Modelos Probabilísticos y Análisis Estadístico
Modelos de Regresión Lineal

Jesid Mauricio Mejía Castro Christiam Alejandro Peña Niño

18 de mayo de 2021

1. Exploración del conjunto de datos

Durante la preparación de los datos se excluyeron las columnas `id` y `zipcode`, pues estas no tendrán ningún efecto en el precio final del inmueble. Por otro lado, se tratarán las variables `waterfron`, `view`, `condition` y `grade` como categóricas.

Adicionalmente, a partir de los Cuadros 1 y 2, puede observarse que las columnas `lat` y `long` presentan muy poca variabilidad. Esto muy probablemente se debe a que la muestra de viviendas es extraída de un sólo condado (King County). El percentil 75 en la variable `yt_renovated` (véase el Cuadro 2) indica que la mayoría de datos es 0 (de hecho, solo el 4 % de los registros tiene este dato). Estas variables se omitirán en el análisis.

Estadística	Mean	Min	Max
price	539,631.800	75,000	7,062,500
bedrooms	3.371	0	33
bathrooms	2.116	0	8
sqft_living	2,082.051	290	10,040
sqft_lot	15,189.620	520	1,651,359
floors	1.493	1	4
sqft_above	1,790.303	290.000	8,860.000
sqft_basement	291.743	0	4,820
yr_built	1,971.054	1,900	2,015
yr_renovated	84.045	0	2,015
lat	47.560	47.156	47.778
long	−122.213	−122.519	−121.315
sqft_living15	1,986.514	399	6,210
sqft_lot15	12,742.440	651	871,200

Cuadro 1: Media, mínimo y máximo para las variables numéricas del conjunto de datos.

Estadística	St. Dev.	Pctl(25)	Pctl(75)
price	362,762.500	322,500	645,000
bedrooms	0.933	3	4
bathrooms	0.766	1.8	2.5
sqft_living	913.916	1,422.8	2,550
sqft_lot	41,885.530	5,071.2	10,713.8
floors	0.539	1	2
sqft_above	825.332	1,200.000	2,216.000
sqft_basement	441.880	0	560
yr_built	29.326	1,952	1,997
yr_renovated	400.894	0	0
lat	0.138	47.471	47.678
long	0.141	−122.327	−122.124
sqft_living15	686.897	1,490	2,360
sqft_lot15	27,371.650	5,100	10,094.8

Cuadro 2: Desviación estándar, percentil 25 y percentil 75 para las variables numéricas del conjunto de datos.

Statistic	Min	Median	Max	Pctl(25)	Pctl(75)
waterfront	0	0	1	0	0
view	0	0	4	0	0
condition	1	3	5	3	4
grade	1	7	13	7	8

Cuadro 3: Estadísticas básicas para las variables categóricas del conjunto de datos.

2. Modelo inicial

Sea y la variable `price`. Las variables independientes se definirán de la siguiente manera:

- x_1: `bedrooms`,
- x_2: `bathhroms`,
- x_3: `sqft_living`,
- x_4: `sqft_lot`,
- x_5: `floors`,
- x_6: `sqft_above`,
- x_7: `yr_built`,
- x_8: `sqft_living15`,
- x_9: `sqft_lot15`,
- x_{10}: `waterfront`,
- x_{11}: `view`,
- x_{12}: `condition`,
- x_{13}: `grade`,

El primer modelo de regresión lineal compuesta, que denominaremos $M1$, tiene la siguiente estructura:

$$\begin{aligned} y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 \\ + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_9 + \beta_{10} x_{10} \\ + \beta_{11} x_{11} + \beta_{12} x_{12} + \beta_{13} x_{13} + \epsilon \end{aligned} \quad (1)$$

Los resultados de este modelo pueden apreciarse en el Cuadro 4. La prueba de hipótesis permite rechazar la hipótesis nula en la ninguno de los coeficientes aporta en la predicción de la variable dependiente ($p < 2{,}2 \cdot 10^{-16}$). Se ve también que la prueba de hipótesis individual para la variable `floors`, cae en la región de rechazo. Por otro lado, la variable `sqft_basement` no genera resultado alguno.

