreasing = TRUE), ]
files <- rownames(matriu_correlacio)[indexs_ordenats[,1]]
columnes <- colnames(matriu_correlacio)[indexs_ordenats[,2]]
matriu_noms <- cbind(files, columnes, matriu_correlacio[indexs_ordenats])
matriu_final <- matrix(matriu_noms, ncol = 3, byrow = FALSE)
colnames(matriu_final) <- c("Nom variable", "Nom variable", "Coef. Correlació")
print(matriu_final)
```

```
##       Nom variable   Nom variable   Coef. Correlació
##  [1,] "price_eur"    "price_eur"    "1"
##  [2,] "bedrooms"     "bedrooms"     "1"
##  [3,] "bathrooms"    "bathrooms"    "1"
##  [4,] "m2_living"    "m2_living"    "1"
##  [5,] "m2_lot"       "m2_lot"       "1"
##  [6,] "floors"       "floors"       "1"
##  [7,] "waterfront"   "waterfront"   "1"
##  [8,] "view"         "view"         "1"
##  [9,] "condition"    "condition"    "1"
## [10,] "m2_basement"  "m2_basement"  "1"
## [11,] "yr_built"     "yr_built"     "1"
## [12,] "yr_renovated" "yr_renovated" "1"
## [13,] "m2_living15"  "m2_living15"  "1"
## [14,] "m2_lot15"     "m2_lot15"     "1"
## [15,] "m2_living15"  "m2_living"    "0.756420259017221"
## [16,] "m2_living"    "m2_living15"  "0.756420259017221"
## [17,] "m2_living"    "bathrooms"    "0.754665278967373"
## [18,] "bathrooms"    "m2_living"    "0.754665278967373"
## [19,] "m2_lot15"     "m2_lot"       "0.718556752433035"
## [20,] "m2_lot"       "m2_lot15"     "0.718556752433035"
## [21,] "m2_living"    "price_eur"    "0.702043721232527"
## [22,] "price_eur"    "m2_living"    "0.702043721232527"
## [23,] "m2_living15"  "price_eur"    "0.585374006317152"
## [24,] "price_eur"    "m2_living15"  "0.585374006317152"
## [25,] "yr_built"     "floors"       "0.578619375159292"
## [26,] "floors"       "yr_built"     "0.578619375159292"
## [27,] "m2_living"    "bedrooms"     "0.576670692502244"
## [28,] "bedrooms"     "m2_living"    "0.576670692502244"
## [29,] "m2_living15"  "bathrooms"    "0.568634289578224"
```

```
## [30,] "bathrooms"    "m2_living15"  "0.568634289578224"
## [31,] "bathrooms"    "price_eur"    "0.525134072745601"
## [32,] "price_eur"    "bathrooms"    "0.525134072745601"
## [33,] "floors"       "bathrooms"    "0.519018991536239"
## [34,] "bathrooms"    "floors"       "0.519018991536239"
## [35,] "bathrooms"    "bedrooms"     "0.51588363761583"
## [36,] "bedrooms"     "bathrooms"    "0.51588363761583"
## [37,] "yr_built"     "bathrooms"    "0.506019438285253"
## [38,] "bathrooms"    "yr_built"     "0.506019438285253"
## [39,] "m2_basement"  "m2_living"    "0.435042973669821"
## [40,] "m2_living"    "m2_basement"  "0.435042973669821"
## [41,] "view"         "waterfront"   "0.401857350697571"
## [42,] "waterfront"   "view"         "0.401857350697571"
## [43,] "view"         "price_eur"    "0.397346474378939"
## [44,] "price_eur"    "view"         "0.397346474378939"
## [45,] "m2_living15"  "bedrooms"     "0.391637523968824"
## [46,] "bedrooms"     "m2_living15"  "0.391637523968824"
## [47,] "floors"       "m2_living"    "0.35332060339984"
## [48,] "m2_living"    "floors"       "0.35332060339984"
## [49,] "m2_living15"  "yr_built"     "0.326228899595712"
## [50,] "yr_built"     "m2_living15"  "0.326228899595712"
## [51,] "m2_basement"  "price_eur"    "0.32383735813766"
## [52,] "price_eur"    "m2_basement"  "0.32383735813766"
## [53,] "yr_built"     "m2_living"    "0.318048768996441"
## [54,] "m2_living"    "yr_built"     "0.318048768996441"
## [55,] "bedrooms"     "price_eur"    "0.308338368688097"
## [56,] "price_eur"    "bedrooms"     "0.308338368688097"
## [57,] "m2_basement"  "bedrooms"     "0.303093375320663"
## [58,] "bedrooms"     "m2_basement"  "0.303093375320663"
## [59,] "m2_living15"  "floors"       "0.296560578164614"
## [60,] "floors"       "m2_living15"  "0.296560578164614"
## [61,] "view"         "m2_living"    "0.284611186216901"
## [62,] "m2_living"    "view"         "0.284611186216901"
## [63,] "m2_basement"  "bathrooms"    "0.283770034004669"
## [64,] "bathrooms"    "m2_basement"  "0.283770034004669"
## [65,] "m2_living15"  "view"         "0.280439081995455"
## [66,] "view"         "m2_living15"  "0.280439081995455"
## [67,] "m2_basement"  "view"         "0.276946578767584"
## [68,] "view"         "m2_basement"  "0.276946578767584"
## [69,] "waterfront"   "price_eur"    "0.266330510522256"
## [70,] "price_eur"    "waterfront"   "0.266330510522256"
## [71,] "floors"       "price_eur"    "0.237207363532409"
## [72,] "price_eur"    "floors"       "0.237207363532409"
## [73,] "m2_living15"  "m2_basement"  "0.200354983394243"
## [74,] "m2_basement"  "m2_living15"  "0.200354983394243"
```

Tenint en compte que solament s'han mostrat aquells coeficients de correlació lineal majors a 0.2, es pot assegurar que la correlació lineal de les variables és positiva, és a dir, quan una de les dues variables augmenta el seu valor, la segona variable també augmenta el seu valor de manera proporcional.

En aquest cas d'estudi, el llindar s'ha establert en 0.2, per tant, aquelles parelles de variables amb un coeficient de correlació lineal proper a 0.2 tindran una correlació dèbil i aquelles parelles amb un coeficient de correlació lineal proper a 1 (o 1 en el cas del càlcul de la correlació lineal amb elles mateixes) tindran una forta correlació.

1.3 Generació dels conjunts d'entrenament i de test

