Universidad de **Valladolid**

Facultad de Ciencias

TRABAJO FIN DE GRADO

Grado en Estadística

CONSTRUCCIÓN DE UN MODELO PARA LA ESTANQUEIDAD DE EDIFICIOS RESIDENCIALES EN ESPAÑA

Autor:
Miguel Sacristán de Frutos

Tutores:
Miguel Alejandro Fernández Temprano
María Pilar Rodríguez del Tío

Agradecimientos

A mis tutores Miguel y Pilar, por su guía y apoyo en este trabajo, y de los que he aprendido mucho tanto en esta ocasión como a lo largo del grado.

A mi familia, en especial a mi abuela, mis padres y mi hermano, que me han acompañado y apoyado durante toda la vida, animándome a seguir adelante y sobreponerme a los retos.

A mis compañeros, sin los que esta etapa no hubiera sido lo mismo, no podría haber tenido un grupo mejor.

Por último a todos los amigos que han estado a mi lado todos estos años, gracias a todos de corazón.

Resumen

La estanqueidad al aire es uno de los factores que más afectan al rendimiento energético de los edificios. Este trabajo tiene como objetivo desarrollar un modelo lineal generalizado que permita identificar y cuantificar las variables más significativas que afectan a la estanqueidad, proporcionando así una herramienta útil para mejorar la eficiencia energética en la construcción y mantenimiento de edificios residenciales. Este modelo será comparado con un modelo lineal ya publicado para analizar si el modelo lineal generalizado supone una mejora. Para ello se utilizará el lenguaje de programación R, y un conjunto de datos representativo de las viviendas construidas en España entre los años 1880 y 2015.

Palabras clave Modelo lineal generalizado, estanqueidad al aire, validación cruzada, análisis estadístico, viviendas

Summary

Airtightness is one of the factors that most affect the energy performance of buildings. This thesis aims to develop a generalized linear model to identify and quantify the most significant variables affecting airtightness, thus providing a useful tool to improve energy efficiency in the construction and maintenance of residential buildings. This model will be compared with a linear model already published to analyze if the generalized linear model is an improvement. For this purpose, the programming language R will be used and a representative dataset of dwellings built in Spain between the years 1880 and 2015.

Keywords Generalized linear model, airtighness, cross validation, statistical analysis, dwellings

Índice general

Índice de figuras

Índice de tablas

Capítulo 1

Introducción

1.1. Contexto

Debido a los nuevos objetivos climáticos europeos, como el Pacto Verde [1] o la agenda 2030, se ha puesto de manifiesto la necesidad de reducir las emisiones y mejorar la eficiencia energética.

Para ello, el rendimiento energético (EP) de los edificios es algo crucial, existiendo unas regulaciones tanto a nivel europeo como nacional que establecen que los edificiones tienen que llegar a ser neutros energéticamente (nZEB). Esto implica un plan de renovación a largo plazo de los edificios existentes que afectará a cerca de 35 millones [2].

Aunque la tasa de renovación de los edificios es muy baja todavía, se estima que en torno al 0.4–1.2 % anual, se necesitan estretegias eficientes para realizar esta transición, pues los edificios existentes en la Unión Europea se espera que sigan en uso hasta 2050 [2]. Estas nuevas tendencias hacen que no solo se preste atención a la reducción de la transmisión de calor a través de la envolvente del edificio si no que la estanqueidad al aire cobre relevancia ya que afecta a la demanda energética para calentar o refrigerar el edificio entre un 10 y un 30 %.

La estanqueidad al aire de los edificios es una variable clave que afecta a su rendimiento energético, confort interior de los edificios (Sherman y Chan, 2004) [3] y es relevante de cara a definir estrategias efectivas de rehabilitación. Medir y analizar la estanqueidad no solo permite identificar fugas de aire y sus fuentes, sino también entender cómo diversas características del edificio influyen en su rendimiento global.

El análisis de estas características cobra especial relevancia cuando disponemos de grandes conjuntos de datos con los que poder desarrollar modelos estadísticos predictivos para la estanqueidad, basados en las características físicas y de diseño de los edificios. Estos modelos, sin dejar de lado la toma de medidas in situ, pueden ser integrados en herramientas que simulen el rendimiento energético de forma que se optimicen procesos de diseño y rehabilitación, optimizando costes y tiempo.

En los últimos años se ha incrementado el interés por los modelos predictivos que estimen la estanqueidad (Bramiana et al., 2016 [4]; Chan et al., 2013 [5]; Khemet y Richman, 2018 [6]; Krstić et al., 2014 [7]). Aunque estas pruebas realizadas físicamente en los edificios siguen siendo

una forma más fiable de medir la estanqueidad, el poder predecirla de forma precisa es crucial para una planificación efectiva.
En España, ya existen modelos predictivos (Fernández Agüera et al. 2016, 2019 [8, 9] ; Ibanez-Puy and Alonso 2019 [10]; Montoya et al. 2010 [11]), diseñados para regiones y tipos de edificios específicos y que sin embargo han aportado resultados fácilmente exportables a otros países del sur de Europa, con características comunes en los sistemas constructivos. Esto también hace los modelos exportables usando otros conjuntos de datos.
En España, y aunque tradicionalmente considerada una fuente de ventilación en viviendas sin sistemas de ventilación controlada, la infiltración de aire ha pasado a estar más controlada debido a la tendencia a edificios más eficientes energéticamente en los que la estanqueidad prima. Desde 2019 y aunque afectando únicamente a viviendas de nueva construcción o rehabilitadas de más de 120 m^2, tanto el Código Técnico de la Edificación (CTE) [12], como otras normativas, han comenzado a exigir límites de estanqueidad. Esto subraya la necesidad de métodos precisos para estimar la estanqueidad, complementando las pruebas in situ y facilitando el cumplimiento de las normativas mediante herramientas de simulación energética como LIDER/CALENER (HULC) [13].

1.2. Objetivos

El objetivo principal de este trabajo es desarrollar un modelo predictivo que permita estimar con precisión la estanqueidad de los edificios residenciales en España. Este modelo debe ser capaz de considerar las diversas características del edificio y las condiciones regionales, proporcionando una herramienta útil para diseñadores y gestores de proyectos de rehabilitación. El presente trabajo es una extensión que trata de perfeccionar y ampliar los estudios **An envelope airtightness predictive model for residential buildings in Spain** [14] y **Airtightness predictive model from measured data of residential buildings in Spain** [15] previamente realizados por mis tutores Pilar Rodríguez y Miguel A. Fernández, conjuntamente con profesores de la Escuela Téccnica Superior de Arquitectura de la Universidad de Valladolid, a través de un modelo lineal (LM).

Aunque los modelos lineales han sido ampliamente utilizados debido a su simplicidad y claridad en la interpretación de las relaciones entre variables, los modelos lineales generalizados, como los que se usarán en este trabajo, son capaces de ofrecer mayor flexibilidad, permitiendo modelar relaciones no lineales y trabajar con distribuciones de datos que no se ajustan a la normalidad.

En los artículos [14, 15] se partía de un modelo lineal en el que la variable respuesta estaba transformada logarítmicamente para cumplir las asunciones de normalidad y linealidad del LM, pero, dado que la variable respuesta parece no seguir una distribución normal, vamos a tratar de usar el GLM. Esto se hace con el objetivo de ver si es posible mejorar la predicción usando directamente la variable original junto con su distribución correspondiente en lugar de transformarla logarítmicamente.

Adicionalmente, en en la 43^{rd} edición de la AIVC Conference [16], se sugirió utilizar un método de validación cruzada para seleccionar el mejor modelo, en lugar de entrenarlo y validarlo con el conjunto completo de los datos. Esta sugerencia se incorpora también en este trabajo por lo que el modelo final que se presentará será elegido en base a una validación 5-fold, como respuesta a las sugerencias de mejora presentadas.

1.3. Herramientas

Para la realización del TFG se han usado las siguientes herramientas:

- ***R***: se trata de un lenguaje de programación y un entorno de software libre diseñado para el análisis estadístico y la visualización de datos. Es ampliamente utilizado por estadísticos, analistas de datos e investigadores para realizar análisis complejos y generar gráficos de alta calidad.
- ***RStudio***: RStudio es un entorno de desarrollo integrado (IDE) para el lenguaje de programación R, utilizado para análisis de datos y desarrollo de software estadístico.
- ***Overleaf***: plataforma de edición en línea para la creación de documentos en LaTeX sin necesidad de instalar un software adicional.

1.4. Asignaturas relacionadas

A lo largo del grado hay asignaturas en las que se ha adquirido la mayor parte del conocimiento para realizar este proyecto, las más relevantes son:

- **Regresión y Anova y Modelos lineales:** estas asignaturas sientan la base de los modelos lineales y las estrategias de validación, desde una perspectiva tanto teórica como práctica. También en Regresión y Anova se presentan estrategias de selección automática de variables que tienen cabida en el objeto de este trabajo.
- **Análisis de Datos y Análisis Multivariante:** se introducen los modelos orientados a la predicción y la metodología para evaluarlos, como la validación cruzada.
- **Modelos Estadísticos Avanzados:** en ella se profundiza en los modelos lineales generalizados, como los usados en este trabajo.
- **Computación estadística:** introducción al lenguaje R, adquiriendo los fundamentos que han permitido el desarrollo técnico de este trabajo en el marco de la estadística.
- **Data Mining:** esta asignatura que cursé en Italia supuso mi primer contacto con los GLM y en ella se profundizó en sus utilidades tanto para regresión como para clasificación.

En este trabajo también hemos utilizado elementos nuevos con respecto a los contenidos del grado como la elección de la función de enlace en el modelo lineal generalizado, dado que la función elegida no es la canónica para la distribución de la variable respuesta que hemos utilizado. Para ello se han aplicado métodos que no conocía apoyándome en la bibliografía proporcionada por los tutores [17]. También he profundizado en métodos para tratar la correlación de las variables categóricas.

Capítulo 2

Marco teórico

2.1. Modelo lineal

El modelo lineal es una técnica estadística utilizada para describir la relación entre una variable dependiente y una o más variables independientes mediante una ecuación lineal. Este modelo permite predecir el valor de la variable dependiente basado en los valores de las variables independientes.
La ecuación del modelo lineal es de la forma:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_n x_n + \epsilon$$

Donde y es la variable dependiente, β_0 es la intersección, β_i son los coeficientes de las variables independientes x_i, y ϵ es el término de error. En el modelo lineal es habitual la construcción de tablas (ANOVA y de coeficientes individuales) en las que podemos ver un resumen de los resultados, incluyendo los coeficientes β_i , el error estándar, el *t-value*, el R^2 indicando la proporción de la varianza en la variable dependiente que explica este modelo, así como el estadístico *F*, que mide la significancia global del modelo.

2.1.1. Hipótesis del Modelo Lineal

Las hipótesis básicas que deben cumplirse en un modelo lineal son las siguientes:

1. **Linealidad:** La relación entre las variables independientes y la variable dependiente debe ser lineal. Esto se puede verificar de las siguientes maneras:

    - Creación de un gráfico de residuos frente a valores predichos. Si los residuos no muestran patrones claros y se distribuyen aleatoriamente alrededor de la línea horizontal, se puede asumir la linealidad.

- Creación de gráficos de dispersión entre cada variable independiente y la variable dependiente para verificar visualmente la relación lineal.

2. **Homogeneidad de varianzas:** La varianza de los errores debe ser constante en todos los niveles de las variables independientes. Esto se puede verificar de las siguientes maneras:

   - Realización del test de Levene.
   - Creación de un gráfico de residuos frente a valores predichos. En este gráfico, los residuos estandarizados se representan en el eje Y y los valores predichos en el eje X. La varianza constante se indica por una dispersión de residuos sin patrón de aumento o disminución a lo largo del eje X. Además, este gráfico puede ayudar a detectar observaciones atípicas, las cuales tienen residuos estandarizados mayores a 3 en valor absoluto.

3. **Normalidad:** Los errores deben distribuirse normalmente. Para verificar esta hipótesis, se puede analizar la distribución de los residuos. Algunas técnicas para esto incluyen:

   - Creación de un histograma de los residuos.
   - Creación de un gráfico *Q-Q plot* (Quantile-Quantile plot).
   - Realización de pruebas de normalidad como el *test de Shapiro-Wilk.*

4. **Independencia:** Las observaciones deben ser independientes entre sí. Esta condición se puede asumir si el muestreo se ha realizado de manera adecuada y aleatoria. Si creamos un gráfico de residuos contra el orden de la observación, los datos no tienen que seguir ningún patrón.

En muchas ocasiones, nos encontramos con que uno o varios de estos supuestos no se cumplen debido a la naturaleza de los datos. Para abordar estos problemas, es común aplicar transformaciones a la variable de respuesta, como por ejemplo, tomar logaritmos. No obstante, estas transformaciones no siempre logran corregir la falta de normalidad, la heterocedasticidad (varianza no constante) o la no linealidad de los datos. Además, el uso de transformaciones puede complicar la interpretación de los resultados obtenidos, lo que añade una capa adicional de dificultad al análisis.

2.2. Modelo Lineal Generalizado

Los modelos lineales generalizados o GLM (McCullagh & Nelder et al.,1989) [17], suponen una alternativa a los modelos lineales cuando no se cumplen algunas de sus hipótesis. Los GLM son, por lo tanto, una extensión de los modelos lineales que nos permiten utilizar distribuciones no normales de los errores, tales como binomiales, Gamma o Poisson entre otras, varianzas no constantes o relaciones no lineales entre la variable respuesta y las predictoras.

En un GLM, se supone que cada resultado Y de las variables dependientes se genera a partir de una distribución particular de la familia exponencial. La media, μ, de la distribución depende de las variables independientes, X, por medio de:

$$E(Y) = \mu = g^{-1}(X\beta)$$

Donde:

1. $E(Y)$ **es el valor esperado de** Y**:**

   - La componente aleatoria Y sigue una distribución de la ya citada familia exponencial. El valor esperado representa el promedio teórico de una variable aleatoria. En un GLM, $E(Y)$ indica el valor medio o esperado de la variable dependiente Y, modelado como μ, la media de la distribución de Y. Esto implica predecir cómo varía la media de Y en función de las variables independientes X.

2. $X\beta$ **es el "predictor lineal", una combinación lineal de parámetros por determinar** β**:**

   - El predictor lineal es una combinación lineal de los valores de las variables independientes ponderadas por sus coeficientes correspondientes, β. Este predictor lineal representa la influencia conjunta de las variables independientes en la media de la variable dependiente. Es una parte fundamental del modelo ya que describe cómo las variables independientes contribuyen al valor esperado de Y.

3. g **es la función de enlace:**

   - La función de enlace, denotada como g, relaciona la media de la distribución de la variable dependiente (μ) con el predictor lineal ($X\beta$). Específicamente, g transforma la media de la distribución de Y para que se relacione linealmente con el predictor lineal. La elección de la función de enlace depende del tipo de datos y la distribución de la variable respuesta. Para cada distribución de la variable respuesta podemos encontrar su propia función de enlace canónica.

Esta fórmula es fundamental en los GLM, ya que define cómo se relacionan las variables independientes con la media de la variable dependiente a través del predictor lineal.

2.2.1. Distribución de la variable respuesta

Como el GLM nos permite realizar modelos cuya variable respuesta se aleja de la normal, en primer lugar tenemos que decidir que distribución es la que sigue nuestra variable. Para ello y aunque hay otras opciones muy relacionadas con la validación del modelo, y que veremos en el punto 2.2.3, los test de bondad de ajuste son la mejor opción. Los dos más relevantes son:

- **Test de Kolmogorov-Smirnov (K-S)** : es una prueba no paramétrica que se utiliza para determinar si una muestra de datos sigue una distribución específica. La prueba compara la función de distribución empírica de la muestra con la función de distribución acumulativa de la distribución teórica.

  Dada una muestra de datos $X_1, X_2, \ldots, X_n$ ordenada de menor a mayor, la función de distribución empírica $F_n(x)$ se define como:

  $$F_n(x) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}$$

  donde $\mathbf{1}_{\{X_i \leq x\}}$ es una función indicadora que toma el valor 1 si $X_i \leq x$ y 0 en caso contrario.

  La función de distribución acumulativa teórica $F(x)$ es la distribución contra la cual se está realizando la prueba.

  El estadístico de Kolmogorov-Smirnov D se define como:

  $$D = \sup_x |F_n(x) - F(x)|$$

  Donde $\sup$ representa el máximo valor absoluto de las diferencias entre la empírica y la teórica en todos los puntos x.
  Para decidir si la muestra sigue o no la distribución teórica, comparamos el valor calculado de D con el valor crítico correspondiente.

  Si D es menor o igual al valor crítico, no hay evidencia suficiente para rechazar la hipótesis nula de que la muestra sigue la distribución teórica. En cambio, si D es mayor que el valor crítico, se rechaza la hipótesis nula y se concluye que la muestra no sigue la distribución teórica.

- **Test de Anderson-Darling (A-D)** : es una prueba de bondad de ajuste que da más peso a las colas de la distribución que el test de Kolmogorov-Smirnov. Es una extensión del test de Cramer-von Mises. Para una muestra de datos ordenada de menor a mayor, $X_1, X_2, \ldots, X_n$, la función de distribución empírica $F_n(x)$ y la función de distribución acumulativa teórica $F(x)$ se utilizan para calcular el estadístico A^2:

  $$A^2 = -\frac{n-1}{n} \sum_{i=1}^{n} \left((2i-1)\left[\log F(X_i) + \log\left(1 - F(X_{n+1-i})\right)\right]\right)$$

  Donde:

  - n es el tamaño de la muestra.
  - X_i son los valores ordenados de la muestra.
  - $F(X_i)$ es la función de distribución acumulativa teórica evaluada en X_i.

