INTRODUCTION TO ECONOMETRICS

Abstract

Econometrics is one of the most important applications to the mathematical statistics and a fundamental tool in the economic research and in the design and analysis of economic policy. The present document develops the basic theory concepts of the econometric modeling for those that begin the study of economics. The specification, assumptions, estimation, hypothesis testing and predictions for the classical regression model are the principal topics presented in this text. The concepts, the tools and their applications developed in this document are relevant for tackling many practical problems in today´s world and for the introduction in advanced econometric courses.

Key words: correlation analysis, least squares estimation, econometric model, hypothesis tests.

JEL classification: C01, C10 y C20

TABLA DE CONTENIDO

INTRODUCCIÓN

El curso de Econometría hace parte del área de métodos cuantitativos en economía y se constituye en una herramienta importante en la investigación económica, el diseño y análisis de política. El contenido y el desarrollo del curso son a nivel introductorio y su interés es la aplicación de los conceptos teóricos. El curso busca proporcionarle al estudiante las bases iniciales para el manejo de los métodos y modelos econométricos, los elementos necesarios para el manejo de la información, análisis de resultados e interpretación de las salidas del computador, y familiarizarlo en sus aplicaciones, tales como la investigación y la evaluación de medidas de política.

El documento se encuentra dividido en siete secciones. La primera presenta la definición de la econometría, sus objetivos, el concepto del modelo y su caracterización. La segunda trata de la organización de datos y la estadística descriptiva. La sección tres aborda los aspectos básicos del análisis de correlación. La sección cuatro presenta el modelo formal de regresión lineal simple. La quinta sección muestra el modelo de regresión lineal múltiple. La sexta sección presenta la teoría relacionada con el incumplimiento de los supuestos del modelo. La sección siete efectúa una introducción al análisis transversal de regresión con variables independientes cualitativas y de variable dependiente cualitativa. Al final del documento se incluye un anexo que desarrolla el procedimiento general de manipulación de datos en el paquete econométrico Eviews 4.1.

1. LA MODELACIÓN Y LA ECONOMETRÍA

1.1. Métodos Cuantitativos de la Economía.

Los métodos cuantitativos de la economía comprenden tres áreas: a) Análisis Matemático y Álgebra Lineal; b) Programación Lineal y Análisis de Insumo-Producto y c) Econometría.

La investigación econométrica se inició con el análisis estadístico de la Demanda por Cournout (1838) y Marshall (1890). Posteriormente Tinbergen en 1939 hizo su aporte a la econometría mediante el estudio del análisis de los ciclos económicos. Sin embargo, en el periodo de 1943-1950 la econometría comienza su desarrollo con los trabajos de la Comisión Cowles. La hipótesis básica es: "los datos económicos se generan por sistemas de relaciones que son, en general estocásticos, dinámicos y simultáneos".

La econometría hoy en día es una herramienta muy importante para el análisis y comportamiento de los fenómenos económicos. Su desarrollo ha sido acelerado debido a la dinámica que han mostrado los adelantos en el análisis matemático, en métodos estadísticos y de computación.

1.2. Definiciones de la Econometría

Dado que en la econometría se asocian la Teoría Económica, las Matemáticas y la Estadística, diferentes definiciones han sido planteadas por los autores, en las que se tratan de relacionar estas tres áreas del conocimiento.

G. Tintner: la econometría consiste en la aplicación de la teoría económica matemática y de los métodos estadísticos a los datos económicos para establecer resultados numéricos y verificar los teoremas.

W.C. Hood y T.C. Koopmans: la econometría es una rama de la economía donde la teoría económica y los métodos estadísticos se fusionan en el análisis de los datos numéricos e institucionales.

T. Havellmo: el método de la investigación econométrica intenta, esencialmente, unir la teoría económica y las mediciones reales, empleando la teoría y la técnica de la inferencia estadística como un puente.

Estas tres definiciones nos indican que la econometría es cuantitativa y que está en estrecho contacto con la realidad.

R. Frisch: la econometría a pesar de nutrirse de la Teoría Económica, de las Matemáticas y de la Teoría Estadística, no es ni "Estadística Económica", ni "Teoría Económica", ni "Economía Matemática".

O. Lange: la econometría es la ciencia que trata de la determinación, por métodos estadísticos, de leyes cuantitativas concretas que rigen la vida económica. Esta combina la Teoría Económica con la Estadística Económica y trata de dar, por métodos matemáticos y de inferencia, una expresión concreta a las leyes generales establecidas por la teoría.

1.3. Objetivo de la Econometría

El objetivo de la econometría es expresar la teoría económica en términos matemáticos, verificar dicha teoría por métodos estadísticos, medir el impacto de una variable sobre otra, predecir los sucesos futuros, o proveer recomendaciones de la política económica.

1.4. El Procedimiento Econométrico

El análisis econométrico involucra las siguientes etapas principales:

1. Especificación del modelo: consiste en usar la teoría, leyes o hipótesis particulares económicas, para investigar las relaciones entre variables y agentes de la economía.
2. Estimación del modelo: trata de la utilización de instrumentos auxiliares como las matemáticas y la estadística para estimar el modelo objetivo.
3. Verificación del modelo: en esta etapa se efectúa la interpretación económica del modelo estimado y se realizan pruebas estadísticas. La fase de

verificación tiene un papel muy importante dado que examina si la expresión cuantificada puede utilizarse adecuadamente con base en la teoría económica.

4. Predicción: el modelo obtenido puede ser utilizado para la predicción y el desarrollo de muchas aplicaciones. Pueden surgir nuevos resultados teóricos, y generarse implicaciones de política económica a partir de las conclusiones del modelo.

1.5. El Modelo

Un modelo es una representación simplificada de la realidad. Los investigadores y los profesionales de diversas áreas del conocimiento trabajan con éstos esquemas, los cuales les permiten estudiar el comportamiento de un fenómeno de interés.

A. Rosenblueth se refirió a los modelos científicos de la siguiente manera: "la construcción de modelos para los fenómenos naturales es una de las tareas esenciales de la labor científica. Mas aún, se puede decir que toda la ciencia no es sino la elaboración de un modelo de la naturaleza. La intención de la ciencia y el resultado de la investigación científica, es obtener conocimiento y el control de alguna parte del Universo".

1.6. El Modelo Económico

Se denomina modelo económico a cualquier conjunto de supuestos que describen una economía o parte de una economía. En este sentido, la teoría económica puede entenderse como la formulación y análisis de modelos cuantitativos. Esta esquematización requiere un planteamiento particular de las interrelaciones entre las variables que intervienen en el fenómeno de estudio.

Las características mínimas que debe satisfacer un modelo económico son las siguientes:

1. Que represente un fenómeno económico real.

2. Que la representación sea simplificada, y
3. Que se haga en forma matemática.

Al definir un modelo económico como un conjunto de relaciones matemáticas (usualmente ecuaciones) que expresan una teoría económica, no se exige necesariamente la especificación concreta del tipo de función que relaciona las variables involucradas. Un ejemplo de un modelo económico es:

$$Y = f(X_1, X_2, \ldots, X_k) \qquad (1)$$

donde Y = cantidad producida; X_i = cantidad del i-esimo insumo, i=1,2,…,k.

Aunque esta ecuación, denominada función de producción, no presenta una estructura muy particular del arreglo de las variables X sobre Y, expresa de forma general la relación entre el producto y los insumos, y que son las cantidades utilizadas de factores las que determinan la magnitud producida, y no lo contrario.

Para establecer una forma concreta de la especificación de un modelo se debe precisar el tipo de relación que existe entre las variables económicas. Un ejemplo de ello es una representación lineal:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k \qquad (2)$$

Está relación puede ser correcta. Sin embargo, cuando no se conoce si el insumo X_2 es determinante en forma lineal sobre Y, puede ocurrir error de especificación. También se debe resaltar que este modelo hace énfasis en un número relativamente pequeño de variables importantes cuya interrelación se puede expresar adecuadamente en un modelo matemático.

1.7. El Modelo Econométrico

El modelo econométrico es el modelo económico que contiene las especificaciones necesarias para su validación empírica. Es usual concebir el modelo econométrico como un modelo conformado por una parte determinística y una parte aleatoria o término de error. El modelo econométrico para el ejemplo expuesto en la ecuación (2) tomaría la forma:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k + \varepsilon \qquad (3)$$

donde $\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$ es la parte determinística y ε es el termino de error o componente estocástico.

Los modelos econométricos por considerar un término aleatorio en su estructura, hacen parte de los modelos probabilísticos. Una diferencia fundamental entre los modelos económicos y los modelos econométricos, es que los primeros son siempre validos, dado que han sido establecidos por la teoría económica y solo persiguen la expresión general de ella. Por otro lado, los modelos econométricos, reflejan el estado de las cosas o de una situación específica y aunque tiene sus bases en la teoría económica sus resultados pueden cambiar de un estudio a otro.

Los modelos econométricos se prueban a través del uso sistemático de la información estadística. Un modelo econométrico permite la inferencia estadística a partir de los datos recopilados, por lo cual éste debe incorporar los elementos aleatorios que se suponen intervienen en la determinación de las observaciones. Estas últimas pueden constituyen una muestra si la aleatoriedad de los datos es garantizada.

Existen diferentes razones por las cuales los modelos econométricos deben considerar el término de error, destacándose como las más importantes las siguientes:

a) Datos: en muchos casos el grado de control que se puede tener sobre las variables de interés es bajo. Adicionalmente, aunque se desea obtener los verdaderos valores de las variables, se debe aceptar que puede existir cierto error en la medición.

   Un ejemplo típico ocurre cuando las personas encuestadas por diferentes motivos revelan un ingreso diferente al real y dicha variable se incorpora al modelo. Otro caso semejante sucede cuando se le pregunta al agricultor sobre la cantidad de fertilizante que aplicó por hectárea a su cultivo en la cosecha pasada; éste dado que no se acuerda de dicha magnitud, provee un dato que diverge del real.

b) Número de variables: el investigador siempre tiene restricciones para incluir todas las variables que explican un fenómeno. Por un lado, no cuenta con completa información, y por otro, aunque disponga de demasiada información su formulación es extremadamente compleja que dificulta su interpretación. Por lo tanto, el procedimiento se basa en incluir aquellas variables más relevantes, dejando fuera del modelo aquellas poco significativas. No obstante, el investigador es consciente de que al no poder incluir todas las variables incurre en cierto margen de error al efectuar la estimación.

c) Disponibilidad de información: muchas veces cuando el investigador quiere incluir una variable importante en el modelo se encuentra con la limitación de cómo cuantificarla. Un ejemplo de ello es la variable “habilidad”; se conoce que ésta teóricamente afecta el salario; sin embargo el investigador tiene que conformarse con incluir otra variable o información adicional que sea semejante y la describa de manera aproximada.

d) Forma funcional: un investigador puede postular que la relación entre las variables de un modelo es de tipo lineal; no obstante, otro investigador podría formular una especificación funcional distinta, por ejemplo cuadrática. Esta es otra fuente de error en la elaboración del modelo, pues no se puede tener total certeza sobre su forma funcional aún cuando la teoría señale algunas directrices para corregirlo.

De acuerdo con lo anterior un procedimiento sugerido para llevar a cabo la formulación de un modelo econométrico es el siguiente: 1) Delimitar el fenómeno de estudio; 2) Tener claridad sobre el objetivo del modelo; 3) Seleccionar las variables relevantes; 4) Establecer las relaciones entre las variables, y 5) Con base en el objetivo planteado, estructurar una especificación y estimar el modelo usando la información y base de datos de las variables.

1.8. Elementos que componen el Modelo

Los elementos que componen el modelo son: las variables, las ecuaciones y los parámetros.

Una variable es una característica de una población que puede tomar diferentes valores. Solo son de interés aquellos valores de la variable que tienen un significado económico. Por ejemplo las variables: precio, producción, ingreso, y cantidad de insumo utilizado tienen región económicamente factible en los números reales positivos.

Una ecuación es una igualdad conformada por una expresión matemática que establece relaciones entre variables. La ecuación contiene no solo las variables de interés sino también los coeficientes que afectan estas mismas. A estas últimas magnitudes se les denomina parámetros desde el enfoque estadístico, los cuales en un modelo lineal actúan como factores de ponderación de cada variable explicativa y

miden el efecto de las fluctuaciones de estas variables sobre la variable dependiente. Los parámetros cumplen un papel muy importante en el modelo, ya que sobre estos el investigador formula pruebas de hipótesis. Al observar la ecuación (3), el coeficiente que no acompaña ninguna variable independiente se le conoce como constante paramétrica o intercepto; en algunos casos su magnitud no tiene interpretación económica.

1.9. Clasificación de las Variables

Desde el punto de vista económico las variables se pueden clasificar como variables endógenas y exógenas. Las variables endógenas son aquellas cuyos valores se determinan o calculan dentro del modelo. En contraste, las variables exógenas se caracterizan por que sus valores están determinados fuera del modelo.

También existen otras clasificaciones de las variables; desde el enfoque de inferencia estadística: variables aleatorias discretas y continuas, y de acuerdo con su rol en expresión matemática: variables dependientes e independientes, explicadas o explicativas. Otro grupo de variables lo constituyen las variables predeterminadas. A este pertenecen las variables exógenas con o sin rezago (o retardo) y las endógenas rezagadas. Una denominación adicional son las variables esperadas o de expectativas, las cuales son gran utilidad en la formulación de modelos dinámicos.

1.10. Clasificación de las Ecuaciones

Bajo la perspectiva económica las ecuaciones se pueden clasificar de la siguiente forma:

a) Ecuaciones de comportamiento: Son aquellas que reflejan el comportamiento de los distintos agentes económicos (consumidores, productores, inversionistas, etc.). Las ecuaciones de comportamiento son las que mayor aporte teórico le

hacen a los modelos. Ejemplos de ecuaciones de comportamiento son: la demanda, la oferta, la inversión, el consumo, el ahorro, etc.

b) Ecuaciones tecnológicas: El ejemplo típico de una ecuación tecnológica es la función de producción, la cual refleja el estado de la tecnología de un sector ó de un país.

c) Ecuaciones institucionales: Reflejan un mandato o voluntad del gobierno o de los estamentos que toman las decisiones en un país. Ejemplo de ecuaciones institucionales son: oferta monetaria, impuestos, subsidios, etc.

d) Ecuaciones de definición: Son ecuaciones o identidades matemáticas y económicas válidas por definición. Generalmente son relaciones contables y la mayoría de los ejemplos de este tipo de ecuaciones se encuentran en las cuentas macroeconómicas. Una ecuación de definición es activo = pasivo + capital, o la ecuación de identidad macroeconómica del Producto Nacional Bruto para una economía con tres sectores.

e) Ecuaciones de equilibrio: Estas garantizan que el modelo tenga solución. Ejemplos de estas ecuaciones son: oferta igual a demanda, o ahorro igual a inversión.

1.11. Clasificación de los Modelos

Según la cobertura económica o subdisciplina, los modelos pueden ser microeconómicos o macroeconómicos. De acuerdo con el número de variables independientes, los modelos se dividen en simples y múltiples. Si se considera el número de ecuaciones se tienen modelos uniecuacionales y multiecuacionales. Con base en el periodo de tiempo, los modelos pueden ser estáticos o dinámicos. Al relacionar el número de variables endógenas con el número de ecuaciones, los modelos se dividen en completos o incompletos.

2. ORGANIZACIÓN DE DATOS Y ESTADÍSTICA DESCRIPTIVA

2.1. Objetivos de la Estadística

La estadística es el lenguaje universal de la ciencia, tanto en sus ramas físicas como sociales. La estadística es un instrumento formal que utilizado de manera rigurosa y con precisión, permite describir resultados y adoptar decisiones respecto a lo que estos evidencian empíricamente. La estadística en su aplicación sigue el método científico y se define como la ciencia de recolectar, clasificar, describir e interpretar datos numéricos, es el lenguaje universal de la ciencia y el estudio de los fenómenos aleatorios. Dentro de sus objetivos fundamentales se encuentra la estimación de una o más características desconocidas de una población, la realización de inferencias y pruebas de hipótesis.

Se considera fundador de la estadística a Godofredo Achenwall, economista alemán (1719-1772), quien siendo profesor de la universidad de Leipzig, escribió sobre el descubrimiento de una nueva ciencia que llamó estadística (palabra derivada de Staat que significa gobierno) y que definió como “el conocimiento profundo de la situación respectiva y comparativa de cada estado”. Desde su aparición la estadística se ha enriquecido continuamente con los aportes de matemáticos, filósofos y científicos.

La teoría general de la estadística es aplicable a cualquier campo científico del cual se toman observaciones. El estudio y aplicación de los métodos estadísticos son necesarios en todos los campos del saber, sean estos de nivel técnico o científico. Las primeras aplicaciones de la estadística fueron los temas de gobierno, luego las utilizaron las compañías de seguros y los empresarios de juegos de azar; posteriormente los comerciantes, los industriales, los educadores, etc. En la actualidad resulta difícil indicar profesiones que no utilicen la estadística.

2.2. Divisiones de la Estadística

La estadística puede dividirse ampliamente en dos áreas: estadística descriptiva o deductiva y estadística inferencial o inductiva. La estadística descriptiva es aquella en la que la mayoría de la personas piensan cuando escuchan la palabra "estadística". Esta es el área de la estadística dedicada a la recolección, presentación, y descripción de datos numéricos, cuyas conclusiones sobre los mismos no van más allá de la información que estos proporcionan. Por otro lado, la inferencia estadística es el método y conjunto de técnicas que se utilizan para obtener conclusiones más allá de los límites del conocimiento aportado por los datos; en otras palabras, busca obtener la información que describe y caracteriza una población a partir de los datos de una muestra.

2.3. Tipos de Variables

En estadística cuando se recopila información, los datos se registran por medio de la observación o medición de una variable aleatoria que proviene de la realización de un experimento. La variable se llama aleatoria, debido a la existencia de distintos resultados posibles del experimento y que no hay certeza total de que al efectuarlo uno de los resultados se obtenga siempre con una probabilidad del 100%. Por lo tanto, el hecho que una variable tome un valor particular es considerado un evento aleatorio.

Aún, cuando las observaciones resultantes no siempre son numéricas en algunos experimentos, estas pueden cuantificarse asignándoles números que indiquen o representen una categorización. Por esta razón, el interés se centra generalmente en variables que pueden representarse numéricamente.

Existen dos tipos de variables aleatorias: discretas y continuas. Las primeras son aquellas cuyo número de valores que pueden tomar es contable (ya sea finito o

infinito) y pueden arreglarse en una secuencia que corresponde uno a uno con los enteros positivos; mientras las segundas toman valores dentro de un intervalo de recta de los números reales. Si se tienen dos variables aleatorias, por ejemplo: el número de hijos por familia y el consumo de energía eléctrica; la primera, se encuentra dentro del grupo de variables aleatorias discretas, y la segunda, dentro del conjunto de variables aleatorias continuas.

2.4. Población y Muestra

El concepto de población y muestra es muy importante en la inferencia estadística, por lo que es conveniente presentar su definición:

- Población: Es la colección completa de individuos, objetos o medidas que tienen una característica en común. La población debe definirse cuidadosamente en cada estudio científico de acuerdo con el interés y objetivo de la investigación.

- Muestra: Es un subconjunto de la población; es decir, ella se compone de algunos de los individuos, objetos o medidas de una población. La muestra es obtenida con el propósito de investigar, a partir del conocimiento de sus características particulares, las propiedades de toda la población. Por ello, es primordial la selección de una muestra representativa de la población. Es necesario formalmente enfatizar en la aleatoriedad de la muestra, es decir sobre la manera de seleccionar los elementos de la población que conformarán la muestra. La palabra “aleatoriedad” para este caso consiste en garantizar que cada elemento de la población tenga la misma probabilidad de ser elegido. Se considera que una muestra es más eficiente, cuando proporciona la mayor información útil al menor costo.

