

Tarea: Modelos de regresión lineal simple y múltiple

Javier Herrero Pérez, Jose Aguilar Milla

2025-12-05

Ejercicio 1

```
library(readxl)
df <- read_excel("boston.xlsx")
```

Apartado 1

Se busca estudiar crim de manera lineal con la variable que tenga mayor relación con esta, para ello se va a estudiar la matriz de correlación eliminando la variable crim:

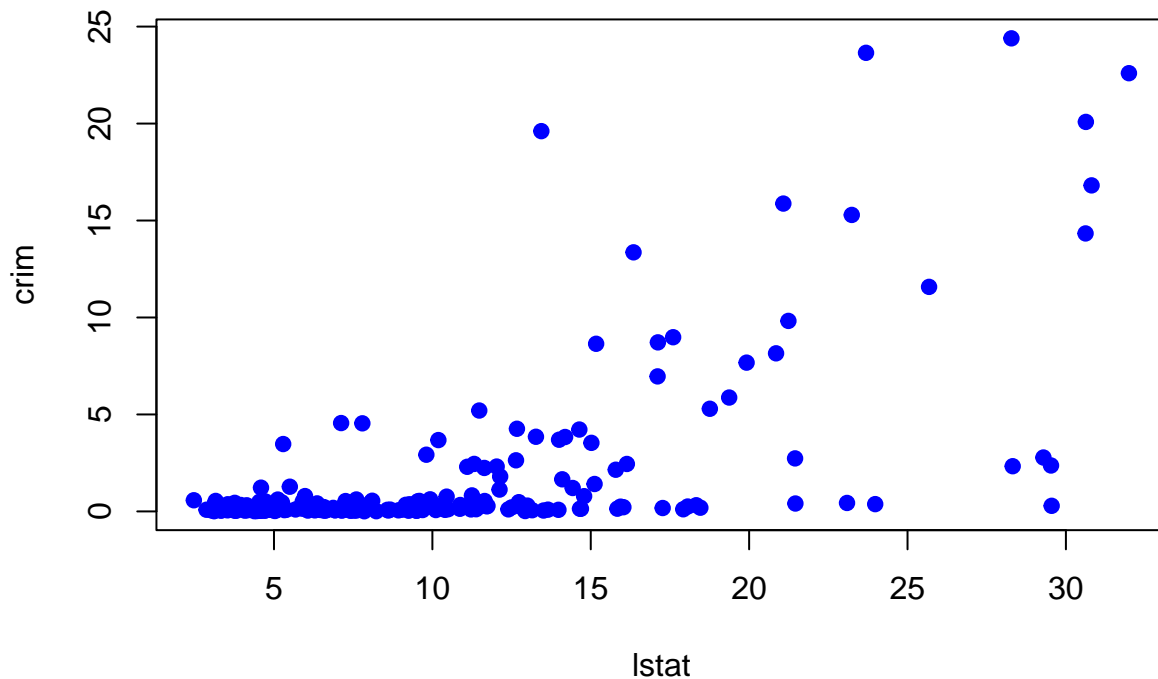
```
cor_matrix <- cor(df$crim, df[, -which(names(df) == "crim")])
cor_matrix
```

```
##           indus        chas        nox        rm        age        dis  ptratio
## [1,] 0.5790421 -0.05326154 0.5206876 -0.323822 0.448534 -0.4633303 0.4261423
##           black      lstat      medv
## [1,] -0.07510281 0.6391068 -0.5231014
```

Se observa que la variable que tiene una mayor relación lineal es lstat, por lo que se entrena el modelo con esta variable:

```
plot(df$lstat, df$crim,
     main = "Diagrama de Dispersión: crim vs. lstat",
     xlab = "lstat",
     ylab = "crim",
     pch = 19, col = "blue")
```

Diagrama de Dispersión: crim vs. lstat



la relación es positiva y moderada. Gráficamente, se observa una tendencia creciente, aunque con mucha dispersión, especialmente para valores altos de crim.

Apartado 2

```
lm1<-lm(crim~lstat,data=df)
summary(lm1)
```

```
##
## Call:
## lm(formula = crim ~ lstat, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9.7449 -1.4171 -0.1196  1.0338 16.4752
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -2.62305    0.45836  -5.723 3.86e-08 ***
## lstat        0.42834    0.03673  11.663 < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.401 on 197 degrees of freedom
## Multiple R-squared:  0.4085, Adjusted R-squared:  0.4055
## F-statistic: 136 on 1 and 197 DF, p-value: < 2.2e-16
```

Se observa que la variable `lstat` es estadísticamente significativa por lo que se puede rechazar la hipótesis nula de $\beta = 0$ para esa variable. En cuanto al coeficiente de correlación es de 0.4085, explicando el 40% de la dispersión de los datos. El p valor asociado a F-statistic es muy pequeño, confirmando que el modelo es globalmente significativo.

Apartado 3

El intervalo de confianza al 90%:

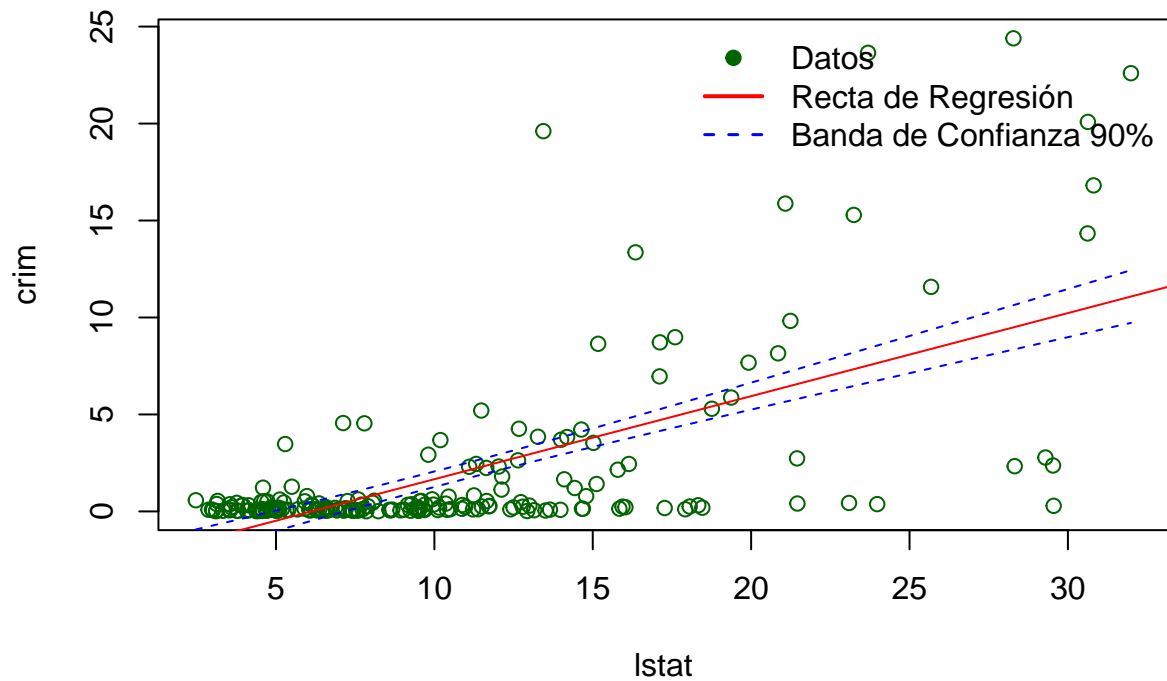
```
confint(lm1,level=0.90)

##              5 %          95 %
## (Intercept) -3.3805543 -1.8655501
## lstat       0.3676485  0.4890385

new_data <- data.frame(lstat = seq(min(df$lstat), max(df$lstat), length.out = 100))
confidence_intervals <- predict(lm1, newdata = new_data, interval = "confidence", level = 0.9)

plot(df$lstat, df$crim,
     main = "Regresión Simple: crim vs. lstat con Bandas de Confianza (90%)",
     xlab = "lstat",
     ylab = "crim", col = "darkgreen")
abline(coef=coef(lm1), col='RED')
lines(new_data$lstat, confidence_intervals[, "lwr"], col = "blue", lty = 2)
lines(new_data$lstat, confidence_intervals[, "upr"], col = "blue", lty = 2)
legend("topright",
     legend = c("Datos", "Recta de Regresión", "Banda de Confianza 90%"),
     col = c("darkgreen", "red", "blue"),
     pch = c(19, NA, NA),
     lty = c(NA, 1, 2),
     lwd = c(NA, 2, 1.5),
     bty = "n")
```

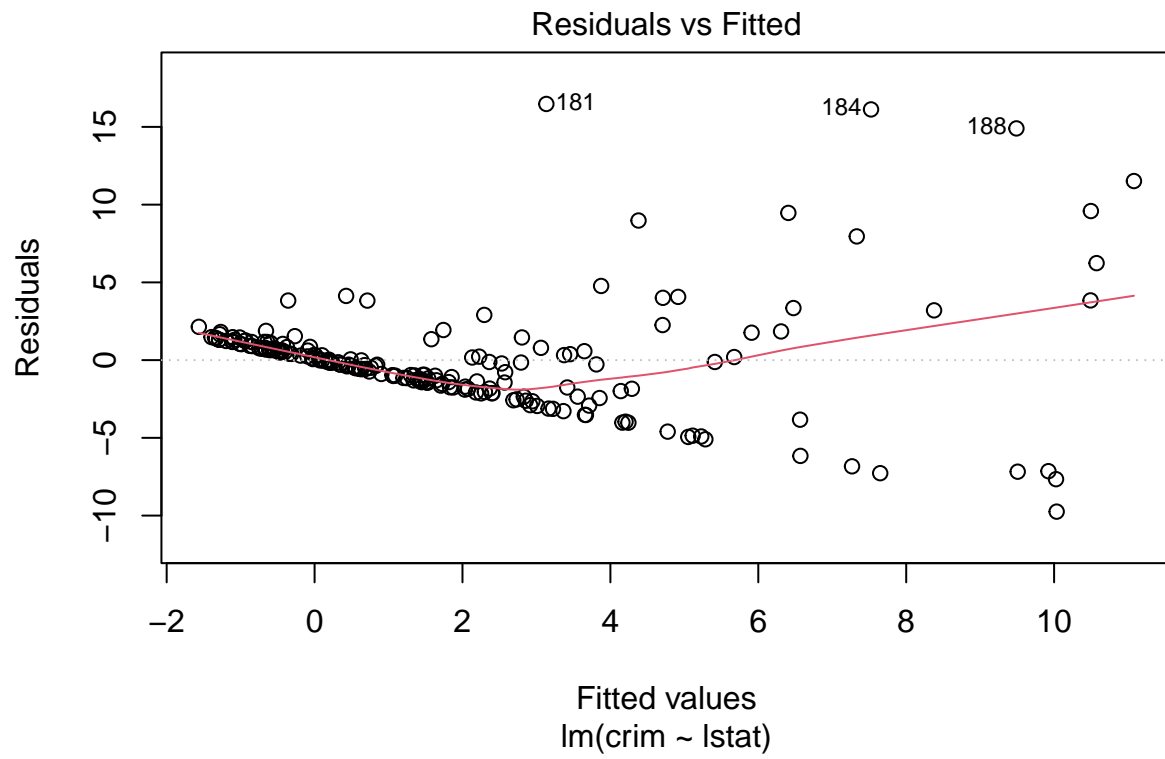
Regresión Simple: crim vs. lstat con Bandas de Confianza (90%)

