1. Demostración.
2. Regresión lineal múltiple.
   Para el conjunto de datos states que contiene datos faltantes, se quitó todo un renglón con la mayoría de los datos faltantes, y para las columnas los faltantes se sustituyeron con la media de esa variable.

   Se examinaron los datos de las variables de interés (metro y energy), observe la Figura 2.1 y note que para la variable energy hay posibles outliers y su rango de valores es más amplio que el de la variable metro.

   También observe la Figura 2.2 y note que la variable de interés energy (energía consumida per cápita) parece tener correlación débil negativa con la variable metro (porcentaje de residentes que viven en áreas metropolitanas), además también parece tener correlación media positiva con otras variables como green, área, toxic, así que también podríamos utilizar alguna de estas variables para pronosticar energy.

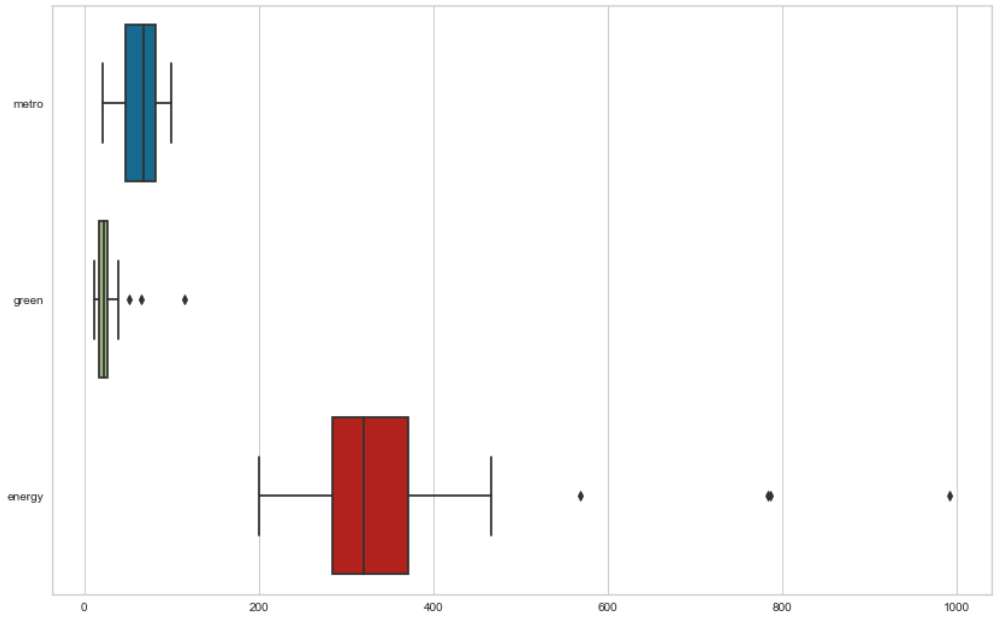

Figura 2.1: Boxplot de las variables metro, green y energy.

