00321
82

UNIVERSIDAD NACIONAL AUTONOMA DE MEXICO

FACULTAD DE CIENCIAS

TRATAMIENTO DE LOS PRINCIPALES PROBLEMAS EN EL ANALISIS DE REGRESION MEDIANTE EL PAQUETE ESTADISTICO ECONOMETRIC VIEWS

T E S I S

QUE PARA OBTENER EL TITULO DE

A C T U A R I O

PRESENTA

GERARDO REYES RUIZ

FACULTAD DE CIENCIAS
UNAM

DIRECTOR DE TESIS:
ACT. FRANCISCO SANCHEZ VILLARREAL

PAGINACION DISCONTINUA

DRA. MARÍA DE LOURDES ESTEVA PERALTA
Jefa de la División de Estudios Profesionales de la
Facultad de Ciencias
Presente

Comunicamos a usted que hemos revisado el trabajo escrito:

Tratamiento de los principales problemas en el análisis de regresión mediante el paquete estadístico Econometric Views

realizado por Gerardo Reyes Ruiz

con número de cuenta 8709294-3 , quien cubrió los créditos de la carrera de: Actuaría

Dicho trabajo cuenta con nuestro voto aprobatorio.

Atentamente

Director de Tesis Propietario	Act. Francisco Sánchez Villarreal
Propietario	M. en A.P. María del Pilar Alonso Reyes
Propietario	Mat. Margarita Elvira Chávez Cano
Suplente	M. en E. Miguel Andrés Cruz Galindo
Suplente	Act. Susana Barrera Ocampo

Consejo Departamental de Matemáticas

M. en C. José Antonio [illegible]

A mi familia y muy en especial a mis padres.

A mi hija Diana Itzel porque a partir del *22/05/2000* ella ha sido el principal motivo por el cual sigo vivo.

A Betán Bautista Guillermo† (donde quiera que esté).

Julio del 2003

INDICE

INTRODUCCION

En 1926 el profesor noruego Ragnar Frisch introdujo el término econometría para denominar a una nueva disciplina que surgía en el campo de la ciencia económica y que estaba constituida por estudios en los que se combinaban la Teoría Económica, la Estadística y las Matemáticas.

La etapa pre-econométrica puede considerarse como aquella en la que la Estadística y la Econometría evolucionaron al sentar las bases que habrían de servir de soporte al nacimiento de la econometría. Esta etapa se inicia en el siglo XVII con William Petty, autor que ha sido considerado por muchos economistas como el padre de la econometría y que además fue un gran impulsor de los primeros trabajos estadísticos. Por lo que respecta a la evolución de la estadística son muchos los autores que se pueden citar ya que han sido muy numerosas e importantes las aportaciones al avance de la metodología estadística en esta época.

Algunas de las principales contribuciones se deben al matemático y físico alemán Karl Gauss, que es considerado como uno de los principales matemáticos de todos los tiempos. A este investigador se debe la distribución normal y la utilización por primera vez, en 1809, del método de mínimos cuadrados aplicado al análisis de la regresión. También es destacable la contribución del estadístico y químico inglés William S. Gosset, conocido con el pseudónimo de Student, a las contribuciones de Ronald A. Ficher, Markov, etc. Gracias a estas contribuciones los estadísticos y economistas pudieron comenzar a aplicar el análisis de regresión y la contrastación de hipótesis de los modelos económicos a comienzos del siglo XX. En 1912 el economista norteamericano Irving Fisher, de la universidad de Yale, trató de organizar a un grupo de investigadores para que estimularan el desarrollo de la teoría económica en su relación con la estadística y las matemáticas. Aunque el primer intento de organización fracasó, la idea de que las tres especialidades tenían que utilizarse en común persistió en los años siguientes. El trabajo de H. Moore "Economic Cycles: their Law and Cause", publicado en 1914, es considerado por muchos autores como el primer trabajo econométrico.

La etapa de aportaciones básicas se caracteriza por el estudio de los principales problemas econométricos, tales como la identificación y estimación de ecuaciones simultáneas, la autocorrelación de las perturbaciones y la multicolinealidad de las regresores. Koopmans, Aitken,

Haavelmo, Anderson, Rubin, Durbin, Watson, Cochrane y Orcutt, son algunos de los autores que destacan por sus aportaciones en esta etapa. En 1931 Alfred Cowles ofreció proporcionar fondos para fundar una revista y una organización de investigación para la Econometric Society. La revista creada por la iniciativa de Cowles fue *Econométrica*, cuyo primer número apareció en 1933 y desde entonces hasta la actualidad ha sido una de las principales fuentes de la información científica de los econometras.

En la década de 1960 se efectúan numerosas aplicaciones prácticas a los modelos de demanda, producción e inversión y se desarrollan los principales modelos macroeconométricos de enfoque Keynesiano derivados del importante impacto de la obra del gran econometra Lawrence R. Klein de la Universidad de Pensylvania, quien alcanzó el Premio Nobel de Economía en 1980. A partir del año 1970, a raíz de la publicación del libro de Box y Jenkins sobre la predicción univariada, la metodología ARIMA se incorpora de forma decidida a los modelos econométricos, especialmente en relación con la especificación de la parte no determinista del modelo, es decir la perturbación aleatoria, en modelos basados en datos temporales con periodicidad trimestral o mensual.

En las décadas de 1980 y 1990 merecen destacarse de forma especial las contribuciones metodológicas relativas a los contrastes tanto de especificación y evaluación de los modelos, los avances en la metodología de los modelos dinámicos, y los numerosos estudios de Econometría Aplicada sobre temas muy diversos, entre los que destacan los estudios de competitividad y los modelos sectoriales.

La herramienta fundamental de la econometría es sin lugar a dudas el análisis de regresión. Los orígenes de la regresión se remontan a 1886 con las investigaciones sobre el efecto que tiene la altura de los padres sobre la altura de sus hijos, realizada por Francis Garton.

El análisis de regresión no se debe confundir con el análisis de correlación, el cual consiste en medir el grado de relación lineal entre las variables, a diferencia del análisis de regresión que mide relaciones de causalidad. Por otro lado el análisis de correlación supone que todas las variables son aleatorias, mientras que cuando se hace regresión se supone que las variables explicativas son fijas en muestreos repetidos.

Para lograr determinar la ecuación de regresión que permita expresar la función $Y= f(X)+e$, existen varias técnicas matemáticas y estadísticas como

son, por mencionar algunas, la de MCO (Mínimos Cuadrados Ordinarios), MV (Máxima Verosimilitud) etc. de los cuales el más utilizado es el Método de Mínimos Cuadrados Ordinarios.

La búsqueda de relaciones causales es el principal objetivo científico. La econometría en su origen estuvo fundamentalmente orientada como una metodología que permitiese contrastar la veracidad de las relaciones causales establecidas por la teoría económica. La metodología econométrica permite estimar diversos modelos teóricos y contrastar cuáles de las variables explicativas incluidas tiene una incidencia significativa sobre la variable explicada.

De esta manera, gran parte de este trabajo está dedicado a revisar las técnicas, mediante el paquete *Econometric Views V3.1*, para determinar la eficacia y las relaciones causales de un modelo econométrico en particular, el modelo econométrico uniecuacional con *K* variables explicativas. Esta delimitación es especialmente relevante, ya que no se pretende, como principal objetivo, crear un nuevo modelo, sino que se analiza el grado de certeza de dicho modelo, para con ello detectar, enfrentar y corregir los problemas de multicolinealidad, heteroscedasticidad y correlación serial.

1 ¿Qué es *econometría*?

1.1 Naturaleza del análisis econométrico

La *econometría* ya es mayor de edad, el sólido cuerpo teórico producido por ella tiene una importancia similar al cúmulo de aplicaciones tanto en economía como en otras ciencias o técnicas. En la primera, tradicionalmente se aplica a la macroeconomía[1], pero en años recientes, también ha estado empleándose en la práctica de todos los demás campos de esta ciencia, entre los que se encuentran: la teoría monetaria, las finanzas gubernamentales, el comercio internacional, el mercado de trabajo y el desarrollo o crecimiento económico. La Historia, la Sociología, la Economía Industrial y la Ciencia Política también utilizan métodos econométricos; se realizan además aplicaciones en áreas tan diversas como educación, derecho, salud y transporte, en fin, es muy extenso el campo de trabajo de la *econometría*.

Ciertamente la *econometría* consiste en una aplicación de métodos estadísticos a un banco de datos generalmente llamado muestra, pero antes de dar una idea más o menos clara de lo que es la *econometría* es necesario saber lo siguiente: el sufijo "metría" significa *medición*, así pues, literalmente la *econometría* significa medir relaciones económicas.

Sin embargo, si bien es cierto que el fenómeno de la medición es una parte importante de la *econometría*, el campo de acción de esta disciplina es mucho más amplio, como se puede apreciar en las siguientes citas textuales:

La econometría puede definirse como la ciencia social en la cual se aplican las herramientas de la teoría económica, las matemáticas y la inferencia estadística, al análisis de los fenómenos económicos.

Arthur S. Goldberger

El arte del econometrista consiste en encontrar el conjunto de supuestos que sean suficientemente específicos y realistas, de tal manera que le permitan aprovechar de la mejor manera posible los datos que tiene a su disposición.

E. Malinvaud

[1]De *macro* y *economía*. Estudio de los sistemas económicos de una nación, región, etc., como un conjunto, empleando magnitudes colectivas o globales, como la renta nacional, las inversiones, exportaciones e importaciones, etc. Se usa en contraposición a microeconomía. Biblioteca de Consulta Microsoft® Encarta® 2003.

Aunque la estadística proporciona una buena parte de las herramientas utilizadas en cualquier estudio de la ciencia, a menudo el econometrista requiere métodos especiales en virtud del carácter *sui generis* de la mayor parte de las cifras económicas, debido a que éstas no son el resultado de un experimento controlado.

El econometrista, como el meteorólogo, generalmente depende de información que no se puede controlar directamente, por tanto, las cifras recolectadas por agencias públicas y privadas, son de características no experimentales. El econometrista toma estos datos como dados, hecho que genera graves problemas que se presentan normalmente en el campo de la matemática estadística. Además, la información puede contener errores de medición, situación que el econometrista puede ayudar a remediar desarrollando métodos especiales de análisis.

Cuando el término *econometría* se acuñó en la década de los treinta, cubría tanto el desarrollo de la teoría pura desde una perspectiva matemática como la estimación práctica de las relaciones económicas. En la actualidad significa principalmente esto último, es decir, al desarrollo matemático de la teoría económica ahora se le denomina *economía matemática*.

La *econometría* trata de la aplicación de la teoría económica, la matemática, y las técnicas estadísticas con el fin de probar hipótesis y estimar, así como pronosticar los fenómenos ya sean estos económicos o no. La *econometría* ha llegado a estar ampliamente identificada con el *análisis de regresión,* es más, se puede decir que la *econometría* ha ido evolucionando a la par del *análisis de regresión.*

De esta manera cabe resaltar que la naturaleza de la *econometría* descansa en la preocupación fundamental del hombre en cuanto a la medición cuantitativa, en el análisis de predicción de fenómenos (que en su gran mayoría son fenómenos económicos) y en la comprobación de las hipótesis relacionadas con los mismos. La misma necesidad de conocer el presente y predecir un futuro no muy lejano hace de la *econometría* una técnica indispensable en el contexto analítico.

1.2 Modelos vs. modelos econométricos

Una particularidad importante de la *econometría* y una fase esencial de cualquier estudio econométrico y no econométrico es la *especificación del modelo,* esto es la construcción y la elaboración de un modelo que represente de manera adecuada el fenómeno que va ha ser estudiado.

Por definición, un ***modelo***[2] es cualquier representación simbólica de un fenómeno real tal como un proceso o sistema real. El fenómeno real está representado por el modelo para explicarlo, predecirlo y controlarlo. A veces el sistema real se denomina *sistema del mundo real* para distinguirlo claramente del sistema modelo que lo representa.

La *modelación* -el arte de construir modelos- es una parte integral en la mayoría de las ciencias, ya sean físicas o sociales, debido a que los sistemas del mundo real bajo consideración, por lo común, son demasiado complejos. El sistema puede ser un electrón que se mueve en un acelerador de partículas, o precios que se colocan en diversos mercados, o bien, la determinación del ingreso nacional. En éstos y en muchos otros casos, los fenómenos del mundo real son tan complicados que únicamente pueden ser tratados en términos de una representación simplificada, esto es, vía un modelo.

Cualquier modelo constituye un compromiso entre la realidad y la maleabilidad. Debe ser una representación "razonable" del sistema del mundo real y por lo tanto "realista" al incorporar los principales elementos del fenómeno que se estudia. Además, debe ser maleable en el sentido que produzca ciertas introspecciones o conclusiones no obtenibles mediante observaciones directas del sistema del mundo real.

Por lo general, para lograr maleabilidad hay que efectuar diversos procesos de idealización, entre los que se incluyen la eliminación de influencias "extrañas" y la simplificación de procesos. Normalmente este proceso de idealización hace que el modelo sea menos "realista", no obstante, el proceso es necesario para asegurar que el sistema modelo pueda ser manipulado en términos razonables.

[2] Intriligator, Michael D. "Modelos econométricos, técnicas y aplicaciones". Editorial Fondo de Cultura Económica, México 1990, pág. 29

El balance adecuado entre realismo y maleabilidad es la esencia de la "buena" modelación. Un "buen" modelo es realista y maleable. Especifica las interrelaciones entre las partes de un sistema en una forma suficientemente detallada y explicada para asegurar que el estudio del modelo conduzca a introspecciones respecto al sistema del mundo real. Al mismo tiempo las especifica en una forma suficientemente simplificada y maleable para asegurar que el modelo pueda ser fácilmente analizado y puedan extraerse conclusiones relacionadas con el sistema del mundo real.

Un tipo de modelo "malo" es aquél altamente realista, pero tan complicado que se vuelve no maleable; en este caso, no hay razón para construirlo. El segundo tipo de modelo "malo" se va al otro extremo: es altamente maleable pero tan idealizado que es irreal al no tomar en cuenta importantes componentes del sistema del mundo real. En este caso, el proceso de idealización ha sido llevado demasiado lejos: influencias que de hecho son importantes fueron excluidas por los supuestos, y/o los procesos del mundo real involucran mayores complejidades de las que fueron postuladas por el modelo. Este extremo puede ser muy peligroso porque las conclusiones alcanzadas a través del modelo pueden o no ser relevantes al sistema del mundo real; el problema radica en que nunca se sabe con antelación si las conclusiones son o no relevantes. En este caso es válido que "un poco de conocimiento es peligroso".

En la medida en que es imposible establecer en forma precisa cómo construir un buen modelo, la modelación es en parte arte y en parte ciencia. Seguir ciertos preceptos generales y conocer intentos previos para modelar un fenómeno son reglas útiles, pero se requiere, ante todo, experiencia para llegar a ser un buen modelador.

Como regla general, los primeros modelos de un fenómeno son muy simples; subrayan la maleabilidad para no abordar la realidad con gran detalle. El caso extremo es la llamada "caja negra" de la figura 1.1, donde no se hace intento alguno por reproducir la realidad. En este caso el modelo sólo hace referencia a los insumos y a los productos del sistema sin considerar al propio sistema. Una descripción caja negra de un aparato de televisión, por ejemplo, tan sólo identificaría los insumos de electricidad y las señales de control emitidas por el operador y los productos en las señales de audio y video. Sólo haría referencia a los insumos y a los productos sin intentar analizar en cómo se relacionan o interactúan ambos.

El proceso de modelación usualmente implica comenzar con una caja negra y luego elaborar lo que está "adentro" de la caja. El modelo inicial o modelo descriptivo, es un modelo simple de caja negra, que sólo maneja insumos y productos. Al rastrear los insumos hacia adelante y los productos hacia atrás, conduce a modelos más elaborados que eventualmente dan lugar a un *modelo analítico*, es decir, un modelo de caja blanca, que trata explícitamente a todas las interconexiones entre los insumos y productos. Un modelo "caja blanca" de un aparato de televisión, por ejemplo, podría consistir en un diagrama complejo de circuitos. El proceso de modelaje generalmente implica un intento continuo por formular más y más modelos analíticos, que sean capaces de analizar más y más diversas interconexiones del sistema del mundo real. Un ejemplo es el desarrollo de modelos de la macroeconomía, comenzando con modelos simples y alcanzando, eventualmente, modelos macroeconométricos altamente detallados.

Figura 1.1 Caja negra

Existen muchos tipos de modelos en cada uno de los campos en los cuales se han aplicado modelos. Entre los tipos más importantes están los modelos *verbales/lógicos*, *los físicos*, *los geométricos* y los *algebraicos*, que implican formas alternativas de representación de un modelo.

Tal vez el tipo de modelo más sencillo y el que usualmente se utiliza primero en cualquier campo de investigación es el *modelo verbal/lógico.* Este enfoque emplea analogías verbales, tales como la metáfora y el símil, el modelo resultante a menudo se denomina un paradigma. Con frecuencia, tales modelos pueden tratar al sistema "como si" fuera, en cierto sentido, propositivo. Así, en la Física, el "principio del menor esfuerzo" establece que una partícula en movimiento actúa como si estuviera minimizando la energía que requiere para moverse. Definitivamente, éste es un modelo en el cual el sistema del mundo real, la partícula en movimiento, está representado en este caso, como una entidad propositiva.

Un segundo modelo es el *modelo físico*. En ciertos casos el sistema del mundo real es físico y puede obtenerse un modelo mediante un ajuste a escala apropiado, hacia arriba o hacia abajo. Así, es común que un alerón para un nuevo avión sea probado por medio de la construcción en pequeña escala y a través de un túnel de viento. Otro ejemplo es el modelo físico de un astrónomo acerca del Sistema Solar. Para considerar los ajustes a escala en la dirección opuesta, los físicos a menudo utilizan modelos físicos del átomo, señalando los neutrones y protones en el núcleo y los electrones en órbita alrededor de él. De modo similar, los biólogos moleculares utilizan modelos físicos tales como los de una molécula de proteína o la molécula de ADN. Estas son versiones de escala hacia arriba del sistema del mundo real en estudio y ciertamente son mucho más manipulables que las entidades reales en consideración.

De enorme importancia para el desarrollo de la teoría en general ha sido el *modelo geométrico,* el cual representa geométricamente las relaciones existentes del fenómeno en estudio. Un modelo geométrico utiliza un diagrama para indicar las interrelaciones entre variables, más que nada este tipo de modelo lo que hace es esquematizar gráficamente el comportamiento del fenómeno, así como la interacción de las "variables" que lo conforman.

El *modelo algebraico*, para propósitos econométricos, es el tipo de modelo más importante; representa al sistema del mundo real a través de una ecuación o un sistema de ecuaciones. Este modelo y los otros tipos a menudo son estudiados utilizando una computadora. Este enfoque en realidad puede involucrar algunos aspectos de todos los tipos de modelos aquí discutidos, incluyendo aseveraciones verbales/lógicas, analogías físicas, tratamientos gráficos y expresiones algebraicas. En general, tal enfoque se emplea cuando los fenómenos se hacen tan complejos y tan no maleables que no pueden ser tratados (es decir, resueltos) analíticamente. Se recurre entonces, a menudo, a la *simulación* del comportamiento del sistema en el modelo, bajo diferentes condiciones o supuestos, utilizando una computadora.

Sin embargo, hay un sinnúmero de ventajas de la representación algebraica de un modelo sobre la geometría, una de ellas, y que tal vez sea la más importante, es la facilidad de manipulación; otra ventaja del modelo algebraico sobre el geométrico es la facilidad de añadir nuevas variables y ecuaciones, así pues, por su propia naturaleza, la geometría está confinada a sólo dos o tres dimensiones. Por su parte, el álgebra no

está restringida de esa forma y por lo tanto, los modelos algebraicos pueden ser ampliados, desagregados y generalizados de muchas maneras. Por supuesto, hay una amplia gama de formas funcionales disponibles para un modelo algebraico; y la elección de una en particular depende de la aceptabilidad teórica, la plausibilidad, la facilidad de estimación, la bondad del ajuste, la capacidad de predicción, etcétera.

Ya en este contexto, un ***modelo econométrico*** se refiere a un tipo especial de modelo algebraico *estocástico,* esto es, que incluye una o más *variables aleatorias*. Representa un sistema a través de un conjunto de relaciones estocásticas entre las variables del sistema.

Pero, ¿qué es un modelo econométrico? Un modelo econométrico es aquél que desarrolla y ofrece una explicación de la teoría relevante para el sistema considerado y es la forma más conveniente para sintetizar esta teoría, para hacer mediciones, prácticas y pruebas.

En el contexto econométrico se estudian tres clases generales de modelos, que se pueden construir con un diferente grado de complejidad y de explicación estructural. Cada uno de estos modelos supone un nivel de comprensión diferente sobre los procesos reales que se tratan de representar mediante el modelo. Las tres clases de modelos son las siguientes:

a) ***Modelos de series de tiempo.*** En esta clase de modelos se supone que no se sabe nada sobre las relaciones causales que en la realidad afectan a la variable que se trata de predecir. En cambio, se examina el comportamiento de una serie de tiempo en el pasado para inferir cuál será su comportamiento en el futuro. El método de las series de tiempo para la obtención de una predicción puede implicar la utilización de un modelo determinístico simple, como la extrapolación lineal o la utilización de un modelo estocástico complejo para predicción adaptativa.

Un ejemplo de la utilización de un análisis de series de tiempo sería la extrapolación de una tendencia pretérita para predecir el crecimiento de la población. Otro ejemplo sería el desarrollo de un modelo lineal estocástico complejo para predecir el número de viajeros de una compañía de aviación. Se han desarrollado modelos de este tipo para predecir la demanda de capacidad de las líneas aéreas, la demanda estacional de teléfonos, el movimiento de los tipos de interés a corto plazo y otras

variables económicas. Los modelos de series de tiempo resultan particularmente útiles cuando se dispone de escasos conocimientos sobre el proceso que se trata de predecir. La estructura limitada de los modelos de series de tiempo hace que estos sean fiables sólo a corto plazo, pero no obstante esta limitación, resultan útiles.

b) ***Modelos de regresión***. Es aquél que consta de una sola ecuación. En esta clase de modelos, la variable objeto de estudio se explica mediante una única función (lineal o no lineal) de una o varias variables explicativas. Un modelo lineal es aquel cuya función describe una línea recta, mientras que los modelos no lineales son aquellos cuya función no describe una línea recta. Se dice que un modelo es lineal en los parámetros cuando éstos estén elevados a la potencia 1.

De esta manera, la ecuación será, en muchos casos, dependiente del tiempo (es decir, el tiempo aparecerá en el modelo de forma explícita), de forma que será posible predecir la respuesta de la variable en estudio a cambios experimentados por una o más de las variables explicativas a lo largo del tiempo.

Un modelo *lineal simple* sería como el que a continuación se muestra:

$$Y_t = \beta_0 + \beta_1 X_t + u_t \qquad \text{con } t \in [1,.....,n]$$

donde:

Y_t = variable explicada
X_t = variable explicativa
u_t = término de error

Mientras que un modelo *de regresión lineal múltiple* explicará una relación entre una variable explicada y *k* variables explicativas. Su forma general se define como:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + u_t \qquad \text{con } t \in [1,.....,n]$$

donde:

Y_t = variable explicada
$X_{1t}, X_{2t}, ..., X_{kt}$ = variables explicativas
u_t = término de error

Un ejemplo de modelo de regresión lineal múltiple sería una ecuación que relacione un tipo de interés en concreto, como por ejemplo, el tipo de interés de los bonos del tesoro nacional, con un conjunto de variables tales como la oferta monetaria, la tasa de inflación y la tasa de cambio del producto nacional bruto. Estos modelos de regresión se utilizan a menudo para predecir no sólo los cambios de los tipos de interés a corto y largo plazo, sino también para otras muchas variables económicas y empresariales.

La decisión de elaborar un modelo de serie de tiempo suele adoptarse en aquellos casos en que sabe muy poco o nada acerca de las variables que influyen sobre la variable en estudio, y cuando se dispone de un gran número de observaciones (de modo que resulte posible hacer cualquier tipo de inferencia) y cuando el objetivo principal es la predicción a corto plazo. Sin embargo, si se dispone de alguna información sobre el proceso en estudio, ya no resultará tan obvio, si se requiere para efectos de predecir, un modelo de serie de tiempo o un modelo de regresión. En algunos casos puede ser aconsejable construir ambos modelos y comparar sus ventajas y/o desventajas respectivas.

c) ***Modelos de simulación multiecuacionales***. Son aquéllos modelos que constan de dos o más ecuaciones. En esta clase de modelos, la variable objeto de estudio puede ser una función de varias variables explicativas, pero dichas variables se relacionan entre sí, así como con la variable en estudio, mediante un conjunto de ecuaciones. La elaboración de un modelo de simulación comienza con la especificación de un conjunto de relaciones individuales, cada una de las cuales se ajusta a los datos disponibles. La simulación es el proceso de resolución simultánea de dichas ecuaciones para un periodo de tiempo cualquiera.

Para ejemplificar un modelo de simulación multiecuacional considérese el siguiente sistema de ecuaciones simultáneas:

i. $P_t = \beta_0 + \beta_1 S_t + u_{1t}$

ii. $S_t = \beta_2 + \beta_3 Q_t + u_{2t}$

iii. $Q_t = \beta_4 + \beta_5 Y_t + u_{3t}$ para toda $t \in [1, \ldots\ldots, n]$

El sistema de ecuaciones está *completo* si hay tantas ecuaciones independientes como variables explicadas. De esta manera, el sistema determina conjuntamente los valores de las variables explicadas en función de las variables explicativas y los valores para el término de error.

Normalmente cada ecuación del sistema tiene un significado propio y representa o una ecuación de comportamiento o una relación bien definida. Dado que cada ecuación representa un aspecto específico de la estructura se llama una *ecuación estructural* y el sistema completo se denomina la *forma estructural* del modelo.

Un ejemplo de modelo de simulación multiecuacional sería, un modelo complejo de la industria textil de los Estados Unidos Mexicanos que contuviese ecuaciones explicativas de variables tales como la demanda de productos textiles, de producción textil, el empleo de mano de obra en la industria textil, las inversiones efectuadas en la industria y los precios de los productos textiles. Estas variables se relacionarían entre sí y con otras variables (como la renta nacional, el índice de precios de consumo, los tipos de interés, etc.) mediante un conjunto de ecuaciones lineales o no lineales. Dados unos supuestos sobre el futuro comportamiento de la renta nacional, los tipos de interés, etc., se podría simular este modelo y obtener una predicción para cada una de las variables.

Los modelos de simulación multiecuacional explican mucho sobre la estructura del proceso en estudio. No sólo se especifican las relaciones individuales, sino que el modelo tiene también en cuenta la interacción de todas estas relaciones. Así, un modelo de simulación de cinco ecuaciones contiene, en realidad, más información que la suma de cinco ecuaciones de regresión individuales. Dicho modelo, no sólo explica las cinco relaciones individuales, sino que describe también la estructura dinámica implicada por la operación simultánea de dichas relaciones.

La elección del tipo de modelo a desarrollar es difícil, e implica un compromiso entre los costos, el tiempo y el trabajo necesarios y la precisión que se desea tengan las predicciones. La construcción de un modelo de simulación multiecuacional puede requerir un elevado gasto de tiempo y de dinero, no sólo en términos de trabajo real, sino también en términos de tiempo de ordenador. Entre las ventajas resultantes de este esfuerzo podrían incluirse la mejor comprensión de las relaciones y de la estructura, así como la posibilidad de efectuar mejores predicciones. No obstante, en algunos casos estas ventajas no serán lo suficientemente

grandes como para compensar el elevado costo. La construcción de modelos multiecuacionales exige disponer de un amplio conocimiento sobre los procesos en estudio, y por esta causa puede ser extremadamente difícil su fabricación.

1.3 Taxonomía de los modelos econométricos

Un modelo econométrico es una construcción teórica-empírica que, como consecuencia, debe de cumplir tanto con los requisitos lógicos como con los empíricos, estos últimos se caracterizan por la generalidad y la validez, por lo tanto, tienen alcance limitado en el espacio y en el tiempo.

Así, un modelo econométrico se especifica por un conjunto de funciones o ecuaciones entre variables relevantes que concurren, en un mismo periodo de tiempo, a explicar un fenómeno de estudio específico.

Las ecuaciones de un modelo econométrico se llaman *estructurales* o *primarias*, por lo que el modelo será estructural o primario. Un modelo es pues una familia de estructuras.

Por ejemplo, si se tiene que:

i. $D_t = \beta_0 + \beta_1 P_t + u_{1t}$

ii. $S_t = \beta_2 + \beta_3 P_{t-1} + u_{2t}$

iii. $Q_t = D_t + S_t$ con $t \in [1, \ldots\ldots, n]$

entonces se referirá a un modelo primario. Pero si en cambio, se tiene una solución específica del modelo, por ejemplo:

1. $D_t = 100 + 3.7 P_t + e_{1t}$

2. $S_t = 7 + .9 P_{t-1} + e_{2t}$

3. $Q_t = D_t + S_t$ para toda $t \in [1, \ldots\ldots, n]$

entonces se estará hablando de una estructura entre *n* posibles. Por lo tanto, un modelo econométrico puede definirse también como un conjunto de estructuras.

Por consiguiente un modelo econométrico tiene los siguientes elementos constitutivos:

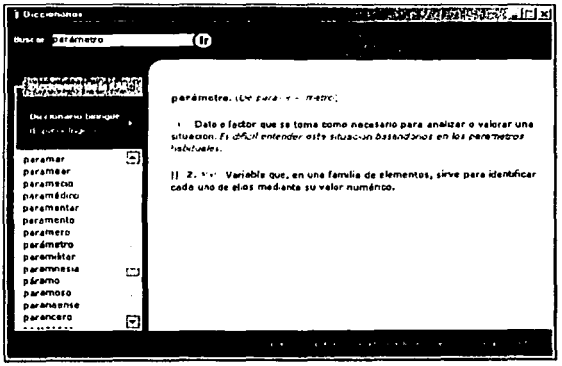
a) Ecuaciones.

b) Variables.

c) Parámetros.

Una *ecuación* o identidad es una relación que se verifica siempre, ya sea por su construcción lógica o por la definición que satisface. Toda ecuación es una relación matemática entre un conjunto de variables, que se satisface para determinados valores numéricos de ellas. Por una variable debe entenderse algo inestable, inconstante y mudable, pero en el contexto matemático una *variable* es una magnitud que puede tomar un valor cualquiera de los comprendidos en un conjunto. Sin embargo de este conjunto de valores, sólo interesan aquellos que tienen significado alguno para la relación matemática en cuestión.

En economía los parámetros (también conocidos como multiplicadores) se definen como los factores de producción correspondientes a cada variable explicativa y miden el efecto de las fluctuaciones de estas variables sobre la variable explicada. Ahora bien, según el Diccionario de la Real Academia Española[3] un parámetro es:



[3] Recurso educativo de la "Biblioteca de Consulta Microsoft Encarta 2003".

De esta manera, un parámetro puede verse como un punto de referencia, sin embargo, para entender mejor este concepto se necesita obligadamente entrar al campo de estudio del ***muestreo***[4].

Una vez dentro de este ámbito, se sabe que el objetivo de cualquier encuesta por muestreo es realizar inferencias acerca de una población de interés, partiendo de la información obtenida en una muestra de dicha población. Las inferencias en las encuestas por muestreo son usualmente dirigidas a la estimación de ciertas características numéricas de la población, tales como la media, el total o la varianza. Estas medidas descriptivas numéricas de la población se denominan *parámetros*. Un *estimador* es una función de variables aleatorias observables y quizás otras constantes conocidas, usado para estimar un *parámetro*.

1.3.1 Importancia y categorización de las variables en un modelo econométrico

Generalmente las variables teóricas incluidas en un modelo econométrico no siempre son observables, por lo cual habrá que elegir las series de datos que mejor se aproximen, o bien adecuar el modelo teórico a esta disponibilidad, redefiniendo las relaciones planteadas.

Debido a que toda ecuación es una relación matemática entre un conjunto de variables, y que se verifica solo para determinados valores numéricos de ellas, la clasificación de las variables de un modelo se hace necesaria por dos motivos:

a) para determinar si el sistema axiomático cumple con las condiciones de validez (consistencia e independencia), y

b) para llevar a cabo una selección óptima de los métodos de estimación de parámetros.

Las variables se clasifican según que el modelo sea *estructural* ó *de decisión*.

[4] Para mayor referencia consultar William G. Cochran "Técnicas de Muestreo" C.E.C.S.A. 1980, Segunda Edición.

1.3.1.1 Clasificación de las variables en modelos econométricos estructurales

Un *modelo econométrico estructural* es pues una familia de estructuras (ecuaciones), en estos modelos las variables son:

i) Endógenas o dependientes.
ii) Predeterminadas.
   - Exógenas.
   - Endógenas con retardo.

iii) Aleatorias o estocásticas.
iv) Expectativas.

Una terminología común utilizada en *econometría* para variables dependientes e independientes es la de ***endógenas*** y ***exógenas*** respectivamente. Variables endógenas son aquellas determinadas dentro del sistema econométrico, y exógenas las que vienen dadas desde fuera del sistema. En un sentido amplio, casi todas las variables son endógenas y las únicas exógenas en las que cabe pensar son el tiempo climatológico, los ciclones, etc. Sin embargo, en cualquier caso esto es un tema de aproximación. Por ejemplo, al estudiar la demanda de gasolina por familias, se puede considerar la cantidad demandada como variable endógena mientras que al precio y el ingreso como variables exógenas, arguyendo que las familias no tienen control sobre ellos. De la misma manera, para determinados propósitos puede considerarse el gasto público y los impuestos como variables exógenas. Sin embargo, a medida que se alargue el periodo de tiempo de las observaciones, estas variables pueden tratarse como variables endógenas.

Así pues, las ***variables endógenas*** son aquellas cuyos valores estimados van a ser determinados por la solución particular del sistema de ecuaciones del modelo mientras que las ***variables predeterminadas*** son aquellas cuyos valores no se obtienen por la solución del modelo sino que provienen de fuera del mismo, ellas contribuyen a explicar el comportamiento de las variables endógenas de un modelo sin ser explicadas por el modelo mismo. Dentro de ellas hay dos tipos: las ***exógenas*** incluyen variables económicas propiamente dichas y variables no económicas. Ambas son explicativas en un modelo dado pero no constituyen objeto de análisis y de explicación en dicho modelo.

El carácter de exógena o endógena de una variable en un modelo depende esencialmente del papel que va a desempeñar en el modelo, es decir, si va a explicar o a ser explicada. La inclusión de variables exógenas con significado económico se justifica por el tema de la investigación (sector, actividad, etc.) y por el periodo que se considera. Las variables exógenas no económicas no presentan este problema, ellas son por ejemplo, la lluvia en la producción agrícola.

Las *variables* ***endógenas con retardo*** intervienen como variables explicativas, ya que a pesar de ser endógenas en el periodo *t*, en el periodo *t-1* su valor es un hecho irreversible, es decir, se convierten en un dato. Las variables endógenas con retardo intervienen con mucha frecuencia en el análisis, sobretodo en la formulación de modelos dinámicos. La importancia se debe al efecto producido por los valores del pasado en los niveles actuales de las variables endógenas.

Las ***variables aleatorias*** *o* ***estocásticas*** son no observables y su introducción caracteriza a los modelos estocásticos o probabilísticos. Estos últimos siguen dominando el contenido de la teoría económica y la economía matemática mientras que los primeros son de uso habitual en *econometría*.

Las variables aleatorias cumplen con la misión de recoger un conjunto de causas que no se encuentran incorporadas explícitamente en un modelo y que se relacionan con:

a) omisión de variables explicativas.
b) errores de especificación del modelo, y
c) errores de medición de la variable endógena.

Toda variable aleatoria está asociada con una distribución de probabilidad, la cual debe hacerse explícita en la elaboración de pruebas de hipótesis.

Las ***variables expectativas*** tampoco son observables y su introducción exige el enunciado de un postulado adicional en el que se especifica su comportamiento en función de variables observables. Son variables expectativas el consumo esperado, el precio normal esperado, etc. Estas variables son comunes en modelos con retardos distribuidos.

1.3.1.2 Clasificación de variables en modelos econométricos de decisión

Los *modelos econométricos de decisión* cumplen una misión muy importante en la programación del crecimiento y desarrollo de un sector, una región o un país. En este tipo de modelos se fijan de antemano ciertos objetivos o metas que se desean alcanzar, por ejemplo: una tasa mínima de crecimiento del PIB, una tasa mínima de alfabetismo, generando una serie de políticas a seguir.

Por ello sus variables se clasifican como:

i) Endógenas.
   - Objetivo.
   - No objetivo.

ii) Exógenas.
   - Controlables.
     1. Instrumentales.
     2. No Instrumentales.
   - No controlables.

Las variables ***objetivo*** son aquellas que se desean o se dejan influenciar, es decir, son aquellas a las cuales se les fija un nivel a alcanzar o un comportamiento en el tiempo. Las variables ***no objetivo*** son aquellas por las que no se tiene interés alguno; así, por ejemplo los niveles de empleo y precios pueden ser variables objetivo.

Las variables exógenas ***controlables*** son aquellas sobre las cuales puede actuar directamente el autor de las decisiones. Por ejemplo, los impuestos es una variable controlable para el gobierno pero no para las empresas o los particulares. Estas variables se convierten en ***instrumentales*** cuando son seleccionadas como medio o instrumento de acción para el logro de los objetivos buscados.

Por ejemplo, supóngase que se quiere utilizar el siguiente modelo para alcanzar una tasa mínima de crecimiento del ingreso nacional:

$$Y_t = C_t + I_t + Z_t \qquad t \in [1,\ldots\ldots,n]$$

y además $C_t = \beta_0 + \beta_1 Y_t + u_t$

donde:

Y_t = ingreso nacional en el periodo t
C_t = consumo privado en el periodo t
I_t = inversión privada en el periodo t
Z_t = inversión pública en el periodo t

Entonces Y_t es una variable endógena objetivo, C_t es una variable endógena no objetivo, la variable exógena controlable e instrumental para el gobierno es Z_t (es decir, la inversión pública) mientras que I_t (la inversión privada) será una variable exógena no controlable por el poder público.

La figura 1.2 muestra esquemáticamente la clasificación de las variables en un modelo decisivo.

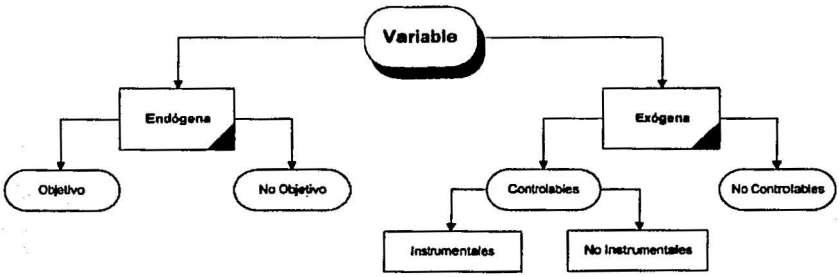


Figura 1.2 Variables en un modelo de decisión

1.4 Etapas para la construcción de un modelo econométrico

Como suele suceder, al tratar de explicar los pasos a seguir para llegar a entender una disciplina, frecuentemente se cae en el error de omitir un determinado número de éstos.

Sin embargo, el proceso que conlleva a una buena "especificación de un modelo econométrico" rigurosamente debe de cumplir con un mínimo de etapas, las cuales son las siguientes:

1) Formulación de teoría y hechos.
2) Planteamiento de la hipótesis acerca del fenómeno en estudio.
3) Especificación de modelos con medición y comprobación estadística.
4) Recolección de datos.
5) Estimación estadística.
6) Inferencia estadística al relacionar a la teoría económica con el análisis empírico.

Lo anterior no quiere decir que un modelo econométrico tenga o deba de cumplir con cada una de las etapas especificadas, de hecho pueden descartarse, aumentarse o incluso repetirse varias veces un cierto número de ellas.

En la figura 1.3 se sintetizan sucintamente los pasos esenciales que un estudio econométrico debe comprender.

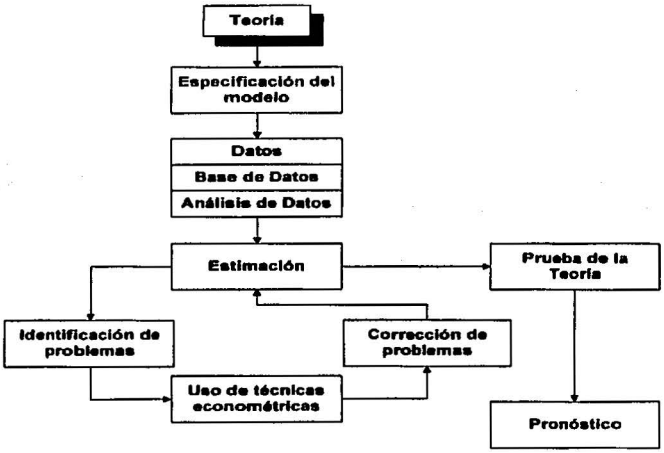


Figura 1.3 Etapas a seguir en la especificación de un modelo econométrico

Sin embargo, cabe mencionar que el objetivo principal de todo modelo econométrico es pues, realizar una "*economía*" del pensamiento sintetizando las características permanentes y relevantes de un fenómeno

en un conjunto completo, consistente y claro con el ánimo de conocer la realidad para actuar sobre y para ella.

1.4.1 Formulación de teoría y hechos

Es importante señalar o tener en cuenta que en la fabricación de cualquier modelo econométrico son dos los componentes fundamentales: *teoría y hechos*. En realidad, un logro muy importante de la *econometría* simplemente es el combinar estos dos elementos.

A la escuela orientada principalmente a la teoría, tan sólo le interesan implicaciones puramente deductivas de ciertos postulados sobre sistemas que contienen fenómenos ya sean económicos o de cualquier otra naturaleza.

Por el contrario, la escuela centrada en los hechos se preocupa exclusivamente por desarrollar y mejorar la información sobre el fenómeno en estudio. Sería realmente difícil el tratar de defender una de estas dos posiciones extremas. En lo que se refiere a la primera escuela, la teoría pura tiene poco contenido práctico intrínseco. Más aún, a menudo se desarrollan teorías contrarias y la manera correcta de elegir entre ellas se basa en la evidencia de que existen en forma de datos; es decir, los datos deben de ser los que guíen el desarrollo de la teoría.

Por lo que respecta a la segunda escuela, los hechos no "hablan por sí mismos" y para poder utilizarlos de manera eficaz deben ser interpretados en términos de una estructura subyacente, incorporada a una teoría. Así, la *econometría* se sirve tanto de teoría como de hechos y, mediante el uso correcto de técnicas estadísticas, logra fusionarlos para estimar relaciones ya sean económicas o no.

La *teoría* es uno de los elementos básicos en cualquier estudio econométrico, pero debe ser formulada de manera que pueda utilizarse. La forma más eficaz con fines econométricos (como se aprecia en la figura anterior) suele ser la de un modelo, en particular, un modelo econométrico. La elección de la teoría es el punto de partida, pues dentro de ese contexto es donde el modelo econométrico propuesto tratará de explicar lo mejor posible un fenómeno en particular.

De esta manera, una vez que la teoría se ha vertido en un modelo econométrico, lo que procede es aplicar técnicas econométricas para estimar las relaciones. Estas técnicas son extensiones de los métodos clásicos de la estadística y muy en particular de la inferencia estadística. El resultado final del proceso es un modelo econométrico estimado en el cual se obtienen ciertas magnitudes llamadas estimadores, los cuales deben contrastarse con los parámetros o los valores esperados de estos parámetros para decidir si la teoría representa una buena aproximación al fenómeno que se estudia.

1.4.2 Planteamiento de la hipótesis acerca del fenómeno en estudio

Con base en la teoría se establece el *planteamiento de hipótesis*, es decir, que es lo que se pretende o hacia donde se quiere llegar con el análisis que se llevará a cabo. La hipótesis debe ser clara y concisa, pues en ella descansa gran parte del análisis econométrico que se pretende realizar.

Para empezar, deben exponerse claramente los objetivos del estudio, así como la revisión de la teoría apropiada y de cualquier conocimiento de la forma en que las variables en cuestión se cree que están relacionadas en la realidad. Con respecto a la última de éstas cuestiones, si, por ejemplo, el comportamiento de una variable está determinado, bien sea total o sólo parcialmente, es entonces evidentemente importante saberlo.

Sin embargo, en la medida en que no sea posible para el estudioso el estar totalmente seguro acerca de algunas medidas, entonces dicho estudioso puede verse obligado a insinuar nada más que una sospecha que no está relacionada ni con la lógica formal ni con la información bien fundamentada. Por ésta razón, se señala frecuentemente que la formulación de hipótesis es tanto un arte como una ciencia y, en virtud de estas cualidades no es sorprendente que el resultado esté, generalmente, lejos de ser claro.

Una *hipótesis* es una suposición muy específica de algo posible o imposible para sacar con ello una consecuencia, es decir, una hipótesis conlleva al planteamiento de una conjetura[5]. Frecuentemente, se toma más de una suposición, pero no todo será probado. Esas suposiciones que no se pretenden poner a prueba son llamadas las *hipótesis conservadoras*. Ellas consisten de todas las suposiciones que se harán acerca del

[5] Diccionario de la Real Academia Española. Multimedia Encarta 2003.

fenómeno en estudio, por supuesto, nunca se asume que dichas suposiciones sean válidas. Las suposiciones que se cree son las más probables son llamadas las *hipótesis de prueba* y como su nombre lo indica, serán las hipótesis a probar. Usualmente las hipótesis de prueba consisten de una expresión de "que un cierto parámetro poblacional es igual a un determinado valor", o "que no excede o cae bajo un cierto valor". En estadística teórica este tipo de hipótesis es conocida como la *hipótesis nula* y se designa como H_0.

La idea de una hipótesis alternativa es completamente importante y requiere de una delicada elaboración. Cuando la hipótesis nula es una proposición a probar, por sí misma implica una contraproposición, de otra manera no tendrá sentido realizar la prueba. A dicha contraproposición se le conoce como la *hipótesis alternativa* y se designa como H_a.

Los criterios para rechazar o no rechazar la hipótesis nula sobre las bases del rango de evidencias no es una garantía de haber llegado a una conclusión correcta, es decir, cuando se lleva a cabo una prueba de hipótesis y la hipótesis nula es rechazada esto no implica que aceptemos a ciencia cierta la hipótesis alternativa, sino que simplemente se dice; no se acepta la hipótesis nula.

La Tabla 1.1 muestra esquemáticamente los tipos de error que se cometen al realizar una prueba de hipótesis.

	Estado del mundo	
Veredicto	***Ho es verdadera***	***Ha es verdadera***
Rechazar Ho	**Error Tipo I**	**Decisión correcta**
No rechazar Ho	**Decisión correcta**	**Error Tipo II**

Tabla 1.1 Tipos de errores en una prueba de hipótesis

Lo que realmente se pretende llevar a cabo en una prueba de hipótesis es minimizar el error tipo I, pues a éste se le asigna una determinada probabilidad de ocurrencia al principio de la prueba, mientras que al error tipo II no. En el contexto de prueba de hipótesis, frecuentemente al error tipo I se le conoce como el *nivel de significancia* de la prueba o "α", y al error tipo II simplemente como "β".

1.4.3 Especificación de modelos econométricos con medición y comprobación estadística.

Disponer de una herramienta informática es clave para esta tarea por la rapidez en la realización de cálculos. De otra manera resultaría impensable su ejecución manualmente bien sea por procedimientos algebraicos de resolución de matrices y determinantes o por cualquier otro método.

No obstante, no es menos cierto que el investigador deberá poseer los conocimientos y experiencia necesaria para definir los algoritmos específicos en función de los elementos que se estiman relevantes como configuradores del valor que interesa estudiar para cada caso. En este sentido cabe criticar la excesiva rigidez de algunas aplicaciones informáticas de tipo "prefabricado" donde el investigador no tiene libertad para definir la matriz de datos a operar ni el cómo operarla; y si bien estos paquetes de software intentan contemplar la máxima causística posible, con frecuencia no funcionan ante la variabilidad de situaciones y combinaciones de las mismas que se dan en la realidad. Los paquetes informáticos más flexibles implican, como contrapartida, la necesidad de conocimientos más profundos por parte del investigador, de técnicas estadísticas más elaboradas para definir el alcance de los procesos operacionales y la obtención de resultados coherentes.

Por esto, en la especificación de los modelos econométricos es preciso poner una especial atención acerca de la "bondad" del modelo concreto a adoptar. Para ello se recomiendan realizar previamente los siguientes pasos:

1. Definir las variables independientes del universo muestral que se estiman, a priori, como más significativas en la definición del modelo. Es decir, hacer una selección, en dado caso de que se cuenten con muchas variables explicativas, de las variables que se *consideren* representarán mejor o se involucren más con la variable a explicar.

2. Obtener conocimiento de la representatividad de la muestra utilizada para el cálculo, es decir, si la tendencia que refleja la ecuación de regresión es significativa o responde a una mera abstracción matemática. Se obtiene mediante el cálculo del llamado coeficiente de correlación lineal, generalmente designado por la letra "*r*" que mide el grado de relación lineal existente entre las variables en su conjunto, a partir de las

varianzas. Es una cantidad adimensional que oscila entre -1 y 1 y en la medida en que su valor absoluto se acerque a la unidad, implicará una relación lineal más perfecta entre las variables involucradas.

3. Calcular la representatividad de las variables independientes con respecto a la dependiente o si, por el contrario, éstas no lo son tanto y existen interferencias entre ellas que distorsionarían el cálculo final. De entrada es conveniente limitar el número de variables posibles máxime si el número de muestras representativas de que se dispone en la base de datos es bajo. En cualquier caso, ante la duda, es conveniente realizar un ejercicio de "tanteo" previo y posteriormente eliminar aquellas que por cálculo nos han parecido menos significativas.