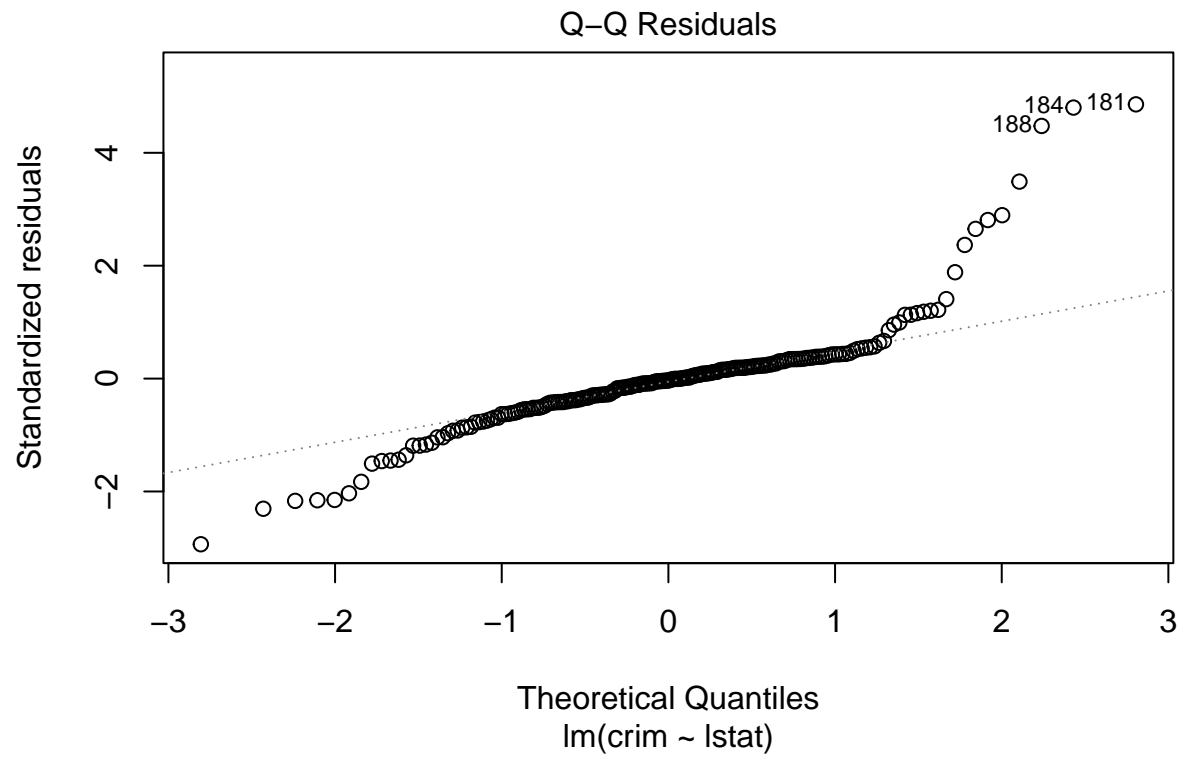


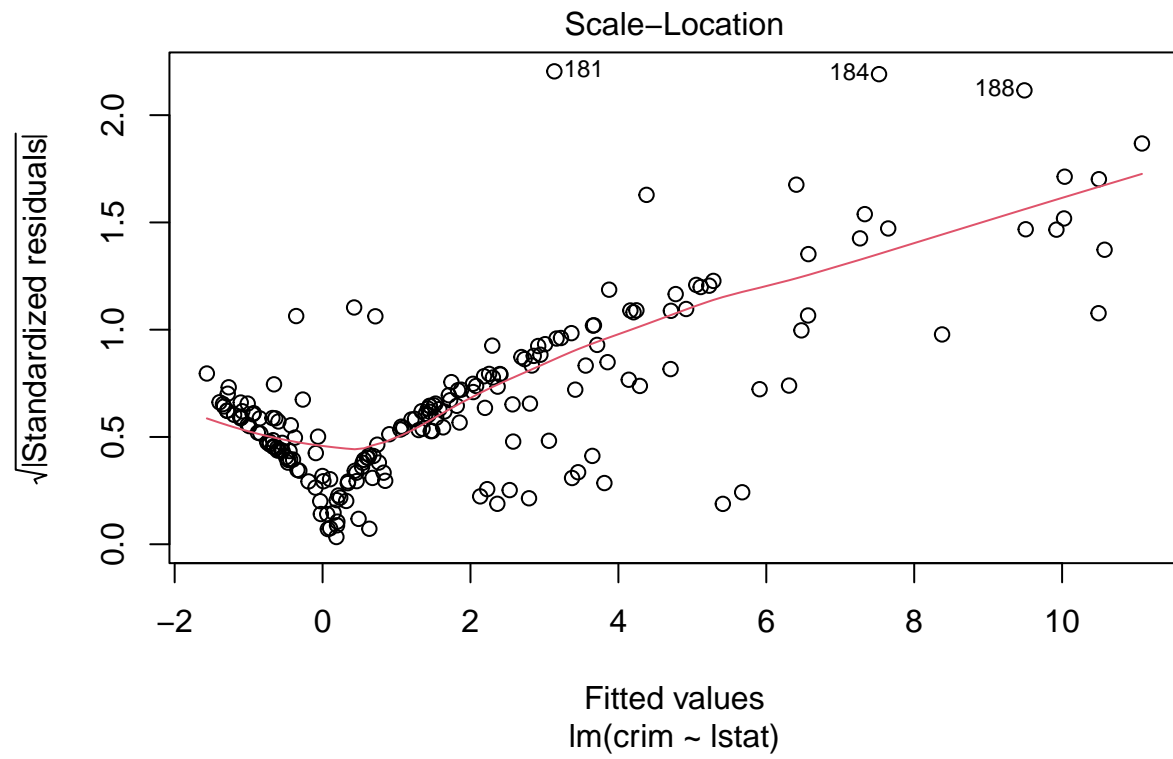
Se observa que la banda de confianza que estima el rango para la media de crim dado un valor de lstat, es relativamente estrecha. Sin embargo, una gran cantidad de valores individuales están excluidos de esta banda. Esto es una consecuencia directa de una varianza residual alta.

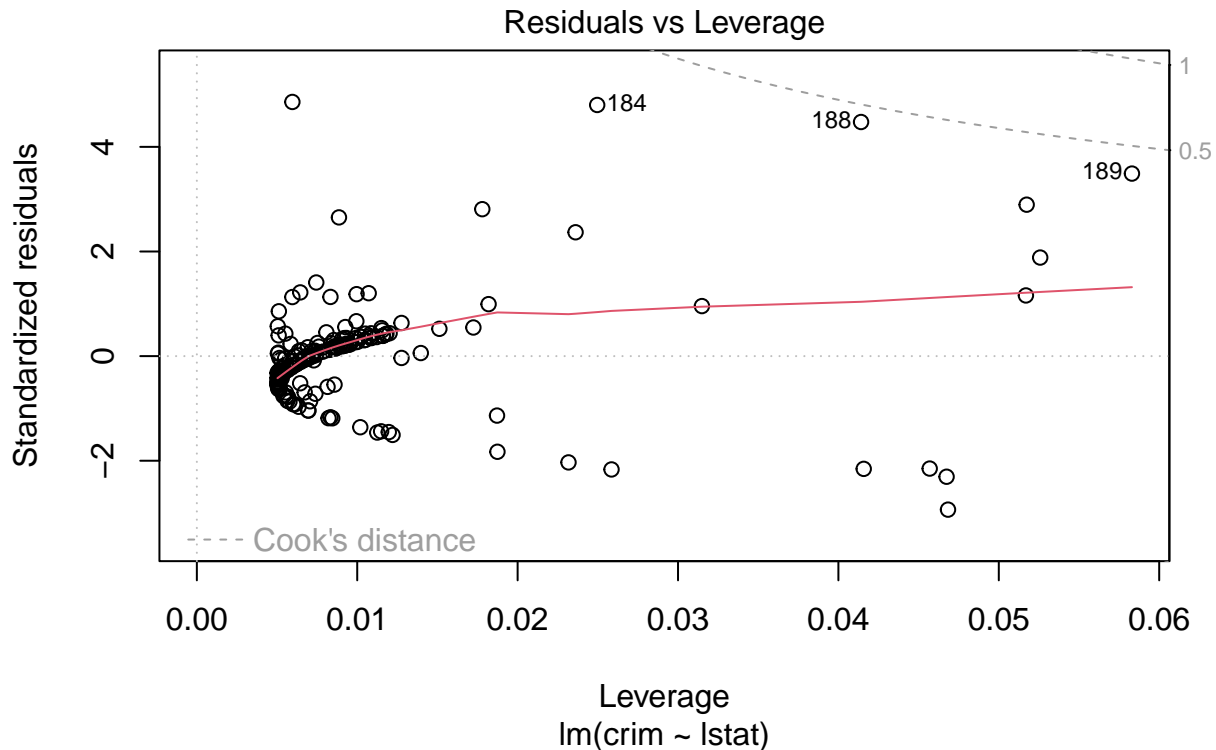
Apartado 4

```
plot(lm1)
```









Tras los resultados obtenidos se puede concluir que los residuos son heterocedasticos, debido a que la dispersión de los residuos no es constante. En la figura de Residuals vs Fitted se aprecia como la línea roja que corresponde al valor medio no es constante si no que presenta un patrón curvo.

Estos resultados justifican el hecho de hacer una transformación logarítmica. También se va a decidir eliminar unos valores outliers, lo que va a mejorar el valor de R^2 y el análisis:

```
df <- read_excel("boston.xlsx")
df<-df[-c(60,178,181),]
df<-df[-c(173,177,127),]
lm1<-lm(log(crim)~lstat,data=df)
summary(lm1)
```

