PROYECTO REGRESIÓN LINEAL

López López Jonathan
Mosqueda García Raúl
Zeferino Alvarado Rodrigo Emmanuel

1.Introducción

Descripción del problema

La venta de inmobiliaria es un mercado demasiado variante e interesante, se sabe o se estima que lo que implica el valor de un inmobiliario es en mayoría la zona y el tamaño.

Para este ejemplo se obtuvieron datos del país de Taiwán para poder explicar la relación que tienen las siguientes propiedades con el valor de un inmobiliario.

1. Edad de la casa (tiempo en años de vida del inmobiliario)
2. Distancia a la estación más cercana de TMR (El Metro de Singapur)
3. Número de tiendas de interés a la redonda
4. Coordenadas geográficas (Latitud y longitud)

Este análisis se llevará a cabo con modelo de regresión lineal múltiple teniendo como variables independientes las siguientes 5:

1. X_2 = Edad de la casa (tiempo en años de vida del inmobiliario)
2. X_3 = Distancia a la estación más cercana de TMR (El Metro de Singapur)
3. X_4 = Número de tiendas de interés a la redonda
4. X_5 = Coordenadas geográficas (Latitud)
5. X_6 = Coordenadas geográficas (longitud)

Y como variable dependiente solo una el valor de los inmobiliario

- Y = Valor de inmobiliario

1.1 Descripción de muestra

La muestra utilizada en este problema fue tomada de Departamento de Ingeniería Civil, Universidad de Tamkang, Taiwán de un estudio realizado en el 2018, con un tamaño de 414 observaciones de las propiedades antes mencionadas.

2. Marco teórico

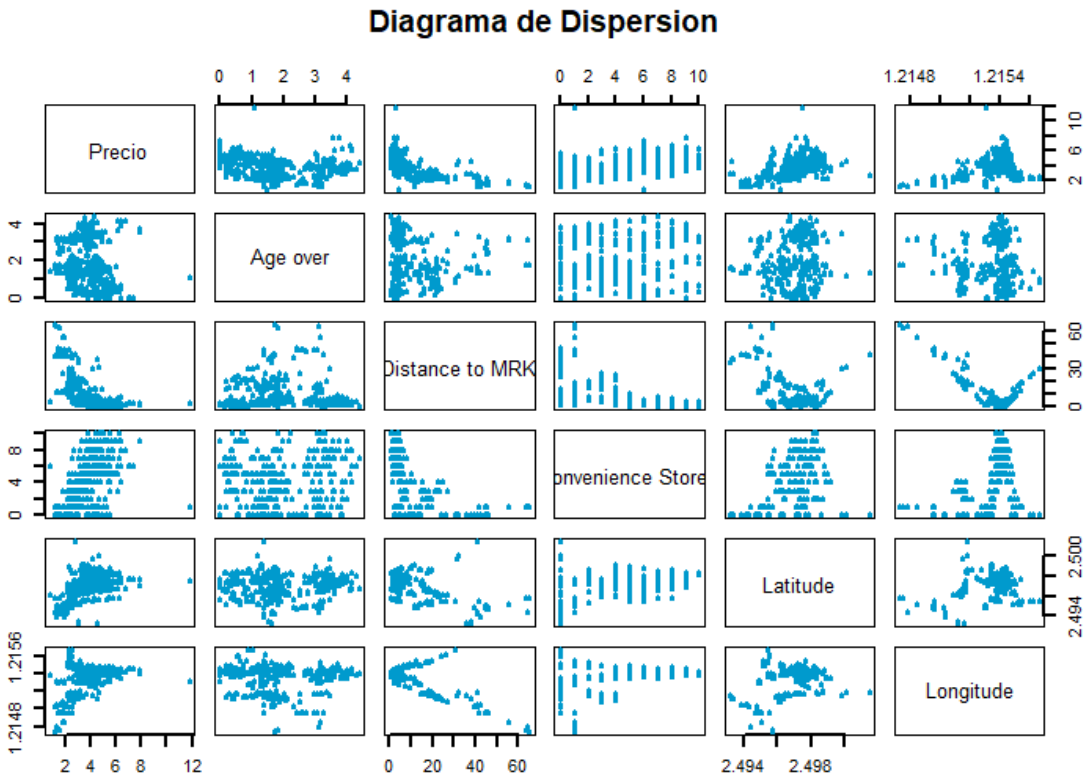
Nueva Taipéi es la ciudad más poblada en la República de China. La zona incluye un tramo importante de la costa norte de Taiwán, su área es de 2.052,57 y su población es de 3,907,000.

Nueva Taipéi se convirtió en un municipio independiente de Taipéi en 2010. La ciudad de Nuevo Taipéi se divide en 29 distritos, que a su vez se dividen en 1.017 villas y que estos a su vez se dividen en 21.683 barrios.

New Taipei Metro es un sistema de tránsito que sirve a Nueva Taipéi, Taiwán, operado por New Taipei Metro Corporation. Actualmente consiste solo en el tren ligero Danhai, con el tren ligero Ankeng y la línea Sanying en construcción, consta de 11 estaciones y una línea, pero el sistema de transporte que sigue siendo el prioritario por la población es el MRT que es un medio de transporte que cubre gran parte del área metropolitana de la ciudad de Taipéi. La red incluye 152,9 kilómetros de rieles con 131 estaciones con un promedio de más de 2,0 millones de viajes en un día entre semana, el sistema ha ayudado a reducir el tiempo de viaje desde un extremo a otro de la ciudad de tres a menos de una hora, y ha aliviado algunos de los problemas de tráfico de la ciudad.

A pesar de estar a un costado de la República de China, Taiwán ha tenido varios percances con dicho país los cuales han ido creciendo en los últimos años derivado a que Taiwán recibe un gran apoyo de E.U. que sigue en una guerra económica con la República de China.