Aunque éstos son tan sólo algunos aspectos que ayudan a la optimización y la construcción del modelo, no es posible olvidar que la mejor herramienta será siempre el sentido común del investigador y la sensibilidad para cimentar las bases de un buen modelo.

1.4.4 Recolección de datos

El otro aspecto básico en un estudio econométrico es un conjunto de *hechos,* término que designa los eventos en el mundo real que están relacionados con el fenómeno bajo investigación. Estos hechos conducen a un conjunto de *datos,* mismos que representan observaciones de hechos relevantes. Sin embargo, antes que nada, los datos deben ser recolectados para después ser seleccionados o "reconfigurados", en una diversidad de formas para adecuarlos al uso requerido por el estudio econométrico. Esta selección requiere distintas reconfiguraciones, tales como los ajustes estacionales o cíclicos, la extrapolación, la interpolación, la combinación de diferentes fuentes de información y, en general, el empleo de otras informaciones para moldear los datos. El resultado es un conjunto de *datos selectos* o *procesados.*

1.4.4.1 ¿Qué son los datos?

Un estudio econométrico implica el uso de datos para estimar un modelo estocástico lineal algebraico, a través de técnicas econométricas. La teoría pura puede tratar al fenómeno o sistema bajo estudio únicamente hasta cierto punto. Ese punto, es de modo clásico, el análisis estático comparativo de los signos de ciertas derivadas parciales, es decir, los

coeficientes de la forma reducida. Para proseguir más allá de este punto, en particular para estimar los valores de los coeficientes, tanto de la forma reducida, como de la forma estructural, se requiere de un conjunto relevante de datos acerca de todas las variables del modelo. Así, por ejemplo, el modelo micro(económico) prototipo requeriría datos sobre los precios, las cantidades el ingreso y la frecuencia con que llueve; en tanto que el modelo macro(económico) prototipo requeriría datos sobre el ingreso nacional, el consumo, la inversión y el gasto del gobierno.

Los datos relevantes en un estudio particular sintetizan los hechos concernientes al fenómeno en investigación. Estos hechos pueden ser de diferentes tipos y originarse de distintas fuentes; la teoría subyacente al fenómeno es usada para determinar la elección entre varias alternativas. Estas son fundamentalmente cuantitativas, cualitativas o una combinación de ambas. Cualquiera que sea su tipo, fuente o naturaleza, se expresan de un modo cuantitativo para llevar a cabo un estudio econométrico. El conjunto de todos esos hechos cuantitativamente expresados es la *información* o *datos* del estudio.

Para estimar un modelo econométrico se requieren datos acerca de todas las variables incluidas en el modelo. Son necesarios los valores adoptados por las variables endógenas, exógenas y, cuando es apropiado, por variables endógenas o exógenas desfasadas o rezagadas, para estimar los parámetros del modelo. En realidad el obstáculo más severo para realizar un estudio econométrico simplemente es la falta de información. Resulta relativamente sencillo construir modelos de todos tipos, tamaños, etc. Pueden ser manejados de diversas formas, sin embargo, encontrar datos para un modelo particular, ya es otra historia.

En general, la información no está disponible o no lo está en la forma deseada. Como resultado, diferentes "proxis" se usan a veces para ciertas variables del modelo. Un ejemplo es la tendencia en el tiempo usada como "proxi" para cambios en las preferencias o variaciones en la tecnología. Más aún, deben realizarse elecciones sobre problemas tales como expresar los datos en cantidades nominales o reales, en cantidades totales o per cápita, niveles (absolutos) o primeras diferencias o diferencias porcentuales, existencias o flujos, etc. Por último, los datos a veces tienen que ser "cocinados" de distintas maneras, como la eliminación de una tendencia y el uso de un ajuste estacional, para poder construir diversas series comparables entre ellas y concentrarse en ciertos fenómenos de interés.

1.4.4.2 Datos cuantitativos vs. cualitativos; variables *dummy*

Los datos pueden ser de distintas clases y por consiguiente pueden obtenerse diferenciaciones entre la variedad de datos disponibles. Aún cuando los datos por cuestiones de definición, son cuantitativos, en realidad representan hechos cuantitativos o cualitativos.

Los hechos cuantitativos, que ya están expresados como números, implican datos bajo la forma de estos números o alguna transformación apropiada de ellos.

Los hechos cualitativos, para los cuales no hay mediciones numéricas, también pueden expresarse en forma de datos. A menudo estos hechos cualitativos se refieren a situaciones en las que ocurren dos resultados mutuamente excluyentes. Así, algo ha ocurrido o no ha sucedido, una actitud o postura se ha adoptado o no, etc. Estos hechos cualitativos puede abarcar variables cualitativas (hombre o mujer, soltero o casado), cambios cualitativos en el tiempo o en el espacio (guerra o tiempos de paz, países industrializados o subdesarrollados), e incluso la agregación de hechos cuantitativos en sus correspondientes cualitativos (burgués o proletario, en vez del nivel cuantitativo del ingreso).

Esta clase de hechos cualitativos, característicamente se expresan como datos numéricos de las variables adecuadas, denominadas *dummy*. La variable dummy adopta uno de los valores posibles; un valor significa una posibilidad cualitativa y el otro valor implica la otra posibilidad. Por convención, la variable dummy asume un valor de cero o de uno, donde uno se refiere a la ocurrencia de un evento o la presencia de una característica y cero se refiere a la no ocurrencia del evento o la ausencia de la característica.

Los *datos en series de tiempo* miden una variable en particular durante periodos de tiempo sucesivo o en diferentes fechas. A menudo el lapso es un año (es decir, datos anuales), pero puede ser un trimestre, un mes o una semana (es decir, datos trimestrales, mensuales o semanales). Es posible emplear un periodo de tiempo más largo para otros propósitos (dos años, cinco años, una década o más). Por lo común, las observaciones son sucesivas y están igualmente espaciadas en el tiempo.

Los *datos en sección cruzada* o de *corte transversal* miden una variable particular en un periodo de tiempo dado, para diferentes entidades. De la misma forma que el "periodo de tiempo" puede asumir diferentes valores en los datos en series de tiempo, la "entidad" puede asumir distintas identidades en los datos en sección cruzada. Por ejemplo, las entidades pueden ser diferentes países o bien pueden referirse a las situaciones de las empresas, las industrias, las familias, o los individuos en una fecha determinada.

En ocasiones, los datos en sección cruzada y los datos en series de tiempo se fusionan o combinan. El resultado puede interpretarse como una sección cruzada en series de tiempo o como una serie de tiempo en secciones cruzadas.

En general, los datos en sección cruzada y en series de tiempo proporcionan estimaciones distintas sobre un modelo. Estos datos y sus estimaciones resultantes generalmente no son comparables. Ninguna de las dos estimaciones es "incorrecta" y cual de las dos usar depende del propósito de la investigación. Por ejemplo, para el análisis estructural, es apropiado emplear datos en sección cruzada, en tanto que para propósitos de predicción a corto plazo, los datos en series de tiempo pueden resultar adecuados. O sea, los datos en series de tiempo reflejan por lo común conductas de corto plazo, mientras que los datos en sección cruzada reflejan comportamientos de largo plazo.

Los *datos en panel* (o datos longitudinales) son un tipo especial de datos combinados sección cruzada-series de tiempo, en el cual se muestra a través del tiempo la misma sección cruzada. Un ejemplo son los datos de la categorización de Nielsen sobre la popularidad de los programas de televisión. Los datos de panel por lo general son *micro datos*, pertenecientes a agentes económicos individuales como las familias o las empresas. No obstante, la mayor parte de los datos disponibles para la investigación econométrica son *macro datos,* pertenecientes a agentes económicos individuales.

En general, los micro datos son preferibles a macro datos porque evitan problemas de agregación y permiten estimar modelos que contienen relaciones aplicables a agentes individuales, pero como suele suceder generalmente, tales datos por lo común no están disponibles porque resulta muy costoso obtenerlos, además de que su publicación puede revelar información confidencial o sobre la propiedad.

1.4.4.3 Datos no-experimentales *vs.* datos experimentales

Los datos *no experimentales* en forma característica se obtienen de observaciones de un sistema no sujeto a control experimental. En contrate, los *datos experimentales* se obtienen a partir de un experimento controlado, esto es, una situación en la cual se aísla el sistema, o proceso bajo investigación, de influencias externas y, a cualquier grado que resulte posible, las influencias en el sistema están sujetas al control del experimentador.

A menudo se ha aseverado que un aspecto importante de la distinción de las ciencias sociales y las ciencias naturales es el tipo de datos que cada una utiliza. Comúnmente, los datos utilizados por las ciencias naturales son experimentales, resultantes de experimentos controlados; mientras que en las ciencias sociales, los datos no son experimentales puesto que en ellos las condiciones subyacentes no están sujetas a control y no pueden ser repetidas. A pesar de ser válida en general, esta distinción no se aplica en todos los casos. Las ciencias naturales de laboratorio, incluyendo la Química y la Física utilizan, de manera significativa, experimentos controlados. Un físico, por ejemplo, que realiza un experimento en física nuclear, a menudo utilizará un acelerador de partículas -un medio ambiente controlado que ofrece datos experimentales que pueden ser repetidos-, sin embargo, los astrofísicos en general no pueden efectuar experimentos de laboratorio sino que, en vez de eso, deben basarse en observaciones sobre las que no tienen control.

1.4.4.4 Problemas con los datos

A pesar de que en algunas situaciones se reúnen datos experimentales, la mayoría de los estudios econométricos deben basarse en datos no experimentales. Los problemas que este tipo de datos presenta pueden ser denominados, utilizando la terminología de la Astronomía, como problemas por "una mala interpretación".

El primero es el *problema de los grados de libertad*, esto es, que los datos disponibles simplemente no incluyen suficiente número de observaciones como para permitir una estimación adecuada del modelo. Bajo el empleo de datos no experimentales es imposible repetir las condiciones que dieron lugar a esos datos de modo que también es imposible generar puntos de

datos adicionales. En algunos casos, los datos disponibles pueden ser inadecuados para estimar un modelo más simple.

En segundo lugar está el *problema de multicolinealidad*, es decir, la tendencia de los datos a agruparse o a moverse juntos en vez de estar "esparcidos". Por ejemplo, en los datos en series de tiempo, las variables propenden a exhibir las mismas tendencias, ya sean éstas cíclicas o seculares, a través del tiempo. Con datos experimentales podría ser posible alterar las condiciones del experimento y obtener una "dispersión" adecuada. Con datos no experimentales estos controles no existen y el sistema del mundo real puede albergar variaciones independientes muy ligeras en los datos y en particular, un alto grado de interdependencia entre ciertas variables.

En tercer lugar está el *problema de correlación serial*, esto es el hecho de que, cuando se utilizan datos en series de tiempo, los cambios subyacentes ocurren muy lentamente a través del tiempo. Así, las condiciones en periodos de tiempo muy cercanos tienden a ser muy similares. En la medida en que el término de perturbación estocástica representa condiciones relevantes al modelo, que no han sido tomadas en cuenta explícitamente en él, tales como variables omitidas, la correlación serial se manifiesta en una dependencia del término de perturbación estocástica en un periodo sobre el de otro periodo.

En cuarto lugar está el *problema de cambio estructural*, aquél en el que puede haber un cambio discontinuo en el mundo real de modo que los datos se refieran a distintas poblaciones. Un ejemplo para los datos en series de tiempo es un periodo de guerra, que a menudo debe ser excluido por no ser representativo.

En quinto lugar está el *problema de errores de medición*, esto se refiere a que los datos que están medidos son sujetos a diversas imprecisiones y desviaciones. De hecho, los datos a menudo son revisados debido a un reconocimiento posterior de estas imprecisiones y desviaciones. De forma más fundamental, las imprecisiones potenciales se provocan por falta de una estricta precisión en la conceptualización. Por ejemplo, las cuentas del PNB (Producto Nacional Bruto) se revisan de vez en cuando sobre la base de tales cambios en la conceptualización. Tales cambios en dicha conceptualización requieren la reconfiguración de los datos para poder hacerlos comparables y consistentes en el tiempo.

Todos estos problemas se analizarán con mayor detalle en secciones posteriores. Cabe señalar que debido a estos problemas, los datos por lo común son depurados en varias formas. No obstante una depuración que ayuda a resolver uno de los problemas, puede agravar otro. Así, por ejemplo, remplazar datos en series de tiempo anuales por datos trimestrales, aumenta el número de puntos de información pero tiende a agravar tanto el problema de multicolinealidad como el de correlación serial.

Eliminar puntos de información no "representativos", como los que pertenecen a periodos poco usuales (años de guerra), ayuda a resolver el problema de cambio estructural, pero agrava los problemas de los grados de libertad y de multicolinealidad. Remplazar variables por sus primeras diferencias eventualmente ayudará a resolver el problema de correlación serial pero agravará el problema de errores en la medición. Es claro que deben hacerse elecciones juiciosas para obtener datos relevantes y que queden listos para su utilización, a partir de un conjunto de datos en bruto.

1.4.5 Estimación estadística

Una vez especificado el modelo econométrico y recolectados los datos apropiados para dicho modelo, la tarea siguiente del econometrista consiste en obtener estimaciones (valores numéricos) de los parámetros del modelo, a partir de la información disponible. Estas estimaciones le confieren un contenido empírico a la teoría.

Así pues, el estimar es calcular la relación que existe entre la variable dependiente y la(s) variable(s) explicativa(s). Por otro lado, la estimación es el proceso de inferir o estimar un parámetro de población (tales como su media o desviación estándar) del correspondiente estadístico de una muestra extraída de la población. Para que sea válida la estimación se debe de basar en una muestra *representativa*. Sólo resta decir que el análisis de regresión es la técnica que frecuentemente se utiliza para obtener dichas estimaciones.

1.4.6 Inferencia estadística al relacionar a la teoría económica con el análisis empírico.

Habiendo obtenido ya estimaciones de los parámetros, la siguiente tarea del econometrista consiste en desarrollar criterios apropiados dirigidos a

establecer si las estimaciones obtenidas están de acuerdo con lo que se espera de la teoría que se está verificando. La refutación o confirmación de las teorías, basándose en la evidencia empírica, se fundamenta en una rama de la teoría estadística conocida como *inferencia estadística* (la inferencia estadística se refiere a la estimación y a la prueba de hipótesis).

La inferencia estadística es uno de los aspectos más importantes y cruciales en el proceso de toma de decisiones, en la economía, la administración y en la ciencia en sí. Una vez que el modelo estadístico ha sido estimado y se ha comprobado que es válido desde un punto de vista estadístico, será necesario iniciar un proceso de "pruebas-imposición de restricciones-reparametrización", que conduzca a la obtención de un modelo teóricamente coherente y que será, finalmente, el modelo econométrico empírico, aquél que podrá ser considerado como la mejor aproximación del proceso mediante el cual se explica el fenómeno en cuestión.

1.5 Propósitos y limitaciones de la *econometría*

La búsqueda de relaciones causales es el principal objetivo en un análisis econométrico. La econometría en su origen estuvo fundamentalmente orientada como una metodología que permitiese contrastar la veracidad de las relaciones causales establecidas por la teoría económica. La metodología econométrica permite estimar diversos modelos teóricos y contrastar cuáles de las variables explicativas incluidas tiene una incidencia significativa sobre la variable explicada. En este sentido existen dos cuestiones importantes que pueden plantearse: la existencia de correlación estadística entre dos variables ¿implica la existencia de una relación causal entre ellas? La aceptación de hipótesis de nulidad de un parámetro ¿implica que no existe relación entre la variable explicativa correspondiente y la variable explicada?

Respecto a la primera cuestión cabe señalar que la existencia de una relación causal importante entre dos variables implica la existencia de un elevado grado de correlación entre ellas, pero lo recíproco no es cierto, ya que puede existir una elevada correlación estadística debida a la casualidad, o debido a que existe algún nexo entre ambas variables, a través de una tercera variable con la que ambas están relacionadas, sin que necesariamente exista una relación causal entre ellas. La existencia

de una relación estadística significativa y estable entre dos o más variables raramente se debe a la casualidad.

Generalmente esta relación se debe a la existencia de algún vínculo casual entre ellas o bien a la relación que ambas mantienen con una tercera variable. La distinción entre estas dos últimas situaciones no siempre es fácil, si la relación existente entre dos variables no se debe a la casualidad, ni a la relación con una tercera variable (lo cual se puede contrastar estadísticamente), entonces es claro que existe una relación causal. El análisis de causalidad es bastante complicado, ya que a veces es difícil precisar el sentido de la causalidad (es decir cual es la variable causa y cuál la variable efecto).

Los principales enfoques econométricos para el análisis de causalidad son el de Wiener-Granger, el de Wu-Hausman, y el basado en el análisis postmuestral del modelo. Los dos primeros son propios de un nivel más avanzado, pero el tercero de estos enfoques puede abordarse con un conocimiento de la metodología econométrica de carácter general. Este enfoque consiste en aceptar la validez de un modelo basado en hipótesis razonables si éste no es refutado en su contrastación con la realidad, tanto en lo que respecta al periodo de estimación del modelo, o periodo muestral, como en lo que concierne a su estabilidad postmuestral y su capacidad predictiva.

Si bien el propósito o la aplicación práctica más frecuente de los modelos econométricos es la predicción, no debe olvidarse que la contrastación de las teorías es también importante, ya que no sólo contribuye a mejorar la capacidad predictiva de los modelos, sino que además permite una mejor orientación de la teoría en cuestión.

1.5.1 Estadística y econometría

Por un lado tenemos a la estadística que se encarga de la colección, presentación, análisis y utilización de datos numéricos para realizar inferencias y alcanzar decisiones ante la incertidumbre que plantean la economía, la administración y otras ciencias sociales o físicas y, por el otro lado a la *econometría* que es la técnica que se abastece del sustento estadístico para así llevar a cabo sus propósitos. Ciertamente sin la estadística no hubiese existido la *econometría*, o sea, la estadística es la herramienta por medio de la cual se construye la *econometría*.

A pesar de que la *econometría* aún no se ha ganado el título de ciencia, su basto campo de trabajo se amplía y fortalece día con día y cada vez resultan más poderosas sus bases estadísticas. Estadística y *econometría*, dos disciplinas totalmente diferentes pero con mucho en común, quizás esto suene contradictorio, pero no lo es, pues, al decir que son diferentes me refiero a que mientras la estadística tiene todo un fundamento matemático que la respalda, la *econometría* carece del mismo y, al decir que tienen mucho en común me refiero a que tanto estadística como *econometría* persiguen el mismo objetivo: explicar lo mejor posible la realidad.

Cualquier estudio econométrico puede tener uno, dos o todos los propósitos primordiales de la *econometría*; el análisis estructural, la predicción y la evaluación de políticas (entiéndase por política ciertos criterios de determinada teoría), mismos que representan los "productos finales" de la *econometría*, del mismo modo que la "teoría" y los "hechos" constituyen sus "materias primas". En este sentido, la figura 1.3 puede ser concebida como un diagrama de flujo que muestra de manera esquemática cómo se combinan y utilizan eventualmente las distintas partes de un estudio econométrico.

El *análisis estructural* es el uso de un modelo econométrico estimado, para efectuar la medición cuantitativa de relaciones económicas. También permite la comparación de teorías contrarias sobre un mismo fenómeno. El análisis estructural representa lo que podría verse como el propósito "científico" de la *econometría*: comprender los fenómenos del mundo real mediante la medición cuantitativa, prueba y validación de relaciones económicas. Es factible que un resultado de este análisis sea un efecto "retroalimentador" sobre la teoría.

La *predicción* es la aplicación de un modelo econométrico estimado, para predecir valores cuantitativos de ciertas variables fuera de la muestra de datos realmente observados. Con frecuencia los pronósticos son la base para tomar decisiones; por ejemplo, la compra de materias primas y el empleo de trabajadores adicionales en una empresa pueden apoyarse en una predicción de que las ventas se incrementarán durante los dos trimestres subsecuentes.

La *evaluación de políticas* es el manejo de un modelo econométrico estimado para elegir entre políticas alternas. Un enfoque presenta explícitamente una función objetivo por maximizar mediante la elección de

políticas y considera al modelo estimado como una restricción de este proceso de optimización. Otro enfoque, a menudo más útil para los hacedores de política, simula diferentes políticas y hacer predicciones condicionadas sobre los valores futuros de las variables relevantes bajo cada opción.

La selección de la política más deseable entre los distintos "futuros candidatos" posibles, indicaría cual política debiera seguirse. En cualquier caso, la selección de una política particular, combinada con los efectos de aquellos eventos exógenos que tienen influencia sobre el sistema, conduce a resultados específicos, y éstos, a su vez, a otra relación de "retroalimentación" que conecta la evaluación de políticas con los hechos.

Estos tres propósitos principales de la *econometría* están íntimamente relacionados. La estructura determinada a través del análisis estructural es utilizada en la predicción que emplea un modelo econométrico, en tanto que la evaluación de políticas que utiliza un modelo econométrico es un tipo de predicción condicionada. Una dificultad que enfrenta la *econometría* consiste en que la teoría especifica la relación funcional exacta entre las diversas variables, aunque en la realidad la medición de la relación funcional entre las variables no es exacta. Otro problema al que se enfrenta la *econometría*, y el cual quizás es el de mayor importancia, es la falta de datos "reales", pues muchas de las veces se cuenta con modelos que explican muy bien al fenómeno de estudio pero sin embargo ya sea por la incapacidad de hacer muestreos convincentes o por el alto costo de los mismos no pueden obtenerse datos confiables.

En fin, posiblemente la lista de limitantes de la *econometría* sea muy extensa, pero lo que menos se pretende es hacer un listado de ellas en este trabajo, sino por el contrario, la *econometría* como parte de la ciencia, también es factible, pues está expuesta a factores externos que no pueden ser contemplados en ningún modelo econométrico. Lo que sí es cierto, y como ya lo había mencionado con anterioridad, la *econometría* se fortalece día con día y se hace indispensable en un estudio predictivo o de simulación.

Finalmente lo que queda decir es que; si el arte de la elaboración de modelos consiste en un conjunto de instrumentos, en su mayoría cuantitativos, que se utilizan para construir y contrastar representaciones matemáticas del mundo real entonces, la *econometría* es un arte.

2 La naturaleza del análisis de regresión

2.1 El método de Mínimos Cuadrados Ordinarios

En la mayoría de los estudios econométricos es imposible el trabajar con poblaciones, por lo que se recurre al empleo de una muestra aleatoria de dicha población. Un objetivo frecuente en la investigación es la especificación de una relación funcional entre dos variables tal como $Y=f(X)$, aquí Y es la ***variable dependiente*** y X es la ***variable independiente***.

La finalidad de trabajar con los datos de la muestra y no con los datos poblacionales es que, éstos últimos no son observables en la realidad, mientras que los primeros sí pueden observarse. Con base en dichos datos muestrales se pretende tener un conocimiento más o menos "confiable" del comportamiento del fenómeno en estudio, es decir, la muestra (siendo ésta aleatoria) es tan sólo una aproximación a la realidad de la población. Esto puede verse más claramente con la ayuda de la figura 2.1.

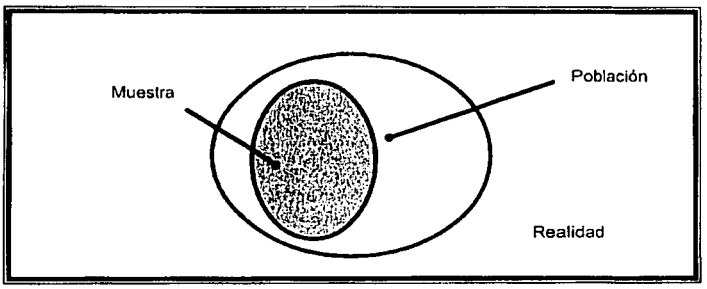


Figura 2.1 Representación gráfica de una muestra

Como es de esperarse, dependiendo de la calidad y tamaño de la muestra serán los resultados obtenidos para especular en cuanto al comportamiento poblacional. Así, una vez especificado el tamaño de la muestra con que se va trabajar el siguiente paso es plantear la relación funcional de los datos observados, es decir, ésta puede ser de forma lineal, exponencial, logarítmica, etc.

El modelo lineal simple, o ***análisis de regresión simple***, se usa para probar hipótesis sobre la relación lineal entre una variable dependiente, **Y**, y una variable independiente o explicatoria, **X**, y para la predicción. Ahora bien, este tipo de modelo es importante no sólo porque puede ser de utilidad para contrastar hipótesis y para predecir, sino también porque constituye la base para el análisis de los modelos de simulación multiecuacionales y de las series temporales. El análisis de regresión lineal simple por lo general comienza representando gráficamente el conjunto de los valores **XY** sobre un diagrama de dispersión y determinando por inspección si allí existe una relación lineal aproximada.

$$Y_i = \beta_0 + \beta_1 X_i \qquad (2.1.1)$$

Ahora bien, como es casi seguro de que los puntos observados a lo largo del tiempo no caigan exactamente en la línea de regresión, la línea de regresión exacta de la ecuación (2.1) debe ser modificada para incluir un término de perturbación aleatorio, error, o término estocástico U_i.

$$Y_i = \beta_0 + \beta_1 X_i + U_i \quad \text{(modelo de regresión lineal poblacional)} \qquad (2.1.2)$$

donde

Y_i	es la variable dependiente
X_i	es la variable independiente
U_i	es el término estocástico del modelo
β_0 y β_1	son los parámetros del modelo lineal

El modelo de regresión lineal poblacional, como su nombre lo indica, pretende encontrar una posible relación lineal en cuanto a las observaciones recabadas de la variable involucrada X y de la variable explicada Y. Debido al hecho de que los errores poblacionales tampoco son observables, la medida de corrección es utilizar un modelo que haga uso de una aproximación a dichos errores, a este tipo de modelo se le conoce como el modelo de ***regresión lineal muestral.***

$$Y_i = b_0 + b_1 X_i + e_i \qquad \text{(modelo de regresión lineal muestral)} \quad (2.1.3)$$

donde

Y_i es el valor de la variable dependiente

X_i es el valor captado para la variable independiente

b_0 es el estimador del verdadero valor poblacional β_0

b_1 es el estimador del verdadero valor poblacional β_1

e_i es el estimador de los errores poblacionales U_i

Supongamos que nos interesa averiguar la relación existente entre dos variables, X e Y. A fin de describir estadísticamente esta relación, necesitamos disponer de un conjunto de observaciones para cada variable y de una hipótesis que establezca la forma matemática explícita de la relación entre X e Y, a éste conjunto de observaciones se le denomina la muestra. Se considerará el caso de que la hipótesis establezca que la mejor forma de describir la relación entre X e Y es mediante una línea recta. Dado el supuesto de linealidad, el objetivo consistirá en especificar una regla mediante la cual sea posible determinar la "mejor línea recta" de ajuste entre X e Y.

Y (calificación media)	X (renta de los padres en miles de pesos)
7	210
6.5	150
4	150
7	90
8	120
7.5	180
8.5	60
6	120

Tabla 2.1 Calificaciones medias y renta familiar

Por ejemplo, supóngase que se desea probar la hipótesis de que la calificación promedio de un estudiante puede predecirse o explicarse en gran medida a partir de los ingresos de sus padres. Para esto se obtienen ocho observaciones de muestra que se indican en la tabla 2.1. Es posible elegir varias rectas que se ajusten a los puntos del diagrama de dispersión correspondientes a las observaciones (figura 2.2). Una de ellas podría ser la que une los puntos desde el valor inferior de X al valor superior de X (línea L1), o bien podría trazarse a ojo una recta que, en apariencia, se ajuste a todos los puntos del diagrama (línea L2).

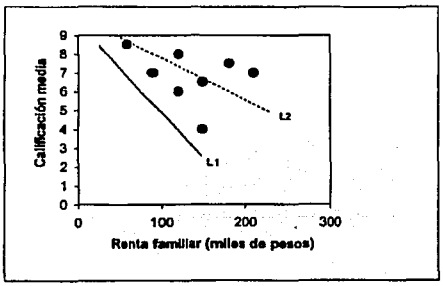


Figura 2.2 Diagrama de dispersión

Un procedimiento mejor consistiría en una recta tal que la suma de las distancias (positivas y negativas) medidas verticalmente entre los puntos del diagrama y la recta fuese igual a cero, estas distancias llamadas *desviaciones*, están representadas en la figura 2.3. Este método nos permite asegurarnos que a todas las desviaciones de igual signo y magnitud se les da la misma importancia. Sin embargo, por desgracia, este procedimiento adolece de la poco deseable propiedad de que dos desviaciones de igual magnitud y distinto signo se anulan.

Este método puede ser mejorado si se hiciera mínimo el valor absoluto de las desviaciones de las observaciones de la muestra con respecto a la línea ajustada. Este procedimiento implica prejuzgar que la importancia de una desviación es proporcional a su magnitud. Aunque la minimización de la suma de las desviaciones en valor absoluto parece interesante no deja de presentar algunas desventajas. En primer lugar, el procedimiento resulta difícil de utilizar desde el punto de vista de los cálculos. En segundo lugar, sería razonable tratar las desviaciones grandes con atención relativamente mayor que las pequeñas. Por ejemplo, una predicción con un error de dos unidades sería probablemente considerada peor que una predicción con dos errores de una unidad cada uno.

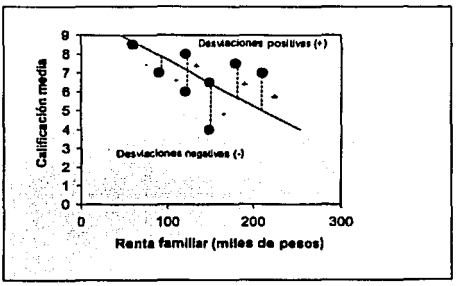


Figura 2.3 Desviaciones

Las relaciones estadísticas se establecen, generalmente, con el propósito de predecir o explicar los cambios que experimenta la variable dependiente cuando se producen cambios en la variable explicativa (o varias variables explicativas para el caso de regresión lineal múltiple).

Para el diagrama de dispersión de la figura 2.2 se puede escribir la ecuación $Y_i = b_o + b_1 X_i$. Puesto que el objetivo es explicar o predecir los cambios experimentados por Y, es lógico que la finalidad sea la minimización de la suma de los cuadrados de las desviaciones verticales de los puntos con respecto a la línea ajustada.

Cuando se decide escribir una ecuación de la forma $Y_i = b_o + b_1 X_i$ en lugar de la forma $X_i = c + d Y_i$ quiere decir que existe un juicio implícito previo de que los cambios en la variable Y son "causados" por cambios experimentados por la variable X y no al revés. En el ejemplo de las calificaciones se ha supuesto implícitamente que la calificación media de un estudiante está determinada por la renta de la familia. Si se combinara este supuesto por el de que la renta familiar viene determinada por la calificación media, la ecuación se escribiría de la forma $X_i = c + d Y_i$ y el criterio para el ajuste variaría de acuerdo con este cambio de supuesto.

Existe un procedimiento factible desde el punto de vista del cálculo, que penaliza los errores grandes relativamente más que los pequeños y que garantiza que la mitad de las desviaciones estará por debajo de la línea ajustada y que la otra mitad estará por encima. Este procedimiento es el denominado de los ***mínimos cuadrados ordinarios***. El criterio es el siguiente; *la "mejor línea de ajuste"* es aquella que minimiza la suma de los cuadrados de las desviaciones de los puntos del diagrama con respecto de la recta (las desviaciones se miden verticalmente).

2.2 Estimadores de Mínimos Cuadrados Ordinarios

El criterio de los **mínimos cuadrados ordinarios** (midiendo las distancias verticalmente) puede expresarse formalmente mediante la siguiente expresión:

$$\text{Minimizar} \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 \qquad (2.2.1)$$

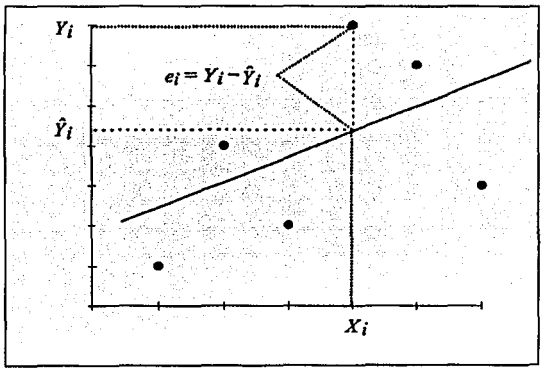


Figura 2.4 Valores ajustados

Donde $\hat{Y}_i = b_o + b_1 X_i$ es el valor ajustado del valor de Y_i correspondiente a una observación, en particular de X_i , y n es el número de observaciones (véase la figura 2.4). Los valores de b_o y b_1 son desconocidos y deberán calcularse de forma que satisfagan el criterio de la ecuación 2.2.1.

El problema consiste en elegir (simultáneamente) unos valores de b_o y b_1 que hagan mínima la expresión de la ecuación 2.2.1. Para ello, basta con utilizar el cálculo elemental de la siguiente manera:

Dado que el *modelo simple o modelo econométrico bivariable* puede escribirse como la ecuación 2.1.3, éste trata de explicar la variable endógena Y_i en términos de un intercepto (b_o), de la variable exógena X_i , y de una variable aleatoria denominada "error" e_i .

Sólo dos de estas variables realmente nos interesan $\mathbf{Y_i}$ y $\mathbf{X_i}$, las cuales sí son observables, es decir, son medibles y por lo tanto se pueden representar por datos.

El modelo 2.1.2 representa el modelo que más se aproxima a la realidad, pero sin embargo los parámetros β_0 y β_1 sólo pueden ser estimados y no hallados con exactitud. Sean b_o y b_1 los estimadores de β_0 y β_1 respectivamente, por lo tanto el residuo puede expresarse como:

$$e_i = Y_i - (b_o + b_1 X_i) \qquad (2.2.2)$$

Es decir, el residuo es la diferencia entre el valor real y el valor estimado (ver figura 2.4).

El método de los mínimos cuadrados es empleado como instrumento para calcular los estimadores b_o y b_1 . En el diagrama de dispersión de la figura 2.4 las observaciones de la variable exógena X_i fueron dibujadas contra las respectivas observaciones de la variable endógena Y_i .

El método de los *mínimos cuadrados* trata de ajustar una línea recta continua, de tal manera que minimice la suma de los cuadrados de los residuos e_i . Así pues, este método consiste en minimizar la función que suma los cuadrados de los residuos, y se puede expresar algebraicamente de la siguiente manera:

$$Q(b_o, b_1) = \sum_{i=1}^{n} e_i^2 = \sum_{i=1}^{n} (Y_i - b_o - b_1 X_i)^2 \qquad (2.2.3)$$

Como e_i está ahora expresado en función de b_0 y b_1,se puede derivar parcialmente con respecto a cada uno de estos estimadores.

Por lo tanto:

$$\frac{\partial Q(b_0,b_1)}{\partial b_0} = \sum_{i=1}^{n} 2(Y_i - b_0 - b_1 X_i)(-1) \qquad (2.2.4)$$

Y de la misma manera:

$$\frac{\partial Q(b_0,b_1)}{\partial b_1} = \sum_{i=1}^{n} 2(Y_i - b_0 - b_1 X_i)(-X_i) \qquad (2.2.5)$$

Para encontrar el mínimo de la función $Q(b_0, b_1)$ es necesario igualar cada una de las derivadas parciales a cero, o sea:

$$\frac{\partial Q(b_0,b_1)}{\partial b_0} = \sum_{i=1}^{n} 2(Y_i - b_0 - b_1 X_i)(-1) = 0 \qquad (2.2.6)$$

$$\frac{\partial Q(b_0,b_1)}{\partial b_1} = \sum_{i=1}^{n} 2(Y_i - b_0 - b_1 X_i)(-X_i) = 0 \qquad (2.2.7)$$

O lo que es igual:

$$\frac{\partial Q(b_0,b_1)}{\partial b_0} = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i) = 0 \qquad (2.2.8)$$

$$\frac{\partial Q(b_0,b_1)}{\partial b_1} = \sum_{i=1}^{n} (Y_i - b_0 - b_1 X_i)(X_i) = 0 \qquad (2.2.9)$$

Las ecuaciones 2.2.8 y 2.2.9 se conocen como **"Ecuaciones Normales"** y pueden ser resueltas para encontrar los valores de b_0 y b_1.

Estas ecuaciones normales se pueden expresar en forma equivalente como:

$$\sum_{i=1}^{n} Y_i = nb_0 + b_1 \sum_{i=1}^{n} X_i \qquad (2.2.10)$$

$$\sum_{i=1}^{n} Y_i X_i = b_0 \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \qquad (2.2.11)$$

Usando el método de sustitución, de la ecuación 2.2.10 puede despejarse b_0 y quedar en función de los otros términos, es decir:

$$b_0 = \frac{\sum_{i=1}^{n} Y_i - b_1 \sum_{i=1}^{n} X_i}{n} = \frac{\sum_{i=1}^{n} Y_i}{n} - b_1 \frac{\sum_{i=1}^{n} X_i}{n} = \overline{Y} - b_1 \overline{X} \qquad (2.2.12)$$

Ahora sustituyendo el valor de b_0 en la ecuación (2.2.11) se tiene el siguiente resultado:

$$\sum_{i=1}^{n} Y_i X_i = \left(\frac{\sum_{i=1}^{n} Y_i}{n} - b_1 \frac{\sum_{i=1}^{n} X_i}{n} \right) \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \qquad (2.2.13)$$

Y en forma equivalente:

$$\sum_{i=1}^{n} Y_i X_i = \frac{\sum_{i=1}^{n} Y_i}{n} \sum_{i=1}^{n} X_i - b_1 \frac{\sum_{i=1}^{n} X_i}{n} \sum_{i=1}^{n} X_i + b_1 \sum_{i=1}^{n} X_i^2 \qquad (2.2.14)$$

Si ahora se multiplica el primer término de la derecha por $\frac{n}{n}$ y tomando a b_1 como factor común para los otros dos términos, se tiene lo siguiente:

$$\sum_{i=1}^{n} Y_i X_i = \frac{n}{n}\sum_{i=1}^{n} X_i \frac{\sum_{i=1}^{n} Y_i}{n} + b_1\left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right) \qquad (2.2.15)$$

O sea que:

$$\sum_{i=1}^{n} Y_i X_i = n\frac{\sum_{i=1}^{n} X_i}{n}\frac{\sum_{i=1}^{n} Y_i}{n} + b_1\left(\sum_{i=1}^{n} X_i^2 - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right) \qquad (2.2.16)$$

Si ahora se multiplica el segundo término del paréntesis por $\frac{n}{n}$ se obtiene lo siguiente:

$$\sum_{i=1}^{n} Y_i X_i = n\frac{\sum_{i=1}^{n} X_i}{n}\frac{\sum_{i=1}^{n} Y_i}{n} + b_1\left(\sum_{i=1}^{n} X_i^2 - \frac{n}{n}\frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n}\right) \qquad (2.2.17)$$

Y en términos de promedios:

$$\sum_{i=1}^{n} Y_i X_i = n\overline{XY} + b_1\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right) \qquad (2.2.18)$$

De donde se puede despejar a b_1 obteniéndose que:

$$b_1 = \frac{\sum_{i=1}^{n} Y_i X_i - n\overline{XY}}{\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)} \qquad (2.2.19)$$

pero $n\overline{XY} = \sum_{i=1}^{n} \overline{XY}$

y $n\overline{X}^2 = \sum_{i=1}^{n} \overline{X}^2$

Por lo que la expresión para b_1 se transforma en:

$$b_1 = \frac{\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} \overline{XY}}{\left(\sum_{i=1}^{n} X_i^2 - \sum_{i=1}^{n} \overline{X}^2\right)} \qquad (2.2.20)$$

Y por las propiedades de las sumas se tiene lo siguiente:

$$b_1 = \frac{\sum_{i=1}^{n} \left(Y_i X_i - \overline{XY}\right)}{\sum_{i=1}^{n} \left(X_i^2 - \overline{X}^2\right)} \qquad (2.2.21)$$

Y finalmente b_1 se convierte en:

$$b_1 = \frac{n\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \qquad (2.2.22)$$

Así pues, los estimadores por el método de Mínimos Cuadrados Ordinarios son:

$$b_0 = \overline{Y} - b_1 \overline{X} \qquad (2.2.23)$$

y

$$b_1 = \frac{n\sum_{i=1}^{n} Y_i X_i - \sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n\sum_{i=1}^{n} X_i^2 - \left(\sum_{i=1}^{n} X_i\right)^2} \qquad (2.2.24)$$

Las expresiones (2.2.23) y (2.2.24) se simplifican en el caso particular de que las medias de muestra de X e Y sean nulas. La ecuación (2.2.23) puede escribirse como:

$$b_0 = \overline{Y} - b_1 \overline{X} = 0 \qquad (2.2.25)$$

Por tanto, cuando las medias de X e Y son iguales a cero, la ordenada en el origen de la recta de regresión ajustada será igual a cero. Para obtener la correspondiente estimación de la pendiente, se dividen el numerador y el denominador de (2.2.24) por n^2.

Obteniéndose

$$b_1 = \frac{\frac{\sum_{i=1}^{n} Y_i X_i}{n} - \frac{\sum_{i=1}^{n} X_i}{n} \frac{\sum_{i=1}^{n} Y_i}{n}}{\frac{\sum_{i=1}^{n} X_i^2}{n} - \frac{\left(\sum_{i=1}^{n} X_i\right)^2}{n^2}} \qquad (2.2.26)$$

Sustituyendo $\frac{\sum_{i=1}^{n} X_i}{n}$ y $\frac{\sum_{i=1}^{n} Y_i}{n}$ por $\overline{X}$ y $\overline{Y}$ se tiene que:

$$b_1 = \frac{\frac{\sum_{i=1}^{n} X_i Y_i}{n} - \overline{XY}}{\frac{\sum_{i=1}^{n} X_i^2}{n} - \overline{X}^2} \qquad (2.2.27)$$

Pero por definición $\overline{X}=\overline{Y}=0$. por tanto:

$$b_1 = \frac{\sum_{i=1}^{n} X_i Y_i}{\sum_{i=1}^{n} X_i^2} \qquad (2.2.28)$$

El hecho de que la ecuación (2.2.28) es menos complicada que la ecuación (2.2.26) sugiere que resultará más cómodo expresar las ecuaciones mínimo cuadráticas en función de las variables expresadas en forma de desviaciones con respecto a sus medias muestrales correspondientes, tanto si dichas medias son nulas como si no lo son.

Para ello, se expresarán los datos en forma de desviaciones con respecto a sus medias respectivas.

Sean

$$x_i = X_i - \overline{X} \qquad \text{y} \qquad y_i = Y_i - \overline{Y}$$

Entonces el estimador b_1 se calculará de la siguiente manera:

$$b_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2} \tag{2.2.29}$$

Y como b_0 depende tan sólo de b_1 entonces su cálculo será el mismo, es decir:

$$b_0 = \overline{Y} - b_1 \overline{X} \tag{2.2.30}$$

De esta manera, la línea recta que mejor se ajusta a los datos viene dada por la siguiente expresión:

$$\hat{Y}_i = b_0 + b_1 X_i \tag{2.2.31}$$

Es de suma importancia mostrar el razonamiento para encontrar los coeficientes b_0 y b_1 que satisfacen la ecuación 2.2.1, debido a que todos los estudios hechos sobre nuevos métodos de estimación descansan sobre el comúnmente llamado ***Método de los Mínimos Cuadrados Ordinarios***.

2.3 Supuestos en el análisis de regresión lineal simple

2.3.1 Supuesto de linealidad

Este supuesto hace referencia a los coeficientes β_0, β_1 y al término estocástico U_i mencionados con anterioridad en este trabajo. El modelo puede no tener una expresión lineal puesto que las variables pueden estar descritas en funciones de otras, como por ejemplo Y = log X. De esta manera, el modelo exige que tenga linealidad en los parámetros pero no en las variables. Este hecho permite tener modelos no lineales y transformarlos por medio de instrumentos matemáticos en modelos lineales.

2.3.2 Supuesto de observabilidad

Este supuesto es de suma importancia pues para poder estimar el modelo es necesario que sea factible medir las variables involucradas en el modelo en cuestión, en otras palabras, es importante contar con observaciones o datos para poder comenzar el estudio del fenómeno.

2.3.3 Supuesto sobre variables independientes

Las variables exógenas involucradas en el modelo son variables matemáticas que toman valores bien definidos y por consiguiente no involucran "algún" término estocástico. Por esto el análisis es condicional a los valores dados en dichas variables, sin embargo hay que tener en cuenta que los valores dados en estas variables son independientes con relación al error U_i o sea, Prob(Y)= f (U_i).

2.3.4 Supuestos sobre el término de error

Las razones por las cuales el modelo incluye el término de error son:

1) Inexactitud de la medición de las variables
2) Imperfección de la teoría, o
3) Imperfecta especificación del modelo

Los supuestos con respecto a estos términos son:

a) Los U_i son variables aleatorias distribuidas normalmente, lo que implica simetría e inmodalidad de la distribución. Es decir $U_i \sim N(0, \sigma^2)$.

b) El valor esperado o promedio de estos errores estocásticos es igual a cero. En símbolos esto quiere decir que $E(U_i)=0$. En otros términos, esto equivale a decir que los errores son compensados, tanto hacia arriba como hacia abajo, y así su suma aritmética es cero.

c) La varianza de U_i es finita y constante para todas y cada una de las observaciones, o sea que cumple con la homoscedasticidad Var $(U_i) = \sigma^2$. Esto quiere decir que las otras causas de variación representadas por el error (variables excluidas; cambios en el comportamiento humano, errores de medición, mala especificación, etc.) permanecen constantes para todas las observaciones.

d) Los errores son independientes para todas las observaciones, esto es que:

$$\forall\ U_i, U_j \quad COV(U_i, U_j) = 0 \text{ para } i \neq j \qquad (2.3.4.1)$$

Esto implica que el conjunto de las causas de variación que representa el error y que actúan en las observaciones, se presenta para cada una independiente de su actuación en las observaciones pasadas o en las siguientes. Esto es muy común en las series de tiempo, donde el valor de X_i depende del valor del periodo anterior X_{i-1}.

Las condiciones anteriores son necesarias para que los valores de b_0 y b_1 sean buenos estimadores de β_0, β_1 respectivamente, es decir que cumplan con las condiciones de *insesgo*, *consistencia* y *eficiencia*.

Pero ¿qué significa que un estimador sea insesgado?¿qué es el sesgo? Bueno, se dice que un estimador es insesgado si la media de su distribución muestral es igual al parámetro verdadero, es decir, la media de la distribución muestral es el valor esperado del estimador. De esta manera, la falta de sesgo significa que $E(\hat{\alpha}) = \alpha$, donde $\hat{\alpha}$ es el estimador del parámetro verdadero α.

El sesgo se define entonces como la diferencia entre el valor esperado del estimador y el parámetro verdadero. O sea, $sesgo = E(\hat{\alpha}) - \alpha$, sin embargo, la falta de sesgo no significa que $\hat{\alpha} = \alpha$, sino que en un muestreo aleatorio simple repetido, se obtendrá, en promedio, la estimación correcta. Se espera que la muestra realmente obtenida se acerque a la media de la distribución muestral del estimador.

Ahora bien, se requieren dos condiciones para que un estimador sea consistente, primero, en la medida que el tamaño de la muestra se incremente, el estimador debe aproximarse más y más al parámetro verdadero (a esta propiedad se le conoce como una carencia asintótica de sesgo), y segundo, cuando el tamaño de la muestra crece indefinidamente la distribución muestral del estimador debe restringirse o llega a ser una línea recta vertical con altura (probabilidad) de 1 sobre el valor del parámetro verdadero.

Por su parte, un estimador eficiente se refiere a aquél que tiene la varianza menor entre todos los estimadores insesgados, es el estimador insesgado con la distribución más compacta o menos extendida. Esto es muy importante porque el investigador podría estar más seguro de que el estimador está más próximo al parámetro poblacional real que se estima. Otra forma de expresar esto es decir que un estimador eficiente tiene el intervalo de confianza más pequeño y es probable que sea más significativo estadísticamente que cualquier otro estimador. Debe notarse que la varianza mínima por sí sola no es muy importante, a menos que se asocie con la falta de sesgo.

2.4 Algunas propiedades de los estimadores de Mínimos Cuadrados Ordinarios

2.4.1 Los estimadores b_0 y b_1 son insesgados y de varianza mínima, es decir, son eficientes.

Recuérdese que $b_1 = \frac{\sum_{i=1}^{n} x_i y_i}{\sum_{i=1}^{n} x_i^2}$ y sea $W_i = \frac{X_i}{\sum_{i=1}^{n} x_i^2}$ entonces es fácil verificar que

$$b_1 = \sum_{i=1}^{n} W_i y_i = \sum_{i=1}^{n} W_i Y_i \qquad (2.4.1.1)$$

si en esta expresión se sustituye el modelo teórico se tiene que:

$$b_1 = \sum_{i=1}^{n} W_i (b_0 + \beta_1 X_i + e_i) \qquad (2.4.1.2)$$

donde

$$b_1 = b_0 \sum_{i=1}^{n} W_i + \beta_1 \sum_{i=1}^{n} W_i X_i + \sum_{i=1}^{n} W_i e_i \qquad (2.4.1.3)$$

usando el hecho de que $\sum_{i=1}^{n} W_i = 0$ y que $\sum_{i=1}^{n} W_i X_i = 1$ entonces se tiene que:

$$b_1 = \beta_1 + \sum_{i=1}^{n} W_i e_i \qquad (2.4.1.4)$$

Ahora bien, tomando el valor esperado en ambos lados se obtendrá lo siguiente:

$$E[b_1] = E\left[\beta_1 + \sum_{i=1}^{n} W_i e_i\right] \qquad (2.4.1.5)$$

y como la esperanza es un *operador lineal* entonces se tiene que:

$$E[b_1]=\left[E[\beta_1]+E\left[\sum_{i=1}^{n}W_i e_i\right]\right] \qquad (2.4.1.6)$$

$$E[b_1]=\beta_1+\sum_{i=1}^{n}W_i E[e_i]=\beta_1 \qquad (2.4.1.7)$$

Esto debido al supuesto de que $E[e_i]=0$.

