Actividad 3: Modelización predictiva

Enunciado A3

Semestre 2021.1

Índice general

En esta actividad se usará el fichero de datos (dat_Air) que contiene información sobre diferentes parámetros sobre la calidad del aire de una determinada ciudad europea en el año 2021. Estos datos han sido medidos en tiempo real en diferentes estaciones distribuidas en distintas zonas. Para nuestro estudio se ha seleccionado los datos recopilados de este año por una de las estaciones móviles. Se muestran las medidas de una serie de variables, tanto meteorológicas como de los principales contaminantes del aire (gases y partículas).

Todas ellas contribuyen para determinar el Índice de Calidad del Aire (ICA).

Las variables del fichero de datos son:

- Estacion: Estación móvil.
- Latitud: Latitud del lugar de medición.
- Longitud: Longitud del lugar de medición.
- Fecha: Fecha de medición.

- Periodo: Mediciones cada hora. Periodo de 1 a 24 horas (diarias).
- SO2: Concentración de SO2 (dióxido de azfre) en μ_g/m^3.
- H2S: Concentración de H2s (ácido sulfhidrico) en μ_g/m^3.
- NO: Concentración de NO (óxido nítrico) en μ_g/m^3.
- NO2: Concentración de (dióxido de nitrógeno) en μ_g/m^3.
- NOX: Concentración de NOX (óxidos de nitrógeno) en μ_g/m^3.
- O3: Concentración de Ozono en μ_g/m^3.
- PM10: Partículas en suspension <10 en μ_g/m^3.
- PM25: Partíulas en Suspension PM 2,5 en μ_g/m^3.
- BEN: Concentración de benceno en μ_g/m^3.
- TOL: Tolueno en μ_g/m^3.
- MXIL: MXileno en μ_g/m^3.
- Dir_Aire: Dirección del viento en grados.
- Vel: Velocidad del viento en m/sg.
- Tmp: Temperatura en grados centígrados.
- HR: Humedad relativa en % de hr.
- PRB: Presión Atmosférica en mb.
- RS: Radiación Solar W/m^2.
- LL: Precipitación en l/m^2.

1 Regresión lineal

La calidad del aire ha sufrido cambios que afectan a nuestro modo de vida, por lo que resulta necesario estudiarlo. Para ello se toman medidas de la emisión de diferentes contaminantes y de factores metereológicos como por ejemplo el viento, la precipitación, radiación solar o la temperatura, con el fin de buscar relaciones entre dichas variables.

En este estudio se quiere demostrar la existencia de relación lineal entre los contaminantes atmosféricos y las variables metereológicas.

1.1 Estudio de correlación lineal

Se pide calcularla matriz de correlación entre las variables siguientes: Contaminantes: O3, NO2 y PM10, junto con las variables metereológicas: Tmp, HR, RS, Vel y Dir.

a) ¿Cual de los contaminantes atmosféricos citados anteriormente, tienen una mayor relación lineal con la RS? Interpretar las relaciones de dicho contaminante con la RS y también con el resto de variables metereológicas.

b) Se toma la media diaria de cada una de las variables del apartado a) y posteriormente se estudia de nuevo la relación pedida en dicho apartado. ¿Existe alguna diferencia en la relación entre las nuevas variables construídas con los valores medios diarios, con respecto a los resultados obtenidos anteriormente?

1.2 Modelo de regresión lineal

Se quiere explicar el nivel de ozono en función de la radiación solar.

a) Estimar por mínimos cuadrados ordinarios un modelo lineal que explique la variable O3 en función de la radiación solar (RS). Se evaluará la bondad del ajuste, a partir del coeficiente de determinación.

b) Para calcular el índice de calidad del aire, se establecen diferentes categorías, según sea la concentración de cada contaminante. En este apartado se tomará como contaminante la concentración de PM10 y se establecerán las siguientes categorías, para construir el PM10_cat (Indice de calidad del Aire, en función de PM10):

Muy buena: valores de (0 a 40],
Buena: valores de (40 a 60],
Mejorable: valores de (60 a 120],
Mala: valores de (120 a 160],
Muy mala: valores de (160 a 724]

Se pide, construir un modelo de regresión lineal, tomando como variable dependiente (O3) y la variable explicativa PM10_cat. Interpretar los resultados.

Nota: Este apartado se podría interpretar también mediante el ANOVA. Dicho modelo se verá en la actividad A4.

