Práctica (PR). Modelos de Regresión

Pérez Efremova, Daniel

Noviembre 2020

Resumen

En esta práctica vamos a utilizar el conjunto de datos *precioscasas.dat* que recoge los precios de un conjunto de viviendas y otras características asociadas. El objetivo es explicar, lo mejor posible, la variabilidad del precio de las viviendas a través de un modelo de regresión múltiple ajustado sobre las características asociadas al precio de una vivienda.

Índice

Introducción al conjunto de datos y las variables **2**

1. Estudio preliminar de los datos **2**

2. Planteamiento del modelo **4**
2.1. Ajuste 4
2.2. Diagnóstico 6

3. Conclusiones **10**

Índice de Códigos

1. Primer modelo general 4
2. Ajuste de un segundo modelo y contraste de regresión sobre las variables con coeficientes poco significativos 5
3. Reajuste del modelo y constraste de normalidad de los residuos . 9

Introducción al conjunto de datos y las variables

Los datos contienen 100 observaciones de las variables (en columnas):

- Primera (y): precios de viviendas en euros (*precio*)
- Segunda (x_1): superficie en metros cuadrados (*superf*)
- Tercera (x_2): numero de cuartos de baño (*cuartosb*)
- Cuarta (x_3): número de dormitorios (*dorm*)
- Quinta (x_4): número de plazas de garaje (*plazasg*)
- Sexta (x_5): edad de la vivienda (*edadv*)
- Séptima (x_6): 1 buenas vistas, 0 vistas corrientes (*vistas*)

1. Estudio preliminar de los datos

Lo primero que vamos a mirar es la correlación entre las variable a través de una matriz de diagramas de dispersión que además contenga los coeficientes de correlación lineal.

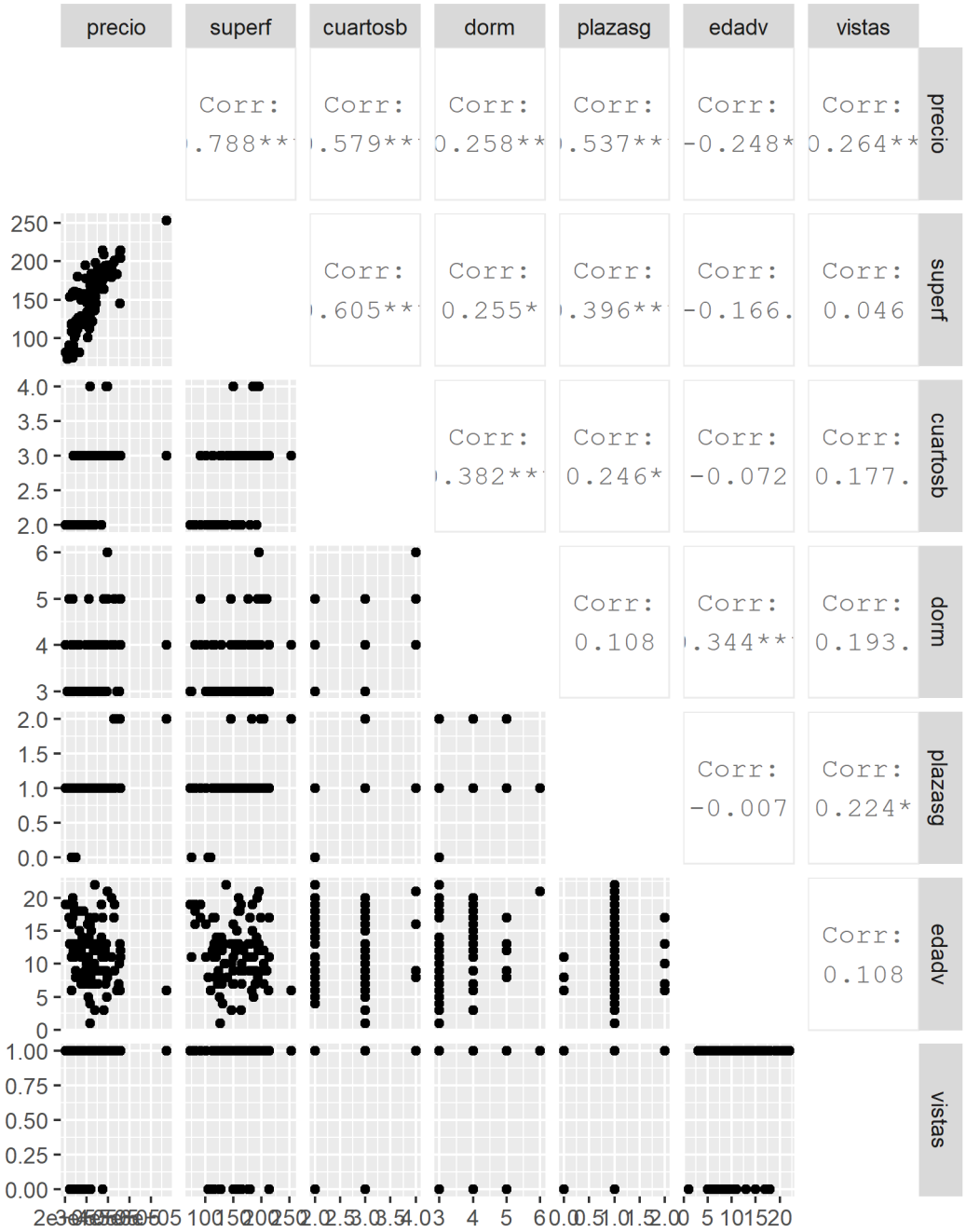
En la Figura 1 vemos que las variables que muestran mayor correlación lineal univariante con los precios de las viviendas son: la superficie (0,788), el número de cuartos de baño (0,579) y el número de plazas de garaje (0,537). Sin embargo, al plantear el modelo, debemos tener cuidado con la multicolinealidad, ya que si incluimos las variables superficie y cuartos de baño, estas presentan un coeficiente de correlación lineal univariante significativamente alto (0,605), lo mismo podemos decir de las variables superficie y número de plazas de garaje ($\approx$ 0,4).

