

# ECONOMETRÍA

## APLICADA UTILIZANDO R

---

Luis Quintana Romero  
Miguel Ángel Mendoza  
Coordinadores



# **ECONOMETRÍA**

## **APLICADA UTILIZANDO R**

**Luis Quintana Romero y Miguel Ángel Mendoza**  
Coordinadores

# **ECONOMETRÍA**

## **APLICADA UTILIZANDO R**

Javier Galán Figueroa

Jorge Feregrino Feregrino

Lucía A. Ruíz Galindo

Luis Quintana Romero

Miguel Ángel Mendoza González

Roldán Andrés Rosales

**Luis Quintana Romero y Miguel Ángel Mendoza**  
Coordinadores

# Econometría aplicada utilizando R

Coordinado por Luis Quintana Romero y

Miguel Ángel Mendoza González

Portada: D. G. Rocío Borrayo

Primera edición, marzo 2016

D.R. © Universidad Nacional Autónoma de México  
Ciudad Universitaria, Delegación Coyoacán,  
C.P. 04510, México, D.F.

D.R. © Facultad de Estudios Superiores Acatlán  
Av. Alcanfores y San Juan Totoltepec s/n,  
C.P. 53150, Naucalpan de Juárez, Estado de México.

Prohibida la reproducción total o parcial por cualquier medio sin la autorización escrita del titular de los derechos patrimoniales.

El libro electrónico *Econometría aplicada utilizando R* fue financiado con recursos PAPIME de la Dirección General de Asuntos del Personal Académico (DGAPA) de la Universidad Nacional Autónoma de México: PE302513 *Libro electrónico y complementos didácticos en medios computacionales, para el fortalecimiento en la enseñanza de la econometría*. Se encuentra disponible de manera libre en el sitio <http://saree.com.mx/econometriaR/>

ISBN EBook: En trámite

Hecho en México

# Contenido

INTRODUCCIÓN .....	10
CAPITULO 1: LA ECONOMETRÍA: SUS USOS Y APLICACIONES EN R .....	15
1. ¿QUÉ ES LA ECONOMETRÍA?.....	15
2. LA METODOLOGÍA ECONOMÉTRICA .....	17
3. EL MODELO ECONOMÉTRICO .....	20
4. ECONOMETRÍA APLICADA Y R .....	22
5. ALGUNOS DESARROLLOS EN R QUE FACILITAN EL USO DE LA ECONOMETRÍA .....	34
REFERENCIAS.....	43
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	44
MATERIAL DE APRENDIZAJE EN LÍNEA .....	44
CAPÍTULO 2: ENFOQUE MATRICIAL DE LA REGRESIÓN LINEAL.....	45
1. EL MODELO MATRICIAL.....	45
2. ANÁLISIS EXPLORATORIO DE LOS DATOS.....	47
3. ESTIMACIÓN POR MINIMOS CUADRADOS ORDINARIOS .....	51
REFERENCIAS.....	55
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	55
MATERIAL DE APRENDIZAJE EN LÍNEA .....	56
CAPITULO 3: EL MODELO DE REGRESIÓN MÚLTIPLE .....	57
1. ESPECIFICACIÓN DEL MODELO DE REGRESIÓN MÚLTIPLE .....	57
2. ESTIMACIÓN DE LOS COEFICIENTES DE REGRESIÓN .....	62
3. LAS PROPIEDADES DE LOS ERRORES .....	69
4. PRUEBAS DE DIAGNÓSTICO .....	75
5. UN EJEMPLO FINAL EN R .....	77
REFERENCIAS.....	82
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	82
MATERIAL DE APRENDIZAJE EN LÍNEA .....	82
CAPITULO 4: ERROR DE ESPECIFICACIÓN.....	83
1. INTRODUCCIÓN .....	83

2. ESPECIFICACIÓN Y SUPUESTOS DEL MODELO GENERAL DE REGRESIÓN LINEAL.....	85
3. SOBREPARAMETRIZACIÓN Y SUBPARAMETRIZACIÓN, CONSECUENCIAS SOBRE LAS PROPIEDADES DE LOS ESTIMADORES .....	87
4. PRUEBA RESET.....	89
5. PRUEBA RESET EN R .....	90
REFERENCIAS.....	93
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	94
MATERIAL DE APRENDIZAJE EN LÍNEA .....	94
CAPITULO 5: NORMALIDAD.....	95
1. INTRODUCCIÓN .....	95
2. MODELO GENERAL DE REGRESIÓN LINEAL.....	96
3. IMPORTANCIA DE LA DISTRIBUCIÓN NORMAL EN LA INFERENCIA ESTADÍSTICA.....	99
4. PRUEBA DE NORMALIDAD DE JARQUE-BERA .....	109
5. PRUEBA JARQUE-BERA EN R .....	110
6. CAUSAS E IMPLICACIONES DE LA NO NORMALIDAD Y POSIBLES SOLUCIONES.....	114
7. CONCLUSIONES .....	114
REFERENCIAS.....	115
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	116
MATERIAL DE APRENDIZAJE EN LÍNEA .....	116
CAPÍTULO 6: MULTICOLINEALIDAD.....	117
1. LA MULTICOLINEALIDAD UN PROBLEMA DE GRADO.....	117
2. PRUEBAS PARA LA DETECCIÓN DE MULTICOLINEALIDAD .....	121
3. UN EJEMPLO PRÁCTICO EN LA DETECCIÓN DE MULTICOLINEALIDAD EN R CON LA FUNCIÓN CONSUMO PARA MÉXICO .....	124
4. SOLUCIONES AL PROBLEMA DE LA MULTICOLINEALIDAD .....	133
REFERENCIAS.....	138
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	139
MATERIAL DE APRENDIZAJE EN LÍNEA .....	139
CAPÍTULO 7: HETEROCEDASTICIDAD .....	140
1. INTRODUCCIÓN .....	140
2. ESTRATEGIAS PARA REALIZAR ESTIMACIONES EN PRESENCIA DE HETEROCEDASTICIDAD ....	141
3. LAS CAUSAS DE LA HETEROCEDASTICIDAD .....	144

4. CONTROL Y DETECCIÓN DE LA HETEROCEDASTICIDAD .....	145
5. EJEMPLO EN R .....	150
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	156
MATERIAL DE APRENDIZAJE EN LÍNEA .....	156
CAPÍTULO 8: AUTOCORRELACIÓN SERIAL.....	157
1. INTRODUCCIÓN .....	157
2. DETECCIÓN DE LA AUTOCORRELACIÓN .....	158
3. PROCEDIMIENTO PARA LA DETECCIÓN DE LA AUTOCORRELACIÓN EN R-STUDIO.....	162
REFERENCIAS.....	173
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	173
MATERIAL DE APRENDIZAJE EN LÍNEA .....	173
CAPITULO 9: ANALISIS DE INTEGRACION: APLICACIONES EN SOFTWARE R.....	174
1. INTRODUCCION .....	174
2. ANALISIS DE INTEGRACIÓN .....	174
3. APLICACIONES EN R.....	182
REFERENCIAS.....	207
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	207
MATERIAL DE APRENDIZAJE EN LÍNEA .....	207
CAPÍTULO 10: COINTEGRACIÓN Y MODELOS DE CORRECCION DE ERROR.....	208
1 INTRODUCCIÓN .....	208
2 EL CONCEPTO DE COINTEGRACIÓN .....	209
3. PRUEBA DE COINTEGRACIÓN DE ENGLE Y GRANGER .....	211
4. ANÁLISIS DE COINTEGRACIÓN DE PHILLIPS-OULIARIS .....	224
5. MODELO DE CORRECCIÓN DE ERROR .....	229
6. COINTEGRACIÓN CON METODOLOGÍA DE JOHANSEN Y JOSELIUS.....	233
REFERENCIAS.....	243
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	243
MATERIAL DE APRENDIZAJE EN LÍNEA .....	244
CAPÍTULO 11: MODELOS VAR .....	245
1. INTRODUCCIÓN .....	245
2. CARACTERÍSTICAS DEL MODELO VAR .....	246
3. UN CASO PARA LA ECONOMÍA MEXICANA .....	248

REFERENCIAS.....	270
ARCHIVO DE DATOS ASOCIADO AL CAPÍTULO .....	271
MATERIAL DE APRENDIZAJE EN LÍNEA .....	271
CAPÍTULO 12: MODELOS ARCH .....	272
1. RIESGO Y VOLATILIDAD .....	272
2. PROCESOS ARCH.....	273
3. VARIANTES DE LOS MODELOS ARCH.....	277
4. UNA APLICACIÓN DEL MODELO ARCH EN R.....	278
REFERENCIAS.....	293
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	294
MATERIAL DE APRENDIZAJE EN LÍNEA .....	294
CAPITULO 13: MODELOS LOGIT Y PROBIT .....	295
1. LA IMPORTANCIA DE LAS VARIABLES CATEGÓRICAS .....	295
2. MODELOS LOGIT Y PROBIT.....	297
3. APLICACIONES EN R.....	303
REFERENCIAS.....	308
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	308
MATERIAL DE APRENDIZAJE EN LÍNEA .....	308
CAPITULO 14: MODELOS PANEL Y SUS APLICACIONES EN R .....	309
1. INTRODUCCION .....	309
2. MODELO PANEL ESTÁTICO GENERAL.....	310
3. ELECCIÓN DE MODELOS ALTERNATIVOS .....	314
4. RESULTADOS DE LOS MODELOS ECONOMÉTRICOS PANEL CON EL PAQUETE PLM DE R.....	315
REFERENCIAS.....	328
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	329
MATERIAL DE APRENDIZAJE EN LÍNEA .....	329
CAPÍTULO 15: ECONOMETRÍA ESPACIAL Y SUS APLICACIONES EN R.....	330
1. INTRODUCCION .....	330
2. VECINDAD Y DEPENDENCIA ESPACIAL .....	332
3. ESTADÍSTICOS DE DEPENDENCIA ESPACIAL .....	344
4. MODELOS ESPACIALES .....	360
REFERENCIAS.....	382

ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	383
MATERIAL DE APRENDIZAJE EN LÍNEA .....	383
CAPÍTULO 16: REPASO BÁSICO DE ESTADÍSTICA Y ÁLGEBRA MATRICIAL .....	384
1. INTRODUCCIÓN .....	384
2. REVISIÓN DE LOS DATOS.....	384
3. VARIABLE ALEATORIA.....	396
4. BREVE REPASO DE ÁLGEBRA DE MATRICES .....	414
REFERENCIAS.....	443
ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO .....	444
MATERIAL DE APRENDIZAJE EN LÍNEA .....	444
LISTA DE AUTORES .....	445

# INTRODUCCIÓN

En este libro de texto los usuarios encontraran una vía práctica para mejorar su comprensión de la econometría, al utilizar aplicaciones a su realidad social, emplear las fuentes de información disponibles en el país y disponer de un formato tecnológico en el que pueden aplicar los conocimientos adquiridos, poner en práctica propuestas propias y realizar trabajo de investigación por su cuenta haciendo uso de medios tecnológicos de uso masivo.

Los capítulos de este libro de texto tienen como eje común la aceptación de que en los últimos veinte años se ha dado una revolución en las técnicas econométricas y en sus aplicaciones. En buena parte estos cambios provienen del reconocimiento de que el paradigma clásico, que actualmente aún predomina en la mayoría de los libros de texto, fue sustentado en supuestos muy discutibles. Los cuestionamientos a la metodología econométrica clásica se desprenden del trabajo de Box y Jenkins (1970) en series de tiempo; Davidson, Hendry, Srba y Yeo (1978) que desarrollaron la idea de modelos de corrección de error (MCE) y que actualmente su propuesta se reconoce como metodología LSE (London School of Economics) o DHDY (por las iniciales de sus autores); los numerosos trabajos de Engle y Granger a partir de los años ochenta en donde se vincula el concepto de cointegración a los MCE; el trabajo del mismo Engle (1982) que dio lugar a los modelos ARCH (heterocedasticidad condicional autorregresiva), los cuales han tenido un gran impacto en el análisis econométrico aplicado al mundo de las finanzas; Los desarrollos de finales de los años noventa en el campo de la Econometría Espacial impulsados por Anselin (1988) y; un sin número de artículos que inspirados en estos trabajos pioneros han cambiado la forma de pensar y hacer econometría en la actualidad.

El reto de este libro es ofrecer a los lectores un enfoque aplicado con el fin de comprender esos nuevos desarrollos en el campo de la econometría y proporcionarles las herramientas teóricas y las técnicas necesarias para su aplicación al estudio de la realidad económica mexicana.

Los libros de texto de econometría que se están publicando recientemente, tanto en Europa como en los Estados Unidos, se vinculan a paquetes computacionales de elevado costo comercial como el EViews, STATA y Microfit, entre otros. Sin embargo, actualmente se ha desarrollado software de uso libre que ha adquirido una gran difusión mundial, uno de ellos es el R, el cual se ha venido utilizando para la modelación econométrica con mucho éxito.

Por tal razón, el presente libro de texto de econometría tiene la peculiaridad de que utiliza ampliamente los desarrollos disponibles libremente en R, además de priorizar la aplicación de los temas que se desarrollan en sus diferentes capítulos. En cada uno de los capítulos del libro se muestran las bases del método o técnica econométrica de que se trate y se aplica inmediatamente al estudio de algún tema relevante de la economía mexicana actual o de otros países.

Los capítulos que conforman este libro presentan un nivel introductorio de cada uno de los temas que se abordan y se priorizan las aplicaciones en R, por lo cual debe considerarse como un libro de econometría básica aplicada. Se ha dejado fuera del texto el tema de los modelos de series de tiempo, ya que por la amplitud de ese tema se requiere de un libro adicional, mismo que ya se encuentra en proceso de preparación con el fin de complementar a la presente obra.

Debemos señalar que este libro de texto forma parte de la producción y edición de tres materiales educativos en el campo de la econometría. Los materiales consisten de un libro electrónico (ebook) de texto, un curso en línea y aplicaciones electrónicas didácticas.

Estos materiales están destinados a profesores y alumnos. En el caso de los profesores es posible emplear el texto electrónico y el curso en línea para los

cursos de actualización del personal docente en econometría. Los profesores pueden utilizar los materiales en la impartición de cursos a nivel licenciatura, ya que los materiales se diseñan de acuerdo a los contenidos de los programas curriculares de econometría y de métodos de pronóstico en diferentes licenciaturas, resolviendo con ello el déficit existente de material actualizado, en español, en soportes electrónicos y con aplicaciones a la realidad del país.

La propuesta es original en la medida en que atiende tres problemas de la enseñanza de la econometría; contar con libros de texto actualizados en formatos tecnológicamente avanzados y en español, incorporar un curso en línea que tenga la virtud de promover el auto aprendizaje y sea complemento de los cursos presenciales, además de proporcionar aplicaciones en formatos tecnológicos que se han difundido ampliamente entre los alumnos.

Los materiales vinculados a este libro de texto se encuentran disponibles de forma libre en la página [www.saree.com.mx/unam](http://www.saree.com.mx/unam). En ese sitio el interesado en el estudio de la econometría encontrará este libro en formato electrónico, presentaciones de power point para cada capítulo, una grabación de video con los procedimientos para aplicar en R lo aprendido en el capítulo, una guía metodológica en MOODLE para avanzar en el estudio de los capítulos y, finalmente, un par de aplicaciones electrónicas para comprender la forma en la que se estiman regresiones.

El libro se integra por dieciséis capítulos cuyo contenido se resume en la siguiente tabla.

CAPÍTULOS	CONTENIDO
<b>CAPÍTULO 1.</b> <b>Metodología econométrica:</b>	Se introduce al lector en la metodología econométrica moderna y en el uso del R
<b>CAPÍTULO 2. Enfoque matricial de la regresión lineal</b>	Se muestra el método de mínimos cuadrados ordinarios en su versión matricial con ejemplos de análisis de la deuda pública en México
<b>CAPÍTULO 3. El modelo de regresión múltiple:</b>	Se desarrolla el modelo de regresión múltiple y la forma en la cual se evalúan sus resultados. Se realizan aplicaciones en R al análisis de las ventas al menudeo en México.
<b>CAPÍTULO 4. Error de especificación.</b>	Se presentan los métodos utilizados para determinar si el modelo econométrico fue especificado incorrectamente debido a un planteamiento no apropiado de la forma funcional. Se realizan aplicaciones en R con el análisis de la demanda de gasolina en los Estados Unidos.
<b>CAPÍTULO 5.</b> <b>Normalidad.</b>	En este capítulo se estudia la importancia e implicaciones del supuesto de normalidad en el modelo de regresión lineal y de manera específica en la inferencia estadística de sus parámetros. Se realizan aplicaciones en R de la prueba Jarque-Bera en un modelo de la demanda de gasolina en los Estados Unidos.
<b>CAPÍTULO 6.</b> <b>Multicolinealidad</b>	Con base en los determinantes del consumo en México se exploran las diferentes pruebas alternativas disponibles en R para detectar y corregir el problema de la multicolinealidad en los modelos econométricos.
<b>CAPÍTULO 7.</b> <b>Heterocedasticidad</b>	Se explican las consecuencias del problema de heterocedasticidad en los modelos econométricos y haciendo uso de un ejemplo sobre distribución de cerveza se muestran las alternativas disponibles en R para realizar pruebas de detección de ese problema.
<b>CAPÍTULO 8.</b> <b>Autocorrelación:</b>	La autocorrelación serial y sus consecuencias es analizada con base en el estudio de las tasas de interés en México. Utilizando R se muestran las pruebas para detectar este problema y las alternativas para su solución.
<b>CAPÍTULO 9. Integración</b>	En este capítulo se aborda uno de los temas más relevantes de la metodología econométrica moderna que es el de identificar el orden de integración de las variables utilizadas en los modelos econométricos. Con base en el R se realizan pruebas de raíz unitaria utilizando como ejemplo el análisis del Producto Interno Bruto de México.
<b>CAPÍTULO 10.</b> <b>Cointegración</b>	Los resultados del capítulo anterior se extienden al estudio de los procesos de cointegración entre las variables del modelo econométrico utilizando en R las técnicas de Engle-Granger y de Johansen, exemplificándolas con ayuda del estudio de la relación de largo plazo entre el consumo y el ingreso en México.
<b>CAPÍTULO 11. Modelos VAR:</b>	Se destaca el uso de modelos VAR para el análisis de la política económica tomando como caso el estudio de la inflación y la oferta monetaria. Se presentan las diferentes rutinas disponibles en R para estimar y realizar pruebas en los modelos VAR.
<b>CAPÍTULO 12. Modelos ARCH:</b>	Los modelos ARCH utilizados para el análisis de la volatilidad y el riesgo son ejemplificados en R con base en el análisis de los procesos inflacionarios en México.
<b>CAPÍTULO 13. Modelos Logit y Probit:</b>	Se desarrollan los modelos Probit y Logit aplicados a casos en los que la variable dependiente es binaria o cualitativa. Con base en el estudio de la diferenciación salarial en México se muestran las rutinas disponibles en R para estimar y realizar pruebas en ese tipo de modelos econométricos.
<b>CAPÍTULO 14. Modelos de panel:</b>	Cuando el fenómeno económico que se está analizando tiene un componente de desagregación de corte trasversal o sección cruzada y otro de series de tiempo se aplican modelos de panel. En este capítulo se estudian las técnicas de panel utilizando R en el análisis de la inflación y el desempleo en México.

<b>CAPÍTULO 15.</b> <b>Econometría espacial:</b>	Uno de los desarrollos más recientes de la econometría es la econometría espacial. En este capítulo se presenta la forma en la que se deben especificar y estimar este tipo de modelos en R y se ejemplifica su uso con el estudio del empleo y el capital humano en la zona centro de México.
<b>CAPÍTULO 16: Repaso básico de estadística, probabilidad y álgebra lineal en R:</b>	Finalmente, se incluye un capítulo opcional en el que se realiza un breve repaso de los elementos básicos de estadística, probabilidad y álgebra lineal indispensables para comprender la base matemática de los diferentes capítulos del libro.

Este libro y los materiales didácticos adicionales que lo acompañan contaron con el apoyo financiero de la Dirección General de Asuntos del Personal Académico de la UNAM a través del proyecto PAPIME PE302513 “Libro electrónico y complementos didácticos en medios computacionales, para el fortalecimiento en la enseñanza de la econometría”.

Los coordinadores del libro agradecen a los profesores José A. Huitrón, Jaime Prudencio, Aída Villalobos y Ángel Reynoso por su apoyo en la revisión de los capítulos y en el diseño de los apoyos didácticos que acompañan al libro. También agradecemos a los alumnos y becarios del proyecto PAPIME; Arturo Abraham Salas, Mónica González, Paola Orozco, Ana Isabel Hernández, Coral Gutiérrez, Eddy Michell López, Jarett Fernando González, Mónica Patricia Hernández, Samarkanda Norma Bustamante, Nataly Hernández, Sarahí Aldana, Brenda Mireya González, Alejandro Corzo, Damaris Susana Mendoza, Nancy Nayeli Morales, Claudia Torres, Edelmar Morales y Carolina Guadalupe Victoria. Todas y todos ellos hicieron una excelente labor de apoyo para el buen éxito del proyecto.

LUIS QUINTANA ROMERO Y MIGUEL ÁNGEL MENDOZA GONZÁLEZ

# **CAPITULO 1: LA ECONOMETRÍA: SUS USOS Y APLICACIONES EN R**

**Luis Quintana Romero y Miguel Ángel Mendoza**

## **1. ¿QUÉ ES LA ECONOMETRÍA?**

Hoy en día la econometría se ha difundido ampliamente entre quienes estudian y buscan realizar aplicaciones de la economía. En general, cualquier licenciatura en economía cuenta, entre su currículo, con uno o más cursos de econometría; hoy en día es usual que la econometría se enseñe con la misma relevancia que se le da a los cursos de microeconomía y macroeconomía. No hay posgrado en economía que deje de incorporar el estudio de la econometría como una disciplina fundamental. Incluso, es posible aseverar que en disciplinas distintas a la economía, como en las matemáticas, algunas ingenierías, la sociología y en la psicología, sus estudiantes reciben algún curso de econometría.

No sólo en la formación académica la econometría está presente, en la vida laboral se realizan todos los días aplicaciones econométricas. En las oficinas gubernamentales se emplean modelos econométricos para realizar pronósticos de variables económicas. En empresas privadas se utilizan algunas técnicas econométricas para proyectar al futuro variables como ventas, precios y demanda, entre otras variables. En el mercado existen numerosos servicios de consultoría que han hecho de la econometría un negocio al ofrecer la venta de pronósticos generados a través de modelos econométricos.

En el mundo de la investigación científica la econometría es un ingrediente indispensable. Diariamente se publican en todo el orbe una gran cantidad de artículos de economía en revistas especializadas, la evidencia empírica que aportan, generalmente, se sustenta en algún modelo económico.

La importancia de esta disciplina es tal que basta escribir en un buscador de internet la palabra "econometrics", para que nos arroje más de nueve millones de referencias.

Con la econometría se busca comprender fenómenos como el de las crisis, identificar sus causas, valorar sus consecuencias futuras y proponer medidas de política para enfrentarlas. Para ello, la econometría utiliza modelos, con estos se busca representar de forma simplificada a los principales factores causales de un problema de interés. La especificación y estimación de esos modelos requiere del conocimiento de teorías económicas, para poder establecer relaciones entre las variables, y de datos, para poder realizar mediciones de dichas relaciones.

No existe una definición única y generalmente aceptable de lo qué es la econometría. Debido a que en ella concurren una gran diversidad de perspectivas teóricas y metodológicas, existen, en consecuencia, diferentes posturas sobre su significado.

A diferencia de lo que ocurre hoy en día, en los años treinta, época en la que se institucionaliza la econometría, existía cierto consenso metodológico. A ese consenso se le identifica como la "metodología de libro de texto" y su definición de econometría era la siguiente:

*La aplicación de métodos estadísticos y matemáticos al análisis de los datos económicos, con el propósito de dar un contenido empírico a las teorías económicas y verificarlas o refutarlas (Maddala, 1996, p.1)*

Bajo esta última conceptualización la econometría aparece, por un lado, como un mero instrumental técnico al ser la aplicación de métodos matemáticos y estadísticos. Por otro lado, es vista prácticamente como la piedra filosofal, al darle el papel de criterio último de verdad al ser la vía para verificar o refutar teorías. El economista aparece en esa definición como un técnico, cuyo único fin es intentar medir lo que la teoría económica ha postulado.

Esta visión de la econometría se ha transformado en los últimos años, en ese sentido vale la pena retomar la definición proporcionada por Aris Spanos:

*"La econometría se interesa por el estudio sistemático de fenómenos económicos utilizando datos observables" (Spanos, 1996, p.3).*

Este es un enfoque moderno con el cual se coincide en este libro, lo que hace a la econometría diferente de otros campos de la economía es la utilización de datos observables. Por lo tanto, la econometría tiene una perspectiva empírica, no se reduce a la teoría y necesariamente hace uso de datos, los cuales no son experimentales sino que son resultado del funcionamiento de la actividad económica. El papel del econométrista no se reduce a medir lo que la teoría económica establece, es un científico social que, a través de un método científico, emprende el estudio de fenómenos económicos. Por lo tanto, no es un observador pasivo de la teoría, al contrario, es capaz de contribuir a la teoría.

La econometría que utilizamos hoy en día se ha ido transformando y modernizando, hasta convertirse en una de las herramientas más potentes a disposición de los economistas y principalmente del análisis empírico de problemas económicos. Esta evolución de la disciplina la sintetiza perfectamente Spanos:

*"En el amanecer del siglo veintiuno, la econometría se ha desarrollado desde los modestos orígenes del "ajuste de curvas" por mínimos cuadrados en los inicios del siglo veinte, hasta un poderoso arreglo de herramientas estadísticas para modelar todo tipo de datos, desde las tradicionales series de tiempo a las secciones cruzadas y los datos de panel." (Spanos, 2006, p. 5)*

## **2. LA METODOLOGÍA ECONOMÉTRICA**

En el apartado previo se estableció que la econometría estudia de forma sistemática los fenómenos económicos. Por lo tanto, utiliza una metodología científica para llevar a cabo esta tarea. Aunque la metodología econométrica no tiene aún un lugar relevante en la discusión de esta disciplina, es un aspecto que debe ser considerado esencial, por ello resulta muy atinada la afirmación de

Spanos (2006) en el sentido de que sin fundamentos metodológicos para guiar la práctica econométrica, no es posible que se logre acumular conocimiento genuino a través de la modelación empírica.

En la medida en que existe una diversidad metodológica en la econometría, resulta difícil establecer un proceso metodológico único. Sin embargo, en términos generales, en el cuadro siguiente se pueden observar las características básicas de los principales enfoques metodológicos, los cuales se distinguen por el papel que le asignan a la teoría y del grado de independencia que le dan a la teoría para la caracterización de los datos Hoover (2006).

Cuadro 1

#### Perspectivas metodológicas en la econometría

<b>Metodología</b>	<b>Período</b>	<b>Autores</b>	<b>Características</b>
Comisión Cowles	Años 40 y 50	Koopmans	Se centró en el problema de identificación y el papel de la teoría para establecer las restricciones de identificación
Vectores Auto Regresivos (VAR)	Años 80	Christoper Sims	Enfoque sin teoría en la estructura de los datos y uso e ecuaciones VAR para modelar impactos en las variables
Calibración	Años 90	Finn Kydland y Edward Prescott	Modelos teóricos de expectativas racionales a los que se les asignan valores numéricos en los parámetros claves
Libro de texto	Años 90 y 2000	Post Comisión Cowles	Resurge la metodología de la Comisión Cowles aplicada a modelos uniecuacionales con métodos instrumentales
London School Economics (LSE)	Años 90 y 2000	Denis Sargan, David Hendry	Especificaciones dinámicas, cointegración y búsqueda de especificaciones parsimoniosas; Anidamiento y metodología de lo general a lo específico

Fuente: Elaboración propia con base en Hoover (2006)

Dentro de estas perspectivas la LSE ha jugado un papel destacado al contraponerse a la de libro de texto y conformar lo que puede denominarse una nueva metodología econométrica. La de libro de texto parte del supuesto de que el modelo teórico es el verdadero modelo y, en consecuencia, coincide con el proceso generador de los datos (PGD). En consecuencia, para esa metodología,

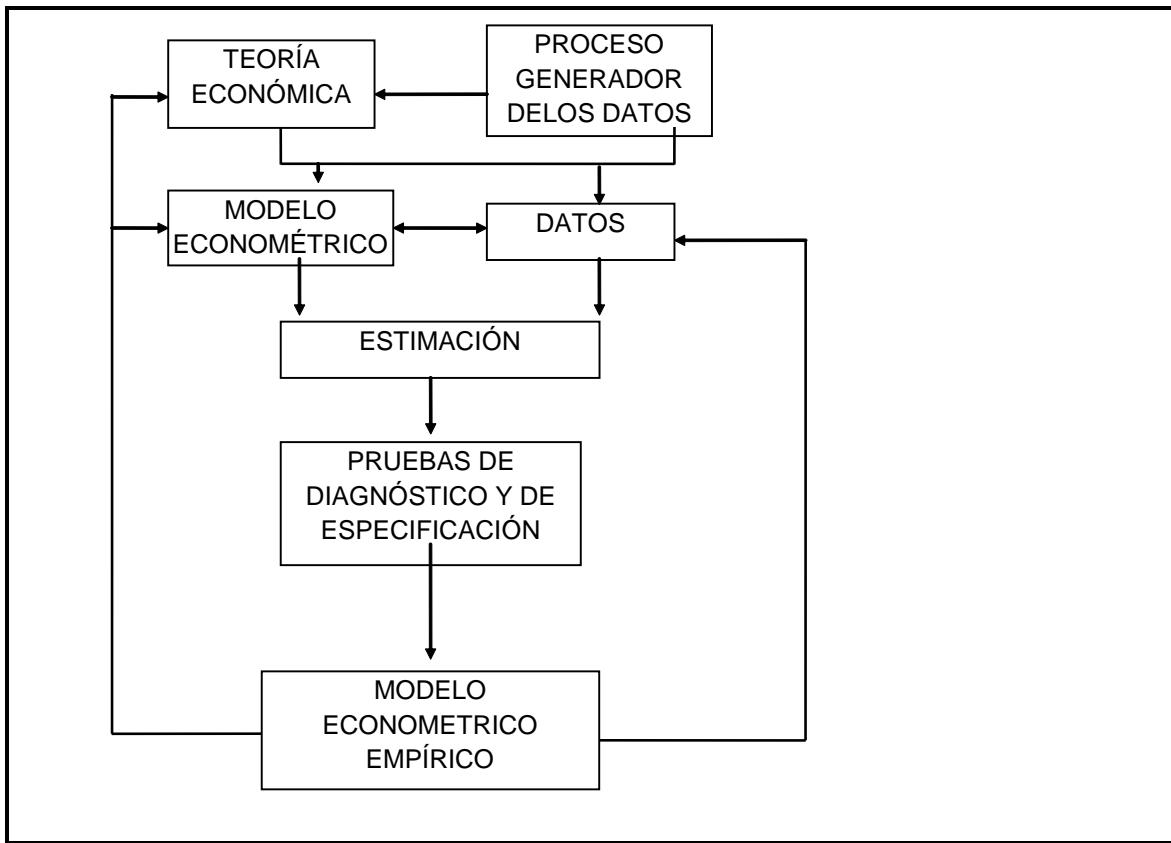
la econometría se reduce a la estimación de los parámetros que la teoría plantea; mide lo que la teoría dice, pero no explica nada.

Al contrario, la LSE parte de la idea de que los modelos son aproximaciones teóricas y empíricas del PGD. La validación de esas aproximaciones se realiza a través de la evaluación de los modelos utilizando una amplia batería de pruebas estadísticas que buscan determinar la congruencia de esas aproximaciones con el PGD. El PGD como fenómeno económico de interés que da lugar a los datos, no es conocido debido a que los datos son observacionales y no experimentales; los datos que se utilizan en los modelos econométricos no son generados en un laboratorio bajo control.

En el esquema siguiente se ejemplifica la metodología LSE o nueva metodología. Ahí se observa que la teoría y los datos tienen la misma importancia y aparecen como punto de partida, además de que las variables teóricas no necesariamente coinciden estrictamente con los datos. También se observa que existe retroalimentación entre el modelo econométrico y las pruebas de diagnóstico y especificación. Los datos, la teoría y el modelo teórico no son tomados como dados, son retroalimentados por el modelo empírico.

**Figura 1**

**Nueva metodología econométrica**



Fuente: Aris Spanos *Statistical Foundation of econometrics*

### 3. EL MODELO ECONOMÉTRICO

Los modelos económicos son una simplificación de la realidad que se compone de relaciones entre variables. Dichas relaciones son no exactas y, por ello, se les llama relaciones estadísticas y pueden describirse en términos probabilísticos. Este tipo de relaciones funcionales pueden expresarse como un modelo estadístico para una variable dependiente  $y_i$  y un conjunto de  $k-1$  variables explicativas o regresores  $X_{ki}$ :

$$y_i = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \quad (1)$$

En donde el término  $u_i$  es un error o perturbación aleatoria y  $\beta_1 \dots \beta_k$  son los parámetros desconocidos a estimar por el modelo.

La estimación de los parámetros de este modelo implica la utilización de variables reales que midan la relación funcional definida. La búsqueda de las variables medibles no es asunto fácil ya que por una parte, la teoría no especifica cuál variable de la contabilidad nacional debe ser utilizada y, por otra parte, la estadística económica disponible no es generada bajo un plan y objetivos de análisis económico, es decir no es controlada por el economista y por ende no necesariamente se ajusta a sus necesidades de estudio de la realidad.

Los modelos econométricos pueden ser uniecuacionales o multiecuacionales. Los modelos uniecuacionales implican la estimación de una sola ecuación los multiecuacionales están formados por más de dos ecuaciones que pueden estar relacionadas entre sí. Los grandes modelos multiecuacionales han perdido importancia debido a la complejidad de su construcción y manejo, además de que el dominio metodológico de modelos más compactos, derivados de las propuestas VAR de formas reducidas, ha llevado a la utilización de modelos de pequeña escala. Sin embargo, aún se siguen actualizando modelos de gran escala para una amplia variedad de países debido a la necesidad de simulaciones de política que requieren los gobiernos, grandes empresas o bancos. Para el caso mexicano la empresa IHS sigue actualizando el primer modelo construido para el país en los años sesenta por CIEMEX una empresa asociada con la firma de modelos WARTHON Econometric Associates International. Actualmente ese modelo genera pronósticos de 800 variables para 25 sectores de la economía (IHS, 2013).

En el apartado anterior se argumentó que la metodología econométrica de libro de texto incorpora el supuesto de “correcta especificación” del modelo. La metodología moderna, al contrario, considera que las variables del modelo son aleatorias y por tanto sus propiedades probabilísticas son compartidas con el término de error.

Para formalizar esta idea consideremos el modelo de regresión como la media condicional de  $y_i$  sobre los valores de  $X_i$ :

$$FRP = E[y_i|X_{ji}] = f(X_{ji}) = \beta_1 + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} \text{ donde } j = 2, 3, \dots, k \text{ } i = 1, 2, \dots, n \quad (2)$$

A esta función se le conoce como función de regresión poblacional (FRP). La estimación de los parámetros de la función requiere de una regla que transforme las variables aleatorias en un estimador de los parámetros desconocidos.

La sustitución de los valores de una muestra particular de realizaciones de las variables aleatorias, en el estimador, genera una estimación de los parámetros desconocidos, la cual depende de la muestra y da lugar a una función de regresión muestral (FRM):

$$FRM = E[y_i|X_{ji}] = f(X_{ji}) = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + \cdots + \hat{\beta}_k X_{ki} \quad (3)$$

El término de error o innovaciones, a diferencia de la metodología tradicional, no es “añadido” a la función de regresión, se obtiene como la diferencia entre  $y_i$  y su media condicional:

$$[u_i|X_{ji}] = y_i - E[y_i|X_{ji}] = FIC \quad (4)$$

Que es conocida como la función de innovación condicional (FIC).

Así la ecuación para  $y_i$  puede escribirse como:

$$y_i = FRP + FIC \quad (5)$$

De esta manera la ecuación tendrá una parte sistemática que se corresponde con FRP y una no sistemática representada por FIC.

#### 4. ECONOMETRÍA APLICADA Y R

El enfoque seguido en este texto es fundamentalmente de econometría aplicada, por ello se centra en las aplicaciones empíricas y se le brinda menor espacio a las discusiones teóricas y conceptuales. Es por lo tanto necesario contar con el

manejo de paquetería computacional que permita la utilización de la metodología econométrica en una amplia variedad de métodos, datos reales y casos prácticos.

El R es un lenguaje y un ambiente para manejo de datos y gráficos en código libre. Dada esas características los desarrollos que se han realizado en R son abiertos y están disponibles gratuitamente, por lo cual su uso se ha difundido ampliamente. El R es difundido libremente por una gran diversidad de sitios espejo del **Comprehensive R Archive Network (CRAN)**. Además de ser gratuitas, los desarrollos para econometría en R se actualizan más rápido que en cualquier otro de los costosos softwares comerciales que se encuentran en el mercado. Esto es así debido a que los usuarios hacen desarrollos, los documentan y los suben al CRAN de R de manera cotidiana.

El R se puede descargar del siguiente vínculo:

<http://CRAN.R-project.org/>

R genera objetos que son números, vectores, matrices, alfa numéricos y cuadros de datos. Los operadores aritméticos a los que usualmente estamos acostumbrados en otros paquetes son los mismos en R; suma (+), resta (-), multiplicación (\*), división (/) y potencia (^). Los ejemplos siguientes están basados en Crawley (2009) y Venables et.al. (2013).

Por ejemplo, podemos generar un objeto número y que contiene el resultado de multiplicar 2 por 5:

```
a <- 2  
b <- 5  
y <- a*b  
> y  
[1] 10
```

También se podría utilizar R como si fuera una calculadora y escribir directamente  $2*5$  y se desplegará el resultado de 10.

Los objetos que hemos creado los podemos listar con las siguientes opciones:

```
objects()  
ls()
```

La ayuda se puede utilizar para obtener referencias de cualquier comando, por ejemplo si queremos saber lo que hace objects basta escribir:

```
help(objects)
```

En seguida R despliega una ventana con toda la documentación del comando, en la cual nos brinda su descripción, uso, argumentos, detalles, referencias y ejemplos de su uso.

Los objetos pueden eliminarse rápidamente, por ejemplo para eliminar a y b basta escribir el siguiente comando:

```
rm(a,b)
```

Para generar un objeto que sea un vector columna podemos usar la opción c;

```
x <- c(5,10,8,7,9)
```

Lo mismo puede hacerse con la función assignment:

```
assign("x", c(5,10,8,7,9))
```

Es posible calcular la media, mean(), la varianza, var(), el valor máximo, max(), el valor mínimo, min() o la longitud del vector, length(). Por ejemplo, si calculamos la media:

```
mean(x)  
[1] 7.8
```

También podríamos generar vectores columna con secuencias de números, por ejemplo si generamos una secuencia del 1 al 10;

```
y<- c(1:10)
> y
[1] 1 2 3 4 5 6 7 8 9 10
```

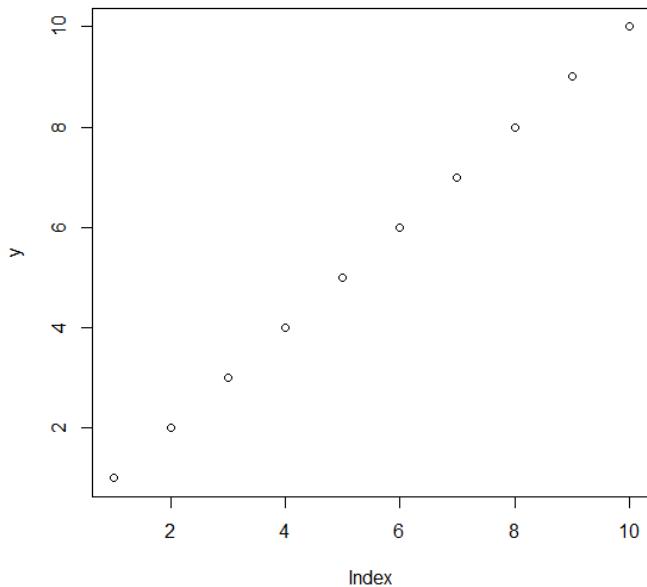
A los elementos de un vector se les pueden asignar nombres, por ejemplo al vector x le asignamos los nombres de los números que contiene:

```
> names(x) <- c("cinco","diez","ocho","siete","nueve")
> x
cinco diez ocho siete nueve
5 10 8 7 9
```

Las gráficas se obtienen usando plot, por ejemplo para realizar una gráfica de los valores del vector y escribimos:

```
plot(y)
```

La gráfica resultante es:



Con el fin de ejemplificar algunas opciones que se utilizarán ampliamente al estimar modelos de regresión vamos a considerar el caso siguiente. Generamos dos vectores con la siguiente información:

```
y <- c(1,2,3,-1,0,-1,2,1,2)
x<- c(0,1,2,-2,1,-2,0,-1,1)
```

Ahora es posible correr la regresión para el modelo:  $y_i = \beta_1 + \beta_2 x_i + u_i$ . Por el momento no se preocupe de las características del modelo, ni de la comprensión del método de estimación ya que eso se aborda en los capítulos siguientes del libro. Aquí simplemente debe aprender que para correr esa regresión se utiliza la función lineal model o lm:

```
lm(y ~ x)
```

Call:

```
lm(formula = y ~ x)
```

Coefficients:

(Intercept)	x
1.0000	0.8125

Los resultados de la regresión se pueden obtener con summary():

```
summary(lm(y ~ x))
```

Residuals:

Min	1Q	Median	3Q	Max
-1.8125	-0.3750	0.1875	0.3750	1.0000

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.0000	0.2938	3.404	0.01138 *
x	0.8125	0.2203	3.688	0.00778 **

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.8814 on 7 degrees of freedom

Multiple R-squared: 0.6602, Adjusted R-squared: 0.6116

F-statistic: 13.6 on 1 and 7 DF, p-value: 0.007782

Ahora ya estamos en condiciones de preparar nuestros datos para utilizarlos en el paquete. La manera más fácil de manejar sus archivos de datos en R es crearlos en una hoja de cálculo como Excel y guardarlos como archivo de texto delimitado por tabulaciones.

Los datos del archivo PWT\_2000.txt fueron guardados en formato de texto delimitado. En el archivo se presentan los datos de la muestra de países de las Penn Tables (2013) con información para el 2000 del PIB per cápita (PIBPC) y de los acervos de capital (K).

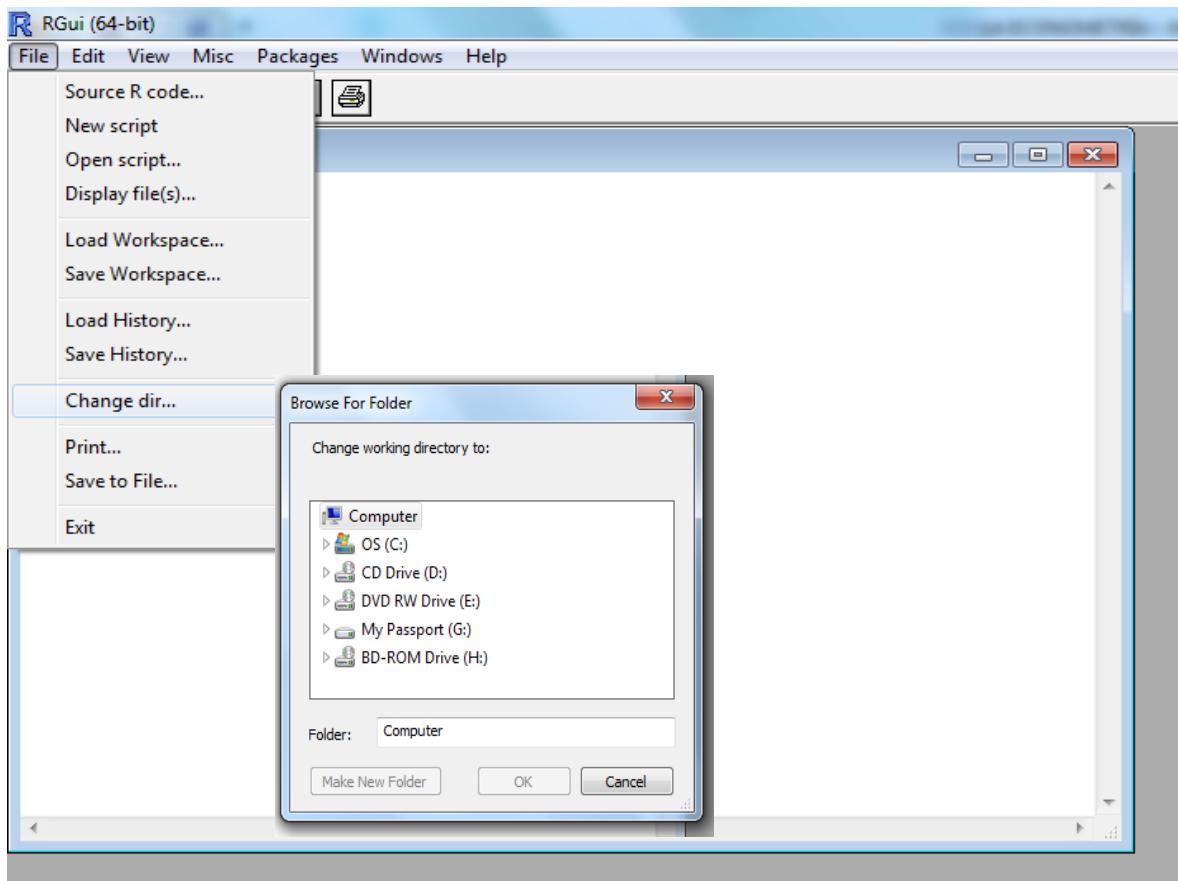
Para abrir esa tabla en R primero se tiene que asegurar que el paquete este direccionado a la carpeta en la que ha guardado su archivo. Para verificar cuál es el directorio actual de trabajo escriba:

```
getwd()
```

Si el directorio que aparece no es el que debe utilizar, puede cambiar de directorio con:

```
setwd("trayectoria del directorio")
```

También puede ir al menú principal de R y en el menú de FILE seleccionar la opción Change directory y en la ventana que se abre buscar la ubicación de su nuevo directorio de trabajo, tal y como se muestra en la imagen siguiente:



Para que sus datos puedan ser cargados en R debe usar el comando para leer tablas (`read.table`) e indicar que la primer línea de su cuadro de datos contiene los nombres de las variables (`header=TRUE`) y que las columnas están separadas por tabulaciones (`sep=""`). Las instrucciones son las siguientes:

```
datos<-read.table("PWT_2000.txt",header=TRUE,sep="")
```

Los datos de la tabla ahora están cargados en un objeto llamado "datos", sin embargo R no puede reconocer cada una de las variables que están en el cuadro: para indicar que las variables están en las columnas se debe usar la siguiente instrucción:

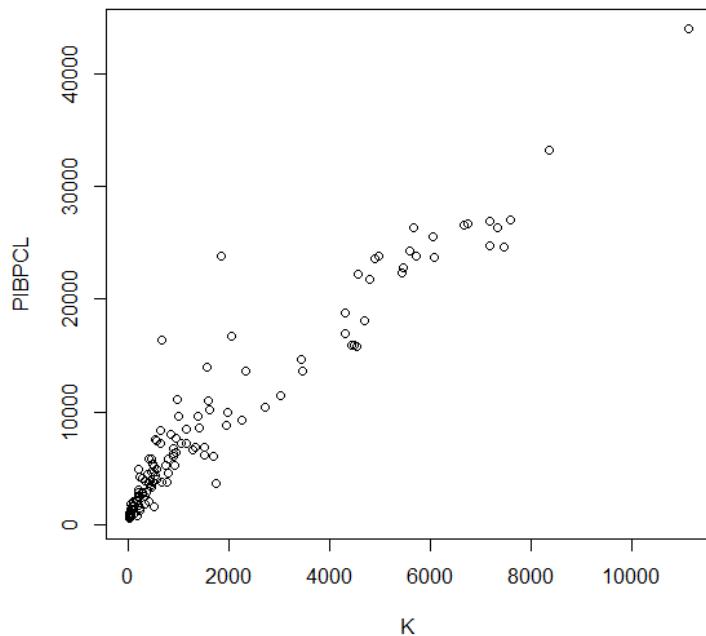
```
> attach(datos)
```

Ahora al pedir un listado a R aparecerá cada una de las variables en la lista:

```
ls()
```

```
The following object(s) are masked from 'datos (position 3)':  
K, PAIS, PIBPCL
```

Una herramienta gráfica que utilizaremos frecuentemente es un diagrama de dispersión. Por ejemplo, se puede solicitar una diagrama de dispersión para visualizar la relación entre el esfuerzo de inversión de los países y su ingreso per cápita:



En la gráfica se puede observar claramente una relación positiva entre el esfuerzo de inversión y el PIB per cápita de los países de la muestra de datos.

Como ya sabemos utilizar el comando de regresión podemos ahora estimar un modelo para explicar el ingreso per cápita de los países en función de su capital, pero ahora guardaremos el resultado en un objeto con nombre PWT:

```
PWT<-lm(PIBPCL ~ K)
```

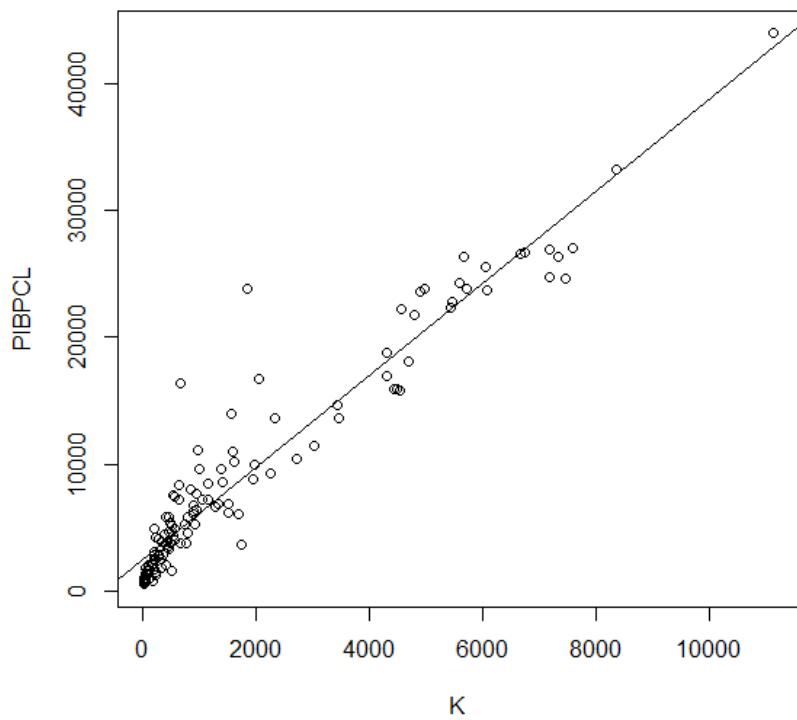
Los resultados del modelo indican que al incrementarse la inversión en un dólar el ingreso de los países se incrementa en 3.64 dólares, tal y como se aprecia en el cuadro de resultados siguientes.

```
> summary(PWT)
Call:
lm(formula = PIBPCL ~ K)
Residuals:
    Min      1Q      Median      3Q      Max 
-5180.3   -1553.1    -591.4     825.3   14757.2 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.364e+03 2.800e+02 8.443   5.06e-14 ***
K           3.641e+00 9.666e-02 37.668  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1 
Residual standard error: 2564 on 131 degrees of freedom
Multiple R-squared:  0.9155,  Adjusted R-squared:  0.9148 
F-statistic: 1419 on 1 and 131 DF, p-value: < 2.2e-16
```

La recta de regresión la podemos añadir al diagrama de dispersión que ya habíamos generado con la siguiente opción:

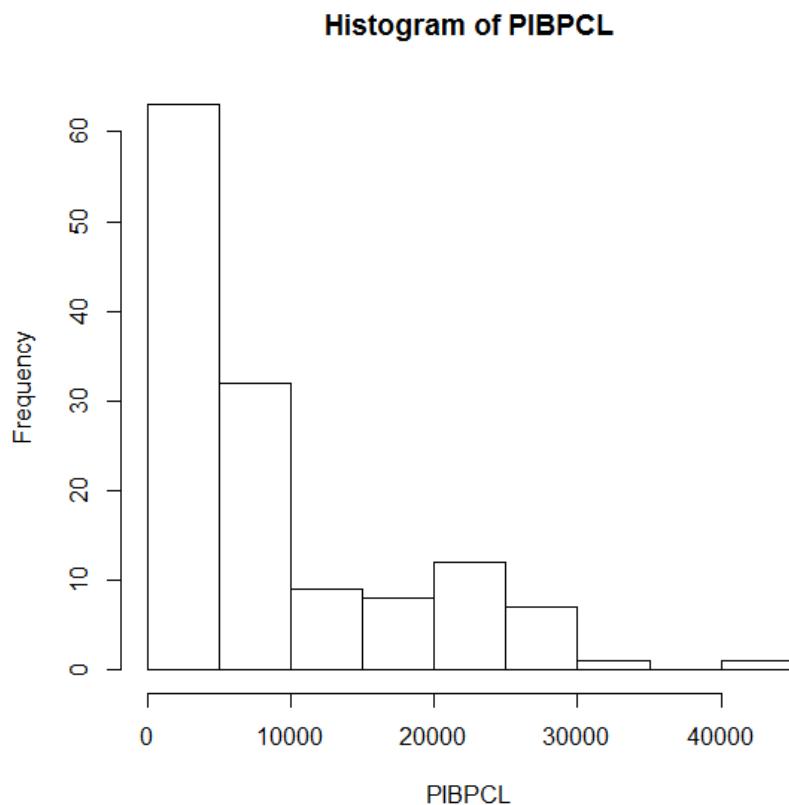
```
abline(PWT)
```

La gráfica resultante es la siguiente:



Otra gráfica que nos va a ser de utilidad es el histograma, en el cual podemos relacionar intervalos de los datos con sus frecuencias. Con la siguiente instrucción generaremos el histograma para los datos del PIB per cápita de los países:

```
hist(PIBPC)
```



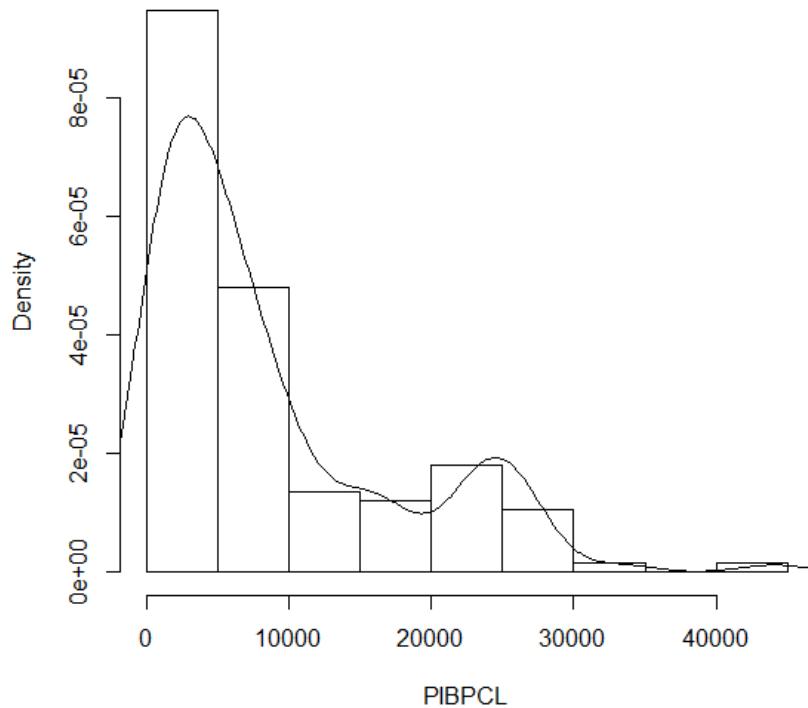
Claramente el histograma muestra que la mayoría de los países se encuentran en los ingresos más bajos de la distribución.

Resulta útil visualizar el histograma en densidades (área bajo la curva igual a la unidad) y añadirle funciones de densidad kernel, lo cual se puede hacer con la instrucción siguiente:

```
hist(PIBPCL,freq=FALSE)
> lines(density(PIBPCL))
```

La gráfica resultante es la siguiente:

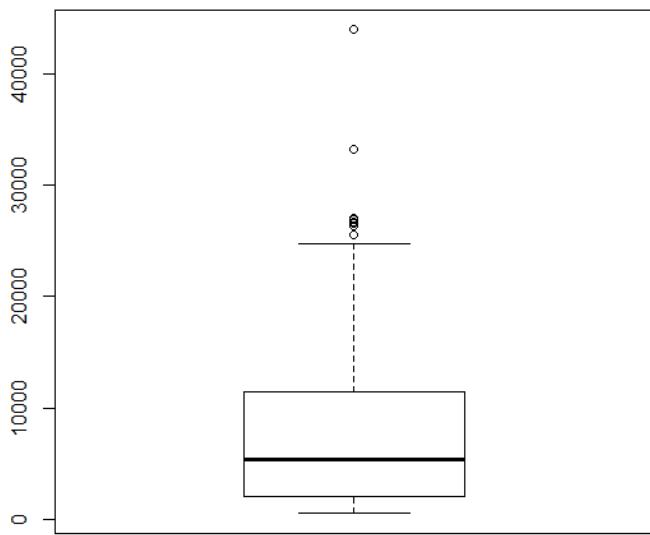
**Histogram of PIBPCL**



Para observar la distribución de los datos es utilizar cajas de box, en las cuales la caja muestra los umbrales para los cuartiles inferior y superior, además de la mediana. Las líneas abajo y arriba de la caja permiten identificar las observaciones extremas. Para obtener este tipo de gráficas se utiliza la instrucción siguiente:

```
boxplot(PIBPCL)
```

La gráfica resultante, muestra un grupo de países con ingresos extremos.

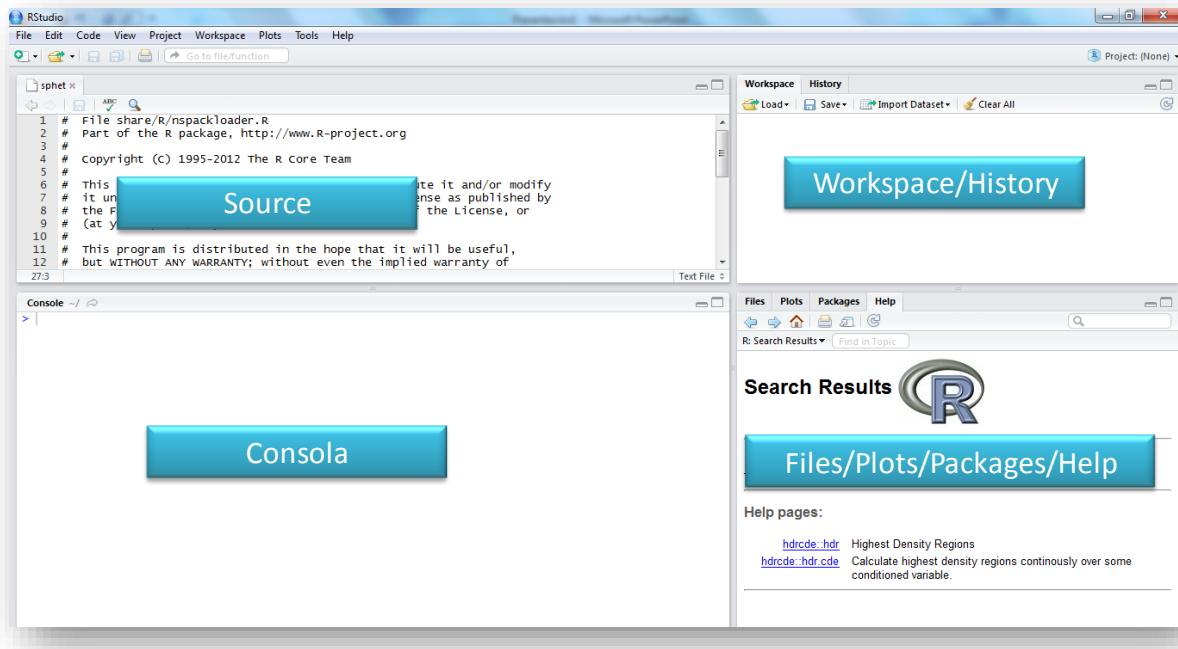


## 5. ALGUNOS DESARROLLOS EN R QUE FACILITAN EL USO DE LA ECONOMETRÍA

En R contamos con interfaces que nos permiten utilizar de forma más amigable los recursos disponibles en ese software. Una de estas interfaces es el RStudio, la cual se puede instalar desde el siguiente vínculo:

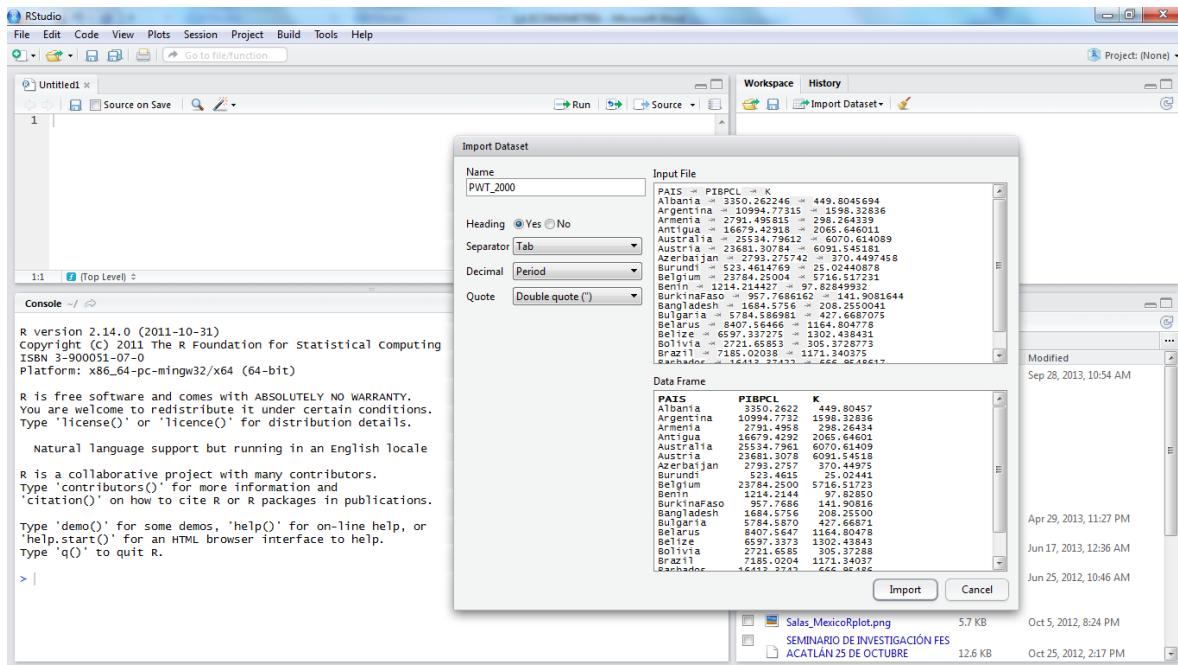
<http://www.rstudio.com>

La primer ventaja de RStudio es que permite visualizar los datos y su historial de trabajo en la ventana de WORKSPACE/HISTORY, al mismo tiempo es posible ver la ventana CONSOLA en la cual se ejecutan los comandos de R, cuanta también con una ventana en la cual se puede visualizar la ayuda (HELP), archivos (FILE), gráficas (PLOT) y paquetería (PACKAGES). La cuarta ventana es la de SOURCE en la que se muestran los archivos de origen.



Usted puede revisar la amplia documentación de este interface en el sitio ya referenciado, simplemente aquí haremos una demostración de las facilidades que nos ofrece.

Por ejemplo, para cargar la base de datos que ya hemos trabajado en el archivo de origen txt, basta con seleccionar de la ventana WORKSPACE la opción import data set y localizar el directorio en el cual está guardado nuestro archivo PWT\_2000.txt, tal y como se muestra en la imagen siguiente. El archivo se despliega en el editor de datos en el formato original delimitado por espacios (Input File) y en el formato de cuadro de datos de R (DataFrame). En la ventana ImportDataSet es suficiente con seleccionar el botón import para que el archivo sea importado al sistema.



Al cargar el archivo al sistema automáticamente se cargará en la ventana SOURCE y en la ventana del WORKSPACE como objeto de datos. En la consola se mostrará la secuencia de comandos usada por R para importar el archivo referido, tal y como se muestra en la siguiente figura.

The screenshot shows the RStudio interface with three main panes:

- Data View (Left):** Displays a table titled "PWT\_2000" with columns PAIS, PIBPCL, and K. The table contains 133 observations of 3 variables. A blue box labeled "Cuadro de datos" points to this pane.
- Object Explorer (Top Right):** Shows the object "PWT\_2000" under the "Data" category, described as "133 obs. of 3 variables". A blue box labeled "Objeto datos" points to this pane.
- Console (Bottom Left):** Shows the R startup message, command history, and the code used to import the data. A blue box labeled "Comandos ejecutados para importar los datos" points to the imported data command.

PAIS	PIBPCL	K
1 Albania	3350.2622	449.88457
2 Argentina	10994.7732	1598.32836
3 Armenia	2791.4958	298.26434
4 Antigua	16679.4292	2065.64601
5 Australia	25534.7961	6070.61409
6 Austria	23681.3078	6091.54918
7 Azerbaijan	2793.2757	370.44975
8 Burundi	523.4615	25.02441
9 Belgium	23784.2500	5716.51723

```

R version 2.14.0 (2011-10-31)
Copyright (C) 2011 The R Foundation for Statistical Computing
ISBN 3-900051-07-0
Platform: x86_64-pc-mingw32/x64 (64-bit)

R is free software and comes with ABSOLUTELY NO WARRANTY.
You are welcome to redistribute it under certain conditions.
Type 'license()' or 'licence()' for distribution details.

Natural language support but running in an English locale

R is a collaborative project with many contributors.
Type 'contributors()' or 'citation()' for more information.

Type 'demo()' for some demos, 'help()' for on-line help, or
'help.start()' for an HTML browser interface to help.
Type 'q()' to quit R.

> PWT_2000 <- read.delim("~/luisq/Investigacion/PAPIIT_PAPIME/capitulos_libro/PWT_2000.txt")
> View(PWT_2000)
> View(PWT_2000)
> View(PWT_2000)
>

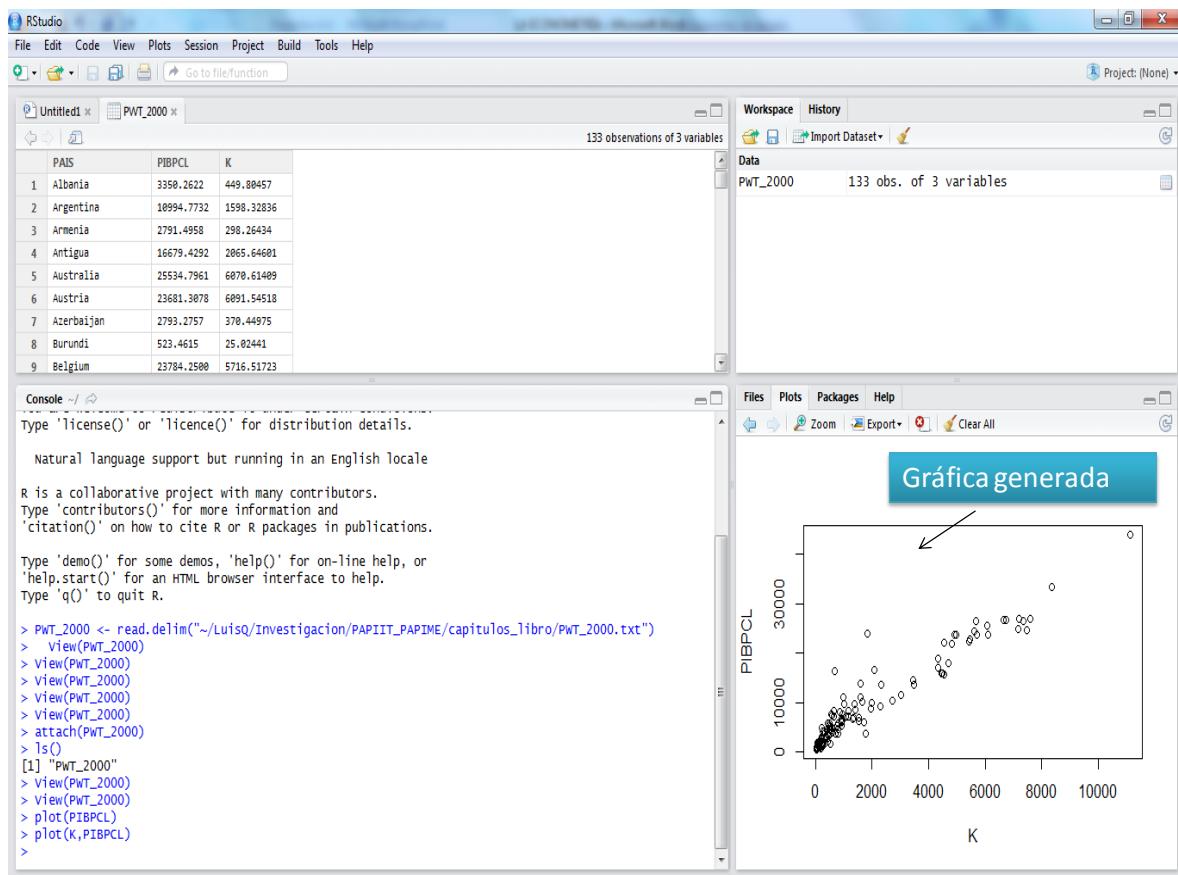
```

Con los datos es posible ahora realizar rápidamente gráficas para su análisis, en la ventana de consola se puede escribir la siguiente instrucción para generar el diagrama de dispersión que ya vimos en la sección previa:

```
plot(K,PIBPCL)
```

Como habrá podido notar, la consola cuenta con auto generación de los códigos de R, en este caso automáticamente se cierra el paréntesis de la instrucción capturada.

En la ventana de PLOTS se visualiza la gráfica que hemos generado y en el menú principal PLOTS permite guardar la gráfica, importarla como PDF o imagen, borrarla o hacerle un zoom poniéndola en una nueva ventana:



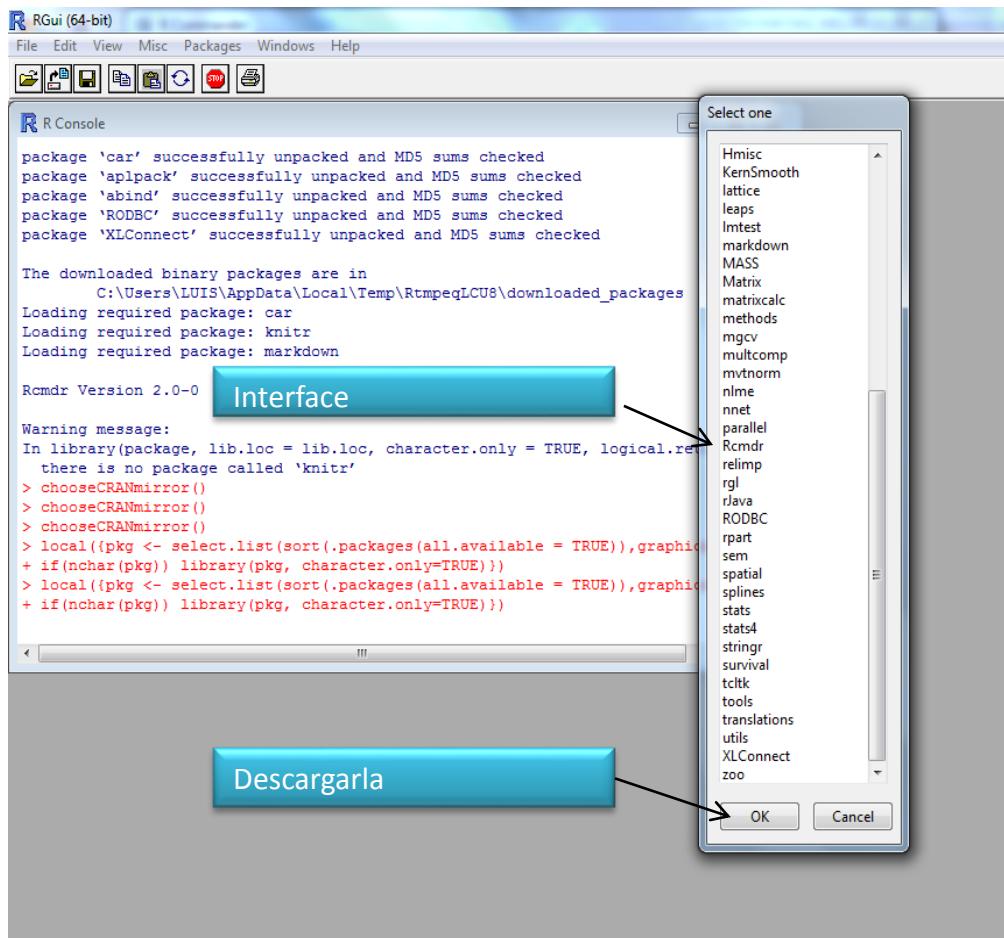
Todos los ejemplos de R que vimos en la sección previa puede ahora replicarlos utilizando el RStudio y se dará cuenta que es más accesible y su visualización en ventanas facilita mucho el trabajo.

Otro de los interfaces que nos será de gran utilidad es el RCommander, que fue desarrollado por John Fox de la Mc Master University en los Estados Unidos.

El RCommander es un paquete estadístico, por lo cual cuenta con todos los elementos para estimar una amplia gama de modelos econométricos (Fox, 2005).

Para instalar el interface es necesario descargarlo de algún espejo del CRAN, en el menú principal de R puede seleccionar PACKAGES/SET CRAN MIRROR y optar por el USA(CA1).

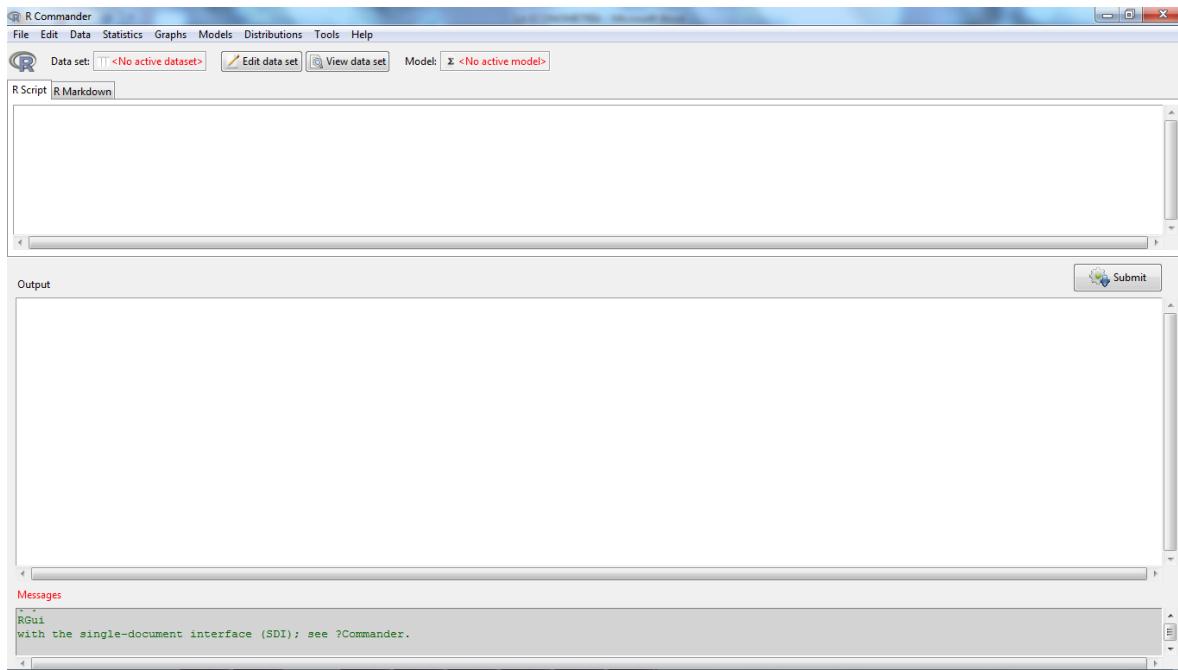
Ahora en la opción del menú principal PACKAGES/LOAD PACKAGES podrá visualizar el interface Rcmdr y al seleccionar OK se descargará, tal y como se muestra a continuación:



Una vez descargado se puede activar con la siguiente instrucción:

```
library(Rcmdr)
```

Si realizó todo correctamente podrá visualizar la ventana del interface del RCommander que aparece a continuación:

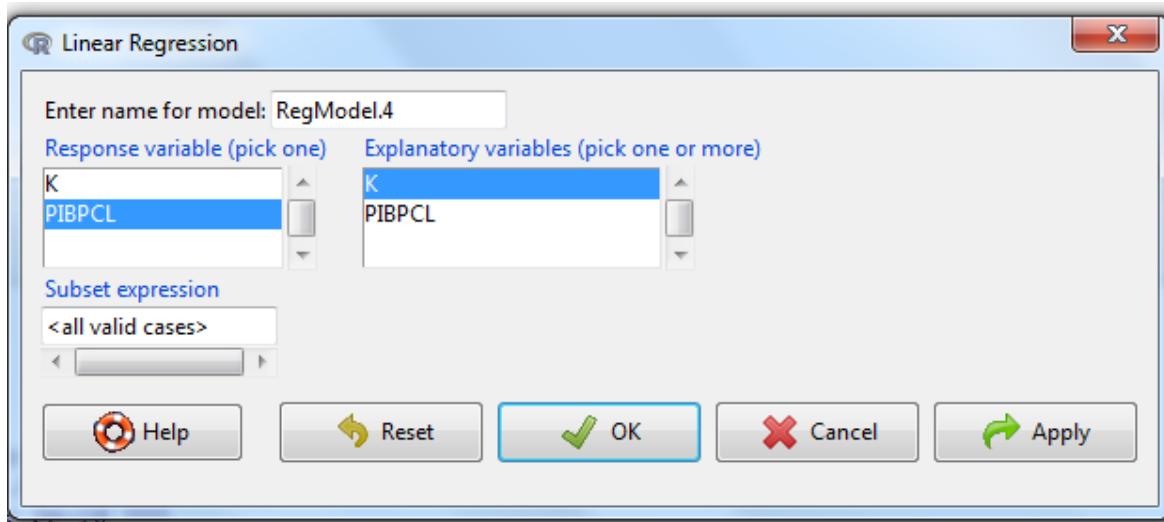


En el menú principal de RCommander con la opción DATA es posible importar nuestra base de datos, usted podrá constatar que las opciones de importación son más amplias que en R al dar la posibilidad de abrir directamente archivos de Excel, Stata, SPSS, SAS y Minitab. Por ejemplo, podemos abrir nuestro archivo txt con las opciones del menú DATA/IMPORT DATA/ FROM TEXT FILE, una vez cargada la base de datos se puede visualizar en el DATASETseleccionando el botón VIEW DATA SET en la segunda línea de botones superiores del interface, el resultado se muestra en la siguiente figura:

The screenshot shows the R Commander interface. On the left, there's an R Script pane with code for reading a dataset from a file. The main area is a 'Dataset' window showing a table with three columns: PAIS, PIBPCL, and K. The table lists 39 countries with their respective values. A 'Submit' button is visible in the top right of the dataset window.

	PAIS	PIBPCL	K
1	Albania	3350.2622	449.80457
2	Argentina	10994.7732	1598.32836
3	Armenia	2791.4958	298.26434
4	Antigua	16679.4292	2065.64601
5	Australia	25534.7961	6070.61409
6	Austria	23681.3077	6091.54518
7	Azerbaijan	2793.2757	370.44975
8	Burundi	523.4615	25.02441
9	Belgium	23784.2500	5716.51723
10	Benin	1214.2144	97.82850
11	BurkinaFaso	957.7686	141.90816
12	Bangladesh	1684.5756	208.25500
13	Bulgaria	5784.5870	427.66671
14	Belarus	8407.5647	1164.80478
15	Belize	6597.3374	1302.43843
16	Bolivia	2721.6585	305.37288
17	Brazil	7185.0204	1171.34037
18	Barbados	16413.3742	666.95486
19	Canada	26922.2121	7191.67689
20	Switzerland	26421.5907	7342.46836
21	Chile	9919.9324	1981.07606
22	China	3746.9831	784.60238
23	Coted'Ivoire	1869.1052	112.86159
24	Cameroon	2042.0278	120.55155
25	Congo	1807.4095	64.07838
26	Colombia	5380.1337	493.36151
27	Comoros	1576.8348	75.84440
28	CapeVerde	4026.5281	573.69322
29	Costa	5863.3682	801.90091
30	CzechRepublic	13673.1313	3477.79971
31	Denmark	26627.1630	6667.25165
32	DominicanRepublic	5270.8389	754.01652
33	Algeria	4893.6792	577.83513
34	Ecuador	3467.2110	465.44818
35	Egypt	4184.3091	250.45181
36	Spain	18054.6479	4693.39272
37	Estonia	9588.4871	1384.34600
38	Ethiopia	635.0624	27.91341
39	Finland	23798.4798	4973.10217

Una vez cargados los datos el interface permite realizar múltiples funciones estadísticas y estimar modelos con el menú STATISTICS o evaluar los modelos estimados con el menú MODELS. Por ejemplo, para correr la regresión entre el PIB per cápita de los países y el capital activamos el menú STATISTICS/FIT MODELS/LINEAR REGRESSION. A continuación se abrirá una ventana con las opciones para seleccionar la variable dependiente y las explicatorias, tal y como se muestra en la imagen siguiente:



Una vez que se selecciona el botón de OK los resultados de la regresión se despliegan en la ventana de resultados (OUTPUT) del RCommander, como se muestra a continuación:

R Commander

Data set: Dataset Model: RegModel4

R Script R Markdown

```
showData(Dataset, placement='20+200', font=getRcmdr('logFont'), maxwidth=80, maxheight=30)
names(Dataset)
summary(Dataset)
RegModel.3 <- lm(PIBPCL~K, data=Dataset)
summary(RegModel.3)
AIC(RegModel.3)
RegModel.4 <- lm(PIBPCL~K, data=Dataset)
summary(RegModel.4)
```

Output

```
Call:
lm(formula = PIBPCL ~ K, data = Dataset)

Residuals:
    Min      1Q  Median      3Q     Max 
-5180.3 -1553.1  -591.4   825.3 14757.2 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 2.364e+03 2.800e+02  8.443 5.06e-14 ***
K           3.641e+00 9.666e-02 37.668 < 2e-16 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

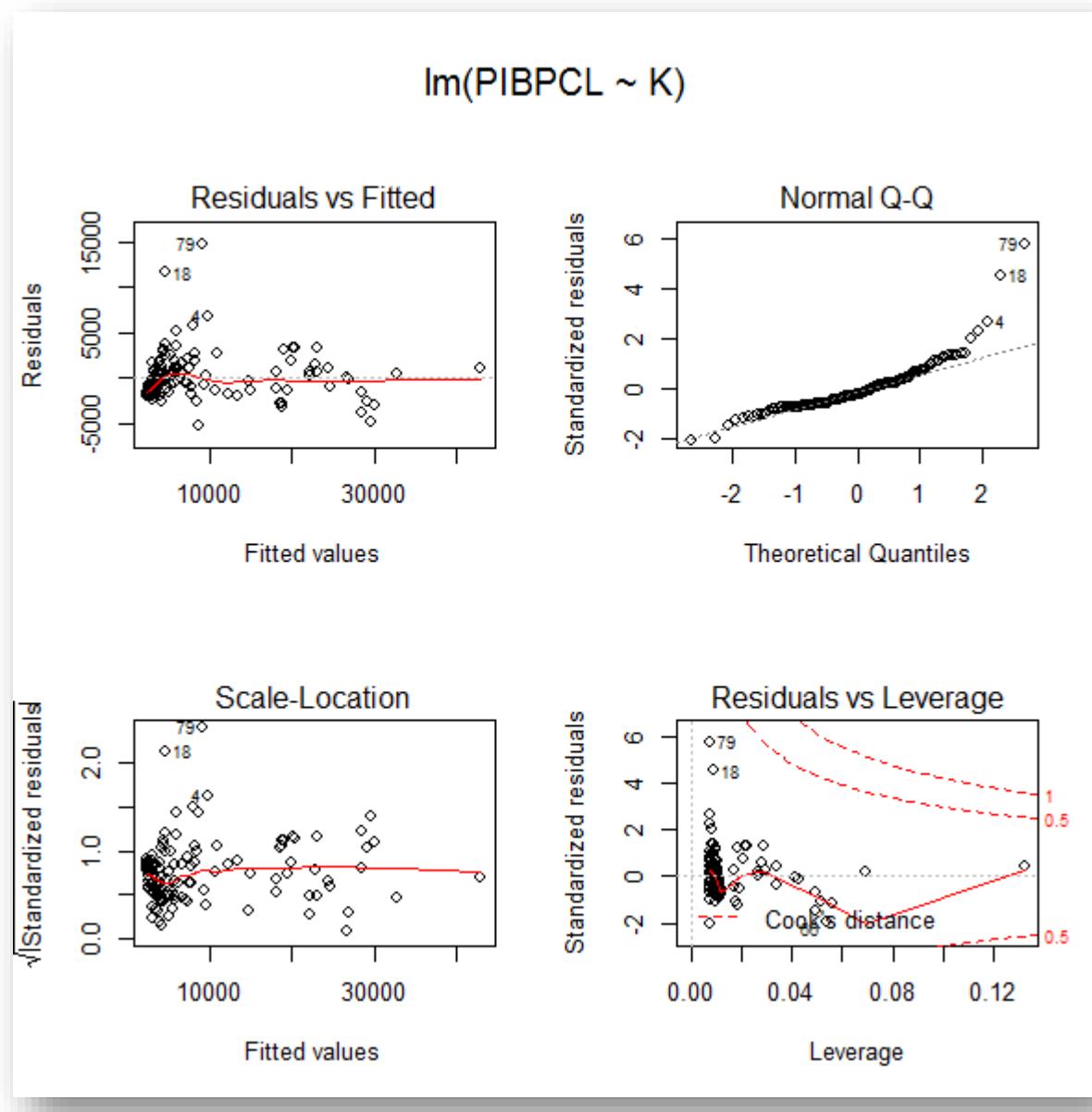
Residual standard error: 2564 on 131 degrees of freedom
Multiple R-squared:  0.9155, Adjusted R-squared:  0.9148 
F-statistic: 1419 on 1 and 131 DF,  p-value: < 2.2e-16
```

Messages

```
[6] ERROR: No explanatory variables selected.
[7] WARNING: There is only one model in memory.
```

Finalmente con el menú MODELS se cuenta con amplias posibilidades para realizar pruebas de hipótesis y diagnósticos de los resultados que serán estudiados en los capítulos siguientes de este libro. Por el momento, en la figura siguiente, simplemente se muestra como ejemplo la forma en que RCommander

despliega una batería gráfica para evaluar los residuales y estimaciones del modelo.



