Segundo Proyecto Estadística. Curso 2019-2020

Daniel Alberto García Pérez D.GARCIA@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Leonel Alejandro García López L.GARCIA3@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Roberto Marti Cedeño R.MARTI@ESTUDIANTES.MATCOM.UH.CU
Grupo C412

Tutor(es):
Msc. Dalia Diaz Sistachs, *Facultad de Matemática y Computación, Universidad de La Habana*

Tema: Estadística, Técnicas de Clasificación, Regresión, Anova.

1. Introducción

El siguiente informe corresponde al trabajo de los autores como parte de la investigación realizada sobre los datos asignados en su segundo proyecto de la asignatura.

Los datos asignados, responden a un estudio realizado sobre las respuestas correspondientes a un sensor de gases en una ciudad italiana (Data/AirQualityUCI.csv). Se tuvieron en cuenta las respuestas de los distintos terminales del sensor, así como las concentraciones de gases existentes en el ambiente.

Por razones desconocidas, existen observaciones incompletas de cada una de las variables presentes en la recopilación, por lo que se hace necesario modificar las mismas para poder realizar un estudio acorde con los requerimientos de cada uno de los métodos a emplear.

1.1 Descripción Inicial

Para la descripción de las características generales de los datos, se empleó el método *skim* presente en la biblioteca *skimr* de r, el cual brinda, entre sus valores principales, la cantidad de datos faltantes, así como los estadísticos descriptivos de cada una de las variables de la muestra.

```
-- Data Summary ------------------------
                           Values
Name                       step1
Number of rows             9357
Number of columns          15
___________________________
Column type frequency:
  factor                   2
  numeric                  13
___________________________
Group variables            None

-- variable type: factor ---------------------------------------------------------------------
# A tibble: 2 x 6
  skim_variable n_missing complete_rate ordered n_unique top_counts
* <chr>             <int>         <dbl> <lgl>      <int> <chr>
1 Date                  0             1 FALSE        391 01/: 24, 01/: 24, 01/: 24, 01/: 24
2 Time                  0             1 FALSE         24 00.: 390, 01.: 390, 02.: 390, 03.: 390
```

Figure 1: Descripción de la fuente de datos mediante la función skim (Parte 1).

Como se puede apreciar de la primera mitad de los datos obtenidos de la función (Figura 1), el set de datos se compone por 9357 observaciones de 15 variables. Estas se componen por 2 de tipo factor y 13 numéricas. Es importante destacar que dado que las dos variables de tipo factor, la fecha y la hora de las mediciones, se descartaron por el equipo para el análisis dado que, las mediciones se realizaron exclusivamente durante tres meses, por lo cual, no se cuenta con información suficiente para caracterizar el resto de los resultados a partir del tiempo.

```
-- variable type: numeric --------------------------------------------------------------------------------------------------------
# A tibble: 13 x 11
   skim_variable n_missing complete_rate    mean      sd      p0     p25     p50     p75    p100 hist
 * <chr>           <int>         <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl>   <dbl> <chr>
 1 CO.GT.           1683         0.820    2.15    1.45    0.1     1.1     1.8     2.9     11.9  ▇▂▁▁▁
 2 PT08.S1.CO.       366         0.961 1100.    217.    647     937    1063    1231    2040    ▁▆▇▃▁
 3 NMHC.GT.         8443         0.0977 219.    204.      7      67     150     297    1189    ▇▂▁▁▁
 4 C6H6.GT.          366         0.961   10.1    7.45    0.1     4.4     8.2    14      63.7  ▇▃▁▁▁
 5 PT08.S2.NMHC.     366         0.961  939.   267.    383     734.    909    1116    2214    ▂▇▅▁▁
 6 NOx.GT.          1639         0.825  247.   213.      2      98     180     326    1479    ▇▂▁▁▁
 7 PT08.S3.NOx.      366         0.961  835.   257.    322     658     806     970.    2683  ▂▇▃▁▁
 8 NO2.GT.          1642         0.825  113.    48.4     2      78     109     142     340   ▁▆▇▂▁
 9 PT08.S4.NO2.      366         0.961 1456.   346.    551    1227    1463    1674    2775    ▁▃▇▆▁
10 PT08.S5.O3.       366         0.961 1023.   398.    221     732.    963    1274.   2523    ▃▇▅▂▁
11 T                 366         0.961   18.3    8.83   -1.9    11.8    17.8    24.4    44.6  ▁▆▇▃▁
12 RH                366         0.961   49.2   17.3     9.2    35.8    49.6    62.5    88.7  ▁▅▇▆▂
13 AH                366         0.961    1.03   0.404   0.185   0.737   0.995   1.31    2.23  ▂▇▅▂▁
```