Por tanto, si planteamos algún modelo que incluya las tres variables deberemos medir los efectos de dos ellas en la tercera para evaluar si podemos eliminar una de las variables por aportar información redundante.

Por otro lado, aunque los demás coeficientes de correlación univariante no sean significativamente altos, no tenemos evidencia clara para descartar alguna variable. Es decir, no podemos descartar que los valores de dos o más de las variables tengan una relación lineal significativa con los precios, mayor de lo que observamos de manera univariante, algo que se hace complicado de medir de forma sencilla y que desarrollaremos más adelante.

Así es que, dada la cantidad moderada de variables, plantearemos un primer modelo preliminar que incluya todas las variables e iremos descartando variables hasta dar con uno razonable.

Figura 1: Matriz de gráficos de dispersión y matriz de correlaciones



2. Planteamiento del modelo

Con la notación de la introducción planteamos un modelo de regresión lineal múltiple (RLM):

$$y_i = \beta_0 + \beta_1 x_{1i} + \cdots \beta_6 x_{6i} + u_i$$

donde y_i son los valores de la variable respuesta (precios que pretendemos explicar), x_{ij} los valores de las variables regresoras y los coeficientes β_j miden el efecto marginal sobre la respuesta de un aumento unitario en x_j cuando el resto de variables regresoras permanecen constantes. Introducimos también la perturbación u_i que será el error del modelo y que debe cumplir algunas especificaciones para el modelo sea válido [1]:

1. $E[u] = 0$
2. $Var(u_i) = \sigma^2 = cte$
3. (independencia en las perturbaciones) $cov(u_i, u_j) = 0$
4. $u \sim N(0, \sigma^2)$
5. El número de observaciones es mayor que la cantidad de variables.
6. Ninguna de las variables explicativas es combinación lineal exacta de las demás.

2.1. Ajuste

Dada la cantidad de variables, usaremos la librería interna de R para ajustar distintos modelos en lugar de implementar los cálculos. Vemos en el siguiente código como ajustar un modelo general que incluya todas las variables y la tabla resumen.

