Examen practica Ejercicio 1

Agustin Lopez Fredes

2023-07-15

Table of Contents

[Ejercicio 1 1](#_Toc140437994)

[1. Construir un modelo lineal simple para explicar el precio en funcion de la superficie y evaluar la bondad del ajuste. 1](#_Toc140437995)

[2. Realizar un analisis diagnóstico y de puntos influyentes e indicar si el modelo es adecuado. 3](#_Toc140437996)

[3. Realizar una transformacion de la variable respuesta para lograr normalidad en la distribución de los residuos. Indicar si el modelo con esta transformacion resulta adecuado. 6](#_Toc140437997)

[4. Eliminar la observacion 64 y ajustar nuevamente el segundo modelo evaluando su validez. 9](#_Toc140437998)

[5. Ajustar un modelo robusto y evaluar el promedio de los errores absolutos cometidos. Comparar con el mejor modelo lineal disponible. 12](#_Toc140437999)

[6. Utilizar un metodo de selección de variables para proponer un modelo multivariado. Analizar el cumplimiento de los supuestos. 13](#_Toc140438000)

[7. Le parece adecuado un modelo GAMLSS en este caso? Justifique. 17](#_Toc140438001)

[8. Resuma sus conclusiones. 17](#_Toc140438002)

## Ejercicio 1

En el archivo preciocasas.xlsx se han registrado respecto de 100 viviendas

### 1. Construir un modelo lineal simple para explicar el precio en funcion de la superficie y evaluar la bondad del ajuste.

library(readxl)  
library(ggplot2)  
library(dplyr)

##   
## Attaching package: 'dplyr'

## The following objects are masked from 'package:stats':  
##   
## filter, lag

## The following objects are masked from 'package:base':  
##   
## intersect, setdiff, setequal, union

datos\_casas <- read\_excel("preciocasas.xlsx")  
  
head(datos\_casas)

## # A tibble: 6 × 7  
## caso impuestos dormitorios banios estrena precio tamanio  
## <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>  
## 1 1 3104 4 2 0 279900 2048  
## 2 2 1173 2 1 0 146500 912  
## 3 3 3076 4 2 0 237700 1654  
## 4 4 1608 3 2 0 200000 2068  
## 5 5 1454 3 3 0 159900 1477  
## 6 6 2997 3 2 1 499900 3153

modelo\_precio\_superficie <- lm(precio ~ tamanio, data = datos\_casas)  
summary(modelo\_precio\_superficie)

##   
## Call:  
## lm(formula = precio ~ tamanio, data = datos\_casas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -236780 -29552 -2507 21639 151675   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -50926.255 14896.373 -3.419 0.000918 \*\*\*  
## tamanio 126.594 8.468 14.951 < 2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 56190 on 98 degrees of freedom  
## Multiple R-squared: 0.6952, Adjusted R-squared: 0.6921   
## F-statistic: 223.5 on 1 and 98 DF, p-value: < 2.2e-16

El test de Wald y de la regresión arrojan el mismo p-valor de 2,2e-16 al ser un modelo de una sola variable, lo que indica que la variable tamanio y el modelo son significativos.

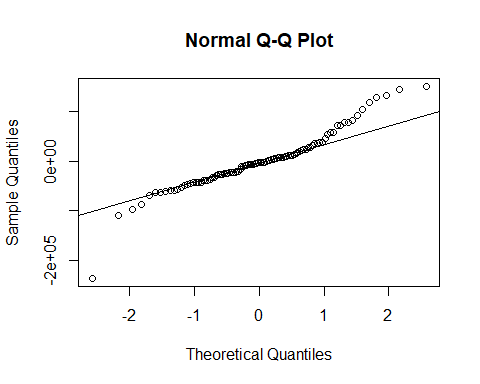
El coeficiente es positivo, lo que indica que con el aumento de tamaño aumenta el precio

