

# DATOS DE PANEL: EXAMEN FINAL

FELIPE. I. TAPPATA

RESUMEN. Este documento contiene la solución al examen final de la materia *Datos de Panel* en la Universidad Torcuato Di Tella, primer trimestre de 2025. El trabajo consiste en la replicación de ciertas tablas y figuras del trabajo de Al-Sadoon et al. (2019), y un análisis de los resultados obtenidos. El código usado fue entregado junto a este documento, y se puede encontrar también en el repositorio <https://github.com/felipetappata/datos-final>. El software usado es Stata, con un uso auxiliar de Python y Bash para el procesamiento de *output* y ayuda en la ejecución simultánea de simulaciones. El archivo `README.md` en la raíz del repositorio contiene instrucciones detalladas para la replicación de los resultados. El texto principal está escrito en español, pero algunos términos como “bias” y “s.e.” se han dejado en inglés para evitar construcciones torpes y resaltar la correspondencia con el trabajo original. Lo mismo ocurre con el código; dado que la documentación y los trabajos originales suelen estar en inglés, los comentarios e instrucciones mantienen el idioma.

**Problema 1.** Reproduzca las tablas 1 a 3 del trabajo de Sadoon et al.

**Solución 1.** Para la construcción de los Cuadros 1, 2 y 3, replicamos los experimentos de Monte Carlo desarrollados por los autores. El proceso generador de datos consiste en un modelo dinámico autorregresivo de primer orden con selección de muestra. Específicamente, generamos la variable de resultado de interés según la ecuación de resultado:

$$y_{it}^* = 2 + \rho y_{i,t-1}^* + \alpha_i + \varepsilon_{it} \quad \text{para } t = 2, \dots, T$$

donde  $\rho$  toma valores de 0.25, 0.50 y 0.75 para capturar diferentes grados de persistencia. Para el período inicial ( $t = 1$ ), utilizamos la forma

$$y_{i1}^* = \frac{2 + \alpha_i + \varepsilon_{i1}}{1 - \rho}.$$

El mecanismo de selección se modela mediante dos especificaciones alternativas. La especificación estática (modelo A) se define como:

$$d_{it}^* = a - z_{it} - \eta_i - u_{it}$$

mientras que la especificación dinámica (modelo B) incluye el rezago de la variable de selección:

$$d_{it}^* = a - 0.5d_{i,t-1}^* + z_{it} - \eta_i - u_{it}$$

En ambos casos,  $d_{it} = \mathbf{1}[d_{it}^* > 0]$ . La constante  $a$  se calibra para que la probabilidad inicial de selección sea  $P(d_{it}^* > 0) = 0.85$  en el modelo A. La variable exógena  $z_{it}$  se distribuye normal con varianza unitaria.

La estructura de correlación entre los componentes de error se especifica mediante las ecuaciones

$$\alpha_i = \alpha_i^0 + 0.5\eta_i,$$

$$\varepsilon_{it} = \varepsilon_{it}^0 + 0.5u_{it},$$

donde  $\alpha_i^0$ ,  $\varepsilon_{it}^0$ ,  $\eta_i$  y  $u_{it}$  son independientes con distribución normal estándar. Esta parametrización implica una correlación de 0.447 entre los componentes de error de las ecuaciones de resultado y selección.

Para cada combinación de parámetros, realizamos 500 replicaciones con tamaños de muestra iniciales de  $N = 500$  y  $N = 5000$ . Generamos series temporales de longitud  $T = 17$  a  $T = 20$  y descartamos las primeras 13 observaciones para mitigar efectos de condiciones iniciales, resultando en paneles con dimensión temporal efectiva  $T = 7$ .<sup>1</sup> Dado que el estimador requiere al menos tres observaciones consecutivas del mismo régimen, una fracción considerable de observaciones no contribuye a la identificación, resultando en una pérdida efectiva de aproximadamente un tercio de las observaciones aun con solo 15 % de selección inicial.

Evaluamos el desempeño de dos estimadores bien conocidos: el estimador GMM-IV en primeras diferencias de Arellano-Bond (AB) y el estimador de sistema (SyS) que combina ecuaciones en niveles y primeras diferencias. Para las ecuaciones en primeras diferencias utilizamos como instrumentos todos los rezagos disponibles desde  $t - 2$  hacia atrás, mientras que para las ecuaciones en niveles empleamos las primeras diferencias rezagadas de la variable dependiente como instrumentos adicionales.