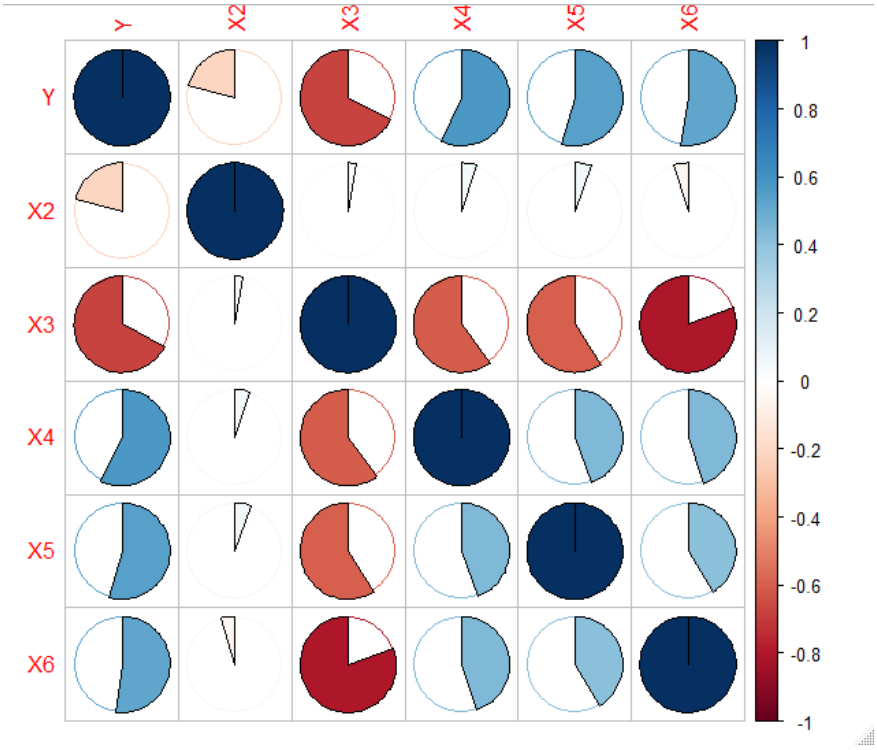
Análisis descriptivo de los datos



El diagrama de dispersión nos permite notar los siguientes puntos:

1. El precio tiene un valor demasiado extremo, el cual puede ser visto en todos los gráficos con las variables explicativas.
2. La distancia a alguna estación del tren (MRK) es inversamente proporcional a la Longitud lo quiere decir que mientras aumenta la longitud la distancia disminuye, esto puede deberse a que la estación está bastante cerca de un punto exacto de la longitud, que se puede ver también como conforme se pasa un punto de longitud, la distancia vuelve a aumentar. Esto puede ocasionar problemas ya que puede ser que estemos presentando un caso de multicolinealidad.
3. Fuera de la distancia a la estación de tren y la longitud, el resto de las variables parecen estar dispersas.

Posteriormente, veremos la tabla y las gráficas de las correlaciones para tener más pruebas que indiquen la existencia de multicolinealidad.



Correlaciones

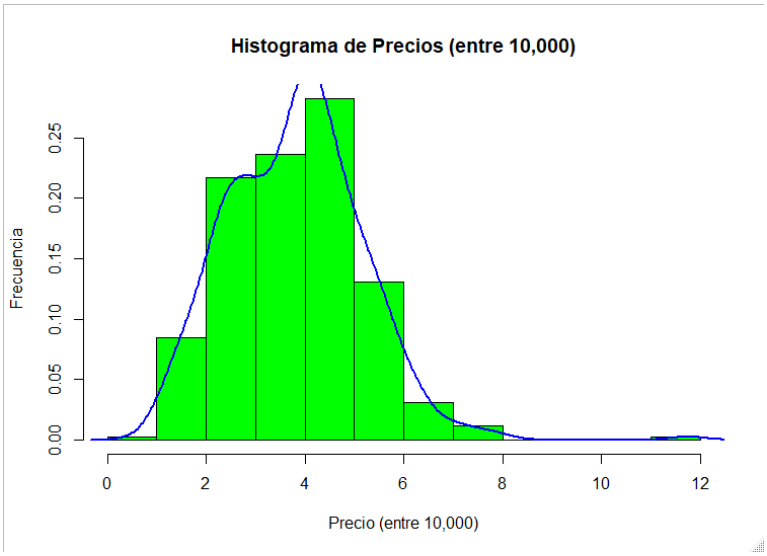
	Y	X2	X3	X4	X5	X6
Y	1	-0.2105670	-0.6736128	0.5710049	0.546306	0.5232865
X2	-0.210567	1	0.0256220	0.0495925	0.054419	-0.0485200
X3	-0.673612	0.0256220	1	-0.6025191	-0.591066	-0.8063167
X4	0.571004	0.0495925	-0.6025191	1	0.444143	0.4490990
X5	0.546306	0.0544199	-0.5910665	0.4441433	1	0.4129239
X6	0.523286	-0.0485200	-0.8063167	0.4490990	0.412923	1

Vemos que naturalmente la diagonal tiene una correlación de uno, pues es la correlación de cada variable consigo misma. El diagrama nos muestra banderas rojas con la variable X# que corresponde a la distancia a la estación de tren más cercana, siendo esta una correlación inversa con la latitud, la longitud, y la cantidad de tiendas de conveniencia cercanas.