```
set.seed(123)
indexs_training <- sample(nrow(subset_estudi_correlacio), 0.8 *
                            nrow(subset_estudi_correlacio))
set_training <- subset_estudi_correlacio[indexs_training, ]
set_test <- subset_estudi_correlacio[-indexs_training, ]
```

1.4 Estimació del model de regressió lineal

L'ajust d'un model de regressió lineal utilitzant el mètode de mínims quadrats ordinaris es du a terme, popularment, amb la funció *lm()*.

```
model <- lm(price_eur ~ ., data = set_training)
```

La variable *price_eur*, es la variable anomenada "de resposta" atès que és la variable que està a l'esquerra de la titlla (*virgulilla* en castellà).

1.4.1

```
prediccions <- predict(model, newdata = set_training)
```

```
coeficient_r <- 1 - (sum((set_training$price_eur - prediccions)^2) /
                       sum((set_training$price_eur - mean(set_training$price_eur))^2))
cat("Coeficient R quadrat:",coeficient_r,"\n")
```

```
## Coeficient R quadrat: 0.609292
```

```
fiv_model_ajustat <- 1 / (1 - coeficient_r)
cat("FIV del model ajustat:",fiv_model_ajustat)
```

```
## FIV del model ajustat: 2.559456
```

Per calcular els valors dels FIV per cada una de les variables predictores del model, cal ajustar un model de regressió lineal incloent totes les variables predictores menys una. És a dir s'han de calcular els valors FIV de excloent a cada model una de les variables predictores del model original.

```
valors_fiv <- data.frame(variable_exclosa = character(ncol(set_training) - 1),
                         fiv = numeric(ncol(set_training) - 1))
for (i in 2:ncol(set_training)) {
  training_aux <- set_training
  columna <- colnames(set_training)[i]
  training_aux <- training_aux[, -i]
  model_sense_variable_i <- lm(price_eur ~ ., data = training_aux)
  prediccions_aux <- predict(model_sense_variable_i, newdata = training_aux)
  coeficient_r_aux <- 1 - (sum((training_aux$price_eur - prediccions_aux)^2) /
                             sum((training_aux$price_eur - mean(training_aux$price))^2))
  fiv <- 1 / (1 - coeficient_r_aux)
  valors_fiv[i, "variable_exclosa"] <- columna
  valors_fiv[i, "fiv"] <- fiv
}
print(valors_fiv[-1, ])
```

```
##    variable_exclosa      fiv
## 2          bedrooms 2.491435
## 3         bathrooms 2.522930
## 4         m2_living 2.175844
```

```
## 5            m2_lot 2.559445
## 6            floors 2.541138
## 7        waterfront 2.474266
## 8              view 2.506456
## 9         condition 2.550368
## 10      m2_basement 2.556748
## 11         yr_built 2.373490
## 12     yr_renovated 2.558392
## 13      m2_living15 2.504313
## 14         m2_lot15 2.549871
```

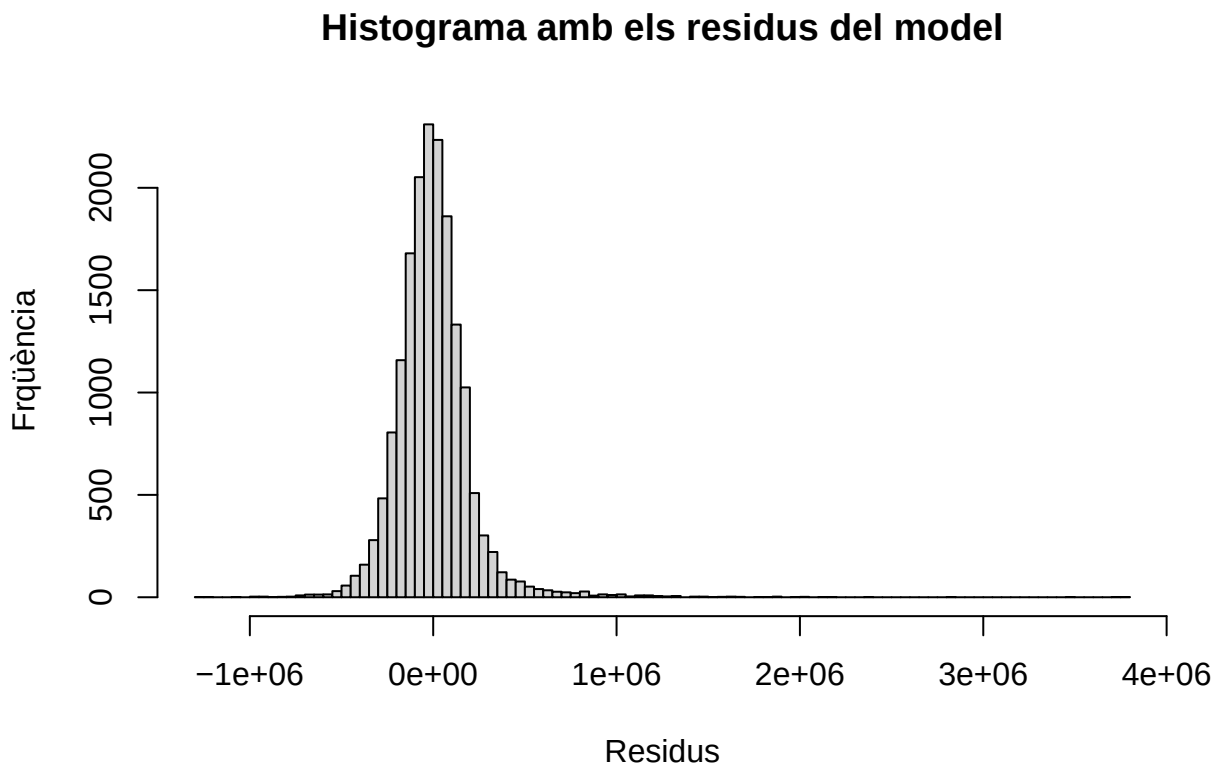
Considerant el FIV del model ajustat i els FIVs dels models individuals, es determina que existeix colinealitat entre les variables. És a dir, l'existència de colinealitat suggereix que les variables predictores estan correlacionades. Per tant, sota la premissa de la seva rellevància teòrica, és a dir, totes les variables incloses en el model són necessàries per a obtenir tots els aspectes importants del fenomen d'estudi (explicar el preu de l'habitatge en funció de les variables seleccionades), i tot i la colinealitat, no es considera excloure cap variable del model.

1.5 Diagnosi del model

Per calcular els residus cal restar els valors observats i els valors reals. Per obtenir els valors predits hi ha dues opcions: la primera utilitzar una crida a la funció *predict()* (utilitzada més amunt per al càlcul del valor de R quadrat) i posteriorment realitzar la resta, o utilitzar la funció *residuals()* que ja retorna directament el càlcul fet.

