Ejercicios para practicar: regresión lineal II

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 (nota: como estás trabajando con datos "conocidos" no es necesario revisar el tipo de las variables, pues lo hiciste el primer día, pero se debe hacer siempre que el conjunto de datos es nuevo).
2. Realiza una partición del conjunto de datos en entrenamiento (80%) y prueba (20%). Explica a continuación de qué sirve esta partición.
3. A continuación, aplica las 6 combinaciones de selección de variables automática (forward, stepwise y backward junto con los criterios AIC yBIC). ¿Cuántos modelos diferentes se generan? ¿Cuántos parámetros tienen? ¿A qué se deben las diferencias observadas (sobre todo entre los modelos generados con AIC y los generados con BIC)?
4. Para determinar qué modelo es mejor para estos datos, aplica validación cruzada sobre los modelos diferentes del apartado 3. Determina qué modelo es preferible teniendo en cuenta el R^2, la estabilidad y el número de parámetros.
5. Para el modelo "ganador", obtén la estimación del R^2 en entrenamiento y prueba, la importancia de las variables y los coeficientes del modelo. Comenta las salidas obtenidas, sin olvidar indicar algo sobre la influencia de las variables input en la objetivo.
6. Representa en varios diagramas de dispersión la variable objetivo frente a las input numéricas para comparar la forma real de su relación y cómo se ha plasmado en el modelo de regresión lineal. Comenta los resultados.
7. NUEVO: Discretiza las variables cuantitativas input, construye los correspondientes modelos stepwise BIC y AIC y compara estos modelos con el "ganador" del apartado 4 a través de validación cruzada repetida. Determina si la discretización es recomendable en este caso y, de ser así, analiza el modelo final (estimación del R^2 en entrenamiento y prueba, la importancia de las variables y los coeficientes del modelo).
8. Utilizando los datos de la partición de entrenamiento, realiza las modificaciones precisas para poder construir modelos penalizados con R. Explica por qué es necesario llevar a cabo este paso y en qué consiste.
9. Construye un modelo LASSO con todos los valores posibles de λ y representa gráficamente los resultados (si copias el código facilitado, cuidado con el *ylim*). Comenta lo que observes y, basándote en dicha información, explica los fundamentos del modelo.
10. Para determinar el modelo penalizado óptimo, aplica validación cruzada en una combinación amplia de α y λ y representa gráficamente los resultados. Indica qué significa la línea horizontal que se añade en el gráfico y comenta los resultados que observas. Determina la combinación de parámetros óptima según la regla *1se*.
11. Construye el modelo con los parámetros α y λ óptimos. Obtén los parámetros beta del modelo y comenta su efecto sobre la variable objetivo. Obtén el R^2 en entrenamiento y prueba y comenta la calidad del modelo, así como su estabilidad.