Desarrollo interrumpido: La paradoja entre la mejora sanitaria y el estancamiento de la esperanza de vida en México (2001-2019)

Luis Erick Palomino Galván
DEMAT (Departamento de Matemáticas)
Universidad de Guanajuato
Guanajuato, Mx.
luis.palomino@cimat.mx

***Abstract*—En este documento analizamos la base de datos *Life Expectative and socio economic* proporcionada por el Word Bank Open Data, cuantificando los factores determinantes de la esperanza de vida en 174 países durante el periodo 2001 a 2019. Posteriormente, se examina el desarrollo de México en el mismo periodo respecto a estos factores determinantes. El análisis demostró la siguiente contradicción: a pesar de las mejoras sistémicas de saneamiento, existe un estancamiento en la Esperanza de vida en México.**

***Index Terms*—Esperanza de vida, Saneamiento, Prevalencia de desnutrición, Bayesian ridge.**

I. Introduction

[1] Word Bank Open Data define la esperanza de vida al nacer como el número de años que viviría un recién nacido si los patrones de mortalidad prevalecientes al momento de su nacimiento se mantuvieran constantes a lo largo de su vida. Esta métrica es clave para evaluar la salud de la población.

El estudio [3] de Acemoglu y Johnson, ha demostrado la relación entre el aumento de la esperanza de vida y la mejora del crecimiento económico (PIB por cápita), controlando los efectos fijos de cada país. Sin embargo, concluyen que es necesario un análisis más profundo para determinar cómo la asignación de la riqueza de un país a través de ciertas inversiones en salud, educación y medio ambiente tiene un efecto general en la determinación de la esperanza de vida.

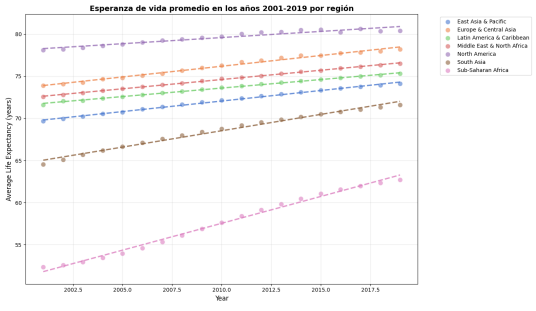

Fig. 1. Regresión lineal a la Esperanza de vida por región (2001-2019)

El objetivo de este documento es cuantificar los factores determinantes de la Esperanza de vida a nivel global durante las últimas dos décadas y contrastar el impacto positivo de las mejoras estructurales frente al bajo crecimiento en la esperanza de vida en México debido al aumento de, identificando a esta última como un freno en el desarrollo.

II. Exploración inicial

La base de datos proporcionada por el World Bank Open Data considera la información de 174 países a lo largo de los años de 2001 a 2019. Los factores que se consideran surgen al plantear las siguientes preguntas:

- ¿Cuál es el impacto del gasto en salud y educación en la esperanza de vida?
- ¿Cómo afecta la prevalencia de la desnutrición y las enfermedades trasmisibles a la esperanza de vida?
- ¿Factores como la corrupción y tasa de desempleo afectan la esperanza de vida?
- ¿El aumento de las emisiones de CO2 reduce la esperanza de vida?¿Es significativo?

A. Definición de las variables

El conjunto de datos considera las siguientes variables:

- **Región:** Región geográfica mundial en la que se ubica el país.
- **Grupo de ingresos:** Clasificación del país según su nivel de ingresos.
- **Esperanza de vida:** Años promedio que se espera que viva una persona al nacer.
- **Prevalencia de desnutrición:** Porcentaje de la población cuyo consumo habitual de alimentos es insuficiente.
- **CO2:** Emisiones de dióxido de carbono (métricas ambientales).
- **Gasto en Salud:** Gasto corriente en salud expresado como porcentaje del PIB (excluye gastos de capital).
- **Gasto en Educación:** Gasto público general en educación expresado como porcentaje del PIB.
- **Desempleo:** Porcentaje de la fuerza laboral que no tiene empleo pero está disponible y buscando trabajo.
- **Corrupción:** Índice de percepción de transparencia, rendición de cuentas y corrupción en el sector público.
- **Saneamiento:** Porcentaje de la población que utiliza servicios de saneamiento gestionados de forma segura (incluye alcantarillado, fosas sépticas o letrinas mejoradas).
- **AVAD por Lesiones:** Años de Vida Ajustados por Discapacidad (DALYs) debido a lesiones; suma de años

perdidos por mortalidad prematura y años vividos con discapacidad.