Los conceptos anteriores pueden tratarse en el siguiente ejemplo: Suponga que se desea conocer el consumo promedio por hogar de energía eléctrica en la ciudad de

Bogotá. Para este caso, la población corresponde a todos los hogares de la ciudad, mientras que la muestra estará constituida por aquellos hogares que pueden ser seleccionados de manera aleatoria, como un grupo representativo de todos los que habitan en Bogotá.

2.5. Parámetros Poblacionales y Estadísticos Muestrales

El término "parámetro" es utilizado para referirse a una característica desconocida de la población, que desea estimarse o evaluarse a través de una prueba de hipótesis, y que describe total o parcialmente su función de probabilidad o función de densidad de probabilidad. Por otro lado, el "estadístico" es una medida numérica de una característica poblacional obtenida a partir de una muestra. Cabe anotar que los estadísticos son fundamentales en la realización de inferencias. El valor promedio y la varianza son ejemplos de tales medidas.

2.6. Medidas de Tendencia Central y de Dispersión

Las medidas de tendencia central se encuentran dentro de las medidas numéricas que se emplean comúnmente para describir conjuntos de datos. La tendencia central de un conjunto de datos es la disposición de éstos para agruparse, ya sea alrededor del centro o de ciertos valores numéricos. A este grupo de medidas pertenecen la media, la mediana y la moda.

Existen otro tipo de medidas numéricas denominadas medidas de dispersión, cuyo objetivo es explorar la variabilidad de los datos, es decir qué tan dispersas son las observaciones en un conjunto de datos. Dentro de estas medidas se encuentran: la varianza, la desviación estándar, el recorrido o rango, entre otras.

2.7. Métodos y Diagnósticos Gráficos.

Los datos en los experimentos son recopilados inicialmente "sin agrupar", para

luego, según el interés del investigador presentarlos “agrupados”, en forma de clases o intervalos. Es importante tener en cuenta que las fuentes de información primaria y secundaria pueden almacenar sus datos “*sin agrupar*” o como datos “*agrupados*”. Con base en lo anterior, es relevante conocer el procedimiento de cálculo de las medidas numéricas para ambos casos. Las expresiones algebraicas que describen la forma de obtener las medidas de tendencia central y de dispersión se muestran en la Tabla No. 1.

Con los datos agrupados de una variable aleatoria es posible construir histogramas de frecuencias, los cuales pueden ser comparados con las representaciones gráficas de distribuciones de probabilidad ya conocidas de variables aleatorias. En la mayoría de los casos, estos histogramas se comparan con la distribución normal, donde por inspección es posible identificar sesgos o apuntamientos en la distribución.

TABLA No. 1. MEDIDAS DE TENDENCIA CENTRAL Y DE DISPERSIÓN.

Medida Numérica	*Datos sin agrupar*	*Datos agrupados*
Media	$\bar{x} = \sum_{i=1}^{n} x_i / n$	$\bar{x} = \sum_{i=1}^{k} f_i x_i / n$, donde $n = \sum_{i=1}^{k} f_i$ Donde f_i es la frecuencia absoluta de la clase i, para todo i = 1,2,…,k clases o intervalos.
Mediana	Valor central de la distribución (el 50% de los datos se encuentran por encima de este valor).	$Mediana = L + c(j/f_m)$ Donde L es el límite inferior de la clase donde se encuentra la mediana, f_m es la frecuencia de esa clase, c es la longitud de ese intervalo y j es el número de observaciones en esta clase necesarias para completar un total de $n/2$.
Moda	Valor más frecuente	Casos: • Punto medio de la clase con frecuencia más alta. • El promedio de los puntos medios de las clases consecutivas con frecuencias iguales más altas. • Puntos medios de las clases no consecutivas con frecuencias iguales más altas.

Medida Numérica	*Datos sin agrupar*	*Datos agrupados*
Varianza	$s^2 = \sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)$ ó $s^2 = \dfrac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}$	$s^2 = \dfrac{\sum_{i=1}^{k} f_i x_i^2 - \dfrac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}}{n-1}$
Desviación Estándar	$s = \sqrt{s^2} = \sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 / (n-1)}$ ó $s = \sqrt{s^2} = \sqrt{\dfrac{\sum_{i=1}^{n} x_i^2 - \dfrac{\left(\sum_{i=1}^{n} x_i\right)^2}{n}}{n-1}}$	$s = \sqrt{s^2} = \sqrt{\dfrac{\sum_{i=1}^{k} f_i x_i^2 - \dfrac{\left(\sum_{i=1}^{k} f_i x_i\right)^2}{n}}{n-1}}$
Recorrido o Rango	*Max-min.*	

2.8 Ejercicios de computador

Considérese el siguiente conjunto de datos hipotéticos de un estudio de demanda:

TABLA No. 2. DATOS HIPOTÉTICOS EN EL ESTUDIO DE DEMANDA DEL BIEN X.

No. de Obs.	DX	PX	PZ	PW	I
1	37	7	5	7	6
2	38	6	7	5	8
3	18	10	3	13	3
4	50	4	9	4	18
5	22	9	3	11	3
6	55	2	12	3	21
7	42	8	5	8	2
8	29	8	5	9	19
9	63	2	18	3	20
10	13	12	2	15	6
11	60	3	9	5	12
12	62	3	10	5	5
13	36	6	5	6	26

Donde:
DX: es la demanda del bien X
PX: es el precio del bien X
PZ: es el precio del bien Z
PW: es el precio del bien W
I: es el ingreso

ESTADÍSTICAS DESCRIPTIVAS

MEDIDAS DE TENDENCIA CENTRAL, DE DISPERSION Y NORMALIDAD

	DX	PX	PZ	PW	I
Mean	40.38462	6.153846	7.153846	7.230769	11.46154
Median	38.00000	6.000000	5.000000	6.000000	8.000000
Maximum	63.00000	12.00000	18.00000	15.00000	26.00000
Minimum	13.00000	2.000000	2.000000	3.000000	2.000000
Std. Dev.	16.89940	3.210560	4.431820	3.811252	8.272599
Sum	525.0000	80.00000	93.00000	94.00000	149.0000
Observations	13	13	13	13	13

3. ANALISIS DE CORRELACION

3.1. Diagrama de Dispersión

Una primera aproximación con el fin de detectar algún tipo de relación entre dos variables (X y Y), consiste en ubicar los pares de valores de en un plano cartesiano hasta conformar la nube de puntos. Un diagrama de dispersión es la representación gráfica de todos los pares de valores en sistema de ejes de coordenadas.

El diagrama de dispersión no es un método estadístico como tal, más bien estaría dentro de los llamados métodos de "inspección preliminar", sin embargo, es una manera simple de visualizar si existe alguna posible relación entre las variables. El diagrama de dispersión puede presentar diferentes formas, tales como las que se presentan en las figuras siguientes:

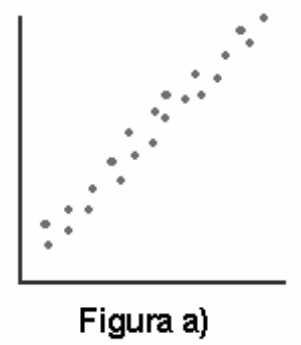


Figura a)

Figura b)

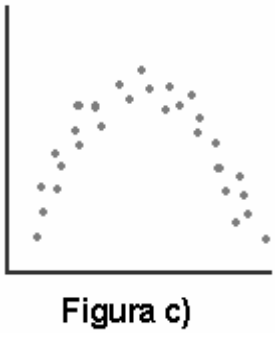


Figura c)

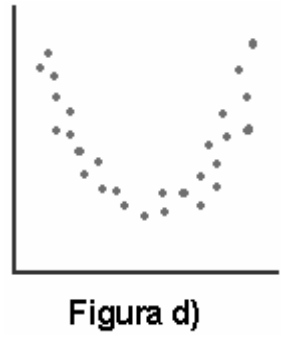


Figura d)

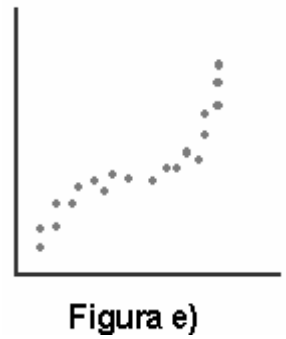


Figura e)

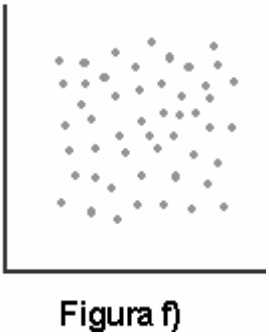


Figura f)

La figura a) muestra una posible relación lineal directa entre las variables; mientras,

la figura b) señala una relación lineal de tipo inversa. Las figura c) y d) revelarían posibles relaciones cuadráticas entre las variables, exhibiendo un máximo y un mínimo para la primera y segunda de estas figuras, respectivamente. La figura e) mostraría una tendencia de tipo cúbico entre las variables. La figura f) es un ejemplo en el cuál no puede identificarse por inspección algún tipo de relación entre las variables, pues aparentemente ella no existe.

3.2. Coeficiente de Correlación Lineal

Si bien es cierto que el diagrama de dispersión permite visualizar la existencia o no de una posible relación lineal entre las variables, el investigador debe soportar sus conclusiones en términos de alguna medida estadística. El coeficiente de correlación lineal es un estadístico que mide el tipo de relación (signo) y la fuerza (magnitud del coeficiente) de asociación lineal entre dos variables. Usualmente el coeficiente de correlación lineal, representado por la letra r, bajo las condiciones de un muestreo aleatorio ideal se considera una buena representación del coeficiente de correlación poblacional (ρ). La fórmula para calcular r es la siguiente:

$$r_{XY} = \frac{\hat{Cov}(X,Y)}{S_X S_Y}$$

$$r_{XY} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}}$$

$$r_{XY} = \frac{\sum x_i y_i - \frac{(\sum x_i)(\sum y_i)}{n}}{\sqrt{\left[\sum x_i^{\,2} - \frac{(\sum x_i)^2}{n}\right]\left[\sum y_i^{\,2} - \frac{(\sum y_i)^2}{n}\right]}} = \frac{\sum x_i y_i - n(\bar{x}\,\bar{y})}{\sqrt{\left[\sum x_i^{\,2} - n(\bar{x})^2\right]\left[\sum y_i^{\,2} - n(\bar{y})^2\right]}}$$

El coeficiente de correlación no tiene unidades y puede tomar valores entre -1 y +1 $(-1 \le r_{XY} \le 1)$. Su interpretación depende del signo y la magnitud que tome. El signo es determinado solamente por el numerador de la fórmula de cálculo; es decir por la

covarianza, la cual mide la asociación lineal absoluta entre las variables; el denominador es siempre positivo dado que en él se encuentran sumas de cuadrados.

Si r tiende a 1 como seria el caso de la figura a) estaría indicando una relación lineal positiva o directa entre las variables. Si r tiende a -1, existiría una relación lineal negativa o inversa entre las variables. Cuando r es exactamente igual a 1 o -1 la relación lineal es perfecta, siendo posible ajustar todos los puntos a través de una línea recta con pendiente positiva (ver figura g) o negativa (ver figura h), respectivamente. Si r es cero no hay relación lineal entre las variables y una línea horizontal une todos los pares de valores localizados en el diagrama de dispersión (ver figura i).

Figura g)

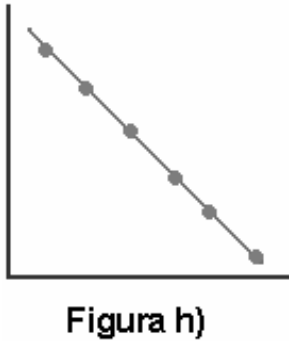


Figura h)

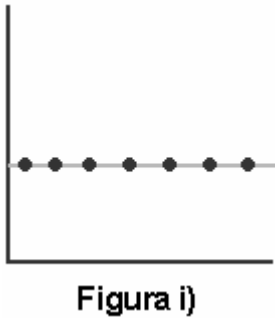


Figura i)

La ventaja principal del coeficiente de correlación lineal es su fácil cálculo e interpretación. Sin embargo, cuando las variables presentan algún tipo de relación no lineal, r no puede medir esta clase de asociación. Así mimo, dado que r calcula la dependencia lineal solo entre pares de variables, no proporciona información sobre la asociación simultánea de más de dos variables.

A continuación se presentan las propiedades del coeficiente de correlación:

1. r es de naturaleza simétrica. Esto indica que el coeficiente de correlación entre X y Y es igual al coeficiente de correlación entre Y y X.

2. r es independiente del origen y la escala. Si se define $X^{*i} = aX_i + c$ y $Y^{*i} = bY_i + d$, donde a>0, b>0, y c y d son constantes, entonces r entre X^* y Y^* (variables transformadas) es igual al r entre X y Y (variables originales).

3. Si X y Y son variables estadísticamente independientes, el coeficiente de correlación lineal entre X y Y es cero. No obstante, si r es cero, esto no implica necesariamente que X y Y sean estadísticamente independientes.

Una de las condiciones para que el coeficiente de correlación se pueda aplicar es que las variables sean continuas y con distribución normal. En caso de que esto no se cumpla como es el caso de variables discretas se debe buscar otra medida estadística para evaluar la dependencia entre las variables. Una alternativa para ello son las tablas de contingencia.

3.3. Pruebas de Hipótesis

La formalidad estadística sugiere realizar pruebas de hipótesis sobre los parámetros poblacionales basándose en los estadísticos encontrados. Por ejemplo, aún cuando el coeficiente de correlación lineal estimado entre dos variables sea diferente de cero, esto no es suficiente para afirmar que el parámetro poblacional ρ es en realidad distinto de cero, pues requiere recordarse que las inferencias se efectúan con base en información muestral y existe un margen de error cuando se realiza este tipo de procedimiento. A continuación se presenta el esquema de prueba de hipótesis para el coeficiente de correlación lineal cuando el investigador desea evaluar si hay o no dependencia lineal entre un par de variables. Por lo tanto, se desea probar si el parámetro poblacional es o no diferente de cero:

Paso 1: Planteamiento de la hipótesis:

Ho: $\rho = 0$

Ha: $\rho \neq 0$

Paso 2: Nivel de significancia. Representa el nivel de error máximo tolerable para realizar la prueba. Este es establecido o definido por el investigador y se denota con la letra α. Los valores de significancia con los cuales se trabajan pueden cambiar de una disciplina o ciencia a otra. Bajo situaciones donde los experimentos tienen una alto grado de control, usualmente se trabaja con niveles del 1% y 5%, (altamente significativo y significativo, respectivamente). En las investigaciones de las ciencias sociales, donde existe un limitado grado de control sobre las variables, pueden encontrarse significancias estadísticas del 10% y en algunas ocasiones hasta un 20%.

Paso 3: El estadístico de prueba. Es una medida estadística calculada a partir de información muestral o experimental para llevar a cabo la prueba. Para el caso de correlación lineal simple, el estadístico de prueba se define como:

$$t_C = \frac{\left(r\sqrt{n-2}\right) - \theta}{\sqrt{1-r^2}} \sim t_{\alpha/2,\, n-2}$$

donde r es el coeficiente de correlación lineal muestral, n es el tamaño de la muestra, n-2 los grados de libertad de la prueba y θ el valor del parámetro poblacional en la hipótesis nula. En este ejemplo particular, θ toma el valor de cero, pero en otras pruebas, de acuerdo con lo que desee evaluar el investigador θ puede corresponder a un valor distinto de cero, entre –1 y 1.

Paso 4: Regiones de decisión. Dado que la hipótesis alterna señala el símbolo $\neq$, se trabaja con los dos lados de la distribución. La región de rechazo estará repartida en los extremos de la función de probabilidad, con un valor de $\alpha/2$ a cada lado. Los valores de los límites derecho e izquierdo que limitan las regiones de rechazo se determinan mediante el uso de la tabla t con sus respectivos grados de libertad. Estos valores de t se denominan estadísticos de contraste. La figura j muestra la región de rechazo (RHo) y aceptación (AHo) de la hipótesis nula de esta prueba:

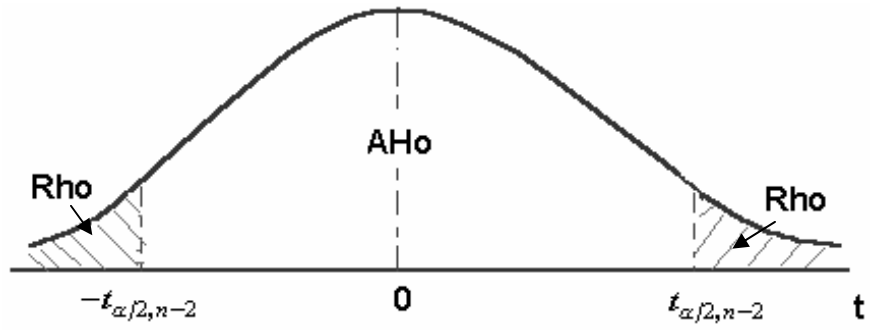


Figura j)

Paso 5: Criterio de decisión y conclusión del investigador. Se debe comparar el estadístico calculado o de prueba (t_C) contra el estadístico tabulado $(t_{\alpha/2,n-2})$. El criterio de decisión esta basado en: 1) si el t calculado es mayor que el t de tablas positivo, cae en la región de rechazo del lado derecho de la distribución y la decisión que se debe tomar es rechazar la hipótesis nula ($\rho \neq 0$); 2) si el t calculado es menor que el t de tablas negativo, el t calculado cae en la región de rechazo del lado izquierdo y la decisión igualmente es rechazar la hipótesis nula ($\rho \neq 0$); y 3) si el t calculado se encuentra entre el -t y t de las tablas, el t calculado cae en la región de aceptación y la decisión es no rechazar la hipótesis nula ($\rho = 0$). Posteriormente, el investigador basado en el criterio de decisión concluye e interpreta

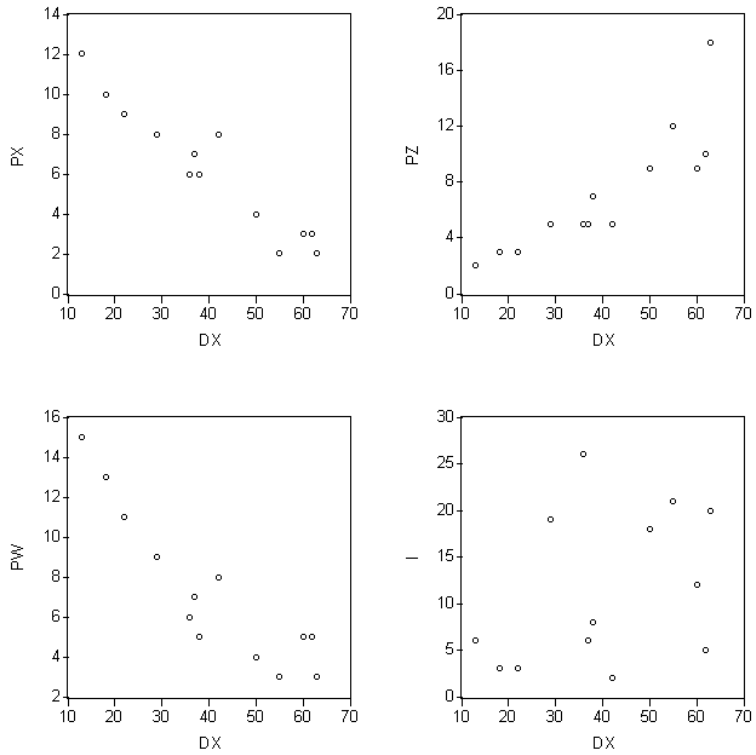
los resultados de la prueba, y plantea las recomendaciones pertinentes.

