

Universidad Nacional de Ingeniería

Escuela Profesional de Estadística

**Predicción de las emisiones de CO₂ de las
edificaciones de Aragón, España en el
periodo 2015 - 2024 mediante un modelo de
regresión lineal múltiple**



Autores:

Sofía Dextre Simangas
Sebastian Matias Romero Davila

Curso: Análisis de Regresión

Docente: Amélida Pinedo Sanchez

Lima, Perú
Junio 2025

Índice

Resumen	3
1. Introducción	3
1.1. Problemática	3
1.2. Objetivo general	3
1.3. Objetivos específicos	3
1.4. Pregunta de investigación	3
2. Antecedentes	3
2.1. Predicción del consumo energético en edificaciones	3
2.2. Modelamiento de la contaminación atmosférica por partículas: Comparación de cuatro procedimientos predictivos en Santiago, Chile	4
2.3. Modelo de regresión lineal simple para pronosticar las emisiones de CO ₂ en el volcán Mauna Loa (Hawái)	4
2.4. Análisis estadístico de las emisiones de CO ₂ por el consumo de combustible gaseoso en ecuador mediante la aplicación de la regresión lineal simple	5
3. Metodología	5
3.1. Base de datos	5
3.1.1. Descripción general	6
3.1.2. Variables explicativas y respuesta	6
3.2. Tratamiento de variables	7
3.2.1. Clasificación ordinal	7
3.2.2. Dummies geográficas y estructurales	8
3.2.3. Creación de nuevas variables (feature engineering)	8
4. Modelo de regresión lineal múltiple	9
4.1. Especificación del modelo	9
4.2. Prueba de significancia global	10
4.3. Diagnóstico del modelo (Durbin-Watson)	11
4.4. Multicolinealidad	11
4.5. Selección de variables	17
5. Análisis de residuos	19
5.1. Tipos de residuos	19
6. Análisis de influencia	25
6.1. i-ésimo elemento de la diagonal de la matriz H o Leverage	25
6.2. Estadística DFFITS	26
6.3. COVRATIO	27
6.4. DFBETAS	28
6.5. Análisis	28
6.6. Modelo con casos de influencia	31
6.7. Modelo sin casos de influencia	31

7. Análisis de variables categóricas	32
7.1. Tipo de edificio	32
7.2. Antigüedad de los edificios	34
8. Validación del modelo	37
9. Resultados	39
9.1. Resumen	39
9.2. Análisis gráfico complementario de los resultados	39
9.3. Interpretación de los resultados	41
10. Conclusiones	43
10.1. Conclusiones por objetivos	43
10.2. Propuestas de mejora o trabajos futuros	43
11. Anexos	45

Resumen

1. Introducción

1.1. Problemática

Las edificaciones residenciales representan una fuente significativa de emisiones de CO_2 , principalmente debido a su consumo energético. En un contexto donde la sostenibilidad energética es una prioridad global, comprender cómo las características técnicas y estructurales de los edificios influyen en sus niveles de emisión resulta crucial. Aragón, una comunidad autónoma ubicada en el noreste de España, ha registrado miles de certificados energéticos entre 2015 y 2024. El análisis de estos registros ofrece una oportunidad única para evaluar el impacto ambiental del parque habitacional regional y detectar patrones relevantes que contribuyan a mejorar las políticas públicas y las prácticas de construcción sostenible.

1.2. Objetivo general

Desarrollar un modelo de regresión lineal múltiple que permita estimar las emisiones de CO_2 en edificaciones residenciales de Aragón, España, utilizando variables relacionadas con el consumo energético, temporales y categóricas provenientes de los certificados de eficiencia energética emitidos entre los años 2015 y 2024.

1.3. Objetivos específicos

- Identificar las variables más significativas de los certificados energéticos para la predicción de emisiones de CO_2 .
- Determinar qué tipos de edificios presentan una mayor contribución a las emisiones de CO_2 .
- Determinar si las viviendas construidas antes del año 1950 tienen una mayor influencia en las emisiones de CO_2 .

1.4. Pregunta de investigación

¿Qué variables explican con mayor fuerza las emisiones de CO_2 en edificaciones residenciales de Aragón, España, durante el periodo 2015–2024, y qué tipo de edificaciones tienen un mayor impacto en los niveles de emisión dentro de esta región?

2. Antecedentes

2.1. Predicción del consumo energético en edificaciones

En la última década, el análisis cuantitativo de emisiones contaminantes se ha consolidado como una herramienta esencial para comprender los impactos ambientales asociados a la actividad humana. Los modelos de regresión, en particular, han sido ampliamente utilizados por su capacidad de capturar relaciones entre variables explicativas (como

consumo de energía, condiciones meteorológicas o características físicas de edificaciones) y la variable respuesta: las emisiones de dióxido de carbono (CO_2). Estos enfoques permiten no solo describir patrones históricos, sino también proyectar tendencias futuras y evaluar escenarios hipotéticos, lo que resulta fundamental para la formulación de políticas públicas y estrategias de mitigación.

2.2. Modelamiento de la contaminación atmosférica por partículas: Comparación de cuatro procedimientos predictivos en Santiago, Chile

Este estudio tiene como objetivo evaluar y comparar la eficacia de cuatro modelos estadísticos para predecir los niveles de contaminación atmosférica por material particulado (MP10) en Santiago de Chile, una de las ciudades más contaminadas de América Latina. Los modelos considerados fueron: regresión lineal múltiple (RLM), regresión por componentes principales (PCR), análisis de correlación canónica (CCA) y análisis de regresión parcial (PLS). El enfoque del artículo es metodológico, orientado a determinar cuál de estos procedimientos ofrece mayor precisión predictiva y robustez frente a colinealidad entre variables independientes.

El conjunto de datos utilizado incluye observaciones meteorológicas diarias como temperatura máxima, temperatura mínima, humedad relativa, presión atmosférica y dirección del viento. Estas variables fueron seleccionadas por su relación empíricamente comprobada con la concentración de partículas contaminantes en el aire. El artículo realiza un análisis comparativo de los modelos utilizando métricas como el error cuadrático medio (MSE), el coeficiente de determinación R^2 y validación cruzada.

Los resultados revelan que la regresión por componentes principales (PCR) ofrece el mejor desempeño global, especialmente en presencia de multicolinealidad severa. A diferencia de la regresión múltiple convencional, que se ve afectada negativamente por la correlación entre predictores, el modelo PCR logra reducir la dimensionalidad del conjunto de variables independientes sin perder capacidad predictiva. Por otro lado, el modelo PLS también mostró buena capacidad de predicción, aunque con una interpretación más compleja.

Enlace al estudio: Ver documento original en línea

2.3. Modelo de regresión lineal simple para pronosticar las emisiones de CO_2 en el volcán Mauna Loa (Hawái)

En este estudio, López Miranda y Romero Ramos (2014) analizan el comportamiento del dióxido de carbono atmosférico medido en el observatorio del volcán Mauna Loa, Hawái, uno de los centros de monitoreo climático más importantes del mundo. Utilizando una serie temporal mensual que abarca desde enero de 1965 hasta diciembre de 1980, los autores aplican un modelo de regresión lineal simple para pronosticar la concentración de CO_2 como función del tiempo.

El modelo considera como variable independiente el tiempo (en meses), y como variable dependiente la concentración de CO_2 en partes por millón (ppm). La ecuación estimada

presenta un coeficiente de determinación $R^2 = 0,8854$, lo que indica que cerca del 89 % de la variación en las emisiones puede explicarse por el paso del tiempo. El modelo también fue sometido a pruebas estadísticas que confirman la validez de los supuestos clásicos de regresión: normalidad de residuos, homocedasticidad, independencia y linealidad. Esto sugiere que el modelo no solo es útil para describir una tendencia creciente sostenida, sino también para realizar pronósticos con cierta precisión a corto plazo.

El valor de este trabajo radica en su simplicidad metodológica y en su claridad didáctica. Aunque se enfoca en un único predictor, demuestra cómo una variable temporal puede ser suficiente para capturar dinámicas ambientales complejas. Además, su aplicación a un conjunto de datos oficial y continuo como el de Mauna Loa refuerza su confiabilidad. Para investigaciones como la presente, que exploran emisiones de CO₂ desde una perspectiva estructural y regional, este antecedente ofrece un punto de partida sólido en cuanto a construcción de modelos, validación estadística y presentación de resultados.

Enlace al estudio: Ver documento original en línea

2.4. Análisis estadístico de las emisiones de CO₂ por el consumo de combustible gaseoso en Ecuador mediante la aplicación de la regresión lineal simple

Luis Alberto De Lucas (2018) desarrolló un estudio que relaciona directamente el crecimiento de la población ecuatoriana con las emisiones de CO₂ generadas por el consumo de combustibles gaseosos. El trabajo, basado en datos del Ministerio del Ambiente entre los años 1979 y 2013, utilizó un modelo de regresión lineal simple para cuantificar esta relación. El modelo resultante obtuvo un coeficiente de determinación ajustado de $R^2 = 0,8161$, lo cual indica una capacidad explicativa notablemente alta para un solo predictor.

El autor realizó un análisis estadístico exhaustivo, que incluyó pruebas de normalidad de residuos, homocedasticidad, y significancia de parámetros, todas las cuales confirmaron la robustez del modelo. El estudio concluyó que el crecimiento poblacional tiene un efecto directo y creciente sobre las emisiones, y recomendó utilizar estos hallazgos como base para políticas energéticas sostenibles. Este antecedente es particularmente relevante para el presente estudio, ya que demuestra la utilidad de modelos simples con variables estructurales agregadas en contextos nacionales.

Enlace al estudio: Ver documento original en línea

3. Metodología

3.1. Base de datos

Revisar el Anexo 1, archivo Excel denominado Base de datos, específicamente la hoja BD_INICIAL.

3.1.1. Descripción general

El presente estudio se basa en una base de datos oficial obtenida del portal del Gobierno de Aragón (<https://www.aragon.es/>), la cual contiene los certificados de eficiencia energética emitidos para edificaciones residenciales en esta comunidad autónoma. La base original consta de 179 853 registros, correspondientes al período 2015–2024.

Cada registro representa un inmueble evaluado y contiene información sobre sus características físicas, tipología, consumo energético y nivel de emisiones de dióxido de carbono (CO₂) por metro cuadrado al año. La información se encontraba originalmente en formato tabular (CSV) y fue sometida a un proceso exhaustivo de limpieza, tratamiento y validación. Entre los pasos realizados destacan:

- Eliminación de registros duplicados y con valores faltantes en variables clave.
- Codificación de variables categóricas mediante técnicas de *one-hot encoding*.
- Conversión y estandarización de unidades.
- Generación de variables derivadas (como días hasta expiración del certificado).
- Validación de supuestos del modelo de regresión: linealidad, homocedasticidad, normalidad de residuos e independencia.

3.1.2. Variables explicativas y respuesta

La variable respuesta del modelo es:

- **Emission_CO2** (kgCO₂/m²/año): emisiones de dióxido de carbono asociadas al consumo energético anual del inmueble.

Las variables explicativas incluyen tanto cuantitativas como cualitativas:

- **Consumo_kWh/m2/año** — Energía consumida por metro cuadrado al año.
- **Clasificación_Emisiones** — Etiqueta (A–G) asignada por nivel de emisiones.
- **Clasificación_Consumo** — Etiqueta (A–G) por nivel de consumo energético.
- **Tipo_edificio** — Tipo de inmueble (bloque, local, unifamiliar, etc.).
- **Estado_edificio** — Estado de conservación o construcción.
- **Año_construccion** — Año de construcción del edificio.
- **Superficie_m2** — Área útil del inmueble.
- **Municipio y Provincia** — Localización geográfica.
- **Coordenadas_gps** — Latitud y longitud del edificio.
- **Días_hasta_expiracion** — Tiempo de vigencia del certificado.

El diccionario de las variables lo puede encontrar en el Anexo 2 denominado "bd_variables".

3.2. Tratamiento de variables

En esta sección se detalla el tratamiento aplicado a las variables del dataset para adecuarlas al análisis y modelamiento posterior. Este tratamiento incluye la conversión de variables categóricas a escala ordinal, la generación de variables dummy y la creación de nuevas variables derivadas.

3.2.1. Clasificación ordinal

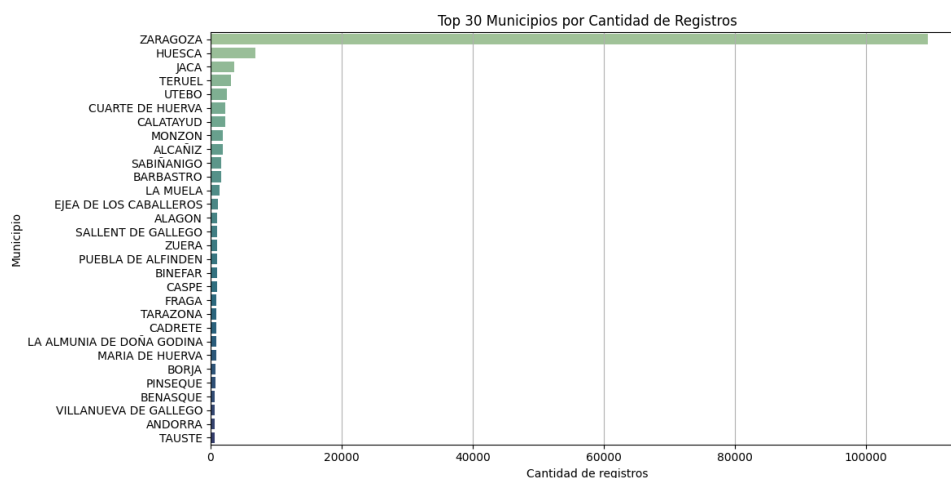
Las variables `Clasificacion_Emisiones` y `Clasificacion_consumo` presentan una escala categórica con un orden inherente, de A (más eficiente) a G (menos eficiente). Para preservar esta estructura ordinal en los modelos de regresión, se asignaron valores numéricos de la siguiente forma:

Categoría	Valor ordinal
A	1
B	2
C	3
D	4
E	5
F	6
G	7

El resultado fue la creación de las variables `Clasificacion_Emisiones_ordinal` y `Clasificacion_consumo_ordinal`, que conservan el orden original pero son tratadas numéricamente en el análisis posterior.

