Análisis de la Variación de Precios de Alquileres en Barcelona (España)

Trabajo Final - Modelos Lineales

Lucas Giúdice
Juan Manuel Karawacki
Bruno Pintos

30 de junio de 2025

Profesores: Fernando Massa e Ignacio Campón

Índice

1. Introducción

El objetivo de este informe es la construcción de un modelo de regresión lineal múltiple que logre explicar y predecir el motivo del aumento de precios de alquiler en la plataforma Airbnb, en la ciudad de Barcelona, provincia de Cataluña, España.

Para ello utilizaremos un conjunto de datos, proporcionado por el curso de Modelos Lineales, con una serie de variables explicativas que se encuentran en principio relacionadas con la variable de respuesta: el precio del alquiler.

Este análisis buscará implementar de forma práctica todos los conocimientos adquiridos durante el curso. Se evaluará la relevancia de variables, la construcción de un modelo eficiente que logre explicar la variabilidad del precio del alquiler con la menor cantidad de variables explicativas posibles, la búsqueda del cumplimiento de los supuestos clásicos de las regresiones lineales y la correcta explicación de las variables utilizadas.

El conjunto de herramientas gráficas jugará un rol fundamental para reafirmar las conclusiones obtenidas durante el informe así como también una forma de resumen concisa y entendible.

2. Análisis Exploratorio de los Datos

En esta sección se presentan las variables explicativas: una breve descripción de las mismas y el tipo de variable que son.

Nombre de la Variable	**Descripción**	**Tipo de Variable**
id	Identificador único del anuncio	Cualitativa
host_id	Identificador del anfitrión	Cualitativa
barrio	Nombre del barrio donde se ubica el alojamiento	Cualitativa
cod_postal	Código postal del alojamiento	Cualitativa
latitud	Coordenada geográfica de latitud	Cuantitativa
longitud	Coordenada geográfica de longitud	Cuantitativa
tipo_habitacion	Tipo de habitación ofrecida	Cualitativa
personas	Cantidad de personas que puede alojar	Cuantitativa
banios	Número de baños disponibles	Cuantitativa
habitaciones	Número de habitaciones disponibles	Cuantitativa
camas	Número de camas disponibles	Cuantitativa
amenities	Lista de servicios incluidos	Cualitativa
precio_euros	Precio por noche en euros	Cuantitativa
estancia_min	Número mínimo de noches requeridas	Cuantitativa
puntuacion	Puntuación promedio de los huéspedes	Cuantitativa

Cuadro 1: Variables del dataset Airbnb - Barcelona

3. Modelo Completo

Si tomamos como primera aproximación al análisis el modelo que toma todas las columnas del conjunto de datos como variables explicativas del precio en euros, resulta entonces:

$$\text{precio\_euros} = \beta_0 + \sum_{i=1}^{14} \beta_i x_{ij} + \varepsilon_j, \quad \text{donde}$$

$X = [id, host\_id, barrio, cod\_postal, latitud, longitud, tipo\_habitacion, personas, banios, habitaciones, camas, amenities, estancia\_min, puntuacion]$

4. Selección de Variables y Preparación del Modelo

Considerando el conjunto de variables que hay en el modelo, es claro que tenemos que tomar algún conjunto menor de variables. De no hacerlo estariamos trabajando con 15 variables de las cuales hay una que tiene cerca de 70 niveles y otra que aporta otros 3. Esto contabilizaria, entre variables y dummies, mas de 85 factores explicativos del precio de alquiler en euros. Claramente no es manejable. Sin contar dummies tenemos 14 variables explicativas y una variable a analizar: $precio_euros$. El objetivo va a ser reducir la dimensionalidad del modelo.

4.1. *Amenities*: Una variable compleja

Si nos adentramos en la variable *amenities* veremos que contiene las comodidades que la vivienda alquilada contiene. Sin embargo la variedad de "valores"que esta toma es absurva para generar un analisis con sentido: desde wifi, microondas, shampoo, detector de monoxido de carbono, etc. Se nos es imposible generar alguna información útil que sea generalizable para todas las observaciones a la vez de significativa para evaluar un precio de alquiler. Decidimos extraerla del conjunto de datos. Nos quedamos con 13 variables explicativas.

4.2. Análisis de datos faltantes (NA)

Si empezamos a indagar en la cantidad de observaciones que tienen al menos un datos omiso, veremos que hay algo menos de 4500 de datos faltantes. Esto considerando un total de 16800 observaciones aproximadamente es una enorme proporción. Sin embargo, rapidamente notamos que el 95 % de dichas faltas corresponden a las variables *puntuación* (con 3900) y *cod_postal* (con 500). Dado que ambas tienen grandes cantidades de datos faltantes, eliminamos sus columnas del conjunto de datos: con tanta cantidad de datos omisos (especialmente en puntuación) el aporte iba a ser escaso. Pasamos de 13 a 11 variables explicativas.

Como anexo, dado que los NA restantes eran menos de 100 en total (de un total de 16800), decidimos de prescindir de las observaciones que los causaban.

4.3. Mecanismos de selecicción de variables

Si bien aplicamos los algoritmos de selección dados en el curso: forward, backward o stepwise, el resultado no fue de utilidad. Esto se dió porque los tres consideraron que las 11 variables restantes eran todas significativas, por lo que no nos aportaron la ayuda que requeríamos para la reducción de la complejidad del modelo.

4.4. Variables identificatorias

Al considerar las variables *id* y *host_id* ambas variable de corte indexatorio. Es decir, ambas sirven como identificadores de usuarios por lo que creemos que el aporte práctico que puede tener para explicar o predecir un precio es pequeño. Optamos por excluirlas dada su poca interpretación practica. Quedamos con 9 variables.

5. Medidas de resumen

A continuación se presenta un resumen descriptivo de las variables numéricas del conjunto de datos. En la tabla se incluyen, para cada variable, el valor mínimo y máximo, los cuartiles primero (Q1) y tercero (Q3), la mediana, la media, la varianza y el coeficiente de variación. Estas medidas permiten analizar tanto la tendencia central como la dispersión de los datos, facilitando la interpretación y comparación entre variables.

Variable	Min	1er Qu.	Mediana	3er Qu.	Max	Media	CV* %
latitud	41.35	41.38	41.39	41.40	41.46	41.39	0.05
longitud	2.11	2.16	2.17	2.18	2.22	2.17	0.82
personas	1.00	2.00	2.00	4.00	18.00	3.36	64.32
banios	0.00	1.00	1.00	1.50	8.00	1.29	41.31
habitaciones	0.00	1.00	1.00	2.00	12.00	1.59	61.94
camas	0.00	1.00	2.00	3.00	30.00	2.24	80.33
precio_euros	7.00	40.00	63.00	107.00	1000.00	92.00	104.57
estancia_min	1.00	1.00	2.00	4.00	365.00	8.46	198.67

Cuadro 2: Medidas descriptivas para variables numéricas

6. Modelos trabajados

En el transcurso de trabajo surgieron dos posturas distintas respecto del manejo de las variables explicativas referentes a la geografía. Por un lado, la variable *barrio*: categórica y con muchisimos niveles, lo que implica dificultad adicional, pero potencialmente clara de interpretar y por tanto útil para explicar. Por otro lado, el conjunto entre las variables *latitud* y *longuitud* que si bien no nos darían a priori una clara interpetación practica, si nos ofrece mayor facilidad para el trabajo ya que reduce niveles (de 68 a 1) y es numérica, lo que la hace ideal para predicciones.Ante la firmeza de las posturas, y la oportunidad de enriquecer el proyecto, planteamos analizar ambos modelos: aquel que trate el factor geográfico de manera

categórica y otro de manera cuantitativa. Esto a su vez implica seleccionar distintas variables para trabajar.

6.1. Modelo Categórico: *Distritos*

Para el modelo categórico la variable predictora referente a la geografía será la variable *barrio*. Por lo tanto prescindiremos de utilizar *latitud* y *longuitud*.