Figure 2: Descripción de la fuente de datos mediante la función skim (Parte 2).

Centrando el análisis en la segunda parte de la respuesta obtenida de *skim* (Figura 2), resalta la cantidad de observaciones faltantes en cada una de las variables, que varían desde 366, hasta 8443. Las variables presentes, son de forma general de 3 tipos, las respuestas de los sensores a determinados compuestos del aire, la concentración de los compuestos presente, y variables generales del ambiente, temperatura, humedad relativa y absoluta.

Para un mejor empleo de los datos, se completaron los datos faltantes con la media de cada una de las variables descritas en los datos.

2. Regresión y ACP

El primer análisis a realizar sobre los datos fue la regresión, para ello se tuvo en cuenta la correlación existente entre cada una de las variables numéricas para corroborar si sería de utilidad realizarla.

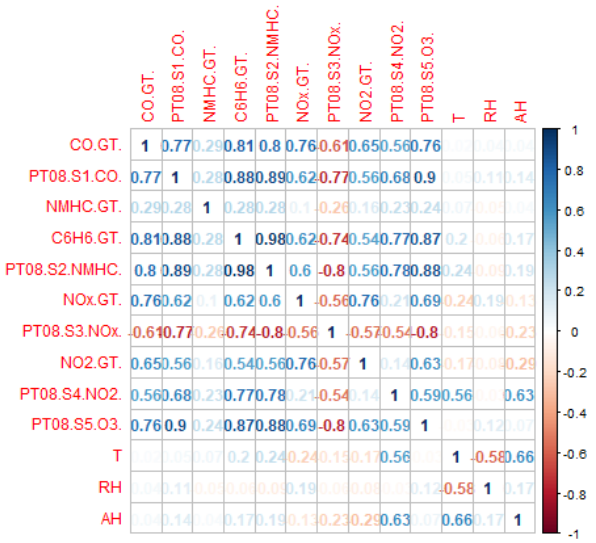


Figure 3: Correlación entre las variables.

La primera característica presente en los datos que se observa a partir de su correlación (Figura 3) es que los datos de los compuestos están altamente correlacionados, por lo que la realización de una regresión lineal sobre los mismos incurriría en el problema de la multicolinealidad.

Como estrategia de solución al problema anterior, se decidió agrupar los datos por sus componentes principales, para así, reducir los datos de ser posible y poder dar una interpretación mas clara a la regresión. Como no se dispone de interés especial por alguna variable se decidió empelar la variable de la respuesta del sensor número 4 relacionado con el dióxido de nitrógeno ($PT08.S4.NO_2$) debido a que presenta pocos datos faltantes y es las mas correlacionada con los datos disponibles.

2.1 ACP

A continuación (Figura 4) se encuentran todas las posibles componentes resultantes del desgloce (Buscar como se escribe) de los datos.

```
Importance of components:
                          PC1    PC2    PC3    PC4    PC5    PC6    PC7    PC8
Standard deviation     2.4959 1.4415 1.1395 0.9612 0.71574 0.63999 0.40913 0.3763
Proportion of Variance 0.5191 0.1732 0.1082 0.0770 0.04269 0.03413 0.01395 0.0118
Cumulative Proportion  0.5191 0.6923 0.8005 0.8775 0.92018 0.95431 0.96826 0.9801
                           PC9    PC10    PC11    PC12
Standard deviation     0.32477 0.29453 0.19141 0.10205
Proportion of Variance 0.00879 0.00723 0.00305 0.00087
Cumulative Proportion  0.98885 0.99608 0.99913 1.00000
```

Figure 4: Posibles componentes principales.

