Actividad 3: Modelización predictiva

Enunciado A3

Semestre 2021.2

Índice

En esta actividad se usará el fichero de datos (dat_Air_Stations) que contiene información de diferentes parámetros sobre la calidad del aire de una determinada ciudad del Norte de España en el año 2018. Los datos nos muestran concentraciones por hora de varios contaminantes atmosféricos (gases y partículas) como SO2, NO2, O3 y PM10, entre otros, monitoreados en cinco estaciones. Por otro lado en dos de las cinco estaciones, se han recogido medidas de variables meteorológicas. El periodo que abarca este estudio está comprendido entre el 1 de Enero 2018 al 31 de Diciembre de 2018. Estos datos han sido medidos en tiempo real.

La contaminación del aire representa un importante riesgo medioambiental y para la salud, tanto en los países desarrollados como en los países en desarrollo, por lo que su estudio es muy necesario.

Las variables del fichero de datos son:

- Estación: Estación móvil.
- Nombre: Nombre de la 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 azufre) 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.
- CO: Concentración de CO en μ_g/m^3.
- O3: Concentración de Ozono en μ_g/m^3.
- PM10: Partículas en suspensión <10 en μ_g/m^3.
- PM25: Partículas en suspensión 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: MiXileno en μ_g/m^3.
- dd: Dirección del viento en grados.
- vv: 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 en W/m^2.
- LL: Precipitación en l/m^2.

---

1 Regresión Lineal

La exposición a la materia particulada (PM10), al ozono (O3), al dióxido de nitrógeno (NO2) y el dióxido de azufre (SO2), plantean graves riesgos para la salud. Las directrices de la OMS sobre la calidad del aire establecen los límites sobre estos principales contaminantes atmosféricos.

PM10: Límite de 45 microgramos de partículas por cada metro cúbico μ_g/m^3. SO2: Límite de 40 μ_g/m^3. NO2: Límite de 25 μ_g/m^3. O3: Límite de 60 μ_g/m^3.

El índice de calidad del aire se calcula de forma individual teniendo en cuenta cada uno de dichos contaminantes. Todos estos valores están referidos a la **media diaria**.

Con referencia a **valores máximos diarios** se tomarán los valores de 100 μ_g/m^3 para O3 y de 120 μ_g/m^3 para NO2. Tanto para PM10 y SO2, se tomarán como referencia únicamente los valores medios diarios para comparar.

1.1 Estudio comparativo entre estaciones.

a) Estudio de los valores medios y máximos diarios de cada contaminante. Para cada una de las estaciones de monitoreo, se calcularán los valores máximos y medios diarios de cada contaminante. Posteriormente se hará una comparativa entre las cinco estaciones en base a dichos valores. Interpretad teniendo en cuenta los límites mencionados anteriormente.

b) Representad gráficamente la evolución de cada uno de los contaminantes en cada estación. Se tomarán los valores máximos diarios.

c) Estudio de correlación lineal.PAra ello se seleccionan las dos estaciones con registros meteorológicos: Estación de Montevil y Estación Avenida Constitución. Para cada una de las estaciones, calcular la matriz de correlación entre los contaminantes citados anteriormente y las variables meteorológicas: Temperatura (TMP), Humedad Relativa (HR), Radiación solar (RS), velocidad del viento (vv), precipitaciones (LL) y Presión barométrica (PRB).Interpretad.

Nota: La matriz de correlación será calculada en base a los valores máximos de cada contaminante.

Nota2: El motivo de estos primeros apartados es tomar un primer contacto sobre las posibles diferencias entre estaciones, así como hacerse una idea de las relaciones existentes entre las variables, pero para construir los modelos de regresión se tomarán los datos por hora.

1.2 Modelo de regresión lineal

Como he mencionado arriba, para construir los modelos de regresión, se tomarán los valores de las variables escogidas por hora, tal como aparecen en la base de datos original.

a) Se pide crear un modelo de regresión lineal, tomando como variable dependiente (O3) y variable explicativa (NO2). Se evaluará la bondad del ajuste, a partir del coeficiente de determinación. Interpretad.

b) Se añade al modelo anterior el nombre de las estaciones (Nombre). Interpretad.

