Regresión Lineal Simple y Múltiple

Javier Adanaqué

17 de Abril, 2017

Resumen

Se analiza cómo se relaciona el valor de las viviendas en Boston frente a características tanto de las viviendas como de la población en general en donde se encuentran las viviendas.

1. Introducción

El objetivo del presente documento es aplicar los métodos de Regresión Lineal Simple y Múltiple aprendidos en clase. Para ello se está usando una base de datos con diferentes características de las viviendas y suburbios/ciudades en Boston, data otorgada en clase junto a otros datasets.

Primero se realizará una exploración rápida de las variables y luego se contruirá el modelo, mostrando el código usado para llegar a ellos.

Nos concentraremos en conocer cómo se relaciona el valor de las viviendas en Boston con respecto a diferentes características de los suburbios en las que se encuentran. Esto sería de mucha utilidad para alguna empresa constructora, para autoridades gubernamentales o para los ciudadanos que se encuentren evaluando comprar o vender una vivienda.

2. Data

Para el análisis se está usando la data `bostonvivienda.txt`, ubicada entre los datasets distribuidos para la clase.

```
valor_viviendas <- read.table("data/bostonvivienda.txt",
                              header = TRUE, stringsAsFactors = FALSE)
valor_viviendas[1:6, c(1:3, 6, 7, 10, 13, 14)]
```

crim	zn	indus	rm	edad	impuesto	lstat	medv
0.00632	18	2.31	6.575	65.2	296	4.98	24.0
0.02731	0	7.07	6.421	78.9	242	9.14	21.6
0.02729	0	7.07	7.185	61.1	242	4.03	34.7
0.03237	0	2.18	6.998	45.8	222	2.94	33.4
0.06905	0	2.18	7.147	54.2	222	5.33	36.2
0.02985	0	2.18	6.430	58.7	222	5.21	28.7

Como se puede observar en las primeras seis observaciones (y algunas columnas), nos

muestra diferentes características sobre los suburbios, incluido el valor medio de las viviendas en el suburbio (ver última columna).

Descripción de variables:

- `crim`: tasa de delincuencia per cápita por ciudad.
- `zn`: proporción de suelo residencial dividido en zonas para lotes de más de 25,000 pies cuadrados.
- `indus`: proporción de acres de negocios no minoristas por la ciudad.
- `chas`: variable ficticia (dummy) Charles River (1 si sale de las vías fluviales; 0 en caso contrario).
- `nox`: concentración de óxidos de nitrógeno (partes por 10 millones).
- `rm`: número promedio de habitaciones por vivienda.
- `edad`: proporción de unidades ocupadas por sus propietarios construidas antes de 1940.
- `dis`: media ponderada de las distancias a cinco centros de empleo de Boston.
- `rad`: Índice de la accesibilidad a las autopistas radiales.
- `impuesto`: tasa de impuestos a la propiedad por el valor total por $ 10,000.
- `ptratio`: proporción de alumnos por profesor por ciudad.
- `negro`: 1000(Bk - 0,63)2, donde Bk es la proporción de negros por la ciudad.
- `lstat`: estatus más bajo de la población (por ciento).
- `medv`: valor mediano de las viviendas ocupadas por sus propietarios en $ 1000s.

3. Análisis exploratorio

A continuación exploraremos algunas variables que parecen relevantes para el valor de las viviendas en Boston. Empezaremos explorando nuestra variable que queremos explicar, `medv`:

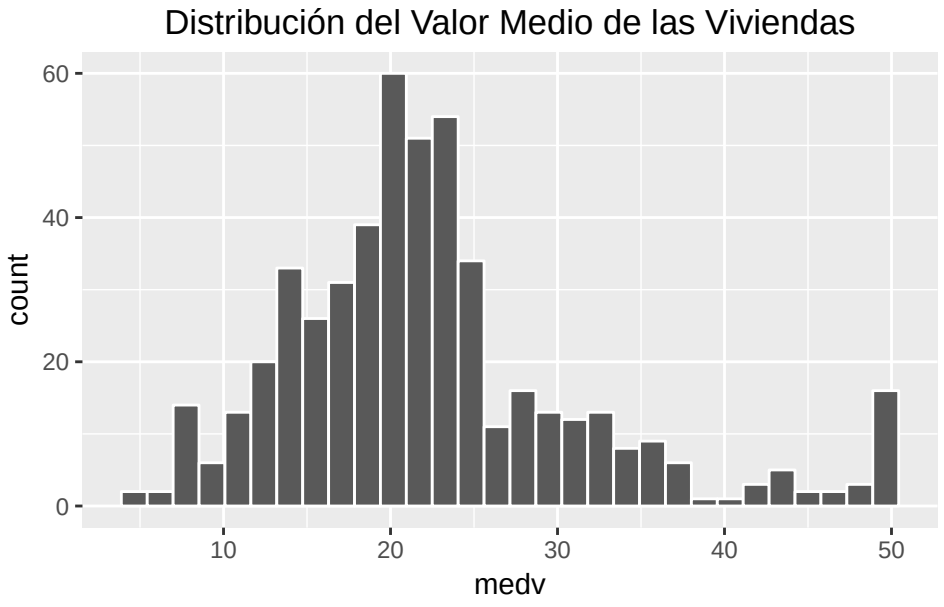

Figura 1: Histograma del Valor Medio de las Viviendas en Boston

Ahora un vistazo a un par de variables más:

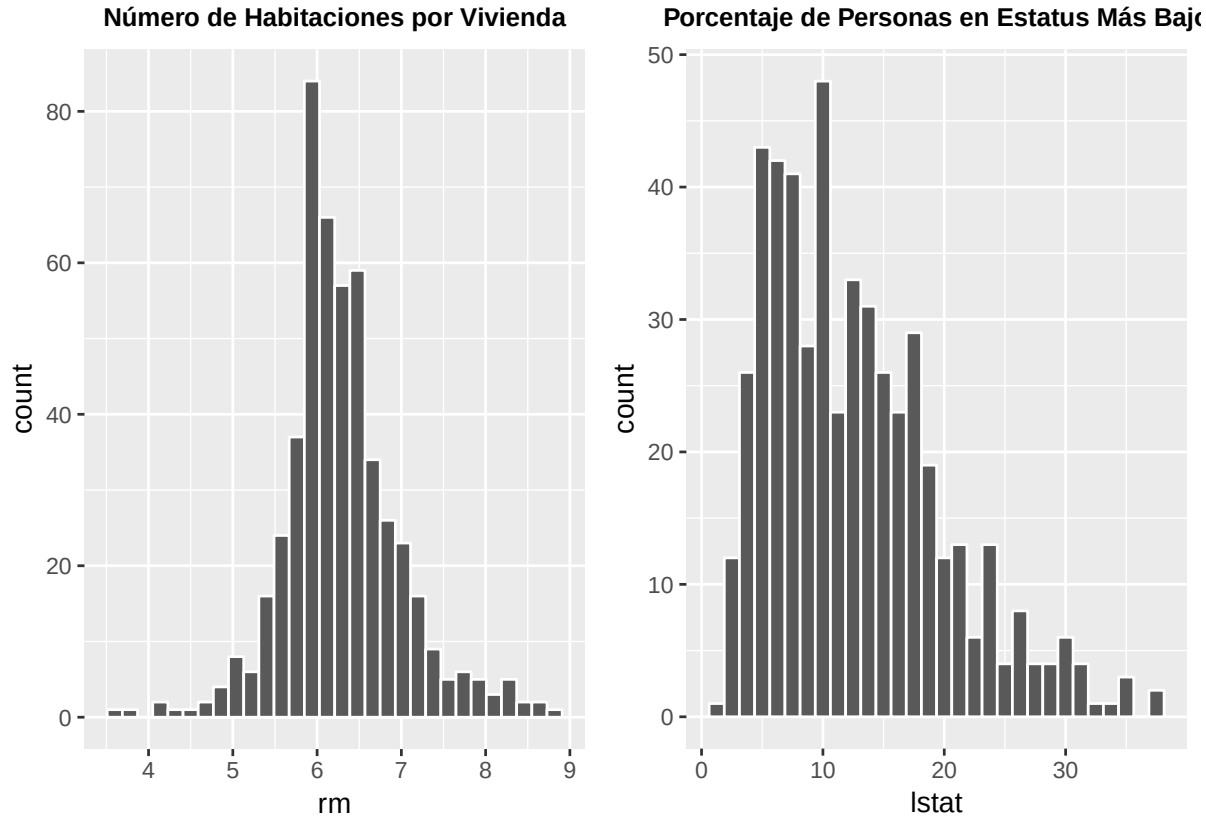

Figura 2: Histograma del Número de Habitaciones y Porcentaje de Personas en Estatus Más Bajo

