Maestría en Econometría - UTDT
Examen Final - Microeconometría II

En años recientes, ha habido un resurgimiento de la literatura de evaluación de impacto utilizando el estimador de diferencia en diferencias cuando hay múltiples períodos temporales, antes y después de la implementación de un programa (o política), y el impacto de dicho programa es heterogéneo. En este contexto de múltiples períodos temporales, se va a estudiar la estimación e inferencia estadística de varios de los estimadores propuestos por la literatura cuando la muestra es "pequeña".

Para resolver el examen, primero, leer el trabajo "How Much Should We Trust Modern Difference-in-Differences Estimates?" de Amanda Weiss y el trabajo "Revisiting Event Study Designs: Robust and Efficient Estimation" de Borusyak, Jaravel y Spiess. Los dos trabajos están colgados en el campus virtual.

El trabajo de Weiss identifica siete estimadores:

- *Estándar two-way fixed effects ("TWFE").*
- *Imai, Kim and Wang Matching Estimator (2023, "IKW").*
- *Callaway and Sant'Anna Aggregated Group-Time Estimator (2021, "CS").*
- *De Chaisemartin and d'Haultfoeuille DIDM Estimator (2020, "DCDH").*
- *Borusyak, Jaravel and Spiess Efficient Estimator (2024, "BJS").*
- *Wooldridge Two-Way Mundlak Regression Estimator (2021, "JW").*
- *Sun and Abraham Dynamic Treatment Effects Estimator (2021, "SA").*

Se van a considerar cuatro de estos estimadores: TWFE, CS, BJS y JW. La sección 4.2 del trabajo de Weiss describe la generación de los datos. Repetir esta generación utilizando como semilla (seed) los últimos 5 números del pasaporte o documento de identidad para que los resultados sean replicables. Obviamente, que los números no van a ser, exactamente, los de las figuras/tablas de Weiss, pero se quiere evaluar si las implicancias de los resultados de este examen coinciden con lo encontrado por Weiss.

***Data-Generating Processes* y Aclaraciones Preliminares**

El archivo *"MenduinaJuan02205.do"* realiza simulaciones de estimaciones de impacto de tratamiento en modelos de diferencias en diferencias (*Difference-in-Difference*) en presencia de efectos homogéneos y heterogéneos. Para mayor claridad, se presenta un resumen cualitativo de este código:

1. CONFIGURACIÓN:

- Define el directorio en el cual se trabajará, instala (de ser necesario) los comandos necesarios y establece una semilla para replicabilidad (*seed*= 02205).
- Define varias variables globales y locales, que determinan, entre otras cosas, el número de simulaciones (*reps*= 100), los modelos a probar (*models*= hom, het), los métodos de estimación (*estims*= TWFE, CS, BJS, JW), los tamaños de muestra (*samples*= 50, 500), períodos (*periods*= 2, 4, … , 30), los valores del verdadero efecto del tratamiento (*ATTs*= 0.2, 0.5, 0.8), etc.