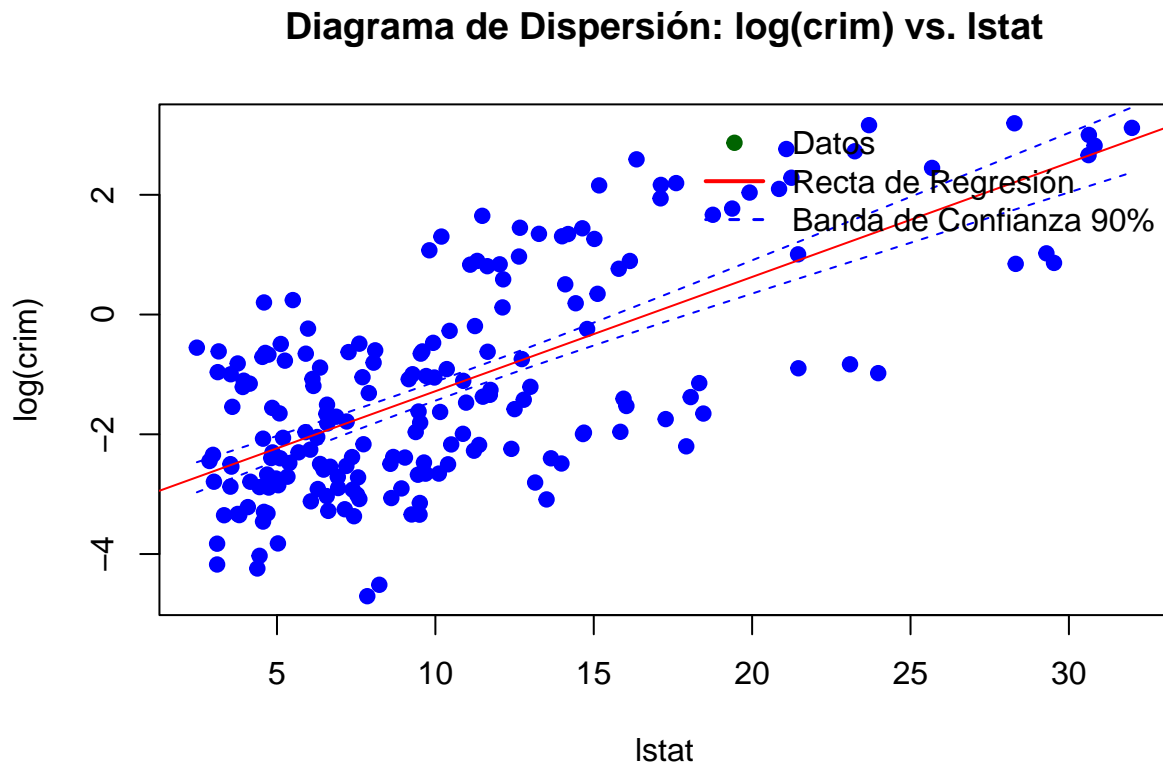
```
##
## Call:
## lm(formula = log(crim) ~ lstat, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -3.01307 -1.08914 -0.04107  1.04539  2.66068
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -3.18914    0.18167  -17.55  <2e-16 ***
## lstat        0.19087    0.01466   13.02  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```



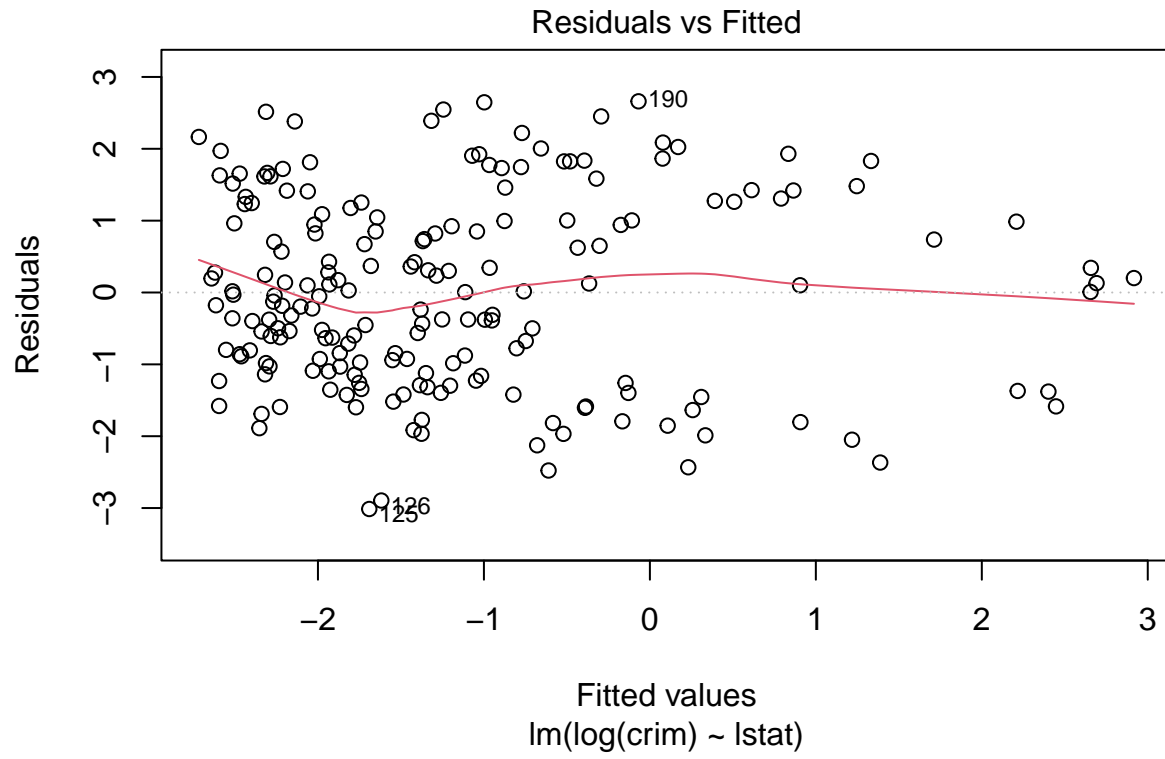
```
##
## Residual standard error: 1.324 on 191 degrees of freedom
## Multiple R-squared:  0.4702, Adjusted R-squared:  0.4674
## F-statistic: 169.5 on 1 and 191 DF,  p-value: < 2.2e-16
```

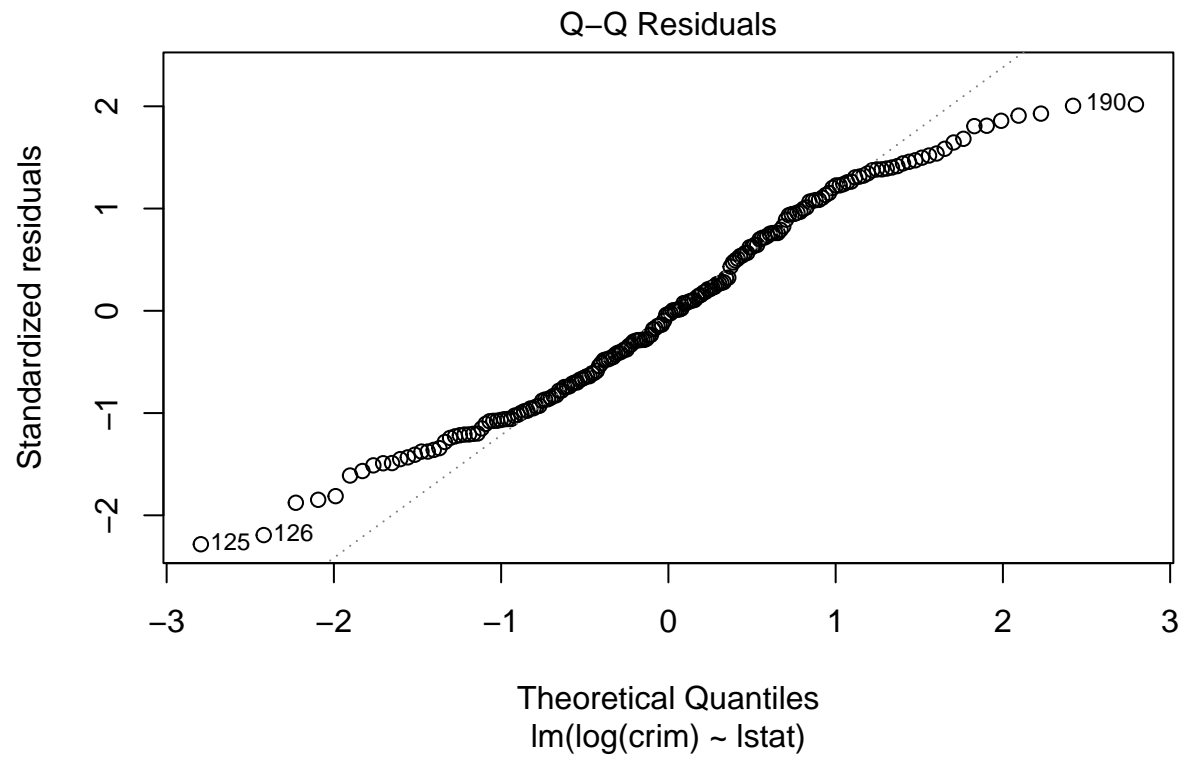
Se ha decidido aplicar el logaritmo únicamente a crim ya que si se aplicaba a ambos el valor de R^2 disminuía.

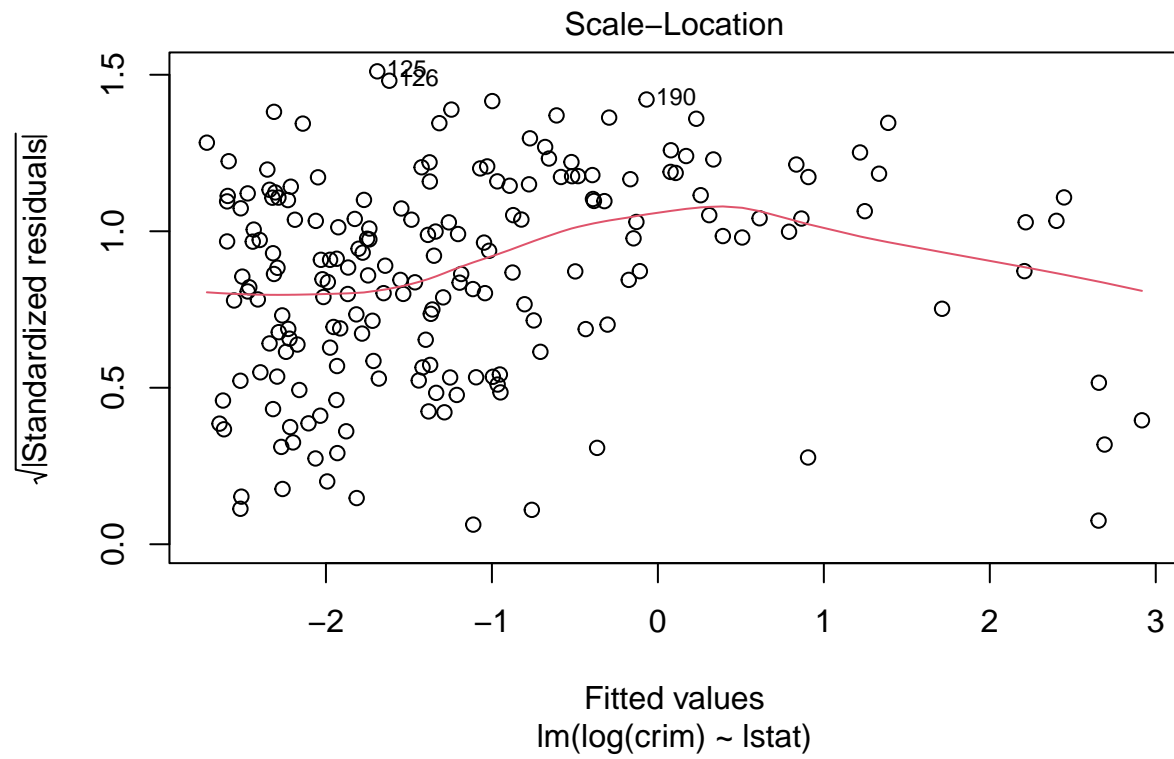
```
new_data <- data.frame(lstat = seq(min(df$lstat), max(df$lstat), length.out = 100))
confidence_intervals <- predict(lm1, newdata = new_data, interval = "confidence", level = 0.9)
plot(df$lstat, log(df$crim),
     main = "Diagrama de Dispersión: log(crim) vs. lstat",
     xlab = "lstat",
     ylab = "log(crim)",
     pch = 19, col = "blue")
abline(coef=coef(lm1), col='RED')
lines(new_data$lstat, confidence_intervals[, "lwr"], col = "blue", lty = 2)
lines(new_data$lstat, confidence_intervals[, "upr"], col = "blue", lty = 2)
legend("topright",
     legend = c("Datos", "Recta de Regresión", "Banda de Confianza 90%"),
     col = c("darkgreen", "red", "blue"),
     pch = c(19, NA, NA),
     lty = c(NA, 1, 2),
     lwd = c(NA, 2, 1.5),
     bty = "n")
```

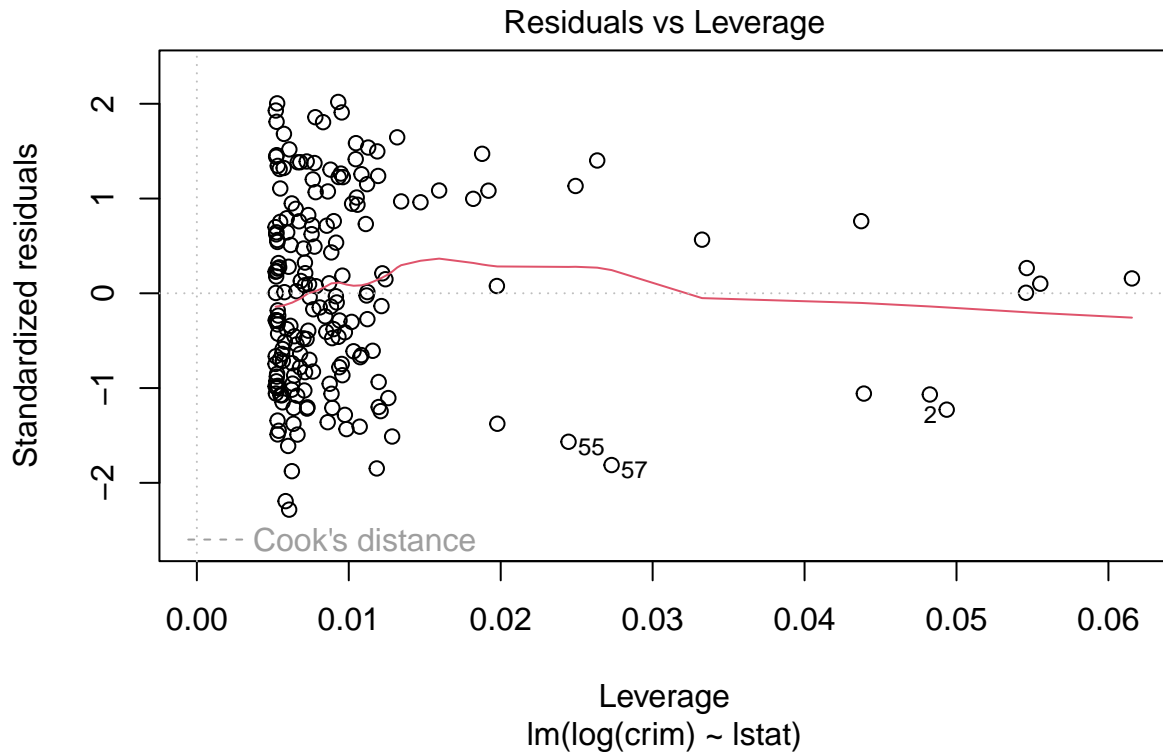


```
plot(lm1)
```









Se observa un mejor ajuste tras la transformación y la eliminación de algunos valores. Ahora la media de los residuos se encuentra en torno al cero y el QQ Residuals presenta una tendencia más clara.

Ejercicio 2

Punto 1

```
df <- read_excel("boston.xlsx")
lm2 <- lm(crim ~ medv, data = df)
```

Punto 2

```
summary(lm2)
```

```
##
## Call:
## lm(formula = crim ~ medv, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -4.1341 -2.4009 -1.0186  0.7805 18.5528
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept)   8.5183     0.8108  10.506 < 2e-16 ***
```

```
## medv          -0.2550      0.0296  -8.615 2.26e-15 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 3.769 on 197 degrees of freedom
## Multiple R-squared:  0.2736, Adjusted R-squared:  0.2699
## F-statistic: 74.21 on 1 and 197 DF,  p-value: 2.258e-15
```

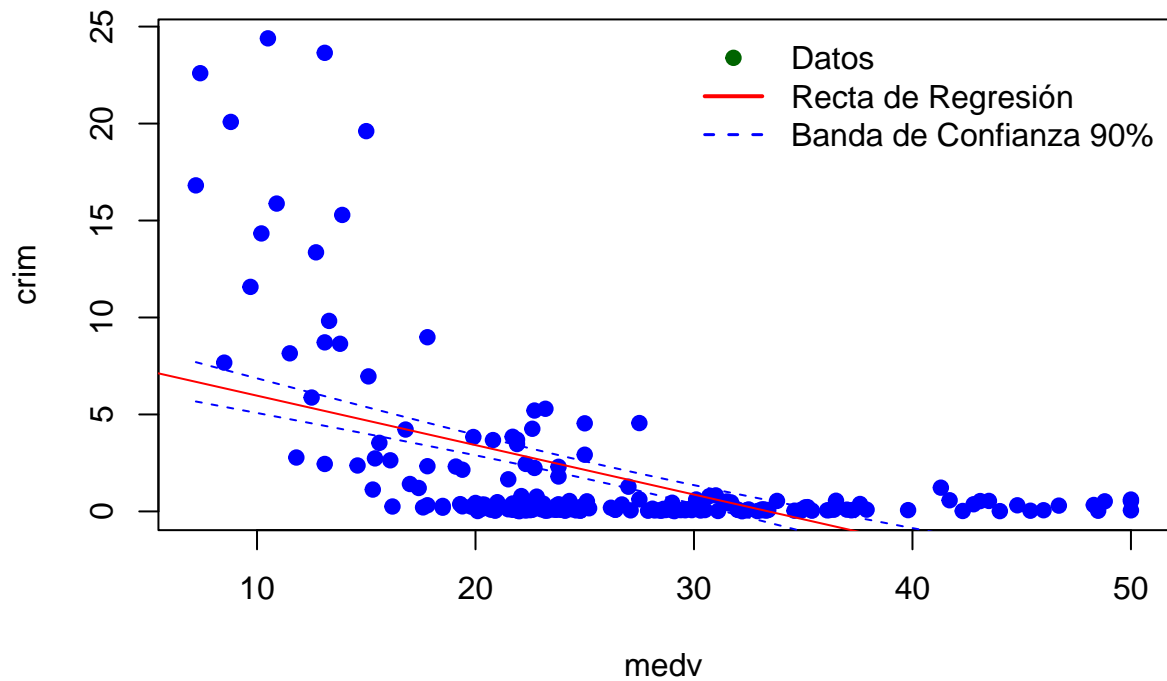
La variable medv es estadísticamente significativa para explicar la variable crim. Por otro lado el p-value de F-statistic es muy bajo por lo que se rechaza la hipótesis nula demostrando que el modelo es globalmente significativo. Este modelo es capaz de captar el 27,36% de la varianza de los datos, aunque es un valor bajo más adelante se desarrollaran transformaciones que subiran este valor. La recta mínima de los cuadrados viene dada por $y = -0.255\beta_1 + 8.518$ por lo que a mayor sea el precio medio de la vivienda menor será la tasa de criminalidad.

Punto 3

