# ejercicio07

Reinaldo Pacheco

2023-12-09

**Análisis de Regresión Lineal Simple con el Dataset de Boston**

Primero, cargamos el conjunto de datos de Boston y seleccionamos las variables de interés: “RM” (número medio de habitaciones por vivienda) y “MEDV” (valor medio de las viviendas).

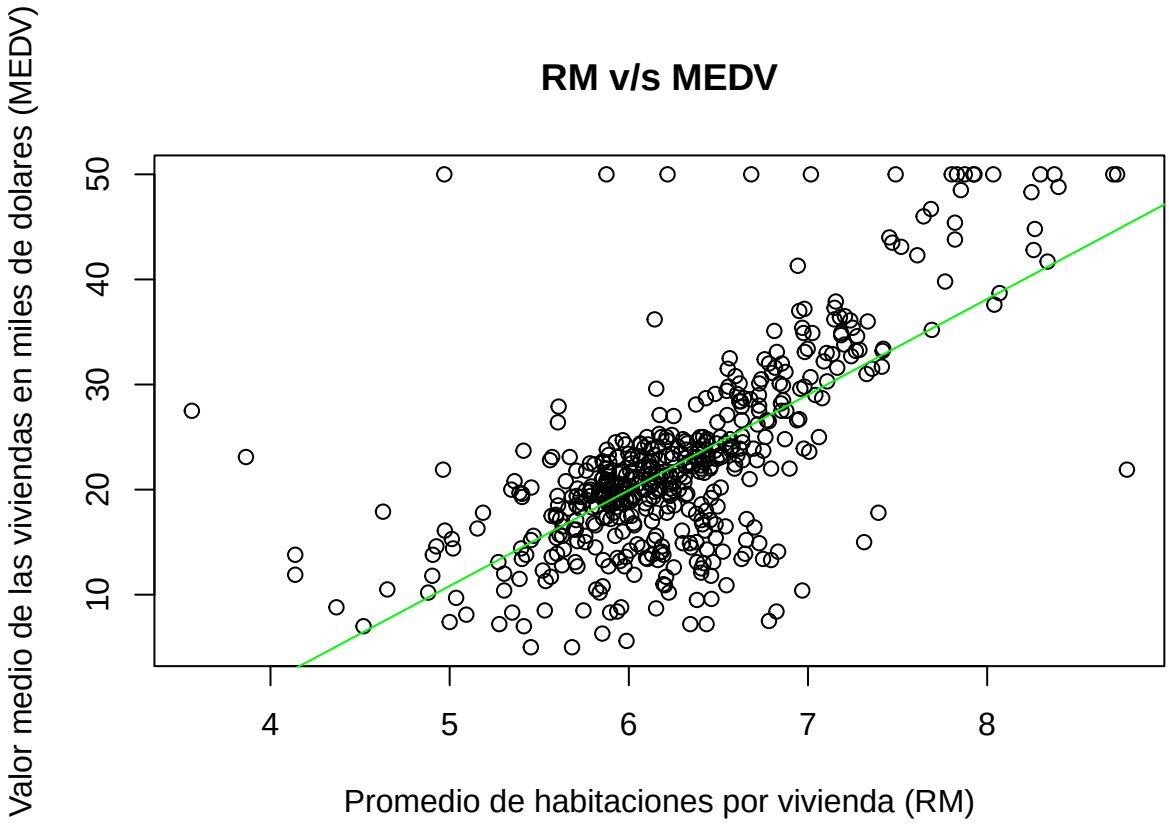
Se eligen estas dos variables para estudiar si existe algún tipo de relación entre la cantidad de habitaciones en una vivienda, con el valor de la vivienda.

```
library(MASS)
data(Boston)
boston_df = Boston[, c("rm", "medv")]
```

**Regresión Lineal Simple y Gráfico** Realizamos una regresión lineal simple entre “RM” y ‘MEDV’ con el fin de investigar la relación entre la media de la cantidad de habitaciones y la media del valor de las viviendas. Se gráfican los resultados.

```
plot(boston_df$rm, boston_df$medv,
     xlab = "Promedio de habitaciones por vivienda (RM)",
     ylab = "Valor medio de las viviendas en miles de dolares (MEDV)",
     main = "RM v/s MEDV")

ajuste = lm(medv ~ rm, data = boston_df)
abline(ajuste, col = "green")
```




Los resultados muestran una relación entre las dos variables estudiadas. Gráficamente, se percibe un coeficiente de relación "Strong positive" lo cual indica que estas dos variables, tienen una relación directa. Estudiando más a fondo los resultados, se puede evidenciar que a medida que aumenta el número de habitaciones por vivienda, el valor medio de las viviendas también aumenta.

**Predicción y Re-cálculo de la Regresión con Nuevos Datos** Se predicen 5 nuevos registros y se recalcula la regresión lineal con estos nuevos datos.

```
lm_fit_boston = lm(medv ~ rm, data = boston_df)

nuevos_valores = c(4, 5, 6, 8, 9)
nuevas_predicciones = predict(lm_fit_boston, newdata = data.frame(rm = nuevos_valores))

datos_boston = data.frame(rm = nuevos_valores, medv = nuevas_predicciones)
nuevos_datos = rbind(boston_df, datos_boston)

nueva_regresion = lm(medv ~ rm, data = nuevos_datos)
summary(nueva_regresion)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = nuevos_datos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
```

```
## -23.346  -2.455   0.029   2.911  39.433
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept) -34.6706     2.5513  -13.59   <2e-16 ***
## rm            9.1021     0.4032   22.57   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 6.584 on 509 degrees of freedom
## Multiple R-squared:  0.5003, Adjusted R-squared:  0.4993
## F-statistic: 509.5 on 1 and 509 DF,  p-value: < 2.2e-16
```

La incorporación de nuevos datos no modifica de manera significativa los coeficientes del modelo. Esto sugiere que el modelo es estable y que las predicciones son correctas con la tendencia observada en los datos originales.

**Análisis de Efectos de una Modificación Extrema** Se modifica un punto del conjunto de datos por un factor extremo y recalculamos la regresión lineal para observar si es que un valor atípico podría afectar el modelo antes construido.

```
valor_extremo = nuevos_datos
valor_extremo$rm[1] = valor_extremo$rm[1] * 10

nueva_regresion = lm(medv ~ rm, data = valor_extremo)
summary(nueva_regresion)
```

```
##
## Call:
## lm(formula = medv ~ rm, data = valor_extremo)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -37.829  -5.486  -1.194   2.521  28.405
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  18.3055     1.0317  17.742  < 2e-16 ***
## rm            0.6620     0.1483   4.464 9.91e-06 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 9.136 on 509 degrees of freedom
## Multiple R-squared:  0.03767,    Adjusted R-squared:  0.03578
## F-statistic: 19.93 on 1 and 509 DF,  p-value: 9.914e-06
```

Al agregar un dato modificado son afectados los coeficientes del modelo de manera significativa. Además el error estandar residual aumenta a 9.136, lo que indica que las predicciones del modelo son menos precisas y el R^2 disminuye a 0.03767, lo que sugiere que la capacidad del modelo para explicar la variabilidad en MEDV es mucho menor al introducir el valor atípico.

En base a este análisis se demuestra que al introducir un valor extremo dentro de los datos se puede tener un impacto considerable en el modelo de regresión lineal.