UNIVERSIDAD DE LA REPÚBLICA
Facultad de Ciencias Económicas y de Administración
Licenciatura en Estadística

Análisis explicativo del precio de apartamentos publicados en Airbnb de Barcelona

Viscailuz, Luciana Miranda, Germán
Junio 2024

Trabajo final de Modelos Lineales

Índice

Introducción

Airbnb es una compañía dedicada a la oferta de alojamientos de carácter vacacional en muchos de los países del mundo que funciona a partir de un programa digital donde los anfitriones pueden publicar sus propiedades para que los clientes puedan verlas y elegir el alojamiento que más se adapte a sus necesidades.

En este proyecto se trabajó con algunas de las propiedades publidcadas en esta plataforma en la ciudad Barcelona, España. Inicialmente, la motivación del proyecto fue poder estimar el precio (en euros) de diferentes apartamentos según las características de cada uno. Para realizar lo mencionado se trabajó con los datos de Airbnb Barcelona, donde se tenía el registro de más de 16.000 apartamentos de dicha ciudad.

La información con la que se contaba era buena pero excesiva, lo que llevó a que algunos datos fueran redundantes, por lo cual una parte importante de este proyecto fue la limpieza de los datos para así disponer de información que permita realizar una buena estimación e interpretación.

Finalmente, el trabajo e interpretación de los datos, tanto sobre como actúan entre si y sus diversos efectos sobre la variable de interés fueron los que guiaron y generaron el interés en este proyecto provocando que el paso a paso sea tan importante como el resultado final.

Los datos

Como se mencionó anteriormente, se disponía de la información de 16.761 apartamentos de Barcelona, donde se nombraban las características que los huéspedes toman en cuenta al momento de elegir su hospedaje y por lo tanto podrían llegar a incidir en su precio, entre estas se destacaba, ubicación, cantidad de camas y baños, cuantas personas se aceptaban, entre otras.

Habían variables cuantitativas pero la mayoría eran cualitativas, e incluso se pasaron a factor algunas de las cuantitativas para su mejor interpretación.

La información de código postal y barrio se decidió resumirla en una variable llamada distrito la cual agrupó los 73 barrios de Barcelona en 10 distritos.

Se decidió prescindir de la latitud y longitud de cada apartamento como de algunas características que se encontraban dentro de la variable amenities, tomando en cuenta finalmente las diez que generalmente se consideran más relevantes.

Se definieron dos nuevas variables, “grupo_habitacion” y “grupo_banios”, donde se agruparon la cantidad de habitaciones y baños respectivamente, para así disminuir la cantidad de categorías de cada una.

Del total de observaciones se operó con 12.848 debido a que las restantes contaban con datos faltantes.

Por último se decicdió trabajar con la transformación logaritmica de la variable respuesta, para así poder solucionar varios problemas relacionados con el diagnóstico que se verá posteriormente.

Finalizada la limpieza y organización de datos se pudo comenzar a trabajar con ellos.

Table 1: Descripción de las variables utilizadas en el informe

distritos	tipo_habitacion	puntuacion	estancia_min	grupo_habitacion
L'Eixample :4585	Entire home/apt:6055	Min. : 20.00	Min. : 1.000	Chica :10639
Ciutat Vella :2870	Private room :6752	1st Qu.: 88.00	1st Qu.: 1.000	Grande: 2209
Sants-Montjuic:1528	Shared room : 41	Median : 93.00	Median : 2.000	NA
Gracia :1062	NA	Mean : 90.98	Mean : 5.977	NA
Nou Barris : 886	NA	3rd Qu.: 97.00	3rd Qu.: 3.000	NA
Sant Martí : 645	NA	Max. :100.00	Max. :365.000	NA
(Other) :1272	NA	NA	NA	NA

Table 2: Descripción de las variables utilizadas en el informe

precio_euros	Wifi	Air_conditioning	personas	grupo_banios
Min. : 8.00	0:3255	0:5494	Min. : 1.000	Muchos: 253
1st Qu.: 40.00	1:9593	1:7354	1st Qu.: 2.000	Pocos :12595
Median : 62.50	NA	NA	Median : 2.000	NA
Mean : 92.45	NA	NA	Mean : 3.411	NA
3rd Qu.: 110.00	NA	NA	3rd Qu.: 4.000	NA
Max. :1000.00	NA	NA	Max. :18.000	NA

Análisis exploratorio

Como parte de la estadística descriptiva se crearon gráficos donde se relacionan algunas de las variables explicativas con la variable de respuesta (precio en euros). Estos graficos permiten obtener interpretaciones de las diferentes relaciones, pero es muy importante destacar que las interpretaciones obtenidas son parciales, debido a que a diferencia del modelo, en cada uno de los gráficos se representa el efecto de una variable sin tomar en cuenta las demás.