1.3 Modelo de regresión lineal múltiple

Se quiere construir un modelo de regresión múltiple con el que podamos predecir la concentración de ozono (O3) en las zonas de Montevil y Avenida de la Constitución.

a) Se pide dos modelos (uno para cada estación) tomando como variable dependiente el nivel de ozono (O3) en función de la concentración de dióxido de nitrógeno (NO2) y diferentes variables meteorológicas como vv (velocidad del viento), RS (radiación solar), HR (humedad relativa) y LL (precipitaciones).

b) Se añade a los modelos anteriores la variable Temperatura (TMP). De ser necesario, se pide comprobar la presencia o no de colinealidad entre las variables (vv) y (TMP). Podéis usar la librería (faraway) y estudiar el FIV (factor de inflación de la varianza). Discutid si sería indicado o no añadir la variable (TMP) a cada uno de los modelos.

1.4 Diagnosis del modelo

Para la diagnosis se escoge el último modelo construído para la estación de Montevil 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). Interpretad los resultados.

1.5 Predicción del modelo

Según el modelo del apartado anterior, calculad la concentración de O3, si se tienen valores de NO2 de 40, vv de 2, RS de 100, HR de 80, LL de 0.10 y TMP de 25.

2 Regresión logística

Para construir las nuevas variables y los modelos de regresión logística, se tomarán los valores de las variables escogidas por hora, tal como aparecen en la base de datos original.

En este apartado se tomarán como contaminantes la concentración de PM10 y de O3. Se procederá a calcular los índices de calidad (icPM10 e icO3) de la forma siguiente:

PM10 recodificada: (**icPM10**)

aceptable: valores de (0 a 45],
mejorable: valores de (45 a 180]

O3 recodificada: (**ic03**)

aceptable: valores de (0 a 60],
mejorable: valores de (60 a 170]

La variable RS también será recodificada:

RS recodificada (**RS_re**):

normal_baja:(0 a 100],
normal_alta: valores de (100 a 700]

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

2.1 Análisis crudo. Cáculo de OR

Se creará una nueva variable con los meses del año a partir de la variable Fecha, llamada **month**.

a) Se calculará las OR (Odds-Ratio) entre cada una de las variables dependientes **icPM10** y **icO3** y las variables explicativas radiación solar recodificada (RS_re) y (month) en la estación de Montevil. 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. ¿Se puede considerar que la radiación solar y el mes del año son factores de riesgo? Justifica tu respuesta e interpreta las OR.

b) Idem para la estación de Avenida Constitución.

2.2 Modelo de regresión logística

Para la estación de Montevil del apartado anterior:

a) Se pide construir un modelo de regresión logística tomando como variable dependiente **icPM10** y variables explicativas (RS_re), (vv) y (PRB). Interpretad y calculad las OR.

b) Se añade al modelo del apartado anterior la variable (month).¿Existe una mejora del modelo?. Justificad e interpretad.

c) Se añadirá al modelo anterior como variable explicativa la variable (TMP). Justificad la presencia o no de una posible interacción con (RS_re). ¿Se podría estar ante una variable de confusión?. Razona tu respuesta.

2.3 Predicción

Según el modelo del apartado b), calculad la probabilidad de que la concentración de PM10 sea o no superior a 45, con unos valores de vv= 0.6, RS_re="Normal_alta",PRB= 1013, en el mes de Agosto.

2.4 Bondad del ajuste

Usa el test de Hosman-Lemeshow para ver la bondad de ajuste, tomando el modelo del apartado b). 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 b). 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 (15%)
- Apartado 1.2 (5%)
- Apartado 1.3 (15%)
- Apartado 1.4 (5%)
- Apartado 1.5 (5%)
- Apartado 2.1 (10%)
- Apartado 2.2 (15%)
- Apartado 2.3 (5%)
- Apartado 2.4 (5%)
- Apartado 2.5 (5%)
- Apartado 3 y Calidad del informe dinámico (15%)