La significancia estadística del coeficiente de correlación en la prueba de hipótesis se afecta por el tamaño de la muestra (n) o mejor aún por los grados de libertad, lógicamente a mayor tamaño de la muestra el valor de r tiene mayor confiabilidad. Si se encuentra un valor de r relativamente bajo y n es grande, es posible que éste sea significativo al comparar el estadístico de prueba con el de contraste o de tablas; alternativamente se puede encontrar un r alto pero no significativo estadísticamente debido a que n es muy pequeño y por consiguiente el número de grados de libertad es bajo.

3.4. Ejercicios de computador

Usando los mismos datos del ejemplo hipotético de demanda planteado en el capítulo anterior, a continuación se presenta el diagrama de dispersión, y la matriz de covarianzas y de correlación de las variables:

DIAGRAMAS DE DISPERSIÓN



MATRIZ DE COVARIANZAS

VARIABLE	DX	PX	PZ	PW	I
DX	263.6213	-47.98225	60.01775	-53.78107	47.89941
PX	-47.98225	9.514793	-11.63905	10.73373	-12.99408
PZ	60.01775	-11.63905	18.13018	-12.65089	16.69822
PW	-53.78107	10.73373	-12.65089	13.40828	-16.18343
I	47.89941	-12.99408	16.69822	-16.18343	63.17160

MATRIZ DE CORRELACION

VARIABLE	DX	PX	PZ	PW	I
DX	1.000000	-0.958056	0.868137	-0.904592	0.371175
PX	-0.958056	1.000000	-0.886170	0.950308	-0.530011
PZ	0.868137	-0.886170	1.000000	-0.811397	0.493410
PW	-0.904592	0.950308	-0.811397	1.000000	-0.556062
I	0.371175	-0.530011	0.493410	-0.556062	1.000000

4. REGRESION SIMPLE LINEAL Y NO LINEAL

4.1. Objetivo del análisis de regresión

El objetivo fundamental del análisis de regresión es el estudio de la dependencia de una variable, llamada explicada, de una o más variables llamadas explicativas. El análisis de regresión se apoya en el concepto matemático de función, en el que se tiene una variable dependiente (variable explicada) y un conjunto de variables independientes (variables explicativas) con el fin de estimar los coeficientes o parámetros de dicha función y efectuar predicciones (encontrar el valor esperado de la variable dependiente cuando se construyen escenarios reflejados en los valores que toman las variables independientes).

Todo procedimiento econométrico sigue los siguientes pasos: la especificación, la estimación, la verificación y la predicción. A continuación se presenta una breve descripción de cada etapa:

1. Especificación: corresponde a la etapa en que el investigador define la forma funcional del modelo que desea utilizar para explicar la variable dependiente siguiendo los lineamientos de la teoría económica.

2. Estimación: durante esta se calculan los valores numéricos de los coeficientes o parámetros del modelo; para ello es necesario apoyarse en los métodos de estimación y la aplicación de rutinas de computador usando paquetes estadísticos (Eviews).

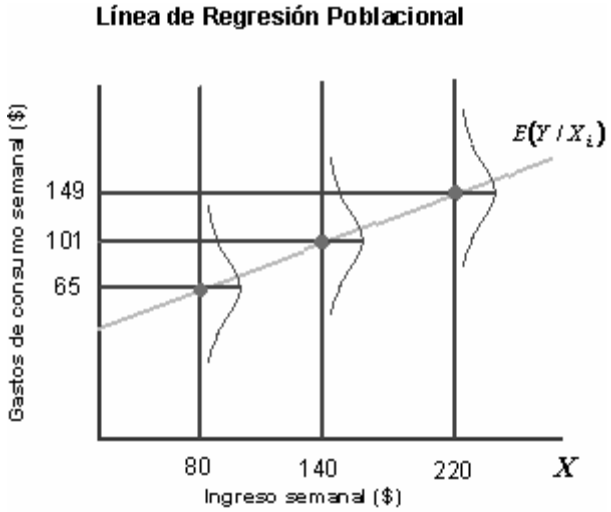
3. Verificación: consiste en corroborar la validez teórica y estadística del modelo, es decir, evaluar si los signos obtenidos para los coeficientes estimados son los esperados y si el modelo cuenta con propiedades estadísticas adecuadas (buen ajuste, alta relevancia y dependencia).

4. Predicción: muchas veces los modelos elaborados por los economistas no tienen solo como objeto mostrar la relación entre variables y la magnitud de dicha relación entre estas a través de una forma funcional, sino que además los modelos tienen implicaciones en términos de predicción. En este sentido puede encontrarse el efecto esperado sobre la variable dependiente para diversos valores de las variables independientes fuera del rango muestral. En este procedimiento la inferencia estadística juega un papel importante.

4.2. Función de regresión muestral y poblacional

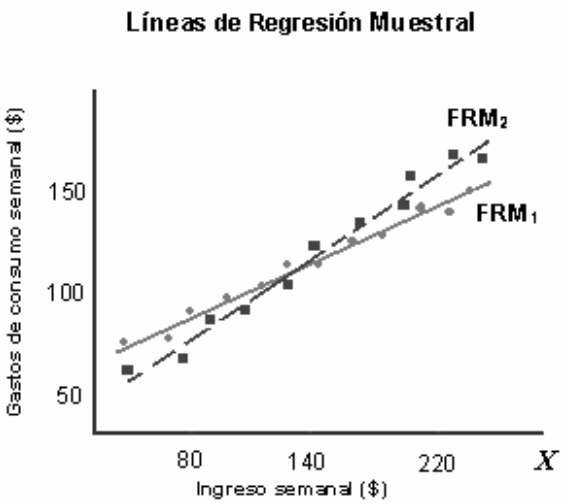
La línea de regresión $E(Y / X_i) = \beta_1 + \beta_2 X_i$ es la unión de los puntos que representan los valores esperados de variable dependiente Y dado los valores de las variables independientes X´s. Esta línea se puede construir a partir del diagrama de dispersión conformado por los datos poblaciones; en este caso la línea de regresión se conoce como la función de regresión poblacional. A continuación se presenta una

gráfica de la línea de regresión poblacional cuando el gasto en consumo de un hogar se desea explicar por el ingreso.



Por otro lado, cuando la línea de regresión es construida con los datos muestrales recibe el nombre de función de regresión muestral. Como todo procedimiento de inferencia estadística, lo que se pretende es que la muestra sea una buena representación de la población. En este sentido, la función de regresión muestral constituye una representación de la función de regresión poblacional. A sí mismo, en la práctica, las muestras de variables aleatorias son usadas para inferir sobre las características de la población.

La siguiente gráfica presenta un ejemplo de dos líneas de regresión muestral para el gasto en consumo semanal de un hogar versus el ingreso.



4.3. Supuestos del modelo de regresión

Los supuestos del modelo junto con los métodos de estimación caracterizan los resultados obtenidos de la regresión (coeficientes, pruebas de hipótesis, intervalos de confianza, predicción, etc.). En particular, los supuestos más importantes del modelo recaen sobre el término del error. Teniendo en cuenta que la función de regresión poblacional puede expresarse también de la forma $\boldsymbol{Y_i = \beta_1 + \beta_2 X_i + u_i}$, el modelo de regresión lineal cuenta con los siguientes supuestos:

Supuesto 1: El modelo de regresión es lineal en los parámetros:

$$Y_i = \beta_1 + \beta_2 X_i + u_i$$

Supuesto 2: Los valores de X son fijos en muestreos repetitivos. Técnicamente esto consiste en que X se supone no estocástica.

Supuesto 3: El valor medio de la perturbación u_i es igual a cero.

$$E[u_i / X_i] = 0$$

Por lo tanto los factores que no están incluidos en el modelo y que por consiguiente, están incorporados en u_i, no afectan sistemáticamente el valor de la media de Y.

Supuesto 4: Homoscedasticidad o varianza constante de u_i. Dado el valor de X, la varianza de u_i es constante para todas las observaciones. Esto es, las varianzas condicionales de u_i son idénticas.

$$Var[u_i / X_i] = E[u_i - E[u_i] / X_i]^2$$
$$Var[u_i / X_i] = E[u_i^2 / X_i]$$
$$Var[u_i / X_i] = \sigma^2$$

La anterior ecuación, establece que la varianza de u_i para cada X_i, es algún número positivo constante igual a σ^2. Nótese que el supuesto 4 implica que las varianzas condicionales de Y_i también son homoscedásticas. Esto es: $Var[Y_i / X_i] = \sigma^2$.

En contraste, si la varianza condicional de la población Y varia con X, esta situación se conoce como Heteroscedasticidad, lo cual puede escribirse como:

$$Var[u_i / X_i] = \sigma_i^2$$

Obsérvese el subíndice sobre σ^2 en esta expresión indica que la varianza de la población Y ahora no es constante.

Supuesto 5: No auto correlación entre las perturbaciones. Dados dos valores cualquiera de X, por ejemplo X_i y X_j $(i \neq j)$, la correlación entre u_i y u_j para todo $i \neq j$ es cero.

$$Cov(u_i, u_j / X_i, X_j) = E[u_i - E[u_i] / X_i][u_j - E[u_j] / X_j]$$
$$Cov(u_i, u_j / X_i, X_j) = E[u_i / X_i][u_j / X_j]$$
$$Cov(u_i, u_j / X_i, X_j) = 0$$

donde i y j son dos observaciones diferentes. Este es también llamado supuesto de no correlación serial. Supóngase que en la función de regresión poblacional $Y_t = \beta_1 + \beta_2 X_t + u_t$, u_t y u_{t-1} están correlacionados positivamente. Entonces Y_t depende no solamente de X_t sino también de u_{t-1}, puesto que u_{t-1} determina en cierta medida a u_t.

Supuesto 6: La covarianza entre u_i y X_i es cero, o $E[u_i, X_i] = 0$.

$$Cov[u_i, X_i] = E[u_i - E[u_i]][X_i - E[X_i]]$$
$$Cov[u_i, X_i] = E[u_i (X_i - E[X_i])] \qquad E[u_i] = 0$$
$$Cov[u_i, X_i] = E[u_i X_i] - E[X_i]E[u_i] \qquad E[X_i] \_ no \_ estocastica$$
$$Cov[u_i, X_i] = E[u_i X_i], \qquad E[u_i] = 0$$
$$Cov[u_i, X_i] = 0$$

Supuesto7: El número de observaciones *n* debe ser mayor que el número de parámetros por estimar.

Supuesto 8: Variabilidad en los valores de X. Se requiere que no todos los valores de X en una muestra dada sean iguales. Así la $Var[X]$ es un número finito positivo.

Supuesto 9: El modelo de regresión esta correctamente especificado. La omisión de variables importantes del modelo o la escogencia de una forma funcional equivocada afectan la validez de la interpretación de la regresión estimada.

Supuesto 10: No hay correlación lineal perfecta entre variables explicativas.

Cuando el modelo de regresión cumple con los anteriores supuestos se le conoce como modelo de regresión clásico y tiene las siguientes propiedades: los estimadores son MELI (mejores estimadores lineales insesgados). Si se agrega el supuesto de normalidad de los errores, los estimadores son MEI (mejores estimadores insesgados) y por lo tanto seguirán distribución normal. Con ello, los intervalos de confianza, las predicciones y las pruebas de hipótesis tienen validez estadística.

4.4. Método de estimación de mínimos cuadrados ordinarios

El objetivo principal de la etapa de estimación es encontrar los valores de los parámetros muestrales. El método de estimación más popular recibe el nombre de mínimos cuadrados ordinarios (MCO). El criterio de este método consiste en proporcionar estimadores de los parámetros que minimicen la suma de los cuadrados de los errores. Operativamente el proceso es construir una función objetivo en términos de la suma de los cuadrados de los errores y mediante optimización (condiciones de primer orden - C.P.O., y condiciones de segundo orden - C.S.O.) obtener las fórmulas de cálculo de los estimadores.

Debido a que la función de regresión poblacional no se puede observar directamente, los estimadores de mínimos cuadrados ordinarios se obtienen a partir de la función de regresión muestral (FRM). La FRM es:

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i$$

$$Y_i = \hat{Y}_i + e_i$$

La suma del cuadrado de los errores puede expresarse como sigue:

$$\sum e_i^2 = \sum \left(Y_i - \hat{Y}_i\right)^2 = \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

De acuerdo con el principio de mínimos cuadrados ordinarios:

$$\min \quad \sum e_i^2 = \min \quad \sum \left(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i\right)^2$$

Derivando la anterior expresión con respecto a $\hat{\beta}_1$ y $\hat{\beta}_2$ e igualando a cero, respectivamente, y resolviendo las ecuaciones normales, se encuentran los estimadores de los parámetros de la regresión:

$$\hat{\beta}_2 = \frac{n\sum X_i Y_i - \left(\sum X_i\right)\left(\sum Y_i\right)}{n\sum X_i{}^2 - \left(\sum X_i\right)^2} = \frac{\hat{Cov}(X,Y)}{\hat{Var}(X)}$$

$$\hat{\beta}_1 = \overline{Y} - \hat{\beta}_2 \overline{X}$$

4.5. Varianzas y errores estándar de los estimadores

Así como existen medidas de dispersión para las variables también las hay para los estimadores, por lo tanto, es necesario siempre presentar una medida de precisión de los estimadores de los parámetros del modelo. Esta medida es el error estándar e indica la confiabilidad de las estimaciones (si son pequeñas dejan ver que los parámetros muestrales van a ser muy parecidos a los poblacionales). La principal utilidad de los errores estándar de los estimadores es la construcción de intervalos de confianza y la prueba de hipótesis. A continuación se presenta la forma de calcular la varianza y error estándar de cada estimador del modelo de regresión lineal simple:

$$Var\left(\hat{\beta}_1\right) = \frac{\sum X_i^2}{n\sum \left(X_i - \overline{X}\right)^2}\sigma^2 \qquad se\left(\hat{\beta}_1\right) = \sqrt{\left[\frac{\sum X_i^2}{n\sum \left(X_i - \overline{X}\right)^2}\right]}\,\sigma$$

y

$$Var\left(\hat{\beta}_2\right) = \frac{\sigma^2}{\sum \left(X_i - \overline{X}\right)^2} \qquad se\left(\hat{\beta}_2\right) = \frac{\sigma}{\sqrt{\sum \left(X_i - \overline{X}\right)^2}}$$

4.6. Intervalos de confianza

En estadística es común efectuar inferencias basadas en estimaciones puntuales y en intervalos. Estas últimas son menos riesgosas debido a que se encuentran dentro de un rango con cierto margen de confiabilidad. En particular, pueden construirse intervalos de confianza para los parámetros del modelo de regresión así como para las predicciones.

Un intervalo de confianza para el parámetro β_2 puede presentarse como sigue:

$$\Pr\left[\hat{\beta}_2 - t_{\alpha/2}\, se\left(\hat{\beta}_2\right) \leq \beta_2 \leq \hat{\beta}_2 + t_{\alpha/2}\, se\left(\hat{\beta}_2\right)\right] = 1 - \alpha$$

donde α es el nivel de significancia estadística y $se\left(\hat{\beta}_2\right)$ es el error estándar de β_2. $100(1-\alpha)$ es el nivel porcentual de confianza del intervalo. Una versión abreviada de esta expresión es: $\hat{\beta}_2 \pm \boldsymbol{t}_{\alpha/2}\, \boldsymbol{se}\left(\hat{\beta}_2\right)$. De la misma forma para β_1:

$$\mathbf{Pr}\left[\hat{\beta}_1 - \boldsymbol{t}_{\alpha/2}\, \boldsymbol{se}\left(\hat{\beta}_1\right) \leq \beta_1 \leq \hat{\beta}_1 + \boldsymbol{t}_{\alpha/2}\, \boldsymbol{se}\left(\hat{\beta}_1\right)\right] = 1 - \alpha$$

$$\hat{\beta}_1 \pm \boldsymbol{t}_{\alpha/2}\, \boldsymbol{se}\left(\hat{\beta}_1\right)$$

Si por ejemplo α es 0.05, la interpretación del intervalo de confianza para β_2 es: dado un nivel de confianza del 95% (en 95 de cada 100 casos) en el largo plazo, el intervalo $\left[\hat{\beta}_2 - t_{\alpha/2}\, se\left(\hat{\beta}_2\right), \hat{\beta}_2 + t_{\alpha/2}\, se\left(\hat{\beta}_2\right)\right]$ contendrá el verdadero valor de β_2.

4.7. Pruebas de hipótesis

En todo modelo de regresión se deben probar hipótesis para evaluar la validez estadística de los resultados. Entre la variedad de pruebas de hipótesis que se pueden efectuar, las pruebas de dependencia y relevancia son las más importantes.

Prueba de relevancia: la prueba de relevancia consiste en evaluar estadísticamente qué tan significativo es un parámetro del modelo, de esta manera puede identificarse si la variable independiente (X) aporta información importante al modelo de regresión. Siguiendo la estructura presentada en el capítulo 2, para cada estimador β_i , i= 1, 2:

Paso 1: Planteamiento de la hipótesis.

Ho: $\beta_i = 0$

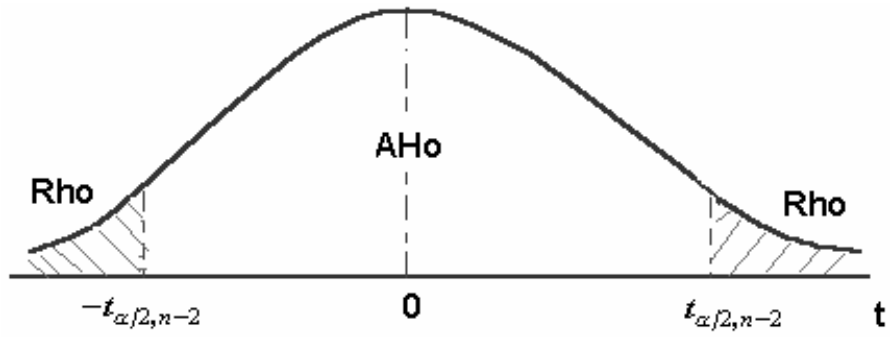
Ha: $\beta_i \neq 0$

Paso 2: Nivel de significancia (α):

Paso 3: El estadístico de prueba. Para la prueba de relevancia en el modelo de regresión, el estadístico de prueba se define como:

$$t_C = \frac{\beta_i}{se(\beta_i)} \sim \quad t_{\alpha/2,\, n-2}$$

Paso 4: Regiones de decisión: La siguiente gráfica muestra la regiones de rechazo y aceptación de la hipótesis nula.



Paso 5: Criterio de decisión y conclusión del investigador: Si $|t_C| > t_{\alpha/2,n-2}$ se rechaza la hipótesis nula. Si la prueba de hipótesis es realizada para β_1 y se rechaza Ho se concluye que el intercepto del modelo es significativo al nivel α. Si la prueba se efectúa para β_2 y se rechaza Ho se concluye que X_i es estadísticamente relevante al nivel α de significancia. Por otro lado, cuando no sea posible rechazar la hipótesis nula, se puede decir que no existe evidencia estadística para afirmar que X_i sea relevante al nivel α de significancia.