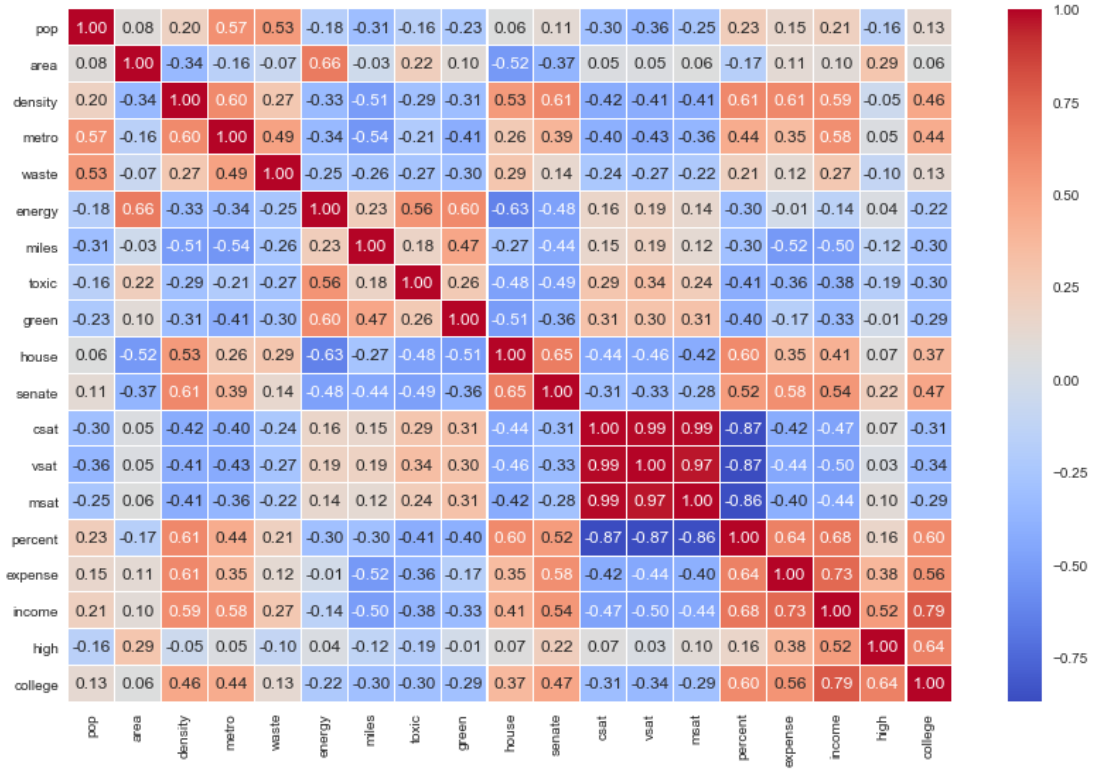

Figura 2.2: Correlación de las variables en el conjunto de datos, tonos rojos intensos son correlación positiva fuerte, tonos azules intensos son correlación negativa fuerte.

Posteriormente se ajustó el modelo de regresión lineal clásico para y (energy) y z_1 (metro):

$$y = -1.8366 z_1 + 441.1054 \qquad (1.1)$$

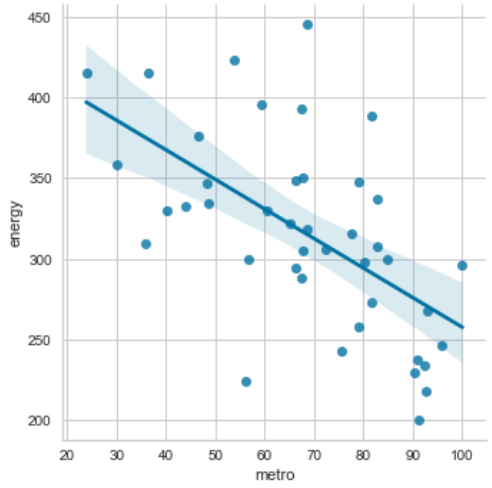

Figura 2.3: Gráfico del modelo 1.1

El modelo nos dice que si no hay algún cambio en el porcentaje de residentes que viven en área metropolitana (metro), el valor la energía consumida per cápita (energy) es 441 y que por cada unidad de cambio en la variable metro la variable energy disminuye 1.83 unidades. También del ajuste del modelo pudimos concluir que la calidad del ajuste no fue muy buena ($R^2 = 0.348$), pero la variable metro si es explicativa en el modelo (el intervalo de confianza [-2.6,-1] no incluye al 0).

Analizando los residuales no muestran simetría o algún patrón por lo que parece que el modelo no necesita de más términos (Observe la Figura 2.4 a), la varianza en lo general parece ser constante, se mantiene entre 100 y -100 (Observe la Figura 2.4 b) y además los residuales también cumplen con el supuesto de normalidad observe la Figura 2.5 y note que los residuales se ajustan en su mayoría a la línea roja (Shaphiro test, W=0.98).

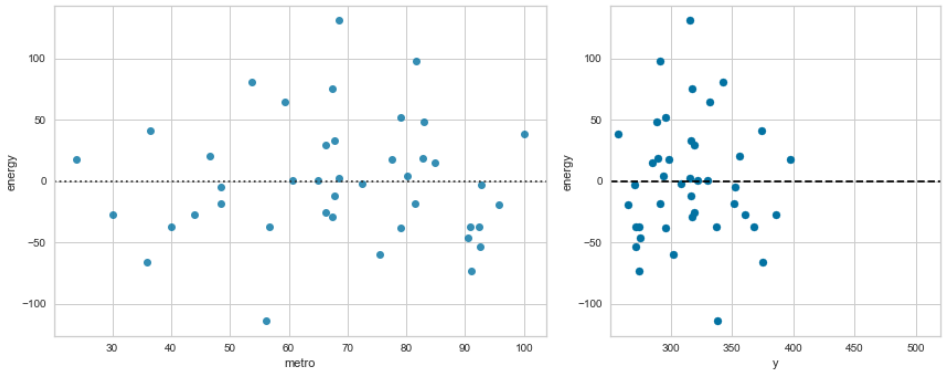

Figura 2.4: a) residuales contra variable metro, b) residuales contra la predicción.