	Variable dependiente:
	price
bedrooms	$-36{,}595.640^{***}$
	(2,214.283)
bathrooms	$39{,}940.160^{***}$
	(3,813.391)
sqft_living	169.008^{***}
	(5.168)
sqft_lot	0.002
	(0.057)
floors	$26{,}740.240^{***}$
	(4,179.062)
waterfront	$583{,}357.500^{***}$
	(21,443.110)
view	$43{,}052.690^{***}$
	(2,511.991)
condition	$18{,}891.770^{***}$
	(2,739.184)
grade	$119{,}160.100^{***}$
	(2,492.602)
sqft_above	-10.422^{**}
	(5.039)
yr_built	$-3{,}515.939^{***}$
	(74.440)
sqft_living15	28.524^{***}
	(3.978)
sqft_lot15	-0.500^{***}
	(0.088)
Constant	$6{,}094{,}389.000^{***}$
	(144,835.800)
Observations	17,289
R^2	0.653
Adjusted R^2	0.653
Residual Std. Error	213,726.000 (df = 17275)
F Statistic	$2{,}502.519^{***}$ (df = 13; 17275)
Note:	*p<0.1; **p<0.05; ***p<0.01

Cuadro 4: Resumen del modelo $M1$

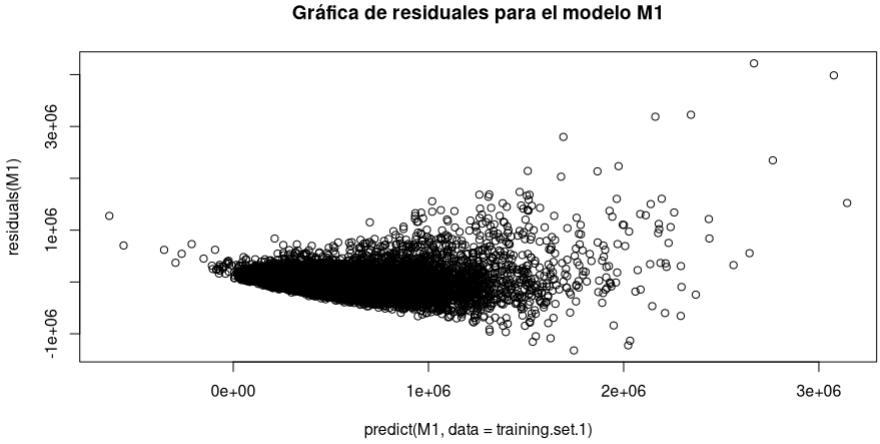


Figura 1: Residuales del modelo M1.

2.1. Ajustes al modelo

El modelo $M1$ presenta un problema que puede evidenciarse mejor al apreciar la gráfica de los residuales en la Figura 1. Los residuales no presentan un comportamiento heterocedástico.

Se plantean entonces algunas transformaciones para corregir este problema. En particular se consideraron las transformaciones cuadrática recíproca y Box Cox. Al modelo con la transformación $\sqrt{y}$ lo llamaremos $M2$, la modelo con la transformación $1/y$ lo llamaremos $M3$ y al modelo con la transformación y^{λ} lo denominaremos $M4$. Las Figuras 2, 3 y 4 muestran los residuales para estos nuevos modelos. El valor de λ máximo calculado con R es de $0.1\bar{0}$

Estas gráficas muestran que los modelos $M2$ Y $M4$ muestran el comportamiento deseado. Sin embargo seleccionaremos el modelo $M4$ por tener mayor coeficiente de determinación.

3. Diagnóstico

3.1. Normalidad de los residuales

Para evaluar la normalidad de los residuales del modelo $M4$ utilizaremos una comprobación visual a través del q-q plot. A partir de la Figura 5 podemos evidenciar un comportamiento normal en los residuales.

3.2. Independencia de los errores

Comprobaremos la independencia de los errores a través del test de Durbin-Watson. Este test arroja que $DW = 2{,}033$ y $p = 0{,}5851$. El valor de p no

