

Taller 1
Econometría Avanzada, 2022-2
Profesor: Manuel Fernández

Carlos A. Ayala - 201911488
Carlos Eduardo Torres - 202225155
Juan Jacobo Trujillo Pradilla - 201923274

Todos los elementos utilizados para la solución de este taller (bases de datos, dofile, outputs y demás), pueden ser consultados en el siguiente enlace:
<https://github.com/CarlosAyala123/Econometria-Avanzada-Taller-3>

Primer ejercicio:

El Instituto de Salud Sexual y Reproductiva (ISSR) realizó un experimento aleatorio para entender el efecto de dar información de planificación familiar en el comportamiento de fertilidad de las parejas. Para ello, se seleccionaron 50.000 hogares al azar, provenientes de 300 municipios rurales. Se seleccionaron 200 hogares por medio de una lotería justa para ser asignados al grupo de tratamiento, y los 100 hogares restantes fueron asignados al grupo de control. Para que los hogares cuenten como efectivamente tratados, se requiere que hayan asistido al menos a 3 sesiones. El grupo de tratamiento se dividió en 2: un grupo a quien se le impartió la información solo a la esposa, y un segundo grupo donde se le dio la información a la pareja.

1. Dada la estructura de panel “2x2”, propongan un modelo de Diferencias en Diferencias (DID) que les permita estimar el efecto de interés de ser tratados en el grupo 1, en el grupo 2 y el efecto combinado. En este caso el efecto que se busca capturar es el de la asistencia efectiva a más de tres sesiones de capacitaciones sobre la probabilidad de usar métodos anticonceptivos para cada grupo de interés.

a) Planteen el modelo de regresión a estimar.

Primera especificación (Grupo 1)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{sola_i} + \beta_3 Post_t * D_{sola_i} + \beta X_{i,t} + e_i$$

Segunda especificación (Grupo 2)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{pareja_i} + \beta_3 Post_t * D_{pareja_i} + \beta X_{i,t} + e_i$$

Tercera especificación (Pooled)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{pooled_i} + \beta_3 Post_t * D_{pooled_i} + \beta X_{i,t} + e_i$$

b) ¿Cuáles son sus supuestos de identificación? ¿Son estos plausibles?

En el caso de diferencias en diferencias, el supuesto principal de identificación requiere que las tendencias en la variable dependiente de los grupos de control y tratamiento tengan tendencias paralelas en el tiempo. Esto es, que el comportamiento de la variable dependiente sea similar en el tiempo, no necesariamente igual en magnitud, para el grupo de control y tratamiento. Así mismo, se requiere que el único medio por el cual se observa la existencia de cambios en la variable dependiente sea a través del tratamiento, y que el efecto sea el mismo para las unidades tratadas. En este caso, debido a que no existen mas de dos periodos de tiempo, es difícil poder ofrecer evidencia suficiente a favor de la existencia de las tendencias paralelas en los grupos de control y de tratamiento. Sin embargo, debido que el entorno del tratamiento se parece a un caso de RCT, es posible implementar la misma lógica para poder determinar el efecto del tratamiento. En este caso, se puede hacer uso de una tabla de balanceo muestral para determinar si existen diferencias significativas en los grupos antes y después del tratamiento. Lo que se esperaría, es que no se encontraran diferencias ni previas ni posteriores al tratamiento, esto comprobando que no existe posibilidad de explicar el cambio en la variable dependiente por fuera del tratamiento. De esta forma, se comprobaría que los resultados potenciales de un periodo a otro deberían ser los mismo en ausencia del tratamiento, y que por lo tanto podría cumplirse la existencia de tendencias paralelas en la tendencia de la variable dependiente.

En este caso, se esperaría que este supuesto se cumpliera sobre los controles. Factores como la presencia de centros de salud, nivel socioeconómico del núcleo familiar o la disponibilidad de recursos públicos para suplir necesidades de la población del municipio pueden traducirse en mayores o menores niveles de conciencia en términos sexuales y reproductivos. Sin embargo, controlando por estos aspectos, se espera que las decisiones sobre el uso de métodos anticonceptivos respondan solo al conocimiento de los beneficios de estos; por lo tanto, no debería haber cambios explicados por algo diferente a la política implementada.

- c) Realice un balance muestral entre los grupos tratados (grupo 1 y grupo 2) antes y después de la intervención para las variables de control relevantes. Interprete.

Tabla 1.1 - Balanceo Muestral grupo 1 en periodo 0

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.948	9.991	-0.043	0.237
Estrato	1.998	1.990	0.008	0.747
Poblacion del municipio	30.109	30.178	-0.069	0.627

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Poblacion en miles

Tabla 1.2 - Balanceo Muestral grupo 1 en periodo 1

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.948	9.991	-0.043	0.237
Estrato	1.998	1.990	0.008	0.747
Poblacion del municipio	30.109	30.178	-0.069	0.627

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Poblacion en miles

Tabla 1.3 - Balanceo Muestral grupo 2 en periodo 0

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.987	10.046	-0.059	0.103
Estrato	1.975	1.992	-0.017	0.491
Poblacion del municipio	30.167	30.101	0.066	0.644

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Poblacion en miles

Tabla 1.4 - Balanceo Muestral grupo 2 en periodo 1

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.987	10.046	-0.059	0.103
Estrato	1.975	1.992	-0.017	0.491
Poblacion del municipio	30.167	30.101	0.066	0.644

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Poblacion en miles

El balanceo muestra corrobora las sospechas expuestas en el literal anterior. Se observa que no existen diferencias significativas entre los grupos de control y tratamiento ni antes ni después del tratamiento. Esto, si bien no representa una prueba formal fuerte sobre la existencia de tendencias paralelas, si ofrece un

ambiente sobre el cual creer que los cambios en la decisión de uso de métodos anticonceptivos no pueden ser explicados por cuestiones diferentes al tratamiento. Así las cosas, se esperaría que, si se observaran más periodos, se pudiera ofrecer mejor evidencia en favor del cumplimiento de tendencias paralelas.

Para el grupo 1 se observa una posible incidencia de la cercanía a un centro de salud sobre el uso de los métodos anticonceptivos. Se podría argumentar que una mujer que no tiene ayuda de su pareja a la hora de decidir usar o no métodos de planificación puede presentar retos mayores al encontrarse lejos de centros de ayuda. Sin embargo, se debe mencionar que aún así, adicional a que la evidencia estadística no es lo suficientemente fuerte, se entiende que la asistencia efectiva al tratamiento supone que los retos existentes por la distancia al centro médico no impiden que la mujer pueda recibir la información en su totalidad, y permite suponer que factores sociales o culturales relacionados a la autonomía de la mujer sobre sus decisiones reproductivas no juegan un papel muy importante en este grupo.

- d) Para cada modelo ¿Qué coeficiente captura el efecto de interés bajo dichos supuestos?

Para los 3 modelos planteados, el coeficiente de interés es el β_3 , el cual corresponde a la interacción entre el periodo de tiempo (Post) y el tratamiento. Este coeficiente captura el efecto de haber sido tratado y de encontrarse en el periodo post. Así, por resultados potenciales, se calcula la diferencia entre el resultado del tratamiento y el contrafactual, que sería el grupo de tratamiento en el evento donde no se hubiera encontrado tratado. El coeficiente se interpreta como el efecto de haber sido tratado sobre la decisión del uso de métodos anticonceptivos.

- e) Estime el efecto para cada uno de los tres grupos de interés (grupo 1, el grupo 2 y efecto combinado) usando la base de datos.

Tabla 1.5 - Efecto de tratamiento efectivo

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento = 1	-0.00147 (0.00638)	-0.00799 (0.00589)	-0.00216 (0.00331)
Tratamiento efectivo en el grupo 1 = 1	0.143*** (0.00859)		
Post = 1, Tratamiento efectivo grupo 1 = 1	0.505*** (0.00944)		

Tratamiento efectivo en el grupo 2 = 1		0.131*** (0.00806)	
Post = 1, Tratamiento efectivo grupo 2 = 1		0.520*** (0.00948)	
Tratamiento efectivo en el grupo 1 o 2 = 1			0.138*** (0.00571)
Post = 1, Tratamiento efectivo grupo 1 o 2 = 1			0.510*** (0.00621)
Constant	0.354*** (0.00409)	0.357*** (0.00453)	0.354*** (0.00230)
Observations	33,042	33,272	100,000
R-squared	0.186	0.181	0.133

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Dado que tenemos un panel “2x2”, no es posible evaluar de forma empírica el cumplimiento del supuesto de tendencias paralelas. Conociendo la aleatorización de la asignación, un asistente de investigación le propone un modelo alternativo de Diferencias en Diferencias (DID) usando la variable de selección a las capacitaciones para medir el impacto sobre el uso de anticonceptivos (ITT).

2. Dado el anterior contexto responda las siguientes preguntas:

- a) Si la idea del asistente es correcta, ¿por qué el coeficiente planteado en el inciso anterior ya no captura el efecto de interés?

La idea detrás de lo dicho por el asistente se refiere al hecho de, al no existir tratamiento efectivo en todas las unidades seleccionadas, pueden existir diferencias preexistentes en los individuos que hayan causado el no terminar de forma efectiva el tratamiento. Así, al calcular el modelo planteado anteriormente, se obtendría el efecto causal de la política entre los efectivamente tratados, y se compararía con el grupo nunca tratado, algo similar a un LATE. Lo que propone el asistente se refiere a un ITT, intento al tratamiento, lo que mide el efecto de haber sido seleccionado para el tratamiento, independientemente de si la unidad fue efectivamente tratada. Esto ayuda a evadir los posibles sesgos que llevaron a los individuos a no terminar el tratamiento de forma efectiva, y logra capturar un efecto de la política sobre la variable de interés.

- b) Si la idea del asistente es correcta, ¿sería posible estimar un efecto consistentemente que cumpla los supuestos necesarios? Si su respuesta es afirmativa, proponga la ecuación que captura dicho efecto.

Al seguir proponiendo un método de diferencias en diferencias, se hace necesario que se siga cumpliendo el supuesto de tendencias paralelas. En este caso, sería necesario realizar nuevamente una tabla de balanceo muestral para determinar si se sigue manteniendo el supuesto aún sobre la asignación y no sobre el tratamiento. A continuación, los resultados.

Tabla 1.6 - Balanceo Muestral grupo 1 en periodo 0

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.963	9.997	-0.034	0.283
Estrato	1.981	2.002	-0.021	0.346
Poblacion del municipio	30.206	30.113	0.092	0.454
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Poblacion en miles				

Tabla 1.7 - Balanceo Muestral grupo 1 en periodo 1

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.963	9.997	-0.034	0.283
Estrato	1.981	2.002	-0.021	0.346
Poblacion del municipio	30.206	30.113	0.092	0.454
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Poblacion en miles				

Tabla 1.8 - Balanceo Muestral grupo 2 en periodo 0

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	10.017	10.045	-0.028	0.369
Estrato	1.981	1.994	-0.013	0.564
Poblacion del municipio	30.193	30.043	0.150	0.225
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Poblacion en miles				

Tabla 1.9 - Balanceo Muestral grupo 2 en periodo 1

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	10.017	10.045	-0.028	0.369
Estrato	1.981	1.994	-0.013	0.564
Poblacion del municipio	30.193	30.043	0.150	0.225
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Poblacion en miles				

En este caso, se observa un mismo comportamiento al observado en la tabla de balanceo muestral realizada con tratamiento efectivo, por lo que es plausible creer que si existieran más periodos de tiempo se podría efectuar mejores pruebas en favor de tendencias paralelas. En este caso, debido que ya no se puede hablar de asistencia efectiva, existe la posibilidad de que la decisión de usar o no métodos anticonceptivos pueda depender de algún factor externo al tratamiento, o en este caso a la intención de tratamiento. Por ende, es necesario ser cauteloso con el resultado obtenido. Confiando en que el planteamiento es correcto, el modelo planteado sería el siguiente.

Primera especificación (Grupo 1)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 P_{sola_i} + \beta_3 Post_t * P_{sola_i} + \beta X_{i,t} + e_i$$

Segunda especificación (Grupo 2)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 P_{pareja_i} + \beta_3 Post_t * P_{pareja_i} + \beta X_{i,t} + e_i$$

Tercera especificación (Pooled)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 P_{pooled_i} + \beta_3 Post_t * P_{pooled_i} + \beta X_{i,t} + e_i$$

- c) Estime el efecto para cada uno de los tres grupos de interés (grupo 1, el grupo 2 y efecto combinado) usando la base de datos.

Tabla 1.10 - Efecto de asignacion a tratamiento

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento = 1	0.000733 (0.00740)	-0.00840 (0.00693)	-0.00108 (0.00360)
Asignado al tratamiento en el grupo 1 = 1	0.0732*** (0.00788)		
Post = 1, Asignacion grupo 1 = 1	0.254*** (0.0101)		
Asignado al tratamiento en el grupo 2 = 1		0.0585*** (0.00801)	
Post = 1, Asignacion grupo 2 = 1		0.259*** (0.0106)	

Asignado al tratamiento en el grupo 1 o 2 = 1			0.0681*** (0.00497)
Post = 1, Asignacion grupo 1 o 2 = 1			0.253*** (0.00646)
Constant	0.354*** (0.00511)	0.361*** (0.00524)	0.355*** (0.00253)
Observations	33,042	33,272	100,000
R-squared	0.073	0.067	0.056

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Sin embargo y dado que se busca capturar es el efecto del tratamiento efectivo, es decir, haber asistido a más de tres sesiones. Usted se aventura a una metodología que involucra utilizar un diferencias en diferencias instrumentado (DID-IV), esta especificación aprovecha explícitamente la aleatoriedad exógena en una tendencia de exposición para estimar el efecto del tratamiento promedio. Para conocer más sobre su implementación le recomendaron leer el artículo de *Hudson, Hull & Liebersohn (2017)*.