Sin embargo, la variable *barrio* no será la variable que utilicemos en sí, sino una transformación de esta: *distritos*. Esto debido a una sencilla razón: *barrio* tiene 68 niveles correspondientes a distintos barrios. Es una variable con un contenido inabarcable. Al utilizar *distritos* estamos redujendo esos niveles de 68 a 10. Claramente es un cambio más que bienvenido, que nos facilitará la manipulación del modleo. Simplemente tuvimos que investigar cuales son cada uno de los distritos de la ciudad de Barcelona e incluir cada uno de los barrios que teniamos dentro de un distrito.

Agregamos entonces *distrito* y eliminamos *barrio*. En este caso entonces quedamos con 7 variables explicativas en total, incluyendo una de mucha menor dimensionalidad que una que previamente tenáimos.

Resulta entonces

$$\text{precio\_euros}_j = \beta_0 + \sum_{i=1}^{7} \beta_i x_{ij} + \varepsilon_j, \quad \text{donde}$$

$$X = [\text{personas, habitaciones, banio, estancia\_min, camas, distritos, tipo\_habitacion}]$$

6.2. Modelo Cuantitativo: *Latitud* y *Longuitud*

Por el contrario, en el caso del modelo cuantitativo (aunque no es tal ya que mantiene la variable $tipo_habitacion$, pero por simplificar) la referencia geográfica será dada por las variables *latitud* y *longuitud*. Es por eso que prescindimos de la utilización de *barrio* y la eliminamos. Nos quedamos con 8 variables explicativas, una mas que la del modelo categórico pero con mucho menor cantidad de niveles.

El modelo se define como:

$$\text{precio\_euros}_j = \beta_0 + \sum_{i=1}^{7} \beta_i x_{ij} + \varepsilon_j, \quad \text{donde}$$

$$X = [\text{personas, habitaciones, banio, estancia\_min, camas, latitud, longitud, tipo\_habitacion}]$$

En este caso, las coordenadas geográficas permiten capturar la variabilidad espacial sin utilizar variables categóricas con muchos niveles como lo sería *barrio*.

7. Diagnóstico del Modelo

Habiendo seleccionado un número más administrable de variables explicativas, procedemos a realizar la evaluación de los cuatro supuestos que tiene que cumplir un modelo para ser considerado lineal: no multicolinealidad, linealidad, homocedasticidad y distribución normal de los errores (normalidad).

Como se han planteado dos modelos alternativos (uno categórico y otro cuantitativo), el análisis de supuestos se presentará de forma diferenciada, indicando claramente a cuál de los dos modelos corresponde cada evaluación.

Tenemos que considerar que como ambos modelos contienen por lo menos una variable categórica, tendremos que evaluar dos de los supuestos; no multicolinealidad y linealidad, para el conjunto de las variables numéricas de los modelos y dos para la totalidad del modelo; homocedasticidad y normalidad.

La parte numérica diferirá en que uno tendrá adicionalmente *latitud* y *longuitud*, mientras que el otro no

7.1. Multicolinealidad

El problema que causa la existencia de multicolinealidad (aproximada) entre las variables del modelo es que la matriz de diseño $\mathbf{X^T X}$ se vuelve cercana a ser no invertible. Esto provocaría una gran inestabilidad en las estimaciones de los coeficientes $\hat{\beta}$, haciendo que pequeñas variaciones en los datos generen grandes cambios en las estimaciones, afectando directamente la precisión de las predicciones.

La multicolinealidad aproximada implica que existe una relación lineal cercana —aunque no exacta— entre las variables explicativas. Si la relación fuera exacta, la matriz $\mathbf{X^T X}$ no sería invertible, y por tanto no existirían estimaciones únicas para los coeficientes.

Recordando que:

$$\hat{\beta}_k \sim \mathcal{N}\left(\beta_k,\ \sigma^2(\mathbf{X^T X})^{-1}\right)$$

se observa que, si el determinante de $\mathbf{X^T X}$ se aproxima a cero, el de su inversa será muy grande, lo que implica alta incertidumbre en las estimaciones, incluso si σ^2 se mantiene constante.

Para diagnosticar la multicolinealidad utilizaremos la herramienta del VIF. La misma mide cuántas veces se incrementa la varianza de un estimador debido a la colinealidad entre predictores y se define como:

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

donde R_j^2 es el coeficiente de determinación obtenido al tomar el modelo con la variable X_j explicada y el resto de las X_i con $i \neq j$ del modelo.

En este análisis, se considerará que una variable presenta problemas de colinealidad cuando su VIF sea mayor o igual a 5. Los resultados obtenidos se resumen en los siguiente cuadro.

7.1.1. Modelo con distritos

Realizando el calculo del VIF

Variable	VIF
personas	4.95
baños	1.36
habitaciones	3.69
camas	4.97
estancia mín.	1.01

Cuadro 3: VIF de variables numéricas seleccionadas

Como vemos ningun valor supera el umbral establecido, por lo que no retiramos ninguna variable.

7.1.2. Modelo con Latitud y Longitud

Como primer paso en el análisis de multicolinealidad, se presenta la matriz de correlación bivariada entre las variables numéricas. Esta herramienta permite detectar relaciones lineales fuertes entre pares de variables que podrían anticipar problemas de colinealidad en el modelo.

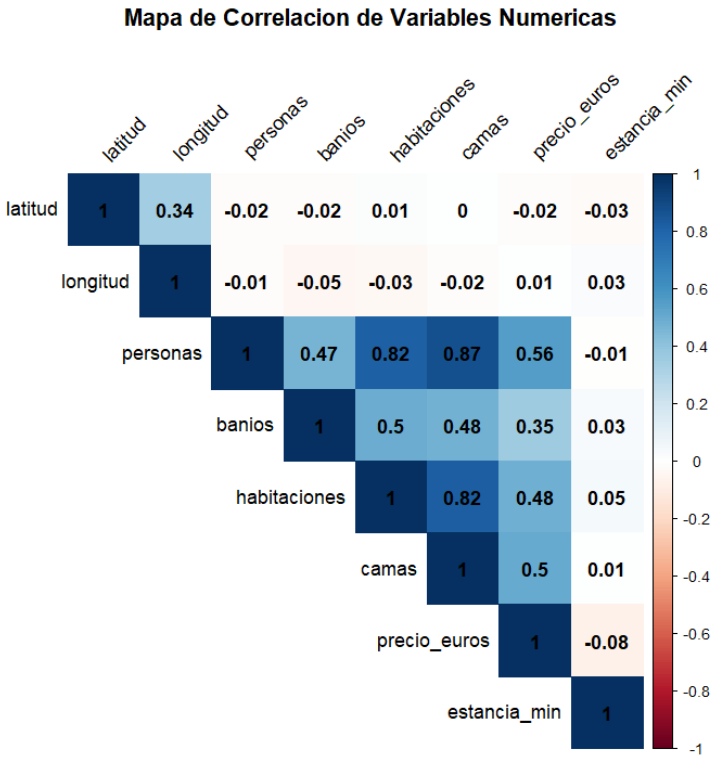

Figura 1: Matriz de correlación entre variables numéricas

El análisis evidencia una alta correlación entre las variables que describen la capacidad del alojamiento (*personas*, *habitaciones*, *baños* y *camas*), lo cual sugiere una posible redundancia informativa. No obstante, ninguna de estas variables muestra una asociación fuerte con el precio del alquiler, lo que sugiere que su variabilidad podría estar explicada por una combinación de factores y no por una única variable, destacándose el potencial rol de la ubicación geográfica.

Calculando el VIF para las variables numéricas del modelos:

Variable	**VIF**
latitud	1.14
longitud	1.14
personas	4.95
baños	1.36
habitaciones	3.69
camas	4.97
estancia_mín.	1.01

Cuadro 4: Tabla VIF

Todos los valores del VIF resultaron ser menores a 5, se concluye que el modelo cumple con el supuesto de no multicolinealidad.

7.2. Linealidad

Se evalúa el cumplimiento del supuesto de linealidad, el cual establece que debe existir una relación lineal entre cada variable explicativa y la variable respuesta. Este supuesto es fundamental en los modelos de regresión lineal, ya que garantiza que el modelo capte adecuadamente las tendencias presentes en los datos. Para su verificación, se analizan los gráficos de residuos versus predichos a nivel global, y de residuos parciales versus cada variable independientea nivel individual, lo que permite detectar posibles patrones no lineales o estructuras sistemáticas que indicarían violaciones al supuesto.

