Taller # 3

Diagnóstico de regresión lineal múltiple

Julián Camilo Riaño Moreno

miércoles, abril 08, 2020

Contents

Descripción de las variables.

Table 1: Organizacion de las variables del taller#3

Variables dadas	Definicion	Tipo de variable (en modelo)	Nombre de variable (en la base de datos)	Unidad
y	Precio de venta de la casa	v_respuesta	precioventa	x/1000
x_1	Impuestos (local, escuela, condado)	v_regresora	impuestos	x/1000
x_2	Numero de banos	v_regresora	num_banos	nÂř
x_3	Tamano de lote	v_regresora	tamano_lote	ft^2
x_4	Espacio vital	v_regresora	espac_vital	ft^2
x_5	Numero de puestos de garaje	v_regresora	num_puestos_gar	nÂř
x_6	Numero de habitaciones	v_regresora	num_hab	nÂř
x_7	Numero de dormitorios	v_regresora	num_dorm	nÂř
x_8	Edad del hogar	v_regresora	edad_hogar	Anos
x_9	Cantidad de chimeneas	v_regresora	cant_chimen	nÂř

En la tabla 1. se describen las caracteristicas de las variables dadas para el ejercicio. Para ajustar las variables definidas se les asignó un nombre para la base de datos que corresponde a la columna "Nombre de variable (en la base de datos)". Además, se definió para este ejercicio la variable respuesta `precioventa`, la cual corresponde a una variable cuantitativa continua; las demás variables se definieron como variables regresoras (`impuestos`, `num_banos`, `tamano_lote`, `espac_vital`, `num_puestos_gar`, `num_hab`, `num_dorm`, `edad_hogar`, `cant_chimen`), de estas `num_banos`, `num_puestos_gar`, `num_hab`, `num_dorm`, `edad_hogar`, son consideradas variables cuantitativas discretas, las son variables cuantitativas continuas.

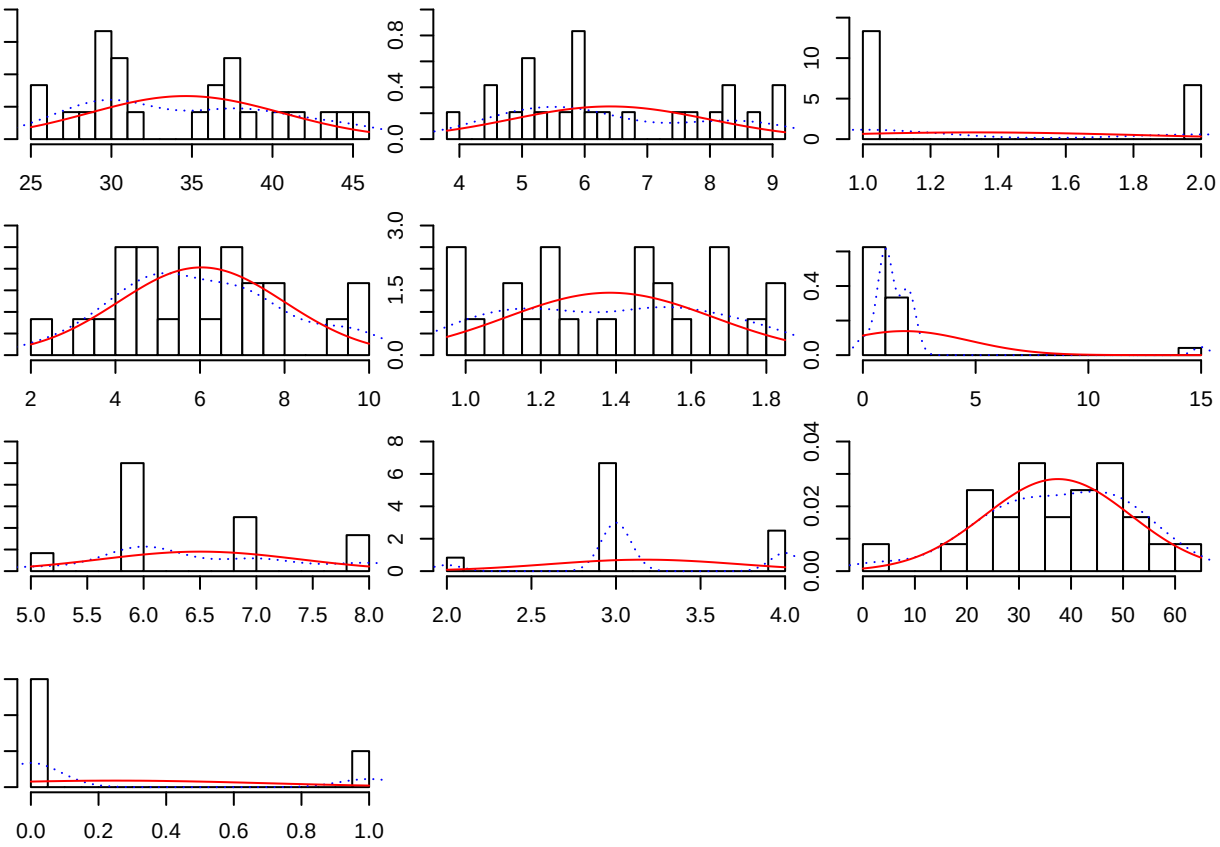

Figure 1: Distribución de las variables dadas

La figura 1, muestra la distribución de las 9 variables que serán análizadas en este documento. En esta figura se deja ver que la variable `num_banos` (x_2) al igual que la variable `cant_chimen` presentan múltiples datos extremos. Las demás variables presenta una distribución normal.

Respuesta a la preguntas taller # 3

Problema

Considere los datos de precios de la vivienda dados en la tabla 1. a. Ajuste un modelo de regresióon múultiple que relacione el precio de venta con los nueve regresores. segundo. Prueba de signifícación de la regresión. ¿Qué conclusiones puedes sacar?. Utilice las pruebas t para evaluar la contribución de cada regresor al modelo. Discute tus hallazgos.

Pregunta #1: Ajuste un modelo de regresión múltiple que relacione el precio de venta con los nueve variables regresoras.

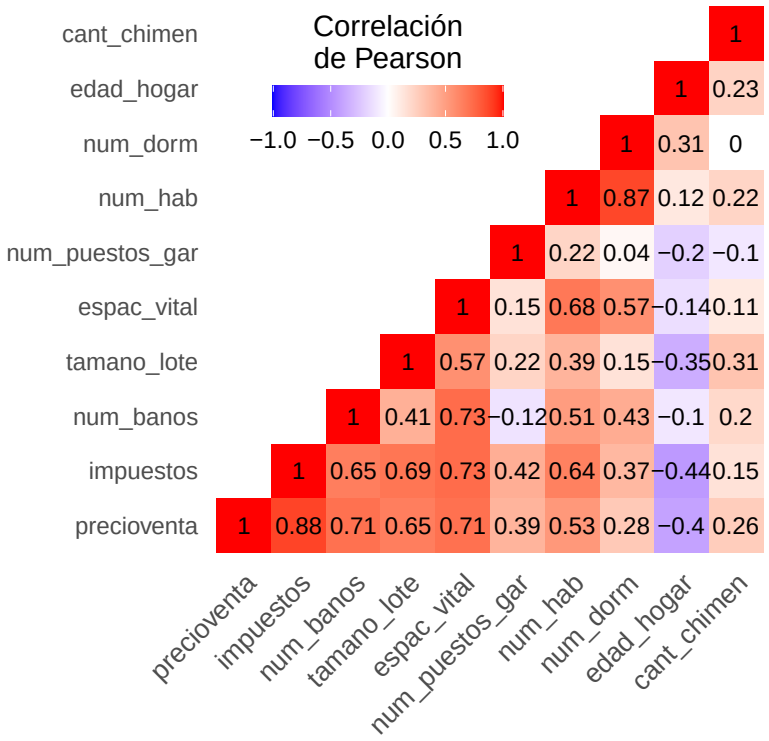

Figure 2: Mapa de calor de correlaciones entre variables

Inicialmente se realizó un análisis de correlación de las variables dadas por medio del coeficiente de correlación de Pearson (coeficientes mostrados en la figura 2). A través de este examen se encontró que la variable regresora `impuestos` (cor=0.81) es la más correlacionada con la variable respuesta `precioventa`. Otras variables con alta correlación con la variable respuesta fueron, `num_banos` y `espac_vital` (cor=0.71), seguidas de `tamano_lote`(cor=0.65). Entre las variables regresoras se encuentran alta correlación entre `num_hab` y `num_dorm` (cor=0.87), seguidas de `espac_vital` con `num_banos` e `Impuestos`(cor=0.73). Las variables con menor correlación son `num_dorm` con `cant_chimen` y `num_puestos_garcon` `num_dorm`.

Pregunta #2: Prueba de signicancia de la regresión. ¿Qué conclusiones puedes sacar?

Se elaboró un modelo de regresión original realizado entre las variable respuesta `precioventa` y las otras nueve variables regresoras. Para evaluar el modelo, se realizó un análisis de la F de Fisher para el modelo que llamaremos original con 9 variable y un análisis de R^2_{adj}.

Table 2: Estadistico F de Fisher: significancia del modelo

F_{stat}	Grados_libertad	$p - value$
10.5619486137295	14	7.694e-05

En la tabla 2. se encuentran los resultado de F de fisher, para el análisis de la siguiente prueba de hipotesis:

$$H_o : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \; para \; cualquier \; j \; dado$$

El estadístico F de Fisher se obtuvo con 14 grados de libertad, donde el $p - value$ demuestra significancia estadística, con lo que se rechaza la hipotesis nula. De tal forma, se puede inferir que almenos una de la variables regresoras (β_i) son diferentes de 0.

