Actividad 3 - Métodos y simulación estadística - Grupo A

Stefania Astudillo Bello

Introducción

Con base en los datos de ofertas de vivienda descargadas del portal Fincaraiz para apartamento de estrato 4 con área construida menor a 200 m^2 (vivienda4.RDS) la inmobiliaria A&C requiere el apoyo en la construcción de un modelo que lo oriente sobre los precios de inmuebles.

El presente análisis se enfoca en encontrar un modelo adecuado para todos los apartamentos que contiene el archivo vivienda4.RDS para esto se excluyen las viviendas de tipo casa el cual arroja 1363 apartamentos en difrentes zonas y precios.

para evaluar el modelo de regresión lineal simple se tienen en cuenta la correlación del área y el precio de los apartamentos.

A continuación se presenta el informe del punto 11 para los directivos de la inmobiliaria

Santiago de Cali, Abril 10 del 2023

Después de aplicar las formulas apropiadas y analizar la información de los precios de los inmuebles, damos respuesta al servicio contratado por el cliente inmobiliaria A&C:

según los datos del archivo enviado vivienda4.RDS, se concluye que para un mismo estrato, el área; a pesar de que esta variable explica un 58% de la variación del precio de las viviendas, no es información suficiente para para lograr encontrar un modelo adecuado que proporcione información detallada de los precios de las viviendas, ya que pueden existir otras consideraciones que no se encuentran en el archivo enviado como por ejemplo No. de habitaciones, años de construcción, ubicación, inseguridad, el cual al ser valoradas se podría ajustar un modelo que podamos orientarlo para un análisis detallado en cuanto al precio de todos los apartamentos que ofrecen para la venta.

cabe resaltar que esta conclusión se da solamente para los apartamentos que se encuentran en el archivo vivienda4.RDS

Como empresa decidimos tener un clasificación aparte de los de tipo casa y por este motivo se analizó solamente los apartamentos, si se requiere realizar un análisis para las viviendas de tipo casa se sugiere realizarlo aparte.

Cordialmente,

Stefania Astudillo

Cientifica de datos

Problema

Con este propósito el equipo de asesores a diseñado los siguientes pasos para obtener un modelo y así poder a futuro determinar los precios de los inmuebles a negociar.

1. Realice un análisis exploratorio de las variables precio de vivienda (millones de pesos COP) y área de la vivienda (metros cuadrados) - incluir gráficos e indicadores apropiados interpretados.

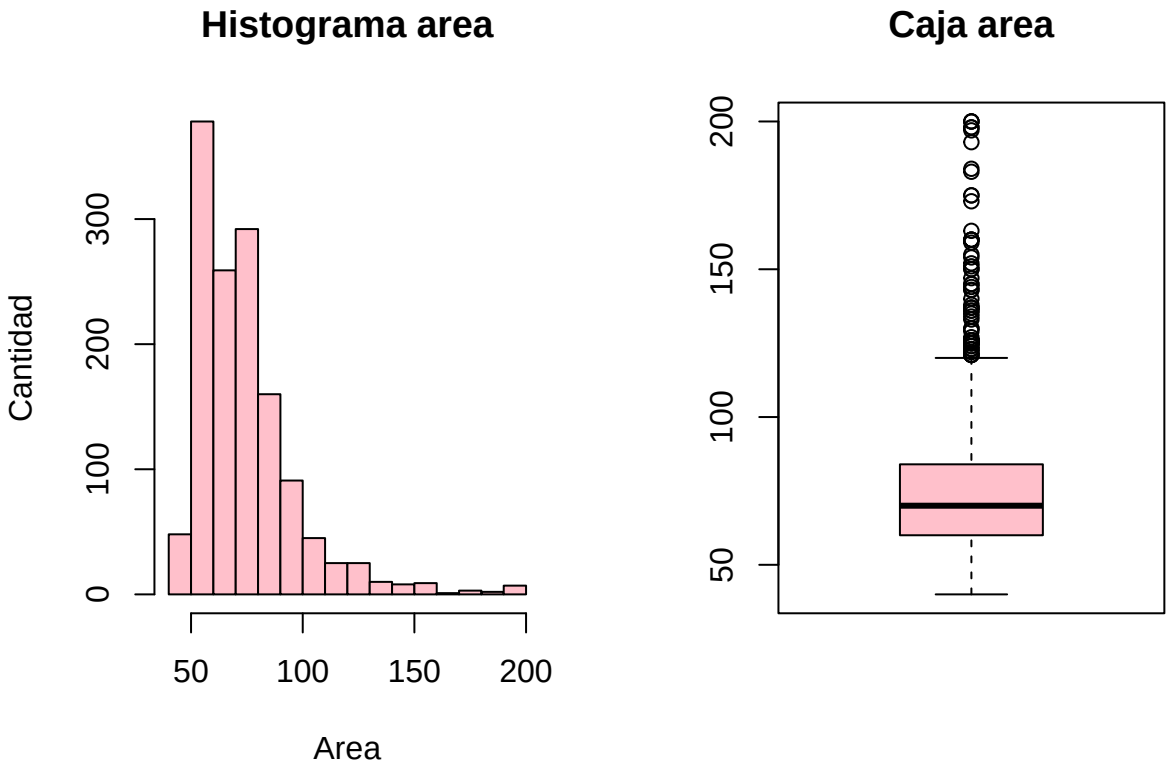
Table 1: Estructura de archivo vivienda4.RDS

