Projecte d'Anàlisi Estadístic

Dataset: House Prices

Bernat Medina Pérez: 1606505

Joan Paz Garcia: 1598851

ÍNDEX

Introducció

Avui dia, el jovent va prenent la iniciativa a l'hora d'independitzar-se a una nova casa. No obstant, estem en una època on ha hagut una inflació important dels preus en aquest sector, fent més difícil la cerca d'una nova casa que s'adeqüi qualitat-preu sense que sigui una quantitat de diners desorbitada.

En aquest estudi estadístic, s'intentarà oferir una ajuda a aquest sector de la població que vol buscar una nova llar i es voldrà informar sobre quins paràmetres són els més importants a tenir en compte per fer la compra de la casa.

Dataset

El dataset amb el que es treballarà serà 'house_prices.csv', el qual consta de variables de tot tipus, des de categòriques fins a contínues. Hi ha 80 variables explicatives i 1 variable explicada, que serà el preu de venta de a casa. Per veure més informació sobre les variables, es pot acudir a la seva font House Prices - Advanced Regression Techniques | Kaggle.

Dades utilitzades al llarg del treball:

- **SalePrice**: És el preu al que es ven la casa. El rang de valors és molt ampli ja que són preus de cases i varien molt.
- **OverallQual**: És un valor entre 0 i 10 que indica la qualitat de la casa en general.
- **GrLivArea**: És l'àrea total del terreny on està la casa en peus quadrats.
- **TotalBsmtSF**: És l'àrea total del soterrani de la casa en peus quadrats.
- **YearBuilt**: És l'any en el que es va construir la casa. Agafa valors de 112 anys diferents.
- **RoofStyle**: És el tipus de sostre de la casa. Pren 6 valors diferents.
- **HouseStyle**: És el tipus de casa. Pren 8 valors diferents.

Objectius

L'obectiu d'aquest treball és analitzar exhaustivament quines característiques d'una casa són més rellevants per al preu d'aquesta. Per tant, s'escollirà com a variable dependent el preu de la casa i es mirarà quines variables independents tenen una influència més gran sobre aquest. D'aquesta manera, es podrà predir per quant es vendria la casa a partir d'unes característiques concretes, facilitant d'aquesta manera la cerca dels clients.

A més, també es voldrà veure quin resultat prediuen diferents models sobre si una casa és cara o barata, i quin resultat prediuen uns altres sobre si una casa és nova o vella.

Treball final

Joan i Bernat

29/12/2022

Exercici 1: Regressió lineal múltiple

En aquest exercici es treballarà amb un model de regressió lineal múltile on s'escolliran les 4 variables explicatives que tinguin més influència sobre la variable dependent 'SalePrice'. S'analitzarà si el model és globalment estadísticament significatiu, si les variables explicatives són estadísticament significatives per explicar la variable target i es construirà un model restringit amb aquelles variables que hagin estat significatives amb la validació anterior o bé es construirà un model amb les variables més significatives.

1.1 Carregar i seleccionar les dades

Per començar, es carrega el dataset escollit i es seleccionen les 4 variables numèriques amb les que es treballaran en aquest exercici. Per fer-ho, seleccionarem totes les columnes numèriques per fer un anàlisi de les correlacions.

A més, la variable dependent escollida (target) ha estat 'SalePrice', ja que és aquella que es voldrà predir i la que es mirarà a l'hora de comprar una casa.

```
dataset_tot <- read.csv("house_prices.csv",sep=",")

# Cargar el paquete dplyr
library(dplyr)
library(magrittr)
numeric_columns <- dataset_tot %>% select_if(is.numeric)
```

Per mirar quines variables explicatives agafar per fer el model, s'agafaran les 4 correlacions més altes que es tenen sobre la variable dependent. En el següent resultat es pot veure com les 4 variables amb més correlació amb 'SalePrice' són: 'OverallQual', 'GrLivArea', 'GarageCars' i 'GarageArea' amb correlacions de 0.7909816, 0.7086245, 0.6404092 i 0.6234314, respectivament.

```
cols <- cor(numeric_columns)
cat('Correlació respecte SalePrice')
```

```
## Correlació respecte SalePrice

sort(cols[,'SalePrice'],decreasing = TRUE)[0:10]

## SalePrice  OverallQual    GrLivArea   GarageCars   GarageArea
1.0000000    0.7909816    0.7086245    0.6404092    0.6234314
##   TotalBsmtSF   X1stFlrSF     FullBath TotRmsAbvGrd    YearBuilt
##   0.6135806     0.6058522    0.5606638    0.5337232    0.5228973
```

Un cop obtingudes les variables independents amb és correlació respecte la variable dependent, s'ha de mirar que entre les variables dependents no hi hagi una correlació molt alta, ja que això implicaria colinealitat. A continuació es miren les correlacions més altes de cada variable escollida prèviament per veure si hi ha colinealitat. Es diu que hi ha una alta correlació entre variables si el coeficient supera el 0.65.

En el següent fragment de codi es pot veure com totes les variables dependents tenen una correlació inferior al 0.65 entre elles excepte 'GarageCars' i 'GarageArea', que tenen una correlació de 0.8824754. Per tant, es pot dir que entre aquestes dues variables independents hi ha colinealitat. Per aquesta raó, una de les dues variables no pot ser escollida per fer el model. S'ha d'escollir a següent variable que tingui més correlació amb la variable dependent. Aquesta nova variable independent és 'TotalBsmtSF', amb una correlació del 0.6135806. Es torna a fer el procés de veure les correlacions amb les demés variables independents escollides i s'obté que cap passa del 0.65.

```
sort(cols[,'OverallQual'],decreasing = TRUE)[0:10]

##  OverallQual    SalePrice   GarageCars    GrLivArea    YearBuilt
GarageArea
##    1.0000000    0.7909816    0.6006707    0.5930074    0.5723228
0.5620218
## YearRemodAdd     FullBath  TotalBsmtSF    X1stFlrSF
##    0.5506839    0.5505997    0.5378085    0.4762238

sort(cols[,'GrLivArea'],decreasing = TRUE)[0:10]

##    GrLivArea TotRmsAbvGrd    SalePrice    X2ndFlrSF     FullBath
OverallQual
##    1.0000000    0.8254894    0.7086245    0.6875011    0.6300116
0.5930074
##    X1stFlrSF BedroomAbvGr   GarageArea   GarageCars
##    0.5660240    0.5212695    0.4689975    0.4672474

sort(cols[,'GarageCars'],decreasing = TRUE)[0:10]
```

```
##   GarageCars   GarageArea    SalePrice  OverallQual    YearBuilt
FullBath
##    1.0000000    0.8824754    0.6404092    0.6006707    0.5378501
0.4696720
##    GrLivArea    X1stFlrSF  TotalBsmtSF YearRemodAdd
##    0.4672474    0.4393168    0.4345848    0.4206222

sort(cols[,'GarageArea'],decreasing = TRUE)[0:10]

##   GarageArea   GarageCars    SalePrice  OverallQual    X1stFlrSF
TotalBsmtSF
##    1.0000000    0.8824754    0.6234314    0.5620218    0.4897817
0.4866655
##    YearBuilt    GrLivArea     FullBath YearRemodAdd
##    0.4789538    0.4689975    0.4056562    0.3715998

sort(cols[,'TotalBsmtSF'],decreasing = TRUE)[0:10]

## TotalBsmtSF   X1stFlrSF   SalePrice OverallQual  BsmtFinSF1  GarageArea
##   1.0000000   0.8195300   0.6135806   0.5378085   0.5223961   0.4866655
##   GrLivArea  GarageCars   BsmtUnfSF   YearBuilt
##   0.4548682   0.4345848   0.4153596   0.3914520
```

A continuació es mostra un resum de com han quedat les correlacios en el dataset filtrat amb les variables seleccionades.

```
library(corrplot)

## corrplot 0.92 loaded

cor(dataset_tot[, c('SalePrice','OverallQual','GrLivArea', 'GarageArea',
'TotalBsmtSF')])

##             SalePrice OverallQual GrLivArea GarageArea TotalBsmtSF
## SalePrice   1.0000000   0.7909816 0.7086245  0.6234314 0.6135806
## OverallQual 0.7909816   1.0000000 0.5930074  0.5620218 0.5378085
## GrLivArea   0.7086245   0.5930074 1.0000000  0.4689975 0.4548682
## GarageArea  0.6234314   0.5620218 0.4689975  1.0000000 0.4866655
## TotalBsmtSF 0.6135806   0.5378085 0.4548682  0.4866655 1.0000000


corrplot(corr = cor(dataset_tot[, c('SalePrice','OverallQual','GrLivArea',
'GarageArea', 'TotalBsmtSF')]), method = "number",
         tl.cex = 0.7,number.cex = 0.8, cl.pos = "n")
```

	SalePrice	OverallQual	GrLivArea	GarageArea	TotalBsmtSF
SalePrice	1.00	0.79	0.71	0.62	0.61
OverallQual	0.79	1.00	0.59	0.56	0.54
GrLivArea	0.71	0.59	1.00	0.47	0.45
GarageArea	0.62	0.56	0.47	1.00	0.49
TotalBsmtSF	0.61	0.54	0.45	0.49	1.00

1.1.1 Explicació de les variables escollides

- **SalePrice**: és la variable dependent. Fa referència al preu de la casa.
- **OverallQual**: és la qualitat del material i de l'acabat.
- **GrLivArea**: és l'àrea habitable i acabada per sobre del nivell del terra d'una casa en peus quadrats.
- **GarageArea**: és l'àrea del garatge en peus quadrats.
- **TotalBsmtSF**: és l'àrea del soterrani en peus quadrats.

```
summary(dataset)
##    SalePrice         OverallQual       GrLivArea   GarageArea
##  Min.   : 55000   Min.   : 4.0   Min.   : 616   Min.   :   0.0
##  1st Qu.:135000   1st Qu.: 5.0   1st Qu.:1240   1st Qu.: 401.5
##  Median :166000   Median : 6.0   Median :1568   Median : 528.0
##  Mean   :195776   Mean   : 6.3   Mean   :1591   Mean   : 520.0
##  3rd Qu.:238782   3rd Qu.: 7.0   3rd Qu.:1874   3rd Qu.: 669.0
##  Max.   :430000   Max.   :10.0   Max.   :3447   Max.   :1052.0
##   TotalBsmtSF
##  Min.   :   0.0
##  1st Qu.: 800.8
##  Median :1035.0
##  Mean   :1129.1
##  3rd Qu.:1453.5
##  Max.   :3200.0

head(dataset)

##      SalePrice OverallQual GrLivArea GarageArea TotalBsmtSF
## 388     125000           6     1125        352         1041
## 543     213250           7     1680        583         1650
```

```
## 836     128000             4     1067         436         1067
## 1326     55000             4     796          0           796
```

1.2 Creació i anàlisi del model

Un cop seleccionades les variables explicatives amb les que es farà el model de regressió lineal múltiple, es crearà un nou dataset filtrat i es farà un sample de 50 d'aquest. Es marcarà la llavor com a 1, d'aquesta manera sempre treballarem sobre el matex sample cada cop que s'executi.

```
set.seed(1)
dataset <- dataset_tot[floor(runif(50, min=1, max=nrow(dataset_tot))),
c('SalePrice','OverallQual','GrLivArea', 'GarageArea', 'TotalBsmtSF')]
model <- lm(dataset$SalePrice~dataset$OverallQual + dataset$GrLivArea +
dataset$GarageArea + dataset$TotalBsmtSF, data = dataset)
summary(model)

##
## Call:
## lm(formula = dataset$SalePrice ~ dataset$OverallQual + dataset$GrLivArea +
##     dataset$GarageArea + dataset$TotalBsmtSF, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87571 -24005    314  21555  69443
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -130049.15   22058.96  -5.896 4.46e-07 ***
## dataset$OverallQual   23841.99    5079.04   4.694 2.53e-05 ***
## dataset$GrLivArea        58.07      12.96   4.481 5.06e-05 ***
## dataset$GarageArea       71.14      31.28   2.274  0.02778 *
## dataset$TotalBsmtSF      40.97      11.88   3.449  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35140 on 45 degrees of freedom
## Multiple R-squared:   0.87,  Adjusted R-squared:  0.8585
## F-statistic: 75.31 on 4 and 45 DF,  p-value: < 2.2e-16
```

Del summary es pot dir que el model lineal múltiple està donat per:

$$Y = -130049.15 + 23841.99X1 + 58.07X2 + 71.14X3 + 40.9X4$$

Correspondència de variables:

- X1: OverallQual
- X2: GrLivArea
- X3: GarageArea
- X4: TotalBsmtSF
- Y: SalePrice

L'intercept és l'ordenada a l'origen i els demés valors són les pendents estimades per cada variable. En altres paraules, es podria dir que, per exemple, per cada unitat addicional de la variable X2, el preu de la casa augmenta en 58.07 €. Aquesta afirmació es pot aplicar de la mateixa manera per totes les variables independents. Veiem que tenen valors alts, ja que el dataset treballa amb un rang de valors alt.