Table 3: R^2 y R^2_{adj} para el modelo

R^2	R^2_{adj}
0.872	0.789

Por otra parte, la tabla 3, muestra los resultados de los R^2 y R^2_{adj} para el modelo de regresión original. Estos resultados ($R^2 \neq R^2_{adj}$). Para el análisis se tendrá en cuenta el R^2_{adj} ya que es el menos sensible a la supresión o adición de variables.De manera que este resultado, permiten inferir que aproximadamente el 79% de la variabilidad de la variable respuesta (`precioventa`) puede ser explicado por las variables regresoras de este modelo; de manera que, la variabilidad restante puede ser consecuencia a otras variables no tenidas encuenta en el modelo o por azar.

Table 4: Intervalos de confianza del 95% para los parametros del modelo original

	IC 2.5%	IC 97.5%
precioventa	8.67	31.8
impuestos	-1.04	3.53
num_banos	0.345	10.1
tamano_lote	-0.823	1.13
espac_vital	-6.83	10.6
num_puestos_gar	-0.00951	1.21
num_hab	-5.69	3.86
num_dorm	-5.67	7.52
edad_hogar	-0.215	0.0598
cant_chimen	-1.12	6.66

La tabla 4. muestra los intervalos de confianza para los parametros del modelo original, destacándose la amplitud encontrada resultadas para las variables `cant_chimen`, `num_dorm` y `num_hab`.

Pregunta #3: Utilice las pruebas t para evaluar la contribución de cada variable regresora al modelo. Discute tus hallazgos.

Table 5: Estimaciones y $t - values$ para el modelo obtenido

	Estimado	ErrorStand	$t - value$	$p - value$
(Intercept)	20.2	5.39	3.75	0.00214
impuestos	1.25	1.06	1.17	0.262
num_banos	5.23	2.28	2.3	0.0376

	Estimado	ErrorStand	$t-value$	$p-value$
tamano_lote	0.156	0.456	0.341	0.738
espac_vital	1.89	4.07	0.466	0.649
num_puestos_gar	0.598	0.283	2.11	0.0532
num_hab	-0.914	2.23	-0.41	0.688
num_dorm	0.926	3.07	0.301	0.768
edad_hogar	-0.0775	0.0641	-1.21	0.246
cant_chimen	2.77	1.81	1.53	0.149

Los resultados mostrados en la tabla 5, donde se muestran las estimaciones del modelo original y sus $t-values$, muestran un estimado del intercepto (`precioventa`) un valor positivo > 0 (20.2); de manera que de estos resultados se puede deducir que algunos estimadores de las variables regresoras, pueden afectar el resultado de la variable respuesta. Al verificar los $t-values$ se observa que solo son significativas `num_banos` ($p-value < 0.05$). Las demás variables regresoras no son significativas ($p-value \geq 0.05$) para establecer su efecto sobre la variable respuesta en el modelo original. Con este modelo en una interpretación prudente solo es posible afirmar que un incremento en una unidad del numero de baños se puede incrementar 5.23/1000 dolares el precio de venta de la propiedad, siempre que las otras variables no cambien.

Pregunta #4: Ajuste un nuevo modelo lineal que contenga sólo las variables regresoras que resultaron significativas en la prueba anterior.

Dado los resultados anteriores se decide realizar un ajuste del modelo aplicando el método *stepwise* y aplicando los estadisticos AIC, BIC, Cp y R^2_{adj}.

Le método AIC (Akaike) fue realizado a través de la función `step`, con un método *stepwise* bidireccional *forward* y *reverse*, a partir de esta estratégia se obtuvo que el **"mejor modelo"** sería uno que incluyera únicamente 5 variables regresoras, a saber: `impuestos`, `num_banos`, `num_puestos_gar`, `edad_hogar`, `cant_chimen`.

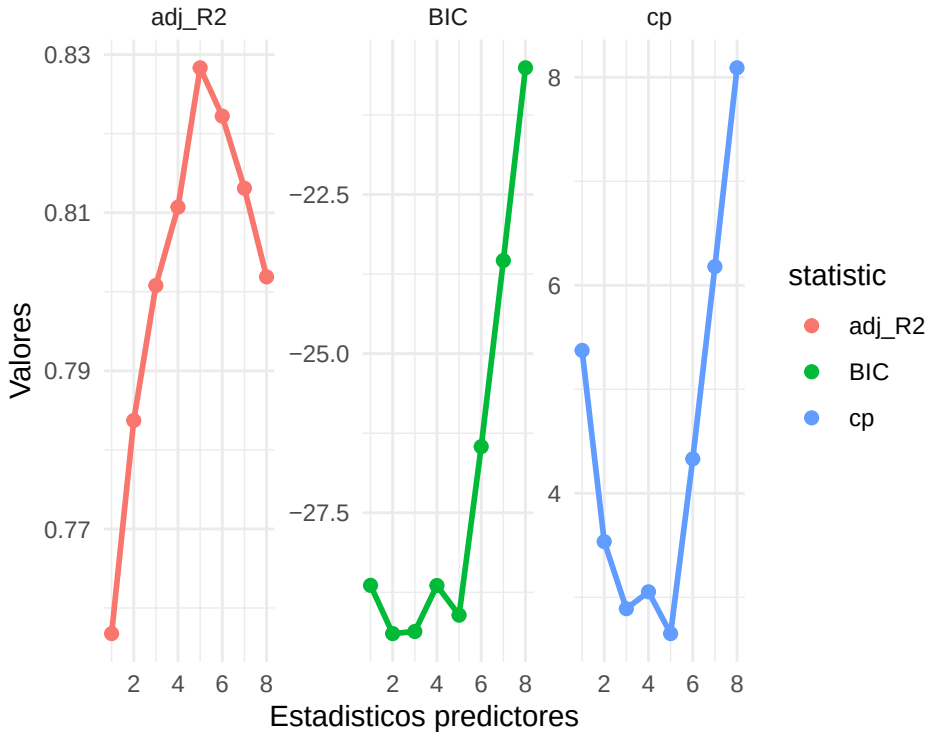

Figure 3: Estadísticos para elección de mejor modelo: adj_R2, BIC, Cp

Además se realizó un análisis utilizando otros estadisticos bajo el método *stepwise* para este caso se utilizó unicamente un direccionamiento *forward* o de introducción progresiva, con estas precisiones los resultados fueron similares al AIC como se puede ver en la figura 3. Es de notar que Para el método R^2_{adj} y Cp muestran que el mejor modelo es el #5 que corresponde al mismo modelo obtenido por AIC. No obstante, estadistico BIC muestra que el mejor modelo es el #2 esto es: `impuestos` y `num_banos`. Para efectos del desarrollo de esta actividad se tomará el modelo de 5 (que será designado como "*mejor modelo*") variables por la consistencia de resultados en los demás estadísticos.

Pregunta #5: Interprete los parámetros estimados del nuevo modelo.

Table 6: Estimadores, $t-values$ y significancia para el mejor modelo sugerido por AIC, Cp, y r^2_{adj}

	Estimado	ErrorStand	$t-value$	$p-value$
(Intercept)	19.7	3.75	5.24	5.54e-05
impuestos	1.35	0.675	1.99	0.0617
num_banos	5.7	1.82	3.13	0.00574
num_puestos_gar	0.575	0.25	2.3	0.0338
edad_hogar	-0.0787	0.0458	-1.72	0.103
cant_chimen	2.54	1.28	1.98	0.0628

La tabla 6 muestra las estimaciones del "mejor modelo" y sus $t-values$, muestran un estimado del intercepto (`precioventa`) con valor positivo > 0 (19.7); de manera que de estos resultados se puede deducir que algunos estimadores de las variables regresoras, pueden afectar el resultado de la variable respuesta. Al verificar los

$t-values$ se observa que solo son significativas nuevamente `num_banos` y `num_puestos_gar` ($p-value < 0.05$). Las demás variables regresoras utilizadas en este modelo no son significativas ($p-value \geq 0.05$), por lo tanto, no es posible establecer que puedan tener efecto alguno sobre la variable respuesta. De acá se se puede inferir que un incremento en una unidad del numero de baños se puede incrementar 5.7/1000 dolares el precio de venta de la propiedad siempre que las otras variables no cambien y/o que un incremento en una unidad del numero de puestos de garaje se puede incrementar 0.575/1000 dolares el precio de venta de la propiedad, de igual manera siempre que las otras variables no cambien.

Table 7: Estadistico F de Fisher: significancia del mejor modelo
Al probar la significancia del "mejor modelo" se estableción nuevamente una prueba de hipotesis de la siguiente manera:

F_{stat}	Grados_libertad	$p - value$
23.1964754219616	18	2.899e-07

$$H_o : \beta_1 = \beta_2 = ... = \beta_k = 0$$

$$H_1 : \beta_j \neq 0 \ para \ cualquier \ j \ dado$$

la tabla 7 muestra los resultaos para el estadístico F de Fisher con 18 grados de libertad, donde el $p-value$ demuestra significancia estadística ($p - value < 0.05$), con lo que se rechaza la hipotesis nula. De tal forma, se puede inferir que almenos una de la variables regresoras (β_i) son diferentes de 0.

