Econometría

ELSA NORMA ELIZALDE ANGELES

Red Tercer Milenio

ECONOMETRÍA

ECONOMETRÍA

ELSA NORMA ELIZALDE ANGELES

RED TERCER MILENIO

AVISO LEGAL

Datos para catalogación bibliográfica

Elsa Norma Elizalde Ángeles

Econometría

ISBN 978-607-733-100-1

Primera edición: 2012

Revisión pedagógica: Dení Stincer Gómez

Revisión editorial: Eduardo Durán Valdivieso

DIRECTORIO

ÍNDICE

Introducción 4

Objetivo de aprendizaje general 5

Mapa conceptual 7

Unidad 1. Introducción Metodológica 8

Mapa conceptual 9

Introducción 10

1.1 Datos 11

1.2 Relaciones 15

1.3 Variables 16

1.4 ¿Qué es la econometría? 19

Autoevaluación 24

Unidad 2. Modelo de regresión lineal clásico 27

Mapa conceptual 28

Introducción 29

2.1 Mínimos cuadrados ordinarios (MCO) 30

2.2 Supuestos 38

2.3 Estimadores 42

2.4 Pruebas de hipótesis 60

2.5 Predicción 73

Autoevaluación 78

Unidad 3. Heterocedasticidad 80

Mapa conceptual 81

Introducción 82

3.1 Causas de la heterocedasticidad 83

3.2 Estimación de MCO con heterocedasticidad 86

3.3 Métodos de correlación 91

Autoevaluación 96
Unidad 4. Autocorrelación 98
Mapa conceptual 99
Introducción 100
4.1 Causas 101
4.2 Estimación de MCO con autocorrelación 104
4.3 Métodos de correlación 107
Autoevaluación 112

Unidad 5. Variables artificiales o cualitativas 114
Mapa conceptual 115
Introducción 116
5.1 Variables cualitativas 117
5.2 Aplicación de las variables cualitativas 119
Autoevaluación 130

Unidad 6. Series temporales 133
Mapa conceptual 134
Introducción 135
6.1 Modelo de regresión con series de tiempo 136
6.2 Estimación 140
6.3 Predicción 144
Autoevaluación 149

Bibliografía 152
Glosario 153

INTRODUCCIÓN

El concepto econometría surge en la segunda década del siglo XX, para hacer referencia a estudios económicos con apoyo de métodos estadísticos. Es con la econometría moderna que se proponen formulaciones que sirven de apoyo para contradecir los planteamientos hechos por la teoría económica.

La econometría es considerada una disciplina, la cual se mueve entre dos teorías: la teoría económica y la teoría estadística. De ahí que se dé en cierta forma un grado de complejidad al tratar de analizar una diversidad de posiciones en torno a la teoría económica, y demostrar sus planteamientos; en tanto, en la teoría estadística se encuentra una gran variedad de técnicas y métodos, en los cuales se pueden encontrar limitaciones.

El presente libro de econometría cuenta con seis unidades en donde se introducen una serie de aplicaciones y formulaciones para determinar cuál es el comportamiento que muestra un modelo planteado.

Una advertencia que se hace es que a partir de la Unidad tres y hasta la seis, se utilice paquetería especial para la solución de modelos econométricos, debido al alto grado de complejidad de los planteamientos y formulaciones establecidas. En este caso, el más recomendable es *E views*, debido a que es uno de los más completos, existen otros como *Shazam*, que también puede ayudar a dar solución al análisis de todas las técnicas vistas en el presente libro.

Este libro es un curso introductorio de econometría dirigido a estudiantes de comercio internacional, el cual les proporciona las herramientas necesarias para el análisis de modelos de regresión.

Al finalizar el estudio de la econometría, el estudiante podrá aplicar métodos econométricos que le permitan evaluar alguna teoría económica o bien para poner en práctica un proyecto comercial. De igual modo, puede hacer pronósticos de variables macroeconómicas o variables relacionadas con el comercio internacional. Es importante mencionar que la econometría es una herramienta fundamental que el estudiante podrá utilizar en su vida profesional.

OBJETIVO DE APRENDIZAJE GENERAL

El objetivo general de aprendizaje del presente libro es proporcionarle al estudiante los elementos necesarios aplicables en la econometría. Que pueda plantear un modelo econométrico y le permita distinguir las variables que se aplican a este.

El libro incluye gráficas, tablas y formulaciones que le permitirán al estudiante establecer una distinción entre un modelo de regresión poblacional y un modelo de regresión muestral. De esta manera, distinguirá el enfoque de mínimos cuadrados ordinarios, determinará la estimación de éstos, y aplicará pruebas de hipótesis a partir de pruebas estadísticas.

El mayor peso que se tiene para el estudio de la econometría se encuentra concentrado en la Unidad dos, en donde se hace una introducción al estudio de mínimos cuadrados ordinarios, y a partir de éste se establecen una serie de supuestos para el modelo de análisis de regresión simple y de ahí una serie de formulaciones que van a permitir llevar a cabo la estimación del modelo. Una vez que se efectúan estas pruebas, se pueden establecer pruebas de significancia para verificar la veracidad o falsedad de la hipótesis.

En las unidades tres y cuatro se analizan dos de las violaciones que se pueden presentar en el modelo clásico, que son la heterocedasticidad y la autocorrelación, determinando cuáles son las causas de su presencia y su corrección.

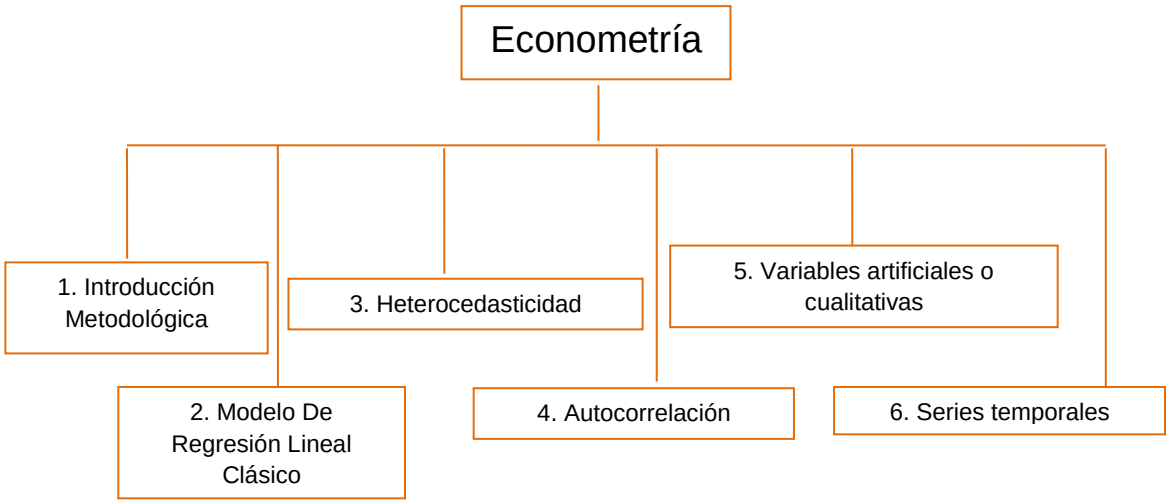
En la Unidad cinco se analiza cómo en un modelo de regresión lineal se pueden aplicar variables cualitativas o artificiales, y convertirlas en valores cuantificables. Y de esta manera verificar qué tendencia pueden mostrar en el modelo.

Por último, se analiza un modelo de regresión con series de tiempo, el cual debe presentar *estacionariedad*, y que de esta manera arroje resultados que cumplan con los supuestos establecidos en el modelo de mínimos cuadrados ordinarios.

Cada sesión cuenta con actividades de aprendizaje y al final de cada Unidad, con una autoevaluación, las cuales le permitirán al estudiante cumplir con los objetivos establecidos en cada Unidad.

En el libro se proporciona, también, un glosario en el cual se incluyen términos que puede consultar el estudiante para facilitar el entendimiento y comprensión de la lectura. Asimismo, se incluye la bibliografía que permitió el desarrollo del presente libro.

MAPA CONCEPTUAL



UNIDAD 1

INTRODUCCIÓN METODOLÓGICA

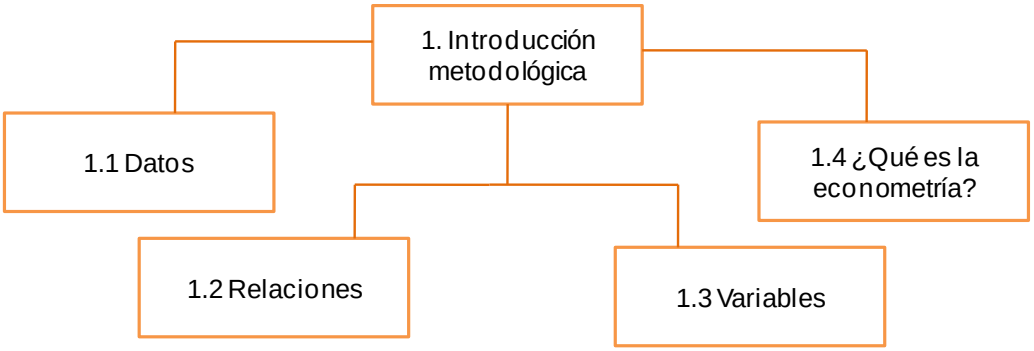
OBJETIVO

El estudiante identificará los tipos de datos que se observan en la economía para llevar a cabo un análisis empírico y distinguirá las diferentes relaciones hay al establecer un modelo. Diferenciará las variables que se aplican en un modelo econométrico y analizará la conformación de la econometría con otras ciencias.

TEMARIO

1.1 DATOS
1.2 RELACIONES
1.3 VARIABLES
1.4 ¿QUÉ ES LA ECONOMETRÍA?

MAPA CONCEPTUAL



INTRODUCCIÓN

Para la elaboración de un modelo econométrico es necesaria la aplicación de métodos estadísticos a datos económicos, de igual forma se debe tener conocimiento de la teoría económica.

De los métodos estadísticos y la teoría económica ya se tiene noción de cursos previos.

En la primera parte de de esta Unidad se determina cuáles son la fuentes de información para obtener datos estadísticos acerca del comportamiento de la economía y aplicar estos datos a un modelo econométrico planteado. Por lo que se establecen los diferentes tipos de datos que se observan en la economía.

La econometría aborda el problema de elaborar modelos que midan las relaciones causales entre variables económicas, por lo que se verifican este tipo de relaciones. Asimismo se analiza que en todo modelo econométrico se utilizan dos tipos de variables, mismas que se ocupan para representar una relación de causalidad entre dos variables.

Por último, se menciona una serie de definiciones acerca de qué es la econometría, haciendo referencia a algunos autores. También se indica cómo a partir de la econometría se puede comprobar el grado de validez de los modelos económicos y se logra usar para explicar el comportamiento de la economía.

1.1 DATOS

El análisis econométrico consiste en la aplicación de métodos estadísticos a datos económicos. Uno de los problemas con los cuales se puede enfrentar el econometrista es la escasa calidad de los datos. Esto se refleja en una disociación, en algunos casos, entre la información disponible y la requerida para comprobar la validez de los modelos teóricos. Esta separación se debe a que ambas actividades las realizan diferentes personas.

Las fuentes de información para obtener datos económicos corresponden a empresas u oficinas de estadística gubernamentales, donde la recolección de información está a cargo, en la mayoría de las veces, de personas que no son especialistas, teniendo con esto imprecisiones en tal información.

Para el caso de México, la recolección de información corresponde al Instituto Nacional de Geografía e Informática (INEGI) y al Banco de México (Banxico), la Secretaria de Trabajo, entre otras entidades, mostrando con esto los datos un alto grado de heterogeneidad. Frente a esto, en la actualidad se observa un flujo creciente de información estadística.

Los datos recolectados por las diversas entidades son de tipo no experimental, lo que implica que no están sujetos al control del investigador, bajo estas circunstancias, lo que se debe hacer es tratar de obtener la mayor información posible de datos imperfectos y reconocer que los resultados de los análisis dependen de los datos, incurriendo con esto en posibles errores de observación, por omisión o por comisión.

Un caso particular son las encuestas que realiza el INEGI, cuando se efectúa el censo poblacional se presentan situaciones en las que los cuestionarios llegan a no tener respuesta, o bien, que sólo se cuente con 50% de las respuestas a las preguntas, sobre todo, que no sean contestadas las de tipo financiero, ocasionando con ello un sesgo en los resultados.

Los datos económicos, por lo general, están disponibles con un nivel de agregación muy alto, como es el caso de las variables macroeconómicas, como el PIB (Producto Interno Bruto), el desempleo, la inflación, etcétera, donde este

nivel de agregación no permite realizar estudios a unidades individuales que podrían construirse en el objetivo fundamental de estudio.

Ante estos problemas, el econometrista debe tener presente que los resultados de la investigación son tan buenos como sea la calidad de la información con la cual trabaje. En ese sentido, es necesario que el investigador ajuste diferentes modelos y seleccione el que más satisfacción le dé.

Los datos que se observan en la economía para realizar el análisis empírico, son de tres tipos, a saber:

1) *Series de corte transversal o de sección cruzada* son las que se recolectan sobre unidades individuales en un momento del tiempo. Por ejemplo, los censos poblacionales que realiza el INEGI cada 10 años, las encuestas que se efectúan, arrojan datos sobre salarios, gastos del consumidor, la participación de la fuerza de trabajo en las diferentes actividades económicas del país, el nivel de educación, etcétera. Cada observación es un nuevo individuo, hogar, empresa, ciudad, en un momento dado del tiempo. Un ejemplo hipotético es el que se muestra en la siguiente tabla, en donde se tiene una muestra de 10 encuestas realizadas a entes individuales:

Observación	Sueldo	Educación	Sexo	Edo. Civil
1	2.10	11	1	0
2	3.00	13	0	1
3	4.08	8	0	0
4	4.24	12	1	1
5	5.07	12	0	1
6	6.03	14	1	0
7	5.07	13	0	0
8	6.03	11	0	1
9	11.12	16	0	1
10	9.08	14	1	0

2) *Series de temporales* es información que se recopila durante un determinado tiempo. Estas series son recolectadas en intervalos muy cortos con movimientos ascendentes y descendentes, simultáneamente. Pueden ser datos anuales, semestrales, trimestrales, mensuales, semanales o diarios. Por ejemplo, estos datos pueden ser las cotizaciones diarias en el mercado secundario de dinero, el índice de inflación mensual, el PNB (Producto Interno Bruto) anual, entre otros. Dada su naturaleza, pueden presentarse problemas al tratar de inferir causa y efecto. Para las series de tiempo se utilizan frecuentemente los números índice que se elaboran tomando el valor adoptado en un año determinado como base y mostrando los siguientes en relación con éste. A continuación se da un ejemplo del comportamiento del PIB en periodos trimestrales del 2008 al 2010.

Producto Interno Bruto México 2011 (base 2003=100)	
2008 (TRIMESTRAL)	**%**
I	2.1
II	2.5
III	1.3
IV	-1.0
2009 (TRIMESTRAL)	**%**
I	-7.4
II	-9.6
III	-5.5
IV	-2.0
2010 (TRIMESTRAL)	**%**
I	4.1
II	7.6
III	5.1
IV	4.4

Fuente: INEGI. Sistema de Cuentas Nacionales de México.

3) *Series longitudinales o de panel* constituye un tipo de datos combinados, consta de una serie temporal para cada miembro del corte transversal en el conjunto de datos, a los cuales se les denomina datos de micropanel. A continuación se da un ejemplo de este tipo de series, incluye información de una muestra de cinco entidades federativas de México: Distrito Federal, Estado de México, Nuevo León, Jalisco y Veracruz.

Observación	Ciudad	Años	Población	Vivienda Propia	Educación
1	1	2005	8,605,239	71.1	9.4
2	1	2010	8,851,080	66.7	10.5
3	2	2005	13,096,686	79.0	7.9
4	2	2010	15,175,862	73.6	9.1
5	3	2005	3,834,141	80.7	8.5
6	3	2010	4,653,458	79.6	9.8
7	4	2005	6,322,002	69.3	7.4
8	4	2010	7,350,682	65.4	8.8
9	5	2005	6,908,975	79.9	6.4
10	5	2010	7,963,194	80.8	7.9

Fuente: INEGI. Sistema de Cuentas Nacionales de México

De esta manera, para llevar a cabo el análisis empírico, se hace uso de datos económicos para contrastar una teoría o estimar una relación. La búsqueda de datos que pueden ser temporales o transversales, dependerá del modelo teórico que se haya planteado. Las fuentes de información disponibles son muy amplias, de organismos gubernamentales o privados, al consultarse con accesibilidad en la web.

ACTIVIDAD DE APRENDIZAJE

Elaborar una tabla de serie longitudinal con 30 observaciones, con dos periodos de tiempo y 4 variables (cuantitativas y cualitativas). Utilizar datos estadísticos de INEGI y Banxico. Entregar a computadora en la siguiente sesión.

1.2 RELACIONES

En econometría se tiene que especificar el modelo matemático con el cual se va a trabajar, una vez que se ha determinado bajo qué teoría económica se va a llevar a cabo el análisis empírico.

Un modelo es simplemente un conjunto de ecuaciones matemáticas. La teoría económica postula una serie de relaciones causales entre diversas magnitudes económicas. La econometría aborda el problema de elaborar modelos que midan las relaciones causales entre variables económicas. Estas relaciones son de tres tipos:

1) Las uniecuacionales constan de una sola ecuación en la que hay una variable dependiente (o determinada) que viene establecida por una o más independientes (o determinantes) o explicativa. Por ejemplo cuando se dice que el consumo (C) depende del nivel de precios (P) y del ingreso disponible (Yd) se expresa como: C = f (P, Yd). C es la variable dependiente, mientras que P y Yd son las variables independientes. Cualquier alteración en los niveles de P y Yd determinarán las variaciones en el consumo (C).

2) Las multiecuacionales parten de un conjunto de ecuaciones. Por ejemplo, si se considera el consumo nuevamente, el gasto que se efectúa para realizarlo es en bienes de consumo inmediato, bienes de uso duradero y en servicios. Cada uno de ellos podrían ser una función del ingreso y la riqueza. De esta manera, se tiene un sistema de ecuaciones C_i, C_d y C_s que están en función del ingreso y de la riqueza. Este conjunto de ecuaciones se pueden tratar separadamente como relaciones uniecuacionales o de manera conjunta.

3) Las simultáneas son cuando dos o más variables vienen determinadas simultáneamente por un cierto número de variables explicativas. Por ejemplo, si se considera el ingreso (Y) y el consumo (C) de la totalidad del mercado, se debe tener en cuenta que los precios y las cantidades vienen determinados simultáneamente por las condiciones de oferta y

demanda, y por otras variables. Tal es el caso del siguiente sistema de ecuaciones:

$$Q = f(P, Y) \quad \text{relación de demanda.}$$

$$Q = g(P, Z) \quad \text{relación de oferta.}$$

Estas dos ecuaciones determinan P y Q, dadas las variables explicativas Y y Z. Donde Y es el ingreso y Z las condiciones climatológicas. Las relaciones de las ecuaciones simultáneas son también multiecuacionales. La diferencia entre estas dos relaciones es la forma en que las variables están interrelacionadas.

ACTIVIDAD DE APRENDIZAJE

Establecer un modelo multiecuacional e indicar cuál es la variable dependiente y cuáles las independientes. Sustentar el modelo bajo una teoría económica. Entregar a computadora en la siguiente sesión.

1.3 VARIABLES

La terminología utilizada en econometría es la variable que se entiende como el concepto económico que se quiere analizar. Normalmente se utilizan variables cuantitativas, es decir, cuyos valores vienen expresados de forma numérica; por ejemplo, como los niveles de precios, el ingreso nacional. También existe la posibilidad de incluir en el modelo econométrico variables cualitativas que se puedan determinar de manera cuantitativa (por ejemplo, hombre, mujer, casado o soltero).

Todo modelo econométrico hace uso de variables dependientes y variables independientes, y se utilizan para representar una relación de causalidad entre dos variables, mismas que reciben la siguiente terminología:

Variable dependiente	Variable independiente
Variable explicada	Variable explicativa
Variable de respuesta	Variable de control o estímulo
Predicha	Predictor
Regresada	Regresor
Variable endógena	Variable exógena

Una variable dependiente o endógena es aquélla que se determina dentro del sistema económico, y una variable independiente o exógena está dada desde fuera del sistema.

Cuando se estudia la dependencia de una variable en una sola variable explicativa, tal como la dependencia entre el consumo y el ingreso disponible, tal estudio se conoce como análisis de regresión simple o en dos variables. Si se estudia la dependencia de una variable en más de una variable explicativa, por ejemplo, la demanda de gasolina por las familias, es la variable endógena y el precio y el ingreso, las variables exógenas; se hace referencia al análisis de regresión múltiple. De esta manera, se puede decir que en el análisis de regresión en dos variables existe sólo una variable explicativa, y cuando se trabaja con más de dos de estas variables, se está frente al análisis de regresión múltiple.

Las variables endógenas pueden además clasificarse como: variables objetivo, que son las que se desean influenciar (niveles de precio y empleo), y variables no objetivo son aquéllas por las que no se está interesado. Las variables exógenas se clasifican como variables instrumento, es decir, que son variables que pueden ser manipuladas específicamente para alcanzar algunos objetivos.