Se tuvieron en cuenta dos criterios tomados de la literatura para la selección de las componentes principales, un criterio de *porcentaje* que debería superar como mínimo el 70% de los datos y el criterio de *Kaiser*. Como podemos observar en las componentes (Figura 4) con las dos primeras ya se cumple el criterio de mas del 70% de los datos, pero para cumplir también con *Kaiser* se extendieron las componentes hasta la tercera.

2.1.1 Descripción de las componentes

Para poder describir detalladamente las características de cada una de las tres componentes en las que se aglomeran los datos de estudio se analizó su matriz de valores propios. (Figura 5)

	PC1	PC2	PC3
CO.GT.	-0.35347648	-0.04638393	0.03504330
PT08.S1.CO.	-0.37073311	0.02593933	-0.10493234
NMHC.GT.	-0.12883560	0.09317803	0.08870589
C6H6.GT.	-0.37361310	0.12266320	0.01402562
PT08.S2.NMHC.	-0.37801424	0.14739559	0.02693871
NOx.GT.	-0.31179657	-0.26951354	0.01283753
PT08.S3.NOx.	0.33774833	-0.09221096	0.09989306
NO2.GT.	-0.29045629	-0.24388495	0.30649399
PT08.S5.O3.	-0.37473906	-0.03834792	-0.08045593
T	-0.02533438	0.66090296	0.14332237
RH	-0.01688014	-0.32692259	-0.75967656
AH	-0.03897675	0.51571297	-0.52035094

Figure 5: Valores propios de las componentes principales.

La primera componente se caracteriza por bajos valores de la concentración de todos los compuestos presentes en el estudio con excepción los hidrocarburos no metánicos ($NMHC.GT$), así como bajos valores de respuesta de todos los sensores, exceptuando a la variable dependiente. Esta componente, la mas numerosa de las analizadas describe el comportamiento mas común de los datos, este resultado puede estar determinado por que las muestras se obtuvieron en una sola zona de la ciudad, en una sola ciudad o en un intervalo de tiempo donde no varían mucho.

La segunda componente se caracteriza por valores altos de la temperatura y humedad relativa. Esta componente describe las situaciones de las horas cercanas al mediodía donde la temperatura es mas elevada.

La última componente se caracteriza por altos valores de la humedad relativa y absoluta. Dada la presencia de altos valores de humedad tanto en la 2da como en la 3ra componente se puede llegar a la conclusión que algunas de las mediciones se encontraron en temporada de lluvias u ocurrió algún evento climatológico.

2.2 Regresión

Posterior a la definición de las componentes se dispuso la creación de un modelo de regresión múltiple, donde la variable dependiente se tomó como la respuesta del cuarto sensor a la concentración de NO_2, y como variables independientes las 3 componentes resultantes del

ACP. Se comprobó una vez mas la utilidad de realizar la regresión mediante los gráficos de dispersión y correlación (Figuras 6 y 7).

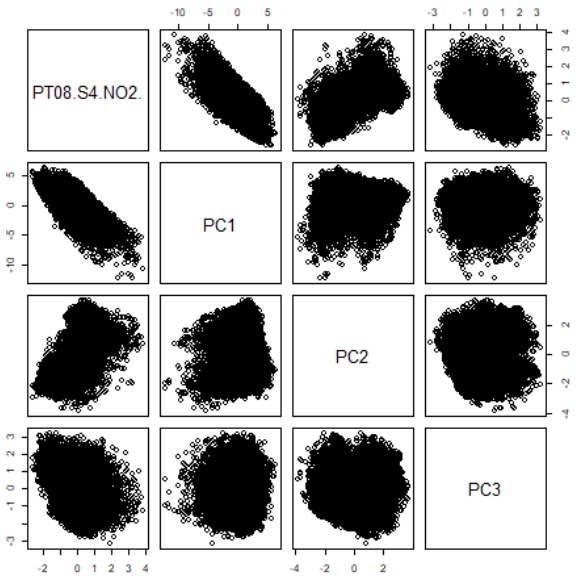


Figure 6: Gráfico de dispersión.