2. PROGRAMA:

- Define un programa llamado *simulation*, que, en primer lugar, genera un panel de datos de dimensión NxT, con efectos fijos por individuo (*alpha*) y efectos fijos por período (*gamma*), junto con un término de error aleatorio (*epsilon*).
- Define la variable *Treated* igual a 1 para el 50% de los individuos con efectos fijos más grandes (id $>$ 25 si N= 50 y id $>$ 250 si N= 500) e igual a 0 en caso contrario.
- Genera las variables de cohorte (*g_hom* y *g_het*) para dos tipos de modelos:
  - Modelo homogéneo: Período aleatorio que varía en cada simulación (t= 2, 3, … , T).
  - Modelo heterogéneo: Período no aleatorio que depende del efecto fijo por individuo. Se considera el cociente $\frac{\frac{N}{2}}{T-1}$ y se asigna la misma cantidad de individuos tratados a cada período (t= 2, 3, … , T); por ejemplo, para N= 50, si T= 6, los id 26 a 30 se tratarán a partir de t= 2, los id 31 a 35 se tratarán a partir de t= 3 y así sucesivamente. Cuando $\frac{\frac{N}{2}}{T-1}$ no es un número entero (lo cual sucede en la mayoría de los casos), entonces, se redondea y es el último período el que va a tener menos individuos tratados; por ejemplo, para N= 50, si T= 3, los id 26 a 38 (13 en total) se tratarán a partir de t= 2 y los id 39 a 50 (12 en total) se tratarán en t= 3.
- Define las variables *Post* asociadas a cada modelo (*Post_hom* y *Post_het*), en función de la variable de cohorte correspondiente (igual 1 si t $>$ *g_hom* e igual 1 si t $>$ *g_het*, respectivamente).
- Define las variables *D_hom* y *D_het* como el producto de *Treated***Post_hom* y *Treated***Post_het*, respectivamente.
- Genera las variables dependientes (*y_hom* e *y_het*) para los dos tipos de modelos:
  - Modelo homogéneo: $Y_{i,t} = \beta D_{i,t} + \alpha_i + \gamma_{i,t} + \epsilon_{i,t}$.
  - Modelo heterogéneo: $Y_{i,t} = \tau_i D_{i,t} + \alpha_i + \gamma_{i,t} + \epsilon_{i,t}$.

- Para cada modelo (homogéneo y heterogéneo) y cada método de estimación (TWFE, CS, BJS y JW), el programa estima el *Average Treatment Effect on the Treated* (ATT) y retorna este valor y el error estándar asociado.

3. SIMULACIONES:

- Realiza simulaciones para diferentes combinaciones de tamaño de muestra (N= 50, 500), cantidad de períodos (T= 2, 4, … , 30) y ATT (0.2, 0.5, 0.8).
- Cada combinación de N, T y ATT se ejecuta *reps* (100) veces para calcular: *ATT Bias* (figura 3), *SE Bias* (figura 5), *Coverage Probability* (tabla 3) y *Statistical Power* (figura 11).
- Guarda los resultados de estos cálculos en matrices de resultados (*results_N`N'_`result'*), una para cada combinación de N y *result* (figure3, figure5, table3 y figure11), es decir, en total, 8 matrices de resultados.
- Convierte estas matrices de resultados en bases de datos para análisis posterior.

4. EJERCICIOS:

- Construye, en función de estas bases de datos con las matrices de resultados, las figuras 3, 5 y 11 y la tabla 3 solicitadas.

Es importante aclarar que, dados los tiempos de ejecución de este algoritmo (con todo lo que implica, es decir, cantidad de combinaciones de N, T y ATT y cantidad de simulaciones), se procedió a realizar 100 simulaciones, en lugar de las 5.000 simulaciones que se realizan en el *paper*. En particular, así como se encuentra el código actualmente y corriendo para N= 50 en un *core* y para N= 500 en otro *core*, el tiempo de ejecución fue, aproximadamente, 7 días (es decir, *céteris páribus*, 5.000 simulaciones se hubieran ejecutado en 350 días). En línea con esto, debido a la menor cantidad de simulaciones, es posible que los resultados no sean tan robustos como los de Amanda Weiss. En particular, todos los resultados asociados al modelo heterogéneo no son satisfactorios, no sólo no coinciden con los del *paper*, sino que son muy distintos. Debido a que todo el código replica *vis-a-vis* lo expresado por la autora en cuanto a la generación de los datos y a las estimaciones, eventualmente, esta discrepancia de resultados se puede deber a la combinación de dos cosas: la menor cantidad de simulaciones y/o una definición incorrecta de la variable de cohorte (*g_het*), que se asumió de la manera descripta anteriormente pero que el *paper* no aclara cómo se construye (en cambio, si bien la variable τ_i tampoco se aclara cómo se construye, se le consultó a Weiss y expresó que se construye como se puede ver en las líneas 91 a 94 del código).