La implementación computacional se encuentra en los *scripts* de Stata: `endogenous.do` y `nonendogenous.do`. Para el estimador de Arellano-Bond, empleamos el comando:

```
xtabond2 y L.y, gmm(L.y, collapse) nolevel
```

La opción `collapse` es utilizada por los autores, siguiendo las recomendaciones de Roodman (2009).<sup>2</sup> Para el estimador de sistema, utilizamos:

```
xtabond2 y L.y, gmm(L.y, lag(2 .)) iv(L.D.y, equation(level))
```

donde `lag(2 .)` especifica el uso de rezagos desde  $t - 2$  hacia atrás como instrumentos GMM, y `iv(L.D.y, equation(level))` incluye la primera diferencia rezagada como instrumento IV para la ecuación en niveles, siguiendo la propuesta de Arellano y Bover (1995).

El Cuadro 1 muestra los resultados de sesgo promedio y error estándar para los dos estimadores, para experimentos Monte Carlo con diferentes valores de  $\rho$  y tamaños de muestra, para las dos especificaciones del modelo de selección, tanto para selección endógena y no.<sup>3</sup>

Los Cuadros 2 y 3 son el resultado de un análisis de sensibilidad para los tamaños de muestra chico ( $N = 500$ ) y grande ( $N = 5000$ ), respectivamente. El análisis de sensibilidad, o robustez, consiste en diseñar cinco “experimentos” que modifican el escenario original para

<sup>1</sup>Los autores mencionan que los resultados no cambian si se generan las 13 observaciones adicionales y la muestra se comienza con una condición individual para cada individuo (Al-Sadoon et al., 2019, pág. 11).

<sup>2</sup>Comentan, de todos modos, que el problema de proliferación de instrumentos no es un problema en la aplicación dada, debido a la cantidad reducida de períodos (Al-Sadoon et al., 2019, pág. 12)

<sup>3</sup>Para generar este cuadro, se corren los scripts `run_parallel_N500.sh` y `run_parallel_N5000.sh` para ejecutar en simultáneo diferentes instancias de `endogenous.do` y `nonendogenous.do` con los parámetros correspondientes. Los resultados se guardan en archivos `csv` y luego se procesan con el script `make_table_1.py` para generar el cuadro.

CUADRO 1. Sesgo promedio en el modelo AR(1) ( $T = 7$ , 500 replicaciones)