A més, també es poden veure els errors de les desviacions típiques de cadascuna de les variables.

1.3 Supòsits de model

Abans de continuar amb el model, s'ha d'assegurar que compleixi amb un mínim de supòsits. Tot i que hi ha diversos supòsits, s'ha decidit treballar amb els següents: linealitat, normalitat, homogeneïtat, independència i presència d'outliers.

1.3.1 Linealitat

Per comprovar si hi ha linealitat es fan vàries proves:

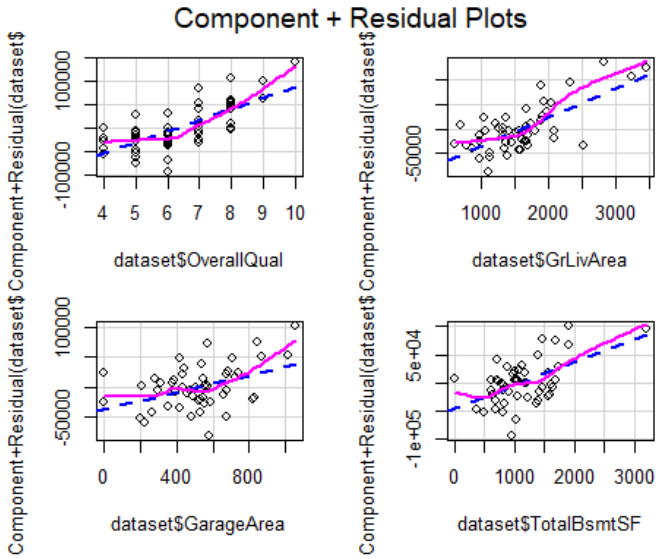
1. Es mira que la mitjana de l'error sigui molt pròxima a 0.
2. Per comprovar visualment, es mira la linealitat perfecte que haurien de seguir les variables i la que realment segueixen (color blau i lila, respectivament).

```
library(car)
```

```
## Loading required package: carData
```

```
##
## Attaching package: 'car'
```

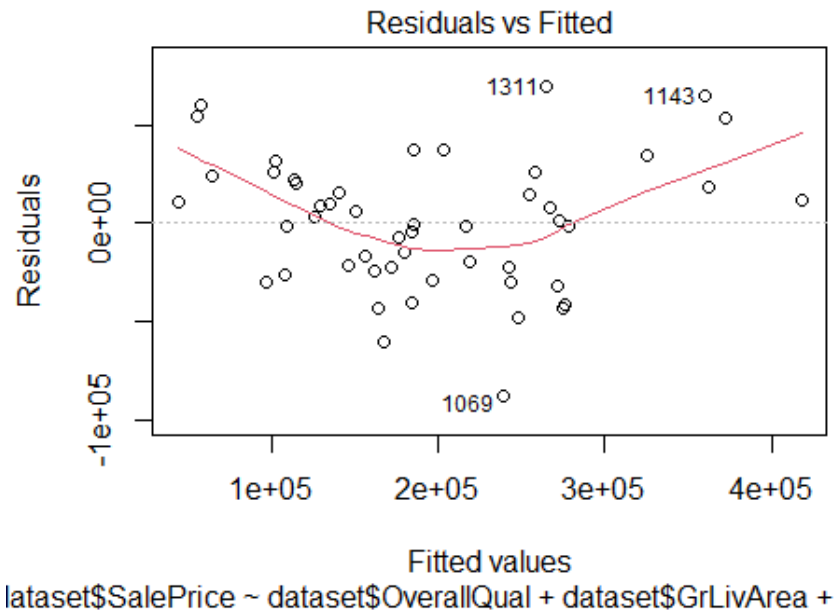
```
## The following object is masked from 'package:dplyr':
##
##     recode

# la mitjana de l'error ha de ser zero
mean(model$residuals)

## [1] -1.090044e-12

# les rectes liles s'han d'aproximar a les discontinues blaves
crPlots(model)
```



En la gràfica anterior, es pot reafirmar com les variables explicatives tenen una correlació positiva respecte la variable explicada, ja que quan una augmenta, l'altra també ho fa.

Aquí es mostra un plot general on ens mostra els errors. Es pot veure que la línia vermella que ens surt no és completament recta però sí que varia poc respecte el 0.

```
plot(model,1)
```



Per tant, s'accepta H0: sí que hi ha linealitat.

- **Multicolinealitat**

Per altra banda, també sha de comprovar que no hi hagi multicolinealitat entre variables independents, fet esmentat anteriorment.

Amb aquesta prova (VIF: Factor d'Inflació de Variança) es comprova l'aparició de multicolinealitat segons els valors resultants. Si el valor és major a 10, vol dir que hi ha multicolinealitat. En aquest cas es pot veure com no hi ha.

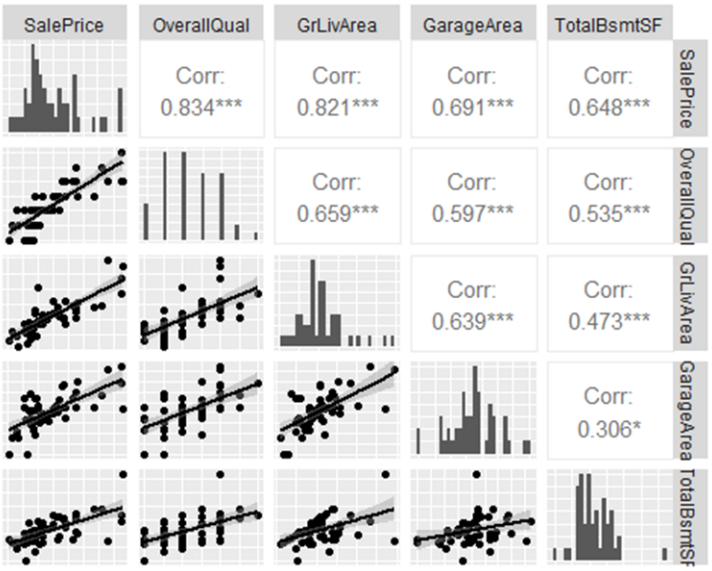
```
vif(model)

## dataset$OverallQual   dataset$GrLivArea  dataset$GarageArea
dataset$TotalBsmtSF
##  2.225003            2.224291            1.884339            1.471255
```

Per últim, es mostra una gràfica resum de la linealitat entre variables, la correlació d'aquestes en el model i la distribució que segueixen.

Tot i que es vegin correlacions altes entre variables explicatives, s'ha comprovat anteriorment que no hi ha multicolinealitat.

Es pot veure com la majoria segueixen una distribució normal. En el següent apartat s'entrarà en més profunditat.

```
library(GGally)
## Loading required package: ggplot2
##
## Attaching package: 'ggplot2'
## The following objects are masked from 'package:psych':
##
##     %+%, alpha
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
ggpairs(dataset, lower = list(continuous = "smooth"),
        diag = list(continuous = "barDiag"), axisLabels = "none")
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```



1.3.2 Normalitat

Per mirar la normalitat es pot analitzar el model o els errors del model. En aquest cas, s'ha estudiat amb els errors.

S'ha estudiat la normalitat utilitzant un shapiro.test. Si el p-value resultant és major a 0.05, es dirà que el model segueix normalitat. Pel contrari, no en seguirà. En el resultat, es pot veure com el p-value és més gran que 0.05, per tant, es té normalitat.

```
library(MASS)
```

```
##
## Attaching package: 'MASS'
```

```
## The following object is masked from 'package:dplyr':
##
##     select
```

```
sresid <- studres(model)
shapiro.test(sresid)
```

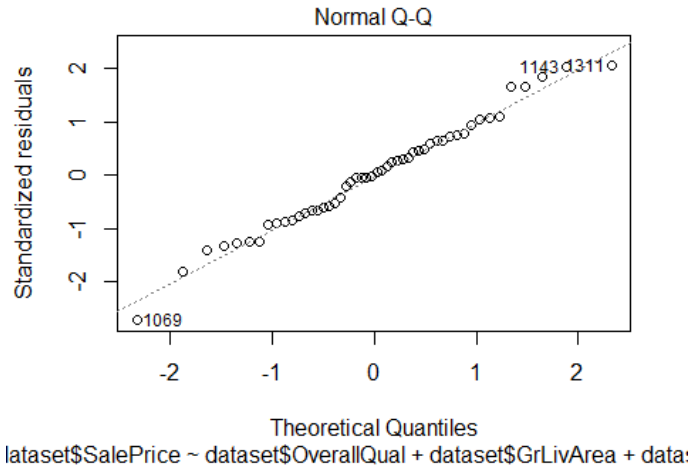
```
##
##  Shapiro-Wilk normality test
##
## data:  sresid
## W = 0.98445, p-value = 0.7478
```

Una altra manera de comprovar aquest supòsit és mostrant plots que mostrin visualment com es comporten les dades i quina distribució segueixen.

En el primer gràfic, es pot veure com els punts estan gaire bé tots per sobre de la recta que haurien de seguir. D'aquesta manera, es pot concloure també que el model segueix una distribució normal.

```
#els punts han d'estar sobre la recta discontinua
plot(model,2)
#la gaussiana s'ha d'aproximar a l'histograma
library(ggfortify)
```

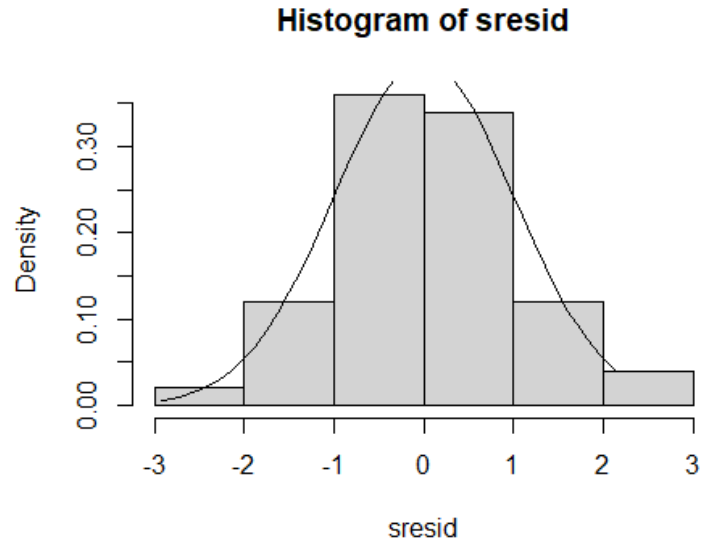
```
## Loading required package: ggplot2
```



lataset$SalePrice ~ dataset$OverallQual + dataset$GrLivArea + data

Per últim, es mostra un histograma on es pot veure clarament com el model segueix una distribució normal.