Código 1: Primer modelo general

```
modelo_general = lm(formula = precio ~ superf + cuartosb + dorm  +
    plazasg + edadv + vistas, data=datos)
summary(general_model)

Residuals:
Min       1Q  Median      3Q     Max
-101248  -23050    -345   18036  141928

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  29844.7    26365.3    1.132  0.26056
superf        1159.3      142.9    8.112 1.98e-12 ***
cuartosb     13284.5     9286.2    1.431  0.15591
dorm          8695.2     6708.7    1.296  0.19814
plazasg      59777.1    14604.0    4.093 9.06e-05 ***
edadv        -3198.4      974.3   -3.283  0.00145 **
```

[1] En atención al teorema de Gauss-Markov puede probarse que, con estas hipótesis, los estimadores que se deducen de aplicar el método de mínimos cuadrados son centrados y de mínima varianza. Aun que es muy complejo probar directamente estas hipótesis, las asumiremos ciertas en ausencia de gran evidencia de que no se cumpla alguna.

```
vistas        34312.9     10963.6    3.130   0.00234 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 38920 on 93 degrees of freedom
Multiple R-squared:  0.7505,     Adjusted R-squared:  0.7344
F-statistic: 46.61 on 6 and 93 DF,  p-value: < 2.2e-16
```

Vemos en el resultado del Código 1, los coeficientes estimados para cada variable, su error estándar de estimación[2] y el estadístico t del contraste individual de regresión. Todas las variables parecen tener una relación positiva con el precio de las viviendas (al incrementar una unidad la variable x_i, si el resto permanece constante, el precio de la vivienda, y, aumenta una cantidad β_i) a excepción de la variable *edadv*, que tiene un efecto negativo en los precios, es decir, al incrementar una unidad la variable *edadv*, si el resto permanece constante, el precio de la vivienda, y, disminuye una cantidad β_{edadv}.

Vemos también la desviación típica residual, los coeficientes de determinación (el usual y el ajustado por grados de libertad) y el estadístico F del contraste de regresión global (contraste que todos los coeficientes sean no nulos). Los coeficientes de determinación son relativamente altos y el estadístico $F = 46{,}61$ es muy alto para compararse con una $\mathbf{F}_{6,93}$ de Snedecor, por lo que el test arroja un $p-$valor muy bajo y todos los coeficientes son significativos, al menos, globalmente.

Algo que debemos tener en cuenta es que los $p-$valores de la variables *cuartosb* y *dorm* son más altos que cualquier nivel de significación usual, por lo que debemos comprobar si son significativas conjuntamente. Para ello planteamos el test:

$$H_0 : \beta_{dorm} = 0, \beta_{cuartosb} = 0$$
$$H_1 : \beta_{dorm} \neq 0, \beta_{cuartosb} \neq 0$$

y para realizar el contraste, ajustaremos un segundo modelo que no tenga en cuenta estas variables. Sean $\triangle VE(i) = VE(k) - VE(i)$ la diferencia de variabilidades explicadas entre los dos modelos y $VNE(k)$ la variabilidad no explicada por el primer modelo, si la hipótesis nula es cierta, entonces ambos términos son independientes, y divididos entre la varianza residual σ^2, se distribuirán como una χ^2, luego, nuestro estadístico de contraste es:

$$F^* = \frac{\triangle VE(i)/i}{VNE(k)/(n-k-1)} \sim F_{(i,n-k-1)}$$

y rechazamos H_0 a nivel de significación α cuando sea $F^* > F_{i,n-k-1;1-\alpha}$.
Vemos en el Código 2 el ajuste del segundo modelo y el resultado del contraste.

[2]Debemos notar que las desviaciones típicas son estables, ninguna es extremadamente grande, por lo que no tenemos evidencia de que exista multicolinealidad entre las variables del modelo.

Código 2: Ajuste de un segundo modelo y contraste de regresión sobre las variables con coeficientes poco significativos

```
particular_model = lm(formula = precio~superf+plazasg+edadv+vistas,
 data=datos)
summary(particular_model)

Residuals:
Min      1Q  Median      3Q     Max
-94251 -20415   -610   18670  136740

Coefficients:
             Estimate  Std. Error t value Pr(>|t|)
(Intercept)  60096.1    23039.6    2.608 0.010567 *
superf        1335.7      117.7   11.345  < 2e-16 ***
plazasg      57908.9    14822.5    3.907 0.000175 ***
edadv        -2658.6      905.4   -2.936 0.004167 **
vistas       39739.8    10881.2    3.652 0.000426 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 39570 on 95 degrees of freedom
Multiple R-squared:  0.7365,    Adjusted R-squared:  0.7254
F-statistic: 66.37 on 4 and 95 DF,  p-value: < 2.2e-16

## calculos del contraste

 # VE primer modelo
VE_gm = sum((general_model$fitted.values - mean(datos$precio))**2)

# VE segundo modelo
VE_pm = sum((particular_model$fitted.values - mean(datos$precio))**2)

# diferencia de variabilidades explicadas entre los modelos
diff_VE = VE_gm-VE_pm

i = 2 # numero de variables que conforman el contraste

# estadistico de contraste
F = diff_VE/2 / sigma(general_model)**2
pvalor = pf(F, i, general_model$df.residual, lower.tail = TRUE)


print(c('estadistico F: ', F))
print(c('pvalor: ', pvalor))

[1] "estadistico F: "  "2.60575574090814"
[1] "pvalor: "         "0.920334015827924"
```