  Para realizar el test de A-D, calculamos el valor de A^2 y lo comparamos con los valores críticos correspondientes para el nivel de significancia elegido y el tamaño de la muestra.

Si el valor de A^2 supera el valor crítico, rechazamos la hipótesis nula de que la muestra sigue la distribución teórica. Por otro lado, si el valor de A^2 es menor o igual al valor crítico, no hay suficiente evidencia para rechazar la hipótesis nula.

El estadístico de Anderson-Darling se centra en las diferencias entre la función de distribución empírica y la función de distribución acumulativa teórica, ponderando más las diferencias en las colas de la distribución.

2.2.2. Elección de la función de enlace

La elección de la función de enlace 3 o link es crucial en un GLM, ya que determina cómo se modela la relación entre las variables independientes y la variable dependiente. Aunque cada distribución tiene su función de enlace canónica, esta no siempre proporciona los mejores resultados. Las funciones de enlace canónicas para cada distribución aparecen en la figura 2.1.

Función de vínculo	Fórmula	Uso
Identidad	μ	Datos continuos con errores normales (regresión y ANOVA)
Logarítmica	$Log(\mu)$	Conteos con errores de tipo Poisson
Logit	$Log(\frac{\mu}{n-\mu})$	Proporciones (datos entre 0 y 1) con errores binomiales
Recíproca	$\frac{1}{\mu}$	Datos continuos con errores gamma
Raíz cuadrada	$\sqrt{\mu}$	Conteos
Exponencial	μ^n	Funciones de potencia

Fuente: Extraído de `https://www.uv.es/lejarza/eaa/teoria/EAA7%20glm.pdf`.

Figura 2.1: Funciones de enlace canónicas para diferentes distribuciones de la respuesta.

Para la elección de la función de enlace, se utilizará la metodología explicada en los capítulos 11.3 y 12.6.3 del libro **Generalized Linear Models P. McCullagh and J.A. Nelder** [17] en la que se trata de encontrar el punto en el que se da la menor *Deviance* a través de una modificación de la función power imitando las transformaciones de Box Cox. Consideramos la función power definida en la figura 2.2.

$$\eta = \begin{cases} \mu^\lambda & \text{for } \lambda \neq 0, \\ \log \mu & \text{for } \lambda = 0, \end{cases}$$

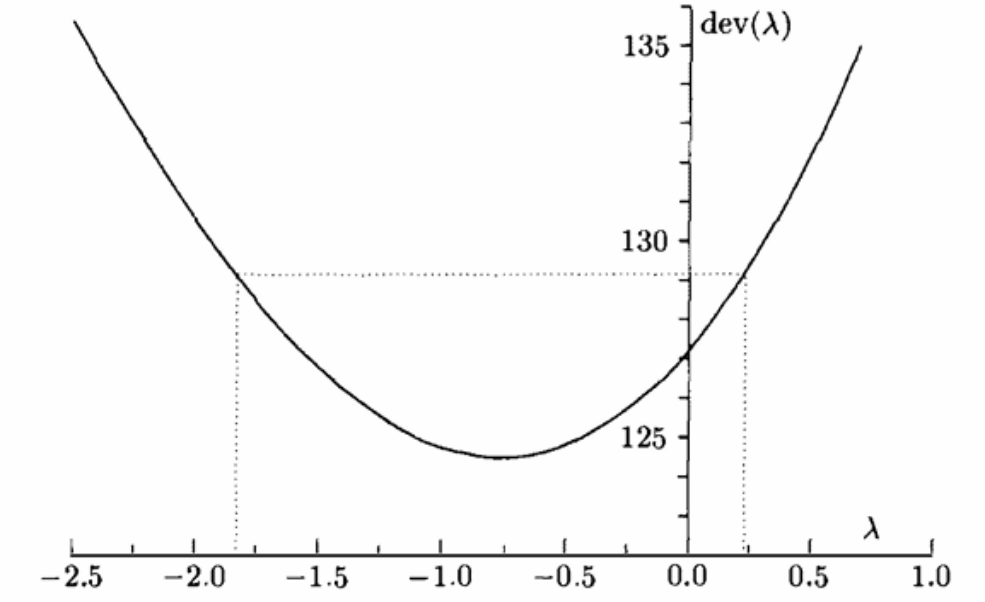
or, in the form having continuity at $\lambda = 0$,

$$\eta = \frac{\mu^\lambda - 1}{\lambda}.$$

Fuente: **Generalized Linear Models P. McCullagh and J.A. Nelder** [17]

Figura 2.2: Función power

De esta forma, representando gráficamente los valores de la *Deviance* del modelo para los diferentes valores de λ, se obtiene una representación gráfica del punto en el que se alcanza la *Deviance* mínima. Esto se puede ver en la figura 2.3.



Fuente: **Generalized Linear Models P. McCullagh and J.A. Nelder** [17]

Figura 2.3: Gráfico de mínima deviance

También se puede usar la estrategia de linearización propuesta por Pregibon en su artículo **Goodness of Link Tests for Generalized Linear Models** [18], en el que a través de expandir la función de enlace con desarrollos de Taylor sobre un λ_0 comparamos si el link que estamos utilizando puede ser mejorado.

2.2.3. Selección de variables y validación del GLM

Una vez que se tiene el modelo con todas las variables, para seleccionarlas se pueden usar métodos de selección automática basados en diferentes parámetros. Tenemos tres formas de realizar esta selección, aunque a lo largo de este trabajo solo se van a usar las dos primeras debido a las limitaciones de los paquetes de R:

- **Selección hacia atrás (Backward):** se comienza con un modelo completo que incluye todas las variables predictoras posibles y se eliminan variables una por una. Los pasos son:

  1. **Inicio:** se parte del modelo completo.
  2. **Iteración:** en cada paso, se elimina la variable que menos contribuye al modelo según el criterio de selección que se haya elegido.
  3. **Parada:** El proceso continúa hasta que no se puedan eliminar más variables sin empeorar el modelo.

- **Selección hacia adelante (Forward):** se comienza con un modelo vacío y se agregan variables predictoras una por una, basándose en algún criterio de selección como el AIC, BIC o p-valor. El procedimiento es el siguiente:

  1. **Inicio:** se parte de un modelo vacío.
  2. **Iteración:** en cada paso, se agrega la variable que mejora más el modelo según el criterio de selección elegido.
  3. **Parada:** el proceso continúa hasta que no se puedan agregar más variables que mejoren significativamente el modelo.

- **Selección mixta (Stepwise):** la selección mixta combina los métodos hacia adelante y hacia atrás. Comienza como el método de selección hacia adelante, pero en cada paso permite la eliminación de variables que se han vuelto no significativas. Los pasos son:

  1. **Inicio:** se parte de un modelo vacío.
  2. **Iteración hacia adelante:** se agrega la variable que mejora más el modelo.
  3. **Iteración hacia atrás:** después de agregar una variable, elimina cualquier variable que haya dejado de ser significativa.
  4. **Parada:** El proceso continúa hasta que no se puedan agregar ni eliminar más variables que mejoren significativamente el modelo.

Estos criterios usados para la selección de variables sirven también para para comparar y validar los diferentes modelos, en nuestro caso se van a usar las siguientes:

- **Criterio de p-valores significativos:** el p-valor es una medida que se utiliza para determinar la significancia estadística de cada variable predictora en el modelo. Se basa en la hipótesis nula, que asume que la variable no tiene efecto sobre la respuesta. Se puede decidir a qué nivel queremos establecer la significación aunque normalmente es a un nivel $\alpha = 0.05$. Un p-valor bajo (menor al nivel que se haya determinado), indica que la variable es significativa y tiene efecto sobre la variable respuesta, mientras que un p-valor alto indica que no es significativa.

  El proceso de selección de variables en base a este criterio consiste en:

  1. Ajustar el modelo con todas las variables.
  2. Evaluar los p-valores de cada variable.
  3. Eliminar iterativamente las variables con los p-valores más altos, reajustando el modelo cada vez, hasta que todas las variables restantes sean significativas.

  Al eliminar variables predictoras no significativas se evita el sobreajuste.

- **Criterio de Información de Akaike (AIC)**:

  - El AIC es una medida de la calidad del modelo que penaliza la complejidad del modelo. Se calcula utilizando la siguiente fórmula:

$$AIC = -2\ln(L) + 2k$$

Donde L es la función de verosimilitud del modelo. El término k es el número de parámetros estimados en el modelo, incluyendo los parámetros de regresión y cualquier otro parámetro de dispersión.

El AIC busca encontrar un equilibrio entre el ajuste del modelo a los datos y la complejidad del modelo. Penaliza los modelos con un mayor número de parámetros, lo que ayuda a evitar el sobreajuste al seleccionar el modelo óptimo entre varios candidatos. Cuanto menor es el AIC mejor es el modelo.

- **Criterio de Información Bayesiano (BIC)**:

  - El BIC es similar al AIC pero penaliza la complejidad del modelo de manera más estricta, teniendo en cuenta también en cuenta el número de observaciones. Se calcula utilizando la siguiente fórmula:

$$\textit{BIC} = -2\ln(L) + \ln(n)k$$

Donde n es el tamaño de la muestra, L al igual que en el caso anterior es la función de verosimilitud del modelo y k el número de parámetros.

Al agregar el término $\ln(n)k$, el BIC penaliza más fuertemente los modelos más complejos que el AIC. Esto lo hace más adecuado para conjuntos de datos pequeños donde la selección del modelo puede ser más difícil y se necesita una penalización más estricta para evitar el sobreajuste.

- **Criterio de la menor *Deviance***: Tal y como se ve en 2.2.2, otro de los criterios que nos pueden conducir a elegir un mejor modelo es la elección del modelo con una menor *Devianza Residual*.

  Para entender la *Devianza Residual* primero tenemos que saber que es la *Devianza Nula*. Esta, es una medida de cuánto peor es el modelo nulo (un modelo que solo incluye la intersección) en comparación con un modelo perfectamente ajustado, es decir, lo bien que se ajustarían los datos si solo consideramos el valor medio de la variable respuesta como predictor. Un valor alto de *Devianza Nula* sugiere que el modelo nulo no se ajusta bien a los datos.

  La *Devianza Residual*, por lo tanto es una medida de cuanto peor es el modelo ajustado, con variables predictoras en comparación con el modelo nulo. Es la diferencia entre la *Devianza Nula* y la devianza del modelo ajustado. Cuanto menor sea la *Devianza Residual*, mejor será el ajuste del modelo a los datos observados. se calcula con la siguiente fórmula:

$$D_{\text{residual}} = D_{\text{nula}} - D_{\text{ajustado}}$$

- **Pseudo $\mathbf{R}^2$**: Dado que el software que se va a utilizar el R y para el GLM no se dispone ni de un R^2 ni de su versión ajustada, vamos a utilizar pseudo R^2, algunas de ellas son:

  - R^2 de McFadden: se trata de una alternativa al "índice de razón de verosimilitud", su fórmula es la siguiente:

$$R^2_{\text{MF}} = 1 - \frac{LL_{\text{Propuesto}}}{LL_{\text{Nulo}}}$$

Donde:

- $LL_{\text{Propuesto}}$ es la log-verosimilitud del modelo ajustado (con todas las covariables).
- LL_{Nulo} es el log-verosimilitud del modelo solo con el *intercept*.

Este pseudo R^2 proporciona una medida de cuánto mejor se ajusta el modelo propuesto en comparación con el modelo nulo.

- R^2 de Cox & Snell: se calcula comparando el logaritmo de la verosimilitud del modelo propuesto con el logaritmo de la verosimilitud de un modelo de referencia.
- R^2 de Nagelkerke: se trata de una versión ajustada del Rcuadrado de Cox & Snell

  Las fórmulas concretas de estos algoritmos el algo que queda fuera de los objetivos de este TFG, pero que pueden ser consultadas aquí [19].

- **Validación cruzada con 5-fold**: se trata de una técnica ampliamente utilizada para evaluar el rendimiento de un modelo predictivo.[20]

  1. **División de los datos**: En primer lugar, se dividen los datos disponibles en cinco partes iguales, o "folds". Cada fold contiene una cantidad equitativa de observaciones.
  2. **Entrenamiento y evaluación**: Luego, el modelo se entrena en cuatro de los folds (el 80 % de los datos) y se evalúa en el quinto fold restante (el 20 % de los datos). Esto se repite cinco veces, de modo que cada fold se utiliza una vez como conjunto de prueba o test y cuatro veces como conjunto de entrenamiento. En nuestro caso este entrenamiento servirá de cara a la selección de variables, utilizando una metodología tanto *backward*, como *forward*.
  3. **Cálculo del rendimiento**: En cada iteración, se calcula una métrica de rendimiento, como el mejor AIC, BIC, menor Deviance o el mejor R^2, utilizando el conjunto de prueba. Al finalizar las cinco iteraciones, en nuestro caso, nos quedaremos con el modelo que mejor R^2 nos haya proporcionado en el subconjunto de prueba.
  4. **Robustez y generalización**: La validación cruzada con 5-fold proporciona una estimación más robusta del rendimiento del modelo en comparación con una sola división de los datos en entrenamiento y prueba. Al utilizar múltiples particiones de los datos, se reduce la variabilidad en la estimación del rendimiento y se obtiene una evaluación más confiable de la capacidad de generalización del modelo a nuevos datos.

  Por lo tanto la validación cruzada con 5-fold es una técnica poderosa para evaluar el rendimiento de un modelo y que proporciona una estimación más robusta de su capacidad de generalización, en vez de darnos un modelo sobreajustado.

- **Análisis de los gráficos de residuales**: Con respecto a los gráficos, se va a analizar los cuatro proporcionados por la función *plot()* de R, esto servirá para detectar anomalías como outliers o puntos de influencia o problemas de falta de normalidad y varianza no constante. Estos análisis se realizan con los residuos del modelo, que son la diferencia entre el valor real y el predicho por el modelo.

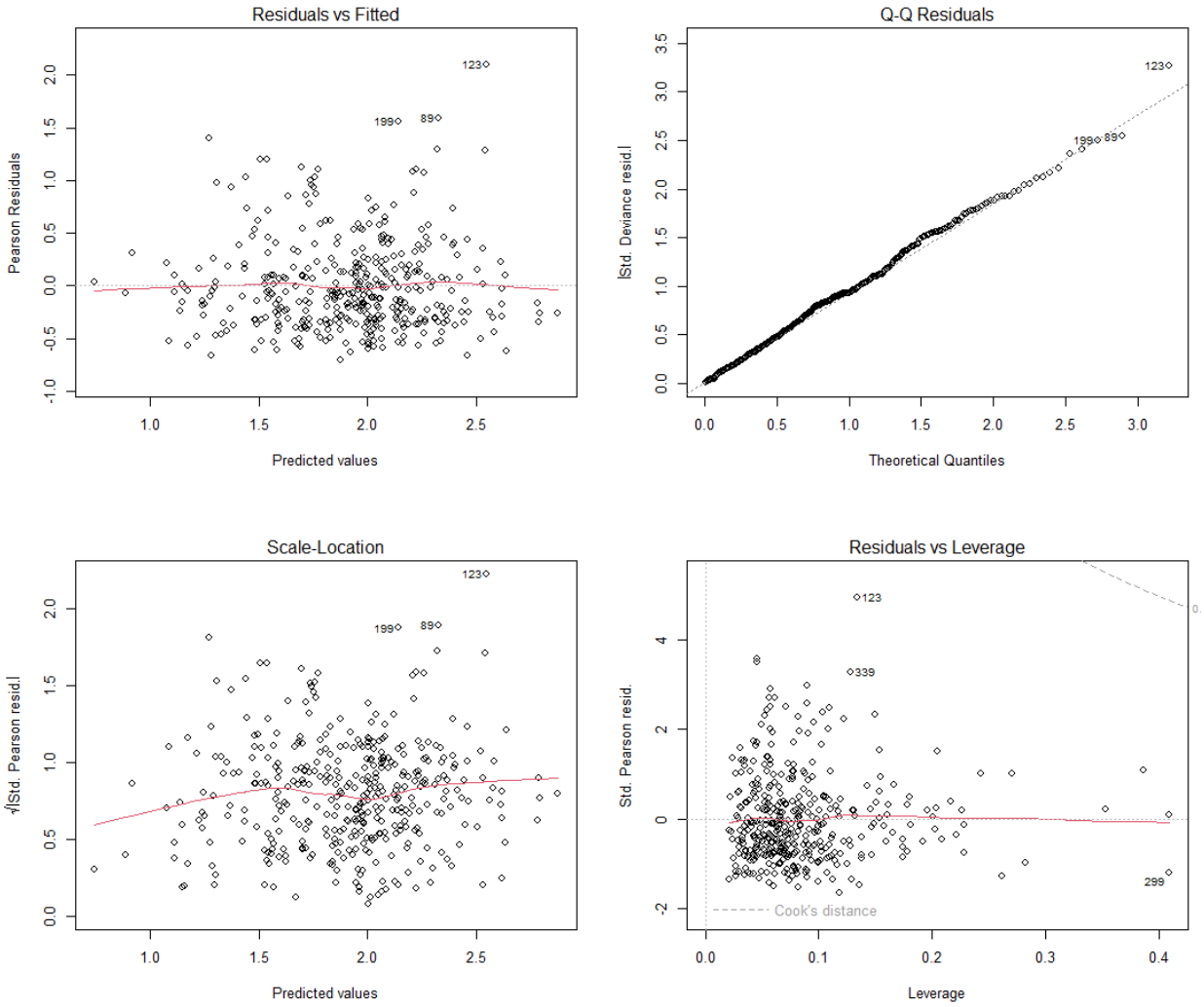


Figura 2.4: Gráficos de residuales de modelo

Los gráficos a analizar, que se pueden observar en la figura 2.4, van a ser los siguientes:

1. **Residuals vs Fitted:** este gráfico muestra los residuos en el eje Y contra los valores ajustados en el eje X. Su propósito es el de detectar problemas de no linealidad, heterocedasticidad (varianza no constante) y outliers.

