Alumno: Sergio Castro Rodríguez

Estudio del precio de la vivienda

# Estudio de los factores que afectan al precio de la vivienda y obtención de un modelo predictivo

El presente estudio trata de analizar el efecto del tamaño de una casa sobre su precio de venta. Además se va a tratar de construir, a partir de las variables que tenemos a nuestra disposición, un modelo predictivo que permita a los comerciales conocer rápidamente cual sería el precio al que deberían vender una vivienda.

# Análisis previo de los datos disponibles

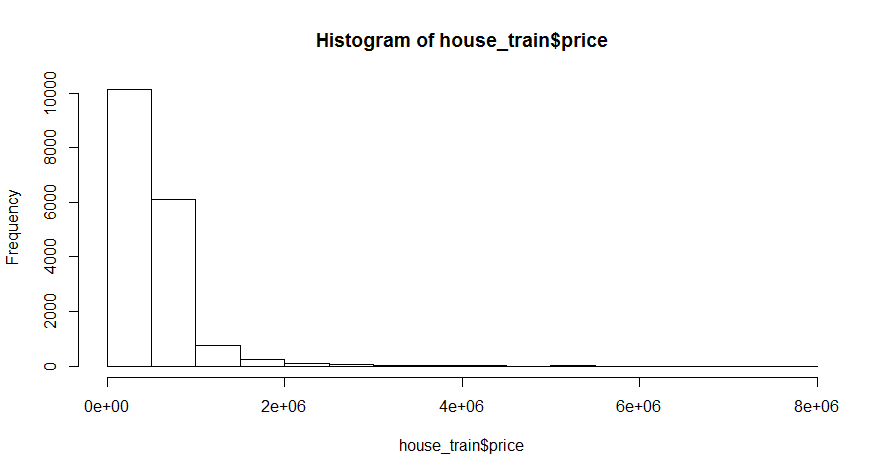
Para tener una idea de los datos que estamos manejando, vamos a hacer un pequeño análisis descriptivo del dataset (house\_train)

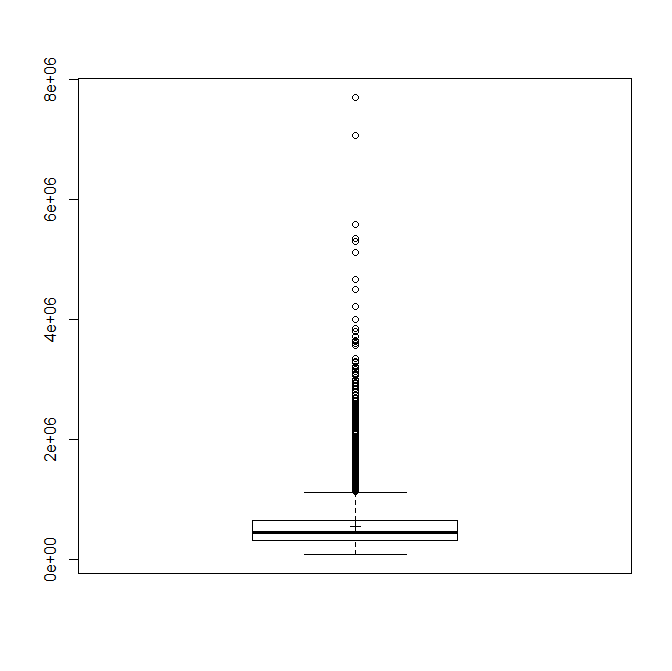
Variables:

