Fundamentos de Análisis de Datos

Trabajo Práctico 2

Profesora: Dra. Andrea Alejandra Rey

Ejercicio 1. Vamos a trabajar con el archivo `MedidasCorporales.xlsx`.

1. ¿Cuántos registros y cuántas variables tiene el conjunto de datos? ¿Todas las variables son numéricas?

2. Regresión lineal simple: ¿cómo influye la altura en el peso?

   a) Realizar una regresión lineal simple y escribir el modelo teórico resultante.

   b) ¿Cuáles son las estimaciones de la ordenada al origen y de la pendiente? ¿Son estos coeficientes de regresión significativos?

   c) Calcular el error estándar residual, el coeficiente de determinación R^2 y su valor ajustado. ¿Qué se podría concluir sobre la bondad de ajuste del modelo?

3. Regresión lineal múltiple: ¿cómo influyen las medidas consideradas en el peso?

   a) Guardar en la variable `n` la cantidad total de registros.

   b) Fijar una semilla igual a 1234 y correr el siguiente comando:

   ```
   muestras <- 1:n %>% 
               createDataPartition(p = 0.8, list = FALSE)
   ```

   Usar la variable `muestras` para separar aleatoriamente el conjunto de registros en conjuntos de entrenamiento y de prueba. ¿Qué porcentaje de los datos integra cada uno de estos conjuntos?

   c) A partir del conjunto de entrenamiento, realizar el modelo de regresión lineal múltiple con todas las variables involucradas.

   d) A partir del conjunto de entrenamiento, realizar el modelo de regresión lineal múltiple con las variables que presenten un nivel de confianza de al menos el 95 %.

   e) Utilizando el conjunto de prueba, calcular el error cuadrático medio:

   $$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2,$$

   donde n es la cantidad total de predicciones, y_i es el valor real de la variable respuesta e $\hat{y}_i$ es el valor predicho. ¿Cuál de los dos modelos muestra un valor menor de MSE?

Ejercicio 2. Vamos a trabajar con el archivo `Dolor.xlsx`. El mismo contiene una muestra de 3504 pacientes que acudieron a un centro de salud presentando dolor en el pecho. Para estos pacientes, se recogieron diversas medidas. En el caso de las variables estrechamiento de arterias coronarias y de tres arterias coronarias, ambas son variables binarias que indican la presencia de estrechamiento en alguna de las arterias coronarias de al menos un 75 % (valor igual a 1) o no (valor igual a 0). En cuanto a la variable sexo, 0 corresponde a masculino y 1 a femenino.

1. Realizar un modelo de regresión logística simple, que estudie la presencia de estrechamiento en alguna arteria coronaria explicada por el colesterol. Escribir la ecuación del modelo resultante y calcular la probabilidad de que una persona con un nivel de colesterol igual a 199 presente estrechamiento arterial.

2. Realizar un modelo de regresión logística múltiple, que estudie la presencia de estrechamiento en alguna arteria coronaria usando todas las variables no categóricas como variables explicativas. ¿Qué puede decirse sobre la significancia de las variables predictoras?

3. Replicar el modelo anterior pero diferenciando entre mujeres y varones. ¿Existen diferencias entre las significancias de las variables explicativas en función del sexo? Justificar la respuesta.

Ejercicio 3. Vamos a trabajar con el archivo `Europa.xlsx`.

1. ¿Cuáles son las variables de interés?

2. Calcular la matriz de covarianza de los datos y analizar si es inversible.

3. ¿Cuál es el mayor autovalor de la matriz de covarianzas?

4. Realizar un PCA y hallar la cantidad necesaria de componentes principales para explicar al menos el 90 % de la varianza total de los datos.

5. Realizar e interpretar un gráfico que visualice la contribución de las variables en las dos primeras componentes principales.

Ejercicio 4. Vamos a considerar el conjunto de datos `JohnsonJohnson` disponible en `R`.

1. ¿Qué tipo de datos mide esta serie de tiempo? ¿Cuál es el período de tiempo analizado?

2. Graficar la serie tiempo, junto con sus descomposiciones aditiva y multiplicativa. ¿Se observa tendencia? ¿Se observa estacionalidad?

3. Analizar la conveniencia de aplicar la transformación de Box-Cox.

4. Usar toda la información de todos los años salvo los dos últimos para realizar un modelo ARIMA automático y uno personalizado, explicando la elección de los órdenes elegidos y teniendo en cuenta lo concluido en el punto anterior. Trabajar con $1 \leq p \leq 14$ y $1 \leq q \leq 30$.

5. ¿Cuáles son los parámetros obtenidos para el modelo ARIMA automático?

6. Predecir las temperaturas del último año utilizando los dos modelos ARIMA hallados. Calcular el criterio de información de Akaike (AIC) y el error de porcentaje medio absoluto (MAPE) en cada caso y decidir, en función de estos valores, qué modelo realiza las mejores predicciones.