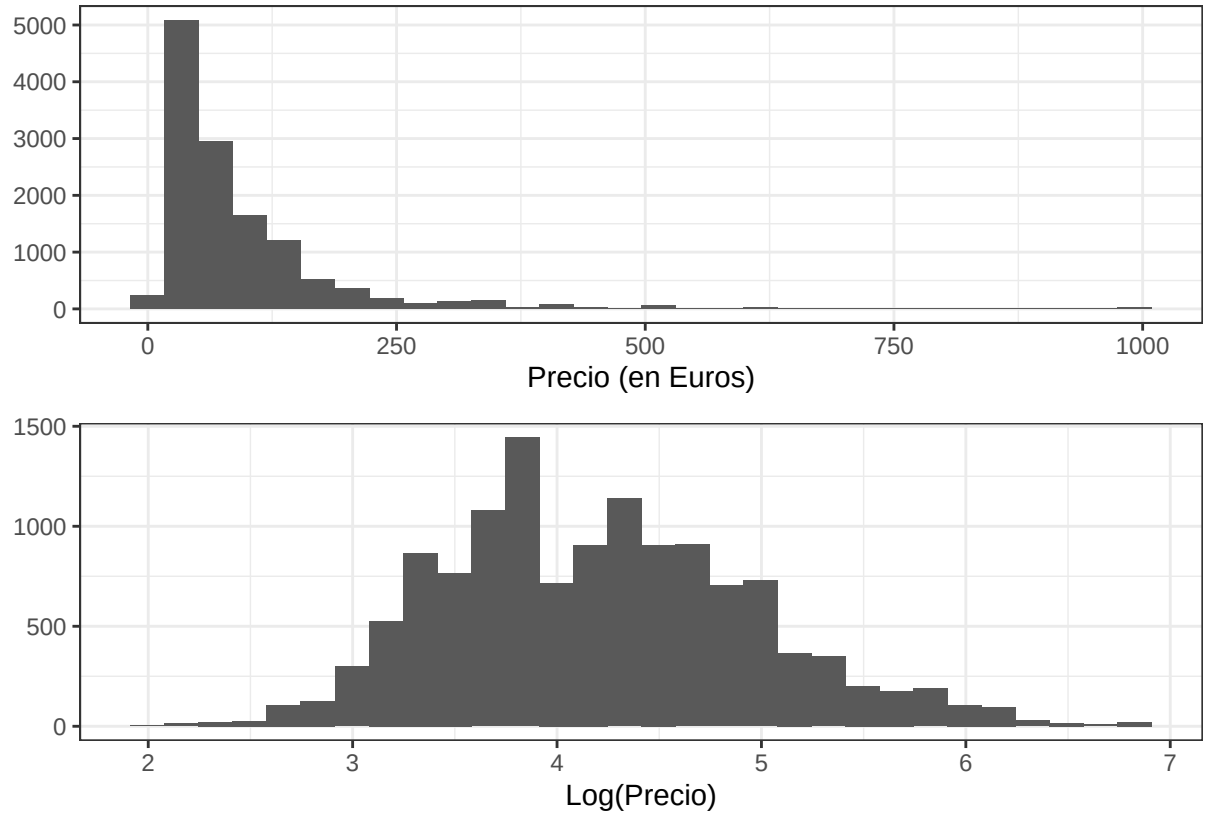

Figure 1: Histograma de la variable de respuesta precio euros

En la Figura 1 se puede observar la distribución de la variable precio y cómo se normaliza al aplicar la transformación de logaritmo.

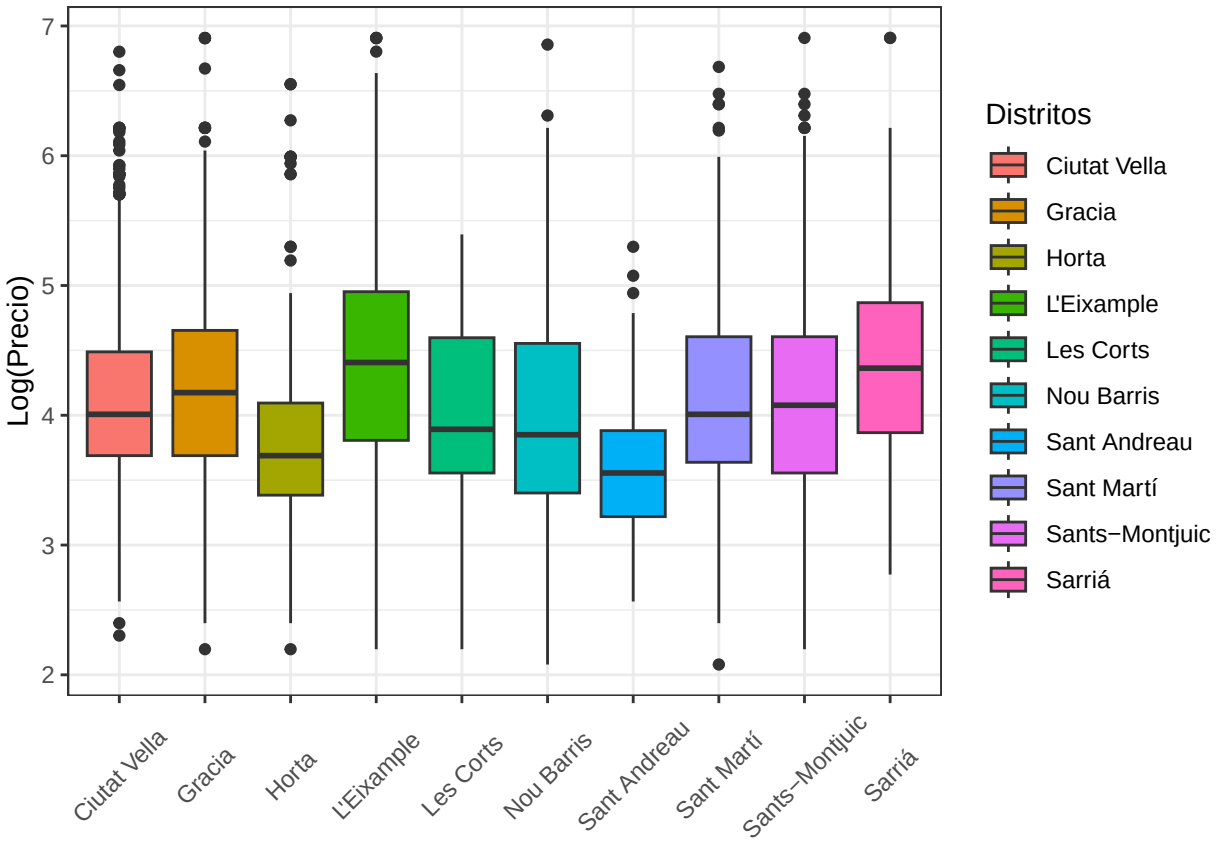

Figure 2: Dispersión de Precio (en euros) por Distrito

En la Figura 2 se muestra la distribución del log(Precio) para cada uno de los Distritos. Se observa que la mayoría de las cajas se encuentran superpuestas entre si, por lo que no se podría afirmar cuál es el distrito que tiene apartamentos con un log(Precio) más alto, pero si se pueden interpretar algunas diferencias entre ellos, como que los apartamentos del distrito L'Eixample tienen un log(Precio) mayor que los que se encuentran en Sant Andreau.