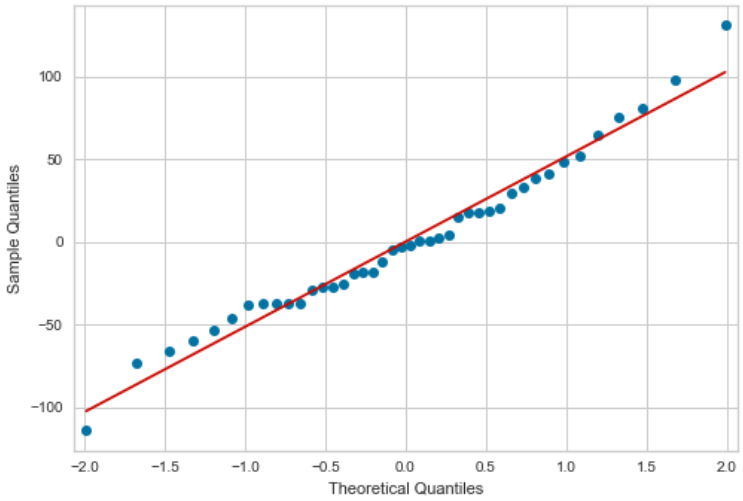

Figura 2.5: QQ-plot de los residuales.

Posteriormente se utilizó también la variable z_2 (green), observe la Figura 2.1 note que también hay posibles datos atípicos además note que parece tener correlación positiva con la variable energy (Figura 2.2).
Nuevamente se ajustó un modelo de regresión lineal clásico para explicar o pronosticar y (energy) con las variables z_1 (metro) y z_2 (green):

$$y = -0.6904z_1 + 5.7493z_2 + 237.0577 \qquad (1.2)$$

El modelo nos dice que si no hay algún cambio en el porcentaje de residentes que viven en área metropolitana (metro), y en la variable green, el valor la energía consumida per cápita (energy) es 237 y que por cada unidad de cambio en la variable metro y en la variable green, la variable energy disminuye en 0.69 unidades y aumenta en 5.7 unidades respectivamente. También del ajuste del modelo pudimos concluir que la calidad del ajuste mejoró ($R^2 = 0.568$), y que ambas variables (metro y green) si son explicativas en el modelo (los intervalos de confianza [-1.3,-0.04] y [3.5,7.9] no incluyen al 0).

Analizando los residuales no muestran simetría o algún patrón por lo que parece que el modelo no necesita de más términos (Observe la Figura 2.6 a), la varianza en lo general parece ser constante, se mantiene entre 100 y -100 (Observe la Figura 2.6 b) y además los residuales también cumplen con el supuesto de normalidad univariada observe la Figura 2.7 y note que los residuales se ajustan en su mayoría a la línea roja (Shaphiro test, W=0.97).

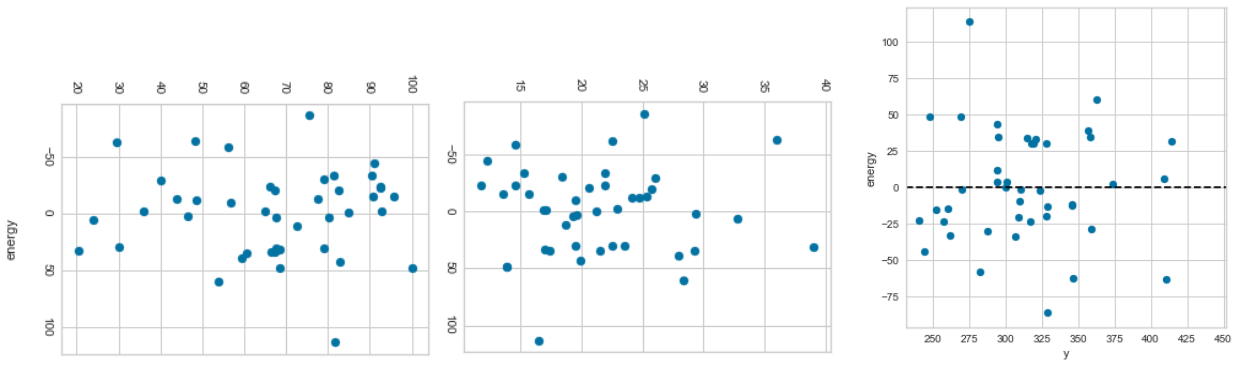

Figura 2.6: a) residuales contra variables metro y green, b) residuales contra la predicción.

