Taller RML (Parte 1) - Grupo 4

Departamento de Estadística - UNALMED

Luis David Hernandez Peréz Daniel Felipe villa Rengifo
Juan Gabriel Carvajal Negrete

Punto 1

Realice una descripción de la base de datos. Contextualice el problema y explique cada una de las variables involucradas en el modelo. <https://www.codersarts.com/post/predict-boston-house-prices-using-python-linear-regression

Descripción de la base de datos

Tomaremos el conjunto de datos de Vivienda que contiene información sobre diferentes casas en Boston. Hay 506 muestras y 13 variables de características en este conjunto de datos. El objetivo es predecir el valor de los precios de la casa utilizando las características dadas.

La descripción de todas las características se proporciona a continuación:

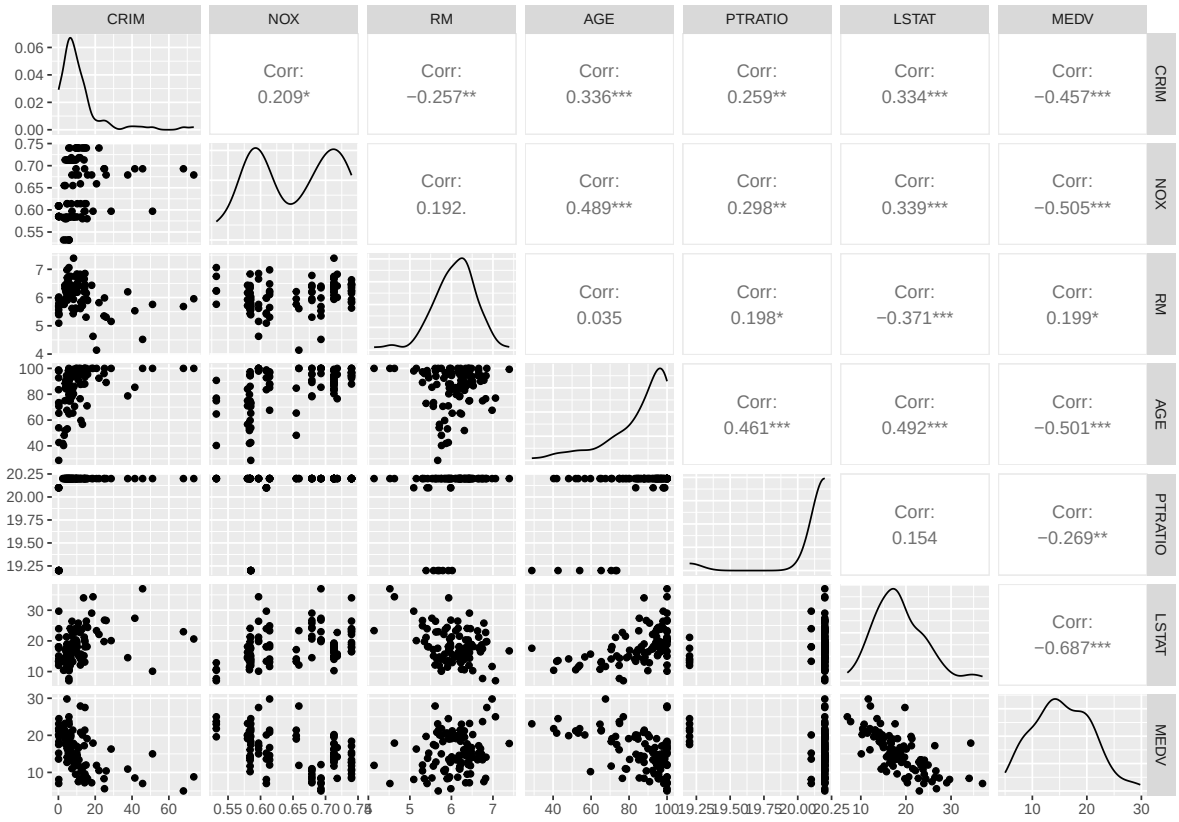
- **CRIM**: tasa de criminalidad per cápita por ciudad
- **ZN**: proporción de terrenos residenciales zonificados para lotes de más de 25 000 pies cuadrados
- **INDUS:** proporción de acres de negocios no minoristas por ciudad
- **CHAS**: variable ficticia de Charles River (= 1 si el terreno limita con el río; 0 en caso contrario) NOX: concentración de óxido nítrico (partes por 10 millones)
- **RM**: número 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 de Boston
- **RAD**: índice de accesibilidad a carreteras radiales
- **TAX**: tasa de impuesto a la propiedad de valor total por cada $10 000

- **PTRATIO**: proporción de alumnos por maestro por ciudad B: $1000(Bk - 0{,}63)^2$, donde Bk es la proporción de [personas de afroamericanas [descendencia] por ciudad
- **LSTAT**: Porcentaje de la población de estatus inferior
- **MEDV**: Valor medio de las viviendas ocupadas por sus propietarios en miles de dólares

Los precios de las viviendas, representados por la variable **MEDV**, constituyen nuestra variable objetivo. El resto de las variables actuarán como características que utilizaremos para predecir el valor de una vivienda.