Prueba de dependencia: esta prueba se lleva a cabo para evaluar si en un modelo de regresión las variables independientes explican estadísticamente en su conjunto la variable dependiente. Se desea que en un modelo de regresión exista una alta dependencia ocasionada por las variables explicativas. Esta prueba de hipótesis como cualquier otra debe seguir una estructura similar a la presentada en el capítulo 2. La hipótesis nula de esta prueba hace referencia a la no existencia de dependencia en el modelo (para el caso de regresión simple como solo hay una variable independiente se desea probar si $\beta_2 = 0$). La hipótesis alternativa argumenta lo contrario, señalando que al menos uno de los coeficientes que acompañan las variables independientes es distinto de cero (en regresión simple esto es equivalente a $\beta_2 \neq 0$).

El estadístico de prueba para el caso de un modelo de regresión lineal simple es $F_C = (t_{n-2})^2 \sim F_{1,n-2}$, donde F_C es el estadístico calculado, que sigue una distribución *F* con un grado de libertad en el numerador y n-2 grados de libertad en el denominador; y t es el estadístico *t* calculado en la prueba de relevancia para β_2. Finalmente, la hipótesis nula es rechazada cuando $F_C > F_{1,n-2}$.

4.8. Predicción

Una aplicación del modelo de regresión es la predicción o el pronóstico de la variable dependiente, de acuerdo con valores dados de las variables independientes. Hay dos tipos de predicciones: la predicción media y la predicción individual. A continuación se presentan estos dos casos:

Predicción media: es la predicción del valor medio condicional de *Y*, correspondiente a un determinado valor de *X*, denotado como X_0, el cual representa un punto sobre la línea de regresión poblacional.

Si se desea predecir $E(Y / X_0)$, la estimación puntual de la predicción media es $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$ y la varianza de $\hat{Y}_0$: $Var(\hat{Y}_0) = \hat{\sigma}^2 \left[\frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$.

Predicción individual: es la predicción de un valor individual de *Y*, correspondiente a un determinado valor de *X*. Si se desea predecir Y_0 / X_0, de igual forma que en la predicción media, la estimación puntual es $\hat{Y}_0 = \hat{\beta}_1 + \hat{\beta}_2 X_0$, sin embargo la manera de calcular la varianza de Y_0 es:

$$Var(Y_0) = \hat{\sigma}^2 \left[1 + \frac{1}{n} + \frac{(X_0 - \bar{X})^2}{\sum (X_i - \bar{X})^2} \right]$$

4.9. El Coeficiente de Determinación

Es importante mencionar que cuando un modelo de regresión es construido con el objeto de predecir, al investigador le interesa encontrar una medida de la bondad de ajuste de los resultados del modelo. Una medida muy común de esta bondad de ajuste es el coeficiente de determinación o R^2, la cual proporciona información

respecto a que tan bien la línea de regresión muestral se ajusta a los datos. Para el caso de un modelo de regresión lineal simple se denota como r^2 y se calcula: $r^2 = (r)^2$, donde r es el coeficiente de correlación lineal entre las variables Y y X. Debido a que el r^2 bajo los supuestos de modelo de regresión clásico se encuentra entre 0 y 1, la manera de interpretarlo es en porcentaje, argumentándose que dicho valor refleja la magnitud porcentual de la variación de la variable *Y* explicada por la variable *X*.

4.10. Modelos de regresión simple no lineal en las variables

En algunos casos el investigador requiere estimar otro tipo de modelos en los que las variables independientes no sean lineales, como por ejemplo variables transformadas en términos logarítmicos, cuadráticos, raíz cuadrada, cúbicos, etc. Las razones para estimar estos nuevos modelos pueden ser: mejorar los resultados en términos de bondad de ajuste, obtener elasticidades directamente de la regresión, o en algunos casos porque la teoría económica lo sugiere. Un ejemplo del modelo no lineal es el conocido como Cobb-Douglas, cuya forma funcional es la siguiente:

$$Y_i = AX_i^{\beta_2} e^{u_i}$$

Para estimar el modelo se efectúa una linealización del modelo original transformándolo en logaritmos. De esta manera:

$$Log\, Y_i = LogA + \beta_2 Log\, X_i + u_i$$

Puede notarse que las variables dependiente e independiente se encuentran transformadas en logaritmos y el término $Log\, A$ es el intercepto de la regresión. Así, con el deseo de obtener los coeficientes de la regresión puede efectuarse la siguiente sustitución:

Sea $YT = Log\, Y_i$, $\beta_1 = Log\, A$ y $XT_i = Log\, X_i$, luego el modelo a estimar toma la forma: $YT_i = \beta_1 + \beta_2 XT_i + u_i$, y los coeficientes del modelo transformado pueden ser obtenidos por el método de mínimos cuadrados ordinarios usando las ecuaciones para los estimadores β_1 y β_2 presentadas en el numeral 4.4.

Teóricamente un modelo Cobb-Douglas es una función con elasticidad constante a lo largo de todo su dominio, siendo diferente de una función lineal, donde la elasticidad depende especialmente de la observación X_i. En este sentido, el modelo Cobb-Douglas permite obtener directamente las elasticidades: el coeficiente $\hat{\beta}_2$ representa la elasticidad de *Y* respecto a *X*, y se interpreta como el aumento (cuando el valor de la elasticidad es mayor que cero) o disminución (cuando el valor de la elasticidad es menor que cero) porcentual en la variable *Y*, ocasionada por el incremento en un 1% de la variable *X*.

4.11. Ejercicios de Computador

Continuando con el ejemplo de datos hipotéticos de demanda presentado en los capítulos anteriores, las siguientes salidas de computador muestran los resultados del modelo de regresión lineal simple de demanda y el modelo no lineal en las variables (doblemente logarítmico) con las respectivas matrices de varianza covarianza de los coeficientes:

MODELO DE REGRESION LINEAL SIMPLE

Dependent Variable: DX
Method: Least Squares
Date: 10/03/06 Time: 16:38
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	71.41791	3.130854	22.81100	0.0000
PX	-5.042910	0.454825	-11.08759	0.0000

R-squared	0.917870	Mean dependent var	40.38462
Adjusted R-squared	0.910404	S.D. dependent var	16.89940
S.E. of regression	5.058427	Akaike info criterion	6.220627
Sum squared resid	281.4646	Schwarz criterion	6.307542
Log likelihood	-38.43407	F-statistic	122.9346
Durbin-Watson stat	2.267643	Prob(F-statistic)	0.000000

Los resultados del modelo lineal muestran que la variable precio cuenta con el signo esperado y es relevante al 1%, 5% y 10% de significancia. El valor del R^2 es 0.918, es decir, el 92% de la variación de la demanda del bien X esta explicada por la variable precio. Adicionalmente se observa la existencia de dependencia conjunta en el modelo al 1%, 5% y 10% de significancia (F_c=122.935). El coeficiente de la variable PX es interpretado como un efecto marginal, por lo tanto, un incremento en una unidad del precio de X disminuye en promedio su demanda en 5.04 unidades, manteniendo todos los demás factores constantes.

MATRIZ DE VARIANZA COVARIANZA DE LOS COEFICIENTES DEL MODELO DE REGRESIÓN SIMPLE

COEFICIENTE	C	PX
C	9.802248	-1.273019
PX	-1.273019	0.206866

MODELO DE REGRESION SIMPLE NO LINEAL EN LAS VARIABLES (DOBLEMENTE LOGARITMICO)

Dependent Variable: LOG(DX)
Method: Least Squares
Date: 10/03/06 Time: 16:48
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.799536	0.208899	22.97538	0.0000
LOG(PX)	-0.722024	0.118417	-6.097307	0.0001

R-squared	0.771676	Mean dependent var	3.597486
Adjusted R-squared	0.750919	S.D. dependent var	0.499106
S.E. of regression	0.249094	Akaike info criterion	0.198665
Sum squared resid	0.682526	Schwarz criterion	0.285581
Log likelihood	0.708676	F-statistic	37.17715
Durbin-Watson stat	2.124425	Prob(F-statistic)	0.000078

Los resultados del modelo doblemente logarítmico indican que la variable logaritmo del precio es significativa (al 1%, 5% y 10%) y exhibe el signo teórico. El R^2 es 0.772, por lo tanto, el 77% de la variación del logaritmo de la demanda del bien X es explicada por el logaritmo de su precio. Adicionalmente existe dependencia conjunta en el modelo (1%, 5% y 10% de significancia). El coeficiente de la variable LOG(PX) es interpretado como una elasticidad, por lo tanto, un incremento en un 1% del precio de X disminuye en promedio su demanda en 0.72%, manteniendo todos los demás factores constantes.

MATRIZ DE VARIANZA COVARIANZA DE LOS COEFICIENTES DEL MODELO NO LINEAL EN LAS VARIABLES (DOBLEMENTE LOGARITMICO)

COEFICIENTE	**C**	**LOG(PX)**
C	0.043639	-0.023345
LOG(PX)	-0.023345	0.014023

5. REGRESION MULTIPLE LINEAL Y NO LINEAL

5.1. Expresión del modelo en forma matricial

En regresión múltiple se supone que las variaciones de Y_i que se pretenden explicar son debidas a "K" variables independientes, es decir $X_1, X_2, \ldots, X_K$ y como en la realidad no pueden presentarse relaciones determinísticas por completo se considera la inclusión del término de perturbación "ε".

Resulta conveniente analizar el modelo clásico de regresión lineal usando el enfoque matricial. Supóngase un modelo lineal de la forma:

$$\boldsymbol{Y} = \beta_1 + \beta_2 \boldsymbol{x}_2 + \beta_3 \boldsymbol{x}_3 + \cdots + \beta_k \boldsymbol{x}_k + \varepsilon$$

Si se tienen n observaciones independientes $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_n$ de Y, podemos escribir y_i como:

$$\boldsymbol{y_i} = \beta_1 + \beta_2 \boldsymbol{x_{i2}} + \beta_3 \boldsymbol{x_{i3}} + \cdots + \beta_k \boldsymbol{x_{ik}} + \varepsilon_i$$

Donde x_{ij} es el valor de la *j*-ésima variable independiente para la *i*-ésima observación, $i = 1,2,3,\ldots,n$. Ahora defínanse las siguientes matrices, con $x_1 = 1$:

$$\mathrm{Y} = \begin{bmatrix} \boldsymbol{y}_1 \\ \boldsymbol{y}_2 \\ \vdots \\ \boldsymbol{y}_n \end{bmatrix} \quad , \quad \mathrm{X} = \begin{bmatrix} \boldsymbol{x}_{11} & \boldsymbol{x}_{12} \cdots \boldsymbol{x}_{1k} \\ \boldsymbol{x}_{21} & \boldsymbol{x}_{22} \cdots \boldsymbol{x}_{2k} \\ \vdots & \vdots \\ \boldsymbol{x}_{n1} & \boldsymbol{x}_{n2} \cdots \boldsymbol{x}_{nk} \end{bmatrix} \quad , \quad \beta = \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} \quad , \quad \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

Por lo tanto las *n* ecuaciones que representan y_i como función de las x_{ij}, los β y ε_i se pueden escribir simultáneamente y de forma compacta como:

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$$

5.2. Supuestos del modelo

Los supuestos del modelo son los siguientes:

1. $\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ (Linealidad en los parámetros).
2. **X** es de tamaño $n \times k$ con rango *k*.
3. $E(\boldsymbol{\varepsilon}/\mathbf{X}) = 0 \quad \Rightarrow \quad E(\mathbf{Y}/\mathbf{X}) = \mathbf{X}\boldsymbol{\beta}$
4. $E(\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}') = \sigma^2\mathbf{I} \quad \Rightarrow \quad Cov(\varepsilon_i \varepsilon_j) = 0 \,, \forall\, i \neq j$.
5. **X** es no estocástica.
6. $(\varepsilon/\mathrm{X}) \sim N(0, \sigma^2\mathrm{I})$

5.3. Método de estimación de mínimos cuadrados ordinarios

Se desea obtener un estimador $\hat{\boldsymbol{\beta}}$ de un vector de parámetros desconocido $\boldsymbol{\beta}$ que minimiza la suma del cuadrado de los errores S, donde:

$$S = \sum \varepsilon^2 = \varepsilon'\varepsilon = (\mathrm{Y} - \mathrm{X}\hat{\beta})'(\mathrm{Y} - \mathrm{X}\hat{\beta})$$

Al minimizar *S* con respecto a $\hat{\boldsymbol{\beta}}$ se encuentra el estimador de mínimos cuadrados ordinarios de regresión múltiple:

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

5.4. Matriz de varianzas y covarianzas de los estimadores

La matriz de varianza-covarianza de los estimadores es relevante en la determinación de los errores estándar de los coeficientes y en la ejecución de pruebas de hipótesis. Para obtener la matriz de varianza-covarianza de los estimadores es necesario calcular previamente la suma de cuadrados de los errores y la varianza del modelo:

1. Suma de cuadrados de los errores. Puede ser calculada así: $SCE = \mathbf{Y'Y} - \boldsymbol{\beta}'\mathbf{X'Y}$.

2. *Varianza del modelo*. Dado que en la mayoría de los casos la varianza es desconocida, se utiliza la información de la muestra para obtener un estimador de la misma: $\hat{\sigma}^2 = (\mathbf{Y'Y} - \boldsymbol{\beta}'\mathbf{X'Y})/(n-k) = SCE/(n-k)$.

Usando la información anterior, la matriz de varianza covarianza de los coeficientes se puede calcular con la siguiente fórmula:

$$\textit{Matriz}\quad \mathbf{var-cov.} = \hat{\sigma}^2(\mathrm{X'X})^{-1}$$

5.5. Pruebas de hipótesis

Para efectuar pruebas de hipótesis es necesario obtener el error estándar de cada uno de los estimadores. Esta medida de dispersión corresponde a la raíz cuadrada de cada uno de los elementos de la diagonal principal de la matriz de varianza – covarianza. A continuación se presentan los aspectos más importantes para efectuar las pruebas de relevancia y dependencia en un modelo de regresión múltiple:

Pruebas de relevancia: En estas pruebas se utilizan los t estadísticos calculados de los estimadores con su respectivo p-valor. A continuación se presenta la forma de obtenerlos:

1. *t – estadísticos.* Los valores de *t* son calculados efectuando el cociente entre el coeficiente estimado y el error estándar respectivo.

2. *p-valores.* Arroja la probabilidad exacta de obtener un valor de *t* mayor que el valor absoluto de *t* obtenido para cada coeficiente. También es conocido como el nivel mínimo de significancia para rechazar la hipótesis nula. Para

obtener dicha probabilidad es necesario el valor del estadístico *t* calculado, el número de grados de libertad $(n-k)$ y el número de colas de la prueba (en este caso dos colas dado que es una prueba de significancia individual).

Prueba de dependencia: Como se mencionó en el capítulo anterior el estadístico utilizado para realizar la prueba es el *F*.

1. *F – estadístico.* Mide la dependencia conjunta en el modelo respecto a las variables explicativas. Puede ser obtenido en la forma matricial de la siguiente manera: $F = \left[(\boldsymbol{\beta}'\mathbf{X}'\mathbf{Y} - \mathbf{n}\overline{\mathbf{Y}}^{\mathbf{2}})(n-k)\right] / \left[(\mathbf{Y}'\mathbf{Y} - \boldsymbol{\beta}'\mathbf{X}'\mathbf{Y})(k-1)\right]$.

2. *p – valor*. Arroja el nivel mínimo de significancia para rechazar la hipótesis nula. En el procedimiento se requiere el valor obtenido de *F*, los grados de libertad del numerador $(k-1)$ y grados de libertad del denominador $(n-k)$.

5.6. Coeficiente de determinación ajustado $\left(\overline{R}^2\right)$

El término "ajustado" se refiere a que es corregido por los correspondientes grados de libertad. El $\overline{\boldsymbol{R}}^2$ mide la bondad de ajuste del modelo de regresión (porcentaje de explicación de la variable dependiente por las variables independientes), así como lo hace el $\boldsymbol{R}^2$ convencional, sin embargo el $\overline{\boldsymbol{R}}^2$ tiene la particularidad de que permite comparar modelos de regresión múltiple en los que se incluyen variables adicionales. No obstante, se debe considerar que la comparación tiene validez cuando en cada modelo la variable dependiente y el tamaño de la muestra sean iguales. La forma de calcular el $\overline{\boldsymbol{R}}^2$ se presenta a continuación:

$$\overline{\boldsymbol{R}}^2 = 1 - \left(1 - \boldsymbol{R}^2\right)\frac{\boldsymbol{n}-1}{\boldsymbol{n}-\boldsymbol{k}}$$

5.7. Intervalos de confianza.

Un intervalo de confianza para el parámetro β_k , $\forall\, k = 1,2,\ldots,K$, tiene la forma:

$$\Pr\left[\hat{\beta}_k - t_{\alpha/2}\, se\left(\hat{\beta}_k\right) \leq \beta_k \leq \hat{\beta}_k + t_{\alpha/2}\, se\left(\hat{\beta}_k\right)\right] = 1-\alpha$$

$$\hat{\beta}_k \pm t_{\alpha/2}\, se\left(\hat{\beta}_k\right)$$

donde α es el nivel de significancia estadística y $se\left(\hat{\beta}_k\right)$ es el error estándar de $\hat{\beta}_k$. El $se\left(\hat{\beta}_k\right)$ se obtiene mediante la fórmula: $se\left(\hat{\beta}_k\right) = \hat{\sigma}\sqrt{(X'X)^{-1}_{kk}}$. Puede notarse, que este intervalo de confianza corresponde a una expresión matemática similar a la presentada en el caso de regresión simple.

5.8. Modelos de regresión múltiple no lineal en las variables

En este numeral, se extenderá el caso de la función tipo Cobb-Douglas desarrollado en el numeral 4.10 al caso de regresión múltiple no lineal en las variables. Considérense más variables independientes X's que pueden explicar la variable Y, por lo tanto, el modelo Cobb-Douglas toma la forma:

$$Y_i = A X_{i2}^{\beta_2} X_{i3}^{\beta_3} \ldots X_{ik}^{\beta_k} e^{u_i}$$

Luego transformando el modelo en logaritmos:

$$Log\, Y_i = Log\, A + \beta_2 Log\, X_{i2} + \beta_3 Log\, X_{i3} + \cdots + \beta_k Log\, X_{ik} + u_i$$

Sea $YT = Log\, Y_i$, $\beta_1 = Log\, A$, $XT_{i2} = Log\, X_{i2}$,..., $XT_{ik} = Log\, X_{ik}$, entonces el modelo a estimar es:

$$YT_i = \beta_1 + \beta_2 XT_{i2} + \beta_3 XT_{i3} + \cdots + \beta_k XT_{ik} + u_i$$

Bajo el esquema matricial los coeficientes del modelo transformado pueden ser obtenidos a través del método de mínimos cuadrados ordinarios usando la fórmula de cálculo presentada en el numeral 5.3. El coeficiente $\hat{\beta}_k \ , \forall\, k = 2,3,\ldots,K$ representa la elasticidad de *Y* respecto a X_k y tiene la misma interpretación que en el caso del modelo de regresión simple doblemente logarítmico del capítulo anterior. Por lo tanto, se tendrán $\boldsymbol{k-1}$ elasticidades en regresión múltiple al estimarse una función tipo Cobb-Douglas. Por otro lado, cabe destacar que ejercicios de estimación diferentes al modelo Cobb-Douglas no permiten obtener directamente elasticidades constantes. Por ello es necesario tener en cuenta la forma que toman las variables en el modelo transformado antes de efectuar interpretaciones de los coeficientes.

5.9. Ejercicios de Computador.

Ejemplo 1.

