

Ejercicios para practicar: regresión lineal I

El conjunto de datos VentaViviendas contiene información sobre el precio de venta de una serie de viviendas, junto con las características básicas de las mismas. Las variables contenidas en el fichero son:

Variable	Descripción
Price (objetivo)	Precio de venta de la vivienda
bathrooms	Número de baños (los medios se refieren a aseos)
superf	Superficie del interior de la vivienda (en m ²)
garden	Superficie del jardín (en hectáreas)
floors	Número de plantas
waterfront	¿Tiene vistas al mar? (1: sí, 0: no)
view	¿Tiene buenas vistas? (1: sí, 0: no)
condition	Estado de la vivienda (de A a D, siendo A el mejor estado)
antig	Antigüedad (en años) de la vivienda
renovated	¿La vivienda ha sido reformada? (1: sí, 0: no)
lat, long	Coordenadas de latitud y longitud de la vivienda

Los ejercicios consisten en lo siguiente (recuerda que debes llevarlos a cabo sobre un archivo qmd para posteriormente generar el html y subirlo al campus virtual):

1. Carga los datos en el entorno de Rstudio a través de la función readRDS. Utilizando el código que consideres (y los datos disponibles), indica qué precio estimarías que tiene una vivienda. A continuación, obtén el histograma de la variable dependiente y comenta el gráfico.
2. Haz un *summary* de los datos para entender bien qué significan y verificar que las variables no tengan errores. Haz un perfil de las viviendas basándote en los estadísticos observados (qué precio suelen tener, tamaño, jardín, etc.).
3. Realiza una partición *Entrenamiento-Prueba* (80-20) de los datos.
4. Construye un primer modelo de regresión lineal para la variable *Price*, en el que incluyas todas las variables input disponibles. ¿Cuántos parámetros tiene? Evalúa la calidad del modelo resultante (indicando si hay sobreajuste o no) y verifica si se cumple la hipótesis de homocedasticidad.
5. Aplica el análisis de tipo II sobre el modelo anterior, explica en qué consiste este análisis y de qué sirve. A continuación, analiza los resultados y extrae las conclusiones pertinentes.
6. Construye un nuevo modelo (que llamaremos modelo2) que contenga únicamente las 3 variables más importantes (usa la información del ejercicio anterior para saber cuáles son). ¿Cuántos parámetros tiene? ¿Este nuevo modelo tiene todos sus parámetros significativos?
7. ¿Todas las variables del modelo son significativas ahora? ¿El orden de importancia de las mismas se ha visto alterado? De ser así, explica a qué se debe el cambio.
8. ¿Ha sido correcta la eliminación de variables llevada a cabo? De no ser así, construye un tercer modelo incluyendo alguna variable de las descartadas.
9. Para el último modelo generado, interpreta el parámetro de tres variables (si es posible, la más importante de cada tipo: cuantitativa, binaria y otra cualitativa con más niveles).
10. Obtén el R² en entrenamiento y prueba y comenta la calidad del modelo, así como su estabilidad.