El Coeficiente de determinación R2 ajustado indica que el modelo explica el 69,21% de la variabilidad en la variable respuesta

Debemos verificar entonces si se cumplen los supuestos del modelo

### 2. Realizar un analisis diagnóstico y de puntos influyentes e indicar si el modelo es adecuado.

#Análisis diagnostico  
#Normalidad  
#Mediante qqplot  
qqnorm(modelo\_precio\_superficie$residuals)  
qqline(modelo\_precio\_superficie$residuals)



#Mediante test de Shapiro Wilk  
shapiro.test(modelo\_precio\_superficie$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: modelo\_precio\_superficie$residuals  
## W = 0.93424, p-value = 8.729e-05

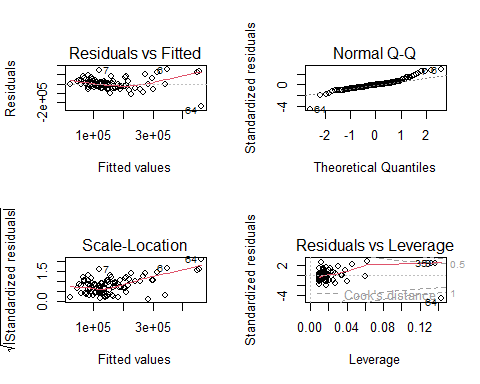
#Homocedasticidad  
lmtest::bptest(modelo\_precio\_superficie)

##   
## studentized Breusch-Pagan test  
##   
## data: modelo\_precio\_superficie  
## BP = 37.299, df = 1, p-value = 1.013e-09

#Autocorrelacion  
lmtest::dwtest(modelo\_precio\_superficie)

##   
## Durbin-Watson test  
##   
## data: modelo\_precio\_superficie  
## DW = 1.4215, p-value = 0.001719  
## alternative hypothesis: true autocorrelation is greater than 0

#Grafico de residuos vs valores ajustados y residuos vs leverage con comando plot(modelo)  
par(mfrow=c(2,2))  
plot(modelo\_precio\_superficie)



El test de Shapiro-Wilk rechaza normalidad de los residuos con un nivel de significación de 5%. Esto también se observa en el qqplot, ya que las observaciones se desvían de la recta El test de Breush-Pagan rechaza homocedasticidad de los residuos con un nivel de significación de 5%. Esto se comprueba en el gráfico de residuos vs fitted-values, donde se ve que hay mayor variabilidad, particularmente dada por la observación 64

El test de Dubin-Watson rechaza autocorrelación de los residuos igual a 0 con un nivel de significación de 5%, por lo que no son independientes

Observamos que ninguno de los supuestos se cumple

Analizamos puntos influyentes:

#Generamos el summary de las distintas métricas  
summary(influence.measures(model = modelo\_precio\_superficie))

## Potentially influential observations of  
## lm(formula = precio ~ tamanio, data = datos\_casas) :  
##   
## dfb.1\_ dfb.tamn dffit cov.r cook.d hat   
## 6 -0.52 0.69 0.75\_\* 0.92\_\* 0.26 0.06\_\*  
## 7 0.20 -0.11 0.29 0.90\_\* 0.04 0.01   
## 9 -0.82 1.00\_\* 1.04\_\* 1.03 0.51 0.14\_\*  
## 22 0.02 -0.02 -0.02 1.09\_\* 0.00 0.06   
## 35 -0.77 0.95 0.98\_\* 1.04 0.46 0.13\_\*  
## 64 1.65\_\* -2.01\_\* -2.08\_\* 0.74\_\* 1.73\_\* 0.14\_\*  
## 76 -0.31 0.43 0.49\_\* 0.97 0.11 0.05

#Por dffits  
n<-length(modelo\_precio\_superficie$precio)  
p<-length(modelo\_precio\_superficie$coefficients)  
which(dffits(modelo\_precio\_superficie)>2 \* sqrt(p / n))