También se resalta la presencia de observaciones atípicas, que tienden a tener un log(Precio) más alto que las mayoría.

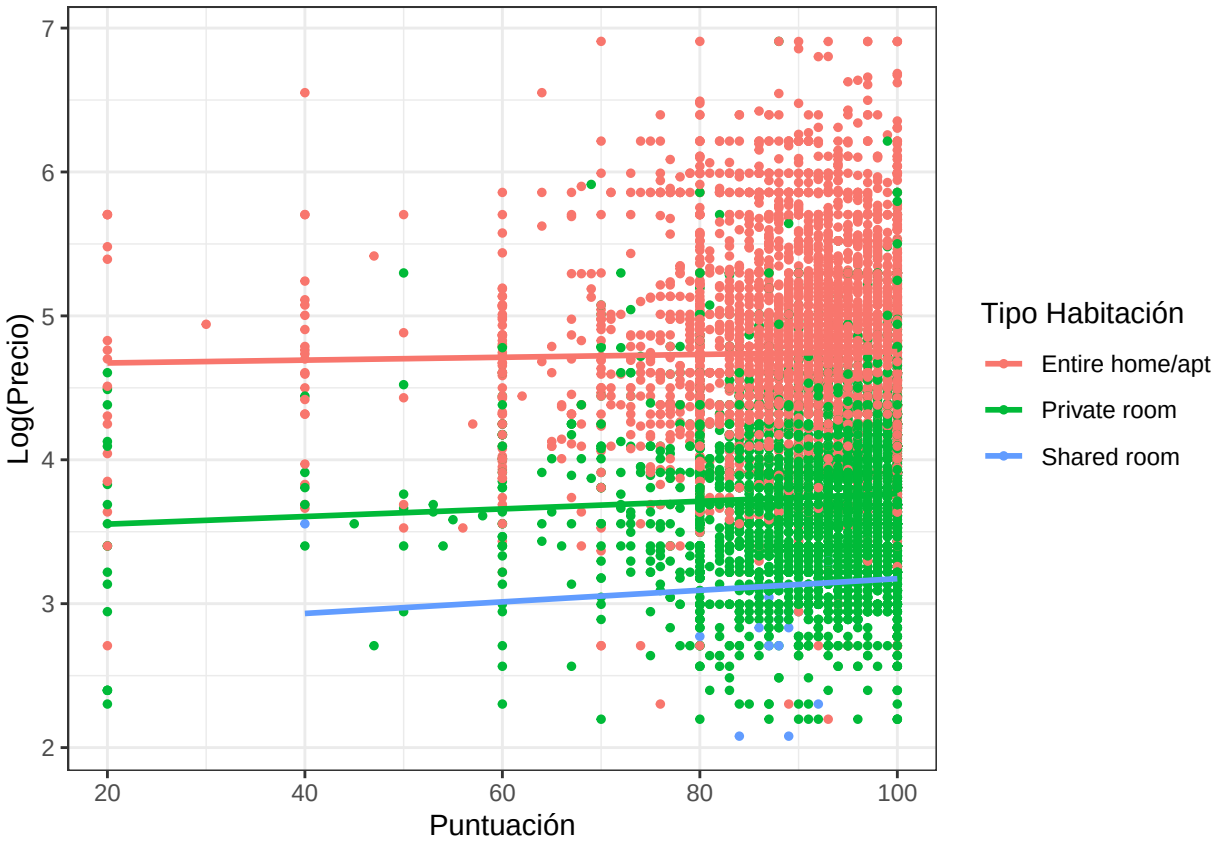

Figure 3: Dispersión de Precio (en euros) según la puntuación, por tipo de alojamiento

En la figura 3 se muestra a partir de un gráfico de dispersión, la relación entre las diferentes combinaciones de las variables explicativas puntuación y tipo de habitación, con la variable respuesta, log(Precio). Lo primero a destacar es que los puntos tienden a estar acumulados en la parte derecha del gráfico, teniendo en su mayoría una puntuación mayor a 60, por lo que permite interpretar que no existe ninguna relación entre puntuación y log(precio). Por otro lado también se puede observar que existe una relación muy marcada entre el tipo de habitación y la variable respuesta, dado que los apartamentos enteros se encuentran en la parte superior derecha del gráfico, es decir, que están relacionados con precios más altos, mientras que el resto se encuentran en el extremo inferior.

Metodología

Selección de variables

A partir de las diferentes combinaciones de las variables explicativas se puede llegar a una gran cantidad de modelos que busquen predecir el precio y cada uno de estos llevará a una predicción diferente del mismo. Uno de los problemas centrales es encontrar cuál de todos estos modelos cumple mejor su objetivo. En este proyecto finalmente se decidió trabajar con quince variables explicativas que formaron parte del modelo incial.

```
mod0 <- lm(log(precio_euros) ~  distritos + tipo_habitacion+personas+ grupo_banios+
             grupo_habitacion+estancia_min+puntuacion+TV+Wifi+Air_conditioning+Elevator+
             Breakfast+Pets_allowed+Patio_or_balcony+check_in_24_hs, data=df_final)
```

Diagnóstico

Luego de la limpieza de datos y estadística descriptiva se comenzó la etapa de diagnóstico, etapa imprescindible debido a que el cumplimiento de todos los supuestos sobre el modelo es el que permite afirmar que las inferencias realizadas sean válidas.

Entre estos supuestos se encuentran:

Multicolinealidad: En esta prueba se busca que ninguna de las columnas de la matriz X sea "casi" combinación lineal de las demás. Cuando esto sucede, el número de condición aumenta, lo que lleva finalmente a que la inversa de X'X sea inestable. Esta inestabilidad es la que finalmente se busca evitar.

