
Munich Personal RePEc Archive

Intermediate economics: Theory and applications

Ramon Antonio, Rosales Alvarez and Jorge Andres, Perdomo Calvo and Carlos Andres, Morales Torrado and Jaime Alejandro, Urrego Mondragon

Facultad de Economia, Universidad de Los Andes

January 2009

Online at https://mpra.ub.uni-muenchen.de/37183/
MPRA Paper No. 37183, posted 08 Mar 2012 15:50 UTC

APUNTES DE CLASE CEDE

ISSN 1909-4442

FUNDAMENTOS DE ECONOMETRÍA INTERMEDIA: TEORÍA Y APLICACIONES

Ramón Antonio Rosales Álvarez
Jorge Andrés Perdomo Calvo
Carlos Andrés Morales Torrado
Jaime Alejandro Urrego Mondragón

1

ENERO DE
2010

Universidad de los Andes
Facultad de Economía

CEDE
Centro de Estudios
sobre Desarrollo Económico

Serie Apuntes de clase Cede, 2010-1
ISSN 1909-4442

Enero de 2010

© 2010, Universidad de los Andes–Facultad de Economía–Cede
Calle 19 A No. 1-37, Bloque W
Bogotá, D. C., Colombia
Teléfonos: 3394949- 3394999, extensiones 2400, 2049, 3233
infocede@uniandes.edu.co
http://economia.uniandes.edu.co

Ediciones Uniandes
Carrera 1ª No. 19 – 27, edificio Aulas 6, A. A. 4976
Bogotá, D. C., Colombia
Teléfonos: 3394949- 3394999, extensión 2133, Fax: extensión 2158
infeduni@uniandes.edu.co
http://ediciones.uniandes.edu.co/

Edición, diseño de cubierta, preprensa y prensa digital:
Proceditor ltda.
Calle 1C No. 27 A – 01
Bogotá, D. C., Colombia
Teléfonos: 2204275, 220 4276, Fax: extensión 102
proceditor@etb.net.co

Impreso en Colombia – Printed in Colombia

El contenido de la presente publicación se encuentra protegido por las normas internacionales y nacionales vigentes sobre propiedad intelectual, por tanto su utilización, reproducción, comunicación pública, transformación, distribución, alquiler, préstamo público e importación, total o parcial, en todo o en parte, en formato impreso, digital o en cualquier formato conocido o por conocer, se encuentran prohibidos, y sólo serán lícitos en la medida en que se cuente con la autorización previa y expresa por escrito del autor o titular. Las limitaciones y excepciones al Derecho de Autor, sólo serán aplicables en la medida en que se den dentro de los denominados Usos Honrados (Fair use), estén previa y expresamente establecidas; no causen un grave e injustificado perjuicio a los intereses legítimos del autor o titular, y no atenten contra la normal explotación de la obra.

UNIVERSIDAD DE LOS ANDES - CEDE

FUNDAMENTOS DE ECONOMETRÍA INTERMEDIA: TEORÍA Y APLICACIONES

Ramón A. Rosales A.
Jorge A. Perdomo C.
Carlos A. Morales T.
Jaime A. Urrego M.

Enero de 2010

Agradecimientos

Los autores expresan sus agradecimientos al Centro de Estudios sobre Desarrollo Económico (CEDE) de la Facultad de Economía de la Universidad de los Andes, por el apoyo financiero para la elaboración y publicación de este documento. Asimismo, a Fabio Sánchez, María del Pilar López, Antonella Fazio, Raquel Bernal, Camilo Bohórquez, Juan Carlos Vásquez, Catherine Rodríguez, Armando Armenta y Gustavo García por facilitar los datos empleados para los estudios de caso. Por último, a todos los profesores del área de econometría y estudiantes que han tomado los cursos de econometría I, II y avanzada en el pregrado y postgrado en economía de la universidad, cuyos aportes han contribuido en la elaboración de este documento.

Los autores
Enero de 2010

Los Autores

Ramón A. Rosales Álvarez Ph.D. en Economía Agrícola de la Universidad de Oklahoma (EEUU). Actualmente se desempeña como director del Centro de Estudios Ganaderos y Agropecuarios (CEGA), y coordinador del área de econometría en la facultad de Economía de la Universidad de los Andes (Colombia).

Jorge A. Perdomo Calvo M.Sc. en Economía del Medio Ambiente y de los Recursos Naturales de la Universidad de Maryland College Park (EEUU) y la Universidad de los Andes (Colombia). Actualmente se desempeña como profesor en las áreas de econometría y economía del transporte en esta última universidad, así como investigador del Centro de Estudios sobre Desarrollo Económico (CEDE).

Carlos A. Morales Torrado M.Sc. en Economía de la Universidad de los Andes (Colombia). Actualmente, se desempeña como profesor en las áreas de econometría y programación en dicha universidad, así como investigador del Centro de Estudios sobre Desarrollo Económico (CEDE).

Jaime A. Urrego Mondragón economista y M.Sc. en Economía de la Universidad de los Andes (Colombia). Actualmente, se desempeña como profesor en el área de econometría en dicha universidad, así como investigador del Centro de Estudios sobre Desarrollo Económico (CEDE), y del Departamento Administrativo Nacional de Estadística (DANE).

FUNDAMENTOS DE ECONOMETRÍA INTERMEDIA: TEORÍA Y APLICACIONES

Ramón A. Rosales Álvarez

Jorge A. Perdomo Calvo

Carlos A. Morales Torrado

Jaime A. Urrego Mondragón

Resumen

La econometría es conjunto de métodos estadísticos inferenciales para el tratamiento cuantitativo de la información económica, que permite apoyar el estudio sobre algunos campos especiales de la economía y los negocios. Este documento presenta los fundamentos intermedios de esta área de estudio, para estudiantes con un conocimiento previo sobre los temas tratados en econometría básica. Particularmente se tratan los temas de endogeneidad, ecuaciones simultáneas, series de tiempo y datos panel. Un aporte importante de estas notas de clase es presentar la teoría junto a ejemplos aplicados, desarrollados con el programa econométrico especializado Stata®.

Palabras Claves: Econometría, modelos de Sección Cruzada, modelos de ecuaciones simultáneas, modelos Probabilísticos, series de tiempo, Modelos con Panel de datos, enseñanza universitaria, manuales.

Clasificación JEL: C01, C21, C3, C25, C22, C23 A23, A33

---

*Esta versión: Enero de 2010.

INTERMEDIATE ECONOMETRICS: THEORY AND APPLICATIONS.

Ramón A. Rosales Álvarez

Jorge A. Perdomo Calvo

Carlos A. Morales Torrado

Jaime A. Urrego Mondragón

Abstract

Econometrics is the area of statistics concerned in analyzing economic data, for both economic and business applications. This document, introduces the intermediate concepts of this area, for students already familiarized with basic econometric theory. In particular, topics concerning endogenity, simultaneous equation models, time series and panel data, are discussed. One special contribution of these class notes is that both theory and applications, using Stata® statistical software package, are developed.

Key words: Econometrics, Cross-Sectional Models, Simultaneous Equation Models, Discrete Regression and Qualitative Choice Models, Time Series Models, Models with Panel Data, Undergraduate Economics Education, Handbooks.

JEL Classification: C01, C21, C3, C25, C22, C23 A23, A33

---

*This version: Janaury 2010.

Contenido

Introducción

La econometría es un conjunto de métodos estadísticos inferenciales para el tratamiento cuantitativo de la información económica, que permite apoyar el estudio sobre algunos campos especiales de la economía y los negocios, destacando entre ellos el estudio de decisiones en producción, demanda, oferta, inversión, entre otras.

También la econometría, además de proporcionar una metodología de trabajo, es una disciplina auxiliar del economista, porque permite contar con un instrumento de análisis en múltiples áreas de aplicación; útil para el trabajo profesional. Por esta razón, los estudiantes e interesados en el tema deben familiarizarse desde la formación básica; este libro les permitirá tener un contacto con los fundamentos intermedios de esta área de estudio.

De acuerdo a lo anterior, antes de iniciar la lectura del libro, el lector debe contar con un conocimiento previo sobre los temas tratados en econometría básica[1], para comprender y familiarizarse con su contexto, debido a que los temas aquí comprendidos suponen previos conocimientos sobre ellos.

Teniendo en cuenta la descripción anterior, el objetivo principal del texto es proveer las diferentes teorías y metodologías de manera sencilla, para estudiar los temas relacionados en un curso de econometría intermedia. Un aporte importante del libro es presentar la teoría y ejemplos aplicados, los cuales fueron desarrollados con el programa econométrico especializado Stata®.

Para abordar el tema de econometría intermedia, el libro se encuentra divido en ocho capítulos de la siguiente manera: capítulo 1, discute lo relacionado con problemas de especificación por omitir variables independientes, formas funcionales incorrectas y endogenidad. Adicionalmente considera sus métodos de corrección entre los que se destacan: variable proxy e instrumental con mínimos cuadrados en dos etapas.

[1] *Véase* Rosales y Bonilla (2006).

El capítulo 2, presenta el tema endogenidad causada por simultaneidad las metodologías de mínimos cuadrados indirectos (MCI), dos y tres etapas (MC2E y MC3E, así como la aplicación de regresiones aparentemente no relacionadas (SUR, seemingly unrelated regression, siglas en inglés). El capítulo 3, contiene aspectos sobre los modelos probabilísticos con variables de respuesta dicótomas (lineales, logit y probit); y sus respectivas estimaciones mediante máxima verosimilitud (MV).

El capítulo 4, comprende la introducción a los conceptos sobre series de tiempo con el fin de proyectar variables dinámicas, procedimiento y aplicación del filtro Hodrick-Prescott, modelos de pronóstico con tendencia determinística y métodos de atenuación exponencial. El capítulo 5 continúa con técnicas de proyección univariadas, abordando todo lo relacionado con la metodología Box-Jenkins (prueba de raíz unitaria, series estacionarias, variables no estacionarias y estacionales).

El capítulo 6, reseña aspectos de series de tiempo con variables dependiente e independientes dinámicas, explorando los modelos autorregresivos, de rezagos distribuidos y expectativas adaptativas, causalidad de Granger y cointegración.

El capítulo 7, abarca lo relacionado con datos de corte transversal agrupados en el tiempo (pruebas de cambio estructural con el estadístico de Chow) y análisis diferencias en diferencias (para realizar evaluación de impacto de un proyecto o política). El capítulo 8, continúa con la relación estática y dinámica mediante panel de datos, estimación agrupada por mínimos cuadrados ordinarios, efectos fijos y aleatorios. Finalmente, se presenta el apéndice sobre los comando en Stata®; utilizados en cada tema a lo largo del documento.

Capítulo 1

Especificaciones inadecuadas de un modelo, variables independientes omitidas no cuantificables y uso de variable aproximada o proxy, endogeneidad y mínimos cuadrados en dos etapas.

1.1 Introducción

A partir de los aspectos abarcados en los cursos de econometría básica, este capítulo ofrece un acercamiento a conceptos, metodologías y prácticas econométricas dirigidas a dos casos particulares. El primero, cuando existe una especificación errónea del modelo y el segundo, bajo el incumplimiento del supuesto de independencia condicional, es decir, cuando se tiene correlación entre el error (e_i) y una o más variables independientes (X_{ij}).

Con este fin, en las siguientes secciones se plantea una discusión sobre el problema de especificación, para entender por qué en ocasiones no son obtenidos los resultados teóricos esperados. Adicionalmente, se analizan las pruebas de Ramsey-RESET, J de Davidson y MacKinnon y multiplicador de Lagrange que pretenden detectar este problema. Así, posteriormente proponer las principales metodologías de corrección.

Asimismo, se explican las causas que originan el incumplimiento del supuesto de independencia condicional –conocido como problema de endogeneidad-; omisión de variables no observables relevantes, doble causalidad, medición errónea de variables y problemas de muestreo. Debido a este problema, también es abierta la discusión sobre por qué los estimadores de mínimos cuadrados ordinarios (MCO) resultan sesgados e inconsistentes. Por esta razón, se introduce el uso de variables instrumentales mediante regresiones en dos etapas (MC2E), que recuperan las propiedades estadísticas (insesgados y consistentes) de los coeficientes. De igual

manera es presentada la prueba de Hausman para identificar endogeneidad en las estimaciones de MCO.

Finalmente, las metodologías expuestas son aplicadas mediante un estudio de caso basado en los datos del artículo de Sánchez, Fazio y Lopez-Uribe (2008), titulado (en inglés) "*Land Conflict, Property Rights, and the Rise of the Export Economy in Colombia, 1850-1925*", el cual pretende otorgar una explicación institucional de la baja integración entre la economía colombiana y el mercado internacional a finales del siglo XIX.

1.2 Discusión sobre especificación de modelos econométricos

Adicional al cumplimiento de los supuestos sobre homoscedasticidad, ausencia de multicolinealidad y autocorrelación residual, en las estimaciones por MCO; por otra parte, conviene especificar apropiadamente el modelo econométrico para obtener resultados apropiados. No obstante, esto último es infringido cuando se trabajan formas funcionales inadecuadas, omiten variables independientes relevantes o inclusión de variables explicativas redundantes. Lo anterior, conduce a sesgos en los estimadores y sus varianzas, generando una inapropiada relación entre variables dependiente e independientes (Gujarati, 2003, 491). A continuación son tratadas las consecuencias de estimar un modelo con especificación errónea y algunos métodos estadísticos para identificar y solucionar el problema.

1.2.1 Causas y consecuencias de especificación inadecuada

Con el fin de estudiar las causas y consecuencias de especificar erróneamente un modelo, considere la siguiente relación económica general (*véase* ecuación 1.1).

$$\begin{aligned} Y_i &= f(X_{i1}, X_{i2} \cdots, X_{ik}) \\ Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \varpi_i \end{aligned} \qquad (1.1)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + e_i \qquad (1.2)$$

En la ecuación 1.1, Y_i es la variable dependiente y X_{ij} [2] un conjunto de variables explicativas linealmente independientes. Esta especificación se conoce como función de regresión poblacional (FRP), donde ϖ_i corresponde a un elemento aleatorio[3]. Los modelos econométricos, pretenden estimar los coeficientes de la FRP, A partir de una expresión equivalente y muestra representativa (*véase* ecuación 1.2). Con estas ecuaciones, a continuación se exponen las causas que originan el problema de especificación y sus consecuencias.

1.2.1.1 Omisión de variables relevantes

Una posible causa mencionada sobre la especificación errónea del modelo, es omitir variables relevantes en una regresión muestral, como consecuencia de escasa disponibilidad de datos, incapacidad para su recolección o algún grado de desconocimiento sobre el planteamiento teórico previo.

Para formalizar lo anterior, a partir de la ecuación 1.2 se concibe un nuevo modelo a estimar con $k-1$ variables explicativas, es decir, omitiendo una variable independiente (*véase* ecuación 1.3).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i \qquad (1.3)$$

En la ecuación 1.3 aparece un nuevo término de error u_i que contiene la variable omitida.[4] Como consecuencia, el intercepto y pendientes ($\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_{k-1}$) son sesgados e inconsistente (*véase* demostración anexo 1), igualmente las varianzas estimadas de los coeficientes ($V(\hat{\beta}_0), V(\hat{\beta}_1), \cdots + V(\hat{\beta}_k)$)[5], invalidando sus intervalos de confianza, pruebas de significancia parcial y global (Gujarati, 2003, 491). Adicionalmente, si la variable omitida esta correlacionada con alguna variable

[2] Con $j = 1, 2, ..., k$

[3] O término de error, en forma de ruido blanco (*véase* capítulo cinco) , que captura todos los determinantes no observables e impredecibles de la variable dependiente Y_i. Para las expresiones poblacionales se utiliza ϖ_i, y para las muestrales e_i. Nuevos términos de error denotados como u_i, υ_i, μ_i y v_i aparecen cuando hay modificaciones sobre la especificación inicial.

[4] Formalmente, $u_i = \beta_k X_{ik} + e_i$, con e_i el termino de error del modelo correcto.

[5] Un estimador está sesgado si no se aproxima al valor poblacional que se desea estimar.

independiente de la ecuación 1.3, se genera otro problema denominado endogeneidad (*véase* sección 1.3).

1.2.1.2 Forma funcional incorrecta

Otra causa de especificación errónea sucede cuando se elige una forma funcional incorrecta, para expresar las variables independientes. Ahora, considere el siguiente modelo lineal con dos variables explicativas X_{i1} X_{i2}, donde X_{i1} explica a la variable dependiente como un polinomio de forma cuadrática (*véase* ecuación 1.4). Si equivocadamente es planteada una relación lineal, conlleva a un problema de especificación (*véase* ecuación 1.5).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1}^2 + \beta_3 X_{i2} + e_i \qquad (1.4)$$

$$Y_i = \beta_{i0} + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad (1.5)$$

Aunque las estimaciones MCO calculan correctamente cada uno de los coeficientes, omitir la forma cuadrática conduce a interpretaciones erróneas del efecto que tiene la variable X_{i1} sobre dependiente Y_i. La diferencia entre el coeficiente obtenido usando una forma funcional incorrecta, en relación al estimador poblacional, corresponde a un sesgo de especificación.[6]

1.2.1.3 Adición de variables independientes redundantes.

La última causa de especificación errónea, es adicionar variables innecesarias como consecuencia de un planteamiento teórico incorrecto; en este caso, considere un conjunto de regresoras adicionales (*véase* ecuación 1.6). Así, cada uno de los estimadores de MCO ($\hat{\beta}_{i-MCO}$) continúan insesgados pero dejan de ser eficientes, lo que aumenta la probabilidad de cometer un error de tipo II[7].

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + e_i \qquad (1.6)$$

---

[6]En este caso particular, el efecto marginal correcto de X_{i1} en Y_i es $\beta_1 + 2\beta_2$; el calculado erróneamente corresponde únicamente a β_1.

[7]Declarar equivocadamente un coeficiente como estadísticamente no significativo.

En resumen, los problemas de especificación por omisión o adición de términos innecesarios pueden conducir a obtener errores estándar equivocados para los parámetros, ocasionado por el sesgo y pérdida de eficiencia en los estimadores por MCO. Todo lo anterior, imposibilita realizar aseveraciones confiables a partir de los resultados obtenidos. A continuación se presentan diversos contrastes estadísticos, que permiten establecer si un modelo está correctamente especificado.

1.2.2 Detección del problema de especificación

Como se discutió anteriormente, los problemas de especificación tienen consecuencias sobre las estimaciones de MCO. Por esta razón, resulta conveniente contar con herramientas que permitan evaluar la idoneidad de un modelo econométrico. Así, las pruebas Ramsey-RESET, J de Davidson-MacKinnon y multiplicador de Lagrange, permiten diagnosticar de manera general la especificación adecuada o inadecuada de un modelo.

1.2.2.1 Prueba Ramsey-RESET

Una primera metodología para detectar especificación errónea en un modelo econométrico es la prueba de Ramsey-RESET, que a través de una regresión auxiliar busca evidenciar estadísticamente si el modelo tiene o no una adecuada especificación; sobre la cual aparecen los polinomios de la variable explicada estimada ($\hat{Y}_i^2$ y $\hat{Y}_i^3$) como nuevas variables independientes. En este caso, considere el modelo lineal general presentado en la sección anterior, con k variables independientes (*véase* ecuación 1.7).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + e_i \qquad (1.7)$$

De la misma forma, la regresión auxiliar (*véase* ecuación 1.8) viene dada por la ecuación inicial 1.6 más un polinomio de los valores estimados ($\hat{Y}_i^2$ y $\hat{Y}_i^3$). Sin embargo, en la práctica $\hat{Y}_i^2$ y $\hat{Y}_i^3$ son suficientes; aunque teóricamente conviene incluir tantas formas no lineales como sea posible de estos valores (Wooldridge, 2009, 303-304).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \delta_1 \hat{Y}_i^2 + \delta_2 \hat{Y}_i^3 + u_i \qquad (1.8)$$

Una vez especificada la regresión auxiliar, la prueba Ramsey-RESET consiste en:

1. Realizar la estimación del modelo en la ecuación 1.7 por MCO.
2. Con los resultados del inciso uno, obtener los valores estimados para la variable dependiente ($\hat{Y}_i$).
3. Estimar una regresión auxiliar de la ecuación 1.7, agregando los nuevos polinomios $\hat{Y}_i^2$ y $\hat{Y}_i^3$ como variables independientes.
4. Ejecutar la prueba estadística F (*véase* ecuación 1.10) sobre los coeficientes (δ_1 y δ_1) que acompañan a $\hat{Y}_i^2$ y $\hat{Y}_i^3$. Si se rechaza la hipótesis nula (*véase* prueba de hipótesis 1.9), donde plantea que estos son conjuntamente iguales a cero, quiere decir que el modelo especificado en la ecuación 1.7 esta especificado incorrectamente.

$$\begin{aligned} H_0 &: \delta_1 = \delta_2 = 0 && \text{Existe evidencia sobre una adecuada especificación.} \\ H_1 &: \delta_1 \neq \delta_2 \neq 0 && \text{Existe evidencia sobre una inadecuada especificación.} \end{aligned} \qquad (1.9)$$

$$F = \frac{(SCE_R - SCE_{NR})/l}{SCE_{NR}/n-k-1} \sim F_{l,n-k-1} \qquad (1.10)$$

Continuando la aplicación Ramsey-RESET, la ecuación 1.9 denota el estadístico F, donde SCE representa la suma al cuadrado de los errores asociado a los subíndices R y NR; que hacen referencia al modelo restringido[8] y no restringido[9] respectivamente. k corresponde al número de coeficientes en el modelo no restringido y n al total de observaciones. La cantidad de restricciones esta denotado por l, para este caso son las dos formas no lineales de Y. Si F calculado supera el valor critico determinado por $F_{l,n-k-1}$ bajo un nivel de significancia dado

[8] El modelo restringido, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + e_i$

[9] El modelo no restringido, $Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + \delta_1 \hat{Y}^2 + \delta_2 \hat{Y}^3 + u_i$

(usualmente 1%, 5% o 10%), entonces los coeficientes δ_1 y δ_2 son conjuntamente significativos o por lo menos uno de ellos es diferente de cero; es decir, existe evidencia estadística sobre especificación incorrecta.

No obstante, bajo este escenario se desconoce la causa sobre especificación incorrecta (por omitir o incluir variables independientes redundantes o forma funcional incorrecta). Por esto, debe probarse uno a uno planteamientos alternativos como: revisión de la teoría económica involucrada, análisis gráficos, u otros estudios (Hill et. al., 2001, 135-138); para conocer la fuente del problema encontrado mediante la prueba Ramsey-RESET.

1.2.2.2 Prueba J de Davidson-MacKinnon

Otra técnica que permite evidenciar especificación errónea en un modelo econométrico, es la prueba Davidson-MacKinnon –también conocida como *J* -; en ella se compara directamente el modelo especificado incorrectamente contra el potencialmente adecuado. En este orden de ideas conviene plantear las funciones que exponen todas las posibles causas sobre especificación incorrecta; es decir, una función de variables independientes omitidas versus la relación sin omisión (*véase* ecuación 1.11), regresión incluyendo variables explicativas irrelevantes comparada con otra sin ellas (*véase* ecuación 1.12) y una forma funcional correcta que ayude a contrastar la equivocada (*véase* ecuación 1.13).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \delta_1 Y_i^{omitida} + e_i \qquad (1.11)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i \rightarrow Y_i^{omitida} \qquad (1.12)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k+1} X_{ik+1} + \delta_1 Y_i^{redundante} + e_i \qquad (1.13)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + u_i \rightarrow Y_i^{redundante} \qquad (1.14)$$

$$Y_i = \beta_0 + \beta_1 \log X_{i1} + \beta_2 \log X_{i2} + \ldots + \beta_k \log X_{ik} + \delta_1 Y_i^{forma\ funcional} + e_i \qquad (1.15)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots + \beta_k X_{ik} + u_i \rightarrow Y_i^{forma\ funcional} \qquad (1.16)$$

A partir de las ecuaciones anteriores, la prueba J de Davidson-MacKinnon consiste en estimar mediante MCO por separado cada una de las segundas especificaciones en las ecuaciones 1.12, 1.14 y 1.16. En segunda instancia, tomar los valores estimados ($\hat{Y}_i$) obtenidos en estas y adicionarlos como variables independientes al primer modelo de las mismas ecuaciones (*véase* ecuaciones 1.11, 1.13 y 1.15).

Finalmente, con los resultados determinar la significancia estadística a nivel parcial de este nueva variable independiente, con el estadístico *t-student*. En general, la metodología se puede diseñar de la siguiente forma:

1. Plantear los modelos de las ecuaciones 1.11, 1.12 y 1.13, teniendo en cuenta que la prueba será efectuada sobre la primera especificación.
2. Realizar la estimación del segundo modelo en cada ecuación por MCO.
3. Obtener los valores estimados ($\hat{Y}_i$) de los segundos modelo.
4. Estimar el primer modelo, agregando el respetivo $\hat{Y}_i$ calculado en el paso tres.
5. Ejecutar una prueba estadística *t-student* de significancia individual, sobre el coeficiente nuevo que acompaña a los valores ajustados ($\hat{Y}_i$). Si el mismo, no resulta estadísticamente igual a cero (se rechaza la hipótesis nula) quiere decir que modelo inicial en la ecuación 1.11, 1.12 o 1.13 esta especificado incorrectamente (*véase* prueba de hipótesis 1.17).

$H_0 : \delta_1 = 0$	Modelo 1.11, 1.12 o 1.13 se encuentra especificado correctamente.	(1.17)
$H_1 : \delta_1 \neq 0$	Modelo 1.11, 1.12 o 1.13 se encuentra especificado incorrectamente por variables independientes omitidas, redundantes o forma funcional; respectivamente.	

Asimismo, el estadístico *t-student* es presentado en la ecuación 1.18; donde $ee(\hat{\delta}_1)$ corresponde al error estándar del coeficiente $\hat{\delta}_1$. Si el valor calculado *t-student*

supera el crítico, determinado por t_{n-1} bajo un nivel de significancia dado (usualmente 1%, 5% o 10%), entonces el coeficiente δ_1 resulta estadísticamente significativo; en otras palabras, es rechazada la hipótesis nula evidenciando especificación incorrecta.

$$t = \frac{\hat{\delta}_1}{ee(\hat{\delta}_1)} \sim t_{n-1} \qquad (1.18)$$

A diferencia de los resultados en la prueba Ramsey-RESET, J de Davidson-MacKinnon permite establecer las causas del problema (por omitir o incluir variables independientes redundantes o forma funcional incorrecta) de acuerdo a la especificación tratada; las cuales pueden ser analizadas individual o simultáneamente en un caso especifico.

1.2.2.3 Multiplicador de Lagrange

Para finalizar con los métodos que ayudan a detectar especificación errónea en un modelo econométrico, se cuenta con la prueba del multiplicador de Lagrange (ML). Entre las alternativas de Ramsey-RESET y J de Davidson-MacKinnon planteadas, ML permite probar el incumplimiento del supuesto de independencia condicional.[10]

Esta técnica consiste en comparar directamente el error estimado (e_i) del modelo especificado incorrectamente contra las variables independientes omitidas, redundantes o con forma funcional adecuada. En otras palabras, plantear (e_i) en función de las explicativas omitidas (*véase* ecuación 1.19), irrelevantes o redundantes (*véase* ecuación 1.20) y con forma funcional correcta (*véase* ecuación 1.21).

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i \rightarrow e_i \\ e_i &= \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_{k-1} X_{ik-1} + \alpha_k X_{ik} + \mu_i \end{aligned} \qquad (1.19)$$

[10]Cuando la covarianza entre al menos una de las variables explicativas (X_{ij}) y el error (u_i) es diferente de cero.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k+1} X_{ik+1} + u_i \rightarrow e_i$$
$$e_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \cdots + \alpha_k X_{ik} + \mu_i \qquad (1.20)$$

$$Y_i = \beta_0 + \beta_1 \log X_{i1} + \beta_2 \log X_{i2} + \ldots + \beta_k \log X_{ik} + u_i \rightarrow e_i$$
$$e_i = \alpha_0 + \alpha_1 X_{i1} + \alpha_2 X_{i2} + \ldots + \alpha_k X_{ik} + \alpha_3 \log X_{i1} + \alpha_4 \log X_{i2} + \ldots + \alpha_{k+1} \log X_{ik} + \mu_i$$
(1.21)

De esta forma, la prueba ML se puede efectuar de la siguiente manera:

1. Plantear y estimar por MCO los modelos iniciales de las ecuaciones 1.19, 1.20 y 1.21, teniendo en cuenta que la prueba será efectuada sobre estos.
2. Con los resultados del numeral uno, obtener los errores estimados (e_i) de cada modelo.
3. Tomar los errores estimados e involucrarlos como variable independiente, para especificar cada modelo auxiliar de las ecuaciones 1.19, 1.20 y 1.21.
4. Estimar por MCO las modelos auxiliares de las ecuaciones 1.19, 1.20 y 1.21.
5. Ejecutar la prueba estadística *ML* (*véase* ecuación 1.23) sobre todos los coeficientes de la regresión auxiliar ($\alpha_0, \alpha_1, \alpha_2, \ldots, \alpha_k$). Si no resultan en conjunto o uno de ellos individualmente estadísticamente igual a cero (se rechaza la hipótesis nula) quiere decir que modelo inicial en la ecuación 1.19, 1.20 o 1.21 esta especificado incorrectamente (*véase* prueba de hipótesis 1.22).

$H_0 : \alpha_1 = \alpha_2 = .. = \alpha_k = 0$	Modelo 1.19, 1.20 o 1.21 se encuentra especificado correctamente.
$H_1 : \alpha_1 \neq \alpha_2 \neq .. \neq \alpha_k \neq 0$	Modelo 1.19, 1.20 o 1.21 se encuentra especificado incorrectamente por variables independientes omitidas, redundantes o forma funcional respectivamente.

(1.22)

$$F = \frac{(SCE_R - SCE_{NR})/l}{SCE_{NR}/n-k-1} \sim F_{l,n-k-1} \qquad (1.23)$$

En la ecuación 1.23 el ML es igual a n (total de observaciones) por el coeficiente de determinación (R^2) de la regresión auxiliar; el cual sigue una distribución ji cuadrado con l grados de libertad que representan el número de los nuevos regresores. Para este caso son las variables explicativas omitidas, redundantes o con forma funcional correcta. Si *ML* calculado supera el valor critico determinado por Ji-cuadrado (χ^2_{n-k-1}) bajo un nivel de significancia dado (usualmente 1%, 5% o 10%), entonces los coeficientes α_1 y α_2 son conjuntamente significativos o por lo menos uno de ellos es diferente de cero; es decir, existe evidencia estadística sobre especificación incorrecta.

Todas las pruebas expuestas anteriormente, permiten establecer si un modelo econométrico esta o no correctamente especificado. Ante una inadecuada especificación, a continuación se presentan algunas técnicas que permiten corregir este problema utilizando teoría económica y variables aproximadas (o proxy, por su nombre en inglés) para remediarla.

1.2.3 Soluciones al problema de especificación incorrecta

La especificación correcta corresponde a uno de los supuestos del modelo clásico de regresión, cuyo cumplimiento permite encontrar coeficientes coherentes con la teoría económica y hacer inferencia estadística para las relaciones entre variables independientes y dependiente de la función estimada. Por esta razón, una vez se detecta un problema de este tipo, con cualquiera de las pruebas tratadas anteriormente, resulta necesario modificar el modelo econométrico inicial; para ello, esta sección presenta dos estrategias: el uso de la teoría económica y variables aproximadas o proxy.

1.2.3.1 Uso de la teoría económica

La primera alternativa para corregir especificación errónea, consiste en retornar a la teoría económica que originó el planteamiento del modelo econométrico; con el fin de identificar omisión de variables independientes relevantes o redundantes y forma funcional incorrecta. En el primer caso, es necesario recolectar los datos faltantes y de esta forma involucrar al modelo las variables explicativas omitidas observables o cuantificables, igualmente excluir las innecesarias. En el segundo, los postulados económicos deben indicar cómo expresar la forma funcional del modelo econométrico o de cada una de las variables implicadas en el mismo y así obtener una regresión correctamente especificada.

No obstante, las variables independientes relevantes excluidas pueden tener las características de no observable y tampoco cuantificables fácilmente. Sin embargo, dada su importancia desde el punto de vista económico descrito econométricamente, no deben ser prescindidas en el análisis; porque se incurre en el problema de especificación incorrecta por variable explicativa omitida. Por esta condición, el tratamiento del problema corresponde al uso de variables aproximadas o proxy.

1.2.3.2 Variables aproximadas o proxy

Adicionalmente al análisis teórico, pueden utilizarse variables aproximadas exógenas, porque en algunas ocasiones el origen del problema de especificación radica en la existencia de variables omitidas no observables o cuantificables. Ésta condición, puede ser la habilidad, el gusto, cultura y calidad de vida entre otros de una persona o sociedad en general; ante esto, igualmente el coeficiente intelectual puede ser una buena aproximación de la habilidad y el índice de desarrollo humano para la calidad de vida.

Tomando en cuenta lo anterior, la variable proxy puede definirse como una representación observable y cuantificable, cercana o relacionada a su determinante no perceptible. La cual puede ser incluida dentro del modelo sustituyendo la variable independiente no observable y de esta forma, capturar el efecto de la variable omitida solucionando la especificación errónea. Adicionalmente, este

método también es pertinente para resolver problemas de endogeneidad (*véase* sección 1.3).

Prosiguiendo el análisis sobre el funcionamiento de una variable proxy, suponga que desde el punto de vista económico debe plantearse un modelo de regresión con cuatro variables independiente, donde se cuenta con dos cuantificables (X_{i1}, X_{i2}) y dos son omitidas (X_{i3}, X_{i4}) por ser no observables (*véase* ecuaciones 1.24, 1.25 y 1.26).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad (1.25)$$

$$u_i = \pi_1 X_{i3} + \pi_1 X_{i4} + e_i \qquad (1.26)$$

Bajo el modelo 1.25, el método consiste en buscar dos variables aproximación (P_{i1} *y* P_{i2}), que teóricamente tengan correlación alta con las variables omitidas (*véase* ecuaciones 1.27 y 1.28). Ésta relación entre variables proxy y no observadas, no puede probarse empíricamente, entonces la cercanía entre X_{i3} y P_{i1}, y entre X_{i4} y P_{i2} debe ser argumentada a teóricamente.

$$corr(P_{i1}, X_{i3}) = 1 \qquad (1.27)$$

$$corr(P_{i2}, X_{i4}) = 1 \qquad (1.28)$$

Una vez establecidas las variables proxy, deben reemplazarse por las no observables (X_{i3}, X_{i4}) en el modelo inicial de la ecuación 1.24 y estimarse la nueva especificación (*véase* ecuación 1.29) mediante MCO; la cual, puede ser examinada con las pruebas conjuntas y parciales (estadísticos *F* y *t-student,* respectivamente) presentados en la sección 1.2.2.

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \beta_3 P_{i1} + \beta_4 P_{i2} + \upsilon_i \qquad (1.29)$$

Ahora, si no existen más variables independientes omitidas, se trabaja la forma funcional correcta y considerando que P_{i1} *y* P_{i2} son una buena aproximación a X_{i3} y X_{i4}, el modelo 1.29 deberá encontrarse bien especificado; obteniendo así

estimadores insesgados y consistentes (*véase* demostración anexo 1). Sin embargo, se puede continuar vulnerando el supuesto de independencia condicional por problemas de endogenidad.

1.3 Endogeneidad

Además de especificarse correctamente el modelo econométrico, es necesario conocer y garantizar exogeneidad en sus variables independientes implicadas, para garantizar el cumplimiento de independencia condicional y obtener estimadores insesgados y consistentes mediante MCO. De lo contrario, cuando se incumple el supuesto por causas distintas a especificación incorrecta; posiblemente puede considerarse endógena. A continuación, se discuten las causas que originan el problema de endogeneidad y consecuencias generadas sobre las estimaciones MCO; posteriormente las estrategias de identificación y corrección.

1.3.1 Causas y consecuencias de la endogeneidad

Una vez conceptualizado el problema de endogeneidad, esta sección describe cuatro posibles causas. Sin embargo, es necesario tener presente que los elementos definidos a continuación no son mutuamente excluyentes, porque en ejercicios empíricos varias fuentes de endogeneidad pueden presentarse simultáneamente.

1.3.1.1 Variable omitida

Una posible primer fuente que origina endogeneidad es un caso particular de variables explicativas omitidas (descrito, *véase* sección 1.2). En este caso, considere un modelo con k variables independientes y una de ellas se omite (*véase* ecuaciones 1.30 y 1.31).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_{k-1} X_{ik-1} + u_i \qquad (1.30)$$

$$u_i = \beta_k X_{ik} + e_i \qquad (1.31)$$

La ecuación 1.31 describe como la variable independiente omitida hace parte del error, adicionalmente ella está relacionada con una de las explicativas (X_{ij}), comportamiento denominado endogeneidad; dada su función con otra exógena del modelo, categorizándola como endógena. Esto implica dependencia condicional, que conduce a estimadores sesgado e inconsistente (*véase* anexo A.1.1 y A.1.2). Aun así, es necesario destacar que de omitir variables independientes no necesariamente resulta endogeneidad; se requiere además correlación con otra variable exógena no excluida.

En el caso particular de variables omitidas, la dirección del sesgo puede ser determinado analíticamente de la siguiente forma: considere un modelo econométrico con dos variables independientes ($k=2$), donde X_{i2} se excluye y está relacionada con X_{i1}[11] (*véase* ecuaciones 1.32, 1.33 y 1.34).

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i \tag{1.32}$$

$$u_i = \beta_2 X_{i2} + e_i \tag{1.33}$$

$$X_{i2} = \alpha_0 + \alpha_1 X_{i1} + \upsilon_i \tag{1.34}$$

$$Y_i = \underbrace{(\beta_0 + \beta_2\alpha_0)}_{\beta_0^*} \pm \underbrace{\left(\beta_1 + \beta_2\alpha_1\right)}_{\beta_1^*} X_{i1} \pm \underbrace{\left(\beta_2\upsilon_i + e_i\right)}_{e_i^*} \tag{1.35}$$

En la ecuación 1.34 α_0, α_1, β_1, β_2 y e_i son los parámetros y termino de error respectivamente de la regresión auxiliar, ahora se reemplaza esta aproximación en 1.33 y posteriormente en 1.32 resultando la ecuación 1.35 o expresión general del sesgo (betas estrella). En este caso, la dirección del sesgo viene determinada por los valores y signos de $\beta_2\alpha_0$ y $\beta_2\alpha_1$[12]; adicionalmente en la expresión 1.35 también el intercepto (β_0) aparece sesgado.

[11] Esto es válido dado que la $Cov(X_{i1}, X_{i2}) \neq 0$

[12] Dependiendo del signo, $\beta_1\alpha_0 > 0$ indica que β_1 se encuentra sobreestimado y $\beta_1\alpha_0 < 0$ β_1 subestimado.

1.3.1.2 Simultaneidad

La segunda causa, por la cual se puede incurrir en un problema de endogeneidad, es denominada simultaneidad y ocurre cuando el planteamiento económico a describir contiene variables dependientes tratadas como independientes y relacionadas entre sí o que se determinan conjuntamente en un proceso.

En estas circunstancias, los estimadores calculados por MCO reflejan una mezcla del efecto entre las diferentes direcciones de causalidad para las variables interrelacionadas. Si el interés es estudiar únicamente una de estas direcciones, debe replantearse el problema a un sistema de varias ecuaciones; con el fin de lograr reflejar las diferentes relaciones que ligan las variables. Este tema se estudiará de manera independiente en el capítulo 2.

1.3.1.3 Error de medición en las variables independientes

En tercer lugar, los problemas de endogeneidad suelen ser ocasionados cuando existen errores de medición en alguna de las variables independientes, surgidos por digitar o responder equivocadamente encuestas, igualmente en la manipulación inadecuada de información secundaria agregada o desagregada.

En este orden de ideas, se tiene un modelo econométrico con dos variables explicativas (*véase* ecuación 1.36); en el mismo, infortunadamente los datos para X_{i2} contienen errores de medición (*véase* ecuación 1.37).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2}^{*} + e_i \qquad (1.36)$$

$$X_{i2}^{*} = X_{i2} + \upsilon_i \qquad (1.37)$$

De esta forma, la ecuación 1.37 expresa como X_{i2}^{*} (hace referencia a los valores recolectados equivocadamente) equivale a la suma entre X_{i2} (indica las observaciones verdaderas -sin errores- para la variable independiente) y el error de medición (υ_i). A partir de lo anterior, remplazando la ecuación 1.37 en 1.36 se puede evidenciar endogeneidad para X_{i2}, dada su relación con υ_i (*véase* ecuaciones 1.38 y 1.39).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad (1.38)$$

$$u_i = e_i + \beta_2 \upsilon_i \qquad (1.39)$$

Igualmente, cuando existe error de medición, los estimadores calculados por MCO están sesgados. Su tamaño y dirección vienen determinados por la estructura del problema, que varía para cada caso particular.

1.3.1.4 Sesgo de selección

Finalmente, la endogeneidad puede surgir por sesgos de selección; este problema aparece cuando los datos no son aleatorios, para las variables independiente o explicativas, resultado de errores en su recolección por parte de los entrevistadores o alguna selección (autoselección) de los encuestados.

En particular, la inadecuada recolección de información puede presentarse al excluir deliberadamente preguntas al momento de aplicar encuestas o poca claridad en la redacción de las mismas. Por ejemplo, para preguntas de selección múltiple, algunas veces ninguna alternativa se ajusta a la condición del encuestado. De esta manera, con un alto porcentaje de datos con respuestas en blanco, probablemente se tenga un problema de no aleatoriedad.

Por otra parte, la clasificación incorrecta de la muestra también conduce a sesgo de selección. Generalmente, puede ocurrir cuando todos los individuos de la población no tienen la misma probabilidad de hacer parte de la muestra; conllevando a una autoselección que es un caso particular de esta fuente de sesgo, comúnmente presentada en los estudios econométricos sobre programas gubernamentales; debido a disponibilidad de tiempo e intereses particulares, porque no todos los individuos tienen la misma propensión a participar en programas públicos.

Al igual que en los casos de variables omitidas, simultaneidad y error de medición, el análisis econométrico por MCO conduce a estimadores sesgados e inconsistentes. Comprendidos los problemas que pueden ocasionar endogeneidad y consecuencias en un modelo econométrico, la siguiente sección introduce el

concepto de variable instrumental con el fin de evaluar la existencia del problema y alternativa para solucionarlo.

1.3.2 Introducción a las variables instrumentales.

Una vez estudiadas las diferentes fuentes que originan endogeneidad, pueden discutirse las estrategias para detectarla y solucionarla. De esta forma, inicialmente el problema es asumido desde el postulado económico plasmado en el modelo, sin ninguna prueba que permita demostrarlo; posteriormente su respectiva solución con variables instrumentales (VI) y luego la aplicación estadística evidenciando el incumplimiento de independencia condicional por endogenidad.

En otras palabras, debe pensarse que el modelo econométrico incurre en endogeneidad, paralelamente plantear su solución (VI) y luego probar realmente la existencia del problema. Diferente al procedimiento convencional efectuado cuando es infringido algún otro supuesto para MCO, donde se lleva a cabo primero las pruebas de detección y posteriormente la medida correctiva.

Por esta razón, es necesario inicialmente introducir el concepto de instrumento o variable instrumental (VI). Definida como aquella relacionada con la explicativa que causa el problema de endogeneidad e independiente al término del error en el modelo. Adicionalmente, la elección de un instrumento se realiza a partir del problema económico planteado econométricamente.

Asimismo, una VI permite transformar el modelo econométrico inicial donde es asumido el problema de endogeneidad; con el fin de obtener estimadores insesgados y consistentes. También, el instrumento surge ante la dificultad de encontrar una variable aproximación de la independiente no observable omitida; paralelamente relacionada con alguna de las otras en el modelo.

Ante esto, como instrumento puede seleccionarse cualquier variable que satisfaga las condiciones de validez y relevancia. En la primera, VI (denotada en las ecuaciones como Z_i) debe ser exógena al modelo econométrico especificado; es decir, independiente del termino de error (*véase* ecuación 1.40) y en la segunda, VI

explica la variable dependiente correlacionada con el error (Wooldridge, 2009, 308) (*véase* ecuación 1.41).

$$\text{cov}(Z_i, e_i) = 0 \qquad (1.40)$$

$$\text{cov}(Z_i, X_i) \neq 0 \qquad (1.41)$$

Para demostrar que la variable instrumental cumple estas condiciones, son utilizadas la prueba de restricciones sobreidentificadas, con el fin de evidenciar la condición 1.40 (*véase* sección 1.3.5) y una regresión auxiliar, donde la VI (Z_i) se utiliza como regresora de la variable endógena (X_i), para comprobar 1.41. Ésta regresión auxiliar, hace parte del proceso de estimación por mínimos cuadrados en dos etapas (MC2E) (*véase* sección 1.3.4). Continuando con el análisis, en las secciones siguientes es explicado el uso de VI, para identificar y estimar modelos con endogeneidad.

1.3.3 Detección endogeneidad: introducción a la prueba de Hausman

A partir del concepto sobre variables instrumentales, esta sección introduce la prueba de Hausman; que permite identificar la existencia de endogeneidad en un modelo. Adicionalmente, también usada para detectar problemas de simultaneidad y efectos fijos en datos longitudinales (*véase* capítulos 2 y 8, respectivamente).

Sin embargo, para aplicar esta prueba estadística es necesario preliminarmente comprender el método de estimación mínimos cuadrados en dos etapas (MC2E), dado que ella permite la comparación estadística entre los estimadores MCO y MC2E (Hill et. al., 2001, 299). De esta forma, a medida que el valor de ambos coeficientes son cercanos, el modelo cumple con independencia condicional; de lo contrario existe al menos un regresor endógeno (Wooldridge, 2009, 527). Por lo anterior, el numeral 1.3.4.2 presenta MC2E y el desarrollo formal de Hausman se encuentra en la sección 1.3.5.

1.3.4 Soluciones a endogeneidad

Conceptualizada la prueba de Hausman para detectar el problema de endogeneidad y sospechando que se cuenta con este problema. Ésta sección, expone las medidas remédiales al mismo; entre las que se destacan el uso de variables aproximación e instrumetales.

1.3.4.1 Variables aproximadas

El método de variable aproximadas presentado anteriormente (*véase* sección 1.2.3), es la primera alternativa para solucionar el problema de endogeneidad, cuando se origina por omisión de variables independientes no observables (X_{i2}) relacionada con otra explicativa (X_{i1}) cuantificable en el modelo. De acuerdo con lo expuesto en la sección 1.2.3, la variable la proxy debe sustituirse por la no observable, sin embargo realizar esta especificación transformará el problema de endogeneidad a uno de colinealidad (*véase* ecuación 1.42, 1.43 y 1.44).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \qquad (1.42)$$

$$u_i = \beta_2 X_{i2} + e_i \qquad (1.33)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 P_i + u_i \qquad (1.44)$$

En relación con lo anterior, la ecuación 1.44 puede estimarse por MCO; aunque en ella la variable aproximada (P_i) tiene relación con la independiente (X_{i1}). Ante esto, debe emplease una variable instrumental y utilizar la metodología MC2E; discutida a continuación.

1.3.4.2 Mínimos cuadrados en dos etapas (MC2E)

La segunda alternativa, para resolver problema de endogeneidad, consiste en eliminar el componente endógeno del sistema mediante el uso de variables instrumentales; así obtener estimadores insesgados mediante MC2E (*véase* anexo 1). Este método es el segundo más usado en la literatura, superado únicamente por los mínimos cuadrados ordinarios (Wooldridge, 2009, 2).

Para comprender esta metodología, considere un modelo de regresión con dos variables independientes (X_{i1} y X_{i2}), endogeneidad sobre X_{i1} por una variable omitida X_{i3}, ($Cov(X_{i1}, X_{i3}) \neq 0$) y se cuenta con un instrumento Z_{i1} relacionado con X_{i1} (*véase* ecuaciones 1.45, 1.46 y 1.47).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \mu_i \quad (1.45)$$

$$\mu_i = \beta_3 X_{i3} + e_i \quad (1.46)$$

$$X_{i1} = \pi_0 + \pi_1 Z_{i1} + \beta_2 X_{i2} + \upsilon_i \quad (1.47)$$

La ecuación 1.47 se refiere a una regresión auxiliar denominada forma reducida de la ecuación estructural, plasmada en 1.45. Así, la primera etapa consiste en estimar 1.47 mediante MCO, donde la variable explicada es X_{i1} y las explicativas son X_{i2} y Z_{i1}, realizar una prueba (*t-student*) de significancia individual sobre el coeficiente (π_0, π_1) que acompaña VI (Z_{i1}); si resulta estadísticamente significativo es validado el instrumento. Posteriormente se modifica el modelo inicial en 1.45, reemplazando los valores observados de la variable endógena (X_{i1}) con los valores estimado ($\hat{X}_{i1}$) (*véase* ecuación 1.48); finalmente la segunda etapa corresponde a estimar 1.48 por MCO.

$$Y_i = \beta_0 + \beta_1 \hat{X}_{i1} + \beta_2 X_{i2} + \mu_i \quad (1.48)$$

El valor resultante para $\hat{\beta}_1$, se denomina estimador de MC2E para X_{i1}. Una vez corregido el problema de endogeneidad con VI mediante MC2E, es posible mostrar que estos nuevas parámetros son insesgados y consistentes (*véase* anexo A.1.4). En términos generales, una estimación por MC2E consiste en:

1. Estimar una regresión auxiliar -forma reducida- mediante MCO, donde la variable endógena X^e es explicada a partir de las exógenas y al menos un instrumento.
2. A partir de la regresión auxiliar del paso uno, realizar la prueba (*t-student*) de significancia parcial sobre el coeficiente que acompaña a VI, con el fin de conocer la validez del instrumento.

3. Si el instrumento es válido, capturar de la forma reducida los valores ajustados de la variable endógena $\hat{X}^e$.
4. Con esta información, remplazar los valores observados de la variable endógena (X^e) por los estimados obtenidos en el paso tres ($\hat{X}^e$) y estimar la segunda etapa, el modelo inicial por MCO; el estimador $\hat{\beta}_{MCO}$ de esta regresión, es insesgado.

Por otra parte, en general debe contarse con igual número de instrumentos (Z) como variables endógenas (X) se tenga en el modelo; cuando existe una o más de una variable endógena. Así, la primera etapa del método MC2E cuenta con varias regresiones auxiliares (una forma reducida por cada endógena). En otras palabras, si existen dos variables independientes que causan problemas de endogeneidad, debe contarse con dos VI e igualmente con dos formas reducidas; una para cada instrumento. Esto, se conoce como condición mínima de orden (*véase* cuadro 1.1).

Cuadro 1.1. Estado del modelo a partir de la condición de orden

Relación entre numero de variables endógenas X^e e instrumentos Z	**Estado del modelo**
$X^e > Z$	No identificado[13]
$X^e = Z$	Justamente identificado
$X^e < Z$	Sobre identificado

Fuente. los autores

Asimismo, deben incluirse las variables exógenas e instrumentales en cada modelo reducido, según el caso para predecir otras endógenas; con el fin de obtener la mejor predicción de la variable de interés (*véase* cuadro 1.2).

[13] Los modelos con más endógenas que instrumentos, incumplen la condición mínima de orden por lo que no pueden ser estimados.

Cuadro 1.2. Diferentes casos del Método de Mínimos Cuadrados en Dos Etapas

Caso	Variables	Primera Etapa de la Regresión	Segunda Etapa de la Regresión
Modelo univariado (*véase* anexo 1.3)	Donde X_{i1} es la variable endógena y Z_{i1} la variable instrumental.	$\hat{X}_{i1} = \pi_0 + \pi_1 Z_1 + \upsilon_i$	$Y_i = \beta_0 + \beta_1 \hat{X}_{i1} + \mu_i$
Una variable endógena y un instrumento	Donde X_{i1} es la variable endógena, $X_{i2} \cdots X_{ik}$ las exógenas y Z_{i1} el instrument0.	$\hat{X}_{i1} = \pi_1 Z_{i1} + \pi_2 X_{i2} + \pi_3 X_{i3} + \cdots + \pi_k X_{ik} + \upsilon_i$	$Y_i = \beta_0 + \beta_1 \hat{X}_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i$
Una variable endógena y más de un instrumento	Donde X_{i1} es la variable endógena, $X_{i2} \cdots X_{ik}$ exógenas y $Z_{i1} \cdots Z_{im}$ instrumentales.	$\hat{X}_{i1} = \pi_1 Z_{i1} + \cdots + \pi_m Z_{im} + \pi_{m+1} X_{i2} + \cdots + \pi_{m+k} X_{ik} + \upsilon_i$	$Y_i = \beta_0 + \beta_1 \hat{X}_{i1} + \beta_2 X_{i2} + \cdots + \beta_k X_{ik} + \mu_i$
Varias variables endógenas y un instrumento por cada una	Donde $X_{i1} \cdots X_{ik}$ son variables endógenas, $X_{i,k+1} \cdots X_{im}$ exógenas y $Z_{i1} \cdots Z_{ik}$ instrumentales. La variable Z_{ij} corresponde al instrumento de X_{ij}	$\hat{X}_{i1} = \pi_1 Z_{i1} + \pi_2 X_{i2} + \pi_3 X_{i3} + \cdots + \pi_k X_{ik} + \upsilon_i$ $\vdots$ $\hat{X}_{ik} = \pi_1 Z_{ik} + \pi_2 X_{i1} + \cdots + \pi_{k-1} X_{i,k-1} + \pi_{j+1} X_{i,k+1} + \cdots + \pi_m X_{im} + \upsilon_i$	$Y_i = \beta_0 + \beta_1 \hat{X}_{i1} + \beta_2 \hat{X}_{i2} + \cdots + \beta_k \hat{X}_{ik} + \beta_{k+1} X_{i,k+1} + \cdots + \beta_m X_{im} + \mu_i$
Varias variables endógenas y más de un instrumento por cada una	Donde $X_{i1} \cdots X_{ik}$ son variables endógenas, $X_{k+1} \cdots X_m$ exógenas y $Z_{i1,j} \cdots Z_{in,j}$ instrumentos para X_j.	$\hat{X}_{i1} = \pi_{1,1} Z_{i1,1} + \cdots + \pi_{n,1} Z_{in,1} + \pi_2 X_{i2} + \pi_3 X_{i3} + \cdots + \pi_k X_{ik} + \upsilon_i$ $\vdots$ $\hat{X}_{ij} = \pi_{1,k} Z_{i1,k} + \cdots + \pi_{n,k} Z_{in,k} + \pi_2 X_{i2} + \cdots + \pi_{j-1} X_{i,j-1} + \pi_{j+1} X_{i,j+1} + \cdots + \pi_k X_{ik} + \upsilon_i$	$Y_i = \beta_0 + \beta_1 \hat{X}_{i1} + \beta_2 \hat{X}_{i2} + \cdots + \beta_k \hat{X}_{ik} + \beta_{k+1} X_{i,k+1} + \cdots + \beta_m X_{im} + \mu_i$

Fuente: los Autores

1.3.5 Prueba de Hausman

A partir de la metodología MC2E, es posible aplicar la prueba de Hausman para identificar existencia de endogeneidad en un modelo. De esta forma, considere un modelo de regresión con una variable independiente (X_{i1}) y otra omitida (X_{i2}) (*véase* ecuaciones 1.49 y 1.50); existiendo relación entre ellas ($Cov(X_{i1}, X_{i2}) \neq 0$). En términos generales, implica endogeneidad.

$$Y_i = \beta_0 + \beta_1 X_{i1} + u_i \qquad (1.49)$$

$$u_i = \beta_2 X_{i2} + e_i \qquad (1.50)$$

La forma estructural en 1.49, representa un modelo de regresión simple con endogeneidad; dada la correlación presumida entre X_{i1} y X_{i2}. Bajo esta condición y de acuerdo con el cuadro 1.2, debe contarse con una VI (Z) que se relacione con X_{i1} y un modelo reducido respectivamente (*véase* ecuación 1.51). Una vez planteadas la forma estructural y reducida aplicar MC2E.

Continuando con el análisis, estimar la primera etapa (ecuación 1.51 por MCO), posteriormente obtener de la ecuación 1.51 los valores estimados del error ($\hat{\upsilon}_i$) y agregarlos como una nueva variable independiente en 1.49 (*véase* ecuación 1.52). Así, X_{i1} será exógena cuando los $\hat{\upsilon}_i$ no estén correlacionados con los errores (u_i) del modelo inicial en 1.52.

$$X_{i1} = \pi_0 + \pi_1 Z_{i1} + \upsilon_i \qquad (1.51)$$

$$Y_i = \beta_0 + \beta_1 X_{i1} + \delta \upsilon_i + \mu_i \qquad (1.52)$$

Bajo este esquema, la prueba hipótesis para endogeneidad equivale a una de significancia parcial sobre el coeficiente δ, que acompaña los residuales obtenidos en la primera etapa, en el modelo 1.52 (*véase* prueba de hipótesis 1.53).

$$\begin{aligned} H_0 &: \delta = 0 \quad \text{No existe endogeneidad} \\ H_1 &: \delta \neq 0 \quad \text{Existe endogeneidad} \end{aligned} \tag{1.53}$$

El estadístico viene dado por una prueba *t-student* (*t*), expuesta en la ecuación 1.54, donde $ee(\delta)$ corresponde al error estándar estimado para δ. Si, *t-student* calculado supera al reportado en las tablas; el coeficiente de interés resulta estadísticamente significativo, implicando endogeneidad para X_{i1}.

$$t = \frac{\delta}{ee(\delta)} \sim t_{N-1} \tag{1.54}$$

En términos generales, la prueba de Hausman consiste en:

1. Especificar la forma estructural y reducida.
2. Aplicar MC2E.
3. Estimar la primera etapa mediante MCO, la forma reducida.
4. Obtener los errores estimados ($\hat{\upsilon}_i$) de la primera etapa.
5. Adicionar los errores obtenidos en 4, como variable explicativa en el modelo estructural y realizar una estimación por MCO de este ultimo.
6. Realizar la prueba de significancia parcial, con el estadístico *t-student*, sobre el coeficiente que acompaña a los residuales estimados en el modelo modificado en 5. Si el estimador es significativo, existe endogeneidad en el modelo.

La prueba de Hausman presentada, permite establecer si un modelo presenta endogeneidad. Una vez evidenciado el problema, a continuación se estudia la prueba de Sargan para determinar restricciones sobreidentificados y evaluar la validez de un instrumento dentro de la metodología MC2E.

1.3.6 Prueba Sargan para restricciones sobreidentificadas

Una vez elegida la VI, identificada la endogeneidad y estimado el modelo por MC2E para la respectiva solución, posiblemente exista más de un instrumento a emplear (Davidson et. al., 2004, 336). Ante esto, es necesario probar mediante la prueba de Sargan la validez de los instrumentos con que se cuentan y utilizados.

Con el fin de comprender esta metodología, considere un modelo de regresión con dos variables independientes (X_{i1} y X_{i2}), una omitida (X_{i3}) relacionada con X_{i1} ($Cov(X_{i1}, X_{i3}) \neq 0$) y dos instrumentos (Z_{i1} y Z_{i2}) (*véase* ecuaciones 1.55, 1.56 y 1.57).

$$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + u_i \quad (1.55)$$

$$u_i = \beta_3 X_{i3} + e_i \quad (1.56)$$

$$\hat{X}_{i1} = \pi_0 + \pi_1 Z_{i1} + \pi_2 X_{i2} + \upsilon_i \quad (1.57)$$

La ecuación 1.55 representa la forma estructural y 1.57 su respectiva forma reducida, dado que se presume endogenidad causada solo por una variable independiente (X_{i1}); razón por la cual, debe constarse con un solo instrumento y una forma reducida, respectivamente. Sin embargo, en este caso existen dos instrumentos; donde uno de ellos debe elegirse, para no sobre identificar VI (*véase* cuadro 1.1).

Bajo esta circunstancia, en primer lugar debe utilizarse únicamente uno de los dos instrumento (Z_{i1}), luego estimar el modelo 1.55 por MC2E (*véase* ecuaciones 1.57 y 1.58). Posteriormente, tomar los errores estimados ($\hat{\mu}_i$) y emplearlos, en una nueva regresión auxiliar (*véase* ecuación 1.59), como variable dependiente en función del instrumento excluido (Z_{i2}) y la exógena X_{i2}. La ecuación 1.59 se estima mediante MCO.

$$Y_i = \beta o + \beta_1 \hat{X}_{i1} + \beta_2 X_{i2} + \mu_i \quad (1.58)$$

$$\hat{\mu}_i = \gamma_0 + \gamma_1 Z_{i2} + \gamma_2 X_{i2} + v_i \quad (1.59)$$

Con los resultados en 1.59, realizar la prueba del multiplicador de Lagrange (*ML*), obtenida de multiplicar el coeficiente de determinación (R^2) de la regresión auxiliar por el número de observaciones (*n*) (*véase* ecuación 1.61). *ML* (nR^2) se distribuye como Ji caudrada (χ_q^2), donde q corresponde al número de instrumentos totales disponibles menos el número de variables endógenas. Si nR^2 supera a χ_q^2 reportado en las tablas bajo un nivel de significancia dado (1%, 5% o

10%), se rechaza la hipótesis nula (*véase* prueba de hipótesis 1.60) concluyendo que alguno de los instrumentos adicionales no es válido[14], caso contrario cuando no es rechazada.

$$\begin{aligned} H_0 &: Cov(Z_{i2}, \hat{\mu}_i) = 0 \quad \text{Instrumento es válido} \\ H_1 &: Cov(Z_{i2}, \hat{\mu}_i) \neq 0 \quad \text{Instrumento no es válido} \end{aligned} \qquad (1.60)$$

$$LM = nR^2 \sim \chi_q^2 \qquad (1.61)$$

En general, la prueba de Sargan para conocer restricciones sobre identificadas o validez de instrumentos consiste en:

1. Especificar la forma estructural y reducida.
2. Identificar las posibles VI a utilizar.
3. Realizar la estimación del modelo por MC2E, usando únicamente una de las variables instrumentales disponibles.
4. Obtener los errores estimados ($\hat{\mu}_i$) derivados de MC2E.
5. Especificar y estimar una regresión auxiliar por MCO, donde los errores estimados $\hat{\mu}_i$ son tomados variable dependiente en función de las exógenas e instrumentos excluidos de la forma reducida.
6. Probar significancia conjunta para los instrumentos excluidos, mediante la prueba del multiplicador de Lagrange (*ML*).
7. Repetir los pasos 1-6, empleando otro de los instrumentos identificados en el paso 2.

Este procedimiento estadístico, concluye la sección teórica de este capítulo. A continuación, se aplican las técnicas en un estudio de caso, para comprender los temas sobre especificación y endogeneidad; en particular el manejo de variables aproximadas, evaluación de especificación, uso de MC2E y pruebas de Hausman y Sargan.

[14] Un instrumento válido implica que no existe correlación con $\hat{\mu}_i$.

1.4 Estudio de caso: derechos de propiedad en Colombia e integración al mercado mundial.

Una vez expuestas las diferentes metodologías relevantes, para detectar y remediar las causas y consecuencias, sobre problemas de especificación y endogeneidad en un modelo econométrico. Su respectiva aplicación, se desarrolla con información cronológica de variables institucionales y económicas en el programa estadístico Stata®.

Con este fin, a continuación son tomados los datos del estudio titulado (en inglés) "*Land Conflict, Property Rights, and the Rise of the Export Economy in Colombia, 1850-1925*" de Sánchez, Fazio y Lopez-Uribe (2008). El artículo, pretende mostrar econométricamente cómo la debilidad de los derechos de propiedad en la frontera de colonización agrícola colombiana, condujo a una baja integración económica entre el país y los mercados mundiales a finales del siglo XIX.

De acuerdo con los autores, la segunda mitad del siglo XIX se caracterizó por un rápido crecimiento económico a nivel mundial y una expansión significativa del comercio entre países. De esta forma, Latinoamérica no fue ajeno a este fenómeno y países como Brasil, Argentina y Chile aumentaron significativamente su producción y niveles de exportaciones. No obstante, en Colombia las exportaciones crecieron por debajo del promedio mundial.

El bajo desempeño del país, es explicado por factores externos como ciclos de precios internacionales desfavorables y una baja demanda de los productos locales en el mercado exterior. Sin embargo, en este artículo Sánchez et al. (2008) tratan de establecer que la baja integración fue resultado de factores internos y no externos, como la mala calidad institucional en algunas regiones de Colombia.

De esta forma, los autores explican que la falta de derechos de propiedad formales en zonas agrícolas alejadas del centro de la nación, junto con las oportunidades de altos ingresos provenientes de una coyuntura mundial apropiada, causaron conflictos por el control de la tierra entre campesinos y terratenientes. Estos

enfrentamientos impactaron negativamente el desempeño económico de estas zonas, conllevando a una producción exportable menor a la potencial.

A partir de lo anterior, se especifica un modelo econométrico lineal (*véase* ecuación 1.50), dada la información con la que cuentan los autores sobre indicadores geográficos (distancia a las principales ciudades, ríos, altitud e indicador de calidad para la tierra) y producción exportable de 760 municipios para Colombia en 1892. Estas variables, permiten mostrar la aplicación y funcionamiento de las diversas metodologías presentadas en el capítulo.

$$PIBpcBienesExportables_i = \beta o + \beta_1 ConflictoTierras_i + \mathbf{X}\boldsymbol{\delta} + e_i \qquad (1.62)$$

En la ecuación 1.62, $PIBpcBienesExportables_i$ corresponde a la producción per cápita de productos exportables del municipio i en 1892 y $ConflictoTierras_i$, es la variable (binaria) de interés que toma valor de uno cuando existe evidencia de conflictos alrededor por la propiedad de la tierra en un determinado municipio y cero, caso contrario. Por último, $\mathbf{X}$ (*véase* cuadro 1.3) corresponde a una matriz de variables de control ($\boldsymbol{\delta}$, con su respectivo vector de coeficientes), relacionadas con la geografía del municipio.

Para comprobar el cumplimiento de la hipótesis plasmada por los autores, mediante los resultados de significancia parcial con la *t-student*, β_1 debe resultar estadísticamente diferente de cero y con signo negativo; dado que teóricamente se espera menor producción exportable como consecuencia de mayores conflictos sobre la tierra.

Sin embargo, el cumplimiento de esta hipótesis no es fácil de probar mediante una regresión lineal simple, puesto que la variable $ConflicoTierras_i$ teóricamente es endógena. En particular, el número de conflictos se relaciona con el nivel de retornos esperados y existencia de derechos de propiedad informales (variables que no se encuentran en el modelo).

Para superar este inconveniente, Sánchez et al. (2008), sugieren usar como VI la distancia entre el municipio y los centros de poder coloniales; empleando MC2E.

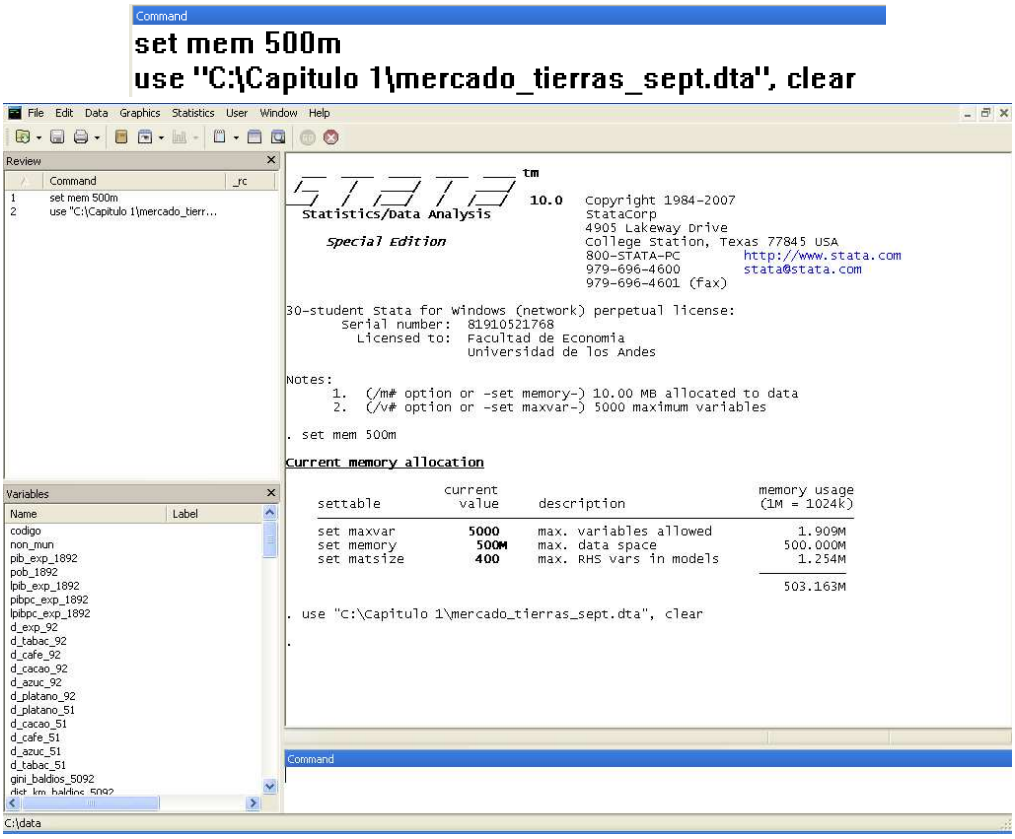
De acuerdo con los autores, existe una relación inequívoca entre calidad institucional y distancia de los centros de mando gubernamentales tradicionales de la región, por lo que este instrumento es relevante.

1.4.1 Análisis general de los datos

Esta primera sección se prepara Stata® para el análisis econométrico y realizar una exploración general de la base de datos a usar, incluyendo una descripción de las variables relevantes utilizadas en el modelo. Este tipo de sondeo es relevante, porque permite familiarizarse con los datos y conocer su calidad, consistencia y veracidad; para lo anterior el procedimiento en Stata® es el siguiente:

1. Determinar la memoria del sistema, a través del comando *set memory* –o *set mem-*. Cuando se desconoce con exactitud el tamaño de la base de datos, puede asignarse 500m de memoria al programa. Esto es suficiente para cargar prácticamente cualquier base.
2. Carga la base de datos, cuyo nombre es *mercado_tierras_sept.dta* (archivo disponible en el CD anexo), con el comando *use* (*véase* figura 1.1).

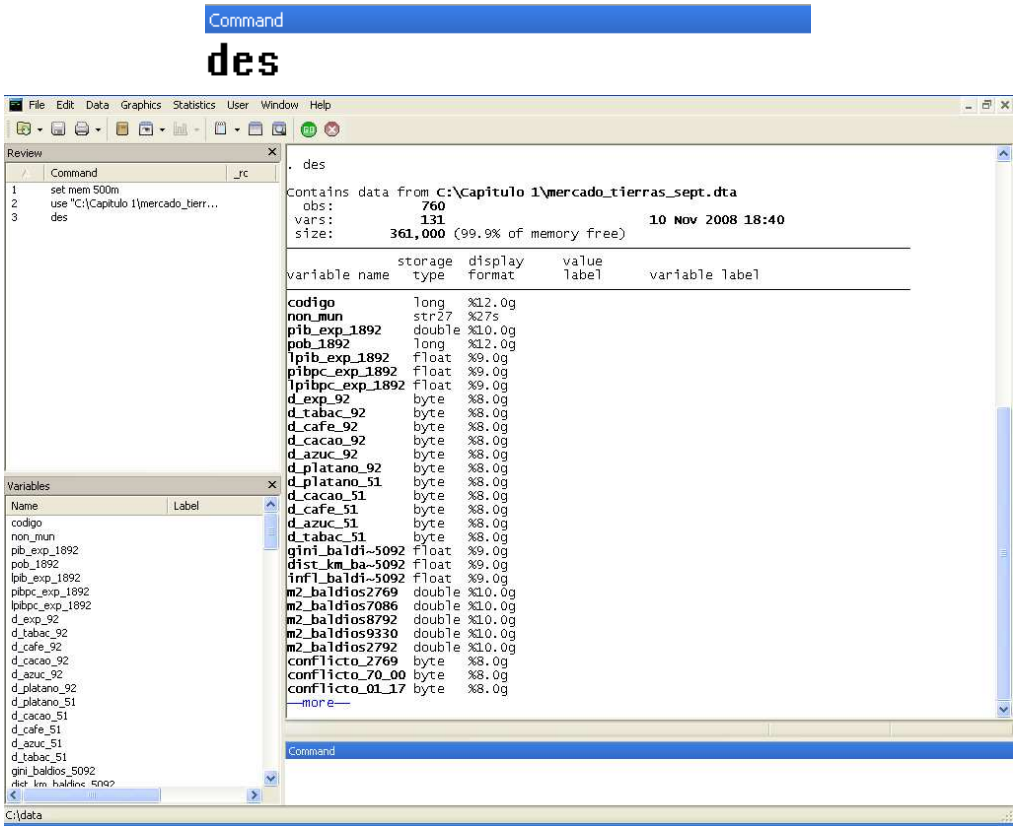
Figura 1.1. Salida comandos set memory y use



Fuente: cálculo autores

3. Para observar el nombre y descripción de las variables disponibles, se utiliza el comando *describe* –o *des*. Adicionalmente, la salida muestra el formato en que están guardadas (*véase* figura 1.2). En este caso, existen 760 observaciones y 131 variables; de las cuales solo son empleadas 17 (*véase* cuadro 1.3). La descripción de las variables no está disponible, aunque los nombres de cada una indican claramente su contenido.

Figura 1.2. Salida comando describe



Fuente: cálculo autores

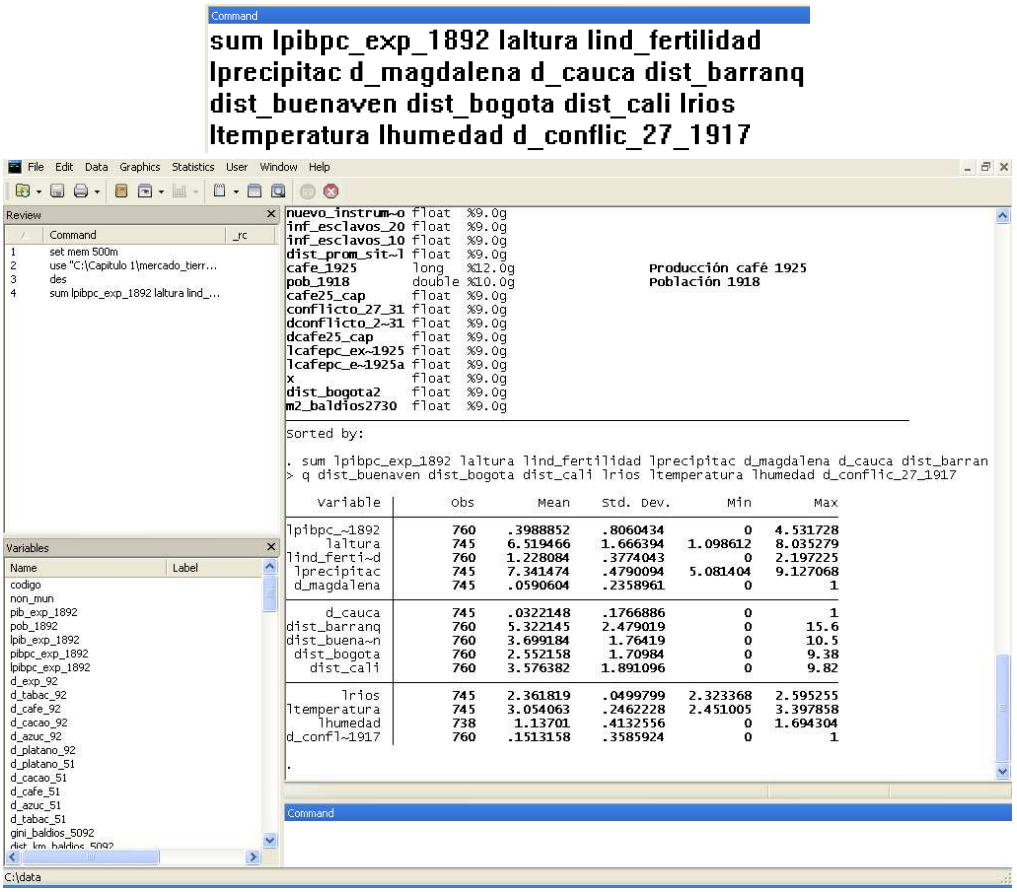
Cuadro 1.3. Variables empleadas en el modelo y su descripción

Variable del Modelo	Variables en la Base	Descripción
$PIBpcBienesExportables_i$	lpibpc_exp_1892	Logaritmo del PIB PC de bienes exportables producidos en la municipalidad i en 1892.
$ConflictoTierras_i$	d_conflic_27_1917	Variable dicótoma; uno, evidencia de conflictos de propiedad en i.
X	laltura, lind_fertilidad, lprecipitac, lrios, ltemperatura, lhumedad, d_magdalena, d_cauca, dist_barranq, dist_buenaven, dist_bogota, dist_cali	Variables geográficas, que registran altitud, fertilidad, precipitación promedio, existencia de ríos, temperatura promedio, índice de humedad, distancia a los ríos magdalena y cauca y a Barranquilla, Buenaventura, Bogotá y Cali, para cada municipalidad.
Instrumentos	inf_indios_1560, inf_esclavos_1800, d_esclavos_mas80	Dicótomas, uno si en la municipalidad existían encomiendas en 1560, o centros con más de 80 esclavos en 1800.

Fuente: los autores

4. Antes de estimar el modelo, es necesario observar las estadísticas descriptivas de las variables relevantes. Para eso, se utiliza el comando *summary* –o *sum*-, que presenta una tabla con el número de observaciones, la media, la desviación estándar, valor mínimo y máximo de cada variable. En este caso, solo aparecen aquellas explícitamente en el modelo (*véase* figura 1.3).

Figura 1.3. Salida comando summary



Fuente: cálculo autores

1.4.2 Estimación del modelo por MCO y pruebas de especificación.

Después del análisis general de las variables, es posible estimar el modelo propuesto (*véase* ecuación 1.62). En esta sección, se realizaran estimaciones a través de mínimos cuadrados ordinarios y verificación su especificación, sin tener en cuenta el posible problema de endogeneidad.

1. Para ejecutar una regresión lineal por mínimos cuadrados ordinarios, se aplica el comando *regress* –o *reg*-. Este comando, muestra además la prueba t de significancia individual para cada una de las variables, una prueba F de significancia conjunta y bondad de ajuste R^2 (*véase* figura 1.4).
2. Para probar la especificación de este modelo, se ejecuta la prueba Ramsey-RESET con el comando *estat ovtest* después de la regresión (*véase* figura 1.4).

Figura 1.4. Salida regresión lineal y prueba Ramsey-RESET

```
reg  lpibpc_exp_1892 laltura lind_fertilidad lprecipitac
d_magdalena d_cauca dist_barranq dist_buenaven
dist_bogota dist_cali lrios ltemperatura lhumedad
d_conflic_27_1917, robust

estat ovtest
```

```
. reg  lpibpc_exp_1892 laltura lind_fertilidad lprecipitac d_magdalena d_cauca dist_barra
> nq dist_buenaven dist_bogota dist_cali lrios ltemperatura lhumedad d_conflic_27_1917, r
> obust

Linear regression                               Number of obs =       738
                                                F( 13,   724) =      8.07
                                                Prob > F      =    0.0000
                                                R-squared     =    0.1091
                                                Root MSE      =    .76909

------------------------------------------------------------------------------
             |               Robust
lpibpc_~1892 |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
     laltura |   .2206594   .0428304     5.15   0.000     .1365728    .304746
lind_ferti~d |  -.0504321   .0922884    -0.55   0.585    -.2316169   .1307527
 lprecipitac |  -.0183078   .0710128    -0.26   0.797    -.1577235   .1211079
 d_magdalena |   .0454098   .1159472     0.39   0.695    -.1822232   .2730427
     d_cauca |  -.2560917   .1465116    -1.75   0.081      -.54373   .0315466
dist_barranq |    .036013   .0335531     1.07   0.283     -.02986    .101886
dist_buena~n |  -.1398699   .2084406    -0.67   0.502    -.5490902   .2693503
 dist_bogota |    .021701   .0268862     0.81   0.420    -.0310832   .0744853
   dist_cali |    .150044   .2266614     0.66   0.508    -.2949481    .595036
       lrios |  -.8265724    .590079    -1.40   0.162    -1.985043   .3318979
ltemperatura |   1.608064   .2289607     7.02   0.000     1.158558   2.05757
    lhumedad |  -.0145888   .0744999    -0.20   0.845    -.1608504   .1316728
d_confl~1917 |   .0971148   .0896939     1.08   0.279    -.0789763   .273206
       _cons |  -4.057601   1.550808    -2.62   0.009    -7.102219  -1.012983
------------------------------------------------------------------------------

. estat ovtest

Ramsey RESET test using powers of the fitted values of lpibpc_exp_1892
       Ho:  model has no omitted variables
                 F(3, 721) =      6.05
                  Prob > F =    0.0005

.
```

Fuente: cálculo autores

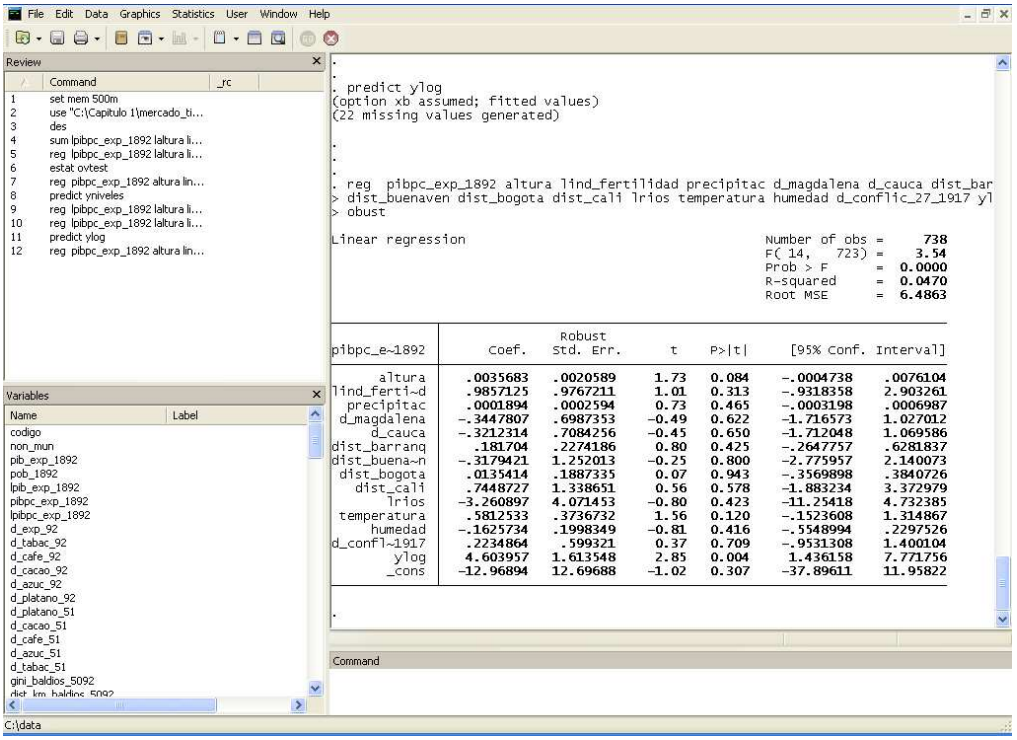
La regresión lineal de la figura 1.4, muestra como la variable de interés resulta no significativa y con el signo contrario al esperado. En relación a la prueba de Ramsey-RESET, el estadístico F tiene un valor de 6.05 con un p-valor de 0.0005[15], lo que indica mala especificación funcional.

3. Para probar alternativas de forma funcional, se compara esta especificación con una alternativa usando la prueba Davidson-MacKinnon. Desafortunadamente, Stata® no cuenta con un comando que ejecute este procedimiento de manera automática, por lo que hay que hacerlo paso a paso (véase figura 1.5).

[15] El p-valor puede interpretarse como la probabilidad de error al rechazar la hipótesis nula.

Figura 1.5. Salida prueba Davidson-MacKinnon

```
Command
reg  pibpc_exp_1892 altura lind_fertilidad precipitac d_magdalena d_cauca
dist_barranq dist_buenaven dist_bogota dist_cali lrios temperatura
humedad d_conflic_27_1917, robust

predict yniveles

reg  lpibpc_exp_1892 laltura lind_fertilidad lprecipitac d_magdalena
d_cauca dist_barranq dist_buenaven dist_bogota dist_cali lrios
ltemperatura lhumedad d_conflic_27_1917 yniveles, robust

reg  lpibpc_exp_1892 laltura lind_fertilidad lprecipitac d_magdalena
d_cauca dist_barranq dist_buenaven dist_bogota dist_cali lrios
ltemperatura lhumedad d_conflic_27_1917 , robust

predict ylog

reg  pibpc_exp_1892 altura lind_fertilidad precipitac d_magdalena d_cauca
dist_barranq dist_buenaven dist_bogota dist_cali lrios temperatura
humedad d_conflic_27_1917 ylog, robust
```



Fuente: cálculo autores

En este ejemplo particular, el valor estimado ($\hat{Y}$) del modelo en logaritmos resulta significativo en el modelo lineal original, aunque viceversa $\hat{Y}$ no es representativo; sobre el modelo en logaritmos, indicando que la función en

logaritmos está correctamente especificada comparada con el modelo lineal, razón por la cual se continua el análisis con esta especificación.

1.4.3 Estimación del modelo por MC2E.

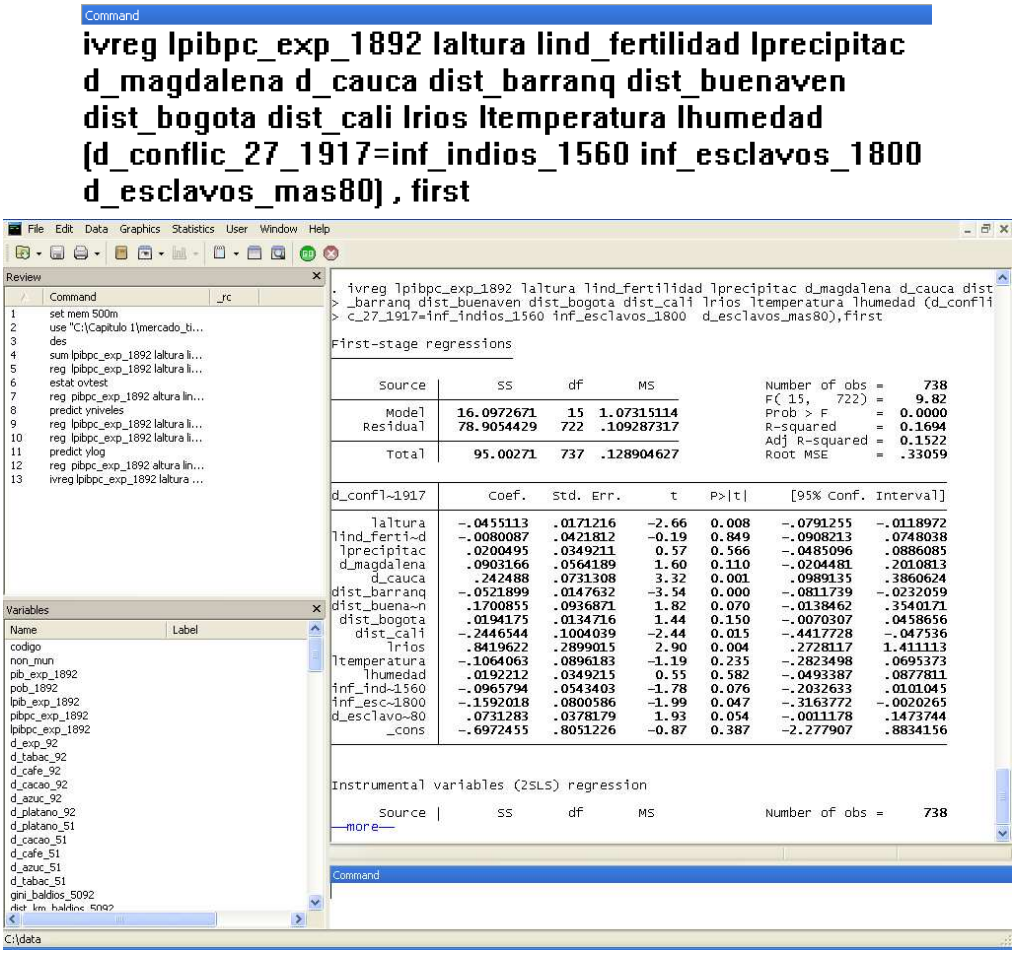
De acuerdo con lo expuesto en la introducción, posiblemente existe endogeneidad en la especificación anterior. Ante esto, resulta necesario estimar el ejercicio por mínimos cuadrados en dos etapas; para lo cual la base de datos contiene variables que representan la ubicación de cada municipalidad en relación a un conjunto de centros institucionales coloniales (encomiendas y centros productivos con presencia de más de 20 esclavos para 1800), útiles como instrumentos. De acuerdo a esto, se plantea el modelo reducido para la primera etapa de la estimación por MC2E (*véase* ecuación 1.63).

$$ConflictoTierras_i = \mathbf{Z}\boldsymbol{\pi} + \mathbf{X}\boldsymbol{\gamma} + \upsilon_i \qquad (1.63)$$

En la ecuación 1.63, $ConflictoTierras_i$ corresponde a la variable endógena, $\mathbf{Z}$ corresponde a un vector con los diferentes instrumentos (inf_indios_1560, inf_esclavos_1800 y d_esclavos_mas80, *véase* cuadro 1.3) y $\mathbf{X}$ las variables exógenas del modelo inicial. El vector $\boldsymbol{\pi}$ corresponde a los coeficientes de los instrumentos y $\boldsymbol{\gamma}$ el de las variables geográficas.

1. Para calcular una regresión por MC2E, se utiliza el comando *ivreg*. Deben listarse en orden las variables dependiente, independientes y endógenas con sus respectivos instrumentos. Para observar el resultado de la primera etapa, se añade la opción *first* (*véase* figura 1.6).

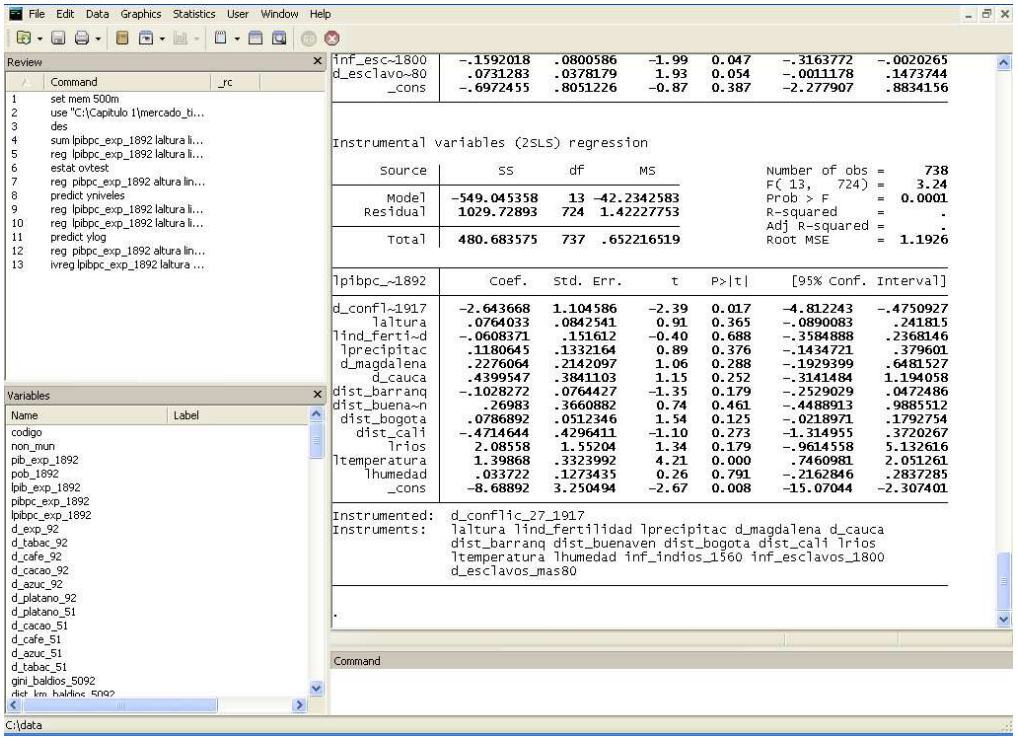
Figura 1.6. Salida regresión por MC2E



Fuente: cálculo autores

Los tres instrumentos usados en esta primera etapa, resultan significativos individualmente con estadísticos t de 1.78, 1.99 y 1.93, y conjuntamente con un F de 9.82. Lo anterior indica su relevancia (*véase* sección 1.3.2), considerándose buenos instrumentos.

Figura 1.7. Salida MC2E (cont.)



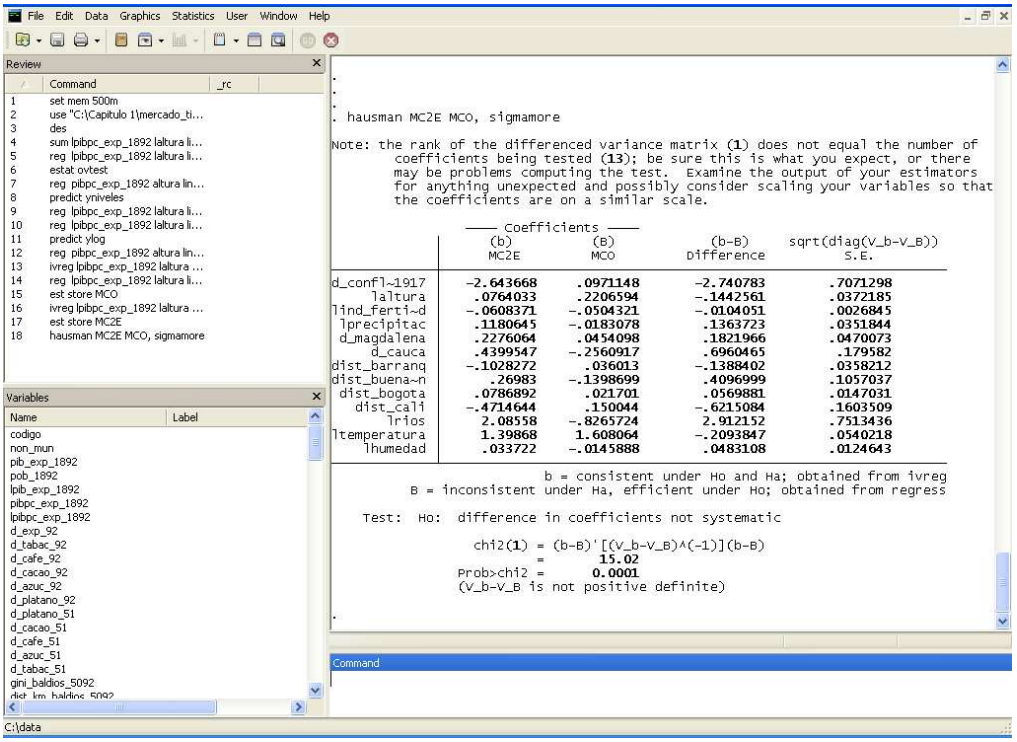
Fuente: cálculo autores

En la segunda etapa (*véase* figura 1.7), se observa como el coeficiente que acompaña a la variable $ConflicoTierras_i$ tiene ahora el signo esperado y es significativo, con estadístico t de 2.39 y p-valor de 0.017. Este resultado confirmaría la hipótesis central del documento.

2. Con el fin de probar la hipótesis de endogeneidad mediante la prueba de Hausman, es necesario estimar el modelo inicial por MCO y MC2E (guardando cada regresión con el comando *estimation store*), para luego comparar estadísticamente los estimadores con el comando *hausman* y la opción *sigmamore* (*véase* figura 1.8).

Figura 1.8. Salida prueba de Hausman

```
reg  lpibpc_exp_1892 laltura lind_fertilidad lprecipitac d_magdalena
d_cauca dist_barranq dist_buenaven dist_bogota dist_cali lrios
ltemperatura lhumedad d_conflic_27_1917

est store MCO

ivreg lpibpc_exp_1892 laltura lind_fertilidad lprecipitac d_magdalena
d_cauca dist_barranq dist_buenaven dist_bogota dist_cali lrios
ltemperatura lhumedad (d_conflic_27_1917=inf_indios_1560 inf_esclavos_
1800  d_esclavos_mas80),first
est store MC2E

hausman MC2E MCO, sigmamore
```

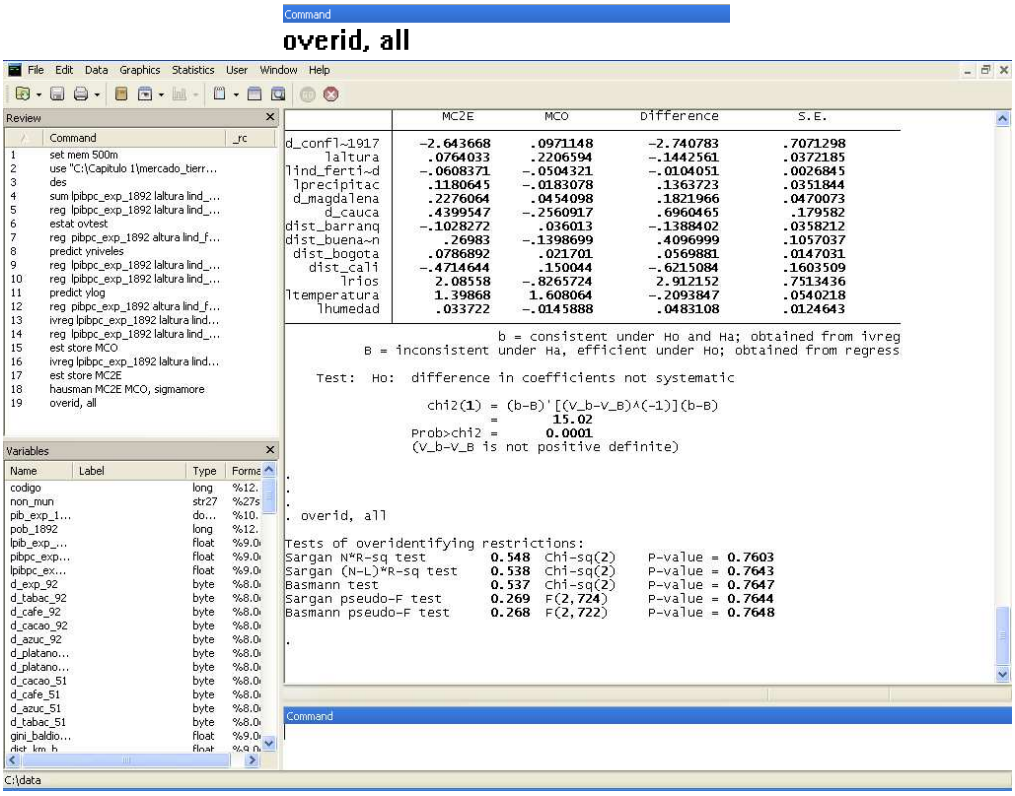


Fuente: cálculo autores

El estadístico χ^2 -reportado en la figura 1.8 como ji cuadrado-, tiene un valor de 15.02; rechazando así, la hipótesis nula de exogeneidad, con p-valor de 0.01%. En conclusión, es posible afirmar que el modelo inicial es endógeno.

3. En la primera etapa de la regresión por MC2E, se comprobó la relevancia de los instrumentos propuestos por Sanchez Et al. (2008). A continuación, para verificar su validez, se analiza una prueba de restricciones sobreidentificadas de Sargan con el comando *overid , all*. (*véase* figura 1.9).

Figura 1.9. Salida prueba de restricciones sobreidentificadas



Fuente: cálculo autores

En la figura 1.9, se observa que con ninguno de los estadísticos reportados se rechaza la hipótesis nula. En particular, el estadístico *ML* estudiado (Sargan Statistic N*R-sq en la salida), tiene un valor de 0.54, con p-valor de 0.7603. De lo anterior, se concluye que los instrumentos usados son exógenos, por lo cual el procedimiento de mínimos cuadrados en dos etapas es correcto y prima sobre MCO.

Conociendo que el estimador de MC2E es confiable, existe evidencia estadística que prueba la hipótesis inicial se cumple; la baja integración de Colombia a los

mercados mundiales para finales del Siglo XIX, fue el resultado de la mala calidad institucional que se vio reflejada en la aparición de conflictos por el control de tierras, en la frontera de colonización agrícola del país.

Resumen

- Especificar apropiadamente el modelo econométrico es necesario para obtener estimadores insesgados y consistentes. El problema de especificación puede ocurrir como consecuencia de omitir variables teóricas relevantes, añadir variables innecesarias o acudir a forma funcional incorrecta.
- Para detectar y establecer si el modelo econométrico estimado mediante MCO presenta problemas de especificación, es posible realizar las pruebas de Ramsey-RESET, J-Davidson-MacKinnon y multiplicador de Lagrange (ML).
- La alternativa más utilizada para corregir problemas de especificación por variables no observadas en el término de error, es el uso de variables aproximadas o proxy.
- El incumplimiento del supuesto de independencia condicional – o problema de endogeneidad- resulta como consecuencia de variables independientes omitidas correlacionadas con las incluidas; errores de medición y muestreo o problemas de doble causalidad entre variables explicativas y explicada. Bajo endogeneidad, los estimadores de MCO resultan sesgados e inconsistentes.
- La primera alternativa para solventar el problema de endogeneidad, consiste en incluir las variables omitidas directamente o a través de aproximaciones.
- Otra alternativa para calcular estimadores insesgados y consistentes, cuando persiste endogeneidad, es a través de MC2E. Para esto, es necesario contar con variables instrumentales.
- Las variables instrumentales deben cumplir con dos condiciones: validez y relevancia. La primera, puede ser verificada a través de una prueba de restricciones sobreidentificadas de Sargan, siempre y cuando se cuente con más de un instrumento candidato. La segunda, se verifica a través de una prueba de significancia en la primera etapa del modelo (reducido).
- La detección de endogeneidad en un modelo econométrico, se realiza mediante la prueba de Hausman. Esta prueba compara los resultados entre

las regresiones por MCO y MC2E, e identifica las diferencias estadísticas entre ambos.

Anexo 1

Anexo 1.1 Prueba de endogeneidad para mínimos cuadrados ordinarios (MCO)

Se parte de un modelo inicial endógeno, en notación matricial (*véase* ecuación A.1.1).

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + e \qquad \text{(A.1.1)}$$

Donde $\mathbf{Y}$ es la variable explicativa; $\mathbf{X}$ es una matriz compuesta por los vectores de las variables explicativas del modelo a consideración; β es el vector de coeficientes acompañan a las variables explicativas); y e es el vector de errores.

En particular, asumiendo que se estima el modelo por MCO, se obtienen estimadores para cada parámetro poblacional matricial (*véase* ecuación A.1.2).

$$\hat{\boldsymbol{\beta}}_{mco} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \qquad \text{(A.1.2)}$$

Para probar que Betas son sesgados, se calcula el valor esperado de la expresión A.1.2.

$$E[\hat{\boldsymbol{\beta}}_{mco}] = E[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})]$$
$$E[\hat{\boldsymbol{\beta}}_{mco}] = E[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + e))]$$
$$E[\hat{\boldsymbol{\beta}}_{mco}] = E[(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'e)]$$
$$E[\hat{\boldsymbol{\beta}}_{mco}] = E[\boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'e)]$$
$$E[\hat{\boldsymbol{\beta}}_{mco}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1}E[(\mathbf{X}'e)]$$

De lo anterior, se espera que $E[(\mathbf{X}'e)] = 0$ con el fin de obtener un estimador de MCO cercano al parámetro poblacional β; equivalente a variables explicativas no relacionadas con el error. Sin embargo, dado que se presume endogeneidad en el modelo representado en A.12 existe una matriz $(\mathbf{C})$ de variables omitidas con un vector de parámetros α, y un vector errores estocásticos u (*véase* ecuación A.1.3).

$$e = \mathbf{C}\boldsymbol{\alpha} + u \qquad \text{(A.1.3)}$$

Continuando con el procedimiento anterior, a partir de la ecuación A.1.3:

$$E[\mathbf{X}'e] = E[\mathbf{X}'(\alpha\mathbf{C} + u)]$$
$$E[\mathbf{X}'e] = E[\mathbf{X}'\alpha\mathbf{C}] + E[\mathbf{X}'u]$$
$$E[\mathbf{X}'e] = \alpha E[\mathbf{X}'\mathbf{C}] + E[\mathbf{X}'u]$$

Lo anterior, evidencia que las variables explicativas en el modelo están recogiendo información de las omitidas, por lo que el término $\alpha E[\mathbf{X}'\mathbf{C}]$ es diferente de cero ($\text{cov}(\mathbf{X}, \mathbf{C}) \neq 0$). Por consiguiente, los estimadores de MCO son sesgados.

Análogamente, para la prueba de consistencia, suponga el modelo de forma matricial en A.1.4.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + e \qquad \text{(A.1.4)}$$

Adicionalmente suponga que:

$$\frac{1}{N}\sum_{i=1}^{N} X_i' X_i \xrightarrow{P} Q \qquad \text{(A.1.5)}$$

Partiendo de la expresión $\hat{\boldsymbol{\beta}}_{mco} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y})$, se prueba la forma en que el estimador se aproxima al parámetro poblacional, cuando N (tamaño de muestra) tiende infinito. Para ello, inicialmente se multiplica por $\frac{1}{N}$ cada lado (*véase* ecuación A.1.6).

$$\hat{\boldsymbol{\beta}}_{mco} = (\frac{1}{N}\mathbf{X}'\mathbf{X})^{-1}(\frac{1}{N}\mathbf{X}'\mathbf{Y}) \qquad \text{(A.1.6)}$$

Reemplazando la expresión de $\mathbf{Y}$ y aplicando el $p\lim$ para $\hat{\boldsymbol{\beta}}_{mco}$, se ratifica el comportamiento del estimador en muestras grandes,

$$\hat{\boldsymbol{\beta}}_{mco} = (\frac{1}{N}\mathbf{X}'\mathbf{X})^{-1}(\frac{1}{N}\mathbf{X}'\mathbf{Y})$$

$$\hat{\boldsymbol{\beta}}_{mco} = (\frac{1}{N}\mathbf{X}'\mathbf{X})^{-1}(\frac{1}{N}\mathbf{X}'(\mathbf{X}\boldsymbol{\beta} + e))$$

$$\hat{\boldsymbol{\beta}}_{mco} = (\frac{1}{N}\mathbf{X}'\mathbf{X})^{-1}(\frac{1}{N}\mathbf{X}'\mathbf{X}\boldsymbol{\beta}) + (\frac{1}{N}\mathbf{X}'\mathbf{X})^{-1}(\frac{1}{N}\mathbf{X}'e)$$

$$\hat{\boldsymbol{\beta}}_{mco} = \boldsymbol{\beta} + (\frac{1}{N}\mathbf{X}'\mathbf{X})^{-1}(\frac{1}{N}\mathbf{X}'e)$$

$$p\lim\hat{\boldsymbol{\beta}}_{mco} = \boldsymbol{\beta} + Q^{-1}(\frac{1}{N}\mathbf{X}'e)$$

Para que el estimador resulte consistente $\frac{1}{N}\mathbf{X}'e$ deber ser igual a cero. Dado el problema de endogeneidad -$Cov(X, \mu) \neq 0$-, el termino $\frac{1}{N}\mathbf{X}'e$ es diferente de cero, con lo cual $p\lim\hat{\boldsymbol{\beta}}_{mco} \neq \boldsymbol{\beta}$.

Esto demuestra que los estimadores de MCO son inconsistentes ante la presencia de endogeneidad.

Anexo 1.2 Variables aproximación o proxy como alternativa para resolver endogeneidad

Retomando la demostración anterior,

$$E[\mathbf{X}'e] = E[\mathbf{X}'(\alpha\mathbf{C} + u)]$$
$$E[\mathbf{X}'e] = E[\mathbf{X}'\alpha\mathbf{C}] + E[\mathbf{X}'u]$$

Por endogeneidad, $E[\mathbf{X}'\alpha C] \neq 0$, existe sesgo en los $\hat{\boldsymbol{\beta}}_{mco}$; sin embargo, al incluir la variable aproximación o proxy, el vector $\mathbf{C}$ queda vacío, permite que los estimadores del modelo recobren sus propiedades. Teniendo en cuenta que u es un componente estocástico, se deduce:

$$E[\mathbf{X}'e] = E[\mathbf{X}'u] = 0 \qquad \text{(A.1.7)}$$

Por lo que,

$$E[\hat{\boldsymbol{\beta}}_{mco}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} E[(\mathbf{X}'e)]$$
$$E[\hat{\boldsymbol{\beta}}_{mco}] = \boldsymbol{\beta} + (\mathbf{X}'\mathbf{X})^{-1} E[(\mathbf{X}'u)]$$
$$E[\hat{\boldsymbol{\beta}}_{mco}] = \boldsymbol{\beta}$$

Los estimadores de mínimos cuadrados ordinarios, resultan ser insesgados.

Para la prueba de consistencia, se tiene:

$$\hat{\boldsymbol{\beta}}_{mco} \xrightarrow{P} \boldsymbol{\beta} + Q^{-1}(\frac{1}{N}\mathbf{X}'e)$$
$$\hat{\boldsymbol{\beta}}_{mco} \xrightarrow{P} \boldsymbol{\beta} + Q^{-1}(\frac{1}{N}\mathbf{X}'(\alpha\mathbf{C} + \mu))$$

Incluyendo la variable, el vector $\mathbf{C}$ queda vacío entonces:

$$\hat{\boldsymbol{\beta}}_{mco} \xrightarrow{P} \boldsymbol{\beta} + Q^{-1}(\frac{1}{N}\mathbf{X}'(e))$$
$$\hat{\boldsymbol{\beta}}_{mco} \xrightarrow{P} \boldsymbol{\beta}$$

Anexo 1.3. Derivación del estimador de variables instrumentales bajo un modelo de regresión simple.

$$\begin{aligned}
&\mathbf{Y} = \mathbf{X\beta} + e \\
&E[\mathbf{Y}] = E[\mathbf{X\beta}] + E[e] \\
&\mathbf{Y} - E[Y] = \mathbf{X\beta} - E[\mathbf{X\beta}] + e - E[e]
\end{aligned} \tag{A.1.8}$$

$$\begin{aligned}
&[\mathbf{Y} - E[\mathbf{Y}]][\mathbf{X} - E[\mathbf{X}]] = [\mathbf{X} - E[\mathbf{X}]]^2\boldsymbol{\beta} + [e - E[e]][\mathbf{X} - E[\mathbf{X}]] \\
&\underbrace{E\left\{[\mathbf{Y} - E[\mathbf{Y}]][\mathbf{X} - E[\mathbf{X}]]\right\}}_{Cov(\mathbf{X},\mathbf{Y})} = \underbrace{E\left\{[\mathbf{X} - E[\mathbf{X}]]^2\right\}}_{Var(\mathbf{X})}\beta + \underbrace{E\left\{[e - E[e]][\mathbf{X} - E[\mathbf{X}]]\right\}}_{Cov(\mathbf{X},e)}
\end{aligned} \tag{A.1.9}$$

$$\begin{aligned}
&Cov(\mathbf{X},\mathbf{Y}) = Var(\mathbf{X})\boldsymbol{\beta} + Cov(\mathbf{X},e) \\
&\frac{Cov(\mathbf{X},\mathbf{Y})}{Var(\mathbf{X})} = \boldsymbol{\beta} + \frac{Cov(\mathbf{X},e)}{Var(\mathbf{X})}
\end{aligned} \tag{A.1.10}$$

Si $Cov(\mathbf{X},e) = 0$, entonces el estimador $\hat{\boldsymbol{\beta}}_{MCO}$ VI es insesgado. Si $Cov(\mathbf{X},e) \neq 0$, existe problema de endogeneidad. Para el segundo caso es posible utilizar un conjunto de variables instrumentales $\mathbf{Z}$ para solucionar el problema. Una vez se tiene el procedimiento de mínimos cuadrados en dos etapas, se tiene a partir de $Cov(\mathbf{Z},\mathbf{Y})$ un nuevo estimador.

$$\begin{aligned}
&Cov(\mathbf{Z},\mathbf{Y}) = Cov(\mathbf{Z},\hat{\mathbf{X}}\boldsymbol{\beta} + u) \\
&Cov(\mathbf{Z},\mathbf{Y}) = Cov(\mathbf{Z},\hat{\mathbf{X}}\boldsymbol{\beta}) + Cov(\mathbf{Z},u) \\
&Cov(\mathbf{Z},\mathbf{Y}) = Cov(\mathbf{Z},\hat{\mathbf{X}})\boldsymbol{\beta} + Cov(\mathbf{Z},u)
\end{aligned} \tag{A.1.11}$$

Si Z es un buen instrumento, éste es exógeno, es decir, $Cov(\mathbf{Z},\mathbf{u}) = 0$. Por tanto el estimador de variables instrumentales tiene la forma.

$$\hat{\boldsymbol{\beta}}_{VI} = \frac{Cov(\mathbf{Z},\mathbf{Y})}{Cov(\mathbf{Z},\hat{\mathbf{X}})}$$

Anexo 1.4. Consistencia de mínimos cuadrados en dos etapas (MC2E)

En forma matricial, es posible obtener el estimador de MC2E fácilmente. Partiendo de un modelo inicial simple, con al menos una variable endógena.

$$\mathbf{Y} = \delta\mathbf{X} + e \qquad \text{(A.1.12)}$$

La primera etapa de la regresión viene dada por

$$\mathbf{X} = \mathbf{Z}\boldsymbol{\pi} + \upsilon \qquad \text{(A.1.13)}$$

Donde $\mathbf{Z}$ es una matriz que incluye los instrumentos de $\mathbf{X}$. El estimador de mínimos cuadrados ordinarios de $\boldsymbol{\pi}$ viene dado por:

$$\hat{\boldsymbol{\pi}}_{mco} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X}) \qquad \text{(A.1.14)}$$

De la primera etapa se obtienen los $\hat{\mathbf{X}}$ que tienen la forma $\hat{\mathbf{X}} = \mathbf{Z}\hat{\boldsymbol{\pi}}$. El estimador de MC2E tiene la forma,

$$\hat{\boldsymbol{\pi}}_{mc2e} = (\hat{\mathbf{X}}'\hat{\mathbf{X}})^{-1}(\hat{\mathbf{X}}'\mathbf{Y})$$
$$\hat{\boldsymbol{\pi}}_{mc2e} = (\hat{\boldsymbol{\pi}}'\mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\pi}})^{-1}(\hat{\boldsymbol{\pi}}'\mathbf{Z}'\mathbf{Y})$$

Reemplazando $\hat{\boldsymbol{\pi}}_{mco} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'\mathbf{X})$,

$$\hat{\boldsymbol{\pi}}_{mc2e} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}$$
$$\hat{\boldsymbol{\pi}}_{mc2e} = [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y} \qquad \text{(A.1.15)}$$

La expresión matricial del estimador de mínimos cuadrados ordinarios, viene dada por $\hat{\boldsymbol{\pi}}_{mc2e}$. Para comprobar que los estimadores son insesgados, se debe obtener el valor esperado de la expresión $\hat{\boldsymbol{\pi}}_{mc2e}$ y revisar que el estimador lleve al parámetro poblacional.

$$E[\hat{\boldsymbol{\pi}}_{\mathbf{mc2e}}] = E[[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Y}]$$
$$E[\hat{\boldsymbol{\pi}}_{\mathbf{mc2e}}] = E[[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(\mathbf{X}\boldsymbol{\delta} + e)]$$
$$E[\hat{\boldsymbol{\pi}}_{\mathbf{mc2e}}] = E[[\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}\boldsymbol{\delta} + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'e)]$$
$$E[\hat{\boldsymbol{\pi}}_{\mathbf{mc2e}}] = E[\boldsymbol{\delta} + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'e]$$
$$E[\hat{\boldsymbol{\pi}}_{\mathbf{mc2e}}] = \boldsymbol{\delta} + [\mathbf{X}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{X}]^{-1}(\mathbf{X}'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}E[\mathbf{Z}'e]$$

Suponiendo que el instrumento usado es válido –o exógeno- se tiene que $E[\mathbf{Z}'e] = 0$, por lo que,

$$E[\hat{\boldsymbol{\pi}}_{mc2e}] = \boldsymbol{\delta} \qquad \text{(A.1.16)}$$

$$\hat{\boldsymbol{\beta}}_{mc2e} = (\mathbf{Z}'\mathbf{X})^{-1}\mathbf{Z}'\mathbf{Y} \qquad \text{(A.1.17)}$$

Esto demuestra que los estimadores de MC2E son insesgados aún ante la presencia de endogeneidad. La prueba de consistencia es análoga, siguiendo los pasos presentados en A.1.1.

Capítulo 2

Modelos de ecuaciones simultáneas

2.1 Introducción

En el capítulo anterior se discutieron los problemas de endogeneidad ocasionados por variables omitidas, errores de medición y sesgo de selección; así como las metodologías sobre variables proxy y mínimos cuadrados en dos etapas (MC2E) para su solución. Continuando con este tema, este capítulo contiene el problema de simultaneidad, correspondiente a un caso especial de endogeneidad.

La especificación teórica de los modelos uniecuacionales considera la variable dependiente, como única endógena. Aun así, en ocasiones se requiere trabajar con varias variables independientes que son endógenas en otras ecuaciones; conduciendo a determinarse simultáneamente o en un sentido bidireccional, originando ecuaciones simultáneas. Característica, que conduce al incumplimiento del supuesto de independencia condicional.