3. Dado el anterior contexto responda las siguientes preguntas, para ello utilice el artículo citado anteriormente.

a) Planteen el modelo de regresión a estimar.

De acuerdo con Hudson, Hull & Liebersohn (2017), la ecuación estructural, la primera etapa y la ecuación reducida son las siguientes.

Primera especificación (Grupo 1)

Ecuación estructural:

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{sola_i} + e_i$$

Primera etapa:

$$D_{sola_i} = \alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{sola_i} + \eta_i$$

Forma reducida:

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 (\alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{sola_i} + \eta_i) + e_i$$

$$\begin{aligned} Planificación_{i,t} \\ &= \beta_0 + \beta_1 Post_t + \beta_2 \alpha_0 + \beta_2 \alpha_1 Post_t + \beta_2 \alpha_2 Post_t P_{sola_i} + \beta_2 \eta_i \\ &+ e_i \end{aligned}$$

$$Planificación_{i,t} = \gamma_0 + \gamma_1 Post_t + \phi Post_t P_{sola_i} + \mu_i$$

Donde $\gamma_0 = \beta_0 + \beta_2 \alpha_0$, $\gamma_1 = \beta_1 + \beta_2 \alpha_1$, $\phi = \beta_2 \alpha_2$ y $\mu_i = \beta_2 \eta_i + e_i$

Segunda especificación (Grupo 2)

Ecuación estructural:

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{pareja_i} + e_i$$

Primera etapa:

$$D_{sola_i} = \alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{pareja_i} + \eta_i$$

Forma reducida:

$$\begin{aligned} Planificación_{i,t} \\ = \beta_0 + \beta_1 Post_t + \beta_2 (\alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{pareja_i} + \eta_i) + e_i \end{aligned}$$

$$\begin{aligned} Planificación_{i,t} \\ = \beta_0 + \beta_1 Post_t + \beta_2 \alpha_0 + \beta_2 \alpha_1 Post_t + \beta_2 \alpha_2 Post_t P_{pareja_i} + \beta_2 \eta_i \\ + e_i \end{aligned}$$

$$Planificación_{i,t} = \gamma_0 + \gamma_1 Post_t + \phi Post_t P_{pareja_i} + \mu_i$$

Donde $\gamma_0 = \beta_0 + \beta_2 \alpha_0$, $\gamma_1 = \beta_1 + \beta_2 \alpha_1$, $\phi = \beta_2 \alpha_2$ y $\mu_i = \beta_2 \eta_i + e_i$

Tercera especificación (Pooled)

Ecuación estructural:

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{pooled_i} + e_i$$

Primera etapa:

$$D_{sola_i} = \alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{pooled_i} + \eta_i$$

Forma reducida:

$$\begin{aligned} Planificación_{i,t} \\ = \beta_0 + \beta_1 Post_t + \beta_2 (\alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{pooled_i} + \eta_i) + e_i \end{aligned}$$

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 \alpha_0 + \beta_2 \alpha_1 Post_t + \beta_2 \alpha_2 Post_t P_{pooled_i} + \beta_2 \eta_i + e_i$$

$$Planificación_{i,t} = \gamma_0 + \gamma_1 Post_t + \phi Post_t P_{pooled_i} + \mu_i$$

Donde $\gamma_0 = \beta_0 + \beta_2 \alpha_0$, $\gamma_1 = \beta_1 + \beta_2 \alpha_1$, $\phi = \beta_2 \alpha_2$ y $\mu_i = \beta_2 \eta_i + e_i$

b) ¿Cuáles son sus supuestos de identificación? ¿Son estos plausibles?

Como hablamos de un DDIV, es necesario que se cumplan 4 supuestos principales según Hudson, Hull & Liebersohn (2017). La restricción de exclusión, que en este caso sería que el instrumento afecte a la variable dependiente únicamente a través del tratamiento, y que el instrumento solo afecte a las unidades dependientes contemporáneas; esto es decir, que instrumento en periodo t , no afecte los resultados potenciales del periodo $t-1$. Adicional a estos, se requiere también la existencia de tendencias paralelas en la senda de crecimiento de los resultados potenciales de la variable dependiente. Finalmente, el supuesto de monotonicidad, el cual establece que el efecto del tratamiento solo se observará en el periodo 1. En términos de resultados potenciales: Sea S_{it} un instrumento dicotómico, $P(S_{i1}^1 \geq S_{i1}^0) > 1$.

Así mismo, pensando en las características de un buen instrumento, usar la asignación al tratamiento cumple con las características necesarias. El instrumento es independiente, debido a que se realizó una aleatorización, la cual se corroboró como bien hecha por medio de las tablas de balanceo muestral, y no responde a las características previas de los grupos de tratamiento y de control. En cuanto a relevancia, se espera que este se cumpla debido que un tratamiento efectivo es función de la asignación, quien no es asignado a tratamiento no podrá ser tratado efectivo como alta probabilidad.

Es plausible que se cumplan estos supuestos como resultado de la naturaleza del instrumento y del tratamiento. Por el lado de la restricción de exclusión, debido a que el instrumento es la asignación a alguno de los grupos de tratamiento, la decisión del individuo solo será afectada a través del tratamiento (asistencia a 3 sesiones). Como se intentó probar en el literal a y b, es plausible suponer la existencia de tendencias paralelas. Por construcción del instrumento, se cumple monotonicidad, y debido a que el tratamiento efectivo es función de la asignación a grupos de tratamiento, se espera que se cumpla la condición de relevancia. Finalmente, la aleatorización en la asignación del tratamiento hace plausible suponer que el instrumento es independiente.

- c) ¿Qué coeficiente representa su efecto de interés bajo dichos supuestos? Interprete.

El coeficiente de interés corresponde a los correspondientes en la ecuación $\phi = \beta_2 \alpha_2$. Así, se busca tener el valor del coeficiente α_2 por medio de la estimación de la primera etapa, y el coeficiente ϕ por medio de la estimación de la ecuación reducida. Luego, el parámetro de interés se observa de la siguiente forma.

$$\beta_2 = \frac{\phi}{\alpha_2} = \frac{E[Y_{i1} - Y_{i0}|Z_i = 1] - E[Y_{i1} - Y_{i0}|Z_i = 0]}{E[S_{i1} - S_{i0}|Z_i = 1] - E[S_{i1} - S_{i0}|Z_i = 0]}$$

En términos de los parámetros del modelo,

$$\beta = \frac{E[Planificación_{i1} - Planificación_{i0}|P_{j_i} = 1] - E[Planificación_{i1} - Planificación_{i0}|P_{j_i} = 0]}{E[D_{j_{i1}} - D_{j_{i0}}|P_{j_i} = 1] - E[D_{j_{i1}} - D_{j_{i0}}|P_{j_i} = 0]}$$

Donde $j \in (sola, pareja, pooled)$.

Hudson, Hull & Liebersohn (2017) demuestran que el parámetro de interés β se puede expresar de la siguiente forma.

$$\beta = \sum_{s>0} \omega_s E[Y_{i1}^s - Y_{i1}^{s-1} | S_{i1}^1 \geq s > S_{i0}^0]$$

Donde

$$\omega_s = \frac{P(S_{i1}^1 \geq s > S_{i0}^0)}{\sum_{r>0} P(S_{i1}^1 \geq r > S_{i0}^0)}$$

Luego, el parámetro captura la respuesta causal promedio (average causal response), que no es más que el efecto promedio por pesos (ω) de la respuesta causal a cambios en una unidad en el tratamiento en el periodo 1. En este caso en particular, el coeficiente se interpreta como el efecto causal del programa dato que se ofreció el programa.

- d) Implemente la metodología anterior planteada en Stata. Para ello utilice el paquete *xtivreg*, que le permite realizar regresión de variables instrumentales para datos panel. En una tabla para tipo de grupo (grupo 1, grupo 2 y combinado) presente los resultados de la primera etapa, segunda etapa y la forma reducida. Interprete.

Tabla 1.11 - Primera Etapa

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento	-0.258*** (0.00346)	-0.248*** (0.00530)	-0.168*** (0.00236)
Post = 1, Asignacion grupo 1 = 1	0.511*** (0.00503)		
Post = 1, Asignacion grupo 2 = 1		0.497*** (0.00612)	
Post = 1, Asignacion grupo 1 o 2 = 1			0.504*** (0.00317)
Constant	0.258*** (0.00346)	0.248*** (0.00306)	0.168*** (0.00149)
Observations	33,042	33,272	100,000
R-squared	0.170	0.165	0.202

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Tabla 1.13 - Segunda Etapa

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Tratamiento efectivo en el grupo 1	0.392*** (0.0101)		
Tratamiento efectivo en el grupo 2		0.378*** (0.0103)	
Tratamiento efectivo en el grupo 1 o 2			0.387*** (0.00628)
Constant	0.354*** (0.00365)	0.356*** (0.00363)	0.354*** (0.00182)
Observations	33,042	33,272	100,000
Number of id_hogar	16,521	16,636	50,000

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Tabla 1.12 - Forma Reducida

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento = 1	0.000733	-0.00840	-0.00108

	(0.00749)	(0.00745)	(0.00371)
Asignado al tratamiento en el grupo 1 = 1	0.0732***		
	(0.00746)		
Post = 1, Asignacion grupo 1 = 1	0.254***		
	(0.0106)		
Asignado al tratamiento en el grupo 2 = 1		0.0585***	
		(0.00745)	
Post = 1, Asignacion grupo 2 = 1		0.259***	
		(0.0105)	
Asignado al tratamiento en el grupo 1 o 2 = 1			0.0681***
			(0.00455)
Post = 1, Asignacion grupo 1 o 2 = 1			0.253***
			(0.00643)
Constant	0.354***	0.361***	0.355***
	(0.00530)	(0.00526)	(0.00262)
Observations	33,042	33,272	100,000
R-squared	0.073	0.067	0.056

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

e) Compare con los resultados obtenidos con aquellos en el inciso a) y b).

En este caso, se encuentran efectos positivos en los 3 grupos: 0.392, 0.378 y 0.354 puntos porcentuales respectivamente. Así, se observa un notable orden en el tamaño de los coeficientes de acuerdo con cuál es la especificación que se observe. La primera especificación, tratamiento efectivo, ofrece coeficientes de magnitud 0.505, 0.520 y 0.510 respectivamente. La segunda, asignación a tratamiento, 0.254, 0.259 y 0.253 respectivamente. Así las cosas, se observa que, en cuanto a tamaño, la segunda especificación ofrece los coeficientes de menor magnitud, seguida por la tercera, y finalmente la primera especificación con los coeficiente de mayor magnitud.

Lo anterior es reflejo del problema de identificación que se enfrenta. Debido a que la decisión de asistir a todas las charlas es endógena, se espera que el efecto capturado no refleje el verdadero efecto del programa debido al sesgo existente. Por ello, tiene sentido que el coeficiente asociado a esta especificación sea el mayor. Por el lado de la segunda especificación, no se está estimando un ATE, sino un ITT. Esto significa que no es el efecto causal de aplicar el programa, sino de ofrecer el programa. Por ende, no se captura el efecto real que el programa tuvo sobre las decisiones de uso de métodos anticonceptivos. Finalmente, el método de IV permite identificar el efecto del tratamiento gracias a que se ofreció el tratamiento. Así, se encuentra un efecto causal del tratamiento sobre los que se

ofreció el tratamiento (LATE), y se determina que este tuvo efectos positivos sobre las decisiones de uso de métodos anticonceptivos.

Una de las participantes del programa le hace conocer que factiblemente el efecto de las capacitaciones puede cambiar dependiendo del sexo del capacitador (i.e. quien imparte la información). Esto debido a que existen sesgos de género marcados y sistemáticos que puedan ser decisivos para determinar el efecto del tratamiento en un contexto en donde la sexualidad y planificación femenina siguen siendo un tema tabú. Es así como usted se plantea explorar la heterogeneidad en los efectos del tratamiento según el sexo del capacitador. Suponga que se cumple tendencias paralelas entre el grupo de control y tratamiento en los tres casos de tratamiento (pooled, grupo 1 y grupo 2).

3. Para medir lo anterior usted plantea una interacción entre el sexo del capacitador con el estado y el tiempo del tratamiento, es decir, una estimación de triple diferencia. Para ello:

a) Planteen el modelo de regresión a estimar.