Linealidad: Este supuesto se basa en la linealidad de la variable log(Precio) y cada una de las variables explicativas. Si en el modelo no hay linealidad, se presentarán problemas de correlación entre los residuos y variabilidad de los mismos. Para verificar el cumplimiento de este supuesto se realizó un análisis gráfico entre los $\hat{Y}$ y $\hat{\epsilon}$, en el cual se busca no encontrar patrones.

Homoscedasticidad: Se busca que el modelo sea homoscedastico, es decir, que la varianza de todos los residuos sea constante. Esto se interpreta como que la varianza no depende de ninguna de las variables explicativas. Esta prueba se suele ver mediante gráficos que relacionan cada variable explicativa con la de respuesta y a partir de una prueba de hipótesis donde se busca no rechazar la hipótesis nula.

Normalidad: Se refiere a que los residuos deben tener una distribución normal. Este supuesto es muy importante debido a que es el que luego permite realizar inferencias. Sin embargo, en los modelos donde el tamaño de muestra es grande, como en este caso, la falta de normalidad de los residuos no generan repercusiones.

En las siguientes lineas del script se pusieron a prueba cada uno de los supuestos antes mencionados.

Multicolinealidad

Se comenzó evaluando el supuesto de multicolinealidad. Para esto se calculó el número de condición, el cual dió 1711,793, es decir, existían problemas de multicolinealidad.

Luego se calculó el VIF para cada una de las variables explicativas donde todas tenían un VIF mayor a 1 pero menor a 5, por lo cual, no se pudo determinar cual es la variable que generó el problema del cumplimiento de este supuesto. De igual forma, esto no provocó grandes problemas debido a que se trabajó con un modelo con muchas variables cualitativas, lo cual hace que la evaluación de este supuesto no tenga tanta relevancia.

Table 3: Valores de la prueb VIF

	Valor VIF
Pets_allowed1	1.023216
tipo_habitacionShared room	1.029174
Patio_or_balcony1	1.037473
puntuacion	1.042523
Breakfast1	1.048010
distritosSant Andreau	1.069657
check_in_24_hs1	1.079611
distritosLes Corts	1.082300
estancia_min	1.094684
distritosHorta	1.115022
grupo_baniosPocos	1.134351
distritosSarriá	1.149912
distritosSant Martí	1.195899
Elevator1	1.197727
distritosNou Barris	1.267226
distritosGracia	1.288725
distritosSants-Montjuic	1.386380
Air_conditioning1	1.635899
grupo_habitacionGrande	1.850647
distritosL'Eixample	1.987565
tipo_habitacionPrivate room	2.873404
personas	3.203203
Wifi1	3.716427
TV1	4.069130

Linealidad

Luego se siguió con el supuesto de linealidad, donde se realizó el gráfico entre residuos y predichos. En el mismo se puede observar que no existe ningún patrón, por lo cual se podría afirmar que se cumple el supuesto de linealidad. Esto se podría dar como una repercusión de la transformación logaritmica, debido a que esta permite solucionar casos donde la variable respuesta no presenta una relación lineal con una o más variables explicativas.

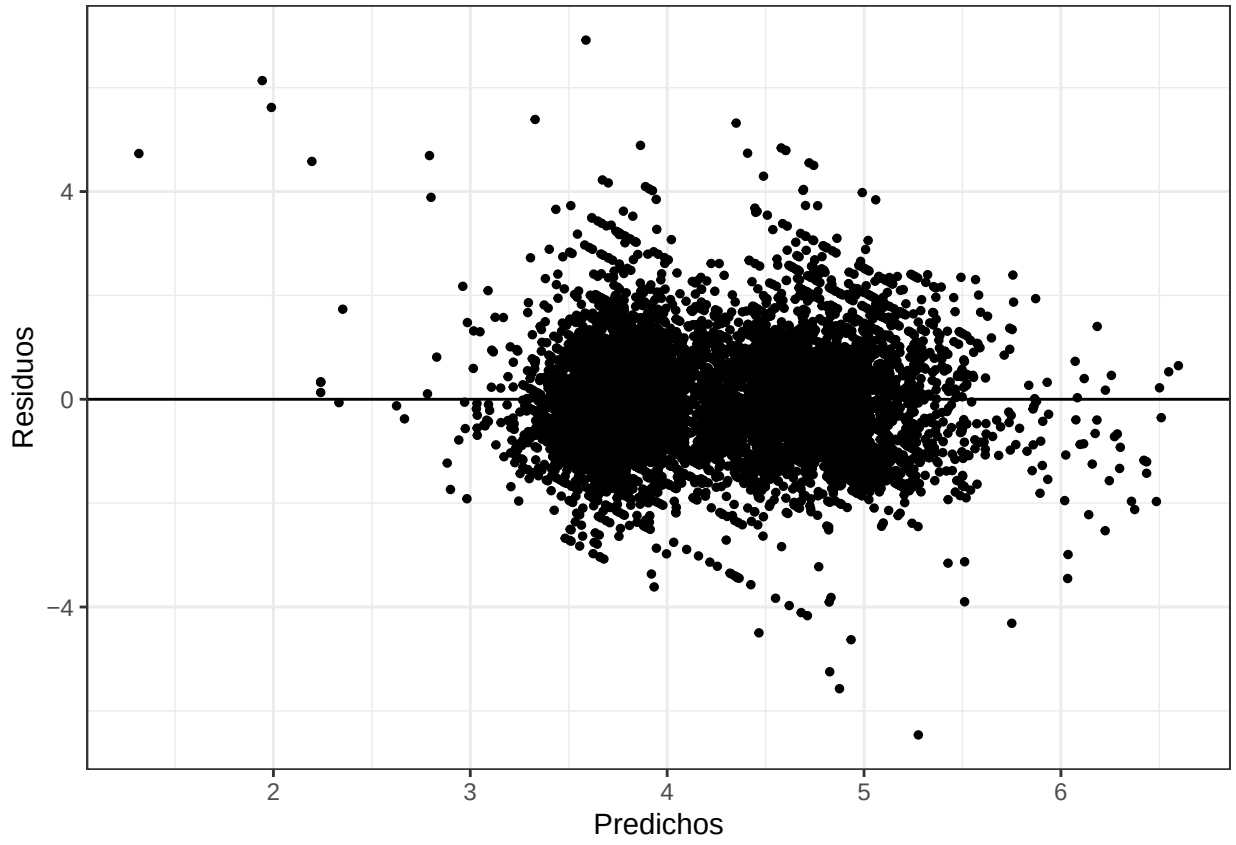

Figure 4: Gráfico de puntos de los residuos del modelo en función de los valores predichos

