Estudios Políticos 59

Artículo

¿Explicar peras a través del comportamiento de las manzanas? Estacionariedad y procesos generadores de datos en series de tiempo

Explaining pears through the behavior of apples? Stationarity and data-generating processes in time-series data.

Edwin Atilano-Robles*

Recibido: 8 de marzo, 2022. Aceptado: 6 de enero, 2023.

Resumen Las series de tiempo que se utilizan en ciencia política tienden a mostrar una fuerte dependencia de sus valores previos. Esto implica que, al incorporarlas en modelos estadísticos, es sencillo que se viole el supuesto de correlación serial. No obstante, al utilizar procesos temporales estacionarios es posible solventar dicha problemática. Esta investigación muestra, a través de simulaciones Monte Carlo, los efectos de especificar modelos estadísticos que incluyen variables no estacionarias. Los diferentes procesos generadores de datos apuntan a que, si las variables no son estacionarias, es más probable que se produzcan inferencias espurias. En consecuencia, se recomienda evaluar el comportamiento de nuestras variables y, de ser el caso, transformarlas para que se comporten de forma estacionaria.

Palabras clave: series de tiempo, estacionariedad, simulaciones Monte Carlo, inferencia, modelos estadísticos.

Abstract Time-series in political science tend to show a strong dependence on their previous values. This implies that, when incorporated into statistical models, it is easy to violate the assumption

* Doctor en Ciencia Política por el CIDE A.C., México. Profesor de Tiempo Completo, Facultad de Estudios Superiores Acatlán, UNAM. Líneas de investigación: metodología, métodos cuantitativos, economía política, regímenes autoritarios. Correo electrónico: edwin_atilano@politicas.unam.mx

of serial correlation. However, by using stationary temporal processes it is possible to manage this problem. This research shows, through Monte Carlo simulations, the effects of specifying statistical models that include non-stationary variables. Different data-generating processes suggest that if the variables are not stationary, spurious inferences are more likely to occur. Consequently, it is recommended to evaluate the behavior of our variables and, if necessary, transform them so that they behave stationarily.

Keywords: time-series, stationarity, Monte Carlo simulations, inference, statistical models.

Introducción

Los modelos de series de tiempo, así como los de datos panel, son una de las estrategias de identificación de potenciales efectos causales más usuales en la ciencia política (Philips 2018). Existen numerosas referencias de la utilización de este tipo de técnicas en diferentes ámbitos de la disciplina. Podría mencionarse, por ejemplo, el análisis de la aprobación presidencial (Canes-Wrone y De Marchi 2002; Newman 2003), las variaciones en la calidad de la democracia así como de sus instituciones (Alizada et al. 2021; Przeworski et al. 2000; Ross 2006; Skaaning, Gerring, y Bartusevičius 2015), análisis de los regímenes autoritarios (Geddes, Wright, y Frantz 2018; Lucardi 2019; Svolik 2012; Wright 2008), el efecto de las instituciones políticas en diferentes resultados económicos (Acemoglu y Robinson 2005; Ansell y Samuells 2011; Iversen y Soskice 2006; 2011), los estudios de opinión pública (Castro Cornejo 2021; McCann y Lawson 2003; Reeskens et al. 2021; Ugues 2018), entre otros.

No obstante, la literatura sobre metodología ha mostrado que es sencillo cometer errores de especificación con este tipo de modelos, especialmente si no se realizan los diagnósticos apropiados para conocer la distribución y comportamiento de nuestras variables (Beck y Katz 1995; Kellstedt y Whitten 2018; Philips 2018; 2021; Philips, Rutherford, y Whitten 2016; Pickup 2014; Williams y Whitten 2011; De Boef y Keele 2008). Este problema surge en virtud de que la especificación de modelos que incorporan una fuente de variación temporal puede violar el supuesto que establece que una variable tiene que distribuirse de manera independiente. En otras palabras, que una variable no tiene dependencia de sus propias observaciones en puntos temporales previos. Al incorporar una variable con estas características a un modelo de regresión, se encontrará que se viola el supuesto de correlación serial, lo que implica que los residuales del modelo se correlacionan con valores previos de sí mismos (Pickup 2014). Como puede intuirse, es difícil deshacerse de este problema *ex ante* en modelos de series de tiempo.

Una manera de solventar la dificultad que representa la correlación serial es que las variables que se introduzcan a los modelos se comporten de manera estacionaria. Es decir, que cumplan al mismo tiempo con tres características: (1) regresión a la media, (2) varianza y (3) covarianza constantes (Philips 2018; 2021). Sin embargo, los procesos temporales que tendemos a analizar en ciencia política tienen una fuerte dependencia de observaciones previas (Philips 2021), lo que conduce a que se comporten de una forma no estacionaria. En otras palabras, las variables tendrían raíz unitaria (Das 2019).

El objetivo principal de este trabajo es mostrar que la incorporación de variables con raíz unitaria en modelos de series de tiempo no solo es una violación al supuesto de correlación serial, sino que presenta problemas más serios en términos de inferencia causal, ya que las series

de tiempo no estacionarias son una fuente donde pueden brotar inferencias espurias. A través de simulaciones Monte Carlo muestro diferentes procesos generadores de datos en los que se observa el efecto de introducir variables no estacionarias en los diferentes valores esperados del coeficiente β_1 (la pendiente) en modelos de regresión para series de tiempo. Los resultados de la simulación de los efectos de interés señalan que modelar procesos temporales con variables no estacionarias facilita que se produzcan inferencias incorrectas. Los hallazgos muestran que esto es especialmente grave cuando se intenta explicar una variable de resultado estacionaria a través de una variable de tratamiento no estacionaria. Esto ocurre porque el intervalo de confianza de la media de la distribución muestral del coeficiente β_1 con un 95% de confianza ni siquiera incluye al verdadero valor que se especificó en el proceso generador de datos. En otras palabras, los modelos con variables independientes no estacionarias permiten que se cometa el error tipo I con una muy alta probabilidad.

El segundo objetivo del documento es acercar discusiones metodológicas con un alto contenido técnico a un público hispanoparlante. Esto es relevante porque los debates metodológicos tienden a darse en inglés, por lo que es primordial tratar de cerrar poco a poco esta brecha. México y los países de habla hispana podrían producir más y mejores inferencias causales al incorporar investigaciones en las que la metodología de la investigación sea el objeto de estudio.

Por lo tanto, la aportación de este estudio es principalmente metodológica. A través del uso de simulaciones se muestra que ignorar los supuestos en los que se sustentan los modelos de series de tiempo podría tener consecuencias serias para las inferencias causales. Este texto busca que especialistas y estudiantes de ciencia política que utilizan estos métodos como estrategia de identificación de efectos causales puedan obtener resultados válidos. Es por esto por lo que se muestran los problemas y las posibles soluciones a través de una serie de diagnósticos y modificaciones en las variables.

El estudio de las series de tiempo

La discusión y el análisis de las series de tiempo como herramienta empírica en ciencia política no son nuevos. De hecho, en la literatura anglosajona puede encontrarse una cantidad importante de textos en los que se analizan problemas relevantes de la ciencia política a través de series de tiempo o bien que estudian metodológicamente los supuestos de las series de tiempo y muestran su comportamiento (Philips 2018; Philips, Rutherford, y Whitten 2016; Philips 2021; Das 2019; Beck y Katz 2011; 1995; Cryer 1986; Pickup 2014; De Boef y Keele 2008). Por ejemplo, De Boef y Keele (2008) advierten que diferentes aplicaciones de las series de tiempo realizadas desde la ciencia política no evalúan las restricciones de los modelos y tienden a presentarse interpretaciones de resultados deficientes.