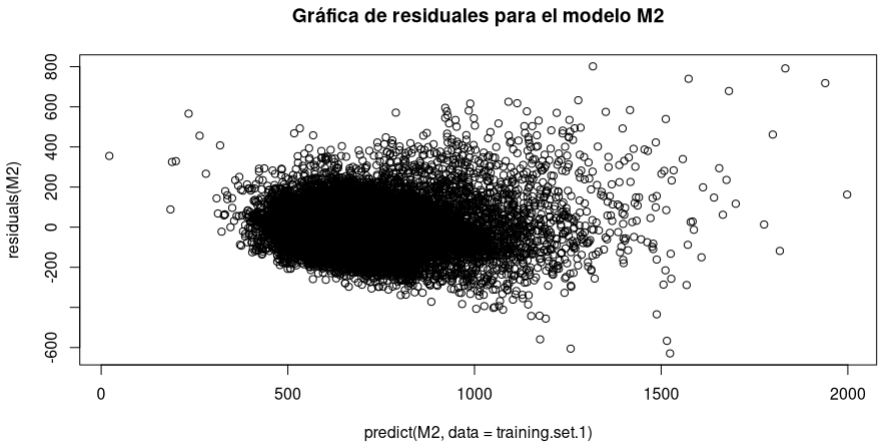


Figura 2: Residuales del modelo M2.

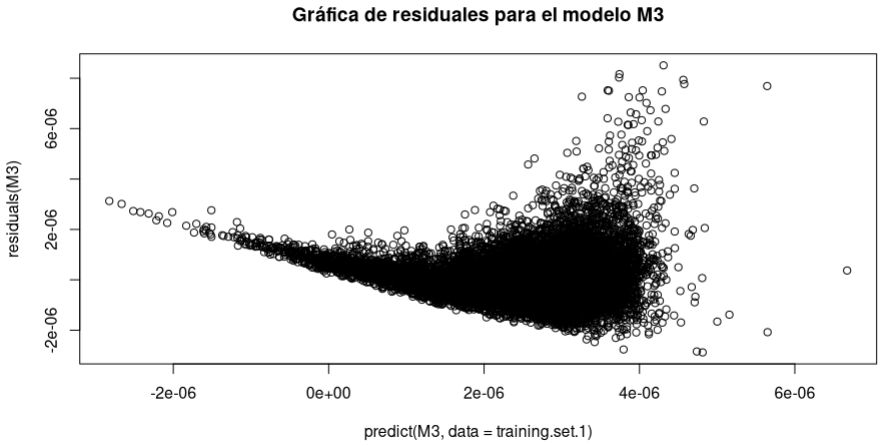


Figura 3: Residuales del modelo M3.

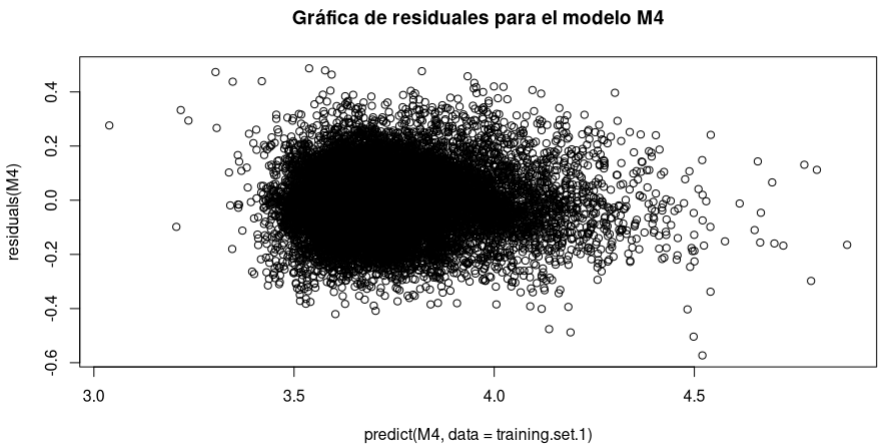


Figura 4: Residuales del modelo M4.

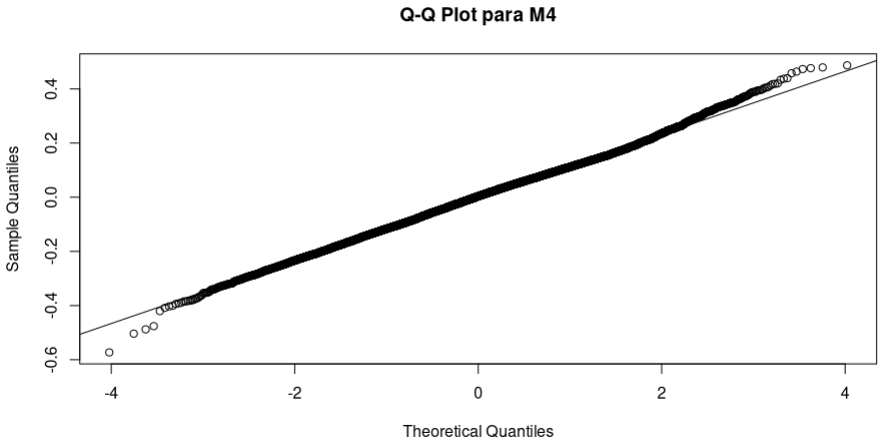


Figura 5: Q-Q Plot para modelo M4.

permite rechazar la hipótesis nula que indica que no existe autocorrelación de primer orden. Es decir, puede que haga falta aplicar un modelo de regresión generalizado.