Homoscedasticidad

Como ya se mencionó, los dos primeros supuestos diagnosticados son importantes pero no tanto para un modelo como el presente, por lo cual se centró la atención en diagnosticar los supuestos restantes, como la homoscedasticidad, donde a partir del test de BREUSCH-PAGAN se pudo observar que en el modelo inicial no se cumplía este supuesto. El p-valor dió muy bajo, lo que provocó que se rechazara H0.

Table 4: Resultados de la prueba de BREUSCH-PAGAN

statistic	p.value	parameter	method	alternative
501.4026	<0,001	24	Koenker (studentised)	greater

Normalidad

Para evaluar la normalidad, se realizó un histograma y un QQ-Plot de los residuos Studentizados.

En el primer gráfico se pudo observar que tenía forma de una campana simétrica (hay que resaltar que la forma del histograma está muy influenciada por el tamaño de los bins), en cambio en el segundo gráfico los puntos se encontraban fuera de la banda, es decir, el primer gráfico dio indicios de que si se cumplia el supuesto de normalidad pero el segundo mostraba que no.

Para definir el cumpliento o no del supuesto se realizó la prueba de Kolmogorov-Smirnov. En esta prueba se obtuvo un p-valor muy chico el cual llevó a que se rechace H0 lo que concluyó que el supuesto de normalidad no se cumplió.

Sin embargo, obtener el cumplimiento de este supuesto no fue de mucha importancia para este modelo debido a la robustez que proporciona el Teorema Central del Límite cuando se trabaja con muchas observaciones.