El comportamiento de la variable dependiente se podría predecir sobre la base del comportamiento de la variable independiente. El problema que puede surgir es la determinación de causalidad, es decir, cuáles son las variables dependientes y cuáles las independientes. En el análisis empírico se puede decir si dos variables están relacionadas, pero no decir si de hecho existe una

relación de dependencia y cuál es la dirección de esa relación, lo que hace necesario una teoría que dé plausibilidad a una relación empírica.

ACTIVIDAD DE APRENDIZAJE

En la siguiente tabla se presentan cifras de tasas de crecimiento anual de desocupación abierta en áreas urbanas (porcentaje con respecto a la Población Económicamente Activa, PEA) en México durante el periodo de enero a diciembre del 2004.

Tasas de crecimiento anual de desocupación abierta
en áreas urbanas, enero a diciembre del 2004.[1]

Mes	Ciudad de México	Guadalajara	Monterrey
Enero	31.43	73.91	41.18
Febrero	34.29	56.52	20.59
Marzo	42.86	4.35	20.59
Abril	25.71	17.39	11.76
Mayo	20.00	52.17	5.88
Junio	40.00	4.35	17.65
Julio	28.57	52.17	14.71
Agosto	65.71	56.52	32.35
Septiembre	51.43	43.48	50.59
Octubre	17.14	91.30	17.65
Noviembre	48.57	13.04	14.71
diciembre	2.86	0.00	32.35

Graficar la tasa de desocupación de cada entidad. Utilizar el eje horizontal para tiempo y el eje vertical para la tasa de desocupación y dar una explicación de que entidad tiene la tasa de desocupación más variable.
Entregar en la siguiente sesión, sugerencia elaborarlo en Excel.

[1] Fuente: Instituto Nacional de Estadística, Geografía e Informática: las cifras desestacionalizadas y de tendencia corresponden a procesos elaborados por el Banco de México, Encuesta Nacional de Empleo Urbano.

1.4 ¿QUÉ ES LA ECONOMETRÍA?

La econometría tiene su origen a principios de los años de 1930, teniendo como objetivo medir los ciclos de los negocios debido a la frecuente presencia de las fases recesivas observadas desde finales del siglo XIX. La econometría considerada “medición de la economía” se ha movido entre la teoría, las matemáticas y la estadística.

El término econometría surgió en 1926 por Ragnar Frisch, economista y estadístico de origen noruego, quien realizó una serie de trabajos en los que aplicó la econometría, series de tiempo y análisis de regresión lineal. Son varios los autores que surgieron a partir de la segunda década del siglo XX y que aplicaron la econometría en sus trabajos de investigación, entre estos autores es conveniente mencionar al economista estadounidense Lawrence Klein, considerado padre de la econometría moderna, a quien se le otorgó el Premio Nobel en 1980 por la creación de modelos econométricos y la aplicación al análisis de las fluctuaciones económicas y políticas económicas.

Así, el método de investigación econométrica se inscribe en la conjugación de la teoría económica y las mediciones verdaderas usando la teoría y la técnica de la inferencia estadística como puente. La inferencia estadística se ha introducido para juzgar la relación entre la teoría económica y comparar esta teoría con algunas mediciones.

La teoría económica se ocupa fundamentalmente de las relaciones entre variables. Variables como la oferta, la demanda, costos, producción, etcétera. La teoría económica se considera como una colección de relaciones entre las variables. De esta manera, es la econometría la encargada de diferenciar estas proposiciones teóricas incorporadas en estas relaciones y de la estimación de los parámetros que aparecen en ellas.

Lo que es cierto es que esta disciplina se le ha definido de diferentes maneras:

> La econometría, es el resultado de la adopción de una posición sobre el papel que juega la economía, consiste en la aplicación de la estadística

matemática a los datos económicos con el objeto de proporcionar no sólo un apoyo empírico, a los modelos construidos por la economía matemática sino una forma de obtener resultados numéricos.[2]

Se puede definir a la econometría como el análisis cuantitativo de fenómenos económicos reales basados en el desarrollo simultáneo de la observación y la teoría, relacionados a través de apropiados métodos de inferencia.[3]

La econometría puede definirse como la ciencia social en la cual se aplican las herramientas de la teoría económica, las matemáticas y la inferencia estadística, al análisis de los fenómenos económicos. [4]

La econometría tiene que ver con la determinación empírica de las leyes económicas.[5]

El arte del econometrista consiste en encontrar consiste en encontrar el conjunto de supuestos que sean suficientemente específicos y realistas, de tal manera que le permitan aprovechar de la mejor manera posible los datos que tiene a su disposición.[6]

De esta serie de definiciones se puede determinar que la econometría evoluciona de la diversidad de la teoría económica. La econometría como disciplina establece planteamientos de validez general al satisfacer las condiciones y supuestos sobre los cuales se construye. Entre las diferentes definiciones que se le han dado a la econometría se puede considerar como el análisis cuantitativo de los fenómenos económicos reales fundamentados en la observación y la teoría, haciendo uso de la inferencia estadística y de la información cuantitativa de las variables económicas. De esta manera, se

[2] Gerhard Tintner, *Methodology of Mathematical Economics and Econometrics*, p. 74.
[3] P. A. Samuelson, T.C. Koopmans, and J. R. N. Stone, "Report of the Evaluative Committee for Econometrica", Econometrica, Vol. 22, No. 2 abril 1954, pp. 141-146.
[4] Arthur S. Goldberger, *Econometric Theory*, p. 1.
[5] H. Theil, *Principles of Econometrics*, p. 1.
[6] E. Malinvaud, *Statistical Methods of Econometrics*, p. 514.

combina la medición con la teoría. Por tanto, la econometría hace una combinación de la estadística matemática, de fórmulas matemáticas y de un modelo económico.

Es a partir de la econometría que se puede comprobar el grado de validez de los modelos económicos, se puede usar para explicar el comportamiento de la economía: de un ente económico, de un agregado económico, a los sectores económicos, entre otros. También se utiliza para observar la evolución de la economía en el tiempo, hacer predicciones cuantitativas y así sugerir medidas de política económica.

Como se puede observar, la econometría es una mezcla de la teoría económica, la estadística matemática, economía matemática y estadística económica.

- La teoría económica es el conjunto de principios o enunciados generales integrados en un cuerpo doctrinario sistematizado que pretenden explicar la realidad económica. Un ejemplo es la teoría microeconómica que establece como ley que cualquier alza en el precio de un bien será seguida de una disminución en la cantidad demandada de éste. Este es un principio económico del comportamiento del consumidor, permaneciendo lo demás constante, pero no hay una medida numérica en cuánto al aumento o la disminución del precio y la demanda del bien. Con la aplicación de la econometría permite hacer estimaciones numéricas a la relación entre el precio y la demanda.
- La estadística matemática proporciona una serie de herramientas a la econometría que le permiten recopilar, organizar, interpretar y analizar datos y establecer conclusiones en algunos problemas planteados. La información obtenida puede presentar errores, por lo que la econometría aplica métodos específicos.
- La economía matemática es una serie de formulaciones matemáticas para explicar la teoría económica haciendo caso omiso a la medición

y verificación de la teoría económica. En este caso, la econometría hace uso de las ecuaciones matemáticas convirtiéndolas en ecuaciones econométricas.

- La estadística económica recopila, clasifica y hace una descripción de la información económica cuantitativa con el uso de gráficas o de manera tabular. Por ejemplo, cifras de la inflación, el desempleo, precios, entre otros. El econometrista se encarga de que esta información valide la teoría económica.

Para que la econometría lleve a cabo el análisis econométrico y dé una explicación acerca de la relación de las variables económicas, toma como punto de partida el modelo económico y lo convierte en modelo econométrico. Para realizar el análisis econométrico se sigue la metodología econométrica, que aunque existen diversas escuelas de pensamiento acerca de ésta, la que a continuación se muestra es la metodología clásica, la cual predomina en la investigación empírica en economía, misma que se realiza dentro de los siguientes lineamientos:

1) Enunciar la teoría o hipótesis.
2) Especificación del modelo econométrico para probar la teoría.
3) Estimación de los parámetros del modelo.
4) Obtención de datos
5) Verificación o inferencia estadística, mediante las pruebas de hipótesis.
6) Predicciones o pronósticos.
7) Utilización del modelo para fines de control o establecimiento de políticas.

ACTIVIDAD DE APRENDIZAJE

Elaborar un esquema de la división de la econometría e indicar a qué se refiere cada concepto.

Asimismo, buscar tres ejemplos de modelos econométricos analizando en cada uno los elementos que los conforman, es decir, cuál es la(s) variable(s) dependiente, la(s) variables independiente(s), qué tipo de ecuación(es) se tiene y qué teoría es la que se aplica. Entregar en hojas blancas en la siguiente sesión.

AUTOEVALUACIÓN

1. Se encarga de recopilar, clasificar y hace una descripción de la información económica cuantitativa con el uso de gráficas o de manera tabular. 2. Proporciona una serie de herramientas a la econometría que le permiten recopilar, organizar, interpretar y analizar datos y establecer conclusiones en algunos problemas planteados. 3. Disciplina que establece planteamientos de validez general al satisfacer las condiciones y supuestos sobre los cuales se construye. 4. Se inscribe en la conjugación de la teoría económica y las mediciones verdaderas usando la teoría y la técnica de la inferencia estadística como puente. 5. Es aquélla que se determina dentro del sistema económico. 6. Hace uso de variables dependientes y variables independientes, y representa una relación de causalidad entre dos variables. 7. Parte de un conjunto de ecuaciones, las cuales se pueden tratar separadamente o de manera conjunta. 8. Es simplemente un conjunto de ecuaciones matemáticas. 9. Son las que se recolectan sobre unidades individuales en un momento del tiempo.	( ) Variable endógena ( ) Estadística económica ( ) Modelo ( ) Econometría ( ) Series de corte transversal ( ) Estadística matemática ( ) Modelo econométrico ( ) Método de investigación econométrica ( ) Series de temporales ( ) Relación multiecuacional

10. Es información que se recopila durante un determinado tiempo, se recolecta en intervalos muy cortos con movimientos ascendentes y descendentes, simultáneamente.	

Respuestas

1. Se encarga de recopilar, clasificar y hace una descripción de la información económica cuantitativa con el uso de gráficas o de manera tabular. 2. Proporciona una serie de herramientas a la econometría que le permiten recopilar, organizar, interpretar y analizar datos y establecer conclusiones en algunos problemas planteados. 3. Disciplina que establece planteamientos de validez general al satisfacer las condiciones y supuestos sobre los cuales se construye. 4. Se inscribe en la conjugación de la teoría económica y las mediciones verdaderas usando la teoría y la técnica de la inferencia estadística como puente.	(5) Variable endógena (1) Estadística económica (8) Modelo (3) Econometría (9) Series de corte transversal (2) Estadística matemática (6) Modelo econométrico (4) Método de investigación econométrica (10) Series de temporales (7) Relación multiecuacional

5. Es aquélla que se determina dentro del sistema económico. 6. Hace uso de variables dependientes y variables independientes, y representar una relación de causalidad entre dos variables. 7. Parte de un conjunto de ecuaciones, las cuales se pueden tratar separadamente o de manera conjunta. 8. Es simplemente un conjunto de ecuaciones matemáticas. 9. Son las que se recolectan sobre unidades individuales en un momento del tiempo. 10. Es información que se recopila durante un determinado tiempo, se recolecta en intervalos muy cortos con movimientos ascendentes y descendentes, simultáneamente.	

UNIDAD 2

MODELO DE REGRESIÓN CLÁSICO

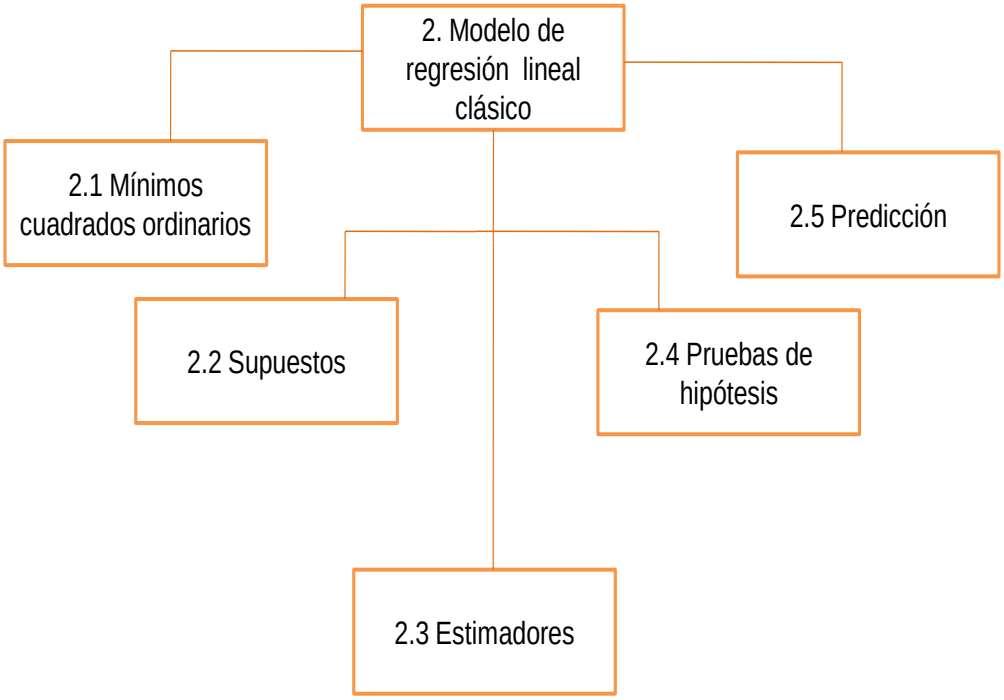
OBJETIVO

El estudiante distinguirá el enfoque de mínimos cuadrados al análisis de regresión e identificará que los estimadores del modelo de mínimos cuadrados ordinarios siguen distribuciones probabilísticas conocidas. Asimismo, el estudiante analizará la estimación y pruebas de hipótesis, y determinará los estadísticos de prueba.

TEMARIO

2.1 MÍNIMOS CUADRADOS ORDINARIOS (MCO)
2.2 SUPUESTOS
2.3 ESTIMADORES
2.4 PRUEBAS DE HIPÓTESIS
2.5 PREDICCIÓN

MAPA CONCEPTUAL



INTRODUCCIÓN

Debe tenerse en cuenta que el modelo de regresión lineal clásico es una abstracción o construcción teórica, pues los supuestos que lo fundamentan pueden llegar a ser considerados rigurosos o poco realistas, pero en la medida en que se progrese en conocimientos, estos supuestos pueden llegar a ser modificados sobre la marcha.

Por medio del enfoque de mínimos cuadrados se efectúa el análisis de regresión, el cual bajo ciertos supuestos produce estimadores lineales insesgados, incluso algunos de esos estimadores presentan varianza mínima.

Se realiza el análisis del problema de estimación puntual de los coeficientes de regresión, se considera la precisión del estimador con la medición del error estándar. Se aplica con esto inferencias acerca de los parámetros (poblacionales) y la aplicación de las pruebas de hipótesis de dichos parámetros.

Un tema a tratar es el problema de la bondad de ajuste de la regresión muestral, el cual se mide por medio del coeficiente de determinación r^2, mismo que será calculado.

Asimismo, se verifica cómo las perturbaciones poblacionales tienen una distribución normal y cómo bajo este supuesto los estimadores del modelo de mínimos cuadrados ordinarios siguen distribuciones probabilísticas conocidas.

Por último, se analizan dos ramas de la estadística clásica como son la estimación y las pruebas de hipótesis, para ello se aplican los intervalos de confianza y la prueba de significancia. También se demuestra cómo la línea de regresión muestral que se obtiene de los datos, puede utilizarse para la predicción o proyección.

2.1 MÍNIMOS CUADRADOS ORDINARIOS (MCO)

El método de mínimos cuadrados es uno que se emplea para estimar o predecir el valor de una variable en función de valores de otra variable, teniendo como antecedente el comportamiento de un conjunto de datos del mismo tipo.

El problema de la predicción lineal se reduce al de ajustar una línea recta a un conjunto de puntos localizados en un diagrama de dispersión. El diagrama de dispersión sirve de base para conocer el tipo de curva que mejor se ajusta a los datos, si esta curva resulta una recta, se llama recta de ajuste.

La recta de ajuste es una línea recta que hace mínima la suma de las desviaciones de cada punto con respecto a la línea recta, esta recta se le conoce como recta de mínimos cuadrados y está representada en la figura 2.1.

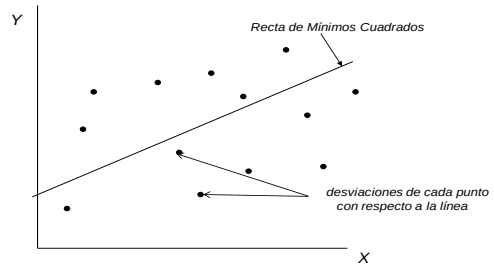


Figura 2.1

Cuando se tienen asignados valores a dos variables distintas obtenidas del mismo elemento de la población o muestreo, se les denomina datos bivariados, esto es que varían aun cuando estén relacionadas de alguna manera. Matemáticamente, los datos bivariados constituyen pares ordenados; llamados *x* y *y*, donde *x* es la variable de entrada y *y* es la variable de salida. La variable de entrada es la que se puede controlar, o la variable a partir de la cual se pueden hacer predicciones. La variable de salida es la que se desea predecir. La predicción es uno de los objetivos principales del análisis de regresión.

Es a partir del diagrama de dispersión que se puede verificar el tipo de relación que existe entre las variables *x*, *y*. Este diagrama consiste en la traza de todos los pares ordenados de datos bivariados sobre un eje coordenado. Donde la variable de entrada, *x*, se localiza en el eje horizontal y la variable de salida, *y*, en el eje vertical.

Esto se puede ejemplificar mediante una muestra aleatoria de ocho alumnos universitarios a quienes se les registraron las faltas y calificaciones en la asignatura de aprender a aprender. En la siguiente tabla aparecen los datos muestrales.

Alumno	1	2	3	4	5	6	7	8
No. de faltas (*x*)	5	4	2	6	4	3	2	7
Calificación (*y*)	7	8	9	5	7	9	10	5

El diagrama de dispersión se muestra en el plano cartesiano, cada par ordenado se presenta mediante un punto, ver figura 2. 2.

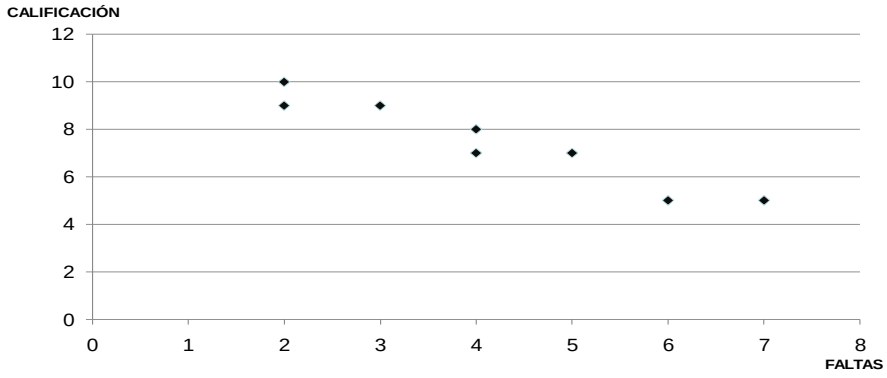


Figura 2.2

Si el diagrama de dispersión indica en general una relación lineal, entonces se ajusta una línea recta a los datos. Esta línea es ubicada por el método de mínimos cuadrados.

Enseguida se muestran algunos diagramas de dispersión en los que se establece una línea de regresión, la cual indica que si tiene pendiente positiva, indica una relación directa entre las variables, una pendiente negativa muestra una relación inversa entre las variables, y una pendiente de cero indica que no tienen relación entre sí.

DIAGRAMAS DE DISPERSIÓN

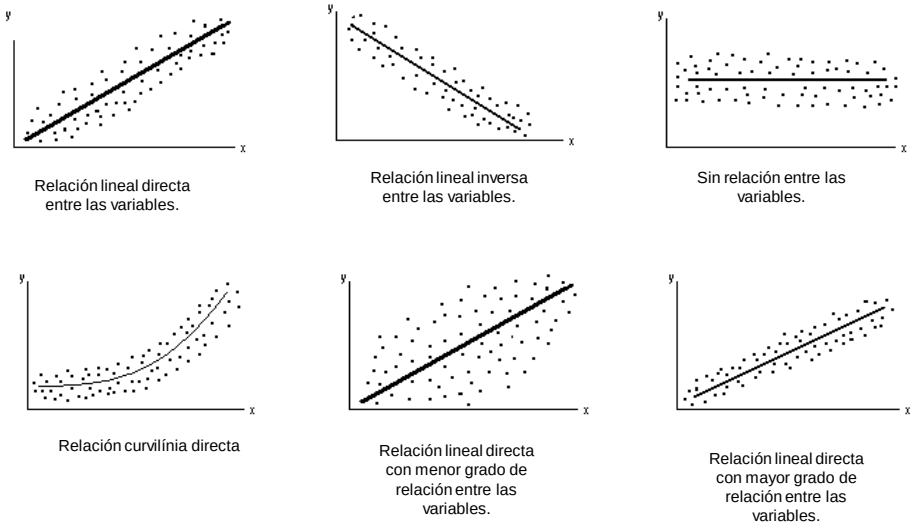


Figura 2.3

Una de las ecuaciones que representan al modelo de regresión simple y que es de las más utilizadas para efectuar la predicción, es la ecuación lineal de la forma:

$$y = mx + b$$

Donde:

m representa la pendiente de la recta.

b es el punto donde la recta intercepta al eje *y*.