				No endogenous selection		Endogenous selection	
Select. Model	$\rho$		(1) AB	(2) SYS	(3) AB	(4) SYS	
<hr/>							
$N = 500$							
A	.25	bias	-0.009 <u>4</u> 9	0.001 <u>2</u> 6	-0.007 <u>3</u> 8	-0.004 <u>4</u> 5	
		s.e.	0.071 <u>4</u> 5	0.04 <u>5</u> 60	0.06 <u>3</u> 23	0.04 <u>5</u> 89	
A	.50	bias	-0.02 <u>3</u> 32	0.001 <u>8</u> 6	-0.01 <u>4</u> 90	-0.01 <u>0</u> 56	
		s.e.	0.114 <u>8</u> 4	0.05 <u>0</u> 38	0.09 <u>1</u> 59	0.05 <u>2</u> 48	
A	.75	bias	-0.1 <u>0</u> 806	0.004 <u>9</u> 2	-0.03 <u>9</u> 82	-0.01 <u>3</u> 73	
		s.e.	0.2 <u>5</u> 894	0.05 <u>6</u> 37	0.1 <u>5</u> 301	0.06 <u>3</u> 24	
<hr/>							
$N = 5000$							
A	.25	bias	-0.001 <u>6</u> 9	0.000 <u>3</u> 7	-0.000 <u>9</u> 7	-0.00 <u>5</u> 53	
		s.e.	0.022 <u>8</u> 5	0.014 <u>2</u> 1	0.020 <u>2</u> 2	0.014 <u>0</u> 3	
A	.50	bias	-0.004 <u>2</u> 8	0.000 <u>2</u> 8	-0.00 <u>2</u> 43	-0.01 <u>2</u> 74	
		s.e.	0.036 <u>9</u> 8	0.016 <u>0</u> 1	0.029 <u>5</u> 4	0.016 <u>2</u> 3	
A	.75	bias	-0.014 <u>8</u> 9	$5.70 \times 10^{-8}$	-0.00 <u>5</u> 73	-0.02 <u>1</u> 87	
		s.e.	0.081 <u>9</u> 9	0.018 <u>3</u> 2	0.048 <u>3</u> 0	0.020 <u>8</u> 4	
<hr/>							
$N = 500$							
B	.25	bias	-0.01 <u>1</u> 37	0.003 <u>3</u> 0	-0.006 <u>9</u> 7	-0.007 <u>2</u> 1	
		s.e.	0.081 <u>8</u> 0	0.05 <u>1</u> 15	0.069 <u>7</u> 9	0.05 <u>1</u> 25	
B	.50	bias	-0.030 <u>9</u> 3	0.003 <u>0</u> 4	-0.014 <u>7</u> 2	-0.014 <u>5</u> 7	
		s.e.	0.132 <u>2</u> 5	0.056 <u>9</u> 1	0.099 <u>3</u> 6	0.059 <u>9</u> 9	
B	.75	bias	-0.142 <u>3</u> 6	0.006 <u>6</u> 8	-0.034 <u>8</u> 0	-0.015 <u>1</u> 9	
		s.e.	0.293 <u>3</u> 3	0.061 <u>5</u> 0	0.153 <u>0</u> 9	0.069 <u>2</u> 2	
<hr/>							
$N = 5000$							
B	.25	bias	0.000 <u>2</u> 0	0.000 <u>4</u> 7	0.000 <u>8</u> 9	-0.008 <u>7</u> 2	
		s.e.	0.026 <u>2</u> 0	0.016 <u>0</u> 4	0.021 <u>5</u> 3	0.015 <u>7</u> 8	
B	.50	bias	-0.001 <u>5</u> 6	0.000 <u>5</u> 4	0.000 <u>3</u> 2	-0.016 <u>3</u> 6	
		s.e.	0.042 <u>6</u> 8	0.018 <u>1</u> 1	0.030 <u>8</u> 1	0.018 <u>6</u> 3	
B	.75	bias	-0.012 <u>6</u> 4	0.000 <u>6</u> 2	-0.001 <u>6</u> 0	-0.025 <u>6</u> 4	
		s.e.	0.097 <u>0</u> 4	0.020 <u>6</u> 6	0.047 <u>5</u> 3	0.024 <u>5</u> 0	

mostrar que los resultados no se modifican significativamente. Los experimentos son los siguientes (Al-Sadoon et al., 2019, págs. 16, 17).

**Experimento I:  $T$  corto.** Consiste en reducir la máxima dimensión longitudinal del panel observado de  $T = 7$  a  $T = 4$ , incrementando el número de observaciones descartadas de 13 a 16. El efecto sobre el sesgo medio es moderado para ambos estimadores y ambos tamaños muestrales,

pero incrementa sustancialmente la varianza, especialmente para el estimador AB con valores altos de  $\rho$ .

**Experimento II: Más selección.** Consiste en aumentar el grado de selección de muestra del 15 % al 25 %, modificando la constante  $a$  de 1.794 a aproximadamente 1.168. Aumenta el sesgo medio muy levemente y la varianza por la reducción en la cantidad de observaciones disponibles.

**Experimento III: Incrementar el ratio de varianzas.** Consiste en aumentar el ratio de la varianza del componente de heterogeneidad individual respecto a la varianza del componente variante en el tiempo de la ecuación de resultado, estableciendo  $\sigma_\eta = 2$  en lugar de  $\sigma_\eta = 1$ . El efecto sobre el sesgo medio es limitado para ambos estimadores, siendo más pronunciado en muestras pequeñas.

**Experimento IV: Reducir la correlación de errores.** Consiste en reducir los parámetros de correlación de  $\theta = \vartheta = 0.5$  a  $\theta = \vartheta = 0.25$ . Se reduce el sesgo medio para ambos estimadores en todos los escenarios considerados, confirmando que menor correlación entre componentes de error facilita la estimación.

