Texto

Descripción generada automáticamente

**FACULTAD DE HUMANIDADES, COMUNICACIÓN Y DOCUMENTACIÓN**

**GRADO EN GESTIÓN DE LA INFORMACIÓN Y CONTENIDOS DIGITALES**

**BIG DATA**

**ANÁLISIS ESTADÍSTICO DE DATOS: BOSTON DATASET**

Pablo Blas Hernández - 100474502

# **ÍNDICE…………………………………………………………………….**

1. **ANÁLISIS DESCRIPTIVO DE LOS DATOS: ESTADÍSTICOS PRINCIPALES**
2. **MATRIZ DE CORRELACIONES ENTRE TODAS LAS VARIABLES**
3. **ANÁLISIS DETALLADO DE LAS VARIABLES**
4. **TABLA DE CONTINGENCIA ENTRE “LSTAT” Y “MEDV”**
5. **ANÁLISIS DE REGRESIONES**

5.1) REGRESIÓN LINEAL SIMPLE

5.1.1) DIAGNÓSTICO DEL MODELO

5.1.1.1) HETEROCEDASTICIDAD VS HOMOCEDASTICIDAD

5.1.1.2) NORMALIDAD DE LOS RESIDUOS

5.1.1.3) INDEPENDENCIA DE LOS RESIDUOS

5.2) REGRESIÓN LINEAL MÚLTIPLE

5.2.1) DIAGNÓSTICO DEL MODELO

5.2.1.1) HETEROCEDASTICIDAD VS HOMOCEDASTICIDAD

5.2.1.2) NORMALIDAD DE LOS RESIDUOS

5.2.1.3) INDEPENDENCIA DE LOS RESIDUOS

**6) ANÁLISIS DEL CRIMEN EN BOSTON EN BASE A LAS VARIBALES “MEDV” Y “LSTAT”**

# **ANÁLISIS DESCRIPTIVO DE LOS DATOS: ESTADÍSTICOS PRINCIPALES**

El dataset sobre el que realizaremos el estudio es el Boston Housing dataset disponible en la librería “MASS” de R. El objetivo del análisis es comprender los factores que afectan al precio de las viviendas en Boston, siendo “medv” (precio medio de la vivienda) nuestra variable target para el estudio. Las variables que componen el dataset son las siguientes:

* **CRIM**: Tasa de criminalidad per cápita por ciudad.
* **ZN**: Proporción de terreno residencial dividido en zonas para lotes de más de 25,000 pies cuadrados (cerca de 2300 metros cuadrados).
* **INDUS**: Proporción de acres de negocios no minoristas por ciudad.
* **CHAS**: Variable ficticia de Charles River que toma el valor 1 si el terreno limita con el río; de lo contrario, toma el valor 0.
* **NOX**: Concentración de óxidos nítricos (partes por 10 millones).
* **RM**: Promedio de habitaciones por vivienda.
* **AGE**: Proporción de unidades ocupadas por sus propietarios construidas antes de 1940.
* **DIS**: Distancias ponderadas a cinco centros de empleo en Boston.
* **RAD**: Índice de accesibilidad a carreteras radiales.
* **TAX**: Tasa de impuesto a la propiedad de valor total por $10,000.
* **PTRATIO**: Relación alumno-profesor por localidad.
* **BLACK**: 1000(Bk - 0.63)^2 donde Bk es la proporción de personas de ascendencia afroamericana por ciudad.
* **LSTAT**: Porcentaje de la población de menor estatus.
* **MEDV (variable target)**: Valor medio de las viviendas ocupadas por sus propietarios en miles de dólares.