- **AVAD por enfermedades transmisibles:** Años de Vida Ajustados por Discapacidad debidos a enfermedades infecciosas o transmisibles.
- **AVAD por enfermedades no transmisibles:** Años de Vida Ajustados por Discapacidad debidos a enfermedades crónicas o no transmisibles.

B. Exploración de la base de datos

Al explorar la base de datos, encontramos que las variables: Lesiones, Enfermedades Transmisibles y Enfermedades no Transmisibles no están estandarizadas, por lo que se descargo la base de datos de la población proporcionada por el World Bank Data, agregamos la variable y calculamos la taza por cada 100,000 habitantes. También encontramos la siguiente cantidad de datos faltantes por variable:

```
life_expectancy_world_bank          150
prevelance_of_undernourishment      635
co2                                 103
health_expenditure_                 131
education_expenditure_             1042
unemployment                        255
corruption                         2282
sanitation                         1198
```

Donde las variables que no aparecen no tienen datos faltantes. Dada la cantidad de datos faltantes, es necesario hacer limpieza de datos. Notemos que el 70% de los datos en corrupción son datos faltantes, en consecuencia excluimos la variable en este análisis. También se eliminaron los países que tienen un porcentaje mayor al 40% de datos faltantes, con el motivo de evitar sesgos en los estimadores y mantener la calidad de la inferencia

III. Imputación de la base de datos

Dado que existen variables de interés para el análisis con datos faltantes, vamos a aplicar técnicas de imputación que preserven la integridad del conjunto de datos. Vamos a describir algunas técnicas consideradas y una justificación de la elección:

- **Imputación simple:** Una manera sencilla de imputar el valor faltante es por la media, mediana o moda de la columna. Sin embargo, subestima la varianza y destruye las correlaciones entre variables, por lo que este método se descartó.
- **Imputación Multivariada:** El método K-NN busca los k vecinos más cercanos y utiliza sus valores para estimar el dato faltante. A priori es una buena manera de imputar los datos, ya que un país esta influenciado por su nivel económico y localización geográfica como lo demuestra Acemoglu y Johnson. Sin embargo, predecir correctamente la relación que hay entre cada variable no es trivial y no maneja la incertidumbre de los parámetros del modelo.
- **Modelos bayesianos:** De manera similar a la regresión, modelamos cada característica con valores faltantes como una función de otras características. Sin embargo, los parámetros se consideran variables aleatorias y se incorpora el conocimiento previo (prior) de los datos observados para obtener una distribución posterior que cuantifica la incertidumbre.

Recordemos que por el origen de nuestros datos, la variables de interés tienen relaciones lineales y con multicolineales (e.j., PIB y Gasto en Salud). Por lo tanto, el modelo bayesiano que elegimos es Ridge, en particular se implementa el algoritmo MICE.

A. MICE: Bayesian Ridge

Como se detalla en [4]Bishop, 2006. Bayesian Ridge supone que la variable objetivo y se genera a partir de una función lineal con ruido Normal. Dado un vector de características X y un vector de pesos w:

$$y = Xw + \epsilon,$$

donde el ruido ϵ sigue una distribución normal $\epsilon \sim \mathcal{N}(0, \alpha^{-1})$, siendo α la precisión del ruido (el inverso de la varianza del error σ^2). Por lo tanto, la verosimilitud de observar y dados X y w es:

$$p(y|X, w, \alpha) = \mathcal{N}(y|Xw, \alpha^{-1}).$$

Para manejar la multicolinealidad, se introduce una regularización mediante un *prior* sobre los pesos w, asumiendo que siguen una distribución gaussiana esférica:

$$p(w|\lambda) = \mathcal{N}(w|0, \lambda^{-1} I_p),$$

donde λ es la precisión de los pesos. Si λ es alto, la varianza es baja, forzando a los pesos a acercarse a cero (alta regularización); si λ es bajo, se permite que los pesos crezcan. A diferencia de Ridge convencional, Bayesian Ridge estima los hiperparámetros α y λ automáticamente a partir de los datos, asumiendo que siguen distribuciones Gamma (priors conjugados):

$$p(\alpha) \sim \text{Gamma}(\alpha_1, \alpha_2) \quad \text{y} \quad p(\lambda) \sim \text{Gamma}(\lambda_1, \lambda_2).$$

Estos hiperparámetros se inicializan con valores pequeños para ser no informativos.

Para obtener el posterior de los pesos w, utilizamos la regla de Bayes:

$$p(w|y, X, \alpha, \lambda) = \frac{p(y|X, w, \alpha)p(w|\lambda)}{p(y|X, \alpha, \lambda)}.$$

La distribución resultante es Gaussiana: $p(w|y, X, \alpha, \lambda) = \mathcal{N}(w|\mu_w, \Sigma_w)$, donde la media μ_w y la matriz de covarianza Σ_w se calculan como:

$$\mu_w = \alpha \Sigma_w X^T y \quad \text{y} \quad \Sigma_w = (\alpha X^T X + \lambda I)^{-1}.$$

Dado que no conocemos los valores reales de α y λ, el algoritmo itera para maximizar la evidencia (verosimilitud marginal). El procedimiento se resume a continuación:

1) **Inicialización:** Se establecen valores iniciales para α y λ.
2) **Paso E (Estimación):** Se calculan μ_w y Σ_w usando los valores actuales de α y λ.

3) **Paso M (Maximización):** Se actualizan α y λ maximizando la verosimilitud marginal, basándose en las nuevas estadísticas de μ_w y la varianza residual.
4) **Convergencia:** Se repiten los pasos 2 y 3 hasta que α y λ converjan.

Gracias al uso de priors, el modelo penaliza pesos excesivamente grandes, lo que lo vuelve robusto ante la multicolinealidad presente entre los indicadores de desarrollo. Esto evita que la imputación se vea sesgada por correlaciones espurias o ruido excesivo en las variables predictoras.

IV. Validación del modelo

Para validar el modelo BayesRidge vamos a gráficar la distribución de los datos originales y los datos imputados, buscando que describa correctamente la distribución original. Consideremos las siguientes distribuciones:

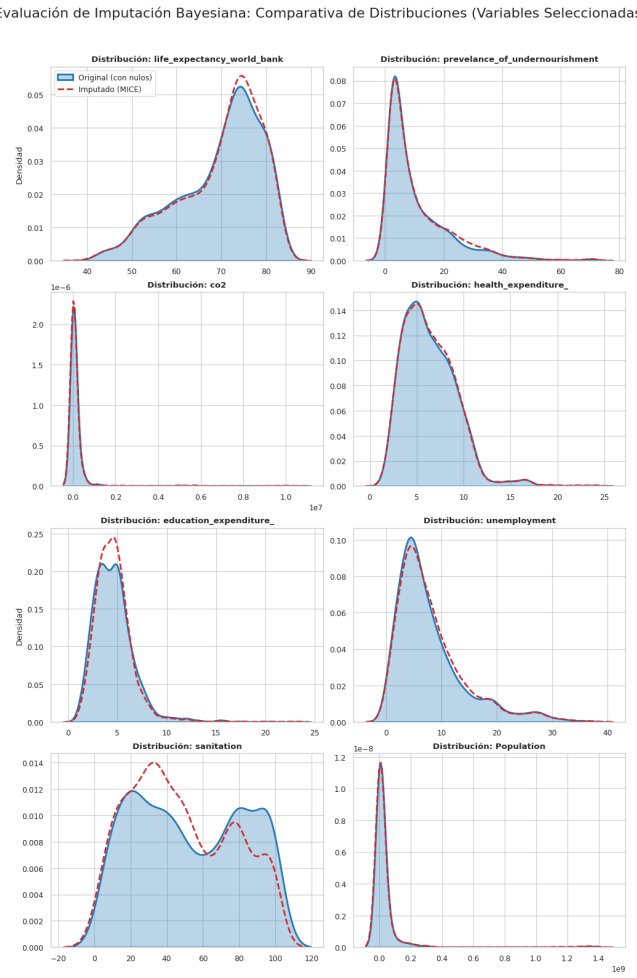

Fig. 2. Distribuciones de los parámetros de interés que tenían al menos un dato faltante en el periodo 2001-2019