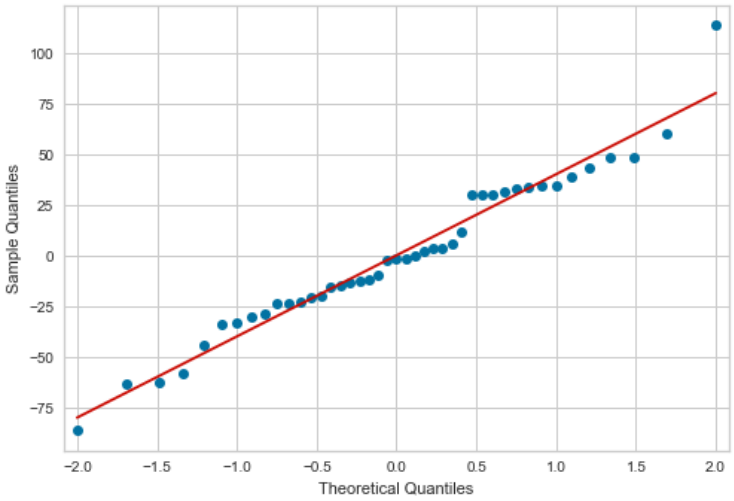

Figura 2.7: QQ-plot de los residuales.

Nota: para el ajuste de los modelos en este y los demás ejercicios, sólo se retiraron outliers (siempre que se cumpliera que las observaciones $n \geq 30$) identificándolos mediante la distancia de Cook para el caso invariado, y para el caso multivariado utilizando el grafico chi cuadrado. También es importante señalar que antes de retirar posibles outliers, se debe consultar con el experto en el tema, sin embargo, aquí se realizó de esta forma para ver el cambio en los modelos después de retirar los datos atípicos, en la mayoría de los casos el modelo mejoró, cumpliendo con los supuestos de residuales distribuidos normal y con varianza constante.

3. Costo de vida.

Para el conjunto de datos costofliving, que enumeran estadísticas del costo de vida para cada uno de los 50 estados de USA, que son alquileres de apartamentos (y_1), costo de casas (y_2) y el índice de costo de vida (y_3) se realizó una regresión lineal multivariada $\left(y=\left[y_1 \quad y_2 \quad y_3\right]\right)$ en términos de las variables de poblaciones estatales $\left(z_1\right)$ e ingresos medios $\left(z_2\right), \left(z=\left[1 \quad z_1 \quad z_2\right]\right)$:

$$\left[y\right]=\left[z\right]\begin{bmatrix} 532.139 & 52.084 & 84.867 \\ 0.008 & 1.4e-3 & -1.17e-4 \\ 4.663 & 2.929 & 3.9e-1 \end{bmatrix} \tag{1.3}$$

Mediante un Test MANOVA (Wilks) para checar la hipotesis nula $H_0: \beta_1=0, \beta_2=0$ al 95%, es decir si estas variables no explican las variables de respuesta $\left(y\right)$ y se determinó que ambas variables $\left(z_1, z_2\right)$ si son explicativas dentro del modelo multivariado. También se examinó la normalidad multivariada de los residuales mediante el grafico chi-cuadrado (Observe la Figura 3.1 y note que es bueno el ajuste de los puntos a la línea, no hay mucha desviación) y además la prueba de Looney & Gulledge también acepta la hipotesis de normalidad multivariada al 95%, por lo que el modelo cumple con los supuestos y es explicativo.

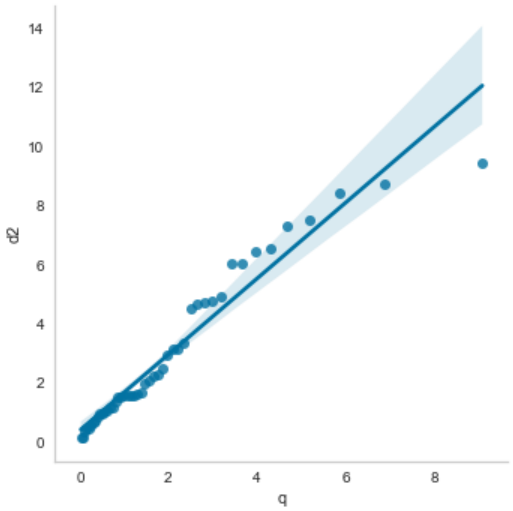

Figura 3.1: Gráfico chi cuadrado de los residuales del modelo 1.3

Posteriormente se ajustaron de forma separada 3 modelos de regresión lineal clásicos:

$$y_1 = 0.0103z_1 + 9.0603z_2 + 298.2664 \qquad R^2 = 0.402 \qquad (1.4)$$

$$y_2 = 0z_1 + 7.4053z_2 - 133.5838 \qquad R^2 = 0.461 \qquad (1.5)$$

$$y_3 = 0z_1 + 0.3864z_2 + 81.1262 \qquad R^2 = 0.171 \qquad (1.6)$$

Podemos observar que los modelos (1.4),(1.5) su calidad de ajuste es media, en comparación con el modelo (1.6) que es débil, además mientras que para el primer modelo ambas variables (z_1, z_2) explican bien a los alquileres de departamentos (y_1), para los modelos (1.5)(1.6), la variable de población (z_1) no explica costo de las casas ni el índice de costo de vida (y_2, y_3), únicamente la variable de ingresos medios (z_2).