7.2.1. Modelo con distritos

Adjuntamos el grafico de residuos contra predichos para las variables numéricas del módelo de distrítos. El primero es el original, el segundo extrayendole el dato atípico de menor precio, y en el tercero extrayendole los 5 menores datos atípicos. Vemos que al ir extrayendole datos atípos el modelo se vuelve cada vez más lineal, por lo que vamos a tomar esta última versión, ya que considerarmos que sacar 5 observaciones en más de 16000 no es significativo, más cuando aprta tanto a la linealidad del modelo. A nuestros efectos la linealidad del modelo se cumple.

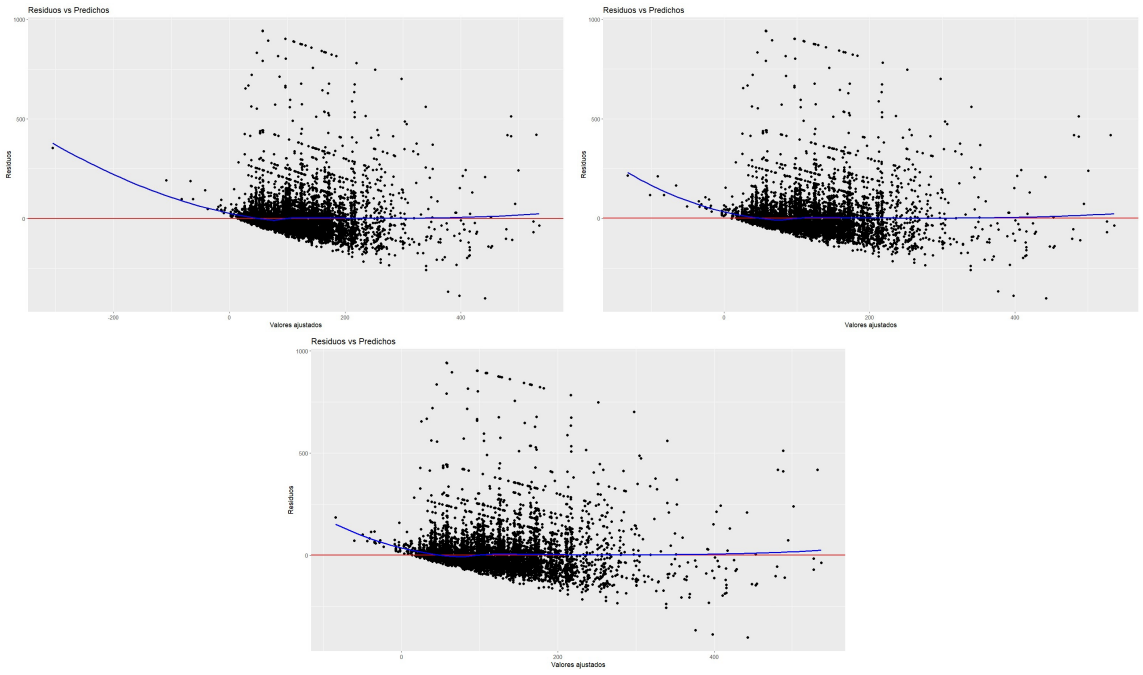

Figura 2: Gráficos Predichos contra Residuos para evaluación de linealidad

A su vez, todas las indagaciones sobre la linealidad de cada una de las variables explicativas son satisfactorias: cumplen con la linealidad. Tal vez podríamos reflexionar sobre *estancia_min* por su leve pendiente ascendente, sin embargo eso se da muy posiblemente por conjunto de datos atípicos que terminan siendo levemente influyentes. No consideramos que esa leve pendiente rompa la linealidad.

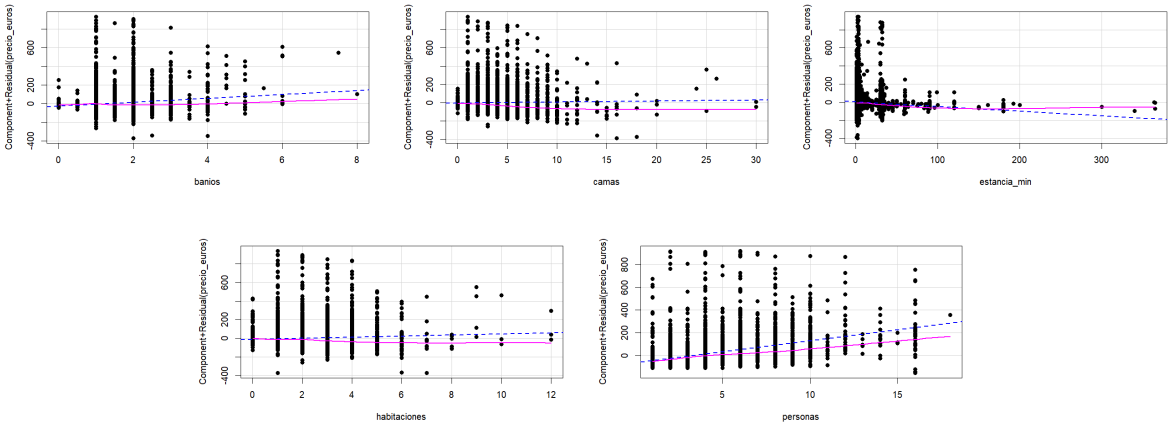

Figura 3: Gráficos de cada variable contra Residuos Parciales para evaluación de linealidad

7.2.2. Modelo con Latitud y Longitud

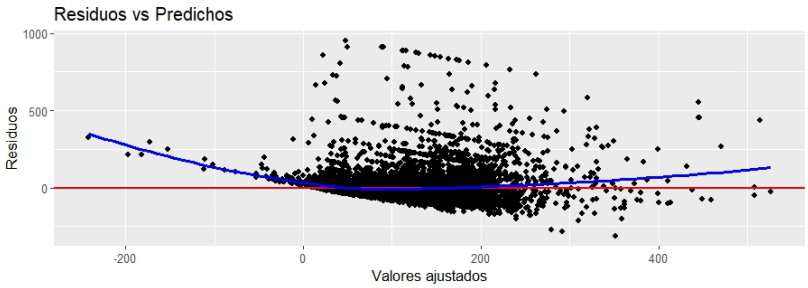

Figura 4: Residuos vs Predichos

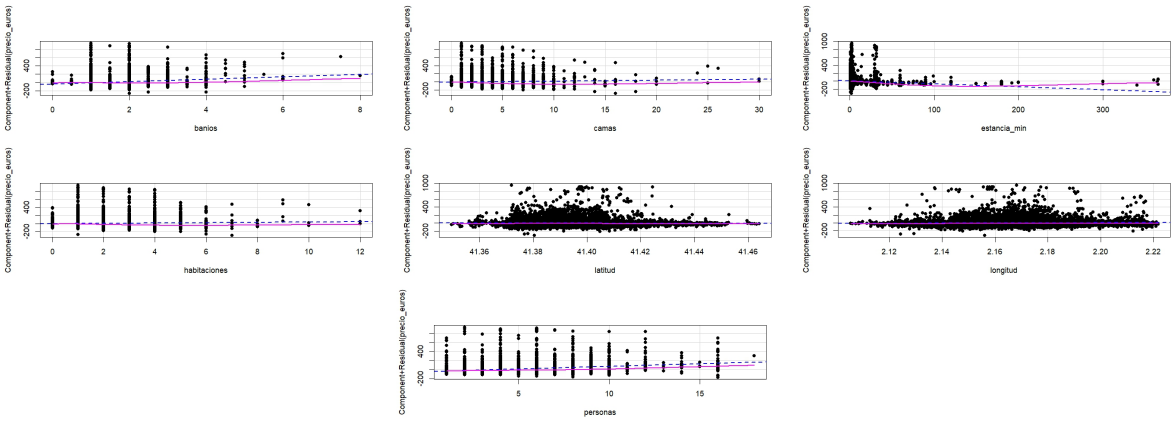

Figura 5: Gráficos Component + Residual para evaluación de linealidad

Del análisis gráfico individual de las variables explicativas mediante gráficos de residuos parciales, se observa que la mayoría de las variables presentan una relación aproximadamente lineal con la variable respuesta (*precio_euros*). En particular, las variables *banios*, *camas*, *habitaciones*, *personas*, *latitud* y *longitud* respetan razonablemente bien el supuesto de linealidad.