De esta manera se ha comprobado que el estimador lineal b_1 es un estimador lineal insesgado, y siguiendo el mismo procedimiento se puede constatar que b_0 también es un estimador lineal insesgado. Ahora se probará que el estimador b_1 (para b_0 se sigue un análisis similar) es de mínima varianza, comprobando así que es un estimador *eficiente*.

Para comprobar que entre todos los estimadores lineales insesgados b_1 es el que tiene la varianza mínima, primeramente se obtendrá la expresión de dicha varianza para b_1.

Siendo $b_1=\beta_1+\sum_{i=1}^{n}W_i e_i$ y sabiendo que $Var[X_i]=E\left[\left(X_i-E(X_i)\right)^2\right]$ entonces:

$Var[b_1]=E\left[\left(b_1-\beta_1\right)^2\right]$ y usando el hecho de que $\sum_{i=1}^{n}W_i e_i=b_1-\beta_1$ se tiene:

$$Var[b_1]=E\left[\left(\sum_{i=1}^{n}W_i e_i\right)^2\right]=E\left[\sum_{i=1}^{n}W_i^2 e_i^2+2\sum_{i<j}\sum W_i W_j e_i e_j\right] \qquad (2.4.1.8)$$

$$Var[b_1]=\sum_{i=1}^{n}W_i^2 E[e_i^2]+2\sum_{i<j}\sum W_i W_j E[e_i e_j]$$

Y sabiendo que $E[e_i^2]=\sigma^2$; $E[e_i e_j]=0$ y que $\sum_{i=1}^{n}W_i^2=\frac{1}{\sum_{i=1}^{n}x_i^2}$ entonces:

$$Var[b_1]=\sum_{i=1}^{n}W_i^2\sigma^2=\sigma^2\sum_{i=1}^{n}W_i^2=\frac{\sigma^2}{\sum_{i=1}^{n}x_i^2} \qquad (2.4.1.9)$$

De esta manera se tiene que $Var[b_1]=\frac{\sigma^2}{\sum_{i=1}^{n} x_i^2}$ y de forma análoga se puede probar que $Var[b_0]=\frac{\sigma^2 \sum_{i=1}^{n} X_i^2}{n \sum_{i=1}^{n} x_i^2}$.

Sin pérdida de generalidad se procederá a demostrar que el estimador b_1 es un estimador eficiente, para esto sólo falta mostrar que entre todos los estimadores lineales insesgados es el que tiene la mínima varianza.

Para esto, recuérdese que $b_1=\sum_{i=1}^{n} W_i y_i$ (es fácil probar que $b_1=\sum_{i=1}^{n} W_i Y_i$) lo cual muestra que b_1 es un promedio ponderado de las Y_i, con las W_i sirviendo como ponderaciones. Ahora bien, si se define un nuevo estimador lineal alterno de b_1 de la siguiente forma:

$b_1^*=\sum_{i=1}^{n} k_i y_i$ donde k_i son también ponderaciones, no necesariamente iguales a W_i.

Entonces se tendrá que:

$$E[b_1^*]=\sum_{i=1}^{n} k_i E[y_i] \text{ lo cual implica que } E[b_1^*]=\sum_{i=1}^{n} k_i E[\beta_0+\beta_1 x_i]$$

Y por las propiedades de la suma se tiene que $E[b_1^*]=\beta_0 \sum_{i=1}^{n} k_i+\beta_1 \sum_{i=1}^{n} k_i x_i$.

Por tanto, para que b_1^* sea insesgado se requiere que $\sum_{i=1}^{n} k_i=0$ y que $\sum_{i=1}^{n} k_i x_i=1$.

De esta manera se tiene que:

$$Var\left[b_1^*\right] = Var\sum_{i=1}^{n} k_i Y_i = \sum_{i=1}^{n} k_i Var\left[Y_i\right] = \sigma^2 \sum_{i=1}^{n} k_i \qquad (2.4.1.10)$$

$$= \sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i}{\sum_{i=1}^{n} x_i^2} + \frac{x_i}{\sum_{i=1}^{n} x_i^2} \right)^2$$

$$= \sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i}{\sum_{i=1}^{n} x_i^2} \right)^2 + \sigma^2 \frac{\sum_{i=1}^{n} x_i^2}{\left(\sum_{i=1}^{n} x_i^2 \right)^2} + 2\sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i}{\sum_{i=1}^{n} x_i^2} \right) \left(\frac{x_i}{\sum_{i=1}^{n} x_i^2} \right)$$

$$= \sigma^2 \sum_{i=1}^{n} \left(k_i - \frac{x_i}{\sum_{i=1}^{n} x_i^2} \right)^2 + \sigma^2 \frac{1}{\left(\sum_{i=1}^{n} x_i^2 \right)^2}$$

Como el segundo término es siempre constante, la varianza de b_1^* puede minimizarse únicamente manipulando el primer término y esto sucede siempre y cuando dicha expresión sea igual a cero, es decir siempre y cuando $k_i = \frac{x_i}{\sum_{i=1}^{n} x_i^2}$ donde se pude apreciar que:

$$Var\left[b_1^*\right] = \sigma^2 \frac{1}{\left(\sum_{i=1}^{n} x_i^2 \right)^2} = Var\left[b_1\right] \qquad (2.4.1.11)$$

Puede decirse que con ponderaciones $w_i = k_i$, que son las ponderaciones de mínimos cuadrados, la varianza del estimador lineal b_1^* es igual a la varianza del estimador b_1 de mínimos cuadrados, de otra manera, $Var(b_1^*) > Var(b_1)$. Con otras palabras, si hay un estimador lineal insesgado de β_1 de varianza mínima, éste debe ser el estimador de mínimos cuadrados. Igualmente se puede demostrar que b_0 es un estimador lineal insesgado con varianza mínima de β_0. En este contexto, cuando un estimador es insesgado y eficiente, entonces se dice que el estimador es MELI (Mejor Estimador Lineal Insesgado).

2.4.2 Los estimadores de Mínimos Cuadrados Ordinarios son los mismos que los de Máxima Verosimilitud.

Supóngase que en el modelo con dos variables $Y_i = \beta_0 + \beta_1 X_i + U_i$ las Y_i tienen una distribución normal e independiente, con una media igual a $\beta_0 + \beta_1 X_i$ y una varianza σ^2. Como resultado de esto, la función de densidad de probabilidad conjunta de $Y_1, Y_2, \ldots, Y_N$, con base en la media y la varianza antes referidas, puede escribirse como:

$$f(Y_1, Y_2, \ldots, Y_N | \beta_0 + \beta_1 X_i, \sigma^2) \qquad (2.4.2.1)$$

Pero en vista de la independencia de las Y_i, esta función de densidad de probabilidad conjunta se torna en el producto de *N* funciones de densidad individual, así:

$f(Y_1, Y_2, \ldots, Y_N | \beta_0 + \beta_1 X_i, \sigma^2)$ será equivalente a:

$$f(Y_1 | \beta_0 + \beta_1 X_i, \sigma^2) f(Y_2 | \beta_0 + \beta_1 X_i, \sigma^2) \cdots f(Y_N | \beta_0 + \beta_1 X_i, \sigma^2) \qquad (2.4.2.2)$$

donde:

$$f(Y_i) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2}\frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2}\right\} \qquad (2.4.2.3)$$

Lo cual no es otra cosa que la función de densidad de una variable con distribución normal con media y varianza dadas. (Nota: exp significa e elevado a la potencia de la expresión encerrada entre { }.)

Sustituyendo (2.4.2.3) para cada Y_i en (2.4.2.2) se obtiene:

$$f(Y_1, Y_2, \ldots, Y_N | \beta_0 + \beta_1 X_i, \sigma^2) = \frac{1}{\sigma^N (\sqrt{2\pi})^N} \exp\left\{ -\frac{1}{2} \Sigma \frac{(Y_i - \beta_0 - \beta_1 X_i)^2}{\sigma^2} \right\} \quad (2.4.2.4)$$

Si $Y_1, Y_2, \ldots, Y_N$ son valores conocidos, la función en (2.4.2.4) se denomina *función de máxima verosimilitud*, expresada como $L(\beta_0, \beta_1, \sigma^2)$, y escrita de la siguiente manera:

$$L(\beta_1, \beta_2, \sigma^2) = \frac{1}{\sigma^N (\sqrt{2\pi})^N} \exp\left\{ -\frac{1}{2} \sum \frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{\sigma^2} \right\} \quad (2.4.2.5)$$

El método de máxima verosimilitud, como su nombre lo indica, consiste en la estimación de parámetros desconocidos de tal manera que la probabilidad de observar un determinado valor de Y es la más elevada (o máxima) posible. En consecuencia, se debe hallar el máximo de la función (2.4.2.5). Este es un sencillo ejercicio de cálculo diferencial. Para diferenciar dicha ecuación es más fácil expresar (2.4.2.5) en términos logarítmicos, de la siguiente forma (*Nota*: ln = logaritmo natural):

$$\ln L = -N \ln \sigma - \frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum \frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{\sigma^2} \quad (2.4.2.6)$$

$$= -\frac{N}{2} \ln \sigma^2 - \frac{N}{2} \ln(2\pi) - \frac{1}{2} \sum \frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{\sigma^2}$$

Diferenciando (2.4.2.6) parcialmente con respecto a β_1, β_2 y σ^2, se obtiene:

$$\frac{\delta \ln L}{\delta \beta_1} = -\frac{1}{\sigma^2}\sum (Y_i - \beta_1 - \beta_2 X_i)(-1) \qquad (2.4.2.7)$$

$$\frac{\delta \ln L}{\delta \beta_2} = -\frac{1}{\sigma^2}\sum (Y_i - \beta_1 - \beta_2 X_i)(-X_i) \qquad (2.4.2.8)$$

$$\frac{\delta \ln L}{\delta \sigma^2} = -\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (Y_i - \beta_1 - \beta_2 X_i)^2 \qquad (2.4.2.9)$$

Igualando estas ecuaciones a cero (la condición de primer orden para optimización) y permitiendo que $\tilde{\beta}_1, \tilde{\beta}_2$ y $\tilde{\sigma}^2$ denoten los estimadores de máxima verosimilitud, entonces se obtendrá que:

$$\frac{1}{\tilde{\sigma}^2}\sum (Y_i - \tilde{\beta}_1 - \tilde{\beta}_2 X_i) = 0 \qquad (2.4.2.10)$$

$$\frac{1}{\tilde{\sigma}^2}\sum (Y_i - \tilde{\beta}_1 - \tilde{\beta}_2 X_i) X_i = 0 \qquad (2.4.2.11)$$

$$-\frac{N}{2\sigma^2} + \frac{1}{2\sigma^4}\sum (Y_i - \beta_1 - \beta_2 X_i)^2 = 0 \qquad (2.4.2.12)$$

De las ecuaciones 2.4.2.10 y 2.4.2.11 se llega a que:

$$\sum Y_i = N\tilde{\beta}_1 + \tilde{\beta}_2 \sum X_i \qquad (2.4.2.13)$$

$$\sum Y_i X_i = \tilde{\beta}_1 \sum X_i + \tilde{\beta}_2 \sum X_i^2 \qquad (2.4.2.14)$$

Las cuales coinciden con las *ecuaciones normales* de mínimos cuadrados que se obtuvieron con anterioridad. Por tanto, los estimadores de máxima verosimilitud son los mismos que los de mínimos cuadrados ordinarios. De hecho esta igualdad no es accidental, pues al examinar la función de verosimilitud 2.4.2.6 se puede observar que el último término entra con signo negativo. Por lo que, al maximizar dicha ecuación en realidad se está minimizando este término, sin embargo, éste es precisamente el enfoque de mínimos cuadrados.

Ahora bien, al sustituir los estimadores de máxima verosimilitud, que son los mismos de mínimos cuadrados ordinarios, en la ecuación 2.4.2.12 y simplificando la expresión resultante, se obtiene el estimador de máxima verosimilitud para $\tilde{\sigma}^2$:

$$\tilde{\sigma}^2 = \frac{1}{N}\Sigma\left(Y_i - \tilde{\beta}_1 - \tilde{\beta}_2 X_i\right)^2$$

$$= \frac{1}{N}\Sigma\left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

$$= \frac{1}{N}\sum e_i^2 \qquad \text{(2.4.2.15)}$$

Esto implica que el estimador $\tilde{\sigma}^2$ de máxima verosimilitud difiere del estimador de mínimos cuadrados $\hat{\sigma}^2 = \sum \frac{e_i^2}{N-2}$, obteniéndose con el método de MV un estimador sesgado hacia abajo de σ^2.

2.4.3 El estimador de la varianza por MCO es insesgado mientras que el de MV es sesgado.

A pesar de que el método de mínimos cuadrados ordinarios (MCO) es mucho más sencillo en la práctica que el de máxima verosimilitud (MV), proporciona un mejor estimador para la varianza del modelo estimado que el de MV. ¿Pero esto qué significa? Pues significa que en un muestreo aleatorio repetido, en promedio, por el método de mínimos cuadrados ordinarios se obtendrá un estimador de la varianza más "real" o más cercano al verdadero valor de dicha varianza que por el método de máxima verosimilitud, de esta manera, el investigador tendrá más certeza en cuanto a la variabilidad de su modelo.

Para ver que el estimador de la varianza por MCO es insesgado, primero se procederá al cálculo de dicha varianza.

Recuérdese que $Y_i = \hat{Y}_i + e_i$ lo cual implica que $Y_i - \overline{Y} = \hat{Y}_i - \overline{Y} + e_i$ y finalmente en términos de desviaciones se obtiene que:

$$y_i = \hat{y}_i + e_i \qquad (2.4.3.1)$$

donde $\hat{y}_i = \hat{Y}_i - \overline{Y}$ y $y_i = Y_i - \overline{Y}$.

Ahora bien, recordando que:

$$Y_i = \beta_0 + \beta_1 X_i + U_i \qquad (2.4.3.2)$$

se obtiene:

$$\overline{Y}_i = \beta_0 + \beta_1 \overline{X}_i + \overline{U}_i \qquad (2.4.3.3)$$

Restando (2.4.3.3) de (2.4.3.2) se tiene lo siguiente:

$$y_i = \beta_1 x_i + (U_i - \overline{U}_i) \qquad (2.4.3.4)$$

Por otra parte, recordando que:

$$\hat{Y}_i = b_0 + b_1 X_i \qquad (2.4.3.5)$$

entonces se tiene:

$$\overline{Y}_i = b_0 + b_1 \overline{X}_i \qquad (2.4.3.6)$$

Volviendo a restar, pero ahora (2.4.3.6) de (2.4.3.5) entonces:

$\hat{Y}_i - \overline{Y} = b_1 (X_i - \overline{X})$ lo cual no es otra cosa más que

$$\hat{y}_i = b_1 x_i \qquad (2.4.3.7)$$

De esta manera, la ecuación (2.4.3.1) es equivalente a:

$$e_i = y_i - b_1 x_i \qquad (2.4.3.8)$$

Si ahora se sustituye (2.4.3.4) en (2.4.3.8) se obtiene que:

$$e_i = \beta_1 x_i + (U_i - \overline{U}_i) - b_1 x_i \qquad (2.4.3.9)$$

lo cual implica

$$e_i^2 = \left[(U_i - \overline{U}_i) - (b_1 - \beta_1) x_i + \right]^2$$

$$\sum e_i^2 = (b_1 - \beta_1)^2 \sum x_i^2 + \sum (U_i - \overline{U}_i)^2 - 2(b_1 - \beta_1) \sum x_i (U_i - \overline{U}_i) \qquad (2.4.3.10)$$

Si ahora se toma el valor esperado en ambos lados

$$E \sum e_i^2 = E\left[(b_1 - \beta_1)^2 \sum x_i^2 + \sum (U_i - \overline{U}_i)^2 - 2(b_1 - \beta_1) \sum x_i (U_i - \overline{U}_i) \right] \qquad (2.4.3.11)$$

Apelando a los supuestos del modelo de regresión lineal clásico, es posible verificar que:

$$E \sum e_i^2 = \sigma^2 + (n-1)\sigma^2 - 2\sigma^2 = (n-1)\sigma^2 - \sigma^2 = (n-2)\sigma^2$$

Si se define $\hat{\sigma}^2 = \frac{\sum e_i^2}{n-2}$ entonces su valor esperado será:

$$E\left(\hat{\sigma}^2\right) = E\left(\frac{\sum e_i^2}{n-2}\right) = \frac{1}{n-2} E\left(\sum e_i^2\right) = \sigma^2 \qquad (2.4.3.12)$$

lo cual muestra que $\hat{\sigma}^2$ de MCO es un estimador insesgado del verdadero σ^2.

A continuación se procederá a mostrar que el estimador $\tilde{\sigma}^2$ de la varianza por el método de Máxima Verosimilitud es sesgado, de hecho, este estimador presenta un sesgo hacia abajo del verdadero valor de σ^2.

Recordando que el estimador de la varianza calculado por este método viene dado por $\tilde{\sigma}^2 = \frac{1}{n}\sum e_i^2$ entonces se tiene que:

$$\begin{aligned} E\left[\tilde{\sigma}^2\right] &= E\left[\frac{1}{n}\sum e_i^2\right] \\ &= \frac{1}{n} E\left[\sum e_i^2\right] \\ &= \left(\frac{n-2}{n}\right)\sigma^2 \\ &= \sigma^2 - \frac{2}{n}\sigma^2 \end{aligned} \qquad (2.4.3.13)$$

lo cual confirma que el estimador de la varianza calculado por el método de MV es sesgado hacia abajo, sin embargo puede apreciarse que cuando el tamaño de la muestra *n* crece indefinidamente entonces el valor de dicho parámetro coincide con el verdadero valor de la varianza σ^2.

2.5 La bondad de ajuste o coeficiente de determinación r^2

2.5.1 Medidas absolutas de la bondad de ajuste

Ya se ha mostrado que con el método de los Mínimos Cuadrados Ordinarios se obtienen los mejores estimadores lineales insesgados de acuerdo con el criterio de la minimización de los residuos, sin embargo esto no es garantía de que la ecuación estimada, por diversas razones, sea la mejor.

Ahora bien, considérense los puntos o parejas de las variables estudiadas para los cuales se estimó la línea $\hat{Y}_i = b_0 + b_1 X_i$ y considérese también la media de X y de Y como se presenta en la figura 2.5.

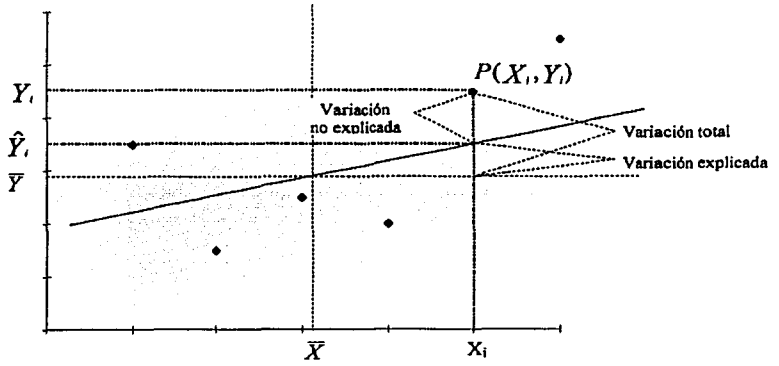


Figura 2.5 Variaciones absolutas

Considérese el punto P, entonces, la desviación a $\overline{Y}$ se escribe como $Y_i - \overline{Y}$ y parte de esta desviación puede ser explicada por la ecuación estimada en la regresión, la cual no da, generalmente, una justificación total de la variación. De esta manera, a la parte no explicada se le llama "variación no explicada", las cuales pueden ser expresadas como:

$(\hat{Y}_i - \overline{Y})$ = desviación explicada

$(Y_i - \hat{Y}_i)$ = desviación no explicada

Por lo tanto para un punto específico P se tiene que:

$$Y_i = \hat{Y}_i + e_i \qquad (2.5.1.1)$$

si ahora se suma para todos los puntos de la regresión se obtiene:

$$\sum Y_i = \sum \hat{Y}_i + \sum e_i \qquad (2.5.1.2)$$

Sin embargo, es fácil comprobar que $\sum e_i = 0$, pues con base en la expresión (2.5.1.1) se tiene que:

$e_i = Y_i - \hat{Y}_i$ o lo que es lo mismo

$e_i = Y_i - (b_0 + b_1 X_i) = Y_i - b_0 - b_1 X_i$

si se vuelve a sumar sobre todos los puntos de la regresión entonces:

$$\sum e_i = \sum (Y_i - b_0 - b_1 X_i)$$

$$= n\overline{Y} - nb_0 - nb_1\overline{X} \qquad \text{pero} \quad b_0 = \overline{Y} - b_1\overline{X}$$

lo cual implica que

$$\sum e_i = n\overline{Y} - n(\overline{Y} - b_1\overline{X}) - nb_1\overline{X}$$

$$= n\overline{Y} - n\overline{Y} + nb_1\overline{X} - nb_1\overline{X} = 0$$

Por lo tanto

$$\sum Y_i = \sum \hat{Y}_i \qquad (2.5.1.3)$$

Y si ahora se dividen ambos lados entre el número total de observaciones de la regresión entonces se observa que:

$$\overline{Y} = \overline{\hat{Y}} \qquad (2.5.1.4)$$

Por otra parte, si se eleva al cuadrado la expresión (2.5.1.1):

$$(Y_i)^2 = (\hat{Y}_i + e_i)^2 = \hat{Y}_i^2 + e_i^2 + 2\hat{Y}_i e_i$$

lo cual implica que

$$\sum Y_i^2 = \sum \hat{Y}_i^2 + \sum e_i^2 + 2\sum \hat{Y}_i e_i \qquad (2.5.1.5)$$

nuevamente es fácil comprobar que $\sum \hat{Y}_i e_i = 0$, pues recordando que $\hat{Y}_i = b_o + b_1 X_i$ entonces

$$\sum \hat{Y}_i e_i = \sum e_i (b_0 + b_1 X_i) = \sum (e_i b_0 + e_i b_1 X_i)$$

$$= \sum e_i b_0 + \sum e_i b_1 X_i$$

$$= b_0 \sum e_i + b_1 \sum e_i X_i$$

pero se sabe que $\sum e_i = 0$ entonces

$$\sum \hat{Y}_i e_i = b_1 \sum e_i X_i$$

por lo tanto $\sum Y_i^2 = \sum \hat{Y}_i^2 + \sum e_i^2$ (2.5.1.6)

y si ahora se resta en ambos lados $n\overline{Y}^2$ se obtiene que

$$\sum Y_i^2 - n\overline{Y}^2 = (\sum \hat{Y}_i^2 - n\overline{Y}^2) + \sum e_i^2 \qquad (2.5.1.7)$$

y como $\overline{Y} = \overline{\hat{Y}}$ entonces

$$\sum\left(Y_i - \overline{Y}\right)^2 = \sum\left(\hat{Y}_i - \overline{\hat{Y}}\right)^2 + \sum e_i^2 \qquad (2.5.1.8)$$

si se define $\hat{y}_i = \hat{Y}_i - \overline{\hat{Y}}$ entonces se tendrá finalmente que en términos de desviaciones

$$\sum y_i^2 = \sum \hat{y}_i^2 + \sum e_i^2 \qquad (2.5.1.9)$$

Lo que en palabras significa que: La suma de los cuadrados totales es igual a la suma de los cuadrados de los residuos más la suma de los cuadrados de los errores.

2.5.2 Coeficiente de determinación

Si la expresión 2.5.1.9 se divide en ambos lados por $\sum y_i^2$ entonces se tiene que:

$$1 = \frac{\sum \hat{y}_i^2}{\sum y_i^2} + \frac{\sum e_i^2}{\sum y_i^2} \qquad (2.5.2.1)$$

Donde la razón $\frac{\sum \hat{y}_i^2}{\sum y_i^2}$ se denomina coeficiente de determinación y el cual es denotado por r^2 . En términos generales, este coeficiente mide la proporción de la variación total que ha sido explicada por la variable explicativa X_i.

Si X_i explicara la totalidad de la variación de Y_i, es decir, si el ajuste fuera perfecto, entonces el coeficiente de determinación sería igual a 1. El otro caso extremo se refiere cuando X_i no tiene ninguna relación con Y_i, o sea que variaciones en X_i no explican las variaciones producidas en Y_i y en este caso $r^2 = 0$.

Por lo tanto, entre más se aproxime el r^2 a 1 más significativa es la variable exógena en la explicación de Y.

2.5.3 El r^2 Ajustado

El coeficiente de determinación se ajusta por grados de libertad ya que el r^2 original está sesgado hacia arriba.

De la expresión 2.5.2.1 se deduce que

$$r^2 = 1 - \frac{\sum e_i^2}{\sum y_i^2} \qquad (2.5.3.1)$$

Por lo que al dividir ambos términos por sus correspondientes grados de libertad se obtiene el r^2 ajustado como:

$$r^2 \text{ ajustado} = 1 - \frac{\sum e_i^2/(n-2)}{\sum y_i^2/(n-1)} = 1 - \frac{S_e^2}{S_y^2} \qquad (2.5.3.2)$$

Donde una expresión alternativa viene dada por:

$$r^2 \text{ ajustado} = 1 - (1 - r^2)\frac{(n-1)}{(n-2)} \qquad (2.5.3.3)$$

3 Violación a los supuestos del análisis de regresión

3.1 La naturaleza de la multicolinealidad

Uno de estos problemas se refiere al término conocido como multicolinealidad, este concepto se atribuye a Ragnar Frisch[6]. Originalmente esto implicaba la existencia de una relación lineal "perfecta o exacta" entre algunas o la totalidad de las variables explicativas de un modelo de regresión.

Hablando estrictamente, el término *multicolinealidad* se refiere a la existencia de más de una relación lineal exacta, mientras que el término *colinealidad* se refiere a la existencia de una sola relación lineal. Sin embargo, esta distinción raramente se mantiene en la práctica, hablándose entonces de multicolinealidad para ambos casos.

Para la regresión en *k* variables, con las variables explicativas X_1, X_2,...., X_k (donde $X_1 = 1$ para todas las observaciones de tal manera que se permita la inclusión de la intersección), se dice que existe una relación lineal exacta si se satisface lo siguiente:

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots\ldots + \lambda_k X_k = 0 \qquad (3.1.1)$$

donde $\lambda_1, \lambda_2, \ldots\ldots, \lambda_k$ son constantes no necesariamente todas iguales a cero.

No obstante, existe aquella situación en donde la multicolinealidad entre las variables explicativas no es perfecta. Es decir, cuando:

$$\lambda_1 X_1 + \lambda_2 X_2 + \ldots\ldots\ldots + \lambda_k X_k + v_i = 0 \qquad (3.1.2)$$

donde v_i es un término estocástico de error.

[6] Ragnar Frisch, *Statistical Confluence Analysis by Means of Complete Regression Systems*, Institute of Economica, Oslo University, publ. No. 5, 1934.

Para apreciar la diferencia que existe entre multicolinealidad perfecta y multicolinealidad menos que perfecta, supóngase que $\lambda_2 \neq 0$ entonces para la expresión 3.1.1 se tiene que:

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \ldots\ldots - \frac{\lambda_k}{\lambda_2} X_{ki} \qquad (3.1.3)$$

lo cual implica que X_2 está exactamente relacionada en forma lineal con otras variables o que se puede derivar a partir de la combinación lineal de las restantes variables X's. En esta situación, el coeficiente de correlación entre la variable X_2 y la combinación lineal del lado derecho de la expresión 3.1.3 será igual a 1.

Similarmente, el $\lambda_2 \neq 0$ implica que la expresión 3.1.2 puede escribirse como:

$$X_{2i} = -\frac{\lambda_1}{\lambda_2} X_{1i} - \frac{\lambda_3}{\lambda_2} X_{3i} - \ldots\ldots - \frac{\lambda_k}{\lambda_2} X_{ki} - \frac{1}{\lambda_2} v_i \qquad (3.1.4)$$

donde se aprecia que X_2 no es exactamente una combinación lineal de las restantes X's debido a que también está determinada por el término del error estocástico v_i.

El punto de partida para el diagnóstico es el modelo de regresión lineal básico[7]:

$$\underset{n\times 1}{Y} = \underset{n\times k}{X}\ \underset{k\times 1}{b} + \underset{n\times 1}{u} \qquad (3.1.5)$$

donde Y y X sintetizan n puntos de datos para la variable dependiente y para las k variables explicativas, respectivamente.

[7] Véase Intriligator, Michael D. "Modelos econométricos, técnicas y aplicaciones". Editorial Fondo de Cultura Económica, México 1990, pág. 176-177.

Entonces, el estimador mínimo cuadrático es:

$$\hat{b} = (X'X)^{-1} X'Y \qquad (3.1.6)$$

que es también el estimador máximo verosímil si los términos de perturbación estocástica están normalmente distribuidos y se cumplen los demás supuestos del análisis de regresión.

En el modelo de los mínimos cuadrados ordinarios, uno de los supuestos fundamentales es el de que la matriz inversa $(X'X)^{-1}$ existe. Esto se traduce en que los vectores columna de las observaciones de las variables independientes no son linealmente dependientes, o lo que es lo mismo, la matriz $(X'X)$ tiene rango completo, es decir, $\rho(X) = k$. Esto también implica que ningún vector de observaciones puede ser expresado como una combinacion lineal de los otros vectores.

Por lo tanto un supuesto básico de los mínimos cuadrados ordinarios es el de que las variables explicativas son independientes, y por lo tanto tienen un efecto sobre la variable dependiente que puede ser separado para cada variable independiente.

De modo que son linealmente independientes las columnas de la matriz de datos sobre las variables explicativas X de dimensión n x k. Bajo este supuesto se deduce que $(X'X)$ es no singular, de manera que puede invertirse para obtener el estimador mínimo cuadrático $\hat{b}$ en (3.1.6). No obstante, si una columna de X es una combinación lineal de otras columnas de la matriz, entonces se viola la condición de rango, es decir $\rho(X) < k$, lo cual implica que $|X'X| = 0$.

De forma tal que $(X'X)$ es una matriz singular y no puede ser invertida. Esta situación se denomina *multicolinealidad perfecta* y en ella las ecuaciones normales de mínimos cuadrados no pueden resolverse para los estimadores $\hat{b}$.

Sin embargo, la situación más característica no es la de *multicolinealidad perfecta* sino aquella considerada como un problema de *multicolinealidad*,

en cuyo caso aunque no puede decirse que $(X'X)$ sea singular, "casi lo es" en el sentido de que $|X'X| \approx 0$.

En este caso, los datos sobre las variables explicativas tienen la propiedad de que no obstante que ninguno es una combinación lineal *exacta* de los otros, los valores de uno o más de ellos, están *casi* dados por tal tipo de combinación lineal en los valores de los otros. Esta situación, bajo la cual las variables explicativas tienden a moverse juntas, ocurre muy a menudo en los estudios prácticos, particularmente en aquellos que utilizan datos en series de tiempo. En realidad, el problema de *multicolinealidad* es uno de los problemas más ubicuos, significativos y difíciles en econometría aplicada. Por su propia naturaleza, los datos económicos tienden a desplazarse juntos, reflejando a menudo factores subyacentes comunes tales como las tendencias y los ciclos.

El problema de *multicolinealidad* es un *problema muestral* que se presenta porque la muestra no ofrece información suficientemente "rica" sobre las variables explicativas como para cumplir con los requisitos del modelo. Si hubiese experimentación controlada disponible, podría generarse un conjunto de datos "más ricos", en el sentido de contener un mayor grado de variación en el comportamiento de las variables explicativas.

3.2 Estimación en presencia de multicolinealidad

¿Porqué supone el modelo clásico de regresión lineal que no existe multicolinealidad entre las X's? El razonamiento es el siguiente; si la multicolinealidad es perfecta entonces los coeficientes de regresión de las variables X's son indeterminados y sus errores estándares son infinitos. Si la multicolinealidad es *menos* que perfecta, los coeficientes de regresión aunque determinados o finitos, poseen errores estándares demasiado grandes (en relación con los coeficientes mismos), lo cual implica que los coeficientes no se pueden estimar con gran precisión o exactitud.

Utilizando la estimación del modelo en forma de desviaciones de sus medias muestrales, veamos como ejemplo el modelo de regresión con tres variables:

$$y_i = b_2 x_{2i} + b_3 x_{3i} + e_i \qquad (3.2.1)$$

se puede demostrar entonces que:

$$b_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\sum x_{3i}^2\right) - \left(\sum y_i x_{3i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2} \tag{3.2.2}$$

$$b_3 = \frac{\left(\sum y_i x_{3i}\right)\left(\sum x_{2i}^2\right) - \left(\sum y_i x_{2i}\right)\left(\sum x_{2i} x_{3i}\right)}{\left(\sum x_{2i}^2\right)\left(\sum x_{3i}^2\right) - \left(\sum x_{2i} x_{3i}\right)^2} \tag{3.2.3}$$

Supóngase ahora que $X_{3i} = \lambda X_{2i}$ donde λ es una constante diferente de cero. Entonces sustituyendo esto en 3.2.2 se obtiene que:

$$b_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\lambda^2 \sum x_{2i}^2\right) - \left(\lambda \sum y_i x_{2i}\right)\left(\lambda \sum x_{2i}^2\right)}{\left(\sum x_{2i}^2\right)\left(\lambda^2 \sum x_{2i}^2\right) - \lambda^2 \left(\sum x_{2i}^2\right)^2} \tag{3.2.4}$$

la cual, claramente es una expresión indeterminada. Esto muestra que si X_2 y X_3 son perfectamente colineales, no existe forma alguna de mantener a X_3 constante a medida que X_2 cambia, también lo hace X_3 en un factor igual a λ. Lo que esto significa entonces es que no existe manera alguna de separar las influencias individuales de X_2 y X_3 de la muestra dada, para fines prácticos, X_2 y X_3 no se pueden diferenciar. En econometría aplicada, éste es un problema muy serio puesto que la idea consiste en separar los efectos parciales de cada X_i sobre la variable dependiente.

El caso de la *multicolinealidad perfecta* es un extremo de tipo *patológico*. Generalmente no existe una relación lineal exacta entre las variables X`s, especialmente con información relacionada con series de tiempo en Economía. Por tanto, regresando al modelo con tres variables, en lugar de tener multicolinealidad exacta, se puede tener que:

$$X_{3i} = \lambda X_{2i} + v_i \tag{3.2.5}$$

donde $\lambda_2 \neq 0$ y v_i es un término de error estocástico, tal que $\sum x_{2i} v_i = 0$.

En este caso es factible la estimación de los coeficientes de regresión b_2 y b_3, por ejemplo, sustituyendo 3.2.5 en la solución para b_2 , en el caso del modelo de tres variables, se obtiene que:

$$b_2 = \frac{\left(\sum y_i x_{2i}\right)\left(\lambda^2 \sum x_{3i}^2 + \sum v_i^2\right) - \left(\lambda \sum y_i x_{3i} + \sum y_i v_i\right)\left(\lambda \sum x_{2i}^2\right)}{\left(\sum x_{2i}^2\right)\left(\lambda^2 \sum x_{2i}^2 + \sum v_i^2\right) - \left(\lambda \Sigma x_{2i}^2\right)^2} \quad (3.2.6)$$

donde se utiliza el hecho de que $\sum x_{2i} v_i = 0$. Ahora, no existe razón alguna para creer *a priori* que la expresión 3.2.6 no se puede estimar. Desde luego que si v_i es suficientemente pequeña, es decir, muy cercana a cero, la expresión 3.2.5 indicará la existencia una colinealidad *casi* perfecta, retornando entonces al caso indeterminado de 3.2.4.

El problema de la multicolinealidad consiste en que se tiene en la especificación del modelo una alta correlación entre los vectores de observaciones de las variables independientes, lo cual se aproxima al caso de una dependencia lineal perfecta, pero sin que ésta llegue a ser perfecta. En este caso, un valor pequeño para el determinante de la matriz $(X'X)$ es obtenido y se puede determinar la matriz inversa $(X'X)^{-1}$.

Sin embargo su sensibilidad es alta, aún para cambios muy pequeños en las observaciones debido a la forma como se calcula la matriz inversa:

matriz inversa = matriz adjunta / determinante

Por lo tanto, cambios pequeños en el valor del determinante pueden inducir cambios grandes en la matriz inversa cuando existe multicolinealidad, y consecuentemente en el vector de coeficientes de los estimadores $\hat{b}$.

Así, la multicolinealidad es un problema en el sentido de que los estimadores del modelo se vuelven altamente sensibles, tanto al conjunto de datos usados para estimarlos, como a la especificación del modelo. También surgen problemas en la interpretación del modelo, pues algunos coeficientes absorben el efecto de otras variables correlacionadas con éstas.

3.3 Consecuencias de la multicolinealidad

Los casos de casi multicolinealidad o de alta multicolinealidad pueden traer consigo las siguientes consecuencias:

a) Varianzas y covarianzas amplias para los estimadores de MCO

Para poder apreciar esto, se tiene que para el modelo de regresión de tres variables las varianzas y las covarianzas de b_2 y b_3 están dadas por:

$$\text{var}(b_2) = \frac{\sigma^2}{\sum x_{2i}^2 (1 - r_{23}^2)} \tag{3.3.1}$$

y

$$\text{var}(b_3) = \frac{\sigma^2}{\sum x_{3i}^2 (1 - r_{23}^2)} \tag{3.3.2}$$

$$\text{cov}(b_2, b_3) = \frac{-r_{23}\sigma^2}{(1 - r_{23}^2)\sqrt{\sum x_{2i}^2 \sum x_{3i}^2}} \tag{3.3.3}$$

donde r_{23} es el coeficiente de correlación entre X_2 y X_3. Es claro que a medida que r_{23} tiende a 1, o sea, a medida que aumenta la colinealidad, las varianzas de los dos estimadores aumentan y, en el límite ellas son infinitas. También es evidente que a medida que r_{23} aumenta hacia 1, la covarianza de los estimadores también aumenta en valores absolutos.

b) Más amplios intervalos de confianza

Debido a la presencia de errores estándar grandes, los intervalos de confianza para los parámetros poblacionales relevantes tienden a ser más amplios. Esto conlleva a que en casos de alta multicolinealidad la información muestral puede ser compatible con un conjunto diverso de hipótesis y por consiguiente, la posibilidad de aceptar una hipótesis falsa (es decir, la probabilidad de cometer el error tipo I) aumenta considerablemente.

c) Razones *t* "no significativas"

Recuérdese de que para evaluar la hipótesis nula de que, por ejemplo $\beta = 0$, se utiliza el estadístico de prueba ***t***: $\frac{b_1}{se(b_1)}$ y comparamos el valor estimado de ***t*** con el valor crítico de ***t***, utilizando la tabla ***t***. Sin embargo, en los casos de alta colinealidad los errores estándar estimados aumentan dramáticamente disminuyendo con esto los valores de t. Por lo tanto, en tales casos se tiende a aceptar con mayor facilidad la hipótesis nula de que el verdadero valor poblacional relevante es cero.

d) Un valor elevado para el r^2 pero pocas razones *t* significativas

Considérese el modelo de regresión lineal con k variables:

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots\ldots\ldots + \beta_K X_{Ki} + u_i$$

Ahora bien, en casos de alta colinealidad es posible encontrar uno o más coeficientes parciales de las pendientes que sean individualmente no significativos en términos estadísticos con base en la prueba ***t***. No obstante, el r^2 en tales situaciones puede ser tan elevado, alrededor de 0.9, que con base en la prueba ***F*** se puede convincentemente rechazar la hipótesis de que $\beta_2 = \beta_3 = \beta_4 = \ldots\ldots = \beta_k = 0$. De hecho, esto es un indicador de multicolinealidad, o sea, valores t no significativos pero un elevado r^2 así como un valor de ***F*** bastante alto.

3.4 ¿Es la multicolinealidad necesariamente mala?

Si la única finalidad del modelo regresivo es el pronóstico o la predicción, entonces la colinealidad no se constituye en un problema serio porque cuanto mayor sea el r^2, mejor será la predicción[8]. Si en una regresión estimada se encontró que $X_2 = 2X_3$ aproximadamente, entonces en una futura muestra utilizada para predecir Y, X_2 debería también ser

[8] R. C. Geary, "Some Results about Relations between Stochastic Variables: A Discussion Document", Review of *International Statistical Institute*, vol. 31, 1963, pp. 163-181.

aproximadamente igual a $2X_3$, lo cual es una condición difícil de cumplir en la práctica, en cuyo caso la predicción se tornará cada vez más incierta. Adicionalmente, si el objetivo del análisis no es solamente la predicción sino también la estimación confiable de los parámetros, la presencia de alta multicolinealidad puede ser un problema serio debido a que como se ha visto, puede producir grandes errores estándar para los estimadores.

Sin embargo, existe un caso en donde la multicolinealidad puede no representar un problema serio. Este es el caso cuando se tiene un elevado R^2 y en donde los coeficientes de regresión son individualmente significativos, como demuestran los altos valores t. Aún así, los diagnósticos de la multicolinealidad, por ejemplo el que arrojaría el índice de condición, indican que existe una seria colinealidad en los datos. Pero ¿Cuándo puede presentarse tal situación? Como menciona Johnston:

Lo anterior puede surgir si los coeficientes individuales resultan estar numéricamente por encima del valor verdadero, de tal manera que el efecto se siga mostrando, a pesar de estar "inflados" los errores estándar y/o debido a que el valor verdadero mismo es tan grande que aún cuando se obtenga una estimación bastante subestimada, ésta continúe siendo significativa[9].

3.5 Métodos para detectar el problema de multicolinealidad

Aunque hoy en día aún no se cuenta con métodos exactos para detectar colinealidad, existen diferentes indicadores de ésta, tales como:

- ☑ La existencia de colinealidad tiene lugar cuando el r^2 es muy alto, pero ninguno de los coeficientes de regresión es estadísticamente significativo, con base en la tradicional prueba ***t***., esto desde luego, es un caso extremo.

- ☑ En el modelo con dos variables explicativas se puede obtener una idea relativamente buena de la colinealidad examinando el coeficiente de correlación simple o de orden cero entre las dos variables. Si esta correlación es alta, entonces existen claras evidencias de colinealidad.

[9] J. Johnston, *Econometric Methods*, 3a. edición, McGraw-Hill Book Company, New York, 1984, p. 249.

☑ Sin embargo, los coeficientes de correlación de orden cero pueden ser engañosos en modelos con más de dos variables explicativas, puesto que existe la posibilidad de tener pequeñas correlaciones de orden cero y existir simultáneamente una alta multicolinealidad. En situaciones como ésta, es necesario examinar los coeficientes de correlación parcial entre todas las variables involucradas en el modelo.

☑ Aún cuando el r^2 es alto pero las correlaciones parciales son bajas, no es posible dictaminar la ausencia de multicolinealidad. En este caso pueden ser superfluas una o más variables. Pero si el r^2 es alto y las correlaciones parciales son también elevadas, entonces la multicolinealidad no puede ser detectada tan directamente.

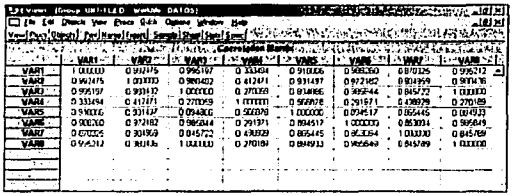
☑ No obstante, se puede proceder a correr una regresión de cada una de las variables independientes sobre las X's restantes del modelo y calcular los correspondientes coeficientes de determinación r_i^2. De esta manera, un alto r_i^2 sugeriría que la variable X_i está altamente correlacionada con el resto de las X's. En consecuencia se puede proceder a eliminar esa X_i del modelo, dado que no produce un serio sesgo de especificación en el modelo.

☑ En el caso de que dos variables sean independientes, se esperaría que al introducir una nueva variable en el modelo no cambie significativamente el valor de los coeficientes de las otras.

Para apreciar el grado de dependencia lineal entre ***n*** variables el paquete estadístico *Econometric Views* ofrece las siguientes herramientas:

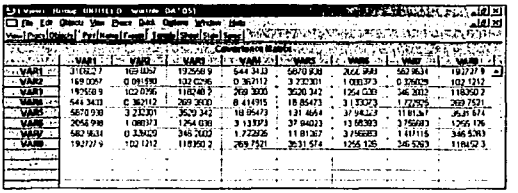
Matriz de correlación

La *matriz de correlación* muestra el coeficiente de correlación parcial entre cada par de variables y tiene un resultado como el siguiente:



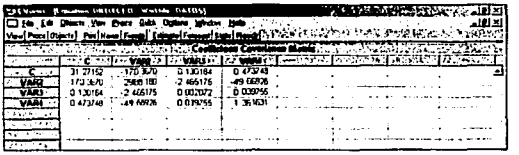
Matriz de varianza-covarianza

La *matriz de varianza-covarianza* muestra el grado de codependencia lineal entre dos variables distintas, y en su diagonal mostrará la varianza de dicha variable. El resultado de salida será como el siguiente:



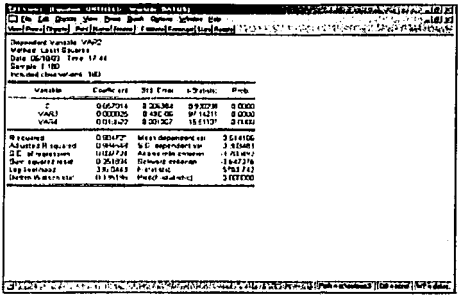
Matriz de covarianza

Mientras que la *matriz de covarianza* para los coeficientes de la regresión VAR1 = C + VAR2 + VAR3 + VAR4 tendrá un resultado como el siguiente:



Prueba de Klein

Para medir el grado de dependencia entre las variables involucradas en la regresión, se deben de hacer tantas regresiones como variables explicativas se hayan considerando en el modelo, por ejemplo, si se quiere apreciar la relación que tiene la variable VAR2 con las variables VAR3 y VAR4 entonces debe de hacerse la regresión VAR2 = C + VAR3 + VAR4, de esta manera el resultado del paquete estadístico será como el siguiente:



Al obtenerse un r^2 muy elevado se deduce que la variable VAR2 está altamente correlacionada con las variables VAR3 y VAR4.

3.6 Medidas del efecto de la multicolinealidad

La manera más simple de medir el efecto de multicolinealidad es calculando la contribución incremental de cada variable independiente en la explicación de la variable dependiente Y, una vez que se hayan incluido otras variables; sumamos estas contribuciones y la comparamos con la contribución conjunta de todas las variables.

Una forma de encontrar estas medidas proporcionales es usando el coeficiente de determinación r^2 para hallar así la contribución total de todas las variables en la explicación de Y y obtener, además, la contribución incremental de una variable X_i dada por:

$$\psi_i = (1 - R^2)\frac{F_i}{n-i} \qquad (3.7.1)$$

donde F_i es el valor del estadístico **F** para la variable X_i en cuestión y n es el número de las observaciones.

El efecto se calcula entonces como:

$$\left[\sum_{i=2}^{n}\psi_i - R^2\right] = \overline{M} \qquad (3.7.2)$$

Solamente en el caso de que todas las variables sean independientes, se tendrá que la suma de las contribuciones incrementales será igual a la contribución total y en consecuencia $\overline{M} = 0$. Para cualquier otro caso el valor de $\overline{M}$ será diferente de cero.

Por otra parte, el valor de $\overline{M}$ puede ser positivo o negativo y cuando su valor absoluto sea grande entonces se tendrá un indicio de que la multicolinealidad existente será severa, o sea, cuando $|\overline{M}|$ se acerque a 1. Este nivel de severidad estará dado por el r^2, pues cuanto más pequeño sea este coeficiente más severa será la multicolinealidad.

3.7 Cómo corregir el problema de la multicolinealidad

Existen varias maneras para tratar el problema de la multicolinealidad, y entre las cuales se citan las siguientes:

- Cambio en la especificación del modelo. Este método consiste en seleccionar la variable más severamente afectada y excluirla del modelo. Sin embargo es preciso tener en cuenta que aunque se reduce el problema de la multicolinealidad, se puede introducir un nuevo error en la especificación del modelo cuando la variable es importante y significativa como variable explicativa de Y.

- Excluir la variable afectada y reemplazarla por otra que mida el mismo fenómeno pero, que no sea colineal con ninguna de las otras variables independientes.

- Una manera de cambiar la especificación del modelo es expresando las interrelaciones entre las variables predeterminadas como ecuaciones separadas y construir así un modelo con ecuaciones simultáneas.

- Redefinir variables. A veces es factible solucionar el problema al transformar o agregar variables. Este procedimiento es adecuado en la medida en que el nuevo modelo refleje todavía las relaciones teóricas del modelo inicial.

Las transformaciones más comunes son primeras diferencias ordinarias o logarítmicas, como a continuación se muestra:

$$i)\ X_t^* = X_t - X_{t-1} \tag{3.7.1}$$

$$ii)\ X_t^* = Ln\left(X_t / X_{t-1}\right) \tag{3.7.2}$$

La transformación logarítmica se usa con frecuencia para eliminar el efecto de tendencias cíclicas en series de tiempo.

El tipo de agregación de variables usado más comúnmente es el de agrupar las variables con multicolinealidad en un índice compuesto que permita una interpretación teórica similar.

- Adquirir datos adicionales, es decir mejorar la muestra en observaciones.
- Uso de técnicas de cálculo más aproximado y evitar así errores de redondeo.