Del gráfico de dispersión (Figura 6), podemos percatarnos de la correlación inversa de la variable dependiente con la primera componente, así como su relacion débil pero lineal con la segunda. Ambos datos se esclarecen con la gráfica de correlación (Figura 7).

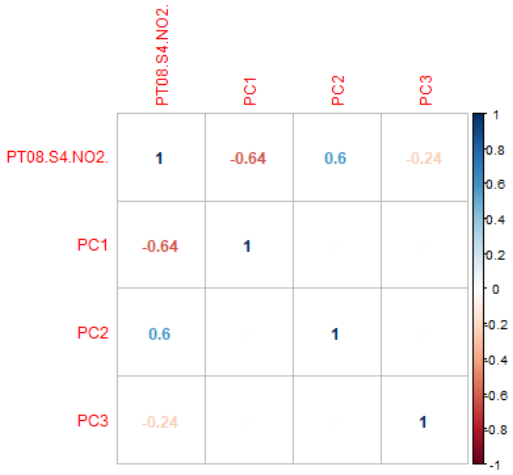


Figure 7: Gráfico de correlación.

Otro detalle significativo se deduce de los resultados de la clasificación en componentes principales, dado que brindan una segmentación en variables independientes, como se muestra en la figura 7. Finalmente antes de la realización de la regresión se tomó como consenso la admisión de los valores de las correlaciones de la primera y segunda componentes con la variable dependiente como lineales.

```
Call:
lm(formula = formula, data = as.data.frame(dataset))

Residuals:
    Min      1Q  Median      3Q     Max
-2.0099 -0.3096 -0.1014  0.3036  1.9046

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)  3.646e-16  4.205e-03    0.00        1
PC1         -2.580e-01  1.685e-03 -153.09   <2e-16 ***
PC2          4.165e-01  2.917e-03  142.76   <2e-16 ***
PC3         -2.142e-01  3.691e-03  -58.03   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.4068 on 9353 degrees of freedom
Multiple R-squared:  0.8346,    Adjusted R-squared:  0.8345
F-statistic: 1.573e+04 on 3 and 9353 DF,  p-value: < 2.2e-16
```

Figure 8: Resumen de la regresión.

Los resultados obtenidos del modelo se reflejan en la figura 8. De los mismos se deduce la siguiente ecuación para determinar el valor de la respuesta del sensor a la concentración de dióxido de nitrógeno.

$$PT08.S4.NO_2 = -0.26 * PC1 + 0.42 * PC2 - 0.22 * PC3$$

Uno de los factores a tener en cuenta es la estandarización de los datos durante el proceso de clasificación, por lo que su efecto se ve reflejado en la ausencia del término independiente en la ecuación de regresión.

En el modelo propuesto por cada unidad de decremento de todas las variables presentes en la componente 1 (Ver sección 2.1.1), el valor de respuesta del sensor al dióxido de nitrógeno disminuye en 0.26 unidades aproximadamente. Por cada unidad de incremento en la temperatura y la humedad absoluta, se incrementa la respuesta del sensor en aproximadamente 0.42 y finalmente por cada unidad de incremento de la humedad relativa y absoluta se decrementa la respuesta en 0.22 unidades aproximadamente.

La precisión del modelo medida en términos del valor de $R - ajustado$ es de 0.8343 lo cual es bastante alto tomando en consideración los datos faltantes y el desprecio de datos resultante del *ACP*. El $p - valor$ de la prueba de $F - statistic$ es menor que 0.05 por lo que podemos asegurar que nuestro modelo produce resultados. El error residual es de 0.4. Todas las variables independientes son de importancia para la estimación de la variable dependiente.

2.2.1 Análisis de los supuestos de la regresión

1. Las variables independientes no están correlacionadas.
2. La media y la suma de los errores es cero.
3. Los errores tienen distribución normal.
4. Los errores son independientes.

5. La varianza de los errores es constante.

El primer requisito de los supuestos del modelo se cumple al emplear como variables independientes el resultado de aplicar el ACP. (Ver gráfico 7).