A continuación, mostramos las principales **medidas de posición y dispersión** de las variables:

**Posición:**

Imagen que contiene texto, gente, sostener, grupo

Descripción generada automáticamente

*Figura 1: “Medidas de posición”*

**Dispersión:**

Tabla

Descripción generada automáticamente

*Figura 2: “Medidas de dispersión”*

# **MATRIZ DE CORRELACIONES ENTRE TODAS LAS VARIABLES**

Una matriz de correlación en R con valores de 1 a -1 representa la fuerza y dirección de la **relación lineal** entre cada par de variables en un conjunto de datos. Si el valor es cercano a 1, significa que existe una fuerte correlación positiva entre las dos variables. Cuando está cerca de -1, las variables tienen una fuerte correlación negativa.

Imagen que contiene Gráfico de rectángulos

Descripción generada automáticamente

*Figura 3: “Heatmap de las correlaciones entre las variables del dataset”*

Es importante tener en cuenta que la correlación no implica **causalidad**. Solo indica que existe una relación lineal entre las variables, pero no determina si una variable causa la otra. Debido a que existen regresores correlacionados entre ellos habrá que comprobar que los modelos estén libres del problema de multicolinealidad.

# **ANÁLISIS DETALLADO DE LAS VARIABLES**

Gráfico, Histograma

Descripción generada automáticamente

Gráfico, Histograma

Descripción generada automáticamente

*Figura 4: “Distribución del valor medio de las viviendas”*

El histograma muestra que la distribución del valor medio de las viviendas (medv) tiende a seguir una distribución sesgada hacia la derecha, lo que nos puede indicar que nuestra variable target “medv” no sigue una distribución normal. La asimetría a la derecha (sesgo positivo) sugiere que la mayoría de las viviendas están en el rango de valores bajos a moderados. Sin embargo, hay una notable presencia de valores altos que contribuyen a esta distribución sesgada positivamente. Esto indica que existen algunas viviendas con valores medios significativamente más altos (como casas de lujos o zonas ricas), que se consideran valores atípicos.

Además, la presencia de un gran número de valores en el extremo derecho (50) sugiere que puede haber un límite en los datos, es decir un techo en los precios de las viviendas en el dataset, y puede distorsionar el análisis al no representar adecuadamente la distribución de valores superiores a 50.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

*Figura 5: “Boxplot de las variables del dataset”*

Como podemos observar en el diagrama de caja de las variables con rango de valores entre 0 y 100, crim y zn obtienen un número de valores atípicos muy elevado y se alejan de tener una distribución asimétrica. En parte también, es el caso de la variable de precios de la vivienda que tiene observaciones atípicas por la cola derecha, es decir, valores superiores al percentil 90 de la distribución.

Gráfico, Gráfico de cajas y bigotes

Descripción generada automáticamente

*Figura 6: “Boxplot de las variables restantes del dataset”*

La variable Black es una transformación del porcentaje de personas de raza negra en un municipio concreto, por lo tanto, la mayoría de los valores están bastante concentrados en un rango pequeño, pero también tiene más valores atípicos que el resto de las variables. Sin embargo, la variable “age” y “tax” son regresores con una distribución concentrada.

El estudio se basa en analizar los factores que más influyen en el precio de las viviendas en Boston (“medv”), por ello calculamos los valores de correlación de las variables que contienen nuestros datos en base al precio de las viviendas.

Coeficiente\_correlacion Variable

lstat -0.7376627 lstat

ptratio -0.5077867 ptratio

indus -0.4837252 indus

tax -0.4685359 tax

nox -0.4273208 nox

crim -0.3883046 crim

rad -0.3816262 rad

age -0.3769546 age

chas 0.1752602 chas

dis 0.2499287 dis

black 0.3334608 black

zn 0.3604453 zn

rm 0.6953599 rm

A continuación, en el siguiente gráfico ordenamos los coeficientes de correlación de los posibles regresores de mayor a menor con respecto a nuestra variable target:

Gráfico, Histograma

Descripción generada automáticamente

*Figura 7: “Gráfico de correlación de las variables respecto a la variable target medv”*

A simple vista, las variables “rm” y “lstat” obtienen valores absolutos significativamente superiores al resto de variables, siendo “lstat” la variable independiente que más influye en el precio medio de la vivienda (“medv”). Podemos obtener las siguientes conclusiones:

* Las casas con más habitaciones ("rm” más alto) son más caras: Las casas con más dormitorios generalmente son más grandes y albergan a más personas, por lo que cuestan más. Es una variable directamente proporcional a la variable target “medv” que indica el precio medio de las viviendas.
* Las áreas con más personas pertenecientes a los estatus más bajos (valores “lstat” más altos) tienen costos más bajos. Una proporción mayor de personas de la clase trabajadora baja es un indicador de menor poder adquisitivo y, por lo tanto, valores de vivienda más bajos. Es una variable inversamente proporcional a la variable target “medv” que indica el precio medio de las viviendas.

Ambas variables deben ser tenidas en cuenta a la hora de realizar una inversión en una vivienda puesto que son las que más influyen en su precio.

Diagrama

Descripción generada automáticamente

*Figura 8: “Gráficos de dispersión para determinar relaciones entre variables”*

“lstat”, “dis” y “rm” pueden tener potencialmente un fit lineal adecuado para la variable target (“medv”). El resto parecen no obtener una relación aparente.

# **TABLA DE CONTINGENCIA ENTRE LSTAT Y MEDV**

Aunque existe una fuerte correlación entre las variables LSTAT y MEDV, tal correlación no implica necesariamente causalidad. Para verificar que existe un argumento convincente que justifique la relación entre variables, se utiliza una tabla de contingencia.

Gráfico

Descripción generada automáticamente

*Figura 9: “Tabla de contingencia entre las variables lstat y medv”*

Podemos generar una serie de insights a partir de los resultados de la tabla:

* **Relación inversa entre lstat y medv**: Observamos una clara tendencia inversa en la tabla, donde a medida que el nivel de estatus más bajo de la población (lstat) aumenta, los precios medios de las viviendas (medv) tienden a disminuir. Esto sugiere que las viviendas tienden a ser más caras en áreas donde el nivel de estatus socioeconómico es más alto.
* **Mayor cantidad de viviendas de precio bajo en áreas con alto lstat**: Notamos que la mayoría de las viviendas en áreas con un nivel de estatus socioeconómico más bajo (lstat en las categorías "Alto" y "Muy alto") tienen precios más bajos (medv en las categorías "Muy bajo" y "Bajo").
* **Menos viviendas de precio alto en áreas con alto lstat**: Por el contrario, hay menos viviendas con precios altos (medv en las categorías "Alto" y "Muy alto") en áreas donde el nivel de estatus socioeconómico es más bajo.

Realizamos una prueba chi-cuadrado donde se considera como **hipótesis inicial o nula Ho** **que no hay diferencias,** lo que equivaldría a afirmar que los resultados son independientes entre sí (una variable no afecta a la otra):

Pearson's Chi-squared test

data: tabla\_contingencia\_final

X-squared = 460.11, df = 9, p-value < 2.2e-16

El p valor de la prueba chi-squared es inferior a 0.05. Por lo tanto, sugiere que hay evidencia suficiente para rechazar la hipótesis nula, es decir, se rechaza la hipótesis nula de independencia entre las variables y se concluye que hay una asociación significativa entre las variables.

**ANÁLISIS DE REGRESIONES**

## **Regresión lineal simple**

En este caso, realizaremos una regresión líneal múltiple del “medv” (precio medio de la vivienda y target de nuestro estudio) y lstat (población con bajo estatus).

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Figura 10: “Diagrama de dispersión y recta de regresión entre medv y lstat”*

En el diagrama de dispersión de las variables medv y lstat (porcentaje de población con bajo estatus) parece existir una relación negativa, pero con una vertiente no lineal.