```
valors_observats <- set_training$price_eur
residus <- valors_observats - prediccions
```

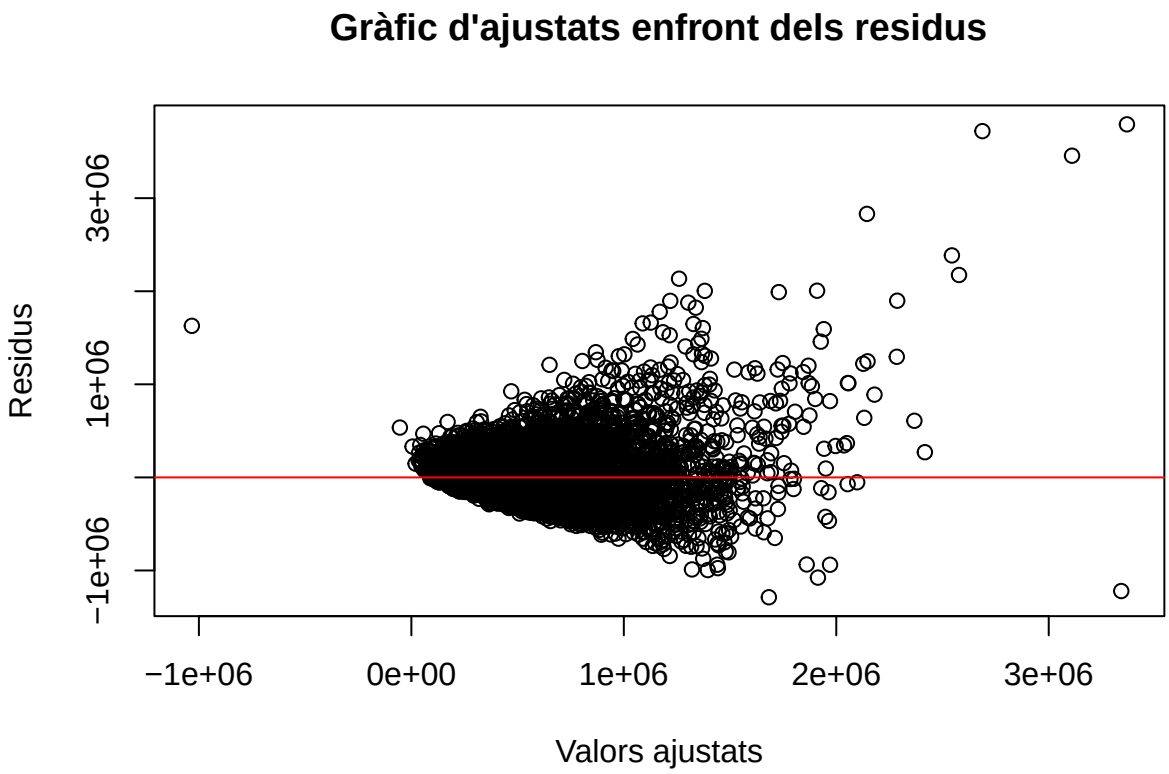
```
hist(residus, breaks = 100, main = "Histograma amb els residus del model",
     xlab = "Residus", ylab = "Frqüència")
```



S'observa que l'histograma s'assembla a una campana al voltant del valor 0. Aquest fet indica que els residus del model segueixen una distribució normal.

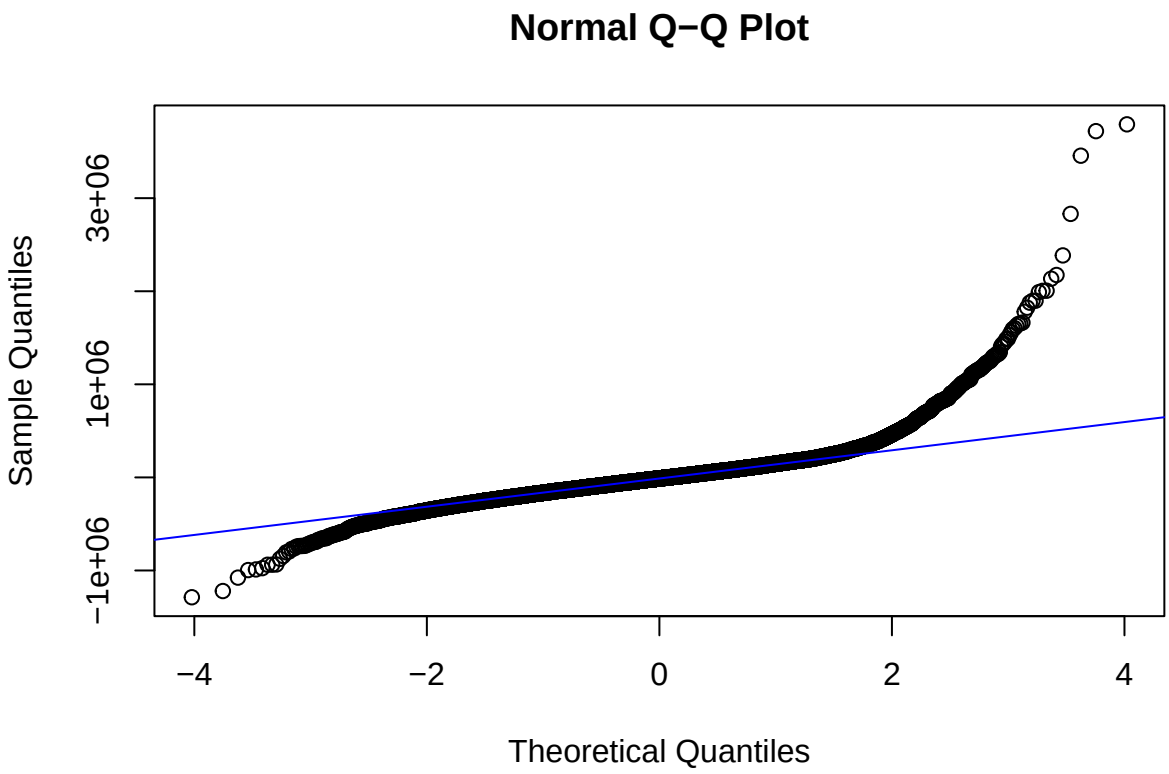
L'esmentada forma de campana al voltant del 0, és un indicador de què el model fa bones prediccions.