Vemos que el segundo modelo, sin las variables *cuartosb* y *dorm*, tiene unas propiedades similares a las del primero, pero ahora, el posible efecto que tuvieran ambas se recoge en la ordenada en el origen (β_0) que ha doblado su valor. Además los coeficientes de determinación apenas han variado.

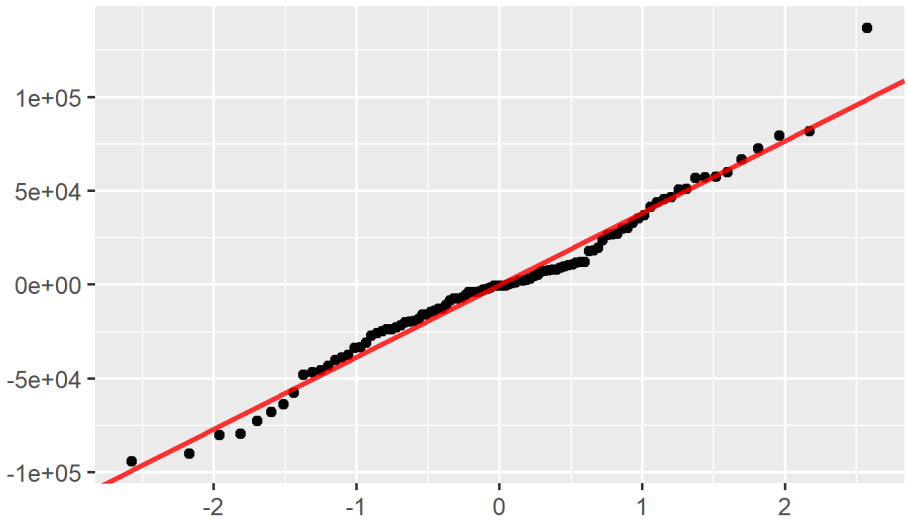
El test planteado arroja un $p-$valor muy alto, por lo que no deberíamos rechazar H_0, es decir, no hay evidencia de que los coeficientes de las variables sean significativos y por tanto, nos quedaremos con el segundo modelo.

2.2. Diagnóstico

Ahora vamos a estudiar si el modelo cumple las especificaciones iniciales. En particular, la hipótesis más relevante es la de normalidad de los residuos.

Vemos en la Figura 2 los cuantiles de los residuos frente a los teóricos de la distribución normal y el cálculo de los coeficientes de asimetría y curtosis.

Figura 2: Gráfico cuantil-cuantil de los residuos del modelo




```
print(c('asimetria:', skewness(particular_model$residuals)))
print(c('curtosis:', kurtosis(particular_model$residuals)))

"asimetria:"   "0.245285595898466"
"curtosis:"    "4.05737138509419"
```

En el gráfico de los cuantiles apreciamos que las colas se desvían mucho de la normalidad aunque esta desviación se produce tanto a la izquierda como a la derecha, por tanto la distribución es simétrica como confirma el coeficiente de asimetría que es relativamente cercano a cero. Las desvaciones mas grandes se producen en la cola derecha, como confirma el coeficiente de kurtosis que es ligeramente mayor a 3.

Podemos plantear un test de normalidad basado en la asimetría y curtosis como proponen Bera y Jarque (1980), especialmente indicado en regresión.

$$H_0 : \text{La distribución de la muestra es normal}$$

$$H_1 : \text{La distribución de la muestra no es normal}$$

Llamando a y k a los coeficientes de asimetría y curtosis, el estadístico de contraste es

$$J = n\left(\frac{a^2}{6} + \frac{(k-3)^2}{24}\right) \sim \chi^2_2$$

y rechazamos H_0 si $J > \chi^2_{1-\alpha}$. Vemos en el siguiente código los cálculos del test y el pvalor asociado.

```
a = skewness(particular_model$residuals)
k = kurtosis(particular_model$residuals)
J = n*((a**2)/6 + ((k-3)**2)/24)

pvalor = pchisq(J, 2, ncp = 0, lower.tail = FALSE, log.p = FALSE)
print(pvalor)

[1] 0.05897668
```

El pvalor del test es muy bajo, pero no lo suficiente como para rechazar con seguridad la hipótesis de normalidad.