La matriz de correlaciones nos permite comprobar que, en efecto, las correlaciones de X#3 son considerables teniendo los valores de:

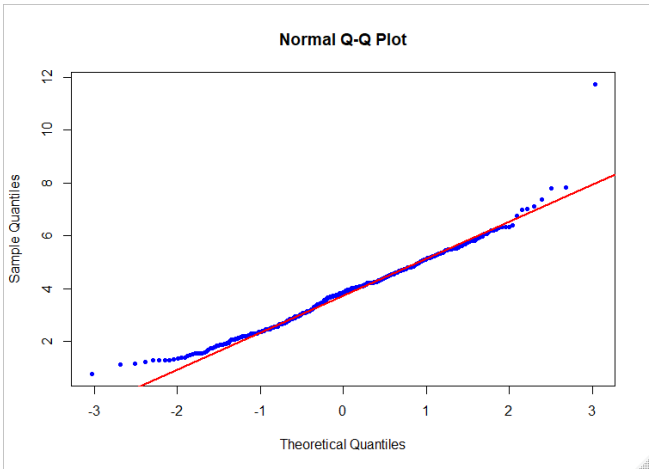
-0.60251914 -0.5910666 -0.80631677

Con las tiendas de conveniencia cercanas, la latitud y la longitud respectivamente. Esto nos indica que posteriormente tenderemos que hacer pruebas formales de correlación para ver la influencia que esta tenga en la modelo lineal.

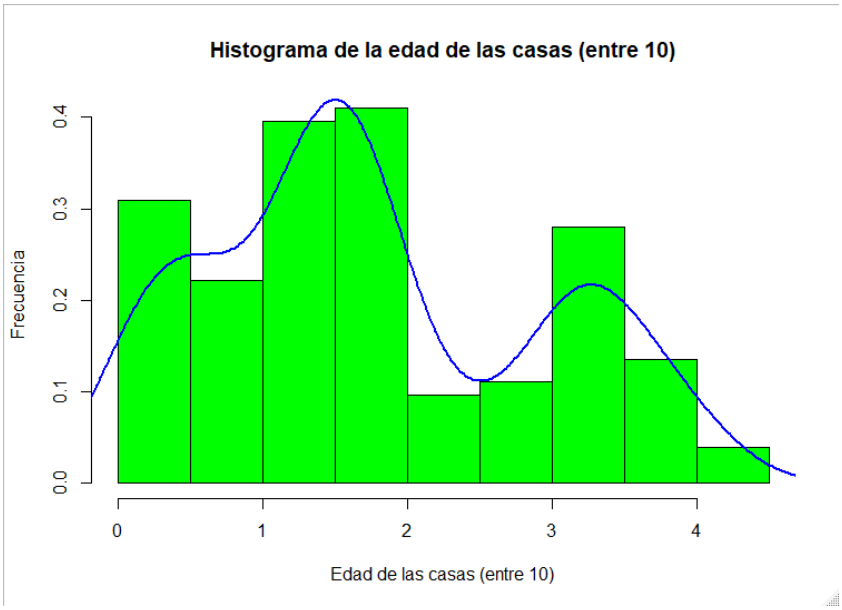


La media está en 3.845. En la gráfica podemos ver que sin embargo la moda está entre 4 y 5, esto se debe al valor extremo que se encuentra hasta 12. Podemos notar que los precios están en su mayoría entre 2 y 4, de ahí la frecuencia de precios disminuye drásticamente después de 5, por esto y el valor extremo está sesgada a la izquierda.

A continuación, veremos el resto de la dispersión de las variables explicativas y especularemos sobre la influencia que estas pueden tener en el precio de las casas.

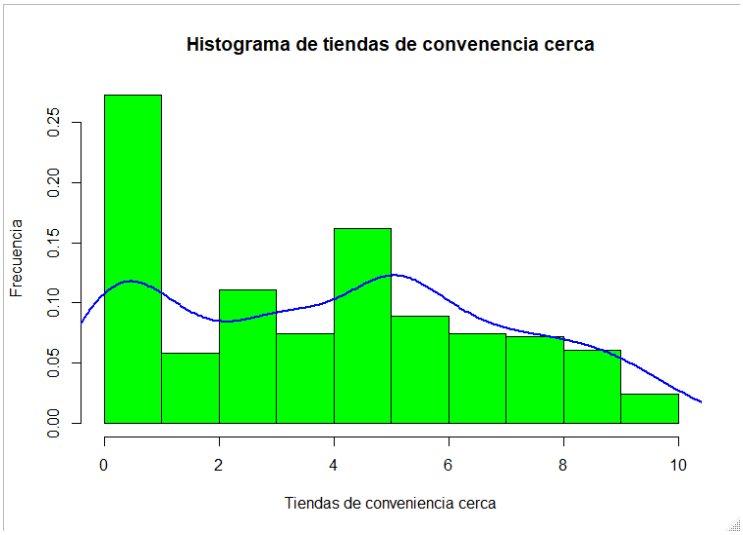


El precio parece ajustarse a la distribución normal, sin embargo, tiene una inclinación positiva y una cantidad considerable de valores significativos. Podemos ver aquí también el valor extremo que se encuentra al final de la muestra. Habrá que hacer analizar qué tan significativos son y ver si podemos omitirlos del análisis.