La variable *estancia_min* muestra una leve pendiente ascendente, sin embargo eso se da muy posiblemente por conjunto de datos atípicos que terminan siendo levemente influyentes. No consideramos que esa leve pendiente rompa la linealidad.

Por otro lado, el gráfico de residuos versus valores ajustados indica una posible violación del supuesto de linealidad global del modelo, evidenciado por una forma en "U" en la dispersión de los residuos. Esto sugiere que, si bien las relaciones individuales pueden ser lineales, el modelo en su conjunto no está capturando adecuadamente todas las relaciones entre variables, lo que motiva a una a transformacion de variables.

7.3. Homoscedasticidad

En esta sección se discutirá lo siguiente:

$$H_0) : \sigma_i^2 = \sigma^2, \ \forall i = 1, ..., n$$
$$H_1) : \text{No } H_0$$

Para evaluar este supuesto se hará uso del test de Breusch-Pagan, que se basa en los siguientes pasos:

- Se supone que la heterocedasticidad se debe a todas las variables del modelo.
- Se estima el modelo $Y = X\beta + \varepsilon$, obteniendo los residuos $\hat{\varepsilon}$ y el estimador de la varianza.
- Se ajusta el modelo $\frac{\hat{\varepsilon}^2}{\hat{\sigma}^2} = X\alpha + \mu$.
- Se define el estadístico $BP = \frac{SCE}{2}$, que bajo H_0 sigue una distribución χ_k^2.
- Se realiza el test contrastando la hipótesis nula: Queremos no rechazar.

7.3.1. Modelo con distritos

Vemos como se rechaza la hipótesis de que la varianza es constante por lo que el modelo no es homocedastico.

Estadístico	**p-valor**	**df**	**Método**
409.16	$4{,}38 \times 10^{-77}$	16	Koenker (studentised)

Cuadro 5: Resultados del test de Koenker (studentised) para heterocedasticidad

7.3.2. Modelo con Latitud y Longitud

Los resultados obtenidos en R utilizando el paquete `lmtest` con la función `bptest` son los siguientes:

Estadístico	**p-valor**	**Método**
368.00	$8{,}75 \times 10^{-74}$	Koenker (studentised)

Cuadro 6: Test de Koenker para heterocedasticidad

Dado que el p-valor es considerablemente inferior al nivel de significancia habitual ($\alpha = 0{,}05$), se rechaza la hipótesis nula de homocedasticidad.

7.4. Normalidad de los residuos

El último supuesto a corroborar es el de la distribución normal de los errores. Si bien este supuesto es de alta relevancia, podemos trabajar a pesar de no cumplirse en presencia de grandes cantidades de datos, como es el caso nuestro. Igualmente realizaremos las pruebas correspondientes.

7.4.1. Modelo con Distritos

Tras realizar distintos contrastes de hipótesis en los cuales buscamos no rechazar, vemos que se va el caso contrario. No podemos afirmar distribución normal de los errores.

Estadístico	p-valor	Método	Hipótesis alternativa
0.208	$< 2{,}2 \times 10^{-16}$	Kolmogorov-Smirnov	two-sided
1310.2	$< 2{,}2 \times 10^{-16}$	Anderson-Darling	no normalidad

Cuadro 7: Resultados de los tests de normalidad

Sin embargo, es importante recordar que dado la cantidad de datos no nos preocuparemos en demasía por esto.

Adicionalmente agregamos el QQ-Plot asociado. El mismo muestra que los residuos se comportan casi normales en el centro, pero presentan desviaciones significativas en las colas, especialmente en la derecha, indicando presencia de outliers o colas pesadas.

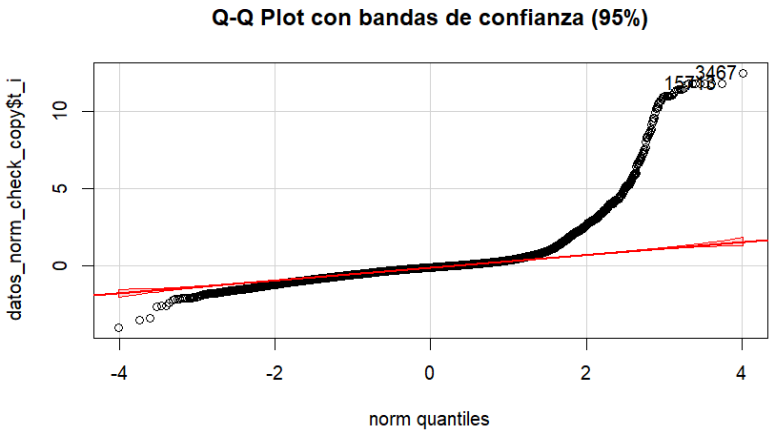

Figura 6: Q-Q-Plot

7.4.2. Modelo con Latitud y Longitud

Para ello se recurrió tanto a herramientas gráficas como a pruebas estadísticas. En primer lugar, se construyó un gráfico Q-Q que compara los cuantiles de los residuos estudentizados del modelo con los cuantiles teóricos de una distribución normal. El gráfico evidencia desviaciones notorias, en especial en las colas, lo cual sugiere una violación del supuesto de normalidad.

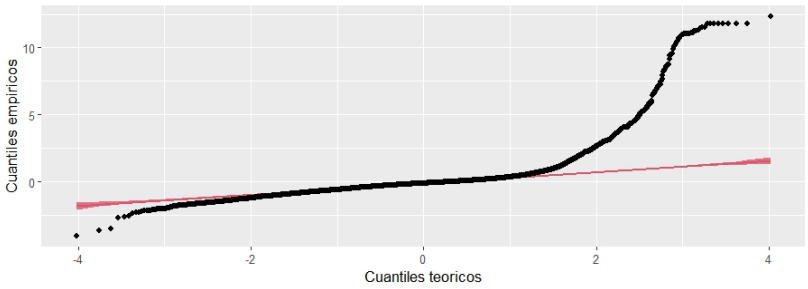

Figura 7: Q-Q-Plot

Para complementar el análisis visual, se realizaron dos pruebas estadísticas: Jarque-Bera y Kolmogorov-Smirnov, ambas aplicadas sobre los residuos estudentizados. En ambos casos, los p-valores fueron extremadamente bajos ($p < 0{,}01$), permitiendo rechazar la hipótesis nula de normalidad de los errores.

Estadístico	**df**	**p-valor**	**Método**
1 281 054	2	$< 2{,}2 \times 10^{-16}$	Jarque-Bera
0.20882	–	$< 2{,}2 \times 10^{-16}$	Kolmogorov-Smirnov

Cuadro 8: Resultados de tests de normalidad

Ambas pruebas conducen a rechazar la hipótesis nula de normalidad, confirmando que los residuos no se distribuyen normalmente. Sin embargo, la cantidad de observaciones nos hace despreocuparnos de este hecho.

8. Conclusiones de ambos Modelos

Propiedad	**Modelo de Distritos**	**Modelo de Lat-Long Original**
No Multicolinealidad	Cumple el supuesto	Cumple, leve correlación entre algunas variables
Linealidad	Cumple tanto global como individualmente	Cumple a nivel individual pero no global
Homocedasticidad	No	No
Normalidad	No, pero no preocupa	No, pero no preocupa

Cuadro 9: Comparación entre el Modelo de Distritos y el Modelo de Lat-Long

9. Modelo Final Latitud y Longitud

Con el objetivo de mejorar el ajuste del modelo y abordar posibles incumplimientos de los supuestos clásicos de la regresión lineal, se implementó una transformación logarítmica total sobre las variables numéricas tanto explicativas como la variable respuesta. Esta estrategia busca estabilizar la varianza, reducir la asimetría de las distribuciones y mejorar la linealidad entre las variables.

En primer lugar, se filtraron observaciones que presentaran valores no positivos en las variables a transformar, dado que el logaritmo natural no está definido para dichos casos. Posteriormente, se aplicó la transformación logarítmica a todas las variables relevantes, agregando una unidad a aquellas que podían tomar el valor cero, como *banios*, *habitaciones*, *camas* y *estancia_min*, para evitar valores indefinidos.