Por lo anterior, este capítulo introduce formalmente los modelos de ecuaciones simultáneas y presenta las diferentes técnicas para identificar el problema de simultaneidad con el respectivo método de estimación más conveniente. En particular se discuten las condiciones de orden y rango, que determinan cuando un sistema multiecuacional puede ser estimado; igualmente los métodos de mínimos cuadrados indirectos (MCI), mínimos cuadrados en dos etapas (MC2E), mínimos cuadrados en tres etapas (MC3E) y sistema de regresiones aparentemente no relacionados (SUR, por seemingly unrelated regression, en ingles).

Finalmente, se aplican las metodologías expuestas mediante dos estudios de caso. El primero, basado en el artículo *"Análisis empírico del fondo de estabilización de precios del azúcar en Colombia"* de Vásquez (2005) y el segundo, a partir de un ejemplo disponible en Hill, et Al. (1993) acerca de la oferta regional en Estados Unidos (E.U.).

2.2 El problema de simultaneidad

En general, los modelos económicos constituidos para representar el funcionamiento de procesos y fenómenos como: formación de precios, crecimiento del PIB y decisiones sobre consumo e inversión de los individuos. Dado que en estos modelos, existe cierta interdependencia entre variables endógenas. Esta sección introduce el uso de sistemas de ecuaciones simultáneas en econometría, para estimar adecuadamente el problema de simultaneidad.

2.2.1 Modelo de ecuaciones simultáneas

Los modelos econométricos estudiados hasta ahora constan de una ecuación, en este capítulo se estudia un sistema conformado por varias ecuaciones simultáneamente, cuyas características principales son las siguientes:

1. Relación bidireccional entre las variables dependiente e independiente.
2. Las variables endógenas se determinan conjuntamente.
3. Se tienen tantas ecuaciones como variables endógenas.
4. Los errores no deben estar correlacionados entre ecuaciones (ausencia de autocorrelación contemporánea).

Desde la perspectiva de causalidad, las definiciones sobre variables dependientes e independientes manejados en modelos uniecuacionales, son reemplazadas por los conceptos de variables endógenas (aquellas que se determinan dentro del modelo) y exógenas (determinadas fuera de este o predeterminadas). Un ejemplo de un modelo con dos ecuaciones simultáneas, dos variables endógenas (Y_{i1} y Y_{i2}) y ninguna variable exógena, se presenta en las ecuaciones 2.1 y 2.2.

$$Y_{i1} = \beta_1 Y_{i2} + e_{i1} \qquad (2.1)$$

$$Y_{i2} = \alpha_1 Y_{i1} + e_{i2} \qquad (2.2)$$

En la ecuación 2.1, el efecto de la variable Y_{i2} sobre Y_{i1}, es capturado por el parámetro poblacional β_1 y viceversa en la ecuación 2.2 viene dado por α_1. La representación anterior, se conoce como *forma estructural* del modelo de ecuaciones simultáneas, donde las ecuaciones y parámetros son denominados *estructurales*. De esta forma, un modelo de ecuaciones simultáneas es considerado completo si se tiene una ecuación estructural por cada variable endógena (Judge et Al., 1988, Cap. 14). Adicionalmente, es posible complementar este modelo incluyendo variables exógenas (X_{i1} y X_{i2}) (*véase* ecuaciones 2.3 y 2.4).

$$Y_{i1} = \beta_1 Y_{i2} + \beta_2 X_{i1} + e_{i1} \qquad (2.3)$$

$$Y_{i2} = \alpha_1 Y_{i1} + \alpha_2 X_{i2} + e_{i2} \qquad (2.4)$$

No obstante, en las ecuaciones anteriores el incumplimiento de independencia condicional para el caso de simultaneidad se manifiesta en la cov(y2, e1)#0 o cov(y1, e2)#0. De igual forma que en endogeneidad, estimar estas ecuaciones individualmente por MCO conduce a coeficientes sesgados e inconsistentes (*véase* anexo 2).

Por otra parte, existe una forma adicional para representar el problema de simultaneidad donde las variables endógenas se constituyen únicamente en función de las exógenas, parámetros estructurales y errores estocásticos; representación conocida como *forma reducida* (Wooldridge, 2009, Cap. 16), igual que en endogeneidad, cada forma estructural tiene relacionada una reducida. Así, remplazando Y_{i2} de la ecuación 2.4 en 2.3 y asumiendo $\beta_1\alpha_1 \neq 1$, es deducida la ecuación 2.7; la cual puede estimarse por MCO. Los coeficientes π_1 y π_2 se conocen como *parámetros reducidos*. Una expresión similar puede obtenerse para representar Y_{i1}, si es reemplazada de la ecuación 2.3 en 2.4.

$$Y_{i1} = \beta_1(\alpha_1 Y_{i1} + \alpha_2 X_{i2} + e_{i2}) + \beta_2 X_{i1} + e_{i1} \qquad (2.5)$$

$$Y_{i1} = \beta_1\alpha_1 Y_{i1} + \beta_1\alpha_2 X_{i2} + \beta_1 e_{i2} + \beta_2 X_{i1} + e_{i1} \qquad (2.6)$$

$$Y_{i1} = \underbrace{\frac{\beta_2}{(1-\beta_1\alpha_1)}}_{\tilde{\pi}_1} X_{i1} + \underbrace{\frac{\beta_1\alpha_2}{(1-\beta_1\alpha_1)}}_{\tilde{\pi}_2} X_{i2} + \underbrace{\frac{\beta_1 e_{i2} + e_{i1}}{(1-\beta_1\alpha_1)}}_{\tilde{\mu}_i} \qquad (2.7)$$

Sin embargo, puede especificarse un sistema con *M* número de ecuaciones y variables endógenas; las cuales pueden representarse fácilmente mediante representación matricial (*véase* anexo 2). Notación, que también aplica en un sistema con dos ecuaciones (*véase* ecuación 2.8).

$$\begin{aligned} &Y_{i1} - \beta_1 Y_{i2} - \beta_2 X_{i1} - e_{i1} = 0 \\ &\underbrace{Y_{i2} - \alpha_1 Y_{i1} - \alpha_2 X_{i2} - e_{i2}}_{\mathbf{Y}\Gamma + \mathbf{X}B + \mathbf{E}} = 0 \end{aligned} \qquad (2.8)$$

En la ecuación 2.8, $\mathbf{Y}$, $\mathbf{X}$ y $\mathbf{E}$ corresponden a matrices que contiene las variables endógenas, exógenas y errores del sistema, respectivamente[16]. $\mathbf{\Gamma}$ es el vector de parámetros de las variables endógenas, $\mathbf{B}$ el de las exógenas. Notación que representa matricialmente las ecuaciones estructurales del modelo. A partir de ella, es posible obtener la forma reducida del mismo (*véase* ecuación 2.9 y Anexo A.2.1).

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\pi} + \mathbf{V} \qquad (2.9)$$

Donde $\boldsymbol{\pi} = -\mathbf{B}\mathbf{\Gamma}^{-1}$ y $\mathbf{V} = -\mathbf{E}\mathbf{\Gamma}^{-1}$

El objetivo principal en un sistema de ecuaciones simultáneas es encontrar los estimadores de la forma estructural a partir de los parámetros de la reducida (*véase* anexo 3). Con este fin, las secciones siguientes presentan el concepto de identificación del modelo y varios métodos de estimación recomendados, según el caso correspondiente.

2.2.2 Sesgo de MCO bajo ecuaciones simultáneas

Una vez introducido el modelo de ecuaciones simultáneas, es posible analizar formalmente el incumplimiento del supuesto de independencia condicional. Recordando lo ilustrado en el capítulo 1, un modelo incumple esta condición

[16] Formalmente, $\mathbf{Y} = \begin{bmatrix} Y_{i1} Y_{i2} \end{bmatrix}$ y $\mathbf{X} = \begin{bmatrix} X_{i1} X_{i2} \end{bmatrix}$

cuando existe relación entre alguna de sus variables independientes y el término de error, conllevando a estimadores sesgados e inconsistentes por MCO. Desarrollando la ecuación 2.7, es posible demostrar que un modelo de ecuaciones simultáneas, es siempre endógeno (*véase* ecuación 2.10).

$$Y_{i1} = \frac{\beta_2}{(1-\beta_1\alpha_1)} X_{i1} + \frac{\beta_1\alpha_2}{(1-\beta_1\alpha_1)} X_{i2} + \frac{\beta_1}{(1-\beta_1\alpha_1)} e_{i2} + \frac{1}{(1-\beta_1\alpha_1)} e_{i1} \qquad (2.10)$$

De la ecuación 2.10, se deduce que existe correlación entre Y_{i1} y e_{i2} siempre y cuando $\frac{\beta_1\alpha_2}{(1-\beta_1\alpha_1)}$ resulte diferente de cero; dado que en la ecuación 2.3 X_{i2} determina a Y_{i2}, razón por la cual α_2 es diferente de cero. De igual forma, cuando Y_{i1} tenga un impacto en Y_{i2}, β_1 será un coeficiente significativo.

En la ecuación 2.10 existe simultaneidad, entonces la aplicación de MCO conduce a estimadores sesgados e inconsistentes o *sesgo de simultaneidad*. Su dirección viene determinado por la forma en que están relacionadas las variables endógenas. Igualmente, la significancia estadística del sesgo puede determinarse mediante la prueba estadística de Hausman, presentada a continuación. Ella, se constituye en la principal herramienta para detectar sesgos de los parámetros en un modelo econométrico.

2.3 Detección del problema: prueba de Hausman

Como se discutió anteriormente, la existencia de endogeneidad genera sesgos en las estimaciones de MCO. Por esta razón, resulta conveniente contar con herramientas que permitan evaluar la existencia de este problema en una ecuación específica. En esta sección, se presenta una generalización de la prueba de Hausman, que permiten diagnosticar el incumplimiento del supuesto de independencia condicional en un modelo de regresión lineal. Este procedimiento es el mismo presentado en el capítulo 1, donde se comparan los estimadores de MCO de una ecuación estructural, que estarían sesgados ante la presencia de este problema, con estimadores obtenidos de alguna otra metodología que garantice insesgabilidad y consistencia, como MC2E.

A continuación, se discute una versión de la prueba de Hausman más general que la ya discutida. En lugar de estimar los errores de una regresión, para luego probar la hipótesis a partir de una regresión auxiliar, se utiliza un estadístico que compara los coeficientes directamente. Ambas metodologías son equivalentes, y deberían conducir a los mismos resultados.

De manera general, la prueba de Hausman plantea que si los estimadores de MCO y MC2E no son estadísticamente diferentes, es posible concluir que el modelo no presenta problema de endogeneidad. Si por el contrario, los estimadores difieren estadísticamente, se asume que éste es el resultado de algún sesgo de endogeneidad y por lo tanto es necesario aplicar una metodología más sofisticada que MCO (*véase* prueba de hipótesis 2.11).

$$\begin{aligned} H_0 &: \beta_{mco} = \beta_{mc2e} && \text{No existe endogeneidad} \\ H_1 &: \beta_{mco} \neq \beta_{mc2e} && \text{Existe endogeneidad} \end{aligned} \qquad (2.11)$$

El estadístico de prueba, que se conoce como estadístico de Hausman, viene dado por la ecuación 2.12. El numerador $(\hat{\beta}_{mc2e} - \hat{\beta}_{mco})^2$ es la distancia entre los estimadores de mínimos cuadrados en dos etapas, y los de mínimos cuadrados ordinarios. El término del denominador $\text{var}\left[\hat{\beta}_{mc2e} - \hat{\beta}_{mco}\right]$ es la varianza conjunta de los estimadores.

$$H = \frac{(\hat{\beta}_{mc2e} - \hat{\beta}_{mco})^2}{\text{var}\left[\hat{\beta}_{mc2e} - \hat{\beta}_{mco}\right]} \sim \chi^2_k \qquad (2.12)$$

Si el valor del estadístico es mayor al valor crítico determinado por χ^2_k bajo el nivel de significancia deseado, se rechaza la hipótesis nula. En ese caso, se afirma que se está ante un problema de endogeneidad y el método de estimación recomendado podría ser MC2E. Si por el contrario no es posible rechazar la hipótesis nula, es posible asumir que no hay ningún sesgo relevante sobre MCO.

Resumiendo, el procedimiento general para la prueba de Hausman es:

1. Realizar la estimación de la ecuación estructural a estudiar por MCO.
2. Realizar estimación por alguna otra metodología, como en este caso MC2E.
3. Construir el estimador de Hausman y se verifica el resultado de la prueba de hipótesis.

En los casos de simultaneidad, donde no es recomendable la aplicación de MCO, se requiere determinar primero el estado de identificación de cada una de las ecuaciones del sistema, con el fin de establecer el método de estimación más apropiado. Esto se expone en la sección siguiente.

2.4 Proceso de identificación

De acuerdo a lo anterior, si se detecta la existencia de simultaneidad a través de la prueba de Hausman, es necesario realizar estimaciones por metodologías alternativas a MCO. Esta sección presenta el conjunto de condiciones, equivalentes a la condición mínima de orden discutida en el capítulo 1, a verificar antes de calcular nuevas estimaciones.

Como se discutió anteriormente, bajo el esquema de ecuaciones simultáneas, la doble causalidad evita encontrar estimadores estructurales mediante MCO. El proceso de identificación que se presenta a continuación, pretende determinar cuándo es posible encontrar los valores de los estimadores de la forma estructural (B y Γ, en la ecuación 2.8) a partir de estimaciones de la forma reducida (π, en la ecuación 2.9). Como ejemplo, en ciertos casos es posible aplicar MCO sobre la especificación de la forma reducida, para luego mediante una transformación, regresar a los parámetros estructurales.

De manera general, una ecuación puede catalogarse en una de tres categorías: en primer lugar se dice no identificada, cuando a partir de la forma reducida no es posible obtener estimaciones de los parámetros estructurales. Alternativamente, puede estar exactamente identificada, cuando a partir de la información de los parámetros de la forma reducida es posible encontrar un único valor para los parámetros de la forma estructural. Por último, una ecuación se dice sobreidentificada cuando a partir de la información disponible en la forma reducida se pueden establecer más de un valor para los parámetros de la forma

estructural. Lo anterior también se cumple para los sistemas de ecuaciones; un modelo de simultaneidad está exactamente identificado, cuando todas sus ecuaciones estructurales lo están.

Con el fin de definir el estado de cada una de las ecuaciones de un modelo, a continuación se presentan las condiciones de orden y rango, que son criterios formales para determinar el estado de identificación de un modelo.

2.4.1 Condición de orden

El primer criterio que se utiliza para definir la identificación del sistema de ecuaciones simultáneas, es la condición de orden, cuyo cumplimiento es necesario aunque no suficiente, para poder obtener estimadores de los parámetros poblacionales a partir de los coeficientes calculados de la forma reducida. En términos de notación, J corresponde al número de variables endógenas y exógenas del sistema que no aparecen ecuación de interés; y M al número total de variables endógenas o ecuaciones en el sistema.

Si en la expresión estructural de interés se tiene $J = M - 1$, se dice que la ecuación está exactamente identificada, lo que implica que a partir de la matriz π (*véase* ecuación 2.9) pueden encontrarse estimadores únicos de los parámetros estructurales del sistema. Si por el contrario $J > M - 1$, la ecuación está sobreidentificada, lo que conduce a varios estimadores de los parámetros estructurales del sistema. Finalmente, si $J < M - 1$, la ecuación se dice no identificada, por lo que no es posible obtener aproximaciones a los parámetros poblacionales.

Esta regla de orden, permite identificar fácilmente cuando un modelo no puede ser estimado (es decir, la no identificación); aun así, que $J \geq M - 1$ no implica necesariamente que la ecuación esté realmente identificada. Por esta razón, ésta se considera una regla aproximada. A continuación se presenta la condición de rango, que aunque es más difícil de calcular, corresponde a un criterio suficiente para determinar el grado de identificación de las ecuaciones en el sistema.

2.4.2 Condición de rango

El segundo criterio de identificación se conoce como condición de rango y, a diferencia de la condición de orden determina con exactitud el estado de identificación de cada una de las ecuaciones estructurales. Aun así, su cálculo es más complejo, al requerir establecer el rango de R, matriz de tamaño $[J \times (M + K)]$ compuesta por las variables exógenas de las ecuaciones que componen el sistema; y de Δ, matriz que contiene los parámetros de las variables endógenas y exógenas del sistema[17].

A partir de lo anterior, la regla de identificación plantea que si el rango de $(R_i\Delta) < M - 1$, la ecuación i no está identificada. Por el contrario, si el rango de $(R_i\Delta) = M - 1$ y rango de $R_i = M - 1$, se afirma que la ecuación i está exactamente identificada. Por último, si el rango de $(R_i\Delta) = M - 1$ y el de $R_i > M - 1$, se deduce que la ecuación i está sobreidentificada.

Una vez se ha determinado, a partir de las condiciones anteriores, el estado de identificación de las ecuaciones estructurales, debe aplicarse un método de estimación alternativo a MCO. A continuación se presentan diferentes metodologías que permiten de obtener estimadores insesgados de los parámetros estructurales.

2.5 Metodologías de estimación de ecuaciones simultáneas.

En esta sección se presentan tres metodologías para estimar los parámetros estructurales de un modelo multiecuacional, que varían en complejidad y precisión. Adicionalmente, se introducen los sistemas de regresiones aparentemente no relacionadas (SUR, por sus siglas en ingles), técnica que, a pesar de no estar diseñada para casos de simultaneidad, permite estimar sistemas de ecuaciones. En los casos más complejos, se usaran estimaciones combinadas de

[17] Formalmente, $\Delta = \begin{pmatrix} \mathbf{B} \\ \mathbf{\Gamma} \end{pmatrix}_{(M+(K\times M))}$

MCO y mínimos cuadrados generalizados (MCG), técnica que se estudia en los cursos básicos de econometría.[18]

2.5.1 Mínimos cuadrados indirectos (MCI)

La primera alternativa para encontrar estimadores de los parámetros estructurales, se conoce como mínimos cuadrados indirectos (MCI), y consiste en aplicar directamente MCO sobre la ecuación reducida, para luego indirectamente deducir las expresiones estructurales. Esta metodología se aplica a ecuaciones que están exactamente identificadas, por lo que se obtienen valores únicos para los parámetros poblacionales.

Para ilustrar el funcionamiento de esta técnica, suponga un sistema biecuacional simple, con dos variables endógenas (Y_{i1} y Y_{i2}), y una exógena (X_{i1}) (*véase* ecuaciones 2.13 y 2.14).

$$Y_{i1} - \beta_0 - \beta_1 Y_{i2} - \beta_2 X_{i1} - e_{i1} = 0 \qquad (2.13)$$

$$Y_{i2} - \alpha_0 - \alpha_1 Y_{i1} - \alpha_2 X_{i1} - e_{i2} = 0 \qquad (2.14)$$

La metodología de MCI se puede desarrollar fácilmente en 3 etapas:

1. De las ecuaciones 2.13 y 2.14 se obtiene la representación de la forma reducida del sistema, que es: $Y = X\pi + V$ (*véase* anexo A.2.2)
2. Se estiman los parámetros de la forma reducida por MCO. Ésta estimación es adecuada porque la forma reducida cuenta con variables exógenas (Gujarati, 2003, 740).
3. A través de la estimación de MCO se derivan los parámetros estructurales del sistema, utilizando la relación $\boldsymbol{\pi} = -\mathbf{B}\boldsymbol{\Gamma}^{-1}$.

Los estimadores de MCI son consistentes y eficientes para las ecuaciones exactamente identificadas. Aun así, al aplicar MCI no se dispone, al menos fácilmente, de la desviación estándar estimada de los parámetros, lo cual resulta

[18] Aquí se asume que el lector ya está familiarizado con el método de MCG. Para una introducción general a este tema, *véase* Guajarti (2003, 379).

inconveniente pues imposibilita la realización de pruebas de hipótesis relativas a los parámetros.

2.5.2 Mínimos cuadrados en dos etapas (MC2E)

En segundo lugar, está el método de MC2E que se presentó como alternativa al problema de endogeneidad en el capítulo 1. Esta metodología hace posible estimar los parámetros de las ecuaciones estructurales de interés directamente, reemplazando las variables endógenas por valores obtenidos a través de regresiones auxiliares, y puede ser aplicado tanto para ecuaciones exactamente identificadas, como para sobreidentificadas.

En este caso, suponga un sistema biecuacional con dos variables endógenas (Y_{i1} y Y_{i2}), y dos exógenas (X_{i1} y X_{i2}), como el presentado al principio de este capítulo (*véase* 2.15 y 2.16):

$$Y_{i1} = \beta_1 Y_{i2} + \beta_2 X_{i1} + e_{i1} \qquad (2.15)$$

$$Y_{i2} = \alpha_1 Y_{i1} + \alpha_2 X_{i2} + e_{i2} \qquad (2.16)$$

Si se desea por ejemplo, investigar el impacto de un cambio en Y_{i2} sobre Y_{i1}, el parámetro de interés a estimar es β_1. Por tanto, la metodología de MC2E consiste en:

1. Extraer el componente exógeno de Y_{i2}, a través de una regresión auxiliar, donde esta variable se explique en función de todas las variables exógenas del sistema X_{i1} y X_{i2}.
2. A partir de la regresión auxiliar de la primera etapa, se calculan los valores ajustados de la variable endógena Y_{i2}.
3. Con esta información, se estima la ecuación estructural de interés (2.16) por MCO, remplazando la variable endógena por los valores predichos en el paso 2 ($\hat{Y}_{i2}$). El estimador $\hat{\beta}_1$ es un estimador insesgado y consistente del parámetro estructural β_1.

Algunas de las características de MC2E son:

1. Puede aplicarse a una ecuación individual en el sistema sin tener en cuenta las otras ecuaciones.
2. Ante ecuaciones exactamente identificadas, arroja los mismos resultados que MCI.
3. A diferencia de MCI, MC2E puede aplicarse a ecuaciones sobreidentificadas.
4. Es fácil de aplicar, ya que solo se necesita saber en número total de variables exógenas o predeterminadas en el sistema sin conocer ninguna otra variable en el mismo.
5. Los errores estándar de MC2E se pueden determinar dado que los coeficientes estructurales son estimados directamente de MCO en la segunda etapa.
6. Si los R^2 en la forma reducida son altos (superiores a 0.80) las estimaciones de MCO y de MC2E serán cercanas.

En términos operativos, para estimar los parámetros de MC2E, también es posible obtener directamente un estimador (*véase* ecuación 2.17). El procedimiento completo de cómo se encuentra esta expresión, se presenta en el anexo 2.3.

$$\hat{\boldsymbol{\pi}}_{mc2e} = [Y_i'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_i]^{-1}(Y_i'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_1 \qquad (2.17)$$

En la ecuación 2.17, Y_1 corresponde a la variable endógena de interés que determina la ecuación a estimar, Y_i corresponde al conjunto de otras variables endógenas adicionales y Z una matriz que incluye todas las variables exógenas del modelo.

2.5.3 Mínimos cuadrados en tres etapas (MC3E)

La última técnica relevante, se conoce como mínimos cuadrados en tres etapas (MC3E), donde se estima el sistema de ecuaciones de forma conjunta en lugar de ecuación por ecuación (como lo hace MCI y MC2E). De manera general, los

métodos de este estilo se denominan de "información completa" debido a que utilizan todas las ecuaciones del sistema conjuntamente. En comparación a las otras técnicas de estimación, aquí la información adicional conduce a estimaciones más eficientes (o de menor varianza).

MC3E es una metodología que parte del método de MC2E, pero tiene en cuenta las correlaciones entre los términos de error de las ecuaciones. El procedimiento general se resume en:

1. Calcular los estimadores MC2E de las ecuaciones identificadas.
2. Utilizar los estimadores del paso uno, para estimar los errores de cada una de las ecuaciones estructurales. Con esta información, se construye la matriz de varianzas y covarianzas de los errores contemporáneos de las ecuaciones estructurales.
3. En la tercera etapa se realiza una estimación por MCG, donde se especifica la matriz encontrada en el paso dos. De esta forma se obtienen los estimadores de MC3E.

Al igual que con MC2E, es posible obtener directamente un estimador de MC3E (*véase* ecuación 2.18). El procedimiento completo de cómo se encuentra esta expresión, se presenta en el anexo 2.4.

$$\hat{\delta}_{MC3E} = (\mathbf{W}'\hat{\mathbf{V}}^{-1}\mathbf{W})^{-1}\mathbf{W}'\hat{\mathbf{V}}^{-1}w \qquad (2.18)$$

En la expresión 2.18, se define $w = P'X'y_i$ y $\mathbf{W} = P'X'Z_i$, con P una matriz de transformación con las variables exógenas o predeterminadas del sistema. Así mismo, V es una matriz que contiene las varianzas de los errores estimados. Ante la inexistencia de correlación serial de los errores, el estimador de MC3E es equivalente al de MC2E. En caso contrario, a través de esta metodología se consiguen estimadores con mayor eficiencia.

2.5.4 Sistema de regresiones aparentemente no relacionadas (SUR)

Las tres técnicas anteriores, están diseñadas para realizar estimaciones de modelos multiecuacionales con simultaneidad. Aun así, no todos los sistemas de ecuaciones presentan este problema: varias ecuaciones pueden estar conectadas no por compartir variables endógenas, sino por una correlación en sus términos de error. Como caso típico, considere un sistema de ecuaciones de oferta de un cultivo; un choque capturado por el error de una de las ecuaciones puede también estar afectando la oferta de otros cultivos, aunque explícitamente las ecuaciones estructurales de ambas no compartan ninguna variable.

Teniendo en cuenta lo anterior, se desarrolla el modelo SUR, que propone estimar un conjunto de ecuaciones aparentemente no relacionadas como una sola regresión; de esta manera se aprovecha la información similar entre las diferentes ecuaciones para mejorar la eficiencia de los estimadores (Zellner, 1962a). Como las ecuaciones de este método no comparten variables, se asume exogeneidad en los regresores.

Al igual que en MC3E, SUR saca provecho del enfoque de MCG para obtener una ganancia en eficiencia con respecto a MCO. Si las ecuaciones resultan no relacionadas, las estimaciones de SUR pasan a ser iguales a las de MCO. Para entender la metodología suponga una expresión reducida de un sistema de ecuaciones.

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E} \qquad (2.19)$$

En la ecuación 2.19, el vector $\mathbf{Y}$ contiene $\mathbf{M}$ variables dependientes, $\mathbf{X}$ es una matriz de K variables dependientes, $\boldsymbol{\beta}$ recoge los parámetros estructurales del sistema y E es el vector de errores. Asuma varianzas constantes pero diferentes entre las ecuaciones del sistema[19], errores con media cero relacionados entre si en un mismo periodo de tiempo[20].

---

[19] Es decir, $Var(E_i) = \sigma_i^2, i = 1, 2, ..., M$

[20] $Cov(E_{it}, E_{jt}) = \sigma_{ij}^2$ con $i, j = 1, 2, ..., M$. En este caso, se asume no existe covarianza entre los errores en distintos periodos.

Los supuestos anteriores, hacen que la aplicación de MCO conduzca a estimadores ($\hat{\boldsymbol{\pi}}_{MCO}$) ineficientes dada la existencia de correlación serial. La metodología SUR consiste en aplicar una transformación $\mathbf{T}$, que modifica el modelo inicial corrigiendo el problema (*véase* ecuación 2.20).

$$\mathbf{T'Y = T'X\beta + T'E} \qquad (2.20)$$

A partir de la ecuación 2.20, y aplicando el estimador tradicional de MCO; es posible obtener una expresión general para $\hat{\beta}_{SUR}$ (*véase* ecuación 2.21). En este caso, $\mathbf{V^{-1} = T'T}$ corresponde a la matriz de transformación

$$\hat{\boldsymbol{\beta}}_{SUR} = (\mathbf{X'V^{-1}X})^{-1}(\mathbf{X'V^{-1}Y}) \qquad (2.21)$$

2.5.5 Resumen de metodologías

A continuación, se presenta un cuadro que compara las diferentes técnicas de estimación presentadas, sus requerimientos y resultados. Esto permite reconocer cual es la metodología apropiada a aplicar a un estudio empírico particular (*véase* cuadro 2.1).

Cuadro 2.1 Metodologías de estimación de ecuaciones simultáneas.

METODOLOGIA	REQUERIMIENTOS	RESULTADOS	
		POSITIVOS	NEGATIVOS
MCI	Modelo exactamente identificado	Se encuentran estimadores para el modelo estructural	No se puedan hacer pruebas de hipótesis con estimadores estructurales
MC2E	Modelo exactamente identificado o sobreidentificado	Se encuentran estimadores insesgados y consistentes	Los estimadores no son eficientes.
MC3E	Modelo exactamente identificado o sobreidentificado	Mejora eficiencia de los estimadores	Si el modelo no tiene correlación de los errores entre las ecuaciones o está exactamente identificado los estimadores de MC3E

			son iguales a los de MC2E.
SUR	Correlación contemporánea de los errores de las ecuaciones	Mejora eficiencia de los estimadores	Si los errores de las ecuaciones no se correlacionan, se tienen las mismas estimaciones de MCO.

Fuente: los autores.

2.6 Estudio de caso: evaluación del fondo de estabilización de precios del azúcar

Una vez estudiados los diferentes métodos de estimación de modelos de ecuaciones simultáneas y de sistemas de regresión aparentemente no relacionados, se puede dar paso a su aplicación en la práctica, mediante estudios de caso.

El caso empírico que se desarrolla a continuación, está basado en el artículo titulado *"Análisis empírico del fondo de estabilización de precios del azúcar en Colombia"* de Vásquez (2005), que pretende probar si el Fondo de Estabilización de Precios del Azúcar –FEPA- ha sido eficaz como mecanismo para promover las exportaciones de azúcar colombiana hacia mercados extranjeros.

El FEPA fue creado como reacción a la crisis de precios del azúcar que ocurrió en 1999 en Colombia, y busca asegurar un ingreso remunerativo para los productores, aumentar las exportaciones y regular la producción nacional de éste a lo largo de los ciclos internacionales. [21]

Dado el gran número de subsidios al azúcar en el mercado internacional, el precio del azúcar al interior de Colombia suele ser mayor que aquel disponible en el extranjero. Para evitar que todos los ingenios productores de azúcar a nivel nacional saturen el mercado local, el fondo compensa las ventas a bajo precio en el

[21]Para más información, véase Prada (2004).

exterior, usando dinero que se obtiene a partir de un canon que se cobra a las ventas locales.

En este trabajo, Vásquez (2005) plantea un sistema de ecuaciones simultáneas que permite modelar el comportamiento del mercado del azúcar, teniendo en cuenta la interacción entre productores, consumidores locales e internacionales y el fondo de estabilización. El modelo viene conformado por tres ecuaciones estructurales; una que describe la oferta, una para la demanda y una para las exportaciones de azúcar producida en Colombia. Dentro de la ecuación de exportaciones, se incluye una variable que pretende capturar el efecto del fondo de estabilización. Si el fondo de estabilización ha sido efectivo para promover las exportaciones, el coeficiente que acompaña a esta variable debería ser positivo y estadísticamente significativo.

El autor plantea que teóricamente existen dos ecuaciones adicionales que deberían tenerse en cuenta al momento de especificar el modelo econométrico que describe el funcionamiento del mercado; una para las importaciones de azúcar que se compran para satisfacer la demanda local, y una para el cambio en los inventarios de azúcar de un periodo a otro. En el artículo se encuentra que para el caso colombiano estas cantidades son despreciables, por lo que es posible excluirlas del modelo sin que el análisis de mercado pierda validez.

Adicionalmente a las tres ecuaciones, el modelo adiciona una condición de equilibrio que relaciona las variables entre sí. El uso de un sistema de ecuaciones simultáneas permite modelar la determinación conjunta de los niveles de producción, consumo y exportaciones. Estas tres serian las variables endógenas. A continuación se presenta la especificación de cada una de las ecuaciones.

En la ecuación usada para representar la demanda, la variable dependiente corresponde a la cantidad de azúcar demandada en el mercado local. Adicionalmente se identifican seis determinantes que actúan como variables independientes: el índice de precios al consumidor de aquellos productos que usan azúcar como insumo (IPC_t) y que sirve de proxy al precio de mercado; la población de Colombia (Pob_t) y el ingreso per cápita ($\frac{Y_t}{Pob_t}$). Adicionalmente se

incluyen variables dicótomas trimestrales, para capturar el componente estacional propio del azúcar mercado (*véase* ecuación 2.22).

$$Q^{D}{}_{t} = \alpha_0 + \alpha_1 IPC_t + \alpha_2 \frac{Y_t}{Pob_t} + \alpha_3 Pob_t + \alpha_4 Trim2_t + \alpha_5 Trim3_t + \alpha_6 Trim4_t + e_2 \qquad (2.22)$$

En esta ecuación estructural, se espera que el precio tenga un efecto negativo sobre la cantidad demandada. El ingreso debería tener un efecto positivo, ya que se espera que el azúcar se comporte como un bien normal. Finalmente, como la ecuación viene dada en niveles, a mayor población se espera una mayor demanda de azúcar.

Para la oferta, la variable dependiente corresponde a la cantidad de azúcar producida para satisfacer tanto al mercado local como la demanda de azúcar colombiana en el mercado internacional. Entre los determinantes de la oferta se encuentra el precio de mercado, que es medido con el índice de precios al consumidor (IPC_t) de aquellos productos que usan azúcar como insumo y el índice de precios del productor de azúcar (IPP_t) que actúa como proxy a los costos de producción que enfrentan las firmas. Finalmente, al igual que en la ecuación estructural de la demanda, $Trim2_t$, $Trim3_t$ y $Trim4_t$ corresponden a dicotomas que pretenden capturar el componente estacional del mercado (*véase* ecuación 2.23).

$$Q^{O}{}_{tt} = \beta_0 + \beta_1 IPC_t + \beta_2 IPP_t + \beta_3 Trim2_t + \beta_4 Trim3_t + \beta_5 Trim4_t + e_1 \qquad (2.23)$$

En la especificación 2.23, se espera que el precio tenga un efecto positivo sobre la cantidad ofrecida. Los costos de producción que enfrentan las firmas deberían tener un efecto negativo, ya que modifican el optimo de producción de la firma dado por $img = cmg$.

Para terminar, en la ecuación usada para representar las exportaciones la variable del lado izquierdo corresponde a la cantidad de azúcar producida por ingenios nacionales que es vendida en el mercado internacional. Como determinantes se identifica en primer lugar la relación entre precios internos y externos, que es aproximada a través del índice de la tasa cambio real para productores no tradicionales (TC_t). Esta variable debería tener una relación positiva y significativa

sobre las exportaciones. Adicionalmente se incluye la variable *FEPA*, el parámetro poblacional de interés que corresponde a una variable dicotoma con el valor de uno en los años en los que el fondo de estabilización estuvo activo. Al igual que en las otras ecuaciones, para capturar el componente estacional del mercado se incluyen las variables trimestrales (*véase* ecuación 2.24).

$$Q^{E}_{tt} = \pi_0 + \pi_1 TC_t + \pi_2 FEPA_t + \pi_3 Trim2_t + \pi_4 Trim3_t + \pi_5 Trim4_t + e_3 \qquad (2.24)$$

La condición de equilibrio de mercado que interrelaciona a estas tres ecuaciones estructurales corresponde a la ecuación 2.25. Esta expresión representa una identidad, por lo que no tiene parámetros a estudiar.

$$Q^{D}_{t} = Q^{O}_{t} - Q^{E}_{t} \qquad (2.25)$$

Para iniciar el procedimiento de estimación, es necesario evaluar las condiciones de identificación para cada una de las cuatro ecuaciones estructurales. Esto no solo permite comprobar si el modelo es estimable, si además establece qué metodología es la más pertinente. La condición de orden se presenta en el cuadro 2.2, y la de rango en el 2.3.

Cuadro 2.2. Condición de orden para identificación.

ECUACIONES	J			M-1	ESTADO DE IDENTIFICACION
	ENDOGENAS	EXOGENAS	TOTAL		
2.22	1	5	6	3	Sobre identificación
2.23	1	4	5	3	Sobre identificación
2.24	1	5	6	3	Sobre identificación

Fuente: los autores.

Cuadro 2.3. Condición de Rango para identificación.

ECUACIONES	Rango (Ri)	Rango (RiΔ)	M-1	ESTADO DE ECUACIONES
2.22	6	3	3	Sobre identificada
2.23	6	3	3	Sobre identificada
2.24	6	3	3	Sobre identificada

Fuente: los autores.

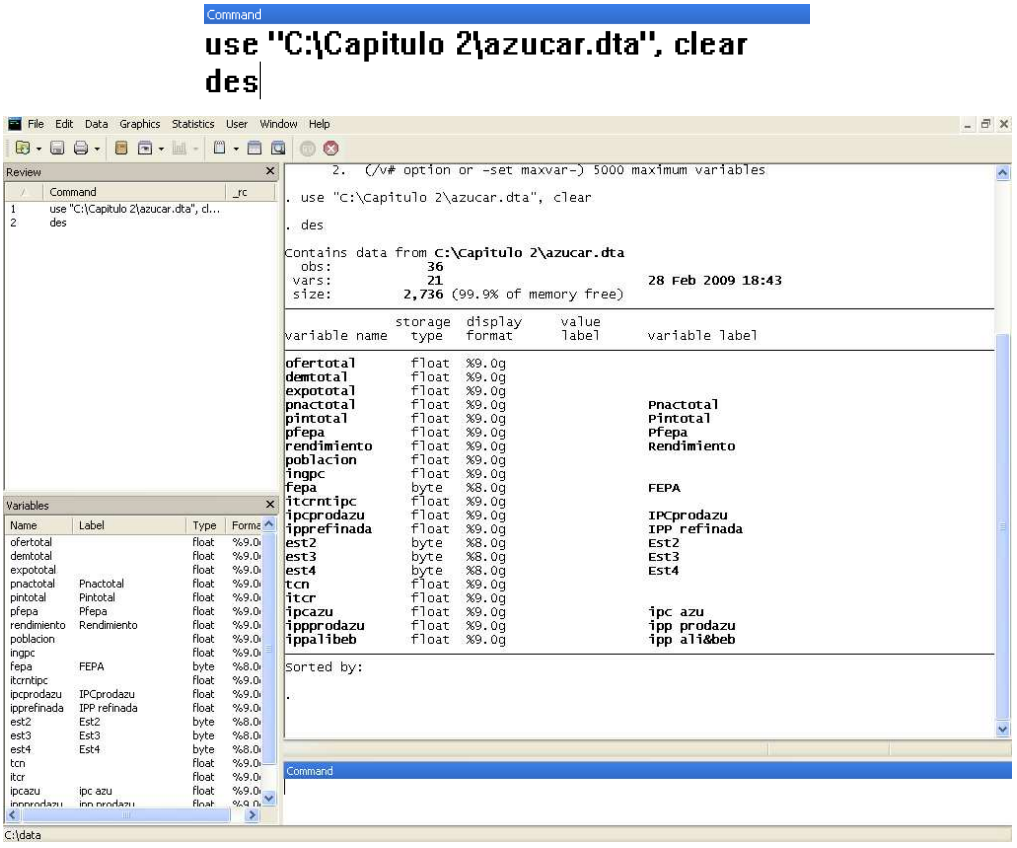
De acuerdo con lo anterior, se dice que las ecuaciones del sistema se encuentran sobreidentificadas, lo que implica que la estimación del modelo se puede realizar a través de MC2E o MC3E. A la identidad no se le aplican las condiciones de orden, pues no cuenta con ningún parámetro adicional. Para estimar estas regresiones conjuntamente, se realizará paso a paso el procedimiento en Stata®.

2.6.1 Análisis general de los datos

Esta primera sección, corresponde al arreglo del programa computacional para el análisis de ecuaciones simultáneas, así como una exploración general de los datos a usar.

1. Inicialmente, es necesario cargar la base de datos usando el comando *use*. Con las variables en memoria, usando el comando *des* es posible obtener la lista de variables disponibles (*véase* figura 2.1).

Figura 2.1. Salida comandos use y describe



Fuente: cálculos autores.

La tabla resultante, muestra como la base cuenta con 36 observaciones para el mercado azucarero colombiano. Estas son observaciones trimestrales, desde el primero de 1996 hasta el último del 2004. Para cada una, hay información sobre 21 indicadores específicos. La variable de interés corresponde a la dicótoma $FEPA_t$ (*véase* cuadro 2.4).

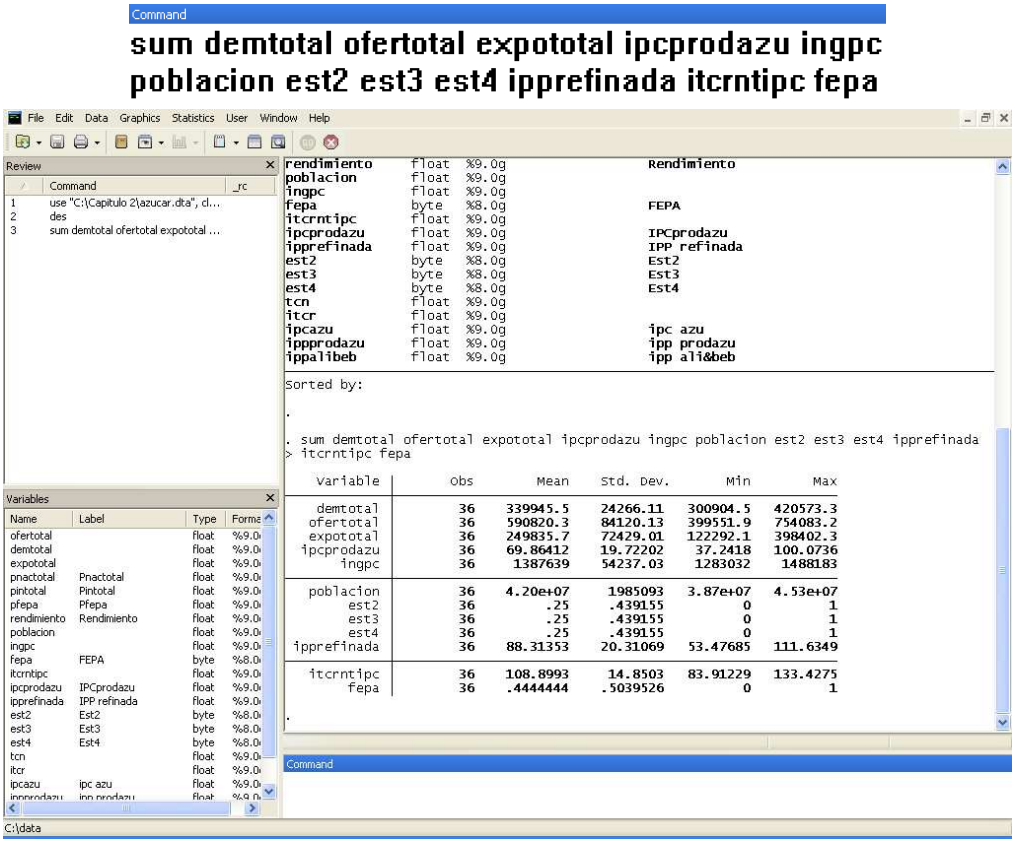
2. Con el comando *sum*, es posible obtener las principales estadísticas descriptivas. En este caso, se pide la tabla solo para las variables de interés, especificando cada una después del comando (*véase* cuadro 2.4 y figura 2.2).

Cuadro 2.4. Variables a usar en el modelo de ecuaciones simultáneas

Tipo de Variable	Variable del Modelo	Variables en la Base	Descripción
Endógenas	Q^D_t Q^O_t Q^E_t	demtotal ofertotal expototal	**1.** La demanda de un bien está en función de los precios y la renta de los consumidores. **2**. La oferta nacional está influenciada por el precio interno, mientras que la oferta de exportaciones lo estará del precio internacional. **3.** Las exportaciones, como bienes producidos en el interior y demandados en el exterior, dependerán de la capacidad de compra del resto del mundo y de los precios internos y externos
Exógenas	TC_t $FEPA_t$ IPC_t IPP_t $\frac{Y_t}{Pob_t}$ Pob_t $Trim2_t$ $Trim3_t$ $Trim4_t$	itcrntipc Fepa, Ipcprodazu ippreﬁnada ingpc población est2 est3 est4	Tasa de Cambio real, dummy que toma valor de 1 si está en periodo antes de 1999 y cero de lo contrario, índice de precios del azúcar, índice de precios al productor, ingreso per cápita, población, trimestre 2, trimestre 3 y trimestre 4.
Instrumentos	Cualquier variable exógena de todo el sistema de ecuaciones	itcrntipc , Fepa, Ipcprodazu, ippreﬁnada, ingpc, población, est2, est3, est4	

Fuente: los autores.

Figura 2.2. Salida comando summary

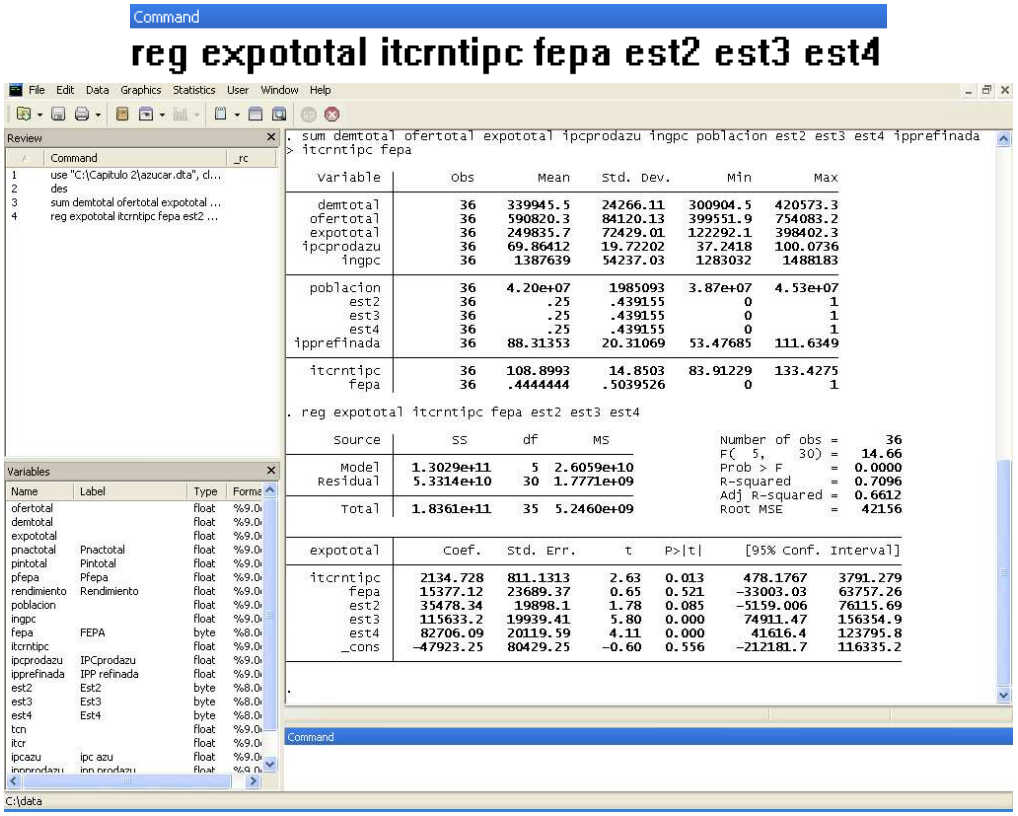


Fuente: cálculos autores.

2.6.2 Estimación del modelo por MCO

1. Se estimará la ecuación de exportaciones por MCO aplicando el comando *reg*, con el fin de comparar los resultados más adelante. Esto permitirá definir si realmente hay evidencia de un sesgo de simultaneidad (*véase* figura 2.3).

Figura 2.3. Salida regresión lineal



Fuente: cálculos autores.

En este caso, la variable de interés tiene el signo esperado pero no es estadísticamente significativa. Dado el planteamiento teórico, es de suponer que este modelo debe ser calculado con alguna técnica específica para modelos de ecuaciones simultáneas.

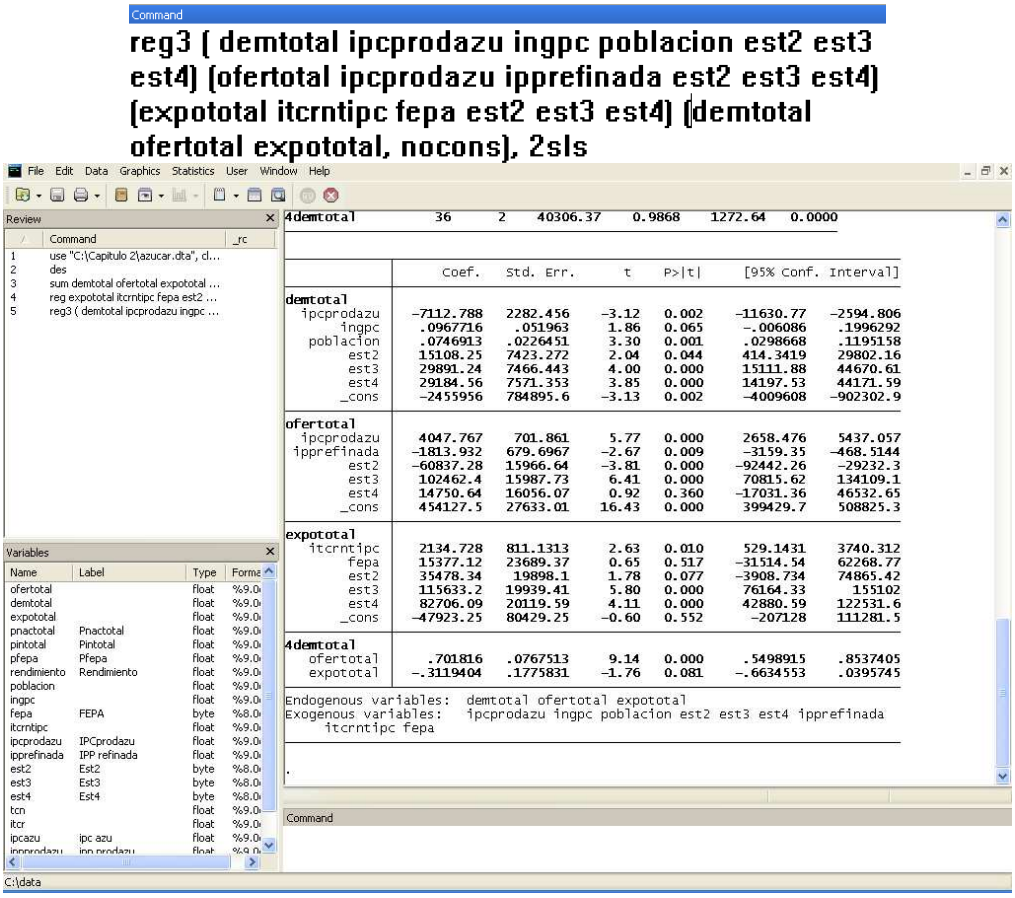
2.6.3 Estimación del modelo por MC2E y MC3E.

En Stata®, las estimaciones de sistemas de ecuaciones simultaneas se realizan a través del comando *reg3*. Este comando permite realizar estimaciones por Mínimos Cuadrados en Dos Etapas (MC2E) y Mínimos Cuadrados en Tres Etapas (MC3E).

1. Inicialmente, se estimará una regresión usando MC2E. En este caso especifico después del comando reg3 y la especificación de cada una de las ecuaciones, se debe añadir la opción *2sls.* El programa estadístico calculará

los parámetros estructurales de cada una de ecuaciones del modelo, usando las variables exógenas de todo el sistema como instrumentos (*véase* figura 2.4).

Figura 2.4. Salida regresión con MC2E



Fuente: cálculos autores.

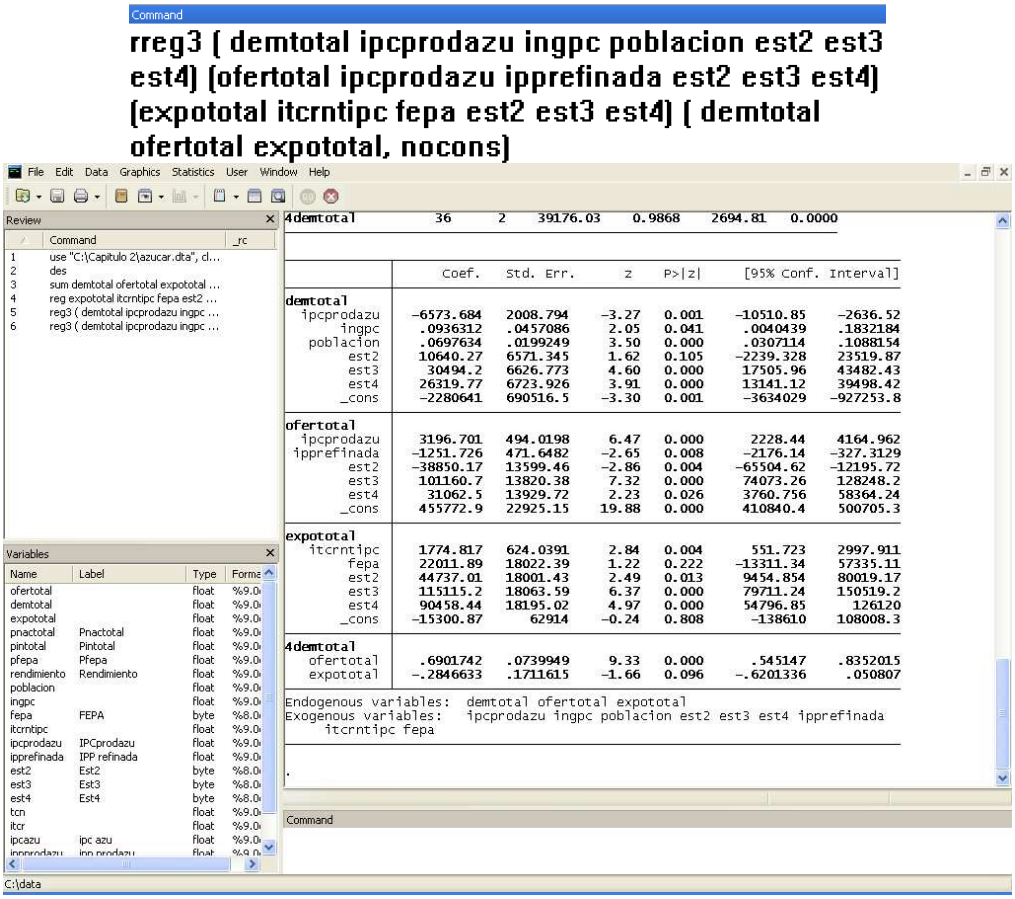
La estimación por MC2E, muestra los signos esperados para cada una de las regresiones. En el bloque de la demanda, el IPC de productos que contienen azúcar presenta una relación negativa y significativa -el estadístico t es de -3.12-, lo que indica el cumplimiento de la ley de la demanda. La población y el ingreso per cápita tienen los signos positivos esperados en el modelo teórico y son significativas, -con estadísticos t son 3.30 y 1.86, respectivamente-.

Asimismo, los resultados son los esperados en el bloque de la oferta. El IPC muestra una relación positiva y significativa, indicando el cumplimiento la ley de la oferta, mientras el IPP del azúcar tiene una relación negativa, reflejando el hecho de que si los costos de producción de los ingenios aumentan, estos reducirán su nivel de producción.

Por último, en el bloque de las exportaciones, el índice de tasa de cambio real de productos no tradicionales deflactado por el IPC resultó positivo y significativo - lo que significa que las exportaciones tienden a crecer a mayor devaluación-. Aun así, la variable de interés, $FEPA_t$ no resulta significativa, aunque si tiene el signo positivo esperado. -el estadístico t es de 0.65-.

2. Si no se especifica ninguna técnica de estimación, el comando *reg3* obtendrá los coeficientes usando el método de MC3E. Como se expuso anteriormente, esta estimación más eficiente que MC2E, lo que aumenta la posibilidad de rechazar la hipótesis nula. Esto es relevante en este caso, ya que hasta ahora ha sido imposible encontrar evidencia estadística de la significancia de $FEPA_t$ (*véase* figura 2.5).

Figura 2.5. Salida regresión con MC3E



Fuente: cálculos autores.

Los signos esperados de las demás variables incluidas en el modelo, se mantienen en esta regresión. Esto ocurre pues la metodología de MC3E obtiene los estimadores con un procedimiento casi idéntico al de MC2E, únicamente mejorando la eficiencia de los coeficientes. Con ninguna de las metodologías usadas, se encontró evidencia estadística de la significancia de la variable $FEPA_t$.

A partir de este análisis, es posible afirmar que el Fondo de Estabilización de Precios del Azúcar no ha logrado ser determinante al momento de fomentar las exportaciones. La dinámica del mercado del azúcar continúa siendo la misma que prevaleció antes de la implementación del fondo.

Para terminar, comparando los resultados obtenidos en las técnicas de ecuaciones simultaneas en relación a los del método de MCO, permite probar la existencia de sesgos de endogeneidad. Se observa que los coeficientes cambian notoriamente de tamaño. En particular, la variable de interés pasa de 15377 a 22011.

2.7 Estudio de caso: análisis regional de la oferta de ganado

El segundo caso empírico de este capítulo, está basado en el ejemplo 17.6 del libro (titulado en inglés) *"Learning and Practicing Econometrics"* de Hill, et Al. (1993). Este ejercicio pretende mostrar el funcionamiento de la metodología de estimación un sistema de regresiones aparentemente no relacionadas (SUR), con datos sobre producción ganadera.

El interés radica en construir un modelo que permita predecir los inventarios de reses en una región particular (C_t) Con este propósito, se construye un modelo lineal donde como variables independientes se toman el precio promedio de la carne (P_t), la cantidad de lluvia del año -como aproximación de la disponibilidad de alimento para los animales- (R_t), y la cantidad de ganado en la región un año atrás (C_{t-1}) (*véase* ecuación 2.26, el superíndice indica la región).

$$C_t^1 = \beta_0 + \beta_1 P_t^1 + \beta_2 R_t^1 + \beta_1 C_{t-1}^1 + e_1 \qquad (2.26)$$

La ecuación 2.26 corresponde a la ecuación de interés en la investigación. Aun cuando podrían obtenerse los estimadores mediante MCO, resulta conveniente aplicar la metodología SUR por términos de eficiencia. En este caso particular, es posible estimar conjuntamente esta expresión junto a expresiones equivalentes para otras regiones del país (*véase* ecuaciones 2.27 y 2.28).

$$C_t^2 = \beta_0 + \beta_1 P_t^2 + \beta_2 R_t^2 + \beta_1 C_{t-1}^2 + e_2 \qquad (2.27)$$

$$C_t^3 = \beta_0 + \beta_1 P_t^3 + \beta_2 R_t^3 + \beta_1 C_{t-1}^3 + e_3 \qquad (2.28)$$

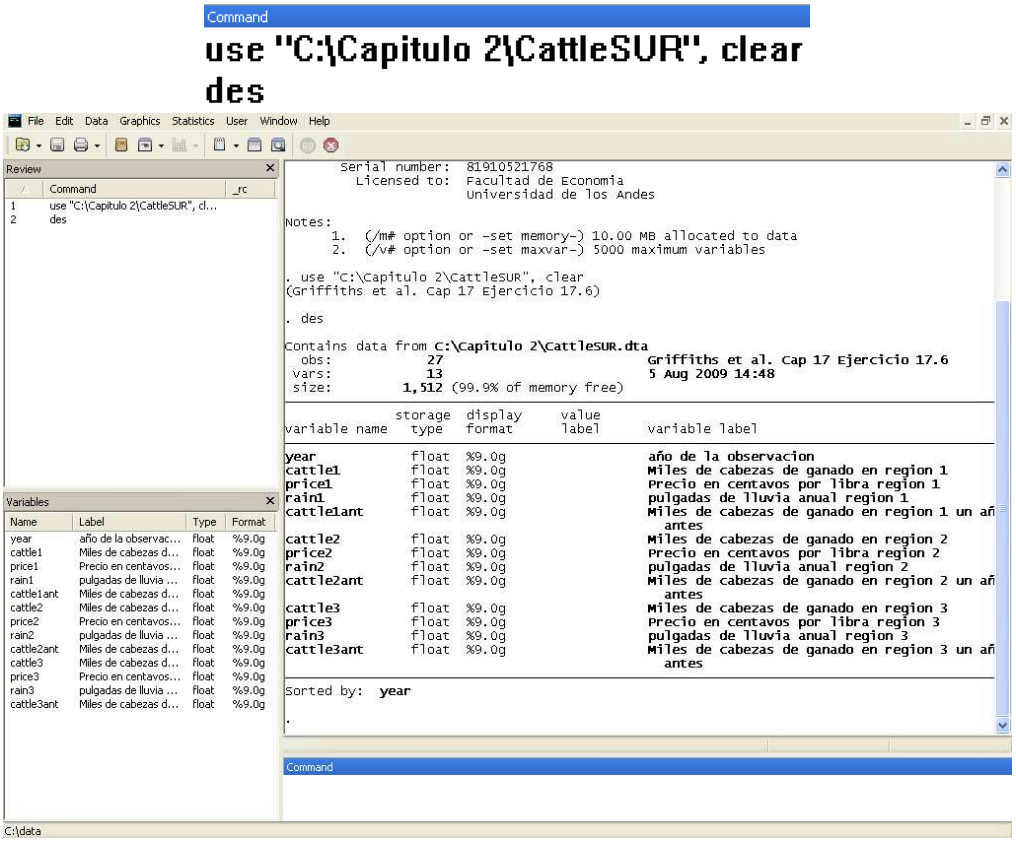
A continuación se presenta el procedimiento para estimar este sistema de ecuaciones, relacionadas únicamente a partir del termino de error, en el programa estadístico Stata®, tal y como se hizo en el caso empírico anterior. En primer lugar

se hace un análisis general de los datos, para luego pasar a analizar las estimaciones.

2.7.1 Análisis general de los datos

1. En primer lugar, usando el comando *use* se carga la base de datos. Con las variables en memoria, es posible obtener la lista de variables disponibles usando el comando *des* (*véase* figura 2.6) Esta base cuenta con una descripción detallada para cada variable.

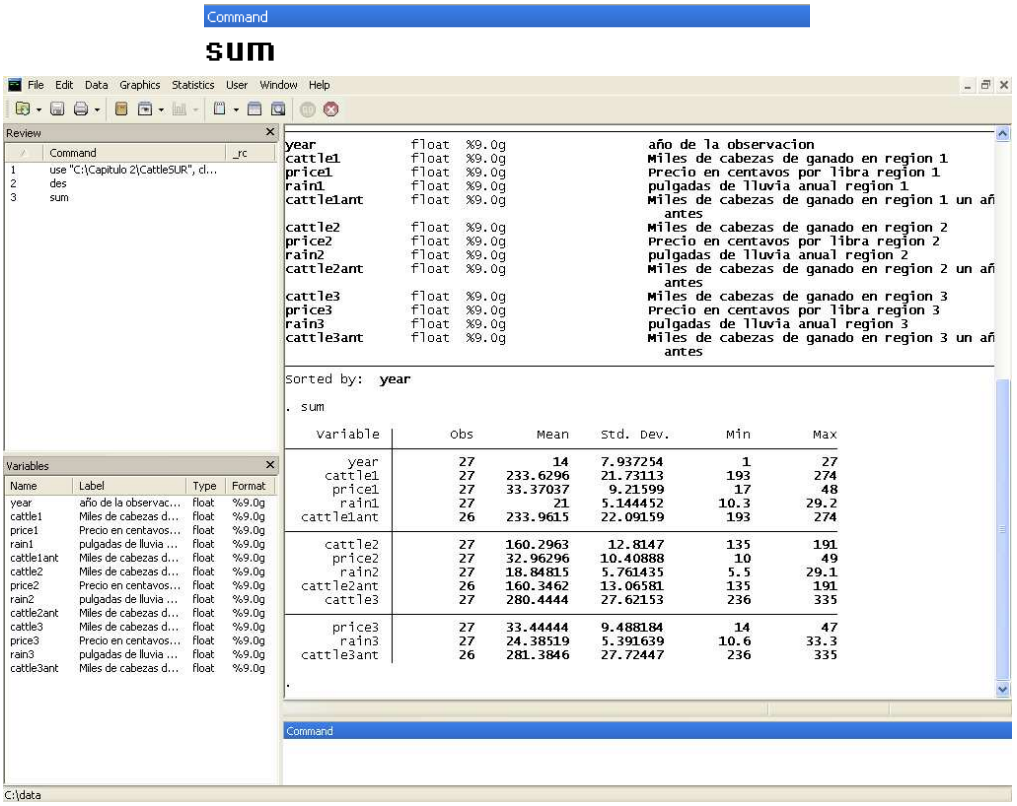
Figura 2.6. Comandos use y describe



Fuente: cálculos autores.

2. Con el comando *sum*, es posible obtener las principales estadísticas descriptivas. (*véase* figura 2.7).

Figura 2.7. Comando summary

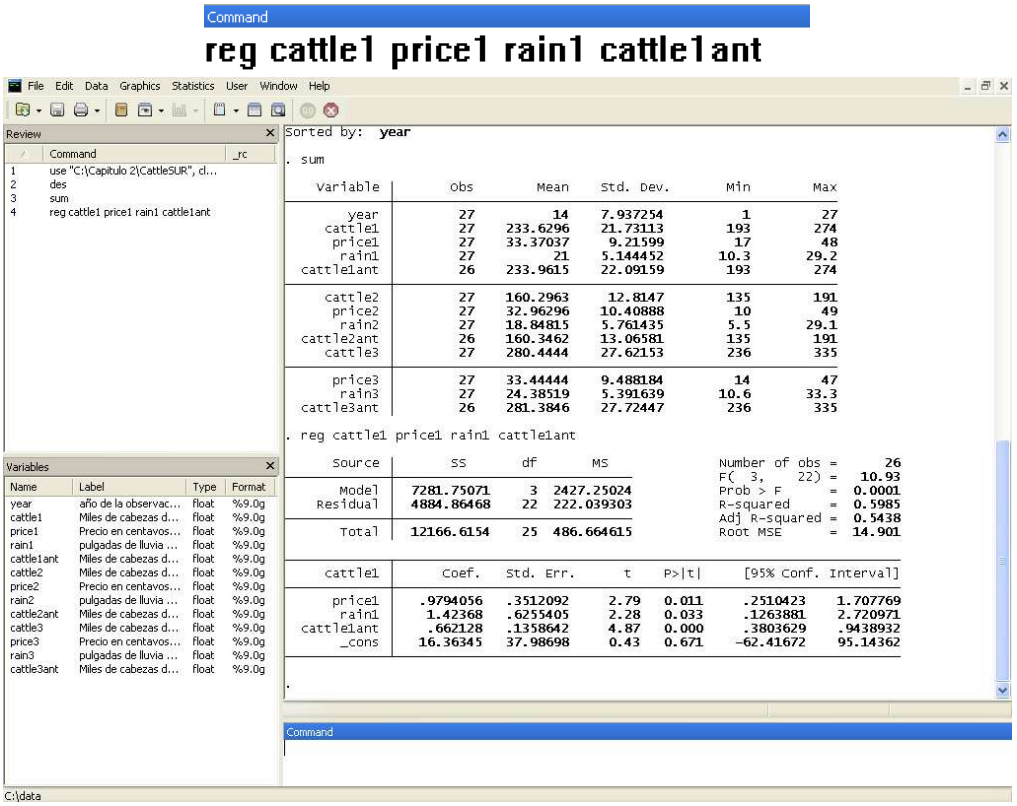


Fuente: cálculos autores.

2.7.2 Estimación del modelo por MCO

1. Antes de pasar a estimar el sistema de regresiones aparentemente no relacionadas, inicialmente se calcularan los coeficientes de la ecuación (2.26) por MCO, con el comando *reg*. Esto permitirá observar cómo cambian los resultados una vez adicionadas las otras ecuaciones estructurales (*véase* figura 2.8).

Figura 2.8. Regresión lineal



Fuente: cálculos autores.

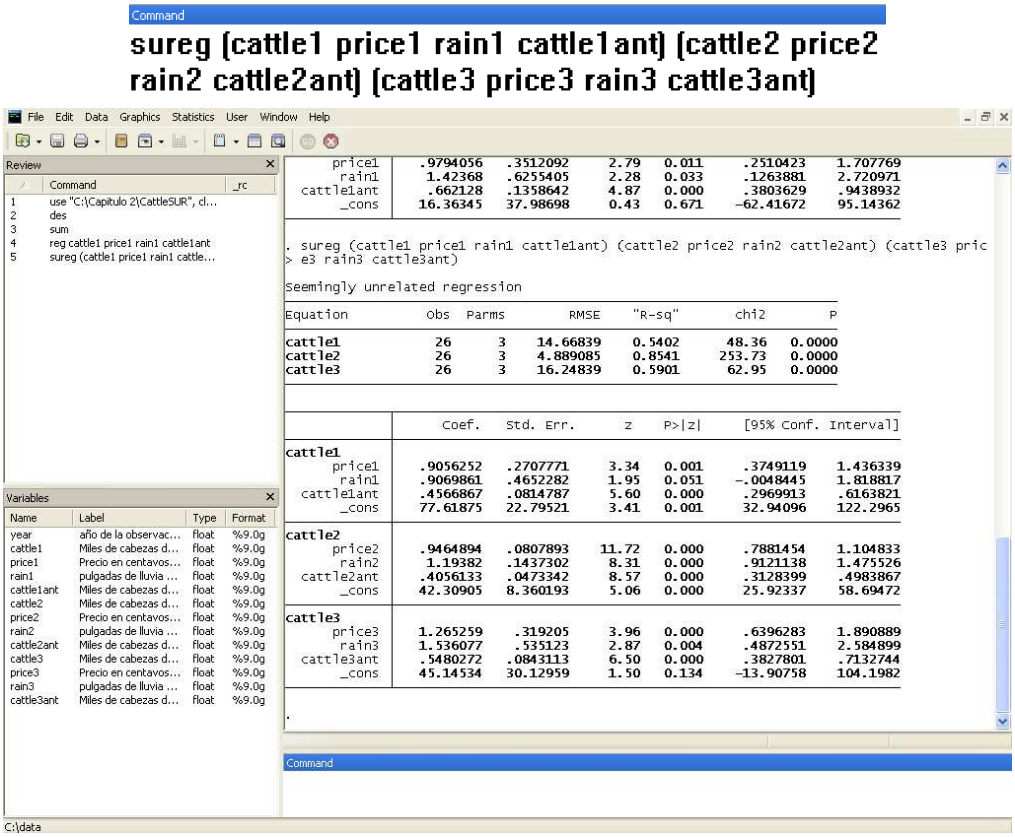
La figura 2.8, muestra como todas las variables independientes de la ecuación de interés aparecen significativas –con estadísticos t de 2.79, 2.28, y 4.87, respectivamente-. En este caso, la constante corresponde al único parámetro no estadísticamente significativo.

2.7.3 Estimación del modelo por SUR

Una vez se ha realizado la estimación por MCO, a continuación se presenta el procedimiento para obtener estimadores usando un modelo SUR. Recordando lo discutido anteriormente, este procedimiento saca provecho del enfoque de MCG para obtener una ganancia en eficiencia.

1. Para realizar una estimación SUR, se utiliza el comando *sureg*. En el programa estadístico es necesario especificar cada una de las ecuaciones del sistema en paréntesis, como se observa en la figura 2.9.

Figura 2.9. Regresión modelo SUR



Fuente: cálculos autores.

La estimación de cada una de las ecuaciones, muestra todas las variables con los signos esperados. Para la ecuación de interés, las independientes así como la constante, aparecen significativas –con estadísticos t de 3.34, 1.95, 5.60 y 3.41, respectivamente-.

El procedimiento presentado, permite realizar estimaciones de sistemas de regresiones aparentemente no relacionadas. Los signos en todas las estimaciones fueron los esperados, con mayor inventario de reses a mayores precios, precipitación y niveles de ganado inicial. Comparando los resultados obtenidos en este caso, se observa una mejora en la eficiencia de cada uno de los estimadores, en relación a los obtenidos por Mínimos Cuadrados Ordinarios. Como muestra, los p-valores estimados para los coeficientes que acompañan al precio y la precipitación en la región de interés, pasan de 0.011 y 0.033, a 0.001 y 0.051 respectivamente.

Resumen

- Para poder capturar la mayor complejidad presente en los fenómenos de simultaneidad, es necesario replantear el problema econométrico como un sistema conformado de múltiples ecuaciones. Esta representación se conoce como forma estructural de un modelo de ecuaciones simultáneas.
- Adicionalmente a la representación estructural, es posible resumir como una única ecuación donde se incluyen todas las variables exógenas del sistema. Esta se conoce como forma reducida.
- La metodología de mínimos cuadrados ordinarios ante ecuaciones simultaneas, no permite obtener estimadores consistentes para los parámetros estructurales de un problema de simultaneidad. Esto hace necesario el uso de nuevas metodologías.
- Como el problema de simultaneidad es un caso particular del de endogeneidad, la prueba de Hausman es la principal herramienta de identificación, comparando los estimadores de mínimos cuadrados ordinarios –que estarían sesgados ante la presencia de este problema- con estimadores obtenidos de alguna otra metodología que garantice parámetros insesgados y consistentes.
- Los paramentos estructurales pueden estimarse siempre y cuando se tengan variables exógenas adicionales en el sistema, que puedan usarse como instrumentos. De acuerdo a la relación entre variables exógenas y endógenas dentro de un modelo, una ecuación estructural puede estar sobreidentificada, identificada, o no identificada. Esto se evalúa a través de las condiciones de orden y rango.
- Para estimar los parámetros de una ecuación estructural, se pueden usar las metodologías de MCI, MC2E o MC3E. La primera, otorga estimadores insesgados pero imposibilita la aplicación de pruebas de hipótesis. Las otras dos, conducen a estimadores consistentes y eficientes de los parámetros estructurales.
- Para sistemas de ecuaciones no endógenas, se aplica la metodología SUR la cual permite obtener estimadores más eficientes que aquellos de MCO, aplicando una versión general de MCG.

Anexo 2

Anexo 2.1 Otros Ejemplos de Ecuaciones Simultáneas

1. Modelo de demanda y oferta de un bien[22]

Función de demanda: $Q_t^d = \alpha_0 + \alpha_1 P_t + \mu_{1t}$ (A.2.1)

Función de oferta: $Q_t^o = \beta_0 + \beta_1 P_t + \mu_{2t}$ (A.2.2)

Equilibro: $Q_t^d = Q_t^o$ (A.2.3)

2. Modelo Keynesiano de determinación del ingreso[23]

Función de consumo: $C_t = \beta_0 + \beta_1 Y_t + u_t$ (A.2.4)

Identidad del ingreso: $Y_t = C_t + I_t (= S_t)$ (A.2.5)

3. Modelo de inflación y apertura comercial[24]

$$Inf = \beta_0 + \beta_1 open + \beta_2 \log(pcnic) + \mu_1 \quad \text{(A.2.6)}$$

$$Open = \alpha_0 + \alpha_1 Infl + \alpha_2 \log(pcinc) + \alpha_3 \log(land) + \mu_2 \quad \text{(A.2.7)}$$

[22] *Véase* (Gujarati, 2003, 692).
[23] *Véase* (Judge et al, 1988, 621).
[24] *Véase* (Wooldridge, 2009, 556).

Anexo 2.2 Notación General

Un sistema conformado por M variables endógenas y K exógenas puede escribirse como sigue (*véase* ecuación A.2.6).

$$
\begin{aligned}
&(1)\gamma_{11}Y_1+\gamma_{21}Y_2+\gamma_{31}Y_3+\ldots+\gamma_{M1}Y_M+\beta_{11}X_1+\beta_{21}X_2+\ldots+\beta_{K1}X_K+e_1=0\\
&(2)\gamma_{12}Y_1+\gamma_{22}Y_2+\gamma_{32}Y_3+\ldots+\gamma_{M2}Y_M+\beta_{12}X_1+\beta_{22}X_2+\ldots+\beta_{K2}X_K+e_2=0\\
&.\\
&.\\
&.\\
&(M)\gamma_{1M}Y_1+\gamma_{2M}Y_2+\gamma_{3M}Y_3+\ldots+\gamma_{MM}Y_M+\beta_{1M}X_1+\beta_{2M}X_2+\ldots+\beta_{KM}X_K+e_M=0
\end{aligned}
\qquad \text{(A.2.8)}
$$

Para simplificar el sistema de ecuaciones se utiliza la representación matricial, con el fin de simplificar las operaciones y ofrecer al lector un mecanismo más claro para entender el fondo del problema de las ecuaciones simultáneas (*véase* ecuación A.2.9).

$$\mathbf{Y\Gamma}+\mathbf{XB}+\mathbf{E}=0 \qquad \text{(A.2.9)}$$

Donde

$$\mathbf{Y}=\begin{pmatrix} Y_{11} & \cdots & Y_{1M} \\ \vdots & \ddots & \vdots \\ Y_{T1} & \cdots & Y_{TM} \end{pmatrix}_{T\times M} \qquad \text{(A.2.10)}$$

En la matriz A.2.10 contiene la información de las M variables endógenas del sistema de ecuaciones. Sus columnas están compuestas por los vectores $\left[Y_1Y_2Y_3\cdots Y_M\right]$.

$$\mathbf{X}=\begin{pmatrix} X_{11} & \cdots & X_{1K} \\ \vdots & \ddots & \vdots \\ X_{T1} & \cdots & X_{TK} \end{pmatrix}_{T\times K} \qquad \text{(A.2.11)}$$

En la matriz A.2.11 contiene la información de las K variables exógenas del sistema de ecuaciones. Sus columnas están compuestas por los vectores $\left[X_1X_2X_3\cdots X_M\right]$.

$$\mathbf{\Gamma} = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1M} \\ \vdots & \ddots & \vdots \\ \gamma_{M1} & \cdots & \gamma_{MM} \end{pmatrix}_{M \times M} \qquad \text{(A.2.12)}$$

En la matriz A.2.12 contiene la información de los parámetros estructurales del sistema de ecuaciones. Sus columnas están compuestas por los vectores $\left[\Gamma_1 \Gamma_2 \Gamma_3 \cdots \Gamma_M\right]$.

$$\mathbf{B} = \begin{pmatrix} \beta_{11} & \cdots & \beta_{1M} \\ \vdots & \ddots & \vdots \\ \beta_{K1} & \cdots & \beta_{KM} \end{pmatrix}_{K \times M} \qquad \text{(A.2.13)}$$

En la matriz A.2.13 contiene la información de los parámetros reducidos del sistema de ecuaciones. Sus columnas están compuestas por los vectores $\left[B_1 B_2 B_3 \cdots B_M\right]$.

$$\mathbf{E} = \begin{pmatrix} e_{11} & \cdots & e_{1M} \\ \vdots & \ddots & \vdots \\ e_{T1} & \cdots & e_{TM} \end{pmatrix}_{T \times M} \qquad \text{(A.2.14)}$$

$$\mathbf{0} = \begin{pmatrix} 0_{11} & \cdots & 0_{1M} \\ \vdots & \ddots & \vdots \\ 0_{T1} & \cdots & 0_{TM} \end{pmatrix}_{T \times M} \qquad \text{(A.2.15)}$$

Finalmente, E Es la matriz que contiene los errores de las ecuaciones y 0 es una matriz de ceros.

Dada la simplificación de las matrices, el sistema quedaría resumidos de la siguiente forma (*véase* ecuación A.2.16, los tamaños de las matrices resultantes se muestran debajo).

$$\underbrace{\underbrace{\mathbf{Y}_{T \times M} \mathbf{\Gamma}_{M \times M}}_{T \times M} + \underbrace{\mathbf{X}_{T \times K} \mathbf{B}_{T \times M}}_{T \times M} + \mathbf{E}_{T \times M}}_{T \times M} = \underset{T \times M}{\mathbf{0}_{T \times M}} \qquad \text{(A.2.16)}$$

Teniendo conocimiento de esta representación, es posible transformar el modelo para que tenga la forma tradicional $Y = X\beta + \varepsilon$. Partiendo de la expresión A.2.17 y multiplicando por Γ^{-1} a ambos lados, se tiene:

$$\mathbf{Y\Gamma + XB + E} = 0 \qquad \text{(A.2.17)}$$

$$\begin{aligned}
&\mathbf{Y}\underbrace{\mathbf{\Gamma\Gamma}^{-1}}_{I} + \mathbf{XB\Gamma}^{-1} + \mathbf{E\Gamma}^{-1} = \underbrace{\mathbf{0\Gamma}^{-1}}_{0} \\
&\mathbf{Y} + \mathbf{XB\Gamma}^{-1} + \mathbf{E\Gamma}^{-1} = \mathbf{0} \\
&\mathbf{Y} = \underbrace{-\mathbf{XB\Gamma}^{-1}}_{X\pi}\ \underbrace{-\mathbf{E\Gamma}^{-1}}_{V} \\
&\rightarrow \mathbf{Y} = \mathbf{X\pi} + \mathbf{V}
\end{aligned} \qquad \text{(A.2.18)}$$

Esta última expresión se conoce como la forma reducida del sistema de ecuaciones simultaneas. Estimando el sistema por medio de MCO se llega a:

$$\hat{\pi}_{MCO} = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{Y}) \qquad \text{(A.2.19)}$$

Anexo 2.3 Estimación por MC2E

En forma matricial, es posible obtener el estimador de MC2E fácilmente. Suponga un modelo completo con dos ecuaciones estructurales, y un conjunto de variables exógenas ($\bar{X}_1$ y $\bar{X}_2$) para cada una (*véase* ecuaciones A.2.20 y A.2.21).

$$Y_1 = \alpha Y_2 + \bar{X}_1\beta + e_1 \qquad \text{(A.2.20)}$$

$$Y_2 = \delta Y_1 + \bar{X}_2\varphi + e \qquad \text{(A.2.21)}$$

A partir de lo anterior, suponiendo que se desea encontrar los estimadores de la primera ecuación del modelo, debe realizarse la primera etapa del proceso para extraer el exógeno de Y_2 (*véase* ecuación A.2.22).

$$Y_2 = Z\pi + \upsilon \qquad \text{(A.2.22)}$$

En la ecuación A.2.22, Z corresponde a una matriz que incluye todas las variables exógenas del modelo, usadas para instrumentar la variable endógena. El

estimador de mínimos cuadrados ordinarios de π, se presenta en la ecuación A.2.23.

$$\hat{\pi}_{mco} = (\mathbf{Z}'\mathbf{Z})^{-1}(\mathbf{Z}'Y_2) \qquad \text{(A.2.23)}$$

De lo anterior se obtienen los $\hat{Y}_2$, a usarse en la segunda etapa. Calculando el estimador de MCO de este segundo paso, y reemplazando la expresión A.2.23, se obtiene el estimador general de MC2E para un sistema biecuacional.

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{mc2e} &= (\hat{\mathbf{Y}}_2'\hat{\mathbf{Y}}_2)^{-1}(\hat{\mathbf{Y}}_2'\mathbf{Y}_1) \\ \hat{\boldsymbol{\beta}}_{mc2e} &= (\hat{\boldsymbol{\pi}}'\mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\pi}})^{-1}(\hat{\boldsymbol{\pi}}'\mathbf{Z}'\mathbf{Y}_1)\end{aligned} \qquad \text{(A.2.24)}$$

$$\begin{aligned}\hat{\boldsymbol{\beta}}_{mc2e} &= [Y_2'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2]^{-1}(Y_2'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_1 \\ \hat{\boldsymbol{\beta}}_{mc2e} &= [Y_2'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2]^{-1}(Y_2'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_1\end{aligned} \qquad \text{(A.2.25)}$$

Para mirar si los estimadores son insesgados, hay que obtener el valor esperado de la expresión $\hat{\pi}_{mc2e}$ y determinar que el estimador sea el parámetro poblacional. La prueba es presentada a continuacion:

$$\begin{aligned}E[\hat{\beta}_{mc2e}] &= E[[Y_2'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2]^{-1}(Y_2'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_1] \\ E[\hat{\beta}_{mc2e}] &= E[[Y_2'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2]^{-1}(Y_2'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'(Y_2\boldsymbol{\delta}+e)] \\ E[\hat{\beta}_{mc2e}] &= E[[Y_2'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2]^{-1}(Y_2'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2\boldsymbol{\delta}+[Y_2'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2]^{-1}(X'Z)(Z'Z)^{-1}Z'e)] \\ E[\hat{\beta}_{mc2e}] &= E[\boldsymbol{\delta}+[Y_2'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2]^{-1}(Y_2'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'e] \\ E[\hat{\beta}_{mc2e}] &= \boldsymbol{\delta}+[Y_2'\mathbf{Z}(\mathbf{Z}'\mathbf{Z})^{-1}\mathbf{Z}'Y_2]^{-1}(Y_2'\mathbf{Z})(\mathbf{Z}'\mathbf{Z})^{-1}E[\mathbf{Z}'e]\end{aligned}$$

Suponiendo que los instrumentos usados son validos, se tiene que $E[Z'e]=0$; lo que implica que el estimador se aproxima de manera satisfactoria a los parámetros poblacionales (*véase* ecuación A.2.27).

$$E[\hat{\boldsymbol{\beta}}_{mc2e}] = \boldsymbol{\delta} \qquad \text{(A.2.27)}$$

Lo anterior, muestra de manera general el procedimiento necesario para encontrar directamente estimadores de MC2E en el contexto de ecuaciones simultaneas. Este

resultado puede generalizarse fácilmente a *m* ecuaciones. La prueba de consistencia, para cada una es análoga a la anterior.

Anexo 2.4 Estimación por MC3E

Suponga un modelo estructural (*véase* ecuación A.2.28).

$$\begin{matrix} y_i = Y_i\beta_i + X_i\gamma_i + u_i \\ y_i = Z_i\delta_i + u_i \end{matrix}, i = 1, 2, ..., M \qquad \text{(A.2.28)}$$

$$\text{Con } Z_i = [Y_i X_i]$$

Se transforma el modelo multiplicando a ambos lados de la ecuación por X, una matriz P (TxK) de las variables predeterminadas del sistema.

$$\mathbf{X'y}_i = \mathbf{X'Z}_i\boldsymbol{\delta}_i + \mathbf{X'}u_i \qquad \text{(A.2.29)}$$

Se sabe que $\mathbf{PX'XP' = I_T}$ por tanto se pude multiplicar el modelo trasformado por P

$$\mathbf{P'X'}\,y_i = \mathbf{P'X'Z_i}\boldsymbol{\delta}_i + \mathbf{P'X'}u_i \qquad \text{(A.2.30)}$$

Si $\mathbf{w_i = P'X'y_i}$, $\mathbf{W_i = P'X'Z_i}$ y $\mathbf{v_i = P'X'}u_i$ el modelo quedaría representado por la siguiente expresión:

$$\mathbf{w}_i = \mathbf{W}_i\boldsymbol{\delta}_i + \mathbf{v}_i \qquad \text{(A.2.31)}$$

Al aplicar MCO a la ecuación A.2.31, se tiene una equivalencia con el estimador de MC2E:

$$\hat{\boldsymbol{\delta}}_i = (\mathbf{W}_i'\mathbf{W}_i)^{-1}(\mathbf{W}_i'w_i) = (\mathbf{Z}_i'\mathbf{XPP'X'Z}_i)^{-1}(\mathbf{Z}_i'\mathbf{XPP'X'}\,y_i) \qquad \text{(A.2.32)}$$

Ahora bien, si se reúnen las M ecuaciones que componen el sistema tendríamos:

$$w = \mathbf{W}\boldsymbol{\delta} + v \qquad \text{(A.2.33)}$$

Donde

$$E[v'v] = \begin{pmatrix} \sigma_{11}I & \cdots & \sigma_{1M}I \\ \vdots & \ddots & \vdots \\ \sigma_{M1}I & \cdots & \sigma_{MM}I \end{pmatrix} = \mathbf{V} \qquad \text{(A.2.34)}$$

Un estimador de la Matriz V puede ser construido con los residuos de los estimadores de MC2E

$$\hat{\sigma}_{ij} = \frac{\hat{u}_i'\hat{u}_j}{T} = \frac{\hat{u}_j'\hat{u}_i}{T} = \hat{\sigma}_{ji} \qquad \text{(A.2.35)}$$

Por lo tanto

$$\hat{\mathbf{V}} = \begin{pmatrix} \hat{\sigma}_{11}I & \cdots & \hat{\sigma}_{1M}I \\ \vdots & \ddots & \vdots \\ \hat{\sigma}_{M1}I & \cdots & \hat{\sigma}_{MM}I \end{pmatrix} \qquad \text{(A.2.36)}$$

De lo anterior se deduce la expresión del estimador de MC3E (*véase* ecuación A.2.37).

$$\hat{\boldsymbol{\delta}}_{MC3E} = (\mathbf{W}'\hat{\mathbf{V}}^{-1}\mathbf{W})^{-1}\mathbf{W}'\hat{\mathbf{V}}^{-1}w \qquad \text{(A.2.37)}$$

Capítulo 3

Modelos de probabilidad: lineal, probit y logit.

3.1 Introducción

Una vez abarcado el tratamiento de los supuestos de MCO en los capítulos anteriores, a continuación se discuten los modelos de regresión probabilísticos en un contexto de corte transversal, caracterizados por contar con variable dependiente discreta o binaria con valores cero y uno[25]. Éstos, difieren del modelo clásico en la estimación e interpretación de resultados. En particular, se presentan las metodologías, pertinencia, ventajas y desventajas para los modelos de probabilidad lineal (MPL), logit y probit; que en economía se han utilizado para explicar el comportamiento individual de los agentes, consumo de bienes durables y análisis de desequilibrios. Los modelos con variables dependientes dicótomas, tienen aplicaciones tanto en los datos de corte transversal como en los datos de series de tiempo

A partir de lo anterior, en primero, se estudia el modelo de probabilidad lineal (MPL) bajo el modelo clásico de regresión lineal, utilizando MCO. En segundo y tercer lugar se introducen los modelos logit y probit, que se estiman mediante máxima verosimilitud (MV). El objetivo principal de estos modelos es encontrar la probabilidad de que un acontecimiento suceda condicionado a un conjunto de regresoras[26].

Finalmente, la aplicación de estos modelos se lleva a cabo en un estudio de caso basado en la información del artículo de Bernal (2008), titulado (en inglés) "*The*

[25] Las variables dicótomas se caracterizan por registrar únicamente dos opciones mutuamente excluyentes entre sí. Por ejemplo, si se tiene una variable que muestre la respuesta a ¿está empleado?

[26] La estimación de modelos de corte transversal clásicos tiene como objetivo la derivación de valor esperado de la variable explicativa, dadas las variables explicativas ($E(Y_i \mid X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ik})$). En los modelos de este capítulo, se busca $P(Y_i = 1 \mid X_{i1}, X_{i2}, X_{i3}, \ldots, X_{ik})$

Informal Labor Market in Colombia: Identification and characterization" el cual está enfocado en derivar los determinantes del trabajo informal en Colombia.

3.2 Modelo de probabilidad lineal

3.2.1 Estimación modelo de probabilidad lineal

La metodología de probabilidad lineal (MPL), basado en los supuestos del modelo clásico de regresión lineal, es una primera aproximación para estimar modelos caracterizados por una variable dependiente dicótoma (*véase* ecuación 3.1). MPL permite entender las particularidades de modelos probabilísticos.

$$Y_i = \mathbf{X}\boldsymbol{\beta} + u_i,\ i = 1, 2, ..., n \qquad (3.1)$$

En la ecuación 3.1, Y_i es un vector con valores cero y uno que describe la variable dependiente; $\mathbf{X}$ una matriz de variables explicativas del modelo; $\boldsymbol{\beta}$ un conjunto de coeficientes y u_i el vector de errores. A partir de la ecuación 3.1, el principal objetivo de MPL es estimar a través de MCO, el valor esperado de la variable dependiente dados los valores de $\mathbf{X}$. Dado que Y_i es una variable dicotoma, este resultado se debe interpretar como la probabilidad condicional que Y_i tome el valor de 1 supeditado a $\mathbf{X}$ (*véase* ecuación 3.2 y anexo 3.1)[27].

$$E(Y_i \mid \mathbf{X}) = \mathbf{X}\boldsymbol{\beta} = P(Y_i = 1 \mid \mathbf{X}) \qquad (3.2)$$

Teniendo en cuenta lo anterior, la ecuación 3.2 muestra la relación existente entre cada una de las X_i y la probabilidad de que ocurra el evento relacionado con la representación (Y_i=1). Dado que la probabilidad se encuentra entre cero y uno, las predicciones de Y_i deberían estar dentro de dicho intervalo (*véase* ecuación 3.3).

$$0 \leq E(Y_i \mid \mathbf{X}) \leq 1 \qquad (3.3)$$

Si la condición 3.3 se cumple, el vector de estimadores ($\hat{\boldsymbol{\beta}}_{MCO}$) captura el cambio marginal en la probabilidad de éxito, teniendo en cuenta los movimientos en las

[27] *Véase* (Gujarati, 2003, 563)

variables X_i. Desafortunadamente, el método de probabilidad lineal no necesariamente cumple con lo anterior, lo que llevas a problemas en inferencias y conclusiones. A continuación se explica este y otros problemas recurrentes al utilizar modelos de probabilidad lineal.

3.2.1 Problemas en el modelo de probabilidad lineal

En primer lugar, de acuerdo a la estimación de MPL, el supuesto sobre la distribución normal de los errores no se cumple. Aunque para las estimaciones por MCO no requiere que los errores se distribuyan de forma normal, este supuesto es necesario para poder hacer inferencias estadísticas. En el caso cuando se tiene un modelo con variable dependiente binaria, se sugiere otro tipo de distribución en los errores y MPL hace caso omiso a esto último. Por tanto, si se tiene que Y_i toma valores de cero y uno, los errores del modelo 3.1 se pueden ver representados en la ecuación 3.4:

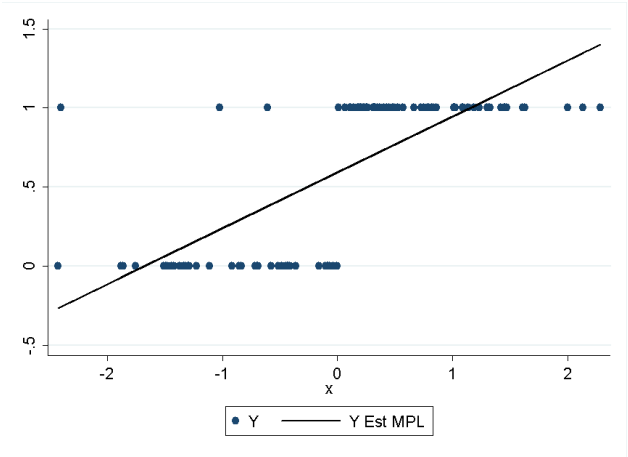
$$u_i = Y_i - \mathbf{X\beta} \qquad (3.4)$$

$$\text{Si } Y_i = 1\text{: } u_i = 1 - \mathbf{X\beta} \qquad (3.5)$$

$$\text{Si } Y_i = 0\text{: } u_i = -\mathbf{X\beta} \qquad (3.6)$$

Las ecuaciones 3.5 y 3.6, muestran que u_i toma dos valores ($1-\mathbf{X\beta}$ y $-\mathbf{X\beta}$), es decir, los errores siguen una distribución binomial. Este resultado deja sin sustento teórico el supuesto de la distribución normal de los errores que asume MPL. Como consecuencia de ello, es inadecuado formular un modelo lineal para estimar aquellos que tienen variable dependiente limitada (*véase* gráfica 3.1).

Gráfica 3.1. Estimación MPL



Fuente: los autores

Es importante destacar que esta dificultad no es crítica, puesto que los estimadores siguen siendo insesgados y en la medida en que se aumente la muestra, es posible encontrar una distribución asintótica normal (prueba basada en el límite central) (Gujarati, 2003, 564).

En segundo lugar, las varianzas de los errores incumplen el supuesto de homoscedasticidad (*véase* anexo A.3.2), las estimaciones no se sitúan dentro del intervalo [0,1] y asume erróneamente una relación lineal entre $\mathbf{X}$ y $P_i(Y_i = 1|\mathbf{X})$ (véase gráfica 3.1).

La crítica más fuerte del modelo consiste en que el efecto marginal es constante para cualquier valor de $\mathbf{X}$, condición difícil de mantener si se asume que teóricamente el efecto marginal es variable.

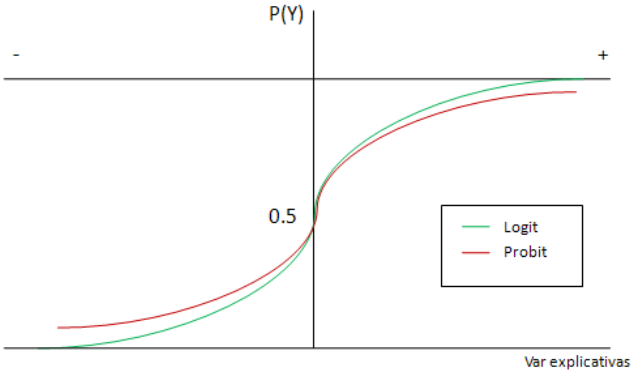
Por lo anterior, es adecuado optar metodologías que mitiguen las falencias presentadas. A continuación, se introducen los modelos logit y probit basados en la estimación por Máxima Verosimilitud (MV), una técnica alternativa que permite distribuciones distintas a la normal, tal y como lo requieren lo modelos probabilísticos.

3.3 Modelos logit y probit.

Una vez identificados los problemas del MPL, se contempla el uso de metodologías correctivas para garantizar estimaciones acordes a los modelos probabilísticos. Dado que se requieren predicciones de probabilidad no lineales, los modelos deben cumplir:

1. A medida que una X_i aumente, $P_i(Y_i = 1 | X_i)$ se incremente dentro intervalo [0,1]
2. La relación de P_i y X_i no debe ser lineal, en especial, debe tener un ajuste a los datos en forma de S[28](*véase* gráfica 3.2) (Greene, 2000, 815).

Gráfica 3.2. Diferencia teórica entre logit y probit



Fuente: los autores

La gráfica 3.2 muestra un mejor ajuste a los datos respecto a las predicciones derivadas de MPL en la gráfica 3.1. Es fácil verificar que **Xβ** se encuentra dentro de un rango de $(-\infty, +\infty)$ y las estimaciones están entre [0,1]. Al mismo tiempo, la probabilidad de que un evento suceda no está linealmente relacionada con $X\beta$.

De acuerdo a lo anterior, y para entender el funcionamiento básico de estos modelos, se utiliza una especificación distinta a la trabajada en los capítulos

[28] Esta forma se parece a la función de distribución acumulativa (FDA). La FDA de una variable aleatoria X es la probabilidad de que la variable tome un valor menor o igual a X_0, donde X_0 es algún valor numérico especificado de X.

anteriores. Ahora, la relación entre X_i y Y_i está determinada por una función F (véase ecuación 3.7).

$$P(Y_i = 1 | \mathbf{X}) = F(\mathbf{X\beta}) \qquad (3.7)$$

En la ecuación 3.7 se asume que F es una función que toma valores en un intervalo abierto (0,1), es decir, que $0 < F(\mathbf{X\beta}) < 1$ para todo $\mathbf{X\beta} \in \mathbb{R}$. Esta se conoce como el modelo base (o index model en inglés), dado que establece el tipo de respuesta que determina $P(Y_i = 1 | \mathbf{X})$ condicionado a $\mathbf{X}$[29]. (Wooldridge, 2002, 457). Adicionalmente, la forma específica que toma $F(\cdot)$ se puede derivar a partir d un modelo de variable latente (*véase* ecuación 3.8).

$$Y_i^* = \mathbf{X\beta} + \upsilon_i, \;\; Y_i = 1 \text{ si } Y_i^* > 0 \qquad (3.8)$$

En la ecuación Y_i^* es conocida como variable latente, Y_i es una variable dicótoma con valores cero o uno; $\mathbf{X}$ es una matriz que contiene todas las variables explicativas del sistema y υ_i es el vector de errores. Por tanto, la probabilidad de tener éxito en un evento está determinada de la siguiente forma:

$$P(Y_i = 1 | \mathbf{X}) = P(Y_i^* > 0 | \mathbf{X}) = P(e_i > -\mathbf{X\beta} | \mathbf{X}) \qquad (3.9)$$

$$P(e_i > -\mathbf{X\beta} | \mathbf{X}) = 1 - F(-\mathbf{X\beta}) = F(\mathbf{X\beta})^{30} \qquad (3.10)$$

Las expresiones de las ecuaciones 3.9 y 3.10 muestran la equivalencia teórica representada en la ecuación 3.7. Este resultado, en primera instancia, genera una relación hipotética entre $\mathbf{X}$ y Y_i^*. Pero esto no es cierto en la realidad, puesto que existe dificultad en la derivación de una unidad de medida para Y_i^*, por ende para evaluar estos modelos se recurre a otras metodologías. En la práctica, se utilizan los modelos logit y probit.

---

[29] Para los modelos econométricos la $F(\cdot)$ es una FDA.

[30] Para más detalles, *véase* (Wooldridge, 2002, 457).

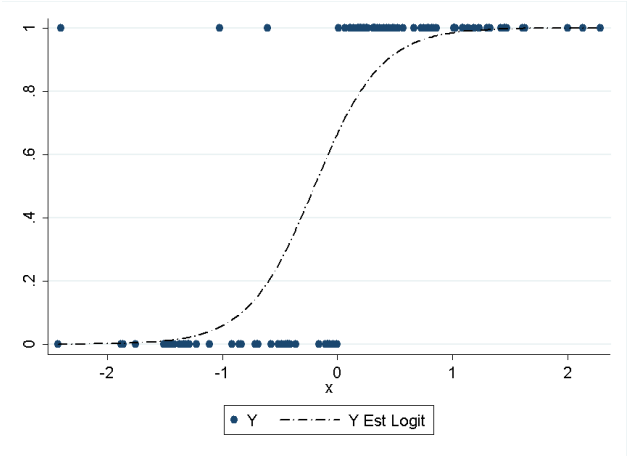
3.3.1 Definición del modelo logit

El modelo logit, es una de las metodologías que permiten estimar apropiadamente los modelos probabilísticos. Este método se basa en la función de probabilidad logística acumulativa, con errores del modelo que siguen una distribución logística (véase ecuación 3.11).

$$P(Y_i = 1 \mid \mathbf{X}) = F(\mathbf{X\beta}) = \frac{e^{\mathbf{X\beta}}}{[1 + e^{\mathbf{X\beta}}]} \qquad (3.11)$$

La ecuación 3.11 muestra que la probabilidad de que suceda un evento ($Y_i = 1$) no sigue una función lineal como se mostraba a través del modelo de probabilidad lineal, sino que tiene una especificación exponencial. Es así como el modelo logit se ajusta a los requerimientos enunciados anteriormente, especialmente que las estimaciones estén dentro del rango [0,1], tal y como lo muestra la gráfica 3.3.

Gráfica 3.3. Modelo logit



Fuente: los autores

La gráfica 3.3 representa una FDA particular (logística), la cual tiene la misma forma de "s" de la gráfica 3.2. Es fácil verificar que la probabilidad predicha se encuentra dentro del rango [0,1] y la probabilidad no está linealmente relacionado

con $\mathbf{X\beta}$ (Gujarati, 2003, 575). Si se comparan las estimaciones de MPL con las del logit se esperan resultados distintos cuando:

1. Existen muy pocas observaciones que representen la respuesta $Y = 1$ ó $Y = 0$

2. Existe mucha variabilidad en una importante variable independiente.

A continuación se introducirá otra metodología considerada para estimar modelos probabilísticos, para reconocer semejanzas, diferencias y pertinencia entre las alternativas de predicción. Además, se estudiara la forma con la que se llevan a cabo las estimaciones de este tipo de modelos.

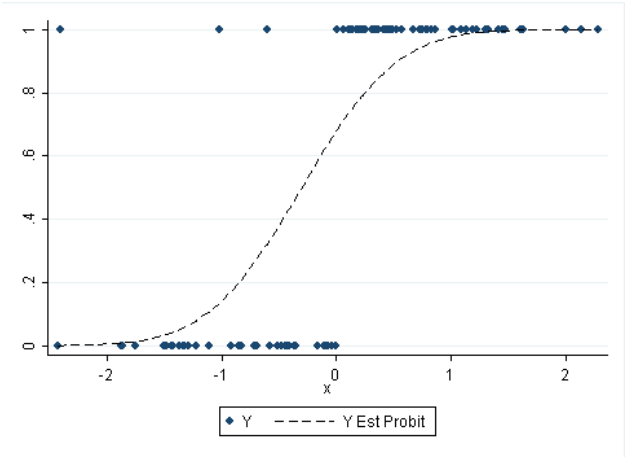
3.3.2 Definición del modelo probit

Una vez se conocen las particularidades del modelo logit, esta sección se encarga de caracterizar el modelo probit (normit), con el fin de identificar similitudes o diferencias en la determinación de modelos probabilísticos. Este último, en particular supone que los errores del modelo siguen una distribución normal. Por lo anterior, la función de probabilidad está dada por:

$$P(Y_i = 1 \mid X) = F(\mathbf{X\beta}) = \int_{-\infty}^{\mathbf{X\beta}} \phi(z)dz = \phi(\mathbf{X\beta})$$
$$\phi(\mathbf{X\beta}) = \frac{1}{\sqrt{2\pi}} e^{-(\mathbf{X\beta})^2/2} \qquad (3.12)$$

La ecuación 3.12 muestra que la probabilidad de que suceda un evento ($Y_i = 1$) está definida a través de una función no lineal. Así como en el modelo logit, ésta especificación se ajusta bien a las características requeridas para la estimación de modelos probabilísticos. La gráfica 3.4 muestra como las estimaciones de este método están dentro del rango de probabilidad deseada.

Gráfica 3.4. Modelo probit



Fuente: los autores

La diferencia con el modelo logit radica en que la distribución normal tiene los extremos ligeramente más angostos (*véase* gráfica 3.4). Esto quiere decir que la probabilidad condicional se aproxima a cero a una tasa mayor (Greene, 2000, 815). A continuación se describirá la técnica utilizada para la estimación de las especificaciones logit y probit.

3.3.3 Estimación máxima verosimilitud

Teniendo en cuenta la definición de las metodologías logit y probit, en esta sección se discute la técnica utilizada en la estimación de modelos probabilísticos. La derivación de estimadores muestrales de estos modelos, se realiza a través de de máxima verosimilitud (MV). Este es un procedimiento estadístico, que tiene como objetivo la derivación de estimadores para un vector $\boldsymbol{\beta}$ de parámetros desconocidos, a través de funciones $f(\mathbf{X},\boldsymbol{\beta})$ que definan X_i y permitan encontrar la probabilidad máxima para la función de densidad probabilística. Para los modelos logit y probit se considera una función conjunta de la siguiente forma:

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} G(Y_i \mid \mathbf{X};\boldsymbol{\beta}) \qquad (3.13)$$

La ecuación 3.13 es una función de verosimilitud definida como la productoria de funciones de densidad de variables independientes, donde $G(Y_i \mid \mathbf{X};\boldsymbol{\beta}) = f(\mathbf{X},\boldsymbol{\beta})$ y está caracterizada por una ecuación binomial (*véase* ecuaciones 3.14 y 3.15):

$$G(Y_i \mid \mathbf{X};\boldsymbol{\beta}) = [P(Y_i = 1 \mid \mathbf{X})]^{Y_i}[P(Y_i = 0 \mid \mathbf{X})]^{1-Y_i} \qquad (3.14)$$

$$G(Y_i \mid \mathbf{X};\boldsymbol{\beta}) = [F(\mathbf{X}\boldsymbol{\beta})]^{Y_i}[1 - F(\mathbf{X}\boldsymbol{\beta})]^{1-Y_i} \qquad (3.15)$$

Cuando se cuenta con observaciones independientes, el cálculo de la función de verosimilitud, donde interviene el producto de las probabilidades individuales, habitualmente se toma el logaritmo de la función, puesto que se transforman los productos en sumas y los cocientes en restas (*véase* ecuaciones 3.16 y 3.17). Teniendo en cuenta que se trabaja con el producto de probabilidades, la función de verosimilitud será siempre menor que 1 y por tanto su logaritmo será negativo.

$$\ln(L(\boldsymbol{\beta})) = \ln(\prod_{i=1}^{n}[F(\mathbf{X}\boldsymbol{\beta})]^{Y_i}[1 - F(\mathbf{X}\boldsymbol{\beta})]^{1-Y_i}) \qquad (3.16)$$

$$\ln(L(\boldsymbol{\beta})) = \sum_{i=1}^{n}\left[Y_i \ln(F(\mathbf{X}\boldsymbol{\beta})) + (1 - Y_i)\ln(1 - F(\mathbf{X}\boldsymbol{\beta}))\right] \qquad (3.17)$$

Las ecuaciones 3.16 y 3.17 son la transformación logarítmica de la ecuación 3.13. Por consiguiente, para encontrar los estimadores de MV es necesario derivar $\ln(L(\boldsymbol{\beta}))$ respecto a $\boldsymbol{\beta}$ para encontrar estimadores ($\hat{\boldsymbol{\beta}}$) de mínima varianza, insesgados y consistentes. Para esto, la función de verosimilitud necesita ser estrictamente cóncava (Gourieroux, 2000, 12), puesto que se desea encontrar un único máximo global de $\ln(L(\boldsymbol{\beta}))$.

De acuerdo a lo anterior, es posible utilizar la técnica expuesta para llevar a cabo la estimación de modelos probabilísticos. Teniendo en cuenta la ecuación 3.17, se utiliza la función de densidad asociada a los métodos logit y probit. Dado que las ecuaciones de máxima verosimilitud asociadas con éstos no son lineales en los parámetros, no es trivial encontrar expresiones analíticas que deriven los estimadores de interés. Por tanto, es necesario utilizar algoritmos numéricos o métodos matemáticos para encontrar los parámetros del modelo que se pretenden estimar (Gourieroux, 2000, 13).

Después de encontrar estimaciones por MV se hace necesario enfatizar el las particularidades originas de ésta técnica. En primera instancia, el resultado que se obtiene es distante a lo que se ha trabajado bajo la técnica de mínimos cuadrados, puesto que la distribución de los errores de los modelos logit y probit exige la utilización del estadístico Z para análisis de significancia individual. Asimismo para probar la significancia global se utiliza el estadístico de razón de verosimilitud que sigue una distribución χ_q^2 con q número de restricciones. En segunda instancia, MV para modelos probabilísticos no soluciona el problema de hetoroscedasticidad identificado en MPL, por tanto es importante que éste se corrija usando el método de mínimos cuadrados generalizados (MCG) que, por medio de la transformación del modelo inicial, consigue mejorar la eficiencia de los estimadores.

Por otro lado, cuando se utilizan por separado los modelos logit y probit se pueden obtener estimadores similares, dado que las funciones de distribución acumulada son parecidas. En relación con lo anterior, para hacer comparaciones entre las dos metodologías hay que tener en cuenta que la distribución logística muestra variación de $\pi^2/3$, por tanto la equivalencia de los $\hat{\beta}_{Logit}$ respecto a los $\hat{\beta}_{probit}$ es $\hat{\beta}_{Logit} = 1.6 \cdot \hat{\beta}_{probit}$ (Greene, 1999, 755).

3.3.4 Pruebas de significancia

Con respecto a las estimaciones de MV, es posible hacer análisis estadísticos después de encontrar los estimadores, en especial, esta técnica permite comparar modelos y corroborar la existencia de dependencia global entre las variables explicativas y la variable dependiente dicótoma de un modelo de interés. Para ello es indispensable especificar los dos modelos que se quieren contrastar, si se tiene en cuenta un modelo con dos variables explicativas (X_{i1} y X_{i2}) la prueba seria:

Modelo sin restricción: $$Y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i1} + u_i \quad (3.18)$$

Modelo con restricción: $$Y_i = \beta_0 \quad (3.19)$$

La ecuación 3.18 muestra el modelo inicial que se estimaría a través de máxima verosimilitud, mientas tanto la ecuación 3.19 representa un modelo caracterizado por no contar con las variables explicativas. Para realizar la comparación entre los estimadores, se utiliza una prueba de hipótesis que compare las funciones lineales de los parámetros de cada uno de los modelos (*véase* ecuación 3.20).

$$\begin{array}{cc} H_0 : \beta_0 = \beta_1 = ... = \beta_k = 0 & \text{Los coeficientes no son significativos conjuntamente} \\ H_1 : \beta_j \neq 0 & \text{Los coeficientes son significativos conjuntamente} \end{array} \qquad (3.20)$$

Dado que las funciones de densidad de probabilidad no se distribuyen normal, como en los modelos de regresión lineal, los estadísticos de la prueba de hipótesis deben ser distintos a las pruebas t o F. Por tanto el estadístico de prueba que se utiliza es una razón de verosimilitud (RV) (*véase* ecuación 3.21)

$$RV = 2(l_{NR} - l_R) \sim \chi_q^2 \qquad (3.21)$$

En la ecuación 3.21, RV se conoce como el cociente de verosimilitud (likelihood ratio en inglés); l_{NR} es el logaritmo natural del modelo no restringido[31] ; l_R es igual al logaritmo natural del modelo restringido[32]; q de la distribución χ^2 hace referencia al número de restricciones del sistema. Seguidamente, la forma de interpretar la prueba es comparar RV con el χ_q^2 de las tablas. Si $RV > \chi_q^2$ entonces se dice que se rechaza la hipótesis nula, en otras palabras, las variables independientes en conjunto son importantes para el modelo que se está estimando. En las siguientes secciones se trataran temas relacionados con pos estimaciones de MV.

[31] $\ln(L(\beta)_{NR})$

[32] $\ln(L(\beta)_R)$

3.3.5 Efectos marginales

La interpretación de los coeficientes para modelos probabilísticos estimados a través de máxima verosimilitud no es igual que el cambio marginal que se tenía en cuenta en un modelo de regresión lineal. Para el MPL los resultados se interpretan como el cambio marginal de la probabilidad dado un cambio en el valor de una variable independiente. Por el contrario, los modelos logit y probit utilizan otra estructura (*véase* ecuación 3.22)

$$\frac{\partial P(Y=1\mid X)}{\partial X_{ij}} = \frac{\partial E(Y\mid X)}{\partial X_{ij}} = \frac{\partial F(X\beta)}{\partial X_{ij}} = f(X\beta)\beta_i \qquad (3.22)$$

La ecuación 3.22 muestra que el efecto de un cambio en X_{ij} sobre la probabilidad que un evento suceda, depende de X a través de la función de densidad de probabilidad $f(X\beta)$. Esta expresión indica el cambio en el logaritmo de las probabilidades asociadas como resultado de un cambio en una variable independiente, manteniendo el resto de las variables explicativas constantes. Teniendo en cuenta que el efecto marginal es variable, lo que se hace es valorarlo con respecto al promedio de las variables.

3.3.6 Bondad de ajuste

El R^2 convencional en las estimaciones de regresiones lineales clásicas, no es compatible con las estimaciones de máxima verosimilitud de modelos probabilísticos. Esto como consecuencia de no tener un ajuste acorde a la dispersión de los datos evidenciados en MPL, por lo que es necesario usar indicadores acordes a la característica específica de los datos. En estas circunstancias, el p-seudo R^2 (o R^2 de McFadden), y el porcentaje de predicciones correctas son buenos indicadores del ajuste de modelos probabilísticos (Gujarati, 2003, 584).

3.3.6.1 P-seudo R^2 o R^2 de McFadden

El estimador de ajuste para modelos probabilísticos se realiza a través del p-seudo R^2, que parte de la especificación de los modelos probabilísticos y ofrece un resultado confiable para determinar la bondad de ajuste de los modelos que se estén trabajando. El estimador aproximado al R^2 toma la siguiente forma especifica:

$$p-seudo \;\; R^2 = 1 - \frac{l_{NR}}{l_R} \qquad (3.23)$$

Donde l_{NR} es igual al máximo del logaritmo natural del modelo no restringido; l_R es igual al logaritmo natural del modelo restringido. Estos dos valores se consiguen estimando cada modelo por separado y extrayendo el resultado de log verosimilitud. Este estimador se interpreta igual que el R^2 de la regresión lineal clásico, pero no se debe sobrevalorar la importancia de éste en modelos para los que la variable dependiente es una dicótoma (Gujarati, 2003, 585).

3.3.6.2 Ajuste del modelo - porcentaje de predicciones correctas

El porcentaje de predicciones correctas es otra medida para establecer si las estimaciones están acorde con los datos observados. Este procedimiento consiste en crea una variable ficticia a partir de la cual se puede contrastar los valores predichos con los observados. Para ello, se predice la probabilidad que $Y_i = 1$ dadas las variables explicativas, en relación a los datos observados para la misma variable cualitativa. Si $F(X\beta) > 0.5$ entonces $Y_i = 1$, y si $F(X\beta) \leq 0.5$ la predicción es $Y_i = 0$. A partir de esta desagregación, se puede obtener un registro de qué tan ciertas son las predicciones de los modelos probabilísticos (*véase* cuadro 3.1).

Cuadro 3.1. Cuadro de predicciones correctas

		Estimado		
Observado		Y=0	Y=1	Total
	Y=0	n_1	n_2	$n_1 + n_2$
	Y=1	n_3	n_4	$n_3 + n_4$
	Total	$n_1 + n_3$	$n_2 + n_4$	N

Fuente: los autores

Las predicciones correctas son las que resultan iguales respecto a los datos observados, esto se ve en el cuadro cuando coinciden $Y = 0$ y $\hat{Y} = 0$; y $Y = 1$ y $\hat{Y} = 1$. El porcentaje de predicciones correctas resulta de contabilizar el número de predicciones correctas y dividirlas en el número total de observaciones (*véase* ecuación 3.24)

$$PPC = \frac{n_1 + n_4}{N} \qquad (3.24)$$

Donde PPC es el porcentaje de predicciones correctas, n_1 numero de predicciones correctas para $Y = 0$ y n_4 son las predicciones correctas para $Y = 1$, y N es el número total de observaciones.

