Segundo Taller

13 de mayo de 2022

Indicaciones generales

- La fecha límite para la entrega del taller será el **lunes 30 de mayo**. Se entregará a través de la plataforma E-Aulas.
- Los talleres y proyectos se pueden realizar por parejas o individualmente.
- Se pide entregar todas las respuestas en un PDF trabajado previamente con **R Markdown**.
- Las respuestas deben estar totalmente justificadas.

1. El dataset *LungCapData.csv* contiene información sobre la capacidad pulmonar de una muestra de 700 observaciones. Además de la capacidad, se recoge información, aparentemente relevante, sobre las observaciones: Age, Height, Smoke, Gender, Caesarean (si nacieron por cesárea).

   a) Realice el ajuste de regresión lineal multiple más adecuado usando como variable independiente `LungCap`. Realice un resumen de los resultados.

   b) Resuma qué indican los betas del modelo.

   c) Determine el intervalo de confianza del 95 % para los coeficientes (use la función `confint()`.

   d) Realice las predicciones para el valor esperado de `LungCap` y los correspondientes intervalos de confianza del 95 % para valores generados en *LungCapDataPred.csv* (Ignore la variable `LungCap`). Sugerencia: use `predict()`. Determine el intervalo de confianza para la predicción (no el valor esperado).

   e) Los valores reales de la variable `LungCap` están en el dataset *LungCapDataPred.csv*. ¿El valor real está dentro del intervalo de confianza? ¿Podemos decir que el modelo ajusta correctamente?

   f) Evalúe el modelo.

   g) Si creáramos un modelo lineal simple que prediga LungCap solamente a partir de Height, ¿qué resultados obtendríamos? Grafique el diagrama de dispersion de `LungCap` y `Height` y la recta de regresión (use `abline`).

2. El dataset *LungCapData.csv* contiene la esperanza de vida por país y año. Además, incluye varias variables explicativas (el detalle se puede consultar en el fichero *descripcionVariables-Vida.txt*).

   a) Haga un análisis exhaustivo intentando predecir la variable esperanza de vida (`Life expectancy`). Imagine que este análisis lo tiene que presentar al CEO de una empresa de seguros que no sabe de estadística. Intente explicar los resultados y el proceso lo más simple posible.

   b) Intente reducir la dimensionalidad y explique el proceso.

c) Vuelva a realizar la regresión de a) pero utilizando como variables explicativas las componentes principales de b). ¿Cambian los resultados? ¿Hay alguna relación entre las variables significativas de a) con cómo están creadas las componentes principales y su significancia en el proceso?

3. (OPCIONAL) La librería `MASS` (cárguela con `library(MASS)`) contiene el dataset de `Boston`, el cual registró la variable `medv` (valor medio de una casa) para 506 barrios en Boston. En este ejercicio, se buscará predecir la variable `medv` usando 13 predictores tales como: `rm` (número promedio de habitaciones por casa), `age` (promedio de la edad de las casas), y `lstat` (porcentaje de hogares con bajo nivel socioeconómico).

   Recuerde que este dataset fue visto en clase y encontrábamos que había problema con los supuestos y con el ajuste del modelo. ¿Se puede mejorar?