Pasamos a realizar la primera prueba con un modelo simple de regresión lineal.

lm.fit =lm(medv~lstat ,data=Boston )

summary(lm.fit)

##

## Call:

## lm(formula = medv ~ lstat, data = Boston)

## Residuals:

## Min 1Q Median 3Q Max

## -15.168 -3.990 -1.318 2.034 24.500

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 34.55384 0.56263 61.41 <2e-16 \*\*\*

## lstat -0.95005 0.03873 -24.53 <2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 6.216 on 504 degrees of freedom

## Multiple R-squared: 0.5441, Adjusted R-squared: 0.5432

## F-statistic: 601.6 on 1 and 504 DF, p-value: < 2.2e-16

Existe una relación lineal negativa demostrada por el coeficiente negativo de -0.95. Este coeficiente es significativo a un nivel de confianza del 99% ya que el p-value asociado es inferior a 0,01. El R2 ajustado es bastante elevado pero el modelo puede aumentar su explicabilidad añadiendo regresores adicionales. En este caso, la variabilidad del precio medio de la vivienda viene explicado en un 54% por la variabilidad del porcentaje de población con bajo estatus. Después de realizar el fit, analizamos el comportamiento de los residuos para discernir si existe algún patrón.

### **Diagnóstico del modelo**

#### **HETEROCEDASTICIDAD VS HOMOCEDASTICIDAD**

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Figura 11: “Residuos vs valores predichos”*

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Figura 12: “Residuos estandarizados vs valores predichos”*

El primer gráfico (grafico residuos vs. valores ajustados) ayuda a identificar si existe una relación entre los residuos y los valores ajustados, en este caso existe un patrón basado en una relación no lineal que no estamos captando y que también se ve reflejado en el segundo gráfico con los residuos estandarizados. En el segundo gráfico se puede identificar que existe cierta varianza en los residuos de los valores ajustados. Por lo que en un principio podemos considerar que se trata de un modelo heterocedástico.

Vamos a realizar una prueba Breusch-Pagan para confirmar si nos encontramos ante un modelo heterocedástico u homocedástico:

> #Breusch-Pagan

> bptest(lm.fit)

studentized Breusch-Pagan test

data: lm.fit

BP = 15.497, df = 1, p-value = 8.262e-05

Obtenemos un valor inferior a p=0.05 por lo que se determina que hay suficiente evidencia para rechazar la hipótesis nula de homocedasticidad y declarar que es un modelo **heterocedástico.** Esto significa que la variabilidad de los errores (residuos) en el modelo no es constante en todos los niveles de lstat. En otras palabras, la varianza de los errores cambia a medida que cambia la variable lstat.

#### **NORMALIDAD DE LOS RESIDUOS**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

*Figura 13: “Alineamiento de los residuos”*

Las características de la distribución de estos residuos también se pueden observar en el gráfico QQ-plot, existe cierta desviación respecto al ideal de distribución normal (puntos en la línea diagonal).

Realizamos un test de normalidad de Shapiro-Wilk para analizar si los residuos siguen una distribución normal

H0: los residuos muestran comportamiento normal.

Shapiro-Wilk normality test

data: resid(lm.fit)

W = 0.87857, p-value < 2.2e-16

p-valor muy inferior a 0.05, por lo que rechazamos H0, los residuos NO se comportan como una normal.

Adicionalmente calculamos Kurtosis y Skeweness, comprobando que la kurtosis queda fuera del rango comprendido entre -2 y 2, por lo que podemos de nuevo rechazar la hipótesis de normalidad.

skewness(residuos)

[1] 1.448435

kurtosis(residuos)

[1] 2.298108

#### **INDEPENDENCIA DE LOS RESIDUOS**

Por último, hay que verificar que los errores no están relacionados con la variable LSTAT.

Para ello realizamos el test de Durbin-Watson, asumiendo como H0 que no existe autocorrelación entre los residuos, por lo que serían independientes.

dwtest(lm.fit) Durbin-Watson test

data: lm.fit

DW = 0.8915, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is greater than 0

Se observa un p-valor mucho menor que 0,05, por lo que rechazamos la hipótesis inicial: los residuos NO son independientes. Esto es lógico pues es un modelo "incompleto" ya que falta información por lo que hay que añadir variables, y es por ello, que realizaremos un modelo múltiple.

