Taller de práctica 6

Regresión lineal simple
(sección 3.2)

1. El siguiente conjunto de datos proviene originalmente del Instituto Nacional de Diabetes y Enfermedades Digestivas y Renales. El objetivo del conjunto de datos es estudiar de forma diagnóstica si un paciente tiene diabetes, en función de ciertas medidas de diagnóstico incluidas en el conjunto de datos. Se impusieron varias restricciones a la selección de estas instancias de una base de datos más grande. En particular, todos los pacientes aquí son mujeres de al menos 21 años de ascendencia indígena Pima

   Los datos en la base (link) contiene las siguientes columnas.

   - `Pregnancies`: Número de embarazos.
   - `Glucose`: Concentración de glucosa plasmática en la sangre a las 2 horas en una prueba de tolerancia oral a la glucosa (mg/dL).
   - `BloodPressure`: Presión arterial diastólica (mm Hg)
   - `SkinThickness`: Grosor del pliegue cutáneo del tríceps (mm)
   - `Insulin`: Insulina sérica de 2 horas (mU/mL).
   - `BMI`: Índice de masa corporal.
   - `DiabetesPedigreeFunction`: Función de pedigrí de diabetes, que califica la probabilidad de diabetes según los antecedentes familiares.
   - `Age`: Edad (años).
   - `Outcome`: variable de clase (0 o 1, o no tiene o tiene diabetes).

   Plantee un modelo para estudiar el porcentaje de diabetes presente en las mujeres a base del índice de masa corporal y la presión arterial separadamente. Luego,

   a) Ajuste cada uno de los modelos y escriba la ecuación de regresión ajustada correspondiente.

   b) Interprete las pruebas de hipótesis de no nulidad de cada parámetro. Considere una confianza del 95 %.

   c) Interprete el R^2 de cada modelo. Compare.

   d) Estudie los supuestos para cada uno de los modelos, utilizando una confianza del 95 %. Comente.

2. Los dos conjuntos de datos están relacionados con las variantes tinto (link) y blanco (link) del vino portugués "Vinho Verde". Debido a cuestiones de privacidad y logística, sólo están disponibles variables fisicoquímicas (entradas) y sensoriales (salidas) (por ejemplo, no hay datos sobre tipos de uva, marca de vino, precio de venta del vino, etc.).

   Las columnas de las bases de datos son las siguientes:

   - `fixed.acidity` (g/L): cantidad de la mayoría de los ácidos involucrados con el vino o fijos o no volátiles (no se evaporan fácilmente).
   - `volatile.acidity` (g/L): cantidad de ácido acético en el vino, que en niveles demasiado altos puede provocar un sabor desagradable a vinagre
   - `citric.acid` (g/L): cantidad de ácido cítrico, que se encuentra en pequeñas cantidades, el cual, puede añadir "frescura" y sabor a los vinos.
   - `residual.sugar` (g/L): cantidad de azúcar que queda después de que se detiene la fermentación. Es raro encontrar vinos con menos de 1 gramo/litro y los vinos con más de 45 gramos/litro se consideran dulces.
   - `chlorides` (g/L): cantidad de sal en el vino.
   - `free.sulfu.dioxide` (ppm): cantidad de la forma libre de SO2 existe en equilibrio entre el SO2 molecular (como gas disuelto) y el ion bisulfito; previene el crecimiento microbiano y la oxidación del vino.
   - `total.sulfur.dioxide` (ppm): cantidad de formas libres y ligadas de S02; En concentraciones bajas, el SO2 es prácticamente indetectable en el vino, pero en concentraciones de SO2 libre superiores a 50 ppm, el SO2 se vuelve evidente en la nariz y el sabor del vino.
   - `density` (g/cm^3): densidad del agua dependiendo del porcentaje de alcohol y contenido de azúcar.

- `pH`: describe qué tan ácido o básico es un vino en una escala de 0 (muy ácido) a 14 (muy básico); la mayoría de los vinos tienen entre 3 y 4 en la escala de pH.
- `sulphates` (mg/L): cantidad de aditivo del vino que puede contribuir a los niveles de dióxido de azufre (S02), que actúa como antimicrobiano y antioxidante.
- `alcohol` (ABV, Alcohol por volumen): el porcentaje de alcohol del vino.
- `quality`: variable de salida (basada en datos sensoriales, puntuación entre 0 y 10), un número cercano a 10 indica un vino de mayor calidad.

A continuación, utilice la base de datos asociada a los datos del vino tinto (si utiliza el comando `read.csv()`, debe agregar el argumento `sep = ";"`) para los enunciados siguientes.

a) Filtre la base de datos para conservar únicamente las columnas `pH`, `fixed.acidity` y `alcohol`. Utilice esta nueva base de datos en los enunciados posteriores.

b) Calcule la matriz de correlación y covarianza entre las variables. Interprete.

c) Elabore dos modelos de regresión lineal simple para estudiar el porcentaje del alcohol del vino, a través de las variables `pH` y `fixed.acidity` respectivamente. Para cada uno de los modelos:

1) Escriba la ecuación poblacional de los datos.
2) Ajuste el modelo y escriba la ecuación de regresión ajustada.
3) Interprete los parámetros estimados.
4) Estudie las pruebas de no nulidad de los parámetros. Utilice una confianza del 95 %.
5) Determine el R^2 e interprete.
6) Estudie los supuestos del modelo. Utilice una confianza del 95 %.

d) Compare el desempeño de ambos modelos mediante el valor del R^2.