**Experimento V: Componentes del error no estacionarios y variantes en el tiempo.** Introduce heterogeneidad temporal multiplicando  $u_{it}$  por un proceso Bernoulli que toma valores 1 o 2, y permitiendo que el parámetro  $\vartheta$  varíe aleatoriamente entre 0.5, 1 y 2. El sesgo medio se reduce significativamente para ambos estimadores, siendo el efecto más notable en muestras pequeñas.

La implementación de estos experimentos se realiza mediante scripts especializados (llamados en el repositorio `experiment1.do` a `experiment5.do`) que modifican parámetros específicos del proceso generador de datos base. Por ejemplo, el Experimento I modifica la variable `T_discard` de 13 a 16, mientras que el Experimento II recalcula la constante  $a$  para lograr la probabilidad de selección deseada. El Experimento III simplemente cambia `sigma_eta` de 1 a 2, y el Experimento IV reduce `theta_param` y `varthetaeta_param` de 0.5 a 0.25. El Experimento V es el más complejo, introduciendo variabilidad temporal mediante:

```
gen bernoulli_process = (runiform() > 0.5) + 1
replace u_it = u_it * bernoulli_process
gen vartheta_random = (0.5, 1, 2)[ceil(runiform()*3)]
replace eps_it = eps_i0 + vartheta_random * u_it
```

Como podemos observar en los cuadros, los resultados confirman las conclusiones principales del trabajo original. En el Cuadro 1, el estimador AB muestra un sesgo que crece sistemáticamente con el parámetro autorregresivo  $\rho$  cuando el tamaño muestral es pequeño ( $N = 500$ ), llegando a ser considerable cuando  $\rho = 0.75$ . Al incrementar el tamaño muestral a  $N = 5000$ , el sesgo promedio del estimador AB se reduce sustancialmente.

Por el contrario, el estimador System muestra un sesgo muy pequeño para  $N = 500$  (nunca excediendo el uno por ciento), y aún menor cuando  $N = 5000$ . Bajo selección endógena, no detectamos cambios significativos en los resultados de sesgo para el estimador AB, mientras que el estimador System siempre muestra un sesgo muy pequeño (entre 1 y 2.25 por ciento

para diferentes valores de  $\rho$ ). Notablemente, el sesgo del estimador System se vuelve más evidente conforme crece el tamaño muestral bajo selección endógena.

Los ejercicios de sensibilidad presentados en los Cuadros 2 y 3 confirman las lecciones principales del análisis: el estimador AB está moderadamente sesgado cuando  $N$  es pequeño o moderado, y su sesgo se acerca a cero cuando  $N$  es grande. El estimador System está siempre moderadamente sesgado pero de manera estable. La reducción de la dimensión temporal (Experimento I) incrementa la varianza estimada pero tiene un efecto muy pequeño sobre el sesgo promedio para ambos estimadores. El incremento en el grado de selección muestral (Experimento II) aumenta el sesgo promedio muy levemente pero incrementa su varianza debido a la reducción significativa en el número de observaciones. Estos resultados implican que el estimador System es especialmente recomendado cuando el tamaño muestral es pequeño o incluso moderado (por ejemplo,  $N \leq 1500$ ) y menos crítico cuando el tamaño muestral es mayor.

**Problema 2.** Reproduzca la figura 1 del trabajo de Sadoon et al.

**Solución 2.** La Figura 1 replica la Figura 1 del trabajo original mediante simulaciones de para diferentes parámetros de persistencia y diferentes tamaños de muestra.<sup>4</sup> En particular, hay un gráfico para cada parámetro de persistencia  $\rho \in \{0.25, 0.50, 0.75\}$ , en el cual se muestra la evolución de los sesgos promedio en las simulaciones Monte Carlo para los estimadores Arellano-Bond (AB) y System (SyS) sobre la muestra completa (*all*) y sobre la muestra con selección endógena (*select*), para cada tamaño de muestra

$$N \in \{200, 400, 600, 800, 1000, 1500, 2000, 2500, 3000, 3500, 4000, 4500, 5000\}.$$

Los resultados confirman los patrones fundamentales identificados en el trabajo original y observables en el Cuadro 1. Para el estimador AB, observamos una reducción dramática (en términos absolutos) del sesgo a medida que aumenta el tamaño muestral, siendo esta mejora particularmente pronunciada para valores altos de  $\rho$ . En el panel 1c con  $\rho = 0.75$ , el sesgo del estimador AB es sustancial para muestras pequeñas pero converge rápidamente hacia cero cuando  $N$  crece.