Esto quiere decir que “lstat” por sí solo no puede explicar el precio medio de las viviendas en Boston y necesita de otras variables que le complementen para poder explicar con certeza dicho precio.

## **Regresión lineal múltiple**

Incorporamos la variable lstat^2 para captar esta información:

Boston$lstat2 <- (Boston$lstat)^2

regfit.comparacion = regsubsets(medv~.,Boston, nvmax=10)

summary(regfit.comparacion)

Regsubsets es un algoritmo que te señala el mejor modelo según el número de regresores máximos que quieres incluir En este caso el mejor modelo de una variable contendría lstat, el mejor modelo de 2 variables utilizaría lstat y rm, el de 3 variables sería lstat, rm y ptratio, etc.

Nota del libro Introduction to Statistical Learning sobre este algoritmo: “The regsubsets() function (part of the leaps library) performs best subset selection by identifying the best model that contains a given number of predictors, where best is quantified using RSS. The syntax is the same as for lm(). The summary() command outputs the best set of variables for each model size”.

Teniendo en cuenta la información del algoritmo realizamos el resumen de regresiones para estas combinaciones:

correlation\_df\_aux = correlation\_df

correlation\_df\_aux$Coeficiente\_correlacion = abs(correlation\_df\_aux$Coeficiente\_correlacion)

correlation\_df\_aux <- correlation\_df\_aux[order(correlation\_df\_aux$Coeficiente\_correlacion), ]

correlation\_df\_aux$Variable

## [1] "chas" "dis" "black" "zn" "age" "rad" "crim"

## [8] "nox" "tax" "indus" "ptratio" "rm" "lstat"

Este orden nos indica las de menor a mayor la significancia de las variables independientes respecto a nuestra variable target “medv” siendo “chas” la variable con menor significación en el precio de la vivienda y “lstat” la variable que más influye en este.

Imagen que contiene Gráfico de dispersión

Descripción generada automáticamente

*Figura 14: “Comparación de modelos de regresión lineal múltiple”*

Podemos observar como la métrica del R2 ajustado, que corrige el R2 estándar para evitar añadir variables que no aporten suficiente información para la variable dependiente, aumenta con cada regresor que añadimos, pero de forma decreciente a partir del modelo con 3 regresores. Las estrellas nos indican el nivel de significación y el numero entre paréntesis representa el error de estandard. Como podemos observar casi todas las variables son significativas y distintas de 0 a un nivel de significación del 99%.

<<<<Modelo final segun el anova empezando por con un modelo con todas las variables medv ~ age + rad + crim + nox + tax + indus + ptratio + rm + lstat + lstat2

<<<<Modelo final segun el anova empezando por con un modelo con pocas variables medv ~ ptratio + rm + lstat + lstat2>>>>>

Escogemos el modelo 5 ya que el test anova no rechaza la nula (p valor es ligeramente superior al 5%), es decir, la diferencia entre el modelo 4 y 5 no es estadísticamente significativa y, por lo tanto, nos quedaríamos con el modelo más simplificado con menos regresores.

## Call:

## lm(formula = medv ~ age + rad + crim + nox + tax + indus + ptratio +

#rm + lstat + lstat2, data = Boston)

##

## Residuals:

## Min 1Q Median 3Q Max

## -15.6943 -2.8079 -0.4346 1.9408 26.4576

##

## Coefficients:

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 37.955365 4.316949 8.792 < 2e-16 \*\*\*