```
plot(fitted(model), residus, main = "Gràfic d'ajustats enfront dels residus",
     xlab = "Valors ajustats", ylab = "Residus")
abline(h = 0, col = "red")
```



En el gràfic de residus enfront dels valors ajustats, es pot observar, que el model presenta certs problemes de dispersió irregular, és a dir, la variància dels residus augmenta a mesura que ho fan els valors ajustats (heteroscedasticitat).

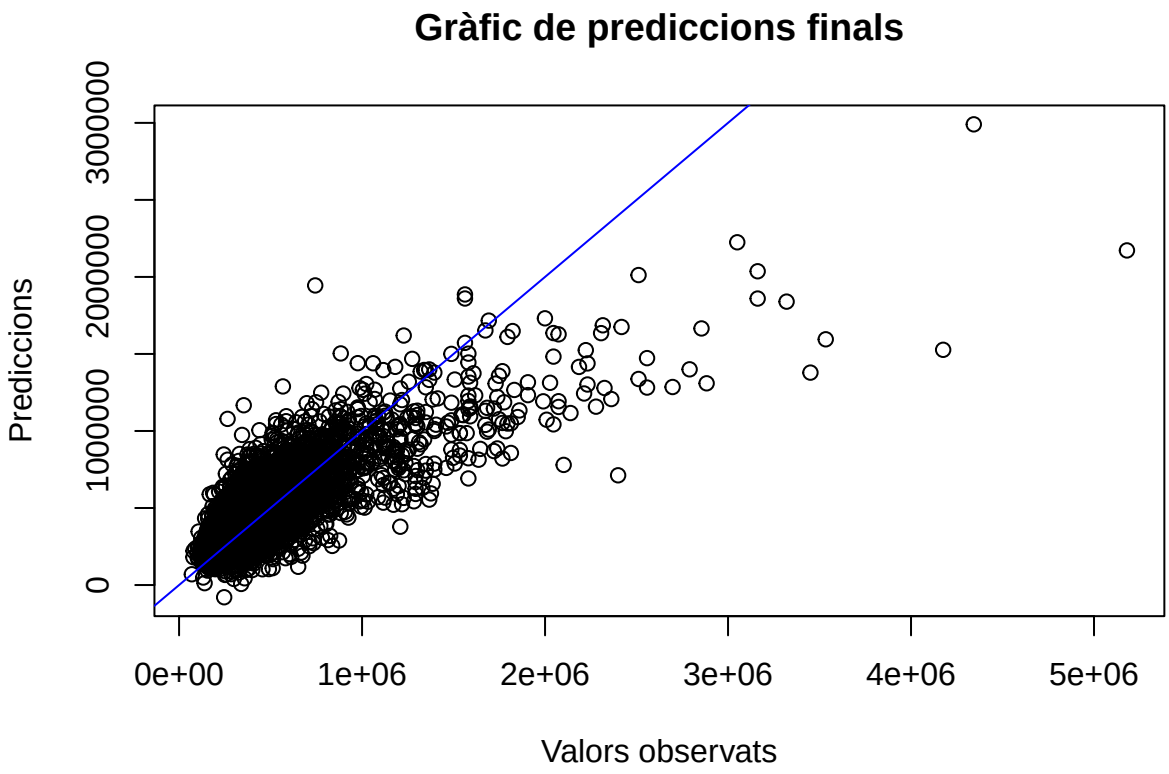
```
qqnorm(residus)
qqline(residus, col = "blue")
```



S'observa que el gràfic QQ presenta dues corbes a les cues. Aquestes curvatures són indicadores d'asimetries, és a dir, llocs on la distribució de les dades no és normal. Atès que la cua dreta és més llarga que l'esquerra es pot assegurar que hi ha una asimetria positiva (quantitat major de valors atípics en l'extrem superior).

1.6 Predicció del model