En el caso particular de la variable `Municipio`, se identificó una alta concentración de registros en un único valor: **Zaragoza**. Por lo que se decidió hacer una dummie que muestre si es de dicho municipio o no.

Esta reducción permite capturar la variabilidad regional sin sobreajustar el modelo con una cantidad excesiva de variables dummy.



3.2.2. Dummies geográficas y estructurales

Para variables categóricas sin orden implícito, como las relacionadas con la localización y el tipo de edificación, se utilizaron variables dummy. Este proceso consiste en transformar cada categoría en una columna binaria (0 o 1). Se aplicó a las siguientes variables:

■ **Provincia**

Provincia	ZARAGOZA	TERUEL
ZARAGOZA	1	0
TERUEL	0	1
HUESCA	0	0

■ **Tipo_edificio**

Tipo_edificio	Unifamiliar	Local	Bloque completo	Edificio completo
Unifamiliar	1	0	0	0
Local	0	1	0	0
Bloque completo	0	0	1	0
Edificio completo	0	0	0	1
Vivienda individual	0	0	0	0

■ **Estado_edificio**

Estado_edificio	Obra terminada	Proyecto nueva construccion	Proyecto reforma
Obra terminada	1	0	0
Proyecto nueva construccion	0	1	0
Proyecto reforma	0	0	1
Existente	0	0	0

■ **Municipio:** Solo se consideró el de mayor moda

Municipio	Municipio_ZARAGOZA
ZARAGOZA	1
Otros municipios	0

Este proceso permite incorporar información categórica de forma compatible con modelos estadísticos que requieren variables numéricas.

3.2.3. Creación de nuevas variables (feature engineering)

Con el objetivo de enriquecer el análisis, se generaron nuevas variables a partir de las existentes. Estas permiten capturar relaciones relevantes no explícitas en los datos originales.

- **Antigüedad del edificio:** diferencia entre el año de emisión del certificado y el año de construcción.

$$\text{Antigüedad} = \text{Anio_emision} - \text{Anio_construccion}$$

- **Consumo por m²:** relación entre el consumo energético y el tamaño del edificio.

$$\text{Consumo_por_m2} = \frac{\text{ConsumoKWh/m2/Anio}}{\text{Superficie_m2}}$$

- **Coordenadas separadas:** la variable `Coordenadas_gps`, que originalmente contiene un string con las coordenadas X e Y, fue dividida en dos columnas numéricas: `coord_x` y `coord_y`, lo cual permite realizar análisis espaciales más precisos.

Coordenadas_gps original	coord_x	coord_y
673219,56 , 4612612,38	673219.56	4612612.38
674903,68 , 4612931,37	674903.68	4612931.37

Tabla 1: Transformación de coordenadas GPS a variables numéricas

Estas variables derivadas pueden mejorar el desempeño predictivo de los modelos al capturar dimensiones relevantes del problema que no estaban explícitas inicialmente.

Culminado el proceso de limpieza y tratamiento de las variables regresoras, la base de datos limpia se encuentra en el Anexo 1, hoja BD_LIMPIO.

4. Modelo de regresión lineal múltiple

En esta sección presentamos el desarrollo del modelo de regresión lineal múltiple propuesto para analizar los factores que influyen en la emisión de dióxido de carbono. La variable dependiente y corresponde a las emisiones anuales de CO₂ por metro cuadrado al año, mientras que las variables explicativas x_1, x_2, \dots, x_k representan atributos físicos de inmuebles, variables de eficiencia energética y factores temporales/geográficos.

4.1. Especificación del modelo

Se plantea el siguiente modelo de regresión lineal múltiple:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_{24} x_{24} + \varepsilon$$

donde la variable respuesta es

y : Emisión de CO₂

Como variables explicativas se consideraron las siguientes:

x_1 : Consumo KWh/m²/Año

x_2 : Año de construcción

x_3 : Superficie en metros cuadrados

x_4 : Año de emisión

x_5 : Días hasta expiración

x_6 : Coordenadas x (latitud)

x_7 : Coordenadas y (longitud)

x_8 : Clasificación de emisiones

x_9 : Clasificación de consumo

$x_{10} - x_{13}$: Variables dummy correspondientes al tipo de edificio:

x_{10} : Edificio completo

x_{11} : Local

x_{12} : Unifamiliar

x_{13} : Bloque completo

$x_{14} - x_{16}$: Variables dummy del estado del edificio:

x_{14} : Proyecto en nueva construcción

x_{15} : Proyecto de reforma

x_{16} : Obra terminada

$x_{17} - x_{19}$: Variables dummy de las provincias de España:

x_{17} : Provincia TERUEL

x_{18} : Provincia ZARAGOZA

x_{19} : Municipio Zaragoza (1: Zaragoza 0:Otros municipios)

x_{20} : Mes de emisión (variable ordinal del mes)

$x_{21} - x_{23}$: Variable dummy de la estación del año:

x_{21} : Otoño

x_{22} : Primavera

x_{23} : Verano

x_{24} : Indicador de antigüedad (1 si el edificio es muy antiguo, 0 en caso contrario)

4.2. Prueba de significancia global

Para determinar si el modelo es estadísticamente significativo en su conjunto, se realiza la prueba F de significancia global, con las siguientes hipótesis:

$$H_0 : \beta_1 = \beta_2 = \dots = \beta_{24} = 0 \quad vs \quad H_1 : \exists \beta_i \neq 0 \quad \text{para algún } i \in \{1, 2, \dots, 24\}$$

Teniendo como estadístico de prueba a

$$F_c = \frac{SCE/p}{SCR/(n-p-1)} \sim F(p, n-p-1)$$

El ajuste por mínimos cuadrados es

$$\hat{y} = \hat{\beta}_0 + \hat{\beta}_1 x_1 + \hat{\beta}_2 x_2 + \dots + \hat{\beta}_{24} x_{24}$$

Tabla ANVA

FV	SC	gl	CM	F_c	p_value
Regresión	86240360.17	24	3593348.34	11539.859	0.000
Residual	55851715.62	179365	311.386		
Total	142092075.8	179389			

Decisión: Como $p_value < 0,05$, se rechaza H_0 .

Conclusión: Bajo un nivel de significancia de 0,05 se concluye que el modelo de regresión es estadísticamente significativo. Esto significa que, en conjunto, las variables independientes incluidas en el modelo permiten predecir de manera significativa las emisiones de CO₂.

4.3. Diagnóstico del modelo (Durbin-Watson)

Calcularemos el estadístico de Durbin-Watson para verificar si los errores del modelo presentan autocorrelación de primer orden, lo cual violaría uno de los supuestos del modelo clásico de regresión. Sean las hipótesis

- H_0 : No existe correlación en la serie ($\rho \approx 0$)
- H_1 : Existe correlación en la serie ($\rho \neq 0$)

Estadístico de prueba

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

Resumen del modelo

R^2	R_{adj}^2	Error estándar de la estimación	Durbin-Watson
0.607	0.607	17.647	1.734

Decisión: Como DW menor que 2, entonces **se rechaza** H_0 .

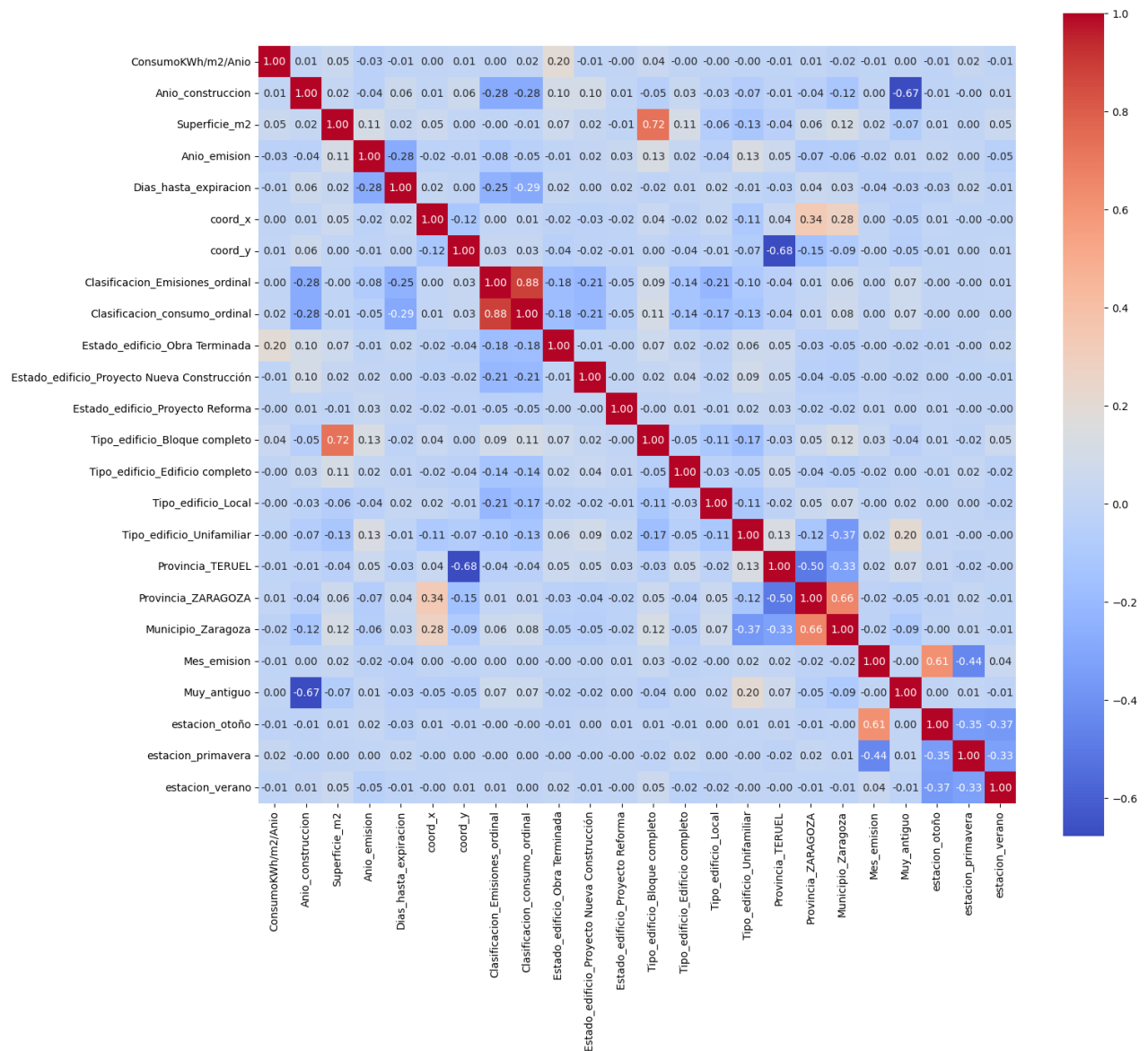
Conclusión: Bajo un nivel de significancia de $\alpha = 0.05$, se concluye que existe evidencia de autocorrelación de primer orden en los residuos del modelo de regresión, dado que el estadístico de Durbin-Watson obtenido es $DW = 1.734$, valor que se encuentra ligeramente por debajo de 2. Esto sugiere una **autocorrelación positiva leve**.

Dado este resultado, evaluaremos también la posible presencia de **multicolinealidad** entre las variables explicativas, ya que ambas problemáticas podrían coexistir y afectar la estabilidad del modelo.

4.4. Multicolinealidad

Para evaluar la presencia de multicolinealidad entre las variables regresoras, se analizarán la matriz de correlaciones, los factores de inflación de la varianza (VIF), la tolerancia y el análisis del sistema propio $X'X$.

Matriz de correlación



Observamos que algunas variables predictoras presentan correlaciones superiores a $\pm 0,6$. Por ello, realizaremos un análisis más detallado de estas variables para evitar posibles problemas de colinealidad en el proceso de selección.