3.4 Estudio de caso: mercado de trabajo informal en Colombia

El caso empírico que se desarrolla en este capítulo está basado en el artículo titulado (en inglés) *"The Informal Labor Market in Colombia: Identification and characterization"* por Bernal (2008). Este trabajo evalúa los determinantes del trabajo informal del mercado laboral Colombiano.

Para entender el enfoque del artículo es indispensable trabajar sobre la base de la definición que el empleo informal se refiere al trabajo no reportado y aquel que evade la regulación formal y deja a los participantes desprotegidos y vulnerables. Con base en ésta definición, el estudio quiere identificar la naturaleza de la informalidad en Colombia, en especial, las características personales y sociodemográficas del grupo de personas que se encuentran realizando trabajos

informales. Esto, con el fin de entender los motivos e incentivos que tienen los trabajadores para pertenecer o no al sistema de trabajo formal.

En esta sección se estimaran los determinantes de la informalidad en Colombia, siguiendo las metodologías de modelos probabilísticos como MPL, logit y probit, basado en el modulo de informalidad aplicado por el DANE en la Encuesta Continua de Hogares (ECH). El modelo empírico que se estudiará tiene la siguiente especificación:

$$Y_i = \mathbf{X}\boldsymbol{\beta} + e_i \qquad (3.25)$$

Donde Y_i es el vector de la variable dependiente dicótoma que toma el valor de uno si representa informalidad y cero en caso contrario; $\mathbf{X}$ matriz de variables que caracterizan la informalidad; e_i es el término del error de la regresión. A continuación se lleva a cabo todo el desarrollo que conlleva a los resultados esperados de los modelos probabilísticos.

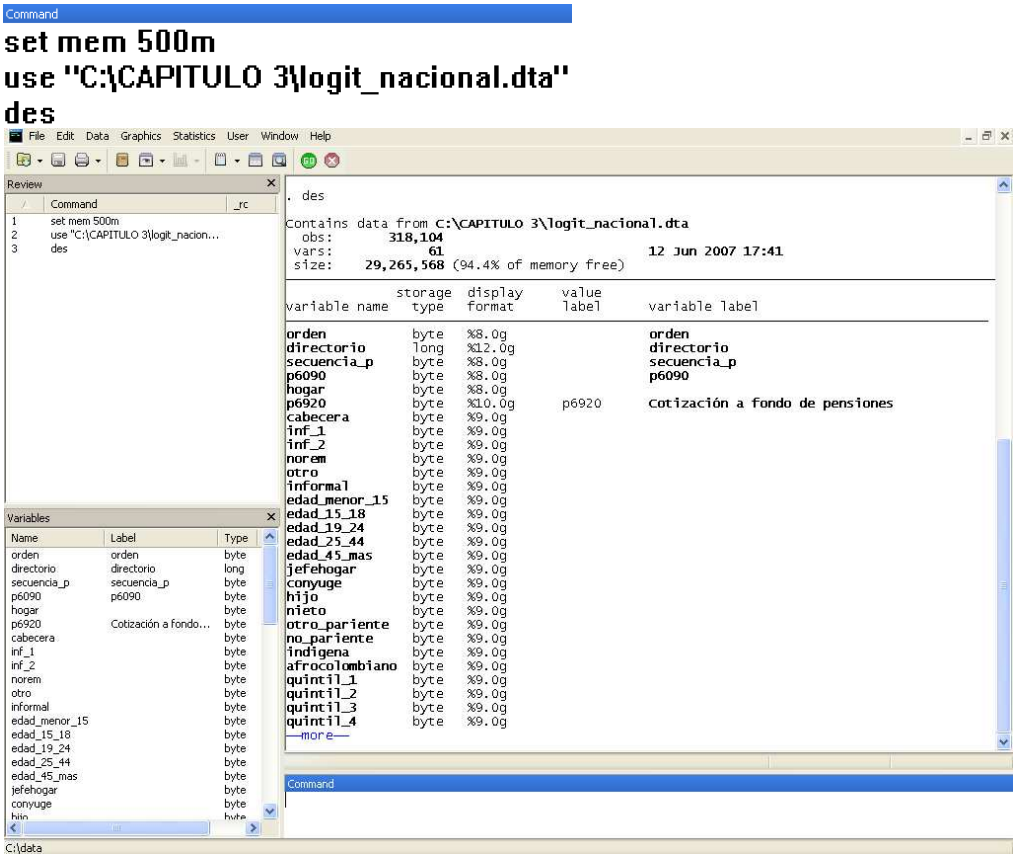
3.4.1 Análisis general de los datos

1. En Stata®, al igual que en los demás capítulos, se determina la memoria con la que se va a cargar la base de datos de interés. Esto se consigue con el comando *set mem*. Para este ejercicio, se utiliza una memoria de 500m tal como se ve en la figura 3.1, debido al tamaño de la información con que se trabaja.

   Una vez se ha asignado la memoria del sistema, se carga la base de datos. Ésta hace parte de la información que utilizó Bernal (2008) y lleva el nombre *capitulo3.dta*. La información hace referencia a la ECH para el año 2006 en el periodo de agosto a diciembre.

2. Para observar las variables que se encuentran disponibles para el ejercicio, se usa el comando *describe* –o *des-*. Este comando hace que STATA muestre una cuadro con la lista de las variables que se encuentran en la base, el formato en que están guardadas, y una descripción de cada una (*véase* figura 3.1).

Figura 3.1. Salida comandos set mem, use y des



Fuente: cálculos autores

En la figura 3.1 se observa que la base cuenta con 318.104 observaciones y 61 variables disponibles para realizar las metodologías pertinentes en cada caso. En el cuadro 3.2 se presentan las variables a usar, en relación a las de la ecuación 3.25.

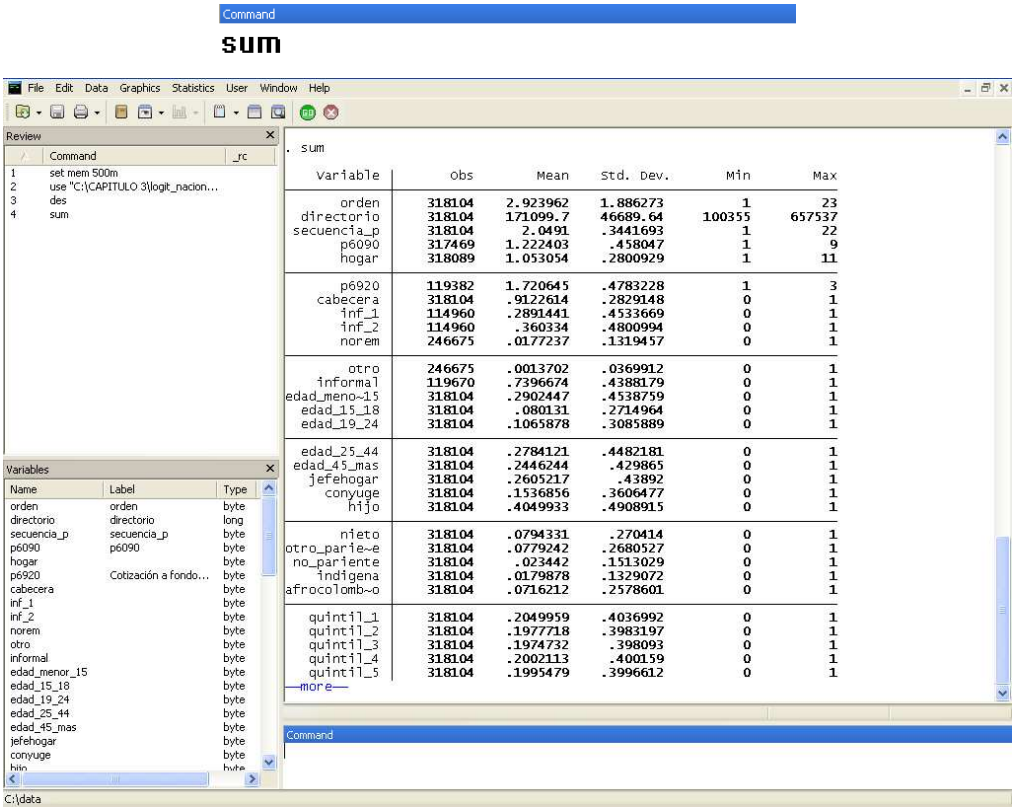
Cuadro 3.2. Variables a usar en el modelo

Variable del Modelo	Variables en la Base	Descripción
Y	informal	Es una dummy que toma el valor de uno cuando se cumplen las características de informalidad y cero en caso contrario.
X	Hombre Edad_15_18 Edad_19_24 Edad_25_44 Edad_45_mas Jefehogar Conyuge Hijo Nieto Otro_pariente Cabecera Educ_primaria Educ_secundaria Educ_superior Indígena Afrocolombiano Quintil_1 Indepte_otros	Conjunto de variables explicativas. Todas son dicótomas que toman el valor de uno si cumplen con la característica descrita en el nombre de la variable, y cero de lo contrario.

Fuente: los autores

3. Antes de pasar a estimar la regresión lineal, es necesario observar las estadísticas descriptivas de las variables a usar. El comando *summary* –o *sum-*, presenta un cuadro con el número de observaciones, la media, desviación estándar y mínimo y máximo de las variables de la base de datos o las consideradas en el cuadro 3.2 (*véase* figura 3.2).

Figura 3.2. Salida comando sum



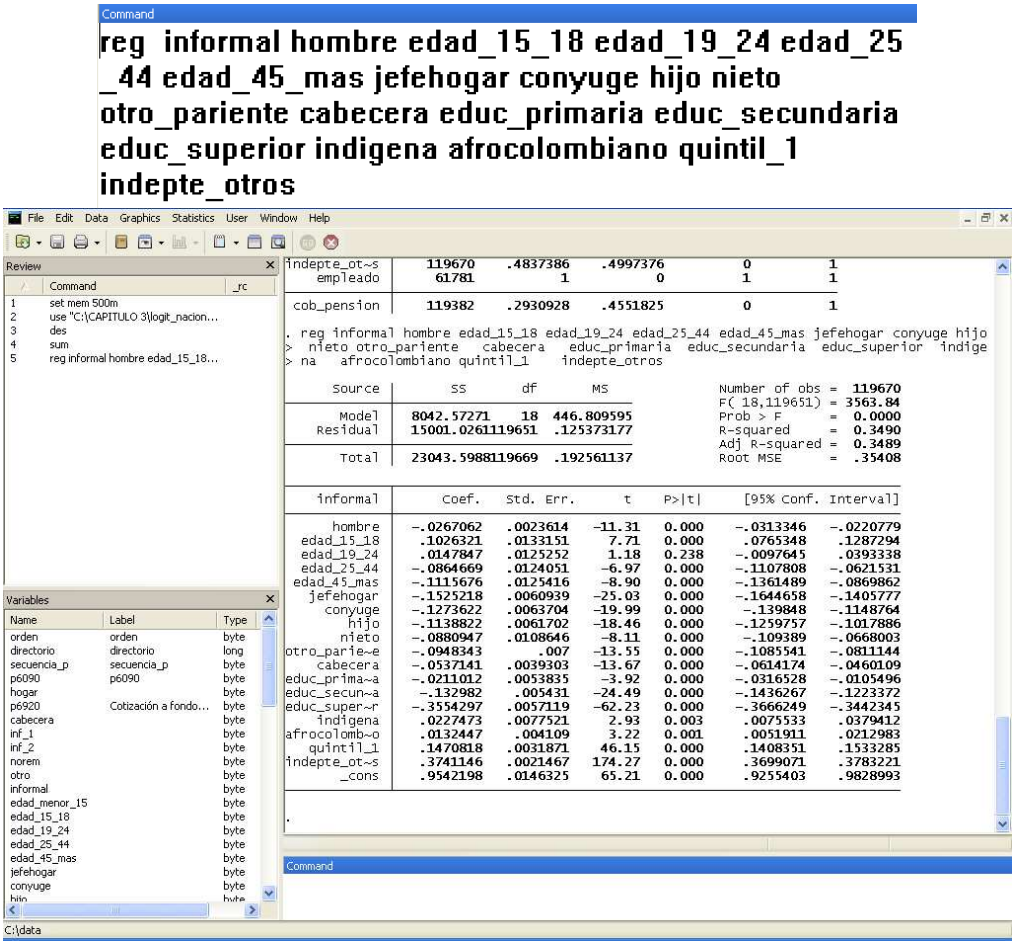
Fuente: cálculos autores

3.4.2 Estimación del modelo MPL

Después de conocer la información necesaria para las estimaciones, se puede pasar a desarrollar el ejercicio empírico propuesto por Bernal (2008). El objetivo es estimar la ecuación 3.25 a través de MPL, teniendo en cuenta las variables del cuadro 3.2.

1. Para llevar a cabo este objetivo, en Stata® se usa el comando *reg* seguido de las variables dependiente y explicativas, bajo un modelo de regresión lineal (*véase* figura 3.3).

Figura 3.3. Salida regresión de modelo de probabilidad lineal



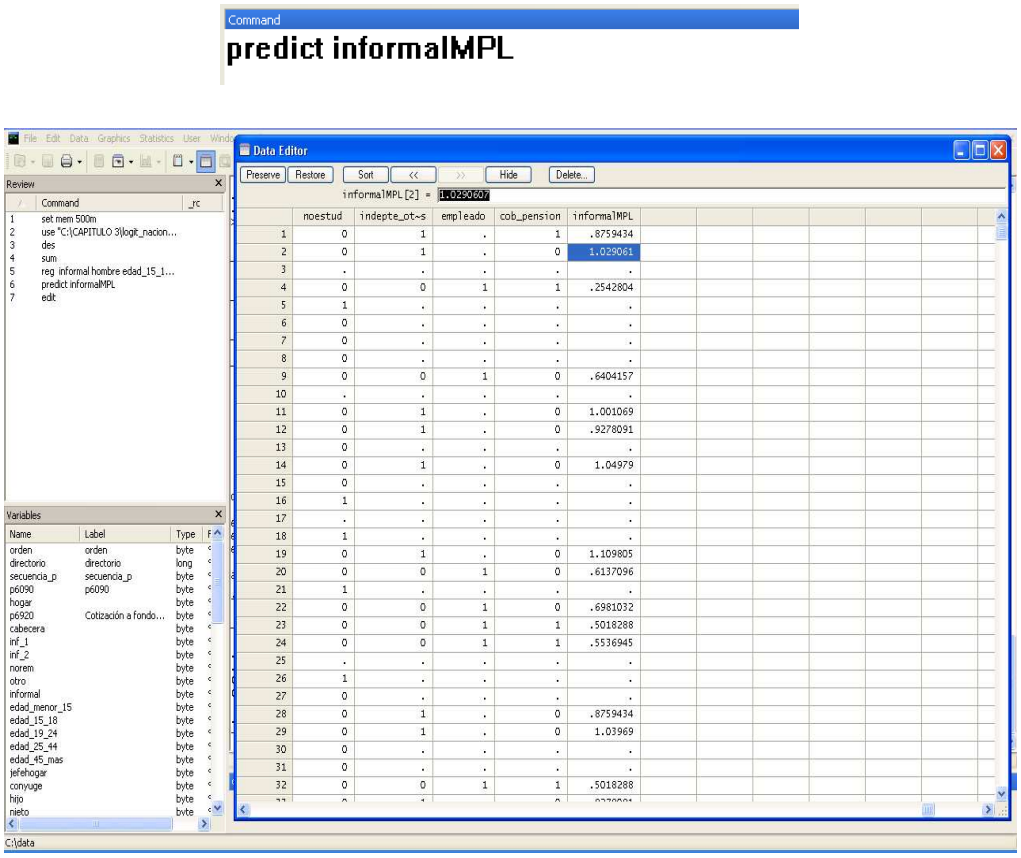
Fuente: cálculos autores

Los resultados de la figura 3.3 indican que ser joven, estar en el quintil más bajo de ingreso, ser afrocolombiano o independiente aumenta la probabilidad de ser informal, respecto a la persona que no esas particularidades. Tener mayor educación, ser hombre o con edad media disminuye en 35, 2.6 o 8.6 puntos porcentuales, respectivamente, la probabilidad de estar en la informalidad respecto a los individuos que no presentan dichas características.

Ahora bien, estas primeras conclusiones deben mirarse con cautela, como se mencionó en la sección 3.2, puesto que las estimaciones por MPL pueden presentar errores, ya sea porque las predicciones del modelo no se ajustan al rango de probabilidad o porque existen problemas de heteroscedasticidad.

2. Para detectar dichos errores, es necesario revisar las predicciones del modelo MPL. Para ello se utiliza el comando *predict* seguido del nombre con el que se crea la nueva variable predicha (*véase* figura 3.4).

Figura 3.4. Salida predicciones del modelo de probabilidad lineal

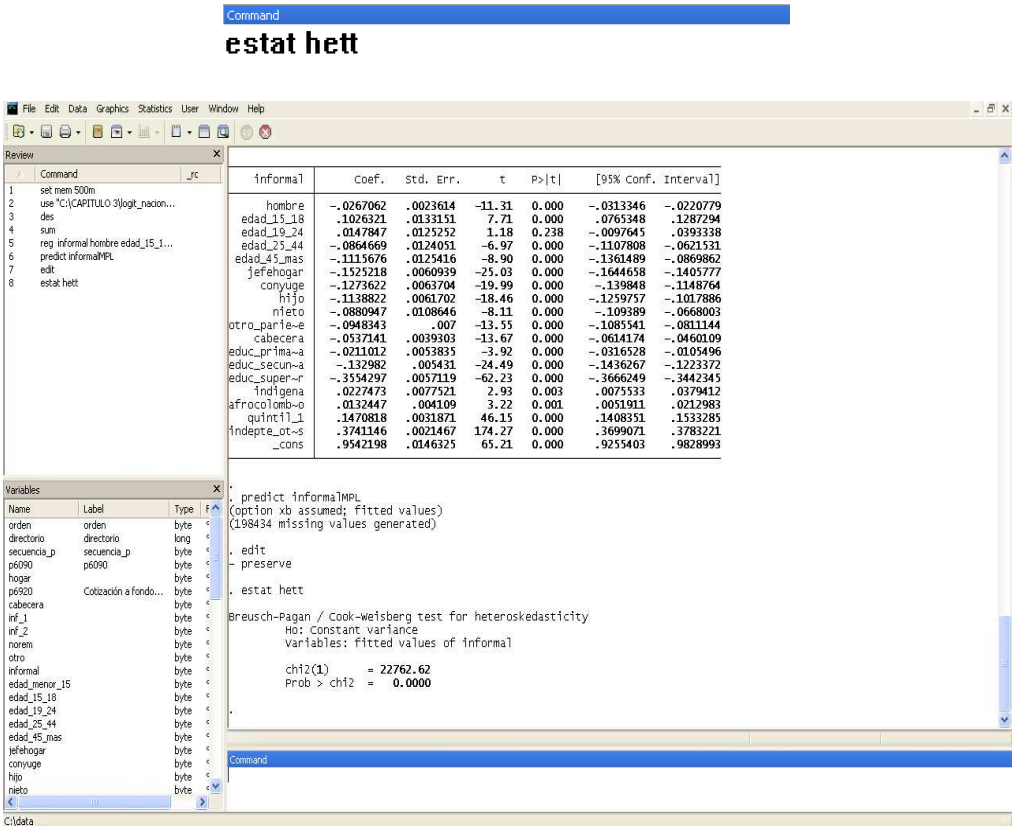


Fuente: cálculos autores

En la figura 3.4 se evidencian las falencias descritas en la sección 3.3.1, dado que las predicciones de la variable dependiente resultan alejadas del intervalo [0,1]. Lo anterior sugiere que el modelo de MPL no se ajusta bien a los requerimientos del modelo de Bernal (2008).

3. Antes de introducir nuevas metodologías para derivar resultados satisfactorios, es pertinente hacer una prueba de heteroscedasticidad al modelo 3.25. Esta prueba se hace a través del comando *estat hett* luego de haber realizado la regresión.

Figura 3.5. Salida prueba de heteroscedasticidad. Comando estat hett



Fuente: cálculos autores

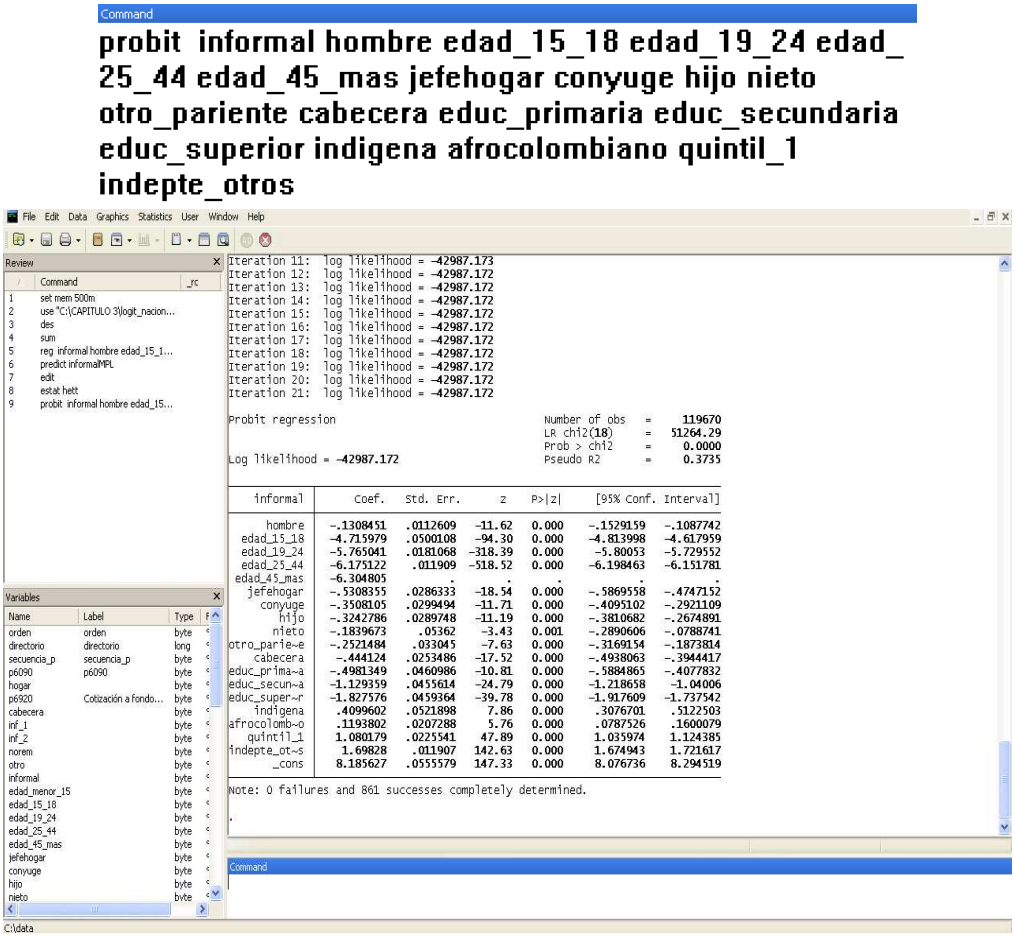
La figura 3.5 muestra el resultado de la prueba que evalúa la hipótesis nula de varianza constante (homoscedasticidad), contra la hipótesis alterna que dice que la varianza no es constante (heteroscedasticidad). En el modelo 3.25, se rechaza la hipótesis nula con un p-valor de 0.000, concluyendo así que existen problemas de heteroscedasticidad. Por consiguiente, es necesario volver a estimar el modelo con errores estándar robustos (opción *robust* después del comando *reg*), o utilizar otras metodologías que garanticen estimaciones precisas y eficientes que generen conclusiones favorables para la investigación de interés.

3.4.3 Estimación del modelo probit

Después de advertir los problemas presentados por la metodología MPL, en esta sección se llevará a cabo una nueva estimación de la ecuación 3.25, esta vez utilizando el modelo probit. Como se estudió en la sección 3.3.2, esta metodología garantiza que las estimaciones estén acorde al rango [0,1] y a los requerimientos de los modelos probabilísticos. A continuación se enunciarán los pasos a seguir para conseguir los resultados bajo este modelo.

1. El comando en Stata® para realizar la estimación del modelo 3.25 con esta metodología, es *probit* seguido de las variables del modelo, tal y como se muestra en la figura 3.6.

Figura 3.6. Salida regresión de modelo probit

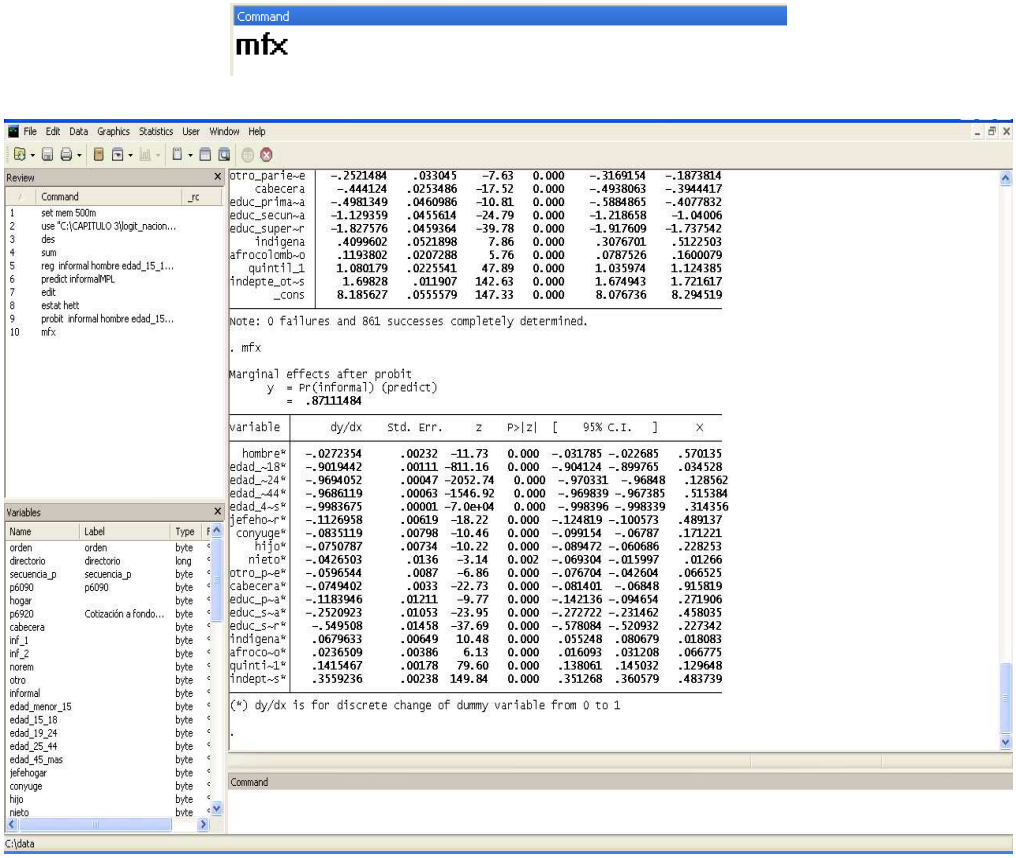


Fuente: cálculos autores

La figura 3.6 muestra las inconsistencias que se presentaban en la estimación de MPL. En los primeros grupos de edad que se tienen en cuenta, los estimadores cambian de signo positivo a negativo, y pasan a ser estadísticamente significativos dado que el p-valor de la prueba de significancia Z es 0.000. En este caso el p-seduo R^2 tiene el valor de 0.37, es decir, un 37% de la varianza de las variables explicativas en conjunto, están explicando el cambio en la probabilidad de la variable informalidad.

2. Los coeficientes que muestra la figura 3.6 no pueden interpretarse de la misma manera que aquellos derivados por MPL. Para conocer los efectos marginales del modelo probit, se utiliza el comando *mfx* justo después del resultado de la estimación. (*véase* figura 3.7).

Figura 3.7. Salida efectos marginales para el modelo probit. Comando mfx



Fuente: cálculos autores

A partir de los resultados de la figura 3.7, se pueden derivar conclusiones sobre cambios marginales en el modelo de informalidad que se está trabajando en esta sección. Por ejemplo, ser afrocolombiano aumenta la probabilidad de ser informal en 2 puntos básicos respecto a las personas que no lo son[33].

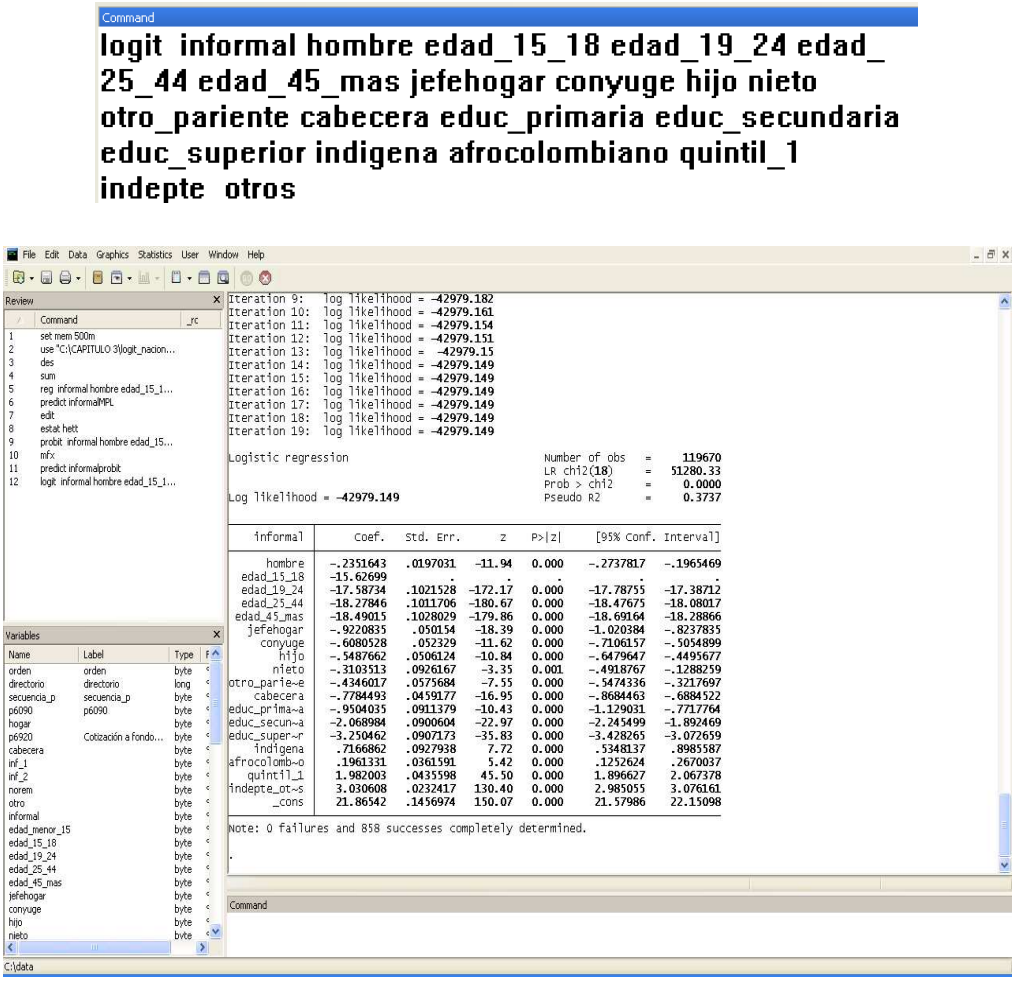
3.4.4 Estimación del modelo logit

Alternativamente a la metodología de probit, el modelo 3.25 puede ser estimado por medio del modelo logit. En esta sección se presenta el proceso de estimación con esta técnica, y se comparan las dos metodologías.

1. La estimación de la ecuación 3.25 a través de este modelo, se consigue utilizando el comando *logit* seguido por las variables explicativa y regresoras. Los resultados del procedimiento anterior están expuestos en la figura 3.8.

[33] Para las demás variables explicativas la interpretación es similar.

Figura 3.8. Salida regresión de modelo logit



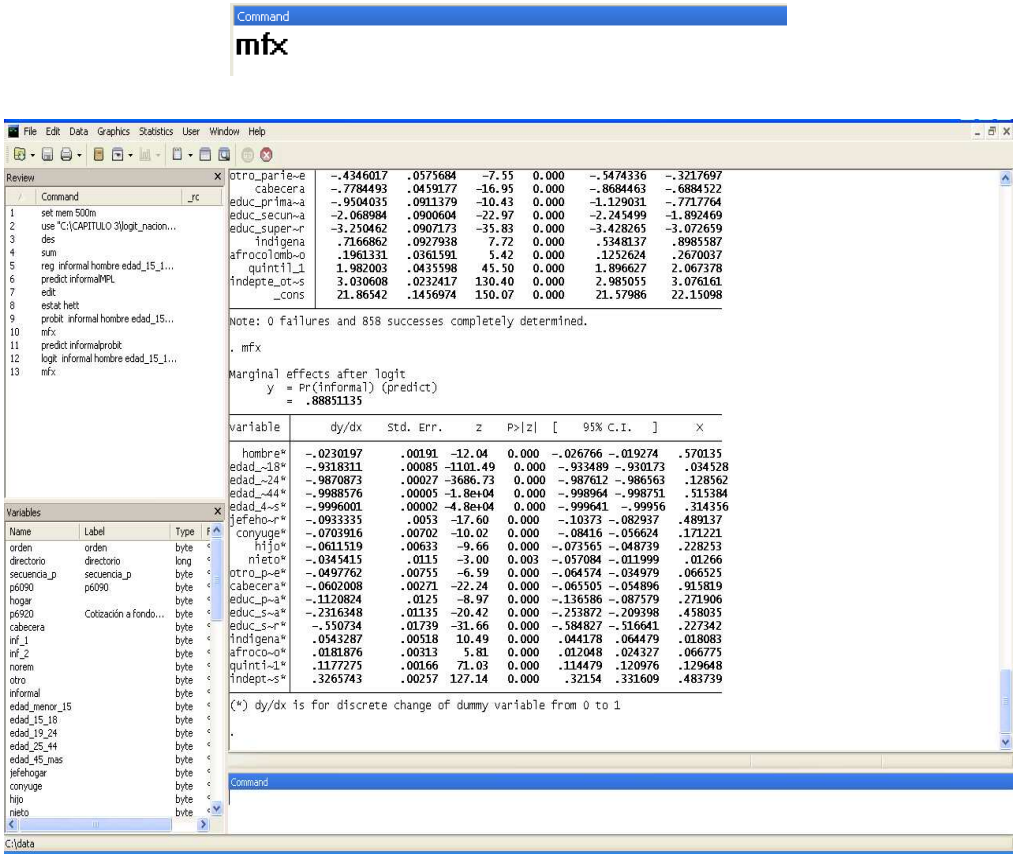
Fuente: cálculos autores

La figura 3.8 muestra la estimación del modelo logit, que resultan similares a las que se mostraron en la figura 3.6 en lo que tiene que ver con los signos, aunque existen algunas diferencias en la magnitud de los coeficientes. La significancia estadística de las variables del modelo se mantiene, puesto que los p-valores que presentan un valor de 0.00. Al mismo tiempo, el ajuste del modelo se mantiene en 0.37 respecto al modelo probit.

2. Al igual que el modelo probit, para poder interpretar los resultados, es indispensable conocer los efectos marginales del modelo logit. Para este fin

se utiliza el comando *mfx* después de la estimación, tal y como se desarrollo anteriormente.

Figura 3.9 Salida efectos marginales para el modelo logit. Comando mfx



Fuente: cálculos autores

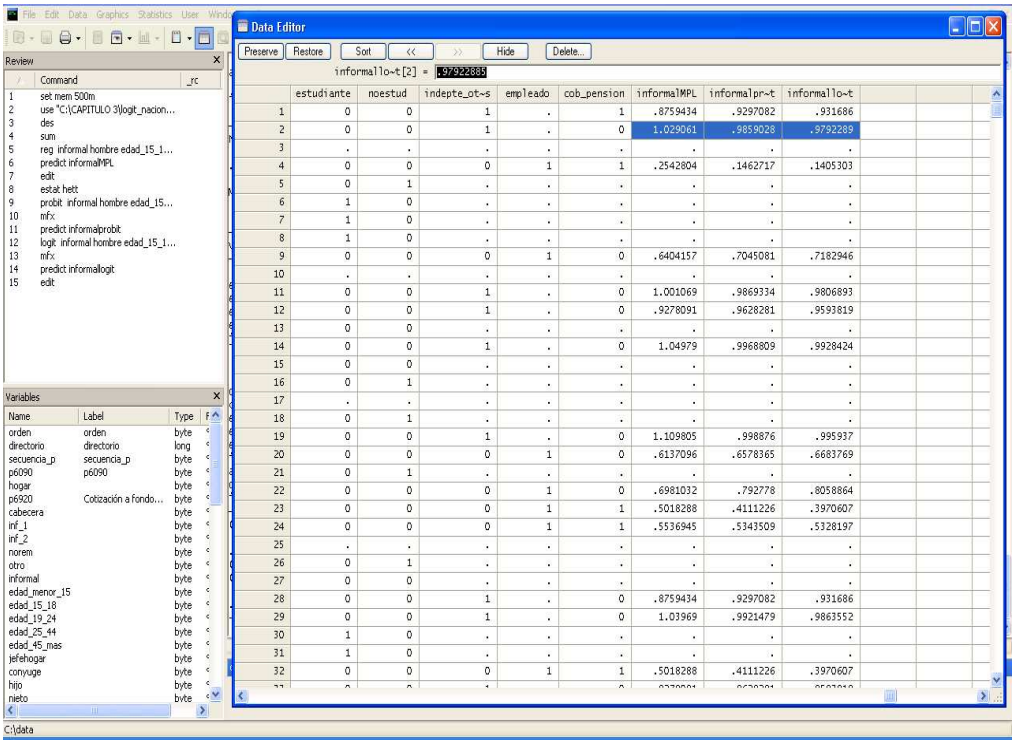
La figura 3.9 muestra los efectos marginales derivados de la estimación de un modelo logit. Como se puede evidenciar los coeficientes de las figuras 3.9 y 3.7 no resultan tan distantes. A partir de esto se pueden derivar conclusiones similares respecto a los determinantes de la informalidad en el mercado laboral de Colombia.

Bernal (2008) utilizó el modelo logit para realizar las primeras aproximaciones a los determinantes de la informalidad en Colombia. Según lo obtenido en la figura 3.9, los hombres tienen 2 puntos porcentuales de posibilidad de ser informales respecto a las mujeres. Así mismo, los

trabajadores más adultos tienen menor expectativa de ser informales que los trabajadores jóvenes. Los trabajadores urbanos tienen 6 puntos porcentuales menos que los trabajadores rurales en la posibilidad de ser informales[34].

Finalmente, se puede evidenciar como estos dos métodos solucionan los problemas de la estimación MPL, en especial a lo que concierne a las predicciones de los modelos probabilísticos (*véase* figura 3.10).

Figura 3.10. Salida predicciones MPL, probit y logit



Fuente: cálculos autores

La figura 3.10 muestra como las predicciones de MPL resultan distantes a las de probit y logit, en las dos últimas se consigue estimaciones dentro del rango [0,1], mientras que las del primero están por fuera del mismo intervalo. Por tanto, las estimaciones de probit y logit resultan más propicias cuando se trabaja con modelos caracterizados por tener variable dependiente limitada.

[34] El resto de variables explicativas en el modelo se pueden interpretar de la misma forma.

3. Para finalizar, es necesario estimar el número de predicciones correctas del modelo que se considere finalmente. De acuerdo a lo anterior, y siguiendo las estimaciones de Bernal (2008), se utilizará la estimación logit para encontrar el porcentaje de predicciones correctas. Para esto, es necesario tabular la variable dependiente observada con la predicha, a través del método expuesto en la sección 3.3.5 y de la siguiente forma (*véase* figura 3.11):

    a. Se escoge el modelo con el cual se realizaran las estimaciones finales del modelo de interés. Para este ejercicio se utiliza el modelo logit.

    b. Se realiza la estimación tal y como se vio en la figura 3.8 y 3.9.

    c. Luego se predice la variable dependiente con el comando *predict* y el nombre de la variable predicha, para el ejercicio se llamará informallogit.

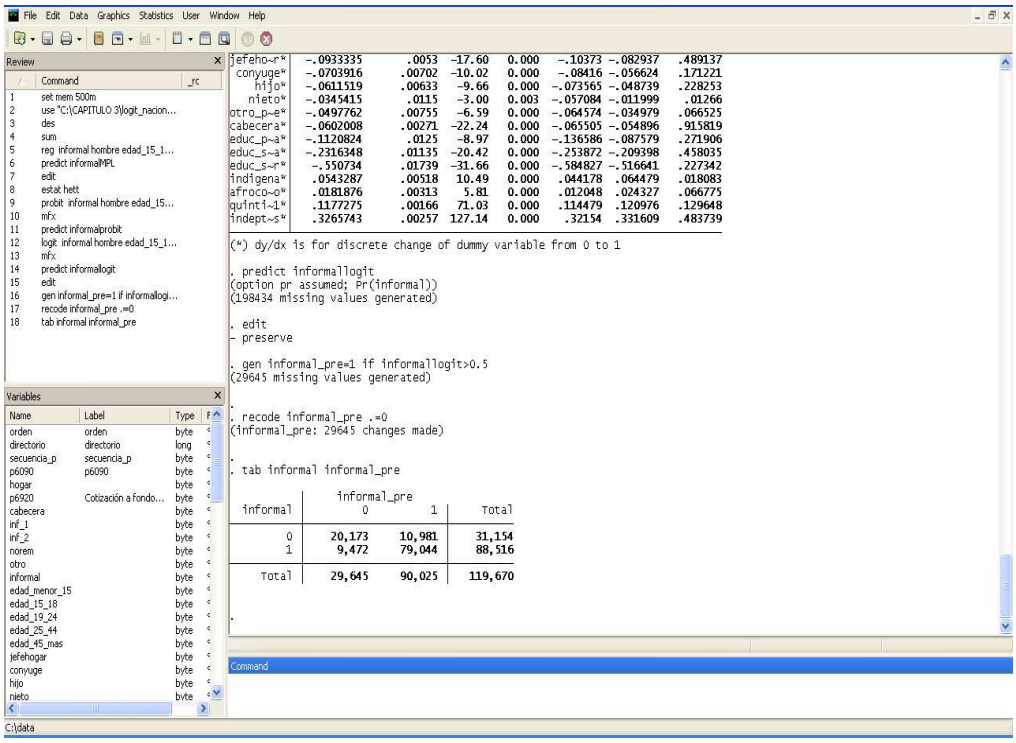
    d. Se genera una variable con la siguiente condición: si la variable dependiente predicha es mayor a 0.5 tome el valor de 1 y 0 en caso contrario. Este proceso se realiza a través del comando *gen* y la condición *if*.

    e. Posteriormente se debe cambiar puntos por ceros, puesto que la variable creada solo tiene los unos dentro de las casillas de las variables. Esto se consigue con el comando *recode* seguido por el nombre de la variable del paso 4 y adicionalmente .=0

    f. Por último, se tabulan los resultados de la variable observada y la predicha con las condiciones del paso 4 y 5. El comando para tabular es *tab* seguido de las variables de interés.

Figura 3.11. Salida porcentaje de predicciones correctas

```
Command
predict informalhat
gen informal_pre=1 if informalhat>0.5
recode informal_pre .=0
tab informal informal_pre
```



```
jefeho~r*   -.0933335    .0053   -17.60   0.000    -.10373   -.082937    .489137
 conyuge*   -.0703916   .00702   -10.02   0.000    -.08416   -.056624    .171221
    hijo*   -.0611519   .00633    -9.66   0.000   -.073565   -.048739    .228253
   nieto*   -.0345415    .0115    -3.00   0.003   -.057084   -.011999     .01266
otro_p~e*   -.0497762   .00755    -6.59   0.000   -.064574   -.034979    .066525
cabecera*   -.0602008   .00271   -22.24   0.000   -.065505   -.054896    .915819
educ_p~a*   -.1120824    .0125    -8.97   0.000   -.136586   -.087579    .271906
educ_s~a*   -.2316348   .01135   -20.42   0.000   -.253872   -.209398    .458035
educ_s~r*    -.550734   .01739   -31.66   0.000   -.584827   -.516641    .227342
indigena*    .0543287   .00518    10.49   0.000    .044178    .064479    .018083
afroco~o*    .0181876   .00313     5.81   0.000    .012048    .024327    .066775
quinti~l*    .1177275   .00166    71.03   0.000    .114479    .120976    .129648
indept~s*    .3265743   .00257   127.14   0.000     .32154    .331609    .483739
------------------------------------------------------------------------------
(*) dy/dx is for discrete change of dummy variable from 0 to 1

. predict informallogit
(option pr assumed; Pr(informal))
(198434 missing values generated)

. edit
- preserve

. gen informal_pre=1 if informallogit>0.5
(29645 missing values generated)

.
. recode informal_pre .=0
(informal_pre: 29645 changes made)

.
. tab informal informal_pre

           |  informal_pre
  informal |         0          1 |     Total
-----------+----------------------+----------
         0 |    20,173     10,981 |    31,154
         1 |     9,472     79,044 |    88,516
-----------+----------------------+----------
     Total |    29,645     90,025 |   119,670

.
```

Fuente: calculo autores

El cuadro que aparece al final de la figura 3.8 muestra el resultado de predicciones correctas del modelo logit. Siguiendo el procedimiento descrito anteriormente, para obtener el ajuste del modelo se tiene en cuenta la información de la casilla (0,0) y la de (1,1) tal y como se sugiere en la ecuación 3.24. Los resultados del modelo que utilizó Bernal (2008) muestran que el 82,8% de predicciones son correctas, con lo que se puede decir que las conclusiones derivadas en la figura 3.9 tienen suficiente validez.

Resumen

- El análisis de modelos probabilísticos demanda nuevas herramientas econométricas. Esto sugiere que los supuestos de corte transversal del modelo básico de MCO, que tiene que ver con la distribución de los errores, no están acorde con la característica de estos modelos.
- La metodología de probabilidad lineal (MPL) trata modelos probabilísticos o de variable dependiente limitada teniendo como base el modelo lineal clásico. El modelo caracterizado por tener variable dependiente dicotóma y estimado a través de MCO, presenta dos problemas: el término del error no se distribuye de forma normal, su varianza es heteroscedástica, las estimaciones se salen del rango de la probabilidad y presenta una relación lineal entre la variable explicativa y las regresoras.
- El modelo logit se basa en la función de probabilidad logística acumulativa que no sigue una función lineal. Para este modelo se tiene en cuenta la función de probabilidad logística. Mientras tanto, el modelo probit (normit) se basa en la función de probabilidad normal acumulativa.
- La estimación de modelos logit y probit se realiza a través de la metodología de máxima verosimilitud. Este es un procedimiento estadístico que supone que los datos siguen algún tipo de modelo matemático definido a través de una ecuación, en la que se desconoce alguno de sus parámetros, siendo éstos calculados o estimados a partir de la información obtenida en un modelo econométrico diseñado para tal fin.
- Las ecuaciones de máxima verosimilitud asociadas con el modelo probit y logit al no ser lineales en los parámetros, no es trivial encontrar expresiones analíticas que resuelvan el sistema. Por consiguiente, se requiere el uso de algoritmos numéricos o métodos matemáticos para encontrar los parámetros del modelo de interés.
- Dos medidas buenas para establecer si las estimaciones están acorde con los datos observados es derivar el p-seudo R^2 y el porcentaje de predicciones correctas.

Anexo 3

Anexo 3.1 Uso del modelo de regresión lineal como modelo probabilístico.

Si se tiene una ecuación de regresión lineal de tipo $Y_i = \beta_0 + \beta_1 X_{1i} + u_i$ y se supone que $E(u_i) = 0$ (para que los estimadores sean insesgados), se puede calcular el valor esperado de Y_i dados los valores de X_{i1} de la siguiente forma:

$$E(Y_i \mid X_{i1}) = \beta_0 + \beta_1 X_{i1} \qquad \text{(A.3.1)}$$

Dado que Y_i es una variable dummy la ecuación A.3.1 se puede interpretar como la probabilidad condicional de que suceda el evento Y_i dado X_{i1}. Para ello se asume que P_i y $1 - P_i$ son las probabilidades que $Y_i = 1$ y $Y_i = 0$, respectivamente. De acuerdo a lo anterior, se puede establecer que Y_i sigue una distribución de probabilidad de binomial. Por consiguiente, dada la definición de esperanza matemática, se tiene:

$$E(Y_i) = 0(1 - P_i) + 1(P_i) = P_i \qquad \text{(A.3.2)}$$

Por lo tanto se puede equiparar la esperanza condicional de la ecuación A.3.1 con la probabilidad condicional de A.3.2. La representación sería la siguiente:

$$E(Y_i \mid X_{1i}) = \beta_0 + \beta_1 X_{1i} = P_i \qquad \text{(A.3.3)}$$

La ecuación A.3.3 demuestra que en los modelos probabilísticos se deja de lado la estimación dl valor esperado de la variable independiente, para encontrar la probabilidad que sucede algún evento sujeto a un conjunto de variables explicativas.

Anexo 3.2 Prueba de heteroscedasticidad para MPL

Se parte de la definición de la varianza para una función de distribución binomial de la siguiente forma:

$$E(u_i^2) = (1 - X_i\beta)^2 P_i + (-X_i\beta)^2 (1 - P_i)$$
$$E(u_i^2) = (1 - X_i\beta)^2 X_i\beta + (-X_i\beta)^2 (1 - X_i\beta) \qquad \text{(A.3.4)}$$
$$E(u_i^2) = (1 - X_i\beta) X_i\beta$$

La ecuación A.3.4 muestra que la varianza del modelo se encuentra relacionada con las variables independientes, a través de la probabilidad de que el evento ocurra. Por tanto, la varianza se puede escribirse de la forma:

$$E(u_i^2) = \sigma^2$$
$$E(u_i^2) = P_i(1 - P_i) \qquad \text{(A.3.5)}$$

La expresión A.3.5 relaciona la varianza con la probabilidad insinuando que el modelo presenta problemas de Heteroscedasticidad.

Capítulo 4

Introducción a series de tiempo

4.1 Introducción

Hasta el momento, los capítulos anteriores han centrado su análisis en estimaciones con información estática en un contexto de corte transversal. Contrariamente, este capítulo examina el comportamiento de sucesos dinámicos mediante una introducción a series de tiempo, exponiendo análisis estadísticos y econométricos para datos económicos recolectados periódicamente (segundo a segundo, minuto a minuto, hora a hora, diaria 5 o 7 días, semanal, mensual, trimestral, semestral, anual, quinquenal entre otras).

Asimismo, se expone diferentes modelos y métodos para pronosticar tendencia de corto y largo plazo en variables macroeconómicas, empresariales o aquellas obtenidas a partir de datos históricos. Esto permite obtener resultados para anticipar posibles acontecimientos futuros desfavorables y mitigar la incertidumbre, sobre las variables económicas exploradas. De igual manera, ayudar a establecer políticas para regular consumo y producción en cualquier actividad, manejar sistemas de inversión y planificación macroeconómica.

De acuerdo con lo anterior, las siguientes secciones contienen algunos conceptos básicos que permiten comprender las diferentes metodologías sobre el análisis en series de tiempo. Donde, se destaca el filtro de Hodrick y Prescott, empleado con el fin de obtener la tendencia y ciclo de una serie; también técnicas econométricas para estimar tendencia incluyendo sus formas funcionales, modelos Box-Cox y predicción.

Adicionalmente, este capítulo incluye técnicas estadísticas para pronosticar variables temporales a través de promedios móviles y métodos de atenuación exponencial simple, doble y Holt-Winters (no estacional, aditivo o multiplicativo). Por último, se aplican las metodologías tratadas, mediante un estudio caso, con datos trimestrales del Producto Interno Bruto (PIB) colombiano.

4.2 Conceptos básicos para Series de Tiempo

Una serie de tiempo (*ST*) es un conjunto de observaciones coleccionadas sucesiva y homogéneamente[35] para una misma variable en periodos específicos. Cada observación es denotada generalmente como Y_t, con $t = 1, 2, \ldots, T$ *(véase* ecuación 4.1). A diferencia de los datos estáticos (corte transversal), las series temporales posibilitan observar la evolución de una variable a lo largo del tiempo; permitiendo analizar su dinámica intertemporal y realizar correlaciones no contemporáneas, en distintos momentos, entre variables dinámicas.

$$Y_t = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_T \end{bmatrix}_{1\times T} \qquad (4.1)$$

A partir de lo anterior y una vez constituida una serie de tiempo, es posible definir un método para proporcionar pronósticos $\hat{Y}_{t+p}$[36] a partir de su propio pasado ($\hat{Y}_{t-1,}\hat{Y}_{t-2,}\cdots,\hat{Y}_{t-p}$)[37]; suponiendo que estos sucesos continuaran en el futuro. En otras palabras, el curso de $\hat{Y}_t$ y su predicción $\hat{Y}_{t+p}$ no se encuentran condicionados a variables independientes[38], porque en este caso también sería necesario predecirlas individualmente $\left(\hat{X}_{t+p,1}, \hat{X}_{t+p,2}, .., \hat{X}_{t+p,k}\right)$[39] para establecer o proyectar $\hat{Y}_{t+p}$. Una vez determinado el significado para una serie de tiempo y con el fin de proseguir con

---

[35] Quiere decir que la variable a estudiar, debe tener la misma periodicidad.

[36] Donde $\hat{Y}$ se refiere a los nuevos valores pronosticados y $t + p$ es el subíndice que representa los p periodos futuros proyectados.

[37] $Y = F(Y_{t-1,}Y_{t-2,}\cdots,Y_{t-p})$, donde $t - p$ es el subíndice que representa los p periodos rezagados de Y_t .

[38] $Y = F(\hat{X}_{t+p,1}, \hat{X}_{t+p,2}, .., \hat{X}_{t+p,k})$. Para todo el conjunto (*J*) de variables independientes con *J*=1,2,…,*k,* como los trabajados en capítulos anteriores de corte transversal, ante la imposibilidad de especificar un modelo estructural con variables exógenas; *véase* Pindyck y Rubinfeld (1998, 488).

[39] Donde $\hat{X}_{(t+p)k}$ se refiere a los nuevos valores pronosticados de cada variable independiente y $t + p$ es el subíndice que representa los p periodos futuros proyectados.

esta conceptualización, a continuación se encuentran componentes y naturaleza para cualquier variable dinámica.

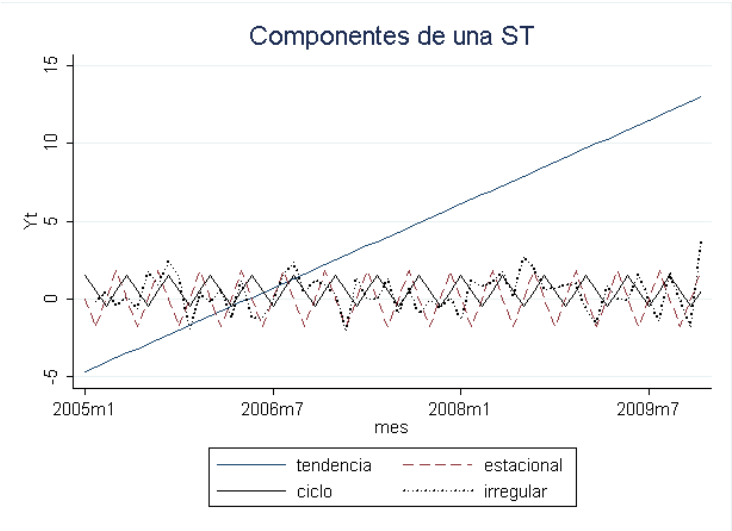
4.2.2 Componentes y naturaleza de una serie de Tiempo

Partiendo del concepto descrito para una serie de tiempo y con el fin de generarle pronóstico, se hace necesario conocer sus componentes y naturaleza. Los primeros corresponden a su tendencia, ciclo, elemento irregular y estacional; mientras la segunda hace referencia si es aditiva o multiplicativa. Estas especificaciones, son concebidas a partir del historial de la variable, porque con él se forma su trayectoria y partir de esta última, mediante una gráfica, se logra identificar las características mencionadas.

De esta manera, conocer el contexto de la variable dinámica ayuda a establecer los periodos donde ocurrieron cambios metodológicos para su medición, afectación de políticas y sucesos económicos coyunturales exógenos que posiblemente alteraron drásticamente el recorrido regular de la serie (cambios estructurales[40]). Esto, con el objetivo de establecer los momentos donde ocurrieron y así entender cómo pueden afectar su predicción $\left(Y_t \rightarrow \hat{Y}_{t+p}\right)$.

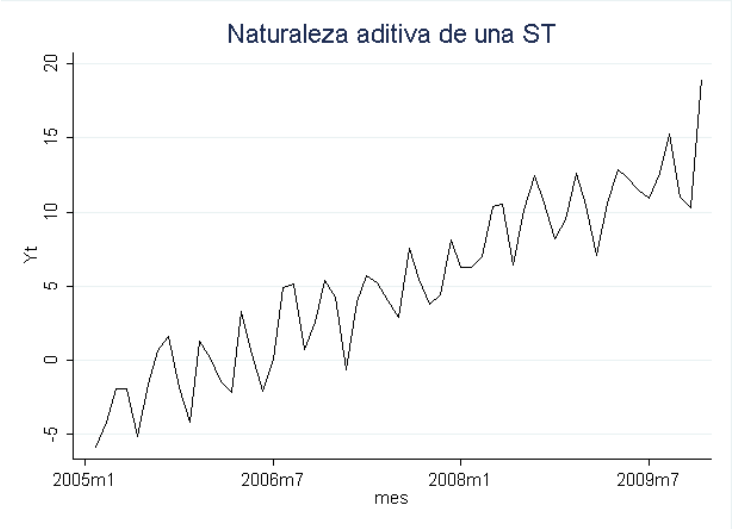
[40]Observaciones atípicas dentro del trayecto regular de una variable.

Gráfica 4.1 Componentes de una ST (Y_t).



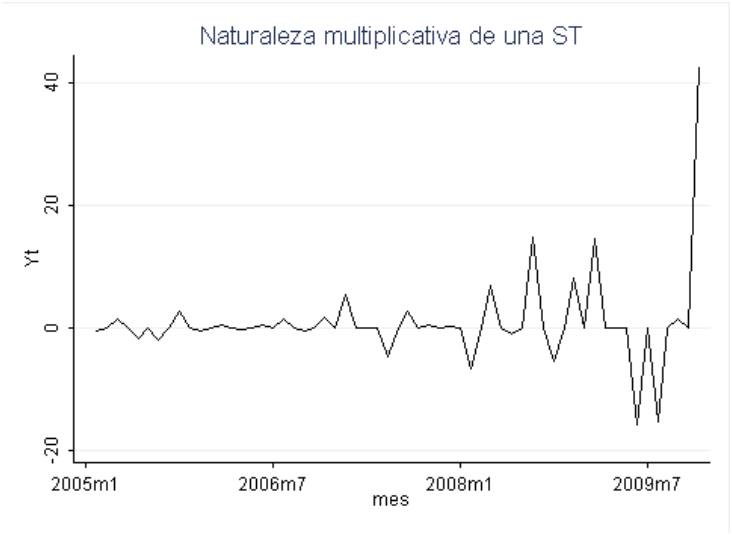
Fuente: cálculos autores, serie de tiempo hipotética con frecuencia mensual entre 2005 y 2009.

Gráfica 4.2. Naturaleza aditiva una ST hipotética



Fuente: cálculos autores

Gráfica 4.3. Naturaleza multiplicativa de una ST hipotética,



Fuente: cálculos autores

Para determinar lo anterior, en la gráfica 4.1 se observa la evolución sobre varias series de tiempo Y_t, generadas hipotéticamente en Stata®[41], cuyos comportamientos históricos establecen sus componentes: tendencia (T_t, línea azul), ciclo (C_t, línea negra), irregular (I_t, línea punteada negra) y estacional (S_t, línea punteada roja). Sin embrago, Y_t puede contener uno, algunos o combinación de todos sus elementos aditiva o multiplicativamente (*véase* gráfica 4.2 y 4.3).

$$Y_t = T_t + C_t + S_t + I_t \qquad (4.2)$$

$$Y_t = T_t * C_t * S_t * I_t \qquad (4.3)$$

La primera especificacion se caracteriza por tener cada componente de forma independiente lo que posibilita descomponer la serie en una suma de los cuatro factores (*véase* gráfica 4.2). La segunda, por otra parte, surge cuando la tendencia (T_t), ciclo (C_t), irregularidad (I_t) y estacionalidad (S_t) son dependientes entre sí;

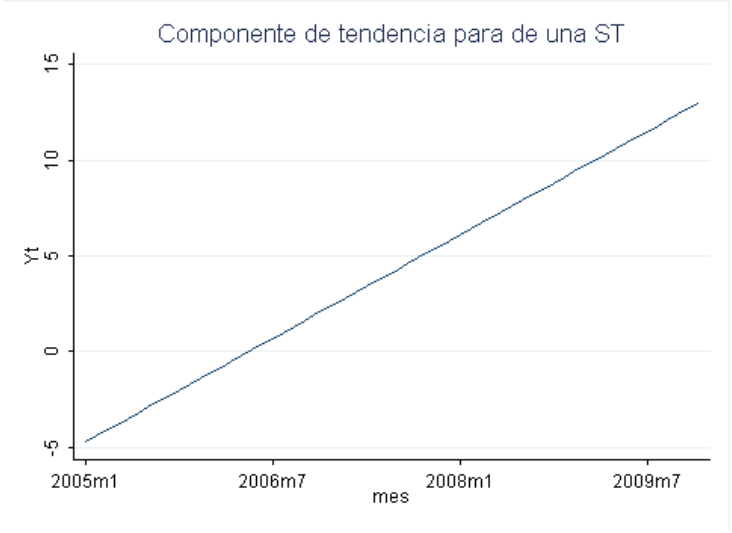
[41]Creadas en STATA 10.0 de la siguiente forma: $\text{tendencia}_t = -5 + 0{,}2T_t$, $ciclo_t = seno\left({}^{T_t\pi}/_{2}\right) + 0.5$ $\text{estacional}_t = 1{,}8 * seno\left({}^{T_t\pi}/_{2}\right)$ e $\text{irregular}_t = 0{,}8 * irregular_{t-1} + \varepsilon_t$; donde ε_t denota el componente estocástico (error) con media cero y varianza uno, T_t los valores de tendencia para el periodo t.

definidas por alta variabilidad (*véase* gráfica 4.3). A continuación se definen y caracterizan individualmente dichos componentes.

4.2.2.1 Tendencia

La tendencia (T_t), se define como el componente de baja frecuencia –o poca volatilidad- que evoluciona lentamente en alguna dirección particular. Cuya interpretación, se relaciona con el comportamiento de largo plazo para una serie y explica los cambios permanentes en sus valores promedios (media aritmética)[42]. Esta conducta, es reflejada comúnmente por las variables económicas como: población, producción, exportaciones e importaciones, entre otras, que presentan recorridos crecientes o decrecientes a lo largo del tiempo. Como ejemplo, en el gráfico 4.4 se observa una tendencia creciente a lo largo del tiempo, creada hipotéticamente en Stata®.

Gráfica 4.4. Componente de tendencia para una ST



Fuente: cálculos autores

Prosiguiendo, la tendencia de Y_t es generada partir de choques persistentes en ella y cambios dinámicos que recaen directamente sobre su comportamiento de largo plazo. Por ejemplo, la variable población suele crecer a un ritmo constante, dado

[42] *Véase* Montenegro (2007, cap 1).

que cada individuo tiene la posibilidad de procrear hijos. Sin embargo, durante periodos bélicos puede cambiar el rumbo, como consecuencia de las muertes ocasionadas, alejándose de su dinámica normal adquirida hacia el largo plazo. Lo anterior no solo sucede para variables formadas naturalmente, sino para las calculadas artificialmente[43], dada la existencia de elementos observables empleados en su estimación que pueden determinar su dirección.

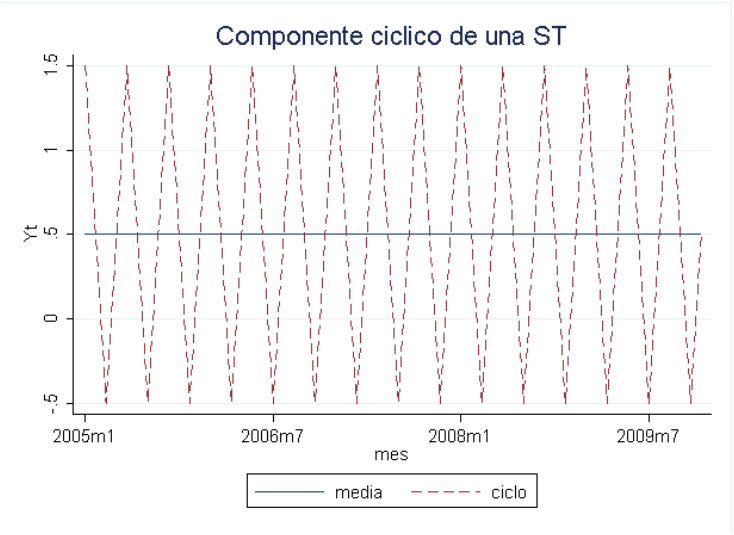
Para el caso del Producto Interno Bruto (PIB), usualmente conserva una tendencia creciente en el tiempo. Esto ocurre, porque sus componentes –consumo, inversión y gasto público – llevan el mismo comportamiento de la población. Ante esto, es posible dividir el PIB sobre la población (PIB per cápita), con el fin de remover la tendencia derivada del efecto poblacional y conocer si el PIB contiene o no este elemento. Análogamente ocurre con el crecimiento de los precios; la serie resultante es el PIB real[44].

4.2.2.2 Ciclo

El segundo competente, es el ciclo que corresponde a una oscilación de largo plazo alrededor de una media (*véase* gráfica 4.5) o tendencia (*véase* gráfica 4.6). Para el primer caso, este movimiento se define por no tener implicaciones permanentes sobre el promedio de Y_t, mientras en el segundo ocurre lo contrario. Sin embargo en ambos, la inclinación ondular siempre logra sus picos máximos por encima y los mínimos abajo del promedio o tendencia respectivamente.

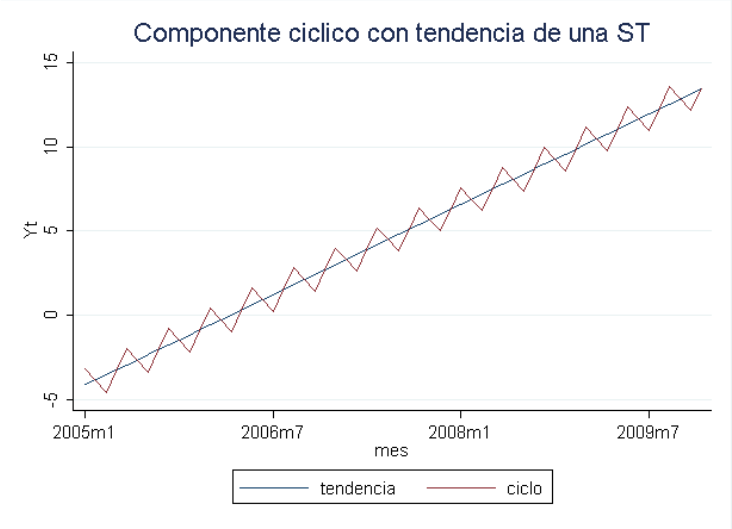
[43] Como el producto interno bruto (PIB), desempleo, tasa de interés, inflación, entre otras.
[44] *Véase* Granger (1993, cap 2).

Gráfica 4.5. Componente cíclico de una ST, alrededor de una media



Fuente: cálculos autores

Gráfica 4.6. Componente cíclico de una ST, alrededor de una tendencia



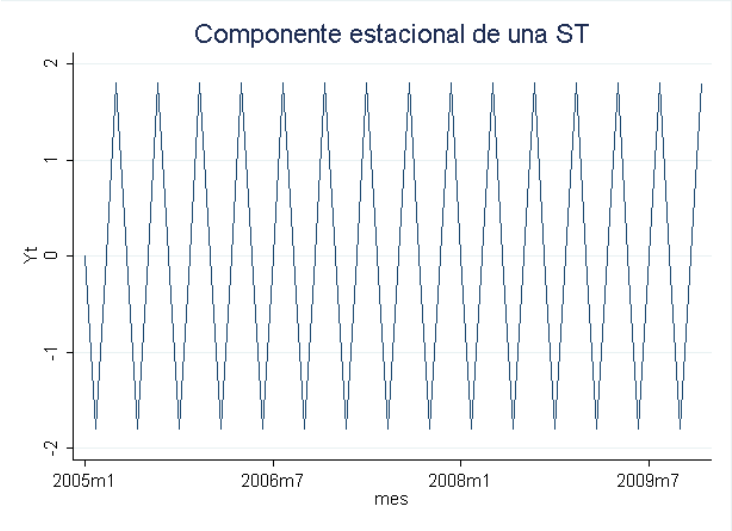
Fuente: cálculos autores

También el ciclo se caracteriza, porque su amplitud cíclica en variables económicas no es definida claramente; dado que su frecuencia puede suceder de manera heterogénea como respuesta a los acontecimientos o fenómenos económicos. Adjuntamente, además de la tendencia, el ciclo puede incluir componente estacional e irregular discutidos a continuación.

4.2.2.3 Componente Estacional

El componente estacional corresponde a los movimientos de una variable sucedidos reiteradamente durante una frecuencia homogénea de tiempo, para series de tiempo cuya periodicidad corresponde es diaria, semanal, mensual, trimestral o semestral. Este elemento se caracterizada por aparecer en un periodo y desvanecerse en el siguiente (*véase* gráfica 4.7).

Gráfica 4.7. Componente estacional de una ST



Fuente: cálculos autores

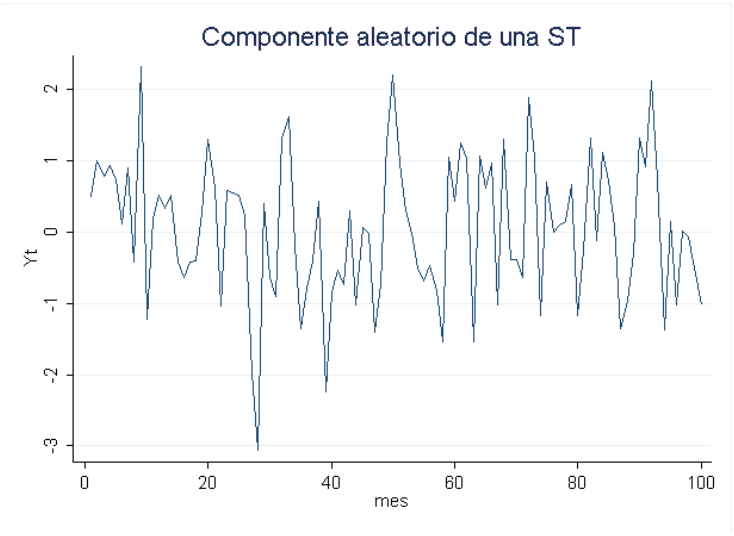
Aunque una serie puede contener movimientos cíclicos y estacionales, gráficamente se puede distinguir entre ambos, porque en el primero sus picos máximos y mínimos pueden presentarse en distintos periodos, mientras en el segundo acontecen con el mismo patrón. En otras palabras, la estacionalidad son oscilaciones de corto plazo y baja persistencia, que ocurren nuevamente después de un lapso equivalente de tiempo[45]. Como muestra, las ventas de bebidas calientes aumentan en los meses de invierno y disminuyen en verano.

[45] *Véase* Makridakis y Wheelwright (1978, cap 1).

4.2.2.4 Componente Irregular

Finalmente Y_t puede contener un componente irregular debido a fenómenos externos impredecibles coyunturales, de índole natural o económica[46]. A diferencia, de los componentes previamente mencionados, este comportamiento irregular no tiene forma definida y sus movimientos son desiguales e impredecibles en el tiempo (*véase* gráfica 4.8).

Gráfica 4.8. Componente irregular de una ST



Fuente: cálculos autores

A diferencia de tendencia, ciclo y estacionalidad, la conducta irregular de una variable no es considerada dentro de los componentes derteminísticos para Y_t, como si los demás, debido a su naturaleza aleatoria[47]. Por otra parte, a continuación se presentan los procedimientos y tratamientos para una variable dinámica; de acuerdo con sus componentes y naturaleza, con el fin de estimar su pronóstico.

[46] Como climáticos, guerras, catástrofes, políticas públicas transitorias, choques especulativos bursátiles, entre otros.
[47] Totalmente al azar, *véase* capítulo 5.

4.3 Filtro Hodrick -Prescott

A partir de los componentes para Y_t y con el fin de tratar, extraer y separar su elemento tendencial y cíclico; esta sección exponre el filtro Hodrick-Prescott (*H-P*)[48]. Este método consiste en obtener una serie suavizada S_t a partir de la original Y_t, mediante una solución al problema de optimización plasmado en la ecuación 4.4. Una vez resuelto, permite estimar tanto el ciclo como la tendencia de la serie.

$$min\sum_{t=1}^{n}\left(Y_t - S_t\right)^2 + \lambda\sum_{t=2}^{n-1}\left[\left(S_{t+1} - S_t\right) - \left(S_t - S_{t-1}\right)\right]^2 \quad (4.4)$$

La ecuación 4.4 se refiere a la suma mínima de la varianza para Y_t alrededor de S_t ($\sum_{t=1}^{n}\left(Y_t - S_t\right)^2$), adicionando la diferencia entre los pares cercanos[49] de S_t ($\sum_{t=2}^{n-1}\left[\left(S_{t+1} - S_t\right) - \left(S_t - S_{t-1}\right)\right]^2$); multiplicado por el parámetro de suavizamiento λ, que representa el grado de atenuación de la nueva serie. Entre mayor valor tome este parámetro, la variable resultante S_t será más cercana a una tendencia lineal $\left(Y_t = \beta_0 + \beta_1 T_t\right)$; caso contrario, la serie suavizada equivaldrá a la original ($S_t = Y_t$).

Los valores sugeridos para λ dependen de la periodicidad de Y_t, y son: 14400 (mensual), 1600 (trimestral) y 100 (anual);. Por otra parte, una vez obtenido el componente cíclico $\left(Y_t - S_t\right)$, a partir del filtro de Hodrick-Prescott, puede ser interpretada como la brecha existente entre su valor real (Y_t) y potencial (S_t).

El filtro de H-P no es una técnica para proyectar Y_t hacia futuro; por esto en el numeral 4.4, se presentan algunos modelos para pronosticar el valor de esta serie cuando contiene tendencia.

[48] *Véase* Montenegro (2007, cap. 1).

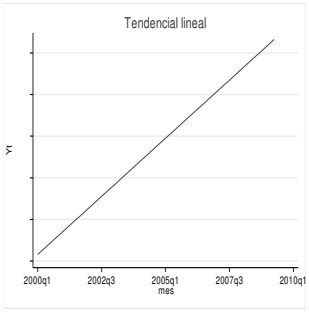
[49] En este caso se quiere que el dato de un periodo sea muy cercano al del periodo siguiente.

4.4 Modelos de pronósticos con tendencia determinística

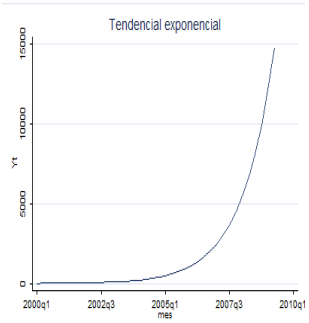
De acuerdo con lo señalado en la sección 4.2.2, gráficamente es posible identificar la tendencia para una serie de tiempo y desde ella definir su forma funcional[50] (*véase* gráfica 4.9). A partir de esta comportamiento tendencial de Y_t, es posible conformar un modelo a partir de una parte determinística ($\beta_{0,}\ \beta_1 T_{t1,}\cdots,\ \beta_k T_{tk}$) y otra estocástica o aleatoria (e_t); combinadas econométricamente (*véase* ecuación 4.5).

Gráfica 4.9 Formas funcionales para modelos deterministas con tendencia

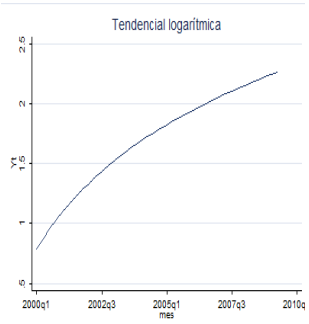
(A)Lineal



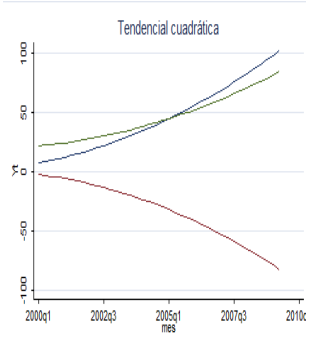
(B) Exponencial



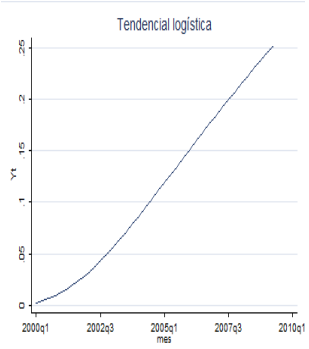
(C) Logarítmico



(D) Cuadrática



(E) Logística



Fuente: cálculos autores.

$$Y_t = F\left(\beta_{0,}\ \beta_1 T_{t1,}\cdots,\ \beta_k T_{tk,}\ \varepsilon_t\right) \quad (4.5)$$

[50] Dependiendo de su trayectoria puede tener una forma funcional implícita lineal, exponencial, logarítmica, cuadrática o logística.

Específicamente, la ecuación 4.5 define a Y_t como el resultado de una combinación entre la reacción constante $(\beta_0, \beta_1, .., \beta_k)$ a lo largo del tiempo y la parte impredecible (e_t) acumulada periodo a periodo. Las funciones implícitas para la tendencia de Y_t están supuestas en la ecuación 4.5 y desagregadas en el cuadro 4.1; ellas, excluyendo la forma Box-Cox, son estimables mediante modelos de regresión por mínimos cuadrados ordinarios (MCO).

Sin embargo, si gráficamente no se logra preestablecer el comportamiento para la tendencia, excluyendo la forma Box-Cox, debe estimarse cada relación expuesta en el cuadro 4.1 y seleccionar la más adecuada; acorde con los resultados de la evaluación para cada modelo, teniendo en cuenta:

1. Signos de los parámetros $(\beta_0, \beta_1, .., \beta_k)$ coherentes con los esperados previamente.
2. Significancia parcial de los coeficientes
3. Significancia global del modelo.
4. Coeficiente de determinación ajustado ($\bar{R}^2$), el valor más cercano a uno.
5. Criterios de Akaike (CA)[51] y Schwarz (CS)[52] más pequeños.
6. El término estocástico e_t distribuido de manera normal.
7. Ausencia de autocorrelación entre los errores, ausencia de multicolinealidad, heteroscedasticidad, endogeneidad y cambio estructural.

[51] $CA = e^{2k/n} \frac{\sum u_t^2}{n} \equiv LNCA = \left(\frac{2k}{n}\right) + LN\left(\frac{\sum u_t^2}{n}\right)$, k se refiere al número de estimadores $(\beta_0, \beta_1, .., \beta_k)$, *LN* es logaritmo natural y n al tamaño de la muestra.

[52] $CS = n^{k/n} \frac{\sum u_t^2}{n} \equiv LNCS = \left(\frac{k}{n}\right) LNn + LN\left(\frac{\sum u_t^2}{n}\right)$, k se refiere al número de estimadores $(\beta_0, \beta_1, .., \beta_k)$, *LN* es logaritmo natural y n al tamaño de la muestra.

Cuadro 4.1 Formas funcionales para pronósticos con modelos de tendencia determinísticas

Función de tendencia	Forma del modelo
Lineal	$Y_t = \beta_0 + \beta_1 T_t + u_t$
Logarítmico; en Y_t (LNY_t) y T_t (LNT_t)	$LNY_t = LN\beta_0 + \beta_1 LNT_t + LNu_t$
Exponencial o log-lineal; logaritmo en Y_t (LNY_t)	$LNY_t = \beta_0 + \beta_1 T_t + u_t$
Lineal-logarítmico; logaritmo en T_t (LNT_t)	$Y_t = LN\beta_0 + \beta_1 LNT_t + LNu_t$
Reciproco en T_t	$Y_t = \beta_0 \pm \frac{\beta_1}{t} + u_t$
Reciproco en Y_t	$\frac{1}{Y_t} = \beta_0 + \beta_1 t + u_t$
Reciproco en Y_t y T_t	$\frac{1}{Y_t} = \beta_0 \pm \frac{\beta_1}{t} + u_t$
Autorregresivo; Y_{t-1} rezagado un periodo de Y_t	$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$
Autorregresivo-logarítmico; LNY_{t-1} rezagado un periodo de LNY_t	$LNY_t = \beta_0 + \beta_1 LNY_{t-1} + u_t$
Cuadrático	$Y_t = \beta_0 \pm \beta_1 T_t \pm \beta_2 T_t^2 + u_t$
Cúbica	$Y_t = \beta_0 \pm \beta_1 T_t \pm \beta_2 T_t^2 \pm \beta_3 T_t^3 + u_t$
Logístico; logaritmo en Y_t (LNY_t)	$LNY_t = \beta_0 - \frac{\beta_1}{t} + u_t;\ \ 0 < \beta_1 < 1;\quad \beta_1 > 1$
Box-Cox	$Y_t^\theta = \beta_0 + \beta_1 T_t^\lambda + u_t$

Fuente: autores, a partir de Mendieta y Perdomo (2007) y Pindyck y Rubinfeld (1998).

Un segundo método consiste en establecer la función implícita a partir del modelo Box-Cox (*véase* cuadro 4.1 y ecuación 4.6). Esto es posible mediante máxima verosimilitud (MV, *véase* capítulo 3), el cual otorga valores para los parámetros de transformación theta y lambda[53] (θ y λ) en Box-Cox. Determinando de esta manera, si la tendencia se ajusta a cualquier forma explícita en el cuadro 4.1 o si es desconocida (denominada Box-Cox).

$$Y_t^\theta = \beta_0 + \beta_1 t^\lambda + \varepsilon_t \quad (4.6)$$

En la ecuación 4.6, $Y_t^\theta = \frac{Y_t^\theta - 1}{\theta}$ y $t^\lambda = \frac{t^\lambda - 1}{\lambda}$ representan una función no lineal para el modelo Box-Cox no restringido (nr), motivo por el cual los parámetros de transformación (θ y λ) y coeficientes (β_0 y β_1) deben originarse a partir de MV, formulada en la ecuación 4.7. Donde, l_{nr} se refiere al logaritmo de la función de

[53] Este coeficiente es distinto al expuesto, en la sección 4.3, sobre atenuación para el filtro Hodrick-Prescott.

verosimilitud[54] no restringida, n representa el tamaño de la muestra, σ^2 la varianza del modelo y π la constante que hace alusión al número pi (3,1416).

$$l_{nr}(\sigma^2, \theta, \lambda, \beta_0, \beta_1) = -\frac{n}{2}LN(2\pi) - \frac{n}{2}LN(\sigma^2) + (\lambda - 1)\sum_{t=1}^{n} LNY_t - \frac{1}{2\sigma^2}\sum_{t=1}^{n}\left(Y_t^{\theta} - \hat{\beta}_0 - \hat{\beta}_1 t^{\lambda}\right)^2 \quad (4.7)$$

Una vez estimados θ y λ, se puede establecer la función de tendencia adecuada subyacente para Y_t (*véase* cuadro 4.2) y con ella realizar su respectivo pronóstico $\left(\hat{Y}_{t+p}\right)$.

Cuadro 4.2 Valores de los parámetros de transformación para determinar la forma de tendencia

Función de tendencia	Condición para θ y λ	Método de estimación
Lineal	$\theta = \lambda = 1$	Mínimos cuadrados ordinarios
Logarítmico	$\theta = \lambda = 0$	Mínimos cuadrados ordinarios
Exponencial o Log-Lineal	$\theta = 1\ ;\ \lambda = 0$	Mínimos cuadrados ordinarios
Lineal-Logarítmico	$\theta = 0\ ;\ \lambda = 1$	Mínimos cuadrados ordinarios
Reciproco en T_t	$\theta = 1\ ;\ \lambda = -1$	Mínimos cuadrados ordinarios
Reciproco en Y_t	$\theta = -1\ ;\ \lambda = 1$	Mínimos cuadrados ordinarios
Reciproco en Y_t y T_t	$\theta = -1\ ;\ \lambda = -1$	Mínimos cuadrados ordinarios
Box-Cox Restringida I	$\theta = \lambda \neq 0$	Máxima verosimilitud
Box-Cox Restringida II	$\theta \neq \lambda \neq 0$	Máxima verosimilitud

Fuente: autores, a partir de Mendieta y Perdomo (2007).

Continuando el proceso selectivo de la forma funcional adecuada para la tendencia de Y_t, anidada Box-Cox, con la prueba de razón de verosimilitud[55] (RV) se elige la más adecuada (*véase* ecuación 4.8). En la ecuación 4.8 l_r significa el logaritmo de la función de verosimilitud restringida a los valores de los parámetros θ y λ del cuadro 4.2 y l_{nr}, la anterior función no restringida (Box-Cox, *véase* ecuación 4.7). RV sigue una distribución chi-cuadrado, χ_2^2, con 2 grados de libertad; refiriéndose al número de condiciones impuestas sobre θ y λ en el modelo Box-Cox (*véase* cuadro 4.2).

$$RV = 2(l_{nr} - l_r) \sim \chi_2^2 \qquad (4.8)$$

[54] *Véase* Greene (1998, 420).
[55] *Véase* Greene (1998, 422).

A pesar del tratamiento anterior, sobre formas funcionales, generalmente los pronósticos concebidos con los modelos determinísticos de tendencia no son tan precisos; porque cuando se estiman las funciones los modelos incumplen algunos supuestos de MCO como: ausencia de autocorrelación en el error y homoscedasticidad; implicando estimadores ineficientes que vulneran la condición de Gauss Markov (MELI).

Adicional a esto, las proyecciones se adhieren al comportamiento de la tendencia y desconocen los otros elementos relevantes de Y_t (ciclo, componente irregular y estacional). Por esta razón, en la sección 4.5 se discuten las metodologías de atenuación exponencial; con el fin de incorporar al pronóstico $(\hat{Y}_{t+p})$ los otros componentes de la serie de tiempo, aunque esta técnica no ayuda aliviar la vulnerabilidad de Gauss Markov ocasionada en la estimación mediante MCO para los modelos determinísticos de tendencia abarcados.

4.5 Pronóstico con métodos de atenuación exponencial

Las técnicas de suavizamiento exponencial emplean la información histórica $(Y_{t-1}, Y_{t-2}, .., Y_{t-p})$ de Y_t para obtener sus pronósticos $(\hat{Y}_{t+p})$. Se aplican generalmente cuando se cuenta con muestras pequeñas (mínimo diez o menos de 30 datos) y pretendiendo proyectar escenarios de largo plazo (más de tres periodos). Éstas son prácticas de tanteo no paramétricas, sin especificación previa de alguna forma funcional o modelos, como los desarrollados en la sección anterior; donde fue necesario aplicar algún método de estimación (MCO o MV) para encontrar sus parámetros.

En este caso la información de Y_t, para estimar sus valores futuros $(\hat{Y}_{t+1}, \hat{Y}_{t+2}, .., \hat{Y}_{t+p})$[56], no está relacionada a la tendencia (T_t) sino con su propio pasado $(Y_{t-1}, Y_{t-2}, .., Y_{t-p})$. Adicionalmente, bajo esta metodología, existen distintos métodos de atenuación (*véase* cuadro 4.4) entre los que se destacan: promedio móvil, promedio móvil doble, atenuación simple, atenuación doble o Holt-Winters (no estacional, aditivo o multiplicativo). Los cuales, se aplican según la naturaleza

[56]Donde $\hat{Y}_{t+p}$ se refiere a los nuevos valores pronosticados de Y_t y $t+p$ es el subíndice que representa sus p periodos futuros proyectados.

de Y_t (tendencia, estacionaria[57], cíclica, estacionales, irregulares o combinación de estas).

[57] Su media aritmética y varianza no están condicionadas (relacionadas) con el tiempo, en otras palabras su promedio es constante y tiene poca variabilidad (varianza constante). Es equivalente a una serie cíclica sin componte irregular, estacionalidad y tendencia; para más detalles *véase* capítulo 5.

Cuadro 4.3. Técnicas de atenuación exponencial.

Nombre de la Técnica	Naturaleza de la Variable	Expresión	Restricción	Equivalencia	Anotación de caso
Promedios móviles (PM)	Estacionaria	$\hat{Y}_{t+1} = \frac{(Y_t + Y_{t-1} + Y_{t-2} + .. + Y_{t-n+1})}{n}$	-	n: número de periodos en el PM	Mayor ponderación a los datos recientes
Promedios móviles doble (PMD)	Tendencia lineal	$\hat{Y}_{t+p} = a_t + b_t p$ $M_t \equiv \hat{Y}_{t+1} = \frac{(Y_t + Y_{t-1} + Y_{t-2} + .. + Y_{t-n+1})}{n}$ $M'_t = \frac{(M_t + M_{t-1} + M_{t-2} + .. + M_{t-n+1})}{n}$	-	$a_t = 2M_t - M'_t$ $b_t = \frac{2}{n-1}(M_t - M'_t)$ p: número de periodos a pronosticar	Mayor ponderación a los datos recientes
Atenuación simple (AS)	Estacionaria	$\hat{Y}_{t+1} = \alpha Y_t + (1-\alpha)\hat{Y}_t$	$0 \leq \alpha \leq 1$	$\hat{Y}_t = \alpha Y_{t-1} + (1-\alpha)\hat{Y}_{t-1}$	$\hat{Y}_t$: corresponde al valor suavizado en t-1
Atenuación doble (AD)	Tendencia lineal	$\hat{Y}_{t+p} = a_t + b_t p$ $A_t = \alpha Y_t + (1-\alpha)A_{t-1}$ $A'_t = \alpha A_t + (1-\alpha)A'_{t-1}$	$0 \leq \alpha \leq 1$	$a_t = 2A_t - A'_t$ $b_t = \frac{\alpha}{1-\alpha}(A_t - A'_t)$	A_t : corresponde al valor suavizado de Y_t A'_t: corresponde al valor doblemente suavizado de Y_t
Holt-Winters no estacional	Tendencia lineal y cíclica	$\hat{Y}_{t+p} = A_t + pT_t$ $A_t = \alpha Y_t + (1-\alpha)(A_{t-1} + T_{t-1})$ $T_t = \beta(A_t - A_{t-1}) + (1-\beta)T_{t-1}$	$0 \leq \alpha \leq 1$ $0 \leq \beta \leq 1$	T_t: estimación de tendencia p: número de periodos a pronosticar	A_t : corresponde al valor suavizado de Y_t
Holt-Winters estacional multiplicativo	Estacional con tendencia y ciclo	$\hat{Y}_{t+p} = (A_t + pT_t)S_{t-L+p}$ $A_t = \alpha\frac{Y_t}{S_{t-L}} + (1-\alpha)(A_{t-1} + T_{t-1})$ $T_t = \beta(A_t - A_{t-1}) + (1-\beta)T_{t-1}$ $S_t = \gamma\frac{Y_t}{A_t} + (1-\gamma)S_{t-L}$	$0 \leq \alpha \leq 1$ $0 \leq \beta \leq 1$ $0 \leq \gamma \leq 1$	S_t: estimación de la estacionalidad L: longitud de la estacionalidad	
Holt-Winters estacional Aditivo	Estacional con tendencia y ciclo	$\hat{Y}_{t+p} = (A_t + pT_t) + S_{t-L+p}$			

Fuente: autores, a partir de Mendieta y Perdomo (2008, Hanke y Reitsch).

Para cada caso expuesto en el cuadro 4.3[58] α, β y γ son coeficientes de suavizamiento que toman valores arbitrarios entre cero y uno; sin embargo para seleccionar el más adecuado, dentro de este rango, se cuenta con distintos indicadores de error para el pronóstico (*véase* cuadro 4.4). Los cuales, se derivan de la diferencia entre valor observado Y_t y pronosticado $\hat{Y}_{t+p}$ $\left(u_t = Y_t - \hat{Y}_{t+p}\right)$, teniendo en cuenta el tamaño de la muestra (n).

De esta manera, los valores entre cero y uno para α, β y γ serán aquellos que garantice el mínimo indicador de error para el pronóstico, entre estos se destacan: el promedio del valor absoluto del error (PVAE), promedio del error al cuadrado (PEC), porcentaje del promedio del valor absoluto del error (PPVAE), raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC) y coeficiente de Theil (CT), *véase* cuadro 4.4.

Cuadro 4.4. Indicadores de error para el pronóstico.

Nombre	Sigla	Expresión
Promedio del valor absoluto del error	PVAE	$PVAE = \frac{\sum_{t=1}^{n} \lvert u_t \rvert}{n}$
Promedio del error al cuadrado	PEC	$PEC = \frac{\sum_{t=1}^{n} u_t^2}{n}$
Porcentaje del promedio del valor absoluto del error	PPVAE	$PPVAE = \sum_{t=1}^{n} \left\lvert \frac{u_t}{Y_t} \right\rvert / n$
Raíz cuadrada del promedio para la suma de errores al cuadrado	RCPSEC	$RCPSEC = \sqrt{\frac{\sum_{t=1}^{n} u_t^2}{n}}$
Coeficiente de Theil	CT	$CT = \frac{\sqrt{\frac{\sum_{t=1}^{n} u_t^2}{n}}}{\sqrt{\frac{\sum_{t=1}^{n} Y_t}{n}} + \sqrt{\frac{\sum_{t=1}^{n} \hat{Y}_t}{n}}}$

Fuente: autores, a partir de Hanke y Reitsch (1996).

[58] Para conocer más detalles sobre conceptualización, características y formulación de los métodos en el cuadro 4.4, *véase* Makridakis y Wheelwright (1978) y Hanke y Reitsch (1996, caps 4 y 5).

Para finalizar y una vez expuestos los métodos de suavizamiento exponencial e indicadores de error para evaluar el pronóstico obtenido mediante el mismo. En la siguiente sección, con información trimestral del PIB colombiano, se encuentra un estudio de caso aplicando y concluyendo cada técnica expuesta en el capítulo (filtro de Hodrick-Prescott y métodos de pronóstico mediante modelos de tendencia determinística y atenuación exponencial).

4.6 Estudio de caso: el producto interno bruto (PIB) colombiano.

Continuando con el desarrollo y aplicación del filtro Hodrick-Prescott (*H-P*) y métodos de pronóstico, usando modelos de tendencia determinística y atenuación exponencial, en estas sección se trabaja con información trimestral (entre 2000-I y 2009-I) sobre la serie de tiempo PIB en millones de pesos; desestacionalizada a precios constantes del año 2000 (*véase* cuadro 4.5 en el anexo 4.1). Variable construida trimestralmente por el Departamento Administrativo Nacional de Estadística (DANE), entidad encargada de recolectar, consolidar y publicar información sobre las cuentas nacionales colombianas desde 1970[59]. Conforme a lo señalado en este capítulo PIB_t equivale a Y_t ($PIB_t = Y_t$).

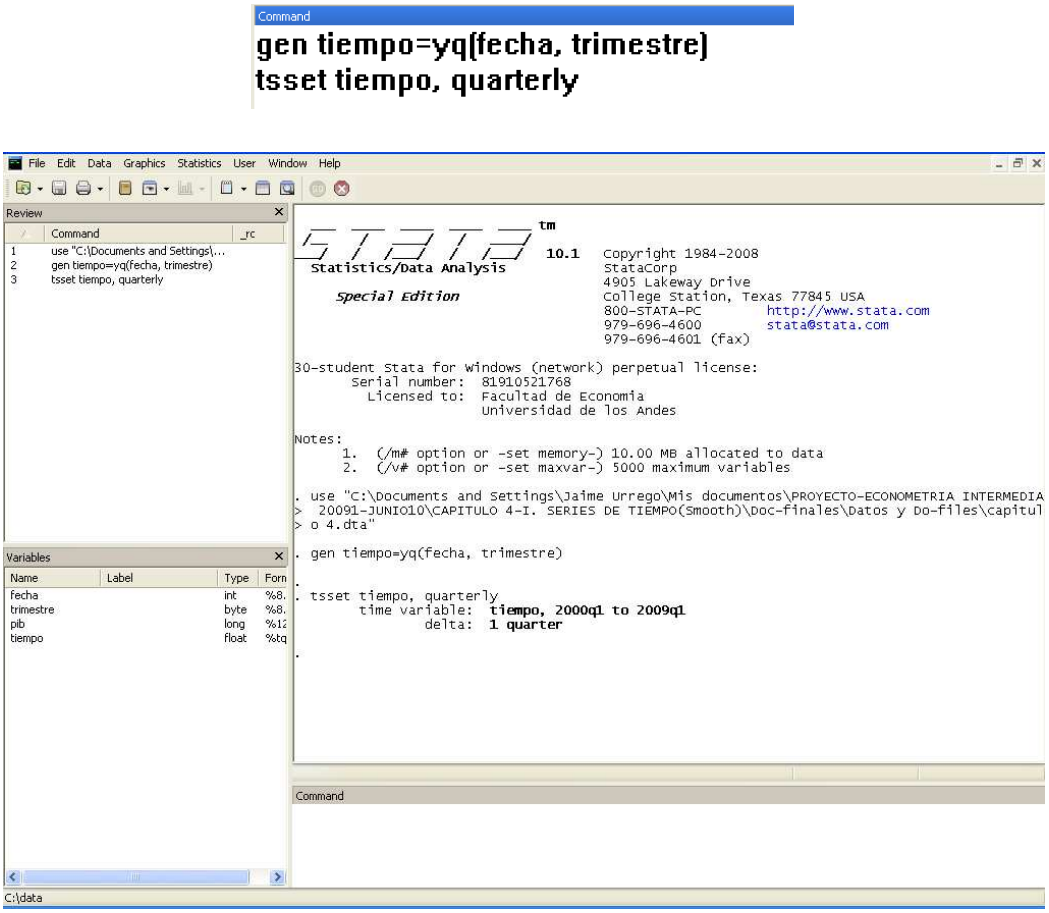
4.6.1 Filtro Hodrick-Prescott

De acuerdo a la periodicidad trimestral del PIB, se cuenta con 37 datos ($n = 37$) como tamaño de muestra; el contenido de esta información se puede apreciar en el cuadro 4.7 del anexo 4.1, mediante la cual se pretende separar la tendencia y encontrar el componente cíclico del PIB colombiano, entre el primer trimestre de 2000 y 2009 (2000-I y 2009-I), empleando el filtro *H-P*. Una vez los datos del PIB están cargados en Stata®, en este programa se realiza paso a paso la aplicación del filtro *H-P* de la siguiente manera:

1- Configurar el programa, para que reconozca el PIB_t como serie de tiempo trimestral, con el comando *gen* y *tsset* (*véase* figura 4.1). En caso de contar con frecuencias temporales distintas, en el cuadro 4.9 del anexo 4.1 se encuentra la programación en Stata® para su debida manipulación.

[59] Departamento Administrativo Nacional de Estadística. Ficha Metodológica Cuentas Nacionales Anuales.

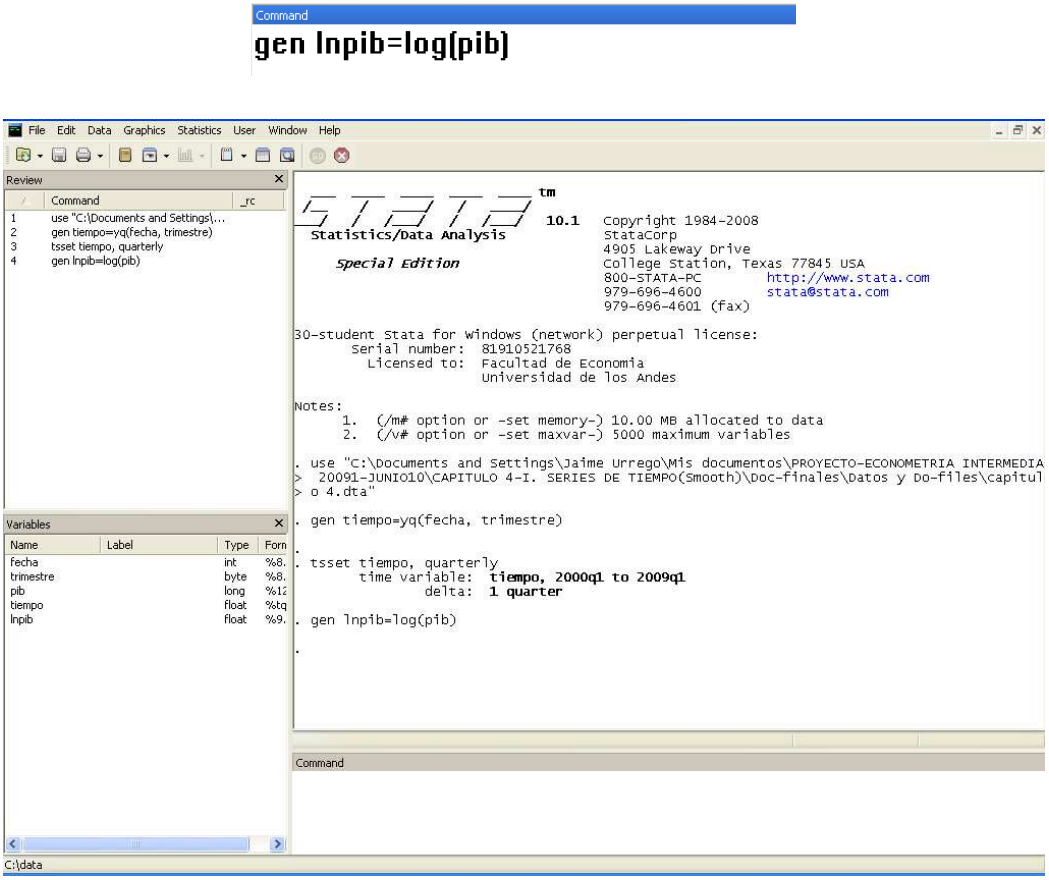
Figura 4.1. Salida de Stata® para especificar una variable como serie de tiempo



Fuente: cálculos autores.

2- Generar una nueva variable, con el comando *gen*, que contenga el logaritmo natural de PIB ($LNPIB_t$) (*véase* figura 4.2).

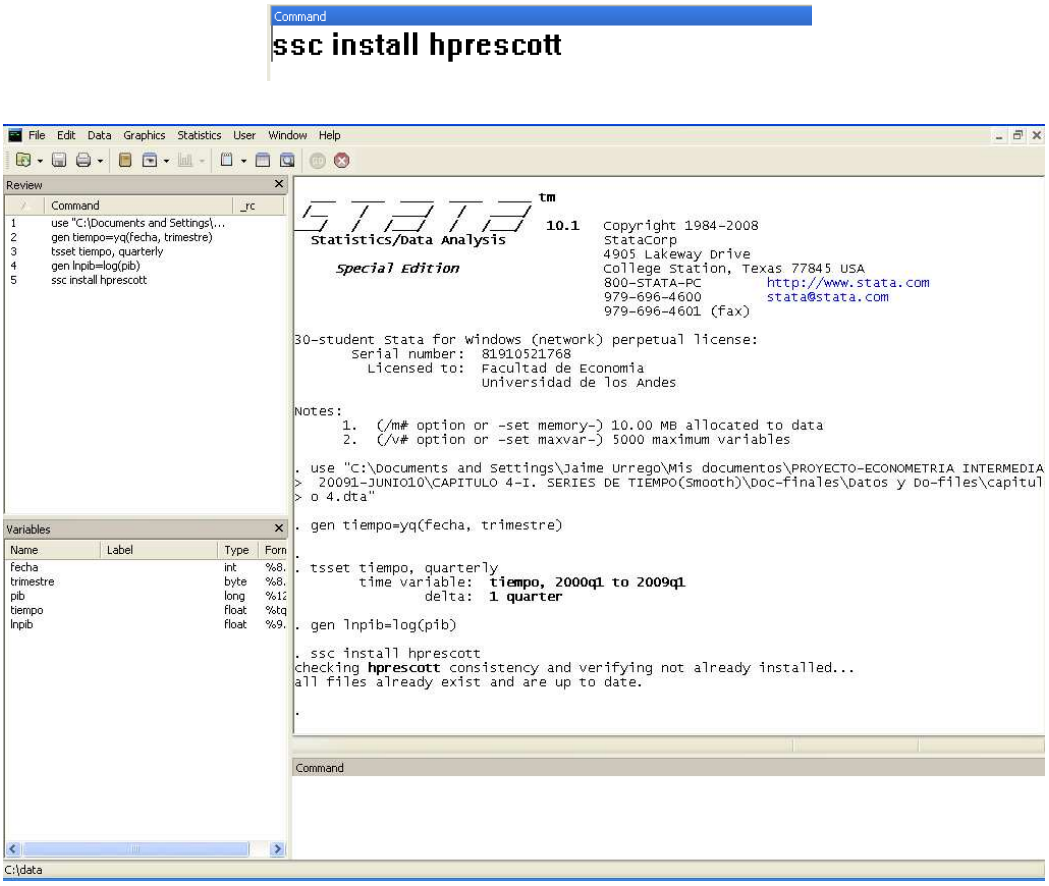
Figura 4.2. Salida de Stata® para crear una variable como logaritmo natural



Fuente: cálculos autores.

3- Descargar e instalar el filtro de Hodrick y Prescott en Stata®, mediante el comando *ssc install hprescott* (*véase* figura 4.3).

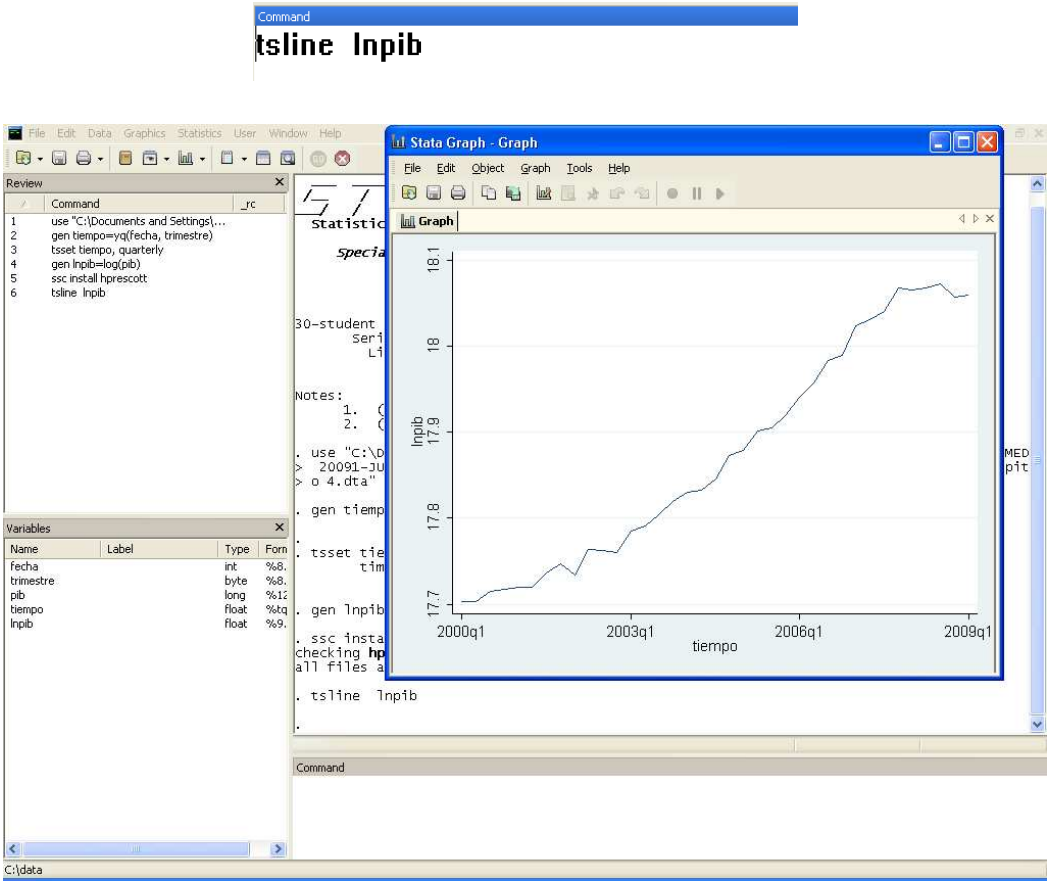
Figura 4.3. Instalación del filtro *H-P*



Fuente: cálculos autores.

4- Graficar el comportamiento del $LNPIB_t$ a través del tiempo, con el comando *tsline,* para conocer sus componentes y naturaleza (*véase* figura 4.4).

Figura 4.4. Salida de Stata® para graficar una serie de tiempo



Fuente: cálculos autores.

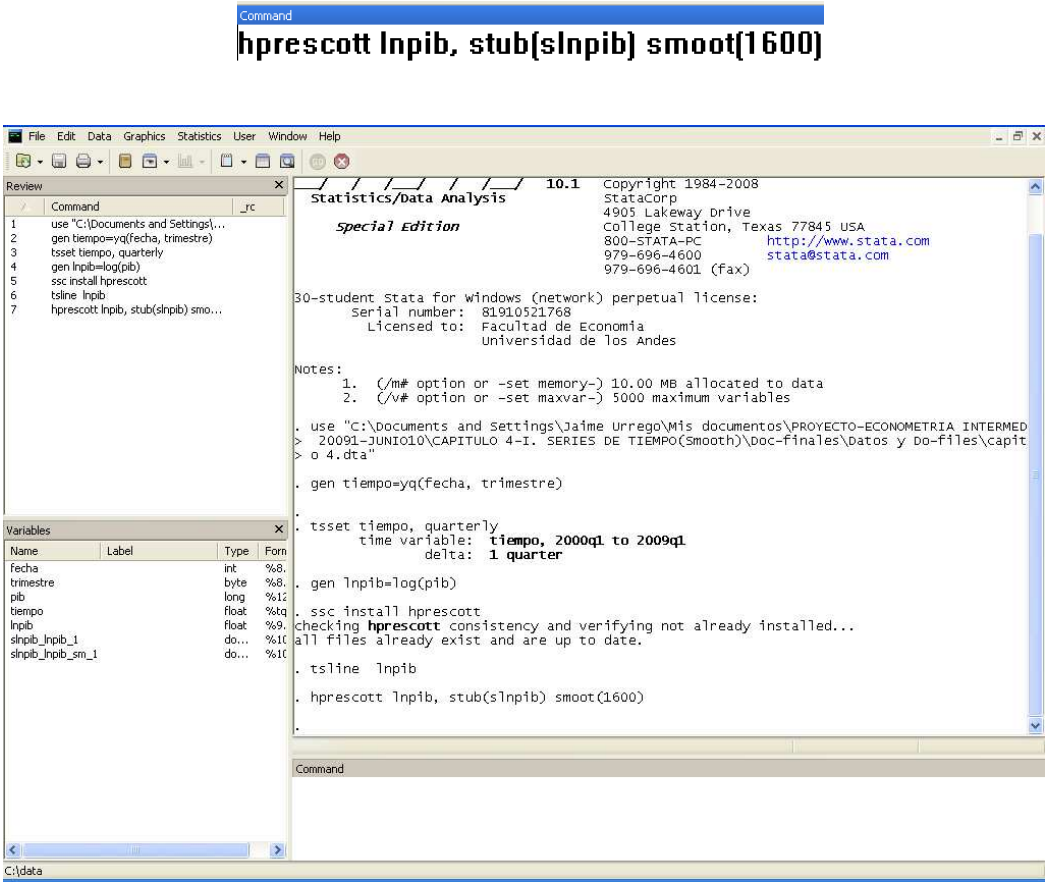
En este caso y acorde con lo señalado en la sección 4.2.2, para el $LNPIB_t$ se observa la combinación entre una tendencia creciente y un componente irregular (*véase* figura 4.4), sin componente cíclico y estacional. Consecuencia de lo anterior, la variable cuestionada presenta media y varianza inestables entre 2000-I y 2009-I; dada la presencia tendencial en ella.

Adicionalmente, su trayectoria normal (creciente) cambia a finales de 2008 y principios de 2009; tomando un curso decreciente (desaceleración). Debido a la posible influencia de la crisis financiera internacional, desatada en estos

periodos, sobre la economía colombiana. Fenómeno económico que exterioriza el componente irregular de la serie.

5- Obtener la tendencia suavizada (S_t) y el ciclo (Y_t-S_t) de la series del $LNPIB_t$, con el comando *hprescott,* esto se consigue agregando la opción *stub*; adicionalmente se puede especificar el valor del parámetro de atenuación λ, programando la instrucción *smooth* (*véase* figura 4.5). Aunque por defecto, como la serie se declaró trimestral este equivale a 1.600.

Figura 4.5. Salida de Stata® para ejecutar el filtro *H-P*



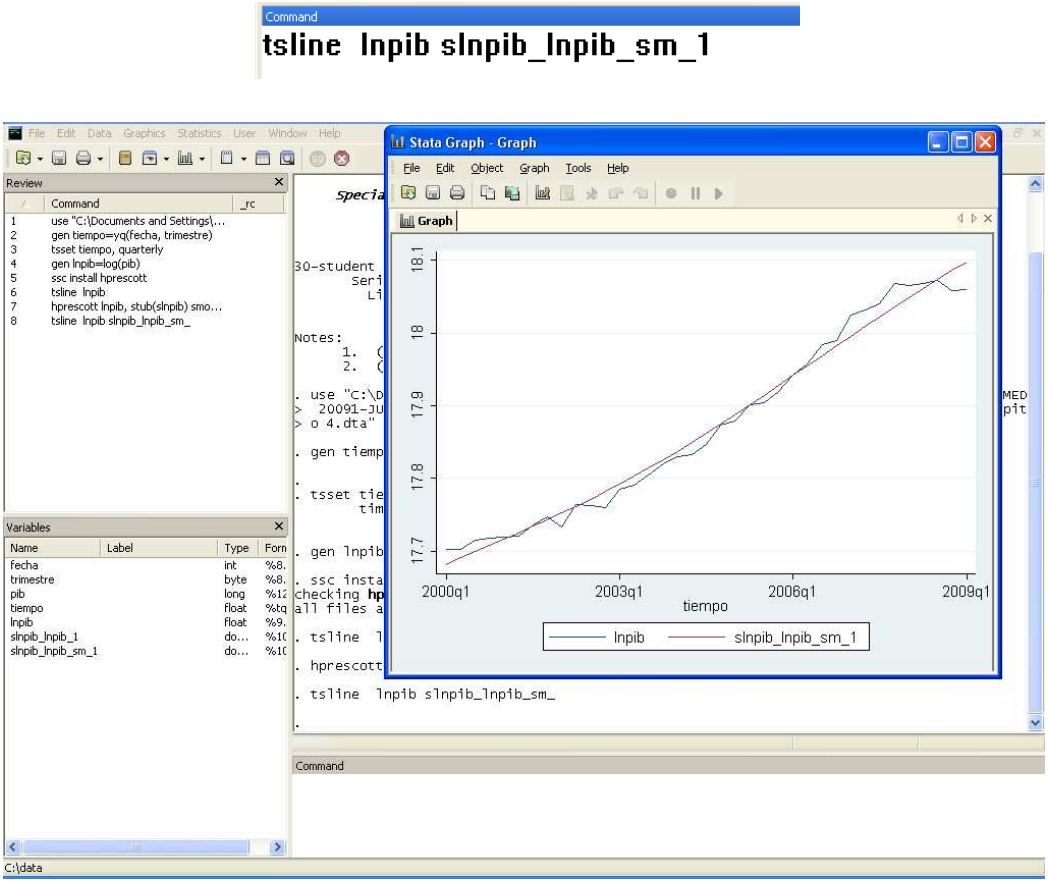
Fuente: cálculos autores.

Observe que, una vez ejecutados los comandos *hprescott*, *stub* y *smooth*, aparecen dos nuevas variables; slnpib_lnpib_1 y slnpib_lnpib _sm_1. Estas, son el resultado obtenido de solucionar la ecuación 4.4 expuesta en la

sección 4.3; correspondiendo al componente cíclico (Y_t-S_t) y tendencia suavizada (S_t) respectivamente, para el $LNPIB_t$ ($LNPIB_t$-S_t).