1.3 Modelo de regresión lineal múltiple

Se quiere explicar el nivel de ozono en función de la radiación solar (RS),concentración de dióxido de nitrógeno (NO2), temperatura (Tmp) y dirección del aire (Dir_Aire).

a) Primero, se añadirá al modelo del apartado a), la variable explicativa (Dir_Aire).¿El modelo ha mejorado?

b) Posteriormente se añade al modelo anterior la variable (NO2). ¿Existe una mejora del modelo?

c) Se toma la variable (Tmp) y se añade al modelo anterior. Se pide comprobar la presencia o no de colinealidad entre las variables (RS) y (Tmp). Podéis usar la librería (faraway) y estudiar el FIV (factor de inflación de la varianza). Según la conclusión obtenida, discutir si sería indicado o no añadir la variable (Tmp) al modelo. De ser afirmativa la respuesta, construye el modelo e interpreta el resultado.

1.4 Diagnosis del modelo

Para la diagnosis se escoge el modelo construído en el apartado b) y se piden dos gráficos: uno con los valores ajustados frente a los residuos (que nos permitirá ver si la varianza es constante) y el gráfico cuantil-cuantil que compara los residuos del modelo con los valores de una variable que se distribuye normalmente(QQ plot). Interpretar los resultados.

1.5 Predicción del modelo

Según el modelo del apartado c), calcular la concentración de O3, si se tienen valores de RS de 180, NO2 de 15, Dir_Aire de 250 grados y Tmp de 20 grados centígrados.

2 Regresión logística

Se quiere estudiar la concentración de O3 del aire de una determinada ciudad.

Primero se creará una nueva variable dicotómica llamada icO3 (índice de calidad del aire basado en O3). Se codificará de la siguiente manera:

buena: valores de (0 a 80],
mejorable: valores de (80 a 100]

Posteriormente se recodificará como valor 0 la categoría " buena". En caso contrario se codificará con el valor 1.

Nota: Dicho índice de calidad se ha recodificado conforme a nuestros datos.

2.1 Estudio de relaciones entre variables. Análisis crudo de posibles factores de riesgo

a) Se visualiza la relación entre icO3 y las variables independientes: RS, Vel y HR. Para ello se recodificaran las variables RS y Vel, dejando la variable cuantitativa HR, tal como está en la base de datos.

Para comprobar si existe asociación entre las variable dependiente y cada una de las variables explicativas, se aplicará el test Chi-cuadrado de Pearson. Un resultado significativo nos dirá que existe asociación.

Se procederán a categorizar las variables explicativas de la siguiente forma:

Radiación solar (RS_cat2):
normal_baja:(0 a 100],
normal_alta: valores de (100 a 700]

Velocidad del viento (Vel_cat2):
flojo: valores de (0 a 3],
moderado: valores de (3 a 10]

b) Posteriormente, para conocer el grado de dicha asociación, se calculará las OR (Odds-Ratio). Importante: Para el cálculo de las OR, se partirá de la tabla de contingencia y se calculará a partir de su fórmula. Debéis implementar dicha fórmula en R. Interpretar las OR calculadas.

2.2 Modelo de regresión logística

a) Estimad el modelo de regresión logística tomando como variable dependiente icO3 y variable explicativa RS_cat2. Calculad la OR a partir de los resultados del modelo y su intervalo de confianza. ¿Se puede considerar que la radiación solar es un factor de riesgo? Justifica tu respuesta.

b) Se crea un nuevo modelo con la misma variable dependiente y se añade al apartado a) la variable TMP. Interpretar si nos encontramos o no ante una posible variable de confusión.

c) Se añade al modelo del apartado a) la variable HR. Estudiar la existencia o no de interacción entre las variables explicativas RS_cat2 y HR. Interpretar.

d) Se crea un nuevo modelo con las variables explicativas RS_cat2 y Dir_Aire. ¿Existe una mejora del modelo?

2.3 Predicción

Según el modelo del apartado d), calculad la probabilidad de que la concentración de O3 sea o no superior a 80, con unos valores de RS_cat2="Normal_alta"y Dir_Aire=40.

2.4 Bondad del ajuste

Usa el test de Hosman-Lemeshow para ver la bondad de ajuste, tomando el modelo del apartado d). En la librería ResourceSelection hay una función que ajusta el test de Hosmer- Lemeshow.

2.5 Curva ROC

Dibujar la curva ROC, y calcular el área debajo de la curva con el modelo del apartado d). Discutir el resultado.

3 Conclusiones del análisis

En este apartado se deberán exponer las conclusiones en base a los resultados obtenidos en todo el estudio.

Puntuación de los apartados

- Apartado 1.1 (10%)
- Apartado 1.2 (10%)
- Apartado 1.3 (15%)
- Apartado 1.4 (5%)
- Apartado 1.5 (5%)
- Apartado 2.1 (10%)
- Apartado 2.2 (15%)
- Apartado 2.3 (10%)
- Apartado 2.4 (5%)
- Apartado 2.5 (5%)
- Apartado 3 y Calidad del informe dinámico (10%)