Tabla 2: Correlación entre pares de variables predictoras

1ra variable predictor	2da variable predictor	Correlación
Clasificacion_Emisiones_ordinal	Clasificacion_consumo_ordinal	0.88
Tipo_edificio_bloque_completo	Superficie_m2	0.72
Provincia_TERUEL	coord_y	-0.68
Muy_antiguo	Anio_construccion	-0.67
Municipio_Zaragoza	Provincia_ZARAGOZA	0.66
estación_otoño	Mes_emision	0.61

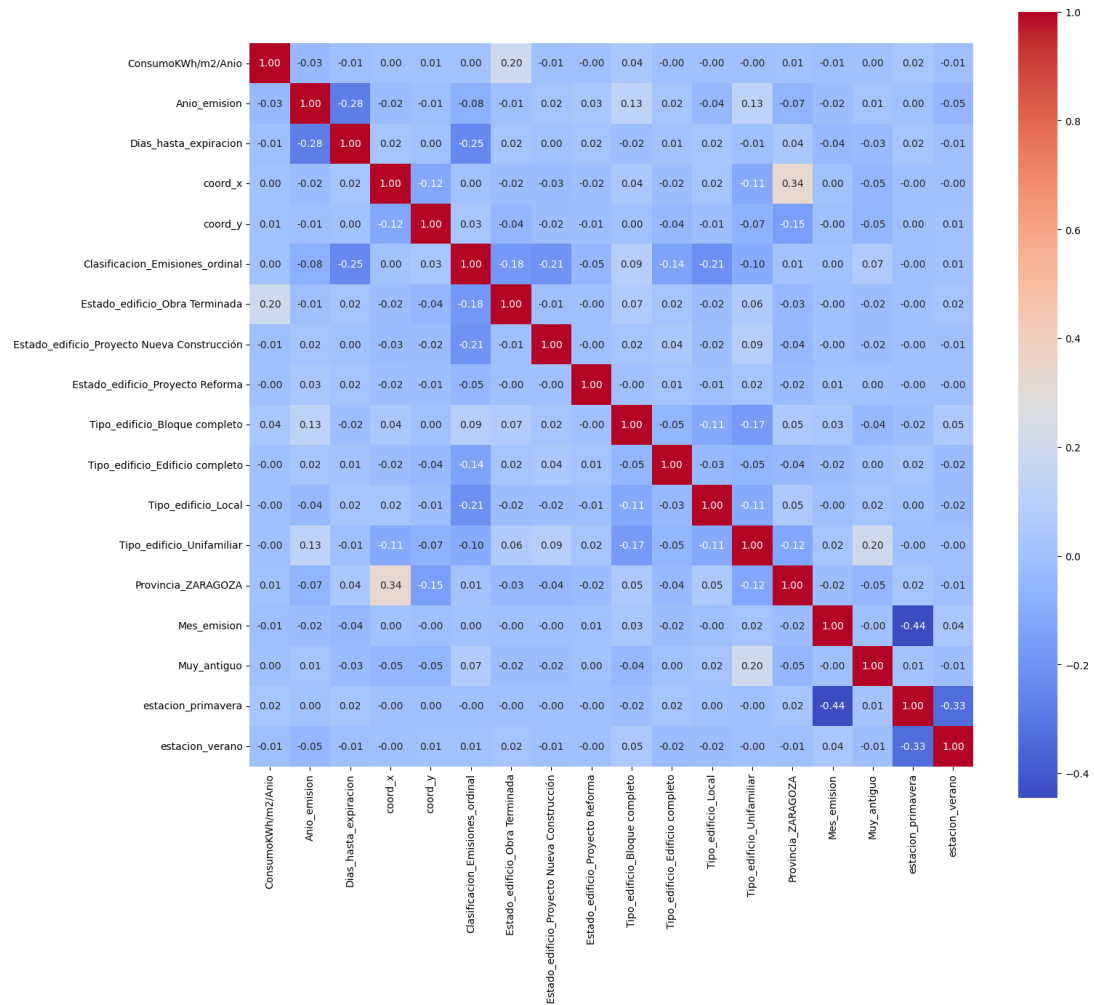
Tabla 3: Estadísticas de colinealidad

Variable	Tolerancia	VIF
ConsumoKWh/m2/Anio	0.953	1.049
Anio_construccion	0.445	2.246
Superficie_m2	0.440	2.271
Anio_emision	0.840	1.191
Dias_hasta_expiracion	0.814	1.228
coord_x	0.729	1.371
coord_y	0.196	5.097
Clasificacion_Emisiones_ordinal	0.210	4.753
Clasificacion_consumo_ordinal	0.209	4.783
Estado_edificio_Obra Terminada	0.901	1.109
Estado_edificio_Proyecto Nueva Construcción	0.934	1.071
Estado_edificio_Proyecto Reforma	0.995	1.005
Tipo_edificio_Bloque completo	0.426	2.345
Tipo_edificio_Edificio completo	0.911	1.097
Tipo_edificio_Local	0.894	1.118
Tipo_edificio_Unifamiliar	0.731	1.368
Provincia_TERUEL	0.142	7.030
Provincia_ZARAGOZA	0.192	5.198
Municipio_Zaragoza	0.444	2.251
Mes_emision	0.528	1.893
Muy_antiguo	0.490	2.042
estacion_otoño	0.409	2.445
estacion_primavera	0.611	1.637
estacion_verano	0.578	1.729

La primera variable candidata a ser retirada es Provincia_TERUEL, considerando su nivel de correlación, el valor del VIF y la tolerancia. A partir de ello, se procedió con el retiro secuencial de otras variables. En el siguiente cuadro se detallan los motivos correspondientes a cada caso.

Prueba	Variable retirada	Motivo
1	Provincia_TERUEL	VIF más alto entre todas las variables.
2	Clasificacion_consumo_ordinal	Presenta la mayor correlación entre todas las variables (0.88) con Clasificacion_Emisiones_ordinal. Esta alta asociación se explica porque ambas clasificaciones fueron aplicadas a la variable respuesta de forma similar, utilizando la misma cantidad de niveles. Se decidió retirar Clasificacion_consumo_ordinal por tener un VIF más elevado que la variable relacionada.
3	estacion_otoño	Mayor VIF entre las variables restantes. Variable dummy proveniente de la variable Mes_emision, con la cual presentaba una correlación de 0.61.
4	Superficie_m2	Segunda mayor correlación (0.72) con la variable Tipo_edificio_bloque_completo. Al mismo tiempo, presentaba mayor VIF entre todas las variables.
5	Municipio_Zaragoza	Además de presentar el mayor VIF entre las variables restantes, muestra una correlación de 0.66 con Provincia_ZARAGOZA, lo cual es esperable, ya que ambas referencia a la capital de Aragón.
6	Anio_construccion	Esta variable tiene el VIF más elevado entre las que aún permanecían en el modelo. A partir de ella se generó la variable dummy Muy_antiguo (con dos categorías), con la que mantiene una correlación negativa de -0.67.

Tras una nueva revisión de la matriz de correlaciones, se confirma que todas las asociaciones entre variables se encuentran por debajo del umbral de $\pm 0,6$.



Ahora los valores actuales de VIF y tolerancia se encuentran dentro de los límites comúnmente aceptados.

Tabla 4: Estadísticas de colinealidad

Variable	Tolerancia	VIF
Clasificación_Emisiones_ordinal	0.754	1.326
Tipo_edificio_Local	0.904	1.106
Tipo_edificio_Unifamiliar	0.846	1.182
Tipo_edificio_Edificio completo	0.961	1.040
coord_v	0.954	1.048
Tipo_edificio_Bloque completo	0.902	1.108
Estado_edificio_Proyecto Nueva Construcción	0.939	1.065
Estado_edificio_Obra Terminada	0.906	1.104
Días_hasta_expiración	0.836	1.196
Año_emisión	0.845	1.183
Muy_antiguo	0.943	1.060
Provincia_ZARAGOZA	0.854	1.171
Estado_edificio_Proyecto Reforma	0.995	1.005
estación_primavera	0.701	1.427
ConsumoKWh/m2/Año	0.956	1.046
coord_x	0.875	1.142
Mes_emisión	0.784	1.275
estación_verano	0.867	1.154

Tabla 5: Análisis de dimensiones

Dimensión	Autovalor	Índice de condición
1	8,518	1,000
2	1,206	2,657
3	1,143	2,730
4	1,053	2,844
5	1,042	2,858
6	1,000	2,918
7	0,967	2,968
8	0,951	2,993
9	0,806	3,251
10	0,785	3,293
11	0,554	3,920
12	0,519	4,050
13	0,204	6,469
14	0,122	8,345
15	0,049	13,229
16	0,041	14,358
17	0,031	16,709
18	0,009	31,256
19	$1,085 \times 10^{-6}$	2801,793

4.5. Selección de variables

Empezamos aplicando el método de selección computacional Stepwise con las 18 variables, bajo los criterios $F_{in} \leq 0,001$ y $F_{out} \geq 0,005$ se obtuvo 16 variables significativas.

Modelo	Variables entradas	Variables eliminadas
1	Clasificación_Emissiones_ordinal	
2	Tipo_edificio_Local	
3	Tipo_edificio_Unifamiliar	
4	Tipo_edificio_Edificio_completo	
5	coord_y	
6	Tipo_edificio_Bloque_completo	
7	Estado_edificio_Proyecto_Nueva_Construcción	
8	Estado_edificio_Obra_Terminada	
9	Dias_hasta_expiracion	
10	Anio_emision	
11	Muy_antiguo	
12	Provincia_ZARAGOZA	
13	Estado_edificio_Proyecto_Reforma	
14	estacion_primavera	
15	ConsumoKWh/m2/Anio	
16	coord_x	

Tabla 6: Resumen del modelo

Modelo	R	R^2	R^2_{adj}	S	Cp_Mallows
1	0,705	0,497	0,497	19,95	450.011
2	0,731	0,534	0,534	19,21	286.268
3	0,754	0,569	0,569	18,47	129.532
4	0,766	0,587	0,587	18,09	49.672
5	0,768	0,589	0,589	18,03	43.228
6	0,769	0,592	0,592	17,98	32.068
7	0,770	0,593	0,593	17,95	28.784
8	0,771	0,595	0,595	17,91	25.616
9	0,772	0,596	0,596	17,89	19.530
10	0,773	0,597	0,597	17,86	16.307
11	0,773	0,598	0,598	17,86	14.801
12	0,773	0,598	0,598	17,85	15.100
13	0,773	0,598	0,598	17,85	15.414
14	0,773	0,598	0,598	17,84	15.750
15	0,773	0,598	0,598	17,84	16.050
16	0,773	0,598	0,598	17,84	16.330

Los valores de **Cp_Mallows** se encuentran en el Anexo 3.

Al analizar la tabla de resumen del modelo, se observa que:

La estabilización de R^2 y R^2_{adj} a partir del modelo 11, alcanzan 0,598 y ya no mejoran significativamente en los modelos siguientes (12-16). Esto indica que las variables adicionales no aportan una mejora sustancial en la capacidad explicativa del modelo.

El error estándar disminuye de 17,855 (Modelo 11) a 17,841 (Modelo 16), una reducción de solo 0,014 en 5 modelos. Esta mejora es insignificante en términos prácticos y no justifica la complejidad añadida.

Entonces, un modelo con 11 variables es más simple, parsimonioso, fácil de interpretar y computacionalmente eficiente.

1. Planteamiento de hipótesis

- H_0 : No existe correlación en la serie ($\rho \approx 0$)
- H_1 : Existe correlación en la serie ($\rho \neq 0$)

2. Estadístico de prueba

$$DW = \frac{\sum_{t=2}^n (e_t - e_{t-1})^2}{\sum_{t=1}^n e_t^2}$$

Número de observaciones: $n = 179,390$

Número de variables regresoras: 18

Resumen del modelo

R^2	R_{adj}^2	Error estándar de la estimación	Durbin-Watson
0.607	0.607	17.647366083099600	1.734

Decisión

Dado que el valor de $DW = 1,734$, el cual está cercano a 2, se concluye que **no se rechaza la hipótesis nula H_0** . Esto indica que no hay evidencia estadística significativa de autocorrelación en los residuos del modelo.

A partir del resultado del estadístico de Durbin-Watson y considerando un nivel de significancia del 5 %, se puede afirmar lo siguiente:

- **Ausencia de autocorrelación de primer orden:** El valor de DW cercano a 2 sugiere que los errores del modelo no presentan autocorrelación positiva ni negativa. Esto es deseable en modelos de regresión, ya que garantiza que los errores sean independientes entre sí.
- **Robustez del modelo:** La evidencia de independencia de errores sugiere que el modelo puede ser confiablemente usado para predicción o análisis explicativo sin necesidad de ajustes adicionales.

Conclusión: Bajo un nivel de significancia del 5 %, se concluye que **no hay evidencia suficiente para afirmar la existencia de autocorrelación positiva de primer orden en los residuos**. Por tanto, se **mantiene la suposición de independencia de los errores**, reforzando la validez estadística del modelo de regresión utilizado.

5. Análisis de residuos

5.1. Tipos de residuos

En el análisis de regresión lineal, los residuos permiten evaluar si el modelo se ajusta adecuadamente a los datos y si se cumplen los supuestos básicos del modelo. En este estudio, donde se estiman las emisiones de CO₂ de edificios residenciales en Aragón, se analizaron los siguientes tipos de residuos:

Residuos ordinarios

Los residuos ordinarios son la diferencia entre el valor observado y_i y el valor ajustado \hat{y}_i por el modelo:

$$e_i = y_i - \hat{y}_i$$

Estos residuos representan el error cometido al predecir una observación con el modelo ajustado.

