36

UNIVERSIDAD NACIONAL AUTÓNOMA DE MÉXICO

ESCUELA NACIONAL DE ESTUDIOS PROFESIONALES
"ACATLÁN"

EXPLICACIÓN DE LA DEMANDA DEL CONSUMO DE LA LECHE UTILIZANDO UN MODELO DE REGRESIÓN LINEAL MÚLTIPLE.

T E S I N A

QUE PARA OBTENER EL TÍTULO DE

LICENCIADO EN MATEMÁTICAS APLICADAS Y COMPUTACIÓN

P R E S E N T A :

NORA SOSA SEDANO

ASESOR: ING. NORA DEL CONSUELO GORIS MAYANS

ABRIL 2002

DEDICATORIAS Y AGRADECIMIENTOS

Gracias Dios mío, por los dones que continuamente nos das, en especial te agradezco el don de la VIDA y el infinito don de la FE...

Gracias Madre Santísima, por tu continua protección maternal, gracias por llevarme hacia Jesús...

Gracias papá, por tu ejemplo para perseverar en la vida, gracias por todo el esfuerzo que has hecho por mí, sin tu ayuda no hubiera podido terminar...

Gracias mamá, por que parte de lo que soy te lo debo a ti, gracias en especial por tus oraciones que me han sostenido en mi camino ...

Gracias a mis hermanas, por aguantar todas mis manías de estudio, recuerden que siempre están en mi corazón...

Gracias madrina, Mary, Coco, Pato, por ser parte de mi familia, a quienes debo tantos recuerdos gratos....

Gracias a mi amiguita Carmen, Sra. Pola y Sr. Pablo, por abrirme las puertas de su casa, por todas las facilidades dadas, a pesar de que se alargó el trabajo mas de lo que esperaba...

Gracias maestra Nora, por su valiosa ayuda en la asesoría de mi trabajo, por su tiempo, sus comentarios, su interés, gracias por todas sus atenciones...

Gracias a todas las personas que me ayudaron en la obtención de información de todo tipo; gracias a mis maestros que han contribuido a mi formación personal y académica; gracias a mis amigos, sobre todo de generación, por su amistad y por hacer de esta etapa una época especial...

Dedico este trabajo a todos aquellos que lo lean, esperando les sea útil y los motive a realizar aplicaciones reales por medio de un trabajo conjunto...

A todos muchas gracias, que Dios y su Santísima Madre los bendiga...

ÍNDICE

INTRODUCCIÓN

CAPITULO I: TEORÍA ECONÓMICA APLICADA A LA DEMANDA DEL CONSUMO DE LECHE 5

A. Naturaleza de la Econometría. 5
A.1 ¿Qué es Econometría?. 5
A.2 Factores a considerar en la construcción de un modelo econométrico. 9
A.2.1 Modelos económicos y modelos econométricos. 9
A.2.2 Elementos de un modelo econométrico. 10
A.3 Etapas de la elaboración de un modelo econométrico. 18
A.4 Fines de la econometría. 22

B. La teoría de la demanda del consumidor. 24
B.1 Concepto de demanda. 24
B.2 Factores que influyen en la demanda. 25
B.3 Elasticidades. 30
B.4 Números Índice. 34

C. Demanda del consumo de la leche en México. 38
C.1 Descripción del bien y sus derivados. 38
C.1.1 Clasificación comercial. 41
C.2 Comportamiento de la actividad lechera en México. 42
C.2.1 Producción del Sector Lechero Nacional. 42
C.2.2 Consumo nacional de leche. 44
C.2.3 Problemática y políticas del Sector Lechero. 46
C.3 Aplicación de los factores que influyen en la demanda de a leche. 49

CAPITULO II: MODELO DE REGRESIÓN LINEAL MÚLTIPLE. 55

A. Naturaleza del Análisis de Regresión. 55
A.1 Origen e interpretación actual del término de regresión. 55
A.2 Relación entre dos variables. 57
A.3 Análisis de Regresión y causalidad. 59
A.4 Análisis de Correlación. 61
A.5 Algunos problemas en la construcción del modelo de regresión lineal. 62

B. Especificación del Modelo de Regresión Lineal Simple. (M.L.S.) 64
B.1 Construcción del Modelo de Regresión Lineal Simple. 64
B.2 Supuestos básicos del Modelo Clásico de Regresión Lineal. 71
B.3 Estimación del modelo. 76
B.3.1 Métodos de estimación. 79
B.4 Transformación de variables. 92

C. El modelo de Regresión Lineal Múltiple. (M.L.M.) 97
C.1 Presentación del Modelo de Regresión Lineal Múltiple General e hipótesis básicas. 97
C.2 Estimación del modelo. 104
C.3 Tipos de contraste de validez de la ecuación. 110
C.3.1 Prueba t de Student. 120
C.3.2 Análisis de Varianza. 122
C.3.3 Coeficiente de Determinación Simple y Ajustado. 125
C.4 Violación de los supuestos básicos del modelo de regresión. 128
C.5 Predicción del modelo. 140
C.6 Variables omitidas y variables irrelevantes. 144

CAPITULO III: FORMULACIÓN Y CONSTRUCCIÓN DEL MODELO PARA LA EXPLICACIÓN DE LA DEMANDA DEL CONSUMO DE LECHE. 147

A. Especificación del modelo de demanda. 147
A.1 Definición de variables y acopio de información. 147

B. Estimación del modelo. 156
B.1 Estimación de los parámetros del modelo de regresión. 156

C. Validación del modelo. 159
C.1 Verificación de supuestos. 159
C.2 Contrastes del modelo. 162

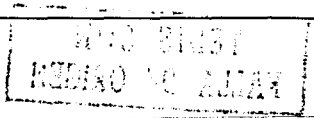
D. Presentación del modelo final. 170
D.1 Interpretación de los resultados. 170
D.2 Explicación de la demanda del consumo de la leche en México. 173

CONCLUSIONES 177

ANEXOS 183

ANEXO A. Denominación de Fracciones Arancelarias y Factores de Conversión. 183
ANEXO B. Introducción al Econometric Views (EVIEWS). 186
ANEXO C. Compendio de Resultados. 195

BIBLIOGRAFÍA 209



INTRODUCCIÓN

La leche, además de ser un producto básico, es uno de los alimentos más completos para el ser humano que se debe incluir en la dieta de todos los mexicanos. En referencia a los productos básicos, en lo personal me llamó mucho la atención saber que en nuestro país hay actualmente importaciones aun de alimentos básicos, como lo son el maíz o el frijol ya que son primordiales en el régimen alimenticio de los sectores bajos. El caso de la leche no es la excepción, pues hasta el año de 1995 ocupábamos el primer lugar en el mundo en la importación de este alimento. El problema de las importaciones no es propio de nuestro país ya que en esta época de globalización el comercio exterior en general se ha visto afectado debido a que pocas naciones son autosuficientes y equilibradas en los mercados de alimentos, en consecuencia la mayoría son netamente exportadoras o importadoras.

En nuestro país, las importaciones de leche son consecuencia del rezago que ha tenido el sector agropecuario[1] en los últimos años, entre otras razones, ante la falta de producción nacional para cubrir la demanda requerida, por el poco apoyo que recibieron los ganaderos de programas encaminados a proteger su industria, al cual se añade la globalización ya que en algunos casos se consigue un mejor precio internacional que no puede competir con el precio nacional. No obstante, en los últimos años el Gobierno le ha dado un mayor impulso a este sector, a través de diferentes programas para volver competitivo al sector agropecuario frente a la perspectiva del comercio exterior.

Asimismo, si bien es cierto que el consumo de la leche es muy importante para nuestra alimentación, está presente también el problema de los salarios, del aumento de precios, etcétera; que puede limitar su consumo. Tomando en cuenta estas consideraciones surge el deseo de conocer cuál es la situación de *la demanda del consumo de la leche* en nuestro país.

Por ser un bien limitado, el tema de la leche se relaciona con la economía (la cual formula los conocimientos y actividades que intervienen en la producción, distribución y *consumo* de los recursos escasos), particularmente con la teoría económica. Cuando se

[1] Referente al campo y a la ganadería.

pretende realizar un análisis económico se encuentra uno con la realidad de que "en la economía todo depende de todo"[2], y allí es donde interviene la teoría económica cuyo principal interés es proporcionar un marco de referencia de las relaciones económicas. En consecuencia, la teoría económica tiene diferentes especializaciones como pueden ser: comercio internacional, finanzas, moneda y banca, economía laboral, teoría del consumidor o del productor, entre otras.

Ahora bien, otro motivo que me llamó la atención es saber que la técnica de regresión lineal tiene una gran aceptación en economía, ya que identificaba el uso de esta técnica con conceptos muy sencillos; por lo cual fue novedoso conocer sobre la utilidad de esta técnica en un área tan importante como lo es la economía. En consecuencia, surge el interés de aprender más sobre la aplicación de la *técnica de regresión lineal*, en particular, para *explicar la demanda del consumo de la leche en nuestro país.*

Al investigar sobre el uso de la técnica de regresión lineal en economía, me introduje en el campo de la econometría, que es la parte de la ciencia económica que se vale de instrumentos matemáticos y estadísticos, para analizar los fenómenos económicos (de manera informal se puede decir que une a la teoría económica, la estadística y las matemáticas para su estudio). Esencialmente, el instrumento estadístico más utilizado en econometría para medir las relaciones económicas y garantizar los resultados del modelo, es el análisis de regresión lineal; empleando principalmente la técnica de mínimos cuadrados y de máxima verosimilitud para la estimación de los parámetros involucrados en sus modelos.

A pesar de que por si misma la técnica de regresión lineal es una buena herramienta, la metodología econométrica ofrece una guía para establecer la relación económica entre las variables que intervienen en la demanda de un producto. De esta manera, el objetivo del trabajo es el de explicar la demanda del consumo de la leche en nuestro país, empleando la técnica de regresión lineal basándome en un estudio econométrico.

De acuerdo a estas ideas, el desarrollo del trabajo se presenta por medio de tres capítulos. El capítulo uno describe algunos tropiezos que tuvo la econometría en sus

[2] BECKER, GARY S.: "*Teoría Económica*", Fondo de Cultura Económica, México, 1977.

inicios para unir las tres disciplinas que la integran, haciendo que se extienda su significado original. Asimismo, se incluyen los elementos básicos que intervienen en la investigación econométrica, y los propósitos principales de su estudio. Como base para la selección de las variables se propone la teoría económica moderna básica para describir la demanda de cualquier bien, esto debido a que me encontré con la limitación de pocas teorías que se enfoquen específicamente a la demanda del consumo de alimentos, y las que hay son muy atrasadas. Se introduce el uso de números índices, como ayuda en la manipulación de los datos, ya que dos variables solo se pudieron recolectar en esta forma; se introduce el concepto de elasticidad, fundamental en la teoría de la demanda, ya que permite medir la sensibilidad de la demanda de un bien ante variaciones en el ingreso y en el precio. También se incluyen las características más sobresalientes de la leche y sus derivados que destacan su importancia alimenticia. Además, se presenta la situàción del sector lechero para contrastar los resultados de la teoría con la realidad de nuestro país. Conjuntamente se incluye una tercera fuente de información de acuerdo a tres entrevistas realizadas en diferentes empresas a los especialistas del área que intervienen en el cálculo de la demanda, con el propósito de obtener información sobre las estadísticas que se realizan en la industria lechera, y contrastar también estos resultados con la teoría económica expuesta y respaldar la selección de las variables en el modelo.

El capítulo dos ofrece un panorama general del origen del análisis de regresión lineal, su uso y sus limitaciones. También se describe la metodología necesaria para aplicar la técnica de regresión lineal, que incluye los supuestos básicos que se deben cumplir para garantizar los resultados de la regresión. Se muestran los dos métodos más utilizados en econometría para la obtención de los parámetros de regresión, así como las funciones que comúnmente se emplean en los modelos económicos, con el fin de establecer la forma funcional de la demanda, como se recomienda en econometría. Se pone especial énfasis en el modelo de regresión lineal múltiple (el cual se presenta en forma matricial para su mayor comprensión) y, en el método de mínimos cuadrados como más apropiado para la obtención de los parámetros del modelo. Además, se incluyen las pruebas necesarias para validar el modelo, por medio de los supuestos básicos (y la solución en caso de incumplimiento), y diferentes herramientas estadísticas para probar la significancia de las variables en forma individual, por grupos de variables, y para probar la posible introducción o la eliminación de variables en el modelo.

El capítulo tres comprende la parte práctica. Con ayuda de los conceptos presentados anteriormente, se lleva a cabo la selección de las variables así como la recopilación de los datos para cada variable que se incluirán para poder explicar la demanda del consumo de la leche. Asimismo, en referencia a otras aplicaciones de trabajos econométricos sobre la demanda de un bien, propuse cuatro ecuaciones (con las mismas variables) para examinar los diferentes resultados, lo cual es viable dado el tamaño muestral y la simplicidad del modelo. A continuación, para elegir cual de los modelos explican con mejor apego a la realidad la demanda de la leche, me auxilié del paquete econométrico "Econometric Views" (EVIEWS) como ayuda en la estimación y validación de los modelos; el motivo de utilizar este paquete fue debido al uso frecuente que encontré en el ámbito econométrico. Se concluye con la presentación de los resultados para el modelo seleccionado, y su uso en la explicación de la demanda del consumo de la leche en nuestro país, apoyado con el concepto de elasticidad precio e ingreso, como marca la teoría.

Al final del trabajo se incluyen las conclusiones y sugerencias, de acuerdo a los resultados obtenidos y a la experiencia en la realización del trabajo. También se presentan los anexos siguientes: uno que indica como hacer la conversión a litros de los diferentes tipos de leche en las importaciones y exportaciones, esto como guía para consultar estos datos, ya que de otra forma es difícil entender como utilizar esta información a simple vista. Otro anexo describe las funciones utilizadas en el desarrollo del trabajo con el paquete econométrico EVIEWS, pues aunque fue de gran ayuda el manual elaborado por un maestro de la Facultad de Economía, al principio me perdí un poco por los ejemplos que se incluyen de aplicaciones econométricas; por ello en el anexo introduzco por lo menos dos formas diferentes (en algunos casos) para realizar una aplicación, como ayuda para que les sea más fácil aprenderlo a quienes no estén tan relacionados con las aportaciones econométricas. Finalmente, un anexo que comprende los resultados obtenidos con la aplicación del EVIEWS, y que se resumen en el tercer capítulo.

CAPITULO I

TEORÍA ECONÓMICA APLICADA A LA DEMANDA DEL CONSUMO DE LECHE.

A. NATURALEZA DE LA ECONOMETRÍA.

A.1. ¿QUÉ ES LA ECONOMETRÍA?.

La *econometría* es una rama de la economía muy reciente, sobre la cual no existe un acuerdo general acerca de sus inicios. Se piensa que un hecho histórico que aceleró su nacimiento fue la crisis económica que se inició en Wall Street (Estados Unidos) en 1929. En esta ocasión, ni los teóricos de la Economía ni los estadísticos fueron capaces de predecir la crisis económica; según dijo el economista y filósofo francés Simiand "los primeros construían *teorías sin hechos* (no contrastaban con la realidad sus hipótesis y teorías), y los segundos *hechos sin teorías* (las extrapolaciones derivadas del comportamiento temporal de sus indicadores, no tenían una base teórico-económica).[1]

Así antes de la crisis de 1929 no existe la colaboración entre los estadísticos, los matemáticos y los economistas. Esto se aprecia en el relato de Charles Roos[2] sobre los comienzos modestos y difíciles que tuvo la creación de la "Econometric Society"[3] debido a la situación que prevalecía entre algunos científicos. Entre los años de 1926 y 1927, Roos trató de editar un trabajo en el que por su propia naturaleza estaban integradas la teoría económica con la estadística a través de las matemáticas superiores (el resultado fue un texto muy técnico y difícil de leer); ofreció el material a una revista de

[1] Citado en Alcaide, 1992.

[2] Para una traducción completa del artículo en español se puede consultar: ALCAIDE, ANGEL: "*Lecturas en Econometría*". Ed. Gredos, Madrid, 1972.

[3] Suele relacionarse el origen formal de la Econometría con la creación de esta sociedad.

economía, otra de estadística y otra de matemáticas, las cuales estaban dispuestas a publicar su estudio con la condición, en cada caso, de que prescindiera de las aportaciones de las dos ramas en que no estaba especializada la revista.

En 1928 Roos le envió al profesor Wilson de la universidad de Harvard, un resumen de su trabajo el cual tuvo éxito y originó una sección dedicada al desarrollo científico de la economía y la sociología en la Asociación Americana para el progreso de la Ciencia. Como continuación de este movimiento científico, el profesor noruego Ragnar Frisch (creador de la palabra *Econometría*) y los norteamericanos Irving Fisher y Charles Roos constituyen una sociedad internacional de económetras denominada la *"Econometric Society"* el 29 de diciembre de 1930 en los Estados Unidos. De esta sociedad se inició mas tarde la publicación de la revista *Econométrica* en el año de 1933, la cual ha sido impulsora de los avances de la econometría. Junto con esta sociedad se crea también la *Cowles Commision for Research in Economics*, fundada en 1932 por Alfred Cowles, que en 1955 se asocia a la universidad de Yale transformándose en la actual *Cowles Fundation for Research in Economic.*

Ambas sociedades estuvieron estrechamente unidas desarrollando varios métodos y técnicas especialmente en los años treinta y cuarenta. En la década de los cincuenta no hubo muchas aportaciones, aunque fue la época en que aparecieron los primeros libros sobre métodos econométricos.

Es a partir de los años sesenta en que las técnicas econométricas tienen un rápido desarrollo, tanto desde el punto de vista teórico, como empírico. De entre la gran cantidad de aportaciones se pueden citar la de Theil en 1961, que generaliza el método de mínimos cuadrados en dos etapas (o bietápicos) para sistemas de ecuaciones simultáneas; y al año siguiente el propio Theil junto con Zeilner, obtienen los estimadores de mínimos cuadrados en tres etapas, aplicados también para ecuaciones simultáneas[4]. Los avances que se han logrado propiciaron la solución a problemas especiales de la econometría, aplicándose también métodos espectrales y bayesianos. Todo esto junto con la aparición de las computadoras, logran generalizar las aplicaciones de los modelos macroeconómicos en muchos países desarrollados.

[4] Ambos métodos de mínimos cuadrados se aplican una vez que se realizó la denominada *identificación* de un modelo resultando estar *sobreidentificado*. Para mayor información consultar cualquier libro de econometría con ecuaciones simultáneas.

Posteriormente en el año de 1970 la técnica de Box & Jenkins (en su publicación: *Time Series Analysis Forecasting and Control*), dio un giro nuevo a la metodología para la modelización de relaciones económicas dinámicas. En los últimos años se han venido publicando los desarrollos de esta técnica en revistas especializadas.

Como se aprecia, la Econometría, ha ido evolucionando tras una serie de aportaciones que le han dado una autonomía propia. En consecuencia su concepto ha venido ampliándose aunque no se debe dejar a un lado su significado literal que significa "*medición económica*". El concepto de medición económica siempre será básico independientemente de los fines que se planteen y los métodos que se apliquen.

En la obra de Alcaide[5], se presentan varias definiciones con aportaciones muy interesantes de la Econometría y su campo de aplicación; aquí se cita parte de los primeros párrafos del artículo primero de los estatutos de la Econometric Society para partir de una idea:

> "*La Econometric Society es una sociedad internacional para el progreso de la teoría económica en sus relaciones con la estadística y las matemáticas.*"

> "*Su objetivo esencial es el de favorecer los puntos de vista teórico y empírico en la explotación de los problemas económicos y que estén inspirados en un estudio metódico* y riguroso..."

que junto con el último párrafo del artículo marca la unidad de los conocimientos teóricos y empíricos tan necesarios en una investigación; y marca la división de la econometría:

> "*Toda actividad susceptible de favorecer mediata e inmediatamente tal unificación de los estudios económicos teóricos y empíricos cae bajo la acción de la sociedad*".

Como se ha dicho, para cumplir con su objetivo, la econometría se sirve de tres ramas: la *teoría económica*, la *estadística* y las *matemáticas*. La forma adecuada en que la econometría combina la teoría económica, las matemáticas y la estadística para sus fines es:

- Como su objetivo lo constituyen los fenómenos económicos, su base es la teoría económica, que guía a la aceptación, el rechazo o la reformulación de las teorías o

[5] Alcaide 1992. Ver bibliografía.

hipótesis económicas. Cabe mencionar que actualmente se aplican los métodos econométricos a otras ciencias sociales.

- Las matemáticas se usan para expresar las afirmaciones verbales de las teorías o hipótesis, por medio de símbolos matemáticos. Por lo tanto, la econometría requiere el uso de *modelos*.
- La otra característica que se resalta es la medición. Así el análisis estadístico aplica técnicas apropiadas para estimar y contrastar las relaciones inexactas y no experimentales entre las variables económicas, utilizando datos económicos apropiados tomados de la realidad para evaluar los resultados por medio de técnicas estadísticas.

Cualquier otra consideración de la combinación de estas ciencias llevaría a otra rama distinta del saber, que aunque son afines a la econometría, tienen sus características propias.

Así, la unión de la economía con las matemáticas constituye la ***economía matemática*** que se ha ocupado de estudiar los modelos que se ajustan a la conducta humana dentro de la esfera de la vida económica, con poco o ningún interés por problemas estadísticos tales como los errores de medición de las variables en estudio. El econometrista puede utilizar las ecuaciones de la economía matemática, aunque requiere todo un arte el convertirlas a ecuaciones econométricas, pues el economista matemático considera como si las relaciones incluidas se cumpliesen exactamente, que no cambian.

La unión de la economía con la estadística forma la ***estadística económica*** cuyo estudio se centra en la recolección, procesamiento y presentación de cifras económicas en forma de gráficas y tablas. Estos datos constituyen la materia prima para los problemas econométricos. La estadística económica se ocupa de la medición de variables tomadas de una en una, la econometría se ocupa de la cuantificación de las relaciones económicas entre variables, de la confrontación de la teoría con la evidencia empírica. El estadista económico no va más allá de la recolección de los datos.

Finalmente la unión de la estadística con las matemáticas constituye la ***estadística matemática*** cuyas técnicas no hacen referencia explícita a variables económicas. El econometrista, aunque puede tomar las herramientas que proporciona el estadista matemático, se ve en la necesidad de desarrollar métodos especiales para poder hacer los análisis de las variables económicas en particular. De hecho, han habido numerosas

aportaciones econométricas, especialmente para sistemas de ecuaciones simultáneas como son los ya mencionados métodos de mínimos cuadrados en dos y tres etapas.

En el presente trabajo al tratar de explicar la demanda del consumo de la leche, se tomará en cuenta la teoría económica que se ha desarrollado para estudios de demanda, por supuesto se necesita el uso de un modelo matemático y técnicas estadísticas para poder saber si son representativos los datos y el modelo. Todo esto lo proporciona la econometría y no sus ramas afines.

A.2. FACTORES A CONSIDERAR EN UN MODELO ECONOMÉTRICO.

A.2.1. Modelos económicos y modelos econométricos.

La palabra *modelo* es ampliamente aceptada como sinónimo de "representación simplificada de la realidad". Dentro de este concepto tienen cabida desde el modelo *mental* (representación no explícita o exteriorizada), el modelo *verbal* (descripción en lenguaje ordinario del modelo mental), el modelo *físico* (representación material mediante objetos de un sistema) hasta el modelo *matemático* (descripción del sistema con la ayuda de un lenguaje matemático); con la única condición de que la representación sea simplificada resaltando lo fundamental y básico. En las ciencias sociales la aplicación de modelos físicos es casi nula.

Un *modelo económico* es una representación simplificada y en términos matemáticos de un conjunto de relaciones económicas con una característica en común que tratan de ser aplicables con validez general. A su vez, se han desarrollado modelos específicos para su aplicación a sistemas reales concretos basándose en la teoría económica[6], estos son los *modelos econométricos* (*Figura. 1.1.*)

Un modelo económico no se puede considerar econométrico ya que poseen ciertas diferencias que se muestran a continuación:

[6] En la ciencia económica existen diversas teorías que han permitido obtener ciertos modelos en forma matemática, pero que están condicionadas al momento del tiempo y lugar en que se apliquen.

- El modelo econométrico exige una especificación estadística más precisa de las variables que lo componen.
- Un modelo econométrico implica medida, necesita de datos estadísticos tomados de la realidad para verificarlos, si es el caso, con la propia teoría.
- El modelo econométrico siempre exige una forma funcional definida, mientras que un modelo económico puede omitir el expresar la forma funcional.
- El necesario carácter aleatorio del modelo econométrico se evita en el modelo económico.
- Comúnmente los modelos econométricos se establecen como relaciones inciertas entre las variables, mientras que los modelos económicos se proponen como relaciones exactas.

Ahora bien, para conocer con precisión si un modelo es aplicable a la realidad, se presenta el apartado siguiente:

A.2.2. Elementos de un modelo econométrico.

El concepto de modelo introducido en la sección anterior, lleva a especificarlo mediante un sistema de relaciones matemáticas. Un modelo econométrico como el siguiente

$$Y_i = \beta_1 + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + u_i \qquad 1.1.$$

se dice *especificado* cuando se establecen con precisión las variables, parámetros y ecuaciones que relacionan las diferentes variables, además incluye datos estadísticos; es decir, las observaciones para cada una de las variables económicas del modelo (Figura 1.1).

A) VARIABLES.

Una primera diferencia entre las variables del modelo 1.1, se tiene al clasificarlas en variables *observables* (Y_i, $X_{2i},\ldots X_{ki}$) y en variables *no observables* (u_i), también llamadas variables latentes.

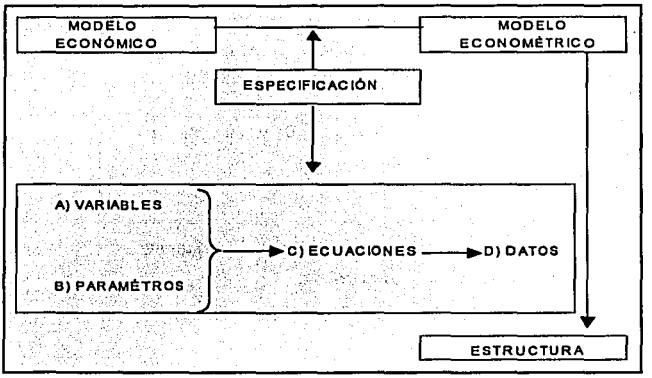


Figura 1.1 Modelos econométricos. Concepto y especificación.

Las variables **observables** se pueden clasificar a su vez en variables *endógenas* y en variables *predeterminadas*, como se muestra en la figura 1.2.

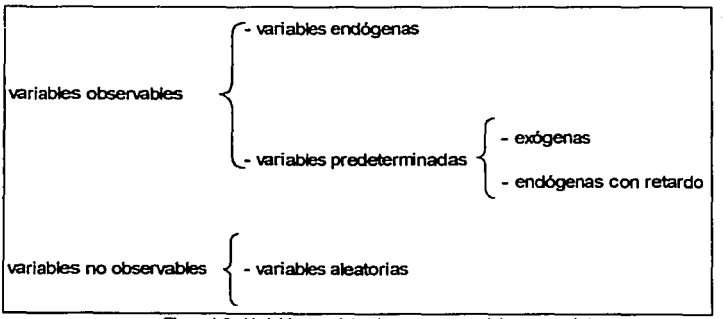


Figura 1.2 Variables que intervienen en un modelo econométrico.

Las variables **endógenas** o variables dependientes, son aquellas variables que explican o influyen en un modelo, a su vez también pueden ser influidas por otras

variables. Por ejemplo, en el siguiente modelo macro de determinación de la renta (ingreso), el consumo está determinado por la renta, y la renta, a su vez, es función del consumo; por tanto, consumo y renta son variables endógenas.

$$C_t = \beta_1 + \beta_2 R_t + u_t \qquad \text{para} \quad \beta_1 > 0 \quad \text{y,} \quad 0 < \beta_2 < 1$$
$$R_t = C_t + I_t \qquad (1.2.)$$

donde:

C_t = consumo per cápita (por persona) en el año t;
R_t = renta per cápita en el año t;
I_t = volumen de inversión en el año t.

Las variables que influyen, pero nunca son influidas por otras variables son llamadas variables **predeterminadas**. Son variables independientes que contribuyen a explicar las variables endógenas. Estas pueden subdividirse a su vez en variables *exógenas* y variables *endógenas con retardo.*

Las variables ***exógenas*** incluyen variables económicas propiamente y variables no económicas. Ambas se encuentran determinadas fuera del fenómeno económico que se trata de modelar, pero son explicativas. A menudo éstas variables juegan un papel muy importante en la determinación de modelos en forma agregada. En la ecuación 1.2. la inversión I es una variable exógena con significado económico. Ejemplos de variables exógenas sin significado estrictamente económico son factores de tipo meteorológico como la lluvia, el tiempo, etc.

Por variables ***endogenas con retardo***, también llamadas retardadas o desplazadas, se entiende aquella variable referida a momentos o periódos de tiempo pasado que influyen en las endógenas del tiempo presente. La introducción de estas variables constituye la construcción de los modelos llamados dinámicos. Un ejemplo es la determinación del consumo en el momento t, mediante el modelo siguiente:

$$C_t = \beta_1 + \beta_2 R_t + \beta_3 C_{t-1} + u_t$$

en la ecuación, la variable consumo aparece como variable endógena en el primer miembro y como predeterminada (endógena con retardo) en el segundo.

Las **variables no observables**, son *variables aleatorias* que constituyen una categoría fundamental en el análisis econométrico. Su introducción caracteriza a los modelos probabilísticos de uso habitual en econometría, en contraste a los modelos deterministas que siguen dominando en la economía matemática.

Las variables aleatorias son propias de la estadística teórica, cuyo análisis de sus supuestos será realizado en el capítulo siguiente. Su misión es recoger el conjunto de causas que no se encuentran explícitamente incorporadas en el modelo, tales como:

1) *Omisión de variable explicativas.* En la especificación de una ecuación se incluyen aquellas variables que se consideran más relevantes, con el propósito de tener un modelo más manejable. Sin embargo, al omitir ciertas variables, no se conoce el aspecto positivo o negativo que influye en forma conjunta sobre la variable endógena. Un principio general que debe observarse en la selección de variables es que la contribución explicativa de las que se excluyen debe ser proporcionalmente inferior a la debida al conjunto de variables incluidas.
2) *Errores de especificación.* Suponiendo que se han incluido las variables explicativas más relevantes, la variable aleatoria recoge los efectos de una especificación incorrecta. Por ejemplo, cuando se especifica que la relación de entre las variables es lineal, pero los datos muestran una relación no lineal (como la logarítmica, exponencial, etc.)
3) *Errores de medida sobre las variables endógenas.* Los datos de las variables endógenas y exógenas, son magnitudes numéricas, obtenidas mediante un proceso de medición humano no exacto. Aún cuando los datos sean válidos, siempre contendrán un pequeño error de medición, error que suele considerarse aleatorio y se le incorpora en la variable aleatoria de cada ecuación de un modelo.

En cualquier caso, independientemente de las causas que originen la aleatoriedad, es de suma importancia la presencia de la variable aleatoria en las ecuaciones de los modelos econométricos[7]. Su aleatoriedad implica el uso obligado de técnicas estadísticas, y cualquier resultado econométrico será establecido en términos probabilísticos.

[7] Excepto en las ecuaciones denominadas como *identidades* (también conocidas como ecuaciones de equilibrio móvil) o las *definiciones*, que se mencionan en la clasificación de ecuaciones con un *criterio económico* (Ver inciso C de esta sección).

B) PARÁMETROS.

Los parámetros son los coeficientes matemáticos que acompañan a las variables en el modelo 1.1, a saber son $\beta_1,\beta_2,...,\beta_K$; sirven para expresar cuantitativamente la influencia de las variables explicativas sobre las correspondientes variables explicadas. Por este motivo suele incluirse un parámetro por variable explicativa en cada ecuación, para posteriormente determinar su valor numérico por algún procedimiento estadístico; lo cual se profundizará más en el capítulo siguiente.

Los parámetros son cantidades fijas[8] dentro de un fenómeno concreto y, por ello, en econometría se les nombra comúnmente como *parámetros estructurales*, en referencia al concepto de estructura que se verá más adelante.

Un aspecto que conviene tener en cuenta en las aplicaciones, es que la teoría económica suele exigir a los parámetros un cierto comportamiento matemático (generalmente acotando su campo de definición) sobre los signos y posible tamaño en una función dada; en tal caso se estará hablando de *restricciones a priori* sobre los parámetros.[9] Por ejemplo, para el modelo de consumo 1.2, se tienen expectativas teóricas a priori, de que β_1 sea mayor o igual a cero, y β_2 esté comprendida entre cero y uno.

C) ECUACIONES.

Las ecuaciones son relaciones que explican la forma en que se relacionan las variables y parámetros del modelo. Por el concepto de parámetros estructurales, también se les conoce en econometría como *ecuaciones estructurales*.

Existen varias formas de clasificación de ecuaciones, para el presente trabajo se muestran a continuación tres formas.

[8] Aunque pueden considerarse también como variables aleatorias con su correspondiente distribución de probabilidad, lo cual se estudia con un enfoque alternativo llamado enfoque bayesiano, el cual ha sido poco utilizado. Si se desea saber sobre métodos bayesianos en econometría, consultar, por ejemplo, MADDALA, G. S.: "*Econometría* ", McGraw-Hill, México, 1985.

[9] El fin de la econometría conocido como análisis estructural, se alcanza cuando se comprueba mediante los resultados del modelo la validez de las restricciones a priori de los parámetros

De acuerdo a su **proceso de estimación**, un criterio para clasificar las ecuaciones es:

a) *Lineales.* Cuando la relación estructural relaciona linealmente las variables y los parámetros.
b) *No lineales.* Cuando se tiene una relación matemática de cualquier otro tipo entre las variables y parámetros.

En el capítulo siguiente se explicará sobre la preferencia de las relaciones lineales o fácilmente linealizables, para facilitar la estimación y contrastación del modelo.

Un **criterio econométrico** que es de utilidad práctica, es clasificarlas según el número de ecuaciones que puede ser:

a) *Uniecuacional.* Cuando el problema se define por una sola ecuación.
b) *Multiecuacional.* Cuando se utiliza un sistema de ecuaciones en la solución de un problema.

Esta clasificación es importante porque encontré que ciertos autores de técnicas de pronósticos clasifican a un modelo econométrico solamente cuando involucra un sistema de ecuaciones, lo cual puede traer alguna confusión. En la actualidad hay libros de econometría exclusivamente para modelos uniecuacionales.

Desde un punto de vista **económico**, las ecuaciones que aparecen en un modelo econométrico se puede clasificar en tres grandes tipos[10] :

a) De *comportamiento.* Son aquellas que describen acciones (el comportamiento) de sujetos o agentes económicos. La función de producción es un ejemplo de este tipo ya que refleja la forma en que se comportan los agentes económicos conocidos como productores.
b) De *restricción.* Reflejan comentarios a la actuación de los agentes económicos. Básicamente definen condiciones que imponen el ordenamiento en el aspecto social (ecuaciones institucionales), el ordenamiento jurídico (ecuaciones legales, por ejemplo, una ecuación de impuesto condicionada a las leyes) o un cierto

[10] Para una clasificación más amplia sobre modelos económicos, ver por ejemplo: Dagum 1980.

proceso técnico o de producción (ecuaciones tecnológicas), sobre el fenómeno de estudio.

c) *Identidades* o de equilibrio móvil y *definiciones.* Expresan relaciones contables o identidades cuantitativas entre magnitudes económicas. No aportan ninguna información adicional y no son susceptibles a ser estimadas o contrastadas; sin embargo, en algunas ocasiones reflejan relaciones de equilibrio en un modelo, y en otras son útiles para completar una especificación econométrica en modelos multiecuacionales.

Un criterio parecido es según el *contenido experimental,* donde las ecuaciones de tipo económico que aparecen en un modelo econométrico pueden ser ecuaciones: de *comportamiento, institucionales, tecnológicas,* de *definición* y de *equilibrio móvil.*

Cada ecuación puede explica un sector (agricultura, manufactura, gobiemo, etc.) o una categoría (productores, consumidores, inversores, instituciones financieras, etc.) de la actividad económica en estudio.

D) DATOS ESTADÍSTICOS.

Para que un modelo económico pueda ser aplicado a la realidad, es necesario disponer de información apropiada por medio de la obtención de datos estadísticos precisos sobre cada una de las variables incluidas en el modelo econométrico, y posteriormente realizar inferencias sobre los parámetros del modelo, es decir, obtener las estimaciones y realizar los contrastes de hipótesis pertinentes.

En la búsqueda y obtención de datos, se dispone de dos fuentes de información: la ofrecida por los anuarios o fuentes estadísticas y la determinada por elaboración propia aplicando técnicas de muestreo. Según la forma de observación en que son recogidos los datos se dividen básicamente en dos tipos: *los datos de series de tiempo y los datos de corte transversal.*

Los datos de una **serie de tiempo** *o series temporales,* se obtienen de un conjunto de observaciones sobre una variable determinada, efectuadas a intervalos regulares de tiempo en *diferentes periodos.* En general la información de series de tiempo puede tener periodicidad diaria, semanal, mensual, trimestral, y anual, aunque para la aplicación de

modelos econométricos lo más común es utilizar las tres últimas, de acuerdo al problema a analizar. Los datos anuales y mensuales sobre los principales indicadores macroeconómicos y sobre las ventas de una empresa, son ejemplos de datos temporales.

La información de **corte transversal**, *series espaciales o cross secction* es la que se obtiene del muestreo de diferentes individuos, empresas, estados o países, para una variable determinada en un *momento dado del tiempo (o el espacio).* A diferencia de las series de tiempo, aquí la unidad básica es el tipo de entidad elegida, que puede ser cualquiera de las mencionadas anteriormente. Ejemplos clásicos de información de corte transversal son las encuestas de ingreso-gasto, los censos, y en general, todo tipo de encuestas. La ventaja que tienen estos datos sobre las series de tiempo consiste en que, si se desea, se puede ampliar el tamaño de la muestra añadiendo el número de observaciones que se requiera. Una posible desventaja que tiene este tipo de información es la limitación en cuanto al número de observaciones, ya que a medida que ésta es más agregada en el tiempo, reduce el tamaño de la muestra.

En algunos casos, pueden combinarse las observaciones de ambos tipos de datos, lo que constituye los denominados **datos de panel** o mezcla de datos que son observaciones sobre distintas unidades (individuos, grupos u objetos) en diversos momentos del tiempo. Debe quedar claro que una parte o todas las unidades sobre las que se recaba información no deben variar en los diversos periodos de tiempo.

Por último, se tienen los modelos econométricos con **datos de carácter cualitativo**. Atributos tales como el sexo ó el estado civil pueden cuantificarse mediante variables ficticias, también conocidas como variables *Dummy*. Existen ocasiones en que no se tiene información sobre algún fenómeno o se observan anomalías cuando se trata de explicar una variable, en este caso se introducen las variables binarias del tipo cero-uno para cuantificar el suceso en estudio. La estacionalidad en las series, se puede modificar mediante estas variables.

- **Estructura económica.**

Cuando se obtienen los valores para los parámetros de las ecuaciones se establecen relaciones en las variables económicas bajo el supuesto de que los parámetros son constantes, es decir, no cambian de valor numérico en todo el periodo

temporal para el que se observan las variables. Esto en terminología econométrica se conoce como **estructura económica** o simplemente estructura, la cual antes de dar por válida, es necesario contrastar convenientemente según las etapas del modelo (que se verán en la sección siguiente).

La introducción del concepto de estructura no tiene una justificación académica, sino que constituye un punto clave para utilizar adecuadamente los modelos[11] describiendo conductas humanas e institucionales, así como leyes tecnológicas. El nombre de estructura económica designa tanto a los parámetros del modelo como a las ecuaciones, como se dijo anteriormente.

El concepto de modelo econométrico es más amplio que el concepto de estructura, al igual que el concepto de modelo económico es más amplio que el de modelo econométrico. Por tanto se hace notar que: de un *modelo económico* pueden surgir tantos *modelos econométricos* como combinaciones diversas de formas funcionales puedan adoptar sus ecuaciones, a su vez, de un modelo econométrico se obtienen tantas *estructuras* como grupos alternativos de datos se pueden tomar (*Figura 1.1.*). Un modelo es una *familia de estructuras.*[12] Por eso cada que hay un cambio estructural se recomienda volver a estimar el modelo.

A.3 ETAPAS DE LA ELABORACIÓN DE UN MODELO ECONOMÉTRICO.

La elaboración de un modelo econométrico se puede dividir en tres grandes etapas: *especificación, estimación y validación.*

Si bien en la construcción de una primera aproximación a un modelo econométrico estas etapas siguen un orden secuencial, al elaborar el modelo es necesario, por regla general, retroceder en más de una ocasión dentro de este orden secuencial. Es decir, el proceso de elaboración de un modelo econométrico no es un proceso que sigue un orden establecido, sino que es necesario confrontar continuamente el modelo con los datos y

[11] El concepto de *estrutura* y de *parámetros estructurales* lo estableció Jacobo Marschak. Ver Alcaide , 1992.

[12] No obstante, también se dice que una ecuación estructural o de comportamiento puede reflejar la *estructura* de un modelo económico, de una economía o el comportamiento de un agente económico, por ejemplo un productor.

con cualquier otra fuente de información, con la finalidad de obtener un modelo econométrico, compatible con los datos, que permita analizar la realidad, ofrezca mejores predicciones o constituya una buena base para tomar decisiones, según el fin que se persiga. Los pasos necesarios para la elaboración de un modelo dentro de la investigación econométrica se presentan en la figura 1.3.

A continuación se describen en forma general, las tres etapas enumeradas anteriormente.

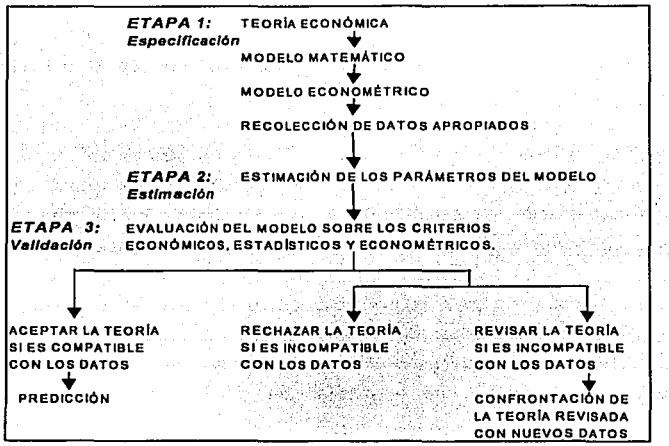


Figura 1.3 Pasos de la investigación econométrica.

1) Etapa de Especificación. Por *especificación* se entiende, de acuerdo a lo visto en la sección anterior, al hecho de "expresar una teoría económica en términos matemáticos"; por tanto, la teoría económica en general es la base que sirve como guía en esta etapa. En ella se encuentran orientaciones sobre qué variables pueden resultar relevantes para explicar un determinado fenómeno y si la influencia de cada variable explicativa es positiva o negativa, así como orientaciones respecto a su magnitud o los límites de variación de esa magnitud. Sin embargo, en muy pocas ocasiones la teoría económica explica la forma precisa de la relación funcional entre las variables para poder

formular adecuadamente el modelo matemático, siendo este un trabajo casi exclusivamente econométrico. También se debe tener presente que las teorías se formulan normalmente en países desarrollados y al aplicarlas a los países en vías de desarrollo es necesario hacer modificaciones de acuerdo a la complejidad de la teoría.

Adicionalmente, para que el modelo quede completamente especificado (pasar de un modelo matemático a un modelo econométrico) deben establecerse hipótesis estadísticas sobre el comportamiento de las variables aleatorias, así como sobre los demás elementos que aparecen en el segundo miembro de la ecuación correspondiente.

Finalmente, es necesario disponer de los datos apropiados, esto es de observaciones sobre todas las variables que aparecen en la especificación del modelo propuesto.

2) Etapa de estimación.[13] La etapa de *estimación* consiste en la obtención de valores numéricos de los coeficientes del modelo econométrico. Es necesario seleccionar el método de estimación apropiado, teniendo en cuenta las implicaciones de esta elección sobre las propiedades estadísticas de los estimadores de los coeficientes.

Los métodos econométricos involucran desde sencillas técnicas estadísticas para la obtención de los parámetros de regresión, como es la de los mínimos cuadrados ordinarios (MCO); hasta técnicas mucho más complejas de estimación para modelos multiecuacionales, como lo es el método de información completa de máxima verosimilitud (MICMV) [14]. La elección de las técnicas de estimación es función de una serie de factores, entre los cuales se pueden mencionar la magnitud del modelo en cuanto al número de variables y ecuaciones y la necesidad que se tenga de trabajar con la forma estructural o la forma reducida.[15]

[13] En los modelos multiecuacionales hay que determinar, como un paso previo a la estimación de los parámetros de una ecuación dada, si es posible estimarla, lo cual se logra con la llamada *identificación del modelo*. El problema de la identificación, trata de determinar los valores de los parámetros estructurales a partir de los coeficientes de la denominada *forma reducida*, a la cual se llega a través de operaciones algebraicas. De acuerdo al resultado obtenido (ecuación exactamente identificada, sobreidentificada o no identificada), da pie para la aplicación de otros métodos de estimación como el de mínimos cuadrados generalizados; por tanto van ligados estos dos pasos en las ecuaciones simultáneas.

[14] Técnicas que se puede consultar en cualquier libro de econometría que incluya modelos multiecuacionales.

[15] La *forma estructural* está constituida por una o varias ecuaciones estructurales o de comportamiento y dos o más variables endógenas por expresión, sin manipulaciones algebraicas. En cambio, en la *forma reducida* del modelo (que es obtenida de la forma estructural), las variables endógenas se encuentren despejadas y son función únicamente de las variables predeterminadas; por tanto se tendrá una ecuación de la forma reducida por cada variable endógena del modelo.

3) Etapa de Validación. En la etapa de validación (algunos autores subdividen esta etapa en verificación y predicción) se evalúan los resultados obtenidos, sobre la base de los *criterios económicos*, *estadísticos* y *econométricos*.

En los *criterios económicos* a priori, se trata de comprobar si las estimaciones de los parámetros del modelo tienen los signos y magnitudes esperados, es decir, si satisfacen las restricciones impuestas por la teoría económica. Si los coeficientes estimados no se ajustan a los que se han postulado, el modelo debe ser revisado o rechazado. Esto se estudia más a fondo si se construye el modelo con el fin de hacer un "análisis estructural", que es uno de los fines de la econometría.[16]

Desde el punto de vista *estadístico*, es conveniente examinar: 1) la proporción de variación de la variable dependiente ante cambios de las variables independientes y, 2) la dispersión o amplitud de cada coeficiente estimado alrededor del parámetro verdadero sea suficientemente estrecha para proporcionar confianza en las estimaciones. Es la parte de la inferencia estadística llamada ``prueba de hipótesis''.

Los criterios *econométricos* se centran tanto a las pruebas referidas a la especificación, como las pruebas que se hacen para satisfacer las suposiciones del modelo de regresión básico. Particularmente en el modelo de regresión lineal simple, las que se refieren al término de error aleatorio. Para ecuaciones con muestras grandes hay una gran cantidad de pruebas para identificar y resolverlo la violación de los supuestos básicos del modelo de regresión; aquí es quizás, donde han habido mayores aportaciones econométricas.

Dentro de la etapa de validación se efectuará la evaluación de la *capacidad predictiva* del modelo estimado. Este paso puede decirse que es el que establece la "contrastación empírica del modelo" (frente a la contrastación estadística), pues una vez conocidos los resultados de la realidad se comparan con los previstos, lo que permite aceptar o rechazar definitivamente el modelo estimado. Una forma de establecer si el modelo es adecuado para predecir sería reservar los datos de uno o dos periodos finales de las observaciones de las variables (no tomándose en cuenta para la estimación), realizar la predicción y comparar posteriormente los valores de la variable endógena con los datos que se habían reservado.

[16] Los otros dos fines son la *predicción* y la *evaluación de políticas*, que se verán en la siguiente sección.

A.4. FINES DE LA ECONOMETRÍA.

La econometría, como cualquier rama del conocimiento científico, tiene por objeto la descripción de lo ocurrido en el pasado para que una vez explicado de la forma más precisa posible, facilite la previsión de lo que ha de acontecer en el futuro y así permitir la toma de decisiones en una situación mucho más favorable ante hechos futuros inciertos. En su inicio la *previsión* de magnitudes económicas fue la finalidad práctica básica de la econometría. Actualmente se le añaden otros dos fines, que suelen conocerse en los textos econométricos como *análisis estructural* y *evaluación de políticas*. Cualquier estudio econométrico puede tener uno, dos o todos estos propósitos.

1) Análisis estructural. Este primer propósito se ha de entender como, el uso de un modelo econométrico estimado para efectuar la medición cuantitativa de relaciones económicas. Esto implica responder a preguntas tales como ¿qué tan sensible es la variable estudiada a los cambios que experimentan las otras variables?. También permite la comparación de teorías contrarias sobre un mismo fenómeno. El análisis estructural representa el propósito "científico" de la econometría: comprender los fenómenos del mundo real mediante la medición cuantitativa, prueba y validación de las relaciones económicas. Cuando hipótesis teóricas y resultados de la estimación coinciden, se dice que el modelo confirma empíricamente la teoría económica.[17]

En concreto, para interpretar adecuadamente ciertos coeficientes (o ciertas combinaciones de coeficientes), hay tres maneras esenciales de hacerlo en el análisis estructural, por medio de los resultados de la *estática comparativa*, las *elasticidades* y los *multiplicadores*.

2) Previsión. Uno de los objetivos más importantes es la previsión o predicción, que es la aplicación de un modelo econométrico estimado para determinar anticipadamente, los valores de ciertas variables fuera de la muestra de los datos realmente observados. Aunque normalmente se realiza un pronóstico sobre otros tiempos o lugares a futuro, también se puede aplicar la previsión a un periodo anterior a la muestra. Con frecuencia los pronósticos son la base para la toma de decisiones, por ejemplo, la compra de materias primas y el empleo de trabajadores adicionales en una

[17] Entendiéndose como modelo correcto aquel que, de acuerdo con los criterios teóricos, mejor se acomoda a la realidad.

empresa, puede apoyarse en una predicción de que las ventas se incrementarán durante los dos trimestres siguientes.