Se obtuvo del modelo que la suma de los errores es $4.3e-13$ y la media de los mismos es $4.6e-17$ por lo que podemos asegurar que la media y la suma de los errores es 0.

```
        Durbin-Watson test

data:  regression
DW = 0.152, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 9: Prueba de Durwin-Watson.

Como se puede observar del resultado de la prueba de independencia de Durwin-Watson (Figura 9), el $p-valor << 0.05$ por lo que podemos rechazar la hipótesis nula y los errores son dependientes. Esto incumple con los supuestos delo modelo por lo que se termina su análisis y se descarta su empleo.

En la carpeta de imágenes adjunta al proyecto se pueden visualizar los resultados obtenidos de la normalidad de los errores y la homocedasticidad.

3. Cúlster, Kmeans y Árbol de Desición

Para el análisis de tipo clúster se tomaron en cuenta los resultados obtenidos del ACP, por lo que el número de particiones del clúster jerárquico se fijo a 3, pero también. se valoró la inclusión de 2 clústers como se puede apreciar en el dendograma (Figura 10)

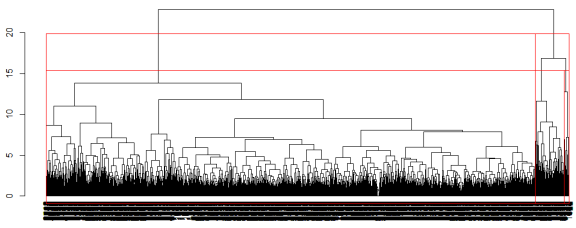


Figure 10: Dendograma del clúster jerárquico con 2 y 3 particiones de los datos.

Uno de los detalles que más denotan de la gráfica anterior recae en la gran acumulación de datos en la primera partición, posiblemente se deba a que los datos son tomados de una misma zona y no variaron mucho durante los tres meses del estudio.

Tomando los resultados anteriores, se dispuso de la ejecución del algoritmo $Kmeans$ cuyo resultado se puede observar en la figura 11. Primero se intentó una aproximación de dos particiones, pero se obtuvo una similitud entre componentes de un 33%, por lo que se

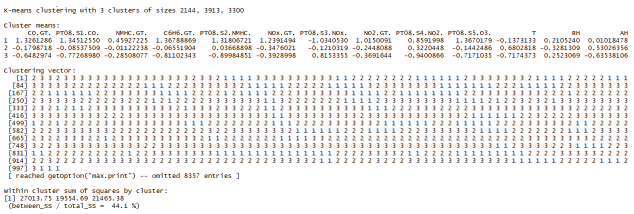


Figure 11: Resumen de la aplicación de Kmeans.

decidió emplear tres particiones para el resultado final del algoritmo.

Para concluir con el proceso de $Kmeans$, se analizó también de forma gráfica la distribución de los datos obtenidos (12). La distribución de los datos en tres particiones de tamaños 2144, 3913 y 3300, se ve reflejado en la gráfica. El color rojo se asigna a la 2da componente, el negro a la tercera y el verde a la primera.

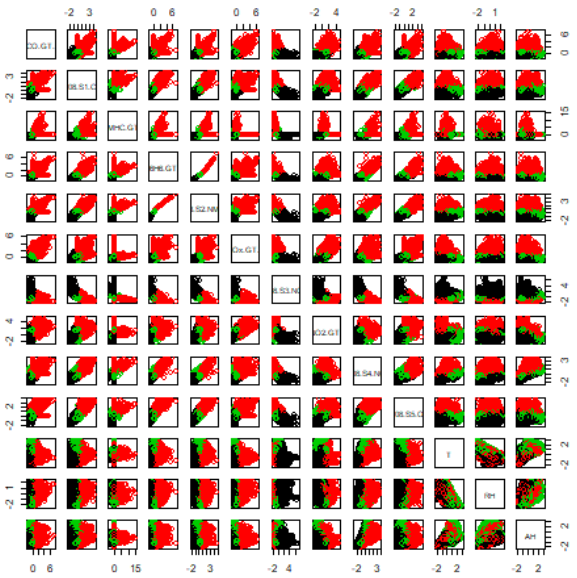


Figure 12: Forma gráfica de Kmeans por variable.