```
hist(sresid, freq=FALSE)
xfit <- seq(min(sresid),max(sresid),length=40)
yfit <- dnorm(xfit)
lines(xfit,yfit)
```



1.3.3 Homogeneïtat

Per comprovar si hi ha homogeneïtat entre les variàncies de l'error es fan diferents proves.

Primer, s'ha realitzat el test de 'Non Constant Variance Test' (ncvTest) on si el p-value és major a 0.05, es considerarà que la variància és homogènia i es complirà aquest supòsit. Pel contrari, es rebutjarà.

En aquest cas, es pot veure com sí que hi ha homogeneïtat entre les variàncies de l'error.

```
#non-constant variance score test
ncvTest(model)
```

```
## Non-constant Variance Score Test
## Variance formula: ~ fitted.values
## Chisquare = 1.22587, Df = 1, p = 0.26821
```

Una altra prova realitzada ha estat el breusch-pagan test. Si el p-value és major a 0.05, es considerarà que la variància és homogènia i es complirà aquest supòsit. Pel contrari, es rebutjarà.

En aquesta prova també s'accepta aquest supòsit.

```
#breusch-pagan test
library(lmtest)
```

```
## Loading required package: zoo
```

```
##
## Attaching package: 'zoo'
```

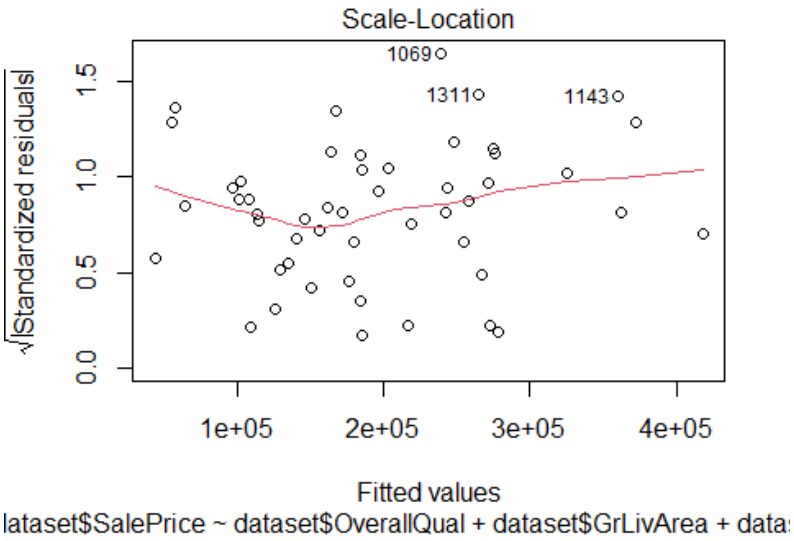
```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
bptest(model)
```

```
##  studentized Breusch-Pagan test
##
## data:  model
## BP = 2.5001, df = 4, p-value = 0.6446
```

Per últim, es mostra un gràfic on es pot apreciar visualment com els errors són homogenis i, per tant, es pot confirmar aquest supòsit.

```
#gràfica la recta del mig ha de ser relativament horitzontal
plot(model, 3)
```



1.3.4 Independència

Per comprovar que es compleix aquest supòsit s'han realitzat varies proves.

$H0$: *els errors són independents*
$H1$: *els errors no són independents*

Prova 1

En la primera prova s'ha comprovat la independència amb el test de Durbin. Com en aquest cas el **p-value és major a 0.05,** es pot dir que els **errors del model són independents**.

```
#test de durbin per comprovar independència
durbinWatsonTest(model)
```

```
##  lag Autocorrelation D-W Statistic p-value
##    1      -0.1114984      2.207599   0.514
##  Alternative hypothesis: rho != 0
```

Prova 2

En la segona prova s'ha fet fet servir el test runs de la llibreria lawstats. Com el test retorna un **p-value major a 0.05**, es pot dir que **es compleix el supòsit d'independència**.

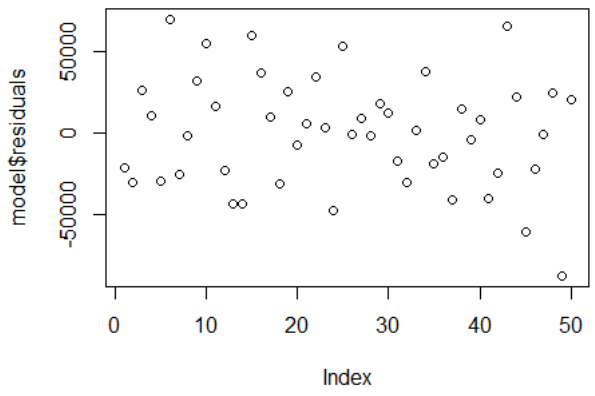
```
#test runs de la llibrerira lawstats tambe per comprobar independencia
library(lawstat)

##
## Attaching package: 'lawstat'

## The following object is masked from 'package:car':
##
##     levene.test

lawstat::runs.test(model$residuals)

##
##  Runs Test - Two sided
##
## data:  model$residuals
## Standardized Runs Statistic = 1.1431, p-value = 0.253
```

Per tant, com els **p-values** d'ambós test són **> 0.05**, **s'accepta la hipòtesi nul·la H_0**. Per tant, es té **evidències estadístiques** per dir que els **errors són independents**.

Prova 3

Es fa un plot dels residuals i es mira si apareix algun patró. En aquest cas, no segueix **cap patró particular**. Per tant, podem dir que **es compleix el supòsit d'independència**.

```
plot(model$residuals)
```

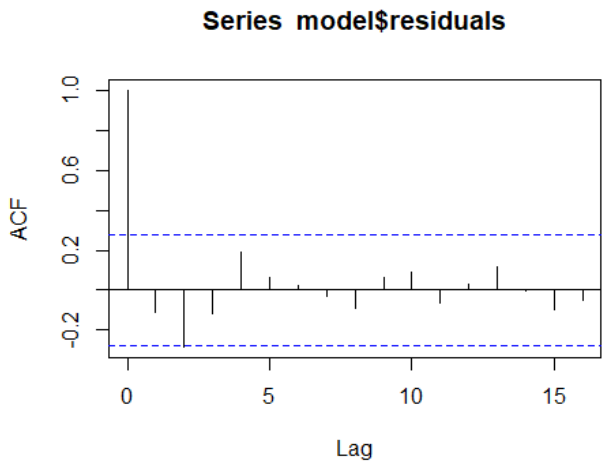


Prova 4

Per últim, s'utilitza la prova Autocorrelation Function (ACF) per mirar si les correlacions estan dins de l'interval de confiança. Si es compleix això, suposarem independència en el model.

En el plot es pot veure com totes les correlacions estan dins de l'interval. La primera de totes és 1, ja que una correlació amb una mateixa variable sempre serà 1. Per tant, podem concloure que el model compleix el supòsit d'independència.

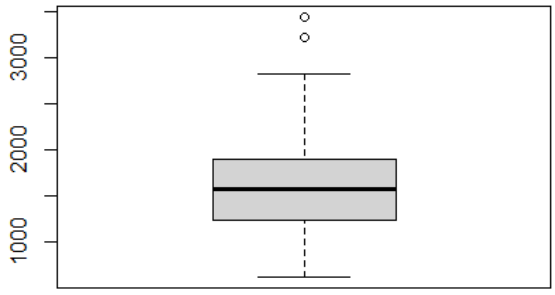
```
acf(model$residuals)
```



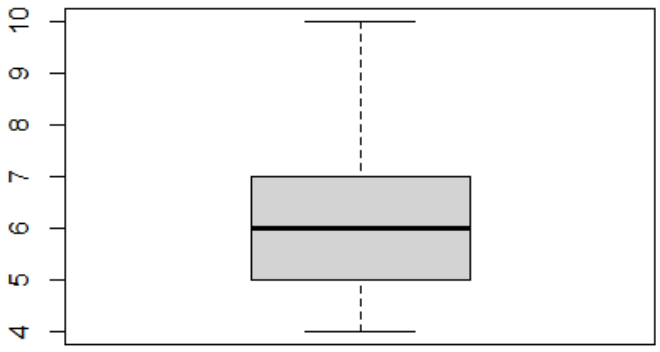
1.3.6 Outliers

En els següents boxplots es pot apreciar com, generalment, no hi ha outliers. Sí que és cert que apareixen alguns però, tot i així, al ser pocs, es pot **acceptar el supòsit d'absència d'outliers**.

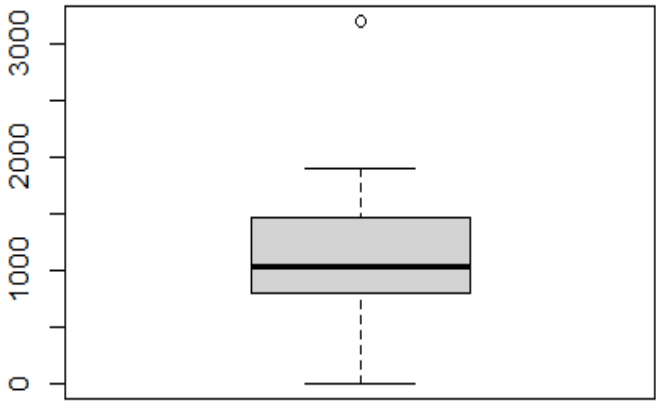
```
# Verificar la ausencia de outliers
boxplot(dataset$GrLivArea)
```



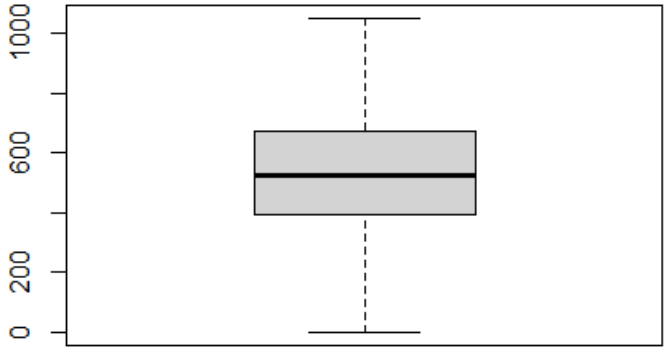
```
boxplot(dataset$OverallQual)
```



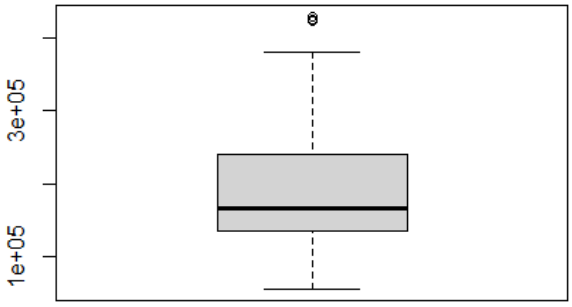