La ecuación anterior representa la recta de regresión o recta de mínimos cuadrados.

Los valores de las constantes *m* y *b* se obtienen mediante las expresiones:

$$m = \frac{cov(x, y)}{var\ (x)}$$

Dado que la recta de regresión pasa por el punto en que se encuentra los puntos coordenados ($\bar{x}, \bar{y}$) que son las medias correspondientes de *y* y *x*, por lo cual esta satisface la ecuación de la recta

$$\bar{y} = m\bar{x} + b$$

Por lo que una vez conocidos $\bar{x}, \bar{y}$ y *m* se despeja *b* en la ecuación anterior, lo cual resulta:

$$b = \bar{y} - m\bar{x}$$

Donde la covarianza es una medida de dispersión conjunta de las dos variables de un conjunto de datos bivariados y se determina mediante la expresión:

$$\text{cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y - \bar{y})}{n}$$

En tanto, la varianza de x se determina con la expresión:

$$\text{var}(x) = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

Las expresiones anteriores se aplican en el ejemplo siguiente:

Obtener la ecuación de la recta de regresión para el siguiente conjunto de datos y estimar un valor para x = 0, x = 6.5 y 9.

x	y
2	8
8	10
3	7
1	5
4	9
5	8
6	9
3	7
2	6
4	8
8	10
5	9

Primero se determina los valores de $\bar{x}$ y $\bar{y}$ esto es:

$\bar{x} = 4.25 \qquad \bar{y} = 8$

Se determina la *cov(x,y)* y la *var(x)*, para lo cual se construye la siguiente tabla.

x	y	$(x_i - \bar{x})$	$(y_i - \bar{y})$	$(y_i - \bar{y})(x_i - \bar{x})$	$(x_i - \bar{x})^2$
2	8	-2.25	0	0	5.0625
8	10	3.75	2	7.5	14.0625
3	7	-1.25	-1	1.25	1.5625
1	5	-3.25	-3	9.75	10.5625
4	9	-0.25	1	-0.25	0.0625
5	8	0.75	0	0	0.5625
6	9	1.75	1	1.75	3.0625
3	7	-1.25	-1	1.25	1.5625
2	6	-2.25	-2	4.5	5.0625
4	8	-0.25	0	0	0.0625
8	10	3.75	2	7.5	14.0625
5	9	0.75	1	0.75	0.5625

51	**96**			**34.00**	

La varianza de x se calcula con la última columna.

$$\text{var}(x) = \frac{\Sigma(x_i - \bar{x})^2}{n}$$

$$\text{var}(x) = \frac{56.25}{12}$$

$$\text{var}(x) = 4.6875$$

La covarianza resulta en:

$$\text{cov}(x,y) = \frac{\Sigma(x_i - \bar{x})(y - \bar{y})}{n}$$

$$\text{cov}(x,y) = \frac{34}{12}$$

$$\text{cov}(x,y) = 2.8333$$

Ahora se obtiene el valor de la pendiente de la recta, *m*.

$$m = \frac{cov(x,y)}{var\ (x)}$$

$$m = \frac{2.8333}{4.6875}$$

$$m = 0.6044$$

El valor de b es:

$$b = \bar{y} - m\bar{x}$$

$$b = 8 - 0.6044(4.25)$$

$$b = 5.4311$$

Sustituyendo los valores anteriores, en la ecuación de la recta.

$$y = mx + b$$

Se obtiene la ecuación de la recta de regresión para el conjunto de datos dado.

$$y = 0.6044x + 5.4311$$

Las estimaciones para x = 0, x = 6.5 y 9 se obtienen sustituyendo estos valores en la ecuación de la recta regresión.

Para x = 0

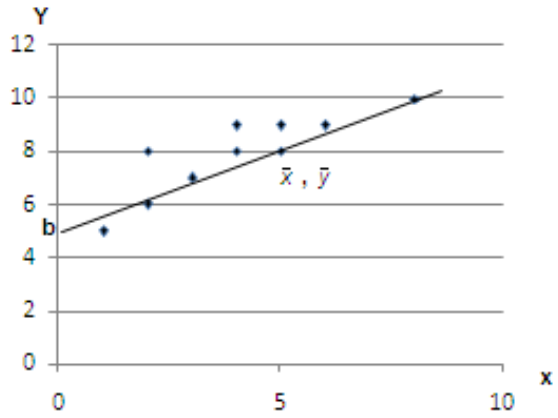
$$y = 0.6044x + 5.4311$$
$$y = 0.6044(0) + 5.4311$$
$$y = 5.4311$$

Para x = 6.5

$$y = 0.6044x + 5.4311$$
$$y = 0.6044(6.5) + 5.4311$$
$$y = 9.3597$$

Para x = 9

$$y = 0.6044x + 5.4311$$
$$y = 0.6044(9) + 5.4311$$
$$y = 10.8707$$



Recta de regresión y diagrama de dispersión

Al establecer un modelo que explique *y* en términos de *x*, no se puede determinar que hay una relación exacta entre las dos variables y que asegure que se están tomando en cuenta otros factores que influyen en *y*, ni asegurar la captura de una relación *ceteris paribus* entre *y* y *x*. Para ello se aplica el modelo de regresión simple en el cual es posible estudiar la relación entre las dos variables, sólo que este modelo cuenta con limitaciones como instrumento general para el análisis empírico. El modelo de regresión simple sirve de ejercicio para el estudio de la regresión múltiple.

Observaciones generales:

ACTIVIDAD DE APRENDIZAJE

a) Trazar un diagrama de dispersión del conjunto de datos proporcionados en la siguiente tabla:

x	2	12	4	6	9	4	11	3	10	11	3	1	13	12	14	7	2	8
y	4	8	10	9	10	8	8	5	10	9	8	3	9	8	8	11	6	9

b) ¿Sería justificado utilizar las técnicas de regresión lineal con estos datos para encontrar la línea de mejor ajuste? Explicar la respuesta.

2.2 SUPUESTOS

El análisis de regresión más sencillo para el caso de dos variables, también se le conoce como modelo de regresión simple, modelo de regresión bivariada o como modelo de regresión de dos variables. Este modelo no es de uso amplio en la econometría aplicada, pero sirve para ilustrar las ideas básicas del mismo.

El análisis de regresión simple permite estimar o predecir el valor medio o promedio (poblacional) de la variable dependiente *y* con base en los valores fijos o conocidos de la variable explicativa *x*. Una ecuación simple que relacione a *Y* con *X* y que dispone de *n* observaciones es:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad 2.1$$

Esta ecuación se define como el modelo de regresión lineal simple donde:

Y es la variable dependiente o explicada cuyo comportamiento se quiere analizar.

X es la variable independiente o explicativa considerada como la causa que crea transformaciones en la variable dependiente.

β son los parámetros cuyo valor se desconoce y se deben estimar. Es mediante esta estimación que se obtiene una medición de las relaciones existentes entre *Y* y *X*.

u es la variable aleatoria llamada término de error o perturbación, y que recoge el efecto conjunto de otras variables no directamente explicadas en el modelo, cuyo efecto individual sobre la variable dependiente no resulta relevante.

i hace referencia a las diversas observaciones para las cuales se establece su validez.

La ecuación (2.1) establece la relación entre *Y* y *X*, si los factores de u_i se mantienen fijos de tal forma que un cambio en u_i sea cero ($\Delta u_i = 0$). De esta manera, el cambio en *Y* es β_1 multiplicada por el cambio en X_i. β_1 es el

parámetro de la pendiente de la relación entre Y y X si se mantienen fijos en u_i los otros factores. En tanto, β_0 es el parámetro de intercepción que también tiene sus usos, pero es poco decisivo para el análisis. La intercepción β_0 de la ecuación no pierde generalidad al suponer que el valor promedio de u_i en la población es cero.

$$E(u_i) = 0 \qquad 2.2$$

La expresión (2.2) no indica ninguna relación entre u_i y X_i, sólo determina la distribución de los factores inobservables de la población. Al relacionar u_i y X_i, variables aleatorias, se puede definir la distribución condicional de u_i dado cualquier valor de X_i. Para cualquier valor de X_i se puede obtener el valor esperado (o promedio) de u_i para aquella parte de la población que describe X_i. El supuesto de que el valor promedio de u_i no dependa de X_i se denota así:

$$E(u_i \mid X_i) = E(u_i) = 0 \qquad 2.3$$

La primera parte de la igualdad supone la media condicional cero, es decir, que para cualquier X, el promedio de los factores inobservables es el mismo Y, por tanto debe ser igual al promedio de u_i para toda la población.

Suponiendo que la media condicional de 2.1 está en función de X_i y utilizando $E(u_i) = 0$ se tiene:

$$E(Y \mid X_i) = \beta_0 + \beta_1 X_i \qquad 2.4$$

La ecuación (2.4) muestra la función de regresión poblacional (FRP) o regresión lineal poblacional, donde E ($Y|X_i$) es una función lineal de X_i. En otras palabras, indica como el valor promedio (poblacional) de Y varía con las X. De manera geométrica, figura 2.4, se puede mostrar la curva de regresión poblacional que es la unión de las medias condicionales o esperanzas de la variable dependiente para los valores fijos de la variable explicativa.

Se puede observar en la figura 2.4 la curva de regresión poblacional que muestra que para cada X_i existe una población de valores de Y, que se suponen normalmente distribuidos, y una media condicional correspondiente. La línea o curva de regresión atraviesa las medias condicionales.

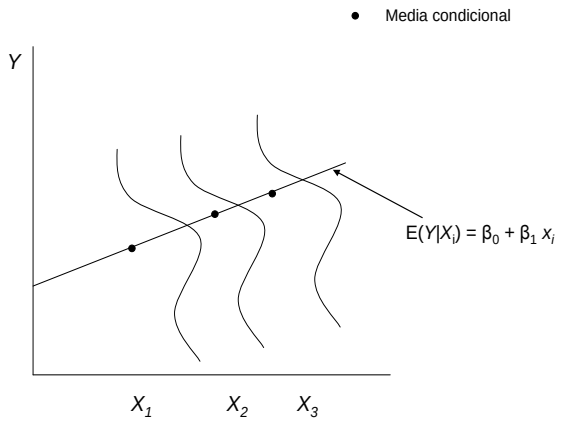


Figura 2.4

Los planteamientos hechos hasta el momento, suponen valores poblacionales de Y correspondientes a los valores fijos de X. Es momento de hacer referencia a problemas de muestreo, pues en la práctica lo que está al alcance es una muestra de valores de Y correspondiente a valores fijos de X.

De manera parecida a la función de regresión poblacional, es posible desarrollar la función de regresión muestral (FRM) para representar la línea de regresión muestral, de manera que la función (2.4) puede escribirse como:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i \qquad 2.5$$

Donde:

$\widehat{\ }$ se lee como “sombrero” o “gorro”

$\widehat{Y}_i$ = estimador de E($Y|X_i$)

$\widehat{\beta}_0$ = estimador de β_0

$\widehat{\beta}_1$ = estimador de β_1

El estimador (Λ) o estadístico (muestral) es un método que dice cómo estimar el parámetro poblacional a partir de la información proporcionada por la muestra que se tenga. La ecuación (2.1) también se puede expresar en su forma estocástica de la siguiente manera:

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \qquad 2.6$$

Conceptualmente e_i es análogo a u_i y denota el término residual (muestral). Debido a las fluctuaciones de una muestra a otra correspondientes a una población, la función de regresión muestral (2.6) es una aproximación para estimar la función de regresión poblacional (2.1). Por tanto, el objetivo primordial del análisis de regresión consiste en estimar la función de regresión poblacional

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad 2.1$$

con base en la función de regresión muestral

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \qquad 2.6$$

Debido a que se dan fluctuaciones entre una muestra y otra, tomada de una misma población, la estimación de la FRP con base en la FRM, es sólo una aproximación.

Hasta el momento, sólo se han considerado algunas ideas fundamentales del análisis de regresión lineal con parámetros desconocidos.

ACTIVIDAD DE APRENDIZAJE

En la siguiente tabla se presenta la cotización (*X*) y el rendimiento al vencimiento *Y* (%) de 50 bonos, donde la cotización se mide en tres niveles: X = 1 (AAA), X = 2 (AA) y X = 3 (A). Cada uno de estos bonos contiene cierto nivel de riesgo, donde AAA significan bonos de alto riesgo, mientras que AA tienen un riego intermedio y A es de bajo riesgo.

Y \ *X*	**1** **AAA**	**2** **AA**	**3** **A**	**Total**
8.5	13	5	0	18
11.5	2	14	2	18
17.5	0	1	13	14
Total	15	20	15	50

Convertir la tabla anterior en una tabla que presente la distribución probabilística conjunta, p(X, Y), es decir, p(*X* = 1, *Y* = 8.5) = 13/50 = .26

2.3 ESTIMADORES

El problema de predicción lineal se reduce al de ajustar una línea recta a un conjunto de puntos localizados a un diagrama de dispersión. El procedimiento que más se emplea para el ajuste de una recta a un conjunto de puntos se conoce como método de mínimos cuadrados ordinarios (MCO).

Retomando la función de regresión de la población en dos variables.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad 2.1$$

Se determinó que bajo esta función no se podía observar directamente, por lo que se estimó a partir de la función de regresión de la muestra

$$Y_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + e_i \qquad 2.6$$

$$= \hat{Y}_i + e_i \qquad 2.7$$

donde $\widehat{Y_i}$ es el valor estimado de Y_i

Transformando la expresión (2.7) como:

$$e_i = Y_i + \hat{Y}_i$$
$$= Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i \qquad 2.8$$

En este caso e_i son los residuos, es decir, las diferencias entre los valores reales y los estimados de *Y*. Dados *N* pares de observaciones de *Y* y *X* se debe de determinar la función de regresión muestral de tal modo que esté tan cerca como sea posible del *Y* real. A partir de esto se adopta el siguiente criterio $\Sigma e_i = \Sigma(Y_i - \hat{Y}_i)$, de tal manera que la suma de los residuos resulte ser tan pequeña como sea posible. A partir de este razonamiento se desprende el diagrama hipotético que se muestra en la figura 2.5.

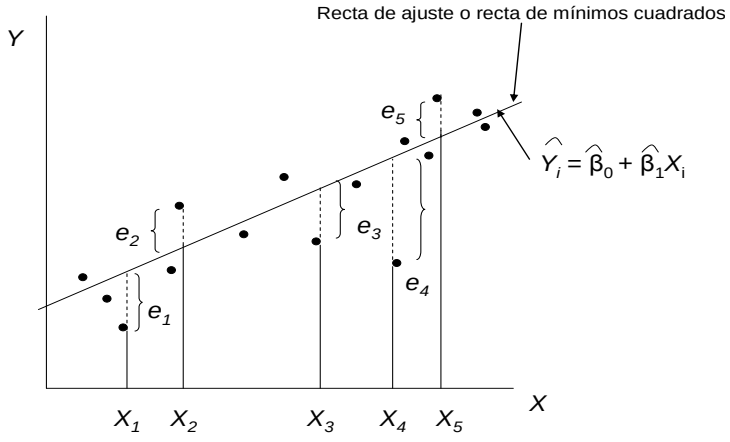


Figura 2.5. Diferencia de los residuos en la recta de los mínimos cuadrados.

Si se adopta el criterio de minimizar Σ e_i, se observa que en la figura 2.5 se le da la misma importancia o peso a los residuos sin importar que tan cerca o que tan dispersas están las observaciones individuales de la línea de ajuste. Este problema se evita estableciendo el criterio de los mínimos cuadrados según el cual la función de regresión muestral puede plantearse en modo tal que

$$\begin{aligned} \Sigma e^2_i &= \Sigma(Y_i - \hat{Y}_i) \\ &= \Sigma(Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \end{aligned} \qquad 2.9$$

resulte ser tan pequeña como sea posible y en donde Σe_i^2 representan los residuos al cuadrado. Al elevar los residuos e_i al cuadrado, se le asigna mayor peso a los residuos que se encuentran más alejados de la recta, tal es el caso de los residuos e_1 y e_4 que se encuentran en la figura 2.5, es importante destacar que cuanto más grandes sean los e_i (en valores absolutos), mayor será Σe_i^2.

Retomando la ecuación (2.8), se tienen dos parámetros desconocidos que estimar, por lo que se esperaría obtener de esta ecuación, buenos estimadores de β_0 y β_1. Dada una muestra, se eligen como estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$, y determinando la media de Y y X se tiene $\bar{Y}$ y $\bar{X}$, respectivamente. De esta manera la 2.8 se reescribe como:

$$\bar{Y} = \hat{\beta}_0 + \hat{\beta}_1 \bar{X} \qquad 2.10$$

Restando 2.10 de 2.6 se tiene

$$\begin{aligned} Y - \bar{Y} &= \hat{\beta}_1 (X - \bar{X}) + e_i \\ y_i &= \hat{\beta}_1 x_i + e_i \end{aligned} \qquad 2.11$$

Donde y_i y x_i representan desviaciones con relación a sus respectivos valores medios (muestrales), $x_i = (X - \bar{X})$ y $y_i = (Y - \bar{Y})$. A la ecuación (2.11) se le conoce como la forma de la desviación para uno de los estimadores de

mínimos cuadrados ordinarios. Al aplicar las siguientes ecuaciones para la estimación de β_0 y β_1

$$\Sigma Y_i = N\hat{\beta}_0 + \hat{\beta}_1 \Sigma X_i \qquad 2.12$$

$$\Sigma Y_i X_i = \hat{\beta}_0 \Sigma X_i + \hat{\beta}_1 \Sigma X_i^2 \qquad 2.13$$

donde *N* es el tamaño de la muestra y al resolver simultáneamente se obtiene.

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} \qquad 2.14$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1 \bar{X} \qquad 2.15$$

De forma alterna se puede estimar β_1 mediante

$$\begin{aligned} \hat{\beta}_1 &= \frac{\Sigma x_i y_i}{\Sigma x_i^2} \\ &= \frac{\Sigma x_i Y_i}{\Sigma X_i^2 - N\bar{X}^2} \\ &= \frac{\Sigma X_i y_i}{\Sigma X_i^2 - N\bar{X}^2} \end{aligned} \qquad 2.16$$

La ecuación (2.14) determina la pendiente estimada, en donde el numerador representa la covarianza muestral entre *X* y *Y*, mientras que el dividendo es la varianza muestral de X. Una consecuencia inmediata es que si *X* y *Y* presentan correlación muestral positiva, entonces $\hat{\beta}_1$ es positiva; si la correlación entre *X* y *Y* es negativa, entonces $\hat{\beta}_1$ es negativa.

Los estimadores dados en (2.14) y (2.15) se denominan estimadores de mínimos cuadrados ordinarios de β_0 y β_1.

Volviendo a la ecuación (2.11), el término de intersección $\hat{\beta}_0$ ha desaparecido, debido a que la línea de regresión muestral pasa siempre por las

medias muestrales de X y Y, esto permite observar que la función de regresión poblacional en forma de desviación se puede expresar de la siguiente manera:

$$\hat{y}_i = \hat{\beta}_1 x_i \qquad 2.17$$

Siendo que las unidades originales de medida de dicha expresión eran $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ como se muestra en (2.5).

Ahora bien, al modelo de regresión simple se le ha denominado también modelo de regresión lineal, la ecuación (2.1) $Y_i = \beta_0 + \beta_1 X_i + u_i$ muestra que es lineal en los parámetros β_0 y β_1. La ecuación (2.1) se entiende como la función de regresión poblacional donde Y_i depende de X_i y de u_i. Al no especificar la forma como se genera X_i y u_i, no se pueden hacer inferencias estadísticas sobre Y_i, β_0 y β_1. A partir de esto, que se desprenden los supuestos en términos de X_i y del término de perturbación u_i, para dar una interpretación válida de estimaciones de la regresión.

$\Rightarrow$ Supuesto 1. Linealidad en los parámetros.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad 2.1$$

En este modelo poblacional, la variable dependiente Y_i se relaciona con la variable independiente X_i y la perturbación u_i, donde los parámetros β_0 y β_1 hacen referencia a la intercepción y la pendiente poblacional. Los datos de las variables aleatorias X y Y, permiten estimar los parámetros β_0 y β_1.

$\Rightarrow$ Supuesto 2. Valores fijos de X.

En un muestreo repetido, los valores de X permanecen fijos, es decir, es una variable no estocástica. Las muestras repetidas no son tan realistas en un contexto no experimental.

$\Rightarrow$ Supuesto 3. Media condicional cero.

$$E(u_i \mid X_i) = 0 \qquad 2.17$$

El valor medio de la perturbación u_i es igual a cero, dado el valor de X_i, esto se puede verificar en la siguiente figura.