Table 8: R^2 y R^2_{adj} para el modelo

R^2	R^2_{adj}
0.866	0.828

De otra parte, la tabla 8, muestra los resultados de los R^2 y R^2_{adj} para el "modelo de regresión original"mejor modelo". Estos resultados son más consistentes que el modelo original ($R^2 \approx R^2_{adj}$) e incluso como era de esperarse por el resultado del estadistico para selección de mejor modelo R^2_{adj}, su resultado es mayor con respecto al resultado del modelo orginal ($R^2_{adj}\ mejor\ modelo = 0.828 > R^2_{adj}\ modelo\ original = 0.789$). De esto se puede inferir que aproximadamente el 83% de la variabilidad de la variable respuesta (`precioventa`) puede ser explicado por las variables regresoras del"mejor modelo" y la variabilidad restante puede ser consecuencia a otras variables no tenidas encuenta en el modelo o por azar.

Pregunta 6: Construya e interprete un gráfíco de residuales contra la respuesta estimada. Es sensato pensar que el modelo cumple los supuestos? Verifíque esto a un nivel de signifícancia del 5%

Para validar el "mejor modelo" de regresión, se realizó un análisis de los residuales para dicho modelo. Para realizar esto se tomaron únicamente los residuales estudentizados Externamente (para efectos prácticos solo se designaran como *residuales estudentizados*).

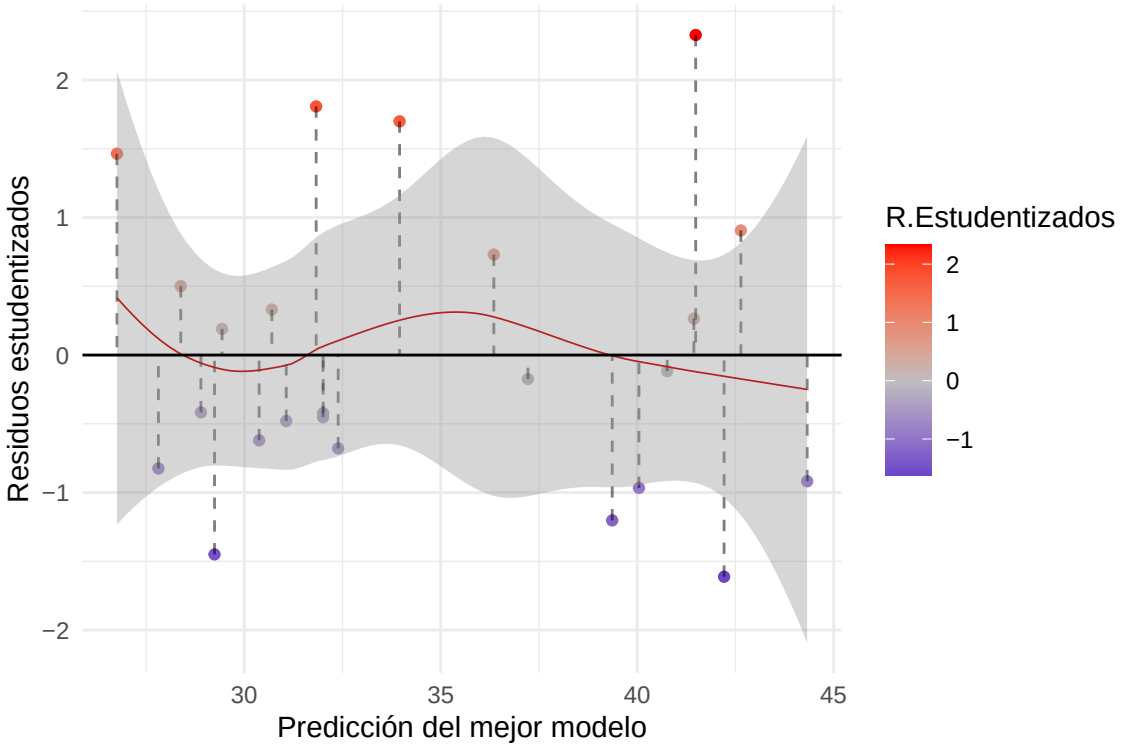

Figure 4: Distribución de los residuales estudentizados

la validación de los supuestos para la regresión, a saber:

- Relación entre residuales estudentiados obtenidos y los predichos o esperados: $\epsilon_i = 0$.
- La distribución de los residuos esperados se distribuyen normalemnte: $\epsilon_i \approx N(\mu = 0, var = \sigma^2)$.
- Homocedasticidad: $var(\epsilon_i) = \sigma^2$.
- No hay autocorrelación: $corr(\epsilon_i, \epsilon_j) = 0$.

Inicialmente se evaluó la premisa $\epsilon_i = 0$. La figura 4, muestra que hay cierta variabilidad entre las relaciones, dadas por algunas observaciones extremas que se alejan de 0.

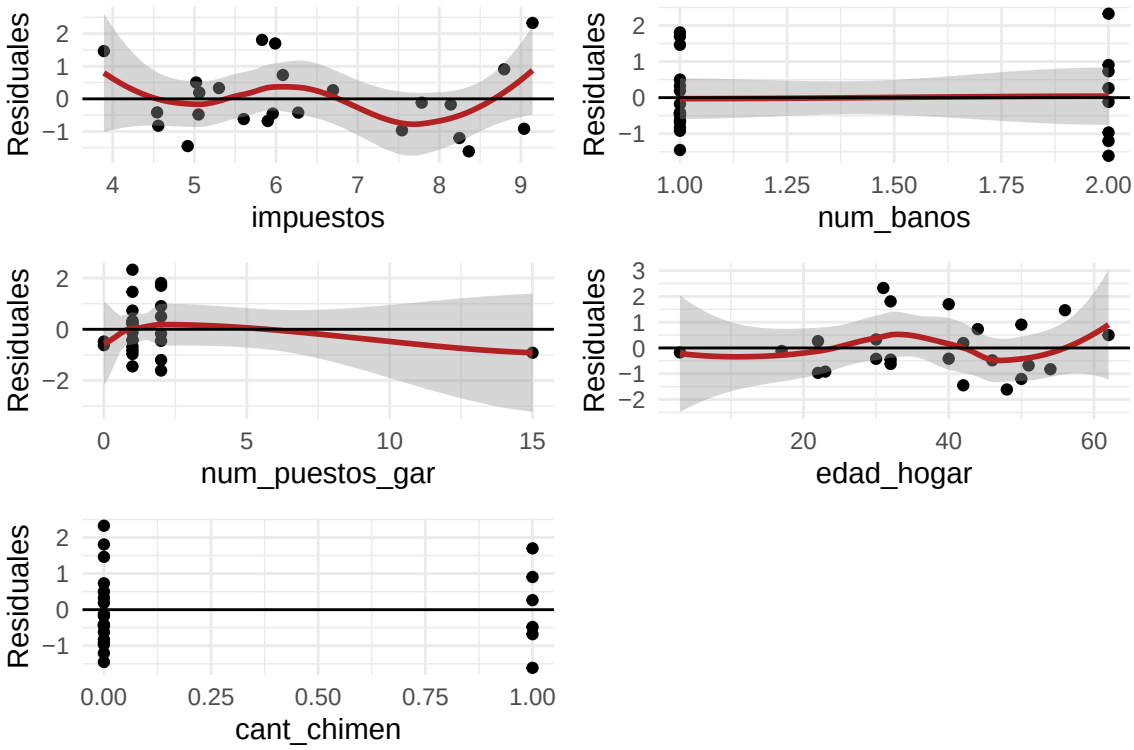

Figure 5: Distribuciones de residuales estudentizados para cada variable del mejor modelo

Además se realizó se realizó una discriminación por variable dada (figura 5), evaluando las relaciones entre los residuales estudentizados y los esperados encontrando que la variable `cant_chimenea`es la que menos se encuentra relación y `num_banos`es la donde la relación es mayor (≈ 0). Las variables con mayor variabilidad en su relación son `impuestos`y `edad_hogar`.

