APRENDIZAJE AUTOMÁTICO
Trabajo-1

Valoración: 18 puntos

Fecha de entrega: 29 de Marzo

Cuestionario (6 puntos): estará disponible más adelante

Ejercicios de Implementación: 12 puntos (Justificar las respuestas en todos los casos)

Ejercicio.-1 (2 puntos)
Usar la base de datos de Boston que es parte de la librería MASS en R

1. Leer la descripción de la base de datos "help(Boston)". Tratar de comprender el problema, identificar las variables del problema y hacer una valoración de la relevancia de las mismas para el estudio.
2. Realizar tres gráficos con las parejas de columnas que considere de más interés. Describir lo que has encontrado justificando la elección de las columnas estudiadas para el problema.
3. ¿Existen predictores asociados con la tasa de crimen per capita? Si es así explicar la relación.
4. Hay algún suburbio de Boston que parezca tener una alta tasa de: a) criminalidad, b) altos impuestos, c) alumnos-por-profesor. Comentar el rango de cada predictor.
5. ¿Cuántos suburbios de este conjunto de datos bordea o cruza el rio Charles?
6. ¿Cuál es la media de la tasa alumnos-profesor entre las ciudades de este conjunto de datos?
7. ¿Qué suburbio de Boston tiene el valor mediano más bajo de propietarios viviendo en sus casas? ¿Cuáles son los valores de los otros predictores para este suburbio, y como se comparan estos valores con el rango global de los otros predictores? Comentar los resultados.
8. ¿Cuántos de los suburbios tienen en promedio más de siete habitaciones por vivienda? ¿más de ocho por vivienda? Haga algún comentario al caso cuyo promedio de habitaciones por vivienda sea mayor de ocho.

Ejercicio-2 (5 puntos)

a) Predecir la ratio de crímenes per-capita usando las otras variables en la base de datos Boston:
   i) Para cada predictor ajustar un modelo de regresión lineal simple con la variable respuesta. Describir los resultados
   ii) ¿En qué modelos existe una asociación estadísticamente significativa entre predictor y respuesta?
   iii) Crear algún gráfico que muestre los ajustes y que valide las respuestas anteriores.

b) Ajustar un modelo de regresión múltiple usando todos los predictores.
   i) Describir los resultados.
   ii) ¿Para qué predictores podemos rechazar la hipótesis nula, H_0: β_j=0?
c) Comparación de los resultados encontrados en los dos puntos anteriores:
   i) Crear un dibujo gráfico 2D donde cada punto del gráfico representa en el eje-x el valor de los coeficientes calculados en la regresión univariante para cada predictor y el eje-Y el valor calculado por la regresión múltiple para ese mismo predictor . Comentar el gráfico.
d) ¿Existe evidencia de asociación no-lineal entre los predictores y la respuesta?
   i) Apoyar la contestación ajustando un modelo lineal cúbico para cada variable predictor ($Y = \beta_0 + \beta_1 X + \beta_2 X^2 + \beta_3 X^3 + \varepsilon$. Comentar los resultados

Ejercicio.-3 (5 puntos)
Usar la base de datos "Auto data set". Leer la base de datos.

1) Realizar una representación gráfica matricial ("scatterplot") que incluya todas las representaciones de cada dos variables del conjunto de datos. Comentar lo que considere que dicha representación nos aporta en la relación entre variables.
2) Calcular la matriz de correlaciones entre variables cuantitativas usando la función cor(). Comentar los valores respecto de las gráficas del punto anterior.
3) Usar la función lm() para realizar una regresión lineal múltiple usando "mpg" como la respuesta y todas la demás variables, excepto "name", como predictores. Usar summary( ) para imprimir los resultados. Comentar los siguientes aspectos del resultado justificando la respuesta.
   a) ¿Existe alguna relación entre los predictores y la respuesta?
   b) ¿Qué predictores parece tener una relación estadísticamente significativa con la respuesta?
   c) ¿Que sugiere el coeficiente para la variable "year"?
4) Usando el modelo ajustado obtener los intervalos de confianza al 95% para los coeficientes.
5) Usar la función plot() para realizar dibujos de diagnóstico sobre la regresión lineal. Comentar cualquier problema que observe en el ajuste.
   a) ¿Se observan valores "outliers " en los residuos?
   b) ¿Considera que hay algún punto con inusual alta influencia sobre el ajuste?
6) Usar los símbolos "*" y ":" de R para ajustar un modelo de regresión lineal con términos de interacción
   a) ¿Hay alguna interacción que sea estadísticamente significativa?