Tras el intento fallido mediante la regresión de establecer una relación entre la respuesta del sensor al dióxido de carbono, se centró el estudio del árbol de decisión sobre el mismo (Figura 13). Pero tras analizar el error de clasificación, este resultó casi uno, de un 99.98%, por lo que se desechó el mismo. Posteriormente, se analizó otro árbol de clasificación, esta vez con una variable con la mayor muestra de datos, el Tiempo. El experimento resultó una vez más con error similar al caso anterior por lo que también se desechó esta aproximación.

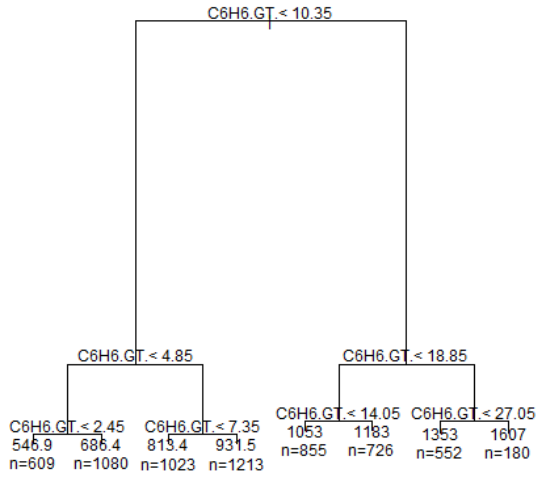
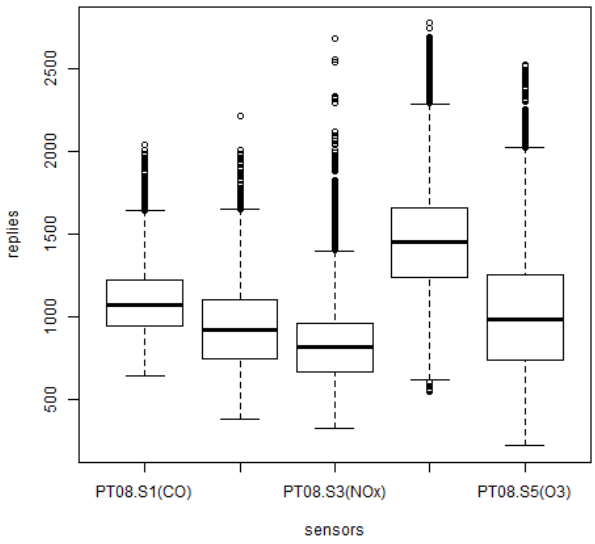


Figure 13: Árbol de decisión sobre la variable PT08.S2.NMHC.

4. ANOVA

Se intentó finalmente realizar el test ANOVA sobre los datos para estudiar el factor sensor dado que todos los sensores miden algún tipo de óxido.

Primeramente se analizó la gráfica de medias y diagrama de cajas simultáneo como se muestra a continuación.



Las etiquetas corresponden a cada uno de los sensores; y el eje replies las concentraciones del compuesto correspondiente.

Se aprecia diferencia entre cada una de las medias de los sensores, así como la presencia de datos aislados cuya cantidad es notable. Este fenómeno de la gran cuantía de datos aislados podemos afirmar que se debe a la sustitución de cada dato faltante por la media del resto. Estos datos remplazados se agrupan cerca de la media (porque tiene el mismo valor) y resulta que los datos reales quedan aislados. Se asegura entonces que el resultado del test va a arrojar diferencias entre las medias.

4.1 Hipótesis y Modelo Estadístico

El modelo estadístico escogido fue le de Clasificación Simple. Tenemos claramente el factor sensor, el cual fue previamente seleccionado para saber si existe diferencia entre la concentración media de los compuestos.

¿ Existe diferencia entre la concentración promedio medida por los sensores ?

La respuesta a esta pregunta es el resultado de contrastar las hipótesis:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5$$

$$H_1 : \text{existen } 1 \leq i < j \leq 5 \text{ tales que } \mu_i \neq \mu_j$$

Donde μ_i denota la concentración media medida por el sensor i. Como estudiamos en conferencias si denotamos y_{ij} como la concentración medida por el sensor i en la réplica j, la misma se puede escribir como: $y_{ij} = \mu_i + e_{ij}$, siendo e_{ij} el error experimental o la perturbación.

4.2 Diferencias entre la concentración promedio medida por cada sensor

Realizamos la prueba de hipótesis planteada anteriormente para un nivel de significación no especificado. Nuevamente apoyándonos en el lenguaje r. Los resultados de la misma son los siguientes (figura 14).