```
prediccions_finals <- predict(model, newdata = set_test)
plot(set_test$price_eur, prediccions_finals, main = "Gràfic de prediccions finals",
     xlab = "Valors observats", ylab = "Prediccions")
abline(0, 1, col = "blue")
```




```
sumatori <- 0
for (i in 1:length(prediccions_finals)) {
  sumatori <- sumatori + ((set_test$price_eur[i] - prediccions_finals[[i]])^2)
}
rmse <- sqrt(sumatori / nrow(set_test))
cat("Valor RMSE:",rmse,"\n")
```

```
## Valor RMSE: 219223.3
```

```
mitjana_preus <- mean(set_test$price_eur)
cat("Mitjana del preu dels habitatges:",mitjana_preus)
```

```
## Mitjana del preu dels habitatges: 502769
```

Com el RMSE està calculat en el preu dels habitatges (variable depenent), es pot interpretar el resultat del RMSE comparant amb el valor mitjà del preu dels habitatges. S'observa que el RMSE és significativament menor que el preu mitjà dels habitatges, per tant, es conclou que el model té bona precisió.

2 Regressió logística

2.1 i 2.2

Per evitar repetir el proces de canvi de tipus dos cops un cop fet el split, s'ajunten els dos apartats.

```
data$price_re <- ifelse(data$price_eur < 500000, 0, 1) # data$price < 500000 ? 0 : 1
```

```
subset_estudi_correlacio_2 <- data[, c("price_re", "bedrooms", "bathrooms", "m2_living",
                                       "m2_lot", "floors", "waterfront", "view",
```

```
                                        "condition", "m2_basement", "yr_built",
                                        "yr_renovated", "m2_living15", "m2_lot15")]
```

Com la variable *price_eur* ha estat codificada en la variable *price_re* i, a més a més, en l'exercici 2.2 s'especifica que la variable de preus sense codificar s'ha d'eliminar, en el subset d'estudi 2 es treuen els preus sense codificar.

```
obtencio_tipus <- function(dades){
  columns_aux <- names(dades)
  type_aux <- sapply(dades, class)
  for (i in seq_along(columns_aux)) {
    cat("La columna", columns_aux[i], "és de tipus", type_aux[i], "\n")
  }
}
```

```
tipus_abans_conversio <- obtencio_tipus(subset_estudi_correlacio_2)
```

```
## La columna price_re és de tipus numeric
## La columna bedrooms és de tipus integer
## La columna bathrooms és de tipus numeric
## La columna m2_living és de tipus numeric
## La columna m2_lot és de tipus numeric
## La columna floors és de tipus integer
## La columna waterfront és de tipus integer
## La columna view és de tipus integer
## La columna condition és de tipus integer
## La columna m2_basement és de tipus numeric
## La columna yr_built és de tipus integer
## La columna yr_renovated és de tipus integer
## La columna m2_living15 és de tipus numeric
## La columna m2_lot15 és de tipus numeric
```

```
subset_estudi_correlacio_2 <- transform(subset_estudi_correlacio_2,
                      price_re = as.factor(price_re),
                      view = as.factor(view),
                      waterfront = as.factor(waterfront))
```

```
tipus_despres_conversio <- obtencio_tipus(subset_estudi_correlacio_2)
```

```
## La columna price_re és de tipus factor
## La columna bedrooms és de tipus integer
## La columna bathrooms és de tipus numeric
## La columna m2_living és de tipus numeric
## La columna m2_lot és de tipus numeric
## La columna floors és de tipus integer
## La columna waterfront és de tipus factor
## La columna view és de tipus factor
## La columna condition és de tipus integer
## La columna m2_basement és de tipus numeric
## La columna yr_built és de tipus integer
## La columna yr_renovated és de tipus integer
## La columna m2_living15 és de tipus numeric
## La columna m2_lot15 és de tipus numeric
```

S'observa el canvi de tipus després d'aplicar la funció *transform()*

```r
set.seed(123)
indexs_training_2 <- sample(nrow(subset_estudi_correlacio_2), 0.8 *
                            nrow(subset_estudi_correlacio_2))
training2 <- subset_estudi_correlacio_2[indexs_training_2, ]
testing2 <- subset_estudi_correlacio_2[-indexs_training_2, ]
```

```r
model_log <- glm(price_re ~ ., data = training2, family = binomial)
```

Atès que la variable dependent solament pot prendre dos valors (1 i 0), s'ajusta el model de regressió logística a un model de regressió logística binomial.

