
**Universidad de San Andrés**

**Departamento de Matemática y Ciencias**

**Maestría en Ciencia de Datos**

**Regresión Avanzada**

**Informe sobre el trabajo práctico integrador**

**-**

Alumnos:

Dominutti, Nicolás

Suárez Gurruchaga, Carlos Roque

Telechea, Hernán

**Introducción**

En el siguiente informe, estudiaremos una base de datos relacionada con el auge del COVID en el mundo, a principios del 2020. Analizaremos las variables presentes, el modelo de regresión y diferentes transformaciones a partir del estudio de sus residuos, interacciones y colinealidad. Las tablas, procedimientos y gráficos empleados para este informe están disponibles en el R Markdown.

**Análisis de variables**

La base de datos contiene 139 registros analizados a partir de 35 variables diferentes. 3 de esas variables son cualitativas ("geoId", "CntrName" y "BCGf"), mientras que las demás son cuantitativas. Arrancamos comparando las variables a través de boxplots y gráficos "pairs", y logramos detectar algunas potenciales relaciones; entre ellas:

- El logaritmo de la variable "muertes.permil" parece tener mayor linealidad que su variable original. Identificamos que una estrategia de vacunación selectiva contra BCG podría tener que ver con el aumento de muertes en diferentes países.
- La relación lineal más fuerte se da entre "Pobla80" y "Hombre80" ($\rho = 0.99$) o "Mujer80" ($\rho = 0.98$). Por el contrario, la más débil parece ser entre las muertes y otras enfermedades, o con la densidad poblacional ($\rho < |0.05|$ para ambos casos).
- Las muertes aumentan a medida que la población es más longeva ($\rho = 0.67$ con "Pobla80").
- Temperaturas más elevadas parecen generar menores chances de mortalidad (80% de los países cuya temperatura promedio en marzo de 2020 fue mayor a 20 grados tuvo menos de 1 muerte logarítmica por mil). Sin embargo, estos datos son de la época en la que la enfermedad aún no estaba propagada en el hemisferio sur (más cálido).
- Ambientes más urbanizados tienen mayor número de muertes, aunque la población es mayor allí.
- En la mayoría de los países, la mortalidad sobre las infecciones fue menor a 0.05. Los países más severos pertenecen al hemisferio norte y son europeos (salvo México, Belice y Yemén). Incluso, a través de un *pairplot*, pudimos ver que la relación entre "Pobla80" y las muertes por mil se fortalece donde la mortalidad del virus sobre los casos totales fue mayor al .1%.

**Aplicación y análisis del modelo lineal**

Como paso siguiente, corrimos un modelo lineal OLS con "l10muertes.permil" como variable de respuesta y las variables regresoras "PoblaDens", "Pobla80", "Urbano", "Tuberculosis", "Camas", "TempMarzo" y "PBI". Del análisis (disponible en el markdown), cuyas variables resultaron significativas y el R2 ajustado parece ser aceptable en 0.6038, identificamos algunos países potencialmente outliers. Revisamos leverage, distancias de Cook, realizamos test de rangos para evaluar heterocedasticidad y KS para evaluar normalidad, y llegamos a la siguiente conclusión: las

observaciones Singapur, Japón y Catar son de las más influyentes por tener mayor leverage y distancia, aunque no son consideradas outliers porque sus residuos estandarizados son similares al resto de los países. Respecto a los residuos, logramos observar cierta normalidad (p-valor = 0.89, no rechazo H0) y heterocedasticidad (p-valor = 0.03, rechazo H0).

Complementamos el análisis de residuos del modelo OLS con una regresión robusta (MM), que nos indica que no existen diferencias significativas entre los betas estimados de ambos modelos; por ende, no habría observaciones viciando los resultados (ninguna es asignada un peso 0). Ante esto, elegimos seguir trabajando con el modelo original.

Seguido a esto, corrimos un análisis *stepwise* para evaluar la importancia de las variables seleccionadas. El criterio mantiene todas las covariables del modelo original e indica que "Pobla80" y "Camas" son las que más influyen en la predicción de muertes logarítmicas. También probamos incluir algunas interacciones que en el *plot* inicial parecían estar más relacionadas, "PBI"*"Camas" y "PBI"*"Pobla80". Sin embargo, al volver a correr *stepwise*, estas no fueron consideradas. Por lo tanto, seguimos con el modelo OLS original.

**Revisión de colinealidad**

Para simplificar el análisis, descartamos variables que, por construcción, sabemos que nos otorgan información similar (eliminamos "geoID", "casos", "muertes", "Mujeres80", "Hombres80", "EnfNoTrans", "HipTen.H", "HipTen.M" y "BCGf"). Sobre las restantes variables, aplicamos el factor de inflación de la varianza (VIF) para estudiar la multicolinealidad y hallamos lo siguiente:

- Países con PBI más altos suelen tener mejor calidad de vida, por lo que esperaríamos una menor probabilidad de muerte por otras enfermedades, como las nutricionales. Además, al poseer una capacidad hospitalaria mayor, habría mayor acceso a tratamientos contra diversas enfermedades.
- Países con mayor infraestructura y más tratamientos disponibles deberían aumentar la expectativa de vida de la población y disminuir la mortalidad neonatal.
- Países con mayor PBI, que invierten más en infraestructura médica y tienen mayor expectativa de vida, fomentarían aún más el desarrollo profesional, entre ellos, de más médicos.
- Por la tan alta correlación, podemos llegar a pensar que la población mayor a 80 años está incluida en la población mayor a 65.
- La mortalidad neonatal estaría correlacionada negativamente con enfermedades transmisibles.