Comparando estos resultados con el modelo multivariado, podemos concluir que en este último se retiraron menos datos atípicos para ajustar un modelo con ambas variables que en conjunto explique bien las 3 métricas de costo de vida, mientras que por separado solo 1 modelo lo explica en conjunto y los otros dos no, además de que para lograr el ajuste se perdió mayor cantidad de datos.

El análisis de los residuales para cada modelo:

Modelo 1

Analizando los residuales, estos no muestran simetría o algún patrón por lo que parece que el modelo no necesita de más términos o que no hay errores de cálculo (Observe la Figura 3.2 a), la varianza en lo general parece ser constante, se mantiene entre 200 y -200 (Observe la Figura 3.2 b) y además los residuales también cumplen con el supuesto de normalidad univariada (Shaphiro test, W=0.97, p=0.58).

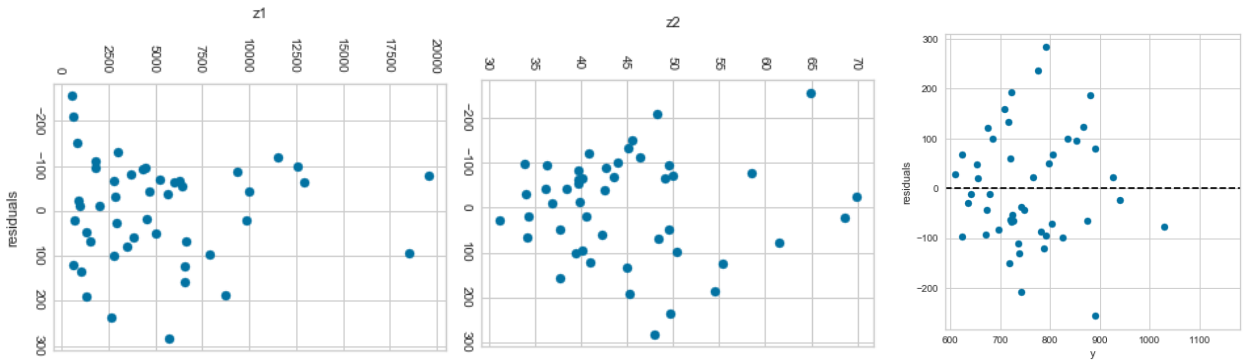

Figura 3.2: a) residuales contra variables (z_1, z_2), b) residuales contra la predicción y_1.

Modelo 2:

Analizando los residuales, estos no muestran simetría o algún patrón por lo que parece que el modelo no necesita de más términos o que no hay errores de cálculo (Observe la Figura 3.3 a), la varianza en lo general parece ser constante, se mantiene entre 100 y -100 (Observe la Figura 3.3 b) y además los residuales también cumplen con el supuesto de normalidad univariada (Shaphiro test, W=0.97, p=0.38).

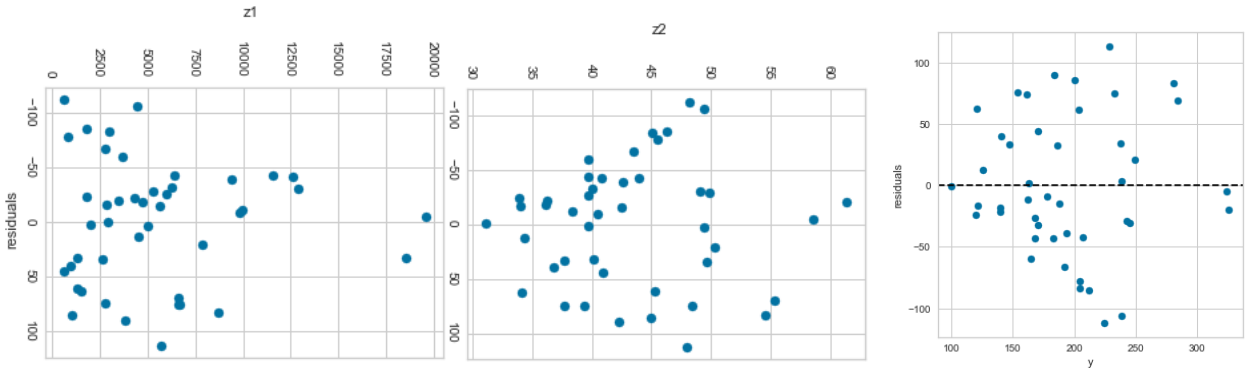

Figura 3.3: a) residuales contra variables (z_1, z_2), b) residuales contra la predicción y_2.