De la misma forma, Philips (2018) señala que la importancia de las pruebas de raíz unitaria en series de tiempo para nuestra disciplina se ha tomado con mayor seriedad en años más recientes. Asimismo, propone una serie de pasos para verificar que las variables que se utilicen en una serie de tiempo sean estacionarias y cómo elegir el mejor modelo con base en el comportamiento de los datos. Asimismo, Philips (2021) enfatiza la importancia del proceso generador de las variables para poder realizar inferencias correctas al utilizar datos que tienen una dinámica temporal. El texto de Philips (2021) es un referente importante para esta propuesta ya que, a través de simulaciones Monte Carlo muestra cómo diferentes modelos tienen mejores o peores

resultados a partir de considerar las dinámicas temporales de los diferentes procesos generadores de datos. A diferencia de la propuesta de Philips (2021), en este texto se muestra la relevancia de los procesos generadores de datos, pero no en la especificación de los modelos, sino en el comportamiento de las propias variables. La estacionariedad de las variables también se aborda en el artículo de Philips (2021), pero sus simulaciones analizan el comportamiento de diferentes especificaciones a partir del corto plazo, largo plazo y procesos de cointegración.

El objetivo de mi propuesta es más modesto, ya que se concentra solo en una parte del proceso: la estacionariedad de las variables. Desafortunadamente, en la literatura hispanoparlante de ciencia política se discute poco sobre cuestiones metodológicas de métodos cuantitativos. Esto no implica la ausencia de referencias para el estudio de estadística en general y de series de tiempo en particular. Es por esto que resulta importante destacar que se pueden encontrar excelentes libros de texto para adentrarse a dichas herramientas (Wooldridge 2012; Nava 2015; Villavicencio 2010; Mauricio 2007; Gujarati 2006; Stock, Watson, y Larrión 2012; Pérez Ramírez 2007). No obstante, no hay discusión en términos de qué pasa con los hallazgos si se violan determinados supuestos.

En todas estas referencias se pueden encontrar explicaciones sobre lo que implica trabajar con series de tiempo y cómo hacerlo de forma correcta. En los textos que se analizaron para esta investigación, se pudo verificar que, de manera consistente, se estudian las propiedades de la variación temporal. Asimismo, se encuentra la metodología para realizar modelos con procesos autorregresivos, cómo manejar variables que tengan algún nivel de integración (es decir que tengan raíz unitaria) o qué son las medias móviles. Estas referencias son particularmente útiles si nuestro objetivo es especificar modelos ARMA o ARIMA.

Ahora bien, a pesar del hecho de que en cada referencia se puede encontrar mención a la importancia de la estacionariedad, los textos diseñados para la docencia no abundan en el análisis metodológico frente a complicaciones en la práctica de la investigación. Por ejemplo, en las referencias antes mencionadas se puede encontrar que al utilizar series de tiempo se recomienda utilizar variables que tengan procesos estacionarios (Wooldridge 2012; Villavicencio 2010), pero las explicaciones brindadas parecen asumir que las variables que se utilizarán en la práctica de la investigación siempre se comportarán así. Esto claramente no ocurre de esta forma.

Al realizar investigación empírica en ciencia política, nos encontraremos con variables que tienen diferentes comportamientos en sus dinámicas temporales y que pueden incorporarse en un mismo modelo, siempre y cuando realicemos los diagnósticos y las modificaciones apropiadas (Atilano Robles 2022). En este sentido, y ante la ausencia de discusión de qué ocurre cuando tenemos variables no estacionarias, la aportación de este artículo es mostrar que será muy probable que obtengamos resultados espurios si se especifican modelos con variables que provienen de un proceso generador de datos estacionario y las manejamos como no estacionarios. De la misma forma, se señala cómo proceder en caso de tener una variable no estacionaria y qué en qué referencias buscar para obtener más información para especificar mejores modelos.

Las series de tiempo en el marco de los métodos cuantitativos

Los modelos de series de tiempo son una de las herramientas que la ciencia política tiene a su alcance para realizar análisis empíricos robustos. No obstante, es fácil olvidar que dichos modelos parten de algunos supuestos que pueden enmarcarse en el campo de los métodos cuantitativos.

En primer lugar, es importante señalar que, usualmente, los métodos cuantitativos son estudios observacionales. Esto implica que la persona que realiza la investigación no tiene control sobre el proceso generador de datos (Keele 2020; Toshkov 2016). Además, a diferencia de un análisis experimental, no es posible asignar aleatoriamente un tratamiento (Huntington-Klein 2022; Kellstedt y Whitten 2018; Toshkov 2016).

De la misma forma, los métodos cuantitativos se sustentan en la premisa de que en el mundo existe cierta "estructura", de tal forma que la información que se obtiene de determinada unidad de análisis podría ser de utilidad para entender otra (Toshkov 2016). No obstante, esta estructura es difusa, parcial y contingente. Si no lo fuera, bastaría con analizar una sola unidad. En otras palabras, los métodos cuantitativos asumen que existe una señal (la estructura) y existe el ruido (la contingencia) y que la investigación consiste en analizar múltiples observaciones para diferenciar los patrones generales de la aleatoriedad (Huntington-Klein 2022).

Ahora bien, en términos de causalidad, los métodos cuantitativos suelen utilizarse para identificar y medir posibles efectos causales (King, Kehoane, y Verba 1994). En otras palabras, cuantificar en qué medida una variable puede causar a otra. No obstante, la causalidad no es observable, por lo que tiene que inferirse (Huntington-Klein 2022; King, Kehoane, y Verba 1994; Morgan y Winship 2014). Si utilizamos una perspectiva contrafactual, la causalidad implica que una variable de tratamiento es causal si y solo si, al modificar la supuesta causa, se modifica también el resultado, manteniendo todos los demás factores constantes (Toshkov 2016). Esta definición nos conduce a una imposibilidad de origen, ya que, para poder atribuir causalidad, tendríamos que observar dos situaciones en donde la única diferencia sea la presencia/ausencia del factor que se atribuye como causa, para así poder comparar el resultado. Esto no es viable en estudios observacionales, por lo que esta situación se conoce como el problema fundamental de la inferencia causal (Huntington-Klein 2022; Morgan y Winship 2014; Toshkov 2016).

Para tratar de manejar este problema y realizar inferencias causales válidas, los métodos cuantitativos pueden seguir una estrategia de condicionamiento, en la cual, se incorporan las variables de control apropiadas a los modelos para tratar de acercarse lo más posible a un escenario contrafactual en donde se mantienen constantes las variables confusoras y puede aislarse el efecto de interés (Kellstedt y Whitten 2018). Esta estrategia es la más usual (Toshkov 2016), sin embargo, tiene una incertidumbre irreductible, ya que es imposible saber si se han controlado todos los factores relevantes. En consecuencia, los métodos cuantitativos tienen la posibilidad de incurrir en el sesgo de variable omitida (Wooldridge 2012). El objetivo es precisamente tratar de minimizarlo.

Las series de tiempo forman parte de esta lógica de inferencia causal, razón por la cual, lo mencionado en esta sección es directamente aplicable a esta herramienta. No obstante, y aunque parezca una verdad de Perogrullo, las series de tiempo tienen una particularidad evidente: su fuente de variación es a lo largo del tiempo (Wooldridge 2012; 2001). Esto no es un asunto menor, ya que la variación temporal, a diferencia de la variación entre unidades, introduce problemas adicionales (Pickup 2014). Uno de esos problemas es el objeto de estudio de este artículo: la no estacionariedad de las variables con las que se especifican los modelos.

En consecuencia, si nuestro objetivo es realizar inferencias causales válidas, necesitamos analizar el comportamiento de nuestras variables y asegurarnos que sigan procesos estacionarios. De otra forma, no solo estaremos violando un supuesto de los modelos, sino que los resultados obtenidos no serán confiables. Para sustentar dicha afirmación, en los apartados subsecuentes mostraré que el problema de correlación serial que se genera por la inclusión de variables con

dinámica temporal no estacionaria en modelos de series de tiempo produce que resultados que no son válidos y que, por lo tanto, podrían llevarnos a inferencias causales espurias.