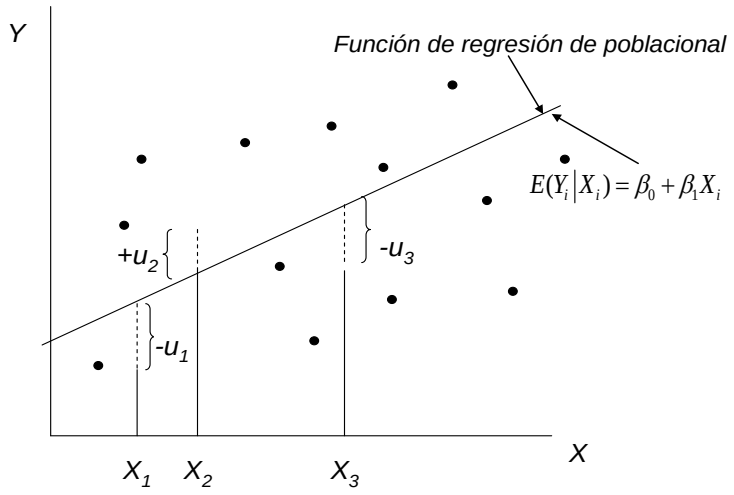


Figura 2.6. Diferencias de los términos de perturbación en la función de regresión de la población.

En la figura 2.6 se muestra la asociación de *Y* con *X*, distribuidos alrededor de su media. Los valores de *Y* corresponden a un *X* dado, se encuentran distribuidos alrededor de su valor medio que se encuentra sobre la línea recta de la función de regresión poblacional. Los puntos localizados por encima o por debajo de la media son los u_i, que se encuentran de manera positiva o negativa. De aquí que el efecto promedio sobre *Y* es cero. Esto es suponiendo que las u_i se distribuyen de manera simétrica. Un aspecto que hay que considerar es que $E(u_i|X_i) = 0$ implica $E(Y_i|X_i) = \beta_0 + \beta_1 X_i$ por lo que ambos supuestos son equivalentes.

⇒ Supuesto 4. Variación muestral en la variable independiente.

En la muestra, las variables independientes X_i no todas son iguales a una misma constante. Se requiere de cierta variación de *X* en la población. Se debe tener en cuenta que para realizar un análisis de regresión, la variación en *Y* al igual que en *X* es fundamental.

⇒ Supuesto 5. Hay independencia o no autocorrelación entre las perturbaciones (u).

$$cov(u_i, u_j) = E(u_i, u_j) = 0 \qquad 2.19$$

Donde *i* y *j* son dos observaciones diferentes (*i* ≠ *j)* y cov significa covarianza. Dados dos valores cualesquiera de *X* (X_i, X_j) para *i* ≠ *j* la correlación u_i, u_j es cero.

⇒ Supuesto 6.Independencia entre u_i y X_i.

$$cov(u_i, X_i) = E(u_i, X_i) = 0 \qquad 2.20$$

Este supuesto afirma que la perturbación u y la variable explicativa *X* no están correlacionadas. Este supuesto se cumple automáticamente si la variable *X* no es aleatoria o estocástica y si el supuesto 3 se mantiene

⇒ Supuesto 7. Homoscedasticidad.

$$Var(u \mid X) = \sigma^2 \qquad 2.21$$

Homoscedasticidad significa igual varianza u_i. La varianza de u_i para cada X_i es número positivo constante igual a σ^2, esto implica igual dispersión o igual varianza. Del mismo modo, se puede decir que la población de *Y* que corresponde a los diferentes valores de *X* tienen misma varianza, lo cual se verifica en la siguiente figura.

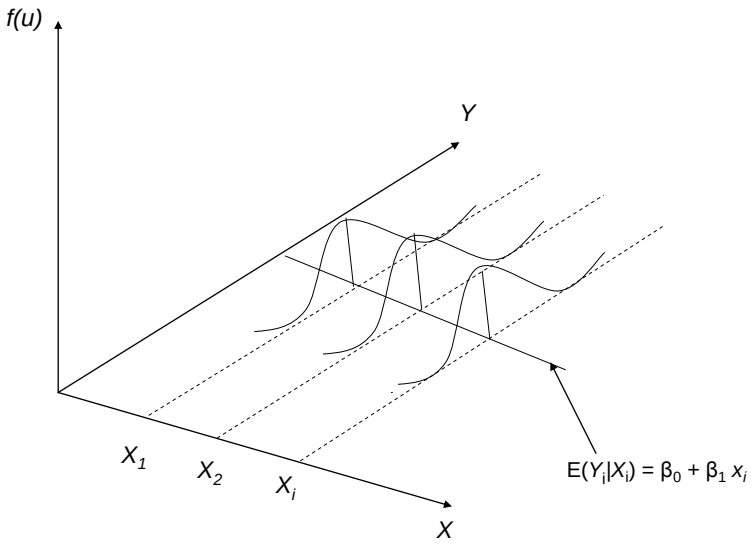


Figura 2.7. Homocedasticidad.

El análisis de corte transversal afirma que la varianza de los factores no observables, u_i, condicionada sobre X_i es constante. A medida que aumenta X, Y promedio también aumenta. La varianza de Y permanece igual para todos los niveles de X.

$\Rightarrow$ Supuesto 8. Heteroscedasticidad.

La heteroscedasticidad se conoce también como dispersión desigual o varianza desigual, la cual se expresa mediante la siguiente ecuación.

$$Var(u_i \mid X_i) = \sigma_i^2 \qquad 2.22$$

La varianza σ^2 de la ecuación (2.21) con subíndice, indica que la varianza de la población Y ya no es constante. Cuando la varianza condicional de la población Y aumenta a medida que X aumenta, se dice que hay una situación de heteroscedasticidad, esto se puede apreciar en la siguiente figura:

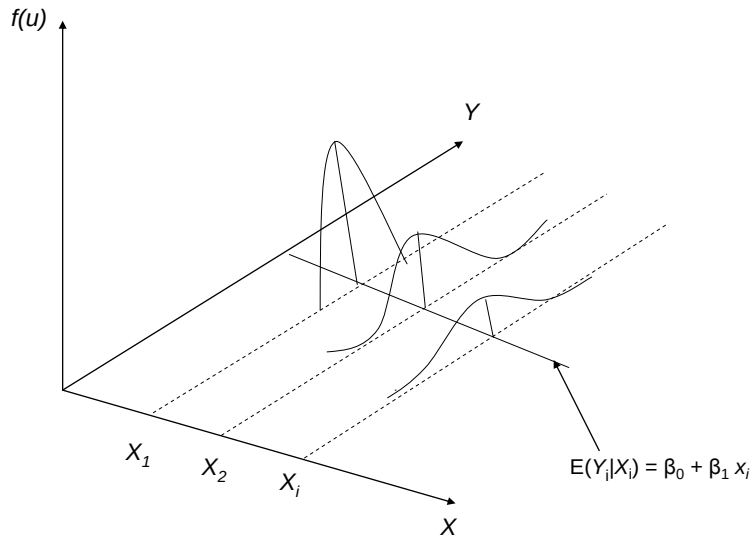


Figura 2.8. Heterocedasticidad.

En este caso, a medida que el valor de *X* aumenta, *Y* promedio también aumenta y, la varianza aumenta con *X*. En la figura 2.8 muestra que la varianza *u* condicional de *X* conforme pasa de X_1 hasta X_i cada vez es menor. Mientras se permita que aumente *Y* promedio con la *X*, se supone que la variabilidad de *Y* alrededor de la media, es constante en todo *X*.

Estos supuestos pertenecen únicamente a la función de regresión poblacional y no a la función de regresión muestral. Así, bajo estos supuestos y de acuerdo con las ecuaciones (2.14) y (2.15) se hace necesario obtener una medida de precisión de los estimadores $\hat{\beta}_0$ y $\hat{\beta}_1$ la cual se mide a través de su error estándar (es).

Este error es la desviación estándar muestral de la distribución muestral del estimador, es decir, es una distribución del conjunto de valores del estimador obtenidos de todas las muestras posibles del mismo tamaño y de la misma población. Bajo los supuestos anteriores y una vez establecidas las ecuaciones (2.14) y (2.15) se determinan los errores estándar de los estimadores de MCO.

$$Var\ (\hat{\beta}_1) = \frac{\sigma^2}{\Sigma x_i^2} \qquad 2.23$$

$$se(\hat{\beta}_1) = \frac{\sigma}{\sqrt{\Sigma x_i^2}} \qquad 2.24$$

$$Var\ (\hat{\beta}_0) = \frac{\Sigma X_i^2}{N\Sigma x_i^2}\sigma^2 \qquad 2.25$$

$$se(\hat{\beta}_0) = \sqrt{\frac{\Sigma X_i^2}{N\Sigma x_i^2}}\sigma \qquad 2.26$$

Donde *var* es la varianza y *se* es el error estándar. Las ecuaciones (2.24) y (2.26) son el error estándar del análisis de regresión simple, que no es válido en presencia de heteroscedasticidad. En la mayoría de los casos el que interesa es $Var\ (\hat{\beta}_1)$. Se puede resumir fácilmente cómo esta varianza depende de la varianza del error, σ^2, también conocida como varianza homoscedástica de u_i en razón del supuesto 7.

Las cantidades que entran en las ecuaciones anteriores se pueden estimar a partir de los datos excepto σ^2, lo que hace necesario establecer su estimador mediante la siguiente fórmula.

$$\hat{\sigma}^2 = \frac{\Sigma e_i^2}{N - 2} \qquad 2.27$$

donde $\hat{\sigma}^2$ es el estimador de MCO del verdadero y desconocido σ^2 y N - 2 es el número de grados de libertad (*g de l*) en los residuos de MCO, es decir, N es el número de observaciones y 2 el número de restricciones independientes (lineales). Puede calcularse fácilmente $\hat{\sigma}^2$ si se conoce Σe_i^2, Σe_i^2 puede calcularse a partir de (2.9), o bien utilizando la siguiente expresión:

$$\Sigma e_i^2 = \Sigma y_i^2 - \hat{\beta}_1^2 \Sigma x_i^2 \qquad 2.28$$

Al comparar 2.9 con 2.18, se puede observar que es más sencillo utilizar la expresión 2.18, puesto que esta se hace para la totalidad de las observaciones. Una expresión alterna puede ser:

$$\Sigma e_i^2 = \Sigma y_i^2 - \frac{(\Sigma x_i y_i)^2}{\Sigma x_i^2} \qquad 2.29$$

Si se sustituye $\hat{\sigma}^2$ en las fórmulas de las varianzas (2.23) y (2.25), se tienen estimadores insesgados de $Var(\hat{\beta}_1)$ y $Var(\hat{\beta}_0)$. Más adelante se necesitarán estimadores de las desviaciones estándar de $\hat{\beta}_1$ y $\hat{\beta}_0$, por lo que el estimador natural de σ es:

$$\hat{\sigma} = \sqrt{\frac{\Sigma e_i^2}{N-2}} \qquad 2.30$$

y se denomina error estándar de la regresión, o bien, error estándar de la estimación. El estimador $\hat{\sigma}$ es un estimador de la desviación estándar de los factores no observables que influyen en *Y*. De manera equivalente estima la desviación estándar de *Y* después de eliminar el efecto de *X*. Por el momento, $\hat{\sigma}$ se usará para estimar las desviaciones estándar de $\hat{\beta}_0$ y $\hat{\beta}_1$. Sin embargo, es importante considerar que $\hat{\beta}_0$ y $\hat{\beta}_1$ son estimadores que no sólo varían de una muestra a otra sino dentro de una muestra dada tienden a depender entre sí. Esta dependencia se mide mediante la covarianza entre ellos.

$$\begin{aligned} \text{cov}(\hat{\beta}_0, \hat{\beta}_1) &= -\bar{X}\,\text{var}(\hat{\beta}_1) \\ &= -\bar{X}\left(\frac{\sigma^2}{\Sigma x_i^2}\right) \end{aligned} \qquad 2.31$$

La covarianza entre $\hat{\beta}_0$ y $\hat{\beta}_1$ depende del signo de $\bar{X}$. Si $\bar{X}$ es positiva, entonces la covarianza será negativa.

Bajo los supuestos del modelo de regresión lineal clásico, los estimadores poseen propiedades óptimas que se resumen en el teorema de Gauss Marcov, al considerar que un estimador es el mejor estimador lineal insesgado, es el que tiene la varianza mínima (eficiente). Tal es el caso del estimador $\hat{\beta}_1$, de mínimos cuadrados ordinarios, que es el mejor estimador lineal insesgado (MELI) de β_1, esto es si es lineal, insesgado y tiene varianza mínima como se muestra en la figura 2.9, en donde los datos están simétricamente distribuidos.

En el inciso (a) de la figura 2.9 se puede verificar cómo el valor promedio o esperado $E(\hat{\beta}_1)$ es igual β_1, lo que afirma que es $\hat{\beta}_1$, es un estimador insesgado de β_1. En tanto, el inciso (b) de la figura 2.9 presenta una distribución muestral con un estimador alterno $\hat{\beta}_1^*$, esto es con fines ilustrativos, en donde el valor esperado de $\hat{\beta}_1^*$ es igual al verdadero β_1. Al sobreponer la gráfica del inciso (a) con la del inciso (b) se obtiene que $\hat{\beta}_1^*$ y $\hat{\beta}_1$ son insesgados, por lo que debe de elegirse el que más cerca posible se encuentre de β_1, es decir, el que menor varianza tiene, a este se le llama estimador MELI.

Por el momento, es suficiente saber que estas propiedades no dependen de ningún supuesto en torno a la forma de la distribución de probabilidad de u_i.

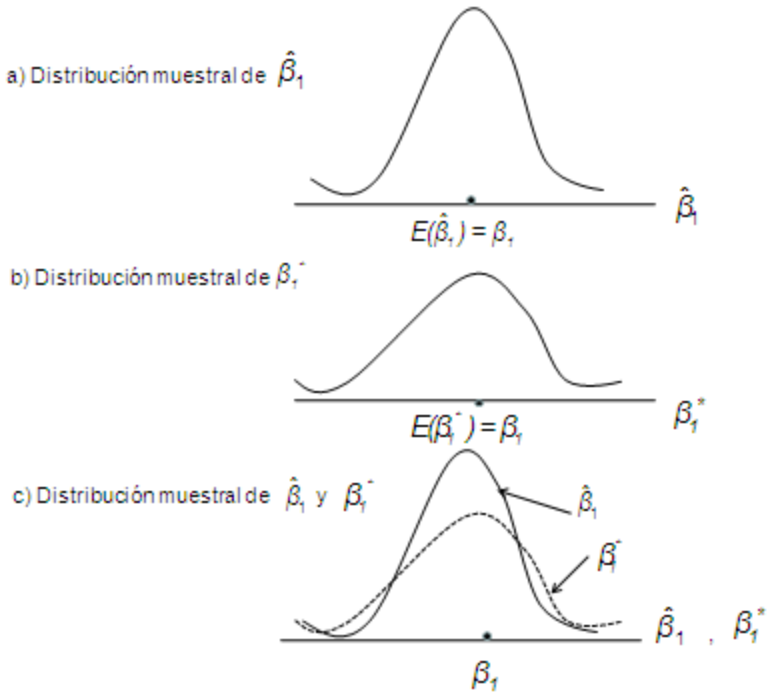


Figura 2.9. Distribuciones muestrales de $\hat{\beta}_1$ y de β_1^*.

Volviendo a la figura 2.6, se determina que si todas las observaciones coincidieran con la línea de regresión se obtendría un ajuste perfecto, lo que pocas veces sucede. Por lo general se presentan e_i positivos y negativos, esperando que los residuos localizados alrededor de la línea de regresión sean lo más pequeños posible. Es a partir de esto que se considera la llamada **bondad del ajuste** de la línea de regresión ajustada a un conjunto de datos, por lo que se busca una medida que ajuste la línea de regresión muestral a los datos. Es así que el coeficiente de determinación r^2 (para dos variables) es la medida que indica qué tan bien la línea de regresión muestral se ajusta a los datos. En caso de una regresión múltiple se determina con r^2.

Para calcular r^2 se procede a retomar la expresión 2.7

$$Y_i = \hat{Y}_i + e_i \qquad 2.7$$

o en forma de desviación

$$y_i = \hat{y}_i + e_i \qquad 2.32$$

elevando al cuadrado ambos lados de 2.32 y sumando sobre toda la muestra, se obtiene

$$\begin{aligned} \Sigma y_i^2 &= \Sigma \hat{y}_i^2 + \Sigma e_i^2 + 2\Sigma \hat{y}_i e_i^2 \\ &= \Sigma \hat{y}_i^2 + \Sigma e_i^2 \\ &= \beta_1^2 \Sigma x_i^2 + \Sigma e_i^2 \end{aligned} \qquad 2.33$$

A partir de una serie de derivaciones de 2.33 se puede determinar con la siguiente fórmula

$$\begin{aligned} r^2 &= \frac{\Sigma \hat{y}_i^2}{\Sigma y_i^2} \\ &= \frac{\beta_1^2 \Sigma x_i^2}{\Sigma y_i^2} \\ &= \beta_1^2 \left(\frac{\Sigma x_i^2}{\Sigma y_i^2} \right) \end{aligned} \qquad 2.34$$

Dado que $\hat{\beta}_1$ de la ecuación 2.14 y la ecuación 2.34 también puede expresarse como:

$$r^2 = \frac{(\Sigma x_i y_i)^2}{\Sigma x_i^2 \Sigma y_i^2} \qquad 2.35$$

La cantidad r^2 es el coeficiente de coeficiente de determinación que representa un valor del ajuste de una línea de regresión, mide la proporción de la variación de la variable dependiente explicada por las variaciones de las variables explicativas. Es claro, por tanto, que $0 £ r^2 £ 1$, donde 1 denota ajuste perfecto y 0 dice que no existe relación alguna entre la variable dependiente y la(s) variable(s) explicativa(s).

Una cantidad muy estrechamente relacionada a r^2, pero conceptualmente diferente, es el coeficiente de correlación, que es una medida de grado de relación entre dos variables. Puede calcularse como:

$$r = \pm\sqrt{r^2} \qquad 2.36$$

o a partir de su definición

$$r = \frac{\Sigma x_i y_i}{\sqrt{\frac{\Sigma x_i^2}{n}}\sqrt{\frac{\Sigma y_i^2}{n}}}$$

$$= \frac{\Sigma x_i y_i}{\sigma_x \sigma_y} \qquad 2.37$$

medida que se conoce como coeficiente de correlación lineal en donde σ_x es la desviación estándar de *X* y σ_y es la desviación estándar de *Y*.

Es momento de poner en práctica los anteriores conceptos. A continuación se ilustran los conceptos básicos de la teoría econométrica mediante un ejemplo numérico.

La siguiente tabla presenta datos muestrales relativos al número de horas de estudio fuera de clase durante un periodo de tres semanas, son ocho alumnos de un curso de econometría y se relaciona con las calificaciones de un examen final de ese periodo. Las calificaciones en examen están descritas en la columna Y_i, mientras que las horas de estudio están registradas en X_i.

Y_i	X_i	$X_i - \bar{X}_i = x_i$	$Y_i - \bar{Y}_i = y_i$	$x_i y_i$	y_i^2	x_i^2	X_i^2	$\hat{Y}_i$	e_i	e_i^2
64	20	-4	-12	48	144	16	400	70,0136054	-6,01360544	36,1634504
61	16	-8	-15	120	225	64	256	64,0272109	-3,02721088	9,16400574
84	34	10	8	80	64	100	1156	90,9659864	-6,96598639	48,5249664
70	23	-1	-6	6	36	1	529	74,5034014	-4,50340136	20,2806238
88	27	3	12	36	144	9	729	80,4897959	7,510204082	56,4031653

92	32	8	16	128	256	64	1024	87,9727891	4,027210884	16,2184275
72	18	-6	-4	24	16	36	324	67,0204082	4,979591837	24,7963349
77	22	-2	1	-2	1	4	484	73,0068027	3,993197279	15,9456245
608	**192**			**440**	**886**	**294**	**4902**	**608**	**0**	**227,496599**

Tabla 2.1.

Con base en esta información, se obtienen los siguientes cálculos:

$\hat{\beta}_0 = 40.0816$ $\quad var(\hat{\beta}_0) = 79.0241$ $\quad se(\hat{\beta}_0) = 8.8895$

$\hat{\beta}_1 = 1.4966$ $\quad var(\hat{\beta}_1) = 0.1289$ $\quad se(\hat{\beta}_1) = 0.3591$

$cov(\hat{\beta}_0, \hat{\beta}_1) = -3.0952$ $\quad \hat{\sigma}^2 = 37.9162$

$r^2 = 0.7432$ $\quad r = 0.8621$ $\quad g\ de\ l = 6$

Por tanto, la línea de regresión estimada es

$$\hat{Y}_i = 40.0816 + 1.4966X_i \qquad 2.38$$

La función de regresión muestral y la línea de regresión asociada a ella se puede interpretar de la siguiente manera: cada punto sobre la línea de regresión proporciona una estimación del valor medio o esperado de $\hat{Y}_i$ correspondiente a un valor seleccionado de (X_i, Y_i), es por tanto, una estimación de $E(Y_i/X_i)$. El valor de $\hat{\beta}_1 = 1.4966$, mide la pendiente de la línea, mientras que $\hat{\beta}_0 = 40.0816$ corresponde a la intersección de la línea con el eje Y e indica el nivel promedio de las calificaciones. En el análisis de regresión interpretación del intercepto no es siempre significativa, por lo que será sólo mejor interpretarla como el medio o promedio.