	zona	estrato	preciom	areaconst	tipo
1	Zona Norte	4	220	52	Apartamento
3	Zona Norte	4	320	108	Apartamento
4	Zona Sur	4	290	96	Apartamento
5	Zona Norte	4	220	82	Apartamento
7	Zona Norte	4	220	75	Apartamento
8	Zona Norte	4	162	60	Apartamento

Table 2: Resúmen área

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
datosArea	40	60	70	75.47836	84	200

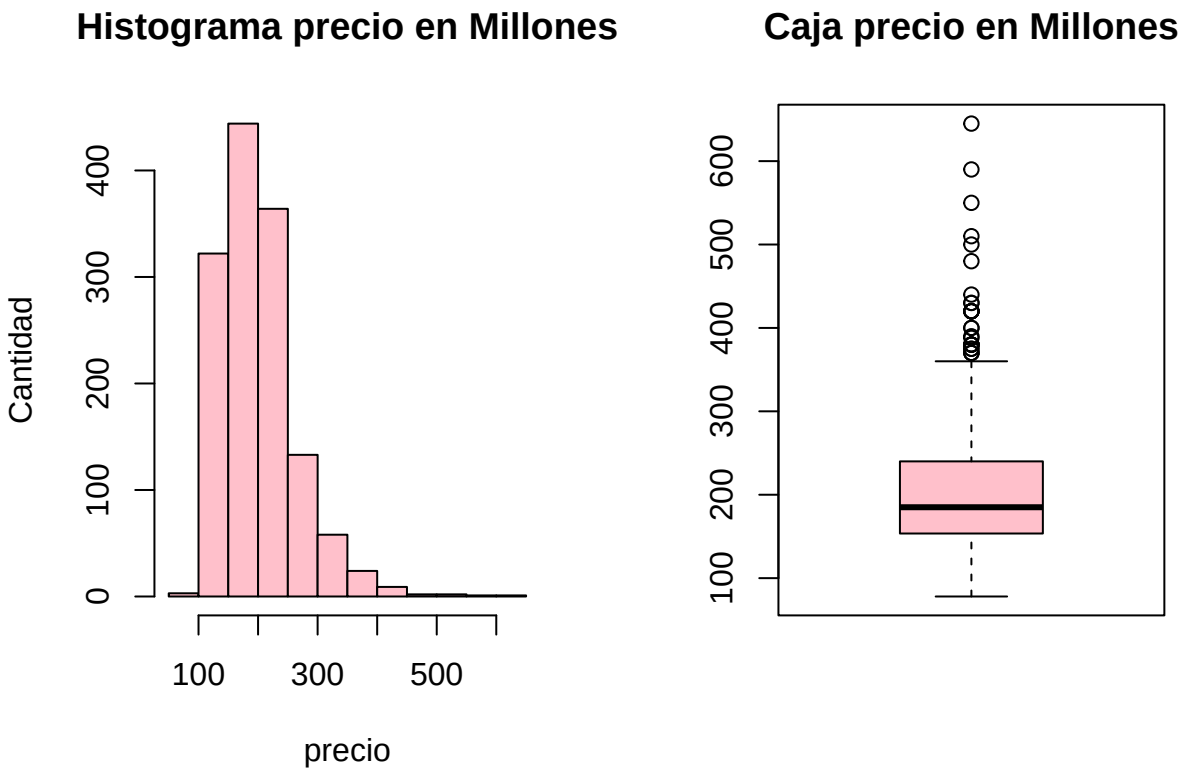
Se lee el archivo vivienda4.RDS el cual contiene 1706 viviendas (casas y apartamentos), se expluyen los de tipo casa para realizar un análisis detallado de un total de 1363 apartamentos con la siguiente estructura:



el promedio del área construida de apartamentos es de 75.4783566 m^2 para áreas que oscilan entre los 40 m^2 (mínimo) y los 200 m^2 (máximo). La distribución del área de las viviendas muestra un sesgo pronunciado a la derecha; podemos observar que el 50% de las viviendas tienen entre 40 m^2 y 70 m^2 mientras que el otro 50% tiene entre 70 m^2 y 200 m^2.

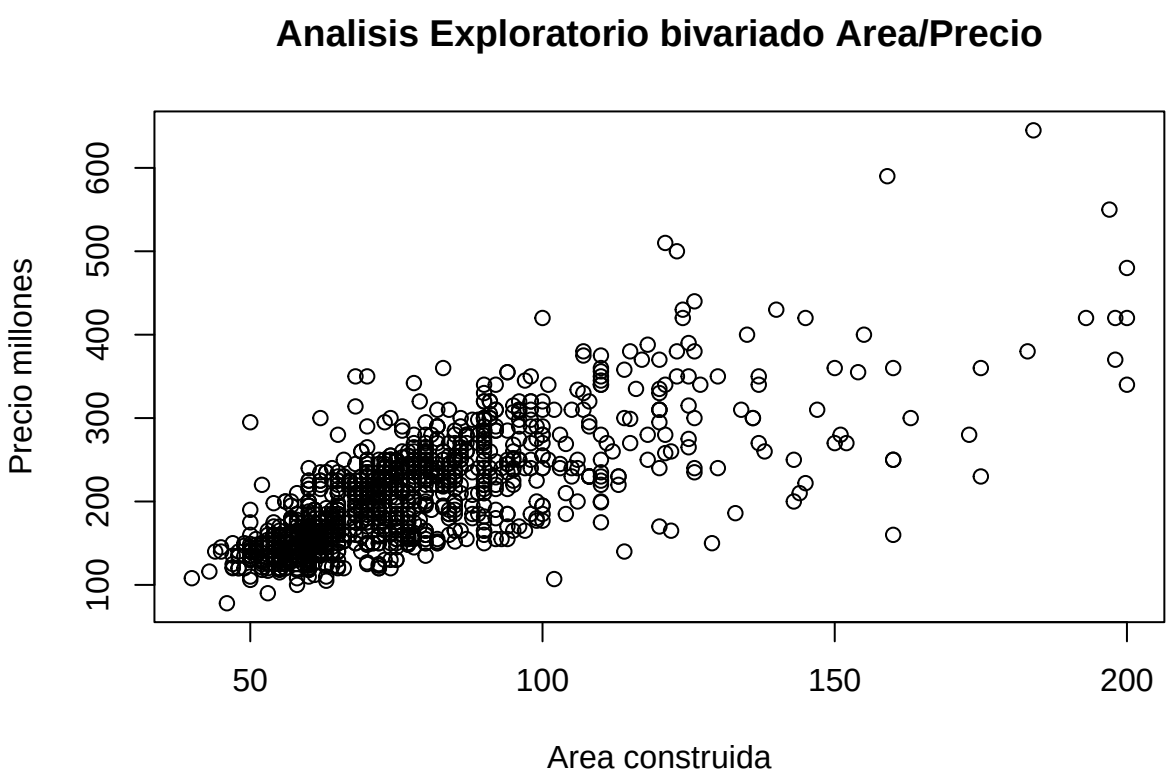
Table 3: Resumen precio en millones

	Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
datosPrecio	78	153.5	185	202.4373	240	645