Punto 2

Realice un análisis descriptivo de las variables que se van a tener en cuenta en el modelo. Concluya.



Correlaciones: MEDV tiene una fuerte correlación positiva con RM (número promedio de habitaciones por vivienda), lo que indica que las viviendas con más habitaciones tienden a tener un valor más alto. Por otro lado, MEDV tiene una fuerte correlación negativa con LSTAT

(porcentaje de la población de estatus inferior), lo que sugiere que las viviendas en áreas con mayor porcentaje de población de bajos ingresos tienden a tener un valor más bajo.

Otras variables: CRIM (tasa de criminalidad): Tiene una correlación negativa moderada con MEDV, lo que indica que las viviendas en áreas con mayor criminalidad tienden a tener un valor más bajo.RM (número de habitaciones): Además de su fuerte correlación con MEDV, RM también muestra una correlación positiva con AGE (proporción de unidades ocupadas por sus propietarios construidas antes de 1940). Esto podría indicar que las viviendas más antiguas tienden a tener más habitaciones.LSTAT (porcentaje de población de bajos ingresos): Además de su correlación negativa con MEDV, LSTAT también muestra correlaciones negativas con RM y una correlación positiva con CRIM. Esto sugiere que las áreas con mayor porcentaje de población de bajos ingresos tienden a tener viviendas más pequeñas y mayor criminalidad.

```
datos4 %>% ggplot(aes(x=MEDV))+
  geom_histogram(aes(y=..density..), fill="dodgerblue") +
  geom_density(linewidth = 1.2) +
  theme_bw() +
  theme(axis.title.y = element_blank())
```

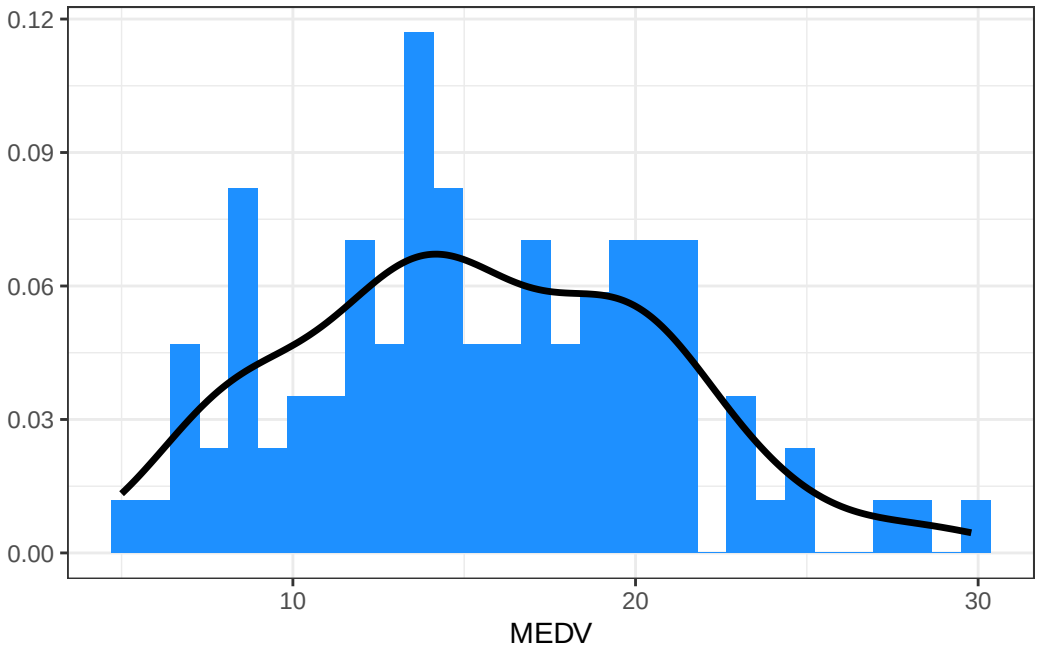


Figura 1: Distribución de la variable respuesta `MEDV`

Vemos que los valores de la variable `MEDV` se distribuyen no normales y en su mayoría agrupados por lo cual podríamos decir que no hay presencia de datos atípicos.