3.8 La naturaleza de la heteroscedasticidad

Uno de los supuestos importantes en el modelo clásico de regresión lineal es que la varianza de cada término de perturbación e_i, condicionado a los valores escogidos de las variables explicativas, es un número constante igual a $\sigma_{e_i}^2 = \sigma^2$. Este es el supuesto de homoscedasticidad o igual (homo) dispersión (cedasticidad), o sea, igual varianza.

La heteroscedasticidad es el problema que se presenta al no cumplirse la condición de que la varianza de e_i sea constante, para todo e_i, con i=1,2,....,n, es decir, cuando no se cumple el supuesto de que $Var(e_i) = \sigma_{e_i}^2 = \sigma^2$.

3.9 Situaciones en las que se presenta el problema

Si no se mantiene la suposición de MCO (Mínimos Cuadrados Ordinarios) de que la varianza del término de error es constante para todos los valores de las variables independientes, entonces se está ante el problema de la heteroscedasticidad (diferente varianza). Las situaciones típicas en las que se presenta este problema pueden ser:

- Cuando cambian las condiciones estructurales del modelo, de manera que la especificación cambia para diferentes grupos de observaciones. La violación del supuesto puede deberse a cambios en los factores estructurales no especificados en el modelo; aunque éstos se suponen constantes, pueden cambiar, debido a variaciones en las leyes, en la tecnología empleada, fluctuaciones en el comportamiento, etc.

- El supuesto de homoscedasticidad se viola frecuentemente cuando en la estimación de un modelo se usan datos transversales. El error que se cometa en la especificación de una sección puede ser muy diferente del error cometido en otra sección[10], y donde los factores no incluidos son importantes en las dos secciones.

[10] Al decir *sección* se está haciendo énfasis a una región, departamento, estrato, etc.

A veces la violación del supuesto de igual varianza está asociado con errores de medición. Cuando existen errores de medición de las variables, de manera que, algunas observaciones, están mejor medidas que otras, la varianza de e_i será más pequeña para aquellas observaciones que posean una mejor medición.

3.10 Consecuencias de la heteroscedasticidad

Al no cumplirse el supuesto de homoscedasticidad las distribuciones y esperanzas de los estimadores por MCO no cambian. Por lo tanto, los estimadores serán todavía insesgados, lineales y consistentes. Pero ahora se presenta un problema, y es el relacionado con la varianza de dichos estimadores.

Al suponer que la varianza es constante para los errores e_i, y como las varianzas cambian para cada una de las diferentes observaciones, entonces no se está haciendo uso de toda la información que se dispone, lo cual indica y permite deducir que los estimadores no serán eficientes.

Esto también implica que $Var(b_i)$ obtenida por el método de MCO será mayor que las varianza de b_i determinada por otros métodos de estimación que usen la información adicional sobre los cambios en la varianza del error.

Normalmente los estimadores mínimo-cuadráticos, en presencia de heteroscedasticidad, son sesgados hacia abajo, es decir, el valor estimado del error estándar del coeficiente es mayor de lo que debería de ser.

El efecto que tiene este sesgo hacia abajo sobre los estimadores mínimo-cuadráticos es que tanto el test ***F*** como el ***t*** de significancia para los coeficientes estarán subestimados, o sea, que se puede aceptar la hipótesis nula H_0 cuando en realidad se debería de rechazar.

En suma, las consecuencias de la heteroscedasticidad son dobles. Las estimaciones de los parámetros de regresión son todavía insesgados pero ineficientes y las estimaciones de las varianzas son sesgadas.

Para poder apreciar esto, considérese un modelo simple sin término constante:

$$y_i = \beta x_i + u_i \qquad Var(u_i) = \sigma_i^2 \qquad (3.10.1)$$

Donde el estimador mínimo cuadrático es $b_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$. Lo cual implica que:

$$b_1 = \frac{\Sigma x_i(\beta x_i + u_i)}{\Sigma x_i^2} = \frac{\beta \Sigma x_i^2 + \Sigma x_i u_i}{\Sigma x_i^2}$$

$$= \beta + \frac{\Sigma x_i u_i}{\Sigma x_i^2} \qquad (3.10.2)$$

Si se satisfacen los demás supuestos sobre los residuos, entonces se tiene que:

$$E(b_1) = E\left(\beta + \frac{\Sigma x_i u_i}{\Sigma x_i^2}\right) = E(\beta) + E\left(\frac{\Sigma x_i u_i}{\Sigma x_i^2}\right)$$

$$= \beta + \frac{\Sigma x_i E(u_i)}{\Sigma x_i^2} = \beta + 0 = \beta$$

De esta manera se ha verificado que el estimador mínimo cuadrático b_1 sigue siendo insesgado en presencia de heteroscedasticidad.

Similarmente se puede apreciar que:

$$V(b_1)=V\left(\frac{\Sigma x_i y_i}{\Sigma x_i^2}\right)=V\left(\beta+\frac{\Sigma x_i u_i}{\Sigma x_i^2}\right)=V\left(\frac{\Sigma x_i u_i}{\Sigma x_i^2}\right)$$

$$=V\left(\frac{x_1u_1}{\Sigma x_i^2}+\frac{x_2u_2}{\Sigma x_i^2}+.....+\frac{x_nu_n}{\Sigma x_i^2}\right)$$

$$=\frac{1}{\left(\Sigma x_i^2\right)^2}\left(x_1^2\sigma_1^2+x_2^2\sigma_2^2+.....+x_n^2\sigma_n^2\right)$$

$$=\frac{\Sigma x_i^2\sigma_i^2}{\left(\Sigma x_i^2\right)^2} \qquad (3.10.3$$

Supóngase ahora que $\sigma_i^2=\sigma^2 z_i^2$ con cada z_i conocida, entonces de la expresión 3.10.1 se tiene el siguiente modelo:

$$\frac{y_i}{z_i}=\beta\frac{x_i}{z_i}+\frac{u_i}{z_i} \qquad (3.10.4)$$

donde $V_i=\frac{u_i}{z_i}$ tiene una varianza constante e igual a σ^2.

Ahora bien, el estimador mínimo cuadrático para este nuevo modelo vendrá dado por:

$$b_1^* = \left(\frac{\sum x_i y_i}{\sum x_i^2}\right)\left(\frac{1/z_i^2}{1/z_i^2}\right) = \frac{\Sigma\left(x_i/z_i\right)\left(y_i/z_i\right)}{\Sigma\left(x_i/z_i\right)^2}$$

$$= \beta + \frac{\Sigma\left[\left(x_i/z_i\right)V_i\right]}{\Sigma\left(x_i/z_i\right)^2} \qquad (3.10.5)$$

y dado que el último término tiene esperanza igual a cero, entonces b_1^* también es insesgado.

Por otra parte, es fácil comprobar que:

$$Var\left(b_1^*\right) = \frac{\sigma^2}{\Sigma\left(x_i/z_i\right)^2} \qquad (3.10.6)$$

$$Var\left(b_1\right) = \sigma^2 \frac{\Sigma x_i^2 z_i^2}{\left(\Sigma x_i^2\right)^2} \qquad (3.10.7)$$

Finalmente, se observa que:

$$\frac{Var(b_1^*)}{Var(b_1)} = \frac{\left(\Sigma x_i^2\right)^2}{\left[\Sigma\left(x_i^2/z_i^2\right)\right]\left(\Sigma x_i^2 z_i^2\right)} \qquad (3.10.8)$$

Sea $a_i = x_i z_i$ y $b_i = \frac{x_i}{z_i}$ entonces la expresión 3.10.8 tiene la forma $\frac{\left(\Sigma a_i b_i\right)^2}{\Sigma a_i^2 \Sigma b_i^2}$, la cual es menor que 1 si a_i y b_i son proporcionales o bien que z_i^2 sea constante, el cual es el caso si los residuos son homoscedásticos.

Esto significa que el estimador de MCO es insesgado, pero menos eficiente, es decir, tiene una varianza más grande.

Si se ignora la heteroscedasticidad, entonces se estaría estimando la varianza del estimador MCO b_1 por la siguiente expresión:

$$Var(b_1) = \frac{\Sigma\left(y_i - b_1 x_i\right)^2}{n-1} \frac{1}{\Sigma x_i^2} \qquad (3.10.9)$$

donde el valor esperado de esta varianza vendrá dado por:

$$E\left[\Sigma\left(y_i - b_1 x_i\right)^2\right] = E\left[\Sigma\left(\{\beta x_i + u_i\} - b_1 x_i\right)^2\right]$$

$$= E\left[\Sigma\left(\{\beta - b_1\} x_i + u_i\right)^2\right]$$

$$= E\left[\Sigma x_i^2(\beta - b_1)^2 + 2(\beta - b_1)\Sigma x_i u_i + \Sigma u_i^2\right]$$

$$= \Sigma x_i^2 E\left[(\beta - b_1)^2\right] + 2(\beta - b_1)\Sigma x_i E(u_i) + \Sigma u_i^2$$

$$= \Sigma \sigma_i^2 - \Sigma x_i^2 E\left[(b_1 - \beta)^2\right]$$

$$= \Sigma \sigma_i^2 - \Sigma x_i^2 E\left[(b_1 - E(b_1))^2\right]$$

$$= \Sigma \sigma_i^2 - \Sigma x_i^2 Var(b_1)$$

$$= \Sigma \sigma_i^2 - \Sigma x_i^2 \left(\frac{\Sigma x_i^2 \sigma_i^2}{(\Sigma x_i^2)^2}\right)$$

$$= \Sigma \sigma_i^2 - \frac{\Sigma x_i^2 \sigma_i^2}{\Sigma x_i^2} \qquad (3.10.10)$$

Entonces si $\sigma_i^2 = \sigma^2$ para todo *i*, la expresión anterior se reduce a $(n-1)\sigma^2$, es decir, se estaría estimando la varianza para b_1 mediante una expresión cuyo valor esperado es:

$$\frac{\Sigma x_i^2 \Sigma \sigma_i^2 - \Sigma x_i^2 \sigma_i^2}{(n-1)(\Sigma x_i^2)^2} \qquad (3.10.11)$$

mientras que la verdadera varianza viene dada por

$$\frac{\Sigma x_i^2 \sigma_i^2}{(\Sigma x_i^2)^2} \qquad (3.10.12)$$

Esto significa que las varianzas estimadas también son sesgadas. Si σ_i^2 y x_i^2 están correlacionadas positivamente, de tal forma que $\Sigma x_i^2 \sigma_i^2 > \frac{\Sigma \sigma_i^2 \Sigma x_i^2}{n}$ entonces el valor esperado de la varianza estimada es más pequeño que la varianza verdadera. Es decir, como ya se había mencionado anteriormente, se estaría subestimado la verdadera varianza del estimador MCO y obteniendo intervalos de confianza más pequeños que los verdaderos. Esto también afecta a los tests de hipótesis sobre el parámetro de la regresión β, o sea, el error tipo I será más alto que el valor esperado.

3.11 Cómo detectar el problema de la no homescedasticidad

Antes que nada, es claro que la heteroscedasticidad es un problema potencialmente serio que se necesita conocer si está presente en una situación determinada. Si se puede detectar su presencia, entonces se pueden tomar acciones correctivas. Sin embargo, antes de examinar los diferentes procedimientos correctivos, primero se tiene que averiguar si está presente o si existe alguna posibilidad de que esté presente la heteroscedasticidad en un caso dado.

Como suele suceder, no existen reglas fijas y seguras para su detección, sino solamente unas cuantas normas muy generales. Esto no puede evitarse, ya que sólo se conoce σ_i^2 cuando conocemos la totalidad de la población Y correspondiente a las X's escogidas.

Por otra parte, debido a que en la mayoría de los fenómenos en estudio no se tiene suficiente control sobre el fenómeno mismo, lo más común es que en dichos estudios sólo se cuente con un valor muestral de Y para cada valor particular de las X's y por eso no existe manera de conocer σ_i^2 a partir de una sola observación de Y. Es así como en la mayoría de los estudios econométricos, la heteroscedasticidad puede ser el resultado de la intuición del investigador de un trabajo educado de prestidigitación, de experiencia empírica previa o de pura especulación.

Ahora bien, cuando no se tiene o no existe información *a priori* o empírica acerca de la heteroscedasticidad, se puede llevar a cabo el análisis de regresión sobre el supuesto de que no existe heteroscedasticidad y luego realizar un examen posterior de los residuos estimados e_i al cuadrado, puesto que éstos son los que se pueden observar y no las perturbaciones u_i, para ver si presentan algún patrón sistemático.

Al examinar los e_i^2 se pueden encontrar diversos patrones de comportamiento a través del tiempo, para tal propósito considérese el siguiente modelo:

$$Y_i = b_0 + b_1 X_i + e_i$$

además supóngase que existen tres grupos de observaciones para los cuales la varianza de e_i difiere.

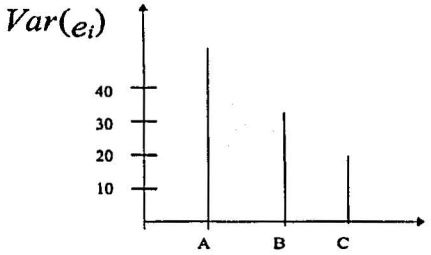


Figura 3.1 Tres grupos de observaciones con distinta varianza

Como u_i es desconocido, la detección del problema se hace con diagramas de dispersión de los residuos e_i.

Si se colocan las observaciones de los grupos A, B y C de manera consecutiva en el tiempo aparecerá un diagrama como el siguiente:

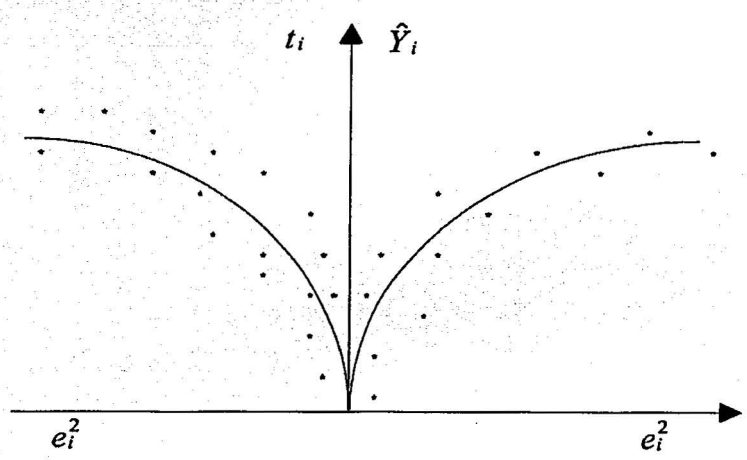


Figura 3.2 Distribución de los e_i^2

La distribución del error de este diagrama conlleva a decir que la varianza presenta un comportamiento creciente a través del tiempo, por lo que cualquier patrón diferente a dos líneas paralelas con respecto al eje de las ordenadas, es suficiente para captar la existencia de la heteroscedasticidad en el modelo.

En la práctica, la técnica de graficar los e_i^2 con respecto al tiempo es limitada, por lo que se han creado métodos informales y de aproximación para detectar la presencia de heteroscedasticidad.

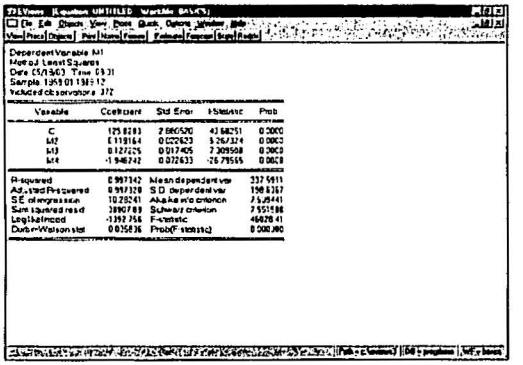
Estos métodos generalmente examinan los residuos obtenidos del procedimiento de mínimos cuadrados ordinarios para buscar en ellos patrones sistemáticos. Si ellos presentan dichos patrones se pueden sugerir maneras de transformar el modelo original bajo consideración, de tal manera que en la ecuación transformada las perturbaciones tengan una varianza constante.

Entre los métodos más usuales para la detección de heteroscedasticidad se encuentran: la prueba de Park, la prueba de Glejser, la prueba de correlación de rango de Spearman y la prueba de Goldfeld-Quandt. Y entre las menos usuales, cada una con diferentes supuestos, se encuentran: la prueba de Bartlett de homogeneidad de la varianza, la prueba de Breush-Pagan, la prueba de pico o prueba máxima, la prueba de heteroscedasticidad general de White, la prueba LM (Multiplicadores de Lagrange) y las pruebas de CUSUM y CUSUM al cuadrado.

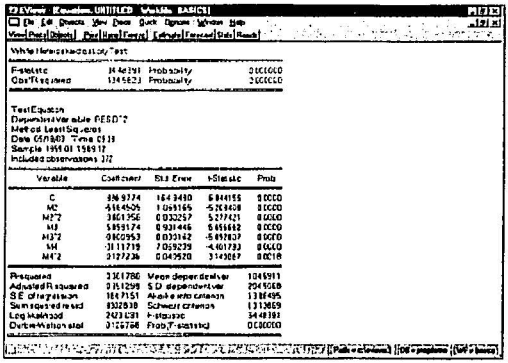
El paquete estadístico *Econometric Views* ofrece las siguientes herramientas para probar la existencia de heteroscedasticidad:

Prueba White para términos independientes

Supóngase que para un periodo determinado se lleva a cabo la regresión M1 = C + M2 + M3 + M4, por lo que el resultado en pantalla será como el siguiente:

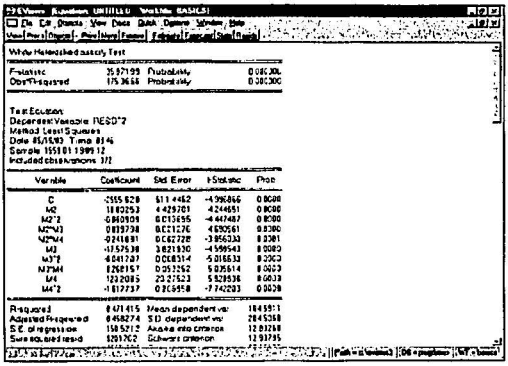


Si en la ventana de la estimación se le da clic en VIEW \ RESIDUAL TESTS \ WHITE HETEROSKEDASTICITY (no cross terms) entonces se obtendrá un resultado como el que a continuación se muestra:



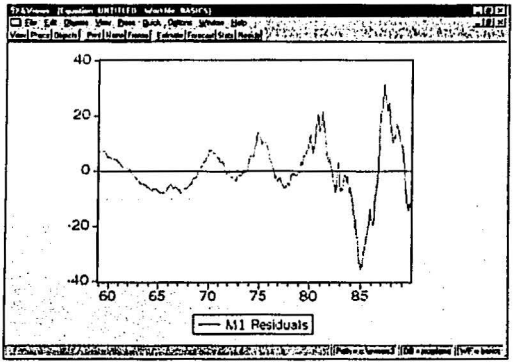
Prueba White con términos cruzados

Supóngase que se lleva a cabo la misma regresión anterior, es decir, suponiendo M1 = C + M2 + M3 + M4, entonces si en la ventana de la estimación se le da clic en VIEW \ RESIDUAL TESTS \ WHITE HETEROSKEDASTICITY (cross terms) se obtendrá un resultado como el que a continuación se presenta:

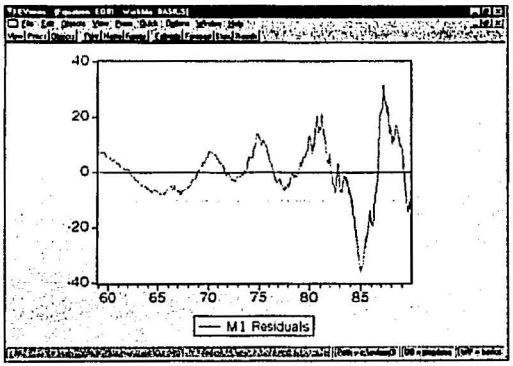


También se pueden visualizar los *residuales* de la estimación gráficamente, esto con la finalidad de ver si existe un patrón sistemático en el comportamiento de la distribución de dichos errores. Para ello solo es

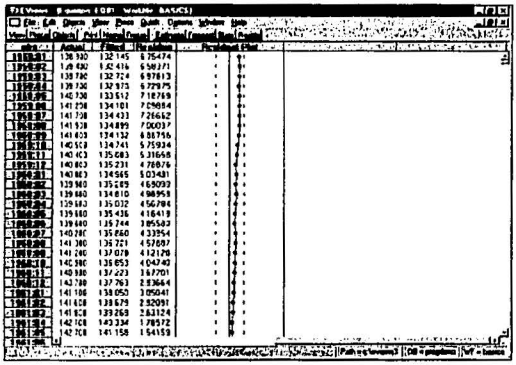
necesario darle un clic en la el botón Resids de la ventana de la estimación, o bien darle clic en la opción VIEW \ ACTUAL, FITTED, RESIDUAL \ RESIDUAL GRAPH, obteniéndose un resultado como el siguiente:



Si lo que se desea visualizar son los *errores estandarizados* entonces en la ventana de la estimación solamente debe darse clic en la opción VIEW \ ACTUAL, FITTED, RESIDUAL \ STANDARDIZED RESIDUAL GRAPH y entonces se mostrará una ventana como la siguiente:

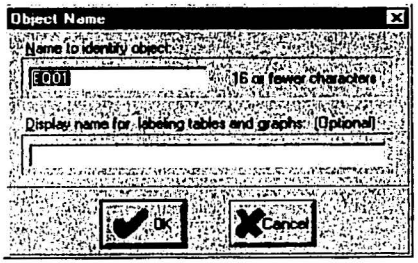


El paquete estadístico *Econometric Views* también ofrece la opción de visualizar los valores reales, los valores estimados y los residuales en un PLOT. Para ello, en la ventana de la estimación, se le da un clic en la opción VIEW \ ACTUAL, FITTED, RESIDUAL \ ACTUAL, FITTED, RESIDUAL TABLE, obteniéndose un resultado como el siguiente:

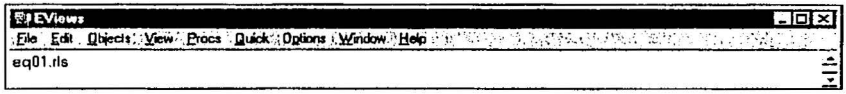


Para mostrar el resultado de la prueba CUSUM es necesario seguir los siguientes pasos:

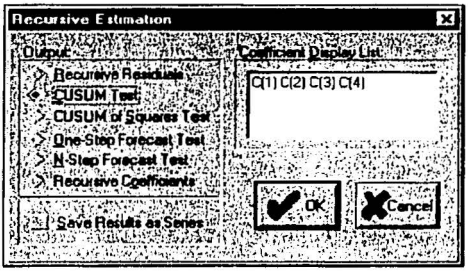
1. Llevar a cabo la regresión deseada.
2. Guardar el resultado de salida, en la ventana de la estimación se le da clic en el botón NAME y entonces se desplegará una ventana como la siguiente para poder darle nombre a la ecuación estimada:



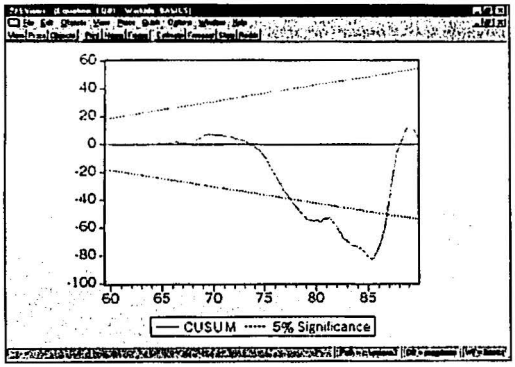
3. En la línea de comando del paquete se teclea



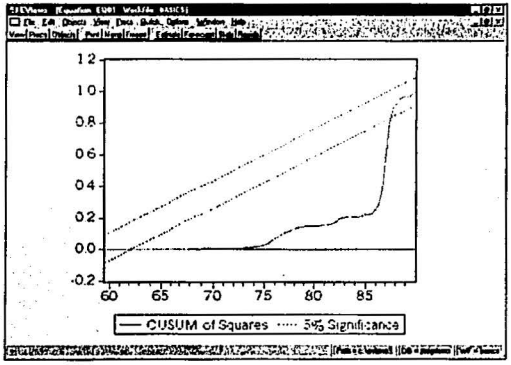
Después de oprimir la tecla ↵ , se desplegará una ventana como la siguiente:



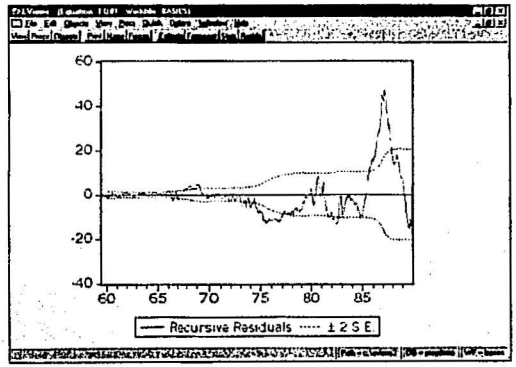
Si se selecciona la opción CUSUM Test entonces se desplegará un resultado como el siguiente:



Si lo que se desea es ver el resultado de la prueba CUSUM al cuadrado entonces solamente se le da clic en la opción CUSUM of Squares Test y el resultado que se desplegará en pantalla será como el siguiente:



Finalmente si lo que se desea visualizar son los Residuales Recursivos, entonces se marca la opción Recursive Residuals y el resultado en pantalla será como el que a continuación se muestra:



El paquete estadístico no se limita únicamente a estas pruebas, también pueden llevarse a cabo varias regresiones muy específicas para probar

otro tipo de comportamiento de la heteroscedasticidad, como por ejemplo la prueba POISSON, en esta prueba se desea visualizar un comportamiento sistemático para los residuales de la siguiente manera:

$$e_i^2 = b_0 + b_1 YEST_i \qquad \text{para } i \in [1,\ldots,n]$$

donde

e_i^2 son los residuales de la observación i al cuadrado

$YEST_i$ es el valor estimado de la variable endógena en el periodo i

si el coeficiente b_1 no es estadísticamente significativo entonces se deduce que no existen evidencias de que los residuales tengan dicho comportamiento sistemático conforme la muestra crece en el tiempo.

Algunas de las pruebas que pueden realizarse de esta manera son: la prueba de heteroscedasticidad simple, la prueba de heteroscedasticidad multiplicativa, la prueba de heteroscedasticidad AMEMYA, la prueba de Park, las pruebas de Glejser y la prueba de Golfield-Quandt. Todas estas pruebas suponen un comportamiento *a priori* de la varianza de los residuales conforme la muestra crece a través del tiempo, sin embargo, la mecánica para todas ellas es mucho muy similar.

3.12 Cómo corregir el problema de la heteroscedasticidad

Algunas de las alternativas para corregir el problema de la heteroscedasticidad son las siguientes:

- Separar las observaciones en subgrupos, de tal forma que éstos tengan la varianza del error e_i aproximadamente igual. Para cada uno de estos subgrupos se corren regresiones y así obtener la varianza de cada regresión, ya que es posible que los coeficientes para cada subgrupo sean diferentes.
- Introducir variables dicotómicas (*dummy*) para controlar la causa de la heteroscedasticidad, por ejemplo, las diferencias entre tecnologías, entre las regiones, etc.

- Hacer uso de los mínimos cuadrados ponderados. Este método consiste en asignarle un mayor peso a las observaciones para las cuales la varianza del error es pequeña y en menor peso a aquellas observaciones para las cuales la varianza es grande. Para lograr esto, el método de los mínimos cuadrados ponderados divide cada una de las observaciones para todas las variables endógenas y exógenas por una medida que corresponde al tamaño de la desviación estándar de la variable para esa observación.

  El esquema de ponderación más comúnmente usado es aquel en el cual se supone que la $Var(e_i)$ es proporcional al cuadrado de las observaciones de alguna de las variables exógenas, normalmente aquella variables donde se presenta el problema de la heteroscedasticidad.

- Procedimiento práctico. Cualquiera de los métodos enumerados, proporciona estimadores más eficientes que en el caso de los MCO. Ello puede lograrse mediante transformaciones en las observaciones para cada variable, para después aplicar el método de los MCO a los datos transformados. Por ejemplo, si en lugar de correr la regresión $Y_i = \beta_0 + \beta_1 X_i + u_i$ corriésemos $\ln(Y_i) = \beta_0 + \beta_1 \ln(X_i) + u_i$ se tiene que muy frecuentemente se reduce el problema de la heteroscedasticidad. Esto debido a que las transformaciones logarítmicas[11] comprimen las escalas en las que se miden las variables.

Finalmente, cabe mencionar que las transformaciones utilizadas para corregir el problema de la heteroscedasticidad son tan sólo especulaciones sobre la naturaleza de σ_i^2, y aquella que logre corregir dicho problema dependerá en gran medida de la naturaleza del mismo y de la severidad de la heteroscedasticidad.

[11] Una ventaja adicional de la transformación logarítmica es que el coeficiente de la pendiente β_1 mide la elasticidad de Y con respecto a X, es decir, el cambio porcentual en Y ante un cambio porcentual en X. Por ejemplo, si Y es el consumo y X es el ingreso entonces β_1 medirá la elasticidad ingreso, mientras que en el modelo original dicho coeficiente mide únicamente la tasa de cambio del consumo promedio por un cambio en una unidad en el ingreso.

3.13 La naturaleza de la correlación serial

Una de las hipótesis básicas del modelo lineal es de que la covarianza de los términos de perturbación es nula, es decir, que dichos términos de perturbación estocásticos son independientes entre sí. En símbolos,

$$Cov(u_i, u_j) = 0 \text{ para toda i diferente de j.}$$

Para un modelo con perturbaciones distribuidas normalmente lo anterior implica que todas estas perturbaciones son independientes dos a dos, para datos transversales esto significa la suposición de que el valor de la perturbación "extraído" para cualquier unidad no viene influido por los valores correspondientes a las demás unidades, y con datos temporales se supone la independencia serial para los términos de perturbación.

Sin embargo, hay circunstancias bajo las cuales la hipótesis de independencia serial del término de perturbación puede no resultar muy pausible. Por ejemplo, se puede hacer una especificación incorrecta en cuanto a la forma de la relación entre las variables, es decir, supóngase que se especifica una relación lineal entre Y y X cuando la verdadera relación es, por ejemplo, cuadrática.

El problema de la correlación serial es un problema frecuente, por no decir típico, cuando se emplean datos en series de tiempo ya que en este caso los términos de perturbación estocástica reflejan en parte variables no incluidas explícitamente en el modelo, mismas que pueden cambiar lentamente a través del tiempo. Así, el término de perturbación estocástica en una observación, estará relacionado con los términos de perturbación estocástica de las observaciones cercanas.

Otra causa de correlación serial es el suavizamiento y otras formas de "reconfigurar" datos, lo cual provoca que los términos de perturbación estocásticos se promedien a lo largo de varios periodos. Una causa general de la correlación serial es una mala especificación del modelo, en particular, la exclusión de variables relevantes para el modelo. En general, en la relación especificada solamente se incluyen ciertas variables importantes y el término de perturbación debe representar, entonces, la influencia de las variables omitidas.

TÍpicamente, el tipo de correlación serial más importante es la correlación serial lineal de primer orden, o sea, la relación lineal entre términos de perturbación estocástica sucesivos, Tal correlación serial de primer orden, adopta la forma de un proceso *Markov* o esquema autorregresivo de primer orden.

$$u_i = \rho u_{i-1} + v_i \text{ para toda } i, |\rho| < 1 \qquad (3.13.1)$$

Aquí ρ es un parámetro desconocido y que comúnmente se le denomina como el *coeficiente de autocorrelación*, también puede interpretarse como el *coeficiente de autocorrelación de primer orden* o, en forma más exacta, el *coeficiente de autocorrelación de 1 rezago.* Este nombre puede justificarse de la siguiente manera; por definición, el coeficiente poblacional de correlación entre u_i y u_{i-1} está dado por:

$$\rho = \frac{E\{[u_i - E(u_i)][u_{i-1} - E(u_{i-1})]\}}{\sqrt{Var(u_i)}\sqrt{Var(u_{i-1})}}$$

y como $E(u_i) = 0$ para cada i y $Var(u_i) = Var(u_{i-1})$ debido a se está reteniendo el supuesto de homoscedasticidad, entonces se tiene que:

$$\rho = \frac{E(u_i u_{i-1})}{Var(u_{i-1})} \qquad (3.13.2)$$

O lo que es lo mismo,

$$\rho = \frac{E(u_i u_{i-1})}{\sigma_u^2} \qquad (3.13.3)$$

En la expresión 3.13.1 v_i es un término de perturbación estocástico residual, que se supone satisface los supuestos del modelo de regresión lineal básico, incluyendo ausencia de correlación serial.
Es decir:

$$E(v_i) = 0 \qquad \text{para toda } i$$

$$E(v_i, v_j) = \begin{cases} \sigma_v^2 \text{ si } i = j \\ \\ 0 \quad \text{si } i \neq j \end{cases} \qquad \forall\ i,\ j$$

Si ahora se considera el modelo:

$$y_i = x_i\beta + u_i \tag{3.13.4}$$

de modo que, desfasando un periodo y multiplicando por ρ,

$$\rho y_{i-1} = \rho x_{i-1}\beta + \rho u_{i-1} \tag{3.13.5}$$

Restando la expresión 3.15.3 a la 3.15.2 entonces se tiene que:

$$y_i - \rho y_{i-1} = (x_i - \rho x_{i-1})\beta + (u_i - \rho u_{i-1}) \tag{3.13.6}$$

O lo que es lo mismo:

$$y_i - \rho y_{i-1} = (x_i - \rho x_{i-1})\beta + v_i \tag{3.13.7}$$

Esta última expresión contempla la variable $y_i - y_{i-1}$ como funciones lineales de las variables en $x_i - \rho x_{i-1}$. Además, presenta la misma forma que el modelo de la expresión 3.15.2, a excepción del remplazo de las variables por sus respectivas diferencias, es decir, $y_i - y_{i-1}$ y $x_i - \rho x_{i-1}$, y el remplazo de u_i por v_i.

Ahora bien, si $\rho = 1$ entonces el modelo transformado de la expresión 3.15.5 implica que:

$$y_i - y_{i-1} = (x_i - x_{i-1})\beta + v_i \qquad (3.13.8)$$

siendo un modelo en que todas las variables se reemplazan por sus primeras diferencias, o sea:

$$\Delta y_i = (\Delta x_i)\beta + v_i \qquad \text{donde } \Delta z_i = z_i - z_{i-1} \qquad (3.13.9)$$

En el otro extremo, si $\rho = -1$ entonces la expresión 3.15.5 toma la siguiente forma:

$$y_i + y_{i-1} = (x_i + x_{i-1})\beta + v_i \qquad (3.13.10)$$

siendo un modelo en el cual todas las variables se reemplazan por promedios móviles de dos periodos:

$$\mathrm{A}y_i = (\mathrm{A}x_i)\beta + v_i \qquad \text{donde } \mathrm{A}z_i = \frac{z_i + z_{i-1}}{2} \qquad (3.13.11)$$

De esta manera, la ecuación 3.13.1 se conoce como ecuación de autorregresión, debido a que el modelo usual de regresión es u_i regresada sobre u_{i-1}, y se denomina como autorregresivo de primer orden debido a que u_i se regresa sobre su pasado con un solo rezago. Si apareciesen dos rezagos, entonces se denominaría autorregresión de segundo orden. Si apareciesen tres rezagos se denominaría autorregresión de tercer orden, etc.

Si los errores u_i satisfacen la expresión 3.13.1 entonces se dice que u_i es AR(1) (es decir, autorregresivo de primer orden) y si los errores u_i satisfacen la siguiente expresión:

$$u_i = \rho_1 u_{i-1} + \rho_2 u_{i-2} + v_i$$

entonces se dice que u_i es AR(2), etcétera.

Ahora considérese nuevamente la expresión 3.13.1, es decir:

$$u_i = \rho u_{i-1} + v_i \text{ para toda } i, |\rho| < 1$$

donde es posible apreciar que:

$$\begin{aligned} u_i &= \rho u_{i-1} + v_i \\ u_{i-1} &= \rho u_{i-2} + v_{i-1} \\ u_{i-2} &= \rho u_{i-3} + v_{i-2} \\ &\vdots \\ u_{i-s} &= \rho u_{i-(s+1)} + v_{i-s} \\ &\vdots \end{aligned}$$

Lo cual implica que:

$$\begin{aligned} u_i &= \rho(\rho u_{i-2} + v_{i-1}) + v_i \\ &= \rho(\rho(\rho u_{i-3} + v_{i-2}) + v_{i-1}) + v_i \\ &= \ldots\ldots \\ &= v_i + \rho v_{i-1} + \rho^2 v_{i-2} + \ldots\ldots \end{aligned}$$

O lo que es lo mismo:

$$u_i = \sum_{s=0}^{\infty} \rho^s v_{i-s}$$

Por consiguiente $E(u_i) = o$ ya que $E(v_i) = 0$ para toda *i*.

Además

$$E(u_i^2) = E(v_i^2) + \rho^2 E(v_{i-1}^2) + \rho^4 E(v_{i-2}^2) + \ldots\ldots$$

y como las v_i son serialmente independientes entonces:

$$E(u_i^2) = (1 + \rho^2 + \rho^4 + \ldots\ldots)\sigma_v^2$$

Por lo que

$$\sigma_u^2 = \frac{\sigma_v^2}{1-\rho^2} \quad \text{para toda } i \qquad (3.13.12)$$

Donde se obtiene que:

$$E(u_i u_{i-1}) = E\left[\begin{array}{l}(v_i + \rho v_{i-1} + \rho^2 v_{i-2} + \ldots\ldots) \\ \times (v_{i-1} + \rho v_{i-2} + \rho^2 v_{i-3} + \ldots\ldots)\end{array}\right]$$

$$= E\left\{\left[v_i + \rho(v_{i-1} + \rho v_{i-2} + \rho^2 v_{i-3} + \ldots)\right](v_{i-1} + \rho v_{i-2} + \rho^2 v_{i-3} + \ldots)\right\}$$

$$= E\left[v_i(v_{i-1} + \rho v_{i-2} + \rho^2 v_{i-3} + \ldots)\right] +$$
$$E\left[\rho(v_{i-1} + \rho v_{i-2} + \rho^2 v_{i-3} + \ldots)^2\right]$$

$$= \rho E\left[\left(v_{i-1}+\rho v_{i-2}+\rho^2 v_{i-3}+\ldots\right)^2\right]$$

$$= \rho \sigma_u^2$$

Finalmente se deduce que:

$$E\left(u_i u_{i-2}\right) = \rho^2 \sigma_u^2 \qquad (3.13.13)$$

y, en general se obtiene,

$$E\left(u_i u_{i-s}\right) = \rho^s \sigma_u^2 \qquad (3.13.14)$$

Así se observa que la relación (3.13.4) no cumple con la hipótesis de independencia serial del término de perturbación. El esquema (3.13.1) es el tipo de esquema autorregresivo más simple posible; tipos más complicados llevarán también, desde luego, al incumplimiento de la hipótesis de independencia serial.

La expresión (3.13.14) se puede volver a escribir en la forma

$$\rho^s = \frac{E\left(u_i u_{i-s}\right)}{\sigma_u^2} \qquad (3.13.15)$$

El primer miembro de esta expresión define el coeficiente de autocorrelación de orden *s* de la serie *u*. El coeficiente de autocorrelación de orden cero para cualquier serie es simplemente la unidad, y para una *serie aleatoria* todos los coeficientes de orden superior serán nulos.

3.14 Estimación en presencia de correlación serial

¿Qué ocurre con los estimadores de mínimos cuadrados ordinarios y sus varianzas si se introduce autocorrelación en las perturbaciones suponiendo que $E(u_i u_j) \neq 0$ para toda i ≠ j, reteniendo los demás supuestos del modelo clásico?

Considérese el modelo 3.13.14 y ahora supóngase que el mecanismo que genera los u_i , debido a que $E(u_i, u_{i+s}) \neq 0$ con $s \neq 0$ es un supuesto demasiado general para que sea de uso práctico. Asúmase, como primera aproximación, que las perturbaciones se generan como en 3.13.1 es decir:

$$u_i = \rho u_{i-1} + v_i \text{ para toda } i, |\rho| < 1$$

Donde v_i es un término de perturbación estocástico residual, que se supone satisface los supuestos del modelo de regresión lineal básico, incluyendo ausencia de correlación serial.

Se debe mencionar que a priori no existe ninguna razón por la cual no se pueda adoptar un esquema AR(2) o AR(3) o cualquier otro mecanismo de generación del error de mayor orden, diferente al mecanismo AR(1) que se ha mencionado, Sin embargo, se utiliza este último no solo por su sencillez sino también porque en muchas aplicaciones ha demostrado ser de gran utilidad.

Ahora bien, el estimador de MCO para b_1 está dado por

$$b_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

pero su varianza, como se muestra en la siguiente sección, dado el esquema AR(1) es ahora:

$$Var(b_1)_{AR(1)} = \frac{\sigma_u^2}{\Sigma x_i^2} + \frac{2\sigma_u^2}{\Sigma x_i^2}\left[\rho \frac{\sum_{i=1}^{n-1} x_i x_{i+1}}{\sum_{i=1}^{n} x_i^2} + \rho^2 \frac{\sum_{i=1}^{n-2} x_i x_{i+2}}{\sum_{i=1}^{n} x_i^2} + \ldots + \rho^{n-1} \frac{x_1 x_n}{\sum_{i=1}^{n} x_i^2}\right]$$

La cual difiere con la expresión tradicional cuando no existe autocorrelación:

$$Var(b_1) = \frac{\sigma_u^2}{\Sigma x_i^2}$$

Comparando estas dos expresiones se observa que la primera es igual a la última más un término que depende de ρ y de las covarianzas entre los valores que tome X, y en general, no se puede decir si la $Var(b_1)$ es inferior o superior a la $Var(b_1)_{AR(1)}$, sin embargo, claro está que si ρ=0 entonces las dos fórmulas coincidirán como deben hacerlo.

Supóngase que se sigue utilizando el estimador de MCO b_1 y se ajusta la fórmula usual de la varianza teniendo en cuenta el esquema AR(1). ¿Cuáles son ahora las propiedades de b_1? Como en el caso de la heteroscedasticidad, los estimadores de MCO en presencia de autocorrelación continúan siendo lineales, insesgados y consistentes, pero dejan de ser eficientes, es decir, de varianza mínima.

También los intervalos de confianza que se deriven probablemente serán más amplios que aquellos basados en el procedimiento de Mínimos Cuadrados Generalizados (MCG). La implicación de este resultado para los procedimientos de prueba de hipótesis es clara: probablemente se declare un coeficiente estadísticamente no significativo (es decir, no diferente de cero), aunque de hecho (es decir, basándose en el procedimiento correcto de MCG) puede serlo.

3.15 Situaciones en las que se presenta el problema

Antes de citar las situaciones en donde se presenta el problema de la autocorrelación de los residuos, es esencial tener en claro ciertos aspectos de terminología. Aunque en la actualidad es práctica común el empleo de los términos *autocorrelación* y *correlación serial* como sinónimos, a veces se suele diferenciar una de otra. Por ejemplo, en ocasiones se define la autocorrelación como la "correlación de rezagos de una serie dada consigo misma, rezagada en un número de unidades de tiempo", mientras que reserva el término de correlación serial para la "correlación de rezagos entre dos series diferentes".

Por tanto, a la correlación entre dos series de tiempo tales como $u_1, u_2, \ldots\ldots, u_{10}$ y $u_2, u_3, \ldots\ldots, u_{11}$, en donde la primera definición corresponde a la última serie rezagada en un periodo de tiempo, mientras que a la correlación entre series de tiempo tales como $u_1, u_2, \ldots\ldots, u_{10}$ y $v_2, v_3, \ldots\ldots, v_{11}$, donde u y v son dos series de tiempo diferentes, se le denomina correlación serial. Aunque la distinción entre estos dos términos puede ser útil, en este trabajo se utilizarán como sinónimos.

Así, este problema está ligado generalmente a:

- Normalmente la autocorrelación de residuos se presenta en datos de series de tiempo, especialmente si el intervalo de tiempo es pequeño, ya que la observación del año j tiene efectos sobre las observaciones de los años siguientes, es decir j+1, j+2, etc.

- El problema de la autocorrelación también ocurre cuando la especificación del modelo excluye una variable cíclica cuando ésta tiene gran importancia como determinante de Y. En este caso el error absorbe el patrón cíclico y consecuentemente los términos de error no serán aleatorios. Ciertamente los signos de los tres primeros errores, por decirlo así, son información que permitirá predecir los errores de las próximas tres observaciones, si se ha detectado una variable cuyo ciclo es de tres años.

- La autocorrelación puede relacionarse con errores de medida y de heteroscedasticidad. Supóngase que una de las variables involucradas en el modelo se mide anualmente y que las demás vienen en datos mensuales. Si se emplea el método de la interpolación en la primera variable para obtener los datos mensuales, es muy probable que el error de medida se presente en los datos mensuales. Esto introduce autocorrelación en el término del error debido a la correlación del error de medida.

- Como los datos transversales conducen a problemas de heteroscedasticidad, también pueden conducir a que se presente el problema de *autocorrelación espacial*, es decir, correlación en el espacio, en lugar de aquella existente a través del tiempo. Sin embargo, es importante tener en cuenta que en análisis de corte

transversal el ordenamiento de los datos debe tener cierto interés lógico o económico para llegar a la conclusión de que se cuenta con la presencia de autocorrelación. Muchos de los factores excluidos pueden influir sobre el término del error. Si estos factores son iguales, entonces existe autocorrelación de residuos debido a la correlación que hay entre estos factores, pues son absorbidos por el término del error.

3.16 Consecuencias de la correlación serial

Si se aplican los mínimos cuadrados ordinarios a un modelo en el que las perturbaciones están autocorrelacionadas, resultan tres consecuencias principales.

- En primer lugar, se obtendrán estimadores insesgados de β, pero las varianzas mustrales de estas estimaciones pueden ser indebidamente grandes con relación a otra obtenidas según un método de estimación ligeramente diferente.

- En segundo lugar, si se aplican las fórmulas de los mínimos cuadrados usuales para las varianzas muestrales de los coeficientes de regresión probablemente se obtendrá una importante subestimación de las mismas. En cualquier caso éstas fórmulas ya no son válidas, ni tampoco son correctamente aplicables las formas de los tests *t* y *F* deducidos para el modelo lineal.

- En tercer lugar, se obtendrán predicciones *no eficientes*, esto es, predicciones con varianzas muestrales innecesariamente grandes.

Considérese a continuación el modelo de regresión de dos variables para explicar las ideas básicas de este análisis:

$$Y_i = b_o + b_1 X_i + u_i$$

$$u_i = \rho u_{i-1} + v_i$$

donde $|\rho|<1$ y v_i satisface los supuestos del modelo de regresión lineal básico, incluyendo ausencia de correlación serial.

Ahora bien, como el estimador b_i de MCO viene dado por:

$$b_1=\frac{\Sigma x_i y_i}{\Sigma x_i^2}$$

pero su varianza, dado el esquema AR(1) es ahora:

$$Var(b_1)_{AR(1)}=\frac{\sigma_u^2}{\Sigma x_i^2}+\frac{2\sigma_u^2}{\Sigma x_i^2}\left[\rho\frac{\sum_{i=1}^{n-1}x_i x_{i+1}}{\sum_{i=1}^{n}x_i^2}+\rho^2\frac{\sum_{i=1}^{n-2}x_i x_{i+2}}{\sum_{i=1}^{n}x_i^2}+...+\rho^{n-1}\frac{x_1 x_n}{\sum_{i=1}^{n}x_i^2}\right]$$

donde $Var(b_1)_{AR(1)}$ representa la varianza de b_1 bajo un esquema autorregresivo de primer orden. Lo cual contrasta con la fórmula tradicional cuando no existe correlación.

$$Var(b_1)=\frac{\sigma_u^2}{\Sigma x_i^2}$$

donde,

$$Var(b_1)_{AR(1)}=Var(b_1)+\frac{2\sigma_u^2}{\Sigma x_i^2}\left[\rho\frac{\sum_{i=1}^{n-1}x_i x_{i+1}}{\sum_{i=1}^{n}x_i^2}+\rho^2\frac{\sum_{i=1}^{n-2}x_i x_{i+2}}{\sum_{i=1}^{n}x_i^2}+..+\rho^{n-1}\frac{x_1 x_1}{\sum_{i=1}^{n}x}\right.$$

como es muy probable de que $\rho \neq 0$ y si las X's están correlacionadas positivamente entonces se tiene que:

$$Var(b_1)_{AR(1)} > Var(b_1)$$

lo cual implica que el estimador mínimo cuadrático b_1, a pesar de ser lineal e insesgado, no continúa teniendo varianza mínima, es decir, deja de ser eficiente.
Por otra parte, para demostrar que b_1 sigue siendo lineal e insesgado, no se requiere que se den los supuestos de correlación serial.

Partiendo del hecho de que

$$b_1 = \frac{\sum x_i y_i}{\sum x_i^2}$$

se obtiene

$$b_1 = \frac{\sum x_i Y_i}{\sum x_i^2} = \sum k_i Y_i \qquad (3.16.1)$$

donde $k_i = \frac{x_i}{(\sum x_i^2)}$

lo cual muestra que b_1 es un estimador lineal, puesto que es una función lineal de Y; en realidad es un promedio ponderado de Y_i, donde k_i representa las diferentes ponderaciones. Se puede demostrar de forma similar que b_0 también es un estimador lineal.