```
> summary(sensors.anova)
              Df    Sum Sq   Mean Sq F value Pr(>F)
sensors        4 2.100e+09 524973580    5901 <2e-16 ***
Residuals  46780 4.162e+09     88963
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Figure 14: Prueba de Hipótesis

Los resultados arrojan un *p-value* menor a 2×10^{-16}, el cual es prácticamente 0, pues el valor es cercano al meno número representable en la aritmética flotante de r. Con lo cual cualquier nivel significación $\alpha \in \{0.01, 0.1, 0.5\}$ es válido, con lo cual rechazamos la hipótesis nula y podemos afirmar que existe diferencia entre la concentración promedio medida por cada sensor.

La mayor concentración la tiene el 4to sensor. Dado el éxito de la prueba de hipótesis, partimos de la existencia de una diferencia, ahora bien, la evidencia visual nos muestra que la media del sensor 4 es superior, incluso si tenemos en cuenta los extremos del intervalo, estos también son superiores a los de los otros sensores.

4.3 Supuestos de normalidad y de igual varianza

La validez de los resultados obtenidos en cualquier análisis de varianza queda supeditada a que los supuestos del modelo se cumplan. Estos supuestos son:

1. Los e_{ij} siguen una distribución normal con media cero.
2. Los e_{ij} son independientes entre sí.
3. Los residuos de cada tratamiento tienen la misma varianza σ^2.

4.3.1 Pruebas Gráficas

Comenzando por las pruebas gráficas. Los resultados se muestran a continuación en las figuras 15, 16 y 17.

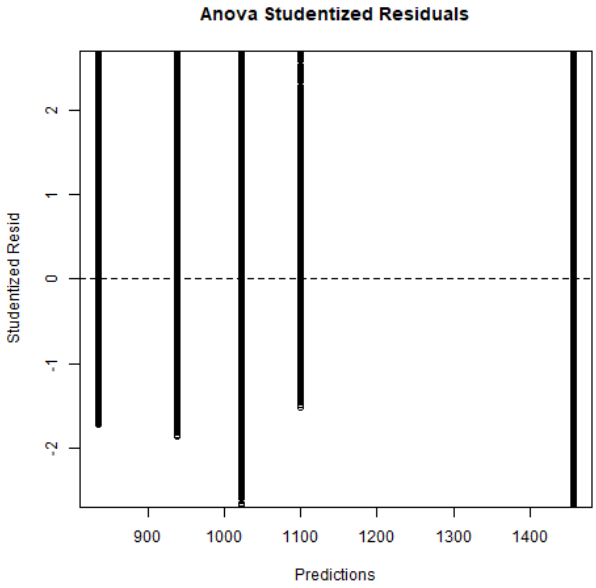


Figure 15: Residuos

Primeramente, en el gráfico estandarizado de residuos (figura 15), notamos los puntos muy dispersos, sin patrón aparente, por lo que podríamos aseguramos el supuesto de varianza constante.

En el gráfico de predichos contra los residuos (figura 16), todos los puntos tienden a estar sobre la una misma recta, aunque no están completamente alineados, y esto se pudiera deber al la sustitución de los datos faltantes o a la cantidad considerable de datos. La cantidad de puntos aberrantes es muy pequeña con respecto al tamaño de los datos por lo que podemos tolerar y seguir afirmando que se cumple el supuesto de normalidad.

Finalmente, en la gráfica del histograma de residuos (figura 17), dicha gráfica se asemeja decentemente a la normal con media 0. Con lo cual afirmamos que se cumple el supuesto de normalidad.