Prueba de normalidad

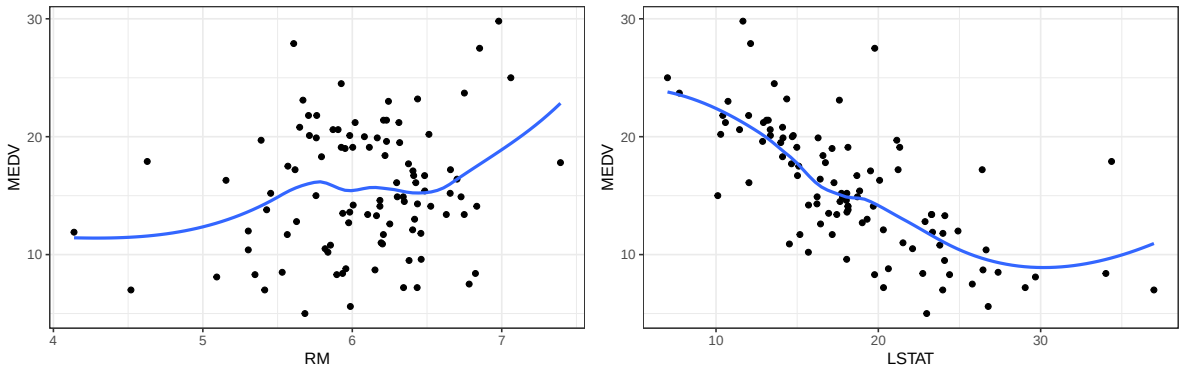
```
shapiro.test(datos4$MEDV)
```

```
    Shapiro-Wilk normality test

data:  datos4$MEDV
W = 0.98536, p-value = 0.3368
```

Distribución: Como se nota en la prueba de shapiro, la variable MEDV se distribuye de manera normal, con algunos valores atípicos en el extremo superior. Esto significa que la mayoría de las viviendas tienen un valor medio cercano a la media, con algunas viviendas que tienen valores significativamente más altos.

Factores influyentes: El análisis sugiere que el valor medio de las viviendas (MEDV) está influenciado por varios factores, incluyendo el número de habitaciones (RM), el porcentaje de población de bajos ingresos (LSTAT), la tasa de criminalidad (CRIM) y posiblemente la edad de la vivienda (AGE). Sabiendo todo esto, podemos ajustar un modelo de regresión múltiple.

Con base en las observaciones anteriores graficaremos la variable`RM` y `LSTAT` usando un diagrama de dispersión, frente a la variable de respuesta `MEDV`.



Podemos notar que los precios de los hogares aumentan a medida que el valor de la variable `RM` aumenta "linealmente". Hay pocos valores atípicos y los datos parecen tener cierto limite en 9.

Los precios tienden a disminuir a medida que aumenta la variable `LSTAT`. aunque no parece seguir exactamente una tendencia lineal "cuadrática o exponencial".

Punto 3

Ajuste un modelo de regresión lineal múltiple, muestre la tabla de parámetros ajustados y escriba la ecuación ajustada. Calcule la Anova del modelo ¿Es significativo el modelo? ¿Qué proporción de la variabilidad total de la respuesta es explicada por el modelo? Opine sobre esto último.

```
# Modelo de regresión lineal múltiple con todas las variables
modelo <- lm(MEDV~.,datos4)
```

Tabla 1: resumen del modelo de regresión

Coeficientes	**Estimación**	**Error Estándar**	**Valor t**	**Pr(>|t|)**
(Intercepto)	59.06933	30.51461	1.936	0.05593 .
CRIM	-0.08634	0.03139	-2.750	0.00715 **
NOX	-21.92000	6.51255	-3.366	0.00111 **
RM	0.26285	0.80549	0.326	0.74492
EDAD	-0.01242	0.02954	-0.421	0.67507
PTRATIO	-1.00533	1.59420	-0.631	0.52984
LSTAT	-0.46531	0.08208	-5.669	1.61e-07 *

El modelo de regresión lineal presentado busca predecir el valor medio de las viviendas (MEDV) en Boston, encontrando que la criminalidad (CRIM), la contaminación (NOX) y el nivel socioeconómico (LSTAT) son los predictores más importantes, con coeficientes negativos y significativos. Mientras que el número de habitaciones (RM), la edad de las viviendas (AGE) y la calidad de la educación (PTRATIO) no muestran una influencia significativa en este modelo. El intercepto (59.06933) representa el valor estimado de una vivienda cuando todas las variables predictoras son cero, pero debe interpretarse con cautela. Se podrían realizar mejoras al modelo, como eliminar variables no significativas, verificar la multicolinealidad, explorar otras variables relevantes (distancia a empleos, áreas verdes), considerar interacciones entre variables y evaluar posibles relaciones no lineales.