La predicción está muy relacionada con la evaluación de políticas a cualquier nivel, ya sea a nivel empresa o gobierno. En general se supone que la predicción es verificable en el sentido de que hay resultados que la refutarían o la validarían. La confianza de la predicción dependerá de:

- El horizonte de predicción
- La constancia de los valores paramétricos estimados a lo largo del horizonte de predición.
- La calidad de las estimaciones de los parámetros del modelo.
- Que el modelo utilizado sea el apropiado y en particular que esté especificado correctamente.

un análisis de predicción es fundamental para hacer cualquier estudio de política económica gubernamental.

3) Evaluación de políticas. Este fin se refiere a una situación en la cual, los encargados de la toma de decisiones deben elegir una política denominada ``plan'', entre un conjunto de alternativas dado. Un enfoque que ha sido muy utilizado, consiste en el uso del modelo estimado para simular los efectos de las diferentes políticas, o en su defecto, de la misma política en condiciones ambientales diferentes. Otro aspecto para la aplicación de este fin, es dividir en políticas de corto a largo plazo; por ejemplo, en el caso de modelos de macroeconómicos, la política de corto plazo se ocupa de la estabilización de la economía dentro de un periodo de uno o dos años. En cambio, la política de largo plazo se encarga del patrón de crecimiento sobre periodos más largos.

En cualquier caso se tiene que la evaluación de políticas está íntimamente relacionada con la predicción, la cual permite analizar los resultados ante diferentes combinaciones de valores de las variables. Combinaciones que será la expresión de los efectos de las diversas políticas alternativas sobre las que la econometría facilita anticipadamente sus resultados, optando por aquélla que sea más adecuada. De hecho, la mayoría de los métodos de evaluación de políticas se basan en un tipo de predicción específico.

B. LA TEORÍA DE LA DEMANDA.

B.1. CONCEPTO DE DEMANDA.

En general la **demanda** indica qué cantidad de un bien o servicio está dispuesto a comprar un consumidor a diferentes niveles de precio.

En economía, según se hable de decisiones microeconómicas o macroeconómicas,[18] se piensa en dos clases de demanda: la *demanda de mercado* o la *demanda agregada*.

Para entender la demanda de mercado hay que comprender que el mercado no es un lugar, el **mercado** es un proceso mediante el cual se ponen en contacto a personas o empresas para intercambiar bienes y servicios por dinero, efectuándose la negociación cuando se establece un acuerdo en el *precio*. El resultado de la interacción de compradores y vendedores es fundamental para la determinación de precios en los bienes o servicios.

La **demanda de mercado** se puede especificar para toda la industria (por ejemplo de calzado) o para una empresa en particular. Por otra parte, la **demanda agregada** representa la cantidad total de bienes y servicios que los diferentes sectores de la economía están dispuestos a gastar en un determinado periodo. Como es de suponer, ambas demandas están relacionadas.

Para comprender las relaciones de demanda de mercado o demanda agregada, es importante conocer la naturaleza de la demanda individual. En este nivel la demanda se determina por dos factores:

1) el valor asociado con la adquisición y uso del bien o servicio y,
2) la capacidad para poder adquirirlo.

[18] La *microeconomía* (teoría de lo pequeño) estudia el comportamiento de los agentes económicos en mercados específicos; a su vez la *macroeconomía* (teoría de lo grande) estudia los fenómenos económicos en su conjunto (lo que se conoce como *agregado*), permitiendo obtener conclusiones aproximadas sobre el crecimiento económico de un país o región. Esta no es la única clasificación ya que la teoría económica también se suele dividir en otras ramas; por ejemplo, una corriente moderna la divide en "teoría de los precios" (por la importancia del mismo como se verá más adelante) y en "teoría del ingreso". Sin embargo, la primera clasificación es más representativa para los conceptos que se manejan enseguida.

Ambos requisitos son básicos para una demanda individual efectiva; el deseo de compra sin el dinero disponible, puede conducir a una necesidad pero no a una demanda.

Existen **dos modelos básicos de demanda individual**. Uno es conocido como la *teoría del comportamiento del consumidor*, que se relaciona con la demanda por productos de consumo personal. Este modelo es apropiado para analizar la demanda individual de bienes y servicios que satisfagan directamente las necesidades del consumidor (demanda directa).

En el otro modelo la demanda individual no está relacionada directamente con el consumo personal final, sino más bien se deriva indirectamente de este; es decir, se adquieren los bienes y servicios porque son insumos importantes en la manufactura y en la distribución de otros productos, no por su valor directo de consumo (demanda derivada); este modelo se conoce como *teoría de la empresa*. Ejemplos de estos bienes son la demanda de trabajadores o de equipos para la producción.

En los dos modelos de demanda individual, las relaciones fundamentales son esencialmente las mismas, independientemente de que difieran los productos y las características individuales que afectan la demanda.

B.2. FACTORES QUE INFLUYEN EN LA DEMANDA.

- **Función de la demanda.**

Los consumidores determinan la cantidad que desean comprar basándose en el precio de su producto, su ingreso y a otros factores que originan la **función de la demanda**. De esta manera la función de la demanda es la relación que existe entre la cantidad demandada y todos los factores (o variables) que influyen en la demanda (y que se explicarán más adelante). En forma general esta función se puede expresar como:

Cantidad del producto X demandado = Q_x = f(precio del producto X demandado, grado de necesidad del bien, ingreso de los consumidores, precio de bienes sustitutos y complementarios, expectativas en cambios de precio, gustos, etc.).

esta función generalizada es solo una lista de variables que pueden influir en la demanda (y que deberá ser especificada de acuerdo a la naturaleza de la relación existente entre la

cantidad demanda y cada una de las variables independientes por medio de una ecuación). Para los economistas el principal determinante de la demanda es el precio ya que a un nivel básico la demanda se expresa como la relación entre el precio y la cantidad. Así, basándose en el precio, P, y la cantidad, Q, los economistas han establecido la **ley de la demanda** la cual señala que: *al disminuir el precio, la cantidad demandada aumenta, al aumentar el precio la cantidad demandada cae*; dicha relación se puede expresar como:

- P	+ Q
+ P	- Q

esta relación indica como varían el precio y la cantidad, en forma inversamente proporcional.

- **Curva de la demanda.**

Este comportamiento se representa gráficamente a través de la **curva de la demanda** o *programa de la demanda*, que es la parte de la función de demanda que expresa la relación existente entre el precio del bien y la cantidad demandada, *manteniendo constantes* los efectos de todas las demás variables (factores) que determinan la demanda, es lo que en Economía se conoce como efecto *ceteris pari bus*.

Para construir la gráfica, el eje horizontal mide las unidades demandadas y el eje vertical el precio por unidad demandada del bien.[19] En la figura 1.4 se ilustran tres niveles de precios y las respectivas cantidades que los consumidores pueden comprar. Por ejemplo, para un precio alto, P_2, la cantidad que se adquiera será baja, Q_0; conforme el precio disminuya a P_1 y P_0, la cantidad que se compre aumentará a Q_1 y Q_2 respectivamente. Como se puede observar, a pesar de que se le llama curva de demanda, a menudo lo que se obtiene es una línea recta.

[19] Ordinariamente se debería especificar la variable dependiente (cantidad demandada) en el eje vertical y la variable independiente (precio) en el eje horizontal; pero la práctica de colocarlas de manera inversa se originó con la teoría de mercados competitivos desde hace muchos años.

Los productores reciben información de la demanda de los consumidores mediante el movimiento de las existencias (cantidad de bienes comprados), y mediante los cambios en los precios de sus productos. A la vez siempre existe un precio lo suficientemente alto para restringir la cantidad demandada de un bien a la cantidad disponible para controlar su existencia.

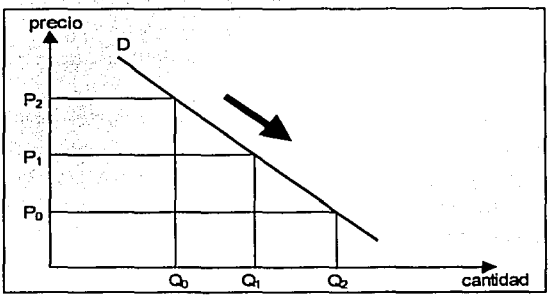


Figura 1.4 Movimiento sobre la curva de la demanda.

- **Factores de la demanda.**

A un nivel general, los economistas han determinado unos **factores que influyen directamente en la demanda**, que ocasionan el desplazamiento de la curva de la misma. En este caso, la magnitud y la dirección del desplazamiento de la curva dependerá del efecto que tengan sobre la demanda el cambio en tales factores. A continuación se describen para entender como pueden afectar a la curva de la demanda.

A) Ingreso del consumidor.

La demanda de la mayor parte de bienes y servicios tiende a aumentar o desplazarse hacia la derecha, conforme aumenta el ingreso monetario de las personas. Sin embargo, en este caso el desplazamiento de la curva de la demanda no sólo depende del aumento o disminución del ingreso, sino del tipo del bien del que se esté hablando. Los bienes se clasifican como *normales, superiores, inferiores o neutros*, de acuerdo al

comportamiento de la demanda y a partir de los cambios en el ingreso de los consumidores.

Se conoce como **bien normal** cuando a medida que aumenta el nivel de ingreso de los consumidores, aumenta su demanda por el bien, y viceversa.

Un **bien superior** *o de lujo* es también un bien normal, pero tiene la particularidad de que el aumento de la cantidad que los consumidores quieren comprar es más que proporcional al aumento del ingreso. Es el caso de un automóvil de marca.

Se considera un **bien inferior** cuando al aumentar el ingreso disminuye la cantidad que se desea consumir, es decir, disminuye la demanda. Dentro de esta categoría se consideran los productos básicos; por ejemplo un consumidor que podía comer principalmente frijoles, al aumentar su ingreso es muy seguro que disminuya su consumo y aumente el de carne o pescado.

Un **bien neutro**, como es de suponer, es aquel que cuando cambia el ingreso no se altera la cantidad consumida.

Los bienes cuya demanda varía directamente con el ingreso, son los bienes normales y superiores. Cabe aclarar que la clasificación de estos bienes depende no solamente del ingreso, sino de los *gustos de las personas*; lo que para alguien podría ser un bien normal, para otra persona será un bien inferior.

B) Cambios en los precios de bienes o servicios sustitutivos y complementarios.

Dos **bienes** son **sustitutivos** o *sustitutos* entre sí cuando al aumentar el precio de uno aumenta la demanda del otro. Es el caso de la margarina que aumentará su demanda si el precio de la mantequilla aumenta, desplazando la curva de la demanda de margarina hacia la derecha; lo contrario sería cierto si el precio del bien sustituto disminuye. Por tanto, la magnitud del cambio depende del grado de sustitutos entre los bienes.

Se dice que dos **bienes** son **complementarios** cuando se consumen conjuntamente, en la compra de uno indiscutiblemente se adquirirá el otro. Como ejemplo

está el uso de automóviles que requieren la gasolina, si aumenta esta es muy probable que disminuya el uso de coches.

Estas clasificaciones dependen de las preferencias de los consumidores, por ejemplo, para las personas que toman el café con crema verá estos productos como bienes complementarios, pero para quien lo toma solo no tendrá necesidad de comprar este suplemento.

C) Cambios en los gustos o preferencias.

Cuando las preferencias de los consumidores se pueden modificar por la publicidad, la promoción o la moda, significa que se demandará mayor cantidad de un producto, por el contrario un cambio desfavorable en los gustos del consumidor provocará una disminución en la cantidad demandada. Es el caso de la ropa, que cambia su demanda continuamente por la moda.

D) Cambios en las expectativas de los precios y en los ingresos.

Cuando los consumidores suponen que a futuro los precios pondrán incrementarse, buscan comprar en ese momento más de ese bien. Por otra parte si los demandantes prevén ganar más a futuro, les llevará a comprar menos en el momento actual, ya que se consideran con la posibilidad de absorber los nuevos precios. Lo contrario también será cierto, para precios e ingresos menores esperados en el futuro.

Estos factores que influyen en la demanda, hay que tomarlos como una base para el cálculo de una demanda particular que podrá contemplar variables no consideradas aquí. Por ejemplo para la demanda de una empresa especifica, las variables que representan las acciones de los competidores, como ***gastos en publicidad*** o el ***precio competitivo***, recibirán mayor peso.

Si la demanda que se desea estudiar es la demanda de mercado, el ***tamaño de la población*** es otro factor importante, el cual puede ser causado por: incremento demográfico, mejoras en los medios de transporte, influencia de la publicidad y la

promoción, grado de preparación escolar, principalmente. Para obtener la curva de demanda de mercado, simplemente se tendrían que sumar las demandas individuales.

En el ámbito macroeconómico la mecánica es similar. La *demanda agregada* es la suma del gasto de los consumidores y de otros agentes; depende principalmente del nivel de precios, de la política monetarista y fiscal y, de otros agentes.

B.3. ELASTICIDADES.

El concepto de elasticidad es análogo a "sensibilidad", el cual es muy utilizado por los economistas. La teoría económica estudia los efectos de los cambios entre las variables y la elasticidad proporciona la sensibilidad de la cantidad demandada de un bien ante una variación en algunos de los factores que intervienen en la demanda. Por ejemplo, la elasticidad proporciona información acerca del comportamiento de productores y consumidores en los diferentes mercados.

La *elasticidad de la demanda* se puede definir como el porcentaje de cambio en la cantidad demandada, que resulta de un cambio de 1% en el valor de una de las variables que determinan la demanda. Para calcular la elasticidad de una variable en particular X, se tiene la siguiente ecuación:

$$\text{elasticidad}\,(\epsilon_X) = \frac{\text{porcentaje de cambio en } Q}{\text{porcentaje de cambio en } X} = \frac{\Delta Q}{\Delta X} \cdot \frac{X}{Q} \qquad 1.3$$

donde:

Q = cantidad demandada
X = cualquier variable
ΔQ = incremento en la cantidad
ΔX = incremento en la variable independiente

Esta ecuación permite medir la sensibilidad de la demanda ante variaciones en cada una de las variables independientes. Sin embargo, es de particular interés medir la *elasticidad del precio, elasticidad del ingreso*.

La **elasticidad precio de la demanda**, ϵ_p, se define como el cambio porcentual de la cantidad demandada originado por cada 1% de cambio en el precio. En virtud de la ley de la demanda, se sabe que el precio y cantidad tienen una relación inversa, por lo que el valor de la elasticidad precio de la demanda siempre tendrá signo negativo. Para ciertos productos esta relación es alta, y para otros baja.

Se pueden dar diferentes tipos de elasticidades en una curva de la demanda que son:

a) Una *demanda es **elástica*** cuando la cantidad demandada es muy sensible a cambios en el precio, entonces el coeficiente resultante será mayor a uno, esto es, a un determinado cambio porcentual en el precio le corresponde un cambio porcentual mayor en la cantidad demandada. Una implicación importante es que si bajara el precio, el gasto del consumidor aumentaría, ya que la cantidad demandada subiría más que proporcionalmente a la disminución del precio.
b) Una *demanda es **inelástica*** cuando la cantidad demandada es muy poco sensible ante cambios en el precio. En este caso la elasticidad de la demanda será menor a uno, es decir, el cambio porcentual en la cantidad demandada es menor al cambio porcentual en el precio. Si el precio disminuyera la cantidad demandada aumentaría más que proporcionalmente al cambio en el precio, por lo que el gasto del consumidor será menor. (Fig. 1.5)
c) De lo anterior se desprende que una demanda cuya elasticidad precio es igual a uno, implica que el cambio porcentual del precio y de la cantidad son iguales (aunque con signo contrario) de tal forma que el gasto del consumidor se mantiene constante a cualquier nivel de precios. En este caso se habla de ***elasticidad unitaria***.
d) Se tiene dos casos extremos: cuando la cantidad demandada responde infinitamente a cambios en el precio, se habla de una demanda ***perfectamente elástica***; y cuando la cantidad demandada no responde ante cambios en el precio se tiene una demanda ***perfectamente inelástica***.

Se hace la observación de que todas las curvas de demanda en forma de línea recta, excepto las perfectamente elásticas o las perfectamente inelásticas, están sujetas a elasticidades variantes en distintos puntos sobre la curva, esto es, cualquier curva de demanda lineal será elástica a algunos niveles de producción, pero inelástica en otros. A medida que la curva de la demanda se aproxima al eje vertical, la proporción P/Q (de la

ec. 1.7) se aproxima al infinito y ϵ_p también se aproxima al infinito negativo. A medida que la curva de demanda se aproxima al eje horizontal, la proporción P/Q se aproxima a cero. En algún punto a lo largo de la curva de la demanda $((\Delta Q/\Delta P) \times (P/Q) = -1.0)$, se tiene el punto para la elasticidad unitaria.

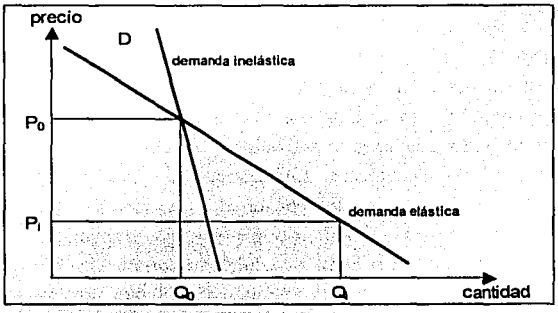


Figura 1.5 Curva de la demanda elástica y de la demanda inelástica.

Hay varios factores que influyen para que la elasticidad precio de la demanda sea alta para un producto y baja para otro; siendo los más importantes:

- Destaca la *disponibilidad de bienes sustitutos* que existen para el bien. Por ejemplo para la sal no existen sustitutos cercanos y la cantidad demandada mostrará poca respuesta ante cambios en el precio, en este caso el valor de la elasticidad tenderá a ser menor que 1; para otros bienes como los duraznos aunque sean muy deseables, si aumenta su precio va a caer mucho la cantidad demandada puesto que hay otras frutas que se pueden consumir. Por tanto la elasticidad precio será mayor a medida que existan mas sustitutos.
- La *proporción del ingreso* que se gasta en un bien, es otro factor que afecta la elasticidad precio de la demanda. La demanda será muy elástica para un bien si una parte considerable del ingreso se destina a ese bien; por otra parte, la demanda de productos que representen una fracción mínima del ingreso, aun cuando suba el precio, no será tan sensible a los cambios en el precio. En consecuencia, la elasticidad de la demanda será muy alta en bienes importantes y

más baja en bienes de menor importancia. Por ejemplo en la elasticidad de la demanda de cerillos, si su precio sube el doble, se consumirá la misma cantidad que antes, ya que no afecta mucho al presupuesto; a diferencia de la demanda de automóviles.

- Otro factor que influye en la sensibilidad de la demanda ante cambios en el precio es el *grado de necesidad* de un bien. Por ejemplo, la insulina para los diabéticos tendrá una demanda muy inelástica ya que es un bien indispensable lo que hace que la cantidad que adquiera no cambie mucho al aumentar el precio de este medicamento.
- *El tiempo* es otro factor a considerar ya que entre más largo sea el periódo estudiado, más elástica será la demanda e vista de que el consumidor tendrá más tiempo para ajustarse y cambiar su patrón de consumo ante cambios en el precio. Es el caso de la demanda de energía eléctrica doméstica que al elevar su precio, a corto plazo tendrá un efecto ligero pues los consumidores al principio serán muy cuidadosos en su uso y su demanda será relativamente inelástica, pero a largo plazo la demanda será más elástica.

Para un productor es importante conocer la respuesta de la cantidad demandada de las curvas de la demanda para fijar los precios de sus productos y maximizar sus utilidades. Si el bien que produce tiene una demanda elástica al aumentar el precio, su ingreso disminuirá; en cambio, si su demanda es inelástica aumentará su ingreso aunque incremente el precio del producto.

La **elasticidad ingreso de la demanda**, ϵ_I, se define como el cambio porcentual en la cantidad demandada originado por el cambio de 1% en el ingreso. La elasticidad ingreso proporciona información para determinar de qué tipo de bien se habla, de acuerdo a la clasificación dada en el ingreso del consumidor.

El ingreso y la cantidad demandada se desplazan en la misma dirección, es decir se encuentran directamente relacionados, por tanto ($\Delta Q/ \Delta I$) y ϵ_I son positivos. Esto no se mantiene para dos *bienes inferiores* en los que, como se dijo, al aumentar el ingreso disminuye la demanda, provocando un desplazamiento de la curva a la izquierda, y la elasticidad será negativa. De manera inversa, los bienes cuya demanda se encuentra positivamente relacionada con el ingreso, son los *bienes superiores o normales*, en donde al aumentar el ingreso aumenta la demanda, desplazando la curva a la derecha.

Para los oferentes es muy importante conocer la elasticidad ingreso del bien que producen, ya que pueden hacer ajustes a las caracteristicas del producto en épocas de auge o recesión. Por ejemplo, en una recesión la demanda de los bienes considerados como normales cae, mientras que los bienes considerados como inferiores se benefician debido al incremento de la demanda de su producto.

La mayoría de los productos presentan una elasticidad positiva respecto al ingreso, sin embargo, la magnitud del coeficiente de elasticidad también es importante. Por ejemplo, supóngase que el valor de ϵ_I para un bien en particular es 0.3 lo que indica que un aumento de 1% en el ingreso causaría que la demanda por ese bien aumente sólo 3/10 de 1%, en consecuencia el producto no estaría manteniendo su importancia relativa en la economía. Otro bien cuya elasticidad ingreso fuera de 2.5, implicaría que la demanda aumenta 2 1/2 veces más rápido que el ingreso. Se concluye entonces que si $\epsilon_I < 1.0$ para un bien en particular los fabricantes de dicho bien no compartirán proporcionalmente los aumentos del ingreso nacional; en cambio cuando $\epsilon_I > 1.0$ la industria obtendrá una participación más que proporcional en los aumentos de los ingresos.

El ingreso puede medirse de muchas formas, ya sea sobre una base per cápita (por persona), por familias o, sobre una base agregada. El producto nacional bruto, el ingreso nacional y el ingreso personal disponible, se han usado como medidas en los estudios de demanda.

B.4 NÚMEROS ÍNDICE.

Los números índice han adquirido una gran importancia como indicadores de los cambios en la actividad económica o en los negocios principalmente, al ser más manejables que grandes volúmenes de datos y fáciles de entender.

En términos generales un **número índice** permite comparar dos magnitudes, ya sea a los largo del tiempo o a través de áreas geográficas diferentes. Existen clases específicas para la construcción de números índice, por ejemplo, están los números índice de precios, índices de calidad (relación temperatura humedad), índices de valor

(calificaciones de una escuela), índices de cantidad (índice de producción industrial) e índices sociológicos (coeficiente de inteligencia I.Q).

Para el presente trabajo se considerarán los *números índices de precios* ya que como se ha mencionado el precio es una variable de vital importancia en el análisis de la demanda. A nivel macro interesa la medición del nivel general de precios para ver su trayectoria (de diferentes bienes) a través del tiempo o para comparar su evolución con la de otros países.

- **Construcción de números índices.**

Los **índices de precios** reflejan el porcentaje de cambio en el precio de uno o más artículos de comercio en un *periodo dado*, de acuerdo a un *periodo base* de referencia. La base de un índice es el periodo de tiempo en el pasado (generalmente de un año o un mes), contra los cuales se miden los cambios. Para seleccionar la base de un número índice deben observarse dos reglas: 1) el periodo seleccionado hasta donde sea posible, debe ser de estabilidad económica, 2) debe ser reciente a fin de que las comparaciones no se afecten sin necesidad por cambios en la tecnología, calidad del producto, gustos, etc.

Existen distintas formas de calcular índices de precio. A un nivel básico, se puede construir un índice de precios para un artículo particular con sólo formar la proporción de el precio pagado en cualquier periodo dado y el precio pagado en el periodo base (que generalmente se toma como 100), y para que quede en términos de porcentaje se multiplica por 100, es decir,

$$\text{Indice de precios en el periodo t} = \left(\frac{\text{precio pagado en el periodo t}}{\text{precio pagado en el periodo 0 (base)}} \right) \times 100$$

el cual es conocido como *índice de precios simple*.

En nuestro país las medidas que más se utilizan como indicadores de precios de productos son el Índice Nacional de Precios al Consumidor (INPC) y Índice Nacional de Precios al Productor (INPP). En la elaboración de un índice de precios al consumidor se contemplan los precios de una canasta de bienes y servicios representativa de la compra

de un consumidor típico. Para el cálculo de un índice de precios al productor se utilizan los precios pagados al productor de los bienes antes de su comercialización, por tanto, no incluye precios de servicios.

El Banco de México tiene a su cargo la elaboración de los índices de precios al productor y al consumidor, para los cuales utiliza un método llamado de Laspeyres, que se puede expresar como

$$\text{Indice de precios año t} = \left(\frac{\text{gasto en la canasta de bienes en el año t}}{\text{gasto en la canasta de bienes en el año 0}} \right) \times 100$$

siendo el gasto la suma de los precios por las cantidades respectivas de cada artículo de la canasta. Por tanto, el índice de Laspeyres mantiene ponderaciones fijas para todos los años en que se calcula; las ponderaciones las determina el Banco de México en base a las encuestas de consumo-gasto que actualmente elabora el INEGI (Instituto Nacional de Estadística, Geografía e Informática)[20].

Algunas aplicaciones de los índices fueron originadas por el interés sobre determinados problemas surgidos en el estudio de la *inflación* la cual se puede definir como el incremento generalizado y sostenido de los niveles de precios. Los índices de precios se usan con frecuencia en la *deflacción* de valores monetarios para expresarlos en términos reales.

Los precios reflejan fielmente el valor del dinero de tal manera que si se multiplica el índice de precios para un grupo determinado de bienes y servicios, el valor del dinero se reduce a la mitad, por lo cual el poder adquisitivo del dinero se expresa como el recíporco del índice de precios.

De esta manera la **deflacción** consiste en corregir el efecto de esa pérdida de valor del dinero. Basándose en esta corrección surgen dos conceptos: el valor monetario o nominal y el valor real de la la unidad monetaria.

[20] Esta encuesta se conoce actualmente como *Encuesta Nacional de Ingresos y Gastos de los Hogares* (ENIGH). Ver nota 4 del capítulo tres.

Se llama **valor monetario** al que se presenta sin efectuar ninguna corrección, es como se encuentra en todos lados; en cambio, el **valor real** es el que corresponde a la unidad monetaria despues de dividirla entre el índice de precios general y de esta manera se le quita la inflación. A una serie de valores nominales se le denomina **valores a pesos corrientes** o de cada año, a una serie de valores reales se le llama **valores a pesos constantes**; en cada caso se debe hacer referencia del año que se tomó como base.

Para convertir cualquier variable nominal a real se usa la siguiente expresión:

$$\text{valor real} = \left(\frac{\text{valor nominal}}{\text{indice de precios}} \right) \times 100$$

En la práctica también se acostumbra deflactar otras variables como es el caso del ingreso.

Otra expresión muy usada en economía, es la **tasa de crecimiento por periódo**, que se puede formular como:

$$\begin{matrix}\text{tasa de crecimiento de variable real} \\ \text{(periodo base a periodo dado)}\end{matrix} = \left(\frac{\text{variable periodo 1 - variable año 0}}{\text{variable año 0}} \right) \times 100$$

la cual indica el crecimiento porcentual real de una variable durante el periódo especificado que puede ser, por ejemplo, los primeros meses del año en curso (como el crecimiento de enero a junio) o comparar con el año anterior (como el crecimiento de junio del 2000 con junio del 2001), etc. También se pueden usar los valores nominales, pero es claro que interesan más los valores reales que no muestran el efecto inflacionario; además también se utiliza para medir los cambios de otras variables que no sean índices de precio, como sería el caso de la tasa de variación de las ventas, la producción, etc.

C. DEMANDA DEL CONSUMO DE LECHE EN MÉXICO.

C.1 DESCRIPCIÓN DEL BIEN Y SUS DERIVADOS.

Se ha dicho que "*de todas las sustancias alimenticias que existen, sólo la leche tiene como única razón de existir, precisamente la de servir como alimento*",[21] dado que la leche sirve para alimentar tanto a un ternerito, como al hombre. La leche de la madre es insustituible ya que por sí sola es capaz de ir formando los músculos, nervios y sangre de las crías.

Existen tantas clases de leche como especie de mamíferos, incluso la de un mismo animal tiene diferente composición en los distintos periodos de la crianza; por ejemplo la leche de cabra -cuyo recién nacido crece más rápidamente que el ser humano- es mucho más rica en proteínas que la leche de mujer.

El periodo conocido como *lactancia* en el cual las crías se alimentan de la leche materna, tiene una duración fija. Una vez concluida esta etapa, sólo el hombre incorpora a su dieta la leche de otros animales o sus derivados. A pesar de ser empleada la leche de otras especies, la de vaca es la que más consume el hombre.

- **Composición.**

La leche de vaca se compone principalmente de agua y de sólidos como se muestra en la tabla 1.1. Los sólidos se dividen en cuatro grupos principales: los lípidos o grasas, las proteínas, la lactosa o azúcar y las cenizas o sales. Cabe aclarar que la variación de estos componentes de la leche está determinada por la raza, el estado nutricio de la misma, la alimentación que recibe, la estación del año y la región en que viven los animales; por eso se pone una variación máxima y mínima. A pesar de esta diversidad, la mayoría de veces la leche de vaca contiene el 13% de lípidos, 3.3.% de proteínas y 4.8% de lactosa, además aporta 60 kilocalorías por cada 100 ml.

[21] Citado en: LOMELÍ, MTRO. JUAN VEGA: "*Manejo y Conservación de Productos Lácteos*", Curso, Universidad Ibero, México, 1996.

Componente	Mínimo	Máximo
Agua	84	89
Sólidos	10.6	17.9
Lípidos	2.6	8.4
Proteínas	2.4	6.5
Lactosa	2.4	6.1
Cenizas	0.6	0.9

Tabla 1.1 Composición de la leche de vaca (g/100 ml).

Una de las sustancias proteínicas es la *caseína* la cual se forma cuando se coagula la leche. La caseína forma aproximadamente el 70% de la proteína total, el otro 30% se encuentra en el suero.

A pesar de contener vitaminas a un nivel bajo, la más abundante es la riboflavina (de otras 10 vitaminas) que está en proporción de 1 mg. por cada medio litro de leche, lo cual cubre una buena parte de la recomendación para los adultos que es de 1.5 a 1.6 mg. diarios. La vitamina A que es la menos abundante, se pierde cuando se descrema la leche, por lo cual se recomienda que se restituya.

La lactosa es un tipo de azúcar muy especial que sólo existe de manera natural en la leche. También se encuentra en la leche un microbio que para nutrirse transforma la lactosa en ácido láctico provocando que la leche se agrie a temperatura ambiente, pero en el interior del estómago produce efectos saludables evitando el desarrollo de otros microbios que alterarían la salud.

Las sales contenidas en la leche están formadas por calcio (120), fósforo (90), sodio (50), potasio (140 mg/ml) y pequeñas cantidades de hierro. El calcio y fósforo pueden interferir si no guardan entre sí cierta proporción que en los primeros meses de vida la relación calcio-fósforo debe ser alrededor de dos, después basta con que sea de uno o un poco más. El primero, como se sabe, es importante para los huesos y el segundo es imprescindible para la regeneración del tejido muscular, además de ser un fortificante del sistema nervioso.

- **Inconvenientes.**

A pesar de ser un alimento con un gran valor nutricional, la leche no es un alimento completo; es pobre en hierro, vitamina C y D. Como aporta más potasio que sodio se debe tomar en cuenta cuando se tiene que restringir o elevar el consumo de ellos. Tampoco es una fuente importante de energía.

Aunque la lactosa que por sí misma no es perjudicial, tiene el inconveniente de que muchas personas no la toleran, favoreciendo algunos dolores del estómago. Para las personas intolerantes a la lactosa, está la opción de consumir yogurt o queso, incluso hoy en día se fabrican las leches "deslactosadas".

- **Derivados Lácteos.**

Los *derivados lácteos* más comunes son: el queso, la leche fermentada, la crema y la mantequilla.

Para muchas personas el queso es el derivado más atractivo, el cual se obtiene de la caseína por medio de un procedimiento de fermentación. Los quesos son en general de fácil digestión, principalmente los blandos o frescos. Poseen un gran valor nutrimental debido a la cantidad considerable de leche que interviene en su elaboración, que puede ser hasta de un 90% en proporción con el agua; conforme aumenta la dureza del queso se incrementan lo sólidos y baja la humedad.

La fermentación de la leche da origen a varios productos, el más conocido es el yogurt. Es una fermentación parecida a la del vinagre o alcohol, utilizando la lactosa.

De los lípidos se obtiene la crema y mantequilla; esta última contiene el doble de grasa que la crema, pero ofrece la ventaja de ser de fácil digestión.

Los inconvenientes por el consumo excesivo de los *derivados lácteos* es la grasa y su contenido de colesterol. En los quesos duros y semiduros, pero sobre todo en la crema y la mantequilla estos lípidos están mas concentrados que en la leche, incluso en la leche descremada casi están ausentes. Sin embargo, no presentan ningún peligro si el resto de la dieta no contiene colesterol o grasas.

Una aclaración importante es que no existe un requerimiento específico de ningún alimento puesto que "no hay alimentos indispensables". Por tanto, la leche y sus derivados pueden ser sustituidos por otros alimentos para las personas en edad adulta, no obstante, en la infancia sigue siendo muy recomendable.

- **Usos.**

Los diferentes componentes de la leche tienen múltiples usos, aunque en un menor porcentaje. Intervienen en el sector alimenticio, por ejemplo, para la elaboración de pan, helados, dulces, chocolates, complementos alimenticios y bebidas; también se utilizan en el sector industrial para adhesivos, zapatería, papelería, plásticos y productos farmaceúticos, entre otros.

C.1.1 Clasificación comercial.

Antes de pasar a la siguiente sección, conviene saber la clasificación que se hace de la leche. En general la leche que se ordeña y que ofrecen los ganaderos, es la **leche no procesada** o bronca; la leche que se vende al público es conocida como **leche procesada** o comercial. A su vez la leche procesada se puede clasificar de distintas formas, una clasificación es determinada por su *origen*, por el *tipo* y por *proceso* (Tabla 1.2)[22].

Por Origen	Leche y extensores o fórmulas lácteas.
Por Tipo (cantidad de grasa)	Entera, parcialmente descremada, semidescremada y descremada.
Por Proceso	Pasteurizada, ultrapasteurizada, en polvo, homogeneizada, reconstituida, deslactosada, evaporada, concentrada, doble concentración y fermentada.

Tabla 1..2 Clasificación comercial de la leche.

[22] Esta clasificación fue proporcionada por una empresa, pero no es rigurosa para efectos estadísticos macroeconómicos, como se verá más adelante.

Algunas consideraciones sobre las clases de leche son las siguientes:

- En la clasificación por origen se tiene que la fórmula láctea ya no es leche solamente, es grasa y una tercera parte es leche. Por ello puede darse a un menor costo, sólo se pide que se especifique en el cartón que es una fórmula láctea.
- En cuanto al tipo de leche, hay que tomar en cuanta que la grasa da un valor agregado a los productos. De esta manera la grasa que se elimina se puede usar, como se mencionó, para la elaboración de crema y mantequilla. Además también se puede vender a otras industrias, como la del chocolate.
- El proceso de pasteurización, también lo hacen los ganaderos debido a la importancia que tiene para la destrucción de los organismos patógenos; aunque no siempre se hace con todo el control de sanidad requerido. Una vez envasada la leche pasteurizada, necesita refrigeración y no dura más de tres días. En cambio la leche la leche ultrapasteurizada, se podría decir que se pasteuriza dos veces, por lo cual ya no necesita hervirse, pudiendo durar hasta tres meses sin refrigeración; aunque una vez abierto el envase, se debe refrigerar.
- Se hace la observación de que se puede combinar el proceso con el tipo, por ejemplo, puede haber una leche pasteurizada semidescremada.

Hay que tener presente que esta clasificación no todas las empresas la utilizan, por ejemplo, la leche deslactosada y la fórmula láctea, son procesos recientes que no manejan todas las empresas. A nivel gubernamental esto es de suma importancia ya que no se encuentran las estadísticas de estos productos por separado, en algunos casos coinciden y en otros no.

C.2 COMPORTAMIENTO DE LA ACTIVIDAD LECHERA EN MÉXICO.

C.2.1 Producción del Sector Lechero Nacional.

En la segunda mitad de la década de los ochenta, la producción nacional del sector lechero experimentó un marcado descenso con la llamada política neoliberal, que en general marginó al sector agropecuario en la asignación de recursos públicos. A partir de 1990 se toma al sector empresarial como eje de la transformación agrícola para volver competitivo al sector agropecuario en el contexto de globalización. Fue hasta el año de

1993 que se superó la producción de 1985 y a partir de 1995 se han registrado incrementos sostenidos.

Actualmente el sector lechero tiene un papel muy importante en la economía del país por los empleos que genera, además de ser un alimento básico. El sector lechero junto con la industria de productos lácteos, genera alrededor de 1.5 millones de empleos y contribuye al país con el 1.3% del Producto Interno Bruto (PIB).[23]

Si bien, la producción de leche se desarrolla en todo el país, los estados que encabezan la producción de leche fresca son: Jalisco, Durango, Coahuila, Veracruz, Chihuahua y Guanajuato, que abastecen el 59% de la producción nacional.

México es un país deficitario en la *producción de leche fresca*, lo que trae como consecuencia que el total requerido se complemente con la importación de leche en polvo descremada; situación que llegó a colocar a nuestro país como el principal importador de leche en el mundo. La mayoría de *leche descremada en polvo*, la traen del extranjero (Estados Unidos, Canadá, Nueva Zelanda y Australia principalmente). También hay importaciones considerables de leche condensada y evaporada; además de derivados de la leche dentro de los que se encuentran el yogurt, quesos duros, sueros, lactosueros, grasa butírica (mantequilla) y otros

A mediados de los noventa, la devaluación y crisis económica, así como una elevación de los precios de leche en polvo internacional, trajo como consecuencia una drástica caída en los niveles de importaciones del país. La dependencia que tiene el sector lechero mexicano del extranjero, se ve reflejado principalmente en:

- El proceso creciente de *agroindustrialización de lácteos*, que es el proceso mediante el cual se transforma la leche para obtener derivados, siendo los más importantes en el ámbito mundial, la leche condensada, evaporada, en polvo, los quesos y mantequilla. Este proceso lleva a una diversificación de la producción y presentación en los productos lácteos.
- El desarrollo tecnológico es superior en los países desarrollados, poniendo en desventaja a los que tienen escasa especialización y condiciones rústicas como es el caso de México.

[23] Indicador económico que indica el total de bienes y servicios que produce un país. Existen varios métodos para calcularlo.

- La producción excesiva de los países desarrollados, en contraste con la insuficiencia de producción nacional explica las grandes cantidades de importaciones de leche, en este caso de nuestro país.

El sistema lechero es sumamente heterogéneo ya que intervienen el aspecto económico, social, cultural y regional. Aún así se tiene una división muy marcada en dos grandes sectores: las *lecherías especializadas* y las *lecherías no especializadas*. Los grandes productores forman las lecherías especializadas, por tanto se encuentran ligados al sistema industrial, cuentan con una tecnología más avanzada, ganado de alta productividad y estrictos controles de sanidad. En contraste, en la lechería no especializada se encuentran los medianos productores que utilizan su producción con dos fines: por un lado, se tiene la pequeña industria artesanal (derivados lácteos); por otro lado, está la parte destinada a los recolectores-vendedores de leche que a su vez la llevan a procesadoras regionales o a empresas reconocidas. También se puede ubicar un grupo de pequeños productores que destinan su producción para el autoconsumo en forma de leche bronca o para la industria casera rústica que por medio de un intermediario la venden al consumidor final.

A pesar de esta situación, dentro del mercado de leche fluida envasada hay un porcentaje de participación muy significativa de las principales empresas industrializadoras y ganaderas el cual se compone por: Grupo LALA con un 20%, Alpura (que solo vende en el D.F.) con un 15%, Grupo GILSA el 13% y el porcentaje restante del 52% está disperso en las demás empresas industrializadoras lácteas.

C.2.2. Consumo nacional de leche.

Tomando en cuenta el total de la producción más las importaciones de leche y restando las exportaciones, se puede estimar el consumo de la leche de la siguiente manera[24]: en México, no obstante que la leche bronca no tiene generalmente un control sanitario suficiente, se consume en un porcentaje del 16.46%, la leche de abasto social forma el 10.23%, la leche para uso industrial (o comercial) lo consume el 73.14% de la población y finalmente un .17% lo constituye la leche fluida, evaporada y condensada que

[24] Las estimaciones las obtuve basándome en información de el documento especial de leche que editó la SAGAR (Secretaría de Agricultura, Ganadería y Desarrollo Rural). Ver bibliografía.

viene de las importaciones. Dentro de la leche comercial[25], el mayor consumo es en leche fluida con un porcentaje del 88.69%, el 8.51% se convierte en leche en polvo, evaporada y condensada y el resto del 2.80% lo constituye el consumo de derivados. A la vez, dentro del consumo de leche fluida el 55% corresponde a la leche pasteurizada-homogeneizada y el resto se reparte en leche ultra pasteurizada, rehidratada y pasteurizada.

Como se recordará, el mercado se ha diversificado con otros *tipos* de leche además de las mencionadas como son: reconstituida con grasa vegetal, deslactosada, fermentada, diversas fórmulas lácteas y en algunos casos leche adicionada con vitaminas y minerales; y con diversas *cantidades de grasa*, como es el caso de la leche entera, descremada y parcialmente descremada. Pero las estadísticas de estos productos no se encuentran.

Es importante señalar que la FAO (Organización de las Naciones Unidas para la Agricultura y la Alimentación) recomienda alcanzar un nivel de 182.5 lt por persona al año,[26] el cual nunca se ha alcanzado en México; como se aprecia en la figura 1.5, algunos años apenas se ha alcanzado un nivel de casi 120 lt por persona.

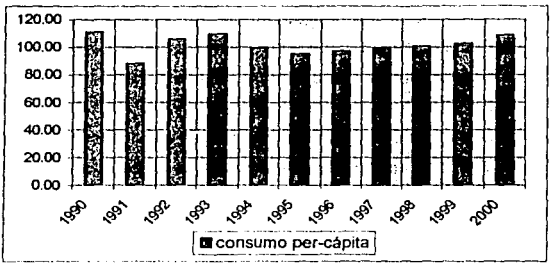


Figura 1.5 Consumo per cápita.

[25] Esta clasificación me la proporcionaron en la empresa A (ver sección C.3 de este capítulo), pero se pueden obtener de los *boletines de la leche* que publica la SAGAR, la cual se transformó a partir del 2000 en SAGARPA (Secretaría de Agricultura, Ganadería, Desarrollo Rural, Pesca y alimentación). Estos boletines han tenido varios cambios como se menciona en el capítulo tres.

[26] Esta cantidad se obtiene basándose en el indicador económico conocido como *Consumo Nacional Aparente* (CNA) que actualmente se calcula sumando la producción nacional, más las importaciones, menos las exportaciones. Cuando se divide entre la población se obtiene el llamado consumo *per cápita* (o por persona), el cual también se conoce como *disponibilidad del producto*.

Por tanto México se considera un país deficitario en el consumo de este alimento a pesar de ser un país "joven" con una media de población alrededor de los 19 años de edad. Esto se debe a que se tiene un mercado predominantemente urbano, pues de acuerdo con datos proporcionados por los programas sociales, la producción industrial está orientada hacia los estratos económicos de medianos y altos ingresos.

A pesar de que la leche pasteurizada tiene el precio más bajo en el mercado, todavía es demasiado caro en relación con los bajos ingresos de la población. No obstante el control que ha tenido el gobierno sobre el precio, este ha aumentado incluso de 2 a 3 veces por año en proporción del salario mínimo. Un hecho sobresaliente que afectó el precio de la leche sucedió a principio de 1995, en que incrementaron desequilibradamente los costos, lo cual junto con la situación del sector lechero en general y las restricciones que representaba el precio internacional (mucho más caro que el precio nacional); provocó que en marzo del mismo año las autoridades otorgaran un aumento en el precio oficial de la leche pasteurizada, en tanto que se liberó el precio de la leche ultrapasteurizada dejándose su fijación a las fuerzas del mercado, argumentando que es consumida en forma sobresaliente por la población de mayor ingreso. Meses después, se liberó también el precio de leche pasteurizada.

En cuanto a la leche de abasto social, tampoco puede llegar a toda la población de bajos recursos. Algunas causas son que el padrón de beneficiarios es limitado y que algunas personas no pueden cubrir la cuota semanal fijada para adquirirla.

En referencia al consumo de leche bronca puede ser limitado por el precio, ya que por lo menos en el Distrito Federal se vende casi al doble del precio que la leche pasteurizada.

C.2.3. Problemática y políticas del Sector Lechero.

Como se apreció en las secciones anteriores, la problemática del sistema nacional lechero es sumamente compleja. Algunos elementos que destacan en esta actividad son:

- **Políticas de precios**. En la búsqueda de la autosuficiencia alimenticia se ha creado una gran brecha entre productores y consumidores que tienen como punto de encuentro, el precio. Respecto a los precios promedio pagados al productor

(precios nominales[27]), se puede apreciar en la figura 1.6 que a partir de 1992 se han ido incrementando principalmente a partir de 1995, sin embargo, al ser deflactados (precios reales) se observa que no han sido de mucho beneficio estos aumentos para ellos, lo cual implica la inevitable elevación de los precios al consumidor, sobre todo ahora que se han liberado los demás precios de la leche. Además hay una gran diversidad de precios dependiendo de las regiones de comercialización, tipos de producción y orientación al consumo. Aquí podrían entrar los extensores que han tenido gran aceptación en algunos segmentos del mercado, por precio más bajo; en realidad los extensores fueron creados para las personas de medianos ingresos, solo que actualmente hay empresas que los venden casi al mismo precio de la leche, lo que puede ocasionar en un futuro que se comercialice menos leche y más extensores al no saber su diferencia.

- **Organización de productores.** La organización de los productores ganaderos es un gran problema debido a la gran heterogeneidad que existe entre ellos.
  Los productores de las *lecherías especializadas* se enfrentan a problemas económicos (eliminación de subsidios en la compra de granos que tenían a través de ASERCA[28], encarecimiento de leche importada, devaluación económica), tecnológicos (casi toda su infraestructura es importada), políticos (subsidios sólo al consumidor y no al productor, en otros países se otorga más bien al productor). La mayor parte de las utilidades las transfieren al exterior y deben pagar por el uso de marcas, patentes y procesos técnicos.
  Los productores de *lecherías no especializadas* tienen una notable desorganización de su producción, atraso tecnológico y baja inversión que, a su vez, trae como consecuencia su baja productividad, frecuentes enfermedades de los animales, una baja calidad en la leche y ganancias bajas.
- **Dependencia externa.** El sector lechero presenta una marcada dependencia externa económica y tecnológica en todas sus fases, desde la producción hasta el consumo. Otro punto importante es que de acuerdo al TLC se debe pensar a futuro en la liberación total de la producción de la leche, lo cual puede tener como consecuencia un enfrentamiento directo de la producción nacional contra la de los países desarrollados, principalmente de Estados Unidos, en un mercado bastante desigual; por lo que se deberá aumentar la calidad, la tecnología y la producción de la leche.

[27] El concepto de variable nominal y variable real se vio en la sección B.4.
[28] Apoyos y Servicios a la Comercialización Agropecuaria. Es un organo desconcentrado de la SAGAR.

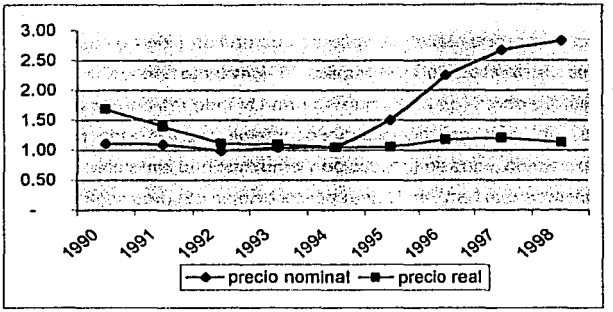


Figura 1.6 Precio promedio pagado al productor.

Como se aprecia, la problemática actual del sector lechero nacional es demasiado compleja, lo que hace suponer que la insuficiencia de la producción y su costo elevado no tienen una solución a corto plazo. Por ello la intervención del gobierno en el futuro de la actividad lechera es de vital importancia.

El Gobierno Federal consciente de la problemática que vive el sector, pero principalmente por el gran potencial productivo que representa, en 1996 instó a la Secretaría de Agricultura, Ganadería y Desarrollo Rural (SAGAR) para que diseñara un programa específico con el objeto de incrementar la producción de leche. De esta manera surge el Programa de Fomento Lechero que es concertado con los organismos de los productores e industriales, que se sustenta en cuatro puntos principales:

- Un procedimiento para evitar que los subsidios de leche en polvo en el mercado internacional disminuyan el precio de la leche nacional, se obtiene con los cupos libres de arancel.
- La implementación del programa Alianza para el Campo, que busca la modernización de la productividad y del hato ganadero, el mejoramiento genético, la asistencia técnica y la salud del animal.
- Establecimiento de una Norma de etiquetado, para eliminar la competencia entre leche y fórmulas lácteas.

- La liberación del precio de leche y productos lácteos, para evitar distorsiones en el mercado.

Esto trajo como consecuencia que las importaciones de leche en polvo que en 1990 representaban el 46.0% del volumen de la producción, representaran en 1998 únicamente el 15.8% del volumen de la producción nacional, lo que trajo como consecuencia que descendiéramos al tercer lugar de la importación de leche descremada en polvo.

C.3. APLICACIÓN DE LOS FACTORES QUE INFLUYEN EN LA DEMANDA DE LA LECHE.