Usando la misma base de datos hipotéticos de demanda de capítulos anteriores a continuación se presentan los resultados de las estimaciones del modelo de regresión múltiple lineal y no lineal en las variables, las matrices de varianza covarianza de los coeficientes, así como la comparación entre los valores observados y predichos de la demanda y sus residuos:

REGRESIÓN LINEAL MÚLTIPLE

Dependent Variable: DX
Method: Least Squares
Date: 10/04/06 Time: 10:31
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	72.73351	10.83288	6.714142	0.0002
PX	-4.588739	1.778623	-2.579939	0.0326
PW	-0.386460	1.214240	-0.318273	0.7584
PZ	0.471929	0.688867	0.685080	0.5127
I	-0.409388	0.201019	-2.036558	0.0761

R-squared	0.947105	Mean dependent var	40.38462
Adjusted R-squared	0.920657	S.D. dependent var	16.89940
S.E. of regression	4.760208	Akaike info criterion	6.242183
Sum squared resid	181.2767	Schwarz criterion	6.459471
Log likelihood	-35.57419	F-statistic	35.81046
Durbin-Watson stat	1.436480	Prob(F-statistic)	0.000037

Los resultados del modelo lineal muestran que la variable precio cuenta con el signo esperado y es relevante al 5% y 10% de significancia. El valor del R^2 es 0.947, es decir, el 95% de la variación de la demanda del bien X esta explicada por las variables independientes. Adicionalmente se observa la existencia de dependencia conjunta en el modelo al 1%, 5% y 10% de significancia (F_c=35.810). El coeficiente de la variable PX es interpretado como un efecto marginal, por lo tanto, un incremento en una unidad del precio de X disminuye en promedio su demanda en 4.59 unidades, manteniendo todos los demás factores constantes.

Vale la pena aclarar que la variable ingreso aún cuando es relevante al 10% de significancia, el signo de su coeficiente no es consistente con la teoría economía relacionada con un bien normal.

MATRIZ DE VARIANZAS Y COVARIANZAS DE LOS ESTIMADORES DEL MODELO DE REGRESIÓN LINEAL MULTIPLE

COEFICIENTE	C	PX	PW	PZ	I
C	117.3513	-11.56233	1.349073	-6.616478	-0.600013
PX	-11.56233	3.163501	-1.826424	0.777287	-0.022642
PW	1.349073	-1.826424	1.474379	-0.192430	0.052888
PZ	-6.616478	0.777287	-0.192430	0.474538	-0.014848
I	-0.600013	-0.022642	0.052888	-0.014848	0.040409

VALORES OBSERVADOS Y ESTIMADOS DE LA DEMANDA Y LOS RESIDUOS A PARTIR DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE

obs	Actual	Fitted	Residual	Residual Plot
1	37.0000	37.8104	-0.81044	| . *| . |
2	38.0000	43.2972	-5.29718	| *. | . |
3	18.0000	22.0098	-4.00977	| .* | . |
4	50.0000	49.7111	0.28890	| . * . |
5	22.0000	27.3714	-5.37143	| *. | . |
6	55.0000	59.4627	-4.46267	| * | . |
7	42.0000	34.4728	7.52721	| . | . *
8	29.0000	27.1267	1.87326	| . | * . |
9	63.0000	62.7036	0.29637	| . * . |
10	13.0000	10.3593	2.64072	| . | * . |
11	60.0000	56.3697	3.63029	| . | *. |
12	62.0000	59.7074	2.29265	| . | * . |
13	36.0000	34.5979	1.40211	| . | * . |

MODELO DE REGRESION MULTIPLE NO LINEAL EN LAS VARIABLES (DOBLEMENTE LOGARITMICO)

Dependent Variable: LOG(DX)
Method: Least Squares
Date: 10/04/06 Time: 10:39
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.042978	1.646616	1.848020	0.1018
LOG(PX)	0.115482	0.301528	0.382991	0.7117
LOG(PW)	-0.336166	0.436617	-0.769933	0.4635
LOG(PZ)	0.671764	0.400484	1.677382	0.1320
LOG(I)	-0.103431	0.084847	-1.219025	0.2576

R-squared	0.902417	Mean dependent var	3.597486
Adjusted R-squared	0.853626	S.D. dependent var	0.499106
S.E. of regression	0.190953	Akaike info criterion	-0.189861
Sum squared resid	0.291703	Schwarz criterion	0.027427
Log likelihood	6.234096	F-statistic	18.49542
Durbin-Watson stat	1.510399	Prob(F-statistic)	0.000418

Los resultados del modelo doblemente logaritmo no son satisfactorios, dado que ninguna de las variables incorporadas como regresores son significativas. Asimismo, las variable LOG(PX) y LOG(I) no presentan los signos esperados, limitando la validez teórica del modelo.

MATRIZ DE VARIANZAS Y COVARIANZAS DE LOS ESTIMADORES DEL MODELO DOBLEMMENTE LOGARÍTMICO

COEFICIENTE	C	LOG(PX)	LOG(PW)	LOG(PZ)	LOG(I)
C	2.711343	-0.229656	-0.565544	-0.638982	-0.061003
LOG(PX)	-0.229656	0.090919	-0.022785	0.064962	0.001807
LOG(PW)	-0.565544	-0.022785	0.190634	0.121451	0.014976
LOG(PZ)	-0.638982	0.064962	0.121451	0.160387	0.008270
LOG(I)	-0.061003	0.001807	0.014976	0.008270	0.007199

VALORES OBSERVADOS Y ESTIMADOS DEL LOGARITMO DE LA DEMANDA Y LOS RESIDUOS A PARTIR DEL MODELO DE REGRESIÓN DOBLEMENTE LOGARÍTMICO

obs	Actual	Fitted	Residual	Residual Plot
1	3.61092	3.50939	0.10153	| . | * . |
2	3.63759	3.80097	-0.16339	| .* | . |
3	2.89037	3.07102	-0.18064	| .* | . |
4	3.91202	3.91411	-0.00209	| . * . |
5	3.09104	3.11501	-0.02396	| . *| . |
6	4.00733	4.10808	-0.10075	| . * | . |
7	3.73767	3.59355	0.14412	| . | * . |
8	3.36730	3.32110	0.04619	| . | * . |
9	4.14313	4.38551	-0.24237	| * . | . |
10	2.56495	2.69990	-0.13495	| . * | . |
11	4.09434	3.84781	0.24653	| . | . *
12	4.12713	4.00914	0.11799	| . | * . |
13	3.58352	3.39174	0.19178	| . | * |

Ejemplo 2.

Ahora considere la siguiente información de una firma sobre los costos de producción y la cantidad producida de un bien para estimar una función de costos cúbica:

TABLA No. 3. COSTOS SEGÚN EL NIVEL DE PRODUCCIÓN

Obs.	Q	CT
1	0	5
2	1	14
3	2	23
4	3	28
5	4	33
6	5	36
7	6	41
8	7	45
9	8	48
10	9	50
11	10	55
12	11	61
13	12	66
14	13	72
15	14	77
16	15	86
17	16	97
18	17	110
19	18	127
20	19	147
21	20	169

Donde:

CT: Costo total de producción
Q: Nivel de producto

ESTADISTICAS DESCRIPTIVAS

	Q	Q2	Q3	CT
Mean	10	136.6667	2100	66.19048
Median	10	100	1000	55
Maximum	20	400	8000	169
Minimum	0	0	0	5
Std. Dev.	6.204837	128.5365	2488.431	43.49899
Observations	21	21	21	21

MODELO DE REGRESION MULTIPLE NO LINEAL EN LAS VARIABLES (FUNCIÓN CUBICA)

Dependent Variable: CT
Method: Least Squares
Date: 27/09/06 Time: 21:48
Sample: 1 21
Included observations: 21

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	4.586862	0.962089	4.767605	0.000200
Q	10.450570	0.427013	24.473660	0.000000
Q2	-0.974658	0.050336	-19.363000	0.000000
Q3	0.043001	0.001653	26.020450	0.000000
R-squared	0.999236	Mean dependent var		66.19048
Adjusted R-squared	0.999101	S.D. dependent var		43.49899
S.E. of regression	1.304364	Akaike info criterion		3.538952
Sum squared resid	28.92322	Schwarz criterion		3.737908
Log likelihood	-33.15899	F-statistic		7408,618
Durbin-Watson stat	0.882959	Prob(F-statistic)		0

MATRIZ DE VARIANZAS Y COVARIANZAS DE LOS ESTIMADORES

COEFICIENTE	C	Q	Q2	Q3
C	0.925616	-0.336905	0.032823	-0.000934
Q	-0.336905	0.182340	-0.020705	0.000640
Q2	0.032823	-0.020705	0.002534	-0.000082
Q3	-0.000934	0.000640	-0.000082	0.000003

6. INCUMPLIMIENTO DE LOS SUPUESTOS DEL MODELO

El cumplimiento de los supuestos del modelo clásico de regresión garantiza que los $\hat{\beta}_k$ obtenidos a través del método de mínimos cuadrados ordinarios sean los mejores estimadores linePÑales insesgados. Cuando tales supuestos son violados, se empiezan a generar problemas en los resultados de la regresión, haciendo que los parámetros obtenidos no cumplan con algunas de las propiedades deseables de un estimador (eficiencia y consistencia). A continuación se describen de manera general los conceptos de multicolinealidad, heteroscedaticidad autocorrelación, y no normalidad, la forma de detectar tales problemas en el modelo estimado y las posibles soluciones a la violación de los supuestos de mínimos cuadrados ordinarios relacionados con estos conceptos.

6.1. Multicolinealidad

La multicolinealidad tiene que ver con la relación lineal entre algún conjunto de variables independientes en un modelo de regresión. Supóngase el siguiente modelo con cuatro variables independientes:

$$Y_t = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_4 + \beta_5 X_5 + \varepsilon_t$$

Cualquier relación lineal entre las variables independientes de este modelo, por ejemplo X_2 con X_3, o X_2 con X5 y X_4 puede generar problemas de multicolinealidad. Por lo general, existen dos tipos de multicolinealidad:

1. Multicolinealidad Perfecta: Para entender el concepto de multicolinealidad perfecta es necesario expresar las variables independientes del modelo en términos de una combinación lineal cuya suma algebraica sea igual a cero. Para el modelo presentado la combinación lineal sería:

$$\lambda_1 X_2 + \lambda_2 X_3 + \lambda_3 X_4 + \lambda_4 X_5 = 0 \ \forall \text{ las T observaciones}$$

Los valores de λ pueden ser positivos o negativos y formar muchas combinaciones. Cuando la suma algebraica para todas las observaciones de la muestra de esta combinación lineal es cero se dice que existe multicolinealidad perfecta. De este caso se exceptúa que simultáneamente los valores de λ sean cero, pues esta es una solución trivial de la ecuación. En otras palabras, la multicolinealidad perfecta se presenta cuando una combinación lineal de uno o más vectores de variables explicativas generan de manera perfecta uno o más vectores idénticos a cualquiera de las variables explicativas en la base de datos.

2. Multicolinealidad Alta: Esta se presenta cuando la colinealidad que existe entre variables independientes es muy fuerte aunque no perfecta.

La multicolinealidad se presenta debido a la tendencia definida de ciertas variables a lo largo de la muestra o a través del tiempo. Tendencias o patrones de comportamiento similares de las variables independientes en un modelo de regresión sustentan la multicolinealidad. La multicolinealidad se puede presentar en datos provenientes de series de tiempo. Por ejemplo, es común encontrarla al regresar variables que tienen que ver con la representación de ciclos económicos. Por ello, antes de efectuar la regresión es útil elaborar diagramas de dispersión entre las variables independientes con el objetivo de analizar el comportamiento tendencial de estas.

El problema de multicolinealidad es un problema ocasionado por las observaciones en los datos recopilados de la muestra. La presencia de multicolinealidad afecta directamente la estimación de los parámetros del modelo.

De acuerdo con el estimador por mínimos cuadrados ordinarios:

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

Si existe multicolinealidad perfecta entre las variables independientes de un modelo de regresión, $(X'X)^{-1}$ no existe. Cuando esto ocurre no es posible estimar $\hat{\beta}$. En presencia de alta multicolinealidad se genera una ampliación del error estándar de $\hat{\beta}$, por lo que el valor de los estadísticos "t" para cada uno de los parámetros del modelo serán mucho menores que en ausencia de multicolinealidad, aumentándose la probabilidad de cometer error de tipo II, es decir, que acepte *Ho* no siendo verdadera. Por consiguiente, el modelo no tiene validez para realizar pruebas de relevancia.

6.1.1. Detección de Multicolinealidad

La detección de multicolinealidad en un modelo puede hacerse por medio de la visualización de contradicciones en los estadísticos que juzgan la bondad del ajuste (R_2), dependencia (F_c) y los estadísticos que permiten evaluar la relevancia de las variables en el modelo (t_c). Otro método de detección es la estimación de $|X'X|$; si el valor obtenido de $|X'X|$ es muy cercano a cero, puede concluirse que es muy probable la existencia de multicolinealidad alta.

No obstante, se encuentran otras pruebas mucho más formales en términos estadísticos. Una de ellas es estimar coeficientes de correlación entre pares de variables independientes y formular pruebas de hipótesis sobre los coeficientes de correlación estimados para comprobar la significancia de la relación lineal en términos estadísticos. Por ejemplo, una vez calculado el coeficiente de correlación lineal entre X_2 y X_3, puede proponerse la siguiente prueba de hipótesis cuya formulación es idéntica a la presentada en el capítulo 2:

Ho: $\rho_{X_2,X_3} = 0$ (No existe relación lineal entre X2 y X3)

Ho: $\rho_{X_2,X_3} \neq 0$ (Si existe relación lineal entre X2 y X3)

El estadístico de prueba es:

$$t_C = \frac{\left(r_{X2,X3}\sqrt{n-2}\right)-\theta}{\sqrt{1-\left(r_{X2,X3}\right)^2}} \sim \quad t_{\alpha/2,n-2}$$

Donde θ es el valor que se desea probar del coeficiente de correlación lineal poblacional. No obstante en la mayoría de los casos este se asume cero con lo cual solo se desea verificar si hay o no correlación entre las variables explicativas.

Si $|t_C| > t_{\alpha/2,n-2}$ a un nivel α de significancia determinado, se rechaza Ho, confirmando la existencia de relación lineal entre X_2 y X_3, es decir el modelo de regresión mostrará multicolinealidad.

El otro método formal consiste en la estimación de regresiones auxiliares que ayudan a evaluar la relación lineal existente entre un conjunto de variables independientes. Para ello, se ejecuta una regresión entre las variables independientes del modelo, por ejemplo X_2 versus (X_3, X_4, X_4, X_5) y luego se analizan los estadísticos resultantes de esta. Si hay relación lineal entre estas variables, el R^2, el F_c y el t_c que acompaña a cada variable independiente de la regresión auxiliar serán altos. Las pruebas de hipótesis sobre relevancia y dependencia estadística en la regresión auxiliar determinan si existe o no multicolinealidad. Es importante tener en cuenta que deben estimarse todas las posibles regresiones auxiliares resultantes de las combinaciones entre las variables independientes o regresores del modelo original. El método de regresiones auxiliares es el más utilizado y recomendado por su sustentación estadística dado que permite evaluar la multicolinealidad ocasionada simultáneamente por la relación lineal entre más de dos variables independientes.

6.1.2. Corrección de Multicolinealidad

La corrección de multicolinealidad en un modelo puede ejecutarse mediante varios métodos:

1. Eliminación de Variables: Esta técnica propone la eliminación de una de las variables independientes relacionadas linealmente. El problema de aplicar esta técnica es que se pueden eliminar variables importantes que teóricamente explican la variable dependiente, presentándose posiblemente sesgo de especificación por omisión de variables.

2. Utilización de Información *a priori*: La información *a priori* comúnmente proviene de estudios anteriores que pueden brindar algún indicio sobre el valor de algún parámetro correspondiente a una de las variables independientes incluida en la ecuación de regresión. Operativamente, el valor *a priori* del parámetro es reemplazado en el modelo original. Luego se procederá a estimar el modelo resultante.

3. Transformación de Variables: Esta técnica plantea una transformación de las variables del modelo original. El más conocido es la transformación en primeras diferencias. Al trabajar con un modelo que incluye datos organizados en series de tiempo se presenta la posibilidad de construir una ecuación de primeras diferencias, asumiendo que con un rezago de cada una de las variables del modelo es posible eliminar la relación lineal que puede existir entre las variables independientes. El modelo original en el periodo t:

$$Y_t = \beta_1 + \beta_2 X_{2t} + \beta_3 X_{3t} + \beta_4 X_{4t} + \beta_5 X_{5t} + \varepsilon_t$$

Luego la ecuación en diferencias es:

$$Y_t - Y_{t-1} = \beta_2 (X_2 - X_{2t-1}) + \beta_3 (X_3 - {}_3X_{3t-1}) + \beta_4 (X_4 - X_{4t-1}) + \beta_5 (X_5 - X_{5t-1}) + \varepsilon_t^*$$

Donde $\varepsilon_t^* = \varepsilon_t - \varepsilon_{t-1}$. Debe tenerse en cuenta que al estimar este nuevo modelo, la interpretación de los coeficientes estimados no es la misma que en el modelo original, debido a que estos ahora representan cambios o diferencias de las variables entre los periodos t y t-1.

4. Aumentar el tamaño de la muestra: Este método consiste en ampliar la muestra o conjunto de datos utilizados para estimar el modelo. Esta es una solución plausible dado que el problema de multicolinealidad es ocasionado fundamentalmente por las observaciones en la muestra. Cuando se incrementa el número de observaciones se piensa que es más difícil reproducir el componente de colinealidad entre los regresores. Sin embargo, en muchos casos no es posible adquirir más información u observaciones de las variables debido a restricciones físicas, técnicas y económicas.

Finalmente, se recomienda que el investigador una vez utilice alguno de estos métodos verifique si el problema de multicolinealidad fue corregido.

6.2. Heteroscedasticidad

El problema de heteroscedasticidad se presenta cuando es violado el supuesto de varianza constante de los errores de la función de regresión. La heteroscedasticidad tiene que ver con la relación entre una o más de las variables independientes del modelo y el cuadrado de los errores estimados a partir de la regresión. Este problema se manifiesta en un crecimiento o decrecimiento de la varianza del modelo.

La presencia de heteroscedasticidad es muy común en regresiones estimadas a partir de datos de corte transversal. Por ejemplo, cuando se recolectan datos provenientes de estratos, de regiones, por tamaño de la familia o por tipo de empresa. En general, puede presentarse en estudios que incluyen grupos con comportamientos marcados a lo largo de toda la muestra; por ejemplo la variable ingreso monetario del hogar según el estrato, pues se puede pensar que la varianza del ingreso monetario del grupo de alta riqueza es más alta que la del grupo de escasos recursos.

El problema de heteroscedasticidad repercute directamente sobre la estimación de los parámetros de la regresión. Los estimadores seguirán siendo insesgados y consistentes pero no eficientes. La heteroscedasticidad causa la subestimación o sobre estimación de la varianza del modelo de regresión, por lo tanto el valor del error estándar de los parámetros, el valor de los estadísticos t y los intervalos de confianza cambian con respecto a los resultados que deberían obtenerse en ausencia de heteroscedasticidad. En este sentido, la presencia de heteroscedasticidad en el modelo de regresión hace que las pruebas de hipótesis no tengan validez estadística o que las inferencias sean erróneas.