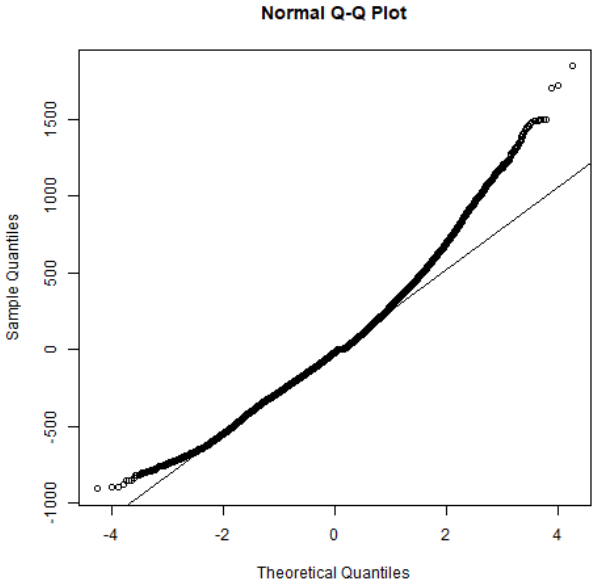


Figure 16: Predichos contra los residuos

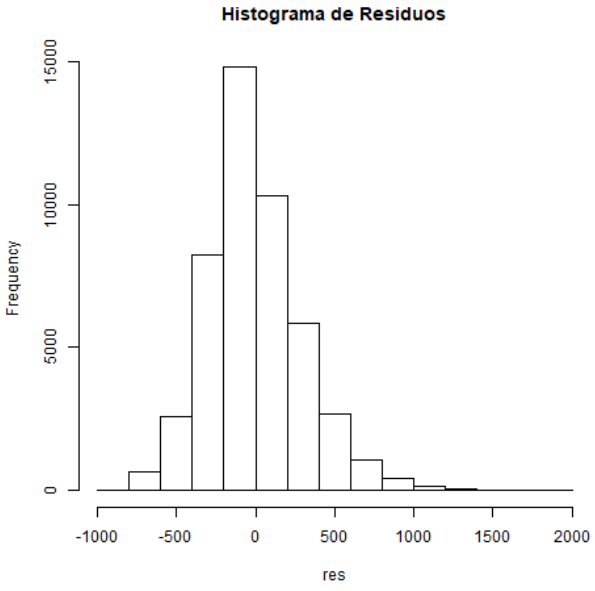


Figure 17: Histograma de residuos

4.3.2 Shapiro-Wilk, Bartlett y Durbin-Watson

El Test de Shapiro-Wilk no se pudo realizar por la gran cantidad de datos. Con lo cual, basándonos en las pruebas gráficas podemos asumir que el mismo se cumple y continuar con los dos test restantes.

```
        Bartlett test of homogeneity of variances

data:  res and df$sensors
Bartlett's K-squared = 4486.3, df = 4, p-value < 2.2e-16
```

Figure 18: Test de igual varianza Bartlett

Como muestra la prueba de la figura 18 (Test de

Bartlett), el *p-value* muestra que es significativa, con lo que rechazamos H_0, y no se cumple es supuesto de varianzas constantes. La prueba deja de tener validez.

```
        Durbin-Watson test

data:  sensors.anova
DW = 0.20914, p-value < 2.2e-16
alternative hypothesis: true autocorrelation is greater than 0
```

Figure 19: Test Durbin-Watson

Finalmente, y de igual forma la prueba de la figura 19 (Test Durbin-Watson) muestra que el test es significativo, por tanto no se cumple el supuesto de independencia.

En conclusión, los dos últimos supuestos fallaron y por ende todos los resultados iniciales no son válidos. A pesar del estado de los datos, se realizó el análisis. Concluimos que estos datos no son lo idóneos par una prueba de ANOVA.

5. Conclusiones

Este equipo considera que la falta de valores presentes en la fuente impide que se pueda llegar a conclusiones específicas sobre los datos. Se recomienda realizar otra recopilación para un futuro estudio. Cualquier análisis con los datos actuales se considera como no confiable.

Todos los códigos, resultados e imágenes presentes en este documento y relacionados con los experimentos se encuentran adjuntos dentro del directorio del proyecto.