Finalmente, veamos cómo se relaciona nuestra variable de interés, `medv`, con respecto a estas dos variables mostradas:

Claramente, existe una relación entre estas variables, positiva con respecto al número de habitaciones y negativa con respecto al porcentaje de la población en el estatus más bajo.

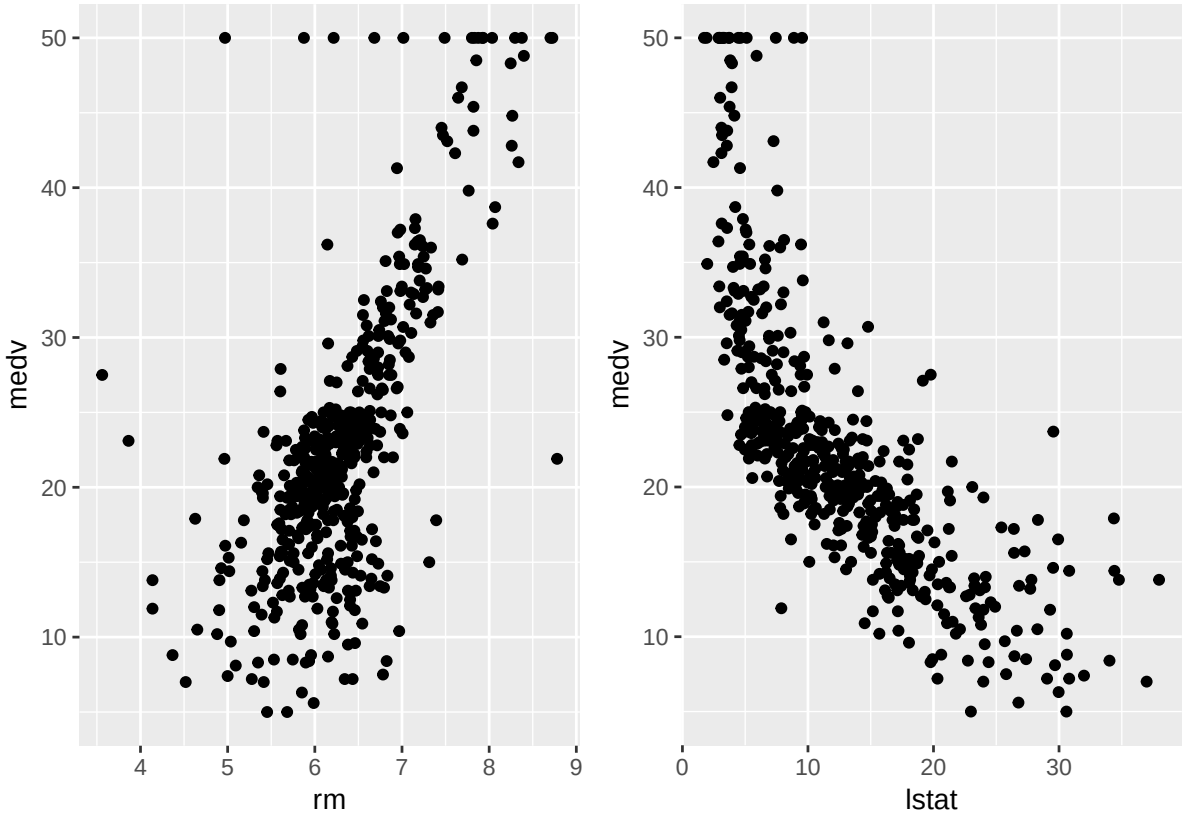

Figura 3: Diagramas de Dispersión. Izquierda: Valor Mediano de las Viendas con respecto al Número de Habitaciones. Derecha: Valor Mediano de las viviendas con respecto al Porcentaje de Personas en Estatus Más Bajo

4. Modelamiento

4.1. Regresión Lineal Simple

En los últimos gráficos de dispersión observamos que existe cierta relación entre el valor de las viviendas en Boston y el número promedio de habitaciones y porcentaje de población en el estatus más bajo.