Modelo 3:

Analizando los residuales, estos no muestran simetría o algún patrón por lo que parece que el modelo no necesita de más términos o que no hay errores de cálculo (Observe la Figura 3.4 a), la varianza parece ser constante, se mantiene entre 10 y -10 (Observe la Figura 3.4 b) y además los residuales también cumplen con el supuesto de normalidad univariada (Shaphiro test, W=0.94).

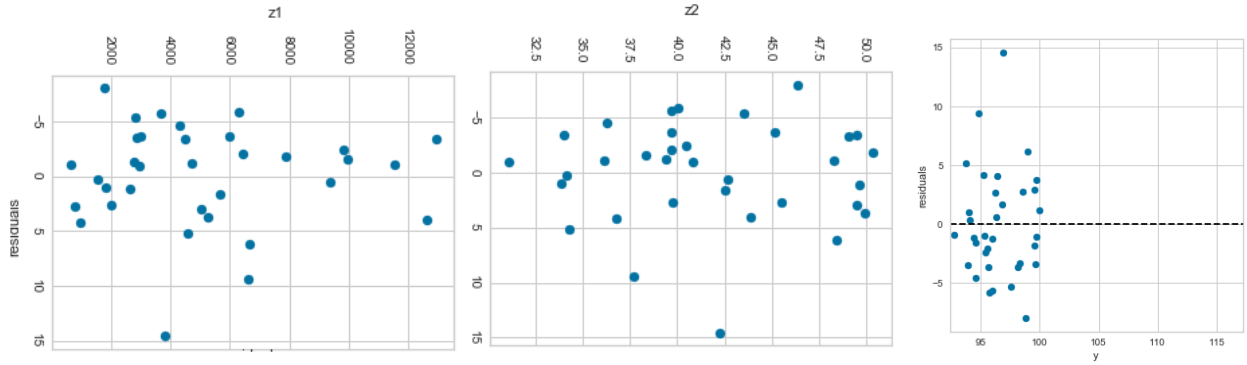

Figura 3.4: a) residuales contra variables (z_1, z_2), b) residuales contra la predicción y_3.

4. Contaminación del aire.

Para los datos de contaminación del aire para los contaminantes $y_1 = NO_2, y_2 = O_3$, se ajustaron modelos de regresión lineal clásico (1.7) y multivariado (1.8) con las variables predictoras relacionadas al clima $z_1 = viento, z_2 = RadiacionSolar$:

$$y_1 = -0.16z_1 + 0.02z_2 + 13.49 \qquad R^2 = 0.023 \tag{1.7}$$

$$[y] = [z]\begin{bmatrix} 5.815 & 2.484 \\ 0.088 & -0.152 \\ 0.041 & 0.095 \end{bmatrix} \tag{1.8}$$

Donde la calidad del ajuste para el modelo (1.7) es muy baja, y las dos variables (z_1, z_2) no fueron significativas (los intervalos de confianza incluyen al 0) para explicar en su conjunto al contaminante $y_1 = NO_2$. Para el caso multivariado (1.8) la variable (z_1) tampoco resultó significativa en el modelo (MANOVA Test al 95%, con Wilks).

Análisis de residuales.

Modelo 1:

Analizando los residuales, estos no muestran simetría o algún patrón por lo que parece que el modelo no necesita de más términos o que no hay errores de cálculo (Observe la Figura 4.1 a), la varianza parece ser constante, se mantiene entre 4 y -4 aunque parece que hay 2 datos atípicos (Observe la Figura 4.1 b), además los residuales también cumplen con el supuesto de normalidad univariada (Shaphiro test, W=0.93).