A partir del nuevo conjunto de datos transformado, se ajustó un modelo de regresión lineal múltiple considerando como variable dependiente el logaritmo del precio (*log_precios_euro*) y como regresores los logaritmos de las variables cuantitativas, excluyendo las variables categóricas *barrio* y *tipo_habitacion*, con el fin de concentrarse en las relaciones lineales entre variables numéricas.

9.1. Multicolinealidad en el modelo transformado

Una vez ajustado el modelo con las variables transformadas logarítmicamente, se procedió a evaluar el cumplimiento del supuesto de no multicolinealidad mediante el cálculo del *Variance Inflation Factor* (VIF). Inicialmente, se excluyeron del modelo las variables categóricas *tipo_habitacion* con el fin de concentrar el análisis en las variables numéricas transformadas.

Los resultados obtenidos se presentan en el siguiente cuadro:

Variable	**VIF**
log_latitud	1.22
log_longitud	1.18
log_personas	3.44
log_banios	1.33
log_habitaciones	2.92
log_camas	4.34
log_estancia_min	1.24

Cuadro 10: Cuadro: Valores del VIF para el modelo logarítmico sin variables categóricas

Como se observa, todas las variables presentan un VIF inferior a 5, se considera que el modelo transformado en escala logarítmica no presenta un problema grave de multicolinealidad.

Posteriormente, se reincorporó la variable categórica *tipo_habitacion* al modelo completo para los análisis posteriores, dado que su exclusión fue únicamente con fines diagnósticos.

9.2. Supuesto de linealidad en el modelo transformado

En esta sección se evalúa el cumplimiento del supuesto de linealidad para el modelo ajustado con las variables transformadas logarítmicamente. Se analizan los gráficos *Component + Residual* (también conocidos como *Partial Residual Plots*) para cada una de las variables explicativas, así como el gráfico de residuos versus valores ajustados.

Los gráficos muestran que la mayoría de las relaciones entre las variables transformadas y la respuesta logarítmica del precio se alinean adecuadamente con la suposición de linealidad. En

particular, variables como *log_latitud*, *log_longitud*, *log_banios*, *log_habitaciones*, y *log_camas* presentan trayectorias suaves y aproximadamente lineales.

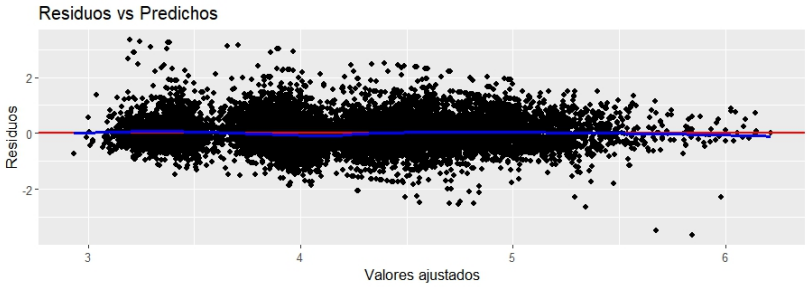

Figura 8: Residuos vs Predichos

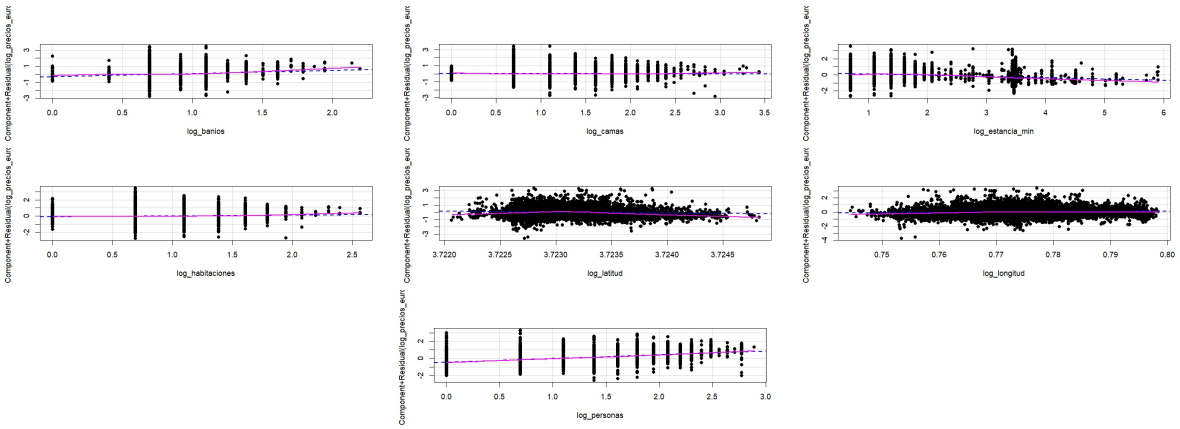

Figura 9: Gráficos Component + Residual para evaluación de linealidad

Conclusión: Con base en los gráficos analizados, se concluye que la transformación logarítmica aplicada mejora sustancialmente el cumplimiento del supuesto de linealidad. Aunque se observan ligeras desviaciones en algunas variables como *log_estancia_min* y *log_personas*, estas no parecen comprometer gravemente la validez del modelo lineal ajustado.

9.3. Homoscedasticidad en el modelo transformado

En el análisis del modelo transformado se aplicó nuevamente el test de **Breusch-Pagan** para evaluar el cumplimiento del supuesto de homocedasticidad. El resultado arrojó un valor p de $2{,}87 \times 10^{-26}$, indicando un claro rechazo de la hipótesis nula.

Conclusión: Al igual que en el modelo original, el modelo con variables logarítmicas *no cumple* con el supuesto de homocedasticidad. A pesar de la transformación aplicada, persiste la presencia de heterocedasticidad en los residuos.

9.4. Normalidad en el modelo transformado

A pesar de haber transformado las variables del modelo utilizando logaritmos, se realizó nuevamente la evaluación del supuesto de normalidad de los residuos. Para ello, se aplicaron los tests de **Jarque-Bera** y **Kolmogorov-Smirnov** sobre los residuos studentizados externamente.

Ambas pruebas arrojaron valores p extremadamente bajos ($< 2{,}2 \times 10^{-16}$), lo cual conduce al rechazo formal de la hipótesis nula de normalidad.

Sin embargo, dado que el conjunto de datos cuenta con más de 15.000 observaciones, se debe considerar que estas pruebas tienden a ser muy sensibles a pequeñas desviaciones respecto a la distribución normal. Por esta razón, si bien el modelo no cumple estrictamente con el supuesto de normalidad, los residuos muestran un comportamiento razonablemente cercano a la normalidad, por lo que en la práctica puede considerarse una *buena aproximación.*

Conclusión: Aunque se rechaza formalmente la normalidad de los errores, la alta cantidad de datos permite continuar con el análisis sin que esto represente una limitación crítica para la validez del modelo.

9.5. Observaciones Atípicas e Influencia

En esta sección se identifican posibles observaciones atípicas y puntos influyentes dentro del modelo ajustado, a partir del análisis de los residuos studentizados y de las medidas de influencia.

Se construyó un gráfico de residuos studentizados frente al número de observación. Las observaciones que superan en valor absoluto el umbral de 3 (i.e. $|t_i| > 3$) son consideradas potencialmente atípicas y fueron destacadas en color naranja, mientras que las restantes se presentan en púrpura. Este criterio permite visualizar valores que se desvían significativamente de la tendencia general del modelo.

Asimismo, se calculó la *distancia de Cook* (D_i), que cuantifica la influencia de cada observación sobre los coeficientes estimados. Para su análisis se utilizó el umbral práctico de $4/n$, siendo n el número total de observaciones. En el gráfico correspondiente se muestra dicho umbral en línea punteada roja. Las observaciones con valores elevados de D_i podrían tener una influencia desproporcionada en los resultados del modelo, por lo que se recomienda su análisis detallado.

