Predicción de calidad de vinos

Tenemos 1000 vinos a los cuales les medimos 11 variables de interés: `fixed acidity`, `volatile acidity`, `citric acid`, `residual sugar`, `chlorides`, `free sulfur dioxide`, `total sulfur dioxide`, `density`, `pH`, `sulphates` y `alcohol`. A su vez, cada uno de estos vinos fue calificado con una nota del 0 al 10 por unos especialistas (que al calificarlos tuvieron en cuenta únicamente su "experiencia sensorial"). Estos datos fueron guardados en la variable `quality`.

El objetivo de este TP es poder explicar mediante una regresión lineal la variable respuesta `quality` en función de las 11 variables explicativas ya mencionadas. Se pueden incluir en el modelo transformaciones de estas variables, interacciones entre ellas (es decir, tomar el producto entre algunas variables explicativas). Se pueden también ignorar algunas de estas 11 variables.

En el conjunto de datos del archivo `vinos_1.csv` figuran las 12 medidas (11 explicativas y la respuesta) para cada uno de los 1000 vinos.

Cuando se propone un cierto modelo, pueden usar la forma que deseen entre las siguientes para "evaluarlo":

- **(Train - Test)** Separar un conjunto de entrenamiento y otro de testeo. Obtener los coeficientes estimados usando únicamente el conjunto de entrenamiento. Con estos coeficientes, predecir la respuesta del conjunto de testeo, llamemos $\widehat{Y}_1, \ldots, \widehat{Y}_{n^*}$ a dichas predicciones, donde n^* es la cantidad de elementos del conjunto de testeo. Luego, considerar, $W = (1/n^*) \sum_{i=1}^{n^*} (Y_i - \widehat{Y}_i)^2$ (en dicha suma sólo intervienen los $Y-$es correspondientes al conjunto de testeo).
- **(Leave One Out Cross-Validation)** Obtener $W = (1/n) \sum_{i=1}^{n} (Y_i - \widehat{Y}_i)^2$ donde $\widehat{Y}_i = \mathbf{X}_i^{\mathrm{T}} \widehat{\boldsymbol{\beta}}^{(-i)}$ y $\widehat{\boldsymbol{\beta}}^{(-i)}$ es el vector de coeficientes estimados usando todas las observaciones salvo la $i-$esima.
- **(K-fold Cross-Validation)** Separar a la muestra en K grupos del mismo tamaño (o lo más parecido que se pueda). Fijar uno de los grupos, digamos el grupo j, y obtener los coeficientes estimados usando todos los grupos restantes. Luego, predecir las respuestas correspondientes al grupo j, obteniendo predicciones $\widehat{Y}_1, \ldots, \widehat{Y}_{n_j}$, donde n_j es la cantidad de observaciones del grupo j y calcular $W_j = \sum_{i=1}^{n_j} (Y_i - \widehat{Y}_i)^2$, donde en dicha suma sólo intervienen $Y-$es del grupo j. Hacer lo mismo variando el j de 1 a K. Finalmente, obtener el valor $W = (1/n) \sum_{j=1}^{K} W_j$.

Si uno consideró varios modelos, se queda con el que tenga menor W.

Se pide lo siguiente:

1. Probar varios modelos. Finalmente, proponer 5 de ellos (en el informe deben decir cuáles son) y comparar los valores W de cada uno de ellos. Pueden ser útiles las funciones `lm` y `predict`.

2. En el archivo `vinos_2.csv` van a encontrar las 11 variables explicativas para cada uno de 599 vinos distintos a los 1000 originales. La variable respuesta para estos vinos nosotros la sabemos pero no va a figurar en el archivo. Usando el modelo que considere más adecuado (obtenido en el item anterior), predecir la calidad de cada uno de estos 599 vinos. Deben enviar, junto con el informe, un archivo `predichos.csv` con estas 599 predicciones (en orden), llamémoslas $\widehat{Y}_1, \ldots, \widehat{Y}_{599}$. Nosotros vamos a computar la cantidad $W = \sum_{i=1}^{599} (Y_i - \widehat{Y}_i)^2$. El que obtenga un menor W se gana una premio fantástico a definir.