6.2.1. Detección de la heteroscedasticidad

A continuación se presentan los métodos para detectar la existencia de heteroscedaticidad:

1. Análisis de residuales: Este método permite evaluar gráficamente si existe heteroscedasticidad causada por una variable independiente en particular o por todo el conjunto de variables independientes. Para el primer caso se elabora un diagrama de dispersión entre X_t y e_t^2 (cuadrado del término de error) donde X_t es el regresor que el investigador supone genera la heteroscedaticidad. En el segundo caso, se construye el diagrama de

dispersión entre Y_t estimado y e_t^2. Si estas gráficas muestran alguna tendencia específica, puede afirmarse que existe heteroscedasticidad en el modelo de regresión. No obstante esta metodología es indicativa y no esta basada en una prueba estadística.

2. Análisis de regresión: Es la utilización de una o más regresiones auxiliares. El procedimiento es similar al planteado para detectar multicolinealidad, con la salvedad de que ahora la regresión no se estima entre las variables independientes, sino entre el cuadrado del término de error y el conjunto de regresores del modelo original. Dentro de este método se encuentran las pruebas de Park, White, Glejser, Breusch-Pagan-Godfrey, y Golfeld – Quandt. A continuación se presenta el procedimiento general para efectuar la prueba de White:

   Si se tiene el siguiente modelo original:

$$Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + \varepsilon_t$$

   Una vez estimado el modelo por el método de mínimos cuadrados ordinarios (MCO), el investigador debe calcular el cuadrado de los errores: $\varepsilon_t^2 = \left(\mathbf{Y_t} - \hat{\mathbf{Y}}_\mathbf{t}\right)^2$, y luego estimar por MCO el siguiente modelo:

$$\varepsilon_t^2 = \alpha_0 + \alpha_1 X_{1t} + \alpha_2 X_{2t} + \alpha_3 X_{1t}^{\ 2} + \alpha_4 X_{2t}^{\ 2} + \alpha_5 X_{1t} X_{2t} + \nu_t$$

   La prueba de hipótesis relacionada con el modelo anterior es:

   Ho: $\alpha_1 = \alpha_2 = \alpha_3 = \alpha_4 = \alpha_5 = 0$ (No hay heteroscedasticidad)

   Ha: $\alpha_1 \neq \alpha_2 \neq \alpha_3 \neq \alpha_4 \neq \alpha_5 \neq 0$ (Si hay heteroscedasticidad)

El estadístico de prueba es $nR^2 \sim \chi_5^2$. En este caso el número de grados de libertad es cinco, que corresponde al número de variables explicativas en la regresión de White. Asimismo, para modelos con más variables explicativas los grados de libertad serán equivalentes al número de regresores en el modelo auxiliar. Si $nR^2 > \chi_{g.l}^2$ a un nivel de significancia α, la hipótesis nula es rechazada, por lo tanto, existe heteroscedasticidad en el modelo original.

Es importante señalar que la prueba de White desarrollada se refiere exclusivamente a la prueba de términos cruzados debido a que incorpora en la regresión auxiliar el término de interacción de las variables independientes del modelo original: $\alpha_5 X_{1t} X_{2t}$. Cuando este componente no es agregado la prueba recibe el nombre de prueba de White sin términos cruzados. Este cambio tiene un efecto directo sobre los grados de libertad de la prueba.

6.2.2. Corrección de heteroscedasticidad

Las medidas correctivas principalmente incluyen dos enfoques: cuando σ^2 es conocida y cuando σ^2 es desconocida.

1. Cuando se conoce σ^2. En este caso se utiliza el método de mínimos cuadrados ponderados (M.C.P) para realizar una transformación de las variables del modelo. Considere el modelo original el cual presenta heteroscedasticidad y σ^2 es conocida:

$$Y_t = \beta_1 + \beta_2 X_t + \varepsilon_t$$

Este método supone la siguiente transformación:

$$Y_t/\sigma = \beta_1/\sigma + \beta_2\, X_t/\sigma + \varepsilon_t/\sigma$$

Donde σ^2 es la desviación estándar del modelo. Se supone que esta transformación permite que el modelo quede libre de heteroscedasticidad. No obstante, para asegurarse de esto puede efectuarse cualquiera de las pruebas de detección presentadas anteriormente.

2. Cuando no se conoce σ^2: Por lo regular es muy difícil tener conocimiento previo de σ^2. Para utilizar el método de mínimos cuadrados ponderados debe recurrirse a supuestos ad hoc, con cierto grado de razonabilidad sobre σ^2 para proceder a la transformación de la regresión original, de tal manera, que el nuevo modelo cumpla con el supuesto de homocedasticidad. Considérese el siguiente modelo:

$$Y_t = \beta_1 + \beta_2 X_t + U_t$$

El investigador presume que la varianza de los errores tiene la siguiente forma:

$$E\left(U_t^{\,2}\right) = \sigma^2 X_t^2$$

Esta expresión es planteada cuando se cree que la varianza de los errores es proporcional al cuadrado de la variable explicativa. Bajo este supuesto el modelo transformado puede presentarse como sigue:

$$Y_t/X_t = \beta_1(1/X_t) + \beta_2 + v_t$$

Donde $v_t = \dfrac{U_t}{X_t}$. Puede verificarse que:

$$E\left(v_t\right) = E\left(U_t/X_t\right) = \left(1/X_t\right)E\left(U_t\right) = 0$$

y que el modelo transformado ahora es teóricamente homocedástico:

$$E\left(v_t^{\,2}\right) = E\left[\left(U_t / X_t\right)^2\right] = \left(1/X_t^2\right) E\left(U_t^{\,2}\right) = \left(1/X_t^2\right)\sigma^2 X_t^2 = \sigma^2 .$$

El método indica que las observaciones de la muestra deben dividirse por la raíz cuadrada de la estructura generadora de la heteroscedasticidad; lo cual para este ejemplo es equivalente a dividir por X_t. Luego el procedimiento indica que el modelo transformado requiere estimarse por MCO. Esta es la razón por la cual el método se denomina mínimos cuadrados ponderados, dado que se ponderan las observaciones originales por un factor. Es conveniente verificar empíricamente si luego de estimar el modelo transformado el problema de heteroscedasticidad fue corregido.

6.3. Autocorrelación

El problema de autocorrelación se presenta en una regresión cuando los errores de las diferentes observaciones están relacionados en el tiempo. Esto indica que el efecto de los errores en el tiempo no es instantáneo sino por el contrario es persistente en el tiempo. La autocorrelación es más común en series ordenadas en el tiempo que en información proveniente de encuestas en un tiempo fijo (sección cruzada). La autocorrelación puede estar relacionada con los ciclos económicos; generalmente ésta se presenta en un modelo con variables macroeconómicas donde en el tiempo ocurre un evidente comportamiento tendencial.

Otra causa de la autocorrelación es la presencia de sesgo de especificación en el modelo; principalmente por omisión de variables importantes, las cuales pasan a formar parte del error de la regresión. La autocorrelación puede ser también

generada en casos donde se usa una forma funcional incorrecta del modelo, esto hace que los datos se ajusten a una forma funcional que no es la más adecuada.

Se argumenta, que la manipulación de información puede llegar a generar también autocorrelación. Un caso típico se presenta en la cuentas nacionales, donde muchos datos son obtenidos a partir de otros, aplicando técnicas de interpolación o extrapolación. Por ejemplo, cuando se convierten datos diarios a semanales. Finalmente, modelos especiales como los de rezagos distribuidos y los autoregresivos pueden originar autocorrelación.

Entre las consecuencias de la autocorrelación se tiene la sobreestimación o subestimación de los estadísticos "t" que juzgan la significancia de las variables independientes en el modelo. Aunque los estimadores siguen siendo insesgados y consistentes son ineficientes. En este sentido se afecta la validez estadística de las pruebas de hipótesis.

6.3.1. Detección de la autocorrelación

Los métodos más comunes para detectar autocorrelación son:

1. Análisis de residuales: este método plantea la construcción de diagramas de dispersión para los errores en función de tiempo o en función de un período inmediatamente anterior. El primer paso es estimar el modelo original por MCO. Luego los errores estimados de la regresión son graficados en un eje de coordenadas para identificar si existe alguna tendencia de los mismos en el tiempo, o de estos con su primer rezago.
2. El estadístico de Durbin – Watson (**d**): Esta prueba es válida para aplicar en errores que se modelan como un proceso autoregresivo de orden 1 "AR(1)" como el mostrado a continuación:

$$\varepsilon_t = \rho\varepsilon_{t-1} + v_t$$

El estadístico "**d**" oscila entre 0 y 4. Si este se aproxima a 0, se dice que existe autocorrelación positiva (relación directa entre los errores), por el contrario si **d** se aproxima a 4, existe autocorrelación negativa (relación inversa entre los errores). El Durbin-Watson (**d**) se estima de la siguiente manera:

$$d = \frac{\sum_{t=2}^{n}(\varepsilon_t - \varepsilon_{t-1})}{\sum_{t=1}^{n}\varepsilon_t^2} = 2(1-\hat{\rho})$$

Donde $\hat{\rho}$ es el coeficiente de autocorrelación de orden 1, el cual puede despejarse directamente **d**:

$$\hat{\rho} = 1 - \frac{d}{2}$$

La hipótesis planteada es:

Ho: $\rho_{\varepsilon_t,\varepsilon_{t-1}} = 0$ (no existe autocorrelación entre los errores)

Ha: $\rho_{\varepsilon_t,\varepsilon_{t-1}} \neq 0$ (hay autocorrelación entre los errores)

El estadístico Durbin- Watson puede ser comparado con su respectivo tabulado, teniendo en cuenta el número de observaciones contenidas en la muestra y el número de regresores. Se debe tener en cuenta que d es utilizado para identificar solo autocorrelación de orden 1 y siempre y cuando el modelo tenga intercepto. Además no puede usarse en el caso de modelos autorregresivos.

Prueba de Breusch-Godfrey. Esta es una prueba similar a la prueba de White. Se diferencia de esta en que la variable dependiente de la regresión

auxiliar es el término de error ε_t y los regresores sus respectivos rezagos hasta el orden deseado por el investigador. Adicionalmente son incluidos los regresores usados en el modelo original. La hipótesis nula corresponde a que todos los coeficientes de autocorrelación de orden (los coeficientes que acompañan a los residuos rezagados en la regresión auxiliar) son iguales a cero, mientras la hipótesis alterna es que al menos uno de ellos es distinto de cero. El estadístico de prueba es $(n-s)R^2 \sim \chi_s^2$, donde s es el número de errores rezagados en la regresión auxiliar. Para probar autocorrelación de orden uno, que es la práctica más común, s será igual a uno. La hipótesis nula es rechazada cuando $(n-s)R^2 > \chi_s^2$ a un nivel de significancia α; en este caso se concluye que hay autocorrelación.

6.3.2. Corrección de la autocorrelación

La corrección del problema de autocorrelación incluye diferentes técnicas que persiguen principalmente la transformación de las variables del modelo con el objetivo de eliminar el patrón tendencial que siguen los errores. Se tienen dos tipos de metodologías de corrección de la autocorrelación:

1. Cuando se conoce el coeficiente de autocorrelación: la transformación recomendada sugiere rezagar un período las variables del modelo y estimar una ecuación de primeras diferencias. Para esto el modelo original debe ser transformado hasta tomar la forma:

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1 (X_{1t} - \rho X_{1t-1}) + \beta_2 (X_{2t} - \rho X_{2t-1}) + U_t^*$$

Esta ecuación es estimada y se propone cualquiera de las técnicas de detección de autocorrelación para averiguar si el problema de autocorrelación fue corregido. Dicha ecuación se conoce como ecuación en diferencias generalizada y consiste en un caso particular del método de mínimos cuadrados generalizados.

2. Cuando no se conoce el coeficiente de autocorrelación: En la mayoría de los casos a nivel empírico el coeficiente de autocorrelación no se conoce. Debido a esto el coeficiente de autocorrelación debe ser estimado partiendo de la suposición de un valor inicial del mismo.

   Una de estos métodos es el procedimiento Cochrane – Occurt: este consiste en la estimación de modelos con sucesivas transformaciones. Es un método iterativo representado en un algoritmo que evalúa durante el proceso la tendencia que sigue el ρ estimado de regresiones sucesivas. Cuando la diferencia de ρ entre un modelo estimado actual y su antecesor es 0.01 se afirma que el coeficiente ρ ha convergido y por consiguiente la tendencia de crecimiento de este se ha eliminado.

   Por otro lado existe el método de corrección a través del Durbin– Watson. Mediante esta técnica, aunque no se conoce ρ , este es posible estimarlo a partir del estadístico "**d**" de la regresión del modelo original. Una vez obtenido el valor de ρ las variables son transformadas para posteriormente estimar la siguiente ecuación de primeras diferencias:

$$Y_t - \rho Y_{t-1} = \beta_0 (1 - \rho) + \beta_1(X_{1t} - \rho X_{1t-1}) + \beta_2(X_{2t} - \rho X_{2t-1}) + U_t^*$$

   Después de aplicar alguno de estos métodos es necesario evaluar de nuevo la presencia de autocorrelación.

6.4. Error de especificación

Uno de los supuestos del modelo clásico de regresión lineal es que el modelo se encuentra bien especificado, es decir que su forma funcional y las variables que lo componen representan la formulación correcta. La teoría económica y algunas

medidas empíricas son útiles para probar si un modelo cuenta con error de especificación.

Existen cuatro tipos de fuentes o razones que generan error de especificación:

1. *Omisión de una variable relevante en el modelo*. Si una variable que afecta de manera importante la variable dependiente del modelo es omitida, se incurre en error de especificación. Esta situación hace que los estimadores sean sesgados.

2. *Inclusión de una variable irrelevante*. En algunos casos los investigadores en su proceso de exploración con el deseo de encontrar un mejor modelo que se ajuste a los datos, incorporan variables explicativas adicionales. Si estas no afectan significativamente a la variable dependiente se comete error de especificación. No obstante, los efectos sobre los coeficientes estimados son menos fuertes que en el caso de la omisión de una variable relevante; los estimadores serán insesgados, pero se obtienen de manera imprecisa. De esta manera, incluir variables irrelevantes afecta los errores estándares de los coeficientes, haciendo que los intervalos de confianza sean más anchos.

   Uso de una forma funcional inadecuada. Consiste en presentar un modelo matemático incorrecto o muy distante del comportamiento de los datos. Por ejemplo, plantear un modelo lineal en las variables cuando los datos se ajustan mejor en realidad a un modelo cuadrático, recíproco o a otra especificación funcional. Una medida empírica para verificar la existencia de una forma funcional inadecuada es la prueba RESET de Ramsey.

3. *Error de medición.* Cuando el valor de las observaciones que se tienen en una muestra no es el real o verdadero, los datos cuentan con error de

medición. Si el error se presenta en la variable dependiente como en la independiente, los estimadores de mínimos cuadrados serán sesgados.

6.4.1. Detención de la forma funcional inadecuada

A continuación se desarrolla el método de detección de una forma funcional inadecuada del modelo mediante la prueba RESET de Ramsey. El caso más simple de la prueba es el siguiente:

Considere el modelo $Y_i = \beta_1 + \beta_2 X_i + \varepsilon_i$, al cual se le desea probar si la forma funcional propuesta es o no inadecuada. Para realizar prueba de Ramsey se estima el modelo original y se predice la variable dependiente $\hat{Y}_i$. Luego se efectúa una regresión auxiliar en la que al modelo original se adicionan los términos $\hat{Y}_i$ cuadráticos o de orden superior, dependiendo de la posible relación que exista entre $\hat{\varepsilon}_i$ y $\hat{Y}_i$. Una gráfica entre $\hat{\varepsilon}_i$ y $\hat{Y}_i$ puede ser útil para identificar los regresores a incluirse. Un ejemplo de la regresión auxiliar es:

$$Y_i = \gamma_1 + \gamma_2 X_i + \gamma_3 \hat{Y}_i^2 + \gamma_4 \hat{Y}_i^3 + v_i$$

Posteriormente se utiliza el siguiente estadístico de prueba:

$$F = \frac{\left(R^2{}_{aux} - R^2{}_o\right)/j}{\left(1 - R^2{}_{aux}\right)/(n-g)} \quad \simeq f_{j,n-g}$$

donde $R^2{}_{aux}$ es el coeficiente de determinación de la regresión auxiliar , $R^2{}_o$ es el coeficiente de determinación del modelo original, j y g son el número de términos $\hat{Y}_i$ y parámetros incluidos en la regresión auxiliar, respectivamente. La hipótesis nula de prueba es que el modelo original esta bien especificado, mientras la

hipótesis alterna afirma lo contrario. Si $F > f_{j,n-g}$ aun nivel α de significancia se concluye que el modelo original esta inadecuadamente especificado.

6.4.2. Corrección de la forma funcional inadecuada

Se pueden mencionar dos mecanismos para encontrar la forma funcional adecuada de un modelo de regresión: el enfoque teórico y el enfoque empírico. El primero hace alusión a revisar nuevamente la teoría económica y a consultar la literatura reciente relacionada con el área de estudio. De este análisis la forma estructural como las variables independientes pueden relacionarse con la variable independiente puede ser ajustada. El segundo consiste en realizar estimaciones del modelo bajo distintas formas funcionales sin desligarse de los fundamentos de la teoría económica hasta aceptar la hipótesis de que el modelo este bien especificado.

6.5. No Normalidad de los errores

Uno de los supuestos claves en el modelo de regresión que permite desarrollar pruebas hipótesis basadas en los estadísticos F y T, es la normalidad de los errores. Si los residuos del modelo no siguen distribución normal se restringe la validez estadística de las pruebas.

6.5.1. Detención de la no normalidad de los errores

En este documento se citan de manera general dos formas de detectar si los residuos del modelo de regresión siguen o no distribución normal:

1. El Histograma de los residuos. Consiste en la construcción de un histograma para los errores estimados del modelo y observar si su polígono suavizado de frecuencias se parece en forma aproximada una distribución normal.

2. Prueba de Normalidad Jarque – Bera. Es una prueba para muestras grandes, basada en los residuos de mínimos cuadrados ordinarios. Requiere calcular la asimetría y curtosis de los residuos.

   Ho: los errores siguen distribución normal
   H1: los errores no siguen distribución normal

   El estadístico de prueba es:

$$JB = n\left[\frac{A^2}{6} + \frac{(K-3)^2}{24}\right] \quad \sim \quad \chi^2_{2\,gl}$$

   Ho es rechazada si JB > $JB > \chi^2_{,2\,gl}$ a un nivel α de significancia. A se denomina asimetría y K curtósis. Estas son medidas descriptivas de una variable aleatoria que hacen referencia al sesgo de la distribución y el apuntamiento de la distribución, respectivamente. Si una variable tiene A= 0 y K = 3, entonces ésta sigue distribución normal. La fórmula de cálculo de estas medidas es la siguiente:

$$A = \frac{E(X-\mu)^3}{[Var(X)]^{\frac{3}{2}}} \quad \text{y} \quad K = \frac{E(X-\mu)4}{[Var(X)]^2}$$

   En este caso X corresponde a los errores del modelo de mínimos cuadrados ordinarios.

6.5.2. Corrección de la no normalidad de los errores

Para corregir la no normalidad de los errores generalmente se usan dos estrategias: aumentar el tamaño de la muestra y buscar una forma funcional adecuada. La primera se basa en el teorema central de límite, y la segunda en que una forma funcional adecuada puede mejorar la distribución de los errores.

