PEC Global: enunciado

UOC - Master BI - Data Mining

Enero 2019

Contents

Introducción

En esta PEC trabajaremos con un dataset extraído de UCI - Machine Learning Repository sobre consumo de combustible de coches (en mpg). El dataset se llama "auto-mpg.txt" y consta de 9 variables:

1. mpg: v. continua
2. cilindros: v. discreta
3. cilindrada: v. continua
4. caballo de vapor (CV): v. continua
5. peso: v. continua
6. aceleración: v. continua
7. año del modelo: v. discreta
8. origen: v. discreta
9. nombre del coche (o ID): factor (único)

Carga de datos

Cargad el fichero de datos e inspeccionar los tipos de variables. En el caso de que alguna variable no sea del tipo adecuado, aplicad la transformación necesaria. Pista: si una variable que debiese ser de tipo numérico tiene otro tipo es posible que no hayamos identificado bien los missings.

Análisis descriptivo

Obtened el número de filas y columnas. Para las variables numéricas describid los valores que toman (media, mediana, cuartiles, ...). Podéis usar gráficos para complementar el análisis. Finalmente, anotad el número de missings para cada variable.

Outliers

Investigad si existen outliers en las variables numéricas. En caso de detectar outliers, eliminadlos del dataset.

Nota: debéis escribir un código R automático (es decir, evitar la inspección y la eliminación manual).

Intervalo de confianza

Calculad el IC al 97 % de mpg.

Nota: Se deben realizar los cálculos manualmente. No se pueden usar funciones R que ya calculen el IC directamente (*t.test* o similar). Sí que podéis usar funciones como *qnorm*, *pnorm*, *qt* y *pt*.

Contraste de hipótesis

Un político afirma que el valor esperado de mpg es, como mínimo, de 25 unidades. Basándonos en nuestros datos, ¿podemos rechazar esta afirmación con un nivel de confianza del 95%?

Nota: Se deben realizar los cálculos manualmente. No se pueden usar funciones R que ya calculen el IC directamente (*t.test* o similar). Sí que podéis usar funciones como *qnorm*, *pnorm*, *qt* y *pt*.

a) Escribid el contraste a realizar.

b) Calculad el estadístico de contraste, el valor crítico y el p-valor.

c) Interpretad el resultado.

Correlación

---

a) Obtened las correlaciones entre variables numéricas. Nota: podéis tener problemas por los missings. Usad: ?cor/help(cor) para obtener ayuda y encontrar una solución.

b) Mostrad las correlaciones con un gráfico. Pista: una opción interesante es usar la librería "corrplot".

c) Interpretad el resultado del apartado anterior.

Regresión lineal

---

a) Estimad un modelo que explique la variable mpg en función de cilindrada, aceleración, año, y origen.

b) Interpretad el resultado del modelo, indicando si los coeficientes son estadísticamente significativos.

c) Predecid el valor para una nueva observación: coche de cilindrada 145, aceleración 15.50, año 76 y origen 2.

ANOVA

---

a) Realizad grupos para la variable año.

   G1: año $<$ 73

   G2: 73 $<=$ año $<$ 76

   G3: 76 $<=$ año $<$ 79

   G4: año $>$ 79

b) Aplicad ANOVA para identificar si existen diferencias en mpg entre los grupos creados. Interpretad el resultado.

c) Explicad el significado de los cálculos SSW, SSB y SST de un análisis de varianza. ¿Cómo se usan estos cálculos para investigar si hay diferencias entre los grupos?

d) En el caso de que se detecten diferencias significativas entre las medias de los distintos grupos, calculad un test a posteriori (post-hoc test) como el test de Tukey. Interpretad el resultado.