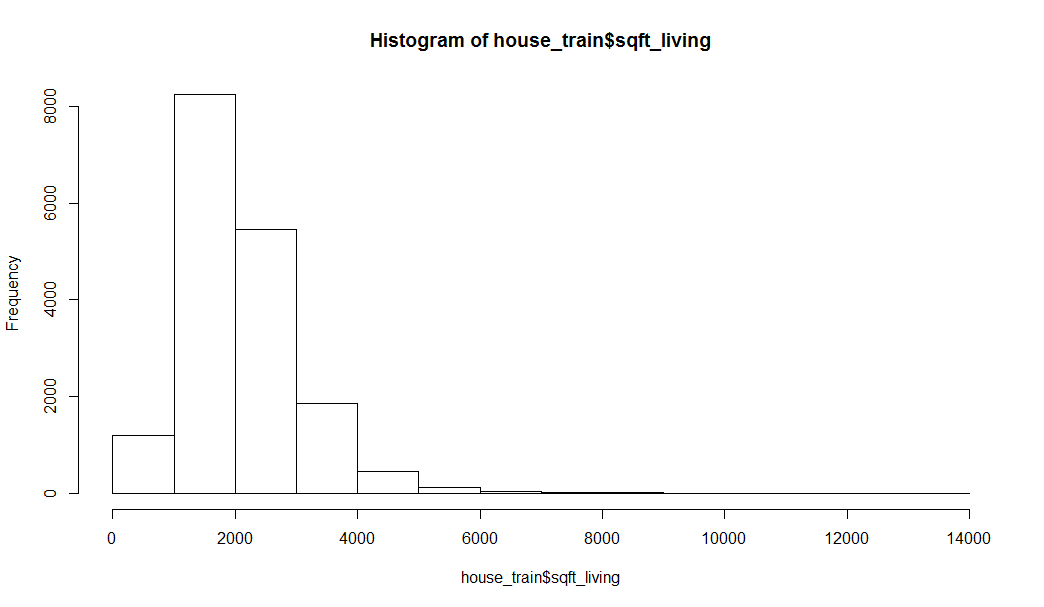
* Id: identificador de la vivienda. Se puede tratar como string puesto que no se van a realizar operaciones sobre él
* date: fecha asociada a la información. El formato en el que lo tenemos no es muy “amigable” por lo que lo vamos a convertir a un formato AAAmmDD para tratarlo con mayor facilidad
* price: precio de la vivienda. Variable numérica continua. Observamos que el valor mínimo es de 75.000 y el máximo de 7.700.000. Hay que destacar que la mediana (450.000) es inferior a la media (539.367) y que el valor del tercer cuartil (640.000) está muy lejos del valor máximo. Eso nos indica la presencia de outliers que estiran de la distribución hacia la derecha.
* bedrooms: número de habitaciones. Variable discreta. La vamos a tomar como factor (valores: 0,1,2,3,4,5,6,7,8,9,10). Hay que destacar que el primer cuartil y la mediana tienen el mismo valor,3. También que hay mucha diferencia entre el tercer cuartil (3) y el valor máximo (10)
* bathrooms: número de baños. Variable discreta. Observamos que hay valores decimales, cosa que no tiene sentido (parece ser que son aseos o baños no completos).La vamos a convertir a factor. Se observa que el primer cuartil ( 1.75) y el tercero (2.5) están muy próximos. La mediana (2.25) y la media (2.15) tienen un valor similar. Esto nos dice que el 50% de los valores, que se concentran entre el Q1 y Q3, son muy parecidos. También vemos que el máximo es 10, pero para que la media sea inferior a la mediana, significa que hay muchas viviendas con valores muy bajos (0).
* sqft\_living=superficie de la vivienda (en pies). Variable numérica continua. Lo que destaca es el valor del máximo, muy alejado del tercer cuartil.
* sqft\_lot: superficie de la parcela (en pies). Variable numérica continua. Lo que destaca es el valor del máximo, muy alejado del tercer cuartil.
* floors: número de plantas. Variable discreta. Observamos que hay valores decimales, cosa que no tiene sentido (parece ser que son buhardillas o sótanos, por ejemplo). La vamos a tomar como factor
* waterfront: indicador de estancia en primera línea al mar. Toma valores 0 y 1. También la vamos a tomar como factor
* view: número de orientaciones de la vivienda. Variable discreta. La vamos a tomar como factor (valores: 0,1,2,3,4). La inmensa mayoría tienen valor 0 (el tercer cuartil es cero) y algunos pocos toman valores superiores.
* condition: campo desconocido
* grade: campo desconocido
* sqft\_above: campo desconocido
* sqft\_basement: campo desconocido
* yr\_built: año de construcción. Variable discreta (desde 1900 hasta 2015)
* yr\_renovated: año de reforma. Variable discreta. Toma el valor 0 si la vivienda no ha sido reformada. En las casas reformadas toma valores entre 1934 y 2015. El tercer cuartil tiene valor 0, por lo que al menos un 75% de las casas no han sido renovadas.
* zipcode: código postal, la tomaremos como factor. No tiene ningún interés como número
* lat: latitud. Variable numérica continua
* long: longitud. Variable numérica continua
* sqft\_living15: campo desconocido
* sqft\_lot15:campo desconocido