## REFERENCIAS

- Crawley, J. Michael (2009), *The R book*, ed. Wiley, Inglaterra.  
Fox, John (2005), *The R Commander: A Basic-Statistics Graphical User Interface to R*, Journal of Statistical Software, vol.14, núm. 9, pp. 1-42.

Hoover D., Kevin (2006), The methodology of econometrics, en Terence Mills y Kerry Patterson, Plagrave Handbook of Econometrics, vol.1, Econometric Theory, Palgrave Mcmillan, pp. 61-87, Reino Unido.

Maddala, G. S. (1996). *Introducción a la econometría*. Ed. Prentice Hall, México.

Spanos, Aris (1996). *Statistical Foundation of econometric modeling*. Ed. Cambridge University Press.

Spanos, Aris (2006), Econometrics in retrospect and prospect, en Terence Mills y Kerry Patterson, Plagrave Handbook of Econometrics, vol.1, Econometric Theory, Palgrave Mcmillan, pp. 3-58, Reino Unido.

Venables, W. N. y D. M. Smith (2013), An introduction to R, ed. R Core Team.

## REFERENCIAS ELECTRÓNICAS

CRAN (2013), <http://www.r-project.org/>

IHS (2013), <http://www.ihs.com/products/global-insight/country-analysis/mexico-economic-forecasts.aspx>

Penn Tables (2013), <https://pwt.sas.upenn.edu/>

RStudio (2013), <http://www.rstudio.com>

## ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO

PWT\_2000.txt

## MATERIAL DE APRENDIZAJE EN LÍNEA

Teórica\_Cap1

Práctica\_Cap1

VideoPráctica\_Cap1

VideoTeoría\_Cap1

# CAPÍTULO 2: ENFOQUE MATRICIAL DE LA REGRESIÓN LINEAL

JAVIER GALÁN FIGUEROA

## 1. EL MODELO MATRICIAL

En este capítulo se considera relevante que el usuario conozca, en primera instancia, las rutinas básicas que son necesarias para estimar los parámetros de la regresión lineal a través del enfoque matricial, utilizando la paquetería del software **R**, los cuales podrán ser utilizados en sus variantes como es el **RStudio**.

Para comenzar, se utilizarán datos de la economía mexicana para el periodo enero de 2009 a diciembre de 2013, con frecuencia mensual y cuya fuente provienen de la página web del Banco de México ([www.banxico.gob.mx](http://www.banxico.gob.mx)), con dicha información permitirá estimar el siguiente modelo:

$$y = f(X_2, X_3) \quad (1)$$

$$y = X\beta + u \quad (2)$$

$$y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + u_t \quad (3)$$

La ecuación (2) es la representación matricial de la regresión lineal, donde **y** es un vector columna de orden  $(n \times 1)$ , **X** es una matriz de orden  $(n \times k)$ ,  **$\beta$**  es un

vector columna de orden  $(k \times 1)$ , por último  $\mathbf{u}$  es un vector columna de orden  $(n \times 1)$ , es decir<sup>1</sup>:

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} \quad (4)$$

$$(n \times 1) \quad (n \times k) \quad (k \times 1) (n \times 1)$$

De la ecuación (3) la variable dependiente,  $\mathbf{y}$ , es el nivel de deuda pública del gobierno mexicano (miles de millones de pesos) que es explicada por el nivel de reservas internacionales, **X2**, (miles de millones de dólares) y por el índice bursátil de la Bolsa Mexicana de Valores, **X3** (miles de unidades).

Para encontrar el modelo en el cual explique el comportamiento de la deuda externa en función de la reserva internacional y del índice bursátil se utilizará los datos que se encuentran en el archivo CAP2\_MCO con extensión CSV (delimitado por comas). Para ejecutarlo en R se hace uso del siguiente código:

```
> deuda<-read.csv("C:/data/cap2_mco.csv", header =T)
> attach(deuda)
```

Si el usuario desea visualizar los datos a través de una lista, basta con escribir:

```
> deuda
```

---

<sup>1</sup> Para el desarrollo correspondiente a la teoría econométrica del presente capítulo se ha consultado los siguientes autores Quintana y Mendoza (2008), Green (2003) y Dinardo (1997).

	y	x2	x3
1	3.15082	84.2291	21.89885
2	3.11256	83.1969	24.33171
3	3.11546	81.5107	24.36838
4	3.19073	82.2459	27.04350
5	3.24860	84.6815	28.12995
6	3.31421	87.8285	29.23224
7	3.38451	88.6224	28.64603
8	3.39667	90.7154	30.95711
9	3.45910	99.8701	32.12047
10	3.48556	98.7063	30.39161
11	3.45893	100.0099	31.63454
12	3.48903	101.6403	33.26643

## 2. ANÁLISIS EXPLORATORIO DE LOS DATOS

Después de haber cargado los datos al programa, se procederá a realizar el siguiente análisis estadístico de las variables.

Si se desea obtener de manera individual los siguientes parámetros: media aritmética, mediana, desviación estándar y varianza de la variable (y) se escribe:

```
> mean(y)
> median (y)
> sd(y)
> var(y)
```

De manera conjunta se puede utilizar:

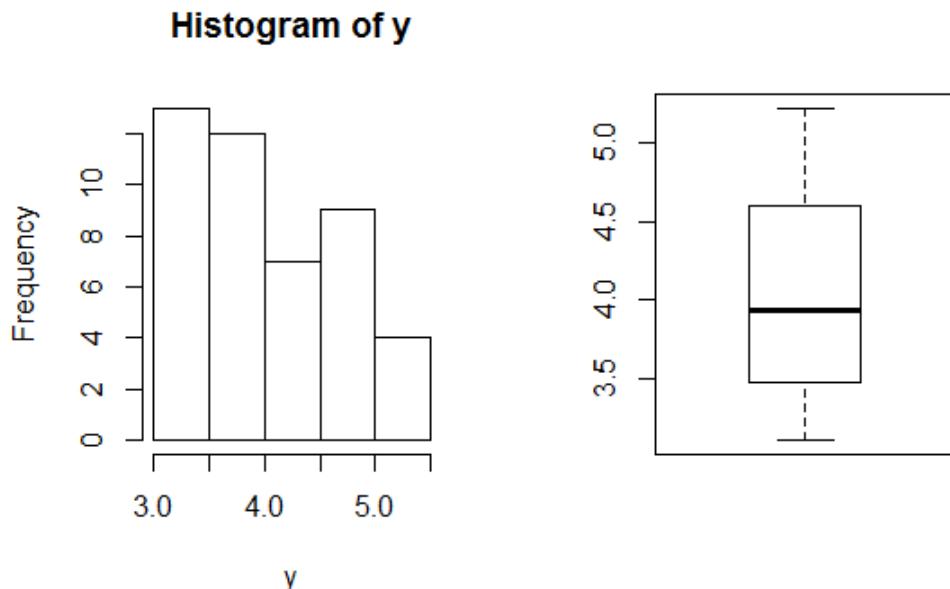
```
> summary (y)
```

El cual arroja como los siguientes resultados para el periodo de estudio: 1) el valor mínimo de la deuda pública es de 3.113, con un máximo de 5.221mil millones de pesos. Con un nivel de endeudamiento medio de 4.054 mil millones de pesos.

```
> summary (y)
Min. 1st Qu. Median Mean 3rd Qu. Max.
3.113 3.486 3.942 4.054 4.603 5.221
```

Del anterior código, el programa R agrupa los datos y calcula los cuartiles donde el primero es 3.486, mientras el segundo o mediana es de 3.942 y el tercero de 4.603. Posteriormente se obtiene el histograma y la gráfica de caja en un sólo gráfico

```
> split.screen(c(1,2))
> hist(y)
> screen(2)
> boxplot(y)
```

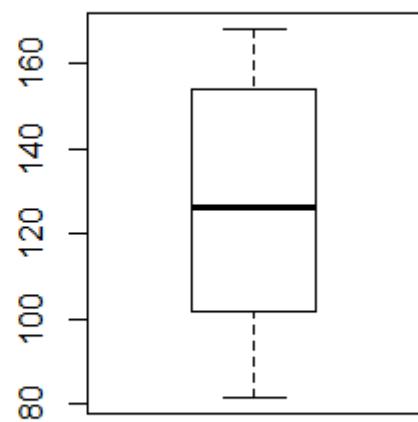
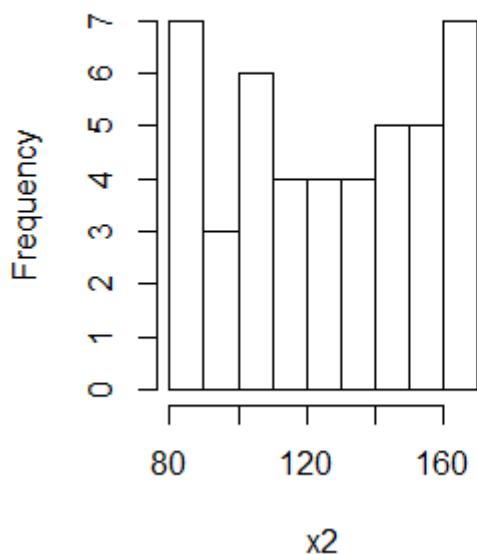


Repitiendo el mismo código para las variables X2 y X3 se tiene los resultados siguientes:

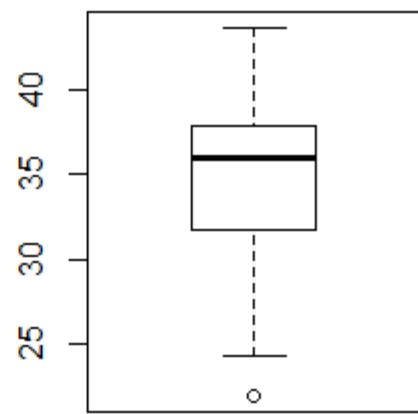
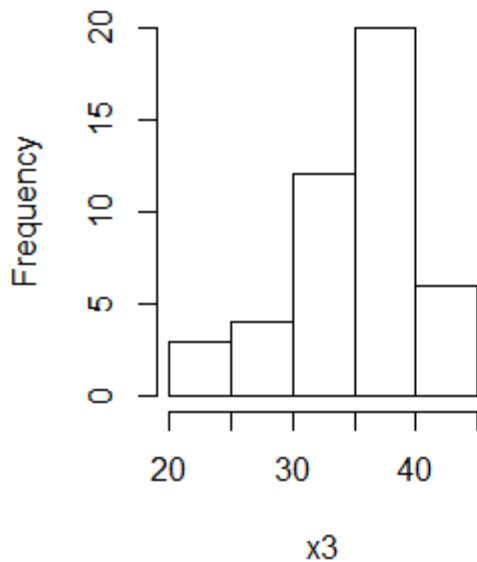
```
> summary (X2,X3)

> summary (X2,X3)
  Min. 1st Qu. Median Mean 3rd Qu. Max.
81510 101600 126500 126000 154100 168300
```

### Histogram of x2



### Histogram of x3



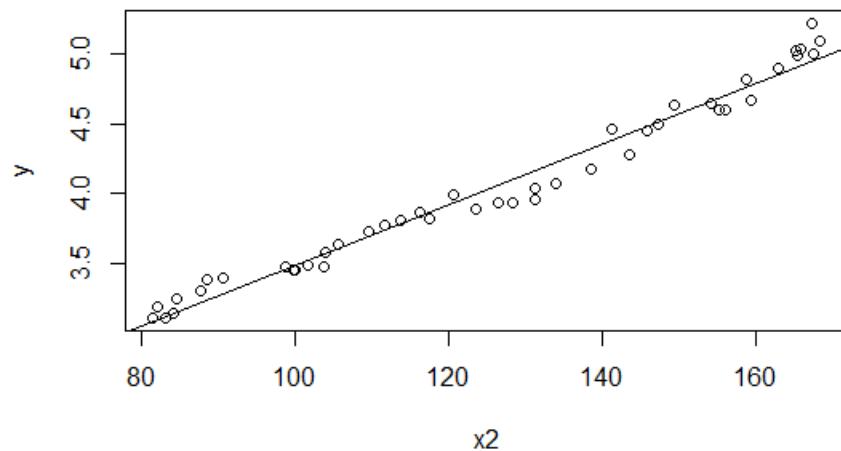
En el análisis de la variable X3 que representa el índice bursátil, se aprecia en su gráfico de caja un valor atípico u outlier que se localiza por debajo del límite inferior, esta observación podría implicar problemas de varianza en el modelo, por el momento sólo se indica su presencia. A continuación se utiliza el siguiente código para obtener la matriz de correlación entre las variables (y, X2, X3).

```
> cor(deuda)
```

```
> cor(deuda)
      y     x2     x3
y  1.0000000 0.9874354 0.8960622
x2 0.9874354 1.0000000 0.9126517
x3 0.8960622 0.9126517 1.0000000
```

De acuerdo a la matriz de correlación, la asociación entre las variables (X2,y) es positiva y del 0.9874 o del 98.74 por ciento. Mientras la asociación entre (X3,y) es de igual manera positiva y del 89.60 por ciento. Por otro lado, las variables (X2,X3) se asocian en 91.26 por ciento. Para obtener los diagramas de dispersión para indicar a nivel gráfico como influye la reserva internacional (X2) y el índice bursátil (X3) al nivel de endeudamiento del gobierno mexicano (y) se prosigue con el siguiente código.

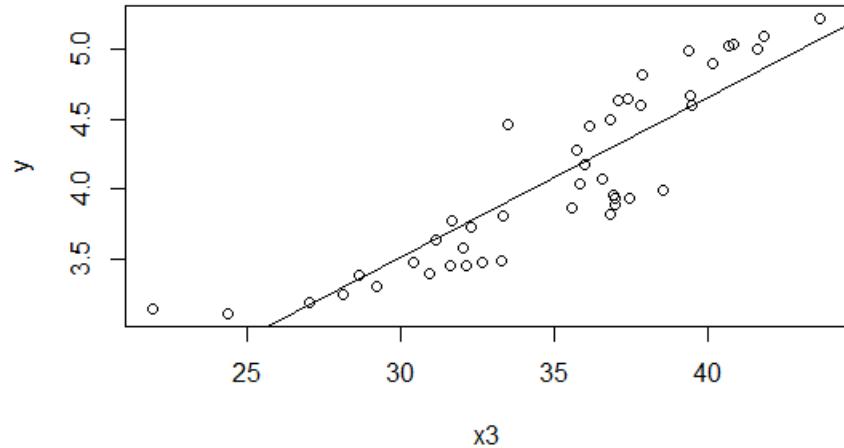
```
> scatter1<-plot(y~x2)
> fit<-lm(y~x2)
> abline(fit)
```



```

> scatter1<-plot(y~x3)
> fit2<-lm(y~x3)
> abline(fit2)

```



### 3. ESTIMACIÓN POR MÍNIMOS CUADRADOS ORDINARIOS

Con el análisis previo se procederá a estimar los parámetros de la ecuación ( 3 ) a través de los Mínimos Cuadrados Ordinarios (MCO). Para ello se considera que el vector  $\beta$  de la ecuación ( 2 ) es estimable a partir de la siguiente expresión<sup>2</sup>:

$$\beta = (X'X)^{-1} X'y \quad (5)$$

Como primer paso se debe especificar en el programa R la matriz  $X$  así como el vector  $y$ . Para ello se sigue el siguiente algoritmo: 1) Para transformar un conjunto

---

<sup>2</sup> Si el lector se encuentra interesado en revisar el proceso de derivación del vector de los estimadores por Mínimos Cuadrados Ordinarios puede consultar los manuales que se encuentran en citados en la sección de referencias del presente capítulo.

de variables a matriz se utiliza el código “**cbind()**”; y 2) Una vez que se ha dado de alta las matrices en R se procede a realizar las operaciones correspondientes para encontrar los componentes del vector  $(X'X)^{-1}X'Y$  los cuales se describen a continuación.

Para crear la matriz **X**, que conforma de acuerdo a la ecuación (4), se utiliza el siguiente código:

```
> X<-cbind(1,X2,X3)
```

Donde las opciones que aparecen dentro del paréntesis indican que el uno hace referencia al intercepto, mientras x2 y x3 a las variables reserva internacional y al índice bursátil. Para el caso para transformar la variable deuda pública (y) a vector se utiliza el mismo código.

```
> y1<-cbind(y)
```

Para estimar el vector  $\beta$  de la ecuación ( 5 ), primero se obtiene el producto  $(X'X)$  para ello se sigue los siguientes pasos:1) transpuesta de **X**; 2) Producto de la transpuesta de **X** por **X**, cabe mencionar, en el programa R el producto de matrices se lleva a cabo mediante el código “%\*%”.

```
> trX<-(t(X))
```

```
> X_X<-trX %*% X
```

```
> X_X
```

```
> X_X
      x2      x3
 45.000 5670.887 1562.692
x2 5670.887 750953.005 202649.321
x3 1562.692 202649.321 55348.376
```

A continuación se obtiene el determinante de la matriz  $(X'X)$ , para determinar si ésta tiene inversa o no. Para obtener la inversa  $(X'X)^{-1}$ , se debe primero activar la librería “**library(MASS)**”, después utilizar el código “**ginv()**”.

```
> det(X_X)
> library(MASS)
> invX_X<-ginv(X_X)
> invX_X
```

```
> invX_X
      [,1]     [,2]     [,3]
[1,] 1.68427699 0.0094969200 -0.0823249040
[2,] 0.00949692 0.0001648509 -0.0008717089
[3,] -0.08232490 -0.0008717089 0.0055340314
```

Una vez que se tiene la inversa  $(X'X)^{-1}$ , se procede a obtener el producto  $X'y$

```
> Xy<-trX %*% y1
> Xy
```

```
> Xy
      y
 182.422
x2 23775.420
x3 6458.080
```

Por último, se procede a calcular al vector beta a través del siguiente código

```
> beta<-invX_X %*% Xy
```

```
> beta
```

```
> beta  
      y  
[1,] 1.381548779  
[2,] 0.022279261  
[3,] -0.003897697
```

Un método de comprobación para tener la certeza que este vector, el cual fue obtenido paso a paso mediante álgebra lineal, se utiliza el código para estimar de manera directa la regresión lineal “**lm(y~x)**”, cabe mencionar que el programa R utiliza el mismo método.

```
> modelo<-lm(y~x2+x3)
```

```
> summary(modelo)
```

Call:  
lm(formula = y ~ x2 + x3)

Residuals:

Min	1Q	Median	3Q	Max
-0.20360	-0.08483	0.01550	0.06820	0.28696

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.381549	0.131886	10.475	2.75e-13 ***
x2	0.022279	0.001305	17.075	< 2e-16 ***

```
x3      -0.003898  0.007560 -0.516   0.609  
---  
Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

Residual standard error: 0.1016 on 42 degrees of freedom  
Multiple R-squared: 0.9752, Adjusted R-squared: 0.974  
F-statistic: 825.3 on 2 and 42 DF, p-value: < 2.2e-16

Se aprecia que el vector beta encontrado coincide con los coeficientes estimados por el código “lm(y~x)”. Por tanto la ecuación estimada se define como sigue:

$$y = 1.381549 + 0.022279X_2 - 0.003898X_3 \quad (6)$$

## REFERENCIAS

Crawley, Michael (2013), *The R Book*, 2<sup>a</sup>. Ed., Wiley, United Kingdom.

Green, William (2003), *Econometric Analysis*, 5<sup>a</sup> Ed., Pearson Education. EUA.

Johnston, J. y J. Dinardo (1997), *Econometrics Methods*, 4<sup>a</sup> Ed., McGraw-Hill. EUA.

Quintana, L. y M. A. Mendoza (2008), *Econometría Básica. Modelos y aplicaciones a la economía mexicana*, Plaza y Valdés Editores, México.

## ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO

cap2\_mco.csv

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap2

Práctica\_Cap2

VideoPráctica\_Cap2

VideoTeoría\_Cap2

# **CAPITULO 3: EL MODELO DE REGRESIÓN MÚLTIPLE**

**Jorge Feregrino Feregrino**

## **1. ESPECIFICACIÓN DEL MODELO DE REGRESIÓN MÚLTIPLE**

El primer paso en la especificación de un modelo econométrico es identificar el objeto de investigación en relación al área de estudios de las ciencias socioeconómicas. En esta etapa, es necesario recopilar información acerca del comportamiento teórico del objeto de investigación para identificar patrones de comportamiento, situar alguna problemática específica y plantear las hipótesis necesarias. La especificación del modelo nos permitirá explorar las hipótesis principales, identificar las relaciones que explican el objeto de estudios y diseñar una propuesta teórica alternativa de acuerdo a los objetivos del usuario.

La identificación del objeto de investigación permitirá realizar una búsqueda exhaustiva de los datos para llevar a cabo una aproximación del comportamiento del fenómeno mediante los hechos estilizados. Una vez identificada la problemática se procede a establecer las relaciones y la selección de las variables. La búsqueda de la información de las variables, la relación teórica y la descripción estadística de estas será útil para determinar la metodología de análisis. En el caso de la mayoría de los hechos socioeconómicos los fenómenos están determinados por un conjunto de variables que puede llegar a ser infinito.

En economía se pueden identificar diversas relaciones teóricas entre variables; por ejemplo la producción para la teoría neoclásica está determinada por la combinación entre capital y trabajo, en la teoría keynesiana el ingreso de una economía cerrada está determinado por el consumo, la inversión y el gasto de gobierno, la tasa de inflación se puede determinar por la brecha del producto y las expectativas de inflación dentro del esquema de metas de inflación; así los ejemplos anteriores representan algunas de las problemáticas que se resuelven a través del establecimiento de relaciones entre variables.

En los modelos econométricos se establecen a priori las relaciones funcionales, con los elementos que se han descrito, para identificar los vínculos fundamentales entre las variables seleccionadas. De esta forma, se establecen las variables independientes y las dependientes. La elección de la variable dependiente y las independientes conformarán una relación funcional múltiple para describir el fenómeno económico mediante la metodología econométrica propuesta.

En el modelo de regresión múltiple las variables exógenas ( $X_j$ ), asociadas a coeficientes lineales constantes ( $\beta_j$ ), indican el efecto condicionado de cada variable independiente sobre la variable dependiente ( $Y$ ), la especificación general del modelo con cuatro variables independientes es la siguiente:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4$$

Por ejemplo: El administrador de una tienda quiere determinar los mejores criterios para elegir la localización de algunas tiendas, una de las primeras sugerencias

para la especificación del modelo es elegir la variable dependiente en este caso serían las ventas

$$Y = \text{Ventas}$$

Posteriormente, se realiza la recomendación sobre la elección de las variables independientes, en este caso la teoría plantea que múltiples variables inciden en el comportamiento de las ventas (Y), se consideran las siguientes:

$$X_1 = \text{Tamaño de la tienda}$$

$$X_2 = \text{Tráfico de personas en la calle}$$

$$X_3 = \text{Tiendas rivales en la zona}$$

$$X_4 = \text{Renta per capita de la población residente en la zona}$$

$$X_5 = \text{Número total de personas que residen en la zona}$$

La especificación sería una forma funcional lineal, donde se busca encontrar el grado de relación entre la variable endógena (Y) con las variables exógenas  $X_1, X_2, \dots, X_5$ . La forma funcional en la mayoría de los modelos, debe incorporar los errores que se generan en la estimación de la relación funcional entre las variables. La relación entre las variables es inexacta, por lo tanto, la evaluación se realiza en términos probabilísticos.

Ejercicio en R: Retomando el ejemplo de localización de tiendas y a fin de estimar el modelo de regresión se debe importar la base de datos a la cual se asignará el nombre “tiendas”, a la columna de datos de la variable dependiente se le asignará

el nombre “ventas”, mientras que los nombres de las variables independientes quedarán de la siguiente forma:

X<sub>1</sub>: “tamaño”

X<sub>2</sub>: “tráfico”

X<sub>3</sub>: “rivales”

X<sub>4</sub>: “renta”

X<sub>5</sub>: “residentes”

El comando para importar los datos desde Excel es el siguiente:

```
tiendas<-read.delim("ruta de acceso",sep=",",header=T,stringsAsFactors=F)
```

La forma funcional reducida de la estimación de la regresión múltiple, al expresarse en términos probabilísticos debe incorporar un término de error ( $\varepsilon_i$  ).

$$\hat{y}_i = b_0 + \sum_{j=1}^k b_j x_{ji} + \varepsilon_i$$

La estimación de una regresión múltiple tiene los siguientes objetivos:

- 1) Estimar los valores de una variable independiente ( $\hat{y}$ ) mediante una función lineal de un número (K) variables independientes observadas  $x_j$ , donde  $j = 1, \dots, K$

La representación es la siguiente,

$$\hat{y}_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki}$$

Donde  $i = 1, \dots, n$  de observaciones.

- 2) Obtener los efectos estadísticos de cada variable independiente, mediante la estimación de los coeficientes  $b_j$ , sobre la variable dependiente ( $\hat{y}$ ). El coeficiente  $b_j$  de cada variable dependiente indica el impacto que tiene una variación unitaria de  $x_j$ , descontando el efecto simultaneo que tienen las otras variables independientes, es decir, se mantiene la independencia entre estas variables.
- 3) Estimar la exogeneidad débil, para mostrar que la distribución marginal de la variable independiente, al no contener información relevante para estimar los parámetros de interés, se puede eliminar.

El modelo de regresión múltiple poblacional sería el siguiente:

$$y_i = \beta_0 + \beta_1x_{1i} + \beta_2x_{2i} + \cdots + \beta_kx_{ki} + \varepsilon_i$$

El modelo de regresión múltiple de una muestra de datos sería el siguiente:

$$y_i = b_0 + b_1x_{1i} + b_2x_{2i} + \cdots + b_kx_{ki} + e_i$$

El modelo de regresión múltiple permite obtener estimaciones simultáneas de  $b_j$  a partir del modelo poblacional  $\beta_j$

## 2. ESTIMACIÓN DE LOS COEFICIENTES DE REGRESIÓN

La estimación de la forma funcional múltiple, parte de los siguientes supuestos sobre los coeficientes a obtener:

- 1) Las variables independientes  $x_{ji}$  son números fijos o bien variables aleatorias  $X_j$ , independientes del término de error  $\varepsilon_i$ .
- 2) El valor esperado de la variable aleatoria ( $\hat{y}$ ) es una función de las variables independientes  $X_j$
- 3) Los términos de error  $\varepsilon_i$  son variables cuya media esperada es igual a cero y la varianza es constante  $\sigma^2$  para todas las observaciones:

$$E[\varepsilon_i] = 0 \quad y \quad E[\varepsilon_i^2] = \sigma^2 \text{ para } (i = 1, \dots, n)$$

- 4) Los términos de error aleatorios  $\varepsilon_i$ , no están correlación entre sí

$$E[\varepsilon_i \varepsilon_j] = 0 \quad \text{para todo } i \neq j$$

- 5) No es posible hallar un conjunto de números que no sean iguales a cero tal

que ,

$$c_0 + c_1 x_{1i} + c_2 x_{2i} + \dots + c_k x_{ki} = 0$$

Esto probaría la ausencia de relación lineal entre las  $X_j$ .

Los primero 4 supuestos están implícitos en la regresión simple, el 5to excluye cualquier posibilidad de relación lineal entre las variables independientes, y nos

permite hacer una selección específica de las variables y su impacto sobre la variable independiente en una regresión múltiple.

El método utilizado para estimar los coeficientes de la regresión múltiple es el de Mínimos Cuadrados Ordinarios (MCO), los coeficientes se obtienen mediante la minimización de los errores o la suma de residuos explicados al cuadrado SCE. En un primer momento los errores en el tiempo están explicados por las desviaciones de la variable independiente observada ( $y_i$ ) en el tiempo en relación a la variable explicada ( $\hat{y}_i$ ):

$$e_i = y_i - \hat{y}_i$$

Para minimizar la SCE se procede de la siguiente forma, matemática la SCE tiene la siguiente representación:

$$SCE = \sum_{i=1}^n e_i^2 = SCE = \sum_{i=1}^n (y_i - \hat{y}_i)^2$$

De la sumatoria se extraen las diferencias elevadas al cuadrado entre los valores de ( $y_i$ ) y los valores de la variable estimada  $\hat{y}_i$ . De igual manera la SCE, se puede expresar en su forma desarrollada para obtener una idea intuitiva sobre la estimación de la forma funcional original:

$$SCE = \sum_{i=1}^n (y_i - (b_0 + b_1x_{1i} + \dots + b_kx_{ki}))^2$$

Por ejemplo: para obtener los resultados de la regresión para dos variables independientes mediante el MCO se procede de la siguiente manera:

$$\hat{y}_1 = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

La SCE resultado de la estimación de  $\hat{y}_1$  en el caso de dos variables independientes ( $b_1 x_{1i}, b_2 x_{2i}$ ) se puede expresar de la siguiente manera, tomando en cuenta el resultado de la relación entre las variables independientes y la variable independiente observada ( $y_i$ )

$$SCE = \sum_{i=1}^n [y_i - (b_0 + b_1 x_{1i} + \dots + b_2 x_{2i})]^2$$

El desarrollo extenso del MCO es resultado de la aplicación de cálculo diferencial donde se debe tener en cuenta un sistema de 3 ecuaciones lineales y 3 incógnitas, ( $b_0, b_1, b_2$ ), las expresiones resultantes son las siguientes:

$$= nb_0 + b_1 \sum_{i=1}^n x_{1i} + b_2 \sum_{i=1}^n x_{2i} = \sum_{i=1}^n y_i$$

$$b_0 \sum_{i=1}^n x_{1i} + b_1 \sum_{i=1}^n x_{1i}^2 + b_2 \sum_{i=1}^n x_{1i} x_{2i} = \sum_{i=1}^n x_{1i} y_i$$

$$b_0 \sum_{i=1}^n x_{2i} + b_1 \sum_{i=1}^n x_{1i} x_{2i} + b_2 \sum_{i=1}^n x_{2i}^2 = \sum_{i=1}^n x_{2i} y_i$$

Ejercicio en R: Utilizando los datos del ejemplo antes mencionado, el comando en R para estimar los coeficientes del modelo de regresión múltiple sería el siguiente:

```
> lm(ventas ~ tamaño + tráfico + rivales + renta + residentes, data=tiendas)
```

De esa forma, el modelo de regresión lineal múltiple estimado es el siguiente:

$$ventas = b_0 + b_1 \text{tamaño} + b_2 \text{tráfico} + b_3 \text{rivales} + b_4 \text{renta} + b_5 \text{residentes}$$

Para almacenar los datos del modelo, a fin de realizar las pruebas pertinentes más adelante, se asigna nombre a los resultados del mismo:

```
>resultado <- lm(ventas ~ tamaño + tráfico + rivales + renta +  
residentes,data=tiendas)
```

La interpretación de los resultados del sistema es la siguiente: en la primera ecuación la variable observada depende de los coeficientes ( $b_1, b_2$ ) asociados a las observaciones de las variables independientes ( $x_{1i}, x_{2i}$ ) y una constante ( $b_0$ ) asociada al número de observaciones ( $n$ ).

En la segunda ecuación, la relación entre la variable independiente y la primer variable dependiente ( $x_{1i}, y_i$ ) esta explicada por la constante asociada a ( $x_{1i}$ ), las observaciones de ( $x_{1i}$ ), elevadas al cuadrado asociadas a ( $b_1$ ) y el comportamiento entre las dos variables independientes ( $x_{1i}, x_{2i}$ ) asociadas a  $b_2$ .

En la tercera ecuación, la relación entre la variable independiente y la segunda variable dependiente ( $x_{2i}, y_i$ ) esta explicada por la constante asociada a ( $x_{2i}$ ), las observaciones de ( $x_{2i}$ ), elevadas al cuadrado asociadas a ( $b_2$ ) y el comportamiento entre las dos variables independientes ( $x_{1i}, x_{2i}$ ) asociadas a  $b_1$ .

En conclusión, de la representación de la regresión múltiple se infiere, que el coeficiente asociado a la variable explicativa correspondiente, es decir, en el caso de la primera variable independiente ( $x_{1i}, b_1$ ), esta explicada por la misma variable

al cuadrado, y en el caso del otro coeficiente ( $b_2$ ) esta explicado por la asociación entre las variables independientes. Lo que se espera, en la regresión es que los dos coeficientes asociados a cada variable independiente expliquen el comportamiento de la variable dependiente de forma significativa. Lo anterior es resultado, de minimizar los errores asociados a la estimación de la variable independiente en relación a la variable observada.

## 2.1 Estimación del MCO múltiple mediante notación matricial

La estimación de los coeficientes de las variables independientes mediante el MCO, en su notación matricial permite visualizar de forma simplificada las operaciones necesarias; esto permite intuir el proceso de estimación de los coeficientes:

$$\hat{y}_1 = b_0 + b_1 x_{1i} + b_2 x_{2i}$$

La notación matricial de la expresión anterior es la siguiente, tenemos, los vectores a estimar la variable independiente (Y) los coeficientes ( $\beta$ ) y los errores en la estimación (e):

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \hat{Y} = \begin{bmatrix} \hat{y}_1 \\ \hat{y}_1 \\ \vdots \\ \hat{y}_1 \end{bmatrix} \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_n \end{bmatrix} e = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix}$$

Las variables independientes ( $X$ ), se organizan matricialmente tomando en cuenta su dimensión expresada mediante  $k - filas$  por  $n - columnas$ , más la constante ( $b_0$ ) representada por una constante numérica igual a (1)

$$\begin{bmatrix} 1 & x_{11} & x_{21} & \dots & x_{k1} \\ 1 & x_{12} & x_{22} & \dots & x_{k2} \\ \vdots & & \ddots & & \vdots \\ 1 & x_{1n} & x_{2n} & \dots & x_{kn} \end{bmatrix}$$

La construcción de la expresión en su forma matricial reducida es la siguiente:

$$Y = X\beta + U$$

La estimación objetivo del modelo, busca obtener los coeficientes estimados del modelo en relación a las variables independientes, para explicar la variable dependiente ( $\hat{Y}$ ) y su notación es la siguiente:

$$\hat{Y} = X\hat{\beta}$$

Donde, la matriz de variables independientes ( $X$ ) está asociada al vector de coeficientes estimados ( $\hat{\beta}$ )

La diferencia entre el modelo estimado en su forma matricial y la variable observada nos permiten obtener los errores derivados de la estimación:

$$Y - \hat{Y} = e$$

Es decir,

$$e = y_i - \hat{y}_i = y_i - (\hat{\beta}_0 + \hat{\beta}_1 x_{1i} + \hat{\beta}_2 x_{2i} + \dots + \hat{\beta}_k x_{ki})$$

Al aplicar, el método de MCO, debemos minimizar la suma de los errores al cuadrado SEC:

$$SEC = \sum_{i=1}^n e_i^2$$

Al minimizar (s) respecto al vector de los coeficientes ( $\beta$ ) tenemos la siguiente notación matricial reducida:

$$\frac{\partial s}{\partial \beta} = -X^T Y - X^T Y + 2 (X^T X \beta)$$

$$\frac{\partial s}{\partial \beta} = -2X^T Y + 2 (X^T X \beta) = \vec{0}$$

Para obtener los coeficientes estimados despejamos  $\beta$

$$\hat{\beta} = (X^T X)^{-1} - X^T Y$$

Entonces ( $\hat{\beta}$ ), es igual a la matriz inversa resultante de la multiplicación entre la matriz transpuesta ( $X^T$ ) y la matriz ( $X$ ), menos la matriz ( $X^T$ ) multiplicada por el vector de ( $Y$ ). El coeficiente estimado ( $\hat{\beta}$ ) representa el efecto de un aumento en una unidad de la variable independiente sobre la respuesta de ( $Y$ ), cuando las otras variables independientes se mantienen constantes.

### 3. LAS PROPIEDADES DE LOS ERRORES

Los estimadores o coeficientes obtenidos tienen propiedades esenciales que permiten una inferencia estadística apropiada, se deduce que la sumatoria de los errores en una serie son igual a cero:

$$\sum_{i=1}^n e_i x_{ij} = 0. j = 1 \dots k$$

La covarianza entre los errores y las variables explicativas a medida que aumenta el número de observaciones es igual cero:

$$Cova = (e_i, x_{ij}) = 0$$

En el caso del sesgo, se define como la diferencia entre la media del estimador y el verdadero valor del parámetro a estimar. En econometría se utiliza la varianza residual de los errores, el cual es insesgado al estar entorno a la misma varianza.

En este caso tenemos:

$$s_r^2 = \frac{1}{n - (k + 1)} \sum_{i=1}^n e_i^2$$

Ejercicio en R: El comando para obtener el vector de residuales de la estimación en el ejemplo:

```
> residuales<- resultado$residuals
```

La interpretación de los fenómenos económicos mediante un modelo econométrico depende de la robustez de los resultados obtenidos en la estimación. La interpretación inicia con la verificación de la eficiencia de los resultados mediante la inferencia estadística. Cuando se realiza la inferencia en un modelo de regresión múltiple se debe verificar la estabilidad de los coeficientes y su poder explicativo del modelo.

La distribución de los coeficientes, al igual que en la regresión simple se distribuyen como una normal, es decir, la media es igual a cero y la desviación estándar es igual a uno.

$$\hat{\beta} \sim N(0,1)$$

Este comportamiento asegura que los coeficientes estimados sigan una trayectoria normal y no sigan un comportamiento errático que genere problemas en la estimación a medida que aumentan las observaciones.

El análisis de probabilidad sobre los coeficientes, para identificar la influencia de cada variable parte de la hipótesis planteada desde el diseño del modelo y su forma funcional. El contraste de hipótesis, se construye mediante una t de Student con k grados de libertad, la prueba muestra las siguientes posibilidades:

La hipótesis nula es

$$H_0: \beta_i = 0$$

La hipótesis alternativa es

$$H_a: \beta_i \neq 0$$

Al aplicar el contraste de hipótesis, cuando la probabilidad de cometer el error tipo I es elevada, es decir, rechazar la  $H_0$  cuando es verdadera y aceptar la  $H_a$  cuando esta última es falsa, entonces, lo correcto es aceptar  $H_0$ ; de ahí se puede inferir que la variable independiente  $X_i$  asociada a su coeficiente tiene un efecto nulo, es decir, no influye sobre la variable dependiente.

El diseño de la prueba es el siguiente, la distribución del valor de los coeficientes cuando se acepta la  $H_0$  se distribuyen de la siguiente forma: para  $n > 30$  observaciones la distribución  $t_{n-k-1}$ , bajo una probabilidad del 95% se encuentra en el intervalo  $[-2,2]$  y entonces se acepta la hipótesis nula. Si  $t > 2$ , se rechaza la hipótesis nula y se puede inferir estadísticamente que las variables independientes influyen en la variable dependiente, es decir se acepta la hipótesis alternativa. El contraste de hipótesis nos señala que la probabilidad de cometer el error tipo I es nulo, por lo tanto, podemos rechazar la hipótesis nula y aceptamos la hipótesis alternativa:

$$H_a: \beta_i \neq 0$$

El criterio del intervalo de confianza está diseñado de la siguiente forma:

$$P\left(\widehat{\beta}_i - t_{\frac{\alpha}{2}}SE(\widehat{\beta}_i) \leq \beta_i \geq \widehat{\beta}_i + t_{\frac{\alpha}{2}}SE(\widehat{\beta}_i)\right) = 1 - \alpha$$

El criterio muestra la probabilidad de que el verdadero  $\beta_i$  se encuentra en el intervalo entre el coeficiente estimado ( $\hat{\beta}_i$ ) y 2 desviaciones estándar (SE) a la derecha y a la izquierda. Cuando tenemos un intervalo de confianza de  $\alpha = .05$ , se plantea que hay un 95% de confianza de que el valor verdadero para cada coeficiente se encuentre dentro del área de aceptación.

Ejercicio R: Retomando nuestro ejemplo, el comando necesario para obtener los estadísticos tales como la probabilidad de los coeficientes del modelo, es el siguiente:

```
>summary(resultado)
```

La matriz de varianzas-covarianzas de los coeficientes en su forma matricial reducida es la siguiente:

$$\text{COV}(\hat{\beta}) = \sigma^2(X^T X)^{-1}$$

De la función anterior es necesaria la estimación de la varianza ( $\sigma^2$ ), en la estimación del modelo, se espera que la varianza de los residuos sea el valor verdadero de la varianza de los estimadores es decir, que la varianza de las variables incluidas en el modelo explique los errores de la estimación:

$$E(\hat{S}_e^2) = \sigma^2$$

Este resultado, nos permite establecer que la elección de las variables en la estimación del modelo, es la especificación correcta, ya que, explica las desviaciones de la variable dependiente respecto a la estimada.

Ejercicio R: En nuestro ejemplo, la matriz de varianzas-covarianzas se obtiene de la siguiente manera:

```
>vcov(resultado)
```

Una forma de medir el poder explicativo del modelo es el contraste F, muestra si las variables explicativas en conjunto explican las variaciones de la variable independiente. Se ha demostrado que los coeficientes  $\beta_1 = \beta_2 = \dots = \beta_k = 0$  y además, siguen una distribución F dado la siguiente forma:

$$\frac{\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y}_i)^2}{k}}{\frac{\sum_{i=1}^n e_i^2}{n-k-1}} \sim F_{k,n-k-1}$$

El resultado muestra la proporción en que la varianza de los coeficientes explica la variación en los errores; cuando se acepta la hipótesis nula se debe a dos factores: 1) las variables no influyen en la variable independiente, 2) existe dependencia no lineal entre la variable explicada y algún regresor. Cuando se rechaza la hipótesis nula en el contraste del test F, muestra que la variable dependiente esta explicada por alguna de las variables independientes. Para conocer de forma específica las variables con poder explicativo relativo a las otras variables es necesario revisar los contrastes individuales mediante la t de student.

En la aplicación de los contrastes de F se presentan los siguientes casos:

- 1) Cuando el contraste F es significativo y todos los coeficientes individuales de acuerdo al contraste de la t de student también son significativos, en este caso todas las variables independientes son significativas para explicar el comportamiento de la variable dependiente.
- 2) Si el contraste F es significativo y sólo algunos de los coeficientes individuales son significativos de acuerdo al contraste de la t de student, las variables no significativas deben ser eliminadas del modelo. Otra solución, es realizar una transformación y estimar nuevamente para verificar si la relación entre las variables no es lineal.
- 3) Cuando el contraste de F es significativo y por el otro lado cuando ninguno de los coeficientes asociados a las variables es significativo de acuerdo al contraste t, entonces podría estar presente un problema de multicolinealidad. Esta última es resultado de una correlación alta entre las variables independientes; entonces, la especificación del modelo requiere una elección eficiente de las variables.

En la tabla ANOVA, podemos evaluar los resultados mediante el Test F:

$$\frac{\hat{S}_e^2}{\hat{S}_r^2}$$

Ejercicio en R: El comando para obtener la tabla ANOVA del ejemplo que se ha desarrollado es el siguiente:

```
>anova(resultado)
```

El Test F muestra la proporción en que la varianza de los errores determina el poder explicativo del modelo. La notación matricial de la prueba, muestra que la diagonal de la matriz conocida, arroja los valores de la varianza ( $\sigma^2$ ):

$$D(X^T X)^{-1} \rightarrow \begin{bmatrix} d_{00} & & & \\ & d_{11} & & \\ & & d_{ii} & \\ & & & d_{kk} \end{bmatrix}$$

De esta forma, la distribución de los coeficientes estimados es la siguiente:

$$\hat{\beta}_i \sim N(\beta_i, \sigma \sqrt{d_{ii}})$$

En donde, la desviación de los coeficientes tienen una distribución normal,

$$\frac{\hat{\beta}_i - \beta_i}{\sigma \sqrt{d_{ii}}} \rightarrow N(0,1)$$

La desviación entre el coeficiente estimado ( $\hat{\beta}_i$ ) y el coeficiente ( $\beta_i$ ) en proporción a la interacción en diagonal conocida se comportan como una normal.

Ejercicio en R: La prueba F en R se realiza con el siguiente comando:

```
>var.test(resultado)
```

#### 4. PRUEBAS DE DIAGNÓSTICO

La información relevante en los modelos de regresión múltiple, está contenida en las variables seleccionadas. Los modelos operan bajo el supuesto de que el modelo contiene todas las variables relevantes para explicar el modelo. En este

sentido la realización de pruebas de diagnóstico sobre la selección eficiente de las variables incluidas en el modelo es necesaria. La omisión de variables relevantes en el modelo, es un problema relevante en la especificación del modelo y en este sentido se pueden generar problemas de multicolinealidad.

Al iniciar el capítulo se planteó que el primer paso es la especificación del modelo, la selección de las variables para la conformación del modelo, se realiza con los referentes que ofrece la teoría económica. Como se ha señalado, las variables referentes en estos modelos no especifican como podrían conformar un modelo econométrico. El primer paso, es revisar la teoría para contrastar las variables relevantes que explican el objeto de estudio desde esa perspectiva. El siguiente paso es realizar una prueba de omisión de variables, supongamos que la teoría señala que la regresión correcta incluye dos variables

$$Y = X_1\beta_1 + X_2\beta_2 + U$$

Finalmente tras un proceso de elección el modelo estimado es:

$$Y = X_1\beta_1 + U$$

El siguiente paso es plantear la hipótesis nula de la omisión de variables:

$$H_0: \beta_2 = 0$$

Posteriormente se realiza un prueba de contraste F para estimar el poder explicativo del modelo, en un caso se estimará la prueba al modelo estimado y una prueba para el modelo que incluye la variable omitida. El rechazo de la hipótesis nula en este caso mostrará que fue omitida una variable relevante.

De igual manera, cuando se incluyen variables irrelevantes en el modelo es necesario realizar pruebas para la especificación del modelo. De hecho, cuando se aplica una metodología donde se parte de la especificación más general se realizan estas pruebas para llegar un modelo más específico.

La prueba de inclusión de variables irrelevantes consiste en probar en la hipótesis

$$H_0: \beta_2 = 0$$

## 5. UN EJEMPLO FINAL EN R

Para ejemplificar un modelo de regresión múltiple retomaremos el modelo de ventas que se utilizó en el primer apartado de este capítulo, pero haremos algunas simplificaciones. Supondremos que las ventas reales se comportan como una función de demanda y que por consiguiente dependerán de los precios de las mercancías y del ingreso por persona de la población.

En el archivo ventas.txt se presentan datos logarítmicos del índice de ventas reales al menudeo (Lventa), el índice de precios al consumidor (Lpr) y el ingreso per cápita (Lingr) aproximado por un índice de remuneraciones reales por persona ocupada.

En RComander utilizamos en el menú principal STATISTICS/Fit models/Linear regression. En la ventana que se abre se selecciona Lventa como variable dependiente y a Lpr y Lingr como variables explicativas.

Los resultados de la regresión se muestran en el siguiente recuadro:

```
lm(formula = LVENTA ~ LINGR + LPR, data = Dataset)
```

Residuals:

	Min	1Q	Median	3Q	Max
	-0.068921	-0.022129	-0.000394	0.025324	0.073677

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	2.41595	0.26438	9.138	7.26e-14 ***
LINGR	0.67508	0.04718	14.307	< 2e-16 ***
LPR	-0.18473	0.04719	-3.915	0.000196 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03236 on 76 degrees of freedom

Multiple R-squared: 0.7293, Adjusted R-squared: 0.7221

F-statistic: 102.4 on 2 and 76 DF, p-value: < 2.2e-16

En los resultados se observa que los coeficientes del modelo son estadísticamente significativos, en todos los casos las pruebas t brindan probabilidades que permiten rechazar la hipótesis nula de que dichos coeficientes son nulos a cualquier nivel de significancia estadística; en el recuadro los niveles de significancia están marcados con asteriscos.

Los valores de los coeficientes se pueden interpretar directamente como elasticidades en la medida en que el modelo se especificó logarítmicamente. Los signos son los esperados y se muestra que el incremento del 10% en el nivel de ingresos reales da lugar a un aumento del 6.75% en las ventas, mientras que el incremento de un 10% en los precios da lugar a una reducción del 1.8% en las ventas.

En la parte inferior del recuadro se muestran los resultados para el coeficiente de determinación y su variante ajustada. En ambos casos se muestra que la variabilidad total en las ventas se explica en más del 70% por la variación de las

variables del modelo, esto implica que hay un ajuste lineal elevado entre las variables.

Finalmente, en el último renglón del recuadro se muestran los resultados para el estadístico F, que tiene un valor elevado de 102 y un p-valor prácticamente de cero, lo cual permite rechazar la hipótesis nula de que las variables del modelo son simultáneamente nulas.

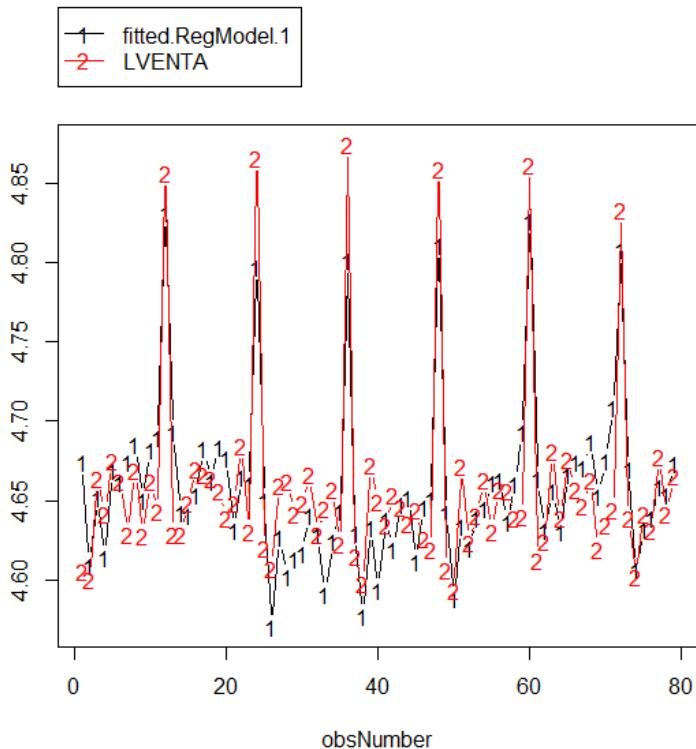
En el menú principal del RCommander al seleccionar MODELS/Hypothesis tests/ANOVA se obtiene la tabla de análisis de varianza, su resultado se muestra en el recuadro siguiente:

```
> Anova(RegModel.1, type="II")
Anova Table (Type II tests)
Response: LVENTA
          Sum Sq   Df   F value    Pr(>F)
LINGR      0.214297   1   204.700 < 2.2e-16 ***
LPR        0.016045   1    15.327  0.0001957 ***
Residuals  0.079563  76
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Para generar los valores estimados de las ventas por la ecuación de regresión, en el menú principal se selecciona MODELS/Add observation statistics to data y en la ventana que se abre se activan las opciones Fitted values que permite obtener los valores estimados y Residuals que incorpora a la tabla de datos los residuales del modelo.

En el menú de graficas del RComander se pueden visualizar los resultados para los valores estimados de las ventas y las ventas observadas. En el menú principal se selecciona GRAPHS/Line graph y en la ventana contextual que se abre se

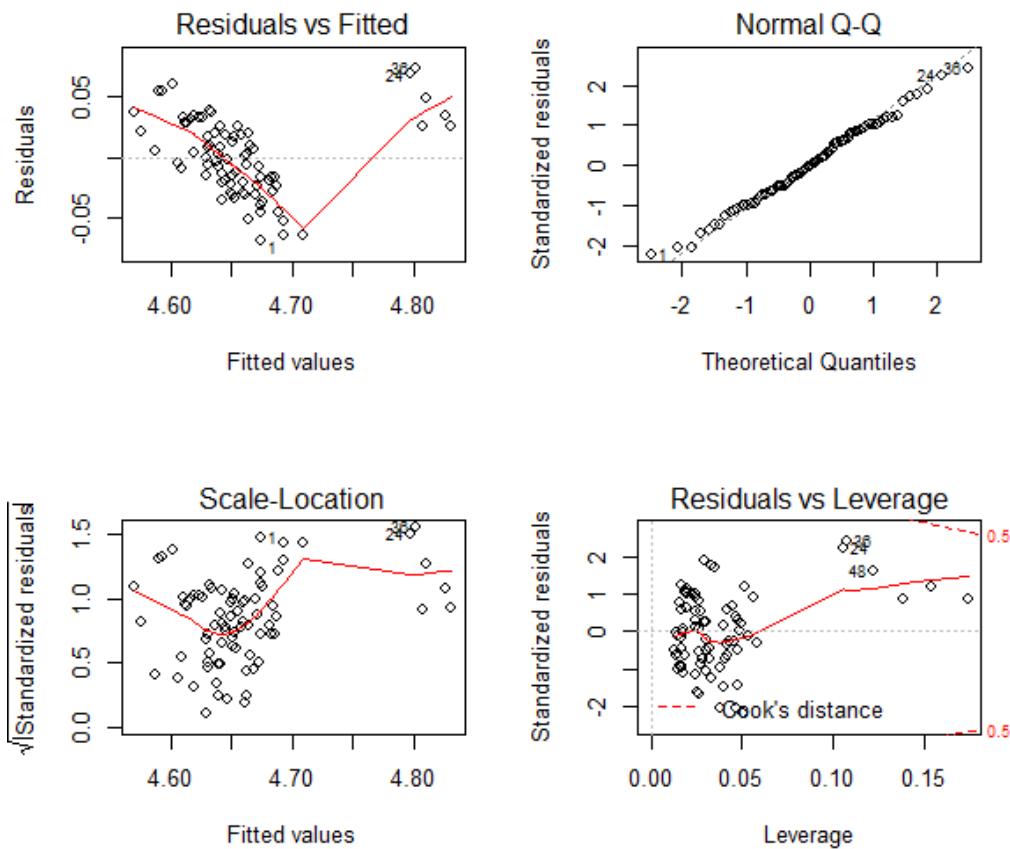
seleccionan los valores para el eje de las X y los valores para el eje de las Y. En el primer caso se seleccionan obsNumber para el eje X, en el eje Y se debe seleccionar la variable dependiente LVENTA y sus valores estimados, que por default el paquete ha guardado en la tabla de datos con el nombre fitted.RegModel.1. La gráfica resultante se muestra a continuación, en ella se aprecia que los valores estimados son relativamente muy próximos a los valores observados de la variable.



Finalmente, en el menú principal MODELS/Graphs/Base diagnostic plots se obtiene un juego de cuatro gráficas para evaluar los residuales de la regresión. En la primera se comparan los residuales del modelo con los valores estimados de la

regresión y en las otras tres se comparan los residuales estandarizados de la regresión.

$\text{lm}(\text{LVENTA} \sim \text{LINGR} + \text{LPR})$



## **REFERENCIAS**

- Crawley, J. Michael (2009), *The R book*, ed. Wiley, Inglaterra.
- Maddala, G. S. (1996), *Introducción a la econometría*. Ed. Prentice Hall, México.
- Quintana Romero, Luis y Miguel Ángel Mendoza (2008), *Econometría básica*, Plaza y Valdés.
- Venables, W. N. y D. M. Smith (2013), *An introduction to R*, ed. R Core Team.

## **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

**ventas.txt**

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap3

Práctica\_Cap3

VideoPráctica\_Cap3

VideoTeoría\_Cap3

# CAPITULO 4: ERROR DE ESPECIFICACIÓN

Lucía A. Ruiz Galindo

## 1. INTRODUCCIÓN

En la elaboración de un modelo econométrico es muy importante la evaluación económica y la econométrica del modelo estimado. En ambas se revisa si la información empírica incorporada al modelo, es decir, con la que éste se estimó, da evidencia a favor o en contra por un lado, de la teoría económica que lo sustentó y por el otro, de los supuestos tanto los que se hacen en su parte determinista como los que se plantean sobre el término estocástico.

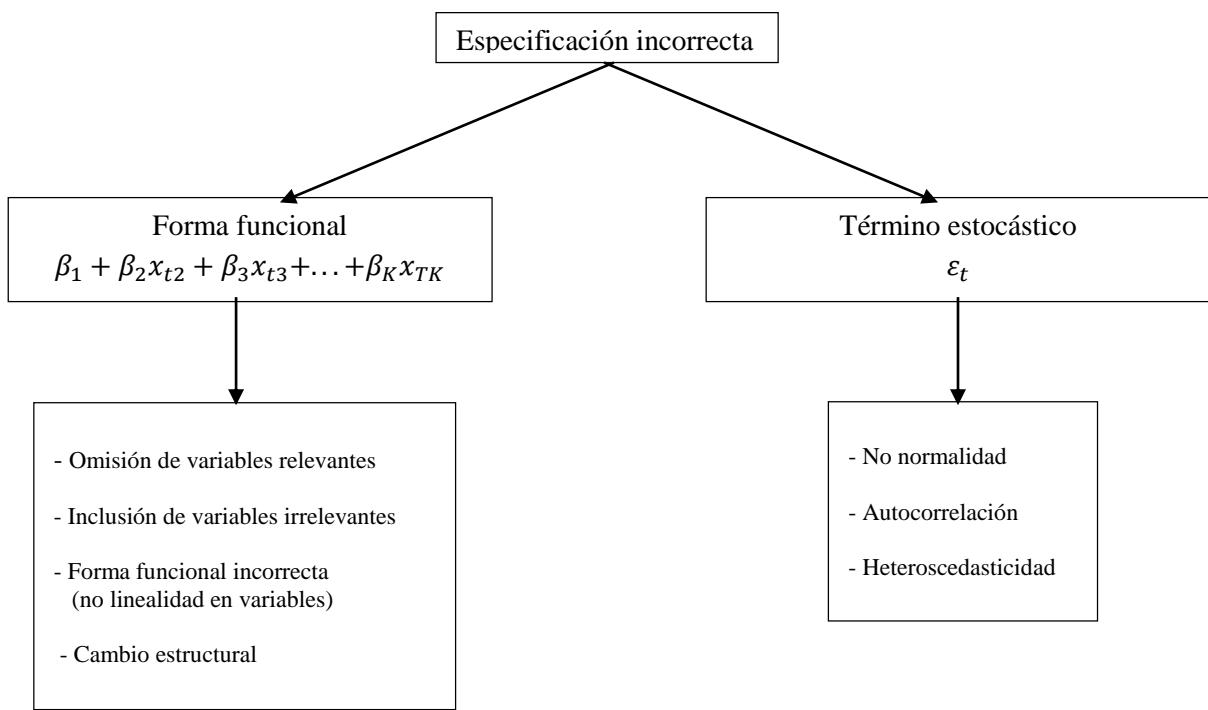
A grandes rasgos, en la evaluación económica se revisa que los signos y las magnitudes de los parámetros estimados sean los propuestos por la teoría económica, mientras que la evaluación econométrica consiste de una variedad de pruebas estadísticas que permiten averiguar si se satisfacen todos los supuestos del modelo. Cuando la evaluación económica es exitosa, pero existen discrepancias entre el resultado de las pruebas y los supuestos del modelo, habrá indicios de que el modelo no está especificado correctamente, hay errores en su especificación.

La especificación incorrecta del modelo puede deberse a una formulación no adecuada de la forma funcional o bien, a que se violan los supuestos del error aleatorio o incluso a la información empírica que se incorpora al modelo para su estimación. La Figura 1 muestra cómo están constituidas las dos primeras fuentes de especificación incorrecta, para cada una de sus componentes existen pruebas estadísticas que permiten decidir cuáles de los supuestos del modelo de regresión no se satisfacen dada la información empírica que se utiliza en su estimación.

En este Capítulo se estudian la especificación incorrecta del modelo ocasionada por un planteamiento no apropiado de la forma funcional (parte

determinista), esta situación generalmente se debe a que se han incluido variables irrelevantes (sobreparametrización), omitido variables relevantes (subparametrización) o bien, a que la forma funcional no es la correcta en lo que respecta a la manera en que se incorporan las variables independientes y en la literatura a esto se le conoce como errores de especificación. Cabe mencionar que los errores de especificación en la forma funcional, también se originan cuando existe cambio estructural en los parámetros, pero este tema no es objeto de estudio de este Capítulo.

Figura 1. Errores de especificación



Por su parte, la existencia de multicolinealidad y la de correlación entre las variables independientes y el término estocástico, también son fuente de especificaciones erróneas del modelo de regresión, pero ellas son debidas a la selección de la información empírica de las variables del modelo.

La especificación del modelo incluyendo sus supuestos, conducen a que los estimadores de los parámetros satisfagan propiedades estadísticas deseables, como son el insesgamiento, la eficiencia y la consistencia. Aquí se estudiará las

consecuencias que tiene sobre las propiedades de los estimadores, la sub y sobreparametrización, además de formular y llevar a cabo en R, la prueba RESET, útil para saber entre otras cosas, si la especificación lineal en las variables es correcta o no.

Este Capítulo en su segunda Sección presenta una exposición sucinta del modelo de regresión lineal, sus supuestos y una breve explicación de la forma en que se incurre en una especificación incorrecta del mismo, en la tercera Sección se estudian las implicaciones que tiene sobre las propiedades de los estimadores de los parámetros, la sobreparametrización y subparametrización del modelo, la cuarta Sección se formula la prueba RESET para analizar si la forma funcional del modelo es correcta o no, en la siguiente Sección se explica la manera en que esa prueba se lleva a cabo en R y se muestran algunos ejemplos de su implementación, y en la sexta y última, se presentan algunas conclusiones.

## **2. ESPECIFICACIÓN Y SUPUESTOS DEL MODELO GENERAL DE REGRESIÓN LINEAL**

En el desarrollo de este Capítulo se considera un modelo de regresión lineal en el que la variable dependiente  $y_t$  es explicada por  $K-1$  variables independientes, esto es,

$$y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + \varepsilon_t \quad (1)$$

donde  $\beta_1, \dots, \beta_K$  son los parámetros del modelo, las  $x_{tk}$ 's son las variables independientes,  $k = 2, \dots, K$ ,  $\varepsilon_t$  es el término o error estocástico y  $t, t = 1, \dots, T$  es un índice que indica el número de la observación y  $T$  es el total de observaciones.<sup>3</sup>

El modelo en (1) se puede formular de manera matricial como sigue

---

<sup>3</sup> En todo lo que sigue, sin pérdida de generalidad, se pensara que las variables están en series de tiempo y por tanto,  $t$  indica un periodo y hay observaciones para  $T$ . Es importante señalar que todo lo que se desarrolla en este Capítulo es válido también para cortes transversales, en cuyo caso,  $t$  representará un individuo.

$$y = X\beta + \varepsilon, \quad (2)$$

donde  $y = (y_1, y_2, \dots, y_T)'$ ,

$$X = \begin{pmatrix} 1 & x_{12} & x_{13} & \dots & x_{1K} \\ 1 & x_{22} & x_{23} & \dots & x_{2K} \\ 1 & x_{32} & x_{33} & \dots & x_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T2} & x_{T3} & \dots & x_{TK} \end{pmatrix},$$

$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T)'$  y  $\beta = (\beta_1, \beta_2, \dots, \beta_K)'$ . Observe que el vector  $y$  está constituido por las  $T$  observaciones de la variables dependiente, la matriz  $X$  de dimensión  $T \times K$ , por las variables independientes, el vector  $\beta$  de dimensión  $K$ , por los parámetros del modelo, y  $\varepsilon$  por los  $T$  términos estocásticos, uno por cada periodo.

El modelo está completamente especificado cuando se plantean sus supuestos. La forma funcional  $\beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK}$ , debe ser lineal en los parámetros, las variables  $x_{tk}$ 's,  $k = 2, \dots, K$ , son las únicas que explican a  $y_t$ , ellas son linealmente independientes y por ello, la matriz  $X$  es de rango completo, y además, los parámetros no cambian en el periodo de estudio, esto es, hay permanencia estructural.<sup>4</sup>

Por su parte, el término estocástico  $\varepsilon_t$ ,  $t = 1, \dots, T$ , tiene media cero, es homoscedástico y no autocorrelacionado y se distribuye de manera normal, todos los supuestos de los errores aleatorios se pueden resumir diciendo que ellos son elegidos de manera no correlacionada de una distribución normal con media y varianza constante o equivalentemente,

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 I),$$

donde  $\mathbf{0}$  e  $I$ , son un vector de ceros de dimensión  $T$  y la matriz identidad de  $T \times T$ , respectivamente, y  $\sigma^2$  es la varianza del término aleatorio, es decir,  $V(\varepsilon_t) = \sigma^2$ ,  $t = 1, \dots, T$ .

---

<sup>4</sup> Los momentos poblacionales están condicionados a la información disponible de las variables en el modelo.

Una vez formulado el modelo de regresión, se procede a estimarlo usando datos de las variables dependiente e independientes, de forma que los estimadores de los parámetros dependen tanto de la especificación del modelo como de la información empírica que se incorpora a él. De manera que, errores en el modelo o incluso en los datos, conducen errores de especificación.

La especificación correcta del modelo conduce a que los estimadores de las  $\beta_k$ 's,  $k = 1, \dots, K$ , las  $\hat{\beta}_K$ , son los mejores estimadores lineales e insesgados, MELI o BLUE por sus siglas en inglés (*Best Linear Unbiased Estimator*), es decir, dentro de los lineales e insesgados son los de mínima varianza, además de que son consistentes. Por su parte, el estimador mínimo cuadrático de  $\sigma^2$  es insesgado, pero su varianza es mayor que la correspondiente al estimador máximo verosímil y éste a pesar de ser más eficiente, es sesgado.

A continuación, en la siguiente Sección, se estudia si estas propiedades prevalecen cuando en el modelo se excluyen variables importantes o cuando se incluyen variables irrelevantes.

### **3. SOBREPARAMETRIZACIÓN Y SUBPARAMETRIZACIÓN, CONSECUENCIAS SOBRE LAS PROPIEDADES DE LOS ESTIMADORES**

Considérense los siguientes modelos

$$\text{M1: } y = X_1\beta^1 + X_2\beta^2 + \varepsilon, \quad \varepsilon \sim N(\mathbf{0}, \sigma^2 I)$$

y

$$\text{M2: } y = X\beta^1 + u, \quad u \sim N(\mathbf{0}, \sigma_1^2 I),$$

donde  $X = (X_1 : X_2)$ ,  $\beta = \begin{pmatrix} \beta^1 \\ \vdots \\ \beta^2 \end{pmatrix}$ ,  $u = X_2\beta^2 + \varepsilon$ ,  $X_1$  tiene las primeras  $K_1$  variables de la matriz  $X$ ,  $X_2$  tiene las  $K_2 = K - K_1$  restantes ( $K = K_1 + K_2$ ), y el vector de parámetros  $\beta$  se plantea de acuerdo a esa división de las variables independientes en  $X$ .

Dados esos modelos, si el correcto o verdadero es M2 y se estima M1, entonces se están incorporando variables irrelevantes para la determinación de  $y$ , esto es, se está sobreparametrizando. Si por el contrario, M1 es el correcto y se estima M2, se están omitiendo en el modelo variables importantes, que pasan a formar parte del término estocástico, en este caso se está subparametrizando. Cada una de esas situaciones tienen consecuencias sobre las propiedades de los estimadores mismos que se plantearán a continuación.<sup>5</sup>

Al sobreparametrizar un modelo se están incluyendo variables que no son importantes en la determinación de  $y$ , de manera que el modelo adecuado es M2, pero el que se estima M1. Observe que en este caso M2 se puede obtener de M1 haciendo  $\beta^2 = 0$  y  $\sigma^2 = \sigma_1^2$  y por ello, algunos autores no lo consideran un error de especificación o una forma incorrecta de especificación, pues solo no incorpora las restricciones mencionadas sobre los parámetros (Davidson y MacKinnon (2004), y Greene (2007)).

En esta situación los estimadores tanto de los parámetros  $\beta^1$  y  $\beta^2$ , como el de  $\sigma^2$  son insesgados y consistentes, y esta propiedad se satisface incluso cuando se imponen la restricciones  $\beta^2 = 0$  y  $\sigma^2 = \sigma_1^2$ . Sin embargo, debe señalarse que inclusión de variables irrelevantes aumenta la varianza de los estimadores de las betas, de manera que ya no serán eficientes de manera relativa.

Por su parte, al subparametrizar un modelo se están omitiendo variables que son importantes en la determinación de la variable dependiente  $y$ . Si se supone que se dejan fuera  $K_2$  variables, esto es, las variables en  $X_2$ , entonces el modelo verdadero es M1, pero se estima M2. En este caso se debe observar en primer lugar, que las variables excluidas se encuentran dentro del término estocástico, por ello su varianza no será estimada de manera correcta y en consecuencia, los intervalos de confianza y las pruebas de hipótesis conducirán a conclusiones erróneas, pues dependen de ese estimador, que además es sesgado. En segundo lugar el estimador del vector  $\beta^1$  en el modelo M2 denotado por  $\tilde{\beta}^1$ , es segado y

---

<sup>5</sup> Un tratamiento riguroso de estos temas se puede estudiar en Kmenta (1997), Jhonston y Dinardo (1997), Davidson y MacKinnon (2004), y Greene (2007), por citar algunos.

eficiente, es decir, su varianza es menor a la correspondiente a  $\hat{\beta}^1$ , que es el estimador de  $\beta^1$  en M1 y por tanto,  $\tilde{\beta}^1$  es más preciso que  $\hat{\beta}^1$ , pero no es insesgado (Davidson y MacKinnon (2004), y Greene (2007)).

#### 4. PRUEBA RESET

Conocer los errores de especificación y en caso de incurrir en ellos, saber sus consecuencias, es importante en la elaboración de un modelo econométrico, igual relevancia tiene el averiguar si ellos se han cometido o no. En esta Sección se estudia la prueba de especificación de Ramsey, denominada RESET, por sus siglas en inglés *Regression Equation Specification Error Test*, debida a Ramsey (1969), que sirve para detectar errores de especificación ocasionados por la omisión de variables independientes, por la posible existencia de correlación entre las variables en  $X$  y  $\varepsilon$  o bien, porque la forma funcional de las variables independientes no es la apropiada.

Así pues, la prueba RESET se usa para analizar si el modelo está bien especificado o no, de manera que las hipótesis a probar son

$$H_0: \text{Forma funcional correcta} \quad \text{vs} \quad H_1: \text{Forma funcional incorrecta.}$$

Esta prueba se realiza una vez que se ha estimado el modelo planteado en (1) y que se ha calculado su ajuste, dado por

$$\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_{t2} + \dots + \hat{\beta}_K x_{tK} \quad (3)$$

y consiste en agregar al modelo inicial, potencias de sus valores ajustados y analizar la significancia estadística conjunta de los parámetros asociados a las potencias de la variable ajustada. De esta manera, el modelo que se debe estimar para efectuar la prueba RESET es

$$y_t = \beta_1 + \sum_{k=2}^K \beta_k x_{tk} + \sum_{i=2}^{m+1} \alpha_i \hat{y}_t^i + v_t, {}^6 \quad (4)$$

---

<sup>6</sup> $v_t$  denota el error estocástico de este modelo y por tanto, tiene los mismos supuestos del modelo de regresión en (1).

en el que se han incorporado  $m$  potencias de la variable ajustada  $\hat{y}_t$ .

Observe que bajo  $H_0$ , los parámetros  $\alpha_i = 0$ , para toda  $i = 2, \dots, m + 1$ , y bajo  $H_1$ , al menos uno de esos parámetros es diferente de cero, en cuyo caso la especificación del modelo no es correcta. Con estas consideraciones, la prueba se puede plantear como

$$\begin{aligned} H_0: \alpha_i &= 0, \quad \forall i = 2, \dots, m + 1 \\ &\text{vs} \\ H_1: \alpha_i &\neq 0, \text{ para al menos una } i = 2, \dots, m + 1. \end{aligned}$$

Bajo  $H_0$  el estadístico de prueba se distribuye como una  $F_{(m, T-K-m)}$ .<sup>7</sup> El número  $m$ , que también representa el número de restricciones lineales bajo la hipótesis nula, se puede determinar usando los criterios de información de Akaike, Schwarz o Hannan-Quinn, utilizados comúnmente para seleccionar entre modelos alternativos, en los que la variable dependiente debe ser la misma.

## 5. PRUEBA RESET EN R

La prueba RESET en R, requiere del paquete lmtest y se efectúa una vez que el modelo ha sido estimado. Mediante la instrucción

```
> library(lmtest)
```

se carga el paquete lmtest, que dicho sea de paso, contiene varias pruebas que son importantes en la evaluación econométrica de un modelo de regresión. En seguida y ya que se dispone de los datos, se estima el modelo y hasta entonces, se hace la prueba RESET introduciendo

```
> resettest(vdep)
```

en donde el argumento vdep es el nombre del objeto donde se guarda el resultado de la estimación. Es importante indicar que esta instrucción introduce por *default* la segunda y tercera potencia de la variable ajustada  $\hat{y}_t$ , de manera que el modelo en la que se basa la prueba RESET es

---

<sup>7</sup> En el Capítulo 5 se explican de manera sucinta, las pruebas de significancia conjunta.

$$y_t = \beta_1 + \sum_{k=2}^K \beta_k x_{tk} + \alpha_2 \hat{y}_t^2 + \alpha_3 \hat{y}_t^3 + v_t.$$

Si se requieren potencias superiores a 3, se introduce la instrucción

```
> resettest(vdep,power=2:m)
```

y si sólo se desea introducir la segunda potencia, se escribe

```
> resettest(vdep,power=2:2)
```

El resultado de la prueba presenta el nombre del objeto en `data`, el estadístico de prueba en `RESET`, los grados de libertad del numerador ( $m$ ), en `df1` y los del denominador ( $T-K-m$ ), en `df2` y el mínimo nivel de significancia al que se rechaza la hipótesis nula, en `p-value`.

*Ejemplo 1.*

La información anual de 1953 a 2004 contenida en el archivo `Gasolina.txt` es usada para estimar un modelo para la demanda de gasolina en USA (Greene, 2003). Se plantea una regresión log-log, en la que se modela la demanda per cápita en función del ingreso per-cápita, del índice de precios de la gasolina y el de los autos nuevos. Estimado el modelo se analiza si la forma funcional es correcta mediante dos pruebas RESET, la primera incorpora de la segunda a la cuarta potencia del ajuste y la segunda, solo la segunda potencia.<sup>8</sup>

```
> library(lmtest)
```

<sup>8</sup> Las variables del archivo son

Año: 1953-2004,  
G: Gasto total en gasolina,  
Pobl: Población  
Pg: Índice de precio de la gasolina,  
Y: Ingreso disponible per-cápita,  
Pan: Índice de precios de los autos nuevos,  
Pau: Índice de precios de los autos usados,  
Ptp: Índice de precios del transporte público,  
Pd: Índice de precios agregado del consumo de bienes durables,  
Pnd: Índice de precios agregado del consumo de bienes no durables,  
Ps: Índice de precios agregado para el consumo de servicios.

Fuente: <http://people.stern.nyu.edu/wgreene/Text/econometricanalysis.htm>

```
> Gasolina <- read.csv("Gasolina.txt")
> View(Gasolina)
> attach(Gasolina)
> cons<-lm(log(G/Pobl)~log(Y)+log(Pg)+log(Pan))
> resettest(cons,power=2:4)
```

#### RESET test

```
data: cons
RESET = 34.05, df1 = 3, df2 = 38, p-value = 7.347e-11
```

```
> resettest(cons,power=2:2)
```

#### RESET test

```
data: cons
RESET = 90.541, df1 = 1, df2 = 40, p-value = 7.933e-12
```

Los resultados de ambas pruebas indican que la forma funcional no es correcta ya que en ambas el *p-value* es menor que cualquiera de los niveles de significancia, comúnmente utilizados, de manera que el modelo debe ser reespecificado.

#### Ejemplo 2

En este ejemplo se presenta una versión del modelo estático para la elasticidad de sustitución Armington para México.<sup>9</sup> La variable explicada en el modelo es la demanda relativa (DRel), que resulta del cociente entre las importaciones totales y la demanda doméstica (diferencia entre el valor bruto de la producción y las exportaciones, ambas a precios de mercado) y la variable explicativa es el precio relativo (PRel), que se obtiene de dividir el índice de precios de la demanda relativa entre el correspondiente a las importaciones.<sup>10</sup>

Una vez estimado el modelo se prueba si la forma funcional es la correcta mediante la prueba de Ramsey, RESET. Ella se realiza considerando primero la segunda y tercera potencia y en seguida se efectúa solo para la segunda potencia, tal y como se muestra a continuación.

```
> library(lmtest)
```

---

<sup>9</sup> Un análisis detallado de este modelo desde sus microfundamentos hasta la especificación final de un modelo dinámico es presentado en Casares, Ruiz-Galindo y Sobarzo (por publicarse).

<sup>10</sup> La estimación incorpora información trimestral del INEGI para el periodo que comprende del primer trimestre de 1993 al primero del 2013, a precios constantes del 2008.

```

Loading required package: zoo

Attaching package: 'zoo'

The following objects are masked from 'package:base':

  as.Date, as.Date.numeric

> Elast <- read.csv("C:/Users/Atzimba/Desktop/Elast.txt")
> View(Elast)
> attach(Elast)

> model<-lm(log(DRel)~log(PRel))
> resettest(model)

      RESET test

data: model
RESET = 0.32523, df1 = 2, df2 = 80, p-value = 0.7233

> resettest(model,power=2:2)

      RESET test

data: model
RESET = 0.25329, df1 = 1, df2 = 81, p-value = 0.6161

```

En ambas pruebas no se rechaza la hipótesis nula, el *p-value*> $\alpha$  ( $\alpha=1\%, 5\%$  o  $10\%$ ), y por tanto hay evidencia a favor de que la forma funcional es correcta. Debe observarse que en la primera prueba RESET de este ejemplo, no se introduce el comando *power* que indica las potencias que se desean incorporar de la variable ajustada y el modelo en (4) se estima con la segunda y tercera potencia, puesto que como ya se mencionó, esas potencias son las que se introducen por *default*.

## REFERENCIAS

Casares, E. R., L. A. Ruiz-Galindo y H. Sobarzo, (por publicarse). “Short and Long Run Armington Elasticities for the Mexican Economy” en A. Pinto y D. Zilberman (editors), Modeling, Dynamics, Optimization and Bioeconomics II, en la serie Springer Proceedings in Mathematics and Statistics.

- Davidson R. y J. G. MacKinnon, (2004). Ed. Oxford University Press, New York.
- Greene, W. H., (2007). Econometric Analysis. Ed. New York University, New York.
- Johnston, J. y J. Dinardo, (1997). *Econometrics Methods* Ed. McGraw-Hill, Singapur.
- Kmenta, J., (1997). *Elements of Econometrics*. Ed. University of Michigan Press
- Ramsey, J. B., (1969). "Tests for Specification Errors in Classical Linear Least Squares Regression Analysis, *Journal of the Royal Statistical Society, Series B*., vol. 31, 2, pp 350-371.

### **Referencias electrónicas**

- Datos (Greene, 2007),  
<http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm>
- INEGI (2013a), "Banco de Información Económica", <http://dgcnesyp.inegi.gob.mx>

### **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

Gasolina.txt

Elast.txt

### **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap4

Práctica\_Cap4

VideoPráctica\_Cap4

VideoTeoría\_Cap4

# CAPITULO 5: NORMALIDAD

Lucía A. Ruiz Galindo

## 1. INTRODUCCIÓN

En la especificación del modelo de regresión lineal se distinguen dos partes, una determinística y otra estocástica. Una vez estimado el modelo y habiendo aprobado la evaluación económica del mismo, se llevan a cabo pruebas de especificación correcta, es decir, pruebas mediante las cuales se verifican los supuestos del modelo, los de su parte determinista, la que corresponde a la combinación lineal de los parámetros, y los de la estocástica, la asociada al término aleatorio.

De manera más específica, las pruebas de especificación correcta consisten en estudiar si la información empírica incorporada en el modelo, la que se utiliza para su estimación, proporciona evidencia a favor o en contra de los supuestos tanto de la parte determinista del modelo como de la aleatoriedad o estocástica. En la primera, generalmente se estudia si las variables independientes son las únicas que explican a la dependiente, si hay permanencia estructural en los parámetros y si la forma funcional en que se han introducido las variables es correcta o no, entre otras, mientras que en el término estocástico, se analizan los supuestos Gauss-Markov, que establecen que los errores aleatorios tienen media cero, son independiente de las variables explicativas, homoscedásticos y no autocorrelacionados, y también se verifica que se satisfaga el supuesto de normalidad, todo ello se lleva a cabo usando pruebas de hipótesis estadísticas. En este contexto, es importante indicar que cuando se realiza la prueba de un supuesto particular, se asume que todos los demás se satisfacen (supuesto *ceteris paribus*).

El objetivo de este Capítulo es doble: estudiar la importancia e implicaciones del supuesto de normalidad en el modelo de regresión lineal y de manera específica en la inferencia estadística de sus parámetros, y presentar en R, aplicaciones de la prueba de Jarque-Bera (Jarque-Bera 1980, 1987), utilizada para detectar si los términos estocásticos en el modelo siguen o no una distribución normal.

En la segunda Sección de este Capítulo se hace una breve presentación del modelo de regresión lineal, en la tercera se desarrollan dos procedimientos de estimación: el de mínimos cuadrados ordinarios y el de máxima verosimilitud y se analizan brevemente las propiedades de los estimadores resultantes, a partir de ellos se estudia la importancia que tiene el supuesto de normalidad de los errores estocásticos en la inferencia estadística y de manera más precisa, en la formulación de intervalos de confianza y de pruebas de hipótesis para todos los parámetros del modelo de regresión lineal, en la cuarta Sección se formula la prueba de Jarque-Bera para analizar si los errores satisfacen el supuesto de normalidad, utilizando para ello los residuos como *proxys* de los errores o términos estocásticos, en la quinta Sección se presenta la forma en que se realiza esta prueba en R y se muestran algunas aplicaciones de la misma, en la sexta se exponen las causas e implicaciones que tendría el hecho de que el supuesto de normalidad no se satisfaga y además, se muestran posibles soluciones, finalmente, en la séptima Sección, se plantean algunas conclusiones.

## **2. MODELO GENERAL DE REGRESIÓN LINEAL**

### **2.1 Especificación del modelo**

Considere que la variable dependiente es explicada por  $K-1$  variables independientes, esto es,

$$y_t = \beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK} + \varepsilon_t \quad (1)$$

donde  $\beta_1, \dots, \beta_K$  son los parámetros del modelo,  $y_t$  es la variable dependiente, las  $x_{tk}$ 's,  $k = 2, \dots, K$ , son las variables independientes,  $\varepsilon_t$  es el término o error

estocástico,  $t$ ,  $t = 1, \dots, T$ , es un índice que indica el número de la observación y  $T$  es el total de observaciones.

El modelo está formulado en el momento o periodo  $t$ , por ello las variables y el término estocástico están indexados con ese subíndice; mientras que el subíndice  $k$  en las variables independientes o explicativas, indica el número de la variable en la ecuación de regresión. Por ejemplo,  $x_{t5}$  y  $x_{tK}$ , señalan la variable 5 y la  $K$ , ambas en el momento  $t$  mientras que  $x_{5k}$  y  $x_{100k}$ , indican la observación 5 y 100 de la variable  $k$ .<sup>11</sup>

En la especificación anterior se distinguen dos partes, la determinista o también conocida como forma funcional, dada por

$$\beta_1 + \beta_2 x_{t2} + \dots + \beta_K x_{tK}$$

y la estocástica, que no es más que el término o error aleatorio  $\varepsilon_t$ . En la parte determinista los parámetros deben de plantearse en forma lineal de manera que el modelo sea lineal en ellos; por su parte, las variables dependiente e independientes, aunque introducidas de manera lineal, pueden no serlo. Debe hacerse notar que de acuerdo a la especificación anterior, los parámetros no cambian al paso del tiempo, no tienen subíndice  $t$ , por ello se dice que hay permanencia estructural o que no hay cambio estructural.

El modelo en (1) se puede formular de manera matricial como sigue

$$y = X\beta + \varepsilon, \quad (2)$$

donde  $y = (y_1, y_2, \dots, y_T)'$ ,

$$X = \begin{pmatrix} 1 & x_{12} & x_{13} & \dots & x_{1K} \\ 1 & x_{22} & x_{23} & \dots & x_{2K} \\ 1 & x_{32} & x_{33} & \dots & x_{3K} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{T2} & x_{T3} & \dots & x_{TK} \end{pmatrix},$$

---

<sup>11</sup> Esta especificación y todo lo que sigue es válido cuando en lugar de variables en series de tiempo se introducen en corte transversal.

$\varepsilon = (\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T)'$  y  $\beta = (\beta_1, \beta_2, \dots, \beta_K)'$ . Observe que el vector  $y$  está constituido por las  $T$  observaciones de la variables dependiente, la matriz  $X$  de dimensión  $T \times K$ , por una columna de unos asociada al término independiente y las  $K-1$  columnas restantes corresponden a las observaciones de las variables independientes, el vector  $\beta$  de dimensión  $K$ , por los parámetros del modelo y  $\varepsilon$  por los  $T$  términos estocásticos, uno por cada periodo.

## 2.2 Supuestos de la forma funcional.

- S1. Linealidad en los parámetros.
- S2. Las  $K-1$  variables independientes son las únicas que explican a la dependiente.
- S3. El número de observaciones  $T$ , es mucho mayor que el de parámetros  $K$ .
- S4. Las variables explicativas son linealmente independientes de manera que ninguna es combinación lineal de otra o de otras y por tanto el rango de  $X$  es  $K$ .
- S5. Los parámetros no cambian en la muestra, es decir, hay permanencia estructural.

## 2.3 Supuestos Gauss-Markov

Los supuestos Gauss Markov son sobre el término estocástico.<sup>12</sup>

SGM1.  $E(\varepsilon_t) = 0, \forall t = 1, \dots, T$ .

SGM2.  $\{\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T\}$  y  $\{x_{1k}, x_{2k}, \dots, x_{Tk}\}$  son independientes  $\forall k = 2, \dots, K$ .

SGM3.  $V(\varepsilon_t) = \sigma^2, \forall t = 1, \dots, T$ .

SGM4.  $Cov(\varepsilon_t, \varepsilon_s) = 0, \forall t, s = 1, \dots, T, t \neq s$ .

SGM5.  $\varepsilon_t$  se distribuye Normal,  $\forall t = 1, \dots, T$ .

