Análisis de la Relación entre la Calidad del Aire y otros factores medioambientales en Vigo

Antón Soto Martínez y Xoel Trasancos Atadell

November 20, 2024

Contents

1 Introducción y Objetivos

El objetivo de este proyecto es analizar la relación entre la calidad del aire y la precipitación en la ciudad de Vigo, utilizando datos proporcionados por Meteogalicia. Para ello, se han seleccionado cuatro estaciones de medición diferentes: VigoCOIA, VigoCIES, VigoPorto y VigoCampus, ubicadas en distintos puntos de la ciudad.

A lo largo del período de estudio, que abarca de mayo a septiembre de 2024, se recopilarán datos de temperatura, precipitación, humedad, presión y velocidad y dirección del viento, con el fin de evaluar su impacto sobre los niveles de contaminación del aire. La calidad del aire será cuantificada a través de concentraciones de distintos contaminantes, tales como CO_2, NO_2 y SO_2, y se construirá un índice de contaminación del aire (AQI) que permita sintetizar la información de estos contaminantes en una única medida.

1.1 Objetivos Específicos

- Estudiar las variaciones en la calidad del aire en función de las condiciones climáticas, particularmente la precipitación, en distintas zonas de Vigo.
- Estudiar las relaciones entre las diferentes estaciones de medición en Vigo y también detectar las diferencias de temperatura, humedad y precipitación.
- Evaluar cómo la temperatura, la humedad, la presión y la velocidad del viento afectan los niveles de contaminación.
- Crear un índice de contaminación del aire basado en los niveles de CO_2, NO_2 y SO_2, permitiendo una comparación más comprensible de la calidad del aire.
- Determinar si existe una correlación significativa entre los datos meteorológicos y el índice de contaminación, aplicando modelos de regresión y análisis multivariante.
- Identificar que mes está asociado a una mejor calidad del aire.

1.2 Obtención de los datos

Los datos proceden de la web de Meteogalicia. Para Vigo tenemos 4 lugares donde se realizan mediciones de precipitación, humedad y temperatura. Para ver datos sobre el viento y presión atmosferica solo hay dos lugares que son en el puerto de Vigo(VigoPorto) y en las islas Cíes (VigoCies). En Meteogalicia [7] podemos descargar los datos en formato XLS.

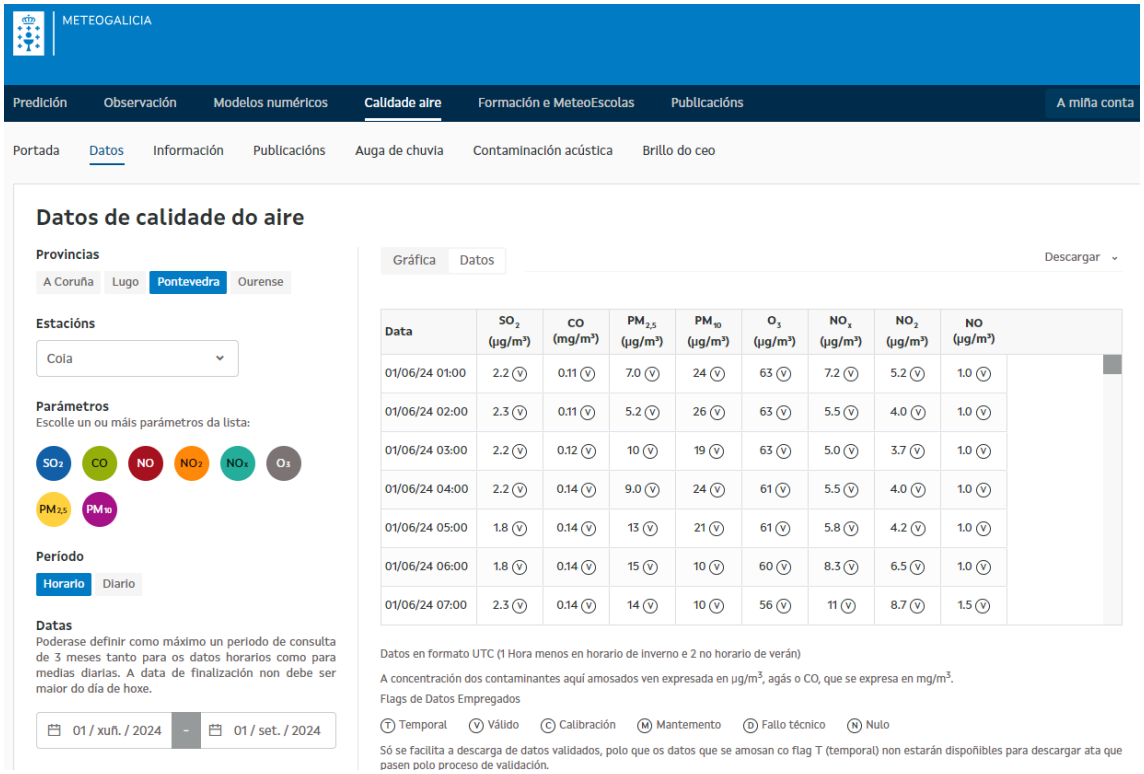

Figure 1: Web Meteogalicia

Después de descargar todos los datos iniciamos el proceso de limpieza de los mismos, que ha sido de las cosas que nos ha llevado más tiempo. Finalmente hemos llegado a este dataset que representamos en forma de tabla:

Day	Rain_coia	Coldday_coia	Humi_coia	Temp_coia	data.AQI_Diario	WindSpeed_porto	Wind_dir_porto	Press_porto	Month
2024-05-01	15.1	0.2	79	10.2	118	12.0	270	1012.9	mayo
2024-05-02	7.0	0.0	77	11.5	112	16.3	225	1015.2	mayo
2024-05-03	11.8	0.0	88	13.3	69	14.8	180	1014.8	mayo
2024-05-04	20.3	0.0	90	14.9	72	16.8	180	1011.9	mayo
2024-05-05	27.2	0.0	92	13.7	85	15.6	225	1010.7	mayo
2024-05-06	0.0	0.0	78	12.8	63	11.1	45	1016.8	mayo

Table 1: Cabecera de los datos meteorológicos selecionados de Vigo

1.3 Problemas encontrados

Para la obtención de los datos que hemos bajado de meteogalicia hemos tenido algún problema que hemos ido solventando durante el trabajo.