```
new_data<-data.frame(medv=seq(min(df$medv),max(df$medv),length.out=100))
coefidence_intervals<-predict(lm2,newdata=new_data,interval = "confidence",level=0.9)

plot(df$medv,df$crim,
     main = "Diagrama de Dispersión: crim vs. medv",
     xlab = "medv",
     ylab = "crim",
     pch = 19, col = "blue")
abline(coef=coef(lm2), col='RED')
lines(new_data$medv, coefidence_intervals[, "lwr"], col = "blue", lty = 2)
lines(new_data$medv, coefidence_intervals[, "upr"], col = "blue", lty = 2)
legend("topright",
     legend = c("Datos", "Recta de Regresión", "Banda de Confianza 90%"),
     col = c("darkgreen", "red", "blue"),
     pch = c(19, NA, NA),
     lty = c(NA, 1, 2),
     lwd = c(NA, 2, 1.5),
     bty = "n")
```

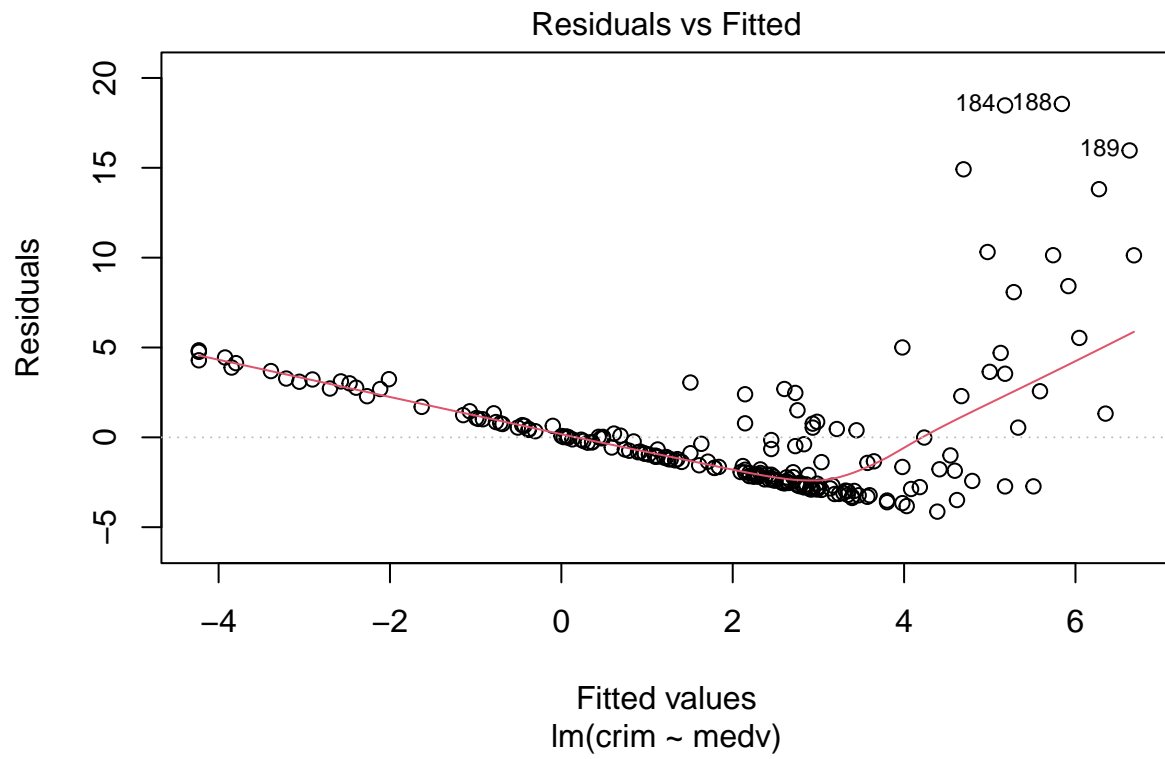
Diagrama de Dispersión: crim vs. medv

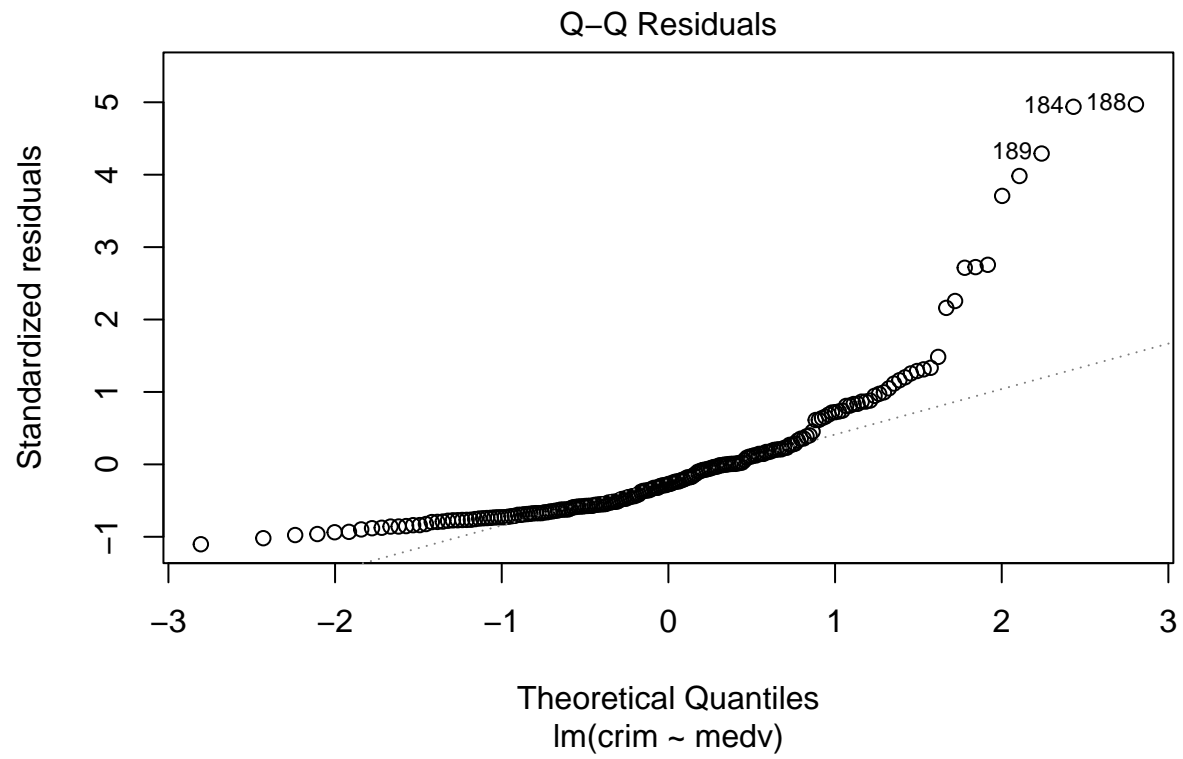


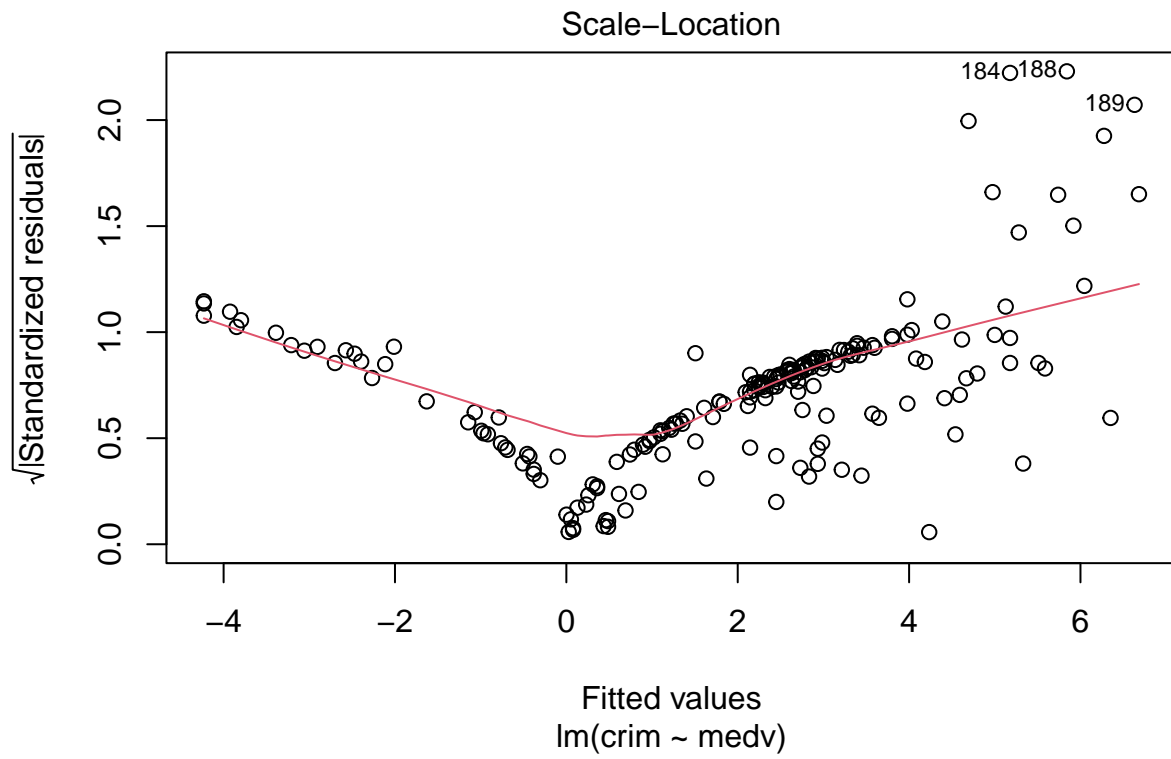
Se observa que regresión no es del todo buena para explicar el conjunto de datos. Veamos a continuación los residuos

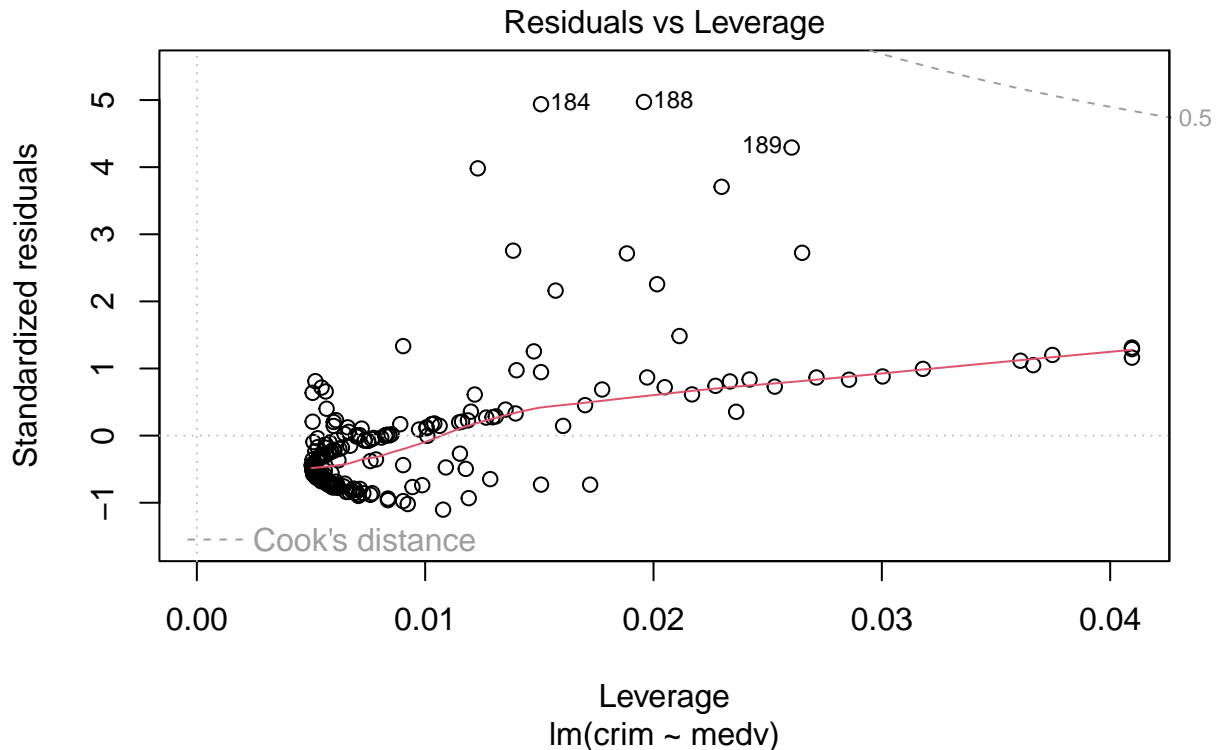
Punto 4

```
plot(lm2)
```









Se observa cómo los residuos presentan heterocedasticidad demostrando que el modelo no es capaz de captar toda la información que hay en el conjunto de datos. En cuanto al QQ-residuals también se observa que no sigue el comportamiento deseado.

Punto 5

Como se ha visto no es adecuado utilizar una regresión lineal, se va a probar a describirlo mediante un polinomio de grado dos para la variable medv. para solucionar el problema de heterocedasticidad en los residuos se va a estudiar el logaritmo de crim. También se ha decidido eliminar una serie de valores outliers para mejorar el valor de R^2 como se muestra en el siguiente chunk:

```
df <- read_excel("boston.xlsx")
df <- df[-c(127,128,34,50),]
lm2 <- lm(log(crim) ~ (medv) + I((medv)**2), data=df)
summary(lm2)
```