3.3. Observaciones inusuales

3.3.1. Valores atípicos

El resultado del test de los p-valores de Bonferroni sugiere que la observación en la fila 13040 es la más extrema.

```
      rstudent unadjusted p-value Bonferroni p
13040 -4.937343         7.9933e-07      0.01382
```

3.3.2. Observaciones con *leverage*

Para visualizar aquellos valores con alto *leverage* utilizaremos el *half normal plot* de la Figura 6. Aquí se evidencia que los registros 13627 y 3635 tienden a ejercer apalancamiento en los datos.

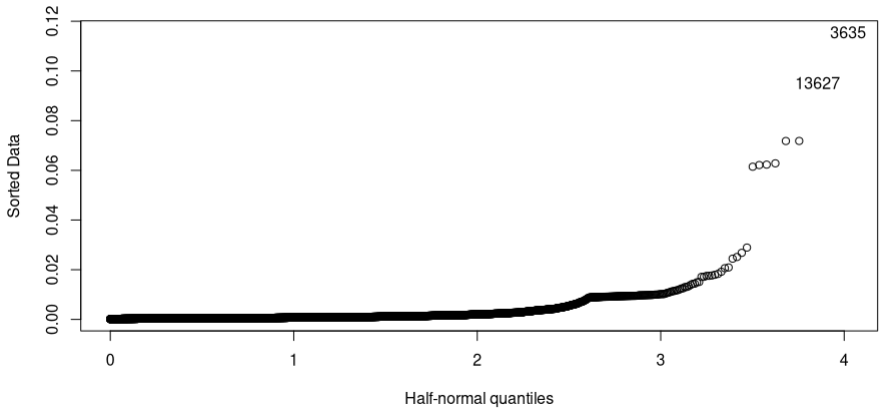


Figura 6: Q-Q Plot para modelo M4.

3.3.3. Observaciones influyentes

Para identificar las observaciones más influyentes en el modelo se halla la distancia de Cook. La Figura 7 no es muy clara en cuanto al efecto visual que debería generar.

El gráfico de influencia de la Figura 8 arroja algo más de información con respecto a aquellas observaciones que más influyen en el modelo.

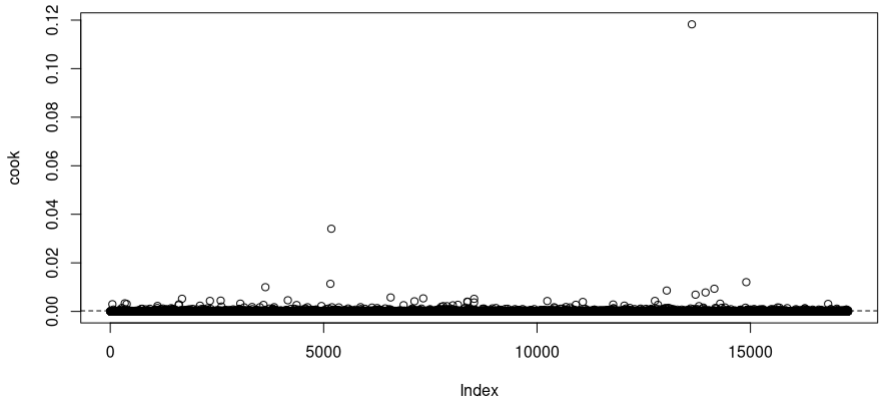


Figura 7: Distancia de Cook para modelo M4.