Los supuestos SGM1, SGM3-SGM5 establecen que los términos estocásticos son elegidos de manera no correlacionada de una distribución normal con media y

---

<sup>12</sup> En todo el documento, los momentos poblacionales y todas las distribuciones están condicionados a la información disponible de las variables en el modelo.

varianza constante, esto último debido a que ellos son homoscedásticos (SGM3). En notación matricial esas condiciones se pueden formular como

$$\varepsilon \sim N(\mathbf{0}, \sigma^2 I),$$

donde  $\mathbf{0}$  e  $I$  son de manera respectiva, un vector de ceros de dimensión  $T$  y la matriz identidad de  $T \times T$ .

### **3. IMPORTANCIA DE LA DISTRIBUCIÓN NORMAL EN LA INFERENCIA ESTADÍSTICA**

En esta Sección se hace una exposición sucinta de como el supuesto de normalidad de los términos estocásticos es utilizado en la inferencia estadística del modelo, es decir, en la estimación puntual de sus parámetros, en el planteamiento de intervalos de confianza y en la formulación de pruebas de hipótesis.

La estimación puntual de los parámetros suele realizarse mediante el método de mínimos cuadrados ordinarios (MCO) y el de máxima verosimilitud (MV), pero el primero no utiliza el supuesto de normalidad, mientras que en el de MV es fundamental. Obtenidos los estimadores, con el propósito de plantear intervalos de confianza y hacer pruebas de hipótesis, es necesario determinar las distribuciones de esos estimadores y como estos dependen de los errores estocásticos, sus distribuciones estarán determinadas por la normalidad. A continuación se presentan los aspectos básicos de la inferencia estadística del modelo de regresión lineal.

#### **3.1 Estimación puntual de los parámetros**

Una vez especificado el modelo de regresión lineal, se estiman los  $K$  parámetros en la ecuación:  $\beta_1, \beta_2, \dots, \beta_K$ , y el asociado a la varianza del término estocástico:  $\sigma^2$ , de manera que el total de parámetros que se deben estimar es  $K+1$ . Los métodos mediante los que se estima el modelo son el de mínimos cuadrados ordinarios (MCO) y el de máxima verosimilitud (MV).

El método de MCO consiste en minimizar la suma de cuadrados de los errores estocásticos, es decir,

$$\min S(\beta_1, \beta_2, \dots, \beta_K) = \sum_{t=1}^T \varepsilon_t^2$$

o equivalentemente en forma matricial,

$$\min S(\beta) = \varepsilon' \varepsilon = (Y - X\beta)'(Y - X\beta).$$

Resolver este problema implica plantear las condiciones de primer orden o ecuaciones normales a partir de las cuales se determina el punto crítico

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'Y$$

y mediante la matriz de segundas derivadas, el hessiano, se analiza que efectivamente en él se alcanza un mínimo. El procedimiento de MCO sólo proporciona el estimador de las betas, no el de la varianza de los errores,  $\sigma^2$ , pero se propone como estimador mínimo cuadrático de la varianza el siguiente

$$\hat{\sigma}_{MCO}^2 = \frac{1}{T-K} \sum_{t=1}^T \hat{\varepsilon}_t^2 = \frac{1}{T} \hat{\varepsilon}' \hat{\varepsilon}$$

donde  $\hat{\varepsilon}_t$  son los residuos, es decir,

$$\hat{\varepsilon}_t = y_t - \hat{y}_t, \quad t = 1, \dots, T.$$

y

$$\hat{y}_t = \hat{\beta}_1 + \hat{\beta}_2 x_{t2} + \dots + \hat{\beta}_K x_{tK}$$

es el ajuste del modelo o también nombrado el ajuste del modelo..

Por su parte, el método de MV como su nombre lo indica, maximiza la función de verosimilitud de los errores o de manera equivalente, su logaritmo, para lo cual se debe de considerar que  $\varepsilon_1, \varepsilon_2, \dots, \varepsilon_T$  constituyen una muestra aleatoria, es decir, un conjunto de variables aleatoria independientes e

idénticamente distribuidas. Como puede observarse, a diferencia del procedimiento de MCO en el que no se hace ningún supuesto sobre la distribución de los errores estocásticos para obtener los estimadores de los parámetros, el de MV debe considerar una distribución de los mismos y por supuesto, esa es la normal.

De esta forma, la función de verosimilitud de los términos estocásticos considera que son seleccionados de manera independiente de una distribución normal y dados los supuestos SGM1, SGM3-SGM5, su media debe ser cero, y su varianza  $\sigma^2$ , esto es,

$$\varepsilon_t \sim N(0, \sigma^2).$$

Por ello, la función de verosimilitud es

$$L(\beta_1, \dots, \beta_K; \varepsilon_1, \dots, \varepsilon_T) = \left( \frac{1}{2\pi\sigma^2} \right)^{\frac{T}{2}} e^{-\frac{1}{2\sigma^2} \sum_{t=1}^T \varepsilon_t^2}$$

y la log-verosimilitud está dada por

$$l(\beta_1, \dots, \beta_K; \varepsilon_1, \dots, \varepsilon_T) = -\frac{T}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{t=1}^T \varepsilon_t^2$$

Nuevamente, obtener la solución de maximizar esta función, implica determinar las condiciones de primer orden y solucionar el sistema de ecuaciones para obtener los puntos críticos, que en este caso están dados por

$$\hat{\beta}_{MV} = (X'X)^{-1}X'Y$$

y

$$\hat{\sigma}_{MV}^2 = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_t^2 = \frac{1}{T} \hat{\varepsilon}' \hat{\varepsilon}.$$

Observe que  $\hat{\beta}_{MCO} = \hat{\beta}_{MV}$ , pero  $\hat{\sigma}_{MCO}^2 \neq \hat{\sigma}_{MV}^2$  los cuáles difieren en sus grados de libertad y por ende, en sus correspondientes propiedades.

El teorema de Gauss-Markov establece que los estimadores mínimo cuadráticos, los de las betas, son los mejores estimadores lineales e insesgados, MELI o BLUE por sus siglas en inglés, *best linear unbaised estimator*, es decir, dentro de los lineales e insesgados son los de mínima varianza. Como  $\hat{\beta}_{MCO} = \hat{\beta}_{MV}$ , entonces los máximos verosímiles también son MELI.

Por su parte, el estimador de MCO de la varianza es insesgado, pero su varianza es mayor que la correspondiente al estimador máximo verosímil, pero éste a pesar de ser más eficiente que el mínimo cuadrático, es sesgado. De aquí en adelante y dado que el estimador mínimo cuadrático de beta es igual al máximo verosímil, se nombrara simplemente beta gorro, es decir,  $\hat{\beta}_k, k = 1, \dots, K$  o bien, en su forma vectorial  $\hat{\beta}$ .

Como puede observarse, el supuesto de normalidad del error aleatorio es de suma importancia para obtener los estimadores máxima verosímiles de los parámetros del modelo de regresión, no así para los mínimos cuadráticos, que prescinde de ese supuesto, y también es útil para determinar distribuciones que adquieren relevancia al formular intervalos de confianza y hacer pruebas de hipótesis para los parámetros del modelo incluyendo la  $\sigma^2$ , tal y como se verá en las siguientes Secciones.

### 3.2 Intervalos de confianza y pruebas de hipótesis

Implicaciones inmediatas del supuesto de normalidad de los errores estocásticos son las que se tienen sobre la distribución de la cantidad pivotal a partir de la cual se plantean los intervalos de confianza, y la del estadístico para llevar a cabo las pruebas de hipótesis, tanto para los parámetros en la especificación del modelo, las  $\beta_k$ , como para la varianza del error aleatorio,  $\sigma^2$ .

El desarrollo de esos dos tipos de inferencia para las betas generalmente se hace bajo dos escenarios, cuando  $\sigma^2$  es conocida y cuando no lo es, pero independientemente de ello se parte del hecho de que los términos estocásticos constituyen una muestra aleatoria, es decir, son variables aleatorias independientes e idénticamente distribuidas, esto es,

$$\varepsilon_t \sim N(0, \sigma^2).^{13}$$

Ese supuesto conduce a dos resultados importantes para hacer inferencia estadística. El primero es que cada  $\hat{\beta}_k$ ,  $k = 1, \dots, K$ , también se distribuye normal y como es insesgada y con varianza  $\sigma^2(X'X)_{ii}^{-1}$ , donde  $(X'X)_{ii}^{-1}$  es el elemento  $i$ -ésimo en la diagonal de  $(X'X)^{-1}$ , se obtiene que

$$\hat{\beta}_k \sim N(\beta_k, \sigma^2 (X'X)_{ii}^{-1})$$

y estandarizando se llega a

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 (X'X)_{ii}^{-1}}} \sim N(0,1). \quad (3)$$

El otro resultado es que

$$(T - K) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi_{T-K}^2, \quad (4)$$

donde  $\hat{\sigma}^2 = \hat{\sigma}_{MCO}^2$  y  $\chi_{T-K}^2$  indica la distribución chi- cuadrada con  $T-K$  grados de libertad.<sup>14</sup> A partir de estas expresiones se formulan los intervalos de confianza y se realizan las pruebas de hipótesis.

<sup>13</sup> Aquí solo se plantean los intervalos de confianza en esos escenarios, para que se note la diferencia en las distribuciones de las cantidades pivotales. Las pruebas de hipótesis se efectuarán solo bajo el supuesto de que la varianza del término estocástico es desconocida, que comúnmente es lo que sucede cuando se hace un modelo.

<sup>14</sup> En este punto es importante recordar que la distribución  $\chi^2$  es el resultado de sumar el cuadrado de variables aleatorias independientes e idénticamente distribuidas de manera normal estándar, en el contexto del

### 3.2.1 Intervalos de confianza para $\hat{\beta}_k$

Partiendo de la expresión en (3) y suponiendo que  $\sigma^2$  es conocida se llega después de un poco de álgebra, al intervalo de confianza

$$P\left(\hat{\beta}_k - z_{\alpha/2}\sqrt{\sigma^2 (X'X)^{-1}_{ii}} \leq \beta_k \leq \hat{\beta}_k + z_{\alpha/2}\sqrt{\sigma^2 (X'X)^{-1}_{ii}}\right) = 1 - \alpha,$$

donde  $z_{\alpha/2}$  es el valor crítico apropiado a una distribución normal y  $\alpha$  es el nivel de significancia.

Cuando  $\sigma^2$  es desconocida, se debe estimar y por ello, su estimador mínimo cuadrático se sustituye en (3) y entonces, la cantidad pivotal ya no se distribuye normal, tiene una distribución *t*-Student (*tS*) con  $T-K$  grados de libertad, es decir,<sup>15</sup>

$$\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{kk}}} \sim tS_{T-K}$$

y después de un poco de álgebra se obtiene el intervalo de confianza

$$P\left(\hat{\beta}_k - \tau_{\alpha/2}\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{ii}} \leq \beta_k \leq \hat{\beta}_k + \tau_{\alpha/2}\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{ii}}\right) = 1 - \alpha,$$

modelo de regresión, esas variables son los errores aleatorios, que se estandarizan para poder utilizar este resultado.

<sup>15</sup> Formalmente, esta cantidad pivotal y su distribución se obtiene mediante el cociente de la expresión con distribución normal en (3) y la raíz cuadrada de la  $\chi^2$  que se encuentra en (4) entre sus grados de libertad, esto es,

$$\frac{\frac{\hat{\beta}_k - \beta_k}{\sqrt{\sigma^2 (X'X)^{-1}_{kk}}}}{\sqrt{\frac{(T-K)}{(T-K)} \frac{\hat{\sigma}^2}{\sigma^2}}} = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{kk}}}$$

Como numerador y denominador son independientes, el cociente tiene una distribución *tS* cuyos grados de libertad son los de la  $\chi^2$ . Recuérdese que el cociente de una distribución normal y la raíz cuadrada de una  $\chi^2$  dividida por sus grados de libertad, tiene una distribución *tS* que hereda los grados de libertad de la chi-cuadrada.

en él,  $\tau_{\alpha/2}$  es el valor crítico asociado a una distribución  $tS$  con  $T-K$  grados de libertad.

### 3.2.2 Pruebas de hipótesis para $\hat{\beta}_k$ <sup>16</sup>

Considere ahora que se quiere probar las siguientes hipótesis

$$H_0: \beta_k = b_k \quad \text{vs} \quad H_1: \beta_k \neq b_k,$$

donde  $b_k$ ,  $k = 1, \dots, K$ , es una constante dada. Bajo el supuesto de que  $\sigma^2$  no es conocida, la expresión en (3) aun cuando se sustituya el estimador de  $\sigma^2$  no es un estadístico de prueba, puesto que el parámetro  $\beta_k$  es desconocido, pero bajo la hipótesis nula,  $\beta_k$  toma el valor  $b_k$  que sí se conoce, de manera que su sustitución en (3) conduce al siguiente estadístico de prueba bajo  $H_0$ ,

$$\tau = \frac{\hat{\beta}_k - b_k}{\sqrt{\hat{\sigma}^2 (X'X)^{-1}_{kk}}} \sim tS_{T-K}$$

y la región crítica, donde se rechaza  $H_0$  a un nivel de significancia  $\alpha\%$ , es

$$|\tau| > c_\alpha \tag{5}$$

donde  $c_\alpha$  es el valor crítico asociado a  $\alpha$ .

Otra forma equivalente, de determinar si la información empírica incorporada al modelo proporciona evidencia a favor o en contra de la hipótesis nula, es mediante el *p-value* o nivel de significancia marginal, se rechaza  $H_0$  si y solo si

$$p\text{-value} < \alpha.^{17} \tag{6}$$

---

<sup>16</sup> Debido a que la varianza del término estocástico generalmente es desconocida, en lo que sigue se hacen las pruebas para el caso en el que no es conocida.

<sup>17</sup> Un análisis detallado de estos aspectos se encuentran en Davidson y MacKinnon (2004) y Spanos (1999).

Dentro de estas pruebas de hipótesis tiene particular relevancia, la prueba de significancia individual, es decir, la que asume bajo  $H_0$  que  $b_k = 0$ , esto es,

$$H_0: \beta_k = 0 \quad \text{vs} \quad H_1: \beta_k \neq 0. \quad (7)$$

Esta es importante porque a través de ella se analiza si  $\beta_k$  es estadísticamente significativo, en cuyo caso, la variable que lo acompaña es importante desde el punto de vista estadístico, en la determinación de la variable dependiente.

### 3.2.3. Pruebas de hipótesis para combinaciones lineales de las betas

La prueba de hipótesis asociada a combinaciones lineales de los parámetros es una prueba conjunta que al igual que los intervalos de confianza y las pruebas de hipótesis estudiadas con anterioridad, basa su desarrollo en la normalidad del término estocástico del modelo de regresión. Considerando que se tienen  $m$  combinaciones lineales de los parámetros beta, las hipótesis a probar son

$$H_0: R\beta = r \quad \text{vs} \quad H_1: R\beta \neq r,$$

donde  $R$  es la matriz de los coeficientes de las combinaciones lineales y es de dimensión  $m \times K$ ,  $\beta$  es el vector que contiene los  $K$  parámetros beta y  $r$  es un vector de dimensión  $m$  con los términos independientes de cada restricción o combinación lineal. Ejemplos de este tipo de pruebas se encuentran en Johnston y Dinardo (1997), así como un estudio exhaustivo de las mismas.

La normalidad del error estocástico conduce también a que

$$\hat{\beta} \sim N(\beta, \sigma^2(X'X)^{-1})$$

y por tanto,

$$R\hat{\beta} \sim N(R\beta, \sigma^2 R(X'X)^{-1}R'),$$

lo cual conduce a

$$R\hat{\beta} - R\beta \sim N(\mathbf{0}, \sigma^2 R(X'X)^{-1}R')$$

donde el vector  $\mathbf{0}$  es de dimensión  $K$  y finalmente,

$$(R\hat{\beta} - R\beta)' [\sigma^2 R(X'X)^{-1}R']^{-1} (R\hat{\beta} - R\beta) \sim \chi_m^2.$$

Esta expresión y la planteada en (4) son formas cuadráticas independientes, cuyo cociente dividido numerador y denominador por sus correspondientes grados de libertad, conduce al estadístico de prueba que se muestra a continuación y que bajo  $H_0$ , se distribuye como una  $F$  con  $m$  y  $T-K$  grados de libertad,

$$f = \frac{1}{m\hat{\sigma}^2} (R\hat{\beta} - r)' [R(X'X)^{-1}R']^{-1} (R\hat{\beta} - r) \sim F_{(m, T-K)}$$

Los criterios de rechazo de la hipótesis nula son igual a los planteados en (5) y (6) usando el valor crítico o el *p-value*, pero ahora se debe usar la distribución  $F$  para decidir si rechazar o no  $H_0$ .<sup>18</sup>

Aquí es importante señalar dos aspectos dentro de este tipo de pruebas. El primero es que al igual que la de significancia individual, es decir, la que considera un sólo parámetro y que fue formulada en (7), la de significancia conjunta también tiene relevancia en la evaluación econométrica del modelo, en ella se considera bajo la hipótesis nula que los  $K-1$  parámetros que son coeficientes de las variables independientes, son cero, de manera que las hipótesis se plantean como

$$H_0: R\beta = \mathbf{0} \quad \text{vs} \quad H_0: R\beta \neq \mathbf{0},$$

donde

---

<sup>18</sup> Los criterios para el rechazo o no de la hipótesis nula siempre son los mismos, pero se debe saber tanto la distribución apropiada del estadístico de prueba para determinar el valor crítico y el *p-value*, como la hipótesis nula.

$$R = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 & 0 \\ 0 & 0 & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \cdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

y  $\mathbf{0}$  es un vector de ceros de dimensión  $m=K-1$ , o equivalentemente,

$$H_0: \beta_k = 0, \forall k = 2, \dots, K$$

vs

$$H_1: \beta_k \neq 0, \text{ para al menos una } k = 2, \dots, K.$$

De esta manera, en caso de que se rechace  $H_0$ , habrá evidencia a favor de que las variables independientes del modelo son estadísticamente diferentes de cero y por tanto, son relevantes estadísticamente para explicar a la variable independiente.

El otro punto a resaltar es que esta es la prueba más general y por tanto, la de significancia individual es un caso particular, en ella la matriz  $R$  es de la siguiente forma

$$R = (0 \quad 0 \quad \cdots \quad 1 \quad \cdots \quad 0 \quad 0)$$

es de dimensión  $K \times 1$  y el uno está en el lugar  $k$ -ésimo de  $R$ .

### 3.2.4 Intervalo de confianza y prueba de hipótesis para $\sigma^2$

A partir de la expresión en (4) dada por

$$(T - K) \frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{T-K}, \tag{8}$$

se puede plantear después de un poco de álgebra, el siguiente intervalo de confianza para la varianza del error estocástico,  $\sigma^2$ ,

$$P\left((T - K) \frac{\hat{\sigma}^2}{\chi_{\alpha/2}} \leq \sigma^2 \leq (T - K) \frac{\hat{\sigma}^2}{\chi_{1-\alpha/2}}\right) = 1 - \alpha,$$

y también se pueden probar las hipótesis

$$H_0: \sigma^2 = s \quad \text{vs} \quad H_1: \sigma^2 \neq s,$$

donde  $s > 0$  es una constante conocida y por tanto, bajo la hipótesis nula el estadístico de prueba es

$$(T - K) \frac{\hat{\sigma}^2}{s^2} \sim \chi_{T-K}^2,$$

que resulta de sustituir en (8) el valor de  $\sigma^2$  bajo la hipótesis nula.

#### 4. PRUEBA DE NORMALIDAD DE JARQUE-BERA

Cuando una variable aleatoria se distribuye normal, su tercer y cuarto momento alrededor de la media también conocidos como sesgo y curtosis, son cero y tres, de manera respectiva. El sesgo igual a cero da cuenta de que la distribución es simétrica, mientras que la curtosis igual a tres plantea que la distribución no es puntiaguda (leptocúrtica), ni achata (platicúrtica), en cuyo caso es normal o mesocúrtica.

Jarque y Bera (1980, 1987) formulan una prueba de normalidad que lleva su nombre, ellos plantean que existen distribuciones que pueden coincidir con la distribución normal, en media y varianza o sea, que su primer momento centrado en cero y su segundo alrededor de la media son los mismos, pero que no necesariamente el tercero y cuarto momentos centrados en la media son iguales. Esa es la razón que los conduce a plantear la prueba de normalidad basada en el sesgo,  $s$ , y la curtosis,  $c$ , de manera que las hipótesis a probar son

$$H_0: \text{Errores normales} \quad \text{vs} \quad H_1: \text{Errores no normales}$$

o equivalentemente,

$$H_0: s = 0, c = 3 \quad \text{vs} \quad H_1: s \neq 0 \text{ y/o } c \neq 3$$

y el estadístico de prueba bajo  $H_0$  es

$$JB = T \left[ \frac{\hat{cs}^2}{6} + \frac{(\hat{cc} - 3)^2}{24} \right] \sim \chi^2_{(2)}$$

donde  $\hat{cs}$  es el coeficiente de sesgo y el  $\hat{cc}$  coeficiente de curtosis dados por

$$\hat{cs} = \frac{\hat{s}}{(\sqrt{\hat{\sigma}^2})^3}, \quad \hat{cc} = \frac{\hat{c}}{(\sqrt{\hat{\sigma}^2})^4}$$

y

$$\hat{s} = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_t^3, \quad \hat{c} = \frac{1}{T} \sum_{t=1}^T \hat{\varepsilon}_t^4.$$

Observe que si el estadístico de prueba  $JB$  es cercano a cero hay evidencia a favor de que los errores se distribuyen de manera normal, en caso contrario, es decir, cuando  $JB$  está alejado de cero, se rechaza la hipótesis nula y las distribuciones de los estimadores de las betas y de la varianza de los errores estocásticos, no tendrán las distribuciones que permiten hacer inferencia estadística del modelo de regresión lineal y por tanto, ellas serán válidas sólo de manera asintótica de acuerdo al teorema de límite central.

## 5. PRUEBA JARQUE-BERA EN R

En la implementación de la prueba de Jarque-Bera en R, es necesario cargar el paquete tseries mediante la instrucción

```
> library(tseries)
```

y una vez que se cuenta en el objeto al que se la va aplicar la prueba se debe introducir

```
> jarque.bera.test(x)
```

en donde el argumento  $x$  es un vector o una serie de tiempo. Esta prueba puede llevarse a cabo para cualquier vector o serie de tiempo del que se desee saber si

se distribuye o no de manera normal. Sin embargo, en el contexto del modelo de regresión, la prueba se realiza sobre los residuales ya que estos son las *proxis* de los errores estocásticos que se suponen son normales, por ello la instrucción para efectuar la prueba de Jarque-Bera para los residuales del objeto llamado model, que guarda los resultados de estimación de la regresión, es

```
> jarque.bera.test(residuals(model))
```

y obviamente, debe de ejecutarse una vez que se estima el modelo. El resultado de la prueba presenta el nombre de la variable en `data`, el estadístico de prueba en `X-squared`, los grados de libertad en `df` y el mínimo nivel de significancia al que se rechaza la hipótesis nula, en `p-value`.

*Ejemplo 1.*

En este ejemplo se genera una variable (vector), que contiene cien números seleccionados de manera aleatoria de una distribución normal y se efectúa la prueba de normalidad para ese variable, pero antes se instala el paquete tseries, tal y como se muestra a continuación.

```
> library(tseries)
```

```
'tseries' version: 0.10-34
```

```
'tseries' is a package for time series analysis and  
computational finance.
```

```
See 'library(help="tseries")' for details.
```

```
> y<-rnorm(100)  
> jarque.bera.test(y)
```

```
Jarque Bera Test
```

```
data: y  
X-squared = 0.46901, df = 2, p-value = 0.791
```

A un nivel de significancia del 5%, la hipótesis nula de normalidad no es rechazada, puesto que  $p\text{-value} > 0.05$ .

*Ejemplo 2.*

La información anual de 1953 a 2004 contenida en el archivo Gasolina.txt es usada para estimar un modelo para la demanda de gasolina en USA (Greene,

2003). Se plantean dos regresiones log-log, en la primera se modela la demanda per-cápita en función del ingreso per-cápita, del índice de precios de la gasolina y el de los autos nuevos y en la segunda, se agrega el índice de precios agregado del consumo de bienes durables, y en ambas se prueba normalidad.<sup>19</sup> Las instrucciones en R son las que se presentan a continuación.

```
> library(tseries)
```

'tseries' version: 0.10-34

'tseries' is a package for time series analysis and computational finance.

See 'library(help="tseries")' for details.

```
> Gasolina <- read.csv("Gasolina.txt")
> View(Gasolina)
> attach(Gasolina)
> cons<-lm(log(G/Pobl)~log(Y)+log(Pg)+log(Pan))
> jarque.bera.test(residuals(cons))
```

Jarque Bera Test

```
data: residuals(cons)
X-squared = 7.3104, df = 2, p-value = 0.02586
```

```
> cons<-lm(log(G/Pobl)~log(Y)+log(Pg)+log(Pan)+log(Pd))
> jarque.bera.test(residuals(cons))
```

Jarque Bera Test

```
data: residuals(cons)
X-squared = 3.6263, df = 2, p-value = 0.1631
```

Con base en los resultados de la prueba de Jarque-Bera de la primera regresión, se rechaza la hipótesis nula de normalidad al 5%, puesto que  $p$ -

<sup>19</sup> Las variables del archivo son

Año: 1953-2004,

G: Gasto total en gasolina,

Pobl: Población

Pg: Índice de precio de la gasolina,

Y: Ingreso disponible per-cápita,

Pan: Índice de precios de los autos nuevos,

Pau: Índice de precios de los autos usados,

Ptp: Índice de precios del transporte público,

Pd: Índice de precios agregado del consumo de bienes durables,

Pnd: Índice de precios agregado del consumo de bienes no durables,

Ps: Índice de precios agregado para el consumo de servicios.

Fuente: <http://people.stern.nyu.edu/wgreene/Text/econometricanalysis.htm>

$p-value < 0.05$ , pero no al 1% de significancia ( $p-value > 0.01$ ), mientras que en la segunda regresión no se rechaza la hipótesis nula y por ello se infiere que los errores son normales.

### Ejemplo 3

En este ejemplo se presenta un modelo estático para la elasticidad de sustitución Armington para México.<sup>20</sup> La estimación incorpora información trimestral del INEGI para el periodo que comprende del primer trimestre de 1993 al primero del 2013, a precios constantes del 2008. La variable explicada en el modelo es la demanda relativa (DRel), que resulta del cociente entre las importaciones totales y la demanda doméstica (diferencia entre el valor bruto de la producción y las exportaciones, ambas a precios de mercado) y las variables explicativas son el precio relativo (PRel), que se obtiene de dividir el índice de precios de la demanda relativa entre el correspondiente a las importaciones, y el producto interno bruto (PIB).

Las siguientes instrucciones en R permiten estimar el modelo de regresión log-log con las variables descritas previamente y efectuar la prueba de normalidad de Jarque-Bera.

```
> Elast <- read.csv("Elast.txt")
> View(Elast)
> attach(Elast)
> model<-lm(log(DRel)~log(PRel)+log(PIB))

> jarque.bera.test(residuals(model))
```

Jarque Bera Test

```
data: residuals(model)
X-squared = 4.9739, df = 2, p-value = 0.08316
```

El  $p$ -value implica que la hipótesis nula de normalidad de los errores aleatorios no se rechaza a un nivel de significancia del 5%, pero si al 10%, puesto que  $p-value > 0.05$  y  $p-value < 0.10$ .

---

<sup>20</sup> Un análisis detallado de este modelo desde sus microfundamentos hasta la especificación final de un modelo dinámico es presentado en Casares, Ruiz-Galindo y Sobarzo (por publicarse).

## **6. CAUSAS E IMPLICACIONES DE LA NO NORMALIDAD Y POSIBLES SOLUCIONES**

Dos son las causas principales de que los residuos del modelo no se distribuyan de manera normal: una es que la muestra no es lo suficientemente grande como para garantizarla y la otra es que si a los datos que se incorporaron al modelo se les hizo alguna transformación, ella no fue la adecuada.

Cuando los datos son pocos y hay posibilidad de obtener más, habrá que incluirlos para obtener una nueva estimación del modelo. Si esto no es posible, habrá que hacer una transformación de la familia Box y Cox, de las cuales la más utilizada es la logarítmica, y que además también puede corregir heteroscedasticidad.

Considere que se quiere transformar a variable  $w$  cuyos valores son positivos, la transformación Box-Cox depende de un parámetro  $\lambda$  y es la siguiente

$$z(\lambda) = \begin{cases} \frac{w^\lambda - 1}{\lambda}, & \lambda \neq 0 \\ \log w, & \lambda = 0 \end{cases}$$

Cuando  $w$  no es positiva, se le suma una constante de manera que se obtengan una nueva variable cuyos valores sí lo sean.

## **7. CONCLUSIONES**

La elaboración de los modelos econométricos conlleva dos tipos de evaluación. Una que se basa en la teoría económica que fue utilizada para la especificación del mismo y en la que se revisan que los signos de los parámetros estimados y su magnitud, entre otros aspectos, coincidan con los que formula la teoría. La otra es la evaluación econométrica, que consiste en analizar la significancia individual y conjunta de los parámetros y verificar si se satisfacen tanto los supuestos del modelo de regresión en su parte determinista como los que se plantean en los términos o errores aleatorios.

Dentro de la evaluación econométrica reviste importancia la normalidad, ya que este supuesto aunque no necesario en la estimación de los parámetros del modelo, resulta indispensable en las otras dos formas de hacer inferencia

estadística, a saber, en el planteamiento de los intervalos de confianza y de las pruebas de hipótesis. A partir de la normalidad de los errores aleatorios, se obtiene las distribuciones apropiadas de las cantidades pivotales para plantear intervalos de confianza y de los estadísticos de prueba para efectuar pruebas de hipótesis.

Por lo anterior, una vez que se ha analizado que los parámetros estimados tienen los signos y magnitudes apropiadas de acuerdo a la teoría económica subyacente, se debe revisar si los residuos del modelo que son las *proxis* de los términos estocásticos, son normales, de no ser así se corre el riesgo de hacer inferencia de manera incorrecta a menos que se tenga una gran cantidad de observaciones para cada variable, en cuyo caso se recurre al teorema de límite central que garantiza normalidad cuando el tamaño de la muestra tiende a infinito, en la práctica esto significa que se tienen muchas observaciones y por tanto, los resultados de inferencia estadística expuestos aquí son válidos de manera asintótica. Sin embargo, si no se puede incrementar el número de observaciones o bien a pesar de haberlo hecho no se obtuvo normalidad, se debe usar una transformación de las variables de la familia Box-Cox y de manera específica, la logarítmica que es la más utilizada en estas situaciones.

## REFERENCIAS

- Casares, E. R., L. A. Ruiz-Galindo y H. Sobarzo, (por publicarse). "Short and Long Run Armington Elasticities for the Mexican Economy" en A. Pinto y D. Zilberman (editors), Modeling, Dynamics, Optimization and Bioeconomics II, en la serie Springer Proceedings in Mathematics an Statistics.
- Davidson R. y J. G. MacKinnon, (2004). Econometric Theory an Methods. Ed. Oxford University Press, New York.
- Greene, W. H., (2007). Econometric Analysis. Ed. New York University, New York.

Jarque, C. M. y A. K. Bera (1980). "Efficint test s for normality, heteroskedasticity and serial independdence of regression residuals", *Economics Letters*, vol. 6, 255-259.

Jarque, C. M. y A. K. Bera (1987). "A Test for Normality of Oservations and Regression Residuals", *International Statistical Review*, vol 55, 2,163-172.

Johnston, J. y J. Dinardo, (1997). *Economerics Methods* Ed. McGraw-Hill, Singapur.

Spanos, A., (1999). *Probability Theory and Statistical Inference. Econometric Modeling with Observational Data*. Ed. Cambridge University, Reino Unido.

## Referencias electrónicas

Datos (Greene, 2007),  
<http://pages.stern.nyu.edu/~wgreene/Text/econometricanalysis.htm>

INEGI (2013a), “Banco de Información Económica”, <http://dgcnesyp.inegi.gob.mx>

## ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO

Gasolina.txt

Elast.txt

## MATERIAL DE APRENDIZAJE EN LÍNEA

Teória\_Cap5

Práctica\_Cap5

VideoPráctica\_Cap5

VideoTeoría\_Cap5

# CAPÍTULO 6: MULTICOLINEALIDAD

Luis Quintana Romero y Miguel Ángel Mendoza

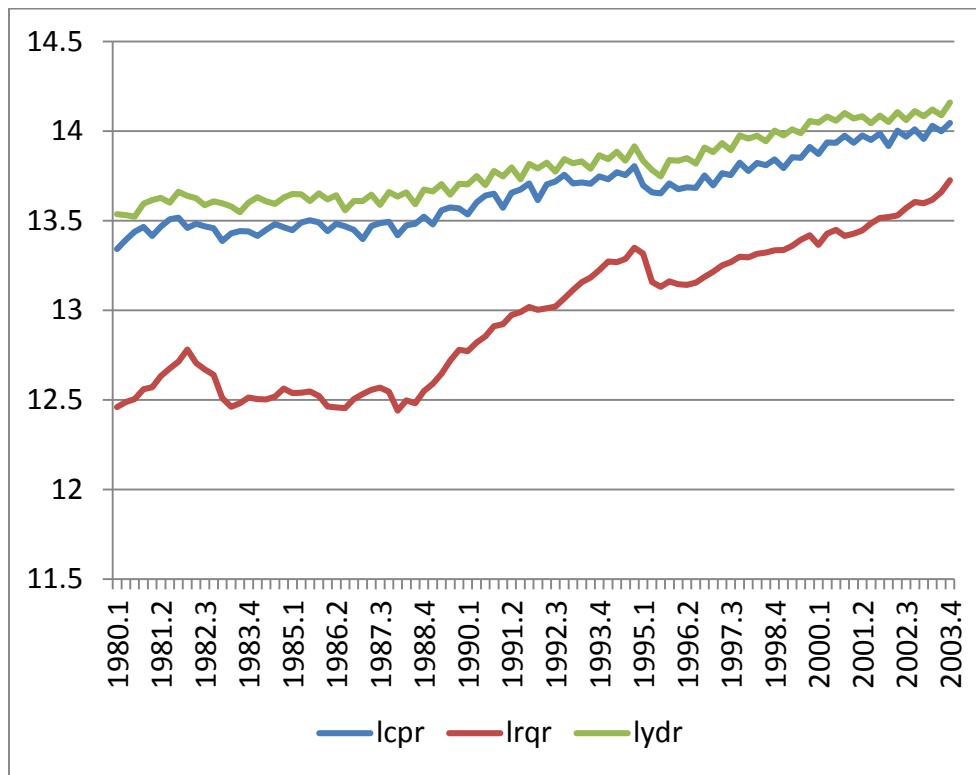
## 1. LA MULTICOLINEALIDAD UN PROBLEMA DE GRADO

La multicolinealidad debe considerarse como un problema de grado que se presenta de manera cotidiana en los modelos econométricos. Esto significa que el comportamiento de buena parte de las variables económicas guarda algún tipo de relación unas con otras y esa relación puede ser de menor o mayor grado. Solamente cuando dicha relación es de mayor grado podría ser un problema dentro de la modelación econométrica tal y como veremos a continuación.

Para ilustrar la relación que existe entre las variables económicas en la gráfica siguiente se muestran los valores logarítmicos trimestrales del consumo privado real, el ingreso nacional disponible real y la riqueza real para la economía mexicana de 1980 a 2003. En la gráfica se observa que el consumo y el ingreso prácticamente tiene el mismo comportamiento, mientras que la riqueza tiene la misma tendencia que las otras dos variables; la gráfica muestra una clara asociación positiva entre las tres variables, lo cual implica que debe de existir algún grado de asociación lineal entre las variables que hemos seleccionado.

**Gráfica 1**

**Consumo, ingreso y riqueza por trimestre en México 1980-2003**



Si bien las variables muestran trayectorias similares existen diferencias entre ellas, por ende están relacionadas de forma aproximada pero no exacta, esto nos permite plantear que la multicolinealidad es la relación perfecta o no, que se da entre variables económicas.

La relación exacta entre las variables se denomina multicolinealidad perfecta, lo cual significa que alguna o algunas de las variables que forman las columnas de la matriz de regresores sería una combinación lineal exacta del resto de columnas. Por ejemplo, si suponemos que la matriz de regresores se compone de tres columnas con las variables  $x_1, x_2, x_3$  se obtendría la siguiente relación lineal:

$$\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 = 0 \quad (1)$$

Siendo las constantes  $\lambda_i$  simultáneamente diferentes de cero, esto es;  $\lambda_i \neq 0 \forall i$ . Lo cual permitiría expresar una variable en términos de las demás, por ejemplo al despejar  $x_1$ :

$$x_1 = \frac{-\lambda_2 x_2 - \lambda_3 x_3}{\lambda_1} \quad (2)$$

Si los coeficientes fueran nulos no habría forma de obtener combinación lineal alguna y las columnas de la matriz de regresores serían linealmente independientes y dicha matriz sería no singular.

La multicolinealidad perfecta en realidad debe considerarse un caso poco frecuente en los modelos econométricos, que de ocurrir tendría como consecuencia la violación del supuesto de rango completo de la matriz de regresores  $[X]$  y en consecuencia tampoco se cumpliría para la matriz  $[X'X]$ , siendo singulares ambas matrices y sus determinantes iguales a cero, lo que daría lugar a la indeterminación de los estimadores de mínimos cuadrados ordinarios para los parámetros del modelo. Esta situación se explica debido a que no estaría definida la matriz inversa  $[X'X]^{-1}$ , que como sabemos es necesaria para obtener los estimadores de mínimos cuadrados ordinarios:  $\hat{\beta} = [X'X]^{-1}[X'Y]$ .

En realidad el problema de la multicolinealidad debe ser visto como un problema de identificación, ya que alternativamente diferentes valores de los parámetros en el modelo generan el mismo valor estimado de la variable dependiente, lo que impide identificar el efecto individual de cada variable.

Resulta más usual que se presente multicolinealidad imperfecta, lo cual intuitivamente implica que los regresores de la regresión se encuentran altamente correlacionadas, pero sin ser esos coeficientes del cien por ciento. En términos de la matriz de regresores, significa que el determinante de la matriz  $[X]$  es cercano a cero, sin embargo ello no impide la obtención de los estimadores de mínimos cuadrados ordinarios, pero se mantiene el problema de identificación debido a que la variación de alguna de las  $X$ 's además de afectar a  $Y$  afectan a las demás variables impidiendo distinguir su efecto individual.

Si suponemos nuevamente que la matriz de regresores se compone de tres columnas con las variables  $x_1, x_2, x_3$  se obtendría la siguiente relación lineal imperfecta entre ellas:

$$\lambda_1 x_1 + \lambda_2 x_2 + \lambda_3 x_3 + v = 0 \quad (3)$$

Siendo las constantes  $\lambda_i$  simultáneamente diferentes de cero, como en el caso previo, pero ahora existe un término de error  $v$ . Debido a esto último, al despejar y expresar una variable en términos de las demás, por ejemplo al despejar  $x_1$ , la combinación lineal que se obtiene ya no es exacta y, por ende, la multicolinealidad ya no es perfecta:

$$x_1 = \frac{-\lambda_2 x_2 - \lambda_3 x_3}{\lambda_1} + \frac{v}{\lambda_1} = \text{combinación lineal} + \text{error} \quad (4)$$

Para tener una idea más precisa de lo que ocurre cuando la colinealidad entre las columnas de la matriz de regresores se incrementa, en el cuadro siguiente se muestra los que sucede con el determinante de la matriz y con los errores estándar de los estimadores de mínimos cuadrados ordinarios al irse incrementando el grado de correlación entre las variables. Para simplificar el asunto se supondrá que la varianza residual es una constante iguala la unidad, por ello  $\sigma^2 = 1$ . Claramente se observa que al ir aumentando la colinealidad entre las columnas de la matriz  $X$ , el determinante disminuye y las varianzas de los estimadores se van incrementando. En el caso límite, cuando las columnas de la matriz son iguales, se tiene multicolinealidad perfecta y el determinante se hace cero por lo que es imposible calcular la matriz inversa necesaria para la obtención de los estimadores de mínimos cuadrados ordinarios y las varianzas de los estimadores tienden a infinito.

**Cuadro 1**  
**Ejemplo matricial de la multicolinealidad**

Matriz X	Determinante	Varianza: $\sigma^2[X'X]^{-1}$
$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$	1	$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
$\begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$	0.75	$\begin{bmatrix} 1.333 & -0.666 \\ -0.666 & 1.333 \end{bmatrix}$
$\begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$	0.36	$\begin{bmatrix} 2.777 & -2.222 \\ -2.222 & 2.777 \end{bmatrix}$
$\begin{bmatrix} 1 & 0.99 \\ 0.99 & 1 \end{bmatrix}$	0.02	$\begin{bmatrix} 50.251 & -49.749 \\ -49.749 & 50.251 \end{bmatrix}$
...	...	...
$\begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}$	0.00	No definida

## 2. PRUEBAS PARA LA DETECCIÓN DE MULTICOLINEALIDAD

Algunas de las pruebas más usuales para detectar multicolinealidad son las siguientes (Quintana y Mendoza, 2008):

### a) Coeficientes t's no significativos y R<sup>2</sup> elevada

Una elevada R<sup>2</sup> junto con uno u algunos coeficientes t poco significativos es una de las pruebas más tradicionales para evaluar multicolinealidad. Del cuadro 1 es fácil comprender que los estadísticos t tenderán a disminuir debido a que su denominador se va incrementando paulatinamente al elevarse la colinealidad entre las variables.

### b) Coeficientes de correlación

Elevados coeficientes de correlación entre pares de variables son un síntoma a favor de la multicolinealidad. Es usual considerar que coeficientes de correlación

entre las variables por encima de 0.8 u 80% son evidencia de correlación seria, sin embargo también existen modelos con multicolinealidad grave y bajos coeficientes de correlación debido a que dicho coeficiente es sensible a transformaciones de las variables.

### c) Regresiones auxiliares y efecto $R^2$ de Theil

Se corren regresiones auxiliares de la variable dependiente contra los k regresores menos uno de ellos, al coeficiente de determinación de esas regresiones se le denomina  $R_i^2$ . El efecto  $R^2$  de Theil (1971) se obtiene con la siguiente expresión:

$$R^2\text{Theil} = R^2 - [\sum_{i=1}^n (R^2 - R_i^2)] \quad (5)$$

donde  $R^2$  es el coeficiente de determinación de la regresión original con todos los regresores y  $R_i^2$  es el coeficiente de determinación de la regresión auxiliar i. Si el efecto de Theil fuera nulo no existiría multicolinealidad, entre mayor sea el efecto más grave es el problema.

### d) Regresiones auxiliares y regla de Klein

La regla de Klein (1967) es un principio práctico, propuesto por el premio Nobel Lawrence Klein. De acuerdo a dicho principio, la multicolinealidad es un problema a considerar si la  $R_i^2$  de alguna regresión auxiliar es mayor que el coeficiente de determinación  $R^2$  de la regresión original.

En este caso, las regresiones auxiliares son diferentes a las de Theil, ya que se efectúan tomando cada uno de los regresores y corriendo regresiones con los regresores restantes. Por ejemplo, si se tuvieran tres regresores  $x_1, x_2, x_3$  en el modelo, las regresiones auxiliares serían las siguientes:

$$x_{1i} = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{3i} + \varepsilon_{1i} \quad (6)$$

$$x_{2i} = \alpha_1 + \alpha_2 x_{1i} + \alpha_3 x_{3i} + \varepsilon_{2i} \quad (6a)$$

$$x_{3i} = \alpha_1 + \alpha_2 x_{2i} + \alpha_3 x_{1i} + \varepsilon_{3i} \quad (6b)$$

siendo  $i=1,2,\dots,n$  y  $\varepsilon_{1i}, \varepsilon_{2i}, \varepsilon_{3i}$  los usuales términos de perturbación aleatoria.

En este caso tendremos tres coeficientes de determinación de las regresiones auxiliares  $R_1^2, R_2^2, R_3^2$  si alguno de ellos es mayor a  $R^2$  el problema de multicolinealidad se puede considerar grave.

#### f) Índice de la condición de número

Este método hace uso de las propiedades de los valores característicos de una matriz, como sabemos el número de valores característicos diferentes de cero es igual al rango de la matriz y el producto de los valores característicos es su determinante.

Para calcular el índice de la condición de número (ICN) se deben obtener los valores característicos de la matriz  $[X'X]$ , a los cuales denominaremos  $\lambda_i$  y se divide el máximo valor característico entre el menor valor característico:

$$ICN = \frac{\sqrt{\lambda_{\text{máximo}}}}{\sqrt{\lambda_{\text{mínimo}}}} \quad (7)$$

Como los valores característicos dependen de las unidades de medida de los datos, es mejor normalizar primero las variables de la matriz  $X$  para después calcular los valores característicos. Si las columnas de  $X$  son ortogonales la condición de número será igual a la unidad. En la práctica una condición de número superior a 20 se considera síntoma de multicolinealidad problemática.

#### g. Factor de inflación varianza

El factor de inflación varianza (VIF) se utiliza como una medida del grado en que la varianza del estimador de mínimos cuadrados es incrementada por la colinealidad entre las variables. El VIF se define de la manera siguiente:

$$VIF = \frac{1}{1-R_i^2} \quad (8)$$

En donde  $R_i^2$  es el coeficiente de determinación de la regresión auxiliar i, tal y como se mostró en el caso previo. Por ejemplo, ante perfecta multicolinealidad  $R_i^2 = 1$ , lo cual hace que el VIF tienda a infinito, si la multicolinealidad es imperfecta y elevada, por ejemplo un  $R_i^2 = 0.9$ , el VIF será igual a 10. Es usual en la práctica que si el VIF resulta mayor a 10 o incluso 5 sea considerado como evidencia de fuerte multicolinealidad.

### **3. UN EJEMPLO PRÁCTICO EN LA DETECCIÓN DE MULTICOLINEALIDAD EN R CON LA FUNCIÓN CONSUMO PARA MÉXICO**

Para tener una idea intuitiva de las implicaciones de la multicolinealidad, en esta sección se realiza primero una simulación con datos artificiales y después se procede a abordar un caso real para México,

Para realizar la simulación se deben generar dos variables, en donde una de ellas es independiente y la otra es una combinación lineal de aquella.

El proceso generador de los datos PGD se puede formular como:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \quad (9)$$

siendo:

$$x_{3i} = \gamma_i + 5x_{2i}$$

$$y_i = 2 + 0.5x_{2i} + 0.1x_{3i} + u_i \quad (10)$$

donde:

$x_{2i}$  y  $x_{3i}$  son series de 1000 variables seudo aleatorias generadas artificialmente con distribución normal, media 0 y varianza unitaria.

$\gamma_i$  es una variable aleatoria normalmente distribuida

$u_i$  es un término de perturbación aleatoria con media cero y varianza constante 0.4

Para construir nuestras variables utilizaremos el generador de números seudoaleatorios de R, por lo cual lo primero que debemos hacer es fijar el valor semilla con el que se generarán los números, en este caso lo fijamos en 50:

```
set.seed(50)
```

Ahora generamos nuestras variables aleatorias con rnorm y corremos la regresión con lm:

```
X2=rnorm(100,0,1)
X3=rnorm(100,0,1)+5*X2
Y=2+0.5*X2+0.1*X3+rnorm(100,0,4)
summary(lm(Y~X2+X3))
```

Los resultados de la regresión se muestran a continuación, en ellos se puede observar que el coeficiente de  $X_3$  no es estadísticamente significativo y la  $R^2$  ajustada es relativamente elevada. Esto significa que debido a la colinealidad entre  $X_2$  y  $X_3$  no es posible separar el efecto de cada una de las variables en la variable dependiente, además de que la varianza del coeficiente de  $X_3$  es muy alta por lo cual el estadístico t es muy bajo.

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.96761	0.03868	50.867	< 2e-16 ***
X2	0.69746	0.20364	3.425	0.000903 ***
X3	0.05881	0.03994	1.472	<b>0.144144</b>
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.3863 on 97 degrees of freedom

Multiple R-squared: 0.8772,      **Adjusted R-squared: 0.8747**

F-statistic: 346.5 on 2 and 97 DF, p-value: < 2.2e-16

Si la colinealidad fuera perfecta entre  $X_2$  y  $X_3$ ,  $X_3$  sería una combinación lineal perfecta de  $X_2$  y el proceso generador podría ser:

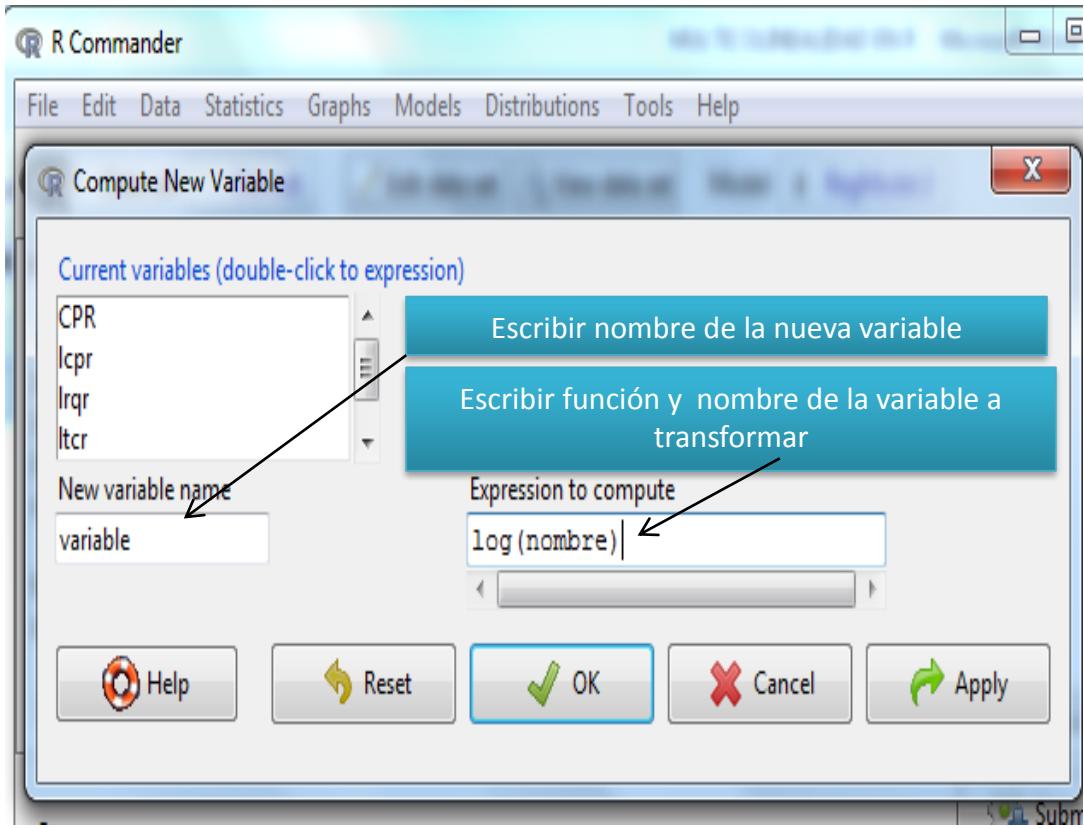
$$x_{3i} = 5x_{2i} \quad (11)$$

Sí incorpora este nuevo proceso en nuestra simulación, el R automáticamente elimina una de las variables y envía una alerta de que uno de los coeficientes no está definido debido a un problema de singularidad en la matriz de regresores, tal y como se observa en el recuadro siguiente:

```
Im(formula = Y ~ X2 + X3)
Residuals:
    Min   1Q Median   3Q   Max
-0.80422 -0.19019  0.01836  0.17085  0.81986
Coefficients: (1 not defined because of singularities)
            Estimate Std. Error t value Pr(>|t|)
(Intercept) 1.94072  0.03649   53.19 <2e-16 ***
X2          1.04002  0.03741   27.80 <2e-16 ***
X3          NA        NA       NA      NA
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.3643 on 98 degrees of freedom
Multiple R-squared: 0.8875, Adjusted R-squared: 0.8863
F-statistic: 772.8 on 1 and 98 DF, p-value: < 2.2e-16
```

En los datos del archivo consumo\_fun.txt se presenta información trimestral para la economía mexicana del consumo privado (CPR), la riqueza real (RQR), y el ingreso disponible real (YPD).

Para utilizar los datos en R los importamos a través del RCommander y una vez cargados en el DATASET realizamos una transformación logarítmica de las variables seleccionando en el menú principal DATA/Manage variables in active dataset/Compute a new variable. Se abrirá una ventana en la cual simplemente en el espacio de New variable name se anota el nuevo nombre de la variable y en el espacio Expression to compute se escribe la función, en este caso log, y en paréntesis el nombre de la variable a transformar, tal y como se muestra en la imagen siguiente.



Con las variables transformadas en logaritmos se estima la siguiente ecuación:

$$lcpr_t = \beta_1 + \beta_2 lrqr_t + \beta_3 lypdr_t + \beta_4 ltcr_t + u_t \quad (12)$$

donde:

$lcprt$  es el logaritmo del consumo privado real en miles de millones de pesos de 1993

$lrqrt$  es el logaritmo de la riqueza real calculada como el cociente del agregado monetario M4 dividido entre el índice de precios al consumidor.

$lyndrt$  es el logaritmo del ingreso nacional disponible real en miles de millones de pesos de 1993

$ltcrt$  es el logaritmo del tipo de cambio real

Los resultados de la regresión se muestran a continuación, de ellos se desprende que un incremento del diez por ciento en la riqueza da lugar a un aumento del 15.4% en el consumo, mientras que una variación de la misma magnitud en el ingreso eleva en 71% al consumo. De los resultados también se observa que el tipo de cambio tiene un efecto negativo, pero éste no resulta estadísticamente significativo.

Call:

lm(formula = lcpr ~ lrqr + ltcr + lypdr, data = Dataset)

Residuals:

Min	1Q	Median	3Q	Max
-0.061536	-0.017314	-0.001635	0.020202	0.072171

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.90203	0.54239	3.507	0.000703 ***
lrqr	0.15401	0.03161	4.873	4.57e-06 ***
ltcr	-0.03185	0.02053	-1.551	0.124223
lypdr	0.71042	0.06637	10.704	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03154 on 92 degrees of freedom

Multiple R-squared: 0.9744, Adjusted R-squared: 0.9735

F-statistic: 1165 on 3 and 92 DF, p-value: < 2.2e-16

En los resultados previos es relevante examinar la posible existencia de multicolinealidad en virtud de la fuerte relación que puede existir entre las tres variables explicativas; la riqueza de los individuos se forma a través de su ingreso y estas dos variables son afectadas sensiblemente por lo que ocurre con los precios de los bienes importados, cuyo efecto es tomado en cuenta por el tipo de cambio.

Una primer evidencia de posible elevada colinealidad entre las variables se deriva de la alta  $R^2$  ajustada de 0.97 y la nula significancia de una de las variables. Para intentar confirmar esta evidencia es preciso realizar algunas exploraciones adicionales.

### a) Coeficientes de correlación

Los coeficientes de correlación entre las variables se calculan con la función `cor` del R;

```
cor(Dataset[,c("lydr","lrqr","ltcr")], use="complete")
```

En RCommander basta seleccionar en el menú principal STATISTICS/Summaries/Correlation matrix. En la ventana que se abre basta seleccionar las variables a correlacionar y el tipo de correlación que en este caso es el Pearson. Los resultados son los siguientes:

```
> cor(Dataset[,c("lrqr","ltcr","lypdr")], use="complete")
      lrqr        ltcr       lypdr
lrqr  1.0000000 -0.528662  0.9632604
ltcr -0.5286620  1.000000 -0.4918170
lypdr 0.9632604 -0.491817  1.0000000
```

En los resultados es posible observar que las correlaciones son muy altas entre el ingreso y la riqueza, 96%, mientras que con el tipo de cambio las correlaciones son relativamente bajas. Por ello, de existir algún problema de colinealidad se deriva de las primeras dos variables.

### b) Factor de inflación-varianza (VIF)

Para calcular el VIF en RCommander seleccionamos del menú principal MODELS/Numeric diagnostics/Variance inflation factors. En los resultados siguientes es posible establecer la existencia de problemas de colinealidad graves en virtud de que las variables de riqueza y de ingreso presentan un VIF muy por arriba de diez unidades.

```
> vif(RegModel.3)
   lrqr      ltcr      lypdr
14.673133  1.396047  13.945404
```

### c) Regresiones auxiliares: La regla de Klein.

Al correr una regresión auxiliar tomando al ingreso como variable dependiente y a la riqueza y el tipo de cambio como explicatorias obtenemos una  $R^2$  ajustada de 0.9267 la cual es inferior a la de 0.9765 del modelo original, tal y como se observa en los resultados del recuadro siguiente. Esto implica que el problema de multicolinealidad no es muy grave.

```
Call:
lm(formula = lypdr ~ lrqr + ltcr, data = Dataset)
Residuals:
    Min      1Q      Median      3Q      Max 
-0.125362 -0.036361  0.004442  0.034763  0.108628 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 7.88543   0.22250  35.439 <2e-16 ***
lrqr        0.45315   0.01519  29.837 <2e-16 ***
ltcr        0.02364   0.03198   0.739   0.462    
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1 
Residual standard error: 0.04928 on 93 degrees of freedom
Multiple R-squared:  0.9283,    Adjusted R-squared:  0.9267 
F-statistic: 602 on 2 and 93 DF,  p-value: < 2.2e-16
```

### d) Regresiones auxiliares: El efecto de Theil.

Con base en los resultados de la regresión auxiliar previa y los de las regresiones auxiliares excluyendo a uno de los regresores se puede calcular el efecto de Theil.

```
Call:
```

**lm(formula = lcpr ~ lrqr + ltcr, data = Dataset)**

Residuals:

	Min	1Q	Median	3Q	Max
	-0.120729	-0.035090	0.002992	0.037276	0.102336

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.50399	0.21225	35.354	<2e-16 ***
lrqr	0.47593	0.01449	32.851	<2e-16 ***
ltcr	-0.01506	0.03051	-0.494	0.623

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04701 on 93 degrees of freedom

Multiple R-squared: **0.9424**, Adjusted R-squared: 0.9412

F-statistic: 760.9 on 2 and 93 DF, p-value: < 2.2e-16

Call:

**lm(formula = lcpr ~ lrqr + lypdr, data = Dataset)**

Residuals:

	Min	1Q	Median	3Q	Max
	-0.063740	-0.020311	0.000018	0.019144	0.069434

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	1.81882	0.54380	3.345	0.00119 **
lrqr	0.16552	0.03096	5.347	6.36e-07 ***
lypdr	0.70255	0.06667	10.537	< 2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03178 on 93 degrees of freedom

Multiple R-squared: **0.9737**, Adjusted R-squared: 0.9731

F-statistic: 1720 on 2 and 93 DF, p-value: < 2.2e-16

Call:

**lm(formula = lcpr ~ ltcr + lypdr, data = Dataset)**

Residuals:

	Min	1Q	Median	3Q	Max
	-0.074269	-0.020489	-0.001975	0.018901	0.082560

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.31390	0.32972	-0.952	0.3436
ltcr	-0.05534	0.02226	-2.486	0.0147 *
lypdr	1.01813	0.02277	44.711	<2e-16 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.03519 on 93 degrees of freedom

Multiple R-squared: 0.9677, Adjusted R-squared: 0.967

F-statistic: 1395 on 2 and 93 DF, p-value: < 2.2e-16

Con los datos del recuadro previo es posible calcular el efecto de Theil utilizando la  $R^2$  original de 0.9744 y las  $R^2$  de las ecuaciones auxiliares de la manera siguiente:

$$0.9744 - (0.9744 - 0.9424) - (0.9744 - 0.9737) - (0.9744 - 0.9677) = 0.935$$

El resultado es la reducción en el efecto individual de la suma de las variables explicatorias debido a la multicolinealidad en relación con el que hubieran tenido de ser independientes las variables.

#### e) La condición de número

En RCommander se pueden calcular los valores característicos para la matriz de regresores del modelo de la ecuación (12). Para ello se debe seleccionar en el menú principal la secuencia de opciones: STATISTICS/Dimensional analysis/Principal component analysis. A continuación se abre una ventana en la que se deben seleccionar las variables que componen la matriz de regresores, que en este caso son lrqr, lypdr y ltcr. También se deben establecer las opciones las cuales permiten analizar la matriz de correlaciones, generar una gráfica de las componentes y sus varianzas, además de permitir añadir las componentes a la tabla de datos.

En el caso de los regresores de la función consumo el R nos presenta las tres raíces características ordenadas de mayor a menor tal y como se observa en el recuadro siguiente.

```
> .PC$sd^2 # component variances  
Comp.1 Comp.2 Comp.3  
2.34925314 0.61498514 0.03576172
```

Al sustituir estos resultados en la fórmula del ICN obtenemos:

$$\text{ICN} = \frac{\sqrt{\lambda_{\text{máximo}}}}{\sqrt{\lambda_{\text{mínimo}}}} = \frac{\sqrt{2.34925314}}{\sqrt{0.03576172}} = 8.105$$

El valor del ICN es inferior al umbral de 20 que se ha definido en la literatura para establecer un grado de multicolinealidad grave, por consiguiente no habría que preocuparse de este problema en el modelo.

#### **4. SOLUCIONES AL PROBLEMA DE LA MULTICOLINEALIDAD**

Una vez que se ha detectado que el grado de multicolinealidad del modelo es grave, se puede optar por una serie de métodos de corrección. Debe señalarse que si el problema de multicolinealidad no es severo más vale no hacer nada, ya que los remediales generalmente pueden implicar problemas más fuertes que el que se buscaba corregir. Debe considerarse que frente a un problema de multicolinealidad los estimadores de mínimos cuadrados ordinarios siguen siendo insesgados, de modo que si el problema no es grave el modelo puede utilizarse sin que afecte en gran medida a la inferencia estadística. Incluso si el objetivo de la modelación no fuera el análisis estructural sino el mero pronóstico, la multicolinealidad no tendría mayor efecto dado que la relación entre las variables se mantiene tanto en el horizonte histórico como en el futuro de las variables.

De cualquier forma, si se quiere hacer algo para resolver el problema los remedios usuales son los siguientes:

##### **a) Imponer restricciones al modelo**

Se deben restringir los parámetros de aquellas variables altamente colineales. Por ejemplo, si las variables  $x_2$  y  $x_3$  son altamente colineales es posible restringir el modelo utilizando información a priori o bien por estimaciones de corte transversal.

Suponga que nuestro modelo es:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \quad (13)$$

con  $i=1,2,\dots,n$

Al aplicar pruebas de detección de multicolinealidad se encontró que esta era grave y se debía a una elevada colinealidad entre  $x_2$  y  $x_3$ . Si, en publicaciones acerca de modelos similares al que se está estimando, existiera evidencia sobre los coeficientes se podría usar esa información para corregir. Por ejemplo, suponga que la evidencia encontrada es que el coeficiente  $\beta_3$  es un medio del coeficiente  $\beta_2$ . Esto nos permite aplicar la siguiente restricción:

$$\beta_3 = 0.5\beta_2 \quad (14)$$

Sustituyendo en el modelo original obtenemos la ecuación restringida:

$$y_i = \beta_1 + \beta_2 x_{2i} + 0.5\beta_2 x_{3i} + u_i \quad (15)$$

$$y_i = \beta_1 + \beta_2(x_{2i} - 0.5x_{3i}) + u_i \quad (15a)$$

$$y_i = \beta_1 + \beta_2 x_{2i}^* + u_i \quad (15b)$$

Donde:  $x_{2i}^* = x_{2i} - 0.5x_{3i}$

Una vez restringido el modelo la multicolinealidad se ha eliminado y al obtener el estimador de MCO  $\hat{\beta}_2$  es posible obtener  $\hat{\beta}_3$  si se sustituye el primero en la restricción (14).

La principal limitante de este método es la carencia de antecedentes empíricos acerca de los coeficientes de interés en los modelos econométricos.

Otra alternativa que implica restringir el modelo original es la estimación de un modelo en corte transversal. Por ejemplo, para el caso que nos ocupa se podría estimar  $\beta_3$  en un modelo de corte transversal y sustituir su valor estimado en el

modelo de series de tiempo. Suponga que en la estimación de corte transversal se obtiene que:

$$\hat{\beta}_3 = 0.5 \quad (16)$$

Se restringe el modelo sustituyendo ese valor en el modelo original:

$$y_i = \beta_1 + \beta_2 x_{2i} + 0.5 x_{3i} + u_i \quad (17)$$

$$y_i - 0.5 x_{3i} = \beta_1 + \beta_2 x_{2i} + u_i \quad (17a)$$

$$y_i^* = \beta_1 + \beta_2 x_{2i} + u_i \quad (17b)$$

Donde:  $y_i^* = y_i - 0.5 x_{3i}$

La limitante de este procedimiento es que la interpretación de los parámetros de corte transversal y series de tiempo puede diferir ampliamente al calcularse sobre conjuntos de datos diferentes.

### b) Componentes principales

El método de componentes principales busca eliminar el problema de multicolinealidad a través de la obtención de un conjunto de variables a partir de las originales y sin implicar grandes pérdidas de información (Everitt y Hothorn, 2006). Las nuevas variables o componentes cumplen con la condición de ser ortogonales entre sí.

El método parte de una forma cuadrática  $\mathbf{x}'\mathbf{A}\mathbf{x}$  que se minimiza sujeta a la condición de normalidad  $\mathbf{x}'\mathbf{x}=1$ :

$$\mathbf{x}'\mathbf{A}\mathbf{x} - \lambda(\mathbf{x}'\mathbf{x} - 1) \quad (18)$$

Donde  $\mathbf{A}$  es una matriz simétrica.

Al derivar con respecto a  $\mathbf{x}$ :

$$2Ax - 2\lambda x = 0 \quad (19)$$

Al factorizar encontramos la ecuación característica:

$$(A - \lambda I)x = 0 \quad (20)$$

Al obtener el determinante de la ecuación característica se genera un polinomio característico y al encontrar sus raíces nos permite obtener los valores característicos  $\lambda_i$ .

Si partimos de la matriz de regresores  $X$ , el método de componentes principales consiste en encontrar una función lineal de las variables originales,  $Z=a'x$ , que maximice la varianza de  $X$  sujeta a la condición de normalidad,  $a'a=1$ . Al resolver el polinomio podemos encontrar la raíz característica máxima y su correspondiente vector característico, el cual es el vector a que necesitamos para encontrar  $Z$ .

La principal limitante de este método es que las nuevas variables  $Z$  pueden no tener interpretación económica alguna.

### c) Eliminar variables

La eliminación de variables sospechosas de colinealidad puede ser otra opción para evitar el problema de multicolinealidad, sin embargo puede llevarnos a un problema más grave como el de variable relevante omitida. En nuestro ejemplo la eliminación de la variable  $x_{3i}$  deja el modelo como:

$$y_i = \beta_1 + \beta_2 x_{2i} + u_i \quad (21)$$

Sin embargo, si la variable omitida fuera relevante se genera un problema de sesgo en los estimadores de MCO.

### d) Transformar variables

La transformación de variables con primeras diferencias o calculando porcentajes es otro remedio que busca diferenciar más las variables entre sí. Sin embargo, su principal limitante es que, por una lado la teoría relevante pudiera estar interesada únicamente en las variables de nivel y no en sus diferencias ni en sus porcentajes,

por otro lado la variable dependiente pudiera estar relacionada con las demás en niveles pero no en porcentajes ni en diferencias.

## 5. UN EJEMPLO PRÁCTICO EN R DE SOLUCIÓN A LA MULTICOLINEALIDAD EN LA FUNCIÓN CONSUMO.

Una vía para buscar corregir cualquier síntoma de multicolinealidad en el modelo que hemos estimado para la función consumo podría ser el de componentes principales. Para lo cual podemos seguir el mismo procedimiento que ya hemos aplicado para calcular los valores característicos de la prueba del ICN, Es decir, se debe seleccionar en el menú principal la secuencia de opciones: STATISTICS/Dimensional analysis/Principal component analysis, pero ahora solamente consideraremos las dos variables que ya hemos confirmado antes guardan una elevada colinealidad entre sí, nos referimos a lrqr y lydr.

Los resultados que se muestran a continuación indican que la componente primera representa el 98.16% de la varianza total, por lo cual si tomamos esa componente para realizar la combinación lineal de los dos regresores prácticamente no habría perdida de información.

```
> .PC <- princomp(~lrqr+lydr, cor=TRUE, data=Dataset)
> unclass(loadings(.PC)) # component loadings
      Comp.1    Comp.2
lrqr  0.7071068 -0.7071068
lydr  0.7071068  0.7071068
> .PC$sd^2 # component variances
      Comp.1    Comp.2
1.96326036 0.03673964
> summary(.PC) # proportions of variance
Importance of components:
      Comp.1    Comp.2
Standard deviation 1.4011639 0.19167588
Proportion of Variance 0.9816302 0.01836982
Cumulative Proportion 0.9816302 1.00000000
```

Si ahora se corre la regresión sustituyendo los dos regresores por la combinación lineal de los mismos en la componente principal primera que ha sido guardada en la tabla de datos con el nombre PC1, es posible replantear el modelo de la ecuación (12) de la siguiente manera:

$$lcpr_t = \beta_1 + \beta_2 PC1_t + \beta_3 ltcr_t + u_t \quad (22)$$

Los resultados de esta regresión se muestran en seguida, de ellos se observa que la variable PC1 es estadísticamente significativa y que representa el efecto combinado de la riqueza y el ingreso en el consumo de los individuos.

```
m(formula = lcpr ~ ltcr + PC1, data = Dataset)
Residuals:
    Min      1Q      Median      3Q      Max 
-0.070837 -0.018441 -0.003601  0.020371  0.070356 
Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 13.686850  0.028197 485.397 <2e-16 ***
ltcr        -0.022570  0.021034  -1.073   0.286    
PC1         0.134105  0.002781   48.221 <2e-16 ***
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
Residual standard error: 0.03273 on 93 degrees of freedom
Multiple R-squared:  0.9721,    Adjusted R-squared:  0.9715 
F-statistic: 1619 on 2 and 93 DF,  p-value: < 2.2e-16
```

## REFERENCIAS

- L. R. Klein, An Introduction to Econometrics , Prentice-Hall, 1962;  
 Theil, H, Principles of Econometrics, Wiley, 1971.  
 Everitt,S. Brian y Torsten Hothorn, A handbook of statistical analysis using R, Chapman / Hall/CRC, 2006.  
 Quintana Romero, Luis y Miguel Ángel Mendoza, Econometría básica, Plaza y Valdés, 2008.

## **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

consumo\_fun.txt

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap6

Práctica\_Cap6

VideoPráctica\_Cap6

VideoTeoría\_Cap6

# CAPÍTULO 7: HETEROCEDASTICIDAD

Jorge Feregrino Feregrino

## 1. INTRODUCCIÓN

El origen de la heterocedasticidad, está asociado a la varianza creciente de las perturbaciones aleatorias de los valores de algunas de las variables, incluidas en el modelo. Dicho de otro modo, podría suponerse que la varianza de la perturbación se compone de una parte constante, homocedastica, y otra parte variable según los valores de una determinada variable. Es muy probable que esta asociación entre el proceso de heterocedasticidad y las variables no sea evidente.

La detección de la heterocedasticidad en la mayoría de los procedimientos es útil para establecer algún tipo de solución que permite corregir este problema.

Los efectos de la heterocedasticidad en los modelos de regresión lineal son los siguientes:

- a) Los estimadores del MCO son lineales insesgados y consistentes, pero en presencia de heterocedasticidad son ineficientes, ya que, la varianza no es la óptima. Cuando las perturbaciones son homocedasticas, la dispersión de los errores en el tiempo, no juega un papel relevante en el sesgo de los estimadores ni su consistencia.

- b) Las varianzas del estimador de Mínimos Cuadrados Ordinarios, no pueden calcularse con la expresión usual cuando se ha detectado heterocedasticidad}:

$$v(\beta) = \sigma^2(X'X)^{-1}$$

La expresión anterior es un estimador sesgado de la varianza de los parámetros; alternativamente, debe utilizarse la siguiente expresión

$$\text{covar} - \text{var}(\hat{\beta}) = \sigma^2[X'X]^{-1}X'\Sigma X[X'X]^{-1}$$

Cuando se realizan estimaciones bajo el supuesto de que las perturbaciones siguen un proceso homocedástico, mediante la aplicación de modelo de regresión con MCO, se cometerá un error de cálculo en la varianza, esto implica, básicamente, que nuestros cálculos sobre la “t” de student, ya no podrán comprarse con los valores de referencia correctos, y lo mismo ocurrirá con el resto de cálculos que tienen origen en la varianza estimada. Por ejemplo, el contraste “F” ya no se distribuirá como una “F” o los contrastes que utilizan como referencia a la j-cuadrada.

## 2. ESTRATEGIAS PARA REALIZAR ESTIMACIONES EN PRESENCIA DE HETEROCEDASTICIDAD

Al suponer en el modelo de regresión la presencia de heterocedasticidad, se puede realizar la estimación, pero debe tenerse cuenta, los problemas relacionados a la aplicación de los contrastes habituales sobre la significancia individual en las variables la t-student y la prueba de significancia conjunta de las

variables explicativas mediante la prueba F. En este sentido, la interpretación e inferencia sobre los valores esperados del modelo debe ser exigente, al ofrecer resultados menos concluyentes, pues los parámetros tenderán a ser amplios. Además, del error de cálculo en la estimación de la varianza de los parámetros, todos aquellos contrastes con base en este estimador serán también incorrectos. Un error frecuente, consiste en suponer un cálculo que implica la utilización de los errores de un modelo heterocedástico, será incorrecto cuando, en realidad, no es así. Cuando se realiza el cálculo de la  $R^2$  mediante una población, implica que no se utilizarán varianzas condicionales a los valores de las variables explicativas, de modo que el cálculo de la  $R^2$  no es afectada por la presencia de heterocedasticidad, de hecho la estimación mediante la siguiente expresión es adecuada en presencia de heterocedasticidad.

$$\tilde{\sigma}^2 = \frac{e'e}{n - k}$$

La presencia de heteroscedasticidad en los modelos de regresión lineal, rompe con una de las restricciones más importantes en la econometría, cuya hipótesis básica señala que la varianza de los errores aleatorios, condicional a los valores de la variable independiente X, es constante:

$$Var(v_t|x_t) = \sigma^2$$

La restricción implica que los valores muestrales de la variable dependiente (y) son iguales las varianzas de los errores ( $v$ ) para los distintos valores de ( $x$ ), es decir. la dispersión en relación a la minimización de los errores, permite

representar los valores de ( $Y$ ) mediante la variable estimada ( $\hat{y}$ ) de manera eficiente, insesgada y consistente.

El análisis de regresión condicional implica, obtener un parámetro estable y útil entre ambas variables, la dispersión entre las variables deben comportarse de la misma forma para evitar problemas de estimación e inferencia econométrica. En términos econométricos los errores de la estimación, no deben crecer a medida que lo hace el tamaño de la muestra de ( $x$ ), la dispersión de los errores en la estimación, debe mantenerse estable y no debe dispersarse en el tiempo.

Desde el punto de vista técnico, la matriz de varianzas en un modelo de regresión ante la presencia de heterocedastidad se representa así:

$$E(UU') = \begin{bmatrix} E(u_1)^2 & \vdots & \vdots \\ E(u_{12})^2 & E(u_2)^2 & \vdots \\ \vdots & \vdots & \vdots \\ E(u_{1n})^2 & E(u_{2n})^2 & E(u_n)^2 \end{bmatrix} = \begin{bmatrix} E(u_1)^2 & 0 & 0 \\ 0 & E(u_2)^2 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & E(u_n)^2 \end{bmatrix} \neq \sigma_i^2 I_n = \sigma^2 \Sigma$$

El estimador en el caso concreto de la presencia de una matriz de varianzas-covarianzas no escalar, donde las perturbaciones aleatorias de la matriz goza de buenas propiedades estadísticas, es lineal, insesgado, eficiente y consistente.

### **3. LAS CAUSAS DE LA HETEROCEDASTICIDAD**

La heterocedasticidad es resultado de la variabilidad de los fenómenos económicos, hay que identificar algunas situaciones específicas, asociadas al riesgo de aparición de este problema. Las causas más frecuentes para la presencia de la heterocedasticidad son las siguientes:

Omisión de las variables en la especificación del modelo: en la selección de las variables del modelo para explicar un fenómeno económico, suelen omitirse variables, ante la imposibilidad de controlar todos los determinantes del variable independiente. Esta restricción es controlada al incluir las perturbaciones aleatorias en el modelo, pero no se puede aseverar que los errores en todo momento cumplan la condición de homocedasticidad. La teoría econométrica, señala que la hipótesis de homocedasticidad se refiere a la varianza constante de las perturbaciones aleatorias, pero no obliga a que las variables explicativas tengan una varianza constante. La inclusión variables exógenas en la especificación del modelo cuya varianza crece en el tiempo, puede influir en la varianza de las perturbaciones y perder su condición de aleatoriedad.

[1] Cambio estructural: Un cambio de estructural puede provocar un ajuste erróneo de los parámetros en la estimación de los conjuntos muestrales. Este problema se reproduce solamente en algunas secciones de la muestra y puede generar diversos desajustes en el modelo, y por tanto, la varianza no constante en todo el período.

[2] Errores en la especificación de la forma funcional: la utilización de una forma funcional incorrecta, puede provocar que la calidad del ajuste de la regresión provoque cambios en las valores de las variables exógenas; es posible ajuste con errores crecientes y alta dispersión. Por ejemplo, la utilización de una función lineal en lugar de una logarítmica potencial, tasa de crecimiento porcentual o una función cuadrática

[3] Fallas en el supuesto de normalidad de las variables explicativas: en la realización del modelo cuando se incluyen variables explicativas cuya distribución no es normal y hay asimetrías en la distribución, los valores de los regresores estarán asociados a una mayor dispersión en las perturbaciones; además, la heteroscedasticidad se puede presentar en variables con un agrupamiento claro alrededor de la media.

[4] La presencia de valores atípicos en la muestra: esto implica desajustes en la varianza de las perturbaciones, por lo regular pertenecen a otro tipo de distribuciones y, por tanto, tienen una varianza diversa.

#### **4. CONTROL Y DETECCIÓN DE LA HETEROCEDASTICIDAD**

Realizar la estimación mediante Mínimos Cuadrados Generalizados, es una solución, aunque esta metodología exige estimar de antemano los valores de las

varianzas heterogéneas relacionadas a la muestra y a las variables explicativas.

Se deben hacer suposiciones simplificadas sobre la aparición eventual de la heterocedasticidad, estas deben permitir determinar la forma de la matriz de perturbaciones, pero un mal diseño no garantizará la eficiencia de la estimación.

Es imposible observar directamente la presencia de heterocedasticidad, en la mayoría de los análisis econométricos, solo se dispone de un par de valores para cada valor ( $y, x$ ), entonces, resulta conceptualmente imposible observar si la varianza de los errores "U" para cada valor de "X" es la misma en toda la muestra.

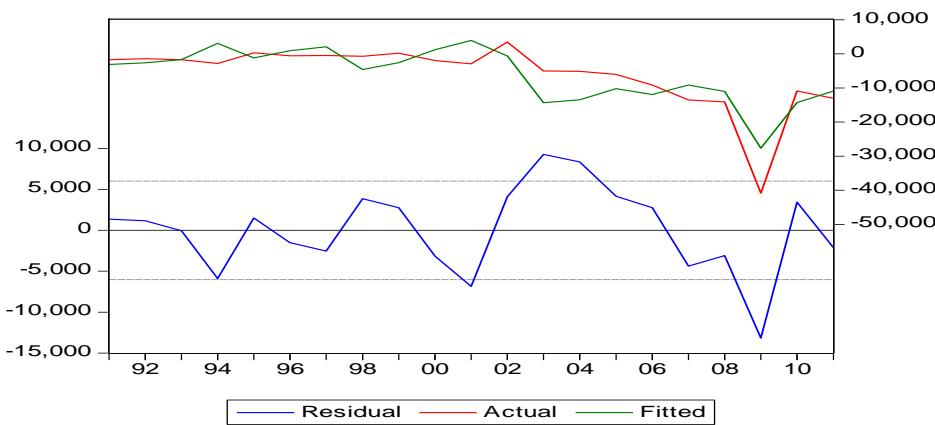
Por tanto, la mayor parte de los métodos se apoyarán en los residuos obtenidos en un modelo previo (estimado generalmente con MCO); estos residuos se utilizarán como una muestra válida de las perturbaciones aleatorias desconocidas.

Antes de cualquier utilizar cualquier método para detectar heterocedasticidad, debe haber un análisis previo de las variables exógenas incluidas, para tratar de identificar mediante análisis descriptivos y gráficos la naturaleza de los fenómenos económicos.

## Contrastes Gráficos

Graficar los errores, permitirá observar una tendencia definida para identificar intuitivamente en el transcurso del tiempo, si los errores crecen en el tiempo y si la varianza de estos errores es heterogénea, es decir, se presentarían mayores valores de los errores en el tiempo. En la siguiente gráfica, podemos observar la posible presencia de heterocedasticidad en los errores de la estimación. Los

errores comienzan a superar las bandas de dispersión a partir de 2002 y este proceso se acelera en 2008, es decir, la varianza crece a medida que el tiempo avanza.



La evolución en el tiempo esta correlacionada con valores de la serie cada vez mayores sobre todo a partir de la crisis de 2008, con lo que el cálculo de la varianza por subperíodos, por ejemplo: entre 1992 y 2000 arrojaría valores significativamente diferentes; es decir, el error estimado es heterocedástico. Evidentemente, este tipo de gráficos sólo tienen sentido si el modelo es temporal.

### Detección de la heterocedasticidad mediante contrastes paramétricos

Hay procedimientos que permiten cuantificar la heterocedasticidad, y valorar su existencia en términos de la probabilidad, recurriendo a distribuciones estadísticas conocidas, este tipo de contrastes se denominan: paramétricos. En este apartado

presentaremos los fundamentos teóricos de los contrastes usuales para la detección de heterocedasticidad en la estimación de los modelos.

### Contraste de Breusch-Pagan

La idea del contraste es comprobar si se puede encontrar un conjunto de variables, que permitan determinar la dinámica de la varianza de las perturbaciones, estimada a partir del cuadrado de los errores del modelo inicial. El proceso a seguir para llevar a cabo este contraste es el siguiente:

- [1] Estimar el modelo inicial, sobre el que se pretende saber si hay o no heterocedasticidad, empleando MCO y determinar los errores.
- [2] Calcular una serie con los errores del modelo anterior al cuadrado estandarizados:

$$\tilde{e}_i^2 = \frac{e_i^2}{\hat{\sigma}^2}$$
$$\hat{\sigma}^2 = \frac{e'e}{n}$$

- 3) Estimar una regresión sobre los determinantes de los errores mediante la incorporación de variables independientes ( $Z$ ), mediante las cuales se busca establecer si este conjunto de variables explican el proceso de heterocedasticidad de las perturbaciones en el modelo original; la estimación propuesta es la siguiente:

$$\tilde{e}_i^2 = \alpha_0 + \alpha_1 z_{1i} + \alpha_2 z_{2i} + \dots + \alpha_p z_{pi} + \varepsilon_t$$

- 4) El modelo es ineficiente si la varianza de la variable dependiente estimada y su error estimado es grande. Entonces, podría afirmarse que el poder explicativo del conjunto de variables  $Z$  sobre la representación de la varianza de las perturbaciones aleatorias es escaso. Mediante el diseño de un contraste calculado

con la sumatoria de los residuales de la estimación planteada en el paso 3, cuando este se encuentre cercano a cero, la probabilidad de que el proceso sea homocedástico es alta. El contraste propuesto sería el siguiente:

$$\frac{\sum \widehat{e}_i^2 * n}{2}$$

Breusch y Pagan, mostraron que el contraste se distribuye como una ji-cuadrada, cuando el proceso del modelo es homocedástico, al revisar el contraste tablas, se toman en cuenta las siguientes hipótesis:

$$H_0: \text{presencia de homocedasticidad}$$

$$H_a: \text{se acepta la presencia de heterocedasticidad}$$

Cuando la probabilidad de cometer el Error Tipo I, es muy alta no se puede rechazar la hipótesis nula, entonces, la varianza de los errores aleatorios es constante, por lo tanto, homocedásticos.

### **El Contraste de White para detectar heterocedasticidad**

El contraste White es considerado una prueba robusta al no requerir supuestos previos como, por ejemplo, la normalidad de las perturbaciones. De igual manera, no es necesario determinar a priori las variables explicativas que determinan heterocedasticidad.

El objetivo de esta prueba es determinar si las variables explicativas del modelo, pueden determinar la evolución de los errores al cuadrado. Es decir; si la dinámica

de las variables explicativas en relación a las varianzas y covarianzas es significativa para determinar el valor de la varianza muestral de los errores.

El proceso de estimación es el siguiente:

1. Estimar el modelo original por MCO, para obtener los errores en la estimación.
2. Estimar una regresión sobre los determinantes de los errores, con la incorporación de todas las variables incluidas en el estimación del primer modelo, estas elevados al cuadrado y sus combinaciones no repetidas:

$$e_i^2 = \alpha_0 + \alpha_1 x_{1i} + \cdots + \alpha_k x_{ki} + \cdots + \alpha_p z_{pi} + \varepsilon_t$$

## 5. EJEMPLO EN R

El ejemplo siguiente se sustenta en el trabajo de [http://ldc.usb.ve/~moises/estadistica/Ej\\_Regresion Lineal Multiple Zoritza.pdf](http://ldc.usb.ve/~moises/estadistica/Ej_Regresion Lineal Multiple Zoritza.pdf).

Un distribuidor de cervezas está analizando el sistema de entregas de su producto; en particular, está interesado en predecir el tiempo sugerido para servir a los detallistas. El ingeniero industrial a cargo del estudio ha sugerido que los factores que influyen sobre el tiempo de entrega son el número de cajas de cervezas y la máxima distancia que debe viajar el despachador.