Los datos crudos los hemos ido amoldando para poder tener un dataset más accesible y que nos permita trabajar con más flexibilidad. En algunas variables nos hemos encontrado valores raros (-9999) que hemos interpretado como fallos de medición del instrumento y los hemos catalogado como NA. Después con el hemos usado la función na.omit para obtener el dataset omitiendo las observaciones catalogadas como NA. En especial este paso ha sido realmente frustrante dado que habíamos realizado los análisis con los datos mal repetidas veces, y hemos tenido

que volver a empezar. Sin embargo, el hecho de persistir es lo que creemos que nos llevará al éxito tanto en este trabajo como en la vida misma.
Volviendo a los datos que figuran como NA, los ubicamos en:
7 observaciones de la variable Wind.dir.cies (consecutivamente desde el 7 al 13 de Junio).
2 observaciones de la variable Wind.dir.campus y WindSpeed.campus (consecutivamente desde el 17 al 18 de Junio.
Además, todas las variables relacionadas con efectos del Sol (SunnyHours.campus, Insolat.campus, Irradia.campus) el 17 de Junio están mal (-9999) debido a un fallo de los sensores.(No se usan estas variables, tampoco están incluidas en la notación)
2 observaciones en Humi.coia el 28 y 29 de Sept

Al principio del trabajo se pensó en investigar sobre la acción humana en la contaminación del aire, sin embargo por falta de medios y de tiempo se abandonó esta investigación.

Para el estudio de la contaminación del aire hemos obtenido los datos de meteogalicia relativos a la estación de Coia debido a que en las otras 3 estaciones no se recogen los datos de contaminación

2 Modelos Estudiados y Análisis de Resultados

2.1 Modelo de Regresión Lineal Múltiple

2.1.1 Presentación del modelo

Inicialmente, partimos de un modelo de regresión lineal múltiple con la fórmula:

$$Y = \beta X + \varepsilon$$

Definimos el vector de coeficientes β como:

$$\beta = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4, \beta_5)$$

donde cada coeficiente está asociado a una variable continua específica:

$$\begin{aligned}
\beta_1 &= \text{lluvia}, \\
\beta_2 &= \text{temperatura}, \\
\beta_3 &= \text{humedad}, \\
\beta_4 &= \text{velocidad_viento}, \\
\beta_5 &= \text{Presión}.
\end{aligned}$$

$$\text{AQI} \sim \text{lluvia} + \text{temp} + \text{hum} + \text{velo} + \text{press}$$

Siendo el indice de calidad del aire la variable respuesta y las demás las explicativas, todas ellas continuas (en futuros analisis incluiremos categóricas).

Al analizar la significancia de los coeficientes, observamos que solo la variable `temp` (temperatura) resultaba muy significativa en el modelo y la `press` (presión) algo significativa. Además con este modelo solo se consigue explicar alrededor de un 13% de la variabilidad (muy poco). En consecuencia, utilizamos un método de selección de variables basado en eliminación hacia atrás (*backward selection*) para simplificar el modelo, descartando las variables no significativas ([4] pg 15 y 16).
Llegando finalmente a un modelo lineal con 3 variables explicativas (Temperatura, humedad y presión) que afectan significativamente a la variable respuesta (AQI). En particular, la temperatura sigue siendo la más significativa con un p valor del orden de $10^(- 5)$. La presión y la humedad tienen un p valor mayor a 0.04

La interpretación de los coeficientes obtenidos para `temp`,`press` y `hum` muestra que ambos presentaban un efecto negativo en la variable de respuesta (índice de calidad del aire). Esto sugiere que un incremento en la temperatura, presión o humedad relativa estaría asociado con una disminución en el índice, aunque se requiere más análisis para interpretar la relación de manera causal.
Un coeficiente negativo para la temperatura sugiere que a medida que la temperatura aumenta, el AQI (índice de calidad del aire) tiende a disminuir, lo cual indicaría una mejor calidad del aire en días más cálidos, al menos en los datos que hemos analizado.

Para fortalecer el análisis, se construyeron intervalos de confianza para el intercepto y para las pendientes correspondientes a las variables `hum`, `press` y `temp`. Estos intervalos permitieron obtener una estimación de la precisión de los coeficientes y facilitar la interpretación del modelo.

```
> confint(mod.m.3, level=0.95)
                              2.5 %        97.5 %
(Intercept)           170.163352 2677.26121080
Temp_coia              -3.702656   -1.28895390
Humi_coia              -0.669005    0.03449673
Vigo_Porto.Press_porto -2.503610   -0.04343403
```

Figure 2: Intervalos de confianza para los estimadores

De manera similar, un coeficiente negativo para la humedad sugiere que a medida que la humedad aumenta, el AQI tiende a decrecer, lo cual también nos presenta una posible mejora en la calidad del aire en días más húmedos, esto nos parece raro debido a que la lluvia casi no afecta.

Se realizaron contrastes de significación para evaluar la contribución de cada variable significativa en el modelo. Adicionalmente, se generó una representación gráfica de los datos y del ajuste del modelo, en la que se observó que los datos no parecían seguir un modelo lineal de manera clara, lo cual plantea dudas sobre la adecuación del modelo lineal.

2.1.2 Validación y Diagnose

Finalmente, se llevó a cabo la validación y el diagnóstico de los modelos ajustados (`mod.hum`), (`mod.pres`) y (`mod.temp`), que son los modelos lineales simples relativos

a las variables explicativas humedad y temperatura respectivamente. Los resultados de esta fase indicaron varios problemas: los residuos no seguían una distribución normal, no presentaban homocedasticidad y se identificaron valores influyentes que afectaban el ajuste del modelo. Estos hallazgos sugieren que el modelo lineal puede no ser la mejor opción para este conjunto de datos y que podrían considerarse enfoques alternativos o transformaciones adicionales.

Sim embargo, cuando consideramos el modelo en el que intervienen como variables explicativas la temperatura, humedad y presión, podemos observar que tampoco conseguimos mejorar las hipotesis de homocedasticidad y normalidad de los errores. Además, calculando las distancias de Cook para nuestros datos el mayor valor obtenido es de 0.131, lo cual nos sugiere que existen datos influyentes ([5]pg14).