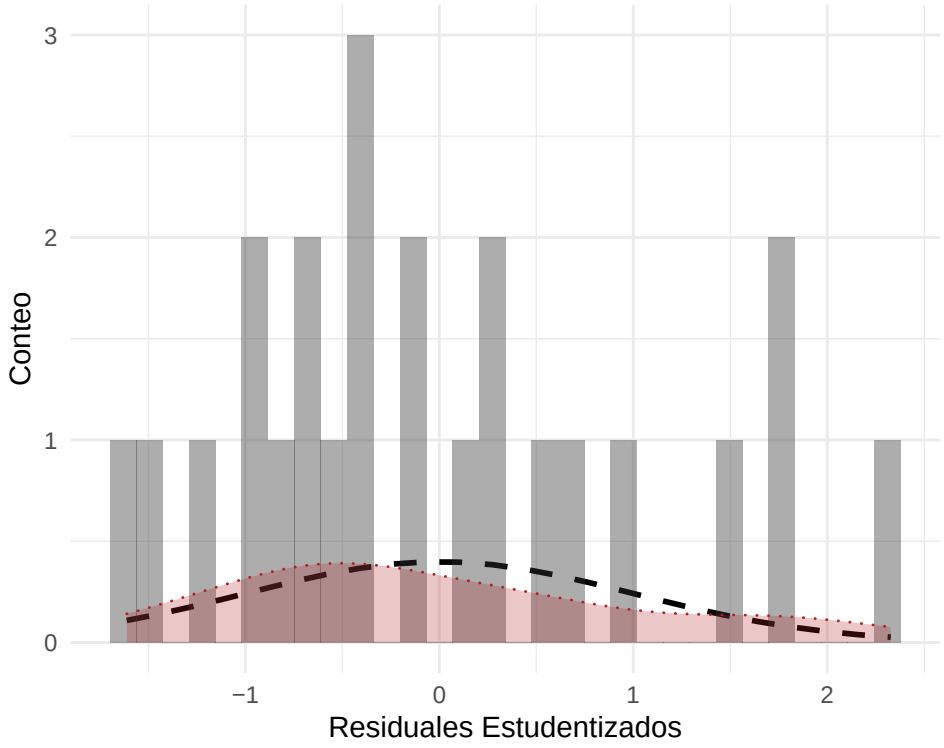

Figure 6: Histograma de distribución de los residuales

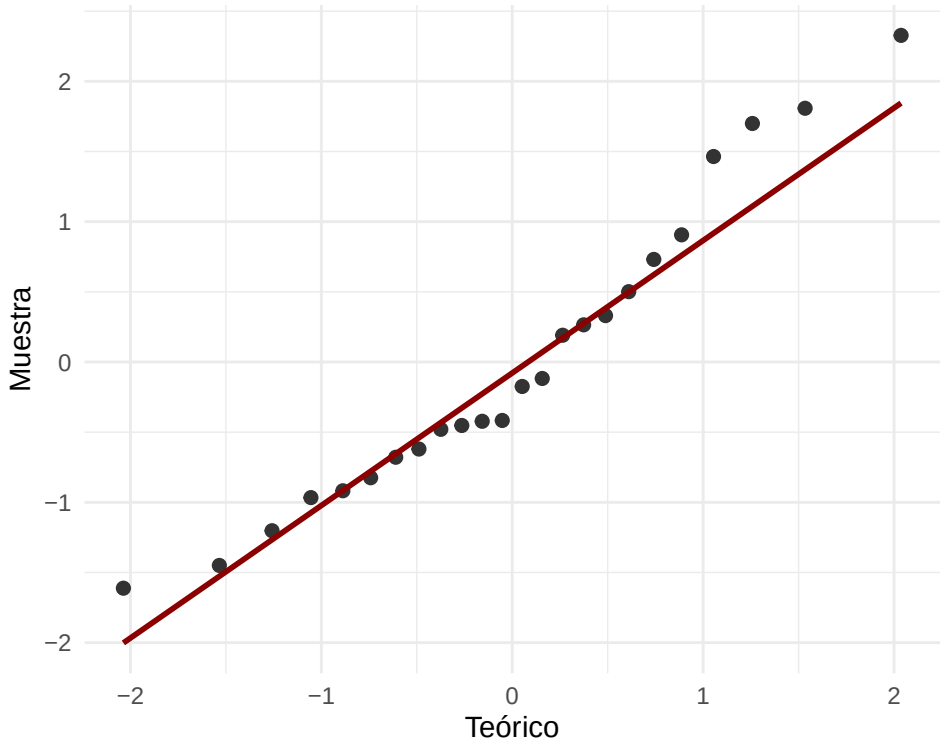

Figure 7: Q-Q Plot: residuales estudentizados

Seguidamente se evaluó la premisa: $\epsilon_i \approx N(\mu = 0, var = \sigma^2)$, incialmente a través de un gráfico de distribución de los residuales estudentizados (figura 6 y 7 respectivamente). En la figura 6 se muestra que a distribución de los residuales estudentizados (linea y sombreado rojo) es levemente asimétrica con una desviación hacia la derecha con respecto a una curva de normalidad (linea negra entrecortada) que cumple con los parametros de la premisa a validar. Por su parte, la figura 7 muestra que las observaciones finales se alejan de la linea central (roja) sugiriendo que estas observaciones son las que pueden estár moviendo la distribución.

Table 9: Test de shapiro-wilk (w) para los residuales

W	$p - value$
0.9505	0.2772

Para comprar la normalidad de los residuales estudentizados se decidió comprobar a través de un test de Shapiro-Wilk, el cual permite constatar la normalidad del conjunto de observaciones, y se formula de la siguiente manera (donde N indica una distribución normal:

$$H_o : el\ conjunto\ de\ datos = N$$

$$H_1 : el\ conjunto\ de\ datos \neq N$$

El resultado obtenido de este estadístico es presentado en la tabla 10, mostrando que su resultado No tiene significancia ($p - value < 0.05$). Por esta razón, no es posible rechazar la hipotesis nula, así que se puede aseverar que los residuales estudentizados siguen una distribución normal (N), por lo cual se valida $\epsilon_i \approx N(\mu = 0, var = \sigma^2)$.

df 5

Table 10: Test de Breusch-Pagan (BP) para los residuales

BP	Grados_libertad	$p - value$
3.176	5	0.6728

Para evaluar si existe homocedasticidad ($var(\epsilon_i) = \sigma^2$) en los residuales estandarizados, se recurrió a realizar un test de Breusch-Pagan, el cual establece si la varianza estimada dependen de las observaciones de manera independiente, se supone homocedasticidad en cuando no hay esta dependecia, y heterocedasticidad en cuanto si existe esta dependencia. Esto se establece de la siguiente forma:

$$H_o : hay\ homocedasticidad$$

$$H_1 : hay\ heterocedasticidad$$

El resultado obtenido del test de Breusch-Pagan es presentado en la tabla 11, mostrando que su resultado No tiene significancia ($p-value < 0.05$). Por esta razón, no es posible rechazar la hipotesis nula, así que se puede aseverar que los hay homocedasticidad, es decir la varianza no depende de las observaciones independientes sino deel modelo en su totalidad, por lo cual se valida ($var(\epsilon_i) = \sigma^2$).

Table 11: Test de Durbin–Watson (DW) para los residuales

DW	$p - value$
2.097	0.5757

Finalmente, Para evaluar si existe autocorrelación $corr(\epsilon_i, \epsilon_j) = 0$ entre los residuales estandarizados, se recurrió a realizar un test de Durbin–Watson, el cual establece que si la autocorrelación designada con el simbolo ρ es igual a 0 ($\rho = 0$) entonces las observaciones no están autocorrelacionados, en cambio si existe autocorrelación ($\rho \neq 0$) entonces algunas de las observaciones están correlacionadas . Esto se establece de la siguiente forma:

$$H_o : \rho = 0$$

$$H_1 : \rho \neq 0$$

La tabla 12 muestra los resultados del test de Durbin–Watson. Este muestra que No existe significancia ($p - value < 0.05$). Por esta razón, no es posible rechazar la hipotesis nula, así que se puede aseverar que los residuales estudentizados no se encuentran correlacionados entre ellos.

Pregunta #7: Realice un análisis de diagnóstico e identifíue (si es que existen) observaciones con alto leverage, incluyentes y extremas en la respuesta.

La figura 8, 9 y 10 corresponden al análisis de los residuales estudentizados para evaluar el efecto de cada una de las observaciones en el mejor modelo de regresión encontrado por los métodos y estadísticos previamente descritos.

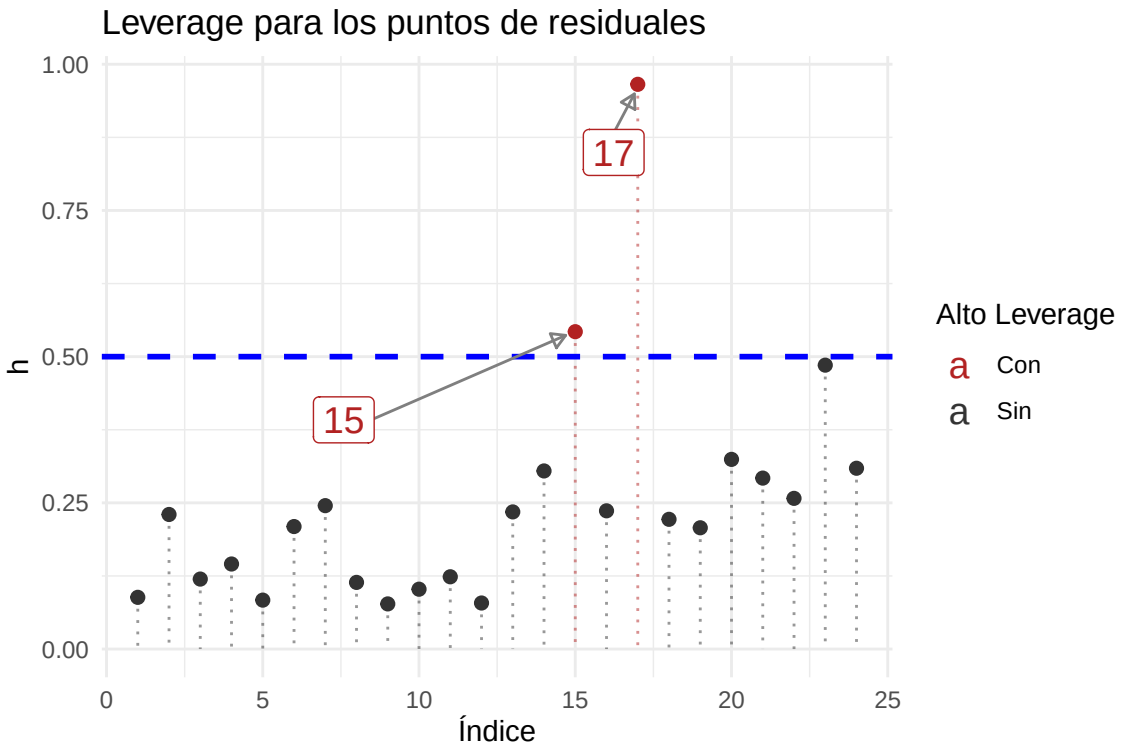

Figure 8: Evaluación de altos *leverage* para el mejor modelo

La figura 8, en el eje x muesta el número de la observación o (índice) y la y los valores correspondientes la diagonal de la matriz H obtenida de los residuales estudentizados y predichos. la linea azul traza el punto donde los valores de la diagonal en H son dos veces la media de estos valores. Como se puede evidenciar la observación número 15 y 17 son superiores a dos veces la media de los valores en H, de manera que se consideran valores con alto *leverage* (o apalancamiento), es decir que estas observaciones pueden afectar la regresión.