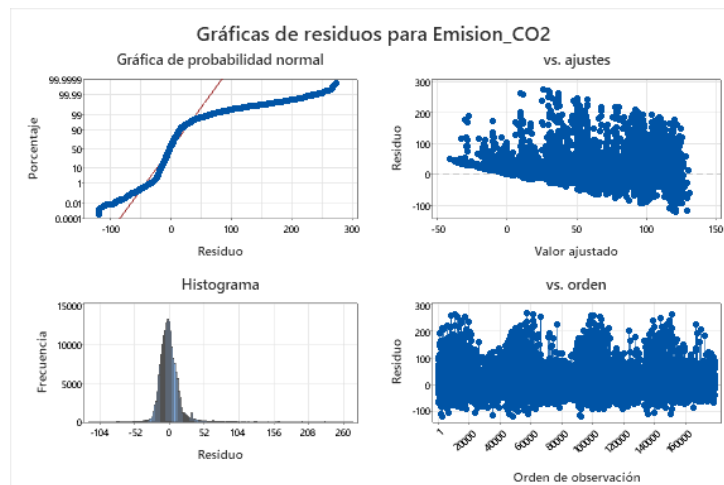


Figura 1: Gráfico de residuos

A continuación, se describen los principales hallazgos obtenidos a partir de las gráficas de diagnóstico del modelo:

- **Gráfica de probabilidad normal:**

Los residuos no siguen una línea recta, especialmente en los extremos. Esto indica una *desviación de la normalidad*, lo que podría afectar la validez de inferencias basadas en hipótesis de normalidad, tales como los intervalos de confianza o las pruebas t .

- **Residuos vs. valores ajustados:**

Se observa una dispersión no constante de los residuos, con un patrón de ensanchamiento en forma de abanico. Esto sugiere que **no se cumple el supuesto de homocedasticidad**. Además, la presencia de cierta curvatura indica que podrían faltar términos no lineales o interacciones en el modelo.

- **Histograma de residuos:**

Aunque los residuos están centrados alrededor de cero, el histograma presenta **asimetría a la derecha y colas largas**, lo cual refuerza la conclusión de que los errores no siguen perfectamente una distribución normal.

- **Residuos vs. orden de observación:**

No se observan patrones sistemáticos evidentes, aunque existen agrupaciones por bloques. Esto podría reflejar alguna **estructura temporal o espacial** en los datos.

Se vio que nuestros residuos no son normales por lo que aplicaremos Box Cox para mejorar la normalidad y estabilizar la varianza

Método

transformación de Box-Cox

λ redondeado	0
λ estimado	0.186742
IC de 95% para λ	(*, *)

El parámetro de transformación estimado por el método de máxima verosimilitud fue:

$$\hat{\lambda} = 0,186742$$

Dado que este valor es cercano a 0, se sugiere utilizar una transformación logarítmica de la variable dependiente:

$$Y^* = \log(Y)$$

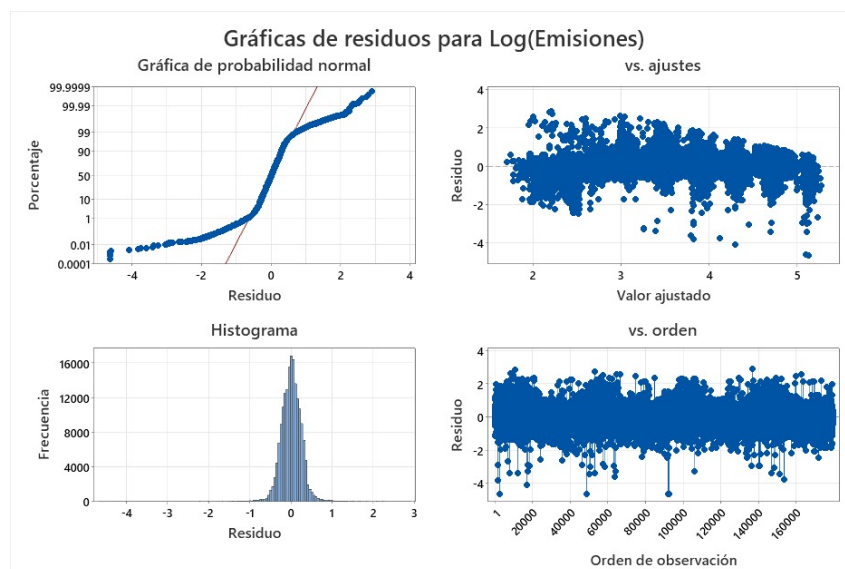
A partir de ahora nuestro target será el logaritmo de las emisiones

Ecuación de regresión

```
Log(Emissiones) = -4.971 + 0.003261 Año_emision + 0.000012 Dias_hasta_expiracion
+ 0.22485 coord_y + 0.406089 Clasificacion_Emisiones_ordinal
- 0.24015 Estado_edificio_Obra Terminada
- 0.37433 Estado_edificio_Proyecto Nueva
+ 0.04873 Tipo_edificio_Bloque completo
+ 0.55493 Tipo_edificio_Edificio completo + 0.47798 Tipo_edificio_Local
+ 0.23491 Tipo_edificio_Unifamiliar + 0.02675 Muy_antiguo
```

Coeficientes

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	-4.971	0.443	-11.21	0.000	
Anio_emision	0.003261	0.000219	14.91	0.000	1.17
Dias_hasta_expiracion	0.000012	0.000001	9.43	0.000	1.19
coord_y	0.22485	0.00494	45.53	0.000	1.01
Clasificacion_Emisiones_ordinal	0.406089	0.000687	590.77	0.000	1.32
Estado_edificio_Obra Terminada	-0.24015	0.00948	-25.34	0.000	1.06
Estado_edificio_Proyecto Nueva	-0.37433	0.00970	-38.59	0.000	1.06
Tipo_edificio_Bloque completo	0.04873	0.00191	25.48	0.000	1.10
Tipo_edificio_Edificio completo	0.55493	0.00553	100.34	0.000	1.04
Tipo_edificio_Local	0.47798	0.00274	174.24	0.000	1.10
Tipo_edificio_Unifamiliar	0.23491	0.00202	116.33	0.000	1.16
Muy_antiguo	0.02675	0.00230	11.66	0.000	1.06



- **Gráfica de probabilidad normal:** Aunque los residuos se alinean en gran parte con la línea de normalidad, hay curvatura en las colas, lo cual sugiere leves desviaciones de la normalidad. Sin embargo, la transformación logarítmica parece haber mejorado esta distribución comparado con el modelo original.
- **Residuos vs. valores ajustados:** No se observa una forma definida o patrón en los residuos, lo cual indica que la **suposición de linealidad y homocedasticidad (varianza constante)** es razonable. Hay una ligera asimetría, pero no es crítica.
- **Histograma de residuos:** Los residuos se distribuyen aproximadamente de forma **simétrica y centrada en cero**, lo que respalda la **normalidad** de los errores.
- **Residuos vs. orden de observación:** No hay un patrón claro, lo que sugiere que **los residuos son independientes** entre sí (no hay autocorrelación).

Conclusión: Las gráficas de residuos respaldan que el modelo de regresión con la transformación logarítmica aplicada a las emisiones cumple los supuestos clásicos del modelo lineal

Residuos estandarizados

Para facilitar la comparación entre residuos de distintas escalas o unidades, se utilizan los residuos estandarizados. Este tipo de residuos tiene media cero y varianza uno. Son útiles para detectar valores atípicos; un valor de $|d_i| > 2$ o 3 puede indicar una observación inusual.

Analizando las observaciones inusuales

$$d_i = \frac{e_i}{\hat{\sigma}}, \quad \text{donde } \hat{\sigma} = \sqrt{CMRes}$$

Gráfica para emisiones CO2

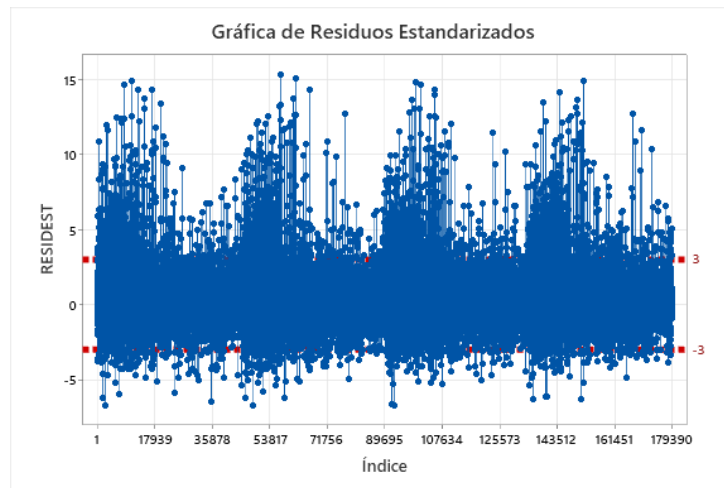


Figura 2: Gráfico de residuos estandarizados

Gráfica para $\log(\text{Emisiones CO2})$

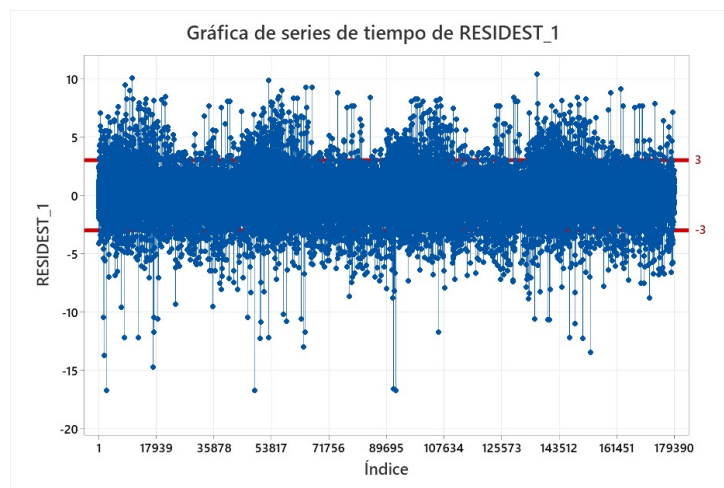


Figura 3: Gráfico de residuos estandarizados

Residuos estudentizados (internos)

Los residuos estudentizados ajustan la varianza de cada observación utilizando su apalancamiento h_{ii} , lo que permite detectar mejor los valores atípicos considerando su posición en el espacio de predictores:

$$r_i = \frac{e_i}{\hat{\sigma} \sqrt{1 - h_{ii}}}$$

Aquí, h_{ii} es el elemento diagonal de la matriz sombrero $H = X(X^T X)^{-1} X^T$, y mide la influencia de la observación i sobre su propia predicción. Los residuos estudentizados son más fiables que los estandarizados cuando se trata de identificar outliers.

Para Emision CO2

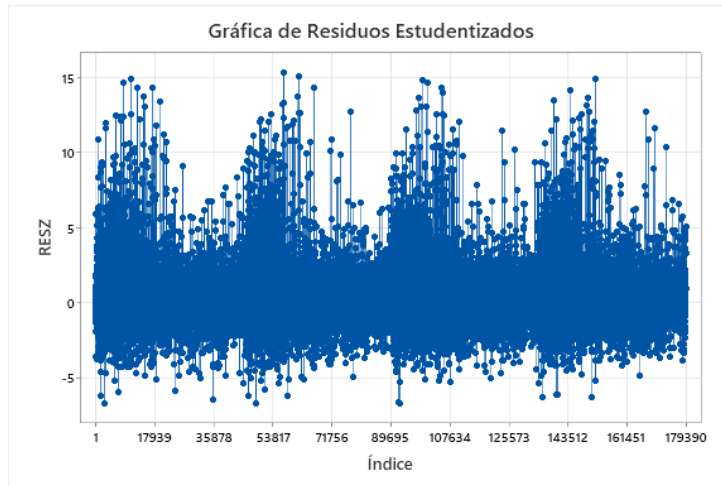


Figura 4: Gráfico de residuos estudentizados

Para log(emision CO2)

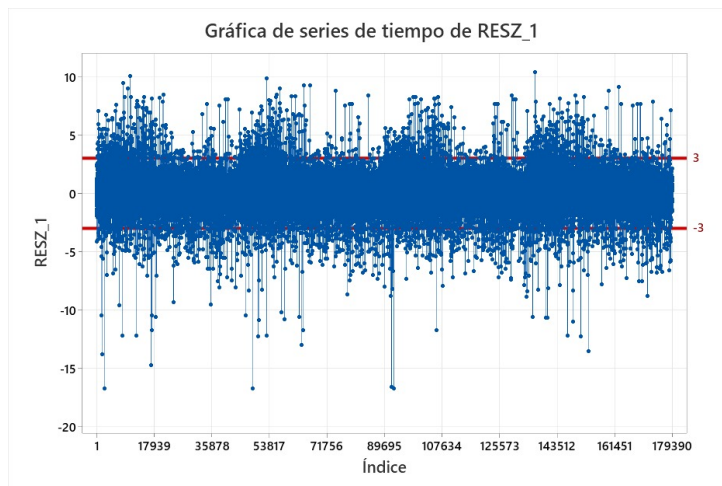


Figura 5: Gráfico de residuos estudentizados

A pesar de que box cox ayudó a la normalización, aún hay muchos residuos extremos, por lo que eliminaremos aquellos mayores a $|3|$

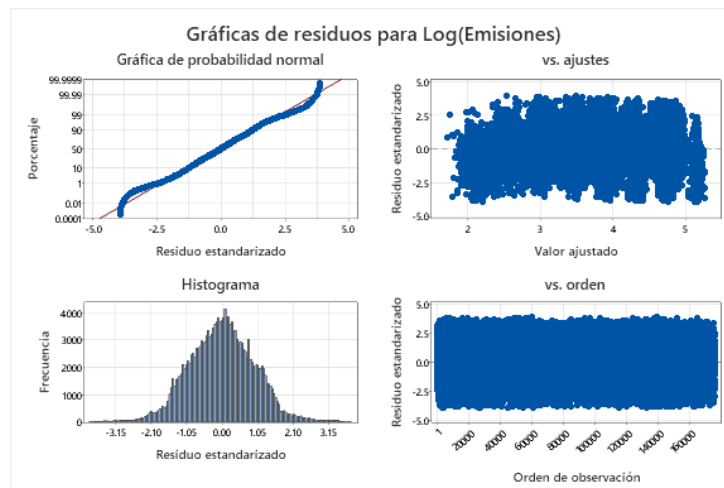


Figura 6: Gráficos de residuos

1. **Gráfica de probabilidad normal:** La mayor parte de los puntos sigue una línea recta, lo que indica que los residuos presentan una distribución aproximadamente normal.
2. **Residuos vs. valores ajustados:** La dispersión de los residuos en torno al cero no muestra patrones evidentes (como forma de U o tendencia) lo cual es bueno.
3. **Histograma de residuos:** El histograma muestra una distribución aproximadamente simétrica y con forma de campana, centrada en cero. Esto es consistente con la suposición de normalidad de los errores.
4. **Residuos vs. orden de observación:** No se observan patrones sistemáticos o tendencias a lo largo del tiempo, lo cual indica independencia entre los residuos.