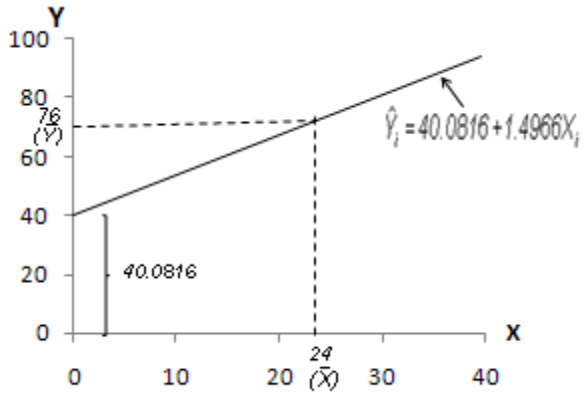


Figura 2.10 Línea de regresión muestral basada en las cifras de la tabla 2.1.

El valor 0.7432 para r^2 significa que cerca de 74% de la variación en las calificaciones de los alumnos se explica por la variable que corresponde a las horas de estudio; puesto que r^2 puede tener un valor máximo de 1 solamente, el r^2 observado sugiere que la línea de regresión muestral se ajusta a la información. El coeficiente de correlación de 0.8621 muestra que las dos variables, calificaciones del examen final y las horas de estudio, están positivamente correlacionadas.

Volviendo al modelo de regresión lineal normal con dos variables, se tiene que los estimadores de mínimos cuadrados ordinarios $\hat{\beta}_1$, $\hat{\beta}_0$ y $\hat{\sigma}^2$ satisfacen propiedades estadísticas como insesgamiento y varianza mínima, por lo que la estimación puntual es sólo la formulación de un aspecto de inferencia estadística, el otro son las pruebas de hipótesis, tema que será tratado en la siguiente sesión.

Hasta el momento, el método de mínimos cuadrados ordinarios no formula ningún supuesto acerca del error poblacional u_i sobre la función de regresión poblacional a partir de la función de regresión muestral.

El error poblacional u_i es independiente de las variables explicativas X_i y se distribuye de manera normal con media cero y varianza σ^2: u_i**:** normal (0, σ^2). Este supuesto de normalidad de u es la suma de muchos factores no observables que influyen en Y.

Bajo el supuesto de normalidad de u_i los estimadores $\hat{\beta}_1$ y $\hat{\beta}_0$ de mínimos cuadrados ordinarios también tienen una distribución normal.

Es importante mencionar como conocimiento y no así como comprobación, debido al alto grado de complejidad matemática, que un método alterno para la estimación puntual es el método de máxima verosimilitud que consiste en la estimación de parámetros desconocidos de manera tal que la probabilidad de observar un determinado valor de *Y* es la máxima posible. Bajo el supuesto de normalidad, los estimadores de máxima verosimilitud generalmente son iguales a los estimadores de mínimos cuadrados ordinarios.

Por último, para llevar a cabo la estimación y las pruebas de hipótesis necesarias de los modelos de regresión lineal, las herramientas son proporcionadas por el método de mínimos cuadrados ordinarios, sumado al supuesto adicional de la normalidad de u_i.

ACTIVIDAD DE APRENDIZAJE

A partir de la información proporcionada en la siguiente tabla, aplicar los conceptos básicos de la teoría econométrica y determinar su análisis una vez obtenidos los resultados. La columna *Y* hace referencia a la tasa salarial, en tanto que la columna *X* hace referencia a la tasa de desempleo, los datos son hipotéticos.

Y	X
1.3	6.2
1.2	7.8
1.4	5.8
1.4	5.7
1.5	5
1.9	4
2.6	3.2
2.3	3.6
2.5	3.3
2.7	3.3
2.1	5.6
1.8	6.8
2.2	5.6

2.4 PRUEBAS DE HIPÓTESIS

Una estimación puntual obtenida de determinada muestra no proporciona suficiente información para evaluar, por ejemplo, una teoría económica. Esta limitación se supera al aplicar intervalos de confianza que forma parte de la *teoría de la estimación*, siendo esta una rama de la estadística clásica.

Como se mencionó, existen fluctuaciones en las distribuciones muestrales de una población, por lo que una sola estimación puede diferir del valor verdadero. Por lo tanto, no se puede basar solo en la estimación puntual, sino que se puede proporcionar la probabilidad de que el verdadero parámetro se encuentre dentro de cierto rango.

A la estimación que incluye un intervalo de valores posibles dentro del cual se encuentra comprendido un parámetro de la población, se le conoce como *estimación del intervalo del parámetro*, en otras palabras, esto conduce a emplear intervalos con centro en el estadístico de la muestra, para calcular el parámetro de la población. Por ello se emplea con frecuencia la estimación por intervalos, ya que se obtiene un valor de mayor exactitud, que cuando se emplean las estimaciones puntuales.

La estimación de intervalos implica el cálculo de un parámetro de la población por medio de un intervalo de la recta real dentro del cual se considera comprendido el valor del parámetro, esto es, la estimación por intervalo se construye en forma tal que la probabilidad de que el intervalo contenga al parámetro, pueda especificarse. La característica fundamental de la estimación por intervalo, es que ilustra con exactitud la estimación del parámetro. Si la longitud del intervalo es muy pequeña, se obtiene exactitud excelente. Tales estimaciones por intervalo reciben el nombre de intervalos de confianza.

Este tipo de intervalos son de gran importancia y por ello se deben analizar con detenimiento. Así, la capacidad para estimar los parámetros de población mediante el uso de datos muestrales, se relaciona en forma directa con el conocimiento que se tiene acerca de la distribución de muestreo del valor estadístico que se está empleando como estimador.

Para un conjunto de datos que se hayan obtenido por muestras de una población normal, es posible que el valor del parámetro se encuentre dentro de cierto rango, alrededor del estimador puntual, digamos, por ejemplo, entre dos y tres errores estándar.

Suponiendo que se desea saber qué tan cerca está $\hat{\beta}_1$ de β_1 se puede tratar de encontrar dos números positivos δ y α, donde este último se encuentra entre 0 y 1, de manera que el intervalo aleatorio ($\hat{\beta}_1 - \delta$, $\hat{\beta}_1 + \delta$) contenga el valor verdadero de β_1 sea $1 - \alpha$. Es decir,

$$Pr(\hat{\beta}_1 - \delta \leq \beta_1 \leq \hat{\beta}_1 + \delta) = 1 - \alpha \qquad 2.39$$

Si este intervalo existe, se le conoce como intervalo de confianza que aquel que proporciona un intervalo de valores, centrado en el valor estadístico de la muestra, en el cual supuestamente se localiza el parámetro de la población con un riesgo de error conocido.

Los niveles de confianza que se utilizan con mayor frecuencia son los de 90%, 95% y 99%. En el caso de la ecuación 2.39, muestra que un estimador de intervalo, en contraste con un estimador puntual, es un intervalo construido con una probabilidad de $1 - \alpha$ de incluir dentro de sus límites el valor verdadero del parámetro. Por ejemplo, si $\alpha = 0.05$, entonces la expresión 2.39 sería la probabilidad de que el intervalo incluya el verdadero parámetro β_1 en 95%, es decir, la probabilidad de que el intervalo contenga β_1 es $1 - \alpha$, el cual recibe el nombre de coeficiente de confianza. En tanto, α es el nivel de significancia y debe encontrarse y debe ser $0 < \alpha < 1$.

Si se conocen el muestreo o las distribuciones probabilísticas de los estimadores y a partir de lo arriba descrito, se pueden hacer afirmaciones sobre el intervalo de confianza similares a la expresión 2.39, en donde se tiene un límite de confianza inferior ($\hat{\beta}_1 - \delta$) y un límite de confianza superior ($\hat{\beta}_1 + \delta$) a los cuales se les denomina valores críticos.

Con anterioridad también se estableció el supuesto de normalidad para u_i y para los estimadores de mínimos cuadrados ordinarios $\hat{\beta}_1$ y $\hat{\beta}_0$ que tienden a una distribución normal con medias y varianzas dadas. Por lo que entonces, se puede utilizar la distribución normal para hacer afirmaciones probabilísticas sobre β_1 dado que se conoce la verdadera varianza poblacional σ^2. Cabe mencionar que muy pocas veces se conoce σ^2, y en la práctica se termina utilizando el estimador insesgado $\hat{\sigma}^2$ por lo que se tiene la ecuación

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1 \sqrt{\sum x_i^2}}{\hat{\sigma}} \qquad 2.40$$

Siendo $\hat{\beta}_1$ el error estándar estimado y la variable *t* que es la distribución *t* de Student, la cual se aproxima a la distribución normal estandarizada y que se puede utilizar en lugar de la distribución normal y por medio de la cual se puede establecer intervalos de confianza para β_1, este valor *t* se determina con N – 2 g de l, quedando establecida en la siguiente ecuación.

$$Pr(-t_{\alpha/2} \leq t \leq t_{\alpha/2}) = 1 - \alpha \qquad 2.41$$

La *t* que se encuentra en el centro de la desigualdad es el valor *t* de 2.40 y $t_{\alpha/2}$ es el valor de la variable *t* obtenida de la distribución *t* para un nivel de significancia de α/2 y *N – 2 g de l*. Si se sustituye el valor de t (2.40) en 2.41 se obtiene

$$Pr\left[-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} \leq t_{\alpha/2}\right] = 1 - \alpha \qquad 2.42$$

Y reordenando en el intervalo de 2.42 se obtiene

$$Pr\left[\hat{\beta}_1 - t_{\alpha/2} se\hat{\beta}_1 \leq \beta_1 \leq \hat{\beta}_1 + t_{\alpha/2} se\hat{\beta}_1\right] = 1 - \alpha \qquad 2.43$$

Esta expresión indica que $\hat{\beta}_1$ se encontrará en una probabilidad del 100 (1 – α) por ciento para el parámetro $\hat{\beta}_1$. Una forma breve de escribir la 2.43 es

$$\hat{\beta}_1 \pm t_{\alpha/2} se\hat{\beta}_1 \qquad 2.44$$

De manera análoga se pueden establecer intervalos de confianza para β_0.

$$Pr\left[\hat{\beta}_0 - t_{\alpha/2} se\hat{\beta}_0 \leq \beta_0 \leq \hat{\beta}_0 + t_{\alpha/2} se\hat{\beta}_0\right] = 1 - \alpha \qquad 2.45$$

O bien, brevemente

$$\hat{\beta}_0 \pm t_{\alpha/2} se\hat{\beta}_0 \qquad 2.46$$

Regresando al ejemplo que se estableció en la sesión 2.3 de calificaciones en examen final-horas de estudio se encontró que $\hat{\beta}_1 = 1.4966$, $se(\hat{\beta}_1) = 0.3591$ y $g\ de\ l = 6$. Si se supone que α = 5%, es decir, un coeficiente de confianza de 95%, entonces la tabla *t* (misma que se podrá consultar en tablas estadísticas como distribución *t)* muestra que para 6 *g de l*, el $t_{\alpha/2}$ crítico es = $t_{0.025}$ = 2.447. Al sustituir los valores en 2.43 se verifica que el intervalo de confianza de 95% para β_1 es el siguiente:

$$0.6178 \leq \beta_1 \leq 2.3754 \qquad 2.47$$

O utilizando 2.44 es:

$$1.4966 \pm 2.447(0.3591)$$

es decir,

$$1.4966 \pm 0.8788 \qquad 2.48$$

La interpretación de este intervalo de confianza es, dado un coeficiente de confianza, de 95% en el largo plazo, en 95 de cada cien casos, intervalos como (0.6178, 2.3754) contendrán el verdadero β_1. Cabe mencionar que no se puede decir que existe una probabilidad de 95% de que el intervalo específico (0.6178, 2.3754) contenga el verdadero valor de β_1 porque este intervalo ahora es fijo, dejando por tanto, de ser aleatorio, en consecuencia, β_1 está o no está en el intervalo, debido a muestreos repetidos. La probabilidad de que el intervalo fijo contenga el verdadero valor de β_1 es, por tanto, 1 o 0.

Ahora bien, bajo la expresión 2.45 se puede verificar fácilmente de que el intervalo de confianza de 95% para β_0 es:

$$18.3289 \leq \beta_0 \leq 61.8343 \qquad 2.49$$

O utilizando 2.46

$$40.0816 \pm (2.447)(8.8896)$$

es decir,

$$40.0816 \pm 21.7527 \qquad 2.50$$

Este intervalo indica que de cada cien casos de intervalos como 2.49 contendrán el verdadero valor de β_0; la probabilidad de que este determinado intervalo fijo incluya el verdadero valor de β_0 es 1 o 0.

Corresponde ahora determinar el intervalo de confianza para σ^2 bajo el supuesto de normalidad.

$$\chi^2 = (N-2)\ \frac{\hat{\sigma}^2}{\sigma^2} \qquad 2.51$$

La prueba de la afirmación de que la 2.51 sigue una distribución normal, es mediante la aplicación de ji-cuadrada (χ^2) con *N – 2 g de l*, esto es para establecer intervalos de confianza para σ^2 como sigue:

$$Pr\left(\chi^2_{1-\alpha/2} \le \chi^2 \le \chi^2_{\alpha/2}\right) = 1-\alpha \qquad 2.52$$

Si se reemplaza χ^2 de la 2.51 en 2.52 se tiene la siguiente expresión.

$$Pr\left[(N-2)\ \frac{\hat{\sigma}^2}{\chi^2_{\alpha/2}} \le \sigma^2 \le N-2\frac{\hat{\sigma}^2}{\chi^2_{1-\alpha/2}}\right] = 1-\alpha \qquad 2.53$$

Esta expresión indica que $\hat{\sigma}^2$ se encontrará en una probabilidad del 100 (1 – α) por ciento para σ^2. Para verificar, nuevamente se retoma el ejemplo anterior, en donde se obtuvo $\hat{\sigma}^2 = 37.9162$ y $g\,de\,l = 6$. Si se le asigna a *α* un valor de 5%, la tabla ji-cuadrado (misma que se podrá consultar en tablas estadísticas como distribución χ^2) para *6 g de l* arroja los valores críticos $\chi^2_{0.025} = 14.4494$, y $\chi^2_{0.975} = 1.2373$. Estos valores muestran que la probabilidad de que un valor ji cuadrado exceda de 14.4497 es del 2.5% y que, sea mayor de 1.2373 es de 97.5%. De esta manera, el intervalo entre estos dos valores corresponde a un intervalo de confianza de 95% para χ^2 esto se muestra en la

siguiente figura 2.11, donde los valores críticos que se obtuvieron se encuentran en las colas de la curva marcados con 2.5%,

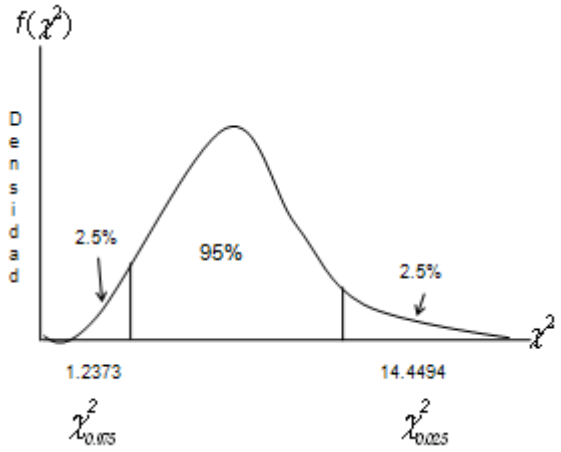


Figura 2.11. Intervalo de confianza del 95% para c^2 (6 g de l).

Al sustituir los datos del ejemplo en la 2.53 se verifica que el intervalo de confianza de 95% sobre σ^2 es el siguiente:

$$15.7443 \pounds \sigma^2 \pounds 183.8584 \qquad 2.54$$

Este intervalo se interpreta como el establecimiento de los límites de confianza de 95% sobre σ^2 y si se mantiene a priori que estos límites incluirán el verdadero valor de σ^2, a largo plazo (es decir, muestreos repetidos) se establece que se está en lo correcto 95% de las veces.

Una vez que se ha considerado la estimación puntual y los intervalos de confianza, se puede pasar a las llamadas pruebas de hipótesis.

El objetivo de la estimación es determinar el valor de cierto parámetro de la población, mientras que el objetivo de las pruebas de hipótesis es decidir si una afirmación acerca de un parámetro de la población es verdadera. Esto conduce a la teoría de la decisión estadística, la cual emplea fundamentalmente la prueba de hipótesis estadística. Las hipótesis estadísticas surgen de las distribuciones de probabilidad de las poblaciones.

Así, una prueba de hipótesis es un método para decidir cuándo aceptar o rechazar una hipótesis, tomando como base una muestra aleatoria de la población de donde se ha de formular la hipótesis.

Si se considera una muestra de la población, la hipótesis tiene una mayor importancia, ya que a partir de la información obtenida de dicha muestra, se ha de decidir si la hipótesis es verdadera o falsa.

Los métodos que deciden si una hipótesis se acepta o se rechaza o el determinar si las muestras observadas difieren significativamente de los resultados esperados, reciben el nombre de *prueba de hipótesis, pruebas de significancia o reglas de decisión*.

Al formular una hipótesis con el fin de aceptarla o de rechazarla, se dice que se tiene una hipótesis nula, la cual se representa por H_o. Cualquier otra hipótesis distinta a la hipótesis dada es llamada una hipótesis alternativa (o alterna), la cual se representa por H_1.

La hipótesis nula es una hipótesis simple, mientras que la hipótesis alternativa es compuesta. Esto es lo que en realidad se conoce como la hipótesis de dos colas o bilateral, es decir, son regiones de significación. Hay que recordar de los cursos de estadística que una distribución normal consta de dos colas. En este caso, una hipótesis alterna bilateral refleje el hecho de que no se tiene una fuerte expectativa teórica o *a priori* sobre la dirección en que se debe de mover la hipótesis alternativa partiendo de la hipótesis nula.

Cabe mencionar que en ocasiones existen altas expectativas teóricas o *a priori* de que la hipótesis alterna es de una cola unilateral o unidireccional. En estos casos, la región crítica es una región a un lado de la distribución, con área igual al nivel de significancia.

Considérese de nuevo el intervalo de confianza de la expresión 2.47, se sabe que en el largo plazo, intervalos como (0.6178, 2.3754) contendrán el verdadero valor de β_1 con una probabilidad de 95%, en consecuencia en muestreos repetidos, tales rangos proporcionan un rango en los límites dentro de los cuales puede encontrarse el verdadero valor β_1 con un coeficiente de confianza de 95%. De esta manera, el intervalo de confianza proporciona un

conjunto de hipótesis nulas posibles, esto es si β_1 bajo H_o cae dentro del intervalo de confianza del 100 ($1 - \alpha$) por ciento, se puede aceptar la hipótesis nula, si se encuentra fuera del intervalo se rechaza.

Una prueba de significancia es un procedimiento mediante el cual se utilizan los resultados de la muestra para verificar la veracidad o falsedad de la hipótesis. Esta prueba consiste en utilizar un estadístico de prueba, estimador, y la distribución muestral de ese estadístico bajo la hipótesis nula.

La prueba *t* de student es utilizada para medir la significancia estadística de los parámetros del modelo, es decir, los betas. Recordando que bajo el supuesto de normalidad la variable

$$t = \frac{\hat{\beta}_1 - \beta_1}{se(\hat{\beta}_1)} = \frac{\hat{\beta}_1 - \beta_1 \sqrt{\sum x_i^2}}{\hat{\sigma}} \qquad 2.40$$

sigue la distribución *t* con $N - 2$ g de l, por lo que se establece el intervalo de confianza como sigue:

$$Pr\left[-t_{\alpha/2} \leq \frac{\hat{\beta}_1 - \beta_1^*}{se(\hat{\beta}_1)} \leq t_{\alpha/2} \right] = 1 - \alpha \qquad 2.50$$

donde β_1^* es el valor de β_1 bajo la H_o y $-t_{\alpha/2}$ y $t_{\alpha/2}$ son los valores de *t* que se obtienen de la tabla *t* para el nivel de significancia de ($\alpha/2$) y $N-2$ g de l.

Reordenando 2.50 se obtiene

$$Pr\left[\beta_1^* - t_{\alpha/2} se(\hat{\beta}_1) \leq \hat{\beta}_1 \leq \hat{\beta}_1 + t_{\alpha/2} se(\hat{\beta}_1) \right] = 1 - \alpha \qquad 2.51$$

Esta expresión proporciona el intervalo en el que $\hat{\beta}_1$ se encontrará con una probabilidad de ($1 - \alpha$) por ciento, dado que $\beta_1 = \beta_1^*$. Mediante las pruebas de hipótesis, el intervalo de confianza de 100 ($1 - \alpha$) por ciento establecido en

2.51 se conoce como región de aceptación de la hipótesis nula, mientras que la(s) región(es) fuera del intervalo de confianza se denomina(n) región(es) de rechazo de la hipótesis nula o región(s) crítica(s).

Bajo el procedimiento de intervalos de confianza se trata de establecer los límites dentro de los cuales se puede encontrar β_1 verdadero pero desconocido, en tanto que el enfoque de la prueba de significancia se plantea un valor hipotético para β_1 , y con esto tratar de comprobar $\hat{\beta}_1$ calculado se encuentra dentro de los límites de confianza con respecto al valor en la hipótesis. Al comparar la expresión 2.43 con 2.51 se observa una estrecha relación entre los enfoques de intervalos de confianza y la prueba de significancia en el análisis de pruebas de hipótesis.