Primera especificación (Grupo 1)

$$\begin{aligned} \text{Planificación}_{i,t} &= \beta_0 + \beta_1 \text{Post}_t + \beta_2 P_{\text{sola}_i} + \beta_3 \text{Sexo}_{\text{capacitador}_i} + \beta_4 \text{Post}_t P_{\text{sola}_i} \\ &+ \beta_5 \text{Post}_t \text{Sexo}_{\text{capacitador}_i} + \beta_6 P_{\text{sola}_i} \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_7 \text{Post}_t P_{\text{sola}_i} \text{Sexo}_{\text{capacitador}_i} + \beta X_{i,t} + e_i \end{aligned}$$

Segunda especificación (Grupo 2)

$$\begin{aligned} \text{Planificación}_{i,t} &= \beta_0 + \beta_1 \text{Post}_t + \beta_2 P_{\text{pareja}_i} + \beta_3 \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_4 \text{Post}_t P_{\text{pareja}_i} + \beta_5 \text{Post}_t \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_6 P_{\text{pareja}_i} \text{Sexo}_{\text{capacitador}_i} + \beta_7 \text{Post}_t * P_{\text{pareja}_i} \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta X_{i,t} + e_i \end{aligned}$$

Tercera especificación (Pooled)

$$\begin{aligned} \text{Planificación}_{i,t} &= \beta_0 + \beta_1 \text{Post}_t + \beta_2 P_{\text{pooled}_i} + \beta_3 \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_4 \text{Post}_t P_{\text{pooled}_i} + \beta_5 \text{Post}_t \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_6 P_{\text{pooled}_i} \text{Sexo}_{\text{capacitador}_i} + \beta_7 \text{Post}_t P_{\text{pooled}_i} \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta X_{i,t} + e_i \end{aligned}$$

b) ¿Cuáles son sus supuestos de identificación? ¿Son estos plausibles?

Nuevamente, se debe mantener el supuesto inicial de tendencias paralelas para poder asegurar que se puede capturar el efecto causal del tratamiento. En este caso en particular, debido a la triple interacción, las tendencias deben mantenerse entre los grupos de control y tratamiento, y adicionalmente en el sexo de los capacitadores. Esto es, las tendencias del uso de métodos anticonceptivos entre individuos, al mirar los grupos que fueron atendidos por encuestador hombre y encuestadora mujer, deben ser paralelas. Así, el supuesto se hace más exigente al requerir tendencias paralelas no solo entre grupos de tratamiento y control, sino también entre la característica a observar.

En este caso, existe alta probabilidad de que se cumpla el supuesto como resultado de que las decisiones de uso de anticonceptivos son independientes de la selección del sexo del encuestador. Adicionalmente, debido a que la proporción de encuestadores hombre y mujer es casi del 50/50, la aleatorización en la asignación del encuestador al hogar sería suficiente para garantizar la existencia del supuesto. Y si no se hubiese asignado por aleatorización, aun así, el supuesto se mantendría.

c) ¿Qué coeficiente representa su efecto de interés bajo dichos supuestos? Interprete.

En este caso, el coeficiente que captura el efecto de interés corresponde al β_7 en las 3 ecuaciones planteadas. Este coeficiente, captura el efecto de haber sido tratado, y de haber sido encuestado por una mujer, y se interpretaría como el efecto que tiene sobre las unidades tratadas el sexo del encuestador. Así, si existen diferencias significativas, se podría concluir que existen sesgos sociales a causa del tabú de la sexualidad respecto a las mujeres, y poder entender la verdadera incidencia del tratamiento sobre el uso de métodos anticonceptivos. En este caso, el coeficiente se interpreta como el efecto de haber tenido una encuestadora mujer respecto a un encuestador hombre sobre los efectos del tratamiento. Así, se observa si el sexo del encuestador influye sobre el tratamiento.

d) Estime el efecto para cada uno de los tres grupos de interés (grupo 1, el grupo 2 y efecto combinado) usando la base de datos. *Planificación_DID.dta* e interprete. Inserte el código utilizado.

Tabla 1.14 - Efecto del sexo del encuestador

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento = 1	-0.00596 (0.00956)	0.00528 (0.0110)	0.00108 (0.00513)
Sexo del capacitador: 1.mujer, 0.hombre = 1	-0.00909 (0.00939)	0.000866 (0.0111)	0.00131 (0.00512)
Post = 1, Sexo capacitador = 1	0.0132 (0.0138)	-0.0274* (0.0162)	-0.00433 (0.00771)
Asignado al tratamiento en el grupo 1 = 1	0.0660*** (0.0112)		
Post = 1, Asignacion grupo 1 = 1	0.269*** (0.0141)		
Sexo capacitador = 1, Asignacion grupo 1 = 1	0.0143 (0.0141)		
Post = 1, Asignacion grupo 1 = 1, Sexo capacitador = 1	-0.0303 (0.0221)		
Asignado al tratamiento en el grupo 2 = 1		0.0549*** (0.0121)	
Post = 1, Asignacion grupo 2 = 1		0.250*** (0.0170)	
Sexo capacitador = 1, Asignacion grupo 2 = 1		0.00696 (0.0167)	
Post = 1, Asignacion grupo 2 = 1, Sexo capacitador = 1		0.0189 (0.0241)	
Asignado al tratamiento en el grupo 1 o 2 = 1			0.0655*** (0.00709)
Post = 1, Asignacion grupo 1 o 2 = 1			0.258*** (0.00947)
Sexo capacitador = 1, Asignacion grupo 1 o 2 = 1			0.00509 (0.00945)
Post = 1, Asignacion grupo 1 o 2 = 1, Sexo capacitador = 1			-0.00853 (0.0139)
Constant	0.358*** (0.00747)	0.360*** (0.00788)	0.354*** (0.00364)
Observations	33,042	33,272	100,000
R-squared	0.073	0.068	0.056

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

En este caso, se observa que los coeficientes de interés para los 3 grupos el coeficiente de interés no es significativo. Sin embargo, es necesario analizar el signo del coeficiente.

Para el grupo 1, tratamiento a la mujer sola, se observa que si el encuestador es de sexo femenino, la probabilidad de uso de métodos anticonceptivos cae en 0,03 puntos porcentuales. Este resultado resulta muy contraintuitivo, en el sentido de que se esperaría que cuando la mujer es la única responsable en el cuidado de la salud reproductiva, una encuestadora mujer facilitaría el recibimiento de la información, ponerse en su lugar y ofrecer información convincente respecto a los beneficios de los métodos anticonceptivos.

Por el lado del grupo 2, donde se ofrecieron capacitaciones a la pareja, se observa un signo positivo, lo que significa que la probabilidad de uso de métodos anticonceptivos incrementó en 0.024 puntos porcentuales. Se espera que mujeres con acompañamiento de su pareja en la decisión de uso de métodos anticonceptivos sean más propensas a su uso, y que la encuesta sea realizada por una mujer permitiría convencer al hombre de los beneficios de los métodos anticonceptivos de forma más efectiva; una mujer hablando de la necesidad de los anticonceptivos para el bienestar sexual y reproductivo de las mujeres.

Finalmente, el grupo 3 tiene un coeficiente negativo, y significa que el haber recibido a una encuestadora mujer significó una reducción de la probabilidad de uso de métodos anticonceptivos en 0.0085 puntos porcentuales.

Codigo empleado:

```
qui reg uso post##sexo##P_s if Tipo==1, vce(cluster id_mun)
outreg2 using "${out}\\Punto_1_d_iv.doc", replace ctitle("Grupo 1")
qui reg uso post##sexo##P_pa if Tipo==2, vce(cluster id_mun)
outreg2 using "${out}\\Punto_1_d_iv.doc", ctitle("Grupo 2")
qui reg uso post##sexo##P_po, vce(cluster id_mun)
outreg2 using "${out}\\Punto_1_d_iv.doc", label ctitle("Efecto combinado") ///
title("Tabla 1.14 - Efecto del sexo del encuestador")
```

Segundo ejercicio

El método de diferencias en diferencias es quizá el método más utilizado en la última década por los economistas en las investigaciones académicas empíricas. Las numerosas aplicaciones condujeron a un interés de desarrollar extensiones a esta metodología. Quizá el desarrollo más importante en esta índole llegó con el método de "estudio de eventos", el cual permitió generalizar diferencias en diferencias a escenarios donde las unidades recibían el tratamiento en distintos momentos del tiempo.

Las primeras aproximaciones de los economistas con los estudios de eventos se basaron en la intuición que se tenían acerca de la metodología de diferencias en diferencias, sin mucho cuidado acerca de si esta generalización era, en principio, siquiera posible. Consecuentemente, los econométricos teóricos empezaron a encontrar varios problemas asociados a la manera en que se utilizaba comúnmente esta metodología, concluyendo que existían principalmente dos problemas: (1) contrario a otras metodologías, no era muy claro el efecto causal que identifica los estudios de eventos; y (2), la estimación por mínimos cuadrados ordinarios de un modelo de panel lineal con efectos fijos, conocida como “two-way fixed effect estimation (TWFE)”, es inadecuada, en tanto no captura un efecto causal salvo para casos muy restrictivos para el análisis empírico.

Quizá el primer trabajo en poner en discusión las falencias del TWFE en estudios de evento es el trabajo de Goodman-Bacon (2021). En breve, este trabajo tiene dos grandes contribuciones. La primera, descompone el estimador de mínimos cuadrados ordinarios que resulta de estimar un TWFE de manera que se puede entender de manera intuitiva lo que se está estimando. Lo segundo, muestra que incluso bajo el cumplimiento de tendencias paralelas, el estimador no recupera un efecto causal en presencia de heterogeneidad en los efectos de tratamiento.

El objetivo de este punto es entender mejor el planteamiento que se hace en Goodman-Bacon (2021) y de paso, repasar algunos temas que vimos en clase acerca de los estudios de eventos.

a) Corran el Do File Sim1 Do que simula una pequeña “base de datos” con la cual nos aproximaremos al problema.

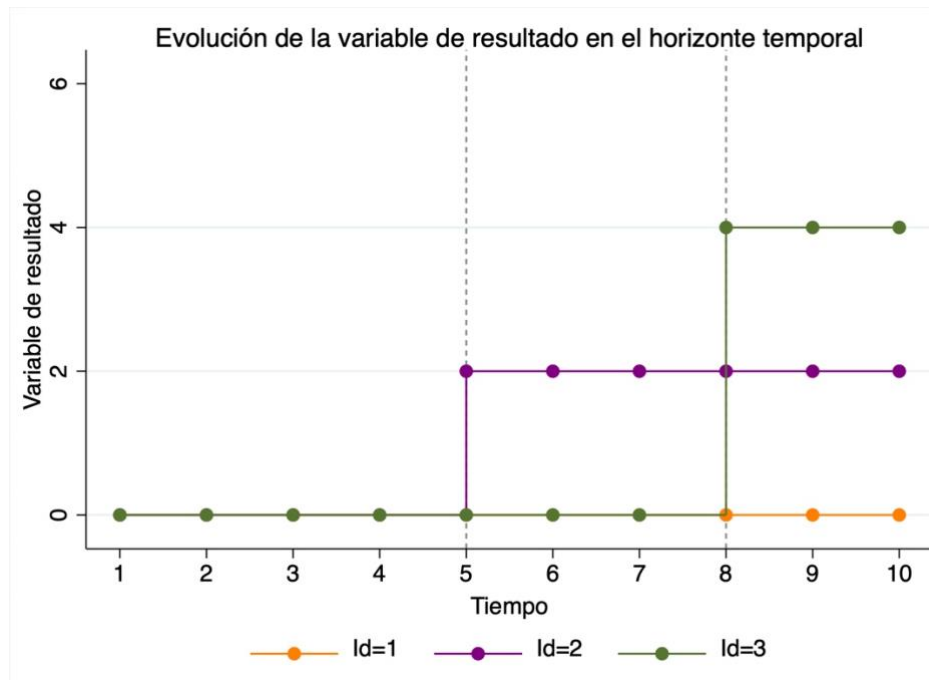
i) Describan en palabras qué hace el código

Solución:

El código crea y organiza un panel de datos. En particular, crea información con respecto al estado del tratamiento de las unidades de análisis (variable D) y con respecto a una variable de resultado (variable Y) para diez períodos de tiempo distintos. Adicionalmente, lo organiza con el fin de que se pueda observar fácilmente si las unidades de análisis fueron tratadas en algún punto del horizonte temporal y en dado caso ver el efecto de dicho tratamiento sobre la variable de resultado.

ii) Ilustren en una gráfica cómo varía la variable dependiente (con nombre Y en la base de datos) en función del tiempo para cada unidad. Ilustren con una línea punteada el momento donde cada unidad es tratada. ¿Cuántas cohortes hay? Describa cada una brevemente.

Solución:



Como es posible evidenciar en la gráfica presentada anteriormente, en la base de datos suministrada hay 3 cohortes distintas: una que no recibe el tratamiento en ningún punto del horizonte temporal, otra que recibe el tratamiento en el período $t=5$ ($id=2$) y una última que recibe el tratamiento en el momento $t=8$ ($id=3$).

iii) Describan matemáticamente $Y_{i,t}(g)$, donde g hace referencia a la cohorte.
¿Cuál es el ATT para cada cohorte tratada?

Solución:

Solución 2a.iii:

primero, el resultado potencial está dado por:

$$y_{it}(g) = \begin{cases} \eta_i + \phi_t + \varepsilon_{it} & \text{cuando } t < G_i \\ \eta_i + \phi_t + \tau_{i,t}(g) + \varepsilon_{it} & t \geq G_i \end{cases}$$

En donde:

η_i, ϕ_t son efectos fijos de unidad y tiempo respectivamente

$G_i = g$ es el momento en el que se trata a la cohorte por primera vez.