El estimador System exhibe un comportamiento cualitativamente diferente: su sesgo permanece relativamente estable y pequeño a través de todos los tamaños muestrales considerados y para todos los valores de persistencia. Esta estabilidad es evidente en los tres paneles, donde las líneas correspondientes al estimador System se quedan alrededor de valores cercanos a cero sin mostrar la tendencia decreciente (en valor absoluto) dramática que caracteriza al sesgo del estimador AB.

---

<sup>4</sup>Como se describe en el README, el proceso de simulación se implementa mediante el script `run_figure1_sims.sh`, que ejecuta simulaciones paralelas para cada combinación de parámetros ( $N, \rho, \text{modelo}$ ). Para cada configuración, el script ejecuta tanto simulaciones con selección endógena como estimaciones en la muestra completa, generando los datos necesarios para las cuatro series mostradas en cada panel: AB (*all*), AB (*select*), System (*all*), y System (*select*). Las simulaciones cubren tamaños muestrales desde  $N = 200$  hasta  $N = 5000$ , permitiendo observar la evolución del sesgo a través de diferentes escalas de muestra. Una vez completadas las simulaciones, el script `fig1.R` procesa los resultados y genera las visualizaciones utilizando el paquete `tikzDevice` de R para compatibilidad directa con L<sup>A</sup>T<sub>E</sub>X. El código genera tres paneles separados correspondientes a  $\rho = 0.25, 0.50, 0.75$ , cada uno como un archivo `.tex` independiente que contiene comandos TikZ.

CUADRO 2. Sesgo promedio en el modelo AR(1). Análisis de sensibilidad para  $N$  chico

		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
Model		AB	SYS	AB	SYS	AB	SYS
Experiment I: Very short $T$ ( $T = 4$ )							
A	bias	-0.00774	-0.00843	-0.01983	-0.02051	-0.04991	-0.03965
	s.e.	0.12400	0.08601	0.19767	0.09995	0.37553	0.13631
B	bias	-0.01572	-0.01832	-0.02915	-0.03412	-0.06525	-0.06328
	s.e.	0.14211	0.09505	0.21703	0.11197	0.40076	0.17570
Experiment II: More sample selection (25 %)							
A	bias	-0.01173	-0.00511	-0.02170	-0.01144	-0.04819	-0.01165
	s.e.	0.07156	0.05283	0.10071	0.06163	0.15277	0.07514
B	bias	-0.01340	-0.00660	-0.02241	-0.01366	-0.04228	-0.00993
	s.e.	0.08385	0.06176	0.11483	0.07377	0.16464	0.08721
Experiment III: Increasing the ratio of variances: $\sigma_\eta/\sigma_\varepsilon = 2$							
A	bias	-0.00982	-0.00600	-0.01803	-0.01665	-0.04587	-0.02102
	s.e.	0.06095	0.04880	0.08754	0.05710	0.14852	0.07023
B	bias	-0.00977	-0.01072	-0.01764	-0.02212	-0.04125	-0.02232
	s.e.	0.06509	0.05149	0.09300	0.06156	0.15150	0.07617
Experiment IV: Reducing the correlation of the errors: $\rho = 0.25$							
A	bias	-0.00814	-0.00043	-0.01816	-0.00172	-0.06140	-0.00046
	s.e.	0.06777	0.04593	0.10310	0.05137	0.19423	0.05927
B	bias	-0.00854	0.00011	-0.02083	-0.00215	-0.06449	0.00042
	s.e.	0.07629	0.05138	0.11493	0.05827	0.20300	0.06423
Experiment V: Non-stationary time-varying error components							
A	bias	-0.00416	-0.00438	-0.00825	-0.01045	-0.01879	-0.01744
	s.e.	0.05014	0.04214	0.06596	0.04669	0.09807	0.05543
B	bias	-0.00134	-0.00501	-0.00575	-0.01221	-0.01744	-0.01919
	s.e.	0.05751	0.04786	0.07497	0.05282	0.10749	0.06033

<sup>1</sup>  $T = 7$ , excepto en el experimento I.<sup>2</sup>  $N = 500$ .<sup>3</sup> Numero de replicaciones: 500.