## named integer(0)

which.max(dffits(modelo\_precio\_superficie))

## 9   
## 9

#Por dfbetas  
which(dfbetas(modelo\_precio\_superficie)[,2]>1)

## 9   
## 9

which.max(dfbetas(modelo\_precio\_superficie)[,2])

## 9   
## 9

#Por leverage  
leverage <- hatvalues(modelo\_precio\_superficie) > (2\*p/n)  
which(leverage)

## named integer(0)

which.max(hatvalues(modelo\_precio\_superficie))

## 64   
## 64

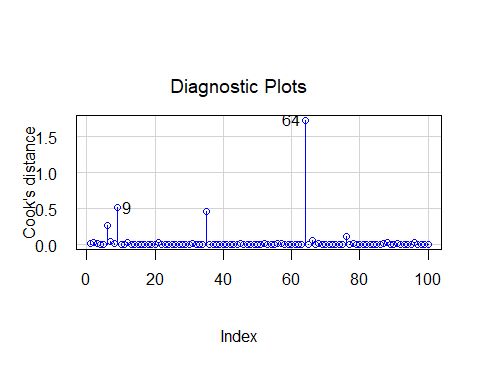
#Por distancias de cook  
dcook<-cooks.distance(modelo\_precio\_superficie)  
dcook\_corte <- dcook > 4/n #adoptamos el valor de corte de 4/n  
which(dcook\_corte)

## named integer(0)

which.max(dcook)

## 64   
## 64

car::influenceIndexPlot(modelo\_precio\_superficie, vars='Cook', las=1,col='blue')

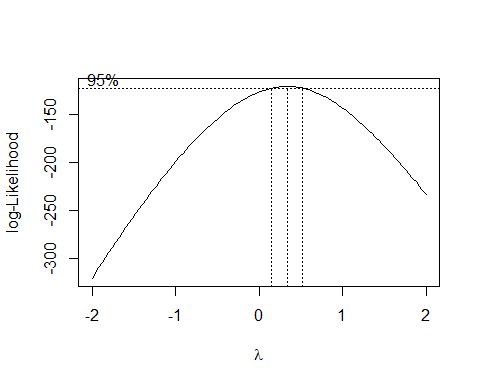


Si bien al adoptar los puntos de corte, sólo dfbetas arroja un punto influyente (la obs 9), podemos observar que los valores máximos de dffits y dfbetas son en la observación 9, mientras que por distancia de cook y leverange son máximos en la observación 64.

En el gráfico de influence plot de distancias de Cook se visualiza la observación 64 como punto influyente, seguido de la observación 9

### 3. Realizar una transformacion de la variable respuesta para lograr normalidad en la distribución de los residuos. Indicar si el modelo con esta transformacion resulta adecuado.

boxcox<- MASS::boxcox(precio ~ tamanio, data = datos\_casas,plotit=TRUE)



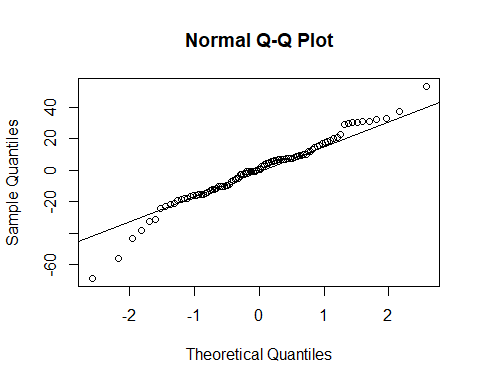
lambda\_bc <- boxcox$x[boxcox$y==max(boxcox$y)]#El lambda arroja 0.34  
lambda\_bc #El lambda es de 0.34

## [1] 0.3434343

modelo\_precio\_superficie.bc <- lm(((precio^lambda\_bc - 1) / lambda\_bc) ~ tamanio, data = datos\_casas)  
summary(modelo\_precio\_superficie.bc)