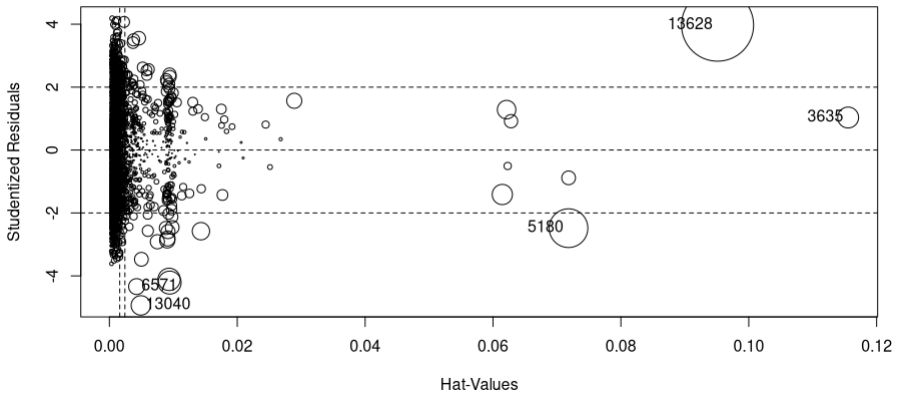


Figura 8: Gráfico de influencia para modelo M4.

3.4. Verificación de la estructura del modelo

EL Cuadro 5 exhibe la matriz de correlación para el conjunto de datos depurado. Inicialmente no se evidencian altos niveles de correlación entre las variables. Aunque se hace llamativa la indagación entre la relación que pueda existir en las variables x_3 (`sqft_living`) y x_8 (`sqft_living15`), o x_3 (`sqft_living`) y x_{13} (`grade`).

	y	x_1	x_2	x_3	x_4	x_5	x_{10}	x_{11}	x_{12}	x_{13}	x_6	x_7	x_8	x_9
y	1	0,31	0,52	0,71	0,09	0,26	0,26	0,39	0,03	0,67		0,06	0,59	0,08
x_1	0,31	1	0,50	0,57	0,03	0,17	-0,01	0,08	0,03	0,35		0,14	0,38	0,02
x_2	0,52	0,50	1	0,75	0,09	0,50	0,06	0,19	-0,13	0,66		0,50	0,57	0,09
x_3	0,71	0,57	0,75	1	0,17	0,36	0,10	0,28	-0,06	0,76		0,32	0,76	0,18
x_4	0,09	0,03	0,09	0,17	1	0	0,02	0,06	-0,01	0,12		0,06	0,14	0,73
x_5	0,26	0,17	0,50	0,36	0	1	0,03	0,04	-0,27	0,46		0,49	0,28	-0,01
x_{10}	0,26	-0,01	0,06	0,10	0,02	0,03	1	0,38	0,01	0,08		-0,02	0,08	0,03
x_{11}	0,39	0,08	0,19	0,28	0,06	0,04	0,38	1	0,04	0,25		-0,05	0,28	0,06
x_{12}	0,03	0,03	-0,13	-0,06	-0,01	-0,27	0,01	0,04	1	-0,15		-0,36	-0,09	-0,01
x_{13}	0,67	0,35	0,66	0,76	0,12	0,46	0,08	0,25	-0,15	1		0,45	0,72	0,12
x_6											1			
x_7	0,06	0,14	0,50	0,32	0,06	0,49	-0,02	-0,05	-0,36	0,45		1	0,33	0,07
x_8	0,59	0,38	0,57	0,76	0,14	0,28	0,08	0,28	-0,09	0,72		0,33	1	0,18
x_9	0,08	0,02	0,09	0,18	0,73	-0,01	0,03	0,06	-0,01	0,12		0,07	0,18	1

Cuadro 5: Matriz de correlación

El Cuadro 6 muestra los gráficos de residuales parciales para cada variable. Esta gráfica sugiere que las variables `bedrooms`, `sqft_lot`, `floors` y `condition` no deberían ser incluidas en el modelo, pues su efecto sobe la variable dependiente no se ajusta al de una línea recta.

Para determinar si existe multicolinealidad entre las variables del modelo, calcularemos el factor de inflación de la varianca (VIF) para las variables. Al observar lel gráfico, se evidencia la fuerte correlación entre las variables `sqft_above` y `sqft_living`. En este caso, basándonos en la matriz de correlación del Cuadro 5, se puede descartar la variable `sqft_living` al tener mayor correlación con el resto de variable y, por tanto, menor significancia estadística.