```r
summary(model_log)
```

```
##
## Call:
## glm(formula = price_re ~ ., family = binomial, data = training2)
##
## Coefficients:
##                Estimate Std. Error z value Pr(>|z|)
## (Intercept)   5.711e+01  2.141e+00  26.668  < 2e-16 ***
## bedrooms     -2.759e-01  3.048e-02  -9.054  < 2e-16 ***
## bathrooms     5.520e-01  5.245e-02  10.525  < 2e-16 ***
## m2_living     1.503e-02  6.190e-04  24.279  < 2e-16 ***
## m2_lot        1.907e-05  8.829e-06   2.160  0.03075 *
## floors        8.018e-01  5.904e-02  13.580  < 2e-16 ***
## waterfront1   3.988e-01  4.391e-01   0.908  0.36378
## view1         6.546e-01  1.652e-01   3.962 7.44e-05 ***
## view2         5.999e-01  1.022e-01   5.871 4.34e-09 ***
## view3         5.376e-01  1.542e-01   3.486  0.00049 ***
## view4         1.525e+00  3.352e-01   4.551 5.34e-06 ***
## condition     2.547e-01  3.536e-02   7.204 5.86e-13 ***
## m2_basement   4.298e-04  7.079e-04   0.607  0.54375
## yr_built     -3.320e-02  1.111e-03 -29.885  < 2e-16 ***
## yr_renovated -3.437e-05  5.621e-05  -0.611  0.54090
## m2_living15   1.347e-02  5.672e-04  23.746  < 2e-16 ***
## m2_lot15     -5.450e-05  1.306e-05  -4.175 2.99e-05 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 22858  on 17289  degrees of freedom
## Residual deviance: 14483  on 17273  degrees of freedom
## AIC: 14517
##
## Number of Fisher Scoring iterations: 6
```

Atès que les variables *waterfront*, *m2_basement* i *yr_renovated* tenen un valor p major a 0.05, es podria considerar que no són significatives i que, per tant, podrien ser eliminades del model, però considerant el que representa cada variable (accés a un llac, superfície de soterrani i any de renovació), no es considera la seva eliminació. La raó de mantenir les tres variables és la seva importància teòrica en l'efecte que suposen en el preu de l'habitatge. És a dir, una casa amb accés a un llac, serà més cara que una casa en un terreny sec (desitjabilitat). Una casa amb soterrani serà més cara que una casa sense soterrani (o un soterrani més petit), ja que el soterrani són metres quadrats habitables. Cal esmentar, que el soterrani, de la mateixa manera que pot augmentar el preu, també el pot fer disminuir. Per exemple, un soterrani gran i acabat (preparat

per fer-hi vida) pot fer augmentar el valor de la casa. Per contra, un soterrani més petit o que no estigui preparat per fer-hi vida pot fer decaure el valor de la propietat. Finalment, l'any de renovació també afecta al preu, pel fet que serà més cara una casa renovada l'any 2012, per exemple, que una casa renovada l'any 1990 o que no s'hagi renovat mai.

```
model_final <- model_log
```

2.3

```
odd_ratio <- exp(coefficients(model_final))
```

```
intervals <- exp(confint(model_final))
```

```
## Waiting for profiling to be done...
```

```
taula_odd <- data.frame(
  Odd_ratio = odd_ratio,
  Limits = intervals
)
print(taula_odd)
```

```
##                 Odd_ratio Limits.2.5.. Limits.97.5..
## (Intercept)  6.327800e+24 9.670157e+22  4.278631e+26
## bedrooms     7.588709e-01 7.147976e-01  8.055088e-01
## bathrooms    1.736749e+00 1.567439e+00  1.925254e+00
## m2_living    1.015143e+00 1.013917e+00  1.016381e+00
## m2_lot       1.000019e+00 1.000002e+00  1.000037e+00
## floors       2.229550e+00 1.986000e+00  2.503240e+00
## waterfront1  1.490004e+00 6.381384e-01  3.598277e+00
## view1        1.924434e+00 1.395834e+00  2.669196e+00
## view2        1.821890e+00 1.492195e+00  2.227560e+00
## view3        1.711823e+00 1.268826e+00  2.323343e+00
## view4        4.597157e+00 2.454712e+00  9.177378e+00
## condition    1.290134e+00 1.203777e+00  1.382783e+00
## m2_basement  1.000430e+00 9.990425e-01  1.001819e+00
## yr_built     9.673433e-01 9.652296e-01  9.694427e-01
## yr_renovated 9.999656e-01 9.998552e-01  1.000076e+00
## m2_living15  1.013560e+00 1.012437e+00  1.014691e+00
## m2_lot15     9.999455e-01 9.999195e-01  9.999707e-01
```

S'observa que les variables *waterfront*, *m2_basement* i *yr_renovated*, contenen el valor 1 en els seus intervals de confiança, per tant, no podem assegurar si són factors de risc o de protecció, atès que un OR d'1 significa que no hi ha associació entre les variables.

Per altra part, es veu que hi ha variables amb un OR major a la unitat. Aquest fet implica que són factors de risc. En aquest model, els factors de risc són:

- *Intercept*
- *bathrooms*
- *m2_living*
- *m2_lot*
- *floors*
- *view1*
- *view2*

- *view3*
- *view4*
- *condition*
- *m2_living15*

Finalment, també s'observen variables amb un Odds-ratio inferior a 1 (factors de protecció):

- *bedrooms*
- *yr_built*
- *m2_lot15*

2.4 Matriu de confusió

```
prediccions_model_final <- predict(model_final, newdata = testing2,
                                   type = "response")
classificar_prediccions <- ifelse(prediccions_model_final >= 0.5, 1, 0)
matriu_confusio <- table(Valor_predit = classificar_prediccions,
                         Valor_real = testing2$price_re)
```

```
print(matriu_confusio)
```

```
##             Valor_real
## Valor_predit    0    1
##            0 2437  549
##            1  301 1036
```

Els valors de la matriu de confusió tenen la següent explicació:

- Quadrant superior esquerra: són els vertaders negatius. És a dir, el valor predit és 0 i el valor real també és 0.
- Quadrant superior dret: són els falsos negatius. És a dir, el model prediu un 0, però el valor real és un 1.
- Quadrant inferior esquerra: són els falsos positius. És a dir, el model prediu un 1, però el valor real és un 0.
- Quadrant inferior dret: són els vertaders positius. És a dir, el valor predit pel model és 1 i el valor real és 1.