Algunas definiciones importantes

La especificación básica de los modelos de regresión estimados a través del procedimiento de Mínimos Cuadrados Ordinarios (MCO) con datos obtenidos de manera aleatoria se obtiene a partir de la siguiente ecuación:

$$y_i=\beta_0+\beta_1 x_i+\varepsilon_i \quad (1)$$

Este tipo de modelos parte del supuesto que el término ε_i , que representa al error aleatorio, tiene un valor esperado igual a cero, varianza constante y se comporta de forma normal e independientemente distribuida, es decir: $\varepsilon_i \sim NID(0,\sigma^2)$. Para obtener una serie de tiempo, podemos modificar ligeramente el modelo básico, de tal forma que la fuente de variación provenga de cambios temporales. En consecuencia, tenemos lo siguiente:

$$y_t=\beta_0+\beta_1 x_i+\varepsilon_t \quad (2)$$

En este caso, la diferencia de la ecuación 1 a la 2 en el subíndice *t* no es cosmética, ya que tiene implicaciones relevantes. Por ejemplo, el supuesto en el que $\varepsilon_t \sim NID(0,\sigma^2)$ no se cumplirá tan fácilmente, ya que tendremos problemas con el componente *I*. Esto ocurre porque una variable que cambia a lo largo del tiempo puede tener dependencia de sus observaciones. En otras palabras, el valor de una variable en el período puede depender del valor de la misma variable en el período *t*-1, el valor en el periodo *t*-1 puede depender de lo que se observó en el periodo *t*-2, y así sucesivamente. De la misma forma, el supuesto de varianza constante (homocedasticidad) en el que $var(\varepsilon_t)=\sigma^2 \forall t$ implica que tendremos volatilidad constante a lo largo del tiempo, lo cual difícilmente ocurrirá.

Por lo tanto, podemos afirmar que los datos de serie de tiempo son una colección de observaciones que ocurren secuencialmente en el tiempo (Das 2019), ya sea, de manera anual, mensual, trimestral, diaria, etc. En este sentido, las series de tiempo provienen de un proceso aleatorio o de uno estocástico. En otras palabras, la información surge de un proceso generador de datos (*data generating process*) en el que los factores subyacentes que explican el comportamiento de una serie de tiempo ocurren completamente al azar o bien por una combinación de aleatoriedad y valores previos de la variable (Das 2019). Por lo tanto, formalmente una serie de tiempo se genera a través de un proceso secuencial a lo largo del tiempo, de tal forma que:

$$\{y_t\} = (y_1, y_2, y_3, \ldots, y_T) \quad (3)$$

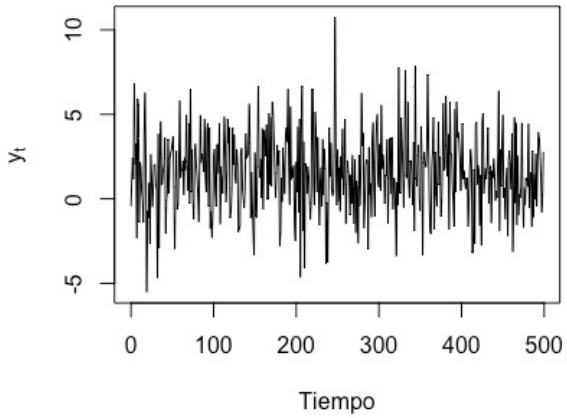
Cada serie de tiempo puede provenir de un proceso estacionario o no estacionario. Para que una variable con cambios temporales sea un proceso estacionario tiene que cumplir tres características: regresión a la media (ausencia de tendencia), varianza y covarianza constantes. La primera característica se puede denotar como $E(y_t)=\mu \forall t$ e implica que los valores tienden a regresar a un punto de equilibrio en la media. La segunda característica puede escribirse como

$E(y_t^2)=\sigma^2$ y se refiere a que la variación de nuestra variable no es volátil ante cambios temporales. Por último, la covarianza constante se denota como $E(y_t , y_{t-s}) = \rho_s$ y señala que la variación conjunta de y_t con el tiempo debe permanecer sin modificaciones abruptas. En contraste, una variable no estacionaria es aquella que incumple con al menos una de las tres características enunciadas.

En la Figura 1 puede observarse el comportamiento estacionario de una variable simulada, a la que simplemente llamaremos y_t. Esta será una de las variables que se utilizarán para realizar los experimentos de la sección siguiente. De la misma forma, en la Figura 2 se muestra un ejemplo de variable simulada no estacionaria. Para este caso, se utilizó una de las estructuras de series de tiempo más famosas: el camino aleatorio (*random walk*). El camino aleatorio se define de la siguiente manera: $z_t = 1.0x_{t-1}+\gamma_t$, en donde . $\gamma_t \sim N(\mu,\sigma^2)$.

Figura 1

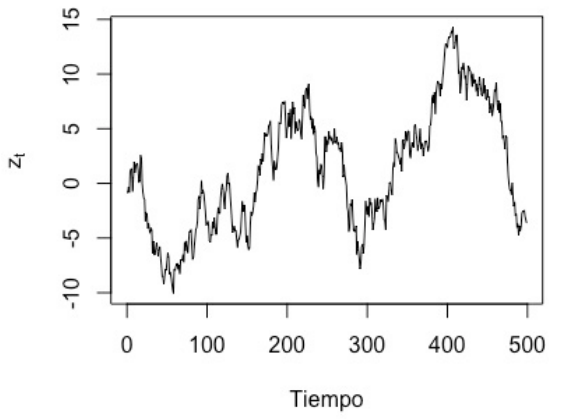
Ejemplo de una variable estacionaria



Fuente: elaboración propia

Figura 2

Ejemplo de una variable no estacionaria



Fuente: elaboración propia

La estructura del camino aleatorio implica que tiene una fuerte dependencia de valores anteriores en el tiempo, ya que el término autorregresivo x_{t-1} tiene un coeficiente igual a uno. No obstante, esta serie de tiempo también se encuentra en función de un segundo término, el cual es completamente aleatorio. En este caso, la inspección visual[1] muestra que no se cumple ninguna de las características de la estacionariedad.

Ante la situación de ambas variables, imaginemos el siguiente escenario: la variable y_t , que se muestra en la Figura 1 proviene de un proceso generador de datos tal que $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$, en donde $x_t \sim N(0,1)$, $\epsilon_t \sim N(0,1)$, x_t es independiente de ϵ_t y x_t es la versión estacionaria de z_t. En otras palabras, z_t es una transformación de x_t para que se comporte como un camino aleatorio. Algunas preguntas que podrían surgir de esta situación son: ¿cuál será el efecto de intentar explicar y_t en función de z_t y no de x_t?, ¿cuáles serán las repercusiones de tener una variable de resultado con raíz unitaria y una de tratamiento estacionaria?, ¿qué ocurre si ambas variables son no estacionarias? En la siguiente sección me dedico a explorar las respuestas a través de simulaciones Monte Carlo.

Simulaciones Monte Carlo

Al introducirnos a los modelos de regresión tendemos a estudiar el Teorema del Límite Central, cuya conclusión es que los estimadores $\hat{\beta}_i$ son insesgados porque el promedio de su distribución muestral es igual al parámetro β (Wooldridge 2012). En otras palabras, al promediar los estimadores de los modelos de regresión en N muestras, se obtendría el verdadero valor poblacional, el cual es el que buscamos conocer. No obstante, esta afirmación parte del supuesto de realizar muestreo repetido, lo cual no es factible en la práctica de la investigación cuantitativa. De hecho, en la práctica de la investigación solo se cuenta con una sola muestra. En consecuencia, para analizar los problemas que surgen al utilizar variables no estacionarias, podemos simular la distribución muestral y generar diferentes estimaciones $\hat{\beta}_i$. En este caso, las simulaciones Monte Carlo involucran conocer el verdadero valor de β de antemano y tener control del proceso generador de datos de las diferentes muestras que se "observarían" en cada experimento individual (Huntington-Klein 2022). A partir de este procedimiento mostraré que al intentar explicar variables que surgen de procesos generadores de datos estacionarios a través de variables no estacionarias (o viceversa) se producen estimadores sesgados o ineficientes. En otras palabras, el promedio de la distribución muestral no es el mismo que el especificado desde el proceso generador de datos o se obtienen estimadores con varianzas amplias.