En esta sección se presentan algunos aspectos sobresalientes de entrevistas realizadas a tres empresas. Las preguntas fueron encauzadas principalmente para saber si aplican o conocen la teoría económica existente del análisis de la demanda de un bien, y si utilizan una técnica estadística para calcular la demanda de la leche. La primera empresa es la paraestatal LICONSA, las otras dos empresas se denominarán empresa A y empresa B. Cabe mencionar que la información obtenida ayudará a identificar qué variables se incluirán en el modelo para explicar la demanda del consumo de la leche.

- **LICONSA (Leche Industrial CONASUPO).**

Esta empresa vende leche rehidratada (es decir, leche en polvo a la que se le agrega agua) y en polvo. Para ello cuenta con un padrón de beneficiarios que incluye a niños menores de 12 años cuyos papás ganan menos de 2 salarios mínimos. Actualmente proporciona 16 litros mensuales por niño beneficiario. La distribución de leche en polvo es a nivel nacional, que dependiendo del nivel geográfico, se distribuye de 2 a 3 veces por semana. La leche líquida se reparte solamente en el Distrito Federal, en el Estado, en Guadalajara y Oaxaca.

Tiene a su cargo un Programa de Abasto Social (PAS) para fusionar plantas de producción y programas de apoyo. La etapa de producción de este programa está integrada por 3 áreas:

- Requisición. De acuerdo al presupuesto, se *estima* a nivel nacional la producción para 1 año.
- Abasto. En base al padrón de beneficiarios, se encarga de la distribución (los beneficiarios son quienes las distribuyen por medio de organismo DICONSA).
- Finanzas. Libera el financiamiento.

En el área de requisición hace poco se elaboró un estudio con el objeto de apoyar a los pequeños y medianos productores de leche, para poner a su disposición diversos servicios que incrementen la producción y calidad de la leche (en base a centros de acopio que captan la leche y venden los insumos necesarios para su producción), entre otras cosas.

El estudio consistió en hacer *entrevistas* a los ganaderos para saber con qué recursos contaban (maquinaria, número de animales, alimento, etc.), asegurándose que si decían por decir ``cuento con 100 vacas'' fuera cierto y no tuvieran sólo 50. Una vez obtenidos los datos de todos los ganaderos se elaboraron dos modelos: uno que contemplaba todas las variables que representaban los gastos y otro que contemplaba todas las variables que representaban los ingresos de los ganaderos. Utilizando una hoja de cálculo, se sumaron cada una de las variables de las dos ecuaciones por separado hasta obtener una sola variable general que representó los gastos y otra que representó sólo los ingresos. Finalmente a este pequeño modelo se le aplicó el método de series de tiempo (Box and Jenkins) y se hizo la estimación del presupuesto para el siguiente año. Obtuvieron muy buenos resultados con estos modelos.

A pesar de que los modelos no va encaminado a calcular la demanda, es útil para saber que métodos matemáticos usan, además de ayuda para la selección de variables por la relación tan estrecha entre la oferta y la demanda. Ellos no necesitan un modelo de demanda ya que ésta supera la oferta que tienen y, como se dijo arriba, ya tienen un padrón de beneficiarios preestablecido.

Apoyándose en la información proporcionada, las variables sobresalientes son: ***el tamaño de la población***, ya que el registro total del padrón que tiene supera su producción y ***el ingreso de los consumidores***, pues se proporciona a las personas que ganen menos de 2 salarios mínimos.

- **Empresa A.**

La empresa A llegó a ser líder en la venta de leche a escala nacional. Las estadísticas para conocer su demanda en el ámbito nacional están a cargo de una agencia especializada; en cambio las estadísticas para su empresa las realizan ellos mismos.

En cálculo de la demanda, toman en cuenta tres escenarios de venta al año: leche fresca, en todos los tipos (es la que dura máximo 3 días), leche ultrapasteurizada, también en todos los tipos (es la que dura hasta 3 meses) y derivados de la leche.

Para medir sus ventas lo hacen de acuerdo a:

- Zonas,
- Marcas,
- Temporadas y,
- Número de espacios asignados en las tiendas.

Con esta información, calculan los *promedios* (o medias) actuales de ventas y los comparan con los promedios del año anterior.

La división por *temporadas* a su vez se subdivide en tres temporadas muy bien ubicadas: la temporada alta, que comprende los periódos de regreso de las vacaciones de Semana Santa, de verano y de invierno; la temporada baja se presenta precisamente en la Semana Santa, las vacaciones de verano y de invierno; y la temporada normal son los demás meses. No toman en cuenta puentes ni días festivos por variar en cada año.

El *número de espacios asignados*, es una variable muy importante pues es como su publicidad y casi se considera como activo. Es de vital importancia el número de fuentes de exhibición que se tengan en las tiendas independientemente de lo que vendan, es decir, el consumo promedio de sus productos lo toman en cuenta en tanto se debe de abastecer a las tiendas, pero si se pierde un espacio asignado representa una pérdida grave. Estos espacios los asignan en las tiendas y no siempre están en proporción con las ventas de los productos, es decir, puede ocupar una marca muchos espacios en las tiendas, pero esto no implica que sea por que vendan más, sin embargo puede influir para que se compre más.

Por otro lado es importante hacer la observación que toman en cuenta algunos de los determinantes clásicos de la demanda, pero no como variables para medir sus promedios, como se explica a continuación:

El ***ingreso de los consumidores*** no lo toman en cuenta ya que tienen ubicado el *mercado por familias*. Por tanto, elaboran sus productos para segmentos específicos del mercado. La clase económicamente alta consume la leche tanto fresca (pasteurizada) como ultrapasteurizada; las clases media alta y popular consumen la fórmula láctea, específicamente la consumen los niños y jóvenes. A la vez la leche entera es para niños de 4 a 11 años de edad y para ancianos. La leche light está ubicada para personas de 20 a 25 años de edad, principalmente amas de casa.

No consideran que los ***precios*** de sus productos se vean afectados por los precios de la competencia o por el precio internacional. Sobre el incremento del precio se hizo el comentario que al ser liberado el precio de la leche ultrapasteurizada, en el ámbito mundial se estaba vendiendo a 1 dólar, siendo que en México estaba casi a la mitad. Para poder fijar el nuevo precio se reunieron las principales empresas llegando a un acuerdo de incrementarlo poco a poco, de otra forma nadie les compraría. El precio lo determinaron ellos tomando en cuenta que la mayoría de insumos son extranjeros, como el cartón que lo traen de Noruega.

Tampoco toman en consideración el precio de los ***bienes complementarios*** como podrían ser los sabores artificiales, el café, la miel, el azúcar o los cereales.

En cuanto a los ***bienes sustitutos***, hay otro comentario que vale la pena incluirlo. En el año de 1995 empezó a disminuir el consumo de leche, por lo que se reunieron las empresas para saber el motivo, acordando cada una realizar un estudio por separado. Al volver a reunirse concluyeron que las causas principales de la disminución del consumo de leche fueron:

- Ya no se tienen los mismos hábitos alimenticios, cambió la idea de los consumidores de estar bien alimentados sólo con determinados alimentos.
- Las amas de casa prefieren el yogurt.
- A los niños también se les da yogurt y jugo.
- Aumentó el consumo de Yakult, bebidas saborizadas y cereal (que lo comen también con yogurt, no sólo con leche).

De los productos que han tenido más influencia para disminuir el consumo de leche consideran principalmente el yogurt y los jugos, que según la teoría serían los bienes sustitutos más importantes.

Respecto a los conceptos de ***elasticidad***, dijeron que a niveles más altos se está constantemente monitoreando el *precio*, pero para fines estadísticos no se considera, como se explicó arriba.

Finalmente consideran que el *clima* y las *temporadas de vacaciones* son los factores que más afectan a la demanda. Respecto al clima, es una variable que afecta sobre todo en la producción, pues pusieron como ejemplo la época de calor en que la vaca genera más grasa que leche. Sobre el consumo de leche en las vacaciones, está como ejemplo diciembre, mes en que la gente acostumbra mas bien tomar ponche que leche.

- **Empresa B.**

La empresa B es de apertura reciente en el país, sin embargo se ha colocado con muy buena aceptación en el mercado mexicano. Esta empresa además de manejar gran diversidad en las presentaciones de leche, también fabrica jugos con presentaciones dirigidas a los niños.

En esta ocasión no se realizó la entrevista personalmente, pero se proporcionó un cuestionario (previamente elaborado) que contestaron, con lo cual se obtuvo la siguiente información.

Al igual que la empresa anterior para saber la demanda nacional, compra estudios a una empresa especializada y la complementan con información gubernamental. Para la demanda de leche de su empresa, elaboran un estudio propio. La estimación de la demanda la calculan para todo el país, para todas las marcas y para todos los tipos de leche.

Para el establecimiento de su demanda, se han basado en un *estudio riguroso del mercado*, así como de la *región*. Particularmente recopilan información de cada empresa y las ventas reales de cada tienda.

En cuanto a las variables que incluyen en sus estudios se encuentran:

- El precio de la leche.
- El ingreso del consumidor.
- Número de compradores y tamaño de la población.
- Nivel educacional.
- Publicidad.
- Inflación.

Como se aprecia, incluyen casi todos los *determinantes de la demanda* excepto los precios de los bienes sustitutivos y complementarios. Además, utilizan para su estudio de la demanda, los conceptos de *elasticidad precio* y *elasticidad ingreso*.

Para ellos los factores que más afectan a la demanda del consumo de leche son: el *precio*, el *ingreso*, la *publicidad* y el *clima*.

CAPITULO II

MODELO DE REGRESIÓN LINEAL MÚLTIPLE.

A. NATURALEZA DEL ANÁLISIS DE REGRESIÓN.

A.1 ORIGEN E INTERPRETACIÓN ACTUAL DEL TÉRMINO DE REGRESIÓN.

La palabra regresión fue introducida por primera vez en 1877 por sir Francis Galton, en sus estudios biológicos sobre la herencia. Galton encontró en sus estudios que los padres altos mostraban tendencia a engendrar hijos altos y que los padres bajos mostraban tendencia a procrear hijos bajos. Sin embargo, la estatura promedio de los hijos de padres altos era menor que la estatura promedio de sus padres, mientras que la estatura de los hijos de padres muy bajos, en promedio, eran más altos que sus padres. Su explicación mostraba que había una tendencia de la estatura *promedio* de los hijos con padres de determinada estatura a moverse o a *regresar* a la estatura *promedio* de la *población*. En expresión de Galton, esto era la "regresión a la mediocridad".

Actualmente el **análisis de regresión** trata de la dependencia de una variable, la variable dependiente Y, a partir de una o más variables independientes $X's$ con el objeto de *estimar* o *predecir* la *media o valor promedio* (poblacional) de la variable dependiente basándose en los valores conocidos (muéstrales)[1] de la o las variables independientes.

Antes de continuar conviene mostrar una clasificación de variables que se usan en el análisis de regresión, según su aplicación, el cual se representa en la tabla 2.1. Los términos del primer renglón, como es de suponer, se utilizan en el análisis de regresión con fines de predicción. Las tres denominaciones siguientes se usan específicamente para los modelos de regresión, siendo todos términos equivalentes. Los términos del

[1] Sobre el concepto de muestra para estimar los valores de la población, se hablará en la sección B.3.

quinto renglón se emplean preferentemente cuando se utiliza análisis de regresión para realizar estudios causa-efecto. La siguiente denominación es propia de los estudios econométricos, sin olvidar que a su vez se hace una subdivisión para las variables X's y Y la cual se presentó en el capítulo anterior. Finalmente, para los problemas de control es más común encontrar los nombres del último renglón.

	Y	$X_0, X_1, \ldots, X_k$
1	predecida (predictando)	predicción (predictores)
2	regresada (regresando)	regresor (regresores)
3	explicada	explicativa (explicatorias)
4	dependiente	independiente
5	de efecto	causales
6	endógena	predeterminada (exógenas)
7	objetivo (respuesta)	de control

Tabla 2.1 Clasificación de variables en el análisis de regresión (Magddala 1996)

Los términos que serán usados con mayor frecuencia en el presente trabajo serán: para Y variable endógena, dependiente o explicada; para X variable exógena, independiente o explicativa. La explicación y el uso de estas variables en un modelo, se presentaron en la sección A.2.2 del capítulo uno.

Siguiendo con el ejemplo de Galton, interesaría descubrir como cambia la estatura *promedio* de los hijos (variable dependiente), a partir de la estatura *promedio* de los padres (variable independiente); es decir, predecir la estatura promedio de los hijos, conociendo la estatura de los padres.

En forma hipotética se muestra un *diagrama de dispersión*[2] en la fig. 2.1. en el cual se observa la distribución de las estaturas de los hijos que corresponden a los valores de las estaturas de los padres en cierta población. Se puede apreciar que a cada una de las estaturas de los padres le corresponde un rango (distribución) de las estaturas de los hijos. Por ejemplo, los papás cuya estatura es de 65 pulgadas, pueden llegar a tener hijos con una estatura desde 60 pulgadas hasta 70 pulgadas

[2] Un *diagrama de dispersión* es una gráfica en la que se traza cada uno de los puntos que representan un par de datos observados (desconocidos) para la variable dependiente e independiente.

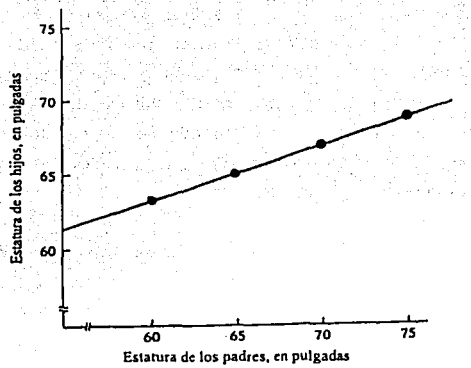


Figura 2.1 Distribución hipotética de las estaturas de los hijos correspondientes a las estaturas de los padres.

En la línea recta que se ha trazado a través de los puntos dispersos, conocida como línea de regresión, se ve claramente que la estatura promedio de los hijos aumenta a medida que aumenta la estatura de los padres. Pero se hace hincapié de que esta relación es cierta solamente en promedio, puesto que ocasionalmente se puede observar que hay hijos altos con padres de baja estatura, y viceversa.

A.2 RELACIÓN ENTRE DOS VARIABLES.

Una relación entre las variables X y Y es el conjunto de todos los valores de X y Y caracterizados por una ecuación determinada. Dentro de la relación que se puede dar entre dos variables se distingue lo que es una *relación funcional* y lo que es una *relación estadística*.

Una **relación funcional** o *determinística* entre dos variables en donde X es la variable independiente y Y la variable dependiente, sería expresada por una fórmula matemática de la forma $Y = f(X)$, en la cual para cada valor de la variable independiente le correspondería un único valor de la variable dependiente, es decir, se tiene una relación exacta. En las ciencias naturales, particularmente en la física clásica, se encuentra esta relación.

Una **relación estadística**, en contraste de una relación funcional, en que no hay una relación perfecta. Para cada valor de X existe una distribución de probabilidad de valores de Y, por lo tanto se manejan variables aleatorias, es decir, variables que tienen una distribución de probabilidad. Así, para cada valor de X, la variable Y puede tomar algún valor determinado (o hallarse en el interior de cierto intervalo) con una probabilidad entre cero y uno. El análisis de regresión se ocupa de la dependencia estadística entre variables, no de la dependencia funcional.

No se debe confundir cuando en el capítulo anterior se habló de encontrar una relación "funcional" para un modelo econométrico; esto se refiere al hecho de que en la teoría económica siempre se representan todas las relaciones en forma determinística, pero como no se espera una explicación perfecta de la realidad, una vez establecida la ecuación ($Y = f(X)$) , se incluye el término del error aleatorio quedando una ecuación de la forma $Y = f(X) + u$, que es la ecuación de regresión que se explicará más adelante.

Para ilustrar estas relaciones, tomando el ejemplo de Galton y suponiendo que la relación fuera exacta o determinística para todos los padres de familia con una estatura determinada, se esperaría que tuvieran todos sus hijos con la misma estatura, como se muestra en el diagrama de dispersión de la figura 2.2. Obsérvese que todos los datos caen sobre la línea de regresión en la relación funcional. Por tanto, el diagrama de dispersión ayuda, por un lado, a ver como están relacionadas ambas variables; por otra parte si es que existe una relación entre dichas variables, saber qué clase de línea o ecuación describe esta relación.

Una situación más real es como la de la figura 2.1. A diferencia de la relación anterior, no todas las observaciones caen directamente sobre la línea, hay valores individuales que difieren del valor promedio. Para una mejor predicción se podría tomar en cuenta también, por ejemplo, la alimentación o el tipo de actividad de los hijos; y aún así existiría una variabilidad aleatoria en la variable dependiente, que no podría ser completamente explicada.

Por último, cabe mencionar que aunque las funciones presentadas en los diagramas de dispersión son lineales (Fig. 2.1 y 2.2), la naturaleza de una relación entre variables, puede tomar formas muy variadas que van desde funciones matemáticas sencillas, hasta otras demasiado complejas conocidas como relaciones no lineales.

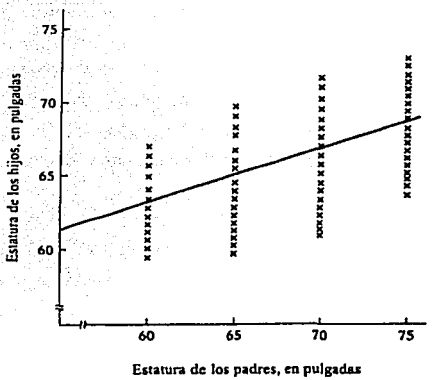


Figura 2.2. Distribución hipotética en una relación determinística de las estaturas de los hijos correspondientes a las estaturas de los padres.

A.3 ANÁLISIS DE REGRESIÓN Y CAUSALIDAD.

La **causalidad** implica que un cambio en las variables independientes causará uno correspondiente en la variable dependiente; es decir, cambios en X influyen sobre cambios en Y pero no al revés. Aunque el análisis de regresión se ocupa del cambio de la dependencia una variable con respecto a otras, no necesariamente implica causalidad. En el análisis de regresión sólo se descubre una *asociación* entre la variable dependiente y las variables independientes, en lugar de detectar una relación causa-efecto, lo cual no impide predecir el valor de una variable con las condiciones de que se tenga información previa sobre otra. Por ejemplo, cuando se calienta un metal y se expande no existe ninguna duda de que se tiene una relación causa-efecto; pero desafortunadamente en la mayoría de los casos (particularmente en Economía) lo anterior no se puede determinar por medio de un análisis estadístico, a menos que se efectúe un experimento rigurosamente controlado.

De esta forma, si existe una relación entre el comportamiento de ambas variables, podría ser el resultado de una verdadera conexión causal, o debido a relaciones de asociación entre las variables, o simplemente podría deberse a una coincidencia, por lo cual hay que tener mucho cuidado.

Mucho se ha discutido sobre el peligro de buscar una relación causa-efecto con los resultados del análisis de regresión. Sin embargo, llama la atención el ver en el libro de Makridakis[3] que clasifica a los modelos de regresión lineal y econométricos[4] dentro de los pronósticos formales[5] con un enfoque causal o estructural, diciendo: "*El objetivo de estos modelos es relacionar la variable que se está pronosticando, con las causas que históricamente han ejercido influencia entre ella y emplear para el pronóstico las relaciones que se establezcan.*"

Un aspecto importante de los métodos de regresión o econométricos, es el conocimiento de la relación que existe entre las diversas series y de la manera en que se comportan las variables en cuestión respecto a otras variables. Así, mientras que puede no haber una relación causa-efecto directa, es típico que exista una relación lógica entre las variables que se incluyan para explicar el modelo. Esto es a lo que se le llama una relación de asociación según lo explicado en el primer párrafo, por eso se clasifican los modelos de regresión dentro del análisis causal.

De acuerdo a lo anterior, se tienen dos ideas para el presente trabajo:

1) Que las relaciones descubiertas por el análisis de regresión en general, son de asociación, pero no necesariamente causales. Dicho de otro modo, el análisis de regresión revela solamente las relaciones estadísticas y que un análisis estadístico por sí mismo, no es una comprobación de las relaciones causales; se requiere trabajo analítico adicional (por medio de un experimento rigurosamente controlado) si se desea saber acerca del patrón causa-efecto que opera en una situación dada. No ha de inferirse causalidad a partir de las relaciones que se observen mediante la regresión.
2) Por otro lado, retomando la idea de descubrir los factores causales en econometria, Kendall y Stuart dicen: "*nuestras ideas de causalidad deben venir de fuera de la*

[3] MAKRIDAKIS, SPYRO; STEVEN, WEELWRIGHTH: "*Manual de Técnicas de Pronóstico*", LIMUSA, México, 1994.

[4] Algunos autores hacen una diferencia entre los modelos de una sola ecuación, llamándolos *modelos de regrasión*, y los modelos de ecuaciones simultáneas denominándolos *modelos econométricos*; sin embargo en libros de econometría no hay tal diferencia, como se explicó anteriormente. Tal vez la confusión sea por que las aplicaciones en econometría se iniciaron en el ámbito macroeconómico, no obstante hoy en día está teniendo auge la aplicación *econométrica* en la empresa con modelos de una sola ecuación.

[5] Llama pronósticos formales a aquellos que utilizan métodos estadísticos. Divide en dos grandes categorías estas técnicas, a saber: 1) los pronósticos con enfoque extrapolativo (lo constituyen los métodos de series de tiempo que incluye curvas de tendencia, método de descomposición, atenuación exponencial, método de Box & Jenkins y modelos bayesianos); y, 2) los pronósticos con enfoque causal o explicativo (como son los modelos de regresión lineal, de ecuaciones simultáneas, método de componentes principales, método de simulación, método de entrada-salida y método de impacto cruzado).

estadística, en última instancia de una u otra teoría "[6]; es decir, dicha relación puede ser sugerida por consideraciones teóricas, específicamente se refiere a la teoría económica la cual sugiere esta relación causal en la cual la variable dependiente se *explica* a través de la(s) ecuación(es) estructural(es) que relaciona ésta con las variables independientes externas, causales de la explicada.[7]

A.4 ANÁLISIS DE CORRELACIÓN.

El principal objetivo del **análisis de correlación** es *medir la fuerza o grado de relación lineal* entre dos variables, mientras que el de regresión se interesa por tratar de *estimar o predecir el valor promedio* de una variable cuando otra *permanece constante* (término que se explicará más adelante) en sus valores. Se dice que dos variables están correlacionadas si los cambios en una variable están asociados con los cambios en la otra variable, así a medida que una variable cambia, se sabrá como cambia la otra.

Los dos análisis tienen diferencias básicas:

- En el análisis de correlación se tratan las dos variables (en el caso simple) de igual forma, no hay distinción entre la variable dependiente e independiente. Aquí se hace la suposición de que ambas variables son aleatorias, es decir, tienen distribución de probabilidad. En un modelo de correlación se pueden intercambiar las posiciones de Y y X, de esta manera se tendrá la misma correlación entre X y Y y de Y con X.
- En el análisis de regresión las variables dependiente e independiente son tratadas en forma diferente. En este análisis puede fijarse la variable independiente en diversos valores específicos, y no es necesario que sea una variable aleatoria. En la figura 2.1, la variable edad está fija en ciertos valores que representan las estaturas de los padres.

Para datos poblacionales, el concepto de correlación se expresa mediante el *coeficiente de correlación* denotado por la letra griega ρ (ro). En datos muestrales se

[6] Tomado de Gujarati (1981).

[7] Existe lo que se conoce como *teoría del análisis causal*. Si se desea profundizar sobre este tema en las series económicas, se puede consultar, por ejemplo: ILLANES, LORENZA "*Causalidad en modelos econométricos de series de tiempo*", Tesis, ITAM, México, 1981.

representa por el ***coeficiente de correlación de Pearson, r***, el cual se calcula con la fórmula:

$$r = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} \qquad 2.1.$$

Los límites del coeficiente de correlación están comprendidos entre

$$-1 \leq r \leq 1$$

el signo aritmético indica la dirección de la relación entre X y Y. Si $r = 1$ expresa una relación lineal perfecta entre X y Y con pendiente positiva. Si $r = -1$ expresa una relación lineal perfecta entre X y Y con pendiente negativa. Si $r = 0$ indica que no existe relación lineal entre las variables.

Por tanto, valores cercanos a la unidad implican una buena correlación o asociación lineal entre X y Y, mientras que valores cercanos a cero indican poca o ninguna correlación. Sin embargo, hay que tener cuidado al interpretar los valores de r; por ejemplo, si se tienen dos valores de r, uno de 0.4 y otro de 0.8, estos indican solamente que hay una correlación positiva de la cual una es más fuerte que la otra, no que $r = 0.8$ tiene una relación lineal dos veces mejor que la dada por $r = 0.4$

A.5 ALGUNOS PROBLEMAS EN LA CONSTRUCCIÓN DEL MODELO DE REGRESIÓN LINEAL.

Antes de aplicar el análisis de regresión, se deben tener en mente algunas consideraciones que pueden influir en los resultados del modelo.

A) Selección de variables independientes. Ya que un modelo es una representación simplificada de la realidad, solamente un número limitado de variables independientes se pueden o se deben incluir. Así un problema central para un modelo de regresión es el de escoger las variables más representativas para el análisis. En econometría se ha dicho que la teoría económica ayuda, sin embargo no siempre se logra saber si las variables que se han incluido son las que mejor representen la realidad, o si

se han omitido variables importantes. Más adelante se hablará sobre los problemas principales que surgen al no escoger correctamente las variables independientes.

B) Forma funcional. Escoger la forma funcional de la relación de regresión va muy ligado al hecho de escoger las variables independientes. En general la teoría indica algunas veces la forma funcional apropiada para un análisis de regresión, no obstante es frecuente que esta no sea conocida por lo cual es un hecho empírico, que en la práctica se debe decidir una vez que los datos han sido recolectados y analizados. Con frecuencia se utilizan las funciones lineales o cuadráticas como primera aproximación a una función de regresión desconocida. Aún en formas complejas se puede obtener una aproximación razonable por medio de una función de regresión lineal; o también se pueden usar dos o más líneas de regresión (piecewise) para aproximar una función de regresión más compleja. Sin embargo, la mayoría de los modelos econométricos son lineales o fácilmente linealizables mediante una sencilla transformación de los datos (como se verá en la sección B.4); por este motivo existe un gran desarrollo de la metodología econométrica relativa a la estimación de modelos lineales.

C) Alcance del modelo. En la formulación del modelo de regresión, frecuentemente es necesario restringir el alcance del modelo en algunos intervalos para la o las variables independientes. La extensión del modelo es determinada o por el diseño de la investigación o por el rango de datos con los que se cuenta. Por ejemplo, una compañía estudia el efecto del precio de un chocolate, cuentan con varios precios en un rango de $4.95 a $7.03; el alcance del modelo podría ser limitado para un nivel de precios comprendido entre $5.00 y $7.00, por experiencias pasadas los investigadores han llegado a la conclusión de que la función de regresión es lineal. Pero la forma de la curva de regresión se hace dudosa fuera del rango considerado, debido a que la investigación intuyó, no probó la naturaleza de la relación estadística para un nivel de precios abajo de $4.95 o arriba de $7.03.

De esta manera, en la práctica la llamada *extrapolación* puede conducir a resultados erróneos. Se debe usar extrapolación solamente cuando se sabe que la relación es lineal en el área determinada, pues fuera de los límites establecidos y conforme los datos se alejen del rango considerado, no se tiene la seguridad de que la curva de regresión sea aproximadamente igual. En la sección C.4. se verán los límites de confianza en los cuales será válida la extrapolación.

B. ESPECIFICACIÓN DEL MODELO DE REGRESIÓN LINEAL SIMPLE (M.L.S).

B.1 CONSTRUCCIÓN DEL MODELO DE REGRESIÓN LINEAL SIMPLE.

La característica diferencial entre los modelos econométricos y económicos, es la introducción del término del error aleatorio. A continuación se presenta el modelo lineal de dos variables ya que expone en forma sencilla las ideas básicas de la regresión; además de ilustrar gráficamente algunas ideas que ayudarán a analizar la especificación de las características del error aleatorio.

Supóngase que se tiene una población con un total de 30 familias y se desea saber la relación entre los gastos de consumo familiar Y y el ingreso disponible familiar diario X; además se sabe que la relación es lineal. Si la relación fuera determinística, a cada valor de la variable independiente X (ingreso disponible), le correspondería un único valor de la variable dependiente Y (consumo), y la relación estaría representada por la ecuación

$$Y_i = a + bX_i \qquad 2.2.$$

siendo a y b los parámetros poblacionales de la relación determinística. El subíndice i indica que se trata del i-ésimo elemento de la población.

Los datos hipotéticos se muestran en la tabla 2.2 la cual indica que todas las familias con el mismo ingreso tienen el mismo consumo. Por ejemplo, el primer renglón indica que hay cinco familias con un ingreso de $100 y que las cinco familias tienen un consumo de $80.

El valor de los parámetros a y b se pueden obtener con las técnicas de la geometría analítica, que al sustituirse en la ecuación queda:

$$Y_i = 10 + 0.70X_i$$

Ingreso Disponible (X)	Consumo (Y)	Número de Familias
100	80	5
140	108	8
180	136	7
200	150	3
220	164	4
260	192	3

Tabla 2.2. Datos de la relación determinística de consumo.

la relación exacta se aprecia más claramente en el diagrama de dispersión de la figura 2.3. Como es de suponer esta relación exacta no se da en la realidad; lo lógico es que todas las familias con el mismo ingreso, consuman de manera diferente; ya sea por sus gustos, por su educación, por sus hábitos alimenticios, etc. Por tanto, una situación común es la reflejada en la tabla 2.3.

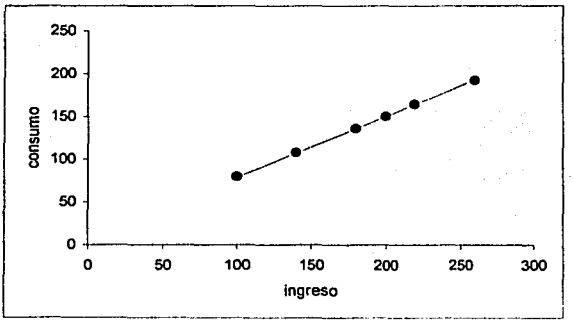


Figura 2.3 Relación determinística.

La interpretación de los datos ahora, tomando también el primer renglón, es que para un ingreso de $100 diarios, hay 5 familias cuyos gastos de consumo varían entre $65 y $80. En otras palabras cada renglón muestra la *distribución* de los gastos de consumo Y correspondientes a un nivel fijo de ingreso X; es decir, es la *distribución condicional* de Y condicionada por los valores de X.

Ingreso Disponible (X)	Consumo (Y)	Valor Medio E(Y)
100	80 85 75 95 65	80
140	108 106 108 110 113 100 116 108	108
180	136 136 140 115 157 135 137	136
200	150 150 152	150
220	164 162 164 166	164
260	192 192 200	192

Tabla 2.3 Datos de la relación estadística de consumo.

Como se recordará el análisis de regresión trata de estimar o predecir el *valor esperado* o promedio poblacional de una variable dependiente dados los valores de una o más variables explicativas basándose en el comportamiento de la primera. Así, la última columna indica el *valor medio o promedio* también conocido como la *media condicional*, que difiere de los valores individuales de la variable dependiente. Los valores promedios de la variable Y aparecen en la figura 2.4 sobre la línea recta con pendiente positiva, indicando que, aunque ocurren variaciones en los gastos de consumo familiar, en *promedio* los gastos de consumo aumentan al aumentar el ingreso, es decir, los valores esperados (condicionales) de Y aumentan al aumentar X.

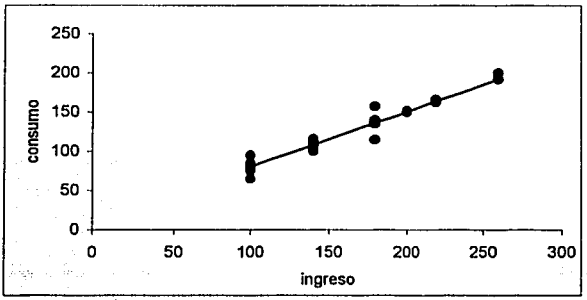


Figura 2.4 Relación estadística.

La línea recta, en el caso hipotético pero que puede tomar otras formas posibles analíticas, se conoce como **línea de regresión**, *curva de regresión* o más precisamente *curva de regresión de X sobre Y.*[8]

- **Ecuación de regresión lineal poblacional.**

Cada media poblacional denotada por $E(Y|X_i)$, es una función de X_i, es decir $E(Y|X_i) = f(X_i)$. Esta ecuación se conoce como *función de regresión poblacional* y muestra cómo el valor promedio de Y varía con las X's.

La forma funcional de la función de regresión poblacional que en algunos casos es un hecho empírico, se puede determinar recurriendo a la teoría. En el ejemplo, el economista afirma que el gasto de consumo está relacionado linealmente con el ingreso, como lo sugiere la teoría económica, por tanto la expresión analítica que relaciona $E(Y|X_i)$ en forma lineal con X es:[9]

$$E(Y|X_i) = \alpha + \beta X_i \qquad 2.3.$$

siendo $E(Y|X_i)$ el valor esperado condicional de Y correspondiente a un valor fijo X_i, α y β son los parámetros desconocidos pero fijos llamados *coeficientes de regresión* donde α es conocida como el intercepto y representa la ordenada al origen de la población, mientras que β es conocida como el coeficiente pendiente porque representa la pendiente real de la población. La ecuación 2.3 se conoce como *función de regresión lineal poblacional.* En forma más simple siendo X dada, la ecuación 2.3 también se puede escribir como:

$$E(Y_i) = \alpha + \beta X_i \qquad 2.4.$$

Si se conocen todos los valores condicionales (es decir, toda la población), como en el ejemplo, es fácil determinar los valores de α y β que en este caso son los mismos que para la relación determinística de la ecuación 2.1, es decir, α=10 y β=0.70.

[8] Desde el punto de vista de la geometría una curva de regresión es simplemente el lugar geométrico de los valores esperados (medias condicionales) de la variable dependiente para los valores fijos de las variables independientes.

[9] Se ha puesto en términos de alfa y beta para diferenciarlo de la ecuación determinística 2.2.

- **Especificación aleatoria.**

Por otra parte, observando la figura 2.4 se aprecia que para un nivel de ingreso dado X_i, el gasto de consumo de una familia está *concentrado* alrededor del consumo promedio de todas las familias para ese mismo X_i, es decir, alrededor de su media condicional.

Dichas diferencias individuales con respecto al comportamiento promedio son tomadas en cuenta por el término de perturbación, que sería simplemente igual al valor observado de la variables Y_i menos el valor hipotético promedio (valor esperado), que vendría dado por la relación:

$$u_i = Y_i - E(Y_i) \qquad 2.5.$$

despejando Y_i queda

$$Y_i = E(Y_i) + u_i \qquad 2.6.$$

en donde la desviación u_i es una variable aleatoria no observable, que puede tomar valores positivos o negativos. Específicamente u_i se conoce como el error aleatorio.[10]

La ecuación expresa que el gasto de una familia dado su nivel de ingreso, es igual al promedio del gasto del consumo de todas las familias con ese mismo nivel de ingreso, más una cantidad (positiva o negativa) que es aleatoria.

Si $E(Y_i)$ se supone lineal en X_i, como en 2.4, la ecuación 2.6 puede escribirse como

$$Y_i = \alpha + \beta X_i + u_i \qquad 2.7.$$

ahora la ecuación establece que el gasto de consumo de una familia está relacionado linealmente con su ingreso más una cantidad aleatoria.

De esta manera se ha llegado a un modelo como el presentado en el capítulo 1 por la ecuación 1.1.[11] sólo que ahora la ecuación 2.7 representa el modelo econométrico

[10] Es común que en los libros de econometría que se le llame *perturbación aleatoria* al error aleatorio.

[11] En la ecuación 2.7 α corresponde al parámetro β_1 de la ecuación 1.1 y β es igual a β_2.

de la relación entre dos variables, que es la forma más simple de la relación estadística entre las variables X y Y. Este modelo se conoce como **modelo de regresión lineal simple** (MLS).

Es un modelo econométrico uniecuacional especificado (de acuerdo a lo visto en la sección A.2.2. del capítulo anterior), ya que la forma funcional de la ecuación 2.7 relaciona entre sí las variables y parámetros correspondientes, incluye un conjunto de datos estadísticos para las variables endógenas y exógenas que deben pertenecer a una misma estructura y, el término del error que es una variable "no observable" por definición, por lo cual no hay información cuantitativa sobre ella.

- **Interpretación de los coeficientes de regresión.**

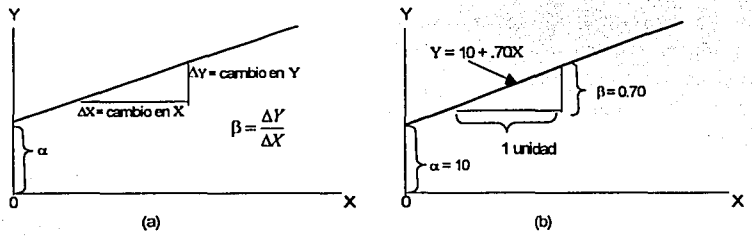
El significado de los coeficientes de regresión, es:

- El coeficiente α, es un factor constante que se incluye en la ecuación, y expresa la intersección de la línea de regresión con el eje Y cuando X es igual a cero.
- La pendiente de la línea β, representa el *cambio promedio* en Y por *cambio unitario* en X (Fig. 2.5.a), es decir, representa la cantidad de cambio promedio en Y (positivo o negativo) para un cambio en particular de una unidad en X.

Tomando la ec. 2.2 del ejemplo de consumo, el valor de $\beta = 0.70$ (Fig. 2.5.b) indica que dentro del rango de X entre \$100 y \$260 diarios, a medida que aumenta X en \$1.00 (una unidad), el aumento en el valor medio o promedio del gasto de consumo es aproximadamente igual a 70 centavos. El valor de $\alpha = 10$ indica el nivel promedio del gasto de consumo cuando el ingreso (diario) es igual a cero; esta interpretación analítica en el análisis de regresión no siempre tiene sentido, pero para el ejemplo podría ser que una familia que no tiene ingresos (por desempleo, por ejemplo), puede mantener ciertos gastos con dinero que tiene ahorrado o que pide prestado. En algunas aplicaciones se interpreta el intercepto como el efecto medio o promedio sobre Y, de todas las variables omitidas en el modelo de regresión.

Finalmente en el modelo, el término del error u_i es una variable aleatoria que toma valores de acuerdo a una determinada *distribución de probabilidad* y representa aquellos

factores no tomados en cuenta explícitamente en la ecuación, pero que afectan a Y, por ejemplo los gustos, el estado civil, etc.



2.5 Representación de los coeficientes de regresión: a) relación lineal positiva, b) relación con datos del ejemplo.

Es preciso hacer notar que la aleatoriedad del término del error implica a su vez, la naturaleza estocástica de la variable dependiente; esto significa simplemente que, dado un valor de la variable independiente, la variable dependiente no toma un valor fijo sino valores diferentes con ciertas probabilidades.

Es evidente que en la realidad no se conocen todos los elementos poblacionales, se cuenta solamente con información parcial, por lo que los parámetros α y β no pueden ser determinados exactamente como se hizo en el ejemplo, sino que deben ser estimados por medio de alguna técnica estadística apropiada (lo cual se verá en la sección B.3).

Además, la especificación completa para el modelo de regresión, no incluye solamente la forma de la ecuación de regresión tal como aparece en la expresión 2.5. Puesto que el término del error aleatorio u_i es no observable y puesto que la variable dependiente es una función de la variable explicativa y del término del error; es necesario la especificación de la distribución de probabilidad de la variable u_i, y de las características de la variable independiente X_i. Esta información está contenida en lo que se conoce como *supuestos del modelo de regresión lineal* que se presentarán a continuación.

B.2. SUPUESTOS BÁSICOS DEL MODELO DE REGRESIÓN LINEAL.

Se considera que el modelo de regresión lineal dado en la ecuación 2.7, debe satisfacer para su correcta aplicación, los siguientes supuestos:

Supuesto 1. El modelo de regresión es lineal en los parámetros como se expresa:

$$Y_i = \alpha + \beta X_i + u_i \qquad 2.7.$$

Se dice que una función es lineal en los parámetros si por ejemplo, α aparece con una potencia de uno y no está multiplicado ni dividido por otro parámetro como sería el caso de $\alpha.\beta$, β/α, etcétera. Una expresión como $Y = \alpha + \beta_1 X_1^2$ es lineal en los parámetros pero no lineal en la variable X. Por lo tanto, de ahora en delante la expresión regresión lineal significará siempre, una regresión lineal en los parámetros, pudiendo ser o no lineal en las variables independientes.[12]

Supuesto 2. El valor esperado del término del error aleatorio es cero (*media nula*).

$$E(u_i \mid X_i) = 0 \qquad i = 1,2,...n \qquad 2.8.$$

Este supuesto establece que el valor de la media de u_i, sobre las X_i dadas es cero; es decir, la variable u_i puede tomar valores positivos o negativos pero en promedio es cero para cada observación, lo que implica una relación exacta en términos de promedio entre X y Y ($E(Y_i) = \alpha + \beta X_i$). En otras palabras, para cada valor de X, los valores de Y están distribuidos normalmente alrededor de su valor medio que sería un punto en la recta de regresión poblacional.

En la figura 2.6 se muestran algunos valores de la variable X y sus poblaciones Y asociadas con cada una de ellas. Como se observa cada población Y correspondiente a un X dado está distribuida normalmente alrededor de su media, con algunos valores de Y

[12] Esto no quiere decir que no haya solución para los modelos de regresión no lineal en los parámetros, sin embargo se ha explicado que la mayoría de modelos econométricos son lineales o fácilmente linealizables.

por encima y por debajo de esta. Las distancias por encima y por debajo de la media no son otra cosa que los u_i,[13] y lo que el supuesto requiere es que el promedio de estas desviaciones deba ser cero para cualquier X dado.

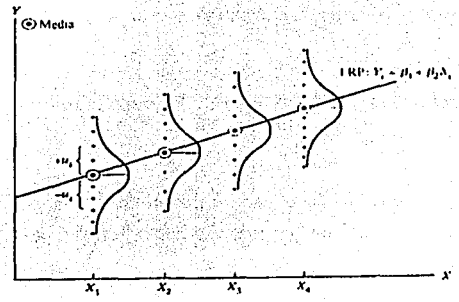


Figura 2.6 Distribución condicional de los errores aleatorios u_i.

Supuesto 3. Todos los errores poseen la misma varianza (*homocedasticidad* de (homos) igual, igual dispersión o (cedasticidad) de igual varianza).

$$\text{Var } (u_i \mid X_i) = \sigma^2 \qquad i = 1,2,\ldots,n. \qquad 2.9.$$

la ecuación expresa que la dispersión de u_i para cualquiera que sea el valor de X_i, es un número positivo constante e igual a σ^2. Planteado de otra forma, se puede observar que las poblaciones Y correspondientes a diversos valores de X tienen la misma varianza (Var $(Y_i|X_i) = \sigma^2$), lo cual se observa en la figura 2.7.

Hay casos en que la varianza de la población de Y cambia al variar X; esta situación de varianza distinta se conoce como *heterocedasticidad* o dispersión desigual o varianza desigual. Simbólicamente

$$\text{Var}(u_i \mid X_i) = \sigma_i^2$$

[13] Para fines ilustrativos se está suponiendo que las u_i están distribuidas simétricamente, no obstante, se reparten normalmente (Tomado de Gujarati (1997)).

en esta ecuación aparece un subíndice en σ^2, el cual indica que la varianza de la población Y ya no es constante como se aprecia en la figura 2.8.

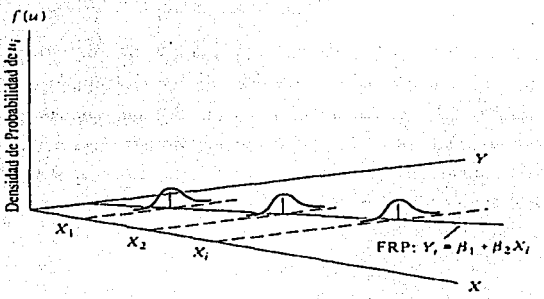


Figura 2.7 Homocedasticidad.

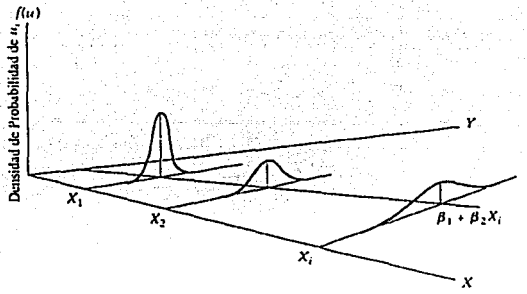


Figura 2.8 Heteroscedasticidad.

Aunque en los dos casos, a medida que aumenta X, aumenta Y; en el segundo habrá variabilidad para los valores de Y a medida que X aumenta, mientras que en el primero no habrá variación.

Supuesto 4. Los términos del error son independientes entre sí (*no autocorrelación* o no-correlación serial)

$$\text{Cov}\,(u_i, u_j | X_i, X_j) = 0 \qquad (i \neq j) \qquad 2.10.$$

Dados dos valores cualquiera de X (X_i y X_j), la correlación entre dos u_i y u_j cualquiera ($i \neq j$) es cero, es decir, no están correlacionados. En la figura 2.9.a se ve que los u están *correlacionados positivamente*; un u positivo está seguido de un u positivo y un u negativo está seguido por un u negativo; mientras que en la figura 2.9.b los u están *correlacionados negativamente*, un u positivo está seguido por un u negativo y viceversa; se tiene un patrón sistemático. Por el contrario, en la figura 2.9.c no hay un orden para los u, lo cual indica cero correlación.

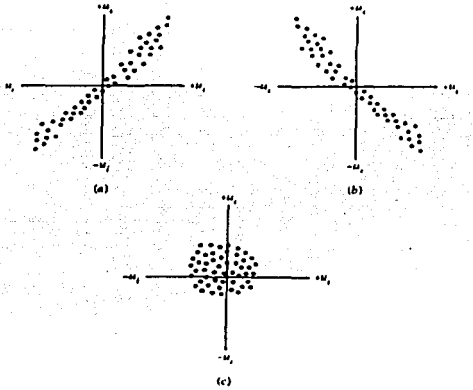


Figura 2.9 Correlación entre errores. (a) correlación serial positiva, (b) correlación serial negativa, (c) correlación cero.

En otras palabras, el considerar este supuesto quiere decir que se considerará el efecto, si este existe, de X sobre Y, sin preocuparse sobre las otras influencias que podrían actuar sobre Y como resultado de las posibles correlaciones entre los u.

Supuesto 5. La variable explicativa X, es fija o determinística *(X no aleatoria*[14]).

El concepto determinístico se usa para indicar que los valores de la variable independiente son fijos, conocidos o controlables, que no cambian de muestra a muestra, es decir que son independientes del muestreo. Lo que equivale a suponer que unos valores de X, previamente elegidos se dejan invariables en el muestreo (se mantiene fija X), cambiando aleatoriamente en un valor para Y. Así el análisis de regresión es un análisis condicional, esto es, condicionado a los valores dados por el valor X.

Supuesto 6. El término del error u_i es una variable aleatoria normalmente distribuida, junto con los supuestos 2,3 y 4 (llamados supuestos estocásticos)[15] lo cual implica que los u_i son independientes y tienen una distribución normal con media cero y varianza constante σ^2. (*normalidad*).

$$u_i \sim N(0, \sigma^2) \qquad i = 1,2,\ldots, n \qquad 2.12.$$

Este supuesto implica que el término del error u_i es continuo y sus valores se extienden entre $-\infty$ y $+\infty$, se distribuye simétricamente alrededor de su media, y su distribución queda totalmente determinada por dos parámetros: μ y σ^2.

El supuesto de normalidad se justifica al considerar cada valor del error aleatorio como el resultado (la suma), de un gran número de pequeñas causas, cada una de las cuales produce una pequeña desviación da la variable dependiente respecto del valor que tomaría si la relación fuese determinística. Este supuesto implica a su vez que la variable dependiente Y_i es también normal por ser una función lineal de u_i; es decir, Y_i se distribuye normalmente con media $(\alpha + \beta X_i)$ y varianza σ^2 lo cual se representa como

$$Y_i \sim N(\alpha + \beta X_i, \sigma^2) \qquad 2.13.$$

[14] En Econometría se les da el nombre de variables aleatorias o *estocásticas*; por ello se tienen en las ecuaciones simultáneas, los modelos que se conocen con el nombre de *regresores estocásticos*.

[15] Los supuestos estocásticos son de considerable importancia. Las situaciones en las cuales no se cumplen son tratadas en la sección C.4.

En resumen para la econometría se define al modelo de regresión lineal simple (siguiendo la notación de la ecuación 1.1), con la ecuación matemática

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad i = 1,2,...,n \qquad 2.14.$$

donde:

Y_i = variables aleatorias no observables (e independientes)
X_i = variables no aleatorias observables
β_1, β_2 = parámetros de regresión desconocidos
u_i = variables aleatorias no observables (e independientes), distribuidas según una normal de media cero y varianza constante.

Cabe aclarar que existe la posibilidad de que no se cumplan uno o más de los supuestos básicos, en tal caso será necesario corregirlos para garantizar los resultados de la regresión. En la sección C.4 se verán los problemas por incumplimiento de los mismos para el modelo de regresión lineal múltiple.

Con los supuestos del modelo de regresión queda cubierta la etapa econométrica de especificación del modelo, ya que estos se utilizan como base para la obtención de estimadores de los parámetros de la regresión; por tanto, un objetivo importante del análisis de regresión es estimar los parámetros desconocidos en el modelo de regresión, como se verá a continuación.

B.3 ESTIMACIÓN DE PARÁMETROS.

En general el procedimiento para estimar los parámetros de una distribución de probabilidad, es suponer que se tiene una muestra aleatoria de tamaño n, a saber $Y_1, Y_2, ..., Y_n$ y utilizar esta información para estimar los parámetros desconocidos. Este procedimiento conocido como el problema de *estimación*, es una rama de la inferencia estadística[16] que puede subdividirse en *estimación puntual* y *estimación por intervalos*. Por tanto, conviene ver algunas definiciones básicas antes de continuar.