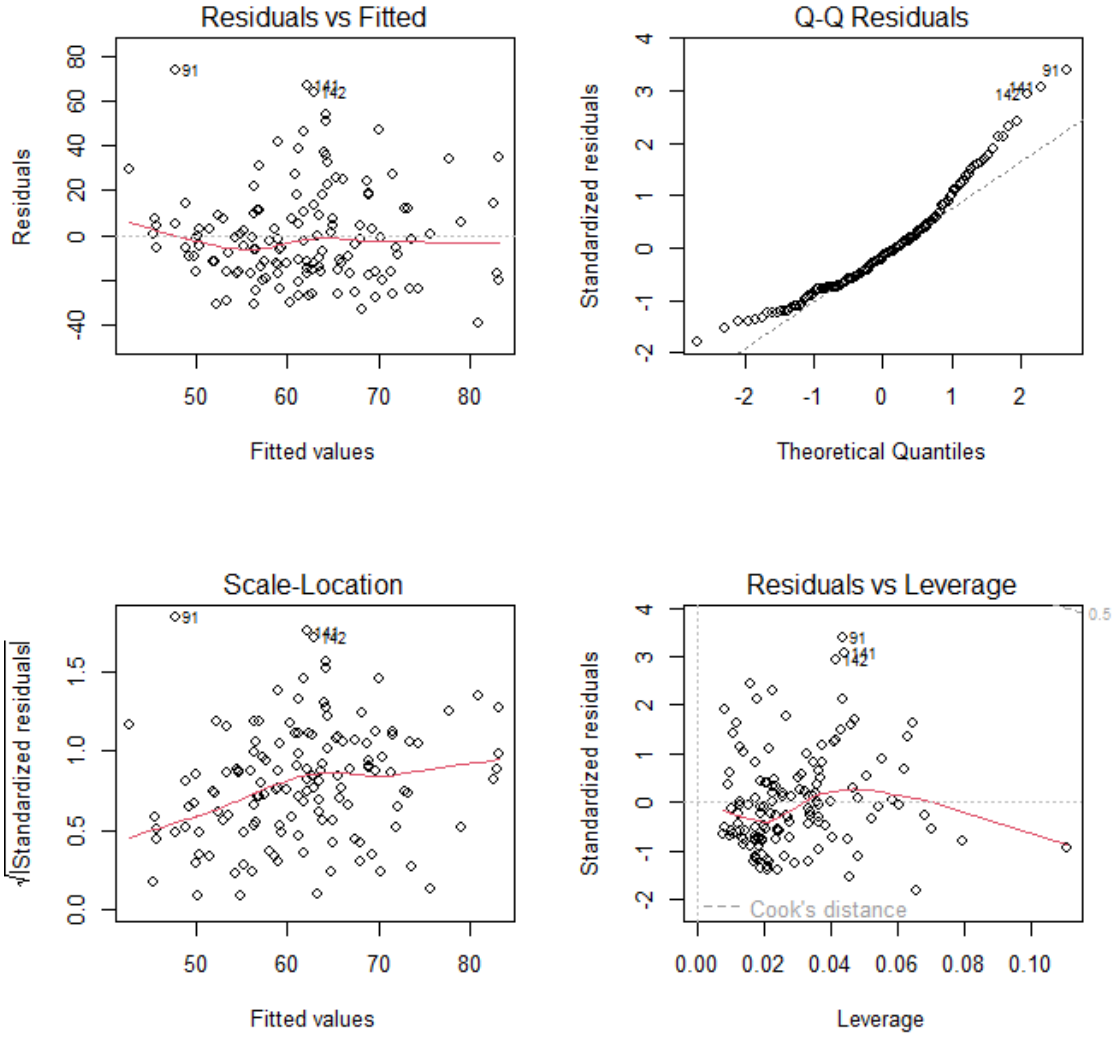

Figure 3: Gráficas para el modelo múltiple

Efectuando el test de Shapiro-Wilk para los residuos studentizaddos y estandarizados llegamos a un pvalor muy bajo por lo tanto rechazamos la normalidad de los residuos. También detectamos la existencia de 6 observaciones atípicas: 88, 91,131, 132, 141 y 142. Creamos un modelo sin atípicos y pasan dos cosas interesantes:

1. Se consigue un R^2 de 0.20 es decir, hemos mejorado la variabilidad explicada.

2. Nuestra variable Presión pasa a ser no significativa con un pvalor de 0.25.

3. Los datos atípicos estaban causando interferencias en el ajuste de nuestro modelo.

2.2 Análisis adicional

Utilizando los datos de las diferentes estaciones de medición de Vigo, vamos a realizar un analisis de los factores meteorológicos en distintas ubicaciones.
Primero empezamos viendo si hay diferencias significativas entre la ubicación Vigo_Campus y Vigo_Porto relativas a la temperatura. Para ello efectuamos un t test. Observamos en la imagen del anexo [6] que el p valor es del orden de 10^{-6} por tanto podemos asegurar que la temperatura del puerto es significativamente distinta a la temperatura del campus.
En el anexo en [7], podemos ver el histograma relativo a las diferencias de temperaturas, así como otros análisis respecto a la humedad y velocidad del viento entre distintas ubicaciones [8, 9, 10, 11].

2.2.1 Conclusiones

En este proyecto hemos llegado a varias conclusiones:
La primera de ellas es que la lluvia no reduce significativamente la contaminación del aire como a priori se podría pensar. Sin embargo después de los estudios realizados hay mejores predictores como pueden ser la temperatura, la presión atmosférica o la humedad relativa.
Además, quitando los valores atípicos del modelo final hemos visto que hemos conseguido aumentar la variabilidad explicada, pese a perder significancia en la presión atmmosférica.
Para nuestro modelo final, las estimaciones de los coeficientes serían las siguientes:

Coeficiente	**Estimación**
(Intercept)	1423.7123
Temp_coia	-2.4958
Humi_coia	-0.3173
Pres_porto	-1.2735222

Table 2: Estimaciones de los coeficientes del modelo

Vemos que para un incremento de la temperatura, la humedad o la presión hacen que disminuya el Índice de calidad del aire.
Finalmente, en la validación y diagnose hemos visto que los datos analizados en nuestro modelo final parecen no seguir una modelo lineal múltiple. Otra conclusión interesante a la que hemos llegado empleando el t test, es que las diferencias entre temperaturas de la zona de montaña (Vigo Campus) y la zona de costa (Vigo Porto) son significativamente distintas. En particular, la temperatura del puerto es significativamente más alta que la del campus durante los meses de Mayo a Septiembre, algo que a priori sería lógico pensar.
Sin embargo, en los analisis que hemos realizado sobre la velocidad del viento en distintas ubicaciones no somos capaces de ver diferencias significativas entre las 4 estaciones. Esto puede llamar la atención debido a que las islas Cíes están a 15 km de la costa de Vigo.