No omito mencionar que la elección del tamaño del muestreo repetido, así como el número de observaciones de cada muestra se realizó de manera arbitraria.[2] No obstante, procuré que en todo momento se cumplieran los supuestos del Teorema del Límite Central para realizar inferencias válidas a través de simulaciones Monte Carlo (Gelman y Hill 2006). De igual forma, se

[1] De la misma forma se puede establecer la prueba de hipótesis Dickey-Fuller para conocer con mayor certeza cómo se comporta nuestra variable. Esta prueba se discute en la sección en la que se abordan las estrategias para evitar las inferencias espurias.

[2] Realicé pruebas adicionales con diferentes tamaños de la simulación y el número de observaciones de cada muestra y al variarlos no hubo cambios sustantivos.

puede observar que las simulaciones fueron correctamente especificadas ya que el promedio de las distribuciones muestrales fue igual al parámetro que se utilizó.

A continuación, planteo cuatro casos para los que se estableció el mismo escenario de simulación:[3] se generaron 1,000 muestras con un tamaño de 500 observaciones cada una en donde la variación se da a lo largo del tiempo. Como se comentó previamente, el proceso generador de datos de la variable y_t es el siguiente:

$$y_t = \beta_0 + \beta_1 x_t + \epsilon_t \text{, en donde } x_t \sim N(0,1) \text{ y } \epsilon_t \sim N(0,1) \quad (4)$$

En términos más específicos, el proceso generador de datos del verdadero modelo es el que se muestra a continuación:

$$y_t = 1.5 + 2x_t + \epsilon_t \quad (5)$$

Esto implica que la distribución muestral de los coeficientes $\hat{\beta}_1$ debería promediar el valor de dos.[4] Como se podrá observar en cada escenario, las implicaciones de modificar las especificaciones de los modelos son diferentes, por lo que se analizará uno a uno para mostrar en dónde se encuentra la mayor problemática.

– Simulación 1: y_t y x_t estacionarias

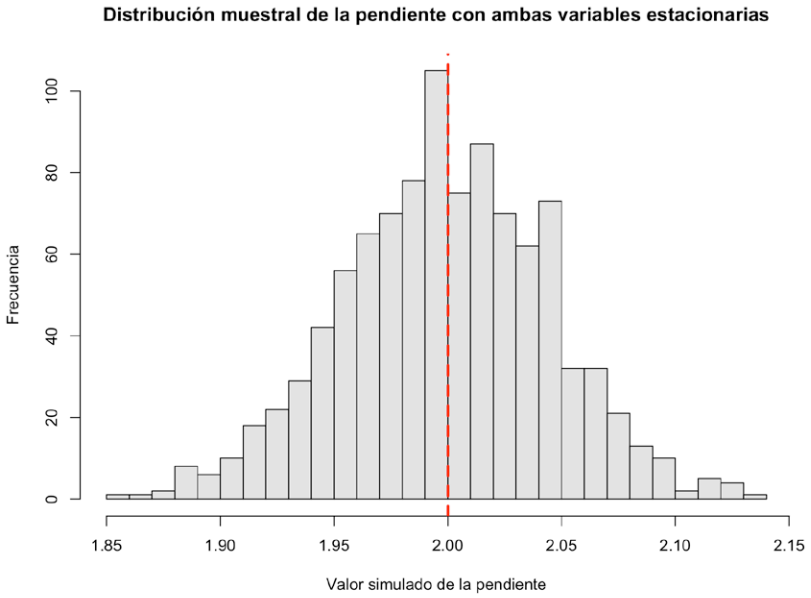
En este caso se utilizó una simulación de los modelos de series de tiempo que era concordante con el proceso generador de datos de y_t. En otras palabras, se simularon 1,000 modelos de regresión con la especificación: $y_t = \beta_0 + \beta_1 x_t + \epsilon_t$ y se extrajeron los coeficientes estimados de la pendiente en cada experimento. Como se mostró anteriormente, la variable x_t se distribuye de manera normal con media igual cero y desviación estándar de uno y se generó independientemente del término de error. Además, esta misma variable es la que se utilizó como parte del proceso generador de datos de y_t, por lo que la distribución muestral claramente señala que los diferentes estimadores $\hat{\beta}_1$ son insesgados, ya que el promedio de la distribución no difiere de lo que se dispuso en el proceso generador de datos.

En la Figura 3 se observa la distribución muestral de esta simulación, en donde la línea roja señala el promedio. De la misma forma, a través de una prueba de hipótesis t de Student es posible verificar que el promedio no difiere sustantivamente de dos. En la Tabla 1 se presentan los resultados de dicha prueba, en la que la hipótesis alternativa es que el promedio de la distribución muestral era diferente de dos. Los resultados señalan que no es posible rechazar la hipótesis nula y, en consecuencia, no se puede descartar con un 95% de confianza que el promedio sea igual a dos. No se esperaba algo diferente en esta primera simulación, ya que se utilizó la variable x_t que formaba parte del proceso generador de datos de y_t y que además se comporta de manera estacionaria.

[3] En el anexo de este artículo se encuentra el código en *R* para poder replicar todos los hallazgos que se presentan a continuación.

[4] La elección de los valores 1.5 y 2 para los parámetros del modelo también fue arbitraria. No obstante, esto no tendría que causar ningún problema en las simulaciones, en virtud de que el objetivo es mostrar las condiciones bajo las cuales se obtiene estimadores insesgados y eficientes con modelos de series de tiempo.

Figura 3



Fuente: elaboración propia

Tabla 1
Resultados de la prueba de hipótesis con ambas variables estacionarias

Estimación promedio	
1.9992	
Intervalo de confianza al 95%	
1.9964	2.0021
Valor del estadístico t	
t=-0.5396	
P-valor	
p=0.5896	

Fuente: elaboración propia

– Simulación 2: y_t estacionaria y z_t no estacionaria

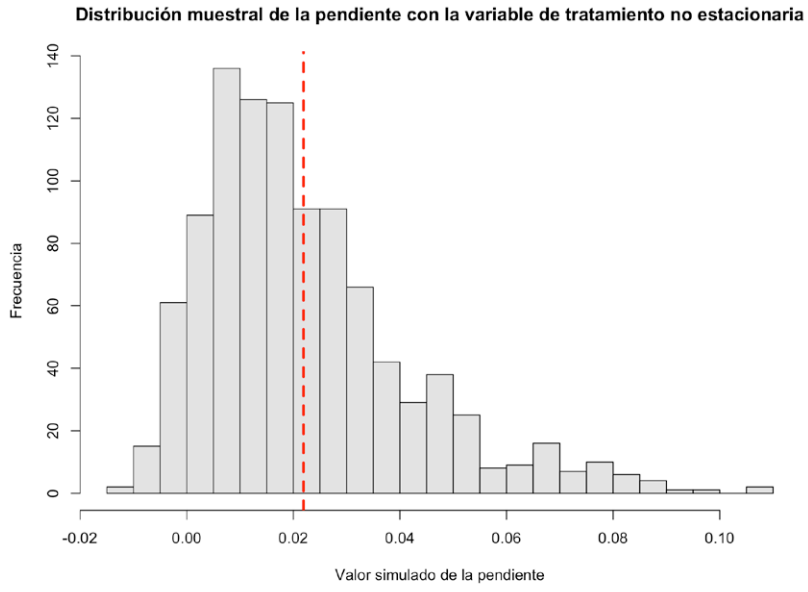
Para el segundo escenario se utiliza nuevamente y_t pero ahora se simulan los 1,000 modelos de regresión con la versión no estacionaria de x_t, es decir, se utiliza z_t, la cual se comporta como el camino aleatorio que se muestra en la Figura 2. La única modificación que se realiza del primer escenario a este es que permitimos que la variable de tratamiento tenga dependencia de sus valores previos. En virtud de que la única diferencia entre x_t y z_t es que la segunda es la versión no estacionaria de la primera, la intuición podría sugerirnos que los resultados de la simulación

del modelo $y_t = \beta_0 + \beta_1 z_t + \epsilon_t$ no deberían ser sustancialmente diferentes a los previos, pero esto no es así.

