MRLM COVID

Equipo X

07 May, 2023

Contents

1 Introducción

1.1 Problema de interés: Análisis de new_deaths_smoothed_per_million con un MRLM en Bélgica, Suecia, Grecia y Portugal.

Queremos analizar el efecto que tuvieron las variables elegidas con respecto las muertes del COVID, y elegimos 4 países de Europa para ello.

Los países fueron seleccionados por dos criterios: - Su GDP per capita, dos países con PIB de 20,000 USD, Grecia y Portugal y dos países arriba de 50,000, Bélgica y Suiza - Los cuatro países tienen la misma cantidad de población, entre 10 millones y 12 millones.

1.1.1 Breve explicación de la base de datos

Glosario de Y y los regresores que utilizamos.

La información está tomando en cuenta los 7 días de la semana.

Smoothed_data: Los datos que estan "smoothed" significa que las cuentas incluyen los datos reportadas pero también se agregan los probables, estamos utilizando estas variables porque nuestro análisis incluye dos países con información reservada, Grecia y Portugal.

Y nuestra variable a analizar - new_deaths_smoothed_per_million: Son las nuevas muertes atribuidas al COVID-19. Por millón de personas

Location: Es la ubicación geografica

Date: Del 2020-02-01 al 2023-02-24

new_cases_smoothed_per_million: Nuevos casos confirmados de COVID-19 por millón de personas.

reproduction_rate: Estimación a tiempo real de la reproducción efectiva del COVID

icu_patients_per_million: Numero de pacientes con COVID que están en cuidados intensivos, por millón de personas

hosp_patients_per_million: Numero de pacientes con COVID en hospitales, por millón de personas.

new_tests_smoothed_per_thousand: Son nuevos exámenes de COVID, por miles.

positive_rate: Son exámenes que salen positivos de COVID

tests_per_case: Pruebas realizadas de COVID

people_vaccinated_per_hundred: Son el total de personas con al menos una vacuna, por 100

new_people_vaccinated_smoothed_per_hundred: Nuevas personas por 100, que reciben una vacuna diariamente.

stringency_index: Son las reacciones políticas basadas en 9 indicadores. en un valor del 0 al 100, 100 siendo el más estricto. Ejemplo de ello sería cerrar escuelas.

excess_mortality_cumulative_per_million: Es la diferencia entre las personas que se reportan como muertas y las proyectadas.

2 Marco teórico

2.1 Conceptos básicos

2.2 Supuestos del modelo

2.3 Método de selección de variables

Al momento de seleccionar variables eliminamos variables que explicaban lo mismo pero tenían algún tipo de transformación lineal. Un ejemplo de estas son total_cases y total_cases_per_million, ya que estamos explicando una variable que ya tiene una transformación en per_million nos vamos a quedar con todos los regresores que están en per_million.

Con las variables que quedaron de esa selección, utilizamos el método "Backward" de selección de variables, el cual utiliza todas las variables y va eliminadolas dependiendo de su p-value, le indicamos a la función a que eliminara las variables que tuvieran un p-value mayor a .05.

Ya por último, con la selección restante, utilizamos el Variance Inflation Factor (VIF), su expresión matemática es (1/(1-Riˆ2)), el cual nos ayuda a cuantificar la intensidad que hay de multicolinealidad entre los regresores. Aceptamos Variables con VIF menores de 10 en el modelo.

2.4 Limitaciones del modelo

Una gran limitación con la que nos encontramos en el MRLM es que aun cuando los datos no se distribuyen normal, asumimos una distribución normal para hacer inferencia. El MRLM es robusto para variables que no se distribuyen normal, no obstante, eso nunca va a ser un modelo exacto apegado a la realidad

El MRLM también asume que hay una linealidad entre la variable dependiente y la independiente, pero cuando hay datos que no son lineales entre sí o que tienen relaciones complejas, el MRLM no logra predecir el comportamiento de las variables de forma acertada.

3 Análisis exploratorio de datos

3.1 Análisis de la base de datos

De la librería de tidyr, utilizamos la función filter, para filtrar la base de datos por los 4 países seleccionados

Utilizamos de la librería dplyr la función de select, para eliminar las variables que la selección backward nos quitó, ya que de esa forma las eliminabamos de la base a modelar.

Los NA, los hicimos 0 ya que en todos los casos si hace sentido que NA sea = a 0 ya que no hay datos negativos.

Seleccionamos new_deaths_smoothed_per_million ya que es una variable que nos explica cuantas personas fueron muriendo debido al COVID y que tanto funcionaron las medidas de restricción, hospitalarias y las vacunaciones para que ese numero diminuyera.

3.2 Selección de la variable explicativa

Fuimos seleccionando los regresores por varios métodos, explicados en el punto 2.3, pero igual hacemos énfasis en las admisiones hospitalarias y las restricciones por país.

4 Modelo de regresión lineal múltiple

4.1 Justificación de la selección del MRLM

4.2 Análisis y significancia de los coeficientes

```
Anova Table (Type II tests)

Response: new_deaths_smoothed_per_million
                                          Sum Sq   Df  F value    Pr(>F)
location                                  2163.1    3 215.4033 < 2.2e-16 ***
date                                       224.6    1  67.0828 3.772e-16 ***
new_cases_smoothed_per_million              47.0    1  14.0357 0.0001827 ***
reproduction_rate                          503.5    1 150.4083 < 2.2e-16 ***
icu_patients_per_million                   160.1    1  47.8385 5.605e-12 ***
hosp_patients_per_million                 1851.5    1 553.1144 < 2.2e-16 ***
new_tests_smoothed_per_thousand            696.3    1 208.0101 < 2.2e-16 ***
positive_rate                               24.2    1   7.2191 0.0072519 **
tests_per_case                             408.2    1 121.9591 < 2.2e-16 ***
people_vaccinated_per_hundred              107.3    1  32.0545 1.637e-08 ***
new_people_vaccinated_smoothed_per_hundred 562.5    1 168.0545 < 2.2e-16 ***
stringency_index                           560.2    1 167.3475 < 2.2e-16 ***
excess_mortality_cumulative_per_million     37.9    1  11.3367 0.0007692 ***
Residuals                                10309.9 3080
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Se puede observar en la tabla ANOVA, que todos los coeficientes seleccionados son significativos, esto es debido a que el p-value es menor a 0.05.

4.3 R^2 y R^2 Ajustada (Bondad de ajuste)