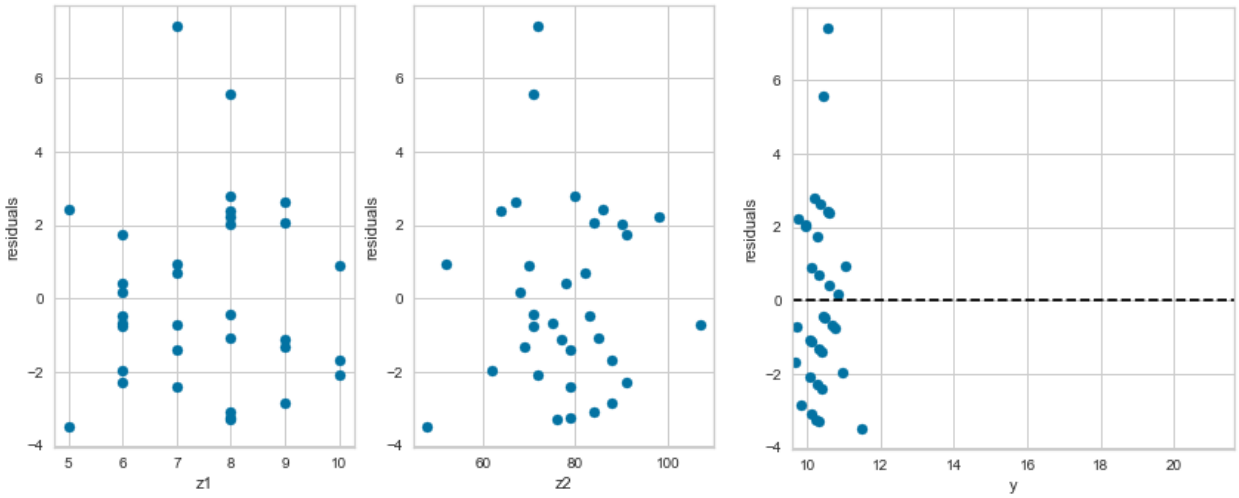

Figura 4.1: a) residuales contra variables (z_1, z_2), b) residuales contra la predicción $\hat{y}_1$.

Modelo 2:

Analizando los residuales, estos no muestran simetría o algún patrón por lo que parece que el modelo no necesita de más términos o que no hay errores de cálculo (Observe la Figura 4.2), la varianza parece ser constante, se mantiene entre [4,-4] y [5,-5] (Observe la Figura 4.3) y además los residuales también cumplen con el supuesto de normalidad multivariada (Figura 4.4)

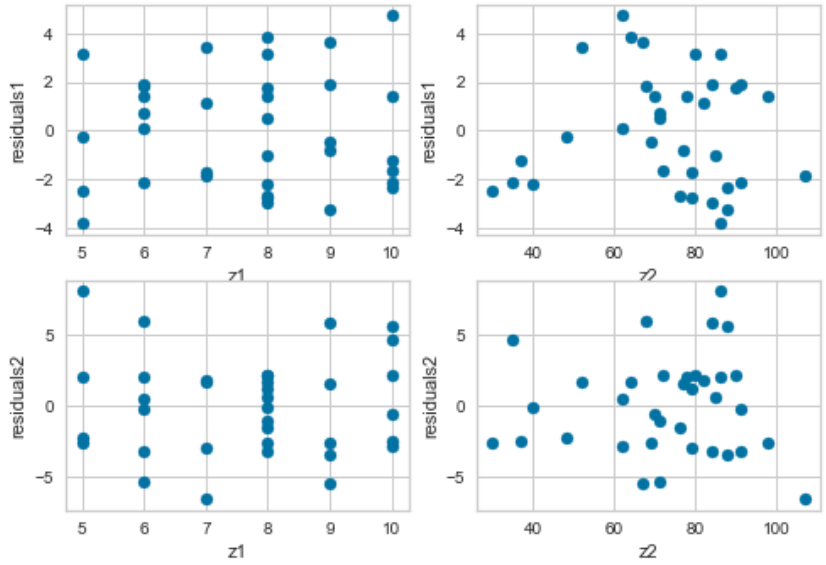

Figura 4.2: residuales contra variables (z_1, z_2)