Como se muestra en la Figura 4, intentar explicar una variable de resultado estacionaria a través de una variable de tratamiento no estacionaria presenta complicaciones serias. A pesar del hecho de que el cambio entre x_t y z_t es solo la estacionariedad y que x_t forma parte del proceso generador de datos de y_t , proceder de esta forma, encontraremos que nuestros estimadores de la pendiente serán sesgados. Los hallazgos de esta simulación muestran que el promedio de la distribución muestral de los estimadores de la pendiente es igual a 0.0219, lo cual se señala en la Figura 4 con la línea roja. No olvidemos que el verdadero valor de la pendiente es de dos, por lo que los hallazgos en este escenario apuntan a una subestimación que tendría consecuencias realmente graves al momento de producir inferencias. En la Tabla 2 se presenta una prueba de hipótesis t de Student con la misma especificación que en el escenario previo. En virtud de que el intervalo al 95% de confianza no incluye el verdadero valor, estaríamos cometiendo un error tipo I: rechazar la hipótesis nula a pesar de que la hipótesis nula no tendría que haberse descartado.

Este escenario muestra las consecuencias más graves de una especificación incorrecta en modelos de series de tiempo. Si el proceso generador de datos de nuestra variable de resultado es estacionario, necesitamos explicarla a través de la variable de tratamiento que, por supuesto, forma parte de dicho proceso, pero, además, tiene que comportarse de manera estacionaria. En caso contrario, corremos un riesgo muy alto de producir inferencias espurias a través de una subestimación sistemática del efecto de nuestra variable de interés.

Figura 4



Fuente: elaboración propia

Tabla 2
Resultados de la prueba de hipótesis con la variable de tratamiento no estacionaria

Estimación promedio	
0.0219	
Intervalo de confianza al 95%	
0.0207	0.2315
Valor del estadístico t	
t=-0.3212	
P-valor	
p=0.0000	

Fuente: elaboración propia

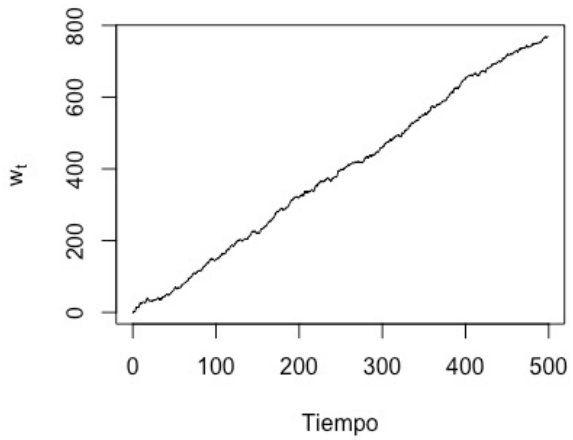
– Simulación 3: y_t no estacionaria y x_t estacionaria

Para este escenario se realiza una modificación a la variable y_t ; ahora será esta la que se comporte como un camino aleatorio y la variable x_t lo hará de manera estacionaria. Esto implica que el proceso generador de datos de la variable y_t modificada, a la que llamaré w_t , es el siguiente:

$$w_t = 1.5 + 1.0\, y_{t-1} + 2x_t + \eta_t \text{ , en donde } \eta_t \sim N(0,1) \quad (6)$$

En la Figura 5 se muestra el comportamiento de w_t , la cual es un camino aleatorio con una fuerte tendencia positiva. Es evidente que no se trata de una variable estacionaria al no tener un regreso a la media. Los resultados de esta simulación sugieren que el tamaño del sesgo en los estimadores de β_1 también será un problema, ya que el promedio de la distribución muestral fue de 2.7370 y el intervalo al 95% de confianza no incluye el valor de dos, por lo que nuevamente estaríamos produciendo una inferencia espuria. Esto implica que la sobre estimación tiene resultados sumamente perjudiciales al alejarse tanto del verdadero valor especificado en el proceso generador de datos.

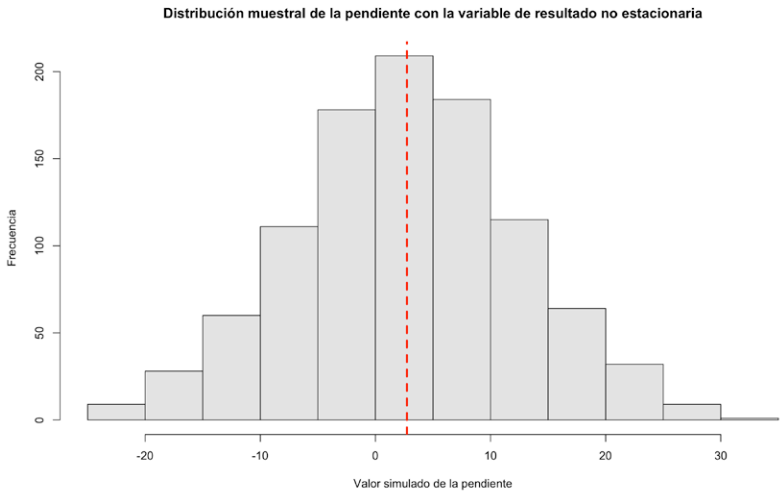
Figura 5
Variable de resultado modificada



Fuente: elaboración propia

En la Figura 6 puede observarse el comportamiento de la distribución muestral simulada bajo estas condiciones, en donde la línea roja indica el promedio. En la misma figura llama la atención algo adicional a la media: tiene una varianza muy amplia. Esto cobra relevancia al compararse con los resultados de la primera simulación (Figura 3). Cuando ambas variables son estacionarias, la desviación estándar de la variable aleatoria que se forma con los estimadores de la pendiente es de 0.0457, mientras que la desviación estándar en la simulación tres es de 9.7415. Esto implica que, además de producir estimadores sesgados, son menos eficientes, en virtud de tener una mayor dispersión. Esto queda claro en la propia Figura 6, ya que el rango de valores es más amplio que en el histograma de la Figura 3. Al igual que con las simulaciones previas, se realizó una prueba de hipótesis bajo las mismas condiciones. Los hallazgos se presentan en la Tabla 3. Es claro que el intervalo al 95% de confianza no incluye el verdadero valor y que es posible rechazar la hipótesis nula en la que se especifica que el promedio es igual a dos.

Figura 6



Fuente: elaboración propia

Tabla 3

Resultados de la prueba de hipótesis con la variable de resultado no estacionaria

Estimación promedio	
2.7370	
Intervalo de confianza al 95%	
2.1324	3.3415
Valor del estadístico t	
t=2.3923	
P-valor	
p=0.0169	

Fuente: elaboración propia

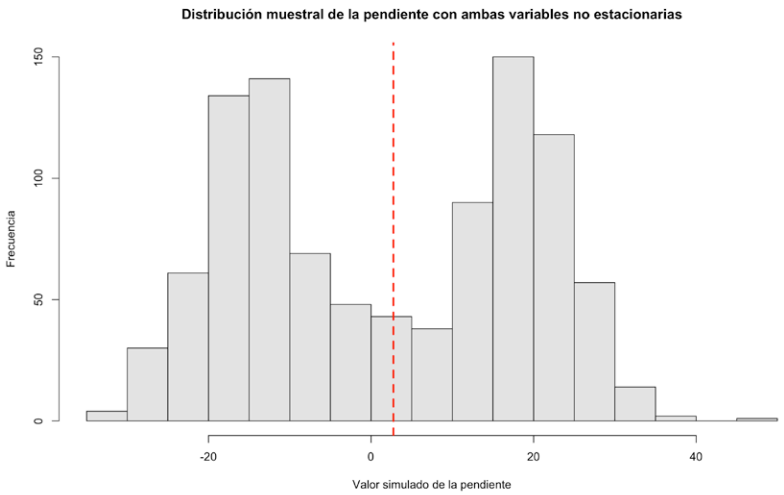
– Simulación 4: w_t y z_t no estacionarias

En el último caso se simularon los modelos de series de tiempo cuando ambas variables tienen raíz unitaria. Los resultados de este escenario son sorprendentes, ya que, como se muestra en la Figura 7, la distribución muestral de los estimadores de β_1 bajo estas condiciones no se comportan de manera normal. De hecho, la distribución es bimodal, lo cual puede ocasionar problemas al momento de generar inferencias.