## age 0.061925 0.011844 5.228 2.53e-07 \*\*\*

## rad 0.312699 0.062477 5.005 7.78e-07 \*\*\*

## crim -0.150608 0.031273 -4.816 1.95e-06 \*\*\*

## nox -9.759260 3.444445 -2.833 0.00479 \*\*

## tax -0.013584 0.003453 -3.934 9.53e-05 \*\*\*

## indus 0.148106 0.055854 2.652 0.00827 \*\*

## ptratio -0.869033 0.119761 -7.256 1.55e-12 \*\*\*

## rm 3.157688 0.400668 7.881 2.08e-14 \*\*\*

## lstat -1.957427 0.130098 -15.046 < 2e-16 \*\*\*

## lstat2 0.038673 0.003394 11.395 < 2e-16 \*\*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##

## Residual standard error: 4.54 on 495 degrees of freedom

## Multiple R-squared: 0.7612, Adjusted R-squared: 0.7564

## F-statistic: 157.8 on 10 and 495 DF, p-value: < 2.2e-16

Debido a que incorporamos un número elevado de regresores, conviene hacer un test de multicolinealidad para comprobar la estabilidad del modelo y esto lo haremos con el Variance Inflation Factor (VIF).

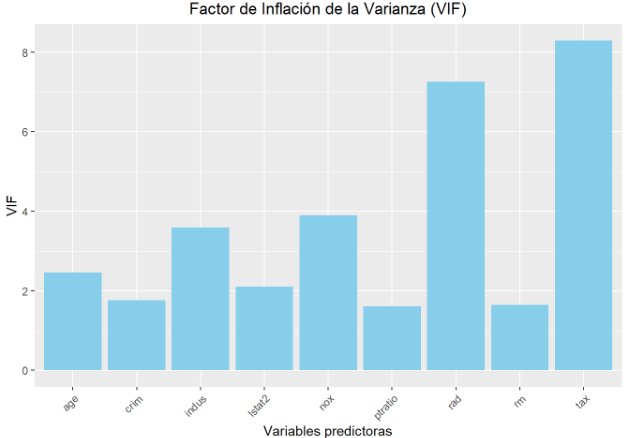
*Gráfico, Gráfico de barras, Histograma

Descripción generada automáticamente*

*Figura 15: “VIF”*

Como podemos observar los valores VIF son superiores a 10 en “lstat” y su transformación al cuadrado. Esto indica de forma evidente multicolinealidad fuerte y complica la interpretación de los coeficientes de regresión debido a la alta correlación entre las variables predictoras. Adicionalmente, los coeficientes individuales pueden ser imprecisos y difíciles de interpretar.

Por lo tanto, para solventar problemas de estabilidad de modelo, eliminamos “lstat” y obtenemos valores más razonables:

**

*Figura 16: “VIF sin la variable lstat”*

### **Diagnóstico del modelo**

#### **HETEROCEDASTICIDAD VS HOMOCEDASTICIDAD**

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Figura 17: “Residuos vs valores predichos”*

*Gráfico, Gráfico de dispersión

Descripción generada automáticamente*

*Figura 18: “Residuos estandarizados vs valores predichos”*

En el primer gráfico, podemos observar cómo aunque obtiene un patrón ligeramente curvado, tiene residuos igualmente distribuidos alrededor de la línea horizontal sin un número significativo (3) de observaciones atipicas destacables.

En el gráfico 2 se realiza un gráfico de Spread-Location con el objetivo de demostrar si el modelo respeta la asunción de homoscedasticidad. Se muestra como los puntos están distribuidos aleatoriamente por encima y por debajo de la línea roja sin un patrón aparente.

Vamos a realizar una prueba Breusch-Pagan para confirmar si nos encontramos ante un modelo heterocedástico u homocedástico:

> #Breusch-Pagan

> bptest(regresionlmultiple5\_aux)

studentized Breusch-Pagan test

data: regresionlmultiple5\_aux

BP = 44.235, df = 9, p-value = 1.277e-06

El p-valor es inferior a 0.05 y, por lo tanto, rechazamos la hipótesis nula de homocedasticidad y existen indicios de heterocestasciticidad y la variabilidad de los errores no es constante en todos los niveles de los regresores. Para corregirlo, utilizamos un ajuste lineal con corrección de errores (errores estándares robustos) y los coeficientes son los siguientes:

coeftest(regresionlmultiple5\_aux, vcov = vcovHC(regresionlmultiple5\_aux, "HC1"))