La demostración de que b_1 es un estimador insesgado puede verse en la sección 2.4 de este trabajo, nótese que en dicha demostración se asume que $E(u_i)=0$ y no que $E(u_i u_j)=0$, es decir, la presencia de correlación serial no juega un papel de suma importancia para esta demostración.

3.17 Cómo detectar el problema de la correlación serial

En la vida real es casi imposible prescindir de errores en las medidas de las encuestas y de los datos. El sesgo de los parámetros estimados es relativamente pequeño cuando el modelo presenta un buen ajuste. Sin embargo, hasta el momento no existe método alguno para detectar la correlación entre las variables explicativas (X_i) y el error observado (e_i) a posteriori. Por lo tanto la solución ideal es evitar cometer errores en la medición (medir las variables lo mejor posible).

Desde luego, es casi imposible omitir errores de medición, por lo que, antes de hacer cualquier cosa, es esencial detectar la presencia de correlación en una situación determinada. A continuación se presentan algunas pruebas de correlación serial que se utilizan comúnmente.

❖ Método gráfico

El supuesto de no autocorrelación en el modelo clásico está relacionado con las perturbaciones poblacionales u_i, las cuales no pueden observarse directamente. Lo que se tiene en cambio son sus valores aproximados, los residuos e_i, los cuales se pueden obtener utilizando el procedimiento normal de MCO. Aunque los e_i no sean lo mismo que los u_i, muy comúnmente un análisis visual de los e_i proporcionan alguna pista sobre la posible presencia de autocorrelación entre los u_i.En realidad, un análisis visual de e_i (o de e_i^2) puede proporcionar una información muy útil no solamente sobre la outocorrelación sino también sobre la heteroscedasticidad, grado de adecuabilidad del modelo o sesgo de especificación.

La importancia de producir y analizar gráficas (de los residuos) como procedimiento estándar de análisis estadístico es fundamental. Además de proporcionar ocasionalmente un resumen de fácil comprensión acerca de un problema complejo, permite adelantar un análisis simultáneo de los datos considerados en su conjunto, mientras que a la vez ilustra claramente el comportamiento de los casos individuales.

Existen diferentes formas de analizar los residuos. Se pueden simplemente graficar contra la variable tiempo, a través de una gráfica de secuencia de tiempo. En forma alternativa, se pueden graficar los *residuos estandarizados* contra la variable tiempo, estos residuos corresponden

sencillamente al dividir los e_i entre el error estándar de la estimación $\hat{\sigma}$ ($\hat{\sigma} = \sqrt{\hat{\sigma}^2}$). Nótese que los e_i, así como $\hat{\sigma}$ se miden en las unidades de la variable dependiente Y. Por tanto al realizar el cociente $\frac{e_i}{\hat{\sigma}}$ se obtendrán número puros, es decir, sin unidades de medición, pudiendo por consiguiente, compararse directamente con los residuos estándar de otra regresión.

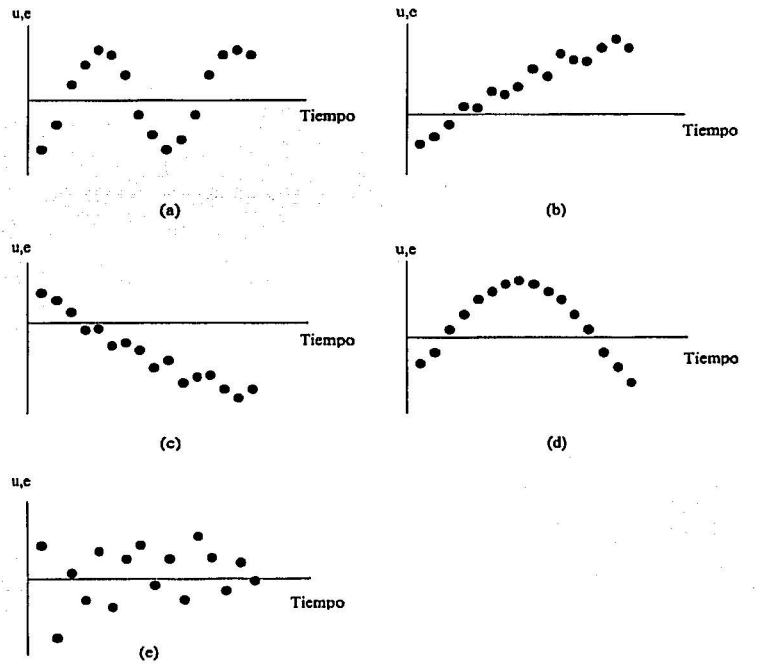


Figura 3.3 Patrones de autocorrelación

❖ La prueba de aleatoriedad o de corridas

Esta prueba, una prueba *no paramétrica*[12], a veces también conocida como prueba de *Geary*[13], se utiliza para probar hipótesis que involucran a dos grupos correlacionados, no hace ningún supuesto a cerca de la forma de la distribución de las diferencias ni pide que todas las observaciones se tomen de la misma población, El único requisito es que el investigador haya logrado en cada pareja un nivel de medición ordinal.

Se supone que la muestra se compone de *n* parejas relacionadas (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n). A cada pareja se le asocia un signo (+) si el valor de X_i sobrepasa al correspondiente Y_i, y un signo (-) si sucede lo contrario y no se asocia ningún signo (de hecho se elimina la pareja) cuando $X_i = Y_i$.

Los supuestos básicos de la prueba son los siguientes:

1. Las variables vibariadas (X_i, Y_i) $i \in [1,2,.....n]$ son mutuamente independientes.
2. La escala es ordinal en cada pareja.
3. Las parejas son internamente consistentes, en el sentido de que si P(+)>P(-) para una pareja, entonces P(+)>P(-) para todas las parejas. La misma afirmación sería válida en los casos P(+)<P(-) y P(+)=P(-).

Se define una *corrida* (llamada también *racha*) como una secuencia ininterrumpida de un símbolo o atributo, en este caso el signo + ó -. Adicionalmente se define la *longitud de la corrida* como el número de elementos en la misma.

Al examinar como se comportan las corridas en una secuencia de observaciones estrictamente aleatoria, se puede llevar a cabo una prueba de aleatoriedad de las corridas. Si existen demasiadas corridas, esto significaría que el signo de los e_i cambia frecuentemente, indicando así una correlación serial negativa. En forma similar, si hay muy pocas corridas esto puede sugerir una correlación positiva.

[12] En las pruebas no paramétricas no se plantean supuestos sobre la distribución con base en la cual se tomaron las observaciones.

[13] Véase R. C. Geary "Relative Efficiency of Count of Sign Changes for Assessing Residual Autoregression in Least Squares Regression", Biometrika vol. 57, pp. 123-127, 1970.

Si se definen

N = número total de observaciones ($N = N_1 + N_2$)
N_1 = número de símbolos + (es decir, residuos positivos)
N_2 = número de símbolos – (es decir, residuos negativos)
n = número de corridas

Entonces, bajo la hipótesis nula de que los eventos sucesivos (en este caso, los residuos) son independientes y suponiendo que $N_1 > 10$ y $N_2 > 10$, el número de corridas tiene una distribución (asintóticamente) normal con:

$$E(n) = \frac{2N_1N_2}{N_1 + N_2} + 1 \tag{3.17.1}$$

$$\sigma_n^2 = \frac{2N_1N_2(2N_1N_2 - N_1 - N_2)}{(N_1+N_2)^2(N_1 + N_2 - 1)} \tag{3.17.2}$$

Para que la hipótesis de aleatoriedad se pueda mantener, se debe esperar que n, el número de corridas en un caso determinado, se encuentre entre el intervalo $E(n) \pm 1.96\sigma_n$, con un 95% de confianza. Por lo que la regla de decisión será la siguiente:

Acéptese la hipótesis de aleatoriedad con un nivel de confianza del 95% si $[E(n) - 1396\sigma_n \leq n \leq E(n) + 1396\sigma_n]$ rechácese la hipótesis nula si el valor estimado de n se encuentra fuera de éstos límites.

❖ La prueba *d* de Durbin-Watson

El modelo más sencillo y de uso más común es aquel en el que los errores u_t y u_{t-1} tienen una correlación igual a ρ. Para este modelo, es posible pensar en probar hipótesis en torno a ρ, con base en una estimación, por ejemplo $\hat{\rho}_1$, la correlación entre residuos de los mínimos cuadrados e_t y e_{t-1}. Para probar la correlación serial de primer orden, considérese la hipótesis nula

$$H_0: \rho = 0 \tag{3.17.3}$$

mediante la cual el modelo con correlación serial de primer orden (3.13.17) se reduce al modelo básico. La prueba *Durbin-Watson* es la prueba de esta hipótesis, más comúnmente conocida como el estadístico ***d*** de *Durbin-Watson*, el cual se define de la siguiente manera:

$$d = \frac{\sum_{i=2}^{i=n}(e_i - e_{i-1})^2}{\sum_{i=1}^{i=n} e_i^2} \qquad (3.17.4)$$

Obsérvese que en el numerador del estadístico ***d*** el número de observaciones es n-1, debido a que se pierde una de ellas al tomar las diferencias consecutivas.

Una gran ventaja del estadístico ***d*** consiste en que se basa en los residuos estimados que se calculan automáticamente en el análisis de regresión. En virtud de esta ventaja, en la actualidad es frecuente incluir en los informes de análisis de regresión el estadístico ***d*** de *Durbin-Watson*, junto con el r^2, el r^2 ajustado, los valores *t*, entre otros. Aunque si bien es cierto que su uso es rutinario, conviene mencionar los supuestos en los cuales se basa:

1. El modelo de regresión incluye el término de intersección. Si este término no está presente, como en el caso de la regresión a través del origen, es esencial volver a correr la regresión incluyendo el término de intersección.
2. Las variables explicativas, son no estocásticas o fijas para muestreos repetidos.
3. Las perturbaciones u_i se generan a través de un esquema autorregresivo de primer orden: $u_i = \rho u_{i-1} + v_i$ para toda $i, |\rho| < 1$.
4. El modelo de regresión no incluye el valor o los valores rezagados de la variable dependiente como una de las variables explicativas.
5. No deben faltar observaciones en los datos, es decir, el estadístico ***d*** no permite la ausencia de observaciones.

Es difícil derivar la distribución probabilística o muestral exacta del estadístico ***d***, debido a que depende de manera complicada de los valores de las X's que están presentes en una muestra dada. Esto es comprensible, puesto que ***d*** se calcula con base en e_i, los cuales dependen de los valores de X dados. Por tanto, a diferencia de las pruebas

t, *F* o χ^2 no existe un único valor crítico que pueda llevar al rechazo o a la aceptación de la hipótesis nula de que no hay correlación serial de primer orden en las perturbaciones u_t.

Con base en la expresión (3.17.4) a continuación se demostrará que los límites de ***d*** están entre 0 y 4.

$$d = \frac{\sum_{t=2}^{t=n}(e_t - e_{t-1})^2}{\sum_{t=1}^{t=n} e_t^2}$$

desarrollando el cuadrado se tiene

$$d = \frac{\sum_{t=2}^{t=n} e_t^2 - 2\sum_{t=2}^{t=n} e_t e_{t-1} + \sum_{t=2}^{t=n} e_{t-1}^2}{\sum_{t=1}^{t=n} e_t^2} \qquad (3.17.5)$$

como $\sum_{t=2}^{t=n} e_t^2$ y $\sum_{t=2}^{t=n} e_{t-1}^2$ difieren en un solo término entonces pueden considerarse aproximadamente iguales, más aún cuando el número de observaciones es muy grande. Por tanto, haciendo $\sum_{t=2}^{t=n} e_t^2 = \sum_{t=2}^{t=n} e_{t-1}^2$ entonces la expresión (3.17.5) se escribe de la siguiente forma:

$$d = \frac{\sum_{t=2}^{t=n} e_t^2 - 2\sum_{t=2}^{t=n} e_t e_{t-1} + \sum_{t=2}^{t=n} e_t^2}{\sum_{t=1}^{t=n} e_t^2} \approx \frac{2\sum_{t=2}^{t=n} e_t^2 - 2\sum_{t=2}^{t=n} e_t e_{t-1}}{\sum_{t=1}^{t=n} e_t^2} \qquad (3.17.6)$$

o lo que es lo mismo

$$d \approx 2\frac{\sum_{t=2}^{t=n} e_t^2}{\sum_{t=1}^{t=n} e_t^2} - 2\frac{\sum_{t=2}^{t=n} e_t e_{t-1}}{\sum_{t=1}^{t=n} e_t^2} \approx 2\left(1 - \frac{\sum_{t=2}^{t=n} e_t e_{t-1}}{\sum_{t=1}^{t=n} e_t^2}\right) \qquad (3.17.7)$$

como las perturbaciones se generan a través de un esquema autorregresivo de primer orden entonces se tiene que:

$$e_i = \rho e_{i-1} + \delta_i \text{ para toda } i, |\rho| < 1 \qquad (3.17.8)$$

si se toman MCO para la expresión (3.17.8) se tiene que

$$\hat{\rho} = \frac{\sum_{i=2}^{i=n} e_i e_{i-1}}{\sum_{i=1}^{i=n} e_i^2} \qquad (3.17.9)$$

por lo que la expresión (3.17.7) se transforma en

$$d \approx 2(1-\rho) \qquad (3.17.10)$$

de aquí se desprenden los siguientes resultados:

Si $\hat{\rho}$ =1 entonces $d \approx 2(1-1) = 0$ (cuanto se tiene correlación serial exacta positiva de primer orden).

Si $\hat{\rho}$ =-1 entonces $d \approx 2(1-(-1)) = 4$ (cuanto se tiene correlación serial exacta negativa de primer orden).

Si $\hat{\rho}$ =0 entonces $d \approx 2(1-0) = 2$ (cuanto no existe correlación serial de primer orden).

Por lo tanto, cuando $-1 \leq \rho \leq 1$ implica que

$$0 \leq d \leq 4 \qquad (3.17.11)$$

Por tanto, como regla general, si se encuentra que ***d*** es igual a 2 en una aplicación, se puede suponer que no existe correlación de primer orden, ya sea positiva o negativa.

La mecánica de la prueba de Durbin–Watson es la siguiente, asumiendo que se satisfacen los supuestos en los que se basa dicha prueba:

1. Correr la regresión de MCO y obtener los residuos de e_i.
2. Calcular ***d*** a partir de (3.17.4).
3. Para un tamaño de muestra dado ***n*** y un número de variables expicativas ***k*** determinado, hallar los valores críticos de d_L y d_U.
4. Seguir las reglas de decisión dadas en la Tabla 3.1.

Hipótesis nula	Decisión	Si sucede
No existe autocorrelación positiva	Rechazar	$0 < d < d_L$
No existe autocorrelación positiva	No hay decisión	$d_L \leq d \leq d_U$
No existe autocorrelación negativa	Rechazar	$4-d_L < d < 4$
No existe autocorrelación negativa	No hay decisión	$4-d_U \leq d \leq 4-d_L$
No existe autocorrelación positiva o negativa	No Rechazar	$d_U < d < 4-d_U$

Tabla 3.1 Prueba de Durbin-Watson: Reglas de decisión.

Para una mejor comprensión de la tabla 3.1 a continuación se muestra una figura que ilustra estas reglas de decisión.

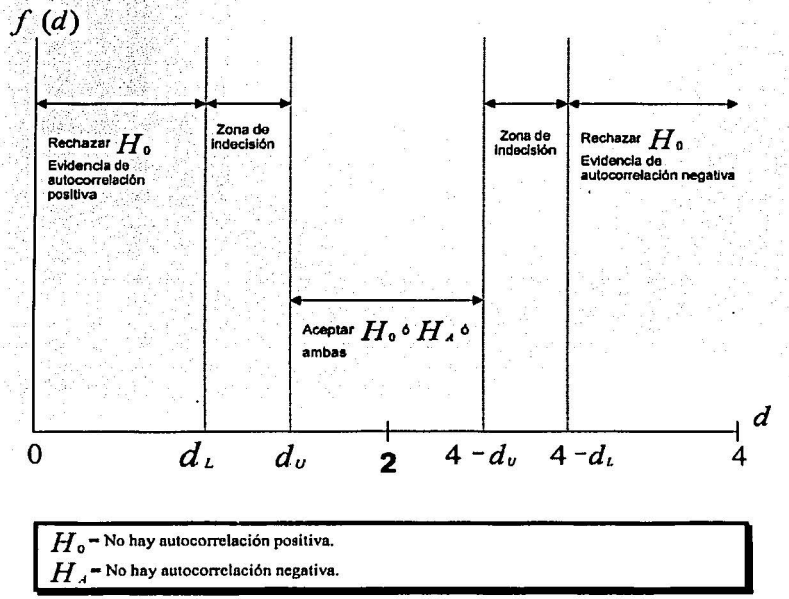
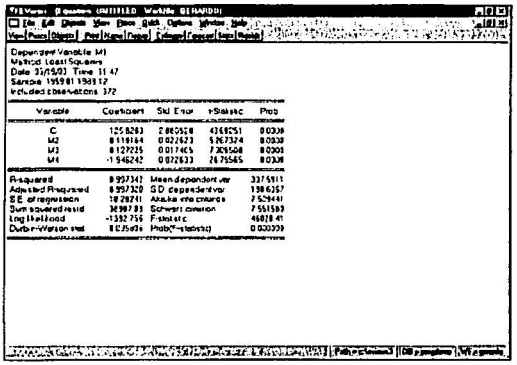


Figura 3.4 Estadístico *d* de Durbin-Watson

En las dos regiones de indecisión que quedan la prueba no sirve, pero en esos casos, se acostumbra rechazar la hipótesis de ausencia de correlación serial de primer orden.

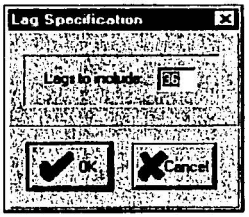
El paquete estadístico *Econometric Views* ofrece las siguientes herramientas para detectar la existencia de correlación serial:

Supóngase que para un periodo determinado se lleva a cabo la regresión M1 = C + M2 + M3 + M4, el resultado en pantalla será como el siguiente:

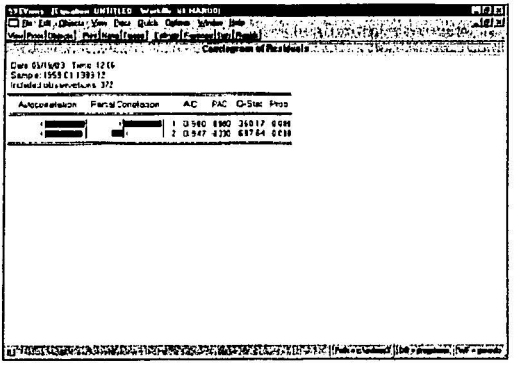


Estadístico Durbin-Watson
El cálculo de este estadístico se presenta en la pantalla de la estimación y se ubica en la segunda columna de resultados y en la fila que corresponde a Durbin-Watson stat.

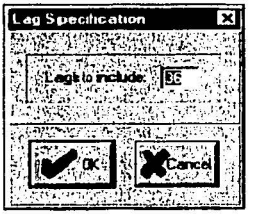
Correlograma *o* estadístico *Q*
Para ver el resultado de esta prueba y con base en la estimación anterior, solamente se le da clic en la opción VIEW \ RESIDUAL TESTS \ CORRELOGRAM- Q-STATISTICS, entonces se desplegará una pantallita como la siguiente:



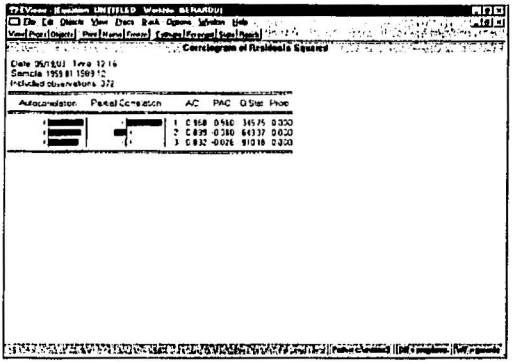
Si se desea que la prueba se haga con dos "*rezagos*" (lags) entonces se teclea el número 2 y el resultado será:



Si lo que se desea ver es el correlograma de los residuales al cuadrado entonces, en la ventana de la estimación, se le da clic en la opción VIEW \ RESIDUAL TEST \ CORRELOGRAM SQUARED RESIDUALS y nuevamente se mostrará la ventana siguiente:

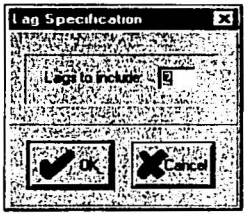


Si se desea que la prueba se haga para tres "*rezagos*" entonces se teclea el número 3 y el resultado será como el que a continuación se muestra:

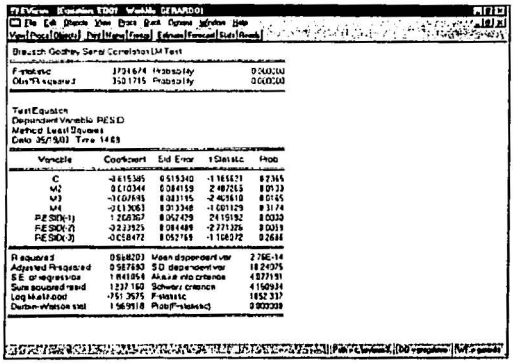


Prueba de Breusch-Godfrey para correlación serial (Multiplicadores de Lagrange)

Para llevar a cabo esta prueba solo es necesario, en la ventana de la estimación, darle clic a la opción VIEW \ RESIDUAL TESTS \ SERIAL CORRELATION LM TEST y entonces se mostrará una ventana como la siguiente:

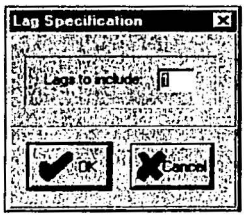


Dependiendo del orden con que se deseé llevar a cabo la prueba entonces será el número que se tenga que teclear, supóngase que se desea probar una correlación de tercer orden, lo que procede es teclear el número 3 y el resultado obtenido será como el siguiente:

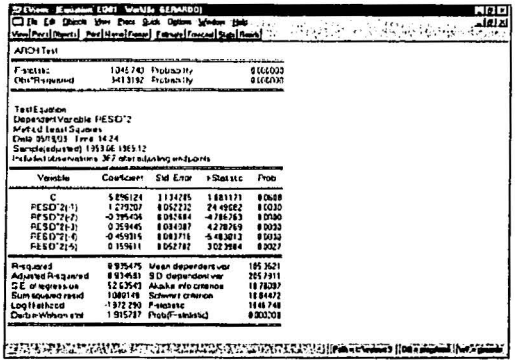


Prueba ARCH

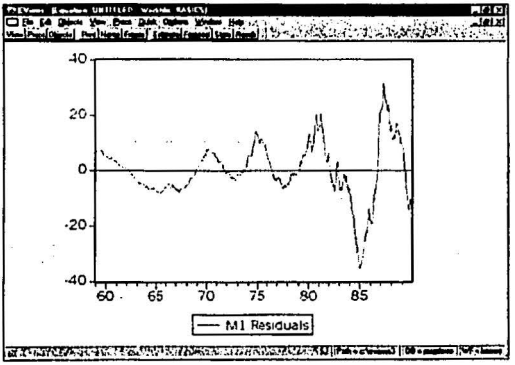
El resultado para esta prueba se obtendrá al darle clic a la opción VIEW \ RESIDUAL TEST \ ARCH LM TEST y en pantalla se observará una ventana como la siguiente:



Dependiendo del grado con que se quiera realizar la prueba entonces será el número tecleado, supóngase que se quiere ver el resultado de la prueba para un orden de 5, una vez tecleado el número 5 en la ventana la pantalla mostrará el siguiente resultado:

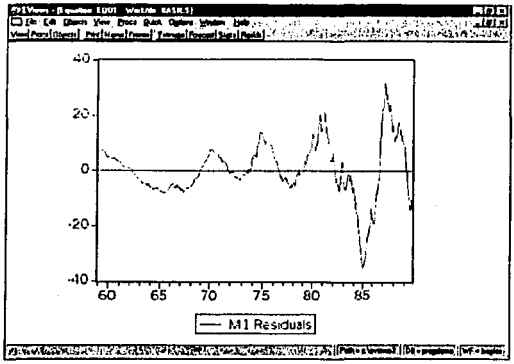


También se pueden visualizar los *residuales* de la estimación gráficamente, esto con la finalidad de ver si existe un patrón sistemático en el comportamiento de la distribución de dichos errores. Para ello solo es necesario darle un clic en la el botón Resids de la ventana de la estimación, o bien darle clic en la opción VIEW \ ACTUAL, FITTED, RESIDUAL \ RESIDUAL GRAPH, obteniéndose un resultado como el siguiente:

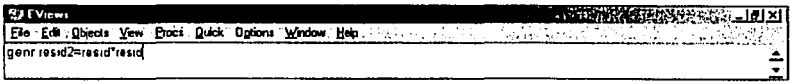


Si lo que se desea visualizar son los *errores estandarizados* entonces en la ventana de la estimación solamente debe darse clic en la opción VIEW \

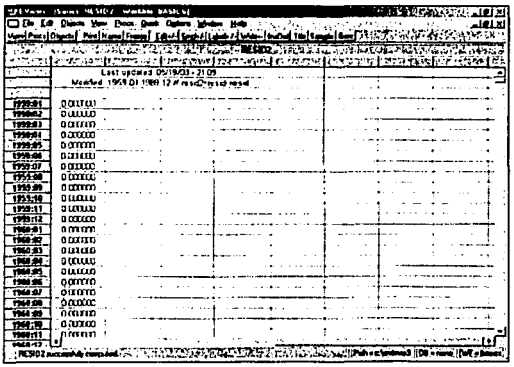
ACTUAL, FITTED, RESIDUAL \ STANDARDIZED RESIDUAL GRAPH y entonces se mostrará una ventana como la siguiente:



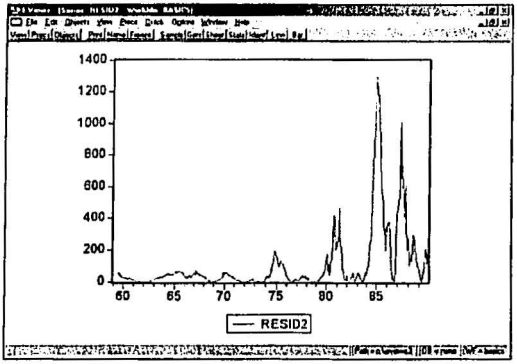
Si se desea visualizar los residuales al cuadrado entonces se debe de generar la serie primero de la siguiente manera:



Una vez generada la serie resid2 se le da doble clic y la pantalla que se mostrará será como la siguiente:



Para ver el gráfico, solamente se le da clic en VIEW \ LINE GRAPH y entonces el resultado en pantalla será como el siguiente:



3.18 Cómo corregir el problema de la correlación serial

Dado que en presencia de correlación serial los estimadores de MCO son ineficientes, es necesario buscar medidas remediales. Sin embargo, el remedio depende del conocimiento que se tenga sobre la naturaleza de la independencia existente entre las perturbaciones.

Si al utilizar la prueba *Durbin-Watson* el diagnóstico es correlación serial de primer orden, existen dos posibles tratamientos para solucionar este problema y cada uno de ellos implica un cambio en la especificación del modelo original.

- El primero consiste en incluir otras variables explicativas. Hasta cierto punto, el término de perturbación estocástica puede estar representando la acción de otras variables, mismas que si fuesen consideradas explícitamente, reducirían la correlación serial.
- El segundo tratamiento, que se utiliza cuando no es fácil disponer de información sobre tales variables adicionales, contempla la estimación no del modelo original sino del modelo transformado (3.13.6). Pero para llevar esto a cabo es necesario antes que nada conocer una estimación del parámetro de otra forma no conocido ρ.

Algunas veces se puede corregir la autocorrelación mejorando la especificación del modelo. Esto sucede cuando la causa de la correlación se debe a que el papel de una variable excluída del modelo es de gran importancia, y además tiene un patrón cíclico muy fuerte. El solo hecho de incluir esta variable, o de efectuar algunas transformaciones sobre algunas de las variables del modelo puede mejorar sustancialmente el modelo y eliminar así la autocorrelación.

Debido a que en la práctica muy pocas veces se conoce el valor de ρ un enfoque para estimarlo es *"estimar "* primero el modelo original poblacional, emplear estas estimaciones para construir las e_i y estimar ρ como la regresión lineal simple del proceso de *Markov* (3.13.1), obteniéndose:

$$\hat{\rho} = \frac{\sum_{i=2}^{i=n} e_i e_{i-1}}{\sum_{i=1}^{i=n} e_i^2}$$

Esta estimación puede ser usada para obtener información sobre el modelo transformado (3.13.1), que en ese caso puede ser estimado utilizando las técnicas comunes de regresión para producir estimaciones de los parámetros β .

Otro enfoque consiste en un procedimiento de dos pasos en el cual se estima ρ mediante la aplicación de MCO a la ecuación (3.13.1) en la siguiente forma:

$$y_i = \rho\, y_{i-1} + (x_i - \rho\, x_{i-1})\beta + v_i \qquad (3.18.1)$$

esto con el fin de obtener $\hat{\rho}$ como el coeficiente estimado de y_{i-1}. El segundo paso consiste entonces en estimar β en el modelo transformado (3.13.1) utilizando esta ρ .

Ante esto, ¿cuál enfoque utilizar? Se complica al tratar de dar una respuesta que cubra todas las contingencias; sin embargo, la preferencia relativa de los diferentes enfoques por lo general es la que tiene el orden bajo el cual han sido presentados en este trabajo. El mejor enfoque es el

que consiste en agregar variables explicativas relevantes. Si esto falla, el siguiente paso mejor es, por lo común, emplear el estimador más sencillo para ρ.

El estimador de (3.17.9) es uno de varios posibles, un estimador alternativo de ρ es el deducido por Theil y Nagar (1971):

$$\hat{\rho} = \frac{1 - \frac{d}{2} + \left(\frac{k}{n}\right)^2}{1 - \left(\frac{k}{n}\right)^2} \qquad (3.18.2)$$

donde

d es el estadístico de *Durbin- Watson.*
k es el número de parámetros que serán estimados.
n es el número de observaciones.

Obsérvese que si ***n>k*** entonces

$$\hat{\rho} \approx 1 - \frac{d}{2} \qquad (3.18.3)$$

De hecho si n es suficientemente grande, $1 - \frac{d}{2}$ es una buena aproximación al estimador en (3.17.9).

Este resultado también puede verificarse usando (3.17.10):

$$d \approx 2(1 - \hat{\rho})$$

lo cual implica que

$$\frac{d}{2} \approx 1 - \hat{\rho}$$

donde finalmente se observa

$$\rho \approx 1 - \frac{d}{2}$$

Así, una vez estimado ρ (independientemente del método) se pueden transformar los datos, y proceder con la estimación normal de MCO. Para poder apreciar esto, considérese el modelo de dos variables (no importa si el modelo tiene más de una variable explicativa porque en realidad la autocorrelación es un problema de las u_i's), también asúmase el hecho de (3.13.1), es decir:

$$u_i = \rho u_{i-1} + v_i \text{ para toda } i, |\rho| < 1$$

De esta manera se tiene lo siguiente

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad (3.18.4)$$

Si la expresión (3.18.4) se cumple para el periodo *i* entonces se cumple para *i-1*, por lo tanto:

$$Y_{i-1} = \beta_0 + \beta_1 X_{i-1} + u_{i-1} \qquad (3.18.5)$$

multiplicando (3.18.5) por ρ se obtiene

$$\rho\, Y_{i-1} = \rho\, \beta_0 + \rho\, \beta_1 X_{i-1} + \rho\, u_{i-1} \qquad (3.18.6)$$

restando (3.18.6) a (3.18.4)

$$Y_i - \rho\, Y_{i-1} = \beta_0(1-\rho) + \beta_1(X_i - \rho\, X_{i-1}) + (u_i - \rho\, u_{i-1}) \qquad (3.18.7)$$

Esta expresión se conoce como *ecuación de diferencia generalizada*, la cual involucra la regresión de Y en X, no en la forma original sino en forma de diferencias, las cuales se obtienen restando una proporción (de

hecho ρ) del valor de una variable en el previo periodo (*i-1*), del valor en el actual periodo (*i*). En este procedimiento de diferenciación se pierde una observación, puesto que la primera de ellas no tiene observación que la anteceda. Para evitar la pérdida de una observación, la primera observación de Y y X se transforman de la siguiente manera[14]:

$$Y_1=\sqrt{1-\rho^2} \text{ y } X_1=\sqrt{1-\rho^2}$$

Ahora bien, sean

$$Y_i^*=Y_i-\rho\ Y_{i-1}$$
$$\beta_0^*=\beta_0\left(1-\rho\right)$$
$$\beta_1^* X_i^*=\beta_1\left(X_i-\rho\ X_{i-1}\right)$$

y por construcción $v_i=u_i-\rho\ u_{i-1}$ entonces (3.18.7) se transforma en:

$$Y_i^*=\beta_0^*+\beta_1^* X_i^*+v_i \qquad (3.18.8)$$

Puesto que v_i satisface todos los supuestos de MCO, entonces se pueden seguir aplicando los MCO a las variables transformadas Y_i^* y X_i^* para obtener los mejores estimadores lineales insesgados (MELI) con todas las propiedades óptimas.

Sin embargo es de suma importancia preguntarse ¿tendrán los coeficientes de regresión estimados las propiedades óptimas usuales del modelo clásico? Sin entrar en detalles técnicos se puede enunciar que: *cuando se utiliza un estimador en lugar del verdadero valor, los coeficientes de MCO estimados tienen las propiedades óptimas usuales sólo asintóticamente, es decir, para muestras grandes. De forma similar, los procedimientos de pruebas de hipótesis convencionales estrictamente hablando son válidos asintóticamente, Por lo que, en muestras pequeñas debe tenerse mucho cuidado al interpretar los resultados estimados.*

[14] Esta transformación se conoce como *"transformación de Prais-Winsten"*.

4 Presentación del paquete estadístico Econometric Views V3.1

El paquete estadístico Econometrics Views V 3.1 es una versión posterior al TSP (Time Series Processor) y que al ser trabajada en el ambiente Windows su ejecución es más amigable e interactiva con el usuario. Dicho paquete fue desarrollado para economistas y lógicamente la mayor parte de sus aplicaciones están enfocadas hacia resolver problemas esencialmente económicos.

Sin embargo, algunas de las áreas en donde Eviews puede ser usada son:

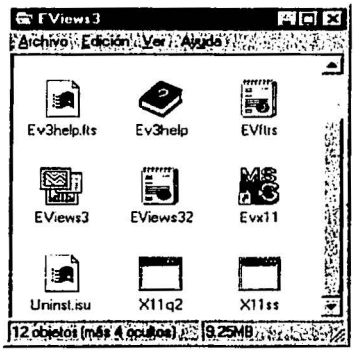
- ✓ Pronósticos de Ventas
- ✓ Análisis de costos y pronósticos
- ✓ Análisis Financiero
- ✓ Pronósticos Macroeconómicos
- ✓ Simulación
- ✓ Análisis de datos científicos y evaluación, entre otras.

Eviews es una nueva versión de un conjunto de herramientas para analizar datos de series de tiempo, originalmente desarrolladas en el software del TSP. El inmediato predecesor de Eviews fue MicroTSP, desarrollado por primera vez en 1981, y como ya se mencionó anteriormente, este paquete fue implementado para economistas y gran parte de sus aplicaciones son en economía, sin embargo esto no limita los usos de dicho paquete para otro tipo de disciplinas.

Los requerimientos de sistema son realmente mínimos, pues para poder tener acceso a este programa tan sólo se requiere disponer de lo siguiente:

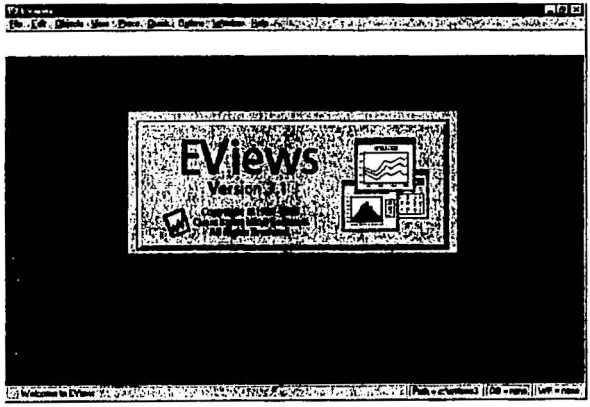
- 🗁 Una PC, 486 o Pentium con Microsoft Windows 95 o posterior.
- 🗁 Un mínimo de 4 megabytes en RAM.
- 🗁 Un monitor VGA, súper VGA o compatible.
- 🗁 Un mouse compatible con Windows
- 🗁 Un mínimo de 10 megabytes libres en disco duro.

Una vez instalado este paquete en la PC, el ícono de trabajo se presentará en una ventana como la siguiente:

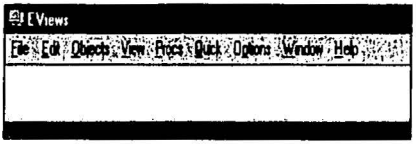


Haciendo doble clic en el ícono 

la pantalla principal que el paquete desplegará será como la siguiente:

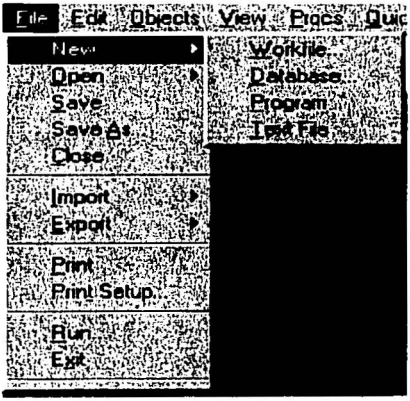


Este paquete contiene en el Menú Principal las siguientes opciones:

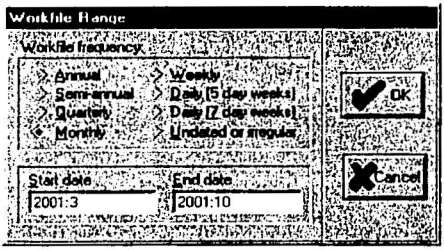


Una vez dentro del paquete, el siguiente paso es la captura de los datos (en el caso de que la base de datos se encuentre en cualquier paquete del ambiente Windows tal como Excel, Lotus for Windows, HGW, etcétera, entonces simplemente se hará un copiado y pegado en forma ordinaria). De otra manera, se recomiendan las opciones de importar y/o exportación de datos que se encuentran en la opción FILE.

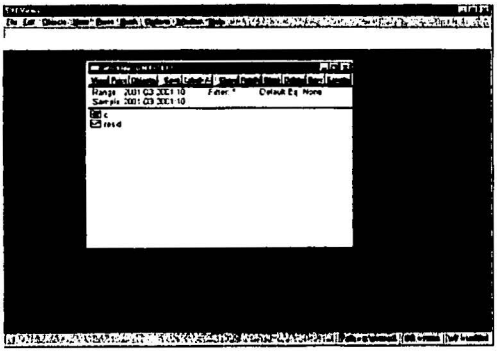
Para crear un archivo de trabajo nuevo, se tiene que hacer clic en la opción FILE/NEW de la barra del Menú Principal y la persiana mostrará las siguientes opciones:



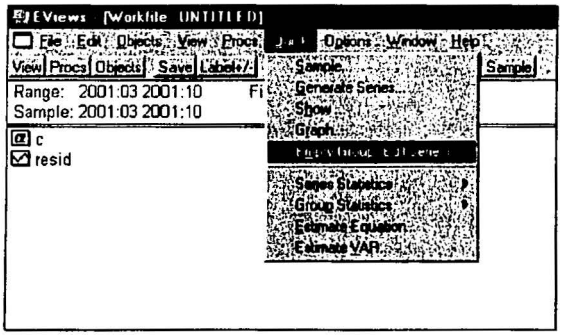
Como puede apreciarse, este submenú proporciona diversas maneras de trabajar con un archivo nuevo: Archivo de texto (Text File), en forma de programa (Program), como una base de datos (Database) y como un archivo de trabajo (Workfile). De esta manera, se selecciona la opción de WORKFILE y se le da clic, a continuación se mostrarán las diferentes maneras con que se pueden trabajar los datos, es decir, estos pueden ser anuales, trimestrales, mensuales, semanales, diarios o sin fecha; también se muestra una ventana de diálogo para poner la fecha de inicio y la fecha final de la serie de datos con que se va a trabajar. Por ejemplo, si los datos son mensuales y comienzan en marzo y terminan en octubre del 2001, entonces la ventana será de la siguiente manera:



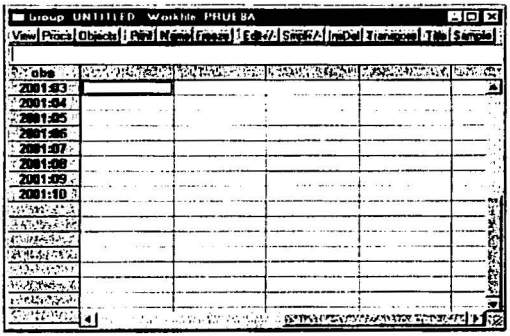
Ya especificado el archivo de trabajo entonces se da clic en la opción de OK o se le da ENTER y ahora la ventana desplegada será como la siguiente:



Ahora para editar las series con que se van a trabajar es necesario seleccionar la siguiente opción de la barra principal QUICK\EMPLY GROUP (EDIT SERIES):



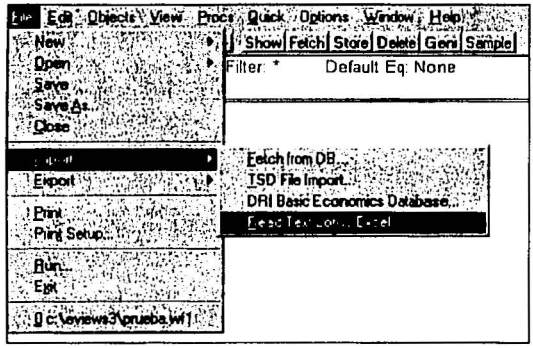
Y entonces se abrirá una ventana mostrando lo siguiente (para guardar este grupo simplemente se hace clic en el recuadro NAME y se le da un nombre, por ejemplo PRUEBA):



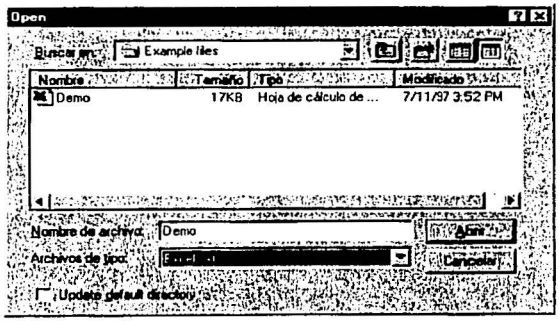
Lo siguiente sería empezar a capturar los datos (en forma de columnas) y para identificar una serie de otra, en la parte donde dice OBS haciendo clic se puede dar el nombre a cada serie de datos (excluyendo algunos nombres internos que utiliza el paquete como C, LS, por citar algunos).

Si los datos no son muchos entonces se procede a su captura, sin embargo cuando éstos son demasiados se recomienda utilizar otros mecanismos, como la importación de datos o el copiado y pegado ordinario del ambiente windows.

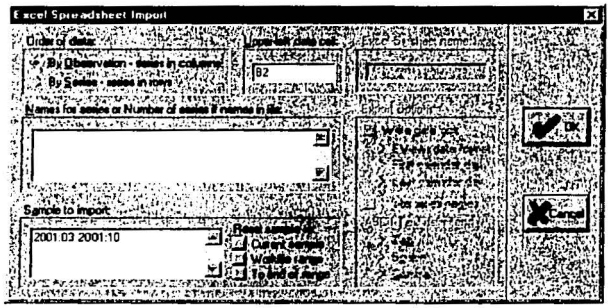
Este paquete ofrece la opción de importar series de datos que se encuentren ya sea en forma de código ASCCI, Lotus, Excel, o bien archivos generados en versiones anteriores de Eviews. Para esto es necesario hacer clic en FILE/IMPORT del menú principal y se selecciona el tipo de formato en que se encuentran los datos que se quieren importar.



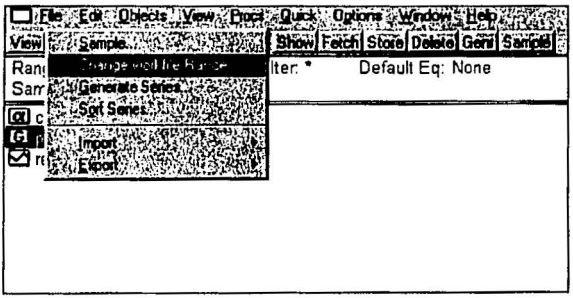
Por ejemplo, si se selecciona la opción donde se encuentra el formato de Excel, la caja de diálogo mostrará lo siguiente:



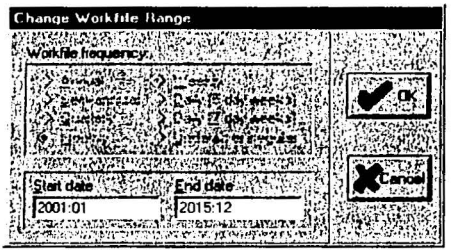
Si se da clic en OK, la caja de diálogo inmediata mostrará la forma en que se pueden extraer los datos, es decir, se pueden importar las series ya sea en forma de columnas o renglones, esto lo hace a partir de la celda izquierda superior hacia la derecha; también en esta caja de diálogo se puede asignar el nombre a las series que se importarán.



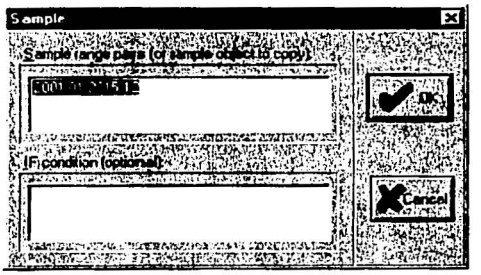
Sin embargo, como el rango de los datos a importar no coincide con el rango del archivo que se creó con anterioridad, entonces lo que se hace es expandir el rango del archivo de trabajo. Para esto, y una vez que se le ha dado clic en la opción CANCEL, se selecciona el botón PROCS del WORKFILE (no el del Menú Principal), y entonces se desplegarán las siguientes opciones:



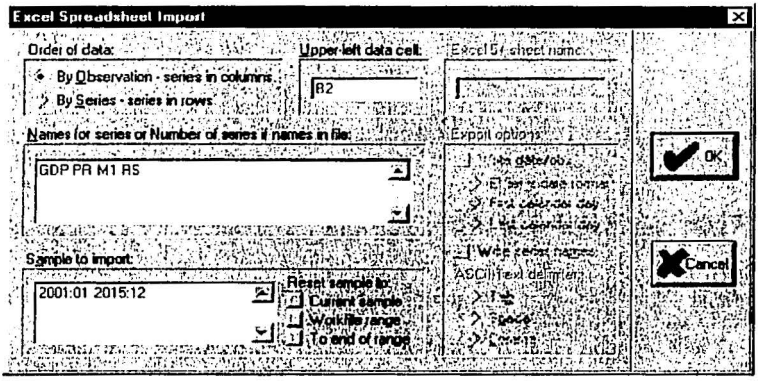
En esta persiana se selecciona la opción CHANGE WORKFILE RANGE, después, en la ventana que se muestra simplemente se le cambia el rango de entrada, supóngase que ahora los datos son de enero del 2001 a diciembre del 2015 (obsérvese que los datos continúan siendo mensuales), entonces la ventana ahora será como la siguiente:



Una vez que se le ha dado clic al botón de OK, ahora se procede a igualar el rango de trabajo, para esto se le da clic al botón de SAMPLE del WORKFILE:

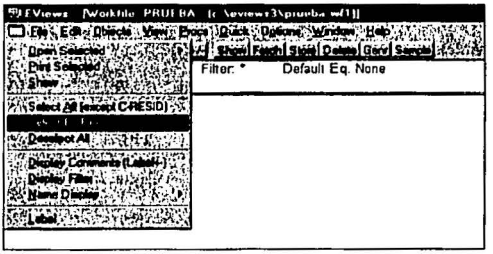


En esta ventana se teclea el rango 2001:1 (espacio) 2015:12 y después se le da clic al botón de OK. Ahora sí ya se pueden importar los datos en formato Excel. Para esto, recuérdese que la última ventana era como la siguiente:

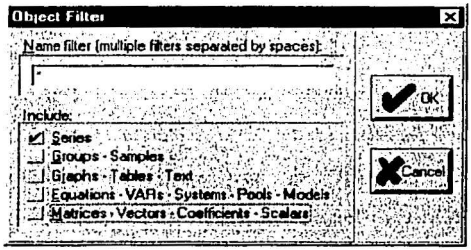


Nótese que en esta ventana aparece ya el nuevo SAMPLE y también se ha tecleado el nombre de las cuatro series que se van ha importar. Aceptadas las opciones (haciendo clic en OK) los datos serán movidos al archivo de trabajo creado con anterioridad.