##   
## Call:  
## lm(formula = ((precio^lambda\_bc - 1)/lambda\_bc) ~ tamanio, data = datos\_casas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -68.844 -11.873 0.338 9.640 53.244   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.007e+02 5.152e+00 19.54 <2e-16 \*\*\*  
## tamanio 4.095e-02 2.928e-03 13.98 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 19.43 on 98 degrees of freedom  
## Multiple R-squared: 0.6661, Adjusted R-squared: 0.6627   
## F-statistic: 195.5 on 1 and 98 DF, p-value: < 2.2e-16

#Análisis diagnostico  
#Normalidad  
#Mediante qqplot  
qqnorm(modelo\_precio\_superficie.bc$residuals)  
qqline(modelo\_precio\_superficie.bc$residuals)



#Mediante test de Shapiro Wilk  
shapiro.test(modelo\_precio\_superficie.bc$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: modelo\_precio\_superficie.bc$residuals  
## W = 0.97273, p-value = 0.03592

#Homocedasticidad  
lmtest::bptest(modelo\_precio\_superficie.bc)

##   
## studentized Breusch-Pagan test  
##   
## data: modelo\_precio\_superficie.bc  
## BP = 3.7537, df = 1, p-value = 0.05269

#Autocorrelacion  
lmtest::dwtest(modelo\_precio\_superficie.bc)

##   
## Durbin-Watson test  
##   
## data: modelo\_precio\_superficie.bc  
## DW = 1.4434, p-value = 0.002441  
## alternative hypothesis: true autocorrelation is greater than 0

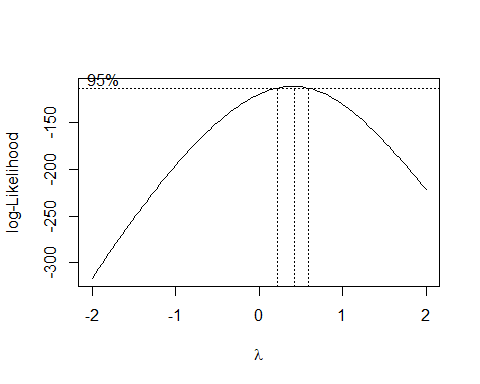
Vemos que el test de la regresión indica que es significativo, pero al realizar análisis diagnóstico observamos que con un nivel de significación de 5% rechaza normalidad y autocorrrelación igual a 0 de los residuos.

El test de Breusch-Pagan arroja un p-valor apenas superior al 5%, por lo que ahora no se rechaza homocedasticidad de los residuos.

Vemos que el modelo aún no cumple los supuestos

### 4. Eliminar la observacion 64 y ajustar nuevamente el segundo modelo evaluando su validez.

#Eliminamos observación 64  
datos\_casas\_sin\_influyente <- datos\_casas[c(-64),]  
  
  
#Realizamos nuevamente transformación de Box y Cox  
  
boxcox<- MASS::boxcox(precio ~ tamanio, data = datos\_casas\_sin\_influyente,plotit=TRUE)



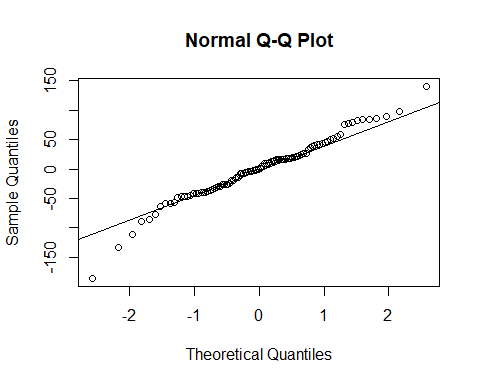
lambda\_bc\_sin\_inf <- boxcox$x[boxcox$y==max(boxcox$y)]#El lambda arroja 0.34  
lambda\_bc\_sin\_inf #El lambda es de 0.4242424