De esta forma, los ATT para $g=2$ y $g=3$ están dados por:

$$ATT(2, t) = E[y_{it}(g) - y_{it}(\infty) | G_i = 5]$$

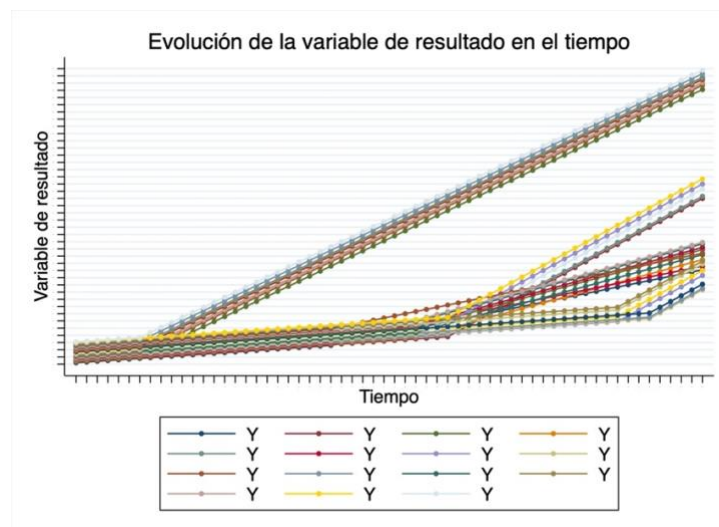
$$ATT(3, t) = E[y_{it}(g) - y_{it}(\infty) | G_i = 8]$$

Nota: $y_{it}(\infty)$ denota el resultado potencial de nunca ser tratado.

d) Para la nueva muestra simulada

- i) Grafique en una única gráfica la variable dependiente de cada unidad en el tiempo.

Solución:



Dada la gráfica anterior y el código proporcionado por el documento *Sim_Do2*, se puede concluir que hay un total de 6 cohortes en la base de datos. Adicionalmente, se puede ver que la evolución de la variable de resultado antes del tratamiento es de naturaleza paralela para los grupos y que a medida que las cohortes se van exponiendo al tratamiento estas cambian su pendiente. En particular, parece que el tratamiento hace que se aumente la pendiente de la variable Y para las cohortes, lo cual indica un posible efecto positivo del tratamiento.

- ii) *Estime el modelo por TWFE. ¿Es coherente el signo encontrado con lo que se espera teóricamente? De acuerdo a lo visto anteriormente, ¿qué puede estar causando este problema?*

Solución:

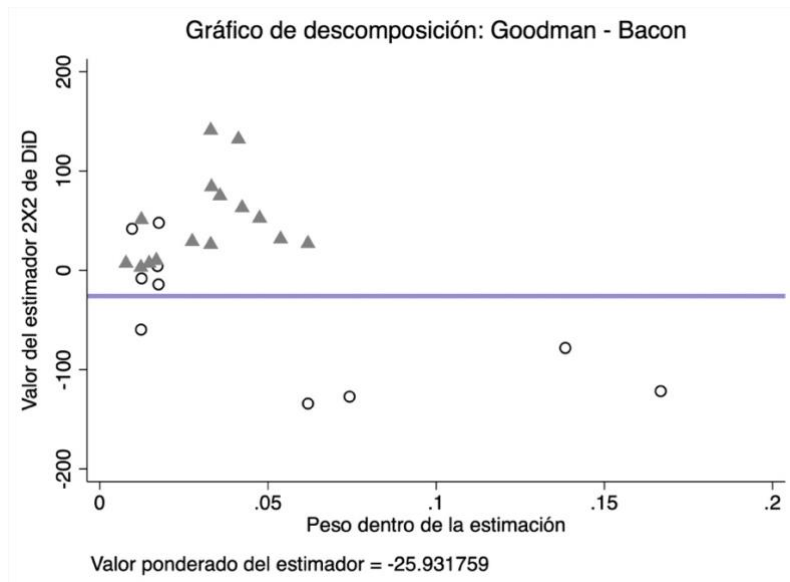
Y	Coeficiente	Errores estándar	P>t
D	-25.932	3.375	0.000
Constante	114.935	1.997	0.000

Tras estimar el modelo por TWFE, es posible observar que el tratamiento efectivamente es estadísticamente significativo. Sin embargo, el efecto de dicho tratamiento sobre la variable de resultado resulta negativo, lo cual contradice el argumento gráfico desarrollado en el punto anterior.

En ese sentido, una posible causa de este comportamiento contraintuitivo del tratamiento sobre la variable de resultado es que, dentro del estimador de TWFE, las comparaciones que indican un efecto negativo del tratamiento están teniendo un mayor peso en la ponderación.

- e) *Para la nueva muestra simulada, presente de manera gráfica la descomposición de Goodman-Bacon. Interprete lo que encontró en el inciso d.ii) a la luz de esta nueva evidencia.*

Solución:



Según la gráfica de descomposición de Goodman – Bacon, en la cual se puede ver el peso de cada comparación 2X2 de diferencias en diferencias dentro de la estimación, así como su respectivo valor, es posible corroborar la hipótesis planteada en el punto anterior. En particular, es posible observar que el valor ponderado del estimador es prácticamente el mismo. Sin embargo, la gráfica indica que hay unas pocas comparaciones que tienen un valor considerablemente negativo y que cuentan con un peso relativo muy grande dentro de la estimación. De esta forma, tras desagregar el estimador de TWFE mediante la metodología de Goodman – Bacon, es posible atribuir el signo negativo del estimador a unas pocas comparaciones con un peso alto dentro de la estimación y que a su vez tienen signo negativo.

f) Usando el estimador de Callaway & Sant’Anna, estime los efectos causales dinámicos del tratamiento en la muestra simulada (i.e, el efecto promedio del tratamiento τ periodos antes o después de que una cohorte es tratada).

¿La estimación parece ser consistente con efectos causales de la simulación? ¿La estimación parece ser consistente con la información que tenemos acerca del cumplimiento de tendencias paralelas?

Solución:

Taller 1
Econometría Avanzada, 2022-2
Profesor: Manuel Fernández

Carlos A. Ayala - 201911488
Carlos Eduardo Torres - 202225155
Juan Jacobo Trujillo Pradilla - 201923274

Todos los elementos utilizados para la solución de este taller (bases de datos, dofile, outputs y demás), pueden ser consultados en el siguiente enlace:
<https://github.com/CarlosAyala123/Econometria-Avanzada-Taller-3>

Primer ejercicio:

El Instituto de Salud Sexual y Reproductiva (ISSR) realizó un experimento aleatorio para entender el efecto de dar información de planificación familiar en el comportamiento de fertilidad de las parejas. Para ello, se seleccionaron 50.000 hogares al azar, provenientes de 300 municipios rurales. Se seleccionaron 200 hogares por medio de una lotería justa para ser asignados al grupo de tratamiento, y los 100 hogares restantes fueron asignados al grupo de control. Para que los hogares cuenten como efectivamente tratados, se requiere que hayan asistido al menos a 3 sesiones. El grupo de tratamiento se dividió en 2: un grupo a quien se le impartió la información solo a la esposa, y un segundo grupo donde se le dio la información a la pareja.

1. Dada la estructura de panel “2x2”, propongan un modelo de Diferencias en Diferencias (DID) que les permita estimar el efecto de interés de ser tratados en el grupo 1, en el grupo 2 y el efecto combinado. En este caso el efecto que se busca capturar es el de la asistencia efectiva a más de tres sesiones de capacitaciones sobre la probabilidad de usar métodos anticonceptivos para cada grupo de interés.

a) Planteen el modelo de regresión a estimar.

Primera especificación (Grupo 1)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{sola_i} + \beta_3 Post_t * D_{sola_i} + \beta X_{i,t} + e_i$$

Segunda especificación (Grupo 2)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{pareja_i} + \beta_3 Post_t * D_{pareja_i} + \beta X_{i,t} + e_i$$

Tercera especificación (Pooled)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{pooled_i} + \beta_3 Post_t * D_{pooled_i} + \beta X_{i,t} + e_i$$

b) ¿Cuáles son sus supuestos de identificación? ¿Son estos plausibles?

En el caso de diferencias en diferencias, el supuesto principal de identificación requiere que las tendencias en la variable dependiente de los grupos de control y tratamiento tengan tendencias paralelas en el tiempo. Esto es, que el comportamiento de la variable dependiente sea similar en el tiempo, no necesariamente igual en magnitud, para el grupo de control y tratamiento. Así mismo, se requiere que el único medio por el cual se observa la existencia de cambios en la variable dependiente sea a través del tratamiento, y que el efecto sea el mismo para las unidades tratadas. En este caso, debido a que no existen mas de dos periodos de tiempo, es difícil poder ofrecer evidencia suficiente a favor de la existencia de las tendencias paralelas en los grupos de control y de tratamiento. Sin embargo, debido que el entorno del tratamiento se parece a un caso de RCT, es posible implementar la misma lógica para poder determinar el efecto del tratamiento. En este caso, se puede hacer uso de una tabla de balanceo muestral para determinar si existen diferencias significativas en los grupos antes y después del tratamiento. Lo que se esperaría, es que no se encontraran diferencias ni previas ni posteriores al tratamiento, esto comprobando que no existe posibilidad de explicar el cambio en la variable dependiente por fuera del tratamiento. De esta forma, se comprobaría que los resultados potenciales de un periodo a otro deberían ser los mismo en ausencia del tratamiento, y que por lo tanto podría cumplirse la existencia de tendencias paralelas en la tendencia de la variable dependiente.

En este caso, se esperaría que este supuesto se cumpliera sobre los controles. Factores como la presencia de centros de salud, nivel socioeconómico del núcleo familiar o la disponibilidad de recursos públicos para suplir necesidades de la población del municipio pueden traducirse en mayores o menores niveles de conciencia en términos sexuales y reproductivos. Sin embargo, controlando por estos aspectos, se espera que las decisiones sobre el uso de métodos anticonceptivos respondan solo al conocimiento de los beneficios de estos; por lo tanto, no debería haber cambios explicados por algo diferente a la política implementada.

- c) Realice un balance muestral entre los grupos tratados (grupo 1 y grupo 2) antes y después de la intervención para las variables de control relevantes. Interprete.

Tabla 1.1 - Balanceo Muestral grupo 1 en periodo 0

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.948	9.991	-0.043	0.237
Estrato	1.998	1.990	0.008	0.747
Poblacion del municipio	30.109	30.178	-0.069	0.627

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Poblacion en miles

Tabla 1.2 - Balanceo Muestral grupo 1 en periodo 1

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.948	9.991	-0.043	0.237
Estrato	1.998	1.990	0.008	0.747
Poblacion del municipio	30.109	30.178	-0.069	0.627

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Poblacion en miles

Tabla 1.3 - Balanceo Muestral grupo 2 en periodo 0

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.987	10.046	-0.059	0.103
Estrato	1.975	1.992	-0.017	0.491
Poblacion del municipio	30.167	30.101	0.066	0.644

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Poblacion en miles

Tabla 1.4 - Balanceo Muestral grupo 2 en periodo 1

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.987	10.046	-0.059	0.103
Estrato	1.975	1.992	-0.017	0.491
Poblacion del municipio	30.167	30.101	0.066	0.644

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$
Poblacion en miles

El balanceo muestra corrobora las sospechas expuestas en el literal anterior. Se observa que no existen diferencias significativas entre los grupos de control y tratamiento ni antes ni después del tratamiento. Esto, si bien no representa una prueba formal fuerte sobre la existencia de tendencias paralelas, si ofrece un

ambiente sobre el cual creer que los cambios en la decisión de uso de métodos anticonceptivos no pueden ser explicados por cuestiones diferentes al tratamiento. Así las cosas, se esperaría que, si se observaran más periodos, se pudiera ofrecer mejor evidencia en favor del cumplimiento de tendencias paralelas.

Para el grupo 1 se observa una posible incidencia de la cercanía a un centro de salud sobre el uso de los métodos anticonceptivos. Se podría argumentar que una mujer que no tiene ayuda de su pareja a la hora de decidir usar o no métodos de planificación puede presentar retos mayores al encontrarse lejos de centros de ayuda. Sin embargo, se debe mencionar que aún así, adicional a que la evidencia estadística no es lo suficientemente fuerte, se entiende que la asistencia efectiva al tratamiento supone que los retos existentes por la distancia al centro médico no impiden que la mujer pueda recibir la información en su totalidad, y permite suponer que factores sociales o culturales relacionados a la autonomía de la mujer sobre sus decisiones reproductivas no juegan un papel muy importante en este grupo.

- d) Para cada modelo ¿Qué coeficiente captura el efecto de interés bajo dichos supuestos?

Para los 3 modelos planteados, el coeficiente de interés es el β_3 , el cual corresponde a la interacción entre el periodo de tiempo (Post) y el tratamiento. Este coeficiente captura el efecto de haber sido tratado y de encontrarse en el periodo post. Así, por resultados potenciales, se calcula la diferencia entre el resultado del tratamiento y el contrafactual, que sería el grupo de tratamiento en el evento donde no se hubiera encontrado tratado. El coeficiente se interpreta como el efecto de haber sido tratado sobre la decisión del uso de métodos anticonceptivos.

- e) Estime el efecto para cada uno de los tres grupos de interés (grupo 1, el grupo 2 y efecto combinado) usando la base de datos.