Vemos como la curva imputada describe muy bien la silueta de la original. Por lo que en términos generales, la imputación ha sido muy buena para 5 de las 7 variables, detectando correctamente las correlaciones lineales con otras variables. Sin embargo, hay dos observaciones:

1) Pico en la linea imputada en la variable gasto en educación. Una posible explicación es que la variable tiene 32% de datos faltantes, por lo que el modelo enfatizo los valores centrales bajo la incertidumbre. Sin embargo, la forma general se mantiene y no hay distorsión significativa.
2) Pico en valores bajos de la variable de saneamiento. Una interpretación del fenómeno es que la variable tiene 37.7% de datos faltantes. Por lo que el modelo descubrió que los datos faltantes no eran aleatorios, sino que detecto que los países con huecos en saneamiento se parecen mucho a los países con bajos ingresos. Así, los países que no reportan saneamiento, probablemente tienen mal saneamiento.

Por lo tanto, creemos que el modelo que se eligió para imputar los datos es correcto. Ahora, nos gustaría ver cuales variables son determinantes para la esperanza de vida y si es necesario eliminar variables que no afecten al estudio.

V. Determinación de factores

Para ver la relación que existe entre las variables consideradas, veremos la siguiente matriz de correlación:

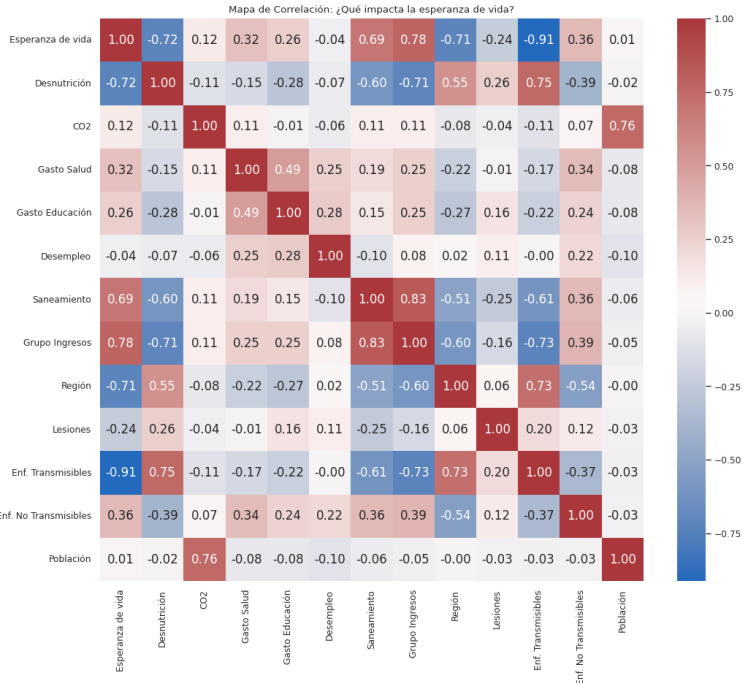

Fig. 3. Matriz de correlación entre las variables a nivel global en el periodo 2001-2019

El hallazgo más significativo es la correlación negativa entre la esperanza de vida y la tasa de enfermedades transmisibles. Esto sugiere que la erradicación de patologías infecciosas es el predictor individual más fuerte de la longevidad a nivel mundial.