Ecuación del modelo

$$\widehat{y}_0 = 59.06933 - 0.08634x_1 - 21.92000x_2 + 0.26285x_3 - 0.01242x_4 - 1.001242x_5 - 1.00533x_6 - 0.46531x_7 + \epsilon_i$$

Tabla 2: Resultados de las pruebas estadísticas

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Prueba1	1	0.13	0.13	0.00	0.9628
Prueba2	1	1739.90	1739.90	31.08	0.0001
Prueba3	1	4138.43	4138.43	73.92	0.0000
Prueba4	1	608.37	608.37	10.87	0.0049
Residuals	15	839.72	55.98		

Tabla de variabilidad del modelo (Tabla Anova)

Comentarios

Punto 4

Pruebe la significancia individual de cada uno de los parámetros del modelo (excepto intercepto), usando la prueba t, y para dos cualesquiera de las predictoras, establezca claramente la prueba de hipótesis y el criterio decisión.

Para realizar el test de significancia individual para cada uno de los parámetros β_jj , esto es, probar

$$H_0 : \beta_j = 0 \text{ vs. } H_1 : \beta_j \neq 0$$

Teniendo en cuenta además que σ^2 es desconocido, mediante el siguiente estadístico de prueba con su distribución bajo H_0:

$$T_0 = \frac{\hat{\beta}_j}{\sqrt{\text{MSE}\, C_{jj}}} \overset{H_0}{\sim} t_{n-k-1};$$

Tabla 3: Resumen del modelo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	59.07	30.51	1.936	0.05593
CRIM	-0.08634	0.03139	-2.75	0.00715
NOX	-21.92	6.513	-3.366	0.00111
RM	0.2628	0.8055	0.3263	0.7449
AGE	-0.01242	0.02954	-0.4205	0.6751
PTRATIO	-1.005	1.594	-0.6306	0.5298
LSTAT	-0.4653	0.08208	-5.669	0

De acuerdo con la Tabla 3, se observa que las variables `CRIM`, `NOX`y `LSTAT` resultan estadísticamente significativas, dado que sus valores p son menores al nivel de significancia establecido ($\alpha = 0.05$).

Para las variables `CRIM` y `AGE`, establezeremos la prueba de hipótesis y el criterio de decision.

Para la variable `CRIM` las hipótesis son las siguientes:

$$H_0 : \beta_1 = 0 \text{ vs. } H_1 : \beta_1 \neq 0$$

Para la variable `AGE` las hipótesis son las siguientes:

$$H_0 : \beta_4 = 0 \text{ vs. } H_1 : \beta_4 \neq 0$$

Y los criterios de desición son:

Rechazo con valor P: si $P\left(|t_{n-k-1}| > |T_0|\right)$ es pequeño;Rechazo con región crítica a un nivel de significancia α

Punto 5

Teniendo en cuenta los resultados anteriores, realice una prueba con sumas de cuadrados extras utilizando el test lineal general; especifique claramente el modelo reducido y el modelo completo, el estadístico de la prueba, su distribución, el cálculo del valor P, la decisión y la conclusión a la luz de los datos. Justifique la hipótesis que desea probar en este numeral.

Del resultado anterior de la Tabla 3 se considera considere el test parcial de la significancia de coeficientes asociados a las variables`RM`,`AGE` y`PTRATIO`.

$$H_0 : \beta_3 = \beta_4 = \beta_5 = 0 \text{ vs. } H_1 : \beta_j \neq 0, \ \textit{para al menos un } j, \ \textit{conj} = 3, 4 \ y \ 5$$

En este caso la matriz L corresponde a,

$$L = \begin{bmatrix} 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

- **Modelo Completo (MF)**

$$Y_i = \beta_0 + \beta_1 \text{CRIM}_i + \beta_2 \text{NOX}_i + \beta_3 \text{RM}_i + \beta_4 \text{AGE}_i + \beta_5 \text{PTRATIO}_i + \beta_6 \text{LSTAT}_i + E_i$$

- **Modelo Reducido (MR)**

$$Y_i = \beta_0 + \beta_1 \text{CRIM}_i + \beta_2 \text{NOX}_i + \beta_6 \text{LSTAT}_i + E_i$$

Por tanto, $SSE(MR) = SSE(X_1, X_2, X_6)$ con n $-$ 3 grados de libertad y SSRMR) = SSR(X3,X4) con 2 grados de libertad. Luego, por la igualdad

Punto 6

Calcule las sumas de cuadrados tipo I (secuenciales) y tipo II (parciales) ¿Cuál de las variables tienen menor valor en tales sumas? ¿Qué puede significar ello?

Expecificando el modelo que estamos trabajando

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 X_{i3} + \beta_4 X_{i4} + \beta_5 X_{i5} + \beta_6 X_{i6} + E_i, \quad E_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2)$$

Utilirizaremos la funcion `anova()` para obtener la suma de cuadrados $SS1$.

