Modelo de Diferencias en Diferencias aplicado en la pandemia por COVID-19

Cristian Alejandro Cedeño Cabrera
Ingeniero Industrial
Universidad de Los Andes, Bogotá

Robinson Galvis Marin
Ingeniero de Sistemas
Institución Universitaria ITM, Medellín

Abstract

Este proyecto tiene como objetivo explicar el modelo de Diferencias en Diferencias (**DiD**) en un análisis cuantitativo. Este modelo es utilizado para evaluar el impacto de un tratamiento o intervención sobre un grupo en comparación con un grupo de control en series temporales de datos. El objetivo es comprender el modelo *DiD* utilizando el software **R**, aplicándolo a un análisis con datos reales, el cual permite simular un diseño experimental con datos observacionales, útil cuando no se pueden realizar experimentos controlados. Esta metodología busca evaluar el impacto de un tratamiento sobre dos grupos a través del tiempo, antes y después de la intervención, es decir, un grupo tratado vs uno no tratado, en este caso, el impacto sobre el ratio de fallecidos a partir de la vacunación en época de pandemia por el *COVID-19*.

1 Problema

El modelo de Diferencias en Diferencias (**DiD**) se utiliza para evaluar el efecto causal de una intervención o tratamiento, comparando dos grupos (tratado y control) antes y después de la intervención. En este enfoque, se mide cómo cambia cada grupo a lo largo del tiempo y luego se comparan esos cambios entre los dos grupos. El análisis se basa en datos de panel, que contienen observaciones repetidas a lo largo del tiempo para cada unidad, lo que permite controlar la variabilidad entre las unidades y a lo largo del tiempo.

Este modelo se considera una herramienta de regresión causal, ya que busca estimar el efecto del tratamiento (en este caso, la aceleración de la vacunación) sobre una variable dependiente (el número de nuevas muertes diarias). La principal ventaja del *DiD* es su capacidad para mitigar los sesgos de selección y otros efectos confusos, controlando factores no observables que podrían afectar tanto al grupo tratado como al grupo de control, siempre y cuando se cumpla la suposición de tendencias paralelas: es decir, que ambos grupos habrían seguido trayectorias similares en ausencia de la intervención.

Para este análisis, se utilizará un conjunto de datos disponible en **Kaggle** que contiene información sobre la relación entre el porcentaje de población vacunada y el número de nuevas muertes diarias por país durante la pandemia de *COVID-19*. El objetivo es examinar si la aceleración de la vacunación tuvo un impacto significativo en la reducción de las muertes diarias, utilizando el modelo *DiD* para comparar los cambios en dos países, *Ecuador* que implementó políticas de vacunación más rápidas (grupo tratado) con *Georgia* que tuvo un ritmo de vacunación más lento (grupo de control). Este enfoque permitirá evaluar de manera más robusta el efecto causal de la vacunación sobre las muertes diarias.

2 Objetivos

Objetivo general

Instruir a los espectadores sobre el uso del modelo de Diferencias en Diferencias (DiD) para realizar un análisis causal con datos observacionales, centrado en el impacto de la aceleración de la vacunación en la reducción de muertes diarias por COVID-19.

Objetivos específicos

- Explicar el funcionamiento y la interpretación de los coeficientes en el modelo *DiD*, usando el contexto de la vacunación como intervención.
- Mostrar cómo ejecutar una regresión *DiD* en **R** utilizando el conjunto de datos sobre la relación entre vacunación y muertes diarias por país.
- Presentar ejemplos prácticos que ilustren los beneficios del modelo *DiD* en el análisis causal, mediante estadísticas descriptivas y gráficos que comparen los cambios en las muertes diarias entre países con diferentes ritmos de vacunación (grupo tratado y grupo control).