Por último, en el siguiente link de Google Drive, se encuentran las figuras realizadas (en .png) y los siguientes archivos .dta:

- *"results_N50_base.dta"* y *"results_N500_base.dta"*:
  bases de datos con los *outputs* de cada una de las 100 simulaciones para tamaño de muestra de N= 50 y N= 500, respectivamente.
- *"results_N50_matriz_figure3.dta"* y *"results_N500_matriz_figure3.dta"*:
  base de datos con la matriz de resultados asociada a la figura 3 para tamaño de muestra de N= 50 y N= 500, respectivamente.
- *"results_N50_matriz_figure5.dta"* y *"results_N500_matriz_figure5.dta"*:
  base de datos con la matriz de resultados asociada a la figura 5 para tamaño de muestra de N= 50 y N= 500, respectivamente.

- *"results_N50_matriz_table3.dta"* y *"results_N500_matriz_table3.dta"*: base de datos con la matriz de resultados asociada a la tabla 3 para tamaño de muestra de N= 50 y N= 500, respectivamente.
- *"results_N50_matriz_figure11.dta"* y *"results_N500_matriz_figure11.dta"*: base de datos con la matriz de resultados asociada a la figura 11 para tamaño de muestra de N= 50 y N= 500, respectivamente.

Ejercicio 1.

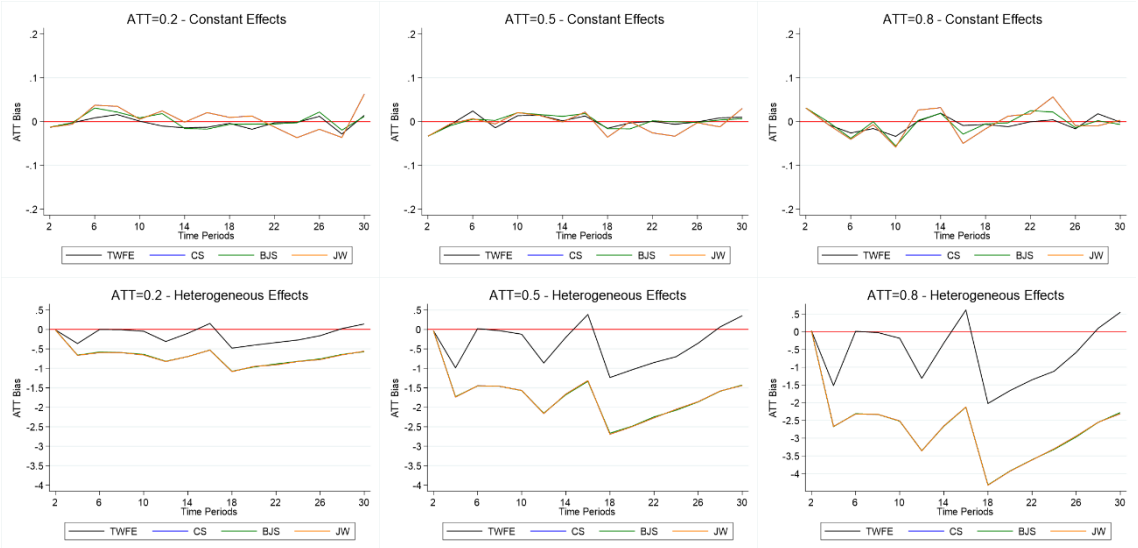
Reproducir la Figura 3 del trabajo de Weiss para los cuatro estimadores. ¿Se encuentran los mismos resultados? En particular, encontrar el sesgo por ponderadores negativos (De Chaisemartin and d'Haultfoeuille, 2021) para el estimador de TWFE cuando el efecto es heterogéneo. Comentar los resultados.