## [1] 0.4242424

modelo\_precio\_superficie.bc\_sin\_infl <- lm(((precio^lambda\_bc\_sin\_inf - 1) / lambda\_bc\_sin\_inf) ~ tamanio, data = datos\_casas)  
summary(modelo\_precio\_superficie.bc\_sin\_infl)

##   
## Call:  
## lm(formula = ((precio^lambda\_bc\_sin\_inf - 1)/lambda\_bc\_sin\_inf) ~   
## tamanio, data = datos\_casas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -185.885 -31.548 0.138 24.891 140.445   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 1.807e+02 1.342e+01 13.47 <2e-16 \*\*\*  
## tamanio 1.092e-01 7.627e-03 14.32 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 50.61 on 98 degrees of freedom  
## Multiple R-squared: 0.6766, Adjusted R-squared: 0.6733   
## F-statistic: 205.1 on 1 and 98 DF, p-value: < 2.2e-16

#Análisis diagnostico  
#Normalidad  
#Mediante qqplot  
qqnorm(modelo\_precio\_superficie.bc\_sin\_infl$residuals)  
qqline(modelo\_precio\_superficie.bc\_sin\_infl$residuals)



#Mediante test de Shapiro Wilk  
shapiro.test(modelo\_precio\_superficie.bc\_sin\_infl$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: modelo\_precio\_superficie.bc\_sin\_infl$residuals  
## W = 0.97461, p-value = 0.05033

#Homocedasticidad  
lmtest::bptest(modelo\_precio\_superficie.bc\_sin\_infl)

##   
## studentized Breusch-Pagan test  
##   
## data: modelo\_precio\_superficie.bc\_sin\_infl  
## BP = 6.7689, df = 1, p-value = 0.009276

#Autocorrelacion  
lmtest::dwtest(modelo\_precio\_superficie.bc\_sin\_infl)

##   
## Durbin-Watson test  
##   
## data: modelo\_precio\_superficie.bc\_sin\_infl  
## DW = 1.42, p-value = 0.001679  
## alternative hypothesis: true autocorrelation is greater than 0

Nuevamente vemos que el test de la regresión indica que el modelo es significativo, pero al realizar análisis diagnóstico observamos que, si bien ahora con un nivel de significación de 5% el test de Shapiro-Wilk no rechaza normalidad de los residuos, el test de Breusch-Pagan rechaza homocedasticidad de los residuos y el test de Durbin-Watson rechaza autocorrelación igual a 0 de los residuos, por lo que aún no cumplimos los supuestos

### 5. Ajustar un modelo robusto y evaluar el promedio de los errores absolutos cometidos. Comparar con el mejor modelo lineal disponible.

modelo\_precio\_superficie\_robusto\_huber<-MASS::rlm(precio ~ tamanio, data = datos\_casas,psi = MASS::psi.huber)#psi.huber por default  
summary(modelo\_precio\_superficie\_robusto\_huber)

##   
## Call: rlm(formula = precio ~ tamanio, data = datos\_casas, psi = MASS::psi.huber)  
## Residuals:  
## Min 1Q Median 3Q Max   
## -212142 -25942 1979 22267 169011   
##   
## Coefficients:  
## Value Std. Error t value   
## (Intercept) -42594.0849 11647.7426 -3.6569  
## tamanio 118.4534 6.6209 17.8908  
##   
## Residual standard error: 37660 on 98 degrees of freedom

#Calculamos el MAE del modelo robusto vs el modelo lineal del punto 4  
  
mae\_robusto\_huber <- mean(abs(datos\_casas$precio - predict(modelo\_precio\_superficie\_robusto\_huber, newdata = datos\_casas)))  
mae\_robusto\_huber