[16] La teoría de *inferencia estadística* consiste en aquellos métodos por los cuales se realizan inferencias o generalizaciones acerca de una población. Las técnicas de inferencia estadística pueden dividirse en dos áreas principales: estimación y prueba de hipótesis (esta última se verá en la sección C.3.1.

Un **estimador puntual** o *estadístico*, denotado por Θ, es simplemente una regla o método que indica cómo combinar los datos con el fin de determinar el valor de un parámetro θ.

Una **estimación puntual** $\hat{\theta}$, utiliza la información muestral y se concreta en un solo número o punto que estima el parámetro de la población de interés, es el valor numérico obtenido después de una aplicación para un parámetro. También es conocido como *estimado* el valor particular obtenido por el estimador el cual generalmente se expresa por medio de una fórmula.

Un **estimador por intervalos** es una regla que indica como calcular dos números con base a los datos muéstrales. Cuando se usa un estimador por intervalo para estimar el parámetro de la población, el par de números que se obtiene se llama **estimación por intervalo** o *intervalo de confianza*.[17]

Para el modelo de regresión lineal, como se supone que el error se distribuye normalmente con media cero, lo único que no se conoce acerca de su distribución es su varianza σ^2. Por lo tanto el modelo basado en los supuestos incluye un total de tres parámetros desconocidos: los parámetros de regresión β_1 y β_2 y la varianza del término de perturbación σ^2, que se deberán de estimar.

Al estimar los valores de β_1 y β_2 por medio de un método de estimación apropiado,[18] se obtiene lo que se conoce como una *línea de regresión estimada* o *recta de regresión de la muestra*. Si β_1 y β_2 se estiman a través de $\hat{\beta}_1$ y $\hat{\beta}_2$, respectivamente, la recta de regresión de la muestra viene dada por la función de regresión muestral

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i \qquad 2.15.$$

siendo $\hat{Y}_i$ el valor estimado de Y_i. La diferencia entre el valor real Y_i y el valor estimado $\hat{Y}_i$ se conoce como **residuo** o residual, y se representa por e_i

[17] El uso del intervalo de confianza se verá en la sección C.3.1.

[18] Un método para transformar los datos en estimadores de los valores de los parámetros poblacionales del modelo.

$$e_i = Y_i - \hat{Y}_i \qquad 2.16.$$

que no debe confundirse con u_i. El error aleatorio u_i es la diferencia entre el valor real Y_i y el valor medio o teórico $E(Y_i)$ que da la relación exacta presentada en la ecuación 2.4.

En general e_i es distinta de u_i porque $\hat{\beta}_1$ y $\hat{\beta}_2$ difieren de los verdaderos valores de β_1 y β_2. En realidad, los residuos e_i pueden considerarse como "estimaciones" del error aleatorio u_i (de forma alternativa, se puede decir que se utiliza la distribución de e_i para estimar los parámetros de la distribución de u_i).

Por tanto, hay que distinguir entre las dos expresiones siguientes:

$$Y_i = \beta_1 + \beta_2 X_i + u_i \qquad \text{(población)} \qquad 2.14.$$

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \qquad \text{(muestra)} \qquad 2.17.$$

estas ecuaciones se pueden considerar como el modelo para una sola observación Y_i, las cuales se representa gráficamente en la figura 2.10 junto con las líneas de regresión lineal poblacional y ajustada.

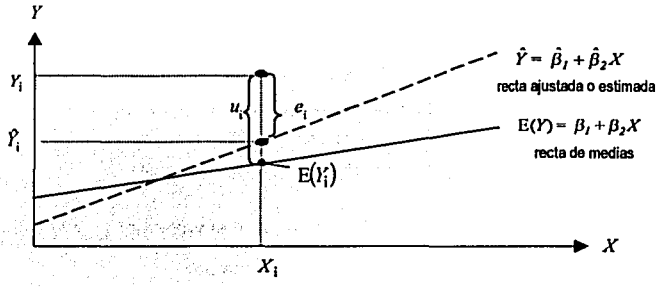


Figura 2.10. Líneas de regresión poblacional y muestral.

Para $X = X_i$ se tiene una observación (muestral) $Y = Y_i$. En términos de la función de regresión muestral el Y_i observado puede expresarse como

$$Y_i = \hat{Y}_i + e_i \qquad 2.18.$$

y en términos de la función de regresión poblacional, como se presentó en la ecuación 2.6, que es

$$Y_i = E(Y_i) + u_i \qquad 2.6.$$

Como se aprecia en la figura 2.10, $\hat{Y}_i$ sobreestima el verdadero $E(Y_i)$ para el X_i dado, mientras que a la izquierda del cruce de las líneas, la línea de la muestra subestima la línea de la población. La sobre o subestimación es inevitable por las fluctuaciones muéstrales.

En el apartado siguiente se presentan dos procedimientos para estimar los parámetros de regresión poblacionales (estructurales), bajo los supuestos mencionados, que se acerque lo más posible a la recta de regresión poblacional.

B.3.1 Métodos de estimación.

Existen en la actualidad varios métodos de estimación (momentos, máxima verosimilitud, transformada inversa, etc.). En econometría los métodos de mínimos cuadrados y máxima verosimilitud han sido la base de las técnicas estadísticas modernas. Como lo indican sus nombres, con estos métodos se obtienen estimaciones de los parámetros de regresión β_1 y β_2 (para el caso simple), mediante minimización o maximización, respectivamente, de ciertas funciones, las cuales son evaluadas utilizando los datos del problema.

A continuación se muestran estos dos métodos de estimación, con sus características principales y las propiedades estadísticas que producen los estimadores mínimo cuadráticos y máximo verosímiles.

- **Método de Mínimos Cuadrados Ordinarios (MCO).**[19]

El **Método de Mínimos Cuadrados Ordinarios** o *directos* se debe a Carl Friedrich Gauss, matemático alemán (1821).

En este método de estimación se busca obtener los estimadores de los parámetros β_1 y β_2 mediante el cálculo de $\hat{\beta}_1$ y $\hat{\beta}_2$ que satisfacen la ecuación

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + e_i \qquad 2.19.$$

$$\text{tal que } \sum_{i=1}^{n} e_i^2 \text{ es un mínimo.}$$

en donde e_i son los residuos. La expresión de residuo en la observación i para las estimaciones específicas de $\hat{\beta}_1$ y $\hat{\beta}_2$, según la ecuación 2.15 y sustituyendo en 2.16 será

$$e_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i \qquad i=1,2,...,n \qquad 2.20.$$

El proceso de elevar al cuadrado es la razón de la terminología de mínimos cuadrados. Al elevar al cuadrado cada residuo y sumando todos los residuos resultantes al cuadrado, da por resultado la **suma de cuadrados de los residuos** o *residual* para cualquier β_1 y β_2 [20]

$$\text{SCR} = S = \Sigma e_i^2 = \Sigma (Y_i - \hat{Y}_i)^2 = \Sigma (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X)^2 \qquad 2.21.$$

El método de mínimos cuadrados implica la minimización de la suma de cuadrados S, mediante la elección de los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$ que resuelven el problema

[19] Como se explicó anteriormente en econometría se les da el calificativo de directos u ordinarios, para diferenciarlos de otros métodos de mínimos cuadrados (como el de mínimos cuadrados generalizados, en dos etapas, etc.), que se utilizan para estimar principalmente los parámetros en los sistemas de ecuaciones simultáneas

[20] Para facilitar el manejo de las sumatorias se omite en adelante el rango de variación, quedando comprendido que va de 1 a n.

$$\min_{\hat{\beta}_1, \hat{\beta}_2} S = S(\hat{\beta}_1, \hat{\beta}_2) = \Sigma e_i^2 = \Sigma (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2$$

Las condiciones necesaria y suficiente para minimizar la suma de cuadrados[21] $S=S(\hat{\beta}_1, \hat{\beta}_2)$, son que ambas derivadas parciales respecto a $\hat{\beta}_1$ y $\hat{\beta}_2$, desaparezcan al igualarse a cero. Así, derivando parcialmente e igualando a cero se obtiene

$$\frac{\partial S}{\partial \hat{\beta}_1} = 2\Sigma(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-1) = 0$$

$$\frac{\partial S}{\partial \hat{\beta}_2} = 2\Sigma(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)(-X_i) = 0 \qquad 2.22.$$

al aplicar las propiedades para las sumatorias ($\Sigma c = nc$ y $\Sigma cX_i = c\Sigma X_i$, donde c es una constante, siendo los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$ las constantes), las ecuaciones pueden escribirse como

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

$$\Sigma Y_i X_i = \hat{\beta}_1 \Sigma X_i + \hat{\beta}_2 \Sigma X_i^2 \qquad 2.23.$$

las cuales se conocen como *ecuaciones normales* de la regresión. Estas ecuaciones son muy importantes ya que a partir de ellas se pueden obtener los valores de los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$ como función de la información conocida de la variable dependiente y de la variable independiente.

Dadas las realizaciones $Y_1, Y_2, \ldots, Y_n$, las ecuaciones pueden resolverse si se dividen entre n. Así, dividiendo la primera ecuación de 2.23

$$\frac{\Sigma Y_i}{n} = \hat{\beta}_1 + \hat{\beta}_2 \frac{\Sigma X_i}{n}$$

[21] Para la explicación de las condiciones necesaria y suficientes para maximizar o minimizar funciones sujetas a diferentes restricciones, ver por ejemplo Intriligator 1971.

el estimador de M.C.O. para β_1 será

$$\hat{\beta}_1 = \frac{\Sigma Y_i}{n} - \hat{\beta}_2 \frac{\Sigma X_i}{n} = \bar{Y} - \hat{\beta}_2 \bar{X} \qquad 2.24.$$

en donde $\bar{X} = \frac{1}{n}\Sigma X_i$ y $\bar{Y} = \frac{1}{n}\Sigma Y_i$ son las medias muéstrales de X y de Y respectivamente. Lo cual significa que la recta muestral pasa a través del punto $(\bar{X}, \bar{Y})$.

Sustituyendo $\hat{\beta}_1$ en la segunda ecuación de 2.23

$$\Sigma Y_i X_i = \left(\frac{\Sigma Y_i}{n} - \hat{\beta}_2 \frac{\Sigma X_i}{n}\right)\Sigma X_i + \hat{\beta}_2 \Sigma X_i^2$$

resolviendo para $\hat{\beta}_2$

$$\hat{\beta}_2 = \frac{\Sigma X_i Y_i - \frac{(\Sigma X_i)(\Sigma Y_i)}{n}}{\Sigma X_i^2 - \frac{(\Sigma X_i)^2}{n}} = \frac{n\Sigma X_i Y_i - (\Sigma X_i)(\Sigma Y_i)}{n\Sigma X_i^2 - (\Sigma X_i)^2} \qquad 2.25.$$

aplicando las propiedades de las sumatorias y tomando en cuenta que $\bar{X}$ y $\bar{Y}$ son constantes, obsérvese que

$$\begin{aligned} n(X_i - \bar{X})^2 &= n(\Sigma X_i^2) - 2n\bar{X}(\Sigma X_i) + n^2\bar{X}^2 \\ &= n(\Sigma X_i^2) - 2n\left(\frac{\Sigma X_i}{n}\right)(\Sigma X_i) + n^2\left(\frac{\Sigma X_i}{n}\right)\left(\frac{\Sigma X_i}{n}\right) \\ &= n(\Sigma X_i^2) - (\Sigma X_i)^2 \end{aligned}$$

es el numerador de la ecuación 2.25. Además

$$\begin{aligned} n\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) &= n(\Sigma X_i Y_i) - n\bar{X}(\Sigma Y_i) - n\bar{Y}(\Sigma X_i) + n^2\bar{X}\bar{Y} \\ &= n(\Sigma X_i Y_i) - \frac{n(\Sigma X_i)}{n} - \frac{n(\Sigma Y_i)(\Sigma X_i)}{n} + n^2\left(\frac{(\Sigma X_i)}{n}\right)\left(\frac{(\Sigma Y_i)}{n}\right) \\ &= n(\Sigma X_i Y_i) - (\Sigma Y_i)(\Sigma X_i) \end{aligned}$$

es el denominador de la ecuación 2.25. Por tanto, 2.25 se puede escribir como:

$$\hat{\beta}_2 = \frac{\Sigma(X_i - \bar{X})\Sigma(Y_i - \bar{Y})}{\Sigma(X_i - \bar{X})^2} \qquad 2.26.$$

ya que la *n* del numerador se cancela con la del denominador.

Así se obtienen los únicos *estimadores mínimo-cuadráticos.* Las características de estos estimadores son:

- que están expresados en función de las observaciones muéstrales en términos de los valores de X_i y Y_i
- que son estimaciones puntuales, es decir, dada una muestra, cada estimador proporcionará un sólo valor para el parámetro β_1 y β_2.

Gauss mostró que bajo el cumplimiento de los cuatro supuestos[22] del llamado *modelo clásico de regresión*[23] que se muestran a continuación

a) Media nula: $E(u_i \mid X_i) = 0 \qquad i = 1,2,..., n$ (supuesto 2)

b) Varianza: $Var(u_i \mid X_i) = \sigma^2 \qquad i = 1,2,..., n$ (supuesto 3 de *homocedasticidad*)

c) Independencia: u_i y u_j son independientes para cualquier i $(i \neq j)$ se tiene que $Cov(u_i, u_j \mid X_i, X_j) = 0$ (supuesto 4 *no auto correlación*)

d) X es una variable no aleatoria (supuesto 5).

el método produce estimadores *lineales* e *insesgados*, además de que son los mejores de todos los estimadores lineales e insesgados (es decir, tienen *varianza mínima*), por lo que se dice que son eficientes[24]. Estas propiedades están contenidas en el conocido teorema de Gauss-Marcov[25]; algunas veces a este teorema se le denomina como MELI de Mejor

[22] Que corresponden a los supuestos 2, 3, 4 de la sección B.2, como se indica. El supuesto 1 de la sección B.2, se incluyó por manejarse modelos lineales en los parámetros.

[23] Es clásico ya que fueron los supuestos originales que sirvieron como base en el análisis de regresión. Debido a las propiedades estadísticas que ofrece, se hizo muy popular, lo que llevó a comparar los modelos de regresión con estos supuestos "gaussianos".

[24] En econometría a los estimadores con varianza mínima y que son insesgados se les llama *eficientes* (para muestras pequeñas o finitas), a pesar de que el concepto de eficiencia puede abarcar otras características, (ver por ejemplo Kmenta).

[25] A Marcov se le debe el enfoque de varianza mínima.

Estimador Lineal Insesgado,[26] por ello los estimadores mínimo-cuadráticos se les conoce también como estimadores MELI.

Para la construcción de intervalos de confianza, las pruebas de hipótesis y la predicción (que se verán en las secciones C.3 y C.5), es necesario suponer que u sigue una distribución de probabilidad, por lo cual se debe incluir este supuesto que en conjunción con los supuestos estocásticos a, b y c, implica que u_i está distribuida normalmente para toda i con media cero y varianza constante

- $u_i \sim N(0,\sigma^2)$ i = 1,2,...,n (supuesto 7, *de normalidad*).

las principales ventajas de este supuesto son: que si u_i se distribuye normalmente también lo hace Y_i (ec. 2.13). Los estimadores $\hat{\beta}_1$ y $\hat{\beta}_2$ también se distribuyen normalmente por ser funciones lineales de Y_i, lo cual se expresa como:

$$\hat{\beta}_1 \sim N(\hat{\beta}_1, \sigma^2_{\hat{\beta}_1})$$
$$\hat{\beta}_2 \sim N(\hat{\beta}_2, \sigma^2_{\hat{\beta}_2}) \qquad \text{cc.2.27.}$$

y están distribuidos independientemente de σ^2, la cual a su vez se distribuye como una χ^2(ji- cuadrada) como se verá más adelante.

Los estimadores mínimo-cuadráticos también poseen todas las propiedades asintóticas (para muestras grandes) deseables, a saber, son asintóticamente insesgados, consistentes[27] y asintóticamente eficientes.

A continuación en base a los supuestos sobre u_i, se encontrarán las medias y las varianzas para los estimadores de β_1 y β_2 dados en 2.27.

[26] En Inglés BLUE, de Best Linear Unbiased Estimator.

[27] Existe el teorema de la consistencia de los mínimos cuadrados el cual se puede consultar por ejemplo, en Intriligator, 1990.

Para encontrar la media de $\hat{\beta}_2$ se utiliza la ecuación 2.26, que multiplicada por Y_i se tiene

$$\hat{\beta}_2 = \frac{\Sigma(X_i - \bar{X})Y_i - \bar{Y}\Sigma(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2}$$

dado que $\bar{Y}$ es una constante y que la suma de las desviaciones del valor medio de una variable es siempre cero $\left[\Sigma(X_i - \bar{X}) = 0\right]$, se simplifica a

$$\beta_2 = \frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - \bar{X})^2} \qquad 2.28.$$

al aplicar el operador E queda

$$\begin{aligned} E(\hat{\beta}_2) &= \frac{\Sigma(X_i - \bar{X})E(Y_i)}{\Sigma(X_i - \bar{X})^2} \\ &= \frac{\Sigma(X_i - \bar{X})(\beta_1 + \beta_2 X_i)}{\Sigma(X_i - \bar{X})^2} \\ &= \frac{\beta_1\Sigma(X_i - \bar{X}) + \beta_2\Sigma(X_i - \bar{X})X_i}{\Sigma(X_i - \bar{X})^2} \end{aligned}$$

nuevamente se toma en cuenta la propiedad $\left[\Sigma(X_i - \bar{X}) = 0\right]$, además como $\left[\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - 2\bar{X}\Sigma X_i + n\bar{X}^2 = \Sigma X_i^2 - 2\bar{X}\Sigma X_i + \bar{X}\Sigma X_i = \Sigma(X_i - \bar{X})X_i\right]$, se concluye que

$$E(\hat{\beta}_2) = 0 + \beta_2 \frac{\Sigma(X_i - \bar{X})X_i}{\Sigma(X_i - \bar{X})X_i} = \beta_2 \qquad 2.29.$$

por tanto $\hat{\beta}_2$ es un estimador insesgado de β_2.

Para encontrar la varianza de $\hat{\beta}_2$ se toma en cuenta que $Y_1, Y_2, \ldots, Y_n$ son independientes, aplicando la propiedad para calcular la varianza de una función lineal de variables independientes, se tiene

$$\text{Var}(\hat{\beta}_2) = \text{V}\left[\frac{\Sigma(X_i - \bar{X})Y_i}{\Sigma(X_i - \bar{X})^2}\right] = \frac{1}{\left[\Sigma(X_i - \bar{X})^2\right]^2}\left[\Sigma(X_i - \bar{X})^2\right]\text{V}(Y_i)$$

$$= \frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} \qquad 2.30.$$

Ahora para probar la insesgadez de $\hat{\beta}_1$ se toma la ecuación 2.24 y aplicándole las propiedades de valor esperado se obtiene

$$\text{E}(\hat{\beta}_1) = \text{E}(\bar{Y} - \hat{\beta}_2\bar{X}) = \text{E}(\bar{Y}) - \bar{X}\text{E}(\hat{\beta}_2) \qquad 2.31.$$

por tanto hay que encontrar $\text{E}(\bar{Y})$. Entonces dado que

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

se tiene

$$\bar{Y} = \frac{1}{n}\sum Y_i = \beta_1 + \beta_2\bar{X} + \bar{u}$$

$$\text{E}(\bar{Y}) = \beta_1 + \beta_2\bar{X} + \text{E}(\bar{u}) = \beta_1 + \beta_2\bar{X}$$

sustituyendo en 2.31, queda

$$\text{E}(\hat{\beta}_1) = \beta_1 + \beta_2\bar{X} - \beta_2\bar{X} = \beta_1 \qquad 2.32.$$

por tanto $\hat{\beta}_1$ también es insesgado.

Al aplicar las propiedades para calcular la varianza de $\hat{\beta}_1$ en 2.24 da

$$\text{Var}(\hat{\beta}_1) = \text{V}(\bar{Y}) + \bar{X}^2\text{V}(\hat{\beta}_2) - 2\bar{X}\,\text{cov}(\bar{Y}, \hat{\beta}_2) \qquad 2.33.$$

ahora se necesita obtener $\text{V}(\bar{Y})$ y $\text{cov}(\bar{Y}, \hat{\beta}_2)$, entonces

$$\text{V}(\bar{Y}) = \text{V}(\bar{u}) = \left(\frac{1}{n}\right)\text{V}(u_i) = \frac{\sigma^2}{n}$$

en la obtención de $\text{cov}(\bar{Y}, \hat{\beta}_2)$, se puede hacer que

$$c_i = \frac{\Sigma(X_i - \bar{X})}{\Sigma(X_i - \bar{X})^2} \qquad \text{donde} \quad \Sigma c_i = 0 \qquad i = 1,2,...,n$$

sustituyendo en 2.28

$$\hat{\beta}_2 = c_i Y_i$$

y

$$\text{cov}(\bar{Y}, \hat{\beta}_2) = \text{cov}\left[\sum\left(\frac{1}{n}\right)Y_i, \sum c_i Y_i\right]$$

$$= \sum\left(\frac{c_i}{n}\right)V(Y_i) + 2\sum\sum_{i<j}\left(\frac{c_i}{n}\right)\text{cov}(Y_i, Y_j)$$

como Y_i y Y_j son independientes $(i \neq j)$ entonces $\text{cov}(Y_i, Y_j)=0$, y se tiene que

$$\text{cov}(\bar{Y}, \hat{\beta}_2) = \sum\left(\frac{1}{n}\right)\sigma^2 = \frac{\sigma^2}{n}\sum c_i = 0$$

sustituyendo estos resultados en 2.33 para encontrar $V(\hat{\beta}_1)$, queda

$$V(\hat{\beta}_1) = \frac{\sigma^2}{n} + \bar{X}^2\left[\frac{\sigma^2}{\Sigma(X_i - \bar{X})^2}\right] + 0 = \sigma^2\left[\frac{1}{n} + \frac{\bar{X}^2}{\Sigma(X_i - \bar{X})^2}\right]$$

$$= \sigma^2\left[\frac{\Sigma(X_i - \bar{X})^2 + n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}\right]$$

como $\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - 2\bar{X}\Sigma X_i + \Sigma \bar{X}^2$ ya que $\Sigma X_i = n\bar{X}$ y $\Sigma \bar{X}^2 = n\bar{X}^2$ al ser $\bar{X}$ una constante, entonces $= \Sigma X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 = \Sigma X_i^2 - n\bar{X}^2$ se obtiene

$$V(\hat{\beta}_1) = \sigma^2\left[\frac{\Sigma X_i^2 - n\bar{X}^2 + n\bar{X}^2}{n\Sigma(X_i - \bar{X})^2}\right] = \frac{\sigma^2 \Sigma X_i^2}{n\Sigma(X_i - \bar{X})^2} \qquad 2.34.$$

El parámetro σ^2, la varianza del error del modelo, refleja la variación aleatoria o la variación del error experimental alrededor de la línea de regresión. Dado que el valor de σ^2 que aparece en las ecuaciones anteriores es desconocido, puede obtenerse un estimador basándose en las observaciones muestrales. Como cada $\hat{Y}_i$ estima la media de Y_i, la diferencia $Y_i - \hat{Y}_i$ representa la desviación de Y_i con respecto a su propia media;

la suma de cuadrados de estas diferencias dividida entre una constante apropiada es la forma en que se determina una σ^2. La constante apropiada es n - 2 que corresponde a los grados de libertad que se pierden al estimar los parámetros β_1 y β_2 antes de obtener $\hat{Y}_i$.

El estimador de σ^2 se denota con S^2 y está dado por

$$S^2 = \frac{\Sigma(Y_i - \hat{Y}_i)^2}{n-2}$$

$$S^2 = \frac{\Sigma e_i^2}{n-2} \qquad 2.35.$$

A continuación se demostrará que es un estimador insesgado para σ^2. Dado que

$$E(S^2) = E\left[\left(\frac{1}{n-2}\right)\Sigma e_i^2\right] = \left(\frac{1}{n-2}\right)E\left(\Sigma e_i^2\right) \qquad 2.36.$$

se debe encontrar $E(\Sigma e_i^2)$ para verificar que $E(S^2)=\sigma^2$. Sustituyendo 2.24 en 2.21, se tiene

$$\begin{aligned} E\left(\Sigma e_i^2\right) = E\left(\Sigma(Y_i - \hat{Y}_i)\right)^2 &= E\left\{\Sigma\left[Y_i - (\bar{Y} - \hat{\beta}_2\bar{X}) - \hat{\beta}_2 X_i\right]^2\right\} \\ &= E\left\{\left[\Sigma(Y_i - \bar{Y}) - \hat{\beta}_2(X_i - \bar{X})\right]^2\right\} \\ &= E\left[\Sigma(Y_i - \bar{Y})^2 - 2\hat{\beta}_2\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) + \hat{\beta}_2^2\Sigma(X_i - \bar{X})^2\right] \end{aligned}$$

de 2.26 se tiene $\Sigma(X_i - \bar{X})(Y_i - \bar{Y}) = \hat{\beta}_2\Sigma(X_i - \bar{X})^2$ y como $\Sigma(Y_i - \bar{Y})^2 = \Sigma Y_i^2 - n\bar{Y}^2$, entonces la ecuación se reduce a

$$\begin{aligned} &= E\left[\Sigma Y_i^2 - n\bar{Y}^2 - 2\hat{\beta}_2\left(\hat{\beta}_2\Sigma(X_i - \bar{X})^2\right) + \hat{\beta}_2^2\Sigma(X_i - \bar{X})^2\right] \\ &= E\left[\Sigma Y_i^2 - n\bar{Y}^2 - \hat{\beta}_2\Sigma(X_i - \bar{X})^2\right] \\ &= \Sigma E(Y_i^2) - nE(\bar{Y}^2) - \Sigma(X_i - \bar{X})^2 E(\hat{\beta}_2^2) \end{aligned}$$

ya que para cualquier variable aleatoria $U, \mathrm{E}(U^2) = V(U) + \left[\mathrm{E}(\bar{Y})\right]^2$, entonces se tiene que

$$
\begin{aligned}
&= \Sigma\left\{\mathrm{V}(Y_i) + \left[\mathrm{E}(Y_i)^2\right]\right\} - n\left\{\mathrm{V}(\bar{Y}) + \mathrm{E}\left[(\bar{Y})^2\right]\right\} - \Sigma(X_i - \bar{X})^2\left\{\mathrm{V}(\hat{\beta}^2) + \mathrm{E}(\hat{\beta}_2)^2\right\} \\
&= \Sigma\sigma^2 + \Sigma(\hat{\beta}_1 + \hat{\beta}_2 X_i)^2 - n\left[\frac{\sigma^2}{n} + (\hat{\beta}_1 + \hat{\beta}_2\bar{X})^2\right] - \Sigma(X_i - \bar{X})^2\left[\frac{\sigma^2}{\Sigma(X_i - \bar{X})^2} + \hat{\beta}_2^2\right] \\
&= n\sigma^2 - 2\sigma^2 + \Sigma(\hat{\beta}_1 + \hat{\beta}_2 X_i)^2 - n(\hat{\beta}_1 + \hat{\beta}_2\bar{X})^2 - \hat{\beta}_2^2\Sigma(X_i - \bar{X})^2
\end{aligned}
$$

como $\Sigma(X_i - \bar{X})^2 = \Sigma X_i^2 - n\bar{X}^2$ y desarrollando el cuadrado de los otros dos términos, da

$$
\begin{aligned}
&= (n\text{-}2)\sigma^2 + n\hat{\beta}_1^2 + 2\hat{\beta}_1\hat{\beta}_2\Sigma X_i + \hat{\beta}_2^2\Sigma X_i^2 - n\hat{\beta}_1^2 - 2\hat{\beta}_1\hat{\beta}_2 n\bar{X} - \hat{\beta}_2^2 n\bar{X}^2 - \hat{\beta}_2^2(\Sigma X_i^2 - n\bar{X}^2) \\
&= (n\text{-}2)\sigma^2 + n\hat{\beta}_1^2 + 2\hat{\beta}_1\hat{\beta}_2 n\bar{X} + \hat{\beta}_2^2\Sigma X_i^2 - n\hat{\beta}_1^2 - 2\hat{\beta}_1\hat{\beta}_2 n\bar{X} - \hat{\beta}_2^2\bar{X}^2 n - \hat{\beta}_2^2(\Sigma X_i^2 - n\bar{X}^2)
\end{aligned}
$$

y sustituyendo en 2.36 en la ecuación simplificada, se muestra que S^2 es un estimador insesgado del verdadero σ^2, esto es:

$$
\begin{aligned}
\mathrm{E}\left[\Sigma(Y_i - \hat{Y})^2\right] &= (n-2)\sigma^2 \\
&= \sigma^2 \qquad\qquad 2.37.
\end{aligned}
$$

la cual es conocida como *varianza residual.*

La **varianza residual** S^2 es una medida absoluta que indica el ajuste de la recta estimada de regresión a las medias de las observaciones de la variable respuesta. Entre más pequeño sea el valor de S^2, mejor será el ajuste del modelo. Bajo el supuesto de normalidad, se tiene que $(n-2)S^2/\sigma^2$ es una variable ji cuadrada con n-2 grados de libertad, independiente del parámetro β.

La raíz cuadrada positiva de S^2, S o $S_{\cdot}$ se conoce como **error estándar estimado**[28] o *error estándar de la regresión*, el cual indica la variabilidad de los puntos observados

[28] De una manera simple, el *error estándar* de un estimador es su desviación estándar. Los paquetes de computadora se refieren comúnmente a los errores estándar estimados como errores estándar solamente, para este caso es el de la regresión, pero también se calculará más adelante el error estándar de los coeficientes de regresión y de la predicción.

alrededor de la línea de regresión; es decir, hasta qué punto los valores observados difieren de los estimados en la línea de regresión. Su cálculo es importante ya que la precisión de un estimador se mide por su desviación estándar y sirve como base para las pruebas de hipótesis y estimación por intervalos.

- **Método de Máxima Verosimilitud**

El **Método de Máxima Verosimilitud** es un método de estimación muy general que se aplica a una gran variedad de problemas.

En el método de mínimos cuadrados no es necesario especificar la distribución de probabilidad de los errores aleatorios para obtener $\hat{\beta}_1$ y $\hat{\beta}_2$; en contraste en la estimación por máxima verosimilitud se incluye el supuesto de que u_i está distribuido normalmente con media 0 y varianza constante σ^2 para toda i=1,2,...,n. La idea de este método se basa, dada una muestra, de elegir los valores de los parámetros desconocidos β_1, β_2 y σ^2, de tal forma que maximice la probabilidad de obtener la muestra observada. Por tanto, es posible obtener estimadores de β_1, β_2 y σ^2 por este método.

Dado que cada Y_i es independiente y *normalmente* distribuido con media $\beta_1 + \beta_2 X_i$ y varianza σ^2 (por la ecuación 2.13), para hallar los estimadores de máxima verosimilitud se utiliza la distribución normal

$$f(X, \mu, \sigma^2) = \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left[-\frac{(X_i - \mu)^2}{2\sigma^2}\right]$$

sustituyendo los valores de los parámetros para el modelo de regresión y tomando en cuenta la independencia de las Ys; la función de densidad de probabilidad conjunta puede escribirse como el producto de n funciones de densidad individuales, que es la función de verosimilitud

$$L(Y_1,Y_2,....,Y_n ; \beta_1,\beta_2,\sigma^2) = \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left[-\frac{(Y_1 - \beta_1 - \beta_2 X_1)^2}{2\sigma^2}\right]\frac{1}{\sqrt{2\Pi\sigma^2}} exp\left[-\frac{(Y_2 - \beta_1 - \beta_2 X_2)^2}{2\sigma^2}\right]$$

$$... \frac{1}{\sqrt{2\Pi\sigma^2}} exp\left[-\frac{(Y_n - \beta_1 - \beta_2 X_n)^2}{2\sigma^2}\right]$$

simplificando

$$= \left(\frac{1}{\sqrt{2\Pi\sigma^2}}\right)^n exp\left[-\frac{1}{2\sigma^2}\sum(Y_i - \beta_1 - \beta_2 X_i)^2\right]$$

para facilitar los cálculos, puede maximizarse el logaritmo natural, ln, de la función de verosimilitud en vez de la propia función ya que ln L y L alcanzan su valor máximo en el mismo punto, por lo que se tiene

$$\max_{\hat{\beta}_1,\hat{\beta}_2,\hat{\sigma}^2} = \ln L(\hat{\beta}_1,\hat{\beta}_2,\sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln 2\sigma^2 - \frac{1}{2\sigma^2}\Sigma(Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i)^2 \qquad 2.38.$$

Ln L se maximizará primero con respecto a $\hat{\beta}_1$, $\hat{\beta}_2$ y después con respecto a σ. Obsérvese que sólo el tercer miembro involucra β_1 y β_2, entonces como $-(1/2\sigma^2)$ siempre tendrá signo negativo, maximizar el tercer miembro equivale a minimizar la suma de residuos al cuadrado $(Y_i - \beta_1 - \beta_2 X_i)^2$. Así los estimadores de máxima verosimilitud de $\hat{\beta}_1$ y $\hat{\beta}_2$ equivalen a los de mínimos cuadrados dados por las ecuaciones 2.25. En cambio, para que los estimadores mínimo cuadráticos de $\hat{\beta}_1$ y $\hat{\beta}_2$ sean estimadores máximo verosímiles, se requiere que u_i se distribuya normalmente. Por tanto tienen las mismas propiedades para muestras finitas bajo el supuesto de normalidad.

Derivando la ec. 2.38 respecto a σ, se obtiene

$$-\frac{n}{2} + \frac{\Sigma e_i^2}{\sigma^2} = 0 \qquad 2.39.$$

donde $Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$ ha sido reemplazado por e_i, dada la ec. 2.20. Resolviendo, el estimador de máxima verosimilitud para la varianza será

$$\hat{\sigma}^2 = \frac{\sum e_i^2}{n} = \frac{\sum (Y_i - \hat{Y}_i)^2}{n} \qquad 2.40.$$

este estimador es sesgado para muestras pequeñas, pero no para muestras grandes que es asintóticamente insesgado.

El método de máxima verosimilitud es un método para muestras grandes, y las propiedades de sus estimadores serán sólo propiedades asintóticas, esto es, consistentes, asintóticamente insesgados y eficientes.

En resumen, estos dos métodos de estimación, bajo el supuesto de normalidad del modelo de regresión lineal, producen estimaciones iguales para los parámetros de regresión, lo cual es válido para regresiones múltiples. Así, los estimadores mínimo cuadráticos son de aplicación más general, ya que son válidos cualquiera que sea la forma en que se distribuya el error. Sin embargo, mientras que el método de máxima verosimilitud proporciona una fórmula para el estimador de la varianza, el de mínimos cuadrados no lo tiene; aunque se puede calcular como se hizo anteriormente obteniéndose la ec. 2.35, este estimador es insesgado para muestras finitas, a diferencia del estimador sesgado que da el método de máxima verosimilitud. Pero, a medida que el tamaño de la muestra n aumenta, la diferencia no es significativa, tienden a estar uno cerca del otro.

B.4 TRANSFORMACIÓN DE VARIABLES.

Las transformaciones de los datos proporcionan un medio de modificar variables por una o dos razones: para mejorar la relación entre las variables o, para corregir el incumplimiento de los supuestos básicos. En cada caso se debe proceder muchas veces por ensayo y error, ponderando los resultados frente a la necesidad de transformaciones adicionales.

En esta sección se presentan algunas de las funciones no lineales más usadas en economía de acuerdo a la aplicación del modelo, así como las transformaciones

apropiadas de las variables para linealizar la ecuación. La corrección en el incumplimiento de los supuestos del modelo de regresión múltiple, se verá en la sección C.4.

- **Transformaciones lineales en el modelo.**

Para este efecto es aconsejable trabajar con un modelo alterno en el cual X o Y (o ambas) entren en una forma no lineal.

A) Función potencial.

Surgen ocasiones en que una ecuación potencial como

$$Y_i = \beta_1 X_i^{\beta_2} u_i$$

puede representar adecuadamente la relación entre X y Y.(fig. 2.11. a). En este caso, si se utiliza una transformación logarítmica

$$\ln Y_i = ln\,\beta_i + \beta_2 ln X_i + ln\,u_i \qquad 2.41.$$

si se escriben las transformaciones de las variables como $Y_i^* = ln\,Y$ y $X_i^* = ln\,X$, entonces la ecuación

$$Y_i^* = \log \beta_i + \beta_2 X_i^* + u_i$$

establece una relación lineal entre logaritmos de las variables X_i y Y_i por lo que se conoce como modelo *doble-logarítmico* o modelo *log-log*, cuya estimación puede realizarse de acuerdo con el procedimiento de mínimos cuadrados.

Se trata de un modelo especialmente interesante en economía, útil para descripción de funciones de producción o de demanda de un producto, cuyo atractivo

radica en el hecho de que el coeficiente de la pendiente de la ecuación β_2 mide la elasticidad de Y con respecto a X, (modelo de elasticidad constante), es decir, el cambio porcentual de Y ante un cambio porcentual de X.

$$\beta_2 = \left(\frac{dY}{dX}\cdot\frac{X}{Y}\right) = \text{elasticidad}_{Y/X} \qquad 2.42.$$

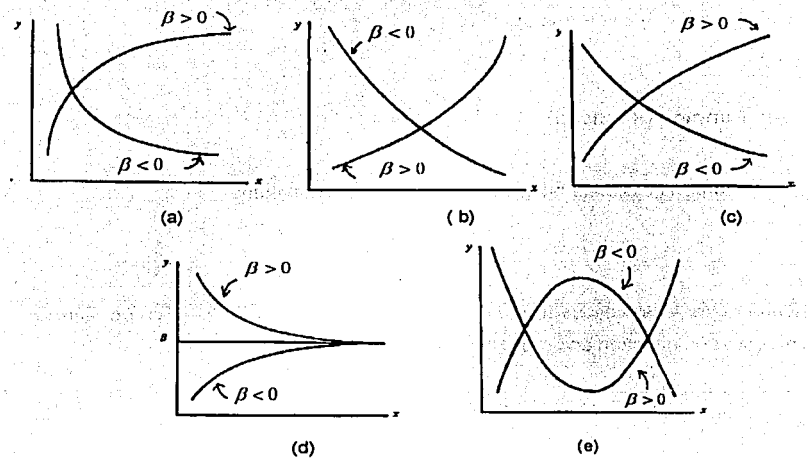


Figura 2.11 Formas funcionales.

B) Funciones exponencial y potencia exponencial.

Cuando interesa analizar la tasa de cambio de la variable Y ante cambios unitarios de X (Fig. 2.11.b) o viceversa, la variación absoluta en la variable Y ante una variación unitaria en la tasa de cambio de X (Fig. 2.11.c), se plantea respectivamente ecuaciones de la forma

$$Y_i = \beta_1 \beta_2^{X_i} e^{u_i}$$

$$e^{Y_i} = \beta_1 X_i^{\beta_2} u_i$$

son fácilmente linealizables, mediante la transformación logarítmica de sus elementos.

$$\ln Y_i = \beta_1 + \beta_2 X_i + u_i \qquad 2.43.$$

$$Y_i = \ln \beta_1 + \beta_2 \ln X_i + \ln u_i \qquad 2.44.$$

si como en el caso anterior se expresan las transformaciones $Y_i^* = \ln Y$ y $X_i^* = \ln X$,

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

$$Y_i = \log \beta_1 + \beta_2 X_i^* + u_i$$

son conocidos como *modelos semi-logarítmicos* en Y ó en X respectivamente; algunos autores los llaman modelo *log-lin* ó *lin-log*. Si se desea calcular los coeficentes de elasticidades, se usan las fórmulas:

$$\beta_2(X)^* \quad \text{y} \quad \beta_2\left(\frac{1}{Y}\right)^*$$

donde * indica que el coeficiente de elasticidad es función del punto en que se calculan de la variable X, de Y o de ambas (a diferencia del modelo doble logarítmico). Cuando no se determinan estos valores, es común que se midan las elasticidades por sus valores medios; por tanto, una manera en que se representaría estas elasticidades es

$$\beta_2(\bar{X}) \qquad 2.45.$$

$$\beta_2\left(\frac{1}{\bar{Y}}\right) \qquad 2.46.$$

Un caso especial del modelo potencia exponencial, se presenta cuando la variable X es el tiempo, entonces la pendiente del modelo permite conocer la tasa anual de crecimiento constante (o decrecimiento) en la variable Y. Por esta razón suele denominarse a la ecuación linealizada *modelo de crecimiento constante*.

C) Función hiperbólica.

La ecuación matemática

$$Y_i = \beta_1 + \beta_2 \frac{1}{X_i} + u_i \qquad 2.47.$$

usada en microeconomía para describir la relación entre la demanda de un artículo Y y el ingreso X (curva de Engel) y a nivel macroeconómico, para expresar la relación entre la tasa de variación de salarios y tasa de desempleo (curva de Phillips), que se muestra en la figura 2.11.d; puede ser estimada considerando la transformación $Z_i^* = 1/X_i$, en cuyo caso el modelo linealizado, conocido como recíproco, es

$$Y_i = \beta_1 + \beta_2 Z_i^* + u_i$$

Para este modelo la elasticidad se mide como:

$$-\beta_2\left(\frac{1}{\overline{XY}}\right) \qquad 2.48.$$

D) Función parabólica.

En ciertas ocasiones en las que se desea representar el coste medio de producción de un bien según el nivel de producto obtenido, se acude a la forma cuadrática (Fig. 2.11.e).

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 X_i^2 + u_i \qquad 2.49.$$

en este caso la linealidad entre X y Y se consigue mediante la transformación de la variable $Z_i^* = X_i^2$, definiéndose la ecuación lineal

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z^* + u_i$$

C. EL MODELO DE REGRESIÓN LINEAL MÚLTIPLE. (M.L.M.)

C.1 PRESENTACIÓN DEL MODELO DE REGRESIÓN LINEAL MÚLTIPLE GENERAL E HIPÓTESIS BÁSICAS.

Una vez presentadas las principales técnicas de estimación para el modelo lineal de dos variables, normalmente se procede a obtener otras medidas descriptivas (de estimación y prueba de hipótesis). Sin embargo esto se hará al analizar el modelo uniecuacional lineal con K variables ($Y, X_2, X_3, \ldots, X_k$) en notación matricial.

Esto se justifica, por un lado, porque los resultados del modelo general se aplican al modelo de dos variables; además que en la práctica la teoría económica rara vez es tan simple como para que sólo una variable explicativa afecte a la variable dependiente. Por otro lado, la notación matricial tiene gran ventaja sobre la escalar ya que se pueden manejar de manera compacta los modelos de regresión con cualquier número de variables, lo cual ayuda a no perder el objetivo perseguido.

Supóngase que la teoría económica establece que una variable Y es función de un conjunto de k-1 variables explicativas $X_2, X_3, \ldots, X_k$. Esto se puede reflejar por extensión del modelo de dos variables, mediante la ecuación lineal:

$$Y = \beta_1 + \beta_2 X_k + \ldots + \beta_k X_k + u$$

donde los $\beta_1, \beta_2, \ldots, \beta_k$ son los parámetros estructurales de la relación. Además incluye el término necesario del error aleatorio u.

Como se sabe, se debe disponer de un conjunto de n observaciones o datos de las variables, todas ellas pertenecientes a la misma relación estructural.

$$\begin{matrix} Y_1 & X_{21} & X_{31} & \cdots & X_{k1} \\ Y_2 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ Y_n & X_{2n} & X_{3n} & \cdots & X_{kn} \end{matrix}$$

se puede entonces escribir las n ecuaciones, una para cada conjunto de observaciones como:

$$\begin{aligned} Y_1 &= \beta_1 + \beta_2 X_{21} + \beta_3 X_{31} + \cdots + \beta_k X_{k1} + u_1 \\ Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3 X_{32} + \cdots + \beta_k X_{k2} + u_2 \\ &\vdots \quad \vdots \quad \vdots \quad \vdots \quad \vdots \\ Y_n &= \beta_1 + \beta_2 X_{2n} + \beta_3 X_{3n} + \cdots + \beta_k X_{kn} + u_n \end{aligned}$$

las cuales se representan por una sola ecuación conocida como **modelo de regresión lineal múltiple** *o modelo lineal general* expresada por la ecuación

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + \ldots, + \beta_k X_{ki} + u_i \qquad i = 1,2,\ldots,n \qquad 1.1.$$

en la cual la variable endógena o dependiente Y es explicada como una función lineal de k-1 variables predeterminadas $X_2, X_3,\ldots,X_k$, de el término del error u y de la i-ésima observación para una población de tamaño n.

- **Supuestos.**

Se mantienen los supuestos básicos del modelo clásico de regresión lineal. El modelo es lineal en los parámetros, las variables explicativas son no aleatorias, el término del error es una variable aleatoria normal con media cero y varianza finita e igual para cada observación, los errores no están correlacionados entre sí, ni con las variables explicativas.

Adicionalmente, para el modelo de más de dos variables, es necesario especificar dos supuestos o condiciones al modelo de regresión lineal múltiple, los cuales se presentan a continuación:

- *Supuesto 8.* El número de observaciones n debe ser mayor que el número de parámetros k del modelo a estimar, $n > k$ (muestra suficiente).
- *Supuesto 9.* No debe existir ninguna relación lineal exacta entre las variables explicativas $X_2, X_3,\ldots,X_k$ (*no multicolinealidad*).

la necesidad de estos supuestos se verá más adelante.

- **Interpretación de la ecuación de regresión.**

Dados los supuestos del modelo de regresión, al tomar la esperanza condicional de Y en ambos lados de la ecuación general, se obtiene

$$E(Y|\ X_2, X_3,...,X_k) = E(Y) = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + ... + \beta_k X_{ki} \qquad 2.50.$$

que es la *media condicional* o el valor esperado de Y condicionado a los valores dados o fijos de las variables $X_2, X_3,...,X_k$. Esta fórmula es equivalente a la expuesta en 2.3 para el caso simple.

- **Significado de los coeficientes de regresión parcial.**

La interpretación de los coeficientes para el modelo de regresión múltiple de $(k-1)$ variables es la siguiente: β_1 es la ordenada al origen, en el análisis de regresión es el valor esperado de la variable explicada Y cuando todas las variables explicativas $X_2, X_3,...,X_k$ son iguales a cero. Los coeficientes restantes $\beta_2, \beta_3,..., \beta_k$ se denominan **coeficientes de regresión parcial** *o coeficientes neto de regresión*, y son coeficientes condicionales dado que se incluyen en la ecuación más de dos variables explicativas con sus respectivos coeficientes. Cada uno representa el cambio de el valor promedio que se produce en Y_i $(E(Y|\ X_2, X_3,...,X_k))$, por unidad de cambio en la variable explicativa de interés y la variable explicada.

Así, un modelo de regresión múltiple de tres variables se representaría con la ecuación

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{3i} + u_i \qquad i = 1,2,..., n$$

cuya interpretación de los parámetros de regresión sería:

- β_1 : es la ordenada al origen o intercepto cuando $X_{2i} = 0$ y $X_{3i} = 0$

- β_2 : mide el cambio en el valor promedio que se produce en Y_i (E (Y| X_{2i}, X_{3i})), por cambio de una unidad en X_{2i}, manteniendo constante X_{3i}. Matemáticamente corresponde a la derivada parcial de Y con respecto a X_{2i}.
- β_3 : mide la variación en el valor promedio de Y_i, ante una variación unitaria en X_{3i}, manteniendo constante X_{2i}. Matemáticamente β_3 es la derivada parcial de (E (Y| X_{2i}, X_{3i})) con respecto a X_{3i}.

Por ejemplo, en una ecuación sobre las ventas de una empresa, en que se hace depender la cantidad vendida Y_i de la publicidad X_{2i} y del precio X_{3i}, el modelo sería igual que el anterior y la interpretación será: β_1 representa el valor que alcanzarían las ventas si no hubiera publicidad X_{2i} y el precio X_{3i} fuera cero; β_2 mide el cambio promedio de las ventas (media de Y) al cambiar la publicidad X_{2i} en una unidad, permaneciendo constante el precio X_{3i}; similarmente β_3 mide la variación en el valor promedio de las ventas ante una variación unitaria del precio X_{3i}, permaneciendo constante la publicidad X_{2i}.

El planteamiento y especificación realizado para el modelo de regresión lineal múltiple, pueden ser descritos alternativamente en *notación matricial*. Así, representando los vectores y las matrices en negrillas, la ecuación 1.1. puede expresarse como:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ 1 & X_{22} & X_{32} & \cdots & X_{k2} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & X_{2n} & X_{3n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_k \end{bmatrix} = \begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_{n1} \end{bmatrix} \qquad 2.51.$$

$$\mathbf{Y} \qquad \mathbf{X} \qquad \beta \qquad \mathbf{u}$$

$$n \times 1 \qquad n \times k \qquad k \times 1 \qquad n \times 1$$

donde:

Y = vector columna (nx1) cuyos elementos son las observaciones de la variable dependiente

X = matriz (nxk) que por columna contiene las n observaciones de cada una de las k-1 variables exógenas X_2, X_3,...,X_k. La primera columna representa el término del intercepto.

β = vector columna (kx1) cuyos elementos son los parámetros desconocidos del modelo

u = vector columna (nx1) que contiene los n errores aleatorios

o en forma más simplificada

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \qquad 2.52.$$

Asimismo, el conjunto de supuestos que afectan al modelo de regresión lineal, expresados en forma matricial pueden escribirse como:

Supuesto a. El valor esperado del vector de los errores **u**, es decir, de cada uno de sus elementos tiene media igual a cero (*media nula*)

$$\mu = E(\mathbf{u}) = \mathbf{0}$$

significa que

$$E(\mathbf{u}) = E\begin{bmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{bmatrix} = \begin{bmatrix} E(u_1) \\ E(u_2) \\ \vdots \\ E(u_n) \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \qquad 2.53.$$

Supuesto b. La matriz de varianzas y covarianzas de los errores **u**, es igual a un escalar (*perturbaciones esféricas*)

$$E(\mathbf{uu}') = \sigma^2\mathbf{I}$$

donde **u**' es el transpuesto del vector columna **u** o vector fila.