En la figura 3, se presenta el *ATT Bias* de los diferentes estimadores (TWFE, CS, BJS, JW) para cada combinación de ATT (0.2, 0.5, 0.8) y modelo (homogéneo y heterogéneo), a lo largo de los diferentes períodos utilizados (T= 2, 4, … , 30), considerando un tamaño de muestra de N= 50.

Por un lado, se puede observar que, para los casos con efectos de tratamiento constantes, todos los estimadores parecen insesgados (análogo a como sucede en el *paper*).

Por otro lado, como se mencionó en la introducción de este trabajo, se observa que, para los casos con efectos de tratamiento heterogéneos, los resultados no son satisfactorios y, en particular, que todos los estimadores tienen un importante sesgo (esto sólo sucede con TWFE en el *paper*).

Figure 3. *Bias for Four Difference-in-Difference Estimators, under Different Data-Generating Processes and ATTs (N=50).*



Fuente: Elaboración propia.

Ejercicio 2.

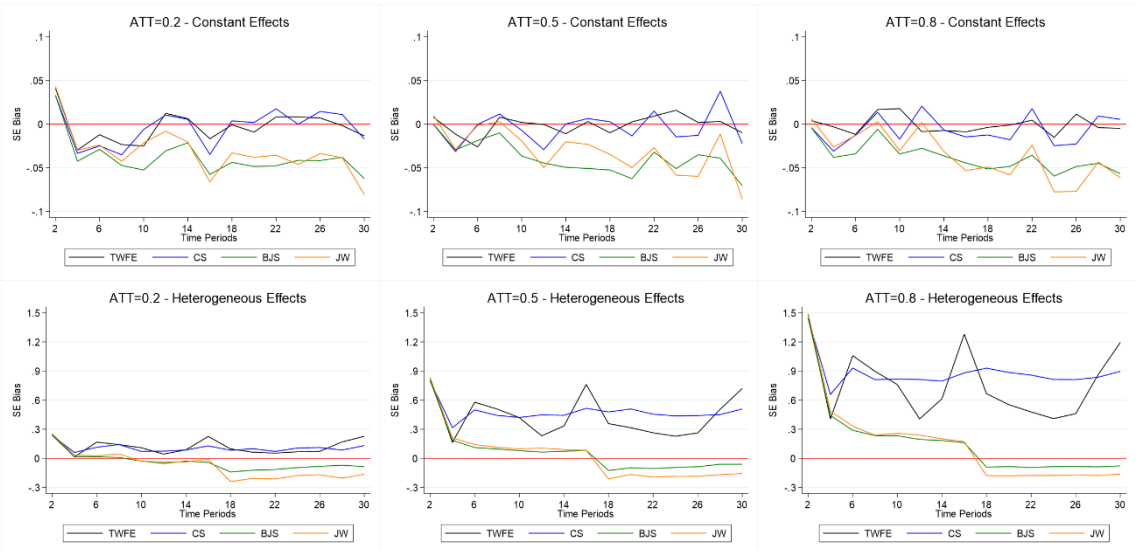
Para evaluar la validez de la inferencia estadística, reproducir la Figura 5 y la Tabla 3 del trabajo de Weiss. ¿Se encuentran los mismos resultados? Comentar los resultados.

En la figura 5, se presenta el *SE Bias* de los diferentes estimadores (TWFE, CS, BJS, JW) para cada combinación de ATT (0.2, 0.5, 0.8) y modelo (homogéneo y heterogéneo), a lo largo de los diferentes períodos utilizados (T= 2, 4, … , 30), considerando un tamaño de muestra de N= 50.

Por un lado, se puede observar que, para los casos con efectos de tratamiento constantes, los errores estándar de los estimadores TWFE y CS parecen insesgados (CS parece diferir respecto al *paper*, donde parece cada vez más sesgado a medida que aumenta T) y los errores estándar de los estimadores BJS y JW parecen sesgados, con un sesgo en aumento a medida que aumenta T (análogo a como sucede en el *paper*).