En general, el modelo presenta un comportamiento adecuado en cuanto a los supuestos del análisis de regresión lineal, aunque se recomienda revisar los valores atípicos extremos (residuos mayores a ± 5) para evaluar su impacto y considerar su posible eliminación.

Durbin Watson

Finalmente, utilizaremos el estadístico Durbin-Watson, que es una prueba que se utiliza para detectar la autocorrelación (especialmente autocorrelación de primer orden) en los residuos de una regresión lineal.

Tabla 7: Interpretación del estadístico Durbin-Watson

Valor de DW	Interpretación
≈ 2	No hay autocorrelación
< 2	Hay autocorrelación positiva
> 2	Hay autocorrelación negativa
≈ 0	Alta autocorrelación positiva
≈ 4	Alta autocorrelación negativa

Calculando el estadístico:

Resumen del modelo^b

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación	Durbin-Watson
1	.888 ^a	.788	.788	.2271250695	1.725

a. Predictores: (Constante), Muy_antiguo, Tipo_edificio_Edificio completo, Estado_edificio_Obra Terminada, Anio_emision, Estado_edificio_Proyecto Nueva, Tipo_edificio_Local, coord_y, Tipo_edificio_Bloque completo, Dias_hasta_expiracion, Tipo_edificio_Unifamiliar, Clasificacion_Emisiones_ordinal

b. Variable dependiente: Log(Emissiones)

DW= 1.7, bastante cercano al dos, lo cual indica que no hay evidencia fuerte de autocorrelación en los residuos.

Residuos PRESS (Predicted Residuals)

Los residuos PRESS se calculan eliminando una observación del conjunto de datos, ajustando el modelo sin ella, y luego prediciendo su valor. Su fórmula es:

$$e_{(i)} = y_i - \hat{y}_{(i)} = \frac{e_i}{1 - h_{ii}}$$

Estos residuos permiten evaluar la capacidad predictiva del modelo para observaciones nuevas. Una diferencia grande entre e_i y $e_{(i)}$ sugiere que la observación i tiene una gran influencia y que el modelo podría estar sobreajustado. Además, se utiliza para calcular la estadística PRESS, útil para comparar modelos:

$$PRESS = \sum_{i=1}^n \left(\frac{e_i}{1 - h_{ii}} \right)^2$$

Un modelo adecuado tendrá un valor de PRESS bajo, indicando buena capacidad de predicción.

$$PRESS = 57199956$$

Para poder interpretar este valor podemos compararlo con la SST.

$$\frac{PRESS}{SST} = \frac{57199951}{142092076} = 0,4026$$

Un valor de PRESS relativamente bajo respecto a la SST indica que el modelo tiene buena capacidad predictiva.

6. Análisis de influencia

6.1. i -ésimo elemento de la diagonal de la matriz H o Leverage

h_{ii} representa el **apalancamiento** de la observación i , es decir, mide qué tan extrema o atípica es esa observación respecto a los valores de las variables independientes. Matemáticamente, el h_{ii} corresponde al elemento i -ésimo de la diagonal de la matriz sombrero H:

$$\mathbf{H} = \mathbf{X}(\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top$$

Un valor de h_{ii} alto indica que la observación tiene una combinación poco común de predictores y, por tanto, puede influir de forma desproporcionada en los resultados del modelo.

Para identificar observaciones con apalancamiento alto, se utiliza:

$$h_{ii} > \frac{2p}{n}$$

donde:

- p es 12, ya que tenemos 11 variables y 1 intersepto
- n es el número total de observaciones 176,725

$$\frac{2 \cdot 11}{176725} = 0,0001245$$

Por lo que las observaciones que cumplan >

$$h_{ii} > 0,0001245$$

Serán clasificadas como una observación con apalancamiento alto, y por ende, se analizará con mayor detalle, ya que podría tener un impacto significativo sobre los coeficientes estimados o sobre las predicciones del modelo.

6.2. Estadística DFFITS

Difference in Fits, mide cuánto cambia la predicción del modelo para la observación i , si eliminamos esa misma observación.

$$\text{DFFITS}_i = t_i \cdot \sqrt{\frac{h_{ii}}{1 - h_{ii}}}$$

Una observación es considerada potencialmente influyente si cumple:

$$|\text{DFFITS}_i| > 2\sqrt{\frac{p}{n}}$$

Por lo tanto, cualquier observación que cumpla:

$$|\text{DFFITS}_i| > 0,0501$$

será considerada influyente sobre su propia predicción y será evaluada en conjunto con otras métricas como el apalancamiento h_{ii} y la distancia de Cook.

6.3. COVRATIO

El estadístico COVRATIO mide el impacto que tiene una observación i sobre la matriz de covarianza de los coeficientes estimados de la regresión

$$\text{COVRATIO}_i = \left(\frac{S_{(i)}^2}{MS_{\text{Res}}} \right)^p \cdot \left(\frac{1}{1 - h_{ii}} \right) \quad (1)$$

Donde:

- $S_{(i)}^2$: Error cuadrático medio al eliminar la observación i .
- MS_{Res} : Error cuadrático medio del modelo completo.
- h_{ii} : Valor de apalancamiento de la observación i .
- p : Número de predictores (incluyendo el intercepto).

Dado que $S_{(i)}^2$ no siempre está disponible directamente, se puede aproximar mediante:

$$\frac{S_{(i)}^2}{MS_{\text{Res}}} \approx 1 - \frac{t_i^2}{p} \quad (2)$$

donde t_i es el residuo eliminado tipificado. Sustituyendo esta expresión en la fórmula del *COVRATIO*, obtenemos:

$$\text{COVRATIO}_i \approx \left(1 - \frac{t_i^2}{p} \right)^p \cdot \left(\frac{1}{1 - h_{ii}} \right) \quad (3)$$

Y para valores pequeños, se puede usar una versión linealizada:

$$\text{COVRATIO}_i \approx 1 - \frac{t_i^2 \cdot h_{ii}}{(1 - h_{ii}) \cdot p} \quad (4)$$

Esta aproximación permite estimar el *COVRATIO* a partir de estadísticas fácilmente disponibles en paquetes como SPSS, sin necesidad de recurrir a la matriz de covarianzas del modelo completo y reducido.

- Si $\text{COVRATIO}_i \approx 1$, el caso no afecta la precisión del modelo.
- Si $\text{COVRATIO}_i \gg 1$, puede subestimar la varianza.
- Si $\text{COVRATIO}_i \ll 1$, indica aumento de la varianza y posible problema de precisión.

Una observación se considera influyente si cumple:

$$|\text{COVRATIO}_i - 1| > 3 \cdot \frac{p}{n}$$

Por lo tanto, toda observación que cumpla:

$$|\text{COVRATIO}_i - 1| > 0,0001867$$

6.4. DFBETAS

La estadística DFBETA mide el cambio en el coeficiente β_j del modelo de regresión al eliminar la observación i :

$$\text{DFBETA}_{ij} = \frac{\hat{\beta}_j - \hat{\beta}_{j(i)}}{SE_{j(i)}}$$

Para identificar observaciones influyentes, se usa el siguiente umbral:

$$|\text{DFBETA}_{ij}| > \frac{2}{\sqrt{n}}$$

Por lo tanto, cualquier observación que cumpla:

$$|\text{DFBETA}_{ij}| > 0,0048$$

es considerada influyente sobre el coeficiente β_j correspondiente.

6.5. Análisis

Para identificar observaciones influyentes en el modelo de regresión, se evaluaron diversos estadísticos de diagnóstico: apalancamiento (h_{ii}), DFFITS, COVRATIO y DFBETAS y se establecieron los siguientes umbrales de decisión:

- **Apalancamiento alto:** Se consideró alta influencia estructural si $h_{ii} > 0,0001245$.
- **Influencia en la predicción:** Se consideró influyente si $\text{DFFITS} > 0,0501$.
- **Influencia sobre la varianza:** Se consideró que una observación incrementa la varianza estimada si $|\text{COVRATIO}_i - 1| > 0,0001867$.
- **Influencia en los coeficientes:** Se consideró influyente si $|\text{DFBETA}_i| > 0,0048$ para cualquier predictor.

A partir de estos criterios, se generaron en SPSS variables binarias que codifican el cumplimiento de cada condición de influencia. Cada variable toma el valor de:

$$\text{Indicador}_i = \begin{cases} 1 & \text{si la observación cumple el criterio de influencia,} \\ 0 & \text{en caso contrario.} \end{cases}$$

Posteriormente, se crearon variables de síntesis para cada estadístico, que toma el valor de 1 si la observación fue marcada como influyente

$$\text{min}(\text{max}(\text{D11110_11a}_6, \text{m1_11a}_6, \text{covm110_11a}_6, \text{D1D11n_11a}_6)$$

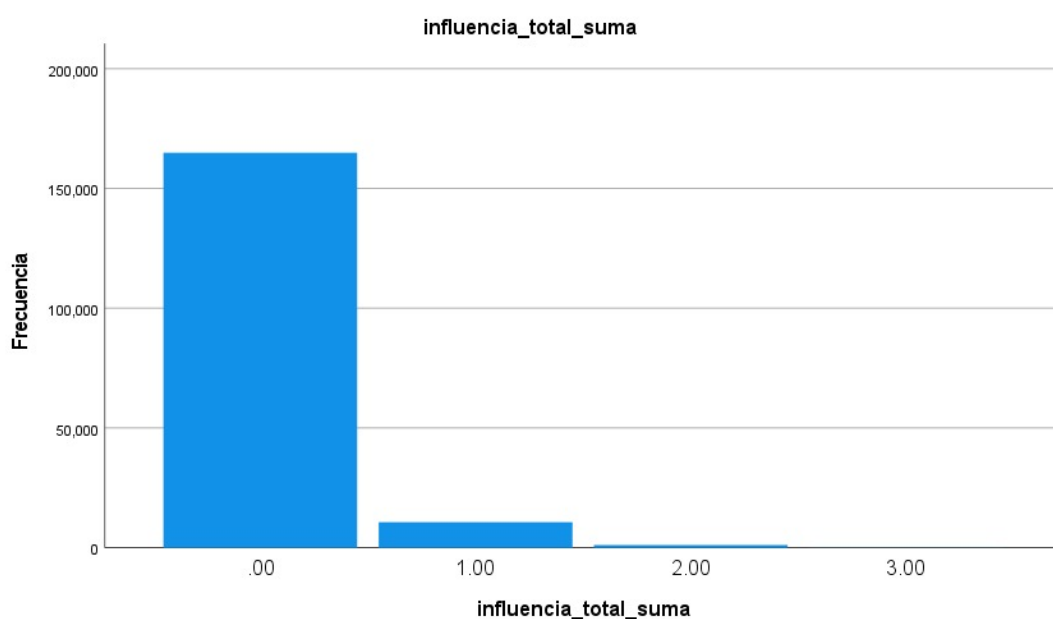
influencia_conjunta: esta variable binaria toma el valor 1 si una observación cumple al menos uno de los siguientes criterios:

Copyright © 2011 John Wiley & Sons, Ltd.

Se construyó la variable `influencia_total_suma` con el objetivo de evaluar la presencia conjunta de observaciones influyentes en el modelo de regresión, considerando los siguientes cuatro criterios: `apalancamiento_alto`, `DFFITS_influyente`, `COVRATIO_influyente` y `algun_dfbeta_influyente`. Esta variable toma valores entre 0 y 4, de acuerdo con el número de condiciones que cumple cada observación.

influencia_total_suma

		Frecuencia	Porcentaje	Porcentaje válido	Porcentaje acumulado
Válido	.00	164911	93.3	93.3	93.3
	1.00	10680	6.0	6.0	99.4
	2.00	1074	.6	.6	100.0
	3.00	60	.0	.0	100.0
	Total	176725	100.0	100.0	



El análisis de frecuencias mostró que el 93.3 % de los casos (164,911 observaciones) no cumplen con ningún criterio de influencia, lo que indica una alta estabilidad del modelo para la mayoría de los datos. Un 6 % (10,680 casos) cumplen un solo criterio, y sólo un 0.6 % (1,074 casos) cumplen dos. Finalmente, apenas el 0.03 % de los casos (60 observaciones) cumplen tres criterios simultáneamente, sin registrarse observaciones que cumplan los cuatro.