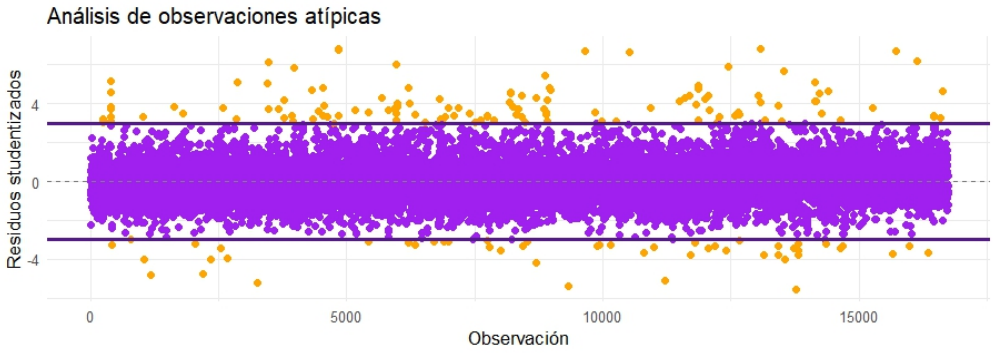

Figura 10: Residuos studentizados por observación

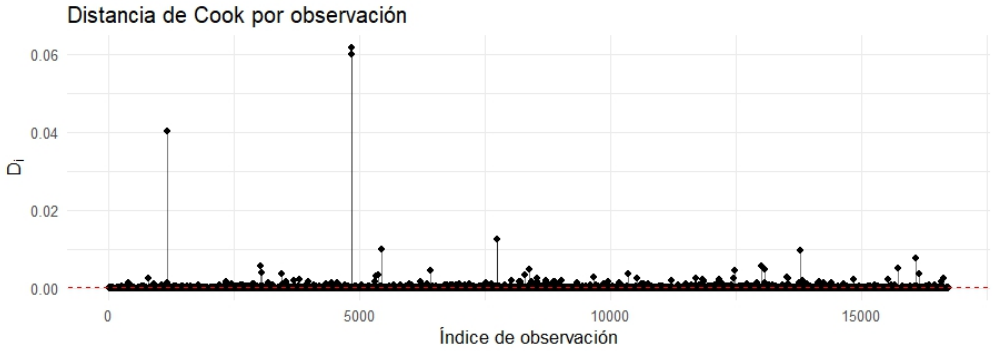

Figura 11: Distancia de Cook por observación

Conclusión: Si bien se detectan algunas observaciones con residuos extremos y valores relativamente elevados de distancia de Cook, no se evidencia un patrón sistemático de observaciones altamente influyentes. Esto sugiere que, en general, el modelo es robusto frente a observaciones individuales. Por lo tanto, se decide **no eliminar ninguna observación**, dado que el tamaño de la muestra es suficientemente grande y la influencia detectada no compromete significativamente la validez global del modelo.

10. Estimación robusta ante heterocedasticidad

Dado que ninguno de los dos modelos logró cumplir con el supuesto de homocedasticidad —ni el modelo de Latitud y Longitud ni el de Distritos—, se decidió utilizar estimadores robustos mediante la librería `sandwich`. Esta herramienta permite ajustar la matriz de varianzas y covarianzas de los coeficientes estimados, aplicando el estimador de varianza heterocedástica consistente de White (también conocido como estimador *sandwich*), lo que proporciona inferencias más confiables frente a la presencia de heterocedasticidad.

El objetivo de este procedimiento es corregir los errores estándar de los coeficientes cuando existe heterocedasticidad, de forma que las pruebas de hipótesis y los intervalos de confianza

resulten válidos. Específicamente, se utilizó la opción `HC0`, que no realiza correcciones de grados de libertad, proporcionando así una estimación robusta de la varianza de los coeficientes.

En el caso del modelo de Distritos:

```
coeftest(mod_limpio, vcov = vcovHC(mod_limpio, type = "HC1"))
```

En el caso del modelo de Latitud y Longitud:

```
coeftest(log_mod, vcov = vcovHC(log_mod, "HC0"))
```

Este enfoque no modifica los coeficientes del modelo, sino únicamente la forma en que se calcula su varianza, permitiendo realizar inferencia más fiable en presencia de heterocedasticidad.

A continuación se presentan los resultados obtenidos al aplicar esta estimación:

10.1. Modelo de Distritos con estimación robusta

Cuadro 11: Resumen del modelo lineal

Variable	**Estimación**	**Error Std.**	**t valor**	**p-valor**
(Intercept)	31.64776	4.47164	7.0774	$1{,}53 \times 10^{-12}$ ***
personas	11.69175	1.04178	11.2229	$< 2{,}2 \times 10^{-16}$ ***
banios	26.89195	1.99840	13.4567	$< 2{,}2 \times 10^{-16}$ ***
habitaciones	3.63959	1.47940	2.4602	0.013897 *
camas	2.52452	1.19389	2.1145	0.034484 *
estancia_min	-0.61979	0.11348	-5.4617	$4{,}79 \times 10^{-8}$ ***
tipo_habitacionPrivate room	-40.8391	2.68989	-15.182	$< 2{,}2 \times 10^{-16}$ ***
tipo_habitacionShared room	-36.6482	13.9730	-2.6228	0.008729 **
distritosEixample	6.64608	1.55983	4.2608	$2{,}05 \times 10^{-5}$ ***
distritosGràcia	2.34573	2.86179	0.8197	0.412415
distritosHorta-Guinardó	0.10614	4.08237	0.0260	0.979257
distritosLes Corts	-9.88934	4.08254	-2.4224	0.015431 *
distritosNou Barris	-9.76152	3.00920	-3.2439	0.001181 **
distritosSant Andreu	-15.2030	2.68169	-5.6692	$1{,}46 \times 10^{-8}$ ***
distritosSant Martí	2.45140	2.22810	1.1002	0.271252
distritosSants-Montjuïc	-7.57899	1.83897	-4.1213	$3{,}78 \times 10^{-5}$ ***
distritosSarrià-Sant Gervasi	3.50145	4.34586	0.8057	0.420429

10.2. Modelo de Latitud y Longitud con estimación robusta

Variable	Estimación	Error Std.	t valor	p-valor
(Intercepto)	443.31	43.76	10.13	$< 2e^{-16}$
log_latitud	−119,16	11.79	−10,11	$< 2e^{-16}$
log_longitud	4.40	0.52	8.41	$< 2e^{-16}$
log_personas	4.34	0.47	9.91	$< 2e^{-16}$
log_banios	3.96	0.24	16.46	$< 2e^{-16}$
log_habitaciones	2.02	0.40	5.04	$6{,}41e^{-7}$
log_camas	−0,93	1.94	−0,48	0.633
log_estancia_min	−1,51	0.33	−4,53	$< 2e^{-16}$
tipo_habitacionEntire	0.82	0.10	8.45	$< 2e^{-16}$
tipo_habitacionPrivate	0.26	0.10	2.74	0.006

Cuadro 12: Coeficientes estimados con errores robustos (HC0)

Como puede observarse, la mayoría de las variables incluidas en el modelo resultan estadísticamente significativas al nivel del 5 %. La única excepción es la variable `log_camas`, cuyo valor p es 0.633, lo cual sugiere que no tiene un efecto significativo sobre el logaritmo del precio de alquiler.

Por lo tanto, se procederá a eliminar esta variable del modelo en la siguiente etapa del análisis, con el objetivo de obtener una especificación más parsimoniosa y precisa.

11. Modelos finales y conclusiones

11.1. Distritos

Luego de realizar las transformaciones necesarias y descartar observaciones atípicas, se estimó un modelo de regresión lineal múltiple utilizando la variable dependiente `precio_euros` y una serie de predictores que describen características estructurales, tipo de habitación y localización geográfica del alojamiento.

Los resultados del modelo se presentan en el siguiente Cuadro. El ajuste global del modelo fue el siguiente:

- **Error estándar residual**: 77.49
- **R-cuadrado**: 0.3518
- **R-cuadrado ajustado**: 0.3512
- **Estadístico F**: 567.1 (gl = 16 y 16717), *p-valor* ¡2.2e^{-16}

Esto implica que el modelo logra explicar aproximadamente el 35.2 % de la variabilidad en el precio en euros, lo cual es moderadamente aceptable dado el contexto heterogéneo de la variable.