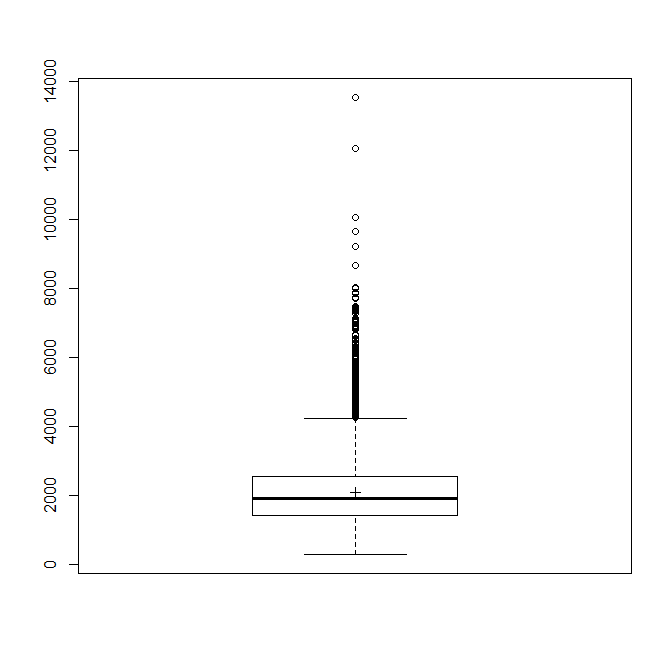
# Estudio del impacto del tamaño de la vivienda en su precio

Se procede a estudiar cómo afecta el tamaño de la vivienda a su precio. Veamos gráficamente como se distribuyen los valores de ambos por separado y la relación que pueden tener





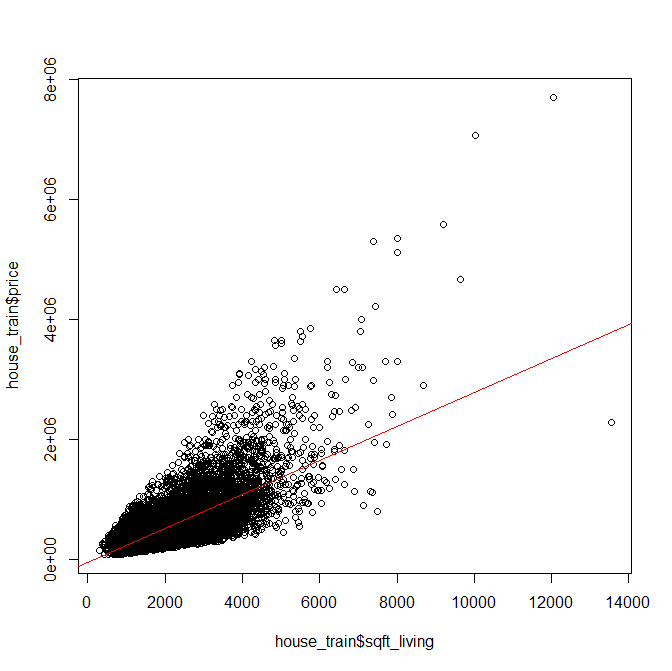




Se corrobora visualmente lo que se sospechaba a la vista del resumen de variables.

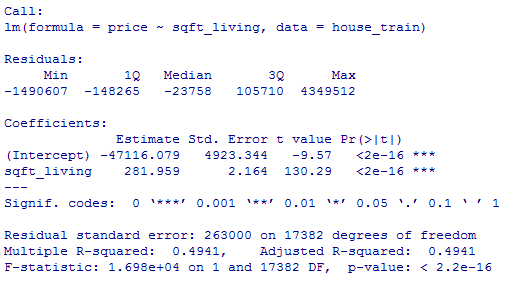
Tanto el precio como el tamaño de la casa como la superficie tienen una cola de valores altos que tiran de la distribución hacia la derecha.

