**Ejercicio Propuesto3  (Resuelto)**

**Se pretende estudiar la posible relación lineal entre el precio de pisos en miles de euros, en una conocida ciudad española y variables como la superficie en m2 y la antigüedad del inmueble en años. Para ello, se realiza un estudio, en el que se selecciona de forma aleatoria una muestra estratificada representativa de los distintos barrios de la ciudad. Los datos aparecen en la siguiente tabla.**

|  |  |  |
| --- | --- | --- |
| **Precio** | **Superficie** | **Antigüedad** |
| **200****120****155****310****320****400****100****80****75****169****110****210****200****180****140****95** | **100****70****120****150****90****227****75****65****80****150****120****100****125****137****90****110** | **20****15****30****20****12****400****100****80****75****169****110****210****200****180****140****95** |

Datos del Ejercicio Propuesto 3

**Se pide:**

1. **Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación**
2. **Contrastar la significación del modelo propuesto**
3. **¿Cuánto será el precio estimado del piso en una conocidad ciudad española para una superficie en 130 m2 y 35 años de antigüedad?**
4. **¿Puede eliminarse alguna variable del modelo? Razona la respuesta**
5. **Coeficiente de determinación y de determinación corregido. Interpretación.**

### ****Solución****

**1. Ajustar un modelo de regresión lineal múltiple. Obtener una estimación de los parámetros del modelo y su interpretación**

Comenzamos introduciendo los datos en R. Para ello, creamos un vector numérico para cada variable que agrupamos en un data.frame.

> Precio <- c(200, 120, 155, 310, 320, 400, 100, 80, 75, 169, 110, 210,200, 180, 140, 95)

> Superficie <- c(100, 70, 120, 150, 90, 227, 75, 65, 80, 150, 120, 100, 125, 137, 90, 110)

> Antigüedad <- c(20, 15, 30, 20, 12, 400, 100, 80, 75, 169, 110, 210, 200, 180, 140, 95)

> datos <- data.frame (Precio, Superficie, Antigüedad)

> datos

 Precio Superficie Antigüedad

1 200 100 20

2 120 70 15

3 155 120 30

4 310 150 20

5 320 90 12

6 400 227 400

7 100 75 100

8 80 65 80

9 75 80 75

10 169 150 169

11 110 120 110

12 210 100 210

13 200 125 200

14 180 137 180

15 140 90 140

16 95 110 95

Podemos consultar los primeros registros del data.frame mediante la orden head para asegurarnos de que la introducción de datos ha sido correcta.

> head(datos)

Precio Superficie Antigüedad

1 200 100 20

2 120 70 15

3 155 120 30

4 310 150 20

5 320 90 12

6 400 227 400

Tras introducir los datos, pasamos a ajustar el modelo de regresión mediante la orden lm.

> reg\_lin\_mul <- lm(Precio ~ Superficie + Antigüedad, data = datos)

> summary(reg\_lin\_mul)

Call:

lm(formula = Precio ~ Superficie + Antigüedad, data = datos)

Residuals:

 Min 1Q Median 3Q Max

-83.357 -48.406 -0.784 33.449 167.897

Coefficients:

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

(Intercept) -19.0512 57.0197 -0.334 0.74362

Superficie 1.9241 0.6186 3.111 0.00828 \*\*

Antigüedad -0.1681 0.2470 -0.680 0.50816

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 70.2 on 13 degrees of freedom

Multiple R-squared: 0.515, Adjusted R-squared: 0.4404

F-statistic: 6.902 on 2 and 13 DF, p-value: 0.009065

A partir de los coeficientes que se muestran en la salida podemos escribir el modelo de regresión lineal estimado, que es

La interpretación de los parámetros es la que sigue:

* El precio de un piso de 0 metros cuadrados de superficie y de 0 años de antigüedad sería de -19.051 miles de euros. En este caso, la interpretación del parámetro constante no tiene sentido.
* Por cada metro cuadrado que aumenta la superficie de un inmueble, su precio se incrementa en 1.924 miles de euros, suponiendo la antigüedad como constante.
* Por cada año que aumenta la antigüedad de un piso, su precio disminuye en 0.168 miles de euros, suponiendo la superficie como constante.

**2. Contrastar la significación del modelo propuesto**

Vamos a contrastar ahora la significación del modelo de regresión lineal que se ha propuesto. Contrastar la significación del modelo equivale a comprobar si hay, al menos, una variable independiente relacionada linealmente con la variable dependiente. En tal caso, el modelo será significativo y tendrá sentido usarlo para la predicción.

El contraste que se debe resolver es el siguiente:

El estadístico de contraste y el p-valor para resolver este contraste son:

F-statistic: 6.902 on 2 and 13 DF, p-value: 0.009065

Dado que el p-valor asociado al contraste es menor que 0.05 (0.00906), rechazamos la hipótesis nula del contraste y concluimos que, al menos una de las dos variables independientes consideradas es relevante a la hora de explicar el precio de las viviendas.

**3. Cuánto será el precio estimado del piso en una conocida ciudad española para una superficie en 130 m2 y 35 años de antigüedad?**

Para responder a esta cuestión, utilizaremos el modelo de regresión lineal que hemos construido.

Bastará con sustituir el valor de las variables Superficie y Antigüedad por 130 y 35, respectivamente.

Por tanto, el precio estimado será

El precio estimado para un piso de 130 m2 de superficie y 35 años de antigüedad será de 236949 euros.

**4. ¿Puede eliminarse alguna variable del modelo? Razona la respuesta**

Vamos a comprobar ahora mediante los contrastes de significación individuales si se puede prescindir del parámetro constante o de alguna de las dos variables independientes que se han considerado. Los contrastes de hipótesis que se deben resolver son:

Para resolver estos contrastes nos fijamos en los p-valores que aparecen en la tabla de coeficientes:

Coefficients:

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

(Intercept) -19.0512 57.0197 -0.334 0.74362

Superficie 1.9241 0.6186 3.111 0.00828 \*\*

Antigüedad -0.1681 0.2470 -0.680 0.50816

En este caso, considerando un nivel de significación del 5%, podemos concluir que el parámetro asociado a la variable *Superficie* es significativamente distinto de 0 (pues su p-valor es inferior a 0.05). Esto significa que la variable *Superficie* es relevante a la hora de explicar linealmente el *Precio* de la vivienda. Por el contrario, los p-valores asociados al parámetro constante y al parámetro que acompaña a la variable *Antigüedad* son superiores a 0.05, lo que nos lleva a rechazar la hipótesis nula del contraste correspondiente. Ello implica que ambos parámetros pueden suponerse iguales a 0 y que la variable *Antigüedad* no está relacionada linealmente con la variable dependiente, por lo que no es útil para predecir el precio de una vivienda. Sería recomendable, por tanto, ajustar un nuevo modelo sin considerar el parámetro constante ni la variable *Antigüedad*.

**5. Coeficiente de determinación y de determinación corregido. Interpretación.**

Los valores del coeficiente de determinación y del coeficiente de determinación corregido son

Multiple R-squared: 0.515, Adjusted R-squared: 0.4404

Para este ejemplo, el coeficiente de determinación vale 0.515, lo que significa que el ajuste de los datos al modelo es moderado. El 51,5% de la variabilidad de la variable Precio se explica por el modelo planteado.

El coeficiente de determinación ajustado vale 0.4404, lo que significa que el ajuste de los datos al modelo es moderado-bajo.