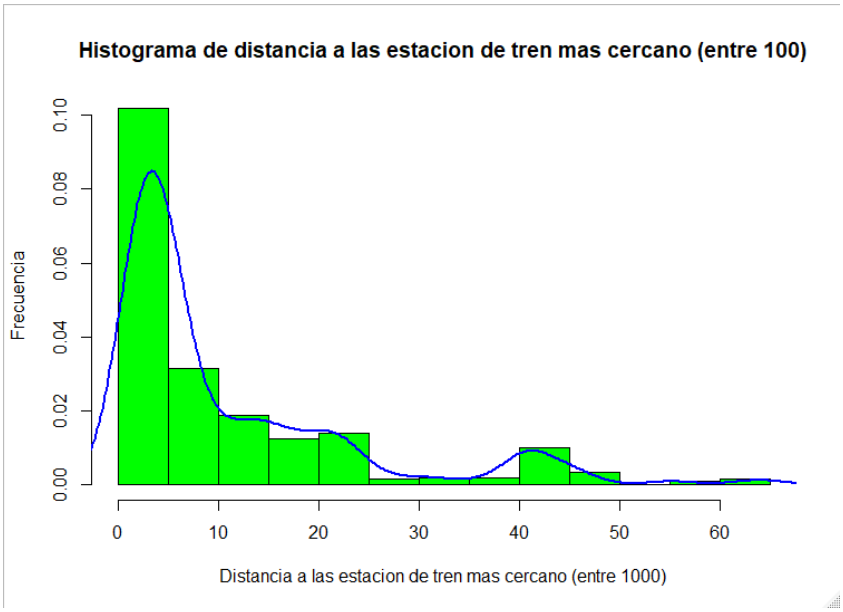
La media está en 1.61 sin embargo tiene una distribución bimodal, la segunda estando alrededor de 3.5. De esto podemos concluir que las casas son relativamente viejas, por lo que se puede pensar que, si no son nuevas, la zona donde se encuentran debe estar muy bien ubicada (esto se reafirma considerando que se encuentran cerca de la estación de tren, y una cantidad considerable de tiendas de conveniencia cercas). Podríamos pensar que entre mas nuevas, mas alto es el precio por la casa.

Existe que la posibilidad de que la edad de las casas influya de manera significante en su precio, sin embargo, en los diagramas de correlación vemos que la de esta con el precio de las casas es muy pequeña. Además, esta sesgada a la izquierda.



La media está en 4 con una moda de 0. Podemos notar de la gráfica que un cuarto de las casas no cuenta con una tienda de conveniencia cerca, sin embargo, el resto cuentan con una cantidad mayor a uno de tiendas cerca, lo cuál puede ser uno de los motivos que justifiquen los precios presentados.

Siendo estas cantidades considerablemente altas, nos inclina a considerar que en efecto estás tienen una influencia en el precio de las casas, lo cual también puede verse en los diagramas de correlación, siendo esta mayor a 0.



La media está en 4.922313. Notemos que el 10% de las casas están muy cerca de la estación, posteriormente la distancia aumenta en cantidades considerables, sin embargo, la porción de casas a la que aumenta es mucho menor que la distancia, por lo que es normal una distribución así dentro de un complejo de casas.

Recordemos que en los diagramas de correlación esta era la que presenta una correlación considerable con el resto de las variables, por lo que no hay que ignorar la posibilidad de que esta variable tenga que ser transformada para ver si es posible una mejor modelación.

Como se pudo ver, la base de datos tiene un buen comportamiento, no se tuvo la necesidad de limpiarlos o completarlos, y parecen indicar que se puede hacer una buena modelación con estos, pues están dispersos entre sí y no parece haber correlaciones entre las variables explicativas, salvo por una variable que se le dará su tratamiento adecuado para que eso cambie.

3.Análisis

3.1 Ajuste del modelo

Al construir el modelo (mediante el método de mínimos cuadrados) notamos que, por el cambio de una unidad en la edad de la casa, la distancia a la estación de tren mas cercana, el numero de tiendas de conveniencia cercanas, la latitud y la longitud, el precio cambia -0.2689168, 0.04259089, 0.116302, 237.7672, -78.05453 unidades para cada una de las covariables respectivamente.

Además, el precio de la casa se relaciona con el resto de las covariables con una diferencia de -494.5595. Esto se resume en el siguiente vector de estimadores:

β = [-494.5595, -0.2689168, -0.04259089, 0.116302, 237.7672, -78.05453].

3.2 Intervalos de Confianza

Con un nivel de confianza α = 5%, obtenemos los siguientes intervalos de confianza para los estimadores:

- β_0 está entre los valores: -1.715546e+03 y 726.42693375,
- β_1 está entre los valores: -3.455744e-01 y -0.19225930,
- β_2 está entre los valores: -5.681016e-02 y -0.02837162,
- β_3 está entre los valores: 7.890858e-02 y 0.15369551,
- β_4 está entre los valores: 1.494086e+02 y 326.12579676,
- β_5 está entre los valores: -1.044222e+03 y 888.11256680

Notamos que, en este caso todos los intervalos de confianza contienen al valor obtenido de los estimadores por el método de mínimos cuadrados. También veamos que tanto el primer como el ultimo intervalo contienen al 0, por lo que podemos decir que las estimaciones de β_0 y β_5 son poco significativos en comparación con las demás.

3.3 Prueba de Hipótesis general

En la prueba de hipótesis se realiza el siguiente contraste de hipótesis:

H0: $\beta=\beta^*$ vs H1 : $\beta \neq \beta^*$

donde β^* representa al vector de β's que obtuvimos anteriormente.

Al realizar la prueba obtenemos un estadístico de prueba igual a 1, el cual está fuera de la región de rechazo dada por ς= {Est. $\leq \lambda_0^{-2/n}$} siendo λ_0 el nivel de significancia 0.05 y n 414.

3.4 Análisis de la varianza a través de la tabla ANOVA

Los valores de la tabla ANOVA son los siguientes:

	Df	Sum Sq	Mean Sq	F value	Pr(>F)	
X2_Age.over10	1	33.90	33.90	42.1840	2.420e-10	***
X3_Distance_to__the_nearest_MRK.over100	1	341.64	341.64	425.0967	< 2.2e-16	***
X4_ConvenienceStores_nearby	1	38.17	38.17	47.4910	2.106e-11	***
x5_Latitude.over10	1	22.99	22.99	28.6106	1.478e-07	***
X6_Longitude.over100	1	0.02	0.02	0.0252	0.8739	
Residuals	408	327.90	0.80			

donde los aspectos a destacar son los siguientes:

- La suma de cuadrados de la regresión es de 436.7181, es la variabilidad entre la línea de regresión y los datos
- La suma de cuadrados del error es de 327.8957, es la variabilidad entre los residuales.
- La suma de cuadrados total es de 764.6138, es la variabilidad de las observaciones.
- El F-value para la mayoría de las variables explicativas es significante, a excepción de la Longitud el cual no pasa la prueba de hipótesis, por lo que es posible que esta variable tenga que ser transformada o removida posterior mente.