El primer paso consiste en importar a R la base de datos en Excel, se deberá convertir el archivo en CSV delimitado por comas. Se le asignará el nombre “distribución”, a la columna de datos de la variable dependiente se le asignará el

nombre “Tiempo”, mientras que los nombres de las variables independientes quedarán de la siguiente forma:

X<sub>1</sub>: “Cajas”

X<sub>2</sub>: “Distancia”

El comando para importar los datos desde Excel es el siguiente:

```
distribución<-  
read.delim("c://Docs//NumCajas.csv",sep=",",header=T,stringsAsFactors=F)
```

Para estimar los coeficientes del modelo de regresión lineal múltiple se deberá utilizar el comando siguiente:

```
>lm(Tiempo ~ Cajas + Distancia ,data=distribución)
```

```
> lm(Tiempo~Cajas+Distancia,data=distribución)  
Call:  
lm(formula = Tiempo ~ Cajas + Distancia, data = distribución)  
Coefficients:  
(Intercept)      Cajas      Distancia  
     2.3112       0.8772       0.4559  
> |
```

El modelo de regresión lineal múltiple estimado es el siguiente:

$$Tiempo = b_0 + b_1 Cajas + b_2 Distancia + u_1$$

Para almacenar los datos del modelo, a fin de realizar las pruebas pertinentes más adelante, se asigna nombre a los resultados del mismo:

```
>modelo <- lm(Tiempo ~ Cajas + Distancia,data=distribución)
```

```
> modelo<-lm(Tiempo~Cajas+Distancia,data=distribución)
> modelo

Call:
lm(formula = Tiempo ~ Cajas + Distancia, data = distribución)

Coefficients:
(Intercept)      Cajas      Distancia
      2.3112       0.8772       0.4559

> |
```

Para realizar las pruebas de hipótesis sobre los parámetros, utilizamos el comando siguiente para obtener las características del modelo:

```
>summary(modelo)
```

```

> summary(modelo)

Call:
lm(formula = Tiempo ~ Cajas + Distancia, data = distribución)

Residuals:
    Min      1Q  Median      3Q     Max 
-9.2716 -0.5405  0.5212  1.4051  2.9381 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  2.3112     5.8573   0.395  0.70007    
Cajas        0.8772     0.1530   5.732 9.43e-05 ***  
Distancia    0.4559     0.1468   3.107  0.00908 **  
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 3.141 on 12 degrees of freedom
Multiple R-squared:  0.7368,    Adjusted R-squared:  0.6929 
F-statistic: 16.8 on 2 and 12 DF,  p-value: 0.0003325

```

También podemos obtener la matriz de covarianzas mediante el comando siguiente:

```

> vcov(modelo)
            (Intercept)      Cajas      Distancia
(Intercept) 34.3079952 -0.676432912 -0.766948764
Cajas       -0.6764329  0.023419588  0.009102869
Distancia   -0.7669488  0.009102869  0.021539183
> |

```

Para determinar el poder explicativo del modelo utilizamos el contraste F, al obtener la tabla ANOVA se pueden evaluar los resultados mediante el test F, tabla ANOVA se obtiene:

```

> anova(modelo)
Analysis of Variance Table

Response: Tiempo
            Df  Sum Sq Mean Sq F value    Pr(>F)    
Cajas        1 236.161 236.161 23.9403 0.0003704 ***
Distancia    1  95.198  95.198  9.6505 0.0090793 **  
Residuals   12 118.375   9.865
---
Signif. codes:  0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
> |

```

## Análisis de residuales mediante análisis gráfico

Las propiedades de los errores se pueden analizar mediante los comandos siguientes, primero para obtener el vector de residuales:

```
>residuales<-modelo$residuals
```

Para revisar los supuestos de normalidad de los residuos, utilizamos los comandos siguientes:

```
>rstint<-rstandard(modelo)
```

Obtiene los residuos estándares del modelo ajustado

Para la visualización de las gráficas se introducen los siguientes comandos:

```
>win.graph()
```

```
>par(mfrow=c(1,3))
```

Por último para observar los gráficos introducimos:

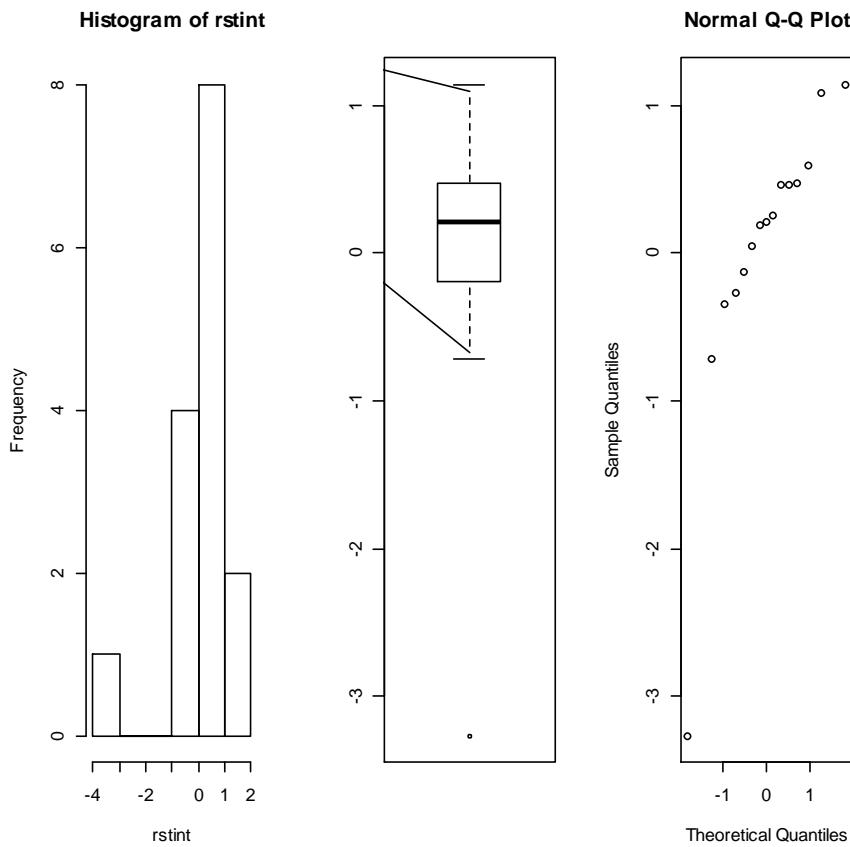
```
>hist(rstint)
```

```
>boxplot(rstint)
```

```
>qqnorm(rstint)
```

```
> rstint<-rstandard(modelo)
> win.graph()
> par(mfrow=c(1,3))
> hist(rstint)
> boxplot(rstint)
> qqnorm(rstint)
> |
```

Los gráficos a obtener son los siguientes:



## Pruebas de diagnóstico

Para detectar heterocedasticidad en el modelo procedemos a realizar los test de Breusch-Pagan y White, el procedimiento a seguir en R:

Para determinar la presencia de heterocedasticidad utilizamos el comando que devuelve el test de Breusch-Pagan:

```
>bptest(modelo)
```

Adicionalmente podemos utilizar el test White, el comando devolverá las matrices de covarianza corregidas para hacer inferencias. La matriz se obtiene mediante el comando:

```
>hccm(modelo)
```

Mientras que los resultados del test utilizando la matriz de covarianza se obtienen con el comando:

```
>coeftest(p,vcov=hccm(modelo))
```

Fuentes: <http://www2.kobe-u.ac.jp/~kawabat/ch08.pdf>

<http://www.r-bloggers.com/heteroscedasticity/>

## ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO

NumCajas.csv

## MATERIAL DE APRENDIZAJE EN LÍNEA

Teoría\_Cap7

Práctica\_Cap7

VideoPráctica\_Cap7

VideoTeoría\_Cap7

# CAPÍTULO 8: AUTOCORRELACIÓN SERIAL

Roldán Andrés-Rosales

## 1. INTRODUCCIÓN

La autocorrelación es un caso particular del modelo de regresión generalizado que se produce cuando las perturbaciones del modelo presentan correlaciones entre ellas. La autocorrelación supone que la matriz de varianzas y covarianzas de las perturbaciones presentan valores distintos de cero en los elementos que están fuera de la diagonal principal (Gujarati, 2004 Griffiths y Judge, 1993).

La autocorrelación puede definirse como “la correlación entre miembros de series de observaciones ordenadas en el tiempo (como en datos de series de tiempo) o en el espacio (como en datos de corte transversal)” (Gujarati, 2004:426); es decir,

$$E(e_i e_j) = 0, i \neq j$$

Planteamos la correlación como la relación existente entre la covarianza y la desviación estándar de x y y, matemáticamente la expresamos (como el cociente entre la covarianza de dos variables dividida entre la raíz cuadrada del producto de sus varianzas) de la siguiente forma:

$$\rho = \frac{\gamma_s}{\gamma_0}$$

donde

$$\gamma_s = E(e_t e_{t-1}); \text{ y } \gamma = E(e^2) = \sigma_e^2$$

Por lo regular, la autocorrelación está asociada a datos de series de tiempo y se define como la correlación existente entre los elementos de una serie de tiempo (Quintana y Mendoza, 2008). Esta autocorrelación puede ser generada por diversas circunstancias i) Errores de especificación como la omisión de variable(s) relevante(s), existencia de relaciones dinámicas no recogidas en el modelo o formulación de una relación funcional

lineal incorrecta;; ii) Existencia de efectos de proximidad entre las observaciones y iii) manipulación de la información<sup>21</sup>.

Las consecuencias de la autocorrelación son similares al de la heteroscedasticidad y son:

- i. El estimador de MCO es todavía lineal e insegado pero no es de mínima varianza y existe otro estimador lineal más eficiente.
- ii. Las varianzas y covarianzas de los estimadores MCO son sesgados.
- iii. Los intervalos de confianza y los estadísticos habituales para el contraste de la hipótesis no son adecuados
- iv. El estadístico  $R^2$  es sesgado.

Por estos motivos, el estimador de MCO deja de ser óptimo, eficiente y los contrastes usuales quedan invalidados. En estos casos, es posible encontrar otro estimador que recoja la información sobre las correlaciones entre las perturbaciones y que sea más eficiente: el estimador cae dentro de los estimadores de mínimos cuadrados generalizados MCG.

Un hecho importante que hay que tener en cuenta cuando detectamos la presencia de autocorrelación es la posibilidad de que dicho fenómeno sea generado por un error de especificación en el modelo más que por la verdadera existencia de correlaciones entre las perturbaciones. Intentar solucionar la presencia de autocorrelación en el modelo en el que existe alguno de los problemas mencionados conduciría, en la práctica, a un modelo en el que no se habrían eliminado los defectos de especificación ni, por supuesto, las consecuencias adversas que pueden originarse (Griffiths y Judge, 1993).

## 2. DETECCIÓN DE LA AUTOCORRELACIÓN

La hipótesis planteada son las siguientes:

$$H_0: \text{No autocorrelación serial}$$

---

<sup>21</sup> En términos económicos y aplicados a una realidad específica, podría implicar que el planteamiento teórico del modelo no se aplica para la economía que se esté probando.

$H_a$ : Autocorrelación serial

Para detectar la autocorrelación se pueden utilizar métodos gráficos y contrastes de hipótesis. Con frecuencia un examen visual de las perturbaciones nos permitirá conocer la presencia de la autocorrelación. Aunque es una forma subjetiva de probar la existencia de la autocorrelación, existen pruebas formales para detectarla.

### Contraste de Durbin-Watson

El contraste de Durbin-Watson es la prueba más conocida para detectar la existencia de la autocorrelación serial. Siguiendo a Griffiths y Judge (1993) y Quintana y Mendoza (2008) podemos expresarla de la siguiente forma:

$$d = \frac{\sum_{t=2}^n (\hat{e}_t - \hat{e}_{t-1})^2}{\sum_{t=1}^n \hat{e}_t^2} = \frac{\sum \hat{e}_t^2 + \sum \hat{e}_{t-1}^2 + 2 \sum \hat{e}_t \hat{e}_{t-1}}{\sum \hat{e}_t^2}$$

Para muestras grandes se puede considerar que las sumatorias de los residuales en el periodo t y en el t-1 son casi iguales por lo que Durbin-Watson sería como:

$$d \cong \frac{2 \sum \hat{e}_t^2 - 2 \sum \hat{e}_t^2 \hat{e}_{t-1}^2}{\sum \hat{e}_t^2} = 2(1 - \hat{\rho})$$

Con la fórmula podemos comprobar la existencia de la autocorrelación serial de primero orden:

si  $\hat{\rho} = -1 \therefore d \approx 4$  existe autocorrelación negativa

si  $\hat{\rho} = 0 \therefore d \approx 2$  no existe autocorrelación serial

si  $\hat{\rho} = 1 \therefore d \approx 0$  existe autocorrelación positiva

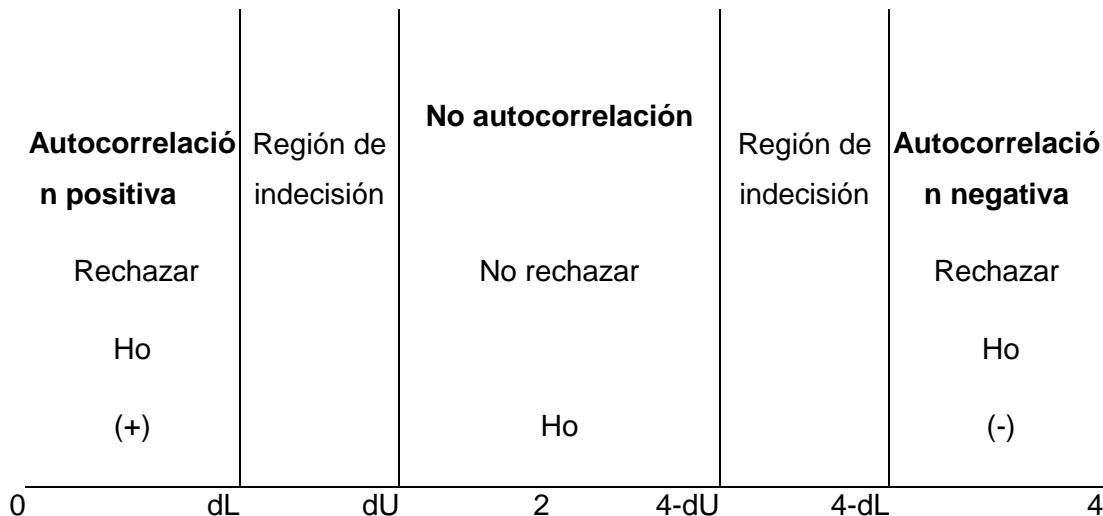
Gráficamente podemos expresar los criterios de rechazo y de indecisión de la hipótesis nula.

Si  $d < d_L$  existe evidencia de autocorrelación serial positiva

Si  $d > 4-dL$  existe evidencia de autocorrelación serial negativa

Si  $D_u < d < 4-dU$  no hay evidencia de autocorrelación

Si  $dL < d < dU$  o  $4-dU < d < 4-dL$  la prueba no es concluyente



A pesar de ser la prueba más conocida y más utilizada para detectar la autocorrelación, sólo permite detectar la autocorrelación serial de primer orden y carece de interpretación cuando incluimos rezagos dentro del modelo, además no permite obtener conclusiones en las regiones de indecisión.

### Prueba de Breusch-Godfrey (prueba LM)

La prueba de Breusch-Godfrey se desarrollaron para determinar si existe o no autocorrelación de orden superior a uno y, consiste en estimar una regresión auxiliar con MCO y hacer un contraste sobre los parámetros de la regresión. Como ejemplo supongamos que estimamos el siguiente modelo:

$$y_t = X_t B + e_t$$

La regresión auxiliar para el contraste de autocorrelación hasta de orden  $p$  en los residuos tiene la forma:

$$e_t = X_t \theta + \sum_{i=1}^p e_{t-i} + \nu_t$$

con estadísticos  $LM=T^*R^2$  y, en muestras grandes  $T \sim \infty$  por lo que  $LM \sim \chi^2(p)$

Las ventajas de la prueba Breusch-Godfrey son las siguientes; i) fáciles de implementar; ii) se puede generalizar para detectar autocorrelación de orden superior y iii) la distribución asintótica del estadístico LM para la prueba de autocorrelación hasta de orden p tiene una distribución de  $\chi^2(p)$

### **Procesos de la perturbación aleatoria:**

Una de las primeras dificultades que surgen cuando las perturbaciones están autocorrelacionadas es el conocimiento del tipo de relación que las une; es decir, la expresión que tienen los elementos de la matriz de varianzas y covarianzas de las perturbaciones.

Lamentablemente, en la mayoría de las ocasiones los elementos de esta matriz no son conocidos, por lo tanto, como la matriz  $\Omega$  no es conocida es necesario estimarla especificando la forma de correlación que siguen las perturbaciones mediante procesos que dependan de un conjunto reducido de los parámetros (Gujarati, 2004).

Los procesos estocásticos más utilizados para especificar las correlaciones entre las perturbaciones son los modelos autorregresivos (AR) y de medias móviles (MA), que incluyen como casos particulares los modelos autorregresivos de orden p, AR(p) y los modelos e medias móviles de orden q MA(q). Podemos expresarlos de la siguiente forma:

Un modelo AR(1) se escribe como

$$y_t = \phi_0 + \phi_1 y_{t-1} + e_t$$

Un modelo AR(2)

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + e_t$$

Un modelo AR(p)

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t$$

Un proceso de medias móviles de orden q, MA(q) se expresan como:

Un modelo MA(1) se escribe como

$$y_t = \theta_0 + \theta_1 e_{t-1} + e_t$$

Un modelo MA(2)

$$y_t = \theta_0 + \theta_1 e_{t-1} + \theta_2 e_{t-2} + e_t$$

Un modelo MA(q)

$$y_t = \theta_0 + \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t$$

Combinando ambos modelos tenemos los llamados “Modelos Autorregresivos y medias móviles” ARMA(p, q), que básicamente es la combinación de los AR y MA.

$$y_t = \phi_0 + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \cdots + \phi_p y_{t-p} + e_t \theta_1 e_{t-1} + \theta_2 e_{t-2} + \cdots + \theta_q e_{t-q} + e_t$$

Los instrumentos que se utilizan para caracterizar cada uno de estos procesos son las funciones de autocorrelación simple y parcial. Se pueden dar diferentes combinaciones como ARIMA, SARMA, SARIMA o SARIMAX, pero no es el objetivo de este capítulo por lo que no lo desarrollaremos.

### 3. PROCEDIMIENTO PARA LA DETECCIÓN DE LA AUTOCORRELACIÓN EN R-STUDIO

Utilizando la información de Quintana y Mendoza (2008) de la tasa de interés sobre la existencia del desplazamiento de la inversión pública en la inversión privada (crowding-out), planteamos el siguiente modelo:

$$TINTER_t = \beta_0 + \beta_1 SALDOPP_t + \beta_2 DEFPART_t + \beta_3 IEPPART_t + e_t$$

donde:

TINTER= tasa de interés real/ tasa de Cetes a 28 días-tasa de inflación

SALDOPP= Saldo de la balanza comercial como porcentaje del PIB

DEFPART= Participación porcentual del déficit presupuestal del PIB

IEPPART= Participación porcentual de la inversión extranjera en el PIB

Existen diversos procedimientos para detectar correlaciones entre las perturbaciones. Dado que éstas no son observables, las variables que se utilizan son los residuos mínimo cuadráticos. El gráfico de los residuos frente al tiempo, o frente a alguna variable y el gráfico de los residuos frente a sí mismos retardados un periodo.

Para poder trabajar en Rstudio debemos de instalar primero los paquetes que utilizaremos para la estimación y las pruebas de autocorrelación. De no hacerlo resultará imposible trabajar con el programa. Para instalar los paquetes podemos escribir los siguientes comandos:

```
install.packages("datasets")
```

```
library(datasets)
```

```
install.packages("Ecdat")
```

```
library(Ecdat)
```

```
install.packages("graphics")
```

```
library(graphics)
```

```
install.packages("lmtest")
```

```
library(lmtest)
```

```
install.packages("stats")
```

```
library(stats)
```

El siguiente paso es anexar la base que debe ser guardado en excel con extensión CSV (separados por comas). Una vez hecho esto debemos de especificar la ruta que tiene el archivo con el siguiente comando:

```
basecorre <- read.csv("/Users/Roldan/Documents/interes.csv")  
attach(basecorre)
```

para trabajar con datos de series de tiempo es importante definirlas como tal, cuando no existe este problema se pueden hacer las pruebas sin ningún inconveniente, pero tratándose de pruebas de series de tiempo como la autocorrelación serial, especificamos la base como series con el siguiente comando:

```
mst<-ts(basecorre, start=c(1980,1), end=c(1999,1), frequency=4)
```

en este caso nuestra información es trimestral por lo que usamos la instrucción anterior. Pero si la base es anual entonces se puede poner el siguiente comando:

```
mst<-ts(basecorre, start=c(1980), end=c(1999))
```

si estuviéramos trabajando con otro tipo de frecuencia en la información tendríamos que seguir el siguiente orden:

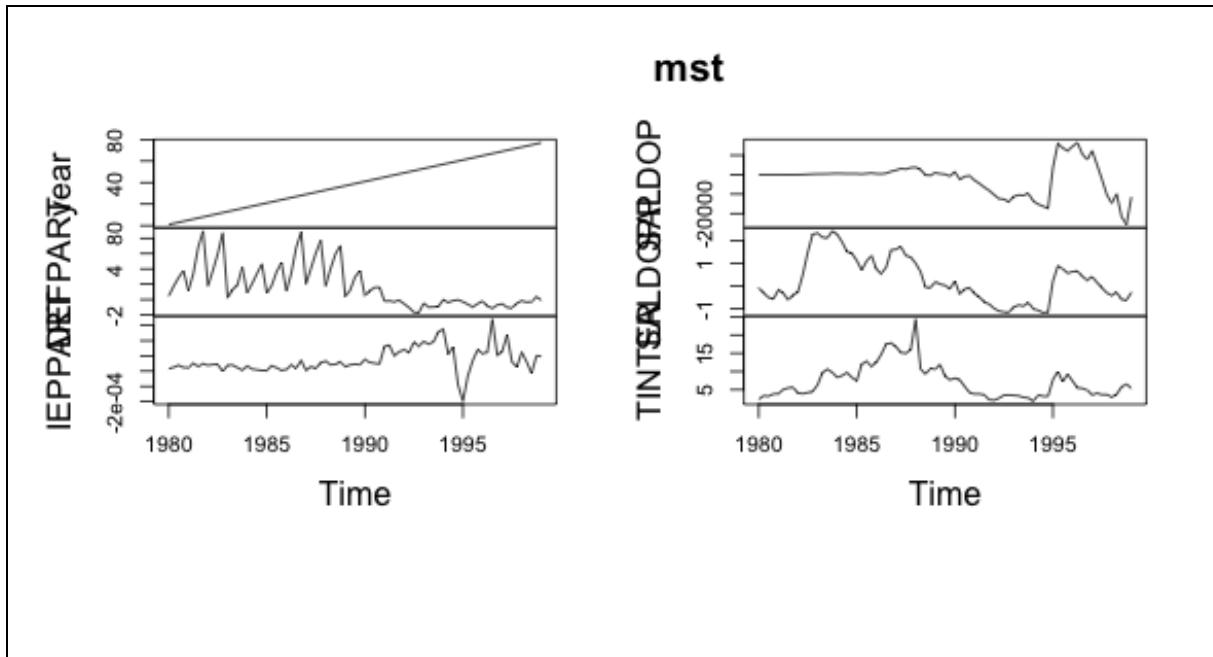
```
ts (nombre-de-la-base, start=c(year,period), end=c(year, period), frequency=) donde  
frequency es el numero de observaciones por unidad en el tiempo y pueden ser 1=anual,  
4=trimestral,12=mensual).
```

Para analizar la base y le pedimos un resumen y la gráfica ponemos los siguientes comandos:

```
> summary (basecorre)
```

year	DEFPART	IEPPART	SALDOP
1980Q1 : 1	Min. :-1.7321	Min. :-1.967e-04	Min. :-25662.96
1980Q2 : 1	1st Qu.:-0.3441	1st Qu.: 2.430e-05	1st Qu.: -6475.92
1980Q3 : 1	Median : 0.9106	Median : 4.560e-05	Median : 49.91
1980Q4 : 1	Mean : 1.7187	Mean : 7.032e-05	Mean : -1736.87
1981Q1 : 1	3rd Qu.: 3.4575	3rd Qu.: 1.158e-04	3rd Qu.: 791.99
1981Q2 : 1	Max. : 8.9393	Max. : 3.398e-04	Max. : 16442.15
(Other):71			
SALDOPP	TINTER		
Min. :-1.17516	Min. : 1.874		
1st Qu.:-0.47966	1st Qu.: 3.638		
Median : 0.02984	Median : 5.637		
Mean : 0.27129	Mean : 7.227		
3rd Qu.: 0.89231	3rd Qu.: 9.684		
Max. : 2.42928	Max. :24.281		

```
>plot(mst)
```



para realizar la regresión por MCO usamos el comando usado en el ejemplo y nos despliega el resultado:

```
> mcor<-lm(TINTER~SALDOPP+DEFPART+IEPPART)
```

```
> summary (mcor)
```

Call:

```
lm(formula = TINTER~SALDOPP+DEFPART+IEPPART)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.2625	-1.7087	-0.4814	1.5587	15.4154

Coefficients:

```

Estimate Std. Error t value Pr(>|t|)

(Intercept) 6.2170 0.6988 8.897 2.91e-13 ***
SALDOPP 2.1562 0.4660 4.627 1.57e-05 ***
DEFPART 0.4167 0.1684 2.474 0.0157 *
IE PPART -4135.2476 5680.6435 -0.728 0.4690

---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

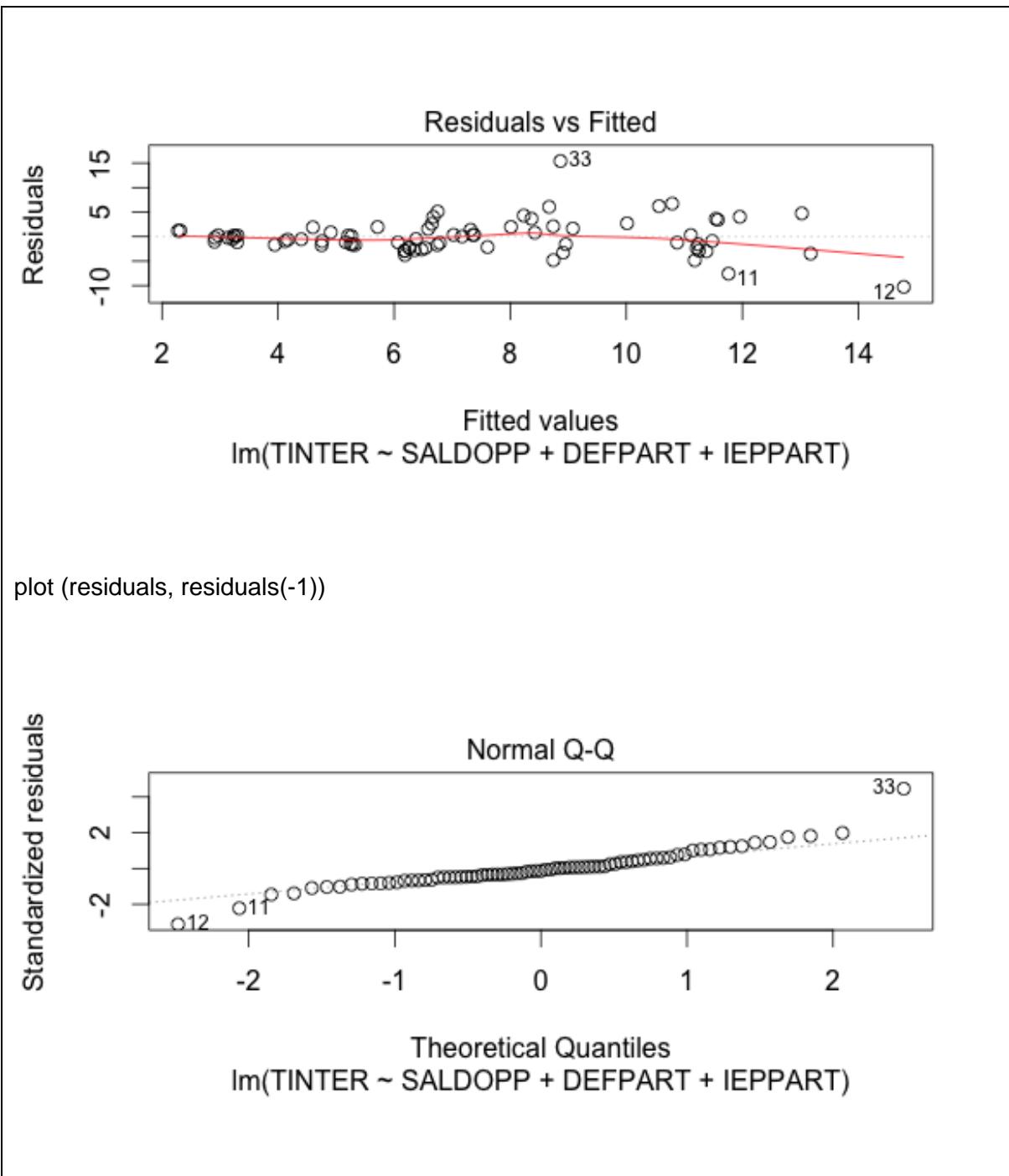
Residual standard error: 3.494 on 73 degrees of freedom
Multiple R-squared: 0.4354, Adjusted R-squared: 0.4122
F-statistic: 18.76 on 3 and 73 DF, p-value: 4.021e-09

```

### Prueba de autocorrelación utilizando gráficas

Para analizar gráficamente la tendencia de la autocorrelación podemos hacerlo con el comando siguiente:

```
plot(mcorm)
```



### Prueba de Durbin Watson

Para la prueba de Durbin Watson para detectar la autocorrelación serial usamos el comando del cuadro y se presenta el resultado de nuestro ejercicio:

```
dwtest(mcor)

Durbin-Watson test

data: mcor
DW = 0.6168, p-value = 1.008e-13
alternative hypothesis: true autocorrelation is greater than 0

bgtest(mcor)

LM test = 37.6807, df = 1, p-value = 8.332e-10
```

Podemos ver que rechazamos la hipótesis nula de no autocorrelación dado que el p-value es menor al 0.05 utilizado. De ahí que concluimos que el modelo presenta al menos autocorrelación de primer orden.

### Prueba de Breusch-Godfrey

```
bgtest(mcor)

LM test = 37.6807, df = 1, p-value = 8.332e-10
```

De la misma forma, podemos ver que tenemos autocorrelación de primer orden con la prueba LM.

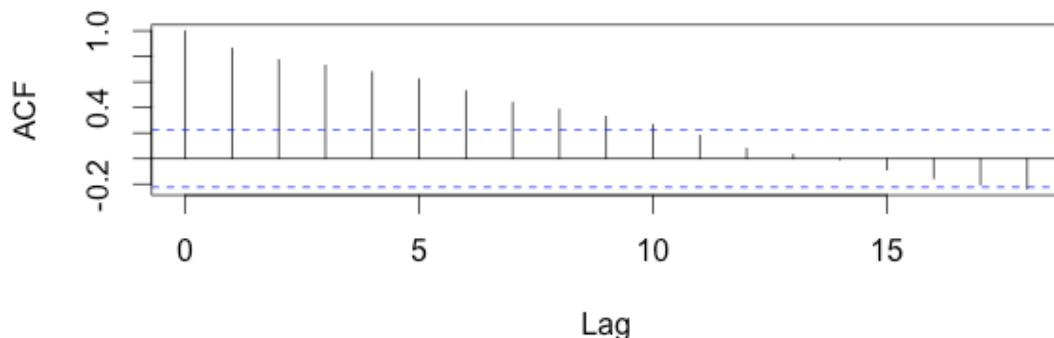
### Correlogramas

Los correlogramas de las funciones de autocorrelación simple y parcial de los residuos.

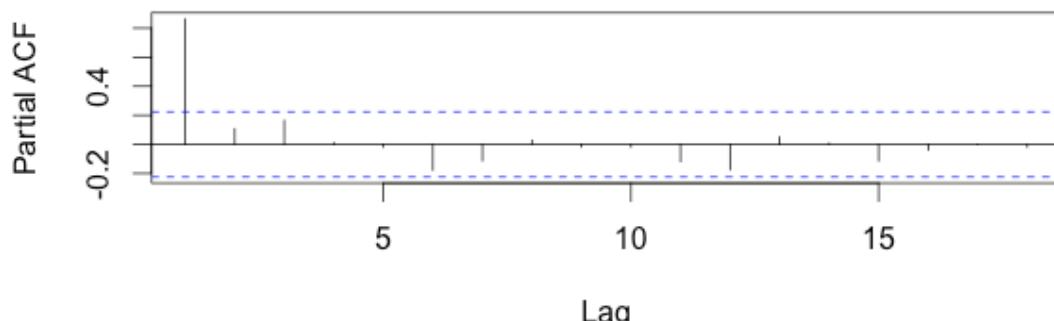
```
>acf (TINTER)
```

```
>pacf(TINTER)
```

**Series TINTER**



**Series TINTER**



Observamos diversos coeficientes de autocorrelación simple fuera de las bandas con un comportamiento de decrecimiento positivo en los primeros coeficientes, mientras que en la función de autocorrelación parcial solamente está fuera de las bandas el primer rezago, este comportamiento podría llevarnos a pensar que se trata de la existencia AR(1).

### Corrección del modelo

Enfrentar el problema de autocorrelación y/o heteroscedasticidad requiere antes de cualquier otra cosa verificar que el modelo no presente ningún error de especificación ya que un modelo mal especificado es una fuente de este tipo de problemas. Si se comprobó

todo lo anterior y el modelo sigue presentando autocorrelación y/o heteroscedasticidad es posible enfrentar el problema con mínimos cuadrados generalizados (MCG). El problema que generalmente se presenta en el modelo de regresión generalizado es que la matriz de varianzas y covarianzas de las perturbaciones es desconocida  $E[e\bar{e}] = \sigma^2 \Omega$ , por lo tanto, se hace necesaria su estimación. Por este motivo, al realizar la estimación obtenemos lo que se conoce como estimadores por mínimos cuadrados generalizados factibles (MCGF).

Si corregimos utilizando el método de Cochrane-Orcutt para agregar un AR(1) en el modelo, donde instalamos primero el paquete orcutt. El comando y resultado es el siguiente:

```
install.packages("orcutt")
library(orcutt)
mcor<-lm(TINTER~SALDOPP+DEFPART+IEPPART)
mcor1<-cochrane.orcutt(mcor)
mcor1;
```

\$Cochrane.Orcutt

Call:

```
lm(formula = YB ~ XB - 1)
```

Residuals:

Min	1Q	Median	3Q	Max
-10.2189	-0.9429	-0.2819	0.9813	9.2605

Coefficients:

Estimate Std. Error t value Pr(>|t|)

XB(Intercept) 7.3582 1.6122 4.564 2.02e-05 \*\*\*

XBSALDOPP 1.2590 0.7163 1.758 0.083 .

XBDEFFPART -0.1122 0.1171 -0.958 0.341

XBIEPPART -71.1856 3900.4015 -0.018 0.985

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 2.231 on 72 degrees of freedom

Multiple R-squared: 0.2718, Adjusted R-squared: 0.2314

F-statistic: 6.72 on 4 and 72 DF, p-value: 0.0001183

\$rho

[1] 0.8358428

\$number.interaction

[1] 12

En este modelo podemos observar podemos aplicar las pruebas de autocorrelación de LM y ver que no tenemos más el problema.

## **REFERENCIAS**

- Griffiths, W. Carter, R. y Judge, G, (1993), *Learning and practicing econometrics*, Wiley.
- Quintana, Luis y Miguel A. Mendoza (2008), *Econometría básica modelos y aplicaciones a la economía mexicana*, Plaza y Valdes Editores, México.
- Guajratí, Damodar, *Principios de econometría*, McGraw Hill, 2009.

## **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

interés.csv

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap8

Práctica\_Cap8

VideoPráctica\_Cap8

VideoTeoría\_Cap8

# **CAPITULO 9: ANALISIS DE INTEGRACION: APLICACIONES EN SOFTWARE R**

**Miguel Ángel Mendoza González y Luis Quintana Romero.**

## **1. INTRODUCCION**

El análisis de integración es parte de la metodología básica de la econometría moderna. La metodología establece que los datos que se observan de cualquier fenómeno económico, social o ambiental provienen de una variable aleatoria que se define por un Proceso Generador de Información (PGI) con un modelo estadístico, probabilístico y muestral. La metodología econométrica moderna establece que cualquier indicador o variable se use en el análisis, siempre y cuando cumpla con las tres condiciones para ser aleatoria: 1) Media finita y constante con respecto al tiempo; 2) Varianza finita y constante respecto al tiempo; y, 3) Covarianza finita, constante con respecto al tiempo, pero que dependa del tiempo en la definición de proceso autorregresivo. A estas tres características también se les conoce como las condiciones de estacionariedad, debido a que para su cumplimientos el equilibrio debe existir y su dinámica debe ser convergente. En el caso de que no se cumpla, se establece que la principal causa es debido a una media no constante y por el equilibrio no es convergente bajo dos situaciones: 1) no existe; o, 2) existe pero no es estable. Al primer caso se le conoce como el problema de existencia del equilibrio por una raíz característica unitaria y al segundo como el problema de divergencia por una mayor que uno. Por tanto, el análisis de integración consiste en analizar si los indicadores o variables de interés cumplen o no con ser aleatoria, con la ayuda de los conceptos de estacionario y las pruebas de raíz unitaria. En el caso de que no se cumpla con ser estacionaria, la metodología establece utilizar transformaciones por medio de la eliminación de tendencias deterministas o estocásticas.

## **2. ANALISIS DE INTEGRACIÓN**

La metodología del análisis de integración se entiende como el procedimiento para entender si un indicador o variable cumple con ser aleatoria o estacionaria. La principal causa de porque un indicador o variable no es estacionaria es que sigue una tendencia

lineal, cuadrática o exponencial del tipo determinística o estocástica. El procedimiento establece que en el caso de que no se cumpla con la condición de ser estacionaria, entonces se elimine la tendencia por medio del método de la regresión o las diferencias. Este último método es el más utilizado debido que la variable transformada se parece más a un proceso aleatorio en varianza y covarianza.

Si al cumplimiento de estacionariedad del indicador o variable ( $y_{i,t}$ ) se identifica por  $I(0)$ , que se escribe como  $y_t \sim I(0)$  y se lee como orden de integración, entonces cuando no se cumple con tal condición se escribe como  $I(d)$ ,  $d > 0$ . En el caso de que la variable tenga un orden de integración igual a uno  $I(1)$ ,  $y_t \sim I(1)$ , significa que al indicador se le aplicó una transformación del tipo  $\Delta y_t$  o  $\Delta \ln y_t$ , para eliminar tendencia lineal o exponencial respectivamente; cuando es de un orden de integración 2,  $y_t \sim I(2)$ , la transformación supone una tendencia cuadrática o cuadrática exponencial  $\Delta^2 y_t$  o  $\Delta^2 \ln y_t$ ; y así, hasta que se requiera aplicar  $d$  diferencias para que la variable transformada  $\Delta^d y_t \sim I(0)$  o  $\Delta^d \ln y_t \sim I(0)$  cumpla con ser estacionaria.

Los métodos más utilizados para identificar el orden de integración de una variable, son los que prueban la existencia o no de raíces unitarias (Unit Root). En la actualidad existen un conjunto de pruebas que tienen variantes alrededor de la prueba básica Dickey-Fuller aumentada (ADF).

## 2.1 Estacionariedad

Para comprender el concepto de estacionariedad procederemos al análisis de las características que presentan series económicas generadas por un proceso estocástico no estacionario.

Muchas series económicas, en particular los precios de las acciones, tienen un comportamiento propio de los procesos de camino aleatorio debido a que no hay posibilidades de arbitraje y por consiguiente el precio actual es igual al precio anterior más un error impredecible.

En el supuesto que nuestra variable  $y_t$  siga un proceso autorregresivo de primer orden AR(1), es decir que su valor actual depende de su valor anterior más un término de perturbación aleatoria:

$$[8.1] \quad y_t = \phi y_{t-1} + u_t, \text{ donde } u_t \sim RB(0, \sigma_u^2)$$

Donde  $u_t$  se supone sigue una distribución normal, con media cero y varianza constante, conocida como ruido blanco:

El coeficiente  $\phi$  indica la trayectoria que sigue la variable  $y_t$  en el tiempo. Si resolvemos recursivamente la ecuación en diferencias tendremos el siguiente resultado:

$$y_t = \phi^t y_0 + \sum_{t=0}^{t-1} u_{t-1}$$

Claramente puede observar que si el coeficiente  $\phi$  es menor a cero nuestra variable oscilará de signo, si es mayor a la unidad provocará un comportamiento explosivo sin límite. En ese sentido lo deseable es que se encuentre en el rango  $0 < \phi \leq 1$ . Si tomamos el caso particular de que sea igual a la unidad tendremos un proceso muy útil para exemplificar una serie no estacionaria y que se conoce como camino aleatoria:

$$y_t = y_{t-1} + u_t$$

El proceso se puede ver como una ecuación en diferencias de primer orden, si la resolvemos de manera recursiva obtenemos la siguiente secuencia:

$$y_1 = y_0 + u_1$$

$$y_2 = y_2 + u_2 = (y_0 + u_1) + u_2$$

.....

$$y_t = y_0 + \sum_{t=0}^t u_{t-1}$$

Es decir, el valor actual de la serie es igual a su valor inicial más todos los choques aleatorios desde que comenzó el proceso. Esto significa que un camino aleatorio es una serie de memoria larga. Cuando se aplica la esperanza para obtener su media, está resulta constante e igual al valor inicial de la serie:

$$E(y_t) = y_0$$

Mientras que su varianza es una función del tiempo:

$$\sigma^2(1) = \sigma_u^2$$

$$\sigma^2(2) = \sigma_u^2 + \sigma_u^2 = 2\sigma_u^2$$

.....

$$\sigma^2(t) = \sigma_u^2 + \sigma_u^2 + \dots + \sigma_u^2 = t\sigma_u^2$$

Por consiguiente se puede decir que la serie no es estacionaria en varianza.

## **2.2 Pruebas de raíces unitarias**

El análisis de raíces unitarias se puede derivar, si se considera que nuestra variable sigue un el modelo AR(1):

$$y_t = \phi y_{t-1} + u_t$$

Con base al operador de rezagos que se aprendió anteriormente, es posible simplificar el modelo como sigue:

$$y_t - \phi y_{t-1} = u_t$$

$$y_t - \phi L y_t = u_t$$

$$(1 - \phi L) y_t = u_t$$

Tenemos entonces un término  $(1-\phi L)$  que es el polinomio de grado uno asociado al proceso autorregresivo de orden 1, que implica una solución homogénea que se resuelve suponiendo que la solución para la variable es igual a  $y_t = \lambda^t$ , donde  $\lambda$  es la raíz característica :

$$\lambda^t - \phi L \lambda^t = 0$$

$$\lambda^t - \phi \lambda^{t-1} = 0$$

$$\lambda^{t-1}(\lambda - \phi) = 0$$

De esto, se puede concluir que  $\lambda = \phi$  y por tanto el parámetro es la raíz característica. Con la ecuación en diferencias que se resolvió recursivamente, se pueden analizar tres posibilidades: 1) Que la raíz característica es menor que uno  $\lambda < 1$  por tanto  $\phi < 1$ , lo cual implica que al aplicar  $\lim_{t \rightarrow \infty} \phi^t = 0$  y la esperanza del proceso, se encuentre que  $E(y_t) = 1/(1 - \phi)$ ; cuando la raíz característica es unitaria o mayor que uno  $\lambda = \phi \geq 1$ , implica que al aplicar el  $\lim_{t \rightarrow \infty} \phi^t = \infty$  y por tanto la esperanza del proceso sea igual a infinito  $E(y_t) = \infty$ .

Las pruebas más ampliamente difundidas para indagar acerca de la existencia de raíces unitarias se deben al trabajo de Dickey y Fuller (1979) y se conocen simplemente

como pruebas ADF. Para aplicar dichas pruebas vamos a partir de nuestro modelo de camino aleatorio puro, pero le restamos el término autorregresivo de los dos lados:

$$y_t - y_{t-1} = \phi y_{t-1} - y_{t-1} + u_t$$

$$\Delta y_t = \gamma y_{t-1} + u_t$$

Donde:

$$\Delta y_t = y_t - y_{t-1}$$

$$\gamma(\phi - 1)$$

De esta manera con una raíz unitaria  $\phi=1$ , el parámetro  $\gamma=0$ . Del modelo propuesto resulta entonces tentador efectuar la prueba de raíz unitaria aplicándole mínimos cuadrados ordinarios al modelo en primeras diferencias y sin constante. El coeficiente estimado del término autorregresivo se podría someter a la prueba usual de significancia estadística t para contrastar la hipótesis nula de raíz unitaria contra la alternativa de estacionariedad:

$$H_0: \gamma=0 \text{ por consiguiente } \phi=1$$

$$H_A: \gamma<0 \text{ por consiguiente } \phi<1$$

A pesar de lo atractivo de esta forma de efectuar la prueba, Dickey y Fuller mostraron que las pruebas t usuales no son adecuadas, ya que el estadístico de prueba no sigue una distribución normal. Los autores obtuvieron los valores críticos a través de simulaciones, denominados tau,  $\tau$ , y encontraron que dependían del tamaño de muestra a utilizar.

La prueba esbozada aquí nos permite afirmar que, de no existir evidencia en contra de la hipótesis nula, la serie no será estacionaria y seguirá una caminata aleatoria

pura. Para dar lugar a la posibilidad de probar si el modelo exhibe tendencia determinística y deriva, los autores propusieron las siguientes tres formas funcionales para efectuar la prueba:

- |   |  |
|---|--|
| a) $\Delta y_t = \gamma y_{t-1} + u_t$                    | Camino aleatorio puro  |
| b) $\Delta y_t = \alpha + \gamma y_{t-1} + u_t$           | Camino aleatorio con constante   |
| c) $\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + u_t$ | Camino aleatorio con constante, tendencia estocástica y determinística |

Al utilizar las pruebas *tau* de DF, los valores críticos son afectados por la inclusión o no de tendencia y constante. Por esta razón Dickey.Fuller desarrollaron tablas pertinentes para cada caso. En los tres casos el parámetro de interés, a partir del cual se realizan las pruebas, es el coeficiente  $\gamma$ . La hipótesis nula en el primer caso es la de camino aleatorio y se contrasta contra la hipótesis alternativa de proceso AR(1) estacionario con media nula. En el segundo caso la hipótesis nula es de camino aleatoria con constante contrastada con la hipótesis de proceso AR(1) estacionario sin tendencia. Finalmente el tercer modelo permite probar la hipótesis nula de camino aleatoria con la hipótesis alternativa de proceso AR(1) estacionario con tendencia determinística.

Si se busca realizar una prueba para el conjunto de parámetros en cada modelo, en lugar de la tau se utiliza una F que se construye a partir de los modelos con y sin restricciones; a estas pruebas se les conoce como  $\phi_1$ ,  $\phi_2$  y  $\phi_3$ . Para la prueba  $\phi_1$  la hipótesis nula es  $\gamma=0$ ; para  $\phi_2$  es  $\alpha=\gamma=0$ ; y para  $\phi_3$  podemos tener dos casos  $\alpha=\beta=\gamma=0$  o bien  $\beta=\gamma=0$ .

El modelo que empleamos para ilustrar las pruebas DF supone que los datos siguen un proceso AR(1), sin embargo en la práctica muchas series pueden no ajustarse a ello. Si suponemos que los datos siguen un proceso más general como un AR(p) tendremos:

$$y_t = \alpha + \phi_1 y_{t-1} + \phi_2 y_{t-2} + \dots + \phi_p y_{t-p} + \nu_t$$

Y, si ahora en la práctica modelamos con un proceso AR(1), los residuales de este último serán:

$$u_t = y_{t-2} + \dots + \phi_p y_{t-p} + \nu_t$$

Razón por la cual existirá autocorrelación en los residuales  $u_t$  y  $u_{t-m}$ , con  $m > 1$ . Problema que, como veremos mas adelante, entre otras cosas lleva a invalidar la inferencia estadística que se podría realizar con el modelo. Para considerar la posibilidad de autocorrelación Dickey-Fuller desarrollaron una prueba DF aumentada conocida como ADF y que utiliza el siguiente modelo:

$$\Delta y_t = \alpha + \beta t + \gamma y_{t-1} + \sum_{i=1}^p \lambda_i \Delta y_{t-i} + u_t$$

Bajo esta nueva regresión es necesario determinar el orden de rezagos que serán considerados en el modelo, pues también de ellos dependerán los valores críticos para efectuar la prueba. Para encontrar la estructura adecuada de rezagos se utilizan los criterios de información de Akaike y Schwarz revisados anteriormente.

El modelo anida las diferentes alternativas de prueba vistas antes, es decir se puede efectuar la prueba de raíz unitaria con o sin constante, con o sin tendencia determinística y con o sin considerar autocorrelaciones. Esta generalidad y amplitud de las posibilidades de prueba puede llevar a que, en algunos casos, se busque obligar a que la prueba otorgue evidencia favorable a alguna intención a priori de quién la está efectuando, para evitar falsear la prueba es conveniente seguir alguna estrategia de prueba.

- i) Graficar los datos. Si la serie original presenta tendencia, se debe incluir tendencia e intercepto.
- ii) Si no parece tener tendencia y su media no es cero, sólo incluir intercepto.
- iii) Si parece fluctuar en torno a su valor medio cero, no incluir ni tendencia e intercepto.

### 3. APLICACIONES EN R

#### **Ejemplo 1. Prueba ADF para análisis de integración del PIB de la economía mexicana**

En este ejemplo se utiliza la librería **urca** para análisis de integración y cointegración para series de tiempo escrita por Bernhard Pfaff y Matthieu Stigler.

```
#Cargar la librería urca
> library(urca)

#Cambiar el directorio de trabajo
>setwd("/Volumes/LACIESHARE/Academico/LibroEconometria_R/Capitulo_8/BaseDatos_Capitulo
8")

# Lectura de la base de datos
> load("BDatos_Integracion.RData")
> summary(BDatos)

  Periodo    PIB_Mex
  1993/01: 1  Min. :7817381
  1993/02: 1  1st Qu.:9506342
  1993/03: 1  Median :10637808
  1993/04: 1  Mean   :107777851
  1994/01: 1  3rd Qu.:12200173
  1994/02: 1  Max.   :13937805
  (Other):81
```

```

# Se asigna el logaritmo de la variable de serie de tiempo al objeto PIB_MEX y se aplica la primera
y segunda diferencia

> lplib_mex <- log(BDatos$PIB_Mex)

> d_lplib_mex <- diff(lplib_mex)

> d_lplib_mex 2<- diff(lplib_mex,2)

# Se asigna la variable que orden en el tiempo

> periodo <- BDatos$Periodo

# Para analizar el comportamiento del PIB de México se graficar la variable del PIB_Mex en el
tiempo

> plot(periodo, lplib_mex, main="Producto Interno Bruto de Mexico")

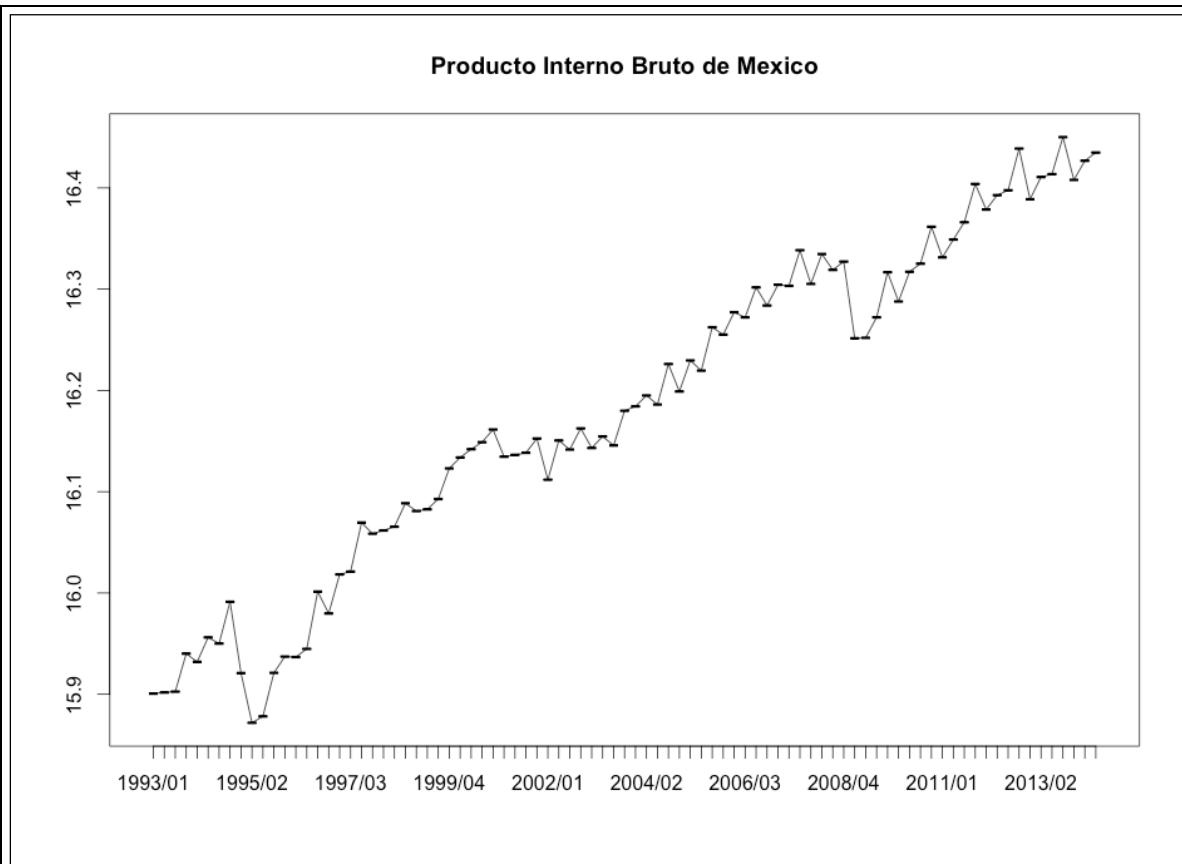
> lines(lplib_mex)

```

En la gráfica 8.1 se observa el comportamiento del PIB trimestral de México, donde destaca la presencia de una marcada tendencia positiva, con algunos caídas importantes en los momentos de crisis económicas en 1995, 2001 y 2009.

Grafica 1: Comportamiento del logaritmo del PIB de México

Serie trimestral para el periodo 1993-2013



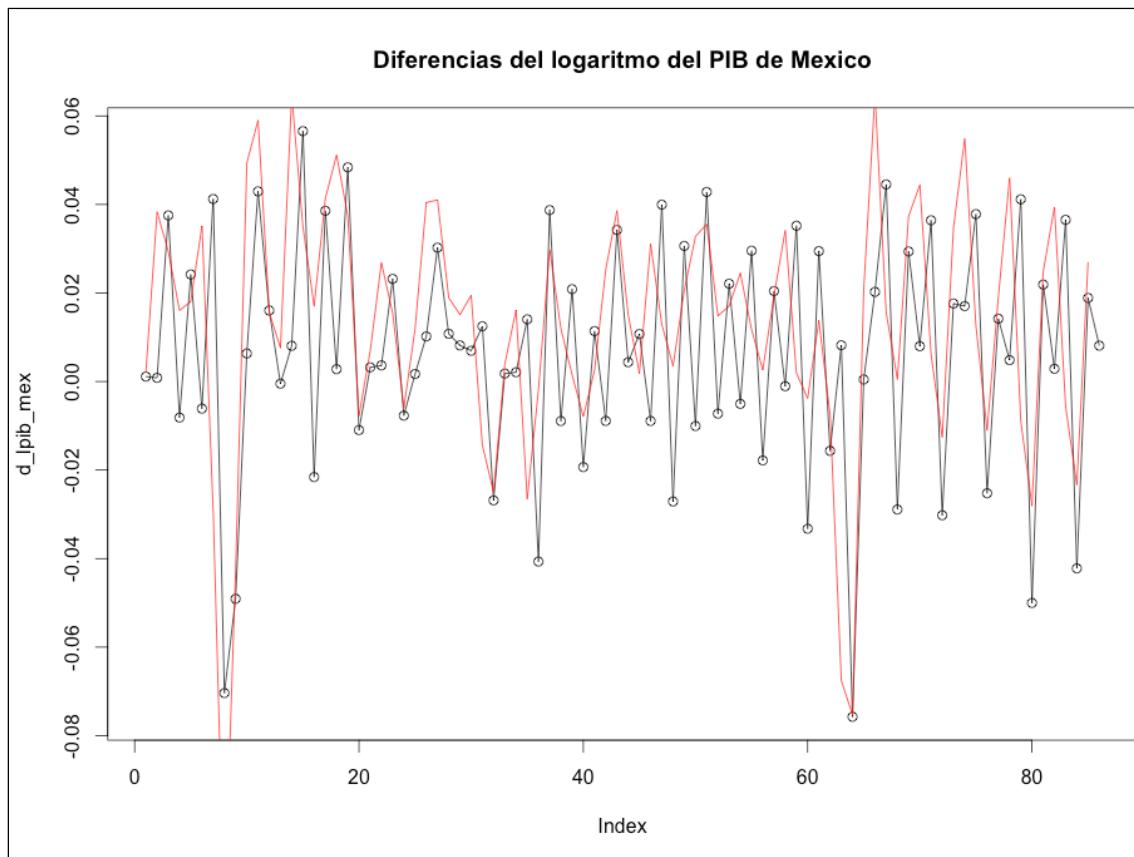
Para construir un gráfica con las primera y segunda diferencia del PIB

```
> plot(d_lpib_mex, main="Diferencias del logaritmo del PIB de México")
> lines(d_lpib_mex, col="black")
> lines(d_lpib_mex2, col="red")
```

De la gráfica 2 se puede observar que con la primera diferencia se estabiliza la media del logaritmo del PIB, con lo cual se infiere que es estacionaria en media. Para revisar si se puede mejorar la estabilidad de la media y la varianza, se aplica la segunda diferencia y la gráfica muestra que no se mejora la estabilidad y por tanto es suficiente aplicar la primera diferencia.

Grafica 2: Comportamiento de la primera y segunda del logaritmo del PIB de México

Serie trimestral para el periodo 1993-2013



Nota: La primera diferencia se representa con la línea negra y la segunda diferencia con rojo

# A continuación se aplica la prueba ADF para establecer si el logaritmo del PIB tiene raíz unitaria y de que tipo:

# Prueba de ADF

```
lc.df <- ur.df(y=PIB_Mex, type='trend',lags=4, selectlags=c("AIC"))
summary(lc.df)
```

```
#####
#
```

```
# Augmented Dickey-Fuller Test Unit Root Test #
```

```
#####
#
```

Test regression trend

Call:

```
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.074714	-0.006856	0.003727	0.012303	0.031288

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	4.7110519	1.2609435	3.736	0.000362 ***
z.lag.1	-0.2961001	0.0793584	-3.731	0.000368 ***
tt	0.0018680	0.0005108	3.657	0.000471 ***
z.diff.lag1	0.1110790	0.1102565	1.007	0.316953
z.diff.lag2	0.1289933	0.0982206	1.313	0.193085
z.diff.lag3	-0.1473676	0.0981777	-1.501	0.137546
z.diff.lag4	0.5604630	0.0978812	5.726	2.01e-07 ***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01893 on 75 degrees of freedom

Multiple R-squared: 0.5459, Adjusted R-squared: 0.5096

F-statistic: 15.03 on 6 and 75 DF, p-value: 3.294e-11

Value of test-statistic is: -3.7312 6.3659 6.9661

Critical values for test statistics:

1pct 5pct 10pct

tau3 -4.04 -3.45 -3.15

phi2 6.50 4.88 4.16

phi3 8.73 6.49 5.47

Con el criterio de Akaike se encontró que el valor mínimo se obtiene con una estructura de cuatro rezagos. Los resultados de la prueba ADF permiten rechazar la hipótesis de raíz unitaria debido a que la tau estimada es mayor en valor absoluto al valor critico del 5% de la prueba: -3.73>-3.45. Lo anterior concuerda con el hecho de tener una serie con tendencia determinista y tal vez sea más adecuado aplicar el método de la regresión que con diferencias para eliminar la tendencia.

Para complementar el análisis anterior, se aplicaron las pruebas ADF con constante y sin constante y tendencia, aplicando los siguientes códigos:

```
lc.df <- ur.df(y=PIB_Mex, type='drift',lags=4, selectlags=c("AIC"))
```

```
summary(lc.df)
```

```
lc.df <- ur.df(y=PIB_Mex, type='none',lags=4, selectlags=c("AIC"))
```

```
summary(lc.df)
```

```
#####
#####
```

```
# Augmented Dickey-Fuller Test Unit Root Test #
```

```
#####
#
```

Test regression drift

Call:

```
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.076920	-0.006254	0.004951	0.011547	0.033179

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.17204	0.23980	0.717	0.47529
z.lag.1	-0.01027	0.01482	-0.693	0.49040
z.diff.lag1	-0.09814	0.10163	-0.966	0.33728
z.diff.lag2	-0.02179	0.09613	-0.227	0.82130
z.diff.lag3	-0.28713	0.09752	-2.944	0.00429 **
z.diff.lag4	0.47634	0.10259	4.643	1.41e-05 ***
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02041 on 76 degrees of freedom

Multiple R-squared: 0.4649, Adjusted R-squared: 0.4297

F-statistic: 13.21 on 5 and 76 DF, p-value: 2.934e-09

Value of test-statistic is: -0.693 2.4612

Critical values for test statistics:

1pct 5pct 10pct

tau2 -3.51 -2.89 -2.58

phi1 6.70 4.71 3.86

#####

# Augmented Dickey-Fuller Test Unit Root Test #

#####

Test regression none

Call:

lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:

Min	1Q	Median	3Q	Max
-0.074402	-0.006846	0.002910	0.011820	0.035865

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
z.lag.1	0.0003611	0.0001714	2.106	0.03845 *
z.diff.lag1	-0.1033179	0.1010588	-1.022	0.30981
z.diff.lag2	-0.0236741	0.0957920	-0.247	0.80546
z.diff.lag3	-0.2906282	0.0970891	-2.993	0.00371 **
z.diff.lag4	0.4754677	0.1022619	4.650	1.35e-05 ***

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02035 on 77 degrees of freedom

Multiple R-squared: 0.488, Adjusted R-squared: 0.4547

F-statistic: 14.68 on 5 and 77 DF, p-value: 4.294e-10

Value of test-statistic is: 2.1061

Critical values for test statistics:

1pct 5pct 10pct

tau1 -2.6 -1.95 -1.61

Los resultados ahora muestran consistencia con el comportamiento del PIB en la gráfica anterior. Esto es, con el estadístico ADF con constante no se puede rechazar la hipótesis de raíz unitaria debido a que la tau estimada es menor en valor absoluto al valor crítico del 5% de la prueba: -0.69<-2.9. En el caso de la prueba ADF sin constante y tendencia, el resultado muestra que la tau es positiva y con ello se indica que se encuentra en el caso de una raíz característica mayor que uno. En los dos casos, los resultados concuerdan con el hecho de tener una serie con tendencia determinista y tal vez sea más adecuado aplicar el método de la regresión que con diferencias para eliminar la tendencia. En conclusión la serie del PIB no es estacionaria y por tanto el orden de integración es mayor a cero.

Para determinar el orden de integración de las variables, se aplican las pruebas ADF con sus tres posibilidades: 1) tendencia y constante; 2) con constante; y, 3) sin constante ni tendencia, a la primera diferencia del logaritmo del PIB.

#Pruebas con la primera diferencia de la variable PIB

```
lc.df <- ur.df(y= d_lpib_mex, type='trend',lags=4, selectlags=c("AIC"))
```

```

summary(lc.df)

lc.df <- ur.df(y= d_lpib_mex, type='drift',lags=4, selectlags=c("AIC"))

summary(lc.df)

lc.df <- ur.df(y= d_lpib_mex, type='none',lags=4, selectlags=c("AIC"))

summary(lc.df)

#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####


```

Test regression trend

Call:

```
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.064879	-0.005980	0.001778	0.011136	0.045258

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	7.109e-03	5.161e-03	1.377	0.1726
z.lag.1	-1.200e+00	2.753e-01	-4.359	4.15e-05 ***
tt	3.026e-06	9.451e-05	0.032	0.9745
z.diff.lag1	2.265e-01	2.546e-01	0.890	0.3765
z.diff.lag2	1.158e-01	2.011e-01	0.576	0.5665

z.diff.lag3	-1.815e-01	1.587e-01	-1.144	0.2564
z.diff.lag4	2.746e-01	1.130e-01	2.431	0.0175 *

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01986 on 74 degrees of freedom

Multiple R-squared: 0.8218, Adjusted R-squared: 0.8074

F-statistic: 56.89 on 6 and 74 DF, p-value: < 2.2e-16

Value of test-statistic is: -4.3589 6.3582 9.5312

Critical values for test statistics:

1pct	5pct	10pct	
tau3	-4.04	-3.45	-3.15
phi2	6.50	4.88	4.16
phi3	8.73	6.49	5.47

#####
#

# Augmented Dickey-Fuller Test Unit Root Test #

#####
#

Test regression drift

Call:

lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)

Residuals:

Min	1Q	Median	3Q	Max
-0.064994	-0.005982	0.001839	0.011256	0.045172

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	0.007248	0.002777	2.610	0.0109 *
z.lag.1	-1.200561	0.273146	-4.395	3.59e-05 ***
z.diff.lag1	0.226823	0.252635	0.898	0.3721
z.diff.lag2	0.116030	0.199591	0.581	0.5628
z.diff.lag3	-0.181313	0.157503	-1.151	0.2533
z.diff.lag4	0.274668	0.112201	2.448	0.0167 *
---				

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.01973 on 75 degrees of freedom

Multiple R-squared: 0.8218, Adjusted R-squared: 0.8099

F-statistic: 69.19 on 5 and 75 DF, p-value: < 2.2e-16

Value of test-statistic is: -4.3953 9.6655

Critical values for test statistics:

1pct	5pct	10pct	
tau2	-3.51	-2.89	-2.58
phi1	6.70	4.71	3.86

```
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####
```

Test regression none

Call:

```
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.058897	-0.001654	0.006114	0.016373	0.050246

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
z.lag.1	-0.7631	0.2238	-3.411	0.00104 **
z.diff.lag1	-0.1450	0.2164	-0.670	0.50490
z.diff.lag2	-0.1523	0.1775	-0.858	0.39351
z.diff.lag3	-0.3580	0.1475	-2.427	0.01760 *
z.diff.lag4	0.2061	0.1132	1.821	0.07257 .

---

Signif. codes: 0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.02047 on 76 degrees of freedom

Multiple R-squared: 0.8057, Adjusted R-squared: 0.7929

F-statistic: 63.01 on 5 and 76 DF, p-value: < 2.2e-16

Value of test-statistic is: -3.4106

Critical values for test statistics:

1pct 5pct 10pct

tau1 -2.6 -1.95 -1.61

Efectuando los tres tipos de pruebas para la variable en segundas diferencias se rechaza la hipótesis de raíz unitaria por lo que se puede concluir que la serie es integrada de orden uno I(1): -4.35>-3.45, -4.39>-2.9 y -3.41>-1.95 respectivamente.

Una de las alternativas a la prueba ADF fue desarrollada por Phillips y Perron (1988), conocida como la prueba PP. Estos autores buscaron incluir la posibilidad de que el término de error no fuera ruido blanco al existir la posibilidad de autocorrelación. A diferencia de la prueba ADF, en donde se busca atender esta posibilidad incorporando una estructura de rezagos en el término autorregresivo, estos autores agregan un factor de corrección a la prueba DF. Partiendo de un modelo AR(1):

$$Y_t = \alpha + \beta Y_{t-1} + u_t$$

El estadístico t, para el coeficiente de la variable autorregresiva en el modelo AR(1), es corregido si existe autocorrelación, por medio del estimador consistente de autocorrelación heterocedástica propuesto por Newey-West (1987):<sup>22</sup>

<sup>22</sup> La fórmula fue tomada del manual de usuario del Eviews

$$w^2 = \gamma_0 + 2 \sum_{j=1}^q \left( 1 - \frac{j}{q+1} \right) \gamma_j$$

Donde:

$$\gamma_j = \frac{1}{T} \sum_{t=j+1}^T \hat{u}_t \hat{u}_{t-j}$$

$$\gamma_0 = \frac{1}{T} \sum_{t=1}^T \hat{u}_t^2$$

Con base en dicho estimador se corrige la  $t_b$  calculada para el coeficiente autorregresivo en el modelo AR(1):

$$t_{pp} = \frac{g_0^{0.5} t_b}{w} - \frac{(w^2 - g_0) T s_b}{2 w \hat{S}}$$

$s_b$  es el error estándar del estimador del coeficiente autorregresivo y  $\hat{S}$  es el error estándar de la regresión.

El estimador de Newey-West es un estimador de la varianza de largo plazo, de la fórmula resulta que sí no existiera autocorrelación el término  $\gamma_j$  sería igual a cero, por lo que el estimador resultaría igual a  $\gamma_0$ . En dicho caso el estadístico  $t_{pp}$  sería igual al  $t_b$  del modelo utilizado.

En la fórmula del estimador aparece un término  $q$ , conocido como rezago de truncación de las autocovarianzas, dicho término es el número de períodos de autocorrelación a incluir. En los paquetes econométricos comerciales el rezago de truncación es elegido automáticamente, de otra forma habría que analizar la función de autocorrelación y determinar cuál es la última autocorrelación significativamente diferente de cero.

Una ventaja de la prueba PP es que resulta flexible para considerar especificaciones diferentes a la AR empleada por el ADF. Por ejemplo resultaría compatible con el supuesto de un proceso MA(1) conocido como medias móviles, en este caso la variable pudo haber sido generada por:

$$Y_t = \mu + u_t + q u_{t-1}$$

$\mu$  es una constante y  $u_t$  es ruido blanco

También podría ser compatible con una mezcla de los procesos AR(1) y MA(1), conocido como ARIMA(1,1):

$$Y_t = \alpha + \beta Y_{t-1} + u_t + q u_{t-1}$$

Generalizando, podríamos considerar el caso en que hubiera  $p$  términos autorregresivos y  $q$  términos de medias móviles y tendríamos un proceso ARMA( $p,q$ ).

Si a una serie de tiempo I(1) se la aplican primeras diferencias para volverla estacionaria y esta última es generada por un proceso ARMA(1,1), entonces la serie original será ARIMA(1,1,1), es decir autoregresiva integrada de medias móviles.

Cuando el parámetro autorregresivo es muy cercana a uno las pruebas ADF y PP tienen problemas para definir la existencia de raíz unitaria, para identificar el orden de integración y por tanto no están correctamente definidas la hipótesis nula de existencia de raíz unitaria y de no estacionario. Para eliminar este tipo de dilemas es posible aplicar una

prueba para confirmar los resultados de las de raíz unitaria, lo que proponen Kwiatkowski-Phillips-Schmidt-Shin (KPSS) es cambiar la hipótesis nula de no estacionario de las pruebas ADF y PP, por la de estacionario.

Para esta prueba se emplea la siguiente regresión:

$$Y_t = \alpha + u_t \quad \text{ó} \quad Y_t = \alpha + bt + u_t$$

Con base en un estimador de la función espectral de los residuales en la frecuencia cero ( $f_0$ ), y una función de los residuales acumulados  $S(t)$ , construye un estadístico de prueba de multiplicadores de Lagrange (LM):

$$\text{LM} = \sum_t S(t)^2 / T^2 f(0)$$

$$\text{donde: } S(t) = \sum_r \hat{u}_r$$

La aplicación de la prueba da lugar a los resultados para complementar y confirmar las pruebas de raíz unitaria ADF y PP.

**Ejemplo 2. Pruebas PP y KPSS que complementa al la prueba ADF para análisis de integración del PIB de la economía mexicana**

Para aplicar este ejercicio se requiere previamente, como en el ejercicio 1, cargar la librería **urca**, cambiar el directorio de trabajo, activar la base de datos y asignar las variables de **lpib\_mex** y primera diferencia **dl\_lpib\_mex**

```
#Cargar la librería urca
```

```

> library(urca)

#Cambiar el directorio de trabajo

>setwd("/Volumes/LACIESHARE/Academico/LibroEconometria_R/Capitulo_8/BaseDatos_Capitulo
8")

# Lectura de la base de datos

> load("BDatos_Integracion.RData")

# Se asigna el logaritmo de la variable de serie de tiempo al objeto PIB_MEX y se aplica la primera
diferencia

> lplib_mex <- log(BDatos$PIB_Mex)

> d_lplib_mex <- diff(lplib_mex)

```

Para complementar los resultados obtenidos con la aplicación de la prueba ADF al logaritmo del PIB, se aplican las pruebas PP con tendencia y constante, y con constante junto a la prueba KPSS que no opciones sobre al tendencia y la constante.

Los resultados de las pruebas PP que muestran a continuación para el caso de tendencia y constante, que el estadístico Z-tau (-4.19) > Z (-3.46) en términos absolutos, por lo que se concluye que se acepta la hipótesis alternativa de raíz no unitaria. Con la prueba PP solamente con constante, se encuentra que el estadístico Z-tau (-0.17269) < Z (-2.89) en términos absolutos, por lo que se concluye lo contrario y la variable tiene una raíz unitaria. Y, de acuerdo a la prueba KPSS el test-statistic (0.1607) > que el valor critico (0.146), lo cual implica que se acepta la hipótesis alternativa de raíz unitaria. Los tres resultados son consistentes con lo encontrado con la prueba ADF y se concluye también que la variable L-pib\_mex tiene raíz unitaria y por tanto no es estacionaria.

```

# Prueba PP con constante y tendencia

> lc.pp <- ur.pp(lplib_mex, type="Z-tau",model="trend", lags="long")
> summary(lc.pp)

```

```
#####
# Phillips-Perron Unit Root Test #
#####
```

Test regression with intercept and trend

Call:

```
lm(formula = y ~ y.l1 + trend)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.074772	-0.011976	0.002096	0.019893	0.041381

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	5.3779514	1.3243709	4.061	0.000110 ***
y.l1	0.6678778	0.0818828	8.157	3.17e-12 ***
trend	0.0020822	0.0005293	3.934	0.000173 ***
---				

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02464 on 83 degrees of freedom

Multiple R-squared: 0.9771, Adjusted R-squared: 0.9765

F-statistic: 1771 on 2 and 83 DF, p-value: < 2.2e-16

Value of test-statistic, type: Z-tau is: -4.1975

```

aux. Z statistics

Z-tau-mu      3.5396
Z-tau-beta    4.7490

Critical values for Z statistics:

      1pct     5pct     10pct
critical values -4.067342 -3.461976 -3.157041

# Prueba PP con constante
> lc.pp <- ur.pp(lpib_mex, type="Z-tau", model="constant", lags="long")
> summary(lc.pp)

#####
# Phillips-Perron Unit Root Test #
#####

Test regression with intercept

Call:
lm(formula = y ~ y.l1)

Residuals:
    Min     1Q   Median     3Q    Max 
-0.079675 -0.015387  0.001721  0.018707  0.046483

Coefficients:

```

(Intercept) 0.27550 0.29008 0.95 0.345  
y.l1 0.98335 0.01793 54.84 <2e-16 \*\*\*  
---  
Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
  
Residual standard error: 0.02668 on 84 degrees of freedom  
Multiple R-squared: 0.9728, Adjusted R-squared: 0.9725  
F-statistic: 3008 on 1 and 84 DF, p-value: < 2.2e-16  
  
Value of test-statistic, type: Z-tau is: -0.7269  
  
aux. Z statistics  
Z-tau-mu 0.7636  
  
Critical values for Z statistics:  
1pct 5pct 10pct  
critical values -3.507211 -2.895068 -2.584427  
  
>  
> #Pruebas KPSS para el logaritmo de la variable PIB  
> lc.kpss <- ur.kpss(lpib\_mex, type="tau", lags="short", use.lag = NULL)  
> summary(lc.kpss)  
  
#####  
# KPSS Unit Root Test #  
#####

Test is of type: tau with 3 lags.

Value of test-statistic is: 0.1607

Critical value for a significance level of:

10pct 5pct 2.5pct 1pct

critical values 0.119 0.146 0.176 0.216

Para complementar la identificación del orden de integración de la prueba ADF, a continuación se presentan los resultados de aplicar las pruebas PP y KPSS a la primera diferencia del logaritmo del PIB (d\_lpib\_mex). Con la prueba PP con tendencia y constante se encontró que el estadístico Z-tau (-15.68) > Z-statistic (-3.46) en términos absolutos, con lo cual se rechaza la hipótesis de raíz unitaria; la misma conclusión se tiene con el resultado de la prueba PP con constante, el estadístico Z-tau (-15.70) > Z-statistic (-2.89) en términos absolutos; y, finalmente con la prueba KPSS se encuentra que el test-statistic (0.038) < que el valor critico (0.146), lo cual confirma el resultado de las pruebas PP y ADF, de que la primera diferencia de la variable lpib\_mex es estacionaria y por tanto su orden de integración es uno ( $I(1)$ ).

# Prueba PP con tendencia y constante

```
> lc.pp <- ur.pp(d_lpib_mex, type="Z-tau",model="trend", lags="long")
> summary(lc.pp)
```

#####
#####

# Phillips-Perron Unit Root Test #

#####
#####

Test regression with intercept and trend

Call:

lm(formula = y ~ y.l1 + trend)

Residuals:

Min	1Q	Median	3Q	Max
-0.086707	-0.010604	0.000978	0.013443	0.050041

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	8.727e-03	2.777e-03	3.142	0.002332 **
y.l1	-3.940e-01	1.015e-01	-3.882	0.000208 ***
trend	-3.409e-05	1.102e-04	-0.309	0.757946

---

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*\*' 0.01 '\*\*' 0.05 '\*' 0.1 '' 1

Residual standard error: 0.02493 on 82 degrees of freedom

Multiple R-squared: 0.1557, Adjusted R-squared: 0.1351

F-statistic: 7.563 on 2 and 82 DF, p-value: 0.0009675

Value of test-statistic, type: Z-tau is: -15.6839

aux. Z statistics

Z-tau-mu 4.6499

Z-tau-beta -0.3625

Critical values for Z statistics:

1pct	5pct	10pct
------	------	-------

```

critical values -4.068637 -3.462585 -3.157396

# Prueba PP con constante

> lc.pp <- ur.pp(d_lpib_mex, type="Z-tau", model="constant", lags="long")
> summary(lc.pp)

#####
# Phillips-Perron Unit Root Test #
#####

Test regression with intercept

Call:
lm(formula = y ~ y.l1)

Residuals:
    Min      1Q  Median      3Q     Max 
-0.085464 -0.009809  0.000877  0.013973  0.051028 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept) 0.008706  0.002761   3.153   0.002250 ** 
y.l1        -0.393307  0.100896  -3.898   0.000196 *** 
---
Signif. codes:  0 '****' 0.001 '***' 0.01 '**' 0.05 '*' 0.1 ' ' 1

```

```
Residual standard error: 0.0248 on 83 degrees of freedom  
Multiple R-squared: 0.1547, Adjusted R-squared: 0.1446  
F-statistic: 15.2 on 1 and 83 DF, p-value: 0.0001957
```

Value of test-statistic, type: Z-tau is: -15.7046

aux. Z statistics

Z-tau-mu 3.5841

Critical values for Z statistics:

1pct 5pct 10pct

critical values -3.508125 -2.895469 -2.584638

```
> #Pruebas KPSS para la primera diferencia del logaritmo de la variable PIB
```

```
> lc.kpss <- ur.kpss(d_lpib_mex, type="tau", lags="short", use.lag = NULL)
```

```
> summary(lc.kpss)
```

```
#####
```

```
# KPSS Unit Root Test #
```

```
#####
```

Test is of type: tau with 3 lags.

Value of test-statistic is: 0.0387

Critical value for a significance level of:

10pct 5pct 2.5pct 1pct

critical values 0.119 0.146 0.176 0.216

## REFERENCIAS

Dickey, D.A., Fuller, W.A. (1979); *Distribution of the estimators for autoregressive time series with a unit root*, *Journal of the American Statistical Association* 74, 427–431.

Kwiatkowski, D.; Phillips, P. C. B.; Schmidt, P.; Shin, Y. (1992). "Testing the null hypothesis of stationarity against the alternative of a unit root". *Journal of Econometrics* 54 (1–3): 159–178

Phillips, P.C.B. and Perron, P. (1988), *Testing for a unit root in time series regression*, *Biometrika*, 75(2), 335–346.

Phillips, P.C.B. and Ouliaris, S. (1990), *Asymptotic Properties of Residual Based Tests for Cointegration*, *Econometrica*, Vol. 58, No. 1, 165–193.

## ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO

Cap9.integracion.R

## MATERIAL DE APRENDIZAJE EN LÍNEA

Teórica\_Cap9

Práctica\_Cap9

VideoPráctica\_Cap9

VideoTeoría\_Cap9

# **CAPÍTULO 10: COINTEGRACIÓN Y MODELOS DE CORRECCIÓN DE ERROR**

**Miguel Ángel Mendoza González y Luis Quintana Romero**

## **1 INTRODUCCIÓN**

En el capítulo anterior se presentaron las pruebas de raíz unitaria para el análisis de integración y en este capítulo se retoman esos elementos para desarrollar la metodología de cointegración, que establece “la combinación lineal entre dos o más variables debe cumplir con la condición de ser estacionaria”, esto es: que la combinación debe tener media, varianza y covarianza constante. El procedimiento de Engle y Granger consiste en utilizar el análisis de integración en la combinación de las variables, con el objetivo de probar si cumplen con la condición de ser estacionaria para establecer que son cointegradas. A la ecuación estática que se utiliza para probar cointegración se le conoce como la relación de equilibrio de largo plazo y para modelar la dinámica de corto plazo al equilibrio de largo plazo, Engle y Granger postulan que es necesario construir el Modelo de Corrección de Error (MCE). Para aplicar las metodologías de cointegración, este capítulo se estructura de la siguiente manera: 1) El análisis de cointegración de Engle-Granger con pruebas de raíz unitaria; 2) Prueba de Phillips y Ouliaris para cointegración; 3) Modelo de Corrección de Error con Engle-Granger; y, 4) Metodología de cointegración de Johansen-Juselius.

## 2 EL CONCEPTO DE COINTEGRACIÓN

La idea de cointegración fue desarrollada por los economistas galardonados con el premio nobel de economía en el año de 2003; Clive Granger y Robert Engle. Su antecedente inmediato fue el trabajo de Granger y Newbold “Spurious regression in econometrics” publicado en 1974 en donde mostraban que la utilización de series no estacionarias podría llevar a una relación de correlación accidental entre ellas. En simulaciones realizadas por estos autores, utilizaron series artificiales no estacionarias generadas a partir de procesos diferentes y completamente independientes. En teoría se hubiera esperado que, en esas regresiones, el valor del coeficiente de determinación fuera muy bajo y la prueba t de las pendientes no indicara significancia estadística. El resultado demostró que en algunas de esas relaciones los coeficientes de determinación eran muy elevados y los estadísticos t no seguían una distribución t bien comportada, lo cual impedía validar la inferencia estadística sobre los parámetros de las regresiones. Por ello, la relación presente entre las variables era de casualidad y no de causalidad.

Tal y como se expuso en secciones en capítulos anteriores, en este tipo de relaciones espurias, Granger detectó que también presentaban un estadístico Durbin-Watson muy bajo e inferior al coeficiente de determinación. Ello se explica por la forma en que se construyeron y relacionaron las series en la regresión. Recuerde que los procesos de raíz unitaria se generaron con las siguientes ecuaciones:

$$[10.1] \quad y_t = y_{t-1} + e_t$$

$$[10.2] \quad z_t = z_{t-1} + v_t$$

Y, posteriormente se estimó la regresión:

$$[10.3] \quad y_t = \beta_1 + \beta_2 z_t + u_t$$

Como los dos procesos son independientes, el valor esperado de las betas es de cero, por consiguiente el término de error es:

$$[10.4] \quad u_t = y_t = f(u_t)$$

Por consiguiente, los términos de error están fuertemente correlacionados reflejando esto en un Durbin-Watson cercano a cero. Esto se confirma observando la función de autocorrelación, las autocorrelaciones son muy elevadas (cercana a la unidad en el rezago uno) y decrecen muy lentamente. Ante el hecho de que gran parte de las series económicas son no estacionarias y, por consiguiente, pueden presentar raíces unitarias, la relación entre ellas puede ser espuria. La diferenciación de un número adecuado de veces de las series podía dar lugar a procesos estacionarios al remover tendencias estocásticas en las series y, de ser así, las técnicas de regresión clásicas podían utilizarse. Sin embargo, la diferenciación de series involucra perdidas de información al sacrificar una observación en cada diferencia y, además, las variables diferenciadas podrían presentar poca relación entre sí y lagunas en su interpretación económica.

Granger y Engle observaron que una excepción se presentaba cuando al combinar series no estacionarias sus residuales si eran estacionarios, a ello le llamaron cointegración. Lo definían como si  $x_t$  y  $y_t$  son  $I(1)$  pero existe una combinación lineal entre ellas, del tipo:

$$[10.5] \quad z_t = m + ax_t + by_t$$

Si  $z_t$  es  $I(0)$ , entonces se dice que  $x_t$  y  $y_t$  están cointegradas y  $[m \ a \ b]$  es un vector de cointegración. En términos muy intuitivos la idea de cointegración supone la existencia de un atractor para las series en el largo plazo, el cual está representado por la combinación lineal  $z_t$  en la definición dada antes. Ello significa que dos series que cointegran exhiben un equilibrio de largo plazo entre sí, dando lugar a la anulación de la tendencia común que presentan entre ellas. Otra vez, de manera intuitiva, en el caso que las variables mantengan una relación de equilibrio lineal entre ellas, que se representa por el vector de cointegración, las desviaciones de ese equilibrio son medidas por  $z_t$  y, dado que son estacionarias o  $I(0)$ , son en consecuencia transitorias.

Si retomamos la discusión acerca de los procesos estacionarios y no estacionarios y la discusión que hemos presentado en esta sección, es posible considerar las siguientes situaciones para dos procesos estocásticos:

- 1) Si  $x_t$  y  $y_t$  son ambos estacionarios, aplica las técnicas de regresión clásica.
- 2) Si son integradas de diferente orden, la regresión clásica no tiene sentido.
- 3) Si son integradas del mismo orden y los residuales contienen tendencia estocástica, la regresión es espuria. Como ya vimos antes, en este caso es posible aplicar primeras diferencias si las series presentan tendencia estocástica.
- 4) Si son integradas del mismo orden y los residuales son una secuencia estacionaria, los dos procesos son cointegrados y aplica la regresión clásica.

### **3. PRUEBA DE COINTEGRACIÓN DE ENGLE Y GRANGER**

Una de las pruebas utilizadas comúnmente para evaluar la existencia de cointegración es la desarrollada por Engle y Granger en su trabajo ya referido de

1987. Para exemplificarla considere que el modelo a estimar es el más general, con  $k$  menos un variables, su forma matricial es:

$$[10.6] \quad Y = XB + U$$

La prueba puede ejecutarse en dos pasos: 1) Realizar pruebas de raíz unitaria a las series de la regresión para verificar que el orden de integración sea  $I(1)$ ; y, 2) Estimar la regresión cointegrante:

$$[10.7] \quad Y = X\hat{B} + \hat{U}$$

Donde se aplican las pruebas de raíz unitaria a los residuales de esta ecuación para verificar su orden de integración. En caso de ser  $I(0)$  no se podrá rechazar la hipótesis nula de cointegración. Al aplicar las pruebas ADF, PP y KPSS a los residuales de la ecuación cointegrante se deben consultar los valores de las tablas de cointegración construidas por MacKinnon (1996).