   Para que un modelo cumpla las asunciones, los residuos deben distribuirse alrededor de la línea horizontal dibujada en rojo sin formar patrones claros. Ejemplos de la violación de las asunciones son un patrón en forma de parábola que indica falta de normalidad, mientras que un “cono horizontal” puede indicar heterocedasticidad.
2. **Standarized residuals vs Fitted:** aquí, al igual que en el anterior, se muestra la raíz cuadrada de los residuos estandarizados en el eje Y contra los valores ajustados en el eje X. Su prósito e interpretación son los mismos que en el caso anterior.
3. **Q-Q Plot:** este gráfico separa los cuantiles teóricos estandarizados de los cuantiles teóricos de una distribución normal, con el fin de verificar la normalidad de los residuos. Si estos siguen una distribución normal, se estarán alineados con la línea diagonal. Si existen desviaciones importantes de esta línea, los residuos no son normales.

4. **Residuals vs Leverage:** este gráfico muestra los residuos estandarizados en el eje Y contra la leverage en el eje X. Su objetivo es identificar puntos influyentes en el modelo, que pueden afectar a la selección de variables y estimación de coeficientes. El gráfico también incluye líneas de Cook's distance (en gris), indicando los valores de influencia potencialmente altos.

Si a través de estos gráficos se diagnostican problemas en el modelo, la solución puede pasar por una transformación de variables, eliminación de observaciones, o la consideración de enfoques diferentes.

2.3. Calculo de correlaciones

El cálculo de correlaciones entre variables es una técnica estadística fundamental en el análisis de datos. Este proceso permite identificar y cuantificar la relación entre dos o más variables. En el contexto de los modelos lineales generalizados (GLM), entender estas relaciones es crucial para construir modelos predictivos precisos y robustos.

2.3.1. Correlación de Pearson

La correlación de Pearson mide la fuerza y la dirección de la relación lineal entre dos variables continuas. Se denota comúnmente por r y se calcula utilizando la siguiente fórmula:

$$r = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum(X_i - \bar{X})^2 \sum(Y_i - \bar{Y})^2}}$$

Donde:

- X_i y Y_i son los valores individuales de las variables X y Y.
- $\bar{X}$ y $\bar{Y}$ son las medias de las variables X y Y.

El coeficiente de correlación de Pearson varía entre -1 y 1. Un valor de 1 indica una correlación positiva perfecta, un valor de -1 indica una correlación negativa perfecta, y un valor de 0 indica que no hay correlación lineal entre las variables.

2.3.2. Correlación de Cramér

La correlación de Cramér, también conocida como V de Cramér, se utiliza para medir la asociación entre dos variables categóricas. Se calcula a partir del estadístico chi-cuadrado (χ^2) y se normaliza para tener un rango entre 0 y 1:

$$V = \sqrt{\frac{\chi^2/n}{\text{mín}(k-1, r-1)}}$$

donde:

- χ^2 es el estadístico chi-cuadrado obtenido de la tabla de contingencia de las dos variables.
- n es el tamaño de la muestra.
- k y r son el número de categorías entre filas y columnas.

Un valor de V de Cramér de 0 indica que no hay asociación entre las variables, mientras que un valor de 1 indica una asociación perfecta. Valores superiores a 0.6 indican una asociación fuerte[21].

2.3.3. Importancia del cálculo de correlaciones en el contexto de los GLM

El cálculo de correlaciones es especialmente relevante en el contexto de los GLM debido a las siguientes razones:

- **Identificación de Relaciones Lineales**: En los GLM, es crucial identificar y modelar las relaciones lineales entre las variables independientes y la variable respuesta. La correlación de Pearson es útil para identificar estas relaciones lineales entre variables continuas.
- **Detección de Multicolinealidad**: La multicolinealidad ocurre cuando dos o más variables independientes están altamente correlacionadas entre sí. Esto puede causar problemas en la estimación de los coeficientes del modelo, como:
  - Coeficientes de regresión inestables y grandes errores estándar.
  - Dificultad para determinar la influencia individual de cada variable en la variable respuesta.
  - Inflación de los errores estándar, lo que reduce la significancia estadística de las variables.

  Calcular las correlaciones entre las variables independientes permite detectar y mitigar la multicolinealidad, eliminando las variables muy relacionadas entre sí.
- **Evaluación de Relaciones No Lineales**: Aunque los GLM, como en el caso de nuestro modelo, pueden manejar relaciones no lineales mediante funciones de enlace, es importante entender la naturaleza de las relaciones entre las variables. La correlación de Cramér es útil para evaluar la asociación entre variables categóricas, lo que puede influir en la especificación del modelo.

- **Mejora del Modelo Predictivo**: identificar y entender las correlaciones ayuda a seleccionar las variables más relevantes para el modelo. Incluir variables que tienen una fuerte correlación con la variable respuesta puede mejorar la precisión del modelo predictivo.

Por ello, el cálculo de correlaciones entre variables es un paso fundamental e imprescindible en la construcción de modelos lineales generalizados. Tanto la correlación de Pearson como la correlación de Cramér aportan información valiosa sobre las correlaciones entre variables, ayudando a construir modelos más precisos y robustos. Al identificar relaciones lineales, detectar multicolinealidad y evaluar asociaciones entre variables, se puede mejorar la especificación y el rendimiento del GLM, asegurando así resultados más fiables e interpretables.

Capítulo 3

Datos

3.1. Obtención

Los datos utilizados de un único dataset, la base de datos nacional INFILES (Feijó-Muñoz et al. 2019) [22]. Este conjunto de datos se compone de 400 observaciones de edificios residenciales de España. Esta muestra se considera representativa y uniformemente distribuida para el territorio nacional.

La distribución de viviendas se hizo siguiendo el criterio de muestreo definido [16, 23]. Como variables de control se utilizaron *zona climática, año de construcción y tipología del edificio*.

- **Zona Climática:** se pueden distinguir cuatro zonas climáticas diferentes, la zona mediterránea (209 observaciones, 52.1 %), la zona continental (129 observaciones, 32.2 %), la zona oceánica (47 observaciones, 11.7 %) y las Islas Canarias (16 observaciones, 4 %).
- **Año de construcción:** el dataset recoge información de viviendas construidas entre los años 1880 y 2015, aunque no están distribuidas uniformemente, dado que los mayores periodos constructivos de España fueron entre 1960 y 1979, encontrando 148 observaciones que significan un 36.9 % de la muestra, y entre 1980 y 2006, con 158 casos que ascienden al 39.4 % de las observaciones. Este periodo acaba con el estallido de la burbuja inmoviliaria y una reducción en la construcción de viviendas a partir de 2007. Es por ello que las viviendas construidas en estos años son las más representadas en nuestro conjunto de datos
- **Tipología:** aquí se va a distinguir entre apartamentos (325 casos, 81 %) y casas unifamiliares (76 casos, 19 %).

La identificación y propósito de cada variable, sobre todo basándose en los nombres de las columnas del dataset resulta una tarea complicada, pues para una persona que no tiene un gran conocimiento arquitectónico, los nombres pueden no llegar a ser muy representativos. Es por ello que aquí ha cobrado especial relevancia un documento proporcionado por los arquitectos en

el que se relacionan los nombres de las columnas con la información a la que hacen referencia, así como una selección de variables con las que trabajar. Gracias a esta selección previa por parte de los arquitectos se ha podido llegar a un conjunto de en torno a 60 variables de las 203 que tiene el conjunto de datos y que son las que se van a usar de cara a realizar el modelo. Estas variables pertenecen a una de estas categorías:

- **Localización:** aquí se pueden ecntrar variables relacionadas con la ubicación geográfica de los edificios, como su sede, o su clima en verano o invierno.
- **Años del Edificio:** en esta categoría están las variables referentes a la antigüedad del edificio.
- **Tipo de edificio:** a esta categoría pertenecen variables tan elementales como el propio nombre de la categoría, *Tipología*, hasta variables como el número de habitaciones o baños, pasando por la altura del edificio.
- **Estado del edificio:** estas variables evalúan el estado en el que se encuentra la vivienda, si presenta problemas, si ha sido reformada, o si la fachada tiene su estado original.
- **Sistema constructivo del edificio:** dentro de esta categoría, las variables muestran las técnicas que se han usado para construir el edificio, como el tipo de persianas, el tipo de aislamiento que se ha usado, las ventanas utilizadas o el tipo de ventilación.
- **Dimensiones:** en esta última categoría están la mayor parte de variables numéricas y hace referencia a las dimensiones de la vivienda, como su superficie útil, su altura libre, o su volumen. También el tamaño de las ventanas.

Esto nos deja la siguiente lista de variables, en la que se muestrean las variables seleccionadas para el modelo, así como las que fueron eliminadas en diferentes etapas del proceso de selección. Las variables se categorizan por tipo y se muestra su inclusión o exclusión del modelo final.

Tipo de Variable	Variable (SPA)	Modelo
Ubicación	zonaclimaticaHE1	**Sí**
	sede	
	sev.Climainvierno	
	sev.Climaverano	
	zon1climáticainfiles	
Antigüedad del edificio	añoconstr1980	
	añoconstrucción	**Sí**
	rangoperiodos	
	normativa	
Tipo de edificio	Tipología	

Continúa en la siguiente página

Tabla 3.1 – continuación de la página anterior

Tipo de Variable	Variable (SPA)	Modelo
	colectivaposicion	
	plantadelavivienda	
	alturadeledificio	
	tipodepromoción	
	númerodeestancias	
	NB4	
	cocinayestarunidos	
Estado del edificio	estadodelavivienda	
	puentestérmicostratados	
	observadasfisurasuotraspatologíasenlavivienda	
	galeríasindependientes	
	GaleriaIntegrada	
	cocina reformada	
	baño reformado	**Sí**
	estadooriginaldelafachada	**Sí**
Sistema constructivo	Falsotecho	**Sí**
	carpinteriaAclase	**Sí**
	carpinteríatipoAmaterial	
	persianasegúncatálogo	
	fachada segun catalogo	
	revestimientoexteriorcontinuo	
	revestimientointeriorcontinuo	
	aislamiento	
	cámaradeaire	
	carpinteríatipoAcorredera	
	carpinteríatipoA doblehoja	
	tipopersiana	
	particiones_interiores	
	heating_system	
	cooling_system	
	tipoventilación	
	rejilladeventilaciónencocina	
	conductos	**Sí**
	campanaextractoraencocina	**Sí**
Dimensiones	carpinteríasenvolvente	**Sí**
	DJ_EZ	
	alturalibrepredeterminadafalsotecho	**Sí**
	superficieútilvivienda	
	volumenvivienda	

Continúa en la siguiente página

Tabla 3.1 – continuación de la página anterior

Tipo de Variable	Variable (SPA)	Modelo
	superficieenvolvente	
	Factordeforma	**Sí**
	volcuartoshúmedos	
	carpinteríasPerímetrototal	
	carpinteríasSuperficietotal	
	FLF	

Tabla 3.1: Resumen de Variables y Resultados del Modelo

No se describen todas las variables debido al gran número, aunque sí mencionamos las menos identificables de la tabla 3.1, NB4, como número de baños, DJ_EZ como porcentaje de superficie opaca exterior sobre la superficie de la envolvente total y FLF como perímetro de carpinterías dividido entre volumen.

Finalmente las variables elegidas son:

- Zona climática: se ha considerado de acuerdo a DB HE1 [24], teniendo en cuenta el invierno(zonas de la A a la E y α, y la severidad en verano (1-4). La severidad climática combina grados-día y radiación solar en cada lugar. Desde la perspectiva internacional, estas zonas tendrían la siguiente equivalencia en la clasificación climática de Köppen-Geiger [25]: $A3 =$ Csa, $B4 =$ BSk-Csa, $C1 =$ Csb-Cfb, $C2 =$ Csa, $C3 =$ BSk, $D2 =$ Csb, $\alpha =$ BSh
- Año de construcción: el año en que el edificio fue construido. Se encuentran edificios construidos entre los años 1880 y 2015, aunque tal y como se describe al principio de este apartado 3.1, podemos encontrar más observaciones entre los años 1960 y 1979 y entre los años 1980 y 2006.
- Baño reformado: esta variable cuenta con dos niveles indicando con un 1 si el baño ha sido reformado y con un 0 si no lo ha sido.
- Estado original de la fachada: indica si la fachada se encuentra en su estado original de construcción o por el contrario ha sido modificada, añadiendo o quitando ventanas, añadiendo balcones o terrazas o modificando las paredes externas entre otras modificaciones.
- Falso techo: indica la presencia de falso techo en las viviendas, diferenciando si se encuentra en pasillo y cuartos húmedos, es decir, cocina y baños, si no tiene falso techo o si por el contrario tiene falso techo en habitaciones que no sean cuartos húmedos.
- Permeabilidad de las ventanas: esta variable, descrita como `carpinteríaAclase`, se evaluó según la norma UNE-EN 12207 [26] y se clasificó como Clase 0 (ventanas no probadas), Clase 1 (hasta 50 m^3/hm^2), Clase 2 (hasta 27 m^3/hm^2), Clase 3 (hasta 9 m^3/hm^2), o Clase 4 (hasta 3 m^3/hm^2). Cabe señalar, sin embargo, que esta información no siempre estaba disponible y podría ser solo una estimación a partir de una inspección visual.

- Conductos: indica la presencia de conductos usados para distribuir el aire caliente, frío o ventilación dentro de un edificio.
- Campana extractora en la cocina: esta variable muestra si la cocina cuenta o no con una campana extractora, tiene tres niveles que hacen referencia a si tiene unidad filtrante, conducto vertical o es a fachada.
- Porcentaje de ventanas: descrita como `carpinteríasenvolvente`, se trata de la suma del área de las puertas y ventanas comparada con la superficie envolvente total. Este parámetro está estrechamente relacionado con A_h en el modelo propuesto por las normativas españolas. Esta es una variable cuantitativa [m^2].
- Altura libre predeterminada falso techo: se trata de la altura libre predeterminada con el falso techo, es decir, la distancia vertical entre el suelo de un espacio interior y el nivel inferior del falso techo instalado. Esta medida se toma antes de la instalación del falso techo y afecta a la percepción del espacio, a la ventilación y a la instalación de sistemas como aire acondicionado. Se mide en metros.
- Factor de forma: esta variable describe la relación entre el volumen de un edificio y su superficie envolvente. Se utiliza para evaluar la eficiencia térmica y energética de un edificio, dado que normalmente una mayor compacidad indica una menor superficie expuesta a cambios térmicos, como pérdidas o ganancias.

3.2. Tratamiento de los datos

Dentro de estas variables, encontramos una mayoría de variables categóricas y una serie de variables numéricas.

Previa a la inclusión de variables en el modelo, para poder trabajar con las variables y que se adapten al modelo necesario hemos tenido que realizar lo siguiente:

- **Detección de outliers:** Como se trata de un conjunto de datos que ya había sido usado anteriormente, la detección de outliers ya estaba hecha, por lo que usando la columna *Atipicos8*, se seleccionaron las viviendas que tenían el valor a 1, dejándonos una muestra de 392 variables.

  Adicionalmente y como se verá en el apartado 4.3, se eliminarán también las observaciones 123, 125 y 289 el primero dado que su desviación residual es alta y los dos últimos por su alto valor de Leverage que afecta significativamente a la selección de variables.

  Por lo tanto en la muestra final vamos a tener 389 viviendas.
- **Tratamiento de valores faltantes:** Con respecto al tratamiento de valores faltantes, no fue necesaria la aplicación de técnicas como la imputación para reemplazar observaciones. Sin embargo, sí que hubo una columna, *colectivaposición*, que debido a su alto porcentaje de valores faltantes (18 %) y los problemas que presentaba de cara a tratarla y compararla con otras variables, fue eliminada de la selección inicial del modelo.

- **Conversión de variables categóricas a factores:** dado que la mayor parte de las variables son categóricas, usando la herramienta R, se conviertieron a factores. Para ello, se tuvo que identificar manualmente las variables categóricas y convertirlas a un factor. Este proceso manual permitió asegurarse de que solo las variables adecuadas fueran convertidas, evitando así errores de clasificación.