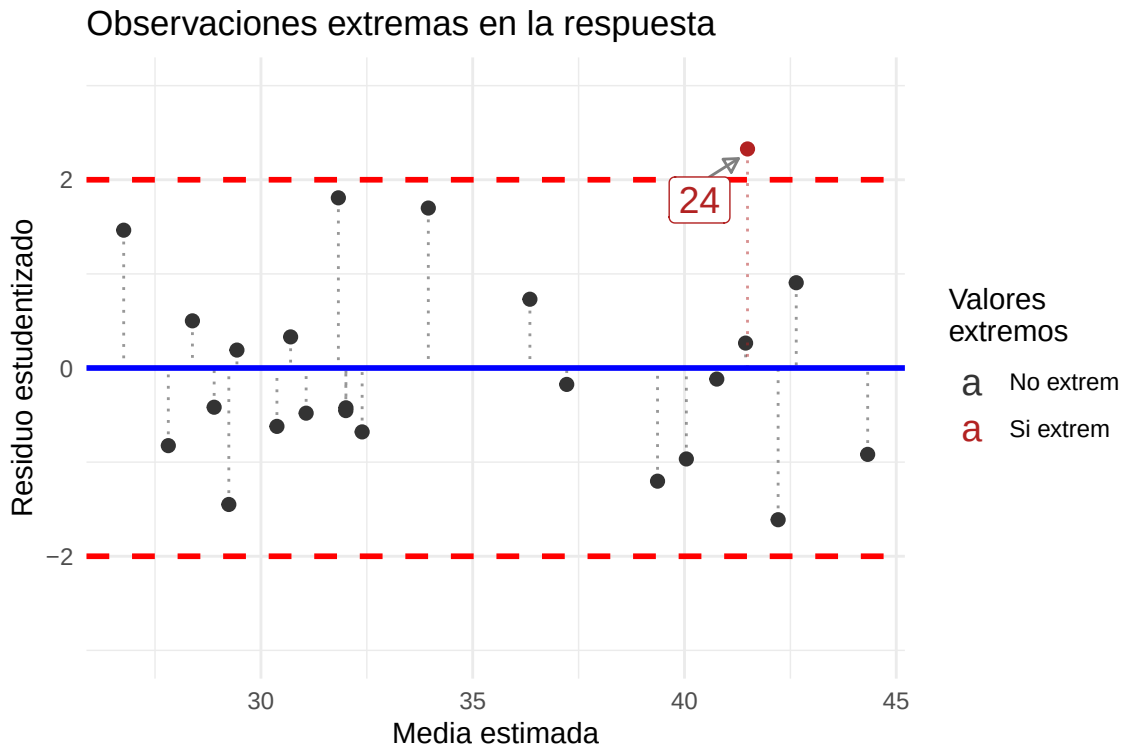

Figure 9: Evaluación de observaciones extremas para el mejor modelo

La figura 9, en el eje x muestra a la media estimada de los residuales estimados por el modelo, y y los valores correspondientes a los valores que toman los residuales estudentizados. la linea azul traza el punto 0 y las lineas rojas entre cortadas, las desviaciones estandar de los residuales estudentizados (2 y -2 sd). Como se puede evidenciar la observación número 24 esta fuera de dos deviaciones estandar por lo cual de considera una observación extrema y se debe evaluar si es influyente.

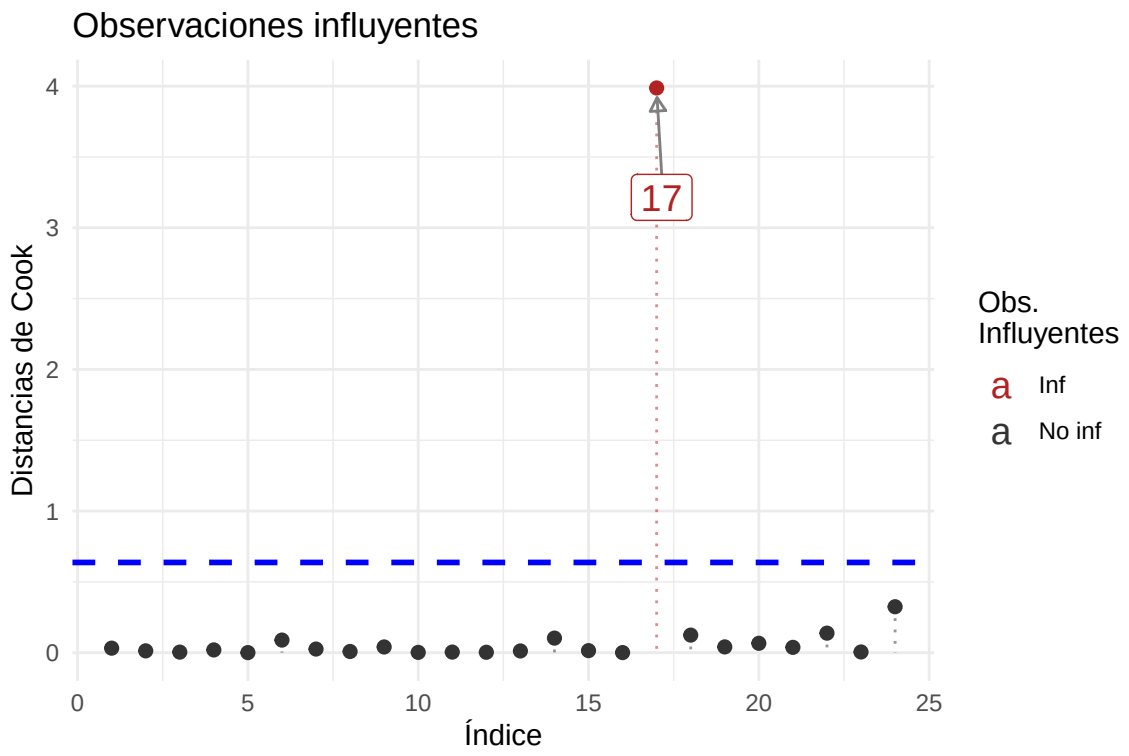

Figure 10: Evaluación de observaciones influyentes para el mejor modelo

Finalmente la figura 10, en el eje x muesta el número de la observación o (índice), y y los valores correspondientes a las distancias de Cook, que evaluan la influencia de las observaciones en el modelo. la linea azul traza el valor que corresponde la desviación 3 veces de promedio de los residuales estudentizados. Como se puede evidenciar la observación número 17 es la observación más influyente en el modelo además que ejerce *leverage* de manera que sería conveniente sustraer esta observación y evaluar nuevamente su efecto en el modelo.

Evaluación de mejor modelo sin la observación influyente

Análisis de estimadores y $t - values$ mejor modelo excluyendo observación influyente #17

La tabla 12 presenta las estimaciones del "mejor modelo" excluyendo observación influyente y sus $t - values$. Allí se evidencia un estimado del intercepto (`precioventa`) con valor positivo > 0 (20.2); de manera que de estos resultados se puede deducir que algunos estimadores de las variables regresoras, pueden afectar el resultado de la variable respuesta. Al verificar los $t - values$ se observa que solo son significativas nuevamente `num_banos` al igual que el modelo original ($p - value < 0.05$). Las demás variables regresoras utilizadas en este modelo no son significativas ($p - value \geq 0.05$), así pues, no es posible establecer que puedan tener efecto alguno sobre la variable respuesta. De acá se se puede inferir que un incremento en una unidad del numero de baños se puede incrementar 6.08/1000 dolares el precio de venta de la propiedad siempre que las otras variables no cambien.

Table 12: Estimadores, $t - values$ y significancia para el mejor modelo sugerido por AIC, Cp, y r^2_{adj}

	Estimado	ErrorStand	$t - value$	$p - value$
(Intercept)	20.2	3.82	5.3	5.92e-05
impuestos	1.07	0.74	1.45	0.165
num_banos	6.08	1.87	3.25	0.00476
num_puestos_gar	1.49	1.03	1.45	0.166
edad_hogar	-0.0926	0.0484	-1.91	0.0729
cant_chimen	2.62	1.29	2.03	0.0578

La tabla 13 muestra los resultados para el estadístico F de Fisher con 17 grados de libertad, donde el $p-value$ demuestra significancia estadística ($p - value < 0.05$), con lo que se rechaza la hipotesis nula. De tal forma, se puede inferir que almenos una de la variables regresoras (β_i) son diferentes de 0.

La tabla 14, muestra los resultados de los R^2 y R^2_{adj} para el "modelo de regresión original"mejor modelo" sin observación influyente. De esto se puede inferir que aproximadamente el 81% de la variabilidad de la variable respuesta (`precioventa`) puede ser explicado por las variables regresoras del "mejor modelo" y la variabilidad restante puede ser consecuencia a otras variables no tenidas encuenta en el modelo o por azar.