Considerando el ejemplo de calificación de examen final-horas de estudio se tiene que $\hat{\beta}_1 = 1.4966$, $se(\hat{\beta}_1) = 0.3591$ y *g de l* = 6. Si se supone que α = 5% $t_{\alpha/2} = 2.447$. Suponiendo que se postula que la hipótesis nula tiene un valor hipotético de $H_0 : \beta_1 = \beta_1^* = 2.7$ y la hipótesis alterna de $H_1 : \beta_1 \neq 2.7$, entonces la ecuación 2.51 se transforma

$$Pr(1.8212 \leq \hat{\beta}_1 \leq 3.5788) = 0.95 \qquad 2.52$$

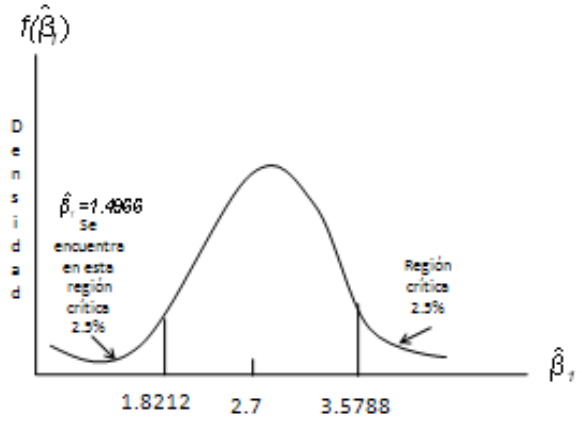


Figura 2.12. Intervalos de confianza del 95% para $\hat{\beta}_1$

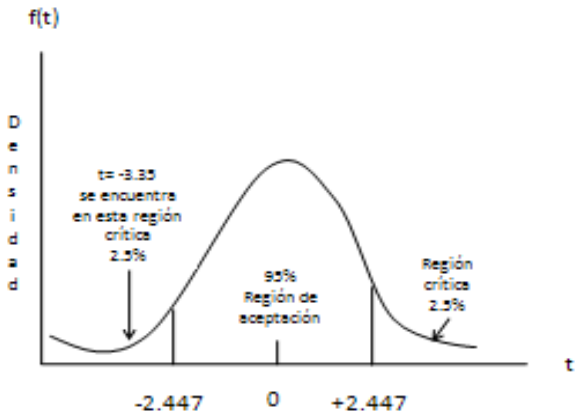


Figura 2.13. Intervalos de confianza del 95 % para t (6 g de l).

Puesto que el valor observado de β_1 se encuentra en la región crítica, se puede rechazar la hipótesis nula de que el verdadero valor de $\beta_1 = 2.7$.

En la práctica no es necesario estimar 2.51. Se puede calcular el valor *t* que aparece en el centro de la doble igualdad de la expresión 2.50 y observar que se encuentra dentro de los valores *t* críticos o fuera de ellos. El valor de *t* se calcula de la siguiente manera:

$$t = \frac{1.4966 - 2.7}{0.3591} = -3.35 \qquad 2.53$$

lo que demuestra que se encuentra dentro de la región crítica de la figura 2.13. Por lo que se concluye que se rechaza la H_0.

Al cálculo anterior se le denomina prueba *t*, que en el caso de las pruebas de significancia, se dice que un estadístico es estadísticamente significativo si el valor del estadístico de prueba se encuentra en la región crítica. Para el ejemplo expuesto, el estadístico *t* es significativo y por tanto se procede a rechazar la hipótesis nula.

Este procedimiento que se acaba de describir se conoce como procedimiento de significancia de dos colas o bilateral, debido a que se consideraron las colas extremas, regiones de rechazo.

En la siguiente tabla se puede resumir la metodología de la prueba *t* de significancia utilizada en las pruebas de hipótesis. La prueba de significancia observa las siguientes reglas de decisión:

Tipo de hipótesis	**H_o: hipótesis nula**	**H_1: hipótesis alterna**	**Reglas de decisión: rechazar la H_o**
De dos colas	$\beta_1 = \beta_1^*$	$\beta_1 \neq \beta_1^*$	$\lvert t \rvert \succ t_{\alpha/2}$, g de l
Cola derecha	$\beta_1 \leq \beta_1^*$	$\beta_1 \succ \beta_1^*$	$t \succ t_\alpha$, g de l
Cola izquierda	$\beta_1 \geq \beta_1^*$	$\beta_1 \prec \beta_1^*$	$t \prec -t_\alpha$, g de l

β_1^* es el valor numérico hipotético de β_1

$\lvert t \rvert$ significa valor absoluto de *t*

t_α ó $t_{\alpha/2}$ significa valor *t* crítico a un nivel de significancia de *α/2*

g de l son grados de libertad, (n – 2) para el modelo de dos variables

El mismo procedimiento es válido para evaluar la hipótesis con respecto a β_0

Tabla 2.2

Continuando con el ejemplo anterior, para la prueba de significancia de σ^2 se considera la siguiente variable:

$$\chi^2 = (N - 2)\ \frac{\hat{\sigma}^2}{\sigma^2} \qquad 2.51$$

donde sigue una distribución χ^2 con *N- 2 g de l*. En el ejemplo hipotético se $\hat{\sigma}^2 = 37.9162$ y los *g de l = 6*, postula que $H_0 : \sigma^2 = 75$ y $H_1 : \sigma^2 \neq 75$ la ecuación 2.51 proporciona el estadístico de prueba para H_0. Al sustituir los valores correspondientes en esta ecuación se encuentra que bajo H_0, $\chi^2 = 3.0333$. Si se supone que α = 5%, los valores críticos de χ^2 son 1.2373 y

14.4494. El valor calculado de χ^2 se encuentra entre estos límites la hipótesis nula se puede aceptar, verificar la figura 2.11. A este procedimiento se le conoce como prueba de significancia ji-cuadrado.

Una prueba de significancia global del modelo, que prueba la hipótesis nula de que el verdadero β_1 es igual a cero, es la prueba estadística *F* de Fisher. Para esto, se debe de proceder a calcular el valor *F* y compararlo con el valor crítico *F* que se obtiene de las tablas *F* para el nivel de significancia que se escoja. Así, la distribución *F* se calcula mediante la siguiente expresión

$$F = \frac{\beta_1^2 \sum x_i^2}{\sum e_i^2 / (N-2)} = \frac{\beta_1^2 \sum x_i^2}{\hat{\sigma}^2} \qquad 2.54$$

Retomando el ejemplo anterior de las calificaciones de los exámenes finales-horas de estudio, se tiene que la razón de *F* de acuerdo con la expresión 2.54 se constituye en un estadístico que sirve para probar la hipótesis nula de que el verdadero β_1 es igual a cero, y para probar se procede a calcular el valor *F* y compararlo con el valor crítico *F* que se obtiene de las tablas *F* (misma que se podrá consultar en tablas estadísticas como distribución *F*) para el nivel de significancia de α = 5%.

El valor calculado de *F* es 17.37. Si α es al 5%, el valor crítico de *F* para 1 y 6 *g de l* es 5.99. El valor de *F* es estadísticamente significativo y, por tanto, se puede proceder a rechazar la hipótesis nula de que las horas de estudio no tienen influencia alguna en las calificaciones del examen al final del periodo de tres semanas.

Si se supone que $H_0 : \beta_1 = 0$, entonces se puede verificar con la ecuación 2.40 que el valor *t* estimado es de 4.167. Este valor *t* tiene 6 *g de l*. Bajo la misma hipótesis nula, el valor *F* fue de 17.37 con 1 y 6 *g de l*. En consecuencia, $(4.167)^2$ = valor *F*, exceptuando los errores de aproximación.

Se concluye entonces que las pruebas *t* y *F* son dos maneras alternas y complementarias para evaluar la hipótesis nula de que $H_0 : \beta_1 = 0$. Hay que tomar en consideración que se puede basar sólo en la prueba *t* sin necesidad de probar la *F*, este es el caso para el modelo de dos variables, pero cuando se trata de una regresión múltiple, la prueba *F* tiene diferentes aplicaciones convirtiéndolo en un método fuerte para evaluar hipótesis estadísticas.

ACTIVIDAD DE APRENDIZAJE

De acuerdo con el ejercicio que realizó en la actividad de la sesión 2.3, tasa salarial-tasa de desempleo, determinar las pruebas de significancia t, ji – cuadrada y la distribución *F*. Asimismo, postular $H_0 : \beta_1 \geq \beta_1^* \geq 0.1$ y $H_1 : \beta_1 \prec \beta_1^* \prec 0.1$. Hacer una interpretación acerca de los datos obtenidos.

2.5 PREDICCIÓN

En la mayoría de los casos, si la relación está bien especificada, no se podrá obtener información suplementaria sobre el modelo, de modo que sólo se debe conformar con un conjunto de estimaciones de los parámetros poco fiables. No obstante, la información estimada sigue siendo satisfactoria para propósitos de predicción. La predicción puede ser individual o media.

Con base en los datos muestrales de la tabla 2.1 se obtiene de la regresión muestral.

$$\hat{Y}_i = 40.0816 + 1.4966 X_i \qquad 2.38$$

donde $\hat{Y}_i$ es el estimador del verdadero $E(Y_i)$ correspondiente a un *X* dado, se puede decir X_0, que es un punto sobre la línea de regresión poblacional misma, verificar la figura 2.4.

Para el caso de la predicción media se retoma el ejemplo de las calificaciones en examen-horas de estudio y para concretar el concepto, se

supone que X_0 = 10 y se quiere predecir $E(Y/X_0 = 10)$. La regresión histórica (2.38) proporciona la estimación puntual de esta predicción media con lo que se obtiene lo siguiente:

$$\begin{aligned} \hat{Y}_0 &= \hat{\beta}_0 + \hat{\beta}_1 X_0 \\ &= 40.0816 + 1.4966(10) \\ &= 55.0476 \end{aligned} \qquad 2.55$$

Puesto que $\hat{Y}_0$ es un estimador, es posible que sea diferente de su verdadero valor. La diferencia entre los dos valores da una idea del error de la predicción o del pronóstico. Para estimar este error se necesita encontrar una distribución muestral de $\hat{Y}_0$. Con la aplicación de la siguiente fórmula es posible demostrar que Y_0 tiene una distribución normal con media $(\beta_0 + \beta_1 X_0)$ y varianza dada por la siguiente fórmula:

$$var(\hat{Y}_0) = \sigma^2 \left[\frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2} \right] \qquad 2.56$$

Al reemplazar el valor desconocido σ^2 por su estimador insesgado $\hat{\sigma}^2$ se tiene la variable

$$t = \frac{\hat{Y}_0(\beta_0 + \beta_1 X_0)}{se(\hat{Y}_0)} \qquad 2.57$$

sigue una distribución *t* con *N – 2 g de l*. De esta manera, se puede encontrar el intervalo de confianza, y así hacer pruebas de hipótesis.

$$Pr\left[\hat{\beta}_0 + \hat{\beta}_1 X_0 - t_{\alpha/2} se(\hat{Y}_0) \le \beta_0 + \beta_1 X_0 \le \hat{\beta}_0 + \hat{\beta}_1 X_0 + t_{\alpha/2} se(\hat{Y}_0) \right] = 1 - \alpha \qquad 2.58$$

$se(\hat{Y}_0)$ se obtiene a partir de la expresión 2.56.

A partir de los datos obtenidos en la tabla 2.1 se determina 2.56.

$$var(\hat{Y}_0) = 37.9162\left[\frac{1}{8} + \frac{(10-24)^2}{294}\right]$$

$$= 30.0169$$

y $$se(\hat{Y}_0) = 5.4788$$

Así, el intervalo de confianza de 95% para el verdadero $E(Y/X_0 = \beta_0 + \beta_1 X_0)$ estará dado por:

$$\left[55.0476 - 2.447(5.4788) \leq E(Y/X_0 = 10) \leq 55.0476 + 2.447(5.4788\right]$$

es decir,

$$\left[41.6409 \leq E(Y/X_0 = 10) \leq 68.4542\right] \qquad 2.59$$

Dado $X_0 = 10$, en muestreos repetidos, 95 de cada 100 intervalos como el de la expresión 2.59 incluirán el verdadero valor promedio. La mejor estimación de este valor medio verdadero es la estimación puntual, *55.0476*. Si se obtienen intervalos de confianza de 95% como el de 2.59 para cada uno de los valores *X* dados en la tabla 2.1 se hallará entonces el intervalo de confianza para la función de regresión poblacional.

Por último, si se quiere predecir un valor individual de *Y* como Y_0 que corresponde a un valor dado de *X* como X_0, es posible probar el mejor estimador lineal insesgado de Y_0 que también está dado por 2.55 pero que su varianza es

$$var(Y_0) = \sigma^2 \left[1 + \frac{1}{N} + \frac{(X_0 - \bar{X})^2}{\Sigma x_i^2} \right] \qquad 2.60$$

Continuando con el mismo ejemplo de calificaciones de examen-horas de estudio, la predicción puntual de Y_0 es de 55.0476, la misma que para $\hat{Y}_0$, y su varianza es 69.5149. Por tanto, el intervalo de confianza de 95% para Y_0 correspondiente a X_0 = 100 está dado por:

$$\left[34.6455 \leq E(Y / X_0 = 10) \leq 75.4497 \right] \qquad 2.61$$

Al comparar el intervalo de la 2.59 con el de la 2.61, se puede apreciar, que el intervalo para Y_0, que este último es más amplio que el del valor medio de Y_0. Esto es porque al calcular los intervalos de confianza de 2.61, condicionados a los valores de X de la tabla 2.1 se tiene un intervalo de confianza de 95% para los valores de X.

Hay diferentes modos de informar los resultados obtenidos de un análisis de regresión, a continuación se hace una descripción del ejemplo de calificaciones de examen-horas de estudio:

$$\begin{array}{llll} \hat{Y}_i = 40.0816 + 1.4966 X_i & & r^2 = 0.7432 & \\ \quad (8.8895) & (0.3591) & g\ de\ l = 6 & 2.62 \\ t = (4.5088) & (4.1674) & F_{1,6} = 17.37 & \end{array}$$

Las cantidades que aparecen en la primera serie de paréntesis corresponden a los errores estándar de los diferentes coeficientes de regresión, en tanto que los coeficientes de la segunda serie de paréntesis corresponden a los valores t, estimados a partir de la ecuación 2.40 bajo la hipótesis nula de que el verdadero valor poblacional de cada coeficiente individual de regresión es cero (4.5088 = 40.0816 ÷ 8.8895).

Esta forma de presentación del informe permite verificar rápidamente si cada uno de los coeficientes de regresión es individualmente significativo desde el punto de vista estadístico.

En el caso del valor de *F*, éste sólo refuerza el estadístico *t* en la prueba de hipótesis de que el verdadero coeficiente de la pendiente de β_1 es cero.

ACTIVIDAD DE APRENDIZAJE

De acuerdo con la información proporcionada en la tabla, de la actividad de aprendizaje de la sesión 2.3, efectuar la predicción individual y media. Asimismo efectuar un informe de los resultados obtenidos del análisis de regresión.

AUTOEVALUACIÓN

Con base en los datos hipotéticos que se presentan en la siguiente tabla, efectuar los cálculos de los estimadores y determinar la línea de regresión estimada. Asimismo, establecer los intervalos de confianza para β_0, β_1 y σ^2. Calcular la prueba *t*, ji-cuadrada y la distribución *F* con un nivel de significancia del 5%.

Y_i	X_i	$X_i - \bar{X}_i = x_i$	$Y_i - \bar{Y}_i = y_i$	$x_i y_i$	y_i^2	x_i^2	X_i^2	$\hat{Y}_i$	e_i	e_i^2
2.8	21									
3.4	24									
3	26									
3.5	27									
3.6	29									
3	25									
2.7	25									
3.7	30									

Respuesta

Y_i	X_i	$X_i - \bar{X}_i = x_i$	$Y_i - \bar{Y}_i = y_i$	$x_i y_i$	y_i^2	x_i^2	X_i^2	$\hat{Y}_i$	e_i	e_i^2
2.8	21	-4,875	-0,4125	2,0109375	0,17015625	23,765625	441	2,71428571	0,085714286	0,00734694
3.4	24	-1,875	0,1875	-0,3515625	0,03515625	3,515625	576	3,02087912	0,379120879	0,14373264
3	26	0,125	-0,2125	-0,0265625	0,04515625	0,015625	676	3,22527473	-0,22527473	0,0507487
3.5	27	1,125	0,2875	0,3234375	0,08265625	1,265625	729	3,32747253	0,172527473	0,02976573
3.6	29	3,125	0,3875	1,2109375	0,15015625	9,765625	841	3,53186813	0,068131868	0,00464195
3	25	-0,875	-0,2125	0,1859375	0,04515625	0,765625	625	3,12307692	-0,12307692	0,01514793
2.7	25	-0,875	-0,5125	0,4484375	0,26265625	0,765625	625	3,12307692	-0,42307692	0,17899408
3.7	30	4,125	0,4875	2,0109375	0,23765625	17,015625	900	3,63406593	0,065934066	0,0043473
				5,8125	**1,02875**	**56,875**	**5413**	**25,7**		**0,43472527**

$\hat{\beta}_0 = 0.5681$ $\quad var(\hat{\beta}_0) = 0.8619$ $\quad se(\hat{\beta}_0) = 0.9284$

$\hat{\beta}_1 = 0.1022$ $\quad var(\hat{\beta}_1) = 0.0013$ $\quad se(\hat{\beta}_1) = 0.3569$

$cov(\hat{\beta}_0, \hat{\beta}_1) = -0.0329$ $\hat{\sigma}^2 = 0.7245$ $g\ de\ l = 6$

$r^2 = 0.5774$

$r = 0.7599$

$$\hat{Y}_i = 0.5681 + 0.1022X_i$$

Intervalos de confianza para β_0, β_1

$$0.0148 \leq \beta_1 \leq 0.1895$$

$$-1.7037 \leq \beta_0 \leq 2.8399$$

Intervalos de confianza para σ^2

$$0.0301 \leq \sigma^2 \leq 0.3534$$

Comentario a tener en cuenta

UNIDAD 3

HETEROCEDASTICIDAD

OBJETIVO

El Estudiante distinguirá cuáles son algunas de las causas de la heterocedasticidad en los modelos de MCO, identificará su detección y corrección.

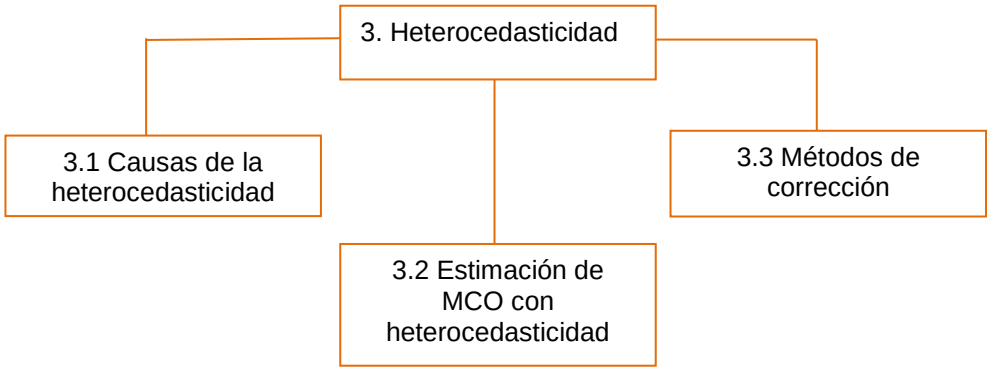
TEMARIO

3.1 CAUSAS DE LA HETEROCEDASTICIDAD

3.2 ESTIMACIÓN DE MCO CON HETEROCESASTICIDAD

3.3 MÉTODOS DE CORRECCIÓN

MAPA CONCEPTUAL



INTRODUCCIÓN

El modelo clásico de regresión lineal es en el que los términos de perturbación u_i tienen todos la misma varianza. Si no se cumple este supuesto, se presenta el fenómeno de heterocedasticidad.

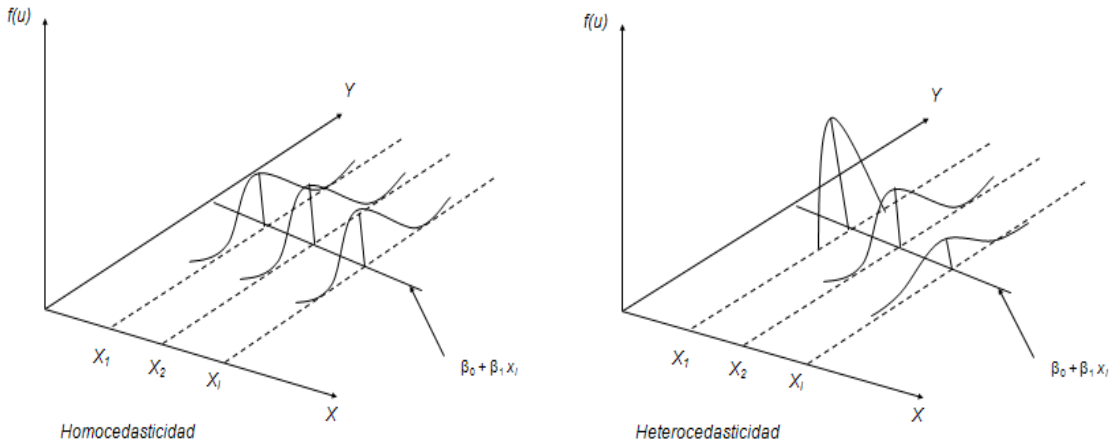
En el apartado 3.1 se analizan algunas de las causas de la heterocedasticidad en el modelo de mínimos cuadrados ordinarios, los cuales en el modelo original muestran varianzas constantes.