#### Ejemplo 1. Metodología de Engle-Granger aplicada a la función Consumo

En este ejemplo se utilizan las librerías **urca** para análisis de integración y cointegración para series de tiempo escrita por Bernhard Pfaff y Matthieu Stigler (2013) y **car** para el análisis de regresión aplicada de Fox y Weisberg (2011).

```
#Activar las librería urca y car
```

```
> library(urca)
```

```

> library(car)

#Cambiar el directorio de trabajo
setwd("/LibroEconometria_R/Capitulo_9/BaseDatos_Capitulo9")

# Lectura de la base de datos
load("Consumo.RData")

summary(Consumo) # Para el resumen de estadísticos básicos de los datos

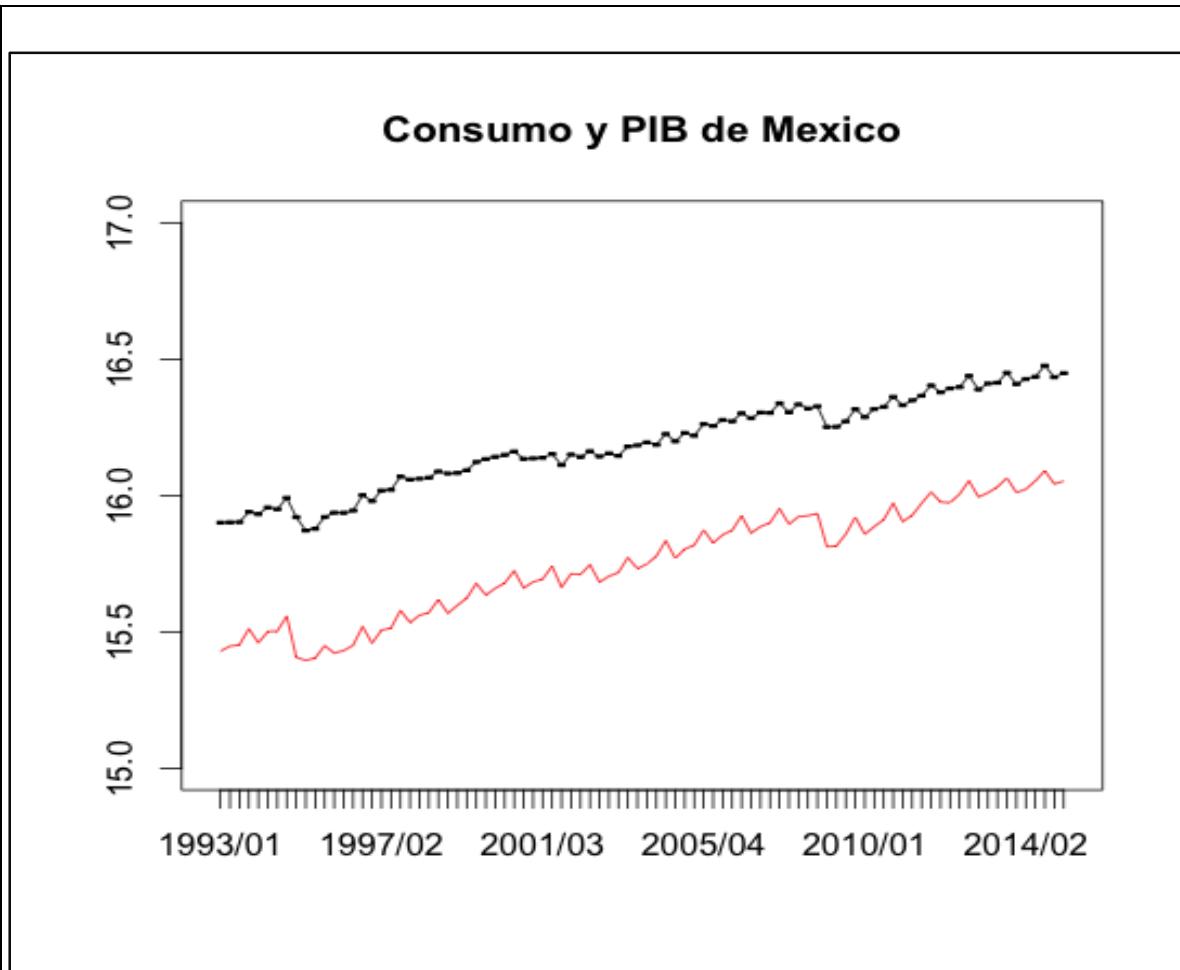
  Periodo      cp_mex      pib_mex
1993/01: 1  Min.   :4855294  Min.   : 7817381
1993/02: 1  1st Qu.:5799294  1st Qu.: 9553135
1993/03: 1  Median :6986111  Median :10694080
1993/04: 1  Mean    :7073960  Mean    :10885032
1994/01: 1  3rd Qu.:8218569  3rd Qu.:12281916
(Other):85  Max.   :9734100  Max.   :14307437
NA's   :1  NA's   :1       NA's   :1

# Se asignan las variables de serie de tiempo al objeto PIB_MEX y aplicar primera y
segunda diferencia

lpib_mex <- log(Consumo$pib_mex)
lcp_mex <- log(Consumo$cp_mex)
dlpib_mex <- diff(lpib_mex)
dlcp_mex <- diff(lcp_mex)

```

```
# Se asigna la variable periodo que ordena en el tiempo  
periodo <- Consumo$Periodo  
  
# Para graficar las variables de los logaritmos del PIB_Mex y el Consumo_Mex en el  
tiempo  
  
plot(periodo, lplib_mex, main="Consumo y PIB de Mexico", ylim=c(15, 17))  
lines(lplib_mex, col="black")  
lines(lcp_mex, col="red")  
  
Como se puede observar en la gráfica siguiente, las series del consumo y del PIB de  
México siguen una tendencias con respecto al tiempo, con caídas muy parecidas en los  
periodos de crisis económicas de 1994-1995 y 2008-2009.
```

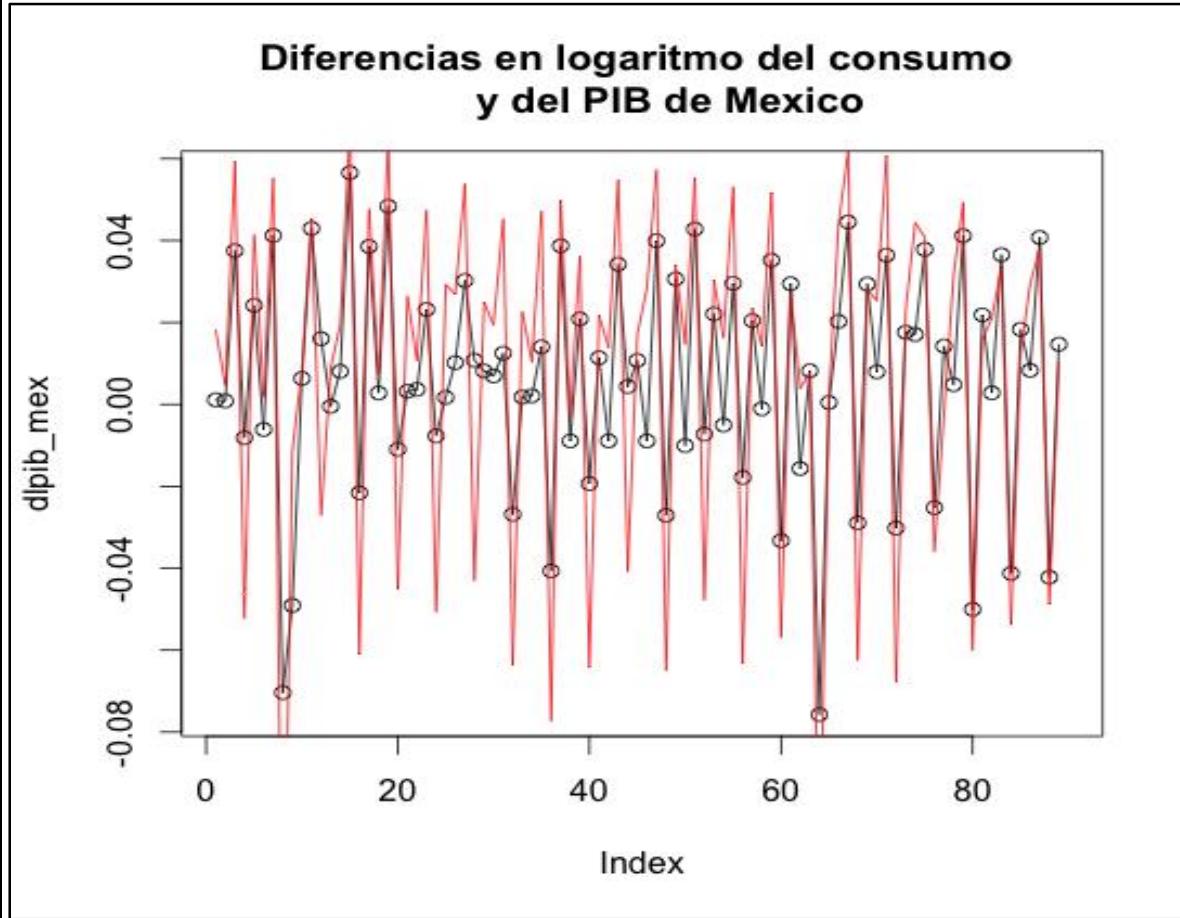


# Para construir un gráfica con las primera diferencia del Consumo y del PIB

```
plot(dlpib_mex, main="Diferencias en logaritmo del consumo
y del PIB de Mexico")
lines(dlpib_mex, col="black")
lines(dlcp_mex, col="red")
```

De la gráfica siguiente con las primeras diferencias de los logaritmos del consumo (rojo) y del pib (negro con círculos), los resultados muestran que las dos transformaciones son series sin tendencia. Un aspecto adicional importante es que el consumo tiene mayor variabilidad que el pib. La primera característica esta relacionada con series estacionarias

en media y complementado con el análisis de integración se puede concluir que las dos variables son i(1). Aunque la segunda característica es importante para el análisis económico, en términos de los supuestos de estacionariedad lo relevante es que las varianzas sean estacionarias con respecto al tiempo y tal supuesto parece cumplirse para las dos series de tiempo.



Para seguir con la metodología de Engle-Granger, en primer lugar se estima la función consumo con elasticidades constantes

```
# Función consumo para probar cointegración tipo Engle y Granger
mod_1 <- lm(lcp_mex ~ lplib_mex)
summary(mod_1)
```

Call:

```
lm(formula = lcp_mex ~ lpib_mex)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.062111	-0.013747	0.000094	0.016850	0.057125

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-3.57902	0.24675	-14.51	<2e-16 ***
lpib_mex	1.19408	0.01524	78.35	<2e-16 ***
---				
Signif. codes:	0 ***	0.001 **	0.01 *	0.05 .
	1	1	1	1

Residual standard error: 0.02409 on 88 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.9859, Adjusted R-squared: 0.9857

F-statistic: 6138 on 1 and 88 DF, p-value: < 2.2e-16

Los resultados muestran que la constante y el parámetro del logaritmo del PIB son altamente significativos y la elasticidad ingreso del consumo es de 1.19.

Para generar un objeto con los residuales y graficarlos, se utilizan los siguientes comandos

```
# Generar los residuales de la ecuación
```

```
res_1 <- residuals.lm(mod_1)
```

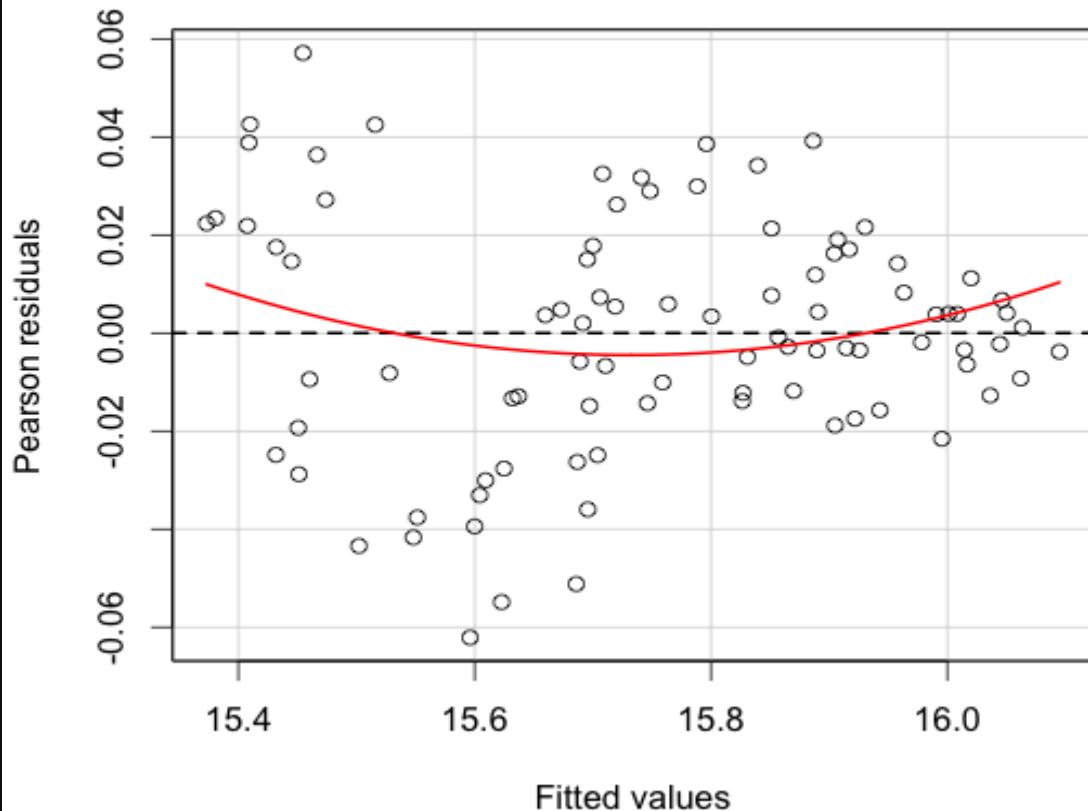
```
summary(res_1)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
-6.211e-02	-1.375e-02	9.414e-05	0.000e+00	1.685e-02	5.713e-02

```
# Grafica de los errores
```

```
residualPlot(mod_1)
```

La gráfica de los residuales muestra un comportamiento de tendencia con pendiente muy similar al cero o sin un comportamiento definido, lo cual es importante para el cumplimiento de la estacionariedad o que sea  $I(0)$ .



En segundo lugar, a los residuales de la función consumo se le aplica las pruebas de raíz unitaria.<sup>23</sup>

```
# Prueba de ADF para cointegracion tipo Granger  
lc.df <- ur.df(y=res_1, type='trend',lags=4, selectlags=c("AIC"))  
summary(lc.df)
```

```
#####
```

```
# Augmented Dickey-Fuller Test Unit Root Test #
```

```
#####
```

Test regression trend

Call:

```
lm(formula = z.diff ~ z.lag.1 + 1 + tt + z.diff.lag)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.0302503	-0.0060666	-0.0004746	0.0071103	0.0283283

Coefficients:

Estimate	Std. Error	t value	Pr(> t )
(Intercept) -3.545e-03	2.568e-03	-1.380	0.17140

<sup>23</sup> En este ejemplo, se muestra solamente las pruebas ADF, las otras se le piden a lector las haga como parte de su aprendizaje.

```

z.lag.1 -1.705e-01 6.408e-02 -2.661 0.00946 **
tt      5.919e-05 4.791e-05 1.236 0.22035
z.diff.lag1 -2.209e-01 9.695e-02 -2.279 0.02540 *
z.diff.lag2 -2.636e-01 9.231e-02 -2.855 0.00551 **
z.diff.lag3 -2.791e-01 8.946e-02 -3.119 0.00254 **
z.diff.lag4 5.527e-01 8.772e-02 6.301 1.63e-08 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

```

Residual standard error: 0.01048 on 78 degrees of freedom

Multiple R-squared: 0.8037, Adjusted R-squared: 0.7886

F-statistic: 53.21 on 6 and 78 DF, p-value: < 2.2e-16

Value of test-statistic is: **-2.6608** 2.6938 3.9284

Critical values for test statistics:

	1pct	5pct	10pct
tau3	-4.04	-3.45	<b>-3.15</b>
phi2	6.50	4.88	4.16
phi3	8.73	6.49	5.47

La primera prueba ADF considera la tendencia y constante, y los resultados muestran que el valor del estadístico (-2.66) es menor en valor absoluto a valor critico al 10 por ciento (-3.15), por lo que los residuales tienen raíz unitaria por lo que se concluye que el consumo y el pib no mantienen una relación de cointegración.

```
# La segunda prueba ADF se aplica solamente con constante
```

```
lc.df <- ur.df(y=res_1, type='drift',lags=4, selectlags=c("AIC"))
```

```
summary(lc.df)
```

```
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####
```

Test regression drift

Call:

```
lm(formula = z.diff ~ z.lag.1 + 1 + z.diff.lag)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.032166	-0.005711	0.000099	0.007042	0.026794

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-0.0007076	0.0011532	-0.614	0.54123
z.lag.1	-0.1597353	0.0636988	-2.508	0.01420 *
z.diff.lag1	-0.2053951	0.0964468	-2.130	0.03632 *
z.diff.lag2	-0.2464100	0.0915675	-2.691	0.00869 **
z.diff.lag3	-0.2590241	0.0882744	-2.934	0.00437 **

```

z.diff.lag4 0.5730218 0.0864515 6.628 3.8e-09 ***
---
Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01051 on 79 degrees of freedom
Multiple R-squared: 0.7998, Adjusted R-squared: 0.7871
F-statistic: 63.13 on 5 and 79 DF, p-value: < 2.2e-16

Value of test-statistic is: -2.5077 3.2558

Critical values for test statistics:
    1pct 5pct 10pct
tau2 -3.51 -2.89 -2.58
phi1 6.70 4.71 3.86

# Por último se aplica la prueba ADF sin tendencia y constante

lc.df <- ur.df(y=res_1, type='none', lags=4, selectlags=c("AIC"))
summary(lc.df)

```

En este caso se encontró que aunque la diferencia es pequeña, el valor del estadístico (-2.51) es menor en valor absoluto a valor crítico al 10 por ciento (-2.58), por lo que la conclusión es la misma que en la prueba anterior, el consumo y el pib no están cointegrados.

```
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####
```

Test regression none

Call:

```
lm(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.032840	-0.006482	-0.000614	0.006472	0.026333

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
z.lag.1	-0.15752	0.06335	-2.487	0.01498 *
z.diff.lag1	-0.19996	0.09566	-2.090	0.03977 *
z.diff.lag2	-0.24062	0.09072	-2.652	0.00964 **
z.diff.lag3	-0.25287	0.08736	-2.895	0.00489 **
z.diff.lag4	0.57956	0.08546	6.782	1.86e-09 ***

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.01047 on 80 degrees of freedom

Multiple R-squared: 0.7989, Adjusted R-squared: 0.7863

F-statistic: 63.56 on 5 and 80 DF, p-value: < 2.2e-16

Value of test-statistic is: -2.4866

Critical values for test statistics:

1pct 5pct 10pct

tau1 -2.6 -1.95 -1.61

Con esta especificación de la prueba, el valor del estadístico (-2.48) es mayor en valor absoluto a valor critico al 10 porciento (-1.61), por lo que ahora la conclusión se modifica y se puede asegurar que el consumo y el pib están cointegrados.

#### 4. ANÁLISIS DE COINTEGRACIÓN DE PHILLIPS-OULIARIS

##### Ejemplo 2. Análisis de cointegración de Phillips y Ouliaris (PO) para función Consumo

Para este ejemplo, se recomienda que tenga activadas las librerías, el cambio de directorio, se asignen las variables previo a la aplicación de la prueba de cointegración de PO.

```
#Cargar la libreria urca
```

```
library(urca)
```

```
library(car)
```

```

#Cambiar el directorio de trabajo

setwd("/Volumes/LACIE
SHARE/Academico/LibroEconometria_R/Capitulo_9/BaseDatos_Capitulo9")

# Lectura de la base de datos

load("Consumo.RData")

summary(Consumo)

# Se asigna la variable de serie de tiempo al objeto PIB_MEX y aplicar primera y segunda
diferencia

lpirb_mex <- log(Consumo$pib_mex)

lcp_mex <- log(Consumo$cp_mex)

# Prueba de Phillips y Ouliaris para cointegracion

#Para aplicar la prueba PO, primero utilizamos el comando cbind que se usa para
integrar variables en un solo objeto

ecb.consumo <- cbind(lcp_mex, lpirb_mex)

# Entonces se aplica la prueba PO del tipo Pz

Lc.po <- ca.po(ecb.consumo, type="Pz")

summary(Lc.po)

#####
# Phillips and Ouliaris Unit Root Test #
#####

```

Test of type Pz

detrending of series none

Response lcp\_mex :

Call:

lm(formula = lcp\_mex ~ zr - 1)

Residuals:

Min	1Q	Median	3Q	Max
-0.158515	-0.029906	0.008515	0.030204	0.061358

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
zrlcp_mex	0.7466	0.1063	7.026	4.49e-10 ***
zrlpib_mex	0.2470	0.1034	2.389	0.0191 *

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04392 on 87 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 5.727e+06 on 2 and 87 DF, p-value: < 2.2e-16

Response lpib\_mex :

Call:

```
lm(formula = lplib_mex ~ zr - 1)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.077564	-0.015059	0.000777	0.020031	0.044753

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
zrlcp_mex	-0.09114	0.06486	-1.405	0.164
zrlpib_mex	1.08906	0.06311	17.257	<2e-16 ***

---

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.02681 on 87 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 1.624e+07 on 2 and 87 DF, p-value: < 2.2e-16

Value of test-statistic is: 10.9944

Critical values of Pz are:

10pct	5pct	1pct	
critical values	33.9267	40.8217	55.1911

Los resultados muestran que el estadístico calculado (10.99) es menor que el valor critico al 10 porciento (33.92) y todos los niveles de significancia, con lo que se concluye que se acepta la hipótesis nula de no cointegración.

```
Lc.po <- ca.po(ecb.consumo, type="Pu")
summary(Lc.po)

#####
# Phillips and Ouliaris Unit Root Test #
#####

Test of type Pu
detrending of series none

Call:
Im(formula = z[, 1] ~ z[, -1] - 1)

Residuals:
    Min      1Q   Median      3Q      Max 
-0.091371 -0.035614  0.007748  0.039446  0.063685 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
z[, -1]  0.9730269  0.0002872  3388     <2e-16 ***
```

---

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.04411 on 89 degrees of freedom

Multiple R-squared: 1, Adjusted R-squared: 1

F-statistic: 1.148e+07 on 1 and 89 DF, p-value: < 2.2e-16

Value of test-statistic is: 9.3937

Critical values of Pu are:

10pct 5pct 1pct

critical values 20.3933 25.9711 38.3413

La conclusión con la prueba **Pz** se confirma con el estadístico **Pu**; el estadístico calculado (9.39) es menor que el valor critico al 10 porciento (20.39) y todos los niveles de significancia, con lo que se concluye que se acepta la hipótesis nula de no cointegración.

## 5. MODELO DE CORRECCIÓN DE ERROR

Engle y Granger demuestran que si existe una doble inferencia: 1) Si un conjunto de variables están cointegradas también existirá un mecanismo de corrección de error (MCE) para representar el proceso generador de los datos 8PGD); y, 2) Si el PGD de las variables tiene una representación de MCE, entonces esas variables estarán cointegradas. A esta inferencia se le conoce como el “Teorema de Representación de Granger”. Ello implica, si retomamos el ejemplo referido al

consumo, que si esta cointegrado con el PIB, es decir que si la relación entre las variables es  $CI(1,1)$ , su posible representación en MCE es:

$$[10.7] \quad \Delta lcp_t = \alpha_1 + \alpha_1 \Delta lpib_t + \delta(lcp_{t-1} - \beta_1 - \beta_2 lpib_{t-1}) + \varepsilon_t$$

Es fácil verificar que la expresión en paréntesis son los residuales obtenida de la función de cointegración, aunque rezagada un período. Esa expresión, muestra el desequilibrio que se presenta en la relación entre las variables y su coeficiente,  $\delta$ , que se conoce como la corrección de error. Dicho coeficiente debe ser negativo y, en valor absoluto, menor a la unidad con el fin de asegurar, en este caso, que los cambios en el consumo sean hacia el equilibrio.

### Ejemplo 3. Modelo de Corrección de Error para el Consumo

```
#Activar las librería urca y car
> library(urca)
> library(car)

#Cambiar el directorio de trabajo
setwd("/LibroEconometria_R/Capitulo_9/BaseDatos_Capitulo9")

# Lectura de la base de datos
load("Consumo.RData")

# Se asigna la variable de serie de tiempo al objeto PIB_MEX y aplican primeras
diferencias
```

```

lpib_mex <- log(Consumo$pib_mex)

lcp_mex <- log(Consumo$cp_mex)

dlpib_mex <- diff(lpib_mex)

dlcp_mex <- diff(lcp_mex)

```

Para poder estimar el modelo de corrección de error es muy importante tener asignadas la variables en logaritmo y primera diferencia del logaritmo, estimar el modelo de largo plazo, guardar los residuales y ahora el rezago de los residuos (res\_1.l).

```

# Modelo de corrección de Error: Metodología de Granger

mod_1 <- lm(lcp_mex ~ lpib_mex)

```

Call:

```
lm(formula = lcp_mex ~ lpib_mex)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.062111	-0.013747	0.000094	0.016850	0.057125

Coefficients:

	Estimate	Std. Error	t value	Pr(> t )
(Intercept)	-3.57902	0.24675	-14.51	<2e-16 ***
lpib_mex	1.19408	0.01524	78.35	<2e-16 ***
---				
Signif. codes:	0 ****	0.001 ***	0.01 **	0.05 *
	.'	0.1 '	'	1

```
Residual standard error: 0.02409 on 88 degrees of freedom
```

```
(1 observation deleted due to missingness)
```

```
Multiple R-squared: 0.9859, Adjusted R-squared: 0.9857
```

```
F-statistic: 6138 on 1 and 88 DF, p-value: < 2.2e-16
```

```
# Se asignan los residuales del modelo estático (mod_1)
```

```
res_1 <- residuals.lm(mod_1)
```

```
# Se aplica un rezago a los residuales y se asignan a (res_1.l)
```

```
res_1.l=lag(res_1) # rezago de los residuales
```

Entonces se estima la función de las diferencias del logaritmo del consumo (dlcp\_mex) con respecto a las diferencias del logaritmo del pib (dlpib\_mex) y de los residuales rezagados un periodo.

```
mod_ce_1 <- lm(dlcp_mex ~ dlpib_mex+res_1.l)
```

```
summary(mod_ce_1)
```

Call:

```
lm(formula = dlcp_mex ~ dlpib_mex + res_1.l)
```

Residuals:

Min	1Q	Median	3Q	Max
-0.054691	-0.011976	0.000953	0.014435	0.049647

Coefficients:

Estimate	Std. Error	t value	Pr(> t )
----------	------------	---------	----------

```
(Intercept) -0.001721  0.002101 -0.819  0.415
dlpib_mex   1.423884  0.076888 18.519  < 2e-16 ***
res_1.l     -0.424926  0.086094 -4.936  3.87e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Residual standard error: 0.01931 on 86 degrees of freedom

(1 observation deleted due to missingness)

Multiple R-squared: 0.8206, Adjusted R-squared: 0.8164

F-statistic: 196.7 on 2 and 86 DF, p-value: < 2.2e-16

Como se observa el parámetro del vector de cointegración (res\_1.l) es negativo y en términos absolutos menos que uno (-0.43), por lo que se acepta que la relación de corto plazo tiende a la de largo plazo. La elasticidad ingreso del consumo de corto plazo es 1.42 y de largo plazo es 1.19, lo cual indica una sobrerreacción de consumo en el corto plazo.

## 6. COINTEGRACIÓN CON METODOLOGÍA DE JOHANSEN Y JOSELIUS

Una alternativa a los procedimientos de evaluación de raíz unitaria y de cointegración que hemos revisado es el método de Johansen (1988). Con base en dicho método es posible probar tanto el orden de integración de un conjunto de variables, como la existencia de cointegración entre las mismas.

El procedimiento se sustenta en los modelos VAR, que con una especificación de un solo rezago, se escribe:

$$Y_t = \mu + A_1 Y_{t-1} + v_t$$

Ahora transformamos el modelo restando  $Y_{t-1}$  de los dos lados:

$$Y_t - Y_{t-1} = \mu + A_1 Y_{t-1} - Y_{t-1} + v_t$$

Agrupando las variables tenemos:

$$\Delta Y_t = \mu + A Y_{t-1} + v_t$$

Donde:

$$A = -I + A_1$$

El lector puede confirmar que esta representación del modelo es la que llamamos Mecanismo de Corrección de Error en la sección anterior; por ello  $A$  es, por tanto, un **Vector de Corrección de Error**: Por ello, el modelo es un Mecanismo de Corrección de Error Vectorial (VECM). También se dará cuenta de la similitud de esta especificación con la de la prueba ADF, incluso podríamos considerarla un ADF multiecuacional.

Si ahora generalizamos la especificación para considerar un modelo VAR( $p$ ), el resultado es:

$$Y_t = \mu + A_1 Y_{t-1} + A_2 Y_{t-2} + \dots + A_p Y_{t-p} + v_t$$

Si se resta  $Y_{t-1}$  hasta  $Y_{t-p}$  de los dos lados, y se rescribe en función de la  $\Delta Y_t$ , entonces:

$$\Delta Y_t = \mu + \Gamma_1 \Delta Y_{t-1} + \Gamma_2 \Delta Y_{t-2} + \dots + \Gamma_{p-1} \Delta Y_{t-p} + A Y_{t-p} + e_t$$

Donde:

$$\Gamma = -I + \Gamma_1 + \dots + \Gamma_i \quad i=1, \dots, p-1$$

$$A = -I + A_1 + \dots + A_p$$

La prueba es sensible a la longitud de los rezagos elegidos en el VECM, por lo tanto deben ser seleccionados óptimamente utilizando los criterios de información que ya hemos visto en capítulos anteriores. En la matriz  $A$  se encuentre la relación de largo plazo. Si su rango es; rango ( $A$ ) =  $r$ , entonces se pueden encontrar las siguientes situaciones:

- a) ***Si  $r = 0$ ,  $A$  es una matriz nula. No existirá ninguna relación de cointegración.***
- b) ***Si  $r = m$ , el proceso multivariante  $Y_t$  es estacionario. Por tanto, habrá  $m-1$  vectores de cointegración linealmente independientes que cancelan la tendencia común. Así  $Y_t$  será estacionario si  $A_{m \times m}$  tiene rango completo.***
- c) ***Si  $0 < r < m$ , se encontrará entre las dos situaciones anteriores, por lo que habrá  $r$  relaciones de cointegración.***

Por lo tanto, el rango de  $A$  mostrará el número de columnas linealmente independientes de esta matriz y ese será también el número de vectores de cointegración existentes entre las variables del VAR. Por otro lado, si  $r > 0$ ,  $A$  puede rescribirse como el producto de dos matrices de dimensión  $(m \times r)$ ,  $A = \gamma \alpha'$ ; siendo  $\alpha$  la matriz de vectores de cointegración,  $\alpha' Y_{t-1}$  representa el término de corrección de error y  $\gamma$  es la matriz de parámetros que mide la velocidad de ajuste.

Johansen demuestra que la estimación máximo verosímil de la matriz de vectores de cointegración,  $\alpha$ , se obtiene a partir del cálculo de las raíces características  $\lambda_i$ ,  $i=1,\dots,m$ .

Para contrastar la hipótesis nula de que hay como máximo  $r$  vectores de cointegración frente a la alternativa de que hay  $m$ ,  $r \leq m$ , el contraste de razón de verosimilitud viene dado por los estadísticos de la traza y de la raíz máxima:

$$\text{Traza} = -2LnQ = -T \sum_{i=r+1}^m (1 - \lambda_i)$$

$$\text{Raíz máxima} = \lambda_r^{max} = -T \cdot Ln(1 - \lambda_r)$$

El contraste de hipótesis consiste en la secuencia:

- 1) La hipótesis nula  $H_0: r=0$  (no cointegración), frente a la alternativa  $H_a: r=1$ .
- 2) En caso de rechazar esta hipótesis (utilizando cualquiera de los dos estadísticos propuestos), se contrasta ahora  $H_0: r=1$  frente a la alternativa  $H_a: r=2$ , y así sucesivamente hasta el momento en que no se rechaza  $H_0$ , o bien hasta que se tuviera que aceptar la hipótesis alternativa de  $r=m$ .

Los valores críticos para los dos estadísticos de prueba, dependen de la inclusión o no de constantes en las ecuaciones; se pueden incluir interceptos en los vectores de cointegración o en el VAR.

Para ejemplificar la prueba de Johansen suponga un VECM de tres variables, tenemos entonces:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \\ a_{31} & a_{32} & a_{33} \end{bmatrix}$$

En el caso de que el rango de la matriz sea  $r=1$ , habrá un vector de cointegración por lo cual:

$$\mathbf{A} = \gamma \alpha' = \begin{bmatrix} \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \end{bmatrix}$$

En caso de haber dos vectores de cointegración  $\gamma$  y  $\alpha$  serán de dimensiones (3x2) y (2x3) respectivamente. Retomando el caso de un solo vector de cointegración, el término de corrección de error se escribe:

$$\mathbf{A} Y_{t-p} = \begin{bmatrix} \gamma_{11} \\ \gamma_{12} \\ \gamma_{13} \end{bmatrix} \begin{bmatrix} \alpha_{11} & \alpha_{12} & \alpha_{13} \end{bmatrix} \begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \end{bmatrix}_{t-p}$$

De esta forma, se podría escribir la representación de mecanismo de corrección de error para cada variable  $\Delta Y_t$  en el VECM.

#### Ejemplo 4. Análisis de cointegración de Johansen-Joselius

```
#Activar las librería urca y car
```

```
> library(urca)
> library(car)
```

```
#Cambiar el directorio de trabajo
```

```

setwd("/LibroEconometria_R/Capitulo_9/BaseDatos_Capitulo9")

# Lectura de la base de datos

load("Consumo.RData")

# Se asigna la variable de serie de tiempo al objeto PIB_MEX y aplican primeras
diferencias

lpib_mex <- log(Consumo$pib_mex)

lcp_mex <- log(Consumo$cp_mex)

# Para aplicar el procedimiento de Johansen, primero se combinan las variables
#en un solo objeto

ecb.consumo <- cbind(lcp_mex,lpib_mex)

En primer lugar se aplica la prueba de la traza de cointegración sin tendencia y constante

# Prueba Johansen de cointegracion de la traza

summary(ca.jo(ecb.consumo, type="trace",ecdet="none",spec=c("longrun"), K=4))

#####
# Johansen-Procedure #
#####

Test type: trace statistic , with linear trend

```

Eigenvalues (lambda):

[1] 0.1751774 0.0169835

Values of teststatistic and critical values of test:

	test	10pct	5pct	1pct
r <= 1	1.47	6.50	8.18	11.65
r = 0	18.04	15.66	17.95	23.52

Eigenvectors, normalised to first column:

(These are the cointegration relations)

	lcp_mex.l4	lpib_mex.l4
lcp_mex.l4	1.000000	1.000000
lpib_mex.l4	-1.280708	-1.037871

Weights W:

(This is the loading matrix)

	lcp_mex.l4	lpib_mex.l4
lcp_mex.d	-0.3578910	-0.08189816
lpib_mex.d	-0.1079621	-0.07962291

Los resultados muestran que para la primera hipótesis  $r=0$  y alternativa  $r=1$ , se acepta la hipótesis alternativa de cointegración: el estadístico es mayor al valor critico del 10 porciento,  $18.04 > 15.56$ . Para la segunda hipótesis  $r<=1$  y alternativa  $r=2$ , se acepta la hipótesis nula de un solo vector de cointegración. Por lo que se puede concluir que existe un vector de cointegración y se representa por [1 -1.28].

```

# Prueba de Johansen de la traza con constante

summary(ca.jo(ecb.consumo, type="trace",ecdet="const",spec=c("longrun"), K=4))

#####
# Johansen-Procedure #
#####

```

Test type: trace statistic , without linear trend and constant in cointegration

Eigenvalues (lambda):

```
[1] 2.017097e-01 1.265562e-01 5.440093e-15
```

Values of teststatistic and critical values of test:

	test	10pct	5pct	1pct
r <= 1	11.64	7.52	9.24	12.97
r = 0	31.01	17.85	19.96	24.60

Eigenvectors, normalised to first column:

(These are the cointegration relations)

	lcp_mex.l4	lpib_mex.l4	constant
lcp_mex.l4	1.000000	1.000000	1.000000

```
lpib_mex.l4 -1.255645 -1.414691 -1.064081  
constant      4.557479  7.198639  1.483342
```

Weights W:

(This is the loading matrix)

```
lcp_mex.l4 lpib_mex.l4    constant  
lcp_mex.d -0.4955292 0.05574003 -1.774483e-11  
lpib_mex.d -0.2795953 0.09201028 -6.438725e-12
```

Los resultados muestran que para la primera hipótesis  $r=0$  y alternativa  $r=1$ , se acepta la hipótesis alternativa de cointegración: el estadístico es mayor al valor critico del 10 porciento,  $31.01 > 17.85$ . Para la segunda hipótesis  $r\leq 1$  y alternativa  $r=2$ , se acepta la hipótesis alternativa nula de dos vector de cointegración. Por lo que se pueden identificar los vectores de cointegración  $[1 \ -1.25 \ 4.55]$  y  $[1 \ -1.41 \ 7.19]$ .

```
# Prueba de Johansen de la traza con tendencia
```

```
summary(ca.jo(ecb.consumo, type="trace", ecdet="trend", spec=c("longrun"), K=4))
```

```
#####
```

```
# Johansen-Procedure #
```

```
#####
```

Test type: trace statistic , with linear trend in cointegration

Eigenvalues (lambda):

```
[1] 3.007210e-01 6.711864e-02 3.632077e-18
```

Values of teststatistic and critical values of test:

	test	10pct	5pct	1pct
r <= 1		5.98	10.49	12.25
r = 0		36.74	22.76	25.32
				30.45

Eigenvectors, normalised to first column:

(These are the cointegration relations)

	lcp_mex.l4	lpib_mex.l4	trend.l4
lcp_mex.l4	1.000000000	1.000000000	1.000000000
lpib_mex.l4	-2.613552539	-0.740358526	-1.944486294
trend.l4	0.008212396	-0.002968024	0.007611815

Weights W:

(This is the loading matrix)

	lcp_mex.l4	lpib_mex.l4	trend.l4
lcp_mex.d	-0.10989875	-0.3329111	-2.471382e-12
lpib_mex.d	0.03206499	-0.2557477	-4.370999e-13

Al igual que en el caso de la opción sin tendencia y constante, se acepta la hipótesis de cointegración con solo vector. El vector de cointegración es [1 -2.61 0.008].

## REFERENCIAS

- Engle, Robert F. y C. W. J. Granger (1987). "Co-integration and Error Correction: Representation, Estimation, and Testing," *Econometrica*, 55, 251–276.
- Hamilton, James D. (1994). *Time Series Analysis*, Princeton University Press
- Johansen, Søren y Katarina Juselius (1990). "Maximum Likelihood Estimation and Inferences on Cointegration—with applications to the demand for money," *Oxford Bulletin of Economics and Statistics*, 52, 169–210.
- Johansen, Søren (1991). "Estimation and Hypothesis Testing of Cointegration Vectors in Gaussian Vector Autoregressive Models," *Econometrica*, 59, 1551–1580.
- Johansen, Søren (1995). *Likelihood-based Inference in Cointegrated Vector Autoregressive Models*, Oxford: Oxford University Press.
- MacKinnon, J.G. (1996) Numerical distribution functions for unit root and cointegration tests, *Journal of Applied Econometrics* 11, 601–618.
- Phillips, P.C.B. and Ouliaris, S. (1990), Asymptotic Properties of Residual Based Tests for Cointegration, *Econometrica*, Vol. 58, No. 1, 165–193.
- Pfaff, B. (2008) *Analysis of Integrated and Cointegrated Time Series with R*. Second Edition. Springer, New York. ISBN 0-387-27960-1

## ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO

Cap10.Cointegracion.R

Consumo.csv

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap10

Práctica\_Cap10

VideoPráctica\_Cap10

VideoTeoría\_Cap10

# CAPÍTULO 11: MODELOS VAR

Javier Galán Figueroa

## 1. INTRODUCCIÓN

Una de las principales tareas del economista en su quehacer diario, es la evaluación de las políticas económicas que son llevadas a cabo por un determinado gobierno para satisfacer sus compromisos adquiridos con anterioridad con el menor costo social. De acuerdo a la literatura de la teoría de juegos, toda política económica que cumple con lo anterior se dice que es creíble ya que los agentes tienen pleno conocimiento sobre las acciones de la autoridad, restringiendo así a la autoridad en caer en un problema de inconsistencia dinámica obligándolo en alcanzar sus objetivos de política de corto y largo plazos (Kydland y Prescott, 1977; Barro y Gordon, 1983).

A nivel empírico los economistas han acudido a los modelos de Vectores Autorregresivos, VAR, como herramienta básica para evaluar las políticas económicas (Galán y Venegas, 2013 y Galán, 2014). Esta metodología econométrica fue planteada inicialmente por el célebre trabajo del Nobel en Economía 2011 Christopher Sims (1980) *Macroeconomic and Reality*, en donde presenta una fuerte crítica hacia los modelos de sistemas ecuaciones y sus principales aplicaciones como son los modelos macroeconómicos o de gran escala.

Sims menciona que la mayor parte de las restricciones que aparecen en los modelos son falsas debido a *i)* no hay conocimiento suficiente en la teoría económica para clasificar a las variables en endógenas y exógenas y *ii)* a priori no se puede establecer restricciones cero. De esta manera Sims propuso el modelo VAR, un sistema de ecuaciones autorregresivas, en donde las variables utilizadas

no se distinguen si son endógenas o exógenas ya que se asume que cada una afecta y es afectada por las demás.

## 2. CARACTERÍSTICAS DEL MODELO VAR

El modelo VAR desarrollado por Sims (1980 y 1986) ha tenido gran popularidad al ser una herramienta muy útil para el análisis empírico de las series de tiempo económicas ya que tiene las siguientes propiedades: *i)* parte de un enfoque ateórico, *ii)* es capaz de separar los efectos pasados que explican al vector de las variables endógenas a través de su pasado o mediante variables autorregresivas. Esto se ilustra de la siguiente manera: un vector autorregresivo de orden uno, VAR(1), se tiene su forma primitiva (Enders, 2010)

$$\begin{aligned} y_t &= b_{10} - b_{12}z_t + \gamma_{11}y_{t-1} + \gamma_{12}z_{t-1} + \varepsilon_{yt} \\ z_t &= b_{20} - b_{21}y_t + \gamma_{21}y_{t-1} + \gamma_{22}z_{t-1} + \varepsilon_{zt} \end{aligned} \quad (1)$$

ó

$$\begin{pmatrix} 1 & b_{12} \\ b_{21} & 1 \end{pmatrix} \begin{pmatrix} y_t \\ z_t \end{pmatrix} = \begin{pmatrix} b_{10} \\ b_{20} \end{pmatrix} + \begin{pmatrix} \gamma_{11} & \gamma_{12} \\ \gamma_{21} & \gamma_{22} \end{pmatrix} \begin{pmatrix} y_{t-1} \\ z_{t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{yt} \\ \varepsilon_{zt} \end{pmatrix} \quad (2)$$

equivalentemente

$$Bx_t = \Gamma_0 + \Gamma_1 x_{t-1} + \varepsilon_t \quad (3)$$

donde el vector  $x_t$  agrupa las variables endógenas, la matriz  $B$  contiene los coeficientes de los efectos contemporáneos del vector  $x_t$ , mientras la matriz  $\Gamma_1$  contiene los coeficientes de los efectos pasados sobre  $x_t$ , por último el vector

□ contiene los efectos estocásticos que afectan a las variables del vector  $x_t$ . A partir de la expresión (3), se obtiene la forma estándar:

$$x_t = \Pi_0 + \Pi_1 x_{t-1} + e_t \quad (4)$$

donde

$$\Pi_0 = B^{-1}\Gamma_0, \quad \Pi_1 = B^{-1}\Gamma_1 \quad y \quad e_t = B^{-1}\varepsilon_t.$$

El término  $e_t$  es un componente residual y es lo que hace la diferencia con la expresión (3). Por otro lado se supone que se cumple la descomposición de Wold donde las variables endógenas del VAR( $p$ ) al cumplir el supuesto de estacionariedad<sup>24</sup> (o ser débilmente estacionarias) es posible invertir la expresión (4) en un vector de medias móviles, VMA( $\infty$ ), permitiendo con ello visualizar a través de la matriz de los multiplicadores de impacto de corto y largo plazo (o funciones impulso respuesta) cómo los choques estocásticos afectan la trayectoria del vector de las variables endógenas, este último aspecto se puede apreciar en las siguientes expresiones:

$$\begin{pmatrix} y_t \\ x_t \end{pmatrix} = \begin{pmatrix} \bar{y} \\ \bar{x} \end{pmatrix} + \sum_{i=0}^{\infty} \begin{pmatrix} \phi_{11}(i) & \phi_{12}(i) \\ \phi_{21}(i) & \phi_{22}(i) \end{pmatrix} \begin{pmatrix} \varepsilon_{yt-i} \\ \varepsilon_{xt-i} \end{pmatrix} \quad (5)$$

ó

$$x_t = \mu + \sum_{i=0}^{\infty} \phi_i \varepsilon_{t-i} \quad (6)$$

---

<sup>24</sup> De acuerdo con Lütkepohl (2005), las variables que comprenden al VAR( $p$ ) son al menos  $I(1)$ .

donde  $\sum_{i=0}^n \phi_{12}(i)$  es el multiplicador de impacto, mientras que  $\sum_{i=0}^{\infty} \phi_{jk}^2(i)$  es el multiplicador total o de largo plazo.

### 3. UN CASO PARA LA ECONOMÍA MEXICANA

Para ejemplificar la estimación de un modelo de Vectores Autorregresivos mediante el programa *RStudio*, se considera un VAR bivariado en donde se considera analizar el comportamiento de la inflación y de la oferta de dinero para el caso de la economía mexicana tomando el periodo del primero mes del año 2000 al cuarto mes del 2014. Cabe mencionar que los datos se obtuvieron del Sistema de Información Estadística del Banco de México.

Posteriormente se aplica logaritmos a las series, para ello se utiliza la base de datos que se encuentra en el archivo *base\_var\_inflacion.csv*. Con esta base se crea el siguiente objeto, *mex\_var*, que será utilizado para la estimación del modelo VAR(p).

```
> mex_var<-read.csv("C:/data/base_var_inflacion.csv", header=T)
> attach(mex_var)
```

A continuación se hace la lista de la base para conocer cómo se encuentra asignado el nombre de las variables y de su ubicación, para ello se utiliza el nombre del objeto que se está trabajando, *mex\_var*.

	M2	INPC
1	28.7740	59.8083
2	29.3109	60.3388
3	29.8290	60.6734
4	30.1613	61.0186
5	30.4289	61.2467
6	30.9230	61.6095
7	31.6270	61.8498
8	31.7013	62.1896
9	32.0940	62.6439

Una vez que se tiene el objeto de trabajo, se procede a dar formato de series de tiempo a la base de datos a cabo por mediante el siguiente código. Comenzando primero por el índice de la oferta monetaria y posteriormente al índice de precios.

```
# Para M2
> tm2=ts(mex_var[,1], start=2000, freq=12)

# Para INPC
> tp=ts(mex_var[,2], start=2000, freq=12)
```

A estas nuevas variables son transformadas en logaritmo mediante el siguiente código:

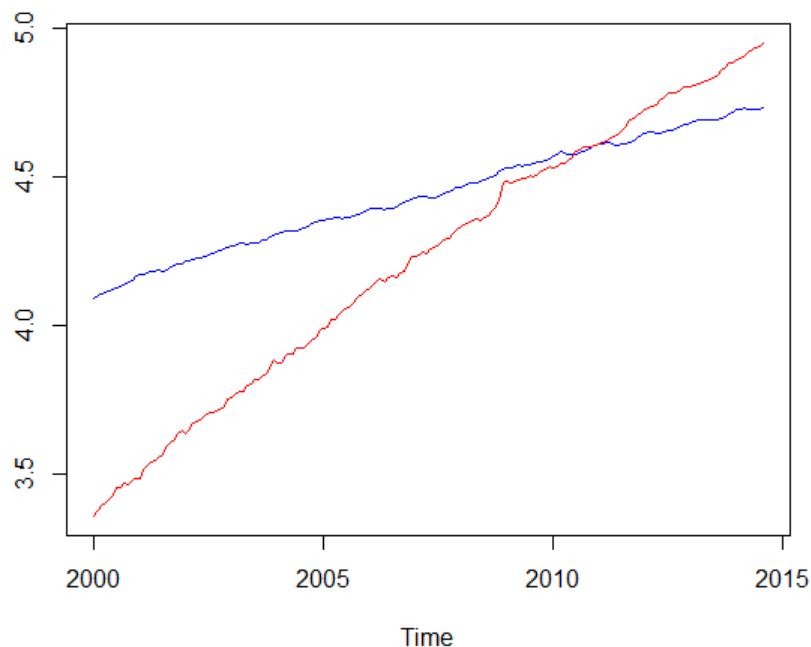
```
# Para M2
> ltm2<-log(tm2)

# Para INPC
> ltp<-log(tp)
```

Una vez que se han transformado en logaritmo las variables, se grafican siguiendo el código siguiente:

```
> ts.plot(ltp, ltm2, col=c("blue", "red"))
```

**Gráfico 1**  
**México: logaritmo del INPC y el logaritmo del índice de M2**  
**(2000:01-2014:04)**



Del anterior gráfico se aprecia tanto el INPC como el índice M2 presenta una trayectoria determinística creciente, por lo que ambas series no satisfacen el supuesto de ruido blanco o que son estacionarias. Para corroborarlo se llevará a continuación las pruebas de raíz unitaria. Para ello se instalará la paquetería de los vectores autorregresivos, vars, posteriormente se activa la librería respectiva.

```
> install.packages("vars")
> library("vars")
```

Para aplicar la prueba de raíz unitaria de Dickey Fuller Aumentada (ADF) se plantea la siguiente Hipótesis nula vs. Hipótesis alternativa:

Ho: La variable x no tiene una raíz unitaria  
Ha: La variable x tiene una raíz unitaria

De esta forma se aplica la prueba ADF sin constante ni tendencia mediante el código siguiente:

```
> adf1_ltp<-summary(ur.df(ltp, lags=1))
> adf1_ltp
```

**Tabla1**  
**Prueba Dickey-Fuller Aumentada para el logaritmo del INPC**  
**Sin constante y tendencia**

```
#####
# Augmented Dickey-Fuller Test Unit Root Test #
#####

Test regression none
```

```

Call:
Im(formula = z.diff ~ z.lag.1 - 1 + z.diff.lag)

Residuals:
    Min     1Q   Median     3Q    Max 
-0.0093954 -0.0015285  0.0003127  0.0018731  0.0069307 

Value of test-statistic is: 5.798

Critical values for test statistics:
    1pct 5pct 10pct 
tau1 -2.58 -1.95 -1.62

```

En la parte inferior de la Tabla 1 se aprecia que el valor estadístico ADF tiene un valor de **5.5798**, el hecho de que sea positivo, indica que el logaritmo del INPC no es estacionario, por lo que se rechaza la hipótesis nula y se acepta la alternativa de que la variable presenta raíz unitaria. Se continua aplicando la prueba ADF para incluyendo la constante y la tendencia respectivamente mediante la siguiente rutina.

```

# La prueba ADF con termino constante o con derivada
> adf2_ltp<-summary(ur.df(ltp, type="drift", lags=12))
> adf2_ltp

# La prueba ADF con tendencia
> adf3_ltp<-summary(ur.df(ltp, type="trend", lags=1))
> adf3_ltp

```

De acuerdo a la Tabla 2 tanto el logaritmo del INPC como el logaritmo del índice de M2, son estacionarias mediante la segunda diferencia, para obtener la primera y segunda diferencia a las series se utiliza la siguiente rutina:

```

# Primera diferencia del logaritmo del indice de precios
> dltp<-diff(ltp)

# Segunda diferencia del logaritmo del indice de precios
> d2ltp<-diff(dltp)

> # Primera diferencia del logaritmo del indice de M2
> dltm2<-diff(ltm2)

> # Segunda diferencia del logaritmo del indice de M2
> d2ltm2<-diff(dltm2)

```

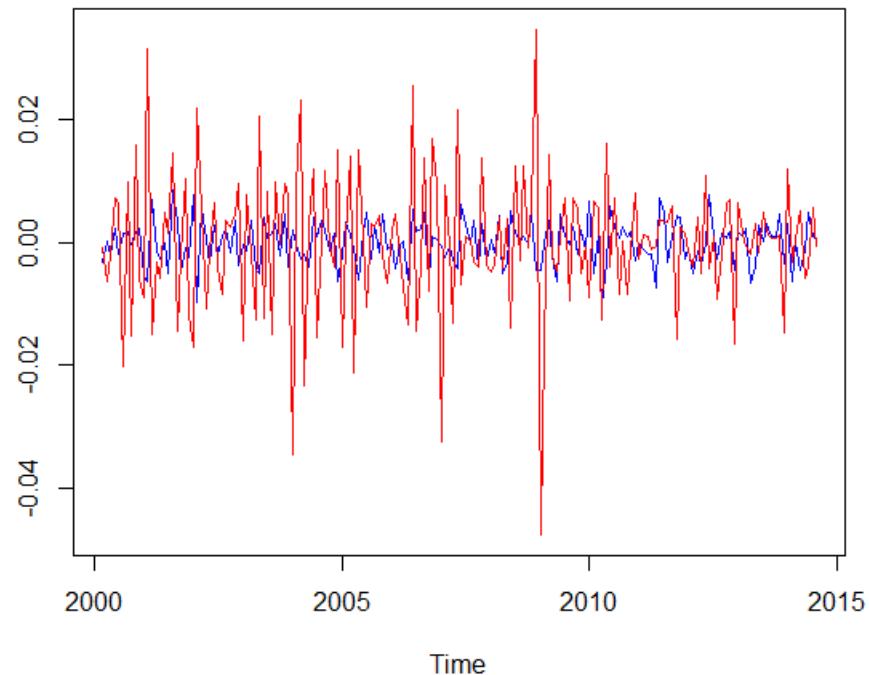
**Tabla 2**  
**Prueba de raíz unitaria Dickey-Fuller Aumentada, ADF**

Variable	Término determinístico	Rezagos	Valor de la prueba	Valor crítico		
				1%	5%	10%
log(INPC)	sin constante y tendencia	1	5.79	-2.58	-1.95	-1.62
	constante	1	-1.28	-3.46	-2.88	-2.57
	constante y tendencia	1	-4.06	-3.99	-3.43	-3.13
$\Delta(\log(\text{INPC}))$	sin constante y tendencia	11	-1.49	-2.58	-1.95	-1.62
	constante	11	-3.89	-3.46	-2.88	-2.57
	constante y tendencia	11	-3.85	-3.99	-3.43	-3.13
$\Delta^2(\log(\text{INPC}))$	sin constante y tendencia	12	-8.11	-2.58	-1.95	-1.62
	constante	10	-12.53	-3.46	-2.88	-2.57
	constante y tendencia	10	-12.57	-3.99	-3.43	-3.13
log(M2)	sin constante y tendencia	1	8.75	-2.58	-1.95	-1.62
	constante	1	-2.13	-3.46	-2.88	-2.57
	constante y tendencia	0	-2.09	-3.99	-3.43	-3.13
$\Delta\log(\text{M2})$	sin constante y tendencia	11	-0.76	-2.58	-1.95	-1.62
	constante	0	-12.62	-3.46	-2.88	-2.57
	constante y tendencia	3	-7.44	-3.99	-3.43	-3.13
$\Delta^2\log(\text{M2})$	sin constante y tendencia	3	-10.57	-2.58	-1.95	-1.62
	constante	3	-10.54	-3.46	-2.88	-2.57
	constante y tendencia	3	-10.51	-3.99	-3.43	-3.13

Graficando las series las segundas diferencias del logaritmo del índice de precios, d2ltp, y el logaritmo del índice de la oferta de dinero, d2ltm2, mediante el siguiente código se obtiene la Gráfica 2.

```
> ts.plot(d2ltp, d2ltm2, col=c("blue", "red"))
```

**Gráfico 2**  
**Segunda diferencia del logaritmo del INPC y del logaritmo del índice de M2**  
**(2000:01-2014:04)**



A continuación se lleva a cabo las pruebas de causalidad en el sentido de Granger para determinar el orden causal entre las variables, para ello se activa la librería *Imtest*. Una vez activa la librería, se aplica de la siguiente manera: *i)* Se verifica la dirección de la causalidad de la oferta de dinero hacia los precios, donde la hipótesis nula implica que la oferta de dinero no causa en el sentido de Granger a los precios:

```
> grangertest(d2ltp~d2ltm2, order=1)
```

**Tabla 3**  
**Causalidad en el sentido de Granger**  
**de la oferta de dinero a los precios**

```
Granger causality test
```

```
Model 1: d2ltp ~ Lags(d2ltp, 1:1) + Lags(d2ltm2, 1:1)
```

```
Model 2: d2ltp ~ Lags(d2ltp, 1:1)
```

Res.Df	Df	F	Pr(>F)
1	170		
2	171	-1	0.8796

Res.Df	Df	F	Pr(>F)
1	170		
2	171	-1	0.8796

De acuerdo a la Tabla 3 se acepta la hipótesis nula que la oferta de dinero no causa en el sentido de Granger a los precios, dado que el *p-value*, 0.3496, es mayor al valor crítico del 5 por ciento.

*ii)* Se verifica la dirección de la causalidad de los precios hacia la oferta de dinero, donde la hipótesis nula implica que los precios no causan en el sentido de Granger a la oferta de dinero.

```
> grangertest(d2ltm2~d2ltp, order=1)
```

**Tabla 4**  
**Causalidad en el sentido de Granger**  
**de la oferta de dinero a los precios**

Granger causality test			
Model 1: d2ltm2 ~ Lags(d2ltm2, 1:1) + Lags(d2ltp, 1:1)			
Model 2: d2ltm2 ~ Lags(d2ltm2, 1:1)			
Res.Df	Df	F	Pr(>F)
1	170		
2	171	-1 0.0909	<b>0.7634</b>

De acuerdo a la Tabla 4 se acepta la hipótesis nula que los precios no causa en el sentido de Granger a la oferta de dinero, dado que el *p-value*, 0.7634, es mayor al valor crítico del 5 por ciento. A continuación aplica la prueba para varios rezagos lo que permite construir la Tabla 5.

**Tabla 5**  
**Causalidad en el sentido de Granger**

Hipótesis nula:	M2 no causa al INPC		INPC no causa al M2	
No. de Rezago	F-statistic	P-value	F-statistic	P-value
Rezago 1	0.879	0.349	0.090	0.763
Rezago 2	0.421	0.656	0.004	0.995
Rezago 3	0.298	0.826	0.254	0.858
Rezago 4	2.710	0.032	0.298	0.878
Rezago 5	2.482	0.033	0.360	0.875
Rezago 6	2.729	0.015	0.623	0.711
Rezago 7	2.468	0.019	1.813	0.088
Rezago 8	2.288	0.024	1.733	0.095
Rezago 9	2.023	0.040	1.542	0.138
Rezago 10	2.158	0.023	1.760	0.073
Rezago 11	1.338	0.209	1.044	0.411
Rezago 12	1.249	0.256	1.012	0.441

De acuerdo a la Tabla 5 se rechaza la hipótesis nula que M2 no causa en el sentido de Granger al INPC en los siguientes rezagos: del 4 al 10. En la misma tabla se acepta la hipótesis nula para todos los rezagos de que el INPC no causa en el sentido de Granger a M2. Con estos resultados se concluye que la dirección causal entre estas variables va de M2 a INPC. A continuación se activa la librería *vars* para identificar el VAR, a través de los criterios de información: Akaike (AIC), Hanna-Quin (HQ), Schwarz (SC) y Error de Predicción Final (FPE) para ello se utiliza los siguientes códigos:

```
> library(vars)
> VARselect(mex_var2, lag.max=12)
```

Antes de la identificación, se debe crear un nuevo objeto en donde contenga las variables transformadas y que sean estacionarias, en este ejemplo el nuevo objeto

se le nombró *mex\_var2*, cabe mencionar, las variables del nuevo objeto se deberán transformar como series de tiempo siguiendo los pasos iniciales del presente capítulo. Para identificar el orden del VAR, en este caso, se utilizó como máximo 12 rezagos y de acuerdo a los criterios de información (AIC, HQ, SC y FPE) de la Tabla 6 indican que se debe utilizar 11 rezagos para la estimación del modelo VAR.

**Tabla 6**  
**Criterio de Rezagos Óptimos del VAR**

\$selection
AIC(n) HQ(n) SC(n) FPE(n)
11 11 2 11
\$criteria
1 2 3 4 5
AIC(n) -2.046463e+01 -2.068196e+01 -2.065764e+01 -2.075149e+01 -2.075348e+01
HQ(n) -2.041820e+01 -2.060458e+01 -2.054931e+01 -2.061220e+01 -2.058323e+01
SC(n) -2.035027e+01 -2.049137e+01 -2.039081e+01 -2.040842e+01 -2.033417e+01
FPE(n) 1.295173e-09 1.042211e-09 1.067940e-09 9.723977e-10 9.706468e-10
6 7 8 9 10
AIC(n) -2.090388e+01 -2.097478e+01 -2.104458e+01 -2.103653e+01 -2.109132e+01
HQ(n) -2.070268e+01 -2.074263e+01 -2.078147e+01 -2.074247e+01 -2.076631e+01
SC(n) -2.040834e+01 -2.040300e+01 -2.039656e+01 -2.031228e+01 -2.029083e+01
FPE(n) 8.353389e-10 7.784501e-10 7.263245e-10 7.326458e-10 6.941159e-10
11 12
AIC(n) -2.130916e+01 -2.130390e+01
HQ(n) -2.095320e+01 -2.091699e+01
SC(n) -2.043244e+01 -2.035094e+01
FPE(n) 5.587598e-10 5.623294e-10

Una vez identificado el VAR se procede con el siguiente código su estimación, VAR. En Tabla 7 se presentan las ecuaciones del VAR de orden 11, VAR(11).

```
> var1<-VAR(mex_var2,p=11)
> var1
```

**Tabla 7**  
**Ecuaciones Estimadas del VAR(11)**

**VAR Estimation Results:**

```
=====
```

**Estimated coefficients for equation d2lm2:**

```
=====
```

Call:

d2lm2 = d2lm2.l1 + d2lp.l1 + d2lm2.l2 + d2lp.l2 + d2lm2.l3 + d2lp.l3 + d2lm2.l4 + d2lp.l4 + d2lm2.l5 + d2lp.l5 +  
d2lm2.l6 + d2lp.l6 + d2lm2.l7 + d2lp.l7 + d2lm2.l8 + d2lp.l8 + d2lm2.l9 + d2lp.l9 + d2lm2.l10 + d2lp.l10 + d2lm2  
.l11 + d2lp.l11 + const

d2lm2.l1	d2lp.l1	d2lm2.l2	d2lp.l2	d2lm2.l3
-7.629557e-01	-1.714439e-01	-8.434147e-01	1.109548e-01	-6.063436e-01
d2lp.l3	d2lm2.l4	d2lp.l4	d2lm2.l5	d2lp.l5
-6.403225e-03	-6.978840e-01	-6.657361e-02	-6.033918e-01	-1.467579e-01
d2lm2.l6	d2lp.l6	d2lm2.l7	d2lp.l7	d2lm2.l8
-5.911774e-01	1.071385e-01	-4.820848e-01	-5.209136e-01	-4.042082e-01
d2lp.l8	d2lm2.l9	d2lp.l9	d2lm2.l10	d2lp.l10
5.843460e-02	-2.957711e-01	-2.264212e-01	-2.607287e-01	-3.519290e-01
d2lm2.l11	d2lp.l11	const		
-1.952367e-01	-4.569120e-02	-1.772894e-05		

**Estimated coefficients for equation d2lp:**

```
=====
```

Call:

```

d2lp = d2lm2.l1 + d2lp.l1 + d2lm2.l2 + d2lp.l2 + d2lm2.l3 + d2lp.l3 + d2lm2.l4 + d2lp.l4 + d2lm2.l5 + d2lp.l5 + d
2lm2.l6 + d2lp.l6 + d2lm2.l7 + d2lp.l7 + d2lm2.l8 + d2lp.l8 + d2lm2.l9 + d2lp.l9 + d2lm2.l10 + d2lp.l10 + d2lm2.l
11 + d2lp.l11 + const

d2lm2.l1    d2lp.l1    d2lm2.l2    d2lp.l2    d2lm2.l3
-0.0153917990 -0.5768874943  0.0225318889 -0.6496238436  0.0258938749

d2lp.l3    d2lm2.l4    d2lp.l4    d2lm2.l5    d2lp.l5
-0.6720337913  0.0795079412 -0.6540234065  0.0019054572 -0.5936082650

d2lm2.l6    d2lp.l6    d2lm2.l7    d2lp.l7    d2lm2.l8
0.0257629817 -0.7731698049 -0.0057926931 -0.6635449616 -0.0051974058

d2lp.l8    d2lm2.l9    d2lp.l9    d2lm2.l10   d2lp.l10
-0.6685683626  0.0220922034 -0.5278803727  0.0046914476 -0.4810432531

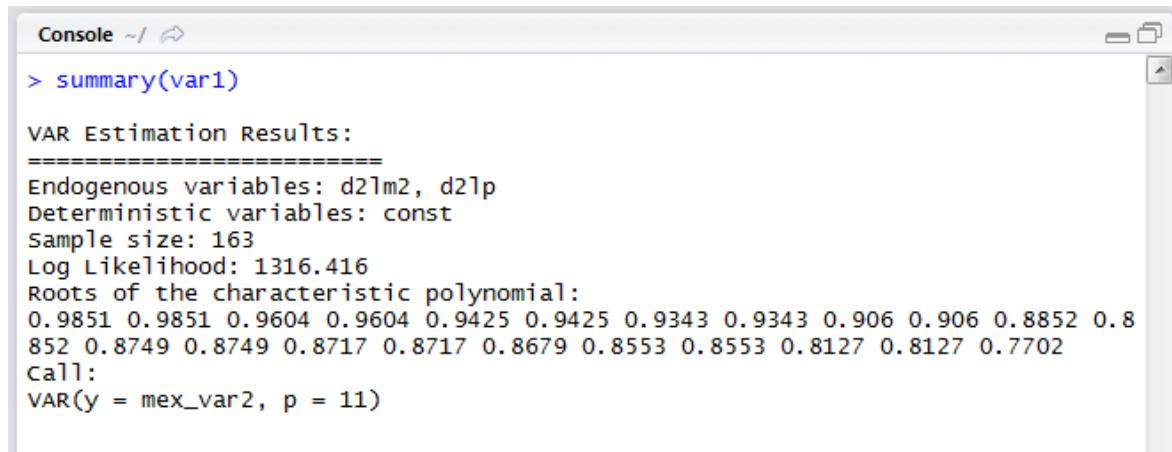
d2lm2.l11   d2lp.l11   const
0.0081195573 -0.4276814667 -0.0001442565

```

Saber si el VAR estimado satisface la condición de estabilidad, se utiliza el código *summary* para que el programa R muestre las raíces del polinomio característico así como los estadísticos necesarios para llevar a cabo la inferencia estadística. De acuerdo a la Tabla 8, se tiene que las raíces del polinomio característico son menores a uno, por lo que el VAR estimado con 11 rezagos satisface la condición de estabilidad.

```
> summary(var1)
```

**Tabla 8**  
**Raíces del Polinomio Característico del VAR Estimado**



```
Console ~/ 
> summary(var1)

VAR Estimation Results:
=====
Endogenous variables: d2lm2, d2lp
Deterministic variables: const
Sample size: 163
Log Likelihood: 1316.416
Roots of the characteristic polynomial:
0.9851 0.9851 0.9604 0.9604 0.9425 0.9425 0.9343 0.9343 0.906 0.906 0.8852 0.852 0.8749 0.8749 0.8717 0.8717 0.8679 0.8553 0.8553 0.8127 0.8127 0.7702
Call:
VAR(y = mex_var2, p = 11)
```

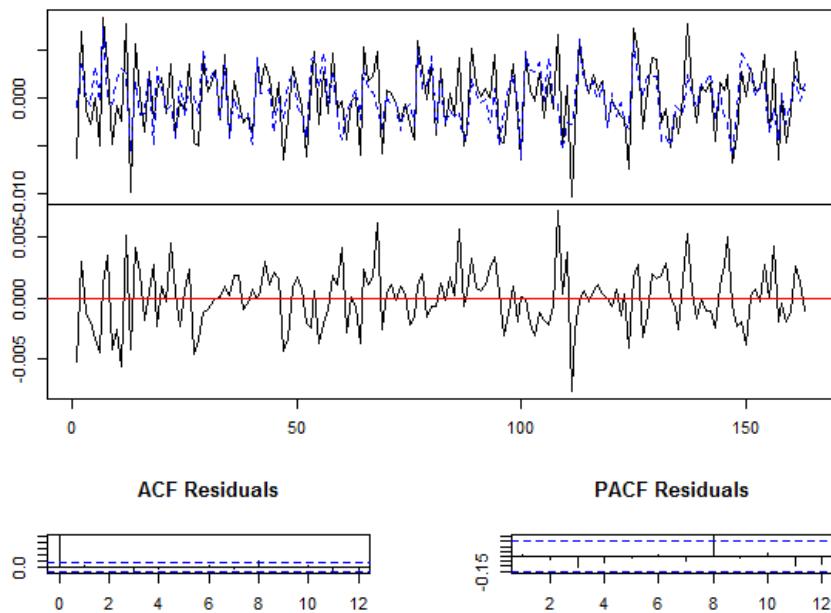
Una vez que se tiene el sistema de ecuaciones del VAR estimados, el usuario puede obtener el grafico de la variable observada versus la estimada, así como de los residuales a través del código *plot*.



```
> plot(var1)
```

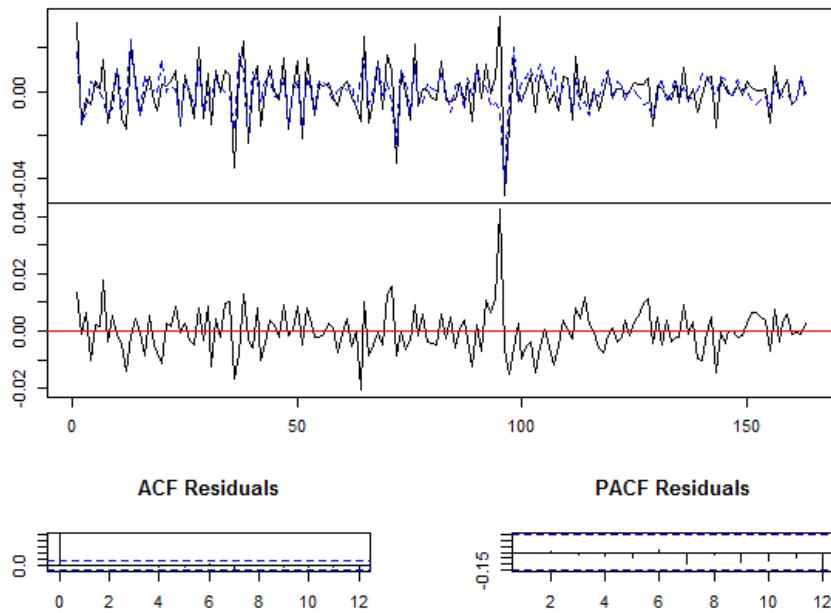
**Grafica 3a**

Diagram of fit and residuals for d2lp



### Grafica 3b

Diagram of fit and residuals for d2lm2



En las Grafica 3a y 3b se muestra el comportamiento de las variables observadas y estimadas así como de los residuales. A continuación se lleva a cabo las pruebas de especificación. En primer lugar se verifica la existencia o no de autocorrelación serial en los residuales mediante el siguiente código:

```
> seriala<-serial.test(var1, lags.pt=11, type="PT.asymptotic")
> seriala$serial
```

**Tabla 9**  
**Prueba de Autocorrelación**

Portmanteau Test (asymptotic)
data: Residuals of VAR object var1
Chi-squared = 27.2801, df = 0, p-value < 2.2e-16

De acuerdo a la prueba de autocorrelación, Tabla 9, permite rechazar la hipótesis nula de que los residuales no están correlacionados y aceptar la hipótesis alterna que hay presencia de correlación serial entre los residuales. Con el siguiente código se verifica si los residuales del modelo VAR estimado los residuales se distribuyen como una normal.

```
> normalidad<-normality.test(var1)
> normalidad$jb.mul
```

**Tabla 10**  
**Prueba de Normalidad en los Residuales**

```
$JB
```

```
JB-Test (multivariate)
```

```
data: Residuals of VAR object var1
```

```
Chi-squared = 201.0673, df = 4, p-value < 2.2e-16
```

```
$Skewness
```

```
Skewness only (multivariate)
```

```
data: Residuals of VAR object var1
```

```
Chi-squared = 25.2089, df = 2, p-value = 3.357e-06
```

```
$Kurtosis
```

```
Kurtosis only (multivariate)
```

```
data: Residuals of VAR object var1
```

```
Chi-squared = 175.8584, df = 2, p-value < 2.2e-16
```

Con la Tabla 10 se tiene evidencia que los residuales no se distribuyen como normal ya que presenta problemas en la curtosis y en el sesgo. Ahora se prosigue con verificar si la varianza de los residuales son o no homocedasticos, por lo que se utiliza el siguiente código:

```
> arch1<-arch.test(var1, lags.multi=11)  
> arch1$arch.mul
```

**Tabla 11**  
**Prueba de Heteroscedasticidad o Varianza Constante**

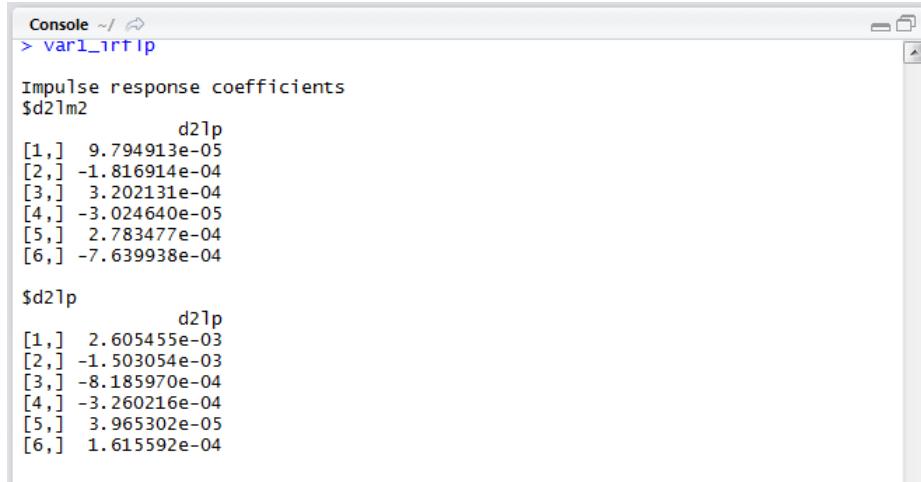
ARCH (multivariate)
data: Residuals of VAR object var1
Chi-squared = 80.1774, df = 99, p-value = 0.917

A través de la prueba de Heteroscedasticidad se tiene que los residuales si satisfacen el supuesto de varianza constante. El modelo VAR(11) estimado tiene problemas de especificación de autocorrelación y de normalidad. Observando las Gráficas 3a y 3b se puede encontrar que existe un punto de ruptura o de cambio estructural, esto implica que se puede introducir una variable *dummy* para ajustar el modelo, en donde los valores de uno indiquen que en ese periodo existe un cambio estructural, mientras los valores cero se refiere la ausencia de cambio estructural.

Una vez que se han llevado a cabo las pruebas de especificación y verificar que el modelo VAR las satisface, en su caso corregirlo, se procede a realizar el análisis impulso respuesta permitiendo con ello observar la trayectoria de la variable de estudio. Para obtener lo anterior se recurre al siguiente código y así obtener los multiplicadores de impacto que se muestran en la Tabla 12.

```
> var1_irfpl<-irf(var1, response="d2lp", n.ahead=8, boot=TRUE)
> var1_irfpl
```

**Tabla 12**  
**Impulso Respuesta de los Precios ante una Innovación de la Oferta**  
**Monetaria**



```
Console ~/ ~
> var1_irflp

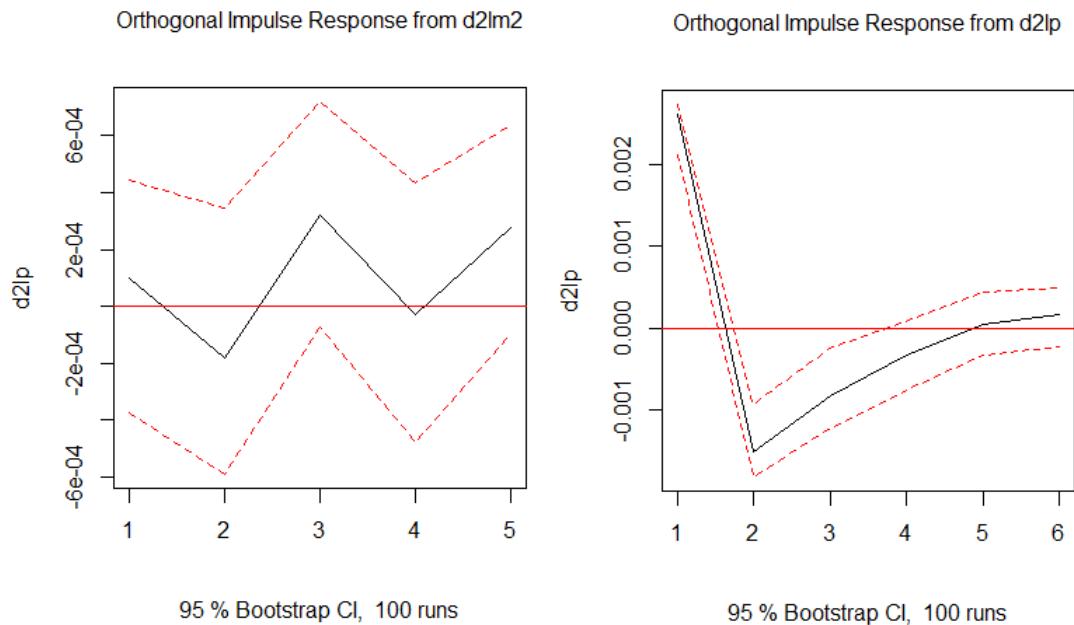
Impulse response coefficients
$d21m2
d21p
[1,]  9.794913e-05
[2,] -1.816914e-04
[3,]  3.202131e-04
[4,] -3.024640e-05
[5,]  2.783477e-04
[6,] -7.639938e-04

$d21p
d21p
[1,]  2.605455e-03
[2,] -1.503054e-03
[3,] -8.185970e-04
[4,] -3.260216e-04
[5,]  3.965302e-05
[6,]  1.615592e-04
```

Para graficar el impulso respuesta este se obtiene con el código siguiente y se muestra en la Gráfica 4. Si el usuario desea continuar graficando las demás variables se repite el proceso para cada variable que se encuentra en el modelo VAR.

```
> plot(var1_irflp)
```

**Gráfica 4**  
**Impulso Respuesta de los Precios ante una Innovación de la Oferta Monetaria**



```
> var1_irflm2<-irf(var1, response="d2lm2", n.ahead=5, boot=TRUE)  
> var1_irflm2
```

**Tabla 13**  
**Impulso Respuesta de la Oferta Monetaria ante una Innovación de los Precios**

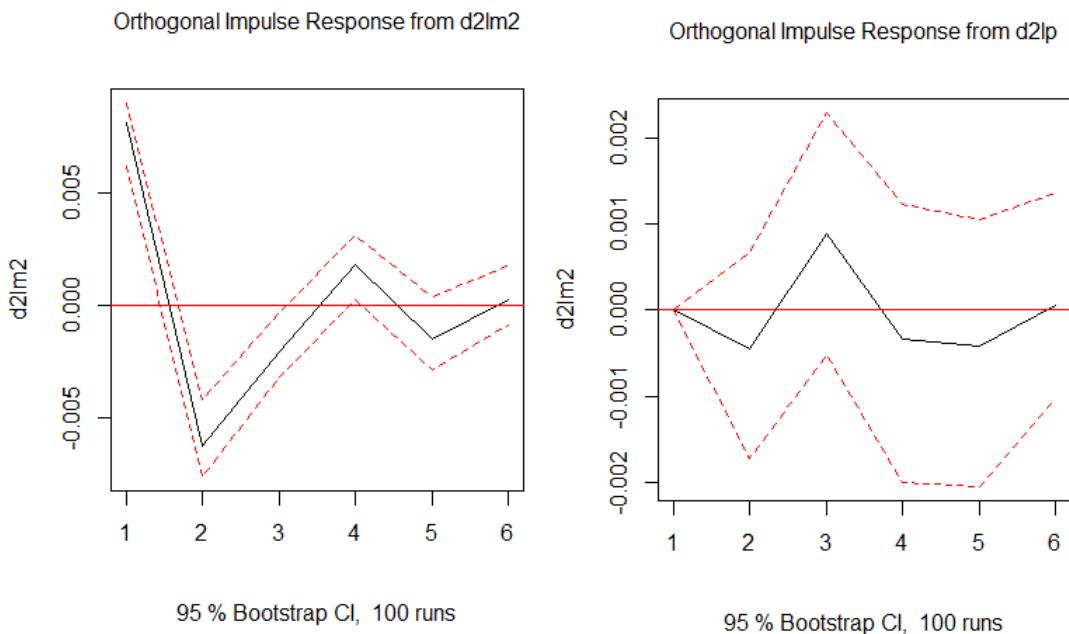
```
Console ~/ ↗
Impulse response coefficients
$d2lm2
d2lm2
[1,] 0.008133275
[2,] -0.006222122
[3,] -0.002070502
[4,] 0.001820286
[5,] -0.001510488
[6,] 0.000251976

$d2lp
d2lp
[1,] 0.000000e+00
[2,] -4.466895e-04
[3,] 8.875815e-04
[4,] -3.435518e-04
[5,] -4.144001e-04
[6,] 5.944598e-05
```

```
> plot(var1_irflm2)
```

**Gráfica 5**

## **Impulso Respuesta de la Oferta Monetaria ante una Innovación de los Precios**



Por último, se utiliza el siguiente código para obtener el análisis de la descomposición de varianza que se presenta en las Tabla 13 y 14.

```
> var1_fevd_d2lp<-fevd(var1, n.ahead=50)$d2lp  
> var1_fevd_d2lp
```

Tabla 14

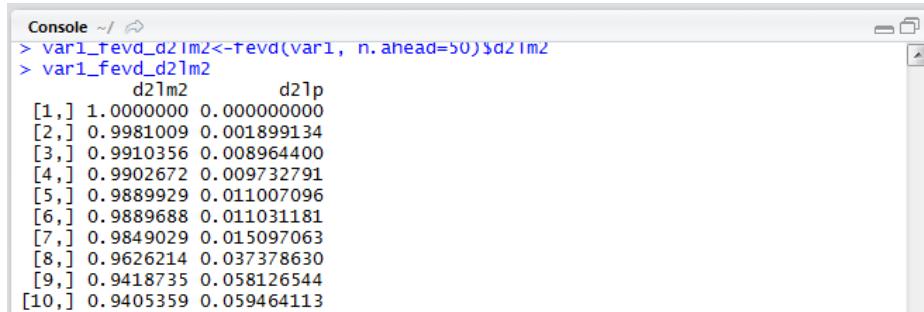
## Descomposición de la Varianza ante una Innovación por parte de los Precios

```
Console ~/ 
> var1_fevd_d2lp<-fevd(var1, n.ahead=50)$d2lp
> var1_fevd_d2lp
      d2lm2      d2lp
[1,] 0.001411304 0.9985887
[2,] 0.004687015 0.9953130
[3,] 0.014716108 0.9852839
[4,] 0.014649629 0.9853504
[5,] 0.022244307 0.9777557
[6,] 0.075732436 0.9242676
[7,] 0.076200320 0.9237997
[8,] 0.075975769 0.9240242
[9,] 0.076736430 0.9232636
[10,] 0.092811289 0.9071887
```

```
> var1_fevd_d2lm2<-fevd(var1, n.ahead=50)$d2lm2
> var1_fevd_d2lm2
```

**Tabla 15**

**Descomposición de la Varianza ante una Innovación por parte de la Oferta de Dinero**



```
Console ~/ ~
> var1_fevd_d2lm2<-fevd(var1, n.ahead=50)$d2lm2
> var1_fevd_d2lm2
      d2lm2      d2lp
[1,] 1.0000000 0.00000000
[2,] 0.9981009 0.001899134
[3,] 0.9910356 0.008964400
[4,] 0.9902672 0.009732791
[5,] 0.9889929 0.011007096
[6,] 0.9889688 0.011031181
[7,] 0.9849029 0.015097063
[8,] 0.9626214 0.037378630
[9,] 0.9418735 0.058126544
[10,] 0.9405359 0.059464113
```

## REFERENCIAS

Barro, Robert y David Gordon (1983), “Rules, discretion and reputation in a model of monetary policy”, *Journal of Monetary Economy*, vol. 12, núm. 1.

Enders, Walter (2010), *Applied Econometric Time Series*, 3<sup>a</sup>. Ed. John Wiley & Sons, Hoboken, New Jersey.

Galán, Javier y Francisco Venegas (2013), “Evolución de la política monetaria en México: Un análisis var estructural. 2000-2011”, *Revista Nicolaita de Estudios Económicos*, vol. 8, núm. 1.

Galán, Javier (2014), “Christopher Sims: modelos, realidad y metodología”, *Equilibrios y Conjeturas, Cuadernos del Seminario de Credibilidad Macroeconomica*, FE-UNAM, año 1, núm. 1,

Kydland, Finn y Edward Prescott (1977), “Rules rather than discretion: The inconsistency of optimal plans”, *The Journal of Political Economy*, vol. 85, núm. 3.

Sims, Christopher (1980), “Macroeconomics and reality”, *Econometrica*, vol. 48, núm. 1, enero.

\_\_\_\_\_ (1986), “Are forecasting models usable for policy analysis?”, Federal Reserve Bank of Minneapolis, *Quarterly Review*, vol. 10, núm. 1, invierno.