Esta distribución sugiere que las observaciones influyentes son muy escasas, lo cual es deseable, ya que implica que el modelo no está siendo afectado significativamente por casos extremos o atípicos.

Compararemos los modelos con y sin estos casos.

6.6. Modelo con casos de influencia

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	.888 ^a	.788	.788	.2271250695

a. Predictores: (Constante), Muy_antiguo, Tipo_edificio_Edificio completo, Estado_edificio_Obra Terminada, Anio_emision, Estado_edificio_Proyecto Nueva, Tipo_edificio_Local, coord_y, Tipo_edificio_Bloque completo, Dias_hasta_expiracion, Tipo_edificio_Unifamiliar, Clasificacion_Emisiones_ordinal

Coeficientes^a

Modelo		Coeficientes no estandarizados		Coeficientes estandarizados	t	Sig.	Correlaciones		
		B	Desv. Error	Beta			Orden cero	Parcial	Parte
1	(Constante)	-10.446	.366		-28.537	<.001			
	Anio_emision	.006	.000	.039	32.761	<.001	-.022	.078	.036
	Dias_hasta_expiracion	2.205E-5	.000	.025	20.534	<.001	-.210	.049	.022
	coord_y	.233	.004	.063	57.158	.000	.068	.135	.063
	Clasificacion_Emisiones_ordinal	.421	.001	.913	725.478	.000	.832	.865	.795
	Estado_edificio_Obra Terminada	-.201	.008	-.028	-24.901	<.001	-.186	-.059	-.027
	Estado_edificio_Proyecto Nueva	-.286	.008	-.038	-34.063	<.001	-.205	-.081	-.037
	Tipo_edificio_Bloque completo	.046	.002	.034	29.406	<.001	.050	.070	.032
	Tipo_edificio_Edificio completo	.553	.005	.126	113.041	.000	-.014	.260	.124
	Tipo_edificio_Local	.521	.002	.254	220.703	.000	.024	.465	.242
	Tipo_edificio_Unifamiliar	.252	.002	.179	151.550	.000	.053	.339	.166
	Muy_antiguo	.018	.002	.011	9.716	<.001	.111	.023	.011

a. Variable dependiente: Log(Emisiones)

6.7. Modelo sin casos de influencia

Resumen del modelo

Modelo	R	R cuadrado	R cuadrado ajustado	Error estándar de la estimación
1	.876 ^a	.768	.768	.2174495521

a. Predictores: (Constante), Muy_antiguo, Anio_emision, coord_y, Tipo_edificio_Local, Tipo_edificio_Bloque completo, Dias_hasta_expiracion, Tipo_edificio_Unifamiliar, Clasificacion_Emisiones_ordinal

Coeficientes ^a						
Modelo		Coeficientes no estandarizados		Coeficientes estandarizados		
		B	Desv. Error	Beta	t	Sig.
1	(Constante)	-12.453	.362		-34.359	<.001
	Anio_emision	.007	.000	.049	38.571	.000
	Dias_hasta_expiracion	2.425E-5	.000	.023	17.292	<.001
	coord_y	.239	.004	.066	55.421	.000
	Clasificacion_Emisiones_ordinal	.426	.001	.909	688.363	.000
	Tipo_edificio_Bloque completo	.046	.002	.036	28.991	<.001
	Tipo_edificio_Local	.556	.003	.272	217.319	.000
	Tipo_edificio_Unifamiliar	.258	.002	.197	156.319	.000
	Muy_antiguo	.024	.002	.014	11.822	<.001

a. Variable dependiente: Log(Emisiones)

■ **Modelo con todos los casos (incluyendo influyentes):**

- R^2 ajustado: 0.788
- Error estándar de la estimación: 0.2271

■ **Modelo sin casos influyentes:**

- R^2 ajustado: 0.768
- Error estándar de la estimación: 0.2174

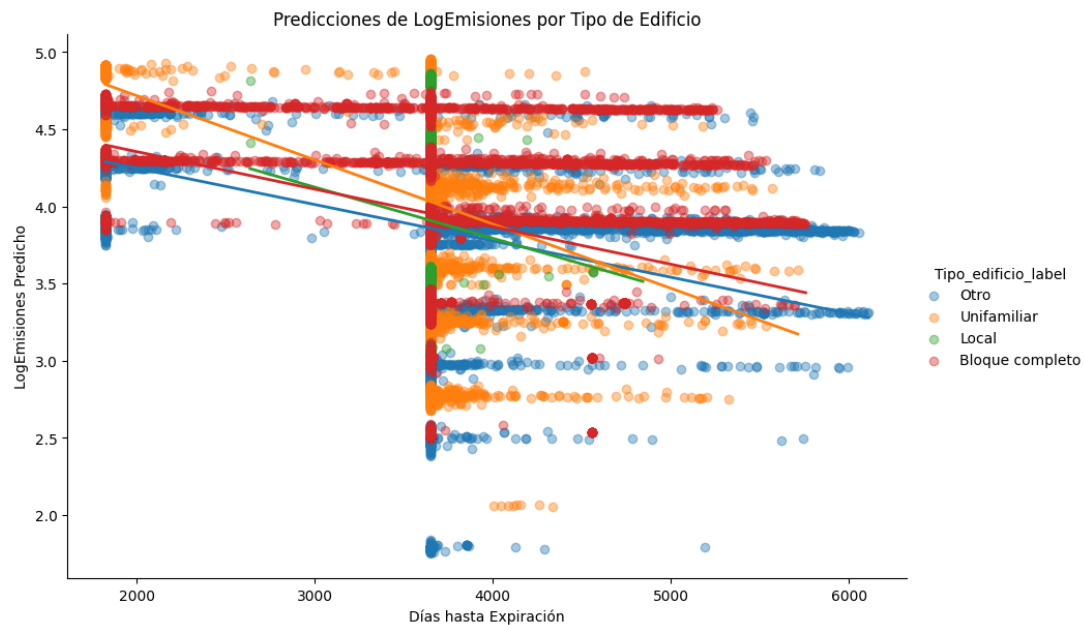
Al remover los casos influyentes, el modelo pierde ligeramente capacidad explicativa (R^2 ajustado disminuye de 0.788 a 0.768), pero también mejora marginalmente en precisión (el error estándar disminuye). Esto sugiere que los casos influyentes estaban aportando información potencialmente distorsionada o inestable por lo que se decide sacarlos ya que su precisión se mantiene alta.

Las bases de datos empleadas, tanto con como sin los valores de influencia, se encuentran en el Anexo 1, hojas BD CON INFLUENCIA y BD SIN INFLUENCIA. Para mayor detalle sobre el código aplicado en el análisis de variables, véase el script en Python incluido en el Anexo 4.

7. Análisis de variables categóricas

7.1. Tipo de edificio

La gráfica muestra que las predicciones de LogEmisiones varían según el tipo de edificio, con diferencias de nivel notables entre categorías y pendientes negativas en casi todos los grupos respecto a los días hasta expiración; se observa que los edificios clasificados como “Bloque completo” tienden a niveles más altos de emisiones predichas, mientras que “Otro (Vivienda individual)” y “Unifamiliar” presentan valores más bajos.



A pesar de que la interacción entre las variables es visualmente evidente, se realizaron pruebas formales que revelaron que dichas interacciones son altamente sensibles a la multicolinealidad, incrementando considerablemente el VIF de las variables originales de las que provienen. Por ello, se evaluaron inicialmente los casos más simples:

- Tipo_Edificio_Local \times Días_Hasta_Expiración (DHE_TEL)
- Tipo_Edificio_Unifamiliar \times Días_Hasta_Expiración (DHE_TEU)
- Tipo_Edificio_Bloque_completo \times Días_Hasta_Expiración (DHE_TEBC)

Coeficientes

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	-12.333	0.362	-34.04	0.000	
Clasificacion_Emisiones_ordinal	0.426195	0.000619	688.16	0.000	1.24
coord_y	0.23940	0.00432	55.45	0.000	1.01
Tipo_edificio_Bloquecompleto	-0.05780	0.00966	-5.98	0.000	40.79
Dias_hasta_expiracion	0.000006	0.000002	3.16	0.002	2.45
Anio_emision	0.006862	0.000179	38.43	0.000	1.15
Muy_antiguo	0.02322	0.00204	11.41	0.000	1.07
Tipo_edificio_Local	-0.027	0.280	-0.10	0.924	13269.99
Tipo_edificio_Unifamiliar	0.0699	0.0175	4.01	0.000	125.49
DHE_TEL	0.000160	0.000077	2.09	0.037	13270.12
DHE_TEU	0.000052	0.000005	10.85	0.000	125.95
DHE_TEBC	0.000029	0.000003	10.81	0.000	39.91

Resumen del modelo

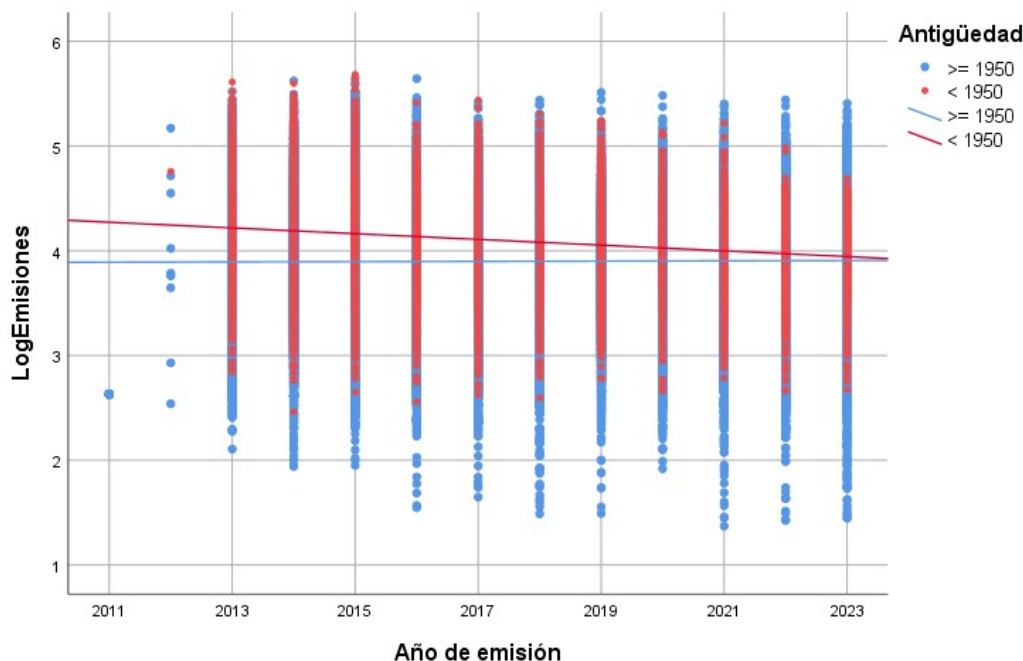
S	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
0.217329	76.84%	76.83%	76.83%

En el primer caso, la interacción resultó significativa únicamente si se eliminaba la variable original `Tipo_Edificio_Local`, obteniéndose el mismo coeficiente de determinación. Esto evidenció que su inclusión era innecesaria, al no aportar mejora alguna al modelo. En los otros dos casos, si bien las interacciones también resultaron significativas y contribuyeron a un ligero aumento del R^2_{ajustado} , generaron un incremento considerable en los VIF, indicando un nivel de multicolinealidad superior al que se considera aceptable para el modelo.

De esta manera, se probó la interacción para las demás variables, observándose el mismo problema. Por tanto, hasta este momento se optó por conservar el modelo finalista sin interacciones adicionales.

7.2. Antigüedad de los edificios

Tenemos aquí una gráfica correspondiente a la relación entre el logaritmo de las emisiones de CO₂ (`LogEmisiones`) y el año de emisión del certificado energético, diferenciando entre edificios construidos antes de 1950 y aquellos construidos a partir de 1950.

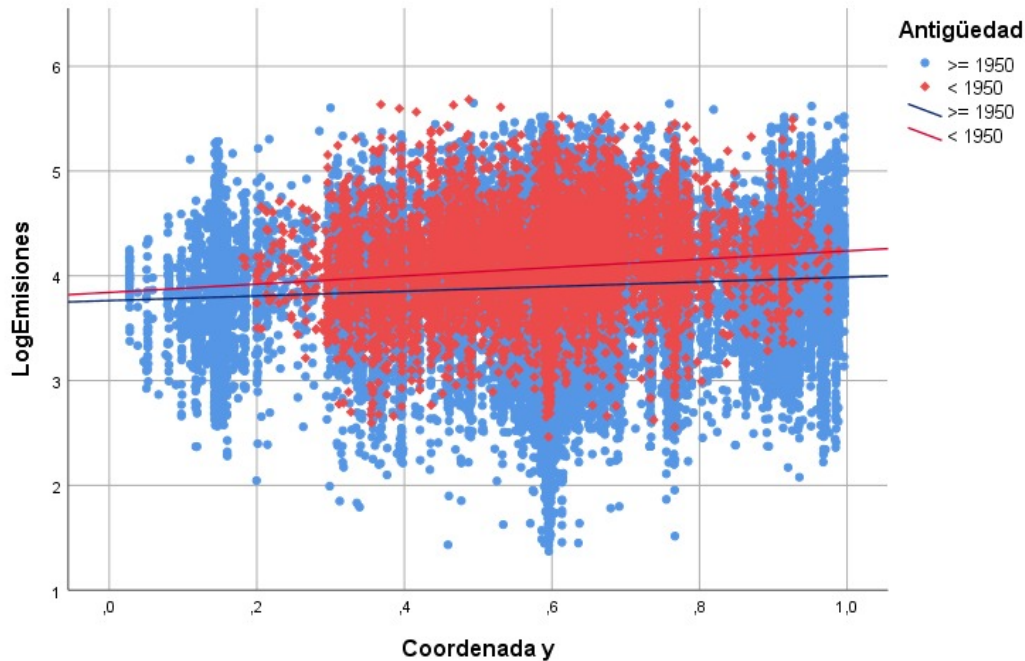


Se observa que, en general, los edificios muy antiguos presentan niveles ligeramente más altos de emisiones en comparación con los más recientes. Además, ambas series muestran una tendencia decreciente, con pendientes ligeramente distintas entre los dos grupos. Esta diferencia sugiere que la antigüedad del edificio podría estar modificando el efecto que tiene el año de emisión sobre las emisiones de CO₂.