El promedio de esta distribución es de 1.8359, lo cual es una ligera subestimación si se compara con el valor especificado en el proceso generador de datos. En esta ocasión el intervalo de confianza sí contiene el verdadero valor de dos. No obstante, preocupa tanto la forma de la distribución como la dispersión de las observaciones. En esta ocasión, la desviación estándar es de 17.3995, por lo que los estimadores son incluso menos eficientes que en la tercera simulación.

En la Tabla 4 se presenta la prueba de hipótesis correspondiente a esta simulación. Nuevamente, el intervalo de confianza al 95% incluye el verdadero valor, por lo que no es posible rechazar la hipótesis nula de que el promedio es igual a dos, lo cual es coincidente con el proceso generador de datos. No obstante, la distribución bimodal y la dispersión de las observaciones permite afirmar que los estimadores que se producen bajo estas condiciones no serán adecuados para producir inferencias correctas.

Figura 7



Fuente: elaboración propia

Tabla 4

Resultados de la prueba de hipótesis ambas variables no estacionarias

Estimación promedio	
1.8359	
Intervalo de confianza al 95%	
0.7562	2.9157
Valor del estadístico t	
t=-0.2982	
P-valor	
p=0.7656	

Fuente: elaboración propia

En la Tabla 5 se muestran los estadísticos descriptivos de las variables aleatorias de los estimadores de la pendiente. Al observar todos los resultados en conjunto queda de manifiesto que la falta de estacionariedad produce serios problemas para las inferencias. No importa en qué variable se encuentre, si hay raíces unitarias, necesitamos dar cuenta de ellas y tratar de modificar nuestras variables. Es por esto por lo que, en la siguiente sección abordo algunas estrategias para lidiar con la no estacionariedad de las series de tiempo.

Tabla 5

Tipo de simulación	Promedio	Mediana	Desviación estándar	Mínimo	Máximo
Primera simulación (ambas variables estacionarias)	1.9992	1.9990	0.0457	1.8520	2.1310
Segunda simulación (variable de resultado estacionaria y variable de tratamiento no estacionaria)	0.0219	0.0176	0.0195	-0.0137	0.1090
Tercera simulación (variable de resultado no estacionaria y variable de tratamiento estacionaria)	2.7370	2.8520	9.7415	-24.8570	32.4030
Cuarta simulación (ambas no estacionarias)	1.8359	1.5040	17.3995	-33.2460	45.0530

Fuente: elaboración propia

¿Qué hacer para evitar inferencias incorrectas?

Como se mostró en la sección anterior, la utilización de variables no estacionarias en modelos de series de tiempo genera diferentes problemas, los cuales dificultan la producción de inferencias correctas, ya que podemos introducir sesgo o bien reducir drásticamente la eficiencia de los estimadores. Por lo tanto, es necesario presentar algunos consejos, así como discutir cuáles podrían ser algunas soluciones a estos problemas. Una vez más, el objetivo es que se puedan generar inferencias más robustas a través de la utilización correcta de los modelos para series de tiempo.

En primer lugar, es necesario mencionar que nada sustituye los diagnósticos previos de nuestras variables. Si trabajamos con series de tiempo, debemos analizar el comportamiento de las variables para identificar procesos estacionarios o no estacionarios. En este sentido, se puede proceder de manera visual, ya que habrá ciertas variables en las que se muestre de manera evidente si hay o no estacionariedad. No obstante, podemos ser más precisos y realizar pruebas de hipótesis Dickey-Fuller, en las que la hipótesis nula es que la variable analizada tiene una raíz unitaria y la hipótesis alternativa es que la variable se comporta como un proceso estacionario.

Una vez que tenemos una idea de la forma en la que se comportan nuestras variables, y si encontramos que alguna no es estacionaria, podemos modificarla para intentar que se comporte como un proceso estacionario. Para realizar esto, se puede obtener la primera diferencia, la cual se define como $\Delta = y_t - y_{t-1}$. Este procedimiento podría ayudarnos a evitar los problemas que se identificaron en la sección previa. No obstante, hay que considerar que una modificación en las

variables puede tener repercusiones en el análisis, por lo que, la interpretación de una variable no estacionaria que se diferenció no es idéntica a una que no se modificó, especialmente a nivel teórico.

Asimismo, existe la posibilidad de que la no estacionariedad de una variable de tratamiento sea causada por una tendencia. Esto quiere decir que la variable puede ser estacionaria si se controla por el nivel de inclinación de ésta. Para diagnosticar esta situación existe la prueba Dickey-Fuller para tendencia, por lo que, si no se puede rechazar la hipótesis nula de raíz unitaria con tendencia, es posible que este problema se corrija una vez que introduzcamos una variable de control con la tendencia en la especificación del modelo.

Otra posibilidad es incorporar las propias dinámicas temporales en los modelos. Todos los experimentos que se discutieron en la sección anterior se sustentan en modelos estáticos, es decir, en modelos en los que se asume que la variable de tratamiento solo tiene un efecto en términos contemporáneos en la variable de resultado. No obstante, existen variables cuyos valores previos podrían afectar los valores contemporáneos de otras. La decisión de incluir estas dinámicas en los modelos es eminentemente teórica, pero hay que considerar las posibles repercusiones empíricas. De igual forma, se podría incorporar un término autorregresivo (y_{t-1}) del lado derecho de la ecuación. Esto implica que consideramos a los valores previos de nuestra variable de resultado como un control más, por lo que el término autorregresivo tiene que afectar tanto a la variable de resultado en términos contemporáneos como a nuestra o nuestras variables de tratamiento.

Por último, es importante que se utilicen los modelos apropiados para cada situación.[5] Esta decisión depende completamente de la forma en la que se comportan nuestras variables y de las posibles dinámicas temporales. Necesitamos transitar más allá de los modelos estáticos, porque las series de tiempo, por naturaleza, pueden tener efectos de largo plazo, los cuales se pueden modelar para generar inferencias más robustas.

Conclusiones

Las series de tiempo son una herramienta fundamental para el análisis empírico de la política. No obstante, en muchas ocasiones, no se consideran las implicaciones técnicas de incorporar variables que tienen un comportamiento no estacionario. Tal como se mostró en este artículo, la no estacionariedad puede traer consigo problemas para nuestras inferencias.

Los resultados de las simulaciones Monte Carlo arrojan que posiblemente el caso más grave es intentar explicar una variable de resultado estacionaria a través de una variable de tratamiento no estacionaria, ya que se introduce un sesgo de tal dimensión que nos conduciría a subestimar sistemáticamente los efectos de interés. De la misma forma, la segunda simulación muestra que el sesgo (positivo en esta ocasión) también será sustancial si se presenta una variable de resultado con raíz unitaria y una variable de tratamiento estacionaria. Las simulaciones tres y cuatro mostraron, además, problemas en términos de eficiencia de los estimadores e incluso de falta de normalidad en la distribución muestral (simulación cuatro). Todos estos problemas en su

[5] Excelentes consejos para encontrar los modelos adecuados en función del comportamiento de nuestras variables se encuentran en Philips (2018, 2021).

conjunto sugieren que necesitamos realizar los diagnósticos apropiados a las variables, así como utilizar los modelos que nos ayuden a producir las mejores inferencias posibles.

Si bien es cierto que los experimentos que aquí se muestran se refieren solo a series de tiempo, los hallazgos pueden generalizarse hacia los modelos para datos panel. Dichas especificaciones pueden agrupar diversas series de tiempo en un mismo análisis, lo que puede traer retos adicionales a los aquí presentados. No obstante, los problemas de estacionariedad no se solucionan al trabajar con series de tiempo agrupadas, al contrario, pueden exacerbarse.

La intención de este artículo es exponer cuáles son algunos de los problemas a los que nos enfrentamos al analizar variables que cambian a lo largo del tiempo, así como ciertas estrategias para solucionarlos. Elementos como los que aquí se presentan, forman parte de los cursos estadísticos de algunas instituciones de educación superior en ciencia política, tanto en México como en América Latina. No obstante, poco se discute en la academia de habla hispana respecto a las consecuencias de una modelación empírica deficiente. Es por eso por lo que necesitamos generar más discusiones metodológicas en español que puedan ser de utilidad tanto para profesionistas como para estudiantes.