En cuanto al precio los datos nos revelan que en promedio los apartamentos cuestan $ 202.437 millones, para valores que oscilan entre los $ 78 millones (mínimo) y los $ 645 millones (máximo). podemos precisar que la mitad de las viviendas se oferta a un precio menor o igual a los $ 185 millones y sólo un 25% están en un rango de costosas superando los $ 240 millones de pesos colombianos.

2. Realice un análisis exploratorio bivariado de datos enfocado en la relación entre la variable respuesta (precio) en función de la variable predictora (área construida) - incluir gráficos e indicadores apropiados interpretados.




```
## [1] 0.7481389
```

Observando la gráfica de dispersión podemos destacar una tendencia directa o creciente en la relación del área construida con el precio de la vivienda, sin embargo, el comportamiento del precio es más disperso para áreas mayores a 100 m^2, lo que puede dificultar la representatividad de un modelo de regresión lineal. Observando el Coeficiente de Correlación de Pearson 0.7481389 podemos determinar que la asociación que mide la relación lineal entre el precio y el área construida es una relación lineal positiva débil.

3. Estime el modelo de regresión lineal simple entre $precio = f(area) + \epsilon$. Interprete los coeficientes del modelo $\beta 0, \beta 1$ en caso de ser correcto.

```
## (Intercept)   areaconst
##   39.046787    2.164733
```

El modelo de regresión lineal se estima en: $Precio = 39.0467873 + 2.1647329 * areaconst$

$\beta 0 = 39.0467873$, este coeficiente nos indica el valor de la variable precio promedio cuando no se tiene una vivienda construida, se podría pensar que es el valor promedio de la base del terreno en ausencia de área construída .

$\beta 1 = 2.1647329$, indica que por cada metro cuadrado construido adicional el valor de la vivienda incrementara en promedio 2.1647329 millones aproximadamente.

4. Construir un intervalo de confianza (95%) para el coeficiente $\beta 1$, interpretar y concluir si el coeficiente es igual a cero o no. Compare este resultado con una prueba de hipótesis t.

```
##              2.5 %   97.5 %
## areaconst 2.06264 2.266826
```

Con un nivel de confianza del 95% y un 5% de error α, podemos determinar que el coeficiente $\beta 1$ para el caso del área construida podría tomar un valor entre los 2.0626399 y los 2.266826 millones por cada metro cuadrado construido. dado que el 0 no esta incluido en el intervalo de confianza, podemos concluir que se rechaza la hipótesis nula y se concluye que hay suficiente evidencia para no aceptar la nulidad de este coeficiente. Si comparamos este resultado con la prueba t que nos da el output del punto anterior, el p-value para el coeficiente $\beta 1$ fue $2.2e-16$ lo cual quiere decir que es menor al nivel de significancia de 0,05 (5%), esto apoya el rechazo de la hipótesis de nulidad y apoya la conclusión obtenida con el intervalo de confianza.

5. Calcule e interprete el indicador de bondad y ajuste R^2.

```
## [1] "Coeficiente de determinación (R2):  0.559711740661315"
```

El indicador de bondad y ajuste para esta relación es de 0.5597117, esto quiere decir que el modelo explica un 55% de la variación del precio de la vivienda. De forma general podemos decir que el modelo tiene poca fuerza para predecir a la variable dependiente.