Los BONUS solo se tendrán en cuenta si se ha obtenido al menos el 50% de los puntos en los ejercicios obligatorios.

BONUS. 1 (1 punto)
Aplique diferentes transformaciones sobre los datos de Auto data set, tales como.: $\log(X)$, $\sqrt{X}$, X^2 y reajuste el modelo. Comente los resultados encontrados.

BONUS.2 (1 punto)

Hacer un estudio sobre la co-linealidad de las variables de la base de datos de Boston. Identificar aquellas variables que considere son co-lineales. Presentar gráficos y resultados que avalen las conclusiones.

BONUS.3 (1 punto)
Hacer un estudio sobre el comportamiento de la varianza del error residual en el modelo de regresión lineal múltiple ajustado la base de datos de Auto. Presentar conclusiones y resultados. (1 punto)

BONUS.4 (2 puntos)
Sea X el vector de variables aleatorias de una muestra.
1.- Demostrar la relación que existe entre la matriz $\boldsymbol{X}^T\boldsymbol{X}$ que aparece en la ecuación de estimación de los coeficientes de un modelo lineal por mínimos cuadrados y la matriz de covarianza del vector X.
2.- Suponer que los elementos del vector X son incorrelados. ¿Cómo es la matriz $\boldsymbol{X}^T\boldsymbol{X}$?
3.- Bajo la hipótesis de variables muestrales X_i incorreladas en un modelo lineal, ¿Cómo afecta la varianza de las X_i a la estimación de los parámetros de $\boldsymbol{\beta}$?

Informe a presentar

Para este trabajo como para los demás proyectos debe presentar un informe escrito con sus valoraciones y decisiones adoptadas en cada uno de los apartados de la implementación. Incluir los gráficos generados. También deberá incluirse una valoración sobre la calidad de los resultados encontrados. (hacerlo en pdf, MS Word o en texto plano). **Sin este informe no se considera que el trabajo ha sido presentado.**

Normas de la entrega de Trabajos: EL INCUMPLMIENTO DE ESTAS NORMAS SIGNIFICA PERDIDA DIRECTA DE 1 PUNTO CADA VEZ QUE SE DETECTE UN INCUMPLIMIENTO.

1. Cada contestación del cuestionario siempre incluirá la correspondiente pregunta.
2. El código se debe estructurar en un único script R con distintas funciones o apartados, uno por cada ejercicio/apartado de la práctica.
3. El path que se use en la lectura de cualquier fichero auxiliar de datos debe ser siempre “datos/nombre_fichero”. Es decir, crear un directorio llamado “datos” en el directorio donde se desarrolla y se ejecuta la práctica.
4. Todos los resultados numéricos serán mostrados por pantalla, parando por cada apartado. No escribir nada en el disco.
5. La práctica deberá poder ser ejecutada de principio a fin sin necesidad de ninguna selección de opciones. Para ello fijar al comienzo los parámetros por defecto que se consideren óptimos.
6. La práctica debe de ejecutarse de principio a fin sin errores.
7. El código debe estar obligatoriamente comentado explicando lo que realizan los distintos apartados y/o bloques.
8. Poner puntos de parada para mostrar imágenes o datos por consola.
9. Todos los ficheros a entregar juntos se podrán dentro de un fichero zip, cuyo nombre debe ser Apellido1_P[1-3].zip.
10. ENTREGAR SOLO EL CODIGO FUENTE, NUNCA LOS DATOS.

Forma de entrega: Subir el zip al Tablón docente de CCIA.