2.3 ANOVA

2.3.1 Presentación del modelo

Considerando los datos disponibles de la variable discreta *Mes*, analicemos nuestros datos siguiendo un modelo ANOVA, cuya parametrización es la siguiente:

$$\begin{cases} Y_{1j} = \mu_1 + \epsilon_{1j}, & j \in 1, \dots, n \\ Y_{ij} = \mu_1 + \alpha_i + \epsilon_{ij}, & i \in 2, \dots, 5, j \in 1, \dots, n \end{cases}$$

donde i se identifica con 1=Mayo, 2=Junio, 3=Julio, 4=Agosto y 5=Septiembre. El μ_1 es la media global del AQI en el mes de Mayo. Los α_i son las desviaciones que experimentan los demás grupos con respecto del grupo de referencia. ([2] pg7 y 8)

Los estimadores de estos parámetros son los siguientes:

$$\begin{cases} \hat{\mu}_1 = \bar{Y}_{1\bullet} = \frac{1}{n} \sum_{j=1}^{n} Y_{1j} \\ \hat{\alpha}_i = \bar{Y}_{i\bullet} - \bar{Y}_{1\bullet} \end{cases}$$

Tendremos que la media global del AQI es de 61.613. Y después para cada mes tenemos las siguientes medias estimadas por el modelo de regresión:

Coeficiente	**Estimación**
Mayo	75.38710
Junio	63.26667
Julio	56.85714
Agosto	44.87097
Septiembre	71.94118

Table 3: medias para cada mes

Los p valores relativos a cada uno de los parametros son muy bajos, excepto para el mes de septiembre.

Lo siguiente es realizar un test F cuya hipotesis nula indica que todas las medias locales son iguales, mientras que la hipotesis alternativa es lo contrario. El p valor obtenido es del orden de 10^{-7} por tanto podemos rechazar la hipotesis nula. Por consiguiente parece razonable estudiar nuestros datos separandolos por mes.

2.3.2 Diagnosis del modelo

Al igual que hicimos en el modelo de regresión lineal múltiple utilizando la distancia de Cook llegamos a un total de 11 observaciones influyentes, más que en el modelo múltiple.
Usando los residuos estudentizados llegamos a 13 observaciones atípicas. Cabe destacar que todos los datos influyentes son atípicos (en este caso, en general no).
Eliminando ahora nuestros datos atípicos, llegamos a un modelo que aumenta hasta 28% su variabilidad explicada. También podemos ver que hay una disminución del pvalor del parámetro relacionado con el mes de septiembre y un incremento del p valor asociado al parámetro del mes de junio.
Podemos observar en los gráficos [12 y 13] del anexo una mejora considerable en el modelo sin atípicos en comparación con los graficos del modelo con atípicos.

Es necesario hacer mención a que los coeficientes del modelo sin los datos atípicos son diferentes al modelo de partida en particular el coeficiente relacionado al mes de septiembre. Veamos cuales son estos nuevos coeficientes:

Coeficiente	**Estimación**
Mayo	72.448
Junio	61.72414
Julio	53.45833
Agosto	43.43333
Septiembre	62.58333

Table 4: Estimaciones de los parametros para el modelo sin atípicos

2.3.3 Validación del modelo

Comencemos estudiando la normalidad de los errores del modelo, para lo cual efectuamos el test de Shapiro-Wilk y, como el p-valor obtenido es muy pequeño, del orden de 10^{-6}, rechazamos la hipótesis de normalidad.
Estudiamos también la homocedasticidad del modelo creando un modelo ANOVA de los residuos absolutos [2] del modelo con respecto a la variable respuesta AQI. Aplicamos ahora el test F para el modelo de los residuos absolutos. En este test, nuestra hipotesis nula será la igualdad de las varianzas respecto a cada mes y la alternativa lo contrario. Nuestro p valor obtenido es 0.001644 por lo que rechazamos la hipótesis de que los errores son homocedasticos.

2.3.4 Conclusión

Una vez que eliminamos los datos atípicos e influyentes y volviendo a ajustar nuestro modelo anova, podemos deducir que cada mes presenta sus propias características respecto a la contaminación del aire. En particular, Agosto es el mes con menos contaminación del aire con unos niveles dentro del umbral catalogado como "BUENO" con una estimación 43.43. El mes con más contaminación es el mes de mayo con diferencia. Esto es plausible ya que anteriormente habíamos analizado que la temperatura hace que el índice decrezca y Agosto popularmente suele ser el mes más cálido.
Sin embargo el modelo ANOVA no cumple con las hipótesis de normalidad y homocedasticidad. Y solo es capaz de explicar un 26% de la variabilidad de los datos.

2.4 ANCOVA

2.4.1 Presentación del modelo

Estudiaremos ahora un modelo más completo utilizando las variables continuas más significativas junto con la variable categórica "MES".
Este modelo será nuestro modelo ANCOVA. Las variables explicativas serán la temperatura y la presión (continuas) por ser las más significativa y el mes (categórica). La variable respuesta sigue siendo el índice de contaminación del aire (AQI). Consideremos inicialmente el modelo más completo posible, un modelo ANCOVA con interacción con nuestras dos variables continuas, parametrizado de la siguiente forma [[9]]:

$$Y_{1j} = \mu_1 + \gamma_1 z_{1j} + \theta_1 w_{1j} + \epsilon_{1j}, \quad j \in \{1, \dots, n\}$$
$$Y_{ij} = \mu_1 + \alpha_i + (\gamma_1 + \delta_i) \cdot z_{ij} + (\theta_1 + \chi_i) \cdot w_{ij} + \epsilon_{ij}, \quad i \in \{2, \dots, 5\} \quad j \in \{1, \dots, n\}$$

donde μ_1 es la media local del mes de mayo, γ_1 es el coeficiente asociado a la temperatura para el mes de mayo y θ_1 es el coeficiente asociado a la presión para el mes de mayo (mayo es el grupo de referencia). α_i es la desviación de cada mes respecto con el mes de referencia, δ_i y χ_i son las desviaciones de los coeficientes correspondientes a cada mes y ϵ_{ij} son los errores correspondientes siguiendo un distribución normal de media igual a 0 y varianza igual a σ^2. ([3] pg 9, 10 y 11)

Partimos de dos modelos:
El primero sin iteración cuyas variables explicativas son la temperatura, presión y el mes, y el segundo con iteración entre las tres variables.

```
mod_t_p_m <- lm(AQI  ~ Temp_coia +Vigo_Porto.Press_porto +Month, data=Newdata )
mod_t_p_m_it <- lm(AQI  ~ Temp_coia*Vigo_Porto.Press_porto*Month, data=Newdata)
```

Figure 4: Modelo con y sin iteración

Aplicando la funcion "summary" en R podemos ver que no hay ninguna iteración significativa, por consiguiente podríamos pensar en que el modelo sin iteración real-

iza un mejor ajuste a nuestro datos ya que la combinación de las variables explicativas no tiene un efecto adicional en la variable respuesta.