Este supuesto indica en forma simplificada que todos los errores son independientes (no autocorrelación) y que poseen la misma varianza (homocedastidad); en econometría comúnmente se denomina supuesto de perturbaciones esféricas, lo cual significa que

$$E(\mathbf{uu'}) = E\begin{bmatrix}(u_1)\\(u_2)\\\vdots\\(u_n)\end{bmatrix}[u_1, u_2, \ldots, u_n] = E\begin{bmatrix}u_1^2 & u_1u_2 & \cdots & u_1u_n\\u_1u_2 & u_2^2 & \cdots & u_2u_n\\\vdots & \vdots & \ddots & \vdots\\u_1u_n & u_2u_n & \cdots & u_n^2\end{bmatrix}$$

aplicando el operador de valor esperado E a cada elemento de la matriz

$$E(\mathbf{uu'}) = \begin{bmatrix}E(u_1^2) & E(u_1u_2) & \cdots & E(u_1u_n)\\E(u_1u_2) & E(u_2^2) & \cdots & E(u_2u_n)\\\vdots & \vdots & \ddots & \vdots\\E(u_1u_n) & E(u_2u_n) & \cdots & E(u_n^2)\end{bmatrix} \qquad 2.54.$$

debido a los supuestos de homocedasticidad ($E(u_i,u_j)=\sigma^2$, para i=j) y de no auto correlación ($E(u_i,u_j)=0$ $i\neq j$), se obtiene

$$E(\mathbf{uu'}) = \begin{bmatrix}\sigma^2 & 0 & \cdots & 0\\0 & \sigma^2 & \cdots & 0\\\vdots & \vdots & \ddots & \vdots\\0 & 0 & \cdots & \sigma^2\end{bmatrix} = \sigma^2\begin{bmatrix}1 & 0 & \cdots & 0\\0 & 1 & \cdots & 0\\\vdots & \vdots & \ddots & \vdots\\0 & 0 & \cdots & 1\end{bmatrix} = \sigma^2\mathbf{I} \qquad 2.55.$$

donde $\mathbf{I}$ es una matriz identidad de n x n.

La matriz 2.54, y su representación matricial en 2.55, se conoce como matriz de *varianza-covarianza* de los errores u_i; los elementos sobre la diagonal principal son las varianzas y el resto de elementos son las covarianzas.

Supuesto c. La matriz **X** (n x k) de observaciones de las variables explicativas es no aleatoria (*insesgadez*).

Es decir, los valores de la matriz **X** se consideran fijos para muestras repetidas.

Supuesto d. El rango de la matriz **X** es igual a k (*rango pleno*)

$$\rho(\mathbf{X}) = k < n \qquad 2.56.$$

donde k es el número de columnas en **X**, y k es menor que el número de observaciones n.

Este supuesto, aplicable a los modelos de más de dos variables, garantiza que el tamaño de la muestra n sea mayor que k (muestra suficiente), además implica que las variables explicativas son independientes, es decir, que no existirá relación lineal exacta entre las variables explicativas del modelo, lo que se conoce como *no multicolinealidad.* Por el contrario, cuando existe una relación lineal exacta, se dice que las variables son colineales.

Como ejemplo, considérese nuevamente el gasto del consumidor en donde la teoría económica presupone que el consumo Y está relacionado linealmente con el ingreso X_2 y además debe incluir la riqueza X_3; es decir, ahora presupone que las dos variables anteriores pueden tener influencia independiente sobre el consumo. Si no fuera así, no tendría sentido incluir ambas variables en el modelo, ya que solamente se debería incluir una variable independiente, no dos; por tanto no se podría evaluar la influencia *separada* del ingreso y de la riqueza sobre el consumo.

De esta manera, la no multicolinealidad requiere que se incluyan solamente aquellas variables que no sean combinación lineal de algunas de las otras variables en el modelo.

Supuesto e. El vector **u** tiene una distribución normal multivariada (normalidad)

$$\mathbf{u} \sim N(\mathbf{0}, \sigma^2 \mathbf{I}) \qquad 2.57.$$

donde $\mathbf{I}$ es una matriz identidad de $n \times n$.

En resumen, el modelo lineal general en notación matricial es

$$\mathbf{Y} = \mathbf{X}\beta + \mathbf{u} \qquad 2.52.$$

en donde:

$\mathbf{Y}$ = vector ($n \times 1$) de variables aleatorias observables.
$\mathbf{X}$ = matriz ($n \times k$) de cantidades fijas, conocidas de rango igual a k ($k < n$).
β = vector ($k \times 1$) de parámetros desconocidos.
$\mathbf{u}$ = vector ($n \times 1$) de variables aleatorias no observables, distribuidas según una normal con media igual a $\mathbf{0}$ y varianza igual a $\sigma^2\mathbf{I}$.

Con esto queda concluida la etapa econométrica de especificación, y se puede pasar a la de estimación, en donde basándose en las observaciones muéstrales, se concretará la relación estructural desconocida que liga a las variables.

C.2. ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO DE REGRESIÓN.

- **Método de Mínimos Cuadrados Ordinarios (M.C.O.)**

Se ha llegado al problema de estimar los parámetros desconocidos del modelo, β_1, β_2, ..., β_k y σ^2. Para ello se utilizará el método de Mínimos Cuadrados Ordinarios, que así como en el caso simple, bajo los supuestos mencionados el método ofrece algunas propiedades estadísticas muy importantes, por lo cual se ha constituído en uno de los más eficaces métodos de análisis de regresión en econometría.

Como se sabe, los estimadores mínimo-cuadráticos son aquellos que minimizan la suma de los cuadrados de las diferencias entre los valores reales y los valores estimados de la variable dependiente.

Si se designa el estimador de β como $\hat{\beta}$, los valores estimados de la variable Y_i serán

$$Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_{2i} + ,...,+ \hat{\beta}_k X_{ki} \qquad 2.58.$$

y en notación matricial

$$\mathbf{Y} = \mathbf{X}\hat{\beta} \qquad 2.59.$$

siendo $\hat{\mathbf{Y}}$ un vector columna (n x 1) cuyos elementos son los valores estimados de la variable dependiente y $\hat{\beta}$ un vector columna (k x 1) cuyos elementos son los valores estimados de los coeficientes del modelo, esto es,

$$\hat{\mathbf{Y}} = \begin{bmatrix} \hat{Y}_1 \\ \hat{Y}_2 \\ \vdots \\ \hat{Y}_n \end{bmatrix}, \qquad \hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \vdots \\ \hat{\beta}_k \end{bmatrix}$$

Falta definir el vector columna **e** (n x 1) que contiene los residuos, es decir, las diferencias entres los valores reales y los estimados de la variable Y,

$$\mathbf{e} = \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = \begin{bmatrix} Y_1 - \hat{Y}_1 \\ Y_2 - \hat{Y}_2 \\ \vdots \\ Y_n - \hat{Y}_n \end{bmatrix} = \mathbf{Y} - \hat{\mathbf{Y}}$$

En el caso de k variables los estimadores mínimo-cuadráticos se obtienen minimizando

$$S = \Sigma e_i^2 = \Sigma (Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_{2i} - ,\ldots,- \hat{\beta}_k X_{ki})^2 \qquad 2.60.$$

donde Σe_i^2 es la suma de cuadrados residual (SCR), que en notación matricial equivale a minimizar **e'e**, puesto que,

$$\mathbf{e'e} = (e_1, e_2, \ldots, e_n) \begin{bmatrix} e_1 \\ e_2 \\ \vdots \\ e_n \end{bmatrix} = e_1^2 + e_2^2 + ,\ldots,+ e_n^2 = \sum e_i^2$$

entonces el objetivo será contrastar el vector β de estimadores que satisfagan la ecuación matricial

$$\mathbf{Y} = \mathbf{X}\hat{\beta} + \mathbf{e} \qquad 2.61.$$

tal que **e'e** sea mínimo.

Para que la ecuación anterior quede en términos de **e** como en 2.60, simplemente se despeja **e**

$$\mathbf{e} = \hat{\mathbf{Y}} - \mathbf{X}\hat{\beta} \qquad 2.62.$$

Desarrollando **e'e** se obtiene

$$\mathbf{e'e} = (\mathbf{Y} - \mathbf{X}\hat{\beta})'(\mathbf{Y} - \mathbf{X}\hat{\beta}) = \mathbf{Y'Y} - \hat{\beta}'\mathbf{X'Y} - \mathbf{Y'X}\hat{\beta} + \hat{\beta}'\mathbf{X'X}\hat{\beta}$$

tomando en cuenta las propiedades de la transpuesta de una matriz [$(\mathbf{X}\beta)' = \beta'\mathbf{X}'$] se tiene que el segundo y tercer miembro son iguales, por lo que la ecuación se puede escribir

$$\mathbf{e'e} = \mathbf{Y'Y} - 2\hat{\beta}'\mathbf{X'Y} + \hat{\beta}\,\mathbf{X'X}\hat{\beta} \qquad 2.63.$$

que es la representación matricial de la ecuación 2.62.

Para minimizar una función de varias variables, en este caso **e'e** función de $\hat{\beta}_1, \hat{\beta}_2, \ldots, \hat{\beta}_k$, la condición de primer orden, como en el caso simple, es que las derivadas parciales sean iguales a cero. En notación matricial se obtiene el vector $\hat{\beta}$ directamente de 2.63 simplemente diferenciando **e'e** respecto a $\hat{\beta}$.[29]
quedando

$$\frac{\partial\, \mathbf{e'e}}{\partial \hat{\beta}} = -2\mathbf{X'Y} + 2\mathbf{X'X}\hat{\beta} = 0$$

y despejando se obtiene el llamado *sistema de ecuaciones normales*

$$(\mathbf{X'X})\hat{\beta} = \mathbf{X'Y} \qquad 2.64.$$

Bajo el supuesto de que la matriz **X'X** es invertible, se multiplican ambos lados por su inversa; y como es sabido que [$(\mathbf{X'X})^{-1}(\mathbf{X'X})$] proporciona la matriz identidad (en este caso de orden $k \times k$), se obtiene la siguiente ecuación, siempre que la inversa exista:

$$\hat{\beta} = (\mathbf{X'X})^{-1}\mathbf{X'Y} \qquad 2.65.$$

[29] Para derivar vectores y matrices, se utilizan las siguientes reglas
$\frac{\partial (a'X)}{\partial X} = a$ y $\frac{\partial (X'AX)}{\partial X} = 2AX = 2X'A$. En este caso el vector a es β y la matriz A es $\mathbf{X'X}$.

que es el estimador mínimo-cuadrático del vector de parámetros β.

Se comprende ahora la necesidad de los supuestos adicionales mencionados anteriormente: para hallar los estimadores mínimo-cuadráticos es necesario hallar la inversa de la matriz **X'X**, y el rango de esta matriz cuadrada debe de ser *k*.

- **Media de $\hat{\beta}$ y matriz de Varianza-Covarianza de $\hat{\beta}$.**

El vector $\hat{\beta}$ es aleatorio ya que según 2.65 es función del vector **Y**, y por lo tanto es función de **e**. Esto significa como en el caso simple, que los valores de $\hat{\beta}$ vienen dados en función de las observaciones muéstrales, así que dependen de la muestra empleada. Las principales características de un vector aleatorio son su esperanza matemática y su varianza, las cuales se necesitan para fines de inferencia estadística.

Combinando las ecuaciones 2.65 y 2.52, se puede expresar $\hat{\beta}$ en función del vector de errores, con el objeto de obtener su esperanza matemática

$$\begin{aligned}\hat{\beta} &= (\mathbf{X'X})^{-1}\,\mathbf{X'}(\mathbf{X}\hat{\beta}+\mathbf{u}) \\ &= (\mathbf{X'X})^{-1}\,\mathbf{X'X}\hat{\beta} + (\mathbf{X'X})^{-1}\,\mathbf{X'u} \\ &= \hat{\beta} + (\mathbf{X'X})^{-1}\,\mathbf{X'u} \qquad\qquad 2.66.\end{aligned}$$

aplicando el operador E y tomando en cuenta el supuesto c (**X** no aleatoria), se tiene

$$\begin{aligned}\mathrm{E}(\hat{\beta}) &= \beta + (\mathbf{X'X})^{-1}\mathbf{X'}\,\mathrm{E}(\mathbf{u}) \\ &= \beta \qquad\qquad 2.67.\end{aligned}$$

ya que E(**u**)=0 según supuesto a. Por tanto, $\hat{\beta}$ es un *estimador insesgado* de β.

Utilizando cálculo matricial se obtiene no sólo las varianzas de los estimadores, sino también sus covarianzas. Como el estimador $\hat{\beta}$ es una variable aleatoria, la matriz de varianzas y covarianzas se define como

$$Var(\hat{\beta}) = V(\hat{\beta}) = E\left[(\hat{\beta} - \beta)(\hat{\beta} - \beta)'\right] \qquad 2.68.$$

De la diferencia de $\hat{\beta}$ y $E(\hat{\beta})$ (que es igual a β), según se deduce de 2.66, resulta

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

por tanto

$$\begin{aligned} V(\hat{\beta}) &= E\left\{\left[(X'X)^{-1}X'u\right]\left[(X'X)^{-1}X'u\right]\right\} \\ &= E\left[(X'X)^{-1}X'uu'X(X'X)^{-1}\right] \\ &= (X'X)^{-1}XE(uu')X(X'X)^{-1} \quad \text{por el supuesto b} \\ &= (X'X)^{-1}X'\sigma^2 IX(X'X)^{-1} \end{aligned}$$

y se tiene

$$= \sigma^2 (X'X)^{-1} \qquad 2.69.$$

donde σ^2 es la varianza homocedástica de u_i y $(X'X)^{-1}$ es la matriz inversa que aparece en la ecuación 2.65.

La matriz $V(\beta)$ es simétrica, siendo sus elementos diagonales las varianzas de los estimadores, y los elementos fuera de la diagonal principal las covarianzas entre los distintos elementos de $\hat{\beta}$.

Así la varianza del estimador $\hat{\beta}_j$ (j=1,2,...,*k*) es igual a la varianza de los errores aleatorios multiplicada por el elemento a_{jj} de la matriz $(X'X)^{-1}$, y la covarianza de los estimadores $\hat{\beta}_j$, $\hat{\beta}_s$ es igual a la varianza de los errores aleatorios multiplicada por el elemento a_{js} de la matriz $(X'X)^{-1}$. Por tanto, las fórmulas para la varianza de un determinado estimador y la covarianza de dos estimadores cualesquiera adopta la forma

$$\begin{aligned} \text{var}(\hat{\beta}_j) &= \sigma^2 a_{jj} \\ \text{var}(\hat{\beta}_j, \hat{\beta}_s) &= \sigma^2 a_{js} \end{aligned}$$

donde a_{jj} y a_{js} representan el jj-ésimo y js-ésimo elemento de la matriz $V(\beta)$.

Sin embargo, no se pueden calcular ni el estimador ni la varianza del parámetro $\hat{\beta}$, puesto que para ello se necesita conocer el valor de la varianza del término del error σ^2. Una estimación insesgada de σ^2 se define de nuevo en términos de la suma de cuadrados residual en base a la establecida en la ecuación 2.35. Ajustando la varianza por los grados de libertad perdidos en la estimación de los residuos (k parámetros) se tiene que un estimador insesgado está dado por la expresión:

$$S_e^2 = \frac{\sum e_i^2}{n-k} = \frac{\mathbf{e'e}}{n-k} \qquad 2.70.$$

Al igual que en el caso simple, la estimación S^2 es una medida de la variación de los errores de predicción o residuales. El cálculo $\mathbf{e'e}$ que aparece en 2.70, se hace fácilmente mediante la expresión obtenida de 2.63 después de sustituir el último $\hat{\beta}$ por su valor dado en 2.65, como se muestra a continuación

$$\begin{aligned} \mathbf{e'e} &= \mathbf{Y'Y} - 2\hat{\beta}'\mathbf{X'Y} + \hat{\beta}'\mathbf{X'X}\left[(\mathbf{X'X})^{-1}\mathbf{X'Y}\right] \\ &= \mathbf{Y'Y} - 2\hat{\beta}\mathbf{X'Y} + \hat{\beta}'\mathbf{X'Y} \\ &= \mathbf{Y'Y} - \hat{\beta}'\mathbf{X'Y} \end{aligned}$$

Así, una vez estimada la varianza residual, basta sustituirla en la matriz de varianzas-covarianzas del vector de estimadores β, para obtener el estimador insesgado de dicha matriz quedando:

$$V(\hat{\beta}) = S_e^2(\mathbf{X'X})^{-1} \qquad 2.71.$$

El resultado más importante de la teoría de los mínimos cuadrados es que no existe otro estimador lineal e insesgado que tenga menor varianza que la de los estimadores MCO dada en 2.69. Así pues, como en la regresión lineal simple, los estimadores MCO son los mejores estimadores lineales e insesgados de β (MELI). Esta propiedad se extiende para todo el vector β, el cual como se explicó en la sección B.3.1 también posee todas las propiedades asintóticas.

C.3 TIPOS DE CONTRASTE DE VALIDEZ DE LA ECUACIÓN.

Cuando se dispone de muchas variables explicativas en un modelo, en econometría se aconseja acudir a un conjunto de criterios, alternativos unas veces, complementarios otras, que globalmente permiten juzgar la mayor o menor idoneidad del modelo uniecuacional. El proceso de contraste y validación del modelo puede realizarse de distintas formas, a saber, en base a los criterios económicos, estadísticos y econométricos que se explicaron en el capítulo anterior (sección A.3.).

En este apartado se presentan las herramientas básicas sobre los criterios estadísticos y econométricos para validar el modelo de regresión[30]; los criterios económicos fueron presentados en el capítulo anterior.

Antes de continuar, se exponen algunos conceptos generales sobre los contrastes de significancia estadística, necesarios para entender las pruebas dadas posteriormente.

- **Pruebas de hipótesis.**

Dentro de la inferencia estadística (junto con la estimación de parámetros poblacionales) se encuentra la prueba de hipótesis cuyo objetivo es probar si la afirmación que se hace sobre un parámetro poblacional –basado en las conclusiones obtenidas para la muestra- es correcta o incorrecta. Tal afirmación se conoce como hipótesis o valor supuesto para el parámetro sobre el que se debe tomar una decisión lo cual involucra algunas pruebas estadísticas. Una *hipótesis estadística* es pues una afirmación o suposición que se hace sobre los parámetros de una distribución de probabilidad de una variable aleatoria.

En la estructura de una prueba de hipótesis, la hipótesis propuesta o hipótesis sometida a análisis se conoce como **hipótesis nula** y se denota con el símbolo H_0; la hipótesis nula suele probarse contra la **hipótesis alternativa** o *hipótesis de investigación* que se denota con H_1.

[30] Para profundizar más sobre estos contrastes en econometría, se puede consultar por ejemplo Pulido 1993.

En un proceso de contraste de hipótesis, hay dos errores que se pueden cometer: **error tipo I,** se comete cuando se rechaza una hipótesis que es cierta; **error tipo II,** se comete al aceptar una hipótesis que es falsa. Esta situación se presenta en la tabla 2.4.

DECISIÓN ESTADÍSTICA	**SITUACIÓN VERDADERA**	
	H_0: cierta	**H_0: falsa**
Aceptar H_0	Decisión correcta $(1 - \alpha)$	Error tipo II β
Rechazar H_0	Error tipo I α	Decisión correcta $(1 - \beta)$

Tabla 2.4 Decisión estadística.

La bondad de una prueba estadística de hipótesis se evalúa mediante las probabilidades de cometer los errores tipo I y tipo II, que se denotan por los símbolos α y β respectivamente.

El *nivel de significación* se refiere a la probabilidad α de cometer un error de tipo I, es decir, rechazar una hipótesis verdadera. El *nivel de confianza* se refiere a la probabilidad $(1 - \alpha)$ de aceptar una hipótesis verdadera.

La probabilidad de cometer un error de tipo II, varía dependiendo del verdadero valor del parámetro de la población. Sin embargo, en la práctica β es desconocido, ya sea porque nunca se calculó antes de realizar la prueba, o porque puede resultar extremadamente difícil calcularla. El hecho de que a menudo β es desconocido explica el porqué se pretende apoyar la hipótesis alternativa por medio de un rechazo de la hipótesis nula. Una probabilidad muy interesante (cuando se puede calcular), es $(1 - \beta)$ denominada potencia de la prueba; *potencia de la prueba* es la probabilidad de rechazar correctamente la hipótesis nula cuando debe ser rechazada, es decir, la hipótesis nula es falsa.

En la práctica se acostumbra emplear un nivel de significación de $\alpha = 0.05$, lo cual significa que al establecer una hipótesis sobre algún parámetro poblacional se está

dispuesto a correr un riesgo del 5% de rechazar dicha hipótesis cuando debería ser aceptada, es decir, la probabilidad de cometer un error de tipo I. En otras palabras, se tiene un 95% de confianza en tomar la decisión correcta. En ocasiones se utiliza un nivel muy alto de significación del 1%.

- **Elementos de una prueba estadística.**

Una prueba estadística involucra los siguientes elementos:

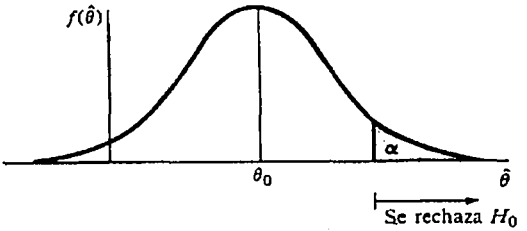
- La hipótesis nula.
- La hipótesis alternativa.
- El estadístico de prueba.
- Región de rechazo (RR).

En referencia a las hipótesis nula y alternativa solo se destaca que la hipótesis nula se debe especificar en la mayoría de las ocasiones en forma contraria a lo que se supone cierto, y entonces la hipótesis alternativa será aquella que se desea apoyar. El **estadístico de prueba** es la variable de decisión que se calcula a partir de los datos de la muestra y se utiliza para tomar una decisión sobre H_0.

El conjunto de todos los valores posibles de la estadística de prueba, se divide en dos: la *región de rechazo* y la *región de aceptación*. La **región de rechazo** consta de aquellos valores del estadístico de prueba que son de tal magnitud que de ser el valor observado del estadístico igual a uno de ellos, la hipótesis nula se rechaza y se acepta la hipótesis alternativa. La **región de aceptación** es el complemento de la región de rechazo; si el valor esperado del estadístico de prueba cae dentro de la región de aceptación, H_0 se acepta o se considera que los datos no proporcionan evidencia suficiente para concluir que es falsa.

Para ilustrar estos elementos, supóngase que se desea probar una hipótesis acerca de un parámetro θ y que se tiene un estimador $\hat{\theta}$ el cual sigue una distribución normal con media θ_0 y varianza $\sigma^2_{\hat{\theta}}$. Si θ_0 es un valor específico de θ, se puede probar

H_0:$\theta = \theta_0$ frente a H_0:$\theta > \theta_0$. La región de rechazo se muestra en la figura 2.12. El valor crítico del estadístico de prueba denotado por C representa el valor de $\hat{\theta}$ que separa las dos regiones de probabilidad. Si se supone cierta la hipótesis nula, la probabilidad de rechazo α es el área bajo la curva normal que queda sobre la región de rechazo. Como se aprecia, la región crítica se determina solamente después de haber planteado H_1.



El tamaño de la muestra debe ser suficientemente grande ($n \geq 30$) para que el estimador se distribuya normalmente. Si se usa $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$ como estadístico de prueba, la región de rechazo puede escribirse como RR = {Z > z_α}. Z es simplemente la desviación respecto a θ_0 de una variable aleatoria normalmente distribuida $\hat{\theta}$, expresada en unidades de desviación estándar $\sigma_{\hat{\theta}}$. Por tanto, una forma de indicar la prueba de hipótesis con α= P(error tipo I) es:

H_0: $\theta = \theta_0$

H_1: $\theta > \theta_0$

Estadístico de prueba: $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$; donde $\sigma_{\hat{\theta}} = \dfrac{\sigma}{\sqrt{n}}$ [31]

[31] La desviación estándar del parámetro θ, será la definida por el *teorema del límite central* el cual enuncia que dada una muestra $X_1, X_2,...,X_n$ de una distribución con media μ y varianza σ^2, la distribución de la media muestral $\overline{X}$ se aproxima a la distribución normal con media μ y varianza σ^2/n conforme n tiende a infinito, o sea que, la variable aleatoria $(\overline{X}-\mu)/(\sigma/\sqrt{n})$ tiene como límite una distribución normal estándar. Esta variable aleatoria se emplea para formular inferencias acerca de μ

Región de rechazo: $Z > z_\alpha$

nótese que se rechaza H_0 si Z cae muy alejado en la cola superior de la distribución normal estándar. La hipótesis alternativa H_1: $\theta > \theta_0$ se denomina alternativa de *cola superior*, y a $RR = \{Z > z_\alpha\}$ se le llama *región de rechazo de cola superior* (o *derecha*).

Una prueba en la que la hipótesis es H_0: $\theta = \theta_0$ frente a H_1: $\theta < \theta_0$ se realiza de manera similar pero ahora se rechaza H_0 para los valores de $\hat{\theta}$ mucho menores que θ_0. El estadístico de prueba sigue siendo Z, pero para un valor de θ dado se rechaza la hipótesis nula cuando $Z < -z_\alpha$. En este caso se denomina a H_1: $\theta < \theta_0$ una alternativa de *cola inferior* y a $RR = \{Z < -z_\alpha\}$ una región de rechazo de cola inferior o *izquierda*). Esta situación se presenta en la figura 2.13.

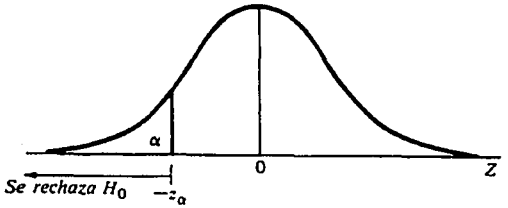


Figura 2.13.

Si se desea probar H_0: $\theta = \theta_0$ frente a H_1: $\theta \neq \theta_0$, entonces se rechazaría H_0 si $\hat{\theta}$ es mucho menor o mucho mayor que θ_0. El estadístico de la prueba sigue siendo Z, pero la región de rechazo se localiza simétricamente en las dos colas de la distribución de probabilidad para Z. Por tanto se rechaza H_0 si $Z < -z_{\alpha/2}$ o $Z > z_{\alpha/2}$; dicho de otro modo se rechaza H_0 si $|Z| > z_{\alpha/2}$. Esta prueba se denomina *prueba de dos colas* (Fig. 2.14).

cuando se conoce el valor de la varianza poblacional σ^2; cuando la varianza poblacional es desconocida, la distribución tendrá varianza S^2/n.

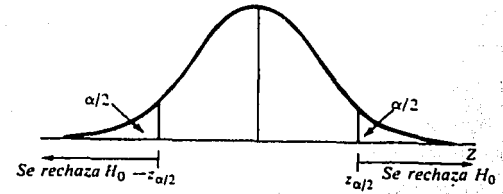


Figura 2.14.

Como frecuentemente el tamaño de la muestra es pequeño (n < 30) o la varianza poblacional es desconocida, se debe utilizar la varianza muestral. En este caso es común usar el estadístico de prueba

$$T = \frac{\hat{0} - 0_0}{\sigma_{\hat{\theta}}} \qquad \text{donde } \sigma_{\hat{\theta}} = \frac{S}{\sqrt{n}}^{32} \qquad 2.72.$$

denominado ***t* de Student** distribuido con n-1 grados de libertad (g. de l.). A continuación se presenta un resumen de las pruebas de hipótesis cuando la muestra es pequeña o la varianza es desconocida:

H_0: $\theta = \theta_0$

H_1:
- $\theta > \theta_0$ (alternativa de cola superior)
- $\theta < \theta_0$ (alternativa de cola inferior)
- $\theta \neq \theta_0$ (alternativa de dos colas)

Estadístico de prueba: $T = \frac{\hat{0} - 0_0}{\sigma_{\hat{\theta}}}$; donde $\sigma_{\hat{\theta}} = \frac{S}{\sqrt{n}}$

Región de rechazo:
- $T > t_\alpha$ (RR de cola superior o cola derecha)
- $T < -t_\alpha$ (RR de cola inferior o izquierda)
- $|T| > t_{\alpha/2}$ (RR de dos colas)

[32] Recuérdese que para el modelo de regresión S representa el error estándar de la regresión.

- **Relación entre procedimientos de intervalos de confianza y prueba de hipótesis.**

A partir de la distribución muestral de Θ es posible determinar $\hat{\theta}_I$ y $\hat{\theta}_S$ de modo que la probabilidad sea igual a un valor fraccionario positivo tal que

$$P\left(\hat{\Theta}_I < \theta < \hat{\Theta}_S\right) = 1 - \alpha$$

donde $0 < \alpha < 1$, entonces se tiene una probabilidad igual $(1 - \alpha)$, de seleccionar una muestra aleatoria que producirá un intervalo que contenga a θ. El intervalo $\hat{\theta}_I < \theta < \hat{\theta}_S$, calculado a partir de la muestra, se llama **intervalo de confianza** de $(1-\alpha)100\%$, la fracción $(1 - \alpha)$ se denomina **coeficiente de confianza** o grado de confianza, y los extremos $\hat{\theta}_I$ y $\hat{\theta}_S$ reciben el nombre de **límites de confianza superior e inferior**. Por ejemplo, para un $\alpha = 0.05$ se tiene un intervalo de confianza de 95%. Idealmente es preferible un intervalo corto con un alto grado de confianza de contener el valor del parámetro.

Como se aprecia, existe una estrecha relación entre los intervalos de confianza y las pruebas de hipótesis. Suponiendo que se de desea probar una hipótesis al nivel de significación del 5%, entonces es posible construir un intervalo de confianza del 95% para el parámetro que se considera y ver si el valor hipotético se encuentra en dicho intervalo. Si es así, no se rechaza la hipótesis, sino no lo es, se rechaza. Esta relación es válida para las pruebas de valores de parámetros.

- **Nivel de significancia observada.**

En lugar de preseleccionar α a niveles arbitrarios y delimitar la región de rechazo, un segundo método es obtener el **nivel de significancia observado** o *valor p* (valor de probabilidad), el cual se define como el mínimo nivel de significación α para el cual los datos observados indican que se tendría que rechazar la hipótesis nula. Es una forma fácil de presentar los resultados ya que diversos paquetes estadísticos imprimen el valor

p de los estadísticos de prueba estimados. Corresponderá al investigador seleccionar el valor máximo de α que se está dispuesto a tolerar; por ejemplo, supóngase que en una aplicación se tiene un valor p de 0.045, entonces será significativo al nivel de 0.05, pero no al nivel de 0.01.

- **Análisis de varianza.**

Para probar la significancia de un grupo de parámetros, es común hacerlo por medio de un **análisis de varianza** [33]. El análisis de varianza es una técnica estadística muy poderosa para contrastar la igualdad de varias medias poblacionales, mediante observaciones muéstrales. Aunque es posible diseñar un análisis de varianza de n-entradas, para el presente trabajo se utilizará un análisis de varianza de dos entradas o de *un solo factor.*

En general, supóngase que se quieren comparar las medias de k poblaciones, *cada una de las cuales se supone que tiene la misma varianza.* Tomando para estas poblaciones muestras aleatorias independientes de tamaño $n_1, n_2,..., n_k$, respectivamente; utilizando x para designar los valores muéstrales actuales de manera que x_{ij} designará la j-ésima observación en la población i-ésima. Ahora, para un solo factor, supóngase que se tiene muestras aleatorias independientes con $n_1, n_2,..., n_k$ observaciones de k poblaciones, siendo $\mu_1, \mu_2,..., \mu_k$ las medias poblacionales, entonces al **análisis de varianza para un solo factor** deberá contrastar la hipótesis nula

$$H_0: \quad \mu_1 = \mu_2 = \ldots = \mu_k$$

si el valor del estadístico de prueba lleva a aceptar H_0, se concluirá que las diferencias observadas entre las medias de las muestras se deben a la variación casual en el muestreo. Si se rechaza H_0, se concluirá que las diferencias entre los valores medios de la muestra son demasiado grandes para deberse únicamente al azar (y por ello no todas las medias de la población son iguales).

[33] ADV, o ANOVA de las siglas en inglés Analisys of variance.

La prueba está basada en un análisis de la variabilidad total de los datos, que está dado por

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}..)^2 \qquad \text{donde} \qquad \bar{x}..= \frac{1}{nk} \cdot \sum_{i=1}^{k}\sum_{j=1}^{n} x_{ij}$$

dado que esta variación total es en realidad, la suma de las desviaciones al cuadrado, también se le llama la **suma de cuadrados total** (SS total). Esta variación total, mide las diferencias entre cada valor y la media $\bar{x}..$ total (gran media). Para separar estas dos contribuciones de la variabilidad total de los datos, se tiene la siguiente descomposición:

$$\sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}..)^2 = \sum_{i=1}^{k}\sum_{j=1}^{n}(\bar{x}_{i.} - \bar{x}..)^2 + \sum_{i=1}^{k}\sum_{j=1}^{n}(x_{ij} - \bar{x}_{i.})^2 \qquad 2.73.$$

donde $\bar{x}_{i.}$ es la media de las observaciones de la *i*-ésima población y $\bar{x}..$ es la media de todas las *nk* observaciones. En general, al primer término de la expresión del lado derecho se le conoce como **variación entre grupos**, mide la diferencia entre la media $\bar{x}_{i.}$ de cada grupo y la media total $\bar{x}..$, ponderadas por el número de observaciones de cada grupo. El segundo miembro de lado derecho se le conoce como **variación dentro de los grupos** y mide la variación de cada valor en relación con la media de su propio grupo y acumula estas diferencias elevadas al cuadrado en todos los grupos.

La descomposición de la suma total de cuadrados en la suma de los dos componentes, constituye la base para el contraste de igualdad de la medias poblacionales de los grupos. Como se sabe la varianza se calcula al dividir la suma de las desviaciones al cuadrado entre sus correspondientes grados de libertad. En el análisis de varianza, la asuma de las desviaciones al cuadrado se representa por las respectivas medias de variación. Se puede demostrar que una estimación insesgada de la varianza poblacional resulta de dividir la suma de cuadrados dentro de los grupos por *n-k* (también conocida como *suma de errores al cuadrado*); y si las medias poblacionales son iguales, otra estimación insesgada de la varianza poblacional se obtendrá dividiendo suma de cuadrados de los grupos por *k*-1.

La varianza dentro de los grupos, denotada por S_w^2, mide la variabilidad en torno a la media de cada grupo; como su variabilidad no se afecta por las diferencias de los

grupos, se puede considerar una media de la variación aleatoria de los valores dentro de un grupo: Por otra parte, la varianza entre los grupos, denotada por S_B^2, tiene en cuenta no sólo las variaciones aleatorias de una observación a otra, sino también mide las diferencias entre un grupo y otro. Si no hay diferencia real entre un grupo y otro, las diferencias en la muestra se pueden explicar por la variación aleatoria, y la varianza entre los grupos S_B^2 debe estar cercana a la varianza dentro de los grupos S_w^2. No obstante, si realmente hay una diferencia entre los grupos S_B^2, la varianza entre los grupos será bastante mayor que la varianza dentro de los grupos S_w^2.

Para probar la hipótesis nula de igualdad de medias, debido a que hay dos varianzas implicadas el contraste se realiza en base a la distribución *F de Snedecor* la cual queda definida como el cociente de dos ji-cuadradas divididas por sus respectivos g.l. e independientes entre sí; es decir

$$F = \frac{S_B^2}{S_w^2}$$

Como se explicó anteriormente, si hubiera una diferencia entre los grupos, la varianza entre los grupos sería bastante mayor que la varianza dentro de los grupos. Por tanto, la regla de decisión sería rechazar la hipótesis nula de que no hay diferencia entre los grupos si

$$F = \frac{S_B^2}{S_w^2} > F_{\alpha(K-1,\ N-k)} \qquad 2.74.$$

para un nivel de significancia α. Por tanto, el análisis de varianza resulta ser una prueba F de una cola.

Los cálculos para llevar a cabo este contraste se presentan generalmente en una tabla de análisis de varianza, como la que se muestra en la tabla 2.5.

A continuación se realizará la aplicación de los conceptos anteriores para probar la significancia de los parámetros en el modelo de regresión lineal.

Fuente de variación	Suma de cuadrados	g.l.	Media cuadrada de la varianza	F
Entre los grupos	SS_B^2	$K-1$	$S_B^2=\frac{SS_B}{k-1}$	$\frac{S_B^2}{S_w^2}$
Dentro de grupos	SS_w^2	$N-k$	$S_w^2=\frac{SS_w}{n-k}$	
Total	SST	$N-1$		

Tabla 2.5 Formato general ANOVA para un solo factor.

C.3.1 Prueba t de Student.

Como se ha dicho para poder hacer pruebas de hipótesis y calcular intervalos de confianza para los coeficientes del modelo de regresión, es necesario usar el supuesto de normalidad del término del error.

Ya que los estimadores mínimo cuadráticos $\hat{\beta}_j$ (j=1,2,...,*k*) son funciones lineales de los errores aleatorios, se distribuyen también normalmente. En notación matricial la media definida por la ecuación 2.67 y la varianza definida en 2.69 se expresan como

$$\hat{\beta} \sim N\left[\beta, \sigma_u^2(\mathbf{X'X})^{-1}\right] \qquad 2.75.$$

Al ser σ^2 desconocida no se puede utilizar la distribución normal, y se cumple que cada elemento de $\hat{\beta}$ se distribuye como una *t* de Student con *n-k*-1 (a semejanza de ecuación 2.72) grados de libertad, esto es

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{S_e^2 a_{jj}}} \qquad j = 1,2,...,k.$$

Dado que la raiz de $S_e^2 a_{jj}$ es *el error estándar del estimador* $\hat{\beta}_j$, que se puede denotar por $S_{\hat{\beta}_j}$, se concluye que

$$\frac{\hat{\beta}_j - \beta_j}{Se(\hat{\beta}_j)} \sim t_{n-k-1} \qquad 2.76.$$

para un nivel de significancia α. El estadístico t puede utilizarse para crear un intervalo de confianza del 100(1 - α)% para el parámetro β_j dado por

$$\hat{\beta}_j \pm t_{\alpha/2} S_e(\hat{\beta}_j) \qquad 2.77.$$

donde $t_{\alpha/2}$ es un valor de la distribución t con n-k-1 g. de l.

Para probar la hipótesis H_0: $\beta_j = \beta_{jo}$ contra la hipótesis alternativa H_1: $\beta_j \neq \beta_{jo}$ con un nivel de significancia de $\alpha/2$ con n-k-1 g. de l, se calcula

$$t = \frac{\hat{\beta}_j - \beta_{jo}}{S_e(\hat{\beta}_j)}$$

y se rechaza si $|t| > t_{\alpha/2}$. O alternativamente se calcula el intervalo de confianza con los valores críticos:

$$h_1 = \beta_{jo} - t_{\alpha/2} S_e(\hat{\beta}_j); \qquad h_2 = \beta_{jo} + t_{\alpha/2} S_e(\hat{\beta}_j)$$

Particularmente es importante contrastar la hipótesis nula que establece si $\beta_j = 0$, puesto que se desea determinar si alguna variable en particular tiene o no influencia significativa en la variable dependiente al ser el parámetro diferente de cero. En este caso se prueba

$$H_0: \quad \beta_j = 0 \qquad j = 1,2, \ldots, k.$$
$$H_1: \quad \beta_j \neq 0 \qquad j = 1,2, \ldots, k.$$

con el estadístico

$$t = \frac{\hat{\beta}_j - 0}{S_e(\hat{\beta}_j)} = \frac{\hat{\beta}_j}{S_e(\hat{\beta}_j)} \qquad j = 1,2, \ldots, k. \qquad 2.78.$$

y la hipótesis se rechaza si

$$|t| > t_{\alpha/2}$$

Recuérdese que si la decisión se basa en un valor p, no es necesario establecer la región crítica y se rechazará H_0 si el valor es mayor o igual a un nivel de significancia α dado.

C.3.2 Análisis de varianza.

En la sección anterior se ha determinado la significancia estadística individual de los coeficientes de regresión, ahora el objetivo es establecer una hipótesis nula para saber si los parámetros involucrados en el modelo contribuyen en forma conjunta con información para la predicción de Y, lo cual se expresa como

$$H_0 : \beta_2 = \beta_3 = \ldots \beta_k = 0$$
$$H_1 : \beta \neq 0, \qquad \text{para algún } j = 2,3,\ldots,k.$$

La hipótesis H_0 indica que la variable dependiente Y, no está relacionada en forma lineal y simultánea con todas las variables independientes (X_2, X_3,..., X_k) que constituyen el modelo. La hipótesis H_1 indicará que en forma lineal y simultánea algunas variables independientes explican satisfactoriamente los cambios de la variable dependiente.

Se hace la aclaración que el hecho de rechazar a H_0 no necesariamente implica que la ecuación estimada de regresión sea útil para efectuar predicciones; se necesita profundizar el análisis antes de que se pueda dar un juicio definitivo sobre la utilidad de la ecuación de regresión.

Con frecuencia el problema de analizar la calidad de la línea de regresión estimada se maneja a través de un enfoque de análisis de varianza. Como se recordará, esta técnica divide la variación total de las observaciones en sus partes componentes de

acuerdo con el modelo propuesto (ec. 2.73). Para el modelo de regresión se realiza la siguiente descomposición de la varianza:

$$\underset{\text{suma total}}{\Sigma(Y_i - \overline{Y})^2} = \underset{\text{suma explicada}}{\Sigma(\hat{Y}_i - \overline{Y})^2} + \underset{\text{suma residual o no explicada}}{\Sigma(Y_i - \hat{Y}_i)^2} \qquad 2.79.$$

simbólicamente

$$\mathbf{SCT = SCE + SCR} \qquad 2.80.$$

siendo **SCT** la **suma de cuadrados total**, que refleja la variación total de las observaciones Y_i con respecto a su media muestral sin considerar la variable de predicción. **SCE** es la **suma de cuadrados explicada** o también conocida como la *suma de cuadrados de la regresión*, que refleja la variación de los valores de Y explicados por el modelo, en este caso la línea recta postulada. **SCR** es la ya conocida **suma de cuadrados residual** o de los *errores* o *no explicada*, que refleja la variación alrededor de la línea de regresión.[34]

Por claridad se hace hincapié en el modelo de regresión simple para explicar el razonamiento que implica el análisis de varianza en la regresión. Esta explicación se extiende al modelo de regresión múltiple, excepto que los valores se basan en diversas variables independientes en lugar de una. De esta manera basándose en la figura 2.10 se construye la figura 2.15

Por ser SCE/σ^2 y SCR/σ^2 valores de variables independientes ji-cuadradas con k y $n-k-1$ g.l. respectivamente; y SCT/σ^2 una variable ji-cuadrada con $n-1$ g.l.; la estadística apropiada para probar H_0 (como en 2.74) será:

$$F = \frac{SCE/k}{SCR/n-k-1} = \frac{SCE/k}{S^2} \qquad 2.81.$$

[34] En los libros de estadística SCR se utiliza para la suma de cuadrados de la regresión y SCE para denotar la suma de cuadrados de los errores. Sin embargo, se ha utilizado esta notación la cual se encuentra en la mayoría de libros de econometría la cual coincide con los nombres del paquete estadístico a usar.

la cual sigue una distribución F con k y $n - k - 1$ g. de l. (se estimarán k parámetros, incluyendo el intercepto). Se rechaza la hipótesis nula al nivel de significancia de α cuando $f > f_{\alpha(k, n-k-1)}$.

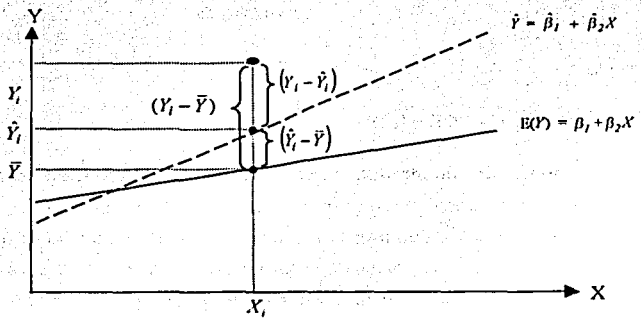


Figura 2.15.

Estos cálculos se resumen en una tabla de análisis de varianza como se muestra en la tabla 2.6. Los *cuadrados medios* no son otra cosa que las diferentes sumas de cuadrados divididas entre sus respectivos grados de libertad; en este caso **CME** denota el **cuadrado medio explicado** y **CMR** representa el **cuadrado medio residual**.

Fuente de variación	Suma de cuadrados	g. l.	Cuadrados medios	F
Debida a la regresión	SCE	k	SCE/k	$\frac{CME}{CMR}$
Debida al error	SCR	$n-k-1$	$S^2 = SCR/(n-k-1)$	
Total	SCT	$n-1$		

Tabla 2.6 Análisis de varianza para probar $\beta = 0$.

Existe una relación básica entre la distribución t con ν grados de libertad y la distribución F con 1 y ν grados de libertad que es

$$t^2_{\alpha/2} = f_\alpha(1, \nu) \qquad 2.82.$$

La prueba t permite probar el rechazo de una alternativa bilateral mientras que, para este procedimiento, la prueba F se limita a probar el rechazo de una alternativa unilateral; es decir, si existe solo una variable independiente en el modelo de regresión, entonces la prueba F equivale a la prueba t de dos extremos realizada sobre β, es por eso que en el análisis de regresión simple no se requiere la prueba F.

C.3.3 Coeficiente de determinación simple y ajustado.

Existen dos estimadores muy importantes en econometría que se utilizan con frecuencia para determinar la capacidad explicativa de un modelo, es decir, de la bondad de ajuste. Estos son: el *coeficiente de determinación múltiple simple* y el *ajustado*.

Basándose en la descomposición de la varianza según la ec. 2.79 y 2.80, el **coeficiente de determinación múltiple** R^2 queda definido por:

$$R^2 = \frac{SCE}{SCT} = \frac{\Sigma(\hat{Y}_i - \overline{Y})^2}{\Sigma(Y_i - \overline{Y})^2} \qquad 2.83.$$

y mide qué proporción de la variación de la variable endógena Y, queda explicada por la variación conjunta de las variables exógenas, X_2, X_3,... ,X_k; dicho de otro modo, R^2 es una medida relativa de qué tanto las variables de predicción incluidas en el modelo explican la variación de las observaciones. La raíz cuadrada de R^2 se conoce como ***coeficiente de correlación múltiple R*** y muestra la medida de la relación entre el conjunto de variables $X's$ y la variable dependiente Y.

Muchas veces se interpreta incorrectamente el coeficiente de determinación. R^2 no puede verificar que la verdadera ecuación de regresión entre X y Y sea estrictamente lineal, solo puede medir cuánto se explica de la variación total por medio de la regresión estimada.

Despejando SCE en 2.80 y sustituyendo en 2.83, se deduce que el coeficiente de determinación puede expresarse también como:

$$R^2 = 1 - \frac{SCR}{SCT} = 1 - \frac{\Sigma c_i^2}{\Sigma(Y_i - \overline{Y})^2} \qquad 2.84.$$

cuyos límites están comprendidos en el intervalo

$$0 \leq R^2 \leq 1$$

entre más cercano a 1 es el valor de R^2, mayor es la cantidad de la variación total que puede explicarse por medio de los términos que aparecen en el modelo. Cuando $R^2 = 1$ indica que el modelo se ajustaría perfectamente a los datos, ya que los valores reales y estimados de la variable dependiente serían iguales ($Y_i = \hat{Y}_i$). Si $R^2 = 0$ el modelo de regresión no explicaría en nada la variabilidad de Y, en este caso el valor de $\hat{Y}_i$ sería igual a $\overline{Y}_i$.

Por sí mismo, R^2 no puede validar el modelo propuesto, ni tener un valor cercano a uno implica necesariamente que la ecuación de regresión estimada sea apropiada para la predicción. Por otro lado, un R^2 muy pequeño no implica la inexistencia de la relación entre la variable dependiente e independiente; podría suceder que la influencia de las variables independientes es poco importante comparada con la de los factores incluidos en él termino de perturbación, pero si afectan a Y.

Cuando se utilizan otros métodos de estimación distintos de MCO, o cuando el modelo no tiene ordenada al origen (β_o), no puede efectuarse la descomposición de la suma de cuadrados indicadas en 2.79 y por tanto 2.83 y 2.84 no son equivalentes. En estos casos la expresión que habitualmente se utiliza para el cálculo de R^2 es la señalada en 2.84.

Existe una estrecha relación entre el coeficiente de determinación y la prueba F utilizada en el análisis de varianza para contrastar la hipótesis básica de que no existe correlación real entre las variables, esto es, H_o: $\rho = 0$. Haciendo uso de la ec. 2.81 y 2.83, se llega a la expresión siguiente (para una demostración ver Gujarati)

$$F = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

obteniéndose el valor de R^2 a un nivel de significancia de α directamente de las tablas estadísticas. Sin embargo, no es necesario obtener el valor de R^2 para un α dado ya que se puede rechazar la hipótesis de nulidad si

$$\frac{R^2/(k-1)}{(1-R^2)/(n-k)} \geq F_{\alpha}(k-1, n-k) \qquad 2.85.$$

Cuando R^2 aumenta, también lo hace F y viceversa; esto es, si $R^2 = 1$ implica que F tiende a infinito, si $R^2 = 0$ implica que $F = 0$; es decir, varían en forma directa. Si se acepta la H_1 de que los parámetros estimados por la regresión no son todos nulos en forma conjunta, equivale a decir que R^2 es significativamente diferente de cero. De esta manera la prueba F que es una medida de la significancia para un conjunto de parámetros en la regresión obtenida, es también una prueba de significancia para R^2.