Concluimos que la falta de normalidad en los residuos se debe a unos pocos datos atípicos de la muestra y deben ser estudiados para ver si han comprometido los resultados. Para detectar estos datos atípicos y/o influyentes usaremos el paquete *olsrr* que contiene diversas funciones para la diagnosis de un modelo de regresión.

En particular estamos interesados en el gráfico de los residuos estudentizados frente a los valores ajustados (Figura 3) para saber qué observaciones tienen residuos excesivamente altos, el estadístico de Cook (Figura 4) para ver la capacidad de atraer la recta de regresión de cada observación y la variación de en la estimación de los coeficientes al sustraer elementos de la muestra (Figura 5) para saber cómo influye en la estimación de cada coeficiente una determinada observación.

Figura 3: residuos estudentizados frente a los valores ajustados

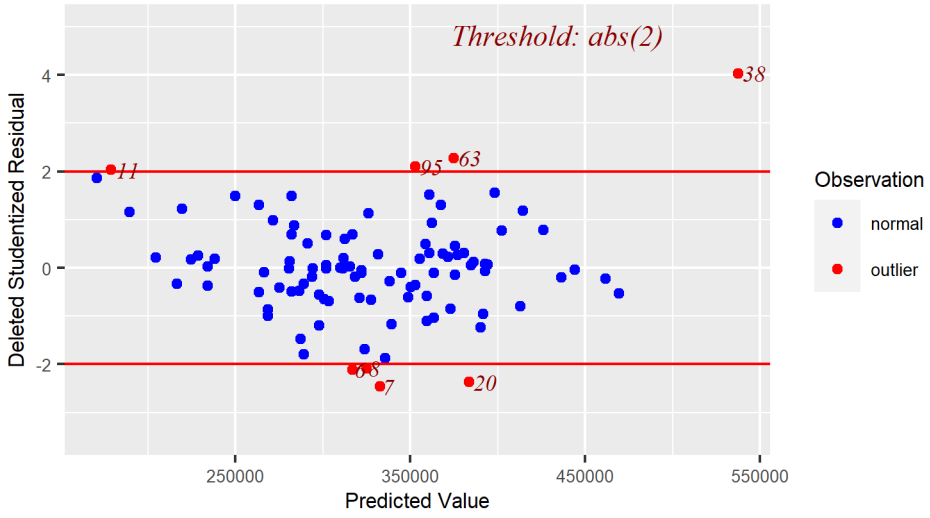
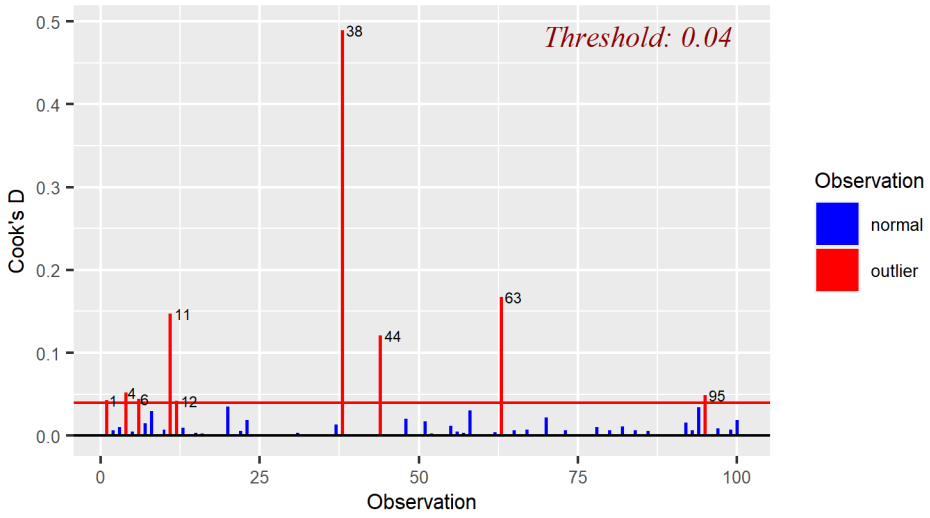


Figura 4: Diagrama de barras del estadístico de Cook



De las Figuras 3 y 4 deducimos que las observaciones 11, 38, 44, 63 y 95 tienen una influencia reseñable en la recta y además un residuo estudentizado muy alto. Esto nos indica que están ejerciendo un efecto tractor sobre la recta no despreciable y que además el error de estimación en esos valores (medido por el residuo estudentizado) es alto, por lo que a pesar de que han atraído la recta, la estimación en esos puntos no es precisa.