Tabla 1.5 - Efecto de tratamiento efectivo

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento = 1	-0.00147 (0.00638)	-0.00799 (0.00589)	-0.00216 (0.00331)
Tratamiento efectivo en el grupo 1 = 1	0.143*** (0.00859)		
Post = 1, Tratamiento efectivo grupo 1 = 1	0.505*** (0.00944)		

Tratamiento efectivo en el grupo 2 = 1		0.131*** (0.00806)	
Post = 1, Tratamiento efectivo grupo 2 = 1		0.520*** (0.00948)	
Tratamiento efectivo en el grupo 1 o 2 = 1			0.138*** (0.00571)
Post = 1, Tratamiento efectivo grupo 1 o 2 = 1			0.510*** (0.00621)
Constant	0.354*** (0.00409)	0.357*** (0.00453)	0.354*** (0.00230)
Observations	33,042	33,272	100,000
R-squared	0.186	0.181	0.133

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Dado que tenemos un panel “2x2”, no es posible evaluar de forma empírica el cumplimiento del supuesto de tendencias paralelas. Conociendo la aleatorización de la asignación, un asistente de investigación le propone un modelo alternativo de Diferencias en Diferencias (DID) usando la variable de selección a las capacitaciones para medir el impacto sobre el uso de anticonceptivos (ITT).

2. Dado el anterior contexto responda las siguientes preguntas:

- a) Si la idea del asistente es correcta, ¿por qué el coeficiente planteado en el inciso anterior ya no captura el efecto de interés?

La idea detrás de lo dicho por el asistente se refiere al hecho de, al no existir tratamiento efectivo en todas las unidades seleccionadas, pueden existir diferencias preexistentes en los individuos que hayan causado el no terminar de forma efectiva el tratamiento. Así, al calcular el modelo planteado anteriormente, se obtendría el efecto causal de la política entre los efectivamente tratados, y se compararía con el grupo nunca tratado, algo similar a un LATE. Lo que propone el asistente se refiere a un ITT, intento al tratamiento, lo que mide el efecto de haber sido seleccionado para el tratamiento, independientemente de si la unidad fue efectivamente tratada. Esto ayuda a evadir los posibles sesgos que llevaron a los individuos a no terminar el tratamiento de forma efectiva, y logra capturar un efecto de la política sobre la variable de interés.

- b) Si la idea del asistente es correcta, ¿sería posible estimar un efecto consistentemente que cumpla los supuestos necesarios? Si su respuesta es afirmativa, proponga la ecuación que captura dicho efecto.

Al seguir proponiendo un método de diferencias en diferencias, se hace necesario que se siga cumpliendo el supuesto de tendencias paralelas. En este caso, sería necesario realizar nuevamente una tabla de balanceo muestral para determinar si se sigue manteniendo el supuesto aún sobre la asignación y no sobre el tratamiento. A continuación, los resultados.

Tabla 1.6 - Balanceo Muestral grupo 1 en periodo 0

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.963	9.997	-0.034	0.283
Estrato	1.981	2.002	-0.021	0.346
Poblacion del municipio	30.206	30.113	0.092	0.454
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Poblacion en miles				

Tabla 1.7 - Balanceo Muestral grupo 1 en periodo 1

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	9.963	9.997	-0.034	0.283
Estrato	1.981	2.002	-0.021	0.346
Poblacion del municipio	30.206	30.113	0.092	0.454
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Poblacion en miles				

Tabla 1.8 - Balanceo Muestral grupo 2 en periodo 0

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	10.017	10.045	-0.028	0.369
Estrato	1.981	1.994	-0.013	0.564
Poblacion del municipio	30.193	30.043	0.150	0.225
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Poblacion en miles				

Tabla 1.9 - Balanceo Muestral grupo 2 en periodo 1

	Y1_mean	Y0_mean	Y1-Y0	p-value
Distancia centro medico	10.017	10.045	-0.028	0.369
Estrato	1.981	1.994	-0.013	0.564
Poblacion del municipio	30.193	30.043	0.150	0.225
*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$ Poblacion en miles				

En este caso, se observa un mismo comportamiento al observado en la tabla de balanceo muestral realizada con tratamiento efectivo, por lo que es plausible creer que si existieran más periodos de tiempo se podría efectuar mejores pruebas en favor de tendencias paralelas. En este caso, debido que ya no se puede hablar de asistencia efectiva, existe la posibilidad de que la decisión de usar o no métodos anticonceptivos pueda depender de algún factor externo al tratamiento, o en este caso a la intención de tratamiento. Por ende, es necesario ser cauteloso con el resultado obtenido. Confiando en que el planteamiento es correcto, el modelo planteado sería el siguiente.

Primera especificación (Grupo 1)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 P_{sola_i} + \beta_3 Post_t * P_{sola_i} + \beta X_{i,t} + e_i$$

Segunda especificación (Grupo 2)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 P_{pareja_i} + \beta_3 Post_t * P_{pareja_i} + \beta X_{i,t} + e_i$$

Tercera especificación (Pooled)

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 P_{pooled_i} + \beta_3 Post_t * P_{pooled_i} + \beta X_{i,t} + e_i$$

- c) Estime el efecto para cada uno de los tres grupos de interés (grupo 1, el grupo 2 y efecto combinado) usando la base de datos.

Tabla 1.10 - Efecto de asignacion a tratamiento

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento = 1	0.000733 (0.00740)	-0.00840 (0.00693)	-0.00108 (0.00360)
Asignado al tratamiento en el grupo 1 = 1	0.0732*** (0.00788)		
Post = 1, Asignacion grupo 1 = 1	0.254*** (0.0101)		
Asignado al tratamiento en el grupo 2 = 1		0.0585*** (0.00801)	
Post = 1, Asignacion grupo 2 = 1		0.259*** (0.0106)	

Asignado al tratamiento en el grupo 1 o 2 = 1			0.0681*** (0.00497)
Post = 1, Asignacion grupo 1 o 2 = 1			0.253*** (0.00646)
Constant	0.354*** (0.00511)	0.361*** (0.00524)	0.355*** (0.00253)
Observations	33,042	33,272	100,000
R-squared	0.073	0.067	0.056

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Sin embargo y dado que se busca capturar es el efecto del tratamiento efectivo, es decir, haber asistido a más de tres sesiones. Usted se aventura a una metodología que involucra utilizar un diferencias en diferencias instrumentado (DID-IV), esta especificación aprovecha explícitamente la aleatoriedad exógena en una tendencia de exposición para estimar el efecto del tratamiento promedio. Para conocer más sobre su implementación le recomendaron leer el artículo de *Hudson, Hull & Liebersohn (2017)*.

3. Dado el anterior contexto responda las siguientes preguntas, para ello utilice el artículo citado anteriormente.

a) Planteen el modelo de regresión a estimar.

De acuerdo con Hudson, Hull & Liebersohn (2017), la ecuación estructural, la primera etapa y la ecuación reducida son las siguientes.

Primera especificación (Grupo 1)

Ecuación estructural:

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{sola_i} + e_i$$

Primera etapa:

$$D_{sola_i} = \alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{sola_i} + \eta_i$$

Forma reducida:

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 (\alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{sola_i} + \eta_i) + e_i$$

$$\begin{aligned} Planificación_{i,t} \\ &= \beta_0 + \beta_1 Post_t + \beta_2 \alpha_0 + \beta_2 \alpha_1 Post_t + \beta_2 \alpha_2 Post_t P_{sola_i} + \beta_2 \eta_i \\ &+ e_i \end{aligned}$$

$$Planificación_{i,t} = \gamma_0 + \gamma_1 Post_t + \phi Post_t P_{sola_i} + \mu_i$$

Donde $\gamma_0 = \beta_0 + \beta_2 \alpha_0$, $\gamma_1 = \beta_1 + \beta_2 \alpha_1$, $\phi = \beta_2 \alpha_2$ y $\mu_i = \beta_2 \eta_i + e_i$

Segunda especificación (Grupo 2)

Ecuación estructural:

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{pareja_i} + e_i$$

Primera etapa:

$$D_{sola_i} = \alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{pareja_i} + \eta_i$$

Forma reducida:

$$\begin{aligned} Planificación_{i,t} \\ = \beta_0 + \beta_1 Post_t + \beta_2 (\alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{pareja_i} + \eta_i) + e_i \end{aligned}$$

$$\begin{aligned} Planificación_{i,t} \\ = \beta_0 + \beta_1 Post_t + \beta_2 \alpha_0 + \beta_2 \alpha_1 Post_t + \beta_2 \alpha_2 Post_t P_{pareja_i} + \beta_2 \eta_i \\ + e_i \end{aligned}$$

$$Planificación_{i,t} = \gamma_0 + \gamma_1 Post_t + \phi Post_t P_{pareja_i} + \mu_i$$

Donde $\gamma_0 = \beta_0 + \beta_2 \alpha_0$, $\gamma_1 = \beta_1 + \beta_2 \alpha_1$, $\phi = \beta_2 \alpha_2$ y $\mu_i = \beta_2 \eta_i + e_i$

Tercera especificación (Pooled)

Ecuación estructural:

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 D_{pooled_i} + e_i$$

Primera etapa:

$$D_{sola_i} = \alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{pooled_i} + \eta_i$$

Forma reducida:

$$\begin{aligned} Planificación_{i,t} \\ = \beta_0 + \beta_1 Post_t + \beta_2 (\alpha_0 + \alpha_1 Post_t + \alpha_2 Post_t P_{pooled_i} + \eta_i) + e_i \end{aligned}$$

$$Planificación_{i,t} = \beta_0 + \beta_1 Post_t + \beta_2 \alpha_0 + \beta_2 \alpha_1 Post_t + \beta_2 \alpha_2 Post_t P_{pooled_i} + \beta_2 \eta_i + e_i$$

$$Planificación_{i,t} = \gamma_0 + \gamma_1 Post_t + \phi Post_t P_{pooled_i} + \mu_i$$

Donde $\gamma_0 = \beta_0 + \beta_2 \alpha_0$, $\gamma_1 = \beta_1 + \beta_2 \alpha_1$, $\phi = \beta_2 \alpha_2$ y $\mu_i = \beta_2 \eta_i + e_i$

b) ¿Cuáles son sus supuestos de identificación? ¿Son estos plausibles?

Como hablamos de un DDIV, es necesario que se cumplan 4 supuestos principales según Hudson, Hull & Liebersohn (2017). La restricción de exclusión, que en este caso sería que el instrumento afecte a la variable dependiente únicamente a través del tratamiento, y que el instrumento solo afecte a las unidades dependientes contemporáneas; esto es decir, que instrumento en periodo t , no afecte los resultados potenciales del periodo $t-1$. Adicional a estos, se requiere también la existencia de tendencias paralelas en la senda de crecimiento de los resultados potenciales de la variable dependiente. Finalmente, el supuesto de monotonicidad, el cual establece que el efecto del tratamiento solo se observará en el periodo 1. En términos de resultados potenciales: Sea S_{it} un instrumento dicotómico, $P(S_{i1}^1 \geq S_{i1}^0) > 1$.

Así mismo, pensando en las características de un buen instrumento, usar la asignación al tratamiento cumple con las características necesarias. El instrumento es independiente, debido a que se realizó una aleatorización, la cual se corroboró como bien hecha por medio de las tablas de balanceo muestral, y no responde a las características previas de los grupos de tratamiento y de control. En cuanto a relevancia, se espera que este se cumpla debido que un tratamiento efectivo es función de la asignación, quien no es asignado a tratamiento no podrá ser tratado efectivo como alta probabilidad.

Es plausible que se cumplan estos supuestos como resultado de la naturaleza del instrumento y del tratamiento. Por el lado de la restricción de exclusión, debido a que el instrumento es la asignación a alguno de los grupos de tratamiento, la decisión del individuo solo será afectada a través del tratamiento (asistencia a 3 sesiones). Como se intentó probar en el literal a y b, es plausible suponer la existencia de tendencias paralelas. Por construcción del instrumento, se cumple monotonicidad, y debido a que el tratamiento efectivo es función de la asignación a grupos de tratamiento, se espera que se cumpla la condición de relevancia. Finalmente, la aleatorización en la asignación del tratamiento hace plausible suponer que el instrumento es independiente.

- c) ¿Qué coeficiente representa su efecto de interés bajo dichos supuestos? Interprete.

El coeficiente de interés corresponde a los correspondientes en la ecuación $\phi = \beta_2 \alpha_2$. Así, se busca tener el valor del coeficiente α_2 por medio de la estimación de la primera etapa, y el coeficiente ϕ por medio de la estimación de la ecuación reducida. Luego, el parámetro de interés se observa de la siguiente forma.

$$\beta_2 = \frac{\phi}{\alpha_2} = \frac{E[Y_{i1} - Y_{i0}|Z_i = 1] - E[Y_{i1} - Y_{i0}|Z_i = 0]}{E[S_{i1} - S_{i0}|Z_i = 1] - E[S_{i1} - S_{i0}|Z_i = 0]}$$

En términos de los parámetros del modelo,

$$\beta = \frac{E[Planificación_{i1} - Planificación_{i0}|P_{j_i} = 1] - E[Planificación_{i1} - Planificación_{i0}|P_{j_i} = 0]}{E[D_{j_{i1}} - D_{j_{i0}}|P_{j_i} = 1] - E[D_{j_{i1}} - D_{j_{i0}}|P_{j_i} = 0]}$$

Donde $j \in (sola, pareja, pooled)$.