**Problema 3.** Comente sobre el sesgo promedio de las estimaciones a medida que aumenta el valor del parámetro de persistencia  $\rho$  para ambos estimadores: Arellano-Bond y SyS. ¿Varían las conclusiones al trabajar con muestras pequeñas o grandes?

**Solución 3.** A partir de la implementación de simulaciones de Monte Carlo siguiendo la metodología de Al-Sadoon et al. (2019), analizamos las propiedades de muestra finita de los estimadores de Arellano-Bond (AB) y System (SyS) bajo diferentes grados de persistencia para el proceso autorregresivo.

CUADRO 3. Sesgo promedio en el modelo AR(1). Análisis de sensibilidad para  $N$  grande

		$\rho = 0.25$		$\rho = 0.5$		$\rho = 0.75$	
Model		AB	SYS	AB	SYS	AB	SYS
Experiment I: Very short $T$ ( $T = 4$ )							
A	bias	0.00057	-0.00523	0.00043	-0.01223	-0.00086	-0.02333
	s.e.	0.04357	0.02738	0.06758	0.03170	0.11401	0.03916
B	bias	0.00161	-0.01003	0.00240	-0.01857	0.00375	-0.03185
	s.e.	0.04521	0.02975	0.06772	0.03524	0.10651	0.04515
Experiment II: More sample selection (25 %)							
A	bias	-0.00072	-0.00549	-0.00193	-0.01346	-0.00452	-0.02358
	s.e.	0.02329	0.01667	0.03333	0.01891	0.05081	0.02382
B	bias	0.00054	-0.00997	-0.00068	-0.01812	-0.00352	-0.02758
	s.e.	0.02710	0.01965	0.03802	0.02326	0.05471	0.03063
Experiment III: Increasing the ratio of variances: $\sigma_\eta/\sigma_\varepsilon = 2$							
A	bias	-0.00055	-0.00979	-0.00169	-0.02268	-0.00468	-0.04042
	s.e.	0.01965	0.01519	0.02907	0.01808	0.04978	0.02540
B	bias	0.00057	-0.01533	-0.00003	-0.02917	-0.00192	-0.04805
	s.e.	0.02072	0.01651	0.02959	0.01979	0.04812	0.02903
Experiment IV: Reducing the correlation of the errors: $\rho = 0.25$							
A	bias	-0.00134	-0.00136	-0.00326	-0.00346	-0.00864	-0.00593
	s.e.	0.02171	0.01409	0.03325	0.01601	0.06119	0.01908
B	bias	0.00058	-0.00215	-0.00044	-0.00418	-0.00438	-0.00625
	s.e.	0.02407	0.01579	0.03653	0.01812	0.06457	0.02171
Experiment V: Non-stationary time-varying error components							
A	bias	0.00111	-0.00402	0.00097	-0.00967	0.00019	-0.01655
	s.e.	0.01590	0.01219	0.02158	0.01361	0.03254	0.01651
B	bias	0.00077	-0.00582	0.00050	-0.01148	-0.00060	-0.01802
	s.e.	0.01726	0.01420	0.02267	0.01573	0.03294	0.01892

<sup>1</sup>  $T = 7$ , excepto en el experimento I.<sup>2</sup>  $N = 5000$ .<sup>3</sup> Numero de replicaciones: 500.

Los resultados del Cuadro 1 revelan patrones claramente diferenciados entre ambos estimadores. Para el estimador AB, el sesgo, en valor absoluto, incrementa sistemáticamente con el parámetro de persistencia  $\rho$ , exhibiendo un crecimiento particularmente pronunciado cuando  $\rho$  se acerca a la unidad. Este comportamiento es consistente con los problemas de instrumentos débiles que caracterizan al estimador en primeras diferencias cuando la serie presenta alta persistencia.