##

## t test of coefficients:

##

## Estimate Std. Error t value Pr(>|t|)

## (Intercept) 19.7367465 7.2496195 2.7225 0.006708 \*\*

## age 0.0060837 0.0149271 0.4076 0.683775

## rad 0.3063803 0.0731980 4.1856 3.364e-05 \*\*\*

## crim -0.1173988 0.0277539 -4.2300 2.784e-05 \*\*\*

## nox -11.9052879 3.9271527 -3.0315 0.002560 \*\*

## tax -0.0148934 0.0027963 -5.3261 1.524e-07 \*\*\*

## indus 0.1046465 0.0530271 1.9735 0.048998 \*

## ptratio -1.1595594 0.1310296 -8.8496 < 2.2e-16 \*\*\*

## rm 5.5153621 0.8408315 6.5594 1.360e-10 \*\*\*

## lstat2 -0.0088438 0.0028628 -3.0892 0.002120 \*\*

## ---

## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

#### **NORMALIDAD DE LOS RESIDUOS**

Gráfico, Gráfico de líneas

Descripción generada automáticamente

*Figura 19: “Alineamiento de los residuos”*

Existe menos desviación respecto al ideal de distribución normal (puntos en la linea diagonal) si lo comparamos con el modelo sin introducir polinomios.

Realizamos un test de normalidad de Shapiro-Wilk para analizar si los residuos siguen una distribución normal:

> #Test de normalidad de Shapiro-Wilk

> shapiro.test(resid(regresionlmultiple5\_aux))

Shapiro-Wilk normality test

data: resid(regresionlmultiple5\_aux)

W = 0.86205, p-value < 2.2e-16

p-valor muy inferior a 0.05, por lo que rechazamos H0, los residuos NO se comportan como una normal.

Adicionalmente calculamos Kurtosis y Skeweness, comprobando que la kurtosis queda fuera del rango comprendido entre -2 y 2, por lo que podemos de nuevo rechazar la hipótesis de normalidad.

> #Kurtosis y Skeweness

> skewness(residuos)

[1] 2.009581

> kurtosis(residuos)

[1] 8.075961

#### **INDEPENDENCIA DE LOS RESIDUOS**

Para dtereminar la independencia de los residuos realizamos el test de Durbin-Watson, asumiendo como H0 que no existe autocorrelación entre los residuos, por lo que serían independientes.

Durbin-Watson test

data: regresionlmultiple5\_aux

DW = 0.85887, p-value < 2.2e-16

alternative hypothesis: true autocorrelation is greater than 0

Se observa un p-valor mucho menor que 0,05, por lo que rechazamos la hipótesis inicial: los residuos NO son independientes.

## **Análisis del crimen en Boston en base a las variables “medv” y “lstat”**

El crimen se representa mediante la variable “crim” (tasa de criminalidad per cápita por ciudad en Boston). Ahora será importante analizar como el crimen afecta a la variable target de este estudio (“medv”) y la variable que más repercute a esta (“lstat”).

Gráfico, Gráfico de dispersión

Descripción generada automáticamente

*Figura 20: “Distribución del crimen en Boston en base a las variables medv y lstat”*

El gráfico de dispersión muestra una relación inversa entre la tasa de criminalidad (“crim”) y el valor medio de las viviendas (“medv”). A medida que aumenta el valor medio de las viviendas, la tasa de criminalidad tiende a disminuir, lo que sugiere que las áreas con viviendas de mayor valor tienden a tener una menor incidencia de crimen.

También se observa una relación positiva entre la tasa de criminalidad y el porcentaje de población de estatus bajo (“lstat”) Las áreas con un mayor porcentaje de población de estatus bajo tienden a tener tasas de criminalidad más altas, lo que indica una asociación entre la pobreza o la desventaja socioeconómica y el crimen.