Trabajo Práctico N°2

Fundamentos de Análisis de Datos
Maestría en Ciencia de Datos

Estudiantes:
Ing. Flores, Matías Gabriel
matflores@itba.edu.ar

Ing. Loiseau, Matías
mloiseau@itba.edu.ar

Docente:
Dra. Rey, Andrea A.

Septiembre 2023

Contents

1 Ejercicio N° 1

Vamos a trabajar con el archivo *MedidasCorporales.xlsx*.

1.1 Primer punto

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

El set de datos de Medias Corporales tiene 507 registros y 24 variables. Verificamos que no haya nulos. Si, todas las variables son numéricas. Esto se puede apreciar en la tabla 1

	Valor
Registros	507
Columnas	24
Nulos	0

Table 1: Cantidad de campos en el set de datos.

1.2 Segundo punto

Consigna: 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 R2 y su valor ajustado. ¿Qué se podría concluir sobre la bondad de ajuste del modelo?

En primer lugar, creamos el modelo de regresión lineal simple y el resultado de la función es

$$\hat{y} = -105.011 + 1.018x \tag{1}$$

donde x representa la altura e $\hat{y}$ representa el peso estimada por el modelo.

La estimación de la ordenada al origen es -105.011 y de la pendiente es 1.018. Estos coeficientes de regresión tienen un alto grado de significancia según los resultados obtenidos.

El error estándar residual nos determina que la estimación promedio del peso puede diferir en 9.308kg. Esto se puede observar en la Tabla 2. En cuanto al coeficiente de determinación R^2 consideramos que es significativamente moderado.

Variable	Valor
Residuo	9.308
R^2	0.5145
Valor ajustado	0.5136
Estadístico-F	354.4
P-Valor	$< 2.2e^{-16}$

Table 2: Parámetros de salida del modelo.

Concluimos que el 51% de las variaciones de los pesos están explicadas por la altura de las personas. Además, podemos concluir que no existe una relación lineal entre la altura y el peso, ya que el estadístico-f observado tiene un valor alto con p-valor menor a 0.05.

1.3 Tercer punto

Consigna: 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:

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

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?

El porcentaje de set de datos de entrenamiento es del 80% y mientras que el de pruebas es del 20%.

A partir de los dos modelos realizados llevamos a cabo el calculo del error cuadrático medio para cada uno. Como resultado obtuvimos que los valores que se detallan en la tabla 3.

Modelo	MSE
1	5.630
2	5.369

Table 3: Errores cuadráticos medios.

El modelo 2 es el que presenta un menor error cuadrático medio. Por lo tanto podemos concluir que es el modelo más eficiente.

2 Ejercicio N° 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.

2.1 Primer punto

Consigna: 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.

$$p(x) = \frac{\exp(-0.7525280 + 0.0062268x)}{1 + \exp(-0.7525280 + 0.0062268x)} \tag{2}$$

$p(199) = 0.6193053$

2.2 Segundo punto

Consigna: 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?

Lo que podemos decir sobre las variables predictoras es que ´´días con síntomas´´ no tiene significacncia, mmientras que las variables de edad y colesterol tienen una fuerte significancia.

2.3 Tercer punto

Consigna: 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.

3 Ejercicio N° 3

Vamos a trabajar con el archivo Europa.xlsx.

3.1 Primer punto

Consigna: ¿Cuáles son las variables de interés?

3.2 Segundo punto

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

3.3 Tercer punto

Consigna: ¿Cuál es el mayor autovalor de la matriz de covarianzas?

3.4 Cuarto punto

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

3.5 Quinto punto

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

4 Ejercicio N° 4

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

4.1 Primer punto

Consigna: ¿Qué tipo de datos mide esta serie de tiempo? ¿Cuál es el período de tiempo analizado?

El dataset contiene las ganancias en dolares trimestrales desde el año 1960 hasta 1980.

4.2 Segundo punto

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

ANALIZANDO GRAFICO ADITIVA 1 y CALCULANDO LA FUERZA DE LA TENDENCIA DA 0.97

Efectivamente hay una fuerza a una tendencia

ANALIZANDO LA ESTACIONALIDAD DE LA ADITIVA Y CALCULANDO LA FUERZA DE LA ESTACIONALIDAD 0.32

Por lo que vemos es una estacionalidad aditiva debil.

ANALIZANDO LA DESCOMP MULTIPLICATIVA 1 Y CALCULANDO LA FUERZA DE TENDENCIA NOS DA 0.99

Efectivamente hay una fuerte tendencia

ANALIZANDO LA DESCOMP MULTI 2 Y CALCULANDO LA FUERZA DE ESTACIONALIDAD NOS DA 0.54

Hay un valor significativo bajo

y por eso tenemos que aplicar la transofrmación box-cox

4.3 Tercer punto

Consigna: Analizar la conveniencia de aplicar la transformación de Box-Cox.

es convenientge utilizarlo para estacionar los datos

4.4 Cuarto punto

Consigna: 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 $1 \leq q \leq 30$.

4.5 Quinto punto

Consigna: ¿Cuáles son los parámetros obtenidos para el modelo ARIMA automático?

4.6 Sexto punto

Consigna: Predecir las ganancias 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.