El plot que enfrenta a las dos variables:

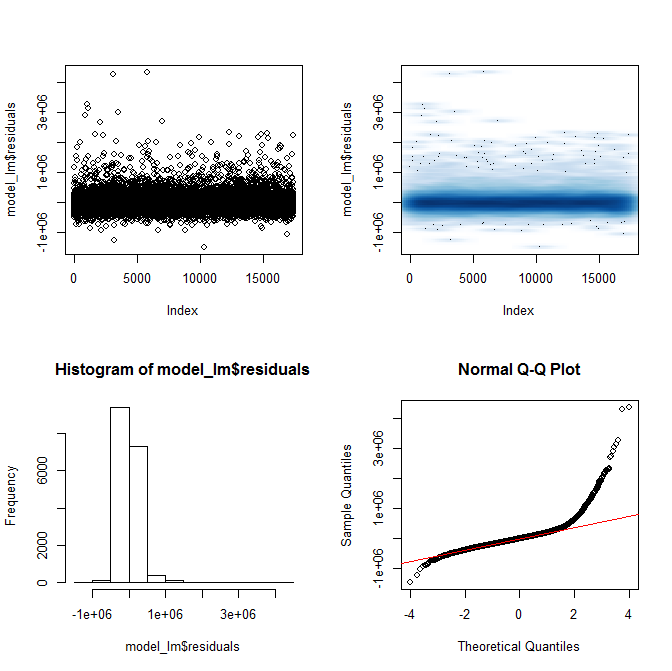


Se puede intuir una asociación lineal.

Primero se realiza el modelo más sencillo entre las dos variables, un modelo lineal:

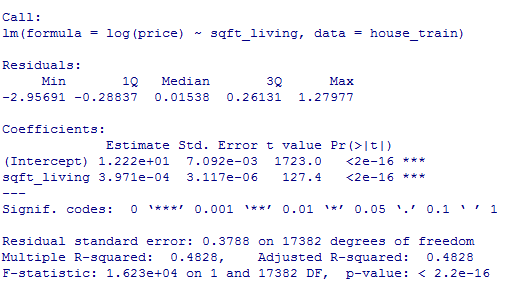


Se observa que la variable sqft\_living es significativa, pero para dar el modelo como bueno se deben estudiar los residuos. Ya en el resumen del modelo se observa que el error de los residuos es muy alto.

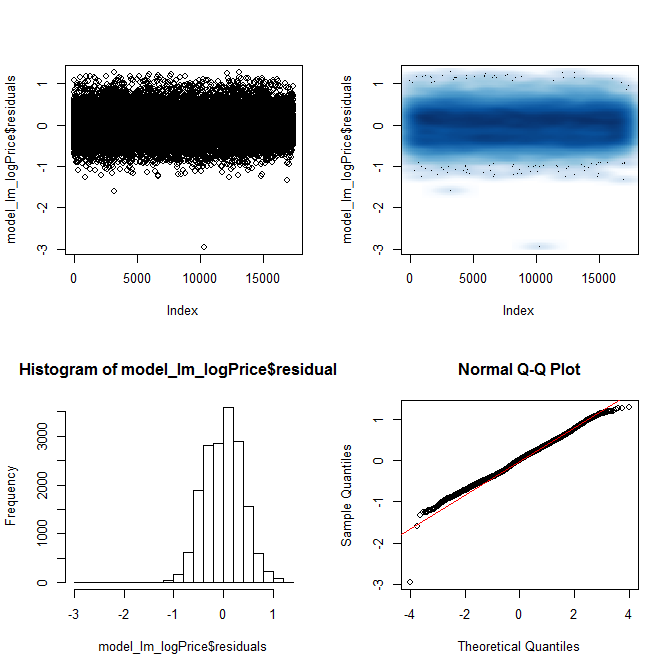


Al ver los residuos se observa que el modelo no se ajusta bien para los valores extremos. En el plot de residuos se aprecia que la distribución de los errores no es uniforme y que hay muchos valores que se encuentran por encima del valor cero. Sí se aprecia que la nube de puntos está en torno al cero pero hay valores por encima que demuestran que los errores no son aleatorios.

Para suavizar el efecto de los datos alejados del centro, se va a realizar una transformación logarítmica sobre el precio de la vivienda. El modelo quedaría de la siguiente forma:

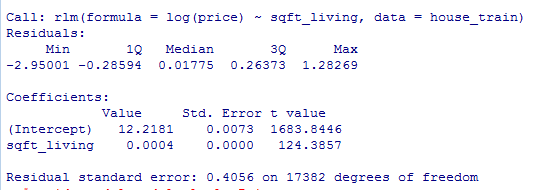


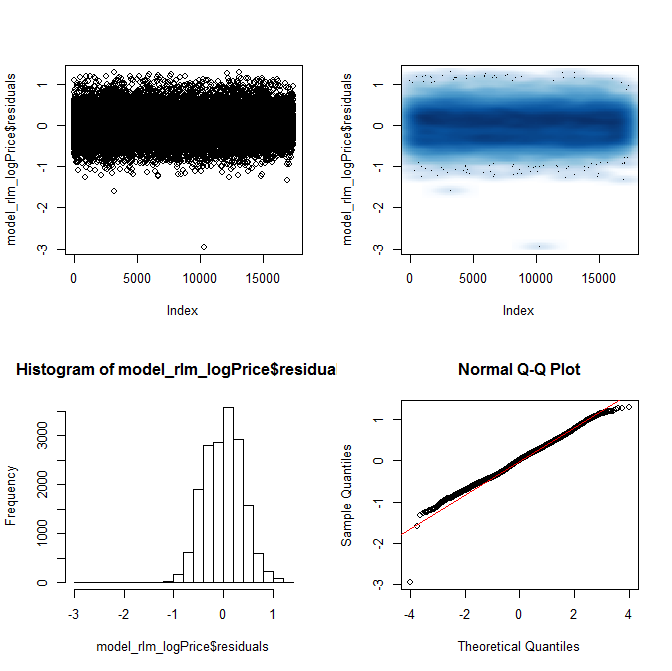
Se observa que el tamaño de la vivienda sigue siendo significativo y el error de los residuos es ahora muy bajo.



En el plot de residuos se observan dos valores que se alejan del resto. Sin tener en cuenta estos dos valores se puede decir, a la vista de los gráficos y de los valores obtenidos en el resumen del modelo, que el ajuste que se ha realizado únicamente con el tamaño de la vivienda es bastante bueno (el R2 es de sólo 0,48, pero no se pueden introducir más variables).

Otra opción a tener en cuenta para intentar evitar el efecto de los outliers es utilizar un modelo robusto. Este mismo modelo pero con técnicas robustas quedaría de la siguiente manera:





No se observa ninguna mejoría al utilizar el modelo lineal robusto, por lo que nos quedamos con el modelo lineal y transformación logarítmica del precio:

*Log(Price)=A + B\*sqft\_living*

El modelo quedaría de la siguiente forma:

Log(Price)= 1.222e+01 + 3.971e-04 \* *sqft\_living*

*El intervalo de confianza al 95% para los coeficientes:*

2.5 % 97.5 %

A=(1.220601e+01, 1.223382e+01)

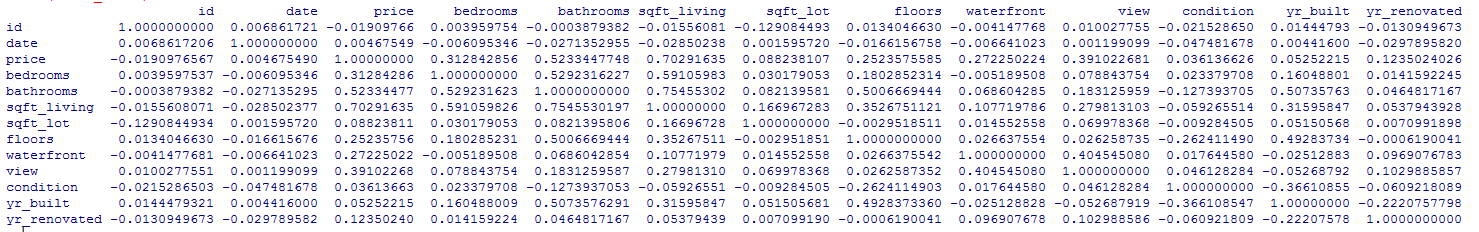
B= (3.909856e-04, 4.032067e-04)

Lo que quiere decir el modelo, en palabras más sencillas, es que por cada pie cuadrado que aumenta de tamaño la vivienda, el precio se incrementa entre un 0.03909856% y un 0.04032067%

Una vez estudiado el efecto del tamaño de la vivienda sobre su precio, el objetivo será encontrar un modelo predictivo que ayude a estimar el precio teniendo en cuenta las variables del dataset.

Para este caso se creará, a partir del dataset original, dos conjuntos de datos, train (80%) y test(20%) y en el que solo tendremos en cuenta las variables de las que conocemos su significado. Tampoco tendremos en cuenta el campo “date”

Vamos a hacer modelos con las variables con mayor correlación con el precio. Para esto, debemos obtener la matriz de correlaciones:



Vamos a generar un modelo con todas las variables que aportan más de 0.2, que son “sqft\_living”, “bathrooms”, “view”, “bedrooms”,”waterfront” y “floors”

Vamos a hacer 4 modelos diferentes que vamos a comparar.