El estimador System muestra un comportamiento cualitativamente diferente. Su sesgo permanece relativamente estable y pequeño para todos los valores de  $\rho$  considerados, sin mostrar el deterioro sistemático que caracteriza al estimador AB. Esta estabilidad refleja la

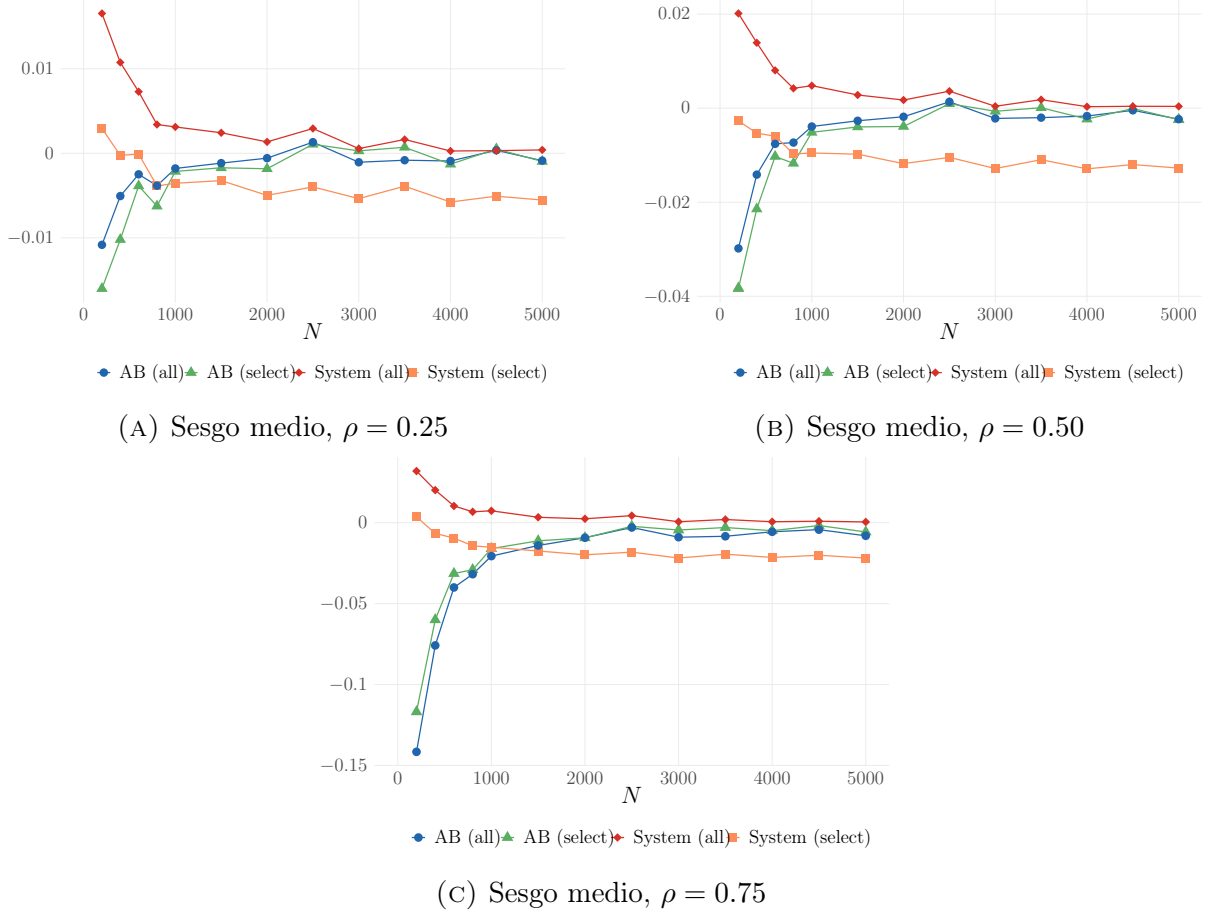


FIGURA 1. Sesgo promedio de estimadores AB y System en la muestra completa ( $N \times T$  observaciones) y la muestra con selección endógena

eficiencia de las condiciones de momentos adicionales en niveles que proporcionan información extra sobre los parámetros de interés.