Por otro lado, a diferencia de lo que sucede en la figura 3, se observa que, para los casos con efectos de tratamiento heterogéneos, los resultados no son tan insatisfactorios y, en particular, que los errores estándar de todos los estimadores tienen un importante sesgo (mayor cuando el ATT es mayor), aunque éste disminuye a medida que aumenta T para los estimadores BJS y JW (análogo a como sucede en el *paper*), no así para los estimadores TWFE y CS (esto no sucede en el *paper*).

Figure 5. *Standard Error Bias for Four Difference-in-Difference Estimators, under Different Data-Generating Processes and ATTs (N=50).*



Fuente: Elaboración propia.

En la tabla 3, se presenta la *Coverage Probability* de los diferentes estimadores (TWFE, CS, BJS, JW) para ATT= 0,5 y ambos modelos (homogéneo y heterogéneo), para una selección de períodos (T= 2, 10, 20, 30), considerando un tamaño de muestra de N= 50. Se consideran intervalos de confianza de 95%.

Por un lado, se puede observar que, para los casos con efectos de tratamiento constantes, la cobertura suele disminuir a medida que aumenta T (análogo a como sucede en el *paper*).

Por otro lado, como se mencionó en la introducción de este trabajo, se observa que, para los casos con efectos de tratamiento heterogéneos, los resultados no son satisfactorios y, en particular, hay muchos valores nulos (esto no sucede en el *paper*).

Table 3. *Coverage Probabilities of DID-Style Methods, Given ATT=0.5 SDs and Constant or Heterogenous Effects - Selection of Time Periods T ∈ {2, 10, 20, 30} (N=50).*

	Constant Effects				**Heterogeneous Effects**			
	TWFE	**CS**	**BJS**	**JW**	**TWFE**	**CS**	**BJS**	**JW**
T= 2	0,95	0,93	0,93	0,95	1,00	1,00	1,00	1,00
T= 10	0,93	0,92	0,87	0,89	1,00	0,17	0,00	0,00
T= 20	0,95	0,97	0,76	0,92	0,14	0,00	0,00	0,00
T= 30	0,95	0,89	0,72	0,70	1,00	0,46	0,00	0,00

Fuente: Elaboración propia.

Ejercicio 3.

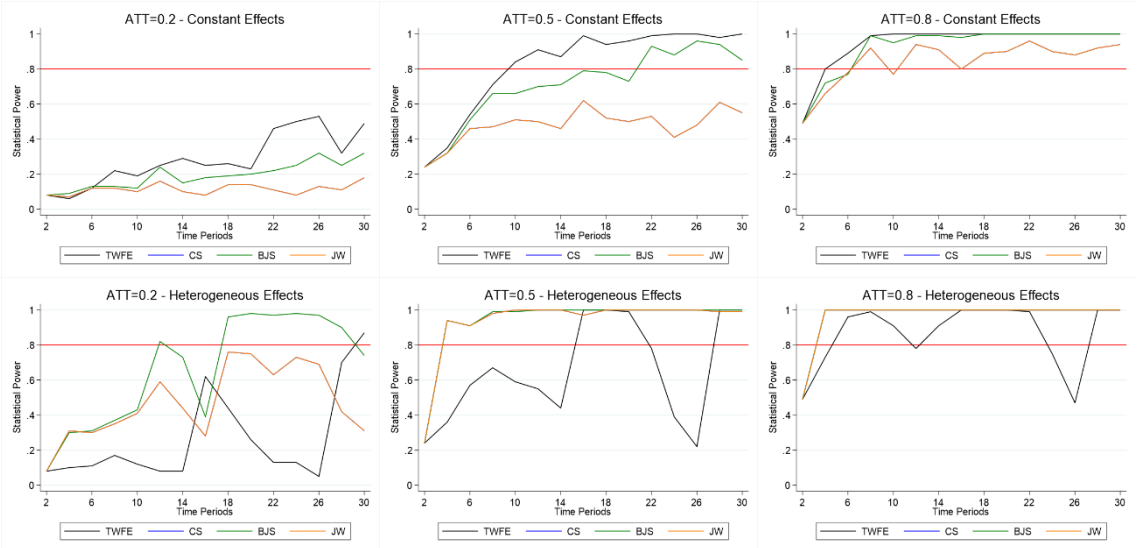
Potencia. Reproducir la Figura 11 de Weiss para los cuatro estimadores. ¿Se encuentran los mismos resultados? Comentar los resultados.