  La conversión a factores permite al modelo distinguir entre los diferentes niveles de cada variable categórica. En modelos estadísticos y de aprendizaje automático, los factores permiten manejar adecuadamente las categorías, asignando niveles a cada categoría y permitiendo que el modelo interprete estas variables correctamente. Cabe mencionar que R realiza automáticamente la asignación de niveles a cada categoría, facilitando así el proceso de preparación de los datos.

- **Eliminación de variables:** La eliminación de variables será descrita posteriormente en la sección 4.3, con las metodologías de detección de variables muy correladas y luego aplicando los métodos citados anteriormente en la sección 2.2.3.

Capítulo 4

Modelo GLM

4.1. Distribución de la respuesta

Dado que el objetivo del presente trabajo es la elaboración de un modelo de regresión generalizada (GLM), la elección de una distribución adecuada para la variable respuesta es crucial para garantizar la validez y la precisión de las inferencias obtenidas en el modelo. De esta forma podemos realizar un modelo predictivo que trabaje directamente con la variable respuesta en su formato original, en lugar de realizar una transformación logarítimica de esta para que su distribución se asemeje a la de una normal, cumpliendo las asunciones de un modelo lineal no generalizado.

La hipótesis inicial es que nuestra variable respuesta, `n50_test` sigue una distribución Gamma ya que este tipo de distribución es adecuada para variables continuas y positivas que pueden presentar una asimetría en sus valores. Para establecer la validez de esta suposición, se plantean las siguientes hipótesis:

$$H_0 : n50\_test \sim \text{Gamma}(\alpha, \beta)$$

$$H_1 : n50\_test \not\sim \text{Gamma}(\alpha, \beta)$$

Donde α y β son los parámetros de forma y escala de la distribución Gamma, respectivamente. La hipótesis nula H_0 establece que la variable respuesta `n50_test` sigue una distribución Gamma con parámetros específicos α y β, mientras que la hipótesis alternativa H_1 establece que `n50_test` no sigue dicha distribución.

Para contrastar estas hipótesis, se han realizado varias pruebas estadísticas de bondad de ajuste, tales como el test de Kolmogorov-Smirnov y el test de Anderson-Darling, que permiten evaluar si la variable respuesta sigue efectivamente una distribución Gamma. Además, se han generado gráficos diagnósticos para visualizar el ajuste de la distribución Gamma a los datos.

A continuación, se detallan los métodos utilizados y los resultados obtenidos de estas pruebas y gráficos diagnósticos:

4.1.1. Pruebas de Bondad de Ajuste

Se llevaron a cabo pruebas de bondad de ajuste para comparar la distribución empírica de la variable respuesta con la distribución Gamma teórica. Estas pruebas evalúan si hay evidencia significativa de que los datos no siguen una distribución Gamma. Para ello se ajustó la variable respuesta a una Gamma y se procedió a realizar los siguientes test:

$$n50\_test \sim \text{Gamma}(3.7014, 0.5182)$$

- **Test de Kolmogorov-Smirnov:** Esta prueba compara la función de distribución empírica de los datos con la función de distribución acumulativa de la distribución Gamma ajustada. El resultado de la prueba, utilizando el paquete `fitdistrplus` [27], arrojó un valor-p de 0.5259. Este valor es mayor al nivel de significancia común 0.05, lo que indica que no se rechaza la hipótesis nula. Por lo tanto, no hay evidencia suficiente para decir que la variable `n50_test` no sigue una distribución Gamma.
- **Test de Anderson-Darling:** Esta prueba es más sensible a las diferencias en las colas de la distribución. Los resultados obtenidos, también utilizando `fitdistrplus`, indicaron que la hipótesis nula fue rechazada. Esto sugiere que, aunque el test de Kolmogorov-Smirnov no encontró evidencia suficiente para rechazar la hipótesis nula, el test de Anderson-Darling sí detectó diferencias significativas, particularmente en las colas de la distribución. Esto puede indicar que la distribución Gamma no captura totalmente los valores extremos de la variable `n50_test`.

4.1.2. Análisis descriptivo y gráficos diagnósticos:

- **Análisis descriptivo de la variable respuesta:** la variable respuesta `n50_test` sigue una distribución con las siguientes características que se pueden observar en la tabla 4.1.

Estadísticas descriptivas de n50_test			
N	Minimum	Maximum	Mean
389	1.1930	29.0740	7.1427
Standard deviation	Lower quartile	Median	Upper quartile
3.9807	4.3698	6.2579	9.1097

Tabla 4.1: Estadísticas descriptivas para la variable n50_test

- **Histograma con Curva de Densidad**: Se generó un histograma (figura 4.1) de la variable `n50_test` sobre el cual se superpuso la curva de densidad de la distribución Gamma ajustada. Este gráfico visualiza el ajuste de la distribución teórica a los datos observados.

Si bien es cierto que existe un intervalo del histograma que se desvía de la curva de la distribución Gamma, la variable respuesta parece ajustarse a una Gamma.

- **QQ-Plot**: Se realizó un QQ-Plot (figura 4.2) para comparar los cuantiles de la variable respuesta con los cuantiles de una distribución Gamma ajustada. En el QQ-Plot, se observó que los puntos se alinean razonablemente bien con la línea de referencia en la parte central del gráfico, aunque se desvian ligeramente en una de las colas indicando que son más pesadas que las esperadas, lo que explica los resultados del test de Anderson-Darling. Esto significa que existen valores extremos más altos que los que la distribución Gamma predice.

Los gráficos de diagnósticos aparecen a continuación:

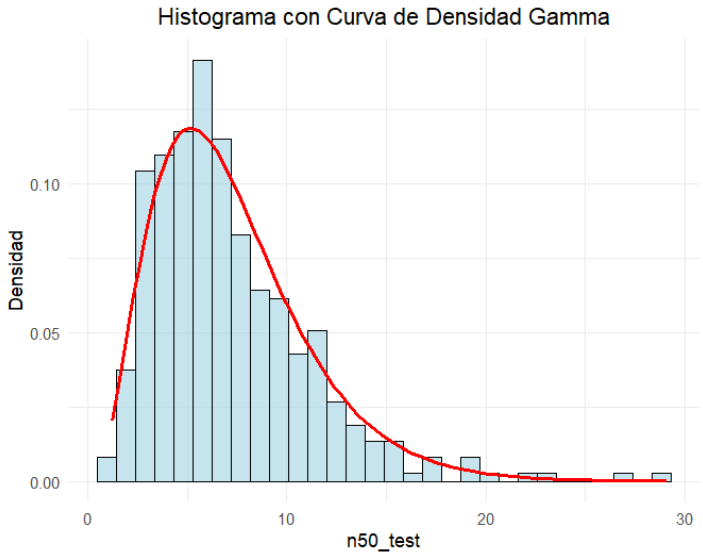


Figura 4.1: Histograma de n50_test y curva de densidad de la Gamma

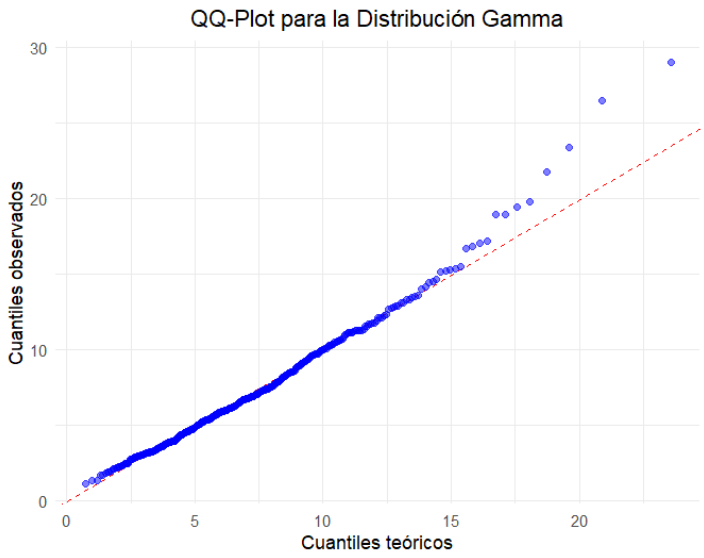


Figura 4.2: QQ-Plot de la distribución Gamma

Por lo tanto, tomando como referencia los resultados del test de Kolmogorov-Smirnov, se puede concluir que no hay evidencia suficiente para rechazar la hipótesis nula de que la variable `n50_test` sigue una distribución Gamma, lo que nos permite utilizar esta distribución en el modelo GLM. No obstante, los resultados del test de Anderson-Darling y los gráficos diagnósticos indican que la distribución Gamma no captura completamente los valores extremos de la variable respuesta. A pesar de estas discrepancias, la elección de la distribución Gamma es justificada en términos generales, aunque se debe considerar la posible influencia de los valores extremos en las inferencias del modelo.

4.2. Elección de la Función de Enlace

En el contexto de la elaboración de un modelo de regresión generalizado (GLM), la elección de la función de enlace es un aspecto fundamental que determina la relación entre la media de la variable respuesta y la combinación lineal de los predictores. La función de enlace define cómo la media de la variable respuesta, μ, se relaciona con el predictor lineal, η. En términos matemáticos, la función de enlace $g(\cdot)$ se define como:

$$\eta = g(\mu)$$

Para una distribución Gamma, la función de enlace canónica es la función inversa. Esto significa que la media de la variable respuesta se modela como la inversa de la combinación lineal de los predictores:

$$\eta = \frac{1}{\mu}$$

Sin embargo, en este trabajo, se optó por utilizar la función de enlace logarítmica. Esta función de enlace se define como:

$$\eta = \log(\mu)$$

La elección de la función de enlace logarítmica estuvo justificada por varias razones:

- **Análisis gráfico de devianza**: Se compararon gráficamente las devianzas residuales de modelos con diferentes funciones de enlace, para ello y partiendo de una adaptación al GLM del modelo inicial, con la selección de variables realizadas en el artículo original en el que se basa este trabajo [15], se comparaba con que valor de la función de enlace se conseguía una menor devianza residual. Para esto hemos utilizado la función `power`. A esta función se le pasó un vector ordenado de 51 posibles valores de lambda comprendidos entre 0 y 0.25 con el objetivo de encontrar el que produjera el menor valor de devianza residual. Aunque hubiese sido deseable obtener la devianza residual para valores negativos

de lambda, ya que el -1 es el valor del link canónico para una gamma, la función `power` presenta la limitación de no poder introducir valores menores a 0.

Una vez tenemos este vector con los diferentes valores de lambda, procedemos a ajustar el modelo para cada uno de los valores de este vector. Durante este proceso, registramos la devianza residual resultante para cada iteración del modelo. El objetivo es determinar el valor de lambda que minimiza la devianza. Al finalizar el ajuste con todos los valores de lambda especificados, identificamos cuál de estos valores produce la devianza mínima. Este valor óptimo de lambda nos indica qué configuración de la función de enlace power() proporciona el mejor ajuste del modelo a los datos observados. Una vez identificamos la devianza mínima, podemos seleccionar la función de enlace cuyo desempeño se asemeje más a este valor mínimo de devianza.

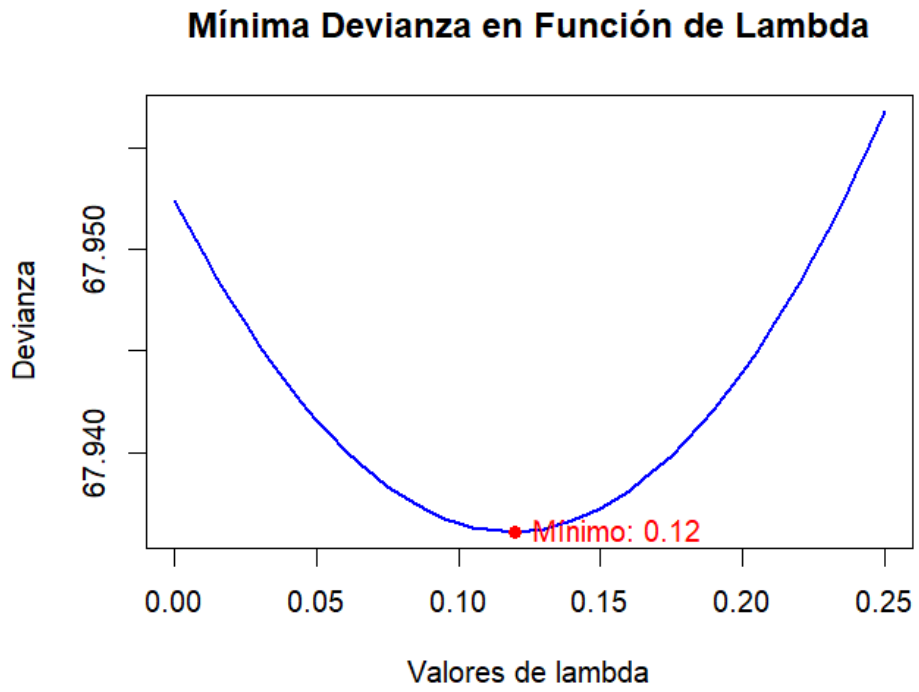


Figura 4.3: Minima Devianza en función de Lambda

Tal y como se observa en la figura 4.3, el valor mínimo de devianza residual obtenido fue 0.12, lo cual sugiere que la función de enlace logarítmica proporciona el mejor ajuste para el modelo, ya que su devianza residual es más cercana a cero en comparación con otras funciones de enlace. Para la función de enlace inversa (canónica), que tiene un valor de -1, y la función identidad, con un valor de 1, se observaron devianzas residuales más altas. Esto indica que estas funciones de enlace no logran ajustar tan bien los datos como lo hace la función logarítmica en términos de minimización de la devianza residual.

- **Comparación de modelos con diferentes enlaces**: Se realizó una evaluación directa de modelos utilizando diferentes funciones de enlace, incluyendo identidad, inversa y logarítmica. Entre estos, el modelo con función de enlace logarítmica mostró los mejores resultados en términos de ajuste y significancia estadística de los predictores, así como

de AIC, BIC y pseudo R^2. Esto consiguió proporcionar estimaciones más estables e interpretables.

Modelo	Función	Link	Res. Dev	AIC	BIC	r2ML
modelo	Gamma	log	67.952	1938.8	2073.8	0.4274676
modeloinv	Gamma	inv	69.125	1945.7	2080.7	0.4172981
modeloid	Gamma	id	68.938	1944.6	2079.6	0.4189209
modelomejordev	Gamma	power(0.12)	67.936	1938.7	2073.7	0.4276088

Tabla 4.2: Resumen de modelos GLM con diferentes funciones de enlace

En la tabla 4.2 se puede comparar en base a los diferentes links para el modelo inicial descrito en la sección 4.5, que el modelo que presenta el menor valor de devianza residual, también obtiene valores mejores para todas las métricas de calidad que se usan para evaluar el modelo.

Con todos estos análisis y comparaciones podemos respaldar la elección de la función de enlace logarítmica sobre la canónica en la especificación del modelo GLM para la variable respuesta `n50_test`.

4.3. Selección de variables

En la elaboración de modelos de regresión, la selección adecuada de variables juega un papel crucial para garantizar la precisión, la interpretabilidad y la generalización del modelo. En este capítulo, se detalla el proceso seguido para seleccionar las variables predictoras relevantes para el modelo de regresión generalizado (GLM).

Al disponer de un dataset de 203 variables, el proceso de selección, tal como se menciona anteriormente en la sección 3.1, pasó por una primera fase de filtrado por parte de nuestros colaboradores de la ETS de Arquitectura en la que nos indicaron las variables que resultaban de interés en el modelo. No obstante, consideramos necesario hacer una selección de variables por métodos estadísticos.

4.3.1. Eliminación de variables con coeficientes Aliased

Los coeficientes aliased se producen cuando las variables predictoras son linealmente dependientes o cuando su diseño experimental genera información redundante, lo que puede llevar a un modelo menos preciso e interpretable. En nuestro caso, al existir coeficientes aliased, no podíamos realizar la tabla de errores de tipo III en la que se ven las variables significativas.

Inicialmente, se procedió a identificar y eliminar aquellas variables categóricas que causaban coeficientes aliased en el modelo.

Para abordar este problema, se realizaron análisis de colinealidad y correlación para variables

categóricas y se eliminaron las variables redundantes, garantizando así que cada predictor en el modelo aportara información única.

En nuestro caso, tuvimos los primeros indicios de aliasing al realizar un heatmap o mapa de calor para las variables categóricas basado en el coeficiente de correlación de la *V de Crámer* 2.3.2.