Ahora, elaboremos un modelo de Regresión Lineal Simple con alguna de las variables. Existen varios métodos para seleccionar las variables más explicativas; sin embargo, en este análisis seleccionaremos sólo una variable a priori, el **Número de Habitaciones Promedio**.

```
lm_viviendas <- lm(medv ~ rm, data = valor_viviendas)
summary(lm_viviendas)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = valor_viviendas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -23.346  -2.547   0.090   2.986  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  -34.671      2.650  -13.08   <2e-16 ***
## rm             9.102      0.419   21.72   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.616 on 504 degrees of freedom
## Multiple R-squared:  0.4835, Adjusted R-squared:  0.4825
## F-statistic: 471.8 on 1 and 504 DF,  p-value: < 2.2e-16
```

Una de las primeras cosas que podemos observar es nuestro coeficiente de determinación ajustado, $R^2_{ajustado}$, que no está mal considerando que sólo hemos seleccionado una variable entre todas las que podrían explicar los precios de las viviendas. Este número nos está diciendo que alrededor del 48 % de la variabilidad en los precios de las viviendas está explicada por el número de habitaciones.

Otro valor muy importante es el `p-value` de nuestro modelo, $< 2{,}2e^{-16}$, lo cual nos dice que nuestro modelo es significativo (rechazamos la Hipotesis Nula). Esto lo podemos validar en la Tabla de Anova con más detalle:

```r
anova(lm_viviendas)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
rm	1	20654.42	20654.41622	471.8467	0
Residuals	504	22061.88	43.77357	NA	NA

Finalmente, pero no menos importante, debemos prestar atención a lo significativo de nuestros coeficiente e intercepto. En este caso, ambos son significativos, también con un valor menor a $2{,}2e^{-16}$.

Con esto, nuestro modelo resultante queda de la siguiente manera:

$$\hat{medv} = -34{,}671 + 9{,}102rm$$

El 9,102 es la pendiente, que nos está diciendo que por cada habitación extra, el precio (mediano) de las viviendas incrementa en $\$9,102$.

El $-34{,}671$ es el intercepto, que nos dice cuál es precio (mediano) de las viviendas, cuando estas no tienen ninguna habitación. En este caso, un valor negativo pareciera con poco sentido, pero si nos fijamos bien en nuestra variable `rm`, el valor mínimo de esta es de alrededor de 4, por lo que el valor **inicial** estimado de una vivienda sería de alrededor de $1737. Otra forma de interpretar este intercepto negativo es que nadie estaría dispuesto a pagar por una vivienda con pocas habitaciones (e.g., de 1 a 3 habitaciones).

Ahora observemos, gráficamente, cómo quedó nuestro modelo:

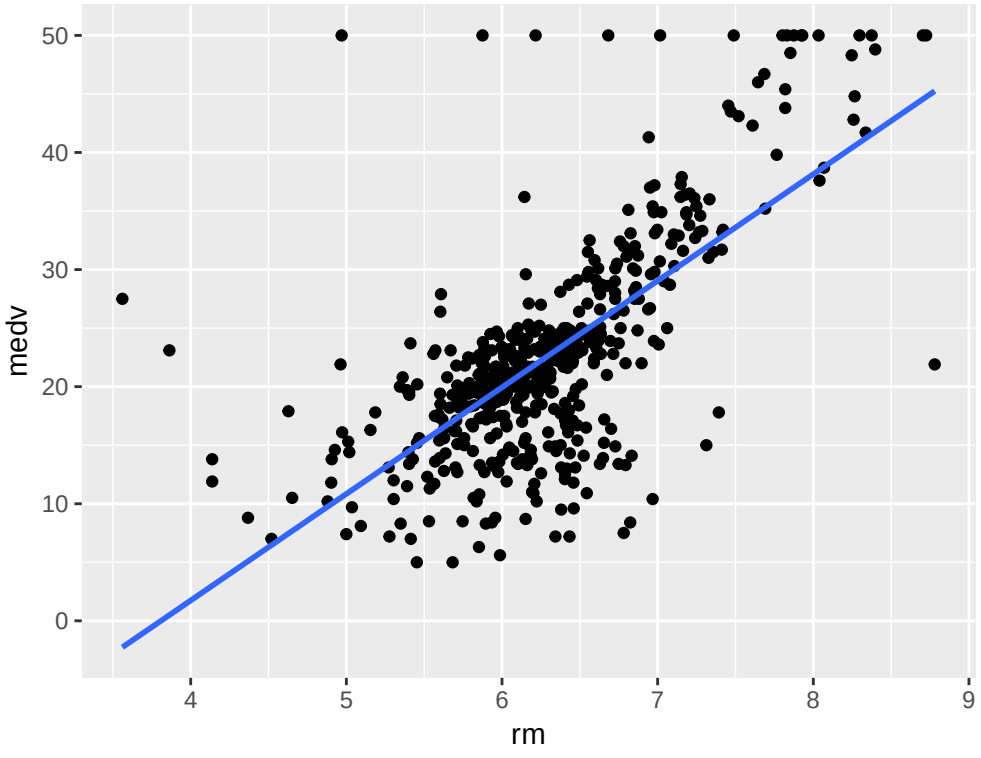

Figura 4: Diagrama de Dispersión y el modelo ajustado (línea azul)

Para acabar, revisemos algunos residuos:

```
residuos_df <- data.frame(medv = valor_viviendas$medv[c(1, 6, 200, 420)],
                          medv_estimado = -34.671 + 9.102*valor_viviendas$rm[c(1, 6,
residuos_df$residuos <- residuos_df$medv - residuos_df$medv_estimado
residuos_df
```

medv	medv_estimado	residuos
24.0	25.17465	-1.17465
28.7	23.85486	4.84514
34.9	28.81545	6.08455
8.4	27.44105	-19.04105

Ahora, revisemos todos los residuos:

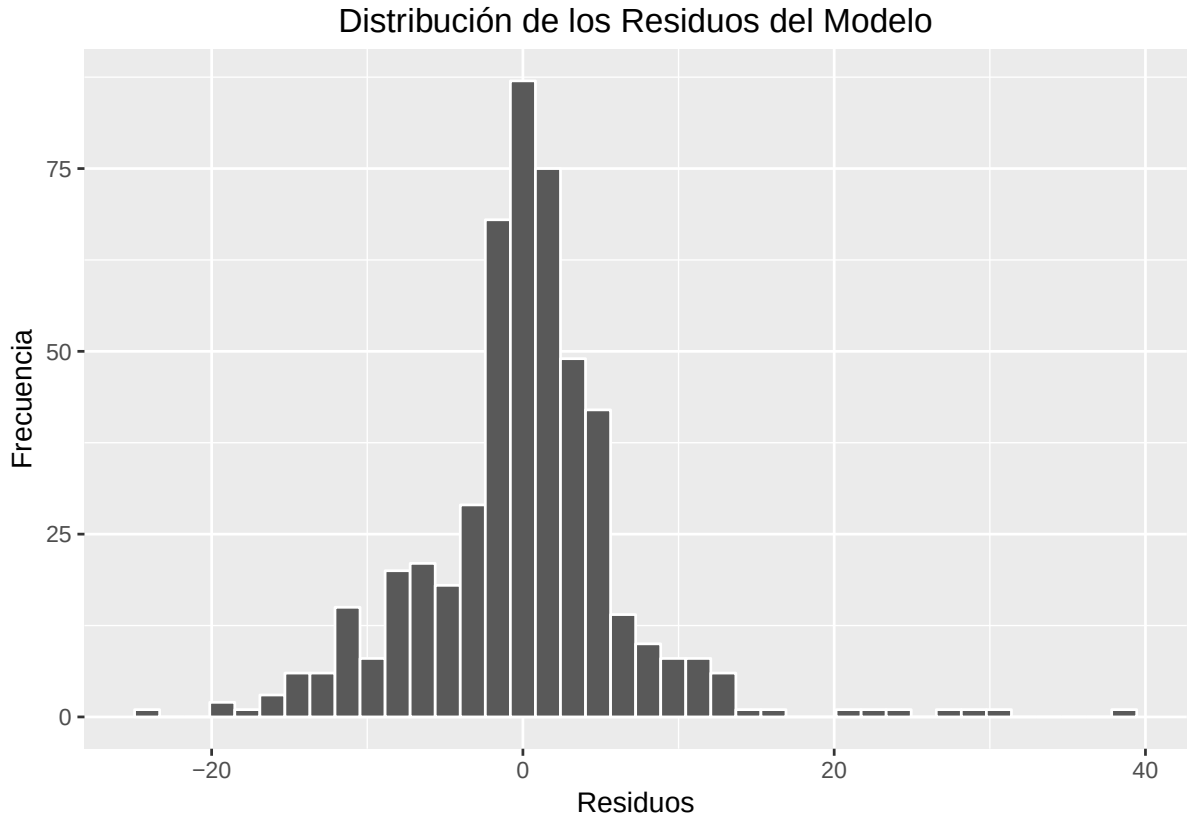

Figura 5: Histograma de los Residuos del Modelo

4.2. Regresión Lineal Múltiple

Ahora extenderemos el modelo analizado en la sección anterior usando una variable adicional que nos permita explicar mejor el precio de las viviendas en Boston.

Usaremos `lstat` como nuestra segunda variable predictora. Recordando que esta variable nos da la proporción de personas en el estatus más bajo de la población, tiene sentido

saber si una mayor proporción de personas pobres influye en el precio de las viviendas; una hipótesis inicial podría ser que sí, que una mayor proporción de personas pobres afectaría negativamente el valor de las viviendas.

Contruyamos el modelo para ver si nuestra hipótesis es cierta y qué tanto nos ayuda a explicar el precio de las viviendas:

```
lm_mult_viviendas <- lm(medv ~ rm + lstat, data = valor_viviendas)
summary(lm_mult_viviendas)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat, data = valor_viviendas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -18.076  -3.516  -1.010   1.909  28.131
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -1.35827    3.17283  -0.428    0.669
## rm           5.09479    0.44447  11.463   <2e-16 ***
## lstat       -0.64236    0.04373 -14.689   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.54 on 503 degrees of freedom
## Multiple R-squared:  0.6386, Adjusted R-squared:  0.6371
## F-statistic: 444.3 on 2 and 503 DF,  p-value: < 2.2e-16
```

Aparentemente, sí nos ayuda a explicar mejor el precio de las viviendas (ver $R^2_{ajustado}$); sin embargo, también nos podemos dar cuenta de que, según el resumen de nuestro modelo, el intercepto es no significativo.

Dado que el intercepto no es significativo, procedemos con eliminarlo de nuestro modelo.

```
lm_mult_viviendas <- lm(medv ~ rm + lstat - 1, data = valor_viviendas)
summary(lm_mult_viviendas)
```

```
##
## Call:
## lm(formula = medv ~ rm + lstat - 1, data = valor_viviendas)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -17.714  -3.498  -1.075   1.877  27.750
##
## Coefficients:
##       Estimate Std. Error t value Pr(>|t|)
## rm     4.90691    0.07019   69.91   <2e-16 ***
## lstat -0.65574    0.03056  -21.46   <2e-16 ***
```

```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 5.536 on 504 degrees of freedom
## Multiple R-squared:  0.9485, Adjusted R-squared:  0.9482
## F-statistic:  4637 on 2 and 504 DF,  p-value: < 2.2e-16
```

Nos damos cuenta de que, efectivamente, eliminar el intercepto mejora considerablemente nuestro modelo. Ahora tenemos un $R^2_{ajustado}$ de 0.9482, indicando que nuestras variables predictoras elegidas explican casi el 95 % de la variabilidad en los precios de las viviendas en Boston.

También nos damos cuenta de que tanto nuestro modelo como los coeficientes son altamente significativos, con un p-valor menor que $2{,}2e^{-16}$.

Con esto, nuestro modelo resultante queda de la siguiente manera:

$$\hat{medv} = 4{,}907rm - 0{,}656lstat$$

Nos damos cuenta de que, efectivamente, `lstat` aporta negativamente a los precios de las viviendas. Cada punto porcentual adicional disminuye el precio estimado de las viviendas en $-0{,}656$ ($\$ - 656$).

Así mismo, nos podemos dar cuenta de que el peso que tiene el número de habitaciones en el precio de las viviendas es menor que el que tenía en nuestro modelo de Regresión Lineal Simple. Ahora, por cada habitación extra, el precio (mediano) de las viviendas incrementa en 4,907 ($\$4,907$). Tiene sentido ahora que no tenemos el intercepto y que tenemos una variable adicional en el modelo (con valor negativo).

Nuevamente, para acabar, revisemos primero algunos residuos y luego todos:

```
residuos_df2 <- data.frame(medv = valor_viviendas$medv[c(1, 6, 200, 420)],
                           medv_estimado = 4.907*valor_viviendas$rm[c(1, 6, 200, 420)]
                                           0.656*valor_viviendas$lstat[c(1, 6, 200, 42
residuos_df2$residuos <- residuos_df2$medv - residuos_df2$medv_estimado
residuos_df2
```

medv	medv_estimado	residuos
24.0	28.99665	-4.996645
28.7	28.13425	0.565750
34.9	31.23496	3.665035
8.4	18.56793	-10.167928

Ahora todos los residuos:

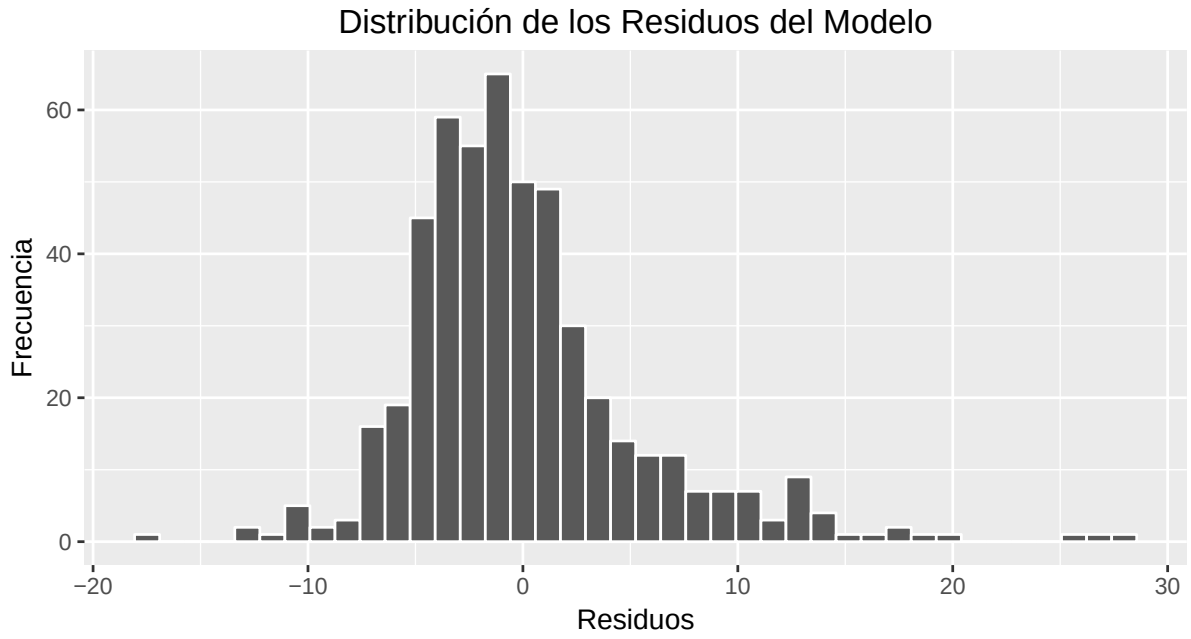

Figura 6: Distribución de los Residuos

5. Conclusiones

El valor mediano de las viviendas en Boston se encuentra altamente explicado por el número de habitaciones promedio por vivienda y por el porcentaje de la población en el estatus más bajo, dejando claro que no sólo las características de las viviendas son importantes, sino que el entorno en el que se encuentran es igual de importante.

Definitivamente, podríamos aplicar más tests a nuestro análisis y considerar la relevancia de mas variables, pero con lo que se ha realizado queda clara la relación entre estas variables y qué tan fuerte esta es.