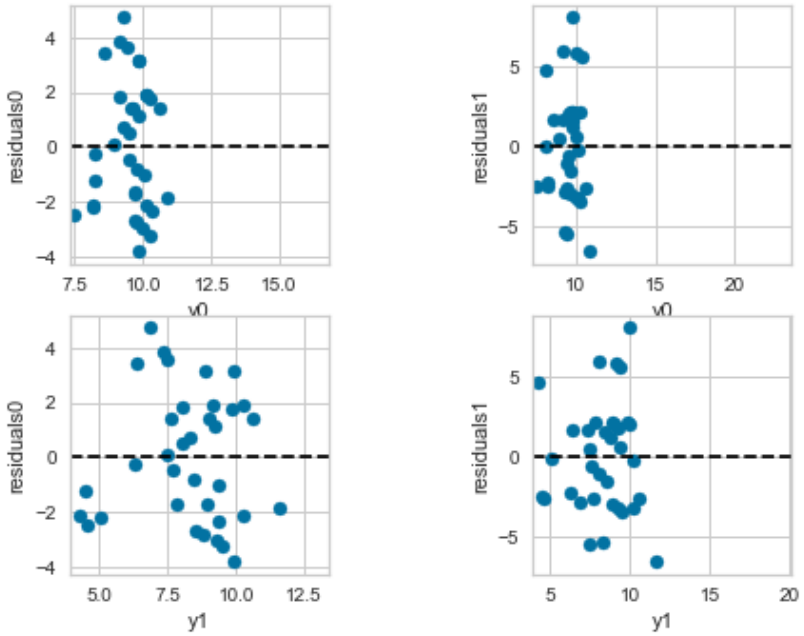

Figura 4.3: residuales contra las predicciones $\hat{y}_i$

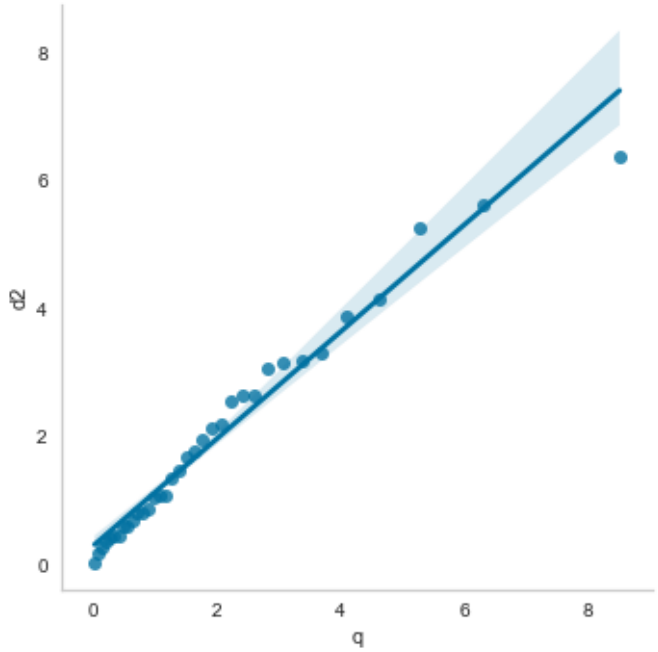

Figura 4.4: Gráfico chi cuadrado. Normalidad multivariada de los residuales, estos se ajustan bien a la línea y también se aplicó la prueba de Looney & Gulledge al 95% para corroborar la hipotesis de normalidad.

Adicionalmente se obtuvieron intervalos de predicción al 95% para $y_1 = NO_2$ y elipse de predicción al 95% para $y_1 = NO_2, y_2 = O_3$ con $z_0 = \begin{bmatrix} 1 & 10 & 80 \end{bmatrix}$, los resultados se muestran en la Figura 4.5, observe que el intervalo de predicción es ligeramente mas angosto que la elipse.

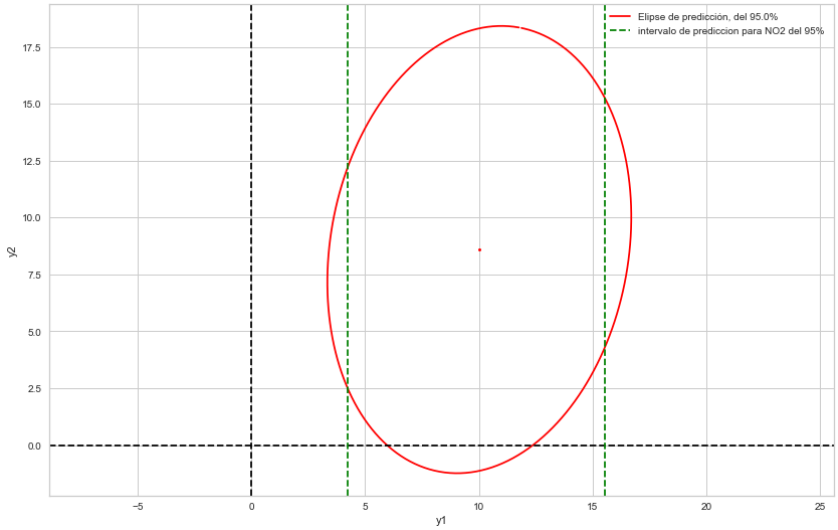

Figura 4.5: Elipse e intervalo de predicción al 95% para $\hat{y}_1, \hat{y}_2$ y $\hat{y}_1$ respectivamente