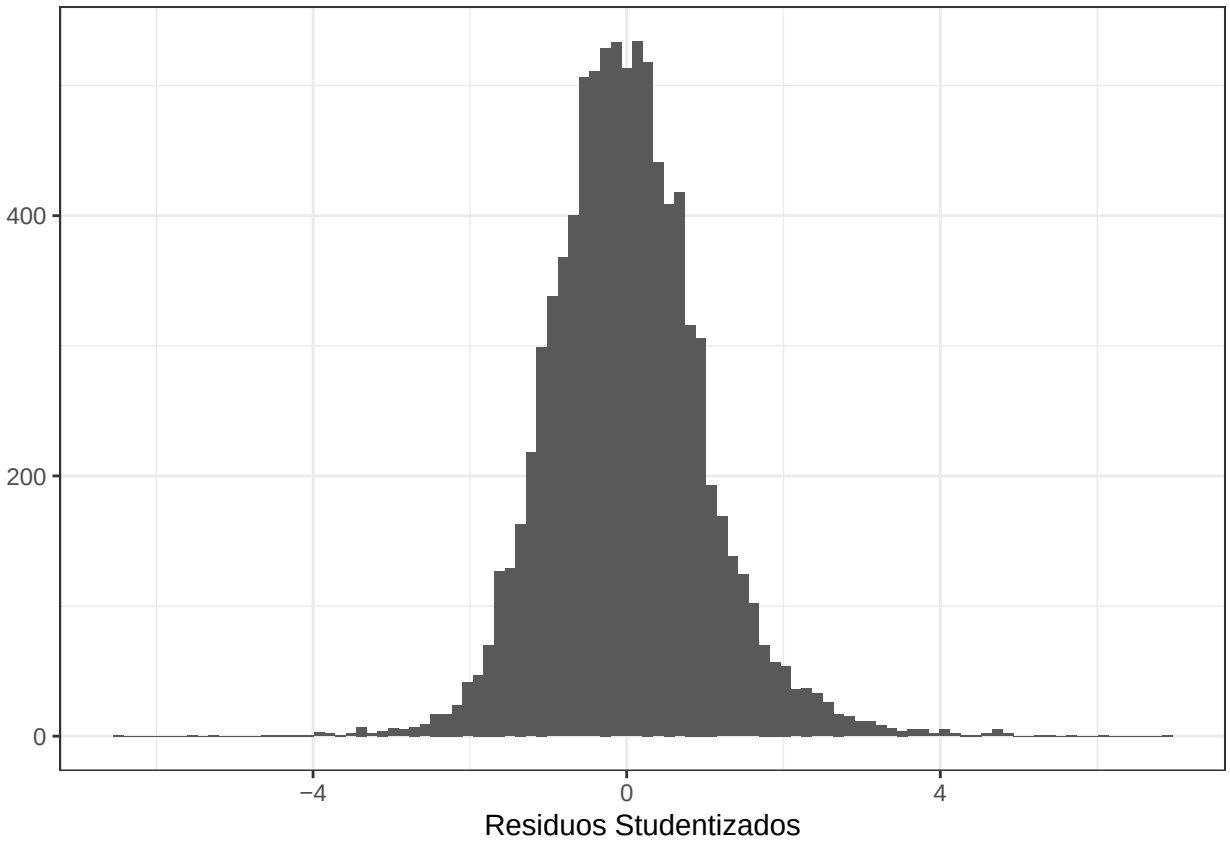

Figure 5: Histograma de los Residuos Studentizados

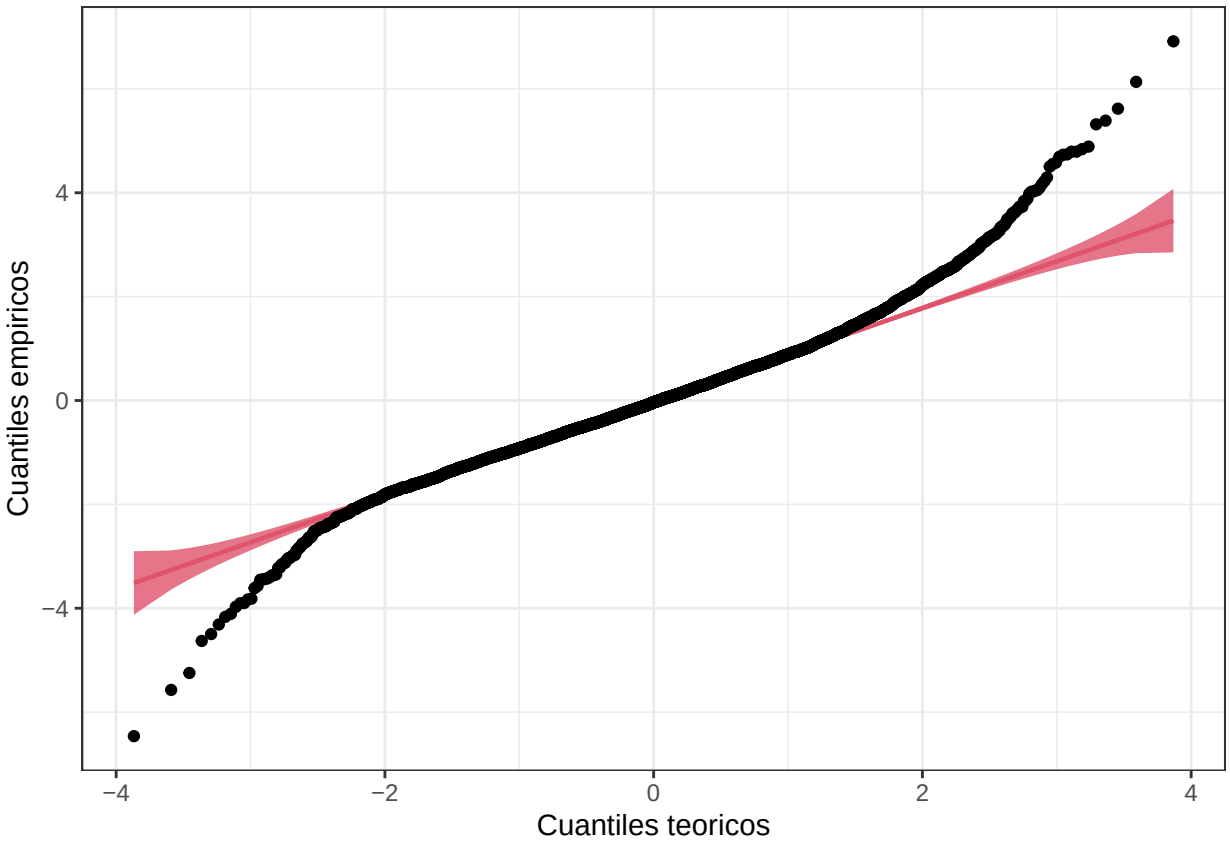

Figure 6: QQ-Plot de los Residuos Studentizados

Table 5: Resultados de la prueba de BREUSCH-PAGAN

P-Valor
<0,001

Atípicos

Por último se realizó el gráfico de la Distancia de Cook para observar si habían observaciones atípicas.

Usualmente en este gráfico se hace la línea roja horizontal al nivel de 4/n, donde n es el número de observaciones. Como en el modelo las observaciones eran muchas, esta linea roja quedaba muy baja, haciendo referencia a que todas las observaciones eran atípicas. Para poder aplicar esto en el modelo, se decidió que la linea esté en el nivel de 4/500, en ese punto las observaciones atípicas terminan siendo las más diferentes.