Los espectadores aprenderán

- El concepto de tendencias paralelas y cómo esta suposición es clave para la validez del modelo *DiD*, particularmente en el contexto de la evaluación del impacto de la vacunación.
- Cómo interpretar los resultados de la regresión *DiD*, centrándose en el coeficiente que refleja el efecto causal de la aceleración de la vacunación sobre las muertes diarias.
- Cómo ejecutar un análisis de Diferencias en Diferencias en **R** utilizando funciones como *lm()* y librerías como *plm()*, adaptadas a datos reales de salud pública.

3 Estado del arte

Uno de los artículos más representativos y con mayor influencia en el modelo de Diferencias en Diferencias (*DiD*) es el trabajo de **Card** y **Krueger** sobre el salario mínimo en los *EEUU*, que utiliza este modelo para estimar el efecto de un aumento en el salario en el empleo en la industria de comida rápida en Nueva Jersey. Este estudio es un referente clave que demuestra cómo *DiD* puede controlar factores comunes en los estados de New Jersey y Pensilvania, y cómo el aumento del salario mínimo no tuvo el efecto negativo sobre el empleo que la teoría económica tradicional sugeriría, desafiando así la hipótesis clásica sobre el empleo y el salario mínimo.

Por otro lado, **Angrist** y **Pischke** en su libro *Mostly Harmless Econometrics (2009)* presentan de manera accesible los fundamentos de las técnicas de econometría aplicada, entre ellas el modelo *DiD*. En este texto, los autores proporcionan una explicación detallada de cómo implementar modelos de estimación causal con datos observacionales, subrayando la importancia de los supuestos subyacentes como las tendencias paralelas, que son clave para la validez de la estimación en *DiD*. Este enfoque resulta fundamental para aquellos que deseen aplicar el modelo en la práctica, ya que ofrece ejemplos concretos y una sólida fundamentación teórica.

Por último, **Athey** e **Imbens** (2017) abordan la estimación causal en el contexto de experimentos naturales y de control observacional, profundizando en las técnicas de *DiD* y otros enfoques como la regresión discontinua y el *matching*. Los autores se centran en la identificación y corrección de sesgos potenciales en los estudios empíricos, destacando la importancia de los métodos robustos que pueden mejorar la precisión y validez de los resultados. Su trabajo es clave para aquellos que buscan perfeccionar las estimaciones causales mediante el control de factores no observables y otras fuentes de sesgo.

4 Metodología

4.1 Dataset

Dataset: Para explicar el modelo de diferencias en diferencias, se utilizará un conjunto de datos disponible en *Kaggle*, que muestra la relación entre el porcentaje de población vacunada y el número de nuevas muertes diarias por país. A través de este conjunto de datos, se analizará si la aceleración de la vacunación tuvo un impacto en la reducción de las muertes diarias.

Campos del dataset:

- country: Nombre del país (String)
- iso-code: Código ISO para cada país (String)
- date: Fecha del registro (date)
- total-vaccinations: Número de todas las dosis de COVID usadas en ese país (int)
- people-vaccinated: Personas con al menos una dosis de la vacuna (int)
- people-fully-vaccinated: Personas con todas las dosis de la vacuna (int)
- New-deaths: Nuevas muertes por COVID (int)
- population: Población total 2021 (bigint)
- ratio: people-vaccinated/population *100 (double)

Nuevas variables, filtros:

- percentdeath: Nuevas muertes diarias sobre la población (double)
- Se usarán los países de Ecuador y Georgia

La comparación se realizará utilizando los datos de los países de Ecuador y Georgia, correspondientes al período de febrero de 2021 a marzo de 2022.