Realizando el test F entre el modelo con y sin iteración obtenemos un p valor de 0.36, por tanto no tenemos pruebas para rechazar la hipotesis nula que nos dice que consideremos el modelo más simple. Para el R^2 si que vemos que el modelo con iteración nos da cerca del 30% mucho más que el modelo sin iteración.

2.4.2 Diagnosis del modelo

Como hicimos en el modelo múltiple y en el Anova vemos los datos con capacidad de influencia según sus leverages (apalancamientos). Tenemos un total de 17 observaciones con capacidad de influencia para el modelo con iteración.Calculando la distancia de Cook para la obtención de los datos influyentes observamos que su máximo toma el valor de 1.22, por lo tanto parece claro que los datos con capacidad de influencia son también influyentes.
A continuación, vemos cuantos de nuestros datos son atípicos. Tenemos 15 observaciones atípicas para nuestro modelo con iteración. Ahora tratamos de eliminar las observaciones atípicas de nuestro modelo. Una vez eliminados los datos atípicos comparamos los modelos (ambos con iteración). El modelo sin atípicos alcanza un R^2 de 0.38, significativamente más alto que el modelo con atípicos.

2.4.3 Validación del modelo

Para validar el modelo ANCOVA, como el modelo que hemos escogido es con iteracción y sin atípicos, debemos hacer una validación marginal por cada mes, creando cinco modelos de regresión lineal múltiple, uno para cada mes.

- **Mayo**:
  Estudiamos si los errores proceden de una distribución normal con el test de Shaphiro Wilk. El p valor obtenido es de 0.07 por tanto podemos asumir que los errores en el mes de mayo no siguen una distribución normal. Utilizamos el test de Breusch-Pagan para ver si los errores son homocedásticos, y en efecto sí lo son ya que aplicando el bptest llegamos a un p valor de 0.84. Concluímos que en el mes de mayo los errores tienen una varianza constante.
  Aplicamos el test de Ramsey y nos da un p valor de 0.009 , por tanto rechazamos que los datos sigan un modelo lineal. Algo que ya se intuía en el gráfico de "Residuals vs fitted" del modelo del mes de mayo.

- **Junio**: Estudiamos si los errores proceden de una distribución normal con el test de Shaphiro Wilk. El p valor obtenido es de 0.54 por tanto podemos asumir que los errores en el mes de junio siguen una distribución normal. Utilizamos el test de Breusch-Pagan para ver si los errores son homocedásticos, y en efecto no lo son ya que aplicando el bptest llegamos a un p valor de 0.039. Concluímos que en el mes de junio los errores no tienen una varianza

constante.
Aplicamos el test de Ramsey y nos da un p valor de 0.34, por consiguiente los datos siguen un modelo lineal. Además, cuando vemos el gráfico de "Residuals vs fitted" del modelo de junio parece claro que los datos siguen un modelo lineal.

- **Julio**: Estudiamos si los errores proceden de una distribución normal con el test de Shaphiro Wilk. El p valor obtenido es de 0.39 por tanto podemos asumir que los errores en el mes de julio siguen una distribución normal. Utilizamos el test de Breusch-Pagan para ver si los errores son homocedásticos, y en efecto sí lo son ya que aplicando el bptest llegamos a un p valor de 0.47. Concluímos que en el mes de julio los errores tienen una varianza constante.
  Aplicamos el test de Ramsey y nos da un p valor de 0.16 por tanto los datos siguen un modelo lineal. Sin embargo, cuando vemos el gráfico de "Residuals vs fitted" del modelo de julio no parece claro que los datos sigan un modelo lineal.

- **Agosto**: Estudiamos si los errores proceden de una distribución normal con el test de Shaphiro Wilk. El p valor obtenido es de 0.0046 por tanto podemos asumir que los errores en el mes de agosto no siguen una distribución normal. Utilizamos el test de Breusch-Pagan para ver si los errores son homocedásticos, y en efecto sí lo son ya que aplicando el bptest llegamos a un p valor de 0.43. Concluímos que en el mes de agosto los errores tienen una varianza constante. Aplicamos el test de Ramsey y nos da un p valor de 0.18 que no es tan bajo como para rechazar que los datos tengan un comportamiento lineal. Sin embargo, cuando vemos el gráfico de "Residuals vs fitted" del modelo de agosto parece claro que los datos no siguen un modelo lineal.

- **Septiembre**: Estudiamos si los errores proceden de una distribución normal con el test de Shaphiro Wilk. El p valor obtenido es de 0.19 por tanto podemos asumir que los errores en el mes de agosto siguen una distribución normal. Utilizamos el test de Breusch-Pagan para ver si los errores son homocedásticos, y en efecto sí lo son ya que aplicando el bptest llegamos a un p valor de 0.60. Concluímos que en el mes de agosto los errores tienen una varianza constante. Aplicamos el test de Ramsey y nos da un p valor de 0.59 entonces los datos siguen un comportamiento lineal. Sin embargo, en el mes de septiembre nos surgen varios problemas debido a que es el mes que más atípicos tiene y las gráficas no nos muestran el comportamiento lineal de los datos de un xeito adecuado.

2.4.4 Conclusión

Una vez eliminados los datos atípicos podemos concluir que el modelo sin iteración es más adecuado para nuestros datos. Esto es algo plausible debido a que la no

dependencia de las variables explicativas.
En primer lugar, no es descabellado pensar que la temperatura y la presión no están relacionadas. En segundo lugar, es cierto que la temperatura puede depender del mes, sin embargo, en nuestro estudio hemos cogido meses próximos y que comparten la estación de verano. Por tanto, parece lógico que no haya relación entre temperatura y mes por el abanico de meses seleccionado.
Finalmente, a pesar que el modelo sin atípicos con iteración nos da un mejor R^2, sometiendo los dos modelos a la función anova. El p valor es de 0.14, lo que nos sugiere que el modelo más simple se ajusta mejor a los datos estudiados.
Para el modelo ANCOVA sin iteración y sin atípicos nos quedan las siguientes estimaciones en los coeficientes:

3 Conclusión

Para finalizar, una vez analizado cada modelo por separado hemos llegado a la conclusión que el modelo que mejor ajusta nuestros datos es el modelo ANOVA, es decir bastaría saber el mes en el que nos encontramos para ajustar los datos y predecir nuevas observaciones. Este hecho, al principio nos causó bastante dolor de cabeza debido a que no veíamos posible que el modelo ANOVA fuese mejor que el ANCOVA en nuestro caso. Sin embargo, llegamos a la conclusión siguiente: Las variables de temperatura y presión están de alguna manera contenidas implícitamente en la variable mes.
El R^2 ajustado para el modelo ANCOVA es más alto que para el ANOVA. Pero esto es debido a que el ANCOVA tienen muchas más variables e iteraciones y por eso es lógico aumentar la variabilidad. A pesar de eso, seguimos prefiriendo el modelo ANOVA debido a su significancia, simplicidad.