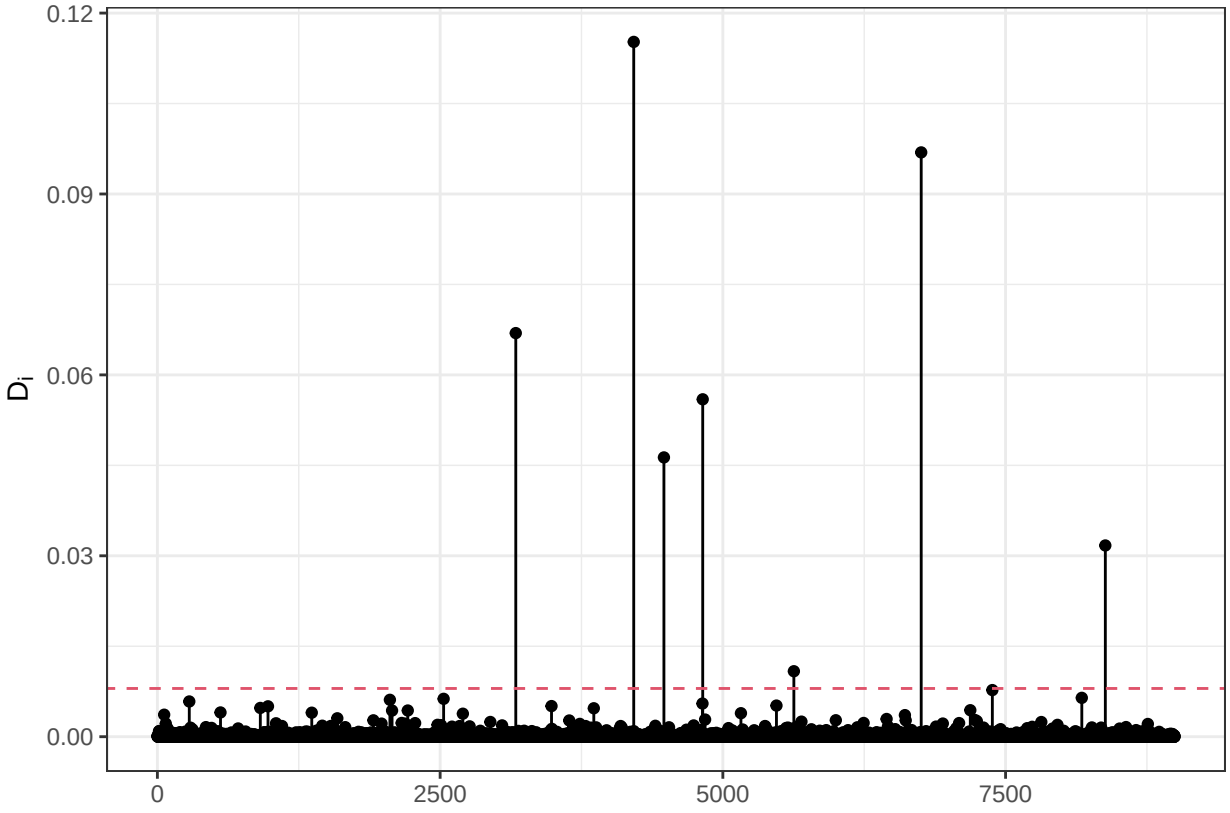

Figure 7: Gráfico de Distancia de Cook por observación del modelo 0

Correción del modelo

A continuación se verán algunos de los cambios realizados en el modelo con respecto al cumplimiento de los principales supuestos junto con los nuevos resultados. Se comenzó eliminando los datos atípicos y creando un nuevo modelo sin tomarlos en cuenta. Este procedimiento se repitió hasta obtener un modelo que se ajuste más a los objetivos iniciales.

```
mod1 <- lm(log(precio_euros) ~  distritos + tipo_habitacion+personas+ grupo_banios+
             grupo_habitacion+estancia_min+puntuacion+TV+Wifi+Air_conditioning+
             Elevator+Breakfast+Pets_allowed+Patio_or_balcony+check_in_24_hs, data=df_2)

Anova(mod1)
```

```
## Anova Table (Type II tests)
```

```
##
## Response: log(precio_euros)
##                   Sum Sq   Df  F value    Pr(>F)
## distritos          67.77    9  33.2597 < 2.2e-16 ***
## tipo_habitacion   225.56    2 498.1404 < 2.2e-16 ***
## personas          202.31    1 893.6037 < 2.2e-16 ***
## grupo_banios        2.55    1  11.2424 0.0008028 ***
## grupo_habitacion    1.83    1   8.0832 0.0044778 **
## estancia_min      133.82    1 591.0658 < 2.2e-16 ***
## puntuacion          3.64    1  16.0812 6.118e-05 ***
## TV                  4.36    1  19.2702 1.148e-05 ***
## Wifi                5.19    1  22.9104 1.724e-06 ***
## Air_conditioning   33.93    1 149.8781 < 2.2e-16 ***
## Elevator            3.15    1  13.9162 0.0001923 ***
## Breakfast           3.79    1  16.7617 4.275e-05 ***
## Pets_allowed        1.71    1   7.5489 0.0060165 **
## Patio_or_balcony    1.46    1   6.4337 0.0112143 *
## check_in_24_hs     11.90    1  52.5559 4.528e-13 ***
## Residuals        2028.99 8962
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

De la salida de Anova, se observó que todas las seleccionadas hasta esta instancia son significativas para el modelo a un nivel de 5%, por lo que se decidió no realizar cambios con respecto a ellas.

A continuación del estudio del anova, se volvieron a realizar las pruebas de diagnóstico con el nuevo modelo. No se obtuvieron resultados muy diferentes a los que ya se tenían, excepto de la última prueba, donde se calculó nuevamente la distancia de cook y se encontraron nuevas observaciones atípicas que fueron eliminadas obteniendo así un nuevo modelo, mod2. La siguiente salida corresponde al summary del modelo 2.

```
##
## Call:
## lm(formula = log(precio_euros) ~ distritos + tipo_habitacion +
##     personas + grupo_banios + grupo_habitacion + estancia_min +
##     puntuacion + TV + Wifi + Air_conditioning + Elevator + Breakfast +
##     Pets_allowed + Patio_or_balcony + check_in_24_hs, data = df_3)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.6847 -0.2983 -0.0184  0.2782  3.3248
##
## Coefficients:
##                                Estimate Std. Error t value Pr(>|t|)
## (Intercept)                   4.0397133  0.0693124  58.283  < 2e-16 ***
## distritosGracia              -0.1428684  0.0205607  -6.949 3.95e-12 ***
## distritosHorta               -0.2495083  0.0309839  -8.053 9.13e-16 ***
## distritosL'Eixample          -0.0319350  0.0147119  -2.171 0.029981 *
## distritosLes Corts           -0.2088467  0.0383750  -5.442 5.40e-08 ***
## distritosNou Barris          -0.2032104  0.0218850  -9.285  < 2e-16 ***
## distritosSant Andreau        -0.3778264  0.0418906  -9.019  < 2e-16 ***
## distritosSant Martí          -0.1098822  0.0251559  -4.368 1.27e-05 ***
## distritosSants-Montjuic      -0.1996524  0.0182161 -10.960  < 2e-16 ***
## distritosSarriá              -0.0331958  0.0299949  -1.107 0.268447
## tipo_habitacionPrivate room -0.5501264  0.0172399 -31.910  < 2e-16 ***
```