## **ARCHIVO DE DATOS ASOCIADO AL CAPÍTULO**

base\_var\_inflacion.csv

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap11

Práctica\_Cap11

VideoPráctica\_Cap11

VideoTeoría\_Cap11

# CAPÍTULO 12: MODELOS ARCH

Luis Quintana Romero y Miguel Ángel Mendoza

## 1. RIESGO Y VOLATILIDAD

El estudio de fenómenos económicos en los cuales hay una gran volatilidad en las variables ha llevado a poner especial interés en la forma en que tal volatilidad puede ser identificada y separada de otros componentes que influyen en el comportamiento de las variables económicas.

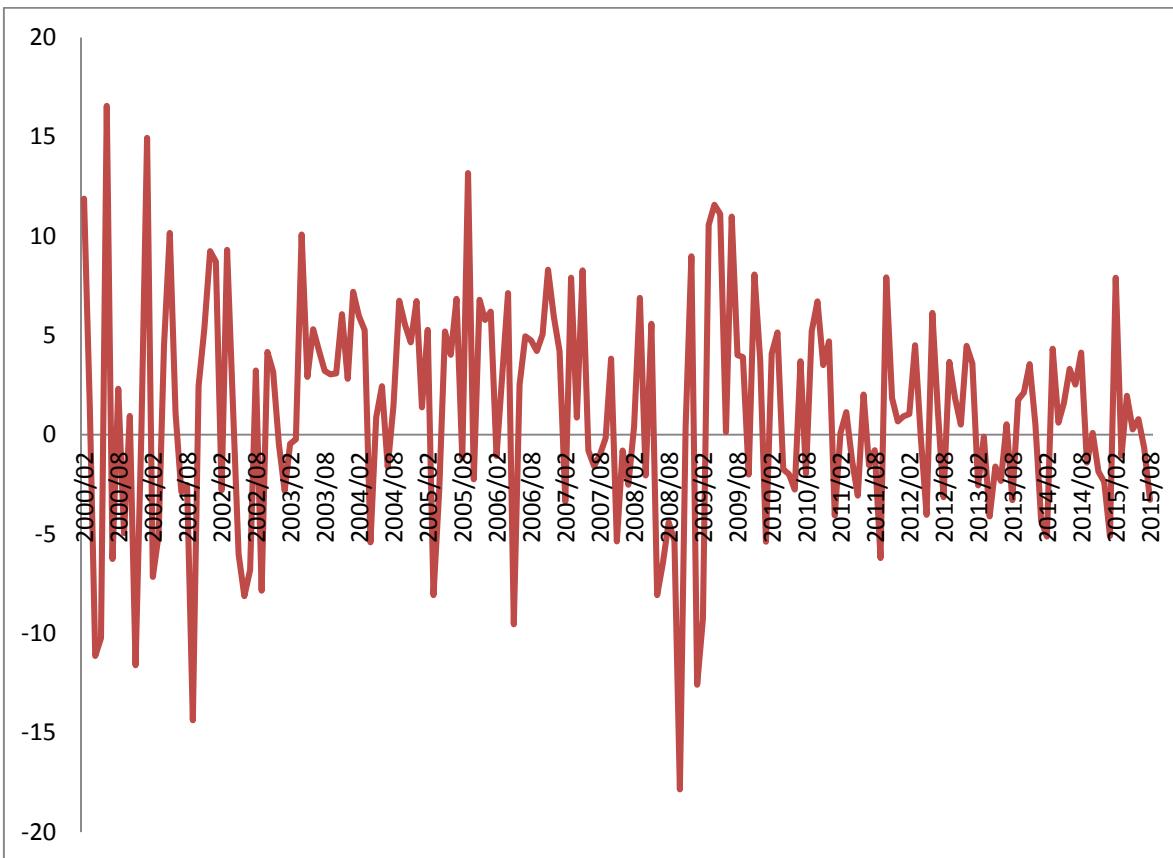
En el análisis de los mercados financieros el estudio del riesgo es altamente relevante. Harry Markowitz (1952), a través de la teoría del portafolio, realizó la fundamentación más influyente sobre la medición del riesgo en esos mercados, entendiendo como riesgo la varianza del rendimiento de un activo.

En la gráfica siguiente se muestra el comportamiento del rendimiento mensual del mercado financiero mexicano a lo largo de 2000 a 2015; el rendimiento está medido como la tasa de crecimiento porcentual del Índice de Precios y Cotizaciones (IPC) de la Bolsa de Valores. En el comportamiento de los rendimientos en el tiempo se observan períodos de volatilidad, en los cuales la intensidad y frecuencia de los cambios con respecto al comportamiento promedio son más elevadas; por ejemplo, en el segundo semestre de 2008 se observa un período de fuertes descensos en los rendimientos que se prolongan prácticamente hasta principios del 2009. De los datos podemos advertir que la volatilidad se agrupa, es decir, que existen momentos en los cuales es más elevada y después tiende a revertirse hacia su comportamiento medio.

Para cualquier inversionista o interesado en el comportamiento del mercado financiero, le resultaría muy útil poder separar esos momentos de volatilidad del comportamiento temporal del rendimiento, ello con el fin de tomar decisiones mejor informadas de inversión con el fin de reducir el riesgo de las mismas o para intentar predecir el comportamiento futuro de los procesos de volatilidad y, de esa manera, minimizar el riesgo esperado a futuro.

Gráfica 1

Rendimiento (IPC) mensual del mercado financiero mexicano; 2000-2015\*



\* Información hasta agosto de 2015

Fuente: INEGI, Banco de Información Económica

La volatilidad en sentido estricto puede definirse como la varianza de la serie de tiempo que se está considerando en el análisis, condicionada a la información pasada.

Los modelos que a continuación exploraremos, justamente tienen como objetivo central facilitar la identificación de la volatilidad en las variables económicas y, de esa manera, formular modelos para su predicción.

## 2. PROCESOS ARCH

En el capítulo 11 de este libro se estudiaron las propiedades de las series de tiempo, aquí simplemente se retomarán algunas de sus características, que son indispensables para formular los modelos de Autocorrelación Condicional Heterocedástica (ARCH).

Para ilustrar el significado de los momentos condicionales y no condicionales de un proceso estocástico consideraremos el caso más simple de un proceso AR(1) estacionario.

$$y_t = \phi_1 y_{t-1} + u_t$$

donde  $0 < \phi_1 < 1$ , siendo  $u_t \sim iid(0, \sigma^2)$

Los momentos condicionales de ese proceso dependen de los valores pasados de  $y_t$ , dado que el proceso es autorregresivo de primer orden la información del pasado se limita a la existente en el período inmediatamente anterior  $y_{t-1}$ .

Por lo tanto, al tomar la esperanza matemática condicional del proceso vamos a obtener:

$$E_{t-1}[y_t] = E[y_t | y_{t-1}, y_{t-2}, \dots] = E[y_t | y_{t-1}] = E[\phi_1 y_{t-1} + u_t] = \phi_1 y_{t-1}$$

Por su parte la varianza condicional sería la siguiente expectativa:

$$var_{t-1}[y_t] = E[(y_t - E(y_t))^2 | y_{t-1}, y_{t-2}, \dots] = E[(y_t - \phi_1 y_{t-1})^2 | y_{t-1}] = E[u_t^2] = \sigma^2$$

Los momentos no condicionales ya no dependen de las realizaciones en el tiempo de  $y_t$ , y toda la información procede del término de perturbación aleatoria.

Para la media no condicional tenemos:

$$y_t - \phi_1 y_{t-1} = u_t$$

$$y_t(1 - \phi_1 L) = u_t$$

$$y_t = (1 - \phi_1 L)^{-1} u_t = (1 + \phi_1 L + \phi_1^2 L^2 + \dots) u_t$$

$$y_t = u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \dots$$

$$E(y_t) = 0$$

Para la varianza no condicional:

$$\begin{aligned} var(y_t) &= E(u_t + \phi_1 u_{t-1} + \phi_1^2 u_{t-2} + \dots)^2 \\ &= E(u_t^2 + \phi_1^2 u_{t-1}^2 + \phi_1^4 u_{t-2}^2 + \dots + \text{productos cruzados}) \\ &= \sigma^2 + \phi_1^2 \sigma^2 + \phi_1^4 \sigma^2 + \dots = \sigma^2 (1 + \phi_1^2 + \phi_1^4 + \dots) = \frac{\sigma^2}{1 - \phi_1^2} \end{aligned}$$

Los productos cruzados en la operación previa son nulos en virtud de que los términos de perturbación aleatoria son independientes en el tiempo.

En su trabajo fundamental sobre los procesos ARCH, Engle (1982) plantea que los intervalos de predicción de un proceso estocástico podrían mejorar si pudiéramos

utilizar más información para la predicción de la varianza. El modelo propuesto por Engle es el siguiente:

$$y_t = u_t$$

$$u_t = v_t h_t^{1/2}$$

$$h_t = \alpha_0 + \alpha_1 u_{t-1}^2$$

siendo  $v_t \sim N(0,1)$  y  $y_t | \Theta_{t-1} \sim N(0, h_t)$  siendo  $\Theta_t$  el conjunto de información disponible en t.

Donde la primera ecuación es la de la media, la segunda es el término de perturbación y la tercera es  $h_t$  la varianza de  $u_t$  condicional a la información disponible en el período t.

En su trabajo original Engle supone que la ecuación de la media  $y_t$  podría tener procesos más complejos involucrando variables explicatorias de la forma  $y_t = x_t \beta$ .

De las tres ecuaciones previas podemos obtener la varianza condicional de la siguiente manera.

$$u_t = v_t \sqrt{\alpha_0 + \alpha_1 u_{t-1}^2}$$

Los momentos no condicionales de esta varianza son los siguientes.

La esperanza matemática:

$$E[u_t] = E\left[v_t \sqrt{\alpha_0 + \alpha_1 u_{t-1}^2}\right] = 0$$

La varianza:

$$\text{var}[u_t] = E[v_t^2 (\alpha_0 + \alpha_1 u_{t-1}^2)] = \alpha_0 + \alpha_1 \text{var}[u_t] = \frac{\alpha_0}{1 - \alpha_1}$$

Claramente tenemos un proceso estacionario en virtud de que  $0 < \alpha_1 < 1$ .

Si ahora obtenemos los momentos condicionales:

Media condicional:

$$E_{t-1}[u_t] = E_{t-1} \left[ v_t \sqrt{\alpha_0 + \alpha_1 u_{t-1}^2} \right] = 0$$

La varianza condicional:

$$\text{var}_{t-1}[u_t] = E_{t-1}[v_t^2 (\alpha_0 + \alpha_1 u_{t-1}^2)] = \alpha_0 + \alpha_1 u_{t-1}^2$$

Destaca en este resultado que la varianza condicional sigue un proceso en el cual la media condicional depende del valor previo del término de perturbación al cuadrado. La varianza condicional será un valor positivo y será estable si se cumple la condición de estacionariedad  $0 < \alpha_1 < 1$  y la de su derivada positiva  $\alpha_0 > 0$ .

El modelo puede generalizarse a procesos AR de mayor orden, si lo generalizamos tendremos un ARCH(p):

$$u = v_t \sqrt{\alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2}$$

La condición de estabilidad es que  $\sum_{i=1}^p \alpha_i < 1$  y las de no negatividad  $\alpha_0 > 0$ ,  $\alpha_i \geq 0$ .

Una representación equivalente para la ecuación de la varianza condicional y que es la que se emplea comúnmente en los libros de texto es la siguiente:

$$h_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2$$

En la ecuación ARCH resulta claro que la varianza es heterocedástica y depende del cuadrado de los choques aleatorios pasados y, en consecuencia, tendrán el mismo impacto en ella tanto los choques positivos como negativos.

También podemos en la ecuación de la media del proceso tener un proceso ARMA más general, por ejemplo un ARMA(p,q):

$$y_t = \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p} + u_t + \theta_1 u_{t-1} + \cdots + \theta_q u_{t-q}$$

Si la varianza condicional es heterocedástica y tiene ambos componentes AR y MA tendremos un proceso generalizado GARCH propuesto por Bollerslev (1986).

Por ejemplo, un proceso GARCH(p,q) se representaría de la siguiente manera:

$$h_t = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2 + \beta_1 h_{t-1}^2 + \cdots + \beta_q h_{t-q}^2$$

En este caso la varianza condicional depende de los cuadrados de las perturbaciones aleatorios y de los cuadrados de la varianza.

En cualquier caso todos los parámetros de la ecuación de la media y de la varianza condicional se tienen que estimar conjuntamente, para lo cual Engle (1982) propone el método de máxima verosimilitud. Esto es posible en la medida que supone que el proceso condicional sigue una distribución normal.

La estimación por MV se puede llevar a cabo dado que se ha supuesto normalidad en el proceso condicional y la función de MV es el producto de las densidades condicionales. Sin embargo, para facilitar la estimación se linealiza la función de verosimilitud con logaritmos y se maximiza entonces esa función que ahora es igual a la suma de las densidades condicionales y se expresa de la manera siguiente:

$$l = \frac{1}{T} \sum_{t=1}^T l_t$$

donde  $l_t = -\frac{1}{2} \log h_t - \frac{1}{2} u_t^2 / h_t$

La maximización de la función y sus resultados se pueden consultar en el trabajo ya referido de Engle (1982) y no es el propósito aquí desarrollarlos.

### 3. VARIANTES DE LOS MODELOS ARCH

Cuando se utiliza un modelo GARCH(1,1) para que tenga varianza finita se debe cumplir la condición  $\alpha_1 + \beta_1 < 1$ . En series financieras la volatilidad es persistente, por lo cual  $\alpha_1 + \beta_1 = 1$  y el proceso se convierte en un IGARCH o GARCH integrado que es estrictamente estacionario.

Si la variable de referencia es sensible a la volatilidad, ésta última tendrá que incorporarse como regresor en la ecuación de la media, el resultado será un modelo ARCH en media o ARCH-M como el siguiente (Wang, 2003):

$$y_t = \lambda_1 x_1 + \cdots + \lambda_m x_m + \gamma h_t + u_t$$

$$h_t = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2$$

En el cual se utilizan m variables exógenas x, las cuales podrían incluir rezagos autorregresivos de y. El mismo modelo ARCHM podría generalizarse a un GARCHM.

Antes se mencionó que los modelos ARCH suponen simetría en los choques aleatorios sean positivos o negativos. Sin embargo, en el mercado financiero y con

muchas variables económicas, las noticias negativas no tienen el mismo peso que las positivas, por lo tanto el efecto de los choques aleatorios debe de ser asimétrico.

Un modelo que captura dicha asimetría es el exponencial generalizado o EGARCH propuesto por Nelson (1991) con la siguiente especificación:

$$\log(h_t) = \alpha_0 + \sum_{j=1}^q \beta_j \log(h_{t-j}^2) + \sum_{i=1}^p \left\{ \alpha_i \left( \left| \frac{u_{t-i}}{\sqrt{h_{t-i}}} \right| - \sqrt{\frac{2}{\pi}} \right) - \zeta_i \frac{u_{t-i}}{\sqrt{h_{t-i}}} \right\}$$

Donde  $\zeta_i$  es el parámetro de respuesta asimétrica y se espera que sea positivo, de forma que choques negativos incrementarán la volatilidad y los positivos la reducirán.

Finalmente, Glosten, Jagannathan y Runkle (1993) consideraron que el EGARCH por su no linealidad era difícil de estimar, propusieron el GARCH de umbral o TGARCH.

$$h_t = \alpha_0 + \sum_{j=1}^q \beta_j (h_{t-j}^2) + \sum_{i=1}^p \{ \alpha_i u_{t-i}^2 + \delta_i u_{t-i}^2 \}$$

En este caso el parámetro  $\delta_i$  captura la respuesta asimétrica, dando lugar a que un choque negativo genera mayor o al menos igual volatilidad que la de un choque positivo.

#### 4. UNA APLICACIÓN DEL MODELO ARCH EN R

Un ejemplo de modelo ARCH en R

Instalar el paquete:

```
install.packages("rugarch")
require(rugarch)
```

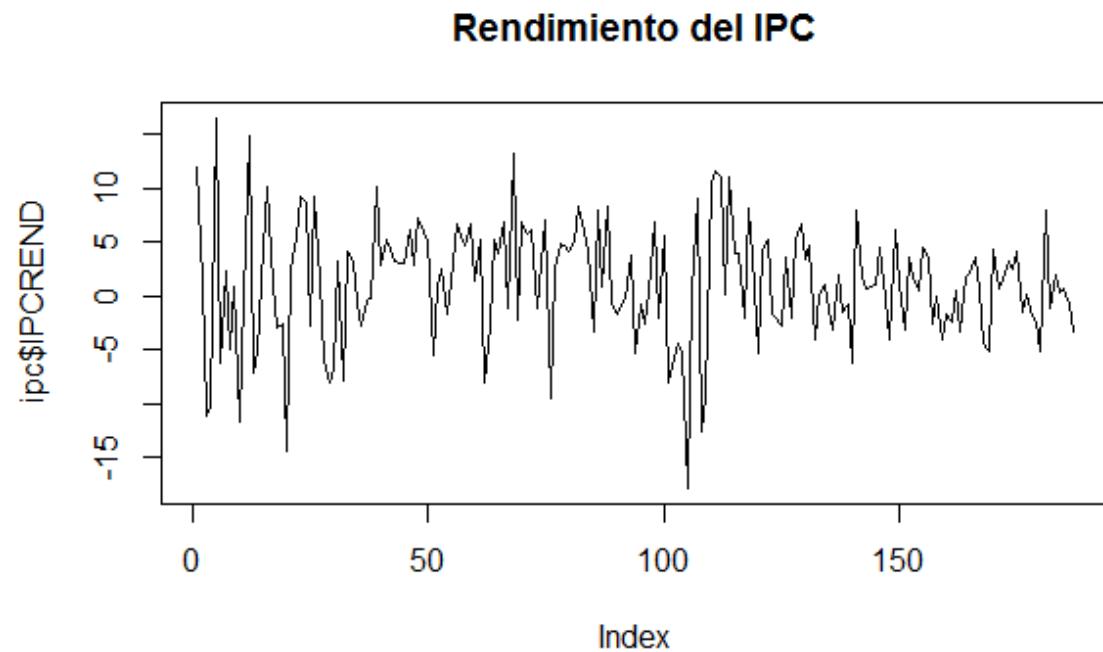
Abrimos el archivo ipc.csv que contiene las observaciones mensuales de los rendimientos del IPC de 2000.02 a 2015.08.

```
ipc <- read.table("ipc.csv", header=TRUE, quote="\")
```

En Rstudio es posible abrir el archivo seleccionando en la ventana de Environment la opción Import Dataset y ubicando la localización del archivo ipc.csv en su computadora.

Visualizamos la variable de rendimiento en un gráfico de línea con la siguiente sintaxis:

```
plot(ipc$IPCREND,type='l',main='Rendimiento del IPC')
```

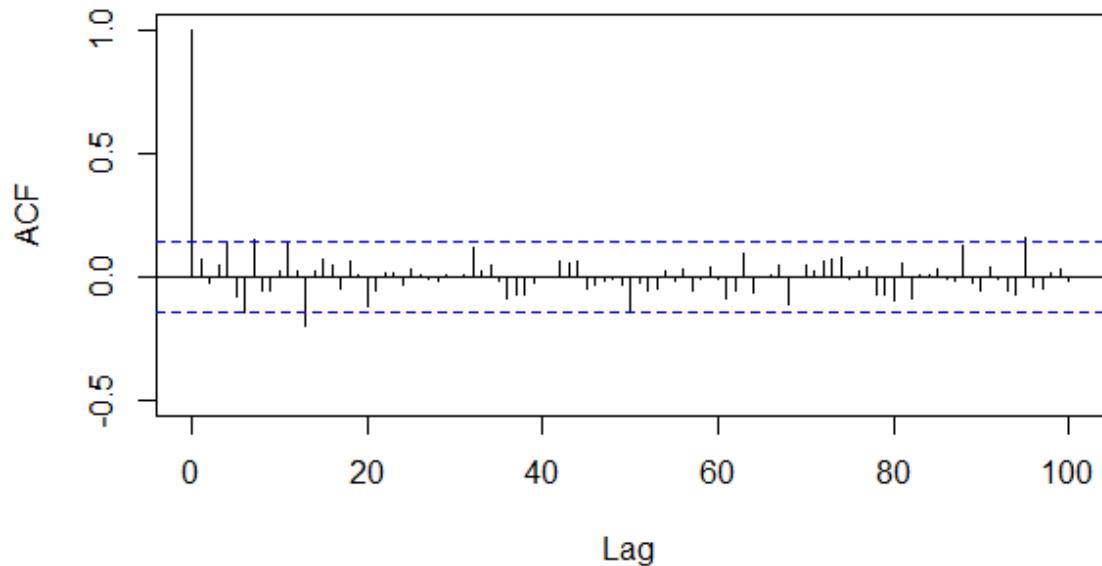


Examinamos la serie con los correlogramas para identificar el tipo de proceso ARIMA que podría representar a la serie:

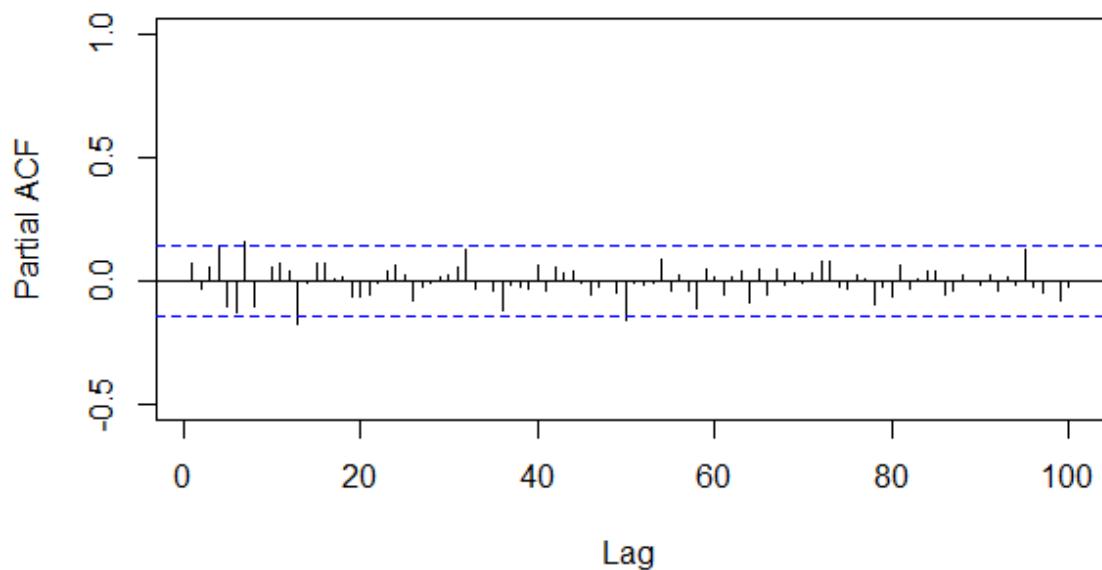
```
acf.ipc=acf(ipc$IPCREND,main='ACF IPC',lag.max=100,ylim=c(-0.5,1))
```

```
pacf.ipc=pacf(ipc$IPCREND,main='PACF IPC',lag.max=100,ylim=c(-0.5,1))
```

**ACF IPC**



**PACF IPC**



En ambos gráficos la serie luce estacionaria, aunque con pequeños brincos en los rezagos cuarto, séptimo y treceavo.

Estimamos el proceso más simple posible con los resultados obtenidos y es un ARMA(2,2)

```
arima22=arima(ipc$IPCREND,order=c(2,0,2))
```

Examinamos los resultados:

```
arima22
```

Call:

```
arima(x = ipc$IPCREND, order = c(2, 0, 2))
```

Coefficients:

	ar1	ar2	ma1	ma2	intercept
-	-0.3970	-0.9680	0.4712	0.9382	1.1560
s.e.	0.0332	0.0473	0.0382	0.0850	0.3942

```
sigma^2 estimated as 27.98: log likelihood = -577.53, aic = 1167.07
```

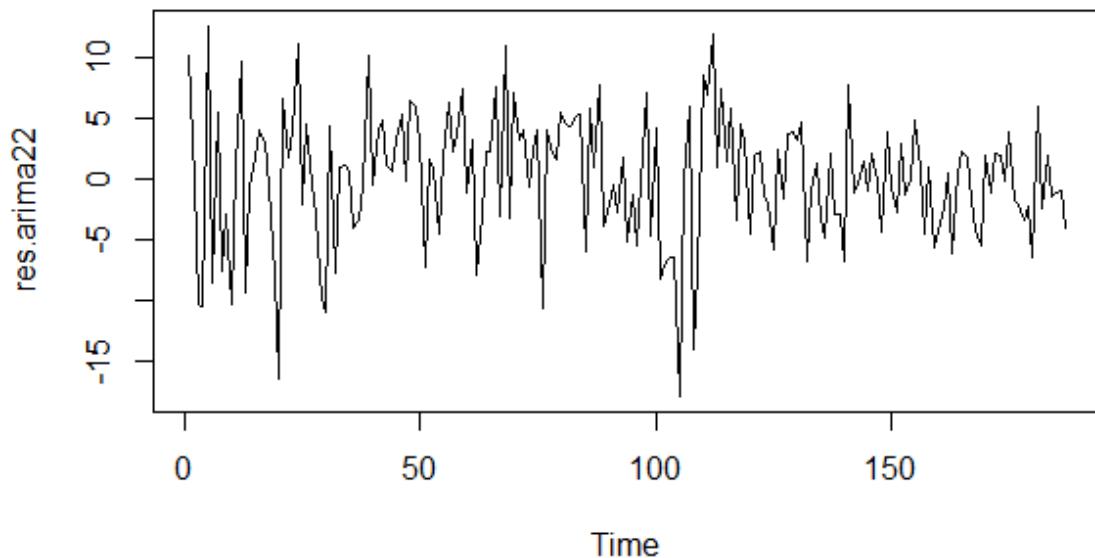
Generamos los residuales del modelo:

```
res.arima22=arima22$res
```

Los graficamos:

```
plot(res.arima22,type='l',main='Residuales')
```

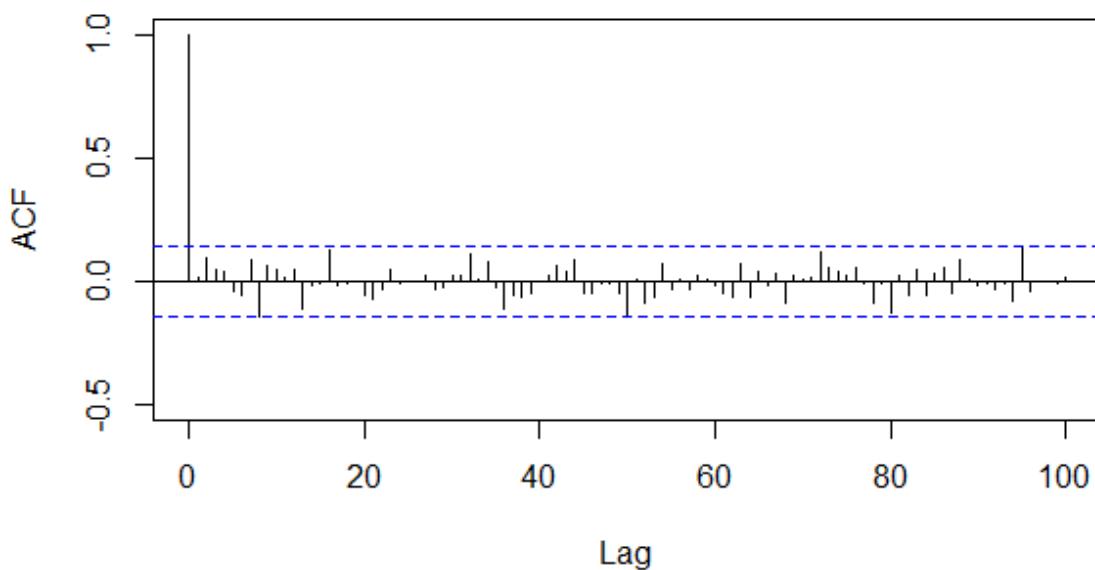
## **Residuales**



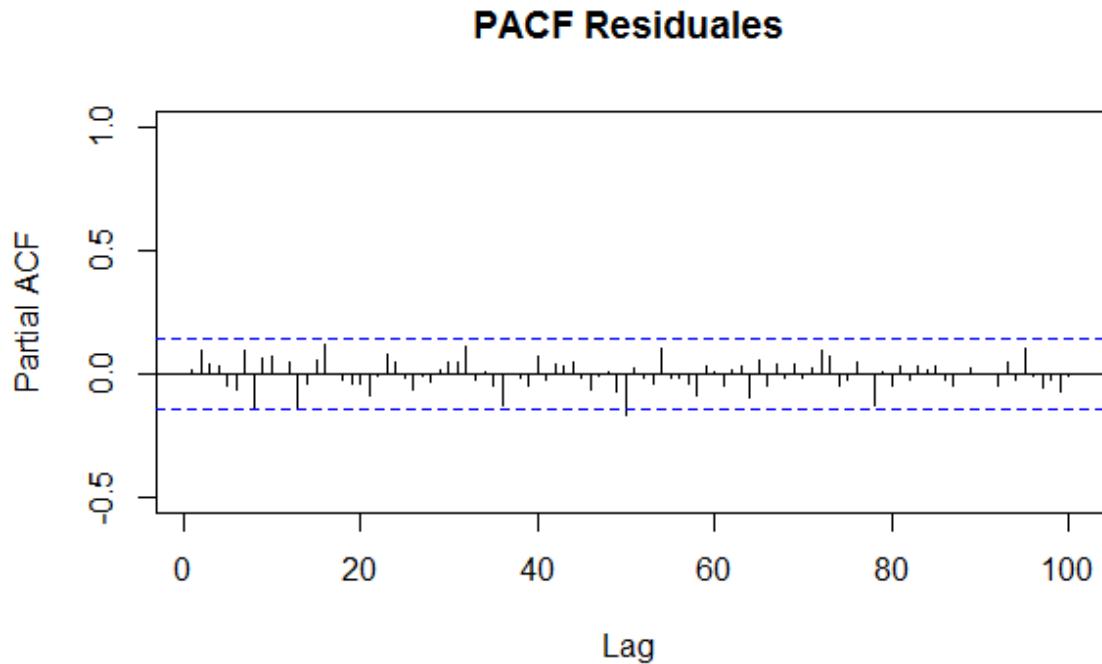
Obtenemos el correlograma de los residuales:

```
acf.res=acf(res.arima22,main='ACF Residuales',lag.max=100,ylim=c(-0.5,1))
```

## **ACF Residuales**



```
pacf.res=pacf(res.arima22,main='PACF Residuales',lag.max=100,ylim=c(-0.5,1))
```

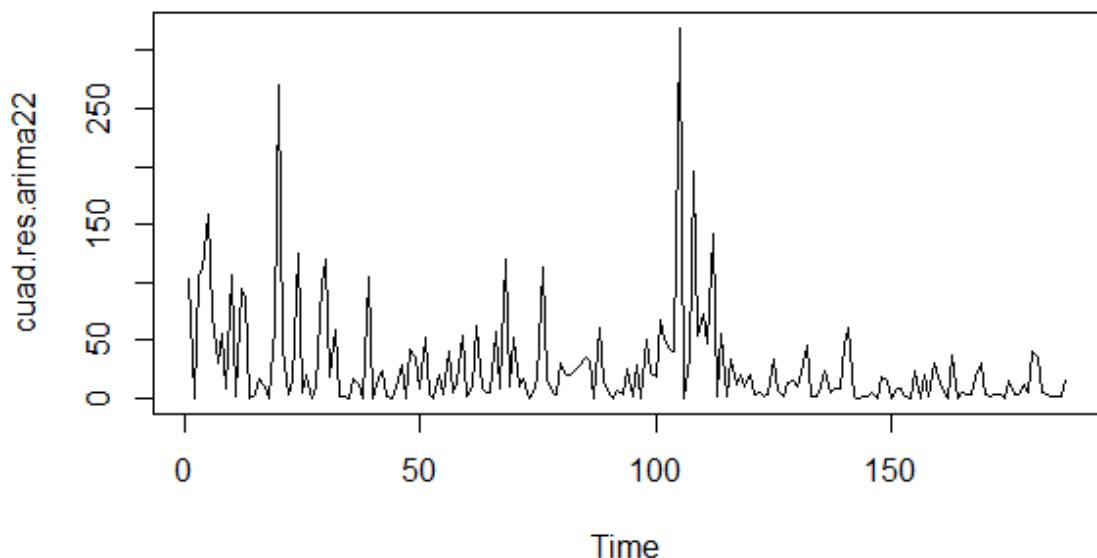


Los correlogramas de los residuales no indican ningún patrón discernible al no existir rezagos significativos, por lo que podríamos considerar que el proceso que siguen es estacionario. Sin embargo, la gráfica de residuales muestra procesos de volatilidad que deben ser examinados.

Para ello examinamos los residuales elevados al cuadrado con el fin de examinar su varianza:

```
cuad.res.arima22=res.arima22^2
par(mfcol=c(3,1))
plot(cuad.res.arima22,main='Residuales al cuadrado')
```

## Residuales al cuadrado

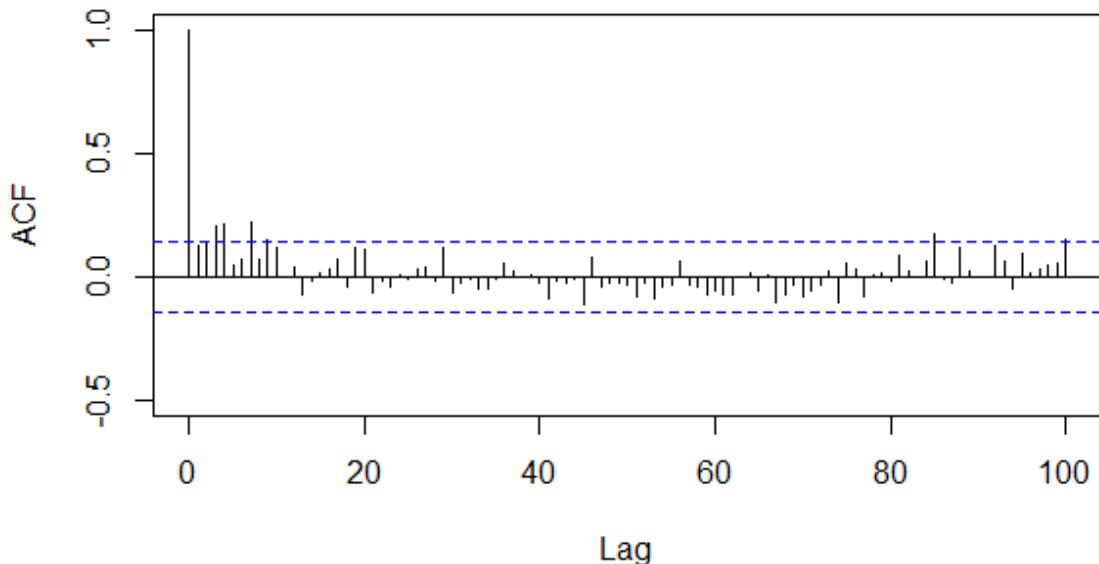


La gráfica muestra claramente procesos agrupados de volatilidad, siendo el más intenso el que aparece después de la observación 100.

Los correlogramas de los residuales al cuadrado son evidencia de la heterocedasticidad presente en la varianza, así que obtenemos los correlogramas:

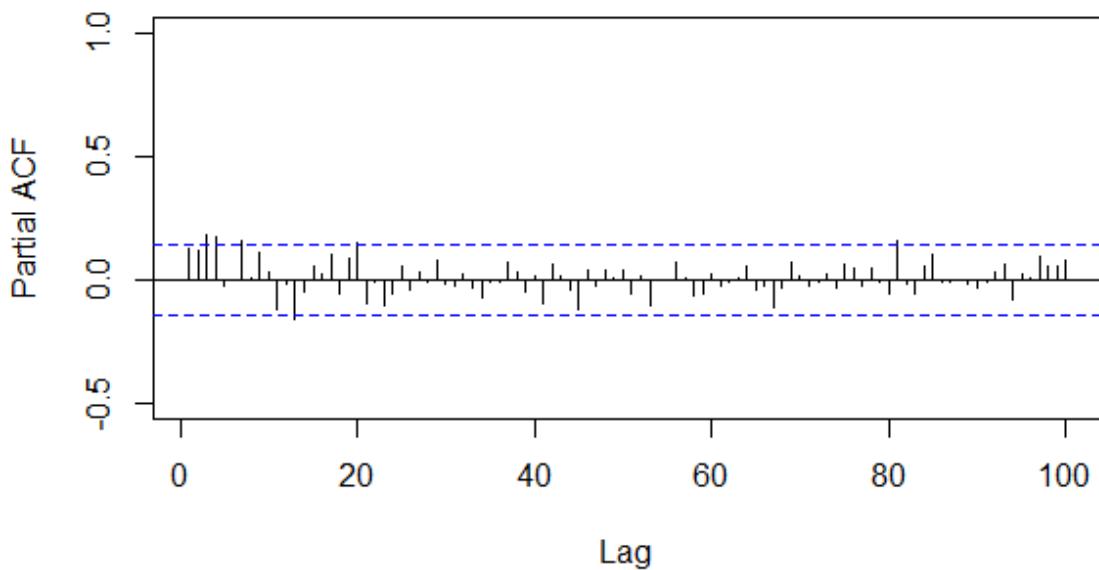
```
acf.res2=acf(cuad.res.arima22,main='ACF Residuales al cuadrado',lag.max=100,ylim=c(-0.5,1))
```

### ACF Residuales al cuadrado



```
pacf.res2=pacf(cuad.res.arima22,main='PACF Residuales al cuadrado',lag.max=100,ylim=c(-0.5,1))
```

### PACF Residuales al cuadrado



Ambas gráficas confirman que en efecto los residuales al cuadrado no son ruido blanco y tienen rezagos significativos, los cuales son indicativos de procesos de volatilidad.

Para modelar la volatilidad utilizamos modelos ARCH, utilizamos el modelo más simple un GARCH(1,1).

Para lo cual utilizamos los comandos ugarchspec y ugarchfit, que son rutinas para generar un modelo GARCH(1,1) con una especificación ARMA(1,1) en la media.

```
spec = ugarchspec()
fit = ugarchfit(data = ipc[,1], spec = spec)
fit
```

La salida de resultados del modelo se muestra a continuación:

```
*-----*
*      GARCH Model Fit      *
*-----*

Conditional Variance Dynamics
-----
GARCH Model   : sGARCH(1,1)
Mean Model    : ARFIMA(1,0,1)
Distribution   : norm

Optimal Parameters
-----
          Estimate Std. Error t value Pr(>|t|)
mu       1.11947  0.382429  2.9273 0.003420
ar1      0.63859  0.435840  1.4652 0.142871
ma1     -0.58482  0.461279 -1.2678 0.204861
omega    1.24751  0.948943  1.3146 0.188636
alpha1   0.14582  0.056392  2.5859 0.009713
beta1    0.80585  0.064351 12.5227 0.000000
```

Robust Standard Errors:

```
          Estimate Std. Error t value Pr(>|t|)
mu       1.11947  0.385841  2.9014 0.003715
ar1      0.63859  0.249164  2.5629 0.010380
ma1     -0.58482  0.259069 -2.2574 0.023984
omega    1.24751  0.696806  1.7903 0.073403
alpha1   0.14582  0.052389  2.7835 0.005378
beta1    0.80585  0.050089 16.0884 0.000000
```

LogLikelihood : -568.9842

Information Criteria

---

Akaike 6.1496  
Bayes 6.2532  
Shibata 6.1476  
Hannan-Quinn 6.1916

Weighted Ljung-Box Test on Standardized Residuals

---

statistic p-value  
Lag[1] 0.1014 0.7501  
Lag[2\*(p+q)+(p+q)-1][5] 1.0955 1.0000  
Lag[4\*(p+q)+(p+q)-1][9] 4.9090 0.4744  
d.o.f=2  
H0 : No serial correlation

Weighted Ljung-Box Test on Standardized Squared Residuals

---

statistic p-value  
Lag[1] 0.5519 0.4576  
Lag[2\*(p+q)+(p+q)-1][5] 1.2246 0.8071  
Lag[4\*(p+q)+(p+q)-1][9] 2.9572 0.7660  
d.o.f=2

Weighted ARCH LM Tests

---

Statistic Shape Scale P-Value  
ARCH Lag[3] 0.0317 0.500 2.000 0.8587  
ARCH Lag[5] 1.6030 1.440 1.667 0.5656  
ARCH Lag[7] 2.8148 2.315 1.543 0.5492

Nyblom stability test

---

Joint Statistic: 1.2262

Individual Statistics:

mu 0.38545  
ar1 0.05193  
ma1 0.05096  
omega 0.26879  
alpha1 0.18367  
beta1 0.27108

Asymptotic Critical Values (10% 5% 1%)

```
Joint Statistic: 1.49 1.68 2.12  
Individual Statistic: 0.35 0.47 0.75
```

#### Sign Bias Test

	t-value	prob	sig
Sign Bias	1.1184	0.2649	
Negative Sign Bias	0.5020	0.6162	
Positive Sign Bias	0.4211	0.6742	
Joint Effect	3.5976	0.3083	

#### Adjusted Pearson Goodness-of-Fit Test:

group	statistic	p-value(g-1)
1	20	25.73
2	30	35.51
3	40	43.37
4	50	41.61

De los resultados se obtiene que los residuales no presentan autocorrelación serial y tampoco hay proceso ARCH en los residuales elevados al cuadrado, por lo que se puede plantear que la modelación fue adecuada.

Con el modelo estimado es posible realizar una predicción de los rendimientos para los siguientes diez períodos, para lo cual utilizamos el comando fit:

```
fit = ugarchfit(data = ipc[,1], spec = spec,out.sample=10)  
> forc=ugarchforecast(fit, n.ahead=10)
```

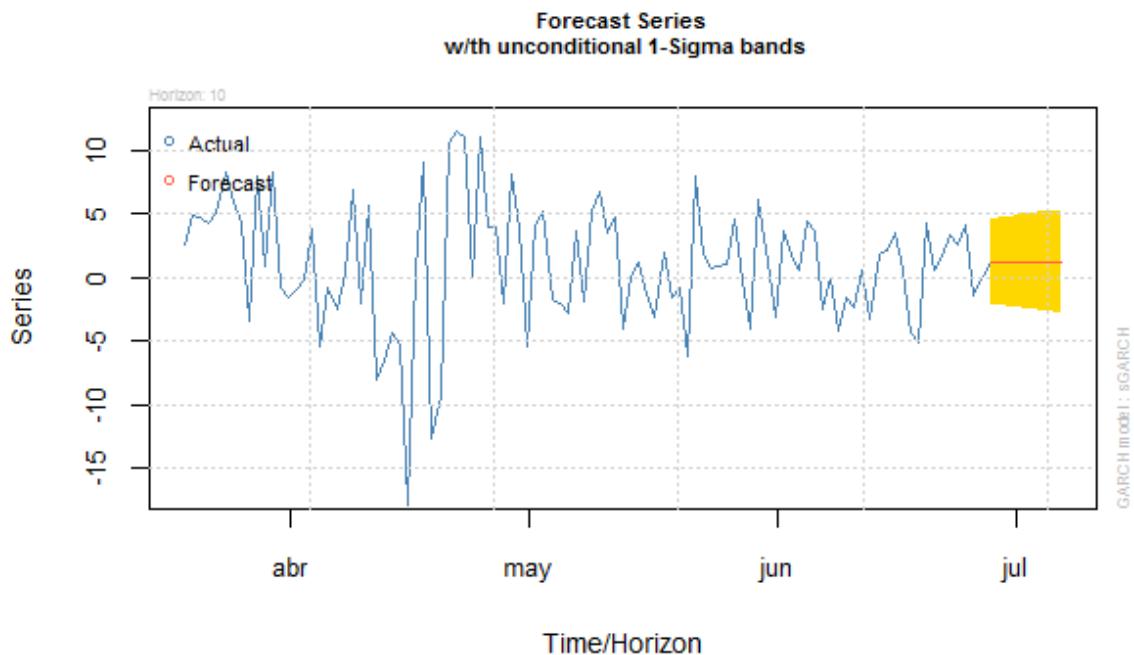
Gráficamos nuestros resultados con el comando plot y se abren las siguientes opciones de gráficas:

```
plot(forc)
```

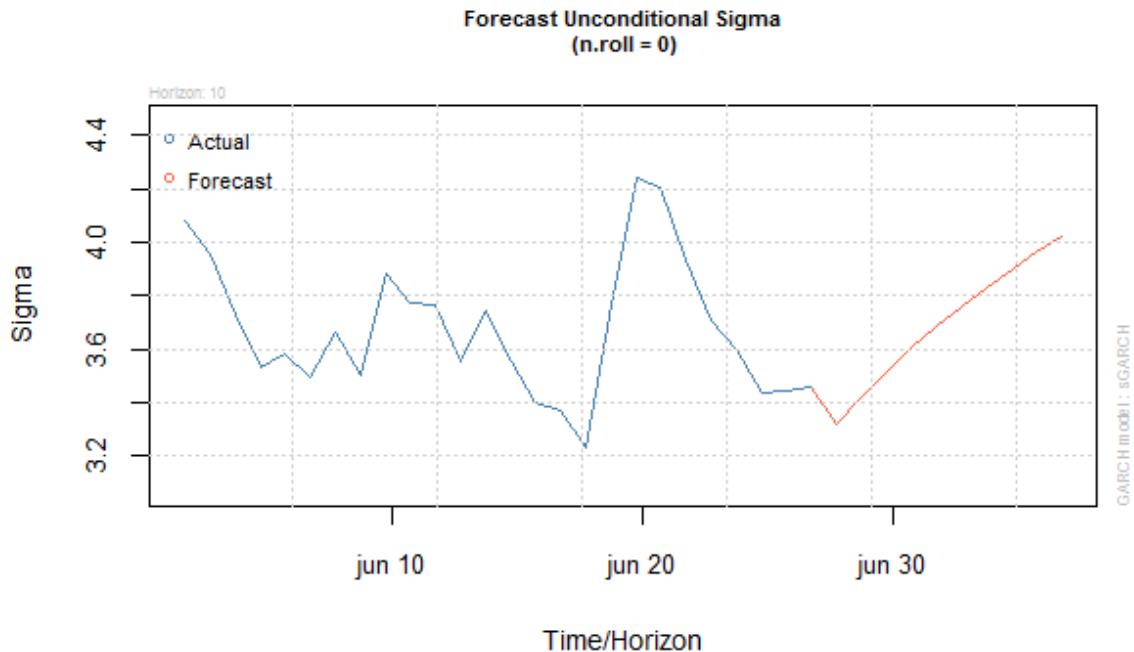
Make a plot selection (or 0 to exit):

- 1: Time Series Prediction (unconditional)
- 2: Time Series Prediction (rolling)
- 3: Sigma Prediction (unconditional)
- 4: Sigma Prediction (rolling)

Por ejemplo, la primera opción con la predicción no condicional nos permite obtener el rendimiento futuro:



O con la opción 4:



En caso de buscar especificaciones diferentes para el modelo GARCH debemos de utilizar las opciones del comando ugarchspec. Por ejemplo, probamos una especificación en la media para un modelo ARMA(2,2) que se corresponde con el que identificamos en el proceso previamente llevado a cabo con la metodología Box-Jenkins.

```
spec2=ugarchspec(variance.model = list(model = "sGARCH", garchOrder = c(1, 1),
submodel = NULL, external.regressors = NULL, variance.targeting = FALSE),
mean.model = list(armaOrder = c(2, 2), include.mean = TRUE, archm = FALSE,
archpow = 1, arfima = FALSE, external.regressors = NULL, archex = FALSE),
distribution.model = "norm", start.pars = list(), fixed.pars = list())
```

```
fit = ugarchfit(data = ipcl[,1], spec = spec2)
fit
```

La salida de resultados es la siguiente, en ella podemos ver que el ajuste del modelo es mejor y que los coeficientes son significativos:

```
*-----*
*      GARCH Model Fit      *
*-----*

Conditional Variance Dynamics
-----
GARCH Model   : sGARCH(1,1)
Mean Model    : ARFIMA(2,0,2)
Distribution   : norm

Optimal Parameters
-----
            Estimate Std. Error t value Pr(>|t|)
mu      1.12934  0.345795  3.2659 0.001091
ar1     -0.42290  0.113351 -3.7308 0.000191
ar2     -0.91474  0.088377 -10.3504 0.000000
ma1      0.47194  0.121415  3.8870 0.000101
ma2      0.88747  0.121691  7.2928 0.000000
omega   1.00344  0.924194  1.0857 0.277593
alpha1   0.13201  0.054373  2.4278 0.015190
beta1    0.82739  0.064640 12.7999 0.000000

Robust Standard Errors:
            Estimate Std. Error t value Pr(>|t|)
mu      1.12934  0.383725  2.9431 0.003250
```

ar1	-0.42290	0.123536	-3.4233	0.000619
ar2	-0.91474	0.095299	-9.5987	0.000000
ma1	0.47194	0.129168	3.6537	0.000258
ma2	0.88747	0.147628	6.0115	0.000000
omega	1.00344	0.789322	1.2713	0.203635
alpha1	0.13201	0.052062	2.5356	0.011226
beta1	0.82739	0.059165	13.9845	0.000000

LogLikelihood : -568.0831

#### Information Criteria

Akaike	6.1613
Bayes	6.2995
Shibata	6.1579
Hannan-Quinn	6.2173

#### Weighted Ljung-Box Test on Standardized Residuals

	statistic	p-value
Lag[1]	0.3833	0.5359
Lag[2*(p+q)+(p+q)-1][11]	4.9924	0.9591
Lag[4*(p+q)+(p+q)-1][19]	8.9131	0.6529

d.o.f=4

H0 : No serial correlation

#### Weighted Ljung-Box Test on Standardized Squared Residuals

	statistic	p-value
Lag[1]	0.1536	0.6951
Lag[2*(p+q)+(p+q)-1][5]	1.0222	0.8546
Lag[4*(p+q)+(p+q)-1][9]	2.6937	0.8083

d.o.f=2

#### Weighted ARCH LM Tests

	Statistic	Shape	Scale	P-Value
ARCH Lag[3]	0.04959	0.500	2.000	0.8238
ARCH Lag[5]	1.61657	1.440	1.667	0.5622
ARCH Lag[7]	2.66222	2.315	1.543	0.5795

#### Nyblom stability test

Joint Statistic:	1.7616
Individual Statistics:	
mu	0.5114

```

ar1 0.1755
ar2 0.1150
ma1 0.1178
ma2 0.2816
omega 0.2998
alpha1 0.1708
beta1 0.2702

```

Asymptotic Critical Values (10% 5% 1%)  
Joint Statistic: 1.89 2.11 2.59  
Individual Statistic: 0.35 0.47 0.75

#### Sign Bias Test

	t-value	prob	sig
Sign Bias	0.3422	0.7326	
Negative Sign Bias	0.2669	0.7898	
Positive Sign Bias	0.7515	0.4533	
Joint Effect	2.9024	0.4069	

#### Adjusted Pearson Goodness-of-Fit Test:

group	statistic	p-value(g-1)	
1	20	12.47	0.8648
2	30	21.07	0.8564
3	40	24.12	0.9704
4	50	36.80	0.9005

El modelo obtenido para la media tiene los siguientes coeficientes y son significativos:

- Constante=mu=1.129
- Términos AR: AR(1)= -0.422 , AR(2)= -0.914
- Términos MA: MA(1)= 0.471, MA(2)= 0.887

Mientras que para el modelo de la varianza tenemos:

- Constante=omega=1.003
- Término ARCH=alpha1=0.132
- Término GARCH=beta1=0.827

Si necesitamos estimar otras especificaciones de la familia de modelos GARCH simplemente modificamos la opción de la lista de modelos en la varianza: model="Tipo de modelo", los modelos que acepta son:

"sGARCH", "fGARCH", "eGARCH", "gjrGARCH", "apARCH", "iGARCH" y "csGARCH".

Cuando utilizamos el tipo de modelos fGARCH tenemos la opción de estimar los siguientes submodelos:

"GARCH", "TGARCH", "AVGARCH", "NGARCH", "NAGARCH", "APARCH", "GJRGARCH" y "ALLGARCH".

## REFERENCIAS

Andersen, T. G. y Bollerslev, T. (1998). " Deutsche Mark Dollar Volatility: Intraday Activity Patterns, Macroeconomic Announcements, and Longer Run Dependencies ". Journal of Finance, 53(1): 219-265.

Baillie R. T., Bollerslev T. y Mikkelsen H. (1996). " Fractionally Integrated Generalized Autoregressive Conditional Heteroskedasticity ". Journal of Econometrics, 74: 3-30.

Bollerslev, T. (1986). " Generalized autoregressive Conditional Heterocedasticity ". Journal of Econometrics, 31: 307-327.

Engle, F. R. (1982). " Autoregressive Conditional Heterocedasticity whit Estimates of the Variance of United Kingdom Inflation ". Econometrica, 50(4), 987-1008.

Engle, F. R. y Patton, A.J. (2001). " What a Good is a Volatility Model? ". Quantitative Finance, 1(2): 237-245.

Engle, R. F. y Ng V. K. (1993). " Measuring and Testing the Impact of News on Volatility ". Journal of Finance, 48(5): 1749-1778.

Engle, R.; Ito T. y Lin W-L. (1990). " Meteor Showers or Heat Waves? Heteroskedasticity Intra-Daily Volatility in the Foreing Exchange Market ". Econometrica, 58(3): 525-542.

Engle, R. y Mezrich J. (1996). " Garch for Groups " , Risk, 9: 36-40

Glosten, L. R.; Jagannathan, R. y Runkle D.E. (1993). " On the Relation between the Expected Value and the Volatility of the Nominal Excess Returns of Stocks ". Journal of Finance, 48: 1779-1801.

Markowitz, Harry (1952). "Portfolio Selection", The Journal of Finance, Vol. 7, No. 1. (Mar., 1952), pp. 77-91.

Nelson, B. D. (1991). " Conditional Heterocedasticity in Asset Returns: A New Approach ". Econometrica, 59(2): 347-370.

## **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

ipc.csv

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap12

Práctica\_Cap12

VideoPráctica\_Cap12

VideoTeoría\_Cap12

# CAPITULO 13: MODELOS LOGIT Y PROBIT

Luis Quintana Romero y Miguel Ángel Mendoza

## 1. LA IMPORTANCIA DE LAS VARIABLES CATEGÓRICAS

En el análisis económico se utilizan variables categóricas, las cuales son indicadoras de la presencia o ausencia de algún atributo. Sobre todo, en la información proveniente de micro datos de individuos, de empresas o de familias es común encontrar este tipo de variables.

En encuestas demográficas es común encontrar variables como el género de las personas que habitan una vivienda, en las encuestas industriales se reporta si la empresa tuvo acceso o no al crédito del sistema financiero y la variable respectiva es simplemente si o no; en todos estos casos las variables se registran en forma binaria, utilizando el número 1 para indicar la presencia del atributo respectivo y el 0 para su ausencia. En inglés es usual denominar a esas variables binarias como *dummies*, término que en castellano se ha naturalizado y muchos hacen referencia a esas variables, de forma indistinta, como binarias o dummies.

Para ilustrar este tipo de variables, en el cuadro siguiente se presentan los datos sobre el género de los trabajadores mexicanos obtenidos de la Encuesta Nacional de Ocupación y Empleo (ENO) para el segundo trimestre de 2015. De esa información se deriva que de un total de 49.6 millones de trabajadores, el 62% son hombres y el 38% son mujeres.

Cuadro 1

Género de la fuerza de trabajo en México, 2015

Género	Frecuencia	Porcentaje
0	18,738,988	37.8
1	30,832,408	62.2
Total	49,571,396	100

Hombre=1

Mujer=0

Fuente: Con base en datos del INEGI, ENOE, segundo trimestre de 2015

Un rápido vistazo a los datos de la ENOE permite observar que la variable género simplemente se va reportando con ceros y unos, tal y como se observa en el cuadro siguiente; por ejemplo, en el hogar 1 se entrevistó a una mujer (género=0), con estado civil soltera (casado=0), mientras que en el hogar 2 es un hombre y está casado.

Hogar	Casado	Género
1	0	0
2	1	1
3	1	0
4	0	1
5	0	0
6	0	0
7	1	1
8	1	0
9	0	1
10	0	0

Fuente: Con base en datos del INEGI, ENOE, segundo trimestre de 2015

En los modelos económicos este tipo de variables binarias suelen incorporarse, sobre todo cuando la información proviene de micro datos. Por ejemplo, en los estudios sobre las remuneraciones de los trabajadores suelen estimarse ecuaciones conocidas como *mincerianas* y que fueron propuestas por Jacob Mincer en su conocido libro publicado en 1974. En ese texto Mincer establece la existencia de una relación positiva entre el salario y la escolaridad de los individuos. Con el fin de analizar algún tipo de discriminación salarial para las mujeres, en las ecuaciones mincerianas se incorpora una variable binaria de género como la que ya se ha descrito antes. El modelo se especifica del siguiente modo:

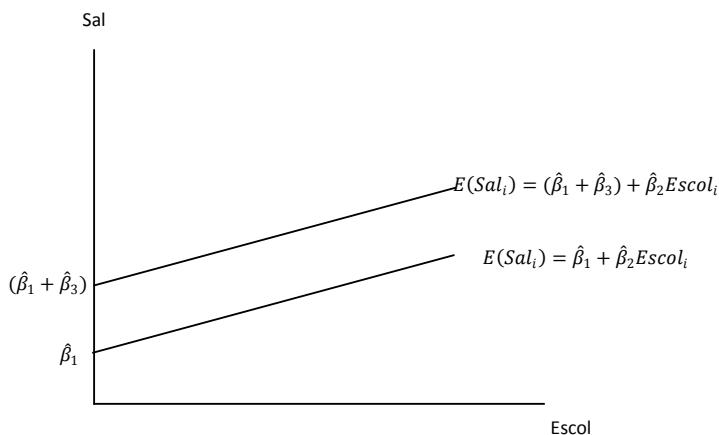
$$sal_i = \beta_1 + \beta_2 Escol_i + \beta_3 Gen_i + u_i \quad (1)$$

donde:

Sal es el salario en pesos, Escol son los años de educación, Gen es una variable binaria con 1 cuando es hombre y 0 cuando es mujer y u es un término de perturbación aleatoria.

Al estimar un modelo como el de la ecuación (1) la recta de regresión tendría un intercepto diferente para hombres y para mujeres (si los valores estimados para los coeficientes  $\beta_1$  y  $\beta_2$  del modelo son positivos y significativos), pero se mantendría la misma pendiente ya que la variable escolaridad no está diferenciada por género.

Gráfica 1 Regresión con variable binaria explicativa



En estos modelos las variables binarias operan como uno más de los regresores de la ecuación. Sin embargo, si estas variables las utilizamos como variables dependientes es necesario considerar otro tipo de modelos, los cuales se revisaran en este capítulo en las secciones siguientes.

## 2. MODELOS LOGIT Y PROBIT

Cuando la variable binaria es la variable dependiente a explicar, el modelo de regresión se interpreta como probabilidades. Retomando el ejemplo de la ecuación *minceriana*, el modelo se podría reformular considerando una variable salarial binaria; igual a la unidad para salarios por encima de la media y cero para salarios por debajo de la media. La especificación del modelo se muestra a continuación.

$$sal_i = \beta_1 + \beta_2 Escol_i + \beta_3 Gen_i + u_i \quad (2)$$

En la ecuación previa la variable  $sal_i$  es igual a 1 cuando el salario del individuo  $i$  está por encima de la media y 0 en otro caso.

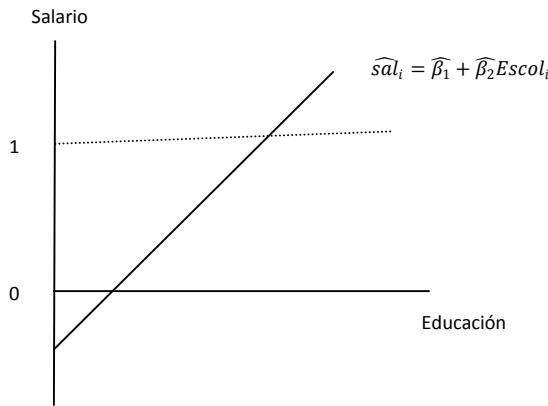
Las preguntas que podríamos responder con esta ecuación son, por ejemplo, ¿cuál es la probabilidad de que el salario se encuentre por encima de la media cuando el individuo tiene cierto número de años de educación? o bien ¿cuál es la probabilidad de que el salario este por encima de la media cuando el individuo es mujer?

Para simplificar la explicación suponga que se estima una versión más restringida de la ecuación previa:

$$sal_i = \beta_1 + \beta_2 Escol_i + u_i \quad (3)$$

El resultado de la estimación por mínimos cuadrados ordinarios podría graficarse de la siguiente manera:

Gráfica 2: Regresión con variable dependiente binaria



Si tomamos las probabilidades tendríamos:

$$P(sal=1|Educ) = F(\widehat{\beta}_1 + \widehat{\beta}_2 Escol_i)$$

Tal como se observa en la gráfica 2, la probabilidad de que el salario sea mayor a la media, dados los años de educación, es una función lineal de la educación. En teoría la función de probabilidad debe tener estrictamente valores entre cero y uno, sin embargo, en la gráfica nada garantiza eso y las probabilidades no tendrían sentido cuando obtenemos valores negativos o mayores a la unidad. Este tipo de modelos se conocen como Modelo de Probabilidad Lineal (MPL), pero su utilidad es limitada dado el resultado mencionado antes.

Para lograr asegurar que las probabilidades estén restringidas a valores entre cero y uno se han sugerido dos modelos fundamentales; el logístico o logit y el probabilístico o probit.

El logístico se especifica a través de una función logística de la siguiente manera:

$$\text{Función logística: } F(z) = \frac{\exp(z)}{[1+\exp(z)]} = P_i \quad (4)$$

donde  $Z_i = \beta_1 + \beta_2 X_{1,i} + \cdots + \beta_k X_{k,i}$

La expresión previa es simplemente una función de distribución acumulada para una variable aleatoria logística Z.

Por lo cual el modelo de regresión Logit quedaría especificado de la siguiente manera:

$$Y_i = \frac{\exp(z)}{[1 + \exp(z)]} + \varepsilon_i$$

Con base en la probabilidad es posible construir la razón de probabilidades (Gujarati, 2014):

$$\frac{P_i}{1-P_i} = \frac{1+\exp(z)}{[1+\exp(-z)]} = \exp(z) \quad (5)$$

Por consiguiente al tomar logaritmos en (5) se obtiene el logit:

$$L_i = \ln \left[ \frac{P_i}{1 - P_i} \right] = Z_i$$

En el ejemplo de los salarios de la ecuación (4) la razón de probabilidades indicaría la razón de la probabilidad de tener un salario por arriba de la media en relación a tenerlo por debajo de la media, dado el nivel educativo.

La regresión logística no supone linealidad como en los modelos de regresión clásica, tampoco requiere del supuesto de normalidad ni del de homocedasticidad (Garson, 2014). Sin embargo, si requiere que las observaciones sean independientes y que las variables explicatorias estén relacionadas linealmente al logito de la variable dependiente, tal y como se expresa esa relación en la ecuación (4).

El probit, por otro lado, se especifica a través de la siguiente función de distribución acumulada normal:

$$\text{Modelo Probit: } F(z) = \Psi(z) = \int_{-\infty}^z \psi(v)dv \quad (5)$$

En donde  $\psi(z)$  es la distribución normal estándar:  $\phi(v) = (2\pi)^{-1/2}\exp(-\frac{v^2}{2})$ .

Por lo cual el modelo de regresión Probit quedaría especificado de la siguiente manera:

$$Y_i = \Psi(z) + \varepsilon_i = \int_{-\infty}^z \psi(v)dv + \varepsilon_i = \int_{-\infty}^z (2\pi)^{-1/2}\exp(-\frac{v^2}{2})dv + \varepsilon_i$$

En general, los resultados de los modelos logit y probit permiten llegar a las mismas conclusiones ya que sus coeficientes sólo difieren en escala; los coeficientes logit son aproximadamente 1.8 veces los que se obtienen en el probit. Tal vez la desventaja más visible de los probit es que sus coeficientes son más difíciles de interpretar y además, debido al supuesto de normalidad, no se recomienda su uso cuando las observaciones se concentran mucho en alguna de las colas de la distribución (Garson, 2012).

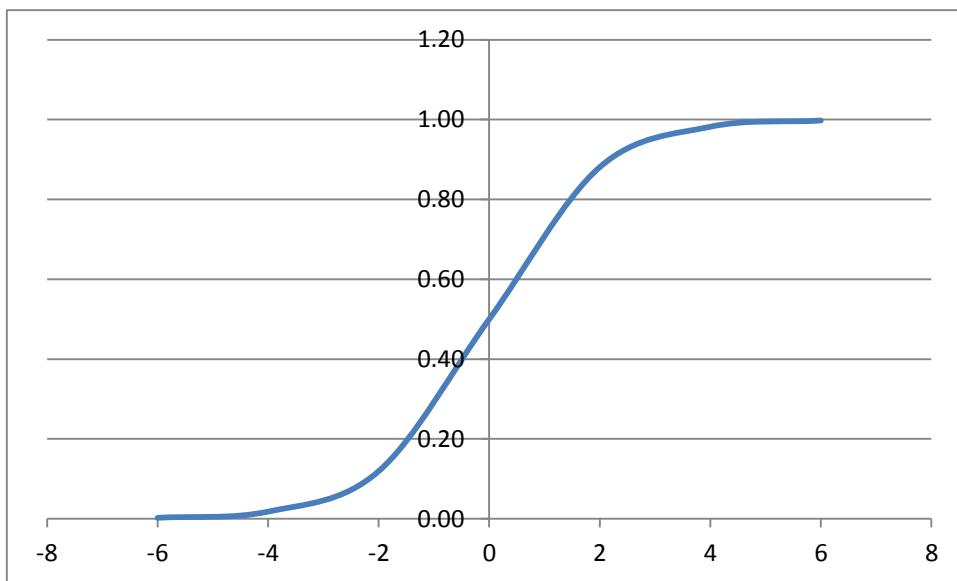
Utilizando los valores para Z que vienen en la siguiente tabla y auxiliándose con una simple calculadora se podría aplicar la fórmula (4) del modelo logístico y construir una función logística.

Cuadro 1: Función logística

Z	Logit
6	1.00
4	0.98
2	0.88
0	0.50
-2	0.12
-4	0.02
-6	0.00

Cómo se puede observar en la gráfica de la función, sus valores están acotados al intervalo de cero y uno.

Gráfica 4: Función logística



En el caso del modelo probit se pueden sustituir las medias de las variables explicatorias en la ecuación (5) para obtener las estimaciones de los valores Z y luego simplemente buscar en la tabla de la normal los niveles de probabilidad que les corresponden.

## **Estimación por MV**

Son modelos estimados por MV debido a su no linealidad. Este método tiene ventajas estadísticas en virtud de que sus estimaciones son consistentes, eficientes y para muestras grandes son insesgadas y su distribución se aproxima a una normal (Garson, 2014).

Para estimarlos es necesario tener la densidad de  $y$  dada  $x$ , la cual es una función binaria de éxito y fracaso:

$$f(y|x_i, \beta) = [F(x_i\beta)]^y [1 - F(x_i\beta)]^{1-y}$$

Al tomar logaritmos tenemos la logMV:

$$l(\beta) = y[F(x_i\beta)] + (1 - y)[1 - F(x_i\beta)]$$

La ecuación se maximiza de manera usual tomando las condiciones de primero y segundo orden, se igualan a cero y se resuelve el sistema de ecuaciones resultante. Sin embargo, es un sistema de ecuaciones no lineales, por lo cual se debe utilizar algún algoritmo de optimización que permita a los estimadores la convergencia.

## **Pruebas de hipótesis**

Se pueden aplicar pruebas de restricciones tipo Wald. Una prueba usual en este sentido consiste en comparar la razón de verosimilitud (LR) del modelo que se está estimando en relación al modelo nulo, en el cual los coeficientes de las variables explicativas están restringidos a ser nulos. Si el LR es significativamente diferente de cero tendremos evidencia de que el modelo que se está estimando es diferente al nulo.

La bondad de ajuste se obtiene con base en el porcentaje correctamente predicho por el modelo: se define un valor predicho de uno si la probabilidad predicha es de

menos 0.5 y de cero en caso contrario. El porcentaje predicho correctamente es el número de veces en que el valor estimado es igual al real.

En ese sentido las R cuadradas son en realidad seudo R cuadradas. Las más usuales son las siguientes.

$$\text{McFadden (1974)} = 1 - \frac{\log MV}{\log MV(0)}$$

Es decir, toma las funciones log verosimilitud no restringida ( $\log MV$ ) y la restringida  $\log MV(0)$  (con sólo la pendiente). Si las variables no explican nada  $\log MV = \log MV(0)$  y por ende la seudo Rcuadrada será cero.

Otras alternativas toman correlaciones entre las variables estimadas y las reales, lo cual es más cercano al espíritu de la R cuadrada en modelos de MCO.

### 3. APPLICACIONES EN R

#### Ejemplo. Modelos probabilísticos logit y probit

Los modelos probabilísticos que se presentan se elaboraron para predecir la probabilidad de obtener ingresos por hora por arriba de la mediana ( $p$ ), de acuerdo a los años de escolaridad, la experiencia y el sexo.

$$\ln(p_i) = \alpha + \beta_1 \text{escolaridad}_i + \beta_2 \text{experi}_i + \beta_3 \text{sexo}_i + u_i$$

Datos

Los indicadores se construyeron con la Encuesta Nacional de Ocupación y Empleo (ENOE) 2015 del INEGI.

# Para llevar estimar los modelos probabilísticos se utilizan las herramientas de la librería de stats.

```
library(stats)
```

```
# Se elige el directorio donde se encuentra la base de datos y el script
```

```
setwd("../LibroEconometria_R/Capitulo_LogitProbit")
```

```
# Lectura de Base de Datos previamente salvada en formato de RData
```

```
load("Capitulo_LogitProbit.RData")
```

```
# Se adjunta la base se de datos para hacerla accesible
```

```
attach(BDatos_1)
```

La base de datos contiene las siguientes variables que se utilizaran para la estimación de los modelos probabilísticos: el ingreso por hora (ing\_x\_hrs) con la cual se construye la variable dummy donde toma el valor de 1 si esta arriba de la media y cero en otro caso; los años de escolaridad con seis años de primaria, tres de secundaria, tres de bachiller, cinco de licenciatura, dos de maestría y cinco de doctorado; la experiencia igual a la edad menos escolaridad y seis años; y, finalmente la variable sexo con uno para hombres y cero mujeres.

```
# Estadísticos básicos de variables
```

```
summary(BDatos_1)
```

ing_x_hrs	ingocup	escolaridad	sexo
Min. : 0.0886	Min. : 16	Min. : 0.000	Min. : 0.0000
1st Qu.: 16.6667	1st Qu.: 3000	1st Qu.: 6.000	1st Qu.: 0.0000
Median : 24.2248	Median : 4300	Median : 9.000	Median : 1.0000
Mean : 34.2161	Mean : 5568	Mean : 9.725	Mean : 0.6085
3rd Qu.: 37.7778	3rd Qu.: 6450	3rd Qu.: 12.000	3rd Qu.: 1.0000
Max. : 3000.0000	Max. : 180000	Max. : 24.000	Max. : 1.0000

exper	capacita
Min. : 0.0	Min. : 0.000
1st Qu.: 14.0	1st Qu.: 0.000
Median : 25.0	Median : 0.000
Mean : 26.4	Mean : 1.278
3rd Qu.: 36.0	3rd Qu.: 0.000
Max. : 92.0	Max. : 98.000

```
# Para generar la variable cualitativa con valor de uno si esta por arriba de la mediana y cero en otro caso
```

```
y <- ifelse(ing_x_hrs>24.22, 1, 0)
```

```
# Para comprobar que se genero una variable dummy
```

```
summary(y)
```

Min.	1st Qu.	Median	Mean	3rd Qu.	Max.
0.0000	0.0000	1.0000	0.5036	1.0000	1.0000

```

# Se estiman los modelos logit y probit

# Modelo logit

mod_logit <- glm(y~escolaridad+exper+sexo, family = "binomial")

summary(mod_logit)

Call:

glm(formula = y ~ escolaridad + exper + sexo, family = "binomial")

Deviance Residuals:

    Min      1Q  Median      3Q     Max 
-2.2072 -1.0861  0.5374  1.0795  2.2437 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -2.4328725  0.0263347 -92.38   <2e-16 ***
escolaridad  0.1859686  0.0017082 108.87   <2e-16 ***
exper        0.0174362  0.0004403  39.60   <2e-16 ***
sexo         0.2789270  0.0126256  22.09   <2e-16 ***

---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 167665  on 120948  degrees of freedom
Residual deviance: 152567  on 120945  degrees of freedom
AIC: 152575

Number of Fisher Scoring iterations: 4

# Modelo probit

mod_probit <- glm(y~escolaridad+exper+sexo, family=binomial(link="probit"))

```

```

summary(mod_probit)

Call:
glm(formula = y ~ escolaridad + exper + sexo, family = binomial(link = "probit"))

Deviance Residuals:
Min      1Q  Median      3Q     Max 
-2.2367 -1.0968  0.5401  1.0796  2.2530 

Coefficients:
            Estimate Std. Error z value Pr(>|z|)    
(Intercept) -1.4116513  0.0150724 -93.66 <2e-16 ***
escolaridad  0.1101217  0.0009769 112.72 <2e-16 ***
exper        0.0095507  0.0002621  36.44 <2e-16 ***
sexo         0.1662554  0.0077044  21.58 <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 167665  on 120948  degrees of freedom
Residual deviance: 152796  on 120945  degrees of freedom
AIC: 152804

Number of Fisher Scoring iterations: 4

Los resultados estadísticos del modelo logit y probit muestran que la escolaridad, la experiencia y el sexo son estadísticamente diferentes de cero y tienen una relación positiva con la probabilidad de obtener ingreso por hora por arriba de la mediana. También se observa que el modelo logit presenta coeficientes mayores a los de modelo probit, aunque no son comparables.

La lectura de los resultados son los siguientes, cuando se utiliza el logaritmo odds:

1. Por una unidad de cambio en años de escolaridad, se incrementa el logaritmo de odds de tener ingreso por arriba de la media por 0.18 en el modelo logit y 0.11 en el modelo probit.

```

2. Por una unidad de cambio en años de experiencia, se incrementa el logaritmo de odds de tener ingreso por arriba de la media por 0.017 en el modelo logit y 0.009 en el modelo probit.

3. El ser hombre (sexo =1), incrementa el logaritmo de odds de tener ingreso por arriba de la media por 0.279 en el modelo logit y 0.166 en el modelo probit.

Si para el análisis se utiliza en lugar del log odds la razón de odds, entonces los resultados son los siguientes:

```
## odds ratios and 95% CI
```

```
exp(cbind(OR = coef(mod_logit), confint(mod_logit)))
```

Waiting for profiling to be done...

	OR	2.5 %	97.5 %
(Intercept)	0.08778431	0.08335809	0.09242306
escolaridad	1.20438447	1.20036800	1.20843265
exper	1.01758906	1.01671217	1.01846834
sexo	1.32171085	1.28941374	1.35483454

```
exp(cbind(OR = coef(mod_probit), confint(mod_probit)))
```

Waiting for profiling to be done...

	OR	2.5 %	97.5 %
(Intercept)	0.2437405	0.2368851	0.2507615
escolaridad	1.1164140	1.1143268	1.1185135
exper	1.0095964	1.0090896	1.0101048
sexo	1.1808747	1.1631579	1.1988686

Ahora, con la razón odds la interpretación es la siguiente:

1. Por una unidad de incremento en años de escolaridad, se incrementa la razón de odds de tener ingreso por arriba de la media por un factor 1.20 en el modelo logit y 1.11 en el modelo probit.
2. Por una unidad de incremento en años de experiencia, se incrementa la razón de odds de tener ingreso por arriba de la media por un factor 1.01 en el modelo logit y 1.009 en el modelo probit.
3. El ser hombre (sexo =1), incrementa la razón de odds de tener ingreso por arriba de la media por un factor de 1.321 en el modelo logit y 1.181 en el modelo probit.

## **REFERENCIAS**

Hosmer, D. & Lemeshow, S. (2000). Applied Logistic Regression (Second Edition). New York: John Wiley & Sons, Inc.

Long, J. Scott (1997). Regression Models for Categorical and Limited Dependent Variables. Thousand Oaks, CA: Sage Publications.

## **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

**Capitulo\_LogitProbit.R**

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap13

Práctica\_Cap13

VideoPráctica\_Cap13

VideoTeoría\_Cap13

# **CAPITULO 14: MODELOS PANEL Y SUS APLICACIONES EN R**

**Miguel Ángel Mendoza González y Luis Quintana Romero**

## **1. INTRODUCCION**

Los modelos panel normalmente se utilizan cuando el fenómeno económico, financiero, social, etc. que se está analizando tiene un componente de desagregación, corte trasversal o sección cruzada y otro de series de tiempo (Cameron, C. y Trivedi, P., 2005; Frees, E., 2004; Greene, W., 1998; Hsiao, C., 2003; Wooldridge, J., 2002). Los modelos panel clásicos más usados son los de efecto común, efecto individual fijo y el de efecto individual aleatorio. Para sus aplicaciones es importante elaborar pruebas de hipótesis entre los supuestos de efectos comunes versus individuales, y entre los efectos individuales fijos o aleatorios. Así los supuestos econométricos básicos de varianza homoscedástica, normalidad, no-autocorrelación serial y/o contemporánea.

Para entender los modelos panel clásicos, este capítulo se estructura de la siguiente manera: 1) La especificación general de un modelo panel y supuestos analíticos; 2) La evaluación del supuesto de consistencia de los estimadores de un modelo pool en comparación con el modelo de efectos individuales fijos; 3) Si el modelo de efectos fijos es mejor que el de pool, entonces evaluar si el modelo de efectos aleatorios es consistente; 4) Con el modelo panel elegido se analizan las implicaciones analíticas.

## 2. MODELO PANEL ESTÁTICO GENERAL

El modelo panel estático general tiene dos fuentes de heterogeneidad entre los elementos  $i$  de sección cruzada, por las constantes individuales  $\mu_i$  y los parámetros de relación individual  $\beta_i$  entre la variable endógena  $y_{i,t}$  y las exógenas  $X_{i,t}$ .

$$y_{i,t} = \mu_i + \beta_i X_{i,t} + \varepsilon_{i,t} \quad (1)$$

donde  $i = 1, 2, 3, \dots, n$  es el identificador de los elementos de la sección cruzada y  $t=12,3,\dots, T$ , el de tiempo.

Con la especificación general se requieren los  $\mu_i$  y los  $\beta_i$ , que son  $i \times i = i^2$  parámetros a estimar al mismo tiempo. Aunque desde el punto de vista analítico es interesante identificar de manera individual las constantes y los parámetros de relación, la complicación aparece con la derivación del método de estimación y en la parte computacional. Por ello, es muy importante analizar modelos panel que simplifiquen la cantidad de parámetros a estimar y que sean analíticamente interesante.

De la especificación general, se pueden aplicar restricciones a los parámetros que conlleva dos grupos de modelos. En la especificación general se establece que cada elemento de la sección cruzada de la variable endógena  $y_{i,t}$ , responde diferente a las variables exógenas  $X_{i,t}$ . La heterogeneidad de los efectos se identifican por  $\beta_i$ , pero al aplicar la restricción  $\beta_1 = \dots = \beta_i = \beta$  se supone que cada elemento de la sección cruzada responde de la misma manera a la variables exógenas; esto es lo que se conoce como respuesta común. El modelo que resulta al aplicar las restricciones, tiene como única fuente de heterogeneidad a las constantes individuales identificadas por las  $\mu_i$ ; como en la ecuación 2.

$$y_{i,t} = \mu_i + \beta X_{i,t} + \varepsilon_{i,t} \quad (2)$$

El segundo grupo de modelos se obtiene al aplicar el siguiente nivel de restricciones a las constantes individuales  $\mu_1 = \dots = \mu_i = \mu$ ; esto es lo que se conoce como efecto común. El nuevo modelo panel que resulta, es una especificación que supone homogeneidad en los elementos de la sección cruzada por condiciones iguales ( $\mu$ ) y respuesta igual ( $\beta$ ) a las variables exógenas; ver ecuación 3.

$$y_{i,t} = \mu + \beta X_{i,t} + \varepsilon_{i,t} \quad (3)$$

## **2.1 Supuestos econométricos y la consistencia de los estimadores**

En el modelo estático general como en las dos versiones con las restricciones de los parámetros, se requiere analizar los supuestos clásicos sobre los errores o innovaciones  $\varepsilon_{i,t}$  de los modelos. Por default, se debe cumplir que la media de los errores por corte transversal y serie de tiempo es igual a cero,  $E[\varepsilon_{i,t}] = 0$ ; que la varianza del modelo, dado las variables exógenas, sea constante para cada sección cruzada, pero puede ser diferente entre ellas,  $E[\varepsilon_{i,t}^2 | x_{i,t}] = \sigma_i^2$ ; y, que no exista correlación serial ni contemporánea  $E[\varepsilon_{i,t} \varepsilon_{j,s}] = 0$ , con  $t \neq s$  e  $i \neq j$ ; estos supuestos se puede resumir en una matriz del siguiente tipo:

$$E[\varepsilon_i \varepsilon_i'] = \Omega = \begin{bmatrix} \varepsilon_1 \varepsilon_1' & \cdots & \varepsilon_1 \varepsilon_N' \\ \vdots & \ddots & \vdots \\ \varepsilon_N \varepsilon_1' & \cdots & \varepsilon_N \varepsilon_N' \end{bmatrix} \text{ donde } \varepsilon_i = [\varepsilon_{i,1} \ \varepsilon_{i,2} \ \dots \ \varepsilon_{i,T}]$$

Si la matriz de varianzas y covarianzas cumple con los supuestos econométricos descritos anteriormente, entonces se puede escribir como  $\Omega = \sigma^2 I_N \otimes I_T$ , los estimadores son insesgados y eficientes, y por tanto el modelo se estima con mínimos cuadrados ordinarios (Greene, W., 1998).

El supuesto de consistencia de los parámetros es relativo y depende de la comparación entre los modelos analizados.

*El modelo de panel con efectos comunes (pooled OLS estimator)*

La especificación tipo pool impone restricciones a los parámetros individuales, al establecer que una constante común ( $\mu_1 = \dots = \mu_i = \mu$ ) y efecto común ( $\beta_1 = \dots = \beta_i = \beta$ ) con respecto a las variables exógenas, como la ecuación 3. El estimador *pooled OLS* se obtiene al apilar (stacking) los datos sobre  $i$  y  $t$  con  $NT$  observaciones y aplicando OLS. Si el modelo esta correctamente especificado y las variables exógenas no están correlacionados con los errores, entonces se puede estimar consistentemente. En otras palabras, si se cumple  $Cov = [\varepsilon_{it}, x_{it}] = 0$  entonces  $N \rightarrow \infty$  o  $T \rightarrow \infty$  son suficientes para la consistencia. El estimador *pooled OLS* es inconsistente si el modelo apropiado es el de efectos fijos, debido a que las constantes individuales que no se incluyeron en el modelo *pool* están correlacionadas con las variables exógenas.

*El modelo de panel con efectos fijos en constante (estimador within)*

La restricción que se elimina con el modelo de efectos fijos es que existe una constante individual para cada elemento de la sección cruzada ( $\mu_i$ ). Desde el punto de vista de los estimadores, el estimador *within* a diferencia del *pooled OLS*, explora las características de los datos panel y mide la asociación entre las desviaciones entre los elementos de las variables exógenas desde sus valores promedio en el tiempo y las desviaciones entre los elementos de la variable endógena desde su valor promedio en el tiempo. Los pasos para la estimación consiste en:

- En primer lugar se comienza con el modelo de efectos individuales en constante, en el cual se prueba el caso de  $\mu_i = \mu$
- Entonces se toma el valor promedio en el tiempo

$$\bar{y}_i = \mu + \beta \bar{X}_i + \bar{\varepsilon}_i$$

- Al modelo de efectos individuales en constante se le resta el modelo promedio en el tiempo, con el cual se obtiene el estimador *within*

$$y_{i,t} - \bar{y}_i = \mu_i + \beta[X_{i,t} - \bar{X}_i] + \varepsilon_{i,t} - \bar{\varepsilon}_i$$

*Modelo Panel con efectos aleatorios en constante (feasible GLS estimator)*

En el modelo de efectos aleatorios, se asume que la constante individual tiene una distribución con media y una desviación estándar  $\mu_i \sim [\mu, \sigma_\mu^2]$ , que junto con los errores o innovaciones  $\varepsilon_{it} \sim [0, \sigma_\varepsilon^2]$  configuran las dos partes aleatorias o probabilísticas del modelo panel con efectos aleatorios.

Al estimador que se utiliza se le conoce como el estimador de mínimos cuadrados generalizado factible (*feasible GLS estimator*), que puede calcularse al estimar con mínimos cuadrados ordinario el siguiente modelo transformado

$$y_{i,t} - \hat{\lambda}\bar{y}_i = (1 - \hat{\lambda})\mu_i + \beta[X_{i,t} - \hat{\lambda}\bar{X}_i] + v_{i,t}$$

Donde  $v_{i,t} = (1 - \hat{\lambda})\mu_i + (\varepsilon_{i,t} - \hat{\lambda}\bar{\varepsilon}_i)$  es iid asintóticamente y  $\lambda = 1 - \frac{\sigma_\varepsilon}{\sqrt{\sigma_\varepsilon^2 + T\sigma_\mu^2}}$ . Notar

que  $\hat{\lambda} = 0$  corresponde a pooled OLS,  $\hat{\lambda} = 1$  corresponde al estimador *within* y cuando  $T \rightarrow \infty$  entonces  $\hat{\lambda} \rightarrow 1$ . El estimador para  $\beta$  es básicamente en dos etapas.

El estimador de efectos aleatorios es totalmente eficiente dentro del supuesto del mismo estimador, aunque la eficiencia la gana en realidad al compararse con el estimador pooled y es inconsistente si el modelo de efectos fijos es el correcto.

### **3. ELECCIÓN DE MODELOS ALTERNATIVOS**

El procedimiento de elección de la especificación de la constante del modelo panel con los estimadores *pooled*, efectos fijos (*within*) o efectos aleatorios (*feasible GLS estimator*), es el siguiente.