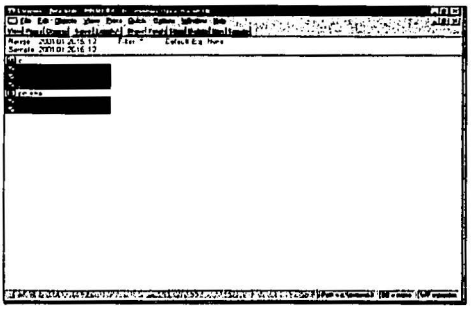
Si ahora se quieren seleccionar únicamente las series entonces se le da clic al botón VIEW del archivo de trabajo, y en la persiana que se despliegue entonces se elige la opción SELECT BY FILTER:



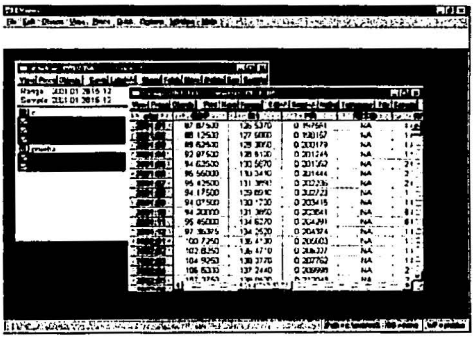
En la ventana que se muestra enseguida se selecciona únicamente la opción SERIES:



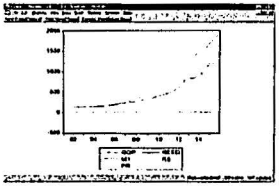
Entonces al darle un clic en OK la ventana será como la siguiente:

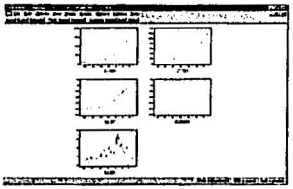


Y haciendo doble clic en la región sombreada se desplegará una ventana mostrando las series que se importaron.



En esta nueva ventana se presentan una serie de opciones para con ellas realizar y/o ejecutar operaciones con los datos seleccionados. Por ejemplo, si se desea ver un gráfico (individual o en grupo) en línea entonces haciendo clic en VIEW/GRAPH/LINE las gráficas serán como las siguientes:





También pueden verse unas estadísticas individuales de estas series (como la media, la mediana, su valor máximo y mínimo, el número de observaciones, la prueba Jarque-Bera de normalidad, el indicador de Kurtosis, entre otras), para ello se le da clic en el botón VIEW del archivo de trabajo, después se selecciona DESCRIPTIVE STATS/INDIVIDUAL SAMPLE y la ventana que se mostrará será como la siguiente:

EViews [Group UNTITLED Workfile PRUEBA]

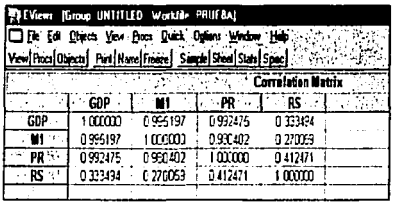
File Edit Objects View Proc Quick Options Window Help

View Proc Object Print Name Freeze Sample Sheet Stats Spec

	GDP	M1	PR	RESID	RS
Mean	632.4190	445.0064	0.514106	NA	5.412928
Median	374.3000	298.3990	0.303002	NA	5.057500
Maximum	1948.225	1219.420	1.110511	NA	15.09733
Minimum	87.87500	126.5370	0.197561	NA	0.814333
Std. Dev.	564.2441	344.8315	0.303493	NA	2.909939
Skewness	0.845800	0.997776	0.592712	NA	0.906782
Kurtosis	2.345038	2.687096	1.829239	NA	4.049883
Jarque-Bera	24.68300	30.60101	20.81933	NA	37.47907
Probability	0.000004	0.000000	0.000030	NA	0.000000
Observations	180	180	180	0	180

Para este grupo de series también se pueden calcular la matriz de correlación y la matriz de covarianza, para ello se le da clic en el botón VIEW del archivo de trabajo, después se selecciona la opción CORRELATIONS ó COVARIANCES, según sea el caso. Entonces para

cada una de éstas opciones se desplegará una ventana como las siguientes:



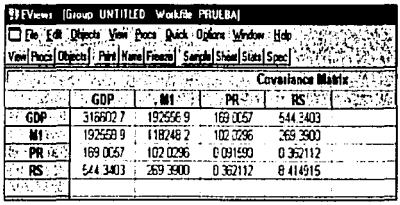
EViews [Group UNTITLED Workfile PRUEBA]

File Edit Objects View Proc Quick Options Window Help

View | Proc | Objects | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Correlation Matrix

	GDP	M1	PR	RS
GDP	1.000000	0.995197	0.992475	0.333494
M1	0.995197	1.000000	0.990402	0.270059
PR	0.992475	0.990402	1.000000	0.412471
RS	0.333494	0.270059	0.412471	1.000000



EViews [Group UNTITLED Workfile PRUEBA]

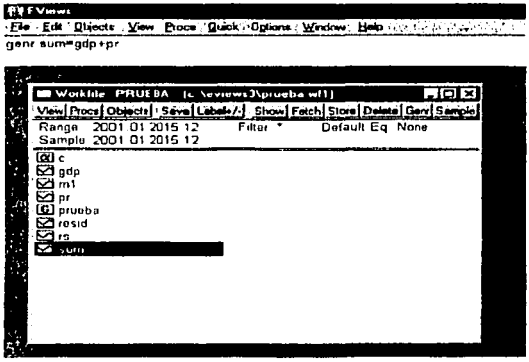
File Edit Objects View Proc Quick Options Window Help

View | Proc | Objects | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Covariance Matrix

	GDP	M1	PR	RS
GDP	3166902.7	1925556.9	169.0057	544.3403
M1	1925556.9	1182482	102.0296	269.3900
PR	169.0057	102.0296	0.091590	0.362112
RS	544.3403	269.3900	0.362112	8.414915

Cabe señalar que también puede utilizarse la línea de comando, que no es más que la línea de estado en donde se pueden ejecutar de manera directa y abreviada los múltiples comandos que nos ofrece este paquete. Por ejemplo, si se quiere generar la serie SUMA=GDP+PR entonces simplemente se teclea GENR SUMA=GDP+PR.



EViews

File Edit Objects View Proc Quick Options Window Help

genr suma=gdp+pr

Workfile PRUEBA [c:\eviews3\prueba.wf1]

View | Proc | Objects | Save | Label/-/ | Show | Fetch | Store | Delete | Genr | Sample

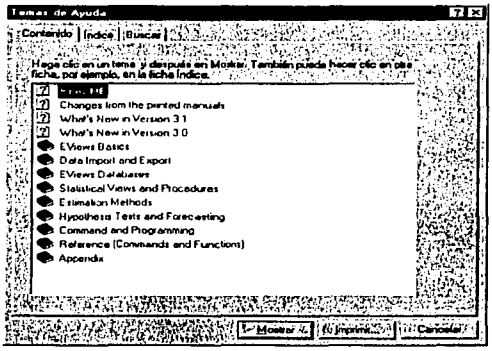
Range 2001 01 2015 12 Filter * Default Eq None

Sample 2001 01 2015 12

c
gdp
m1
pr
prueba
resid
rs
suma

Como puede apreciarse, la dinámica y flexibilidad de trabajar con Eviews es enorme, y aunque la finalidad de este trabajo es mostrar como se pueden hacer regresiones y corregir aquellos problemas que un análisis regresivo conlleva, aquí es imposible mostrar a detalle lo que se puede hacer con este paquete, sin embargo, cabe señalar que si no se cuenta con el Manual del Usuario, esta versión contiene un amplio y muy especializado soporte de ayuda, el cual se obtiene oprimiendo en cualquier momento la tecla F1 ó simplemente se le da clic en la opción **"HELP"** del

Menú Principal y entonces se dispondrá de las librerías que integran dicha ayuda mostrándose una ventana como la siguiente:



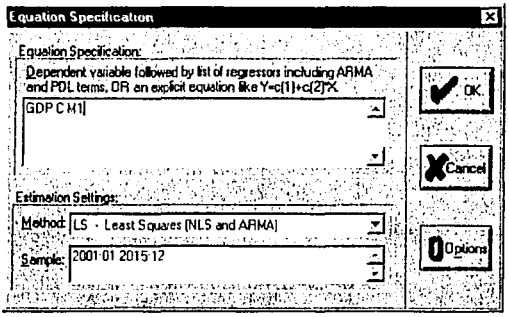
Por último, y muy brevemente, se muestra un ejemplo de como se puede llevar a cabo una regresión lineal con una serie de datos, para esto supóngase que la serie GDP está en función de las series M1, PR y RS, además supóngase que dicha relación puede expresarse de la siguiente manera:

$$GDP = C(1) + C(2)*M1$$

Donde

$C(1)=b_0$ es la ordenada al origen y
$C(2)=b_1$ es la pendiente de la recta o bien, en términos económicos, es el "multiplicador" de la variable M1.

Realmente el hacer regresiones lineales en este paquete es relativamente sencillo, al hacer clic en QUICK del Menú Principal se selecciona la opción de ESTIMATE EQUATION y entonces se presenta una caja de diálogo como la siguiente:



En esta caja en primer lugar se presenta un espacio para teclear la relación lineal que se desea, siendo la variable dependiente (GDP) la que se teclea primero seguida por la letra C en el caso de que se deseé una regresión con intercepto al origen y por último se teclea la variable independiente (M1), después de esto se escoge la opción Least Squares (Mínimos Cuadrados) y finalmente se selecciona el periodo (Sample) en el cual se va a llevar a cabo la regresión. Una vez especificado lo que se desea se le da clic al botón de OK y entonces el resultado de la estimación será como el siguiente:

EViews - [Equation UNTITLED Workfile PRUEBA]

File Edit Objects View Procs Quick Options Window Help

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: GDP
Method: Least Squares
Date: 04/17/01 Time: 21:59
Sample: 2001:01 2015:12
Included observations: 180

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-92.24297	6.751931	-13.66172	0.0000
M1	1.628430	0.012006	135.6364	0.0000

R-squared	0.990417	Mean dependent var	632.4190
Adjusted R-squared	0.990364	S.D. dependent var	564.2441
S.E. of regression	55.38936	Akaike info criterion	10.87770
Sum squared resid	546100.6	Schwarz criterion	10.91318
Log likelihood	-976.9931	F-statistic	18397.24
Durbin-Watson stat	0.077568	Prob(F-statistic)	0.000000

Como suele suceder en la mayoría de los casos, la ejecución del comando es relativamente sencilla, sin embargo, cuando se pretenden interpretar los resultados es donde surgen algunas confusiones. Sin tratar de confundir y de tal manera que la siguiente interpretación de los resultados sea clara y sencilla entonces se puede decir que; la línea de regresión lineal que mejor se ajusta al comportamiento de la variable GDP, siendo esta explicada por la variable M1, viene dada por la siguiente expresión:

$$\begin{array}{rl} GDP = -92.24297 + 1.628430\ M1 & r^2=0.990417 \\ (6.751931)\quad (0.012006) & \\ t\ (-13.66172)\quad (135.6364) & \end{array}$$

De esta manera, y una vez obtenidos estos resultados el siguiente paso es la interpretación de lo que se observa a simple vista. En primer lugar lo que se aprecia es que el coeficiente b_0= -92.24297 implica que; aún sin presentar cambios la variable M1, la serie GDP muestra una tasa negativa promedio de -92.24297 puntos porcentuales en cuanto a su valor tomado para el periodo de enero del 2001 a diciembre del 2015. De la misma manera puede verse que el coeficiente b_1= 1.628430 siendo este positivo implica que por cada unidad porcentual en que se incremente la variable M1 entonces la serie GDP aumentará en 1.628430 unidades porcentuales.

También puede apreciarse que la bondad de ajuste para este modelo uniecuacional es muy alta y logra explicar el 99.0417% del comportamiento de la variable GDP, es decir, el fenómeno en estudio es explicado correctamente tan sólo por la variable en consideración. Los errores estándar de los coeficientes son uno bajo y uno considerablemente grande, lo cual implica un intervalo pequeño y uno más grande para dichos coeficientes respectivamente; de la misma manera se observa que tanto la variable en consideración (M1) como el coeficiente independiente son estadísticamente significativos (conjuntamente) en la ecuación lo cual se confirma con una probabilidad del estadístico *F* menor al 5% para ambos coeficientes, mientras que el estadístico *Durbin-Watson* muestra claramente la existencia de correlación serial negativa (de primer orden) entre las variables involucradas.

A groso modo se puede decir que el modelo lineal no es representativo, sin embargo algunas medidas de corrección se mostrarán en el siguiente capítulo de este trabajo.

5 Ejemplo con el paquete estadístico Econometric Views V3.1

En este capítulo se mostrará y se analizará un banco de datos que de ninguna manera presenta algún sustento económico ni mucho menos refleja teoría económica alguna, simplemente estos datos serán usados con la finalidad de explicar mediante el paquete estadístico Eviews algunas de las pruebas y procedimientos estadísticos en un análisis econométrico, por lo que la cimentación de un ***marco teórico*** no se aplica para estos datos.

Sin sustento teórico alguno, lo que procede es la ***especificación del modelo*** o relación funcional que existe entre las variables involucradas. Para esto, se integró un ***banco de datos*** con cuatro series de tiempo que se nombraron VAR1, VAR2, VAR3 y VAR4 con 180 observaciones cada una de ellas.

El objetivo es tratar de explicar el comportamiento de la VAR1 mediante las variables restantes, es decir:

$$VAR1 = f(VAR2, VAR3, VAR4)$$

De esta manera, el ***modelo lineal propuesto*** viene dado por la siguiente expresión:

$$VAR1_t = b_0 + b_1\, VAR2_t + b_2\, VAR3_t + b_3\, VAR4_t + U_t \quad \text{con } t=1,2,\ldots,180$$

En este modelo lineal propuesto se asume lo siguiente:

a) El modelo a considerar es un modelo lineal en parámetros.
b) En el modelo no existe multicolinealidad entre las variables explicativas, o sea, la VAR2 no depende del comportamiento de la variable VAR3 ni de la variable VAR4 en el tiempo t, etc.
c) Las variables endógenas (VAR2, VAR3, VAR4) no son variables aleatorias y para fines de este estudio se tomarán como valores constantes.
d) El término estocástico U_t tiene una función de distribución de probabilidad normal con media cero y varianza igual a σ_u^2 I.

BANCO DE DATOS

EViews [Group GROUP01 Workfile DATOS]

File Edit Objects View Procs Quick Options Window Help

View | Proc | Object | Print | Name | Freeze | Edit+/- | Smpl+/- | InsDel | Transpose | Title | Sample

obs	VAR1	VAR2	VAR3	VAR4
1	87.87500	0.197561	126.5370	1.640000
2	88.12500	0.199167	127.5060	1.677667
3	89.62500	0.200179	129.3950	1.828667
4	92.67500	0.201246	128.5120	1.920667
5	94.62500	0.201952	130.5870	2.047333
6	95.55000	0.201444	130.3410	2.002667
7	95.42500	0.202236	131.3890	2.021667
8	94.17500	0.202723	129.8910	1.486333
9	94.07500	0.203416	130.1730	1.083667
10	94.20000	0.203841	131.3650	0.814333
11	95.49000	0.204291	134.6270	0.869667
12	97.36375	0.204374	134.2520	1.036333
13	100.7250	0.205603	136.4130	1.256333
14	102.8250	0.206227	136.4710	1.614333
15	104.9250	0.207762	138.3770	1.861333
16	106.6000	0.209998	137.2440	2.349333
17	107.2750	0.212048	138.0530	2.379333
18	109.6750	0.213329	138.3750	2.506667
19	109.8750	0.216140	138.9930	2.596667
20	112.1250	0.217055	139.0070	3.003667
21	114.3000	0.220072	139.5410	3.171667
22	114.7250	0.221468	139.5250	3.157000
23	116.5500	0.222636	139.9280	3.382333
24	115.4250	0.222957	138.4180	3.343333
25	113.4750	0.226529	139.6330	1.838000
26	114.6000	0.226404	139.6550	1.017667
27	118.0167	0.227799	143.1710	1.710667
28				

Path = c:\eviews3 | DB = none | WF = datos

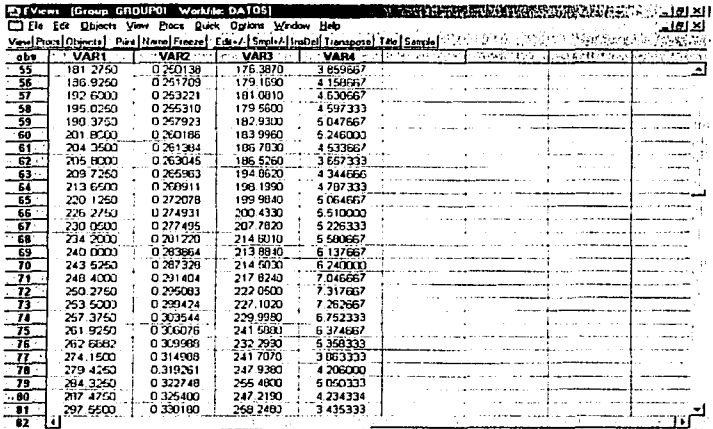
EViews [Group GROUP01 Workfile DATOS]

File Edit Objects View Procs Quick Options Window Help

View | Proc | Object | Print | Name | Freeze | Edit+/- | Smpl+/- | InsDel | Transpose | Title | Sample

obs	VAR1	VAR2	VAR3	VAR4
28	121.2750	0.228756	144.1120	2.787667
29	124.0750	0.229238	145.8600	2.800333
30	127.3250	0.229074	146.1400	3.019333
31	127.4000	0.229405	147.3960	3.533000
32	120.4250	0.230256	145.4830	4.299333
33	131.0250	0.231353	145.6990	3.943000
34	131.5250	0.230202	145.5990	3.092333
35	132.2250	0.233170	147.5250	2.390333
36	133.9250	0.234030	146.5270	2.360667
37	132.0250	0.234534	149.3190	2.376667
38	134.7250	0.235122	149.7360	2.324667
39	137.3750	0.235574	152.2370	2.324667
40	140.6250	0.236245	152.5250	2.476000
41	143.8250	0.237472	153.2740	2.739000
42	145.6500	0.237990	153.5320	2.716000
43	147.4750	0.238642	154.7270	2.868000
44	149.2000	0.239351	156.5950	2.803333
45	150.5250	0.240005	157.0800	2.909000
46	152.7250	0.240692	158.0560	2.941333
47	156.9250	0.241155	160.6090	3.280667
48	158.1500	0.242978	162.1270	3.499333
49	162.3250	0.243484	162.7750	3.538000
50	164.5750	0.244041	163.8810	3.481333
51	167.3000	0.245164	169.9330	3.504000
52	169.0000	0.246450	170.7350	3.685000
53	173.9250	0.247677	170.0200	3.899667
54	177.0500	0.248814	170.2630	3.879000
55				

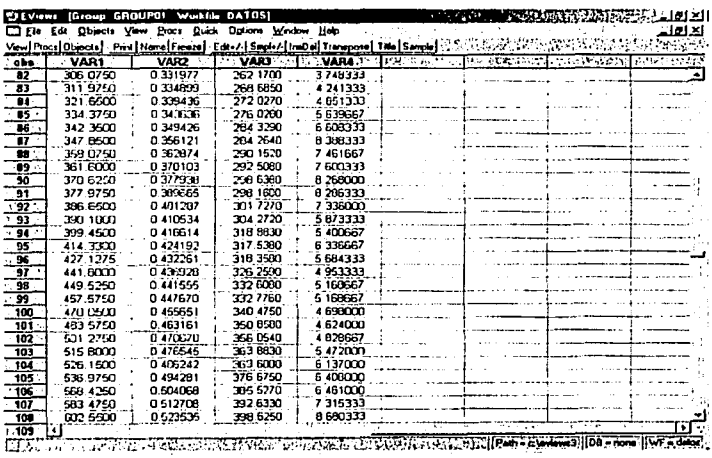
Path = c:\eviews3 | DB = none | WF = datos



View | Proc | Objects | Print | Name | Freeze | Edit+/- | Smpl+/- | InsDel | Transpose | Title | Sample

obs	VAR1	VAR2	VAR3	VAR4
55	181.2750	0.250138	176.3870	3.859667
56	186.9250	0.251709	179.1690	4.158667
57	192.6000	0.253221	181.0810	4.630667
58	195.0250	0.255310	179.5600	4.597333
59	190.3750	0.257923	182.9300	5.047667
60	201.8000	0.260186	183.9960	5.246000
61	204.3500	0.261384	186.7030	4.533667
62	205.8000	0.263045	186.5260	3.657333
63	209.7250	0.265983	194.0620	4.344666
64	213.6500	0.269911	198.1990	4.797333
65	220.1250	0.272078	199.9840	5.064667
66	226.2750	0.274931	200.4330	5.510000
67	230.0500	0.277495	207.7820	5.226333
68	234.2000	0.281220	214.8010	5.580667
69	240.0000	0.283864	213.8840	6.137667
70	243.5250	0.287328	214.6030	6.240000
71	248.4000	0.291404	217.8240	7.046667
72	250.2750	0.295083	222.0500	7.317667
73	253.5000	0.299424	227.1020	7.262667
74	257.3750	0.303544	229.9980	6.752333
75	261.9250	0.306076	241.5980	6.374667
76	262.6667	0.309968	232.2930	5.358333
77	274.1500	0.314988	241.7070	3.863333
78	279.4250	0.319261	247.9380	4.206000
79	284.3250	0.322748	256.4800	5.050333
80	287.4750	0.326400	247.2190	4.234334
81	297.5500	0.330180	268.2480	3.435333
82				

Path = c:\eviews3 | DB = none | WF = datos



View | Proc | Objects | Print | Name | Freeze | Edit+/- | Smpl+/- | InsDel | Transpose | Title | Sample

obs	VAR1	VAR2	VAR3	VAR4
82	306.0750	0.331977	262.1700	3.749333
83	311.9750	0.334899	268.6850	4.241333
84	321.6500	0.339436	272.0270	4.651333
85	334.3750	0.343636	276.0200	5.639667
86	342.3500	0.349426	284.3290	6.608333
87	347.8500	0.356121	284.2640	8.388333
88	359.0750	0.362874	290.1520	7.461667
89	361.6000	0.370103	292.5080	7.600333
90	370.6250	0.377938	296.6360	8.268000
91	377.9750	0.389665	298.1600	8.206333
92	386.6500	0.401207	301.7270	7.336000
93	390.1000	0.410534	304.2720	5.873333
94	399.4500	0.416614	318.8830	5.400667
95	414.3000	0.424192	317.5380	6.336667
96	427.1375	0.432261	318.3500	5.684333
97	441.8000	0.439328	326.2590	4.953333
98	449.5250	0.441555	332.6000	5.160667
99	457.5750	0.447670	332.7760	5.168667
100	470.0500	0.455651	340.4750	4.698000
101	483.5750	0.463161	350.8500	4.624000
102	501.2750	0.470870	356.0540	4.828667
103	515.8000	0.476545	363.8830	5.472000
104	526.1500	0.486242	363.6000	6.137000
105	536.9750	0.494281	376.6750	6.408000
106	569.4250	0.504068	385.5770	6.461000
107	583.4750	0.512708	392.6330	7.315333
108	602.5500	0.523536	398.6250	8.680333
109				

Path = c:\eviews3 | DB = none | WF = datos

EViews - [Group GROUP01 Workfile DATOS]

File Edit Objects View Procs Quick Options Window Help

View | Procs | Objects | Print | Name | Freeze | Edit+/- | Smpl+/- | InsDel | Transpose | Title | Sample |

obs	VAR1	VAR2	VAR3	VAR4
109	616.1500	0.636119	403.0990	9.357667
110	630.5500	0.646451	414.6290	9.372334
111	648.2250	0.658725	426.9140	9.631333
112	662.5500	0.669177	434.9300	11.80967
113	680.5500	0.681791	434.6280	13.45867
114	679.4825	0.695176	443.6310	10.04933
115	695.8500	0.610007	454.0170	9.235333
116	727.7755	0.625788	465.7540	13.70967
117	760.0750	0.641522	465.1870	14.36900
118	767.9264	0.654000	469.4660	14.82900
119	791.9250	0.666463	472.3550	15.08733
120	796.4000	0.670682	483.6490	12.02267
121	794.7000	0.688663	483.3400	12.89500
122	807.0750	0.697210	492.8670	12.36900
123	814.7750	0.706641	496.7580	9.705334
124	824.7500	0.714332	514.6210	7.936000
125	840.2250	0.720759	533.9010	8.081333
126	867.3000	0.728273	546.2220	8.420667
127	890.8250	0.734653	556.1000	9.186666
128	916.1250	0.741922	560.7670	8.793333
129	947.7500	0.750158	586.7270	9.133333
130	969.9000	0.756830	591.1820	9.843333
131	986.5500	0.762470	587.6940	10.34333
132	999.1500	0.768030	601.9560	8.973333
133	1020.300	0.776792	611.5350	8.183333
134	1033.650	0.782520	626.9290	7.523334
135	1055.375	0.787652	662.2570	7.103333
136				

Path = c:\eviews3 | DB = none | WF = datos

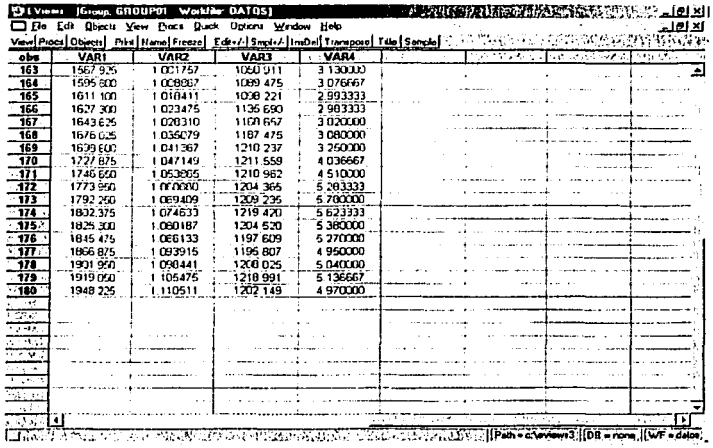
EViews - [Group GROUP01 Workfile DATOS]

File Edit Objects View Procs Quick Options Window Help

View | Procs | Objects | Print | Name | Freeze | Edit+/- | Smpl+/- | InsDel | Transpose | Title | Sample |

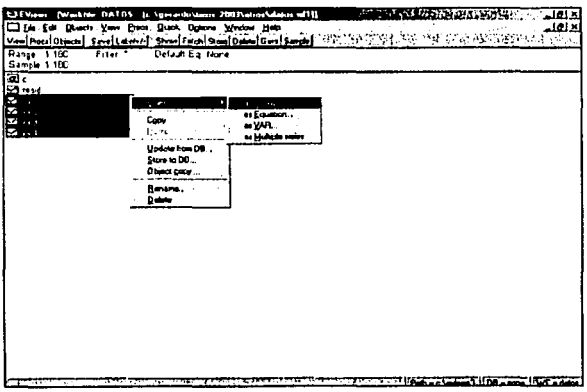
obs	VAR1	VAR2	VAR3	VAR4
136	1071.325	0.794516	671.3910	7.146667
137	1093.525	0.798070	696.9050	6.886667
138	1096.375	0.802191	729.4950	6.130000
139	1110.075	0.808463	743.6910	5.533333
140	1125.450	0.814540	796.0230	5.340000
141	1141.400	0.820895	779.5550	5.533333
142	1161.275	0.826824	797.9650	5.733333
143	1180.675	0.833310	807.1650	6.033333
144	1209.000	0.840956	808.0580	6.003334
145	1224.525	0.846646	823.8310	5.760000
146	1250.075	0.856631	842.6710	6.230000
147	1273.650	0.866620	844.2460	6.993333
148	1301.350	0.874446	840.1540	7.703333
149	1329.225	0.884528	839.7360	8.533334
150	1353.275	0.893900	843.5020	8.440000
151	1371.750	0.901281	844.7470	7.850000
152	1394.425	0.908788	852.3150	7.633333
153	1416.175	0.920650	863.0090	7.756667
154	1437.225	0.931833	875.8290	7.766667
155	1445.575	0.941421	882.5480	7.493333
156	1445.450	0.961110	887.7420	7.023334
157	1455.475	0.962695	900.8960	6.053333
158	1473.100	0.969090	914.3580	5.593333
159	1487.575	0.977045	940.5670	5.406667
160	1501.500	0.983000	959.7510	4.583333
161	1530.450	0.991272	1002.641	3.910000
162	1560.275	0.997892	1014.978	3.723333
163				

Path = c:\eviews3 | DB = none | WF = datos



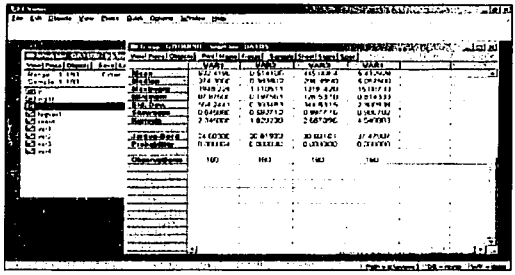
obs	VAR1	VAR2	VAR3	VAR4
163	1567 925	1 001757	1090 911	3 130000
164	1595 800	1 008887	1089 475	3 076667
165	1611 100	1 010411	1098 221	2 993333
166	1627 300	1 023475	1135 690	2 983333
167	1643 625	1 020310	1160 667	3 020000
168	1676 025	1 035079	1187 475	3 080000
169	1699 600	1 041367	1210 237	3 250000
170	1727 875	1 047149	1211 559	4 036667
171	1746 650	1 053895	1210 982	4 510000
172	1773 950	1 060000	1204 365	5 283333
173	1792 250	1 069409	1209 235	5 700000
174	1802 375	1 074633	1219 420	5 623333
175	1825 300	1 080187	1204 520	5 380000
176	1845 475	1 086133	1197 609	5 270000
177	1866 875	1 093915	1196 807	4 950000
178	1901 950	1 098441	1206 025	5 040000
179	1919 050	1 105475	1218 991	5 136667
180	1948 225	1 110511	1202 149	4 970000

Para esto, se ha creado un grupo de trabajo (group01) de la siguiente manera:

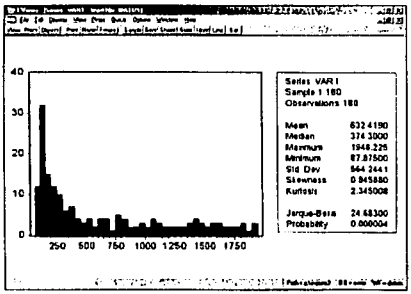


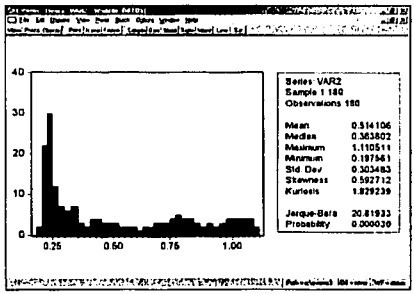
A continuación se muestran las estadísticas descriptivas para cada una de las variables con rango en forma individual, para ello se debe de seguir la

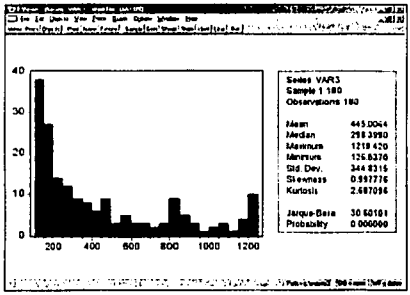
siguiente instrucción VIEW\DESCRIPTIVE STATS\INDIVIDUAL SAMPLES:

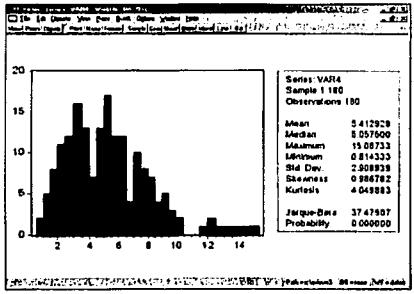


Y de forma individual se obtiene



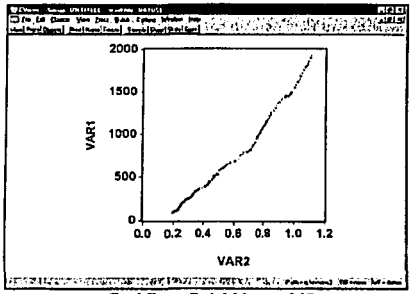




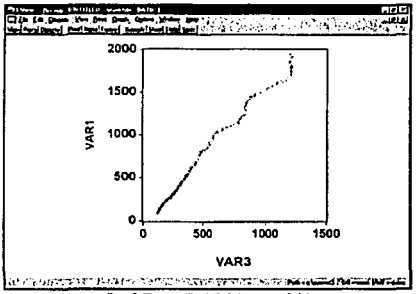


ANALISIS DE LOS GRAFICOS

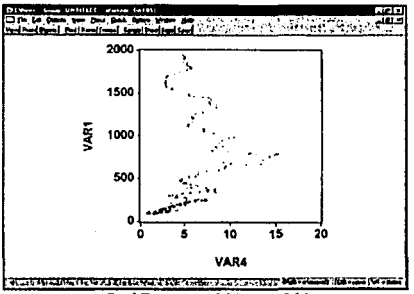
Para visualizar como se relacionan las variables explicativas con la variable dependiente a continuación se muestran tres gráficos que a groso modo dan una idea del comportamiento entre dichas variables:



Gráfica 5.1 V1 vs. V2



Gráfica 5.2 V1 vs. V3



Gráfica 5.3 V1 vs. V4

A simple vista se puede decir que la VAR1 presenta una relación positiva con la VAR2 y la VAR3, mientras que con la VAR4 no muestra algún patrón específico de comportamiento.

Esto puede corroborarse al hacer uso de la matriz de correlación que a continuación se muestra:

EViews [Group UNTITLED Workfile DATOS]

File Edit Objects View Proc Quick Options Window Help

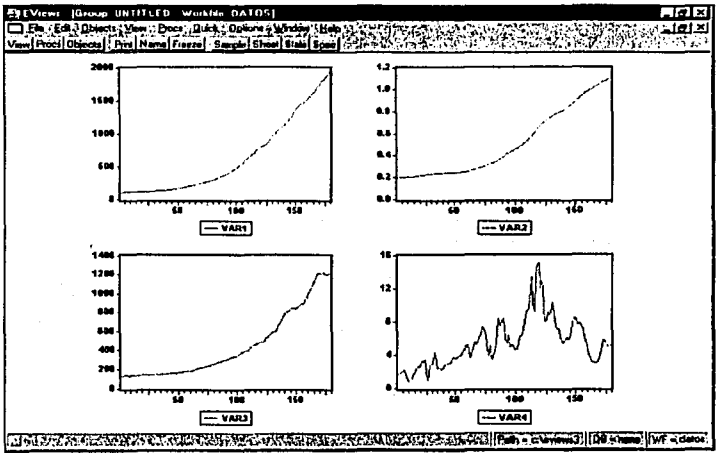
View Proc Objects Print Name Freeze Sample Sheet Stats Spec

Correlation Matrix

	VAR1	VAR2	VAR3	VAR4
VAR1	1.000000	0.992475	0.995197	0.333494
VAR2	0.992475	1.000000	0.980402	0.412471
VAR3	0.995197	0.980402	1.000000	0.270059
VAR4	0.333494	0.412471	0.270059	1.000000

Esta matriz de correlación confirma la relación positiva entre las variables VAR1, VAR2 y VAR3, pues el coeficiente de correlación tomado de 2 en 2 es superior a 0.9, mientras que de la VAR4 se puede decir que no muestra ninguna relación con la VAR1, de hecho con ninguna de las variables en cuestión, lo cual sugiere que la VAR4 podría ser descartada del modelo.

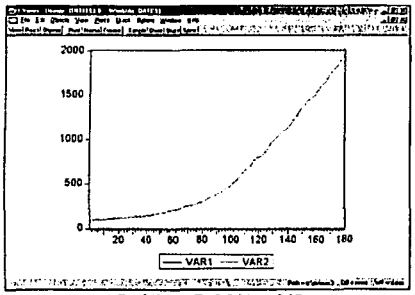
El comportamiento a través del tiempo y la posible relación entre éstas variables puede verse mejor con la ayuda del siguiente gráfico múltiple:



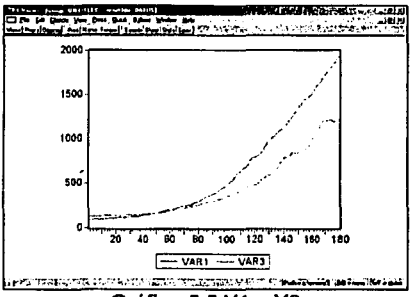
Gráfica 5.4 Variables VAR1, VAR2, VAR3 y VAR4

Es claro que la VAR1 presenta un comportamiento de tipo exponencial positiva, por lo que esto sugiere una mejor adecuación de los datos, es decir, trabajar con datos transformados. Simultáneamente puede apreciarse que las variables VAR2 y VAR3 presentan el mismo comportamiento exponencial que la VAR1 (aunque con otro tipo de escala), mientras que la VAR4 por su comportamiento irregular no parece mostrar ningún tipo de interés hacia la VAR1.

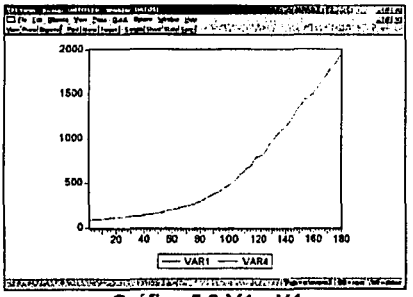
Ahora se mostrará el comportamiento gráfico de la variable VAR1 con cada una de las demás variables:



Gráfica 5.4 V1 y V2



Gráfica 5.5 V1 y V3



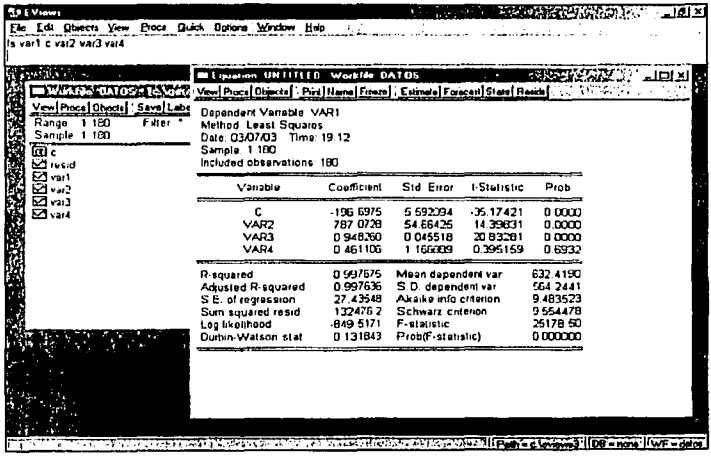
Gráfica 5.6 V1 y V4

Aunque no puede apreciarse fácilmente la relación entre las variables VAR1 y VAR2, por tener ésta última una escala mucho menor que la primera, se sabe que la variable VAR2 presenta un comportamiento similar al de la variable VAR1 y también al de la variable VAR3. Finalmente, y como se mencionó en el párrafo anterior, la variable VAR1 presenta (gráficamente) mejor relación con las variables VAR2 y VAR3, mientras que definitivamente la VAR4 desconoce a la variable VAR1.

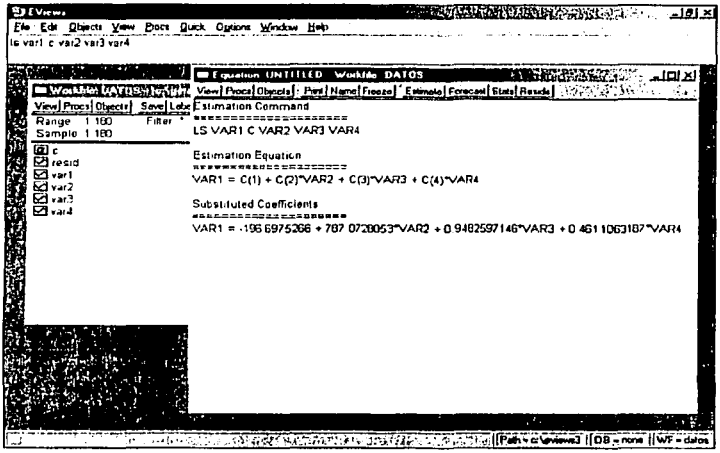
ESTIMACION

Ya que se tiene una idea de cómo se interrelacionan las variables en estudio, a continuación se muestra la regresión con los datos sin transformar (datos en bruto o datos originales), para esto, en la línea de comando se teclea la expresión LS VAR1 C VAR2 VAR3 VAR4 y después se oprime la tecla  .

El resultado de la estimación será el siguiente:



Para ver la representación del modelo, lo que se tiene que hacer es darle un clic en el botón VIEW y escoger la opción REPRESENTATIONS, entonces la pantalla que se desplegará será como la siguiente:



También puede verificarse la matriz de covarianza de los coeficientes estimados, para esto se le da un clic en el botón VIEW y se selecciona la opción COVARIANCE MATRIX, por lo que se desplegará una ventana como la siguiente:

EViews - [Equation EQ01 Workfile DATOS]

File Edit Objects View Procs Quick Options Window Help

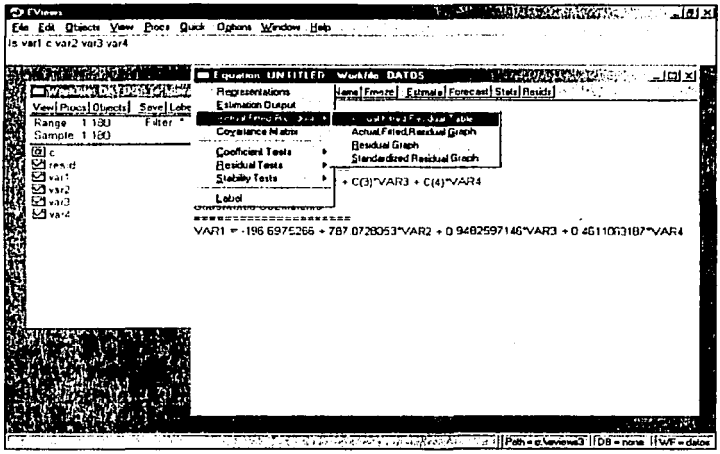
View Procs Objects Print Name Freeze Estimate Forecast Stats Resids

Coefficient Covariance Matrix

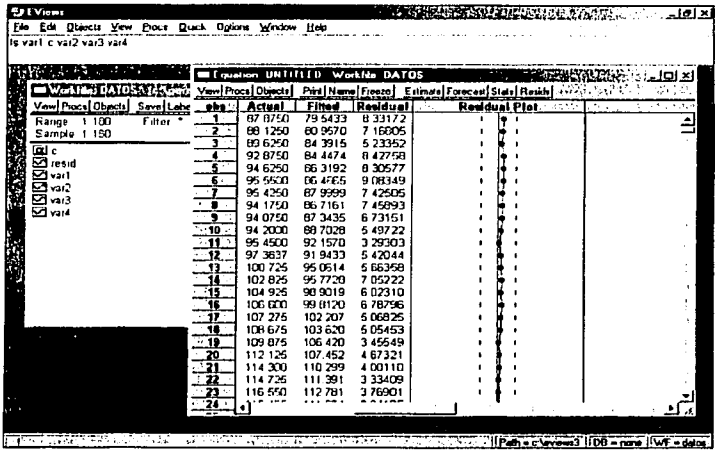
	C	VAR2	VAR3	VAR4
C	31.27152	-170.3670	0.130184	0.473748
VAR2	-170.3670	2988.180	-2.465175	-49.66926
VAR3	0.130184	-2.465175	0.002072	0.039755
VAR4	0.473748	-49.66926	0.039755	1.361631

Aquí puede observarse lo siguiente: el término constante **C** tiene una relación en sentido opuesto con el coeficiente de la variable VAR2, mientras que con los coeficientes de las variables VAR3 y VAR4 tiene una relación en el mismo sentido, es decir positiva. El coeficiente de la variable VAR2 se relaciona negativamente con los coeficientes de VAR3 y VAR4, mientras que el coeficiente de VAR3 tiene una relación en igual sentido con el coeficiente de la variable VAR4.

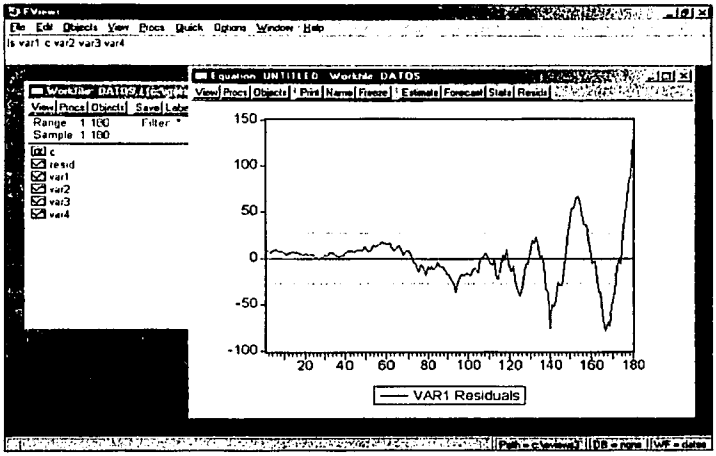
También pueden apreciarse los valores que toma cada observación en el modelo propuesto, su estimado y los residuales de la siguiente manera:



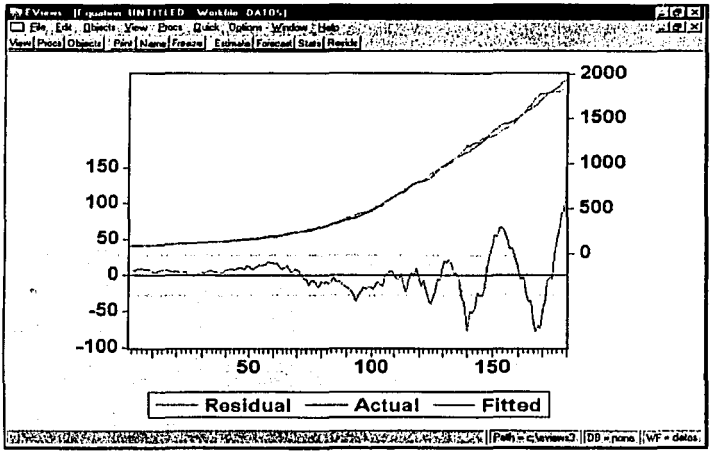
Por lo que se podrá apreciar en pantalla algo como lo que a continuación se muestra:



Y gráficamente los residuales son de la siguiente manera:



Ahora bien, para visualizar el ajuste del modelo y el comportamiento de los residuales en el periodo de estudio, tan sólo se le da un clic en el botón **RESIDS** y entonces la pantalla será:



Gráficamente se pueden apreciar dos cosas; al comienzo se tiene un buen ajuste del modelo mientras que al transcurso del tiempo éste pierde sensibilidad y por consiguiente no capta la suficiente información de los datos en cuestión, esto lo confirma el comportamiento creciente de la dispersión de los residuales al traspasar la línea delimitada por el error estándar de la regresión, además de presentar estos últimos un patrón definido a través del tiempo.

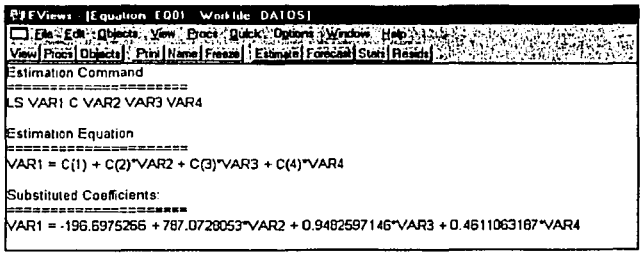
Interpretación de la estimación

Como se había comentado con anterioridad, el llevar a cabo regresiones en este "paquete" es relativamente sencillo, sin embargo, la interpretación y validación son un poco más elaborados. De esta manera y partiendo del modelo lineal propuesto, la interpretación de los resultados es la siguiente:

El objetivo es encontrar la mejor "*recta*" que se ajuste al acomodo de los datos de la VAR1 conforme transcurre el tiempo mediante la siguiente expresión:

$$VAR1_t = b_0 + b_1\, VAR2_t + b_2\, VAR3_t + b_3\, VAR4_t + U_t \qquad \text{con } t=1,2,....,180$$

Así, con base en el banco de datos aquí expuesto, la estimación hecha para la variable VAR1 en Eviews viene dada como:



Es decir, la variable VAR1 es explicada de la siguiente manera:

$$VAR1_t = -196.6975266 + 787.0728053\, VAR2_t + 0.9482597146\, VAR3_t + 0.4611063187\, VAR4_t + e_t$$

ES=	(5.592094)	(54.66425)	(0.045518)	(1.166889)
t=	(-35.17421)	(14.39831)	(20.83281)	(0.395159)

con t=1,2,3,.....,180

A partir de esta descripción se desprende la siguiente interpretación:

1. El valor promedio esperado en el periodo t=1 a t=180 asciende a –197 unidades en las que se mida la variable **VAR1**.
2. Por cada unidad que se incremente la variable **VAR2** esto conllevará a un incremento de 787 unidades de la variable **VAR1**, esto si las variables **VAR3** y **VAR4** permanecen constantes.
3. Si las variables **VAR2** y **VAR4** permanecen constantes entonces el incremento de una unidad de la variable **VAR3** producirá un incremento de casi una unidad de la variable **VAR1**.
4. De la misma manera, si las variables **VAR2** y **VAR3** permanecen constantes entonces el incremento de una unidad en la variable **VAR4** repercutirá en un incremento de media unidad de la variable **VAR1**.