En la figura 11, se presenta el *Statistical Power* de los diferentes estimadores (TWFE, CS, BJS, JW) para cada combinación de ATT (0.2, 0.5, 0.8) y modelo (homogéneo y heterogéneo), a lo largo de los diferentes períodos utilizados (T= 2, 4, … , 30), considerando un tamaño de muestra de N= 50. Se considera la hipótesis nula (falsa) ATT= 0 y α= 0,05.

Por un lado, se puede observar que, para los casos con efectos de tratamiento constantes, la potencia aumenta a medida que aumenta T y también cuando el ATT es mayor (análogo a como sucede en el *paper*).

Por otro lado, como se mencionó en la introducción de este trabajo, se observa que, para los casos con efectos de tratamiento heterogéneos, los resultados no son satisfactorios y, en particular, hay mucha volatilidad en las series (esto no sucede en el *paper*).

Figure 11. *Statistical Power (with True Variance) for Four Difference-in-Difference Estimators, under Different Data-Generating P (N=50).*



Fuente: Elaboración propia.

Ejercicio 4 (Bonus).

Calcular los errores estándar vía block bootstrap y reproducir la Tabla 4 del trabajo de Weiss. Para el estimador CS, también reportar los errores estándar usando el procedimiento de WildBootstrap y comparar los resultados. La respuesta a este ítem de la pregunta no es obligatoria para aprobar el examen.

Ejercicio 5.

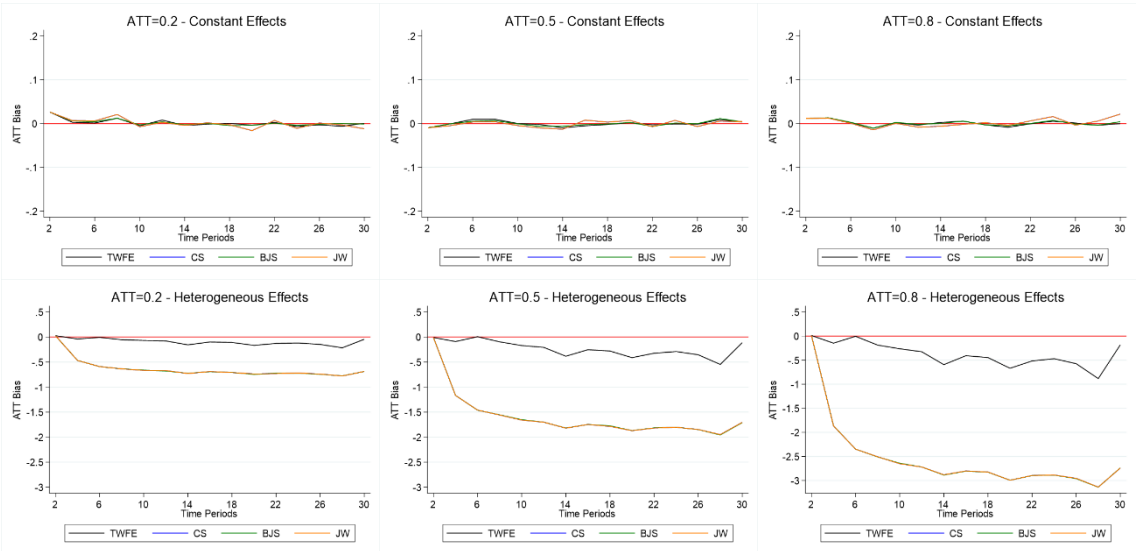
Repetir todos los ejercicios, pero, ahora, usando una muestra de corte transversal de 500 unidades en lugar de 50. ¿Cómo cambian los resultados? ¿Qué aprendizaje deja este ejercicio?