Hudson, Hull & Liebersohn (2017) demuestran que el parámetro de interés β se puede expresar de la siguiente forma.

$$\beta = \sum_{s>0} \omega_s E[Y_{i1}^s - Y_{i1}^{s-1} | S_{i1}^1 \geq s > S_{i0}^0]$$

Donde

$$\omega_s = \frac{P(S_{i1}^1 \geq s > S_{i0}^0)}{\sum_{r>0} P(S_{i1}^1 \geq r > S_{i0}^0)}$$

Luego, el parámetro captura la respuesta causal promedio (average causal response), que no es más que el efecto promedio por pesos (ω) de la respuesta causal a cambios en una unidad en el tratamiento en el periodo 1. En este caso en particular, el coeficiente se interpreta como el efecto causal del programa dato que se ofreció el programa.

- d) Implemente la metodología anterior planteada en Stata. Para ello utilice el paquete *xtivreg*, que le permite realizar regresión de variables instrumentales para datos panel. En una tabla para tipo de grupo (grupo 1, grupo 2 y combinado) presente los resultados de la primera etapa, segunda etapa y la forma reducida. Interprete.

Tabla 1.11 - Primera Etapa

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento	-0.258*** (0.00346)	-0.248*** (0.00530)	-0.168*** (0.00236)
Post = 1, Asignacion grupo 1 = 1	0.511*** (0.00503)		
Post = 1, Asignacion grupo 2 = 1		0.497*** (0.00612)	
Post = 1, Asignacion grupo 1 o 2 = 1			0.504*** (0.00317)
Constant	0.258*** (0.00346)	0.248*** (0.00306)	0.168*** (0.00149)
Observations	33,042	33,272	100,000
R-squared	0.170	0.165	0.202

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Tabla 1.13 - Segunda Etapa

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Tratamiento efectivo en el grupo 1	0.392*** (0.0101)		
Tratamiento efectivo en el grupo 2		0.378*** (0.0103)	
Tratamiento efectivo en el grupo 1 o 2			0.387*** (0.00628)
Constant	0.354*** (0.00365)	0.356*** (0.00363)	0.354*** (0.00182)
Observations	33,042	33,272	100,000
Number of id_hogar	16,521	16,636	50,000

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

Tabla 1.12 - Forma Reducida

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento = 1	0.000733	-0.00840	-0.00108

	(0.00749)	(0.00745)	(0.00371)
Asignado al tratamiento en el grupo 1 = 1	0.0732***		
	(0.00746)		
Post = 1, Asignacion grupo 1 = 1	0.254***		
	(0.0106)		
Asignado al tratamiento en el grupo 2 = 1		0.0585***	
		(0.00745)	
Post = 1, Asignacion grupo 2 = 1		0.259***	
		(0.0105)	
Asignado al tratamiento en el grupo 1 o 2 = 1			0.0681***
			(0.00455)
Post = 1, Asignacion grupo 1 o 2 = 1			0.253***
			(0.00643)
Constant	0.354***	0.361***	0.355***
	(0.00530)	(0.00526)	(0.00262)
Observations	33,042	33,272	100,000
R-squared	0.073	0.067	0.056

Standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

e) Compare con los resultados obtenidos con aquellos en el inciso a) y b).

En este caso, se encuentran efectos positivos en los 3 grupos: 0.392, 0.378 y 0.354 puntos porcentuales respectivamente. Así, se observa un notable orden en el tamaño de los coeficientes de acuerdo con cuál es la especificación que se observe. La primera especificación, tratamiento efectivo, ofrece coeficientes de magnitud 0.505, 0.520 y 0.510 respectivamente. La segunda, asignación a tratamiento, 0.254, 0.259 y 0.253 respectivamente. Así las cosas, se observa que, en cuanto a tamaño, la segunda especificación ofrece los coeficientes de menor magnitud, seguida por la tercera, y finalmente la primera especificación con los coeficiente de mayor magnitud.

Lo anterior es reflejo del problema de identificación que se enfrenta. Debido a que la decisión de asistir a todas las charlas es endógena, se espera que el efecto capturado no refleje el verdadero efecto del programa debido al sesgo existente. Por ello, tiene sentido que el coeficiente asociado a esta especificación sea el mayor. Por el lado de la segunda especificación, no se está estimando un ATE, sino un ITT. Esto significa que no es el efecto causal de aplicar el programa, sino de ofrecer el programa. Por ende, no se captura el efecto real que el programa tuvo sobre las decisiones de uso de métodos anticonceptivos. Finalmente, el método de IV permite identificar el efecto del tratamiento gracias a que se ofreció el tratamiento. Así, se encuentra un efecto causal del tratamiento sobre los que se

ofreció el tratamiento (LATE), y se determina que este tuvo efectos positivos sobre las decisiones de uso de métodos anticonceptivos.

Una de las participantes del programa le hace conocer que factiblemente el efecto de las capacitaciones puede cambiar dependiendo del sexo del capacitador (i.e. quien imparte la información). Esto debido a que existen sesgos de género marcados y sistemáticos que puedan ser decisivos para determinar el efecto del tratamiento en un contexto en donde la sexualidad y planificación femenina siguen siendo un tema tabú. Es así como usted se plantea explorar la heterogeneidad en los efectos del tratamiento según el sexo del capacitador. Suponga que se cumple tendencias paralelas entre el grupo de control y tratamiento en los tres casos de tratamiento (pooled, grupo 1 y grupo 2).

3. Para medir lo anterior usted plantea una interacción entre el sexo del capacitador con el estado y el tiempo del tratamiento, es decir, una estimación de triple diferencia. Para ello:

a) Planteen el modelo de regresión a estimar.

Primera especificación (Grupo 1)

$$\begin{aligned} \text{Planificación}_{i,t} &= \beta_0 + \beta_1 \text{Post}_t + \beta_2 P_{\text{sola}_i} + \beta_3 \text{Sexo}_{\text{capacitador}_i} + \beta_4 \text{Post}_t P_{\text{sola}_i} \\ &+ \beta_5 \text{Post}_t \text{Sexo}_{\text{capacitador}_i} + \beta_6 P_{\text{sola}_i} \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_7 \text{Post}_t P_{\text{sola}_i} \text{Sexo}_{\text{capacitador}_i} + \beta X_{i,t} + e_i \end{aligned}$$

Segunda especificación (Grupo 2)

$$\begin{aligned} \text{Planificación}_{i,t} &= \beta_0 + \beta_1 \text{Post}_t + \beta_2 P_{\text{pareja}_i} + \beta_3 \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_4 \text{Post}_t P_{\text{pareja}_i} + \beta_5 \text{Post}_t \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_6 P_{\text{pareja}_i} \text{Sexo}_{\text{capacitador}_i} + \beta_7 \text{Post}_t * P_{\text{pareja}_i} \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta X_{i,t} + e_i \end{aligned}$$

Tercera especificación (Pooled)

$$\begin{aligned} \text{Planificación}_{i,t} &= \beta_0 + \beta_1 \text{Post}_t + \beta_2 P_{\text{pooled}_i} + \beta_3 \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_4 \text{Post}_t P_{\text{pooled}_i} + \beta_5 \text{Post}_t \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta_6 P_{\text{pooled}_i} \text{Sexo}_{\text{capacitador}_i} + \beta_7 \text{Post}_t P_{\text{pooled}_i} \text{Sexo}_{\text{capacitador}_i} \\ &+ \beta X_{i,t} + e_i \end{aligned}$$

b) ¿Cuáles son sus supuestos de identificación? ¿Son estos plausibles?

Nuevamente, se debe mantener el supuesto inicial de tendencias paralelas para poder asegurar que se puede capturar el efecto causal del tratamiento. En este caso en particular, debido a la triple interacción, las tendencias deben mantenerse entre los grupos de control y tratamiento, y adicionalmente en el sexo de los capacitadores. Esto es, las tendencias del uso de métodos anticonceptivos entre individuos, al mirar los grupos que fueron atendidos por encuestador hombre y encuestadora mujer, deben ser paralelas. Así, el supuesto se hace más exigente al requerir tendencias paralelas no solo entre grupos de tratamiento y control, sino también entre la característica a observar.

En este caso, existe alta probabilidad de que se cumpla el supuesto como resultado de que las decisiones de uso de anticonceptivos son independientes de la selección del sexo del encuestador. Adicionalmente, debido a que la proporción de encuestadores hombre y mujer es casi del 50/50, la aleatorización en la asignación del encuestador al hogar sería suficiente para garantizar la existencia del supuesto. Y si no se hubiese asignado por aleatorización, aun así, el supuesto se mantendría.

c) ¿Qué coeficiente representa su efecto de interés bajo dichos supuestos? Interprete.

En este caso, el coeficiente que captura el efecto de interés corresponde al β_7 en las 3 ecuaciones planteadas. Este coeficiente, captura el efecto de haber sido tratado, y de haber sido encuestado por una mujer, y se interpretaría como el efecto que tiene sobre las unidades tratadas el sexo del encuestador. Así, si existen diferencias significativas, se podría concluir que existen sesgos sociales a causa del tabú de la sexualidad respecto a las mujeres, y poder entender la verdadera incidencia del tratamiento sobre el uso de métodos anticonceptivos. En este caso, el coeficiente se interpreta como el efecto de haber tenido una encuestadora mujer respecto a un encuestador hombre sobre los efectos del tratamiento. Así, se observa si el sexo del encuestador influye sobre el tratamiento.

d) Estime el efecto para cada uno de los tres grupos de interés (grupo 1, el grupo 2 y efecto combinado) usando la base de datos. *Planificación_DID.dta* e interprete. Inserte el código utilizado.

Tabla 1.14 - Efecto del sexo del encuestador

VARIABLES	(1) Grupo 1	(2) Grupo 2	(3) Efecto combinado
Dummy del periodo post tratamiento = 1	-0.00596 (0.00956)	0.00528 (0.0110)	0.00108 (0.00513)
Sexo del capacitador: 1.mujer, 0.hombre = 1	-0.00909 (0.00939)	0.000866 (0.0111)	0.00131 (0.00512)
Post = 1, Sexo capacitador = 1	0.0132 (0.0138)	-0.0274* (0.0162)	-0.00433 (0.00771)
Asignado al tratamiento en el grupo 1 = 1	0.0660*** (0.0112)		
Post = 1, Asignacion grupo 1 = 1	0.269*** (0.0141)		
Sexo capacitador = 1, Asignacion grupo 1 = 1	0.0143 (0.0141)		
Post = 1, Asignacion grupo 1 = 1, Sexo capacitador = 1	-0.0303 (0.0221)		
Asignado al tratamiento en el grupo 2 = 1		0.0549*** (0.0121)	
Post = 1, Asignacion grupo 2 = 1		0.250*** (0.0170)	
Sexo capacitador = 1, Asignacion grupo 2 = 1		0.00696 (0.0167)	
Post = 1, Asignacion grupo 2 = 1, Sexo capacitador = 1		0.0189 (0.0241)	
Asignado al tratamiento en el grupo 1 o 2 = 1			0.0655*** (0.00709)
Post = 1, Asignacion grupo 1 o 2 = 1			0.258*** (0.00947)
Sexo capacitador = 1, Asignacion grupo 1 o 2 = 1			0.00509 (0.00945)
Post = 1, Asignacion grupo 1 o 2 = 1, Sexo capacitador = 1			-0.00853 (0.0139)
Constant	0.358*** (0.00747)	0.360*** (0.00788)	0.354*** (0.00364)
Observations	33,042	33,272	100,000
R-squared	0.073	0.068	0.056

Robust standard errors in parentheses

*** p<0.01, ** p<0.05, * p<0.1

En este caso, se observa que los coeficientes de interés para los 3 grupos el coeficiente de interés no es significativo. Sin embargo, es necesario analizar el signo del coeficiente.

Para el grupo 1, tratamiento a la mujer sola, se observa que si el encuestador es de sexo femenino, la probabilidad de uso de métodos anticonceptivos cae en 0,03 puntos porcentuales. Este resultado resulta muy contraintuitivo, en el sentido de que se esperaría que cuando la mujer es la única responsable en el cuidado de la salud reproductiva, una encuestadora mujer facilitaría el recibimiento de la información, ponerse en su lugar y ofrecer información convincente respecto a los beneficios de los métodos anticonceptivos.

Por el lado del grupo 2, donde se ofrecieron capacitaciones a la pareja, se observa un signo positivo, lo que significa que la probabilidad de uso de métodos anticonceptivos incrementó en 0.024 puntos porcentuales. Se espera que mujeres con acompañamiento de su pareja en la decisión de uso de métodos anticonceptivos sean más propensas a su uso, y que la encuesta sea realizada por una mujer permitiría convencer al hombre de los beneficios de los métodos anticonceptivos de forma más efectiva; una mujer hablando de la necesidad de los anticonceptivos para el bienestar sexual y reproductivo de las mujeres.

Finalmente, el grupo 3 tiene un coeficiente negativo, y significa que el haber recibido a una encuestadora mujer significó una reducción de la probabilidad de uso de métodos anticonceptivos en 0.0085 puntos porcentuales.