Con base en los resultados de la regresión se puede decir lo siguiente:

1. El ajuste del modelo r^2, o la bondad del modelo, representa el 99.76% del comportamiento de la variable **VAR1**, el cual a simple vista es demasiado bueno, pues quiere decir que las variables explicativas representan casi en su totalidad a la variable **VAR1** en el periodo de estudio.
2. El r^2 ajustado es muy similar al r^2 lo cual no contradice la representatividad del modelo.
3. El error estándar del modelo asciende a 27.5 unidades en promedio durante el periodo t=1 y t=180.
4. Los estadísticos ***t's*** para los coeficientes de **C, VAR2 y VAR3** son estadísticamente representativos individualmente, mientras que el estadístico ***t*** para el coeficiente de **VAR4** implica no representatividad. Esto se confirma con la probabilidad calculada para dichos parámetros, pues al ser menor que el 5% implica significancia estadística para los parámetros **C, VAR2 y VAR3**, mientras que para **VAR4** no se cumple.
5. El error estándar promedio para cada valor estimado de la variable **VAR1** asciende a ± 564 unidades.
6. El estadístico ***F*** implica que en conjunto los parámetros de **C, VAR2, VAR3** y **VAR4** son estadísticamente significativos, lo cual se sostiene al observar que la probabilidad de dicho estadístico es menor al 5%.
7. Finalmente la prueba **d** de *Durbin-Watson* al ser muy cercana a 0 implica la existencia de correlación serial de primer orden.

Si lo que realmente importara fuese el ajuste del modelo a los datos, éste modelo estimado cumple con ello, pues el ajuste es del 99% para el periodo de estudio, sin embargo a continuación se mostrará la validez del mismo.

EVALUACION

Evaluación Estadística

Para los resultados obtenidos en la regresión hecha mediante el paquete Eviews se probará los siguiente:

$$H_0: \beta_i = 0 \qquad vs. \qquad H_A: \beta_i \neq 0 \qquad \forall\ i = 1,2,3,4$$

Esta prueba de hipótesis se hará para confirmar la significancia de los parámetros estimados de forma individual, donde el estadístico de prueba será en función de una distribución de probabilidad ***t*** de *student* como la siguiente:

$$t_C = \frac{\hat{\beta}_i - \beta_i}{ES(\hat{\beta}_i)} = \frac{estimador - parámetro}{\text{error\_estándar\_del\_estimador}} \qquad \text{donde } t_C \sim t_{(n-k)}$$

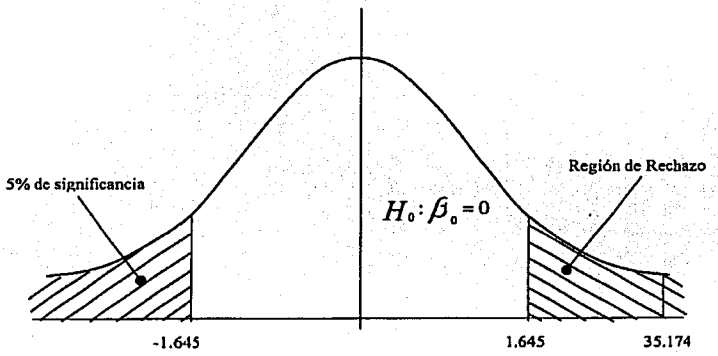
a) Para el coeficiente de **C** (β_0) se tiene

$$t_C = \frac{-196.6975266 - 0}{5.592094} = 35.1742167 \text{ (}\boldsymbol{t}\text{ calculada)}$$

$$t_{180-4}^{0.05} = t_{176}^{0.05} = 1.645 \text{ (}\boldsymbol{t}\text{ de tablas con un nivel de significancia del 5\%)}$$

de esta manera se tiene que $t_C > t_{176}^{0.05}$ lo cual implica que se rechaza la hipótesis nula $H_0: \beta_0 = 0$ a un nivel de significancia del 5%, entonces β_0 es estadísticamente significativo en el modelo.

Gráficamente



b) Para el coeficiente de **VAR2** (β_1) se tiene

$$t_c = \frac{787.0728053 - 0}{54.66425} = 14.3983097 \text{ (}t\text{ calculada)}$$

$$t^{0.05}_{180-4} = t^{0.05}_{176} = 1.645 \text{ (}t\text{ de tablas con un nivel de significancia del 5\%)}$$

de esta manera se tiene que $t_c > t^{0.05}_{176}$ lo cual implica que se rechaza la hipótesis nula $H_0: \beta_1 = 0$ a un nivel de significancia del 5%, por lo que β_1 es estadísticamente significativo en el modelo.

c) Para el coeficiente de **VAR3** (β_2) se tiene

$$t_c = \frac{0.9482597146 - 0}{0.045518} = 20.8326313 \text{ (}t\text{ calculada)}$$

$$t^{0.05}_{180-4} = t^{0.05}_{176} = 1.645 \text{ (}t\text{ de tablas con un nivel de significancia del 5\%)}$$

de esta manera se tiene que $t_c > t_{176}^{0.05}$ lo cual implica que se rechaza la hipótesis nula $H_0: \beta_2 = 0$ a un nivel de significancia del 5%, entonces β_2 es estadísticamente significativo en el modelo.

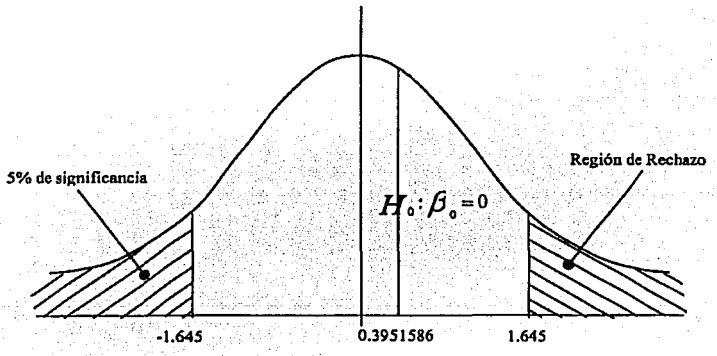
d) Para el coeficiente de **VAR4** (β_3) se tiene

$$t_c = \frac{0.4611063187 - 0}{1.166889} = 0.3951586 \text{ (}t\text{ calculada)}$$

$$t_{180-4}^{0.05} = t_{176}^{0.05} = 1.645 \text{ (}t\text{ de tablas con un nivel de significancia del 5%)}$$

de esta manera se tiene que $t_c < t_{176}^{0.05}$ lo cual implica que no se rechaza la hipótesis nula $H_0: \beta_3 = 0$ a un nivel de significancia del 5%, por lo que β_3 no es estadísticamente significativo en el modelo.

Gráficamente



Ahora se mostrará la significancia conjunta de los coeficientes estimados para **C, VAR2, VAR3 y VAR4**. Esto se basará en el estadístico de prueba F_c cuya función de distribución es una $F_\alpha(k-1, n-k)$.

Para los resultados obtenidos en la regresión hecha mediante el paquete Eviews se probará los siguiente:

$$H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0 \quad vs. \quad H_A: \beta_i \neq 0 \quad \forall\ i = 1,2,3,4$$

Aquí la ***F*** calculada viene dada por la siguiente expresión

$$F_c = \frac{ESS/(k-1)}{RSS/(n-k)} \sim F_\alpha(k-1, n-k)$$

donde

ESS (Suma de Cuadrados de la Estimación)
RSS (Suma de Cuadrados de los Residuales)
K es e número de coeficientes estimados en la regresión
n es el número total de observaciones o periodo de estudio

de esta manera la regla de decisión es, si

$$F_c > F_\alpha(k-1, n-k)$$

entonces se rechaza $H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ a un nivel de significancia α.

Se sabe que la Suma Total de Cuadrados es igual a la Suma de Cuadrados de la Estimación más la Suma de Cuadrados de los Residuales, es decir:

$$TSS = ESS + RSS$$

Lo cual implica que

$$ESS = TSS - RSS$$

Entonces

$$\boldsymbol{ESS} = 56988477.511188 - 132476.2 = 56856001.311188$$
$$\boldsymbol{RSS} = 132476.2$$
$$\boldsymbol{k} = 5$$
$$\boldsymbol{n} = 180$$

por lo que

$$F_c = \frac{56856001.311188/(5-1)}{132476.2/(180-4)} = \frac{18952000.4370627}{752.705681818182} = 25178.500567823$$

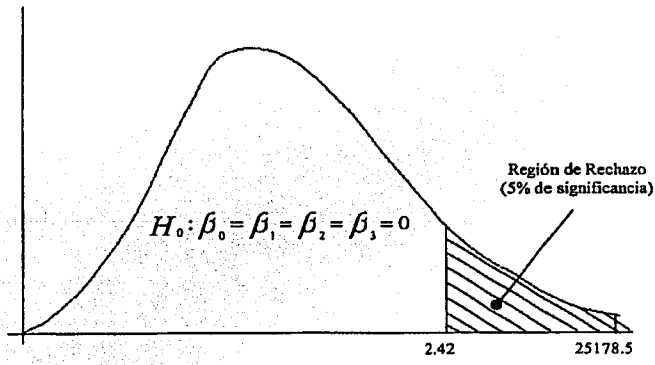
$$F_\alpha(k-1, n-k) = F_{0.05}(4,176) \approx 2.42$$

de aquí se deduce que

$$F_c > F_\alpha(k-1, n-k)$$

lo cual implica que se rechaza $H_0: \beta_0 = \beta_1 = \beta_2 = \beta_3 = 0$ a un nivel de significancia del 5%, por lo que los coeficientes estimados en conjunto son estadísticamente significativos para el modelo.

Gráficamente



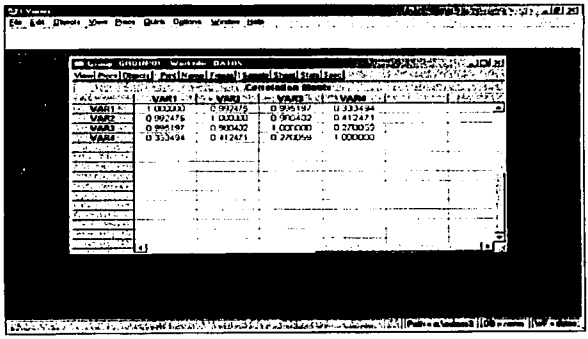
Hasta este momento se ha mostrado y confirmado lo siguiente:

1. Los coeficientes estimados para las variables **C, VAR2** y **VAR3** son estadísticamente significativos en el modelo, en forma individual.
2. El coeficiente estimado de la variable **VAR4** no es representativo en el modelo, y como se había mencionado con anterioridad, dicha variable puede ser omitida en el modelo.
3. Los coeficientes estimados son en conjunto estadísticamente significativos para el modelo, por lo que la variable **VAR1** está explicada en un 99% por las variables en cuestión. Dicho porcentaje implica un buen ajuste de la *línea de regresión* con respecto a la dispersión de los datos de la variable **VAR1**, concluyendo la representatividad del modelo propuesto, sin embargo y aunque es una proporción muy alta también puede ser un indicador de problemas de otra naturaleza.

Evaluación Econométrica

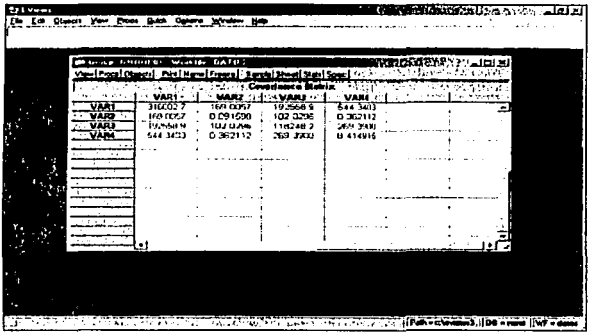
Multicolinealidad

Para evaluar la existencia de dependencia lineal entre las variables en estudio primero se analizarán las correlaciones parciales haciendo uso de la matriz de correlación siguiente:

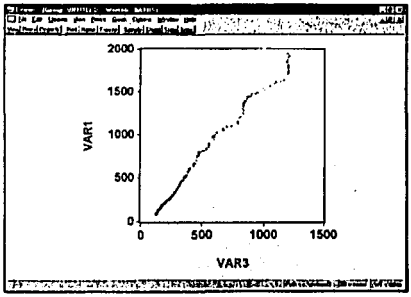


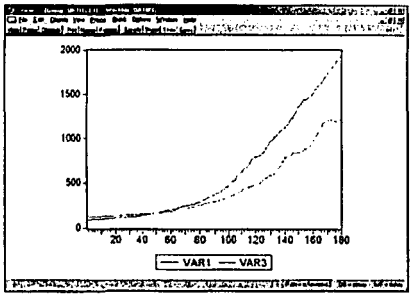
Como puede observarse, las variables **VAR1**, **VAR2** y **VAR3** presentan una alta dependencia lineal, esto debido a una correlación muy cercana a 1, destacando ligeramente la relación entre las variables **VAR1** y **VAR3**, mientras que la variable **VAR4** casi no presenta una codependencia lineal con las demás variables, de hecho con la variable **VAR2** es con la que muestra mayor interrelación 0.412471.

Esto puede corroborarse al hacer uso de la matriz de varianza-covarianza siguiente:



Aquí se confirma la codependencia lineal, y que de hecho resulta ser positiva, entre las variables **VAR1** y **VAR3** al obtenerse una covarianza demasiado alta, este hecho se venía apreciando desde el análisis de los gráficos 5.2 y 5.5 respectivamente.





Ahora se hará uso de la prueba de *Klein*, pues ésta se considera más general que el análisis de correlaciones parciales. Para aplicar la prueba de *Klein* es necesario observar lo siguiente:

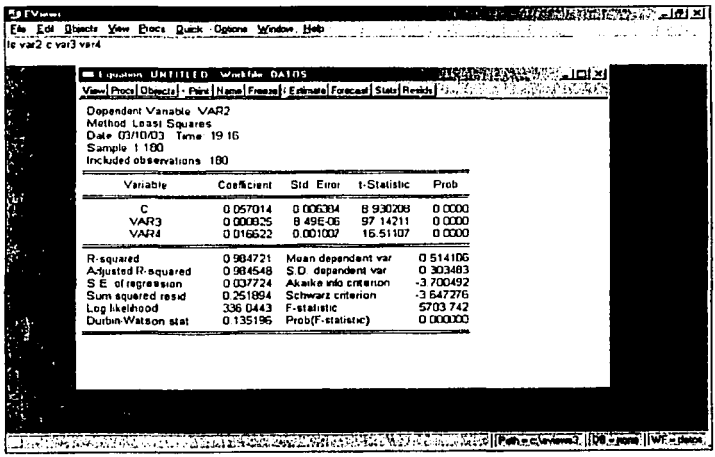
- ☑ El coeficiente de determinación r^2 es alto y las ***t's*** calculadas en la estimación son bajas.
- ☑ Si el r^2 de la regresión principal es menor o igual a cada coeficiente de determinación parcial de las variables explicativas, es decir

$$r^2 \leq r_j^2$$

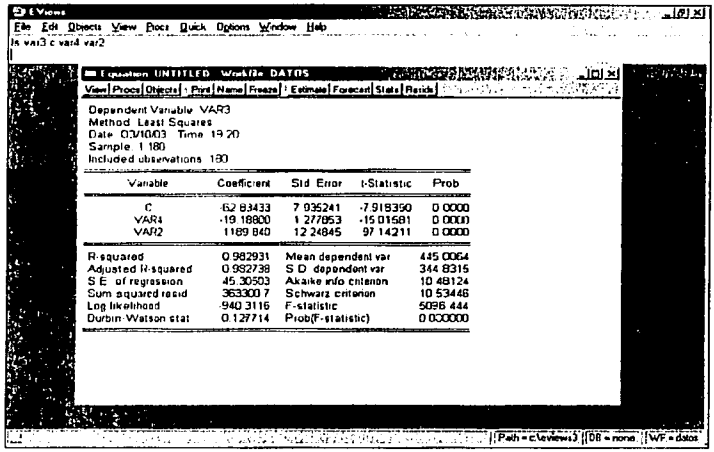
donde r_j^2 es el coeficiente de determinación de la variable *j* contra las demás variables explicativas.

Para el modelo se obtuvo un $r^2 = .997675$ ahora se mostrarán las regresiones parciales y sus respectivos coeficientes de determinación.

En la regresión parcial LS VAR2 C VAR3 VAR4 se obtuvo:



En la regresión parcial LS VAR3 C VAR2 VAR4 se obtuvo:



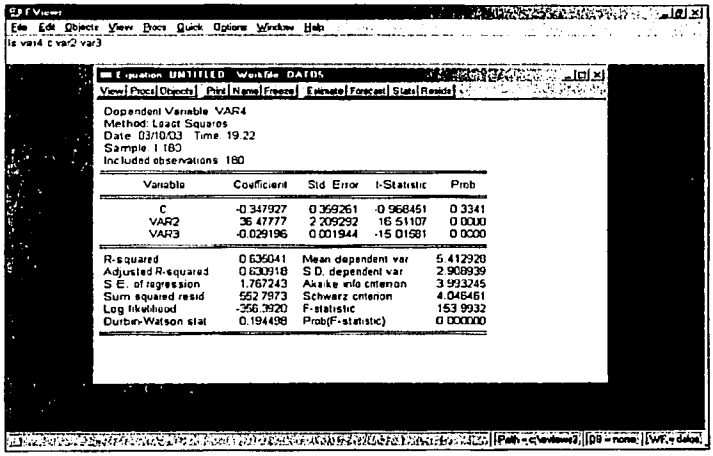
Equation: UNTITLED Workfile: DATOS

Dependent Variable: VAR3
Method: Least Squares
Date: 03/10/03 Time: 19:20
Sample: 1 180
Included observations: 180

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	62.83433	7.935241	-7.918350	0.0000
VAR4	-19.18800	1.277853	-15.01581	0.0000
VAR2	1189.840	12.24845	97.14211	0.0000

R-squared	0.982931	Mean dependent var	445.0064
Adjusted R-squared	0.982738	S.D. dependent var	344.8315
S.E. of regression	45.30503	Akaike info criterion	10.48124
Sum squared resid	363300.7	Schwarz criterion	10.53446
Log likelihood	-940.3116	F-statistic	5096.444
Durbin-Watson stat	0.127714	Prob(F-statistic)	0.000000

Y en la regresión parcial LS VAR4 C VAR2 VAR3 se obtuvo;



Equation: UNTITLED Workfile: DATOS

Dependent Variable: VAR4
Method: Least Squares
Date: 03/10/03 Time: 19:22
Sample: 1 180
Included observations: 180

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.347927	0.359261	-0.968451	0.3341
VAR2	36.47777	2.209292	16.51107	0.0000
VAR3	-0.029196	0.001844	-15.01581	0.0000

R-squared	0.635041	Mean dependent var	5.412928
Adjusted R-squared	0.630918	S.D. dependent var	2.908939
S.E. of regression	1.767243	Akaike info criterion	3.993245
Sum squared resid	552.7873	Schwarz criterion	4.046461
Log likelihood	-356.3920	F-statistic	153.9932
Durbin-Watson stat	0.194498	Prob(F-statistic)	0.000000

De esta manera se obtuvieron los siguientes resultados:

$$r^2 = .997675$$
$$r_1^2 = .984721$$
$$r_2^2 = .982931$$
$$r_3^2 = .635041$$

aunque puede decirse que $r^2 \leq r_j^2$ para j = 1,2,3 se aprecian evidencias de multicolinealidad severa en el modelo lineal propuesto, sobre todo entre las variables **VAR2** y **VAR3**, esto debido a que los coeficientes de determinación parciales más altos fueron para r_1^2 y r_2^2, mientras que el valor alto del estadístico *t* en cada regresión parcial respectivamente, confirma la representatividad de cada variable y por consiguiente la existencia de una multicolinealidad muy alta entre dichas variables.

Por lo tanto, y como se mencionó con anterioridad, se confirma la codependencia lineal entre las variables **VAR1**, **VAR2** y **VAR3**.

Heteroscedasticidad

Para verificar si hay evidencia de heteroscedasticidad en el modelo lineal propuesto se usará primero la prueba *Goldfeld-Quandt*:

Siendo
n = 180 (número de observaciones para cada variable explicativa)
m =20 (número total de observaciones excluidas transversalmente)

entonces la prueba de hipótesis será

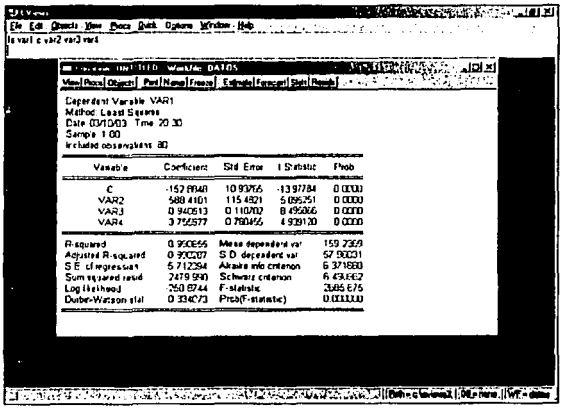
H_0: No existe heteroscedasticidad *vs.* H_A: Existe heteroscedasticidad

y el estadístico de prueba vendrá dado por la siguiente expresión:

$$F_c = \frac{RSS_1}{RSS_2} \sim F_\alpha\left(\frac{n-m}{2}, \frac{n-m}{2}\right)$$

en donde RSS_1 es la suma de residuales más grande y RSS_2 es la suma de residuales más pequeña de las dos regresiones resultantes a consecuencia de excluir ***m*** datos.

Para el periodo t = 1-80 se obtuvo la siguiente estimación:

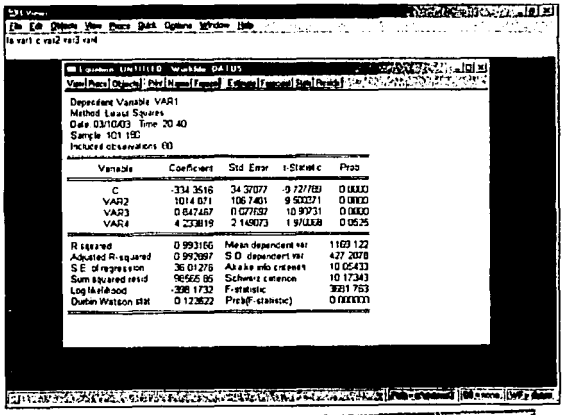


Dependent Variable: VAR1
Method: Least Squares
Date: 03/10/03 Time: 20:30
Sample: 1 80
Included observations: 80

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-152.8948	10.93076	-13.97784	0.0000
VAR2	568.4101	115.4821	5.095251	0.0000
VAR3	0.940613	0.110702	8.496066	0.0000
VAR4	3.755577	0.780466	4.811120	0.0000

R-squared	0.950656	Mean dependent var	159.7369
Adjusted R-squared	0.950207	S.D. dependent var	57.80031
S.E. of regression	5.712394	Akaike info criterion	6.371880
Sum squared resid	2479.990	Schwarz criterion	6.434642
Log likelihood	-250.8744	F-statistic	2165.675
Durbin-Watson stat	0.334073	Prob(F-statistic)	0.000000

Y para el periodo t = 101-180 se obtuvo la siguiente estimación:



Dependent Variable: VAR1
Method: Least Squares
Date: 03/10/03 Time: 20:40
Sample: 101 180
Included observations: 80

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-334.3516	34.37077	-9.727189	0.0000
VAR2	1014.071	106.7401	9.500271	0.0000
VAR3	0.847467	0.077592	10.99731	0.0000
VAR4	4.233819	2.149073	1.970068	0.0525

R-squared	0.993166	Mean dependent var	1169.122
Adjusted R-squared	0.992897	S.D. dependent var	427.2078
S.E. of regression	36.01276	Akaike info criterion	10.05433
Sum squared resid	98565.86	Schwarz criterion	10.17343
Log likelihood	-398.1732	F-statistic	3691.753
Durbin-Watson stat	0.123822	Prob(F-statistic)	0.000000

De esta manera se tiene

$RSS_1 = 98565.85$ para t = 1-80
$RSS_2 = 2479.99$ para t = 101-180
y

$$F_c = \frac{98565.85}{2479.99} = 39.74445461$$

mientras que

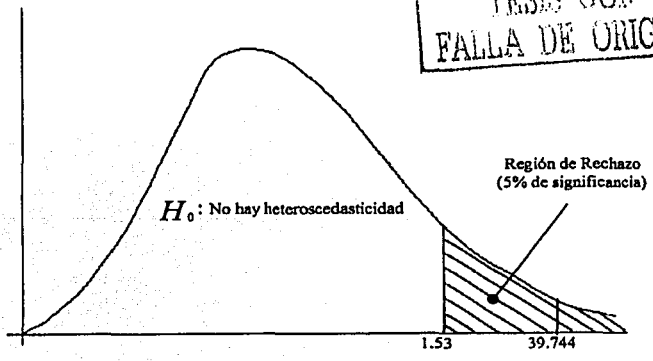
$$F_\alpha = \left(\frac{n-m}{2}, \frac{n-m}{2}\right) = F_{0.05}\left(\frac{180-20}{2}, \frac{180-20}{2}\right) = F_{0.05}(80,80) \approx 1.53$$

Por lo tanto

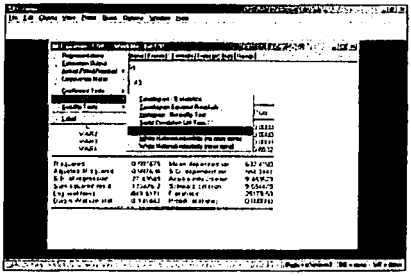
$$F_c > F_\alpha\left(\frac{n-m}{2}, \frac{n-m}{2}\right)$$

entonces se rechaza la hipótesis H_0: No existe heteroscedasticidad a un nivel de significancia del 5%, por lo que existen evidencias de que la varianza de los residuales del modelo lineal propuesto no permanece constante en el periodo de estudio.

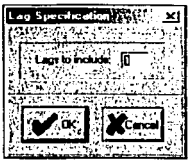
Gráficamente



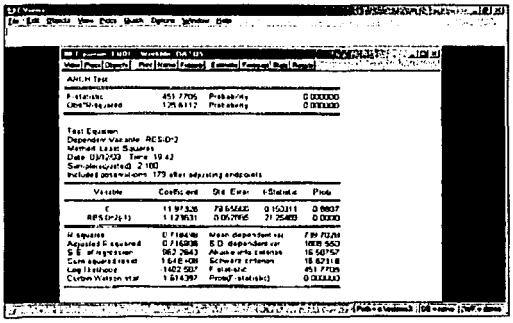
Ahora se mostrará el resultado de la prueba ARCH con un rezago, para esto debe seguirse la siguiente instrucción en el paquete:



Después de seleccionar el tipo de prueba deseada se debe especificar el número de rezagos con los cuales se hará dicha prueba, en este caso se teclea 1 (este número lo pone de forma predeterminada):

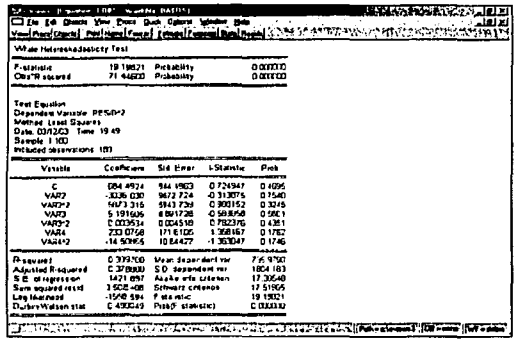


Finalmente al darle clic en el botón de OK el resultado para esta prueba será el siguiente:

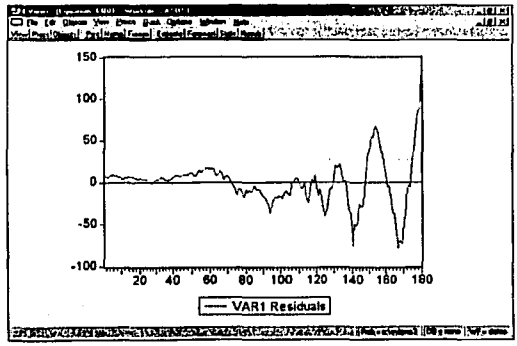


La probabilidad del estadístico ***F*** al ser menor que el 5% implica que existen evidencias de heteroscedasticidad en los residuales del modelo lineal propuesto.

Estos dos resultados anteriores son confirmados por la prueba ***White*** la cual al calcular un estadístico ***F*** con una probabilidad menor al 5% implica el hecho de que la varianza no permanece constante durante el periodo de estudio t = 1-180.



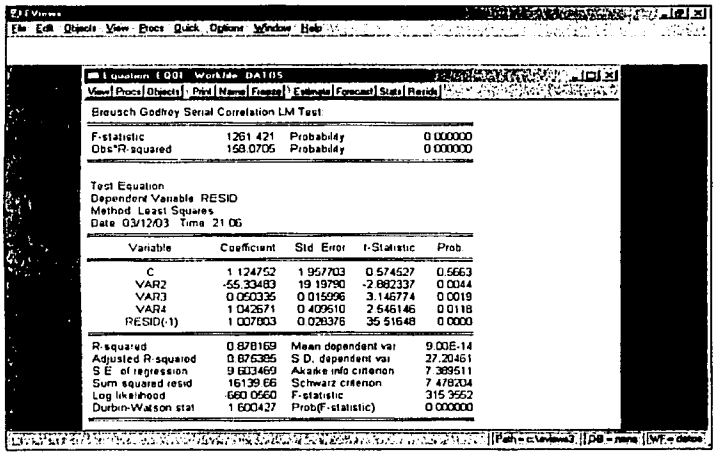
Estas afirmaciones se predecían al observarse el gráfico de los residuales, los cuales presentan un comportamiento creciente conforme transcurren las observaciones del periodo en estudio.



Correlación Serial

La prueba de *Durbin-Watson* para este modelo fue de ***d*** = 0.131843 la cual al ser muy cercana a cero implica la presencia de correlación serial positiva de primer orden, sin embargo a continuación se muestra el resultado de la prueba LM (Multiplicador de Lagrange) con un rezago:

En este contexto, la hipótesis nula que se pretende probar es $H_0: \rho = 0$ vs. la hipótesis alternativa $H_A: \rho > 0$.



Equation: EQ01 Workfile: DATOS

Breusch-Godfrey Serial Correlation LM Test

F-statistic	1261.421	Probability	0.000000
Obs*R-squared	158.0705	Probability	0.000000

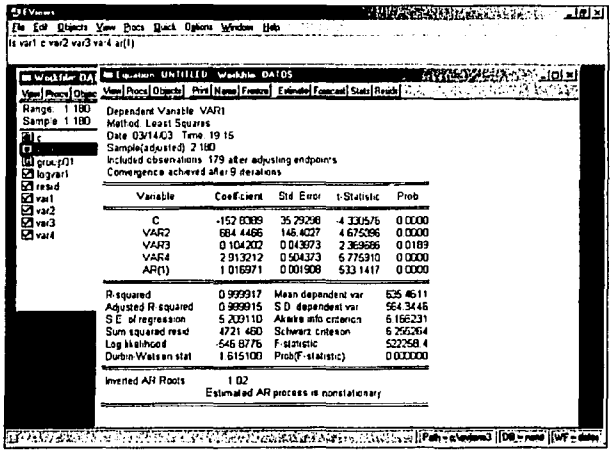
Test Equation
Dependent Variable: RESID
Method: Least Squares
Date: 03/12/03 Time: 21:06

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.124752	1.957703	0.574527	0.5663
VAR2	-55.33483	19.19790	-2.882337	0.0044
VAR3	0.050335	0.015996	3.146774	0.0019
VAR4	1.042671	0.409510	2.546146	0.0118
RESID(-1)	1.007803	0.028376	35.51648	0.0000

R-squared	0.878169	Mean dependent var	9.00E-14
Adjusted R-squared	0.876395	S.D. dependent var	27.20461
S.E. of regression	9.603469	Akaike info criterion	7.389511
Sum squared resid	16139.66	Schwarz criterion	7.478204
Log likelihood	-660.0560	F-statistic	315.3552
Durbin-Watson stat	1.600427	Prob(F-statistic)	0.000000

En esta prueba se observa que el estadístico ***F*** calculado tiene una probabilidad menor al 5%, además se observa la significancia de la variable RESID(-1) al tener una ***t*** calculada mayor a 2 y tener una probabilidad de ocurrencia menor al 5%, esto implica que la "información" no captada por el modelo en el periodo t-1 sí influye en el periodo t, lo que induce a confirmar la existencia de correlación serial de primer orden.

A continuación se presenta una regresión del modelo en donde se le anexó un AR(1), esto con la finalidad de tratar de eliminar el problema de correlación serial de primer orden.



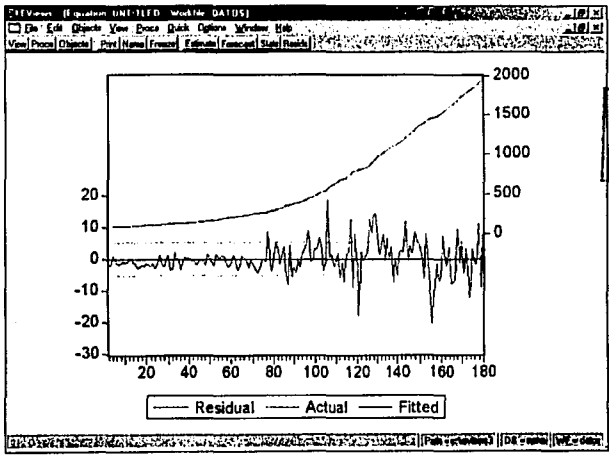
Dependent Variable: VAR1
Method: Least Squares
Date: 03/14/03 Time: 19:15
Sample(adjusted): 2 180
Included observations: 179 after adjusting endpoints
Convergence achieved after 9 iterations

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-152.8089	35.29298	-4.330576	0.0000
VAR2	684.4466	146.4027	4.675096	0.0000
VAR3	0.104202	0.043973	2.369686	0.0189
VAR4	2.913212	0.504373	6.775910	0.0000
AR(1)	1.016971	0.001908	533.1417	0.0000

R-squared	0.999917	Mean dependent var	635.4611
Adjusted R-squared	0.999915	S.D. dependent var	564.3446
S.E. of regression	5.209110	Akaike info criterion	6.166231
Sum squared resid	4721.460	Schwarz criterion	6.256264
Log likelihood	-546.8776	F-statistic	522258.4
Durbin-Watson stat	1.615100	Prob(F-statistic)	0.000000

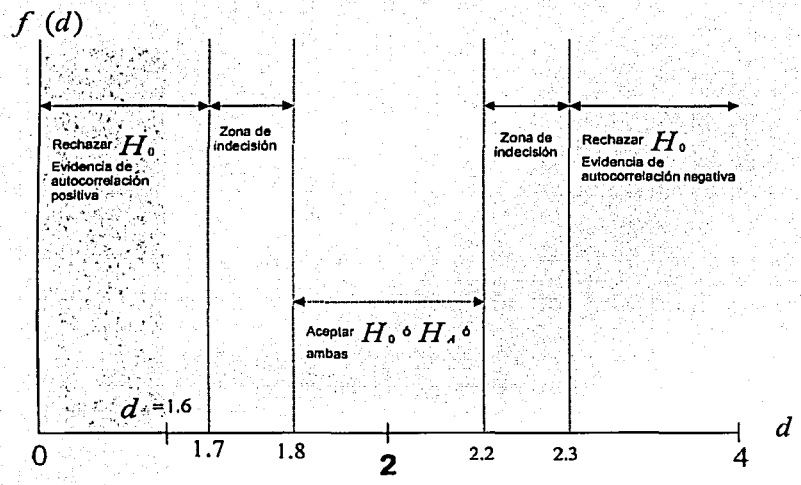
Inverted AR Roots 1.02
Estimated AR process is nonstationary

Ahora el modelo estimado y los residuales se visualizan de la siguiente manera:



Los puntos de significancia d_L y d_U a un nivel de significancia del 5% para n=200 y k= 4 son 1.728 y 1.810 respectivamente. Donde se deduce que $1.615100 = d < d_L = 1.728$.

Gráficamente



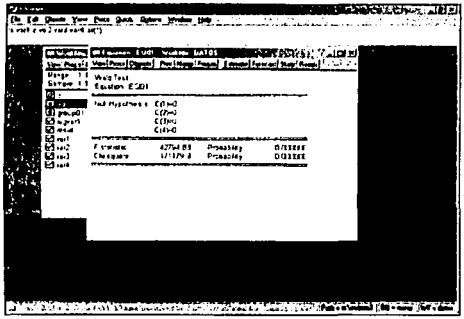
Lo cual implica que, aunque se corrige un poco el problema de la correlación serial de primer orden con un modelo autorregresivo de orden uno, aún existen evidencias de este problema entre las variables explicativas, esto sugiere la proposición de otro método regresivo (como el de mínimos cuadrados generalizados) o bien tomar la decisión de excluir variables y anexar otras.

Otros resultados de importancia

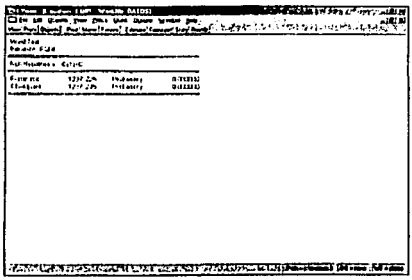
A continuación se presentan algunos resultados complementarios para establecer y afirmar algunos criterios que serán de suma importancia en la toma de decisiones que se harán el reformular el modelo lineal propuesto.

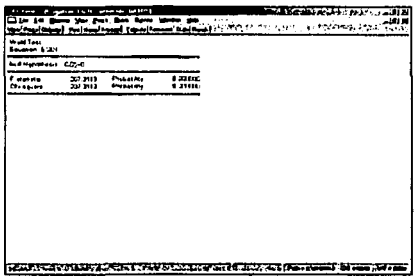
Pruebas sobre los coeficientes estimados

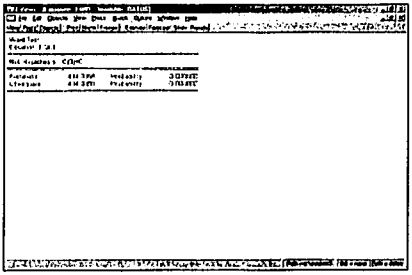
La prueba de *Wald* para establecer restricciones en los coeficientes estimados siendo la hipótesis nula $H_0: b_0 = b_1 = b_2 = b_3 = 0$ mostró el siguiente resultado:

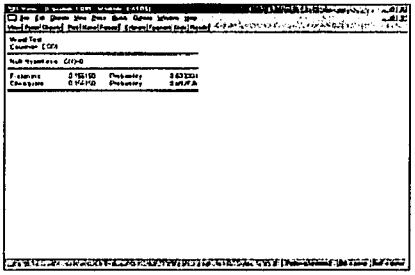


El estadístico ***F*** al presentar una probabilidad de ocurrencia menor al 5% implica que los coeficientes calculados son estadísticamente significativos en conjunto. Sin embargo los resultados individuales mostraron lo siguiente:

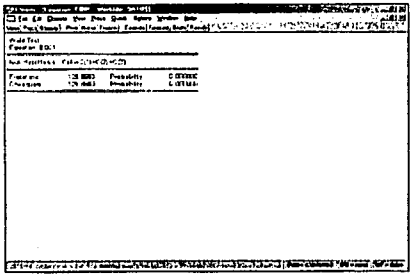


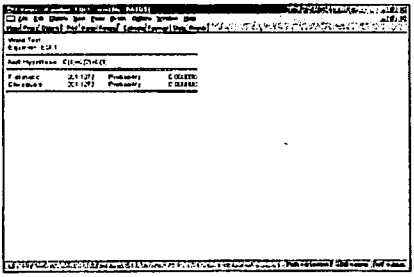




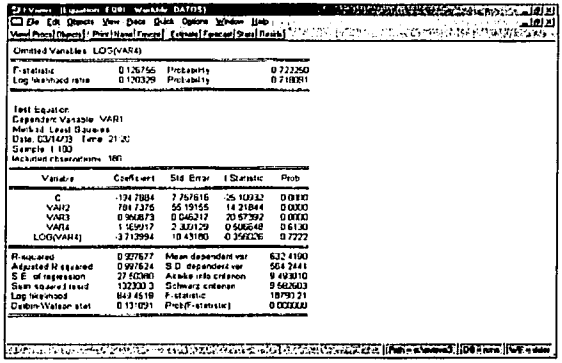


Nuevamente se confirma la significancia de los coeficientes calculados para las variables **VAR2** y **VAR3** mientras que el coeficiente de la variable **VAR4** no muestra representatividad en el modelo, sin embargo éste coeficiente podría ser una combinación lineal de los demás, como lo sugieren los siguientes resultados:



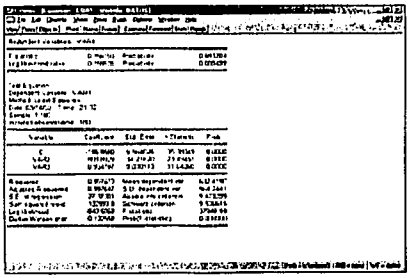


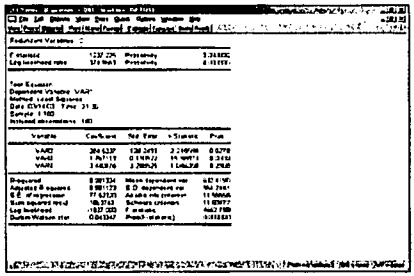
Para verificar si la transformación logarítmica de **VAR4** (=LOG(**VAR4**)), omitida en la regresión principal, es relevante para el modelo se obtuvo el siguiente resultado para la prueba de *omisión* de variables:



Al observarse una estadístico ***F*** con una probabilidad mayor al 5% implica que la variable LOG(**VAR4**) si se anexará al modelo lineal propuesto no sería estadísticamente significativa, es decir, no provocaría ningún impacto.

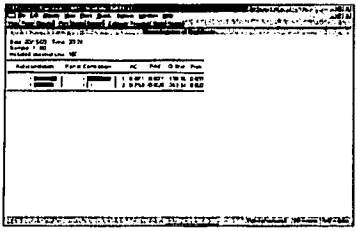
Finalmente, el siguiente resultado confirma contundentemente que dicha variable resulta ser *redundante* en el modelo, mientras que la ordenada al origen o la constante **C** sí resulta significativa para el modelo.



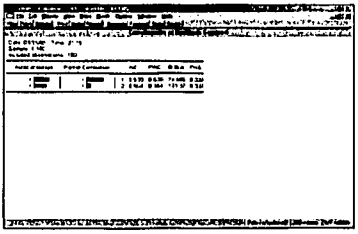


Pruebas sobre los residuales

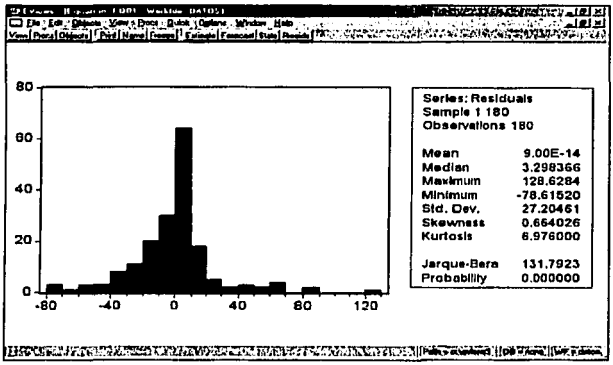
Una prueba que permite mostrar un orden superior de correlación serial es la sobre llamada *prueba de formateo* o más comúnmente conocida como prueba Q. . Esta es una prueba estadística para la hipótesis nula de que en los residuos no existe correlación de orden k. Para el modelo se hizo la prueba para $k = 2$ y el resultado fue el siguiente:



Aquí se observa que existen fuertes evidencias de correlación serial de grado 2, pues los valores calculados para los renglones 1 y 2 de **AC** (autocorrelación) sobrepasan por mucho el error estándar que los delimita, aunque cabe señalar que existe mayor correlación de grado uno. Mientras que estos valores para la correlación parcial (**PAC**) implican: primero, que la variable dependiente se relaciona mucho con las variables explicativas en el tiempo t y segundo, que en el tiempo **t-1** ésta relación disminuye considerablemente. Finalmente se observa que el estadístico Q en ambos casos tiene una probabilidad de ocurrencia menor al 5%, por lo que en ambos casos se rechaza la hipótesis nula a un 5% de significancia. Este mismo patrón puede apreciarse para el comportamiento de los residuales al cuadrado como a continuación se muestra:

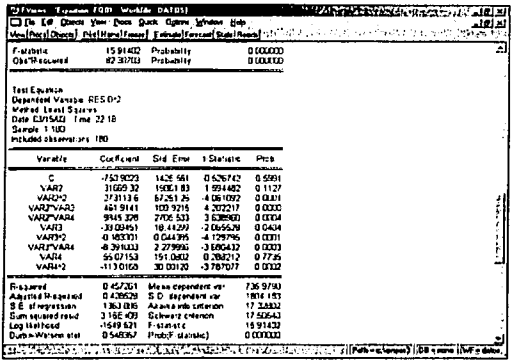


Ahora se mostrará si los residuos se distribuyen como una función de distribución de probabilidad normal, para esto se hizo uso de la prueba estadística *Jarque-Bera*, y el resultado fue el siguiente:



La estadística de *oblicuidad* ***S*** (Skewness) al ser muy cercana a cero indica que la distribución tiene una simetría considerable, sin embargo al ser positiva indica que la cola superior de la distribución es un poco más gruesa que la cola inferior. El estadístico ***K*** (Kurtosis) al ser mucho mayor que 3, de hecho es más del doble, implica que las colas de la distribución son muchísimo más gruesas que las de una distribución normal. Finalmente se observa que el valor calculado para el estadístico *Jarque-Bera* tiene una probabilidad de ocurrencia menor al 5%, por lo que se rechaza la hipótesis nula de que los residuales se distribuyen normalmente.

Para concluir con las pruebas de los residuales se presenta el resultado de la prueba *White* con términos cruzados, esto con la finalidad de confirmar la existencia de heteroscedasticidad en el modelo lineal propuesto. La estimación presentó lo siguiente:



F-statistic	15.91402	Probability	0.000000
Obs*R-squared	82.30703	Probability	0.000000

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 03/15/02 Time: 22:18
Sample: 1 100
Included observations: 100

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-752.9023	1426.661	-0.527742	0.5991
VAR2	31609.32	19081.83	1.656482	0.1127
VAR2^2	-3731.136	6725.125	-0.554892	0.5804
VAR2*VAR3	461.9141	100.9215	4.577217	0.0000
VAR2*VAR4	9945.209	2705.533	3.676420	0.0004
VAR3	-39.09451	18.48299	-2.115529	0.0374
VAR3^2	-0.183001	0.044395	-4.122795	0.0001
VAR3*VAR4	-49.39033	22.27999	-2.216432	0.0293
VAR4	55.07153	151.0802	0.364212	0.7165
VAR4^2	-113.0168	30.00120	-3.767077	0.0002

R-squared	0.457261	Mean dependent var	736.9790
Adjusted R-squared	0.428528	S.D. dependent var	1816.153
S.E. of regression	1363.035	Akaike info criterion	17.30832
Sum squared resid	3.16E+09	Schwarz criterion	17.50643
Log likelihood	-1649.621	F-statistic	15.91402
Durbin-Watson stat	0.548567	Prob(F-statistic)	0.000000

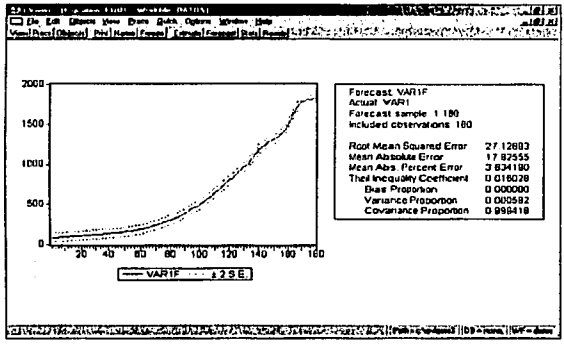
El hecho de tratar de explicar los residuales al cuadrado con base en una combinación lineal de las variables explicativas y sus "*cruces*" entre ellas es irrelevante, este hecho lo confirma el estadístico calculado ***F*** al mostrar una probabilidad de ocurrencia menor al 5%, por lo que la hipótesis nula $H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots\ldots\ldots = \sigma_m^2$ $(m \leq n)$ no se sostiene y se rechaza a un nivel de significancia del 5%.

Sin embargo puede apreciarse que al incluir un término cruzado en la ecuación original no disminuiría el problema de no homoscedasticidad, pues los valores de los estadísticos ***t*** calculados para los cruces de las variables independientes y sus cuadrados son estadísticamente significativos para explicar el comportamiento de los residuales al cuadrado, a pesar de que apenas y se obtuvo un $r^2 = 45.7\%$, es decir, la varianza no depende del comportamiento de las variables explicativas y sus posibles cruces conforme el número de observaciones crece.

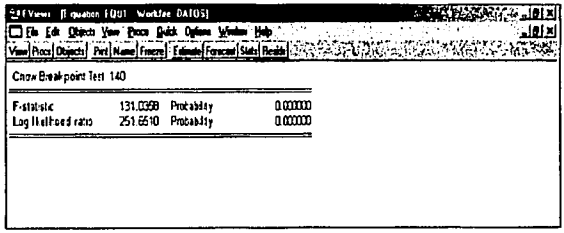
Pruebas sobre la estabilidad del modelo

La prueba estadística de *punto de rompimiento* de *CHOW* se basa en la idea de estimar separadamente dos o más rangos (conjunto de observaciones) pertenecientes al periodo de estudio para observar si existen diferencias significativas en las ecuaciones estimadas para cada uno de ellos.