Cabe mencionar dos deficiencias importantes en el coeficiente de determinación: que R^2 no va ha disminuir nunca al añadir más variables en el modelo, es decir, R^2 tiende a aumentar al agregar más variables aunque estas no expliquen nada desde el punto de vista teórico; y que R^2 es un estimador sesgado, este sesgo es negativo y se mantiene aún para muestras muy grandes, esto es, a medida que aumenta el número de observaciones, resulta más difícil un buen ajuste.

Para solucionar este problema, se suele utilizar el **coeficiente de determinación ajustado** o *corregido* $\bar{R}^2$ que se obtiene al ajustar por sus grados de libertad las varianzas

$$\bar{R}^2 = 1 - \frac{SCR/(n-k)}{SCT/(n-1)} = 1 - \frac{n-1}{n-k}(1-R^2) \qquad 2.86.$$

este coeficiente de determinación $\bar{R}^2$ es también sesgado, pero su sesgo es menor que el de R^2. Se puede observar también que el $\bar{R}^2$ ajustado puede disminuir si la variable o variables adicionales producen solamente un pequeño aumento en R^2. De esta forma, conforme aumenta el número de variables k, no está claro cual va a ser la variación que

experimente el coeficiente de determinación ajustado. Además, el R^2 se hace más pequeño a medida que se tiene menor número de datos por variable independiente, por tanto, su valor tenderá invariablemente a ser menor que el valor del coeficiente de determinación estándar; así, de las expresiones 2.84 y 2.86 resulta que la relación entre ambos coeficientes es tal que

$$0 \leq \bar{R}^2 \leq R^2 \leq 1$$ [35]

Es importante comparar este coeficiente siempre en modelos con unidades homogéneas de la variable dependiente, pues con frecuencia se pasa por alto las transformaciones en dicha variable. Por ejemplo, sería un error comparar, el $\bar{R}^2$ de un modelo logarítmico en la variable dependiente, contra el $\bar{R}^2$ de un modelo lineal en la variable dependiente.

C.4 VIOLACIÓN DE LOS SUPUESTOS BÁSICOS DEL MODELO DE REGRESIÓN.

Como se recordará, cuando no se cumple alguno de los supuestos del modelo de regresión lineal origina que los estimadores mínimo-cuadráticos puedan no ser los mejores estimadores lineales e insesgados (MELI), asimismo pueden perder validez las pruebas estadísticas, como son los intervalos de confianza y pruebas de hipótesis.

En regresión múltiple se deben evaluar los supuestos no solo de las variables individuales, sino del valor teórico en sí mismo y de su relación con las variables independientes. Cada supuesto origina un problema específico para la econometría, que en caso de presentarse, deberá ser corregido según la gravedad del mismo.

En este apartado se verán los problemas que surgen en la violación de los supuestos así como su detección y corrección; a saber, en los supuestos de *no linealidad*, *heterocedasticidad*, *autocorrelación*, de *falta de normalidad* y de *multicolinealidad*, por ser los más importantes en el modelo que se está tratando. Si no se cumple el supuesto

[35] Debido al proceso de ajuste puede ocurrir que valores, que por definición deben estar comprendidos entre 0 y 1, den valores negativos; en tal caso debe interpretarse como nulo el coeficiente de determinación ajustado.

de nulidad de la media, afecta al intercepto que resulta ser insesgado; y cuando no se puede determinar que las X's sean no aleatorias, afecta en el uso de ecuaciones simultáneas pues surge el problema de los llamados "regresores estocásticos".[36]

Las gráficas de los residuales y de las variables independientes constituyen el método básico en la identificación del incumplimiento de los supuestos para el conjunto de la relación. Los residuales a menudo proporcionan información que permite modificar y mejorar un modelo de regresión, pues además de la ayuda para examinar los supuestos, también detectan valores en los datos fuera de lo normal.

La gráfica de residuos más común se forma con los residuales frente a los valores de la predicción de la variable dependiente; es decir, e_i^2 contra $\hat{Y}_i$. Para un modelo de regresión simple, se pueden trazar los residuales respecto de las variables dependientes o independientes, dado que están directamente relacionados. En la regresión múltiple, sin embargo, sólo los valores dependientes pronosticados representan el efecto total del valor teórico de regresión; por tanto, a no ser que el análisis residual pretenda concentrarse en una sola variable, se usan las variables dependientes pronosticadas. Hay algunos programas de regresión que calculan automáticamente los residuos y su representación gráfica, lo que es de gran ayuda.

La figura 2.16 contiene unas gráficas de residuos que muestran los supuestos discutidos en las secciones siguientes. Una gráfica de interés especial es la gráfica de la figura 2.16.a, en que se cumplen todos los supuestos y es la gráfica de no correlación de residuos.

- **No linealidad.**

La linealidad de la relación entre variables dependientes o independientes, representa el grado de cambio en la variable dependiente asociado con la variable independiente. El concepto de correlación está basado en una relación lineal, siendo por tanto, un supuesto crítico del análisis de regresión.

[36] Por tanto, este tema se puede consultar en cualquier libro de econometría con ecuaciones simultáneas.

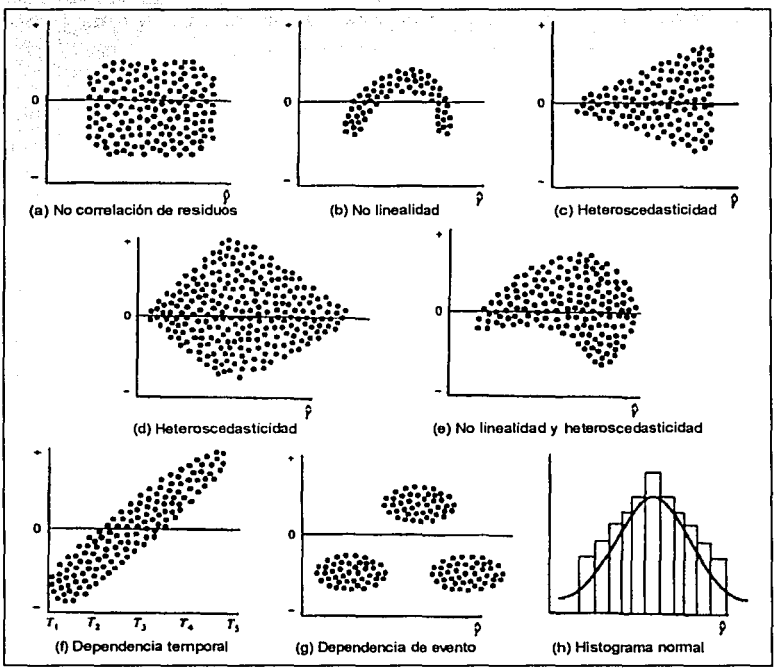


Figura 2.16.

La forma más habitual para evaluar la linealidad en la regresión simple, es examinar las gráficas de dispersión de las variables para identificar cualquier relación no lineal en los datos. En la figura 2.16.b se muestra una forma típica de residuos que indica la existencia de una relación no lineal no incorporada en el modelo. Cualquier modelo curvilíneal de los residuos indica que su corrección aumentará tanto la precisión de la predicción como la validez de los coeficientes estimados. Las soluciones mediante transformaciones de los datos para conseguir la linealidad se presentaron en la sección B.4.

En la regresión múltiple el examen de los residuales muestra los efectos combinados de todas las variables independientes, pero no se puede examinar el efecto de cualquier variable independiente separadamente en un gráfico de residuales.

- **Heterocedasticidad.**

La presencia de varianzas desiguales (heterocedasticidad) es uno de los supuestos que habitualmente no se cumple. El problema de heteroscedasticidad tiende a estar presente cuando se trabaja con datos transversales, ya que en estos estudios es común la variación en los valores que toman las unidades muestrales (individuos o familias, empresas de tamaño diferente, divisiones geográficas variables, etc.). Por el contrario, en las series de tiempo, los cambios experimentales en las variables, son graduales.

También puede generarse a causa de una *especificación incorrecta* del modelo (ya sea por la omisión en la ecuación de una variable explicativa fundamental; o por error al especificar la función de la ecuación, como puede ser especificando una función lineal cuando se trataba de una no lineal); o por un *cambio estructural.*

Independientemente de cual sea la causa que produce la heteroscedásticidad, los estimadores mínimo-cuadráticos siguen siendo insesgados y consistentes pero ya no tienen varianza mínima, aún cuando el tamaño de la muestra aumente indefinidamente (eficiencia asintótica). Además, dado que las varianzas para los parámetros estimados ya no es la mínima, esto origina intervalos de confianza para β muy anchos; por otra parte la pruebas de significancia estadística t y F para los parámetros, resultan incorrectas.

No existen reglas fijas ni procedimientos únicos para la detección del problema de heterocedásticidad. Los hay desde métodos gráficos, hasta test estadísticos basados en comparaciones de varianzas de subgrupos de residuos. Son métodos muy generales ya que lo más común en los estudios econométricos, es tener sólo una muestra de Y para cada valor de X, por lo cual no se puede conocer la varianza poblacional.

La gráfica de residuales de la figura 2.16.a de no correlación de residuos, muestra el patrón de comportamiento de los residuos cuando la varianza es constante. Sin embargo, la forma más común es la triangular en cualquier dirección como se muestra en 2.16.c. También puede esperarse una forma de diamante (Fig. 2.16.d) en donde se espera más variación en los valores intermedios que en los extremos. Muchas veces los incumplimientos de los supuestos ocurren simultáneamente como se presenta en la figura 2.16.e en el caso de no linealidad y heterocedásticidad. Por tanto, las soluciones para cada uno de los problemas de incumplimiento corrigen a menudo otros problemas.

Si existe heterocedasticidad se puede aplicar dos enfoques:

1) *Cuando se conoce* σ^2, se utiliza el llamado método de mínimos cuadrados ponderados el cual minimiza la importancia de los errores con los valores extremos, ponderándolos en proporción inversa a sus varianzas.

2) *Cuando se desconoce* σ^2, que es el caso más común, se tienen varias alternativas, algunas de las más utilizadas en econometría son: la determinación del modelo con variables expresadas en forma de porcentajes o razones, es decir, *variables deflactadas*, de gran eficiencia en aplicaciones prácticas. O la *transformación de datos* de las variables (que dependerá de la naturaleza del problema) para estabilizar la varianza.

Una de las transformaciones más utilizadas en econometría es la **transformación logarítmica**, ya que al aplicar el logaritmo a la ecuación de regresión comprime la escala en que se miden las variables, mejorando el ajuste del modelo. Además en esta transformación el coeficiente pendiente mide la *elasticidad* de Y con respecto a X, que como se recordará es lo mismo que el cambio porcentual en Y debido a un cambio porcentual en X, a diferencia del coeficiente en el modelo original que mide solo la tasa de cambio medio de Y por una unidad de cambio en la variable X. Por ello son de gran aplicación en econometría los modelos logaritmicos.

En general se aplicarán transformaciones según el tipo de problema o según la severidad de la heterocedasticidad. Finalmente un problema que puede surgir al aplicar transformaciones a un modelo de regresión múltiple, es que sin la

transformación hay veces que las variables originales no presentan un problema de correlación, pero al aplicarla se puede provocar este problema por lo menos en algunas variables.

- **Autocorrelación.**

Otro problema común en econometría, es el de autocorrelación o correlación serial el cual viola el supuesto de que los términos del error no son independientes entre sí. Cuando la correlación entre los términos del error se presenta en series de tiempo se denomina *autocorrelación temporal*; y cuando se presenta en datos de corte transversal o series de espacio se le llama *autocorrelación espacial* (correlación en el espacio, no en el tiempo). Por convención se utiliza el subíndice t para series de tiempo y el subíndice i para datos de corte transversal.

El problema de autocorrelación se presenta más frecuentemente en las series de tiempo, debido a que los datos de la serie estadística se ordenan cronológicamente y, particularmente, porque el intervalo de tiempo entre estos datos es constante y corto, como lo es el día, la semana o el mes.

El fenómeno de autocorrelación puede tener diferente origen, citándose entre las causas más importantes:

- *Inercia* de comportamiento de una variable temporal, se refiere a que los valores que toma dicha variable en determinados instantes, están relacionados con los valores alcanzados en otros instantes del tiempo anterior (los conocidos ciclos económicos). Si los valores de la variable explicativa están relacionados entre sí, los del error también lo estarán, es decir, son dependientes.
- *Errores de especificación* del modelo, que es una causa común de heterocedasticidad, se presenta concretamente cuando se *omite una variable* en la especificación del modelo, o cuando se comete un error en la *definición de la forma de la función*, e incluso se da por el *cambio estructural.*
- *Errores de medición* de las variables, que son los errores de sesgo por grupos de datos (unos con sesgo positivo y otros con sesgo negativo) que suponen

sobre o infravaloraciones por periodos de magnitudes de las variables. Son provocados por ciertas prácticas estadísticas de manipulación de datos: extrapolación o interpolación, alisado de series, procedimientos de estimación que varían en el tiempo, etc.

Ante la presencia de autocorrelación los estimadores mínimo-cuadráticos son insesgados y consistentes, pero ineficientes; por tanto, al igual que en la heteroscedasticidad, los intervalos de confianza serán más anchos y las pruebas de significancia serán menos fuertes. Además, si se persiste en ignorar el supuesto, la varianza residual para S^2 subestimará la verdadera pero desconocida σ^2, incluso σ^2 si no está subestimada, las varianzas y errores stándar de los estimadores tienden a subestimar las verdaderas varianzas y errores estándar como resultado de lo anterior; las pruebas t y F proporcionarán conclusiones erróneas sobre los parámetros por lo que pierden validez; aunque los estimadores MCO sean insesgados, para una muestra en particular tienden a dar una visión distorsionada de los verdaderos valores poblacionales, es decir, se vuelven insensibles a las fluctuaciones muestrales. Por ende, la autocorrelación es un problema serio que distorsiona la efectividad del método de mínimos cuadrados.

La autocorrelación como se ilustró en la figura 2.9, puede ser positiva, negativa o cero; siendo la más frecuente en los fenómenos económicos, la positiva. Para la detección se puede recurrir al método gráfico que relaciona los errores de la regresión contra el tiempo, proporcionando el comportamiento de éstos a lo largo del tiempo, como se muestra en la figura 2.16.f. Otra forma habitual se presenta en la figura 2.16.g, la cual ocurre cuando las condiciones básicas del modelo cambian pero no se incluyen en el modelo, por ejemplo, las estaciones del año.

Para corregir la autocorrelación se suele recurrir a la transformación de los datos, tales como modelos de primeras diferencias o modelos de regresión especialmente formulados.

- **No normalidad.**

Quizá el incumplimiento de no normalidad de las variables independientes, dependientes o ambas, es el que más frecuentemente se viola. En este caso los contrastes de significación estadística son inválidos o reducidos en su significación, ya que se requiere la normalidad para los estadísticos t y F.

El diagnóstico más simple para el conjunto de variables independientes en la ecuación, es un *histograma* de residuos donde se puede comprobar visualmente si la distribución de los datos observados se aproxima a la distribución normal, como se aprecia en la figura 2.16.h. Aunque atractivo por su simplicidad, este método es particularmente difícil en muestras pequeñas, donde la construcción del histograma puede distorsionar la representación visual.

Una aproximación de mayor confianza es la **gráfica de cuantiles-cuantiles**[37] **normales**, el propósito de una gráfica de cuantiles es describir, en forma de muestra, los valores de los datos en forma ordenada $Y_{(i)}$ (graficados en el eje horizontal) contra el cuantil correspondiente de la distribución normal $q_{0,1}(f_i)$, donde $f_i = \frac{i - \frac{3}{8}}{n + \frac{1}{4}}$. Una relación cercana a una línea recta sugiere que los datos provienen de una distribución normal; la intersección en el eje horizontal es una estimación de la media de la población y la pendiente es una estimación de la desviación estándar σ.

Para corregir la no normalidad se recurre a la transformación de los datos. Sin embargo, la no normalidad es muchas veces un resultado de otras violaciones de los supuestos; por ello se recomienda examinar este supuesto después o junto con los análisis y soluciones para las otras violaciones.

La falta de normalidad también puede deberse a la presencia de **datos atípicos** o *datos fuera de intervalo*; es decir, valores de Y que al parecer no concuerdan con el modelo. Dado que casi todos los valores de Y deben estar a una distancia de 3σ o menos

[37] Los *cuantiles* son medidas que suelen utilizarse para resumir o descubrir propiedades de conjuntos de datos cuantitativos. Se aplican cuando los valores ordenados de una variable han de ser divididos en grupos homogéneos en cuanto al tamaño.

de $E(Y)$ (los valores medios de Y), se espera que en su mayoría estén a una distancia de $3S$ o menos de $\hat{Y}$. Si un residual es mayor en valor absoluto que $3S$, se considerará como dato fuera de intervalo.

Hay diversas doctrinas sobre la eliminación o conservación de datos atípicos, algunas coinciden en la opinión de mantenerlos a menos que exista prueba de que son equivocaciones y no son representativos de las observaciones de la población. Pero cuando representan un segmento de la población, recomiendan mantenerlos para asegurar su generalidad al conjunto de la población; si se eliminan se corre el riesgo de mejorar el análisis pero limitar su generalidad.

- **Multicolinealidad.**

Colinealidad es la asociación, medida como correlación, entre dos variables independientes. **Multicolinealidad** se refiere a la correlación entre tres o más variables independientes; esto es, se refiere a la relación lineal exacta o perfecta entre las variables independientes del modelo de regresión. Es muy común en la práctica utilizar estos términos indistintamente.

En notación matricial esto quiere decir que una columna de la matriz **X**, es una combinación lineal de otras columnas de la matriz, y el rango de **X** no podrá ser K (como máximo k -1), con lo que se viola la condición de rango dada en 2.56, quedando ahora

$$\rho(\mathbf{X}) < k$$

y tomando en cuenta las propiedades de las matrices

$$|\mathbf{X}'\mathbf{X}| = 0$$

entonces la matriz $\mathbf{X}'\mathbf{X}$ es no singular, no tiene inversa, y por lo tanto el sistema de ecuaciones normales de 2.64 no pueden resolverse y ser obtenidos los estimadores para β.

La situación habitual, por las propias características de las variables económicas, es que se generen situaciones de alta colinealidad entre las variables explicativas, sin llegar a ser *exacta*. En términos matriciales se dice que **X'X** es casi singular

$$|\mathbf{X'X}| \approx 0$$

este problema es conocido como de *multicolinealidad imperfecta*. En este caso, aunque los datos no son combinación lineal *exacta* de los otros, los valores de uno o más de ellos están dados *casi* todos por el mismo tipo de combinación lineal.

En la situación de multicolinealidad perfecta los coeficientes de regresión de las variables explicativas son indeterminados y sus errores estándar infinitos. Si la multicolinealidad es imperfecta, los coeficientes de regresión aunque son determinados, poseen errores estándar muy grandes (en relación con los propios coeficientes), lo que implica que no se pueden estimar con precisión dichos coeficientes. La multicolinealidad disminuye el poder predictivo de cualquier variable independiente individual en la medida en que está asociado con las otras variables independientes.

En cuanto a las propiedades de los estimadores, en el caso de multicolinealidad alta, siguen siendo insesgados y tienen la mínima varianza, es decir son eficientes; pero a pesar de ser eficientes esto no quiere decir que tengan una varianza muy pequeña en relación con el valor del estimador en una muestra dada. Lo anterior tiende a generar intervalos de confianza amplios sobre los parámetros, debido a la presencia de errores estándar muy grandes. Además que los errores estándar y los estimadores mínimo-cuadrático son muy sensibles ante cambios ligeros en los datos.

El problema de multicolinealidad es uno de los problemas más significativos y difíciles en la econometría, ya que se presentan en los datos provenientes de las series de tiempo. Por tanto es un problema esencialmente muestral (la muestra no es muy rica), no un problema de especificación. Cabe aclarar que si el objetivo primordial es la predicción, entonces la multicolinealidad no impide que se obtengan buenos pronósticos, siempre y cuando se mantenga en un futuro la colinealidad existente de la muestra dada. Por el contrario, si la colinealidad entre las variables explicativas de la muestra no se presenta para muestras futuras, la predicción será inciena. La tarea del investigador es,

por tanto, (1) valorar el grado de multicolinealidad y (2) determinar su impacto en los resultados y las soluciones pertinentes en caso de ser necesarios.

Existen varios procedimientos para detectar la presencia de multicolinealidad, algunos de ellos son los siguientes:

- El medio más simple y obvio de identificar la colinealidad es un examen de la *matriz de correlación* de las variables independientes (utilizando el coeficiente de correlación dado en 2.1), en el que un coeficiente alto generalmente mayor o igual que 0.9, indicará presencia de multicolinealidad. Cuando uno o más de estos coeficientes de correlación se desvían sustancialmente de cero, puede ser bastante difícil encontrar el subconjunto más eficaz de variables para incluirlo en la ecuación de predicción. Este procedimiento muestra solamente la correlación entre dos variables. La ausencia de valores elevados de correlación no asegura una falta de colinealidad; podría ocurrir que con coeficientes de correlación relativamente pequeños, existiera multicolinealidad entre el conjunto de variables explicativas del modelo o entre un subgrupo de ellas.
- Otra forma de evaluar su presencia, es cuando los coeficientes t resultan muy bajos y poco o ninguno tendrá un valor diferente de cero, al mismo tiempo el R^2 podrá ser elevado (entre 0.7 y 1) y probablemente la prueba F deberá ser rechazada al ser todos los coeficientes iguales a cero.
- Una alternativa, más costosa y elaborada, consiste en estimar diferentes ecuaciones alternativas, por supuesto basándose en la teoría económica.

Así como no existe una forma única para detectar el problema, tampoco hay una única solución para corregir la presencia de multicolinealidad. Algunas alternativas son:

- *Ampliar la información.* Ya que la multicolinealidad es un problema muestral, es posible que en otra muestra con las mismas variables, no sea tan seria. Esta solución es recomendada cuando el tamaño de muestra de aplicación es relativamente pequeño y se disponga de datos adicionales correspondientes para las variables. Una variante es recurrir a diferentes fuentes de información en la obtención de los datos estadísticos para cada variable; específicamente se refiere a combinar datos de corte transversal y series de tiempo (mezcla de datos). Pero surge el problema de interpretarlos, no obstante que ha tenido muchas aplicaciones

esta técnica en situaciones en que los estimadores de corte transversal no varían de una muestra a otra.

- *Cambiar de modelo.* Dentro de esta solución se encuentra la *eliminación* de las variables que generen la multicolinealidad, y la *transformación* de las variables en el modelo. El primer caso, es la solución más simple, eliminar una de las variables colineales; su inconveniente es que puede acarrear un problema grave que es el de cometer un error de especificación del modelo al omitir una variable importante que diga la teoría económica que deba de incluirse, a la vez que crea sesgos en todos los coeficientes estimados. Con la transformación de variables, lo que se busca es definir las variables explicativas en tal forma que la relación entre ellas disminuya; particularmente se utiliza la transformación por incrementos o primeras diferencias, que disminuye el efecto de altas correlaciones entre las variables. Sin embargo, a la vez que se puede reducir la correlación de las variables explicativas, también se puede reducir la relación de estas con la variable endógena.
- *Convivir con la multicolinealidad.* Es una de las soluciones adoptadas en econometría aplicada, útil cuando el fin primordial es la predicción. Pero si el propósito es el análisis estructural, particularmente aclarar las influencias aisladas de las variables independientes, entonces la multicolinealidad es un problema grave que hay que corregirse.

Una tendencia que existe en el análisis de regresión, es que cuando se obtiene un t no significativo para los coeficientes, se culpa a la multicolinealidad de la falta de significancia, pudiendo ser un problema de error de especificación. Por tanto, es conveniente revisar el modelo desde el punto de vista teórico antes de tratar de corregir la multicolinealidad.

C.5 PREDICCIÓN DEL MODELO.

Habitualmente el principal problema de cualquier modelo econométrico se centra en su uso con fines predictivos, es decir, estimar un valor de la variable dependiente para ciertos valores dados de las variables independientes. Existen dos tipos de predicción que pueden efectuarse: la predicción del valor medio o promedio de Y correspondiente a un determinado valor de X igual a X_0 (esto es, estimar $E(Y|X_0)$) y la predicción de un valor individual o particular de Y dado X_0. Estas predicciones se encuentran con nombres muy diferentes según el libro de texto que se consulte (en algunos ni siquiera se considera como predicción el cálculo del valor medio de Y), aquí se le llamará simplemente *predicción de la media de Y* y *predicción de un valor individual de Y.*

- **Predicción de la media de Y.**

Considérese el siguiente vector de valores

$$\mathbf{X}'_0 = (1, X_{20}, X_{30}, \ldots, X_{k0})^{38} \qquad 2.87.$$

para los cuales se desea predecir el valor medio de la variable explicada de Y denotado por la ecuación:

$$E(Y|X_0) = \mathbf{X}_0\beta$$

que por simplificación como en la ecuación 2.4, se puede escribir como

$$E(Y_0) = \mathbf{X}_0\beta \qquad 2.88.$$

que representa el valor esperado de Y_0 correspondiente a un valor $X=X_0$. Dado que β se distribuye normalmente con media y varianza tal como se indicó en la ecuación 2.75, se tiene que

$$\hat{Y}_0 = \mathbf{X}_0\hat{\beta} \qquad 2.89.$$

[38] Se incrementan las condiciones de las X's por el numero 1 para facilitar el uso de la notación matricial.

también tiene una distribución normal y es un estimador insesgado para la predicción *media* de Y_0, que es conocido como *predictor mínimo cuadrático* $\hat{Y}_0$, con un error de predicción dado por

$$e_0 = \hat{Y}_0 - E(Y_0) \qquad 2.90.$$

tal predictor es el único y el mejor predictor lineal insesgado de la variable dependiente; específicamente este predictor tiene la varianza mínima (para una demostración ver Johnston) cuya fórmula en notación matricial es simplemente una función de $\sigma^2(\mathbf{X'X})^{-1}$ y el vector $\mathbf{X_0'}$, es decir

$$\sigma^2_{\hat{Y}_0} = \mathbf{X_0'}\text{cov}(\hat{\beta})\mathbf{X_0} = \sigma^2\,\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0} \qquad 2.91.$$

reemplazando σ^2 por su estimador insesgado, se da lugar a la función conocida como *estimador de la varianza del predictor*

$$S^2_{\hat{Y}_0} = S_e^2\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0} \qquad 2.92.$$

Siguiendo un proceso idéntico al descrito para la estimación por intervalos de los parámetros, se tiene que la expresión

$$\frac{\hat{Y}_0 - E(Y_0)}{S_{\hat{Y}_0}} = \frac{\hat{Y}_0 - X_0\beta}{S_e\sqrt{\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0}}} \qquad 2.93.$$

sigue una distribución t con n-k-1 grados de libertad para un nivel de significancia de $\alpha/2$. El intervalo de confianza de 100(1 - α)% para el valor medio de $E(Y_0)$ será

$$\hat{Y}_0 - t_{\alpha/2}\,S_e\sqrt{\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0}} < E(Y|X_0) < \hat{Y}_0 + t_{\alpha/2}\,S_e\sqrt{\mathbf{X_0'}(\mathbf{X'X})^{-1}\mathbf{X_0}} \qquad 2.94.$$

donde $t_{\alpha/2}$ es un valor de la distribución t con n-k-1 grados de libertad. El término que multiplica a $t_{\alpha/2}$ es llamado *error estándar de la predicción* (algunos autores lo llaman *error estándar del pronóstico*), el cual proporciona un intervalo de predicción sobre la media de la variable dependiente a niveles dados de la variable independiente X_0. Usualmente este término aparece en la impresión de paquetes de computadora para regresión.

Cuando se obtienen los intervalos de confianza para cada uno de los valores de la muestra se encuentra lo que se conoce como *bandas de confianza* para la función de regresión poblacional. Gráficamente se ilustra en la figura 2.17 los intervalos de confianza o bandas de confianza para la media de Y. Como es de esperarse, las mejores predicciones se lograrán para valores de X que se muevan en torno de $\bar{X}$.

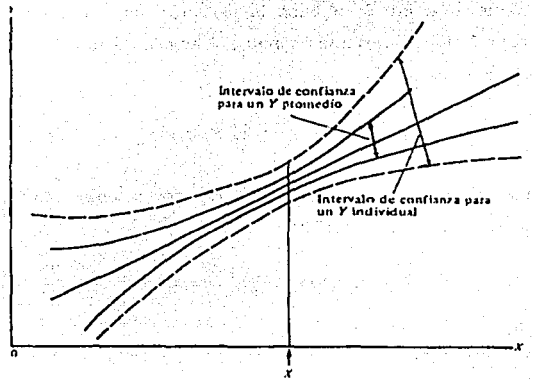


Figura 2.17 Intervalos para el valor medio y el valor individual.

- **Predicción de un valor particular de Y.**

Aun cuando resulta de interés considerar la respuesta media de Y para un valor dado de X, el interés principal es poder usar la ecuación de predicción basada en los datos observados para poder predecir una respuesta observada. Para obtener la predicción de un valor individual de Y, Y_0, también se puede emplear la ecuación 2.89 para estimar el valor individual de Y. El error de la predicción en este caso será el definido por:

$$e_0 = \hat{Y}_0 - Y_0 \qquad 2.95.$$

la cantidad ($\hat{Y}_0$ - Y_0) es igual a la diferencia entre $\hat{Y}_0$ y el valor esperado de Y_0 cuando $X=X_0$ más un error aleatorio u que representa la diferencia entre el valor esperado de Y y el valor real de Y. Dada la independencia entre Y_0 y $\hat{Y}_0$, supone una varianza del error

$$\sigma^2_{\hat{Y}_0-Y_0} = \sigma^2\left[1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right] \qquad 2.96.$$

sustituyendo por S^2, el estimador insesgado será

$$S^2_{\hat{Y}_0-Y_0} = S^2_e\left[1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0\right] \qquad 2.97.$$

De esta manera, el intervalo de predicción del 100(1-α)% para un solo valor pronosticado Y_0 se puede determinar por el estadístico

$$T = \frac{\hat{Y}_o - Y_o}{S_{\hat{Y}_o-Y_o}} = \frac{\hat{Y}_o - Y_o}{S_e\sqrt{1 + \mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0}} \qquad 2.98.$$

el cual tiene una distribución t con n-k-1, que conduce a una prueba de hipótesis de que el nuevo punto de datos (Y, X) es generado por la misma estructura. El intervalo de predicción del 100 (1- α)% para una respuesta individual Y_0 será

$$\hat{Y}_0 - t_{\alpha/2}S_e\sqrt{1+\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0} < Y_0 < \hat{Y}_o + t_{\alpha/2}S_e\sqrt{1+\mathbf{X}_0'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}_0} \qquad 2.99.$$

donde $t_{\alpha/2}$ es un valor de la distribución t con n-k-1 grados de libertad.

Obsérvese que el intervalo de confianza para un Y_0 individual, es más ancho que para el valor medio de Y_0, lo cual se ilustra nuevamente en la figura 2.18. En general una de las características importantes que muestran en las bandas de confianza es que la amplitud es menor cuando $X_0=\bar{X}$. Esto reafirma el cuidado que se debe de tener al extrapolar la línea de regresión histórica cuando se tratar de predecir Y_0, o un $\hat{Y}_0$ asociado a un X_0 dado, que esté más o menos lejos de la media muestral $\bar{X}$.

C.6 VARIABLES OMITIDAS Y VARIABLES IRRELEVANTES.

En la práctica, aunque se especifica que Y depende de $X_1, X_2, \ldots, X_k$, no todos los coeficientes de estas variables pueden estimarse con una precisión razonable. En algunos casos se sabe cuál es el modelo correcto, pero no se puede implementar debido a que no se encuentra la información necesaria disponible. Otras veces se sabe cuáles variables se debe incluir en un modelo, pero no se sabe la forma funcional exacta en que las variables deben aparecer en el modelo. Finalmente, se presenta el caso de no saber realmente cuál es el modelo correcto. Todas estas razones llevan a cometer un *error de especificación* del modelo.

Existen algunas guías generales para juzgar la calidad de un modelo, como se ha presentado en las secciones anteriores. También hay que tener presente que un buen modelo no debe presentar ni demasiadas variables, ni debe olvidar las que sean verdaderamente relevantes; es decir, debe cumplir el principio de *parsimonia* según el cual un fenómeno debe ser descrito con el mínimo número de elementos posibles. En esta sección se presentan dos pruebas para saber si se ha cometido el error de omitir una variable relevante, o en su defecto, incluir una variable innecesaria.

- **Variables omitidas.**

Cuando se omite una variable propia del modelo, las consecuencias son muy serias, los estimadores de los coeficientes de la variable retenida en el modelo son sesgados e inconsistentes. Además, las varianzas y los errores estándar son estimados en forma incorrecta dañando así los procedimientos usuales de prueba de hipótesis.

Existe una prueba conocida como *variables omitidas* (omitted variables) para detectar este problema, la cual permite agregar un conjunto de variables a una ecuación existente y preguntar si el conjunto aporta alguna contribución significativa a la explicación de la variable dependiente. En este caso la hipótesis nula establece que el conjunto de variables no son significativas, es decir, los coeficientes son igual a cero. Bajo la hipótesis de que los errores son independientes y normalmente distribuidos, se

utiliza el estadístico F, con número de grados de libertad para el numerador igual al número de variables que se desea incluir en el conjunto.

Por ejemplo, para investigar simultáneamente la importancia de que se deba incluir X_4 y X_5 en un modelo con tres variables el modelo, se prueba la hipótesis:

$$H_0: \ \beta_4 = \beta_5 = 0$$
$$H_1: \ \beta_4 \text{ y } \beta_5 \text{ no son ambas cero,}$$

y se calcula[39]

$$f = \frac{[R(\beta_4, \beta_5 | \beta_1, \beta_2, \beta_3)]/2}{S^2}$$

que prueba lo favorable de X_4 y X_5 en el modelo. El valor de S^2 es el cuadrado medio del error para el modelo que contiene las tres variables. El número de grados de libertad asociados con el numerador, para este caso, será igual a dos. Si $f < f\alpha\ (2,\ n-K-1)$ para un nivel de significancia preseleccionado, X_4 y X_5 no se incluyen en la ecuación de regresión.

- **Variables irrelevantes.**

Las consecuencias de incluir variables irrelevantes o redundantes en el modelo, son menos serias. Los estimadores de los coeficientes de las variables relevantes como también los de las variables irrelevantes, continúan siendo insesgados y consistentes, la varianza del error se sigue estimando de forma correcta. El único problema es que las varianzas estimadas tienden a ser más grandes, haciendo menos precisa la estimación de los parámetros y los intervalos de confianza tienden a ser más grandes.

Para detectar la presencia de variables irrelevantes, se utiliza la prueba denominada *variables redundantes* (Redundant variables) que muestra la significancia estadística de un subconjunto de variables que se han incluido; es decir, se prueba si tienen coeficientes cero, y de esta forma pueden ser eliminadas pues no reportan

[39] La cantidad de variación de la respuesta que se atribuye a X_4 y X_5 en presencia de las variables $X_1, X_2, \ldots, X_k$ se puede escribir como $R(\beta_4, \beta_5 | \beta_1, \beta_2, \beta_3)$.

información al modelo. La prueba del estadístico F en este caso tiene grados de libertad en el numerador igual al número de variables que se piensa deben ser eliminadas. Al igual que en la prueba anterior, no es necesario calcular el valor del F, cuando se reporta el valor de probabilidad p.

Recuérdese que se puede utilizar el valor p para probar la significancia de la o las variables que se desean incluir o eliminar en un modelo, aceptándose si es menor a un nivel de significancia preseleccionado.

En general, hay que tener presente que en muchos problemas en que se quiere establecer las variables que se deben incluir en el modelo final y, cuando el número de variables es pequeño, pueden utilizarse las pruebas t individuales. En este caso, se puede aplicar la relación básica entre la distribución t con v grados de libertad y la distribución F con 1 y v grados de libertad dada por la ecuación 2.82.

CAPITULO III

FORMULACIÓN Y CONSTRUCCIÓN DEL MODELO PARA LA EXPLICACIÓN DE LA DEMANDA DEL CONSUMO DE LECHE.

A. ESPECIFICACIÓN DEL MODELO DE DEMANDA.

A.1 DEFINICIÓN DE VARIABLES Y ACOPIO DE INFORMACIÓN.

Recordando lo visto en el capítulo uno, especificación se refiere al hecho de expresar una teoría económica en términos matemáticos. En esta sección se efectúa la definición de variables de acuerdo a la información obtenida en las entrevistas con los especialistas del área; a la vez se contrastarán estas variables con la teoría económica básica para calcular la demanda de un bien, y se complementará con la información referente a la situación actual del consumo nacional de leche. Estas tres fuentes de información fueron expuestas en el primer capítulo.

De acuerdo a la información obtenida en las entrevistas con los expertos de área, se concluye que los factores que influyen en la demanda del consumo de la demanda de la leche son:

- Precio de la leche.
- Ingreso de los consumidores.
- Precio de bienes sustitutos.
- Cambios en los gustos.

- Tamaño de la población.
- Inflación.
- Nivel educacional.
- Publicidad.
- Clima.

Como se señaló en el capítulo 1, la teoría económica elemental del consumidor individual dice que el **precio del producto** es el principal indicador para saber el tipo de demanda que se está estudiando. Habrá que observar la evolución de esta variable ya que como se mencionó, el precio de la leche fue liberado a partir de 1997 por lo que pudiera ser no significativa. El **ingreso del consumidor** es otra variable que incluye la teoría económica; para este modelo se espera que sea una variable representativa, pues las empresas tienen orientada su demanda para segmentos específicos, tanto para los estratos de medianos a altos ingresos (por la parte comercial), como para las personas de bajos recursos económicos (por la parte subsidiada). De esta manera, el precio del producto y el ingreso del consumidor, son las variables base ya que se incluyen en casi todos los modelos econométricos que se han elaborado para calcular el consumo de alimentos (ver, por ejemplo, Intrilligator 1990). Sobre estas variables se utilizará el concepto de "elasticidad" (ver capítulo 1) como ayuda para explicar la demanda de la leche.

El **precio de los bienes sustitutivos**, como se recordará, se determina cuando al aumentar el precio de un bien aumenta la demanda del otro. En el caso particular de la leche, difícilmente una persona que acostumbra tomarla, si sube su precio preferirá tomar refresco o café. Lo que sí puede suceder es que con el tiempo se deje de consumir la misma cantidad por diversos motivos, por ejemplo, la edad o los hábitos alimenticios, entre otros. Una sustitución que podría ser significativa sería respecto a la misma leche y las fórmulas lácteas que han disminuido el consumo de leche; sin embargo, los porcentajes en sus ventas varían todavía mucho como para tomarse en consideración, pero en un futuro si podría ser una variable de peso. El **precio de los bienes complementarios** se forma cuando dos productos se consumen conjuntamente. Hoy en día existen las bebidas *de leche saborizadas* para quienes prefieren agregar a su leche un sabor artificial, quedando el café y los cereales como posibles productos complementarios pero, como se sabe, no son complementarios para todas las personas,

sólo para un segmento. Por tanto no se toman en cuenta estos dos factores que señala la teoría para calcular la demanda de la leche.[1]

En los **cambios de gustos o preferencias** entrarían los jugos y los yogures que han provocado una disminución en el consumo de la leche; pero en México no existen estadísticas sobre estas variables que son conocidas como *órdenes de preferencia*. De esta manera, no se puede incluir dicha variable.

El **tamaño de la población** se incluye pues ya sea con fines gubernamentales o comerciales se piensa en producir más para cubrir la demanda del consumo de la leche, desde los niños hasta las personas de la tercera edad; es decir, debe existir un consumo mínimo por habitante. De esta forma su introducción es importante como variable significativa para cubrir la demanda nacional de leche.

La **inflación** en un país como el nuestro, es determinante incluirla, por lo cual se hará uso de variables deflactadas que como se vio anteriormente, absorben el efecto de la inflación en las variables.

El **nivel educacional** es una variable cualitativa que se expuso en el sentido de que la gente hoy en día busca una alimentación más sana, como se manifiesta por la tendencia creciente que hay por los alimentos *light*, por citar un ejemplo. Un comentario del director de la Cámara Nacional de Industriales Lácteos (CANILEC), es que la parte del mercado que más consume leche es la clase media pues una persona que vive en una zona residencial, por mantenerse en forma consume poca leche; lo que no ocurre en clases más bajas que no cuidan tanto que si nutre y que no, sino que se come lo que les gusta, es decir mas "antojitos". Por otra parte, esta variable contempla también que no se conocen las diferencias entre los diversos productos de leche, como sería el caso de una fórmula láctea (extensor) o una leche deslactosada, por citar origen o proceso de leche. Por tanto es una variable a considerar en la demanda de una empresa en particular ante la necesidad de orientar sus productos hacia un segmento específico del mercado,

[1] Para saber si un producto es sustitutivo o complementario, se aplica el concepto de "elasticidad cruzada", el cual no se vio por no aplicarse. Para mayor información se puede encontrar en cualquier libro de economía o administración que incluya los factores de la demanda.

pero no es significativo para tomarse en cuenta en la demanda total del mercado del sector lechero. Por tanto, no se incluye esta variable.

El **clima** es otra variable cualitativa que se menciona en las entrevistas y que es importante considerarla en la demanda de una empresa pensando que hay *temporadas* de mayor venta de leche a lo largo del año. Sin embargo en la demanda del mercado de leche, está la contraparte de leche subsidiada con un porcentaje de consumo fijo a lo largo del año. Basándose en esta información, tampoco se considera esta variable.

Lo mismo ocurre con la **publicidad** no se toma en cuenta para la demanda del mercado del sector lechero, ya que por ejemplo, la leche bronca se consume cerca de un 30%, sin ninguna publicidad. Por otro lado, en la empresa A se dijo que hay lugares en provincia que son muy regionalistas pues solamente consumen la leche propia del lugar. Por tanto, es una variable muy importante que se debe incluir para cubrir un segmento específico del mercado como consecuencia de la diversificación de sus productos y para darlos a conocer; de esta manera, es un factor a considerar para la demanda de una empresa en particular solamente, pero que no se incluirá en este modelo.

De esta forma el modelo que se propone para calcular la demanda del consumo de leche en el ámbito nacional, incluye las siguientes variables explicativas:

- Precio de la leche
- Ingreso de los consumidores
- Población

- **Acopio de información.**

La muestra que se considera es a partir del año 1990 hasta el año 2000 por las razones citadas posteriormente. Aunque la mayoría de observaciones se pueden obtener de forma mensual (y con ello tener mas datos), son manejadas en un periodo de tiempo anual considerando que los cambios económicos en este sector no se dan en una forma tan rápida. De esta forma se obtiene un tamaño muestral de 11 datos.

La obtención de las observaciones para las variables seleccionadas se obtuvieron con algunos tropiezos, como se comenta a continuación.

El Consumo Nacional Aparente (CNA) es el indicador económico que se utiliza a nivel macroeconómico para obtener la cantidad demandada de leche. Aunque es recomendable utilizar otros datos que reflejen mejor los cambios en el consumo real, es la única información que se puede conseguir y la proporciona la *Secretaría de Agricultura y de Ganadería* (SAGAR)[2].

El CNA, como se explicó anteriormente, se forma actualmente con la suma de la producción más las importaciones menos las exportaciones. Los valores de la producción están reportados en miles de litros y no se encuentran en forma continua para recabar los datos de la muestra del año de 1990 al año 2000. Las observaciones de 1990 a 1998 de la producción de leche las obtuve del documento especial que editó la SAGAR[3]; las observaciones partir del año 1998 hasta noviembre del año 2000 las obtuve en forma mensual de los *boletines de la leche* que edita la SAGAR[4], y para el mes de diciembre apliqué la tasa de crecimiento (vista en el capítulo uno) de forma mensual. Cabe aclarar que se presentan algunos meses de mayor productividad de leche, los cuales se repiten a través de los años y que influirían en el consumo aparente si se eligiera un periodo de tiempo más corto.

Las importaciones son reportadas toneladas. Para la recopilación se utiliza en primer lugar, la denominación de las fracciones arancelarias de importación apropiadas a un tipo específico de leche (fluida, en polvo, evaporada, etc.); también obtenidas del documento especial de la SAGAR y complementadas con las que incluye el *"Boletín de la leche"*, las cuales se resumen en la tabla A.1 del anexo A. Además se utilizan unos factores de conversión para convertir a litros los productos de acuerdo al tipo específico de leche. Los factores que se necesitan en este trabajo se muestran en la tabla A.2, también del anexo A.

[2] Como se mencionó en el capítulo uno, se transformó en SAGARPA, ver nota 24 para el significado de las siglas. En este trabajo se nombra como SAGAR, tomando en cuenta que la mayoría de información fue publicada en el periodo de esta Secretaría.

[3] Ver bibliografía.

[4] Estos boletines han sufrido algunos cambios desde su primera edición en el 94, en cuanto al contenido, en cuanto a las dependencias que lo elaboran, y en cuanto a su publicación que se hacía mensualmente; a partir de noviembre del 99 se publica bimestralmente. Ver bibliografía.

A diferencia de las importaciones que son reportadas en toneladas, la SAGAR reporta las exportaciones en kilogramos, por lo que se ve el poco peso que tendrán en el Consumo Nacional Aparente. Se omiten las exportaciones de leche condensada y evaporada ya que prácticamente es una empresa la que realiza estas exportaciones; por otra parte, se tiene el inconveniente de que a partir del año 1996 hasta el año 1998 se dejaron de reportar los datos de las exportaciones en leche condensada y evaporada; además, eran tomadas en cuenta en los años anteriores dentro de la misma fracción arancelaria. Fue hasta el año de 1999 que se introducen más especificaciones, por separado, para las exportaciones en forma mensual, aunque anualmente si existen series de datos para todas las especificaciones que contiene el documento especial de la SAGAR. Por otra parte, como se recordará, somos deficitarios en la producción de leche y no tiene sentido tomar en cuenta las exportaciones de leche condensada y evaporada. Las fracciones arancelarias de exportación para los productos que se incluyen, se dan en la tabla A.3 del anexo A. Como se observa en el anexo, hasta el año de 1999 la única fracción arancelaria que se tenía para leche (aparte de la leche condensada y evaporada que estaban integradas en una sola), es la 0402 2999, lo que confirma el poco peso que representan en las exportaciones. Los factores de conversión a litros por grupo de productos, son mostrados en la tabla A.4, del anexo A.

La obtención del precio de la leche fue un gran obstáculo que se debía resolver por la importancia que puede tener en la demanda de un producto. En el boletín que publica la Secretaría de Ganadería de la SAGAR, se incluyen los precios de algunos productos de la leche basados en información de PROFECO, pero esta información solamente se reporta para algunos años y únicamente de precios en el área metropolitana. En PROFECO, debido a que obtienen el precio para cerca de 1000 artículos, estos no se mantienen vigentes en las computadoras por más de una semana; por otra parte, no cuenta con información histórica al público ya que su biblioteca no está en uso desde 1995; además, me aclararon que PROFECO verifica los precios solamente en el D.F. En SAGAR me aconsejaron buscar en SNIM (Sistema Nacional de Mercados), hablé pero me explicaron que esta información la pasaron a la SECOFI a partir de 1996, año en que se liberó el precio. En SECOFI (después de decirme que tampoco tienen esa información) me recomendaron ir al Banco de México en donde finalmente obtuve las observaciones.

El Banco de México maneja los precios de la leche en forma de números índices, de manera agregada y, de varios tipos y procesos. Los precios se pueden consultar en los indicadores económicos en la parte del *Índice Nacional de Precios al Consumidor* (INPC) como más representativos a nivel nacional, ya que se basan en el gasto en lácteos proporcionado por la actual ENIGH que aplica el INEGI[5]. Además, el *Diario Oficial de la Federación* publica mensualmente toda la serie de precios de las distintas marcas y presentaciones, que actualmente abarca 46 localidades en donde se realiza la muestra, arrojando gran cantidad de datos mensualmente; por ello, fue necesario hacer una solicitud para obtener solamente los índices de precios de la leche de toda la República en forma más compacta. Solo para los meses de octubre y noviembre se utilizó la tasa de crecimiento de agosto y septiembre para completar las observaciones con las otras series.

Hay que considerar que el Banco de México tuvo un cambio de base que hasta 1994 era base (1978=100), este cambio aparece a partir de noviembre de 1995 en el Diario Oficial. Para convertir el índice de precios de la base anterior a la nueva base (1994=100), se divide el correspondiente índice mensual entre la constante C=37394.134 y el resultado se multiplica por 100, lo cual se indica en el mismo Diario. Finalmente para ver la evolución real de la variable y que no influya el tiempo sobre ella, se hace la deflacción correspondiente, dividiendo los precios corrientes (base 1994=100) entre el INPC general (en la misma base), con lo cual se obtienen los precios constantes. Todos estos conceptos se explicaron en el capítulo 1.

El ingreso de las personas también es obtenido en el Banco de México a través del INPC de los indicadores económicos anuales, en la especificación por *estrato de ingreso*. Se examina solamente los índices entre uno y tres salarios mínimos, que corresponden al porcentaje de la población de mayor consumo de leche según las entrevistas realizadas, lo cual coincide con la información del INEGI[6], en la que se muestra que la mayor parte de la población recibe estos niveles de salarios. Con la conversión de base correspondientes a (1994=100), esta variable se deflacta de igual forma que el índice de precios de la leche con el INPC general, con la finalidad de conocer su evolución libre de la inflación a lo

[5] Consultar la nota 19 del capítulo uno para ver las siglas de la ENIGH. Esta encuesta la elaboraba la Secretaría de programación y Presupuesto (SPP), en la cual se basaron para los índices de precios base 1978=100. Con el tiempo ha ido aumentando el tamaño de la muestra cubriendo más localidades; como es de suponer utiliza datos de corte transversal.

[6] En el anuario estadístico de 1998 en la sección *Encuesta Nacional de Empleo* (ENE), y ENIGH 98.

largo del tiempo (precios constantes). Como se aprecia en la tabla 3.1, estos índices están por debajo de los precios de la leche lo cual hace suponer que tal vez no sea tan representativa como se desearía.