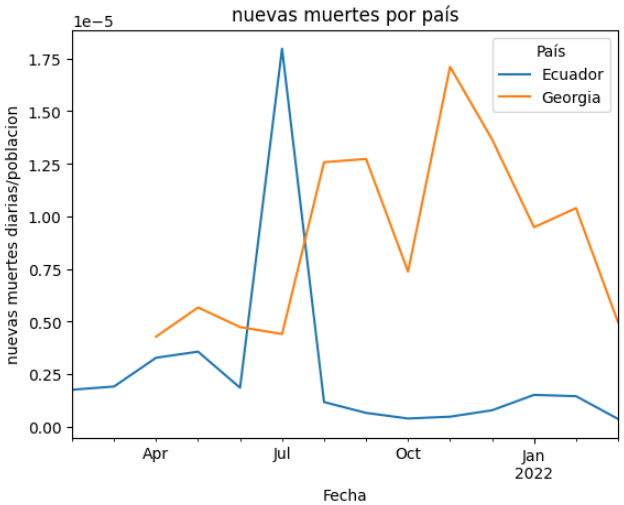

Figure 1: Nuevas muertes diarias por país

La Figura 1 muestra el porcentaje de nuevas muertes diarias por mes sobre la población total en los países de Ecuador y Georgia, se identifica que a partir del mes de agosto del 2021, Ecuador empieza a tener un porcentaje de muerte diarias controlada y cercano al 0%

Para complementar la gráfica anterior, se evidencia en la figura 2 que en agosto del 2021, Ecuador llegó al 60% de la población vacunada, se quiere identificar si la rápida vacunación de la población ayudó a controlar y disminuir la muertes por COVID.

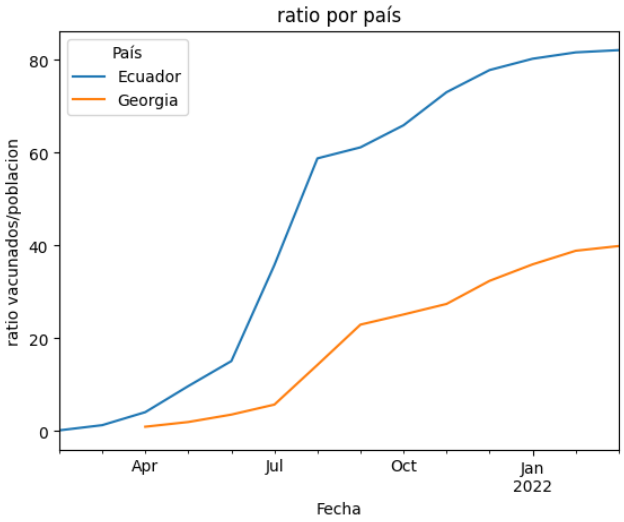

Figure 2: Porcentaje de vacunados

4.2 Modelo

El modelo básico de *DiD* puede implementarse en **R** de la siguiente manera:

did_model <- lm(Y Post * Treated + factor(Time) + factor(ID), data = dataset)

4.3 Comentarios adicionales

Evaluación del modelo

- Se compararán los resultados de diferentes especificaciones del modelo para verificar la robustez de las estimaciones.
- Se puede realizar el **Test de Hausman** para comprobar la validez de los efectos fijos frente a los efectos aleatorios en el modelo.

Posibles sesgos y supuestos

- El modelo *DiD* está sujeto a la suposición de tendencias paralelas. Si no se cumple, los resultados pueden estar sesgados. Se deben realizar pruebas para verificar este supuesto y considerar técnicas adicionales como **matching** para mejorar la validez del modelo.

5 Fuente

[1] Card, D., Krueger, A. B. (1994). Minimum wages and employment: A case study of the fast food industry in New Jersey and Pennsylvania. *American Economic Review*, 84(4), 772-793.

[2] Angrist, J.D., Pischke, J.S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press.

[3] Athey, S., Imbens, G.W. (2017). The econometrics of randomized experiments. *Annual Review of Economics*, 9, 1-29.

[4] Karaji, S. (2021). COVID vaccination vs death dataset. Kaggle. `https://www.kaggle.com/datasets/sinakaraji/covid-vaccination-vs-death/data`

[5] Wikipedia contributors. (n.d.). Difference in differences. Wikipedia, The Free Encyclopedia. Retrieved November 10, 2024, from `https://en.wikipedia.org/wiki/Difference_in_differences`