## [1] 38969.75

#Para el modelo de box y cox debemos antitransformar las predicciones  
pred\_bc\_sin\_infl <- (predict(modelo\_precio\_superficie.bc\_sin\_infl, newdata = datos\_casas) \* lambda\_bc\_sin\_inf + 1)^(1/lambda\_bc\_sin\_inf)  
mae\_bc\_sin\_influyente <- mean(abs(datos\_casas$precio - pred\_bc\_sin\_infl))  
mae\_bc\_sin\_influyente

## [1] 37002.1

mae\_modelo\_original <- mean(abs(datos\_casas$precio - predict(modelo\_precio\_superficie, newdata = datos\_casas)))  
mae\_modelo\_original

## [1] 39090.43

Vemos que el mean absolute error es mayor en el modelo robusto de Huber con respecto al modelo aplicando Box y Cox donde se eliminó la observación 64, pero es ligeramente menor al del modelo original con todas las observaciones y sin transformar

### 6. Utilizar un metodo de selección de variables para proponer un modelo multivariado. Analizar el cumplimiento de los supuestos.

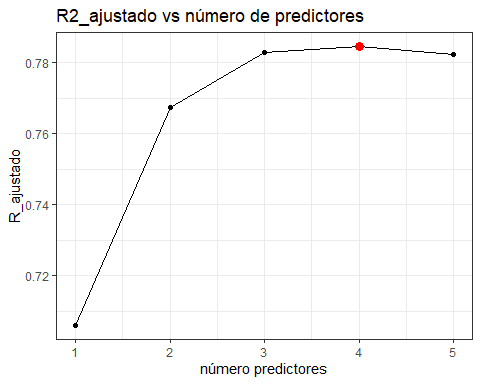
#Dejamos fuera la variable caso que sería un indice  
mejores\_modelos <- leaps::regsubsets(precio ~ . -caso, data = datos\_casas, nvmax = 5)   
summary(mejores\_modelos)

## Subset selection object  
## Call: regsubsets.formula(precio ~ . - caso, data = datos\_casas, nvmax = 5)  
## 5 Variables (and intercept)  
## Forced in Forced out  
## impuestos FALSE FALSE  
## dormitorios FALSE FALSE  
## banios FALSE FALSE  
## estrena FALSE FALSE  
## tamanio FALSE FALSE  
## 1 subsets of each size up to 5  
## Selection Algorithm: exhaustive  
## impuestos dormitorios banios estrena tamanio  
## 1 ( 1 ) "\*" " " " " " " " "   
## 2 ( 1 ) "\*" " " " " " " "\*"   
## 3 ( 1 ) "\*" " " " " "\*" "\*"   
## 4 ( 1 ) "\*" "\*" " " "\*" "\*"   
## 5 ( 1 ) "\*" "\*" "\*" "\*" "\*"

which.max(summary(mejores\_modelos)$adjr2)

## [1] 4

p <- ggplot(data = data.frame(n\_predictores = 1:5, R\_ajustado = summary(mejores\_modelos)$adjr2), aes(x = n\_predictores, y = R\_ajustado)) +   
 geom\_line() +   
 geom\_point()   
# Se identifica en rojo el máximo   
p <- p + geom\_point(aes(x=n\_predictores[which.max(summary(mejores\_modelos)$adjr2)], y=R\_ajustado[which.max(summary(mejores\_modelos)$adjr2)]), colour = "red", size = 3)   
p <- p + scale\_x\_continuous(breaks = c(0:5)) + theme\_bw() +   
 labs(title = "R2\_ajustado vs número de predictores", x = "número predictores")   
p



El mejor modelo por R2 ajustado es el de 4 predictores, usando las variables impuestos, dormitorios, estrena y tamaño Generamos este modelo y evaluamos los supuestos

mejor\_modelo <- lm(precio ~ impuestos + dormitorios + estrena + tamanio, data = datos\_casas)  
summary(mejor\_modelo)