```
##
## Call:
## lm(formula = log(crim) ~ (medv) + I((medv)^2), data = df)
##
## Residuals:
```

	Min	1Q	Median	3Q	Max
	-2.5453	-0.8791	-0.1476	0.8548	3.4657

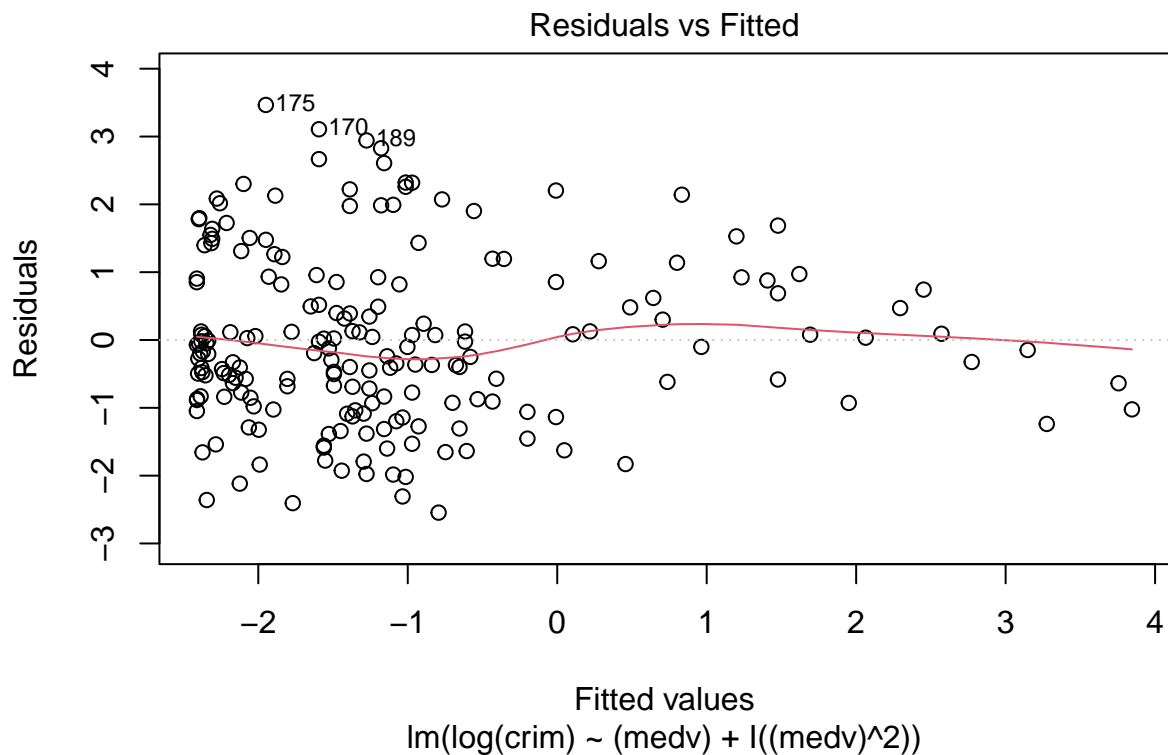
```
##
## Coefficients:
```

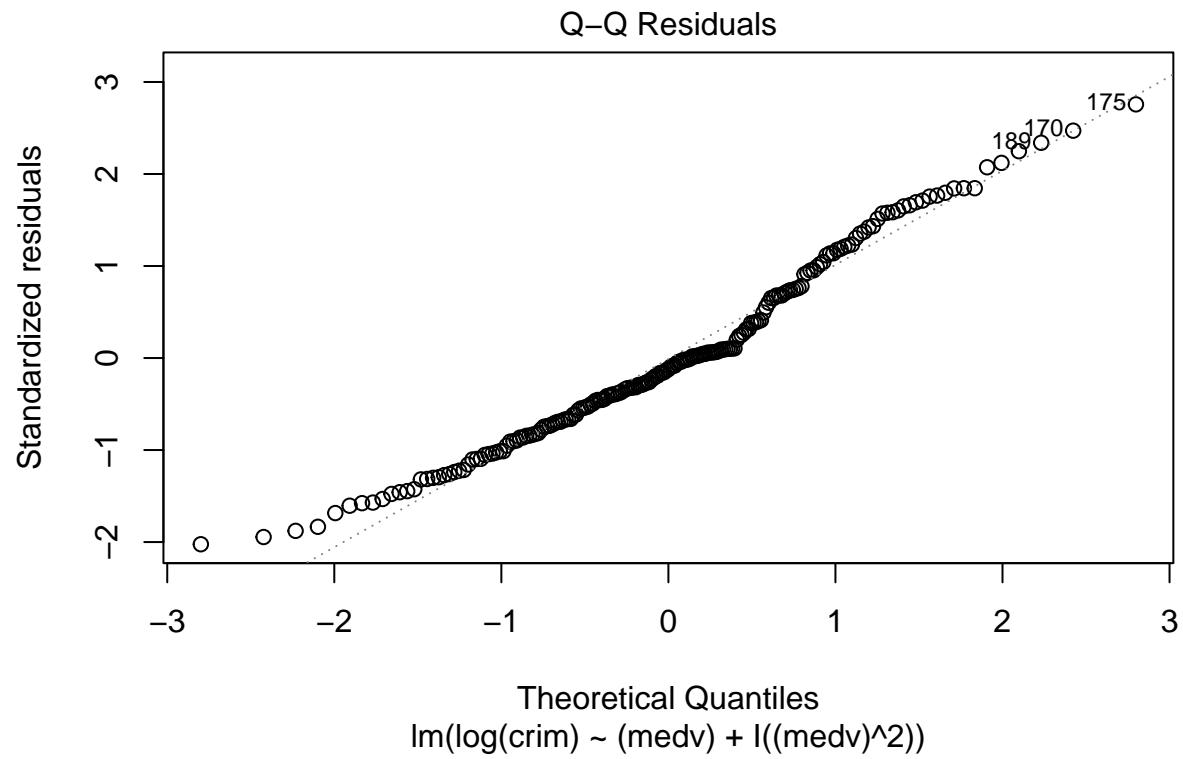
	Estimate	Std. Error	t value	Pr(> t)
(Intercept)	7.4947677	0.6475937	11.573	<2e-16 ***
medv	-0.5648700	0.0477090	-11.840	<2e-16 ***

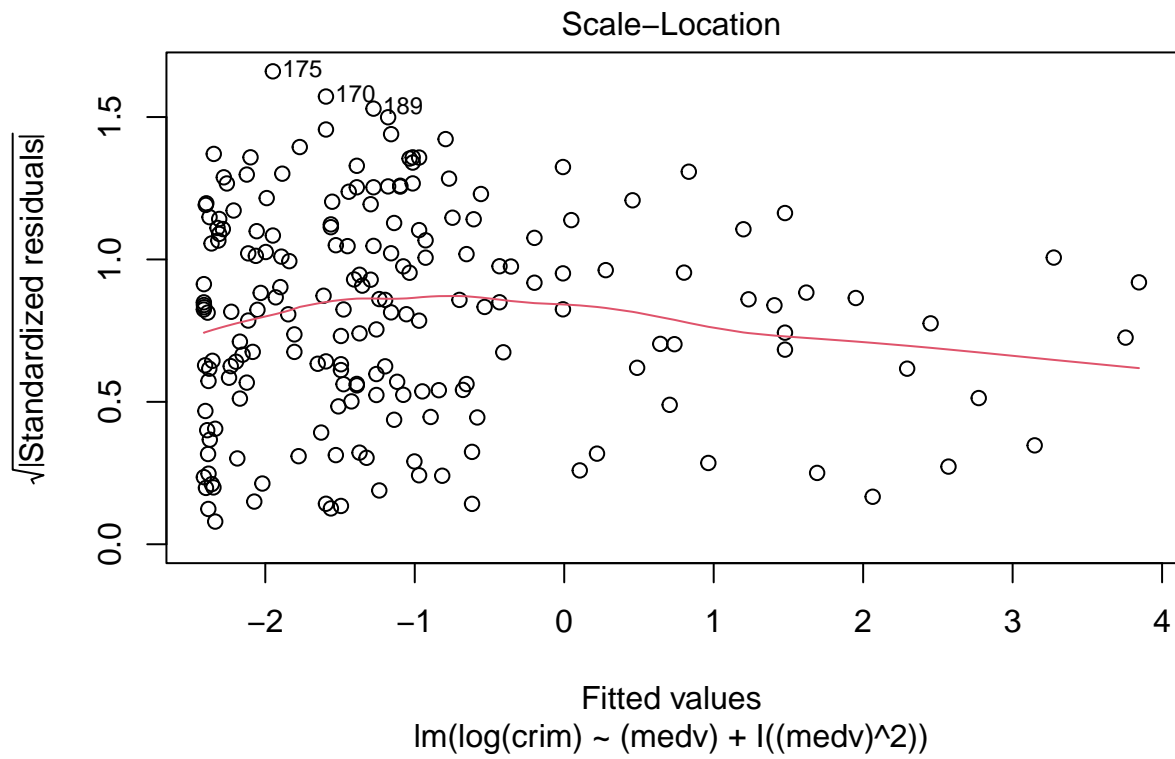
```
## I((medv)^2) 0.0080528 0.0008307 9.694 <2e-16 ***
## ---
## Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.262 on 192 degrees of freedom
## Multiple R-squared: 0.5285, Adjusted R-squared: 0.5236
## F-statistic: 107.6 on 2 and 192 DF, p-value: < 2.2e-16
```

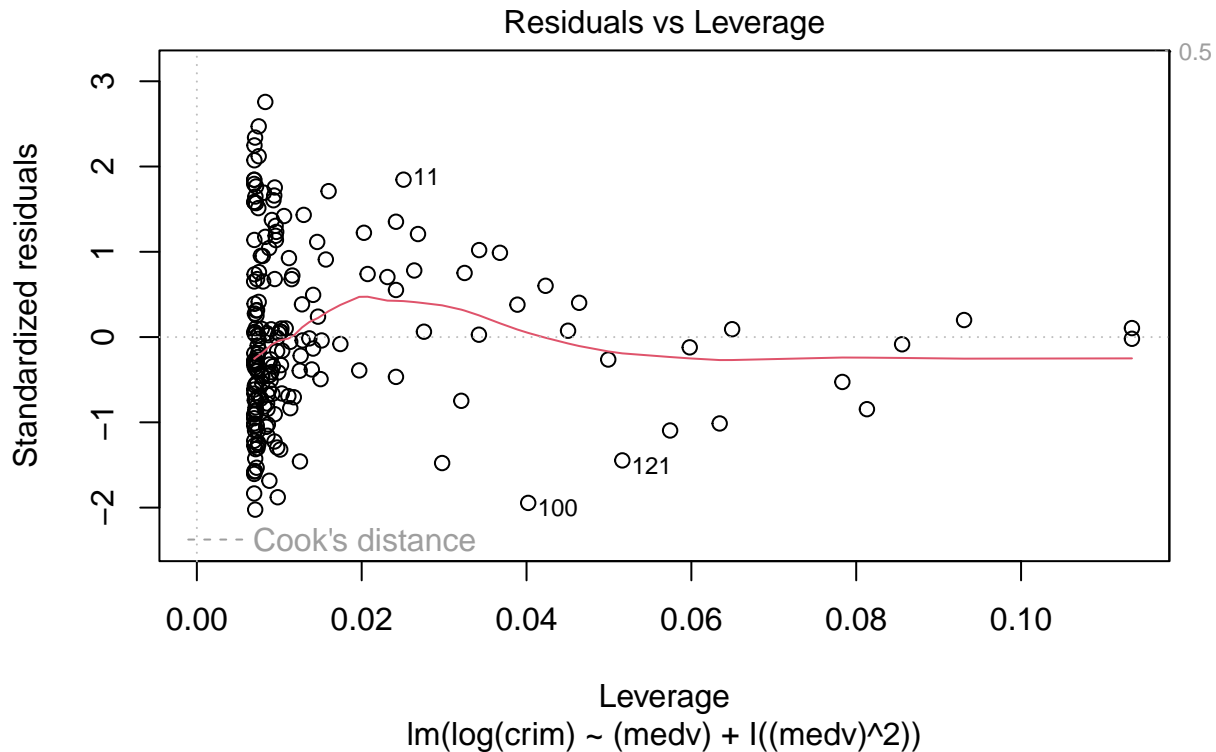
Con este nuevo ajuste se observa que tanto medv como $I((medv)^2)$ tienen significancia estadística con un p value muy bajo. El valor de F-statistic sigue siendo bajo. También se observa que el valor de R^2 a aumentando hasta 0.5285, es decir, este nuevo modelo explica un 25.5% más de varianza que el modelo anterior, lo cual es una mejora significativa.

```
plot(lm2)
```









Se observa ahora que los residuos presentan un mejor comportamiento que para el modelo anterior, siendo más homocedásticos y con la media con valores en torno al 0. Aunque todavía se observa mucha acumulación de puntos para valores bajos, se ha conseguido mejorar el modelo. En cuanto al QQ-residuals ahora sigue un comportamiento más parecido al deseado.

Apartado 6

```
medv_prediccion <- data.frame(medv = c(10, 30, 100))
IC_log_crim <- predict(lm2, newdata=medv_prediccion, interval = "confidence", level=0.90)
IP_log_crim <- predict(lm2, newdata=medv_prediccion, interval = "prediction", level=0.90)
```

```
IC_crim <- exp(IC_log_crim)
IP_crim <- exp(IP_log_crim)
print("Intervalo de confianza:")
```

```
## [1] "Intervalo de confianza:"
```

```
print(IC_crim)
```

```
##          fit          lwr          upr
## 1 1.417309e+01 9.019219e+00 2.227204e+01
## 2 1.103784e-01 9.048294e-02 1.346485e-01
## 3 4.962122e+13 4.053690e+10 6.074132e+16
```

```
print("Intervalos de predicción:")
```

```
## [1] "Intervalos de predicción:"
```

```
print(IP_crim)
```

```
##           fit           lwr           upr
## 1 1.417309e+01 1.676449e+00 1.198225e+02
## 2 1.103784e-01 1.357461e-02 8.975129e-01
## 3 4.962122e+13 3.003750e+10 8.197304e+16
```

Se observa que para 10k y 30k dolares el precio es lo esperado debido a que estos valores se encuentran dentro del rango del conjunto de datos, siendo más sencillo calcularlo. Sin embargo, para el caso de 100k el valor se dispara debido a que se está extrapolando y sale del orden de 10^{13} con intervalos tan grandes debido al uso del modelo cuadrático y la transformación logarítmica. Por lo tanto, al tener buscar estudiar un valor tan lejos del rango de datos con el que se entrena el modelo provoca que el error aumente siendo un resultado poco significativo y con mucho error, como se observa en los intervalos de predicción y de confianza.

Ejercicio 3

Apartado 1

La metodología RegSubsets consiste en ajustar todos los posibles modelos de regresión lineal que se pueden contruir con k variables predictoras donde k va de 1 hasta el número de variables. Para cada número de variables se selecciona el mejor modelo utilizando como criterio R^2 , BIC y CP.

Para el análisis se va a utilizar log(crim), ya que produce el modelo más robusto.