Tabla 4: Sumas de cuadrados tipo I

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
CRIM	1	583.851029	583.851029	48.911687	0.0000000
NOX	1	490.821090	490.821090	41.118173	0.0000000
RM	1	112.756068	112.756068	9.446056	0.0027767
AGE	1	116.388172	116.388172	9.750333	0.0023899
PTRATIO	1	2.483889	2.483889	0.208086	0.6493352
LSTAT	1	383.573149	383.573149	32.133556	0.0000002
Residuals	93	1110.126202	11.936841	NA	NA

De la Tabla 4 podemos concluir que:

- La variable `PTRATIO` tiene la menor suma de cuadrados, lo que implica que su efecto en la variable `MEDV` es muy bajo cuando se incluye después de las demás variables.Esto podría sugerir que su efecto ya está explicado por las otras variables del modelo o que no tiene relacion directamente con la variable respuesta.
- Las variables `CRIM`, `NOX` y `LSTAT` son las más importantes, ya que explican una gran proporción de la varianza en `MEDV` y son altamente significativas.

Para la $SS2$ utilirizaremos la funcion `Anova()` para obtener la suma de cuadrados

Tabla 5: Sumas de cuadrados tipo 2

	Sum Sq	Df	F value	Pr(>F)
CRIM	90.304276	1	7.5651738	0.0071517
NOX	135.228379	1	11.3286572	0.0011107
RM	1.271095	1	0.1064850	0.7449152
AGE	2.111003	1	0.1768477	0.6750666
PTRATIO	4.747061	1	0.3976815	0.5298357
LSTAT	383.573149	1	32.1335564	0.0000002
Residuals	1110.126202	93	NA	NA

De la Tabla 5 podemos concluir que:

Las variables CRIM, NOX y LSTAT son las variables con mayor impacto en la variabilidad de MEDV,en particular la variable LSTAT tiene la mayor contribución.

- Las variables `RM`, `AGE` y `PTRATIO` tienen una baja contribución con respecto a la variable `MEDV` una vez que se controlan las otras variables, lo que sugiere que pueden no ser necesarias en el modelo.

Punto 7

Construya y analice gráficos de los residuales estudentizados vs. Valores ajustados y contra las variables de regresión utilizadas. ¿Qué información proporcionan estas gráficas?

La Figura 2 nos dice que:

- E modelo es adecuado, ya que no hay patrones visibles (es decir, los residuales se distribuyen aleatoriamente alrededor de cero).Esto sugiere que los errores tienen una varianza constante (homocedasticidad) y que el modelo es lineal
- Aunque la mayoría de los puntos están cerca de la línea cero, hay algunos puntos que se encuentran bastante alejados (outliers), especialmente en la parte superior del gráfico.

De la Figura 3 podemos decir que:

- En la mayoría de los gráficos, los residuos se distribuyen de manera aleatoria alrededor de cero, lo que sugiere que el modelo es adecuado para estas variables. No hay patrones claros que sugieran problemas de heterocedasticidad o no linealidad.
- El gráfico de la variable `PTRATIO` con respecto a los residuales presenta unos posibles outliers, lo cual se debe investigar más a fondo para ver si es un error de medición o si representa un caso especial.