```
vertaders_negatius <- matriu_confusio[1, 1] # Predit = real = 0
falsos_negatius <- matriu_confusio[1, 2] # Predit = 0; real = 1
falsos_positius <- matriu_confusio[2, 1] # Predit = 1; real = 0
vertaders_positius <- matriu_confusio[2, 2] # Predit = real = 1
sensibilitat <- vertaders_positius / (vertaders_positius + falsos_negatius)
especificitat <- vertaders_negatius / (vertaders_negatius + falsos_positius)
cat("Sensibilitat:",sensibilitat*100,"%\n")
```

```
## Sensibilitat: 65.36278 %
```

```
cat("Especificitat:",especificitat*100,"%")
```

```
## Especificitat: 89.00657 %
```

Es pot observar que el model té una alta especificitat, la qual cosa implica que és bo detectant els casos negatius. En altres paraules, el model detecta correctament el 89.00657% dels casos en els quals els habitatges tenen un preu inferior a 500000 euros.

Pel que fa a la sensibilitat del model, el raonament és paregut. És a dir, el model és capaç de predir correctament el 65.36278% dels casos en els quals els habitatges tenen un preu igual o superior a 500000 euros.

2.5

```
prediccio_individual <- predict(model_final, newdata = training2[3, ], type = "response")
cat("Probabilitat de que l'habitatge de la tercera fila tingui un preu superior a
      500000 euros:",prediccio_individual*100,"%")
```

```
## Probabilitat de que l'habitatge de la tercera fila tingui un preu superior a
##       500000 euros: 81.67841 %
```

S'observa que l'habitatge de la tercera fila de *training2*, té una probabilitat del 81.7%, aproximadament, de tenir un preu superior a 500000 euros.

2.6

A

La *devianza* nul · la representa amb quina eficàcia la variable de resposta es prediu mitjançant un model que inclou totes les variables independents.

```
# logLik -> Càlcul de la log likelihood function
devianza_residual <- -2 * logLik(model_final)
cat("Devianza residual:",devianza_residual[1])
```

```
## Devianza residual: 14483.01
```

La *devianza* nul · la representa amb quina eficàcia la variable de resposta es prediu mitjançant un model que inclou només el punt d'intersecció de la línia de regressió amb l'eix Y (intercept).

```
model_nul <- model_log <- glm(price_re ~ 1, data = training2, family = binomial)
devianza_nula <- -2 * logLik(model_nul)
cat("Devianza nul·la:",devianza_nula[1])
```

```
## Devianza nul·la: 22857.61
```

Es pot observar que la *devianza* residual és menor a la *devianza* nul · la, i com es sabut, per a que un model sigui considerat "bo", la *devianza* residual ha de ser menor que la nul · la, per tant, atesa la inferioritat de la *devianza* residual, és conclou que el model és bo.

B

```
chi_quadrat <- devianza_nula - devianza_residual
cat("Valor de Chi quadrat:", chi_quadrat)
```

```
## Valor de Chi quadrat: 8374.604
```

Els graus de llibertat del model es defineixen com el nombre d'observacions d'aquest menys el nombre de variables que estan sent estimades.

El nombre d'observacions correspon al nombre de files del conjunt de dades usat per entrenar el model i el nombre de variables que estan sent estimades. Per obtenir les variables hi ha dues opcions: la primera és imprimir el *summary()* i comptar totes les variables que allí apareixen (inclòs l' *intercept*). La segona opció és a través del mateix model. Com s'observa a continuació, per obtenir el nombre de variables esmentat, cal obtenir el nombre de coeficients del model.

```r
#Graus de llibertat = total files - variables sent estimades
graus_llibertat_model <- nrow(training2) - length(model_final$coefficients)
cat("Nombre de graus de llibertat del model:",graus_llibertat_model)
```

```
## Nombre de graus de llibertat del model: 17273
```

```r
probabilitat <- pchisq(chi_quadrat, df = graus_llibertat_model, lower.tail = FALSE)
cat("Probabilitat associada a l'estadístic de contrast:",probabilitat)
```

```
## Probabilitat associada a l'estadístic de contrast: 1
```

Tot i que el valor Chi-quadrat és bastant alt (8374.604), el valor de la probabilitat d'1 suggereix que no hi ha suficient evidència per a poder afirmar amb seguretat que les prediccions que es puguin fer amb el model ajustat són millors que les que es puguin fer amb el model nul (model ajustat solament amb *intercept*).

2.7

```r
chooseCRANmirror(ind = 1)
install.packages("pROC")
```

```
## Installing package into 'C:/Users/mcr99/AppData/Local/R/win-library/4.3'
## (as 'lib' is unspecified)
```

```
## package 'pROC' successfully unpacked and MD5 sums checked
```

```
## Warning: cannot remove prior installation of package 'pROC'
```

```
## Warning in file.copy(savedcopy, lib, recursive = TRUE): problem copying
## C:\Users\mcr99\AppData\Local\R\win-library\4.3\00LOCK\pROC\libs\x64\pROC.dll to
## C:\Users\mcr99\AppData\Local\R\win-library\4.3\pROC\libs\x64\pROC.dll:
## Permission denied
```

```
## Warning: restored 'pROC'
```

```
##
## The downloaded binary packages are in
##  C:\Users\mcr99\AppData\Local\Temp\RtmpaAlXNy\downloaded_packages
```