```
library(leaps)
```

```
## Warning: package 'leaps' was built under R version 4.5.2
```

```
library(readxl)
df <- read_excel("boston.xlsx")
regfit<-regsubsets(log(crim)~.,data=df)
res.summary <- summary(regfit)
res.summary
```

```
## Subset selection object
## Call: regsubsets.formula(log(crim) ~ ., data = df)
## 10 Variables (and intercept)
##           Forced in Forced out
## indus      FALSE      FALSE
## chas       FALSE      FALSE
## nox        FALSE      FALSE
## rm         FALSE      FALSE
## age        FALSE      FALSE
## dis        FALSE      FALSE
## ptratio    FALSE      FALSE
## black      FALSE      FALSE
## lstat      FALSE      FALSE
## medv       FALSE      FALSE
## 1 subsets of each size up to 8
## Selection Algorithm: exhaustive
##           indus chas nox rm  age dis ptratio black lstat medv
## 1  ( 1 ) " "   " "  "*" " " " " " " " " " " " "
## 2  ( 1 ) " "   " "  "*" " " " " " " "*" " " "
## 3  ( 1 ) "*"   " "  "*" " " " " " " "*" " " "
## 4  ( 1 ) "*"   " "  "*" " " "*" " " "*" " " "
## 5  ( 1 ) "*"   "*"  "*" " " "*" " " "*" " " "
```



```
## 6 ( 1 ) "*" "*" "*" "*" " " " " "*" " " "*" " "
## 7 ( 1 ) "*" "*" "*" "*" "*" " " "*" " " "*" " "
## 8 ( 1 ) "*" "*" "*" "*" "*" "*" "*" " " "*" " "
```

Para 8 variables se han seleccionado todas menos nox y black.

```
par(mfrow = c(1, 3))
```

```
# Gráfico  $R^2$  ajustado
```

```
plot(res.summary$adjr2, xlab = "Número de Variables", ylab = " $R^2$  Ajustado", type = "l")
```

```
points(which.max(res.summary$adjr2), res.summary$adjr2[which.max(res.summary$adjr2)], col = "red", cex = 2)
```

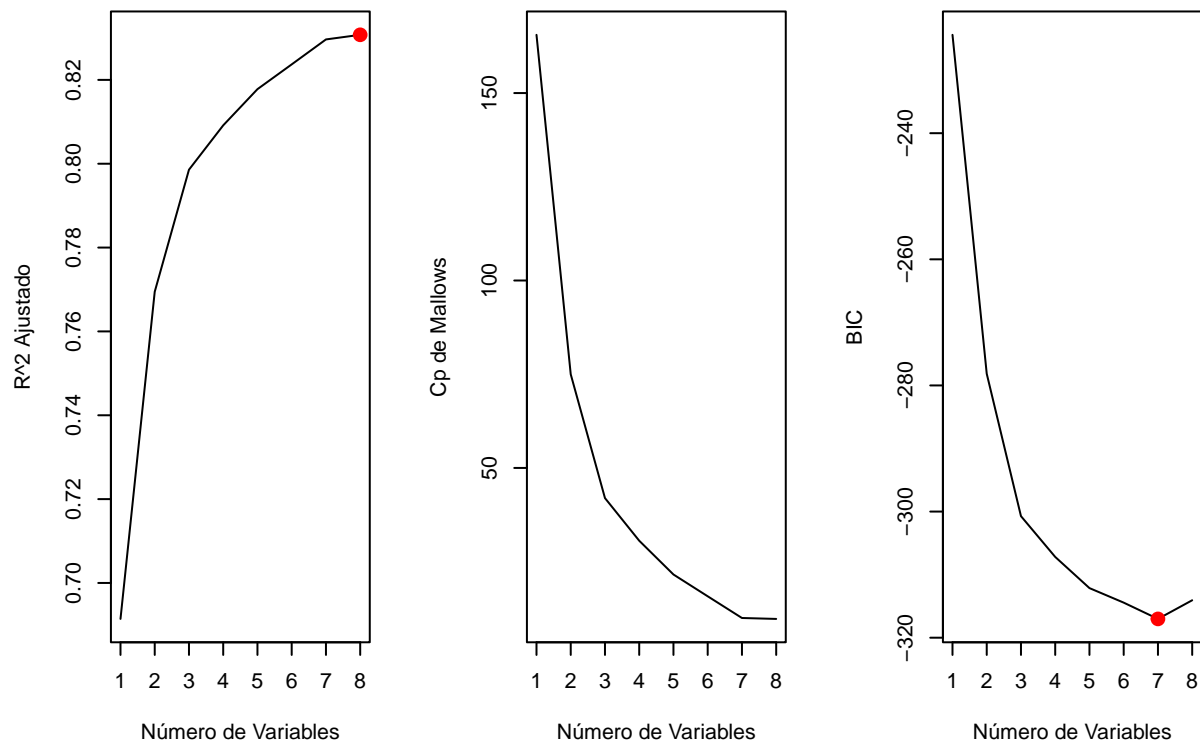
```
# Gráfico Cp de Mallows
```

```
plot(res.summary$cp, xlab = "Número de Variables", ylab = "Cp de Mallows", type = "l")
```

```
# Gráfico BIC
```

```
plot(res.summary$bic, xlab = "Número de Variables", ylab = "BIC", type = "l")
```

```
points(which.min(res.summary$bic), res.summary$bic[which.min(res.summary$bic)], col = "red", cex = 2, pch = 1)
```



```
par(mfrow = c(1, 1))
```

Las gráficas de RegSubsets para log(crim) muestran que el R^2 se maximiza con 8 variables, pero el criterio BIC y Cp alcanzan su punto óptimo con 7 variables.

Por lo tanto, se va a hacer el modelo de 7 variables.

```

modelo_final <- lm(log(crim) ~.-black-medv-dis, data = df)
summary(modelo_final)

##
## Call:
## lm(formula = log(crim) ~ . - black - medv - dis, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.20643 -0.51599  0.03419  0.55653  2.34177
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -11.837342   0.924091 -12.810  < 2e-16 ***
## indus        0.102717   0.016710   6.147 4.51e-09 ***
## chas        -0.552478   0.159412  -3.466 0.000653 ***
## nox         6.227401   0.850136   7.325 6.51e-12 ***
## rm          0.354721   0.095766   3.704 0.000278 ***
## age         0.008112   0.002921   2.777 0.006033 **
## ptratio     0.187359   0.028125   6.662 2.81e-10 ***
## lstat       0.042877   0.013566   3.161 0.001831 **
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.763 on 191 degrees of freedom
## Multiple R-squared:  0.8357, Adjusted R-squared:  0.8296
## F-statistic: 138.8 on 7 and 191 DF,  p-value: < 2.2e-16

```

Todos los p-value de las variables son menores a 0.05 por lo que se rechaza la hipótesis nula de que $\beta_i = 0$ con i asociado a la variable, esto confirma que todas las variables seleccionadas son estadísticamente significativas. En cuanto al coeficiente R^2 nos indica que este modelo es capaz de capturar el 83,57% de la varianza de los datos, además de que el p value de F-statistic es también despreciable.

Las variables con coeficientes positivos (aumenta la criminalidad) son:

- nox: Un aumento de la contaminación está asociado a un aumento de la tasa de criminalidad
- rm: No se espera que cuanto mayor sea el tamaño de las viviendas haya más criminalidad. Este resultado puede deberse a un efecto de colinealidad, en el que el modelo ya ha asignado los efectos negativos de la riqueza a otras variables.
- ptratio: A más alumnos por profesor la criminalidad aumenta.
- indus: A mayor industria más criminalización.
- age: Viviendas antiguas conlleva más criminalidad.
- lstat: A mayor porcentaje de población de clase baja la criminalidad aumenta.

Y el que tiene coeficiente negativo es chas.

Por lo tanto, ¿tienen todos sentido? No, los coeficientes de rm y age son contraintuitivos ya que deberían de ser negativos y son positivos. Estos signos que parecen erróneos ocurren porque en la regresión lineal múltiple el coeficiente de cada predictor aísla su efecto sobre los demás. Es posible que estas variables estén afectadas por la colinealidad de los otros predictores con más peso.

Apartado 2

El método de stepwise selecciona el modelo óptimo de manera automática dado un conjunto de variables predictoras.

```
df <- read_excel("boston.xlsx")
lm3<-lm(log(crim)~.,data=df)
lm3_stepwise<-step(lm3)

## Start:  AIC=-99.03
## log(crim) ~ indus + chas + nox + rm + age + dis + ptratio + black +
##      lstat + medv
##
##           Df Sum of Sq   RSS   AIC
## - black     1    0.4009 108.72 -100.296
## <none>                108.32  -99.031
## - medv     1    1.1444 109.47  -98.939
## - age      1    1.4288 109.75  -98.423
## - dis      1    2.1688 110.49  -97.086
## - lstat    1    4.1362 112.46  -93.574
## - chas     1    7.1735 115.50  -88.270
## - rm       1    8.4905 116.81  -86.014
## - indus    1   16.7519 125.07  -72.415
## - nox      1   17.6453 125.97  -70.999
## - ptratio  1   19.3452 127.67  -68.331
##
## Step:  AIC=-100.3
## log(crim) ~ indus + chas + nox + rm + age + dis + ptratio + lstat +
##      medv
##
##           Df Sum of Sq   RSS   AIC
## <none>                108.72 -100.296
## - medv     1    1.1676 109.89 -100.170
## - age      1    1.3744 110.10  -99.796
## - dis      1    2.0229 110.75  -98.627
## - lstat    1    3.8623 112.58  -95.349
## - chas     1    6.9877 115.71  -89.900
## - rm       1    8.2187 116.94  -87.794
## - indus    1   18.6164 127.34  -70.843
## - nox      1   19.3651 128.09  -69.677
## - ptratio  1   19.5725 128.30  -69.355

summary(lm3_stepwise)

##
## Call:
## lm(formula = log(crim) ~ indus + chas + nox + rm + age + dis +
##      ptratio + lstat + medv, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.15193 -0.57653  0.02534  0.57184  2.10224
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.541730   1.132765  -9.306   < 2e-16 ***
```