A continuación, se repiten los ejercicios (1 a 3), pero, ahora, usando una muestra de corte transversal de 500 unidades en lugar de 50. En términos generales, ante este aumento del tamaño de muestra, se puede observar que mejora (disminuye) el *ATT Bias* y el *SE Bias*, y mejora (aumenta) la *Coverage Probability* y el *Statistical Power*.

Como aprendizaje de este ejercicio, se tiene que, al aumentar el tamaño de muestra, mejora la estabilidad de las estimaciones y la precisión estadística, posibilitando menores sesgos (ver figuras 3' y 5'), mayor probabilidad de cobertura (ver tabla 3') y mayor potencia estadística (ver figura 11'), lo cual fortalece la interpretación y la confiabilidad de los hallazgos de las simulaciones. Cabe aclarar que, aquí, para los casos con efectos de tratamiento heterogéneos, los resultados tampoco son satisfactorios.

Ejercicio 1.

Figure 3'. *Bias for Four Difference-in-Difference Estimators, under Different Data-Generating Processes and ATTs (N=500).*

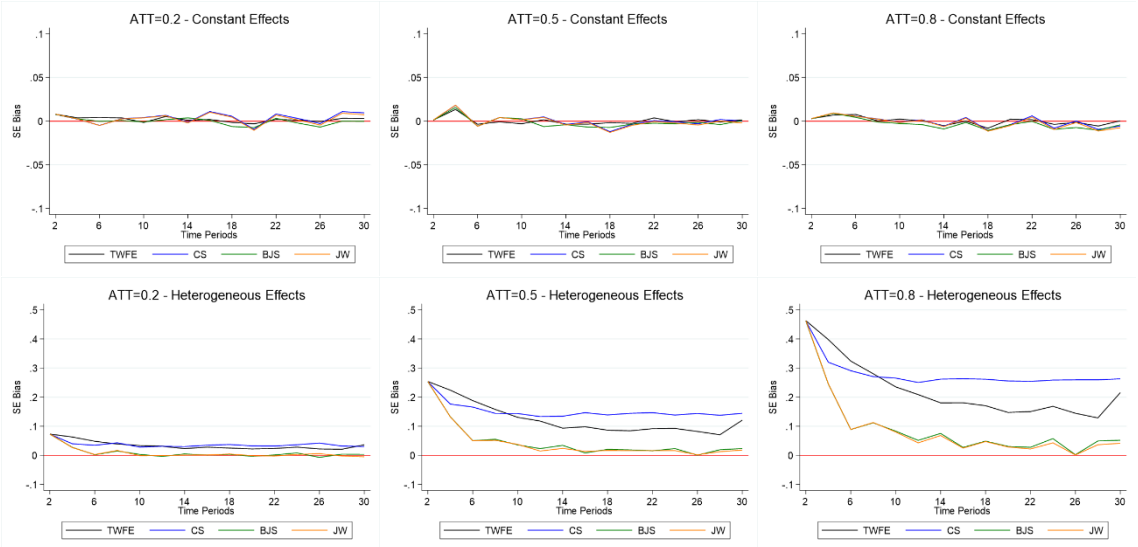


Fuente: Elaboración propia.

Se puede observar que, ahora, todos los estimadores parecen más insesgados (el sesgo se encuentra más cercano a 0).

Ejercicio 2.