Por tanto, consideramos adecuado evaluar la interacción entre las variables `Muy_antiguo` y `Anio_emision` en el modelo, con el objetivo de verificar si el impacto del año de emisión varía según la antigüedad del edificio. Esto nos permitirá comprobar si la tendencia temporal de reducción de emisiones es diferente entre edificios antiguos y recientes.

Por otro lado, tenemos esta gráfica que muestra la relación entre las emisiones de CO₂ (`LogEmisiones`) y la coordenada geográfica *y*, diferenciando entre edificios construidos

antes y después de 1950.



Se observa que ambos grupos presentan una tendencia ligeramente creciente en las emisiones conforme aumenta la coordenada y . Además, existe superposición entre ambos grupos a lo largo del eje horizontal. Aunque no se aprecia una diferencia sustancial en la tendencia según la antigüedad, de todas formas evaluaremos la interacción entre las variables `Muy_antiguo` y `coord_y` en el modelo.

Entonces, el modelo propuesto es

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \beta_9 x_1 x_8 + \beta_{10} x_3 x_8 + \varepsilon$$

x_1 : Año de emisión (`Anio_emision`)

x_3 : Coordenada y (`coord_y`)

x_8 : Antigüedad de la edificación (`Muy_antiguo`)

Para probar si los dos modelos de regresión tienen distintas pendientes las hipótesis serían

$$H_0 : \beta_9 = \beta_{10} = 0 \quad vs \quad H_1 : \beta_9 \neq 0 \text{ y/o } \beta_{10} \neq 0$$

Sea $n = 164911$, $p = 10$, $k = 2$ y nivel de significancia $\alpha = 0,05$.

El estadístico de prueba es

$$F_c = \frac{SCE(\beta_{10}, \beta_9 \mid \beta_8, \beta_7, \dots, \beta_0)}{k \hat{\sigma}_{MCO}^2} \sim F(k, n - p - 1),$$

Construimos nuestra tabla ANVA con la siguiente identidad

$$SCT^* = SCE(\beta_{10}, \beta_9) + SCE(\beta_{10}, \beta_9 \mid \beta_8, \beta_7, \dots, \beta_0) + SCR$$

FV	SC	gl	CM	F_0	p_value
$SC(\beta_8, \beta_7, \dots, \beta_0)$	25825.4	8	3228.175		
$SC(\beta_{10}, \beta_9 \mid \beta_8, \beta_7, \dots, \beta_0)$	6.6	2	3.3	69.85	0.000
Residual	7790.5	164900	0.0472437		
Total	33622.5	164910			

Decisión : Dado que $p < 0.05$, entonces **se rechaza H_0**

Conclusión : Bajo un nivel de significancia de 0.05 se concluye que la interacción contribuye significativamente al modelo.

Sin embargo, al analizar nuevamente los valores del VIF de las variables, se obtuvo lo siguiente:

Coefficientes

Término	Coef	EE del coef.	Valor T	Valor p	FIV
Constante	-13.507	0.375	-36.04	0.000	
Clasificacion_Emissiones_ordinal	0.425550	0.000619	687.58	0.000	1.24
coord_y	0.23388	0.00447	52.34	0.000	1.08
Tipo_edificio_Bloquecompleto	0.04553	0.00158	28.83	0.000	1.09
Dias_hasta_expiracion	0.000025	0.000001	17.71	0.000	1.21
Anio_emision	0.007413	0.000185	40.13	0.000	1.23
Muy_antiguo	14.15	1.30	10.92	0.000	432484.95
Tipo_edificio_Local	0.55624	0.00256	217.36	0.000	1.11
Tipo_edificio_Unifamiliar	0.26032	0.00166	156.80	0.000	1.13
MA_AE	-0.007021	0.000642	-10.94	0.000	432185.69
MA_CY	0.0758	0.0172	4.41	0.000	27.79

Resumen del modelo

S	R-cuad.	R-cuad. (ajustado)	R-cuad. (pred)
0.217356	76.83%	76.83%	76.83%

A pesar de que algunas de las interacciones evaluadas resultaron ser estadísticamente significativas, se observa que persiste una alta sensibilidad al incremento del VIF, lo cual refuerza la presencia de multicolinealidad en el modelo. Esta situación compromete la estabilidad e interpretabilidad de los coeficientes estimados.

Por tanto, se optó por conservar únicamente las ocho variables que explican mejor la variabilidad en las emisiones. De esta forma, se define el modelo finalista.

$$\log(y) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + \beta_5 x_5 + \beta_6 x_6 + \beta_7 x_7 + \beta_8 x_8 + \varepsilon$$

x_1 : Año de emisión

x_2 : Días hasta expiración

x_3 : Coordenadas y (longitud)

x_4 : Clasificación de emisiones

x_5 : Edificio de tipo bloque completo

x_6 : Edificio de tipo local

x_7 : Edificio de tipo unifamiliar

x_8 : Antigüedad de la edificación (Muy_antiguo)

8. Validación del modelo

La validación de nuestro modelo se llevará a cabo mediante la comparación entre el R^2 obtenido en el conjunto de entrenamiento y el $R^2_{predicción}$ calculado en el conjunto de prueba.

- Si $R^2_{pred} \approx R^2$ de entrenamiento, el modelo generaliza bien.
- Si $R^2_{pred} \ll R^2$ de entrenamiento, hay sobreajuste (*overfitting*).
- Si R^2_{pred} es cercano a 0 o negativo, el modelo no predice mejor que la media.

¿Cómo funciona el entrenamiento de modelos en Python?

El proceso de entrenamiento para ajustar modelos lineales en `python`, como `LinearRegression` de `sklearn`, se realiza mediante la **Descomposición en Valores Singulares (SVD)**. Esta técnica permite resolver el problema sin necesidad de invertir la matriz $X'X$, lo que mejora la estabilidad numérica y hace que el método sea más robusto ante problemas como matrices mal condicionadas o columnas casi colineales.

$$\hat{\beta} = \arg \min_{\beta} \varepsilon' \varepsilon \Rightarrow (X'X)\hat{\beta} = X'Y$$

Con **SVD**, se descompone $X = U\Sigma V'$, y se obtiene:

$$\hat{\beta} = V\Sigma^{-1}U'Y$$

Donde:

- U : matriz $n \times n$, con los vectores ortogonales de las observaciones.
- Σ : matriz diagonal $n \times p$, con los valores singulares.
- V : matriz $p \times p$, con los vectores ortogonales de las variables originales.

¿Es la estimación igual a MCO?

Sí, la estimación es igual a la de **Mínimos Cuadrados Ordinarios (MCO)** si la matriz X tiene **rango completo**:

$$\hat{\beta}_{\text{MCO}} = (X'X)^{-1}X'Y$$

Esta estimación es equivalente a:

$$\hat{\beta}_{\text{SVD}} = V\Sigma^{-1}U'Y$$

Ambas expresiones dan el mismo resultado, pero **la versión con SVD es más estable numéricamente**, especialmente cuando hay colinealidad o diferencias grandes en las escalas de las variables.

Código en Python para validar el modelo

```
# Importación de librerías necesarias

# Para manipulación de datos con DataFrames
import pandas as pd

# Para ajustar el modelo de regresión lineal
from sklearn.linear_model import LinearRegression

# Para dividir los datos en entrenamiento y prueba
from sklearn.model_selection import train_test_split

# Para calcular el R² de predicción sobre datos de prueba
from sklearn.metrics import r2_score

# 1. Variables predictoras y respuesta
X = df_3[["Clasificacion_Emissiones_ordinal",
          "Tipo_edificio_Local",
          "Tipo_edificio_Unifamiliar",
          "coord_y",
          "Tipo_edificio_Bloquecompleto",
          "Dias_hasta_expiracion",
          "Anio_emision",
          "Muy_antiguo"]]
y = df_3["LogEmissiones"]

# 2. Separar en entrenamiento y test (80% - 20%)
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2, random_state=42)

# 3. Ajustar modelo de regresión lineal
modelo = LinearRegression()
modelo.fit(X_train, y_train)

# 4. Calcular R² en entrenamiento
r2 = modelo.score(X_train, y_train)

# 5. Calcular R² ajustado
n = X_train.shape[0] # número de observaciones
p = X_train.shape[1] # número de predictores

r2_ajustado = 1 - (1 - r2) * (n - 1) / (n - p - 1)

# 6. Calcular R² de predicción (en test)
y_pred = modelo.predict(X_test)
r2_pred = r2_score(y_test, y_pred)

# 7. Mostrar resultados
print(f"R² : {r2:.6f}")
print(f"R² ajustado : {r2_ajustado:.6f}")
print(f"R² de predicción: {r2_pred:.6f}")
```

R^2 : 0.768845
 R^2 ajustado : 0.768831
 R^2 de predicción: 0.764966

El modelo ajustado presenta un coeficiente de determinación R^2 de 0,768845, lo que indica que aproximadamente el 76,9% de la variabilidad observada en las emisiones de CO₂ es explicada por las variables incluidas en el modelo. Por su parte, el R^2_{ajustado} es de 0,768831, lo cual refleja una penalización mínima por el número de predictores utilizados, confirmando que el modelo es parsimonioso. Asimismo, el coeficiente de determinación en el conjunto de prueba ($R^2_{\text{predicción}} = 0,764966$) es muy similar al obtenido en el entrenamiento. Esto se resume en la siguiente relación:

$$R^2_{\text{pred}} \approx R^2$$

lo cual indica que el modelo generaliza bien y tiene una adecuada capacidad predictiva en datos no utilizados durante el ajuste.

9. Resultados

9.1. Resumen

Este proyecto desarrolló un modelo de regresión lineal múltiple con el objetivo de predecir las emisiones de CO₂ de edificios residenciales en Aragón, España. Dado que la variable de emisiones presentaba una distribución fuertemente sesgada, se aplicó una transformación logarítmica para mejorar el cumplimiento de los supuestos clásicos de regresión: normalidad, homocedasticidad y linealidad.

El modelo final se construyó a partir de un conjunto de variables seleccionadas mediante el método stepwise, entre las que destacan: clasificación energética del edificio, tipo de inmueble, año de emisión y ubicación geográfica. Estas variables explican con buen nivel de ajuste la variación en el logaritmo de las emisiones de CO₂.

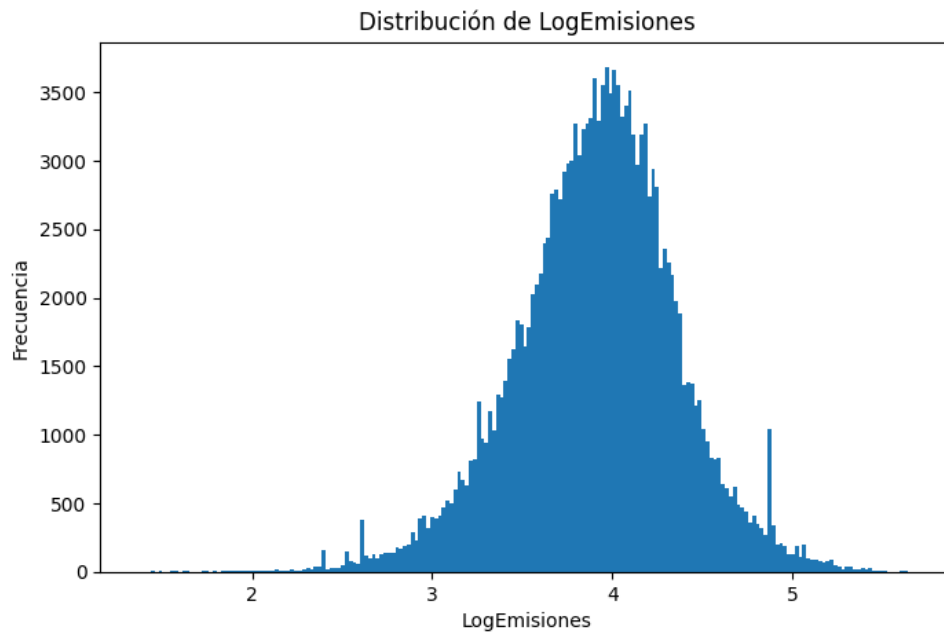
Durante el análisis de calidad del modelo, se aplicaron pruebas de influencia (apalancamiento, DFFITS, DFBETAS y COVRATIO) para identificar observaciones con un impacto desproporcionado sobre los coeficientes. Se excluyeron todos los casos que cumplieran al menos un criterio de influencia. Como resultado, el modelo ajustado final —calculado sin los valores influyentes— alcanzó un **R² ajustado de 0.7688**, lo que indica que el 76.9 % de la variabilidad en las emisiones logarítmicas puede explicarse mediante las variables seleccionadas. El error estándar de la estimación también disminuyó ligeramente tras la depuración, lo que refuerza la precisión del modelo.