```
## tipo_habitacionShared room  -1.1448646  0.0923910 -12.392  < 2e-16 ***
## personas                     0.1205344  0.0041590  28.981  < 2e-16 ***
## grupo_baniosPocos           -0.1485515  0.0387097  -3.838 0.000125 ***
## grupo_habitacionGrande       0.0569183  0.0180285   3.157 0.001599 **
## estancia_min                -0.0125994  0.0004734 -26.614  < 2e-16 ***
## puntuacion                   0.0023316  0.0005351   4.357 1.33e-05 ***
## TV1                          0.0977819  0.0218924   4.466 8.05e-06 ***
## Wifi1                       -0.1086236  0.0221644  -4.901 9.71e-07 ***
## Air_conditioning1            0.1580737  0.0128916  12.262  < 2e-16 ***
## Elevator1                    0.0413311  0.0112736   3.666 0.000248 ***
## Breakfast1                   0.0907214  0.0218366   4.155 3.29e-05 ***
## Pets_allowed1                0.0439681  0.0161156   2.728 0.006379 **
## Patio_or_balcony1            0.0301949  0.0117814   2.563 0.010395 *
## check_in_24_hs1              0.1142929  0.0157051   7.277 3.69e-13 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4722 on 8954 degrees of freedom
## Multiple R-squared:  0.6107, Adjusted R-squared:  0.6097
## F-statistic: 585.3 on 24 and 8954 DF,  p-value: < 2.2e-16
```

Table 6: Tabla comparativa de métricas de los modelos definidos

Varianza_Mod0	Varianza_Mod1	Varianza_Mod2	R2_Mod0	R2_Mod1	R2_Mod2
0.482	0.476	0.472	0.595	0.605	0.611

Se compararon los resultados de los tres modelos definidos y se observó que las pruebas tuvieron resultados similares en todos los casos. A su vez, a medida que se fueron aplicando las correcciones la varianza fue disminuyendo modelo a modelo mientras que el R^2 fue aumentando.

Una vez obtenidos los resultados deseados, dentro de un margen aceptable, se decidió no seguir iterando las pruebas y tomar como modelo final el número 2 ("mod2").

Resultados

A nivel global, se observó que el p-valor de la prueba F es $< 0,001$ lo que indica que el modelo es significativo a todo nivel de significación. Además la proporcion de la variabilidad de "Log(Precio_Euros)" explicada por las variables explicativas es de un 61,07%.

Posteriormente, analizando cada una de las variables, se observó que todas las seleccionadas son significativas para el modelo a un nivel de 5%, aunque si el nivel fuera al 1%, la variable "Patio_or_balcony" ya no lo sería.

Dado que el objetivo es poder explicar el precio en euros, para la interpretación de los coeficientes, previamente se deshizo el cambio de variable aplicado, permitiendo obtener las siguientes conclusiones.

Table 7: Tabla de coeficientes del modelo 2

	Coeficientes	Exp_Coef	Porcentaje
(Intercept)	4.040	56.810	5581.0
distritosGracia	-0.143	0.867	-13.3
distritosHorta	-0.250	0.779	-22.1
distritosL'Eixample	-0.032	0.969	-3.1
distritosLes Corts	-0.209	0.812	-18.8
distritosNou Barris	-0.203	0.816	-18.4
distritosSant Andreau	-0.378	0.685	-31.5
distritosSant Martí	-0.110	0.896	-10.4
distritosSants-Montjuic	-0.200	0.819	-18.1
distritosSarriá	-0.033	0.967	-3.3
tipo_habitacionPrivate room	-0.550	0.577	-42.3
tipo_habitacionShared room	-1.145	0.318	-68.2
personas	0.121	1.128	12.8
grupo_baniosPocos	-0.149	0.862	-13.8
grupo_habitacionGrande	0.057	1.059	5.9
estancia_min	-0.013	0.987	-1.3
puntuacion	0.002	1.002	0.2
TV1	0.098	1.103	10.3
Wifi1	-0.109	0.897	-10.3
Air_conditioning1	0.158	1.171	17.1
Elevator1	0.041	1.042	4.2
Breakfast1	0.091	1.095	9.5
Pets_allowed1	0.044	1.045	4.5
Patio_or_balcony1	0.030	1.031	3.1
check_in_24_hs1	0.114	1.121	12.1

Observando por ejemplo el distrito de Gracia, se interpreta que es un 13,3% más barato que el distrito de referencia ("Ciutat Vella"), dejando todas las demás variables constantes. Para el resto de los distritos, la interpretación es análoga.

Si se mira el coeficiente de la variable "Estancia_min" se puede concluir que por cada día que aumente la estancia mínima, se espera que el precio decrezca un 1,3%. Lo mismo sucede para la variable "Puntuación", donde por cada punto que aumente la puntuación, se espera que el precio aumente un 0,2%

En ambos casos, siempre se deja todas las demás variables constantes.

Finalmente para comprobar la efectividad real del modelo, se realizaron predicciones con datos nuevos.

Table 8: Tabla predicciones de las primeras 10 observaciones

precio_euros	predicciones	precio_predic	delta_precio
33	3.624227	37.49574	-4.50
210	4.712469	111.32664	98.67
75	4.620906	101.58603	-26.59
85	4.789032	120.18500	-35.19
30	3.239149	25.51200	4.49
45	2.174781	8.80026	36.20
180	5.525650	251.04947	-71.05
116	4.576555	97.17904	18.82
250	5.052166	156.36070	93.64
110	4.548917	94.53002	15.47

Con estos valores, se calculó el R^2, obteniendo así casi un 40% de variabilidad del precio explicada por el modelo.

Table 9: R^2 calculado para los nuevos datos

Valor R Cuadrado
0.397

Conclusiones

Luego de realizadas las pruebas y predicciones se logró concluir que el modelo si es funcional y cumple con su objetivo, sin embargo, hay varios supuestos que no se cumplen lo cual lleva a pensar que es un modelo que todavía puede ser mejorado, ya sea agregando variables que sean relevantes que no estuvieran incluidas en los datos iniciales o trabajando de diferente forma con las presentes.

"Todos los modelos son incorrectos, pero algunos son útiles" George Edward Pelham Box

Bibliografía

Carmona, Francesc (2003). Modelos Lineales (notas de curso). Departament d'Estad´ıstica. Farraway, Julian (2014). Linear Models with R, second edition. Chapman Hall/CRC. Rencher, Alvin y Bruce Schaalje (2008). Linear Models in Statistics, second edition. John Wiley Sons, Inc. Peña, Daniel (2010). Regresi´on y Dise˜no de Experimentos. Alianza Editoria https://es.wikipedia.org/wiki/Distritos_de_Barcelona