```
boxplot(dataset$TotalBsmtSF)
```




```
boxplot(dataset$GarageArea)
```



```
boxplot(dataset$SalePrice)
```



En conclusió, es pot dir que el model compleix tots els supòsits

1.3.7 Resum dels supòsits de model

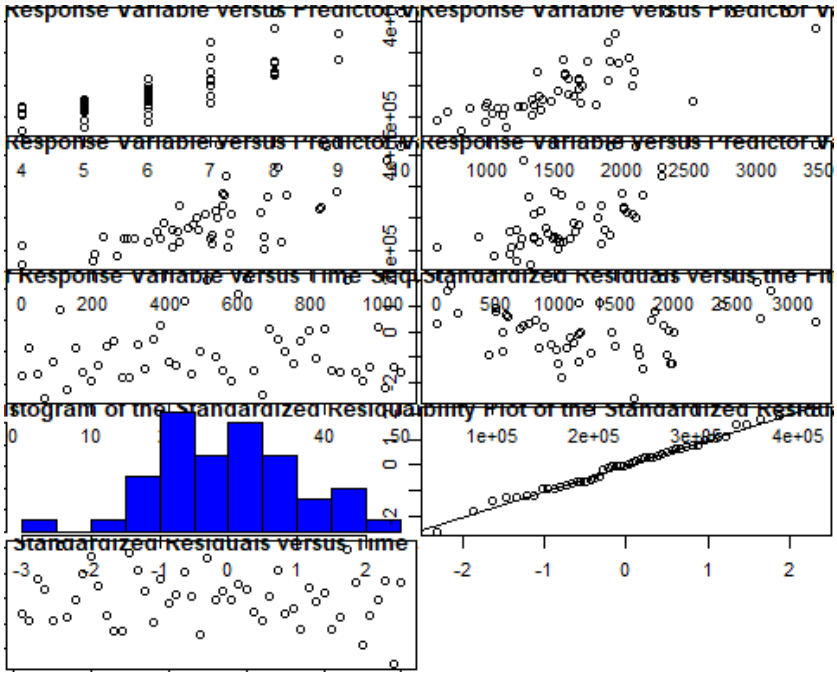
Una altra manera per comprovar tots els supòsits és fer un gvlma del model. No obstant, no és una bona idea fiar-se d'aquesta prova perquè poden veure's modificats/intoxicats per un sol cas atípic. Per aquesta raó s'han fet l'anàlisi dels supòsits un per un anteriorment.

No obstant, en els resultats obtinguts, es pot veure com s'accepten 3 supòsits (value < p-value) i es rebutgen 2 (value > p-value).

```
library(gvlma)
#resum comprovació supostos + les dades de la regressió
summary(gvlma(model))

##
## Call:
## lm(formula = dataset$SalePrice ~ dataset$OverallQual + dataset$GrLivArea +
##     dataset$GarageArea + dataset$TotalBsmtSF, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87571 -24005    314  21555  69443
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -130049.15   22058.96  -5.896 4.46e-07 ***
## dataset$OverallQual   23841.99    5079.04   4.694 2.53e-05 ***
## dataset$GrLivArea        58.07      12.96   4.481 5.06e-05 ***
## dataset$GarageArea       71.14      31.28   2.274  0.02778 *
## dataset$TotalBsmtSF      40.97      11.88   3.449  0.00123 **
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35140 on 45 degrees of freedom
## Multiple R-squared:   0.87,  Adjusted R-squared:  0.8585
## F-statistic: 75.31 on 4 and 45 DF,  p-value: < 2.2e-16
##
##
## ASSESSMENT OF THE LINEAR MODEL ASSUMPTIONS
## USING THE GLOBAL TEST ON 4 DEGREES-OF-FREEDOM:
## Level of Significance =  0.05
##
## Call:
##  gvlma(x = model)
##
##                       Value   p-value                   Decision
## Global Stat        13.89779 0.0076285 Assumptions NOT satisfied!
## Skewness            0.02373 0.8775716    Assumptions acceptable.
## Kurtosis            0.06758 0.7948929    Assumptions acceptable.
## Link Function      13.48955 0.0002399 Assumptions NOT satisfied!
## Heteroscedasticity  0.31693 0.5734569    Assumptions acceptable.
```

```
par(mar = c(0.1, 0.1, 0.1, 0.1))
plot(gvlma(model))
```



1.4 Intervals de confiança

En aquest apartat, es tractaran els intervals de confiança per la variable dependent i l'interval de confiança general per totes les variables. En aquest cas, s'ha volgut centrar en l'interval de confiança per la mitjana, en l'interval de confiança per la variància i en l'interval de confiança per la proporció.

- Interval de confiança per totes les variables del model.

```
confint(model)
##                               2.5 %       97.5 %
## (Intercept)          -1.744782e+05 -85620.12407
## dataset$OverallQual   1.361228e+04  34071.69376
## dataset$GrLivArea     3.197070e+01     84.17173
## dataset$GarageArea    8.131007e+00     134.15173
## dataset$TotalBsmtSF   1.704617e+01     64.89285
```

- Interval de confiança per la mitjana

L'interval de confiança per la mitjana de la variable 'SalePrice' és el següent. També es pot veure com si es fa una mitjana de la variable, efectivament està dins de l'interval.

```
# Calcular el intervalo de confianza para la media
t_test <- t.test(dataset$SalePrice)
confint_mean <- t_test$conf.int
confint_mean

## [1] 169229.0 222322.2
## attr(,"conf.level")
## [1] 0.95

mean(dataset$SalePrice)

## [1] 195775.6
```

- Interval de confiança per la variància

L'interval de confiança per la mitjana de la variable 'SalePrice' és el següent. També es pot veure com si es mira la variància de la variable, efectivament està dins de l'interval.

```
# Calcular el intervalo de confianza para la varianza
library(stests)

##
## Attaching package: 'stests'
```

```
## The following object is masked from 'package:stats':
##
##     var.test

var_test <- var.test(dataset$SalePrice)
confint_var <- var_test$conf.int
confint_var

## [1]  6088360524 13549057281
## attr(,"conf.level")
## [1] 0.95

var(dataset$SalePrice)

## [1] 8725293278
```

- Interval de confiança per la proporció

En aquest cas, s'ha volgut esbrinar quina proporció de cases valien menys que el preu mitjà de totes les cases.

L'interval de confiança per la proporció de la variable 'SalePrice', amb el threshold com a mitjana de preus, és el següent. Es pot veure com si es mira la proporció de preus que estan sota del preu mitjà, efectivament està dins de l'interval.

```
# Calcular el intervalo de confianza para la proporción
prop_test <-
prop.test(length(dataset[dataset$SalePrice<mean(dataset$SalePrice),]),length(
dataset$SalePrice))
confint_prop <- prop_test$conf.int
confint_prop

## [1] 0.03740462 0.22591184
## attr(,"conf.level")
## [1] 0.95

length(dataset[dataset$SalePrice<mean(dataset$SalePrice),])/length(dataset$Sa
lePrice)

## [1] 0.1
```

1.5 Validació global

Per comprovar la validació global s'aplicaran diferents mètodes:

1.5.1 Comprovació manual

El que es farà en aquest apartat és aplicar la teoria que s'ha vist a classe i trobar el F_0 de forma manual.

Es trobaran tots el elements de l'anova manualment i es mostraran els resultats obtinguts en una taula ANOVA feta a mà (mostrada en el codi com a comentari).

1. **Plantejament d'hipòtesis**

$$H_0: \beta_1 = \beta_2 = \cdots = \beta_4 = 0$$

$$H_1: \beta_i \neq \beta \quad i \neq j$$

2. $\alpha = 0.05$
3. **Estadístic de contrast (ANOVA)**

FV	gl	SS	MS	F_0
Regression	k = 4	371975267875	92993816969	75.3134
Error	n-k-1 = 45	55564102765	1234757839	
Total	n-1 = 49	427539370640		

4. **Regió de rebuig**

Per veure si el model és globalment estadísticament significatiu, s'ha de comparar el resultat de la F_0 obtinguda amb els valors de la taula de Fisher.

Per rebutjar H0 s'ha de complir:

- $F(0.05, 4, 45) < F_0 = 0.1752145 < 75.3134$

SÍ es compleix. Per tant, **rebutgem H_0**.

Per tant, amb un 95% de confiança, es pot dir que **el model és globalment estadísticament significatiu.**

```r
#estimado --> yt = beta0 + beta1*X1t + beta2*X2t + beta3*X3t + beta4*X4t

#mult = (Xt*X)
#solve(mult) = (Xt*X)**-1
u <- rep(1,nrow(dataset))
X1 <- dataset$OverallQual
X2 <- dataset$GrLivArea
X3 <- dataset$GarageArea
X4 <- dataset$TotalBsmtSF
y <- dataset$SalePrice
matriz = t(rbind(u,X1,X2,X3,X4)) #t: transposta; en R se lee la transpuesta
al revés --> así que para ponerla normal hacemos la transpuesta

#multiplicar matrices en R
#A%*%B
mult = t(matriz)%*%matriz
mult
```

```
##        u     X1        X2       X3       X4
## u     50    315     79531    25998    56457
## X1   315   2091    528563   173287   375507
## X2 79531 528563 142858153 45338021 96669558
## X3 25998 173287  45338021 15895200 31049264
## X4 56457 375507  96669558 31049264 76624111
```

```r
det = det(mult)
det
```

```
## [1] 5.499634e+23
```

```r
inv = solve(mult) #inversa
inv
```

```
##                 u            X1            X2            X3            X4
## u   3.940834e-01 -6.230562e-02 -1.890460e-06  2.959779e-05  5.366313e-06
## X1 -6.230562e-02  2.089204e-02 -1.786931e-05 -4.126609e-05 -1.721156e-05
## X2 -1.890460e-06 -1.786931e-05  1.360044e-07 -1.373762e-07 -2.695333e-08
## X3  2.959779e-05 -4.126609e-05 -1.373762e-07  7.926449e-07  3.254534e-08
## X4  5.366313e-06 -1.721156e-05 -2.695333e-08  3.254534e-08  1.142611e-07
```

```r
#matriz2 = Xt*y
matriz2 = t(matriz)%*%y
matriz2
```

```
##           [,1]
## u      9788780
## X1    67294626
## X2 17740914484
## X3  5786202454
## X4 12572507272
```

```r
#m_Def = (Xt*X)**-1  *  Xt*y
m_Def = inv%*%matriz2
m_Def #beta
```

```
##             [,1]
## u  -130049.14552
## X1   23841.98863
## X2      58.07122
## X3      71.14137
## X4      40.96951
```

```r
#yt*y
yt_y = t(y)%*%y
yt_y
```

```
##              [,1]
## [1,] 2.343944e+12
```

```r
SSE = yt_y - (t(m_Def) %*% t(matriz) %*% y)
SSE
```

```
##             [,1]
## [1,] 55564102765
```

```r
sumatori(yi)**2)
SST = yt_y - (sum(y)**2)/nrow(dataset)
SST
```

```
##              [,1]
## [1,] 427539370640
```

```r
SSReg = SST - SSE
SSReg
```

```
##              [,1]
## [1,] 371975267875
```

```r
#SSE = yt*y - betat*xt*y
#SST = yt*y - (sumatori(yi)**2)/n
#SSReg = SSE - SST

#VALIDACIÓN GLOBAL --> con fisher
#1) H0: betai = 0 (beta0 = beta1 = beta2 = 0),  i = 0,..,2
#   H1: betai != betaj,                         i != j


#2) alpha = 0.05
#3) ANOVA
k = (ncol(dataset)-1)
n = nrow(dataset)
```

```
MSReg = SSReg/k
MSE = SSE/(n-k-1)
f0 = MSReg/MSE
# ANOVA
# FV            gL            SS            MS            F0
# Regression    k = 4         371975267875  92993816969   75.3134
# Error         n-k-1 = 45    55564102765   1234757839
# Total         n-1 = 49      427539370640


#F alpha (k,n-k-1) -> F 0.05 (4,45) =
Fisher = qf(0.05,k,n-k-1)
Fisher
```

```
## [1] 0.1752145
```

```
Fisher_reves = qf(0.95,k,n-k-1)
Fisher_reves
```

```
## [1] 2.578739
```

1.5.2 Comprovació amb codi

Amb la comanda summary(model) també es pot saber directament si el model és globalment estadísticament significatiu.

Si es mira la última línia que retorna aquest fragment de codi (F-statistic) es pot veure com el F_0 pren el valor de 75.31, igual que el valor que s'ha trobat fent la validació manuelment. No obstant, aquí s'ha de tenir en compte el p-value que es retorna en aquesta mateixa línia. Pren el valor de 2.2e-16. Com el p-valor < alpha (0.05), **rebutgem la hipòtesi nul·la.**

Per tant, tenim evidències per dir que el model, amb un 95% de confiança, **és globalment estadísticament significatiu**.