En la Figura 5 confirmamos que cuando extraemos de la muestra alguna de estas observaciones influyentes, los coeficientes del modelo varían en mayor medida que cuando extraemos alguna de las no influyentes, existiendo así cierta evidencia de que estas observaciones parecen haberse generado por un procedimiento distinto al resto. Vamos eliminar estas observaciones del ajuste, aunque esto no quiere decir que las eliminemos sin más, sino que deben ser estudiadas por separado para ver la causa de tal heterogeneidad.

El ajuste y resumen del nuevo modelo lo vemos en el siguiente código.

Código 3: Reajuste del modelo y constraste de normalidad de los residuos

```
datos2 = datos[-c(11, 38, 44, 63, 95),]

particular_model = lm(formula =
precio~superf+plazasg+edadv+vistas, data=datos2)
print(summary(particular_model))

a = skewness(particular_model$residuals)
k = kurtosis(particular_model$residuals)
```

```
J = n*((a**2)/6 + ((k-3)**2)/24)

pvalor = pchisq(J, 2, ncp = 0, lower.tail = FALSE, log.p = FALSE)
print(pvalor)

[1]
Residuals:
Min      1Q  Median      3Q     Max
-90234 -17824    3349   17107   65393

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  70007.1    23130.4   3.027  0.00322 **
superf        1305.6      104.8  12.453  < 2e-16 ***
plazasg      45644.0    16361.1   2.790  0.00644 **
edadv        -2508.0      798.6  -3.140  0.00228 **
vistas       40929.3     9544.6   4.288 4.52e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 34040 on 90 degrees of freedom
Multiple R-squared:  0.7393,    Adjusted R-squared:  0.7277
F-statistic:  63.8 on 4 and 90 DF,  p-value: < 2.2e-16

[2] 0.1345386
```

Podemos observar cambios importantes en el modelo:

- La ordenada en el origen ha aumentado su valor y el coeficiente de plazas de garaje (*plazasg*) ha disminuido.
- La desviación típica residual ha disminuido.
- El pvalor del test de normalidad de los residuos ahora sí podría considerarse significativo (aun que sigue sin ser excesivamente alto) y los residuos pueden considerar normales.

Por lo que escogeremos este modelo por tener unas propiedades mucho mas deseables que el modelo ajustado con las observaciones atípicas.

3. Conclusiones

En base a los resultados anteriores y los datos disponibles, concluimos que el conjunto de variables que mejor explica la variabilidad del precio de las viviendas es *superf*, *plazasg*, *edadv* y *vistas*.

Además existen evidencias de que la relación entre éstas y el precio de las viviendas es lineal:

$$precio = 70007{,}1 + 1305{,}6 \cdot superf + 45644 \cdot plazasg - 2508 \cdot edadv + 40929{,}3 \cdot vistas$$

y la interpretación de este modelo es la siguiente [3]:

[3] Notemos que la ordenada en el origen en nuestro problema carece de interpretación práctica. Si todas las variables toman valores nulos, entonces una vivienda sin superficie no tiene sentido ($superf = 0$).

- Si todas las demás variables permanecen constantes, un incremento de una unidad de superficie (*superf*) incrementa el precio promedio de la vivienda en 1305,6 euros.
- Si todas las demás variables permanecen constantes, un incremento de una unidad en la cantidad de plazas de garaje (*plazasg*) incrementa el precio promedio de la vivienda en 45644 euros.
- Si todas las demás variables permanecen constantes, un incremento de una unidad en la edad de la vivienda (*edadv*) disminuye el precio promedio de la vivienda en 2508 euros.
- Si todas las demás variables permanecen constantes, que la vivienda tenga buenas vistas ($vistas = 1$) incrementa el precio promedio en 40929,3 euros.

Figura 5: Variación de los coeficientes del modelo eliminado las observaciones