4 Notación

Para el siguiente proyecto vamos a utilizar la siguiente notación:
AQI := Índice de calidad del aire
Day := Día en el que se tomó la observación
Rain_x := Lluvia en la ubicación x medidos en L/m^2
Temp_x := Temperatura en la ubicación x en °C
Humi_x := Porcentaje de Humedad relativa en la ubicación x
WindSpeed_x := Velocidad del viento en la ubicación x en m/s
ColdDay_x := Horas de frío (¡7 ºC)
Wind_dir_x := Dirección del viento en la ubicación x en º (de $0^º$ a $360^º$)
Press_x := Presión atmosférica registrada en la ubicación x en milibares
Month := Mes (De mayo a septiembre)
La ubicación x podrá ser alguna de estas 4 : (Coia, Porto, Cies, Campus)

En el documento evitaremos emplear unidades para facilitar la lectura y gestionar el espacio de una manera más eficiente.

5 Bibliografía

References

[1] AirNow. *Technical Assistance Document for the Reporting of Daily Air Quality*, 2023.

[2] Rosa M. Crujeiras and Manuel Febrero. Análisis de la varianza, 2024. Apuntes de clase, Tema 4: Asignatura de Modelos de Regresión y Análisis Multivariante, Grado en Matemáticas, USC.

[3] Rosa M. Crujeiras and Manuel Febrero. Análisis de la varianza, 2024. Apuntes de clase, Tema 5: Asignatura de Modelos de Regresión y Análisis Multivariante, Grado en Matemáticas, USC.

[4] Rosa M. Crujeiras and Manuel Febrero. Construcción de un modelo de regresión, 2024. Apuntes de clase, Tema 3: Asignatura de Modelos de Regresión y Análisis Multivariante, Grado en Matemáticas, USC.

[5] Rosa M. Crujeiras and Manuel Febrero. Diagnosis de observaciones atípicas o influyentes, 2024. Apuntes de clase, Tema 2 :Asignatura de Modelos de Regresión y Análisis Multivariante, Grado en Matemáticas, USC.

[6] Rosa M. Crujeiras and Manuel Febrero. El modelo lineal general, 2024. Tema 1: Apuntes de clase, Asignatura de Modelos de Regresión y Análisis Multivariante, Grado en Matemáticas, USC.

[7] Meteogalicia. Web meteogalicia.

[8] Ministerio para la Transición Ecológica y el Reto Demográfico. Óxidos de nitrógeno y calidad del aire, 2023. Consultado el 6 de noviembre de 2024.

[9] Douglas C. Montgomery, Elizabeth A. Peck, and G. Geoffrey Vining. *Introduction to Linear Regression Analysis*. Wiley, 2021.

[10] Wikipedia. Navaja de ockham.

A Anexo

En este anexo vamos a incluir gran parte del trabajo realizado. Lo primero será explicar la creación del índice de calidad del aire.

A.1 Creación del Índice de Calidad del Aire

A partir de los datos recogidos de los gases contaminantes en la estación de medición de Vigo_Coia, hemos obtenido el Índice de Calidad del Aire, también conocido como AQI (Air Quality Index). Este índice proporciona una medida estandarizada de la calidad del aire, permitiendo evaluar el nivel de contaminación en una escala comprensible.

Para el cálculo del AQI, utilizamos diferentes fuentes de información, entre ellas el [1] *Technical Assistance Document for the Reporting of Daily Air Quality*(`https://document.airnow.gov/technical-assistance-document-for-the-reporting-of-daily-air-qu pdf`) y datos proporcionados por Meteogalicia. Estas fuentes nos proporcionaron los rangos y puntos de corte necesarios para estandarizar los valores de concentración de cada contaminante, adaptando el cálculo del AQI al contexto de la estación de Vigo_Coia.

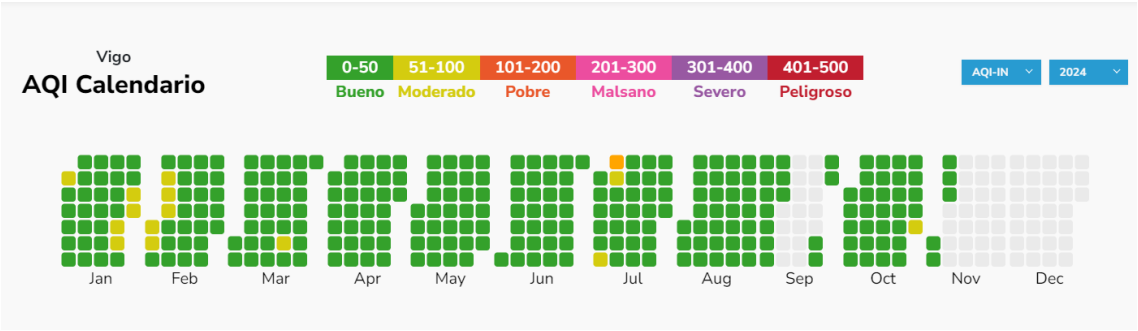

Figure 5: Indicadores del AQI

El proceso de cálculo del AQI siguió los pasos detallados, que se muestran en el fragmento de código adjunto en el script de Meteodata.R en el apartado de "Creación del índice (AQI)". Cada paso incluye la clasificación de los contaminantes y su conversión en el índice de calidad del aire mediante interpolación entre los valores de referencia, siguiendo los estándares establecidos.

