Regresión Avanzada
Trabajo Práctico Integrador

Contexto:

Consideramos el conjunto de datos `COVID.txt` que contiene información relativa a la mortalidad en diferentes países en la primera ola de la epidemia de COVID19 ocurrida durante el año 2020. El objetivo de este trabajo es construir un modelo lineal que nos permita comprender cuales fueron los factores determinantes en las muertes que se produjeron por este virus, que tuvo una ámplia variabilidad en diferentes países. Para realizar este análisis contamos con datos provenientes de 139 países, es importante destacar que la base de datos original contaba con 71 países más que fueron eliminados de la misma ya que sus registros se encontraban incompletos.

El modelo se basa en una serie da variables demográficas, geográficas y de salud publica:

- `Hombres80` población de hombres mayor a 80 años (% de la población masculina).
- `Mujeres80` población de mujeres mayor a 80 (% de la población femenina).
- `Pobla80:` promedio entre `Female80` y `Male80`.
- `Pobla65:` población mayor a 65 años (% de la población).
- `PoblaMid:` población entre 15 y 64 años (% de la población).
- `PoblaData:` población en 2018 (en 100 millones de personas).
- `PoblaDens:` densidad poblacional (cientos de personas por km cuadrado de superficie)
- `Mujeres:` Población femenina (% de la población total)
- `Urbano:` Población urbana (% de la población total)
- `ExpectVida:` Esperanza de vida al nacer, total (años)
- `NeontlMort:` Tasa de mortalidad neonatal, neonatal (por 1000 nacidos vivos)

- `DisMort`: Mortalidad por enfermedades cardiovasculares, cáncer, diabetes o enfermedad renal crónica entre las edades exactas de 30 y 70 (%)
- `Lesion`: Causa de muerte por lesión (% del total)
- `EnfNoTrans`: Causa de muerte por enfermedades no transmisibles (% del total)
- `EnfTrans`: Causa de muerte por por enfermedades transmisibles y materna, prenatal y condiciones nutricionales (% del total)
- `PBI`: producto bruto interno per cápita PPP (miles de dólares internacionales corrientes)
- `Tuberculosis`: Incidencia de tuberculosis (por 1000 personas)
- `Diabetes`: Prevalencia de la diabetes (% de la población de 20 a 79 años)
- `Medicos`: Médicos (cada 1000 personas)
- `Camas`: Camas de hospital (cada 1000 personas)
- `ImmunSaramp`: Inmunización de sarampión (% de chicos entre 12 y 23 meses)
- `TempMarzo`: Temperatura promedio en marzo.
- `HipTen.H`: Prevalencia bruta de hipertensión en 2010 en hombres
- `HipTen.M`: Prevalencia bruta de hipertensión en 2010 en mujeres
- `HipTen`: Promedio de `HT.women` y `HT.men`
- `BCG`: Estrategia de inmunización 0 = selectiva, 1 = todos.
- `BCGf`: Es la variables `BCG` escrita como un factor.
- `Tiempo`: número de días entre el primer caso de COVID-19 registrado y el 31 de diciembre de 2019.

Por otra parte como variable de respuesta tenemos

- `l10muertes.permil`: $\log_{10}(muertes.permil+1)$ donde `muertes.permil` es el número de muertos cada millón de habitantes.

Ambas variables figuran en la base de datos.

1. Hacer un análisis exploratorio de las variables. Graficar los boxplots de las variables `muertes.permil` y `l10muertes.permil` en relación a `fBCG` y entender porque conviene `l10muertes.permil`

2. Hacer un análisis exploratorio con el resto de las variables.

3. Considerando únicamente como variables regresoras `PoblaDens, Pobla80,Urbano, Tuberculosis, Camas, TempMarzo, PBI`

   (a) Ajustar un modelo lineal para predecir las muertes, elegir como variable de respuesta la que resulta más adecuada.

   (b) Analizar el leverage para todos los países. Que se observa?

   (c) Analizando los residuos. Determinar si hay observaciones o variables que tienen alta influencia.

   (d) Analizar posibles interacciones con la variable PBI.

   (e) Seleccionar un modelo a partir de todo lo realizado anteriormente.

   (f) Construir intervalos de predicción de nivel 95% para cada uno de los países analizados. Graficar los valores predichos, si les sale con sus intervalos de predicción versus los valores observados.

4. Considerando todas las variables, analizar problemas de colinealidad.

   (a) En primer lugar hay problemas evidentes por construcción de las variables, mostrar como se evidencian y solucionarlos. Obtener el mejor modelo posible.

   (b) Realizar un procedimiento de selección de variables, se pueden considerar las dos variables de respuesta posibles y analizar con cuál conviene seguir adelante.

   (c) Comparar el modelo obtenido en el item 4 (b) con el obtenido en 3 (d).

5. En base al modelo final que obtienen, desde la perspectiva de la política pública qué se puede sugerir cambiar a corto plazo?