6. ¿Cuál sería el precio promedio estimado para un apartamento de 110 metros cuadrados? Considera entonces con este resultado que un apartamento en la misma zona con 110 metros cuadrados en un precio de 200 millones sería una atractiva esta oferta? ¿Qué consideraciones adicionales se deben tener?.

```
##        fit      lwr      upr
## 1 277.1674 192.0449 362.2899
```

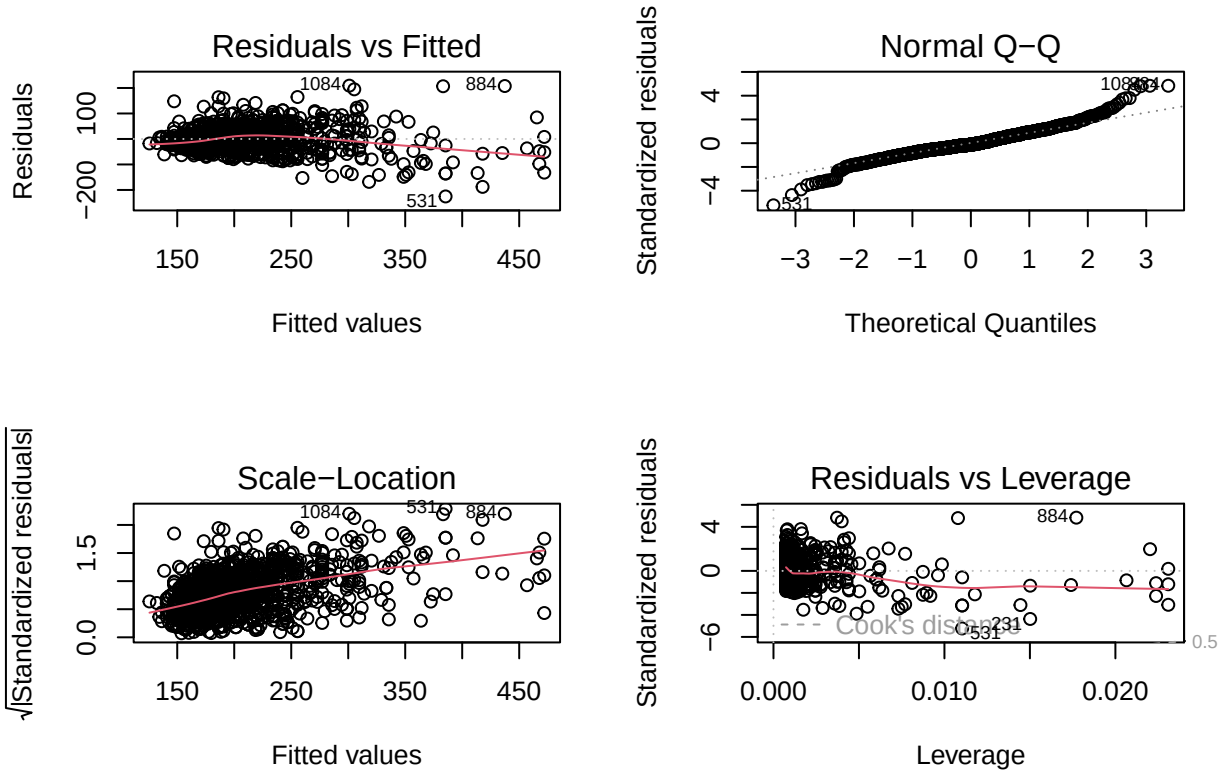
Con base en los datos obtenidos en los puntos anteriores, se puede establecer la siguiente ecuación de regresión del modelo, para estimar los precios de las viviendas en función del área.

Precio = 39.0467873*(Area) + 2.1647329.

El Precio para un apto de 110 mts2 es = 277.1674088 millones aproximadamente. por lo tanto se considera que 200 millones es una buena oferta para el comprador porque se estaría adquiriendo por debajo del precio estimado con un ahorro de 77 Millones aproximadamente

Otras consideraciones: El modelo debería tener en cuenta otras variables que no han sido consideradas para la toma de decisiones a la hora de escoger la vivienda, como por ejemplo: No. de habitaciones, años de construcción, ubicación, inseguridad, etc.

7. Realice la validación de supuestos del modelo por medio de gráficos apropiados, interpretarlos y sugerir posibles soluciones si se violan algunos de ellos. Utilice las pruebas de hipótesis para la validación de supuestos y compare los resultados con lo observado en los gráficos asociados.