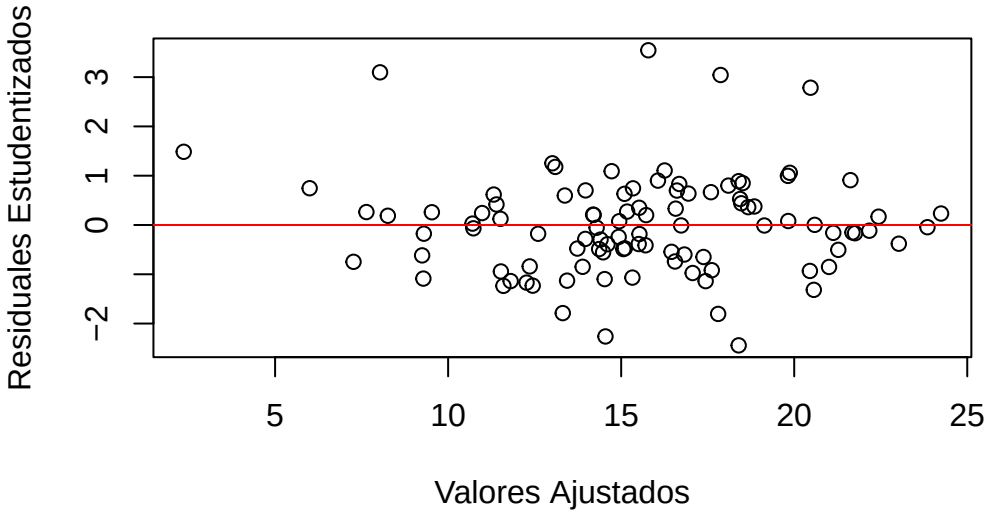


Figura 2: Residuales Estudentizados vs Valores ajustados

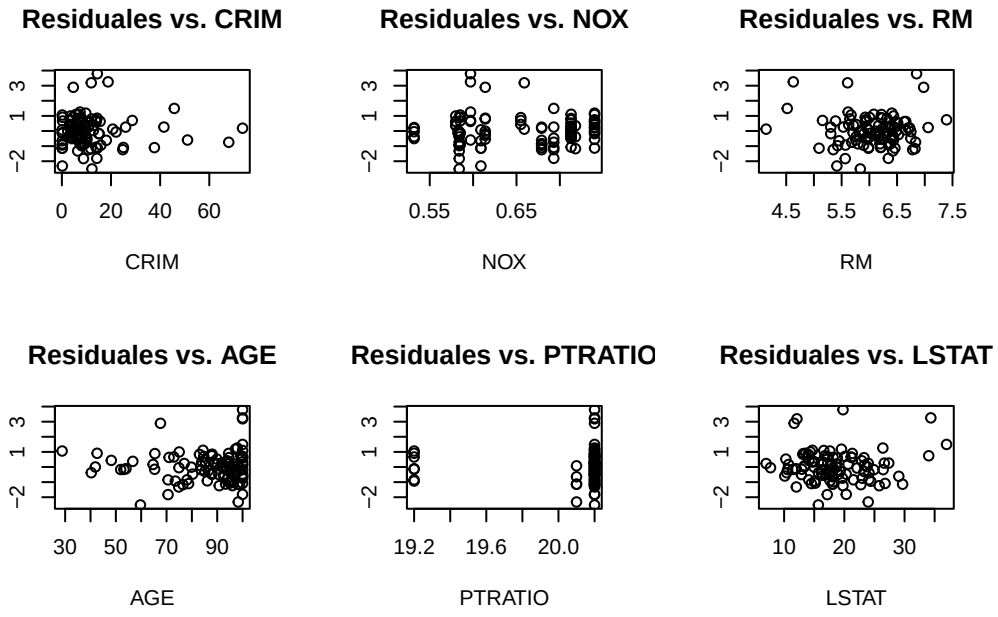


Figura 3: Residuales Estudentizados vs Cada Variable

- El gráfico de variable `RM` con respectoa los residuales tiene un rango más estrecho, pero igualmente los residuales parecen distribuidos aleatoriamente.

Punto 8

Construya una gráfica de probabilidad normal para los residuales estudentizados. ¿Existen razones para dudar de la hipótesis de normalidad sobre los errores en este modelo?

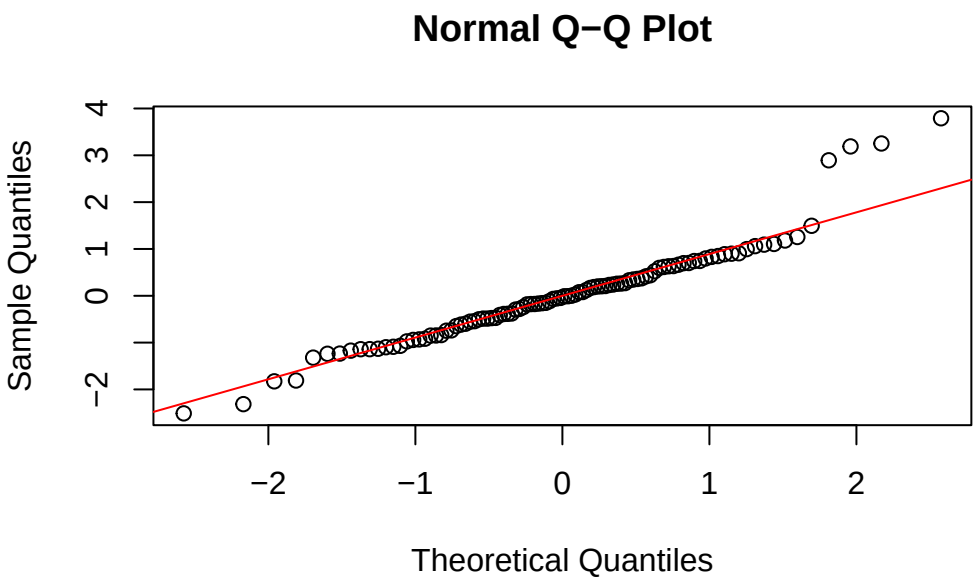


Figura 4: Gráfico Q-Q de Residuales Estudentizados

De la Figura 4 vemos que hay algunos puntos que se alejan significativamente de la línea recta (en particular en los extremos), lo cual hay posibles razonespara dudar de la normalidad de los errores.

Punto 9

Diagnostique la presencia de observaciones atípicas, de balanceo y/o influenciales y concluya.

Diagnóstico de observaciones atípicas

Para las observaciones atipicas se considerada atípica si su residuo estudentizado está fuera del rango [-2, 2] o [-3, 3] dependiendo del nivel de tolerancia. Para este caso consideraremos observacion atipicas si esta fuera del rango de [-3, 3].

```
outliers <- which(abs(residuales_estudentizados) > 3)
outliers
```

```
408 410 413
  8  10  13
```

En este caso, las observaciones 8, 10, 13, 408, 410 y 413 cumplen con este criterio, por tanto son posibles outliers.