La mayor parte de los datos de la población son obtenidos del documento especial de la SAGAR los cuales coinciden con los de CONAPO (Consejo Nacional de la Población) y son complementados con los que incluye la ENE (ver nota 6) y los censos que elabora el INEGI. Estos son los únicos valores que no se encuentran en forma mensual, probablemente por el esfuerzo de recolección que representa y por que los cambios que repercuten no se dan en intervalos de tiempo tan cortos. Por tanto, se considera la misma cantidad anual para cada trimestre expresada en miles de personas para hacer más manejables las cifras.

La tabla 3.1 muestra las observaciones que se obtuvieron para cada variable del modelo en un periodo de tiempo anual, asi como las unidades en que se manejan.

AÑO	*CONSUMO DE LA LECHE (miles de litros)*	*PRECIO REAL DE LA LECHE (base 1994=100)*	*INGRESO REAL (base 1994=100)*	*POBLACIÓN (miles de personas)*
1990	8,996,194	122.33	100.07	81,249
1991	7,298,407	116.96	101.49	83,120
1992	8,992,319	111.10	100.07	85,050
1993	9,518,527	104.28	99.71	87,030
1994	8,818,320	100.00	100.00	89,066
1995	8,626,249	106.96	101.65	91,158
1996	9,013,277	110.11	102.92	93,181
1997	9,410,764	110.91	102.78	94,732
1998	9,638,158	111.74	102.55	96,254
1999	9,964,357	114.59	102.76	97,585
2000	10,539,434	111.12	102.40	97,361

Tabla 3.1 Datos para el análisis de la demanda del consumo de la leche.

Como se aprecia en el acopio de información, aunque se hubiera querido tener una muestra mayor, el *"Boletín de la Leche"* se creó recientemente en el 94, y la información no es tan actual pues se publica normalmente con un bimestre de retraso; además no se tiene información antes del año de 1990, según me informaron a causa del temblor del 85 se perdió mucha información; por esta razón, el inconveniente para la obtención de una muestra mayor fue debido al consumo. Y aunque también se pensó en una muestra de periodicidad más corta como trimestral o mensual, se sabe que los cambios económicos en la industria lechera (como en todo el sector agropecuario), no se producen de manera tan rápida; además de que la población se reporta solamente en forma anual; por último, la mayor productividad de leche en algunos meses afectaría al consumo aparente. Por tanto, se encontró una limitación en cuanto a la periodicidad y en cuanto al tamaño muestral.

- **Forma funcional de la ecuación.**

Junto a la consideración de las variables que explican el modelo de la demanda, otro aspecto en la especificación en la metodología econométrica, es la referente al tipo de función que relaciona las variables.

En al capítulo 2 se presentaron las transformaciones más usuales en Economía en caso de que el modelo lineal no se ajuste a la distribución de los datos. La función lineal es la que, desde el punto de vista teórico, menos refleja las características que definen la demanda de un bien, pero es aceptable para un grupo reducido de observaciones; por tanto, esta será la primera aproximación que se tomará en cuenta, por ser la ecuación más sencilla de trabajar. El resto de las ecuaciones suelen explicar más adecuadamente las variaciones la demanda, adaptándose mejor unas que otras en cada caso particular, según el tipo de bien que se esté tratando de analizar (de lujo, duradero, etc.), o según la estructura especial de los datos. En particular las formas logarítmicas o semilogarítmicas, han sido las que mejor se ajustan en los análisis de la demanda que aparece en los trabajos econométricos.

De esta manera, las funciones que se probarán para explicar la demanda del consumo de la leche serán: la lineal, la logarítmica y las semilogarítmicas, tanto en la variable dependiente como en las variables independientes.

B. ESTIMACIÓN DEL MODELO.

B.1 ESTIMACIÓN DE LOS PARÁMETROS DEL MODELO DE REGRESIÓN.

Para el proceso de estimación de acuerdo con la información de la sección anterior el modelo queda integrado de la forma siguiente:

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

donde:

Y_t = Consumo de leche en miles de litros en el año t

X_{2t} = Precio de la leche a precios constantes (base 1994=100) en el año t

X_{3t} = Ingreso de los consumidores a precios constantes (base 1994=100) en el año t

X_{4t} = Población en miles de personas en el año t

β_1 = Parámetro del intercepto que mide el valor esperado del consumo de leche cuando las variables del precio, el ingreso y la población son igual a cero.

β_2 = Parámetro que mide el cambio en el valor promedio que se produce en el consumo de leche, por cambio de unidad en el precio, manteniendo constantes las demás variables

β_3 = Parámetro que mide el cambio en el valor promedio del consumo de leche, por unidad de cambio en el ingreso, permaneciendo constantes las demás variables

β_4 = Parámetro que mide la variación en el valor promedio del consumo de leche, por unidad de cambio en la población, manteniendo constantes las demás variables

u_t = Término del error aleatorio en el año t

Según las clasificaciones dadas en el capítulo 1, las variables mencionadas tienen las siguientes características: la variable consumo es una variable endógena, la variable precio, ingreso y población son todas variables predeterminadas exógenas. La ecuación

empleada se conoce como ecuación de comportamiento de los consumidores. El modelo es un modelo de regresión lineal múltiple uniecuacional.

- **Expectativas sobre los signos de los parámetros del modelo.**

El signo para el coeficiente del precio de la leche se espera que sea negativo de acuerdo con teoría, la cual indica que al disminuir el precio aumenta la demanda de un bien. Hay que considerar por la deflacción de esta variable puede tener otro comportamiento al esperado.

En el signo del coeficiente del ingreso existe la posibilidad de que sea positivo siempre y cuando el nivel de ingresos aumente proporcionalmente al consumo de la leche (de acuerdo a las observaciones se aprecia que el nivel de ingresos ha aumentado muy poco); de ser así habría un desplazamiento de la curva de la demanda hacia la derecha. El problema que puede surgir es que el aumento en el índice de precios es menor para los salarios que para el precio de leche, por lo que pudiera no ser tan significativa esta variable.

Finalmente el signo correspondiente al coeficiente de la población también se espera que sea positivo, considerando que ante un aumento de la población, corresponda un aumento en el consumo de leche.

- **Estimación.**

Considerando el tamaño muestral y la sencillez del modelo para explicar la demanda de la leche, se utilizará el método de Mínimos Cuadrados Ordinarios para estimar los parámetros del modelo de regresión, ajustando las cuatro ecuaciones propuestas anteriormente (ecuación lineal, semilogarítmicas y logarítmica) aplicables en estudios de demanda. Para realizar la estimación se emplea un programa utilizado por los economistas conocido como Econometric Views (EVIEWS).[7]

[7] En el anexo A se muestran las herramientas empleadas en este trabajo con EVIEWS.

Los valores de los parámetros de regresión para cada modelo propuesto se muestran en la tabla 3.2. Estos valores son tomados del anexo C sección C.1 con los resultados de la estimación de las ecuaciones. Por simplificación se utilizan las designaciones dadas en el capítulo 2 (sección B.4), llamando modelo lin-log cuando hay una transformación logarítmica en las variables independientes, modelo log-lin cuando la transformación logarítmica es únicamente en la variable dependiente y modelo log-log al aplicar la transformación logarítmica a todas las variables en la ecuación.

TIPO DE ECUACIÓN	*COEFICIENTES*			
	β_{1t}	β_{2t}	β_{3t}	β_{4t}
Lineal	64,076,536	86,574.37	-881,536.3	275.6195
Lin-log	100,828,174	10,661,131	-93,369,416	25,369,543
Log-lin	22.38654	0.009495	-0.100806	.0000311
Log-log	27.11842	1.173794	-10.68025	2.867661

Tabla 3.2. Coeficientes estimados de las ecuaciones propuestas.

TIPO DE ECUACIÓN	*ECUACIÓN ESTIMADA*
Lineal	$Y_t = 64,076,536 + 86,574.37\ X_{2t} - 881,536.3\ X_{3t} + 275.6195\ X_{4t}$
Lin-log	$Y_t = 100,828,174. + 10,661,131\ \ln X_{2t} - 93,369,416\ \ln X_{3t} + 25,369,543\ \ln X_{4t}$
Log-lin	$\ln Y_t = 22.38654 + 0.009495\ X_{2t} - 0.100806\ X_{3t} + .0000311\ X_{4t}$
Log-log	$\ln Y_t = 27.11842 + 1.173794\ \ln X_{2t} - 10.68025\ \ln X_{3t} + 2.867661\ \ln X_{4t}$

Tabla 3.3 Sustitución de coeficientes en las ecuaciones propuestas.

La interpretación de los coeficientes se realiza al elegir la ecuación que mejor se ajuste a los datos de acuerdo a los contrastes utilizados en las secciones siguientes, esto por tener interpretación diferente los coeficientes de cada ecuación estimada. En la tabla 3.3 se presenta la sustitución de los parámetros estimados para cada ecuación propuesta.

C. VALIDACIÓN DEL MODELO.

El proceso para seleccionar la ecuación que mejor explique la demanda del consumo de leche dentro de las cuatro ecuaciones propuestas, se hará basándose en la verificación de los supuestos básicos del modelo de regresión lineal así como en los contrastes estadísticos presentados en la sección C.3 del capítulo anterior.

C.1 VERIFICACIÓN DE SUPUESTOS.

Como se recordará es importante verificar el cumplimiento de los supuestos como garantía de que los estimadores sean MELI, y para poder llevar cabo con efectividad las pruebas estadísticas. Para ello, se emplean en esta sección los procedimientos descritos en el capítulo 2 para muestras pequeñas, debido al tamaño muestral utilizado de 11 observaciones para cada variable.

En la regresión múltiple la linealidad del modelo así como la homocedasticidad se pueden comprobar en la misma gráfica de residuales contra el valor estimado de Y. Las gráficas para cada ecuación se muestran en el anexo C sección C.3, con los datos de las tablas de la sección C.2.

Al observar las gráficas para el modelo lineal, se aprecia que no hay un patrón de los residuales (como el presentado en la gráfica b de la figura 2.17, del capítulo dos) que indique la necesidad de emplear una transformación, sin embargo, las gráficas de las ecuaciones propuestas presentan un patrón de comportamiento similar al que muestra el modelo lineal. Por tanto, se deberá de evaluar con los criterios siguientes qué modelo será el mejor, si el modelo lineal o con alguna transformación de las variables.

Lo mismo ocurre al examinar la igualdad de varianzas, comparando las gráficas de los residuales con las que incumplen este supuesto (figuras 2.17 c, d y e del capítulo 2), se puede observar que no presentan una tendencia de la varianza de los residuales que indique heterocedasticidad. El problema de heterocedasticidad se presenta muchas

veces cuando se utilizan variables reales, en este caso puede ser que el uso de variables deflactadas haya eliminado este problema.

El incumplimiento de la autocorrelación se detecta con las gráficas de los residuos contra el tiempo, las cuales se pueden consultar en el anexo C sección C.4. Como se observa en las gráficas no hay un patrón de comportamiento que revele correlación positiva o negativa de los residuos a lo largo del tiempo. Lo que sí se notan, son datos fuera de intervalo para el año 96 en los modelos lineal y lin-log y en los años 91 y 96 para los otros modelos, sin embargo se examinará qué tanto afectan o no la distribución de los residuales con las gráficas para detectar la normalidad.

Para verificar el supuesto de normalidad de los residuos, se empleó la gráfica de cuantiles-cuantiles normales. Las gráficas de los residuos para cada una de las ecuaciones se muestran en el anexo C sección C.5. Como se recordará, a medida que los datos se acercan a la línea recta, indica que los datos se distribuyen normalmente. Observando las gráficas se aprecia que los residuales tienden a presentar un comportamiento parecido a la distribución normal, lo cual garantiza la significancia estadística de los contrastes utilizados posteriormente. Si bien, las ecuaciones presentan una diferencia casi imperceptible, hay una mayor preferencia para los modelos lin-log y log-lin, a pesar de los datos atípicos que se detectaron en las gráficas de los residuales a lo largo del tiempo. En realidad, en los modelos log-lin y log-log ya no aparecen los datos atípicos que se detectaron en los años 1991 y 1996; y en las ecuaciones lineal y lin-log sigue apareciendo un dato atípico pero en el año de 1990 (antes en el año de 1996). En este caso son más representativos los datos atípicos de las gráficas para detectar autocorrelación, los cuales se decide mantener ya que reflejan dos años en que el consumo de leche disminuyó de manera más sobresaliente que concuerda con la realidad.

Aunque para el modelo de regresión múltiple se recomienda probar la normalidad con los valores estimados, también se hace la prueba para cada variable, no siendo tan rigurosos los resultados obtenidos. No obstante que el resultado es el mismo para algunas variables, se presentan los resultados de cada variable por modelo en las gráficas C.5.6-8 del mismo anexo. En orden de severidad las variables que reflejan menos normalidad son: la variable que describe el ingreso, después la variable del

consumo, enseguida la variable precio y finalmente la variable que describe la población es la que tiene una distribución mas cercana a la distribución normal. Hay dos datos atípicos en el caso del ingreso (correspondientes a los años de 1993 y 2000, siendo más representativo el del 93 en que hubo una disminución del ingreso), uno para la población (que corresponde al año 2000 y también refleja una disminución en la población que no se tuvo en los otros años) y uno para el consumo de leche (en el año de 1990, que antecede a la disminución más grande de consumo en 1991), sin embargo, estos datos no disminuyen tampoco la normalidad de las variables por separado.

La independencia lineal se verifica con la matriz de correlación para las variables independientes de cada modelo; en este caso las matrices correspondientes se incluyen en el anexo C sección C.6. Al examinar los coeficientes de correlación muestral, se aprecia que las variables no reportan correlación entre ellas excepto la variable que describe la población con la variable del ingreso, con una correlación entre .79 (en los modelos lin-log y log-log) y .80 (en los modelos lineal y log-lin). Sin embargo, esta cantidad es aceptable para poder hacer el análisis apropiado de las variables. Más adelante también se comprobará con los resultados de la prueba *F*.

En la tabla 3.4. se expone un resumen con los resultados de la verificación de los supuestos para cada modelo. Como se aprecia, todos cumplen los supuestos básicos del modelo de regresión lineal múltiple.

ECUACIÓN	Linealidad	Homocedas-ticidad	No correlación	Normalidad	No colinealidad
Lineal	si	si	si	si	si
Lin-log	si	si	si	si	si
Log-lin	si	si	si	si	si
Log-log	si	si	si	si	si

Tabla 3.4. Resultados de la verificación de los supuestos del modelo de regresión lineal múltiple.

C.2 CONTRASTES DEL MODELO.

Una vez que se ha examinado el cumplimiento de los supuestos básicos, se procede a la comprobación de la significancia estadística de las variables incluidas en los modelos. Para ello se emplean nuevamente los reportes de la regresión principal del anexo C sección C.1, y resumidos en la tabla 3.5.

ECUACIÓN	R^2	R^2 ajustado	Valor ρ de prueba F	Valor ρ de prueba t		
				X_{2i}	X_{3i}	X_{4i}
Lineal	.8715	.8164	.0017	.0125	.0026	.0004
Lin-log	.8787	.8267	.0014	.0079	.0020	.0003
Log-lin	.8581	.7973	.0024	.0181	.0032	.0005
Log-log	.8666	.8094	.0019	.0115	.0025	.0004

Tabla 3.5 Resultados de significación estadística.

En primer lugar se analizan los resultados para los valores del coeficiente de determinación como medida para explicar el porcentaje explicado por las variables independientes en el modelo. Los resultados para todas las ecuaciones estimadas son buenos. Hay que recordar que no es comparable el coeficiente de determinación cuando la variable dependiente no es la misma en los modelos. De esta forma se compara el R^2 del modelo lineal junto con el del modelo lin-log observándose un ajuste ligeramente superior en el modelo lin-log con un porcentaje de explicación del consumo de leche por el precio, el ingreso y la población con un 87.87%. De los modelos que tienen una transformación logarítmica en la variable dependiente, se tiene un mejor ajuste en el modelo log-log con un porcentaje del 86.66% para explicar el consumo de la demanda de leche con las variables del precio, ingreso y población.

El coeficiente de determinación ajustado es particularmente útil al aumentar o disminuir variables independientes o para diferentes tamaños muestrales, por ajustarse al número de grados de libertad. A pesar de que en las ecuaciones estimadas no se han realizado cambios en el tamaño de la muestra o en las variables, se incluye para compararlo con el coeficiente de determinación estándar que en algunas ocasiones suele ser mucho mayor que el ajustado. En este caso, las conclusiones del R^2 ajustado son

también más favorables para el modelo lin-log y modelo log-log con un porcentaje de explicación del consumo de leche del 82.67% y 80.94% respectivamente por las variables incluidas.

Para la prueba F se toma en cuenta el valor de su probabilidad con un nivel de significación del 5%, y será mejor la ecuación que reporte menor probabilidad. Como se recordará la hipótesis alternativa establece que por lo menos uno de los parámetros incluidos en el modelo no son cero. Al examinar los resultados, se aprecia que todos los valores de probabilidad de la prueba F son muy significativos, es decir, se rechaza la hipótesis nula y se concluye con un nivel de confianza del 95%, que al menos uno de lo coeficientes no es cero en conjunto; siendo mejor para el modelo lin-log y para el modelo lineal con un valor de probabilidad del .0014 y .0017, respectivamente. A la vez los resultados de la prueba F también indican la ausencia de multicolinealidad entre las variables de las ecuaciones estimadas (confirmando lo analizado en la matriz de correlación), lo cual es muy importante tomando en cuenta que se busca un modelo explicativo, no predictivo de la demanda del consumo de la leche y con la presencia de multicolinealidad no se podrían separar los efectos de cada variable en el modelo de regresión.

La prueba t se aplica para comprobar si de manera individual las variables independientes incluidas en los modelos son significativas para explicar la variable dependiente al ser su coeficiente diferente de cero. La hipótesis nula también se verifica con los valores de probabilidad reportados a un nivel de significancia del 5%. Al observar las probabilidades, se advierte que las tres variables propuestas influyen significativamente en el consumo de leche con un nivel de confianza del 95%, sobresaliendo en orden de importancia, (independientemente del tipo de ecuación empleada), la variable población, después la variable del ingreso y finalmente la variable que describe el precio. Ahora bien, si se analizan los resultados individuales por ecuación, se tiene que las probabilidades para el modelo lin-log (precio=.0079, ingreso=.0020 y población=.0003) y para el modelo log-log (precio=.0115, ingreso=.0025 y población =.0004) son ligeramente menores, y por tanto más significativas, que las de los otros dos modelos.

Basándose en los resultados anteriores, se observa que la ecuación lin-log y log-log son las que en conjunto presentan ligeramente mejor ajuste de los datos frente a la ecuación lineal y la ecuación log-lin, excepto para la prueba *F*. Para elegir entre estos dos modelos se examinan las gráficas del valor actual, estimado y residual elaboradas con los datos de las tablas C.2.2 y C.2.4 del anexo C.2, que se muestran en las figuras 3.1 y 3.2 respectivamente.

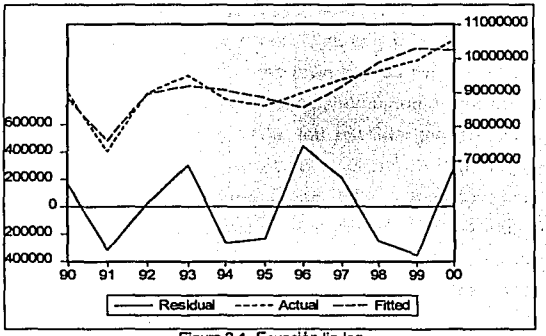


Figura 3.1 Ecuación lin-log.

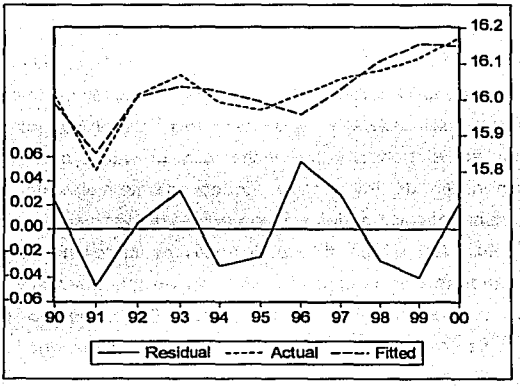


Figura 3.2 Ecuación log-log.

Como se aprecia hay un mejor ajuste de los datos en el modelo lin-log influenciado por los residuales con un dato fuera de intervalo para el año 1996 a diferencia del modelo log-log en el cual hay dos datos fuera de intervalo para los años 1991 y 1996, los cuales ya se habían señalado anteriormente. Por tanto, aunque el modelo log-log sigue siendo un buen modelo, se elige utilizar el modelo lin-log para explicar la demanda del consumo de la leche por tener residuales un poco más pequeños y valor de probabilidad menor en la prueba *F*.

A continuación se verá qué pasa con el modelo lin-log si se elimina una de las variables propuestas, considerando si hay un incremento significativo que mejore el ajuste del modelo; para ello se aplica la prueba de *variables redundantes* del EVIEWS. Se comienza con la variable LX_2, para probar la hipótesis H_o de que la regresión no es significativa. Los resultados de la regresión se consideran en la tabla 3.6.

Redundant Variables: LX2

F-statistic	13.49488	Probability	0.007927
Log likelihood ratio	11.81692	Probability	0.000587

Test Equation:
Dependent Variable: Y
Method: Least Squares
Date: 10/27/01 Time: 10:14
Sample: 1990 2000
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	33057159	70794482	0.466945	0.6530
LX3	-43510635	22326688	-1.948817	0.0872
LX4	15521273	4372093.	3.550079	0.0075

R-squared	0.644720	Mean dependent var	9165091.
Adjusted R-squared	0.555900	S.D. dependent var	831459.2
S.E. of regression	554091.1	Akaike info criterion	29.51505
Sum squared resid	2.46E+12	Schwarz criterion	29.62356
Log likelihood	-159.3328	F-statistic	7.258731
Durbin-Watson stat	1.663819	Prob(F-statistic)	0.015932

Tabla 3.6 Regresión sin LX_2 ajustada por las otras variables.

En primer lugar, el valor de probabilidad para la prueba F parcial de LX_2 en presencia de las otras variables, es muy significativo (0.008) lo cual lleva a aceptar la hipótesis alternativa de que la regresión es significativa con el precio. Como se recordará, el coeficiente de determinación ajustado es particularmente útil en estas pruebas; en esta ocasión bajó mucho su valor comparado con los resultados de la tabla 3.5, explicando solo con un 55.59% las dos variables incluidas en el modelo. Adicionalmente, tomando en cuenta el tamaño muestral, es útil verificar también los valores de los t individuales; como se observa, para coeficientes del ingreso su valor de probabilidad deja de ser significativo con un 95% de confianza y para el coeficiente de la población aunque aumentó su probabilidad, sigue siendo significativo. Por tanto, se prueba omitir la variable del ingreso en presencia del precio y la población, lo cual se puede consultar en la tabla 3.7.

Redundant Variables: LX3			
F-statistic	23.01510	Probability	0.001972
Log likelihood ratio	16.01370	Probability	0.000063

Test Equation:
Dependent Variable: Y
Method: Least Squares
Date: 10/29/01 Time: 10:53
Sample: 1990 2000
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-97950772	46642514	-2.100032	0.0689
LX2	951917.7	4028952.	0.236269	0.8192
LX4	8993899.	3353238.	2.682153	0.0278

R-squared	0.479687	Mean dependent var	9165091.
Adjusted R-squared	0.349608	S.D. dependent var	831459.2
S.E. of regression	670545.7	Akaike info criterion	29.89657
Sum squared resid	3.60E+12	Schwarz criterion	30.00509
Log likelihood	-161.4311	F-statistic	3.687676
Durbin-Watson stat	2.066405	Prob(F-statistic)	0.073293

Tabla 3.7 Regresión sin LX_3, ajustada por LX_2 y LX_4.

De igual manera la probabilidad de F para LX_3 ajustada por LX_2 y LX_4, es muy significativa (0.002) lo cual lleva a aceptar la hipótesis alternativa de que la regresión

también es significativa con el ingreso. Por ello se puede observar que los valores de los estadísticos caen al eliminar esta variable, particularmente el coeficiente de determinación ajustado solo explica el 34.96% con el precio y la población (lo que se comprueba también con la probabilidad de la F de la regresión total de .0733), y el valor de probabilidad de la prueba t para el precio (0.8192) deja de ser significativa al nivel del 5%. Con estos resultados, se prueba a eliminar las variables del precio y del ingreso que han dejado de ser significativas al eliminarse una de ellas en el modelo. Esto se aprecia en la tabla 3.8.

Redundant Variables: LX2 LX3

F-statistic	11.61227	Probability	0.005978
Log likelihood ratio	16.09019	Probability	0.000321

Test Equation:
Dependent Variable: Y
Method: Least Squares
Date: 10/27/01 Time: 10:16
Sample: 1990 2000
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	-91280278	35125985	-2.598654	0.0288
LX4	8802043.	3078050.	2.859617	0.0188

R-squared	0.476056	Mean dependent var	9165091.
Adjusted R-squared	0.417840	S.D. dependent var	831459.2
S.E. of regression	634398.4	Akaike info criterion	29.72171
Sum squared resid	3.62E+12	Schwarz criterion	29.79405
Log likelihood	-161.4694	F-statistic	8.177409
Durbin-Watson stat	2.042909	Prob(F-statistic)	0.018792

Tabla 3.8 Regresión sin LX_2 y LX_3 ajustada por LX_4.

Nuevamente el valor de probabilidad de la prueba F parcial, también es significativo con el valor de 0.005, lo cual lleva a aceptar la hipótesis alternativa que la regresión es significativa con el precio y con el ingreso. Del mismo modo, se puede observar que el valor del coeficiente de determinación ajustado cae, explicando en este caso un 41.78% el modelo con la población (obsérvese que el valor de F en la regresión total si es significativo), no obstante la probabilidad de la prueba t para la población (0.0188) sigue siendo significativa para un 95%, de confianza (pero no para un 99%).

Finalmente se prueba la regresión con una variable tendencial, tomando en consideración que el tiempo puede ser un factor importante en este modelo. Los valores de esta variable, denotada por la letra *T*, se muestran en la tabla del anexo C. La prueba a utilizar en EVIEWS es la de variables omitidas, cuyos resultados se presentan en la tabla 3.9; en este caso la H_o establece que la regresión no es significativa si se incluye esta variable al ser su coeficiente igual a cero.

Omitted Variables: T

F-statistic	0.064752	Probability	0.807634
Log likelihood ratio	0.118076	Probability	0.731131

Test Equation:
Dependent Variable: Y
Method: Least Squares
Date: 10/27/01 Time: 10:22
Sample: 1990 2000
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.42E+08	1.69E+08	0.837648	0.4343
LX2	9990300.	4083026.	2.446788	0.0500
LX3	-89895217	24972105	-3.599825	0.0114
LX4	20605548	19167922	1.075002	0.3237
T	81285.86	319439.6	0.254464	0.8076

R-squared	0.879950	Mean dependent var	9165091.
Adjusted R-squared	0.799917	S.D. dependent var	831459.2
S.E. of regression	371916.9	Akaike info criterion	28.79368
Sum squared resid	8.30E+11	Schwarz criterion	28.97455
Log likelihood	-153.3653	F-statistic	10.99482
Durbin-Watson stat	2.212399	Prob(F-statistic)	0.006297

Tabla 3.9 Regresión aumentando el tiempo.

En contraste con los resultados anteriores, el valor de probabilidad de la prueba *F* parcial por ser mayor, es desfavorable en presencia del tiempo, ajustada por las otras variables con un valor de probabilidad de 0.8076, lo cual lleva a aceptar la hipótesis nula de que el coeficiente es cercano a cero y la regresión no es significativa con el tiempo. Por ello, aunque el coeficiente de determinación ajustado reporta un porcentaje de explicación que es del 79.99%, al examinar los valores de probabilidad de la prueba *t* para la población (0.327) deja de ser significativa y para el precio apenas lo es (.0500).

Con las pruebas realizadas, se concluye que las variables del precio, del ingreso y de la población, son significativas para explicar la demanda de la leche y la eliminación de alguna de ellas disminuye los demás valores de los estadísticos de una u otra forma.

Posteriormente se correrá la regresión eliminando las observaciones del año 1996, (año en el que se detectó un dato fuera de intervalo), con la finalidad de examinar los resultados a pesar de que no reportaron los residuales falta de normalidad. Para eliminar estos datos, se corrió la regresión con valores del 1 al 10 con la correspondencia en años de: 1=1990, 2=1991, 3=1992, 4=1993; 5=1994; 6=1995; 7=1997, 8=1998, 9=1999; 10=2000, y no halla confusión en los años de la muestra restringida. Los resultados se tienen en la tabla 3.10.

Dependent Variable: Y
Method: Least Squares
Date: 10/29/01 Time: 14:09
Sample: 1 10
Included observations: 10

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	1.52E+08	48453319	3.140519	0.0201
LX2	12872512	2709221.	4.751370	0.0032
LX3	-1.15E+08	19810248	-5.789292	0.0012
LX4	28582952	3640016.	7.852425	0.0002

R-squared	0.925137	Mean dependent var	9180273.
Adjusted R-squared	0.887705	S.D. dependent var	874826.4
S.E. of regression	293157.6	Akaike info criterion	28.30398
Sum squared resid	5.16E+11	Schwarz criterion	28.42502
Log likelihood	-137.5199	F-statistic	24.71546
Durbin-Watson stat	2.566208	Prob(F-statistic)	0.000892

Tabla 3.10 Regresión eliminando datos atípicos del año 1996.

Como se esperaba, los resultados de la regresión en conjunto han mejorado sobre todo del coeficiente de determinación ajustado, que indica un porcentaje de explicación del 88.77% del consumo; de igual manera la probabilidad para la prueba F, es muy significativa (.0008) ya que están relacionadas. De manera individual, sobresale nada mas la probabilidad para la variable del precio (.0032).

A pesar del aumento en los resultados de los estadísticos, al examinar la figura 3.4, se aprecia un patrón de los residuales con tendencia creciente a lo lago del tiempo y

la presencia de dos datos atípicos para los años de 1997 y 1999, lo cual no es real tomando en cuenta el comportamiento del consumo para esos años que fue creciente si se elimina el del año 1996. De esta manera, se confirma el hecho de mantener este año, pues aunque disminuyen los valores de la regresión, reflejan el comportamiento real del consumo de leche.

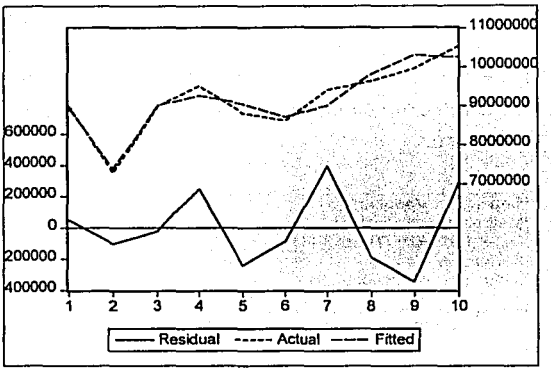


Figura 3.4 Valores actuales, estimados y residuos de la regresión sin datos atípicos del 96.

D. PRESENTACIÓN DEL MODELO FINAL.

D.1 INTERPRETACIÓN DE LOS RESULTADOS.

De acuerdo a los resultados establecidos en la sección anterior se selecciona el modelo lin-log para explicar la demanda del consumo de la leche. En economía este modelo expresa el cambio absoluto en Y debido a un cambio porcentual (o tasa de crecimiento porcentual) en las X's[8]; es decir, expresa el cambio absoluto del consumo de la leche ante un cambio porcentual en el precio, el ingreso y la población. Las

[8] La interpretación del coeficiente de la pendiente, por ejemplo, de β_2 usualmente se representa como: β_2=(cambio en Y) / (cambio en ln X) = (cambio en Y)/(cambio relativo en X). Dado que un cambio en el logaritmo de un número es un cambio relativo, simbólicamente se expresa como: $\beta_2 = (\Delta Y)/(\Delta X/X)$ y en forma equivalente $\Delta Y = (\beta_2/100)\Delta X$, que expresa como el cambio absoluto en Y es igual a β_2 veces el cambio relativo en X. Si éste último es multiplicado por 100, entonces la ecuación proporciona el cambio absoluto en Y ocasionado por un cambio porcentual en X.

observaciones para el logaritmo de cada variable se muestran en la sección C.7 del anexo C, cuya ecuación estimada se presenta a continuación:

$$Y_t = 100{,}828{,}174. + 10{,}661{,}131.87 \ln X_{2t} - 93{,}369{,}416.2 X_{3t} + 25{,}369{,}543.27 \ln X_{4t}$$

Con este modelo se obtienen las siguientes conclusiones:

El modelo lin-log cumple con todos los supuestos básicos del modelo de regresión lineal, incluso se vio que a pesar de la presencia de un dato atípico en el año 1996 no se viola el supuesto de normalidad. Como se explicó, este dato es representativo de la realidad pues fue el año en que se liberó el precio de la leche, propiciando cambios un poco más bruscos en el consumo de la leche (primero un decremento y después un aumento el cual se ha sostenido a la fecha); por ello, es necesario mantener este dato para explicar el comportamiento real la demanda del consumo de leche, pues si se eliminara, ya no describiría la demanda de la situación actual. Es de notar que tampoco se viola el supuesto de multicolinealidad en un grado mayor que pueda afectar la explicación de los coeficientes estimados para todas las variables.

En base a las pruebas estadísticas, el valor del coeficiente de determinación indica que las variables del precio, ingreso y población explican alrededor del 87.87% la demanda del consumo de leche, es decir, que si influyen en conjunto estas variables en la demanda y el resto se explica por otros factores aleatorios no incluidos. La prueba F con un valor de probabilidad ρ de .0013 revela con un 95% de confianza que al menos uno de los coeficientes involucrados en el modelo es diferente de cero; es decir, se tiene una probabilidad muy baja de cometer un error de tipo I. A la vez el resultado de la prueba F indica que las variables independientes no están relacionadas linealmente y no afectan a la variable independiente, es decir, confirma la ausencia de multicolinealidad entre las variables. Individualmente la significancia de las variables que describen el precio, el ingreso y la población, se examina de acuerdo a los valores ρ de la distribución t con un valor de probabilidad .0079, .0020 y .0003 respectivamente, lo cual indica con un 95% de confianza que efectivamente influyen por separado estas variables para explicar la demanda del consumo de la leche pues sus coeficientes estimados son significativamente diferentes de cero; dicho de otro modo, se tiene una probabilidad muy baja de cometer un error de tipo I. Estos resultados manifiestan que las variables del precio, del ingreso y de la población deben permanecer en el modelo, pues ya sea de manera individual o de

manera conjunta, ejercen una influencia significativa en la explicación de la demanda del consumo de la leche.

Los signos de los coeficientes no son los esperados, excepto el de la variable que se refiere a la población; estos cambios en los signos en el precio y en el ingreso se explican en la siguiente sección con el uso de las elasticidades.

Para hacer la interpretación de los coeficientes en el modelo lin-log en forma porcentual, como hacen los economistas, se debe multiplicar el valor de los coeficientes estimados de la pendiente por 0.01 (o dividirlo entre 100), y obtener así el cambio absoluto en la variable dependiente ocasionado por un cambio porcentual en las variables independientes. De esta manera, la interpretación de los coeficientes estimados en la ecuación lin-log es como sigue:

- El coeficiente del intercepto expresa que se estima un consumo promedio de leche alrededor de 1,008,281 litros al año, cuando se da un valor de cero al precio, al ingreso y a la población.
- La interpretación del coeficiente para la variable que describe el precio, LX_2, indica que ante un cambio porcentual de un peso en el precio real, la cantidad demandada aumentará en promedio en 106,611.32 miles de litros al año.
- El coeficiente de la variable del ingreso, LX_3, muestra que ante un incremento porcentual en el ingreso real de un peso, hay una disminución promedio de 933,694.16 miles de litros en el consumo de leche al año.
- Finalmente el coeficiente de la variable que describe la población, LX_4, expresa que hay un aumento en promedio del consumo de leche de 253,695.43 miles de litros al año como resultado del incremento en la tasa de crecimiento porcentual de la población.

La variable que describe la población influye en la demanda del consumo de la leche. El signo positivo que se obtuvo en la ecuación estimada, refleja que la cantidad demandada está en proporción con el aumento de las personas. Si bien es claro que a mayor población le corresponde un mayor consumo, la población ha sufrido un cambio a partir del año de 1996, año en que se ha incrementado pero en una tasa más baja, en comparación de la demanda del consumo de leche que ha crecido en una tasa mayor.

A pesar de que en México no se ha alcanzado el consumo per-cápita recomendado por la FAO de 182.5 litros al año, si ha aumentado el consumo de la leche en estos últimos años sobre todo a partir de 1996, como se vio en la gráfica 1.7 (capítulo uno) en el año 2000 en que se tuvo un nivel de consumo de 108.25 litros por persona al año; sin embargo, no se ha superado el nivel de 1990 de casi 120 litros al año.

Finalmente, aunque en teoría se sabe que la leche es considerada un producto básico, en la realidad en nuestro país no se consume como tal ya sea por el precio, por el ingreso, por que no se le da la importancia a su consumo o simplemente por decisión personal; no obstante se incrementa su consumo en cierta proporción con el aumento de la población.

D.2 EXPLICACIÓN DE LA DEMANDA DEL CONSUMO DE LA LECHE EN MÉXICO.

En el estudio de la demanda interesa conocer de forma particular las elasticidades precio e ingreso de la misma. Para la obtención del coeficiente de elasticidad precio e ingreso en el modelo lin-log, se aplican las fórmulas $\beta_2\left(\frac{1}{Y}\right)$ y $\beta_3\left(\frac{1}{Y}\right)$ respectivamente (según lo visto en la sección B.4 del segundo capítulo). Los coeficientes obtenidos de acuerdo a la ecuación estimada se muestran en la tabla 3.11.

Elasticidad	*Coeficiente*
ϵ_p	1.17
ϵ_I	-10.22

Tabla 3.11 Coeficientes de elasticidad precio e ingreso.

El coeficiente de la *elasticidad precio*, ϵ_p, de 1.17 al ser mayor que uno en valor absoluto, muestra que la leche es un bien elástico respecto al precio, es decir, es muy sensible a los cambios en el precio real. A un cambio del 1% en el precio real, le corresponde un cambio porcentual mayor en la cantidad demandada. La demanda elástica sugiere que si bajara el precio, el gasto del consumidor aumentaría ya que la

cantidad demandada subiría más que proporcionalmente a la disminución del precio. Adicionalmente el coeficiente de elasticidad alto quiere decir también que no se considera indispensable el consumo de la leche, a pesar de ser un producto básico, por los cambios tan altos que ocurren en la demanda respecto al precio.

El ser tan alta la elasticidad precio puede deberse a distintos factores, de acuerdo a la información obtenida algunos de ellos son:

- En primer lugar la disponibilidad de productos sustitutos, no por que sean comparables, pero de acuerdo a las entrevistas, los jugos y los yogures han disminuido el consumo de leche; además muchas personas prefieren el refresco, café e incluso el agua se toma en una cantidad considerable en lugar de la leche.
- En segundo lugar, también puede influir la proporción del ingreso que se gasta, pero esto se comprobará con la elasticidad ingreso.
- El tiempo es otro factor que se refleja en la demanda elástica, considerando que se toma un periodo anual en el cual ya se ajustaron los consumidores a los cambios en el precio.
- Finalmente, un factor que influye para que una demanda sea más elástica es lo competitivo de la industria lechera, a diferencia de las industrias que pueden absorber casi todo el mercado.

A largo plazo se esperaría también que la diversificación de productos haga más elástica la demanda.

El signo del coeficiente para la *elasticidad ingreso*, ϵ_I, por ser negativo indica que la leche es inelástica respecto al ingreso, es decir, no hay cambios bruscos al aumentar o disminuir el ingreso. Al aumentar el ingreso real disminuye la cantidad demandada y origina un desplazamiento de la curva de la demanda hacia la izquierda. El signo negativo también indica que la leche es un bien inferior (como ocurre con los productos básicos, contrario a lo que señaló la elasticidad precio). Por ser un bien inferior señala que si aumentara el ingreso del consumidor, se sustituiría la leche por otro bien mejor; esta situación es válida para algunos estratos, pues si aumentara el ingreso del consumidor preferirían productos más caros como sería el yogurt.

Por otro lado, a pesar del incremento mínimo en el nivel de ingresos y la liberación del precio de la leche a partir de 1996, la demanda del consumo de la leche ha aumentado en las clases más altas que han mantenido el consumo de leche, aunque para los estratos de ingresos bajos (en el caso de leche subsidiada) no se puede hablar de un aumento en el consumo de leche. Esta situación se apreció hace poco con el aumento de 50 centavos que se quiso aplicar al precio de la leche que proporciona LICONSA; para los estratos más altos esta cantidad podría no ser relevante, pero para ellos implicaría tal vez dejar de comprar un kilo de tortillas por comprar leche, lo cual no es posible; este caso refleja la dependencia de la demanda ante variaciones en el ingreso para ciertos niveles.

La magnitud del coeficiente, supone que ante un aumento por cada unidad porcentual en el ingreso real, causará una disminución en la cantidad demanda aproximadamente del 10.22 por ciento. Para los industriales de la leche, este resultado implica que no compartirán proporcionalmente los aumentos del ingreso nacional. Esto supondría también que la industria de la leche no es tan sensible a los cambios de la actividad económica, lo que implicaría que es aprueba de recesiones (por ser un producto básico aumentaría su demanda); pero como no podría participar en una economía de pleno crecimiento, se desearía en ingresar a otras industrias que proporcionen mejores oportunidades de desarrollo. Es un problema que normalmente se presenta en áreas de la agricultura (con productos alimenticios con una elasticidad precio menor que uno), en la cual se ven afectados principalmente los ingresos de los granjeros que se mantienen en proporción de los ingresos de los trabajadores urbanos.

En el caso de la leche, se dijo que el Gobierno ha tenido en estos últimos años un interés especial por impulsar el sector lechero, por lo cual se implementaron procedimientos específicos a partir de 1996, provocando que en los últimos años se haya percibido el desarrollo del sector de manera creciente. Por tanto, en la industria lechera si influyen los cambios de la economía nacional, pero se necesitarían más observaciones (a partir de 1996) para apreciar qué tan significativos son los cambios en el modelo propuesto. Hay que recordar que el sector lechero es muy heterogéneo y por ello, se necesitaría un modelo particular para explicar la demanda en cada región, por productos etc.; sin embargo el modelo refleja la situación actual del sector lechero en conjunto en el periodo examinado.

A continuación se establece un intervalo de confianza para el precio y el ingreso, con la finalidad de establecer un rango de las elasticidades. Aplicando la ecuación 2.77, se considera un intervalo de confianza del 95%; el valor de la t para 7 grados de libertad es de 2.3646 y los demás valores se obtienen de la tabla C.1.2, del anexo C, con lo cual se establecen los intervalos que se muestran en la tabla 3.12.

Elasticidad	*Intervalos*
ϵ_p	.94 ≤ 1.17 ≤ 1.39
ϵ_I	-5.17 ≤ -10.22 ≤ 15.21

Tabla 3.12 Coeficientes de elasticidad precio e ingreso.

Como se observa el cambio es principalmente para la elasticidad precio que puede llegar a ser inelástica con un valor de .94 (menor que uno), lo cual es cierto para los estratos más altos, pero en una mínima parte. En cambio la elasticidad ingreso sigue siendo inelástica (con signo negativo), aunque la magnitud del 5.17 se puede acercar a la elasticidad en un momento dado, lo que implicaría una mayor influencia en la demanda de la leche ante los cambios económicos en nuestro país.

CONCLUSIONES

Conforme a la información de las tres fuentes de consultadas (teoría económica, situación del sector lechero nacional y entrevistas), y de acuerdo a la información disponible; la especificación de los modelos quedó establecida de la siguiente manera: como variables principales el precio, el ingreso y la población y, como formas funcionales las ecuaciones lineal, lin-log, log-lin y log-log, para explicar la demanda del consumo de la leche en nuestro país. Los resultados obtenidos para las ecuaciones propuestas, fueron muy similares reflejando en muchos casos una diferencia casi imperceptible que pudo ser consecuencia del tamaño muestral utilizado, lo cual dificultó la selección de la relación funcional. Sin embargo, el conjunto de pruebas aplicadas me ayudaron a elegir el modelo lin-log como más conveniente para la explicación de la demanda del consumo de la leche.

El primer criterio para validar la calidad de los resultados de la regresión, es la verificación de los supuestos básicos del modelo de regresión lineal múltiple. El modelo lin-log cumplió con todos los supuestos básicos lo que garantiza los resultados del análisis de regresión, así como la aplicación de las pruebas estadísticas. Aunque detecté la presencia de un dato atípico para el año de 1996, no lo eliminé ya que es representativo de la realidad tomando en cuenta que en ese año se liberó el precio de la leche ultrapasteurizada y pasteurizada provocando un cambio en el consumo de la leche, y observé que no afectó la normalidad en los residuos. Es de notar que las variables independientes en la ecuación cumplen el supuesto de no multicolinealidad, necesario para poder separar los efectos de cada variable entre sí y hacer uso del modelo con fines explicativos aplicando el concepto de elasticidad.

El segundo criterio se basa en la aplicación de contrastes estadísticos, los cuales verificaron el uso de las variables independientes tanto en forma individual (prueba t en términos probabilísticos), como en conjunto (la prueba F reportada en forma de probabilidad y el coeficiente de determinación en forma de porcentaje), cuyos resultados fueron más significativos, lo cual reafirmó la selección de las variables incluidas.

Adicionalmente, apliqué la prueba de "variables redundantes" y la prueba de "variables omitidas", para ver si mejoraba el ajuste del modelo, pero los resultados no

fueron significativos (prueba *F* parcial). Corrí también la regresión eliminando el dato atípico del año 1996, pero observé un comportamiento creciente de los residuales a lo largo del tiempo y dos datos atípicos que no son representativos del consumo real de la leche.

Por consiguiente, todos los resultados obtenidos confirmaron el comportamiento semilogarítmico con el modelo lin-log, sin ningún cambio en las variables seleccionadas para explicar la demanda del consumo de la leche.

Los coeficientes del precio y del ingreso tuvieron signos contrarios a lo que marca la teoría económica, lo cual es justificado con la interpretación de las elasticidades, con las observaciones obtenidas y por los cambios que se han dado en el sector lechero, como se menciona posteriormente.

La demanda del consumo de leche depende de la población. Hay que recordar que México no ha logrado alcanzar el consumo mínimo por habitante recomendado por la FAO, es decir, la demanda supera la oferta y el reto es cubrir esta demanda nacional en proporción con el incremento poblacional, ya sea por la parte de la leche subsidiada o por la parte de la leche comercial.

El resultado de la elasticidad precio reflejó tener una demanda elástica, indicando que el consumo de leche responde de manera muy sensible a los cambios en el precio. Dada la situación del sector lechero, consideré que entre los factores que más influyen para que la demanda sea elástica se encuentran: el no considerarse un bien indispensable para todas las personas, pues hay productos que sin ser propiamente sustitutos como marca la teoría, se consumen en lugar de la leche; la proporción del ingreso que se gasta en el consumo de la leche; el periodo de tiempo influyó tomando en cuenta que en un año los consumidores ya se ajustaron al precio; finalmente lo competitivo de la industria lechera, hace una demanda elástica. Al establecer intervalos de confianza, observé que la demanda puede ser inelástica ante el precio, lo cual es cierto para las clases más altas pero en una mínima parte.

El resultado de la demanda elástica ante el precio, confirmó lo señalado en una de las entrevistas referente al precio de la leche que es más alto en otros países, a

diferencia del nuestro, en que no se puede incrementar su precio tan bruscamente sin que afecte la cantidad consumida, lo que es propio de los productos con una demanda elástica. En este modelo, el control del precio de la leche influyó para que el signo del coeficiente del precio sea positivo en el periodo estudiado.

La demanda resultó ser inelástica respecto al ingreso por tener signo negativo, lo cual indicó que, además de no ser muy sensible a los cambios en el ingreso, se trata de un bien básico. La inelasticidad de la leche ante el ingreso, es válido para las clases mas altas, sin embargo, para los estratos bajos es determinante el ingreso, como se explicó en el caso de LICONSA. Por otra parte, el coeficiente del ingreso reflejó las circunstancias para la mayoría de los industriales del sector lechero cuya situación en general es muy desfavorable. Si la intervención del Gobierno logra mejorar la situación, sobre todo de los pequeños y medianos productores, la elasticidad ingreso cambiaría a signo positivo y colocaría a la leche como un bien normal o superior, dependiendo del valor del coeficiente. No obstante, si se realizara en este momento un estudio por empresas, el signo de elasticidad ingreso sería positivo para alguna de ellas.

Hay que considerar que el periodo estudiado contempla el cambio casi imperceptible en el nivel del ingreso (libre de inflación) sobre todo a partir del año de 1996; simultáneamente en este año ocurre la liberación del precio de la leche ultra pasteurizada y poco tiempo después el de leche pasteurizada; por tanto, estos cambios también se reflejan en la demanda que aparece desfavorable ante las variaciones en el ingreso; por ello al establecer intervalos de confianza los resultados siguen indicando la inelasticidad respecto al ingreso y en una menor proporción el acercamiento a la elasticidad. En caso de que siga en aumento el consumo de la leche y mejoren los ingresos, los signos y los demás resultados del modelo cambiarían a futuro si se tomara una muestra a partir del año 1996 en que se han dado cambios significativos en el sector lechero.

Como se puede uno dar cuenta, para el caso de la leche los resultados de un modelo de demanda no serán los mismos de una empresa a otra, de una región a otra, ni para toda la industria, dado lo heterogéneo del sector lechero. Tampoco serán los mismos resultados para diferentes periodos de estudio. Todo esto implica que los resultados obtenidos con este modelo no se pueden generalizar.

En este trabajo se presentaron varias limitaciones que dificultó la explicación de la demanda del consumo de leche. La primera limitación fue la falta de teorías económicas sobre el consumo de alimentos, pues las que hay son muy atrasadas. Otra limitación fue en cuanto a la periodicidad, ya que en forma agregada no es útil manejar un periodo menor y el tamaño muestral obstaculizó la evaluación de los cambios estructurales que se han dado, sobre todo a partir de 1996 (en una muestra mayor estos cambios estructurales se pueden detectar con algunas pruebas que incluye el paquete econométrico).