En síntesis, el modelo muestra un equilibrio adecuado entre ajuste, interpretabilidad y solidez estadística. Ofrece una base fiable para el análisis de eficiencia energética y puede servir como apoyo en la toma de decisiones en políticas públicas de sostenibilidad en el ámbito residencial.

9.2. Análisis gráfico complementario de los resultados

Como parte del análisis complementario, se incluyen representaciones gráficas que permiten ilustrar de forma visual los resultados obtenidos.

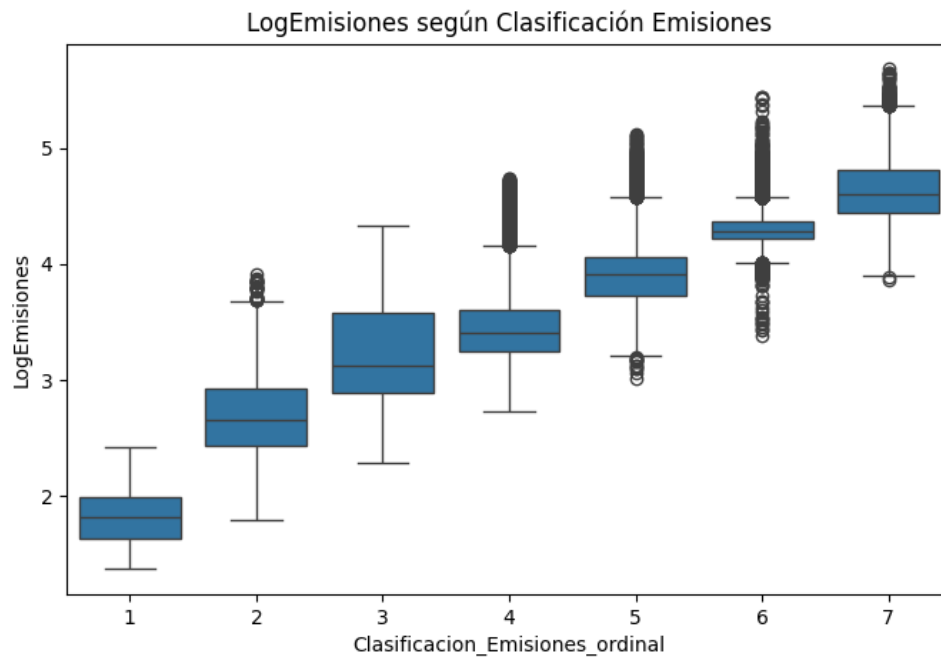
La distribución de LogEmisiones es aproximadamente normal con una leve asimetría a la derecha; en general, presenta una forma adecuada para el análisis de regresión.



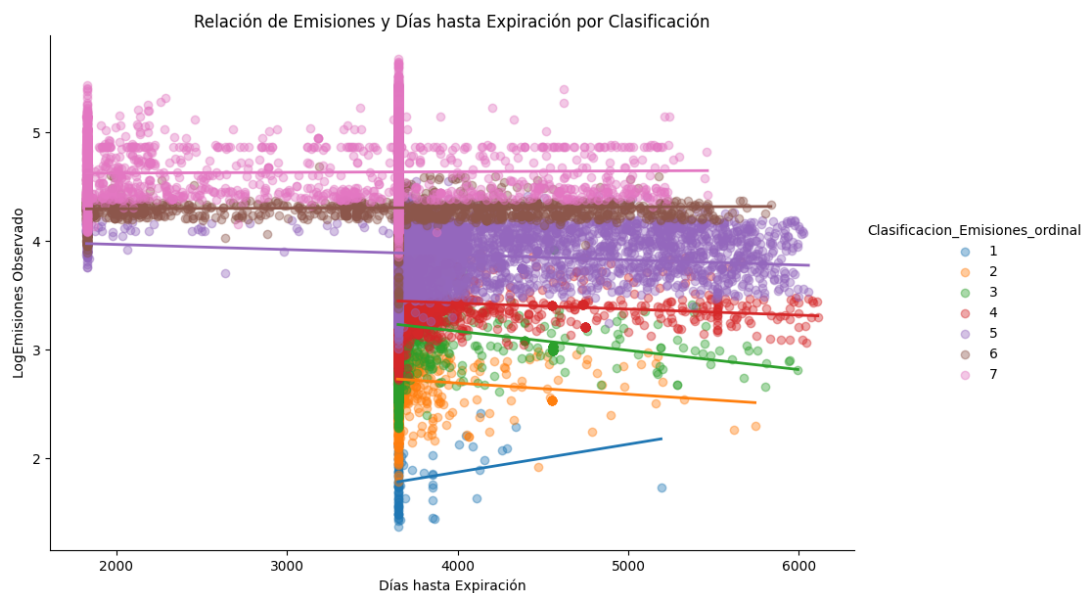
Recordando que teníamos una variable llamada clasificación, que a pesar que ya tiene una transformación, la dummizaremos para ver si obtenemos mejores resultados

Clasif_2	Clasif_3	Clasif_4	Clasif_5	Clasif_6	Clasif_7
0	0	0	1	0	0
0	0	0	1	0	0
0	0	1	0	0	0
0	0	0	1	0	0
0	0	0	0	0	1
...
0	0	0	1	0	0
0	0	0	0	1	0
1	0	0	0	0	0
0	0	0	1	0	0
0	0	0	0	0	1

Revisando su relación con logEmisiones, se evidencia que puede tratarse de rectas con distinta pendiente



La gráfica muestra que las predicciones de LogEmisiones aumentan de forma parecida con el año en todas las clasificaciones, pero cada grupo tiene un nivel diferente de emisiones; es decir, las líneas son casi paralelas y se diferencian sobre todo por la altura, con pequeñas desviaciones en las pendientes también. Pero se ve claramente que las rectas por clasificación son distintas, por lo que tratar la variable como dummie es la decisión correcta.



9.3. Interpretación de los resultados

Ahora exploraremos los resultados de una manera más detallada:

OLS Regression Results			
=====			
Dep. Variable:	LogEmisiones	R-squared:	0.776
Model:	OLS	Adj. R-squared:	0.776
Method:	Least Squares	F-statistic:	4.406e+04
Date:	Sun, 29 Jun 2025	Prob (F-statistic):	0.00
Time:	22:32:56	Log-Likelihood:	20662.
No. Observations:	164911	AIC:	-4.130e+04
Df Residuals:	164897	BIC:	-4.116e+04
Df Model:	13		
Covariance Type:	nonrobust		

El modelo presenta un coeficiente de determinación $R^2 = 0,776$, lo que indica que aproximadamente el 77.6 % de la variabilidad en LogEmisiones es explicada por las variables independientes incluidas. El estadístico F obtenido ($F = 44,060$) es muy superior al valor crítico correspondiente a un nivel de significancia de $\alpha = 0,05$, lo que permite rechazar la hipótesis nula de que todos los coeficientes poblacionales son iguales a cero. Por tanto, se concluye que el modelo es globalmente significativo y posee un alto poder explicativo.

=====						
	coef	std err	t	P> t	[0.025	0.975]

Intercept	-12.4533	0.362	-34.359	0.000	-13.164	-11.743
Anio_emision	0.0069	0.000	38.571	0.000	0.007	0.007
Dias_hasta_expiracion	2.425e-05	1.4e-06	17.292	0.000	2.15e-05	2.7e-05
coord_y	0.2394	0.004	55.421	0.000	0.231	0.248
Muy_antiguo	0.0241	0.002	11.822	0.000	0.020	0.028
Tipo_edificio_Local	0.5564	0.003	217.319	0.000	0.551	0.561
Tipo_edificio_Unifamiliar	0.2584	0.002	156.319	0.000	0.255	0.262
Tipo_edificio_Bloquecompleto	0.0458	0.002	28.991	0.000	0.043	0.049
Clasificacion_Emisiones_ordinal	0.4259	0.001	688.363	0.000	0.425	0.427
=====						

- Todas las variables del modelo resultaron estadísticamente significativas con p-valores menores a 0.001.
- La variable *Clasificacion_Emisiones_ordinal* presentó un coeficiente positivo de 0.4259, indicando que cada incremento en la clasificación de emisiones se asocia con un aumento promedio de 0.426 unidades en LogEmisiones, siendo este efecto uno de los más relevantes del modelo.
- Las variables de tipo de edificio mostraron efectos positivos significativos respecto a la categoría base, destacando los edificios clasificados como “Local” (coeficiente 0.5564) y “Unifamiliar” (0.2584).
- El año de emisión y los días hasta expiración presentaron efectos positivos de menor magnitud, pero también significativos, con coeficientes de 0.0069 y 2.425e-05 respectivamente.
- La variable *Muy antiguo* tuvo un coeficiente positivo de 0.0241, lo que sugiere una ligera asociación entre la antigüedad y mayores emisiones estimadas.
- El intervalo de confianza al 95 % de todos los coeficientes no incluye el valor cero, confirmando la robustez estadística de los efectos estimados.

10. Conclusiones

10.1. Conclusiones por objetivos

Objetivo 1: Identificar las variables más significativas de los certificados energéticos para la predicción de emisiones de CO₂

En base a los resultados se concluye que las variables más significativas fueron las de clasificación de emisiones, que mostraron los coeficientes más altos y p-valores menores a 0.001, indicando que conforme aumenta la categoría ordinal, el LogEmisiones crece considerablemente. Otras variables relevantes fueron el tipo de edificio, el año de emisión y la coordenada Y, todos con efectos estadísticamente significativos, aunque de menor magnitud.

Objetivo 2: Determinar qué tipos de edificios presentan una mayor contribución a las emisiones de CO₂

Los resultados del modelo de regresión, respaldados por los gráficos de predicciones, indican que el tipo de edificio tiene un efecto relevante sobre los niveles de LogEmisiones. En particular, los edificios clasificados como “Bloque completo” y “Local” presentan valores predichos superiores de manera consistente respecto a otras categorías, mientras que “Unifamiliar” muestra un patrón decreciente más pronunciado conforme aumenta el tiempo hasta expiración. La categoría “Otro” se asocia con menores emisiones estimadas. Estas diferencias son estadísticamente significativas y evidencian la relevancia del uso y tipología del edificio en la predicción de emisiones de CO₂.

Objetivo 3: Evaluar si las viviendas construidas antes del año 1950 tienen una mayor influencia en las emisiones de CO₂.

En base a los resultados obtenidos, se concluye que las edificaciones construidas antes del año 1950 presentan una mayor influencia en las emisiones de CO₂. Esta afirmación se fundamenta, en primer lugar, en la gráfica de `logEmisiones` frente a las coordenadas geográficas `y`, donde se aprecia una concentración de valores altos de emisión en zonas asociadas a construcciones más antiguas. Asimismo, en la visualización de `logEmisiones` en función del `Año_Emisión`, se observa que las edificaciones construidas antes de 1950 tienden a agruparse de manera más compacta en niveles elevados de emisión, en comparación con las más recientes.

Aunque se detecta una tendencia general decreciente en las emisiones con el paso de los años, la evidencia gráfica respalda que las viviendas muy antiguas contribuyen de forma más significativa a niveles altos de emisión de CO₂, lo cual justifica su consideración dentro del modelo como una variable relevante.

10.2. Propuestas de mejora o trabajos futuros

A partir de los resultados obtenidos y las limitaciones identificadas, se plantean diversas líneas de mejora y futuras investigaciones que pueden fortalecer y ampliar el alcance del presente trabajo.

Una primera propuesta consiste en replicar el modelo en otras comunidades autónomas o

a nivel nacional, con el fin de evaluar su capacidad de generalización. Comparar diferentes contextos geográficos permitiría identificar patrones estructurales y normativos en relación con las emisiones de CO₂ en la edificación.

Además, se propone explorar técnicas estadísticas alternativas como modelos machine learning (e.g., árboles de decisión, bosques aleatorios o regresión regularizada), que podrían ofrecer mejoras en predicción sin requerir supuestos estrictos de normalidad y linealidad.

También se recomienda incorporar nuevas variables dependiendo del lugar en el que se haga los estudios que podrían enriquecer el análisis, tales como indicadores de ocupación, hábitos de consumo energético, calidad del aislamiento, sistemas de climatización o fuentes de energía renovable utilizadas.

11. Anexos

1. Anexo 1. Base de datos

Archivo en formato Excel que contiene las siguientes hojas:

- **BD_INICIAL:** Base de datos original antes de la limpieza.
- **BD_LIMPIO:** Base de datos tras el tratamiento y depuración de variables regresoras.
- **BD CON INFLUENCIA:** Base de datos con los valores de influencia incluidos.
- **BD SIN INFLUENCIA:** Base de datos resultante tras excluir las observaciones influyentes.

2. Anexo 2. Diccionario de variables

Tablas descriptivas de las variables utilizadas en el modelo.

3. Anexo 3. Cp de Mallows

Resultados del Cp de Mallows para los diferentes modelos propuestos.

4. Anexo 4. Analisis_CO2_sin_valores_de_influenciaipynb Colab

Código en Python utilizado para el tratamiento de datos, detección de observaciones influyentes y análisis de variables.