Curso: Estadística Multivariada

Tarea 7

Fecha de entrega: martes 17 de marzo de 2020

Instrucciones

- Subirla a la plataforma en un zip que contenga el código y el archivo pdf con los resultados

1 Problemas

1. Recuerde que la matriz **H** está definida por $\boldsymbol{H} = \boldsymbol{Z}(\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'$ con elementos diagonales h_{jj}.
   (a) Muestre que $\boldsymbol{H}$ es una matriz idempotente.
   (b) Muestre que $0 < h_{jj} < 1$, $j = 1, 2, ..., n$, y que $\sum_{j=1}^{n} h_{jj} = r+1$, donde r es el número de variables independientes en el modelo de regresión (De hecho,$(1/n) \leq h_{jj} < 1$).

2. Regresión lineal múltiple.
   (a) Utilice el conjunto de datos *states.txt*.
   (b) Ajuste un modelo que pronostique la energía consumida per capita (energía) con respecto al porcentaje de residentes que viven en áreas metropolitanas (metro). Reporte lo siguiente:
      i. Examine / grafique los datos antes de aplicar el modelo
      ii. Escriba el modelo e interpretelo
      iii. Grafique el modelo para buscar desviaciones de los supuestos del modelado
   (c) Seleccione uno o más predictores adicionales para agregar al modelo y repita los pasos anteriores. ¿Es este modelo significativamente mejor que el modelo con la variable metro solo como único predictor?

3. Los datos del archivo "costofliving.txt" enumeran algunas estadísticas del costo de vida para cada uno de los 50 estados de los USA. Los tres costos son: alquileres de apartamentos, costo de casas y el índice de costo de vida.
   (a) Realiza una regresión lineal multivariada para explicar estas tres métricas en términos de las poblaciones estatales e ingresos medios. ¿Son útiles estas variables independientes para explicar conjuntamente las variables de costo?
   (b) Ajusta tres modelos de regresión lineal de manera separada y verifica la utilidad de las variables independientes en cada uno ellos. Compara los resultados con los obtenidos en el inciso (a)

4. Considere los datos de contaminación del aire (datoscontaminacion). Sea $Y_1 = NO_2$ y $Y_2 = O_3$, las dos respuestas (contaminantes) correspondientes a las variables predictoras Z_1=viento y Z_2=radiacion solar

(a) Desarrolle un análisis de regresión usando únicamente la primera respuesta Y_1

   i. Sugiera y ajuste modelos de regresión lineal apropiados
   ii. Analice los residuales
   iii. Construya un intervalo de predicción del 95% para NO_2 correspondiente a $z_1 = 10$ y $z_2 = 80$

(b) Desarrolle un análisis de regresión multivariada múltiple usando las respuestas Y_1 y Y_2

   i. Sugiera y ajuste modelos de regresión lineal apropiados
   ii. Analice los residuales
   iii. Construya un elipsoide de predicción del 95% para ambas NO_2 y O_3, para $z_1 = 10$ y $z_2 = 80$. Compare esta elipse con el intervalo de predicción en la parte (a)iii. Comente