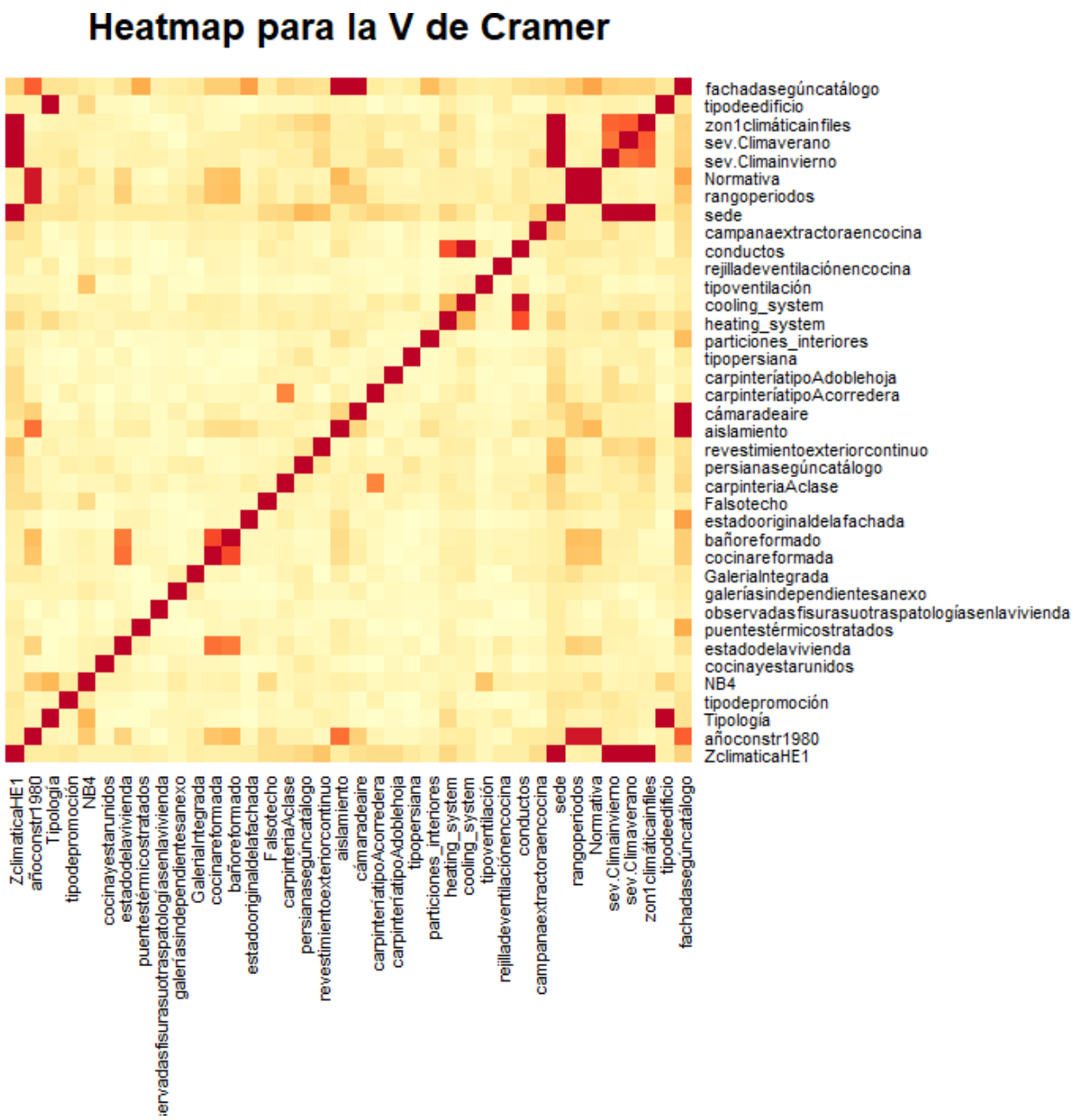


Figura 4.4: Heatmap de Cramer para las correlaciones

El aliasing se detectó también a través de la imposibilidad para calcular los errores de tipo III indicando que ciertas variables estaban confundiéndose con otras, haciendo difícil determinar sus impactos individuales en la variable dependiente.

De esta forma y tras usar tanto el mapa de calor (figura 4.4), como la tabla de errores de

tipo III, se eliminaron las siguientes variables: `añoconstr1980`, `sede`, `zon1climáticainfiles`, `zonaclimaticaHE1`, `añoconstrucción`, `sev.Climainvierno`, `sev.Climaverano`. Tras esto se eliminaron los problemas de coeficientes Aliased en el modelo que nos impedían calcular la tabla de errores de tipo III.

4.3.2. Eliminación de variables numéricas con alta correlación

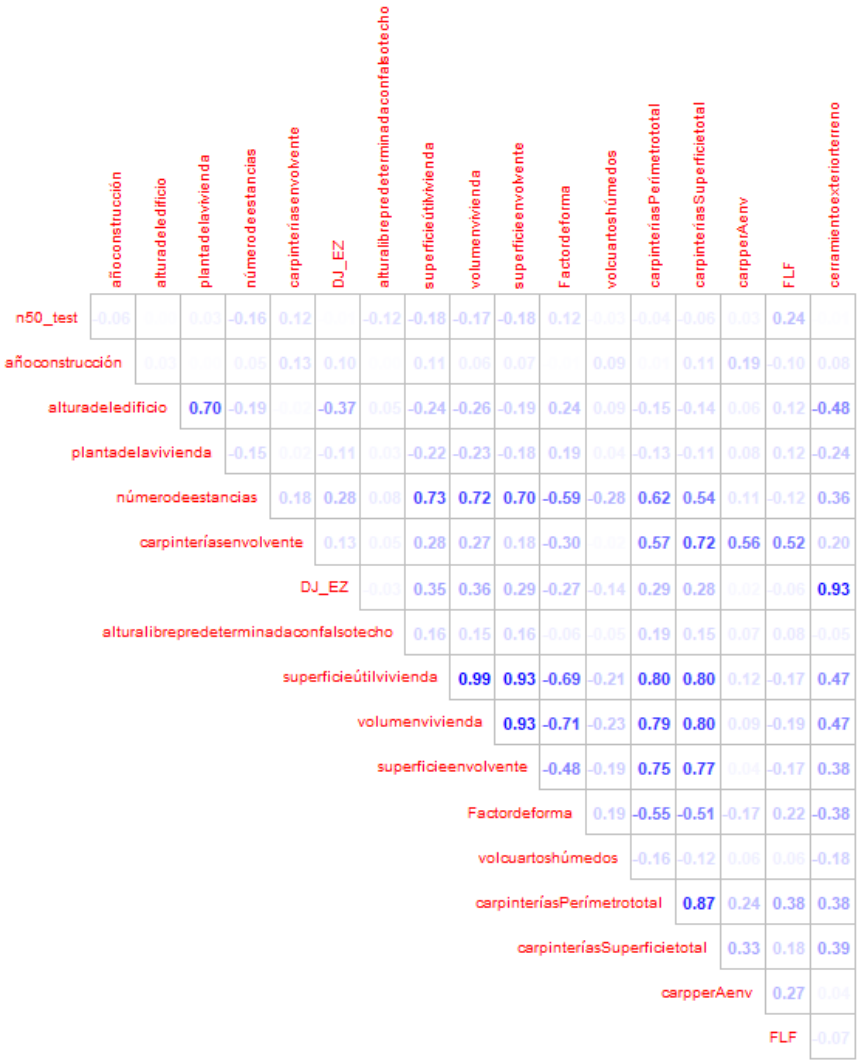


Figura 4.5: Matriz de correlaciones de Pearson

Posteriormente, se llevó a cabo un análisis para identificar y eliminar variables numéricas con alta correlación entre sí. La alta correlación entre variables puede introducir multicolinealidad en el modelo, lo cual dificulta la interpretación de los coeficientes y puede llevar a peores estimaciones de los parámetros del modelo.

La elección de realizarlo previo a una selección de variables backward o forward se debió a la menor cantidad de variables numéricas y a una mayor facilidad de cara a calcular las correlaciones para este tipo de variables frente a las categóricas con las herramientas que proporciona

R, como la matriz de correlación.

Para ello, a partir de la matriz de correlaciones de la figura 4.5, se fueron eliminando las variables muy correladas entre sí, con un valor superior a 0.8 o inferior a -0.8. De esta forma, de las 18 variables numéricas de las que se partía, nos quedamos con 12, eliminando:

En primer lugar `volumenvienda` debido a su correlación de casi 1 con `superficieenvolvente` y `superficieútilvivienda`. Eliminamos `cerramientoexteriorterreno` por su correlación con `DJ_EZ` y se elimina `carpinteríasSuperficietotal` ya que estaba muy correlada con varias variables. Se elimina también `colectivaposicion` por su alto número de valores faltantes como se ha mencionado anteriormente.

Finalmente se eliminan `superficieútilvivienda` y `carpinteríasPerímetrototal` dada su alta correlación con `superficieenvolvente`.

También, podemos observar en la figura 4.5, que la variable respuesta `n50_test` no está a priori muy relacionada con casi ninguna de las variables numéricas predictoras.

Una vez obtenido el dataset que se va a utilizar, planteamos un modelo, en el que aplicaremos la metodología de validación cruzada descrita en el siguiente apartado 4.4

4.4. Modelo final a partir de validación cruzada

Modelado a partir de las variables seleccionadas

La validación cruzada es una técnica esencial para evaluar la capacidad predictiva y la robustez de un modelo estadístico. En nuestro estudio, se implementa la validación cruzada 5-fold para asegurar que el modelo generalice bien a datos no observados, minimizando el riesgo de sobreajuste y proporcionando una estimación más fiable de su rendimiento.

El dataset final, compuesto por 12 variables numéricas y 31 categóricas, se utilizará para construir el modelo inicial. Este modelo servirá de base para aplicar métodos más avanzados de selección de variables.

La finalidad de este modelo es mejorar la predicción del modelo inicial [15]. Para ello se utiliza una metodología, que aunque es poco ortodoxa, es adecuada para nuestro caso. Se ha considerado que las interacciones mejoran la calidad del modelo, pero en un modelo con esa cantidad de variables, incluir todas las posibles interacciones resulta imposible, por lo que la decisión fue plantear un modelo con esas 43 variables, y a través de la metodología selección de variables tanto backward como forward, quedarnos con una lista de variables más reducida en la que aplicar ya las interacciones y realizar la selección final.

Como el modelo va a ser validado con una metodología 5-fold en la que nos quedaremos con el modelo que ofrezca el mejor $pseudoR^2$. Esta metodología se repetirá 5 veces para la selección forward y 5 para la backward. Para ello y previo a realizar esta selección, en primer lugar se

realiza una división del dataset de 289 observaciones en 5 grupos de tamaño equivalente, en nuestro caso 3 grupos de 79 observaciones, uno de 77 y uno de 75. Esta selección se hizo automáticamente a través del paquete de R `caret` [28].

Tras obtener esa selección de variables, para cada uno de los 5 subconjuntos de datos de entrenamiento y test se hace lo siguiente:

1. A partir de las observaciones del subconjunto de entrenamiento correspondiente, se procede a utilizar una selección automática de variables 2.2.3 para eliminarlas hasta quedarnos con las que tienen un p-valor inferior a 0.15. Esto se hace porque el objetivo de esta primera selección es obtener un subconjunto más reducido de variables en las que sí poder incluir las interacciones, simplificando el modelo. De esta forma nos quedamos con variables que, aunque su nivel de significación sea bajo, pudieran llegar a ser significativas en una interacción, descartando variables con un p-valor muy alto (superior a 0.15).

2. Posteriormente, se introducen términos de interacción entre las variables categóricas y numéricas seleccionadas en el paso anterior. Se ajusta un nuevo modelo que incluye estas interacciones y se realiza una selección final de variables aplicando nuevamente el método backward o forward según proceda con un umbral de p-valor de 0.1. Para esta fase se usan también las observaciones del subconjunto de entrenamiento. De esta forma, se seleccionan las variables significativas, verificando manualmente que tanto ellas como las interacciones tengan un p-valor menor a 0.1, dado que el método no siempre funciona correctamente.

3. Una vez se tienen las variables finales con las que va a contar el modelo, se ajusta el GLM y se comprueba en el subconjunto de test el $pseudoR^2$ que proporciona. El modelo elegido va a ser el que presente un mejor valor para esta métrica.

4. Finalmente, a partir de la selección de variables elegida, en base al mejor $pseudoR^2$ en el conjunto de validación, se construye un GLM en el que el conjunto de validación pasa a ser el 100 % de las observaciones, siendo este el modelo final.

Esto se realiza tanto para backward como para forward, obteniendo dos modelos finales que serán los que se comparen finalmente.

La selección de variables nos proporcionó los $pseudoR^2$ que se pueden observar en la tabla 4.3. En ella vemos los $pseudoR^2$ proporcionados por cada modelo aplicando el conjunto de entrenamiento y el de test. De esta forma los modelos que mejores resultados proporcionan para el conjunto de validación son el modelo 5 tanto para la selección backward como para la selección forward.

El número final de variables incluyendo interacciones de cada modelo se puede observar en la tabla 4.4. Las variables elegidas en los modelos finales (Modelo 5) forward y backward se pueden observar en la tabla 4.5.

	Backward		Forward	
	Train	Test	Train	Test
Modelo1	0.566	0.566	0.556	0.479
Modelo2	0.603	0.744	0.576	0.631
Modelo3	0.495	0.694	0.469	0.605
Modelo4	0.512	0.713	0.499	0.675
Modelo5	0.488	0.832	0.451	0.758

Tabla 4.3: pseudoR^2 de los modelos generados con selección automática de variables

	Backward	Forward
	Número de Variables	Número de Variables
Modelo1	17	16
Modelo2	23	19
Modelo3	17	13
Modelo4	15	13
Modelo5	17	14

Tabla 4.4: Número de variables de los modelos generados con selección automática de variables

Backward	**Forward**
ZclimaticaHE1	ZclimaticaHE1
añoconstrucción	añoconstrucción
bañoreformado	bañoreformado
Falsotecho	Falsotecho
carpinteriaAclase	carpinteriaAclase
tipopersiana	
conductos	conductos
campanaextractoraencocina	campanaextractoraencocina
carpinteríasenvolvente	carpinteríasenvolvente
ALPDFT[1]	ALPDFT
Factordeforma	Factordeforma
ZclimaticaHE1:ALPDFT	
añoconstrucción:ALPDFT	
tipopersiana:ALPDFT	
campanaextractoraencocina:ALPDFT	
carpinteriaAclase:Factordeforma	carpinteriaAclase:Factordeforma
conductos:Factordeforma	conductos:Factordeforma
	estadooriginaldelafachada
	ZclimaticaHE1:añoconstrucción

[1] ALPDFT: alturalibrepredeterminadaconfalsotecho

Tabla 4.5: Comparación de variables entre Backward y Forward

Observamos que ambas selecciones comparten 10 variables en común y 2 de los términos de interacción, por lo que son modelos bastante similares.

De cara a la elección del modelo final entre estos dos últimos seleccionados, en los que a

priori el elegido mediante una selección backward presenta un mejor rendimiento en el subconjunto de test, vamos a evaluarlos en el total de las observaciones y en base a las métricas descritas en la sección 2.2.3 se elegirá el que mejores resultados proporcione.

A continuación, en la tabla 4.6, se presentan los valores de las métricas para cada modelo evaluados en el total de observaciones:

	Num Var	Devianza Residual	Df	AIC	BIC	pseudo R^2
Modelo Backward	17	57.811	354	1863.817	2006.506	0.485
Modelo Forward	14	59.223	358	1865.439	1992.274	0.472

Tabla 4.6: Comparación de métricas entre Modelo Backward y Modelo Forward

Se puede observar que para ambos modelos los resultados proporcionados por las métricas son muy similares, proporcionando una mejor significación el modelo con la selección Backward, de un 1.3 % más, y teniendo un menor valor de devianza residual y de AIC que el modelo Forward, sin embargo este último proporciona un mejor BIC y los valores de las otras métricas son muy cercanos, a pesar de tener menos variables, pudiendo llegar a tener un menor sobreajuste.