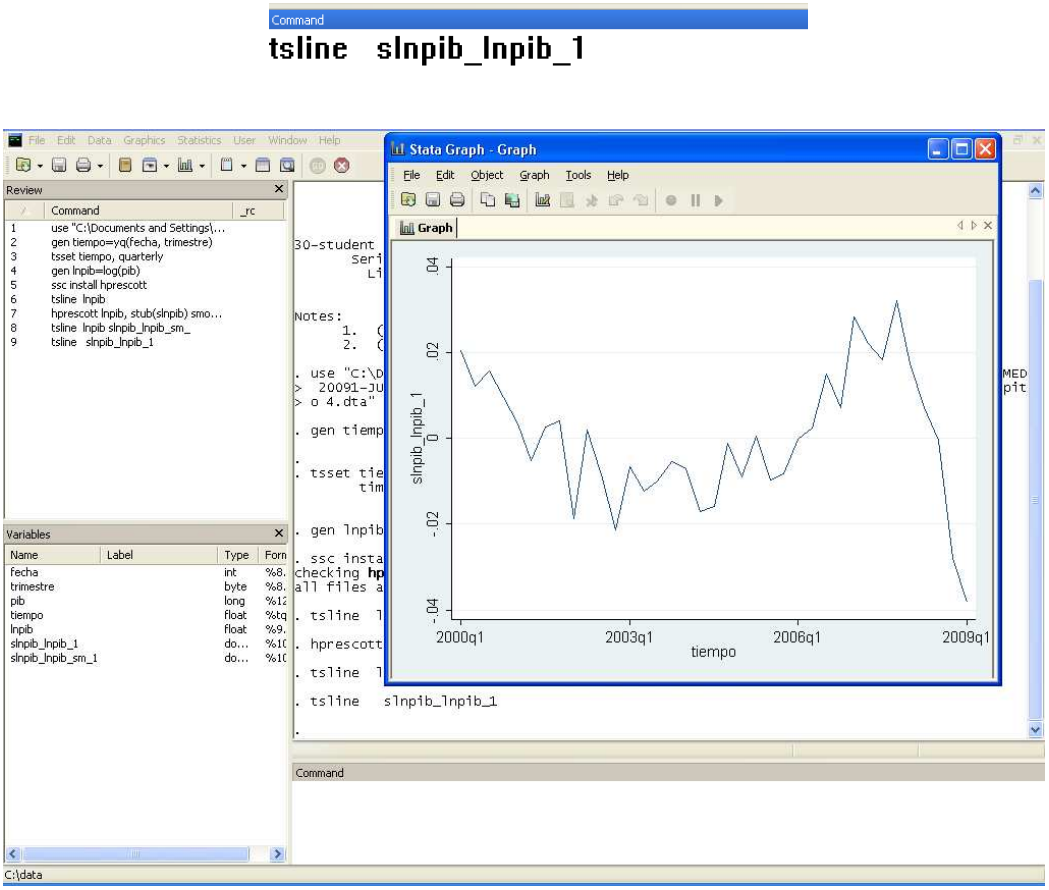
6- Graficar con el comando *tsline*, la tendencia suavizada (S_t, slnpib_lnpib _sm_1) y el ciclo (Y_t-S_t, slnpib_lnpib_1) o fluctuación de corto plazo para el $LNPIB_t$. Con el fin de conocer su trayectoria (*véase* línea roja figura 4.6, y línea azul 4.7).

Figura 4.6. Salida de Stata® para graficar la tendencia suavizada con el filtro *H-P*



Fuente: cálculos autores.

Figura 4.7. Salida de Stata® para graficar el ciclo obtenido con el filtro *H-P*



Fuente: cálculos autores.

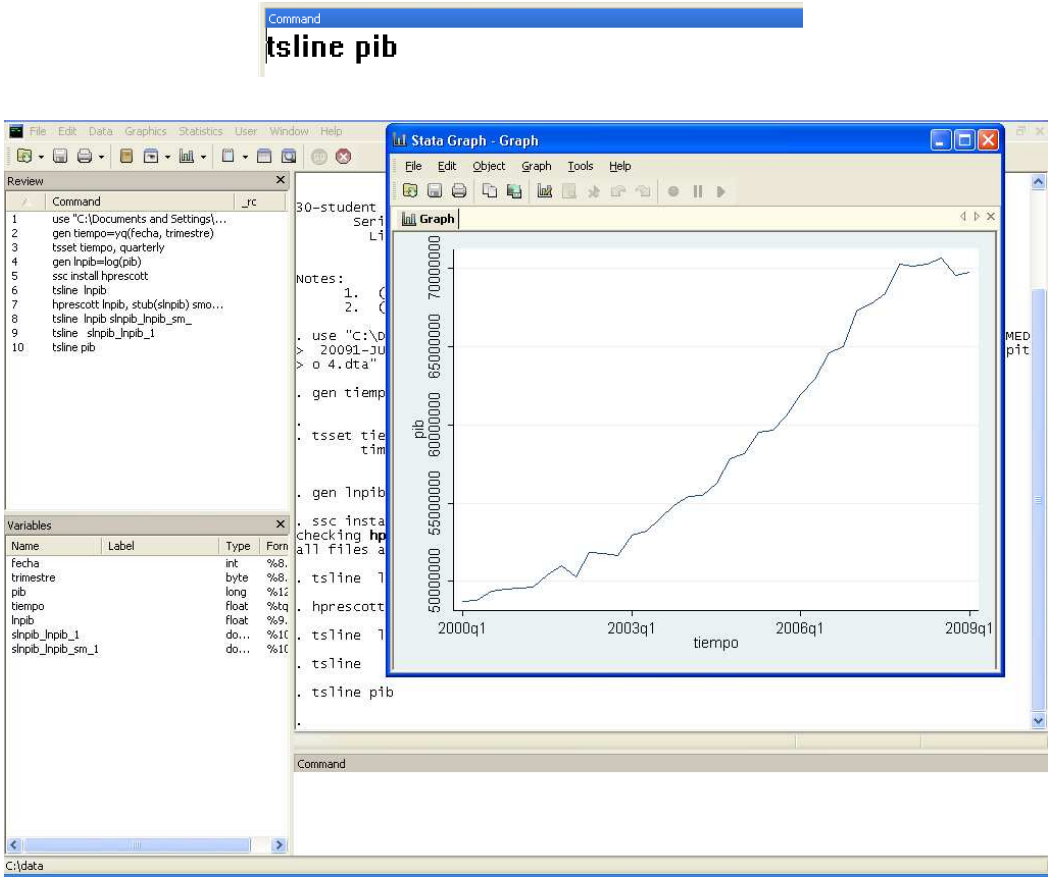
En la figura 4.7 se puede observar el movimiento cíclico para el $LNPIB_t$, donde su trayectoria es decreciente a finales de 1998 y comienzo de 1999. Confirmando la desaceleración económica colombiana, explicada posiblemente por la crisis financiera internacional; evidenciando así el componente irregular de la variable analizada. Con esto, se culmina el análisis de caso y evidencia del filtro Hodrick-Prescott; el siguiente tema corresponde a la aplicación para los modelos de pronósticos con tendencia determinística.

4.6.2 Modelos de pronósticos con tendencia determinística

Continuando con la serie de tiempo PIB (PIB_t) se pretende proyectarla un periodo ($\widehat{PIB}_{t+1}$ =2009-II), mediante los modelos de pronósticos con tendencia determinística; realizando paulatinamente la metodología en el programa Stata® de la siguiente manera:

1- Graficar el comportamiento del PIB_t a través del tiempo, con el comando *tsline* (*véase* figura 4.8), para conocer la forma funcional de la tendencia.

Figura 4.8. Salida de Stata® para graficar una serie de tiempo



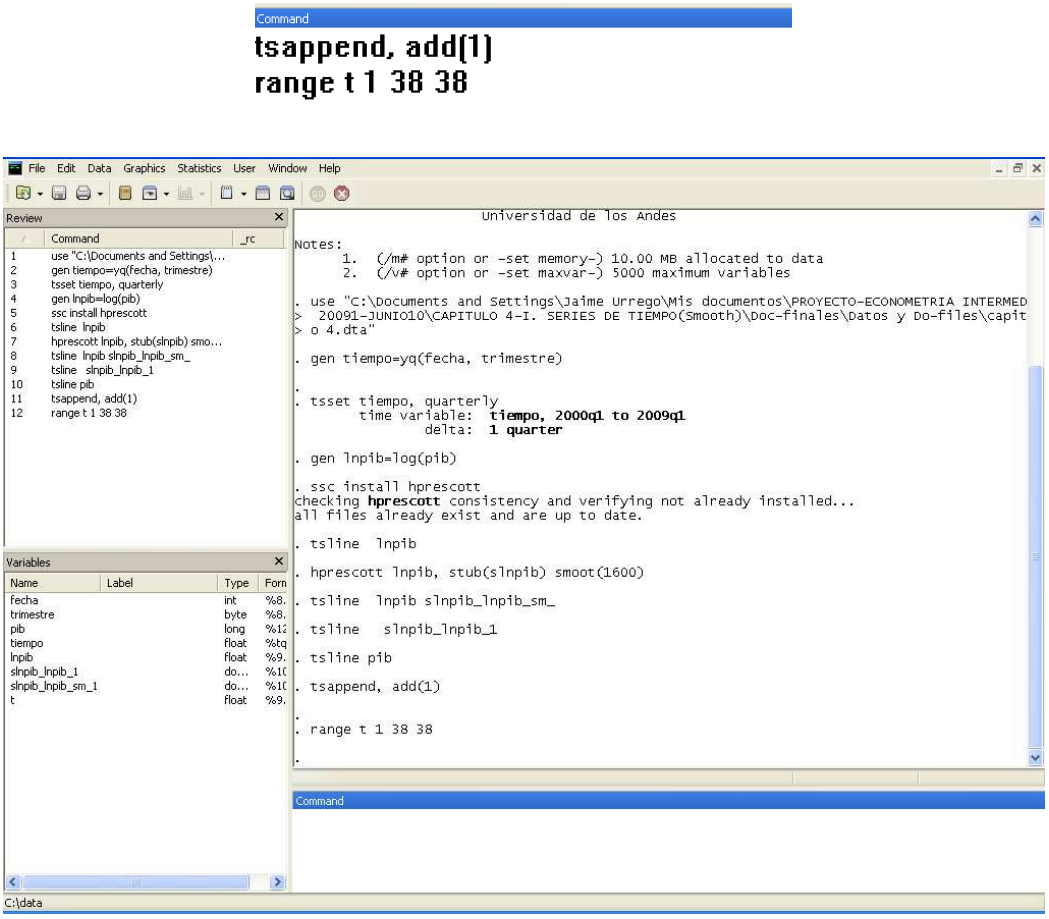
Fuente: cálculos autores.

En este caso, el PIB_t muestra una tendencia creciente lineal hasta el 2008-I, donde existe un punto de inflexión (*véase* figura 4.8). Como se mencionó

anteriormente, la variable cuestionada presenta media y varianza inestables entre 2000-I y 2009-I; dada la presencia tendencial e irregular en ella.

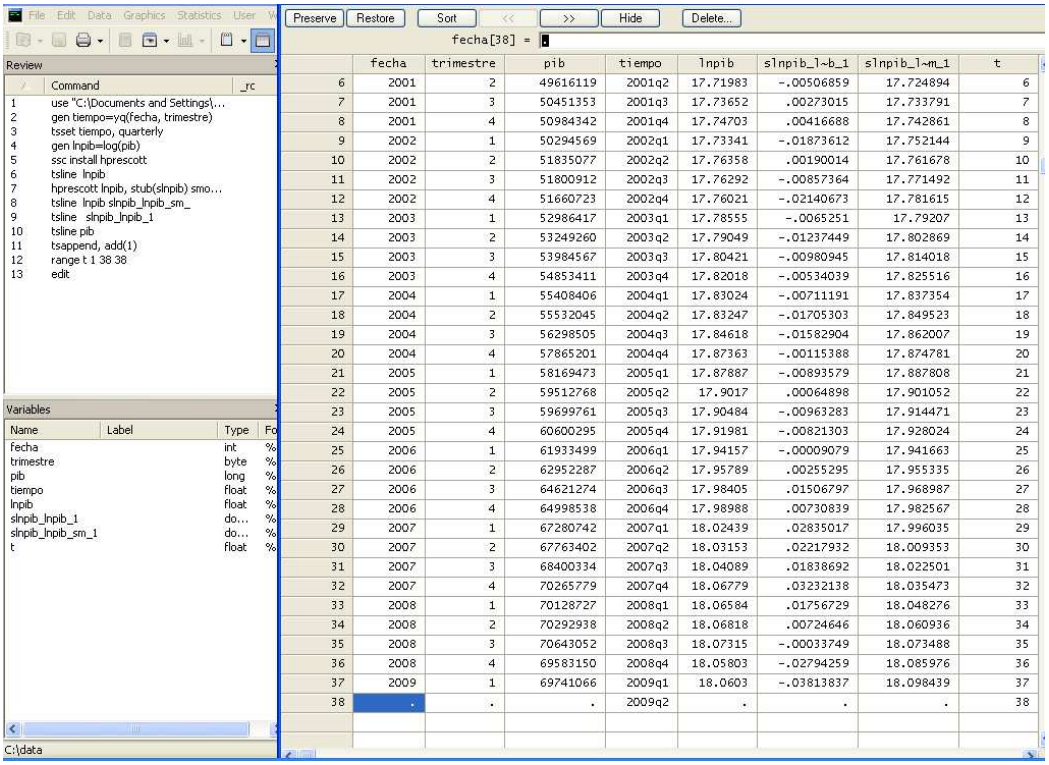
2- Adicionar una observación con el comando *tsappend, add(1)* (*véase* figura 4.9), para poder involucrar el nuevo valor del pronóstico ($\widehat{PIB}_{t+1}$ =2009-II) y así tener ahora una muestra de 38 observaciones (*n=38*). Posteriormente, generar una nueva variable de tendencia (*t*), con el comando *range* (*véase* figura 4.9), que contenga una serie que acumule la tendencia desde un valor inicial de 1 hasta 38; incluyendo este nuevo dato para la tendencia (*t+1=38*) y conseguir $\widehat{PIB}_{t+1}$ (*véase* figura 4.10).

Figura 4.9. Salida de Stata® para adicionar datos y generar T_t



Fuente: cálculos autores.

Figura 4.10. Salida de Stata® con información sobre el PIB colombiano, en millones de pesos a precios constantes del año 2000, y la tendencia.



Review

	Command	_rc
1	use "C:\Documents and Settings\...	
2	gen tiempo=yq(fecha, trimestre)	
3	tsset tiempo, quarterly	
4	gen lnpib=log(pib)	
5	ssc install hprescott	
6	tsline lnpib	
7	hprescott lnpib, stub(slnpib) smo...	
8	tsline lnpib slnpib_lnpib_sm_	
9	tsline slnpib_lnpib_1	
10	tsline pib	
11	tsappend, add(1)	
12	range t 1 38 38	
13	edit	

Variables

Name	Label	Type	Fo
fecha		int	%...
trimestre		byte	%...
pib		long	%...
tiempo		float	%...
lnpib		float	%...
slnpib_lnpib_1		do...	%...
slnpib_lnpib_sm_1		do...	%...
t		float	%...

C:\data

Preserve | Restore | Sort | << | >> | Hide | Delete...

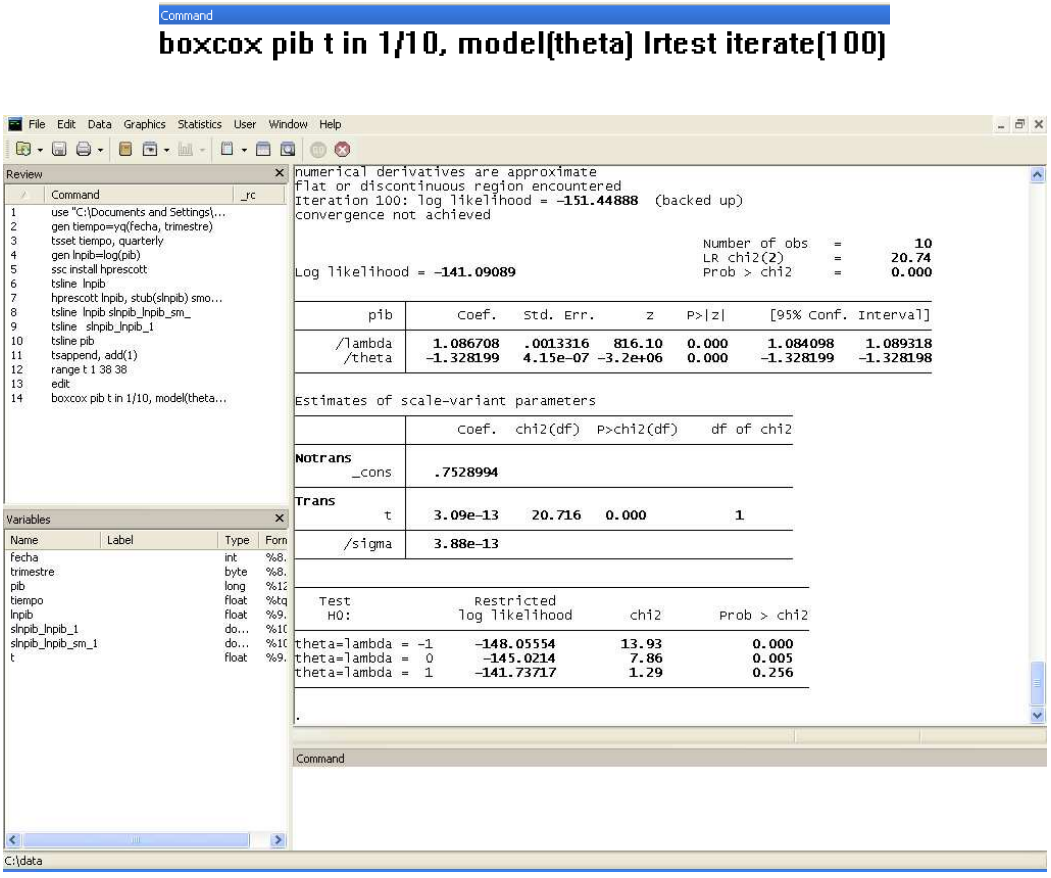
fecha[38] =

	fecha	trimestre	pib	tiempo	lnpib	slnpib_l~b_1	slnpib_l~m_1	t
6	2001	2	49616119	2001q2	17.71983	-.00506859	17.724894	6
7	2001	3	50451353	2001q3	17.73652	.00273015	17.733791	7
8	2001	4	50984342	2001q4	17.74703	.00416688	17.742861	8
9	2002	1	50294569	2002q1	17.73341	-.01873612	17.752144	9
10	2002	2	51835077	2002q2	17.76358	.00190014	17.761678	10
11	2002	3	51800912	2002q3	17.76292	-.00857364	17.771492	11
12	2002	4	51660723	2002q4	17.76021	-.02140673	17.781615	12
13	2003	1	52986417	2003q1	17.78555	-.0065251	17.79207	13
14	2003	2	53249260	2003q2	17.79049	-.01237449	17.802869	14
15	2003	3	53984567	2003q3	17.80421	-.00980945	17.814018	15
16	2003	4	54853411	2003q4	17.82018	-.00534039	17.825516	16
17	2004	1	55408406	2004q1	17.83024	-.00711191	17.837354	17
18	2004	2	55532045	2004q2	17.83247	-.01705303	17.849523	18
19	2004	3	56298505	2004q3	17.84618	-.01582904	17.862007	19
20	2004	4	57865201	2004q4	17.87363	-.00115388	17.874781	20
21	2005	1	58169473	2005q1	17.87887	-.00893579	17.887808	21
22	2005	2	59512768	2005q2	17.9017	.00064898	17.901052	22
23	2005	3	59699761	2005q3	17.90484	-.00963283	17.914471	23
24	2005	4	60600295	2005q4	17.91981	-.00821303	17.928024	24
25	2006	1	61933499	2006q1	17.94157	-.00009079	17.941663	25
26	2006	2	62952287	2006q2	17.95789	.00255295	17.955335	26
27	2006	3	64621274	2006q3	17.98405	.01506797	17.968987	27
28	2006	4	64998538	2006q4	17.98988	.00730839	17.982567	28
29	2007	1	67280742	2007q1	18.02439	.02835017	17.996035	29
30	2007	2	67763402	2007q2	18.03153	.02217932	18.009353	30
31	2007	3	68400334	2007q3	18.04089	.01838692	18.022501	31
32	2007	4	70265779	2007q4	18.06779	.03232138	18.035473	32
33	2008	1	70128727	2008q1	18.06584	.01756729	18.048276	33
34	2008	2	70292938	2008q2	18.06818	.00724646	18.060936	34
35	2008	3	70643052	2008q3	18.07315	-.00033749	18.073488	35
36	2008	4	69583150	2008q4	18.05803	-.02794259	18.085976	36
37	2009	1	69741066	2009q1	18.0603	-.03813837	18.098439	37
38	.	.	.	2009q2	.	.	.	38

Fuente: Dane y cálculos autores.

3- Estimar la función Box-Cox, mediante el comando *boxcox* (*véase* figura 4.11), para conocer la mejor forma funcional ajustada a la tendencia del PIB, conforme con los cuadros 4.1 y 4.2 expuestos.

Figura 4.11. Salida de Stata® para el modelo Box-Cox



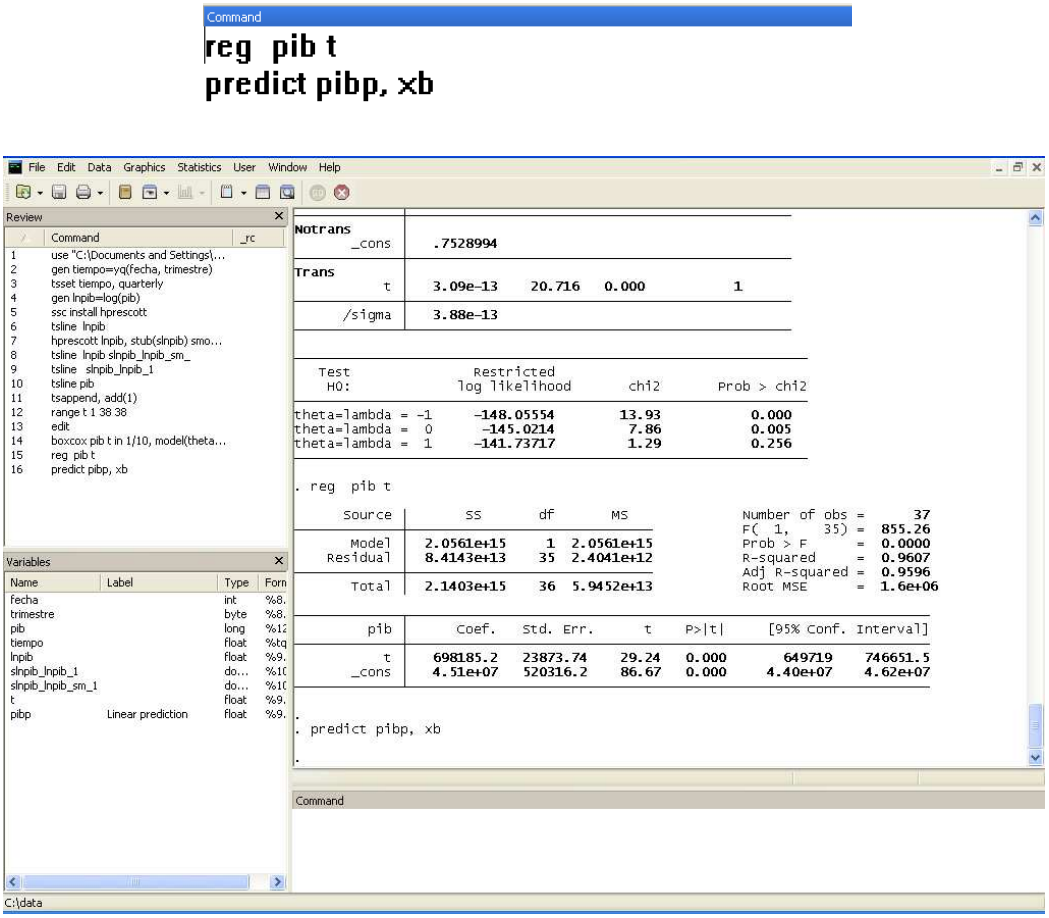
Fuente: cálculos autores.

De acuerdo con los resultados obtenidos en la prueba de Razón de Verosimilitud -*RV*- (*véase* probabilidad de 0.256 en la figura 4.11), se puede concluir que el modelo de tendencia determinística lineal es el más adecuado para pronosticar el PIB. En otras palabras, sus parámetros de transformación θ y λ son simultáneamente iguales a uno (*véase* resultados en la figura 4.11).

4- Por los resultados del numeral anterior ($\theta = \lambda = 1$, *véase* cuadro 4.2 en la sección 4.4), se puede estimar el modelo de tendencia determinística lineal mediante MCO en Stata® utilizando el comando *reg* (*véase* figura 4.10A). Seguido del pronóstico con la opción *predict pibp, xb* (*véase* figura 4.10B) y finalmente graficar (instrucción *tsline*)

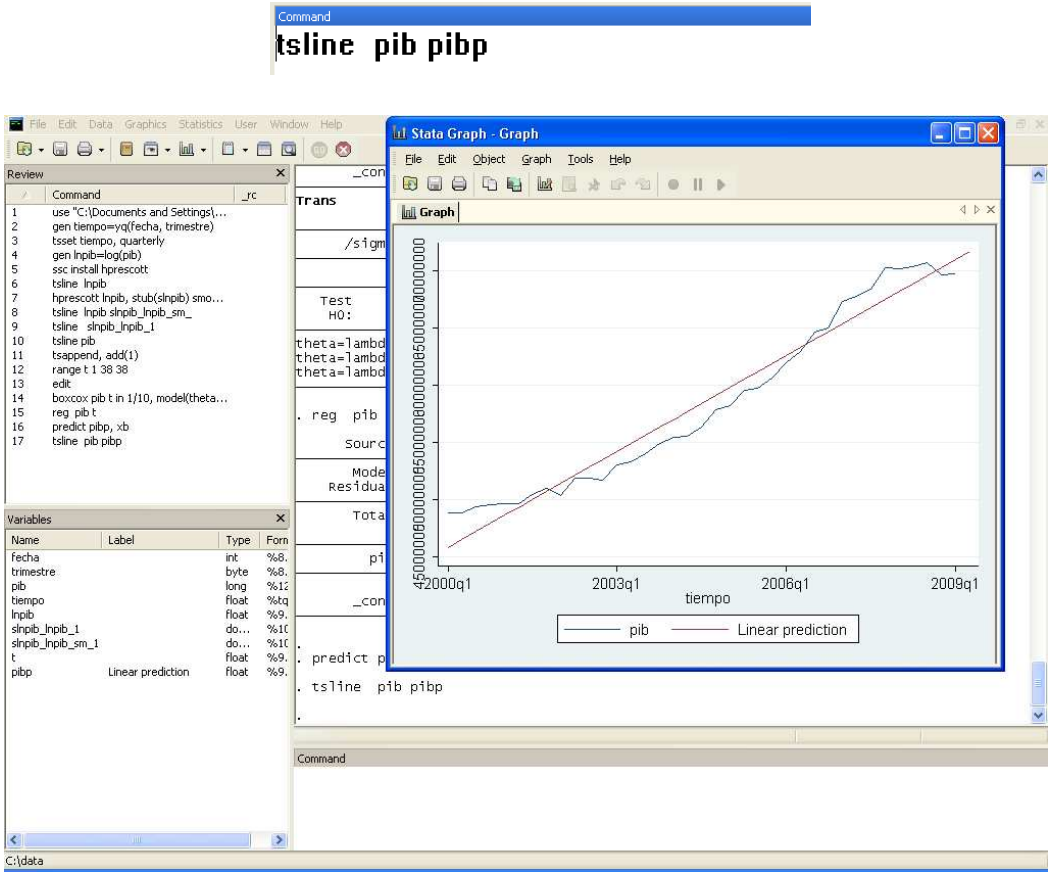
paralelamente las series observada (PIB_t) y proyectada ($\widehat{PIB}_{t+1}$); para conocer si el ajuste y tendencia de la predicción (*ppib*) es similar a la original (*véase* línea roja figura 4.13 y 4.14).

Figura 4.12. Salida de Stata® con la regresión lineal y pronóstico del PIB



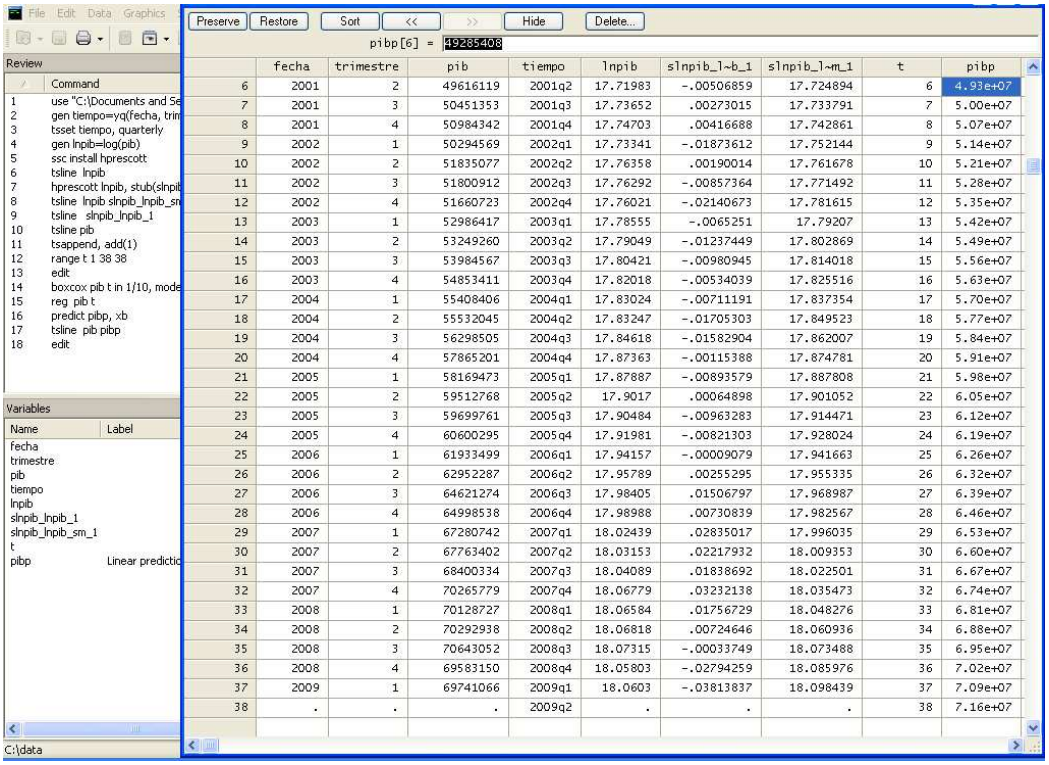
Fuente: cálculos autores.

Figura 4.13. Salida de Stata® con representación del pronóstico



Fuente: cálculos autores.

Figura 4.14. Salida de Stata®, con información sobre el pronóstico del PIB Colombiano.



	fecha	trimestre	pib	tiempo	lnpib	slnpib_l~b_1	slnpib_l~m_1	t	pibp
6	2001	2	49616119	2001q2	17.71983	-.00506859	17.724894	6	4.93e+07
7	2001	3	50451353	2001q3	17.73652	.00273015	17.733791	7	5.00e+07
8	2001	4	50984342	2001q4	17.74703	.00416688	17.742861	8	5.07e+07
9	2002	1	50294569	2002q1	17.73341	-.01873612	17.752144	9	5.14e+07
10	2002	2	51835077	2002q2	17.76358	.00190014	17.761678	10	5.21e+07
11	2002	3	51800912	2002q3	17.76292	-.00857364	17.771492	11	5.28e+07
12	2002	4	51660723	2002q4	17.76021	-.02140673	17.781615	12	5.35e+07
13	2003	1	52986417	2003q1	17.78555	-.0065251	17.79207	13	5.42e+07
14	2003	2	53249260	2003q2	17.79049	-.01237449	17.802869	14	5.49e+07
15	2003	3	53984567	2003q3	17.80421	-.00980945	17.814018	15	5.56e+07
16	2003	4	54853411	2003q4	17.82018	-.00534039	17.825516	16	5.63e+07
17	2004	1	55408406	2004q1	17.83024	-.00711191	17.837354	17	5.70e+07
18	2004	2	55532045	2004q2	17.83247	-.01705303	17.849523	18	5.77e+07
19	2004	3	56298505	2004q3	17.84618	-.01582904	17.862007	19	5.84e+07
20	2004	4	57865201	2004q4	17.87363	-.00115388	17.874781	20	5.91e+07
21	2005	1	58169473	2005q1	17.87887	-.00893579	17.887808	21	5.98e+07
22	2005	2	59512768	2005q2	17.9017	.00064898	17.901052	22	6.05e+07
23	2005	3	59699761	2005q3	17.90484	-.00963283	17.914471	23	6.12e+07
24	2005	4	60600295	2005q4	17.91981	-.00821303	17.928024	24	6.19e+07
25	2006	1	61933499	2006q1	17.94157	-.00009079	17.941663	25	6.26e+07
26	2006	2	62952287	2006q2	17.95789	.00255295	17.955335	26	6.32e+07
27	2006	3	64621274	2006q3	17.98405	.01506797	17.968987	27	6.39e+07
28	2006	4	64998538	2006q4	17.98988	.00730839	17.982567	28	6.46e+07
29	2007	1	67280742	2007q1	18.02439	.02835017	17.996035	29	6.53e+07
30	2007	2	67763402	2007q2	18.03153	.02217932	18.009353	30	6.60e+07
31	2007	3	68400334	2007q3	18.04089	.01838692	18.022501	31	6.67e+07
32	2007	4	70265779	2007q4	18.06779	.03232138	18.035473	32	6.74e+07
33	2008	1	70128727	2008q1	18.06584	.01756729	18.048276	33	6.81e+07
34	2008	2	70292938	2008q2	18.06818	.00724646	18.060936	34	6.88e+07
35	2008	3	70643052	2008q3	18.07315	-.00033749	18.073488	35	6.95e+07
36	2008	4	69583150	2008q4	18.05803	-.02794259	18.085976	36	7.02e+07
37	2009	1	69741066	2009q1	18.0603	-.03813837	18.098439	37	7.09e+07
38	.	.	.	2009q2	.	.	.	38	7.16e+07

Fuente: cálculos autores.

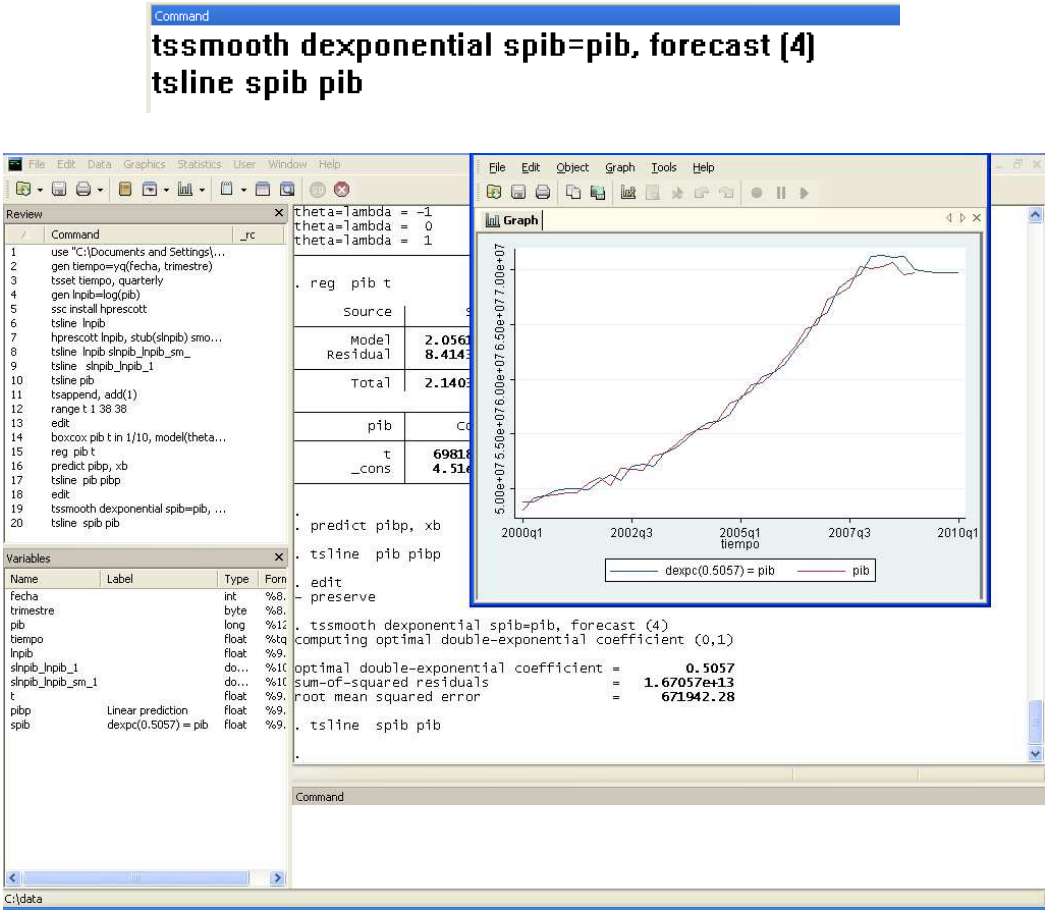
Finalizando la aplicación para los modelos de pronósticos con tendencia determinística en las figuras 4.13 y 4.14 se observa gráfica y numéricamente como diverge el valor del PIB proyectado (70900000 *y* 71600000) con el observado (69'741.066), no se adhiere al comportamiento de su tendencia. Recordando, que generalmente las predicciones concebidas con estos modelos no son tan precisas. Por esto, el siguiente tema corresponde al mismo ejemplo aplicando métodos de suavizamiento exponencial.

4.6.3 Pronóstico con métodos de atenuación exponencial

Retomando de nuevo el PIB trimestral (PIB_t) y aplicando los métodos de suavizamiento exponencial doble (AD) y Holt-Winters no estacional, se pretende pronosticarlo cuatro periodos ($\widehat{PIB}_{t+4}$ =2010-I); debido a que previamente se identificó tendencia creciente en la serie. Por esto, fueron descartadas las otras técnicas expuestas en el cuadro 4.3 (promedios móviles, atenuación simple y Holt-Winters estacional) no apropiadas para este caso. Para lo anterior se aplican gradualmente estas metodologías, en el programa Stata®, de la siguiente manera:

1- Realizar atenuación exponencial doble (AD), recordando que se debe adicionar previamente los periodos a pronosticar con el comando con el comando *tsappend, add(4)*. Posteriormente la instrucción *tssmooth dexponential spib = pib, forecast(4)* (*véase* figura 4.15) para conocer la nueva variable suavizada (*spib*) que contienen la predicción hasta el 2010-I. Igualmente graficar (*tsline*) *spib* y saber si es similar a la original (*véase* línea roja figura 4.15 y 4.17 en el anexo 4.1).

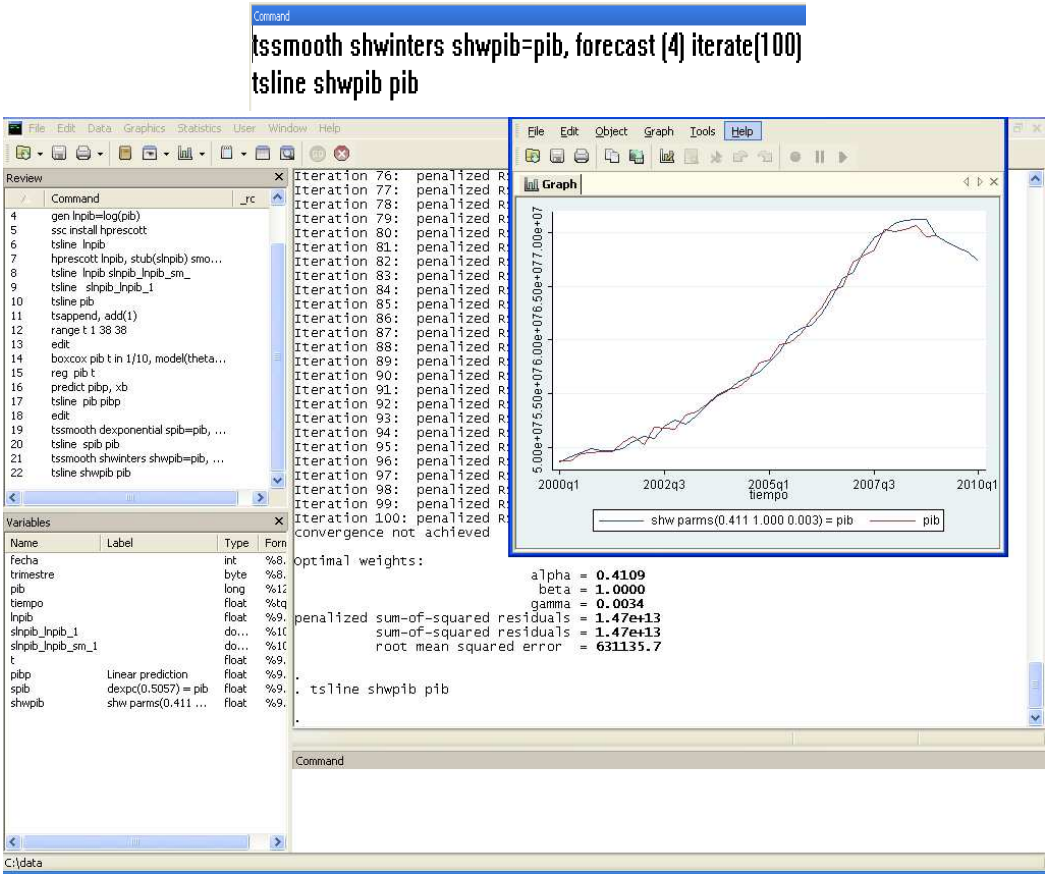
Figura 4.15. Salida de Stata® aplicando atenuación exponencial doble



Fuente: cálculos autores.

2- Aplicar atenuación Holt-Winters no estacional, mediante la instrucción *tssmooth shwinters spibhw = pib, forecast(4) iterate(100)* (*véase* figura 4.16) para conocer la nueva variable suavizada (*spibhw*) que contienen la predicción hasta el 2010-I. Igualmente graficar (*tsline*) *spibhw* y saber si es similar a la original (*véase* línea roja figura 4.16 y 4.17 anexo 4.1). Para series temporales con otras naturalezas (ciclo, estacionaria o estacional), en el cuadro 4.7 del anexo 4.1 se encuentra la programación en Stata® para aplicar el método de atenuación adecuado, de acuerdo a esta.

Figura 4.16. Salida de Stata® aplicando atenuación exponencial Holt-Winters no estacional



Fuente: cálculos autores.

De acuerdo con lo expuesto en la sección 4.5 (*véase* cuadro 4.4), figuras 4.16 y 4.17 (anexo 4.1), el pronóstico de mejor ajuste es el obtenido mediante Holt-Winters no estacional. Porque su raíz cuadrada del promedio para la suma de errores al cuadrado[60] (RCPSEC= 631135.7) es inferior al suavizamiento doble (RCPSEC= 671942.28).

Además los resultados de sus predicciones ($\widehat{PIB}_{t+1,2009-II}$ =69'000.000, $\widehat{PIB}_{t+2,2009-III}$ =68'600.000, $\widehat{PIB}_{t+3,2009-IV}$ =68'100.000 y $\widehat{PIB}_{t+4,2010-I}$ =67'300.000) gráficamente se adhieren mejor a la tendencia de la serie original (PIB_t). Obtenidos a partir de los parámetros de suavizamiento (α=0.4109 y β=1), entre cero y uno, que

[60] *Véase* cuadro 4.4 de este capítulo.

garantizan el menor valor para los indicadores de error del pronóstico (RCPSEC), expuestos en el cuadro 4.4. Por otra parte, los resultados del pronóstico reflejan la posible desaceleración en que puede entrar la economía colombiana duarte el 2009 y 2010.

Con este estudio de caso finaliza la introducción a series de tiempo, aplicación del filtro *H-P*, modelos de pronósticos con tendencia determinística y métodos de predicción con atenuación exponencial. En el próximo capítulo se encuentra la metodología Box-Jenkins (BJ), también para realizar proyecciones, a partir de series estacionarias, procesos autorregresivos y de media móvil.

Resumen

- Una serie de tiempo se define como un conjunto de observaciones coleccionadas sucesiva y homogéneamente para una misma variable en periodos específicos. En el análisis de este tipo de variables, se desarrollan modelos donde su comportamiento actual es función de su tendencia o propio pasado; el cual otorga información sobre la trayectoria que continuara en el futuro.
- Una serie de tiempo generalmente contiene cuatro componentes que pueden desagregarse de la siguiente manera: tendencia -componente de baja variabilidad que evoluciona en alguna dirección particular-; ciclo oscilación de largo plazo independiente alrededor de la tendencia-; componente estacional -movimientos de una serie sucedidos con una frecuencia definida de tiempo- y componente irregular –ocasionados por choques exógenos impredecibles e inesperados alterando bruscamente el curso normal de la variable-.
- La tendencia, ciclo, estacionalidad y componente irregular de una serie pueden relacionarse entre sí aditiva o multiplicativamente. Esta interrelación, define la naturaleza de la serie y sus diversas dinámicas del comportamiento a lo largo del tiempo.
- El filtro Hodrick-Prescott permite separar la tendencia y componente cíclico para una serie de tiempo a partir de una variable suavizada. La brecha entre la tendencia y atenuación, se puede interpretada como el ciclo que corresponde a la diferencia entre el su valor real y potencial.
- El método más sencillo para pronosticar una serie de tiempo, es empleando los llamados modelos de tendencia determinística. Los cuales, proyectan la variable de acuerdo a la forma funcional subyacente para su tendencia.
- Otra técnica para pronosticar series de tiempo son las de suavizamiento exponencial, ellas tienen en cuenta su propio pasado, tendencia, ciclo, estacionalidad y naturaleza de la variable dinámica para de esta forma proyectarla. Esta metodología, obtiene sus parámetros de atenuación (α, β y γ), entre cero y uno, a partir de tanteo; seleccionado los valores de acuerdo al mínimo indicador de error para el pronóstico que genere.

- Dentro de las técnicas de suavizamiento exponencial más destacadas se encuentran: promedios móviles, atenuación simple, doble y Holt-Winters (no estacional, aditiva y multiplicativa).

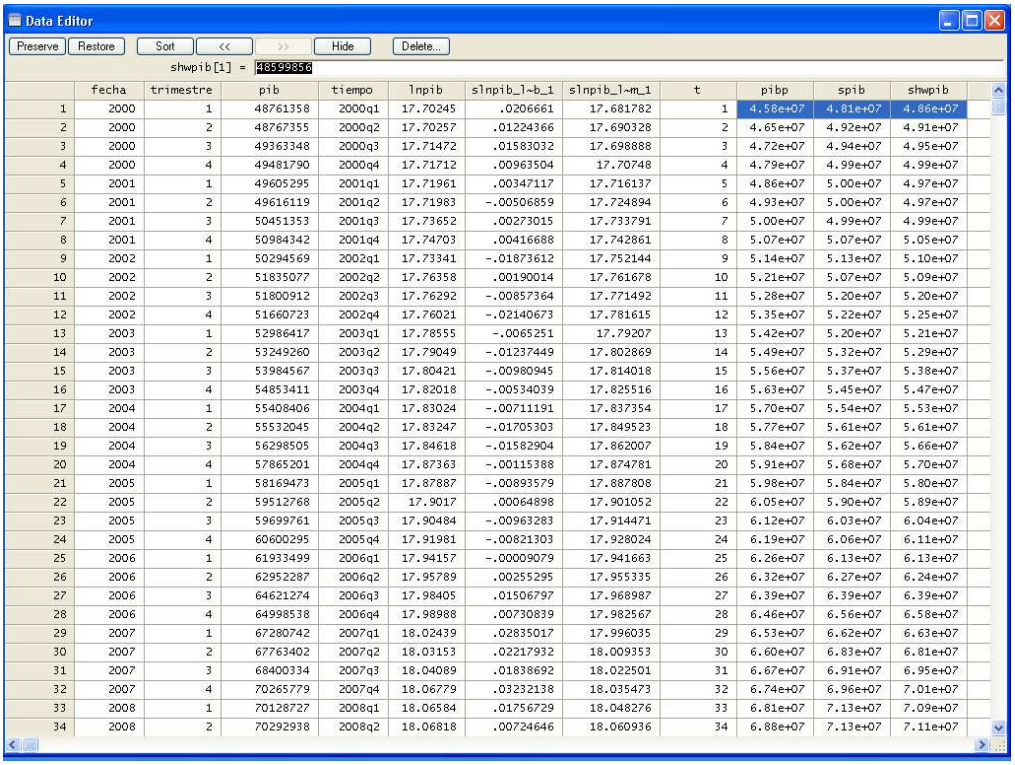
Anexo 4

Cuadro 4.5. Información sobre el PIB Colombiano, en millones de pesos a precios constantes del año 2000.

Fecha	Trimestre	PIB	Fecha	Trimestre	PIB
2000	1	48,761,358	2004	4	57,865,201
2000	2	48,767,355	2005	1	58,169,473
2000	3	49,363,348	2005	2	59,512,768
2000	4	49,481,790	2005	3	59,699,761
2001	1	49,605,295	2005	4	60,600,295
2001	2	49,616,119	2006	1	61,933,499
2001	3	50,451,353	2006	2	62,952,287
2001	4	50,984,342	2006	3	64,621,274
2002	1	50,294,569	2006	4	64,998,538
2002	2	51,835,077	2007	1	67,280,742
2002	3	51,800,912	2007	2	67,763,402
2002	4	51,660,723	2007	3	68,400,334
2003	1	52,986,417	2007	4	70,265,779
2003	2	53,249,260	2008	1	70,128,727
2003	3	53,984,567	2008	2	70,292,938
2003	4	54,853,411	2008	3	70,643,052
2004	1	55,408,406	2008	4	69,583,150
2004	2	55,532,045	2009	1	69,741,066
2004	3	56,298,505	-	-	-

Fuente: DANE.

Figura 4.17. Salida de Stata®, con información sobre el pronóstico del PIB Colombiano, con todas las metodologías expuestas en el capítulo.



Data Editor

Preserve | Restore | Sort | << | >> | Hide | Delete...

shwpib[1] = 48599856

	fecha	trimestre	pib	tiempo	lnpib	slnpib_l~b_1	slnpib_l~m_1	t	pibp	spib	shwpib
1	2000	1	48761358	2000q1	17.70245	.0206661	17.681782	1	4.58e+07	4.81e+07	4.86e+07
2	2000	2	48767355	2000q2	17.70257	.01224366	17.690328	2	4.65e+07	4.92e+07	4.91e+07
3	2000	3	49363348	2000q3	17.71472	.01583032	17.698888	3	4.72e+07	4.94e+07	4.95e+07
4	2000	4	49481790	2000q4	17.71712	.00963504	17.70748	4	4.79e+07	4.99e+07	4.99e+07
5	2001	1	49605295	2001q1	17.71961	.00347117	17.716137	5	4.86e+07	5.00e+07	4.97e+07
6	2001	2	49616119	2001q2	17.71983	-.00506859	17.724894	6	4.93e+07	5.00e+07	4.97e+07
7	2001	3	50451353	2001q3	17.73652	.00273015	17.733791	7	5.00e+07	4.99e+07	4.99e+07
8	2001	4	50984342	2001q4	17.74703	.00416688	17.742861	8	5.07e+07	5.07e+07	5.05e+07
9	2002	1	50294569	2002q1	17.73341	-.01873612	17.752144	9	5.14e+07	5.13e+07	5.10e+07
10	2002	2	51835077	2002q2	17.76358	.00190014	17.761678	10	5.21e+07	5.07e+07	5.09e+07
11	2002	3	51800912	2002q3	17.76292	-.00857364	17.771492	11	5.28e+07	5.20e+07	5.20e+07
12	2002	4	51660723	2002q4	17.76021	-.02140673	17.781615	12	5.35e+07	5.22e+07	5.25e+07
13	2003	1	52986417	2003q1	17.78555	-.0065251	17.79207	13	5.42e+07	5.20e+07	5.21e+07
14	2003	2	53249260	2003q2	17.79049	-.01237449	17.802869	14	5.49e+07	5.32e+07	5.29e+07
15	2003	3	53984567	2003q3	17.80421	-.00980945	17.814018	15	5.56e+07	5.37e+07	5.38e+07
16	2003	4	54853411	2003q4	17.82018	-.00534039	17.825516	16	5.63e+07	5.45e+07	5.47e+07
17	2004	1	55408406	2004q1	17.83024	-.00711191	17.837354	17	5.70e+07	5.54e+07	5.53e+07
18	2004	2	55532045	2004q2	17.83247	-.01705303	17.849523	18	5.77e+07	5.61e+07	5.61e+07
19	2004	3	56298505	2004q3	17.84618	-.01582904	17.862007	19	5.84e+07	5.62e+07	5.66e+07
20	2004	4	57865201	2004q4	17.87363	-.00115388	17.874781	20	5.91e+07	5.68e+07	5.70e+07
21	2005	1	58169473	2005q1	17.87887	-.00893579	17.887808	21	5.98e+07	5.84e+07	5.80e+07
22	2005	2	59512768	2005q2	17.9017	.00064898	17.901052	22	6.05e+07	5.90e+07	5.89e+07
23	2005	3	59699761	2005q3	17.90484	-.00963283	17.914471	23	6.12e+07	6.03e+07	6.04e+07
24	2005	4	60600295	2005q4	17.91981	-.00821303	17.928024	24	6.19e+07	6.06e+07	6.11e+07
25	2006	1	61933499	2006q1	17.94157	-.00009079	17.941663	25	6.26e+07	6.13e+07	6.13e+07
26	2006	2	62952287	2006q2	17.95789	.00255295	17.955335	26	6.32e+07	6.27e+07	6.24e+07
27	2006	3	64621274	2006q3	17.98405	.01506797	17.968987	27	6.39e+07	6.39e+07	6.39e+07
28	2006	4	64998538	2006q4	17.98988	.00730839	17.982567	28	6.46e+07	6.56e+07	6.58e+07
29	2007	1	67280742	2007q1	18.02439	.02835017	17.996035	29	6.53e+07	6.62e+07	6.63e+07
30	2007	2	67763402	2007q2	18.03153	.02217932	18.009353	30	6.60e+07	6.83e+07	6.81e+07
31	2007	3	68400334	2007q3	18.04089	.01838692	18.022501	31	6.67e+07	6.91e+07	6.95e+07
32	2007	4	70265779	2007q4	18.06779	.03232138	18.035473	32	6.74e+07	6.96e+07	7.01e+07
33	2008	1	70128727	2008q1	18.06584	.01756729	18.048276	33	6.81e+07	7.13e+07	7.09e+07
34	2008	2	70292938	2008q2	18.06818	.00724646	18.060936	34	6.88e+07	7.13e+07	7.11e+07

Fuente: cálculos autores.

Cuadro 4.6. Transformación para variables de tiempo.

Caso	Solución	Comando
Variable de tiempo viene guardada en la forma: Día-Mes-Año	Transformar la variable original a una nueva variable	gen nuevaVariable = date(variableOriginal, "DMY")
Existen múltiples variables de tiempo; Dia Mes Año	Consolidar las variables originales en una única nueva variable	gen nuevaVariable = mdy(Mes, Dia, Año)

Fuente: los autores.

Cuadro 4.7. Casos particulares del comando tssmoth.

Metodología de Suavizamiento.	Comando
Promedios móviles (PM)	tssmooth ma pib_ma=pib
Atenuación simple (AS)	tssmooth exponential pib_as=pib
Atenuación doble (AD)	tssmooth dexponential pib_ad=pib
Holt-Winters	tssmooth hwinters pib_hw=pib
Holt-Winters estacional	tssmooth shwinters pib_shw=pib

Fuente: los autores.

Capítulo 5

Metodología Box-Jenkins para pronosticar series de tiempo, mediante procesos autorregresivos y media móvil

5.1 Introducción

Prosiguiendo el tratamiento sobre series de tiempo y una vez realizada su introducción, este capítulo realiza una discusión sobre los modelos Arima[61] (Autoregressive Intregrated Moving Average, siglas en inglés), incursionada por Box-Jenkins (BJ) a partir de procesos autorregresivos (AR, Autoregressive, siglas en inglés) y media móvil (MA, Moving Average, siglas en inglés). Este procedimiento, permite especificar y estimar modelos para generar pronósticos de corto plazo con muestras representativas (grandes).

Por otra parte los pronósticos, obtenidos mediante modelos Arima, tratan e incluyen todos los componentes de la serie (tendencia, ciclo, estacionalidad e irregularidad). Mientras con atenuación exponencial, definidas en el capitulo anterior, no se considera su elemento irregular. En otras palabras, la diferencia entre estas metodologías consiste en el elemento irregular considerado en el primero; adicional a los otros elementos involucrados bajo la segunda técnica.

Razón por la cual, este capítulo se centra en modelar el componente señalado, permitiendo así predecir $\left(\hat{Y}_{t+p}\right)$ el comportamiento de la variable (Y_t) en el corto plazo. Para esto, en su contenido se encuentran temas relacionados con el análisis de variables estacionarias, ruido blanco y ergódicas; funciones de autocorrelación simple y parcial, pruebas de correlogramas, raíz unitaria y estadísticos Box-Pierce y Ljung-Box. Igualmente, las formas para detectar y estimar procesos AR, MA, ARMA, Arima y Sarima.

[61]Autorregresivo integrado de media móvil.

Para lo anterior, las siguientes secciones presentan algunos conceptos básicos, usados a lo largo del capítulo y el siguiente; también describen los procesos generadores de datos (PGD) como los autorregresivos (AR) y media móvil (MA); estos dos últimos empleados para modelar el componente irregular de una serie. Adicionalmente en ellas, se discute la metodología Box-Jenkins; que considera todo el proceso necesario para efectuar pronósticos, así como sus ventajas y desventajas. Por último, se aplica esta metodología (BJ) mediante un estudio caso con los datos del capítulo 4, sobre el producto interno bruto (PIB) colombiano, e información mensual del índice de precios al consumidor (IPC) en Colombia.

5.2 Conceptos básicos

Esta sección expone algunos conceptos para el estudio de otros temas relacionados con series de tiempo, dirigidos a comprender los modelos de procesos autorregresivos y media móvil, metodología Box-Jenkins (BJ) y técnicas de rezagos distribuidos; discutidas en el siguiente capítulo. Entre estos, se encuentran los procesos estocásticos, estacionariedad[62], ruido blanco y condiciones de ergodicidad; ampliando un poco más los presentados en el capítulo 4.

5.2.1 Proceso estocástico discreto, estacionariedad, ruido blanco y ergodicidad.

A partir de la ilustración sobre una serie de tiempo[63], expuesta en capítulo 4, se deriva el proceso aleatorio o estocástico[64]. Entendido como el ejercicio de acumular información estadística, para variables de interés, por parte de organizaciones especializadas. Esta composición, se convierte en un proceso estocástico discreto (PED) porque los valores recopilados se obtienen sin ser predeterminados; subyacen de alguna metodología para su cálculo o monitoreo. Esta labor, permite

[62]Previamente se debe entender que el concepto de estacionariedad difiere a estacionalidad, tratado en el capítulo anterior; es distinto una variable estacionaria a una estacional.

[63]La cual se define como un conjunto finito de observaciones recogidas en momentos consecutivos y homogéneos de tiempo. La frecuencia es un valor entero positivo t= 0 ,1,…,T; característica que lo hace discreto.

[64]Evento totalmente independiente, no relacionado con ningún suceso; en otras palabras se da por el azar.

coleccionar datos o variables aleatorias ordenadas en el tiempo[65]; abriendo paso a cualquier aplicación con series de tiempo.

Además del concepto anterior y con el propósito de realizar pronósticos mediante BJ, los PED deben cumplir con la condición de estacionariedad[66]. Concebida, como aquella cuya distribución conjunta e incondicional[67] se mantienen constantes a lo largo del tiempo. En otras palabras, su media o promedio aritmético ($E[Y_t]$), varianza ($Var[Y_t]$) y covarianza de sus rezagos $\left(Cov\left[Y_t, Y_{t-p}\right]\right)$ no están condicionados con el tiempo[68].

Asimismo, cuando todos sus momentos (media, varianza, asimetría y curtosis) son finitos e independientes del tiempo; determinan una serie temporal estricta o fuertemente estacionaria. Esto, puede ser probado analíticamente y estadísticamente para los PED, sobre una muestra determinada. Sin embargo, esta característica no permitirá especificar un modelo Arima para pronosticarla.

Por esta razón, el PED únicamente debe cumplir con media y varianza finitas e independientes del tiempo[69]; condición denominada como estacionariedad débil o suave. Debido a que un PED fuertemente estacionario, equivale a una serie de tiempo ruido blanco o totalmente estocástica; la cual no puede pronosticarse, porque su condición es aleatoria o al azar[70].

Como ejemplo, para un PED fuerte estacionario, se tiene el lanzamiento de un dado durante varias ocasiones. Registrando cada resultado obtenido y suponiendo que no existe ningún tipo de alteración sobre el dado (no se hace trampa en su lanzamiento), es posible formar así una serie y con ella obtener un PED. Estos valores, son formados del azar razón que impide encontrar un pronóstico antes del

---

[65] Gujarati (2003, 771).

[66] Hace referencia a estacionamiento o parqueo.

[67] Función de distribución de probabilidad conjunta $p(Y_1, Y_2, .., Y_T)$; resultante de la colección de datos individuales Y_1, Y_2,.., Y_T. De esta misma forma la predicción un periodo adelante $\left(\hat{Y}_{T+1}\right)$ se obtiene por una función de distribución condicional $p\left(\hat{Y}_{T+1} | Y_1, Y_2, .., Y_T\right)$; en otras palabras, está sujeto a las observaciones pasadas (Pindyck y Rubinfeld, 1998, 519).

[68] Greene (1998, 612).

[69] Gujarati (2003, 772).

[70] Cualquier serie de tiempo con esta condición es impredecible, como los juegos de azar.

siguiente lanzamiento; dada su naturaleza totalmente estocástica. En pocas palabras, series de tiempo para loterías y juegos de azar son impredecibles bajo escenarios normales sin ningún tipo de fraude.

A partir del concepto expuesto sobre estacionariedad fuerte, se puede contextualizar un ruido blanco; definiéndolo, como una serie de tiempo netamente aleatoria la cual no se puede pronosticar. Dado que ambas comparten las mismas características: media, varianza y covarianza entre rezagos no condicionados con el tiempo.

Otro concepto relacionado con estacionariedad es la condición ergódica, cuyo cumplimiento se manifiesta cuando la media y varianza de una muestra (serie de tiempo) convergen a sus mismos momentos poblacionales. Dicha condición, resulta útil para distinguir entre variables predecibles y no pronosticables (series ruido blanco)[71]. Desafortunadamente, no existe un procedimiento formal que permita deducir si una serie de tiempo cumple o no la condición ergódica; por esta razón, en la práctica una variable débilmente estacionaria se asume como ergódica[72].

Para finalizar la conceptualización, las series de tiempo también pueden resultar no estacionarias (es lo más común) en media, varianza y covarianza; porque las mismas, se encuentran condicionadas con el tiempo. Este caso predomina en la mayor parte de series temporales (PED), razón por la cual a continuación se exponen métodos para detectar estacionariedad y en caso de no contar con esta característica, cuál es el tratamiento adecuado para conseguir que ella se convierta en débilmente estacionaria; con el fin de pronosticarla mediante la metodología BJ.

[71] Greene (2000, 721).
[72] Montenegro (2007, 11).

5.3 Métodos para detectar estacionariedad débil o fuerte (ruido blanco) y alternativas de conseguir variables con estacionariedad débil.

Teniendo en cuenta lo hasta ahora expuesto en el capítulo y abriendo paso hacia la metodología BJ para pronósticos univariados, en esta sección se considera el estudio de estacionariedad débil y fuerte a través del análisis gráfico, correlograma y pruebas de raíz unitaria de Dickey-Fuller; que son las más convencionales, en teoría y práctica, para detectarla.

Asimismo el correlograma para conocer si el PED es ruido blanco, aunque en esto también son utilizas las pruebas Box-Pierce y Ljung-Box. Enfatizando también sobre la transformación en primeras diferencias, muy empleada para conseguir que un PED se convierta en estacionario; cuando con las pruebas a continuación, previamente se ha determinado su ausencia.

5.3.1 Análisis gráfico para detectar estacionariedad

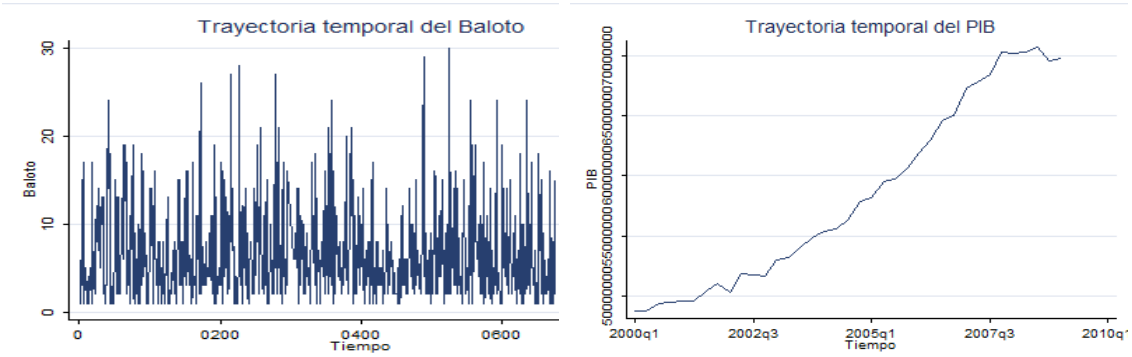
Uno de los métodos más sencillos para detectar que un PED puede ser fuertemente estacionario, es mediante una gráfica a través del tiempo con forma senoidal (cíclica), sin tendencia y movimientos similares a los de un electrocardiograma. Bajo estas condiciones, el comportamiento del baloto[73] (Y_t) -representado en la gráfica 5.1 (A)- es un PED fuertemente estacionario (ruido blanco).

En otras palabras, si el componente y naturaleza de la serie temporal es tendencial la variable no resulta estacionaria, porque bajo esta forma su media y varianza son inestables a lo largo del tiempo. Este comportamiento, puede apreciarse en la gráfica 5.1 (B) para el PIB colombiano (expuesto en el capítulo 4), el cual es un PED no estacionario; porque presenta tendencia creciente y no se observa la forma senoidal (cíclica).

[73] Juego de baloto colombiano en línea más comprado.

Gráfica 5.1. Condición de estacionariedad

(A) Baloto (B) PIB



Fuente: Baloto y Dane.

Aunque la gráfica es una primera aproximación, para determinar si el PED es o no estacionario, solo es un método exploratorio no concluyente; debido a que pueden existir periodos con y sin tendencia para el PED. Conduciendo a posibles conclusiones erróneas sobre la estacionariedad y gráficamente también es complejo conocer si es fuerte (ruido blanco) o débilmente estacionario. Para evitar estos juicios de valor y concluir objetivamente sobre la estacionariedad, también existe la prueba gráfica del correlograma y estadístico de raíz unitaria Dickey-Fuller; expuestos a continuación.

5.3.2 Análisis gráfico del correlograma para detectar estacionariedad y ruido blanco

Otra prueba gráfica derivada de la función de autocorrelación simple (FAS=$\hat{\rho}_p$) [74] para detectar estacionariedad, en una serie temporal o PED, se denomina correlograma. Donde FAS, depende de su función de autocovarianza ($\gamma_p = Cov[Y_t, Y_{t-p}]$) y se estima como lo expresa la ecuación 5.1. En otras palabras, es el cociente entre la función de autocovarianza y varianza ($Var[Y_t] = \gamma_0$) de Y_t. Cada valor de rho estimado $(\hat{\rho}_p)$ se gráfica, originando así el correlograma para FAS.

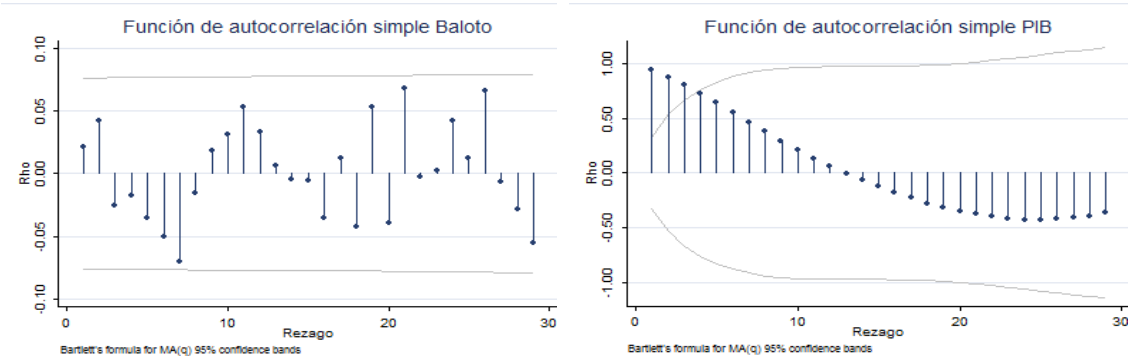
$$\widehat{\boldsymbol{\rho}}_{\boldsymbol{p}} = \frac{Cov[Y_t, Y_{t-p}]}{Var[Y_t]} = \frac{\gamma_p}{\gamma_0} \quad (5.1)$$

[74] El subíndice p señala la longitud del rezago o número de retardos analizados para ρ, para T=1, 2,…, $t - p$; generalmente equivale al 25% de la muestra. En otras palabras si n=80 entonces p es 20 ($p = 20$).

En la ecuación 5.1, $\hat{\rho}_p$ toma valores únicamente entre menos uno y uno $(-1 < \hat{\rho}_p < 1)$ y se comporta como una distribución normal $\left[\hat{\rho}_p \sim N(\mu, \sigma^2) \rightarrow \hat{\rho}_p \sim N\left(0, \frac{1}{n}\right)\right]$; con media igual a cero $(\mu = 0)$ y varianza $\left(\sigma^2 = \frac{1}{n}\right)$. A partir de esto, las gráficas 5.2 (A) y 5.2 (B) exponen el comportamiento del FAS, mediante el correlograma, para el baloto y PIB respectivamente; graficados en el aparte anterior.

Gráfica 5.2. Correlograma de la función de autocorrelación simple (FAS)

(A) Baloto (B) PIB



Fuente: Baloto, Dane y cálculos autores.

De esta manera, la gráfica 5.2 (A) muestra como el PED del baloto resulta estacionario, porque sus valores de $\hat{\rho}_p$ se mueven senoidalmente dentro de su intervalo de confianza[75]. Mientras la gráfica 5.2 (B), indica no estacionariedad para el PIB dado que exterioriza estimaciones $\hat{\rho}_p$ decrecientes exponencialmente de mayor a menor (entre 1 y -1), llegando a cero, tomando valores negativos y la mayor parte de ellos se encuentran fuera de su intervalo de confianza.

Figura 5.1. Valores del FAS y FAP

(A) Baloto

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial Autocor]
1	0.0208	0.0210	.28719	0.5920		
2	0.0420	0.0420	1.4624	0.4813		
3	-0.0259	-0.0287	1.9108	0.5911		
4	-0.0178	-0.0155	2.1221	0.7133		
5	-0.0355	-0.0379	2.9647	0.7054		
6	-0.0502	-0.0487	4.6572	0.5885		
7	-0.0705	-0.0728	8.0018	0.3324		
8	-0.0153	-0.0241	8.1585	0.4181		
9	0.0178	0.0244	8.3723	0.4971		
10	0.0307	0.0377	9.0106	0.5311		
11	0.0533	0.0568	10.929	0.4493		
12	0.0328	0.0387	11.656	0.4737		
13	0.0058	0.0001	11.679	0.5541		
14	-0.0051	-0.0096	11.697	0.6307		
15	-0.0053	-0.0135	11.715	0.7004		
16	-0.0357	-0.0297	12.582	0.7030		
17	0.0123	0.0300	12.685	0.7570		
18	-0.0428	-0.0218	13.937	0.7332		
19	0.0532	0.0568	15.875	0.6656		
20	-0.0394	-0.0834	16.94	0.6569		
21	0.0677	0.0299	20.088	0.5157		
22	-0.0022	-0.0020	20.092	0.5773		
23	0.0026	0.0176	20.096	0.6361		
24	0.0418	0.0434	21.305	0.6207		
25	0.0121	0.0229	21.405	0.6698		
26	0.0657	0.0611	24.393	0.5535		
27	-0.0063	0.0015	24.421	0.6069		
28	-0.0281	-0.0234	24.97	0.6295		
29	-0.0553	-0.0737	27.097	0.5665		

(B) PIB

LAG	AC	PAC	Q	Prob>Q	[Autocorrelation]	[Partial Autocor]
1	0.9412	1.0052	35.511	0.0000		
2	0.8783	0.1043	67.316	0.0000		
3	0.8048	-0.2125	94.806	0.0000		
4	0.7241	-0.7382	117.73	0.0000		
5	0.6402	-0.1761	136.21	0.0000		
6	0.5479	-0.3655	150.19	0.0000		
7	0.4635	0.0103	160.52	0.0000		
8	0.3787	-0.5961	167.66	0.0000		
9	0.2893	0.1294	171.97	0.0000		
10	0.2139	0.2376	174.42	0.0000		
11	0.1351	-0.0998	175.43	0.0000		
12	0.0620	0.2574	175.65	0.0000		
13	-0.0051	-0.1551	175.65	0.0000		
14	-0.0673	-0.7149	175.94	0.0000		
15	-0.1245	-1.2071	176.96	0.0000		
16	-0.1797	-0.3723	179.17	0.0000		
17	-0.2280	.	182.92	0.0000		
18	-0.2768	.	188.74	0.0000		
19	-0.3151	.	196.7	0.0000		
20	-0.3462	.	206.88	0.0000		
21	-0.3753	.	219.58	0.0000		
22	-0.3967	.	234.72	0.0000		
23	-0.4154	.	252.51	0.0000		
24	-0.4256	.	272.61	0.0000		
25	-0.4301	.	294.85	0.0000		
26	-0.4241	.	318.44	0.0000		
27	-0.4111	.	342.83	0.0000		
28	-0.3976	.	368.18	0.0000		
29	-0.3657	.	392.31	0.0000		

Fuente: cálculos autores.

Por otra parte, las figuras 5.1 (A) y 5.1 (B) exhiben los valores del FAS ($\hat{\rho}_p$, *véanse* columnas AC)[76], graficados en los correlogramas anteriores, para el baloto (senoidal entre 0.0208 y -0.0553) y PIB colombiano (decrecen exponencialmente, desde 0.9412 hasta -0.3617) respectivamente. En ellos, numéricamente pueden apreciarse los comportamientos y conclusiones sobre estacionariedad descritos a partir de las gráficas 5.2 (A) y 5.2 (B).

[75] De acuerdo a las propiedades para $\hat{\rho}_p \sim N(\mu, \sigma^2) \rightarrow \hat{\rho}_p \sim N\left(0, \frac{1}{n}\right)$ y de la distribución normal estándar, este intervalo de confianza del 95% para cada rho estimado equivale a $prob\left(\hat{\rho}_p - 1{,}96 * \sqrt{\frac{1}{n}} \leq \rho_p \leq \hat{\rho}_p + 1{,}96 * \sqrt{\frac{1}{n}}\right) = 0{,}95$.

[76] AC (Autocorrelation, sigla en inglés).

Además, también se observan los estimadores de $\hat{\rho}_p$[77] (*véanse* columnas PAC)[78] para la función de autocorrelación parcial (FAP) y prueba Ljung-Box (*véanse* columnas $\boldsymbol{Q}$)[79]; adicionalmente, los resultados descritos para FAS y $\boldsymbol{Q}$ ayudan a comprobar si el PED es ruido blanco. Por el momento, FAP no juega un papel predominante en este análisis.

$$\boldsymbol{FAS} \rightarrow Y_t = \rho_0 \pm \rho_1 Y_{t-1} + u_{t1};\ Y_t = \rho_0 \pm \rho_2 Y_{t-2} + u_{t2}\ ;..;\ Y_t = \rho_0 \pm \rho_p Y_{t-p} + u_{tp} \quad (5.2)$$

$$\boldsymbol{FAP} \rightarrow Y_t = \rho_0 \pm \rho_1 Y_{t-1} \pm \rho_2 Y_{t-2} \pm .. \pm \rho_p Y_{t-p} + u_t \quad (5.3)$$

No obstante, la diferencia entre FAS y FAP puede observarse en las ecuaciones 5.2 y 5.3; el primero para cada rezago es semejante a un modelo de regresión lineal simple y la segunda a uno múltiple, aunque ninguna es función de variables independientes y sus respectivos estimadores $(\hat{\rho}_0, \hat{\rho}_1, \hat{\rho}_2, .., \hat{\rho}_p)$ son obtenidos empleando MCO, su comportamiento también se exponen en un correlograma; esto quiere decir que se conciben dos correlogramas, uno para el FAS y otro del FAP (*véase* figura 5.1 (A) y 5.1 (B)) y u_t hace referencia al error (perturbación aleatoria del modelo).

De esta forma y una vez los coeficientes estimados del FAS, por medio de ellos, se establece si la serie Y_t (baloto y PIB) o PED resulta ruido blanco o fuertemente estacionario; empleando una prueba de significancia conjunta, que determine si el conjunto de rho estimados $(\hat{\rho}_0, \hat{\rho}_1, \hat{\rho}_2, .., \hat{\rho}_p)$ estadísticamente equivalen a cero; hipótesis (*véase* 5.4 y 5.5) soportada en las pruebas $\boldsymbol{Q}$ de Box-Pierce y Ljung-Box – $\boldsymbol{LB}$– (*véase* ecuaciones 5.6 y 5.7).

H_0: $\hat{\rho}_1 = \hat{\rho}_2 = \cdots = \hat{\rho}_p = 0$; la serie Y_t o PED (baloto/PIB) es ruido blanco, derivando automaticamente en estacionariedad fuerte para ella (donde *p*=29, *véase* lag figura 5.1). (5.4)

H_1: $\hat{\rho}_1 \neq \hat{\rho}_2 \neq \cdots \neq \hat{\rho}_p \neq 0$; la serie Y_t o PED (baloto/PIB) no es ruido blanco, involucrando que posiblemente puede ser debilmente estacionaria. (5.5)

[77] Observe que los valores $\hat{\rho}_p$ para el FAS y FAP son diferentes, en otras palabras $\hat{\rho}_{p\_FAS} \neq \hat{\rho}_{p\_FAP}$.
[78] PAC (Partial Autocorrelation, sigla en inglés).
[79] Estadístico $\boldsymbol{Q}$ de Box Pierce.

$$\boldsymbol{Q} = n\sum_{t=1}^{p}\hat{\rho}_p^2 \rightarrow \boldsymbol{Q}\sim\chi_p^2 \text{ (5.6)}$$

$$\boldsymbol{LB} = n(n+2)\sum_{t=1}^{p}\left(\frac{\hat{\rho}_p^2}{n-p}\right) \rightarrow \boldsymbol{LB}\sim\chi_p^2 \text{ (5.7)}$$

En las ecuaciones 5.6 y 5.7 n equivale al tamaño de la muestra, $\hat{\rho}_p^2$ cada estimador al cuadrado y p al número de rezago calculado y evaluado, ambos siguen una distribución chi-cuadrado $\left(\boldsymbol{Q}, \boldsymbol{LB}\sim\chi_p^2\right)$ con p grados de libertad (número de rezago calculado y evaluado). Si el valor calculado de $\boldsymbol{Q}$ o $\boldsymbol{LB}$ excede el critico χ_p^2[80], es rechazada H_0, concluyendo que Y_t es ruido blanco y por tanto es posible pronósticarla; caso contrario no se rechaza H_0 cuando estos valore son menores al critico, sin embargo aunque no es ruido blanco tampoco implica que Y_t es estacionaria.

Estas estipulaciones, expuestas en las hipótesis nula (H_0) y alterna (H_1), concluyen que toda variable dinamica ruido blanco es fuertemente estacionaria, por ende no predecible; pero no todo PED estacionario es ruido blanco; debido a la condición de estacionariedad debil que esta puede presentar para poder pronósticarla. De acuerdo con esto y los resutados en la figura 5.1 (A), no es rechazada H_0 para el baloto; estableciendo que es ruido blanco (fuertemente estacionaria), no pronósticable. Mientras el PIB no es ruido blanco porque es rechazada H_0; adicionalmente tampoco es estacionario, de acuerdo con el comportamiento tendencial en la gráfica 5.1 (B) y caida exponencial del FAS (*veáse* correlograma gráfico 5.2 (B) y valores autocorrelation figura 5.1 (B)).

Aunque lo anterior condujó a conclusiones preliminares sobre la estacionaredad para Y_t, el correlograma FAS también resulta una prueba exploratoria y no concluyente en este aspecto. Razon, por la que se prosigue con el análisis de raíz unitaria de Dickey-Fuller.

[80] De igual manera si es tomada la decisión con la probabilidad $\boldsymbol{Q}$ o $\boldsymbol{LB}$ (*véase* figura 5.1, columna prob>Q), comparada contra su nivel de significancia (α) al 1, 5 o 10 por ciento respectivamente; si prob < α → rechazo H_0, prob > α → no rechazo H_0.

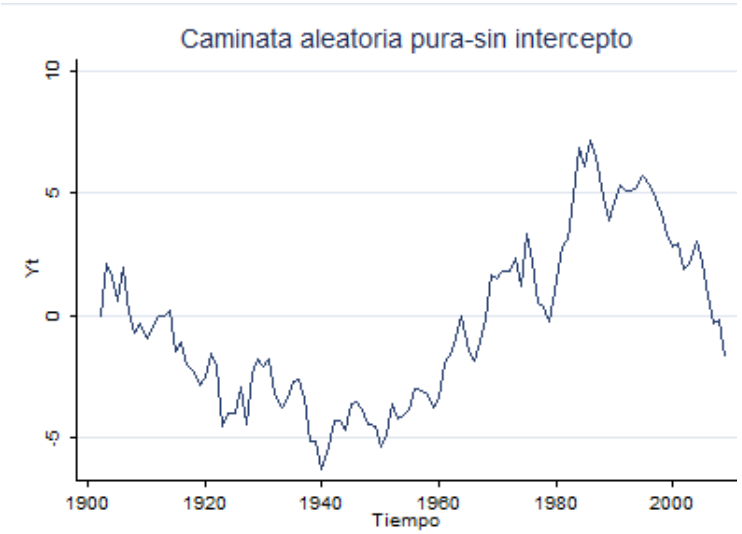
5.3.3 Análisis de raíz unitaria Dickey-Fuller (DF) para detectar estacionariedad

Una prueba estadística formal concluyente para estacionariedad es la de raíz unitaria, desarrollada por Dickey-Fuller[81]. Previo abordarla analíticamente, en primera instancia se parten de los conceptos caminata aleatoria (CA) y proceso integrado (PIN), con el fin de poseer elementos teóricos necesarios para su respectivo entendimiento y aplicación.

$$Y_t = Y_{t-1} + u_t \text{ (5.8), dado que } Y_t = \mu + Y_{t-1} + u_t,\ \mu = 0 \text{ y } \hat{\rho} = 1$$

Así, la forma más sencilla de representar una caminata aleatoria (*véase* ecuación 5.8) está dada por el cambio sucesivo[82] en Y_t, con media cero ($\mu = 0$) más el termino del error (u_t) o perturbación aleatoria del modelo (Pindyck y Rubinfeld, 1998, 519). La ecuación 5.8, indica que la variable Y_t es igual a su valor rezagado un periodo (Y_{t-1}) más un choque aleatorio (u_t)[83].

Gráfica 5.3. Caminata aleatoria pura, sin intercepto (media igual a cero)



Fuente: cálculos autores, serie de tiempo hipotética con frecuencia anual entre 1902 y 2009.

[81] Gujarati (2003, 788).

[82] $(Y_{t-1}, Y_{t-2}, Y_t, Y_{t+1}, Y_{t+2}, .., Y_{t+p})$

[83] Donde el valor esperado del error es igual a cero $\{E(u_t) = 0, E(u_1) = 0, .., E(u_p) = 0\}$, igualmente para sus covarianzas entre los rezagos de ella $\{E(u_t u_p)\} = 0$ para $t \neq p$.

Adicionalmente en la ecuación 5.8, se puede determinar que su media es constante ($\mu = 0$), pero varianza y covarianza están condicionadas con t $\{(\gamma_0 = t\gamma_0), [\gamma_p = \gamma_{0_{t-1}}]\}$. Estas dos últimas características señalan que la *caminata aleatoria* (*véase* gráfica 5.3), expuesta en la ecuación 5.8, no es un proceso estacionario (*véase* demostración en el anexo 5.1).

$$Y_{t+1} = Y_t + E(u_{t+1}) = Y_t \quad (5.9)$$

$$Y_{t+2} = Y_{t+1} + E(u_{t+2}) = Y_t + E(u_{t+1}) + E(u_{t+2}) = Y_t \quad (5.10)$$

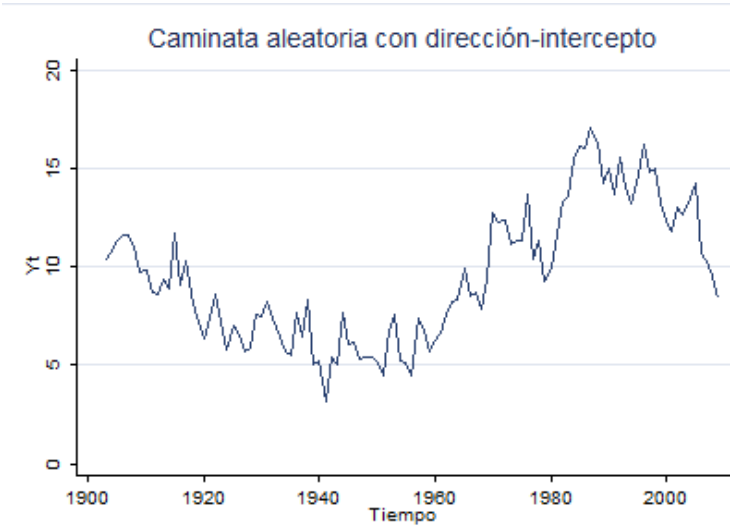
$$Y_{t+3} = Y_{t+2} + E(u_{t+3}) = Y_t + E(u_{t+1}) + E(u_{t+2}) + E(u_{t+3}) = Y_t \quad (5.11)$$

De este modo, los choques estocásticos ($u_{t+1}, u_{t+2}, u_{t+3}$), en una *caminata aleatoria* son persistentes; caracterizándola así por tener memoria infinita[84] y el pronóstico ($\hat{Y}_{t+p}$) siempre convergerá a Y_t (*véase* ecuaciones 5.9-5.11). Otro tipo de *caminata aleatoria* es la que lleva dirección (tendencia creciente o decreciente en Y_t)[85], recuerde que el componente tendencial describe el comportamiento de la media (μ) en Y_t. Ésta, es denominada como *caminata aleatoria con deriva, media, variaciones o intercepto* ($\mu = \alpha$, *véase* ecuación 5.12 y gráfica 5.4).

$$Y_t = \alpha + Y_{t-1} + u_t \quad (5.12)$$

[84] $(Y_{t-p}, .., Y_{t-2}, Y_{t-1}, Y_t, Y_{t+1}, Y_{t+2}, .., Y_{t+p})$.
[85] Creciente si $\delta > 0$ y decreciente $\delta < 0$.

Gráfica 5.4. Caminata aleatoria con variaciones (intercepto)



Fuente: cálculos autores, serie de tiempo hipotética con frecuencia anual entre 1902 y 2009.

Adicional a las dos anteriores, también se presentan otra clase de *caminatas aleatorias*, destacando las que contienen raíz unitaria[86] ($\rho = 1$), tendencia estacionaria[87] (βT_t) y diferencia estacionaria (ΔY_t). Las mismas, se encuentra definidas en las ecuaciones 5.13, 5.14 y 5.15 respectivamente. En esta última, se puede apreciar que la primera diferencia es igual al ruido blanco (u_t, término aleatorio o error); en otras palabras la primera diferencia (ΔY_t)[88] de Y_t es estacionaria.

$$Y_t = \rho Y_{t-1} + u_t \quad (5.13), \text{ donde } -1 \leq \hat{\rho} \leq 1$$

$$Y_t = \delta + \beta T_t + \rho Y_{t-1} + u_t \quad (5.14)$$

$$Y_t = Y_{t-1} + u_t \Rightarrow Y_t - Y_{t-1} = u_t \Rightarrow \Delta Y_t = u_t \quad (5.15)$$

[86]Tenga en cuenta que a diferencia de la caminata aleatoria pura expuesta (ecuación 5.6), ahora aparece el coeficiente rho (ρ) multiplicando el primer rezago (Y_{t-1}) de Y_t.

[87]En conclusión una caminata aleatoria, raíz unitaria y un proceso no estacionario son considerados equivalentes (sinónimos).

[88]Todos los aspectos sobre primera diferencia y ecuaciones en diferencia se encuentran en el anexo 5.1, es recomendable su lectura previa antes de continuar con lectura de este capítulo.

Teniendo en cuenta la ecuación 5.12, δ y β son los parámetros del modelo en la ecuación 5.14 que representan media (μ) y constante de tendencia (T_t) respectivamente; u_t hace referencia al error (perturbación aleatoria o ruido blanco en cada ecuación). Una vez definido el concepto de camita aleatoria, expuesto hasta el momento, es abordado a continuación el tema de proceso integrado (PIN); para posteriormente realizar el análisis sobre estacionariedad mediante la prueba de raíz unitaria Dickey-Fuller (DF).

Considerando lo anterior, un proceso integrado se refiere al orden de diferenciación donde la serie temporal resulta por lo menos débilmente estacionaria; entonces si la variable debe diferenciarse d veces para lograr estacionariedad, la serie es integrada de orden d $[Y_t \sim \mathrm{I}(d)]$. En otras palabras, cuando ΔY_t es estacionario significa que la serie Y_t es integrada de orden uno $[Y_t \sim \mathrm{I}(1)]$ o su primera diferencia es integrada de orden cero $[\Delta Y_t \sim \mathrm{I}(0)]$. Si la serie en su nivel (Y_t) resulta estacionaria, sin diferenciarla, se denomina integrada de orden cero $[Y_t \sim \mathrm{I}(0)]$.

Comprendidos los conceptos sobre caminata aleatoria y proceso integrado, se prosigue con el análisis de estacionariedad mediante raíz unitaria Dickey-Fuller (DF). Ésta prueba es considerada concluyente, porque permite realizar una hipótesis (*véase* 5.17 y 5.18) formal para probar si la serie temporal es o no estacionaria; tomando inicalmente el PED o CA con raíz unitaria de la ecuación 5.13 y restando en ambos lado Y_{t-1} resulta la ecuación 5.16: en ella, $\delta = \rho - 1$ y equivale al coeficiente del modelo.

$$Y_t - Y_{t-1} = \rho Y_{t-1} + u_t - Y_{t-1} \rightarrow \Delta Y_t = (\rho - 1)Y_{t-1} + u_t \rightarrow \Delta Y_t = \delta Y_{t-1} + u_t \quad (5.16)$$

H_0: $\delta = 0$; $\rho = 1$; la serie Y_t (baloto/PIB) contiene raíz unitaria, equivale a decir que es una caminata aleatoria o simplemente no es estacionaria. (5.17)

H_1: $\delta \neq 0$; $\rho \neq 1$; la serie Y_t (baloto/PIB) no contiene raíz unitaria equivale a decir que no es una caminata aleatoria o simplemente es estacionaria. (5.18)

Dada la ecuación 5.16 e hipótesis 5.17 y 51.8, ésta (H_0) es o no rechazada por medio del estadisco tau[89] (τ); en este caso τ debe tomar un valor negativo, porque $-1 \leq \hat{\rho} \leq 1$, para ser comparado en valor absoluto $(|\tau|)$ con los valores criticos de MacKinnon[90] al 1, 5 y 10 por ciento, también en valor absoluto. Caso contrario si τ resulta positivo, significa que $\hat{\rho} > 1$ y por ende Y_t es una serie de tiempo explosiva, no estacionaria; bajo esta situación para τ, no se hace necesario realizar la prueba de hipótesis sobre estacionariedad de Y_t dado que automaticamente no es estacionaria.

Ahora con τ negativo, el criterio para decidir rechazar H_0 es aplicando $|\tau|$ y comparlo contra el valor absoluto de la tabla[91] MacKinnon; si resulta mayor al 1, 5 y 10 por ciento, debe rachazarse H_0, concluyendo así que Y_t cumple la condición de estacionariedad debil (en media y varianza). Contrariamente, no se rechaza H_0 cuando $|\tau|$ es menor a estos valore.

No obstante, la decisión de rechazar H_0 siempre se realiza con el estadístico τ porque las pruebas parciales convencionales (con t-student) resultan inadecuadas para concluir sobre la hipótesis de estacionariedad para Y_t; debido al sesgo en el estimador $(\hat{\delta})$. Adicional a esto, también se pueden cometer error tipo I o II en la hipótesis cuestionada; por el sesgo de especificación implícito en el modelo de caminata aleatoria con raíz unitaria (ecuación 5.16) dado que el mismo puede contener tendencia e intercepto y termino del error (u_t) puede estar correlacionado con sí mismo.

$$\Delta Y_t = \delta Y_{t-1} + \beta_1 \Delta Y_{t-1} + \beta_2 \Delta Y_{t-2} + \cdots + \beta_p \Delta Y_{t-p} + u_t \rightarrow \Delta Y_t = \delta Y_{t-1} + \sum_{\beta=1}^{p} \beta_p \Delta Y_{t-p} + u_t \quad (5.19)$$

Por lo anterior, es necesario especificar y probar las caminatas aleatorias con intercepto y tendencia (*véase* cuadro 5.1) de las ecuaciones 5.12 y 5.14, además aplicar la prueba Dickey-Fuller aumentada (DFA) para corregir el problema de autocorrelación residual. Ésta última, consiste en involucrar rezagos

[89] $\tau = \frac{\hat{\delta}}{s_{\hat{\delta}}}$. El parámetro estimado $(\hat{\delta})$, sobre su error estándar $(S_{\hat{\delta}})$, se le atribuye a Dickey y Fuller (1979) y se compara con los valores críticos de la tabla de MacKinnon.

[90] *Véase* los valores en Engle y Granger (1991, cap. 13).

[91] *Ibid. Op .Cit.*

$(\Delta Y_{t-1}, \Delta Y_{t-2}, .., \Delta Y_{t-p})$ de la variable dependiente (ΔY_t) en DF (*véase* ecuación 5.19). Con este fin, el cuadro 5.1 contiene las formas DF y DFA para ayudar a determinar estacionariedad.

Cuadro 5.1. Formas de los modelos para realizar pruebas de raíz unitaria (DF y DFA).

Orden de integración para la serie	Tipo de modelo	Forma del modelo $\boldsymbol{\delta = \rho - 1}$
0	No contiene intercepto y tendencia	$\Delta Y = \delta Y_{t-1} + u_t$
0	Contiene intercepto pero no tendencia	$\Delta Y_t = \alpha + \delta Y_{t-1} + u_t$
0	Contiene intercepto y tendencia	$\Delta Y_t = \alpha + \delta Y_{t-1} + \beta t + u_t$
0	Aumentado con intercepto y tendencia	$\Delta Y_t = \alpha + \delta Y_{t-1} + \beta_1 t + \sum \beta_p \Delta Y_{t-p} + u_t$
1	No contiene intercepto y tendencia	$\Delta^2 Y_t = \delta \Delta Y_{t-1} + u_t$
1	Contiene intercepto pero no tendencia	$\Delta^2 Y_t = \alpha + \delta \Delta Y_{t-1} + u_t$
1	Contiene intercepto y tendencia	$\Delta^2 Y_t = \alpha + \delta \Delta Y_{t-1} + \beta t + u_t$
1	Aumentado con intercepto y tendencia	$\Delta^2 Y_t = \alpha + \delta \Delta Y_{t-1} + \beta_1 t + \sum \beta_p \Delta^2 Y_{t-p} + u_t$
d	No contiene intercepto y tendencia	$\Delta^{d+1} Y_t = \delta \Delta^d Y_{t-1} + u_t$
d	Contiene intercepto pero no tendencia	$\Delta^{d+1} Y_t = \alpha + \delta \Delta^d Y_{t-1} + u_t$
d	Contiene intercepto y tendencia	$\Delta^{d+1} Y_t = \alpha + \delta \Delta^d Y_{t-1} + \beta t + u_t$
d	Aumentado con intercepto y tendencia	$\Delta^{d+1} Y_t = \alpha + \delta \Delta^d Y_{t-1} + \beta_1 t + \sum \beta_p \Delta^{d+1} Y_{t-p} + u_t$

Fuente: autores, a partir de Mendieta y Perdomo (2008, 129), Pulido y Pérez (2001), Gujarati (2003) y Greene (1998).

En el cuadro 5.1, $\Delta Y_t = Y_t - Y_{t-1}$, $\Delta^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2}$ y $\Delta^{d+1} Y_t$ son ecuaciones en primera, segunda y *d+1* diferencias[92]; δ es el parámetro asociado a los rezagos (Y_{t-1} *o* ΔY_{t-1}), en los periodos inmediatamente anteriores; α intercepto; t tendencia; β coeficientes que acompañan a las demás variables en cada modelo. Si δ estadísticamente (de acuerdo con $|\tau|$) toma valor de cero, se concluye que la serie tiene raíz unitaria o no es estacionaria. Ante esto, se prosigue con la transformación de una serie a estacionaria (cuando $\delta = 0$) y orden de integración.

5.3.4 Transformación de una serie no estacionaria a estacionaria y orden de integración

Ante posibles conclusiones, de no estacionariedad[93] para una serie temporal, en las pruebas gráficas (trayectoria en el tiempo y FAS) y raíz unitaria Dickey-Fuller (aumentada) presentadas, es necesario recurrir algún proceso de transformación con el fin de obtenerla estacionaria y así llevar a cabo la metodologia Box-Jenkins (BJ); cuya característica principal es que Y_t resulte débilmente estacionaria (no ruido blanco).

Para lo anterior y cuando Y_t no es estacionaria en media (μ) y varianza (γ_0), porque contiene tendencia[94], ella puede ser removida mediante la diferenciación ($\Delta^d Y_t$) en algún orden para Y_t, hasta conseguir su estacionariedad. Aunque esta practica generalmente corrige la media, en ocaciones su varianza sigue condicionada con el tiempo; logrando que la serie resulte estacionaria en media pero no en varianza.

$$Y_t = Y_{t-1} + u_t \Rightarrow Y_t - Y_{t-1} = u_t \Rightarrow \Delta Y_t = u_t \Rightarrow (1 - L)Y_t = u_t \quad (5.20)$$

$$(1 - L)^d Y_t = u_t \quad (5.21)$$

$$LNY_t = LNY_{t-1} + u_t \Rightarrow LNY_t - LNY_{t-1} = u_t \Rightarrow \Delta LNY_t = u_t \Rightarrow (1 - L)LNY_t = u_t \quad (5.22)$$

$$(1 - L)^d LNY_t = u_t \quad (5.23)$$

[92]Todos los aspectos sobre primera diferencia y ecuaciones en diferencia se encuentran en el anexo 5.3, es recomendable su lectura previa antes de continuar con lectura de este capítulo.
[93]Esta característica prima sobre las series de tiempo que se trabajan en la práctica.
[94]Bajo tendencia su media y varianza esta condicionadas con el tiempo (*t*).

En este último caso, aplicar logaritmo a la variable (LNY_t) y adicionalmente la diferencia en logaritmo $(\Delta^d LNY_t)$, logra obtener a Y_t estacionaria en media y varianza. Las primeras difencias (ΔY_t) económicamente representa el valor en que subio o disminuyó la variable entre un periodo y otro; mientras la primera diferencia logartimica (ΔLNY_t) obtiene la tasa de crecimiento o disminución porcentual (*véase* ecuaciones[95] 5.20-5.23).

Si la variable Y_t resulta integrada de orden uno $[Y_t \sim \mathrm{I}(1)]$, estacionaria en sus primeras diferencias, es porque la serie contiene una raíz unitaria. Si es integrada de orden dos $[Y_t \sim \mathrm{I}(2)]$, Y_t involucra dos raíces unitarias y así sucesivamente si es integrada de orden d $[Y_t \sim \mathrm{I}(d)]$; tiene implícitas d raíces unitarias y se deberá diferenciar d veces (Gujarati, 2003, 794).

Sin embargo, las pruebas para detectar estacionariedad (gráfica, FAS y DFA) y respectivo orden de diferenciación o integración, donde resulta ser estacionaria Y_t, aplican en series que no contienen componente estacional; el tratamiento sobre desestacionalización, estacionariedad y predicción para variables con estacionalidad son expuestas en la sección 5.5. Una vez desarrollado el tema de estacionariedad, con el fin de aplicar la metodología de pronóstico Box-Jenkins para una serie no estacional, a continuación se expone ésta técnica bajo el contexto de modelos univariados ARIMA.

5.4 Modelos univariados (Arima) y metodología Box-Jenkins.

Los componentes tendencia y ciclo (abarcados hasta el momento) son elementos determinísticos o semideterminísticos en una serie temporal; mientras el irregular es su característica aleatoria o estocástica (no determinística). Consideración que dificulta un poco más los análisis para predecir una variable recopilada a través del tiempo, haciendo más complejos los tratamientos determinísticos tendencia y ciclo.

En el mismo contexto, los métodos de suavizamiento exponencial del capítulo 4 no consideran su elemento estocástico (irregular); ante esto, los modelos univariados

[95]Todos los aspectos sobre operador de rezago (L), polinomios de rezago $\{A(L)\}$ y su relación con ecuaciones en diferencia se encuentran en el anexo 5.2, es recomendable su lectura previa antes de continuar con lectura de este capítulo.

Arima (Autoregressive Intregrated Moving Average, siglas en inglés), incursionado por Box-Jenkins (BJ), incluyen los componentes determinísticos -tendencia y ciclo- y aleatorio (irregular) para proyectar Y_t. A partir de procesos autorregresivos (AR, Autoregressive, siglas en inglés) y media móvil (MA, Moving Average, siglas en inglés)

De esta forma, la metodología Box-Jenkins pretende obtener la predicción $(\hat{Y}_{t+p})$[96], de un PED débilmente estacionario, en el corto plazo (máximo tres periodos[97]); a partir de una muestra representativa (más de 30 datos[98] $-n-$). Básicamente, el empleo BJ, se puede resumir de la siguiente manera:

1. Familiarizarse con la serie: conocer el contexto histórico y políticas que han afectado la serie a lo largo del tiempo con el fin de conocer los posibles cambios estructurales surgidos por distintos fenómenos económicos, política o metodología de medición o recolección del PED.
2. Análisis de estacionariedad: es el primer análisis a realizar con esta metodología, la variable de estudio debe resultar débilmente estacionaria[99] -no ruido blanco[100]-; aplicando las pruebas: gráficas (trayectoria temporal y correlograma del FAS), Q de Ljung-Box y raíz unitaria (DFA). Asimismo, determinar el orden de integración para la variable en cuestión[101].
3. Identificar el proceso generador de datos (PGD), sobre los correlogramas (FAS y FAP) de la variable estacionaria, a través de los términos autorregresivos (AR) y medía móvil (MA).

[96] Donde $\hat{Y}$ se refiere a los nuevos valores pronosticados y $t+p$ es el subíndice que representa los p periodos futuros proyectados.
[97] Representados en los $p = 3$ periodos futuros proyectados.
[98] Pulido y Pérez (2001, 643).
[99] Su media aritmética y varianza no están condicionadas con el tiempo.
[100] La media aritmética, varianza y covarianza entre sus rezagos no están condicionadas con el tiempo. Es totalmente aleatoria y no se puede pronosticar bajo estructura Arima; dado que no permite conocer el proceso generador de datos (AR y MA), para la variable aleatoria ordenada en el tiempo.
[101] Integrada de orden cero (estacionaria en su nivel original), uno (estacionaria en primeras diferencias $\Delta Y_t = Y_t - Y_{t-1}$) o dos (estacionaria en segundas diferencias $\Delta^2 Y_t = Y_t - 2Y_{t-1} + Y_{t-2}$).

4. Especificación y estimación[102] del modelo AR, MA, ARMA (para variables integradas de orden cero), ARIMA (para variables integrada de algún orden) o SARIMA (ARIMA estacional, *véase* sección 5.5).
5. Validación del modelo especificado y estimado: pruebas de significancia parcial (con el estadístico Z de la distribución normal estándar) y global (con el estadístico Wald o razón de verosimilitud), raíces invertidas de AR y MA (menores a 1)[103], residuales ruido blanco (no autocorrelacionados), distribución normal en residuales (*iid*, idéntica e independientemente distribuidos) y ausencia de heteroscedasticidad en la varianza residual.
6. Realizar el pronóstico, de corto plazo, con el modelo correctamente especificado y validado mediante las pruebas del numeral cinco.
7. Validación de la predicción $\left(\hat{Y}_t \; o \; \hat{Y}_{t+p}\right)$: mediante la gráfica de los valores observados Vs proyectados e indicadores de error para el pronóstico (*véase* cuadro 4.4 en el capítulo 4): el promedio del valor absoluto del error (PVAE), promedio del error al cuadrado (PEC), porcentaje del promedio del valor absoluto del error (PPVAE), raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC) y coeficiente de Theil (CT).
8. Utilización de los resultados para tomar decisiones.

A partir de lo anterior se cubren, en detalle desde el numeral tres, cada uno de los incisos presentados. Además, al final de esta sección es presentado un esquema que resume el proceso de la metodología BJ.

5.4.1 Identificación de términos AR, MA, ARMA y Arima como proceso generador de datos

Los modelos univariantes (Arima) se denominan modelos de series de tiempo estocásticas, en otras palabras, ellos suministran descripciones sobre la naturaleza aleatoria del proceso generador de la muestra (o datos, PGD). Esta descripción, no es resultado de una relación causa efecto (como los modelos de regresión

[102]Mediante métodos de mínimos cuadrados no lineales, máxima verosimilitud o Yulke Walker si el proceso es únicamente AR (autorregresivo), *véase* inciso 5.4.2.
[103]*Véase* más detalles anexo 5.4.1 y 5.5.

convencionales), sino es una función que añade la aleatoriedad del proceso (Pindyck y Rubinfeld, 1998, 514).

Por esta razón, los procesos autorregresivos (AR), media móvil (MA) y comprobacion de sus parámetros (ϕ y θ), dentro del circulo unitario[104], son utilizados para construir modelos univariados (Arima) de series de tiempo. Los cuales, buscan explicar el movimiento de Y_t y asi poder pronósticarla $\left(\hat{Y}_{t+p}\right)$; a partir de su propio pasado (AR[105], $Y_{t-1}, Y_{t-2}, .., Y_{t-p}$, *véase* ecuación 5.24), rezagos del error (MA[106], $u_t, u_{t-1}, u_{t-2}, .., u_{t-q}$, *véase* ecuación 5.25) o combinación de ambos (ARMA, $Y_{t-1}, Y_{t-2}, .., Y_{t-p}, u_t, u_{t-1}, u_{t-2}, .., u_{t-q}$, *véase* ecuación 5.26).

$$Y_t = f\left(Y_{t-1}, Y_{t-2}, .., Y_{t-p}\right) \text{ (5.24)}$$

$$Y_t = f\left(u_t, u_{t-1}, u_{t-2}, .., u_{t-q}\right) \text{ (5.25)}$$

$$Y_t = f\left(Y_{t-1}, Y_{t-2}, .., Y_{t-p}, u_t, u_{t-1}, u_{t-2}, .., u_{t-q}\right) \text{ (5.26)}$$

Contiguamente, el orden de integración (I) de la serie (Y_t) y el PGD (AR, MA o ARMA) especifica el modelo a estimar. Así, $Y_t \sim$ I(0) -debilmente estacionaria en niveles- posiblemente su estructura para realizar la predicción es AR (p), MA (q) o ARMA (p, q). Ahora, si $Y_t \sim$ I(1) o $Y_t \sim$ I(d) -estacionaria en primeras diferencias o se diferencio d veces para obtener su estacionariedad- seguramente tendrá una forma Arima (p, d[107], q), ARI (p, d) o IMA (d, q).

Para lo anterior, supongamos que una vez realizadas la pruebas de estacionariedad, de la sección 5.3, se determina que Y_t es débilmente estacionaria en su nivel (integrada de orden cero $Y_t \sim$ I(0)). Así que la estructura de su proceso generedor de datos (PGD) puede estar dado por un AR (p), MA (q) o ARMA (p,q)

[104]Es la región factible donde ϕ y θ se encuentra entre -1 y 1, garantizando estacionariedad para Y_t e invertibilidad del proceso, como se muestra en el anexo 5.4.1.

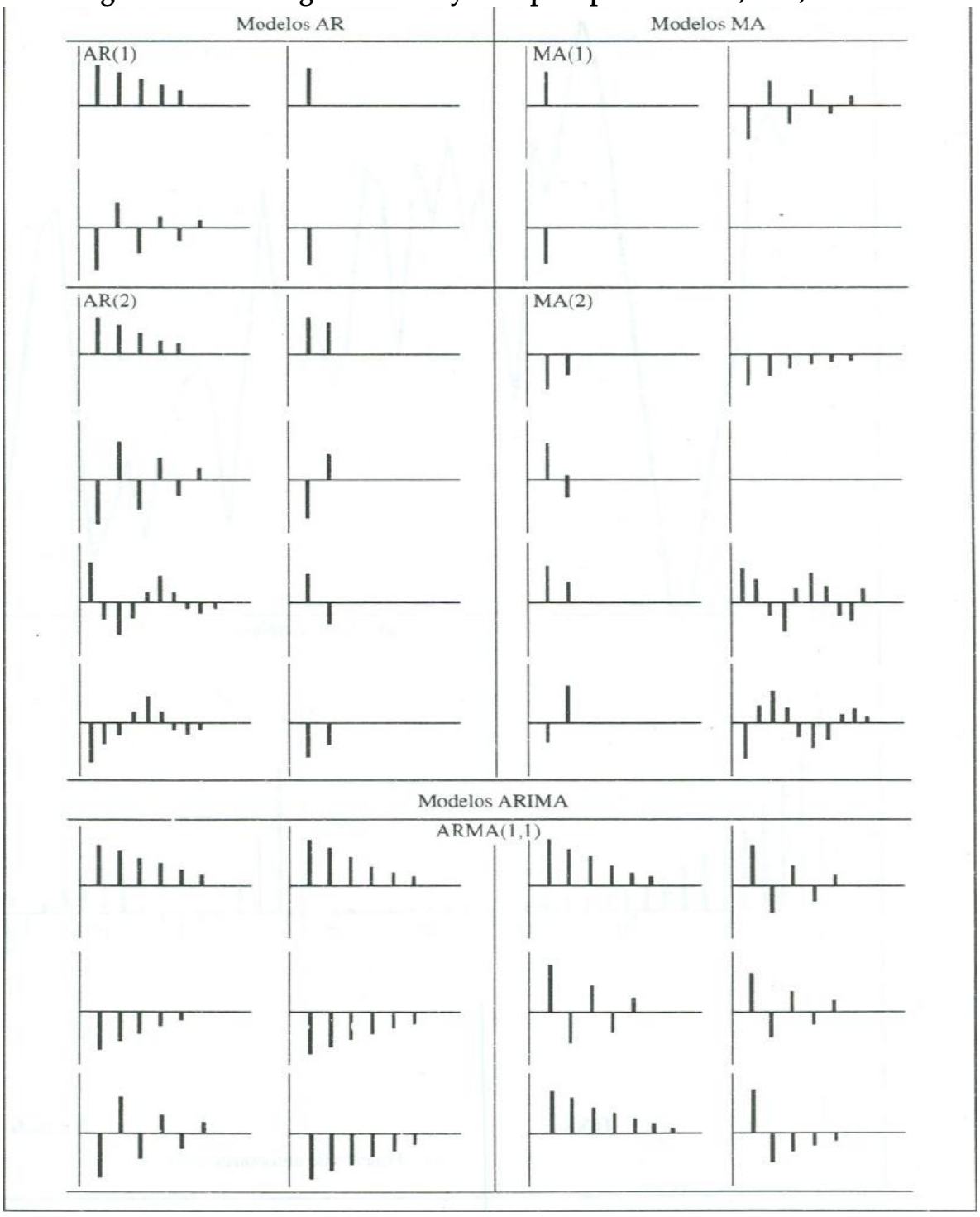
[105] Donde $t-p$ es el subíndice que representa los p periodos rezagados de Y_t o momento ocurridos en el pasado de Y_t.

[106] Donde $t-q$ es el subíndice que representa los q periodos rezagados de u_t o momento ocurridos en el pasado de u_t.

[107] Donde d se refiere al número de diferencias ($\Delta^d Y_t$) realizadas para convertir a Y_t en una serie débilmente estacionaria.

como lo exponen las ecuaciones 5.24, 5.25 y 5.26. Las cuales, pueden observarse y determinarse en los correlogramas FAS y FAP de la figura 5.2; para Y_t.

Figura 5.2. Correlogramas FAS y FAP para procesos AR, MA, ARMA



Nota:Funciones de autocorrelación simple (FAS) a la izquierda y parcial (FAP) a la derecha, para las distintas formas AR, MA y ARMA.

Fuente: Pulido y García (2001, 649), tomada textualmente de su libro.

De acuerdo con la figura 5.2, el PGD para Y_t a partir de un proceso autorrgresivo de orden uno -AR (1)- se puede determinar de la siguiente manera:

1. Cuando simultáneamente el comportamiento del FAS decrece exponencialmente y FAP resalta su primer rezago fuera de su intervalo de confianza, para los valores positivos de rho estimado ($0 \leq \hat{\rho} \leq 1$).
2. Cuando simultáneamente el comportamiento el FAS tiene movimientos senoidales y FAP resalta su primer rezago fuera de su intervalo de confianza, para los valores negativos de rho estimado ($0 \geq \hat{\rho} \geq -1$).

$$Y_t = \delta + \phi Y_{t-1} + u_t \Rightarrow (1 - \phi L)Y_t = \delta + u_t \Rightarrow \phi(L)Y_t = \delta + u_t \quad (5.27)$$

Con cualquiera de los dos comportamientos anteriores para el FAS y FAP, se puede especificar y estimar el modelo AR (1) de la ecuación 5.27. Donde δ y ϕ son los coeficientes intercepto y componente friccional[108] respectivamente, a su vez δ equivale a la media[109] (μ) de Y_t, u_t el término aleatorio o error (ruido blanco), L operador rezago y $\phi(L)$ polinomio de rezago[110].

También en la figura 5.2, el PGD para Y_t a partir de un proceso autorrgresivo de orden dos -AR (2)- se puede determinar de la siguiente manera:

1. Cuando simultáneamente el comportamiento del FAS decrece exponencialmente y FAP resalta sus dos primeros rezagos fuera de su intervalo de confianza, para los valores positivos de rho estimado.
2. Cuando simultáneamente el comportamiento el FAS tiene moviemientos senoidales y FAP resalta sus dos primeros rezagos fuera de su intervalo de confianza, intercalando estos dos valores -negativo y positivo de rho estimado- o solamente hacia el lado de los negativos.

[108] ϕ no tiene interpretación económica, pero equivale al peso e influencia positiva o negativa que tiene su pasado inmediatamente anterior Y_{t-1} sobre el comportamiento actual Y_t. En otras palabras, el valor friccional que va de atrás hacia delante.

[109] *Véase* demostración en el anexo 5.4.

[110] Todos los aspectos sobre operador rezago (L), polinomios de rezago $\{\phi(L)\}$ y su relación con ecuaciones en diferencia se encuentran en el anexo 5.2 y 5.3.

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + u_t \Rightarrow (1 - \phi_1 L - \phi_2 L^2) Y_t = \delta + u_t \Rightarrow \phi(L) Y_t = \delta + u_t \quad (5.28)$$

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + u_t \Rightarrow \left(1 - \phi_1 L - \phi_2 L^2 - .. - \phi_p L^p\right) Y_t = \delta + u_t \Rightarrow \phi(L) Y_t = \delta + u_t \quad (5.29)$$

Con los dos comportamientos anteriores para el FAS y FAP, se puede especificar y estimar el modelo AR (2) de la ecuación 5.28. Donde δ, ϕ_1 y ϕ_2 son los coeficientes, intercepto y componentes friccionales respectivamente; a su vez δ equivale a la media (μ) de Y_t; u_t el término aleatorio o error (ruido blanco); L operador rezago y $A(L)$ polinomio de rezago[111], así mismo la ecuación 5.29 representa un AR (p).

Igualmente, en la figura 5.2 es posible posible determinar el PGD para Y_t a partir de un proceso de media móvil de orden uno -MA (1)-, de la siguiente manera:

1. Cuando simultáneamente el comportamiento del FAP tiene movimientos senoidales y FAS resalta su primer rezago fuera de su intervalo de confianza, para los valores positivos de rho estimado.
2. Cuando simultáneamente el comportamiento el FAP no tiene movimientos (no existen porque estadísticamente todos sus valores de rho son iguales a cero, $\hat{\rho} = 0$) y FAP resalta su primer rezago fuera de su intervalo de confianza, para los valores negativos de rho estimado.

$$Y_t = \delta - \theta u_{t-1} + u_t \Rightarrow Y_t - \delta = (1 - \theta L) u_t \Rightarrow Y_t - \delta = \theta(L) u_t \quad (5.30)$$

A partir de lo anterior, para el FAS y FAP, se puede especificar y estimar el modelo MA (1) de la ecuación 5.30. Donde δ y θ son los coeficientes intercepto y componente friccional[112] del residuo respectivamente, u_t el término aleatorio o error (ruido blanco), L operador rezago y $\theta(L)$ polinomio de rezago. Además, en la figura 5.2 también se apreciar el PGD para Y_t a partir de un proceso proceso de media móvil de orden dos -MA (2)-, de la siguiente manera:

[111] Para más detalles de procesos AR *véase* demostración en el anexo 5.4.

[112] θ no tiene interpretación económica, pero equivale al peso e influencia positiva o negativa que tiene el pasado inmediatamente anterior (ε_{t-1}) de ε_t sobre el comportamiento actual Y_t. Recoge todos los choques exógenos actuales y anteriores que no son fácilmente controlables para explicar Y_t.