```r
library(pROC)
```

```
## Type 'citation("pROC")' for a citation.
```

```
##
## Attaching package: 'pROC'
```

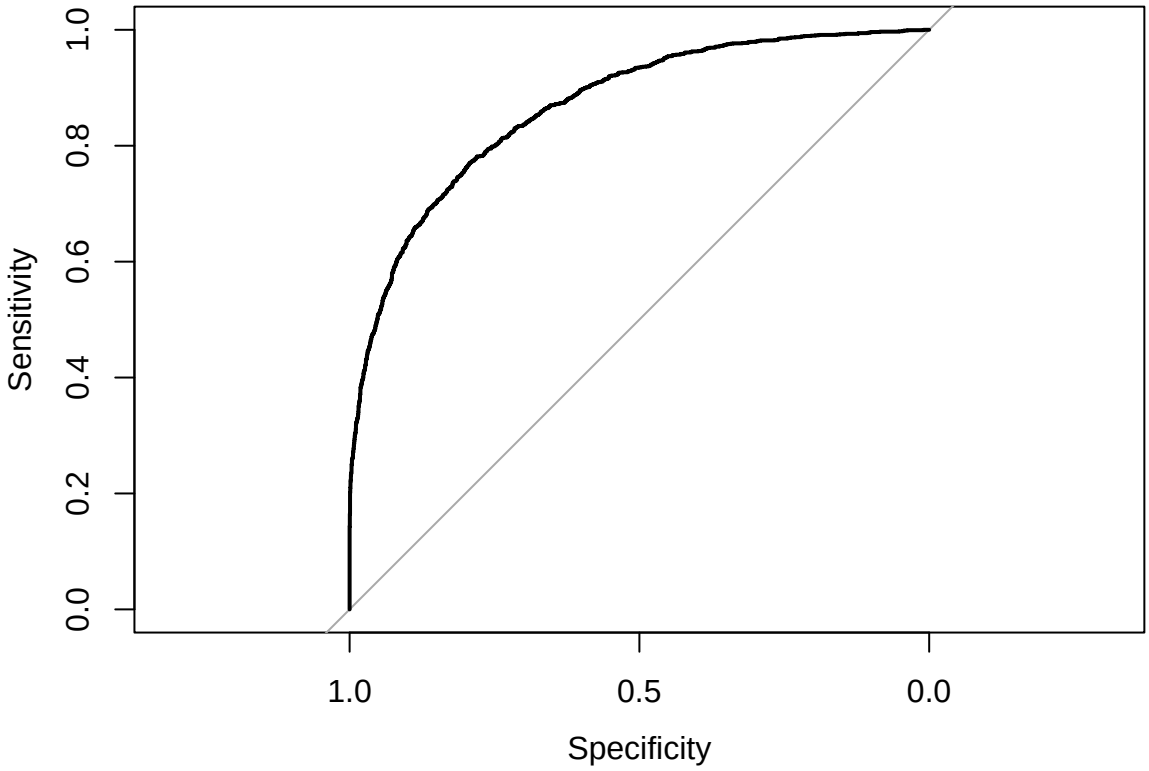
```
## The following objects are masked from 'package:stats':
##
##     cov, smooth, var
```

```r
corba_roc <- roc(testing2$price_re, prediccions_model_final)
```

```
## Setting levels: control = 0, case = 1
```

```
## Setting direction: controls < cases
```

```r
plot(corba_roc)
```




```
auc(corba_roc)
```

```
## Area under the curve: 0.8669
```

Es pot observar que 0.8 <= AUROC <= 0.9, per tant, i segons vist en els materials docents, una AUROC de 0.8669 implica que el model es capaç de discriminar molt be les classes.

3

En primer lloc, s'enumeren les conclusions del model lineal. La primera conclusió extreta és tot i que de vegades, per qüestió de càlculs, s'hagi de treure alguna variable del model de regressió, no sempre és adient fer-ho, ja que aquestes variables considerares per a descartar poden tenir una certa importància teòrica.

La segona conclusió del model lineal, és que com el gràfic mostrar en l'apartat 1.5 té forma de campana al voltant del 0, es pot concloure que el model és bo. En observar altres tipus de gràfics, però, s'hi poden veure alguns problemes com per exemple valors atípics (QQ-plot). Si s'observa el gràfic de valors ajustats enfront de residus, es pot observar que hi ha una zona del mapa que concentra una gran quantitat de punts. Un gràfic ideal tindria els punts distribuïts de manera aleatòria i sense que aquests formin cap patró. Per tant, el fet que molts punts es concentrin en una zona indica algun tipus de problema amb el model.

Al moment de fer prediccions, es pot observar, numèricament, que el valor de l'error és menor a la mitjana dels preus dels habitatges, cosa que indica que el model està actuant bé.

Passant al model de regressió logística, es pot concloure que de totes les variables que formen part del model final, no tenen el mateix efecte. Es poden dividir en dues categories: de risc i de protecció. Si s'imagina una balança es pot definir com a variable de risc aquella que posa pes al costat de la balança que afavoreix l'esdeveniment. Per contra, les variables de protecció són aquelles que afegeixen pes al costat de la balança

que afavoreix que l'esdeveniment no passi. Per tant, es podria dir, en aquest cas d'estudi, que les variables de risc són aquelles que intervenen perquè l'habitatge tingui un preu més elevat.

També es pot concloure que el model és millor predient els casos en què els habitatges tenen un preu igual o inferior a 500000 euros.

Finalment, es conclou que el model de regressió logística fa bones prediccions atesa l'àrea existent sota la corba de l'apartat 2.7. En aquest tipus de gràfiques com més panxa tingui més àrea hi haurà sota la corba i per tant millors prediccions farà el model.