También fue determinante la falta de información para recopilar los datos sobre algunas variables. Respecto a la falta de información de las variables incluidas, no hay reportadas series de datos mayores para el consumo, y respecto al precio de la leche únicamente lo pude obtener sin poder separar los tipos y procesos de la leche. En cuanto a la falta de información de otras variables, en esta era de globalización hubiera sido importante incluir el precio internacional de la leche por la dependencia tan fuerte que tiene el sector con el mercado exterior, sin embargo, las series consultadas (en la SAGAR) han ido cambiando a lo largo del tiempo, lo que constituiría una muestra no muy homogénea; otra variable a incluir en estudios de demanda de acuerdo a la teoría, son los precios de los bienes que se consideran sustitutos, complementarios, y los cambios en los gustos o preferencias, para detectar más características del bien; sin embargo, no existen series de datos de estas variables que consideré significativas en el modelo de la demanda y, si existieran, por el tamaño muestral no sería viable introducirlas ya que afectarían los resultados del modelo.

La falta de información es muy significativa también para realizar estudios de la demanda de la leche en particular, dado que las diferentes dependencias no tienen una igualdad de información entre sí. Por ejemplo, para el caso de los extensores sería importante estar al corriente de su comportamiento a futuro por el peligro que representan para desplazar la demanda de la leche ante un precio más bajo, pero no hay información por separado. También me comentaron que a futuro se piensa que prevalezca solamente el proceso de leche ultrapasteurizada, en este caso también sería recomendable primero realizar un estudio nacional, que no se puede llevar a cabo sin la información necesaria de los otros procesos.

La experiencia del trabajo realizado, me confirma la gran necesidad de contar con información suficiente, uniforme y de calidad para poder hacer estudios que garanticen los resultados de los modelos, particularmente para aplicar las teorías económicas de acuerdo a la problemática de nuestro país, recordando que las aportaciones econométricas se orientan al uso de grandes muestras de datos y de ecuaciones simultáneas, lo cual no siempre es aplicable debido a la falta de información. Por tanto, es una oportunidad para los alumnos de MAC desarrollar bases de datos uniformes y confiables, para el buen uso de los modelos; o en su defecto, la búsqueda de técnicas aplicables para muestras pequeñas.

Tomando en cuenta que se necesita que pasen varios años para aumentar la información, recomendaría la aplicación de las técnicas econométricas en las empresas, dado que tienen un mejor control de la información de sus productos. Por otra parte, me di cuenta en las entrevistas realizadas, que dos de ella compran estudios a la misma compañía privada para conocer la demanda en el ámbito nacional; y sería interesante conocer cómo realizan sus estudios ante la falta de información y, del mismo modo, es un campo en el que se puede desarrollar un alumno de MAC, realizando modelos.

A corto plazo, pienso que convendría un trabajo conjunto de los alumnos de MAC y los estudiantes de economía. A los estudiantes de economía les sería útil para aprovechar mejor las herramientas que proporcionan las pruebas estadísticas y la paquetería. Para los alumnos de MAC, sería un gran apoyo contar con más herramientas que ayuden, entre otras cosas, a la identificación de las variables económicas, así como el manejo de términos económicos en general, para poder hacer estudios más precisos -ya sea en el ámbito empresarial, nacional o internacional-, respaldados con los conocimientos que proporciona la carrera.

Pienso que la econometría tiene un gran futuro en nuestro país, pero depende de la colaboración de las tres ramas que la integran para no caer en los conflictos que se mencionaron en la breve reseña del capítulo uno, de querer trabajar solamente desde el punto de vista de cada área.

ANEXOS

ANEXO A. DENOMINACIÓN DE FRACCIONES ARANCELARIAS Y FACTORES DE CONVERSIÓN.

Aquí se incluyen las denominaciones de las fracciones arancelarias, indispensables para consultar la información de las importaciones y exportaciones de la leche fluida y en polvo; ya que por sí mismas no se entienden de la forma en que se encuentran en los boletines de la leche, dificultando que esta información se pueda obtener directamente sin estas denominaciones.

Aunque en apariencia parezcan muchas fracciones (tablas A.1 y A.3), al agruparse por tipo de leche (tablas A.2 y A.4), solamente se está hablando de tres tipos de leche (leche fluida y en polvo -sin grasa y con grasa-), recordando que se excluyó la leche evaporada y condensada. Por ejemplo, para la leche fluida solo se encuentra en los boletines de la leche las seis fracciones que se indican en la tabla A.2, correspondientes a las características de la fracción 04 a la 0401 3099 en la tabla A.1, sin especificar que se trata de la leche fluida.

Asimismo, las tablas A.2 y A.4 contienen los factores de conversión necesarios para hacer la conversión a litros de la leche en polvo (la leche fluida ya viene en litros); estos factores de conversión son tomados como referencia, ya que varían en función de las características de los sólidos, al tipo de leches y al país de origen. Para hacer la conversión a litros, por ejemplo de la leche en polvo sin grasa, simplemente se suman las fracciones correspondientes a este tipo (0402 1001 y 0402 1099) de acuerdo a las cantidades mensuales que reportan los boletines de la leche, y se multiplica la cantidad resultante por el factor apropiado, que en este caso será de 8.33.

A.1. IMPORTACIONES.

FRACCIÓN	*DENOMINACIÓN*
04	Leche y productos lácteos, productos comestibles de origen animal no expresadas en otras partidas
0401	Leche y nata (crema) sin concentrar, sin adición de azúcar no de otros edulcorantes
0401 10	Con un contenido de materias grasas en peso, menor o igual al 1%
0401 1001	En envases herméticos
0401 1099	Los demás
0401 20	Con un contenido de materias grasas, en peso superior al 1%, pero inferior o igual al 6%
0401 2001	En envases herméticos
0401 2099	Los demás
0401 30	Con un contenido de materias grasas, en peso superior al 1% y mayor que 6%
0401 3001	En envases herméticos
0401 3099	Los demás
0402	Leche y nata (crema) concentradas o con adición de azúcar o de otros edulcorantes
0402 10	En polvo, gránulos o demás en formas sólidas, con un contenido de materias grasas en peso menor o igual al 1.5%
0402 1001	Leche en polvo o pastillas
0402 1099	Los demás
0402 21	Sin adición de azúcar ni de otros edulcorantes
0402 2101	Leche en polvo o en pastillas
0402 2199	Los demás
0402 29	Las demás
0402 2999	Las demás

Tabla A.1 Denominación de las fracciones arancelarias de importación de leches.

GRUPO	*FRACCIONES ARANCELARIAS*	*FACTORES DE CONVERSIÓN*
Leche fluida	0401 1001, 0401 1099, 0401 2001, 0401 2099, 0401 3001, 0401 3099	1
Leche en polvo (grasa menor o igual al 1.5%)	0402 1001, 0402 1099	8.33
Leche en polvo (grasa mayor que 1.5%)	0402 2101, 0402 2199, 0402 2999	11.5

Tabla A.2 Fracciones arancelarias y factores de conversión a litros por grupos de importación.

A.2. EXPORTACIONES.

FRACCIÓN	*DENOMINACIÓN*
0	Leche y productos lácteos, productos comestibles de origen animal no expresadas en otras partidas
0401	Leche y nata (crema) sin concentrar, sin adición de azúcar no de otros edulcorantes
0401 10	Con un contenido de materias grasas en peso, menor o igual al 1%
0401 20	Con un contenido de materias grasas, en peso superior al 1%, pero inferior o igual al 6%
0401 30	Con un contenido de materias grasas, en peso superior al 1% y mayor que 6%
0402	Leche y nata (crema) concentradas o con adición de azúcar o de otros edulcorantes
0402 10	En polvo, gránulos o demás en formas sólidas, con un contenido de materias grasas en peso menor o igual al 1.5%
0402 21	Leche en polvo sin adición de azúcar ni de otros edulcorantes
0402 29	Las demás
0402 2999	Las demás

Tabla A.3 Denominación de las fracciones arancelarias de exportación de leches.

GRUPO	*FRACCIONES ARANCELARIAS*	*FACTORES DE CONVERSIÓN*
Leche fluida	0401 10, 0401 20, 0401 30	1
Leche en polvo (grasa menor o igual al 1.5%)	0402 10	8.33
Leche en polvo (grasa mayor que 1.5%)	0402 21, 0402 29 ó 0402 2999 (a partir del 98)	11.5

Tabla A.4 Fracciones arancelarias y factores de conversión a litros por grupos de exportación.

ANEXO B. INTRODUCCIÓN AL ECONOMETRIC VIEWS (EVIEWS)

El *Econometric Views* o *EVIEWS*, es una nueva versión del programa TSP (Time Series Processor) lanzado en 1981. EVIEWS ha sido desarrollado por economistas por lo que la mayoría de sus aplicaciones son en economía, específicamente está orientado hacia la estimación de modelos econométricos. EVIEWS cuenta con técnicas avanzadas las cuales se basan en investigaciones publicadas en libros y revistas principalmente de econometría, por lo que no siempre son fáciles de entender a simple vista. La mayoría de sus herramientas están pensadas para utilizarse con grandes cantidades de datos; asimismo, aunque puede utilizar datos de corte transversal, EVIEWS trabaja principalmente con series de tiempo.

Algunas de las aplicaciones con que dispone EVIEWS son:

1) Lectura de archivos de datos y escritura en formatos de hojas de cálculo.
2) Manipulación de series por medio de fórmulas matemáticas.
3) Descripción estadística.
4) Gráficas de datos.
5) Regresión con Mínimos Cuadrados Ordinarios.
6) Regresión con Mínimos Cuadrados en dos etapas.
7) Regresión con Mínimos Cuadrados no lineales.
8) Modelos autorregresivos y promedios móviles.
9) Rezagos polinomiales distribuidos.
10) Estimación de modelos Logit y Probit.
11) Opciones binarias.
12) Predicción.
13) Solución de modelos de ecuaciones simultáneas.
14) Estimaciones lineales y no lineales de ecuaciones simultáneas y simulaciones.

En este apartado se proporciona una descripción solamente de las herramientas empleadas en el desarrollo del trabajo, utilizando la versión 3.1 de EVIEWS. Como apoyo en el manejo de las demás aplicaciones, se puede consultar el manual elaborado por el

maestro Gustavo Vargas (ver bibliografía), que es una traducción basada en la ayuda que incluye el paquete el cual viene en inglés.

- **Introducción de datos.**

Para abrir un archivo nuevo de trabajo, se da clic en la opción *FILE/ NEW /WORKFILE* de la pantalla del menú principal de EVIEWS. Antes de continuar, se hace la observación de un área blanca, llamada *COMAND WINDOW*, en la cual se pueden teclear directamente los comandos sin abrir las ventanas, lo que es de ayuda; en esta introducción se verán algunos comandos para utilizarse en esta área.

Al abrir el archivo de trabajo se despliega una ventana llamada *WORKFILE RANGE* en la que se debe especificar si la frecuencia de los datos es anual, semestral, trimestral, mensual, semanal o diaria; *Start Date* es la fecha de inicio de las series y *End Date* es la fecha final de las series. Si los datos tienen una frecuencia diferente o irregular se elige *Undate OR Irregular* y solo se especifica el fin de la serie. Si se elige un periodo de tiempo regular, EVIEWS los identifica con reglas específicas; en este caso el periodo es anual y abarca hasta el siglo XXI por lo que se debe poner el año completo 1990 a 2000. Para una muestra comprendida en el siglo XX se pueden poner solamente los últimos dos dígitos del año, como 90 o 96.

Una vez especificado el periodo, se presiona *OK* y se despliega una ventana llamada *Workfile: UNTITLE* con una barra de herramientas en la parte superior del archivo. Este archivo de trabajo también inicia con un vector de coeficientes precedido de un icono identificado con la letra griega alfa, α; también inicia con una serie de residuales denominada *RESID*, en esta caso identificada con un dibujo como icono, ☑. EVIEWS maneja distintos niveles para objetos como son: series, grupos, ecuaciones, vectores autorregresivos, gráficas, etc.; cada uno identificado con un icono en particular.

Para empezar a trabajar se necesitan editar las de series de las variables incluidas, por lo que se puede dar clic en el menú principal en *QUICK/ Empty Group* y aparece una hoja de cálculo llamada *Group: UNTITLE* en la que se introducen manualmente las observaciones para cualquier numero de series, con la frecuencia ya definida (para este trabajo es de 1990 al 2001). Al introducir la primera observación de

cada serie, automáticamente aparece el nombre SER01, SER02, etc. en el archivo de trabajo de manera individual, los cuales se pueden cambiar, como se explica más adelante. Si ya se tienen los datos en una hoja de cálculo como EXCELL, se pueden llenar las series utilizando las opciones de copiar y pegar.

Una vez que se han llenado las series de datos se da un clic en *Edit +/-*, si se desean salvar todas las series en un solo grupo se da clic en la barra de herramientas en la opción *NAME* la cual abre la caja de diálogo *Object name*, sino se elige un nombre por default aparece como GROUP01 en el archivo de trabajo (con el icono). Si se desea salvar todo el archivo de trabajo se hace clic en el menú principal en *FILE/ SAVE* y se procede como con cualquier archivo en Windows.

Para cambiar los nombres de cada serie, se selecciona el nombre situado en el archivo de trabajo (identificadas también con el icono) y con un clic derecho se elige *Rename*, que abre la caja titulada *Object name*. Es recomendable que el nombre sea corto para trabajar con las series ya que en esa caja también hay la opción de dar un titulo más especifico que identifique las series en las tablas y las gráficas; por ejemplo, un nombre puede ser PIB y la leyenda será Producto Interno Bruto.

En este trabajo se dieron los siguientes nombres:

Y: Serie de la variable que describe el consumo nacional aparente.
X2: Serie de la variable que describe el precio.
X3: Serie de la variable que describe el ingreso.
X4: Serie de la variable que describe la población.
Glineal: Grupo de variables sin ninguna transformación, pertenecientes a la forma lineal.

Finalmente, para abrir un archivo de trabajo ya existente, se da clic en *FILE/ OPEN /WORKFILE*. Para abrir una serie individualmente de un archivo de trabajo como **Y**, se da doble clic en el nombre y se abre automáticamente; de igual manera para abrir un grupo ya salvado en el archivo de trabajo como **Glineal** (o cualquier otro objeto como una ecuación, gráfica, tabla, etc.), dar doble clic en el nombre y se abrirá el grupo.

- **Generación de series.**

Como fue necesario aplicar la transformación logaritmo a las variables para las ecuaciones alternativas, se *generaron* las nuevas series a partir de las ya existentes. Para ello se presentan dos opciones:

1) Utilizar el menú principal en *QUICK/ GENERATE SERIES*, que abre una caja de diálogo llamada *Generate series by ecuation* y el formato para el logaritmo de la variable **X2** se escribe: **LX2=log(X2)**, en donde **LX2** es el nombre que se le da a la nueva variable con transformación logaritmo (obviamente puede escogerse cualquier otro nombre) y **log(X2)** es la función logaritmo natural aplicada a la variable **X2**.
2) Introducir la función en el área del *COMAND WINDOW*. Aquí se genera la serie directamente en el área del *COMAND WINDOW* escribiendo: **genr LX=log(X)**.

Las series que se generaron fueron creadas con el nombre de **LY**, **LX2**, **LX3** y **LX4**.

- **Estimación de ecuaciones alternativas.**

Una vez generadas las series, se procede a la estimación de las ecuaciones propuestas; para las que se muestran dos opciones:

1) En el área del *COMAND WINDOW* escribir: **LS Y C X2 X3 X4**, donde **LS** es la opción del método de mínimos cuadrados ordinarios y **C** es el intercepto. Dar enter y se despliega el reporte de estimación en la caja llamada *Equation: UNTITLE*.
2) Sombrear las series a estimar (**Y**, **X2**, **X3** y **X4**), dar un clic y abrir en la opción *as Equation*. Si hay series no contiguas seleccionarlas presionando *control* de manera individual en cada serie. Se abre la caja de diálogo *Equation specification* en donde se escribe solamente **Y C X2 X3 X4**, sin **LS** pues lo abre por default (en esta caja se pueden elegir otras opciones diferentes, aparte del método de mínimos cuadrados), se da clic en *OK* y se despliega la caja con el reporte de la estimación.

Todo esto para el caso lineal. Antes de estimar las otras ecuaciones, es necesario salvar la ecuación (en este caso se salvó con el nombre de **lineal**), dado que se borra la

ecuación anterior cada que se estima una nueva, identificándose con el icono . Además, para varias aplicaciones es más laborioso volver a estimar cada ecuación y de esta manera solo se da doble clic en el archivo de trabajo en el nombre de la ecuación con que se desea trabajar.

La elección de variables para la estimación de los otros modelos propuestos es:

Y C LX2 LX3 LX4 para la ecuación **lin-log** y al presionar en la barra de herramientas *Procs/ Make Regressor Group* se salvó como grupo con el nombre **Glin-log**.

LY C X2 X3 X4 para la ecuación **log-lin**, se salvó como grupo con el nombre **Glog-lin**.

LY C LX2 LX3 LX4 para la ecuación **log-log** y se salvó como grupo con el nombre **Glog-log**.

- **Uso de residuales.**

Si se desea ver la tabla de los valores actuales, los residuales y los valores estimados, se da clic en la barra de herramientas del archivo de trabajo en la opción *VIEW/ Actual, Fitted, Residual/ Actual, Fitted, Residual Table*. Al elegir esta opción se nota que también se despliega la gráfica de esta tabla, pero para obtenerla sin los datos de la tabla y percibirla más claramente se hace clic en *VIEW/ Actual, Fitted, Residual/ Actual, Fitted, Residual Graph* o más directamente se obtiene esta gráfica al presionar *RESIDS* en la barra de herramientas de la ecuación.

Para salvar el vector de residuales de cada ecuación estimada se da clic en la barra de herramientas de la ecuación en la opción *PROCS/ Make Residuals Series* y aparece la caja de diálogo *Make residuals*, por default da el nombre de Resid01, Resid02, etc. En este caso se conservaron los nombres proporcionados por EVIEWS; es decir, **Resid01**, **Resid02**, **Resid03**, **Resid04**, que como se dijo, se identifican con el icono .

Una forma de salvar los valores estimados, $\hat{Y}$, de cada ecuación, es dando clic en la opción *Forecast* de la barra de herramientas de la ecuación y aparece la caja de diálogo con el mismo nombre; en el cuadro *Forecast name* aparece el nombre de la

variable dependiente y la letra **F** por default; es decir **YF**. En el trabajo se nombraron como **YF**, **YF2**, **LYF3** y **LYF4** a los valores estimados de las ecuaciones propuestas lineal, lin-log, log-lin y log-log. Al seleccionar en la misma caja *Output/ Do Graph* y presionar *OK*, aparece una gráfica de los valores estimados y más o menos dos errores estándar al 95% de confianza. Si se observa en el archivo de trabajo aparecen las series estimadas **YF**, **YF2**, etc. Se hace la observación que, dando clic en la selección Output/ Forecast: evaluation, aparece una tabla de resultados para evaluar el periódo de predicción, dado que el trabajo pretende encontrar un modelo explicativo no se utiliza esta tabla de resultados.

De esta manera, para graficar los residuales contra $\hat{Y}$, se puede elegir una de las siguientes alternativas:

1) Dar clic en el menú principal en *QUICK/ Graph* abriéndose la caja de diálogo Series list; para el caso lineal se pone **YF** y **Resid01** (es importante colocarlos en este orden), luego se despliega la caja *Graph* en la que se selecciona *Graph Type/ Scatter Diagraman*. Para que aparezca con puntos y sea más representativa la gráfica, seleccionar en *Show Options/ Line Graphs/ Symbols only*, presionar *OK* y se despliega la gráfica.
2) Seleccionar en el archivo de trabajo primero la serie de los valores estimados **YF** y después de los residuales **Resid01**, dar un clic derecho y elegir en el menú que aparece *Open/ as Group*, luego dar clic en la barra de herramientas en *View/ Graph/ Scatter/ Simple Scatter*, y aparece la gráfica de puntos como en el trabajo.

Si se quiere graficar únicamente los residuales contra el tiempo (sin los valores actuales y estimados), se presentan dos posibilidades:

1) Dar clic en la barra de herramientas de la ecuación en *VIEW/ Actual, Fitted, Residual/ Residual Graph*. Esta opción también gráfica los residuales uniéndolos con líneas; para que aparezca graficada con puntos se selecciona la gráfica con el mousse y dando un clic derecho, aparece un menú, se elige *Options* y se despliega la caja de diálogo *GRAPH OPTIONS*, se selecciona *Line Graphs/ Symbols only* para obtener solo los puntos.

2) También se podría seleccionar en el archivo de trabajo el nombre de los residuales que se desea graficar **Resid01,** dar doble clic (o un clic derecho y elegir en el menú que aparece *Open/ as Group*), luego dar clic en la barra de herramientas del grupo en *View/ Line Graph*. Para que aparezcan con puntos seleccionar la gráfica y proceder como en el inciso anterior. La desventaja de esta opción es que no aparecen los intervalos de confianza de los residuales como en la opción anterior.

Si se desea graficar los residuales en forma de Cuantiles-cuantiles normales por medio de:

- Una ecuación, seleccionar en el archivo de trabajo el nombre de los residuales (para el caso lineal **Resid01**) dando doble clic; en la barra de herramientas del grupo elegir la opción *View/ Distribution Graph/ Quantile-Quantile.../ Normal Distribution* y presionar *OK*.
- El grupo de variables de la ecuación, se selecciona el nombre del grupo (para el caso lineal **Glineal**) en el archivo de trabajo dando doble clic; luego en la barra de herramientas *View/ Múltiple Graphs/ Distribution Graphs/ Quantile-Quantile.../ Normal Distribution* y presionar *OK* para que se desplieguen las gráficas de todas las variables de la ecuación lineal.

Todas las gráficas se identifican con el icono .

- **Usos de variables.**

La obtención de la matriz de correlación para el modelo lineal se obtiene seleccionando en el archivo de trabajo las variables **X2**, **X3** y **X4**, luego con un clic derecho elegir *Open/ as Group*, y en la barra de herramientas presionar *View/ Correlations*. La diferencia en la matriz de correlación que se obtiene al abrir el grupo **Glineal,** es que incluye la correlación con la variable **Y**. Esta matriz se salvó como una tabla, por lo que el icono correspondiente será .

Para aplicar la prueba de variables redundantes en una ecuación como la lineal, se elige en la barra de herramientas la opción *View/ Coefficient Test/ Redundant Variables-Likelihood Ratio*, abriéndose la caja de diálogo *Omitted-Redundant Variable Test* allí se

escribe la o las variables que se piensa no son tan significativas, presionar *OK* y se despliega el reporte de regresión con el valor de la prueba *F* y la probabilidad de la o las variables seleccionadas.

La generación de la variable tendencial se creó directamente utilizando el COMMAND WINDOW, tecleando **t=@trend**, sin la palabra *genr*.

La obtención de los coeficientes de elasticidad precio e ingreso, se calculan en el modelo lin-log mediante las fórmulas $\beta_2\left(\frac{1}{Y}\right)$ y $\beta_3\left(\frac{1}{Y}\right)$ respectivamente. En EVIEWS aparecerá el resultado como una serie con el mismo valor. Para generar la serie de la elasticidad precio, se introduce en el *COMMAND WINDOW* como **genr ep=10661131.87/@mean(y)**, para la serie de elasticidad ingreso se utiliza la fórmula **genr ei=993369416.2/@mean(y),** apareciendo el nombre de estas series en el archivo de trabajo. Para ver el valor de cada serie, solo se da doble clic en el nombre del archivo de trabajo.

- **Otras operaciones complementarias.**

Aunque se pueden ir imprimiendo los resultados obtenidos, aquí fue necesario copiarlos y pegarlos para que quedaran insertados en el texto. El procedimiento fue:

- Para copiar una tabla, presionar *EDIT/ COPY* en el menú principal y aparece una caja de diálogo llamada *Copy* Precisión seleccionar *Formated- Copy.Numbers as they appear* y presionar *OK* y en el lugar que se desee colocar presionar *PEGAR* del menú *EDICIÓN.*
- Si se desea copiar una gráfica en un texto, dar *control* C y seleccionar en la caja de diálogo *Copy Graph of Metafile* que aparece, la opción *Copy to clipboard* y elegir si se copia a color o en blanco y negro, presionar *OK* y la opción PEGAR del menú *EDICIÓN.*

Algunas veces se copian con mejor presentación si se elige en la barra de herramientas la opción *Freeze* para "congelar" la gráfica. Esta opción es muy útil

también para hacer modificaciones en la apariencia de una gráfica o tabla como sería el tipo y tamaño de letra, agregar algún letrero, o eliminar alguna parte no necesaria. A estas gráficas o tablas congeladas también se les puede dar un nombre para salvarlas en el archivo de trabajo.

Cuando se tienen varias ventanas abiertas (de gráficas, ecuaciones, grupos, archivos de trabajo, etc.), es fácil localizar una ventana en particular presionando en el menú principal *WINDOW*; allí se indica el nombre de todas las ventanas abiertas. Para activar otra ventana, simplemente se selecciona con el mousse y se activa directamente.

Una vez terminada la sesión de trabajo, una forma rápida de cerrar todas las ventanas es presionando en el menú principal *WINDOW/ CLOSE ALL OBJET* y se cerrarán todos los objetos; si se desea que también se cierre el archivo de trabajo presionar *WINDOW/ CLOSE ALL*.

ANEXO C. RESUMEN DE RESULTADOS

C.1 ESTIMACIÓN DE ECUACIONES ALTERNATIVAS.

Dependent Variable: Y
Method: Least Squares
Date: 09/04/01 Time: 11:18
Sample: 1990 2000
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	64076536	14510330	4.415925	0.0031
X2	86574.37	25958.49	3.335108	0.0125
X3	-881536.3	192621.2	-4.576527	0.0026
X4	275.6195	43.04917	6.402436	0.0004

R-squared	0.871486	Mean dependent var	9165091.
Adjusted R-squared	0.816408	S.D. dependent var	831459.2
S.E. of regression	356260.4	Akaike info criterion	28.68000
Sum squared resid	8.88E+11	Schwarz criterion	28.82469
Log likelihood	-153.7400	F-statistic	15.82288
Durbin-Watson stat	2.222433	Prob(F-statistic)	0.001681

Tabla C.1.1. Ecuación lineal.

Dependent Variable: Y
Method: Least Squares
Date: 09/04/01 Time: 13:39
Sample: 1990 2000
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	100828174	47923689	2.103932	0.0734
LX2	10661132	2902142.	3.673538	0.0079
LX3	-93369416	19462480	-4.797406	0.0020
LX4	25369543	3827336.	6.628512	0.0003

R-squared	0.878655	Mean dependent var	9165091.
Adjusted R-squared	0.826650	S.D. dependent var	831459.2
S.E. of regression	346181.2	Akaike info criterion	28.62260
Sum squared resid	8.39E+11	Schwarz criterion	28.76729
Log likelihood	-153.4243	F-statistic	16.89553
Durbin-Watson stat	2.211838	Prob(F-statistic)	0.001379

Tabla C.1.2. Ecuación lin-log.

Dependent Variable: LY
Method: Least Squares
Date: 09/04/01 Time: 13:39
Sample: 1990 2000
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	22.38654	1.730220	12.93855	0.0000
X2	0.009495	0.003095	3.067667	0.0181
X3	-0.100806	0.022968	-4.388905	0.0032
X4	.0000311	5.13E-06	6.063692	0.0005

R-squared	0.858099	Mean dependent var	16.02698
Adjusted R-squared	0.797284	S.D. dependent var	0.094351
S.E. of regression	0.042481	Akaike info criterion	-3.204246
Sum squared resid	0.012632	Schwarz criterion	-3.059557
Log likelihood	21.62335	F-statistic	14.11005
Durbin-Watson stat	2.215940	Prob(F-statistic)	0.002363

Tabla C.1.3. Ecuación log-lin.

Dependent Variable: LY
Method: Least Squares
Date: 09/04/01 Time: 13:39
Sample: 1990 2000
Included observations: 11

Variable	Coefficient	Std. Error	t-Statistic	Prob.
C	27.11842	5.702650	4.755406	0.0021
LX2	1.173794	0.345339	3.398966	0.0115
LX3	-10.68025	2.315926	-4.611653	0.0025
LX4	2.867661	0.455431	6.296581	0.0004

R-squared	0.866567	Mean dependent var	16.02698
Adjusted R-squared	0.809382	S.D. dependent var	0.094351
S.E. of regression	0.041194	Akaike info criterion	-3.265779
Sum squared resid	0.011878	Schwarz criterion	-3.121090
Log likelihood	21.96179	F-statistic	15.15364
Durbin-Watson stat	2.195864	Prob(F-statistic)	0.001913

Tabla C.1.4. Ecuación log-log.

C.2. TABLA DE VALORES: ACTUAL, ESTIMADO Y RESIDUAL.

obs	Actual	Fitted	Residual
1990	8,996,194	8,845,650	150,544.
1991	7,298,407	7,644,648	-346,241.
1992	8,992,319	8,921,050	71,269.0
1993	9,518,527	9,193,693	324,834.
1994	8,818,320	9,128,670	-310,350.
1995	8,626,249	8,853,289	-227,040.
1996	9,013,277	8,564,025	449,252.
1997	9,410,764	9,184,186	226,578.
1998	9,638,158	9,878,289	-240,131.
1999	9,964,357	10,306,753	-342,396.
2000	10,539,434	10,295,753	243,681.

Tabla C.2.1 Ecuación lineal.

obs	Actual	Fitted	Residual
1990	8,996,194	8,835,520	160,674.
1991	7,298,407	7,618,916	-320,509.
1992	8,992,319	8,968,858	23,461.0
1993	9,518,527	9,213,808	304,719.
1994	8,818,320	9,082,506	-264,186.
1995	8,626,249	8,860,809	-234,560.
1996	9,013,277	8,567,782	445,495.
1997	9,410,764	9,190,855	219,909.
1998	9,638,158	9,883,874	-245,716.
1999	9,964,357	10,309,785	-345,428.
2000	10,539,434	10,283,294	256,140.

Tabla C.2.2 Ecuación lin-log.

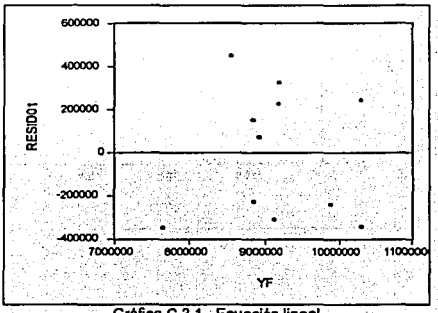
obs	Actual	Fitted	Residual
1990	16.0123	15.9895	0.02284
1991	15.8032	15.8536	-0.05040
1992	16.0119	16.0011	0.01074
1993	16.0688	16.0343	0.03444
1994	15.9923	16.0278	-0.03546
1995	15.9703	15.9927	-0.02236
1996	16.0142	15.9575	0.05667
1997	16.0574	16.0275	0.02984
1998	16.0812	16.1060	-0.02472
1999	16.1145	16.1533	-0.03876
2000	16.1706	16.1535	0.01717

Tabla C.2.3 Ecuación log-lin.

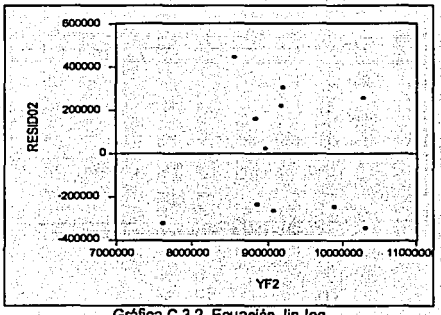
obs	Actual	Fitted	Residual
1990	16.0123	15.9884	0.02393
1991	15.8032	15.8505	-0.04733
1992	16.0119	16.0065	0.00541
1993	16.0688	16.0366	0.03215
1994	15.9923	16.0227	-0.03036
1995	15.9703	15.9935	-0.02315
1996	16.0142	15.9579	0.05634
1997	16.0574	16.0282	0.02912
1998	16.0812	16.1066	-0.02539
1999	16.1145	16.1537	-0.03920
2000	16.1706	16.1521	0.01849

Tabla C.2.4 Ecuación log-log.

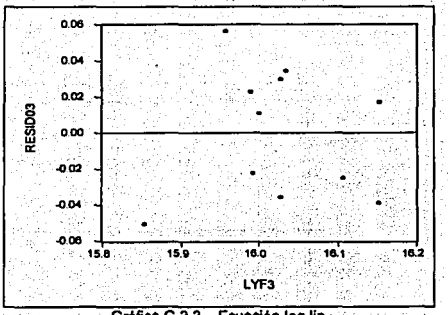
C.3 GRÁFICAS DE RESIDUALES CONTRA $\hat{Y}$.



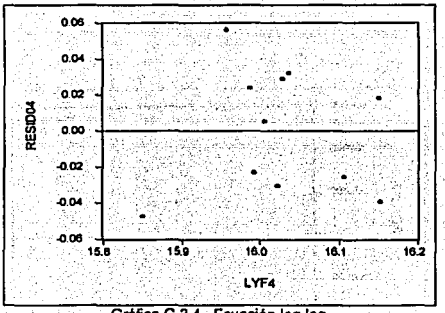
Gráfica C.3.1 Ecuación lineal.



Gráfica C.3.2 Ecuación lin-log.

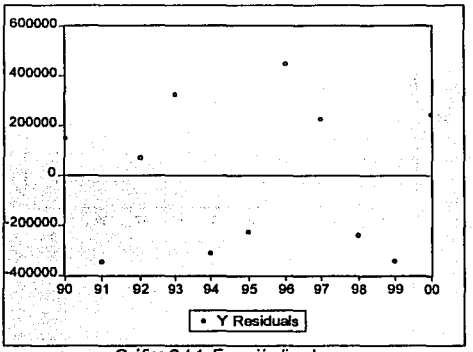


Gráfica C.3.3 Ecuación log-lin.

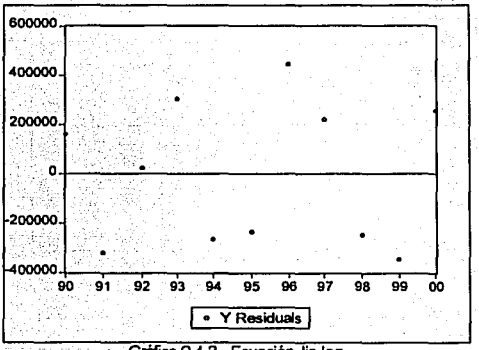


Gráfica C.3.4 Ecuación log-log.

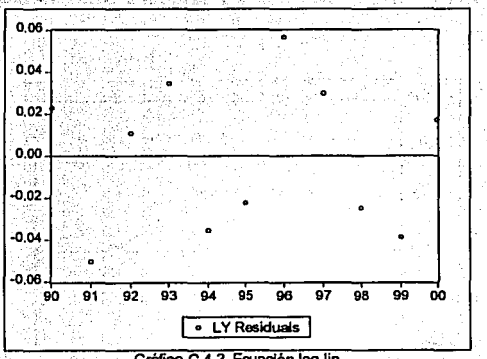
C.4 GRÁFICAS DE RESIDUALES CONTRA EL TIEMPO.



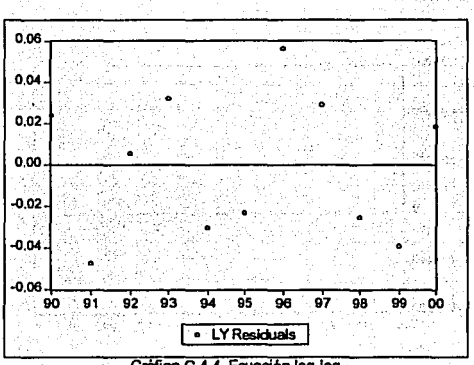
Gráfica C.4.1 Ecuación lineal.



Gráfica C.4.2 Ecuación lin-log.

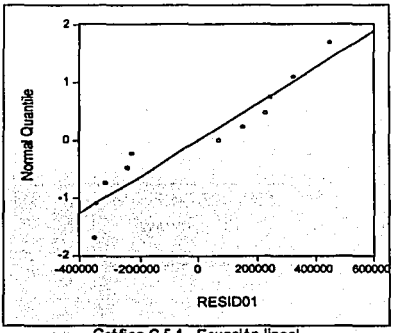


Gráfica C.4.3 Ecuación log-lin.

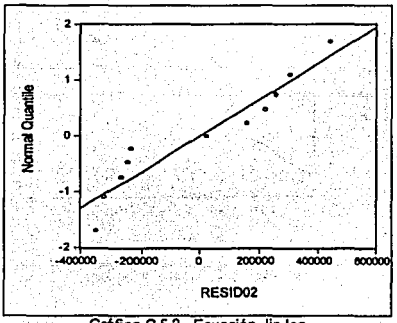


Gráfica C.4.4 Ecuación log-log.

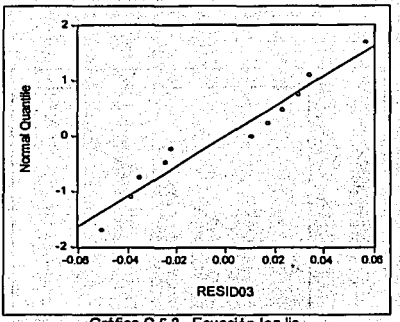
C.5. GRÁFICAS DE CUANTILES-CUANTILES NORMALES.



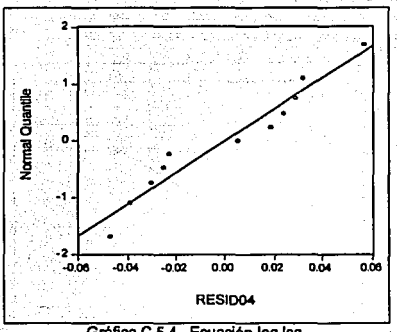
Gráfica C.5.1 Ecuación lineal.



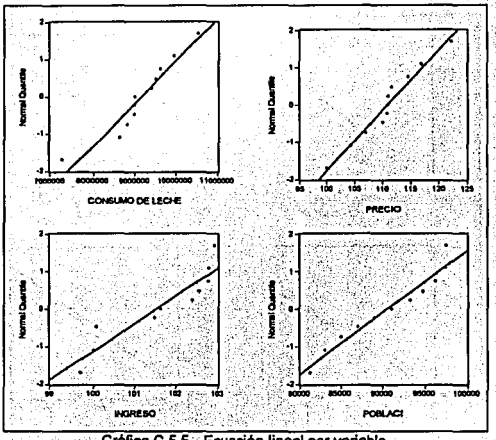
Gráfica C.5.2 Ecuación lin-log.



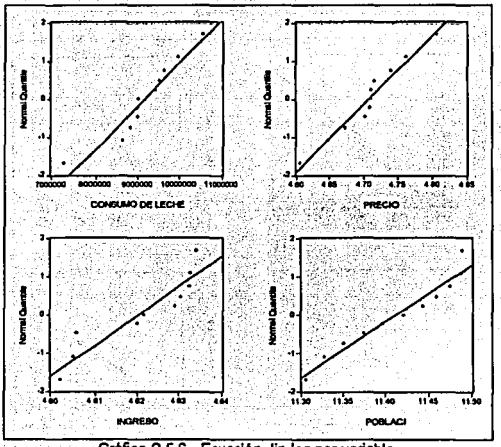
Gráfica C.5.3 Ecuación log-lin.



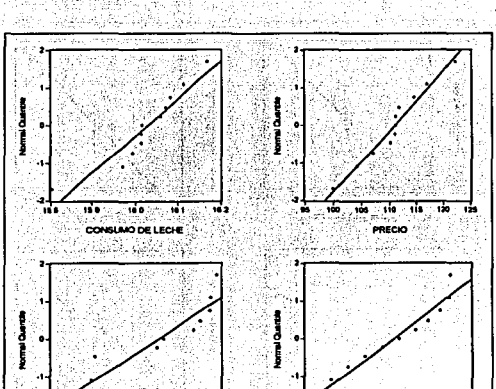
Gráfica C.5.4 Ecuación log-log.



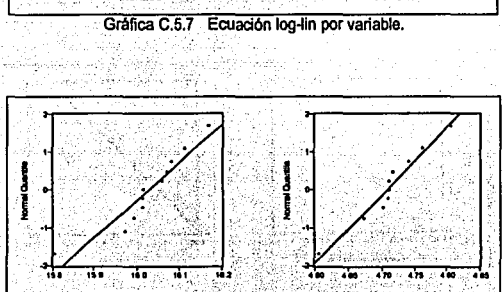
Gráfica C.5.5 Ecuación lineal por variable.



Gráfica C.5.6 Ecuación lin-log por variable.



Gráfica C.5.7 Ecuación log-lin por variable.



Gráfica C.5.8 Ecuación log-log por variable.

C.6. MATRIZ DE CORRELACIÓN.

	X2	X3	X4
X2	1.000000	0.200160	-0.242540
X3	0.200160	1.000000	0.794611
X4	-0.242540	0.794611	1.000000

Tabla C.6.1 Ecuación lineal.

	LX2	LX3	LX4
LX2	1.000000	0.225078	-0.242160
LX3	0.225078	1.000000	0.788604
LX4	-0.242160	0.788604	1.000000

Tabla C.6.2 Ecuación lin-log.

	X2	X3	X4
X2	1.000000	0.200160	-0.242540
X3	0.200160	1.000000	0.794611
X4	-0.242540	0.794611	1.000000

Tabla C.6.3 Ecuación log-lin.

	LX2	LX3	LX4
LX2	1.000000	0.225078	-0.242160
LX3	0.225078	1.000000	0.788604
LX4	-0.242160	0.788604	1.000000

Tabla C.6.4 Ecuación log-log.

C.7 DATOS DE LA VARIABLE TENDENCIAL.

	Modified: 1990 2000 // t=@trend
1990	0.000000
1991	1.000000
1992	2.000000
1993	3.000000
1994	4.000000
1995	5.000000
1996	6.000000
1997	7.000000
1998	8.000000
1999	9.000000
2000	10.00000

Tabla C.7.1.

C.8 DATOS TRANSFORMADOS PARA LA ECUACIÓN LIN-LOG.

obs	Y	LX2	LX3	LX4
1990	8,996,194.	4.806722	4.605870	11.30527
1991	7,298,407.	4.761832	4.619960	11.32804
1992	8,992,319.	4.710431	4.605870	11.35099
1993	9,518,527.	4.647080	4.602266	11.37401
1994	8,818,320.	4.605170	4.605170	11.39713
1995	8,626,249.	4.672455	4.621536	11.42035
1996	9,013,277.	4.701480	4.633952	11.44230
1997	9,410,764.	4.708719	4.632591	11.45881
1998	9,638,158.	4.716175	4.630350	11.47475
1999	9,964,357.	4.741361	4.632396	11.48848
2000	1,0539,434	4.710629	4.628887	11.48743

Tabla C.8.1.

BIBLIOGRAFÍA

REFERENCIA DE TEXTOS

- ALCAIDE INCHAUSTI, ANGEL; ÁLVAREZ VÁZQUEZ, NELSON: " *Econometría: Modelos Deterministicos y Estocásticos. Teoría* " Centro de Estudios Ramón Areces, Madrid, 1992.
- CANAVOS, GEORGE C.: "*Probabilidad y Estadística: Aplicaciones y Métodos*", McGraw-Hill, México, 1988.
- DAGUM, CAMILO: "*Introducción a la Econometría*". Editorial Siglo XXI, México, 1980.
- ESCOBAR ITURBE, MA. CRISTINA: "*Modelos y Economía Matemática*", Serie de investigación 9, Departamento de Economía, División de Ciencias Sociales y Humanidades, Unidad Iztapalapa, UAM, México,1993.
- GUISÁN, MA. DEL CARMEN: "*Econometría*", McGraw-Hill / Interamericana, Madrid, 1997.
- GÓMEZ, MA. DEL SOCORRO: Hernández, Silvia C.: "*Introducción a la Economía: Un enfoque Aplicado*", McGraw-Hill, México, 1995.
- GUJARATI, DAMODAR: "*Econometría*". McGraw-Hill, Colombia, 1990.
- HAIR JR., JOSEPH F. : ANDERSON, ROLP E.; TATHAM, RONALD L.; BLACK, WILLIAM C.: "*Análisis Multivariante*". Prentice Hall, Madrid, 1999.
- HERNÁNDEZ ALONSO, JOSÉ: "Introducción a la *Econometría*". Editorial ESIC, Madrid, 1995.
- HERNÁNDEZ ALONSO, JOSÉ: "*Ejercicios de Econometría*". Editorial ESIC, Madrid, 1992.
- HERNÁNDEZ TINAJERO, ALEJANDRO: "*Innovación Tecnológica en la Producción de Leche como una alternativa en la Seguridad Alimentaria: El caso de los Altos de Jalisco*", Tesis, Facultad de Economía, UNAM, México, 1996.
- INTRILLIGATOR, MICHAEL D.: "*Modelos Econométricos, Técnicas y Aplicaciones*". Fondo de Cultura Económica, México, 1990.
- KAZMIER, LEONARD; DIAZ MATA, ALFREDO: "*Estadística Aplicada a la Administración y Economía*" Serie Shaum's, McGraw-Hill, México, 1991.

- LAVASTIDA LÓPEZ, NAPOLEÓN: "*Estadística I*", Instituto Politécnico Nacional, México, 1991.
- LÓPEZ CASUSO, RAFAEL: "*Cálculo de Probabilidades e Inferencia Estadística: con tópicos de Econometría*", Universidad Católica Andrés Bello, Caracas, 1996.
- LOREDO PÉREZ, FELIPE: "*El Mercadeo de la Leche en México (1979-1984)*", Tesis, Facultad de Economía, UNAM, México, 1986.
- MADDALA, G. S.: "*Introducción a la Econometría*", Prentice-Hall Hispanoamericana, México, 1996.
- MENDENHALL, WILLIAM; WACKERLY, DENNIS D.; SCHEAFFER, RICHARD L.: "*Estadística Matemática con Aplicaciones*", Grupo Editorial Iberoamérica, México, 1994.
- MENDENHALL, WILLIAM; REINMUTH, JAMES E.: "*Estadística para Administración y Economía*", Grupo Editorial Iberoamérica, México, 1981.
- NETER, JOHN; WASSERMAN, WILLIAM; KUTNER, MICHAEL H.: "*Applied Linear Regresión Models*", Irwin, Boston, 1989.
- NETER, JOHN; WHITMORE, G.A.: "*Fundamento de Estadística aplicada a los Negocios*", Compañía Editorial Continental, México, 1978.
- NEWBOL, PAUL; "*Estadística para Negocios y la Economía*", Prentice may, España, 1997.
- NOVALES CINEA, ALFONSO: "*Econometría*", McGraw-Hill, Madrid, 1993.
- PAPPAS, J.L; BRIGHAM, E.F.: "*Fundamentos de Economía y Administración*", McGraw-Hill, México, 1992.
- PETERSON, WILLIS L.: "*Principios de Economía Micro*", CECSA, México, 1988.
- PETERSON, WILLIS L.: "*Principios de Economía Macro*", CECSA, México, 1988.
- PINEDA, OCTAVIO L.: "*Métodos y Modelos Econométricos: Una Introducción*" Colección de textos politécnicos. Serie económica, LIMUSA, México, 1998.
- PULIDO, ANTONIO: "*Modelos Econométricos*". Ediciones Pirámide, Madrid, 1993.
- RODAS CARPIZO, A.: "*Economía Básica*", LIMUSA, México, 1990.
- SALAS, JAVIER: "*Econometría Aplicada a los Países en Desarrollo: El caso Mexicano*", Fondo de Cultura Económica, México, 1990.
- SALVATORE, DOMINICK: "*Econometría*", SHAUMS, McGraw-Hill, México, 1991.

- TEH-WEI, HU: "*Econometría: Un Análisis Introductorio*", Fondo de Cultura Económica, México, 1979.
- URIEL, EZEQUIEL; CONTRERAS, DULCE; MOLTÓ, MA. LUISA; PEINÓ, AMADO: " *Econometría: El Modelo Lineal* ", AC, Madrid, 1992.
- VARGAS SANCHEZ, MTRO. GUSTAVO: "*Aplicaciones Econométricas con Econometric View*" , Facultad de Economía, Centro de Educación Continua, UNAM, 1996.
- WALPOLE, RONALD E.: "*Probabilidad y Estadística*". McGraw-Hill, México, 1993.

REFERENCIA HEMEROGRÁFICA

- "*Boletín Mensual de Leche*", Vol. IV-VII, Centro de Estadística Agropecuaria SAGAR, 1996-1999.
- "*Boletín Bimestral de Leche*", Vol. VI-VII, Centro de Estadística Agropecuaria SAGAR, 2000.
- "*Boletín Bimestral de Leche*", Vol. VII, Centro de Estadística Agropecuaria SAGARPA, 2000.
- BOURGES RODRÍGUEZ, HÉCTOR; MORALES DE LEÓN, JOSEFINA: "*La leche y sus Derivados en la Dieta*", Instituto Nacional de Nutrición Salvador Zubirán. En: reproducción del Suplemento de Cuadernos de Nutrición No. 4, julio-agosto 1996.
- "*Claridades Agropecuarias*", no. 33, ASERCA, México, mayo 1996.
- "*Claridades Agropecuarias*", no. 77, ASERCA, México, enero 2000.
- García Bojalil, Dr. Carlos; Lastra Marín, Ing. Ignacio de J.; Peralta Arias, Lic. Ma. de los Ángeles; Olivera Cázares, MVZ Enrique; Ortega Marquéz, Lic. Armando; Pérez Frías, Lic. Humberto; Segura Martínez, Ing. César; Treviño Rodríguez, Lic. Florencio; Velásquez Valencia, MVZ Ma. Teresa.:"*Situación Actual y Perspectivas de la Producción de Leche de Ganado Bovino en México 1990-2000*", SAGAR, octubre de 1999.
- "*Leche Boletín Mensual*", Vol. II-III, SARH-INEGI, 1993-1995.
- "*Leche Boletín Mensual*", Vol. IV, Centro de Estadística Agropecuaria CEA, SAGAR-INEGI, 1995.