1. Cuando simultáneamente el comportamiento del FAP crece exponencialmente de los valores negativos hacia positivos de rho y FAS resalta sus dos primeros rezagos fuera de su intervalo de confianza, para los valores negativos de rho estimado.
2. Cuando simultáneamente el comportamiento el FAP no tiene movimientos (no existen porque estadísticamente todos sus valores de rho son iguales a cero, $\hat{\rho} = 0$) o tienen forma senoidal y FAP resalta sus dos primeros rezagos fuera de su intervalo de confianza, intercalando los valores negativo y positivo de rho estimado o solamente hacia el lado de los postivos.

$$Y_t = \delta - \theta_1 u_{t-1} - \theta_2 u_{t-2} + u_t \Rightarrow Y_t - \delta = (1 - \theta_1 L - \theta_2 L^2) u_t \Rightarrow Y_t - \delta = \theta(L) u_t \quad (5.31)$$

$$Y_t = \delta - \theta_1 u_{t-1} - \theta_2 u_{t-2} - .. \theta_2 u_{t-q} + u_t \Rightarrow Y_t - \delta = \left(1 - \theta_1 L - \theta_2 L^2 - .. - \theta_q L^q\right) u_t \Rightarrow Y_t - \delta = \theta(L) u_t \quad (5.32)$$

Con las caracteristicas anteriores para el FAS y FAP, se puede especificar y estimar el modelo MA (2) de la ecuación 5.31. Donde δ, θ_1 y θ_2 son los coeficientes, intercepto y componentes friccionales del residuo respectivamente; u_t el término aleatorio o error (ruido blanco); L operador rezago y $\theta(L)$ polinomio de rezago, también la ecuación 5.32 representa un MA (q).

En la figura 5.2, se puede observar por último el PGD para Y_t a partir de un proceso autorrgresivo de media móvil de orden uno -ARMA (1,1)- determinado de la siguiente manera:

1. Cuando simultáneamente el comportamiento del FAS y FAP decrecen o crecen exponencialmente, desde los valores positivos hacia negativos de rho o viceversa, y ambos resaltan su primer rezago fuera de su intervalo de confianza, para los valores positivos o negativos de rho; según su dirección.
2. Cuando simultáneamente el comportamiento del FAS y FAP tienen movimientos senoidales y resaltan su primer rezago fuera de su intervalo de confianza, intercalando los valores negativo y positivo de rho estimado o solamente hacia alguno de estos lados.

$$Y_t = \delta + \phi Y_{t-1} - \theta u_{t-1} + u_t \Rightarrow (1 - \phi L)Y_t - \delta = (1 - \theta L)u_t \Rightarrow \phi(L)Y_t - \delta = \theta(L)u_t$$
(5.33)

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} - \theta_1 u_{t-1} - \theta_2 u_{t-2} - .. - \theta_q u_{t-q} + u_t \Rightarrow$$
$$\left(1 - \phi_1 L - \phi_2 L^2 - .. - \phi_p L^p\right)Y_t - \delta = \left(1 - \theta_1 L - \theta_2 L^2 - .. - \theta_q L^q\right)u_t \Rightarrow \phi(L)Y_t - \delta = \theta(L)u_t$$ (5.34)

Ante estos dos comportamientos para el FAS y FAP, se puede especificar y estimar el modelo ARMA (1,1) de la ecuación 5.33. Donde δ, ϕ y θ son los coeficientes, intercepto y componentes friccionales; u_t el término aleatorio o error (ruido blanco); L operador rezago, $\phi(L)$ y $\theta(L)$ polinomios de rezago AR y MA respectivamente, también la ecuación 5.34 representa un ARMA (p,q).

Una vez realizado un análisis general para identificar el PGD de Y_t, a través de las estructuras AR, MA y ARMA; en el cuadro 5.2 se encuentran resumidos los principales comportamientos del FAS y FAP que caracterizan dichos procesos, con el fin de especificar y estimar, mediante mínimos cuadrados no lineales, máxima verosimilitud o Yule-Walker si es un caso especifico de un AR, el modelo más adecuado para generar el pronóstico de la serie.

Cuadro 5.2. Procesos AR, MA o ARMA según comportamiento del FAS y FAP

FAS	FAP	Proceso es:
Todos las correlaciones iguales a cero	Todos las correlaciones iguales a cero	Ruido blanco
Caída exponencial directa a cero	Primera correlación positiva y significativa. Todas las demás estadísticamente cero.	AR (1)
Caída exponencial oscilatoria a cero	Primera correlación negativa y significativa. Todas las demás estadísticamente cero.	AR (1)
Caída a cero; puede ser oscilatoria	*p* correlaciones significativas (pueden ser positivas, negativas o intercaladas). Todas las demás estadísticamente cero.	AR (p)
Primera correlación positiva y significativa. Todas las demás estadísticamente cero.	Caída oscilatoria a cero	MA (1)
Primera correlación negativa y significativa. Todas las demás estadísticamente cero.	Caída directa a cero	MA (1)
q correlaciones significativas (pueden ser positivas, negativas o intercaladas). Todas las demás estadísticamente cero.	Caída a cero; puede ser oscilatoria	MA (q)
Caída exponencial directa a cero a partir de la primera correlación	Caída exponencial directa a cero a partir de la primera correlación	ARMA (1,1)
Caída exponencial oscilatoria a cero a partir de la primera correlación	Caída exponencial oscilatoria a cero a partir de la primera correlación	ARMA (1,1)
Caída a cero que puede ser oscilatoria a partir del rezago *q*	Caída a cero que puede ser oscilatoria a partir del rezago *p*	ARMA (p,q)

Fuente: Enders (2004, 85).

No obstante, si $Y_t \sim \mathrm{I}(0)$, $Y_t \sim \mathrm{I}(1)$ o $Y_t \sim \mathrm{I}(d)$ resulta ruido blanco, mediante el correlogramas FAS y prueba Ljung-Box, cualquiera de ellas no es predecible; porque su PGD es inobservable o imposible especificar las estructuras AR, MA o ARMA para un modelo univariado (Arima). Como el caso expuesto sobre el baloto, en la figura 5.1, el cual resultó ser fuertemente estacionario en niveles $\{Baloto_t \sim \mathrm{I}(0)\}$.

Figura 5.3. Valores del FAS y FAP para el baloto

LAG	AC	PAC	Q	Prob>Q	-1 0 1 [Autocorrelation]	-1 0 1 [Partial Autocor]
1	0.0208	0.0210	.28719	0.5920		
2	0.0420	0.0420	1.4624	0.4813		
3	-0.0259	-0.0287	1.9108	0.5911		
4	-0.0178	-0.0155	2.1221	0.7133		
5	-0.0355	-0.0379	2.9647	0.7054		
6	-0.0502	-0.0487	4.6572	0.5885		
7	-0.0705	-0.0728	8.0018	0.3324		
8	-0.0153	-0.0241	8.1585	0.4181		
9	0.0178	0.0244	8.3723	0.4971		
10	0.0307	0.0377	9.0106	0.5311		
11	0.0533	0.0568	10.929	0.4493		
12	0.0328	0.0387	11.656	0.4737		
13	0.0058	0.0001	11.679	0.5541		
14	-0.0051	-0.0096	11.697	0.6307		
15	-0.0053	-0.0135	11.715	0.7004		
16	-0.0357	-0.0297	12.582	0.7030		
17	0.0123	0.0300	12.685	0.7570		
18	-0.0428	-0.0218	13.937	0.7332		
19	0.0532	0.0568	15.875	0.6656		
20	-0.0394	-0.0834	16.94	0.6569		
21	0.0677	0.0299	20.088	0.5157		
22	-0.0022	-0.0020	20.092	0.5773		
23	0.0026	0.0176	20.096	0.6361		
24	0.0418	0.0434	21.305	0.6207		
25	0.0121	0.0229	21.405	0.6698		
26	0.0657	0.0611	24.393	0.5535		
27	-0.0063	0.0015	24.421	0.6069		
28	-0.0281	-0.0234	24.97	0.6295		
29	-0.0553	-0.0737	27.097	0.5665		

Fuente: cálculos autores.

Este ejemplo es retomado en la figura 5.3., donde se puede observar que FAS y FAP del baloto carecen de movientos, dado que el valor estimado para cada rho es estidísticamente igual a cero. Razón, por la cual no se puede determinar su PGD mediante las estructuras AR, MA y ARMA señaladas en la figura 5.3, que permitan especificar y estimar un modelo univariado para su pronóstico.

Para finalizar la etapa de identificacion del PGD, también las trayectorias para el FAS y FAP de Y_t en la figura y cuadro 5.2 aplican en los correlogramas para series integradas de orden uno o d débilmente estacionarias; que hubiesen sido diferenciadas con el fin de obtener estacionariedad debíl. Una vez especificado el modelo, la estimación de los procesos (AR, MA, ARMA o Arima) es el siguiente paso a seguir dentro de la metodología Box-Jenkins.

5.4.2 Métodos para estimar modelos AR, MA, ARMA y Arima

La estimación de los coeficientes para los procesos autorregresivos (AR), media móvil (MA) y ARMA (p,q) puede ser efectuada mediante distintos métodos como: mínimos cuadrados no lineales, máxima verosimilitud y Yule-Walker, según el caso especificado. Es decir, determinando el grado de integración (d) para Y_t $\{Y_t \sim \mathrm{I}(d)\}$; el orden p de la estructura AR y q de MA. Asimismo, en el cuadro 5.3 se exponen los modelos con las características señaladas.

Cuadro 5.3. Plataformas Arima para predecir una serie de tiempo.

Orden de integración para la serie Y_t	**Orden del proceso y modelo ARIMA (p, d, q)**	**Forma del modelo**
0	AR(1) → ARIMA (1,0,0)	$Y_t = \delta + \phi Y_{t-1} + u_t$
0	MA(1) → ARIMA (0,0,1)	$Y_t = \delta - \theta u_{t-1} + u_t$
0	ARMA(1, 1) → ARIMA (1, 0, 1)	$Y_t = \delta + \phi Y_{t-1} - \theta u_{t-1} + u_t$
0	AR(p) → ARIMA (p,0,0)	$Y_t = \delta + \phi_1 Y_{t-1} + .. + \phi_p Y_{t-p} + u_t$
0	MA(q) → ARIMA (0,0,q)	$Y_t = \delta - \theta_1 u_{t-1} - .. - \theta_q u_{t-q} + u_t$
0	ARMA(p, q) → ARIMA (p, 0, q)	$Y_t = \delta + \phi_1 Y_{t-1} + .. + \phi_p Y_{t-p} - \theta_1 u_{t-1} - .. - \theta_q u_{t-q} + u_t$
1	ARIMA (1, 1, 0)	$\Delta Y_t = \delta + \phi \Delta Y_{t-1} + \Delta u_t$
1	ARIMA (0,1, 1)	$\Delta Y_t = \delta - \theta \Delta u_{t-1} + \Delta u_t$
1	ARIMA (1, 1, 1)	$\Delta Y_t = \delta + \phi \Delta Y_{t-1} - \theta \Delta u_{t-1} + \Delta u_t$
1	ARIMA (p, 1, 0)	$\Delta Y_t = \delta + \phi_1 \Delta Y_{t-1} + .. + \phi_p \Delta Y_{t-p} + \Delta u_t$
1	ARIMA (0, 1,q)	$\Delta Y_t = \delta - \theta_1 \Delta u_{t-1} - .. - \theta_q \Delta u_{t-q} + \Delta u_t$
1	ARIMA (p, 1, q)	$\Delta Y_t = \delta + \phi_1 \Delta Y_{t-1} + .. + \phi_p \Delta Y_{t-p} - \theta_1 \Delta u_{t-1} - .. - \theta_q \Delta u_{t-q} + \Delta u_t$
d	ARIMA (1, d, 0)	$\Delta^d Y_t = \delta + \phi \Delta^d Y_{t-1} + \Delta^d u_t$
d	ARIMA (0,d, 1)	$\Delta^d Y_t = \delta - \theta \Delta^d u_{t-1} + \Delta u_t$
d	ARIMA (1, d, 1)	$\Delta^d Y_t = \delta + \phi_1 \Delta^d Y_{t-1} - \theta_1 \Delta^d u_{t-1} + \Delta^d u_t$
d	ARIMA (p, d, q)	$\Delta^d Y_t = \delta + \phi_1 \Delta^d Y_{t-1} + .. + \phi_p \Delta^d Y_{t-p} - \theta_1 \Delta^d u_{t-1} - .. - \theta_q \Delta^d u_{t-q} + \Delta^d u_t$

Fuente: autores, a partir de Mendieta y Perdomo (2008, 130); Pérez y Pulido (2001); Gujarati (2003) y Greene (1998).

De acuerdo con cada modelo especificado en el cuadro 5.3, la estimación de los coeficientes $\hat{\delta}$, $\hat{\phi}$ y $\hat{\theta}$ se obtiene mediante Yule-Walker, mínimos cuadrados no lineales y máxima verosimilitud, expuestos a continuación.

5.4.2.1 Método Yule-Walker para procesos AR de orden p

La técnica Yule-Walker (YW) es empleada únicamente para estimar los parámetros $\hat{\phi}_1$, $\hat{\phi}_2$,…,$\hat{\phi}_p$ del proceso AR de algún orden (1,2, .., *p*), *véase* ecuación 5.35. Obteniéndolos, a partir de los rho estimados ($\hat{\rho}$) en la función de autocorrelación simple (FAS), expuesta en el numeral 5.3.2 (*véase* ecuación 5.36)[113]. Adicionalmente ésta metodología, suministra los valores iniciales requeridos bajo mínimos cuadrados no lineales (MCNL) o máxima verosimilitud (MV) (*véase* incisos 5.4.2.2 y 5.4.2.3).

$$Y_t = \delta + \phi_1 Y_{t-1} + .. + \phi_p Y_{t-p} + u_t \quad (5.35)$$

$$\rho_p = \phi_1 \rho_{p-1} + \phi_2 \rho_{p-2} + .. + \phi_p \rho_{p-g} \quad (5.36)$$

De las ecuaciones 5.35 y 5.36, es desarrollado el siguiente sistema (*véase* ecuación 5.37) para estimar matricialmente $\hat{\phi}_1$, $\hat{\phi}_2$,…,$\hat{\phi}_p$ en un proceso AR (p).

$$\rho_1 = \phi_1 + \phi_2 \rho_1 + .. + \phi_p \rho_{p-1}$$

$$\rho_2 = \phi_1 \rho_1 + \phi_2 + .. + \phi_p \rho_{p-2} \quad (5.37)$$

……

$$\rho_p = \phi_{p-1} + \phi_2 \rho_{p-2} + .. + \phi_p$$

El sistema expresado en la ecuación 5.37, puede transformarse matricialmente (*véase* ecuaciones 5.38, 5.39 y 5.40) de la siguiente forma:

$$\begin{bmatrix} \rho_1 \\ \rho_2 \\ \vdots \\ \rho_p \end{bmatrix}_{p\text{x}1} = \begin{bmatrix} 1 & \rho_1 & \cdots & \rho_{p-1} \\ \rho_1 & 1 & \cdots & \rho_{p-2} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p-1} & \rho_{p-2} & \cdots & 1 \end{bmatrix}_{pxp} \begin{bmatrix} \phi_1 \\ \phi_2 \\ \vdots \\ \phi_p \end{bmatrix}_{p\text{x}1} \quad (5.38)$$

[113] g equivale a rho estimado en el rezago *t-p-1*, para más detalles *véase* desarrollo en anexo 5.4.

$$\boldsymbol{\rho} = \boldsymbol{R}\boldsymbol{\phi} \text{ (5.39)}$$

$$\widehat{\boldsymbol{\phi}} = \widehat{\boldsymbol{R}}^{-1}\widehat{\boldsymbol{\rho}} \text{ (5.40)}$$

Las ecuación 5.39 corresponde a la expresión matricial de la ecuación 5.38, donde $\boldsymbol{\rho}$ es un vector columna de tamaño *p*x1 que contienen todos los valores poblacionales de rho para FAS; $\boldsymbol{R}$ es una matriz cuadrada de tamaño *p*x*p*, también con los rho poblacionales y su diagonal principal son uno y $\boldsymbol{\phi}$ es un vector columna de tamaño *p*x1 donde se encuentran los parámetros poblacionales de proceso AR (p) para la ecuación 5.35.

A partir de lo anterior, aplicando Yule-Walker, mediante el despeje de $\boldsymbol{\phi}$ en 5.39 se consigue los estimadores $\widehat{\boldsymbol{\phi}}$ del proceso AR (p) en la ecuación 5.40. En ella, $\widehat{\boldsymbol{R}}^{-1}$ la inversa estimada de la matriz cuadrada $\boldsymbol{R}$ y $\widehat{\boldsymbol{\rho}}$ un vector columna de tamaño *p*x1 que contienen todos los valores estimados de rho para FAS. Una vez calculados los valores de $\widehat{\boldsymbol{\phi}}$, a través de YW, ellos son tomados para inicializar el método no lineal en el siguiente inciso.

5.4.2.2 Método mínimos cuadrados no lineales para procesos AR y MA de orden *p* y q

El método de mínimos cuadrados no lineales (MCNL) es utilizado con el fin de calcular los otros coeficientes ($\hat{\theta}_1$,$\hat{\theta}_2$,…,$\hat{\theta}_q$) asociados al componente de media móvil (MA); cuando la especificación del modelo corresponde a un MA (q), ARMA (p,q) o Arima (p,d,q), de acuerdo con el cuadro 5.3. No obstante, previamente deben obtenerse los estimadores $\widehat{\boldsymbol{\phi}}$ en un proceso AR bajo el método Yule-Walker (previamente explicado); requeridos preliminarmente para iniciar MCNL.

Asimismo, el objetivo es encontrar estimadores AR y MA ($\hat{\phi}_1$, $\hat{\phi}_2$,…,$\hat{\phi}_p$, $\hat{\theta}_1$,$\hat{\theta}_2$,…,$\hat{\theta}_q$) que minimicen la suma de errores al cuadrado (SEC, *véase* ecuación 5.41). Por esto, los mismos deben obtenerse a través de MCNL, debido a la no linealidad (θ^{-1}) en los parámetros MA($\hat{\theta}_1$,$\hat{\theta}_2$,…,$\hat{\theta}_q$); situación que puede observarse en la ecuación 5.42, haciendo $\delta = 0$ y despejando el polinomio de rezago ($\theta(L)$) para el componente MA desde la ecuación 5.34. Ante esto, se incumple el supuesto de

linealidad en los coeficientes, razón por cual es imposible aplicar mínimos cuadrados ordinarios (MCO).

$$SEC = \sum_{n=d+p+1}^{T}\left(u_t|\hat{\phi}_1, \hat{\phi}_2, .., \hat{\phi}_p, \hat{\theta}_1, \hat{\theta}_2, .., \hat{\theta}_q\right)^2 (5.41)$$

$$\theta^{-1}(L)\phi(L)Y_t = u_t \ (5.42)$$

En este orden de ideas, MCNL soluciona el problema planteado en la ecuación 5.41 a partir de los métodos numéricos: ensayo error (búsqueda directa), optimización directa o linealización iterativa (expansión de una serie de Taylor mediante Gauss Newton o Newton-Raphson)[114]. Ahora, intuitivamente los estimadores de un MA (q) pueden obtenerse a partir de los encontrados en AR (p) con YW (*véase* ecuación 5.43) de la siguiente forma:

$$\hat{\theta}_1 = -\hat{\phi}_1$$

$$\hat{\theta}_2 = \hat{\phi}_1\hat{\theta}_1 - \hat{\phi}_2$$

$$\hat{\theta}_3 = \hat{\phi}_1\hat{\theta}_1 + \hat{\phi}_2\hat{\theta}_2 - \hat{\phi}_3 \ (5.43)$$

….

$$\hat{\theta}_q = \hat{\phi}_1\hat{\theta}_1 + \hat{\phi}_2\hat{\theta}_2 + \hat{\theta}_3\hat{\phi}_3 + \cdots - \hat{\phi}_p$$

Por otra parte, adicional a YW y MCNL, también se cuenta con el método de máxima verosimilitud (MV) -expuesto en la siguiente sección- para calcular los estimadores MA, ARMA o Arima; aun así, MV previamente requiere los valores iniciales de YW en los AR y resuelve igualmente el problema no lineal con los métodos numéricos de MCNL.

5.4.2.3 Método de máxima verosimilitud para procesos AR y MA de orden p y q

Otro método adecuado estadísticamente para estimar los parámetros de los procesos MA, ARMA o Arima es máxima verosimilitud (MV). Cuyos estimadores ($\hat{\phi}_1$, $\hat{\phi}_2$,…,$\hat{\phi}_p$, $\hat{\theta}_1$,$\hat{\theta}_2$,…,$\hat{\theta}_q$) deben maximizar la función de verosimilitud, con

[114] Para más detalles algebraicos *véase* (Pindyck y Rubinfeld, 1998, cap.10) y Gujarati (2003, cap.14).

respecto a la varianza ($\sigma_{u_t}^2$) [115] del error (u_t, termino estocástico o ruido blanco del modelo). Equivalentemente se maximiza el logaritmo de la función de verosimilitud (l) de la ecuación 5.44 y 5.45.

$$l\left(\sigma_{u_t}^2, \hat{\phi}_1, \hat{\phi}_2, \dots, \hat{\phi}_p, \hat{\theta}_1, \hat{\theta}_2, \dots, \hat{\theta}_q\right) = -\frac{n-d-p}{2} LN(2\pi) - \frac{n-d-p}{2} LN\left(\sigma_{u_t}^2\right) - \frac{1}{2\sigma_{u_t}^2} \sum_{n=d+p+1}^{T} \left(Y_t - \hat{\phi}_1 Y_{t-1} - \dots - \hat{\phi}_p Y_{t-p} + \hat{\theta}_1 u_{t-1} + \dots + \hat{\theta}_q u_{t-q}\right)^2 \quad (5.44)$$

$$\left(\sigma_{\varepsilon_t}^2, \theta^{-1}(L)\phi(L)\right) = -\frac{n-d-p}{2} LN(2\pi) - \frac{n-d-p}{2} LN\left(\sigma_{\varepsilon_t}^2\right) - \frac{1}{2\sigma_{\varepsilon_t}^2} \sum_{n=d+p+1}^{T} (\theta^{-1}(L)\phi(L)Y_t)^2 \quad (5.45)$$

La ecuación 5.45 o 5.44 no tiene solución analítica para encontrar $\hat{\phi}_1$, $\hat{\phi}_2$,…,$\hat{\phi}_p$, $\hat{\theta}_1$,$\hat{\theta}_2$,…,$\hat{\theta}_q$, por esto se requiere de los métodos Yule-Walker y numéricos de MCNL con el fin de resolver el problema y lograr estimadores AR y MA que maximicen el logaritmo de la función de verosimilitud. Este último método, finaliza la etapa de estimación de los modelos univariados bajo la metodología Box-Jenkins, la cual prosigue con su validación.

5.4.3 Validación del modelo AR, MA, ARMA o Arima estimado con YW, MCNL o MV

La verificación del modelo estimado es el siguiente paso en la técnica BJ, donde en etapas previas de este procedimiento se expuso la identificación del PGD, para una serie débilmente estacionaria, a través de una estructura AR, MA, ARMA o Arima (mediante los comportamientos simultáneos de los correlogramas FAS y FAP) y respectiva estimación; por cualquiera de los métodos (Yule-Walker, MCNL o MV) de la sección anterior.

De esta manera, en las etapas de identificación y estimación del PGD puede tenerse la posibilidad de contar con varias especificaciones AR, MA, ARMA o Arima. En consecuencia, debe realizarse la validación de cada proceso estimado con el fin de conocer cuál es que minimiza la suma de errores al cuadrado; concediéndole así, la categoría del mejor el modelo estimado, entre el conjunto de las distintas posibilidades con que se cuenta. Para lo anterior, a continuación se plantea una

[115] $\sigma_{u_t}^2 = \frac{\sum_{n=d+p+1}^{T} \hat{\varepsilon}_t^2}{n-d-p-q}$.

posible forma de validar[116] cada uno de los modelos para realizar el proceso de selección, acorde con:

1. Pruebas de significancia parcial (con el estadístico Z de la distribución normal estándar), los estimadores son relevantes individualmente.
2. Pruebas de significancia conjunta o global (con el estadístico Wald o razón de verosimilitud), los estimadores son relevantes globalmente.
3. Criterio de Akaike (CA); el valor más pequeño, como lo señalado en la sección 4.4 del capítulo 4.
4. Criterio de Schwarz (CS); el valor más pequeño, como lo señalado en la sección 4.4 del capítulo 4.
5. El término estocástico (u_t) se distribuya de manera normal (con el estadístico Jarque-Bera), como lo señalado en la sección 4.4 del capítulo 4.
6. El término estocástico (u_t) resulte ruido blanco o de igual forma ausencia de autocorrelación entre los errores (con el estadístico $\boldsymbol{Q}$ de Box-Pierce o Ljung-Box, expuesto en la sección 5.3 inciso 5.3.2).
7. La obtención de las raíces del polinomio característico, para verificar el cumplimiento de la condición de estacionariedad. Observar que las raíces invertidas de AR y MA son menores que uno. Las raíces pueden obtener analíticamente a partir del polinomio característico (*véase* anexo 5.5) del modelo a estimar y son calculadas por la mayoría de programas computacionales de análisis estadístico.

Comprendida la etapa de validación, que ayuda a seleccionar la mejor especificación dentro del conjunto de posibilidades y así obtener el modelo con la menor suma de errores al cuadrado para predecir Y_t, los siguientes apartados contienen el pronóstico y su respectiva evaluación, continuando así con la metodología Box-Jenkins.

[116] Tenga en cuenta que dentro de estos no se debe tener presente el coeficiente de determinación R^2 o el ajustado $\bar{R}^2$. Porque los estimadores no son obtenidos mediante MCO, sino con YW, MCNL o MV.

5.4.4 Pronóstico con el modelo validado y seleccionado

Una vez se cuenta con un modelo estimado que cumpla los criterios de verificación anteriores, puede ser utilizado para realizar pronósticos ($\hat{Y}_{t+p}$). Por simplicidad, considere un modelo AR (1), *véase* cuadro 5.3, con el fin de predecir Y_t un periodo adelante ($\hat{Y}_{t+1}$). Para esto, basta con plantear la expresión 5.46 en términos *t*+1 (*véase* ecuación 5.47).

$$Y_t = \delta + \phi Y_{t-1} + u_t \quad (5.46)$$

$$\hat{Y}_{t+1} = \hat{\delta} + \hat{\phi} Y_t \quad (5.47)$$

A partir de la ecuación 5.47, se realiza la proyección reemplazando los estimadores ($\hat{\delta}, \hat{\phi}$) obtenidos por YW y el ultimo valor observado de Y_t . Así sucesivamente, si se quiere predecir dos ($\hat{Y}_{t+2}$) y tres ($\hat{Y}_{t+3}$) periodos adelante Y_t (*véanse* ecuaciones 5.48 y 5.49), con un modelo AR (1).

$$\hat{Y}_{t+2} = \hat{\delta} + \hat{\phi} \hat{Y}_{t+1} \quad (5.48)$$

$$\hat{Y}_{t+3} = \hat{\delta} + \hat{\phi} \hat{Y}_{t+2} \quad (5.49)$$

$$\hat{Y}_{t+p} = \hat{\delta} + \hat{\phi}_1 \hat{Y}_{t+p-1} + .. + \hat{\phi}_p Y_t - \hat{\theta}_1 \hat{u}_{t+q-1} - .. - \hat{\theta}_q \hat{u}_t \quad (5.50)$$

$$\Delta \hat{Y}_{t+1} = \hat{\delta} + \hat{\phi} \Delta Y_t - \hat{\theta} \Delta u_t$$

$$\Delta \hat{Y}_{t+1} = \hat{Y}_{t+1} - Y_t \quad (5.51)$$

$$\hat{Y}_{t+1} = \Delta \hat{Y}_{t+1} + Y_t$$

$$\Delta^d \hat{Y}_{t+p} = \hat{\delta} + \hat{\phi}_1 \Delta^d \hat{Y}_{t+p-1} + .. + \hat{\phi}_p \Delta^d Y_t - \hat{\theta}_1 \Delta^d \hat{u}_{t+q-1} - .. - \hat{\theta}_q \Delta^d \hat{u}_t$$

$$\Delta^d \hat{Y}_{t+p} = \Delta\left(\Delta^{d-1} \hat{Y}_t\right) \quad (5.52)$$

Asimismo, si se requiere un pronóstico de *p* periodos adelante ($\hat{Y}_{t+p}$) a partir de un modelo ARMA (p,q) o Arima (1,1,0 o p,d,q), será necesario utilizar como insumo el

valor de los errores estimado en el futuro y actual ($\hat{u}_{t+q-1}$ y $\hat{u}_t$ o $\Delta^d\hat{u}_{t+q-1}$ y $\Delta^d\hat{u}_t$, *véase* ecuación 5.50, 5.51 y 5.52) y los periodos previamente proyectados ($\hat{Y}_{t+p-1}$o $\Delta^d\hat{Y}_{t+p-1}$). Adicionalmente, cuando la serie es integrada debe resolverse la ecuación en diferencia resultante del modelo Arima; para despejar de ella el valor pronosticado de Y_t en niveles que se pretende encontrar (*véase* ecuación 5.51 y 5.52). Una vez proyectada la serie con el modelo seleccionado, la metodología BJ finaliza con la validación del pronóstico.

5.4.5 Validación del pronóstico

El último paso de la metodología Box-Jenkins, consiste en validar el pronóstico obtenido con la manipulación algebraica del modelo; en el aparte anterior. En primer lugar debe realizarse una gráfica de trayectoria temporal comparativa entre los valores observados y proyectados, para observar si ellos llevan comportamiento similar en el tiempo; esto, permite comparar intuitivamente la calidad de este pronóstico.

Sin embargo, para seleccionar la mejor predicción se cuenta con el menor valor obtenido en los indicadores de error para el pronóstico, expuestos en el capítulo 4 (*véase* cuadro 4.4): promedio del valor absoluto del error (PVAE), promedio del error al cuadrado (PEC), porcentaje del promedio del valor absoluto del error (PPVAE), raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC) y coeficiente de Theil (CT).

Aunque, un menor valor de los indicadores de error para el pronóstico señalan una buena proyección; Theil es el más empleado en BJ, este coeficiente se encuentra en el intervalo cero-uno; en el caso de predicción perfecta este valor tenderá o será igual a cero, ideal para concluir sobre una buena calidad en la predicción; caso contrario si tiende o es igual a uno.

$$prob\ \left(\hat{Y}_{t+p} - n\left[1 + \sum_{p=1}^{p}\hat{\phi}_p^2 + \sum_{q=1}^{q}\hat{\theta}_p^2\right]^{0.5}\sigma_{u_t} \leq Y_{t+p} \leq \hat{Y}_{t+p} + n\left[1 + \sum_{p=1}^{p}\hat{\phi}_p^2 + q{=}1q\theta p20.5\sigma ut{=}0{,}95 \right. \quad (5.53)$$

Por último, también es posible calcular un intervalo de confianza al 95% para el pronóstico (*véase* ecuación 5.53); donde este se hace mayor conforme aumenta el

número de periodos ($\hat{Y}_{t+p}$) a proyectar. En términos generales una buena predicción debe conducir a intervalos de confianza pequeños, contrario ocurre cuando se quieren proyectar más de tres periodos con la metodología Box-Jenkins.

No obstante, la metodología BJ descrita hasta el momento aplica en series débilmente estacionarias que no contienen componente estacional. Para esta última condición, previo a proyectar la variable mediante la técnica BJ hay que desestacionalizarla; razón por cual, en la siguiente sección es abordado el tema sobre modelos univariados Sarima bajo Box-Jenkins para series estacionales.

5.5 Modelos univariados (Sarima) y metodología Box-Jenkins.

Los análisis realizados sobre estacionariedad y PGD, no están diseñados para trabajar con series estacionales. Al igual que la tendencia, el componente estacional es un obstáculo al momento de aproximarse al componente irregular que se desea modelar para realizar pronóstico. A continuación se encuentra el tratamiento de variables estacionales no estacionarias[117].

5.5.1 Uso de series Estacionales y ajuste estacional (desestacionalización)

Cuando el componente estacional es prevaleciente en la serie, previo a realizar las pruebas de estacionariedad (gráfica, FAS y DF) y aplicación de la metodología BJ para pronosticarla, deberá desestacionalizarse. Debido a que esta condición, invalida las conclusiones derivadas de la gráfica, FAS y DF; sin importar el grado de diferenciación o integración de la variable para garantizar estacionariedad.

En otras palabras, las pruebas gráficas y de raíz unitaria (DF) aplicadas sobre una serie diferenciada, y que aun así conserva estacionalidad, pueden señalar estacionariedad debíl sobre ella; pero es una conclusion errada dada la presencia de raíces unitarias estacionales[118], desconocidas en la transformación inicial. Por esto, deberá aplicarse una difencia estacional (*véase* ecuación 5.54) o

[117] Hay que tener en cuenta que los conceptos de estacionalidad y estacionariedad son diferentes.

[118]Las raíces unitarias abarcadas hasta el momento son denominadas regulares, para series sin componente estacional. En estas últimas se conocen como estacionales, sus raíces unitarias implícitas; debido al componente estacional que caracteriza la serie a tratar o pronosticar.

desestacionalizarla preliminarmente a cualquier tipo de análisis o pronóstico que se le quiera realizar.

$$\Delta(Y_t - Y_{t-s}) \text{ o } \Delta_s^d Y_t = \Delta_s(Y_t - Y_{t-s}) \text{ (5.54)}$$

Otra alternativa para eliminar el componente estacional, es usar una de las metodologías de suavizamiento exponencial estacional (Holt-Winters multiplicativo y aditivo) presentadas en el capítulo 4. Así, se procede a realizar los análisis de estacionariedad mediante las gráficas, correlograma (FAS) y raíz unitaria (DFA) en la serie sin estacionalidad (desestacionalizada). Una vez concluida estacionariedad débil para ella, posteriormente se puede emplear la metodología BJ para su respectiva predicción; ahora con modelos univariantes Sarima (estacional autorregresivo integrado de media móvil, *véase* cuadro 5.4) o Arima estacional $(p, d, q)\ (P, D, Q)_s$ [119].

[119] Puede observase que las letras minúsculas p,d,q son las tratadas en el modelo Arima de la sección anterior, mientras las mayúsculas se relacionan con la estacionalidad así: *P* se refiere al orden del proceso Autorregresivo estacional (SAR), *D* al orden de integración o diferenciación estacional $(D)_s$ y *Q* al orden de la media móvil estacional (SMA).

Cuadro 5.4. Plataforma Sarima para predecir una serie de tiempo estacional.

Orden de integración para la serie Y_t	Orden del proceso y modelo Sarima	Forma del modelo
d-S	SARIMA (p, d, q) $(P,D,Q)_s$	$\Delta_s^d Y_t = \delta + \phi_1 \Delta_s^d Y_{t-1} + .. + \phi_p \Delta_s^d Y_{t-p} - \theta_1 \Delta_s^d u_{t-1} - .. - \theta_q \Delta_s^d u_{t-q} + \Delta_s^d u_t$

Fuente: autores, a partir de Mendieta y Perdomo (2008, 130); Pérez y Pulido (2001); Gujarati (2003) y Greene (1998).

De esta forma, la metodología Box-Jenkins es similar a la expuesta en la sección 5.4; teniendo en cuenta que el procedimiento debe realizar se sobre la serie desestacionalizada de la siguiente manera:

1. Desestacionalizar la serie estacional, mediante una diferencia estacional (*véase* ecuación 5.54).
2. Análisis de estacionariedad: la variable desestacionalizada debe resultar débilmente estacionaria -no ruido blanco-; aplicando las pruebas: gráficas (trayectoria temporal y correlograma del FAS), $\boldsymbol{Q}$ de Ljung-Box y raíz unitaria (DFA) sobre la variable desestacionalizada. Asimismo, determinar el orden de integración para la serie en cuestión.
3. Identificar el proceso generador de datos (PGD), sobre los correlogramas (de acuerdo con los comportamientos FAS y FAP simultáneamente, *véase* inciso 5.4.1) de la variable sin estacionalidad (desestacionalizada) débilmente estacionaria, a través de los términos autorregresivos (AR) y media móvil (MA).
4. Especificación y estimación[120] (*véase* inciso 5.4.2) del modelo AR, MA, ARMA, SAR, SMA (para variables integradas de orden cero), Sarima (para variables integrada de algún orden con estacionalidad).
5. Validación del modelo especificado y estimado (*véase* inciso 5.4.3): pruebas de significancia parcial (con el estadístico Z de la distribución normal estándar); los estimadores son relevantes individualmente y global (con el estadístico Wald o razón de verosimilitud); los coeficientes

[120]Mediante métodos de mínimos cuadrados no lineales, máxima verosimilitud o Yulke Walker si el proceso es únicamente AR (autorregresivo), *véase* anexo 5.4.

son relevantes globalmente, raíces invertidas de AR y MA (menores a 1), criterios de Akaike (CA) y Schwarz (CS), residuales ruido blanco (no autocorrelacionados, con el estadístico $\boldsymbol{Q}$ de Box-Pierce o Ljung-Box), distribución normal en residuales (*iid*, idéntica e independientemente distribuidos, con el estadístico Jarque-Bera).

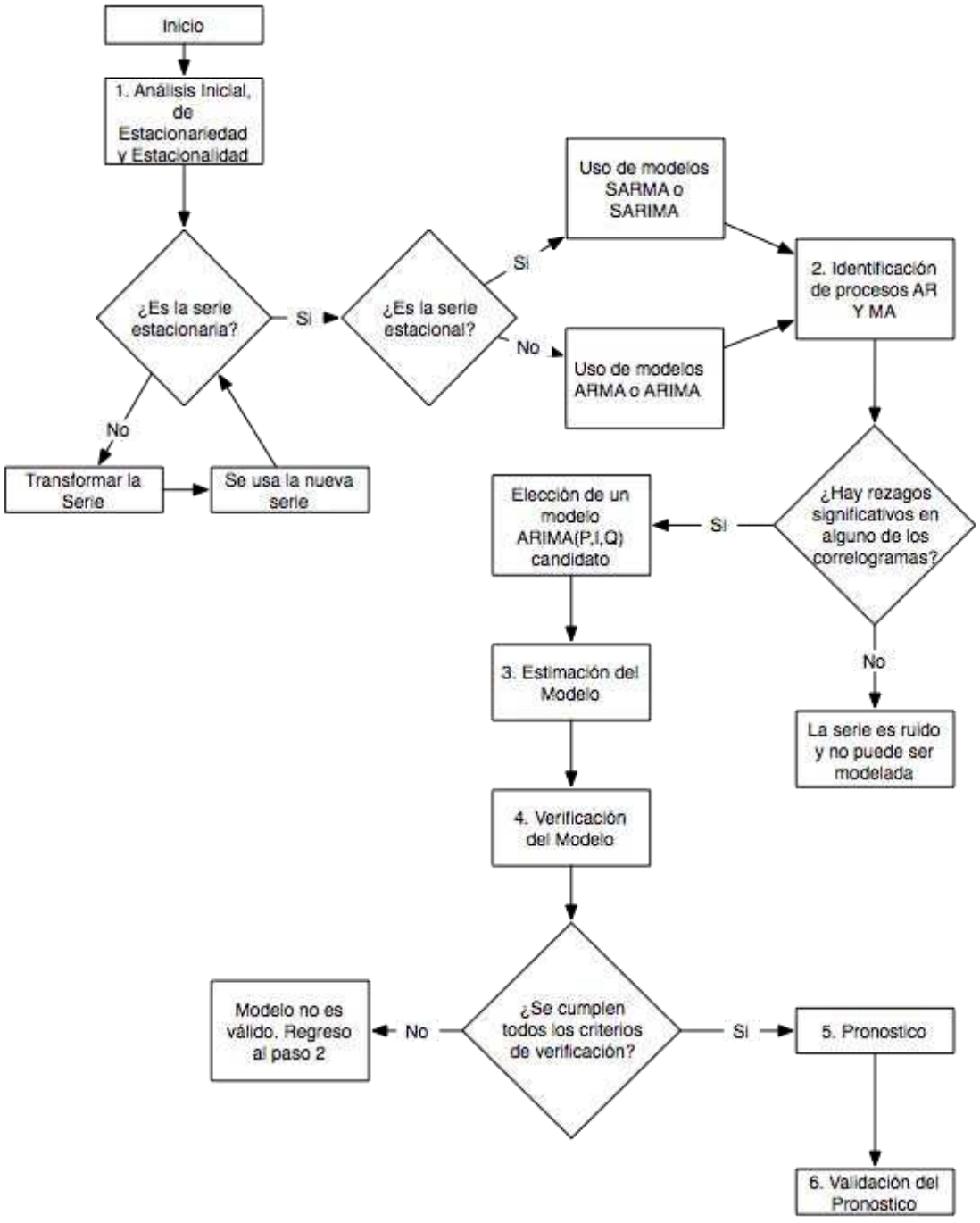
6. Realizar el pronóstico, de corto plazo, con el modelo Sarima correctamente especificado y validado mediante las pruebas del numeral cinco (*véase* inciso 5.4.4).
7. Validación de la predicción $\left(\hat{Y}_t \; o \; \hat{Y}_{t+p}\right)$: mediante la gráfica de los valores observados Vs proyectados e indicadores de error para el pronóstico (*véase* inciso 5.4.).

Una vez culminado el tratamiento sobre series estacionales, a continuación es presentado a través de un esquema el resumen de la metodología Box-Jenkins y principales ventajas y desventajas sobre los modelos Arima abarcados en este capítulo.

5.5.2 Resumen de la metodología Box-Jenkins

Para finalizar la metodología de Box-Jenkins, en este inciso se presenta un diagrama de flujo (*véase* figura 5.4) resumiendo su procedimiento y todos los aspectos abarcado en el capítulo sobre la misma. Igualmente, los diferentes pasos se encuentran numerados de la misma manera como se presentaron desde el inicio de esta sección.

Figura 5.4. Diagrama de flujo con el procedimiento BJ.



Fuente: los autores.

5.6 Ventajas y desventajas de los modelos Arima.

Entre las ventajas y desventajas de los modelos Arima descritos bajo la metodología Box-Jenkins descrita, se puede destacar lo siguiente:

- En general es una buena herramienta para realizar pronósticos de corto plazo.
- A diferencia de las metodologías sobre tendencia determinística y suavizamiento exponencial, existen procedimientos formales que permiten la comparación entre los diferentes modelos que se pueden especificar y estimar; así evaluar la precisión de un pronóstico determinado.
- Una desventajas, es que estos modelos toman bastante tiempo en estimarse y usualmente requieren significativos recursos computacionales. En particular para el proceso de estimación se hace necesario usar diferentes métodos numéricos que solo están disponibles en programas especializados de última generación.
- Otra desventaja, es requieren una cantidad relativamente grande de datos para poder llegar a estimaciones validas. Al igual que los modelos econométricos de corte transversal, se recomiendan al menos 40 o más observaciones de la serie; esta cantidad en series anules pueden no encontrarse disponibles.
- Por último, estos modelos no cuentan con un sistema sencillo para incorporar nuevos datos; esto hace que el modelo tenga que reajustarse periódicamente -o incluso volver a estimarse totalmente después de unos periodos- lo cual es complejo, cuando las series son diarias o semanales[121].

Para cerrar el capítulo y una vez expuestos los modelos Arima y Sarima mediante la metodología BJ, en la siguiente sección se encuentra un estudio de caso con su aplicación. Para lo anterior, se tomó la información trimestral del PIB colombiano, utilizada en el capítulo 4 y datos mensuales del índice de precios al consumidor en Colombia (IPC).

[121] *Véase* Hanke y Wichern (2006, 426).

5.7 Estudio de caso: PIB colombiano

A continuacion, se realizan todos los pasos de la metodología Box-Jenkins, que permiten modelar una serie y obtener sus pronósticos. Para esta sección, tomado como ejemplo, la serie trimestral del producto interno bruto (PIB) colombiano (para evidenciar el caso del modelo Arima), utilizada en el capítulo 4. En la siguiente, usando los datos mensuales (1992-I hasta 2009-VI) del índice de precios al consumidor en Colombia (IPC), para evidenciar el caso del modelo Sarima.

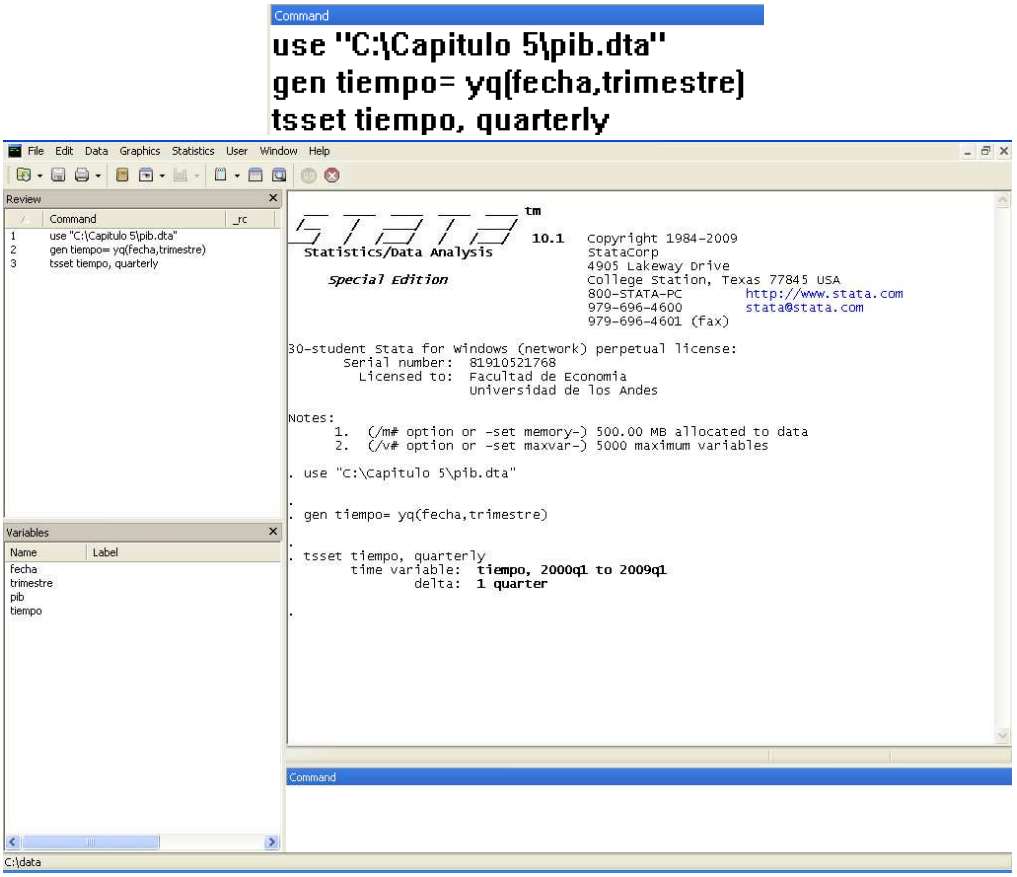
Con la información del PIB colombiano se pretende estimar un modelo Arima para realizar su predicción dos periodos adelante, para el II y II trimestre de 2009 ($\widehat{PIB}_{t+1}$ =2009-II y $\widehat{PIB}_{t+2}$ =2009-III), paso a paso bajo la metodología Box-Jenkins de la siguiente manera:

5.7.1 Análisis de estacionariedad

5.7.1.1 Análisis gráfico para detectar estacionariedad

1- Cargar la base de datos y configurar el programa, para que reconozca el PIB_t como serie de tiempo trimestral, con el comando *gen* y *tsset* (*véase* figura 5.5).

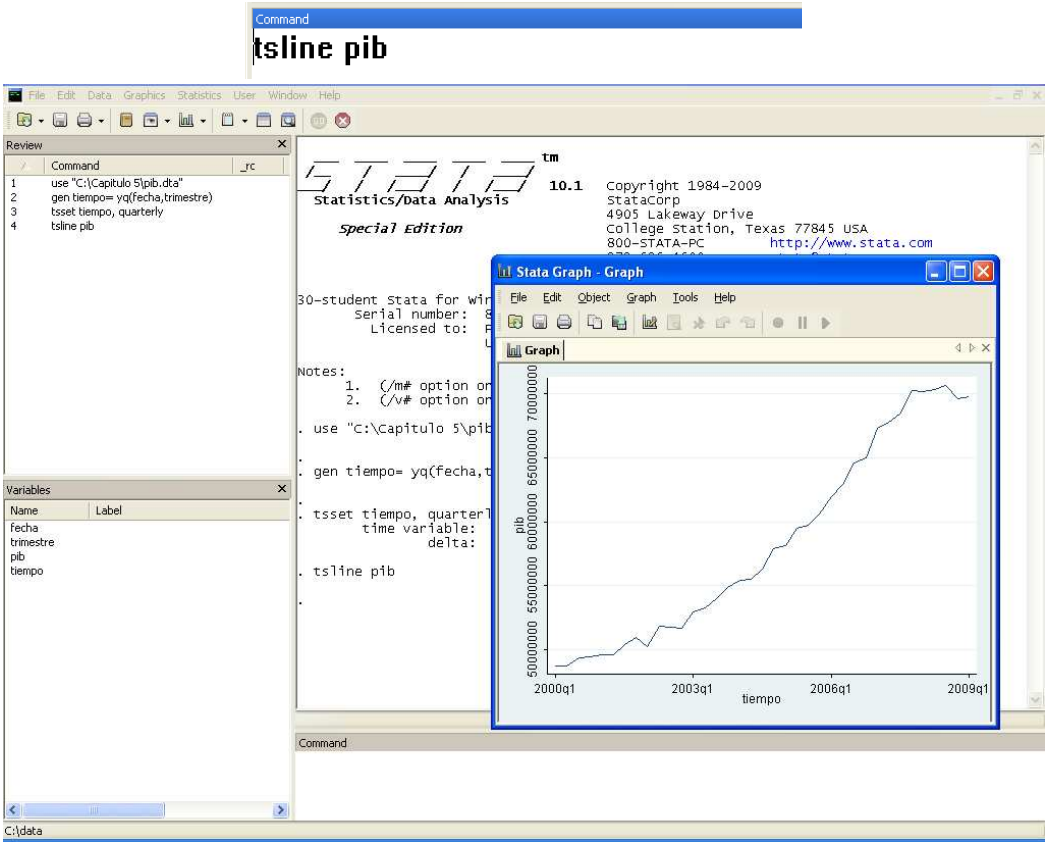
Figura 5.5. Salida de Stata® para especificar una variable como serie de tiempo



Fuente: cálculos autores.

2- Graficar el comportamiento del PIB_t a través del tiempo, con el comando *tsline* (*véase* figura 5.6), para conocer la forma funcional de la tendencia.

Figura 5.6. Salida de Stata® para graficar una serie de tiempo



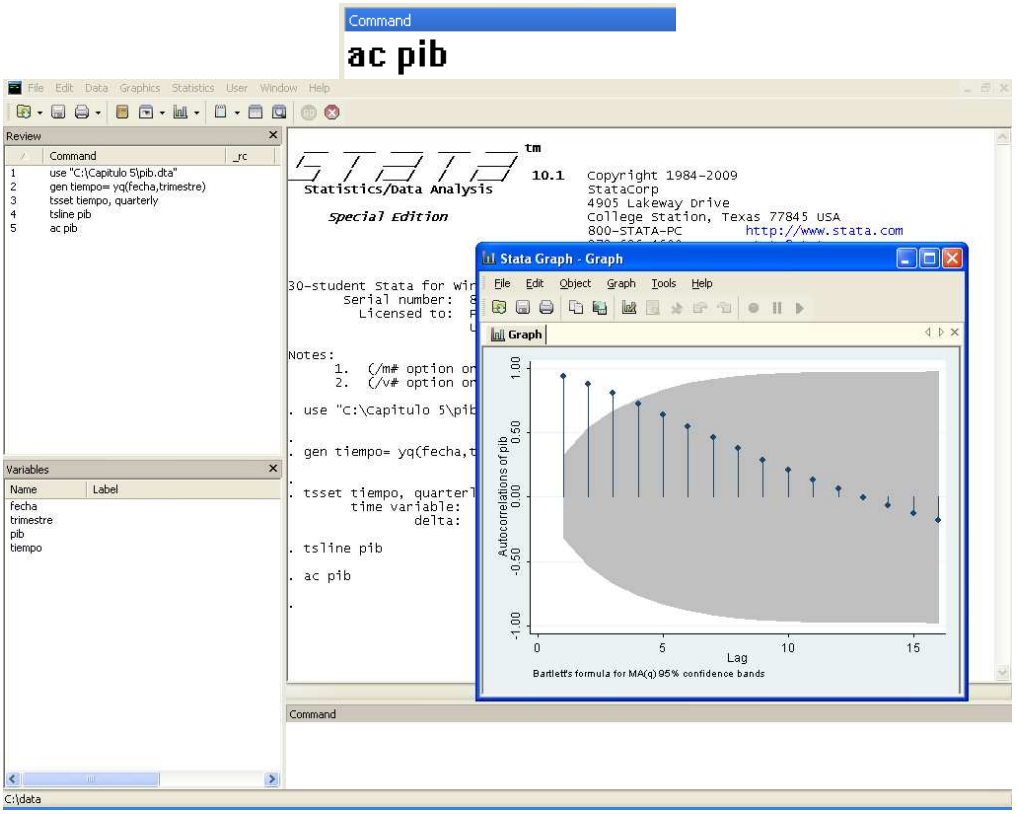
Fuente: cálculos autores.

En este caso, el PIB_t muestra una tendencia creciente lineal hasta el 2008-I, donde existe un punto de inflexión (*véase* figura 5.6). Como se mencionó anteriormente, la variable cuestionada presenta media y varianza inestables entre 2000-I y 2009-I; dada la presencia tendencial e irregular en ella. Conllevando a que la serie (PIB) no es estacionaria en media y varianza.

5.7.1.2 Análisis gráfico mediante el correlograma de la función de autocorrelación simple (FAS) para detectar estacionariedad

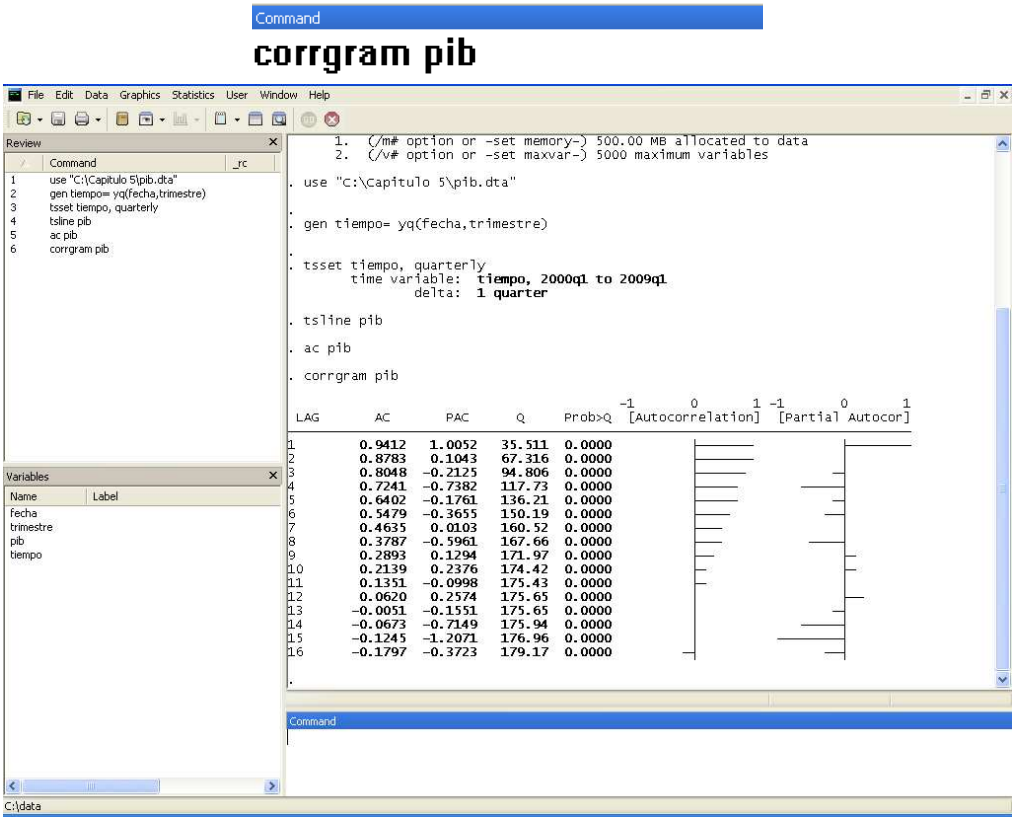
1- Graficar los rho ($\hat{\rho}_p$) estimados del FAS, con el comando *ac* y *corrgram* (*véase* figura 5.7 y 5.8).

Figura 5.7. Salida de Stata® para graficar el FAS del PIB



Fuente: cálculos autores.

Figura 5.8. Salida de Stata® para graficar el FAS del PIB



Fuente: cálculos autores.

De esta manera, la figura 5.7 y 5.8 indica no estacionariedad para el PIB dado que exterioriza estimaciones $\hat{\rho}_p$ decrecientes exponencialmente de mayor a menor (entre 1 y -1), llegando a cero, tomando valores negativos y la mayor parte de ellos se encuentran fuera de su intervalo de confianza. Por otra parte, las figuras 5.8 exhibe los valores del FAS ($\hat{\rho}_p$, *véanse* columnas AC), graficados en el correlograma anterior, para el PIB colombiano (decrecen exponencialmente, desde 0.9412 hasta -0.3617) respectivamente.

H_0: $\hat{\rho}_1 = \hat{\rho}_2 = \cdots =: \hat{\rho}_p = 0$; La serie PIB es ruido blanco, implicando automaticamente estacionariedad en ella (donde p=16).

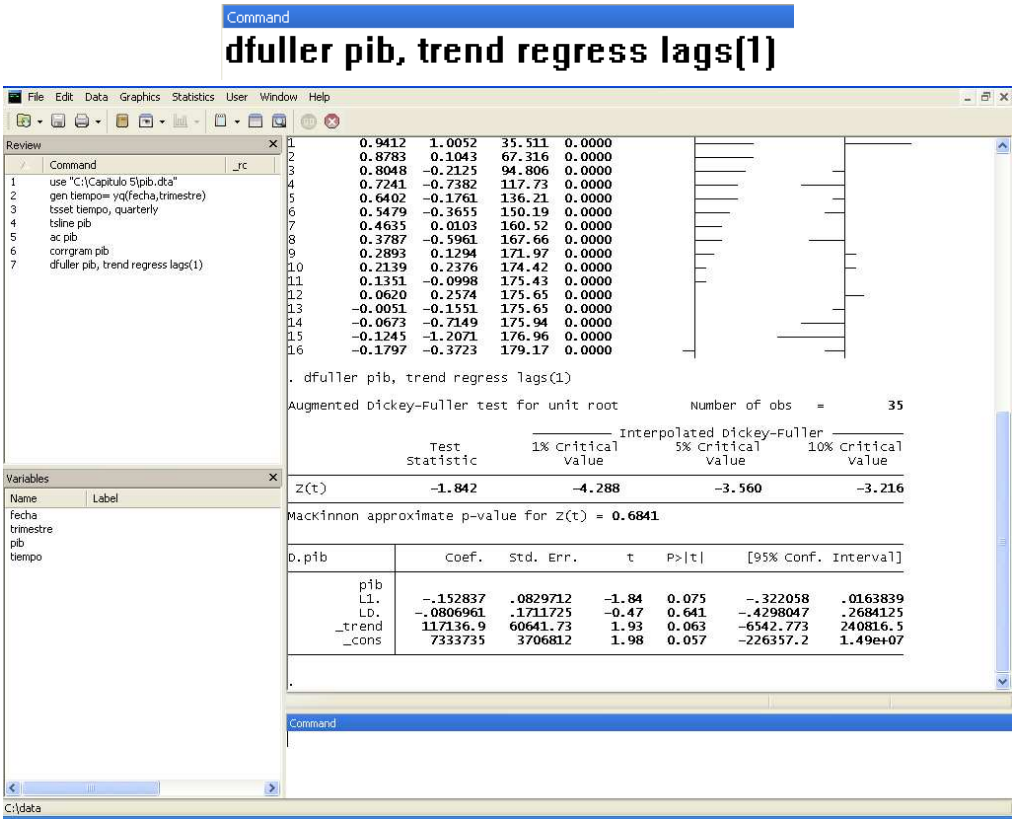
H_1: $\hat{\rho}_1 \neq \hat{\rho}_2 \neq \cdots \neq: \hat{\rho}_p \neq 0$; La serie PIB no es ruido blanco, implicando que posiblemente es estacionaria.

Adicionalmente, los resultados descritos para FAS y ***Q*** ayudan a comprobar que el PIB no es ruido blanco. *Véase* ultima columna en la figura 5.8 que contiene las probabilidades del estdistico **Q** Ljung-Box**,** indicando (con nivel de significancia del 5%) que se rechaza la hipotesis nula de ruido blanco para el PIB.

5.7.1.3 Análisis de raíz unitaria Dickey-Fuller aumentado (DFA) para detectar estacionariedad

1- Estimar tau de DFA con tendencia e intercepto, con el comando *dfuller pib, trend regress lags(1)* (*véase* figura 5.9).

Figura 5.9. Salida de Stata® para realizar DFA



Fuente: cálculos autores.

$\Delta PIB_t = \alpha + \delta PIB_{t-1} + \beta t + \gamma \Delta PIB_{t-1} + u_t$, con intercepto y tendencia, adicionalmente con un rezago para ΔPIB_{t-1}; $\delta = (\rho - 1)$.

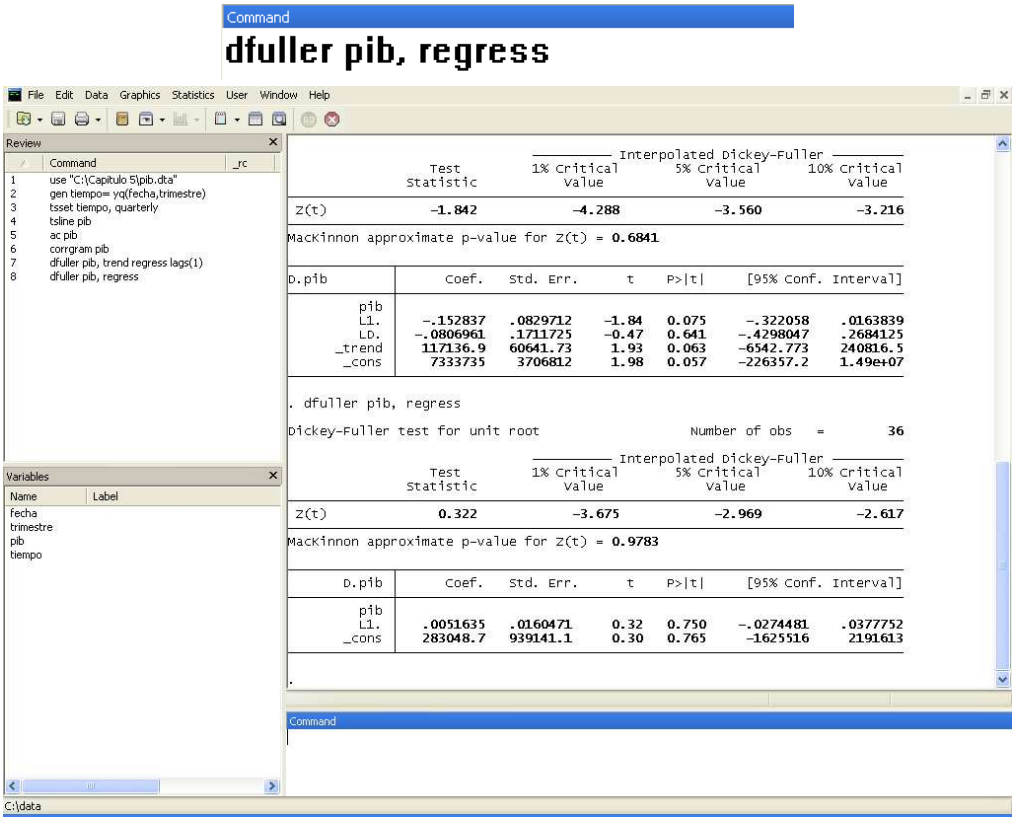
$H_0: \delta = 0$; $\rho = 1$; la serie PIB contiene raíz unitaria, equivale a decir que es una caminata aleatoria o simplemente no es estacionaria.

$H_1: \delta \neq 0$; $\rho \neq 1$; la serie PIB no contiene raíz unitaria equivale a decir que no es una caminata aleatoria o simplemente es estacionaria.

En este caso (*véase* figura 5.9) $\tau = -1.842$ y su probabilidad igual a 0.6841. Indicando que no se rechace la hipotesis nula (a un nivel de significancia del 1%, 5% y 10%), por tanto el PIB no es estacionario en su nivel. Descartando que el mismo es integrada de orden cero PIB~I(0). En este caso el valor de tau (τ) es negativo y su valor absoluto es 1.842 ($|\tau| = 1.8242$), significa que $-1 < \hat{\rho} < 1$. Valor, comparable con lo valores absolutos criticos de MacKinnon ($|1\%| = 4.288$, $|5\%| = 3.560, |10\%| = 3.216$); $|\tau| < |1\%|, |5\%|$ y $|10\%|$, ratificando el no rechazo de la hipotesis nula. Igualmente el modelo debe estar especificado con intercepto y tendencia debido a que sus valores estaditicos (1.98 y 1.93) son estadísticamente significativos.

2- Estimar tau de DFA sin tendencia e intercepto, con el comando *dfuller pib* (*véase* figura 5.10).

Figura 5.10. Salida de Stata® para realizar DFA sin tendencia e intercepto



Fuente: cálculos autores.

$\Delta PIB_t = \delta PIB_{t-1} + u_t$, sin intercepto y tendencia, $\delta = (\rho - 1)$.

H_0: $\delta = 0$; $\rho = 1$; la serie PIB contiene raíz unitaria, equivale a decir que es una caminata aleatoria o simplemente no es estacionaria.

H_1: $\delta \neq 0$; $\rho \neq 1$; la serie PIB no contiene raíz unitaria equivale a decir que no es una caminata aleatoria o simplemente es estacionaria.

En este caso (*véase* figura 5.10) el modelo se encuentra mal especificado $\tau = 0.322$ y su probabilidad igual a 0.9783. Indicando que no se rechaze la hipotesis nula, por tanto el PIB no es estacionario en su nivel. Sin embargo, en este caso el valor de tau (τ) es positivo, automaticamente no se rechaza la hipotesis nula; porque $\hat{\rho} > 1$ y este valor debe estar entre $-1 \leq \hat{\rho} \leq 1$; mostrando que la serie PIB es explosiva (media, varianaza y covarianza estan concionadas con el tiempo). Impidiendo asi

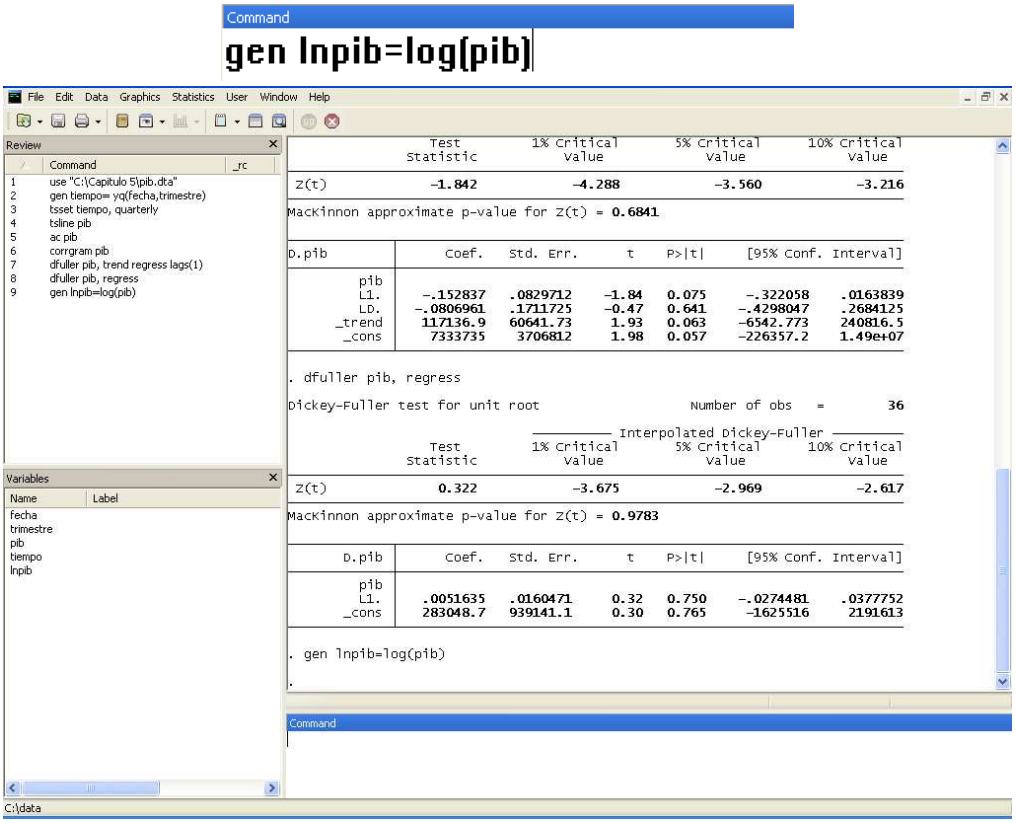
que τ tome valores negativo, los cual imposibilita evaluarlo en valor absoluto con los valores absolutos criticos (1%, 5% y 10%) de MacKinnon.

Entonces, el valor de tau (τ) obtenido con el modelo bien especifcado de la figura 5.5 prima sobre la anterior en la figura 5.10, porque su tendencia, intercepto y el rezago son significativos al 10% de nivel de significancia. Una vez determinado que una serie no es estacionaria, exiten transformaciones con ecuaciones en diferencia para lograr estcionariedad en la variable PIB.

5.7.1.2 Transformación en primeras diferencia logarítmicas para convertir el PIB en una serie estacionaria

1- Generar una nueva variable, con el comando *gen*, que contenga el logaritmo natural de PIB ($LNPIB_t$) (*véase* figura 5.11).

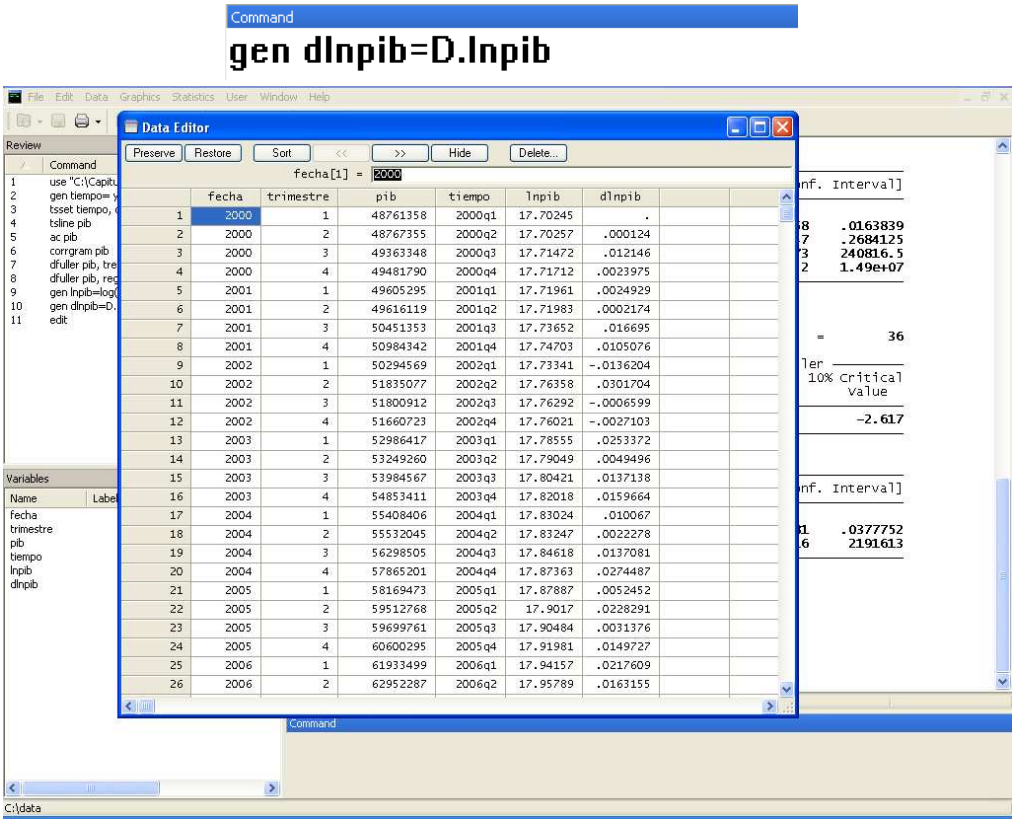
Figura 5.11. Salida de Stata® para crear una variable como logaritmo natural



Fuente: cálculos autores.

2- Generar dos nueva variable, con el comando *gen*, que contenga el primer rezago del logaritmo natural de PIB ($LNPIB_{t-1}$) y primera diferencia logarítmica ($\Delta LNPIB_t$, *véase* figura 5.12).

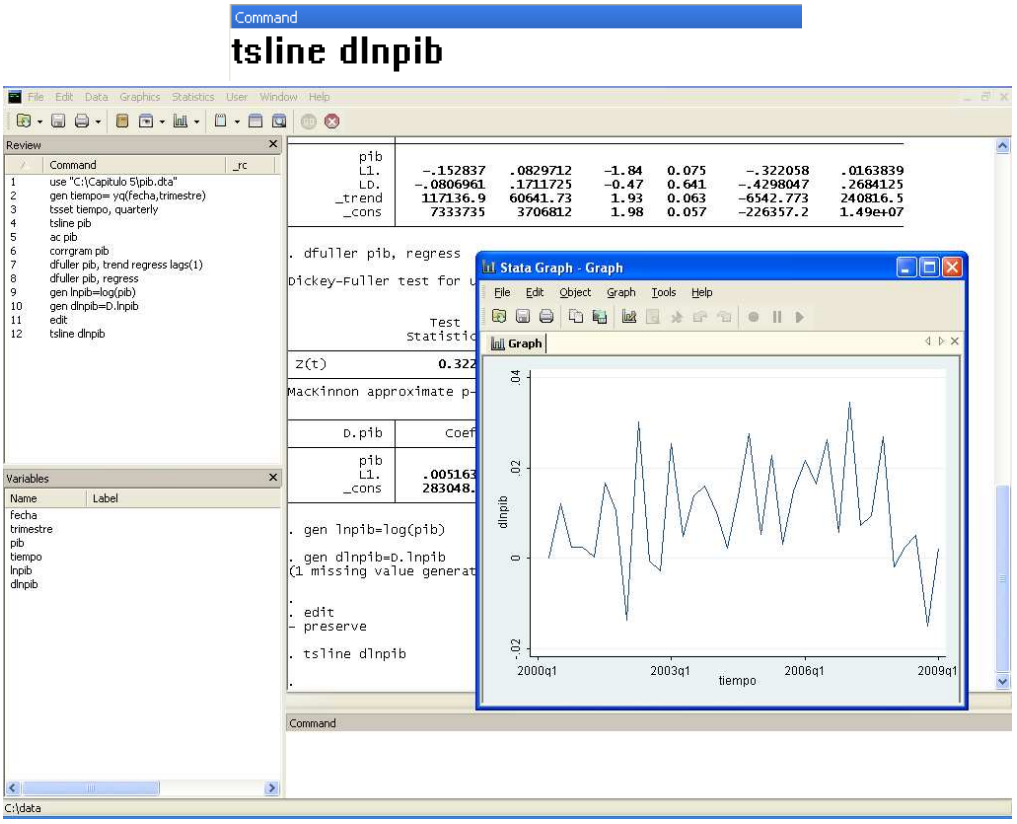
Figura 5.12. Salida de Stata® para crear una en primeras diferencias del logaritmo natural



Fuente: cálculos autores.

3- Graficar el comportamiento de la primera diferencia logarítmica de PIB_t ($\Delta LNPIB_t$) a través del tiempo, con el comando *tsline* (*véase* figura 5.13).

Figura 5.13. Salida de Stata® para graficar una serie de tiempo en primeras diferencias logarítmica

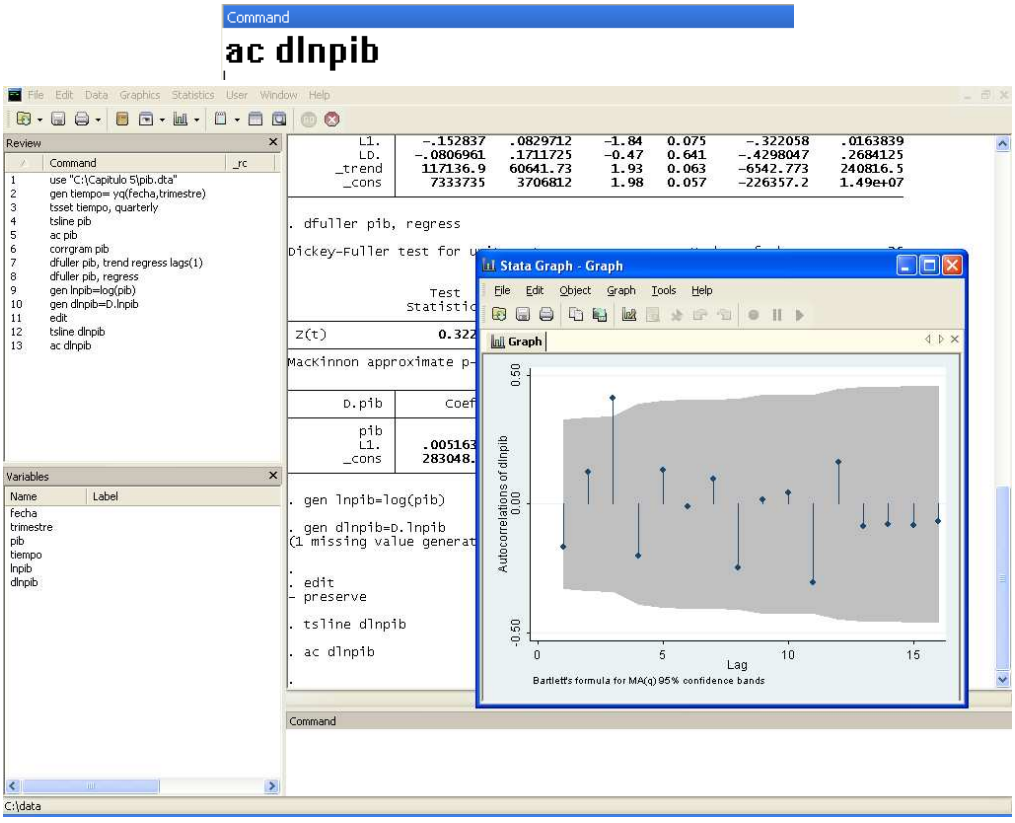


Fuente: cálculos autores.

En este caso, la primera diferencia en logaritmo del PIB_t ($\Delta LNPIB_t$) no muestra una tendencia, sino movimiento senoidales (*véase* figura 5.13); aparentemente la variable cuestionada presenta media y varianza estables entre 2000-II y 2009-I. Donde presumiblemente la serie en primeras diferencias logarítmica es estacionaria en media y varianza.

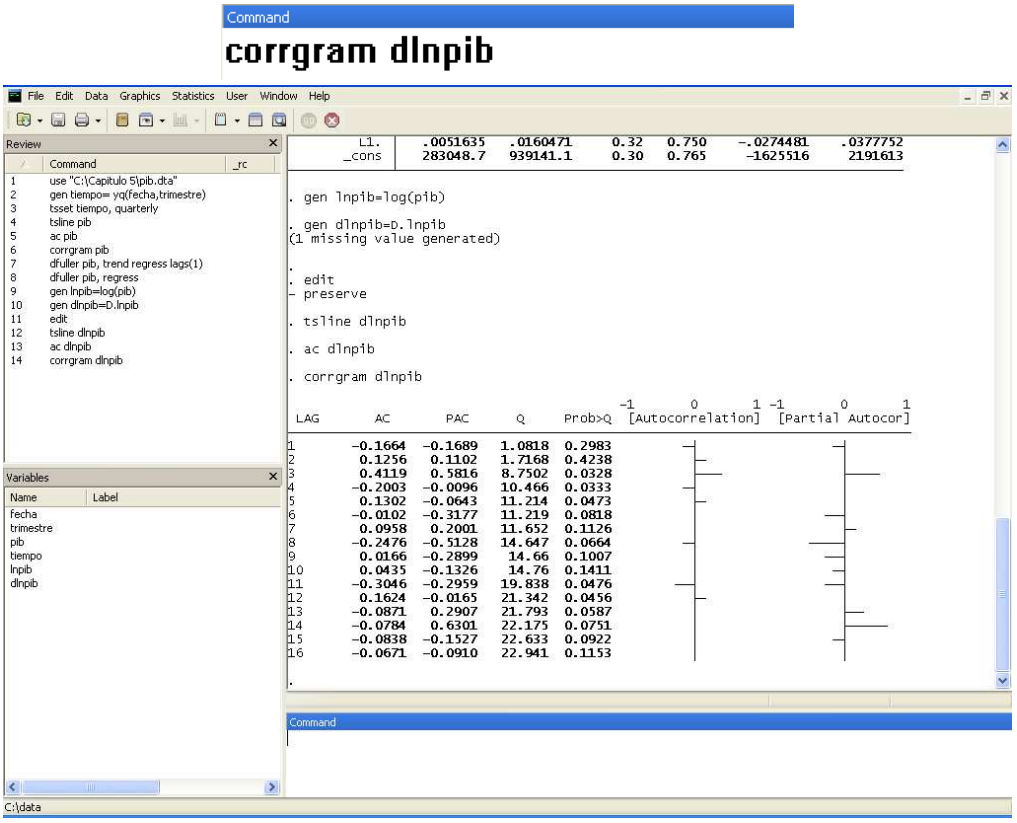
4- Graficar los rho ($\hat{\rho}_p$) estimados del FAS para la series en primeras diferencias logarítmicas del PIB ($\Delta LNPIB_t$), con el comando *ac* y *corrgram* (*véase* figura 5.14 y 5.15).

Figura 5.14. Salida de Stata® para graficar el FAS de $\Delta LNPIB_t$



Fuente: cálculos autores.

Figura 5.15. Salida de Stata® para graficar el FAS de $\Delta LNPIB_t$



Fuente: cálculos autores.

De esta manera, la figura 5.14 y 5.15 indica posible estacionariedad para el PIB dado que exterioriza estimaciones $\hat{\rho}_p$ senoidales intercaladas y la mayor parte de ellos se encuentran dentro de su intervalo de confianza. Por otra parte, las figuras 5.10 exhibe los valores del FAS ($\hat{\rho}_p$, *véase* columnas AC), graficados en el correlograma anterior, para el PIB colombiano (intercalados, desde-0.1664 hasta -0.0671) respectivamente.

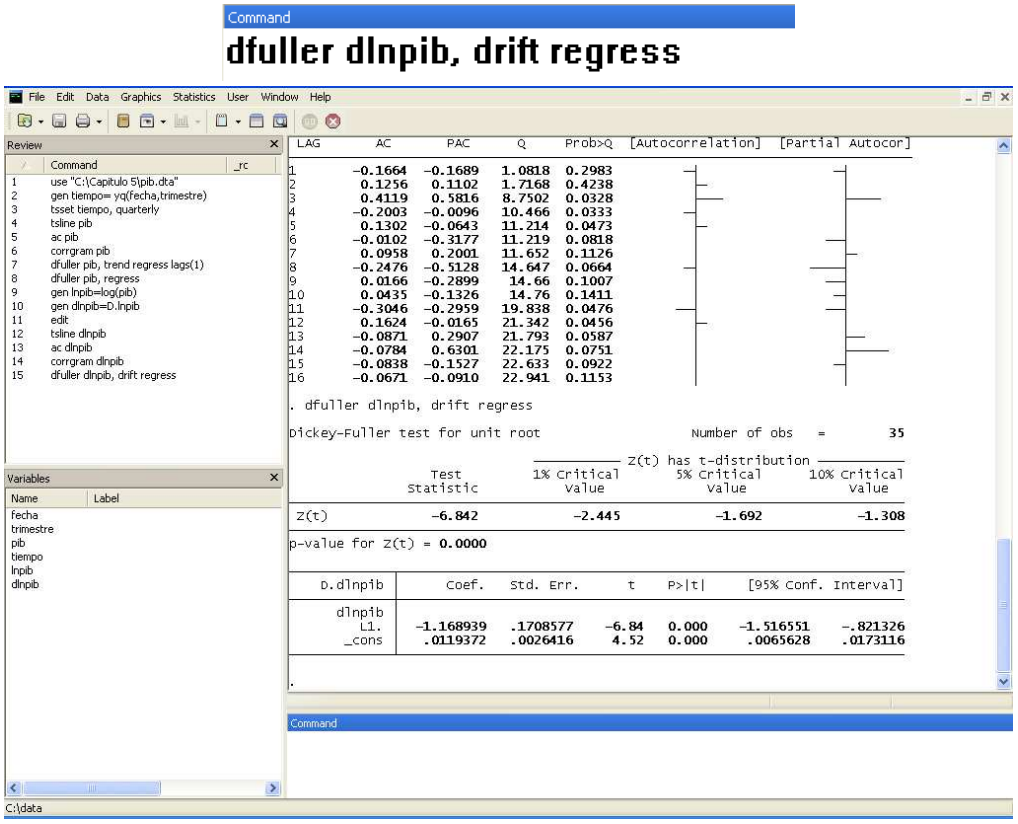
$H_0: \hat{\rho}_1 = \hat{\rho}_2 = \cdots =: \hat{\rho}_p = 0$; La serie $\Delta LNPIB_t$ es ruido blanco, implicando automaticamente estacionariedad en ella (donde p=16).

$H_1: \hat{\rho}_1 \neq \hat{\rho}_2 \neq \cdots \neq: \hat{\rho}_p \neq 0$; La serie $\Delta LNPIB_t$no es ruido blanco, implicando que posiblemente es estacionaria.

Adicionalmente, los resultados descritos para FAS y $\boldsymbol{Q}$ ayudan a comprobar que el $\Delta LNPIB_t$ no es ruido blanco. *Véase* ultima columna en la figura 5.15 que contiene las probabilidades del estdistico **Q** Ljung-Box**,** indicando (con nivel de significancia del 5%) que se rechaza la hipotesis nula de ruido blanco para el $\Delta LNPIB_t$. Este resultado, indica que a la serie PIB en primeras diferencias logarítmicas es posible encontrar su PGD para poder predercila, caso contrario hubiese pasado sino se rechaza la hipotesis nula.

5- Estimar tau de DF con intercepto para $\Delta LNPIB_t$, con el comando *dfuller dlnpib,drift regress* (*véase* figura 5.16 y ecución 5.55).

Figura 5.16. Salida de Stata® para realizar DF de $\Delta LNPIB_t$



Fuente: cálculos autores.

$$\Delta^2 lnPIB_t = \alpha + \delta \Delta lnPIB_{t-1} + u_t \text{ con intercepto; } \delta = (\rho - 1) \quad (5.55)$$

H_0: $\delta = 0$; $\rho = 1$; la serie $\Delta LNPIB_t$contiene raíz unitaria, equivale a decir que es una caminata aleatoria o simplemente no es estacionaria.

H_1: $\delta \neq 0$; $\rho \neq 1$; la serie $\Delta LNPIB_t$no contiene raíz unitaria equivale a decir que no es una caminata aleatoria o simplemente es estacionaria.

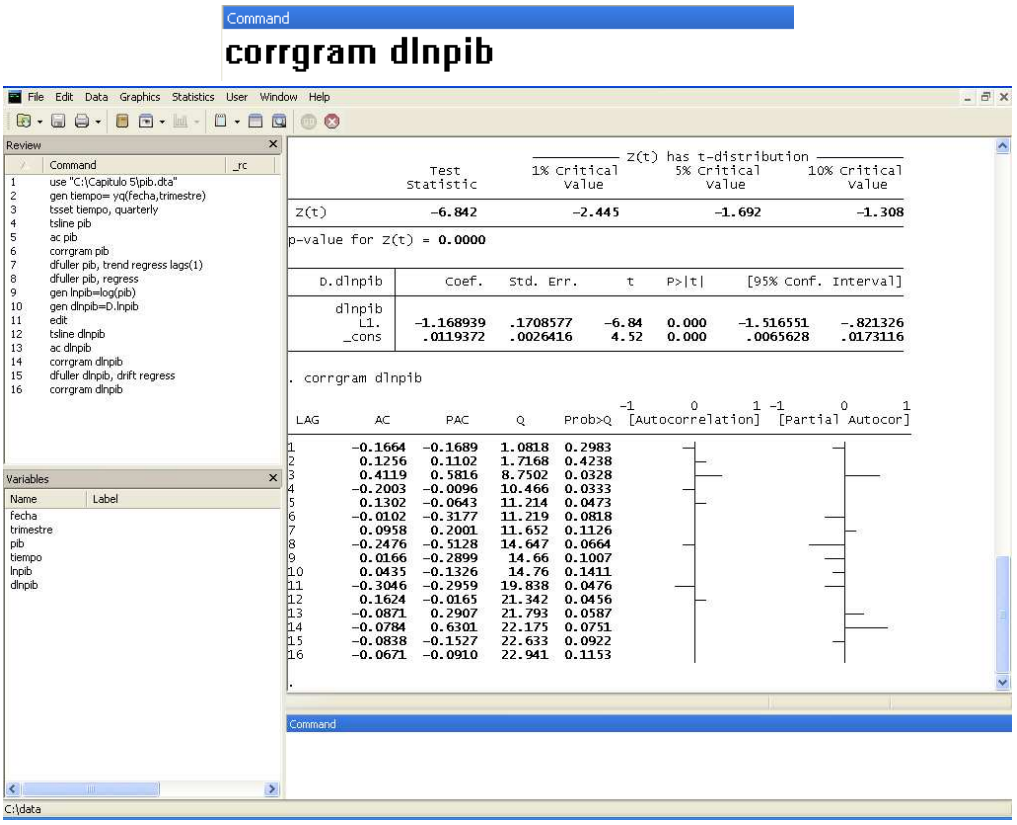
En este caso (*véase* figura 5.16) $\tau = -6.842$ y su probabilidad igual a cero; Indicando que se rechaze la hipotesis nula (a un nivel de significancia del 1%, 5% y 10%), por tanto el PIB es estacionario en su primera diferencia logaritmica. En otras palabras, es integrado de orden uno LNPIB~I(1). En este caso el valor de tau (τ) es negativo y su valor absoluto ($|\tau| = 6.842$), significa que $-1 < \hat{\rho} < 1$. Valor, comparable con lo valores absolutos criticos de MacKinnon ($|1\%| = 2.445$, $|5\%| = 1.692, |10\%| = 1.308$); $|\tau| > |1\%|, |5\%|$ y $|10\%|$, ratificando el rechazo de la hipotesis nula; igualmente el modelo debe estar especificado con intercept, sin tendencia y rezagos.

Estos resultados con los anteriores de ruido blanco hacen que la serie $\Delta LNPIB_t$ resulta débilmente estacionaria, por lo cual es posible encontrar su PGD a través de estructuras AR y MA que permita pronosticarla. Dado que resultó integrada de orden uno, sin componente estacional, la especiación del modelo para proyectarla es un Arima (*p*, *1*, *q*).

5.7.2 Identificación del proceso generador de datos (PGD)

1- Graficar los rho ($\hat{\rho}_p$) estimados del FAS y FAP para la serie en primeras diferencias logarítmicas del PIB ($\Delta LNPIB_t$), con el comando *corrgram* (*véase* figura 5.17).

Figura 5.17. Salida de Stata® para graficar el correlograma FAS y FAP de $\Delta LNPIB_t$



Fuente: cálculos autores.

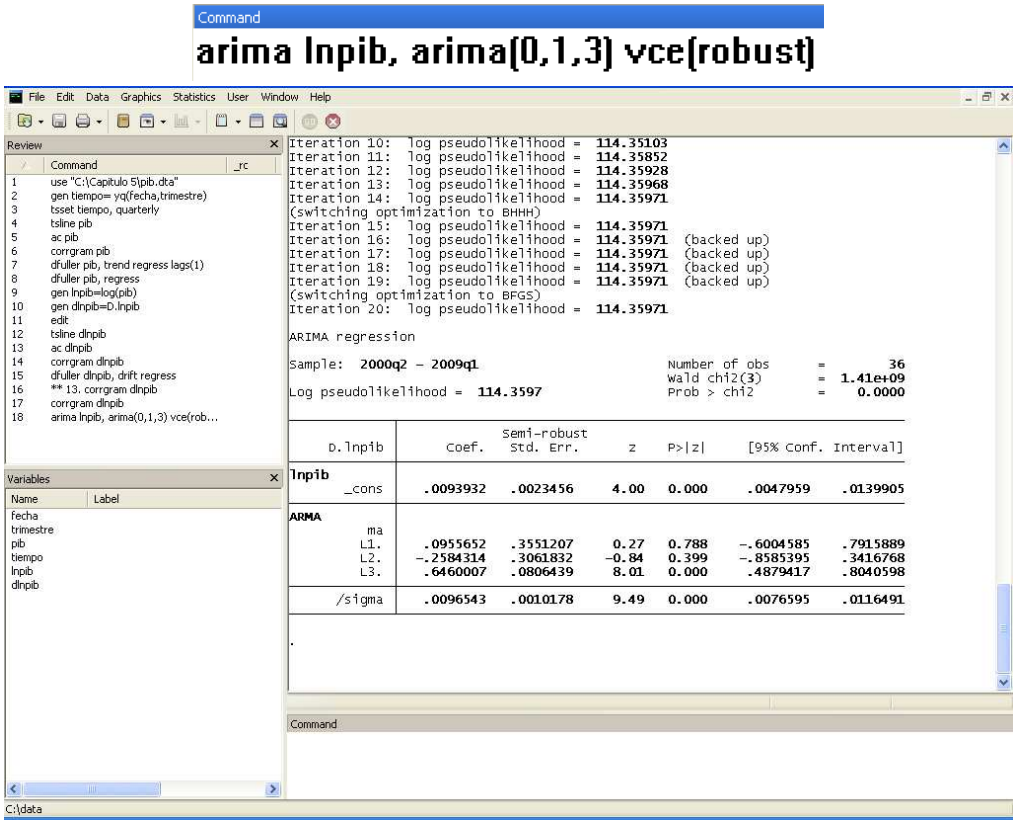
De esta manera y acorde con la figura 5.17 y cuadro 5.2 indica que el PGD para $\Delta LNPIB_t$ viene dado por un MA (3), dado el movimiento senoidal en FAP y el tercer rezago significativo en el FAS, fuera del intervalo de confianza. Entonces, el modelo a especificar y estimar para proyectar el PIB debería ser un Arima (0,1,3)

5.7.3 Estimación del modelo mediante máxima verosimilitud.

1- Estimar los parámetros $\hat{\theta}$ para el modelo Arima (0,1,3), ecuación 5.56, con el comando *arima lnpib, arima(0,1,3) vce(robust)* y raíces de polinomio mediante *armaroots* (véase figura 5.18 y 5.19).

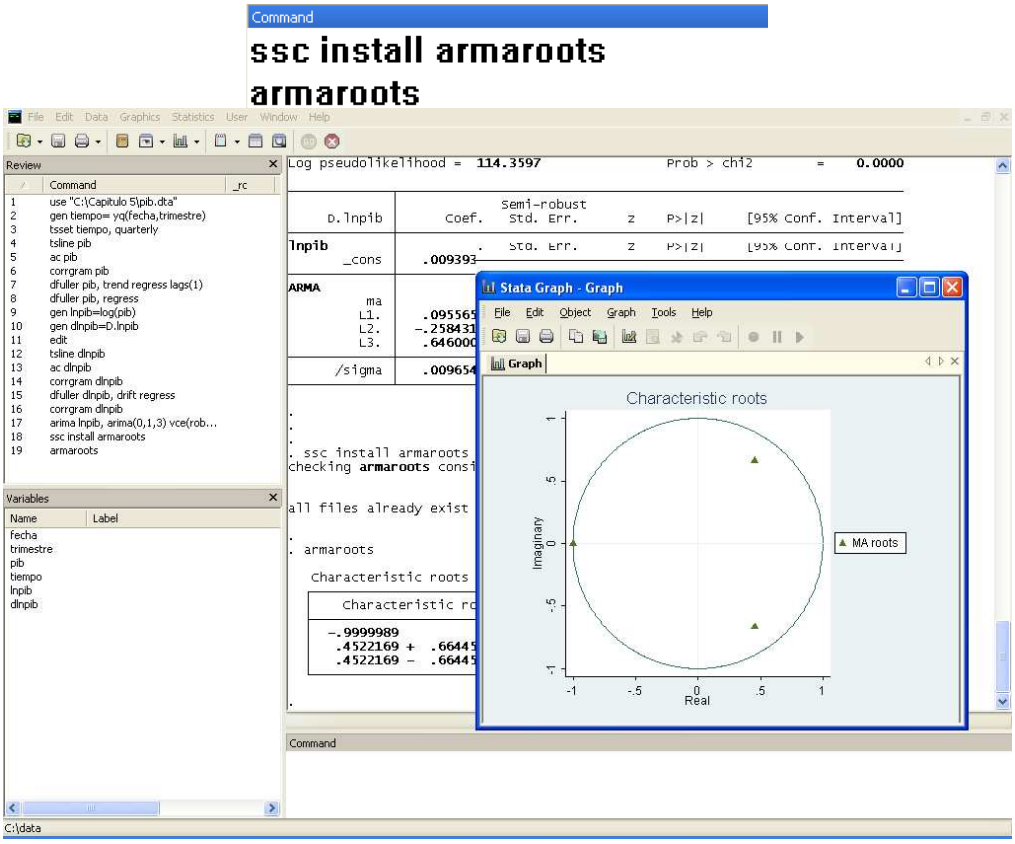
$$\Delta LNPIB_t = \delta - \theta_1 \Delta lnu_{t-1} - \theta_2 \Delta lnu_{t-2} - \theta_3 \Delta lnu_{t-3} + \Delta lnu_t \quad (5.56)$$

Figura 5.18. Salida de Stata® con los resultados de la estimación para el modelo Arima (0,1,3)



Fuente: cálculos autores.

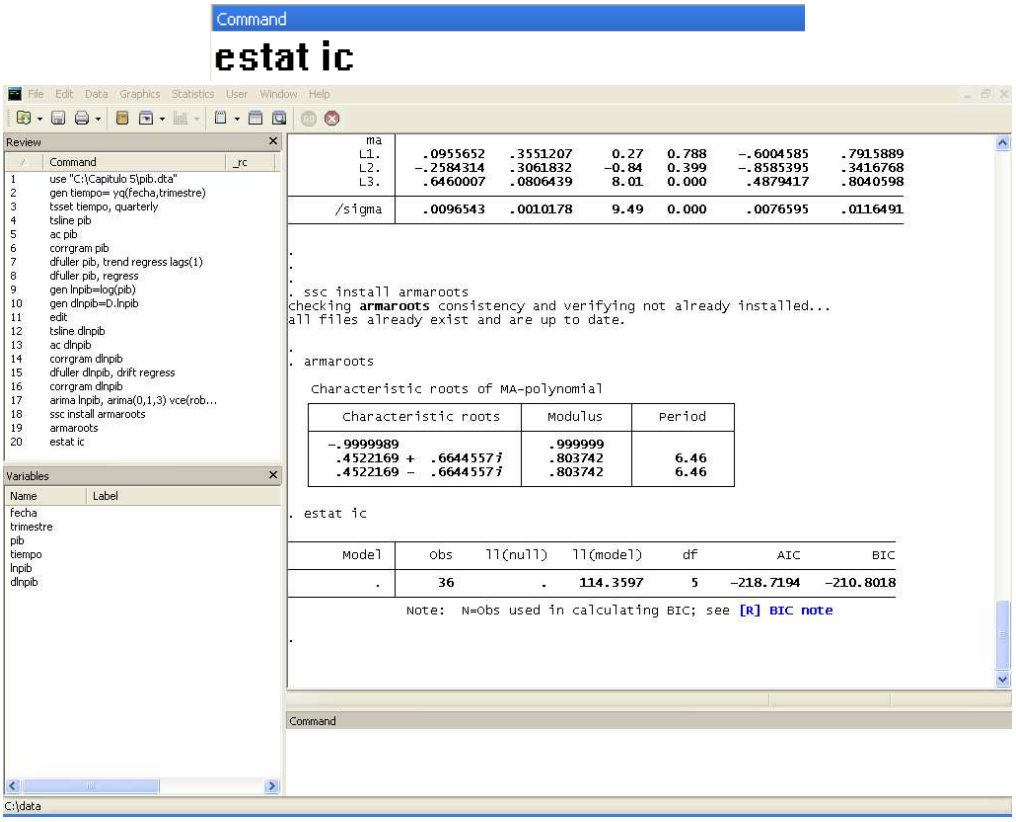
Figura 5.19. Salida de Stata® con los resultados de la estimación para el modelo Arima (0,1,3), raíces de polinomio característico y circulo unitario



Fuente: cálculos autores.

2- Estimar el criterio de Akaike en Stata®, mediante el comando *estat ic* (*véase* figura 5.20).

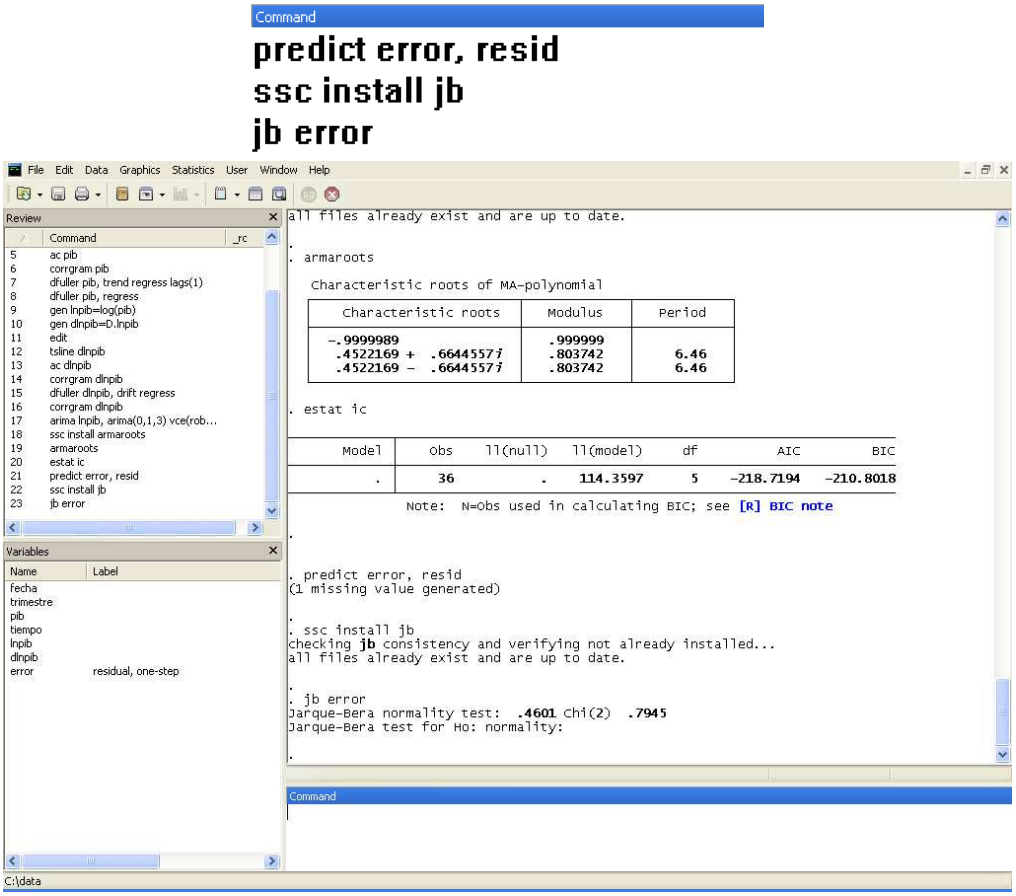
Figura 5.20. Salida de Stata® con los resultados de la estimación para el modelo Arima (0,1,3) y criterio de Akaike (AIC)



Fuente: cálculos autores.

3- Descargar y ejecutar la ayuda para instalar la prueba Jarque Bera en Stata®, mediante el comando *ssc install jb* (*véase* figura 5.21). Una vez realizada al estimación con la instrucción *arima lnpib, arima (0,1,3) vce(robust)*, se captura el error del modelo con el comando *predict error, residual* e inmediatamente después se ejecuta *jb error*; para comprobar si se distribuyen normalmente.

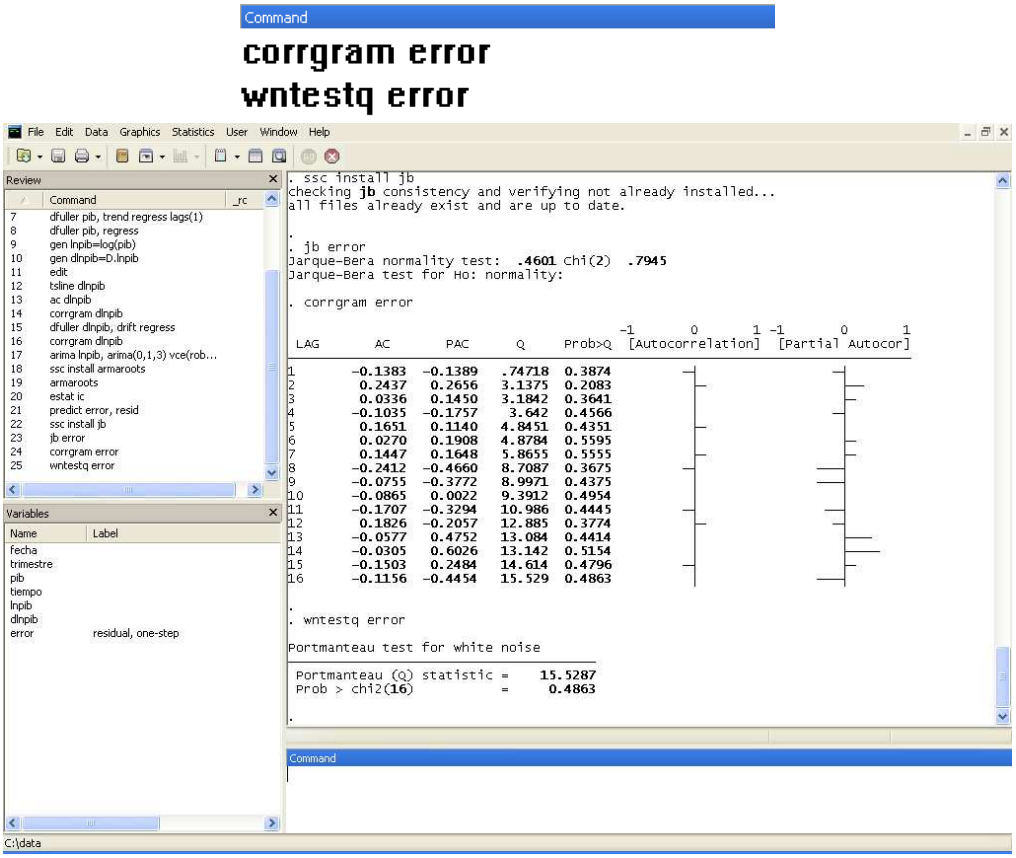
Figura 5.21. Salida de Stata® con los resultados de la estimación para el modelo Arima (0,1,3), y prueba de normalidad JB



Fuente: cálculos autores.

4- Realizar el correlograma (FAS y FAP), cálculo del estadístico Ljung Box y ***Q*** para determinar si el error es ruido blanco en Stata®, mediante los comandos *corrgram error* y *wntestq error* (*véase* figura 5.22) respectivamente.

Figura 5.22. Salida de Stata® con los resultados de las pruebas Ljung Box y Q de ruido blanco para el error



Fuente: cálculos autores.

5.7.4 Validación del modelo estimado.

$$\Delta\widehat{LNPIB}_t = 0.0093764 + 0.50420\Delta lnu_{t-3} \quad (5.57)$$

A partir de la figura 5.21, se plante la ecuación 5.57 con los valores estimados para $\hat{\delta} = 0.0093$ y $\hat{\theta} = 0.50420$ los cuales resultan estadísticamente significativos individualmente de acuerdo con la probabilidad Z (igual a cero, $p>|Z|$). También con la probabilidad Chi-cuadrado (igual a cero, $p>chi2$), el estadístico de Wald indica significancia conjunta, *véase* figura 5.14. En otras palabras, con los resultados anteriores los parámetros estimados son estadísticamente diferentes de cero e importantes para explicar el PGD de acuerdo con lo esperado previamente.

Los tres valores de las raíces de polinomio característico para el proceso MA son menores a uno, aunque ésta estructura por naturaleza es estacionaria el resultado lo confirma; señalando también que el modelo no se encuentra sobreparametrizado (no se estiman demasiados coeficientes que sobrecarguen el modelo, así ellos resulten significativos). Adicionalmente, estos tres valores se encuentran dentro del circulo unitario, indicando que la variable dependiente es estacionaria (*véase* figura 5.19).

Prosiguiendo el análisis, el criterio Akaike (AIC=-222,2156, *vése* figura 5.20) tiene un valor muy bajo aunque debería especificarse otro modelo Arima para ser comparado y seleccionar el valor AIC más pequeño entre ellos. Ahora, en cuanto a normalidad del error y de acuerdo con los la probabilidad Jarque Bera (0.6346, *véase* figura 5.21), los errores siguen una distribución normal, como debería suceder.

H_0: $\hat{\rho}_1 = \hat{\rho}_2 = \cdots =: \hat{\rho}_p = 0$; los errores son ruido blanco (donde p=16).

H_1: $\hat{\rho}_1 \neq \hat{\rho}_2 \neq \cdots \neq: \hat{\rho}_p \neq 0$; los errores no son ruido blanco.

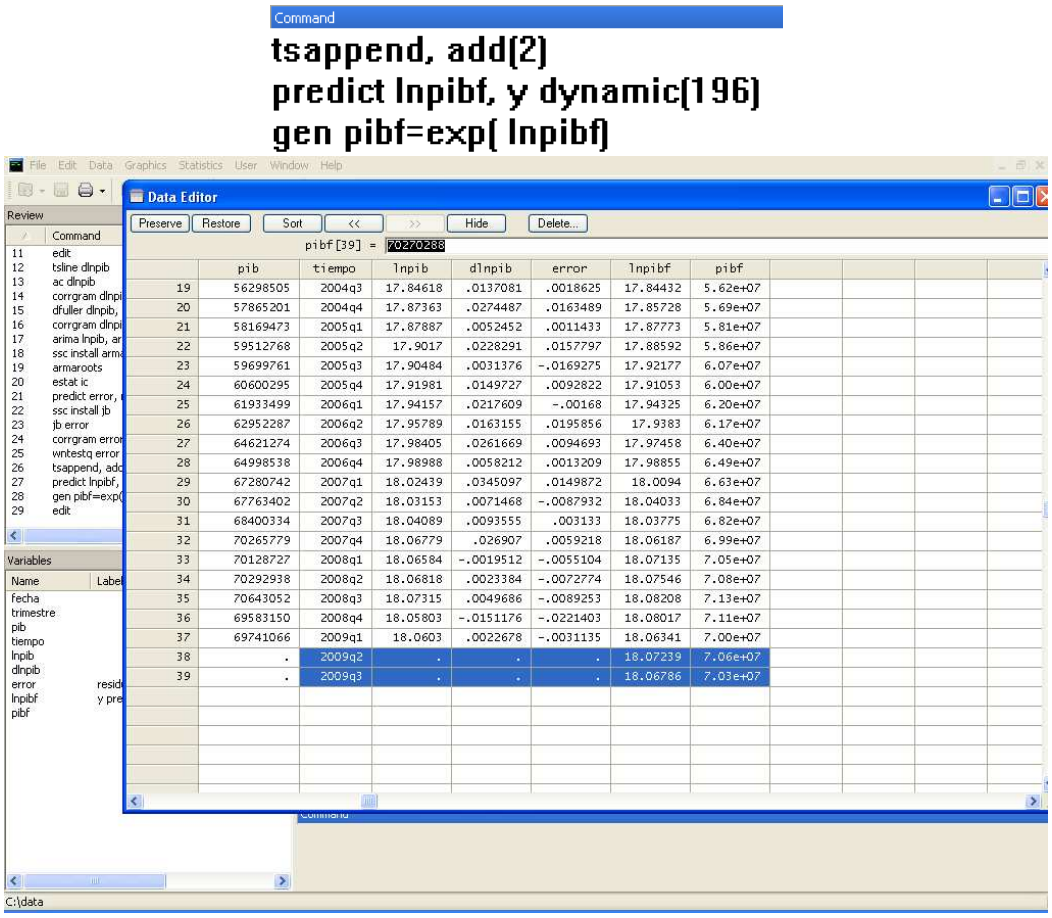
Finalmente y de acuerdo con las hipótesis anteriores, los errores son ruido blanco; dado que los resultados descritos del FAS, Ljung-Box y $\boldsymbol{Q}$ ayudan a comprobarlo. *Véase* en la figura 5.22 la columna que contiene las probabilidades del estdistico **Q** y $p>chi2$ (16) mayores al nivel de significancia 5% (0.05); indicando que no se rechaza la hipotesis nula de ruido blanco para el error.

5.7.5 Pronóstico con el modelo estimado y validado.

1- Adicionar previamente los periodos a pronosticar con el comando con el comando *tsappend*, add(2). Posteriormente, proyectar la variable retornando la primera diferencia a predecir en su nivel logarítmico inicial mediante la instrucción *predict lnpibf, y dynamic(196)* (las opciones y y dynamic(196), indican que la predicción se debe hacer sobre el modelo en logaritmos (antes de ser diferenciado), y las estimaciones dinamicas a partir de la última observación real disponible). Por último sacar el anti exponencial para

retornar a la variable PIB original utilizando *gen pibf=exp(lnpibf)* (con el comando *edit* observar los nuevos valores para el PIB, *véase* figura 5.23).

Figura 5.23. Salida de Stata® para pronosticar el PIB dos periodos



	pib	tiempo	lnpib	dlnpib	error	lnpibf	pibf
19	56298505	2004q3	17.84618	.0137081	.0018625	17.84432	5.62e+07
20	57865201	2004q4	17.87363	.0274487	.0163489	17.85728	5.69e+07
21	58169473	2005q1	17.87887	.0052452	.0011433	17.87773	5.81e+07
22	59512768	2005q2	17.9017	.0228291	.0157797	17.88592	5.86e+07
23	59699761	2005q3	17.90484	.0031376	-.0169275	17.92177	6.07e+07
24	60600295	2005q4	17.91981	.0149727	.0092822	17.91053	6.00e+07
25	61933499	2006q1	17.94157	.0217609	-.00168	17.94325	6.20e+07
26	62952287	2006q2	17.95789	.0163155	.0195856	17.9383	6.17e+07
27	64621274	2006q3	17.98405	.0261669	.0094693	17.97458	6.40e+07
28	64998538	2006q4	17.98988	.0058212	.0013209	17.98855	6.49e+07
29	67280742	2007q1	18.02439	.0345097	.0149872	18.0094	6.63e+07
30	67763402	2007q2	18.03153	.0071468	-.0087932	18.04033	6.84e+07
31	68400334	2007q3	18.04089	.0093555	.003133	18.03775	6.82e+07
32	70265779	2007q4	18.06779	.026907	.0059218	18.06187	6.99e+07
33	70128727	2008q1	18.06584	-.0019512	-.0055104	18.07135	7.05e+07
34	70292938	2008q2	18.06818	.0023384	-.0072774	18.07546	7.08e+07
35	70643052	2008q3	18.07315	.0049686	-.0089253	18.08208	7.13e+07
36	69583150	2008q4	18.05803	-.0151176	-.0221403	18.08017	7.11e+07
37	69741066	2009q1	18.0603	.0022678	-.0031135	18.06341	7.00e+07
38	.	2009q2	.	.	.	18.07239	7.06e+07
39	.	2009q3	.	.	.	18.06786	7.03e+07

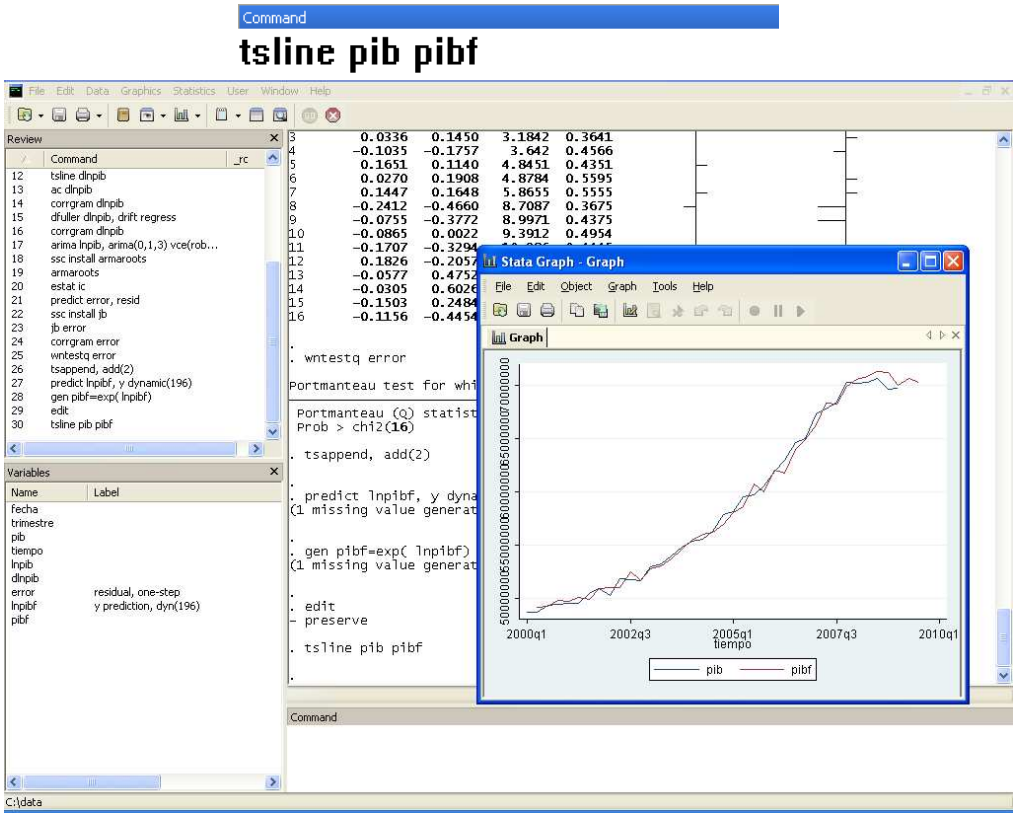
Fuente: cálculos autores.