```
summary(model)
```

```
##
## Call:
## lm(formula = dataset$SalePrice ~ dataset$OverallQual + dataset$GrLivArea +
##     dataset$GarageArea + dataset$TotalBsmtSF, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87571 -24005    314  21555  69443
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
```

```
## (Intercept)         -130049.15   22058.96  -5.896 4.46e-07 ***
## dataset$OverallQual   23841.99    5079.04   4.694 2.53e-05 ***
## dataset$GrLivArea        58.07      12.96   4.481 5.06e-05 ***
## dataset$GarageArea       71.14      31.28   2.274  0.02778 *
## dataset$TotalBsmtSF      40.97      11.88   3.449  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35140 on 45 degrees of freedom
## Multiple R-squared:   0.87,  Adjusted R-squared:  0.8585
## F-statistic: 75.31 on 4 and 45 DF,  p-value: < 2.2e-16
```

1.6 Validació Individual

En la validació individual es comprova si cada variable és estadísticament significativa dins del model. Per fer-ho, es farà un anàlisi de les variàncies, també conegut com a ANOVA. Es comprova que els **p-valor** de cada variable siguin **< 0.05** perquè s'està utilitzant un 95% de confiança. Per tant, podem dir que es **rebutja la hipòtesi nul·la per cadascuna de les variables**. Aleshores, per cada variable, es pot dir que, amb un 95% de confiança, hi ha evidències estadístiques per dir que cadascuna de les **variables són significatives pel model**.

```
summary(model)
```

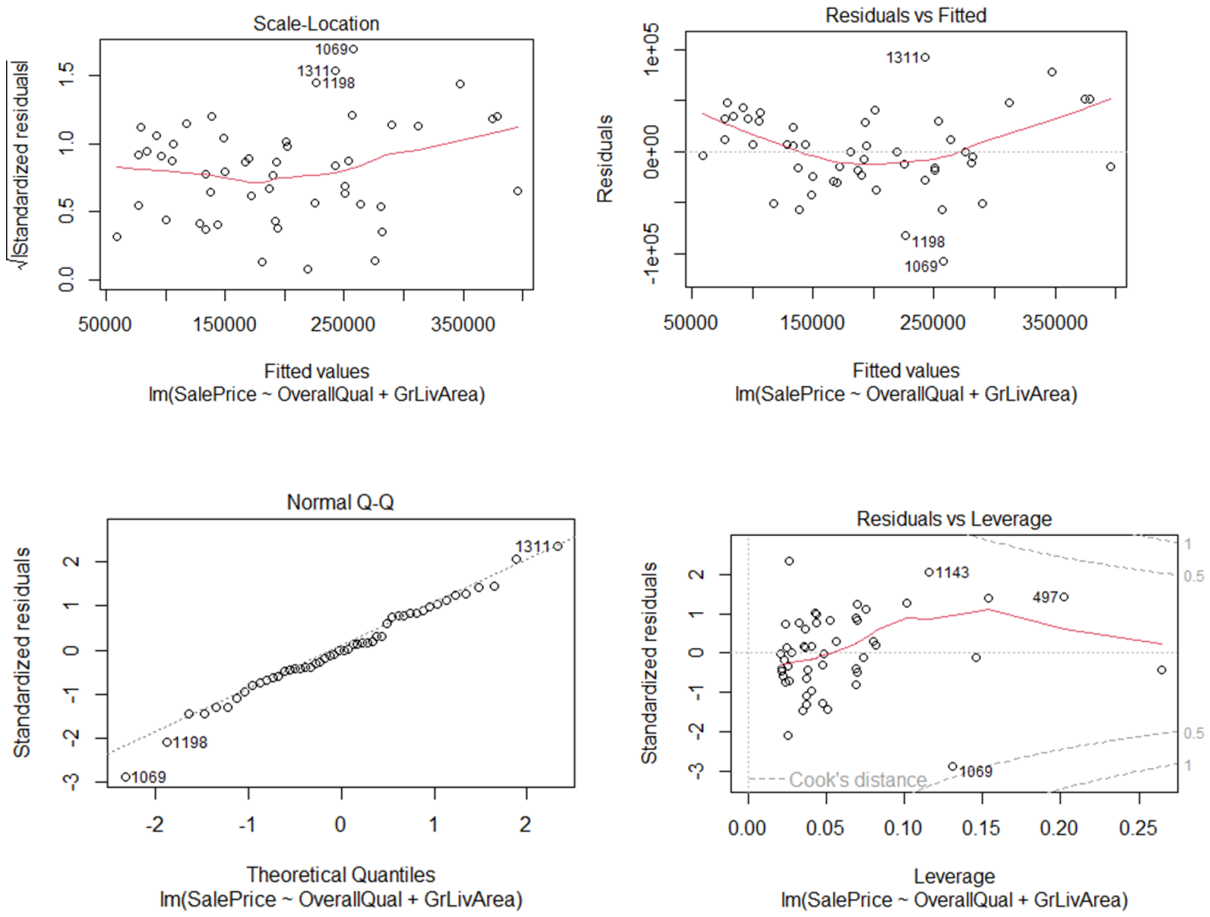
```
##
## Call:
## lm(formula = dataset$SalePrice ~ dataset$OverallQual + dataset$GrLivArea +
##     dataset$GarageArea + dataset$TotalBsmtSF, data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -87571 -24005    314  21555  69443
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)         -130049.15   22058.96  -5.896 4.46e-07 ***
## dataset$OverallQual   23841.99    5079.04   4.694 2.53e-05 ***
## dataset$GrLivArea        58.07      12.96   4.481 5.06e-05 ***
## dataset$GarageArea       71.14      31.28   2.274  0.02778 *
## dataset$TotalBsmtSF      40.97      11.88   3.449  0.00123 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 35140 on 45 degrees of freedom
## Multiple R-squared:   0.87,  Adjusted R-squared:  0.8585
## F-statistic: 75.31 on 4 and 45 DF,  p-value: < 2.2e-16
```

Es pot veure que les 4 variables son estadísticament significatives amb un 95% de confiança. Però si s'agafen les més significatives hi ha 2 que estan al voltant del zero. OverallQual i GrLivArea.

1.7 Model Restringit

En aquest apartat es farà un model únicament amb les variables estadísticament significatives. Com que totes ho son, s'agafaran les més significatives, OverallQual i GrLivArea.

```
model_millor <- lm(SalePrice~OverallQual+GrLivArea,data = dataset)
summary(model_millor)

##
## Call:
## lm(formula = SalePrice ~ OverallQual + GrLivArea,
##     data = dataset)
##
## Residuals:
##    Min     1Q Median     3Q    Max
## -94491 -23124  -5734  26939  82337
##
## Coefficients:
##               Estimate Std. Error t value Pr(>|t|)
## (Intercept) -132705.60   23005.05  -5.769 6.43e-07 ***
## OverallQual   32774.01    5140.62   6.375 7.26e-08 ***
## GrLivArea      77.58       13.12   5.914 3.63e-07 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 36700 on 46 degrees of freedom
## Multiple R-squared:  0.8551, Adjusted R-squared:  0.8457
## F-statistic: 90.49 on 3 and 46 DF,  p-value: < 2.2e-16
plot(model_millor)
```



1.8 Comparació i predicció de models

En aquest apartat es compara el model original i el nou fet amb les variables més significatives. La comparació esta basada en els errors comesos per cada model.

El segon model dona més error tant en MSE, RMSE i MAE i un menor valor per R^2. Per tant, és millor el model inicial amb les 4 variables que el segon amb 2 variables.

Tot i així, es pot observar com el model restringit compleix els supòsits en l'apartat anterior.

```
### predicció models ###
prediction <- function(y_pred, y_true) {
  mse <- mean((y_true - y_pred)^2)
  rmse <- sqrt(mse)
```

```
  mae <- mean(abs(y_pred - y_true))
  cat(paste0("\nMSE: ", round(mse, 5),'\n'))
  cat(paste0("RMSE: ", round(rmse, 5),'\n'))
  cat(paste0("MAE: ", round(mae, 5),'\n'))
  correl <- cor(y_pred, y_true)
  cat(paste0("R-squared entre y_pred i y_true: ", round(correl, 5),'\n'))
}
set.seed(2)
df_test   <-   dataset_tot[floor(runif(50,   min=1,   max=nrow(dataset_tot))),
c('GrLivArea','OverallQual','TotalBsmtSF', 'GarageArea', 'SalePrice')]


#comparació dels 2 models

#el millor model té errors més elevats. Per tant, és millor el model amb
totes les variables escollides inicialment.

#model sencer
pred_model_sencer = predict(model, newdata = df_test)
prediction(pred_model_sencer,dataset$SalePrice)

##
## MSE: 1111282055.29961
## RMSE: 33335.8974
## MAE: 26770.20656
## R-squared entre y_pred i y_true: 0.93276

#nou model
pred_nou_model = predict(model_millor, newdata =
df_test[,c('OverallQual','GrLivArea')]))
prediction(pred_nou_model,dataset$SalePrice)

##
## MSE: 16400486709.6908
## RMSE: 128064.38502
## MAE: 92961.50281
## R-squared entre y_pred i y_true: 0.12499
```

Exercici 2: Regressió logística i Regressió amb Poisson

Aquest segon exercici consisteix en fer dos models. El primer serà una regressió logística per la qual s'utilitzarà la funció glm amb una família binomial. L'altre serà Poisson i s'utilitzarà també la funció glm però amb una família de tipus Poisson.

El dataset no tenia cap columna amb dades binàries així que s'ha hagut de fer una. A partir de la columna de SalePrice, s'ha agafat la mitja i s'ha creat una columna anomenada 'Overpriced' on si la casa és més cara que la mitja (variable 'SalePrice' per sobre de 195775.6) hi haurà un 1 i si està per sota hi haurà un 0.

Les dades escollides per predir si el preu de la casa és excessiu (Overpriced) són l'àrea de la zona on està construïda la casa (GrLivArea), la qualitat general (OverallQual) i l'àrea total del soterrani (TotalBsmtSF).

2.1 Regressió logística

La regressió logística serveix per explicar una variable binària a partir d'unes dades. No es pot utilitzar regressió lineal ja que no ens estaria dient si és 0 o 1 sinó un valor intermig. El que fa el model és obtenir una probabilitat entre 0 i 1, i a partir d'un threshold decidir si és una classe o altre. En el cas d'aquest treball el threshold serà la probabilitat arrodonida, és a dir, si <0.5 serà 0 i si >=0.5 serà 1

2.1.1 Interpretació dels paràmetres

Els paràmetres del model binomial indiquen la probabilitat logarítmica de pertànyer a la classe 1 (preu superior a 195775.6).