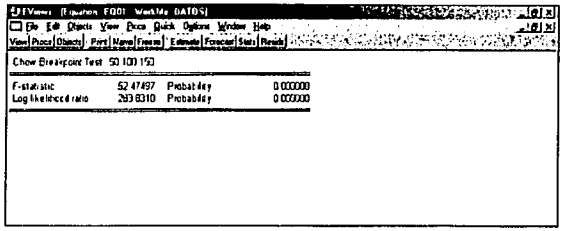
Antes de realizar la prueba se muestra el comportamiento de los valores estimados, esto con la finalidad de poder apreciar a simple vista algún cambio brusco en el modelo:



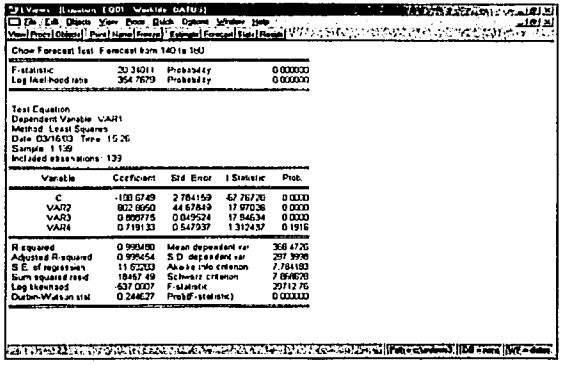
Aunque no puede apreciarse fácilmente un cambio brusco de los datos estimados se observa que por la observación 140 el modelo cambia ligeramente. Tomándose como un posible punto de rompimiento la observación t=140 se obtuvo el siguiente resultado:



Al obtenerse un estadístico ***F*** con una probabilidad de ocurrencia menor al 5%, conlleva a decir que existen evidencias de un posible punto de rompimiento en el modelo lineal propuesto a partir de la observación 140. Sin embargo al hacer una prueba para los puntos t=50, t=100 y t=150 se obtuvo:

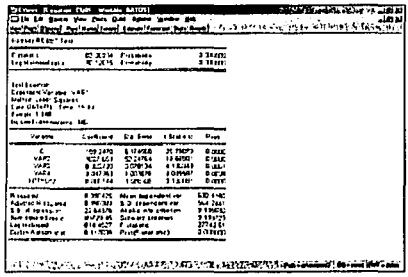


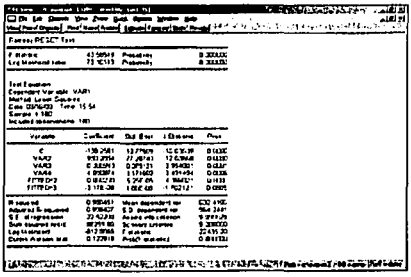
Esto implica que para 4 puntos existen diferencias en cuanto a la tendencia lineal del modelo. No es de extrañarse este resultado pues como puede observarse gráficamente los valores estimados reflejan un comportamiento casi exponencial. Ahora se mostrará la prueba *CHOW* pero para el periodo estimado a partir de t=140. El resultado fue el siguiente:



Este resultado confirma el rechazo de la hipótesis nula de la no existencia de un posible cambio estructural antes y después de la observación t=140.

Finalmente se presenta el resultado de la prueba estadística de *RAMSEY RESET* para 1 y 2 predicciones y de esta manera corroborar la posible existencia de un error en la especificación del modelo. La estimación de este estadístico para cada uno de los casos fue la siguiente:





En ambos casos se obtuvo un estadístico ***F*** que rechaza la hipótesis nula H_0: *No existen evidencias de un rompimiento estructural* en cuanto a la linealidad del modelo propuesto se refiere.

De esta manera se concluye la revisión tanto estadística como econométrica, a continuación se procederá a corregir los problemas que así lo requieran y a reformular el modelo lineal propuesto.

REFORMULACION

Con base en los resultados obtenidos para el modelo lineal propuesto las medidas remediables y correctivas son las siguientes:

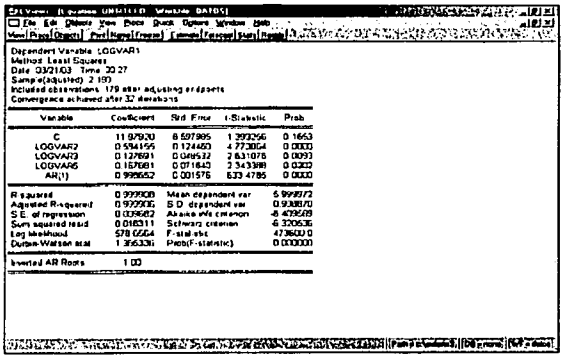
- Se excluyó la variable **VAR4** por no ser representativa en el modelo.
- Se cambió la escala de las variables y se usó la escala logarítmica en base **e**, esto con la finalidad de *suavizar* los datos y evitar el comportamiento exponencial.
- Se anexó la variable **VAR5** por lo que el modelo se transforma en:

$$\text{var 1} = e^{\log(e\,\text{var 2 var 3 var 5})}$$

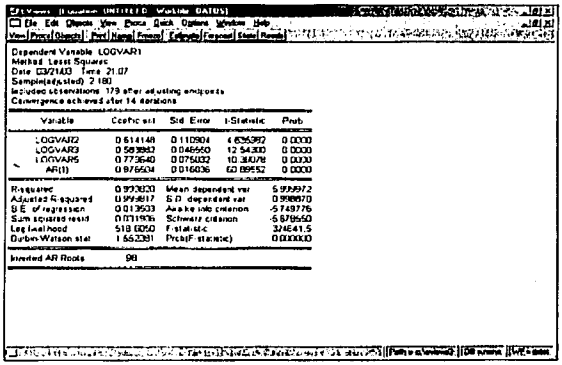
- Se llevó a cabo un modelo autorregresivo de primer orden, es decir, los términos de perturbación estocásticos siguen un patrón de comportamiento de la siguiente manera:

$$u_i = \rho u_{i-1} + v_i \text{ para toda } i, |\rho| < 1$$

donde v_i es un término de perturbación estocástico residual, que se supone satisface los supuestos del modelo de regresión lineal básico, incluyendo ausencia de correlación serial. De esta manera se obtuvo el siguiente resultado:

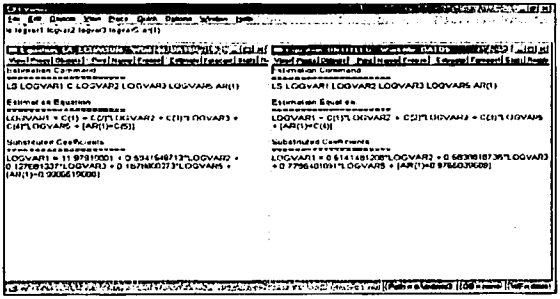


En este modelo lineal en parámetros puede apreciarse que los coeficientes para todas las variables son significativos individualmente, excepto para la constante ***C***. Si se hace la regresión con las mismas variables pero sin el término constante se obtiene lo siguiente:



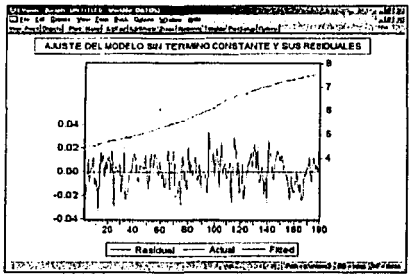
Qué se gana y qué se pierde en el modelo al omitir el término constante?

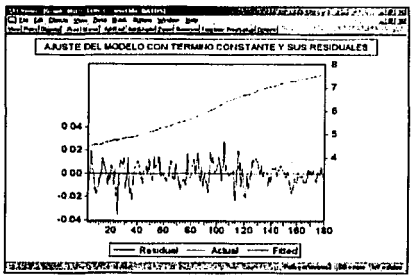
Partiendo de las dos estimaciones, sus representaciones son:



Representatividad no se pierde, pues para el modelo lineal con el término constante se obtuvo un $r^2 = .999908$ mientras que para el modelo lineal sin el término constante se obtuvo un $r^2 = .999820$.

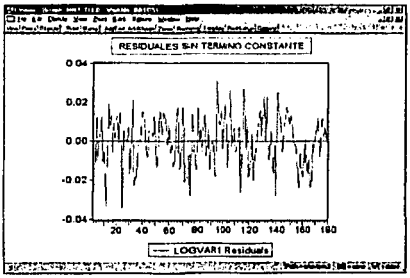
Esto puede confirmarse al ver los gráficos de ajuste de los modelos respectivos:

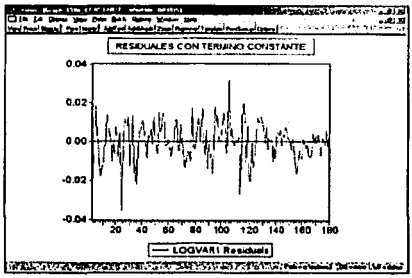




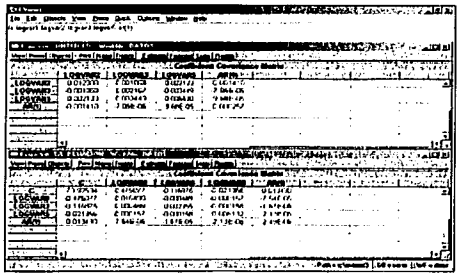
Sin embargo puede apreciarse que el comportamiento de los residuales con término constante son menos "*agresivos*" que los residuales obtenidos de la regresión sin término constante.

De hecho se observa que los residuales con término constante son más *suaves* y pareciera que convergen (o a estabilizarse) a cero desde el término 140, mientras que para los residuales sin término constante pareciera no existir algún patrón de comportamiento sistemático.

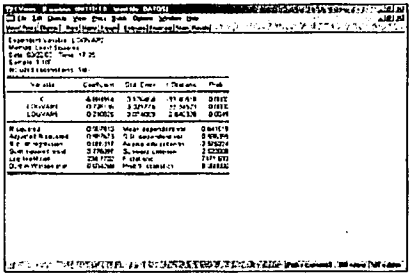


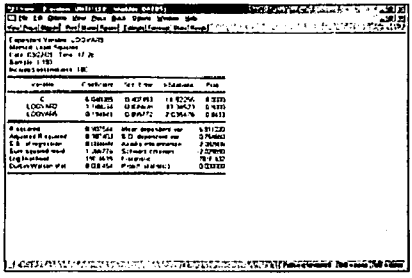


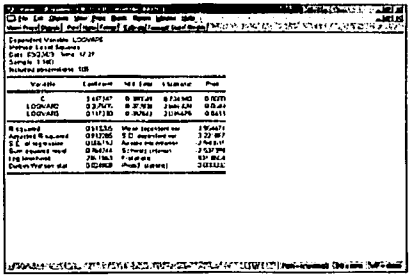
A continuación se muestra la matriz de covarianzas para ambos casos:



En la regresión con el término constante puede apreciarse una ligera disminución en la dependencia lineal (negativa) entre los coeficientes estimados en el modelo. Esto aparentemente implicaría la ausencia de multicolinealidad, sin embargo al llevar a cabo las regresiones entre ambas variables se obtuvieron las siguientes regresiones parciales:

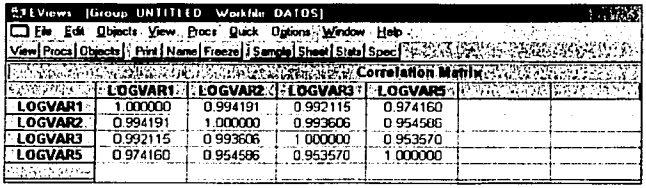






Estos resultados implican que aún con el cambio de escala se sigue preservando una alta multicolinealidad entre las variables explicativas, pues a pesar de que los coeficientes parciales de determinación en todos los casos son menores que el coeficiente de determinación de la regresión principal, éstos son demasiado elevados.

Este resultado puede confirmarse con la matriz de correlación entre las variables explicativas que a continuación se presenta:


EViews [Group UNTITLED Workfile DATOS]

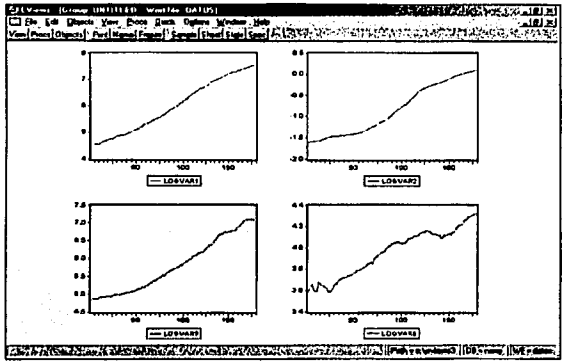
File Edit Objects View Procs Quick Options Window Help

View | Procs | Objects | Print | Name | Freeze | Sample | Sheet | Stats | Spec

Correlation Matrix

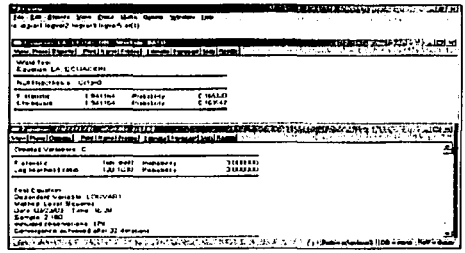
	LOGVAR1	LOGVAR2	LOGVAR3	LOGVAR5
LOGVAR1	1.000000	0.994191	0.992115	0.974160
LOGVAR2	0.994191	1.000000	0.993606	0.954586
LOGVAR3	0.992115	0.993606	1.000000	0.953570
LOGVAR5	0.974160	0.954586	0.953570	1.000000

También puede apreciarse gráficamente una gran relación lineal entre las variables que integran el modelo transformado:



Por otra parte se llevó a cabo para la regresión con término constante la prueba de *Wald* para probar la hipótesis nula $H_0 : c(1) = 0$ (es decir que el término constante es igual a cero) mientras que para la regresión sin término constante se llevó a cabo la prueba de variables omitidas para

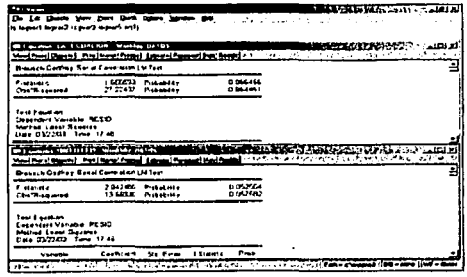
probar si la constante ***C*** es significativa en el modelo y los resultados fueron:



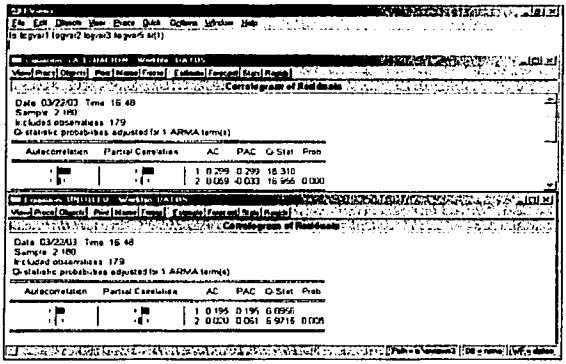
Con base en el primer resultado se deduce que la constante es igual a cero mientras que en el segundo resultado se concluye que dicha constante al ser omitida del modelo no causa cambio alguno.

Debido a que la prueba ***d*** de *Durbin-Watson* no es válida en el modelo de regresión sin el término constante, para el modelo con dicho término constante se obtuvo un estadístico ***d***=1.385336 el cual implica la existencia de correlación serial negativa de primer orden.

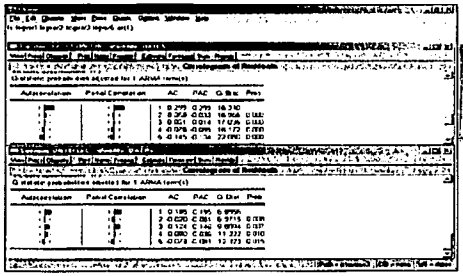
Sin embargo, se llevó a cabo la prueba de Multiplicadores de Lagrange (LM) obteniéndose hasta un grado superior de correlación, de hecho para el modelo con término constante hasta un orden de 16 y para el término sin constante hasta un orden de 6. Estos resultados ya no se obtienen cuando:



Ahora se presenta el resultado de la prueba ***Q*** para ambos casos, primero se observa el resultado para la regresión con término constante y después para la que no lo tiene:

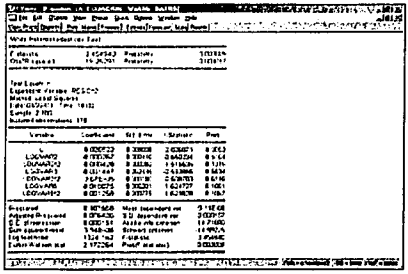


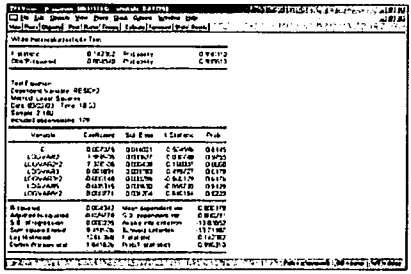
En ambos resultados se obtuvo que no existen indicios de correlación serial de segundo orden. De hecho ésta prueba se llevó a cabo hasta un orden de 5, obteniéndose los siguientes resultados:



Como puede apreciarse, con los resultados de dicha prueba no se rechaza la hipótesis nula de que no existe correlación de grado superior.

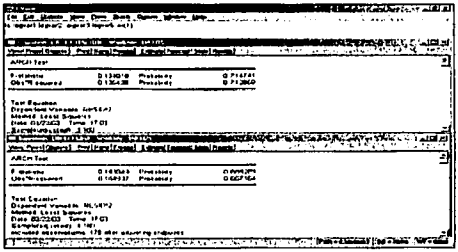
Otra prueba de comparación fue la de heteroscedasticidad, y en la cual se obtuvieron las salidas que a continuación se muestran:



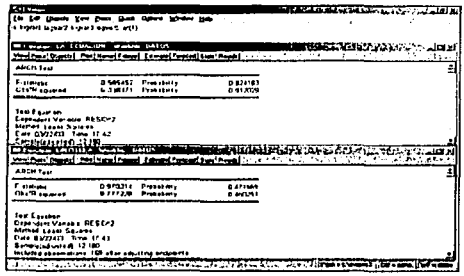


Aunque en ambas pruebas se aprecia que los estadísticos ***t*** no son muy elevados, se confirma que la regresión con el término constante presenta problemas de no homoscedasticidad mientras que la regresión sin el término constante no presenta problemas de heteroscedasticidad.

Asimismo se llevó a cabo la prueba ARCH de grado 1 para comprobar la existencia de heteroscedasticidad de grado uno en ambos modelos y se obtuvieron los siguientes resultados:

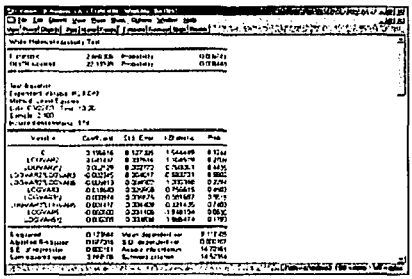


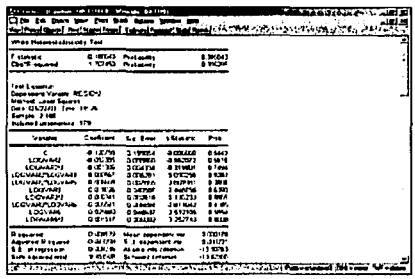
Esto implica que para ambos casos no existen evidencias de heteroscedasticidad de primer orden. Se llevó también a cabo esta prueba hasta el orden 10 y los resultados fueron:



Estos resultados implican que no existen evidencias de heteroscedasticidad de orden superior en los residuos de ambas regresiones, por lo que según esta prueba, la conducta de la varianza de los residuales no muestra .un comportamiento sistemático a lo largo de la muestra.

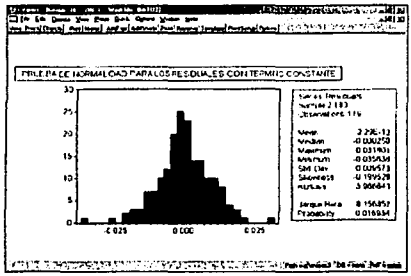
Sin embargo para la prueba de heteroscedasticidad usando términos cruzados estos resultados no se sostienen, en dicha prueba se obtuvo lo siguiente:

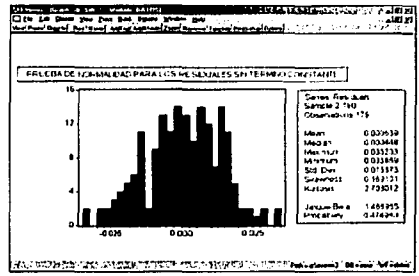




La primera salida que representa el resultado de la regresión con el término constante muestra problemas de heteroscedasticidad mientras que la segunda salida que representa el resultado de la regresión sin el término constante no muestra dicho problema.

Finalmente se presenta la prueba de *Jarque–Bera* para comprobar la normalidad de los residuales para cada regresión:





Es evidente que para la regresión con término constante los residuales no se distribuyen como una función normal, mientras que para la regresión sin término constante sus residuales se distribuyen como una función de distribución de probabilidad normal con media 0 y varianza igual a σ^2.

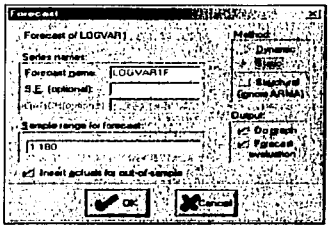
Tal vez éste sea el resultado que diferencie a los dos modelos, pues a pesar que existen diferencias (y coincidencias) sustantivas entre ellos, el supuesto de normalidad en los residuos garantiza que los estimadores son *eficientes* (insesgados y con varianza mínima) y *consistentes* (a medida que el tamaño de la muestra aumenta indefinidamente, los estimadores convergen hacia sus valores poblacionales verdaderos).

Por lo que se ha llegado a mostrar que el modelo lineal sin término constante es el que mejor explica el comportamiento de la variable **LOGVAR1** para $t \in [1,2,.....,180]$, es decir:

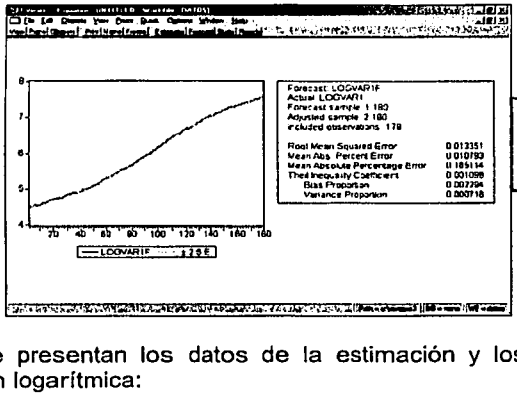
```
Estimation Equation:
=====================
LOGVAR1 = C(1)*LOGVAR2 + C(2)*LOGVAR3 + C(3)*LOGVAR5 + [AR(1)=C(4)]

Substituted Coefficients:
=====================
LOGVAR1 = 0.5141481208*LOGVAR2 + 0.5838818736*LOGVAR3 + 0.7796401691*LOGVAR5 + [AR(1)=0.9765039689]
```

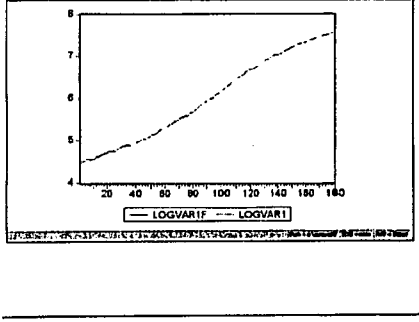
Para obtener los valores estimados puntualmente se le da clic en el botón FORECAST y entonces se mostrará una ventana como la siguiente:

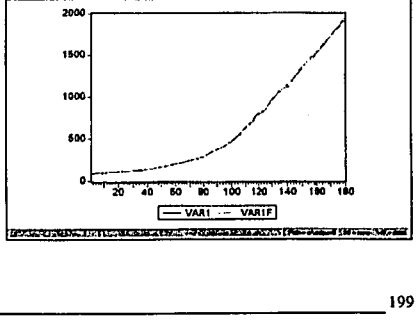


Con esto se genera la serie **LOGVAR1F**, que no es más que los valores estimados para cada observación por el modelo lineal propuesto en el rango t=1,....,180. Si se seleccionan las opciones de **Do graph** (hacer gráfico) y **Forecast evaluation** (presentar la evaluación de la estimación) entonces se obtendrá:



Finalmente se presentan los datos de la estimación y los datos sin la transformación logarítmica:





A pesar de que el modelo sin el término constante describe estructuralmente mejor el comportamiento de la variable **LOGVAR1**, éste presenta problemas de correlación serial de grado superior y multicolinealidad. Se ha mostrado que dicho modelo tiene problemas de correlación serial de hasta grado 6 (según la prueba de LM), pero qué tan severa es la multicolinealidad?

Para medir el efecto de multicolinealidad se calculó la contribución incremental de cada variable independiente en la explicación de la variable dependiente **Y** (=**LOGVAR1**), una vez que se fueron incluyendo las otras variables. Con base en las expresiones (3.7.1) y (3.7.2) se obtuvieron los siguientes resultados:

Regresión	r^2	F_l	ψ_l
logvar2	0.988416	15188.44	0.96587649
logvar2 logvar3	0.989854	8634.112	0.46638411
logvar2 logvar3 logvar5	0.996172	15268.53	0.31322674

El efecto se calcula entonces como $\overline{M} = \left[\sum_{l=2}^{n} \psi_l - R^2\right]$ donde el R^2 es el coeficiente de determinación de la regresión del modelo lineal sin término constante, o sea $R^2 = 0.999820$. Por lo que:

$$\overline{M} = 1.74548735 - 0.999820 = 0.74566735 \approx 0.75$$

Este resultado implica que el grado de colinealidad entre las variables explicativas aunque alto, no puede considerarse de severidad.

REFINAMIENTO DE LA REFORMULACION

Se ha llegado a obtener un modelo lineal en parámetros que satisface en gran medida el comportamiento de la variable **VAR1**, sin embargo aún presenta insensibilidad estructural, y muy en particular resalta el problema de la correlación serial de grado superior.

La búsqueda de la mejor representación conlleva en la mayoría de los casos a modificar resultados que con modelos anteriores se habían obtenido, sin embargo, esa búsqueda del mejor modelo depende mucho de las necesidades y objetivos específicos que se tengan.

La finalidad de este trabajo es mostrar las herramientas que el paquete estadístico Eviews ofrece para el análisis y evaluación de un modelo propuesto en particular, la profundidad tanto econométrica como estadística dependerá obviamente de la magnitud del problema en estudio.

A manera de concluir con el análisis del modelo sin término constante a continuación se presenta un refinamiento del mismo modelo que satisface mejor los supuestos del análisis de regresión simple. Teniendo en cuenta la matriz de correlación siguiente:

EViews [Group UNTITLED Workfile DATOS]

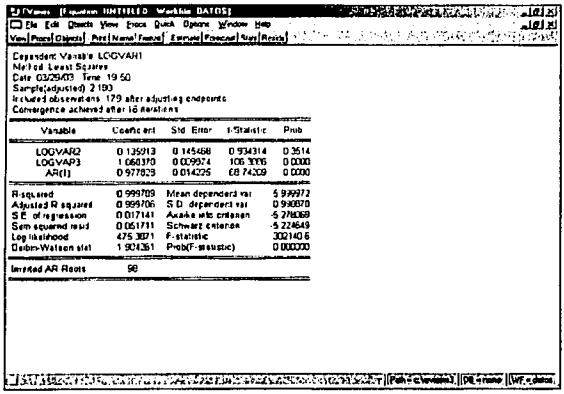
File Edit Objects View Proc Quick Options Window Help

View Proc Object Print Name Freeze Sample Sheet Stats Spec

Correlation Matrix

	LOGVAR1	LOGVAR2	LOGVAR3	LOGVAR5
LOGVAR1	1.000000	0.994191	0.992115	0.974160
LOGVAR2	0.994191	1.000000	0.993606	0.954586
LOGVAR3	0.992115	0.993606	1.000000	0.953570
LOGVAR5	0.974160	0.954586	0.953570	1.000000

Se observa que las variables **LOGVAR2** y **LOGVAR3** presentan una alta interrelación, mientras que la variable **LOGVAR5** es la que presenta una relación menor, aunque muy ligera, con las demás variables en estudio, por lo que omitiendo ésta variable (a pesar de que esta variable resultó ser significativa) del modelo se obtiene la siguiente estimación:



Dependent Variable LOGVAR1
Method Least Squares
Date 03/29/03 Time 19 50
Sample(adjusted) 2 100
Included observations 179 after adjusting endpoints
Convergence achieved after 16 iterations

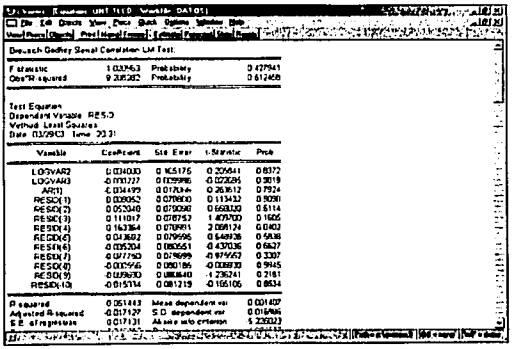
Variable	Coefficient	Std Error	t-Statistic	Prob
LOGVAR2	0 135913	0 145468	0 934314	0 3514
LOGVAR3	1 060370	0 009974	106 3096	0 0000
AR(1)	0 977829	0 014225	68 74209	0 0000

R-squared	0 999709	Mean dependent var	5 999972
Adjusted R-squared	0 999706	S D dependent var	0 930870
S E of regression	0 017141	Akaike info criterion	-5 278069
Sum squared resid	0 051711	Schwarz criterion	-5 224649
Log likelihood	475 3871	F-statistic	302140 6
Durbin-Watson stat	1 904361	Prob(F-statistic)	0 000000

Inverted AR Roots 98

Lo primero que se observa en cuanto a las variables explicativas es que la variable **LOGVAR2** no es significativa en el modelo, sin embargo su participación permite mejores resultados dentro del modelo que con la variable **LOGVAR5**, la variable **LOGVAR3** sí es significativa y el modelo autorregresivo de primer orden **AR(1)** resulta dar buenos resultados para explicar el comportamiento de los residuales.

Para detectar el problema de correlación serial de orden superior entre las variables **LOGVAR2** y **LOGVAR3** se realizó la prueba LM de orden 10, el resultado fue:



Breusch-Godfrey Serial Correlation LM Test:

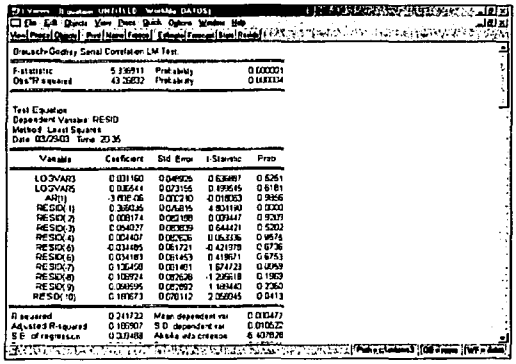
F-statistic	1 032463	Probability	0 427941
Obs*R-squared	9 308382	Probability	0 612468

Test Equation
Dependent Variable RESID
Method Least Squares
Date 03/29/03 Time 20 31

Variable	Coefficient	Std Error	t-Statistic	Prob
LOGVAR2	0 034000	0 165176	0 205841	0 8372
LOGVAR3	-0 001727	0 025995	-0 066438	0 9618
AR(1)	-0 034499	0 017044	-0 261612	0 7924
RESID(-1)	0 009952	0 078600	0 113432	0 9098
RESID(-2)	0 053048	0 079260	0 669329	0 5114
RESID(-3)	0 111017	0 078752	1 409700	0 1605
RESID(-4)	0 163764	0 078951	2 058124	0 0403
RESID(-5)	0 051602	0 079696	0 646936	0 5438
RESID(-6)	-0 035204	0 080651	-0 436036	0 6627
RESID(-7)	-0 077293	0 079699	-0 979557	0 3307
RESID(-8)	-0 007946	0 080186	-0 099093	0 9145
RESID(-9)	-0 009870	0 080840	-0 122041	0 2181
RESID(-10)	-0 015114	0 081219	-0 186106	0 8634

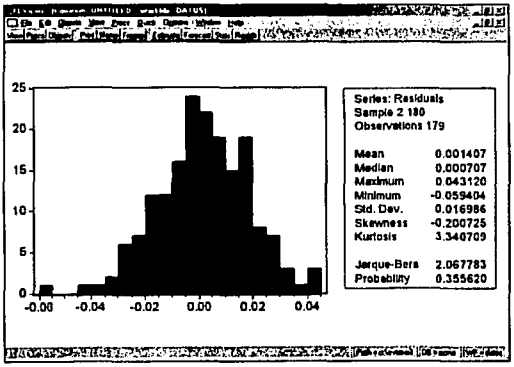
R-squared	0 051483	Mean dependent var	0 001407
Adjusted R-squared	-0 017127	S D dependent var	0 016896
S E of regression	0 017131	Akaike info criterion	-5 230213

Mientras que para las variables **LOGVAR3** y **LOGVAR5** el resultado fue:

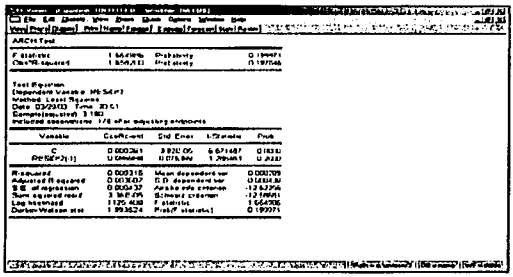


Con esto se muestra que la variable que presenta el problema de correlación serial en el modelo es la variable **LOGVAR5**, por lo que al ser omitida se soluciona dicho problema.

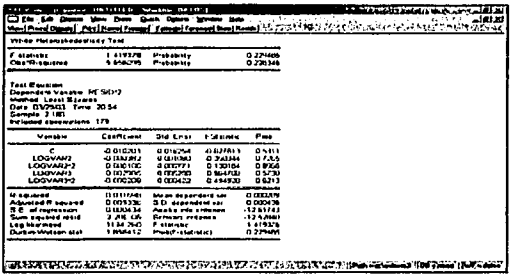
Por otra parte, este modelo tiene unos residuales que tienen una función de distribución normal con media 0 y varianza igual a σ^2, garantizando así la consistencia de los estimadores, esta afirmación puede hacerse con base en el resultado de la prueba *Jarque-Bera*:



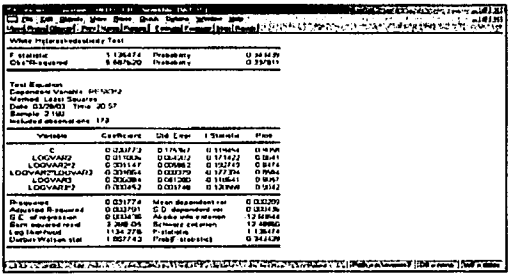
La prueba *ARCH(1)* muestra que no existen indicios de heteroscedastiscidad.



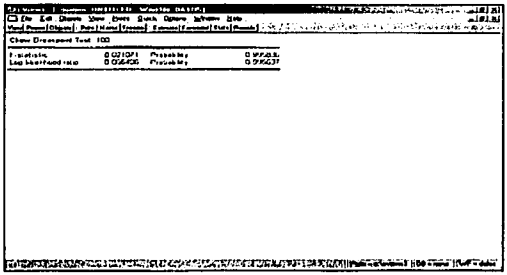
La prueba *WHITE* sin términos cruzados confirma que los residuales no presentan un patrón sistemático de comportamiento conforme la muestra crece.



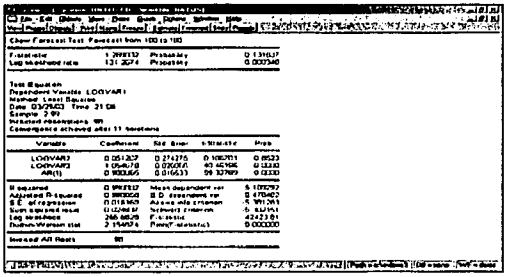
La prueba *WHITE* con términos cruzados sostiene los resultados de no heteroscedasticidad.



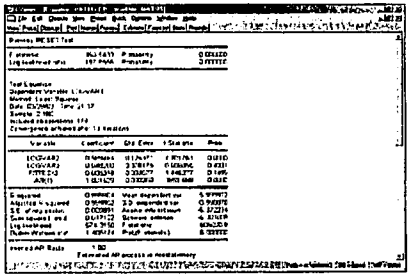
La prueba *CHOW BREAKPOINT* conlleva a decir que al menos en la observación 100 no existe un punto de rompimiento o cambio brusco en los datos.

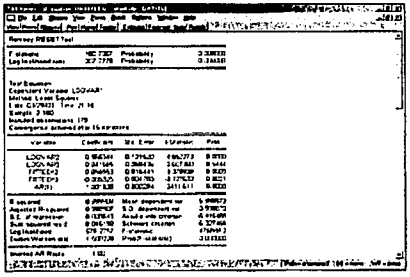


De la prueba *CHOW FORECAST* se deduce que para las observaciones posteriores a 100 no existen evidencias de un cambio estructural para el modelo.

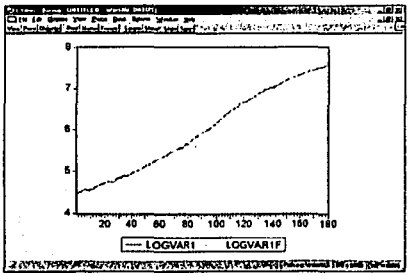


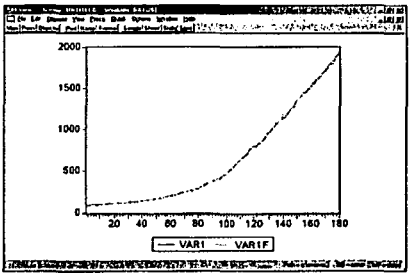
Sin embargo la prueba *RAMSEY RESET* para 1 y 2 predicciones implica que la tendencia estructural del modelo no se sostiene:





Finalmente se muestra el ajuste del modelo tanto para los datos con la transformación logarítmica como para los datos originales:





Se ha encontrado un modelo que cumple con los supuestos del análisis de regresión más satisfactoriamente que los modelos anteriores, más sin embargo si lo que se desea es un modelo predictivo, definitivamente ésta no será su función principal, debido a que la estructura del modelo hacia futuras observaciones no se sostiene. Más sin embargo, si lo que se desea es un modelo para realizar pronósticos a corto plazo, entonces el modelo que involucra a la variable **LOGVAR5** será el adecuado, aún cuando éste presenta problemas internos más severos, conserva la estructura hacia futuras observaciones.

CONCLUSIONES

Actualmente la especificación y evaluación de un modelo econométrico involucran a un amplio conjunto de técnicas, a las cuales dicho modelo puede y debe someterse en muy diferentes etapas durante el proceso de construcción y subsiguiente empleo.

"Existen una gran cantidad de contrastes o índices que intentan medir la bondad de los distintos modelos, sin embargo la elección del mejor indicador se debe hacer dependiendo de los objetivos que persigue el investigador ya que estos indicadores de bondad son planteados para la búsqueda de soluciones de problemas específicos" (Pulido, 1993 p.187).

En términos más amplios, es posible contrastar la eficacia de los modelos de regresión con relación a otros modelos más complejos e incluso con otros enfoques metodológicos alternativos, es decir, comparando costes, tiempo requerido, etc.

Sin embargo, de qué sirve el diseño de un buen modelo econométrico? De que sirven los mínimos cuadrados ordinarios? Qué sentido tendría el análisis de regresión si no se cuenta con una base de datos? En base a qué se harían pronósticos? Estas preguntas y algunas otras no tendrían ningún sentido si no se cuenta ante todo con una base de datos, y en gran medida todo el análisis que se haga dependerá de la calidad de los datos que se recaben y/o generen. Evidentemente una base de datos bien estructurada conllevará a un estudio econométrico que refleje mucho mejor la realidad del fenómeno en estudio.

Desgraciadamente el integrar bases de datos confiables resulta complicado, pues muchas de las veces es muy costoso llevar a cabo un estudio de muestreo serio. Por lo que al carecer de los medios para construir bases de datos confiables el único recurso es utilizar las bases de datos que ya han sido generadas por instancias incluso ajenas al estudio del fenómeno que se está analizando.

Este trabajo ha mostrado varias técnicas para corregir un modelo econométrico muy particular, sin embargo, éstas son utilizadas una vez que ya ha sido estimado el modelo, por lo que el análisis de los datos hace necesario un estudio aparte, no sin antes mencionar que un modelo econométrico jamás corregirá los errores que de ante mano se adquirieron por el uso de unos datos malos.

Una vez que se ha mencionado cuan importantes son los datos en un estudio econométrico, en este trabajo se ha visto que al realizar la estimación de un modelo, los resultados que deben de apreciarse primero para darse una idea de cómo está el modelo estimado, en cuanto a eficacia se refiere, son: el contraste de signos, la bondad de ajuste o R^2, el estadístico ***t*** de student y el estadístico ***d*** de Durbin-Watson.

a) Signo esperado del parámetro de regresión. Un contraste básico para todo parámetro es que su signo coincida con el que se había planteado originalmente en el modelo, basado en el conocimiento teórico de las relaciones económicas. Es decir, si el modelo está planteado en términos lineales directos, los parámetros del modelo que se estiman son aproximaciones del efecto que tiene una variación de la variable exógena sobre la variable explicada por el modelo. Aunque para el modelo analizado en este trabajo no se hizo ningún supuesto sobre los signos esperados, es relevante mencionar este criterio para considerarlo en un análisis puramente económico.

b) Coeficiente de determinación o R^2. El coeficiente de determinación se puede interpretar como la proporción de variación de la variable endógena que queda explicada por la regresión, es decir la explicación que dan las variables exógenas que se han incluido en el modelo.

c) Contraste de significación de un parámetro individual (estadísticos ***t***). Un problema relacionado con la selección de modelos, en el que los métodos tradicionales presentan inconvenientes, es el de decidir si se retiene o no una nueva variable añadida como regresor.

d) Estadístico ***d*** de Durbin-Watson. Las causas de la autocorrelación son varias y algunas de estas son: variable excluida, funciones incorrectas (potencia de las variables explicativas), retardo, inercia, etc.

Con base en estos criterios y en las pruebas econométricas que se hicieron al modelo de este trabajo las conclusiones son las siguientes:

I. Disponer de una herramienta informática es clave para la construcción y evaluación de modelos econométricos hoy en día, debido principalmente a la rapidez en la realización de los cálculos. De otra

manera resultaría impensable su ejecución manualmente bien sea por procedimientos algebraicos de resolución de matrices y determinantes o por cualquier otro método.

En este trabajo se utilizó el paquete estadístico Econometric Views en su versión 3.1 para el ambiente Windows, esta herramienta sirvió para llevar a cabo tanto el diseño como el análisis de un modelo econométrico en específico: el modelo uniecuacional múltiple (y lineal en parámetros).

II. A pesar de que no se puede emitir un juicio inmediato sobre la validez de un modelo a partir del R^2, generalmente se establece que un $R^2 = 0.8$ (equivalente a un 80%) es suficiente para considerar que un modelo es "bueno". En consecuencia, un valor de R^2 elevado no es en sí mismo una garantía de que el modelo sea correcto.

En situaciones especiales el uso del R^2 no solo es desaconsejado sino incorrecto o inadecuado. Así, es incorrecto comparar los coeficientes de determinación de modelos que poseen diferentes variables dependientes o que estén relacionadas por una transformación no lineal. Esto último es lo que ocurre, por ejemplo, cuando en uno de los modelos aparece la variable dependiente con rezagos y en el otro en diferencias o en logaritmos.

III. El estadístico *t* se basa en la metodología llamada de prueba de hipótesis, la hipótesis nula es Ho: $b_i = 0$, es decir que el efecto de la variable X_i sobre Y_i no existe.

Un parámetro puede considerarse estadísticamente significativo, a niveles de confianza de aproximadamente un 95%, si el valor del estimador supera dos veces su desviación estándar.

IV. La disposición de información confiable hará que un modelo econométrico, cualquiera que este sea, refleje con mayor realidad y veracidad al fenómeno que sea de interés y de estudio.

V. La elección del modelo econométrico que más convenga dependerá en gran medida de la necesidad del estudio y los compromisos

adquiridos antes de su construcción y posterior especificación. Si un modelo se construye para predecir, un criterio natural de selección es el de la minimización de los errores de predicción fuera de la muestra. La racionalidad de este criterio descansa en el hecho de que no tiene mucho mérito que un modelo prediga bien dentro del periodo muestral.

VI. Sólo se puede hacer un buen modelo de aquello que se conoce con suficiente profundidad.

Finalmente, es importante mencionar que la econometría sigue experimentado un gran desarrollo día con día, tanto en su metodología como en sus aplicaciones. Algunos factores a mencionar son los siguientes:

- El continuo aumento de la disponibilidad de datos estadísticos, y la creación de bases de datos confiables, gracias a la labor de diversos organismos nacionales e internacionales.

- La generalización del uso de ordenadores y los avances experimentados en el desarrollo de programas informáticos destinados a los problemas econométricos, que facilitan el uso de técnicas complejas.

- Los efectos de las innovaciones y la internacionalización sobre las sociedades actuales, ya que aumentan la competitividad e impulsan a muchas empresas e instituciones a basar sus decisiones en estudios cuantitativos bien fundamentados, lo que repercute sobre una mayor demanda de estudios econométricos.

BIBLIOGRAFIA

- Banguero Lozano, Harold "El Modelo Simple" (Econometría) Conferencia No. 3, El Colmex
- Banguero Lozano, Harold "Mínimos Cuadrados Generalizados" (Econometría) Conferencia, El Colmex
- Banguero Lozano, Harold "Modelo de Regresión Lineal Múltiple" (Econometría) Conferencia, El Colmex
- Banguero Lozano, Harold "Problemas en el Modelo Lineal" (Econometría) Conferencia No. 5, El Colmex
- Banguero Lozano, Harold "Teoría y Modelos" (Econometría) Conferencia, El Colmex
- Biblioteca de Consulta Microsoft® Encarta® 2003
- Carrascal Arranz, Ursicino, González González Yolanda, Rodríguez Prado Beatriz "Análisis Econométrico con Eviews" Editorial RA-MA, Madrid 2001
- Cassoni E., Adriana "Pruebas de Diagnóstico en el Modelo Econométrico" Documentos de Trabajo I y II, Primera Edición Centro de Investigación y Docencia Económicas A. C. (CIDE), 1991
- Christ, Carl F. "Modelos y Métodos Econométricos" Editorial Limusa, México, 1979
- Cochran, William G. "Técnicas de Muestreo" Compañía Editorial Continental S. A. (CECSA) Segunda Edición en Español de la Tercera Edición en Inglés, 1980
- Darnell, Adrian C. & J. Lynne Evans "The Limits of Econometrics" Aldershot, England: Edward Elgar, Gower, 1990
- Fernández, Viviana "Ayudantía de Eviews" (Documento de trabajo) Centro de Economía Aplicada (CEA), Universidad de Chile, 2003
- Gandolfo, Giancarlo "Métodos y Modelos Matemáticos de Dinámica Económica" Editorial TECNOS, Madrid, 1976
- Geary, R. C. "Some Results About Relations Between Stochastic Variables: A Discussion Document" Review of International Statistical Institute, vol. 31, 1963, pp. 163-181
- Geary, R. C. "Relative Efficiency of Count of Sign Changes for Assessing Residual Autoregression in Least Squares Regression" Biometrika vol. 57, 1970, pp. 123-127
- Guerrero M., Víctor "Una introducción a la metodología" El Colmex
- Gujarati, Damodar N. "Econometría" McGrawHill Segunda Edición, 1994
- Hernández Alonso, José "Ejercicios de Econometría" Editorial ESIC Segunda Edición, Colección Universitaria, 1992

- Hu, Teh-Wei "Econometría: un Análisis Introductorio" Fondo de Cultura Económica, 1979, pp. 7-11
- Intriligator, Michael D. "Modelos Econométricos, Técnicas y Aplicaciones" Editorial Fondo de Cultura Económica, México, 1990
- Jhonston, John "Métodos de Econometría" Editorial Vicens-Vives Barcelona, 1995
- Johnston, Jhon "Econometric Methods" McGrawHill Book Company, 3ª Edition, New York, 1984
- Kendrick, David A. "Stochastic control for economic models" McGrawHill, New York, 1981
- Kmenta, Jan "Elements of Econometrics" Mac Millan, Second Edition 1986
- Maddala, G. S. "Introduction to Econometrics", Mac Millan Publishing Company, Second Edition, New York, 1992
- Manual de Econometric Views, Guía del Usuario
- Notas del Diplomado intitulado "Econometría" de la Facultad de Economía, UNAM, 1995
- Notas en clase de las materias Estadística I y Estadística II impartidas por el Prof. Francisco Sánchez Villarreal en la Facultad de Ciencias, UNAM
- Pindyck, R. S. y D. L. Rubinfeld "Econometric Models and Economic Forecasts", McGrawHill, 1983
- Pulido San Román, Antonio "Modelos Econométricos" Ediciones Pirámide S.A., Madrid, 1993
- Ragnar, Frisch "Statistical Confluence Analysis by Means of Complete Regression Systems" Institute of Economica, Oslo University, Publ. No. 5, 1934
- Raj, Des "Teoría del Muestreo" Fondo de Cultura Económica, México, 1992
- Romero Cortés, José Carlos "Análisis de Regresión y Correlación Lineal Simple" Ensayos. Primera Edición de la Universidad Autónoma Metropolitana, Unidad Azcapotzalco, 1989
- Salvatore, Dominick "Econometría", Serie de Compendios Schaum, McGrawHill, 1996
- Scheaffer, Richard L., Mendenhall William & Ott Lyman "Elementos de Muestreo" Grupo Editorial Iberoamérica S. A de C. V. 1987, pp. 14-15
- Shumpeter, J. A. "Historia del análisis económico" Editorial Ariel, 1971
- Spanos, Aris "Statistical Foundations of Econometric Modelling" Cambridge University Press, Cambridge, 1986
- Wonnacott, Ronald J. & Wonnacott Thomas H. "Econometría" Editorial Aguilar, Edición Española, Madrid, 1982