De acuerdo con la gráfica #1 Residuals vs Fitted, se observa que los residuos tienen un comportamiento agrupado a la izquierda y pocos a la derecha por lo tanto al no estar distribuidos de manera uniforme no se cumple el supuesto de Homoscedasticidad, y en cuanto a la gráfica QQ-Plot se observan que aunque la mayoría de los datos concuerdan con la linea roja no todos los datos son distribuidos normalmente por lo tanto no se cumple normalidad. En cuanto al gráfico Scale-Location no se cumple el supuesto de No autocorrelación porque los errores están correlacionados. El gráfico Residuals vs Leverage no cumple el supuesto de Outliers ya que se evidencia datos atípicos es decir que hay errores estandarizados distanciados del resto

8. De ser necesario realice una transformación apropiada para mejorar el ajuste y supuestos del modelo.

```
mod_lin_lin <- lm(aptos$preciom~aptos$areaconst)
mod_lin_log <- lm(aptos$preciom~log(aptos$areaconst))
mod_log_lin <- lm(log(aptos$preciom)~aptos$areaconst)
mod_log_log <- lm(log(aptos$preciom)~log(aptos$areaconst))

summary(mod_lin_lin)
```

```
##
## Call:
## lm(formula = aptos$preciom ~ aptos$areaconst)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -225.404  -23.902   -4.754   25.763  209.021
```

```
##
## Coefficients:
##                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)     39.04679    4.09977   9.524   <2e-16 ***
## aptos$areaconst  2.16473    0.05204  41.595   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 43.34 on 1361 degrees of freedom
## Multiple R-squared:  0.5597, Adjusted R-squared:  0.5594
## F-statistic:  1730 on 1 and 1361 DF,  p-value: < 2.2e-16
```

```
summary(mod_lin_log)
```

```
##
## Call:
## lm(formula = aptos$preciom ~ log(aptos$areaconst))
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -196.252  -21.338   -1.579   22.096  261.436
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           -635.532     19.092  -33.29   <2e-16 ***
## log(aptos$areaconst)  195.419      4.445   43.97   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 41.98 on 1361 degrees of freedom
## Multiple R-squared:  0.5868, Adjusted R-squared:  0.5865
## F-statistic:  1933 on 1 and 1361 DF,  p-value: < 2.2e-16
```

```
ad.test(aptos$preciom)
```

```
##
##  Anderson-Darling normality test
##
## data:  aptos$preciom
## A = 29.913, p-value < 2.2e-16
```

```
ad.test(log(aptos$areaconst))
```

```
##
##  Anderson-Darling normality test
##
## data:  log(aptos$areaconst)
## A = 20.559, p-value < 2.2e-16
```

```
summary(mod_log_lin)
```