Codigo empleado:

```
qui reg uso post##sexo##P_s if Tipo==1, vce(cluster id_mun)
outreg2 using "${out}\\Punto_1_d_iv.doc", replace ctitle("Grupo 1")
qui reg uso post##sexo##P_pa if Tipo==2, vce(cluster id_mun)
outreg2 using "${out}\\Punto_1_d_iv.doc", ctitle("Grupo 2")
qui reg uso post##sexo##P_po, vce(cluster id_mun)
outreg2 using "${out}\\Punto_1_d_iv.doc", label ctitle("Efecto combinado") ///
        title("Tabla 1.14 - Efecto del sexo del encuestador")
```

Segundo ejercicio

El método de diferencias en diferencias es quizá el método más utilizado en la última década por los economistas en las investigaciones académicas empíricas. Las numerosas aplicaciones condujeron a un interés de desarrollar extensiones a esta metodología. Quizá el desarrollo más importante en esta índole llegó con el método de "estudio de eventos", el cual permitió generalizar diferencias en diferencias a escenarios donde las unidades recibían el tratamiento en distintos momentos del tiempo.

Las primeras aproximaciones de los economistas con los estudios de eventos se basaron en la intuición que se tenían acerca de la metodología de diferencias en diferencias, sin mucho cuidado acerca de si esta generalización era, en principio, siquiera posible. Consecuentemente, los econométricos teóricos empezaron a encontrar varios problemas asociados a la manera en que se utilizaba comúnmente esta metodología, concluyendo que existían principalmente dos problemas: (1) contrario a otras metodologías, no era muy claro el efecto causal que identifica los estudios de eventos; y (2), la estimación por mínimos cuadrados ordinarios de un modelo de panel lineal con efectos fijos, conocida como “two-way fixed effect estimation (TWFE)”, es inadecuada, en tanto no captura un efecto causal salvo para casos muy restrictivos para el análisis empírico.

Quizá el primer trabajo en poner en discusión las falencias del TWFE en estudios de evento es el trabajo de Goodman-Bacon (2021). En breve, este trabajo tiene dos grandes contribuciones. La primera, descompone el estimador de mínimos cuadrados ordinarios que resulta de estimar un TWFE de manera que se puede entender de manera intuitiva lo que se está estimando. Lo segundo, muestra que incluso bajo el cumplimiento de tendencias paralelas, el estimador no recupera un efecto causal en presencia de heterogeneidad en los efectos de tratamiento.

El objetivo de este punto es entender mejor el planteamiento que se hace en Goodman-Bacon (2021) y de paso, repasar algunos temas que vimos en clase acerca de los estudios de eventos.

a) Corran el Do File Sim1 Do que simula una pequeña “base de datos” con la cual nos aproximaremos al problema.

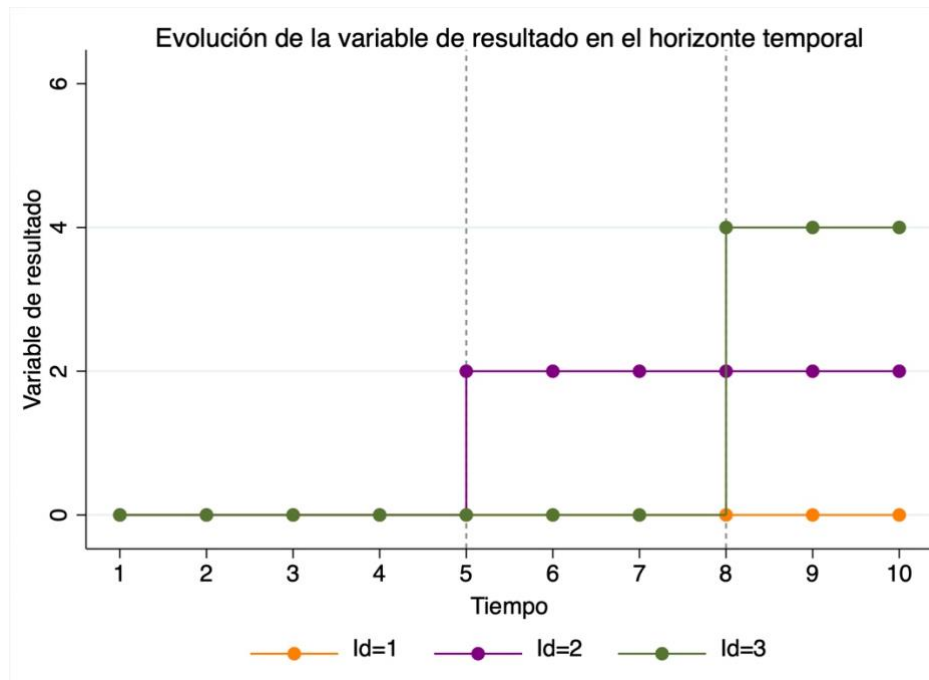
i) Describan en palabras qué hace el código

Solución:

El código crea y organiza un panel de datos. En particular, crea información con respecto al estado del tratamiento de las unidades de análisis (variable D) y con respecto a una variable de resultado (variable Y) para diez períodos de tiempo distintos. Adicionalmente, lo organiza con el fin de que se pueda observar fácilmente si las unidades de análisis fueron tratadas en algún punto del horizonte temporal y en dado caso ver el efecto de dicho tratamiento sobre la variable de resultado.

ii) Ilustren en una gráfica cómo varía la variable dependiente (con nombre Y en la base de datos) en función del tiempo para cada unidad. Ilustren con una línea punteada el momento donde cada unidad es tratada. ¿Cuántas cohortes hay? Describa cada una brevemente.

Solución:



Como es posible evidenciar en la gráfica presentada anteriormente, en la base de datos suministrada hay 3 cohortes distintas: una que no recibe el tratamiento en ningún punto del horizonte temporal, otra que recibe el tratamiento en el período $t=5$ ($id=2$) y una última que recibe el tratamiento en el momento $t=8$ ($id=3$).

iii) Describan matemáticamente $Y_{i,t}(g)$, donde g hace referencia a la cohorte.
¿Cuál es el ATT para cada cohorte tratada?

Solución:

Solución 2a.iii:

primero, el resultado potencial está dado por:

$$y_{it}(g) = \begin{cases} \eta_i + \phi_t + \varepsilon_{it} & \text{cuando } t < G_i \\ \eta_i + \phi_t + \tau_{i,t}(g) + \varepsilon_{it} & t \geq G_i \end{cases}$$

En donde:

η_i, ϕ_t son efectos fijos de unidad y tiempo respectivamente

$G_i = g$ es el momento en el que se trata a la cohorte por primera vez.

De esta forma, los ATT para $g=2$ y $g=3$ están dados por:

$$ATT(2, t) = E[y_{it}(g) - y_{it}(\infty) | G_i = 5]$$

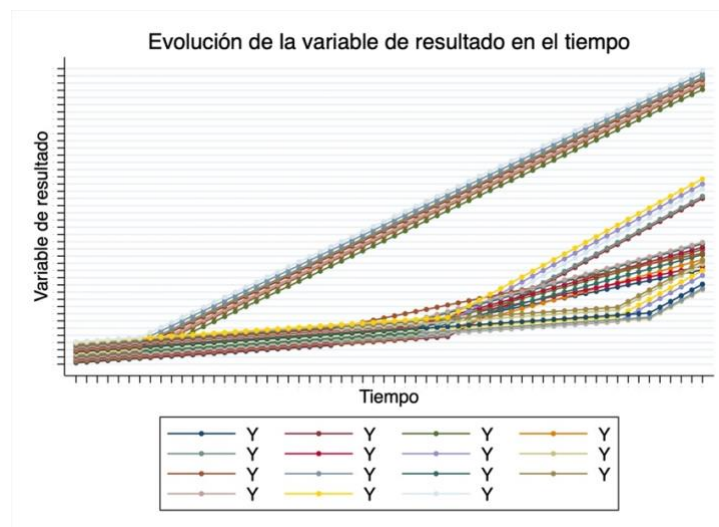
$$ATT(3, t) = E[y_{it}(g) - y_{it}(\infty) | G_i = 8]$$

Nota: $Y_{it}(\infty)$ denota el resultado potencial de nunca ser tratado.

d) Para la nueva muestra simulada

- i) Grafique en una única gráfica la variable dependiente de cada unidad en el tiempo.

Solución:



Dada la gráfica anterior y el código proporcionado por el documento *Sim_Do2*, se puede concluir que hay un total de 6 cohortes en la base de datos. Adicionalmente, se puede ver que la evolución de la variable de resultado antes del tratamiento es de naturaleza paralela para los grupos y que a medida que las cohortes se van exponiendo al tratamiento estas cambian su pendiente. En particular, parece que el tratamiento hace que se aumente la pendiente de la variable Y para las cohortes, lo cual indica un posible efecto positivo del tratamiento.

- ii) *Estime el modelo por TWFE. ¿Es coherente el signo encontrado con lo que se espera teóricamente? De acuerdo a lo visto anteriormente, ¿qué puede estar causando este problema?*

Solución:

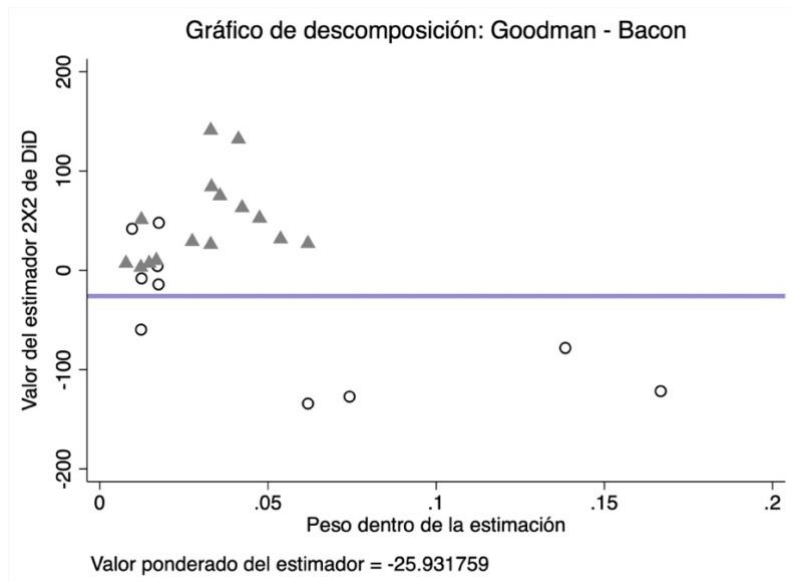
Y	Coeficiente	Errores estándar	P>t
D	-25.932	3.375	0.000
Constante	114.935	1.997	0.000

Tras estimar el modelo por TWFE, es posible observar que el tratamiento efectivamente es estadísticamente significativo. Sin embargo, el efecto de dicho tratamiento sobre la variable de resultado resulta negativo, lo cual contradice el argumento gráfico desarrollado en el punto anterior.

En ese sentido, una posible causa de este comportamiento contraintuitivo del tratamiento sobre la variable de resultado es que, dentro del estimador de TWFE, las comparaciones que indican un efecto negativo del tratamiento están teniendo un mayor peso en la ponderación.

- e) *Para la nueva muestra simulada, presente de manera gráfica la descomposición de Goodman-Bacon. Interprete lo que encontró en el inciso d.ii) a la luz de esta nueva evidencia.*

Solución:



Según la gráfica de descomposición de Goodman – Bacon, en la cual se puede ver el peso de cada comparación 2X2 de diferencias en diferencias dentro de la estimación, así como su respectivo valor, es posible corroborar la hipótesis planteada en el punto anterior. En particular, es posible observar que el valor ponderado del estimador es prácticamente el mismo. Sin embargo, la gráfica indica que hay unas pocas comparaciones que tienen un valor considerablemente negativo y que cuentan con un peso relativo muy grande dentro de la estimación. De esta forma, tras desagregar el estimador de TWFE mediante la metodología de Goodman – Bacon, es posible atribuir el signo negativo del estimador a unas pocas comparaciones con un peso alto dentro de la estimación y que a su vez tienen signo negativo.

f) Usando el estimador de Callaway & Sant’Anna, estime los efectos causales dinámicos del tratamiento en la muestra simulada (i.e, el efecto promedio del tratamiento τ periodos antes o después de que una cohorte es tratada).

¿La estimación parece ser consistente con efectos causales de la simulación? ¿La estimación parece ser consistente con la información que tenemos acerca del cumplimiento de tendencias paralelas?

Solución:

Tercer ejercicio

El objetivo de este ejercicio es indagar sobre las distintas estructuras de errores estándar que comúnmente se emplean al estimar un modelo de regresión lineal. En particular, queremos evaluar cómo distintas metodologías para la estimación de los errores estándar se comportan en diversos contextos que pueden hacer que uno u otro enfoque resulte más apropiado. El análisis se centrará en estimar y efectuar inferencia estadística sobre el siguiente modelo:

$$Y_i = \beta * X_i + \varepsilon_i, \quad \beta = 10 \quad (0.1)$$

Donde inicialmente los X_i son independientes e idénticamente distribuidos con distribución normal estándar, mientras que en los siguientes puntos variaremos la estructura de la distribución de los términos de error ε_i .

1. Inicialmente, considere un escenario homocedástico en el que $\varepsilon_i \sim \mathcal{N}(0,1)$ iid. En este contexto, estudiaremos los efectos de suponer erróneamente un modelo heterocedástico (pero sin autocorrelación).