Table 13: Estadistico F de Fisher: significancia del mejor modelo

F_{stat}	Grados_libertad	$p - value$
20.2777398941543	17	1.275e-06

Table 14: R^2 y R^2_{adj} para el modelo

0.856	0.814

Análisis de residuales del mejor modelo excluyendo observación influyente #17

La figura 11, 12 y 13 corresponden al análisis de los residuales estudentizados para evaluar el efecto de cada una de las observaciones en el mejor modelo excluyendo la observación influyente del "mejor modelo". A partir de esto es de notar que la observación 23 aparece como extrema e influyente pero sin alto *leverage*. En su lugar, presentan alto *leverage* la observación 15 y 22. De estas la observación 15 previamente había presentado alto *leverage* en el "mejor modelo", sin embargo, aparece en este modelo como no influyente.

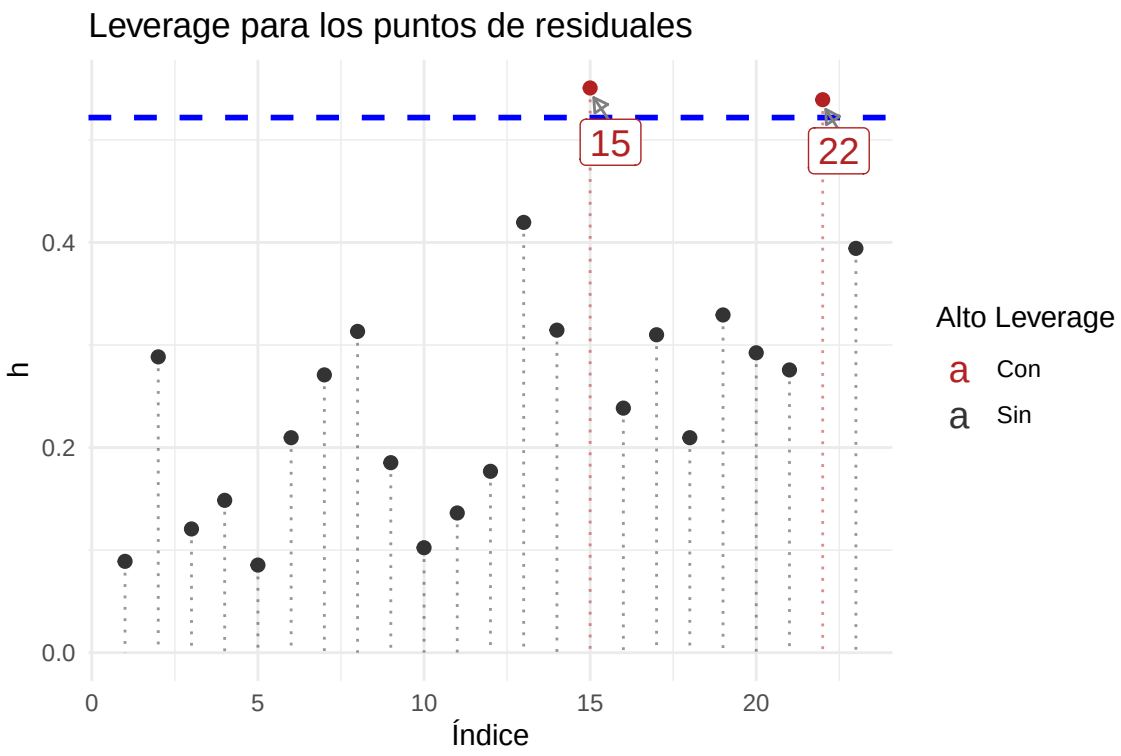

Figure 11: Evaluación de altos *leverage* para el mejor modelo sin observación influyente

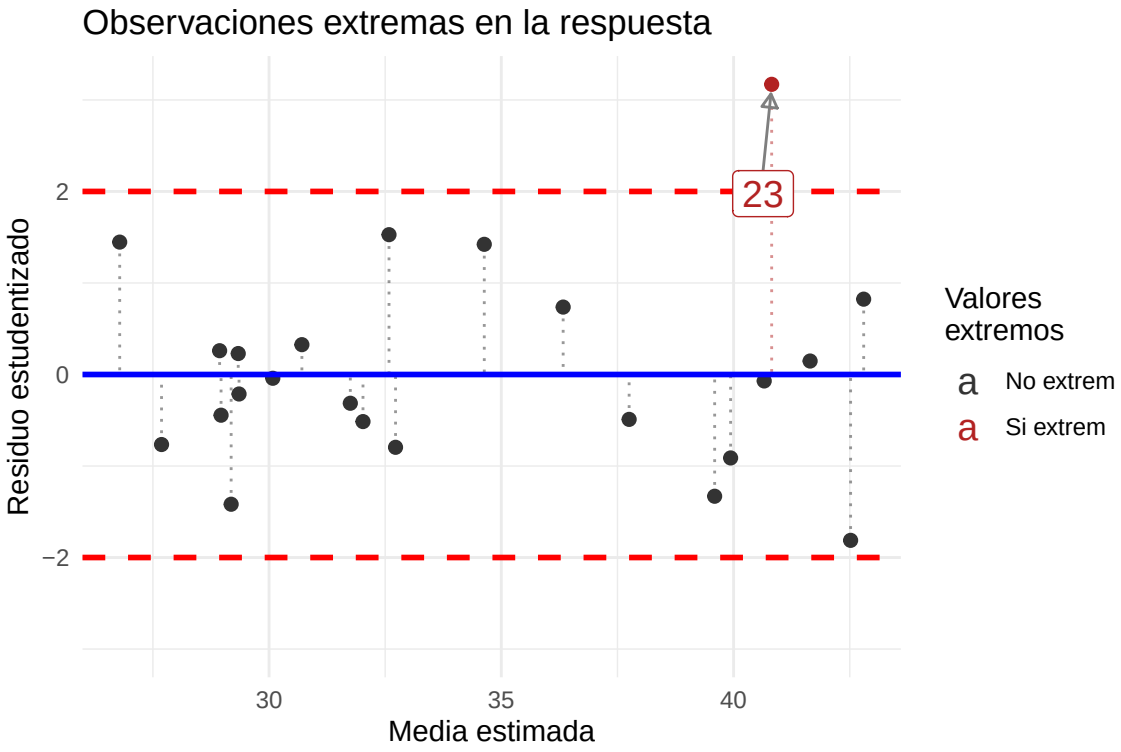

Figure 12: Evaluación de observaciones extremas para el mejor modelo sin observación influyente

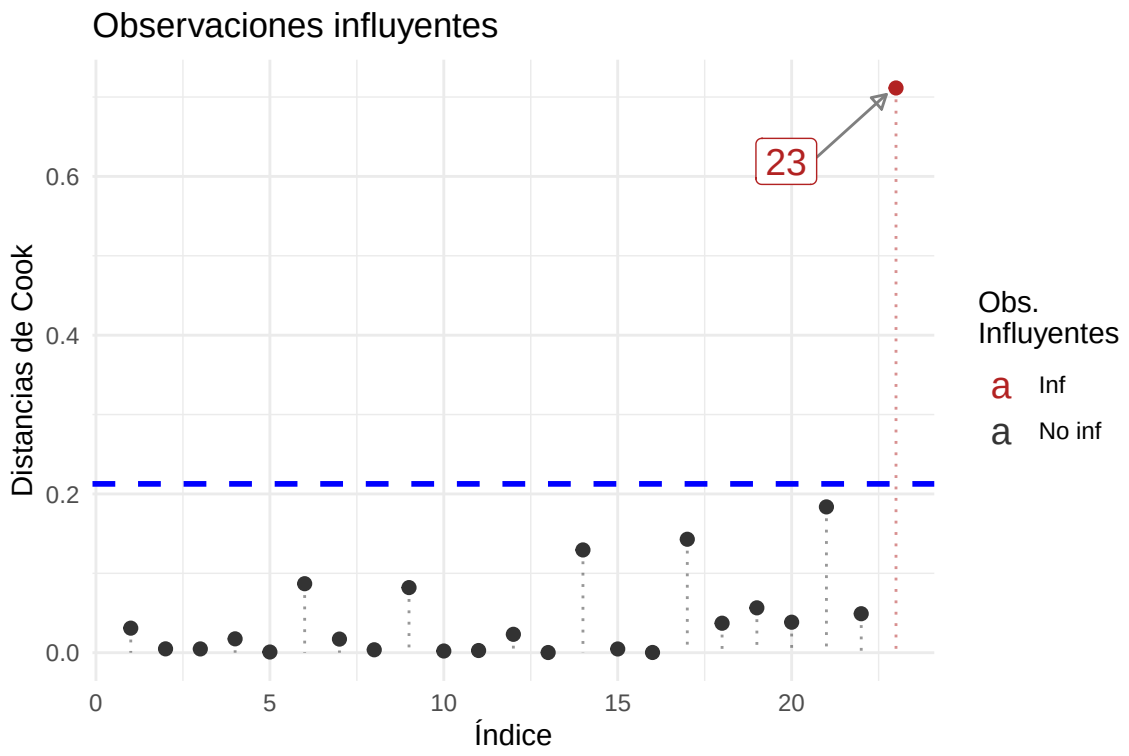

Figure 13: Evaluación de observaciones influyentes para el mejor modelo sin observación influyente

Anexo (mejor modelo por BIC)

Se realizó mejor modelo por BIC con dos variables regresoras `impuestos` y `num_banos`, con todas las observaciones.