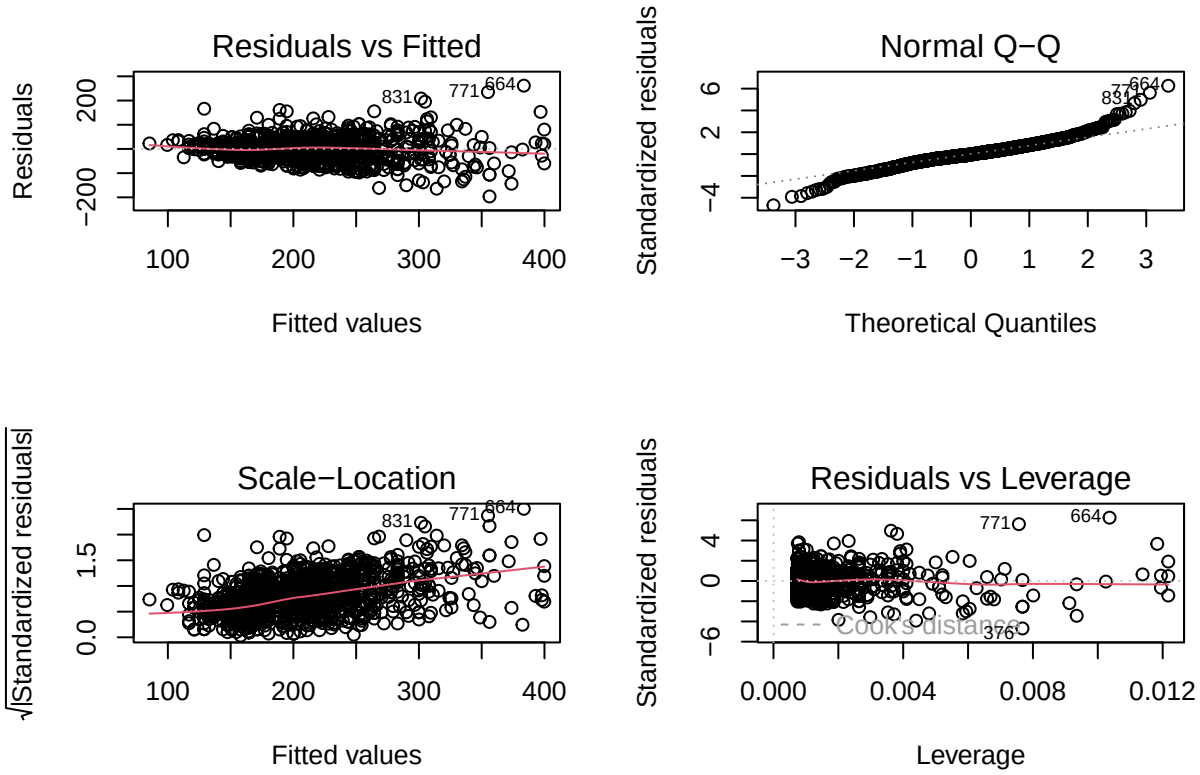
```
##
## Call:
## lm(formula = log(aptos$preciom) ~ aptos$areaconst)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.98857 -0.13188 -0.01249  0.15595  0.66387
##
## Coefficients:
##                   Estimate Std. Error t value Pr(>|t|)
## (Intercept)     4.5512586  0.0194001  234.60   <2e-16 ***
## aptos$areaconst 0.0094530  0.0002463   38.38   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.2051 on 1361 degrees of freedom
## Multiple R-squared:  0.5198, Adjusted R-squared:  0.5195
## F-statistic:  1473 on 1 and 1361 DF,  p-value: < 2.2e-16
```

```
summary(mod_log_log)
```

```
##
## Call:
## lm(formula = log(aptos$preciom) ~ log(aptos$areaconst))
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -0.8890 -0.1119  0.0028  0.1343  0.7538
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)           1.48373    0.08703   17.05   <2e-16 ***
## log(aptos$areaconst)  0.88175    0.02026   43.52   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1914 on 1361 degrees of freedom
## Multiple R-squared:  0.5819, Adjusted R-squared:  0.5816
## F-statistic:  1894 on 1 and 1361 DF,  p-value: < 2.2e-16
```

el r2 ajustado aumentó para la ecuación lineal - logaritmo, se procede a análizar sus gráficos

9. De ser necesario compare el ajuste y supuestos del modelo inicial y el transformado.



En el modelo inicial que fue el lineal su R2 fue de 0.5597 pero en el modelo que resultó el mejor r2 ajustado es decir el lineal logaritmo aumentó en 0.5865, para este ultimo modelo transformado tampoco se cumple los supuestos ya que al comprobar su p-value con el test de normalidad resultó menor a 0.05%, por lo tanto se concluye que a pesar de mejorar su r2 ninguno de los modelos cumplen los supuestos de Homoscedasticidad, Normalidad, No autocorrelación y el supuesto de Outliers.

10. Estime varios modelos y compare los resultados obtenidos. En el mejor de los modelos, ¿se cumplen los supuestos sobre los errores?

Se estimaron diferentes modelos pero ninguno es acorde para dar una información detallada a la inmobilidaria, si se consideran otras variables a fututo se podría estimar un modelo de regresión lineal multiple dependiente de multiples variables.