- a) En primer lugar, consideraremos dos posibles estimadores de la varianza del vector de $\hat{\beta}$, que viene dada por

$$\text{var}(\hat{\beta}) = (X'X)^{-1}X'\Sigma X(X'X)^{-1}$$

Donde Σ es la matriz de varianzas y covarianzas del vector de ε_i 's y X es la matriz de variables explicativas.

Bajo errores homocedásticos, se supone $\Sigma = \sigma^2 I$, de manera que solo es necesario estimar σ^2 , lo cual se hace por medio de la fórmula

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^n \hat{\varepsilon}_i^2}{n - k}$$

donde k es el número de parámetros y $\hat{\varepsilon}_i$ son los residuales de la regresión.

En cambio, bajo errores heterocedásticos, se hace necesario estimar directamente la matriz de varianzas y covarianzas Σ . Una alternativa usual son los errores tipo Huber-White o *hc1* (*heteroskedasticity consistent 1*, los cuales son la opción predeterminada de Stata al agregar la opción “*robust*”). En particular este estimador toma la forma:

$$\hat{\text{var}}(\hat{\beta})_{hc1} = (X'X)^{-1}X'\text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2)X(X'X)^{-1}$$

donde el estimador de la matriz Σ es la matriz diagonal compuesta por los residuales del modelo al cuadrado ($\hat{\Sigma}_{hc1} = \text{diag}(\hat{\varepsilon}_1^2, \dots, \hat{\varepsilon}_n^2)$). En otras palabras, el estimador puntual del elemento Σ_{ii} es $\hat{\varepsilon}_i^2$.

- a) Simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ acorde al modelo (1). Para cada una, estimen el modelo (1) y calculen el intervalo de confianza del 95 % para $\hat{\beta}$ empleando errores estándar usuales y *hc1*. Para cada caso, calcule el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero. Presenten e interpreten sus resultados.

Tabla 1: 1000 observaciones

VARIABLES	(1)	(2)
	Errores estandar	hc1
% de éxito del parámetro en intervalo	94.7 %	94.3 %
Número de observaciones	1000	1000

Los intervalos de confianza y los errores estándar son muy parecidos, por tanto, se podría decir que no se tienen unos cambios muy importantes entre los errores estandar y los robustos a heterocedasticidad cuando se tiene una muestra homocedástica. Pues el aumento en la precisión del parámetro en el intervalo es muy pequeño.

- b) Los errores $hc1$ resultan problemáticos cuando se cuenta con muestras pequeñas por dos razones principales. Por una parte, aunque el estimador obtenido de Σ es consistente, su convergencia es relativamente lenta. Por otra parte, se ha probado que es especialmente sensible a la presencia de datos atípicos. Para visualizar estos problemas, repitan el procedimiento del inciso anterior, esta vez con muestras de tamaño 10. ¿Cómo se comparan sus resultados?

Tabla 2: 10 observaciones

VARIABLES	(1) Errores estandar	(2) hc1
% de éxito del parámetro en intervalo	100 %	90 %
Número de observaciones	10	10

Se presentan cambios mayores en el porcentaje de éxito del parámetro, en este caso se puede ver que al usar errores robustos se tienen menos casos de éxito respecto a los errores estándar normales. Esto puede dar cuenta de las falencias de los errores $hc1$ con algunos datos atípicos que toman más relevancia cuando se tiene una muestra pequeña.

- c) Una solución al problema de muestra pequeña al que se enfrentan los errores $hc1$ fue propuesta por Davidson y MacKinnon (1993), quienes sugieren usar un estimador alternativo que contiene un factor de corrección para muestras finitas¹ $\hat{\Sigma}_{hc3} = \text{diag}(\hat{\epsilon}_1^2/(1 - h_{11})^2, \dots, \hat{\epsilon}_n^2/(1 - h_{nn})^2)$, en donde h_{ii} se refiere al i -ésimo elemento de la diagonal de la matriz $X(X'X)^{-1}X'$. La intuición detrás de esta corrección es que cuantifica, en cierta medida, qué tan influyente es una observación a la hora de calcular $\hat{\beta}$. De esta manera, este estimador se hace más resistente a datos atípicos². Repita el procedimiento del inciso b, pero esta vez emplee errores convencionales y $hc3$. Reporte sus resultados y compare.

Tabla 3: 10 observaciones

VARIABLES	(1) Errores estandar	(2) hc3
% de éxito del parámetro en intervalo	94.7 %	93.8 %
Número de observaciones	1000	1000

En este caso los errores estándar robustos $hc3$ presentan una mayor variación en el porcentaje de éxito respecto al de los errores $hc1$, en este caso se tiene una diferencia de casi un punto porcentual, mientras que con los errores $hc1$ no se veían cambios significativos.

2. Suponga ahora que $\varepsilon_i \sim \mathcal{N}(0, X_i^2)$ independientes.

- a) Simulen 1000 muestras $\{X_i, Y_i\}_{i=1}^{1000}$ acorde al modelo (1). Para cada una, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ empleando errores estándar usuales, $hc1$ y $hc3$. Para cada caso, calcule el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero. Presenten e interpreten sus resultados. Comparen con lo obtenido en el punto 1.

Tabla 4: 1000 observaciones

VARIABLES	(1) Errores estandar	(2) hc1	(3) hc3
% de éxito del parámetro en intervalo	62.4 %	94.4 %	94.7 %
Número de observaciones	1000	1000	1000

En el caso de esta distribución, se tienen mejoras muy considerables con errores estándar robustos ($hc1$) y los errores estándar robustos de Davidson y MacKinnon ($hc3$). El porcentaje de éxito cambia en casi 32 puntos porcentuales respecto a los errores estándar normales. Por otro lado, se puede ver que las diferencias entre $hc1$ y $hc3$ no son tan grandes, lo que puede dar una visión de que los cambios fundamentales entre estos dos métodos ocurren con muestras pequeñas.

¹Puede revisar este [link](#) o [este](#) para más información.

²Este procedimiento es, de hecho, el default en R al usar errores robustos.

- b) Repitan el anterior procedimiento con muestras de tamaño 10. Presenten e interpreten sus resultados. Comparen con lo obtenido en el punto 1.

Tabla 5: 10 observaciones

VARIABLES	(1) Errores estandar	(2) hc1	(3) hc3
% de éxito del parámetro en intervalo	40 %	90 %	100 %
Número de observaciones	10	10	10

En este caso resaltan las diferencias entre los errores estandar tradicionales son más grandes respecto al caso de 1000 muestras, y además, se evidencia una mayor diferencia en el porcentaje de éxito entre los errores hc1 y los hc3, siendo estos últimos los que mejores resultados se obtuvieron.

3. Otro problema común es la autocorrelación o correlación serial, la cual tiene que ver con la existencia de grupos de unidades u observaciones dentro de la muestra cuyos términos de error se encuentran correlacionados. Este panorama rompe con el supuesto clásico de independencia entre los términos de error y plantea nuevos problemas que exploraremos a continuación.

- a) Simulen 1000 muestras $\{X_i, Y_i\}_{n=1}^{1000}$ acorde a los siguientes pasos:

- Simule las primeras 100 observaciones de la muestra siguiendo el modelo 1 tomando $\varepsilon_i \sim \mathcal{N}(0, 1)$ iid.
- Por medio del comando **expand**, produzca 10 copias de los datos (de esta forma construya una base con 1000 observaciones en total)

Explique por qué este procedimiento induce un problema de autocorrelación. Para cada muestra, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ empleando errores estándar usuales, *hc1*, *hc3* y cluster³. Los errores cluster permiten que las covarianzas de errores de observaciones pertenecientes a un mismo cluster sean estimadas individualmente (i.e. da libertad sobre el valor de estas covarianzas), mientras que las demás covarianzas son restringidas a 0. En cada caso, calcule el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero. Presenten e interpreten sus resultados.

Tabla 6: 1000 observaciones

VARIABLES	(1) Errores estandar	(2) hc1	(3) hc3	(4) cluster
% de éxito del parámetro en intervalo	46 %	50 %	43 %	93 %
Número de observaciones	1000	1000	1000	1000

Esto genera un problema de autocorrelación porque al repetir 10 veces los valores generados inicialmente voy a tener muestra que depende de una observación anterior. Al ser observaciones iguales el error puede estar correlacionado con una de las creadas repetidamente pues se repiten secuencialmente en mis datos, (primero 10 veces el primer elemento inicial, luego 10 veces el segundo y así sucesivamente) lo que genera que los errores de las observaciones adyacentes estén correlacionados, estas son las que debemos seleccionar para hacer los clusters. Como se puede ver en los resultados de la columna 4, al hacer los clusters se obtienen mejoras sustanciales en el porcentaje de éxito del parámetro.

- b) ¿Es el modelo heterocedástico sin autocorrelación un caso específico de errores con estructura cluster?

Si, pues al tratarse de un modelo heterocedástico la varianza de los errores es distinto para cada x , y en la estructura cluster de los errores si la varianza del estimador agrupado es menor que la del estimador no agrupado, significa que las sumas agrupadas de $e_i \cdot x_i$ tienen menos variabilidad que los $e_i \cdot x_i$ individuales. Es decir, cuando se suman los $e_i \cdot x_i$ dentro de un conglomerado, parte de la variación se anula y la variación total es menor.

- c) Un problema con los errores clusters es que el número de entradas de la matriz varcov de los errores que hay que estimar puede ser muy alto si el número de clusters es relativamente bajo.

³Pista: piense cuáles son las observaciones que se correlacionan para escoger el nivel de cluster

Como consecuencia, con pocos clusters se hace necesario estimar un mayor número de parámetros simultáneamente, lo que puede afectar la calidad de dicha estimación. Para ver esto, simulen 1000 muestras $\{X_i, Y_i\}_{i=1}^{1000}$ para $j = 10, 20, 50, 100$ según el siguiente procedimiento:

- Simulen las primeras j observaciones.
- Dupliquen las base de datos tantas veces sea necesario para producir 1000 observaciones en total.
- Para cada muestra, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ empleando errores estándar cluster.

Para cada j , calcule el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero. Gráfique el j empleado contra el porcentaje de aciertos respectivo e interprete.

4. Una alternativa al problema de seleccionar la estructura de los errores para estimar su matriz de varianza y covarianza es a través de métodos no paramétricos. En particular, la metodología de bootstrap nos permite recuperar una estimación del error estándar incluso en contextos en los que no es claro cuál de todos los estimadores previamente vistos es más adecuado.

- a) Simulen 1000 muestras $\{X_i, Y_i\}_{i=1}^{1000}$ acorde al modelo (1) cuando $\varepsilon_i \sim \mathcal{N}(0, 1)$ iid. Para cada una, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ del 95 % empleando errores estándar bootstrap con 100 repeticiones. Calculen el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero.

Tabla 7: 1000 observaciones

VARIABLES	(1) Errores estandar
% de éxito del parámetro en intervalo	95.2 %
Número de observaciones	1000

- b) Simulen 1000 muestras $\{X_i, Y_i\}_{i=1}^{1000}$ acorde al modelo (1) cuando $\varepsilon_i \sim \mathcal{N}(0, X_i^2)$ independientes. Para cada una, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ del 95 % empleando errores estándar bootstrap con 100 repeticiones. Calculen el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero.

Tabla 8: 1000 observaciones

VARIABLES	(1) Errores estandar
% de éxito del parámetro en intervalo	93.7 %
Número de observaciones	1000

- c) Simulen 1000 muestras $\{X_i, Y_i\}_{i=1}^{1000}$ acorde a los siguientes pasos:

- Simule las primeras 100 observaciones de la muestra siguiendo el modelo 1 tomando $\varepsilon_i \sim \mathcal{N}(0, 1)$ iid.
- Por medio del comando **expand**, produzca 10 copias de los datos (de esta forma construya una base con 1000 observaciones en total)

Para cada muestra, estimen el modelo (1) y calculen el intervalo de confianza $\hat{\beta}$ del 95 % empleando errores estándar bootstrap con 100 repeticiones. Calculen el porcentaje de las veces en que en dicho intervalo se encuentra el parámetro verdadero.

Tabla 9: 1000 observaciones

VARIABLES	(1) Errores estandar
% de éxito del parámetro en intervalo	39 %
Número de observaciones	1000

d) Comparen sus resultados con lo obtenido en los anteriores incisos.

Para comenzar podemos ver que los resultados del 4.a son similares a los del 1.a en el que se había obtenido un porcentaje de acierto del 94.7%, mientras que utilizando los errores bootstrap se tiene un acierto del 95.2%. Por otro lado, en el caso de la distribución con desviación estándar x^2 , obtuvo mejores resultados que utilizando los errores estandar clásicos y resultados similares de solo un punto porcentual de diferencia respecto a los errores robustos hc1 y hc3. Finalmente, donde más se tuvieron diferencias y problemas en el porcentaje de éxito del parámetro en el intervalo fue en el caso donde se crean 100 observaciones y luego se reproducen copias exactas debido a que este solo tuvo un 39% de éxito, teniendo el nivel más bajo respecto a los otros cuatro tipos de errores. En este caso, el que mejores resultados obtuve fue el de cluster, mientras que los otros, incluído el bootstrap tuvieron porcentajes de éxito muy bajos.