Cuadro 13: Resumen del modelo lineal

Variable	Estimación	Error Std.	t valor	p-valor
(Intercept)	31.64776	4.47164	7.0774	$1{,}53 \times 10^{-12}$ ***
personas	11.69175	1.04178	11.2229	$< 2{,}2 \times 10^{-16}$ ***
banios	26.89195	1.99840	13.4567	$< 2{,}2 \times 10^{-16}$ ***
habitaciones	3.63959	1.47940	2.4602	0.013897 *
camas	2.52452	1.19389	2.1145	0.034484 *
estancia_min	-0.61979	0.11348	-5.4617	$4{,}79 \times 10^{-8}$ ***
tipo_habitacionPrivate room	-40.8391	2.68989	-15.182	$< 2{,}2 \times 10^{-16}$ ***
tipo_habitacionShared room	-36.6482	13.9730	-2.6228	0.008729 **
distritosEixample	6.64608	1.55983	4.2608	$2{,}05 \times 10^{-5}$ ***
distritosGràcia	2.34573	2.86179	0.8197	0.412415
distritosHorta-Guinardó	0.10614	4.08237	0.0260	0.979257
distritosLes Corts	-9.88934	4.08254	-2.4224	0.015431 *
distritosNou Barris	-9.76152	3.00920	-3.2439	0.001181 **
distritosSant Andreu	-15.2030	2.68169	-5.6692	$1{,}46 \times 10^{-8}$ ***
distritosSant Martí	2.45140	2.22810	1.1002	0.271252
distritosSants-Montjuïc	-7.57899	1.83897	-4.1213	$3{,}78 \times 10^{-5}$ ***
distritosSarrià-Sant Gervasi	3.50145	4.34586	0.8057	0.420429

Interpretaciones destacadas de los coeficientes

- **personas**: Un incremento de una persona adicional en la capacidad del alojamiento se asocia con un aumento promedio de 11.69 euros en el precio, manteniendo constantes las demás variables.
- **baños**: Cada baño adicional eleva el precio estimado en 26.89 euros, lo cual refleja un impacto significativo y positivo.
- **habitaciones y camas**: Ambas variables también tienen efectos positivos y estadísticamente significativos sobre el precio.
- **estancia mínima**: Su coeficiente negativo sugiere que cuanto mayor sea el mínimo de noches requeridas, menor será el precio. Un día más de estancia mínima reduce el precio en promedio 0.62 euros, indicando menor flexibilidad.
- **tipo de habitación**:
  - Las *Private room* tienen un precio 40.84 euros menor que la referencia (*Entire home/apt*).
  - Las *Shared room* son 36.65 euros más baratas en promedio, aunque esta estimación tiene mayor error estándar.
- **Ubicación**:
  - Los distritos como *Eixample* presentan una prima positiva de 6.65 euros respecto al distrito base.

- En cambio, zonas como *Sant Andreu*, *Les Corts* y *Sants-Montjuïc* muestran coeficientes negativos, lo que sugiere precios más bajos en esas zonas.
- Varios coeficientes asociados a barrios no resultaron significativos, lo que indica que su influencia puede ser menor o más difícil de captar linealmente.

Considerando el Anova del modelo para corrobar la significación de cada variable a nivel global:

Cuadro 14: Tabla ANOVA tipo II para el modelo con *precio_euros* como variable respuesta

Variable	**Suma de Cuadrados**	**gl**	**F**	**Pr($>F$)**
camas	72 730	1	12.19	0.0004813 ***
baños	2 488 576	1	417.16	<2.2e-16 ***
estancia_min	2 588 308	1	433.88	<2.2e-16 ***
habitaciones	73 272	1	12.28	0.0004584 ***
distritos	449 292	9	8.37	1.469e-12 ***
tipo_habitacion	3 692 905	2	309.52	<2.2e-16 ***
personas	1 471 046	1	246.59	<2.2e-16 ***
Residuals	99 695 929	16 712	—	—

Vemos que todas las variables son significativas en su conjunto para el modelo. También llama la atención el alto número asignado a los residuos del modelo, es decir la variación del precio de alquiler no explicada por el modelo. Sin embargo hay que tener en cuenta que como ya vimos en el summary, este modelo solo llega a explicar el 35 % por lo que termina teniendo sentido.

Conclusión

El modelo estimado ofrece una explicación razonable de los precios de alojamientos en Barcelona según variables estructurales y de ubicación. Si bien el modelo es intuitivo y fácil de interpretar, vemos que el modelo no explica gran parte de la variabilidad del precio de alquileres. Ganamos explicabilidad perdemos exactitud.

11.2. Latitud y Longitud

Luego de eliminar la variable `log_camas`, que no resultó significativa en el modelo anterior, se estimó nuevamente el modelo utilizando la función `lm()` y se aplicó la corrección robusta de errores estándar con `sandwich`. Los resultados obtenidos se presentan a continuación:

Variable	Estimación	Error Std. (HC0)	t valor	p-valor
(Intercepto)	443.98	43.69	10.16	$< 2e^{-16}$
log_latitud	−119,34	11.77	−10,14	$< 2e^{-16}$
log_longitud	4.40	0.52	8.41	$< 2e^{-16}$
log_personas	4.31	0.13	32.85	$< 2e^{-16}$
log_banios	3.83	0.22	17.54	$< 2e^{-16}$
log_habitaciones	2.97	0.54	5.45	$5{,}12e^{-08}$
log_estancia_min	−1,54	0.33	−33,35	$< 2e^{-16}$
tipo_habitacionEntire	0.82	0.10	8.45	$< 2e^{-16}$
tipo_habitacionPrivate	0.26	0.10	2.76	0,0058

Cuadro 15: Modelo final con estimación robusta (HC0)

Además, el resumen del modelo arroja los siguientes valores de ajuste global:

- **Error estándar residual**: 0.504
- **R-cuadrado**: 0.5507
- **R-cuadrado ajustado**: 0.5505
- **Estadístico F**: 2563.0 (gl = 8 y 16724), **p-valor** $< 2{,}2e^{-16}$

Estos resultados indican que el modelo explica aproximadamente un 55 % de la variabilidad del logaritmo del precio en euros, lo cual se considera razonable dadas las características heterogéneas de los alojamientos analizados.

Interpretaciones destacadas de los coeficientes

- **log_personas**: Un aumento del 1 % en la cantidad de personas que puede alojar el inmueble se asocia con un incremento aproximado del 0.043 en el logaritmo del precio, manteniendo las demás variables constantes.
- **log_banios**: Un 1 % más en la cantidad de baños se asocia con un aumento del 0.0383 en el log del precio, lo que refleja un impacto positivo y significativo.
- **log_habitaciones**: También tiene un efecto positivo: un 1 % de aumento se traduce en un incremento aproximado del 0.0297 en el logaritmo del precio.
- **log_estancia_min**: Su coeficiente negativo indica que a mayor duración mínima exigida por el anfitrión, menor es el precio del alojamiento. Es decir, un 1 % de aumento en este valor implica una reducción del 0.0154 en el log del precio.
- **tipo_habitacion**: Considerando como referencia la categoría “Shared room” (habitación compartida), se observa que:
  - “Entire home/apt” incrementa significativamente el log del precio en 0.82 unidades.

- "Private room" lo incrementa en 0.26 unidades.

- **log_latitud y log_longitud**: A pesar de ser estadísticamente significativas, sus interpretaciones directas no son triviales debido a la escala logarítmica y la complejidad geográfica involucrada.

P ara el caso de log_latitud y log_longitud, si bien ambas variables resultan estadísticamente significativas, su interpretación directa no es sencilla debido a su transformación logarítmica y a la complejidad espacial de la ciudad de Barcelona. Por ello, se optó por realizar un análisis complementario mediante un mapa georreferenciado que relaciona la ubicación de los alojamientos con el nivel socioeconómico de los barrios.

Para ello, se clasificaron los barrios de Barcelona en tres niveles socioeconómicos: *Alto*, *Medio* y *Bajo*, en base a información proveniente del Instituto de Estadística de Cataluña. Esta clasificación se implementó en R mediante vectores predefinidos que agrupan los nombres de los barrios según su nivel de ingreso relativo. Posteriormente, se creó una nueva variable categórica (`nivel_ingreso`) dentro del dataset original, asignando a cada observación su categoría correspondiente.