```
## indus      0.097858  0.017202  5.689 4.80e-08 ***
## chas      -0.552395  0.158494 -3.485 0.000611 ***
## nox       5.503853  0.948607  5.802 2.72e-08 ***
## rm        0.435544  0.115229  3.780 0.000210 ***
## age       0.005145  0.003329  1.546 0.123850
## dis      -0.096980  0.051716 -1.875 0.062303 .
## ptratio   0.171231  0.029355  5.833 2.32e-08 ***
## lstat     0.037078  0.014309  2.591 0.010312 *
## medv     -0.017998  0.012633 -1.425 0.155899
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7585 on 189 degrees of freedom
## Multiple R-squared:  0.8393, Adjusted R-squared:  0.8317
## F-statistic: 109.7 on 9 and 189 DF,  p-value: < 2.2e-16
```

El modelo obtenido con stepwise tiene un R^2 ligeramente mejor y un AIC más bajo. Sin embargo, el modelo RegSubsets considero que es mejor porque simplifica más la complejidad. Si la diferencia de R^2 es mínima, se prefiere el modelo más simple, en este caso RegSubsets.

El modelo de stepwise explica un 83,93% de la varianza.

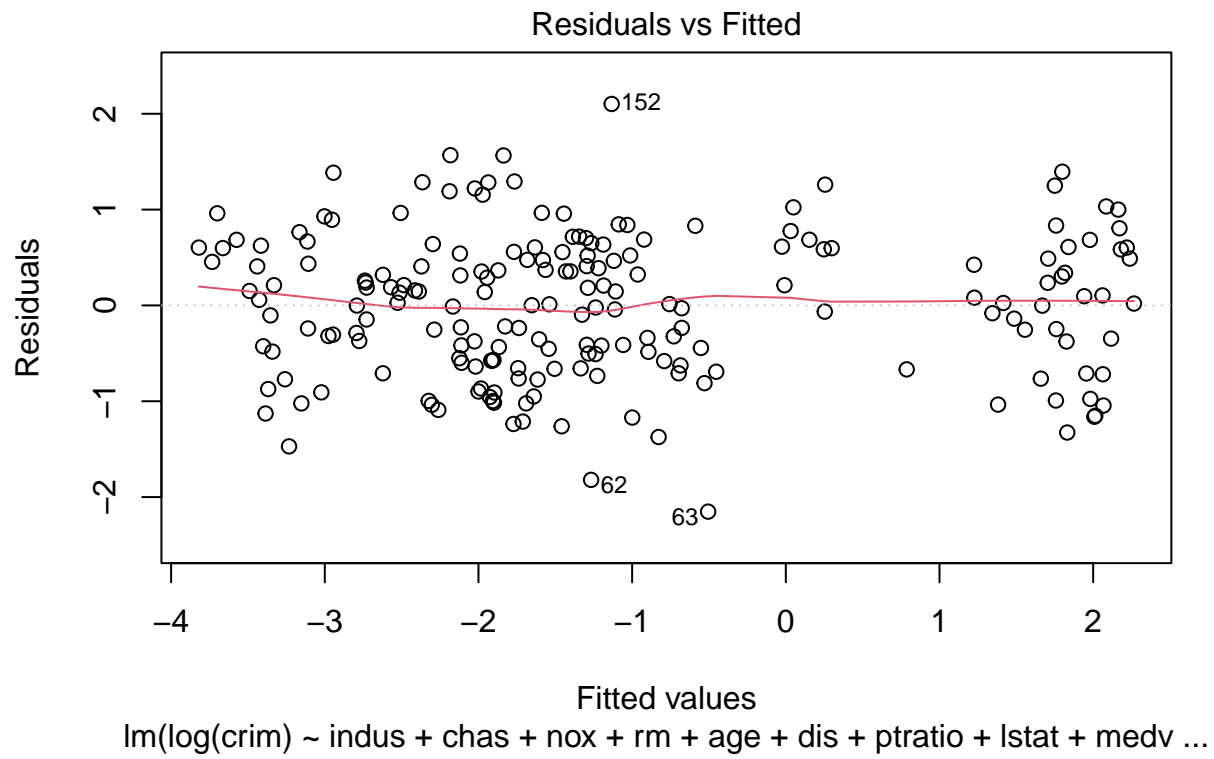
En cuanto el efecto de la variables chas sobre crim, se observa que el coeficiente de chas es negativo y estadísticamente significativo.

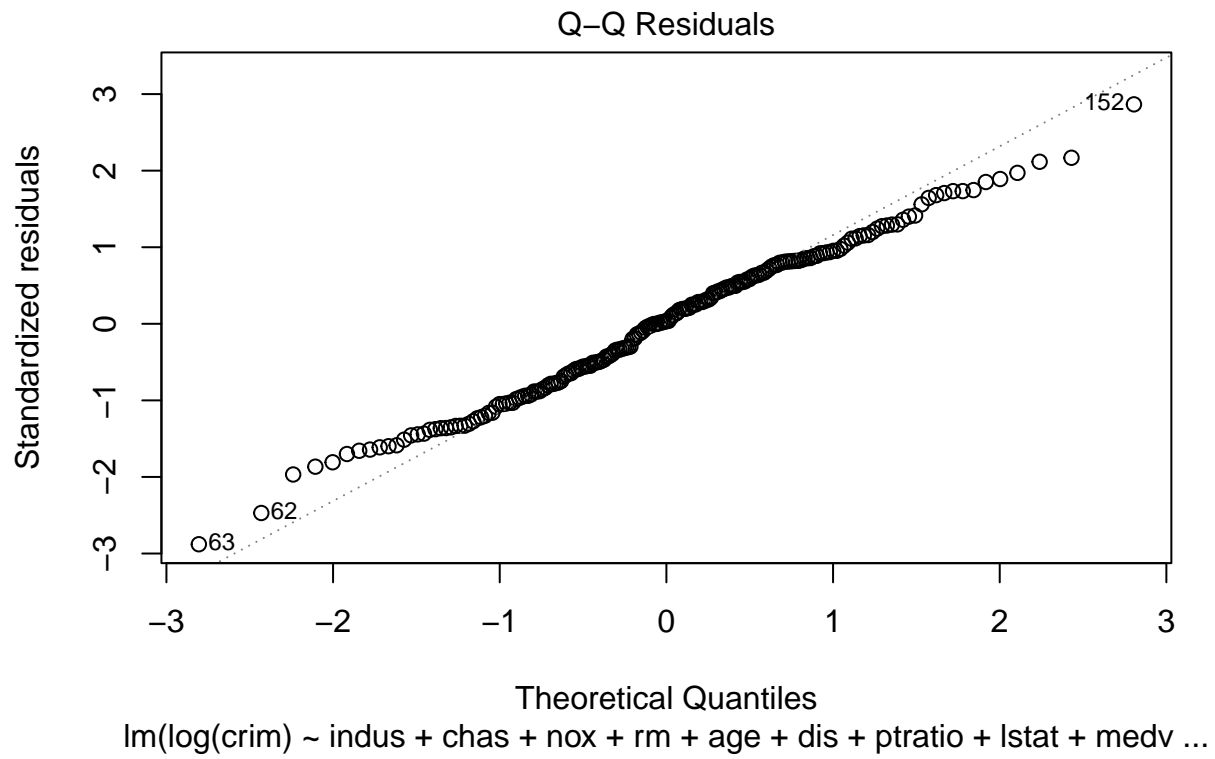
Apartado 3

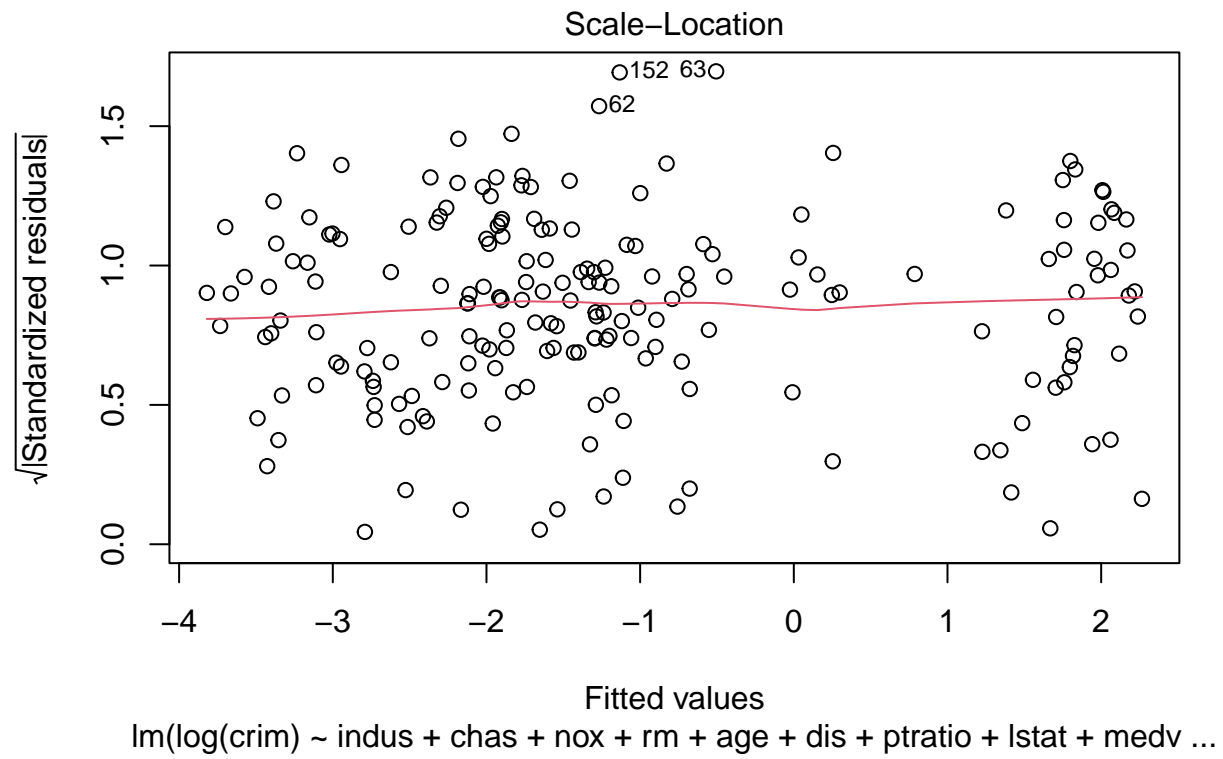
En el apartado anterior se ha podido observar cómo hay variables que no son estadísticamente significativas. En este apartado se va a realizar el diagnóstico de ese modelo sin emprender ninguna acción de eliminar esas variables.

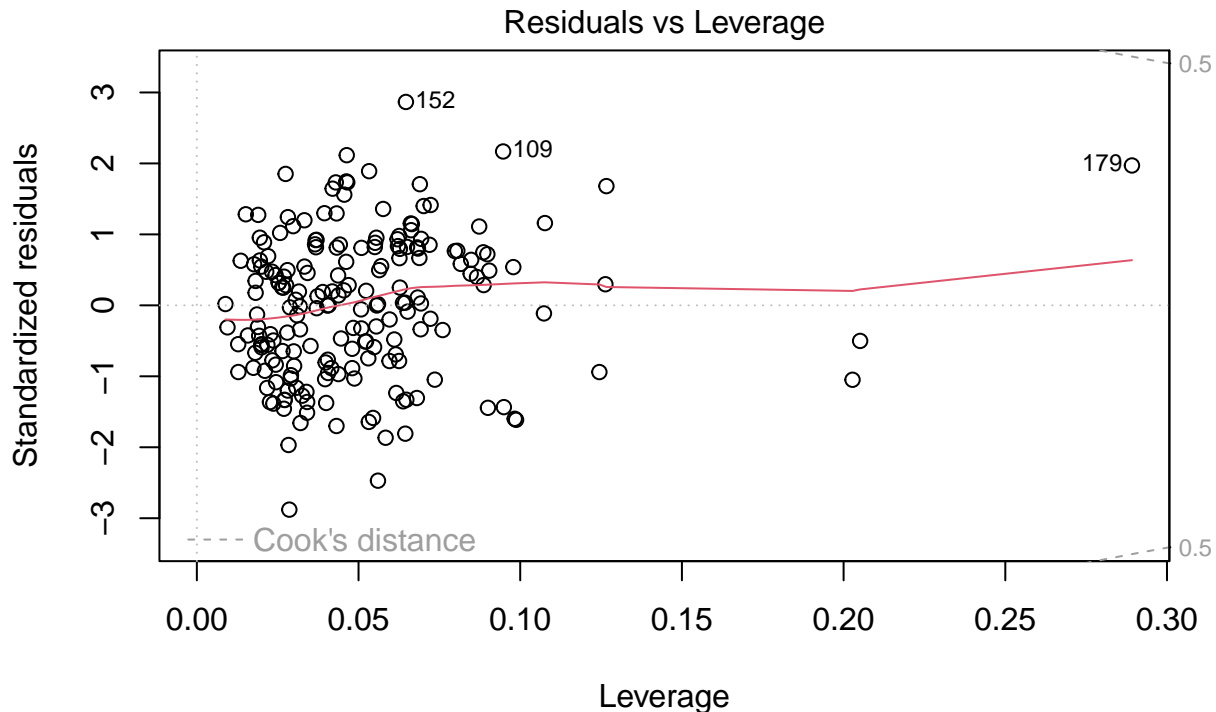
El problema principal de este modelo es que se incluyen variables no significativas según los resultados vistos con la función `summary()`. Esto sugiere que el modelo ha tenido en cuenta las variables significativas y las que no lo son (age, medv) no es estadísticamente relevante. Estova en contra del principio de parsimonia.

```
plot(lm3_stepwise)
```









$\text{lm}(\log(\text{crim}) \sim \text{indus} + \text{chas} + \text{nox} + \text{rm} + \text{age} + \text{dis} + \text{ptratio} + \text{lstat} + \text{medv} \dots)$

La gráfica de Residuals vs Fitted, la línea roja de la media de los residuos está prácticamente plana en torno a cero. Se ha mejorado la homocedasticidad de los residuos. En cuanto al QQ-residuals se observa que presenta una tendencia esperada, aunque se observa una ligera desviación del supuesto de normalidad de los residuos, sobretudo para valores extremos.

Por último, en la figura de Residuals vs Leverage se identifican valores outliers con altos residuos que pueden estar afectando la robustez del modelo.

Apartado 4

Uno de los principales problemas que se tiene con stepwise es que utiliza age y medv. El problema de la heterocedasticidad en los residuos ya estaba resuelto por haber usado el $\log(\text{crim})$, si no se hubiera puesto desde un principio hubiera sido otra mejora a considerar