En la figura 5.23 se puede apreciar que el nuevo valor proyectado para el PIB en el segundo trimestre de 2009 (2009-II) corresponde a 70’088.648. El mismo, se involucra como un nuevo valor observado en la serie original y se reestima el modelo Arima (0, 1, 3) de acuerdo a los procedimiento anteriores, para predecir 2009-III, siempre y cuando los resultados anteriores analizados no se alteren.

5.7.6 Validación del pronóstico.

1- Graficar (comando *tsline*) *pib* y *pibf,* para conocer si es similar el original al proyectado (*véase* línea roja figura 5.24).

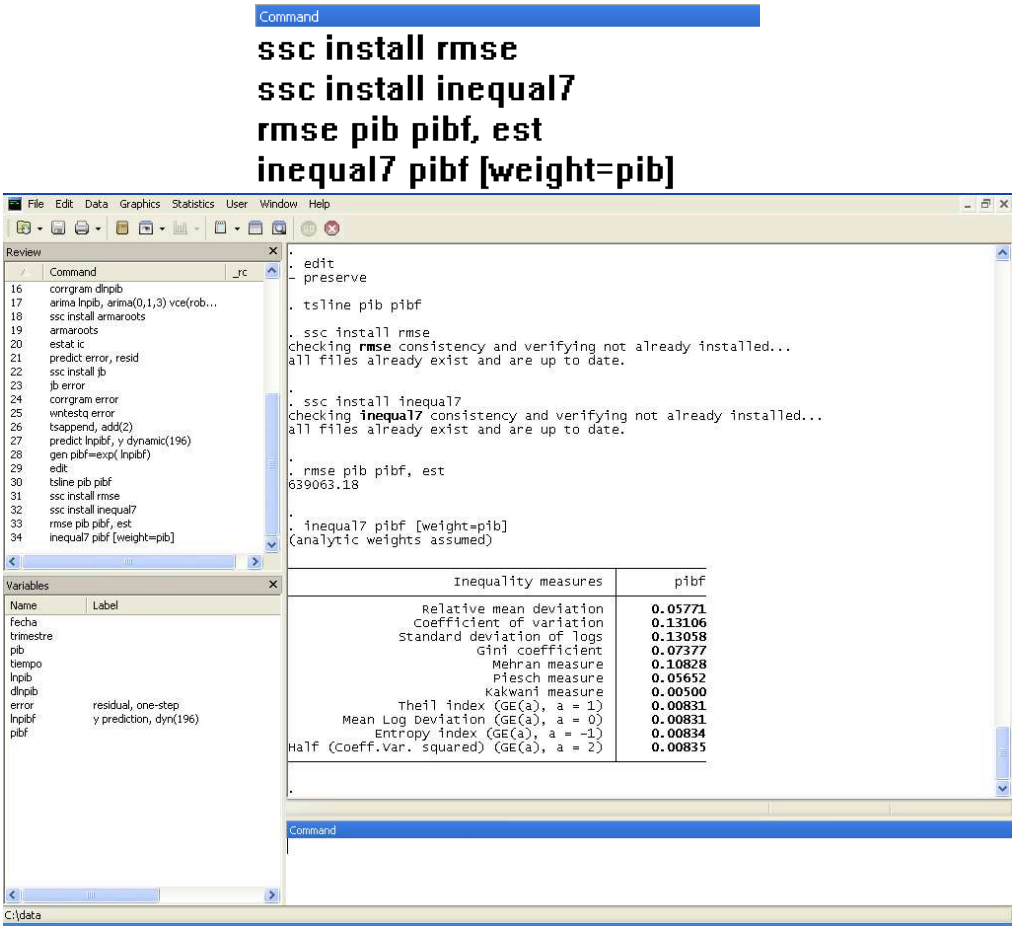
Figura 5.24. Salida de Stata® con las gráficas observada y proyectada del PIB



Fuente: cálculos autores.

2- Descargar y ejecutar la ayuda para instalar la raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC) y coeficiente de Theil (CT) en Stata®, mediante los comandos *ssc install rmse* y *ssc install inequal7* respectivamente (*véase* figura 5.25). Una vez realizada al estimación se ejecuta la instrucción *rmse pib pibf, est* y *inequal7 pibf [weight=pib]*.

Figura 5.25. Salida de Stata® con los resultados de la raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC) y coeficiente de Theil (CT, *inequal7*)



Fuente: cálculos autores.

En la figura 5.25, se aprecia que el valor observado y proyectado del PIB llevan trayectorias similares, no obstante los valores pronosticados del los últimos periodos sobrepasan los observados, presumiendo que su pronóstico puede estar por encima en 628854,5 (RCPSEC, *rmse*) al real que pueda presentarse en este periodo. Por tanto, a 70'088.648 se le debe restar 628854,5 para tener una mejor aproximación de la proyección a la futura observada en 2009-II.

También en la figura 5.25, la predicción está bien ajustada de acuerdo con el coeficiente de Theil (0.00831) cercano a cero, ocurriría lo contario cuando este

tienda a uno. Con este estudio de caso finaliza el procedimiento de la metodología Box-Jenkins para series no estacionales, a continuación es analizado el IPC en Colombia que incluye estacionalidad.

5.8 Estudio de caso: IPC colombiano

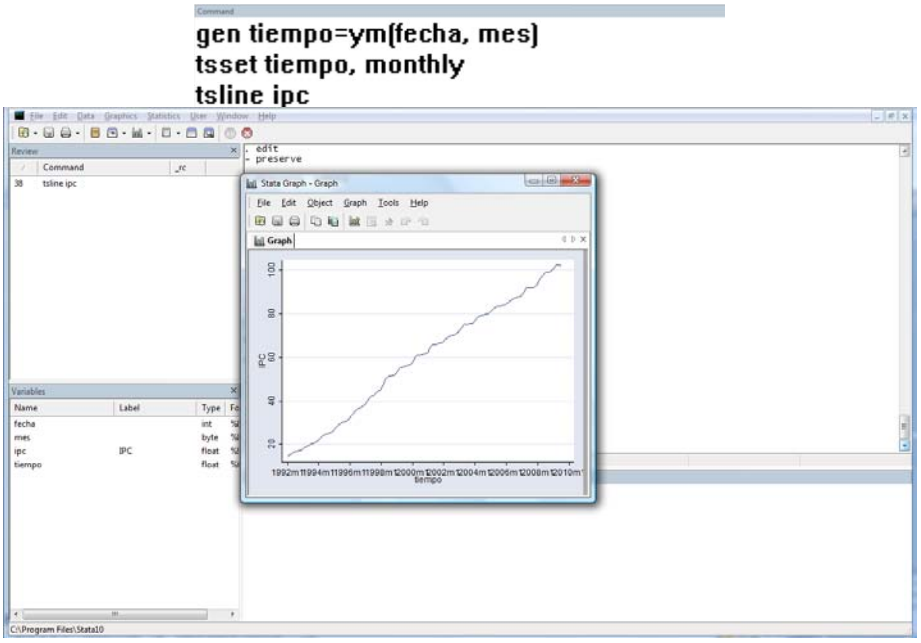
Con la información del IPC colombiano se pretende estimar un modelo Sarima para realizar su predicción un periodo adelante, para el mes de agosto de 2009 ($\widehat{IPC}_{t+1}$ =2009-VIII), paso a paso bajo la metodología Box-Jenkins de la siguiente manera:

5.8.1 Análisis de estacionalidad y desetacionalización con estacionariedad implícita falsa

5.8.1.1 Análisis gráfico para detectar estacionalidad

1- Configurar el programa para que reconozca el IPC_t como serie de tiempo trimestral, con el comando *gen* y *tsset* (*véase* figura 5.26). Luego, graficarla con la instrucción *tsline*.

Figura 5.26. Especificación de serie de tiempo y graficas



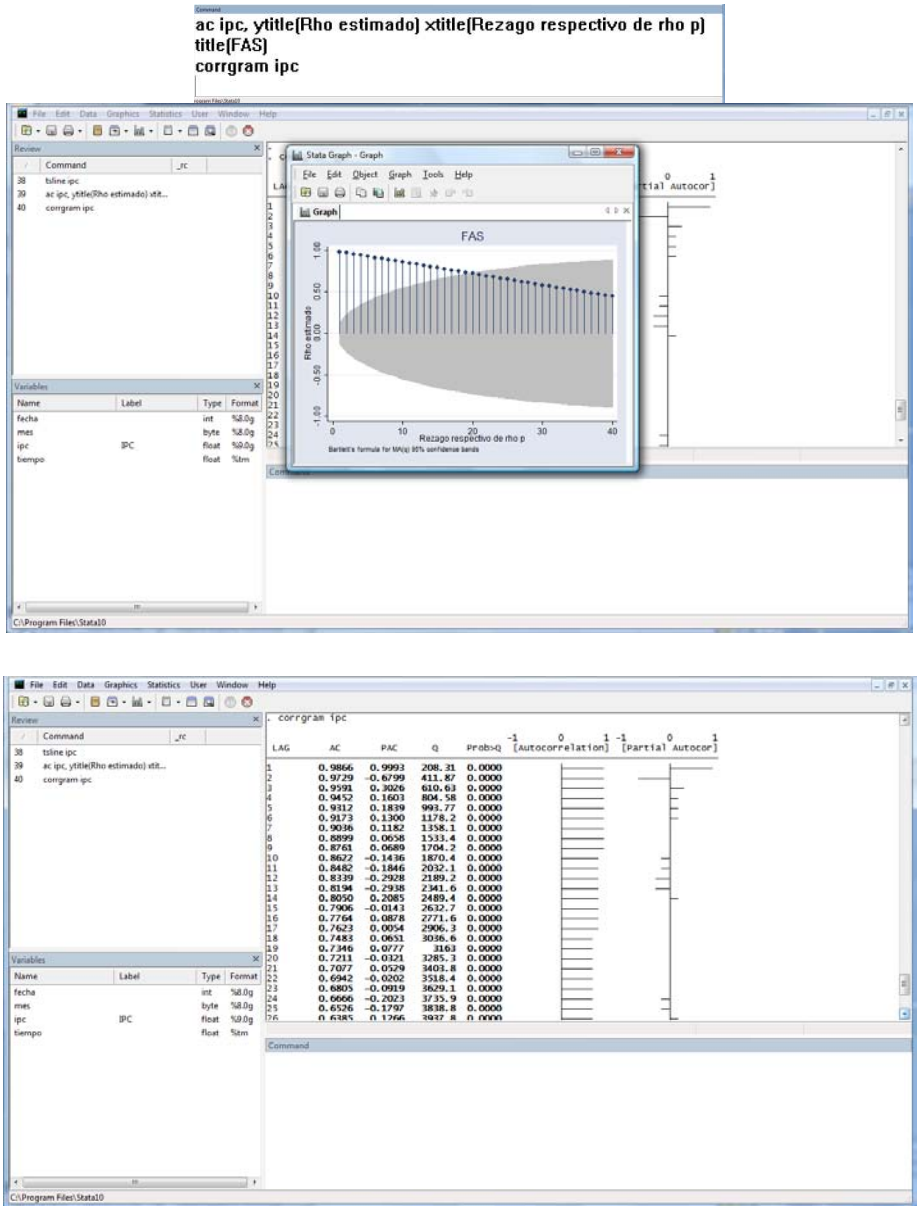
Fuente: cálculos autores.

En este caso, el IPC_t muestra una tendencia creciente lineal, con leves movimientos ondulatorio (*véase* figura 5.26). Como se mencionó anteriormente, la variable cuestionada presenta media y varianza inestables entre 1992-I y 2009-VII; dada la presencia tendencial, estacional e irregular en ella. Conllevando a que la serie (IPC) es estacional, no es estacionaria en media y varianza.

5.8.1.2 Análisis gráfico mediante el correlograma de la función de autocorrelación simple (FAS) para detectar estacionariedad

1- Graficar los rho ($\hat{\rho}_p$) estimados del FAS, con el comando *ac* y *corrgram* (*véase* figura 5.27).

Figura 5.27. Salida de Stata® para graficar el FAS del IPC



Fuente: cálculos autores.

De esta manera, la figura 5.27 indica no estacionariedad para el IPC dado que exterioriza estimaciones $\hat{\rho}_p$ decrecientes exponencialmente de mayor a menor (entre 1 y 0), llegando a cero, tomando valores negativos y la mayor parte de ellos se encuentran fuera de su intervalo de confianza. Por otra parte, exhibe los valores del FAS ($\hat{\rho}_p$, *véanse* columnas AC), graficados en el correlograma anterior, para el

IPC colombiano (decrecen exponencialmente, desde 0.9866 hasta 0.6385) respectivamente.

$H_0: \hat{\rho}_1 = \hat{\rho}_2 = \cdots =: \hat{\rho}_p = 0$; La serie IPC es ruido blanco, implicando automaticamente estacionariedad en ella (donde p=40).

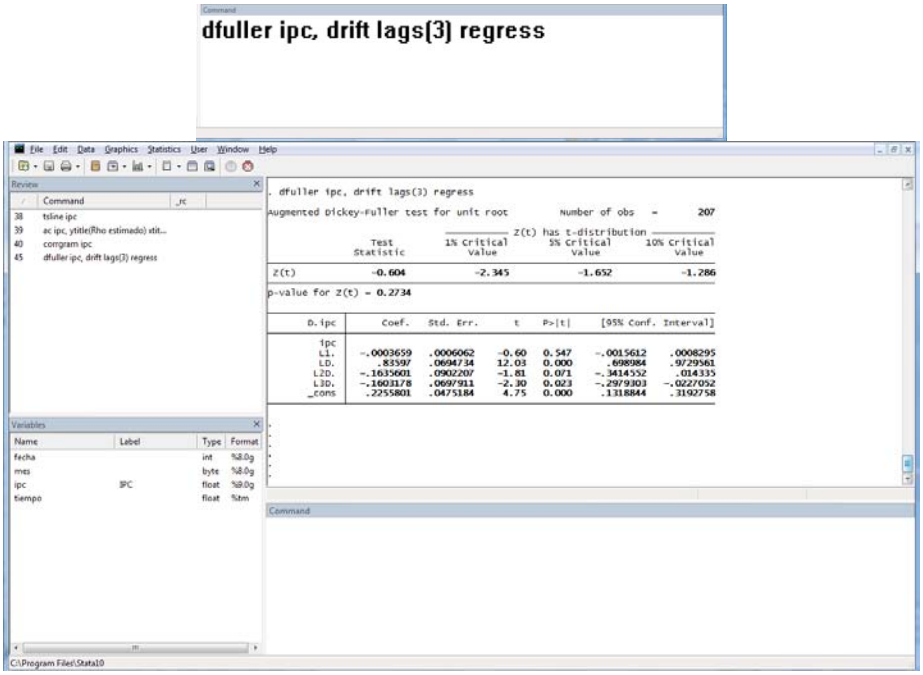
$H_1: \hat{\rho}_1 \neq \hat{\rho}_2 \neq \cdots \neq: \hat{\rho}_p \neq 0$; La serie IPC no es ruido blanco, implicando que posiblemente es estacionaria.

Adicionalmente, los resultados descritos para FAS y $\boldsymbol{Q}$ ayudan a comprobar que el IPC no es ruido blanco. *Véase* ultima columna en la figura 5.27 que contiene las probabilidades del estdistico **Q** Ljung-Box**,** indicando (con nivel de significancia del 5%) que se rechaza la hipotesis nula de ruido blanco para el PIB.

5.8.1.3 Análisis de raíz unitaria Dickey-Fuller aumentado (DFA) para detectar estacionariedad

1- Estimar tau de DFA con intercepto y tres rezagos, con el comando *dfuller ipc, drift lags(3) regress* (*véase* figura 5.28 y ecuación 5.58).

Figura 5.28. Salida de Stata® para realizar DFA



Fuente: cálculos autores.

$\Delta IPC_t = \alpha + \delta IPC_{t-1} + \gamma_1 \Delta IPC_{t-1} + \gamma_2 \Delta IPC_{t-2} + \gamma_3 \Delta IPC_{t-3} + u_t$, con intercepto y tres rezagos; $\delta = (\rho - 1)$ (5.58)

H_0: $\delta = 0$; $\rho = 1$; la serie IPC contiene raíz unitaria, equivale a decir que es una caminata aleatoria o simplemente no es estacionaria.

H_1: $\delta \neq 0$; $\rho \neq 1$; la serie IPC no contiene raíz unitaria equivale a decir que no es una caminata aleatoria o simplemente es estacionaria.

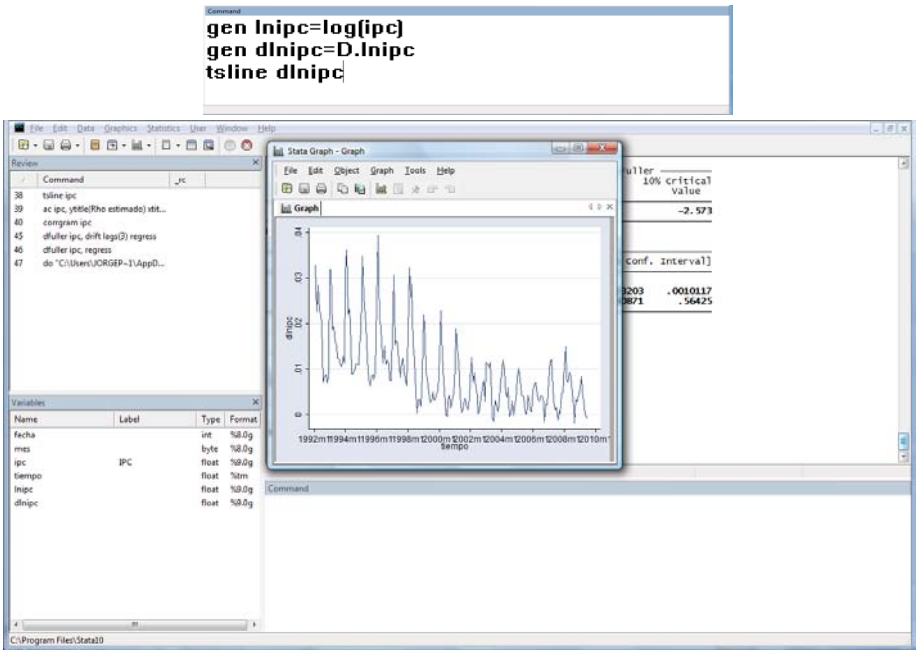
En este caso (*véase* figura 5.28) $\tau = -0.604$ y su probabilidad igual a 0.3734. Indicando que no se rechaze la hipotesis nula (a un nivel de significancia del 1%, 5% y 10%), por tanto el IPC no es estacionario en su nivel. Descartando que el mismo es integrada de orden cero $IPC \sim I(0)$. En este caso el valor de tau (τ) es negativo y su valor absoluto es 0.604 $(|\tau| = 0.604)$, significa que $-1 < \hat{\rho} < 1$. Valor, comparable con lo valores absolutos criticos de MacKinnon $(|1\%| = 2.345,\ |5\%| = 1.652, |10\%| = 1.286)$; $|\tau| < |1\%|, |5\%|$ y $|10\%|$, ratificando el no rechazo de la hipotesis nula. Igualmente el modelo debe estar especificado con intercepto y tres

rezagos debido a que sus valores estaditicos (4.75, -2.30, -1.81 y 12.03) son estadísticamente significativos.

5.8.1.4 Transformación en primeras diferencia logarítmicas para convertir el IPC en una serie estacionaria y detectar estacionalidad

1- Generar una nueva variable, con el comando *gen*, que contenga la primera diferencia logarítmica de IPC ($\Delta LNipc_t$) y graficarla con el comando *tsline* (*véase* figura 5.29).

Figura 5.29. Salida de Stata® para crear una variable en primera diferencias logarítmicas



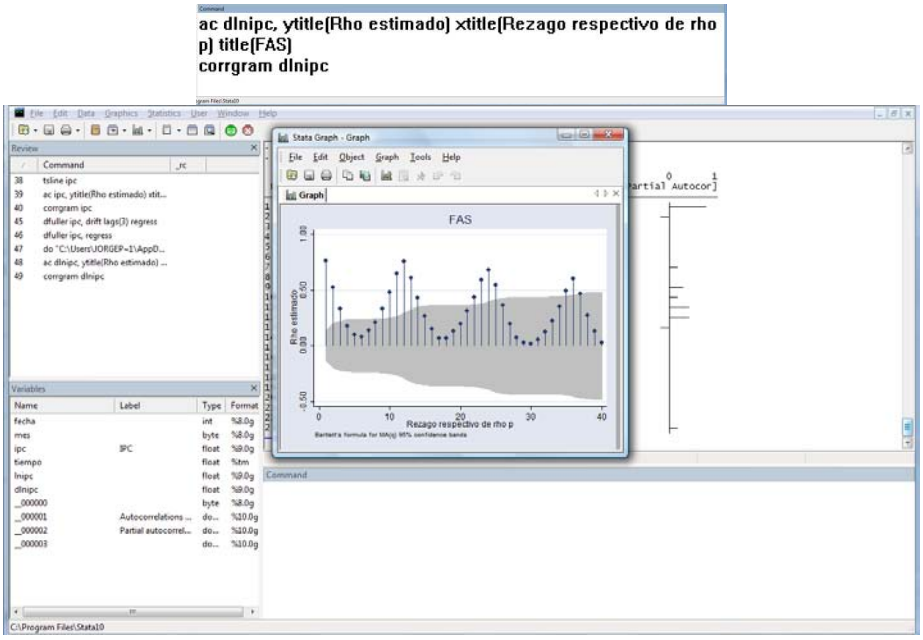
Fuente: cálculos autores.

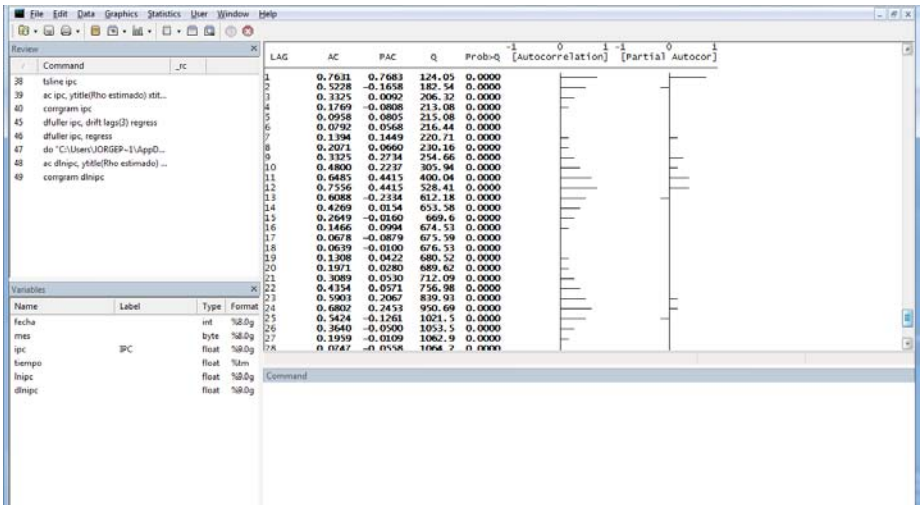
En este caso, la primera diferencia en logaritmo del IPC_t ($\Delta LNipc_t$) muestra una tendencia decreciente con movimientos senoidales, reflejando el componente estacional (*véase* figura 5.29). Económicamente, ésta transformación ($\Delta LNipc_t$) representa el incremento porcentual mes a mes de los precios al consumidor, en otras palabras es la tasa de inflación mensual. Aparentemente la variable cuestionada presenta media y varianza estables entre 1992-II y 2009-VII. Donde

presumiblemente la serie en primeras diferencias logarítmica es estacionaria en media y varianza, pero el componente estacional hace que la serie no resulte estacionaria, así aparentemente tenga movimiento senoidal sin tendencia.

2- Graficar los rho ($\hat{\rho}_p$) estimados del FAS para la series en primeras diferencias logarítmicas del IPC ($\Delta LNipc_t$), con el comando *ac* y *corrgram* (*véase* figura 5.30).

Figura 5.30. Salida de Stata® para graficar el FAS de $\Delta LNipc_t$



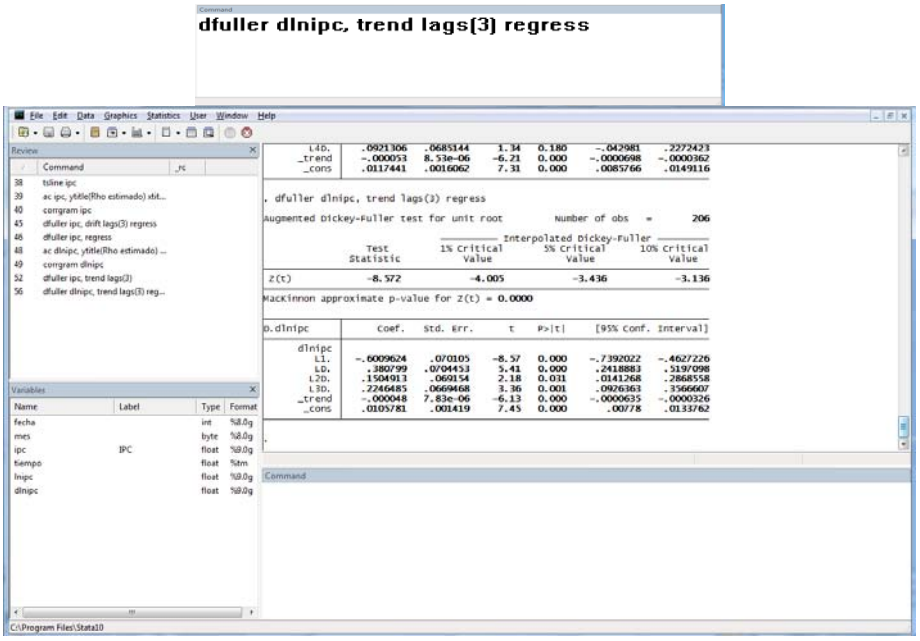


Fuente: cálculos autores.

De esta manera, la figura 5.30 indica posible estacionariedad para el IPC dado que exterioriza estimaciones $\hat{\rho}_p$ senoidales intercaladas, sin embargo la mayor parte de ellos se encuentran fuera su intervalo de confianza. Adicionalmente, los valores del FAS ($\hat{\rho}_p$, *véase* columnas AC) se intercalan (desde 0.7631 hasta 0.0747); aunque en los rezagos 12 y 24 sobresalen repetitivamente cada 12 meses. Detectando aun mejor el componente estacional del IPC cada diciembre.

3- Estimar tau de DFA con intercepto, tendencia y tres rezagos para $\Delta LNipc_t$, con el comando *dfuller dlnipc, trend lags(3) regress* (*véase* figura 5.31).

Figura 5.31. Salida de Stata® para realizar DFA de $\Delta LNPIB_t$



Fuente: cálculos autores.

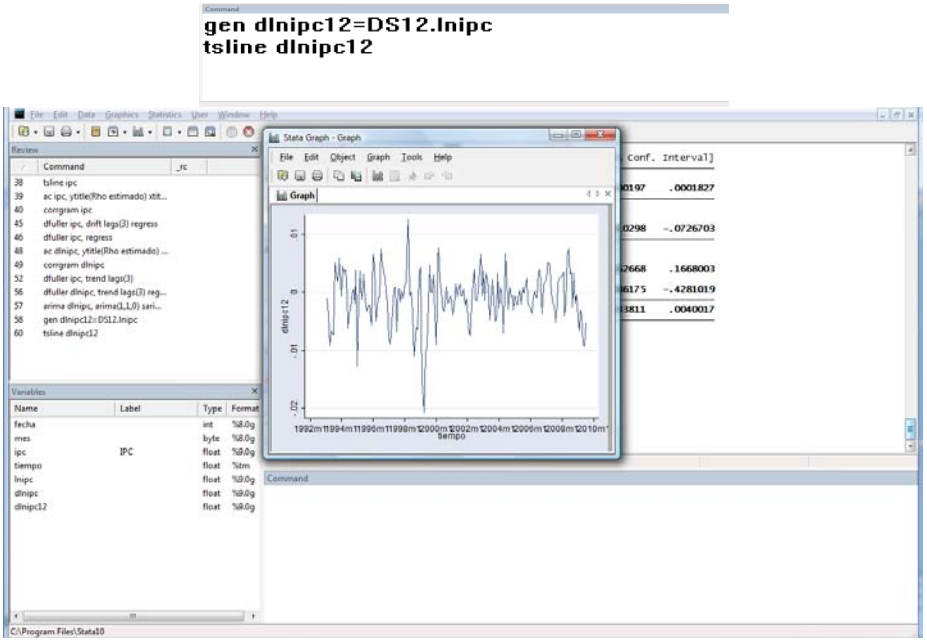
En este caso (*véase* figura 5.31) $\tau = -8.572$ y su probabilidad igual a cero; Indicando que se rechaze la hipotesis nula (a un nivel de significancia del 1%, 5% y 10%), por tanto el IPC es estacionario en su primera diferencia logaritmica. En otras palabras, es integrado de orden uno LNIPC~I(1). En este caso el valor de tau (τ) es negativo y su valor absoluto ($|\tau| = 8.572$), significa que $-1 < \hat{\rho} < 1$. Valor, comparable con lo valores absolutos criticos de MacKinnon ($|1\%| = 4.005$, $|5\%| = 3.436, |10\%| = 3.136$); $|\tau| > |1\%|, |5\%|$ y $|10\%|$, ratificando el rechazo de la hipotesis nula; igualmente el modelo debe estar especificado con intercepto, tendencia y tres rezagos.

No obstante a lo anterior, el análisis realizado con las pruebas de raíz unitaria DFA es invalido por la presencia de estacionalidad en el IPC; dado que la series contiene raices estacionarias y no regulares como la expuesta en la figura 5.31. Entonces, no importa si la serie resulta ser estacionaria con DFA, el correlograma muestra un componete estacional cada 12 meses, por esta razón se debe desestacionalizar el IPC. Realizando una diferencia estacional a la primera diferencia logarítmica ($\Delta LNipc_t$) y trabajar con el IPC desestacionalizado.

5.8.1.5 Diferencia estacional (*s=12*) para desestacionalizar el IPC y estacionariedad débil

1- Generar una nueva variable, con el comando *gen* y *DS12*, que contenga la diferencia estacional cada 12 periodos ($\Delta_{12}LNipc_t$) de la primera diferencia logarítmica de IPC ($\Delta LNipc_t$) y graficarla con el comando *tsline* (*véase* figura 5.32).

Figura 5.32. Salida de Stata® para crear una variable en primera diferencias estacionales logarítmicas



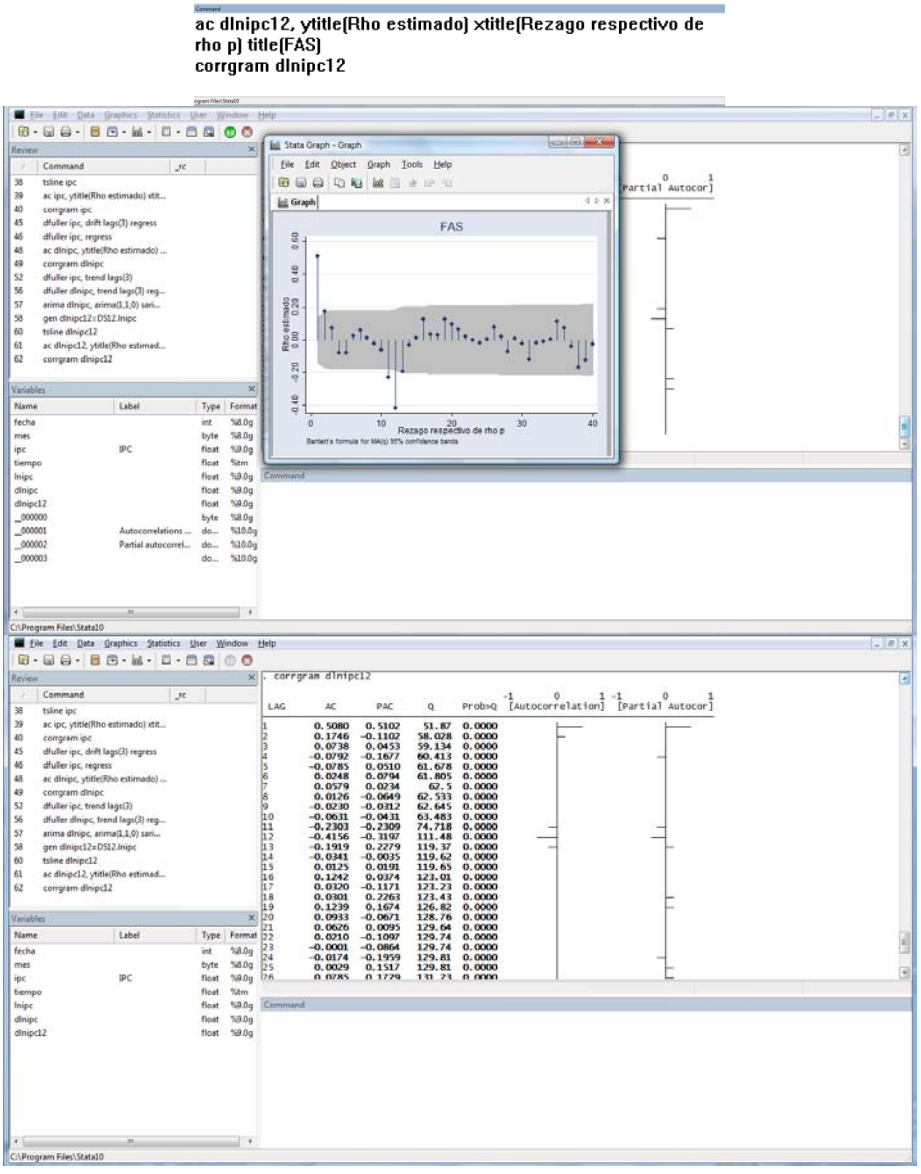
Fuente: cálculos autores.

En este caso, la diferencia estacional cada 12 periodos ($\Delta_{12}LNipc_t$) de la primera diferencia logarítmica de IPC ($\Delta LNipc_t$) o IPC desestacionalizado muestra movimientos senoidales sin tendencia (*véase* figura 5.32). Económicamente, ésta transformación ($\Delta_{12}LNipc_t$) representa el incremento porcentual anual, mes a mes, de los precios al consumidor; en otras palabras es la tasa de inflación anualizada. Aparentemente la variable cuestionada presenta media y varianza estables entre 1992-II y 2009-VII. Donde presumiblemente la serie en primeras diferencias

estacionales logarítmica es estacionaria en media y varianza, desvaneciendo así el componente estacional.

2- Graficar los rho ($\hat{\rho}_p$) estimados del FAS para la series en primeras diferencias estacionales logarítmicas del IPC ($\Delta_{12} LNipc_t$), con el comando *ac* y *corrgram* (*véase* figura 5.33).

Figura 5.33. Salida de Stata® para graficar el FAS de $\Delta_{12} LNipc_t$



Fuente: cálculos autores.

De esta manera, en la figura 5.33 se puede apreciar como desvaneció por completo la estacionalidad cada diciembre, indicando posible estacionariedad débil para el IPC dado que exterioriza estimaciones $\hat{\rho}_p$ senoidales intercaladas y la mayor parte de ellos se encuentran dentro de su intervalo de confianza. Adicionalmente, los valores del FAS ($\hat{\rho}_p$, *véase* columnas AC) se intercalan (desde 0.5080 hasta 0.0785), respectivamente.

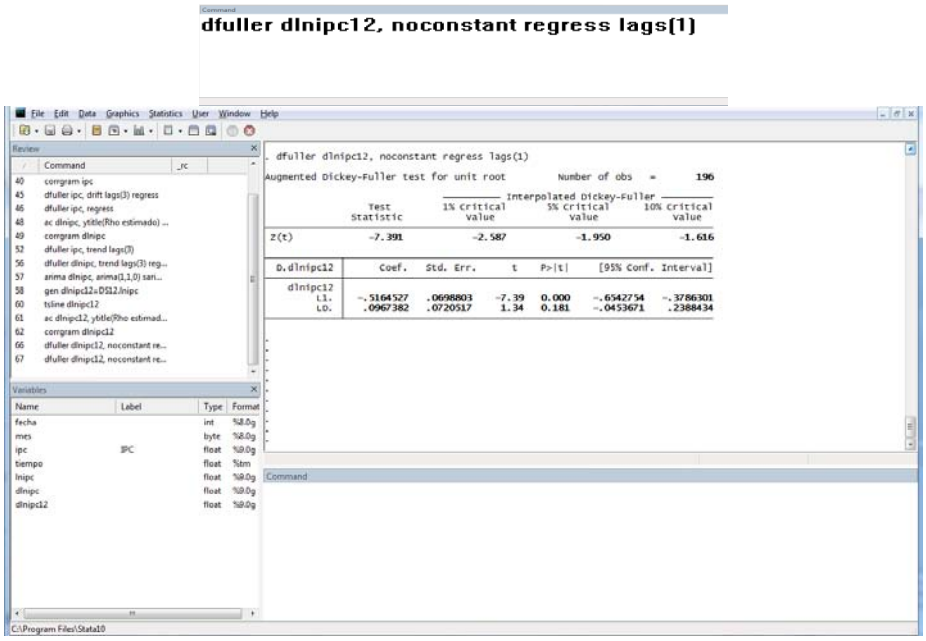
$H_0: \hat{\rho}_1 = \hat{\rho}_2 = \cdots =: \hat{\rho}_p = 0$; La serie $\Delta_{12} LNipc_t$ es ruido blanco, implicando automaticamente estacionariedad en ella (donde p=28).

$H_1: \hat{\rho}_1 \neq \hat{\rho}_2 \neq \cdots \neq: \hat{\rho}_p \neq 0$; La serie $\Delta_{12} LNipc_t$no es ruido blanco, implicando que posiblemente es débilmente estacionaria.

Adicionalmente, los resultados descritos para FAS y $\boldsymbol{Q}$ ayudan a comprobar que el $\Delta_{12} LNipc_t$ no es ruido blanco. *Véase* ultima columna en la figura 5.33 que contiene las probabilidades del estdistico **Q** Ljung-Box**,** indicando (con nivel de significancia del 5%) que se rechaza la hipotesis nula de ruido blanco para el $\Delta_{12} LNipc_t$. Este resultado, indica que a la serie IPC, en primeras diferencias estacionales logarítmicas, es posible encontrar su PGD para poder predercila, caso contrario hubiese pasado sino se rechaza la hipotesis nula.

3- Estimar tau de DFA sin intercepto y tendencia, con un rezagos para $\Delta_{12} LNipc_t$, con el comando *dfuller dlnipc12, noconstant lags(1) regress* (*véase* figura 5.34 y ecuación 5.59).

Figura 5.34. Salida de Stata® para realizar DFA de $\Delta_{12}LNPIB_t$



Fuente: cálculos autores.

$\Delta_{12}^2 lnipc_t = \delta\Delta_{12} lnipc_{t-1} + \gamma\Delta_{12}^2 lnipc_{t-1} + u_t$ con un rezago, sin intercepto y tendencia; $\delta = (\rho - 1)$ (5.59).

H_0: $\delta = 0$; $\rho = 1$; la serie $\Delta_{12}LNipc_t$contiene raíz unitaria, equivale a decir que es una caminata aleatoria o simplemente no es estacionaria.

H_1: $\delta \neq 0$; $\rho \neq 1$; la serie $\Delta_{12}LNipc_t$no contiene raíz unitaria equivale a decir que no es una caminata aleatoria o simplemente es estacionaria.

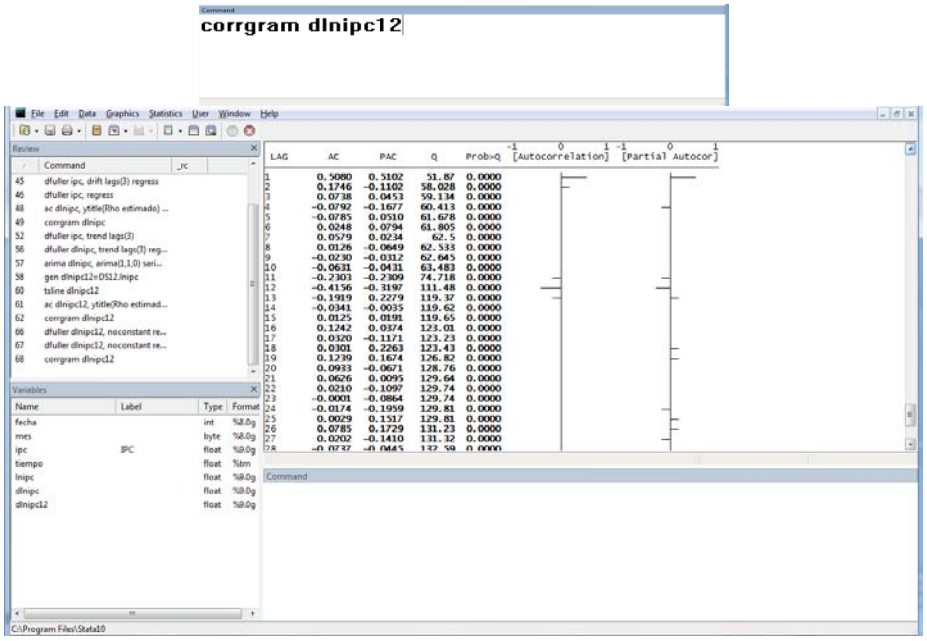
En este caso (*véase* figura 5.34) $\tau = -7.391$ y su probabilidad igual a cero; Indicando que se rechaze la hipotesis nula (a un nivel de significancia del 1%, 5% y 10%), por tanto el IPC es estacionario en su primera diferencia estacionales logaritmica. En otras palabras, es integrado de orden uno $\Delta_{12}LN$IPC~I(1). En este caso el valor de tau (τ) es negativo y su valor absoluto ($|\tau| = 7.391$), significa que $-1 < \hat{\rho} < 1$. Valor, comparable con lo valores absolutos criticos de MacKinnon ($|1\%| = 2.587$, $|5\%| = 1.950, |10\%| = 1.616$); $|\tau| > |1\%|, |5\%|$ y $|10\%|$, ratificando el rechazo de la hipotesis nula; igualmente el modelo debe estar especificado con un rezago, sin intercepto y tendencia.

Estos resultados con los anteriores de ruido blanco hacen que la serie $\Delta_{12} LN\text{IPC}_t$ resulta débilmente estacionaria, por lo cual es posible encontrar su PGD a través de estructuras AR, MA, SMA y SAR que permita pronosticarla. Dado que la series desestacionalizada resultó integrada de orden uno, la especiación del modelo para proyectarla es un Sarima $(p, 1, q)$ (P,1,Q)12.

5.8.2 Identificación del proceso generador de datos (PGD)

1- Graficar los rho ($\hat{\rho}_p$) estimados del FAS y FAP para la serie en primeras diferencias estacionales logarítmicas del IPC ($\Delta_{12} LNipc_t$), con el comando *corrgram* (*véase* figura 5.35).

Figura 5.35. Salida de Stata® para graficar el correlograma FAS y FAP de $\Delta_{12} ipc_t$



Fuente: cálculos autores.

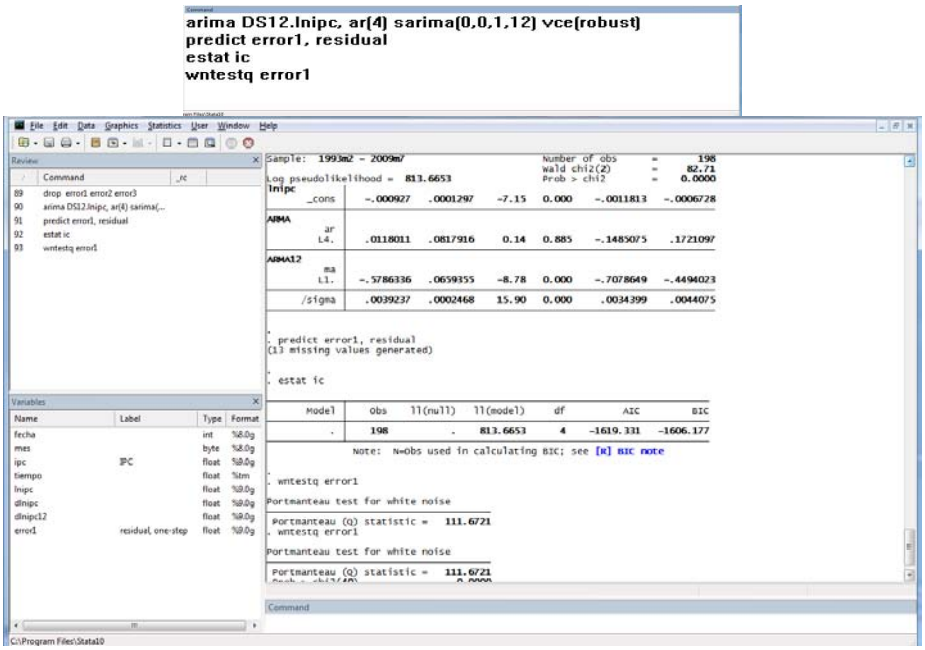
De esta manera y acorde con la figura 5.35 y cuadro 5.2 indica que el PGD para $\Delta_{12} ipc_t$ viene dado por un MA (1) y SMA (1)12 dado el movimiento senoidal en FAP, el primer y doceavo rezagos significativos en el FAS, fuera del intervalo de confianza. Entonces, el modelo a especificar y estimar para proyectar el IPC debería ser un Sarima (0,1,1) (0,1,1)12. Ésta ultima parte, significa una diferencia estacional y un componente MA cada 12 periodos ($s=12$) dentro del PGD del IPC.

No obstante y acorde con la figura y cuadro 5.2, la especificación se convierte en un ensayo error; por esto, también posiblemente se puedan estimar las especificaciones Sarima (4, 1, 0) $(0,1,1)_{12}$ y Sarima (1, 1, 4) $(0,1,1)_{12}$. Al final, en la validación del modelo, se elige en el que tenga el criterio Akaike más pequeño.

5.8.3 Estimación de los modelos mediante máxima verosimilitud.

1- Estimar los parámetros $\hat{\phi}$ y $\hat{\theta}$ para el modelo Sarima (4, 1, 0) $(0,1,1)_{12}$, con el comando *arima DS12.lnipc, ar(4) sarima(0,0,1,12) vce(robust)*. Adicional el criterio de Akaike y prueba ruido blanco en los errores mediante las instrucciones *estat ic*, *predict error1* y *wntestq error1*, respectivamente (*véase* figura 5.36).

Figura 5.36. Salida de Stata® con los resultados de la estimación para el modelo Sarima (4, 1, 0) $(0,1,1)_{12}$

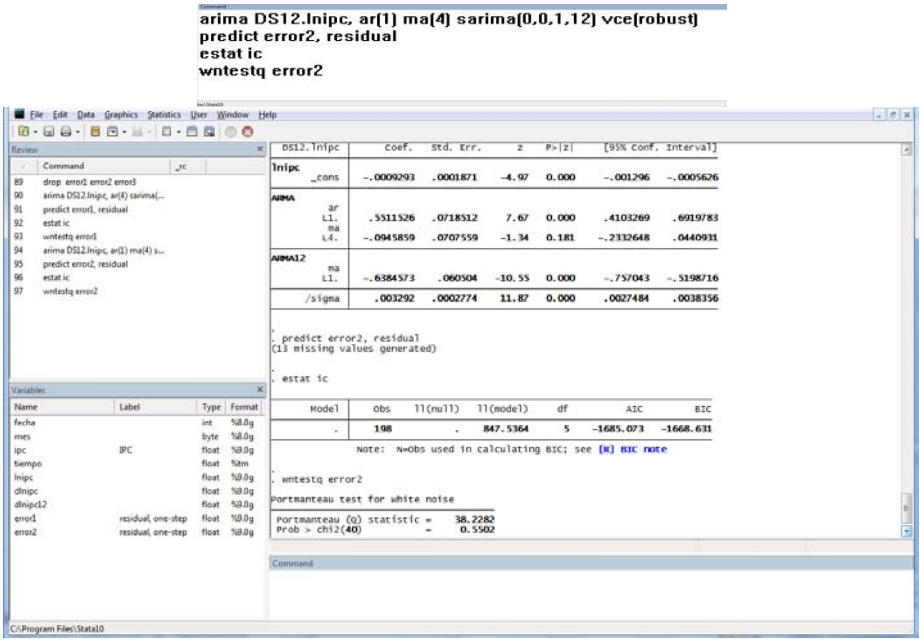


Fuente: cálculos autores.

2- Estimar los parámetros $\hat{\phi}$ y $\hat{\theta}$ para el modelo Sarima (1, 1, 4) $(0,1,1)_{12}$, con el comando *arima DS12.lnipc, ar(1) ma(4) sarima(0,0,1,12) vce(robust)*. Adicional el criterio de Akaike y prueba ruido blanco en los errores mediante las

instrucciones *estat ic*, *predict error2* y *wntestq error2*, respectivamente (*véase* figura 5.37).

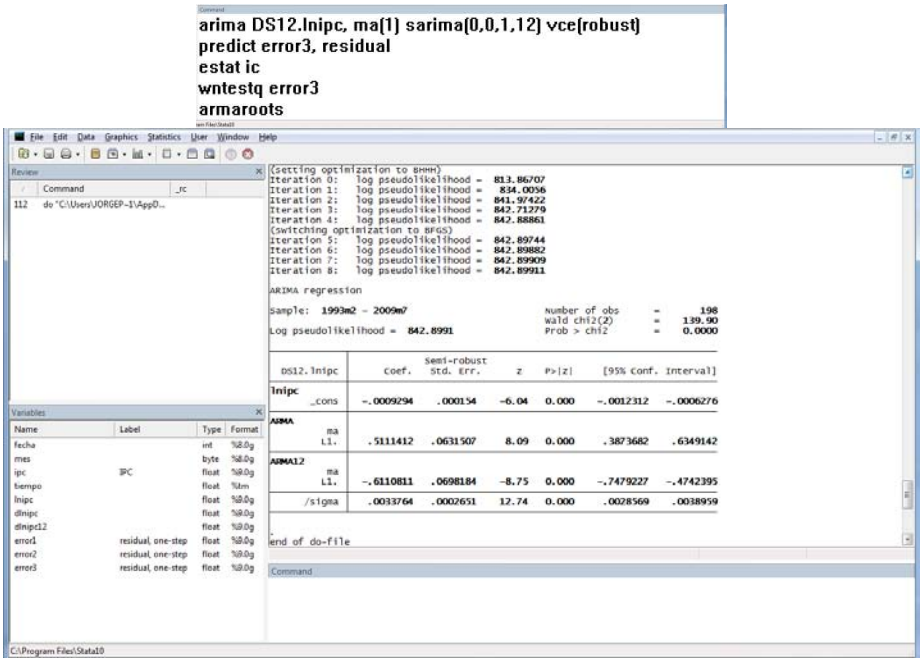
Figura 5.37. Salida de Stata® con los resultados de la estimación para el modelo Sarima (1, 1, 4) $(0,1,1)_{12}$

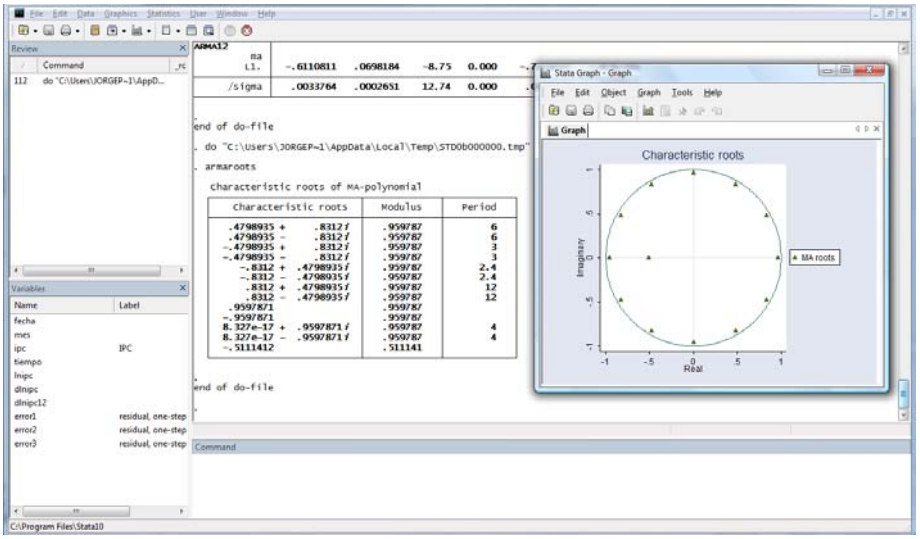


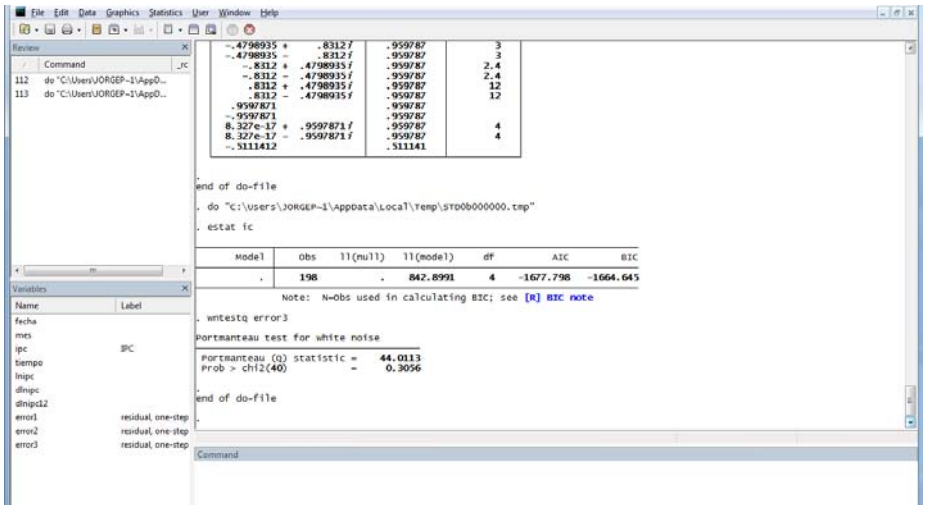
Fuente: cálculos autores.

3- Estimar los parámetros $\hat{\phi}$ y $\hat{\theta}$ para el modelo Sarima (0, 1, 1) $(0,1,1)_{12}$, con el comando *arima DS12.lnipc, ma(1) sarima(0,0,1,12) vce(robust)*. Adicional el criterio de Akaike, prueba ruido blanco en los errores, raíces de polinomio característico y circulo unitario; mediante las instrucciones *estat ic*, *predict error3*, *wntestq error3* y *armaroots*, respectivamente (véase figura 5.38 y ecuación 5.60).

Figura 5.38. Salida de Stata® con los resultados de la estimación para el modelo Sarima (0, 1, 1) (0,1,1)12







Fuente: cálculos autores.

5.8.4 Validación del modelo estimado.

$$\Delta_{12}\widehat{LNipc}_t = -0.0009294 + 0.5111\Delta lnu_{t-1} - 0.6110\Delta_{12} lnu_{t-1} \quad (5.60)$$

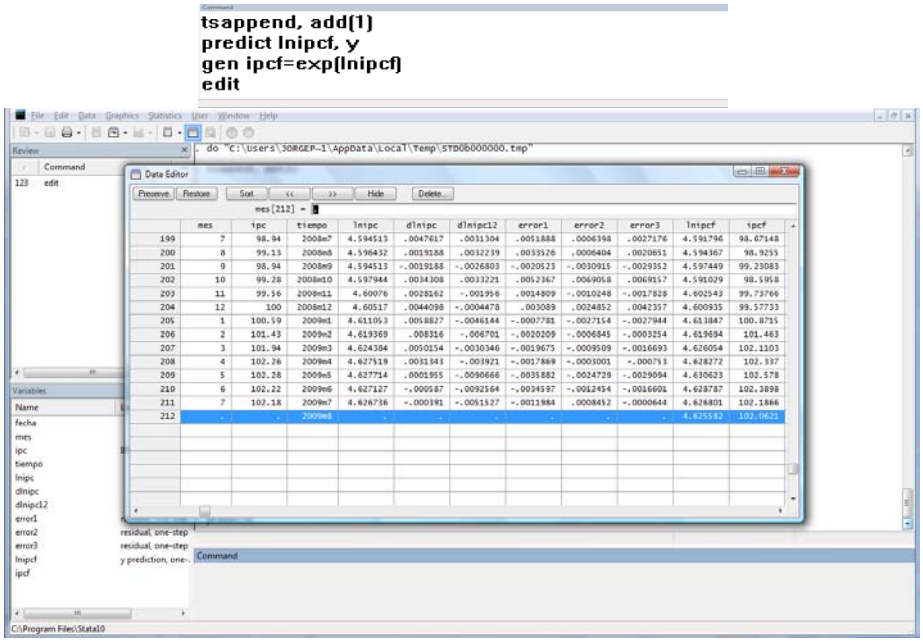
A partir de los resultados de las figuras 5.36, 5.37 y 5.38, fue seleccionado el modelo Sarima (0,1,1) $(0,1,1)_{12}$, porque los valores estimados de sus parámetros resultaron estadísticamente significativos individualmente de acuerdo con la probabilidad Z (igual a cero, $p>|Z|$). También significativos conjuntamente con la probabilidad Chi-cuadrado (igual a cero, $p>chi2$), del estadístico de Wald, criterio Akaike más pequeño que el primer modelo y residuales ruido blanco.

Adicionalmente los 13 valores de las raíces de polinomio característico para los procesos MA y SMA son menores a uno, aunque ésta estructura por naturaleza es estacionaria el resultado lo confirma; señalando también que el modelo no se encuentra sobreparametrizado (no se estiman demasiados coeficientes que sobrecarguen el modelo, así ellos resulten significativos). Adicionalmente, estos 13 valores se encuentran dentro del círculo unitario, indicando que la variable dependiente es estacionaria.

5.8.5 Pronóstico con el modelo estimado y validado.

1- Adicionar previamente los periodos a pronosticar con el comando con el comando *tsappend*, add(1). Posteriormente, proyectar la variable retornando la primera diferencia estacional a predecir en su nivel logarítmico inicial (mediante la instrucción *predict lnipcf, y*); por último sacar el anti exponencial para retornar a la variable IPC original utilizando *gen ipcf=exp(lnipcf)* (con el comando *edit* observar los nuevos valores para el IPC, *véase* figura 5.39).

Figura 5.39. Salida de Stata® para pronosticar el IPC un periodo



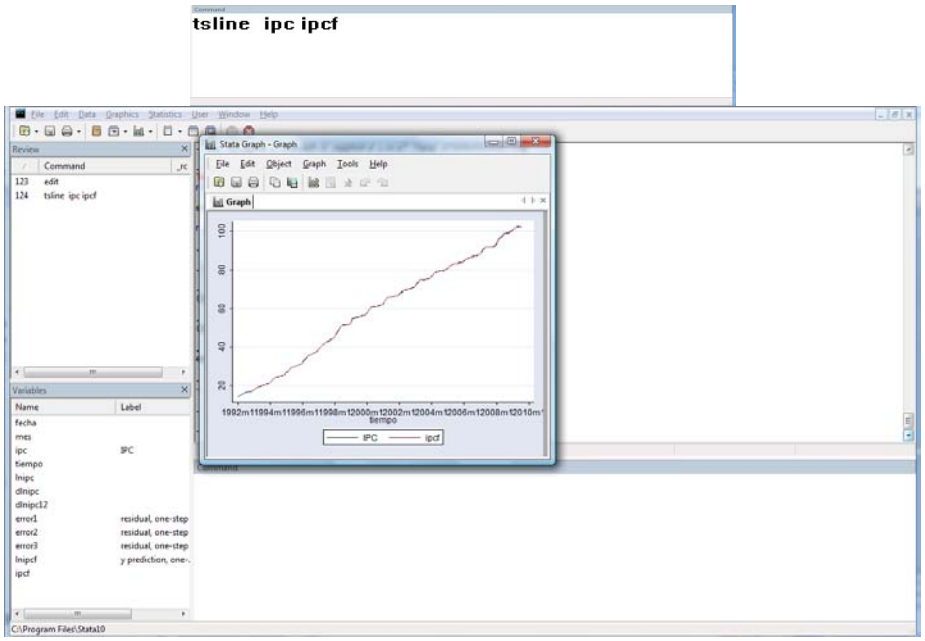
Fuente: cálculos autores.

En la figura 5.39 se puede apreciar que el nuevo valor proyectado para el IPC en agosto de 2009 (2009-VIII) corresponde a 102.0621. El mismo, se involucra como un nuevo valor observado en la serie original y se reestima el modelo Sarima (0,1,1) $(0,1,1)_{12}$ de acuerdo a los procedimiento anteriores, para predecir 2009-IX y 2009-X, siempre y cuando los resultados anteriores analizados no se alteren.

8.5.6 Validación del pronóstico.

1- Graficar (comando *tsline*) *ipc* e *ipcf*, para conocer si es similar el original al proyectado (*véase* línea roja figura 5.40).

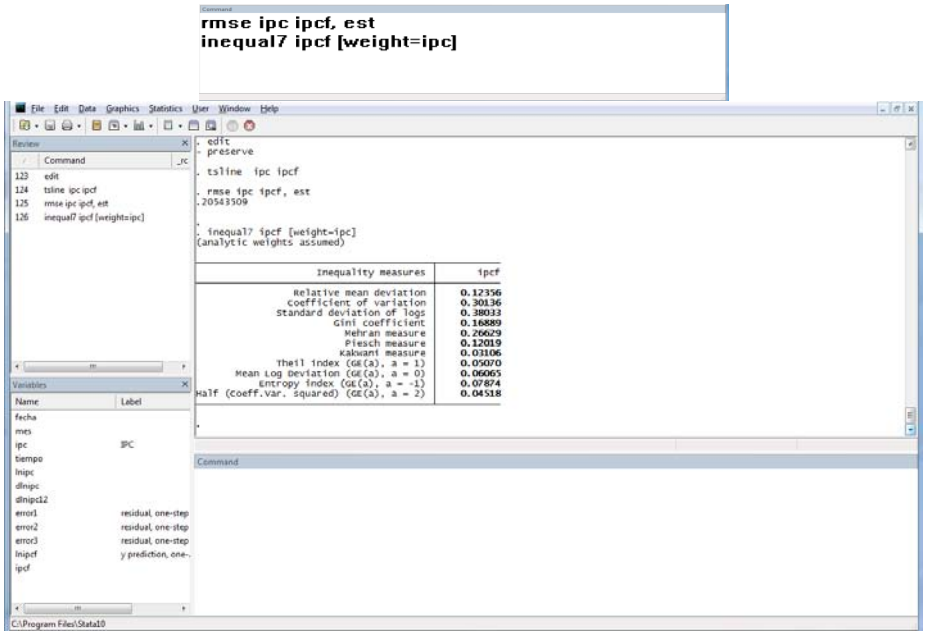
Figura 5.40. Salida de Stata® con las gráficas observada y proyectada del IPC



Fuente: cálculos autores.

2- Ejecutar la raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC) y coeficiente de Theil (CT) en Stata®, mediante los comandos *rmse ipc ipcf, est* y *inequal7 ipcf [weight=ipc]*, respectivamente (*véase* figura 5.41).

Figura 5.41. Salida de Stata® con los resultados de la raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC) y coeficiente de Theil (CT, *inequal7*)



Fuente: cálculos autores.

En la figura 5.40, se aprecia que el valor observado y proyectado del IPC llevan trayectorias similares, casi están sobrepuestas, presumiendo que su pronóstico puede estar por encima en 0.20543509 (RSPSEC, *rmse*) al real que pueda presentarse en este periodo. Por tanto, a 102.0621 se le debe restar 0.20543509 para tener una mejor aproximación de la proyección a la futura observada en 2009-VIII.

También en la figura 5.41, la predicción está bien ajustada de acuerdo con el coeficiente de Theil (0.005070) cercano a cero, ocurriría lo contario cuando este tienda a uno. Con este estudio de caso finaliza el procedimiento de la metodología Box-Jenkins para series estacionales, en el próximo capítulo se encuentra algunos aspectos sobre series de tiempo con variables endógena y exógenas dinámicas, causalidad y cointegración entre ellas.

Resumen.

- La modelación de series de tiempo a partir de procesos autorregresivos y media móvil, bajo la metodología Box-Jenkins, permite construir modelos que generan pronósticos para generar pronósticos de corto plazo con muestras representativas (grandes).
- Las series se toman como procesos aleatorios o estocásticos, caracterizadas por un mecanismo generador de datos desconocido. Siempre que estas cumplan las condiciones de estacionariedad y ergodicidad, será posible realizar una aproximación al funcionamiento interno de la serie y así obtener pronósticos a partir de sus valores históricos.
- La existencia de una tendencia, implica necesariamente un incumplimiento de la condición de estacionariedad. Por esta razón, en algunos casos surge la necesidad de transformar la serie inicial para poder aplicar correctamente la metodología de Box-Jenkins. Las series que requieren de una transformación se conocen como series integradas.
- Los procesos autorregresivos (AR) relacionan el valor contemporáneo de la serie, directamente con sus rezagos. Por otra parte, los de media móvil (MA) relacionan el valor presente de la variable con un conjunto de residuales pasados. Estos dos elementos pueden ser combinados en modelos que se conocen como ARMA, para las series estacionarias en niveles y Arima para variables integradas.
- Las funciones de autocorrelación simple y parcial indican cual es la forma más parsimoniosa de modelar una serie específica, mediante procesos autorregresivos, media móvil puro o una combinación de ambos.
- Cuando se cuenta con series que tienen un componente estacional, no se puede aplicar la metodología de Box-Jenkins convencional directamente. En este caso, se hace necesario modelar adicionalmente el componente estacional, desestacionalizando la variable –empleando diferencia estacional- y modelando la variable desestacionalizada con estructuras Sarima.
- La metodología de Box-Jenkins resume todos los pasos necesarios para obtener un pronóstico de una serie de tiempo. En primer lugar, se efectúa un análisis general sobre la serie, donde se reconoce el cumplimiento de las

condiciones de estacionariedad y ergodicidad, así como la posible existencia de un componente estacional. En segundo lugar se identifica un modelo candidato el cual se estima y verifica. Por último, generar un pronóstico que a su vez también es validado gráficamente y a través de raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC) y coeficiente de Theil (CT).

Anexo 5

En el anexo cinco se encuentran algunas demostraciones y notaciones matemáticas sobre el desarrollo, pruebas y estimación de los modelos Arima expuesto en este capítulo; bajo la metodología Box-Jenkins.

A.5.1 Caminatas aleatorias no estacionarias en varianza y covarianza

Para series que contienen raíces unitarias regulares su media es constante ($\mu = 0$), *véase* a continuación su desarrollo partiendo desde la ecuación 5.61 y terminando en la 5.62.

$$Y_t = \mu + Y_{t-1} + u_t \text{ (5.61)}$$

$$E(Y_t) = E(\mu) + E(Y_{t-1}) + E(u_t)$$

Dado que $E(Y_{t-1}) = E(Y_t) = E(Y_{t+1}) = \mu$ y $E(u_t) = 0$

Entonces

$$E(Y_t) - E(Y_{t-1}) = E(\mu) + E(u_t)$$

$$\mu - \mu = \mu$$

$$\mu = 0$$

$$Y_t = Y_{t-1} + u_t \text{ (5.62)}$$

Para series que contienen raíces unitarias regulares, su varianza se condiciona con el tiempo ($\gamma_0 = T\gamma_0$), razón por la cual no es estacionaria. *Véase* a continuación su desarrollo partiendo desde la ecuación 5.63 y terminando en la 5.64.

$$Y_t = Y_{t-1} + u_t \text{ (5.63)}$$

$$\sigma_{Y_t}^2 = \gamma_0 = E[(Y_t - \mu)^2] = E[(Y_{t-1} + u_t - \mu)^2]$$

$$\gamma_0 = E[(Y_{t-1} - \mu)^2] + E(\varepsilon_t^2) + 2E(Y_{t-1} - \mu)E(u_t)$$

$$\gamma_0 = \gamma_{0_{t-1}} + \sigma_{u_t}^2$$

$$Y_{t-1} = Y_{t-2} + u_{t-1}$$

$$\gamma_{0_{t-1}} = E[(Y_{t-1} - \mu)^2] = E[(Y_{t-2} + u_{t-1} - \mu)^2]$$

$$\gamma_{0_{t-1}} = E[(Y_{t-2} - \mu)^2] + E(u_{t-1}^2) + 2E(Y_{t-2} - \mu)E(u_{t-1})$$

$$\gamma_{0_{t-1}} = \gamma_{0_{t-2}} + \sigma_{u_t}^2$$

reemplazando $\gamma_{0_{t-1}}$ en γ_0

ahora $\gamma_0 = \gamma_{0_{t-1}} + \sigma_{u_t}^2$ se convierte en $\gamma_0 = \gamma_{0_{t-2}} + \sigma_{u_t}^2 + \sigma_{u_t}^2$

$$\gamma_0 = \gamma_{0_{t-2}} + 2\sigma_{u_t}^2$$

así sucesivamente se tiene que:

$$\gamma_0 = \gamma_{0_{t-T}} + T\sigma_{u_t}^2 \text{ (5.64)}$$

Para series que contienen raíces unitarias regulares, su covarianza se condiciona con el tiempo $(\gamma_p = \gamma_{0_{t-1}})$, razón por la cual no es estacionaria. *Véase* a continuación su desarrollo partiendo desde la ecuación 5.65 y terminando en la 5.66.

$$Y_t = Y_{t-1} + u_t \text{ (5.65)}$$

$$\gamma_p = cov(Y_t, Y_{t-1}) = E[(Y_t - \mu)(Y_{t-1} - \mu)]$$

$$\gamma_p = E[(Y_{t-1} + u_t - \mu)(Y_{t-1} - \mu)]$$

$$\gamma_p = E[(Y_{t-1} - \mu)^2 + u_t(Y_{t-1} - \mu)]$$

$$\gamma_p = \gamma_{0_{t-1}} \text{ (5.66)}$$

como se conoce que $\gamma_{0_{t-1}} = \gamma_{0_{t-2}} + \sigma_{u_t}^2$ y $\gamma_0 = \gamma_{0_{t-2}} + 2\sigma_{u_t}^2$ entonces γ_p queda condicionado con $\gamma_{0_{t-1}}$ y este último se condiciona con T

A.5.2 Operador y polinomio empleado en series de tiempo

El operador de rezago, L[122] (*Lag*, siglas en inglés) es definido como un operador lineal para obtener equivalentemente un único valor de Y_t (*véase* ecuación 5.70). Donde L^pse antepone a Y_t simplificando su rezago para p periodos (Enders, 1995, 45). De esta forma, la relación en variables con rezagos de orden uno, dos, tres o p; se pueden apreciar en las ecuaciones 5.67, 5.68 y 5.69 respectivamente.

$$LY_t = Y_{t-1} \qquad (5.67)$$

$$L^2Y_t = L(LY_t) = Y_{t-2} \qquad (5.68)$$

$$L^3Y_t = L(L^2Y_t) = Y_{t-3} \qquad (5.69)$$

.

.

$$L^pY_t \equiv Y_{t-p} \Rightarrow L^pY_t = L(L^{p-1}Y_t) = Y_{t-p} \quad (5.70)$$

Usando el operador de rezago, L, la ecuación $Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + u_t$ se puede reescribir como se expresa en la ecuación 5.71.

$$\left(1 - \phi_1 L - \phi_2 L^2 - .. - \phi_p L^p\right) Y_t = \delta + u_t \qquad (5.71)$$

$$A(L)Y_t = \delta + u_t \qquad (5.72)$$

A partir de la ecuación 5.71, reescribiéndola en la ecuación 5.72, se puede definir y deducir el polinomio $A(L)$, del operador de rezago (L). Donde $A(L)$ es el polinomio de $\left(1 - \phi_1 L - \phi_2 L^2 - .. - \phi_p L^p\right)$, $\phi_1; \phi_2 .. \phi_p$ los parámetros[123] a estimar del modelo, δ

[122]Significa rezago en castellano.

[123]Constantes que denotan el peso o ponderan la importancia de los rezagos a los que se encuentran asociados (Guerrero, 2003, 11).

la media y u_t el término aleatorio. En conclusión, un polinomio de rezago se puede definir como la expresión general de los operadores de rezago; estos polinomios son empleados tradicionalmente, en los análisis de series de tiempo, porque permiten expresar de manera concisa y simple modelos que representan fenómenos reales.

A.5.2.1 Propiedades del operador de rezago

Las principales propiedades del operador rezago son las siguientes:

1- $L^0 = 1$ equivalente a $L^0 Y_t = Y_t$, estrictamente $L^0 = \mathbf{I}$; donde **I** se refiere a una matriz identidad.
2- El rezago de una constante (α) es una constante (α): $L\alpha = \alpha$.
3- Se cumple la ley distributiva en operadores de rezago así:

$$\left(L^p + L^i\right)Y_t = L^p Y_t + L^i Y_t = Y_{t-p} + Y_{t-i}.$$

4- Se cumple la ley asociativa en operadores de rezago así:

$$L^p L^i Y_t = L^{p+i} Y_t = Y_{t-p-i}.$$

5- El crecimiento o valores negativos de L, son los operadores hacia adelante: $L^{-i} Y_t = Y_{t+i}$; definiendo p=-i de tal forma que $L^{-i} Y_t = Y_{t-p} = Y_{t+i}$.
6- Para $|\alpha| < 1$, la suma infinita $(1 + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \cdots)Y_t = \frac{Y_t}{(1-\alpha L)}$, aparentemente esta propiedad no puede observarse intuitivamente; pero retomando las propiedades 3 y 4, y multiplicando cada lado por $(1 - \alpha L)$ se tiene lo siguiente:

$$(1 - \alpha L)(1 + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \cdots)Y_t = \frac{Y_t}{\cancel{(1-\alpha L)}}\cancel{(1 - \alpha L)}$$

$$(1 - \alpha L)(1 + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \cdots)Y_t = Y_t$$

$$(1 - \alpha L + \alpha L - \alpha^2 L^2 + \alpha^2 L^2 - \alpha^3 L^3 + \alpha^3 L^3 + \cdots)Y_t = Y_t$$

dado que $|\alpha| < 1$, la expresión $\alpha^n L^n Y_t$ converge a cero cuando $n \to \infty$; por ende, $(1-0)Y_t = Y_t \Rightarrow Y_t = Y_t$, ambos lados de la ecuación son iguales, respectivamente también $(1 + \alpha L + \alpha^2 L^2 + \alpha^3 L^3 + \cdots)Y_t = \frac{Y_t}{(1-\alpha L)}$

7- Para $|\alpha| > 1$, la suma infinita $[1 + (\alpha L)^{-1} + (\alpha L)^{-2} + (\alpha L)^{-3} + \cdots]Y_t = \frac{-\alpha L Y_t}{(1-\alpha L)}$

equivalentemente $\frac{Y_t}{(1-\alpha L)} = -(\alpha L)^{-1} \sum_{p=0}^{\infty} (\alpha L)^{-p} Y_t$; multiplicando cada lado por $(1 - \alpha L)$ se tiene lo siguiente:

$$(1 - \alpha L)[1 + (\alpha L)^{-1} + (\alpha L)^{-2} + (\alpha L)^{-3} + \cdots]Y_t = -\alpha L Y_t$$

$$[1 - \alpha L + (\alpha L)^{-1} - 1 + (\alpha L)^{-2} - (\alpha L)^{-1} + (\alpha L)^{-3} - (\alpha L)^{-2} + \cdots]Y_t = -\alpha L Y_t$$

dado que $|\alpha| > 1$, la expresión $\alpha^{-n} L^{-n} Y_t$ converge a cero cuando $n \to \infty$; por ende, $-\alpha L Y_t = -\alpha L Y_t$, ambos lados de la ecuación son iguales, respectivamente también $[1 + (\alpha L)^{-1} + (\alpha L)^{-2} + (\alpha L)^{-3} + \cdots]Y_t = \frac{-\alpha L Y_t}{(1-\alpha L)}$

A.5.3 Ecuaciones en diferencia para series de tiempo

El entendimiento y desarrollo de ecuaciones en diferencia es importante, en la técnica Box-Jenkins, por emplearse como instrumento para construir modelos de series de tiempo, entendimiento de la teoría económica (implícita en este tipo de modelos) e indicar el proceso generador de datos (**PGD**) para la serie temporal. Adicionalmente, las ecuaciones en diferencia describen las consecuencias dinámicas de los eventos propios de una variable a lo largo del tiempo (Hamilton, 1994, 1).

Las ecuaciones en diferencia son un caso especial de ecuaciones diferenciales $\left(\frac{\partial Y_t}{\partial t} = \frac{\Delta Y_t}{\Delta t}\right)$, también se caracterizan por ser lineales, no lineales, homogéneas, no homogéneas, de orden uno y superior. 1) lineal, cuando el termino Y_t solo esta elevado a la primera potencia, 2) no homogénea, si en el término al otro lado de la igualdad es diferente de cero y 3) de primer orden cuando se refiere a una primera diferencia[124].

[124]Ver más detalles en Chiang (1987, capítulos 16 y 17).

Las ecuaciones en diferencia se originan cuando la variable de tiempo (t), tiene naturaleza discreta. Debido a que t solo toma valores enteros, ante esta circunstancia siempre $\Delta t = 1$: razón que conduce a lo siguiente $\frac{\partial Y_t}{\partial t} = \frac{\Delta Y_t}{\Delta t} = \frac{\Delta Y_t}{1} = \Delta Y_t$; indicando la primera diferencia para Y_t y tomado distintos valores, dependiendo que periodos consecutivos se involucren en la diferenciación.

Δ, se define como el operador de diferencia, vinculado estrechamente con el de rezago (L). Se emplea para describir una ecuación como $Z_t = Y_t - Y_{t-1}$; donde Z_t se denomina variable de resultado y Y_t de flujo. La ecuación 5.74 muestra las equivalencias utilizando Δ y su vínculo con L (operador rezago), se encuentra expresado en las ecuaciones 5.73 y 5.74 respectivamente.

$$\Delta Y_t = Y_t - Y_{t-1} \Rightarrow \Delta Y_t = Z_t \qquad (5.73)$$

$$\Delta = 1 - L \Rightarrow \Delta Y_t = (1 - L)Y_t \qquad (5.74)$$

$$\Delta^d Y_t = (1 - L)^d Y_t \qquad (5.75)$$

En le ecuación 5.75, d hace referencia al orden de la diferencia para Y_t, es así como en las ecuaciones 5.73 y 5.74 Δ esta elevado a la potencia uno; denotando una ecuación de primera diferencia. Mientras la ecuación 5.75, indica una ecuación diferenciada d veces. De esta misma forma, se pueden obtener ecuaciones en primeras $(\Delta Y_t, \Delta Y_{t+1}, .., \Delta Y_{t-2})$, segundas $(\Delta^2 Y_t, \Delta^2 Y_{t+1})$ y tercera $(\Delta^3 Y_t)$ diferencias (*véase* ecuación 5.76-5.82).

$$\Delta Y_{t+1} = Y_{t+1} - Y_t \qquad (5.76)$$

$$\Delta Y_{t+2} = Y_{t+2} - Y_{t+1} \qquad (5.77)$$

$$\Delta Y_{t-1} = Y_{t-1} - Y_{t-2} \qquad (5.78)$$

$$\Delta Y_{t-2} = Y_{t-2} - Y_{t-3} \qquad (5.79)$$

$$\Delta^2 Y_t \equiv \Delta(\Delta Y_t) = \Delta(Y_t - Y_{t-1}) = \Delta Y_t - \Delta Y_{t-1} = (Y_t - Y_{t-1}) - (Y_{t-1} - Y_{t-2}) = Y_t - 2Y_{t-1} + Y_{t-2} \quad (5.80)$$

$$\Delta^2 Y_{t+1} \equiv \Delta(\Delta Y_{t+1}) = \Delta(Y_{t+1} - Y_t) = \Delta Y_{t+1} - \Delta Y_t = (Y_{t+1} - Y_t) - (Y_t - Y_{t-1}) = Y_{t+1} - 2Y_t + Y_{t-1} \quad (5.81)$$

$$\Delta^3 Y_t \equiv \Delta(\Delta^2 Y_t) = \Delta(Y_t - 2Y_{t-1} + Y_{t-2}) = \Delta Y_t - 2\Delta Y_{t-1} + \Delta Y_{t-2} = (Y_t - Y_{t-1}) - 2(Y_{t-1} - Y_{t-2}) + (Y_{t-2} - Y_{t-3}) = Y_t - 3Y_{t-1} + 3Y_{t-2} - Y_{t-3} \quad (5.82)$$

Ecuaciones en diferencia, igualmente las diferenciales, tienen como objetivo principal analizar la dinámica de una serie temporal; cuya solución busca hallar la trayectoria temporal de Y_t.

A.5.4 Modelos AR, MA y ARMA para series estacionarias en niveles

Los procesos autorregresivos (AR) son los rezagos de Y_t que explican su comportamiento (*véase* ecuación 5.83). igualmente, de orden uno o p se denotan como AR(p); así cuando son de primer orden AR(1) es representado por la ecuación 5.84, donde $|\phi| < 1$ asegurando estacionariedad para Y_t. En este sentido se puede obtener la media (μ), varianza (γ_0), covarianza (γ_p) y función de autocorrelación simple ($\hat{\rho}_p$); como se observa en las ecuaciones desde 5.84 hasta 5.88, dado que $E(Y_t) = E(Y_{t-1}) = \mu$.

$$Y_t = f\left(Y_{t-1}, Y_{t-2}, .., Y_{t-p}\right) \quad (5.83)$$

$$Y_t = \delta + \phi Y_{t-1} + u_t \Rightarrow (1 - \phi L)Y_t = \delta + u_t \Rightarrow \phi(L)Y_t = \delta + u_t \quad (5.84)$$

$$E(Y_t) = E(\delta) + \phi E(Y_{t-1}) + E(u_t)$$

$$\mu = \delta + \phi\mu + 0 \Rightarrow \mu - \phi\mu = \delta \Rightarrow (1 - \phi)\mu = \delta \Rightarrow \mu = \frac{\delta}{(1-\phi)} \quad (5.85)$$

$$\sigma_{Y_t}^2 = \gamma_0 = E[(Y_t - \mu)^2] = E[(\delta + \phi Y_{t-1} + u_t - \mu)^2]\ ,\ \delta = 0 \text{ por ende } \mu = 0$$

$$\gamma_0 = E[(\phi Y_{t-1})^2] + E(u_t^2) + 2E(\phi Y_{t-1})E(u_t)$$

$$\gamma_0 = E(\phi^2 Y_{t-1}^2) + E(u_t^2) + 2E(\phi Y_{t-1})E(u_t)$$

$$\gamma_0 = \phi^2 \gamma_0 + \sigma_{u_t}^2$$

$$\gamma_0 - \phi^2 \gamma_0 = \sigma_{u_t}^2$$

$$(1 - \phi^2)\gamma_0 = \sigma_{u_t}^2$$

$$\gamma_0 = \frac{\sigma_{u_t}^2}{(1-\phi^2)} \quad (5.86)$$

$$\gamma_1 = cov(Y_t, Y_{t-1}) = E[(\phi Y_{t-1} + u_t)(Y_{t-1})]$$

$$\gamma_1 = [E(\phi)E(Y_{t-1}) + E(u_t)]E(Y_{t-1})$$

$$\gamma_1 = \phi\gamma_0 \Rightarrow \gamma_1 = \frac{\phi\sigma_{u_t}^2}{(1-\phi^2)}$$

$$\gamma_2 = cov(Y_t, Y_{t-2}) = E[(\phi Y_{t-1} + u_t)(Y_{t-2})], \text{ como } Y_{t-p} = \delta + \phi Y_{t-p-1} + u_{t-p}$$

$$\gamma_2 = E[\phi(\phi Y_{t-2} + u_{t-1}) + u_t\ ]E(Y_{t-2})$$

$$\gamma_2 = E[\phi^2 Y_{t-2} + \phi u_{t-1} + u_t\ ]E(Y_{t-2})$$

$$\gamma_2 = \phi^2\gamma_0 \Rightarrow \gamma_2 = \frac{\phi^2\sigma_{u_t}^2}{(1-\phi^2)}$$

$$\gamma_p = \phi^p\gamma_0 \Rightarrow \gamma_p = \frac{\phi^p\sigma_{u_t}^2}{(1-\phi^2)} \quad (5.87)$$

$$\hat{\rho}_p = \frac{\gamma_p}{\gamma_0} = \phi^p \quad (5.88)$$

En la ecuación 5.88 $\hat{\rho}_p$ se refiere a la función de autocorrelación simple, iniciando $\hat{\rho}_0 = 1$ y así sucesivamente disminuye geométricamente. Adicionalmente ϕ^p señala

que el proceso tiene memoria infinita, así su valor actual Y_t depende de sus valores pasados (Y_{t-1},..,Y_{t-p}). Por otra parte, es posible obtener la condición de estacionariedad del AR (1) de esta misma manera a partir de polinomio de rezagos, dado por $\phi(L) = (1 - \phi L)$ con una única raíz $L = \frac{1}{\phi}$. Para que esta raíz sea -en valor absoluto- mayor a uno, se debe cumplir la condición $|\phi| < 1$.

Asimismo, un proceso autorregresivo de orden p AR(p), describe el valor actual de Y_t mediante promedio ponderado de las observaciones pasadas mas el termino aleatorio o error (ruido blanco, u_t). La ecuación 5.88 expresa un AR(p), donde $\sum_{t=1}^{p} \phi_p < 1$, garantizando de esta forma la estacionariedad en Y_t y partir de ella se deriva la forma de su media (μ), varianza (γ_0), covarianza (γ_p) y función de autocorrelación simple ($\hat{\rho}_p$); como se observa en las ecuaciones desde 5.89 hasta 5.92.

$$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + u_t \; ; \; \delta = \left(1 - \phi_1 - \phi_2 - .. - \phi_p\right)\mu$$

$$\Rightarrow \left(1 - \phi_1 L - \phi_2 L^2 - .. - \phi_p L^p\right) Y_t = \delta + u_t \Rightarrow \phi(L) Y_t = \delta + u_t \quad (5.89)$$

$$\mu = \frac{\delta}{(1 - \phi_1 - \phi_2 - .. - \phi_p)} \quad (5.90)$$

$$\sigma_{Y_t}^2 = \gamma_0 = E[(Y_t - \mu)^2], \; \delta = 0 \text{ por ende } \mu = 0$$

$$E(Y_t)^2 = E\left[Y_t\left(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + u_t\right)\right]$$

$$\gamma_0 = E(Y_t)^2 = \phi_1 \gamma_1 + \phi_2 \gamma_2 + .. + \phi_p \gamma_p + \sigma_{u_t}^2 \quad (5.91)$$

$$\gamma_1 = cov(Y_t, Y_{t-1}) = E\left[\left(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + u_t\right) Y_{t-1}\right]$$

$$\gamma_1 = \phi_1 \gamma_0 + \phi_2 \gamma_1 + .. + \phi_p \gamma_{p-1} \Rightarrow \gamma_1 = \frac{\phi_1 \gamma_0 + .. + \phi_p \gamma_{p-1}}{1 - \phi_2}$$

$$\gamma_2 = cov(Y_t, Y_{t-2}) = E\left[\left(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + u_t\right) Y_{t-2}\right]$$

$$\gamma_2 = \phi_1 \gamma_1 + \phi_2 \gamma_0 + .. + \phi_p \gamma_{p-1}$$

$$\gamma_p = cov\left(Y_t, Y_{t-p}\right) = E\left[\left(\phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + u_t\right) Y_{t-p}\right]$$

$$\gamma_p = \phi_1\gamma_{p-1} + \phi_2\gamma_{p-2} + .. + \phi_p\gamma_0 \quad (5.92)$$

A partir de lo anterior, en el cuadro 5.5 se presenta la forma de la media (μ), varianza (γ_0), covarianzas ($\gamma_1, \gamma_2, .., \gamma_p$) y función de autocorrelación simple ($FAS \rightarrow \hat{\rho}_p = \frac{\gamma_p}{\gamma_0}$) para los procesos AR (1), AR (2) y AR (p). Deducidos, de la misma forma como lo expresado desde la ecuación 5.84 hasta 5.92.

Cuadro 5.5. Formas de la media, varianza, covarianzas y FAP en procesos AR.

Orden de integración para la serie Y_t	0	0	0
proceso	**AR(1); $\|\boldsymbol{\phi}\| < 1$**	**AR(2); $\sum_{t=1}^{2}\|\boldsymbol{\phi_p}\| < 1$**	**AR(p); $\sum_{t=1}^{p}\|\boldsymbol{\phi_p}\| < 1$**
Forma del modelo	$Y_t = \delta + \phi_1 Y_{t-1} + u_t$	$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + u_t$	$Y_t = \delta + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + .. + \phi_p Y_{t-p} + u_t$
Media $\boldsymbol{\mu}$	$\frac{\delta}{(1-\phi_1)}$	$\frac{\delta}{(1-\phi_1-\phi_2)}$	$\frac{\delta}{(1-\phi_1-\phi_2-..-\phi_p)}$
Varianza $\boldsymbol{\sigma^2_{Y_t} = \gamma_0}$	$\frac{\sigma^2_{u_t}}{(1-\phi^2)}$	$\phi_1\gamma_1 + \phi_2\gamma_2 + \sigma^2_{\varepsilon_t}$ $\gamma_0 = \frac{(1-\phi_2)\sigma^2_{u_t}}{(1+\phi_2)\left[(1-\phi_2)^2 - \phi_1^2\right]}$	$\phi_1\gamma_1 + \phi_2\gamma_2 + .. + \phi_p\gamma_p + \sigma^2_{u_t}$
Covarianza $\boldsymbol{\gamma_1}$	$\frac{\phi\sigma^2_{u_t}}{(1-\phi^2)}$	$\phi_1\gamma_0 + \phi_2\gamma_1$ $\gamma_1 = \frac{\phi_1\gamma_0}{(1-\phi_2)}$	$\phi_1\gamma_0 + \phi_2\gamma_1 + .. + \phi_p\gamma_{p-1}$ $\gamma_1 = \frac{\phi_1\gamma_0 + .. + \phi_p\gamma_{p-1}}{1-\phi_2}$
Covarianza $\boldsymbol{\gamma_2}$	$\frac{\phi^2\sigma^2_{u_t}}{(1-\phi^2)}$	$\phi_1\gamma_1 + \phi_2\gamma_0$	$\phi_1\gamma_1 + \phi_2\gamma_0 + .. + \phi_p\gamma_{p-1}$
Covarianza $\boldsymbol{\gamma_p}$	$\frac{\phi^p\sigma^2_{u_t}}{(1-\phi^2)}$	$\phi_1\gamma_{p-1} + \phi_2\gamma_{p-2}$	$\gamma_p = \phi_1\gamma_{p-1} + \phi_2\gamma_{p-2} + .. + \phi_p\gamma_0$ $\gamma_k = \phi_1\gamma_{k-1} + \phi_2\gamma_{k-2} + .. + \phi_p\gamma_{k-p}$
FAS=$\widehat{\rho}_p = \frac{\gamma_p}{\gamma_0}$	ϕ^p	$\rho_1 = \frac{\phi_1}{1-\phi_2}$ $\rho_2 = \phi_2 + \frac{\phi_1^2}{1-\phi_2}$ $\rho_p = \phi_1\rho_{p-1} + \phi_2\rho_{p-2}$	$\rho_1 = \phi_1 + \phi_2\rho_1 + .. + \phi_p\rho_{p-1}$ $\rho_p = \phi_1\rho_{p-1} + \phi_2\rho_{p-2} + .. + \phi_p$ $\rho_k = \phi_1\rho_{k-1} + \phi_2\rho_{k-2} + .. + \phi\rho_{k-p}$

Fuente: autores, a partir de Pindyck y Rubinfeld (1998).

Por otra parte, los procesos de media móvil (MA) son los rezagos de u_t que explican el comportamiento de Y_t (*véase* ecuación 5.93). Cuando resulta de orden uno o q se denota como MA (q), así de primer orden MA (1) representado por la ecuación 5.94; una caracteristica prevaleciente en los procesos MA es estacionariedad por naturaleza, ésta condición hace posible que todo proceso AR se represente de manera equivalente mediante un MA; en otras palabras, garantiza la invertibilidad de AR.

A partir de lo anterior, se puede expresar la idea de dualidad entre estacionariedad para un proceso AR e invertibilidad para un proceso MA; en este sentido, todo proceso MA es estacionario, mientras todo proceso AR es invertible (Guerrero, 2003, 83). Adcionalemente para procesos Ma (1) y MA (q) el resultado de su media (μ), varianza (γ_0), covarianza (γ_p) y FAS ($\hat{\rho}_p$), es derivado de la siguiente manera (*véase* ecuaciones de 5.94 hasta 5.99):

$$Y_t = f\left(u_{t-1}, u_{t-2}, .., u_{t-q}\right) \quad (5.93)$$

$$Y_t = \delta - \theta u_{t-1} + u_t \Rightarrow Y_t - \delta = (1 - \theta L)u_t \Rightarrow Y_t - \delta = \theta(L)u_t \quad (5.94)$$

$$E(Y_t) = E(\delta) + \theta E(u_{t-1}) + E(u_t)$$

$$E(Y_t) = E(\delta) = \mu \quad (5.95)$$

$$\sigma^2_{Y_t} = \gamma_0 = E[(Y_t - \mu)^2] = E[(\delta - \theta u_{t-1} + u_t - \mu)^2]\,, \mu = \delta = 0$$

$$\gamma_0 = E[(\theta u_{t-1})^2] + E(u_t^2) + 2E(\theta u_{t-1})E(u_t)$$

$$\gamma_0 = E(\theta^2 u_{t-1}^2) + E(u_t^2) + 2E(\theta u_{t-1})E(u_t)$$

$$\gamma_0 = \theta^2 \sigma^2_{u_t} + \sigma^2_{u_t}$$

$$\gamma_0 = (1 + \theta^2)\sigma^2_{u_t} \quad (5.96)$$

$$\gamma_1 = cov(Y_t, Y_{t-1}) = E[(u_t - \theta u_{t-1})(u_{t-1} - \theta u_{t-2})],\ \mu = \delta = 0$$

$$\gamma_1 = -\theta \sigma^2_{u_t} \quad (5.97)$$

$$\gamma_p = cov(Y_t, Y_{t-p}) = E\left[(u_t - \theta u_{t-1})(u_{t-p} - \theta u_{t-p-1})\right]$$

$\gamma_p = 0$ (5.98), para todo p>1

$$\hat{\rho}_1 = \frac{\gamma_1}{\gamma_0} = \frac{-\theta}{1+\theta^2} \text{ y } \hat{\rho}_2 = \frac{\gamma_2}{\gamma_0} = 0, \ldots, \hat{\rho}_p = \frac{\gamma_p}{\gamma_0} = 0 \quad (5.99)$$

Asimismo, un proceso de media móvil de orden q MA(q), describe el valor actual de Y_t mediante promedio ponderado de las observaciones pasadas de u_t. La ecuación 5.100 expresa un MA(q), donde $\sum_{t=1}^{q} \theta_q < \infty$, garantizando la condición de dualidad y a partir de ella se deriva la forma de su media (μ), varianza (γ_0), covarianza (γ_p) y función de autocorrelación simple ($\hat{\rho}_p$); como se observa en las ecuaciones desde 5.101 hasta 5.105.

$$Y_t = \delta - \theta_1 u_{t-1} - \theta_2 u_{t-2} - .. \theta_q u_{t-q} + u_t \Rightarrow Y_t - \delta = \left(1 - \theta_1 L - .. - \theta_q L^q\right) u_t \Rightarrow Y_t - \delta = \theta(L) u_t \ (5.100)$$

$$E(Y_t) = \mu \ (5.101)$$

$$\sigma_{Y_t}^2 = \gamma_0 = E[(Y_t - \mu)^2] = E\left[(\theta_1 u_{t-1} - \theta_2 u_{t-2} - .. \theta_q u_{t-q} + u_t)^2\right], \mu = \delta = 0$$

$$\gamma_0 = E\left[(\theta_1 u_{t-1} - \theta_2 u_{t-2} - .. \theta_q u_{t-q} + u_t)(\theta_1 u_{t-1} - \theta_2 u_{t-2} - .. \theta_q u_{t-q} + u_t)\right]$$

$$\gamma_0 = \sigma_{u_t}^2 \left(1 + \theta_1^2 + .. + \theta_q^2\right) \ (5.102)$$

$$\gamma_1 = cov(Y_t, Y_{t-1})$$
$$= E\big[(\theta_1 u_{t-1} - \theta_2 u_{t-2} - .. \theta_q u_{t-q} + u_t)(\theta_1 u_{t-2} - \theta_2 u_{t-3} - .. \theta_q u_{t-q-1} + u_{t-1})\big]$$

$$\gamma_1 = \sigma_{u_t}^2\big(\theta_1 + \theta_2\theta_1 + .. + \theta_q\theta_{q-1}\big) \quad (5.103)$$

$$\gamma_p = cov\big(Y_t, Y_{t-p}\big)$$
$$= E\big[(\theta_1 u_{t-1} - \theta_2 u_{t-2} - .. \theta_q u_{t-q} + u_t)(\theta_1 u_{t-p-1} - \theta_2 u_{t-p-2} - .. \theta_q u_{t-p-q} + u_{t-p-q-1})\big]$$

$\gamma_p = 0$ (5.104), para todo p>1

$$\hat{\rho}_1 = \frac{\gamma_1}{\gamma_0} = \frac{\theta_1 + \theta_2\theta_1 + .. + \theta_q\theta_{q-1}}{1 + \theta_1^2 + .. + \theta_q^2} \text{ y } \hat{\rho}_2 = \frac{\gamma_2}{\gamma_0} = 0, \ldots, \hat{\rho}_p = \frac{\gamma_p}{\gamma_0} = 0 \quad (5.105)$$

En las ecuaciones 5.99 y 5.105 $\hat{\rho}_p$ se refiere a la función de autocorrelación simple, iniciando en un valor para $\hat{\rho}_1$ y luego desvanece totalmente porque para todo *p>1* el valor de rho es cero. Esto indica que los procesos MA tienen memoria finita o de un solo periodo, así su valor actual de Y_t depende únicamente de su pasado inmediatamente anterior Y_{t-1}. Sucesivamente, quiere decir que el proceso olvida más de un periodo hacia atrás.

A partir de lo anterior, en el cuadro 5.6 se presenta la forma de la media (μ), varianza (γ_0), covarianzas ($\gamma_1, \gamma_2, .., \gamma_p$) y función de autocorrelación simple ($FAS \rightarrow \hat{\rho}_p = \frac{\gamma_p}{\gamma_0}$) para los procesos MA (1), MA (2) y MA (q). Deducidos, de la misma forma como lo expresado desde la ecuación 5.94 hasta 5.105.

Cuadro 5.6. Formas de la media, varianza, covarianzas y FAP en procesos MA.

Orden de integración para la serie Y_t	0	0	0
proceso	**MA(1)**	**MA(2); $\sum_{q=1}^{2}\boldsymbol{\theta_q} < \infty$**	**MA(q); $\sum_{t=1}^{q}\boldsymbol{\theta_q} < \infty$**
Forma del modelo	$Y_t = \delta - \theta_1 u_{t-1} + u_t$	$Y_t = \delta - \theta_1 u_{t-1} - \theta_2 u_{t-2} + u_t$	$Y_t = \delta - \theta_1 u_{t-1} - \theta_2 u_{t-2} - .. \theta_q u_{t-q} + u_t$
Media $\boldsymbol{\mu}$	δ	δ	δ
Varianza $\boldsymbol{\sigma^2_{Y_t} = \gamma_0}$	$\sigma^2_{u_t}(1+\theta^2)$	$\sigma^2_{u_t}(1+\theta_1^2+\theta_2^2)$	$\sigma^2_{u_t}\left(1+\theta_1^2+\theta_2^2+..+\theta_q^2\right)$
Covarianza $\boldsymbol{\gamma_1}$	$-\theta_1\sigma^2_{u_t}$	$-\theta_1\sigma^2_{u_t} + \theta_2\theta_1\sigma^2_{u_t} = -\theta_1(1-\theta_2)\sigma^2_{u_t}$	$\sigma^2_{u_t}\left(-\theta_1 + \theta_1\theta_2 + \theta_2\theta_3 + .. + \theta_q\theta_{q+1}\right)$
Covarianza $\boldsymbol{\gamma_2}$	0	$-\theta_2\sigma^2_{u_t}$	$\sigma^2_{u_t}\left(-\theta_2 + \theta_1\theta_3 + .. + \theta_q\theta_{q+2}\right)$
Covarianza $\boldsymbol{\gamma_p}$	0	0	0
FAS=$\boldsymbol{\hat{\rho}_p = \frac{\gamma_p}{\gamma_0}}$	$\rho_1 = \frac{-\theta_1}{(1+\theta^2)}$ $\rho_p = 0$	$\rho_1 = \frac{-\theta_1(1-\theta_2)}{(1+\theta_1^2+\theta_2^2)}$ $\rho_2 = \frac{-\theta_2}{(1+\theta_1^2+\theta_2^2)}$ $\rho_p = 0$	$\rho_1 = \frac{\left(-\theta_1 + \theta_1\theta_2 + \theta_2\theta_3 + .. + \theta_q\theta_{q+1}\right)}{\left(1+\theta_1^2+\theta_2^2+..+\theta_q^2\right)}$ $\rho_2 = 0$ $\rho_p = 0$

Fuente: autores, a partir de Pindyck y Rubinfeld (1998).

Finalmente se tiene la combinación de procesos AR y MA, donde se expresan como ARMA (p,q) de acuerdo con la combinación de su respectivo orden. Continuando con la dinamica anterior, en el cuadro 5.7 se presenta la forma de la media (μ), varianza (γ_0), covarianzas ($\gamma_1, \gamma_2, .., \gamma_p$) y función de autocorrelación simple

($FAS \rightarrow \hat{\rho}_p = \frac{\gamma_p}{\gamma_0}$) para la forma más sencilla ARMA (1,1). Deducidos, de la misma forma como lo expresado para AR y MA individualmente.

Cuadro 5.7. Formas de la media, varianza, covarianzas y FAP en procesos ARMA (1,1).

Orden de integración para la serie Y_t	0
proceso	ARMA(1,1)
Forma del modelo	$Y_t = \delta + \phi_1 Y_{t-1} - \theta_1 u_{t-1} + u_t$
Media μ_{Y_t}	$\frac{\delta}{(1-\phi_1)}$
Varianza $\sigma^2_{Y_t} = \gamma_0$	$\frac{1+\theta_1^2-2\phi_1\theta_1}{1-\phi_1^2}\sigma^2_{u_t}$
Covarianza γ_1	$\frac{(1-\phi_1\theta_1)(\phi_1-\theta_1)}{1-\phi_1^2}\sigma^2_{u_t}$
Covarianza γ_2	$\phi_1\gamma_1$
Covarianza γ_p	$\phi_1\gamma_{p-1}$
FAS=$\hat{\rho}_p = \frac{\gamma_p}{\gamma_0}$	$\rho_1 = \frac{(1-\phi_1\theta_1)(\phi_1-\theta_1)}{1+\theta_1^2-2\phi_1\theta_1}$ $\rho_p = \phi_1\rho_{p-1}$

Fuente: autores, a partir de Pindyck y Rubinfeld (1998).

Todo el proceso anterior, también aplica en series integradas o desestacionalizadas; simplemente es realizar un equivalente en su orden de diferenciación de esta manera $\Delta Y_t = Y_t - Y_{t-1} \Rightarrow \Delta Y_t = Z_t$ y reemplazar en todo los procesos $Y_t, .., Y_{t-p}$ por $Z_t, .., Z_{t-p}$. A continuación se realiza una pequeña reseña matematica sobre la

dualidad que todo proceso AR se represente de manera equivalente mediante un MA; en otras palabras, garantiza la invertibilidad de AR.

A.5.4.1 Condición de dualidad, estacionariedad e invertibilidad

Los procesos autorregresivos y de media móvil están estrechamente ligados, realmente es posible transformar cualquier proceso AR (p) en un MA y un MA en AR (p); esto se conoce como invertibilidad y es posible dado que: todo proceso MA por naturaleza es estacionario y $\sum_{t=1}^{p}|\phi_p| < 1$, para que converja en ambos sentidos y pueda ser llevada a cabo la metodología Box-Jenkins (*véase* ecuación 5.106, 5.107, 5.108 y 5.109). Así, empleando operador y polinomio de rezago en un AR (p) se tiene:

$$\left(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p\right) Y_t = u_t \text{ (5.106)}$$

$$\phi(L) Y_t = u_t \text{ (5.107)}$$

$$Y_t = \phi^{-1}(L) u_t \text{ (5.108)}$$

$$Y_t = \left(1 - \phi_1 L - \phi_2 L^2 - \ldots - \phi_p L^p\right)^{-1} u_t \text{ (5.109)}$$

De esta manera, la ecuación 5.109 representa la estimación no lienal de un modelo MA derivado a partir de un AR (p), como se observa en el procedimiento desde la ecuación 5.106 hasta 5.109 No obstante, si $\sum_{t=1}^{p}|\phi_p| \geq 1$ el proceso AR no converge caracterizandolo así como no estacionario e imposibilitando de esta manera aplicar la técnica BJ.

A.5.5 Círculo unitario y estacionariedad

De acuerdo con la condición $\sum_{t=1}^{p}|\phi_p| < 1$ para que converja un AR (p), los valores de cada ϕ_p deben ser admisibles para que las raíces complejas estén dentro del circulo unitario garantizando la estabilidad del proceso y modelo. Para lo anterior, suponga un proceso AR (2) (*véase* ecuación 5.110) cuyo polinomio se expresa en la ecuación 5.111 sus raíces a partir de una solución homogénea equivalen a la expresiones 5.112 y 5.113.

$$(1 - \phi_1 L - \phi_2 L^2) Y_t = u_t \text{ (5.110)}$$

$$1 - \phi_1 L - \phi_2 L^2 = 0 \text{ (5.111)}$$

$$\lambda^2 - \phi_1 \lambda - \phi_2 = 0 \text{ (5.112)}$$

$$\lambda = \frac{\phi_1 \pm \sqrt{\phi_1^2 + 4\phi_2}}{2} \text{ (5.113)}$$

$$\sqrt{\phi_1^2 + 4\phi_2} < 2 \pm \phi_1, \text{ donde } \phi_1 \pm \phi_2 < 1 \text{ (5.114)}$$

A partir de la ecuación 5.113 sus raíces en valor absoluto deben ser inferiores a uno (*véase* ecuación 5.114), cumpliendo así con la condición exigida, dado que el circulo por debajo donde se tiene las raíces complejas está determinado por $\sqrt{\phi_1^2 + 4\phi_2} = 0$.

Capítulo 6

Modelos de rezagos distribuidos y autorregresivos, causalidad de Granger y cointegración.

6.1 Introducción

En los dos últimos capítulos se explicaron los conceptos y metodologías básicas de series de tiempo, los cuales pretendían modelar el comportamiento a lo largo del tiempo de una variable. El presente capítulo explora el análisis de relaciones temporales entre más de dos series de tiempo. Estas nuevas metodologías se conocen como modelos econométricos dinámicos. A diferencia de los modelos econométricos estáticos, los dinámicos incorporan el efecto del tiempo o el rezago que requiere la variable dependiente para responder a los cambios en las variables independientes. (Gujarati, 2003, 633). El reconocimiento de un factor temporal en los modelos de regresión simple, conduce a la aceptación que los fenómenos económicos son dinámicos, y no estáticos (Pena, _,2).

De acuerdo a lo anterior, en las secciones se estudiaran los modelos de rezagos distribuidos y autorregresivos. Los primeros, consideran que la respuesta total de la variable dependiente ante los cambios en las independientes se alcanza después de ciertos periodos de tiempo. Los segundos, están caracterizados por incluir, en las variables explicativas, rezagos de la variable dependiente, adicional a las variables explicativas de interés. Asimismo, se tratan conceptos los conceptos de variables de expectativas, multiplicadores y elasticidades de corto y largo plazo, periodo de ajuste, criterios de Akaike y Schwarz, trasnformaciones de Koyck y Almon, expectativas adaptativas, ajuste parcialy Nerlove. Adicionalmente, este capítulo discute causalidad de Granger y cointegración.

Finalmente, para comprender los modelos presentados en este capítulo, se desarrolla un estudio de caso basado en la información sobre oferta de azúcar extraída de Hill, Griffiths y Judge (2001).

6.2 Introducción a modelos con variables rezagadas

El estudio del comportamiento de las variables económicas requiere el entendimiento de las relaciones temporales de las mismas; reconociendo la respuesta de éstas a eventos exógenos y endógenos con cierta periodicidad o rezago. Los modelos econométricos que integran rezagos en las variables explicativas, se conocen como modelos de regresión dinámica (*véase* ecuación 6.1).

$$Y_t = \alpha + \sum_{p=0}^{\infty} \beta_p X_{t-p} + u_t \qquad (6.1)$$

En la ecuación 6.1, Y_t es una serie de tiempo que figura como variable dependiente del modelo, X_{t-p} son las variables explicativas del modelo, teniendo en cuenta todos los rezagos de la serie de tiempo ($p = 0,1,2....$) y u_t es el error del modelo. Bajo esta especificación, un cambio en la variable independiente X_{t-p} a cualquier punto del tiempo puede afectar a Y_t de la forma $E(Y_s \mid X_t, X_{t-1},...)$ en los periodos presentes. Si se cree que los efectos de la variable explicativa toman demasiado tiempo, los modelos a tener en cuenta son los de rezagos infinitos, en el caso contrario se utilizan los modelos de rezagos finitos (Greene, 2003, 560)[125].

Partiendo de lo anterior, para medir el efecto de un cambio marginal de una variable independiente (X_{t-p}) sobre la dependiente se utilizan los efectos marginales, puesto que reflejan las secuelas que tiene la variable explicativa en el corto y largo plazo. El coeficiente β_0 es el estimador que captura los efectos de corto plazo y se conoce como el *"multiplicador de corto plazo o de impacto"*. Asimismo, si el cambio en X_{t-p} se mantiene constante a través de los periodos, las sumas parciales de los coeficientes a lo largo de los periodos[126], se conoce como el *"multiplicador de largo plazo o de equilibrio"*

[125] En la realidad los efectos de las variables tienen persistencia en periodos cortos de tiempo.

[126] $\sum_{p=0}^{k} \beta_p = \beta_0 + \beta_1 + \beta_2 + ... + \beta_k = \beta$

En relación con lo anterior, la inclusión de rezagos está justificada en el comportamiento del mercado y agentes económicos. Las razones se han denominado como: psicológicas, dentro de este grupo se destacan los hábitos de consumo determinados por el ingreso; tecnológicas, como por ejemplo actualización de capital; e institucionales como contratos. En el caso en el que se modele un evento económico equivocadamente como estático sin considerar las características fundamentales de los individuos, los resultados empíricos serán incorrectos. A continuación, se expondrán los procedimientos que se siguen cuando se trabaja con modelos dinámicos.

6.2.1 Operadores de rezago y diferencia para modelos dinámicos

Una forma de manipular los modelos con variables rezagadas es utilizar los operadores de rezago, esto permite trasnformar algebráicamente las expresiones para encontrar modelos simplificados, por ejemplo: modelos autorregresivos (considerados en el capítulo 5) o rezagos distribuidos (*véase* ecuación 6.2).

$$L^p X_t = X_{t-p} \qquad (6.2)$$

En la expresión 6.2, X_{t-p} es una variable rezagada p periodos; L es el operador de rezago. De acuerdo a esto, se pueden desarrollar operaciones relacionadas, como la de primeras diferencias. La transformación consiste en rezagar la variable un periodo y restarla a la variable inicial (*véase* ecuación 6.3). Esto última se puede utilizar para reescribir el modelo de regresión dinámica de la ecuación 6.1 (*véase* ecuación 6.4)

$$\Delta X_t = X_t - X_{t-1} = (1-L)X_t \qquad (6.3)$$

$$Y_t = \alpha + \sum_{p=0}^{\infty} \beta_p L^p X_t + u_t = \alpha + B(L)X_t + u_t \qquad (6.4)$$

En la ecuación 6.4, $B(L)$ es un polinomio que agrupa todos los rezagos. A su vez, éste se puede reescribir a través de la expansión de Maclaurin o de Taylor [127] (*véase* ecuación 6.5).

$$Y_t = \alpha + \frac{1}{1-bL} X_t + u_t \qquad (6.5)$$

La ecuación 6.5 se conoce como la *forma de promedio móvil o la forma de rezago distribuido* (*véase* sección 6.3). Si se multiplica la ecuación 6.5 por $(1-bL)$ se consigue la expresión de la *forma de modelo autorregresivo* (*véase* ecuación 6.6 y sección 6.3).

$$Y_t = \alpha(1-bL) + \beta X_t + bY_{t-1} + (1-bL)u_t \qquad (6.6)$$

En la ecuación 6.6, el operador de rezago es útil para representar los modelos de variable rezagada. Las siguientes secciones mostrarán los conceptos de modelos dinámicos con más detalle, con el fin de ofrecer una justificación en la teoría económica.

6.3 Modelos de rezagos distribuidos y autorregresivos

Como se discutió anteriormente, los modelos dinámicos cuentan con rezagos de las variables independientes como variables explicativas. Esto permite corroborar efectos de un cambio marginal en alguna variable independiente sobre la variable dependiente, no solo en el periodo presente sino también a largo del tiempo. Un ejemplo claro de esto es la curva J del ajuste gradual de los flujos comerciales. Se suele observar con frecuencia que la cuenta corriente de un país empeora inmediatamente después de una depreciación real de la moneda local, y comienza a mejorar sólo algunos meses más tarde (Krugman y Obstfeld, 2006, 468). Un modelo general de este estilo, formalmente tendría una variable independiente (X) con r rezagos (*véase* ecuación 6.7):

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_r X_{t-r} + u_t \qquad (6.7)$$

---

[127] $B(L) = 1 + bL + (bL)^2 + \ldots = \sum_{i=0}^{\infty} (bL)^i = \frac{1}{1-bL}$, con $|b| < 1$.

La ecuación 6.7 relaciona dos series de tiempo, si la variable X cambia en una unidad en el periodo t, el efecto esperado en la variable dependiente seria β_0 para el mismo periodo. Si se evalúa para $t-1$ el efecto seria capturado por β_1, si es para $t-2$ seria β_2 (siempre y cuando todo lo demás se mantenga constante en el tiempo). Estas interpretaciones resultan útiles cuando se requieren hacer análisis de los efectos en el tiempo relacionados con cambios en las variables explicativas del modelo.

Por otro lado, se puede derivar otra especifiación de rezagos distribuidos, denominados modelos autorregresivos (ADL, por Autorregresive Distributeg Lags en inglés). Éstos son una ampliación de los procesos autorregresivos expuestos en el capítulo 5. A diferencia de los últimos, estos modelos utilizan dos series de tiempo, al tiempo que incluye uno a más rezagos de la variable dependiente como una explicativa. Se conocen también como modelos dinámicos, puesto que señalan la trayectoria en el tiempo de la variable dependiente en relación con sus valores pasados (*véase* ecuación 6.8).

$$Y_t = \alpha + \sum_{i=0}^{r} \beta_i X_{t-i} + \sum_{g=1}^{p} \gamma_i Y_{t-g} + u_t \qquad (6.8)$$

En la ecuación 6.8, se supone que u_t no presenta problemas de autocorrelación y heteroscedasticidad. Utilizando las expresiones de polinomios de rezagos desarrollados en la sección 6.2.1, la ecuación 6.8 se puede reescribir como:

$$A(L)Y_t = \alpha + B(L)X_t + u_t \qquad (6.9)$$

Donde $A(L)$ es un polinomio de orden p y $B(L)$ es un polinomio de orden r. Bajo esta especificación el modelo autorregresivo se escribe como ARDL (p,r)[128]. Los rezagos de la ecuación 6.9 implican un conjunto de respuestas dinámicas que afectan el comportamiento de la variable dependiente Y_t en el corto, medio y largo plazo (Johnston y Dinardo, 1997, 245).

[128] *Véase* (Greene,2003, 571)

Para conocer empíricamente los resultados de estos modelos, se deben desarrollar las estimaciones de los mismos a través de regresión lineal. A priori se puede establecer un resultado sencillo de MCO, con el cual se derivan los estimadores que indican los cambios de variables en distintos periodos del tiempo. Aunque parecen básicas, las estimaciones pueden llegar a presentar dificultades, como:

1- La escogencia del número de rezagos[129] que influye en la estimación, a medida que se tienen en cuenta más rezagos se pierden grados de libertad.
2- Los rezagos están altamente correlacionados, originando multicolinealidad en el modelo.
3- Puede existir problema de autocorrelación en los errores por la presencia de variables dependientes rezagadas.