```
La R^2 es [1] 0.7015499

La R^2 ajustada es [1] 0.7000964
```

4.4 Supuestos y validación del modelo

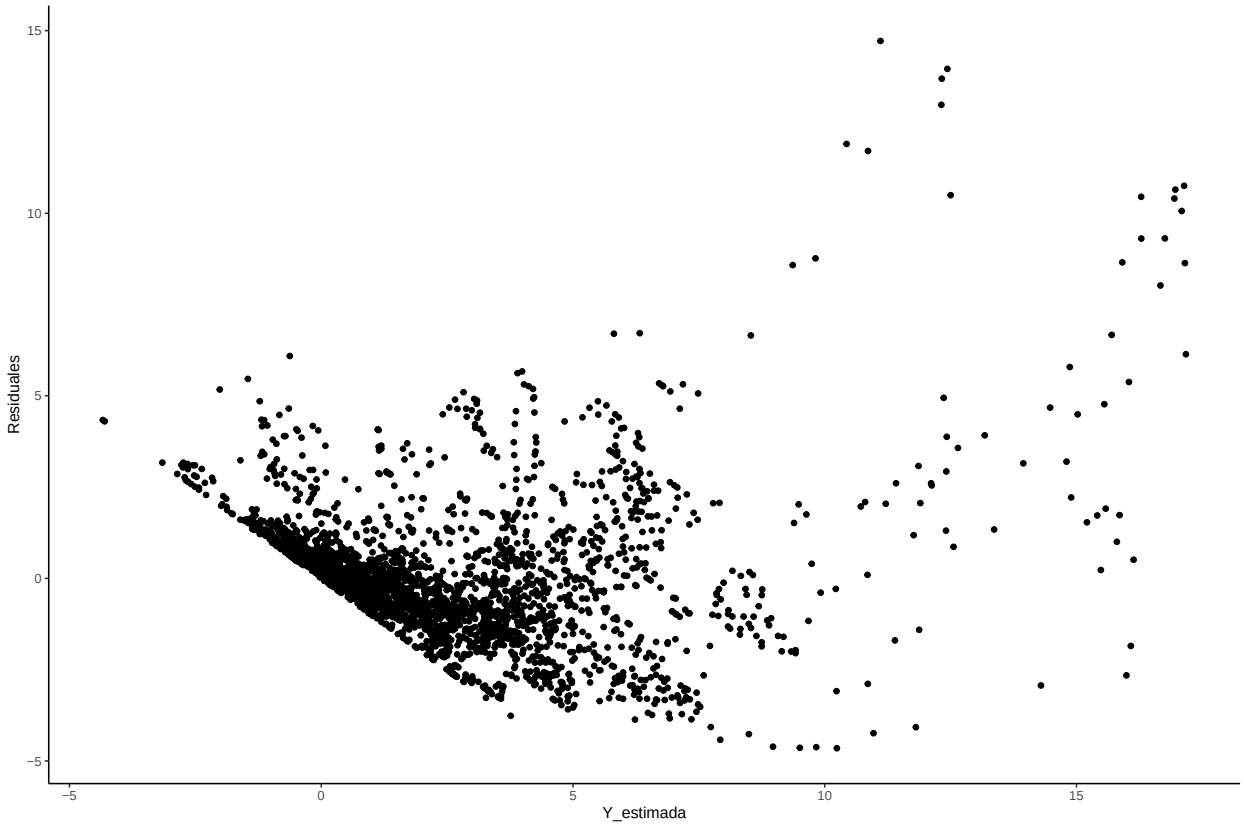
4.4.1 Análisis de residuales

4.4.1.1 Comprobación de la linealidad de la Fn de regresión Eso lo comprobamos con las R^2 ya que sabemos que matemáticamente es SSE/SST y es el porcentaje de variabilidad que es explicada por el modelo.

```
Una R^2 ajustada de [1] 0.7000964
```

nos dice que hay una bondad de ajuste del 70% de nuestro modelo con Y

4.4.1.2 Heterocedasticidad Comprobamos heterocedisticidad (la varianza de los errores es constante), lo comprobamos con un gráfico comparando los residuales con las Y observadas ($\hat{y}$), para esto tenemos que hacer un DF con ambos vectores obtenidos de nuestro modelo



Se puede observar que no hay un patrón en sí en el gráfico, con esto podemos asumir que hay heterocedisticidad.

4.4.1.3 Independencia en los errores

4.4.1.4 Presencia de errores atípicos

4.4.1.5 Verificación de la normalidad en los errores

5 Predicciones

6 Conclusiones

6.1 Porqué es útil el modelo

6.2 Cómo mejorar el modelo

7 Bibliografía

Edouard Mathieu, Hannah Ritchie, Lucas Rodés-Guirao, Cameron Appel, Charlie Giattino, Joe Hasell, Bobbie Macdonald, Saloni Dattani, Diana Beltekian, Esteban Ortiz-Ospina and Max Roser (2020) - "Coronavirus Pandemic (COVID-19)". Published online at OurWorldInData.org. Retrieved from: 'https://ourworldindata.org/coronavirus' [Online Resource]