Las diferencias entre muestras pequeñas ( $N = 500$ ) y grandes ( $N = 5000$ ) son sustanciales y asimétricas entre estimadores. Para muestras grandes, el sesgo del estimador AB se reduce drásticamente en todos los casos, acercándose a cero como es esperado por su consistencia. Esta mejora es especialmente notable para valores altos de  $\rho$ , donde la ganancia en precisión es considerable. Es decir, la *mejora* del sesgo del estimador AB cuando aumenta el tamaño de la muestra es más pronunciada para valores altos de  $\rho$ , como podemos ver al comparar la figura 1 que tiene  $\rho = 0.25$  con la figura 1c que tiene  $\rho = 0.75$ .

El estimador System mantiene un sesgo bajo y relativamente estable, independientemente de cual sea el tamaño muestral.

Las desviaciones estándar exhiben el comportamiento esperado para ambos estimadores: disminuyen al incrementarse el tamaño muestral. Sin embargo, el estimador AB consistentemente presenta mayor variabilidad que el System, especialmente para valores altos de  $\rho$ .

La presencia de selección endógena introduce modificaciones modestas pero sistemáticas en las propiedades de ambos estimadores. Para el estimador AB, que teóricamente debería mantener su consistencia bajo selección endógena, los efectos son generalmente pequeños y



en varias especificaciones del experimento se reducir el sesgo en términos absolutos cuando hay selección endógena respecto al caso de selección no endógena.

El estimador System muestra un incremento más sistemático del sesgo bajo selección endógena, particularmente evidente en muestras grandes y valores altos de  $\rho$ . Este comportamiento es consistente con la invalidez teórica de las condiciones de momentos en niveles bajo selección endógena, aunque la magnitud, a nivel práctico, del sesgo permanece manejable.

Los experimentos reportados en los Cuadros 2 y 3 confirman la robustez general de nuestros hallazgos principales. La reducción de la dimensión temporal (Experimento I) intensifica los problemas de ambos estimadores pero preserva su ordenación relativa. Los experimentos que modifican las correlaciones entre componentes de error o introducen heterogeneidad adicional generalmente mantienen las conclusiones cualitativas, aunque pueden alterar las magnitudes específicas del sesgo.

En línea con las conclusiones del trabajo original, nuestros resultados muestran que en muestras pequeñas, el estimador System presenta una performance consistentemente superior al AB, exhibiendo menor sesgo y varianza para todos los valores de  $\rho$  considerados. Esta ventaja es particularmente pronunciada para valores altos del parámetro de persistencia, donde el estimador AB sufre problemas severos de instrumentos débiles.

En muestras grandes, la brecha entre estimadores se reduce sustancialmente. La presencia de selección endógena no altera cualitativamente estas conclusiones, aunque si podemos decir en este caso que para muestras grandes en las especificaciones estudiadas, con selección endógena el estimador AB muestra un sesgo menor en términos absolutos que el estimador system.

En la práctica, si estamos trabajando con paneles dinámicos potencialmente desbalanceados, estos resultados sugieren que el uso directo de estimadores GMM-IV estándar, particularmente el System en muestras pequeñas y con parámetros de persistencia mayores, puede proveer estimaciones razonablemente precisas sin necesidad de corrección explícita por selección muestral, exhibiendo menor sesgo que el estimador AB.

## REFERENCIAS

- Arellano, Manuel y Olympia Bover (1 de jul. de 1995). “Another Look at the Instrumental Variable Estimation of Error-Components Models”. En: *Journal of Econometrics* 68.1, págs. 29-51. ISSN: 0304-4076. DOI: 10.1016/0304-4076(94)01642-D. URL: <https://www.sciencedirect.com/science/article/pii/030440769401642D> (visitado 14-07-2025).
- Roodman, David (2009). “How to Do Xtabond2: An Introduction to Difference and System GMM in Stata”. En: *Stata Journal* 9.1, págs. 86-136. URL: <https://ideas.repec.org/a/tsj/stataj/v9y2009i1p86-136.html> (visitado 14-07-2025).
- Al-Sadoon, Majid M., Sergi Jiménez-Martín y Jose M. Labeaga (2019). *Simple Methods for Consistent Estimation of Dynamic Panel Data Sample Selection Models*. 1631. Department of Economics and Business, Universitat Pompeu Fabra. URL: <https://ideas.repec.org/p/upf/upfgen/1631.html>.