3.5 Medidas de bondad de ajuste del modelo

Para las medidas de bondad de ajuste tenemos los siguientes resultados:

- $R^2 = 0.5711617$
- R^2 ajustada = 0.5659063

donde notamos que solo aproximadamente el 57% de la variabilidad de los datos están representados por el modelo. Como estos valores son considerablemente chicos, muestra una decadencia en el modelo para representar los datos, por lo que posteriormente tendremos que hacer transformaciones a las variables para mejorar este resultado. En cuanto a la R^2 ajustada podemos observar que es muy cercana a la R^2 por lo que podríamos decir que no estamos sobre parametrizando el modelo y no necesitaremos quitar ninguna variable.

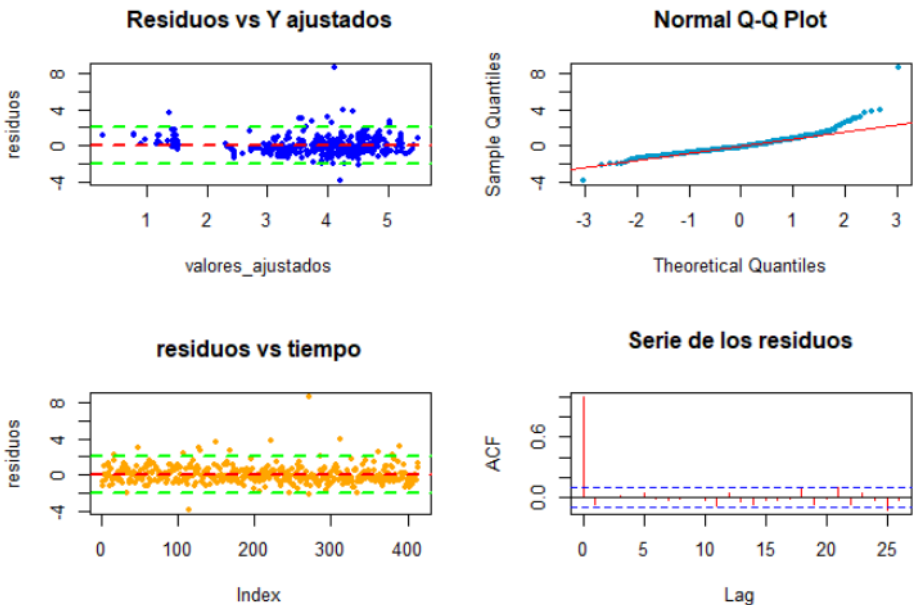
3.7 Multicolinealidad y Factores de Inflación de la Varianza

Para la multicolinealidad veremos los VIF para ver si existe entre algunas variables. Los VIF para cada parámetro son:

X2	X3	X4	X5	X6
1.014249	4.282985	1.613339	1.599017	2.923881

donde todos los factores son mayores que 1 por lo que hay multicolinealidad, sin embargo, debido a que ninguno de es mayor que 5, los coeficientes de regresión asociados están bien estimados.

3.8 Análisis de residuales



De las gráficas anteriores podemos notar en los Residuos contra los valores ajustados y el Q-Qplot, que existen una cantidad significante de valores extremos en los residuales, lo cual se pude ver de nuevo en la gráfica contra el tiempo, sin embargo, en la serie de los residuos no parece haberlos. Podemos suponer que los residuos no se comportan de manera normal, mayormente por el Q-Qplot donde vemos que tiene colas ligeras.

Dentro de las pruebas formales, se analizaron los siguientes conceptos:

- **Homocedasticidad** (Pruebas Breusch-Pagan y Prueba de corridas de Varianza no constante): Para ambas pruebas obtuvimos los P-values: 0.4805 y 0.20579 respectivamente, los cuales son mayores a nuestro nivel de significancia α = 0.05, por lo que no hay pruebas suficientes para rechazar la hipótesis nula, i.e. el modelo cumple con el supuesto de homocedasticidad o varianza constante.

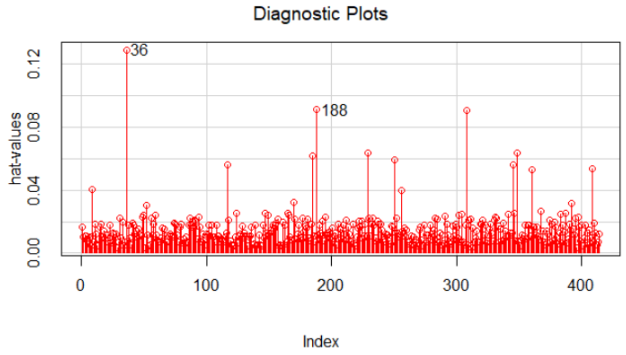
- **Correlación** (Pruebas Durwin-Watson y Breusch-Godfrey): Para ambas pruebas obtuvimos los P-values: 0.9376 y 0.1087 como ambos valores son mayores a α = 0.05, por lo que no hay pruebas suficientes para rechazar la hipótesis nula de que la correlación entre las variables es 0.