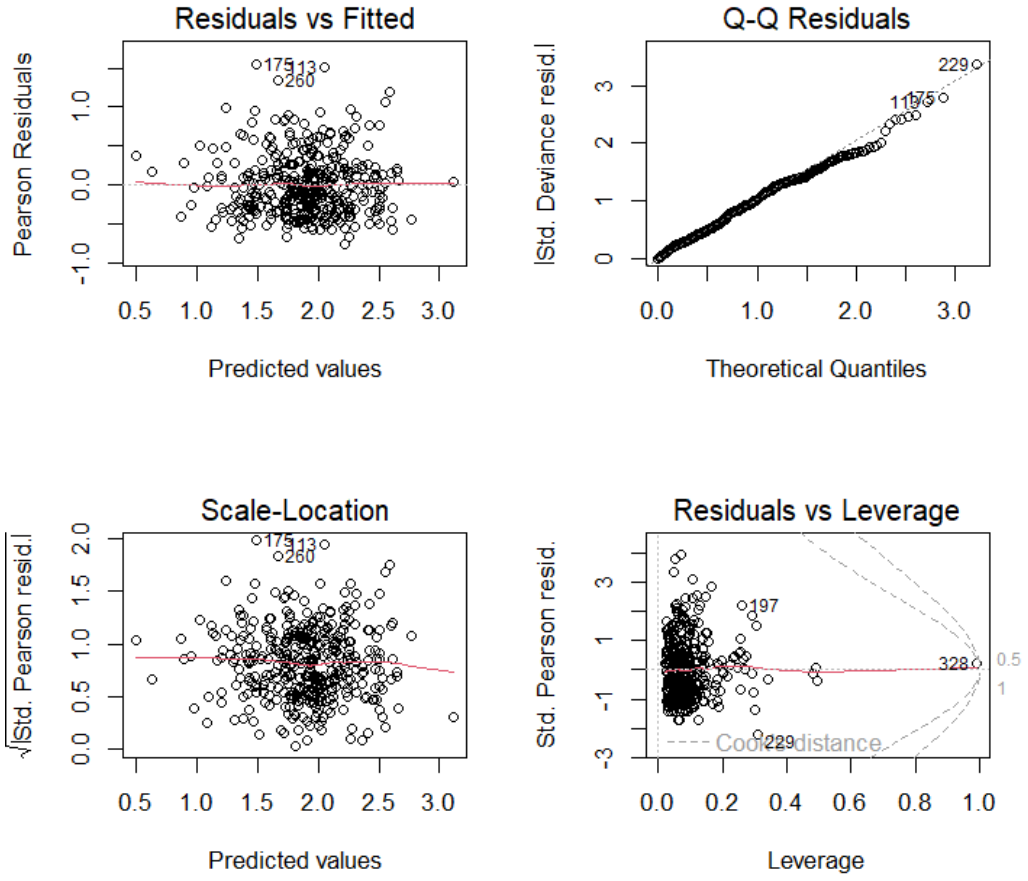


Figura 4.6: Gráficos de residuales del modelo Backward

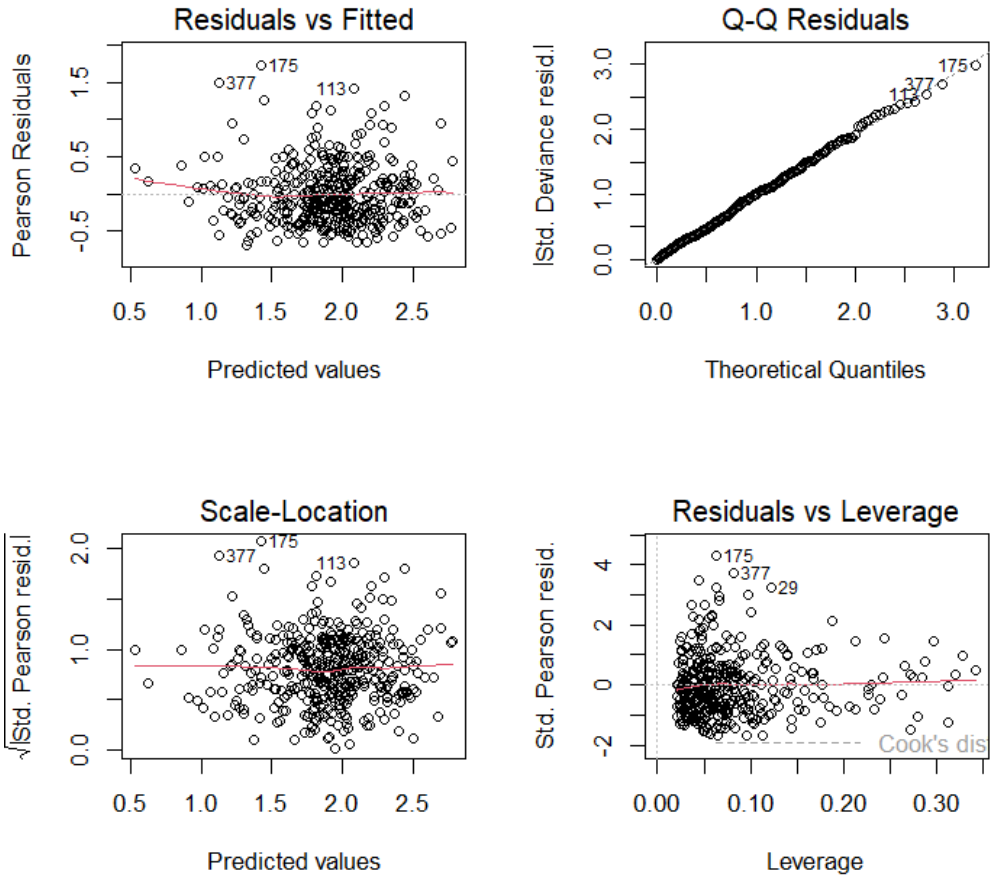


Figura 4.7: Gráficos de residuales del modelo Forward

En cuanto a los gráficos de residuales (figuras 4.6 y 4.7), podemos observar que los de Residuales vs Predichos no presentan casi diferencias entre ambos modelos, y cumplen las asunciones que buscamos de cara a validar el modelo. Sin embargo el Q-Q plot de los residuos del modelo Backward presenta más desviaciones de la línea diagonal que el de los residuales del modelo Forward.

El gráfico en el que hay más diferencias es el de Residuales vs Leverage, en el que el del modelo Backward tiene un valor muy alto de leverage para la observación 328, pudiendo llegar a ser un punto de influencia que nos lleve a obtener coeficientes sesgados en el modelo que interactuen con esa observación y estimaciones que no sean representativas en la mayoría de datos. Se ha probado a eliminar esta observación del modelo, obteniendo la misma selección de variables, y resultados similares, por lo que no parece influir, aunque al hacer el gráfico de residuales seguimos obteniendo puntos con un levarage muy cercano a uno.

Sin embargo y dado que el modelo Forward también incluye esta observación, pero no presenta ese alto leverage se puede considerar que el modelo forward es más robusto frente a esa observación específica.

Todo esto, nos lleva a elegir el modelo Forward como modelo final, al no presentar los pro-

blemas del Backward, ser más sencillo y proporcionar resultados similares.

Una vez que nos decantamos por el modelo que usa la selección forward, se procede a comprobar si las variables pueden llegar a presentar algún problema de multicolinearidad no detectado en la fase anterior de selección de variables. Los resultados resultantes aparecen en los siguientes gráficos:

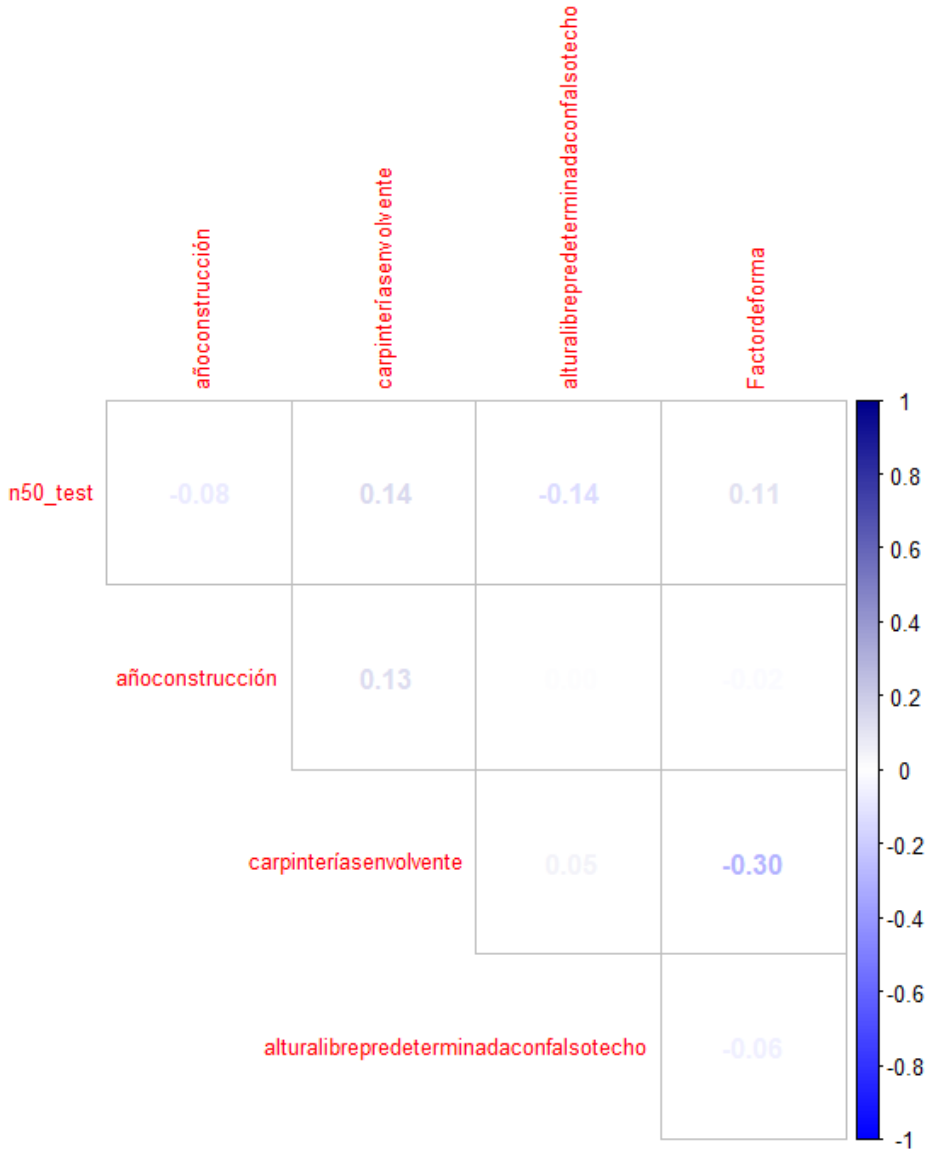


Figura 4.8: Correlación de Pearson para las variables numéricas

En el caso de la figura 4.8, no existen problemas de multicolinearidad, existiendo valores muy bajos para la correlación entre variables numéricas, pudiéndose observar valores de casi 0 entre `añoconstrucción` con `alturalibrepredeterminadafalsotecho` y `Factordeforma`. Se puede ver también que la correlación entre la respuesta y las variables predictoras es baja.

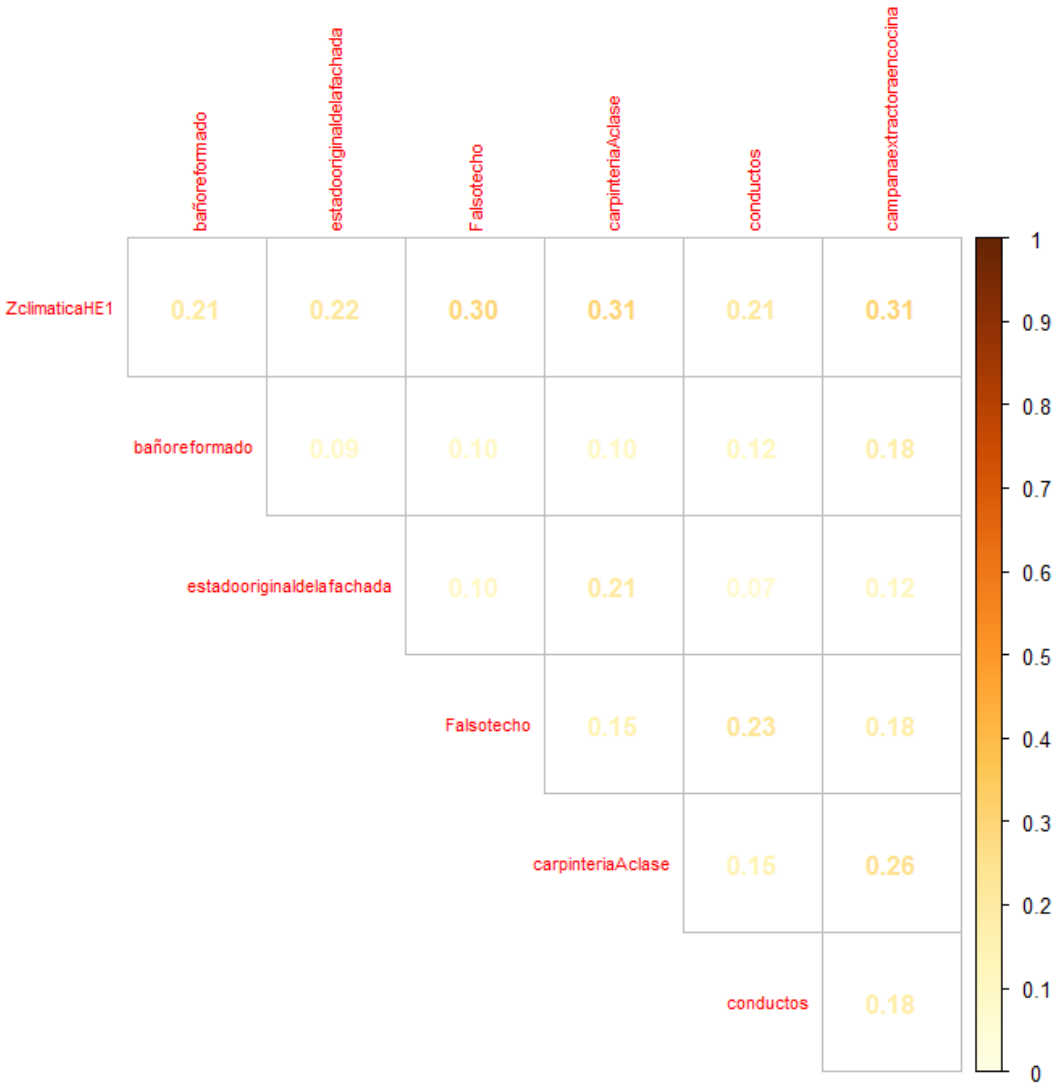


Figura 4.9: V de Cramér para las variables categóricas

En cuanto a las variables categóricas que aparecen en la figura 4.9, tal y como se describe en la sección 2.3.2 para la V de Cramér, los valores de máximo 0.31 no nos indican una asociación fuerte entre variables, por lo que no encontramos problemas de multicolinearidad tampoco entre estas variables.

Por todo esto, elegimos un modelo con 11 variables principales y tres interacciones que nos explica un 47.2 % de la variabilidad de la variable respuesta. La tabla de coeficientes se puede observar en la tabla 4.7 y la tabla ANOVA de este modelo se puede observar en la tabla 4.8. En esta última, podemos ver la variabilidad de la variable respuesta dependiendo de cada variable explicativa, de la interacción incluida en el modelo, y si estas son significativas o no. La variable `alturalibrepredeterminadaconfalsotecho` ha dejado de ser significativa al incluir el conjunto total de variables. `Conductos` ha dejado de ser significativa a nivel 0.05 pero sí a nivel 0.1.

Las interacciones que hemos elegido como significativas expresan que los efectos de las variables numéricas son significativamente diferentes para los distintos niveles de la variable

categórica con la que tienen la interacción. Este es el caso de la variable `Factordeforma` no es significativa pero su interacción con `carpinteriaAclase` y con `conductos` sí que lo es. Los efectos de `añocontrucción` también son diferentes para los múltiples niveles de `ZclimaticaHE1`.

La tabla 4.7 nos sirve para estimar el valor de `n50_test` si dispusieramos de valores de entrada para nuestras variables predictoras, en el que por ejemplo en el caso de que la zona climática sea Valladolid, el valor de la variable respuesta se decrementaría en 33.48 unidades a igualdad del resto de variables.

Cabe destacar que las variables excluidas en nuestra selección, no tienen por que ser irrelevantes, si no que su efecto ha podido ser capturado por otras variables que sí hemos incluido en el modelo.

Variables	**Estimado**	**Std. Error**	**Pr(>|t|)**	**Significancia**
(Intercept)	10.091144	3.554145	0.004780	**
carpinteriaAclase-Clase2	-1.120702	0.560033	0.046132	*
carpinteriaAclase-Clase3	-1.473943	0.573899	0.010625	*
carpinteriaAclase-Clase4	-2.798300	0.725999	0.000137	***
ZclimaticaHE1Bilbao-C1	6.884544	5.678690	0.226179	
ZclimaticaHE1Canarias-$\alpha 3$	-3.209006	17.565461	0.855146	
ZclimaticaHE1Madrid-C3	-7.981988	4.786625	0.096277	.
ZclimaticaHE1Málaga-A3	8.992100	10.808432	0.405990	
ZclimaticaHE1Sevilla-B4	-8.416036	5.403978	0.120264	
ZclimaticaHE1Valladolid-D2	-33.484837	13.027313	0.010563	*
conductos-Sí	-0.756567	0.411551	0.066842	.
carpinteríasenvolvente	0.054909	0.011539	2.84×10^{-6}	***
Factordeforma	-0.296678	0.409276	0.468996	
campanaextractoraencocina-Conductovertical	-0.400547	0.087054	5.84×10^{-6}	***
campanaextractoraencocina-A fachada	-0.287488	0.086128	0.000933	***
bañoreformado-Sí	-0.216239	0.049232	1.48×10^{-5}	***
añoconstrucción	-0.003889	0.001791	0.030549	*
Falsotecho-No	-0.066536	0.068516	0.332157	
Falsotecho-Sí	0.250152	0.065057	0.000143	***
estadooriginaldelafachada-Sí	0.376491	0.149957	0.012491	*
alturalibrepredeterminadaconfalsotecho	-0.027537	0.029403	0.349618	
conductos-Sí:Factordeforma	0.758643	0.315605	0.016735	*
ZclimaticaHE1Bilbao-C1:añoconstrucción	-0.003699	0.002877	0.199491	
ZclimaticaHE1Canarias-$\alpha 3$:añoconstrucción	0.001220	0.008831	0.890167	
ZclimaticaHE1Madrid-C3:añoconstrucción	0.003991	0.002425	0.100716	
ZclimaticaHE1Málaga-A3:añoconstrucción	-0.004651	0.005444	0.393563	
ZclimaticaHE1Sevilla-B4:añoconstrucción	0.004249	0.002733	0.120889	
ZclimaticaHE1Valladolid-D2:añoconstrucción	0.016603	0.006576	0.012002	*
carpinteriaAclase-Clase2:Factordeforma	0.745710	0.438865	0.090154	.
carpinteriaAclase-Clase3:Factordeforma	0.923370	0.450421	0.041091	*
carpinteriaAclase-Clase4:Factordeforma	1.799154	0.564710	0.001569	**

Nota: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$
Los niveles que no aparecen tienen parámetro 0, pues son la clase referencia de la variable

Tabla 4.7: Ecuación del GLM

Variables	LR Chisq	Df	Pr(>Chisq)	Significancia
carpinteriaAclase	17.5121	3	0.0005544	***
ZclimaticaHE1	16.1978	6	0.0127305	*
conductos	3.1658	1	0.0751947	.
carpinteríasenvolvente	24.2377	1	8.515×10^{-7}	***
Factordeforma	0.5664	1	0.4516859	
campanaextractoraencocina	22.7983	2	1.121×10^{-5}	***
bañoreformado	19.3545	1	1.086×10^{-5}	***
añoconstrucción	4.9959	1	0.0254081	*
Falsotecho	17.1358	2	0.0001901	***
estadooriginaldelafachada	5.7081	1	0.0168870	*
alturalibrepredeterminadaconfalsotecho	0.8870	1	0.3462808	
conductos:Factordeforma	5.4448	1	0.0196269	*
ZclimaticaHE1:añoconstrucción	16.4186	6	0.0116750	*
carpinteriaAclase:Factordeforma	11.8561	3	0.0078929	**

Nota: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$

Tabla 4.8: Tabla ANOVA del GLM

4.5. Comparación con el modelo inicial

Puesto que uno de los objetivos de este trabajo era el de mejorar el modelo inicial [15] a través de un GLM que utilizara como variable respuesta la variable sin transformar, a continuación se muestra la comparativa entre ambos modelos.