6.6. Ejercicios de computador.

Considere el mismo modelo de demanda presentado en el ejemplo 5.9. Se desea efectuar la prueba de multicolinealidad, heteroscedasticiad, autocorrelación, forma funcional inadecuada y normalidad de los errores. El modelo original es:

MODELO DE DEMANDA LINEAL

Dependent Variable: DX
Method: Least Squares
Date: 10/04/06 Time: 10:31
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	72.73351	10.83288	6.714142	0.0002
PX	-4.588739	1.778623	-2.579939	0.0326
PW	-0.386460	1.214240	-0.318273	0.7584
PZ	0.471929	0.688867	0.685080	0.5127
I	-0.409388	0.201019	-2.036558	0.0761

R-squared	0.947105	Mean dependent var	40.38462
Adjusted R-squared	0.920657	S.D. dependent var	16.89940
S.E. of regression	4.760208	Akaike info criterion	6.242183
Sum squared resid	181.2767	Schwarz criterion	6.459471
Log likelihood	-35.57419	F-statistic	35.81046
Durbin-Watson stat	1.436480	Prob(F-statistic)	0.000037

A. *Prueba de Multicolinealidad*

A continuación se presentan la matriz de correlaciones y cuatro regresiones auxiliares con el objeto de identificar la posible existencia de asociación lineal simultánea entre las variables independientes PX, PZ, PW e I:

MATRIZ DE CORRELACION

VARIABLE	PX	PZ	PW	I
PX	1.000000	-0.886170	0.950308	-0.530011
PZ	-0.886170	1.000000	-0.811397	0.493410
PW	0.950308	-0.811397	1.000000	-0.556062
I	-0.530011	0.493410	-0.556062	1.000000

REGRESIÓN DE PX EN FUNCIÓN DE PZ, PW, I

Dependent Variable: PX
Method: Least Squares
Date: 10/05/06 Time: 09:48
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	3.654915	1.624021	2.250535	0.0510
PZ	-0.245705	0.099796	-2.462068	0.0360
PW	0.577343	0.121443	4.754039	0.0010
I	0.007157	0.037598	0.190365	0.8532

R-squared	0.942092	Mean dependent var	6.153846
Adjusted R-squared	0.922789	S.D. dependent var	3.210560
S.E. of regression	0.892115	Akaike info criterion	2.857216
Sum squared resid	7.162817	Schwarz criterion	3.031046
Log likelihood	-14.57190	F-statistic	48.80600
Durbin-Watson stat	1.511932	Prob(F-statistic)	0.000007

REGRESIÓN DE I EN FUNCIÓN DE PW, PX Y PZ

Dependent Variable: I
Method: Least Squares
Date: 10/05/06 Time: 09:51
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	14.84857	17.26790	0.859894	0.4122
PW	-1.308833	1.965637	-0.665857	0.5222
PX	0.560323	2.943420	0.190365	0.8532
PZ	0.367451	1.135704	0.323545	0.7537

R-squared	0.317173	Mean dependent var	11.46154
Adjusted R-squared	0.089564	S.D. dependent var	8.272599
S.E. of regression	7.893449	Akaike info criterion	7.217603
Sum squared resid	560.7588	Schwarz criterion	7.391434
Log likelihood	-42.91442	F-statistic	1.393498
Durbin-Watson stat	2.341199	Prob(F-statistic)	0.306773

REGRESIÓN DE PW EN FUNCIÓN DE PX, PZ, I

Dependent Variable: PW
Method: Least Squares
Date: 10/05/06 Time: 10:00
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-0.915011	2.958162	-0.309317	0.7641
PX	1.238775	0.260573	4.754039	0.0010
PZ	0.130516	0.184036	0.709189	0.4962
I	-0.035872	0.053873	-0.665857	0.5222

R-squared	0.911829	Mean dependent var	7.230769
Adjusted R-squared	0.882439	S.D. dependent var	3.811252
S.E. of regression	1.306773	Akaike info criterion	3.620658
Sum squared resid	15.36890	Schwarz criterion	3.794489
Log likelihood	-19.53428	F-statistic	31.02476
Durbin-Watson stat	1.807371	Prob(F-statistic)	0.000045

REGRESIÓN DE PZ EN FUNCIÓN DE PX, PW, I

Dependent Variable: PZ
Method: Least Squares
Date: 10/05/06 Time: 10:01
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	13.94298	2.424163	5.751669	0.0003
PX	-1.637987	0.665289	-2.462068	0.0360
PW	0.405510	0.571794	0.709189	0.4962
I	0.031290	0.096710	0.323545	0.7537

R-squared	0.797402	Mean dependent var	7.153846
Adjusted R-squared	0.729869	S.D. dependent var	4.431820
S.E. of regression	2.303399	Akaike info criterion	4.754308
Sum squared resid	47.75081	Schwarz criterion	4.928139
Log likelihood	-26.90300	F-statistic	11.80764
Durbin-Watson stat	1.468360	Prob(F-statistic)	0.001791

Se puede observar en la matriz de correlación que los coeficientes de asociación lineal de las variables son bastante altos. En cuanto a las cuatro regresiones auxiliares, la primera señala la existencia de una fuerte relación lineal de PZ y PW con PX al 5% de significancia; mientras la segunda no evidencia relación lineal simultánea entre PX, PZ y PW con I. La tercera y la cuarta regresión exhiben asociación lineal de PX con PW al 1% de significancia, y de PX con PZ al 5 % de significancia, respectivamente. Estas salidas econométricas presentan dependencia conjunta y un alto R^2, excepto en la regresión del ingreso. Lo anterior induce a afirmar la presencia de multicolinealidad en el modelo de demanda. Esta es generada por la estrecha dependencia lineal entre las variables explicativas de los precios PZ, PW y PX.

B. *Prueba de Heteroscedasticidad*

Con el objeto de verificar si los errores del modelo tienen varianza constante se desarrolla la prueba de Heteroscedasticidad de White (sin términos cruzados):

PRUEBA DE WHITE

White Heteroskedasticity Test:

F-statistic	0.714607	Probability	0.682884
Obs*R-squared	7.648476	Probability	0.468539

Test Equation:
Dependent Variable: RESID^2
Method: Least Squares
Date: 10/05/06 Time: 09:43
Sample: 1 13
Included observations: 13

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-35.46343	183.3230	-0.193448	0.8560
PX	-8.022919	26.87381	-0.298540	0.7802
PX^2	1.668258	2.339704	0.713021	0.5152
PZ	12.89670	20.37802	0.632873	0.5612
PZ^2	-0.473441	0.701913	-0.674500	0.5370
PW	6.476469	18.61942	0.347834	0.7455
PW^2	-0.834719	1.273948	-0.655222	0.5481
I	-5.747052	4.094526	-1.403594	0.2331
I^2	0.175578	0.151639	1.157870	0.3113

R-squared	0.588344	Mean dependent var	13.94436
Adjusted R-squared	-0.234967	S.D. dependent var	16.39096
S.E. of regression	18.21512	Akaike info criterion	8.848341
Sum squared resid	1327.162	Schwarz criterion	9.239460
Log likelihood	-48.51422	F-statistic	0.714607
Durbin-Watson stat	1.980651	Prob(F-statistic)	0.682884

El nR^2 de la prueba no es significativo al 1%, 5% y 10%. En este sentido, no se

puede rechazar la hipótesis nula de homoscedasticidad, es decir el modelo original no presenta heteroscedasticidad.

C. *Prueba de Autocorrelación*

La prueba de correlación serial Breusch-Godfrey arroja los siguientes resultados:

PRUEBA BREUSCH-GODFREY

Breusch-Godfrey Serial Correlation LM Test:

F-statistic	0.641758	Probability	0.449407
Obs*R-squared	1.091746	Probability	0.296085

Test Equation:
Dependent Variable: RESID
Method: Least Squares
Date: 10/05/06 Time: 09:26
Presample missing value lagged residuals set to zero.

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	0.348096	11.09241	0.031381	0.9758
PX	0.196375	1.836272	0.106942	0.9178
PZ	0.037411	0.706375	0.052962	0.9592
PW	-0.171777	1.260744	-0.136250	0.8955
I	-0.047910	0.214196	-0.223672	0.8294
RESID(-1)	0.305911	0.381864	0.801098	0.4494

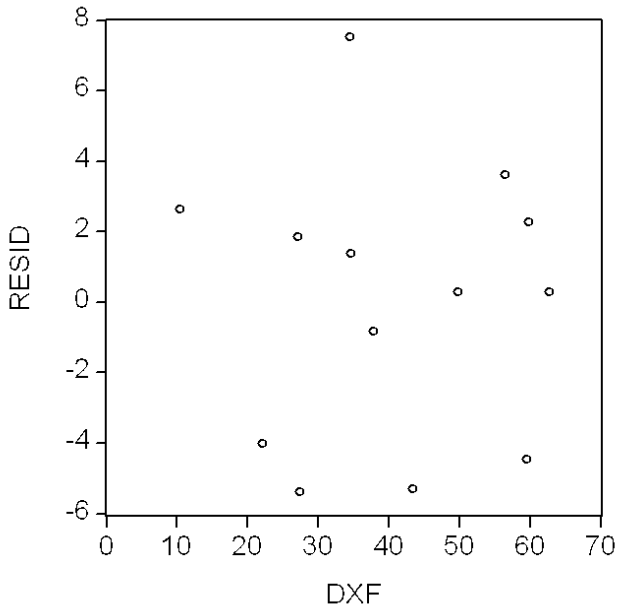
R-squared	0.083980	Mean dependent var	-1.01E-14
Adjusted R-squared	-0.570319	S.D. dependent var	3.886694
S.E. of regression	4.870508	Akaike info criterion	6.308311
Sum squared resid	166.0530	Schwarz criterion	6.569057
Log likelihood	-35.00402	F-statistic	0.128352
Durbin-Watson stat	1.837887	Prob(F-statistic)	0.980963

El estadístico de prueba cuyo valor es 1,0 y el primer rezago de los residuos no son siginificativos al 1%, 5% y 10%. Por lo tanto no hay evidencia estadística para afirmar que existe autocorrelación de orden uno.

D. *Prueba sobre forma funcional inadecuada*

Para verificar si la forma funcional del modelo original de demanda es adecuada se presenta a continuación la gráfica del residuo (RESID) y la variable demanda estimados (DXF), y la prueba RESET de Ramsey:

GRÁFICA DEL RESIDUO Y LA DEMANDA ESTIMADOS



PRUEBA RESET DE RAMSEY

Ramsey RESET Test:

F-statistic	0.240234	Probability	0.639025
Log likelihood ratio	0.438664	Probability	0.507768

Test Equation:
Dependent Variable: DX
Method: Least Squares
Date: 10/05/06 Time: 09:42
Sample: 1 13
Included observations: 13

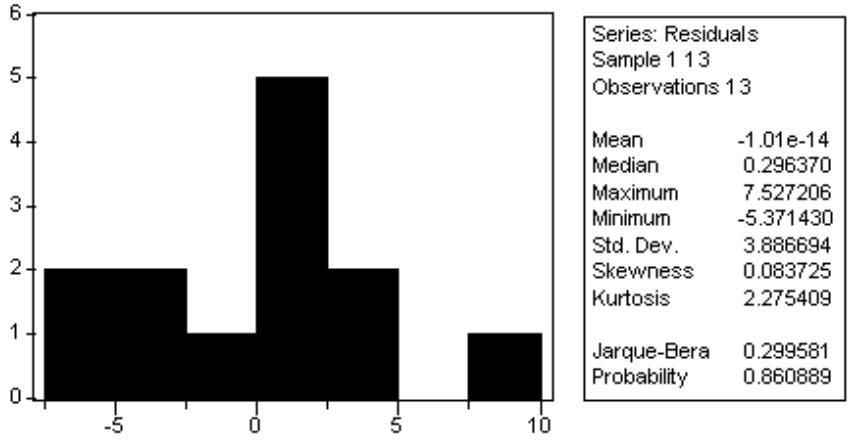
Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	52.50613	42.81102	1.226463	0.2597
PX	-2.010581	5.582463	-0.360160	0.7293
PZ	-0.058048	1.301350	-0.044606	0.9657
PW	-1.026028	1.825321	-0.562108	0.5916
I	-0.245060	0.396301	-0.618368	0.5559
FITTED^2	0.005793	0.011819	0.490137	0.6390

R-squared	0.948860	Mean dependent var	40.38462
Adjusted R-squared	0.912331	S.D. dependent var	16.89940
S.E. of regression	5.003739	Akaike info criterion	6.362286
Sum squared resid	175.2618	Schwarz criterion	6.623031
Log likelihood	-35.35486	F-statistic	25.97566
Durbin-Watson stat	1.543094	Prob(F-statistic)	0.000221

La gráfica no muestra un comportamiento tendencial entre el error y la variable dependiente estimada de la demanda, con lo cual el aporte de ésta última sobre el componente no explicado en el modelo de demanda sería casi nulo. Este análisis es robustecido por la prueba de Ramsey, pues al 1%, 5% y 10 % de significancia (F_c = 0,6390) no puede afirmarse que el modelo de demanda cuenta con una forma funcional inadecuada, es decir se acepta la hipótesis de una buena especificación.

D. *Prueba de Normalidad*

Para probar si los residuos del modelo de demanda siguen distribución normal a se realiza el histograma de los errores y la prueba Jarque Bera. Los resultados de estos procedimientos son los siguientes:



Aunque el histograma de frecuencias de los errores no dilucida una muy clara similitud con la distribución normal, exhibe una alta concentración de observaciones cercana a cero, y frecuencias comparativamente bajas hacia los extremos. El valor de prueba Jarque Bera no es significativo al 1%, 5% y 10%, indicando que no se puede rechazar la hipótesis nula. Por lo tanto se puede afirmar que estadísticamente los errores siguen distribución normal.

7. INTRODUCCIÓN A VARIABLES CUALITATIVAS

7.1. Regresión con variables independientes cualitativas

En algunos casos ciertas características tomadas de una población y recopiladas a través de una muestra no corresponden a variables cuantitativas. Por ejemplo, si se encuesta un conjunto de personas, puede ser importante preguntar información respecto a su sexo, la raza, la región de origen, estado civil, el estrato económico, etc. Estas variables son denominadas variables cualitativas y su tratamiento o

análisis en los modelos econométricos tiene una connotación diferente a las variables cuantitativas.

Por medio de asignaciones numéricas específicas, de escala ordinal o cardinal, las variables cualitativas pueden ser registradas en un modelo econométrico. Un ejemplo de escala cardinal es la variable "género", donde el investigador puede asignar a esta variable una serie de observaciones numéricas como se describe a continuación:

$$GENERO = \begin{cases} 1 & Si \quad es \quad \text{hom}\,bre \\ 0 & Si \quad es \quad mujer \end{cases}$$

En este sentido, a todas las observaciones correspondientes al sexo femenino le será asignado el número cero y a las de sexo masculino el número 1. Esta nueva variable recibe el nombre de variable *dummy*, debido particularmente a que solo podrá tomar dos valores, el uno o el cero.

Por otro lado, un ejemplo de una variable que puede ser representada en escala ordinal es el "estrato" económico. El nombre de ordinal se refiere a que en la estructura de registro el orden tiene gran relevancia. El investigador puede conformar la variable de la siguiente manera:

$$ESTRATO = \begin{cases} 1 & Si \quad el \quad hogar \quad pertenece \quad al \quad estrato \quad 1 \\ 2 & Si \quad el \quad hogar \quad pertenece \quad al \quad estrato \quad 2 \\ 3 & Si \quad el \quad hogar \quad pertenece \quad al \quad estrato \quad 3 \\ 4 & Si \quad el \quad hogar \quad pertenece \quad al \quad estrato \quad 4 \\ 5 & Si \quad el \quad hogar \quad pertenece \quad al \quad estrato \quad 5 \\ 6 & Si \quad el \quad hogar \quad pertenece \quad al \quad estrato \quad 6 \end{cases}$$

Cuando la variable se construye de esta manera recibe el nombre de variable categórica ordenada.

Considérese el siguiente modelo de regresión lineal para un conjunto de hogares:

$$TRABAJO_t = \beta_1 + \beta_2 SALARIO_t + \beta_3 GENERO_3 + \beta_4 ESTRATO_t + \varepsilon_t$$

Donde:

TRABAJO: Número de horas trabajadas al mes

SALARIO: Ingreso laboral

GENERO: Sexo del jefe de familia

ESTRATO: Nivel de estrato económico del hogar.

Si el coeficiente β_3 es positivo se interpreta como el número de horas de trabajo mensual adicionales que ofrece el hogar cuando el jefe de familia es hombre comparado cuando el jefe de familia es mujer, manteniendo los demás factores constantes. Por otro lado, en cuanto al coeficiente de la variable estrato, cuando β_4 es positivo, este manifiesta que hogares con un nivel de estrato más alto ofrecen al mercado laboral más horas de trabajo al mes.

También existen variables categóricas no ordenadas. Por ejemplo considere la siguiente variable cualitativa que representa la región a la que pertenecen un conjunto de hogares:

$$\text{Re}\,gión = \begin{cases} 1 & Oriental \\ 2 & Pacífica \\ 3 & Central \\ 4 & Atlántica \\ 5 & El\ resto \end{cases}$$

Sin embargo la combinación de categorías de región se puede convertir en un sistema de variables dummy: D1=1 si pertenece a Oriental y D1=0 en otro caso, D2=1 si pertenece a Pacífica y D2=0 en otro caso, etc. Se debe tener en cuenta que el número de variables dummy de este sistema es igual al número de categorías menos uno. Si se incluyen un número de variables dummy igual al número de categorías se genera multicolinealidad perfecta en el modelo, por lo que los estimadores no se pueden estimar.

Utilidad de las variables dummy:

1. Sirven para mostrar cambio en intercepto
2. Sirven para mostrar cambio en pendiente
3. Sirven para mostrar cambio en pendiente y en intercepto

De acuerdo con ello, las dummy pueden ser entonces utilizadas para mostrar la existencia de cambio estructural. Por ejemplo para periodos de tiempo (D=1 para datos tomados en o después de 1970 y D=0 para datos antes de 1970).

Ejemplo de un modelo log – lin con una variable dummy:

$$LnW = \alpha + \beta_1 X + \beta_2 D + u$$

Donde W es el salario, X es educación y D es igual a 1 si el individuo es de raza blanca.

Supóngase que el valor estimado de β_2 es 0.26 y es significativo. Tenga en cuenta que la prueba de significancia de una variable dummy se efectúa de la misma forma como cualquier otra variable explicativa. El β_2 estimado muestra cuanto más gana porcentualmente un individuo por ser blanco que otro de diferente raza, suponiendo las demás variables constantes.

El salario esperado de un individuo blanco es: $LnW\ blanco = K + 0.26$, donde K es todo lo que permanece constante. Note que D = 1. El salario promedio de un individuo de otra raza es: $LnW\ otra\ raza = K$. Observe que D = 0. La diferencia entre $LnW\ blanco$ y $LnW\ otra\ raza$ es 0.26. Por lo tanto:

$$LnW\ blanco - LnW\ otra\ raza = 0.26$$

$$\frac{W\ blanco}{W\ otra\ raza} = e^{0,26}$$

$$W\ blanco = 1.297\, W\ otra\ raza$$

Lo anterior que quiere decir que el salario de una persona de raza blanca es 29.7% más alto que el de otra raza. Cuando la variable dependiente del modelo esta transformada en logaritmo este procedimiento resulta útil, el cual se resume en la siguiente fórmula: Efecto $= 100\left[e^{\hat{\beta}_j} - 1\right]$, donde $\hat{\beta}_j$ es el coeficiente de la variable dummy de interés.

En algunos casos el $\hat{\beta}_2$ del ejemplo se interpreta directamente multiplicando por 100 (como un modelo log-lin) sin necesidad de aplicar el antilogaritmo. Así para el ejemplo: 0.26*100=26%. Entonces el salario de una persona de raza blanca es 26% más alto que el de otra raza. No obstante, el resultado obtenido por el primer procedimiento es más exacto.