Diagnóstico de observaciones de balanceo

Primeramente calculamos los valores de leverage (diagonales de la matriz H), también llamada matriz de proyección).Las observaciones con h_{ii} grandes y residuales r_i también grandes probablemente serán influenciales.

La observación i es un punto de balanceo si $hii > 2(k+1)/n$, pero si $2(k+1)/n > 1$, este criterio no funciona pues los h_{ii} siempre son menores que 1.

```
# Calcular valores de leverage
leverage <- hatvalues(modelo)

# Umbral para leverage alto
umbral_leverage <- 2 * ((length(coef(modelo)) + 1) / nrow(datos))

# Identificar observaciones con leverage alto
balanceo <- which(leverage > umbral_leverage)
print(balanceo)
```

```
401 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422
  1   4   5   6   7   8   9  10  11  12  13  14  15  16  17  18  19  20  21  22
423 424 425 427 428 430 431 432 433 435 436 437 438 439 440 441 442 443 444 445
 23  24  25  27  28  30  31  32  33  35  36  37  38  39  40  41  42  43  44  45
446 447 448 451 452 454 455 458 460 461 462 463 464 465 466 468 469 470 472 473
 46  47  48  51  52  54  55  58  60  61  62  63  64  65  66  68  69  70  72  73
474 475 476 477 478 481 482 483 484 485 486 488 489 490 491 492 493 494 495 496
```

```
 74  75  76  77  78  81  82  83  84  85  86  88  89  90  91  92  93  94  95  96
497 498 499 500
 97  98  99 100
```

Las observaciones 406, 407, 411, 415, 419, 494, 495, 496, 6, 7, 11, 15, 19, 95 y 96 tienen valores de leverage altos, es decir, estos puntos tienen una influencia considerable en el ajuste del modelo debido a su posición en el espacio de las variables predictoras.

Diagnóstico de observaciones influenciales

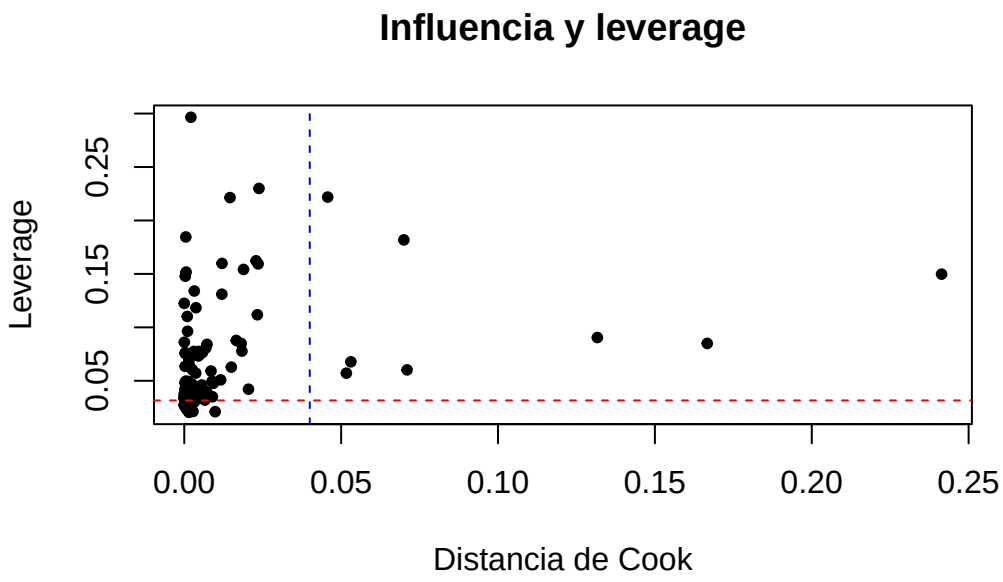
Para este diagnostico utlizaremos la distacia de Cook lo uqe hace es medir la influencia de la observación i sobretodos los valores ajustados de la respuesta, para $i = 1, 2, ..., n$.Una observación es influyente si su distancia de Cook supera el umbral de $4/n$.

```r
# Calcular la distancia de Cook
cook <- cooks.distance(modelo)

# Umbral de influencia
umbral_cook <- 4 / nrow(datos4)

# Identificar observaciones influyentes
influenciales <- which(cook > umbral_cook)
influenciales
```

```
408 410 413 415 427 474 490 496
  8  10  13  15  27  74  90  96
```