Bajo los estimadores originales del modelo de MCO se hace una transformación del modelo en presencia de heterocedasticidad, en la que se aplica el método de mínimos cuadrados generalizados que es equivalente al método de cuadrados ponderados, a los cuales se les considera que son MELI.

Por último, para la detección y corrección de la heterocedasticidad se verifican dos métodos, informal y formal, el gráfico y el de Goldfeld-Quandt.

3.1 CAUSAS DE LA HETEROCEDASTICIDAD

El supuesto de homocedasticidad, introducido en la Unidad 2, plantea que la varianza de las perturbaciones u_i, de la función de regresión poblacional, tienen la misma varianza. Esta homocedasticidad se necesita para justificar las pruebas *t* y *F* y los intervalos de confianza para la estimación de los mínimos cuadrados ordinarios (MCO) del modelo de regresión lineal.



La heterocedasticidad es la existencia de una varianza no constante en las perturbaciones de un modelo de regresión lineal. Formalmente esto se escribe en símbolos como sigue:

$$E\left(u_i^2\right)=\sigma_i^2$$

El subíndice de σ^2, indica que las varianzas condicionales de u_i (= varianza condicional de Y_i) no continúan siendo constantes. Ante esto, existen múltiples causas por las cuales se explica la variación en las varianzas. Sólo se mencionarán las causas que se consideran más básicas.

- Aparece porque ha omitido una variable relevante en el modelo especificado. Cuando se ha omitido una variable en la especificación, ésta quedará parcialmente recogida en el comportamiento de las perturbaciones aleatorias. Antes de realizar

el análisis de heterocedasticidad es importante verificar que se han incluido en el modelo todas las variables relevantes, pues si se ha omitido una variable importante en el modelo, puede dar como resultado, estimadores sesgados y varianzas no eficientes.

- Procede de la naturaleza de los datos. Los datos de los cuales se dispone, puede ser que estén ordenados por agentes o unidades económicas (datos de panel o longitudinales) y hacen referencia a la existencia de dos dimensiones en los datos. Esto significa que los datos de los cuales se disponen pueden ser agrupados, agregados o promediados sobre un conjunto de individuos, empresas, sectores, por lo cual pueden poseer características individuales, dando lugar a diferente variabilidad. Ante esto, se puede esperar distinta dispersión.
- El tamaño de las unidades que se comparan. Las unidades pueden ser familias, empresas o países. Por ejemplo, si se compara el nivel de ingreso y se tienen dos tipos de familias, una con altos ingresos y otra con bajos ingresos, se esperaría que la mayor variabilidad la presentará la familia con altos ingresos, pues es de suponer que la familia de bajos ingresos ya cuenta con cierto nivel de consumo para cubrir sus necesidades básicas. Frente a esto, se podría esperar que la varianza de *Y* aumentará con el tamaño o magnitud del ingreso familiar. Por lo que se sugiere proponer para este modelo perturbaciones heterocedásticas, en la explicación de los agentes económicos, estas perturbaciones pueden ser gustos, características individuales, familiares, etcétera, lo que permite explicar por qué éstas pueden mostrar una variabilidad.
- Modelos de aprendizaje por error. Un ejemplo es cuando una persona comete errores, en la medida que aprende, se puede decir que sus errores de comportamiento disminuyen, por lo que puede esperarse que la varianza tienda a disminuir. En otra

situación, puede ser que un estudiante que destina tiempo de estudio para una asignatura, entre mayor sea el lapso de dedicación, los resultados en un examen pueden ser favorables al disminuir el número de errores y, por ende, la varianza.

- La forma de la función que sea aplicada de manera incorrecta. En este caso, puede ser que se utilice una función lineal en lugar de una logarítmica potencial, esto provoca que la calidad de la regresión varíe según los valores de la variable exógena. Esto es, que se ajusten bien el valor pequeño y mal los valores grandes, es decir, que en las zonas de peor ajuste existan no sólo errores mayores, sino que también éstos estén más dispersos.
- La distribución de las variables explicativas, es decir, como se encuentran los datos alejados de la media, puede ser que se encuentren alejados a la derecha de la media (o a la izquierda de la media). Esto conduce a que se dé una transformación en las variables para corregir el problema y hacer que éstas se hallen de manera uniforme alrededor de la media.

En todo caso, cual sea el origen del problema, en muchas cuestiones es posible asociar la varianza no constante de las perturbaciones aleatorias a los valores de alguna de las variables incluidas en el modelo. En este sentido, cuando se da la existencia de la heterocedasticidad y no hay una explicación estimable para la misma, resulta a menudo de utilidad someter a los datos a algunas transformaciones sencillas que tiendan a estabilizar la varianza.

El problema de heterocedasticidad se presenta principalmente en modelos de corte transversal, y de igual forma en observaciones con series de tiempo. De esta manera, la heterocedasticidad puede presentarse debido a cómo se especificó el modelo, a cómo se obtuvo la información y de las decisiones en relación al tratamiento de los datos.

Las varianzas condicionales que no son constantes en el modelo aparecen cuando la combinación lineal de todos los regresores generan errores, manifestándose con esto heterocedasticidad en el modelo.

ACTIVIDAD DE APRENDIZAJE

Investigar otras causas de la presencia de heterocedasticidad en los modelos de regresión lineal, explicar cada una, en hojas blancas para entregar en la siguiente sesión.

3.2 ESTIMACIÓN DE MCO CON HETEROCEDASTICIDAD

En la Unidad 2 se obtuvo la estimación de mínimos cuadrados ordinarios y se consideraron bajo los supuestos de Guss-Marcov que los estimadores son el mejor estimado lineal insesgado (MELI), donde a la varianza mínima se le conoce como un estimador eficiente.

De esta manera, nuevamente regresamos al modelo de dos variables, se suponen válidas las suposiciones de Guss-Marcov.

$$Yi = \beta_0 + \beta_1 X_i + u_i$$

Donde el estimador de MCO para β_1 es igual a:

$$\hat{\beta}_1 = \frac{\Sigma x_i y_i}{\Sigma x_i^2} = \frac{\Sigma x_i Y_i}{\Sigma X_i^2 - N\bar{X}^2} = \frac{\Sigma X_i y_i}{\Sigma X_i^2 - N\bar{X}^2} \qquad 3.1$$

Su varianza está dada ahora por

$$Var\,(\hat{\beta}_1) = \frac{\Sigma x_i^2 \sigma_i^2}{\left(\Sigma x_i^2\right)^2} \qquad 3.2$$

La cual es diferente a la fórmula de varianza de β_1 obtenida bajo el supuesto de homocedasticidad y que es igual a

$$Var\,(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma x_i^2} \qquad 3.3$$

Debido a que $\sigma_i^2 = \sigma^2$ para cada i, por lo que las dos fórmulas son idénticas.

Recordando, se tiene que $\hat{\beta}_1$ es el mejor estimador lineal insesgado si se cumplen los supuestos del modelo clásico, incluyendo el de homocedasticidad. Cuando se elimina el supuesto de homocedasticidad por el supuesto de heterocedasticidad $\hat{\beta}_1$ no continúa siendo el mejor estimador y la varianza mínima no está dada por 3.2, por lo que surge la conveniencia de buscar estimadores alternativos que verifiquen mejores propiedades que los de MCO. Este es el caso de los estimadores de mínimos cuadrados generalizados (MCG). Para ver cómo se logra esto, se continúa con el ya conocido modelo de dos variables.

$$Y_i = \beta_0 + \beta_1 X_i + u_i \qquad 3.4$$

Para facilitar el manejo algebraico, se expresa de la siguiente manera:

$$Y_i = \beta_0 X_{0i} + \beta_1 X_i + u_i \qquad 3.5$$

donde $X_{0i} = 1$ para cada i. La expresión 3.4 y 3.5 son idénticas.

Suponiendo que se conocen las varianzas heterocedásticas σ_i^2, se divide 3.5 entre σ_i, se obtiene

$$\frac{Y_i}{\sigma_i} = \beta_0\left(\frac{X_{0i}}{\sigma_i}\right) + \beta_1\left(\frac{X_i}{\sigma_i}\right) + \left(\frac{u_i}{\sigma_i}\right) \qquad 3.6$$

simplificando se tiene la siguiente forma:

$$Y_i^* = \beta_0^* X_{0i}^* + \beta_1^* X_i^* + u_i^* \qquad 3.7$$

la notación β_0^* y β_1^* son los parámetros transformados para distinguirlos de los parámetros normales de MCO β_0 y β_1. Donde el término de error de transformación u_i^*:

$$\begin{aligned} var\left(u_i^*\right) = E\left(u_i^*\right)^2 &= E\left(\frac{u_i}{\sigma_i}\right)^2 \\ &= \frac{1}{\sigma_i^2}E\left(u_i^2\right) \text{ dado } \sigma_i^2 \text{ se conoce} \\ &= \frac{1}{\sigma_i^2}\left(\sigma_i^2\right) \text{ dado } E\left(u_i^2\right) = \sigma_i^2 \\ &= 1 \end{aligned} \qquad 3.8$$

se tiene que es una constante, por lo que la varianza del término de perturbación transformado u_i^* es ahora homocedástica, lo que indica que si se debe de aplicar el método de MCO al modelo transformado 3.6 y así obtener estimadores que sean MELI. De igual forma los β_0^* y β_1^* estimados serán ahora MELI, a pesar de que los estimadores de MCO $\hat{\beta}_0$´y $\hat{\beta}_1$ no lo sean.

El procedimiento de transformar las variables originales para que se satisfagan los supuestos del modelo clásico y de aplicar a continuación MCO se conoce como el método de mínimos cuadrados generalizados (MCG), así los

estimadores obtenidos de esta manera se conocen como estimadores MCG, los cuales son MELI.

Para estimar $\hat{\beta}_0^*$ y $\hat{\beta}_1^*$ se hace uso de la función de regresión muestral aplicada a 3.6

$$\frac{Y_i}{\sigma_i} = \hat{\beta}_0^* \left(\frac{X_{0i}}{\sigma_i} \right) + \hat{\beta}_1^* \left(\frac{X_i}{\sigma_i} \right) + \left(\frac{e_i}{\sigma_i} \right)$$

$$Y_i^* = \hat{\beta}_0^* X_{0i}^* + \hat{\beta}_1^* X_i^* + e_i^* \qquad 3.9$$

Así, para obtener los estimadores de MCG se determina la siguiente expresión

$$\Sigma e_i^{*2} = (Y_i^* - \hat{\beta}_0^* X_{0i}^* - \hat{\beta}_1^* X_i^*)^2 \qquad 3.10$$

o bien

$$\Sigma \left(\frac{e_i}{\sigma_i} \right)^2 = \left[\left(\frac{Y_i}{\sigma_i} \right) - \hat{\beta}_0^* \left(\frac{X_{0i}}{\sigma_i} \right) - \hat{\beta}_1^* \left(\frac{X_i}{\sigma_i} \right) \right]^2 \qquad 3.11$$

bajo ciertas técnicas de cálculo que se aplican para obtener 3.11, el estimador de MCG para β_1^* es

$$\hat{\beta}_1^* = \frac{(\Sigma w_i)(\Sigma w_i X_i Y_i) - (\Sigma w_i X_i)(\Sigma w_i Y_i)}{(\Sigma w_i)(\Sigma w_i X_i^2) - (\Sigma w_i X_i)^2} \qquad 3.12$$

y su varianza está dada por

$$var(\hat{\beta}_1^*) = \frac{(\Sigma w_i)}{(\Sigma w_i)(\Sigma w_i X_i^2) - (\Sigma w_i X_i)^2} \qquad 3.13$$

donde $w_i = 1 / \sigma_i^2$.

Con la técnica de MCO se minimiza la expresión

$$\Sigma e_i^2 = (Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i)^2 \qquad 3.13$$

mientras que en MCG se minimiza la expresión 3.11, la cual se puede expresar como

$$\Sigma w_i e_i^2 = \Sigma w_i (Y_i - \hat{\beta}_0^* - \hat{\beta}_1^* X_i)^2 \qquad 3.14$$

donde $w_i = 1 / \sigma_i^2$, siendo idénticas las expresiones 3.11 y 3.14.

En MCG se minimizó la suma ponderada de los residuos al cuadrado, donde w_i representa las ponderaciones. A la expresión 3.14 se le conoce como técnica de mínimos cuadrados ponderados (MCP) y los estimadores. A los estimadores obtenidos de esta manera y dados en las expresiones 3.12 y 3.13 se conocen como estimadores de MCP, siendo éstos un caso especial de una técnica de estimación más general conocida como MCG. En presencia de heterocedasticidad se pueden utilizar indistintamente los dos términos MCP y MCG.

El método de MCG pondera cada residuo cuadrado por medio de la inversa de la varianza condicional de u_i dada X_i esto es para cualquier conjunto de ponderaciones positivas. Para realizar MCP y obtener variables transformadas, se puede realizar de manera efectiva con paquetes de regresión.

ACTIVIDAD DE APRENDIZAJE

Mencionar qué otra paquetería hay, además de la ya mencionada en la introducción del libro, y que es la apropiada para la detección y estimación de

los MCO, en presencia de heterocedasticidad. Dar una explicación a lo obtenido. Entregar en hojas blancas en la siguiente sesión.

3.3 MÉTODOS DE CORRECCIÓN

Hasta el momento, sólo se ha demostrado cómo se lleva a cabo la estimación de MCO con la presencia de heterocedasticidad. Antes de referirse a los métodos de corrección, es conveniente mencionar uno de los diversos contrastes para la determinación de la heterocedasticidad.

Un modelo de regresión lineal ajustado supone la detección de heterocedasticidad, la cual puede ser probada por exámenes informales, como las gráficas, y por análisis formales, como el de Goldfeld y Quandt .

El método gráfico permite observar si los errores varían con algunos regresores, como *X* que se encuentra en el eje de las abscisas, en tanto que en el eje de las ordenadas se coloca *Y* y/o los residuos al cuadrado (e_i^2 ,u_i^2).

Ahora bien, cuando no hay información a priori o empírica acerca de la heterocedasticidad, se puede llevar a cabo el análisis de regresión bajo el supuesto de que no existe heterocedasticidad, y luego realizar un examen posterior (post mortem) de los residuos estimados al cuadrado e_i^2 para ver si presentan algún patrón sistemático. Aunque e_i^2 y u_i^2 no son la misma cosa, se puede utilizar como aproximación. Al examinar los e_i^2 se pueden encontrar patrones como los que aparecen en la figura 3.1.

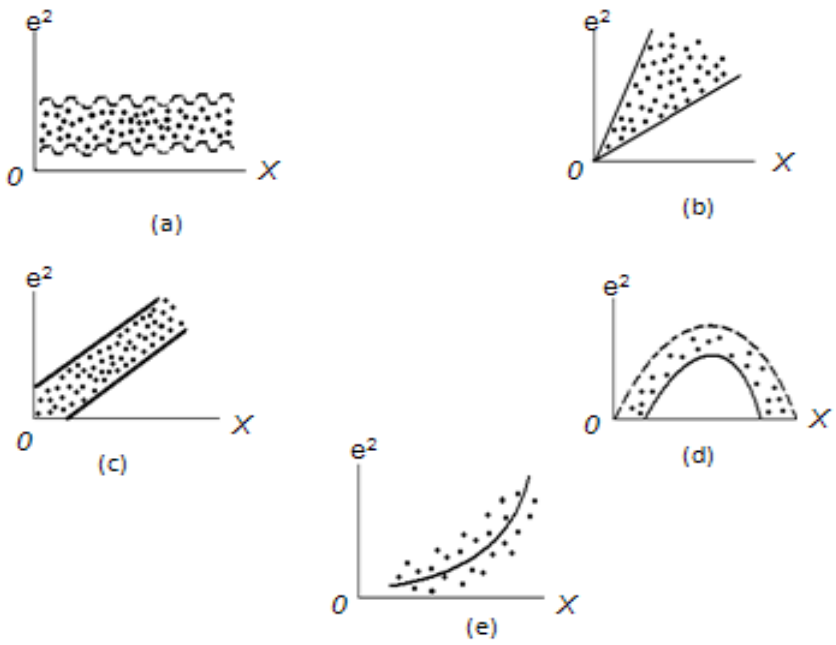


Figura 3.1.

Diagrama de dispersión de los residuos estimados.

Este tipo de comportamiento que se presentan en la figura 3.1, regularmente se observa en las series económicas, las cuales pueden mostrar cierta conducta, es decir, pueden presentar tendencias negativas o positivas de acuerdo con las condiciones de la propia economía de un país, estas condiciones se dan en el tiempo. Asimismo, esto se refleja de la suma de los errores, siendo significativo en el comportamiento de las de las variables que se encuentran en el modelo.

En la figura 3.1 los e_i^2 se grafican contra la variable *X*, incluida en el modelo. En el caso de la figura 3.1 *a* se puede observar que no hay un patrón sistemático entre las dos variable, lo cual sugiere la inexistencia de heterocedasticidad en la información. En tanto, las figuras 3.1 *b* a *e* presentan patrones definidos. Por ejemplo, la figura 3.1 *c* presenta patrones definidos de que la varianza del término de perturbación está linealmente relacionada con la variable *X*, en tanto que la figura 3.1 *d* y *e* indican una relación cuadrática entre e_i^2 y la variable *X*.

Cabe mencionar que la prueba gráfica resulta insuficiente debido a la combinación lineal de algunas o todas las variables, por lo que es posible no detectar la heterocedasticidad. De ahí que sea necesario aplicar pruebas formales para su detección.

Otro contraste para detectar la heterocedasticidad, y que sugiere ya un método formal, es el de Goldfeld y Quandt. Este contraste es válido para cualquier tamaño muestral, con la ventaja de que puede guiar sobre la forma de heterocedasticidad presente.

El contraste se basa en separar las observaciones muestrales en dos grupos, realizar dos regresiones separadas para cada uno y comparar las varianzas estimadas en ambos grupos, lo que equivale a considerar dos muestras, tomadas de dos poblaciones, que además de tener medias diferentes pueden diferir en la varianza. Los pasos a seguir para realizar esta prueba son los descritos a continuación:

- Se ordena la variable dependiente en forma decreciente, también se pueden utilizar las variables independientes.
- De la muestra de datos, se divide en tres estratos o subgrupos iguales, siendo la muestra superior la que contiene los valores más grandes y la inferior la que contiene los datos pequeños de la regresión. La muestra central se utiliza al efectuar un análisis subsecuente.
- De la submuestra superior se estima la regresión y con los residuos se calcula la suma al cuadrado de estos (SCR1), lo mismo se hace con el estrato inferior, SCR2. De esta manera, se efectúa la prueba de contraste *F*, bajo la hipótesis nula de homocedasticidad, con m_1-k_1 grados de libertad en el numerador y m_2-k_2 grados de libertad en el denominador. Donde m_1 es el número de observaciones de la submuestra superior, k_1 es el número de parámetros de la submuestra superior, m_2 es el de observaciones de la submuestra inferior, k_2 es el de parámetros de la submuestra inferior. Una vez efectuado esto se

verifica el valor en tablas de *F*, si este valor resulta menor al calculado, se rechaza la hipótesis nula de homocedasticidad.

- Finalmente se calcula el estadístico *F*.

$$F = \frac{SRC2 / m_2 - k_2}{SRC1 / m_1 - k_1}$$

el cual bajo hipótesis nula de homocedasticidad, seguirá la distribución *F* de *Snedecor* indicada. Si el modelo es homocedástico, el valor de ese cociente de ese cociente, no debe separarse de 1, por lo que si se aleja de dicho valor significativamente, es decir, si supera el valor crítico dado por la distribución, se rechazará la hipótesis nula una vez elegido el nivel de significación del contraste.

Si no se rechaza la hipótesis nula, no significa que no exista heterocedasticidad, dado que esta podría estar asociada a otra variable, por lo que hay que repetir el proceso con otras variables.

En algunos casos esta prueba de contraste no arroja resultados claros, ya que puede ser afectada por la forma funcional del modelo y por el tipo de distribución de los datos.

En repetidas ocasiones ya se ha mencionado que la heterocedasticidad viene producida por la dependencia de la varianza de las perturbaciones aleatorias de una o más variables que, a su vez, pueden estar presentes en el modelo o no. Los distintos métodos de detectar este problema servían para probar, en el caso en el que ésta realmente se presentará.

Si las varianzas con heterocedasticidad σ_i^2 se conocen, el método más directo para resolver el problema consiste en utilizar la técnica de MCP, que minimizan la importancia de las perturbaciones con valores externos ponderándolas en proporción inversa a sus varianzas.

Se han creado algunos métodos informales y de aproximación para detectar la presencia de la heterocedasticidad, los cuales examinan los residuos

obtenidos del procedimiento de MCO normales y así poder sugerir maneras de transformar el modelo original, de tal manera que en la ecuación transformada las perturbaciones tengan una varianza constante.

ACTIVIDAD DE APRENDIZAJE

Investigar otras pruebas de corrección de la heterocedasticidad y explicar cada una de ellas, Entregar en hojas blancas en la siguiente sesión.