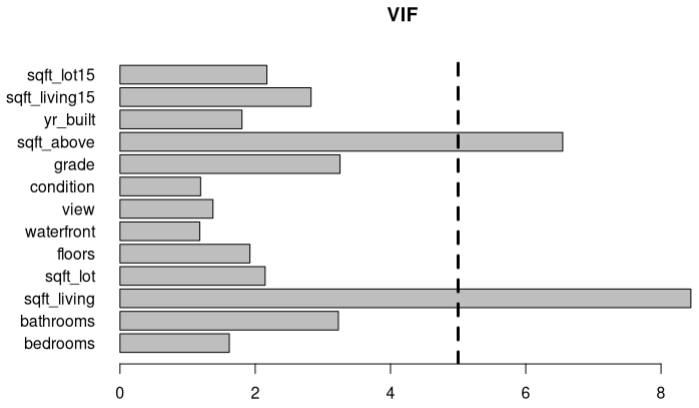
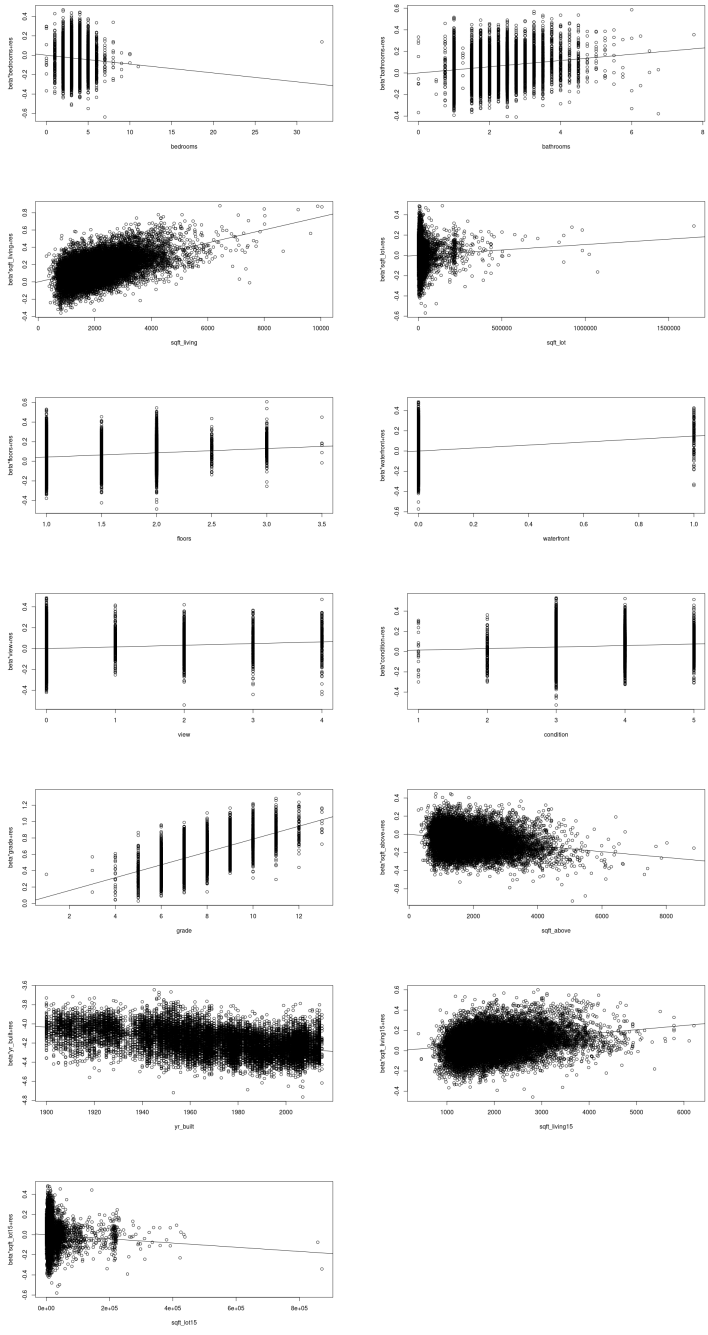


Figura 9: Factor de inflación de la varianza (VIF)

Por último, evaluamos del cuadrado medio del error (RMSE) a través de la siguiente fórmula:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n} (\hat{Y}_i - Y_i)^2}{n}}$$

Para el caso del modelo $M4$ este valor es 0.1163. Este valor indica que la diferencia entre los valores predichos y los reportados no es muy alta.



Cuadro 6: Gráficas de residuales parciales para cada coeficiente.