```
df <- read_excel("boston.xlsx")
```

```
lm3 <- lm(log(crim) ~ . - age - medv, data=df)
```

```
lm3_stepwise <- step(lm3)
```

```
## Start: AIC=-97.68
```

```
## log(crim) ~ (indus + chas + nox + rm + age + dis + ptratio +
```

```
## black + lstat + medv) - age - medv
```

```
##
```

```
## Df Sum of Sq RSS AIC
```

```
## - black 1 0.3643 111.64 -99.029
```

```
## <none> 111.28 -97.679
```

```
## - dis 1 4.2892 115.56 -92.153
```

```
## - chas 1 6.7139 117.99 -88.021
```



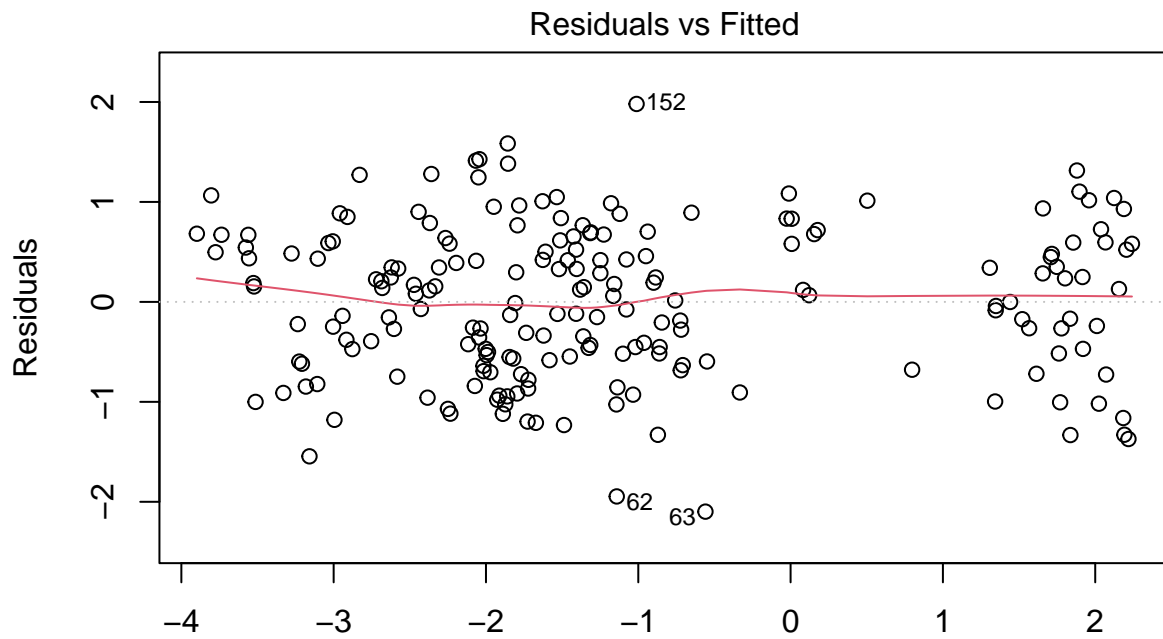
```
## - lstat      1      8.9616 120.24 -84.265
## - rm         1      8.9951 120.27 -84.210
## - nox        1     20.3865 131.66 -66.202
## - indus      1     20.9445 132.22 -65.360
## - ptratio    1     23.6580 134.93 -61.317
##
## Step:  AIC=-99.03
## log(crim) ~ indus + chas + nox + rm + dis + ptratio + lstat
##
##           Df Sum of Sq    RSS    AIC
## <none>                111.64 -99.029
## - dis      1      4.0374 115.68 -93.959
## - chas     1      6.5496 118.19 -89.684
## - lstat    1      8.6171 120.26 -86.233
## - rm       1      8.6759 120.31 -86.135
## - nox      1     22.2070 133.85 -64.927
## - indus    1     23.2133 134.85 -63.436
## - ptratio  1     24.3359 135.97 -61.786

summary(lm3_stepwise)

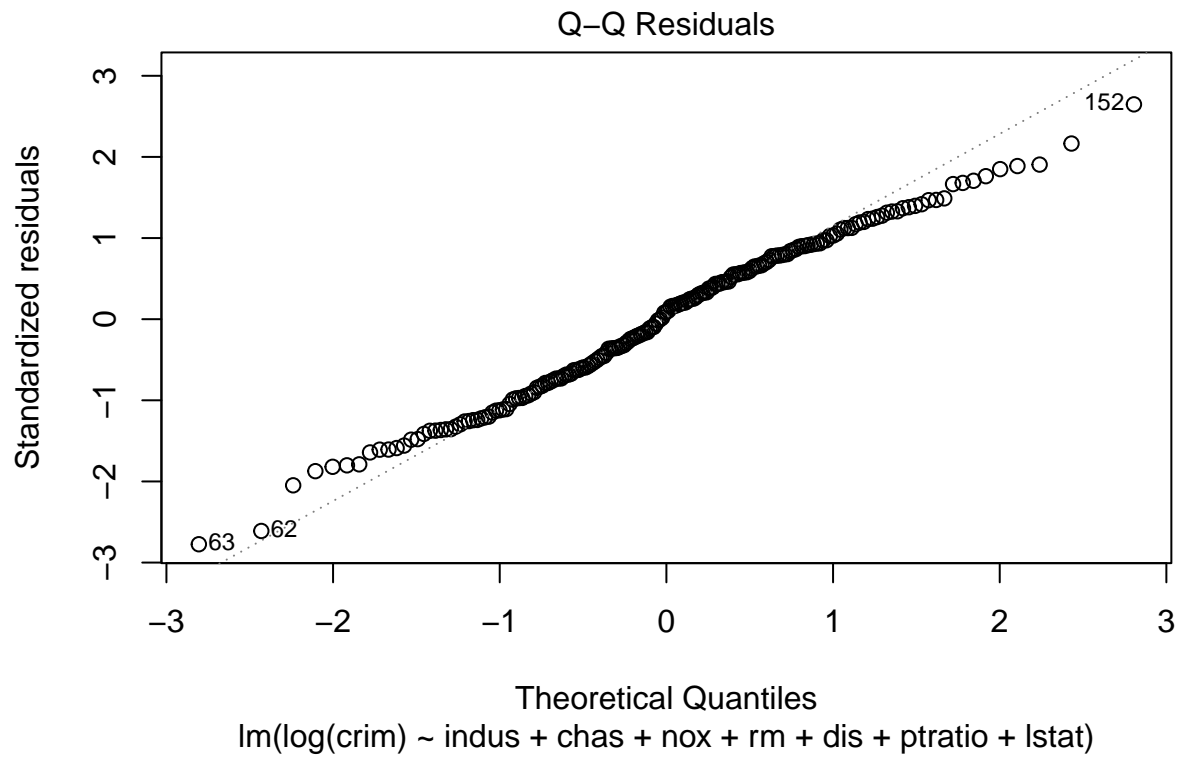
##
## Call:
## lm(formula = log(crim) ~ indus + chas + nox + rm + dis + ptratio +
##      lstat, data = df)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.09840 -0.55955  0.06672  0.58558  1.98025
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -10.75672     1.10360  -9.747  < 2e-16 ***
## indus         0.10555     0.01675   6.302 1.98e-09 ***
## chas        -0.53299     0.15922  -3.347 0.000982 ***
## nox          5.82243     0.94461   6.164 4.13e-09 ***
## rm           0.36711     0.09529   3.853 0.000159 ***
## dis        -0.11431     0.04349  -2.628 0.009281 **
## ptratio      0.18262     0.02830   6.453 8.83e-10 ***
## lstat        0.05026     0.01309   3.840 0.000167 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.7645 on 191 degrees of freedom
## Multiple R-squared:  0.835, Adjusted R-squared:  0.829
## F-statistic: 138.1 on 7 and 191 DF,  p-value: < 2.2e-16
```

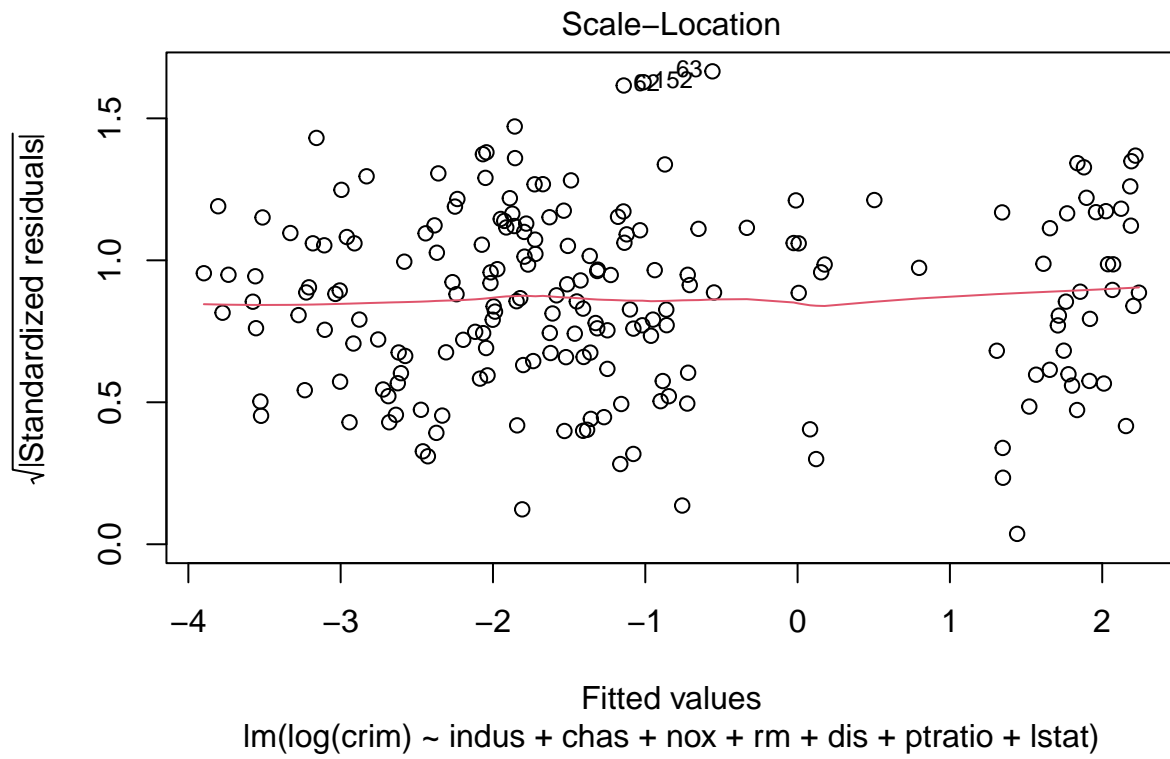
Se observa que ahora el valor de AIC sale ligeramente más alto que antes tras haber eliminado estos valores, y también el coeficiente de R^2 , sin embargo, ahora todas las variables son estadísticamente significativas. Tras eliminar estas variables se obtiene un modelo más simple y con coeficientes más robustos.

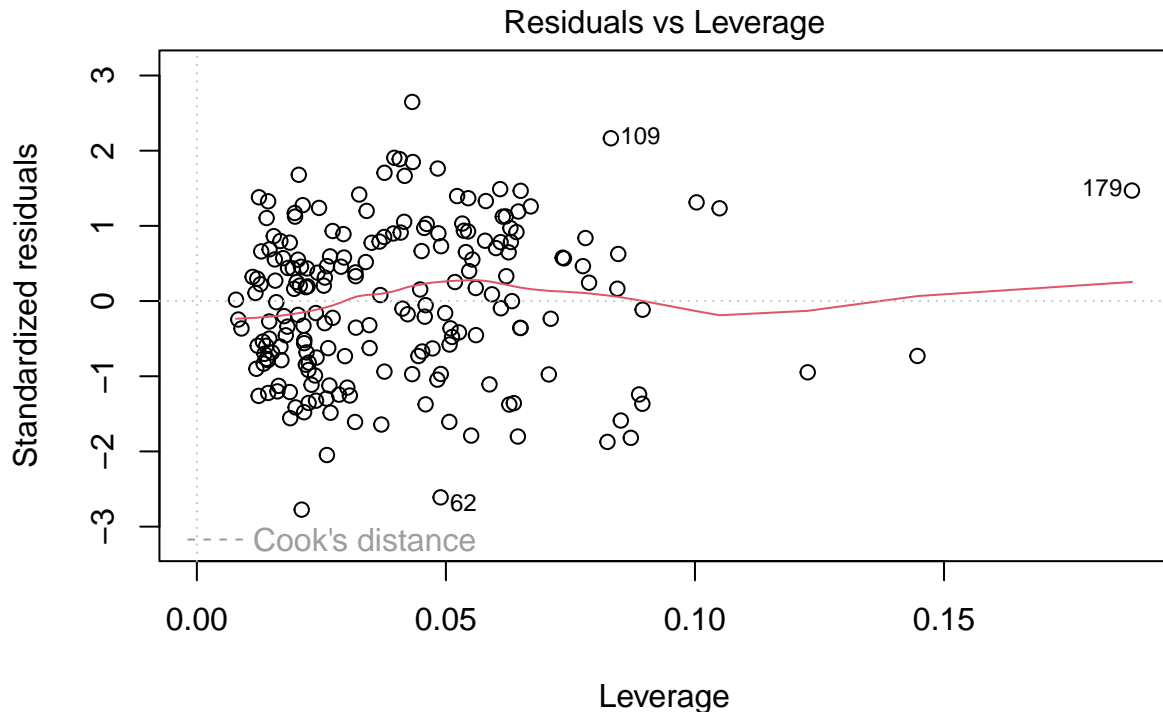
```
plot(lm3_stepwise)
```



lm(log(crim) ~ indus + chas + nox + rm + dis + ptratio + lstat)







$\text{lm}(\log(\text{crim}) \sim \text{indus} + \text{chas} + \text{nox} + \text{rm} + \text{dis} + \text{ptratio} + \text{lstat})$

Tras el reajuste, los problemas de no linealidad y heterocedasticidad están resueltos (gracias a que se utiliza $\log(\text{crim})$) y eliminar las variables más débiles ha ayudado a que el modelo sea más preciso para interpretar.

Apartado 5

```
medv_indus <- median(df$indus)
medv_nox <- median(df$nox)
medv_rm <- median(df$rm)
medv_ptratio <- median(df$ptratio)
medv_lstat <- median(df$lstat)
medv_dis <- median(df$dis)

new_data_median <- data.frame(indus = medv_indus, chas = 0, nox = medv_nox, rm = medv_rm, dis = medv_dis, ptratio = medv_ptratio, lstat = medv_lstat)

IP_log_crim <- predict(lm3_stepwise,
                      newdata = new_data_median,
                      interval = "prediction",
                      level = 0.95)
IP_crim <- exp(IP_log_crim)

print(IP_crim)

##          fit          lwr          upr
## 1 0.1908101 0.04189432 0.8690552

print(median(df$crim))
```

```
## [1] 0.24103
```

Para la variable categórica chas se ha decidido utilizar 0 por conveniencia.

El resultado obtenido es lógico, dentro del intervalo de confianza se encuentra el valor de la mediana por lo que se puede concluir que el modelo es robusto capaz de explicar aproximadamente un 84% de la varianza de los datos que permite predecir la tasa de criminalidad esperada para un determinado barrio.