EStimadores y $t-values$ para el mejor modelo por BIC

La tabla 15 presenta las estimaciones del "mejor modelo por BIC" y sus $t-values$. Allí se evidencia un estimado del intercepto (`precioventa`) con valor positivo > 0 (13.1); de manera que de estos resultados se puede deducir que algunos estimadores de las variables regresoras, pueden afectar el resultado de la variable respuesta. Al verificar los $t-values$ se observa que solo es significativa la variable `impuestos`. En este caso la variable `num_banos`que había sido significativa en los otros dos modelos es no son significativas ($p-value \geq 0.05$) en este. Por esto, se puede establecer que `impuestos` puedan tener efecto alguno sobre la variable respuesta. De acá se se puede inferir que un incremento en una unidad de impuestos se puede incrementar 2.71/1000 dolares el precio de venta de la propiedad siempre que las otras variables no cambien.

Table 15: Estimadores, $t-values$ y significancia para el mejor modelo sugerido por BIC$

	Estimado	ErrorStand	$t-value$	$p-value$
(Intercept)	13.1	2.43	5.41	2.31e-05
impuestos	2.71	0.485	5.6	1.49e-05
num_banos	3.08	1.59	1.93	0.0666

La tabla 16 muestra los resultados para el estadístico F de Fisher con 21 grados de libertad, donde el $p-value$ demuestra significancia estadística ($p - value < 0.05$), con lo que se rechaza la hipotesis nula. De tal forma, se puede inferir que almenos una de la variables regresoras (β_i) son diferentes de 0.

La tabla 17, muestra los resultados de los R^2 y R^2_{adj} para el "modelo de regresión original"mejor modelo por BIC". De esto se puede inferir que aproximadamente el 78% de la variabilidad de la variable respuesta (`precioventa`) puede ser explicado por las variables regresoras del"mejor modelo por BIC" y la variabilidad restante puede ser consecuencia a otras variables no tenidas encuenta en el modelo o por azar. Esto es esperado por que el resultado del estadístico R^2_{adj} para selección de modelo daba como mejor el seleccionado como "mejor modelo" o sea el modelo de 5 variables.

Table 16: Estadistico F de Fisher: significancia del mejor modelo por BIC

F_{stat}	Grados_libertad	$p - value$
42.6719279014923	21	4.007e-08

Table 17: R^2 y R^2_{adj} para el mejor modelo por BIC

R^2	R^2_{adj}
0.803	0.784

Validación del mejor modelo BIC

Para determinar la validez del mejor modelo obtenido por BIC se evaluaron los mismos parametros que el modelo original.

Inicialmente se evaluó la premisa $\epsilon_i = 0$. La figura 14, muestra que hay cierta variabilidad entre las relaciones, dadas por algunas observaciones extremas que se alejan de 0 al igual que el "mejor modelo" obtenido por otros estadísticos.

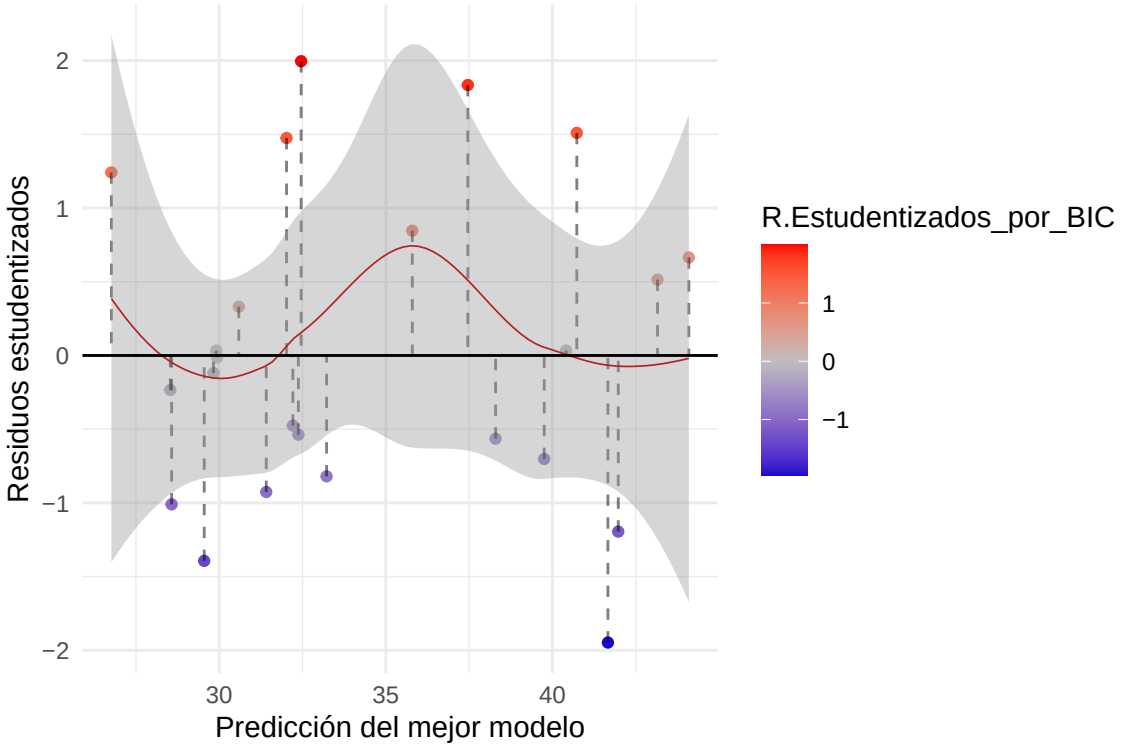

Figure 14: Distribución de los residuales estudentizados del modelo por BIC

A continuación se evaluó la premisa: $\epsilon_i \approx N(\mu = 0, var = \sigma^2)$, incialmente a través de un gráfico de distribución de los residuales estudentizados (figura 15), este muestra que a distribución de los residuales estudentizados para este modelo (linea y sombreado rojo) continua levemente asimétrica con una desviación hacia la derecha con respecto a una curva de normalidad (linea negra entrecortada) que cumple con los parametros de la premisa a validar, caso similar a lo ocurrido en el "mejor modelo" de 5 variables.

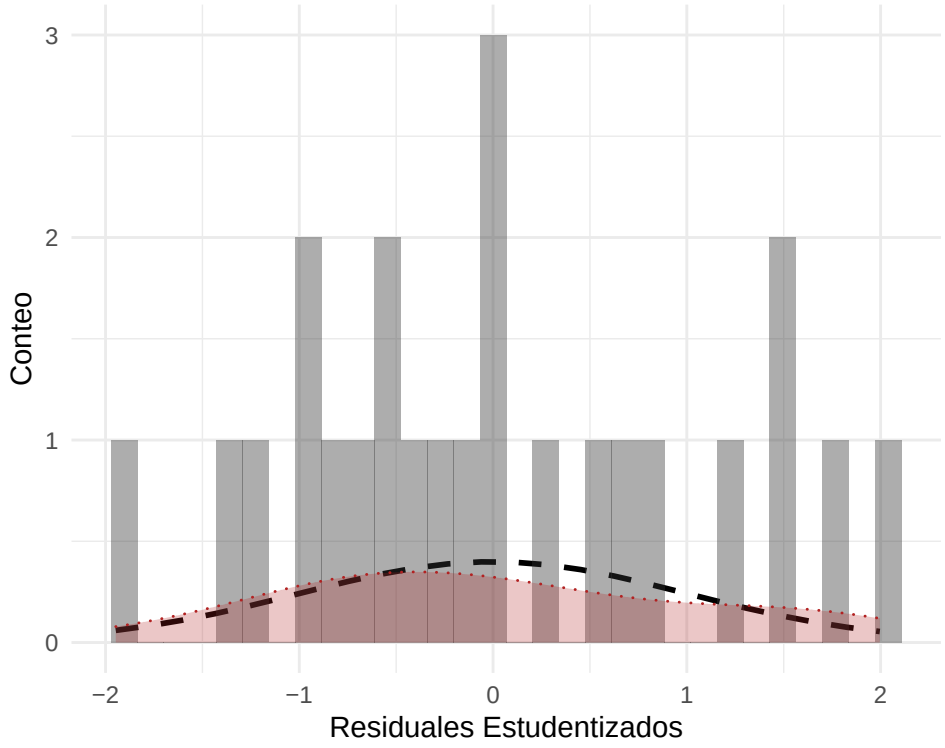

Figure 15: Histograma de distribución de los residuales del modelo por BIC

Table 18: Test de shapiro-wilk (w) para los residuales

W	$p-value$
0.9709	0.6897

La tabla 18, muestra el resultado del test de shapiro-Wilk cuyo resultado No tiene significancia ($p-value < 0.05$). Por esta razón, no es posible rechazar la hipotesis nula, así que se puede aseverar que los residuales estudentizados para este modelo obtenido por BIC siguen una distribución normal (N), por lo cual se valida $\epsilon_i \approx N(\mu = 0, var = \sigma^2)$.

Table 19: Test de Breusch-Pagan (BP) para los residuales

BP	Grados_libertad	$p-value$
0.3459	2	0.8412

La tabla 19, muestra el resultado del test de Breusch-pagan cuyo resultado No tiene significancia ($p-value < 0.05$). Por esta razón, no es posible rechazar la hipotesis nula, así que se puede aseverar que los hay homocedasticidad, es decir la varianza no depende de las observaciones independientes sino deel modelo en su totalidad, por lo cual se valida ($var(\epsilon_i) = \sigma^2$).