*Selección de variables*

Una vez analizados todos los puntos anteriores, decidimos volver a correr un modelo de selección de variables para confirmar nuestros análisis parciales. En resumen, optamos por la transformación logarítmica de las muertes por mil como variable regresora porque el R2 ajustado es mayor y hay

mayor número de estimadores significativos, aunque no todos. Por ello, aplicamos el método *stepwise* bajo el criterio Akaike para estudiar la colinealidad. Este modelo, que terminamos eligiendo como mejor, redujo la cantidad de covariables de 23 a 10, aumentó el R2 ajustado (de 0.67 a 0.69), bajó el RSE (de 0.43 a 0.42) y la cantidad de variables no significativas bajó (de 16 a 2). Aún con un método backward llegamos a la misma conclusión.

*Métodos de regularización*

Como último paso del análisis, intentamos reducir la dimensionalidad del problema a partir de la regularización Lasso, con el ECM como métrica de estudio. Obtenemos como valores óptimos 0.0039 (lambda.min) y 0.3470 (lambda+1se). Con lambda.min, no logramos reducir la dimensionalidad pero sí lo hacemos con lambda+1se: nos quedamos con las covariables "casos.permil", "Pobla80" y "ExpectVida". Elegimos añadir la variable BCG porque, por el tipo de problema que se está analizando, consideramos que es importante incluir cuestiones de vacunación.

Notamos que, al utilizar el método de regularización Lasso, el R2 no supera lo obtenido con el criterio *stepwise* (0.69 vs 0.63). Por lo tanto, vamos a conservar este último como el modelo definitivo. A esto, aplicamos un análisis de residuos y finalmente detectamos un outlier (alto leverage y alto residuo), Catar, que tiene un valor extremo de "casos,permil" pero uno de los más bajos de "muertes.permil". Tal diferencia podría deberse a algún error en la recolección de datos. Tomando este punto de evaluación como disparador, podríamos utilizar distintas técnicas para intentar mejorar la performance de nuestro mejor modelo. Estas pruebas están detalladas en el markdown y el Anexo 1.

**Conclusiones**

Si pensáramos este análisis como punto de partida para analizar potenciales políticas públicas, observamos que la estrategia de inmunización es un factor muy importante para disminuir las muertes por COVID. Sugeriríamos a las autoridades que opten por una vacunación total (no selectiva) y que se concientice sobre los beneficios de vacunarse. Creemos necesario tener más centros de vacunación.

Los aislamientos preventivos y campañas de concientización sobre cuidados especiales (uso de barbijos, higiene de manos, distancia social, etc.) también podrían ser acciones efectivas, dado que otro factor influyente en las muertes es el total de casos por mil habitantes.

Por último, la población más afectada por el virus es la de mayor edad (más de 65 años). Sugerimos implementar políticas de concientización sobre la vulnerabilidad de este sector. También sugerimos escalar los programas de asistencia para personas mayores y promover voluntariados en torno a esta actividad (por ejemplo, que alguien compre por esa persona para evitar aglomeraciones).

**Anexo 1: pruebas de optimización de modelo ante valores atípicos**

Prueba 1: eliminamos la observación de Catar

Logramos que la regresión aumente su R2 ajustado en +10%, aunque solo 4 de los betas estimados resultan significativos (vs. 8 del último modelo elegido). Esto indicaría la presencia de multicolinealidad pero como ya trabajamos esto en secciones anteriores, puede ser otra la pregunta a responder. Por ejemplo, ¿existen nuevos outliers que no hayamos detectado hasta el momento?

Prueba 2: eliminamos la observación de Singapur

Una observación del último modelo propuesto tiene alto leverage, bajo residuo estandarizado y una larga distancia de Cook. Este caso es Singapur (considerado un punto extremo anteriormente), país con la mayor densidad poblacional del set de datos (+79). Puede ser que, debido a este punto, nuestro modelo no lo ajuste bien a dicho país (su beta estimado para la variable PoblaDens es de -0.023).

Al eliminar la observación, volver a ajustar y correr un proceso de selección de variables, observamos que el R2 ajustado mejora de 0.796 a 0.798 y la mayoría de las variables resultan significativas. Los puntos atípicos ya no son tan visibles; el modelo parece performar bien.

Prueba 3: regresión robusta

Volvemos al modelo previo a eliminar Catar y Singapur, para luego correr una regresión robusta y analizar los residuos. Hay 5 valores atípicos (vs 2), que no eliminaríamos pero lo probamos. El modelo OLS mejora 1.6% (0.7978 vs 0.8112). Con la regresión robusta, los betas casi no varían respecto del OLS y no observamos outliers. Este termina siendo nuestro mejor modelo pero, como eliminamos 5 valores (Canadá, Catar, Estados Unidos, India y Japón), no lo recomendamos.

Prueba 4: transformaciones

La última acción por realizar para corregir el problema de outliers es transformar alguna variable. Corrimos un test de Box-Tidwell sobre “casos.permil”, que nos sugiere aplicar un logaritmo (lambda=0 cae dentro de los intervalos de confianza). Encontramos un mejor modelo, con un R2 ajustado que aumentó 4.5% (0.8116 a 0.8475) y no eliminamos observaciones. Los residuos parecen ser normales y homocedásticos, y no encontramos outliers al correr el modelo robusto.

Finalmente, elegimos el modelo OLS luego de esta transformación (R2 ajustado aumenta 40%, RSE cae 62%, no hay multicolinealidad y los outliers parecen estar controlados). “Casos.permil” es definitivamente una variable preponderante (tiene el beta estimado más grande) pero antes no la consideramos. Pasó lo mismo con “BCG”. En cuanto a los intervalos de predicción, el último modelo los tiene más estrechos, se acerca bastante a los valores reales de la variable de respuesta.