#### **3.1 Modelo de efectos individuales (IE) versus el modelo Pool**

En sentido estricto, se tiene que comparar los modelos de efectos individuales (fijos y aleatorias) con respecto al modelo pool. Sin embargo, es tradición comparar en esta primera fase el modelo de efectos fijos con el modelo pool, para comparar la eficiencia del primero.

##### *Prueba pooling*

Para ello, se utiliza una prueba de restricción de parámetros entre los dos modelos y se analizan las hipótesis:

$$H_h: \forall \mu_i = 0$$

$$H_a: \mu_1 \neq 0, \dots, \mu_i \neq 0$$

Para analizar las hipótesis, se utiliza un estadístico  $\lambda^2(k)$  con lo k grados de libertad definidos por la cantidad de efectos individuales; a esta prueba se le conoce como *pooling*.

#### **3.2 Modelo de efectos aleatorios (EA) versus el modelo de efectos fijos (EF)**

En el caso de que el modelo de *efectos fijos* sea eficiente en comparación con el modelo *pooled*, entonces se puede analizar si el modelo de efectos aleatorios es eficiente en comparación con el modelo de efectos fijos. Para probar la consistencia del modelo panel con efectos aleatorios, se utiliza la prueba de *Hausman*.

### *Prueba de Hausman*

Las hipótesis que se utiliza para analizar la consistencia se resume en:

$H_0$ : Estimador EA es consistente con respecto al estimado EF

$H_a$ : Estimador EF es consistente con respecto al estimador EA

El estadístico para probar se define como

$$\lambda^2(k) : H = [\beta_{EF} - \beta_{EA}]^T [Cov(\beta_{EF}) - Cov(\beta_{EA})]^{-1} [\beta_{EF} - \beta_{EA}]$$

Donde  $\beta$  es el vector de coeficientes compuesto con los parámetros  $[\beta \gamma]$  del modelo, Cov es la matriz de varianza-covarianza y k es el número de coeficientes.

## **4. RESULTADOS DE LOS MODELOS ECONOMÉTRICOS PANEL CON EL PAQUETE PLM DE R.**

### **4.1 La curva de Philips para las ciudades de México**

El modelo de la curva de Phillips tradicionalmente explica la inflación con base a las expectativas, a los factores de demanda y de oferta (Varela y Torres, 2009). La especificación de un modelo estático general de la curva de Phillips para las  $i$  ciudades de México, se puede escribir como en la ecuación 4.

$$\pi_{i,t} = \mu_i + \beta_i(u_{i,t} - u_{i,t}^*) + \gamma_i z_{i,t} + \varepsilon_{i,t} \quad (4)$$

Donde la inflación para cada ciudad  $\pi_{i,t}$  se explica por una constante individual  $\mu_i$ , por el exceso de demanda que se deriva al observar que la tasa de desempleo

actual se encuentra por arriba de la tasa de desempleo natural o potencial  $u_{i,t}^*$ , que en la literatura se le conoce como el componente del desempleo actual que no está correlacionado con la inflación de largo plazo (NAIRU), y por variables de oferta y/o de política monetaria  $z_{i,t}$ .

## 4.2 Aplicación de los modelos panel con R

En la página de R (<http://www.r-project.org>) se describe como un software libre y en desarrollo, para computar estadística y gráficas. Esta compilado y corre en una variedad de plataforma UNIX, Windows y MacOS. De las ventajas de esta plataforma es que existen un número cada vez mayor de paquetes, rutinas o programas, con los cuales se puede hacer econometría aplicada. Para la aplicación en R, se utiliza el paquete de econometría de datos panel *plm* desarrollado por Croissant y Millo (2008).

### Datos

La base de datos fue construida por Mendoza, M.A. (2013) y contiene la inflación (INF), medida por la tasa de crecimiento del índice nacional de precios al consumidor por ciudad, la tasa de desocupación (U) por ciudad, la tasa de desocupación natural (UN) por ciudad, estimada con el filtro *Hodrick-Prescot*, y la tasa de interés medida con CETES a 28 días (CETES28).

El índice nacional de precios al consumidor y los CETES28 se construyeron con base al Banco de México (BANXICO) y la tasa de desocupación al INEGI.

### *Análisis Exploratorio*

Con el objetivo de llevar a cabo el análisis exploratorio, desde la consola de *R*, se escribió el siguiente comando para tener activo la base de datos construida y guardada previamente en el formato de *R* (*RData*).

*Comandos en R:*

```
> load("C:/R/InflacionDesempleoCiudades.RData")
```

La instrucción *load* tiene que incluir el lugar donde se localiza el archivo (C:/R/) y su nombre (InflacionDesempleoCiudades.RData).

Para obtener el resumen de los estadísticos de la base de datos, se utiliza el comando *summary* con el nombre de la tabla en la base de datos entre paréntesis, que en este caso se le asignó el nombre de Datos.

*Comandos en R*

```
> summary(Datos)
```

Ciudad	Periodo	Nom_Ciudad	INF
Min. : 1.00	Min. :1995	LEON : 17	Min. : 2.30
1st Qu.: 7.75	1st Qu.:1999	ACAPULCO : 16	1st Qu.: 4.40

Median :14.50	Median :2002	AGUASCALIENTES: 16	Median : 5.50
Mean :14.50	Mean :2002	CAMPECHE : 16	Mean :11.17
3rd Qu.:21.25	3rd Qu.:2006	CD DE MEXICO : 16	3rd Qu.:16.30
Max. :28.00	Max. :2010	CHIHUAHUA : 16	Max. :41.70
(Other) :351			
U	UN	CETES28	U_UN
Min. :0.700	Min. :1.012	Min. : 4.300	Min. :-1.565e+00
1st Qu.:2.300	1st Qu.:2.829	1st Qu.: 6.732	1st Qu.:-5.513e-01
Median :3.350	Median :3.470	Median : 8.120	Median :-6.304e-02
Mean :3.531	Mean :3.531	Mean :14.319	Mean :-6.691e-14
3rd Qu.:4.500	3rd Qu.:4.167	3rd Qu.:17.500	3rd Qu.: 4.226e-01
Max. :8.000	Max. :6.989	Max. :48.620	Max. : 2.794e+00

En la tabla se incluyen en las dos primeras columnas las variables que identifican las ciudades (sección cruzada) y el periodo (tiempo) del formato panel, donde los valores mínimos y máximos indican 28 ciudades y una serie anual de 1995 a 2010. También se incluye una columna con el nombre de las 28 ciudades y adicionalmente a las variables INF, UN y CETES28 que se describieron anteriormente, se incluye la variable U\_UN que es la diferencia de la tasa de desocupación observada y la tasa de desocupación natural (NAIRU).

Con el objetivo de analizar la relación entre inflación, la diferencia de la tasa de desocupación observada y la tasa natural, y la tasa de interés de Cetes, se construyeron las matrices de diagramas de dispersión para los períodos 1995-2010.

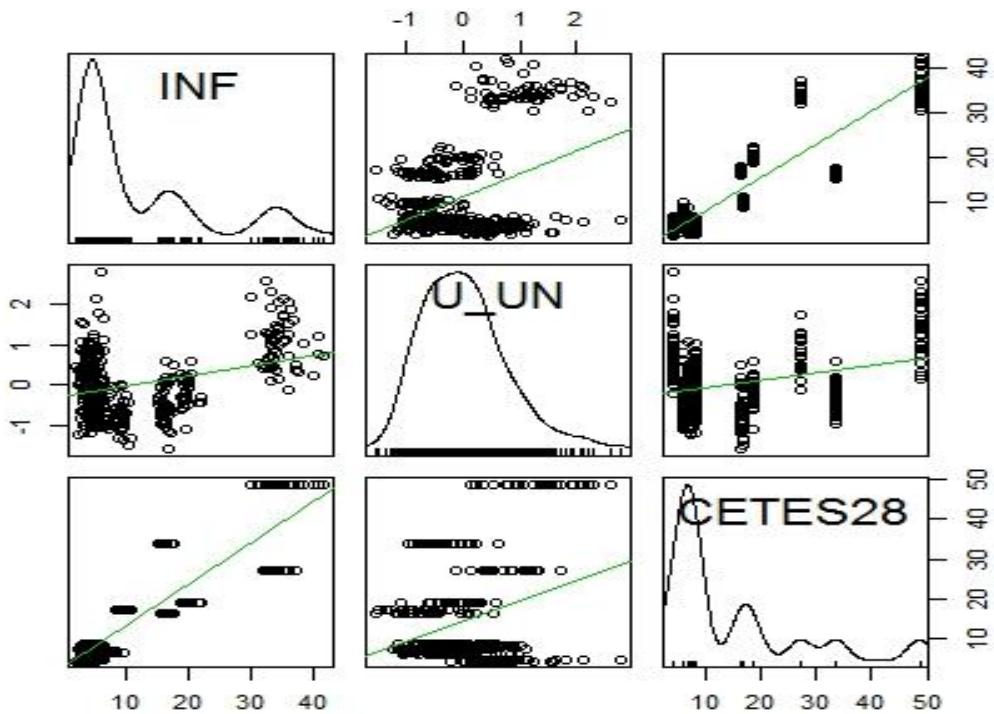
### *Matrices de diagramas de dispersión 1995-2010*

El comando para construir matriz en *R* incluye la especificación de las variables que se analizaran en los diagramas de dispersión, el tipo de línea de relación que en este caso es por medio de una línea de regresión, no se incluye una línea suavizada ni su correspondiente amplitud, en la diagonal se especifica la función de densidad para cada variable de la matriz.

### *Comandos en R*

```
> scatterplotMatrix(~INF+U_UN+CETES28, reg.line=lm, smooth=FALSE,  
spread=FALSE,span=0.5, diagonal = 'density', data=Datos, Periodo < 2001)
```

Figura 1: Matriz de diagramas de dispersión 1995-2010 entre inflación, desempleo y tasa de interés



En la diagonal de la matriz, se observa que la inflación y la tasa de interés presentan más de una moda y el desempleo solo una, pero es interesante como en los tres casos las modas con las mayores concentraciones se localizan en los niveles más bajos. Con respecto al diagrama de dispersión entre inflación y desempleo (segundo diagrama de la primera fila de la matriz) la relación es claramente positiva y se identifican tres grandes concentraciones relacionadas con las modas de la inflación. En cuanto a la relación inflación y tasa de interés (tercer diagrama de la primera fila), parece que la relación es positiva y muy fuerte.

### *Análisis confirmatorio*

En este apartado se estiman los tres tipos de modelos panel (*pool*, *efectos fijos* y *aleatorios*) y se aplican las pruebas de pooling y Hausman para tomar la decisión del mejor modelo en términos de consistencia.

#### *Estimación Pool*

El comando para estimar el modelo pool incluye de derecha a izquierda, la especificación del modelo (*pooling*), la fuente de información que en nuestro caso se encuentra en Datos que está definida dentro de la base de datos de *InflacionDesempleoCiudades.RData*, la especificación de la función de inflación con respecto al diferencial del desempleo y los CETES a 28 días (*INF~U\_UN+CETES28*), el comando que llama a la programación de los modelos panel (*plm*) y la instrucción para asignar el resultado (<-) en un objeto llamado *modelo.pool*.

#### *Comandos en R*

```
> modelo.pool <- plm(INF~U_UN+CETES28, data = Datos, model = "pooling")
```

Una vez, que se estimó el modelo panel tipo *pool* se puede observar los resultados con el comando *summary* y entre paréntesis el objeto asignado.

*Comandos en R*

```
> summary(modelo.pool)
```

Oneway (individual) effect Pooling Model

Call:

```
plm(formula = INF ~ U_UN + CETES28, data = Datos, model = "pooling")
```

Balanced Panel: n=28, T=16, N=448

Residuals :

Min. 1st Qu. Median 3rd Qu. Max.

-9.530 -2.500 -0.958 1.540 16.100

Coefficients :

	Estimate	Std. Error	t-value	Pr(> t )
--	----------	------------	---------	----------

(Intercept)	0.876198	0.356340	2.4589	0.01432 *
-------------	----------	----------	--------	-----------

U_UN	1.445834	0.319420	4.5264	7.71e-06 ***
------	----------	----------	--------	--------------

CETES28	0.718582	0.019378	37.0820	< 2.2e-16 ***
---------	----------	----------	---------	---------------

---

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

```
Total Sum of Squares: 46897  
Residual Sum of Squares: 9964.2  
R-Squared : 0.78753  
Adj. R-Squared : 0.78226  
F-statistic: 824.704 on 2 and 445 DF, p-value: < 2.22e-16
```

### *Estimación con Efectos Fijos*

Para estimar el modelo de efectos fijos se utiliza la misma línea de comandos que en el caso anterior pero con dos modificaciones: 1) En la especificación del tipo de modelo, se cambia “pooling” por “within”; y, 2) la estimación del modelo se asignan a un nuevo objetivo llamado modelo.ef. De la misma manera, para ver los resultados se utiliza el comando *summary*.

### *Comandos en R*

```
> modelo.ef <- plm(INF~U_UN+CETES28, data = Datos, model = "within")
```

```
> summary(modelo.ef)
```

Oneway (individual) effect Within Model

Call:

```
plm(formula = INF ~ U_UN + CETES28, data = Datos, model = "within")
```

Balanced Panel: n=28, T=16, N=448

Residuals :

Min. 1st Qu. Median 3rd Qu. Max.

-9.740 -2.570 -0.959 1.490 16.000

Coefficients :

	Estimate	Std. Error	t-value	Pr(> t )
--	----------	------------	---------	----------

U_UN	1.445834	0.328766	4.3978	1.389e-05 ***
------	----------	----------	--------	---------------

CETES28	0.718582	0.019945	36.0279	< 2.2e-16 ***
---------	----------	----------	---------	---------------

Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 46848

Residual Sum of Squares: 9915.4

R-Squared : 0.78835

Adj. R-Squared : 0.73556

F-statistic: 778.484 on 2 and 418 DF, p-value: < 2.22e-16

### *Modelo de Efectos Aleatorios*

Finalmente para estimar el modelo de efectos aleatorios, se especifica el modelo con “*random*” y el mecanismo para estimar la varianza, que en este caso se utiliza

el método de amemiya. Como en los casos anteriores se utiliza el comando *summary*, para ver los resultados de la estimación.

### *Comandos en R*

```
> modelo.ea <- plm(INF ~ U_UN+CETES28, data = Datos, model = "random",random.method="amemiya")
```

```
> summary(modelo.ea)
```

Oneway (individual) effect Random Effect Model

(Amemiya's transformation)

Call:

```
plm(formula = INF ~ U_UN + CETES28, data = Datos, model = "random",  
random.method = "amemiya")
```

Balanced Panel: n=28, T=16, N=448

Effects:

```
var std.dev share
```

```
idiosyncratic 23.72 4.87 1
```

```
individual 0.00 0.00 0
```

theta: 0

Residuals :

Min. 1st Qu. Median 3rd Qu. Max.

-9.530 -2.500 -0.958 1.540 16.100

Coefficients :

Estimate Std. Error t-value Pr(>|t|)

(Intercept) 0.876198 0.356340 2.4589 0.01432 \*

U\_UN 1.445834 0.319420 4.5264 7.71e-06 \*\*\*

CETES28 0.718582 0.019378 37.0820 < 2.2e-16 \*\*\*

---

Signif. codes: 0 '\*\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

Total Sum of Squares: 46897

Residual Sum of Squares: 9964.2

R-Squared : 0.78753

Adj. R-Squared : 0.78226

F-statistic: 824.704 on 2 and 445 DF, p-value: < 2.22e-16

Como se mencionó previamente, el procedimiento para elegir los modelos es el modelo *pool* versus el de efectos fijos y en segundo lugar el modelo de efectos fijos versus el modelo de efectos aleatorios.

*Modelo de efectos fijos (ef) versus el modelo Pool (pool)*

*Comandos en R*

```
> pooltest(modelo.pool, modelo.ef)

F statistic

data: INF ~ U_UN + CETES28
F = 0.0763, df1 = 27, df2 = 418, p-value = 1
alternative hypothesis: instability
```

Los resultados muestran que el mejor modelo es de tipo *pool*, por lo que se puede concluir que es *consistente* con respecto al modelo de *efectos fijos*.

*Modelo de efectos aleatorios (ea) versus el modelo de efectos fijos (ef)*

En sentido estricto, como el modelo *pool* fue mejor que el de *efectos fijos* no es necesario hacer la prueba de *Hausman* para elegir entre el modelo de efectos fijos y aleatorios. Sin embargo, para fines ilustrativos se muestran los comandos y los resultados de la prueba de *Hausman*.

*Comandos en R*

```
> phtest(modelo.ea, modelo.ef)
```

Hausman Test

```
data: INF ~ U_UN + CETES28  
chisq = 0, df = 2, p-value = 1  
alternative hypothesis: one model is inconsistent
```

Aunque la prueba de *Hausman* indica que el mejor modelo es el de efectos aleatorios, no tiene sentido debido a que el modelo *pool* es el adecuado.

### *Consideraciones finales sobre los resultados*

Con el objetivo de probar en el nivel más general la hipótesis sobre desempleo e inflación, en este trabajo se estimaron modelos panel con el paquete *plm* del software *R* en sus tres especificaciones alternativas (*pool*, efectos fijos o aleatorios). Con las pruebas *pooling* y de *Hausman* se encontró que el modelo *pool* es consistente con respecto a los modelos de efectos fijos y aleatorio y de acuerdo a los resultados econométricos se concluyó que tanto la tasa de desempleo bajo el mecanismo NAIRU como la tasa de interés, tienen un efecto positivo y homogéneo sobre el proceso inflacionario de las ciudades para el periodo 1995-2010.

## **REFERENCIAS**

- Cameron, C. y Trivedi, P. (2005) *Microeometrics, Methods and Applications*, primera edición, Cambridge University Press.
- Croissant y Millo (2008), Panel Data Econometrics in R: The *plm* Package, en Journal of Statistical Software.
- Frees, E. (2004) *Longitudinal and Panel Data, Analysis and Applications in the Social Sciences*, Cambridge University Press.
- Greene, W. (1998) *Análisis Econométrico*, Prentice Hall, Tercera edición.

Hsiao, C. 2003. *Analysis of Panel Data*. Cambridge University Press: segunda edición.

Mendoza, M.A. (2013) *Inflación y desempleo en ciudades de México: una evaluación con modelos panel y aplicaciones en software R*,

Varela, R. y Torres, V. (2009) Estimación de la tasa de desempleo no aceleradora de la inflación en México, Análisis Económico, Núm. 57, vol. XXIV.

Wooldridge, J. (2002) *Econometric Analysis of Cross Section and Panel Data*, Massachusetts Institute of Technology.

## **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

InflacionDesempleoCiudades.RData

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap14

Práctica\_Cap14

VideoPráctica\_Cap14

VideoTeoría\_Cap14

# CAPÍTULO 15: ECONOMETRÍA ESPACIAL Y SUS APLICACIONES EN R

Miguel Ángel Mendoza González y Luis Quintana Romero

## 1. INTRODUCCION

Como estudioso de los fenómenos económicos, sociales o ambientales seguramente se habrá percatado que cada vez se realiza una mayor difusión de información georeferenciada. Es decir, las variables aparecen vinculadas a su dimensión espacial y pueden ser manejadas en potentes mapas en los que se van superponiendo capas de información.

A la par de la difusión de datos georeferenciados, también se han desarrollado modernos paquetes computacionales, conocidos como *Sistemas de Información Geográfica* (GIS por sus siglas en inglés), ello ha permitido dar impulso a una novel subdisciplina de la econometría conocida como econometría espacial.

La econometría espacial fue definida a principios de los años setenta por Jean Paelinck como el creciente cuerpo de la literatura en ciencia regional que trata primordialmente con la estimación y prueba de problemas encontrados en la implantación de modelos econométricos multirregionales.<sup>25</sup>

---

<sup>25</sup> Véase, Luc Anselin (1988) Spatial Econometrics: Methods and Models, Kluwer Academic Publishers, p.7.

Luc Anselin (1988) uno de los pioneros y grandes impulsores de la econometría espacial considera que el campo de esta disciplina está formado por:

“...aquellos métodos y técnicas que, sustentados en una representación formal de la estructura de la dependencia y heterogeneidad espacial, provee el medio para llevar a cabo la adecuada especificación, estimación, prueba de hipótesis y predicción para modelos en la ciencia regional.”<sup>26</sup>

Los métodos desarrollados por la econometría espacial permiten atender problemas de violación a los supuestos del modelo de regresión, que no es posible resolverlos en el marco de los modelos estadísticos estándar. Estos problemas son típicos en los datos espaciales y se refieren a:

- 1) Dependencia espacial entre observaciones: Correlación espacial.
- 2) Heterogeneidad espacial entre observaciones: Heteroscedasticidad espacial.

El caso al que se le ha dedicado mayor atención es al primero, debido a que el segundo ha podido estudiarse en el marco de modelos de panel y otras técnicas similares en donde la heterocedasticidad y el cambio estructural juegan un papel relevante.

En este capítulo se abordarán los siguientes temas:

- Vecindad
- Dependencia espacial

---

<sup>26</sup> Ibid, p.10.

- Estadísticos de dependencia espacial
- Regresión espacial
- Selección de modelos espaciales

## **2. VECINDAD Y DEPENDENCIA ESPACIAL**

Usualmente cuando el economista maneja series económicas, sociales o ambientales lo hace desde una perspectiva en la cual toma como dadas las coordenadas de localización geográfica de las variables: Es decir, por ejemplo, cuando analiza los precios o la producción no hace referencia a su ubicación geográfica específica; se aísla a la variable de su contexto espacial.

Obviar el contexto espacial significa una perdida importante de información, sólo piense lo que ocurriría si, por ejemplo, un estudio de criminalidad en una ciudad no considerará el efecto que tiene la situación que priva en las ciudades vecinas.

En ese sentido, los datos generalmente presentan algún tipo de dependencia o autocorrelación espacial, la cual puede definirse como la existencia de una relación funcional entre lo que ocurre en un punto del espacio y lo que sucede en otro lugar, lo cual se explica fundamentalmente por razones de interacción humana con su entorno físico-ambiental.<sup>27</sup>

La dependencia espacial implicaría que al tomar en consideración una variable, para diferentes localidades, esperaríamos características más similares en localidades vecinas, que en aquéllas separadas por grandes distancias. La dependencia espacial puede ser positiva o negativa, de ser positiva la presencia de un atributo en una localidad se extendería a las regiones vecinas y, en caso de ser negativa, obstaculizaría su presencia en sus vecindades.

---

<sup>27</sup> Véase Anselin, op.cit. p.10

Los datos espaciales se pueden clasificar de acuerdo con el objeto espacial al que se refieren y al nivel de medida de las variables. Dicha clasificación puede ilustrarse matricialmente como en la figura 1:<sup>28</sup>

Figura 1

Matriz de datos espaciales

$z_1(1)$	$z_2(1)$	...	$z_k(1)$	$s(1)$	Caso 1
$z_1(2)$	$z_2(2)$	...	$z_k(2)$	$s(2)$	Caso 2
.	.	.	.	.	.
.	.	.	.	.	.
$z_1(n)$	$z_2(n)$	...	$z_k(n)$	$s(n)$	Caso n

Donde tenemos k variables  $\{z_1, z_2, \dots, z_k\}$  medidas en la localización  $s(i)$  donde  $i=1,2,\dots,n$ . Si incorporamos el factor de temporalidad, podríamos tener una matriz de este tipo para cada período del tiempo. Las relaciones entre las variables y localizaciones clasificadas en la matriz de datos pueden establecerse a través de conectividad o vecindad.

#### *Matriz de vecindad por contigüidad*

La noción de vecindad se puede establecer de forma binaria; en tal caso, si dos unidades espaciales tienen una frontera común se les asigna un uno, en caso contrario se le asigna un cero. Bajo esta sencilla idea, una variable particular podría referenciarse en un mapa, a partir del cual es posible establecer sus fronteras y, en consecuencia, identificar sus vecindades. Luc Anselin (1988)

---

<sup>28</sup> La matriz de datos espaciales fue retomada del libro de Robert Haining, (2003) Spatial Data Analysis, Theory and practice.

plantea diferentes medidas de vecindad, las cuales se asemejan a un tablero de ajedrez y que podemos apreciar en la figura 2:

Figura 2

Diferentes vecindades

		B		
	B	A	B	
		B		

TORRE

	C		C	
		A		
	C		C	

ALFIL

	C	B	C	
	B	A	B	
	C	B	C	

REINA

La vecindad entre puntos también puede ser de orden superior, sí se consideran series de bandas concéntricas alrededor de la localidad bajo consideración.

Figura 3

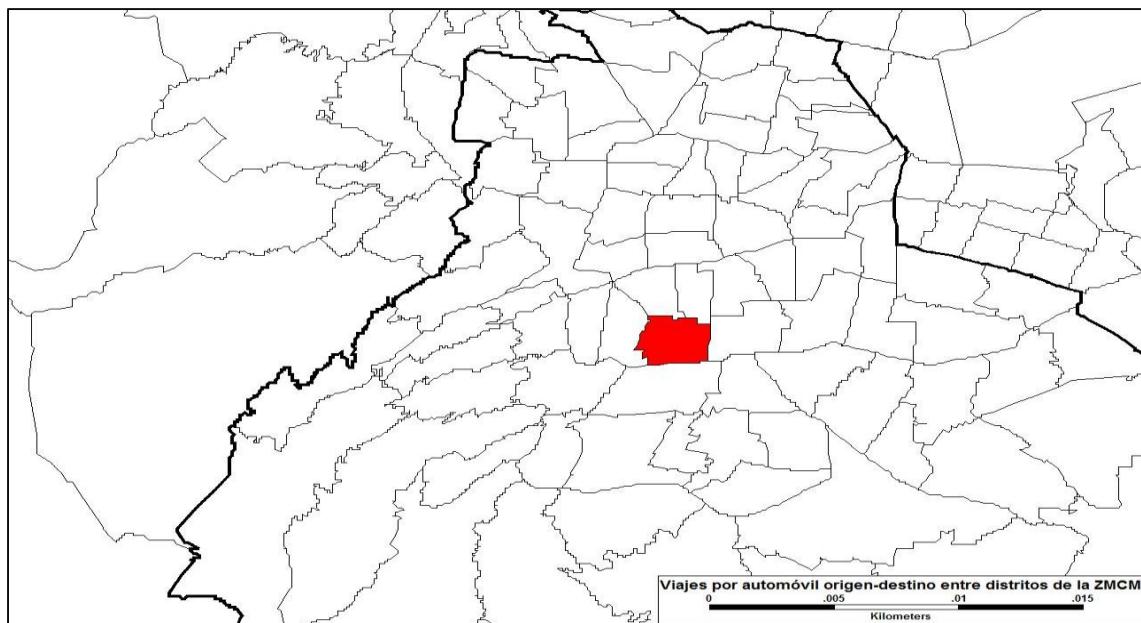
Vecindad de orden superior

		D		
	C	B	C	
D	B	A	B	D
C	B	C		
		D		

Por ejemplo, en la figura 3 y considerando vecindad tipo torre, las celdas C y D son contiguas de segundo orden a la celda A, y son contiguas de primer orden a B.

En un mapa geográfico, como en la figura 4, es posible construir cualquiera de los tipos de matrices de vecindad descritas anteriormente:

Figura 4: Distritos de viajes de origen-destino de la ZMVM



Fuente: Elaboración propia con Encuesta Origen-Destino, INEGI (2007)

### Ejemplo 1. Construcción analítica de una matriz de vecindad

Para ilustrar la forma en la que se construye una matriz binaria de vecindades retomamos el ejemplo presentado por Anselin (1988) en su libro ya citado anteriormente. Suponga que la localización de diferentes variables podría ubicarse en un mapa cuadriculado como el siguiente:

	1		<sup>2B</sup>		3			
	<sup>4B</sup>		<sup>5A</sup>		<sup>6B</sup>			
	7		<sup>8B</sup>		9			

A cada localidad le asignamos un número y tomando como punto de referencia la localidad 5 calculamos vecindades tipo torre. Por ejemplo, la localidad 1 y la 3 no tienen vecindad, por ello se le asigna un cero en la matriz de vecindades. La vecindad de una localidad consigo misma es contabilizada también con un cero. La matriz de contactos resultante aparece en la figura siguiente:

Localidad	1	2	3	4	5	6	7	8	9
1	0	1	0	1	0	0	0	0	0
2	1	0	1	0	1	0	0	0	0
3	0	1	0	0	0	1	0	0	0
4	1	0	0	0	1	0	1	0	0
5	0	1	0	1	0	1	0	1	0
6	0	0	1	0	1	0	0	0	1
7	0	0	0	1	0	0	0	1	0
8	0	0	0	0	1	0	1	0	1
9	0	0	0	0	0	1	0	1	0

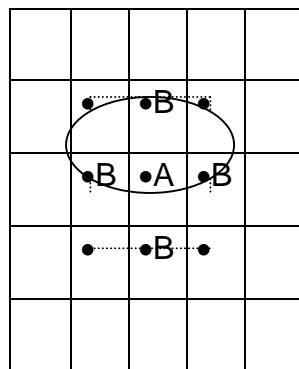
### *Matriz de vecindad por distancia*

La matriz de vecindades binarias es limitada, ya que únicamente considera la vecindad física, por lo cual no contabiliza la posibilidad de interacción entre regiones alejadas. Por ello, han sido propuestas otras medidas de vecindad alternativas, sustentadas en distancias de diferente tipo y cuya matriz,  $\mathbf{W}$ , es conocida como la matriz de pesos o contactos espaciales.<sup>29</sup>

Anselin plantea que, en caso de que la unidad espacial sea un sistema urbano, la vecindad puede ser obtenida de la trayectoria más corta en una red o gráfica formada por una conexión de puntos. Por ejemplo, en la figura 5, la distancia más corta entre los puntos es representada por la línea punteada y la vecindad por el círculo que conecta los puntos y tiene como centroide a la localidad A.

Figura 5

Vecindad por distancia más corta



---

<sup>29</sup> Anselin (1988) expone ampliamente las medidas propuestas por Cliff y Ord, Dacey, Bodson y Peters para construir diferentes tipos de matrices de contactos.

Considerando los centroides como punto de referencia para medir las distancias geográficas, Fotheringham, Brunsdon y Charlton (2000) proponen las siguientes medidas de distancias:

### I. Localización en el plano cartesiano

En un sistema cartesiano, la distancia se mide por el teorema de Pitágoras y

la localización es por medio de las coordenadas geográficas: *latitud* y *longitud*.

#### 1. *Distancia Euclíadiana*

Con base a las coordenadas de latitud ( $x$ ) y la *longitud* ( $y$ ), la distancia entre los centroides de las localidades  $i$  y  $j$ :

$$d_{i,j} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2}$$

La distancia euclíadiana entre dos localidades  $i$  y  $j$  con coordenadas  $(x_{i,1}, x_{i,2})$ ,  $(x_{j,1}, x_{j,2})$ , se puede escribir también como:

$$d_E(i, j) = \left[ \sum_{k=1}^2 (x_{ik} - x_{jk})^2 \right]^{1/2}$$

La distancia puede ser generalizada a  $m$  dimensiones.

$$d_E(i, j) = \left[ \sum_{k=1}^m (x_{ik} - x_{jk})^2 \right]^{1/2}$$

## 2. Métrica de Minkowski

En el caso de que  $p=2$  es la distancia euclíadiana, si  $p=1$  es la distancia conocida como *Manhattan* o distancia *taxicab*.

$$d_E(i, j) = \left[ \sum_{k=1}^m |x_{ik} - x_{jk}|^p \right]^{\frac{1}{p}}$$

## II. Localización en el globo o superficie de la tierra

En el caso de considerar la superficie de la tierra en lugar del plano cartesiano, se necesita de los cálculos geométricos:

### 1. Trigonometría esférica (curvatura de la tierra)

$$S_{ij} = R \cdot \text{arcos}[\cos(90^\circ - \Phi_i) \cos(90^\circ - \Phi_j) + \sin(90^\circ - \Phi_i) \sin(90^\circ - \Phi_j) \cos(\lambda_j - \lambda_i)]$$

$R$  es el radio de la tierra, arcoseno (arcos), coseno (cos), seno (sen), la latitud y longitud de la locación i son  $(\Phi_i, \lambda_i)$

### 2. Mercator (proyección a una forma cilíndrica)

$$x = R\lambda$$

$$y = R \ln(\tan(\pi/4 + \phi/2))$$

Donde  $R$  es el radio de la tierra,  $\ln$  es el logaritmo natural, tangente ( $\tan$ ),  $\phi$  es la latitud y  $\lambda$  es la longitud.

### 3. Lambert (proyección a un área cilíndrica)

$$x = R\lambda$$

$$y = R \sin \phi$$

#### **Ejemplo 2. Librerías de R, transformación de formatos de capas de polígonos a R y lectura de bases de datos (DataFrame)**

Las librerías que se utilizan para la estimación de los modelos espaciales son: Tools for Reading and Handling Spatial Objects (**maptools**); Spatial Dependence; Weighting Schemes, Statistics and Models (**spdep**); ColorBrewer Palettes (**RColorBrewer**); y Choose Univariate Class Intervals (**classInt**). Los cuales deben ser instalados previamente.

Para comenzar el ejercicio, lo primero que se hace es activar las librerías

```
library(maptools)
```

```
library(spdep)
```

```
library(RColorBrewer)
```

```
library(classInt)
```

### Elegir y fijar el directorio de trabajo

```
setwd("/Capitulo_14/BaseDatos_Capitulo14_R")
```

Para leer y transformar formatos shape de cartografía de polígonos a R, se aplica el comando de `readShapePoly` a los archivos de `Zona_Centro` y se graban en el objeto empleo.

```
empleo <- readShapePoly("Zona_Centro.shp")
```

En el objeto empleo ahora se puede consultar el contenido de la base de datos

```
> summary(empleo)
```

Object of class `SpatialPolygonsDataFrame`

Coordinates:

	min	max
x	-87.84584	-87.17459
y	24.56797	25.16453

Is projected: NA

proj4string : [NA]

Data attributes:

	ID	CVEGEO	NOM_ENT
Min.		:266.0 09002 : 1	Distrito Federal: 16
1st Qu.		:684.2 09003 : 1	Morelos : 33
Median		:727.5 09004 : 1	M\332xico :125

Mean	:714.4	09005	:	1
3rd Qu.	:770.8	09006	:	1
Max.	:927.0	09007	:	1
(Other):168				
	NOM_MUN	POBTOT	POBMAS	
Zacualpan	: 2	Min. : 4051	Min. : 2012	
-lvaro Obreg <be>n</be>	: 1	1st Qu.: 18469	1st Qu.: 9054	
Acambay	: 1	Median : 44852	Median : 22188	
Acolman	: 1	Mean : 148300	Mean : 71778	
Aculco	: 1	3rd Qu.: 136118	3rd Qu.: 67958	
Almoloya de Alquisiras	: 1	Max. :1815786	Max. :880998	
(Other)	:167			
.....				

Para poder analizar la distribución del empleo y del capital humano en los municipios de la zona centro del país, primero se generan las variables. Para ello se asignan el logaritmo natural de las variables de población ocupada (POCUPADA), los años de escolaridad en población igual o mayor de 15 años (ESCOLA\_15.) y su logaritmo.

# Logaritmo del Empleo

```
empleo <- log(empleo$POCUPADA)
```

# Capital Humano y logaritmo del capital humano (años de escolaridad promedio)

```
ch <- empleo$ESCOLA_15
```

```
lch <- log(empleo$ESCOLA_15)
```

```
> summary(empleo$POCUPADA)
Min. 1st Qu. Median Mean 3rd Qu. Max.
1158 6426 14470 59570 49050 752300
> summary(empleo$ESCOLA_15)
Min. 1st Qu. Median Mean 3rd Qu. Max.
3.130 5.270 6.165 6.373 7.008 12.680
> summary(lempleo)
Min. 1st Qu. Median Mean 3rd Qu. Max.
7.054 8.768 9.580 9.882 10.800 13.530
> summary(lch)
Min. 1st Qu. Median Mean 3rd Qu. Max.
1.141 1.662 1.819 1.821 1.947 2.540
```

Los indicadores estadísticos muestran que la media del empleo es de 59,570 trabajadores y la mediana de 14,470, lo cual indica que la función de distribución se sesga hacia la izquierda. Mientras que para el caso del capital humano la media y la mediana son muy parecidos; 6.373 y 6.156 años de escolaridad respectivamente.

### 3. ESTADÍSTICOS DE DEPENDENCIA ESPACIAL

Para la medición de dependencia espacial se han propuesto numerosos estadísticos, uno de los más utilizados es el índice de Moran (1948), que se define en la fórmula siguiente:

$$I = \frac{R}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (x_i - \bar{x})(x_j - \bar{x})}{\sum_i (x_i - \bar{x})^2}$$

donde  $x_i$  es la variable cuantitativa en la región  $i$ ,  $\bar{x}$  es su media muestral,  $w_{ij}$  son los pesos de la matriz  $\mathbf{W}$ ,  $R$  es el tamaño de muestra (Regiones); y,

$$E(I) = \frac{-1}{R - 1}$$

$$V(I) = \frac{RS_4 - S_3 S_1 (1 - 2R)}{(R - 1)(R - 2)(R - 3)(\sum_i \sum_j w_{ij})^2}$$

$$S_1 = \frac{1}{2} \sum_i \sum_j (w_{ij} + w_{ji})^2$$

$$S_2 = \sum_i \left( \sum_j w_{ij} + \sum_j w_{ji} \right)^2$$

$$S_3 = \frac{R^1 \sum_i (x_i - \bar{x})^4}{(R^1 \sum_i (x_i - \bar{x})^2)^2}$$

$$S_4 = (R^2 - R + 3)S_1 - RS_2 + 3 \left( \sum_i \sum_j w_{ij} \right)^2$$

Bajo la hipótesis nula de no autocorrelación, el estadístico de Moran es asintóticamente normal:

$$I^* = \frac{I - E(I)}{\sqrt{V(I)}}$$

El índice de Moran sigue una distribución normal estandarizada en muestras grandes (Vaya y Moreno, 2000), de forma tal que un valor positivo (negativo) significativo del índice  $Z(I)$  llevará al rechazo de la hipótesis nula de no autocorrelación espacial y a la aceptación de autocorrelación espacial positiva (negativa).

Es posible graficar la información del índice en un diagrama de dispersión de Moran. Dicho diagrama, presenta en el eje horizontal a la variable  $x$  normalizada y en el eje vertical a la variable multiplicada por la matriz de pesos  $W$ , lo cual da lugar al retardo espacial de dicha variable. La visualización de un patrón aleatorio en la gráfica brinda evidencia de la ausencia de autocorrelación espacial.

### *Dependencia espacial*

La *dependencia temporal*, como la correlación serial, es *unidireccional* (el pasado explica el presente), mientras que la *dependencia espacial* es *multidireccional* (una región puede estar afectada no solamente por otra región contigua o vecina sino por otras que la rodean, al igual que ella puede afectar a las otras). Este hecho imposibilita la utilización del operador rezago  $L$ ,  $L^p Y_t = Y_{t-p}$ , presente en el contexto temporal, para el análisis de la dependencia espacial. La solución consiste en utilizar la matriz  $\mathbf{W}$  de efectos espaciales como operador de rezago espacial, que

se puede leer como una media ponderada de los valores vecinos y se define como:

$$WY = \sum_{j=1}^N w_{ij} y_j$$

donde  $y_j$  es el valor que toma el atributo medido en la vecindad  $j$ ,  $w_{ij}$  es un ponderador cuya suma es la unidad.

### Ejemplo 3. Correlación espacial, estadístico y diagrama de dispersión de Moran en municipios de zona centro de México

Para el análisis de correlación espacial se debe elaborar previamente el ejemplo 2, en específico : activas las librerías, cambiar el directorio de trabajo, la lectura de la cartografía en R y la asignación de las variables de trabajo. En este ejercicio se construye la matriz de vecindad tipo Queen estandarizada, se grafica la red de conexión de los centroides, se calcula el estadístico y se gráfica el diagrama de dispersión de Moran.

Lo primero que se genera es la matriz con valores de unos y ceros, de acuerdo a la cartografía.

```
# Construir lista de vecinos tipo Queen de polígonos
```

```
> pr.nb <- poly2nb(empleo, queen=TRUE)
```

En segundo lugar, la matriz se estandariza y transforma en una lista

```
# Matriz de ponderación W estandarizada
```

```
> wqueen <- nb2listw(pr.nb, style="W")
```

Para revisar las características de la matriz se aplica el summary

# Características de la Matriz W tipo Queen

```
> summary(wqueen)
```

Characteristics of weights list object:

Neighbour list object:

Number of regions: 174

Number of nonzero links: 950

Percentage nonzero weights: 3.137799

Average number of links: 5.45977

Link number distribution:

1 2 3 4 5 6 7 8 9 10 11 14

3 3 19 41 32 26 20 16 9 3 1 1

3 least connected regions:

76 94 120 with 1 link

1 most connected region:

116 with 14 links

Weights style: W

Weights constants summary:

n	nn	S0	S1	S2
---	----	----	----	----

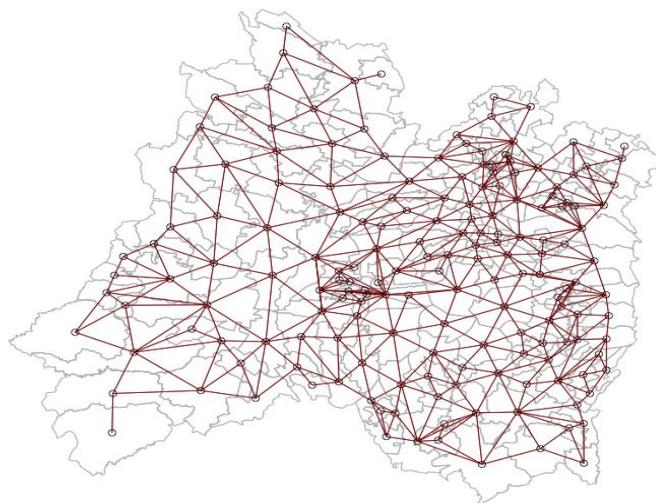
W	174	30276	174	70.54671	724.6509
---	-----	-------	-----	----------	----------

La información generada muestra que la matriz esta conformada con 174 municipios, que de los 174\*174 posibles combinaciones (30,276) 950 links no son ceros, lo cual representa el 3.13% del total de combinaciones; la cantidad promedio de vecinos por municipio es de 5.45; la distribución de los link muestra que tres municipios tienen solamente un vecino, que la mayor cantidad de municipios (41) tienen 4 vecinos y solamente un municipio tiene el máximo de vecinos (14); también nos muestra cuales son los tres municipios con un vecino solamente -los municipios con identificador oid 76, 94 y 120- y el municipio con 14 vecinos -el municipio 116.

Para poder visualizar las conexiones geográficas identificadas se muestra en la siguiente gráfica la red que se construye con los centroides de cada municipio con sus vecinos, de acuerdo a la matriz W tipo Queen.

```
# Grafica con la conexion espacial  
> cent <- coordinates(empleo)  
> plot(empleo, border="grey", lwd=1.5)  
> plot(pr.nb,cent, add=T, col="darkred")
```

Figura 14.1: Red de conexión entre municipios y su vecinos  
en la zona centro de México



Nota: Conexión con base a los centroides y matriz W tipo Queen

Del mapa se observa que los municipios con un vecino son del estado de México, se localizan en la periferia al norte, noreste y suroeste de la región: .... Nopaltepec y Tlatlaya respectivamente. El municipio de Tianguistenco, también del Estado de México, tiene la mayor cantidad de conexiones geográficas y se localiza en el centro de la región.

Para probar si el empleo y el capital humano tienen dependencia espacial, se aplica la prueba de correlación espacial de Moran al logaritmo del empleo y al capital humano y su logaritmo. La hipótesis nula es que la correlación sea cero, lo cual implica que el indicador que se está analizando este aleatoriamente distribuido en la región de estudio<sup>30</sup>; contra la hipótesis alternativa de correlación espacial diferente de cero.

---

<sup>30</sup> Los p-valores se obtienen con un proceso de aleatorización (randomisation), lo cual permite simular la distribución del índice de Moran.

```

# Estadistico de Moran

> moran_lempelo <- moran.test(lempelo, wqueen,randomisation=TRUE,
alternative="two.sided", na.action=na.exclude)

> moran_ch <- moran.test(ch, wqueen,randomisation=TRUE,
alternative="two.sided", na.action=na.exclude)

> moran_lch <- moran.test(lch, wqueen,randomisation=TRUE,
alternative="two.sided", na.action=na.exclude)

#Ver resultados

> print(moran_lempelo)

> print(moran_ch)

> print(moran_lch)

```

Moran's I test under randomisation

data: lempelo

weights: wqueen

Moran I statistic standard deviate = 12.466, p-value < 2.2e-16

alternative hypothesis: two.sided

sample estimates:

Moran I statistic	Expectation	Variance
0.587496960	-0.005780347	0.002264968

Para el caso del logaritmo del empleo se encontró que el coeficiente de correlación de Moran es de 0.5874, lo cual indica que la dependencia global es positiva, y de acuerdo al que el p-value (0.0000000000000022) es menor que

0.05, se acepta que el coeficiente de correlación es estadísticamente diferente de cero.

Moran's I test under randomisation

data: ch

weights: wqueen

Moran I statistic standard deviate = 14.652, p-value < 2.2e-16

alternative hypothesis: two.sided

sample estimates:

Moran I statistic	Expectation	Variance
0.687457813	-0.005780347	0.002238704

Moran's I test under randomisation

data: Ich

weights: wqueen

Moran I statistic standard deviate = 14.843, p-value < 2.2e-16

alternative hypothesis: two.sided

sample estimates:

Moran I statistic	Expectation	Variance
0.698986435	-0.005780347	0.002254493

Para el capital humano se calculó el coeficiente de correlación de Moran para el nivel y su logaritmo para analizar sus diferencias. Los resultados muestran que el coeficiente de correlación de Moran es muy parecido sin y con logaritmos; de 0.6874 y 0.6989 respectivamente; y, que en los dos casos son estadísticamente diferente de cero.

El diagrama de dispersión se utiliza para visualizar la correlación entre el indicador de interés -por ejemplo el logaritmo del empleo- y el rezago espacial multiplicado por el mismo indicador ( $W_{\text{empleo}}$ )- que se calcula en el coeficiente de Moran. Para generar el diagrama de dispersión, se señalan las medias del logaritmo del empleo y de su rezago espacial ( $W_{\text{empleo}}$ ), que divide a los municipios en cuatro grupo (cuadrantes) y que se identifican por un movimiento contrario a las manecillas del reloj, en: Cuadrante I (High-High), superior derecho del diagrama de dispersión, con municipios que se caracterizan por presentar valores numéricos por arriba de la media del indicador y tener vecinos con la misma característica (arriba de la media); En el cuadrante II (Low-High), superior izquierdo del diagrama de dispersión, se identifican los municipios con indicador con valores por debajo de la media y vecinos con la característica contraria (arriba de la media); El cuadrante III (Low-Low), inferior izquierdo del diagrama, contiene a los municipios con indicador por debajo de la media y vecinos con la misma característica; y, finalmente el cuadrante IV (High-Low) con valores por arriba de la media y vecinos.

```
# Grafica de diagrama de dispersión de Moran
```

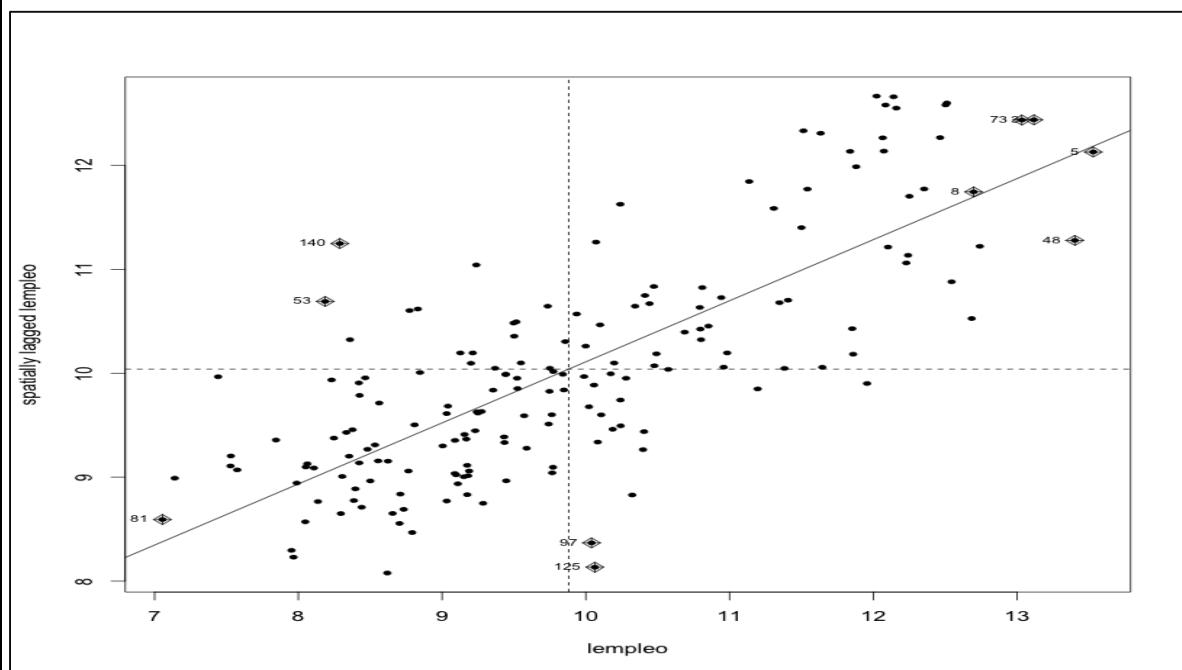
```
> moran.plot(empleo, wqueen, pch=20)  
> moran.plot(lch, wqueen, pch=20)
```

En la figura 14.2 se presenta el diagrama de dispersión para el logaritmo del empleo y su rezago espacial, donde se muestra una relación positiva, con coeficiente de correlación global de 0.59 (coeficiente de Moran), que indica la predominancia de municipios en el primer y tercer cuadrante y relativamente pocos en el segundo y cuarto cuadrante. En cada cuadrante se identifican aquellos municipios que ejercen una influencia inusual. Por ejemplo, en el primer cuadrante se identificaron a los municipios 2, 5, 8, 48 y 73 que se caracterizan por

observar los mayores empleos en relación de la media y tener vecinos también con alto empleo; en el segundo cuadrante, se identificaron a los municipios 53 y 140 por ser los de menor empleo con vecinos con alto empleo; en el tercer cuadrante destaca el municipio 81 por tener pocos empleos y tener vecinos de municipios que tampoco ofrecen empleos; y, finalmente en el cuarto cuadrante están los municipios 97 y 125 con empleo ligeramente por arriba de la media y municipios vecinos con las peores condiciones con respecto a la generación de empleo.

Figura 14.2: Diagrama de dispersión entre el logaritmo del empleo y su rezago espacial

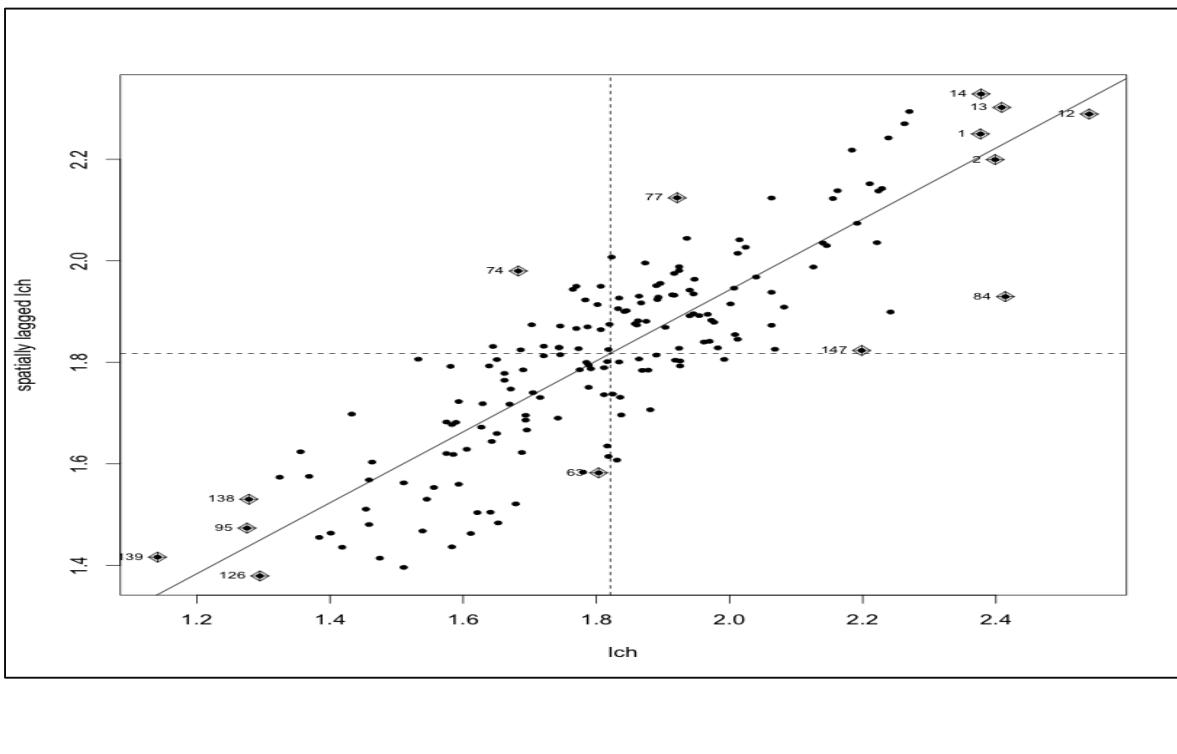
en la zona centro de México



Para el caso del logaritmo del capital humano y su rezago espacial, la relación positiva con un coeficiente de correlación global de Moran de 0.70. En este caso, se tienen nueve municipios en el primer cuadrante con comportamientos con influencia inusual; uno municipio en el segundo cuadrante; dos municipios en el tercer cuadrante; y, ninguno en el cuarto cuadrante.

Figura 14.3: Diagrama de dispersión entre el logaritmo del capital humano y su rezago espacial

en la zona centro de México



### *Indicador Local de Asociación Espacial (LISA)*

En procesos en los cuales existen patrones de agrupación local o **clúster**, el índice de Moran no los puede detectar, dado que sólo evalúa la dependencia global de todas las regiones. Como alternativa se han propuesto estadísticos locales, tal es el caso del índice local de Moran que se calcula en cada región o localidad y su definición es la siguiente:

$$I_i = \frac{z_i}{\sum_i z_i^2 / N_j} \sum_i w_{ij} z_j$$

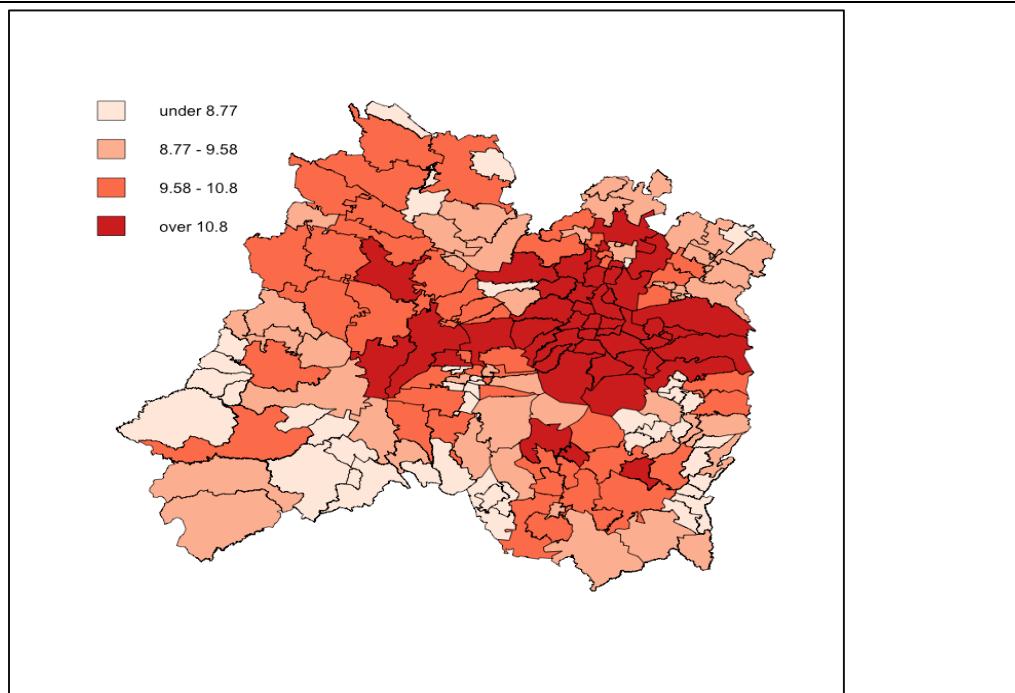
donde  $z_i$  es el valor de la variable correspondiente en la región  $i$ ,  $N_j$  es el conjunto de regiones vecinas a  $i$ . Un valor elevado, positivo (negativo) y significativo del estadístico da lugar a la existencia de un clúster alrededor de la región  $i$  de valores similares elevados (bajos). Con base en el índice local,  $I_i$ , es posible encontrar su contribución al índice global,  $I$ , y detectar sus valores extremos lo cual lo convierte en un LISA.

#### Ejemplo 4. Análisis de correlación espacial local (LISA) en la zona centro de México

Este ejemplo se aplica el análisis LISA, pero antes se construyen mapas temáticos con la distribución de los municipios por cuartiles de empleo y de capital humano. Para elaborar estos mapas en R, primero se tiene que utilizar un método para estratificar y otro para asignarle tonos de colores. Para la estratificación se aplica el método de cuartiles y para la asignación de color se define una escalar de tonos de colores rojos para el empleo y azules para el capital humano.

```
> # Mapa de quintiles del logaritmo del empleo  
  
> brks <- round(quantile(lempleo, probs=seq(0,1,0.25)), digits=2)  
  
> colours <- brewer.pal(4,"Reds")  
  
>  
  
> plot(empleo, col=colours[findInterval(empleo, brks, all.inside=TRUE)],  
+      axes=F)  
  
> legend(x=-87.9, y=25.2, legend=leglabs(brks), fill=colours, bty="n")  
  
> invisible(title(main=paste("EMPLEO", sep="\n")))  
  
> box()
```

Figura 4: Distribución del empleo en la zona centro de México



Nota: Estratos de cuartiles del logaritmo del empleo

Del mapa anterior se desprende que existe una gran heterogeneidad en la distribución del empleo en los municipios de la zona centro del país. En primer lugar, existe una fuerte asociación espacial del empleo entre municipios en niveles de empleo alto y medios altos, las cuales fundamentalmente forman manchas en el centro-norte del Distrito Federal y otras dos en el noreste y noroeste de la zona centro. En segundo lugar, se observa que las regiones de concentración del empleo bajo y medio-bajos se agrupan formando una mancha que se distribuye fundamentalmente hacia el estado de Guerrero y en municipios del Estado de México alrededor del D.F.

```
> # Mapa de quintiles del logaritmo del capital humano
> brks <- round(quantile(lch, probs=seq(0,1,0.25)), digits=2)
> colours <- brewer.pal(4,"Blues")
>
```

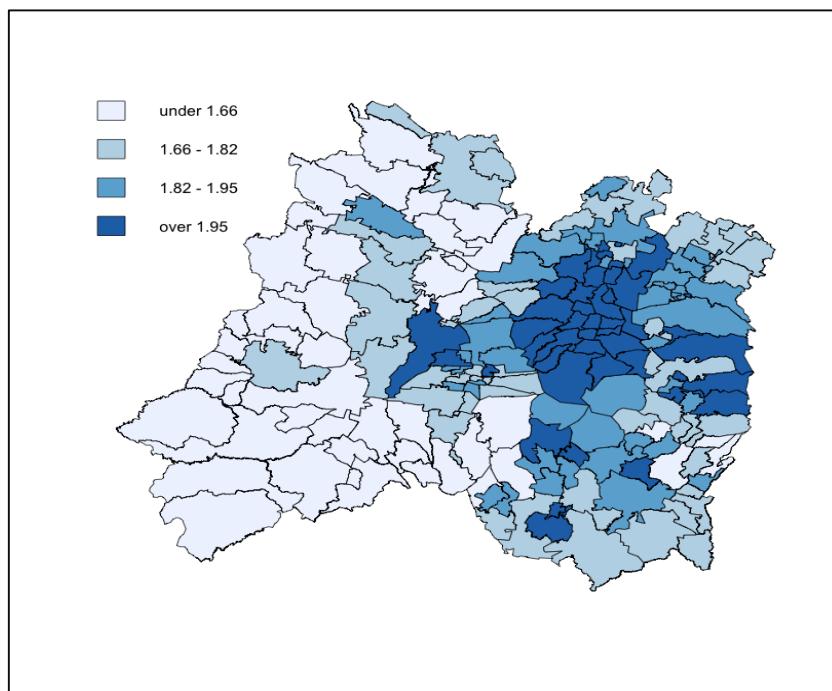
```

> plot(empleo, col=colours[findInterval(lch, brks, all.inside=TRUE)],
+      axes=F)
> legend(x=-87.9, y=25.2, legend=leglabs(brks), fill=colours, bty="n")
> #invisible(title(main=paste("CAPITAL HUMANO", sep="\n")))
> #box()

```

El mismo procedimiento se aplica al logaritmo del capital humano, los resultados se muestran en la siguiente figura.

**Figura 5: Distribución del capital humano en la zona centro de México**



Nota: Estratos de quintiles del logaritmo del capital humano

En este mapa se observa que la dependencia espacial es notoriamente más elevada que la visualizada antes para el empleo de los municipios de la zona centro. Los manchones en azul más oscuros dan cuenta de una fuerte asociación espacial entre los municipios con población con mayores años de estudio, lo mismo sucede con las manchas más claras que indican asociación entre los municipios del Estado de México cercanos a Guerrero.

Para evaluar estadísticamente la asociación espacial detectada en los mapas con estratos de la variable del logaritmo del empleo se aplica el análisis LISA.

```
# Valores de referencia z de la distribución t
```

```
> z <- c(1.65, 1.96)
```

```
> zc <- c(2.8284, 3.0471)
```

```
>
```

```
# Estimación de índice de Moran local (li)
```

```
> f.li <- localmoran(lempleo, wqueen)
```

```
> zli <- f.li[, "Z.li"] # Asignación de la distribución Z del li
```

```
> mx <- max(zli)
```

```
> mn <- min(zli)
```

```
# Mapa de significancia para los z-scores
```

```
> pal <- c("white", "green", "darkgreen")
```

```
> z3.li <- classIntervals(zli, n=3, style="fixed", fixedBreaks=c(mn, z, mx))
```

```
> cols.li <- findColours(z3.li, pal)
```

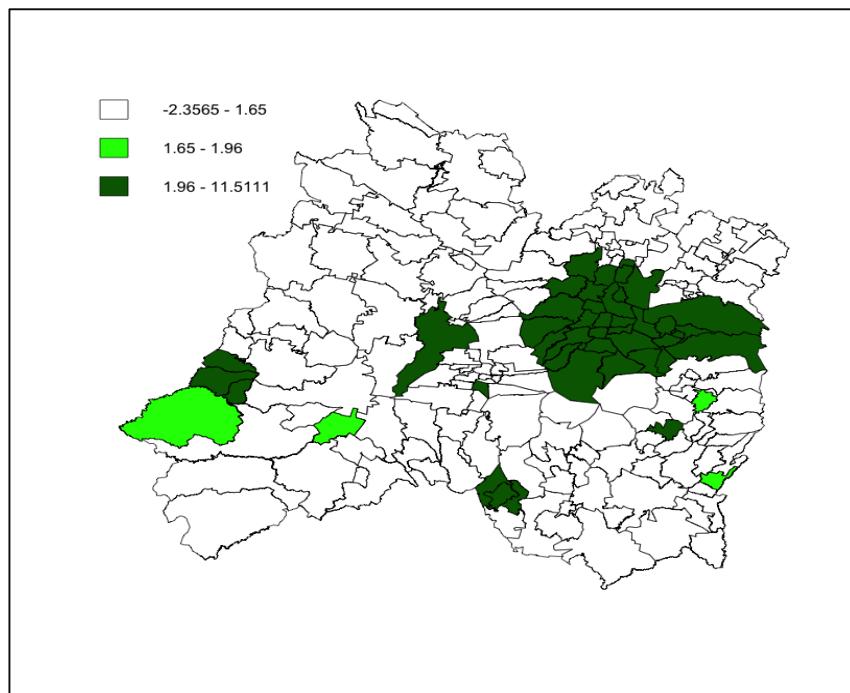
```
> plot(empleo, col=cols.li)
```

```
> brks <- round(z3.li$brks,4)
```

```
> leg <- paste(brks[-4], brks[-1], sep=" - ")
```

```
> legend(x=-87.9, y=25.2, fill=pal, legend=leg, bty="n")
```

**Figura 6: Mapa LISA con significancia para Z<sub>Li</sub> del logaritmo del empleo**



El mapa LISA de significancia anterior muestra las regiones que contribuyen al índice global de Moran y que conforman entre sí *clúster* significativos de dependencia espacial del logaritmo del empleo. Al combinar este resultado con los del mapa 14.4 de distribución de los cuartiles del logaritmo del empleo, se pueden identificar el clúster de (High-High) conformado por municipios del D.F. y el estado de México al noreste de la región y por el municipio de Toluca; y, el clúster (Low-Low) que se concentra en municipios del Estado de México frontera con el estado de Guerrero.

## 4. MODELOS ESPACIALES

Confirmada la dependencia espacial de los datos, es necesario especificar un modelo de regresión espacial que tome en cuenta dicha dependencia. Para plantear una especificación general prototipo, se combinaron las estrategias de Anselin (1988), Lesage y Pace (2009) y Ehorst (2010) para datos de corte transversal como los que hemos analizado en el caso 2.

El modelo general planteado es:

$$y_i = \rho W_1 y_i + \beta X_i + \theta W_2 X_i + \varepsilon_i$$

$$\varepsilon_i = \lambda W_3 \varepsilon_i + u_i$$

con  $u_i \sim N(0, \Omega)$  siendo los elementos diagonales de  $\Omega_{ij} = h_i(z\alpha)$  con  $h_i > 0$ .

donde  $y_i$  es el vector de la variable endógena,  $X_i$  es una matriz de variables exógenas y el término de error  $\varepsilon_i$  que incorpora una estructura de dependencia espacial autorregresiva,  $W_1$ ,  $W_2$  y  $W_3$  son matrices de pesos espaciales.

A partir de esta especificación podemos tener cinco casos:

- 1) Modelo de regresión clásico sin efectos espaciales:  $\rho = 0, \lambda = 0, \theta = 0$

$$y_i = \beta X_i + \varepsilon_i$$

$$\varepsilon_i = u_i$$

2) Modelo autorregresivo:  $\rho \neq 0, \lambda = 0, \theta = 0$

$$y_i = \rho W_1 y_i + \beta X_i + \varepsilon_i$$

$$\varepsilon_i = u_i$$

3) Modelo de error espacial autorregresivo:  $\rho = 0, \lambda \neq 0, \theta = 0$

$$y_i = \beta X_i + \varepsilon_i$$

$$\varepsilon_i = \lambda W_3 \varepsilon_i + u_i$$

que se puede reescribir en su forma final como

$$y_i = \beta X_i + (I - \lambda W_3)^{-1} u_i$$

4) Modelo Durbin Espacial:  $\rho \neq 0, \lambda = 0, \theta \neq 0$

La estrategia de Durbin sobre el factor común se aplica al modelo de Rezago Espacial, como:

$$y_i = \rho W_1 y_i + \beta X_i + \theta W_1 X_i + u_i$$

- 5) Modelo mixto autorregresivo espacial con errores espaciales autorregresivos (SARMA):  $\rho \neq 0, \lambda \neq 0, \theta = 0$

$$y_i = \rho W_1 y_i + \beta X_i + (I - \lambda W_3)^{-1} u_i$$

- 6) Modelo Error Durbin Espacial:  $\rho = 0, \lambda \neq 0, \theta \neq 0$

La estrategia de Durbin sobre el factor común se aplica al modelo de Error Espacial, con los siguientes pasos:

- a) De la primera ecuación despejar los errores y sustituir en la segunda

$$y_i - \beta X_i = \lambda W_3(y_i - \beta X_i) + u_i$$

- b) Al despejar  $y_i$ , se obtiene:

$$y_i = \lambda W_3 y_i + \beta X_i + \theta W_3 X_i + u_i$$

donde  $\theta = -\beta \lambda$

#### 14.4.1. *Métodos de Estimación*

Al igual que en el modelo de regresión clásico, la presencia de autocorrelación espacial dará lugar a que los estimadores de mínimos cuadrados ordinarios sean

insesgados, pero ineficientes, por lo cual no se cumple el teorema de Gauss-Markov. En los casos 2, 4, 5 y 6 la especificación considera rezagos autorregresivos de la variable dependiente, en consecuencia los estimadores de mínimos cuadrados ordinarios serán sesgados e inconsistentes. La estimación del modelo espacial se realiza a través del método de máxima verosimilitud en concordancia con el modelo espacial específico que se seleccione.

De acuerdo a Lesage y Pace (2009) la estrategia de estimación de los modelos Durbin Espacial (SDM) y Rezago Espacial (SAR) por sus siglas en inglés, es la siguiente:

#### El modelo SDM

$$y = \rho W y + \alpha i_n + X\beta + WX\theta + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

donde 0 representa un vector de ceros de  $n \times 1$  y  $i_n$  un vector de unos  $n \times 1$  asociados con el término de la constante  $\alpha$ . Este modelo puede ser escrito de forma compacta con  $Z = [i_n \ X \ WX]$  y  $\delta = [\alpha \ \beta \ \theta]'$  y entonces definir el caso del modelo SAR cuando  $Z = [i_n \ X]$  y  $\delta = [\alpha \ \beta]'$

#### El modelo SAR

$$y = \rho W y + Z\delta + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

Si el valor del parámetro rho ( $\rho$ ) fuera conocido por decir  $\rho^*$ , el modelo se puede escribir como

$$y - \rho^* W y = Z\delta + \varepsilon$$

Por lo que se puede resolver el problema de estimación de  $\delta$  como

$$(I_n - \rho^* W)y = Z\delta + \varepsilon$$

$$\hat{\delta} = (Z'Z)^{-1}Z'(I_n - \rho^* W)y$$

También se encuentra la estimación de la varianza

$$\hat{\sigma}^2 = n^{-1}e(\rho^*)'e(\rho^*) \text{ donde } e(\rho^*) = y - \rho^* W y - Z\hat{\delta}$$

donde  $e$  son los errores de estimación.

Lo anterior indica que el método de estimación se concentra en el log de verosimilitud con respecto a los parámetros de  $\beta$  y  $\sigma^2$  y por tanto la maximización de la verosimilitud se convierte a un problema de optimización univariante en el parámetro  $\rho$ .

Propuesta para estimar al mismo tiempo todo:

1. Estimar la función de log-verosimilitud concentrada con respecto a los parámetros  $\beta$  y  $\sigma^2$ , para obtener soluciones muy cercanas a las condiciones de primer orden junto con rho.

2. Sustituir las estimaciones de  $\beta$  y  $\sigma^2$ , por lo que la función de log-verosimilitud depende de la muestra de datos y el parámetro desconocido rho.
3. En este punto la función de log-verosimilitud esta concentrada con respecto rho, por lo que se usa para encontrar la estimación de máxima verosimilitud  $\hat{\rho}$  que será usada a su vez en la estimación de  $\hat{\beta}(\hat{\rho})$  y  $\hat{\sigma}^2(\hat{\rho})$  en la siguiente vuelta.

La función de verosimilitud para SDM y SAR toma la forma siguiente

$$\ln L = -\left(\frac{n}{2}\right) \ln(\pi\sigma^2) + \ln|I_n - \rho W| - \frac{e'e}{2\sigma^2}$$

$$e = y - \rho Wy - Z\delta$$

$$\rho \in (\min(\omega)^{-1}, \max(\omega)^{-1})$$

donde  $\omega$  es el vector de  $n \times 1$  raíces características de la matriz  $W$ . Dado que la matriz siempre esta construida para tener raíces máximas de 1, entonces  $\rho \in (\min(\omega)^{-1}, 1)$  el cual es un subconjunto del empleado en la práctica  $\rho \in [0,1]$ .

La función de log-verosimilitud concentrada en los valores de  $\ln L(\rho)$  se escribe como

$$\ln L(\rho) = \kappa + \ln|I_n - \rho W| - (n/2)\ln(S(\rho))$$

$$S(\rho) = e(\rho)'e(\rho) = e_0'e_0 - 2\rho e_0'e_d + \rho^2 e_d'e_d$$

$$e(\rho) = e_0 - \rho e_d$$

$$e_0 = y - Z\delta_0$$

$$e_d = Wy - Z\delta_d$$

$$\delta_0 = (Z'Z)^{-1}Z'y$$

$$\delta_d = (Z'Z)^{-1}Z'Wy$$

La optimización es con respecto al parámetro rho y una vez estimado  $\hat{\rho}$  con máxima verosimilitud se llega a la estimación con máxima verosimilitud de  $\hat{\delta}$  y  $\hat{\sigma}^2$

$$\hat{\delta} = \delta_0 - \hat{\rho}\delta_d$$

$$\hat{\sigma}^2 = n^{-1}S(\hat{\rho})$$

$$\hat{\Omega} = \hat{\sigma}^2[(I_n - \hat{\rho}W)'(I_n - \hat{\rho}W)]^{-1}$$

### Estimación del modelo de Error Espacial (SEM)

Con una estrategia parecida, se obtiene la solución para SEM

$$y = X\beta + u$$

$$u = \lambda W u + \varepsilon$$

$$\varepsilon \sim N(0, \sigma^2 I_n)$$

$$\ln L = -\left(\frac{n}{2}\right) \ln(\pi\sigma^2) + \ln|I_n - \lambda W| - \frac{e'e}{2\sigma^2}$$

$$e = (I_n - \lambda W)(y - X\beta)$$

Para un valor dado de  $\lambda$ ,

$$\beta(\lambda) = (X(\lambda)'X(\lambda))^{-1}X(\lambda)'y(\lambda), \text{ donde}$$

$$X(\lambda) = (X - \lambda WX)$$

$$y(\lambda) = (y - \lambda Wy)$$

$$\sigma^2(\lambda) = e(\lambda)'e(\lambda)/n$$

$$e(\lambda) = y(\lambda) - X(\lambda)\beta(\lambda)$$

La función de log-verosimilitud concentrada en los parámetros  $\beta$  y  $\sigma^2$

$$\ln L(\lambda) = \kappa + \ln|I_n - \lambda W| - (n/2)\ln(S(\lambda))$$

$S(\lambda) = e(\lambda)'e(\lambda)$  no es cuadrático, se necesita todo un proceso simultáneo

$$\hat{\beta} = \beta(\hat{\lambda})$$

$$\hat{\sigma}^2 = n^{-1}S(\hat{\lambda})$$

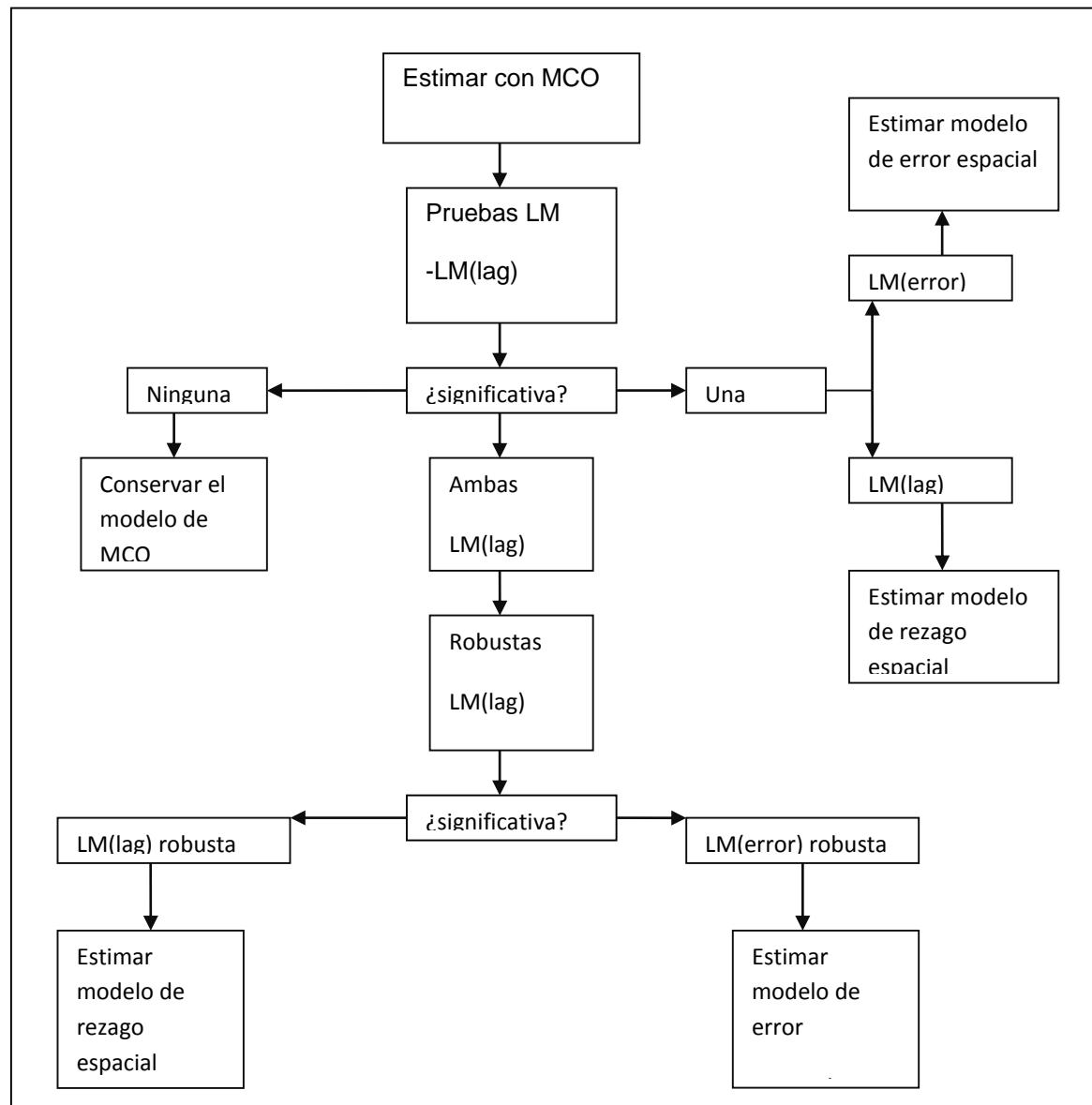
$$\hat{\Omega} = \hat{\sigma}^2 \left[ (I_n - \hat{\lambda}W)'(I_n - \hat{\lambda}W) \right]^{-1}$$

### **Estrategia de Selección de modelos: de lo particular a lo general**

Anselin (2005) propone seguir un proceso de decisión para seleccionar entre el modelo clásico y los modelos espaciales SAR, SEM y SARMA, utilizando la

estrategia que se muestra en la figura 10 y los estadísticos de contraste para las prueba de hipótesis de los tipos de dependencia espacial.

Figura 10: Estrategia de Selección de modelos: de lo particular a lo general



Fuente: Anselin, Luc (2005) Exploring Spatial Data with OpenGeoDa: A Workbook, consultado en: <http://sal.uiuc.edu/>

### *Contrastes de autocorrelación espacial*

Estos contrastes se aplican después de estimar el modelo clásico para analizar la presencia de algún tipo de dependencia espacial. La hipótesis nula es que el tipo de dependencia espacial es igual a cero, contra la hipótesis alternativa de que es diferente de cero.

#### 1. *Test I de Moran*

Mide el efecto de autocorrelación espacial en los residuos  $e_i$  en un modelo no-espacial o clásico, sin distinguir estructuras de Rezago o Error Espacial:

$$I = \frac{N}{S_0} \frac{\sum_{(2)} w_{ij} e_i e_j}{\sum_{i=1}^N e_i^2} = \frac{N}{S_0} \frac{e' W e}{e' e}$$

La inferencia se hace con el valor z estandarizado. El primer y segundo momento

$$E[I] = \frac{N}{S_0} \frac{\text{tr}(MW)}{N - K}$$
$$E[I]^2 = \frac{\left(\frac{N}{S_0}\right)^2 \text{tr}(MWMW') + \text{tr}(MW)^2 + [\text{tr}(MW)]^2}{(N - K)(N - K + 2)}$$

Se distribuye como una  $\chi^2$  con un grado de libertad

#### 2. *Test LM-ERR: Error espacial*

Se basa en el principio de los multiplicadores de Lagrange y fue propuesto por Burridge (1980):

$$LM - ERR = \frac{\left[ \frac{e' We}{s^2} \right]^2}{tr[W' W + W^2]}$$

Se distribuye como una  $\chi^2$  con un grado de libertad

### 3. Test LM-EL: Error espacial (robusto)

El estadístico  $LM-ERR$  se ajusta por una mala especificación local de la dependencia espacial, como es el caso de una variable endógena rezagada (Anselin, 1996):

$$LM - EL = \frac{\left[ \frac{e' We}{s^2} - T_1(R\tilde{J}_{\rho-\beta})^{-1} \frac{e' Wy}{s^2} \right]^2}{T_1 - T_1^2(R\tilde{J}_{\rho-\beta})}$$

$$\text{con: } (R\tilde{J}_{\rho-\beta})^{-1} = \left[ T_1 + \frac{(WX\beta)' M (WX\beta)}{s^2} \right]^{-1}$$

$$T_1 = tr(W' W + W^2)$$

Se distribuye como una  $\chi^2$  con un grado de libertad

### 4. Test LM-LAG: Rezago Espacial

Por rezago espaciales de la variable endógena (Anselin, 1988):

$$LM - LAG = \frac{\left[ \frac{e'Wy}{s^2} \right]^2}{(R\tilde{J}_{\rho-\beta})}$$

Se distribuye como una  $\chi^2$  con un grado de libertad

##### 5. Test LM-LE: Rezago Espacial (Robusto)

El estadístico es robusto ante la presencia de dependencia local del error espacial (Anselin, 1988):

$$LM - LE = \frac{\left[ \frac{e'W_1y}{s^2} - \frac{e'W_1e}{s^2} \right]^2}{(R\tilde{J}_{\rho-\beta} - T_1)}$$

Se distribuye como una  $\chi^2$  con un grado de libertad

##### 6. Test SARMA: Rezago y Error Espacial

Es robusto ante la presencia de dependencia local y del error espacial (Anselin, 1988):

$$SARMA = \frac{\left[ \frac{e'Wy}{s^2} - \frac{e'We}{s^2} \right]^2}{(R\tilde{J}_{\rho-\beta} - T_1)} + \frac{\left[ \frac{e'We}{s^2} \right]^2}{T_1}$$

Se distribuye como una  $\chi^2$  con dos grados de libertad.

**Ejemplo 5. Modelos espaciales para el empleo determinado por capital humano**

La teoría neoclásica del empleo por capital humano, aplicada a lo local, permite plantear que las localidades con mayor capital humano obtendrán mayor cantidad de empleo.

$$\ln(\text{emp}_i) = \alpha + \beta \ln(\text{ch}_i) + u_i$$

donde  $\ln(\text{emp}_i)$  es el logaritmo natural del empleo y  $\ln(\text{ch}_i)$  es el logaritmo del capital humano (años de escolaridad) de la región  $i$ . Un coeficiente positivo para la beta estimado será evidencia a favor de la determinación del empleo por capital humano. Como las variables están en logaritmos, el parámetro beta mide la elasticidad.

Para estimar este modelo se utiliza primero una estimación OLS

```
> ModeloEmpleo_OLS <- lm(lempleo ~ lch , data=empleo)  
> summary(ModeloEmpleo_OLS)
```

Call:

```
lm(formula = lempleo ~ lch, data = empleo)
```

Residuals:

Min	1Q	Median	3Q	Max
-4.3948	-0.8243	0.0286	0.7611	2.8910

Coefficients:

Estimate	Std. Error	t value	Pr(> t )
----------	------------	---------	----------

```
(Intercept) 3.8775 0.6763 5.734 4.33e-08 ***
Ich         3.2969 0.3680 8.960 5.26e-16 ***
---
Signif. codes: 0 '****' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.2 on 172 degrees of freedom
Multiple R-squared: 0.3182,    Adjusted R-squared: 0.3142
F-statistic: 80.27 on 1 and 172 DF, p-value: 5.264e-16
```

Dado el valor pequeño del p-valor de beta podemos concluir que es estadísticamente significativa y dado que la elasticidad es de 3.29, se concluye que el empleo es altamente sensible al capital humano en la región.

Las pruebas de diagnóstico al modelo se muestran a continuación, donde los diferentes estadísticos de prueba contrastan la hipótesis nula de no autocorrelación espacial o de proceso aleatorio.

```
> #Prueba de Moran a residuales del modelo OLS
> I_Moran <- lm.morantest(ModeloEmpleo_OLS,wqueen)
> print(I_Moran)
```

Global Moran's I for regression residuals

```
data:
model: lm(formula = lempleo ~ Ich, data = empleo)
weights: wqueen
```

Moran I statistic standard deviate = 8.0244, p-value = 5.1e-16

alternative hypothesis: greater

sample estimates:

Observed Moran's I	Expectation	Variance
0.368395591	-0.009877828	0.002222204

# Pruebas de Multiplicadores de Lagranges

```
#  
lm.LMtests(columbus.lm,col.listw,test=c("LMerr","RLMerr","LMIlag","RLMIlag","SAR  
MA"))  
  
>  
lm.LMtests(ModeloEmpleo_OLS,wqueen,test=c("LMerr","RLMerr","LMIlag","RLMIla  
g","SARMA"))
```

Lagrange multiplier diagnostics for spatial dependence

data:

model: lm(formula = lempleo ~ lch, data = empleo)

weights: wqueen

LMerr = 58.244, df = 1, p-value = 2.32e-14

Lagrange multiplier diagnostics for spatial dependence

data:

```
model: lm(formula = lempelo ~ lch, data = empleo)
```

```
weights: wqueen
```

```
RLMerr = 0.0032553, df = 1, p-value = 0.9545
```

Lagrange multiplier diagnostics for spatial dependence

```
data:
```

```
model: lm(formula = lempelo ~ lch, data = empleo)
```

```
weights: wqueen
```

```
LMIlag = 70.722, df = 1, p-value < 2.2e-16
```

Lagrange multiplier diagnostics for spatial dependence

```
data:
```

```
model: lm(formula = lempelo ~ lch, data = empleo)
```

```
weights: wqueen
```

```
RLMlag = 12.482, df = 1, p-value = 0.0004109
```

Lagrange multiplier diagnostics for spatial dependence

```
data:  
model: lm(formula = lempleo ~ lch, data = empleo)  
weights: wqueen  
  
SARMA = 70.726, df = 2, p-value = 4.441e-16
```

En los resultados, el índice de Moran presenta un p-valor muy pequeño (2.32e-14) lo cual permite rechazar la hipótesis nula de no autocorrelación espacial. El LM-lag y el LM-lag robusto presentan la hipótesis alternativa específica de modelo de rezago espacial, mientras que el LM-error establecen como hipótesis alternativa al modelo de error espacial, pero de acuerdo con el y LM-error robusto no se puede aceptar la hipótesis considerando que el parámetro rho de rezago espacial. De acuerdo con el p-valor del estadístico SARMA existe la posibilidad de considerar un modelo con rezago y error espacial.