```
mediana <- median(dataset$SalePrice)
quantil3 <- quantile(dataset$SalePrice)[4]
mitjana <- mean(dataset$SalePrice)
dataset$Overpriced <- ifelse(dataset$SalePrice>mitjana,1,0)
reg_log <- glm(Overpriced~GrLivArea + OverallQual + TotalBsmtSF, family =
'binomial', data = dataset)
summary(reg_log)
```

```
##
## Call:
## glm(formula = Overpriced ~ GrLivArea + OverallQual + TotalBsmtSF,
##     family = "binomial", data = dataset)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q       Max
## -1.21082  -0.09850  -0.00706   0.03691   2.51997
##
## Coefficients:
```

```
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -32.879069  12.962525  -2.536   0.0112 *
## GrLivArea     0.002351   0.002182   1.078   0.2812
## OverallQual   3.528727   1.525170   2.314   0.0207 *
## TotalBsmtSF   0.005255   0.002745   1.914   0.0556 .
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.029  on 49  degrees of freedom
## Residual deviance: 13.032  on 46  degrees of freedom
## AIC: 21.032
##
## Number of Fisher Scoring iterations: 8
```

2.1.2 Intervals de confiança

La funció confint ensenya els intervals de confiança per als paràmetres d'un model. Aquests intervals proporcionen una mesura sobre els paràmetres que pot servir per avaluar la significació estadística dels resultats. Si algun interval per algun paràmetre conté el valor zero, hi ha la possibilitat que el paràmetre no tingui cap efecte en el model.

En aquest cas els paràmetres GrLivArea i TotalBsmtSF poden no tenir significació estadística en el model.

```
confint(reg_log)
```

```
##                      2.5 %        97.5 %
## (Intercept) -70.511019109 -15.479276914
## GrLivArea    -0.001972228   0.007308426
## OverallQual   1.343478062   7.751237641
## TotalBsmtSF   0.000938634   0.012585562
```

2.1.3 Matriu de variàncies i covariàncies

La funció vcov té com a output la matriu de covariàncies del model. En la diagonal principal hi ha les variàncies del model i la resta de valors son les covariàncies entre els paràmetres.

```
print(vcov(reg_log)[col(vcov(reg_log))==row(vcov(reg_log))])
```

```
## [1] 1.680271e+02 4.760901e-06 2.326142e+00 7.536880e-06
```

La matriu de correlació dels paràmetres es treu de la de covariàncies, utilitzant la funció cov2cor.

Es pot veure com el valor absolut de la correlació de les variables dependents és > 0.5 i la resta de casos < 0.5.

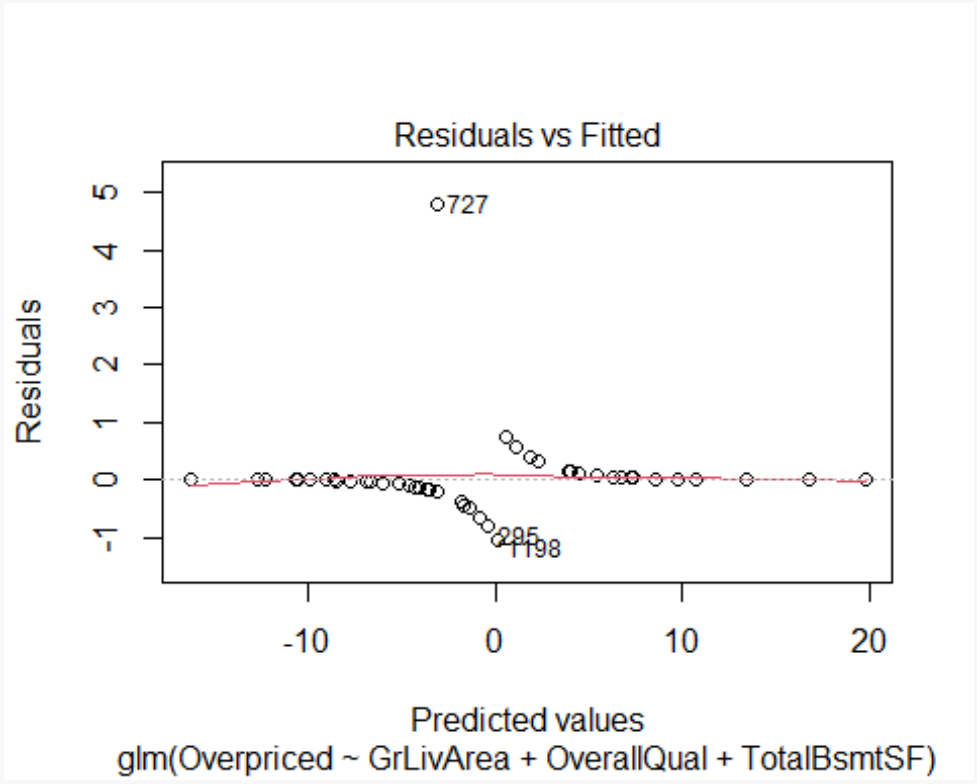
```
print(cov2cor(vcov(reg_log)))
```

```
##             (Intercept)  GrLivArea OverallQual TotalBsmtSF
## (Intercept)   1.0000000 -0.4508734  -0.9253768  -0.6990754
## GrLivArea    -0.4508734  1.0000000   0.1425571   0.2071901
## OverallQual  -0.9253768  0.1425571   1.0000000   0.5366162
## TotalBsmtSF  -0.6990754  0.2071901   0.5366162   1.0000000
```

2.1.4 Residuals

El P-valor del Breusch-Pagan test és >0.05 per tant s'accepta H0. Hi ha evidències estadístiques per dir que el residu és homoscedàstic..

```
plot(reg_log,1)
```




```
bptest(reg_log)
```

```
##
##  studentized Breusch-Pagan test
##
```

```
## data:  reg_log
## BP = 1.9471, df = 3, p-value = 0.5835
```

2.1.5 Prediccions

Com ja s'ha vist en el punt 1.8, en regressió lineal s'utilitza R^2 per avaluar la qualitat del model. En regressió logística no serveix R^2 per tant s'haurà d'utilitzar una altra métrica. Es farà servir la McFadden R^2. Aquesta métrica funciona en el rang del 0 al 1. Un model que s'apropi més al 1 serà millor que un proper al 0.

En el cas del model logístic amb el que s'està treballant dona un valor de 0.8084.

```
library(pscl)

## Classes and Methods for R developed in the
## Political Science Computational Laboratory
## Department of Political Science
## Stanford University
## Simon Jackman
## hurdle and zeroinfl functions by Achim Zeileis

pR2(reg_log)["McFadden"]

## fitting null model for pseudo-r2

##  McFadden
##  0.8084371
```

En quan a les prediccions, s'agafarà un dataset de test amb el que s'avaluarà el model i, posteriorment, es farà una 'confusion matrix' per veure la seva accuracy i altres métriques d'avaluació.

El model té un accuracy del 60%. No és gaire bon resultat. Per tant, és recomanable no utilitzar aquest model.

Altres métriques com la sensitivity o la specificity donen valors com 65.5% i 52.4%, respectivament.

```
library(e1071)
set.seed(10)
test <- dataset_tot[floor(runif(50, min=1, max=nrow(dataset_tot))),
c('GrLivArea', 'OverallQual', 'TotalBsmtSF', 'SalePrice')]
test$Overpriced = ifelse(test$SalePrice > mediana, 1, 0)

true_labels <- factor(test$Overpriced,levels = c(0:1))
# Predcicó
```

```
y_hat <- factor(round(predict(reg_log, new_data = test[,c('GrLivArea',
'OverallQual', 'TotalBsmtSF')],type='response')),levels = c(0,1))
conf_mat = confusionMatrix(data = true_labels, reference = y_hat, dnn =
c('0bservacions','Prediccions'))
print(conf_mat)
```

```
## Confusion Matrix and Statistics
##
##             Prediccions
## 0bservacions  0  1
##            0 19 10
##            1 10 11
##
##                Accuracy : 0.6
##                  95% CI : (0.4518, 0.7359)
##     No Information Rate : 0.58
##     P-Value [Acc > NIR] : 0.4461
##
##                   Kappa : 0.179
##
##  Mcnemar's Test P-Value : 1
##
##             Sensitivity : 0.6552
##             Specificity : 0.5238
##          Pos Pred Value : 0.6552
##          Neg Pred Value : 0.5238
##              Prevalence : 0.58
##          Detection Rate : 0.3800
##    Detection Prevalence : 0.58
##       Balanced Accuracy : 0.5895
##
##        'Positive' Class : 0
##
```

```
predict(reg_log, type = 'response')
```

```
##          388          543          836         1326          295
1311
## 2.681999e-02 9.883686e-01 2.373270e-05 3.021240e-06 4.011665e-01
9.981459e-01
##         1379          965          918           91          301
258
## 1.060478e-03 6.384039e-01 4.939264e-05 8.181689e-08 2.541520e-03
9.823971e-01
##         1003          561         1124          727         1048
1448
## 9.993399e-01 9.490026e-03 4.886041e-05 4.178930e-02 4.501126e-04
9.987403e-01
##          555         1135         1364          310          951
```

```
184
## 8.672579e-01 4.183175e-02 5.767122e-03 9.999985e-01 1.726779e-04
9.796527e-01
##          390          564           20          558         1269
497
## 9.999999e-01 1.598315e-01 1.254177e-03 4.664121e-06 9.993121e-01
1.000000e+00
##          704          875          721          272         1208
976
## 1.135477e-04 1.204721e-04 9.998160e-01 9.092494e-01 7.556159e-01
1.990730e-01
##         1159          158         1056          601         1198
945
## 9.993693e-01 9.827953e-01 1.350655e-01 9.954454e-01 5.195559e-01
1.416097e-02
##         1143          807          773         1152           35
697
## 9.999389e-01 4.171607e-04 2.476402e-02 1.654253e-02 9.999786e-01
2.618807e-05
##         1069         1011
## 3.038191e-01 2.109701e-04
```

2.2 Regressió amb Poisson

La regressió de Poisson és una tècnica estadística utilitzada per analitzar les dades d'un recompte. Es basa en el model de Poisson, que suposa que la mitjana i la variància d'un recompte són iguals. Per al dataset estudiat, el que farà Poisson és comptar quantes cases son Overpriced (1) suposant unes condicions com l'àrea del terreny , la qualitat de la casa i l'àrea del soterrani.

2.2.1 Interpretar els paràmetres

A l'hora d'interpretar, es pot veure que a diferència de la regressió logística, amb Poisson hi ha només una variable que és estadísticament significativa (OverallQual). Les altres dues donen uns P-valors molt elevats.

Els paràmetres de regressió de Poisson, són probabilitats expressades amb logaritmes, per tant, per saber les probabilitats reals s'ha de fer l'exponencial. Es pot veure que amb les dues variables que no són estadísticament significatives l'exponencial dona pràcticament 1 això vol dir que quan augmenten, la casa té un 0% de possibilitats de ser molt cara. En canvi, la variable OverallQual en dona 1.673. Vol dir que per cada unitat que augmenti OverallQual, la casa té un 60% més de possibilitats de ser molt cara. Aquest resultat té molt sentit, ja que com millor és la qualitat d'una casa, més cara serà.