Bibliografía

Acemoglu, Daron, y James A. Robinson. (2005). *Economic Origins of Dictatorship and Democracy*. New York: Cambridge University Press.

Alizada, Nazifa, Rowan Cole, Lisa Gastaldi, Sebastian Hellmeier, Palina Kolvani, Jean Lachapelle, Anna Lührmann, Seraphine F. Maerz, Shreeya Pillai, y Staffan I. Lindberg. (2021). *Autocratization Turns Viral. Democracy Report 2021*. University of Gothenburg: V-Dem Institute.

Ansell, Ben W., y David Samuells. (2011). Inequality and Democratization: Individual-Level Evidence of Preferences for Redistribution under Autocracy. Vol. Annual Meeting of the American Political Science Association. Seattle, Washington: American Political Science Association.

Atilano Robles, Edwin. (2022). Cooperación legislativa entre oposición y gobierno en México. Un análisis de series de tiempo. *Revista Perfiles Latinoamericanos*. 30 (59), 1–30.

Beck, Nathaniel, y Jonathan N. Katz. (1995). What to do (and not to do) with Time-Series Cross-Section Data. *The American Political Science Review*. 89 (3), 634–647. https://doi.org/10.2307/2082979.

Beck Nathaniel (2011). Modeling Dynamics in Time-Series–Cross-Section Political Economy Data. *Annual Review of Political Science*. 14 (1), 331–52. https://doi.org/10.1146/annurev-polisci-071510-103222.

Canes-Wrone, Brandice, y Scott De Marchi. (2002). Presidential Approval and Legislative Success. *Journal of Politics*. 64 (2), 491–509. https://doi.org/10.1111/1468-2508.00136.

Castro Cornejo, Rodrigo. (2021). How Do Campaigns Matter? Independents, Political Information, and the Enlightening Role of Campaigns in Mexico. *International Journal of Public Opinion Research*. 33 (4), 779–798. https://doi.org/10.1093/ijpor/edaa029.

Cryer, Jonathan D. (1986). *Time series analysis*. Vol. 286. Springer.

Das, Panchanan. (2019). Time Series: Data Generating Process. *Econometrics in Theory and Practice: Analysis of Cross Section, Time Series and Panel Data with Stata 15.1*, editado por Panchanan Das, 247–259. Singapore: Springer. https://doi.org/10.1007/978-981-32-9019-8_9.

De Boef, Suzanna, y Luke Keele. (2008). Taking Time Seriously. *American Journal of Political Science*. 52 (1), 184–200. https://doi.org/10.1111/j.1540-5907.2007.00307.x.

Geddes, Barbara, Joseph Wright, y Erica Frantz. (2018). *How Dictatorships Work: Power, Personalization, and Collapse*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781316336182.

Gelman, Andrew, y Jennifer Hill. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Analytical Methods for Social Research. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511790942.

Gujarati, Damodar N. (2006). *Principios de econometría*. España: McGraw-Hill/Interamericana de España.

Huntington-Klein, Nick. (2022). The Effect: An Introduction to Research Design and Causality. Routledge & CRC Press.

Iversen, Torben, y David Soskice. (2006). Electoral Institutions and the Politics of Coalitions: Why Some Democracies Redistribute More than Others. *American Political Science Review.* 100 (2), 165–81.

Iversen, Torben (2011). Inequality and Redistribution: A Unified Approach to the Role of Economic and Political Institutions. *Revue économique.* 62 (4), 629–49.

Keele, Luke. (2020) Differences-in-Differences: Neither Natural nor an Experiment. *The SAGE Handbook of Research Methods in Political Science and International Relations*. Luigi Curini y Robert Franzese (eds), 822–34. SAGE Publications Ltd.

Kellstedt, Paul M., y Guy D. Whitten. (2018). *The Fundamentals of Political Science Research*. Cambridge: Cambridge University Press. https://doi.org/10.1017/9781108131704.

King, Gary, Robert O. Kehoane, y Sidney Verba. (1994). *Designing Social Inquiry. Scientific Inference in Qualitative Research*. Princeton, New Jersey: Princeton University Press.

Lucardi, Adrián. (2019). Strength in Expectation: Elections, Economic Performance, and Authoritarian Breakdown. *The Journal of Politics.* 81 (2), 552–70. https://doi.org/10.1086/701723.

Mauricio, José Alberto. (2007). Análisis de series temporales. España. Universidad Complutense de Madrid.

McCann, James A., y Chappell Lawson. (2003). An Electorate Adrift? Public Opinion and the Quality of Democracy in Mexico. *Latin American Research Review.* 38 (3), 60–81.

Morgan, Stephen L., y Christopher Winship. (2014). *Counterfactuals and Causal Inference: Methods and Principles for Social Research*. Analytical Methods for Social Research. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9781107587991.

Nava, Alejandro. (2015). *Procesamiento de series de tiempo*. México: Fondo de Cultura Economica.

Newman, Brian. (2003). Integrity and Presidential Approval, 1980–2000. *Public Opinion Quarterly.* 67 (3), 335–67. https://doi.org/10.1086/377242.

Pérez Ramírez, Fredy O. (2007). *Introducción a las series de tiempo. Métodos paramétricos*. Colombia: Universidad de Medellín.

Philips, Andrew Q. (2018). Have Your Cake and Eat It Too? Cointegration and Dynamic Inference from Autoregressive Distributed Lag Models. *American Journal of Political Science* 62 (1), 230–44. https://doi.org/10.1111/ajps.12318.

Philips, Andrew Q. (2021). How to Avoid Incorrect Inferences (While Gaining Correct Ones) in Dynamic Models. *Political Science Research and Methods.* julio, 1–11. https://doi.org/10.1017/psrm.2021.31.

Philips, Andrew Q., Amanda Rutherford, y Guy D. Whitten. (2016). Dynamic Pie: A Strategy for Modeling Trade-Offs in Compositional Variables over Time. *American Journal of Political Science.* 60 (1), 268–83.

Pickup, Mark. (2014). *Introduction to Time Series Analysis*. SAGE Publications.

Przeworski, Adam, Fernando Limongi, José Antonio Cheibub, y Michael E. Álvarez. (2000). *Democracy and Development: Political institutions and Material Well-Being in the World, 1950-1990*. Cambridge: Cambridge University Press.

Reeskens, Tim, Quita Muis, Inge Sieben, Leen Vandecasteele, Ruud Luijkx, y Loek Halman. (2021). Stability or change of public opinion and values during the coronavirus crisis? Exploring Dutch longitudinal panel data. *European Societies*. 23 (sup1), S153–71. https://doi.org/10.1080/14616696.2020.1821075.

Ross, Michael. (2006). Is Democracy Good for the Poor? *American Journal of Political Science*. 50 (4), 860–874.

Skaaning, Svend-Erik, John Gerring, y Henrikas Bartusevičius. (2015). A Lexical Index of Electoral Democracy. *Comparative Political Studies*. 48 (12), 1491–1525. https://doi.org/10.1177/0010414015581050.

Stock, James H, Mark W Watson, y Raúl Sánchez Larrión. (2012). *Introducción a la Econometría*.

Svolik, Milan W. (2012). *The Politics of Authoritarian Rule*. Cambridge: Cambridge University Press.

Toshkov, Dimiter. (2016). *Research Design in Political Science*. Political Analysis. London, UK: Palgrave Macmillan.

Ugues, Antonio. (2018). Public Perceptions of Clean Elections in Mexico: An Analysis of the 2000, 2006, and 2012 Elections. *Journal of Politics in Latin America*. 10 (2), 77–98. https://doi.org/10.1177/1866802X1801000203.

Villavicencio, Jhon. (2010). Introducción a series de tiempo. *Puerto Rico*.

Williams, L. K., y G. D. Whitten. (2011). Dynamic simulations of autoregressive relationships. *Stata Journal*. 11 (4), 577–88.