Ejemplo de un modelo con dos variables dummy:

$$Y = \alpha_0 + \beta_1 X + \alpha_1 D_1 + \alpha_2 D_2 + u$$

Donde Y es gasto en salud, X es el ingreso y D_1 es igual a 1 si el individuo tiene bachillerato y 0 en caso contrario, D_2 es igual a 1 si el individuo tiene universidad y 0 en caso contrario.

$$E(Y/X, D_1 = 0, D_2 = 0) = \alpha_0 + \beta_1 X$$
$$E(Y/X, D_1 = 1, D_2 = 0) = \alpha_0 + \beta_1 X + \alpha_1$$
$$E(Y/X, D_1 = 0, D_2 = 1) = \alpha + \beta_1 X + \alpha_2$$

En este caso solo se generan cambios en intercepto, dado que α_1 y α_2 se agregan a la constante α_0 en su respectivo modelo.

Ejemplo de un modelo con interacción de variables dummy:

$$Y = \alpha_0 + \beta_1 X + \alpha_1 D_1 + \alpha_2 D_2 + \alpha_3 D_1 D_2 + u$$

Donde Y es gasto en medicamentos, X es el ingreso y D_1 es igual a 1 si el individuo es mujer y 0 en caso contrario, D_2 es igual a 1 si el individuo tiene bachillerato.

$$E(Y/X, D_1 = 0, D_2 = 0) = \alpha_0 + \beta_1 X$$
$$E(Y/X, D_1 = 1, D_2 = 0) = \alpha_0 + \beta_1 X + \alpha_1$$
$$E(Y/X, D_1 = 0, D_2 = 1) = \alpha_0 + \beta_1 X + \alpha_2$$
$$E(Y/X, D_1 = 1, D_2 = 1) = \alpha_0 + \beta_1 X + \alpha_1 + \alpha_2 + \alpha_3$$

De nuevo este esquema generan solo cambios en intercepto, dado que α_1, α_2 y α_2 se agregan a la constante α_0 en su respectivo modelo.

Ejemplo de un modelo con variable dummy e interacción de una dummy con una variable continua:

$$Y = \alpha_0 + \beta_1 X + \alpha_1 D + \beta_2 DX + u$$

Donde Y es consumo, X es el PIB y D es igual a 1 si la observación pertenece o es mayor al año 1993 y 0 en caso contrario.

$$E(Y/X, D = 0) = \alpha_0 + \beta_1 X$$
$$E(Y/X, D = 1) = \alpha_0 + \beta_1 X + \alpha_1 + \beta_2 X$$

La última expresión puede escribirse como:

$$E(Y/X, D=1) = (\alpha_0 + \alpha_1) + (\beta_1 + \beta_2)X$$

Esta ecuación es útil para mostrar si existe cambio estructural en intercepto y/o pendiente, dependiendo de la significancia de los coeficientes α_1 y β_2. Por ejemplo si $\alpha_1 = \beta_2 = 0$ el resultado sugiere la no ocurrencia de cambio en pendiente e intercepto en la regresión.

7.2. Regresión con variable dependiente cualitativa

Existen otra clase de modelos en econometría llamados modelos de variable dependiente cualitativa. Estos se dividen en dos clases: los modelos de probabilidad y los modelos de elección discreta para más de dos alternativas.

En los modelos de probabilidad, la variable dependiente solo puede tomar dos valores cero o uno. Por ejemplo, si el jefe de familia tiene empleo o no, así la variable "¿ESTA EMPLEADO?" toma el valor de uno si tiene trabajo o cero en caso contrario. Existen tres formas generales de estimar este tipo de modelos: 1) mínimos cuadrados ordinarios, el cual es conocido como el modelo de probabilidad lineal, siendo el menos utilizado por no cumplir en la mayoría de los casos con los axiomas de la probabilidad; 2) el modelo *logit*, donde la función de distribución que siguen los errores es logística; y 3) el modelo probit, cuando las perturbaciones se asumen con distribución normal.

Dentro de los modelos de elección discreta con más de dos alternativas, se encuentran: el modelo *logit multinomial*, el modelo *probit multinomial* y el modelo *nested logit*. En cada uno de estos, la variable dependiente es categórica, pero a diferencia de los anteriores modelos, esta puede tomar más de dos valores u organizarse en especie de ramas o brazos. Por ejemplo, a un investigador le

puede interesar el tipo de transporte que las personas utilizan para llegar a su lugar de trabajo: bus, automóvil, taxi, transmilenio, bicicleta, etc.; cada una de estas alternativas es distinta.

La forma funcional de los modelos con variable dependiente cualitativa y su interpretación, resulta ser más compleja que la de los modelos con variables independientes cualitativas. Todos los modelos de este tipo, a excepción del modelo de probabilidad lineal no son lineales en los parámetros y se estiman por el método de máxima verosimilitud.

BIBLIOGRAFÍA

1. Canavos, G. (1991), Probabilidad y Estadística. Aplicaciones y Métodos. McGraw Hill. México.
2. Freund, J; Miller, I. y Miller M. (2000), Estadística Matemática con Aplicaciones. Prentice Hall. Pearson Educación. Sexta edición. México.
3. Greene, W. (1998), Análisis Econométrico. Prentice Hall. Tercera Edición.
4. Gujarati, D. (2003), Basic Econometrics, McGraw Hill. Fourth edition.
5. Hamilton, J. (1994), Times Series Analysis. Princeton: Princeton University Press.
6. Judge, G.; Carter Hill, R.; Griffiths, W., Lütkepohl, H. and Lee, T. (1988), Introduction to the Theory and Practice of Econometrics. John Wiley and Sons. Second edition.
7. Novales, Alfonso . 1997. *Econometría*, McGraw Hill, Bogotá.
8. Maddala, G.S. (1983), Limited-Dependent and Qualitative Variables in Econometrics, Cambridge University Press.
9. Mason y Lind. 2001. Estadística para Administración y Economía. Editorial Alfaomega.
10. Mendenhall, W.; Wackerly, D. y Scheaffer R. (1994), Estadística Matemática con Aplicaciones. Grupo Editorial Iberoamérica S.A. Segunda edición.
11. William E. Griffiths, R. Carter Hill, George G. Judge (1993), Learning and Practicing Econometrics. John Wiley & Sons, New York.
12. Wooldridge, Jeffrey M. (2002), Introductory Econometrics: a modern approach, South-Western College Publishing. Second edition.
13. Wooldridge, Jeffrey M. (2002), Econometric Analysis of Cross Section and Panel Data, MIT Press.

ANEXOS

ANEXO 1.
REGRESIÓN LINEAL MÚLTIPLE EN EL PAQUETE ESTADÍSTICO EVIEWS 4.1

Ejemplo

De acuerdo con la siguiente base de datos:

TABLA No. 4. VARIABLES PARA LA ESTIMACIÓN DE LA DEMANDA LINEAL DEL BIEN X.

Obs.	DX	I	PW	PX	PZ
1980	22	10	8	3	9
1981	20	11	9	5	9
1982	19	13	12	6	8
1983	18	14	13	8	7
1984	16	16	15	9	6
1985	14	17	17	10	5
1986	13	18	19	11	4
1987	11	19	21	13	3
1988	9	20	23	15	2
1989	7	21	24	17	2
1990	6	23	25	18	1
1991	5	25	27	20	1

Estime la función de demanda del bien X, teniendo en cuenta la siguiente especificación del modelo:

$$\boldsymbol{Dx} = \beta_0 + \beta_1 \boldsymbol{I} + \beta_2 \boldsymbol{Pw} + \beta_3 \boldsymbol{Px} + \beta_4 \boldsymbol{Pz} + \boldsymbol{U}$$

Donde:

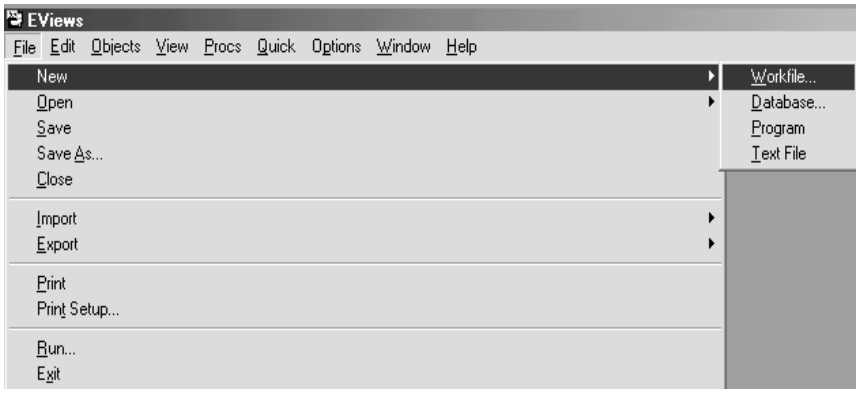
Dx: Cantidad demandada del bien X.
I: Ingreso.
Pw: Precio del bien W.
Px: Precio del bien X.
Pz: Precio del bien Z.
U: Término de error

Desarrollo

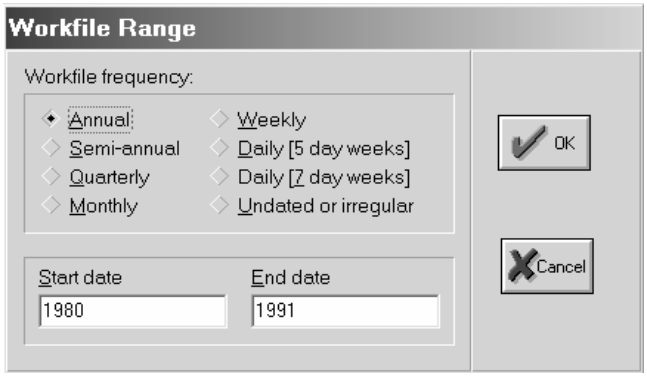
Este ejercicio será desarrollado en el paquete estadístico *Econometrics Views 4.1*. A continuación se muestra todo el procedimiento para estimar el modelo de demanda lineal siguiendo los supuestos del modelo clásico de regresión lineal normal.

A. Importar la base de datos.

Este paquete estadístico puede importar datos en hoja electrónica guardados con extensión wks, wk1 y Excel. Una vez se inicia la sesión en E-views se debe generar un nuevo archivo de trabajo.

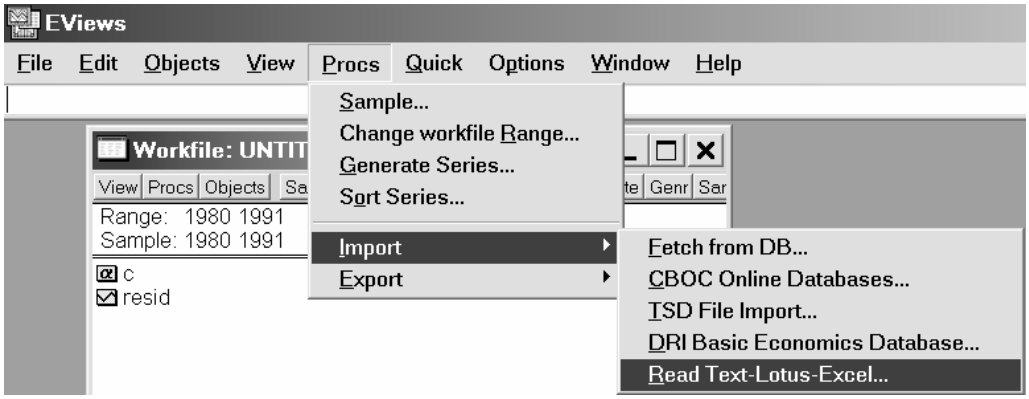


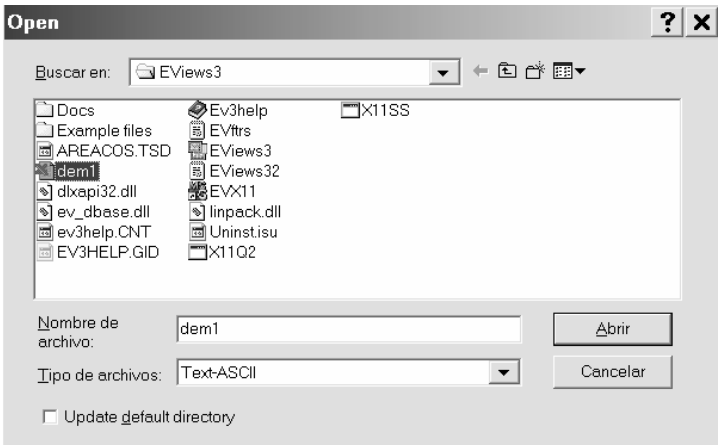
El programa requiere que se seleccione el tipo de frecuencia que caracterizan los datos.



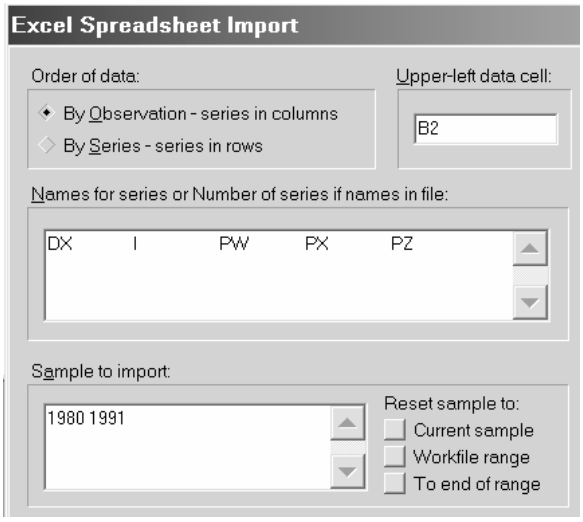
Debido a que los datos presentan una frecuencia anual se elige la opción "*Annual:*" indicando el periodo inicial y final.

Posteriormente el procedimiento es importar los datos que se encuentran en hoja electrónica (Excel).



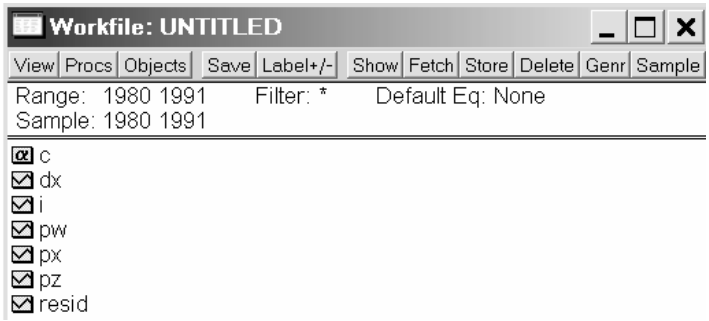


Se selecciona el archivo a importar; en este caso corresponde a dem1.xls del subdirectorio donde se haya almacenado.



Las variables deben ser incluidas en el orden que se encuentran en la base de datos separadas por espacios y con sus nombres correspondientes. Por ejemplo: *Dx, I, Pw, Px* y *Pz*.

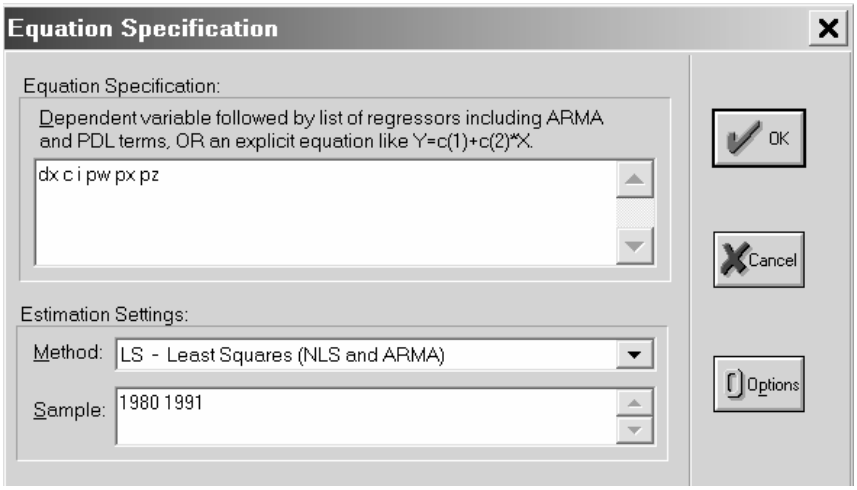
Cuando los datos son importados el programa muestra la siguiente ventana con el respectivo nombre de las variables:



De esta manera la base de datos ha sido importada con todas sus variables. Las observaciones pueden ser vistas al seleccionar las columnas deseadas y pulsando el link *show*.

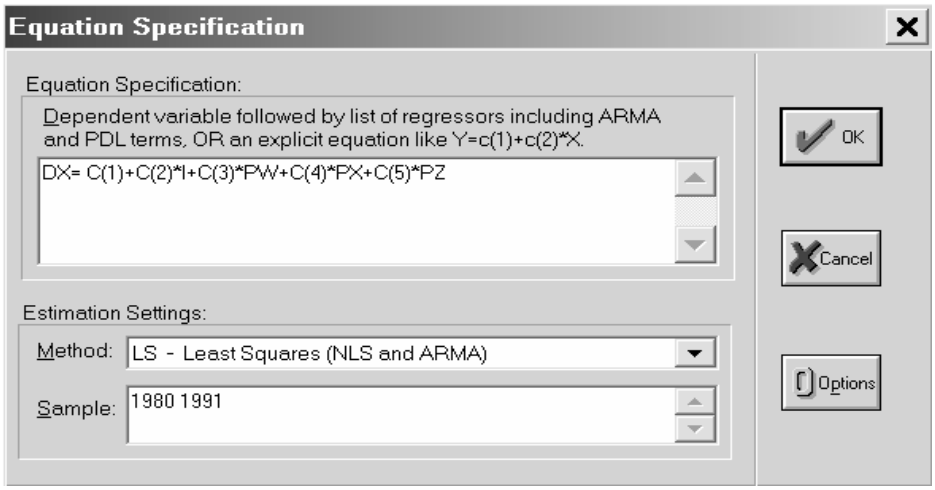
B. Estimación del modelo

Usando el link *quick* y *Estimate Equation* es arrojada una ventana donde la ecuación del modelo debe ser incorporada.



En dicha ventana las variables pueden introducirse separadas por espacios empezando por la variable dependiente y luego las independientes incluyendo la constante cuando no se efectúa regresión al origen.

En esta ventana también el modelo puede introducirse escribiendo la ecuación con los símbolos $(=,*,+)$ nombrando los coeficientes como C(1), C(2), ..., C(n).



Aplicando O.K. de acuerdo con la primera modalidad de estimación, el resultado de es el siguiente:

Equation: UNTITLED Workfile: UNTITLED

View | Procs | Objects | Print | Name | Freeze | Estimate | Forecast | Stats | Resids

Dependent Variable: DX
Method: Least Squares
Date: 01/23/04 Time: 18:15
Sample: 1980 1991
Included observations: 12

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	19.53532	6.615221	2.953087	0.0213
I	0.199881	0.265745	0.752154	0.4765
PW	-0.220685	0.364751	-0.605029	0.5643
PX	-0.712364	0.208323	-3.419519	0.0111
PZ	0.480277	0.491722	0.976725	0.3612

R-squared	0.995856	Mean dependent var	13.33333
Adjusted R-squared	0.993487	S.D. dependent var	5.789227
S.E. of regression	0.467195	Akaike info criterion	1.610199
Sum squared resid	1.527901	Schwarz criterion	1.812243
Log likelihood	-4.661191	F-statistic	420.5069
Durbin-Watson stat	1.986112	Prob(F-statistic)	0.000000