La principal diferencia entre ambos modelos es la variable respuesta utilizada, mientras que en el inicial se usa el logaritmo de la variable n50_test, en nuestro modelo se usa la variable en su escala original. La implicación más inmediata de esto es que para el primer modelo podemos utilizar un modelo lineal o LM y para el segundo la opción elegida es un GLM. Otro punto a tener en cuenta es que para el modelo inicial se usa el conjunto de datos con 392 observaciones, puesto que no hizo falta eliminar las observaciones 123, 125 y 289 como se comentó anteriormente en el apartado 3.2.

Tabla 4.9: Comparación de variables entre los modelos inicial y final

Variables en el modelo inicial	Variables en el modelo final
ZclimaticaHE1	ZclimaticaHE1
añoconstr1980	añoconstrucción
estadodelavivienda	
persianasegúncatálogo	
carpinteríatipoAmaterial	
carpinteriaAclase	carpinteriaAclase
Falsotecho	Falsotecho
Tipología	

Continúa en la siguiente página

Tabla 4.9 – Continuación de la página anterior

Variables en el modelo inicial	**Variables en el modelo final**
heating_system	
NB4	
DJ_EZ	
carpinteríasenvolvente	
	Factordeforma
	estadooriginaldelafachada
	carpinteríasenvolvente
	conductos
	campanaextractoraencocina
	bañoreformado
	alturalibrepredeterminadaconfalsotecho
añoconstr1980*DJ_EZ	
Tipología*DJ_EZ	
	conductos:Factordeforma
	ZclimaticaHE1:añoconstrucción
	carpinteriaAclase:Factordeforma

Variables	**Sum Sq**	**Df**	**Pr(>F)**	**Significancia**
(Intercept)	8.732	1	3.093×10^{-11}	***
ZclimaticaHE1	9.226	6	2.070×10^{-8}	***
añoconstr1980	2.612	1	0.0002061	***
estadodelavivienda	1.120	1	0.0145346	*
persianasegúncatálogo	1.841	3	0.0204160	*
carpinteríatipoAmaterial	1.985	3	0.0144438	*
carpinteriaAclase	4.688	3	2.038×10^{-5}	***
Falsotecho	3.172	2	0.0002380	***
Tipología	1.227	1	0.0105634	*
heating_system	1.832	4	0.0447626	*
NB4	2.844	4	0.0046274	**
DJ_EZ	0.599	1	0.0734000	.
carpinteríasenvolvente	2.904	1	9.260×10^{-5}	***
añoconstr1980:DJ_EZ	2.541	1	0.0002510	***
Tipología:DJ_EZ	1.112	1	0.0149089	*
Residuals	66.681	359		

Nota: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$

Tabla 4.10: Tabla ANOVA del LM

En la tabla 4.9, donde se muestran las variables de ambos modelos, se puede observar que únicamente coinciden las variables `ZclimaticaHE1`, `carpinteriaAclase` y `Falsotecho`. `añoconstrucción` se consideró en el modelo inicial categorizada y en el final como variable numérica. El resto de variable e interacciones son diferentes.

El modelo inicial se presenta en la tabla ANOVA 4.10 y en la tabla 4.11, de esta forma obtenemos un modelo con 12 variables y 2 interacciones con un R^2 de 0.4274.

Variables	Estimado	Std. Error	Pr(>|t|)	Significancia
(Intercept)	2.591863	0.378017	3.09×10^{-11}	***
ZclimaticaHE1Bilbao-C1	-0.356665	0.094726	1.94×10^{-4}	***
ZclimaticaHE1Canarias-$\alpha3$	-0.630074	0.154624	5.67×10^{-5}	***
ZclimaticaHE1Madrid-C3	-0.054603	0.072342	0.450868	
ZclimaticaHE1Málaga-A3	-0.284267	0.100454	0.004919	**
ZclimaticaHE1Sevilla-B4	-0.085030	0.078419	0.278957	
ZclimaticaHE1Valladolid-D2	-0.576613	0.126418	6.99×10^{-6}	***
añoconstr1980$\geq$1981	0.329296	0.087812	2.06×10^{-4}	***
estadodelavivienda-reforma integral	-0.136814	0.055713	0.014535	*
persianasegúncatálogo-P.02	-0.051758	0.110079	0.638507	
persianasegúncatálogo-P.03	-0.318012	0.158078	0.044994	*
persianasegúncatálogo-Sin persiana	-0.195431	0.118487	0.099944	.
carpinteríatipoAmaterial-aluminio	0.002886	0.205631	0.988811	
carpinteríatipoAmaterial-madera	0.226616	0.209894	0.281014	
carpinteríatipoAmaterial-PVC	-0.070890	0.215917	0.742860	
carpinteriaAclase-Clase2	-0.274590	0.086020	0.001537	**
carpinteriaAclase-Clase3	-0.341782	0.092742	2.64×10^{-4}	***
carpinteriaAclase-Clase4	-0.596492	0.119611	9.57×10^{-7}	***
Falsotecho-No	-0.048358	0.062841	0.442087	
Falsotecho-Sí	0.264361	0.069383	1.63×10^{-4}	***
Tipología-unifamiliar	-0.411548	0.160116	0.010563	*
heating_system-otros	-0.088219	0.186440	0.636375	
heating_system-sin calefacción	-0.187202	0.102943	0.069821	.
heating_system-superficie radiante	-0.302289	0.182413	0.098359	.
heating_system-unidades puntuales	-0.261495	0.085587	0.002417	**
NB4-1	-0.263389	0.259158	0.310158	
NB4-2	-0.427361	0.260737	0.102079	
NB4-3	-0.519750	0.268672	0.053836	.
NB4-4	-0.610133	0.280302	0.030154	*
DJ_EZ	0.003648	0.002032	0.073400	.
carpinteríasenvolvente	0.045459	0.011497	9.26×10^{-5}	***
añoconstr1980$\geq$1981:DJ_EZ	-0.009943	0.002688	2.51×10^{-4}	***
Tipología-unifamiliar:DJ_EZ	0.008957	0.003661	0.014909	*

Nota: *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$, . $p < 0.1$
Los niveles que no aparecen tienen parámetro 0, pues son la clase referencia de la variable

Tabla 4.11: Ecuación del LM

Finalmente, el modelo lineal presenta los siguientes gráficos de residuos (figura 4.10), en los que podemos ver como sí se respetan las asunciones del modelo 2.1.1, en el que en el gráfico de residuos frente a valores predichos, estos no presentan ningún patrón y están distribuidos aleatoriamente. En el Q-Q plot las observaciones no se alejan de la línea diagonal. Adicionalmente, no hay observaciones con un Leverage alto, por lo que posiblemente no existan problemas relacionados con los puntos de influencia.

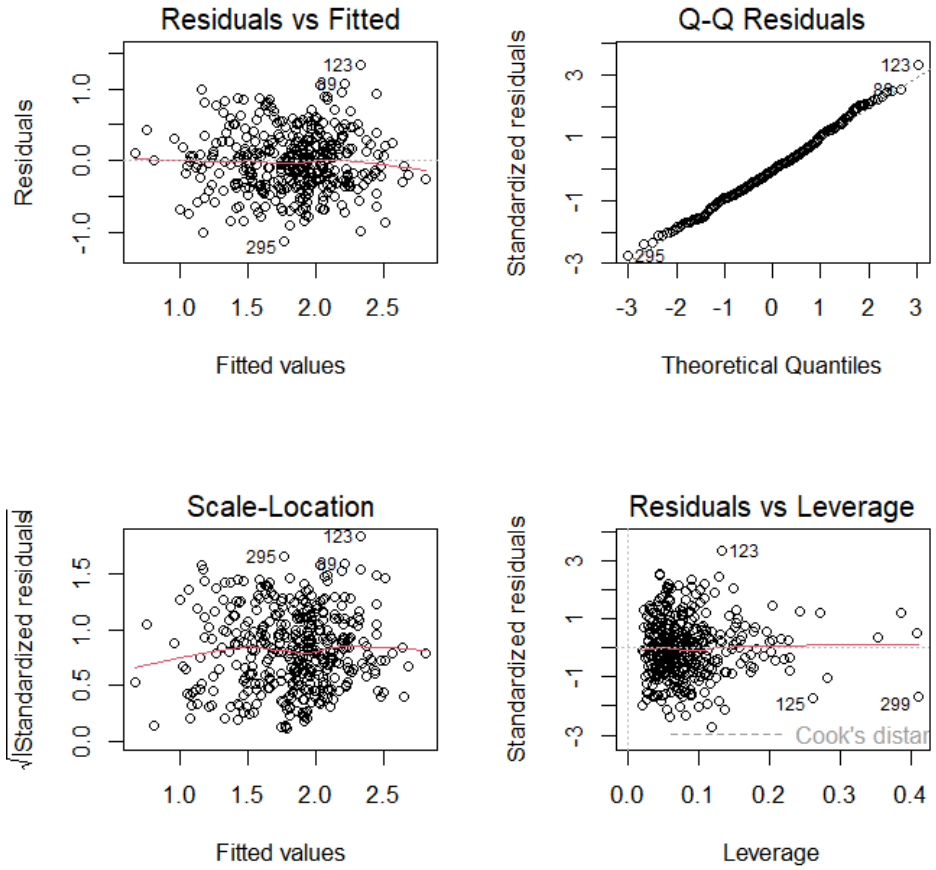


Figura 4.10: Gráficos de residuales del modelo lineal

Se probó también un modelo intermedio (GLM) con las mismas variables del modelo inicial, pero usando la variable respuesta sin transformar y una función de enlace logarítmica para la gamma. Los resultados de esta comparación entre estos tres modelos aparece en la tabla 4.12. Los AIC y BIC de los modelos varían mucho entre el modelo lineal y los GLM debido al cálculo de la verosimilitud que aplican, por lo que no son comparables, pero se puede observar que hay una mejora de casi 5 puntos en el porcentaje de la variabilidad explicada, y una mejora de la devianza y del BIC, el criterio que más penaliza el sobreajuste con respecto al GLM realizado con las variables iniciales.

Modelo	Función	Link	Res. Dev	AIC	BIC	R^2
Modelo Lineal Inicial	Gaussian		66.681	486.1	621.1	0.428
GLM Inicial Logarítmico	Gamma	log	67.952	1938.8	2073.8	0.427
Forward	Gamma	log	59.223	1865.4	1992.3	0.472

Tabla 4.12: Comparación de métricas entre los modelos inicial y final

Podemos concluir que el modelo final elegido mejora en un 10 % el R^2 anterior, indicando una mayor capacidad para explicar la variabilidad en los datos. Esta mejora es significativa, ya

que un aumento en el R^2 implica que el modelo captura mejor las relaciones entre las variables independientes y la dependiente.

Al mantener el mismo número total de predictores, 14, aunque en el GLM tenemos una variable menos y un término de interacción más que en el LM, minimizamos el riesgo de caer en un modelo sobreajustado. El término de interacción adicional en el GLM permite capturar relaciones no lineales y dependencias complejas entre las variables, mejorando así la precisión del modelo sin aumentar el número total de predictores.

Adicionalmente, la metodología de validación cruzada garantiza que el modelo no solo se ajusta bien a un conjunto específico de datos, sino que también generaliza a otros datos no considerados. Esto reduce la variabilidad y establece una medida más confiable de su desempeño. La validación cruzada es una técnica robusta que utiliza múltiples subconjuntos de datos para entrenar y validar el modelo, asegurando que la evaluación del rendimiento del modelo sea más precisa y menos sesgada.

La validación cruzada también ayuda a evitar el sobreajuste al proporcionar una evaluación más equilibrada del rendimiento del modelo, asegurando que el modelo elegido no solo funciona bien en los datos de entrenamiento, sino también en nuevos datos. Esto es crucial en aplicaciones del mundo real donde el modelo debe ser capaz de generalizar y realizar predicciones precisas en diferentes conjuntos de datos.

En resumen, el modelo GLM final no solo mejora el R^2 del modelo inicial, sino que también se beneficia de una estructura más robusta y generalizable, validada a través de una metodología rigurosa de validación cruzada. Esto nos permite tener mayor confianza en la capacidad del modelo para desempeñarse bien en diversos escenarios y con datos nuevos.

Capítulo 5

Conclusiones y trabajo futuro

Este trabajo surgió como una continuación de un estudio previamente realizado por mis tutores [14, 15]. El objetivo fue mejorar el modelo existente mediante el uso de un modelo lineal generalizado (GLM), utilizando la variable respuesta en su formato original. A través de este GLM, se buscó determinar las variables que influyen en la estanqueidad al aire de los edificios residenciales de España. Para desarrollar este modelo se empleó la herramienta R, utilizando datos de una muestra representativa con 203 variables y 392 valores. Para la elección del modelo final, se utilizó un método de validación cruzada, tal y como se propuso en la conferencia [15].

Para la selección del GLM final, se realizaron varios pasos previos, empleando métodos estadísticos descriptivos e inferenciales para filtrar valores y variables que pudieran afectar negativamente la capacidad predictiva del modelo.

En primer lugar, se comprobó si la variable respuesta seguía una distribución Gamma mediante pruebas de bondad de ajuste, como la de Kolmogorov-Smirnov. Los resultados no proporcionaron suficiente evidencia para rechazar la hipótesis nula de que la distribución seguía una Gamma. Además, los gráficos descriptivos mostraron una similitud con la distribución Gamma. Por tanto, se decidió que la distribución de la respuesta sería Gamma.

Una vez decidida la distribución, se eligió la función de enlace adecuada para el modelo. La función elegida fue la logarítmica, ya que proporcionaba mejores resultados que la canónica. La metodología para esta decisión se obtuvo del libro [17], en sus secciones 11.3 y 12.6.

El proceso de selección de variables fue clave para obtener un GLM comprensible y preciso. Partiendo del dataset de 203 variables, se refinaron las variables predictoras mediante varios pasos clave. En primer lugar, como se ha descrito en la sección 4.3.1 se eliminaron las variables categóricas que causaban coeficientes aliased, asegurando información única en cada predictor. Luego, se eliminaron variables numéricas con alta correlación para reducir la multicolinealidad, ver sección 4.3.2. Resultando un conjunto de 12 variables numéricas y 31 variables categóricas. También se eliminaron 3 observaciones que consideramos outliers y posibles puntos de influencia en la selección de variables del modelo. Esto nos dejó 389 observaciones.

Finalmente y tal y como se describe en la sección 4.4, se redujo la cantidad de variables, aplicando una metodología automática para incluir las interacciones. Este enfoque permitió identificar las variables y sus interacciones más significativas, resultando un modelo más simple y robusto. A fin de hacer hincapié en este último atributo, se aplicó validación cruzada 5-fold para elegir el modelo que mejores resultados proporcionaba en el conjunto de test, obteniendo así nuestro modelo final.

Entre la selección backward y forward, se eligió la forward, ya que explicaba un porcentaje similar de la variabilidad y evitaba problemas detectados en los gráficos residuales del modelo backward, como desviaciones en el Q-Q plot y puntos de influencia con alto leverage.

Esta selección de variables mejoró la precisión, la interpretabilidad y la calidad del modelo de regresión. El modelo final proporciona un R^2 de 0.472, mejorando en un 10 % el modelo anterior, y manteniendo el número total de términos entre variables y términos de interacción. Esto evita un sobreajuste y permite capturar relaciones más complejas entre las variables.

El modelo final, elegido mediante validación cruzada, ofrece más robustez frente a la introducción de nuevas observaciones, proporcionando mejores resultados en comparación con modelos cuyo conjunto de entrenamiento y validación es el mismo, reduciendo nuevamente el riesgo de sobreajuste.

Por todo esto podemos concluir que el utilizar un GLM con la variable respuesta sin transformar, sí que ha supuesto una mejora sobre el modelo en el que se utiliza un LM con la variable transformada. Adicionalmente este modelo es más robusto que el modelo inicial sin incrementar la complejidad, debido a la metodología de validación cruzada.