```
poiss<- glm(Overpriced~GrLivArea + OverallQual + TotalBsmtSF, family =
'poisson', data = dataset)
summary(poiss)

##
## Call:
## glm(formula = Overpriced ~ GrLivArea + OverallQual + TotalBsmtSF,
##     family = "poisson", data = dataset)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.0762  -0.5685  -0.4367   0.3142   1.2046
##
## Coefficients:
##               Estimate Std. Error z value Pr(>|z|)
## (Intercept) -5.2508741  1.2077090  -4.348 1.38e-05 ***
## GrLivArea    0.0002771  0.0004070   0.681  0.49599
## OverallQual  0.5148163  0.1790061   2.876  0.00403 **
## TotalBsmtSF  0.0002198  0.0003995   0.550  0.58222
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for poisson family taken to be 1)
##
##     Null deviance: 36.435  on 49  degrees of freedom
## Residual deviance: 17.885  on 46  degrees of freedom
## AIC: 67.885
##
## Number of Fisher Scoring iterations: 5

exp(poiss$coefficients)

## (Intercept)   GrLivArea OverallQual TotalBsmtSF
## 0.005242934 1.000277165 1.673331094 1.000219783

anova(poiss, test = 'Chisq')

## Analysis of Deviance Table
##
## Model: poisson, link: log
##
## Response: Overpriced
##
## Terms added sequentially (first to last)
##
##
##             Df Deviance Resid. Df Resid. Dev Pr(>Chi)
## NULL                           49     36.435
## GrLivArea    1   8.9222        48     27.513 0.002817 **
## OverallQual  1   9.3309        47     18.182 0.002253 **
```

```
## TotalBsmtSF  1   0.2973        46     17.885 0.585585
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

2.2.2 Intervals de confiança

La funció confint ensenya els intervals de confiança per als paràmetres d'un model. Aquests intervals proporcionen una mesura sobre els paràmetres que pot servir per avaluar la significació estadística dels resultats. Si algun interval per algun paràmetre conté el valor zero, hi ha la possibilitat que el paràmetre no tingui cap efecte en el model.

En aquest cas els paràmetres GrLivArea i TotalBsmtSF poden no tenir significació estadística en el model.

```
confint(poiss)
```

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

```
##                      2.5 %       97.5 %
## (Intercept) -7.7863209281 -3.026506088
## GrLivArea   -0.0006058035  0.001009632
## OverallQual  0.1604694492  0.868573120
## TotalBsmtSF -0.0006093729  0.001015571
```

2.2.3 Matriu de variàncies i covariàncies

La funció vcov te com a output la matriu de covariàncies del model. En la diagonal principal hi ha les variàncies del model i la resta de valors son les covariàncies entre els paràmetres.

```
print(vcov(poiss)[col(vcov(poiss))==row(vcov(poiss))])
```

```
## [1] 1.458561e+00 1.656877e-07 3.204318e-02 1.595606e-07
```

Es pot veure com el valor absolut de la correlació de les variables dependents es > 0.5 i la resta de casos < 0.5.

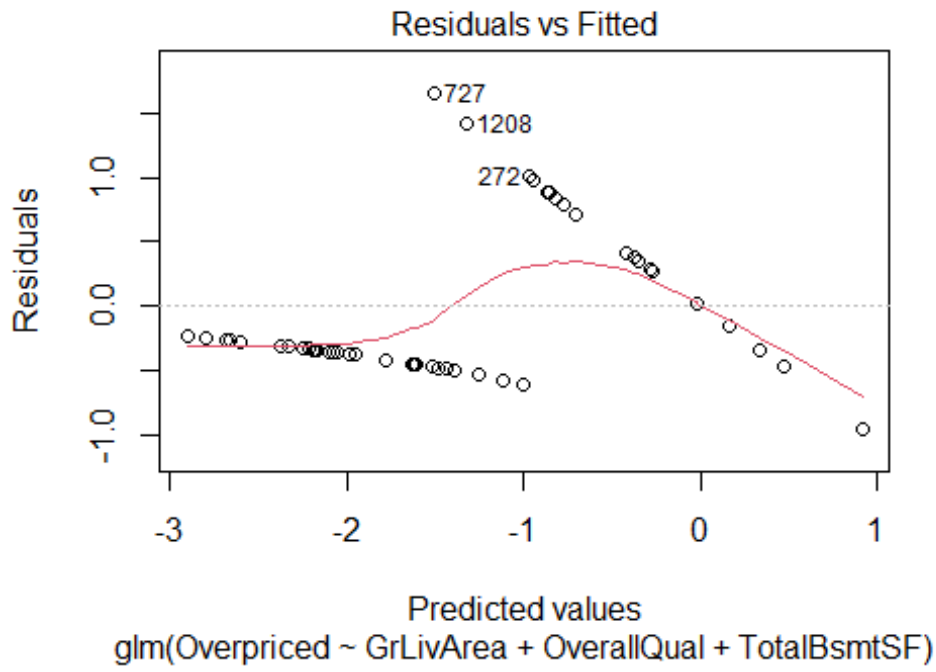
```
print(cov2cor(vcov(poiss)))
```

```
##             (Intercept)  GrLivArea OverallQual TotalBsmtSF
## (Intercept)  1.00000000 -0.1324435  -0.7847560  0.02889614
## GrLivArea   -0.13244347  1.0000000  -0.3374668 -0.33138132
## OverallQual -0.78475600 -0.3374668   1.0000000 -0.25450951
## TotalBsmtSF  0.02889614 -0.3313813  -0.2545095  1.00000000
```

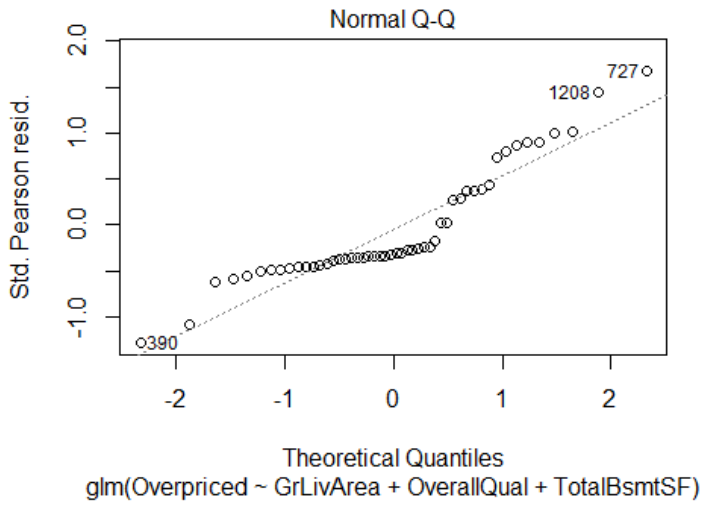
2.2.4 Residuals

Es pot veure en els plots que els residus són homoscedàstics. Però per confirmar s'ha fet el Breusch-Pagan test (dona 0.5835, que és major a 0.05) i es pot concloure que hi ha evidències estadístiques per dir que els residus són homocedàstics.

```
plot(poiss,1)
```




```
plot(poiss,2)
```



```
bptest(poiss)
```

```
##
##  studentized Breusch-Pagan test
##
## data:  poiss
## BP = 1.9471, df = 3, p-value = 0.5835
```

2.2.5 Prediccions

Per a les prediccions s'utilitzarà el mateix dataset de test utilitzat en l'apartat 2.1.5 i es farà un anàlisi de les mètriques obtingudes.

```
y_hat <- factor(round(predict(poiss, new_data =
test[,c('GrLivArea','OverallQual', 'TotalBsmtSF')], type =
'response')),levels = c(0:1))
```

Aquest model dona unes métriques molt més pobres que les de la regressió logística. Per començar, **l'accuracy** es del **45.8%**, un 15% menys que la logística. En aquest model la **sensitivity** dona un **52.6%** i la **specificity** un **20%**, valors molt més baixos que el model anterior.

Analitzant amb més atenció la confusion matrix es veu que el model ha previst 18 valors com a cars quan realment eren barats i 8 valors com a barats quan realment eren cars. La resta dels 50 valors de test els ha predit bé.

```
conf_mat = confusionMatrix(data = true_labels, reference = y_hat)
print(conf_mat)
```

```
## Confusion Matrix and Statistics
##
##           Reference
## Prediction  0  1
##          0 20  8
##          1 18  2
##
##                Accuracy : 0.4583
##                  95% CI : (0.3137, 0.6083)
##     No Information Rate : 0.7917
##     P-Value [Acc > NIR] : 1
##
##                   Kappa : -0.2
##
##  Mcnemar's Test P-Value : 0.07756
##
##             Sensitivity : 0.5263
##             Specificity : 0.2
```

```
##          Pos Pred Value : 0.7143
##          Neg Pred Value : 0.1
##              Prevalence : 0.7917
##          Detection Rate : 0.4167
##    Detection Prevalence : 0.5833
##       Balanced Accuracy : 0.3632
##
##        'Positive' Class : 0
##

predict(poiss, type = 'response')

##        388        543        836       1326        295       1311
1379
## 0.19761380 0.44087833 0.06984807 0.06104838 0.24778281 0.49555091
0.16824916
##        965        918         91        301        258       1003
561
## 0.38955894 0.07413893 0.05483659 0.12866627 0.42734791 0.70451401
0.13736289
##       1124        727       1048       1448        555       1135
1364
## 0.09729896 0.22245807 0.11249161 0.75198088 0.42280651 0.21930859
0.19489336
##        310        951        184        390        564         20
558
## 1.40195570 0.10566473 0.46260950 2.49223824 0.23633907 0.12498157
0.06846846
##       1269        497        704        875        721        272
1208
## 0.98304033 1.59272607 0.12300544 0.10917469 0.76584887 0.38095601
0.26680750
##        976       1159        158       1056        601       1198
945
## 0.32749779 0.70661751 0.65910920 0.22815248 0.68922277 0.36798490
0.19998031
##       1143        807        773       1152         35        697
1069
## 0.97388253 0.11193404 0.19665689 0.14250766 1.17129472 0.09341423
0.28469740
##       1011
## 0.11459639
```

Exercici 3: Disseny de blocs aleatoris i família binomial

3.1 Família binomial

Per aquesta part del treball s'ha creat una altre variable per utilitzar com a variable a predir en la familia de binomials. Aquesta variable és Newhouse i surt a partir de la columna YearRemodAdd del dataset. YearRemodAdd és l'any en el que es va fer l'última reforma a la casa i si aquest és superior a la mediana de la mateixa columna, es posarà un 1, és a dir que és una casa nova. Si és inferior a la mediana es posarà un 0, es a dir que considerem que la casa es vella.

Per a la família binomial s'utilitzaran dues variables de tipus categòriques (HouseStyle i RoofStyle) per a predir una variable binomial (NewHouse).

```
set.seed(36)
encode_ordinal <- function(x, order = unique(x)) {
  x <- as.numeric(factor(x, levels = order, exclude = NULL))
  x
}
house_cat <- house_prices_tot[floor(runif(50, min=1,
max=nrow(house_prices_tot))), c('RoofStyle', 'HouseStyle','YearRemodAdd')]
house_cat$RoofStyle <- encode_ordinal(house_cat$RoofStyle)
house_cat$HouseStyle <- encode_ordinal(house_cat$HouseStyle)
median(house_prices_tot$YearRemodAdd)
```

```
## [1] 1994
```

```
# Com que necesitem una variable binomial, separarem les cases entre les que
son prèvies i posteriors al 1994 (mediana).

house_cat$NewHouse = ifelse(house_cat$YearRemodAdd > 1994, 1, 0)
```

Primer es farà una anova per veure si les variables són estadísticament significatives.

```
anova = aov(NewHouse~HouseStyle+RoofStyle,data=house_cat)
summary(anova)
```

```
##              Df Sum Sq Mean Sq F value Pr(>F)
## HouseStyle    1   0.03  0.0305   0.120  0.731
## RoofStyle     1   0.36  0.3598   1.417  0.240
## Residuals    47  11.93  0.2538
```

Com que P-value > 0.05 en el cas de HouseStyle es pot concloure que no hi ha evidències estadístiques per dir que el HouseStyle afecti a la variable a predir (NewHouse). El mateix passa amb la variable RoofStyle, que el seu P-value és inferior a 0.05 i per tant no hi ha evidències estadístiques per dir que el RoofStyle afecti a la variable a predir.