En vista de estas dificultades, es posible encontrar una expresión básica que conlleve a los mismos resultados esperados en un principio. Para este fin, se tendrán en cuenta dos transformaciones: la de Koyck y Almon. Asimismo se discutirán los modelos de expectativas adaptativas, de ajuste parcial (Nerlove) y se tendrá en cuenta la prueba durbin h para detectar problemas de autocorrelación.

6.3.1 Modelo de Koyck

La transformación de Koyck es la primera aproximación al manejo de modelos de rezagos distribuidos. Es un método alternativo para estimar los modelos de rezagos distribuidos infinitos, imponiendo a priori condiciones sobre los coeficientes β_i. Considere el siguiente modelo de rezago distribuido infinitos:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + ... + u_t \qquad (6.10)$$

En la ecuación 6.10, Y_t es una serie de tiempo que figura como variable dependiente del modelo, X_{t-p} son los rezagos de la variable explicativa del modelo. Si todos los estimadores tienen el mismo signo, Koyck supone que todos en su conjunto se pueden reducir geométricamente de la siguiente forma:

[129] Para determinar la longitud del rezago es posible utilizar los criterios de Akaike y Schwartz referenciados en el capítulo 5.

$$\beta_p = \beta_0 \lambda^p \text{, con } p = 1, 2, \ldots \qquad (6.11)$$

Donde $0 < \lambda < 1$ es la tasa de descenso del rezago distribuido y $1 - \lambda$ es la velocidad de ajuste. Como se ve en la ecuación 6.11, el coeficiente β_p depende de β_0 y λ. Entre más cerca esté λ a uno, más lento será el ajuste y por tanto los valores del pasado lejano de X tienen un impacto más notable sobre Y_t. Mientras que si λ está más hacia cero la velocidad de ajuste es mayor, es decir, los valores de X más presentes tienen efectos mayores sobre el comportamiento de Y_t (Gujarati, 2003, 641). Utilizando la expresión 6.11, el modelo 6.10 se puede reescribir como:

$$Y_t = \alpha + \beta_0 X_t + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \ldots + u_t \qquad (6.12)$$

La ecuación 6.12 es el primer paso para transformar el modelo de rezagos infinitos. Hay que anotar que los nuevos coeficientes del modelo dejan de ser lineales y por tanto el análisis de regresión clásica resulta inválido. Por tanto, el modelo necesita una restructuración que permita la estimación pertinente. Siguiendo con el proceso, resulta conveniente rezagar un periodo la ecuación 6.12 (*véase* ecuación 6.13).

$$Y_{t-1} = \alpha + \beta_0 X_{t-1} + \beta_0 \lambda X_{t-2} + \beta_0 \lambda^2 X_{t-3} + \ldots + u_{t-1} \qquad (6.13)$$

Se multiplica por λ a ambos lados de la ecuación 6.13:

$$\lambda Y_{t-1} = \lambda \alpha + \lambda \beta_0 X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \beta_0 \lambda^3 X_{t-3} + \ldots + \lambda u_{t-1} \qquad (6.14)$$

Y por último, se restan las ecuaciones 6.12 y 6.14 obteniendo la siguiente expresión:

$$Y_t = \alpha(1 - \lambda) + \beta_0 X_t + \lambda Y_{t-1} + \upsilon_t \qquad (6.15)$$

En la ecuación 6.15, $\upsilon_t = u_t - \lambda u_{t-1}$ es un promedio móvil del modelo inicial. Por tanto, esta ecuación se conoce como la transformación de Koyck y resulta ser un modelo autorregresivo. Bajo este método se logra simplificar la estimación de un modelo de infinitos parámetros a uno en el que hay que estimar solo tres. Aunque la expresión 6.15 resulta favorable en la eliminación del efecto de

multicolinealidad, el nuevo modelo con Y_{t-1} como variable explicativa genera otro inconveniente y es la correlación entre los errores $\upsilon_t = u_t - \lambda u_{t-1}$.

Una vez se tiene claro la estructura de los modelos de rezagos distribuidos, es posible caracterizar algunos resultados derivados de la estimación de éstos a través de la transformación de Koyck. El rezago medio o mediano se utiliza para tipificar la estructura de los rezagos.

El rezago mediano (RM) corresponde al tiempo que se requirió para que se diera el 50 por ciento del cambio total en Y_t cuando se tiene un cambio marginal en X_t. En el modelo autorregresivo toma la forma: $RM = -\frac{\log(2)}{\log(\lambda)}$. Mientras tanto el rezago medio (RME) es el promedio ponderado de todos los rezagos involucrados y está representado para el método de Koyck por: $RME = \frac{\lambda}{1-\lambda}$. Estos dos rezagos resulta ser una medida que resume la velocidad con la cual se siente el efecto en Y_t cuando cambia X_t.

Adicionalmente, con la información de los modelos dinámicos es posible derivar las elasticidades de corto y largo plazo. Recordando el concepto de elasticidad, ésta es una medida de la capacidad de respuesta porcentual de una variable ante variaciones de otra. En el caso particular de los modelos dinámicos, las elasticidades hacen referencia al ajuste que tienen las variables consideradas a lo largo del tiempo.

De acuerdo a lo anterior, las elasticidades se derivan, preferiblemente de un modelo especificado doblemente logarítmico, a partir de los estimadores de un modelo econométrico. Considerando la ecuación 6.7 la elasticidad en el corto plazo es igual al coeficiente que acompaña a la variable X_{t-1} . Mientras que la de largo plazo es igual a la razón entre el coeficiente β_0 y el coeficiente β_r.

6.3.2 Modelo de expectativas adaptables

En el modelo de Koyck explicado en la sección 6.3.1, se transformaba el modelo de rezagos distribuidos en uno autorregresivo. Aún así, esta transformación no cuenta con un sustento teórico para su aplicación en el análisis económico. Por esta razón, se trata una aplicación razonable que es el modelo de expectativas adaptativas (Cagan, 1956).

Considere un modelo dinámico donde se tenga en cuenta una variable explicativa caracterizada por ser los valores esperados en el largo plazo (*véase* ecuación 6.16). Esto es porque existen modelos que representan dicho razonamiento como un componente importante en la especificación de los modelos econométricos.

$$Y_t = \beta_0 + \beta_1 X_t^* + e_t \quad \text{con} \quad X_t^* - X_{t-1}^* = \gamma(X_t - X_{t-1}^*) \qquad (6.16)$$

En la ecuación 6.16, Y_t es la variable dependiente, X_t^* es la variable independiente esperada de X_t, γ es el coeficiente de esperanza y e_t el término del error. La idea básica del modelo de expectativas adaptables es que se especifica el modelo con base en la información del pasado cercano, asumiendo que en economía los agentes se toman cierto tiempo en cambiar sus perspectivas sobre lo que ocurre en la realidad, debido a la incertidumbre del futuro (Gujarati, 2003, 459).

Ahora si se rezaga un periodo la ecuación 6.16 y dicho resultado se multiplica por $(1-\gamma)$ se obtiene la siguiente expresión:

$$(1-\gamma)Y_{t-1} = (1-\gamma)\beta_0 + \beta_1(1-\gamma)X_{t-1}^* + (1-\gamma)e_{t-1} \qquad (6.17)$$

Finalmente se restan la expresión 6.16 incluyendo la expresión de X_t^* a 6.17 y se obtiene un representación del modelo autorregresivo para expectativas adaptativas (*véase* ecuación 6.18).

$$Y_t = \beta_0 + \gamma\beta_1 X_t + (1-\gamma)Y_{t-1} + \upsilon_t \ \text{con} \ \upsilon_t = e_t - (1-\gamma)e_{t-1} \qquad (6.18)$$

La ecuación 6.18 es una expresión de un modelo autorregresivo que además tiene como variable independiente a X_t que es el valor real observado de la variable.

Adicional a los modelos de expectativas adaptativas también existen otros basados en modelos autorregresivos que están en sustentados en otro razonamiento teórico (véase sección 6.3.3)

6.3.3 Modelo de ajuste parcial

El modelo de ajuste parcial (Neverloe, 1956) es otro razonamiento de la transformación de Koyck, en la que se trae a la realidad económica los modelos dinámicos. Esta es una técnica basada en la idea de ajuste en el tiempo de variables en la economía, en especial aquellas relacionadas con costos.

De forma general, lo modelos de ajuste parcial se especifican de acuerdo a los efectos de las variables explicativas sobre un valor óptimo de la variable que se desea explicar (Y_t^*) (*véase* ecuación 6.19).

$$Y_t^* = \alpha_0 + \beta_1 X_t + e_t \quad (6.20)$$

En la ecuación 6.20, Y_t^* es la variable explicativa óptima, X_t es una serie de tiempo que explica a Y_t^* y e_t es el termino del error. De acuerdo a lo anterior, no es posible alcanzar el óptimo todas las veces, por tanto

$$Y_t - Y_{t-1} = \gamma(Y_t^* - Y_{t-1}) \quad (6.21)$$

La ecuación 6.21 muestra el diferencial de la variable dependiente en dos periodos de tiempo (t y $t-1$). Teniendo en cuenta lo anterior, el modelo a estimar estaría dado por:

$$Y_t = \gamma\alpha_0 + \beta_1\gamma X_t + (1-\gamma)Y_{t-1} + \gamma e_t \quad (6.22)$$

La expresión 6.22 resulta ser un modelo autorregresivo de rezagos distribuidos similar al expuesto en la transformación de Koyck, indicando que es una aproximación coherente al uso de modelos dinámicos. Asimismo se caracteriza la velocidad de ajuste a través del coeficiente γ representando la dinámica del modelo.

De acuerdo a las características presentadas, se piensa en la similitud en las aproximaciones generadas de la transformación de Koyck. En apariencia los modelos de expectativas adaptativas y los de ajuste parcial son parecidos, pero el enfoque que cada uno trabaja es distinto. Para el primero, se está modelando la incertidumbre de los agentes económicos a lo que suceda en el futuro, por consiguiente lo mejor que puede esperar es que suceda es lo mismo que ya se sabe de tiempo atrás. Mientras el segundo, se refiere a la rigidez técnica e institucional en variables económicas que no permiten generar cambios contemporáneos cuando se dan eventos exógenos.

6.3.4 Modelo de Almon

El modelo de Almon es una extensión de la transformación de Koyck que flexibiliza el supuesto que los estimadores β se reducen geométricamente a la expresión de la ecuación 6.10, puesto que es un supuesto fuerte y restrictivo. En realidad no solo se presentan relaciones ascendentes (descendentes) entre variables de series de tiempo. También se pueden dar casos en los que la serie sigue un comportamiento cíclico. Por tanto, es conveniente ajustar los modelos con curvas acordes con el comportamiento de las series de tiempo. El método que trata lo descrito se conoce como el modelo de Almon (Judge et al, 1985, 729).

Partiendo de lo anterior, el método de Almon parte de la base que se tiene un modelo finito de rezagos distribuidos como el siguiente:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_k X_{t-k} + u_t \qquad (6.23)$$

Siguiendo el teorema de Weierstrass, que establece que una función continua, en un intervalo cerrado y acotado, puede ser aproximada uniformemente mediante un polinomio de un grado apropiado (Montalvo, 2003, 31), Almon (1965) supone que β_i puede ser aproximado mediante a un polinomio en i de forma general (*véase* ecuación 6.24)[130]:

$$\beta_i = a_0 + a_1 p + a_2 p^2 + \ldots + a_m p^m \qquad (6.24)$$

[130] *Véase* (Gujarati, 2003, 664)

La ecuación 6.24 es un polinomio de grado m en p. Se supone adicionalmente que $m < p$. Para ilustrar el método de Almon, se supone que β_i se ajusta a un polinomio de grado 2, por tanto la ecuación 6.24 quedaría $\beta_i = a_0 + a_1 i + a_2 i^2$. Si se reemplaza en 6.23 el modelo quedaría

$$Y_t = \alpha + a_0 Z_{t0} + a_1 Z_{t1} + a_2 Z_{t2} + u_t \qquad (6.25)$$

En la ecuación 6.25, $Z_{t0} = \sum_{p=0}^{k} X_{t-p}$, $Z_{t1} = \sum_{p=0}^{k} pX_{t-p}$, $Z_{t2} = \sum_{p=0}^{k} p^2 X_{t-p}$. En el esquema de Almon, se estima un nuevo modelo de regresión de Y_t sobre las variables Z_{tj}. A partir de las estimaciones de los parámetros que acompañan a las Z_{tj} se obtienen los valores de los β. Bajo esta perspectiva, la estimación por MCO es factible; siempre y cuando se satisfagan los supuestos del modelo de regresión lineal.

6.3.5 Detección de autocorrelación en modelos autorregresivos.

Una vez se han estudiado las características fundamentales de los modelos de rezagos distribuidos y autorregresivos, es pertinente realizar pruebas que garanticen el cumplimiento de los supuestos básicos de regresión lineal. En especial, estos modelos están expuestos a problemas de autocorrelación. Bajo esta perspectiva, en esta sección se desarrolla la prueba h propuesta por Durbin (1970), que sustituye la prueba d de Durbin y Watson[131], para verificar la correlación serial del modelo autorregresivo de orden uno. La prueba de hipótesis consiste en corroborar si existe o no correlación serial (*véase* prueba de hipótesis 6.26).

$$\begin{aligned} H_0 &: \rho = 0 && \text{No existe correlación serial} \\ H_1 &: \rho \neq 0\,. && \text{Existe correlación serial} \end{aligned} \qquad (6.26)$$

De acuerdo a lo anterior se deriva un nuevo estadístico de prueba específico para evaluar correlación serial en modelos autorregresivos (*véase* ecuación 6.27).

[131] La prueba d de Durbin y Watson se encuentra sesgada para modelos autorregresivos. Para más detalles *véase* (Gujarati, 2003, 655)

$$h = \hat{\rho}\sqrt{\frac{n}{1-(n\,\mathrm{var}(\hat{\gamma}_i))}} \qquad (6.27)$$

En la ecuación 6.27, n es el tamaño de la muestra; $\mathrm{var}(\hat{\gamma}_i)$ es la varianza del coeficiente del rezago de la variable dependiente que está como variable explicativa en el modelo inicial 6.24 asumiendo que $i = 1$[132]; y que $\hat{\rho}$ viene dado del estimador de Durbin y Watson $d = 2(1-\hat{\rho})$. Durbin demostró que cuando $\rho = 0$, el estadístico h sigue la distribución normal estándar (asintóticamente). La prueba se contrasta como normal estándar, si h tiende a cero quiere decir que no existe autocorrelación[133].

Una vez que se han desarrollado las pruebas para descartar o corregir problemas de autocorrelación, los modelos autorregresivos se pueden estimar por MCO y con la seguridad de que las conclusiones derivadas de ese proceso resultan correctas al momento de analizar el comportamiento en el tiempo de las relaciones entre las variables económicas.

6.4 Prueba de causalidad de Granger

En las secciones anteriores se explicaron los modelos pertinentes para evaluar las relaciones en el corto y largo plazo entre las variables en consideración. Dichas relaciones no son de causalidad, es decir, la relación entre variables no significa influencia. Bajo la prueba de causalidad de Granger se pretende encontrar algún tipo de causalidad entre las variables.

La prueba de causalidad de Granger pretende determinar si las observaciones pasadas de una variable de series de tiempo permiten pronosticar a otra. Ésta indica, de acuerdo a los datos, si una variable causa a otra. Asimismo sirve para establecer si existe exogeneidad en el modelo, y es semejante a decir que no existe causalidad en el sentido de Granger.

---

[132] No importa cuántos valores rezagados de Y se incluyan en el modelo de regresión, para la prueba h solo es necesario tener en cuenta la varianza del primer rezago.

[133] Existe una prueba más potente que sirve para muestras grandes y pequeñas que se conoce como la prueba de Breusch-Godfrey.

En esta sección se estudiará un caso simple, en el que se quiere probar si X causa en el sentido de Granger a Y (esta prueba también es válida en el sentido contrario). Para llevarla a cabo empíricamente se debe determinar, en primera instancia, los modelos a contrastar. El modelo restringido (*véase* ecuación 6.28) es aquel que no tiene en cuenta los rezagos de la variable X que es la de interés, al mismo tiempo se consideran los rezagos de la variable dependiente. El modelo no restringido (*véase* ecuación 6.29) está caracterizado por los rezagos de la variable de interés y los de la variable dependiente como variables explicativas.

Modelo restringido: $$Y_t = \alpha + \beta_1 X_t + \sum_{i=1}^{p} \gamma_i Y_{t-i} + e_t \qquad (6.28)$$

Modelo no restringido: $$Y_t = \alpha + \sum_{i=0}^{r} \beta_i X_{t-i} + \sum_{i=1}^{p} \gamma_i Y_{t-i} + u_t \quad (6.29)$$

El procedimiento práctico consiste en:

1. Estimar la ecuación 6.28 por MCO y obtener la suma de cuadrados del error del modelo restringido (SCE_R).
2. Estimar la ecuación 6.29 por MCO y obtener la suma del cuadrados del error del modelo no restringido (SCE_{NR}).
3. Realizra una prueba de hipótesis que indica en la hipótesis nula que X no causa en el sentido de Granger a Y .

$$\begin{aligned} &H_o : \beta_1 = \beta_2 = ... = \beta_r = 0 && X \text{ no causa a } Y . \\ &H_1 : \beta_j \neq 0 . && X \text{ causa a } Y . \end{aligned} \qquad (6.30)$$

4. Para generar conclusiones sobre las hipótesis, se utiliza el estadístico de prueba F y se concluye bajo los criterios de decisión de la misma.

$$F_c = \frac{(SCE_R - SCE_{NR})/r}{SCE_{NR}/N-k} \sim F_{r,n-k} \qquad (6.31)$$

En la ecuación 6.31, r es igual al número de rezagos de la variable X , k es el número de parámetros estimados del modelo no restringido y n es el

número total de observaciones. Si $F_c > F_{r,n-k}$ entonces se dice que se rechaza H_0, con lo que se concluye que los términos del rezago para X conjuntamente son significativos. Por tanto se dice que X causa en el sentido de Granger a Y. Estos resultados se obtienen al suponer que las variables que se están contrastando son estacionarias y los errores no están correlacionados (Gujarati, 2003, 673).

6.5 Cointegración

Otra metodología que utiliza modelos de series de tiempo bivariados es la cointegración. Esta técnica aprovecha la relación entre dos series integradas para encontrar relaciones de corto plazo. De esta forma se pueden analizar políticas económicas, al tiempo que se logran hacer proyecciones. Por tanto, la relación lineal entre dos series no estacionarias puede resultar un buen mecanismo para conseguir eliminar las tendencias estocásticas de dos series relacionadas. La diferencia entre un par de series integradas puede ser estacionaria, y esta propiedad se conoce como cointegración (Granger, 2004).

De acuerdo a lo anterior, se requieren dos series no estacionarias que se combinen linealmente para conseguir errores estacionarios. En términos económicos, dos variables serán cointegradas si existe una relación a largo plazo, o de equilibrio, entre ambas (Gujarati, 2003, 796). En la práctica, muchos pares de series macroeconómicas parecen tener dicha propiedad, tal como se deduce de la teoría económica. Sin embargo existen otras series integradas que no cumplen dicha propiedad, y consecuentemente, cuando ocurre, se conoce como una *regresión espúrea*[134].

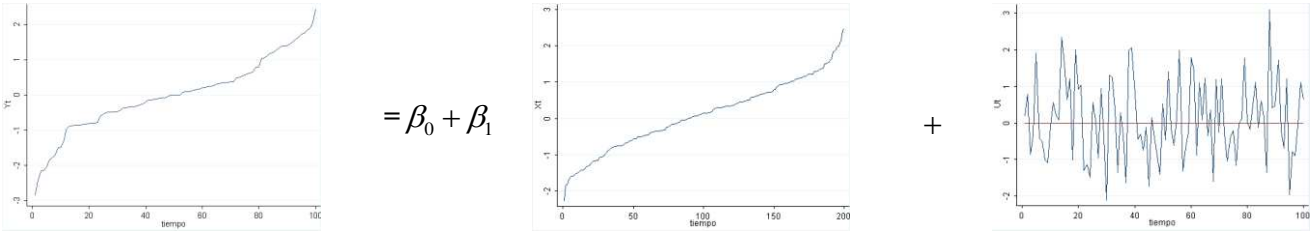
El análisis de cointegración es esencial cuando se tiene una combinación de variables que presenten una similitud en el orden de integración (*véase* capítulo 5), en especial, cuando las series de tiempo son $I(1)$ (véase gráfica 6.1). Si lo anterior se

[134] El término de correlación espúrea fue acuñado por Karl Pearson en 1924, y decía que una correlación puede describirse como espúrea si es inducida en el método de datos y no está presente en la información original.

cumple, una combinación lineal de estas variables, que sea estacionaria, se conoce como regresión cointegrante (*véase* ecuación 6.32)

Gráfica 6.1 Cointegración de la ecuación 6.32

Y_t $= \beta_0 + \beta_1$ X_t $+$ u_t



$I(1)$ $I(1)$ $I(0)$

Fuente: cálculos autores a partir de Montenegro (2007)

$$Y_t = \beta_0 + \beta_1 X_t + u_t \quad (6.32)$$

Una vez se tiene claro que las series son potencialmente cointegradas, y cuando se garantice que los errores del modelo son estacionarios, la metodología de estimación básica de MCO es útil para encontrar qué tipo de relación existe entre las variables de un modelo de regresión lineal como en la ecuación 6.31.

La corroborar estadísticamente que existe cointegración se realiza una prueba de estacionariedad de Dickey-Fuller a los errores predichos de la ecuación lineal 6.31 (*véase* prueba de hipótesis 6.33).

$$\begin{array}{ll} H_0 : u_t \sim I(1) & X \text{ no causa a } Y. \\ H_1 : u_t \sim I(0) & X \text{ causa a } Y. \end{array} \quad (6.33)$$

Al igual que en el capítulo 5, la prueba de hipótesis 6.33 se constata con el estadístico de prueba tau (τ) (véase ecuación 6.34).

$$\tau = \frac{\hat{\delta}}{ee(\hat{\delta})} (6.34)$$

En la ecuación 6.34, $\hat{\delta}$ es el estimador de que hace referencia a la combinación lineal de $\hat{\beta}$ de la ecuación 6.32[135]. El estadístico τ se compara con los valores críticos de la tabla de MacKinnon de acuerdo a los niveles de 0.01, 0.05 o 0.1. Una propiedad útil de las predicciones basadas en la cointegración es que, cuando se prolongan de alguna manera hacia adelante, las predicciones de las dos series forman una razón constante, tal y como se espera por parte de algunas teorías económicas asintóticas (Catalán).

[135] $u_t = Y_t - \beta_0 - \beta_1 X_t$

6.6 Estudio de caso: oferta de azúcar

Adicional a la discusión de los conceptos de los modelos econométricos dinámicos, a continuación se expone un ejercicio empírico para ilustrar el funcionamiento de modelos de rezagos distribuidos y autorregresivos. Lograr entender los efectos y cambios en el tiempo es una de las ventajas de los modelos de variables rezagadas, para este caso se evaluarán las elasticidades de corto y largo plazo para un modelo de oferta de Caña de azúcar.

Los datos del ejercicio empírico fueron tomados del libro "*Undergraduate Econometrics*" de Hill, Griffiths y Judge (2001). La información a utilizar corresponde al área sembrada de caña de azúcar (como Proxy de la oferta) y el precio que recibe el productor en Banglasesh.

En esta sección se trabajará con distintas especificación de la oferta de azúcar, para ilustrar las diferentes metodologías consideradas dentro del capítulo. En primer lugar, se tendrán en cuenta modelos de rezagos distribuidos, luego modelos de expectativas adaptativas y ajuste parcial, después modelo Almon, posteriormente se generara la causalidad de Granger y finalmente cointegración (*véase* cuadro 6.1)

Cuadro 6.1. Modelos de rezagos distribuidos para la producción de caña de azúcar

Modelo	Especificación
Koyck/Almon	$Q_t = \alpha + \beta_0 P_t + \beta_1 P_{t-1} + \beta_2 P_{t-2} + \beta_3 P_{t-3} + u_t$
Expectativas adaptativas/Ajuste parcial	$Q_t = \alpha + \lambda_1 P_{t-1} + \lambda_2 Q_{t-1} + \nu_t$
Causalidad Granger	$Q_t = \alpha + \varphi_1 P_t + \varphi_2 Q_{t-1} + \upsilon_t$
Cointegración	$Q_t = \alpha + \gamma_1 P_t + \mu_t$

Fuente: los autores

Donde Q_t es el área cultivada de caña de azúcar en el periodo t, P_t y P_{t-1} es el precio de la caña de azúcar que recibe el productor en el periodo t y $t-1$, y Q_{t-1} es el área cultivada de caña de azúcar en $t-1$ (*véase* cuadro 6.2).

Cuadro 6.2. Variables a usar en las ecuaciones del modelo de rezagos distribuidos

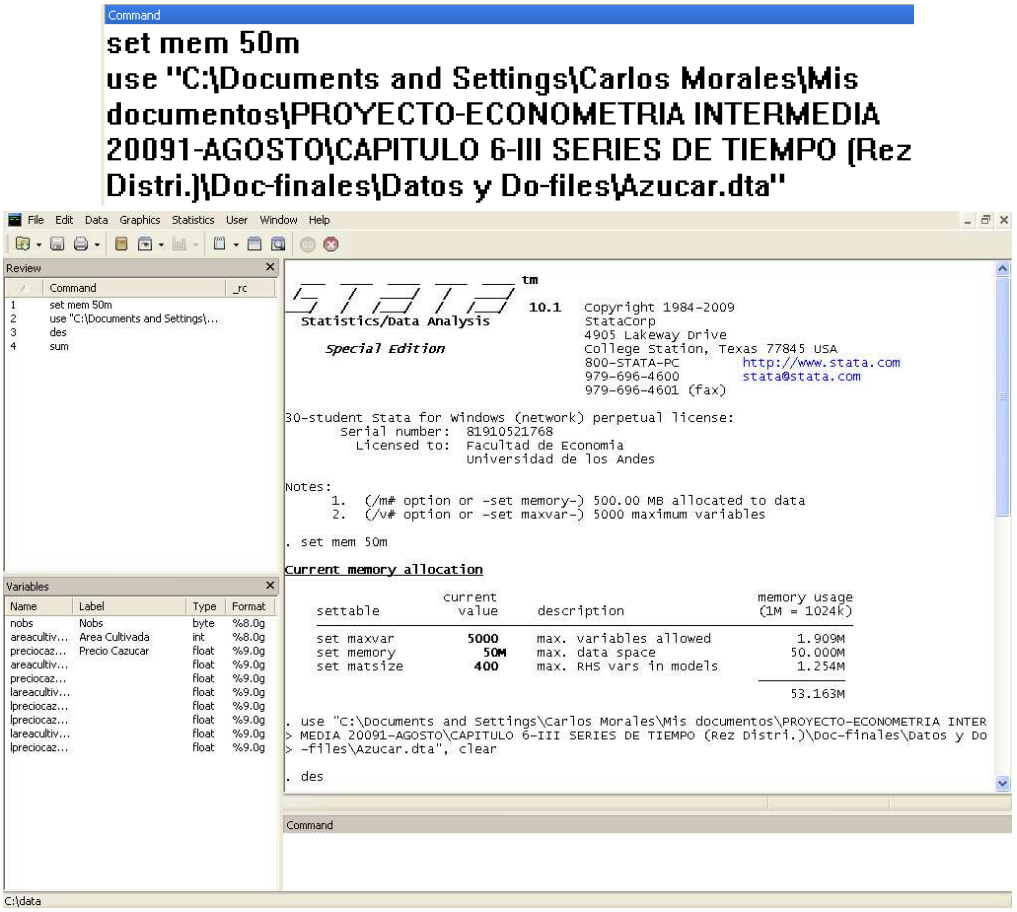
Tipo de Variable	Variable del Modelo	Variables en la Base	Descripción
Dependiente	Q_t	lareacultivada	Logaritmo del área cultivada de caña de azúcar en Bangldesh
Independientes	$P_t, P_{t-1}, P_{t-2}, Q_{t-1}$	Lpreciocazucar, Lpreciocazucar_1, Lpreciocazucar_2, lareacultivada_1	**1.** Las tres primeras hacen referencia al logaritmo del precio de la caña de azúcar para el periodo t, t-1 y t-2 **2.** La última variable es el logaritmo del área cultivada de caña de azúcar en t-1.

Fuente: los autores

6.6.1 Análisis general de los datos

1. A partir de lo anterior y para estimar los modelos por medio de Stata®, se establece la memoria necesaria y se carga la base de datos (*véase* figura 6.1).

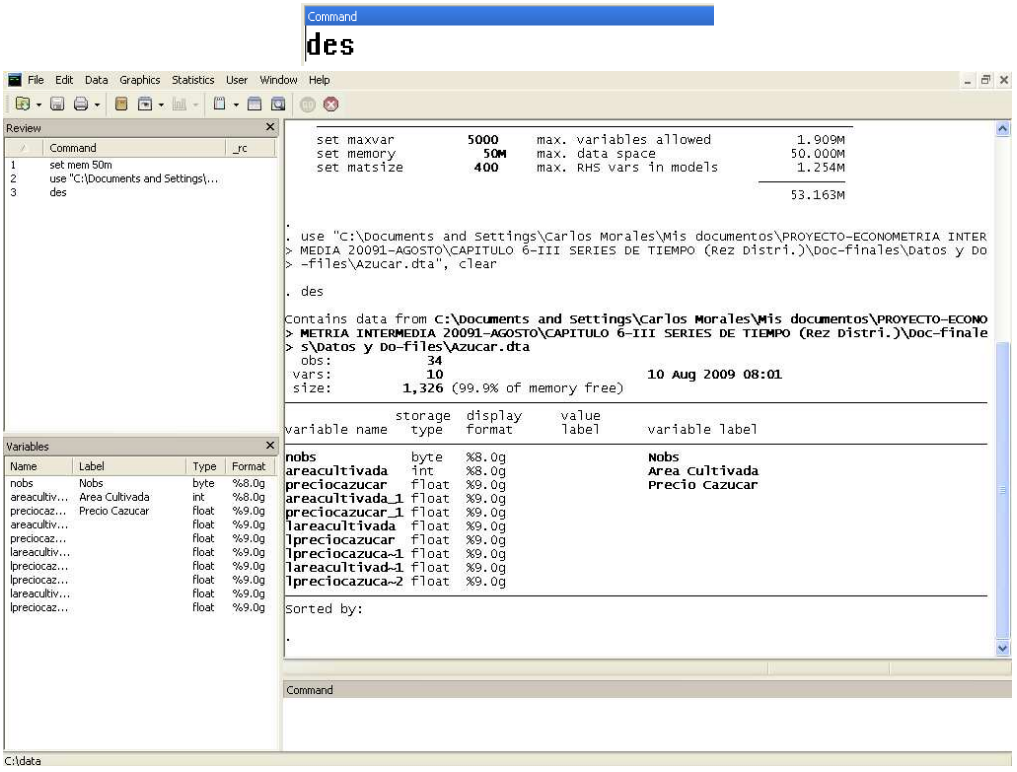
Figura 6.1. Salida comandos set mem y use.



Fuente: cálculos autores

2. Con las variables en memoria, a través del comando *des,* se puede determinar cuáles son las variables disponibles para las diferentes estimaciones en consideración (*véase* figura 6.2).

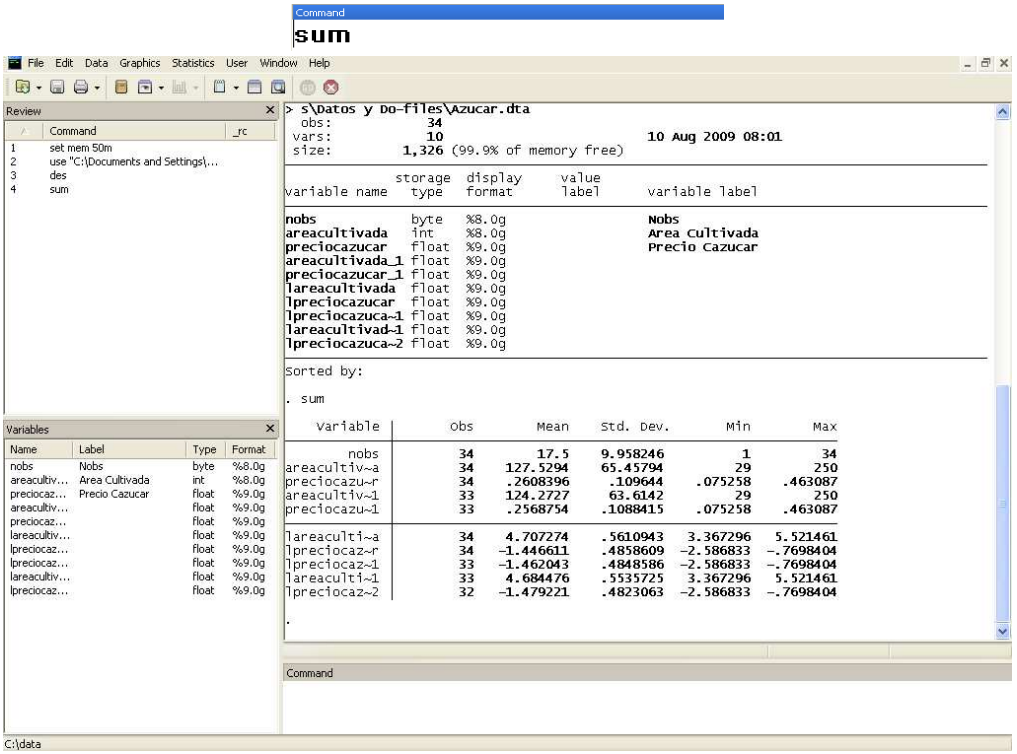
Figura 6.2. Salida comando des



Fuente: cálculos autores

3. Si se requiere hacer un análisis descriptivo de los datos, se utiliza el comando *sum* (*véase* figura 6.3).

Figura 6.3. Salida comando sum

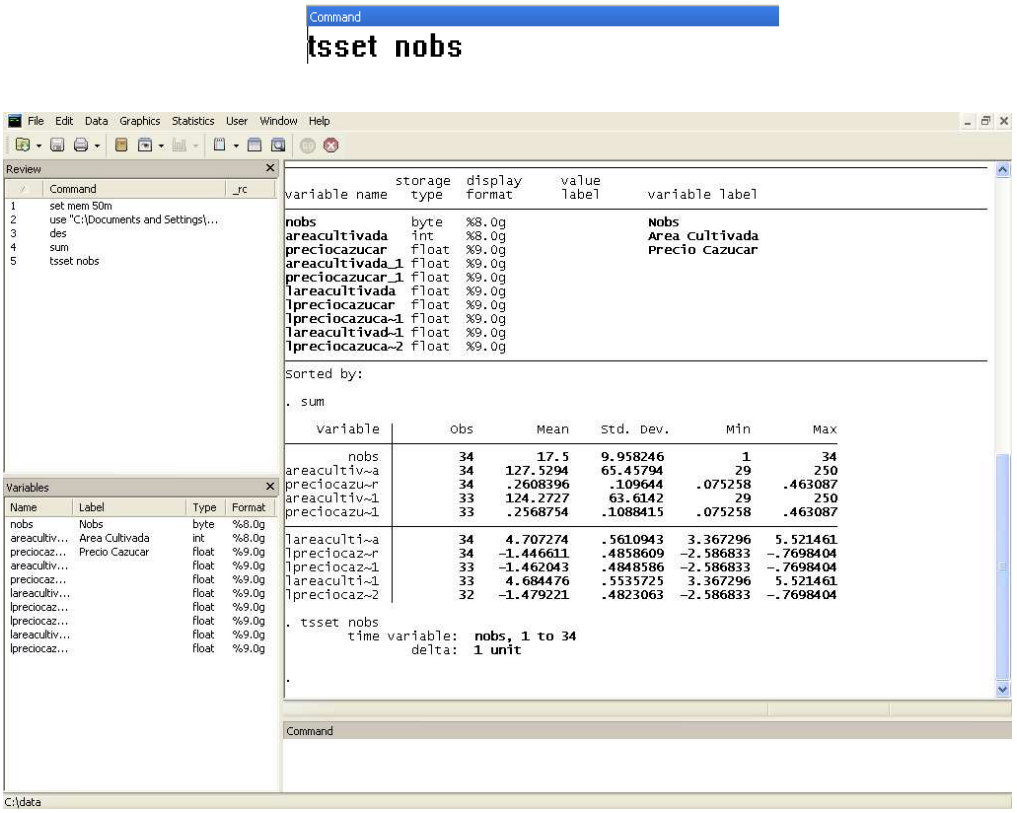


Fuente: cálculos autores

La figura 6.3 muestra que la base de datos de la producción del azúcar cuenta con 34 observaciones y 9 variables para utilizar en los diferentes modelos. El resumen de los datos básicos para estimar las ecuaciones del cuadro 6.1 se encuentran descritos en el cuadro 6.2.

4. Finalmente es necesario nombrar la base de datos como serie de tiempo para operaciones que se desarrollaran a lo largo del estudio de caso (*véase* figura 6.4)

Figura 6.4 Salida declaración datos como serie de tiempo.

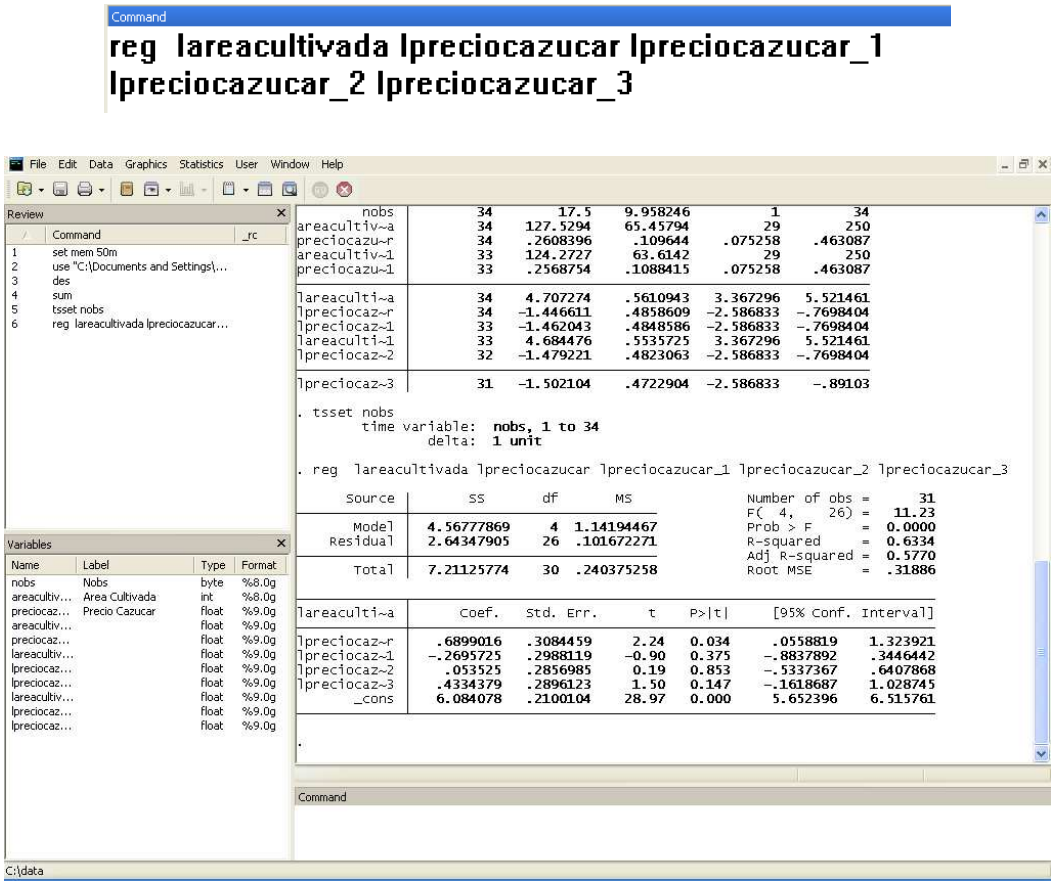


Fuente: cálculos autores

6.6.2 Estimación del modelos de rezagos distribuidos por medio de Koyck y Almon

1. Para estimar la primera ecuación del cuadro 6.1 en Stata®, se realiza a través de MCO con el comando *reg* (*véase* figura 6.5).

Figura 6.5. Salida estimación modelo rezagos distribuidos



Fuente: cálculos autores

En la figura 6.5 se muestra una regresión de modelo de rezagos distribuidos para la oferta de caña de azúcar. Se utilizaron tres rezagos del precio para establecer cuáles son los efectos de corto plazo. La variable contemporánea del precio resulta significativa al 5%, indicando que los cambios en precios afectan en el mismo periodo a la oferta de azúcar.

2. En primera instancia se utiliza el método de transformación de Koyck para ilustrar de forma alternativa el modelo de rezagos distribuidos de la figura 6.5. Siguiendo el procedimiento descrito en la sección 6.3.1 se transforma el modelo inicial (*véase* cuadro 6.1) a partir de la expresión $\beta_k = \beta_0 \lambda^k$, donde λ es el factor de ajuste en el tiempo y k el número de rezagos (*véase* ecuación 6.35).

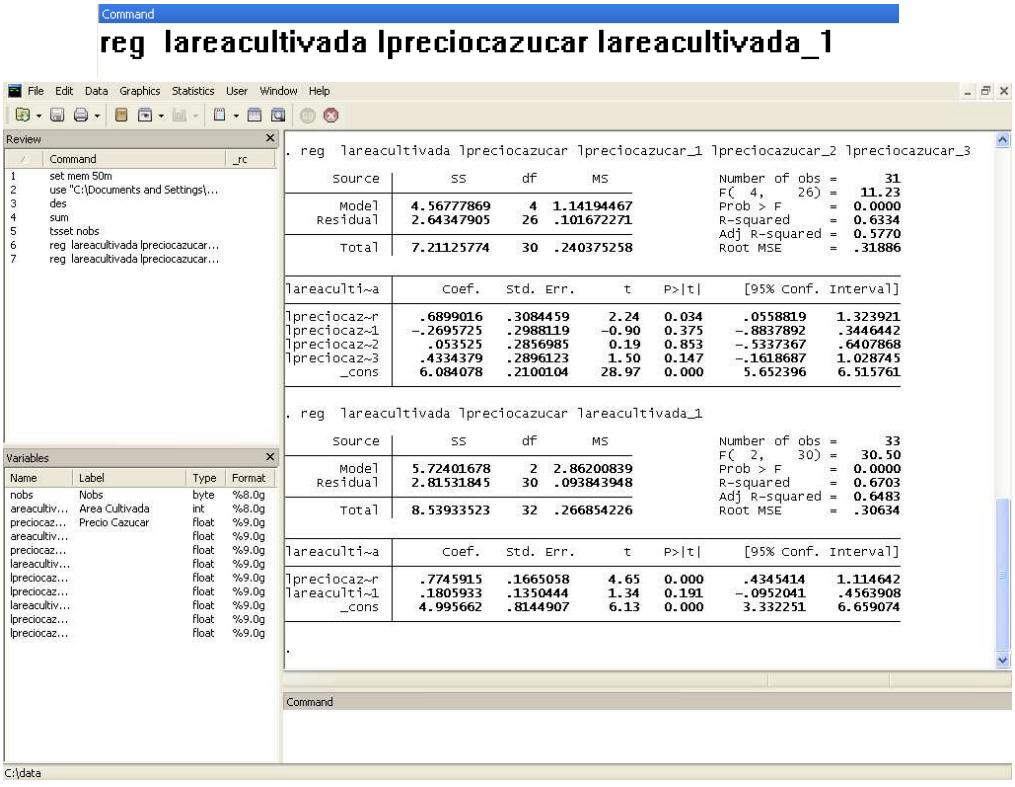
$$Q_t = \alpha + \beta_0 P_t + \beta_0 \lambda P_{t-1} + \beta_0 \lambda^2 P_{t-2} + \beta_0 \lambda^3 P_{t-3} + u_t \qquad (6.35)$$

Al finalizar el procedimiento de transformación (*véase* sección 6.3.1) se obtiene un modelo autorregresivo de orden uno (*véase* ecuación 6.36).

$$Q_t = \alpha(1-\lambda) + \beta_0 P_t + \lambda Q_{t-1} + e_t \quad \text{con} \quad e_t = u_t - \lambda u_{t-1} \quad (6.36)$$

De acuerdo a lo anterior, se estima la ecuación 6.36 con solo tres parámetros (en vez de uno de 5) por medio de MCO con el comando *reg* (véase figura 6.6).

Figura 6.6. Salida estimación método Koyck

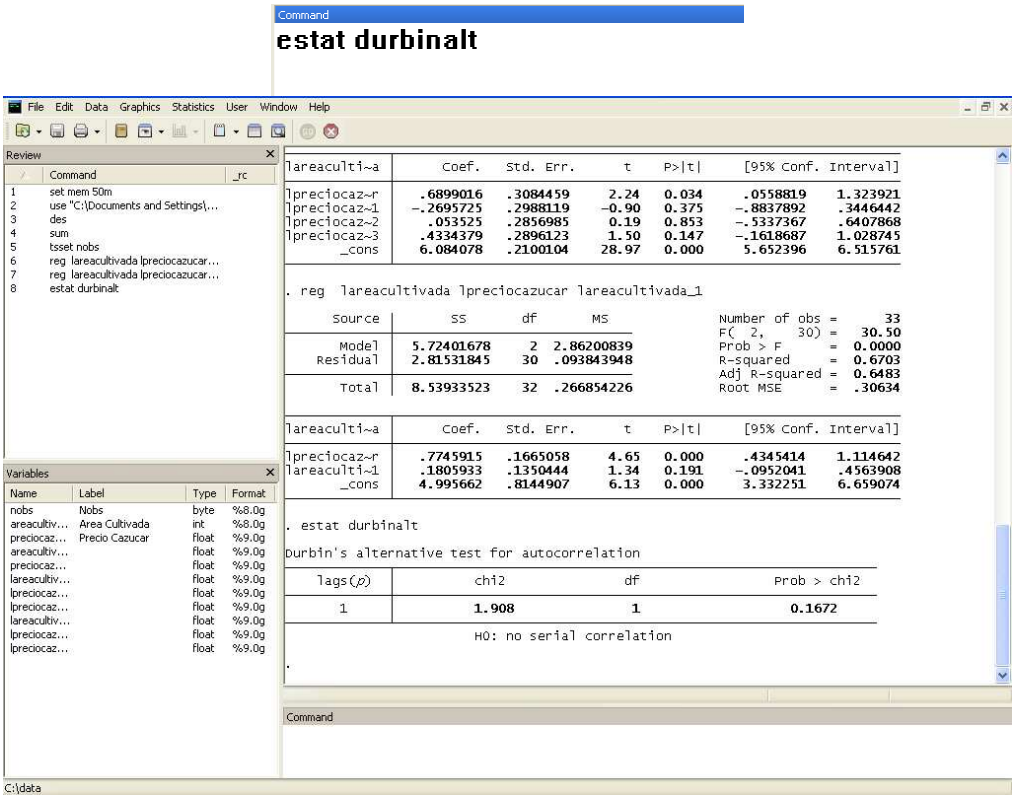


Fuente: cálculos autores

3. De acuerdo a la sección 6.3.1 los modelos autorregresivos pueden presentar problemas de autocorrelación. Con el fin de determinar si el modelo

presenta o no problemas de acutocorrelación, se lleva a cabo la prueba Durbin h basada en la estimación de la figura 6.6 (*véase* figura 6.7)

Figura 6.7. Salida prueba durbin h para autocorrelación



Fuente: cálculos autores

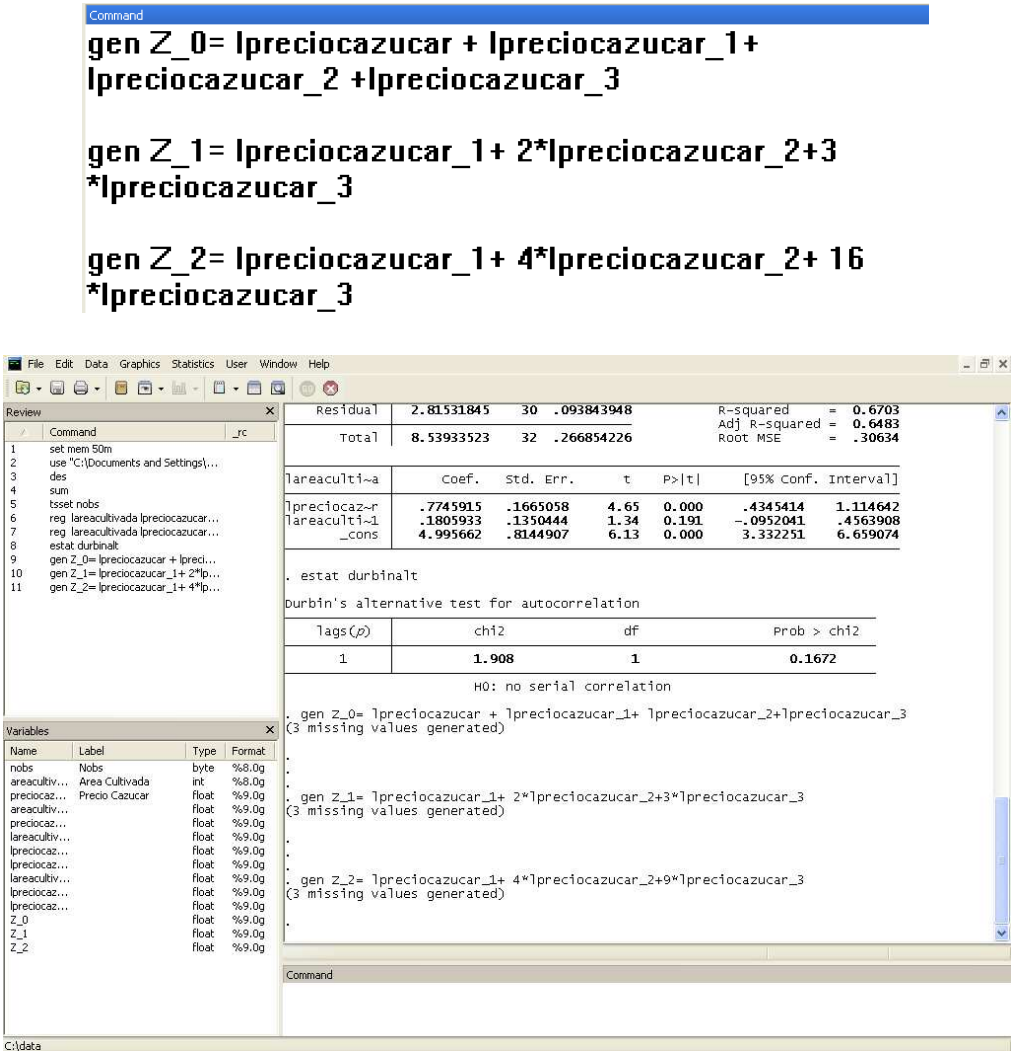
A través de la prueba de la figura 6.7, se concluye que no se puede rechazar la hipótesis nula que dice que no existe autocorrelación en el modelo. Con esto se concluye con seguridad que las estimaciones de la figura 6.6 están correctas.

4. De igual forma, es posible utilizar el método de Almon para estimar un modelo autorregresivo básico. Si se supone que el modelo inicial (*véase* cuadro 6.1) se ajusta a un polinomio de grado dos $\beta_i = a_0 + a_1 i + a_2 i^2$ por tanto dicho modelo se puede representar como:

$$Q_t = \alpha + a_0 Z_{t0} + a_1 Z_{t1} + a_2 Z_{t2} + e_t \quad (6.37)$$

En la ecuación 6.35, $Z_{t0} = \sum_{p=0}^{k} p_{t-p}$, $Z_{t1} = \sum_{s=0}^{k} sP_{t-s}$, $Z_{t2} = \sum_{s=0}^{k} s^2 P_{t-s}$. Empíricamente se deben generar las variables Z_{ti} de acuerdo a lo estipulado anteriormente (véase figura 6.8).

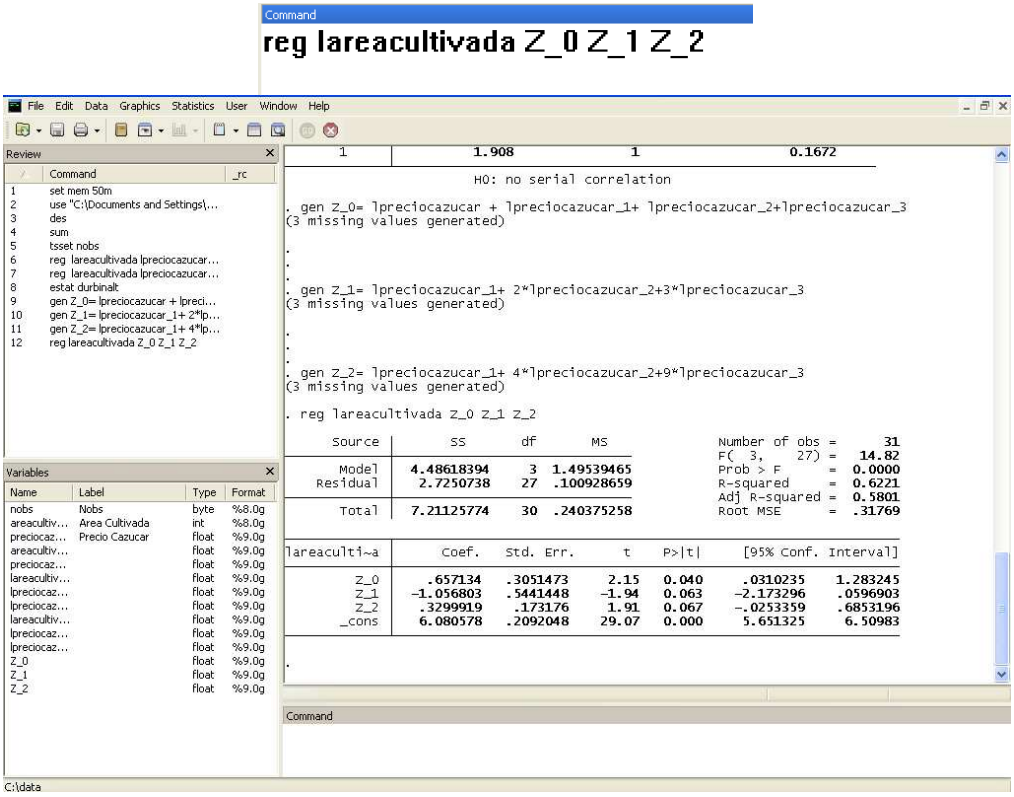
Figura 6.8. Salida variables Z generadas.



Fuente: cálculos autores

5. Una vez generadas las variables Z_{ti}, es posible estimar el modelo de Almon de la ecuación 6.37 por medio de MCO con el comando *reg* (véase figura 6.9)

Figura 6.9. Salida estimación modelo Almon.



Fuente: cálculos autores

Los coeficientes que aparecen en la salida 6.9 son utilizados para calcular los coeficientes β_i iniciales por medio del polinomio de grado dos ($\beta_i = a_0 + a_1 i + a_2 i^2$). $\hat{a}_0, \hat{a}_1, \hat{a}_2$ son respectivamente 0.5637, -0.5445, 0.0914. A partir de esto se tiene:

$$\hat{\beta}_0 = \hat{a}_0 = 0.56 \qquad (6.38)$$

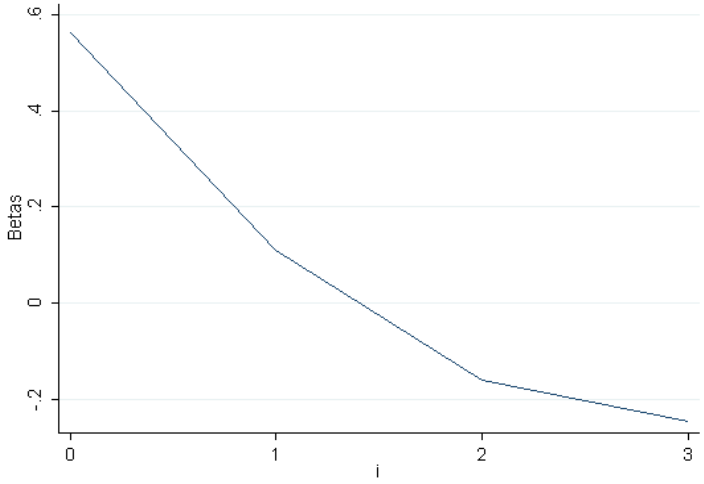
$$\hat{\beta}_1 = \hat{a}_0 + \hat{a}_1 + \hat{a}_2 = 0.1106 \qquad (6.39)$$

$$\hat{\beta}_2 = \hat{a}_0 + 2\hat{a}_1 + 4\hat{a}_2 = -0.1597 \qquad (6.40)$$

$$\hat{\beta}_3 = \hat{a}_0 + 3\hat{a}_1 + 9\hat{a}_2 = -0.2472 \qquad (6.41)$$

Estos estimadores (6.38, 6.39, 6.40 y 6.41) presentan los mismos signos de la estimación de rezagos distribuidos (*véase* figura 6.9). Asimismo es posible mostrar que el polinomio de grado dos utilizado para el método de Almon se ajusta bien al modelo de oferta de caña de azúcar (*véase* gráfica 6.2).

Gráfica 6.2 Ajuste modelo Almon



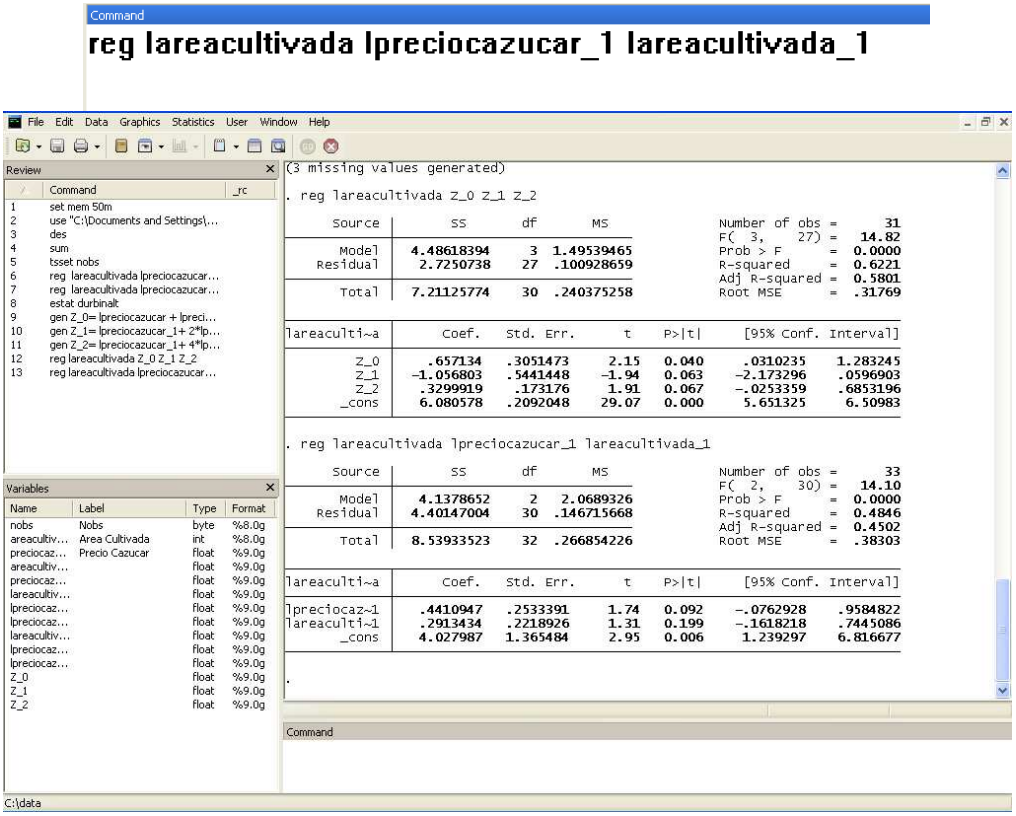
Fuente: cálculo autores

6.6.3 Estimación de expectativas adaptables

Para esta regresión se utiliza la segunda ecuación del cuadro 1, en la que se tiene en cuenta como variables explicativas el precio en el periodo $t-1$ y el primer rezago del área cultivada.

1. Se realiza la estimación del modelo de expectativas adaptables por medio del comando *reg (véase* ecuación 6.10)

Figura 6.10. Salida estimación modelo expectativas adaptables



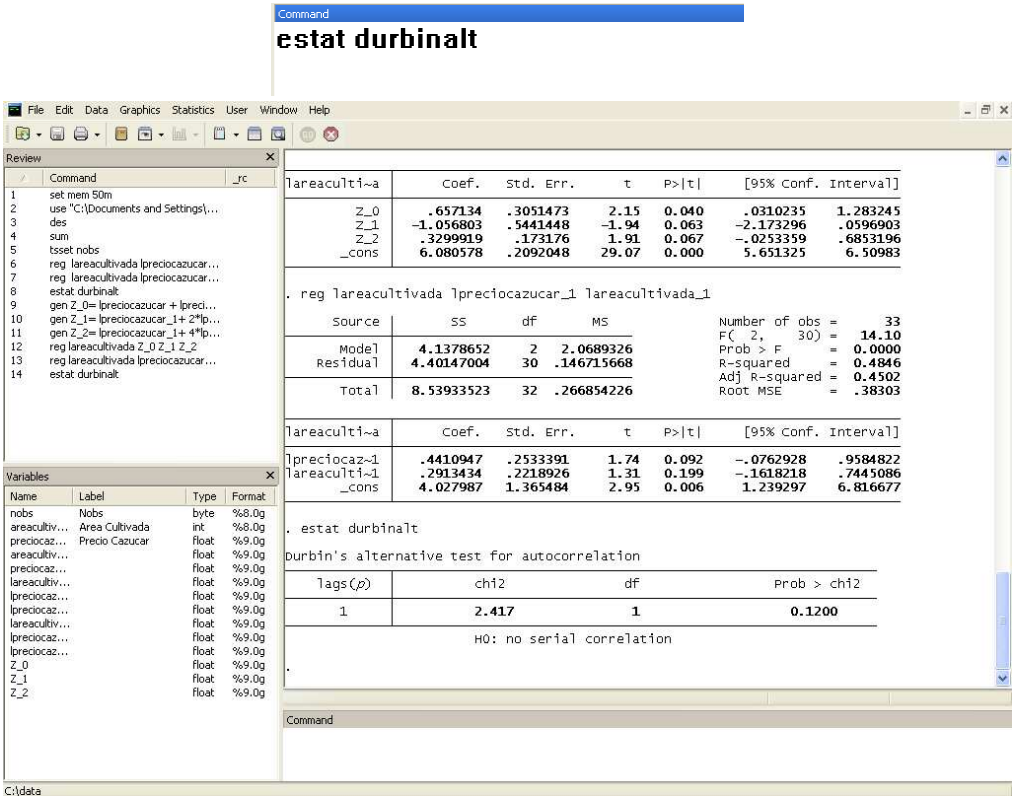
Fuente: cálculos autores

La figura 6.10 muestra que el precio en el periodo $t-1$ es positivo y significativa al 10%, validando la hipótesis de la oferta. Mientras el primer rezago del área cultivada no resulta significativo al 10%, indicando que el porcentaje de área cultivada no tiene en cuenta el periodo anterior.

De acuerdo a lo anterior, las elasticidades se derivan, preferiblemente del modelo especificado doblemente logarítmico, a partir de los estimadores de la figura 6.10. Considerando el cuadro 6.1, la elasticidad en el corto plazo es igual al coeficiente que acompaña a la variable del precio que tiene un valor de 0.4410 con lo que se dice que variaciones en el precio no afectan en gran medida la cantidad cultivada de caña de azúcar. Mientras que la de largo plazo no se puede derivar dado que no se tienen más rezagos en consideración.

2. Con el fin de determinar si el modelo presenta o no problemas de acutocorrelacion , se lleva a cabo la prueba Durbin h basada en la estimación de la figura 6.10 (*véase* figura 6.11).

Figura 6.11. Salida prueba durbin h para autocorrelación



Fuente: cálculos autores

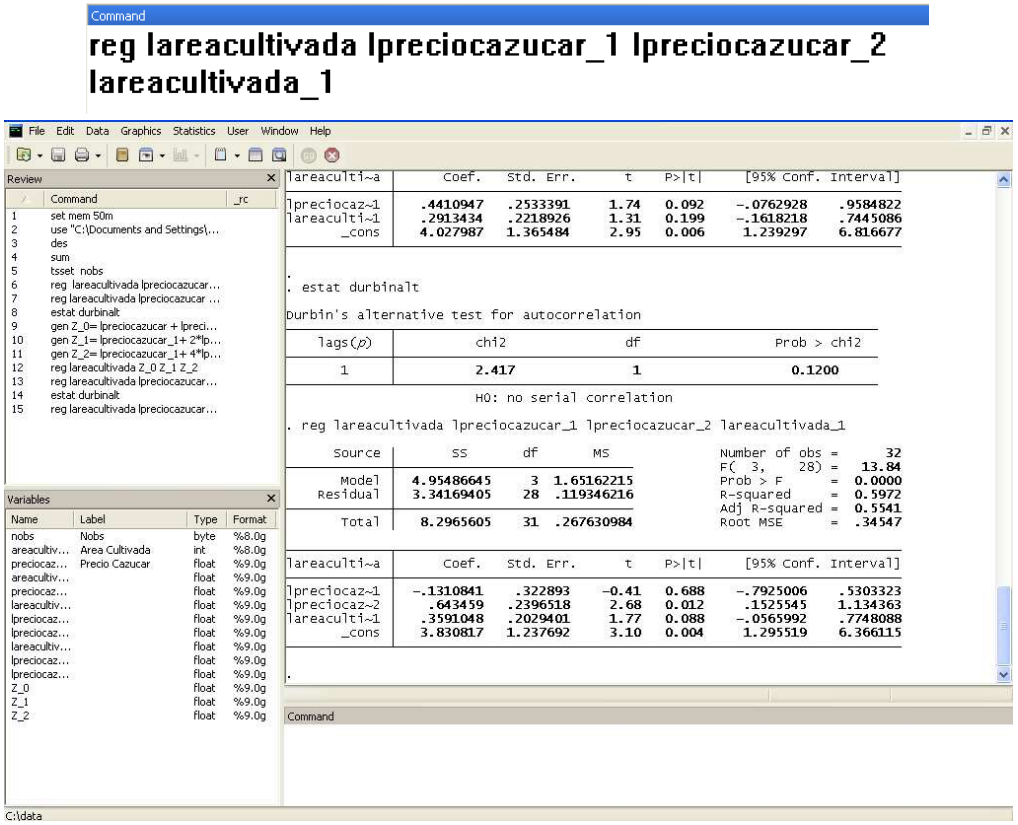
A través de la prueba de la figura 6.9, se concluye que no se puede rechazar la hipótesis nula que dice que no existe autocorrelación en el modelo. Con esto se concluye con seguridad que las estimaciones de la figura 6.10 están correctas.

6.6.4 Prueba para causalidad de Granger.

Una vez se han derivado los resultados de modelos autorregresivos y de rezagos distribuidos, ahora se desarrolla la prueba de Granger para establecer si precio del azúcar causa en el sentido de Granger al área cultivada (*véase* cuadro 6.1).

1. Siguiendo los pasos de la sección 6.5 la estimación del modelo restringido está dada por la figura 6.10, de allí se conoce el valor de SCE_R=4.4014.
2. Adicionalmente, se requiere el valor de SCE_{NR}. Para ello se realiza la estimación del modelo no restringido de la forma: $Q_t = \alpha + \beta_1 P_{t-1} + \beta_2 P_{t-2} + \lambda Q_{t-1} + u_t$ (véase figura 6.12).

Figura 6.12. Salida estimación modelo no restringido



Fuente: cálculos autores

De la figura 6.12 se deriva el valor de SCE_{NR}=3.3416. De acuerdo a este resultado se puede derivar la prueba de Granger.

3. Teniendo en cuenta SCE_R y SCE_{NR} se puede llevar a cabo la prueba de hipótesis a través del estadístico de prueba F. Si se utilizan los resultados de cada estimación, el estadístico de prueba calculado quedaría de la siguiente forma:

$$F_c = \frac{(4.4014 - 3.3416)/1}{3.3416/33} = 10.4723 \qquad (6.42)$$

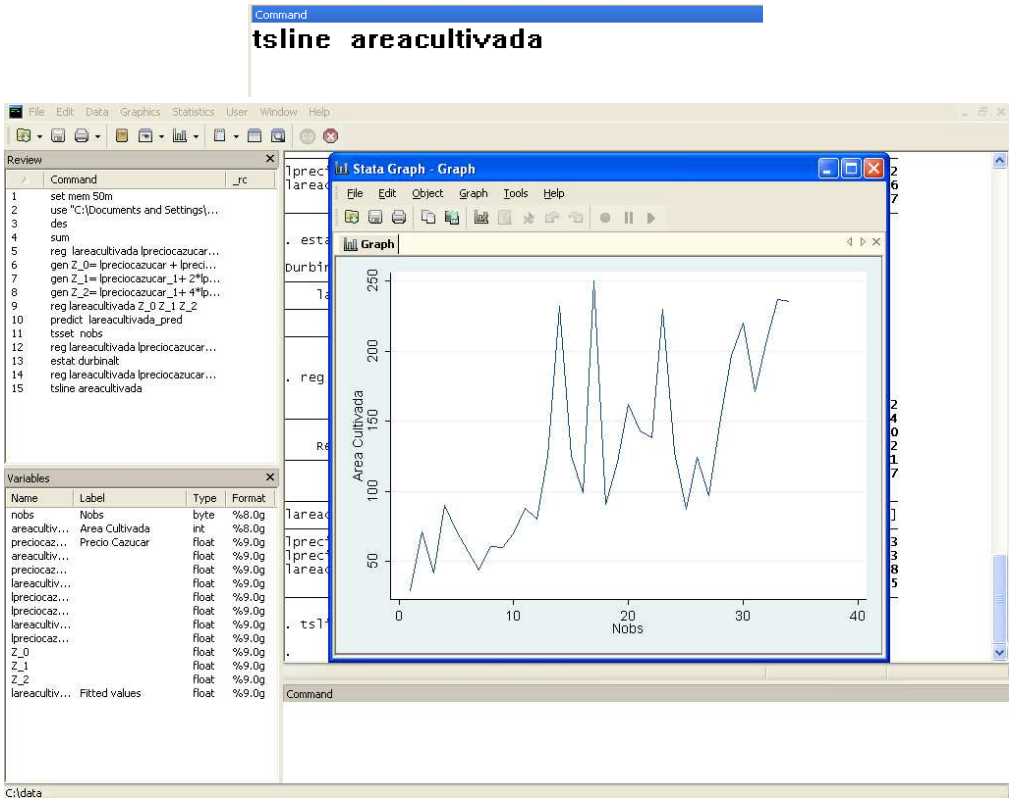
De acuerdo a la ecuación 6.42, y teniendo en cuenta los datos, $F_c = 10.4723$, mientras que $F_{1,33} = 4.17$, la conclusión de la prueba de causalidad es que precio causa en el sentido de Granger al área cultivada, puesto que $F_c > F_{r,n-k}$ y por tanto se puede rechazar la hipótesis nula.

6.6.5 Prueba para cointegración.

Finalmente, con el modelo dinámico se puede llevar a cabo una prueba de cointegración sobre las dos series de tiempo que se han trabajado en esta sección (Q_t y P_t, en niveles). Para continuar con lo descrito en la sección 6.5:

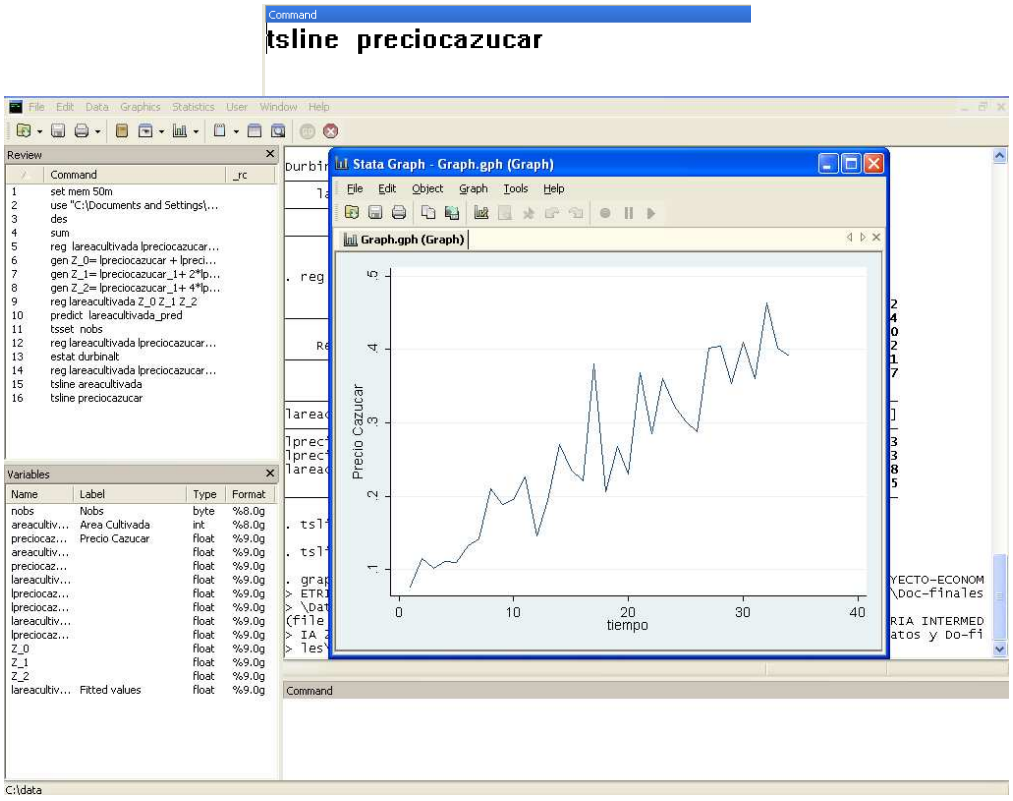
1. Se grafica por separado cada una de las series para determinar que no sean estacionarias y se pueda configurar una regresión lineal en la que el resultado sea estacionario. Para graficar series de tiempo se utiliza el comando *tsline* seguido de la variable de interés. (*véase* figura 6.13 y figura 6.14).

Figura 6.13. Salida gráfica área cultivada de caña de azúcar.



Fuente: cálculos autores

Figura 6.14. Salida gráfica precio caña de azúcar.



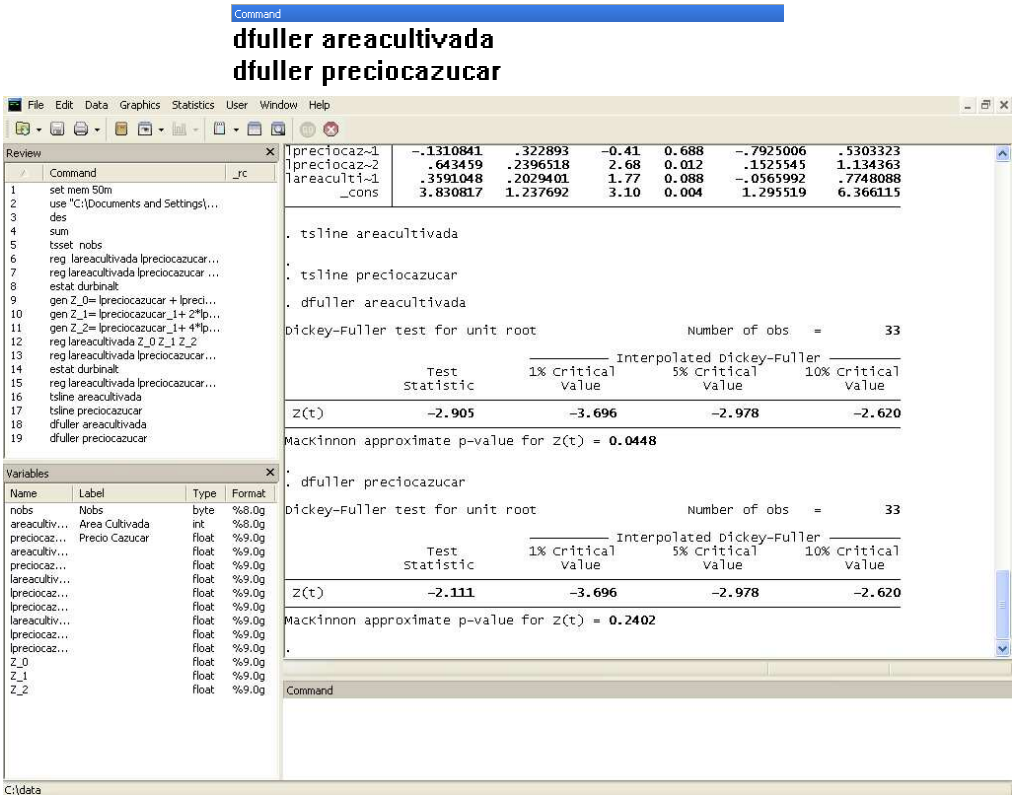
Fuente: cálculos autores

En las figuras 6.13 y 6.14 se evidencia la no estacionariedad de las series de área cultivada y precio de la caña de azúcar.

De acuerdo a esto, se puede llevar a cabo una combinación lineal entre estas dos series para permitir que los errores sean estacionarios. Si lo anterior se consigue, se dice que las series están cointegradas.

2. Se debe probar estadísticamente que cada una de las series sea I (1). Para ello se realiza un test de estacionariedad para verificar dicha característica por medio del comando *dfuller* seguido por el nombre de las variables a evaluar (*véase* figura 6.15).

Figura 6.15. Salida prueba Dickey-Fuller para areacultivada y preciocazucar



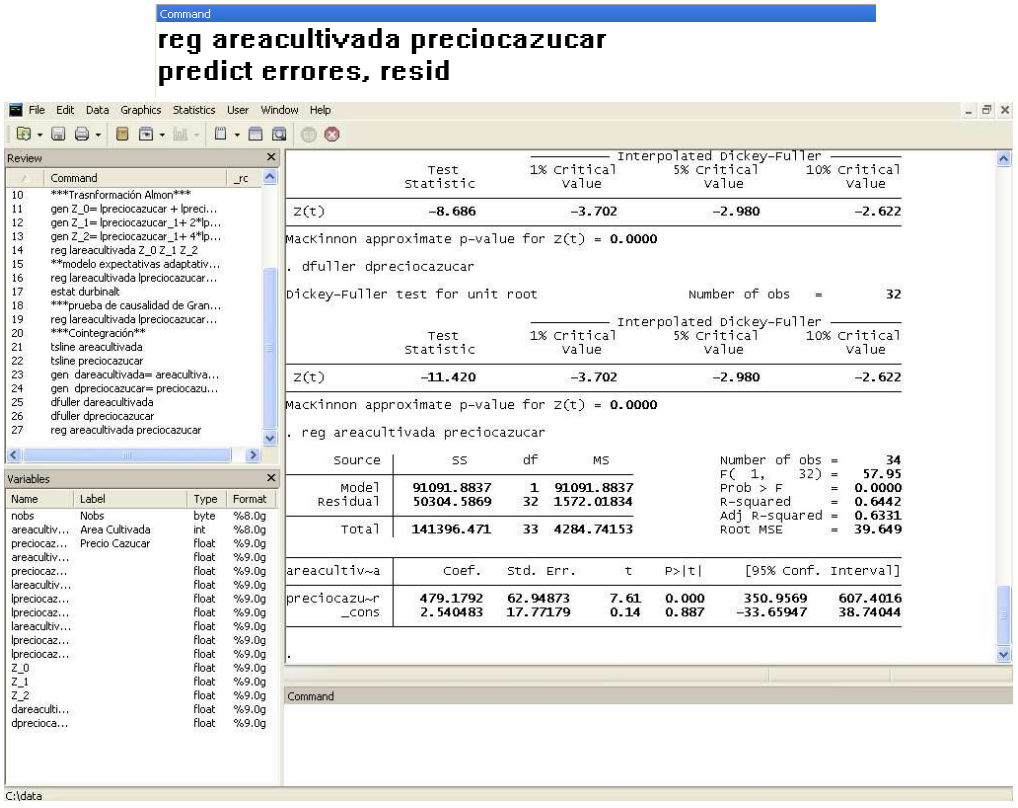
Fuente: cálculos autores

La figura 6.15 muestra la prueba de Dickey-Fuller para las dos series candidatas a cointegración. El *areacultivada* y *preciocazucarresulta* no estacionaria al 1% de significancia. Por tanto se dice que por lo menos se debe sacar primera diferencia (I(1)) para que las serie resulten estacionarias[136].

3. De acuerdo a lo anterior, es posible realizar una regresión lineal entre el área cultivada y el precio para derivar errores estacionarios. Esto se lleva a cabo con el comando *reg.* (*véase* figura 6.16).

[136] El lector puede rectificar que las series trabajas son I (1), t sacando la primera diferencia a cada serie de tiempo y realizando la misma prueba de estacionariedad.

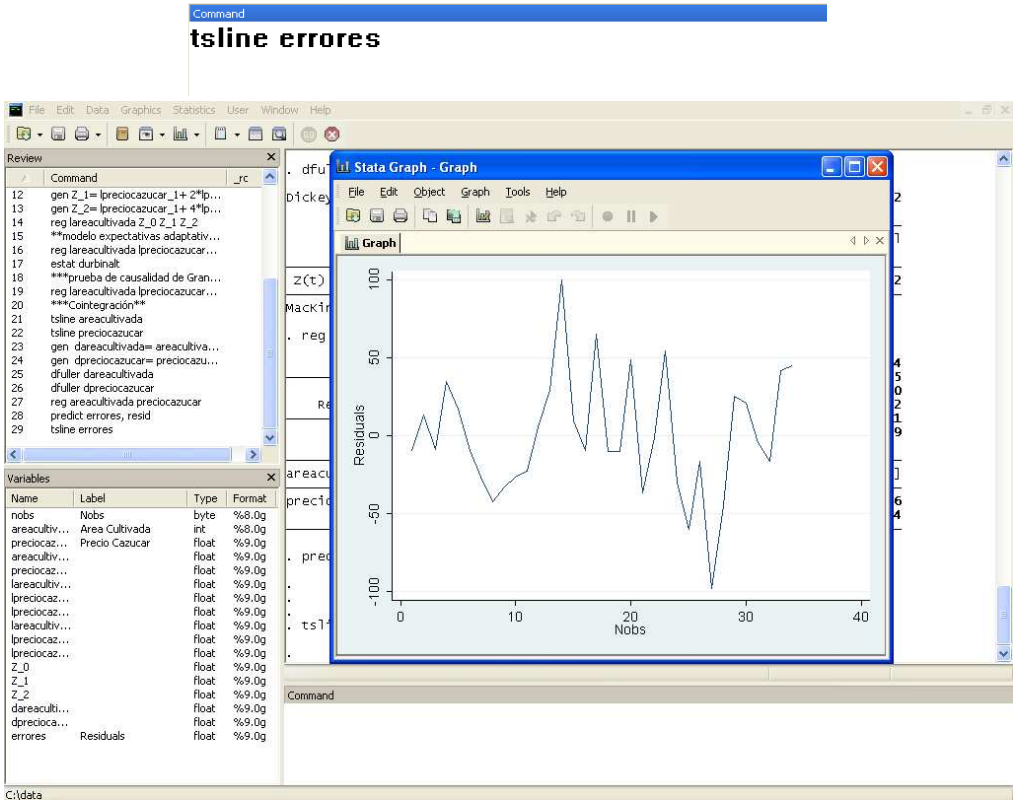
Figura 6.16. Salida gráfica precio caña de azúcar.



Fuente: cálculos autores

4. Luego de la estimación y predicción de los errores, se prueba que éstos últimos tengan un comportamiento estacionario. Para ello se grafican los errores predichos con el comando *tsline* (*véase* figura 6.17).

Figura 6.17. Salida gráfica de los errores predichos

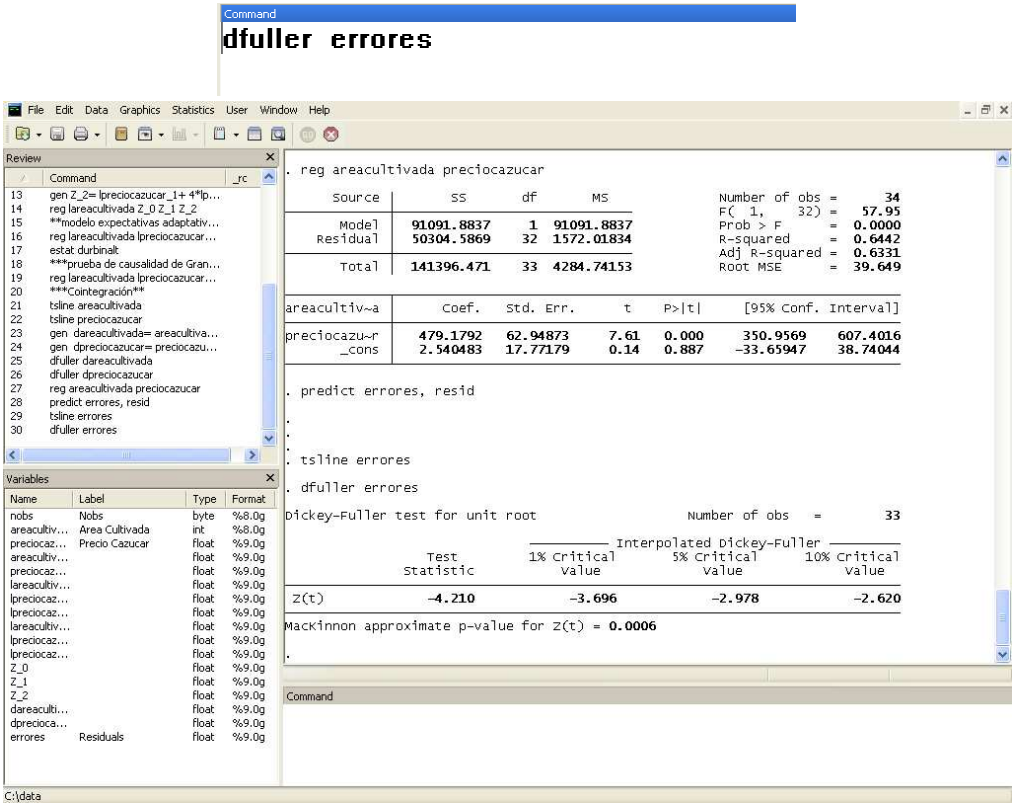


Fuente: cálculos autores

De acuerdo a lo descrito anteriormente, se encontró que los errores predichos de la regresión lineal son estacionarios gráficamente. Con esto se intuye que las series Q_t y P_t están cointegradas.

5. Finalmente para validar las conclusiones, se realiza una prueba de Dickey-Fuller a través del comando *dfuller* seguido de la variable de interés (*véase* figura 6.18).

Figura 6.18. Salida prueba de estacionariedad Dickey-Fuller



Fuente: cálculos autores

La figura 6.18 muestra las intuiciones planteadas en los numerales anteriores, esto es que los errores son estacionarios dado que mediante la prueba Dickey-Fuller se rechaza la hipótesis nula de no estacionariedad con u p-valor inferior al nivel de significancia de 5%.

Resumen

- Los modelos econométricos que integran rezagos en las variables explicativas, se conocen como modelos de regresión dinámica Existen dos tipos de modelos con variables rezagadas: Los modelos de rezagos distribuidos y los modelos autorregresivos. En el primero, los valores actuales y rezagados de los regresores son variables explicativas. En el segundo, los valores rezagados de la variable dependiente aparecen como variables explicativas.
- La escogencia del número de rezagos que influye en la estimación o la correlación entre los rezagos, puede ocasionar inconvenientes en las conclusiones que se deriven de una estimación. Las transformaciones de Koyck y Almon se presentan como una herramienta útil para solucionar las dificultades mencionadas.
- La transformación de Koyck es un método alternativo para estimar los modelos de rezagos distribuidos infinitos que impone a priori condiciones sobre los coeficientes β_i. Bajo esta transformación se encuentra un modelo de tipo autorregresivo de primer orden.
- El método de Almon surge como una alternativa al modelo de Koyck, puesto que este ultimo puede presentar problemas de autocorrelación con lo que las estimaciones pueden presentar inconsistencias. Almon supone que β_i puede ser aproximado mediante a un polinomio.
- Los modelos autorregresivos de rezagos distribuidos son una ampliación de los procesos autorregresivos simples. En estos modelos se utilizan dos series de tiempo, al tiempo que incluye uno a más rezagos de la variable dependiente entre sus variables explicativas.
- Los modelos autorregesivos tienden a presentar problemas de correlación serial por la existencia de la variable dependiente rezagada como variable explicativa del modelo. Para detectar estos inconvenientes es indispensable seguir una prueba de autocorrelación. La prueba h de Durbin es una alterativa para probar autocorrelación, siempre y cuando exista una muestra grande.
- La causalidad de Granger es una prueba que consiste en determinar si las observaciones pasadas de una variable de series de tiempo permiten

pronosticar a otra. Ésta indica, de acuerdo a los datos, si una variable causa a otra. Asimismo sirve para establecer si existe exogeneidad en el modelo.

- Cointegración significa que a pesar de que un conjunto de series no sean estacionarias individualmente, una combinación lineal entre ellas puede ser estacionaria.
- Por medio de la cointegración se puede aprovechar la relación entre dos series integradas para encontrar relaciones de corto plazo. De esta forma se pueden analizar políticas económicas, al tiempo que se logran hacer proyecciones.

Capítulo 7

Modelos para datos de corte transversal agrupados en el tiempo y estimador diferencia en diferencia.

7.1 Introducción

En capítulos anteriores se trataron metodologías que basan sus procesos en muestras de tipo corte transversal o series de tiempo. En primer caso, los métodos simples de MCO y sus derivados, sirven para muestras recolectadas en un mismo momento del tiempo, es decir, una "foto" a la realidad, por lo que las variables son estáticas. Segundo, las series de tiempo muestran el comportamiento de una variable en un periodo de tiempo, dejando a un lado las relaciones con otras variables. En cambio este capítulo se enfocara en la explicación de metodologías que combinan los dos tipos de muestras anteriores, pasando de una "foto" a un "video" que muestra la dinámica de las variables en el tiempo, por medio de la agrupación de datos de corte transversal a través del tiempo.

Teniendo en cuenta lo anterior, es importante señalar que la utilización de datos de corte transversal a lo largo del tiempo ofrece mayores beneficios que los de corte transversal simple. En primer lugar, el incremento en el tamaño de la muestra permite obtener estimadores más precisos (consistentes) y estadísticos de prueba más confiables. Una mayor cantidad de datos implica además más variabilidad entre ellos, menor colinealidad entre las variables, más grados de libertad y mayor eficiencia en las estimaciones (Hsiao, 2002). En segundo lugar, esta metodología permite investigar si las relaciones entre las variables han cambiado con el paso del tiempo, por medio de la prueba de Chow.

Partiendo de las características presentadas, al final del capítulo se expondrán las metodologías a través de un estudio de caso basado en la información del artículo de Rodríguez, Sánchez y Armenta (2007), titulado "*Hacia una mejor educación rural:*

Impacto de un programa de intervención a las escuelas en Colombia", cuyo objetivo es evaluar el impacto que tuvo el Programa de Educación Rural (PER) en las tasas de eficiencia y calidad de la educación en los centros educativos públicos donde se aplicó.

7.2 Unión de corte transversal y series de tiempo

Este capítulo estudia la metodología econométrica que combina los procedimientos con muestras de corte transversal y series de tiempo, con el fin de estudiar características antes desapercibidas del tiempo sobre las relaciones entre las variables explicativas y explicadas. A partir de lo anterior, y para el resto del capítulo, se tendrán en cuenta dos dimensiones, una que identifica a la unidad de corte transversal (i)[137] y otra para el tiempo (t). Por tanto, se requerirán nuevos procedimientos para evidenciar las características particulares que ofrece este tipo de conformación de datos (Dougherty, 2007, 1).

Antes de llevar a cabo una caracterización formal de estos modelos, hay que hacer alusión a la naturaleza de las muestras conformadas de esta forma. De acuerdo a ello, existen dos tipos de agrupación:

1. Las muestras que agrupan datos en el tiempo, es decir, muestras aleatorias de corte transversal para la misma población en diferentes periodos del tiempo, pero no necesariamente tienen en cuenta la misma muestra en cada uno de ellos.

2. Los paneles que agrupan las mismas unidades de corte transversal en diferentes periodos del tiempo, o sea, se agrupan datos de la misma muestra de corte transversal en distintos momentos del tiempo (*véase* capítulo 8).

La primera caracterización o agrupación, es el interés de este capítulo; y es suma importancia para las evaluaciones de cambio estructural o evaluaciones de impacto para políticas económicas realizadas en un momento determinado, y que son primordiales para áreas de estudio económicas como la evaluación social de proyectos. Para entender este tipo de conformación de los datos, es pertinente

[137] Individuos, hogares, municipios, ciudades, departamentos, países, etc.

tener en cuenta un nuevo modelo teórico, sobre el cual se explicaran las características particulares de este tipo de agrupación (*véase* ecuación 7.1).

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + u_{it} \qquad (7.1)$$

En la ecuación 7.1, Y_{it} es la variable dependiente del modelo, X_{it} y K_{it} son las variables independientes y u_{it} es el término del error. Adicionalmente es importante señalar que i hace referencia a las unidades de corte transversal de la muestra, (1 a n); t representa el periodo en el que se encuentra expresado el conjunto de variables para cada unidad de corte transversal i y dentro de un rango finito (1 a T). Una vez se tiene presente la estructura del nuevo método, a continuación se discutirán las particularidades del mismo haciendo énfasis en los beneficios de su utilización y el procedimiento práctico que se sigue para derivar los resultados de interés.

7.3 Corte transversal a lo largo del tiempo.

Las muestras de corte transversal a lo largo del tiempo es la primera aproximación a las metodologías con paneles de datos. Ésta, permite estudiar relaciones entre las variables en distintos periodos del tiempo, facilitando la revisión de efectos originados por choques exógenos sobre una o varias variables del modelo econométrico de interés. Como se mencionó anteriormente, estos beneficios no estaban presentes en los métodos de corte transversal, puesto que las muestras son tomadas en un determinado periodo de tiempo y las relaciones entre variables son estudiadas bajo ese contexto, sin permitir comparaciones temporales.

Teniendo claro lo anterior, es importante entender cómo se estima un conjunto de variables recolectas en el tiempo. Esta agrupación de datos permite establecer con mayor claridad cuáles son los tipos de relación que se pueden generar entre las variables independientes y la explicada (Wooldridge, 2009, 445). La estimación se realiza a través de mínimos cuadrados agrupados (MCA) que, simplemente, traduce el procedimiento de MCO a un conjunto de variables evaluadas en distintos periodos del tiempo.

7.3.1 Introducción a mínimos cuadrados agrupados (MCA)

Mínimos cuadrados agrupados (MCA o pooled OLS en inglés) es una metodología econométrica que sirve para investigar si las relaciones entre las variables (explicativas y explicadas) han cambiado con el paso del tiempo. Consiste en realizar estimaciones usando todo el conjunto de datos, sin hacer ninguna distinción entre grupos. Partiendo de la expresión de la ecuación 7.1, y para reconocer el funcionamiento del mecanismo de estimación, se utiliza el enfoque matricial (*véase* ecuación 7.2)

$$Y_i = Z_i\beta + \varepsilon_i \text{, donde } Z_i = [X_{it} K_{it}] \quad (7.2)$$

En la ecuación 7.2, Y_i es el vector de la variable dependiente; Z_i una matriz de variables explicativas y ε_i es el vector de errores del modelo. Ahora bien, para obtener los estimadores, se debe estimar la ecuación 7.2 a través de MCO. Estos resultan consistentes y eficientes como consecuencia del incremento del tamaño de la muestra, respecto a muestras de corte transversal. Por tanto los estimadores estarían representados en la siguiente expresión:

$$\hat{\beta}_{MCA} = (Z_i{}' Z_i)^{-1}(Z_i{}' Y_i) \quad (7.3)$$

De acuerdo a la expresión 7.3, se puede establecer con mayor precisión si las relaciones evaluadas en un modelo de interés varían como consecuencia del paso del tiempo. Para conseguir conclusiones al respecto, se deben llevar a cabo pruebas de cambio estructural de Chow.

7.3.2 Prueba de Cambio Estructural de Chow

Teniendo en cuenta la ecuación 7.2 y asumiendo que $t = 1, 2$, el modelo estructural puede desglosarse en dos grupos, correspondientes a dos momentos del tiempo. De esta forma se puede evaluar qué pasa en cada año con las variables de interés (*véase* ecuaciones 7.4 y 7.5).

$$\text{Grupo 1:} \qquad Y_{i1} = \beta_0 + \beta_1 X_{i1} + \beta_2 K_{i1} + u_{i1} \quad (7.4)$$

Grupo 2: $$Y_{i2} = \beta_0 + \beta_1 X_{i2} + \beta_2 K_{i2} + u_{i2} \quad (7.5)$$

Al estimar las ecuaciones 7.4 y 7.5 a través de MCO en forma agrupada, y por separado, lo que se quiere ver es si existe algún efecto del tiempo sobre las variables. Esto se consigue comparando los estimadores del grupo 1 y 2. Si se encuentra una diferencia estadística entre ellos, es porque existe cambio estructural. En el caso en que los estimadores sean los mismos, se dice que las relaciones de las variables no han cambiado en el tiempo.

Reconociendo lo anterior, es importante evaluar, en el caso en que intuitivamente se detecte cambio estructural, la fuente de variación de las relaciones en el tiempo. La prueba de Chow permite establecer cuál es la causa del cambio estructural, esto es: cambio en el intercepto, en pendiente, o una combinación de las dos anteriores. Cada uno de estos casos se discuten a continuación.

7.3.2.1 Cambio en intercepto

La primera causa de cambio estructural se debe a cambio en el intercepto, es decir, que el paso del tiempo permitió que las variables se desplazaran de manera proporcional en los periodos. Para reconocer este efecto, se transforma el modelo descrito por la ecuación 7.1, y se añade una variable dummy que separe la muestra en dos periodos. Esta variable se denomina D_2 y toma el valor de uno si $t = 2$, y cero si $t = 1$. Bajo esta especificación, el modelo quedaría expresado de la siguiente forma:

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + \delta_1 D_2 + u_{it} \quad (7.6)$$

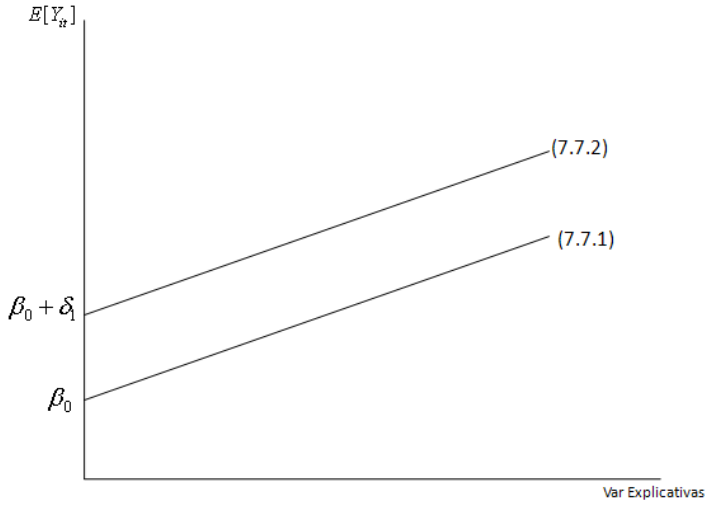
Manteniendo las mismas variables iniciales, los valores esperados para Y_{it} están determinados por los periodos en los que están agrupados los datos (*véase* ecuaciones 7.7.1 y 7.7.2). Estos definen el pronóstico de la variable dependiente dado el cambio en el tiempo y manteniendo constantes las variables explicativas.

Si t=1: $$E[Y_{i1} \mid X_{i1}, K_{i1}, D_2 = 0] = \beta_0 + \beta_1 X_{i1} + \beta_2 K_{i1} \quad (7.7.1)$$

Si t=2: $$E[Y_{i2} \mid X_{i2}, K_{i2}, D_2 = 1] = (\beta_0 + \delta_1) + \beta_1 X_{i2} + \beta_2 K_{i2} \quad (7.7.2)$$

A través de las ecuaciones 7.7.1 y 7.7.2 se puede identificar la diferencia en los estimadores de cada ecuación. Por ejemplo, la ecuación 7.7.1 muestra teóricamente tres estimadores ($\beta_0, \beta_1, \beta_2$), mientras que la ecuación 7.7.2 solo dos (β_1, β_2)iguales a 7.7.1 y uno distinto ($\beta_0 + \delta_1$). La diferencia radica en el coeficiente que acompaña a la variable D_2 y que hace referencia al cambio de periodo en la muestra. Si δ_1 resulta significativo estadísticamente, se dice que hay un cambio estructural debido al cambio en intercepto, es decir, un desplazamiento positivo (o negativo) de la curva referente al valor esperado de la variable dependiente Y_{it} (*véase* gráfica 7.1).

Gráfica 7.1 Cambio en intercepto del modelo estructural



Fuente: los autores

Por lo tanto, se debe probar la significancia individual del estimador δ_1 en la ecuación 7.6, por medio de una prueba de hipótesis como la siguiente:

$$H_0 : \delta_1 = 0 \quad \text{No existe cambio en intercepto} \qquad (7.8)$$

$$H_1 : \delta_1 \neq 0 \quad \text{Existe cambio en intercepto} \qquad (7.9)$$

Para encontrar respuesta a la hipótesis de la expresión 7.8, se utiliza un estadístico t que permite verificar la significancia individual del coeficiente que acompaña a

la dummy de tiempo. Si $|t_c| > |t_{\alpha/2,n-k-1}|$, entonces se rechaza H_0, asegurando que δ_1 es significativo. Este resultado lleva a concluir que hubo un cambio estructural en el modelo y que el paso del tiempo modificó los efectos de las variables del modelo pero no las relaciones entre ellas.

7.3.2.2 Cambio en pendiente

No solamente puede suceder que existe un cambio en intercepto, también pueden existir cambios de pendiente como consecuencia de introducir la interacción de D_2 con una de las variables explicativas, por ejemplo X_{it}, como variable explicativa para el modelo de la ecuación 7.1. A través de la nueva variable se quiere verificar si la relación entre alguna variable explicativa y la dependiente cambia en el tiempo (*véase* ecuación 7.10).

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + \delta_2 X_{it} \cdot D_2 + u_{it} \qquad (7.10)$$

De acuerdo al nuevo modelo de la ecuación 7.10, los valores esperados para Y_{it} dado $t = 1, 2$ (*véase* ecuaciones 7.11.1 y 7.11.2), adicionando la interacción $X_{it} \cdot D_2$ y suponiendo que las variables explicativas se mantienen constantes, muestra cómo cambian las relaciones entre la variable explicativa (X_{it}) y la dependiente (Y_{it}) a través de dos periodos.

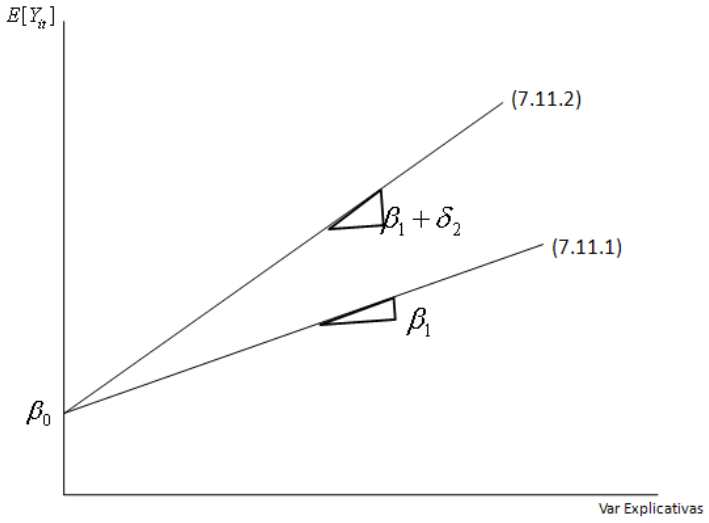
$$\text{Si t=1: } E[Y_{i1} \mid X_{i1}, K_{i1}, D_2 = 0] = \beta_0 + \beta_1 X_{i1} + \beta_2 K_{i1} \qquad (7.11.1)$$

$$\text{Si t=2: } E[Y_{i2} \mid X_{i2}, K_{i2}, D_2 = 1] = \beta_0 + (\beta_1 + \delta_2) X_{i2} + \beta_2 K_{i2} \qquad (7.11.2)$$

Así como en las ecuaciones 7.7.1 y 7.7.2, las expresiones 7.11.1 y 7.11.2 identifican si las relaciones entre las variables del sistema varían como consecuencia del tiempo. Lo anterior se consigue comparando los estimadores correspondientes a cada año de la muestra. La ecuación 7.11.1 muestra teóricamente tres estimadores ($\beta_0, \beta_1, \beta_2$), mientras que la ecuación 7.11.2 muestra dos estimadores (β_0, β_2) iguales y uno distinto ($\beta_1 + \delta_2$) respecto a 7.11.1. La diferencia está en la existencia del coeficiente que acompaña a la variable $X_{it} \cdot D_2$. Si δ_2 resulta significativo estadísticamente, se dice que hay un cambio estructural debido al cambio en

pendiente, es decir, genera un movimiento positivo (o negativo) de la curva que hace referencia al valor esperado de la variable dependiente Y_{it} (*véase* gráfica 7.2).

Gráfica 7.2 Cambio en pendiente en modelo estructural



Fuente: los autores

Como en la prueba de cambio en intercepto, se quiere verificar si estadísticamente δ_2 es igual a cero o no. Por tanto se requiere de una prueba de significancia individual que lleve a concluir la existencia de cambio estructural (*véase* prueba de hipótesis).

$$H_0 : \delta_2 = 0 \quad \text{No hay cambio en pendiente} \qquad (7.12)$$

$$H_1 : \delta_2 \neq 0 \quad \text{Existe cambio en pendiente} \qquad (7.13)$$

Para encontrar la conclusión correcta, se utiliza un estadístico t para verificar la significancia individual del coeficiente que acompaña a la variable que relaciona a X_{it} con D_2, en la ecuación 7.10. Si $|t_c| > |t_{\alpha/2,n-k-1}|$ se dice que se rechaza H_0 y que δ_2 es significativo. Por tanto, existe un cambio estructural. Para este caso la relación entre Y_{it} y X_{it} cambia con el paso del tiempo.

7.3.2.3 Cambio en intercepto y pendiente

Esta prueba está caracterizada por mezclar las dos pruebas descritas anteriormente. El nuevo modelo contiene la dummy de tiempo (D_2) y la interacción de ésta con una variable explicativa ($X_{it} \cdot D_2$). Con esta especificación se quiere probar si la relación entre la variable explicada y una independiente cambia en el tiempo, asimismo si el cambio de periodo tiene algún efecto sobre el modelo a estimar. Por lo tanto, el nuevo modelo estaría dado por la ecuación 7.14.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + \delta_1 D_2 + \delta_2 X_{it} \cdot D_2 + u_{it} \qquad (7.14)$$

Teniendo en cuenta la ecuación 7.14, los valores esperados para Y_{it} dado $t = 1, 2$ (*véase* ecuaciones 7.15.1 y 7.15.2) incluyendo dos nuevas variables (D_2 y $X_{it} \cdot D_2$) y manteniendo constantes las variables explicativas, refleja el comportamiento de las variables en dos periodos de tiempo distinto.

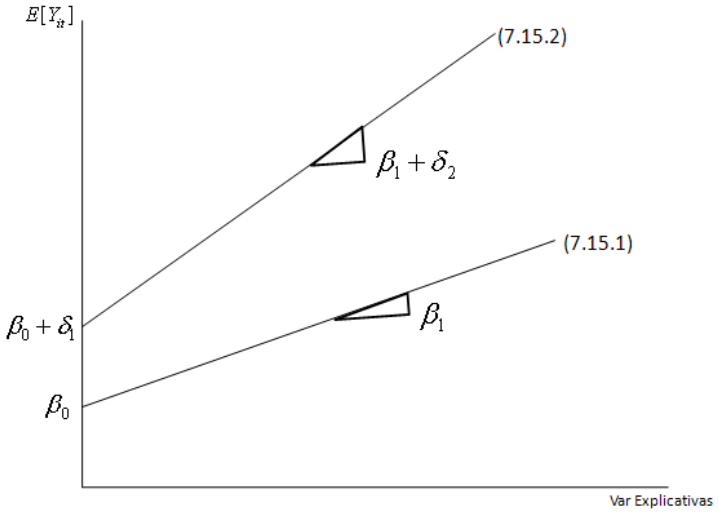
Si t=1: $E[Y_{i1} \mid X_{i1}, K_{i1}, D_2 = 0] = \beta_0 + \beta_1 X_{i1} + \beta_2 K_{i1}$ (7.15.1)

Si t=2: $E[Y_{i2} \mid X_{i2}, K_{i2}, D_2 = 1] = (\beta_0 + \delta_1) + (\beta_1 + \delta_2) X_{i2} + \beta_2 K_{i2}$ (7.15.2)

Las expresiones 7.15.1 y 7.15.2 muestran cuál es el valor esperado de Y_{it} cuando se evalúa el modelo en $t = 1$ o $t = 2$ teniendo en cuenta la inclusión de la variable dicótoma D_2 y la interacción $X_{it} \cdot D_2$. De esta forma se puede identificar si las relaciones entre las variables del sistema varían como consecuencia del tiempo y/o el cambio originado por el paso del tiempo.

Si conjuntamente δ_1 y δ_2 resultan conjuntamente significativos estadísticamente, se dice que existió un cambio estructural debido al cambio en intercepto y pendiente, es decir, genera un desplazamiento y movimiento positivo (o negativo) de la curva que hace referencia al valor esperado de la variable dependiente Y_{it} (*véase* gráfica 7.3).

Gráfica 7.3 Cambio en intercepto y pendiente en modelo estructural



Fuente: los autores

De acuerdo a la gráfica 7.3, hay que tener en cuenta los coeficientes δ_1 y δ_2, al tiempo es necesario verificar si los dos, al mismo tiempo, son iguales a cero o no. Para ello se utiliza una estadístico de prueba F que permite evaluar la significancia conjunta de dos estimadores. Formalmente la prueba de hipótesis sería:

$H_0 : \delta_1 = \delta_2 = 0$ No existe cambio ni en intercepto ni en pendiente (7.16)

$H_1 : \delta_1 \neq \delta_2 \neq 0$ Existe o cambio en intercepto, y/o en pendiente (7.17)

A partir de lo anterior, vale pena enfatizar en la forma que toma el estadístico de prueba utilizado en este análisis, puesto que permite diferenciar los procedimientos anteriores del que se está llevando a cabo ahora. La F está expresada de la siguiente manera:

$$F_c = \frac{(SCE_R - SCE_{NR})/j}{SCE_{NR}/n-k-1} \sim F_{j,n-k-1} \quad (7.18)$$

Donde SCE es la sumatoria de los errores al cuadrado y los subíndices R y NR hacen referencia al modelo restringido[138] y no restringido[139], respectivamente. *j* es igual al número de restricciones, k es el número de coeficientes en el modelo no restringido y n es el número total de observaciones. De acuerdo a la prueba de hipótesis, si $F_c > F_{j,n-k-1}$ se dice que se rechaza H_0, con lo que se puede decir que δ_1 y δ_2 son conjuntamente significativos. Si lo anterior se cumple, entonces existe un cambio estructural en el modelo.

7.3.3 Estimador diferencia en diferencia

Una vez que se ha estudiado de forma práctica el método de agrupación de datos de corte transversal a lo largo del tiempo, se introduce el estimador diferencia en diferencia (DD) como una alternativa para evaluar efectos de choques exógenos sobre las variables explicativas de algún modelo econométrico, de forma directa. Este procedimiento es útil para explicar el impacto de alguna política económica. A la vez, es utilizada comúnmente en áreas de estudio económico como la evaluación de proyectos.

Partiendo de lo anterior, el DD se basa en experimentos naturales (o cuasiexperimentos), que ocurren cuando un evento exógeno, al modelo, cambia el contexto en que las unidades de corte transversal se comportan. Lo anterior puede determinar que las relaciones económicas entre las variables involucradas en determinado estudio sean distintas con el paso del tiempo. Para evaluar dichas variaciones, siempre se tiene en cuenta un grupo de control, que no es afectado por el choque exógeno, y un grupo de tratamiento, que si lo está. Los dos deben ser escogidos de forma aleatoria para evitar sesgos de selección[140].

Asimismo, para revisar las diferencias relativas entre grupos, es pertinente dividir la muestra, que caracterice los datos de corte transversal en distintos periodos de tiempo, teniendo en cuenta escenarios antes y después de ocurrido el evento

[138] Modelo restringido: $Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + u_{it}$

[139] Modelo no restringido: $Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + \delta_1 D_2 + \delta_2 X_{it} \cdot D_2 + e_{it}$

[140] Para más detalles, *véase* (Gujarati ,2003, 453)

exógeno. Por lo cual, se tienen en cuenta los grupos de control y tratamiento en cada periodo de tiempo (véase ecuación 7.19).

$$Y_{it} = \alpha_0 + \alpha_1 D_2 + \alpha_2 D_T + \alpha_3 D_2 \cdot D_T + u_{it} \qquad (7.19)$$

En la ecuación 7.19, D_T toma el valor de uno si la unidad de corte transversal está en el grupo de tratamiento y cero si es del grupo de control. D_2 toma el valor de uno si $t = 2$, y cero en el otro caso. De esta manera, si se estima la ecuación 7.19 por MCA se obtienen los estimadores de diferencia en diferencia de la forma:

$$\hat{\alpha}_{DD} = (\overline{Y}_{T2} - \overline{Y}_{C2}) - (\overline{Y}_{T1} - \overline{Y}_{C1}) \qquad (7.20)$$

$$\begin{aligned} \hat{\alpha}_{DD} &= [(\hat{\alpha}_0 + \hat{\alpha}_1 + \hat{\alpha}_2 + \hat{\alpha}_3) - (\hat{\alpha}_0 + \hat{\alpha}_1)] - [(\hat{\alpha}_0 + \hat{\alpha}_2) - (\hat{\alpha}_0)] \\ \hat{\alpha}_{DD} &= [\hat{\alpha}_2 + \hat{\alpha}_3] - [\hat{\alpha}_2] \\ \hat{\alpha}_{DD} &= \hat{\alpha}_3 \end{aligned} \qquad (7.21)$$

El estimador α_3 de la ecuación 7.19 captura la diferencia, primero, entre el grupo de control y el de tratamiento en cada uno de los periodos del experimento. En segunda instancia, se establece la diferencia entre los dos periodos de tiempo. A partir de estas dos diferencias, se llega a concluir que el efecto de un choque exógeno en el modelo está determinado por el coeficiente que acompaña la variable $D_2 \cdot D_T$, ésta hace referencia al grupo de tratamiento en el periodo después de ser afectadas las unidades de corte transversal, y al mismo tiempo es el efecto de un cambio estructural.

Ahora bien, el modelo se puede construir bajo las variables de corte transversal a lo largo del tiempo como las que se tenían en la ecuación inicial 7.1 de este capítulo. Si se quiere evaluar el efecto de un choque exógeno (cambio estructural o impacto) sobre una variable explicativa, simplemente se plantea un modelo igual al de la ecuación 7.14 teniendo en cuenta la iteración de la variable explicativa y la dummy de tiempo (véase ecuación 7.20).

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \delta_1 D_2 + \delta_2 X_{it} \cdot D_2 + u_{it} \qquad (7.20)$$

Los efectos de un cambio en la variable de interés X_{it} como consecuencia de una política se puede evidenciar por medio del estimador $\hat{\delta}_2$. Este estimador guarda las propiedades de MCO siempre y cuando se cumplan los supuestos del modelo de regresión clásico, en especial, que no se genere endogeneidad en el sistema.

Es importante aclarar que el estimador diferencia en diferencia funciona bien cuando se cuenta con información de corte transversal agrupada en el tiempo, o lo que es lo mismo, cuando no se cuenta con información para la misma unidad de corte transversal para los periodos en los que están recogidos los datos. En el caso en el que se tenga la misma muestra a lo largo del tiempo, se tendría un panel de datos y se utilizarían diferentes técnicas para llegar a las conclusiones esperadas. Estas metodologías se trabajaran con suficiente detalle en el capítulo 8.

7.4 Estudio de caso: impacto de un programa de intervención a las escuelas rurales en Colombia.

Luego de estudiar los conceptos y metodologías para muestras de corte transversal agrupada en varios periodos, esta sección desarrolla un ejercicio empírico con el que se pretende poner en práctica lo trabajado en este capítulo. Para ello se utiliza el artículo "Hacia una mejor educación rural: Impacto de un programa de intervención a las escuelas en Colombia" escrito por Rodríguez, Sánchez y Armenta (2007). Este trabajo evalúa el impacto que tuvo el Programa de Educación Rural (PER) en las tasas de eficiencia y calidad de la educación en las escuelas rurales que accedieron a dicho programa.