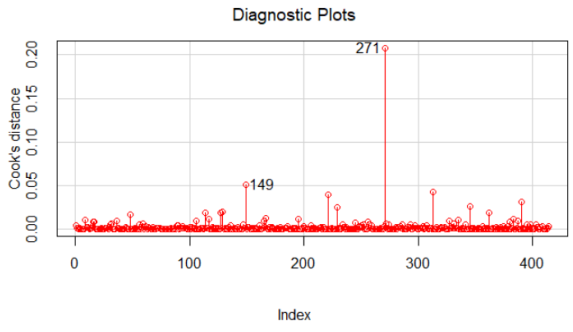
- **Normalidad** (Pruebas Anderson Darling para normalidad y Shapiro Wilk para normalidad): Para ambas pruebas obtuvimos el P-value 2.2e-16, el cual es menor a α = 0.05, por lo que hay pruebas suficientes para rechazar la hipótesis nula de que los residuos se distribuyen de forma normal. Como ya nos esperábamos al ver las gráficas no se cumplió el supuesto de normalidad por lo que será necesario manipular la base de datos mediante transformaciones para lograr que el nuevo modelo de regresión cumpla con todos los supuestos, ya que al no cumplir este supuesto las predicciones e intervalos de confianza podrían verse seriamente afectados.

A continuación, analizáremos los valores extremos o outliers para ver cuáles de estos son influyentes, los que no entonces podemos pensar que están debido a un error de obtención de los datos y considerar cuales pueden ajustarse o no al modelo.

```
> outlierTest(modelo)
     rstudent unadjusted p-value Bonferroni p
271  9.480495         2.1321e-19   8.8268e-17
313  4.062973         5.8150e-05   2.4074e-02
114 -3.935919         9.7459e-05   4.0348e-02
221  3.912786         1.0691e-04   4.4261e-02
```

Las muestras 271, 313, 114 y 221 son consideradas outliers por lo que procedemos a ver si estas observaciones son valores influyentes o no.





De las distancias de Cook vemos que los valores correspondientes a las posiciones 149 y 271 sus respectivas distancias de Cook no superan el 1 por lo que no son considerados valores influyentes, de igual forma los hat-values nos muestran que los correspondientes a las posiciones 36 y 188 tampoco lo son, por lo tanto, no los consideraremos valores influyentes. Por lo que procedemos a quitar los outliers de nuestra muestra.

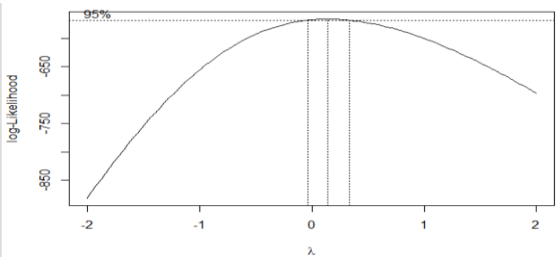
Sin estos datos ahora $R^2 = 0.6475$ y R^2 ajustada = 0.6431, notemos que se ha hizo una mejora considerable en cuanto a la explicación de la variabilidad de los datos, además que se hace disminuye la diferencia entre las medidas de bondad de ajuste.

Ahora al construir el modelo sin estos datos, los nuevos valores para β_0, β_1, β_2, β_3, β_4 y β_5 son: -737.6703, -0.2853, -0.0385, 0.1223, 229.3621 y 139.1949, respectivamente. Las diferencias entre las betas con los outliers y sin los outliers son: 473.5119 -0.006964441 -0.006754434 -0.01418547 -23.57558 -341.0407 respectivamente. Podemos observar que las mayores diferencias en los valores de las β's es en β_0, β_4 y β_5.

Al hacer las pruebas formales para normalidad tenemos P-values de 1.857e-09 y 1.584e-09, por lo que se sigue sin cumplir el supuesto de normalidad y debemos de hacer transformaciones en las variables.

4.Transformación

4.1 Transformación de Box-Cox y transformación elegida



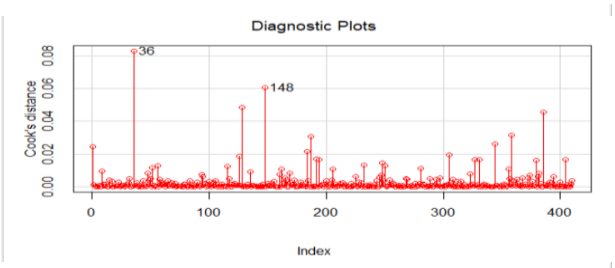
Vemos que la transformación de Box-Cox nos da $\lambda \approx 0$, por lo que realizamos la transformación Log(Y) y hacemos el modelo con esa transformación. Debido a que nuestro actual interes es mejorar el supuesto de normalidad de los residuos, procedemos a inmediatamente realizar las pruebas formales donde obtenemos los P-values 1.6e-05 y 5.045e-07, como aún no se cumple el supuesto, realizaremos más transformaciónes para que se cumplan y proceder con el seguimiento del modelo.

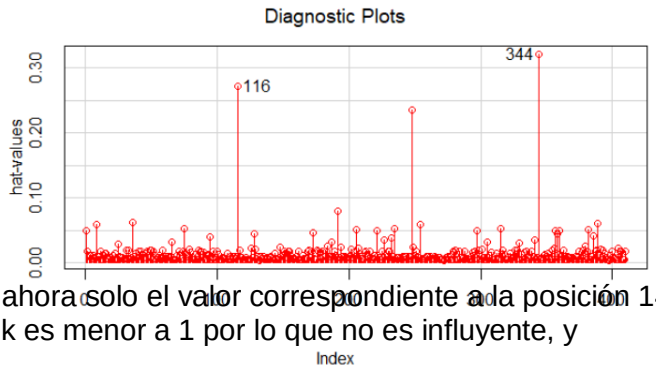
La solución la encontramos al elevar todas las variables de respuesta a la 5, obteniendo los siguientes P-values: 0.0703 y 0.1017 para las pruebas de normalidad.