B Gráficos e ilustraciones

```
> t.test( Temp_campus, Temp_porto,alternative = "two.sided", var.equal = TRUE) #pvalue muy bajo

        Two Sample t-test

data:  Temp_campus and Temp_porto
t = -4.5525, df = 304, p-value = 7.679e-06
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.640768 -1.046813
sample estimates:
mean of x mean of y
 17.30458  19.14837
```

Figure 6: t test entre temperaturas

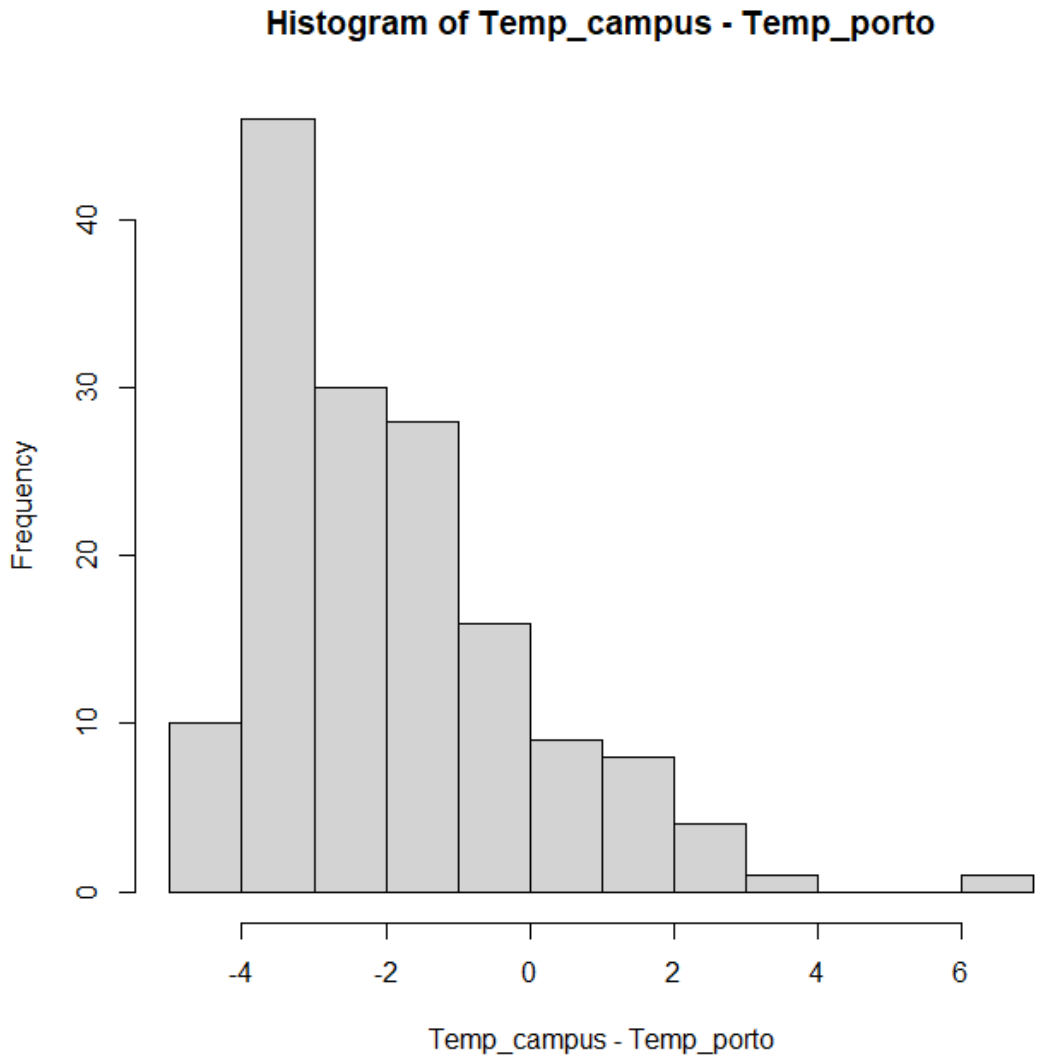

Figure 7: Histograma de diferencias

```
> t.test( Rain_campus, Rain_porto,alternative = "two.sided", var.equal = TRUE)

        Two Sample t-test

data:  Rain_campus and Rain_porto
t = 0.99477, df = 304, p-value = 0.3206
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6859746  2.0885890
sample estimates:
mean of x mean of y
 2.761438  2.060131
```

Figure 9: ttest entre la precipitación del campus y puerto

```
> t.test(Temp_cies, Temp_porto)

        Welch Two Sample t-test

data:  Temp_cies and Temp_porto
t = -3.9848, df = 300.51, p-value = 8.481e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.8267888 -0.6189629
sample estimates:
mean of x mean of y
 17.92549  19.14837
```

Figure 10: ttest entre temperaturas de las islas cíes y el puerto

```
> t.test( Humi_campus, Humi_porto,alternative = "two.sided", var.equal = TRUE)

        Two Sample t-test

data:  Humi_campus and Humi_porto
t = 0.71359, df = 304, p-value = 0.476
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.941425  4.150576
sample estimates:
mean of x mean of y
 78.64706  77.54248
```

Figure 8: ttest entre humedades de Coia y Puerto

```
> t.test(WindSpeed_cies, WindSpeed_porto)

        Welch Two Sample t-test

data:  WindSpeed_cies and WindSpeed_porto
t = -1.2555, df = 272.85, p-value = 0.2104
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -1.6404464  0.3628547
sample estimates:
mean of x mean of y
 11.30719  11.94599
```