Wooldridge, Jeffrey M. (2012). *Introductory econometrics: a modern approach*. Fifth edition. Ohio: South-Western Cengage Learning.

Wooldridge, Jeffrey M., (1960-2001). *Econometric Analysis of Cross Section and Panel Data*. MIT Press.

Wright, Joseph. (2008). Do Authoritarian Institutions Constrain? How Legislatures Affect Economic Growth and Investment. *American Journal of Political Science*. 52 (2), 322–343. https://doi.org/10.2307/25193816.

Anexo. Código para replicar los hallazgos en *R*

```
# Directorio de trabajo
setwd("/directorio/propio/aquí/")

# Librerías que se necesitan
library(pacman)
library(tidyverse)
# Limpiar el ambiente
rm(list=ls())
# Parametros y semilla
set.seed(12345)  # Semilla para replicar
beta_0 = 1.5  # Intercepto
beta_1 = 2.0  # Pendiente
n = 500  # Tamaño de la muestra
M = 1000  # Numero de experimentos

# Vectores para almacenar
intercepto <- rep(0, M)
pendiente <- rep(0, M)
```

```
intercepto_2 <- rep(0, M)
pendiente_2 <- rep(0, M)

intercepto_3 <- rep(0, M)
pendiente_3 <- rep(0, M)

intercepto_4 <- rep(0, M)
pendiente_4 <- rep(0, M)
# Simulaciones Monte Carlo
for (i in 1:M){ # M es el número de iteraciones

# Generar datos
  U_i = rnorm(n, mean = 0, sd = 1) # Error 1
  E_i = rnorm(n, mean = 0, sd = 1) # Error 2
 X_i = rnorm(n, mean = 0, sd = 1) # Variable de tratamiento est.
 Y_i = beta_0 + beta_1*X_i + U_i # Variable de resultado est.
Y_t = cumsum(Y_i) + beta_1*X_i + E_i # Variable de resultado no est.
  X_t = cumsum(X_i) # Variable de tratamiento no est.
  time = (0:499) # Tiempo

# Crear un data.frame
data_i = data.frame(Y = Y_i, Yt = Y_t, X = X_i, Xt = X_t, time = time)

# Modelo 1
  ols_i <- lm(data = data_i, Y ~ X)

# Ambas variables estacionarias
  pendiente[i] <- ols_i$coefficients[2]
  intercepto[i] <- ols_i$coefficients[1]

coeficientes_est <- data.frame(int = intercepto, pend = pendiente)

# Modelo 2
  ols_ii <- lm(data = data_i, Y ~ Xt)

# X con raiz unitaria
  pendiente_2[i] <- ols_ii$coefficients[2]
  intercepto_2[i] <- ols_ii$coefficients[1]

coeficientes_est2 <- data.frame(int = intercepto_2, pend = pendiente_2)

# Modelo 3
  ols_iii <- lm(data = data_i, Yt ~ X)
```

```
                    # Y con raiz unitaria
                    pendiente_3[i] <- ols_iii$coefficients[2]
                    intercepto_3[i] <- ols_iii$coefficients[1]

    coeficientes_est3 <- data.frame(int = intercepto_3, pend = pendiente_3)

                                # Modelo 4
            ols_iv <- lm(data = data_i, Yt ~ Xt)

                    # Ambas con raiz unitaria
                    pendiente_4[i] <- ols_iv$coefficients[2]
                    intercepto_4[i] <- ols_iv$coefficients[1]

    coeficientes_est4 <- data.frame(int = intercepto_4, pend = pendiente_4)

}
# Figura 1
plot(data_i$time, data_i$Y, type = "l",
                main = "Ejemplo de una variable estacionaria",
                                xlab = "Tiempo",
  ylab = expression(y[t]))

# Figura 2
plot(data_i$time, data_i$Xt, type = "l",
                main = "Ejemplo de una no variable estacionaria",
                                xlab = "Tiempo",
  ylab = expression(z[t]))

#Figura 3
hist(coeficientes_est$pend, breaks = 20, freq = T,
      main = "Distribución muestral de la pendiente con ambas variables estacionarias",
                xlab = "Valor simulado de la pendiente",
                                ylab = "Frecuencia",
                                col = "gray90")
abline(v = 1.9992, col = "red", lty = 2, lwd = 2.5)

# Figura 4
hist(coeficientes_est2$pend, breaks = 20, freq = T,
  main = "Distribución muestral de la pendente con la variable de tratamiento no estacionaria",
                xlab = "Valor simulado de la pendiente",
                                ylab = "Frecuencia",
                                col = "gray90")
abline(v = 0.0219, col = "red", lty = 2, lwd = 2.5)
```

```
# Figura 5
plot(data_i$time, data_i$Yt, type = "l",
     main = "Variable de resultado modificada",
     xlab = "Tiempo",
     ylab = expression(w[t]))

# Figura 6
hist(coeficientes_est3$pend, breaks = 20, freq = T,
     main = "Distribución muestral de la pendiente con la variable de resultado no estacionaria",
     xlab = "Valor simulado de la pendiente",
     ylab = "Frecuencia",
     col = "gray90")
abline(v = 2.7370, col = "red", lty = 2, lwd = 2.5)

# Figura 7
hist(coeficientes_est4$pend, breaks = 20, freq = T,
     main = "Distribución muestral de la pendiente con ambas variables no estacionarias",
     xlab = "Valor simulado de la pendiente",
     ylab = "Frecuencia",
     col = "gray90")
abline(v = 1.8359, col = "red", lty = 2, lwd = 2.5)

# Pruebas de hipótesis

# Tabla 1
t.test(coeficientes_est$pend, alternative = c("two.sided"), conf.level = 0.95, mu = 2)
#
##
#
##
##  One Sample t-test
#
#
##  data:  coeficientes_est$pend
##  t = -1.03, df = 999, p-value = 0.3033
##  alternative hypothesis: true mean is not equal to 2
##  95 percent confidence interval:
##  1.995841 2.001296
##  sample estimates:
##  mean of x
##  1.998568
# Tabla 2
t.test(coeficientes_est2$pend, alternative = c("two.sided"), conf.level = 0.95, mu = 2)
#
##
#
##
##  One Sample t-test
#
#
##  data:  coeficientes_est2$pend
##  t = -3238.6, df = 999, p-value < 2.2e-16
##  alternative hypothesis: true mean is not equal to 2
```

```
##  95 percent confidence interval:
##  0.01985766 0.02225580
##  sample estimates:
##  mean of x
## 0.02105673
# Tabla 3
t.test(coeficientes_est3$pend, alternative = c("two.sided"), conf.level = 0.95, mu = 2)
# #
##  One Sample t-test
# #
##  data: coeficientes_est3$pend
##  t = 2.3923, df = 999, p-value = 0.01693
##  alternative hypothesis: true mean is not equal to 2
##  95 percent confidence interval:
##  2.132447 3.341457
##  sample estimates:
##  mean of x
##  2.736952
# Tabla 4
t.test(coeficientes_est4$pend, alternative = c("two.sided"), conf.level = 0.95, mu = 2)
# #
##  One Sample t-test
# #
##  data: coeficientes_est4$pend
##  t = -0.29816, df = 999, p-value = 0.7656
##  alternative hypothesis: true mean is not equal to 2
##  95 percent confidence interval:
##  0.7562282 2.9156654
##  sample estimates:
##  mean of x
##  1.835947
# Estadísticos descriptivos de la Tabla 5
summary(coeficientes_est$pend)
##  Min. 1st Qu. Median Mean 3rd Qu. Max.
##  1.831 1.968 1.998 1.999 2.028 2.140
summary(coeficientes_est2$pend)
##  Min. 1st Qu. Median Mean 3rd Qu. Max.
## -0.016382 0.007403 0.016487 0.021057 0.030511 0.112686
summary(coeficientes_est3$pend)
##  Min. 1st Qu. Median Mean 3rd Qu. Max.
## -24.857 -3.594 2.852 2.737 9.048 32.403
summary(coeficientes_est4$pend)
##  Min. 1st Qu. Median Mean 3rd Qu. Max.
## -33.246 -14.448 1.504 1.836 18.139 45.053
```