Estas observaciones tienen un impacto considerable en el modelo. Se debe investigar para decidir si se deben eliminar, ajustar, o conservar, dependiendo de si representan errores o datos extremos válidos. Si representan errores o casos extremos no representativos de la población general, podrían excluirse para mejorar la precisión del modelo. Sin embargo, si son casos válidos, es importante reconocer su impacto y considerar métodos robustos que puedan reducir su influencia sin eliminarlas.

```r
# Gráfico de leverage vs. distancia de Cook
plot(cook, leverage, main = "Influencia y leverage",
     xlab = "Distancia de Cook", ylab = "Leverage", pch = 20)
abline(h = umbral_leverage, col = "red", lty = 2)  # Umbral de leverage
abline(v = umbral_cook, col = "blue", lty = 2)    # Umbral de Cook
```



Punto 10

Suponga que se establece que hay un error de digitación en un máximo de 10 observaciones. Ajuste el modelo sin esas observaciones y presente solo la tabla de parámetros ajustados resultante. ¿Cambian notablemente las estimaciones de los parámetros, sus errores estándar y/o la significancia? ¿Qué concluye al respecto? Evalúe el gráfico de normalidad para los residuales studentizados de este ajuste. ¿Mejoró la normalidad? Concluya sobre los efectos de estas observaciones.

Supongamos que tenemos 10 observaciones que se consideran como errores de digitación, son las 10 ultimas filas.

```
# se eliminaron las 10 ultimas observacione
datos_limpios <- datos4[-c(91:100), ]
```

```
# Ajustar el modelo sin las observaciones problemáticas
modelo_ajustado <- lm(MEDV ~., data = datos_limpios)
```

Tabla 6: Resumen del Nuevo Ajuste del Modelo

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-1257.86511	523.86822	-2.40111	0.01858
CRIM	-0.12208	0.03302	-3.69736	0.00039
NOX	-26.45174	6.56241	-4.03080	0.00012
RM	-0.48198	0.82745	-0.58250	0.56181
AGE	0.03297	0.03258	1.01191	0.31452
PTRATIO	64.41754	26.00079	2.47752	0.01526
LSTAT	-0.48668	0.08398	-5.79529	0.00000

De la Tabla 6 notamos que si hubo cambios notables las estimaciones de los parámetros en sus errores estándar algunas variables no sufrieron cambios notables , pero otras si como es el caso de la variable `PTRATIO` y con respecto a la significancia la variable que sufrio cambios en su significancia fue `PTRATIO`.

```
residuales_estudentizados_ajustado <- rstudent(modelo_ajustado)
```

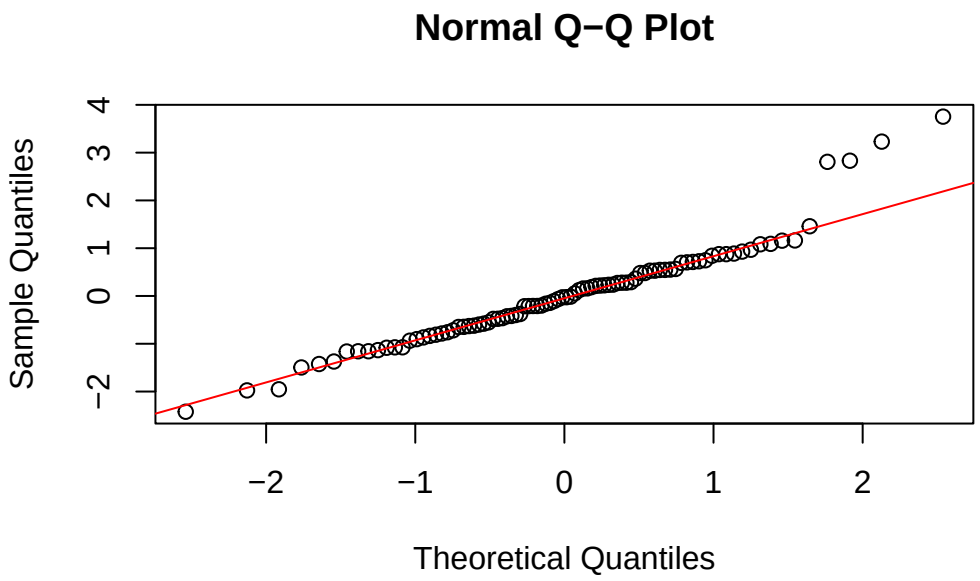


Figura 5: Gráfico Q-Q de Residuales Estudentizados Nuevo Ajuste

De la Figura 5 vemos que hay algunos puntos que se alejan significativamente de la línea recta (en particular en los extremos), lo cual hay posibles razones para dudar de la normalidad de

los errores, por tanto no mejoro la normalidad.

Podemos concluir que, en este caso, las observaciones problemáticas no tuvieron una influencia significativa en los resultados del modelo, ya que las estimaciones de los parámetros , sus errores estándar sufrieron cambios pero no tan notables excepto en una variable y l a significancia estadísticapermanecieron prácticamente inalterados tras su eliminación. Esto nos sugiere que el modelo es robusto frente a dichas observaciones.