Figure 11: ttest velocidad viento en Cíes y Porto

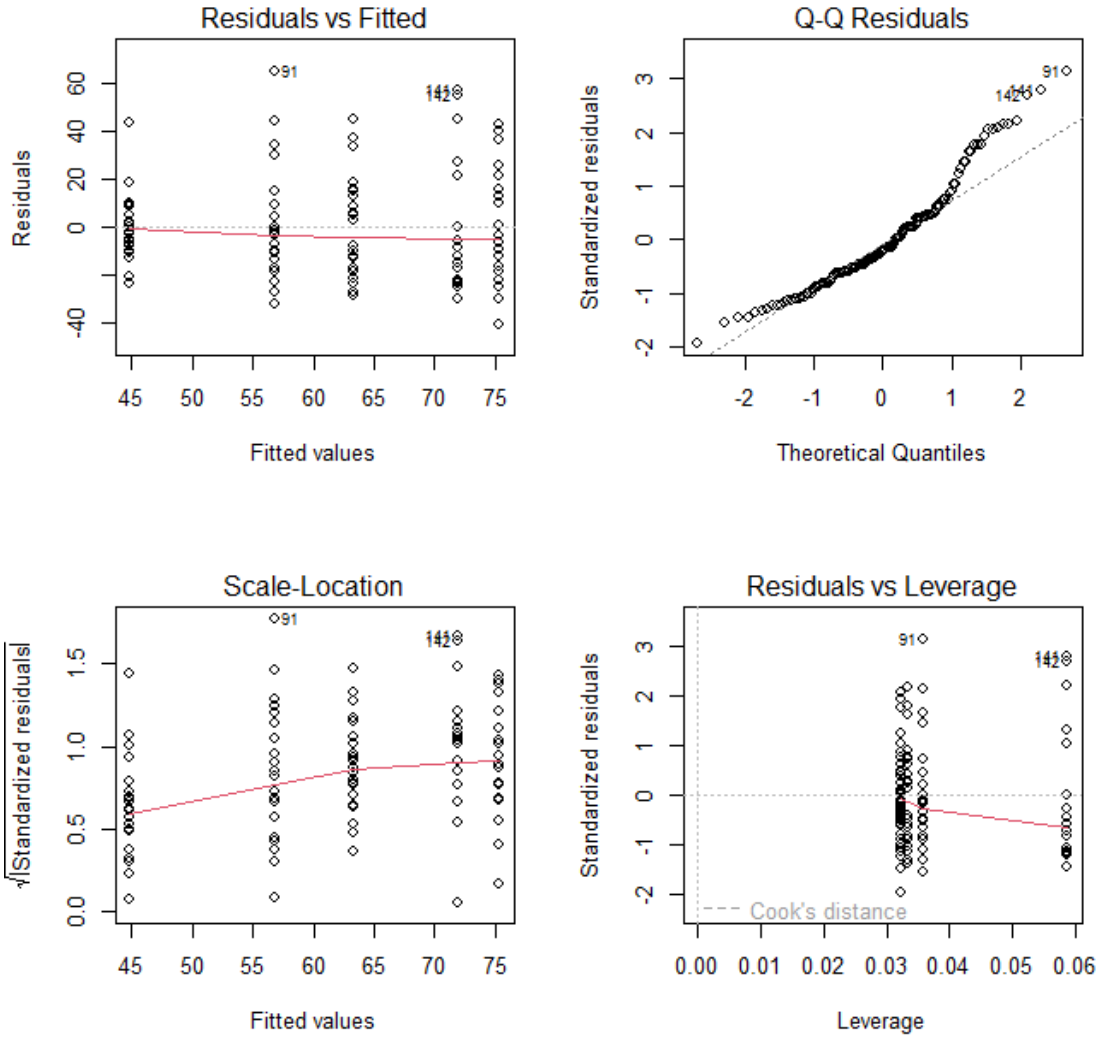

Figure 12: Modelo con atípicos

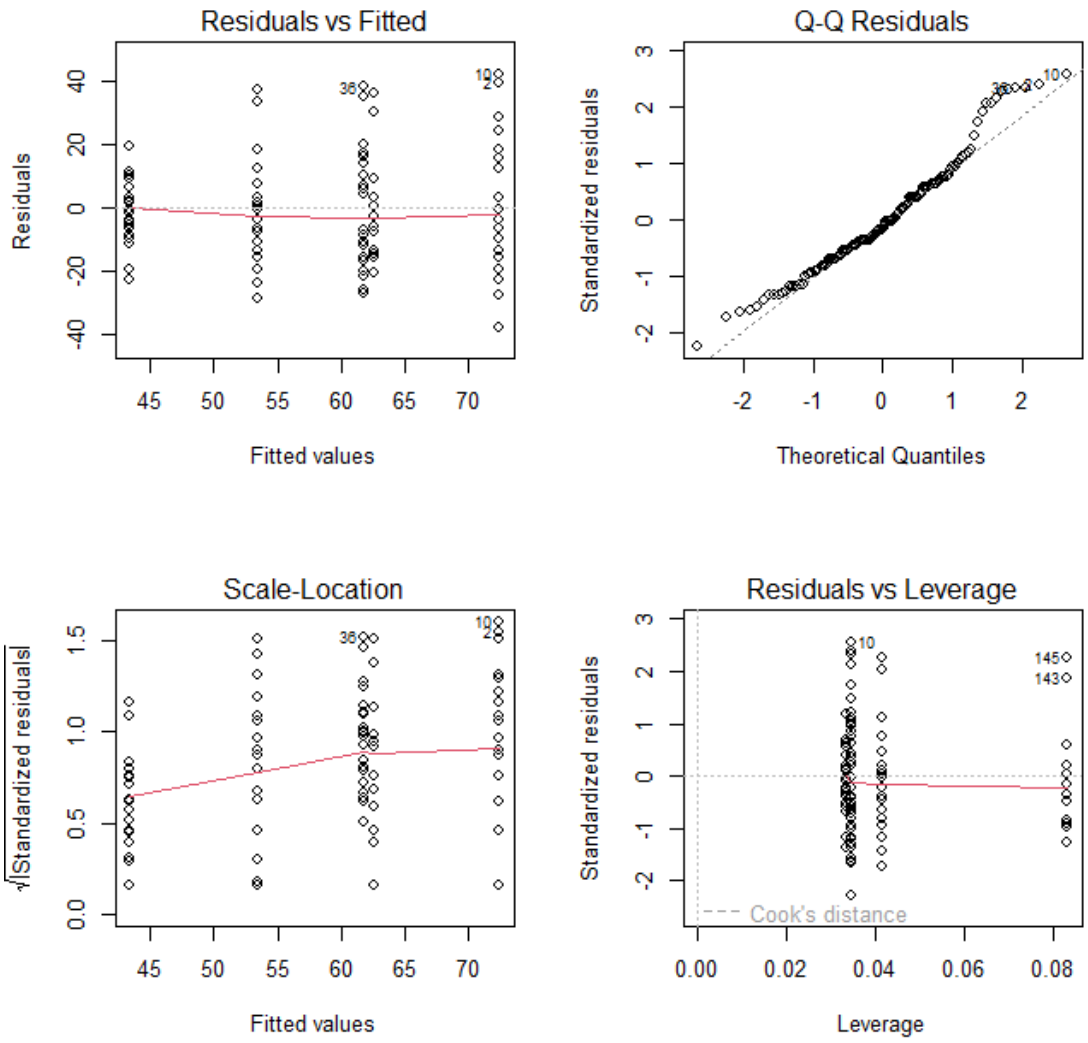

Figure 13: Modelo sin datos atípicos