4.2 Influencia de los datos

Habiendo hecho las transformaciones, realizaremos un análisis de residuales para mejorar la estimación del modelo.

```
> outlierTest(modelo_op)
    rstudent unadjusted p-value Bonferroni p
148 3.887231        0.00011854       0.0486
```





Con la función de outlierTest vemos que ahora solo el valor correspondiente a la posición 148 es uno, además vemos que su distancia de Cook es menor a 1 por lo que no es influyente, y podemos ignorarlo de la muestra.

4.3 Construcción del nuevo modelo

Con los cambios realizados, en el nuevo modelo obtenemos un nuevo modelo correspondiente al siguiente vector de Bethas:

β = -2.874e+02, -9.065e-05, -2.684e-10, 2.747e-06, 6.849e-01, 8.379e+01].

Hacer comentarios sobre la nueva betha

Notemos las significativas diferencias entre este vector de Bethas y el original:

```
> diff_bethas
  (Intercept)       X2_def       X3_def       X4_def       X5_def       X6_def
207.17046111   0.26882619   0.04259089   -0.11629930 -237.08231475 161.84681991
```

Podemos observar que la diferencia más grande es en X5 y X6 la cual la primera es negativa y la segunda positiva.

4.4 Intervalos de confianza

Con un nivel de confianza α = 5%, obtenemos los siguientes intervalos de confianza para los estimadores:

- β_0 está entre los valores: -3.342301e+02 y -2.405480e+02
- β_1 está entre los valores: -1.807393e-04 y -5.550933e-07,
- β_2 está entre los valores: -5.476663e-10 y 1.093205e-11,
- β_3 está entre los valores: 1.543389e-06 y 3.949923e-06,
- β_4 está entre los valores: 5.736126e-01 y 7.961399e-01,
- β_5 está entre los valores: 6.559255e+01 y 1.019920e+02

Notamos que, en este caso todos los intervalos de confianza contienen al valor obtenido de los estimadores por el método de mínimos cuadrados, también podemos ver que el intervalo de β_2 contiene al 0, por lo que podremos decir que las estimaciones de β_2 no son muy significativas.

4.5 Prueba de hipótesis general

Utilizando la misma prueba que la primera ocasión, no hay pruebas suficientes para rechazar H0 ($\beta=\beta^*$) pues el estadístico de prueba tiene un valor de 1.

4.6 Análisis de la varianza a través de la tabla ANOVA

Los valores de la tabla ANOVA son los siguientes:

```
> anova(modelo_def)
Analysis of Variance Table

Response: Y_def
           Df  Sum Sq Mean Sq  F value Pr(>F)
X2_def      1  0.1232  0.1232   2.0969 0.1484
X3_def      1 11.0473 11.0473 187.9977 <2e-16 ***
X4_def      1  6.1032  6.1032 103.8609 <2e-16 ***
X5_def      1 12.7225 12.7225 216.5054 <2e-16 ***
X6_def      1  4.8138  4.8138  81.9193 <2e-16 ***
Residuals 403 23.6814  0.0588
---
```

donde los aspectos a destacar son los siguientes:

- La suma de cuadrados de la regresión es de 34.81, es la variabilidad entre la línea de regresión y los datos
- La suma de cuadrados del error es de 23.6814, es la variabilidad entre los residuales.
- La suma de cuadrados total es de 58.4238, es la variabilidad de las observaciones.
- El F-value para la mayoría de las variables explicativas es significante, a excepción de la edad de la casa (X_2) el cual no pasa la prueba de hipótesis.

4.7 Medidas de bondad de ajuste del modelo

Para las medidas de bondad de ajuste tenemos los siguientes resultados:

- R^2 = 0.5951296
- R^2 ajustada = 0.5901064

Si comparamos las medidas de bondad notaremos que obtuvimos un pequeño incremento, esto quiere decir que de tener un 57% pasamos a un 59%, a pesar de esto el modelo sigue siendo poco predecible para los datos, aquí podremos decir que puede que exista una mejor transformación o un diferente modelo.

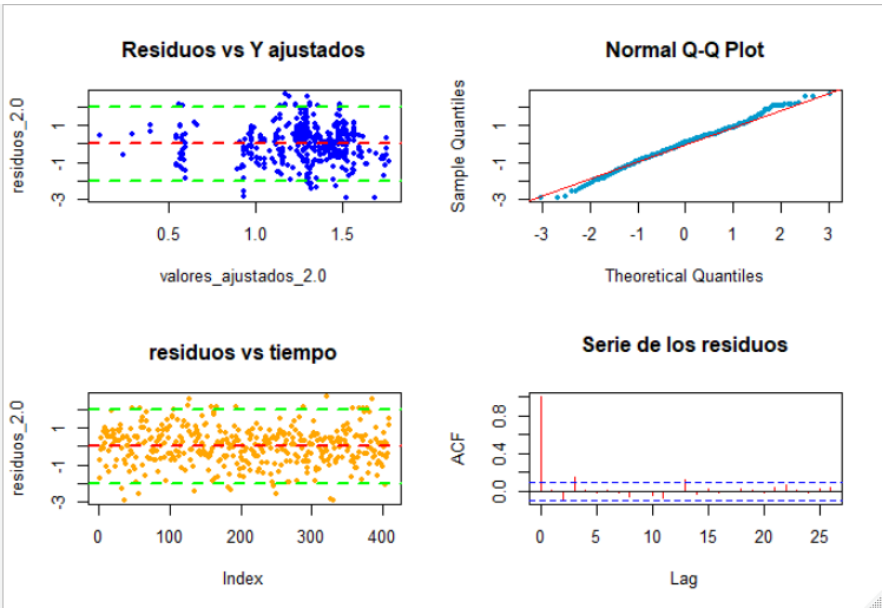
4.8 Multicolinealidad y factores de inflación de la varianza

Para la multicolinealidad veremos los VIF para ver si existe entre algunas variables. Los VIF para cada parámetro son:

X2	X3	X4	X5	X6
1.026383	1.508143	1.127179	1.278676	1.676333

donde todos los factores son mayores que 1 por lo que hay multicolinealidad, sin embargo, debido a que ninguno de es mayor que 5, los coeficientes de regresión asociados están bien estimados y si comparamos con el modelo original, observamos un cambio significativo en la variable x_3, lo cual nos dice que tenemos una mejor estimación en esta variable.