Es fa el model binomial amb les dues variables HouseStyle i RoofStyle.

```
model_binomial <- glm(NewHouse~HouseStyle+RoofStyle,data=house_cat, family =
'binomial')
print(summary(model_binomial))
```

```
##
## Call:
## glm(formula = NewHouse ~ HouseStyle + RoofStyle, family = "binomial",
##     data = house_cat)
##
## Deviance Residuals:
##     Min       1Q   Median       3Q      Max
## -1.5308  -1.0319  -0.9479   1.2632   1.4258
##
## Coefficients:
##             Estimate Std. Error z value Pr(>|z|)
## (Intercept)  -1.5674     1.1316  -1.385    0.166
## HouseStyle    0.1226     0.2271   0.540    0.589
## RoofStyle     0.8776     0.7362   1.192    0.233
##
## (Dispersion parameter for binomial family taken to be 1)
##
##     Null deviance: 68.593  on 49  degrees of freedom
## Residual deviance: 67.008  on 47  degrees of freedom
## AIC: 73.008
##
## Number of Fisher Scoring iterations: 4
```

Com que les variables són categòriques per comprovar la seva significació estadística s'ha de fer un test amb una X^2. Al fer el test es veu que el P-valor de cap de les dues variables és inferior a 0.05. Per tant, no hi ha evidències estadístiques per dir que cap de les dues variables són estadísticament significatives.

```
print(anova(model_binomial, test = 'Chisq'))
```

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: NewHouse
##
## Terms added sequentially (first to last)
##
##
##            Df Deviance Resid. Df Resid. Dev Pr(>Chi)
```

```
## NULL                          49     68.593
## HouseStyle  1  0.12353        48     68.469   0.7252
## RoofStyle   1  1.46181        47     67.008   0.2266
```

3.2 Disseny de blocs complets aleatoris

La idea d'aquest apartat és fer un disseny de blocs complets aleatoris, el qual consisteix en controlar les possibles influències no dessitjades en un estudi experimental. Consisteix en agrupar de forma aleatòria els participants als diferents grups de tractament. Això ajuda a assegurar que qualsevol diferència observada entre els grups sigui deguda únicament al tractament en qüestió i no a altres variables no controlades.

A Rstudio el disseny de blocs es fa fent una ANOVA de les variables a evaluar però han d'estar amb el datatype factor. La variable HouseStyle té 8 nivells i la variable RoofStyle té 6 blocs.

```
estil_casa <- factor(house_cat$HouseStyle)
estil_sostre <- factor(house_cat$RoofStyle)
aov_disseny_blocs <- aov(NewHouse~estil_sostre+estil_casa, data = house_cat)

anova(aov_disseny_blocs)
```

```
## Analysis of Variance Table
##
## Response: NewHouse
##              Df  Sum Sq Mean Sq F value  Pr(>F)
## estil_sostre  1  0.3200 0.32000  1.3982 0.24338
## estil_casa    4  1.9298 0.48246  2.1080 0.09587 .
## Residuals    44 10.0702 0.22887
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Quan es rebutja la hipòtesi nul·la (H_0) en un disseny de blocs, significa que les dades recopilades suggereixen una diferència significativa entre els tractaments o entre els grups de blocs. Això indica que el factor de disseny utilitzat (els blocs) afecta d'alguna manera el resultat de la variable dependent. No obstant això, és important tenir en compte que el rebuig de H_0 no significa necessàriament que el factor de disseny estigui causant la diferència observada, només que hi està associat.

En aquest cas, com els p-value < 0.05, es pot dir que no hi ha diferència significativa.

```
predict(model_binomial, test = 'response')
```

```
##         909         986        1171         380        1109          30
## -0.56719380 -0.56719380 -0.44458317 -0.32197254 -0.32197254 -0.56719380
##        1394         638         131         526         415         891
## -0.19936191 -0.19936191 -0.32197254 -0.56719380 -0.32197254 -0.19936191
##         765          76         111         605         844         161
##  0.31043668 -0.44458317 -0.19936191 -0.56719380 -0.56719380 -0.56719380
##         746         247         453         446         482        1191
##  0.55565794 -0.32197254  0.55565794  0.31043668  0.31043668 -0.56719380
##         988         406        1348        1147         733         615
##  0.31043668 -0.56719380 -0.56719380 -0.56719380 -0.32197254 -0.07675128
##         807        1014        1248         291       446.1         771
## -0.44458317  0.31043668 -0.44458317 -0.32197254  0.31043668  0.80087920
##         913         358         456        1405        1367        1075
## -0.56719380 -0.56719380  0.31043668 -0.19936191 -0.32197254 -0.56719380
##         237         740        1350         922         175         432
## -0.56719380 -0.32197254 -0.32197254 -0.19936191 -0.56719380 -0.19936191
##         324         875
## -0.56719380 -0.19936191
```

Conclusions

En conclusió, en aquest treball s'ha analitzat l'ús de diverses tècniques estadístiques per analitzar relacions entre variables. En primer lloc, s'ha aprofundit en la regressió lineal múltiple, una eina àmpliament utilitzada a la recerca per modelar relacions entre variables quantitatives. S'ha discutit com la regressió lineal múltiple permet la incorporació de múltiples variables predictores al model, cosa que pot proporcionar una comprensió més completa de la relació entre les variables. Tot i això, també s'ha destacat la importància de comprovar els supòsits del model abans d'usar-lo, ja que un incompliment d'aquests pot afectar la validesa de les conclusions.

A més, s'ha abordat l'ús de la regressió logística, una tècnica estadística que es fa servir per modelar relacions entre variables categòriques i quantitatives. S'ha discutit com la regressió logística és una eina valuosa per predir la probabilitat d'un esdeveniment binari ocórrer, i com es pot fer servir per analitzar problemes en diferents camps.

En segon lloc, s'ha analitzat la regressió amb Poisson, una tècnica estadística que es fa servir per modelar relacions entre variables quantitatives i comptar esdeveniments. S'ha discutit com la regressió amb Poisson és una eina important per analitzar problemes relacionats amb freqüències d'esdeveniments, i com es pot fer servir per analitzar problemes en diferents camps.

Finalment, s'ha explorat l'ús de dissenys de blocs per controlar les variables que poden afectar el resultat. S'ha discutit com els dissenys de blocs són una eina important per controlar les variables aleatòries i millorar la precisió dels resultats. En general, aquest treball ha demostrat la importància de seleccionar la tècnica estadística adequada per al problema en qüestió i la necessitat de comprovar els supòsits del model abans de fer-lo servir. En general, el treball ha mostrat la importància de les tècniques estadístiques en l'anàlisi de dades i la seva capacitat per proporcionar informació valuosa i ajudar en la presa de decisions.

En aquest estudi, s’han arribat als següents resultats i conclusions:

Conclusions Regressió Lineal Múltiple

1. S’ha creat un model de regressió lineal múltiple que té com a variable dependent el preu de les cases. Aquests s’han predit a partir d’un conjunt de variables que en principi influenciaven més a la variable predita.

2. Les variables amb més influència per predir el preu de les cases són: OverallQual, GrLivArea, GarageArea i TotalBsmtSF.

3. El model compleix tots els supòsits: Linialitat, Multicolinialitat, Normalitat, Homoscedasticitat, Independència i Absència d’outliers.

4. En la validació global s'arriba a la conclusió de que el model és globalment estadísticament significatiu. Per altra banda, en la validació individual, s'han considerat totes les variables estadísticament significatives.

5. S'ha creat un model restringit amb les dues variables més significatives. Ha resultat ser pitjor, ja que la R^2 pren un valor bastant més baix. En altres paraules, el model restringit és capaç d'explicar el 12.50% la variabilitat observada en el preu de les cases, mentre que el model amb les quatre variables escollides ho fa amb un 93.28%.

6. Per tant, el millor model a seguir és el que utilitza 4 variables explicatives.

Conclusions Regressió Logística

1. S'ha creat un model de Regressió Logística que prediu si el preu d'una casa serà superior al preu mitjà de totes les cases o no, per tant, cara o barata.
2. Únicament té com a variable estadísticament significativa l'OverallQual. Per tant, en aquest cas, la variable que tindrà més influència en si una casa sobrepassa el preu mitjà o no és l'OverallQual.
3. Mitjançant McFadden R2, s'observa que es té un valor de 0.81. Per tant té un ajust prou bo del model i, per tant, una bona capacitat predictiva.
4. Amb la confussion matrix, sent TP cases es veu com hi ha 19 True Positive (barata), 11 True Negative (cara). S'han predit 10 cases cares com a barates i s'han predit 10 cases barates com a cares, el que comporta tenir una accuracy del 60%. Per tant, es pot dir que no prediu tot lo bé que es voldria.

Conclusions Poisson

1. S'ha creat un model de Poisson que prediu si el preu d'una casa serà superior al preu mitjà de totes les cases o no, a l'igual que en el model de Regressió Logística.

2. Únicament té com a variable estadísticament significativa l'OverallQual. Per tant, en aquest cas, la variable que tindrà més influència en si una casa sobrepassa el preu mitjà o no és l'OverallQual.

3. Amb la confussion matrix es veu com hi ha 20 True Positive (barata), 2 True Negative (cara). S'han predit 8 cases cares com a barates i s'han predit 18 cases barates com a cares, el que comporta tenir una accuracy del 45.83%. Per tant, es pot dir que prediu pitjor que el model de Regressió Logística.

Conclusions Model Binomial

1. S'ha creat un model Binomial que estudia si la casa es pot considerar nova o no, agafant com a threshold l'any central de la mediana dels valors de la variable YearRemodAdd. Com a variables independents categòriques es tenen: HouseStyle i RoofStyle.
2. Ni HouseStyle ni RoofStyle són estadísticament significatives pel model, és a dir, no són vàlides per explicar la variable dependent.

Conclusions Disseny de Blocs Complets Aleatoris

1. Com cap de les variables té un p-value inferior a 0.05, es pot dir que es rebutja H0. Per tant, no hi ha diferència significativa.

Bibliografia

Informació variada:
https://rpubs.com/Joaquin_AR/226291

Supòsits de model en R-studio:
https://youtu.be/jtaZ9J3iiDM

Anàlisi documentació paquets R:
https://www.rdocumentation.org/

Poisson:
https://bookdown.org/jaimeisaacp/bookglm/regresi%C3%B3n-de-poisson.html

Logística:
https://idaejin.github.io/courses/R/2019/euskaltel/regresion-logistica.html#ejemplo-predecir-el-salario-de-los-trabajadores

Vídeo summary, anova:
https://online.stat.psu.edu/stat485/lesson/12/12.2

Matriu variàncies i covariàncies:
How to Create a Covariance Matrix in R - Statology

Model Lineal:
https://www.institutomora.edu.mx/testU/SitePages/martinpaladino/modelos_lineales_con_R.html