Table 20: Test de Durbin–Watson (DW) para los residuales

DW	$p-value$
2.161	0.5897

La tabla 20, muestra el resultado del test de Durbin–Watson cuyo resultado No tiene significancia ($p-value < 0.05$).Por esta razón, no es posible rechazar la hipotesis nula, así que se puede aseverar que los residuales estudentizados para el modelo obtenido por no se encuentran correlacionados entre ellos.

Análisis de residuales del mejor modelor BIC

La figura 16, 17 y 18 corresponden al análisis de los residuales estudentizados para evaluar el efecto de cada una de las observaciones en el mejor modelo producido por el estadístico BIC. A partir de estos resultados, es importante notar que la observación 17 al igual que el "mejor modelo" de 5 variables, presenta alto *leverage* y es influyente. Además llama la atención que la observación 15 previamente había presentado alto *leverage* en el "mejor modelo" aparece en este modelo como una observación influyente. Además la observación # 15 presenta alto *leverage* en el "mejor modelo" sin observación influyente (figura 12) de manera que es importante proponer un modelo que excluya esta observación. Para este nuevo modelo no se obtienen observaciones etremas.

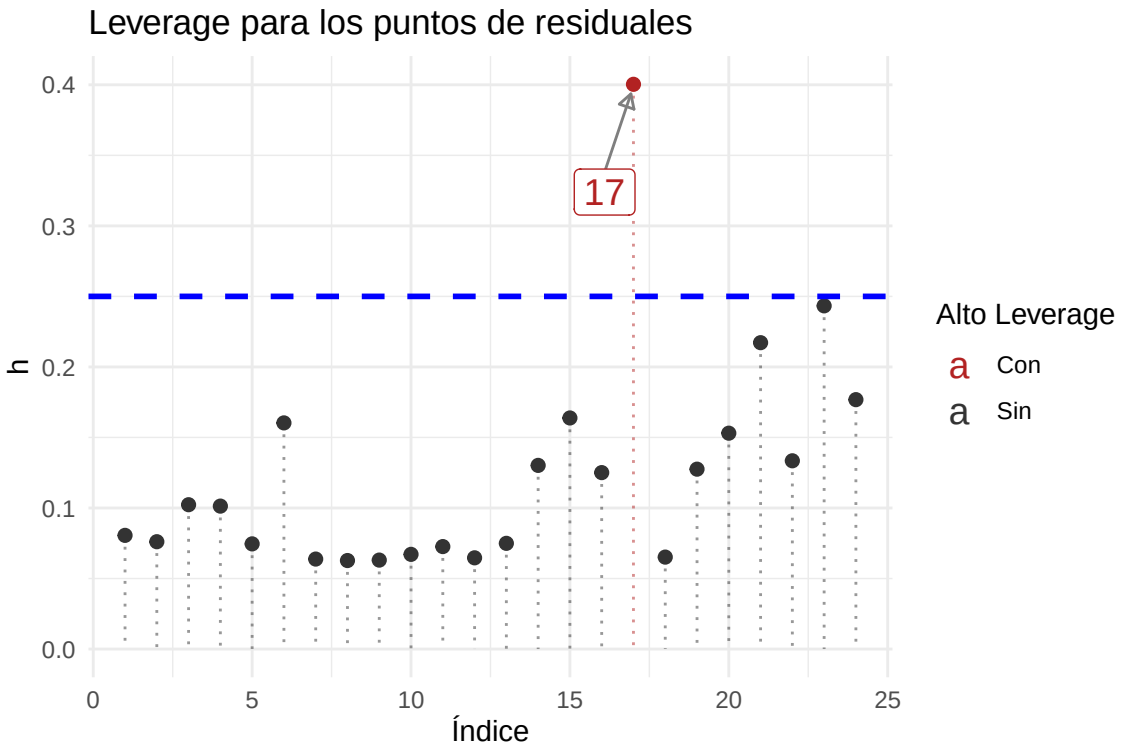

Figure 16: Evaluación de altos *leverage* para el mejor modelo por BIC

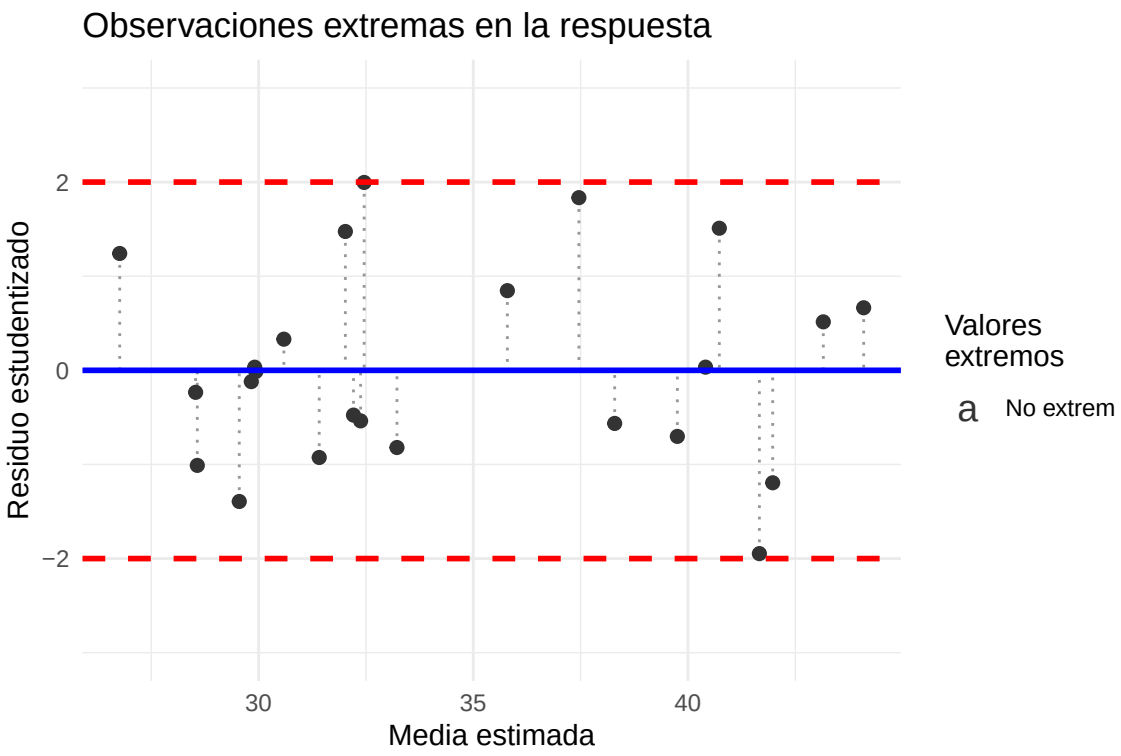

Figure 17: Evaluación de observaciones extremas para el mejor modelo por BIC

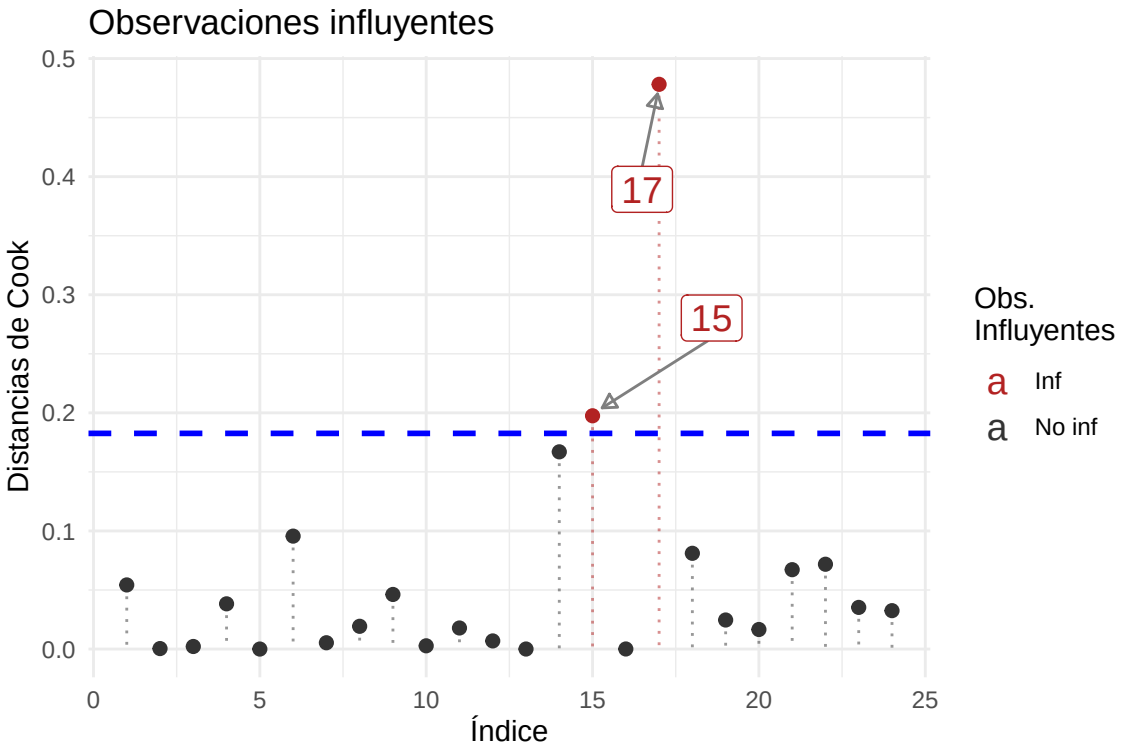

Figure 18: Evaluación de observaciones influyentes para el mejor modelo por BIC