También observamos un fuerte agrupamiento de variables socioeconómicas, tales como grupo de ingresos y acceso a saneamiento muestran correlaciones positivas fuertes con la

esperanza de vida. Lo cual confirma que la infraestructura básica y la capacidad económica no son factores aislados, sino condiciones para romper el ciclo de mortalidad por causas previsibles. Palomino

VI. Esperanza de vida en México

Una vez que hemos identificado los factores determinantes en la esperanza de vida a nivel mundial, podemos contrastar los hallazgos con la evolución que ha tenido México en estos años. Palomino La exploración de las series de tiempo de

Fig. 4. Variables determinantes en México evaluadas en los años 2001 y 2019.

México (Fig. 4) revela una contradicción. Mientras que los indicadores de desarrollo de Saneamiento y Prevalencia de Desnutrición muestran una mejora sistémica, la esperanza de vida sufrió un estancamiento e incluso un retroceso entre 2005 y 2015.

Para demostrar que este comportamiento es atípico en el contexto global, utilizamos el algoritmo K-NN en el periodo 2005-2015 con el objetivo de identificar un conjunto de países con condiciones socioeconómicas e infraestructurales idénticas a las de México. Considerando las variables Saneamiento, Gasto en salud, Prevalencia de Desnutrición y Enfermedades Transmitirles, en el año 2005 K-NN determina que los vecinos de México son: Suriname, Costa Rica, Morocco, Colombia y China.

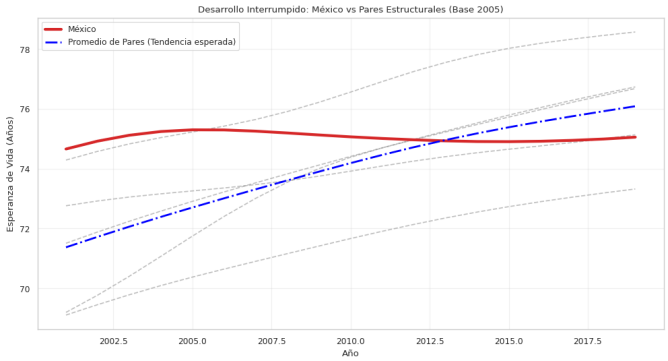

Fig. 5. Esperanza de vida en México y el promedio de la esperanza de sus vecinos del año 2005

Notamos que México está por encima respecto al resto de países vecinos, teniendo casi 2.5 años más de Esperanza de vida que el promedio en el periodo 2000 a 2005. Sin embargo, la gráfica revela tres fases críticas que confirman la hipótesis del desarrollo interrumpido:

- **Fase de Crecimiento (2000-2005):** México muestra una tendencia positiva, coherente con las mejoras en infraestructura sanitaria.
- **El Punto de Inflexión (2006-2010):** Coincidiendo con el inicio de la crisis de seguridad nacional, la esperanza de vida en México sufre una desaceleración súbita, volviéndose asintótica, mientras que sus vecinos estructurales mantienen una pendiente de crecimiento positiva constante.
- **Punto de inflexión (2013-2019):** Ocurre un fenómeno de inversión, el promedio de los países vecinos supera a México. Para 2019, la brecha se ha invertido; México no solo perdió su ventaja inicial, sino que finalizó el periodo por debajo de la tendencia esperada.

VII. Conclusión

Este análisis sirve como una evidencia preliminar de un desarrollo interrumpido. Sin embargo, es necesario de realizar estudios especializados sobre el país para confirmar inequívocamente el peso de otros factores como la violencia sobre la Esperanza de vida.

México presenta un caso de estudio para la ciencia de datos aplicada a la política. Demuestra que el crecimiento económico y la infraestructura, por sí solos, son insuficientes para garantizar el bienestar de la población.

References

[1] The World Bank, "World Bank Open Data," 2024. [Online]. Available: https://data.worldbank.org/.

[2] M. Roser, H. Ritchie, E. Ortiz-Ospina, and L. Rodés-Guirao, "Our World in Data," 2024. [Online]. Available: https://ourworldindata.org/.

[3] D. Acemoglu and S. Johnson, "Disease and development: The effect of life expectancy on economic growth," *Journal of Political Economy*, vol. 115, no. 6, pp. 925–985, Dec. 2007.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. New York: Springer, 2006.