Con la información anterior, se estiman los modelos espaciales de Rezago, Error, SARAR y Durbin.

El primer modelo que se estimó es el de modelo de rezago espacial. Los resultados del modelo muestran los siguientes resultados: el parámetro rho de rezago espacial es 0.67, positivo y de acuerdo al p-value (6.6613e-16) estadísticamente significativo, lo cual implica que el rezago espacial del logaritmo del empleo es importante. El parámetro de beta también es significativo pero la elasticidad es mucho menor (1.32) en comparación del 3.3 del modelo OLS, que es consistente cuando se incorpora la dinámica espacial al modelo.

```
# Estimar el Modelo Rezago Espacial  
> ModeloEmpleo_lag <- lagsarlm(lempleo ~ lch , data=empleo,wqueen)  
> summary(ModeloEmpleo_lag)
```

```
Call:lagsarlm(formula = lempleo ~ lch, data = empleo, listw = wqueen)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.1783841	-0.5655122	-0.0045741	0.5698724	2.4373878

Type: lag

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.73366	0.61410	1.1947	0.2322
lch	1.32893	0.32970	4.0307	5.562e-05

Rho: 0.67009, LR test value: 65.113, p-value: 6.6613e-16

Asymptotic standard error: 0.06377

z-value: 10.508, p-value: < 2.22e-16

Wald statistic: 110.42, p-value: < 2.22e-16

Log likelihood: -245.0376 for lag model

ML residual variance (sigma squared): 0.87624, (sigma: 0.93607)

Number of observations: 174

Number of parameters estimated: 4

AIC: 498.08, (AIC for lm: 561.19)

LM test for residual autocorrelation

test value: 0.11791, p-value: 0.73132

La segunda posibilidad consiste en estimar un modelo de Error Espacial. Para este modelo el parámetro lambda de error espacial es 0.71, positivo y de acuerdo al p-value (1.9318e-14) es estadísticamente significativo, lo cual implica que el error espacial del logaritmo del empleo es importante. El parámetro de beta es significativo y la elasticidad ligeramente mayor al de modelo de rezago espacial.

```
# Estimar el modelo de Error Espacial
```

```
> ModeloEmpleo_err <- errorsarlm(lempleo ~ Ich , data=empleo,wqueen)  
> summary(ModeloEmpleo_err)
```

```
Call:errorsarlm(formula = lempelo ~ Ich, data = empleo, listw = wqueen)
```

Residuals:

Min	1Q	Median	3Q	Max
-3.229495	-0.603003	0.050277	0.613023	2.550370

Type: error

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	6.47006	0.94208	6.8679	6.516e-12
Ich	1.64366	0.49579	3.3152	0.0009157

Lambda: 0.71817, LR test value: 58.604, p-value: 1.9318e-14

Asymptotic standard error: 0.061095

z-value: 11.755, p-value: < 2.22e-16

Wald statistic: 138.18, p-value: < 2.22e-16

Log likelihood: -248.2923 for error model

ML residual variance (sigma squared): 0.89031, (sigma: 0.94356)

Number of observations: 174

Number of parameters estimated: 4

AIC: 504.58, (AIC for lm: 561.19)

La tercera opción consiste en estimar el modelo que incorpora tanto rezago como error espacial. De acuerdo a que rho es 0.70 y significativo (p-value: 3.6575e-09), lambda -0.07 pero no significativo y el parámetro beta significativo, se puede concluir que al combinar los procesos de rezago y el error espacial, predomina el primero por lo que la mejor opción es el modelo de rezago espacial. Esta conclusión ya se había identificado de la aplicación de la prueba LM robusta de error espacial.

#### Estimar modelo SARAR

```
> ModeloEmpleo_sarar <- sacsarlm(lempleo ~ lch , data=empleo, wqueen,  
type="sac")  
  
> summary(ModeloEmpleo_sarar)
```

Call:sacsarlm(formula = lempiego ~ lch, data = empleo, listw = wqueen,  
type = "sac")

Residuals:

Min	1Q	Median	3Q	Max
-3.142804	-0.581190	-0.016665	0.575159	2.415785

Type: sac

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.56642	0.74371	0.7616	0.446294
Ich	1.24960	0.41976	2.9770	0.002911

Rho: 0.70235

Asymptotic standard error: 0.11906

z-value: 5.899, p-value: 3.6575e-09

Lambda: -0.073671

Asymptotic standard error: 0.24171

z-value: -0.30479, p-value: 0.76053

LR test value: 65.214, p-value: 6.8834e-15

Log likelihood: -244.9869 for sac model

ML residual variance (sigma squared): 0.86265, (sigma: 0.92879)

Number of observations: 174

Number of parameters estimated: 5

AIC: 499.97, (AIC for lm: 561.19)

Otra alternativa es modelo durbin de rezago espacial, que considera el modelo de rezago espacial y le incorpora el rezago espacial de la variable de logaritmo del capital humano. Los resultados muestran que esta alternativa no aporta mayor información al modelo de rezago espacial, debido a que el parámetro del rezago espacial del logaritmo del capital humano (lag.lch) no es significativo (p-value: 0.48242).

#Estimar el modelo de Durbin Rezago Espacial

```
> ModeloEmpleo_lag_durbin <- lagsarlm(lempleo ~ lch , data=empleo,wqueen,
type="mixed")  
> summary(ModeloEmpleo_lag_durbin)
```

Call:lagsarlm(formula = lempleo ~ lch, data = empleo, listw = wqueen,  
type = "mixed")

Residuals:

Min	1Q	Median	3Q	Max
-3.059210	-0.599797	-0.023718	0.585009	2.417886

Type: mixed

Coefficients: (asymptotic standard errors)

	Estimate	Std. Error	z value	Pr(> z )
(Intercept)	0.54567	0.67171	0.8124	0.41659
lch	1.03039	0.55554	1.8547	0.06363
lag.lch	0.50523	0.71928	0.7024	0.48242

Rho: 0.6515, LR test value: 52.636, p-value: 4.0157e-13

Asymptotic standard error: 0.06901

z-value: 9.4407, p-value: < 2.22e-16

Wald statistic: 89.127, p-value: < 2.22e-16

Log likelihood: -244.8002 for mixed model

ML residual variance (sigma squared): 0.88042, (sigma: 0.93831)

Number of observations: 174

Number of parameters estimated: 5

AIC: 499.6, (AIC for lm: 550.24)

LM test for residual autocorrelation

test value: 0.024694, p-value: 0.87513

## REFERENCIAS

Anselin, L. (1988) Spatial Econometrics: Methods and Models. Kluwer Academic Publishers, Dordrecht, The Netherlands.

Anselin, L. (2012) OpenGeoDa 1.2 User's Guide. Spatial Analysis Laboratory (SAL). Department of Agricultural and Consumer Economics, University of Illinois, Urbana-Champaign, IL.

Fotheringham, Brunsdon y Charlton (2000) Quantitative Geography: Perspectives on Spatial Data Analisys.

Haining, Robert (2003)patial Data Analysis: Theory and Practice, 1st edition, Cambridge University Press

LeSage, J. y Pace, K. (2009) Introduction of Spatial Econometrics, Taylor & Francis Group, LLC.

### **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

Zona\_Centro.dbf

Zona\_Centro.shp

Zona\_Centro.shx

### **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teórica\_Cap15

Práctica\_Cap15

VideoPráctica\_Cap15

VideoTeoría\_Cap15

# **CAPÍTULO 16: REPASO BÁSICO DE ESTADÍSTICA Y ÁLGEBRA MATRICIAL**

**Luis Quintana Romero y Miguel Ángel Mendoza**

## **1. INTRODUCCIÓN**

Para facilitar la comprensión de los temas tratados en este curso, a continuación se abordan una serie de herramientas y conceptos básicos de estadística, probabilidad y álgebra lineal.

El objetivo es permitirle al alumno recordar los elementos que ya ha aprendido en ese campo a lo largo de los cursos básicos de matemáticas que ha tomado previamente. Además, de brindar la oportunidad de practicar un poco, antes de entrar de lleno al estudio de la econometría.

En caso de que así lo deseé, es posible profundizar los temas aquí tratados a través de la consulta de la amplia gama de textos de matemáticas para economistas disponibles hoy día en las librerías (Chiang, 1987, Kholer, 1997, Weber, 1999).

## **2. REVISIÓN DE LOS DATOS**

Antes de iniciar cualquier análisis es fundamental conocer los datos que se van a utilizar, por ello a continuación se muestran algunas formas para su descripción.

### a) Localización

La medida de localización más usual es la media aritmética de una muestra.

Para exemplificar esta medida tomamos los datos del cuadro 1 que corresponden al Producto Interno Bruto de los estados mexicanos (en este archivo las variables que están en las columnas son los estados) y que se encuentran en el archivo pib\_estados.txt. Considerando esta información, los PIB promedio entre 2003 y 2011 para el Distrito Federal y Chiapas son:

$$\text{PIB medio en el DF: } \bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{17,375,377}{9} = 1,930,597 \text{ millones de pesos de 2008}$$

$$\text{PIB medio en Chiapas: } \bar{X} = \frac{\sum_{i=1}^n X_i}{n} = \frac{1,850,724}{9} = 205,636 \text{ millones de pesos de 2008}$$

Donde:  $X_i$ = valores muestrales y  $n$ = número de observaciones muestrales

El PIB promedio de México en ese mismo período ha sido 11,380,499 millones de de 2008 tal y como se observa en el cuadro 1.

### Cuadro 1.

### **PIB en millones de pesos de 2008 para los estados mexicanos**

Año	Chiapas	DF	Nacional
2003	199,555.37	1,710,591.68	10,119,898.14
2004	194,308.50	1,782,074.82	10,545,909.78
2005	195,008.15	1,830,742.76	10,870,105.27
2006	202,567.16	1,933,232.06	11,410,946.02
2007	199,816.04	1,990,454.43	11,778,877.72
2008	207,208.94	2,029,146.99	11,941,199.48
2009	204,472.32	1,949,101.89	11,374,629.55
2010	220,648.93	2,033,881.57	11,960,871.07
2011	227,138.20	2,116,150.87	12,422,056.85
Promedio	205,635.96	1,930,597.45	11,380,499.32

Fuente: INEGI, PIB de los estados

Para calcular la media en R Commander seleccionamos en el menú principal las opciones; STATISTIC/Summaries/Numerical summary. En la ventana que se abre se seleccionan los estados(opción DATA) y el estadístico a calcular (opción Statistics) que en este caso es la media. Los resultados se muestran a continuación:

```
> numSummary(pib_estados2[,c("CHIS", "DF")], statistics=c("mean",
+ "quantiles"), quantiles=c(0,.25,.5,.75,1))

      mean      0%     25%     50%    75%   100% n
CHIS 205636 194308.5 199555.4 202567.2 207208.9 227138.2 9
DF   1930597 1710591.7 1830742.8 1949101.9 2029147.0 2116150.9 9
```

Otras dos medidas de centralización muy utilizadas son la moda y la mediana. La moda es el valor con la mayor frecuencia en los datos, mientras que la mediana es el valor medio de un conjunto ordenado de datos. En nuestro ejemplo, la moda no es única ya que el PIB de un año no se repite con el mismo valor en otro período. Por otra parte, para calcular la mediana podemos ordenar los datos de menor a mayor, por ejemplo para el 2011, y obtendremos que la mediana es el valor del PIB de 248,500 millones de pesos correspondiente al promedio de Querétaro y Sinaloa, pues justo antes y después de esos dos estados tenemos el 50% de los datos respectivamente. Es importante hacer notar que la mediana no se verá afectada por la presencia de valores extremos ya que sólo depende del orden de los valores, más no de sus magnitudes.

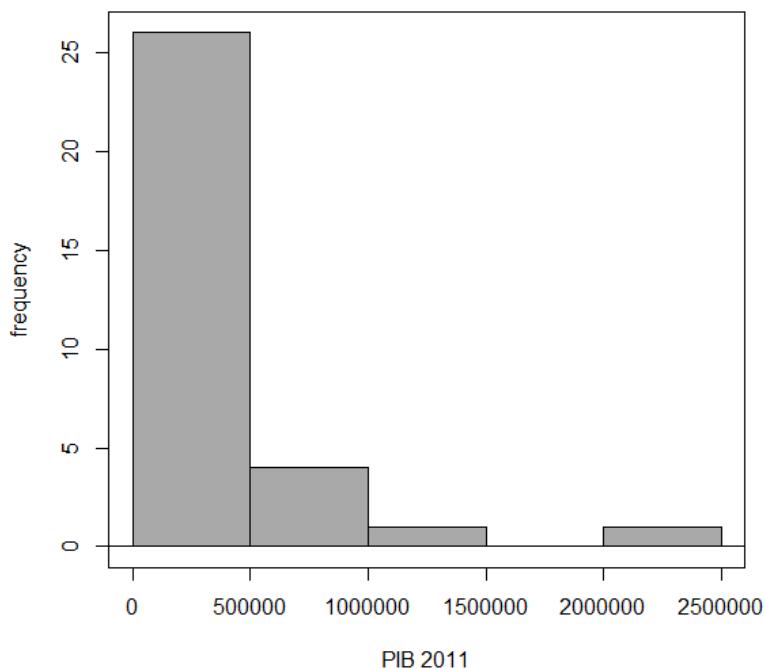
Para calcular la mediana utilizando RCommander se escribe la función `mean(pib_estados$PIB11)` en la ventana RScript, posteriormente se activa el botón Submit y en la ventana Output se obtiene el siguiente resultados:

```
> median(pib_estados$PIB11)  
[1] 248499
```

La distribución de los datos se puede visualizar utilizando histogramas, los cuales presentan los porcentajes de observaciones que quedan dentro de un intervalo en particular. Por ejemplo, en RCommander al seleccionar en el menú principal la opción GRAPHS/Histogram se abre una ventana en la cual dentro de la opción Data se elige PIB11, el resultado se muestra a continuación y en éste se observa que el grueso de los estados se encuentran en un intervalo de 0 a 500,000 millones de pesos y sólo una entidad federativa presenta un PIB por encima de 2,000,000 millones de pesos:

**Gráfica 1**

**Histograma del PIB estatal de 2011 en millones de pesos de 2003 (frecuencias)**

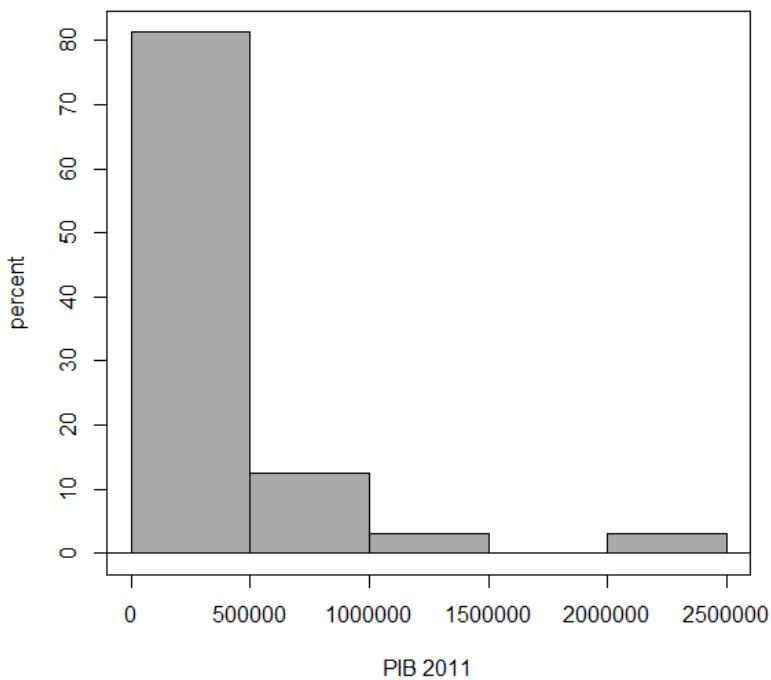


Fuente: Elaborado en RCommander con datos del INEGI

El histograma se puede visualizar también en porcentajes en lugar de frecuencias, para lo cual simplemente en OPTIONS se elige la opción *percentages*, en el histograma resultante ahora es claro que poco más del 80% de los estados mexicanos tienen un PIB que no rebasa los 500,000 millones:

**Gráfica 2**

**Histograma del PIB estatal de 2011 en millones de pesos de 2003 (porcentajes)**



Fuente: Elaborado en RCommander con datos del INEGI

En los dos histogramas la forma de la gráfica muestra un gran sesgo positivo con relación al valor medio del PIB hacia valores bajos del mismo, por tal razón es relevante conocer que tan dispersos se encuentran los datos respecto a su media. Una de las medidas de dispersión más frecuente es la varianza muestral ( $S^2$ ) y su raíz cuadrada ( $S$ ) es la desviación estándar. La varianza y la desviación estándar para el PIB de los estados en los años de 2003 y 2011 son las siguientes:

PIB 2003

$$S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n - 1} = \frac{3468376923222.16}{32 - 1} = 111,883,126,555.55 \text{ (mill. de pesos al cuadrado)}$$

$$S = \sqrt[2]{111,883,126,555.55} = 334489.35 \text{ (mill. de pesos)}$$

$$\text{PIB 2011} \quad S^2 = \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{n-1} = \frac{5039942063659.1}{32-1} = \\ 162,578,776,247.07 \text{ (mill. de pesos al cuadrado)}$$

$$S = \sqrt[2]{5039942063659.1} = 403210.59 \text{ (mill. de pesos)}$$

En los resultados precios es posible constatar que la varianza mide la dispersión pero eleva al cuadrado las unidades de medida originales, lo cual resulta difícil de interpretar; en nuestro ejemplo tenemos millones de pesos al cuadrado. Así que para interpretar ese valor, en las unidades de medida originales, es mejor considerar la desviación estándar.

Una alternativa de medición de la dispersión que tiene la virtud de no hacer referencia a la unidad de medida es el coeficiente de variación (CV), el cual se calcula dividiendo la desviación estándar entre la media. Por ejemplo, para el caso del PIB tendríamos el resultado siguiente:

$$\text{PIB 2003} \quad CV = \frac{S}{\bar{X}} \times 100 = 105.77$$

$$\text{PIB 2011} \quad CV = \frac{S}{\bar{X}} \times 100 = 103.87$$

El resultado del coeficiente de variación muestra que el PIB en 2003 varía más que el de 2011, pese a que este último presenta una varianza mayor.

Además de las medidas de centralización y dispersión ya revisadas, podemos obtener mediciones de la forma de las distribuciones de probabilidad de nuestros datos, las cuales nos aportan información sobre la asimetría de su distribución. Las medidas de forma con las que usualmente vamos a trabajar son el sesgo y la curtosis.

El sesgo (SK), mide la simetría de una distribución de frecuencias de nuestros datos con relación a la media, mientras que la curtosis (KS) mide el achatamiento o agudeza en relación con la forma que presenta una distribución normal de los datos.

$$\text{SESGO} \quad SK = \frac{1}{n} \sum_{i=1}^n \left[ \frac{X_i - \bar{X}}{S} \right]^3$$

$$\text{CURTOSIS} \quad KS = \frac{1}{n} \sum_{i=1}^n \left[ \frac{X_i - \bar{X}}{S} \right]^4$$

De esta forma el sesgo y la curtosis para los datos del ejemplo son:

PIB 2003    SK=2.76    KS=9.29

PIB 2011    SK=2.92    KS=10.60

Las distribuciones simétricas tienen sesgo igual a cero y si además son mesocúrticas tienen una curtosis igual a 3, tal y como ocurre con las distribuciones normales; curtosis superiores a 3 se consideran leptocúrticas e inferiores a 3 son platocúrticas. De acuerdo a los resultados de nuestro ejemplo, el valor positivo que tiene el indicador de sesgo indica que la cola derecha de la distribución de los

datos es larga y que gran parte de las observaciones se ubican al lado izquierdo de la media. Por su parte, las curtosis de nuestros datos son mucho mayores a 3, lo cual indica que son leptocurticas y por consiguiente presentan un pico más afilado que el de la normal lo cual es resultado de la elevada concentración de los datos en los valores bajos del PIB.

Para obtener los estadísticos previos en RCommander solo hay que entrar al menú STATISTICS y seleccionar Summaries/Numerical Summaries; en la ventana que se abre se debe elegir en Data el PIB para 2003 y 2011, y en Statistics seleccionar Standard Deviation , Coefficient of Variation, Skewness y Kurtosis. Los resultados son los siguientes:

```
> numSummary(pib_estados[,c("PIB03", "PIB11")], statistics=c("sd", "quantiles", "cv", "skewness", "kurtosis"), quantiles=c(0,.25,.5,.75,1), type="2")
```

	<b>sd</b>	<b>cv</b>	<b>skewness</b>	<b>kurtosis</b>	<b>n</b>
<b>PIB03</b>	334489.4	1.057685	2.756137	9.296279	32
<b>PIB11</b>	403210.6	1.038696	2.924704	10.601004	32

La distribución de los datos se puede suavizar si se supone que éstos provienen de una muestra aleatoria continua. Para ello se utilizan las funciones de densidad kernel cuyo estimador se define a continuación (Everitt y Horthorn, 2006):

$$\hat{f}(x) = \frac{1}{hn} \sum_{i=1}^n \frac{1}{h} K\left(\frac{x - x_i}{h}\right)$$

Donde K es una función kernel y h es la amplitud de banda o parámetro de suavizamiento.

La función kernel cumple con la condición:

$$\int_{-\infty}^{\infty} K(x)dx = 1$$

Las funciones kernel más usuales son las siguientes:

a) Rectangular

$$K(h) = \begin{cases} \frac{1}{2} & |x| < 1 \\ 0 & \text{en otro caso} \end{cases}$$

b) Triangular

$$K(h) = \begin{cases} 1 - |x| & |x| < 1 \\ 0 & \text{en otro caso} \end{cases}$$

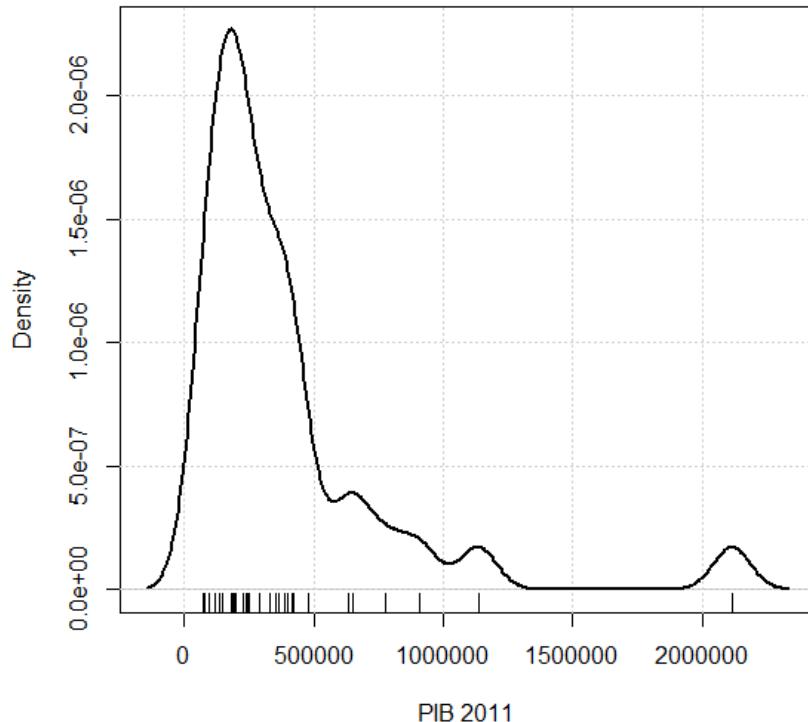
c) Gaussiano

$$K(h) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$$

En el menú principal de RCommander en la opción GRAPHS seleccionamos Density estimate y en la ventana que se abre se selecciona en DATA el PIB de 2011 y en OPTIONS la función kernel Gaussiana, el resultado se muestra en el gráfico siguiente.

Gráfica 3

**Función de densidad kernel del PIB estatal de 2011 en millones de pesos de 2003**



## 2.3 Probabilidad

La teoría de la probabilidad es fundamental para realizar inferencias de la población a partir de datos muestrales. En esta sección se establecerán algunas de las definiciones básicas.

La probabilidad del evento A está definida por:  $\Pr(A)$  y cumple con los siguientes axiomas definidos por el matemático ruso Kolmogorov:

La probabilidad para cualquier evento A es un número positivo entre cero y uno:  
 $0 \leq \Pr(A) \leq 1$ .

La probabilidad de que ocurra el evento seguro Z es la unidad:  $\Pr(Z) = 1$ .

La probabilidad de la unión de los eventos  $A_1, A_2, A_3, \dots$  es igual a la suma de las probabilidades de los eventos individuales siempre y cuando sean mutuamente excluyentes (no pueden ocurrir al mismo tiempo):

$$\Pr(\cup A_i) = \sum \Pr(A_i) \text{ si } A_i \cap A_j = \emptyset \text{ donde } \emptyset \text{ es el evento imposible}$$

Los tres axiomas permiten establecer los siguientes teoremas:

Si el evento A es un subconjunto del evento B entonces la  $\Pr(A)$  es menor o igual a la  $\Pr(B)$ :  $A \subset B \rightarrow \Pr(A) \leq \Pr(B)$ .

Para cualquier evento A la probabilidad de su complemento,  $\Pr(A^c)$ , es igual a:  $1 - \Pr(A)$ .

La probabilidad del evento imposible es cero:  $\Pr(\emptyset) = 0$ .

La probabilidad de la unión de dos eventos es:

$$\Pr(A \cup B) = \Pr(A) + \Pr(B) - \Pr(A \cap B)$$

Si los eventos A y B son independientes la probabilidad de ocurrencia de un evento no influye en la del otro s la siguiente probabilidad condicional:  $\Pr(A|B) = \Pr(A)$ . Si no son independientes la probabilidad condicional se define de la siguiente manera:

$$\Pr(A|B) = \frac{\Pr(A \cap B)}{\Pr(B)}$$

### 3. VARIABLE ALEATORIA

Variables como el PIB o la inflación no son conocidas antes de que sean generadas y reportadas por las agencias gubernamentales, lo cual implica que su resultado nunca es conocido de antemano, por ello a este tipo de variables se les conoce como variables **aleatorias**.

Para ejemplificar lo que significa una variable aleatoria considere el lanzamiento de dos monedas al aire. El conjunto de resultados posibles para este experimento está dado por:

$\Omega = \{(s,s), (s,a), (a,s), (a,a)\}$  siendo  $\Omega$  el conjunto universal o de resultados, donde s= sol y a= águila

Podemos definir la variable aleatoria  $X$  como el número de soles que aparecen, así que  $X$  puede tomar los valores 2, 1 y 0.

Formalmente una variable aleatoria es una función medible que vincula al conjunto de resultados con el conjunto de los números reales. Las variables aleatorias son discretas o continuas. Una variable aleatoria discreta es aquella que puede tomar sólo un número finito de valores o bien infinito pero que pueden ser contados. Una variable aleatoria continua puede tomar un número infinito de valores en cualquier intervalo.

Las probabilidades asociadas a una variable aleatoria se calculan a través de su función de distribución probabilística. En nuestro ejemplo, al lanzar dos monedas y donde  $X =$  salga sol, sus probabilidades se muestran en el cuadro siguiente.

### Cuadro 1

Función de distribución de una variable aleatoria discreta

$x$	$\Pr(X=x)$	$\Pr(X \leq x)$
0	1/4	1/4
1	2/4	3/4
2	1/4	4/4
$\Sigma \Pr =$	4/4=1.0	

También es posible determinar la probabilidad de que la variable aleatoria tome valores a lo mucho iguales a  $x$ , esto es la función de densidad acumulada y se define como:  $\Pr(X) \leq x$ .

Para variables aleatorias continuas no es relevante calcular la probabilidad de que  $X$  tome un valor particular  $x$  debido a que en un intervalo  $a,b$  existe un número infinito de puntos. Por ello, la cuestión relevante es calcular la probabilidad de que  $X$  tome valores en el intervalo  $a,b$  donde  $a < b$ ;

$$\Pr(a \leq x \leq b) = \int_a^b f(x)dx$$

La función de densidad acumulada de una variable aleatoria continua se define como:

$$\Pr(X \leq x) = \int_{-\infty}^x f(x)dx$$

Para una variable aleatoria discreta, su media o valor esperado es un promedio ponderado de sus posibles resultados, en donde el ponderador es la probabilidad asociada a cada valor, tal y como se indica a continuación:

$$\mu_x = E(X) = pr_1x_1 + pr_2x_2 + \dots + pr_nx_n = \sum pr_i x_i$$

donde  $pr_i$  es la probabilidad de  $x_i$  y  $E$  es el operador de esperanza matemática.

Su varianza es definida por la siguiente expresión:

$$Var(X) = \sigma_X^2 = E[X - E(X)]^2 = E[X - \mu_x]^2 = E(X)^2 - (\mu_x)^2$$

Para una variable aleatoria continua la media es igual a:

$$\mu_x = E(X) = \int_{-\infty}^{\infty} xf(x)dx$$

Su varianza es:

$$Var(X) = \sigma_X^2 = \int_{-\infty}^{\infty} (x - \mu_x)^2 f(x)dx$$

El operador de valor esperado (E) que hemos utilizado tiene algunas propiedades interesantes:

$$E(k) = k \quad \text{siendo } k \text{ una constante cualquiera}$$

$$E(X + Y) = E(X) + E(Y)$$

$$E(XY) = E(X)E(Y) \quad \text{si } X, Y \text{ son independientes}$$

Asimismo la varianza tiene las propiedades siguientes:

$$Var(kX) = k^2 Var(X) \quad \text{siendo } k \text{ una constante cualquiera}$$

$$Var(X + Y) = Var(X) + Var(Y) \quad \text{si } X, Y \text{ son independientes}$$

La distribución conjunta de una variable aleatoria X y una variable aleatoria Y para el caso discreto es igual a la lista de probabilidades de ocurrencia de todos los resultados para X y Y.

Por ejemplo, en el siguiente cuadro aparecen datos del número de personas de cuatro entidades federativas del país a los que se les preguntó su grado de felicidad y sus opiniones se clasificaron en muy feliz, feliz, poco feliz y nada feliz.

	1. Baja California	2. Chihuahua	3. San Luis Potosí	4. Chiapas	Total
1. Muy feliz	300	100	80	10	<b>490</b>
2. Feliz	100	100	70	20	<b>290</b>
3. Poco feliz	30	20	50	10	<b>110</b>
4. Nada feliz	70	80	200	260	<b>610</b>
<b>Total</b>	<b>500</b>	<b>300</b>	<b>400</b>	<b>300</b>	<b>1500</b>

Con base en esa información se pueden definir dos variables aleatorias, por un lado la de entidad federativa de procedencia X y por otra la del grado de felicidad Y.

Tomando como base los números con los que se ordenaron los encabezados de la información y dividiendo los datos entre el valor total de 1500, obtenemos un cuadro de probabilidades:

X Y	1	2	3	4	Total
1	0.200	0.067	0.053	0.007	<b>0.327</b>
2	0.067	0.067	0.047	0.013	<b>0.193</b>
3	0.020	0.013	0.033	0.007	<b>0.073</b>
4	0.047	0.053	0.133	0.173	<b>0.407</b>
Total	<b>0.333</b>	<b>0.200</b>	<b>0.267</b>	<b>0.200</b>	<b>1.000</b>

Así, la probabilidad de que una persona sea muy feliz es:  $P(Y=1)=0.327$ .

La probabilidad de que una persona provenga de Chiapas es:  $P(X=4)=0.2$ .

La probabilidad de que provenga de Chiapas y que sea muy feliz es:  $P(X=4 \text{ y } Y=1)=0.007$ .

Así la función de densidad conjunta de dos variables aleatorias puede definirse como:

$$f_{xy}(x,y) := P(X=x, Y=y)$$

Y la función de densidad acumulada:

$$F_{xy}(x,y) := P(X \leq x \text{ y } Y \leq y)$$

Por ejemplo, la función de densidad acumulada para  $F(4,1)$  con los datos del cuadro anterior es:

$$F(4,1)=P(X \leq 4 \text{ y } Y \leq 1)=0.327$$

Para una variable continua las funciones de densidad y la acumulada son respectivamente:

$$f(x,y)=\partial^2 F(x,y)/ \partial x \partial y$$

$$F(x,y)=P(X \leq x \text{ y } Y \leq y)$$

La función de densidad marginal de  $X$  es igual a la función de densidad de  $X$ . Por ejemplo: la función de densidad marginal de  $X$  en  $X=3$  es igual a 0.267. O lo que es lo mismo igual a  $f(3,1)+f(3,2)+f(3,3)+f(3,4)$  es decir sumamos todos los valores de  $Y$  manteniendo constante el valor de  $X$ .

Las funciones de densidad condicionales se pueden expresar como un cociente de la conjunta dividida por la marginal de la variable condicionante, es decir la podemos obtener con la siguiente expresión:

$$f(A|B)=\frac{f_{AB}(A,B)}{f_B(B)}$$

Aplicando la conjunta al ejemplo de la felicidad, es posible calcular la probabilidad condicional de que una persona sea muy feliz dado que proviene de Chiapas, esto lo podemos expresar de la siguiente manera:

$$f(Y=1|X=4) = \frac{f_{YX}(Y=1, X=4)}{f_X(X=4)} = \frac{0.007}{0.2} = 0.035$$

Las variables resultarían independientes si la conjunta puede expresarse como el producto de las marginales, es decir que:

$$f(A,B) = f(A)f(B)$$

En el ejemplo anterior se puede verificar que no hay independencia estadística dado que:

$$f_{YX}(Y=1, X=4) \neq f_Y(Y=1)f_X(X=4)$$

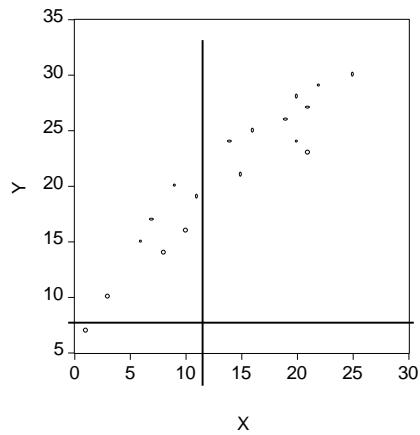
El operador de esperanza matemática puede utilizarse para describir las funciones de distribución conjuntas. Por ejemplo, la covarianza entre X y Y estará dada por la siguiente expresión:

$$\text{Cov}(X,Y) = E(X-E(X))(Y-E(Y))$$

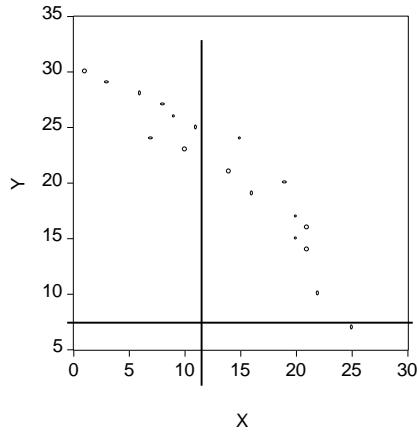
La covarianza mide la asociación lineal entre las dos variables, observe que en la gráfica 5 panel (a) las dos variables están siempre abajo o arriba de sus medias, presentando una covarianza positiva, mientras que en la (b) cuando una está arriba de su media la otra está debajo de su media mostrando una covarianza negativa.

Gráfica 5

### Covarianza



(a)



(b)

El problema que tiene la covarianza como medida de asociación lineal es que depende de las unidades de medida de X y Y, para evitar ese problema se normaliza dividiéndola entre las desviaciones estándar de las dos variables. A esta covarianza, liberada de unidad de medida, se le conoce como el coeficiente de correlación de Pearson y se denota de la siguiente forma:

$$\rho(X, Y) = \text{Cov}(X, Y) / \sigma_X \sigma_Y$$

El coeficiente de correlación tiene un rango que va de -1 a +1.

### 3.1 Algunas funciones de distribución particulares

En muchos casos prácticos surgen variables aleatorias con distribuciones de probabilidad similares a algunas distribuciones utilizadas cotidianamente. Veamos las más importantes para los modelos econométricos.

#### Distribución Binaria

La variable aleatoria X toma únicamente valores 0 y 1 con probabilidades p y 1-p, a estas variables se les denomina variables aleatorias Bernoulli. Donde  $P(X=1)=p$  y la  $P(X=0)=1-p$

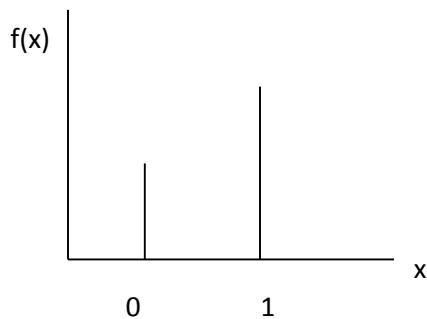
Por lo tanto, la función de probabilidad es:

$$f(x)=P(X=x)=p^x(1-p)^{1-x}$$

Un importante grupo de modelos econométricos utilizan variables dependientes cualitativas que son codificadas como uno y cero, estos modelos se conocen como logit, probit y tobit. Su función de distribución se podría representar con una gráfica de bastón (véase la gráfica 6):

Gráfica 6

Función de distribución de variables cualitativas



Por ejemplo, considere que la probabilidad de abordar el metrobús vacío en la estación Ciudad Universitaria sea igual a 0.4. Es posible encontrar la probabilidad de abordar vacío el metrobús en miércoles y viernes pero no en los demás días. Suponiendo que la probabilidad de abordar el metrobús vacío en un día cualquiera es independiente de abordarlo en cualquier otro día, la secuencia de ese evento está dada por:

$$X = [0,0,1,0,1,0,0]$$

La probabilidad de  $X$  está dada por:

$$P(X) = p^2(1 - p)^5 = 0.4^2(0.6)^5 = 0.0124416$$

## **Distribución Binomial**

La variable aleatoria mide el número de éxitos en n experimentos de éxito-fracaso, iguales e independientes.

Su función de densidad probabilística es definida con la siguiente expresión:

$$f(x) = P(X = x) = C_x^n p^x (1 - p)^{n-x}$$

Donde:  $C_x^n = n! / x! (n - x)!$

Por ejemplo, considere tres consumidores que van a una tienda de ropa y obtenga la probabilidad de que ninguno haga una compra si la probabilidad de compra es de 0.3. Aplicando la fórmula:

$$f(0) = P(X = 0) = C_0^3 0.3^0 (1 - 0.3)^{3-0} = 0.343$$

En R se puede calcular esa probabilidad utilizando la siguiente instrucción: `dbinom(x,n,p)`. Para el caso previo en R se obtiene el resultado escribiendo:

```
> dbinom(0,3,0.3)
```

[1] 0.343

## Distribución Poisson

La variable aleatoria toma una serie de valores en un tiempo o espacio en donde el número de valores es igual a cualquier entero entre cero e infinito. La muestra,  $n$ , es grande tal que  $np=\lambda$  y la probabilidad de acierto,  $p$ , es pequeña:

$$f(x) = P(X = x) = \frac{\lambda^x e^{-\lambda}}{x!}$$

Donde  $\lambda$  es un número dado

Por ejemplo, suponga que se toman datos de los clientes que llegan a una ventanilla de banco durante 15 minutos en un día de la semana por la tarde. Se considera que la probabilidad de que un cliente llegue es la misma en cualquier par de períodos de igual duración y que la llegada o no de un cliente es independiente de la llegada o no en cualquier otro período de tiempo. Si por los registros del banco se sabe que en 15 minutos llegan 10 clientes, es posible plantear una función de probabilidad de Poisson:

$$f(x) = P(X = x) = \frac{10^x e^{-10}}{x!}$$

Si queremos saber la probabilidad de que lleguen exactamente cinco clientes, tenemos:

$$f(5) = P(X = 5) = \frac{10^5 e^{-10}}{5!} = 0.0378$$

En R podemos obtener el mismo resultado con la función dpois(x,λ). En nuestro caso se escribe en R:

```
> dpois(5,10)
```

```
[1] 0.03783327
```

## Distribución Normal

La distribución Normal es sin duda la distribución más usual que se emplea en estadística. La densidad de probabilidad de esa función está dada por la fórmula siguiente:

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

donde:

$\sigma$  es la desviación estándar,  $\mu$  es la media,  $\pi$  es 3.1416 y  $e$  es el número 2.71828

La función normal depende de los valores de la media,  $\mu$ , y de la desviación estándar,  $\sigma$ , que al variar generan una familia infinita de distribuciones normales.

Para evitar construir una tabla de probabilidades cada que varían los parámetros de la función, lo que se hace es estandarizarla: Una variable aleatoria es estandarizada si se le resta su media y se divide entre su desviación estándar, lo que da como resultado una variable con media cero y varianza igual a la unidad.

La fórmula de la distribución normal estándar es:

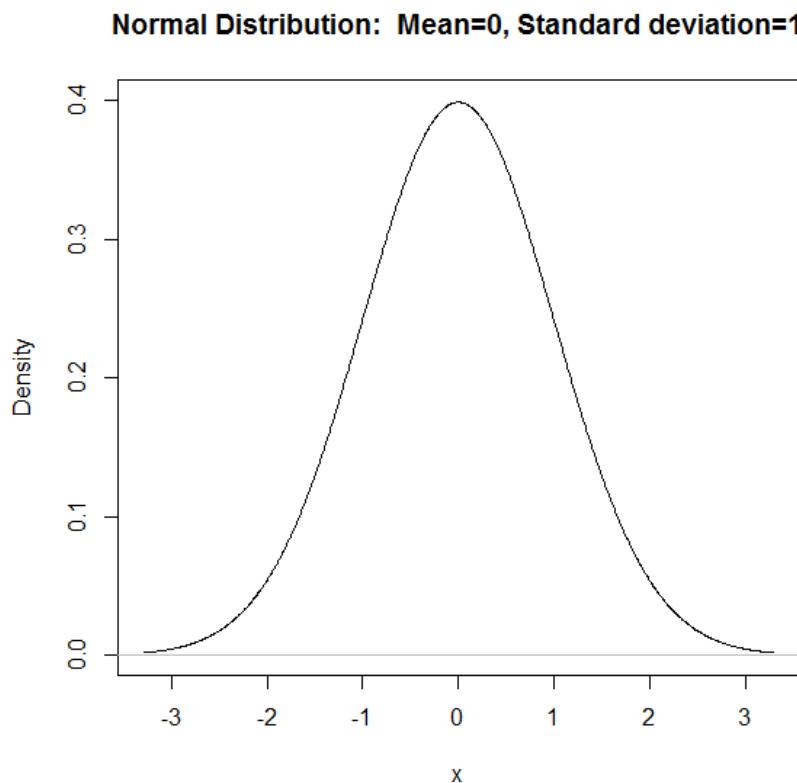
$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$$

donde  $z = \frac{x-\mu}{\sigma}$  y  $-\infty \leq z \leq +\infty$

En RCommander podemos fácilmente generar la función de densidad de una normal estandarizada, para lo cual se selecciona en el menú principal la opción DISTRIBUTIONS/Continous distributions/Normal distributions/Plot normal distribution. En la ventana que se abre se selecciona para una media de cero y una varianza unitaria. El resultado es el siguiente:

Gráfica 7

Distribución Normal Estandarizada



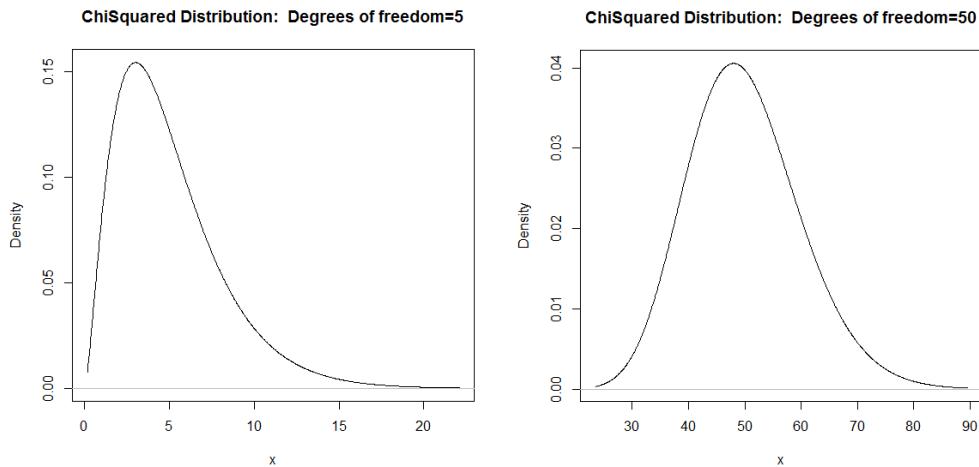
Aproximadamente el 68% del área o probabilidad de la curva se encuentra a más menos una desviación estándar de la media, el 95.5% de probabilidad se encuentra a más menos dos desviaciones y casi toda el área entre más menos tres desviaciones.

La distribución normal es muy utilizada ya que es simétrica y puede caracterizarse por su media y varianza. Además de que una propiedad importante de dos o más variables aleatorias distribuidas normalmente, con la misma media y varianza, es que su suma ponderada se distribuye también como una normal.

En el análisis econométrico se utilizan, con gran frecuencia, tres distribuciones derivadas de la normal:

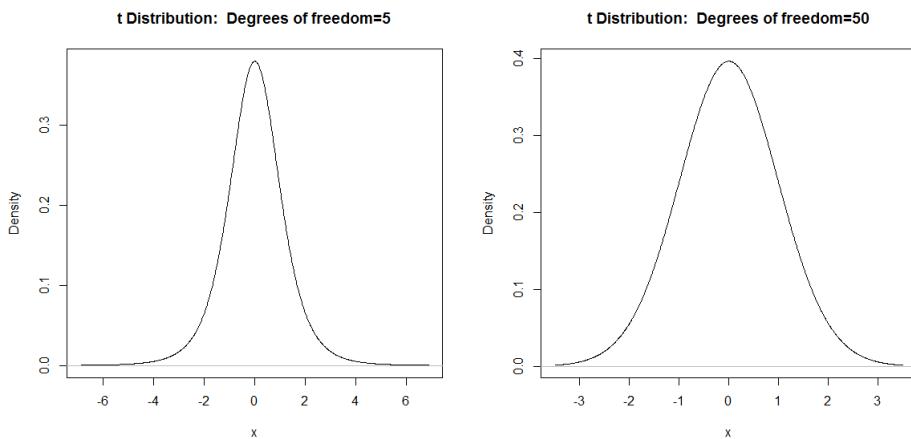
La distribución ji cuadrada,  $\chi^2$ : Sea  $Z$  una variable aleatoria con distribución normal estandarizada  $Z \sim N(0,1)$ . Al elevarla al cuadrado tenemos  $Z^2 \sim \chi^2(1)$ , se cumple entonces que  $Z_1^2 + Z_2^2 \sim \chi^2(2)$  sí son independientes y normalmente distribuidas, por lo tanto  $\sum_{i=1}^{gl} Z_i \sim \chi^2(gl)$ .

En RComnader utilizando la misma opción del menú DISTRIBUTIONS es posible generar la gráfica de densidad para la Ji-cuadrada simplemente anotando el número de grados de libertad en la ventana que se despliega para este tipo de distribuciones. El resultado de una ji cuadrada con 5 y 50 grados de libertad se muestra en las dos gráficas siguientes:



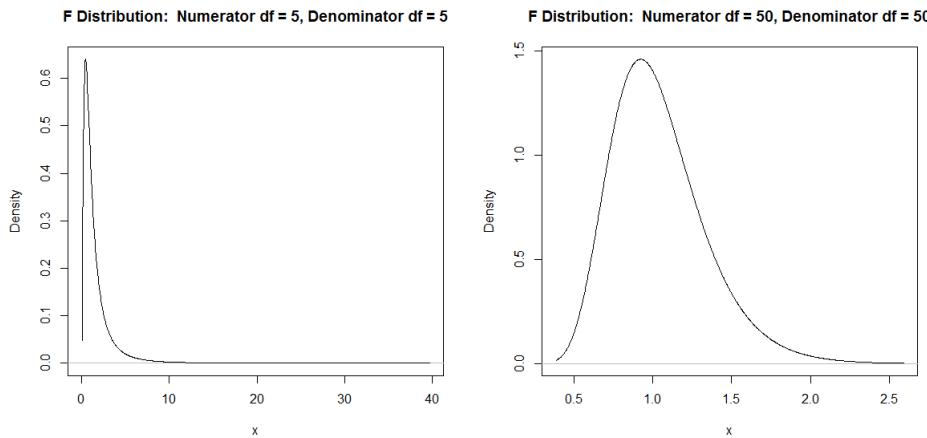
La distribución t de student: Es una combinación de dos variables independientes tal que;  $t(gl) = Z / (\sqrt{W}/\sqrt{gl})$  donde  $W \sim \chi^2(gl)$ .

En RComnader utilizando la misma opción del menú DISTRIBUTIONS es posible generar la gráfica de densidad para la t, simplemente anotando el número de grados de libertad en la ventana que se despliega para este tipo de distribuciones. El resultado de una t con 5 y 50 grados de libertad se muestra en las dos gráficas siguientes:



La distribución F de Fisher es una combinación de dos variables ji-cuadrada tal que;  $F(gl_1, gl_2) = (Z_1^2 / gl_1) / (Z_2^2 / gl_2)$ . Es fácil concluir que  $F(1, gl_2) = t(gl_2)^2$ .

En RComnader utilizando la misma opción del menú DISTRIBUTIONS es posible generar la gráfica de densidad para la F, simplemente anotando el número de grados de libertad que corresponden al numerador y al denominador en la ventana que se despliega para este tipo de distribuciones. El resultado de una t con 5 y 50 grados de libertad tanto en el numerador como en el denominador se muestra en las dos gráficas siguientes:



En todos los casos es fácil constatar que cuando el número de grados de libertad aumenta en estas funciones tienden a parecerse más a una normal.

Utilizando RCommander además de la generación de las gráficas de densidad, también es posible generar funciones de distribución acumuladas, probabilidades y realizar simulaciones para calcular la media y la desviación estándar.

#### 4. BREVE REPASO DE ÁLGEBRA DE MATRICES

El uso de álgebra de matrices nos brinda un modo de representar y manejar datos de forma compacta. Una matriz no es más que un arreglo rectangular de elementos o números en renglones o columnas. Por ejemplo, el siguiente arreglo de datos es una matriz:

$$\mathbf{B} = \begin{bmatrix} 50 & 80 & 100 \\ 30 & 40 & 50 \\ 60 & 50 & 80 \end{bmatrix}$$

De forma general, podemos representar una matriz en un cuadro de entradas y salidas como el siguiente:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \cdots & \cdots & \cdots & \cdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}$$

La matriz **A** es de dimensión mxn,  $\dim(\mathbf{A})=(m \times n)$  es decir tiene m por n elementos arreglados en m renglones y n columnas.

O bien de una forma más compacta podemos representarla como:

$$\mathbf{A} = \{a_{ij}\}$$

La matriz **A** transpuesta se representa por **A'** y se obtiene permutando las filas por las columnas de **A**.

Por ejemplo, suponiendo una matriz **A** particular como la siguiente:

$$\mathbf{A} = \begin{bmatrix} 10 & 5 & 8 \\ 20 & 10 & 30 \\ 30 & 40 & 1 \end{bmatrix}$$

La matriz transpuesta para A es:

$$\mathbf{A}' = \begin{bmatrix} 10 & 20 & 30 \\ 5 & 10 & 40 \\ 8 & 30 & 1 \end{bmatrix}$$

En R se pueden trabajar las matrices A y B que ya hemos visto. Para introducir la matriz A podemos escribir:

```
A<-matrix(c(10, 20, 30, 5, 10, 40, 8, 30, 1), nrow=3, ncol=3, byrow=T)
```

Las opciones que se utilizaron para generar la matriz A fueron:

c() que es el vector de datos

nrow que define el número de renglones

ncol es el número de columnas

byrow=T le indica que los datos se incorporan por fila, si byrow=F los datos se incorporan por columna.

Para la matriz B podemos escribir:

```
B<-matrix(c(50, 80, 100, 30, 40, 50, 60, 50, 80), nrow=3, ncol=3, byrow=T)
```

Podemos visualizar las matrices con la siguiente instrucción:

```
> list(A)
```

```
[[1]]
```

```
,1] [,2] [,3]
```

```
[1,] 10 20 30
```

```
[2,] 5 10 40
```

```
[3,] 8 30 1
```

```
> list(B)
```

```
[[1]]
```

```
,1] [,2] [,3]
```

```
[1,] 50 80 100
```

[2,] 30 40 50

[3,] 60 50 80

La transpuesta de la matriz A la obtenemos con la siguiente instrucción:

> t(A)

[,1] [,2] [,3]

[1,] 10 5 8

[2,] 20 10 30

[3,] 30 40 1

Para matrices que tienen las mismas dimensiones se pueden realizar las operaciones básicas de suma y resta. Si comparamos las matrices **A'** y **B** nos daremos cuenta que son de la misma dimensión pero no son iguales, es decir **A' ≠ B**. Para que dos matrices sean iguales no solamente deben ser de la misma dimensión sino además deben ser iguales elemento a elemento, en este caso cada elemento  $a_{ij}$  en la matriz **A'** debiera ser igual a cada elemento  $b_{ij}$  en la matriz **B** para todo i, j.

Para sumar dos matrices se debe sumar elemento a elemento; si **A'** y **B** tienen el mismo orden, **A'+B** es una nueva matriz **C** del mismo orden en donde:

$$C_{ij} = a_{ij} + b_{ij} \text{ para toda } i, j$$

$$\mathbf{A}' + \mathbf{B} = \mathbf{C} = \begin{bmatrix} 10 & 5 & 8 \\ 20 & 10 & 30 \\ 30 & 40 & 1 \end{bmatrix} + \begin{bmatrix} 50 & 80 & 100 \\ 30 & 40 & 50 \\ 60 & 50 & 80 \end{bmatrix} = \begin{bmatrix} 60 & 85 & 108 \\ 50 & 50 & 80 \\ 90 & 90 & 81 \end{bmatrix}$$

En R obtenemos esa misma suma con la instrucción siguiente:

```
> t(A)+B
```

```
[,1] [,2] [,3]
[1,] 60 85 108
[2,] 50 50 80
[3,] 90 90 81
```

La resta o sustracción entre dos matrices **A** y **B** requiere al igual que la suma de la conformabilidad de las matrices, y el resultado es: **C**= $a_{ij}-b_{ij}$ .

$$\mathbf{A}' - \mathbf{B} = \mathbf{C} = \begin{bmatrix} 10 & 5 & 8 \\ 20 & 10 & 30 \\ 30 & 40 & 1 \end{bmatrix} - \begin{bmatrix} 50 & 80 & 100 \\ 30 & 40 & 50 \\ 60 & 50 & 80 \end{bmatrix} = \begin{bmatrix} -40 & -75 & -92 \\ -10 & -30 & -20 \\ -30 & -10 & -79 \end{bmatrix}$$

En R obtenemos esa misma resta con la instrucción siguiente:

```
> t(A)-B
```

```
[,1] [,2] [,3]
[1,] -40 -75 -92
[2,] -10 -30 -20
[3,] -30 -10 -79
```

Otra operación básica con matrices es la multiplicación escalar, si se multiplica a la matriz  $A'$  por el escalar  $z$  el resultado es:  $zA' = za_{ij}$ . Por ejemplo, para duplicar la matriz  $A'$  ésta debe multiplicarse por el escalar  $z=2$ , tal y como se indica a continuación:

$$zA' = 2A' = \begin{bmatrix} 20 & 10 & 16 \\ 40 & 20 & 60 \\ 60 & 80 & 2 \end{bmatrix}$$

En R obtenemos el producto escalar con la instrucción siguiente:

```
> 2*t(A)
[,1] [,2] [,3]
[1,] 20 10 16
[2,] 40 20 60
[3,] 60 80 2
```

Una matriz de dimensión  $1 \times n$  es un vector renglón y lo representamos generalmente con letras minúsculas:

$$\mathbf{v} = [v_{11} \quad v_{12} \quad \dots \quad v_{1n}]$$

Un vector de dimensión  $n \times 1$  es un vector columna, por ejemplo el siguiente vector **c**:

$$\mathbf{c} = \begin{bmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{n1} \end{bmatrix}$$

Para cualquier par de vectores  $\mathbf{v}$  y  $\mathbf{c}$  el producto interno de vectores es:

$$\mathbf{v}\mathbf{c} = [v_{11} \ v_{12} \ \dots \ v_{1n}] \begin{bmatrix} c_{11} \\ c_{21} \\ \vdots \\ c_{n1} \end{bmatrix} = v_{11}c_{11} + v_{12}c_{21} + \dots + v_{1n}c_{n1} = \sum v_{1k}c_{k1}$$

con  $k=1, \dots, n$

Por ejemplo, para los siguientes vectores el producto interno es:

Por ejemplo, considere los siguientes vectores  $\mathbf{v}$ ,  $\mathbf{c}$ :

$$\mathbf{v} = [1 \ 4 \ 6] \quad \mathbf{c} = \begin{bmatrix} 2 \\ 5 \\ 10 \end{bmatrix}$$

El producto interno está dado por:

$$\mathbf{v}\mathbf{c} = [1 \ 4 \ 6] \begin{bmatrix} 2 \\ 5 \\ 10 \end{bmatrix} = [2 + 20 + 60] = [82]$$

Al multiplicar dos vectores se multiplica cada elemento del vector fila  $i$ -ésima por su correspondiente elemento en la  $j$ -ésima columna del segundo vector y posteriormente se suman esos productos.

La multiplicación matricial es una generalización de la multiplicación vectorial repetida. Por ejemplo, el producto de las matrices **A'** y **B** está dado por:

$$\mathbf{A}'\mathbf{B} = \begin{bmatrix} 10 & 5 & 8 \\ 20 & 10 & 30 \\ 30 & 40 & 1 \end{bmatrix} \begin{bmatrix} 50 & 80 & 100 \\ 30 & 40 & 50 \\ 60 & 50 & 80 \end{bmatrix} = \begin{bmatrix} 1130 & 1400 & 1890 \\ 3100 & 3500 & 4900 \\ 2760 & 4050 & 5080 \end{bmatrix}$$

El primer elemento de la matriz final anterior se obtiene por producto interno de vectores de la manera siguiente:

$$\mathbf{a}_1 \mathbf{b}_1 = [10 \quad 5 \quad 8] \begin{bmatrix} 50 \\ 30 \\ 60 \end{bmatrix} = [500 + 150 + 480] = [1130]$$

La multiplicación matricial la podemos llevar a cabo en R con la siguiente instrucción:

```
> t(A) %*% B
 [,1] [,2] [,3]
[1,] 1130 1400 1890
[2,] 3100 3500 4900
[3,] 2760 4050 5080
```

Se debe ser cuidadoso con R debido a que tiene dos formas de multiplicar estas matrices, la que acabamos de hacer que sigue el cálculo de álgebra lineal, mientras que `t(A)*B` hubiera dado lugar a una multiplicación elemento a elemento como la que se presenta enseguida:

```
> t(A)*B
```

[,1]	[,2]	[,3]
[1,] 500 400 800		
[2,] 600 400 1500		
[3,] 1800 2000 80		

De la multiplicación matricial se puede deducir:

Si una matriz **A** es de dimensión mxn y una matriz **B** es de orden nxp, el producto **AB** es una matriz de orden mpx.

Para poder realizar la multiplicación debe cumplirse que la dimensión columna de la primer matriz sea igual a la dimensión fila de la segunda matriz.

Existe un tipo de matrices con la peculiar característica que multiplicada por sí misma, cualquier número de veces, es siempre igual a la matriz original, a estas matrices se les llama idempotentes:

$$\mathbf{A} = \mathbf{A}^2 = \mathbf{A}^3 \dots \mathbf{A}^n$$

A partir de las operaciones elementales de matrices que hemos desarrollado hasta este momento, podemos establecer un conjunto de propiedades:

- 1) La suma de matrices es commutativa:  $\mathbf{A}+\mathbf{B}=\mathbf{B}+\mathbf{A}$
- 2) La multiplicación de matrices en general no es commutativa. En particular podemos decir que una matriz cuadrada commuta consigo misma y con la matriz identidad:  $\mathbf{AB}\neq\mathbf{BA}$     $\mathbf{AA}=\mathbf{AA}$        $\mathbf{AI}=\mathbf{IA}$

La matriz identidad en su diagonal principal tiene únicamente elementos escalares iguales a la unidad y ceros fuera de ella. Dado que es una matriz cuadrada se suele representar como  $I_n$  cuando es de dimensión  $n \times n$  o simplemente como  $I$  cuando no hay confusión acerca de su dimensión:

$$I_n = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & 1 \end{bmatrix}$$

En R se puede generar una matriz identidad con la función `diag(m)`, siendo  $m$  el orden de la matriz. Por ejemplo, generamos una matriz identidad de cuatro por cuatro:

```
> diag(4)
[,1] [,2] [,3] [,4]
[1,] 1 0 0 0
[2,] 0 1 0 0
[3,] 0 0 1 0
[4,] 0 0 0 1
```

La suma y la multiplicación de matrices cumplen con la propiedad asociativa:

Suma:  $(A+B)+C=A+(B+C)$

Multiplicación:  $(AB)C=A(BC)$

La multiplicación y la multiplicación escalar cumplen con la propiedad distributiva:

Multiplicación:  $\mathbf{A}(\mathbf{B}+\mathbf{C})=\mathbf{AB}+\mathbf{AC}$

Multiplicación escalar:  $z(\mathbf{A}+\mathbf{B})=z\mathbf{A}+z\mathbf{B}$

También las matrices transpuestas cumplen con ciertas propiedades que conviene conocer:

La transpuesta de una matriz transpuesta es igual a la matriz original:  $(\mathbf{A}')'=\mathbf{A}$

La transpuesta de una suma de matrices es igual a la suma de sus transpuestas:  
 $(\mathbf{A}+\mathbf{B})'=\mathbf{A}'+\mathbf{B}'$

La transpuesta del producto de matrices es igual al producto inverso de sus transpuestas:  $(\mathbf{AB})'=\mathbf{B}'\mathbf{A}'$  o bien  $(\mathbf{ABC})=\mathbf{C}'\mathbf{B}'\mathbf{A}'$

Una operación muy útil para evaluar la existencia de solución a un sistema de ecuaciones es el determinante de una matriz. Consideraremos primero una sencilla matriz de dimensión 2x2:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

El determinante de  $\mathbf{A}$  es una función de los valores  $a_{ij}$  de la matriz:

$$\det(\mathbf{A}) = |\mathbf{A}| = a_{11}a_{22} - a_{12}a_{21}$$

Para cualquier matriz cuadrada, independientemente de su orden, el determinante se obtiene como:

$$C_{ij} = (-1)^{i+j} \det(M_{ij})$$

En donde  $C$  es llamado el cofactor del elemento  $(i,j)$  y  $M$  es el menor o la submatriz que se obtiene al eliminar la fila  $i$  y la columna  $j$  de la matriz  $A$ .

Entonces podemos reescribir el determinante de nuestra matriz  $A$  de orden 2 como:

$$\det(A) = |A| = a_{11}C_{11} - a_{12}C_{12}$$

Por ejemplo, considerando la matriz  $A'$  el menor  $M_{11}$  se obtiene eliminando la primer fila y la primer columna:

$$M_{11} = \begin{bmatrix} 10 & 30 \\ 40 & 1 \end{bmatrix}$$

Generalizando se puede calcular el determinante de  $A$  por cofactores, el resultado sería el siguiente, si es que comenzamos la expansión por el primer renglón:

$$|A| = 10C_{11} + 5C_{12} + 8C_{13}$$

Los menores son:

$$M_{11} = \begin{bmatrix} 10 & 30 \\ 40 & 1 \end{bmatrix} M_{12} = \begin{bmatrix} 20 & 30 \\ 30 & 1 \end{bmatrix} M_{13} = \begin{bmatrix} 20 & 10 \\ 30 & 40 \end{bmatrix}$$

Los cofactores son:

$$C_{11} = [(10)(1) - (40)(30)] = -1190$$

$$C_{12} = -[(20)(1) - (30)(30)] = 880$$

$$C_{13} = [(20)(40) - (30)(10)] = 500$$

El determinante de **A** es:

$$|A| = 10(-1190) + 5(880) + 8(500) = -3500$$

Para calcular el determinante en R escribimos:

```
> det(t(A))
```

```
[1] -3500
```

Ahora estamos en condiciones de calcular la inversa  $A^{-1}$  de una matriz **A** cuadrada y no singular, que es la matriz única que cumple con la relación:

$$AA^{-1} = I = A^{-1}A$$

La matriz inversa juega la misma función que el reciproco en el álgebra ordinaria.

La inversa la obtenemos con la siguiente fórmula:

$$\mathbf{A}^{-1} = \frac{1}{|\mathbf{A}|} adj(\mathbf{A})$$

En donde la matriz adjunta de  $\mathbf{A}$ ,  $Adj(\mathbf{A})$ , es la matriz transpuesta de cofactores de  $\mathbf{A}$ .

Retomando la matriz  $\mathbf{A}'$  del ejemplo, su matriz de cofactores tendrá 9 elementos, que podemos calcular utilizando la fórmula de cofactores ya vista antes:

$$\mathbf{C} = \begin{bmatrix} c_{11} & c_{13} & c_{14} \\ c_{21} & c_{22} & c_{23} \\ c_{31} & c_{32} & c_{32} \end{bmatrix} = \begin{bmatrix} -1190 & 880 & 500 \\ 315 & 230 & -250 \\ 70 & -140 & 0 \end{bmatrix}$$

De este modo la adjunta de  $\mathbf{A}'$  es:

$$adj(\mathbf{A}') = \mathbf{C}' = \begin{bmatrix} -1190 & 315 & 70 \\ 880 & -230 & -140 \\ 500 & -250 & 0 \end{bmatrix}$$

Utilizando la fórmula para calcular la inversa obtenemos (las cifras de la matriz han sido redondeadas):

$$\mathbf{A}'^{-1} = -\frac{1}{3500} \begin{bmatrix} -1190 & 315 & 70 \\ 880 & -230 & -140 \\ 500 & -250 & 0 \end{bmatrix} = \begin{bmatrix} 0.34 & -0.90 & -0.02 \\ -0.25 & 0.07 & 0.04 \\ -0.14 & 0.07 & 0 \end{bmatrix}$$

En R obtenemos la inversa de  $A'$  con la siguiente instrucción:

```
> solve(t(A))
 [,1]      [,2]      [,3]
[1,] 0.3400000 -0.09000000 -0.02
[2,] -0.2514286  0.06571429  0.04
[3,] -0.1428571  0.07142857  0.00
```

Otras dos operaciones con matrices que nos serán de utilidad son la traza y el producto Kronecker.

La traza de una matriz es la suma de los elementos de su diagonal principal y tiene las siguientes propiedades:

$$\text{tr}(A+B)=\text{tr}(A)+\text{tr}(B)$$

$$\text{tr}(AB)=\text{tr}(BA)$$

$$\text{tr}(ABC)=\text{tr}(CAB)=\text{tr}(BCA)$$

$$\text{tr}(k)=k \text{ donde } k \text{ es una constante}$$

Por ejemplo, la traza de la matriz A del ejemplo es:

$$\text{tr}(A)=10+10+1=21$$

Es posible calcular la traza en R, por ejemplo para la matriz A tendríamos:

```
> sum(diag(A))
```

```
[1] 21
```

El producto Kronecker nos permite multiplicar matrices que por sus dimensiones no son conformables para la multiplicación. Sea **A** una matriz cualquiera con dimensiones (n,k) y **B** una matriz (m,p). El producto Kronecker de **A** por **B** es:

$$K = A \otimes B = [a_{ij}B] = \begin{bmatrix} a_{11}B & a_{12}B & \dots & a_{1k}B \\ a_{21}B & a_{22}B & \dots & a_{2k}B \\ \dots & \dots & \dots & \dots \\ a_{n1}B & a_{n2}B & \dots & a_{nk}B \end{bmatrix}$$

Para nuestras matrices **A** y **B** su producto Kronecker se obtiene en R con la siguiente instrucción:

```
> kronecker(A,B)
```

	[,1]	[,2]	[,3]	[,4]	[,5]	[,6]	[,7]	[,8]	[,9]
[1,]	500	800	1000	1000	1600	2000	1500	2400	3000
[2,]	300	400	500	600	800	1000	900	1200	1500
[3,]	600	500	800	1200	1000	1600	1800	1500	2400
[4,]	250	400	500	500	800	1000	2000	3200	4000
[5,]	150	200	250	300	400	500	1200	1600	2000
[6,]	300	250	400	600	500	800	2400	2000	3200
[7,]	400	640	800	1500	2400	3000	50	80	100
[8,]	240	320	400	900	1200	1500	30	40	50

[9.] 480 400 640 1800 1500 2400 60 50 80

Ahora veremos algunos productos matriciales que usaremos con frecuencia y que llamaremos formas cuadráticas.

La primera es una suma de cuadrados, si definimos un vector columna **u**:

$$\mathbf{u} = \begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{bmatrix}$$

El producto de la transpuesta de ese vector por el vector original da lugar a un escalar que es la suma de cuadrados de los elementos del vector **u**:

$$\mathbf{u}'\mathbf{u} = [u_1 \ u_2 \ \dots \ u_n] \begin{bmatrix} u_1 \\ u_2 \\ \dots \\ u_n \end{bmatrix} = u_1^2 + u_2^2 + \dots + u_n^2 = \sum_{i=1}^n u_i^2$$

Si definimos una función **z** como función de los elementos de una matriz podemos plantearla como:

$$z = \mathbf{x}'\mathbf{A}$$

Ahora si queremos obtener una suma ponderada de cuadrados partimos de una función como:

$$z = \mathbf{x}' \mathbf{A} \mathbf{x}$$

En donde:  $\mathbf{A}$  es una matriz diagonal cuadrada de dimensiones  $n \times n$   
 $\mathbf{x}$  es un vector columna de  $n$  elementos

$$\mathbf{x}' \mathbf{A} \mathbf{x} = [x_1 \ x_2 \ \dots \ x_n] \begin{bmatrix} a_{11} & 0 & \dots & 0 \\ 0 & a_{22} & \dots & 0 \\ \dots & \dots & \dots & \dots \\ 0 & 0 & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = a_{11}x_1^2 + a_{22}x_2^2 + a_{33}x_3^2 + \dots + a_{nn}x_n^2$$

Si la matriz  $\mathbf{A}$  no es diagonal y esta dada por:

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix}$$

La función cuadrática  $z = \mathbf{x}' \mathbf{A} \mathbf{x}$  será:

$$\mathbf{x}' \mathbf{A} \mathbf{x} = [x_1 \ x_2 \ \dots \ x_n] \begin{bmatrix} a_{11} & a_{12} & \dots & a_{1n} \\ a_{21} & a_{22} & \dots & a_{2n} \\ \dots & \dots & \dots & \dots \\ a_{n1} & a_{n2} & \dots & a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} =$$

$$a_{11}x_1^2 + a_{12}x_1x_2 + a_{13}x_1x_3 + \dots + a_{1n}x_1x_n +$$

$$a_{22}x_2^2 + a_{21}x_2x_1 + a_{23}x_2x_3 + \dots + a_{2n}x_2x_n +$$

$$\dots + \dots + \dots + \dots + \dots +$$

$$a_{nn}x_n^2 + a_{n1}x_nx_1 + a_{n3}x_nx_3 + \dots + a_{nn}x_nx_n +$$

Si la matriz  $\mathbf{A}$  es simétrica el resultado será:

$$= a_{11}x_1^2 + 2a_{12}x_1x_2 + 2a_{13}x_1x_3 + \dots + 2a_{1n}x_1x_n$$

$$+ a_{22}x_2^2 + 2a_{23}x_2x_3 + \dots + 2a_{2n}x_2x_n$$

....

.....

$$\dots + a_{nn}x_n^2$$

Las funciones de un vector o una matriz consideradas antes nos permiten definir la derivada vectorial y las siguientes reglas de derivación:

Para una función  $z = \mathbf{x}'\mathbf{c}$  en donde  $\mathbf{x}$  es un vector columna de dimensión  $n \times 1$  y  $\mathbf{c}$  es un vector columna  $n \times 1$ , la derivada vectorial de  $z$  con respecto a  $\mathbf{x}$  es:

$$dz/d\mathbf{x} = \mathbf{c}$$

Para una función  $z = \mathbf{x}'\mathbf{A}\mathbf{x}$  en donde  $\mathbf{A}$  es una matriz cuadrada  $n \times n$ :

$$dz/d\mathbf{x} = (\mathbf{A} + \mathbf{A}')\mathbf{x}$$

Para una función  $z = \mathbf{x}'\mathbf{A}\mathbf{x}$  en donde  $\mathbf{A}$  es una matriz cuadrada y simétrica:

$$dz/d\mathbf{x} = 2\mathbf{A}\mathbf{x}$$

Por ejemplo si definimos una matriz  $\mathbf{A}$  cuadrada y simétrica de  $2 \times 2$ :

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$$

De acuerdo a lo que vimos en las formas cuadráticas el producto:

$$\mathbf{x}'\mathbf{A}\mathbf{x} = a_{11}x_1^2 + 2a_{12}x_1 x_2 + a_{22}x_2^2$$

Construimos el vector de derivadas:

$$dz/d\mathbf{x} = \begin{bmatrix} \partial z / \partial x_1 \\ \partial z / \partial x_2 \end{bmatrix} = \begin{bmatrix} 2a_{11}x_1 + 2a_{12}x_2 \\ 2a_{12}x_1 + 2a_{22}x_2^2 \end{bmatrix} = 2 \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = 2\mathbf{A}\mathbf{x}$$

El rango de una matriz no nula  $A$  es igual a  $k$  si el determinante de al menos uno de sus menores cuadrados de orden  $k$  es distinto de cero, siendo nulos los correspondientes a todos los menores cuadrados de orden  $(k+1)$  si es que existen.

Veamos un ejemplo, sea:

$$\mathbf{A} = \begin{bmatrix} 1 & 2 & 3 \\ 2 & 3 & 4 \\ 3 & 5 & 7 \end{bmatrix} \quad \mathbf{M} = \begin{vmatrix} 1 & 2 \\ 2 & 3 \end{vmatrix} = -1 \neq 0$$

$$|\mathbf{A}| = 1(+1) - 2(14-12) + 3(10-9) \\ = 1-4+3 = 0$$

El determinante de la matriz  $\mathbf{A}$  es igual a cero, por lo cual la matriz  $\mathbf{A}$  no es de rango completo igual a tres. Si tomamos un menor cualquiera de dimensiones 2x2, por ejemplo la submatriz  $\mathbf{M}$ , encontramos que su determinante es diferente de cero, por lo tanto el rango de la matriz  $\mathbf{A}$  es igual a dos.

Una matriz  $\mathbf{A}$  de orden  $n$  es regular si su rango  $k=n$ , es decir si su determinante es  $|\mathbf{A}| \neq 0$  en caso contrario se llama singular.

Algunas propiedades interesantes de las matrices se obtienen a través de sus valores y vectores característicos, por ello veremos ahora como se obtienen.

Una matriz  $\mathbf{A} = [a_{ij}]$  ( $i, j = 1, 2, \dots, n$ ) puede ser transformada, utilizando un vector  $\mathbf{x}$  que se convierte mediante esa transformación en el vector  $\lambda\mathbf{x}$ , tal que :

$$\mathbf{Ax} = \lambda\mathbf{x}$$

A esta transformación se le conoce como ecuación característica y la podemos factorizar como:

$$\mathbf{Ax} = \lambda \mathbf{x}$$

$$\lambda \mathbf{x} - \mathbf{Ax} = (\lambda \mathbf{I} - \mathbf{A}) \mathbf{x}$$

Lo cual involucra las siguientes matrices:

$$(\lambda \mathbf{I} - \mathbf{A}) \mathbf{x} = \begin{bmatrix} \lambda - a_{11} & -a_{12} & \dots & -a_{1n} \\ -a_{21} & \lambda - a_{22} & \dots & -a_{2n} \\ \dots & \dots & \dots & \dots \\ -a_{n1} & -a_{n2} & \dots & \lambda - a_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix} = \mathbf{0}$$

Para lograr hacer la transformación es necesario resolver la ecuación característica para  $\lambda$  y  $\mathbf{x}$ . Esto lo hacemos si tomamos el determinante  $|\lambda \mathbf{I} - \mathbf{A}|$ , que es un polinomio de grado  $n$  y que se le conoce como polinomio característico de la matriz  $\mathbf{A}$ , en donde las raíces del polinomio son las raíces características  $\lambda$ .

Las raíces se llaman eigenvalores y los vectores correspondientes son eigenvectores.

Veamos un ejemplo, consideremos una matriz cuadrada y simétrica  $\mathbf{A}$ :

$$\mathbf{A} = \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix}$$

Construimos la ecuación característica:

$$(\lambda\mathbf{I} - \mathbf{A}) = \begin{bmatrix} \lambda-4 & -2 \\ -2 & \lambda-1 \end{bmatrix} = \mathbf{0}$$

Obtenemos el polinomio característico:

$$\begin{aligned} |\lambda\mathbf{I} - \mathbf{A}| &= (\lambda-4)(\lambda-1) - 4 = 0 \\ &= \lambda^2 - \lambda - 4\lambda + 4 - 4 \\ &= \lambda^2 - 5\lambda \end{aligned}$$

Podemos factorizar:

$$|\lambda\mathbf{I} - \mathbf{A}| = \lambda(\lambda-5) = 0$$

Las raíces que satisfacen la ecuación son:

$$\lambda_1 = 0$$

$$\lambda_2 = 5$$

Sustituimos cada una de las raíces en la ecuación característica para resolver en términos de  $\mathbf{x}$ .

Para  $\lambda_1 = 0$  el vector característico es:

$$\begin{bmatrix} 0 & -4 \\ -2 & 0 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

$$\begin{bmatrix} -4 & -2 \\ -2 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Que es el sistema de ecuaciones:

$$-4x_1 - 2x_2 = 0$$

$$-2x_1 - x_2 = 0$$

Despejando para las  $x$ 's:

$$x_2 = -4/2 x_1 = -2x_1$$

$$x_1 = -1/2x_2$$

Resulta claro que cualquier valor para  $x_1$  y  $x_2$  satisface las ecuaciones, así que debemos normalizarlas para obtener una solución única. La normalización es:

$$x_1^2 + x_2^2 = 1 \quad \text{y matricialmente } \mathbf{x}'\mathbf{x} = 1$$

$$x_1^2 + (-2x_1)^2 = 1$$

$$x_1^2 + 4x_1^2 = 1$$

$$x_1^2 = 1/5$$

$$x_1 = 1/5^{1/2}$$

$$x_2 = -2/5^{1/2}$$

Por tanto para  $\lambda = 0$  el eigen vector es:

$$\mathbf{x} = \begin{bmatrix} 1/5^{1/2} \\ -2/5^{1/2} \end{bmatrix}$$

Seguimos el mismo procedimiento para  $\lambda = 5$

$$(\lambda\mathbf{I} - \mathbf{A}) = \begin{bmatrix} 5-4 & -2 \\ -2 & 5-1 \end{bmatrix}$$

Sustituyendo en la ecuación característica:

$$\begin{bmatrix} 1 & -2 \\ -2 & 4 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

Tenemos un sistema de dos ecuaciones:

$$x_1 - 2x_2 = 0$$

$$-2x_1 + 4x_2 = 0$$

Despejando :

$$x_1 = 2x_2$$

$$x_2 = 1/2x_1$$

Cualquier par de valores satisface las ecuaciones, así que debemos normalizar

$$x_1^2 + x_2^2 = 1 \quad \text{y matricialmente } \mathbf{x}'\mathbf{x} = 1$$

$$(2x_2^2) + x_2^2 = 1$$

$$4x_2^2 + x_2^2 = 1$$

$$5x_2^2 = 1$$

$$x_2 = 1/(5)^{1/2}$$

Entonces:

$$x_1 = 2/(5)^{1/2}$$

El vector característico es:

$$\mathbf{x} = \begin{bmatrix} 2/(5)^{1/2} \\ 1/(5)^{1/2} \end{bmatrix}$$

Se cumple para los eigen vectores de una matriz simétrica que son ortogonales:

Para cada  $i \neq j$   $x_i' x_j = 0$

De acuerdo con nuestro ejemplo:

$$x_i' x_j = [1/(5)^{1/2} \quad -2/(5)^{1/2}] \begin{bmatrix} 2/(5)^{1/2} \\ 1/(5)^{1/2} \end{bmatrix} = 2/5 - 2/5 = 0$$

Con los vectores característicos se puede formar una matriz  $\mathbf{T}$  de transformación:

$$\mathbf{T} = [T_1, T_2, \dots T_k]$$

En el ejemplo:

$$\mathbf{T} = \begin{bmatrix} 1/(5)^{1/2} & 2/(5)^{1/2} \\ -2/(5)^{1/2} & 1/(5)^{1/2} \end{bmatrix}$$

La matriz de transformación nos permite diagonalizar a la matriz **A** en una matriz cuyos elementos de la diagonal principal son los valores característicos de **A**.

Si premultiplicamos **A** por **T'** y la posmultiplicamos por **T** podemos diagonalizarla:

$$\mathbf{T}'\mathbf{A}\mathbf{T} = \Lambda$$

En el ejemplo:

$$\begin{aligned} \Lambda &= \begin{bmatrix} 1/(5)^{1/2} & -2/(5)^{1/2} \\ 2/(5)^{1/2} & 1/(5)^{1/2} \end{bmatrix} \begin{bmatrix} 4 & 2 \\ 2 & 1 \end{bmatrix} \begin{bmatrix} 1/(5)^{1/2} & 2/(5)^{1/2} \\ -2/(5)^{1/2} & 1/(5)^{1/2} \end{bmatrix} = \\ &= \begin{bmatrix} 0 & 0 \\ 0 & 25/5 \end{bmatrix} = \begin{bmatrix} 0 & 0 \\ 0 & 5 \end{bmatrix} = \Lambda \end{aligned}$$

En R podemos generar una nueva matriz de 2x2 con los valores de la matriz A, para no confundirla con la matriz A que hemos empleado antes ahora le llamaremos AE. La instrucción en R para generar la matriz AE es la siguiente:

```
> list(AE)
```

```
[[1]]
```

```
[,1] [,2]
```

```
[1,] 4 2
```

```
[2,] 2 1
```

Los valores y vectores característicos los obtenemos con la instrucción:

```
> eigen(AE)
```

```
$values
```

```
[1] 5 0
```

```
$vectors
```

```
[,1] [,2]
```

```
[1,] -0.8944272 0.4472136
```

```
[2,] -0.4472136 -0.8944272
```

Con base en los valores característicos se puede obtener:

1) Rango ( $\mathbf{A}$ ) = Rango ( $\Lambda$ )

Y el rango de  $\Lambda$  es el número de valores diferentes de cero en su diagonal principal; en este caso el rango ( $\mathbf{A}$ ) = 1

2) El determinante de una matriz es igual al producto de sus raíces características, en nuestro ejemplo obtuvimos dos raíces características cuyo producto es igual a cero.

3) Los valores característicos también nos permiten determinar el signo de una matriz:

Si todas las raíces  $\lambda$  son positivas,  $\mathbf{A}$  es positiva definida.

Si son negativas es negativa definida

Si algunas son cero y las demás negativas es seminegativa definida

Si algunas son cero y las demás son positivas es semipositiva definida.

Si tiene raíces positivas y negativas es indefinida.

## REFERENCIAS

Chiang, Alpha C. (1987). *Métodos Fundamentales de Economía Matemática*. Mc Graw-Hill.

Kohler, Heinz (1996). *Estadística para Negocios y Economía*. Ed. CECSA.

Weber, Jean E. (1999). *Matemáticas para Administración y Economía*. Ed. Harla.

Hatekar, R. Neeraj (2010) Principles of econometrics, an introduction (using R), Sage Texts.

Anderson, David R. et. al. (1999) Statistics for business and economics, International Thompson Publishing.

Everitt,S. Brian y Torsten Hothorn, A handbook of statistical analysis using R, Chapman / Hall/CRC, 2006.

Quintana Romero, Luis y Miguel Ángel Mendoza, Econometría básica, Plaza y Valdés, 2008.

## **ARCHIVOS DE DATOS ASOCIADO AL CAPÍTULO**

PIB\_estados.txt

PIB\_estados2.txt

## **MATERIAL DE APRENDIZAJE EN LÍNEA**

Teória\_Cap16

Práctica\_Cap16

VideoPráctica\_Cap16

VideoTeoría\_Cap16

## **LISTA DE AUTORES**

Javier Galán Figueroa, profesor de la Licenciatura en Economía de la Facultad de Estudios Superiores Acatlán (FESA) UNAM, email: [javier.galanf@gmail.com](mailto:javier.galanf@gmail.com).

Jorge Feregrino, profesor e investigador en el Tecnológico de Estudios Superiores de Coacalco (TESCo) y del Departamento de Economía de la Facultad de Estudios Superiores Acatlán (FESA), UNAM, email: [jorferegrino@yahoo.com](mailto:jorferegrino@yahoo.com).

Lucía A. Ruiz Galindo, profesora e investigadora en el Departamento de Economía de la Universidad Autónoma Metropolitana Azcapotzalco (UAM-A), email: [laruizq@prodigy.net.mx](mailto:laruizq@prodigy.net.mx).

Luis Quintana Romero, profesor de la Licenciatura y el Posgrado en Economía, adscrito al programa de Investigación de la Facultad de Estudios Superiores Acatlán (FESA) UNAM, email: [luquinta@apolo.acatlan.unam.mx](mailto:luquinta@apolo.acatlan.unam.mx).

Miguel Ángel Mendoza González, profesor del Posgrado en Economía, Facultad de Economía de la UNAM, email: [mendozag@unam.mx](mailto:mendozag@unam.mx).

Roldán Andrés, investigador del Centro de Investigación en Geografía y Geomática "Ing. Jorge L. Tamayo", A.C. (Centro Geo) y profesor de la Facultad de Estudios Superiores Acatlán (FESA), UNAM, email: [roldandres@yahoo.com.mx](mailto:roldandres@yahoo.com.mx).

*Econometría aplicada utilizando R*

Es un libro electrónico disponible libremente en el sitio:

<http://saree.com.mx/econometriaR/>