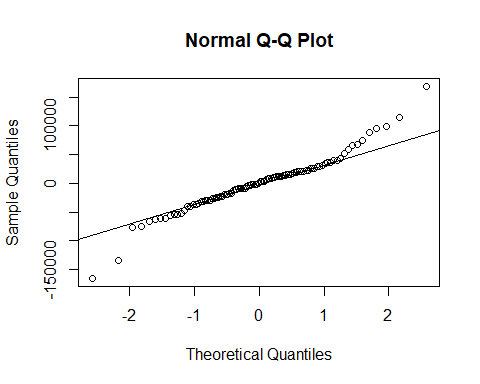
##   
## Call:  
## lm(formula = precio ~ impuestos + dormitorios + estrena + tamanio,   
## data = datos\_casas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -182119 -23980 -1690 21492 163181   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 2955.343 22827.885 0.129 0.897   
## impuestos 37.991 6.735 5.641 1.74e-07 \*\*\*  
## dormitorios -11607.319 8871.787 -1.308 0.194   
## estrena 41626.816 16794.918 2.479 0.015 \*   
## tamanio 67.587 13.240 5.105 1.70e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47000 on 95 degrees of freedom  
## Multiple R-squared: 0.7933, Adjusted R-squared: 0.7846   
## F-statistic: 91.15 on 4 and 95 DF, p-value: < 2.2e-16

Al entrenar el modelo observamos que la variable dormitorio no es significativa según el test de Wald, por lo que generamos el modelo sin la misma

mejor\_modelo <- lm(precio ~ impuestos + estrena + tamanio, data = datos\_casas)  
summary(mejor\_modelo)

##   
## Call:  
## lm(formula = precio ~ impuestos + estrena + tamanio, data = datos\_casas)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -165501 -25426 1449 20536 168747   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -21353.776 13311.487 -1.604 0.11196   
## impuestos 37.231 6.735 5.528 2.78e-07 \*\*\*  
## estrena 46373.703 16459.019 2.818 0.00588 \*\*   
## tamanio 61.704 12.499 4.937 3.35e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 47170 on 96 degrees of freedom  
## Multiple R-squared: 0.7896, Adjusted R-squared: 0.783   
## F-statistic: 120.1 on 3 and 96 DF, p-value: < 2.2e-16

#Análisis diagnostico  
#Normalidad  
#Mediante qqplot  
qqnorm(mejor\_modelo$residuals)  
qqline(mejor\_modelo$residuals)



#Mediante test de Shapiro Wilk  
shapiro.test(mejor\_modelo$residuals)

##   
## Shapiro-Wilk normality test  
##   
## data: mejor\_modelo$residuals  
## W = 0.95335, p-value = 0.001388

#Homocedasticidad  
lmtest::bptest(mejor\_modelo)

##   
## studentized Breusch-Pagan test  
##   
## data: mejor\_modelo  
## BP = 37.778, df = 3, p-value = 3.149e-08

#Autocorrelacion  
lmtest::dwtest(mejor\_modelo)

##   
## Durbin-Watson test  
##   
## data: mejor\_modelo  
## DW = 1.5108, p-value = 0.006163  
## alternative hypothesis: true autocorrelation is greater than 0

En el modelo el test de regresión muestra que el modelo es significativo, y el test de Wald indica que todas las variables lo son, pero nuevamente rechazamos todos los supuestos (normalidad, homocedasticidad y autocorrelación igual a 0 de los residuos)

### 7. Le parece adecuado un modelo GAMLSS en este caso? Justifique.

Dado que uno de los supuestos que se incumplió en forma frecuente es el de homocedasticidad de los residuos, considero que sería adecuado un modelo GAMLSS donde se modele tanto el valor esperado de la respuesta como su escala

### 8. Resuma sus conclusiones.

Observamos que las variables de mayor poder predictivo son impuestos, estrena y tamaño. No obstante, en prácticamente todos los modelos entrenados no se cumplieron los supuestos, por lo que lo recomendable es entrenar un modelo robusto con las variables mencionadas