El PER, en su planteamiento original, buscaba diseñar y ejecutar proyectos educativos en instituciones rurales para alcanzar cuatro objetivos principales: primero, aumentar la cobertura y calidad educativa; segundo, fortalecer la capacidad de gestión de los municipios e instituciones educativas en la identificación de necesidades, manejo de información, planeación y evaluación; tercero, mejorar las condiciones de convivencia en la institución educativa; y cuarto, diseñar mecanismos que permitieran una mejor comprensión de la situación de la educación media técnica rural. Con esta finalidad, el proyecto tendría una duración de diez años y se implementaría en tres etapas, cada una de tres años y medio. Las primeras experiencias del PER comenzaron en el año 2002, y un año después se había implementado en más de 1,800 sedes en 12 departamentos del país.

De acuerdo a lo anterior, el artículo pretende comparar los resultados académicos que obtuvieron los estudiantes que fueron intervenidos por el PER respecto a los mismos que hubiesen alcanzado las personas si no participaran en el programa. Partiendo de la metodología descrita en este capítulo, la mejor forma de estimar los efectos de una política de este tipo, es utilizando datos de corte transversal a lo largo del tiempo, puesto que permiten analizar los efectos del PER sobre la eficiencia y la calidad de la educación en población rural colombiana.

Bajo la metodología de diferencias en diferencias, los resultados de un grupo de escuelas no participantes en el PER se utilizan como control para los valores del grupo de tratamiento. Por tanto el modelo a seguir está dado por:

$$Y_{it} = \delta_0 + \delta_1 Esc.PER_{it} + \delta_2 A.2004 + \delta_3 Esc.PER_{it} \cdot A.2004 + \mathbf{X\beta} + u_{it} \qquad (7.21)$$

En la ecuación 7.21, Y_{it} es la variable de interés para la evaluación141; $Esc.PER_{it}$ se refiere a las escuelas intervenidas por el programa PER; $Esc.PER_{it} \cdot A.2004$ corresponde a las escuelas intervenidas en el año 2004; y $\mathbf{X}$ es una matriz de controles de la regresión, con $\boldsymbol{\beta}$ el vector de coeficientes.

La hipótesis central de la evaluación es que el programa de educación rural tuvo un impacto positivo sobre el crecimiento en matricula escolar y la tasa de aprobación, y negativo sobre la tasa de reprobación y la tasa de deserción. Para corroborar lo anterior, se llevará a cabo el procedimiento a través del programa computacional Stata®.

7.4.1 Análisis general de los datos

A partir de la información expuesta anteriormente, se describirán los pasos a seguir para determinar los resultadas de interés consignados en el artículo de Rodríguez *et al* (2007). A continuación están los requerimientos básicos para utilizar la base de datos.

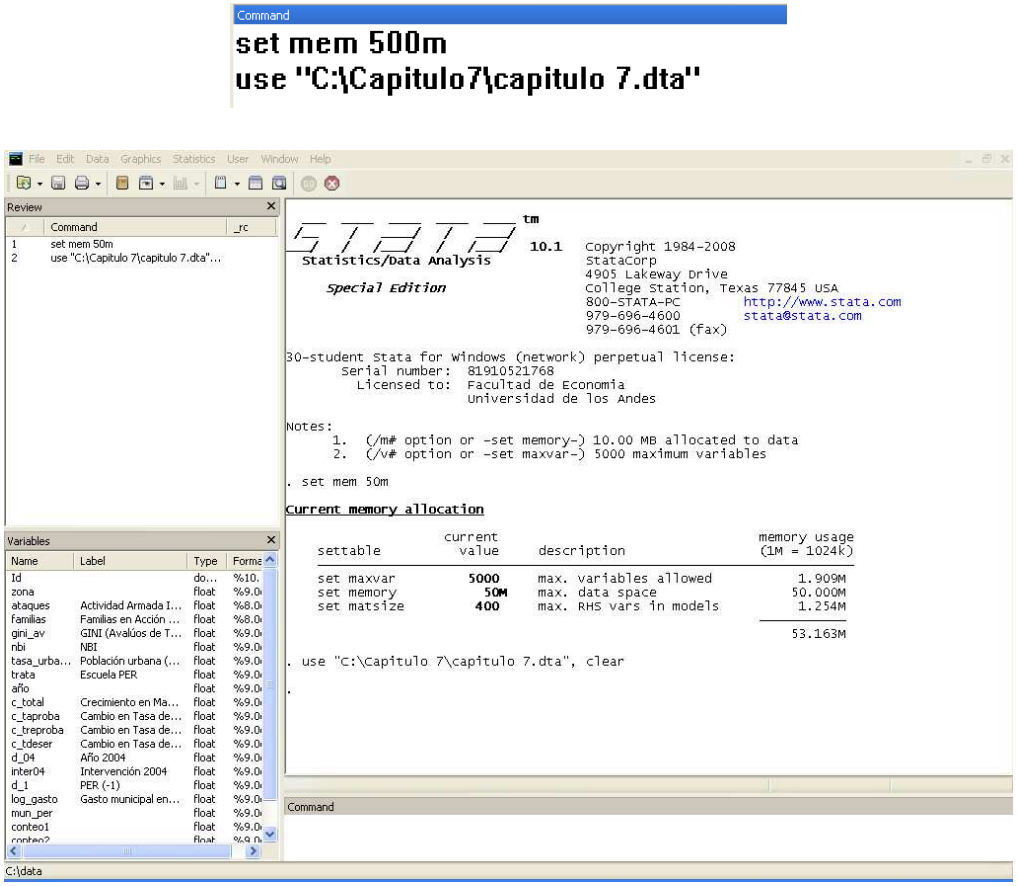
1. Se debe determinar la memoria con la que se van a cargar los datos de interés. Esto se consigue con el comando *set mem.* Para este caso se utiliza 50m.

2. Una vez se ha asignado la memoria del sistema, se puede proceder a cargar la base de datos. La base usada en este trabajo lleva el nombre *capitulo7.dta*,

141 Crecimiento en matricula escolar, cambio en la tasa de aprobación, cambio en la tasa de reprobación y cambio en la tasa de deserción, en diferentes regiones.

y es tomada del trabajo de Rodríguez *et al* (2007). La información hace referencia a datos censales de matrícula, indicadores de eficiencia y calidad de las escuelas rurales de Colombia (*véase* figura 7.1)

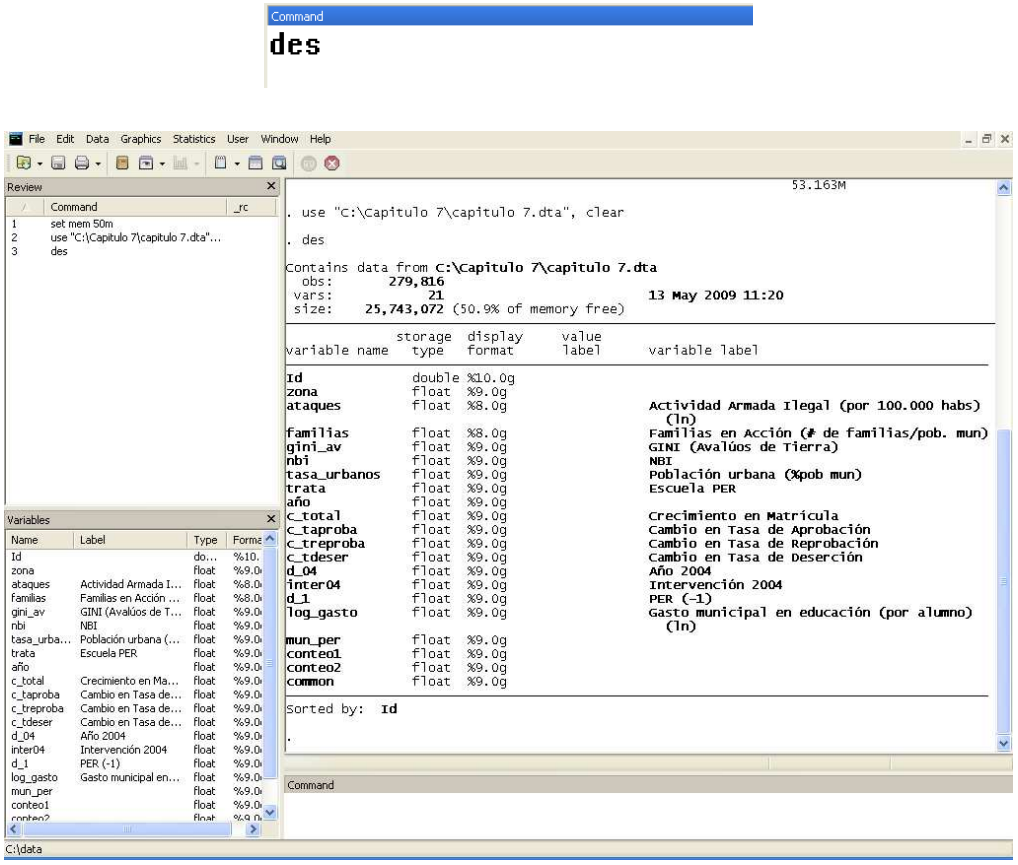
Figura 7.1. Salida comandos set memory y use



Fuente: cálculos autores

3. Para observar las variables que se encuentran disponibles, se usa el comando *describe* –o *des-*. Este comando genera un cuadro con la lista de las variables que se encuentran en la base de datos, el formato en que están guardadas, y una descripción de cada una (*véase* figura 7.2).

Figura 7.2. Salida comando describe



Fuente: cálculos autores

De la figura 7.2 se observa que la muestra cuenta con 279.816 observaciones y 21 variables disponibles para realizar las estimaciones pertinentes en cada caso. En el cuadro 7.1 se presentan las variables a usar, para estimar la ecuación 7.21.

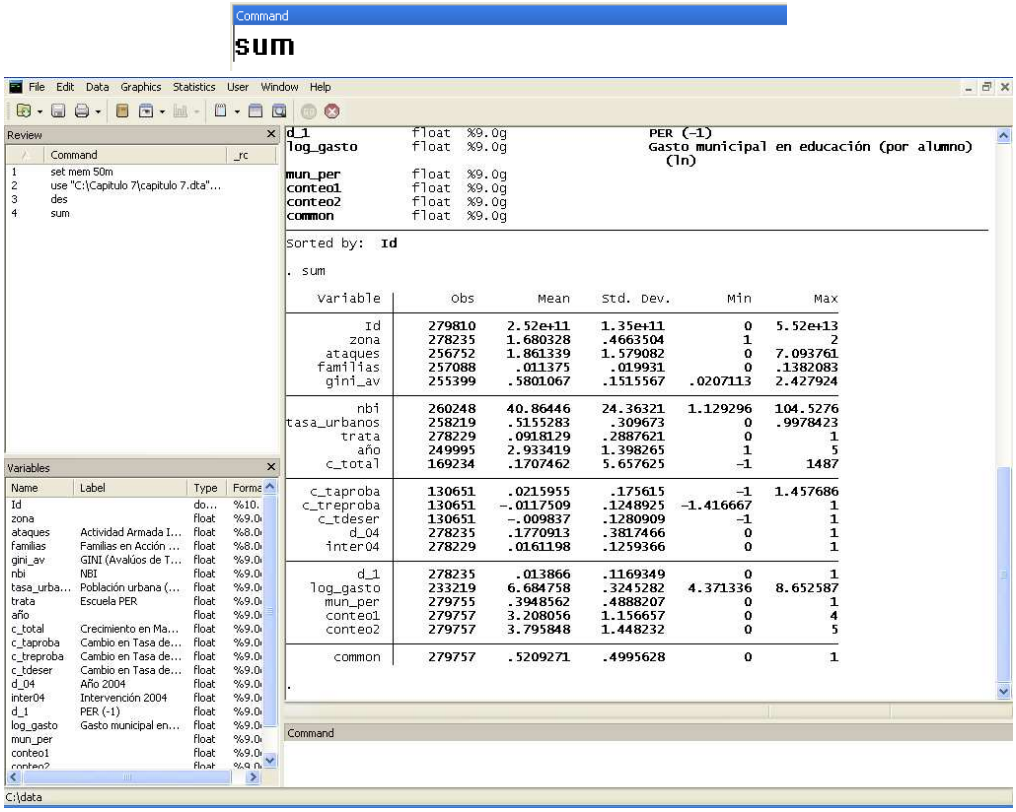
Cuadro 7.1. Variables a usar en el modelo

Variable del Modelo	Variables en la Base	Descripción
Y_{it}	C_total, C_taproba C_treproba, C_tdeser	Crecimiento en la matricula, cambio en la tasas de aprobación, reprobación y deserción
$Esc.PER_{it}$	trata	Variable dicótoma que toma un valor de uno la escuela hace parte del programa PER y cero cuando no.
$A.2004$	d_04	Dicótoma que hace referencia al año 2004.
X_{it}	log_gasto, familias, ataques, gini_av, nbi, tasa_urbanos, d_1	Gasto municipal en educación (por alumno, en log), porcentaje de Familias en Acción, actividad armada ilegal (por 100.000 habitantes, en log), GINI (avalúos de tierra), NBI, población urbana (en porcentaje)

Fuente: los autores

4. Antes de pasar a estimar la regresión lineal, es necesario observar las estadísticas descriptivas de las variables. El comando *summary* –o *sum-*, presenta un cuadro con el número de observaciones, la media, desviación estándar y mínimo- máximo de las variables especificadas (*Véase* figura 7.3).

Figura 7.3. Salida comando summary



Fuente: cálculos autores

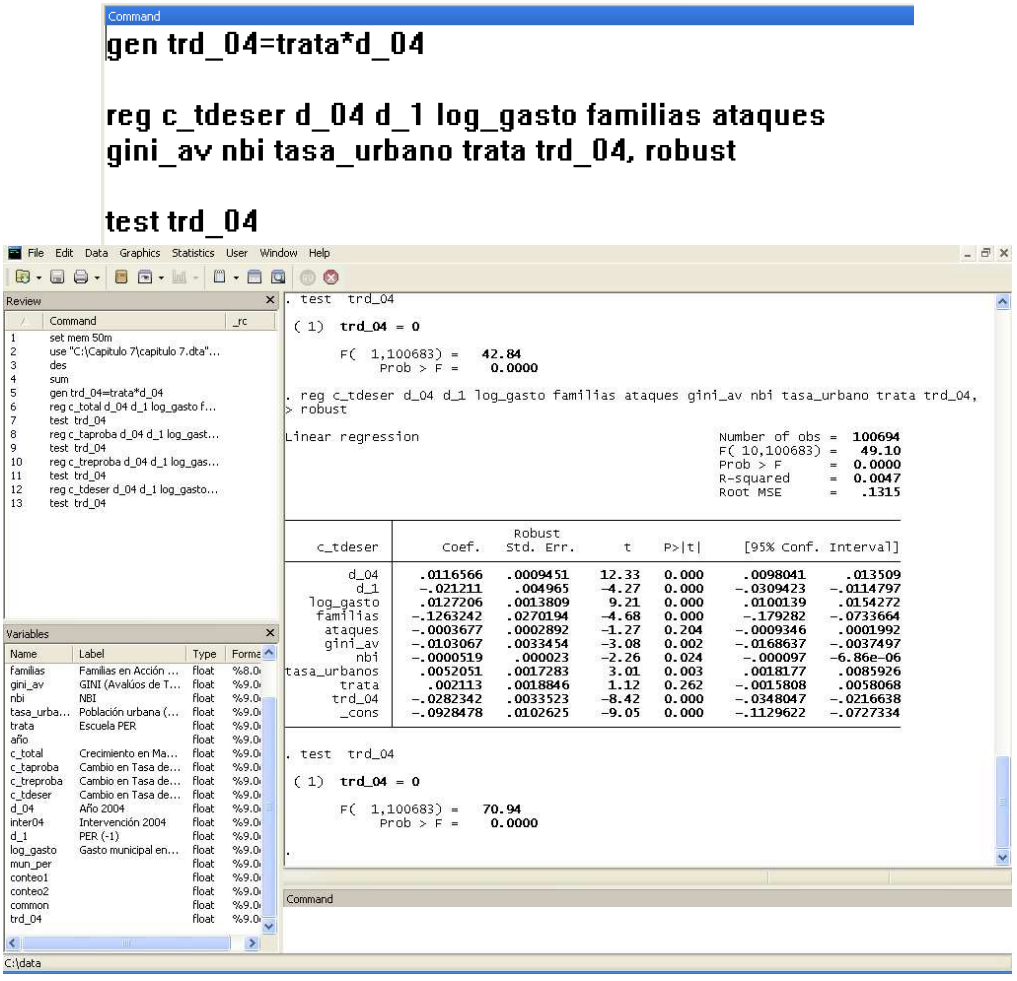
7.4.2 Estimación del modelo diferencias en diferencias.

Una vez que se sabe con qué información se cuenta para realizar las estimaciones, se pasa a desarrollar el ejercicio empírico propuesto por Rodríguez *et al* (2007). El objetivo es estimar la ecuación 7.21 a través de la metodología de diferencias en diferencias. Para esto, es necesario hacer visible la variable de interés, que en este caso es la intervención de las escuelas rurales por el PER en el año 2004.

1. Partiendo de lo anterior, es necesario generar una variable que capture la iteración entre $Esc.PER_{it}$ y $A.2004$. Esto se consigue a través del comando *gen*, seguido por el nombre de la nueva variable a conseguir y la multiplicación entre las dos variables en cuestión.

2. A partir de la variable generada, se ejecuta la regresión de la ecuación 7.21 por MCA. En Stata® se utiliza del mismo modo que en una regresión simple el comando *regress* –o *reg*-(*véase* figura 7.4).

Figura 7.4. Salida regresión en MCA con la variable C_tdeser



Fuente: cálculos autores

Una vez se tiene las regresiones de todo el modelo bajo una de las cuatro especificaciones de interés, es pertinente establecer si las hipótesis planteadas inicialmente son ciertas o no. Para la ecuación 7.21, el coeficiente de interés es δ_3, dado que es el estimador del efecto que tuvo el programa PER sobre las escuelas rurales en el año 2004.

5. De acuerdo al paso 4, es conveniente evaluar la significancia estadística de δ_3, esto para validar algún efecto del PER sobre las escuelas rurales. Para ello se utiliza la prueba de Chow. La prueba de hipótesis a seguir es la siguiente:

$$\begin{array}{ll} H_0 : \delta_3 = 0 & \text{PER no tiene efectos sobre la tasa de deserción escolar} \\ H_0 : \delta_3 \neq 0 & \text{PER tiene efectos sobre la tasa de deserción escolar} \end{array} \qquad (7.22)$$

Esta prueba se hace, una vez realizadas las regresiones, a través del comando *test* que utiliza el estadístico de prueba F para evaluar la significancia individual del estimador[142]. Asimismo, se puede desarrollar el mismo procedimiento para cada una de las variables dependientes con las que cuenta el artículo y se evalúa el efecto particular de cada modelo.

Si $F_c > F_{j,n-k-1}$, se rechaza H_0, con lo que se dice que δ_3 es significativo. De acuerdo a la figura 7.4, la prueba arroja un $F_c = 70.94$ con un p-valor de 0.000, esto quiere decir que $F_c > F_{j,n-k-1}$. De esta manera, se valida la existencia de un efecto del PER sobre la tasa de deserción escolar y dado que el coeficiente de la iteración $Esc.PER_{it} \cdot A.2004$ es negativo, corrobora la hipótesis inicial de un impacto negativo del PER en 2004 sobre la tasa de deserción para el 2004.

[142] Cuando se trata de la prueba de significancia individual la prueba F muestra los mismos resultados que la prueba t .

Resumen

- Los datos de corte transversal agrupados en el tiempo, son una muestra que mezcla datos de corte transversal con los de series de tiempo, a través de la unión de dos o más muestras representativas obtenidas en momentos diferentes del tiempo Esto sirve para estudiar la relación entre variables en distintos periodos de tiempo, en especial para hacer evaluaciones de impacto de políticas económicas.
- Este tipo de agrupación ofrece mayores beneficios que las muestras de corte transversal. La principal es el incremento el tamaño de la muestra, permitiendo obtener estimadores más precisos (consistentes) y estadísticos de prueba más confiables. La mayor cantidad de datos implica una mayor variabilidad entre ellos, menor colinealidad entre las variables, más grados de libertad y mayor eficiencia en las estimaciones. Asimismo, es posible investigar si las relaciones entre las variables han cambiado con el paso del tiempo.
- Existen dos tipos agrupación de los datos de corte transversal en el tiempo: 1. Las muestras que agrupan datos en el tiempo, es decir, muestras aleatorias de corte transversal de la misma población en diferentes periodos del tiempo, pero no necesariamente se tiene en cuenta la misma muestra en cada uno de ellos. 2. Los paneles que agrupan las mismas unidades de corte transversal en diferentes periodos del tiempo.
- La prueba de Chow es la representación teórica de la metodología de diferencia en diferencia. Esta analiza los cambios estructurales en el tiempo como consecuencia de cambios exógenos en el modelo de interés.
- El estimador diferencia en diferencia resume todo el proceso que debe seguir un estudio cuando está enfocado en evaluar impactos exógenos sobres las variables del modelo en distintos periodos del tiempo.

Capítulo 8

Modelos para datos panel o longitudinales

8.1 Introducción

En contraste con la unión de muestras corte transversal a lo largo del tiempo el tiempo presentados en el capítulo anterior, a continuación se estudian metodologías diseñadas para bases de datos compuestas por un conjunto único de unidades de sección cruzada (países, regiones, personas, etc.), rastreadas periodo a periodo (mensualmente, trimestralmente, anualmente, etc.). Ésta información, conocida como paneles de datos o bases longitudinales, corresponden a una sola muestra representativa observada con regularidad.

Usar datos con estas características en estudios econométricos, tiene al menos cuatro ventajas. En primer lugar, las observaciones repetidas en el tiempo permiten eliminar efectos cíclicos, útil para probar modelos teóricos de largo plazo. Adicionalmente, posibilitan solucionar problemas de endogeneidad resultantes de variables omitidas constantes periodo a periodo. Por último, al igual que en la unión de cortes transversales, permiten de identificar y medir efectos no detectables en muestras de corte transversal, y mejoran la precisión de las estimaciones (Baltagi, 2005, 3-6).

Con respecto a lo anterior, este capítulo discute inicialmente las características de este tipo de muestras, para luego centrarse en las consecuencias y metodologías de panel de datos. Específicamente, se presentan las estimaciones de efectos entre grupos, efectos aleatorios y efectos fijos; comunes en los artículos recientes de economía.

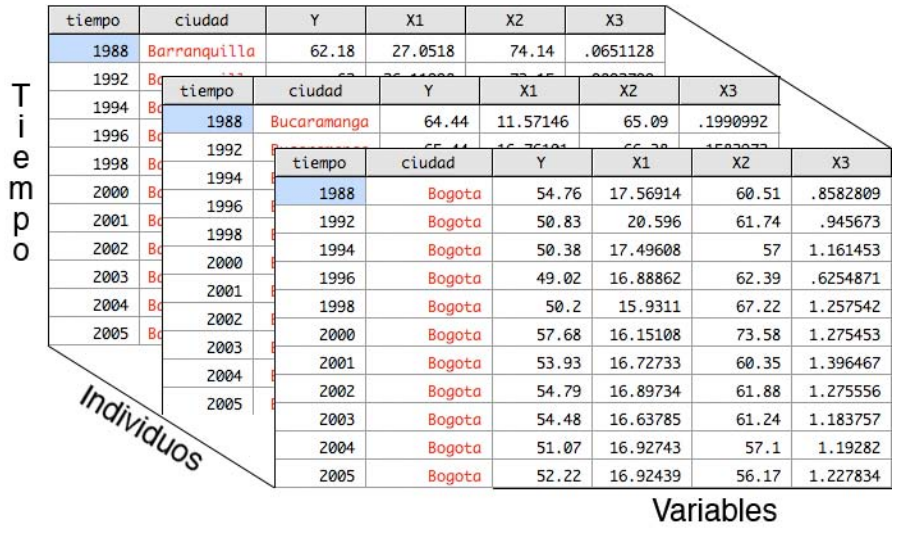
Al igual que en los capítulos anteriores, inicialmente se exponen los temas relevantes de manera formal, para luego aplicarlos a un estudio de caso. En esta oportunidad, los datos y técnicas provienen del artículo titulado "*Informalidad*

regional en Colombia. Evidencia y determinantes" de García (2008) con el cual se pretende estudiar los diferenciales regionales en el grado de informalidad laboral en Colombia, utilizando datos de tipo longitudinal

8.2 Organización de los paneles de datos

A diferencia de los datos de corte transversal y series temporales presentados anteriormente, este capítulo discute el uso de paneles compuestos por la unión de observaciones recolectadas en el tiempo, para un conjunto determinado de unidades de corte transversal. La gráfica 8.1 es una representación tridimensional que permite comprender conceptualmente la estructura de este tipo de información. En este caso, los datos cuentan con una dimensión que corresponde al tiempo –el eje vertical, identificado por la primera columna-, otra para los individuos –el eje diagonal, caracterizado por la segunda columna-, y una tercera para las variables –en el eje horizontal, a partir de la tercera columna-. Por esta razón, una base de datos longitudinales puede interpretarse tanto como un seguimiento de la misma muestra a lo largo del tiempo, o como una unión de series de tiempo para varios individuos.

Gráfica 8.1. Visión tridimensional de un panel



Fuente: los autores

Aunque la visión tridimensional es útil conceptualmente, para el uso de programas computacionales, los datos deben estar ordenados como una matriz, con filas que representen observaciones (una para cada individuo) y columnas que contengan valores para cada variable. Si se observan n unidades de sección cruzada durante T periodos, y para cada observación se cuenta con k variables, el conjunto de datos tendrá nkT valores.

Como ahora no existen uno sino dos elementos que identifican a cada observación -la unidad de corte transversal y el momento del tiempo-, los paneles deben organizarse en una forma particular. A continuación se presentan dos posibles formas de ordenar los datos:

1. *Agrupar las filas por unidad*: suponga que la matriz de datos contiene inicialmente todas las observaciones de la primera unidad de corte transversal, para todos los momentos del tiempo. A continuación, inician las observaciones del siguiente individuo, y así sucesivamente. De esta forma, la matriz de datos queda ordenada como un conjunto de series temporales apiladas verticalmente, con n secciones (una por cada individuo), cada una compuesta de T filas (por periodo de tiempo) (*véase* gráfica 8.2).

Gráfica 8.2. Panel con filas agrupadas por unidad.

ciudad	tiempo	Y	X1	X2	X3
Barranquilla	1988	62.18	27.0518	74.14	.0651128
Barranquilla	1992	62	26.11998	73.15	.0802799
Barranquilla	1994	57.92	21.86649	68.96	.2278566
Barranquilla	1996	59.15	21.9932	73.29	.6506485
Barranquilla	1998	65.18	21.04429	72.63	.4847876
Barranquilla	2000	68.8	20.62614	75.43	.8509239
Barranquilla	2001	62.46	19.10799	67.44	.8866857
Barranquilla	2002	63.41	19.55784	69.21	.8054969
Barranquilla	2003	61.75	19.9551	68.98	.8423997
Barranquilla	2004	62.92	19.83938	67.85	.6831863
Barranquilla	2005	63.19	19.67786	65.82	.8248166
Bucaramanga	1988	64.44	11.57146	65.09	.1990992
Bucaramanga	1992	65.44	16.76101	66.38	.1582972
Bucaramanga	1994	66.2	15.14647	63.26	.1696348
Bucaramanga	1996	62.9	15.51173	70.26	.2250799
Bucaramanga	1998	67.76	15.4124	69.08	.2615195
Bucaramanga	2000	67.5	18.27614	78.43	.3816909
Bucaramanga	2001	69.11	17.8723	75.59	.3040712
Bucaramanga	2002	65.31	17.0489	72.38	.2443494
Bucaramanga	2003	68.25	19.2539	72.63	.5954379
Bucaramanga	2004	64.37	19.25272	69.57	.6428825
Bucaramanga	2005	64.51	20.95028	70.81	.7209885

Fuente: los autores a partir de García (2008)

2. *Agrupar las filas por periodo*: En este caso, la matriz de datos estará compuesta por T secciones (uno por cada unidad de tiempo), cada uno con n filas (correspondientes a cada individuo). De esta forma, el primer bloque contiene las observaciones recolectadas en el primer periodo, para cada uno de los individuos; el segundo, las observaciones para el periodo siguiente; y así sucesivamente. Aquí, la matriz de datos corresponde a un conjunto de muestras de sección cruzada, apiladas verticalmente (*véase* gráfica 8.3).

Gráfica 8.3. Panel con filas agrupadas por periodo.

tiempo	ciudad	Y	X1	X2	X3
2003	Bogota	54.48	16.63785	61.24	1.183757
2003	V/CENCIO	75.64	5.287313	76.67	1.668204
2003	C/CUTA	74.97	5.077655	80.39	.6812333
2003	PEREIRA	61.78	13.34973	62.65	.9026697
2003	Medellin	60.15	17.95798	57.5	.9696133
2003	Bucaramanga	68.25	19.2539	72.63	.5954379
2003	CALI	61.44	18.55664	67.56	.8420485
2003	Barranquilla	61.75	19.9551	68.98	.8423997
2003	MANIZALES	61.74	12.0883	65.39	1.680004
2003	PASTO	69.2	3.288382	75.19	1.52958
2004	V/CENCIO	70.55	5.787243	75.24	1.450779
2004	PEREIRA	62.19	13.75955	64.56	.8966156
2004	Medellin	55.35	18.39003	53.12	.9807917
2004	C/CUTA	74.83	5.673062	80.1	.9215235
2004	CALI	62.19	18.64841	67.01	.910323
2004	Bogota	51.07	16.92743	57.1	1.19282
2004	Bucaramanga	64.37	19.25272	69.57	.6428825
2004	PASTO	68.86	3.267123	72.32	1.450947
2004	MANIZALES	60.36	13.65252	62.29	1.662287
2004	Barranquilla	62.92	19.83938	67.85	.6831863

Fuente: los autores a partir de García (2008)

Los primeros paneles contenían información sobre la evolución de un conjunto de indicadores económicos de interés -como el PIB, inversión y consumo- para un grupo determinado de países. Estos se conocen como paneles macro, y usados generalmente para evidenciar hipótesis sobre crecimiento, convergencia, ciclos y

estabilidad macroeconómica. En estas bases de datos, las dimensiones de tiempo y de individuos son de similar magnitud.

Un segundo tipo de panel, es aquel que registra información sobre un grupo particular de individuos o firmas. Estudios con este tipo de datos aparecieron por primera vez en la literatura relacionada con la producción agrícola, y posteriormente, su uso se expandió a otras corrientes de la literatura económica. Estas bases de datos son más complejas de construir, dado que es costoso seguir periodo a periodo un conjunto de firmas o individuos. En este caso, la dimensión de tiempo suele ser menor que la muestra de individuos (Arellano, 2003, 1-2).

Adicionalmente a la diferenciación entre paneles macro y micro, ellos pueden catalogarse de acuerdo a la disponibilidad de datos. Aquellas bases de datos donde existen observaciones para todas las unidades de sección cruzada en momentos del tiempo sucesivos, son denominadas completas o balanceadas. Aquí, el tamaño de la muestra será nT, que corresponde al número de individuos multiplicado por el número de periodos. Cuando por el contrario, hay información faltante para ciertos individuos o periodos, el panel está incompleto o desbalanceado (Gujarati, 2003, 617).

Una vez presentada la organización de los datos, a continuación se explican las metodologías econométricas apropiadas para el análisis de estos datos. Por simplicidad durante el resto del capítulo se asume un panel microeconómico balanceado. Inicialmente se describe el estimador de efectos entre grupos, útil para obtener estimaciones de largo plazo; posteriormente se presenta el modelo con efectos fijos en el término de error, finalizando con las metodologías de efectos aleatorios y fijos.

8.3 Estimación de dinámicas de largo plazo – efectos entre grupos

Una primera metodología particular del análisis de datos longitudinales, es la estimación de dinámicas entre grupos. En análisis de corte transversal tradicional, cada observación captura información de su nivel de largo plazo y su componente cíclico, algo inadecuado para obtener correlaciones de largo plazo. El estimador de efectos entre grupos que se presenta a continuación, conocido en ingles como

estimación *between*, permite obtener únicamente la información de largo plazo de las variables al separa su componente cíclico del tendencial.

En términos generales, un estimador de efectos entre grupos reduce el problema de un panel longitudinal a un corte transversal, empleando el cálculo promedio de las variables al interior de cada individuo. Como las variables estacionarias se encuentran distribuidas alrededor de su valor de largo plazo, esta tecnica permite eliminar el efecto de los ciclos. El procedimiento general para obtener este estimador, es:

1. Calcular el promedio de la variable dependiente y de las independientes, a lo largo del tiempo, para cada una de las unidades de sección cruzada.
2. Realizar una estimación de MCO, donde se usan como variables, los promedios calculados en el paso 1.

Para analizar en detalle el funcionamiento de esta metodología, considere un modelo de datos longitudinales, donde Y_{it} es la variable dependiente, X_{it} y K_{it} las explicativas, e_{it} un error aleatorio variante tanto entre individuos como en el tiempo (*véase* ecuación 8.1). Al igual que en capítulo 7, los subíndices it denotan los individuos y el momento del tiempo, respectivamente.

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_1 K_{it} + e_{it} \quad (8.1)$$

A partir de la ecuación 8.1, el proceso de estimación consiste en calcular el promedio de las variables a usar, para todos los periodos de tiempo (*véase* ecuación 8.2). Con las medias resultantes ($\bar{Y}_{it}$, $\bar{X}_{it}$, $\bar{Z}_{it}$), se plantea una nueva especificación a estimarse por MCO (v*éase* ecuación 8.3).

$$\bar{Y}_i = \frac{1}{T}\sum_{t=1}^{T} Y_{it} \qquad \bar{X}_i = \frac{1}{T}\sum_{t=1}^{T} X_{it} \qquad \bar{K}_i = \frac{1}{T}\sum_{t=1}^{T} K_{it} \qquad (8.2)$$

$$\bar{Y}_i = \beta_0 + \beta_1 \bar{X}_i + \beta_2 \bar{K}_i + u_{it} \quad (8.3)$$

El modelo de regresión caracterizado por la ecuación 8.3, permite encontrar la correlación condicional de largo plazo entre la variable dependiente y las

independientes. Los estimadores de MCO ($\hat{\beta}_{MCO}$) de esta ecuación, se conocen como estimadores entre grupos ($\hat{\beta}_{EG}$) del modelo inicial 8.1. En general, este procedimiento transforma el modelo a una estimación de sección cruzada. Para garantizar que $\hat{\beta}_{EG}$ sea un estimador insesgado, consistente y eficiente de los parámetros poblacionales, es necesario asumir los supuestos de exogeneidad , homoscedasticidad, no multicolinalidad ni autocorrelación residual.

En la práctica el estimador de efectos entre grupos caracterizado por la ecuación 8.3 es poco usado: la pérdida de observaciones deteriora la precisión en las estimaciones y la reducción del problema a uno de corte transversal elimina la posibilidad de analizar el dinamismo de las variables en el tiempo. La importancia de este estimador radica en que es utilizado para construir otros estimadores; en particular, efectos aleatorios y fijos al interior de grupos que se discuten a continuación.

8.4 El problema de efectos fijos en el término de error

A partir del estimador de efectos entre grupos expuesto anteriormente, se pueden construir metodologías que permitan superar los problemas típicos del uso de datos longitudinales; la idea de correlación serial de los errores, y endogeneidad por variables omitidas constantes en el error. Para comprender el origen de estas problemáticas, a continuación se presenta el modelo de regresión lineal con una descomposición del término de error en un efecto constante, y uno variable en el tiempo. Posteriormente, se expone cada metodología y su relevancia en el análisis de datos longitudinales.

8.4.1 Modelo con término de error compuesto.

Con el fin de comprender las problemáticas particulares que surgen del uso de paneles longitudinales, es necesario replantear el modelo clásico de regresión mostrando los diferentes componentes del término de error. Como representación, suponga el modelo bivariado utilizado en la sección anterior donde Y_{it} es la variable dependiente y X_{it} y K_{it} las independientes (*véase* ecuación 8.4). A partir

de esto, considere el termino de error u_{it}, como la suma de tres términos independientes: C_i, D_t y ε_{it} (*véase* ecuación 8.5).

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + u_{it} \quad (8.4)$$

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + \underbrace{C_i + D_t + \varepsilon_{it}}_{u_{it}} \quad (8.5)$$

En primer lugar, C_i corresponde a un efecto fijo por individuo invariante periodo a periodo, que se conoce como la heterogeneidad no observada de la muestra, y corresponde a vector conformado por las variables constantes en el tiempo capturadas por el error. Estas variables suelen ser características no observables de individuos, tales como habilidades congénitas no cuantificables.

Análogamente, D_t corresponde a un efecto con varianza periodo a periodo, pero invariante entre individuos, es decir, rasgos comunes a toda la muestra, que cambian de un periodo a otro. Por ejemplo: el clima y lugar donde reside la población, hacen parte de este término.

Finalmente, el resto de variables del error cambiantes tanto entre individuos como a lo largo del tiempo, se denotan ε_{it}. Como todos estos términos corresponden a un único término de error, el valor esperado del modelo continúan siendo una suma ponderada de las variables independientes (*véase* ecuación 8.6).

$$E[Y_{it}] = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} \quad (8.6)$$

El resto del capítulo está centrado en los efectos de la heterogeneidad no observada (C_i), efecto variante entre individuos pero constante en el tiempo. Las consecuencias y metodologías presentadas a continuación, son análogas para casos cuando se tiene un efecto invariante entre individuos D_t.[143]

[143] Para más detalles, *véase* Baltagi (2005).

8.4.2 Efectos aleatorios

8.4.2.1 Correlación serial resultante de efectos constantes en el error

Un primer problema de las estimaciones usando datos longitudinales, es la posible correlación serial entre los errores de diferentes periodos encuestados para la muestra; inconveniente por la existencia del efecto constante por individuo C_i.

Para observar el origen de este problema, suponga el modelo de regresión bivariada caracterizado por la ecuación 8.5, con un efecto fijo (C_i) en el error. Para este caso, se asume que no existen efectos variantes únicamente en el tiempo (*véase* ecuación 8.7)

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + u_{it} \qquad \text{con} \qquad u_{it} = C_i + \varepsilon_{it} \qquad (8.7)$$

Suponiendo que el panel longitudinal cuenta con información para al menos dos periodos de tiempo, se deduce que:

En $t = 1$: $\quad Y_{i1} = \beta_0 + \beta_1 X_{i1} + \beta_2 K_{i1} + u_{i1} \qquad \text{con} \qquad u_{i1} = C_i + \varepsilon_{i1} \qquad (8.8)$

En $t = 2$: $\quad Y_{i2} = \beta_0 + \beta_1 X_{i2} + \beta_2 K_{i2} + u_{i2} \qquad \text{con} \qquad u_{i2} = C_i + \varepsilon_{i2} \qquad (8.9)$

Las ecuaciones 8.8 y 8.9 muestran como los errores del modelo usualmente están correlacionados serialmente, dada la existencia de C_i en u_{i1} y u_{i2}. Esto causa estimadores de MCO menos eficientes en comparación a los que se obtendrían sin autocorrelación residual.

En particular, es importante recordar como en la construcción de pruebas t usadas para evaluar la significancia de un coeficiente, son necesarios los errores estándar. Entre mayor es la varianza –resultante de la autocorrelación residual- menor posibilidad de encontrar la verdadera significancia para una variable, y mayor probabilidad de cometer error tipo II - declarar un coeficiente estadísticamente no significativo, cuando en realidad lo es-. Para solucionar la correlación bajo paneles longitudinales se utiliza el estimador de efectos aleatorios, presentado a continuación (Baltagi, 2005, 13-18).

8.4.2.2 Estimador de efectos aleatorios

Cuando el término constante en el tiempo C_i causa un problema de autocorrelación residual, debe aplicarse la metodología de estimación por efectos aleatorios, correspondiente a un caso particular de mínimos cuadrados generalizados (MCG). En esta sección particular, es presentada una transformación que permite encontrarlo aplicando una regresión tradicional de MCO.

En relación con lo anterior, esta metodología tiene dos supuestos. Primero, que el efecto fijo –o la heterogeneidad no observada- realmente existe, esto quiere decir, que cierta fracción de las variables no observadas capturadas por el error, son constantes en el tiempo. Si este no es el caso, el problema de autocorrelación residual sería inexistente, por lo cual el uso de MCO debería ser suficiente para encontrar una estimación correcta. Para comprobar específicamente esta condición, es posible aplicar una prueba estadística conocida como Breusch-Pagan (*véase* sección 8.5.2).

Adicionalmente, el efecto fijo debe ser independiente de las variables explicativas del modelo; en otras palabras, no debe estar generando problemas de endogeneidad. El incumplimiento de este supuesto, lleva al problema de efectos fijos presentado más adelante (*véase* sección 8.4.3).

De esta forma, el procedimiento general para obtener el estimador de efectos aleatorios, es:

1. Transformar el modelo inicial a un nuevo modelo ponderado, sin correlación residual.
2. Realizar una regresión por MCO del modelo transformado en 1. Estos estimadores son los coeficientes de efectos aleatorios del modelo inicial.

A partir de lo anterior, el primer paso consiste en aplicar una transformación al modelo, a través de un ponderador λ, función de la heterogeneidad no observada y del error variante entre periodos y entre individuos (σ_c^2 y σ_ε^2). Esto para obtener el estimador de mínimos cuadrados generalizados (MCG) a través de una regresión lineal de mínimos cuadrados ordinarios (MCO). Formalmente, para el modelo definido en la ecuación 8.7, se plantea:

$$\lambda \bar{Y}_i = \beta_1 \lambda \bar{X}_i + \beta_2 \lambda \bar{K}_i + \lambda C_{it} + \lambda \varepsilon_{it} \qquad (8.10)$$

Con,

$$\bar{Y}_i = \frac{1}{T}\sum_{i=1}^{T} Y_{it} \qquad \bar{X}_i = \frac{1}{T}\sum_{i=1}^{T} X_{it}$$
$$\bar{K}_i = \frac{1}{T}\sum_{i=1}^{T} K_{it} \qquad \lambda = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_c^2}} \qquad (8.11)$$

Con esta información, el modelo queda transformado a:

$$\underbrace{(Y_{it} - \lambda \bar{Y}_i)}_{Y^*} = \beta_1 \underbrace{(X_{it} - \lambda \bar{X}_i)}_{X^*} + \beta_2 \underbrace{(K_{it} - \lambda \bar{K}_i)}_{K^*} + u_{it} \qquad (8.12)$$

La regresión 8.12 puede ser estimada directamente por MCO. Los β_{MCO} de la ecuación 8.12, corresponden a los estimadores de efectos aleatorios β_{EA} del modelo inicial 8.7. La validez de esta transformación puede verificarse, partiendo de la formula convencional de mínimos cuadrados generalizados (Greene, 2000, 568-570):

$$\hat{\beta}_{MCG} = (X^t \Omega^{-1} X)^{-1} X^t \Omega^{-1} Y \quad (8.13)$$

con

$$\Omega^{-1} = \begin{pmatrix} \Sigma & 0 & 0 & & 0 \\ 0 & \Sigma & 0 & & 0 \\ 0 & 0 & \Sigma & & 0 \\ & & & \ddots & \\ 0 & 0 & 0 & & \Sigma \end{pmatrix} \quad \Sigma = \begin{pmatrix} \sigma^2{}_\varepsilon + \sigma^2{}_c & \sigma^2{}_c & & \sigma^2{}_c \\ \sigma^2{}_c & \sigma^2{}_\varepsilon + \sigma^2{}_c & & \sigma^2{}_c \\ & & \ddots & \sigma^2{}_c \\ \sigma^2{}_c & \sigma^2{}_c & & \sigma^2{}_\varepsilon + \sigma^2{}_c \end{pmatrix} \qquad (8.14)$$

En la expresión 8.14, Σ corresponde a la matriz de varianza-covarianza de los estimadores, que en este caso está compuesta por la suma de $\sigma^2{}_\varepsilon$ y $\sigma^2{}_c$ en la diagonal, y únicamente de $\sigma^2{}_c$ en todos los otros lugares.

8.4.3 Endogeneidad resultante de efectos fijos en el error

Asumir que el efecto fijo del error no está correlacionado con alguna de las variables explicativas del modelo, es un supuesto estricto. En realidad, los modelos econométricos en base a paneles longitudinales de individuos tienen muchas características no observadas, y la correlación entre efectos constantes y variables independientes es un problema común. En esta sección son presentadas alternativas de estimación para ser aplicadas bajo la condición tratada.

La correlación entre variables independientes y C_i es un caso particular del incumplimiento del supuesto de independencia condicional ó endogeneidad discutido en el capítulo 1. Garantizar su cumplimiento es necesario para obtener estimadores de MCO insesgados, consistentes y eficientes, que se aproximen correctamente a los parámetros poblacionales.

En el contexto de paneles de datos, contar con información para cada individuo en momentos diferentes del tiempo, otorga nuevas alternativas para resolver la endogeneidad, que no se tienen en muestras de corte transversal. En general, suponga el modelo bivariado anterior, con endogeneidad en X_{it} (*véase* ecuaciones 8.15 y 8.16).

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 K_{it} + u_{it} \quad \text{con} \quad u_{it} = C_i + \varepsilon_{it} \tag{8.15}$$

$$\text{cov}(C_i, X_{it}) \neq 0 \tag{8.16}$$

Con el fin de subsanar el problema de endogeneidad, a continuación se presentan tres metodologías que permiten resolverlo:

1. Uso de variables dicótomas variables independientes,
2. Estimador de efectos fijos por primeras diferencias, y
3. Estimador de efectos fijos al interior de los grupos.

Aquí no se tratan temas de endogeneidad resultante de una correlación de las variables independientes y el componente del error que varia tanto entre

individuos como en el tiempo (ε_{it}); en esos casos es necesario aplicar variables instrumentales en el contexto de panel.

8.4.3.1 Uso de variables dicótomas variables independientes,

La alternativa más simple para resolver este tipo de endogeneidad, es capturar el efecto de la heterogeneidad no observada a través de una o más variables independientes. Usando esta técnica, el efecto fijo deja de estar en el término del error y pasa a hacer parte de la especificación del modelo, eliminando el incumplimiento del supuesto de independencia condicional.

Partiendo de efectos fijos iguales por individuo periodo a periodo, es posible capturarlos dentro del modelo de regresión, como cambios de intercepto. Esto equivale a incluir una variable dicótoma por individuo, dentro de las regresoras. El procedimiento general de esta metodología es:

1. Transformar el modelo inicial, eliminando el intercepto y agregando una variable dicótoma por cada individuo de la muestra.

2. Realizar una regresión de MCO, del modelo creado en 1.

El nuevo modelo a estimar corresponde al inicial sin intercepto, más un vector $\mathbf{D}$ de n variables dicótomas[144] (*véase* ecuación 8.17). En el caso en que se decida mantener el intercepto β_0, el vector debe estar compuesto por $n-1$ variables para no caer en multicolinealidad perfecta. En la ecuación 8.17, $\boldsymbol{\delta}$ agrupa todos los coeficientes que acompañan a la matriz.

$$Y_{it} = \beta_1 X_{it} + \beta_2 K_{it} + \boldsymbol{\delta}\mathbf{D} + \varepsilon_{it} \quad (8.17)$$

Aunque esta metodología es sencilla y efectiva para obtener estimadores consistentes, es en términos estadísticos costosa, pues cada variable dicótoma nueva deteriora la precisión de las estimaciones. En las bases de datos con una

[144] Una por cada individuo de la muestra. La variable dicótoma i, tomara el valor de uno para el individuo i y cero para todos los demás.

dimensión de individuos muy grande, la pérdida de grados de libertad resultante puede incluso imposibilitar el cálculo de estimadores. Como alternativa, se usan los estimadores de primeras diferencias y de efectos al interior de grupos presentadas a continuación, corrigendo el problema con menor perdida de grados de libertad.

8.4.3.2 Estimador de efectos fijos por primeras diferencias

Otra forma alternativa para solventar el problema de endogeneidad es eliminar el efecto constante C_i del error, mediante el estimador de primeras diferencias, o el de efectos al interior de grupos (*véase* sección 8.4.3.3).

Bajo primeras diferencias, la metodología consiste en restar el primer rezago en el tiempo a cada observación individual de la base de datos, para eliminar el término constante del error. El procedimiento general, viene dado por:

1. Restarle a cada observación por individuo, los valores observados en el periodo anterior.

2. Realizar una estimación de MCO, donde se usan como variables las primeras diferencias que resultan al aplicar el paso 1.

Las primeras diferencias del modelo 8.15, están definidas por la ecuación 8.18, donde tanto el intercepto, como el efecto fijo por individuo depositado en el término de error fueron eliminados. De manera general, al restar las observaciones del mismo individuo en dos momentos del tiempo, todas las variables constantes desaparecen de la especificación del modelo.

$$\underbrace{Y_{it} - Y_{it-1}}_{Y_i^*} = \beta_1 \underbrace{(X_{it} - X_{it-1})}_{X_i^*} + \beta_2 \underbrace{(K_{it} - K_{it-1})}_{K_i^*} + \underbrace{(\varepsilon_{it} - \varepsilon_{it-1})}_{\varepsilon_i^*} \qquad (8.18)$$

Al estimar la ecuación 8.18, los estimadores de MCO ($\hat{\beta}_{MCO}$) resultantes, se conocen como estimadores de PD ($\hat{\beta}_{PD}$) del modelo inicial caracterizado por la expresión 8.15. A diferencia del control por variables dicótomas, esta metodología

corrige el problema de endogeneidad perdiendo únicamente un grado de libertad, de haber rezagado las variables un periodo en el tiempo.

8.4.3.3 Estimador de efectos fijos al interior de grupos

Finalmente, la tercera forma para remover el problema de endogeneidad, es eliminar el efecto fijo del error, restándole a cada variable su media muestral ($\bar{Y}_i$ $\bar{X}_i$ $\bar{K}_i$). Este procedimiento, tiene el mismo efecto que primeras diferencias, porque todo elemento constante en el tiempo desaparece cuando se restan observaciones del mismo individuo (Baltagi, 2005, 10-13). El procedimiento general para obtener este estimador es:

1. Calcular el promedio, de la variable dependiente y las independientes, para cada individuo en el tiempo.
2. Restarle a los valores contemporáneos de cada variable, el promedio calculado en 1.
3. Realizar una estimación de MCO, a partir de las variables resultantes del paso 2.

El primer paso consiste en calcular los promedios de la variable dependiente y las variables independientes del modelo base (véase ecuación 8.19). Con esta información, se transforma la especificación inicial, restándole a cada una de las observaciones su promedio. El nuevo modelo a estimar, viene dado por la ecuación 8.20.

$$\bar{Y}_i = \frac{1}{T}\sum_{i=1}^{T} Y_{it} \qquad \bar{X}_i = \frac{1}{T}\sum_{i=1}^{T} X_{it} \qquad \bar{K}_i = \frac{1}{T}\sum_{i=1}^{T} K_{it} \tag{8.19}$$

$$\underbrace{Y_{it} - \bar{Y}_i}_{Y_i^*} = \beta_1 \underbrace{(X_{it} - \bar{X}_i)}_{X_i^*} + \beta_2 \underbrace{(K_{it} - \bar{K}_i)}_{K_i^*} + \underbrace{(\varepsilon_{it} - \bar{\varepsilon}_i)}_{\varepsilon_i^*} \tag{8.20}$$

Aplicando esta metodología es posible eliminar el efecto fijo sin perder ningún grado de libertad, corrigiendo el problema de endogeneidad. Los estimadores de

MCO ($\hat{\beta}_{MCO}$) de la ecuación 8.20, se conocen como estimadores de efectos fijos al interior de grupos ($\hat{\beta}_{EF}$) del modelo inicial 8.15.

Los estimadores de primeras diferencias y de efectos al interior de grupos, son idénticos cuando $T = 2$. Para $T > 3$, cuando los errores ε_{it} no están autocorrelacionados, se usan efectos al interior de grupos, transformación que no implica la pérdida de grados de libertad. Si por el contrario, se cree que los errores de un momento del tiempo se relacionan con los errores pasados o futuros, el modelo de primeras diferencias resulta más conveniente.

En ambos casos, las transformaciones aplicadas eliminan la posibilidad de analizar variables constantes en el tiempo. Si el interés está en variables de este tipo, el control por variables dicótomas es la única alternativa para resolver la endogeneidad.

No obstante, en datos panel es posible conocer mediante pruebas estadísticas la existencia de problemas de autocorrelacion residual y endogeneidad por heterogeneidad no observada. La siguiente sección, expone las pruebas estadísticas relevantes para este propósito.

8.5 Identificación del estimador apropiado

Una vez comprendidas teóricamente las diferentes metodologías disponibles para estimar un modelo con datos longitudinales, se pueden discutir los criterios de selección aplicados en la práctica para identificar cuál es la conveniente en cada caso. Esta sección presenta una técnica basada en argumentos teóricos y estadísticos, que permiten determinar el modelo más apropiado; entre efectos al interior de grupos, aleatorios y fijos.

8.5.1 Elección entre dinámicas de largo plazo y datos a través del tiempo

El primer paso para elegir la metodología apropiada, en un modelo econométrico puntual, consiste en determinar la relevancia del estimador de dinámicas de largo plazo. Para esto, se analiza teóricamente con antelación, el propósito de las estimaciones econométricas a realizar. Si el objetivo es obtener coeficientes de una

dinámica de largo plazo, y no hay interés en analizar el comportamiento intertemporal de las variables, es conveniente optar por el estimador entre grupos (*véase* sección 8.3). Si por el contrario se desea continuar usando la base completa, se debe elegir entre mínimos cuadrados agrupados, u otra de las metodologías de datos panel (*véase* sección 8.4).

El costo de usar un estimador entre grupos, es la pérdida de un número muy grande de observaciones, deteriorando la precisión en las estimaciones. Cuando se elige esta alternativa, es necesario contar con suficientes observaciones para obtener estimaciones precisas, aún después de haber eliminado la dimensión temporal del problema.

8.5.2 Elección entre mínimos cuadrados agrupados y efectos fijos ó aleatorios

Cuando se decide continuar usando la base completa y no usar un estimador entre grupos, es necesario analizar la posible existencia de un efecto constante en el término de error. Esto se consigue identificando posibles inconsistencias en una estimación inicial por MCO, que no puedan ser atribuidas a otros problemas del modelo –como heteroscedasticidad-, que indican la necesidad de emplear efectos aleatorios, o alguna metodología de efectos fijos. Si las estimaciones se consideran sin inconsistencias, la estimación por MCO (o MCA siguiendo al capítulo 7), es la apropiada. Para esto, se utiliza la prueba estadística de Breusch y Pagan, que pretende identificar la existencia de un efecto fijo. Esta prueba, expuesta en detalle a continuación es complementaria al análisis de inconsistencias, y no criterio único para rechazar el uso de mínimos cuadrados ordinarios.

8.5.2.1 Prueba de Breusch-Pagan

La prueba de Breusch y Pagan consiste en identificar la existencia de autocorrelación residual entre los términos de error de un modelo estimado por MCA, en distintos momentos del tiempo; lo anterior, bajo datos longitudinales, es equivalente a probar la existencia de efectos constantes en el término de error. La prueba consiste en,

1. Realizar la estimación del modelo a estudiar por MCO.

2. Obtener los errores calculados $\hat{\mu}_{it}$, dada la regresión en 1.

3. Construir el estimador de Lagrange (LM) y verificar el resultado de la prueba de hipótesis.

La prueba de hipótesis, viene expresada en la ecuación 8.21, y el estadístico de prueba en 8.22. Cuando el valor del estimador sea mayor a aquel reportado en la tabla de valores críticos de la distribución χ^2 con un grado de libertad, bajo el nivel de significancia deseado, se rechaza la hipótesis nula. En ese caso se confirma la existencia de un componente fijo en el error, y es necesario aplicar efectos aleatorios, o alguna metodología de efectos fijos. Si por el contrario no es posible rechazar la hipótesis nula, se asume no existe un término fijo en el error y se utiliza MCO (Greene, 2008, 205-208).

$$\begin{array}{ll} H_0 : \sigma_u^2 = 0 \rightarrow Corr(\sigma_{ut,}^2 \sigma_{us}^2) = 0 & \text{No hay evidencia de efectos constantes en el error. Usar MCO.} \\ H_1 : \sigma_u^2 \neq 0 \rightarrow Corr(\sigma_{ut,}^2 \sigma_{us}^2) \neq 0 & \text{Hay evidencia de efectos constantes en el error, elegir entre EA y EF.} \end{array} \qquad (8.21)$$

$$LM = \frac{nT}{2(T-1)} \left[\frac{\sum_{i=1}^{n} \left[\sum_{t=1}^{T} e_{it} \right]^2}{\sum_{i=1}^{n} \left[\sum_{t=1}^{T} e_{it}^2 \right]^2} - 1 \right]^2 \sim \chi^2_1 \qquad (8.22)$$

Sin embargo, para casos donde no se logra demostrar la existencia de efectos aleatorios, se realizan estimaciones por medio de MCO. De lo contrario, se continúa el proceso de identificación comparando entre efectos aleatorios y efectos fijos.

8.5.3 Elección entre efectos aleatorios y efectos fijos

Como se discutió en la sección 8.4, en los problemas donde existe correlación entre el término fijo del error y al menos una de las variables independientes, se debe aplicar una de las metodologías de efectos fijos. Cuando por el contrario, no hay problema de endogeneidad, es conveniente aplicar el estimador de efectos aleatorios. Solucionar este dilema es equivalente a la disyuntiva entre aplicar MCO o MC2E en un problema de corte transversal (*véase* capítulo 1). Por esto, debe solucionarse con una prueba de Hausman.

8.5.3.1 Prueba de Hausman

Para elegir entre estimadores de efectos aleatorios y fijos, se utiliza la prueba de Hausman discutida en los capítulos 1 y 2, que plantea que una desigualdad estadística entre los estimadores indica la existencia de endogeneidad (*véase* prueba de hipótesis 8.23).

$$\begin{array}{ll} H_0 : \beta_{EA} \approx \beta_{EF} & \text{No hay evidencia de endogeneidad.} \\ & \text{Usar Efectos Aleatorios.} \\ H_1 : \beta_{EA} \neq \beta_{EF} & \text{Hay evidencia de endogeneidad.} \\ & \text{Usar Efectos Fijos.} \end{array} \qquad (8.23)$$

En términos generales, la prueba consiste en:

1. Realizar la estimación del modelo a estudiar por efectos aleatorios.

2. Realizar la estimación por alguna de las metodologías de efectos fijos.

3. Construir el estimador de Hausman y verificar el resultado de la prueba de hipótesis.

Los primeros dos pasos consisten en estimar el modelo por efectos aleatorios y por alguna metodología de efectos fijos (véanse secciones 8.3.2 y 8.3.3). Con el valor de

estos estimadores, se construye el estadístico de prueba, definido en la ecuación 8.24.

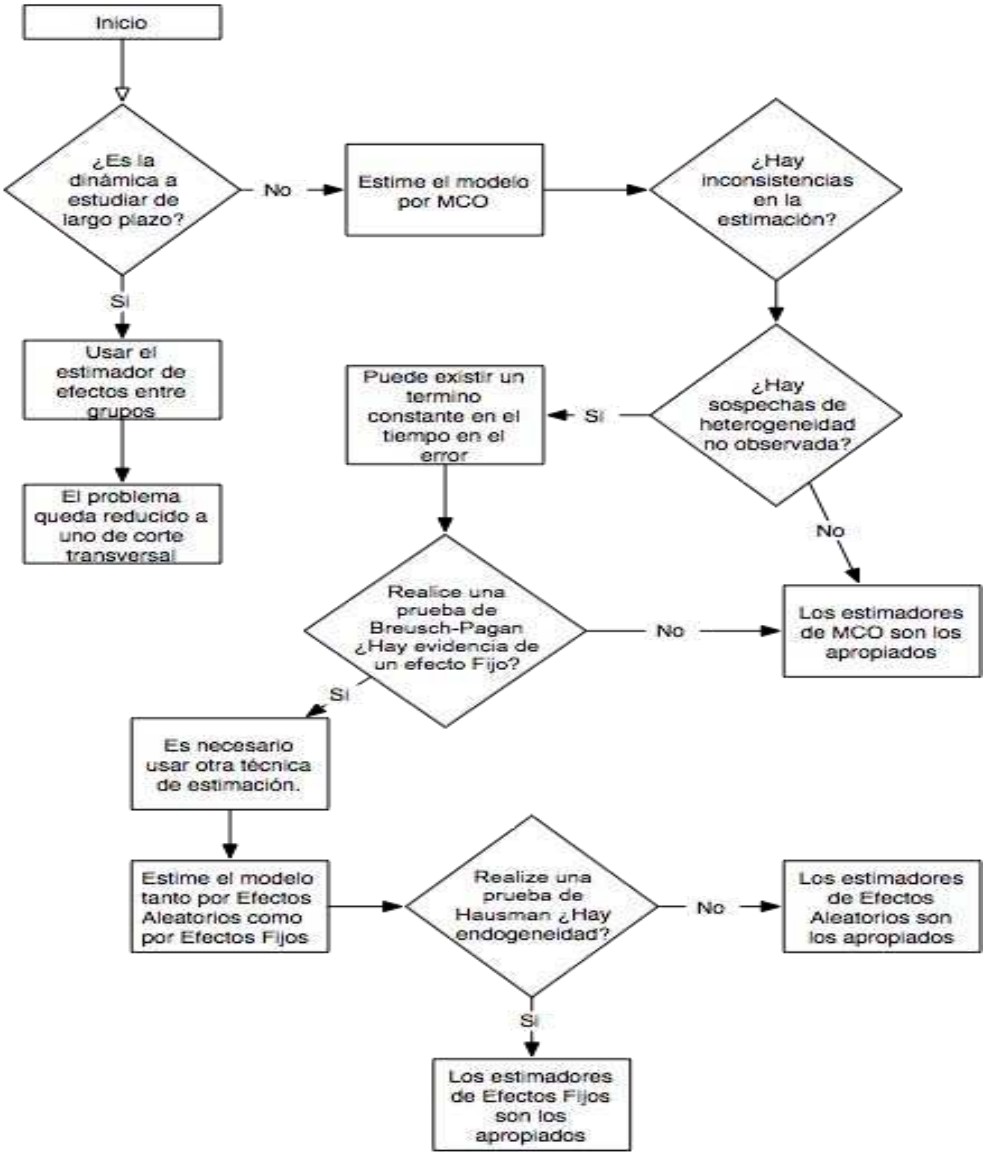
$$H = \frac{(\hat{\beta}_{EF} - \hat{\beta}_{EA})^2}{\text{var}\left[\hat{\beta}_{EF} - \hat{\beta}_{EA}\right]} \sim \chi^2{}_k \qquad (8.24)$$

H es el estadístico de Hausman; $\hat{\beta}_{EF}$ corresponde a los estimadores de efectos fijos, y $\hat{\beta}_{EA}$ a los de efectos aleatorios. Cuando el valor del estadístico es mayor al valor que se reporta en la tabla de valores críticos de la distribución χ^2, queda rechazada la hipótesis nula. En ese caso, se afirma la existencia de un problema de endogeneidad y resulta necesario aplicar una de las metodologías de efectos fijos. Si por el contrario no es posible rechazar la hipótesis nula, es posible asumir que no hay ningún sesgo relevante, y conviene usar efectos aleatorios (Greene, 2008, 208-209).

8.5.4 Resumen del proceso de identificación

A continuación se presenta un esquema que resume el proceso de identificación del estimador apropiado, con todos los conceptos presentados a lo largo de esta sección (*véase* grafica 8.4).

Gráfica 8.4. Esquema de identificación del estimador idóneo para un problema de datos longitudinales.



Fuente: los autores

8.6 Estudio de caso: informalidad regional en Colombia

El estudio de caso desarrollado a continuación, está basado en el artículo titulado "*Informalidad regional en Colombia. Evidencia y determinantes*" de García Cruz (2008), que pretende estudiar los diferenciales regionales en el grado de informalidad laboral en Colombia, utilizando datos de tipo longitudinal. En particular, identifica la relación entre informalidad, importancia relativa del sector industrial, y grado de burocratización del Estado.

En ese estudio, el autor describe cómo en Colombia gran proporción de la población trabajadora se encuentra desempleada o trabajando fuera de la formalidad. Usando datos del Departamento Administrativo Nacional de Estadística –DANE- , se encuentra que para el 2006, cerca de seis de cada diez trabajadores colombianos se encuentran laborando en la informalidad.

Para medir la informalidad, se utilizan dos indicadores. El primero sigue la definición DANE, que entiende por informalidad a los trabajadores que laboran por cuenta propia (no profesionales ni técnicos), aquellos empleados en el servicio doméstico, los trabajadores familiares sin remuneración, y los empleadores y empleados en empresas de hasta diez trabajadores. La segunda definición, proveniente de diferentes trabajos disponibles en la literatura internacional, asocia la informalidad con la ausencia de seguridad social en salud, pensión o del salario mínimo vigente como ingreso laboral. Aquí, se realizará el análisis únicamente con el primer indicador.

Las fuentes de información son los módulos de informalidad aplicados por el DANE en la Encuesta Nacional de Hogares (ENH) y en la Encuesta Continua de Hogares (ECH) en los meses de junio para el período 1988-2006. Los indicadores construidos corresponden a series bianuales de 1988 hasta el 2000 -con excepción de 1990-, y anuales desde el 2001 hasta el 2006, para un conjunto de diez áreas metropolitanas (Barranquilla, Bogotá, Bucaramanga, Cali, Cúcuta, Manizales, Medellín, Pasto, Pereira y Villavicencio).

Para capturar las diferencias locales sobre la informalidad laboral, se hace un análisis con regresiones tipo panel. La variable dependiente corresponde a uno de los dos indicadores de informalidad, y las variables independientes son el grado de desarrollo industrial, y el grado de burocratización el cual intenta capturar el elemento institucional (véase ecuación 8.30).

$$TI_{it} = \beta o_i + \beta_1 PPIB_{it} + \beta_2 Gasto_{it} + e_{it} \qquad (8.30)$$

En la ecuación 8.30, TI_{it} corresponde a la tasa de informalidad en la región i para el año t. $PPIB_{it}$ es el grado de desarrollo industrial, medido como la participación porcentual del PIB industrial sobre el PIB departamental. $Gasto_{it}$ corresponde al gasto en nómina medido en términos per cápita, y pretende capturar la ineficiencia del Estado.

En la literatura se ha documentado que ciudades con mayor desarrollo industrial, de mayor tamaño, con mercados grandes y buena infraestructura, presentan menor crecimiento de las actividades informales; por esta razón, el autor espera que la variable desarrollo industrial tenga una relación inversa con el grado de informalidad laboral. Respecto a la variable de eficiencia estatal, espera una relación directa, pues una mayor burocracia desincentiva el trabajo en el sector formal.

8.6.1 Análisis general de los datos

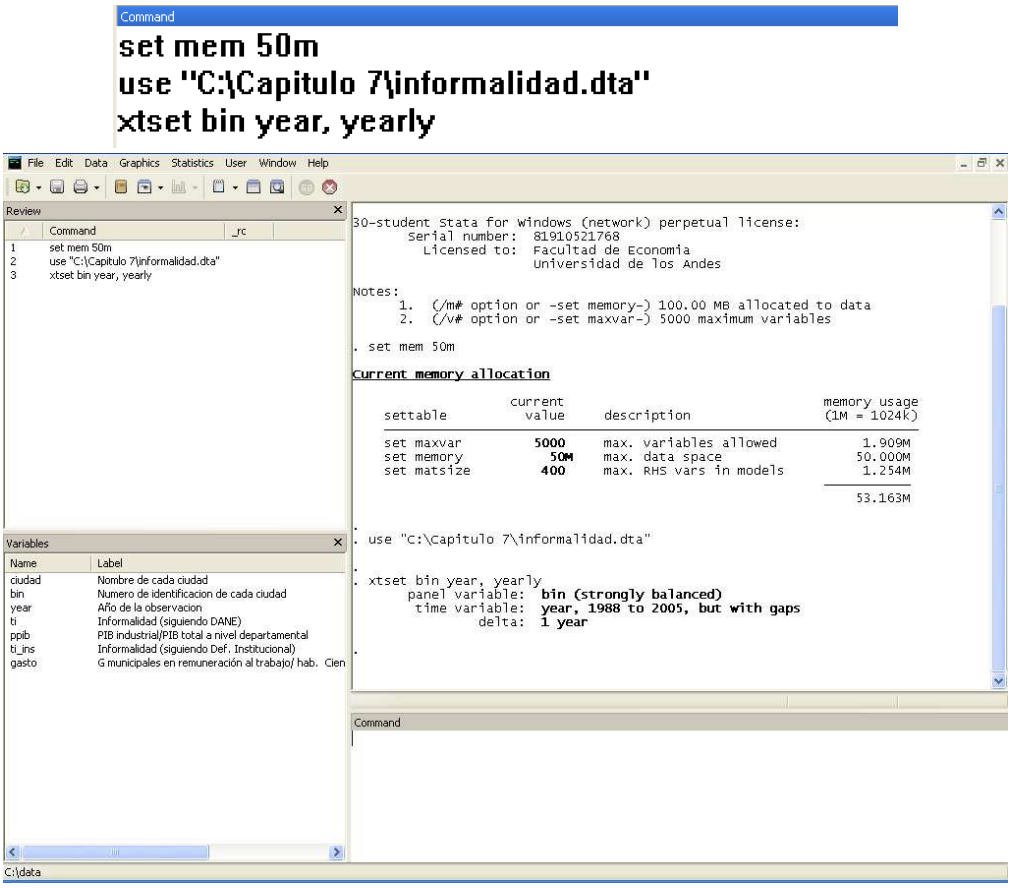
Esta primera sección, expone los comandos que preparan el programa computacional para el análisis econométrico con datos longitudinales, y presenta una exploración general de la base de datos a usar, y las variables relevantes.

1. Para realizar el análisis en Stata®, se debe determinar la memoria –con el comando *setmem*- y cargar la base de datos –con el comando *use*-. En este caso, la base lleva el nombre de *informalidad.dta*.

2. Como este programa viene predeterminado para trabajar con datos de corte transversal, también es necesario revelar que la base de datos es un panel.

Esto se consigue a través del comando *xtset* indicando la variable que identifica a cada uno de los individuos, y la que mide el paso del tiempo. Adicionalmente es conveniente especificar la frecuencia en que están registrados los datos; en este caso con la opción *yearly*, que indica datos anuales (*véase* figura 8.1).

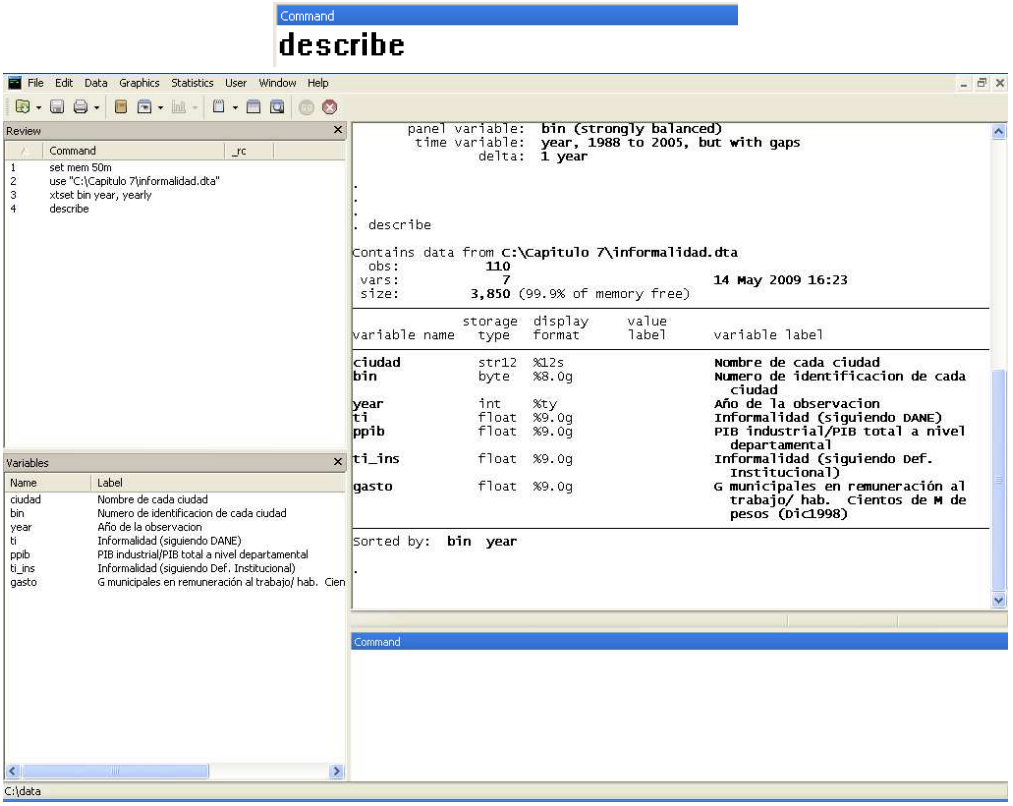
Figura 8.1. Salida comandos setmem, use y xtset



Fuente: cálculo autores

3. La tabla de variables disponibles puede verse con el comando *describe*. En este caso, las variables a usar corresponden a *ti* (la tasa de informalidad), *ppib* (la proxy al desarrollo industrial) y *gasto* (que mide eficiencia estatal) (*véase* figura 8.2).

Figura 8.2. Salida comando describe



Fuente: cálculo autores

4. Para el análisis de datos longitudinales existen varios comandos descriptivos adicionales. Por ejemplo, el comando *xtdescribe* muestra la disponibilidad de datos para cada unidad de corte transversal. Esto es particularmente útil para analizar si el panel a usar está balanceado. A continuación se resumen algunos comandos descriptivos útiles (*véase* cuadro 8.1)

Cuadro 8.1 Transformación de variables de tiempo

Comando	Descripción
xtdescribe	Análisis de desbalance del panel
xtsum	Estadísticas descriptivas
xttab	Tabla de frecuencias
xtline	Gráfico de series de tiempo por individuo

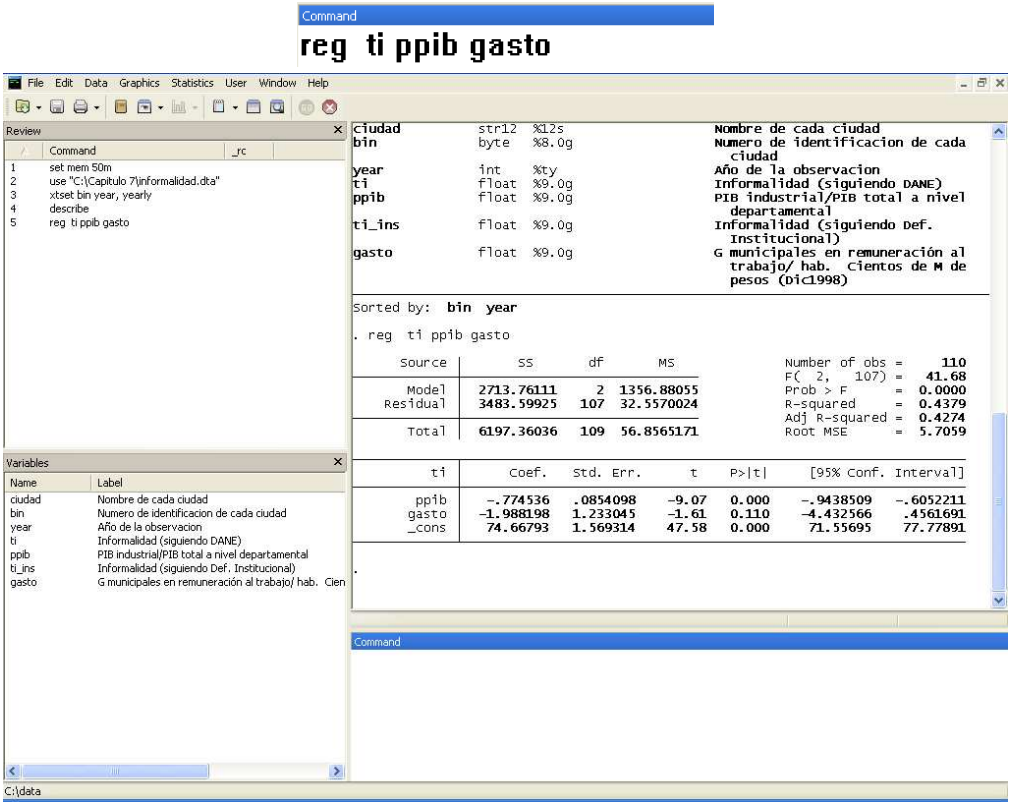
Fuente: los autores, a partir de Cameron y Trivedi (2009).

8.6.2 Estimaciones e identificación del modelo apropiado

Para identificar el modelo a usar, se sigue el esquema presentado en la sección 8.4.1. Esta metodología permite deducir cuál es el modelo econométrico apropiado para un ejercicio particular. En este caso no interesa reducir el problema a uno de corte transversal, por lo que directamente se estima el modelo por mínimos cuadrados ordinarios.

1. La estimación por MCO, es equivalente a las efectuadas en capítulos anteriores, sin realizar diferenciaciones por individuo. Esto se consigue usando el comando *reg* seguido por la variable dependiente *ti*, e independientes *ppib y gasto* (*véase* figura 8.3).

Figura 8.3. Salida regresión por mínimos cuadrados agrupados



Fuente: cálculo autores

En la figura 8.3 la variable que captura el desarrollo industrial, tiene el signo esperado (negativo) y es significativa al explicar la tasa de informalidad con un estadístico t de 9.07, el cual equivale a una significancia mayor al 1%. En cambio, el gasto en nómina medido en términos per cápita en cambio, resulta no ser significativo con un t estadístico de 1.61-y presenta un signo contrario al esperado. Esta inconsistencia hace pensar, que la heterogeneidad no observada entre ciudades puede estar sesgando el los resultados del modelo.

Para estimar una regresión siguiendo las metodologías propuestas en este capítulo se usa el comando *xtreg*, que bajo diversas especificaciones puede calcular los estimadores de efectos aleatorios y fijos entre y al interior de grupos (*véase* cuadro 8.2). Las estimaciones por primeras diferencias deben realizarse manualmente.

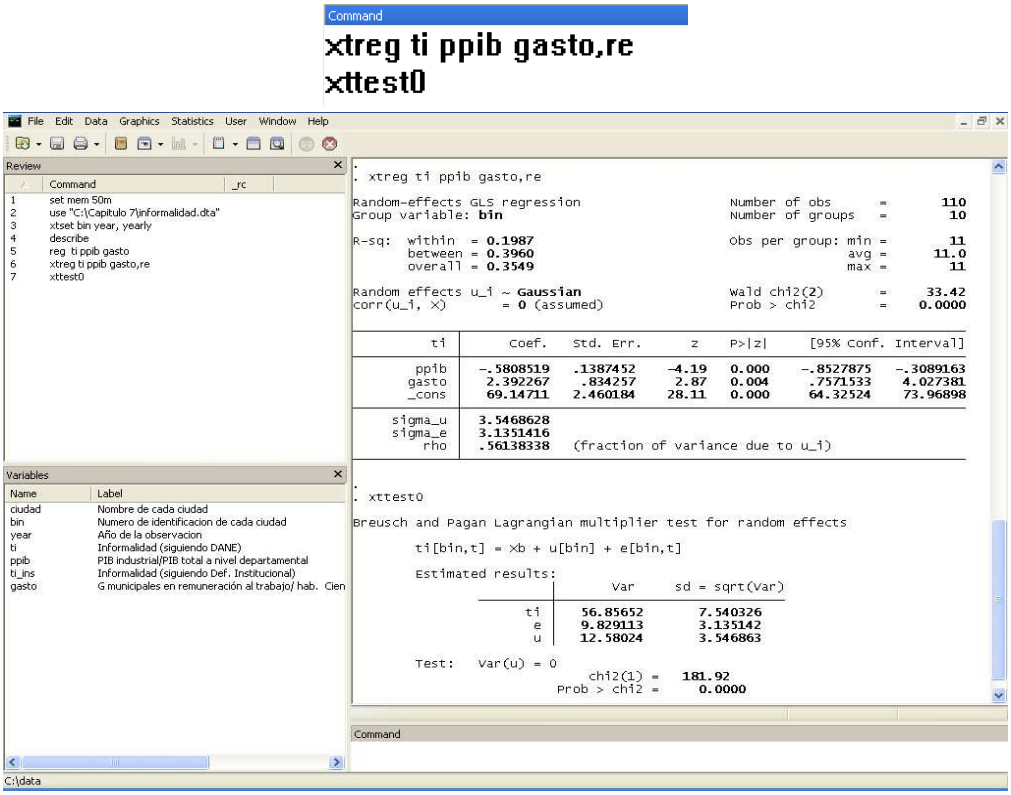
Cuadro 8.2 Casos particulares del comando xtreg

Descripción	Comando
Estimador de efectos entre grupos	xtreg ti ppib gasto, be
Estimador de efectos aleatorios	xtreg ti ppib gasto, re
Estimador de efectos fijos al interior de grupos	xtreg ti ppib gasto, fe

Fuente: los autores, con base a Cameron y Trivedi (2009).

2. Siguiendo el esquema de identificación, el siguiente paso consiste en realizar una prueba de Breusch-Pagan para probar la existencia de correlación entre los términos de error del modelo. Para esto, se realiza una regresión de efectos aleatorios, y se prueba la evidencia de autocorrelación entre los términos de error con el comando *xttest0* (*véase* figura 8.4).

Figura 8.4. Salida regresión por efectos Aleatorios y prueba de Breusch-Pagan



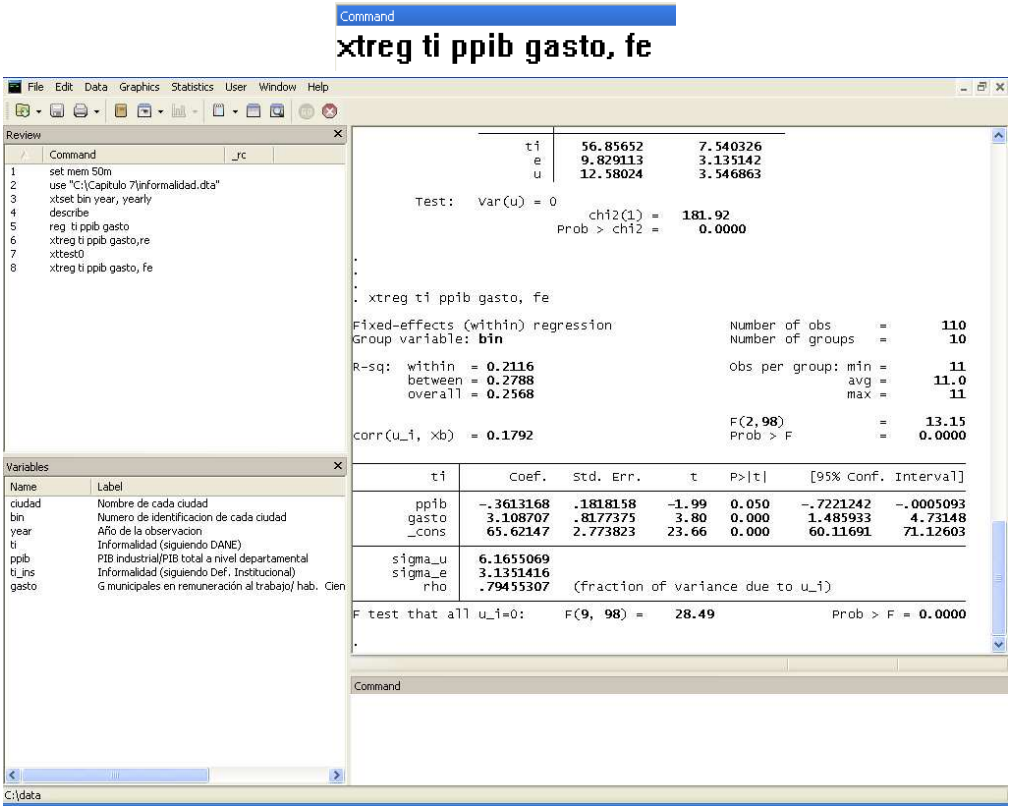
Fuente: cálculo autores

La salida de la regresión por efectos aleatorios, que se observa en la figura 8.4, tiene la misma estructura de aquella que arroja una regresión lineal simple, aunque muestra algunas estadísticas adicionales. En la parte superior, se muestra el número de observaciones totales (nT) así como el número de individuos o grupos (n). En lugar del estadístico F, se presenta un estadístico análogo (χ^2) que se interpreta de la misma manera. Para el estadístico de bondad de ajuste (R^2), se muestran tres diferentes medidas que indican cómo explica el modelo, la varianza al interior de cada individuo (*within*), entre individuos (*between*) y de manera general (*overall*). Los estadísticos *sigma_u* y *sigma_e* corresponden respectivamente a un estimador de la desviación estándar del componente constante del error y del error de ruido blanco tradicional.

Con respecto a la prueba de Breusch-Pagan, en este caso se rechaza la hipótesis nula al 1%, lo que muestra evidencia estadística de heterogeneidad no observada en el término error. Esto implica la necesidad de usar alguna de las metodologías que tengan en cuenta la existencia de efectos constantes en el tiempo. En la estimación por efectos aleatorios, tanto la variable que captura el desarrollo industrial como el gasto en nómina tienen el signo esperado, y son significativas al 1% al explicar la tasa de informalidad.

3. La regresión por efectos fijos al interior de grupos se realiza tambien con el comando *xtreg*, pero con la opción *fe* en lugar de *re*. Para esta estimación, ambas variables independientes tienen el signo esperado y son significativas -con estadísticos t de -1.99 y 3.80- (*véase* figura 8.5).

Figura 8.5. Salida regresión por efectos fijos al interior de grupos



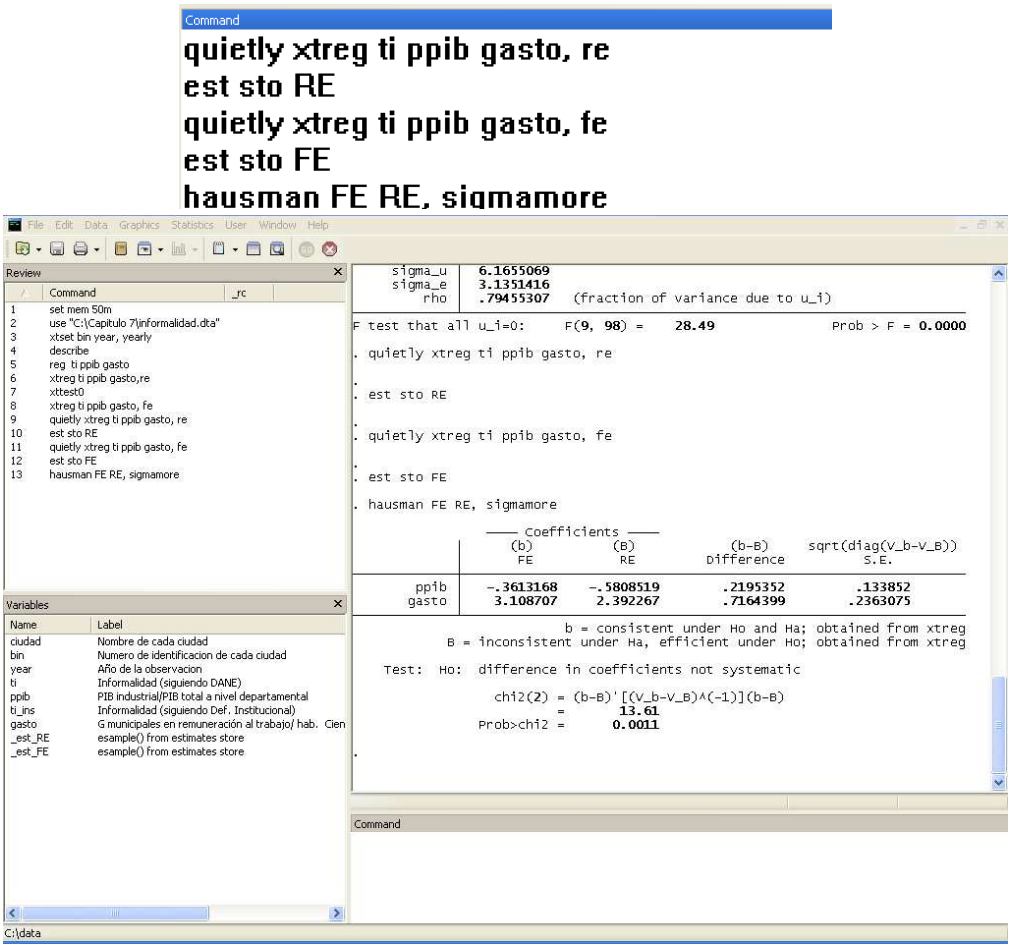
Fuente: cálculo autores

4. El último paso para identificar el estimador apropiado, es realizar una prueba de Hausman para probar la endogeneidad resultante de la correlación de una de las variables independientes con el componente fijo del término de error. Esta prueba se realiza en dos pasos. Primero, se estiman una vez más las regresiones por efectos aleatorios y efectos fijos, pero guardando en cada caso los estimadores en la memoria del programa. Esto se consigue ejecutando el comando *estimates store –o est sto-* después de cada regresión.

   A continuación, se ejecuta la prueba con el comando *hausman* -utilizando la opción *sigmamore-*, indicando a continuación el nombre de las variables en donde se guardaron los estimadores. La variable que guarda los resultados del estimador por efectos fijos debe ir primero, y luego la de efectos aleatorios. En este caso se utilizó el comando *quietly* en cada una de las

estimaciones para que Stata® no muestre en pantalla los resultados. Los estimadores de las regresiones se guardaron como RE y FE (*véase* figura 8.6).

Figura 8.6. Salida prueba de Hausman



Fuente: cálculo autores

Esta prueba de Hausman rechaza la hipótesis nula de estimadores de efectos aleatorios consistentes, a una significancia del 1%, lo que muestra que hay una fuerte evidencia estadística de endogeneidad. En este sentido, el estimador adecuado es el de efectos fijos al interior de grupos cuyos resultados se encuentran en la figura 8.5.

Resumen

- Los paneles de datos –o bases longitudinales- contienen información sobre la evolución de un conjunto de unidades de corte transversal a lo largo del tiempo, a las que a los que se les hace un seguimiento periodo a periodo.
- El uso de datos longitudinales es útil, ya que mejora la precisión de las estimaciones, permite separar el componente de largo plazo de la de corto plazo en las observaciones, posibilita eliminar problemas de endogeneidad resultante de variables no observables fijas, y hace posible analizar la dinámica de un conjunto de variables en el tiempo.
- Las bases longitudinales pueden dividirse en paneles micro que registran información sobre hogares o firmas y los macro que siguen a un conjunto de regiones o países. Adicionalmente, se habla de paneles balanceados, donde los datos están completos y reportados con una temporalidad constante, y desbalanceados.
- Para probar empíricamente modelos teóricos de largo plazo, la metodología de estimación de efectos entre grupos, permite reducir el problema de un panel longitudinal a uno de corte transversal de largo plazo.
- Un problema particular de las estimaciones usando datos longitudinales, es la correlación serial entre los términos de error de los diferentes periodos. Para solventar este problema se usa la estimación de efectos aleatorios.
- Adicionalmente, es posible que la correlación entre efectos fijos y variables creen un problema de endogeneidad. Para solucionarlo, es posible utilizar variables dicótomas que capturen el efecto de la heterogeneidad no observada, o metodologías de efectos fijos -primeras diferencias o estimador de efectos entre grupos-, que transformando el modelo inicial, eliminan la endogeneidad.
- Para lograr establecer qué modelo es el adecuado, es necesario seguir un criterio de selección. La prueba de Breusch-Pagan se usa para identificar si el efecto constante del término de error causa correlación serial. Para establecer si existe endogeneidad, se usa la prueba de Hausman.

Apéndice

Manual comandos Stata®

A.1 Introducción

Este anexo, resume los diversos comandos del paquete estadístico Stat® utilizados a lo largo del presente libro. Esta pensado como un manual de referencia tanto para estudiantes que se inician en la aplicación de técnicas econométricas, como para usuarios experimentados que puedan no tener presente la sintaxis exacta con que se da una orden en particular en este programa computacional.

Para mantener la coherencia con el resto del libro, los diferentes comandos vienen organizados por tema, cubriendo separadamente variables instrumentales y ecuaciones simultáneas, modelos con variable dependiente dicótoma, series de tiempo, y panel de datos. Adicionalmente se presenta una lista de comandos generales, cuyo uso suele ser frecuente en los trabajos investigativos de economía.

Cada uno de los comandos se presenta en una tabla, donde se da una breve descripción de su función, el tipo de variables que se deben incluir, y una lista de opciones comunes. Adicionalmente se presenta la sintaxis en la forma de un ejemplo, que debería permitir al lector transponer con facilidad el comando expuesto al trabajo con una base de datos particular. La notación a usar en estos casos se describe a continuación.

Comandos y opciones	Ordenes que comprenda el paquete estadístico. Pueden ser comandos u opciones.
x1, x2, x3, x4.	Variables que se encuentran en la base de datos, y que serán usadas como variables independientes en los modelos econométricos.
y1, y2, y3,	Variables que se encuentran en la base de datos, y que serán usadas como variables independientes en los modelos econométricos.

A.2 Comandos generales

Comando	**Función**	**Variables a incluir**	**Opciones principales**
set mem	Especifica cuántos megabytes (MB) de memoria usara Stata® para este proyecto.	Ninguna.	La opción *perm*, hace al cambio permanente, lo cual cambia la memoria predeterminada. Stata® iniciara todos los proyectos futuros con esta cantidad de memoria.
Ejemplo: *set mem 50M, perm*			

Comando	Función	Variables a incluir	Opciones principales
gen	Permite crear una nueva variable.	Es necesario incluir el nombre de la nueva variable. Se incluyen otras para definir el contenido de la nueva variable.	Ninguna.
Ejemplo: *gen x3 = x2 – x1*			

Comando	Función	Variables a incluir	Opciones principales
drop	Permite eliminar una variable	La o las variables a eliminar.	Ninguna.
Ejemplo: *drop x2 x3*			

Comando	Función	Variables a incluir	Opciones principales
replace	Permite remplazar los datos contenidos en una variable.	La variable a modificar. Otras se incluyen, cuando en base a estas se va a crear el nuevo contenido.	Ninguna.
Ejemplo: *gen x3 = 0* *replace x3 = x2 – x1*			

Comando	Función	Variables a incluir	Opciones principales
for var	Permite repetir el mismo procedimiento para un conjunto de variables.	La lista de variables a las que se aplica el mismo procedimiento. *X* corresponde a la variable general.	Ninguna.
Ejemplo: *for var x1 x2 x3: gen X_sq = X^2*			

Comando	Función	Variables a incluir	Opciones principales
des	Muestra la descripción de cada variable y su formato.	Opcionalmente indicar una lista de variables a describir. Si no se especifica ninguna, se describen todas las disponibles.	Ninguna.
Ejemplo: *des x1 x2 x3 y1 y2 y3*			

Comando	Función	Variables a incluir	Opciones principales
sum	Presenta un resumen de estadísticas descriptivas de las variables especificadas.	Opcionalmente indicar una lista de variables. Si no se especifica ninguna, se muestra una tabla para todas las disponibles.	La opción *detail*, hace que se presenten estadísticas adicionales.
Ejemplo: *sum x1 x2 x3, detail*			

Comando	Función	Variables a incluir	Opciones principales
tab	Muestra una tabla con las frecuencias de una variable cualitativa.	La variable cualitativa de interés.	La opción *missing*, hace que se presenten los valores no disponibles. La opción *gen (a)*, crea una variable dummy para cada categoría. El valor en paréntesis (*a*) corresponde a un sufijo que tendrán las variables creadas.
Ejemplo: *tab x1, missing gen(a)*			

Comando	Función	Variables a incluir	Opciones principales
tab	Muestra una tabla cruzada entre dos variables cualitativas.	Las variables cualitativas de interés.	Ninguna.
Ejemplo: *tab x1 x2*			

Comando	Función	Variables a incluir	Opciones principales
display	Muestra el resultado de un cálculo matemático.	Ninguna.	Ninguna.
Ejemplo: *display 125*2*			

Comando	Función	Variables a incluir	Opciones principales
corr	Calcula una matriz de correlación o de covarianza entre dos variables.	Indicar las variables a usar.	La opción *cov*, hace que presente la matriz de varianza-covarianza. Sin esta opción, Stata® presenta la matriz de correlaciones.
Ejemplo: *corr x1 x2, cov*			

Comando	Función	Variables a incluir	Opciones principales
reg	Ejecuta una regresión por MCO	Indicar la variable dependiente seguida por las variables independientes. Incluir constante al final en el caso en que se utilice la opción *hascons.*	La opción *nocons*, hace que se realice las estimaciones sin constante. La opción *hascons*, hace que use una constante entregada por el usuario. La opción *robust* hace que se estime la varianza con el estimador de White.
Ejemplo: *reg y x1 x2 x3, robust*			

Comando	Función	Variables a incluir	Opciones principales
predict	Predice los valores ajustados, o los errores después de un modelo.	Una nueva variable que guardara los valores ajustados.	La opción *resid* hace que el comando calcule los errores, en lugar de los valores calculados de la variable dependiente.
Ejemplo *reg y x1 x2 x3* *predict, resid*			

Comando	Función	Variables a incluir	Opciones principales
ssc install	Instala un componente adicional.	Ninguna.	Ninguna.
Ejemplo: *ssc install ivreg2* *ssc install hprescott*			

A.3 Especificación, endogeneidad y simultaneidad

Comando	Función	Variables a incluir	Opciones principales
estat ovtest	Ejecuta una prueba Ramsey-RESET.	Ninguna. El comando únicamente es válido después de una regresión.	Ninguna.
Ejemplo: *reg y x1 x2 x3, robust* *estat ovtest*			

Comando	Función	Variables a incluir	Opciones principales
ivreg	Ejecuta una regresión por Mínimos Cuadrados en dos etapas.	Indicar la variable dependiente seguida por las variables independientes. Las variables endógenas se presentan en paréntesis igualadas a sus respectivos instrumentos. Incluir constante al final en el caso en que se utilice la opción *hascons*.	La opción *first* hace que adicionalmente se presente la primera etapa del modelo. La opción *nocons*, hace que se realice las estimaciones sin constante. La opción *hascons*, hace que use una constante entregada por el usuario. La opción *robust* hace que se estime la varianza con el estimador de White.
Ejemplo: *ivreg y x1 x2 (x3 = z1), first robust*			

Comando	Función	Variables a incluir	Opciones principales
ivreg2	Ejecuta una regresión por Mínimos Cuadrados en dos etapas y presenta pruebas adicionales	Indicar la variable dependiente seguida por las variables independientes. Las variables endógenas se presentan en paréntesis igualadas a sus respectivos instrumentos. Incluir constante al final en el caso en que se utilice la opción *hascons.*	La opción *first* hace que adicionalmente se presente la primera etapa del modelo. La opción *nocons*, hace que se realice las estimaciones sin constante. La opción *hascons*, hace que use una constante entregada por el usuario. La opción *robust* hace que se estime la varianza con el estimador de White.
Ejemplo: *Ivreg2 y x1 x2 (x3 = z1), first robust*			

Comando	Función	Variables a incluir	Opciones principales
reg3	Estima un modelo de ecuaciones simultaneas	Indicar una a una las diferentes ecuaciones del modelo en paréntesis. Cada ecuación se escribe empezando por la variable dependiente seguida de las independientes.	Las opciones *ols, 2sls* y *sure* obliga al comando a realizar las estimaciones por MCO, MC2E o SUR. La opción *first* obliga al comando a presentar la primera etapa del modelo. Las opciones *nocons* y *hascons*, se utilizan dentro de cada ecuación, para que se realicen las estimaciones sin constante o con una constante determinada. Las opciones *endog* y *exog* permiten especificar las variables endógenas y exógenas.
Ejemplo: *reg3 (y1 x1 x2 x3) (y2 x4 x5) (y3 x1 x5, nocons), endog (y1, y2, y3)*			

Comando	Función	Variables a incluir	Opciones principales
sureg	Estima un modelo SUR.	Indicar una a una las diferentes ecuaciones del modelo en paréntesis, escribiendo cada ecuación de la manera estándar.	Las opciones *nocons* y *hascons*, se utilizan dentro de cada ecuación, para que se realicen las estimaciones sin constante o con una constante determinada. La opción *constraints(constraints)* permite especificar restricciones. La opción *small*, se utilizan para muestras pequeñas.
Ejemplo: *sureg (y1 x1 x2 x3) (y2 x4 x5)*			

Comando	Función	Variables a incluir	Opciones principales
est store	Guarda el conjunto de estimadores de una regresión.	Ninguna. El comando únicamente es válido después de una regresión.	Ninguna.
Ejemplo: *reg y x1 x2 x3* *est store MCO*			

Comando	Función	Variables a incluir	Opciones principales
hausman	Ejecuta una prueba de Hausman.	Los estimadores obtenidos en las regresiones de MC2E y MCO.	La opción *sigmamore* obliga al comando a calcular los errores estimados a partir del modelo de MCO. Recomendado para una prueba Hausman de endogeneidad.
Ejemplo: *reg y x1 x2 x3* *est store MCO* *ivreg y x1 x2 (x3 = z1)* *est store MC2E* *hausman MC2E MCO, sigmamore*			

Comando	Función	Variables a incluir	Opciones principales
overid	Ejecuta una prueba de restricciones sobre identificadas.	Ninguna. El comando únicamente es válido después de una regresión.	La opción *chi2* obliga al comando a calcular la prueba usando un χ^2. Es el default. La opción *f* obliga al comando a calcular la prueba usando un pseudo-F. La opción *all* hace que se reporten 5 versiones del estadístico automaticamente.
Ejemplo: *ivreg y x1 x2 (x3 = z1)* *overid, all*			

A.4 Modelos de probabilísticos: lineal, probit y logit.

Comando	Función	Variables a incluir	Opciones principales
probit	Estima un modelo Probit.	Indicar la variable dependiente dicótoma seguida por las variables independientes.	La opción *nocons*, hace que se realice las estimaciones sin constante. La opción *robust* hace que se estime la varianza con el estimador de White.
Ejemplo: *probit y x1 x2 x3*			

Comando	Función	Variables a incluir	Opciones principales
logit	Estima un modelo Logit.	Indicar la variable dependiente dicótoma seguida por las variables independientes.	La opción *nocons*, hace que se realice las estimaciones sin constante. La opción *robust* hace que se estime la varianza con el estimador de White.
Ejemplo: *logit y x1 x2 x3*			

Comando	Función	Variables a Incluir	Opciones Principales
mfx	Estima los efectos marginales.	Ninguna. El comando únicamente es válido después de una regresión.	La opción *varlist(x1 x2)*, hace que calculen los efectos marginales únicamente para las variables x1 y x2. Las opciones *eyex*, *dyex* y *eydx*, se utilizan para especificar que se desean elasticidades en la forma d(lny)/d(lnx), d(y)/d(lnx) o d(lny)/d(x), respectivamente.
Ejemplo: *logit y x1 x2 x3* *mfx*			

A.5 Series de tiempo

Comando	Función	Variables a incluir	Opciones principales
tsset	Declara una base de datos como serie de tiempo.	Una variable que registre el paso del tiempo.	Las opciones *daily, weekly, monthly, quarterly, halfyearly* y *yearly*, permiten especificar la frecuencia en que se encuentran los datos.
Ejemplo: *tsset year, yearly*			

Comando	Función	Variables a incluir	Opciones principales
tsline	Realiza una grafica de una o más variables con respecto al tiempo.	Las variables de interés. Todas se mostraran simultáneamente, en un mismo gráfico.	Ninguna.
Ejemplo: *tsline y1 x1*			

Comando	Función	Variables a Incluir	Opciones Principales
hprescott	Aplica un filtro de Hodrick-Prescott a una variable determinada.	Las variables de interés.	La opción *stub(a)*, especifica que las nueva series se guardaran en una variable con el sufijo *a*. Para este comando siempre es necesario aplicar esta opción. La opción *smooth* define el parámetro de suavizamiento.
Ejemplo: *hprescott y, stub(a) smooth(6.25)*			

Comando	Función	Variables a Incluir	Opciones Principales
tssmooth	Calcula una nueva serie a través de una metodología de suavizamiento.	En primer lugar es necesario incluir el tipo de suavizamiento que se desea realizar. *ma* corresponde a un promedio móvil, *exponential* a una atenuación simple, d*exponential* a una atenuación doble, y hwinters y shwinters a suavizamientos Holt-Winters. Adicionalmente es necesario añadir el nombre de una nueva variable donde se guardara la serie suavizada. Por último la variable a suavizar, después de un signo de igualdad.	La opción *coef(a,b)*, especifica los parámetros a usar.
Ejemplo: *tssmooth dexponential y_ad=y*			

Comando	Función	Variables a Incluir	Opciones Principales
dfuller	Ejecuta una prueba dikey-fuller de estacionariedad.	Las variables de interés.	La opción *trend*, incluye una tendencia en la prueba. La opción *regress*, muestra la tabla de regresión. La opción *lags(a)*, especifica el numero de rezagos. La opción *drift*, incluye un termino de desviación. La opción *nocons*, elimina la constante.
Ejemplo: *dfuller y, trend*			

Comando	Función	Variables a Incluir	Opciones Principales
ac	Muestra un grafico de autocorrelacion simple.	Las variables de interés.	Ninguna.
Ejemplo: *ac y*			

Comando	Función	Variables a Incluir	Opciones Principales
pac	Muestra un grafico de autocorrelacion parcial.	Las variables de interés.	Ninguna.
Ejemplo: *pac y*			

Comando	Función	Variables a Incluir	Opciones Principales
Corrgram	Muestra una tabla de correlaciones	Las variables de interés.	Ninguna.
Ejemplo: *corrgram y x1 x2*			

Comando	Función	Variables a Incluir	Opciones Principales
arima	Estima un modelo ARIMA.	Las variables de interés.	La opción *arima(p,i,q)*, especifica el numero de términos autorregresivos, de media móvil y el orden de integración. La opción *sarima(p,i,q,s)*, especifica el número de términos autorregresivos, de media móvil, el orden de integración y la periodicidad estacional
Ejemplo: *arima y, arima(1,0,1)* *arima y, arima(1,0,1) sarima(0,0,1,12)*			

Comando	Función	Variables a Incluir	Opciones Principales
armaroots	Muestra las raíces del polinomio característico.	Ninguna El comando únicamente es válido después de una regresión arma	Ninguna
Ejemplo: *armaroots*			

Comando	Función	Variables a Incluir	Opciones Principales
tsappend	Adiciona periodos a pronosticar en la base de datos	Ninguna	La opción *add(a)* permite especificar cuántos periodos adicionar.
Ejemplo: *tsappend, add(a)*			

Comando	Función	Variables a Incluir	Opciones Principales
rmse	Calcula la raíz cuadrada del promedio para la suma de errores al cuadrado (RCPSEC)	La serie inicial y la serie pronosticada.	La opción *est* esa la misma muestra y mismos grados de libertad en la última regresión.
Ejemplo: *Rmse y yhat*			

Comando	Función	Variables a Incluir	Opciones Principales
inequal7	Calcula el coeficiente de Theil	La serie pronosticada. ponderada por la variable original	Ninguna
Ejemplo: *inequal7 yhat (weight=y)*			

Comando	Función	Variables a Incluir	Opciones Principales
wntestq	Aplica una prueba de normalidad de los errores.	La serie de interés a evaluar.	La opción *lags(a)*, especifica el numero de rezagos a usar.
Ejemplo: *arima y, arima(1,0,1)* *predict e, resid* *wntesq e*			

Comando	Función	Variables a Incluir	Opciones Principales
jb	Aplica una prueba de normalidad Jarque-Bera.	La serie de interés a evaluar.	Ninguna.
Ejemplo: *arima y, arima(1,0,1)* *predict e, resid* *jb e*			

Comando	Función	Variables a Incluir	Opciones Principales
wntestb	Aplica una prueba de normalidad de los errores.	La serie de interés a evaluar.	La opción *table*, especifica que se desea obtener una tabla, y no una grafica.
Ejemplo: *arima y, arima(1,0,1)* *predict e, resid* *wntesb e*			

Comando	Función	Variables a Incluir	Opciones Principales
estat durbinalt	Aplica una prueba autocorrelación a modelos autorregresivos	Ninguna. El comando únicamente es válido después de una regresión.	Ninguna
Ejemplo: *reg y x1 x2 y_1* *estat durbinalt*			

A.6 Panel de datos

Comando	Función	Variables a Incluir	Opciones Principales
xtset	Declara una base de datos como de datos panel.	Una variable que registre el paso del tiempo. Una variable que identifique a cada individuo	La opción *chi2* obliga al comando a calcular la prueba usando un χ^2. Es el default. La opción *f* obliga al comando a calcular la prueba usando un pseudo-F. La opción *all* hace que se reporten 5 versiones del estadístico automáticamente.
Ejemplo: *xtset id year*			

Comando	Función	Variables a Incluir	Opciones Principales
xtreg	Estima un modelo de datos panel.	Indicar la variable dependiente seguida por las variables independientes.	La opciones *fe re* y *be* hace que se estimen modelos por efectos al interior del grupo, efectos aleatorios y efectos entre grupos. Para efectos fijos o aleatorios, la opción *robust* hace que se estime la varianza con el estimador de White.
Ejemplo: *xtreg y1 x1 x2 x3, fe*			

Comando	Función	Variables a Incluir	Opciones Principales
est store	Guarda el conjunto de estimadores de una regresión.	Ninguna. El comando únicamente es válido después de una regresión.	Ninguna.
Ejemplo: *xtreg y x1 x2 x3, fe* *est store FE*			

Comando	Función	Variables a Incluir	Opciones Principales
xttest0	Aplica una prueba de Breuch-Pagan para efectos aleatorios.	Ninguna. El comando únicamente es válido después de una regresión por efectos aleatorios.	Ninguna.
Ejemplo: *xtreg y x1 x2 x3, re* *xttest0*			

Comando	Función	Variables a Incluir	Opciones Principales
hausman	Ejecuta una prueba de Hausman.	Indicar variable que guarda los estimadores obtenidos en las regresiones de EF y EA (en ese orden).	La opción *sigmamore* obliga al comando a calcular los errores estimados a partir del modelo de MCO. Recomendado para una prueba Hausman de endogeneidad. La opción *constant* obliga al comando a incluir la constante en la comparación.
Ejemplo: *xtreg y x1 x2 x3, re* *est store RE* *ivreg y x1 x2 x3, fe* *est store FE* *hausman FE RE*			

Bibliografía

Alonso, J; Contera, M; y Orozco, B. (2006). Sector Público y Déficit Fiscal. Apuntes de economía, Universidad ICESI.

Arellano, M. (2003) Panel data econometrics. Oxford University Press.

Baltagi, B. H. (2005) Econometric analysis of panel data. Ed 3. J. Wiley & Sons.

Banco Mundial. World Development Indicators.

Bernal, R. (2008). The informal labor in Colombia: Indentification and characterization. Facultad de Economía, Universidad de Los Andes.

Breusch, T., y Pagan, A. (1980) The LM test and its applications to model specification in econometrics. Review of Economic Studies, 47.
Cameron, A. C. y Trivedi, P. (2009) Microeconometrics: Methods and applications. Cambridge.

Catalán, H. Teoría de la cointegración. UNAM.

Chiang, A. (1988) Métodos Fundamentales de Economía Matemática, Ed. 3.
McGraw-Hill.

Davidson, R. y Mackinnon, J. (1999). Foundations of econometrics. Draft.

Davidson, R. y Mackinnon, J. (2004). Econometric theory and methods. Oxford University Press.

Dougherty, C. (2007) Introduction to econometrics. 3ed. Oxford University Press

Enders, W. (2004). Applied econometric time series. Ed 2. New York: John Wiley & Sons

Engle R. E. y Granger W. J. (1991). Long-Run economic relationship: Reading in cointegration, Oxford University Press, New York

Garcia, G. (2008). Informalidad regional en Colombia. Evidencia y determinantes. Revista Desarrollo y Sociedad N° 61. Universidad de Los Andes.

Gourieroux, C. (2000). Econometrics of qualitative dependent variables. Cambridge University Press.

Granger, C. (1993) Forecasting in business and economics. Ed. 2. Elsevier Science.

Granger, C. (2004) Análisis de series temporales, cointegración y aplicaciones. Revista Asturiana de Economía.

Greene, W. (1999) Análisis Econométrico. Ed. 3. Prentice Hall.

Greene, W. (2000) Econometric analysis. Ed. 4. Prentice-Hall

Greene, W. (2003). Econometric analysis. Ed. 5. Pearson Prentice-Hall

Guerrero, Victor (2003) Análisis Estadístico de Series de Tiempo Económicas. Ed 2. Editorial Thomson

Gujarati, D. (2003). Econometría. Ed. 4. Mc Graw Hill.

Hanke, J y Reitsch, A. (1996). Pronósticos en los Negocios. Ed 5. Prentice Hall

Hamilton, J. (1994). Times Series Analysis. Princeton: Princeton University Press.

Hausman, J. y Taylor, W. (1981) Panel data and unobservable individual effects. Econometric Society

Hill, Griffiths y Judge (1993) Learning and practicing econometrics. Ed. 1. New York: John Wiley

Hill, Griffiths y Judge (2001) Undergraduate econometrics. Ed. 2. New York: John Wiley

Hsiao, C. (2002). Analysis of panel data. Econometric society monographs. Cambridge.

Johnston, J. y Dinardo, J. (1997) Econometrics methods. McGraw Hill

Juan, A; Kizys, R; y Manzanedo, L. Modelos de ecuaciones simultáneas. Universitat Obertura de Catalunya.

Judge, G, Hill, C, Griffiths, W, Lütketpohl, H, Lee, T. (1985) The theory and practice of econometrics. New York: John Wiley.

Judge, G, Hill, C, Griffiths, W, Lütketpohl, H, Lee, T. (1988) Introduction to the theory and practice of econometrics. 2nd Sub edition New York: John Wiley & Sons.

Krugman, P y Obstfeld, M. (2006). Economía internacional: Teoría y política. Pearson. Addison Wesley.

Maddala, G. S. (1983). Limited-dependet and qualitative variables in econometrics. Cambridge University.

Mahía, R. (2006) Breve apunte sobre la estimación de modelos multiecuacionales.

Makridakis, S. y Wheelwright, S. C. (1978) Forecasting methods and applications. Ed. 2. New York: John Wiley & Sons

Mendieta, J. C. y Perdomo, J. A. (2007). Especificación y estimación de un modelo de precios hedónico espacial para evaluar el impacto de Transmilenio sobre el valor de la propiedad en Bogotá. Documento CEDE 2007-22. Universidad de los Andes.

Mendieta, J. C. y Perdomo, J. A (2008) Fundamentos de economía del transporte: Teoría, metodología y análisis de política. Colección CEDE 50 años. Universidad de los Andes.

Montalvo, F. (2003) Cálculo diferencial e integral en varias variables. Universidad de Extremadura.

Montenegro, A. (2007) Series de tiempo. Ed. 5. Pontificia Universidad Javeriana.

Mundlak, Y. (1978) On the pooling of time series and cross section data. Econometric Society.

Muñoz, J. y Kikut, A. (1994) El Filtro de Hodrick y Prescott: Una técnica para la extracción de la tendencia de una serie. Banco Central de Costa Rica.

Oczkowski, E. (2003) Two-stage least squares (2SLS) and structural equation models (SEM) Charles Sturt University.

Pena, B. El uso de retardos en los modelos econométricos uniecuacionales: Modelos autorregresivos y modelos con retardos escalonados. Instituto Nacional de Estadística de España.

Perron, B, y Roger moon, H. (2006) Seemingly unrelated regressions.

Pindyck, R. y Rubinfeld, D. (1981). Econometric models and econometric forecasts. Mc Graw Hill.

Pindyck, Robert S y Rubinfield, Daniel L (2000). Econometría modelos y pronósticos. Ed. 4. McGraw-Hill.

Prada, T. (2004) Incorporación del fondo de estabilización de precios del azúcar en Colombia. Ilades-Georgetown University, School of Economics and Bussines.

Pulido, A. (1987). Modelos econométricos. Ed. 1. Pirámide.

Pulido, A y Garcia, J (2001) Modelos Econométricos. Ed 1. Madrid: Ediciones Pirámide.

Rodríguez, C, Sánchez, F y Armenta, A. (2007). Hacia una mejor educación rural: Impacto de un programa de intervención a las escuelas en Colombia. Universidad de los Andes: Documento CEDE 2007-13.

Rosales, R y Bonilla, J. (2006). Introducción a la econometría. Apuntes de clase N°3. CEDE. Facultad de Economía. Universidad de los Andes.

Sánchez, F., Fazio, A. y López, M. (2008). Land conflict, property rights, and the rise of the export economy in Colombia, 1850-1925. Universidad de los Andes: Documento CEDE 2008-16.

Shiskin, J. (1978). Seasonal adjustment of sensitive indicators. Keynote address US department of label.

Vásquez, J. C. (2005). Análisis empírico del fondo de estabilización de precios del azúcar en Colombia. Memoria de grado maestría. Facultad Economía, Universidad de los Andes.

Wooldridge, J. (2002). Econometric analysis of cross section and panel data. MIT Press.

Wooldridge, J. (2009) Introductory econometrics. Ed. 4 Australia: South-Western/Cengage learning.

Zellner, A. (1962) An efficient method of estimating seemingly unrelated regressions and tests for aggregation bias. Journal of the American Statistical Association N° 57.

Zellner, A., y Theil, H. (1962) Three-stage least squares: Simultaneous estimation of simultaneous equations. Econometrica N° 30.