4.11 Análisis de residuales



Podemos observar en el Q-Qplot que seguimos teniendo el problema de las colas, pero un poco las ligeras, si vemos en la gráfica de residuos vs tiempo vemos valores extremos los que vemos igual al grafica residuos vs y ajustada y en este caso podemos observar ligeramente estos valores en la serie de los residuos en 6 y 13 aproximadamente,

Dentro de las pruebas formales, se analizaron los siguientes conceptos:

- **Homocedasticidad** (Pruebas Breusch-Pagan y Prueba de corridas de Varianza no constante): Para ambas pruebas obtuvimos los P-values: 0.000282 y 0.59717 respectivamente, de los cuales solo la prueba ncv tiene un P-value mayor a nuestro nivel de significancia $\alpha = 0.05$, sin embargo, de las gráficas podemos notar que los residuales no son precisamente aleatorios, por lo que concluímos que la muestra cumple con el supuesto de homocedasticidad.

- **Correlación** (Pruebas Durwin-Watson y Breusch-Godfrey): Para ambas pruebas obtuvimos los P-values: 0.3715 y 0.8203, como ambos valores son mayores a $\alpha = 0.05$, por lo que no hay pruebas suficientes para rechazar la hipótesis nula de que la correlación entre las variables es 0.

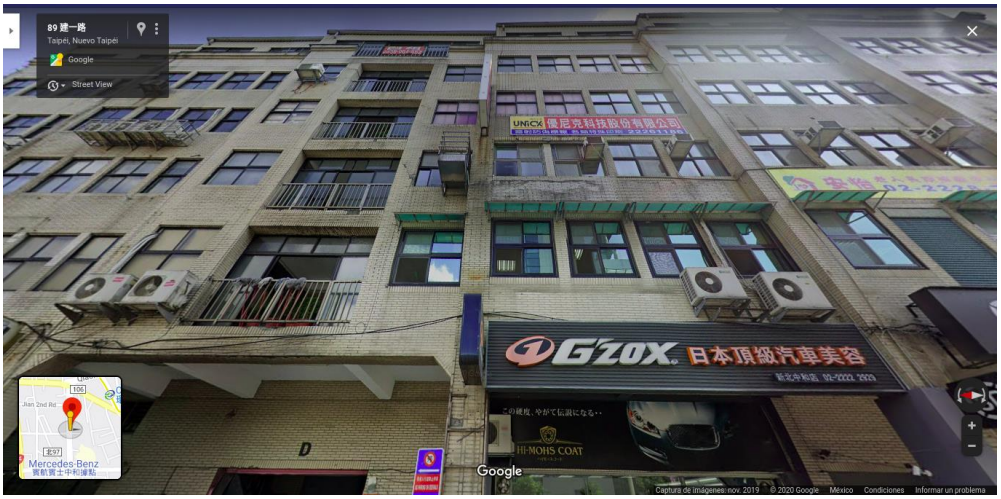
- **Normalidad** (Pruebas Anderson Darling para normalidad y Shapiro Wilk para normalidad): Para ambas pruebas obtuvimos los P-values: 0.09948 y 0.1327, como ambos valores son mayores a $\alpha = 0.05$, por lo que no hay pruebas suficientes para rechazar la hipótesis nula de que los residuales siguen una distribución normal estandarizada (lo cual es de esperarse, pues para esto se hicieron las transformaciones de las variables explicativas).

5. Ejemplo

A continuación, tomaremos como ejemplo una casa verdadera de la ciudad de Nueva Taipéi de la cual extraeremos los datos necesarios y estimaremos su precio utilizando el modelo estimado.

La casa (departamento) que utilizaremos se encuentra cerca de una zona medianamente comercial y de un parque, la dirección es: 91-83, Jianyi Road, Zhonghe District, New Taipei City, Taiwán 235, la cual cuenta con las siguientes propiedades:

- **Latitud:**25.04776
- **Longitud:**121.53185
- Tiendas de conveniencia :9 incluyendo Seven y gasolineras
- El metro más cercano es: Circular Line Qiaohe Station y está a 550 metros
- Aproximadamente una edad de 18 años



Aplicando en modelo antes estimado obtenemos las siguientes variables:

$$ln(Y) = \beta_0 + \sum_{i=1}^{5} \beta_i X_i$$

- X_1 = 18.895
- X_2= 5032.84375
- X_3 = 59049
- X_4 = 98.59263
- X_5 = 2.651243

Precio de la vivienda de la unidad de área: -2.874e+02 -9.065e-05X_1 -2.684e-10X_2 +2.747e-06X_3 +6.849e-01X_4 +8.379e+01X_5

⇨ Log(Y)=-2.874e+02 -9.065e-05 * 18.895 -2.684e-10 * 5032.84375 +2.747e-06 * 59049 +6.849e-01 *98.59263 +8.379e+01 * 2.651243 = 2.434237

Para terminar, elevamos e a este resultado para obtener Y

$$e^{2.434237} = 11.4069$$

por lo que el precio por unidad estimado de esta vivienda es $11.4069 (10000 Nuevo dólar de Taiwán / Ping, donde Ping es una unidad local, 1 Ping = 3,3 metros) con un asertividad del 59%.