A partir de este TFG surgen diversas propuestas para trabajos futuros. Dado que anteriormente se han publicado modelos predictivos sobre la estanqueidad al aire de los edificios residenciales, este trabajo podría publicarse como una continuación de los artículos previos a los que amplía [14, 15]. Además, sería interesante explorar la aplicación de modelos predictivos más complejos, utilizando técnicas de aprendizaje automático, que no han sido probados hasta ahora. Si se dispusiera de un nuevo conjunto de observaciones que ampliara el ya utilizado, sería posible evaluar y validar el rendimiento de este modelo.

Apéndices

Apéndice A

Código del modelo final

A.0.1. Introducción

A continuación se incluirá el código utilizado para obtener el modelo final. Dado que el modelo final ha sido obtenido mediante validación cruzada, el código a utilizar es el mismo para todos los modelos, solo hay que modificar los conjuntos de entrenamiento y test para cada caso, y la metodología de selección backward o forward según corresponda.

A.0.2. Obtención de los datasets de entrenamiento y test

```
# Cargar el paquete caret
library(caret)

# Crear indices para las particiones usando un vector del tamano del dataset total
set.seed(123)  # Para reproducibilidad
folds <- createFolds(1:nrow(datosp1), k = 5, list = TRUE, returnTrain = FALSE)

# Lista para almacenar los datasets
datasets <- list()

# Iterar sobre cada particion para crear los subconjuntos de datos
for(i in 1:length(folds)) {
  # Indices del conjunto de datos para la particion actual
  indices <- folds[[i]]

  # Crear subconjunto de datos
  datasets[[i]] <- datosp1[indices, ]
}

test_data <- list()
train_data <- list()

for(i in 1:length(datasets)) {
  # Conjunto de validacion
  test_data[[i]] <- datasets[[i]]
  # Conjunto de entrenamiento (todos los demas subconjuntos)
  train_data[[i]] <- do.call(rbind, datasets[-i])}
```

A.0.3. Obtención del modelo final

En primer lugar se van a instalar los paquetes `car` [29] para poder realizar las tablas ANOVA y `pscl` [30] para obtener el pseudoR^2.

A continuación aparece el código R para la obtención del modelo final por validación cruzada y metodología de selección forward. Para ello, tras obtener los datasets de entrenamiento y validación en la sección anterior, se aplica la metodología descrita en la sección 4.4. Si quisiéramos realizar la selección backward, en las líneas 16 y 94 habría que cambiar el parámetro *direction* por "backward".

```
1  train5<-train_data[[5]]
2  test5<-test_data[[5]]
3
4  #Partimos del modelo con todas las observaciones, usando el subconjunto de entrenamiento
       correspondiente
5  modelocompleto5 <- glm(n50_test ~ ZclimaticaHE1 + anoconstr1980 + Tipologia +
       anoconstruccion + alturadeledificio + plantadelavivienda +
6                         tipodepromocion + numerodeestancias + NB4 + cocinayestarunidos +
                              estadodelavivienda + puentestermicostratados +
7                         observadasfisurasuotraspatologiasenlavivienda +
                              galeriasindependientesanexo + GaleriaIntegrada +
                              cocinareformada +
8                         banoreformado + estadooriginaldelafachada + Falsotecho +
                              carpinteriaAclase + carpinteriatipoAmaterial +
9                         persianaseguncatalogo + revestimientoexteriorcontinuo +
                              aislamiento + camaradeaire + carpinteriatipoAcorredera +
10                        carpinteriatipoAdoblehoja + tipopersiana + particiones_interiores
                               + heating_system + #cooling_system +
11                        rejilladeventilacionencocina + conductos +
                              campanaextractoraencocina + carpinteriasenvolvente +
12                        DJ_EZ + alturalibrepredeterminadaconfalsotecho +
                              superficieenvolvente + Factordeforma + volcuartoshumedos +
13                        carpperAenv + FLF
14                       , family = Gamma(link = "log"), data = train5)
15 #Aplicamos el algoritmo de seleccion automatica y nos quedamos con las variables con p-
       valores menores a 0.15
16 stepCriterion(modelocompleto5,direction = "forward", criterion ="p-value",levels = c
       (0.15,0.15))
17
18 #Tras obtener una lista reducida de variables, incluimos manualmente las interacciones
19 mod5forwinterinic<-glm(n50_test ~
20                        carpinteriaAclase +
21                        ZclimaticaHE1 +
22                        conductos +
23                        banoreformado +
24                        campanaextractoraencocina +
25                        persianaseguncatalogo +
26                        Falsotecho +
27                        alturalibrepredeterminadaconfalsotecho +#
28                        carpinteriatipoAmaterial +
29                        carpinteriasenvolvente +
30                        anoconstruccion +
31                        Factordeforma +
32                        particiones_interiores +
33                        estadooriginaldelafachada +
34                        volcuartoshumedos +
35
```

```
                          carpinteriaAclase *alturalibrepredeterminadaconfalsotecho+
                          ZclimaticaHE1 *alturalibrepredeterminadaconfalsotecho+
                          conductos *alturalibrepredeterminadaconfalsotecho+
                          banoreformado *alturalibrepredeterminadaconfalsotecho+
                          campanaextractoraencocina *alturalibrepredeterminadaconfalsotecho
                              +
                          persianaseguncatalogo *alturalibrepredeterminadaconfalsotecho+
                          Falsotecho *alturalibrepredeterminadaconfalsotecho+
                          carpinteriatipoAmaterial *alturalibrepredeterminadaconfalsotecho+
                          particiones_interiores *alturalibrepredeterminadaconfalsotecho+
                          estadooriginaldelafachada *alturalibrepredeterminadaconfalsotecho
                              +

                          carpinteriaAclase *carpinteriasenvolvente+
                          ZclimaticaHE1 *carpinteriasenvolvente+
                          conductos *carpinteriasenvolvente+
                          banoreformado *carpinteriasenvolvente+
                          campanaextractoraencocina *carpinteriasenvolvente+
                          persianaseguncatalogo *carpinteriasenvolvente+
                          Falsotecho *carpinteriasenvolvente+
                          carpinteriatipoAmaterial *carpinteriasenvolvente+
                          particiones_interiores *carpinteriasenvolvente+
                          estadooriginaldelafachada *carpinteriasenvolvente+

                          carpinteriaAclase *anoconstruccion+
                          ZclimaticaHE1 *anoconstruccion+
                          conductos *anoconstruccion+
                          banoreformado *anoconstruccion+
                          campanaextractoraencocina *anoconstruccion+
                          persianaseguncatalogo *anoconstruccion+
                          Falsotecho *anoconstruccion+
                          carpinteriatipoAmaterial *anoconstruccion+
                          particiones_interiores *anoconstruccion+
                          estadooriginaldelafachada *anoconstruccion+

                          carpinteriaAclase *Factordeforma+
                          ZclimaticaHE1 *Factordeforma+
                          conductos *Factordeforma+
                          banoreformado *Factordeforma+
                          campanaextractoraencocina *Factordeforma+
                          persianaseguncatalogo *Factordeforma+
                          Falsotecho *Factordeforma+
                          carpinteriatipoAmaterial *Factordeforma+
                          particiones_interiores *Factordeforma+
                          estadooriginaldelafachada *Factordeforma+

                          carpinteriaAclase *volcuartoshumedos+
                          ZclimaticaHE1 *volcuartoshumedos+
                          conductos *volcuartoshumedos+
                          banoreformado *volcuartoshumedos+
                          campanaextractoraencocina *volcuartoshumedos+
                          persianaseguncatalogo *volcuartoshumedos+
                          Falsotecho *volcuartoshumedos+
                          carpinteriatipoAmaterial *volcuartoshumedos+
                          particiones_interiores *volcuartoshumedos+
                          estadooriginaldelafachada *volcuartoshumedos

                        , family = Gamma(link = "log"), data = train5)

#Aplicamos de nuevo seleccion automatica, esta vez con criterio de p-valores menores a 0.1
stepCriterion(mod5forwinterinic,direction = "forward", criterion ="p-value",levels = c
    (0.1,0.1))
```

Código del modelo final

```
95
96 #Verificamos el modelo obtenido con el subconjunto de entrenamiento
97 mod5forwintermid<-glm(n50_test ~carpinteriaAclase + ZclimaticaHE1 + conductos +
       carpinteriasenvolvente + Factordeforma +
98                         campanaextractoraencocina + banoreformado + anoconstruccion +
                               Falsotecho + estadooriginaldelafachada +
99                         alturalibrepredeterminadaconfalsotecho + conductos:Factordeforma +
                                ZclimaticaHE1:anoconstruccion +
100                        carpinteriaAclase:Factordeforma  ,
101                      family = Gamma(link = "log"),data = train5)
102 Anova(mod5forwintermid, type = "III")#son todas significativas
103 pR2(mod5forwintermid)#0.4508514
104
105 #Probamos el modelo obtenido con el conjunto de test
106 mod5forwinterfin<-glm(n50_test ~carpinteriaAclase + ZclimaticaHE1 + conductos +
       carpinteriasenvolvente + Factordeforma +
107                        campanaextractoraencocina + banoreformado + anoconstruccion +
                               Falsotecho + estadooriginaldelafachada +
108                        alturalibrepredeterminadaconfalsotecho + conductos:Factordeforma +
                                ZclimaticaHE1:anoconstruccion +
109                        carpinteriaAclase:Factordeforma ,
110                      family = Gamma(link = "log"),data = test5)
111 #Evaluamos las metricas
112 summary(mod1forwinterfin)
113 Anova(mod5forwinterfin, type = "III")
114 AIC(mod5forwinterfin)#379.3343
115 BIC(mod5forwinterfin)#455.1566
116 pR2(mod5forwinterfin)#0.7580953
117
118 #Visualizamos los graficos de residuales
119 par(mfrow=c(2,2))
120 plot(mod5forwinterfin)
121
122 ################################################
123 ################################################
124 #Verificamos el modelo final elegido con la metodologia de seleccion forward, con las
       variables del quinto modelo y el total de los datos
125 modforwinterfin<-glm(n50_test ~carpinteriaAclase + ZclimaticaHE1 + conductos +
       carpinteriasenvolvente + Factordeforma +
126                        campanaextractoraencocina + banoreformado + anoconstruccion +
                               Falsotecho + estadooriginaldelafachada +
127                        alturalibrepredeterminadaconfalsotecho + conductos:Factordeforma +
                                ZclimaticaHE1:anoconstruccion +
128                        carpinteriaAclase:Factordeforma ,
129                      family = Gamma(link = "log"),data = datosp2)
130 #Evaluamos las metricas
131 summary(modforwinterfin)
132 Anova(modforwinterfin, type = "III")
133 AIC(modforwinterfin)#1865.439
134 BIC(modforwinterfin)#1992.274
135 pR2(modforwinterfin)#0.4720110
136 #Visualizamos los graficos de residuales
137 par(mfrow=c(2,2))
138 plot(modforwinterfin)
```

Bibliografía

[1] E. Commission, “Communication from the commission to the european parliament, the council, the european economic and social committee and the committee of the regions stepping up europe’s 2030 climate ambition investing in a climate-neutral future for the benefit of our,” 2020. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0562

[2] ——, “Renovation wave: the european green deal factsheet, 2020,” 2020. [Online]. Available: https://doi.org/10.2833/535670

[3] R. M.H.Sherman, “Building airthighness: Research and practice,” 2004, http://refhub.elsevier.com/S0360-1323(22)00666-7/sref14.

[4] J. H. C.N. Bramiana, A.G. Entrop, “Relationships between building characteristics and airtightness of dutch dwellings,” *Energy Proc*, 2016, https://doi.org/10.1016/j.egypro.2016.09.103.

[5] R. M.H.Sherman, “Analysis of air leakage measurements of us houses,” *Energy Build. 66*, 2013, https://doi.org/10.1016/j.enbuild.2013.07.047.

[6] R. R. B. Khemet, “A univariate and multiple linear regression analysis on a national fan (de)pressurization testing database to predict airtightness in houses, build,” *Environ 146*, 2018, https://doi.org/10.1016/j.buildenv.2018.09.030.

[7] H. Krstic, Z. Koskiand, I. Otkovic, and M.Spani, “Application of neural networks in predicting airtightness of residential units,” *Energy Build 84*, 2014, https://doi.org/10.1016/j.enbuild.2014.08.007.

[8] J. Fernández-Agüera, S. Domínguez-Amarillo, J. Sendra, and R. Suárez, “An approach to modelling envelope airtightness in multi-family social housing in mediterranean europe based on the situation in spain,” *Energy Build 128*, 2016, https://doi.org/10.1016/j.enbuild.2016.06.074.

[9] ——, “Predictive models for airtightness in social housing in a mediterranean region,” *Sustain. Cities Soc. 51*, 2019, https://doi.org/10.1016/j.enbuild.2016.06.074.

[10] J. A. M. Ibanez-Puy, “Airtightness in spanish residential buildings. case study,” 2019, https://doi.org/10.1109/ICE.2019.8792809.

[11] M. Montoya, E. Pastor, F. Carrié, G. Guyot, and E. Planas, “Air leakage in catalan dwellings: developing an airtightness model and leakage airflow predictions,” *Build and Environ*, 2010, https://doi.org/10.1016/j.buildenv.2009.12.009.

[12] Ministerio de Fomento, *Código Técnico de la Edificación (CTE)*, Gobierno de España, 2019. [Online]. Available: https://www.codigotecnico.org/

[13] ——, *Herramienta Unificada LIDER/CALENER (HULC)*, Gobierno de España, 2019. [Online]. Available: https://energia.gob.es/desarrollo/EficienciaEnergetica/CertificacionEnergetica/DocumentosReconocidos/Paginas/HULC.aspx

[14] I. Poza-Casado, P. Rodríguez-del Tío, M. Fernández-Temprano, M. Padilla-Marcos, and A. Meiss, “An envelope airtightness predictive model for residential buildings in spain,” *Building and Environment*, 2022, https://www.sciencedirect.com/journal/building-and-environment.

[15] ——, “Airtightness predictive model from measured data of residential buildings in spain,” 2023, https://www.aivc.org/resource/airtightness-predictive-model-measured-data-residential-buildings-spain.

[16] I. Poza-Casado, A. Meiss, M. A. Padilla-Marcos, and J. Feijó Muñoz, “Preliminary analysis results of spanish residential air leakage database„” 2018.

[17] P.McCulagh and J. Nelder, *Generalized Linear Models*. Chapman and Hall, 1989.

[18] D. Pregibon, *Goodness of Link Tests for Generalized Linear Models*. Journal of the Royal Statistical Society, 1979, https://www.jstor.org/stable/2346405.

[19] What are pseudo r-squareds? [Online]. Available: https://stats.oarc.ucla.edu/other/mult-pkg/faq/general/faq-what-are-pseudo-r-squareds/

[20] J. A. Rodrigo. (2016) Validación de modelos predictivos: Cross-validation, oneleaveout, bootstraping. [Online]. Available: https://rpubs.com/Joaquin_AR/238251

[21] V. O. Rosmary Isea. Coeficiente v de cramer. [Online]. Available: https://mariafatimadossantosestadistica1.files.wordpress.com/2018/06/coeficientes-v-de-cramer-y-c-de-pearson.pdf

[22] J. Feijó-Muñoz, A. Meiss, I. Poza-Casado, M. Ángel Padilla-Marcos, M. Rabanillo-Herrero, A. R. del Val, R. Gonzalez-Lezcano, C. Pardal, V. E. Iribarren, R. A. de Larriva, J. Fernandez-Aguera, V. J. del Campo Díaz, M. J. Dios-Viéitez, and M. M. Calderín, *Permeabilidad al aire de los edificios residenciales en España. Estudio y caracterización de sus infiltraciones*, 2019.

[23] I. P. Casado, “Airtightness performance of the building envelope of dwellings in spain. characterisation and energy impact of air infiltration,” *Universidad de Valladolid*, 2021.

[24] M. de Fomento and G. de España, “Código técnico de la edificación (cte). documento básico he 1: Limitación de la demanda energética,” 2019. [Online]. Available: https://www.codigotecnico.org/images/stories/pdf/ahorroEnergia/DBHE.pdf

[25] A. E. de Meteorología (AEMET), “Atlas climático ibérico (iberian climate atlas), ministerio de medio ambiente y medio rural y marino de españa„” 2011. [Online]. Available: http://www.aemet.es/documentos/es/conocermas/publicaciones/Atlas-climatologico/Atlas.pdf

[26] AENOR, “Une-en 12207 ventanas y puertas, permeabilidad al aire.” 2017.

[27] M. L. Delignette-Muller and C. Dutang, “fitdistrplus: An r package for fitting distributions,” 2015, r package version 1.0-14. [Online]. Available: https://cran.r-project.org/web/packages/fitdistrplus/index.html

[28] M. Kuhn, *caret: Classification and Regression Training*, 2023, r package version 6.0-93. [Online]. Available: https://CRAN.R-project.org/package=caret

[29] J. Fox and S. Weisberg, *car: Companion to Applied Regression*, 2023, r package version 3.1-2. [Online]. Available: https://CRAN.R-project.org/package=car

[30] S. Jackman, *pscl: Political Science Computational Laboratory*, 2022, r package version 1.5.5. [Online]. Available: https://CRAN.R-project.org/package=pscl