Luego, se utilizó el paquete `ggmap` para obtener un mapa base de la ciudad de Barcelona, y se superpusieron los puntos de los alojamientos en función de sus coordenadas (`latitud` y `longitud`) y su nivel socioeconómico. La codificación de colores se realizó de la siguiente manera: verde para barrios de alto ingreso, dorado para ingreso medio y rojo para ingreso bajo.

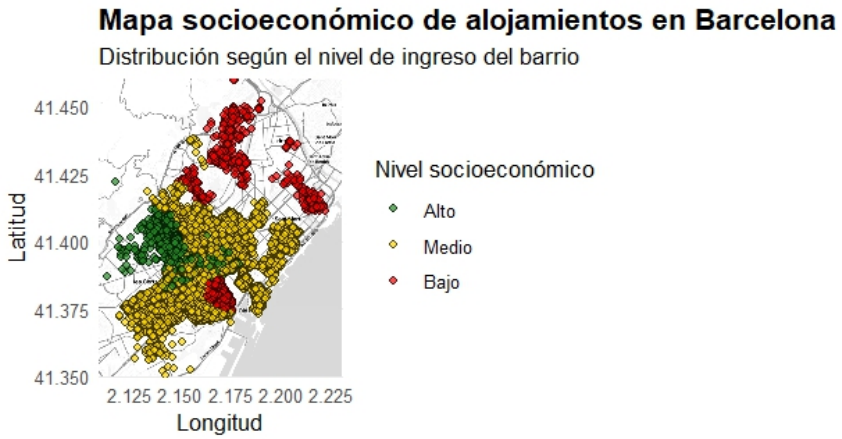

Figura 12: Mapa socioeconómico de alojamientos en Barcelona según ubicación

En el gráfico se observa cómo los alojamientos ubicados en zonas de nivel socioeconómico alto (en verde) se concentran en determinadas áreas del norte y del oeste de la ciudad, mientras que los de nivel bajo (en rojo) se localizan principalmente en la franja este y noreste. Esta

distribución geográfica permite contextualizar el efecto significativo de la latitud y longitud sobre el precio, ya que actúan como *proxies* espaciales del nivel de ingreso del barrio.

Conclusión

El modelo final presenta un ajuste razonable y una especificación robusta, siendo capaz de explicar una parte sustancial de la variación del precio de los alojamientos en función de características estructurales, ubicación y tipo de habitación. Las estimaciones robustas permiten además asegurar inferencias confiables a pesar de la presencia de heterocedasticidad. Ganamos precisión y conseguimos explicabilidad, sin embargo tuvimos que complejizar significativamente el uso de las variables utilizando transformaciones lineales para hacerlo funcionar.

Bibliografía y paquetes utilizados

El análisis fue realizado íntegramente en el lenguaje de programación `R`, utilizando una variedad de paquetes especializados en estadística, visualización y diagnóstico de modelos lineales. A continuación, se listan las librerías empleadas, especificando aquellas que también se utilizaron en el modelo con clasificación de barrios por nivel socioeconómico (modelo "distrito").

- **readxl** [1]: Permite importar datos desde archivos Excel. Utilizado para cargar la base original de Airbnb. También fue empleado en el modelo distrito.
- **dplyr** [2]: Paquete central del tidyverse para manipulación eficiente de datos (filtrado, mutación, agrupación). También fue empleado en el modelo distrito.
- **tidyverse** [3]: Conjunto de paquetes (incluye `ggplot2`, `dplyr`, etc.). Utilizado como entorno base. También aplicado al modelo distrito.
- **ggplot2** [4]: Librería para gráficos estadísticos personalizados. Fundamental para visualizaciones de residuos, normalidad, e influencia. También utilizada para generar el mapa socioeconómico en el modelo distrito.
- **ggmap** [5]: Utilizada para descargar mapas base y crear visualizaciones georreferenciadas. Fundamental para el análisis espacial del modelo distrito.
- **car** [6]: Herramientas para diagnóstico de regresión: `vif`, `crPlot`, etc. Utilizada también en el modelo distrito.
- **corrplot** [7]: Gráficos de matrices de correlación para análisis exploratorio de relaciones entre variables numéricas.
- **lmtest** [8]: Permite aplicar pruebas estadísticas sobre modelos, como `coeftest`. También utilizado en el modelo distrito.
- **sandwich** [9]: Estimadores robustos de varianzas-covarianzas (tipo `HC0`) ante heterocedasticidad. También usado en el modelo distrito.

- **tseries** [10]: Contiene la prueba de Jarque-Bera y Kolmogorov-Smirnov para contrastar normalidad de residuos.
- **qqplotr** [11]: Complementa a `ggplot2` para construir gráficos Q-Q con bandas de confianza.
- **skedastic** [12]: Permite aplicar el test de Breusch-Pagan para detectar heterocedasticidad.
- **leaps** [13]: Algoritmos de selección de variables mediante regresión escalonada. Utilizado en pruebas complementarias.
- **faraway** [14]: Funciones útiles para modelado y análisis de regresión.
- **HH** [15]: Herramientas para análisis visual y comparación de modelos. Se exploraron algunos gráficos de diagnóstico.
- **glmnet** [16]: Proporciona técnicas de regularización como LASSO y Ridge. Aplicado en fases exploratorias del proyecto.
- **nortest** [17]: Contiene pruebas complementarias de normalidad, como Anderson-Darling.

Clasificación socioeconómica de barrios. Para asignar a cada barrio un nivel socioeconómico (alto, medio o bajo), se utilizó información del **Instituto de Estadística de Cataluña (Idescat)**, específicamente del indicador de renta media por distrito para la ciudad de Barcelona. La fuente oficial es:

- **Idescat - Indicadores socioeconómicos territoriales:** `https://www.idescat.cat/pub/?id=ist&n=14075&geo=mun:080193&lang=es`

La información fue utilizada para agrupar los barrios del dataset de Airbnb según su nivel de ingreso relativo, permitiendo generar el mapa georreferenciado incluido en el análisis.

Referencias

[1] Hadley Wickham and Jennifer Bryan. *readxl: Read Excel Files*. R package version 1.4.3.

[2] Hadley Wickham, Romain François, Lionel Henry, Kirill Müller. *dplyr: A Grammar of Data Manipulation*. R package version 1.1.4.

[3] Hadley Wickham. *tidyverse: Easily Install and Load the Tidyverse*. R package version 2.0.0.

[4] Hadley Wickham. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York, 2016.

[5] David Kahle and Hadley Wickham. *ggmap: Spatial Visualization with ggplot2*. R Journal, 5(1), 2013.

[6] John Fox and Sanford Weisberg. *An R Companion to Applied Regression*. Third Edition. Sage, 2019.

[7] Taiyun Wei and Viliam Simko. *corrplot: Visualization of a Correlation Matrix*. R package version 0.92.

[8] Achim Zeileis and Torsten Hothorn. *Diagnostic Checking in Regression Relationships*. R News 2(3), 2002.

[9] Achim Zeileis. *sandwich: Robust Covariance Matrix Estimators*. R package version 3.0-0.

[10] Adrian Trapletti and Kurt Hornik. *tseries: Time Series Analysis and Computational Finance*. R package version 0.10-54.

[11] Jakub Głowacki. *qqplotr: Quantile-Quantile Plot Extensions for 'ggplot2'*. R package version 0.0.6.

[12] Matthew Lovett. *skedastic: Heteroskedasticity Diagnostics for Linear Models*. R package version 1.0.0.

[13] Trevor Hastie, Robert Tibshirani, Jerome Friedman. *leaps: Regression Subset Selection*. R package version 3.1.

[14] Julian Faraway. *Linear Models with R*. Second Edition. Chapman and Hall/CRC, 2014.

[15] Richard Heiberger. *HH: Statistical Analysis and Data Display*. R package version 3.1-49.

[16] Jerome Friedman, Trevor Hastie, Rob Tibshirani. *Regularization Paths for Generalized Linear Models via Coordinate Descent*. Journal of Statistical Software, 33(1), 2010.

[17] Gross, J. and Ligges, U. *nortest: Tests for Normality*. R package version 1.0-4.