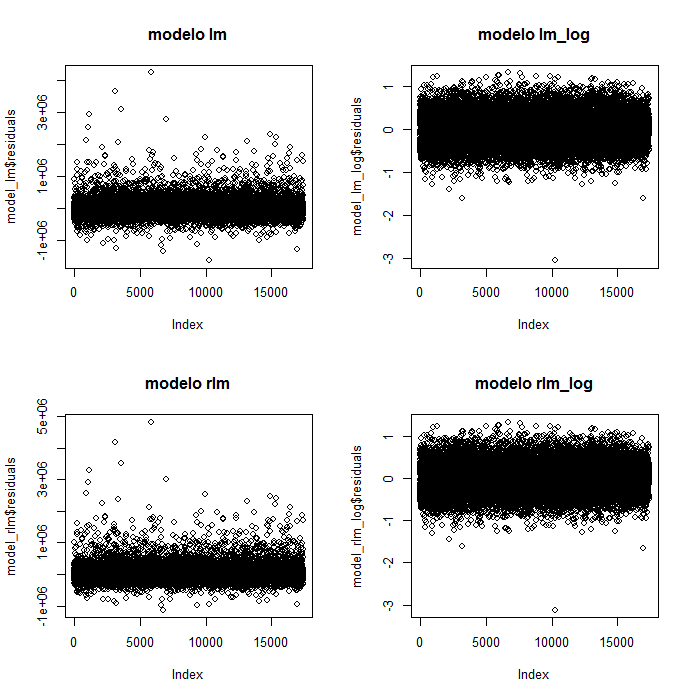
Como hemos visto en el punto anterior, el modelo con el logaritmo del precio se ajustaba mejor que el que utilizaba el precio sin transformar. También tenemos en cuenta que hay outliers en la muestra, por lo que también haremos modelos robustos y no robustos.

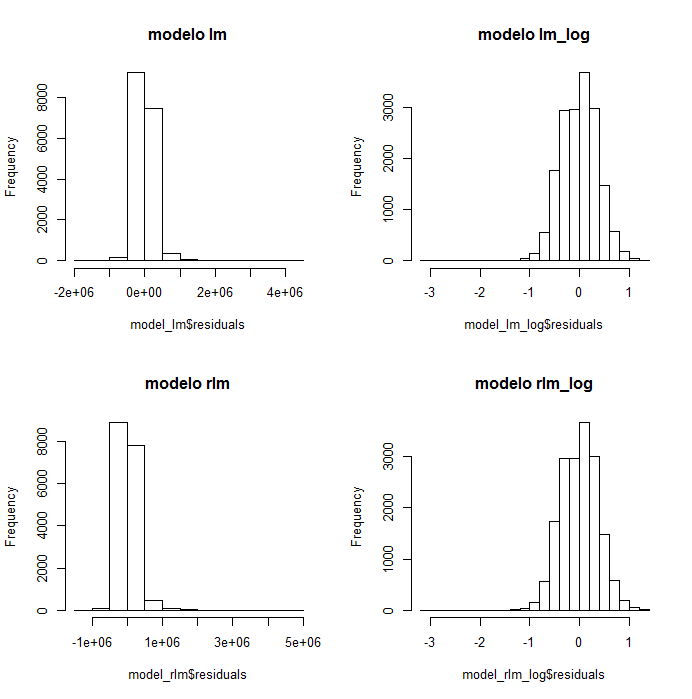
La comparativa de los 4 modelos respecto a su AIC y BIC es la siguiente:

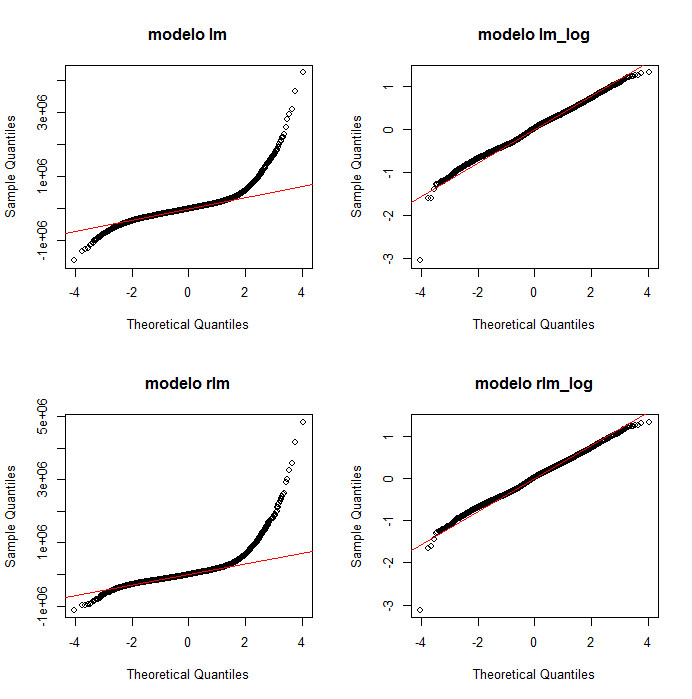


Vemos que los dos modelos que utilizan los logaritmos tienen un AIC y BIC muy inferior a los modelos sin transformación. Incluso se observa que el modelo no robusto tiene unas métricas algo inferiores.

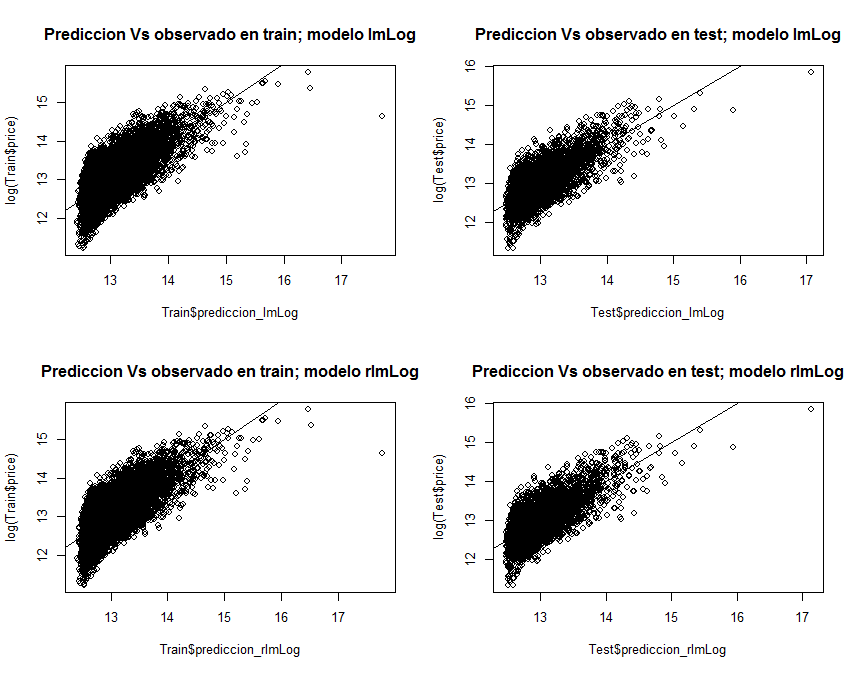
Veamos ahora cómo se comportan lo residuos de estos cuatro modelos:





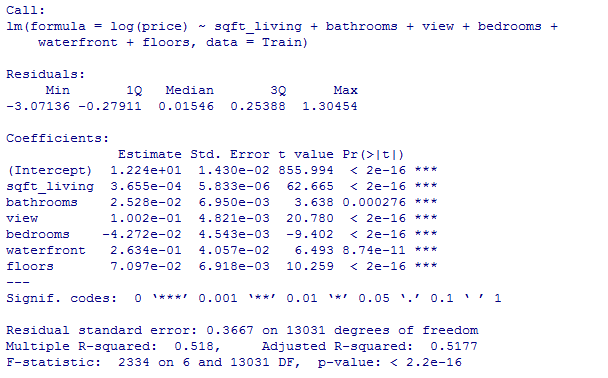


Veamos ahora cómo funciona la predicción el los datos de test:



En este caso seguimos teniendo resultados muy similares.

Optamos por quedarnos con el modelo lineal (no robusto) con transformación logarítmica del precio.



Ahora vamos a estimar con este modelo, el precio de venta de las viviendas que se encuentran en el fichero “house\_test.csv” y generar el fichero “house\_test\_price.csv”.

Hay que tener en cuenta que el modelo utiliza el logaritmo del precio, por lo que debemos deshacer esta transformación para obtener el precio estimado.