Figure 5'. *Standard Error Bias for Four Difference-in-Difference Estimators, under Different Data-Generating Processes and ATTs (N=500).*



Fuente: Elaboración propia.

Se puede observar que, ahora, los errores estándar de todos los estimadores parecen insesgados.

Table 3'. *Coverage Probabilities of DID-Style Methods, Given ATT=0.5 SDs and Constant or Heterogenous Effects - Selection of Time Periods T ∈ {2, 10, 20, 30} (N=500).*

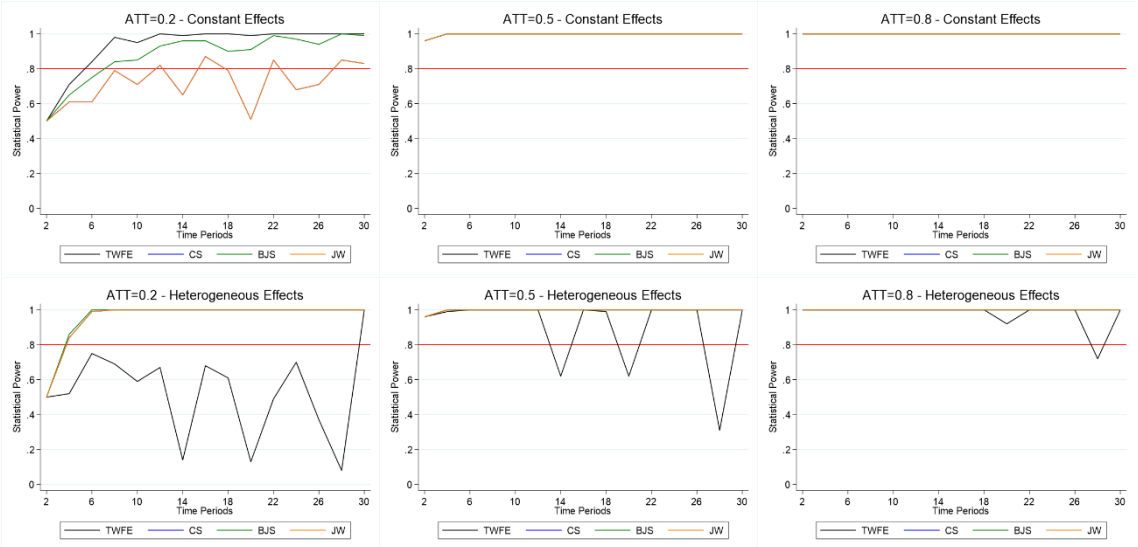
	Constant Effects				Heterogeneous Effects			
	TWFE	**CS**	**BJS**	**JW**	**TWFE**	**CS**	**BJS**	**JW**
T= 2	0,94	0,94	0,94	0,94	1,00	1,00	1,00	1,00
T= 10	0,93	0,94	0,95	0,94	1,00	0,00	0,00	0,00
T= 20	0,94	0,92	0,94	0,91	0,00	0,00	0,00	0,00
T= 30	0,95	0,97	0,95	0,94	1,00	0,00	0,00	0,00

Fuente: Elaboración propia.

Se puede observar que, ahora, la cobertura tiende al nivel de confianza utilizado (95%).

Ejercicio 3.

Figure 11'. *Statistical Power (with True Variance) for Four Difference-in-Difference Estimators, under Different Data-Generating P (N=500).*



Fuente: Elaboración propia.

Se puede observar que, ahora, la potencia es mucho mayor y es igual a 1 incluso para T pequeños.

Referencias.

Borusyak, K., Jaravel, X. y Spiess, J. (2024). Revisiting Event Study Designs: Robust and Efficient Estimation. *Review of Economic Studies*, 91, 3253-3285. https://doi.org/10.1093/restud/rdae007

Weiss, A. (2024). How Much Should We Trust Modern Difference-in-Difference Esimates? https://doi.org/10.31219/osf.io/bqmws