AUTOEVALUACIÓN

I. Relacionar las siguientes columnas e indicar en el paréntesis la respuesta que corresponde a la afirmación.

1. Si la heterocedasticidad se origina en la combinación lineal de todas o de algunas de las variables incluidas, donde el ensayo será insuficiente, no podrá ser detectada. 2. El procedimiento de transformar las variables originales para que se satisfagan los supuestos del modelo clásico y de aplicar a continuación MCO se conoce como… 3. Es la que supone que ya se ha ajustado el modelo de regresión lineal. 4. Es aquella en que se requiere una transformación de las variables para corregir tales asimetrías. 5. Cómo es considerado el contraste de Goldfeld y Quandt para detectar la heterocedasticidad. 6. Es la existencia de una varianza no constante en las perturbaciones de un modelo de regresión lineal. 7. Este contraste se basa en separar las observaciones muestrales en dos grupos, realizar dos regresiones separadas para cada uno. 8. Plantea que la varianza de las perturbaciones u_i, de la FRP, tienen la misma varianza. 9. Cuál es la técnica que resuelve el problema en que las varianzas con heterocedasticidad σ_i^2 se conocen 10. Son aquellas que pueden mostrar cierta conducta, es decir tendencias negativas o positivas de acuerdo a las condiciones de un país.	( ) Series económicas ( ) Detección de hetereocedasticidad ( ) MCP ( ) Método formal ( ) Heterocedasticidad ( ) Prueba gráfica ( ) Goldfeld y Quandt ( ) Causa de heterocedasticidad ( ) Supuesto de homocedasticidad ( ) MCG

Respuestas

I. Relacionar las siguientes columnas e indicar en el paréntesis la respuesta que corresponde a la afirmación.

1. Si la heterocedasticidad se origina en la combinación lineal de todas o de algunas de las variables incluidas, donde el ensayo será insuficiente, no podrá ser detectada. 2. El procedimiento de transformar las variables originales para que se satisfagan los supuestos del modelo clásico y de aplicar a continuación MCO se conoce como… 3. Es la que supone que ya se ha ajustado el modelo de regresión lineal. 4. Es aquella en que se requiere una transformación de las variables para corregir tales asimetrías. 5. Cómo es considerado el contraste de Goldfeld y Quandt para detectar la heterocedasticidad. 6. Es la existencia de una varianza no constante en las perturbaciones de un modelo de regresión lineal. 7. Este contraste se basa en separar las observaciones muestrales en dos grupos, realizar dos regresiones separadas para cada uno. 8. Plantea que la varianza de las perturbaciones u_i, de la FRP, tienen la misma varianza. 9. Cuál es la técnica que resuelve el problema en que las varianzas con heterocedasticidad σ_i^2 se conocen 10. Son aquellas que pueden mostrar cierta conducta, es decir tendencias negativas o positivas de acuerdo a las condiciones de un país.	(10) Series económicas (3) Detección de hetereocedasticidad (9) MCP (5) Método formal (6) Heterocedasticidad (1) Prueba gráfica (7) Goldfeld y Quandt (4) Causa de heterocedasticidad (8) Supuesto de homocedasticidad (2) MCG

UNIDAD 4

AUTOCORRELACIÓN

OBJETIVO

El estudiante distinguirá cuáles son algunas de las causas de la autocorrelación en los modelos de MCO, identificará su detección y corrección.

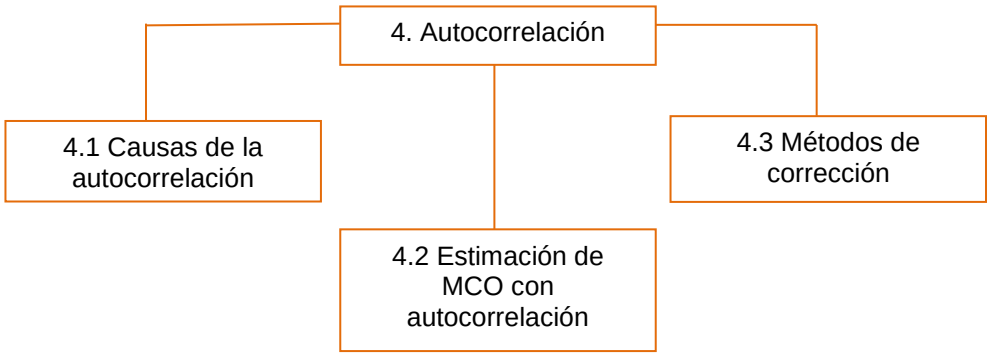
TEMARIO

4.1 CAUSAS DE LA AUTOCORRELACIÓN

4.2 ESTIMACIÓN DE MCO CON AUTOCORRELACIÓN

4.3 MÉTODOS DE CORRECCIÓN

MAPA CONCEPTUAL



INTRODUCCIÓN

En el modelo clásico de regresión lineal en el que los errores y perturbaciones u_i entran en la función de regresión poblacional, se encuentran bajo el supuesto de que son aleatorios o no correlacionados, cuando se viola este supuesto es porque existe autocorrelación o correlación serial.

En la presente Unidad se verifican cuáles son las causas de la autocorrelación y los nombres que recibe ésta, de acuerdo con la serie de datos que se utilicen, pueden ser datos de corte transversal, o los más usuales que son las series temporales.

Asimismo, se analiza la estimación de mínimos cuadrados ordinarios, pues a pesar de que estos estimadores continúan siendo insesgados y consistentes, dejan de ser eficientes en presencia de autocorrelación. Puesto que las perturbaciones no se pueden observar, en la práctica se asume que pueden ser generadas por algún mecanismo factible, por lo que se utiliza el esquema autorregresivo de primer orden de Markov, el cuál se plantea en la segunda parte de la Unidad.

Por último, se examina una prueba de contraste para detectar la autocorrelación, el estadístico Durbin-Watson, para modelar el comportamiento de las perturbaciones. Lo cual exige un conjunto de suposiciones para el modelo lineal clásico y se hace énfasis en sus limitaciones.

4.1 CAUSAS DE LA AUTOCORRELACIÓN

La autocorrelación es la correlación entre los términos de error de un modelo de regresión. Su efecto es que invalida uno de los supuestos que fundamentan el procedimiento de mínimos cuadrados ordinarios y, por lo tanto, hace necesario una modificación de tal procedimiento. Con frecuencia se le llama también correlación serial, ambos términos se utilizan de manera indistinta.

El modelo de regresión serial supone que tal autocorrelación no existe en las perturbaciones u_i, lo cual se expresa

$$E(u_i, u_j) = 0 \qquad i \neq j \qquad 2.19$$

El modelo clásico supone que el término de perturbación asociado a alguna observación no está influenciado por el término de perturbación asociado a cualquier otra observación. La autocorrelación se presenta entre dos series de tiempo tales como $u_1, u_2, \ldots, u_{10}$ $u_2, u_3, \ldots, u_{11}$, en donde la primer serie se encuentra rezagada en un periodo. En tanto, una correlación serial se presenta en dos series de tiempo diferente como son: $u_1, u_2, \ldots, u_{10}$, y $v_2, v_3, \ldots, v_{11}$. Tintner define a la autocorrelación como: "La autocorrelación es una correlación de rezagos de una serie dada consigo misma, rezagada en un número de unidades de tiempo, mientras que el término de correlación serial para la correlación de rezagos entre dos series diferentes."[7]

En datos de serie temporal es común la presencia de la autocorrelación o correlación serial de las perturbaciones, esto se debe a algún fallo en la especificación del modelo. A continuación se mencionan algunas de las causas de autocorrelación:

- La existencia de tendencias y ciclos en los datos, esto se puede observar en la mayoría de las variables económicas que tienden a presentar cierta inercia. Por ejemplo, en un momento de recesión de la economía, las

[7] Gerhard Tintner, *Econometrics*, p. 187.

variables se encuentran en bajos niveles, al surgir un proceso de recuperación o activación de la economía, estas variables comienzan a moverse hacia arriba.

Así, con el movimiento el valor de una serie en un punto en el tiempo es mayor al valor anterior. Esto indica que la mayoría de las variables económicas no son estacionarias, significa que si la variable endógena del modelo tiene una tendencia creciente o muestra un comportamiento cíclico que no es explicado por las exógenas, el término de error recogerá ese ciclo o tendencia. Por lo tanto, en las regresiones con series de tiempo, es probable que las observaciones sean interdependientes.

- El caso de variables explicativas omitidas. Puede ser que en un modelo planteado originalmente se haya excluido una variable que se consideró no era candidata a entrar en el modelo, y al realizar una revisión de sus resultados, se da cuenta que esas variables debieron de haber sido incluidas para eliminar el patrón de correlación observado entre los residuos. La omisión de variables relevantes provoca sesgos en el estimador.
- La especificación incorrecta de la forma funcional del modelo. Esto es que cuando al describir observaciones se aplica un modelo lineal, cuando esto pudo haber sido de manera cuadrática, con esto, los residuos muestran comportamientos no aleatorios, es decir, están correlacionados. Puede ser que los residuos presenten lapsos positivos (o negativos), seguidos de lapsos negativos (o positivos) y así sucesivamente. Aquí la causa del problema en los residuos es un error de especificación en la forma funcional.

El problema de autocorrelación es más común en datos de series de tiempo, es decir, observaciones en diferentes periodos. Los datos temporales presentan una ordenación natural (ayer, hoy, mañana y cada uno antecede al otro) por lo que al considerar dos perturbaciones sucesivas, se sabe que se

hace referencia a dos periodos sucesivos. Aunque la autocorrelación predomina en las serie de tiempo, se puede presentar también en los datos de corte transversal, denominada como autocorrelación espacial.

En el caso de la autocorrelación espacial o correlación en el espacio, el ordenamiento de los datos debe de tener cierto orden lógico o económico, un ejemplo puede ser el sueldo de un grupo de individuos tomados de cierta población y compararlo con el nivel de educación con el que cuentan. Puesto que es posible que los niveles de sueldo difieran de un individuo a otro, los residuos estimados de regresión pueden presentar un patrón sistemático asociado con los diferentes niveles educativos. Ante esto, la correlación también se puede presentar con datos de corte transversal.

Es importante mencionar que la autocorrelación puede ser positiva o negativa, aunque en el caso de series temporales, y con la aplicación de variables económicas, se puede presentar una autocorrelación positiva, debido a que se mueven hacia arriba durante periodos prolongados. Véase la figura 4.1 en que la gráfica (a) tiene una tendencia positiva, (b) una autocorrelación negativa.

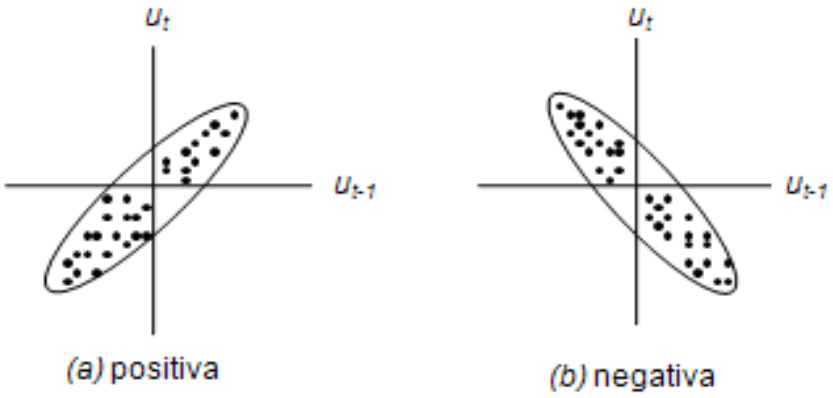


Figura 4.1.

Autocorrelación positiva y negativa.

ACTIVIDAD DE APRENDIZAJE

Investigar otras causas de la presencia de autocorrelación en los modelos de regresión lineal, explicar cada una, en hojas blancas para entregar en la siguiente sesión.

4.2 ESTIMACIÓN DE MCO CON AUTOCORRELACIÓN

Nuevamente se considera el modelo de regresión con dos variables para presentar las ideas básicas del presente análisis, considerando $Y_t = \beta_0 + \beta_1 X_t + u_t$, donde t corresponde a los datos u observaciones en el periodo *t*, en este caso se utilizan series de tiempo, tema que se tratará con mayor profundidad en la Unidad 6, pero que, por el momento, sirve para explicar la autocorrelación.

Como una aproximación se puede asumir que las perturbaciones se generan de la siguiente manera:

$$u_t = \rho u_{t-1} + \varepsilon_t \qquad -1 \prec \rho \prec 1 \qquad 4.1$$

donde ρ se conoce como coeficiente de autocovarianza y ε_t es la perturbación estocástica que satisface los supuestos de MCO tradicionales, es decir,

$$\begin{aligned} E(\varepsilon_t) &= 0 \\ var(\varepsilon_t) &= \sigma^2 \\ cov(\varepsilon_t, \varepsilon_{t+s}) &= 0 \qquad s \neq 0 \end{aligned} \qquad 4.2$$

La expresión 4.1 se conoce como esquema autorregresivo de primer orden de Markovy, se denota como *AR* (1). Se dice que la expresión 4.1 es autorregresivo debido a que se interpreta como la regresión de u_t sobre sí misma rezagada un periodo. En tanto, ε_t es un ruido blanco, esto es en la terminología de series temporales, que cumple con las hipótesis básicas: media cero, varianza constante y autocorrelaciones nulas y, en donde la condición de

que el parámetro ρ no sea mayor que uno en valor absoluto garantiza que u_t es estacionario.

El estimador de mínimos cuadrados ordinarios para β_1 que generalmente se ha empleado está dado por:

$$\hat{\beta}_1 = \frac{\Sigma x_t y_t}{\Sigma x_t^2} \qquad 4.3$$

En tanto, la varianza determinada bajo el esquema de *AR* (1) es:

$$var(\hat{\beta}_1)_{AR1} = \frac{\sigma^2}{\Sigma x_t^2} + \frac{2\sigma^2}{\Sigma x_t^2}\left[\rho \frac{\sum_{t=1}^{N-1} x_t x_{t+1}}{\sum_{t=1}^{N} x_t^2} + \rho^2 \frac{\sum_{t=1}^{N-2} x_t x_{t+2}}{\sum_{t=1}^{N} x_t^2} + \ldots + \rho^{N-1} \frac{x_1 x_N}{\sum_{t=1}^{N} x_t^2}\right] \qquad 4.4$$

donde $var(\hat{\beta}_1)_{AR1}$ representa la varianza de $\hat{\beta}_1$ bajo un esquema autorregresivo de primer orden, caso contrario a la expresión 4.4, está la fórmula tradicional cuando no existe autocorrelación.

$$var(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma x_t^2} \qquad 4.5$$

Si se hace una comparación entre 4.4 y 4.5 la primera muestra que es igual a la última más un término que depende de ρ y de las covarianzas que tome *X*. Bajo esta afirmación, no se puede decir si la varianza $var(\hat{\beta}_1)$ es superior o inferior a $var(\hat{\beta}_1)_{AR1}$, lo que sí se puede mencionar es que si ρ es cero, las dos fórmulas coincidirán.

El estimador de MCO de $\hat{\beta}_1$ dada en la expresión 4.3 y la varianza dada en la expresión 4.4 bajo el esquema de *AR* (1), demuestran que $\hat{\beta}_1$ sigue

siendo lineal e insesgado, la diferencia es que no continúa teniendo varianza mínima, por lo que hace a que no sea eficiente.

De esta manera, continuando con el modelo de dos variables y suponiendo un proceso *AR* (1) se puede demostrar que el MELI de β_1 está dado por la siguiente expresión.

$$\beta_1^{MCG} = \frac{\sum_{t=2}^{N}(x_t - \rho x_{t-1})(y_t - \rho y_{t-1})}{\sum_{t=2}^{N}(x_t - \rho x_{t-1})^2} \qquad 4.6$$

El subíndice *t* va ahora de *t = 2* a *t = N*. Su varianza está dada por:

$$var\ \beta_1^{MCG} = \frac{\sigma^2}{\sum_{t=2}^{N}(x_t - \rho x_{t-1})^2} \qquad 4.7$$

El estimador β_1^{MCG}, como lo sugiere el superíndice se obtiene utilizando el método de MCG. El estimador de MCG de β_1 dado en 4.6 incorpora el parámetro de autocorrelación ρ, en tanto que la fórmula de MCO dada en 4.3 simplemente no la tiene en cuenta. Esta razón es por la cual el estimador de MCG es MELI y no el de MCO, pues el estimador de MCG utiliza al máximo la información disponible.

En conclusión, bajo autocorrelación es el estimador de MCG dado en 4.6, el que es MELI, estando su varianza mínima determinada por 4.7 y no así por 4.4 y 4.5.

ACTIVIDAD DE APRENDIZAJE

Mencionar qué otra paquetería hay, además de la ya mencionada en la introducción del libro, *E views*, y que es la apropiada para la detección y estimación de los MCO, en presencia de autocorrelación. Dar una explicación a lo obtenido. Entregar en hojas blancas en la siguiente sesión.

4.3 MÉTODOS DE CORRECCIÓN

La autocorrelación es un problema serio, por lo cual es necesario detectar su presencia en una situación determinada, por lo cual existen diferentes pruebas de correlación serial que se utilizan comúnmente, y básicamente son instrumentos estadísticos y gráficos. En la práctica, no se sabe *a priori* si hay autocorrelación y cuál puede ser el proceso más adecuado para modelizarla.

Como se mencionó, existen varios contrastes de autocorrelación que se construyen usando los residuos del modelo original, MCO. Una de las pruebas más conocidas para detectar la autocoorrelación serial es la desarrollada por los estadísticos Durbin y Watson, comúnmente conocida como el estadístico *d* de Durbin-Watson, la cual se define de la siguiente manera:

$$d = \frac{\sum_{t=2}^{t=N} (e_t - e_{t-1})^2}{\sum_{t=1}^{t=N} e_t^2} \qquad 4.8$$

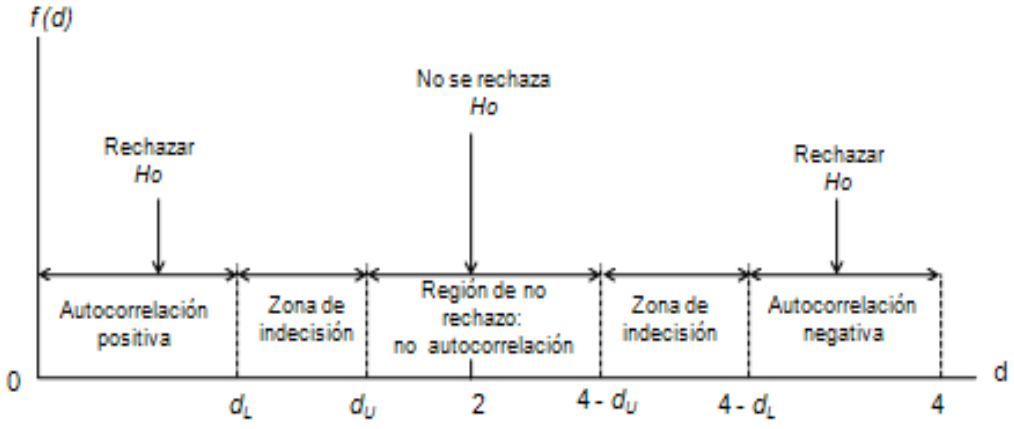
Esta razón es la suma de las diferencias al cuadrado de residuos sucesivos. La ventaja del estadístico *d* es que se basa en los residuos estimados que se calculan automáticamente en el análisis de regresión. El estadístico *d* se basa en los siguientes supuestos:

- El modelo de regresión incluye el término de regresión.
- Las variables explicativas, *X*, no son fijas.
- Las perturbaciones u_t se generan mediante el esquema autorregresivo *AR* (1), $u_t = \rho u_{t-1} + \varepsilon_t$.
- El modelo de regresión no incluye el o los valores rezagados de la variable dependiente como una de las variables explicativas.
- No hacen falta observaciones en los datos.

A partir del estadístico *d* se puede interpretar que:

- El valor de *d* será próximo a cero, cuando las diferencias entre los residuos en un periodo son pequeñas, presentándose con esto autocorrelación positiva.
- El estadístico *d* será próximo al límite superior cuando los residuos son prácticamente iguales, sólo que con signos contrarios, por lo que hay autocorrelación negativa.
- Cuando el estadístico *d* presenta un valor intermedio, es decir, que la relación entre los residuos es intermedia, hay ausencia de autocorrelación.

Si al calcular la expresión 4.8 cae fuera de los valores críticos entre el límite inferior, d_L, y el límite superior d_u, se puede tomar la decisión de la posible presencia de correlación serial positiva o negativa tal como lo muestra la figura 4.2. Estos límites van a depender del número de observaciones *N* (fluctuar entre 6 y 200) y del número de variables explicativas (hasta máximo 20) y no del valor que tomen estas variables explicativas. El procedimiento específico de prueba se puede explicar con la figura 4.2 que muestra los límites de *d* que están entre 0 y 4.



Estadístico *d* de Durbin-Watson

Figura 4.2

Esto es,

Si	Hipótesis nula	Decisión
$0 < d < d_L$	Autocorrelación positiva con un esquema *AR* (1)	Rechazar H_o
$d_L \leq d \leq d_u$	No existe autocorrelación positiva	No hay decisión
$4 - d_L < d < 4$	Autocorrelación negativa con un esquema *AR* (1)	Rechazar H_o
$4 - d_u \leq d \leq 4 - d_L$	No existe autocorrelación negativa	No hay decisión
$d_U < d < 4 - d_u$	No existe autocorrelación	No rechazar H_o

Tabla 4.1

Reglas de decisión

Para establecer los límites de variación del estadístico *d*, la fórmula 4.8 se puede desarrollar obteniéndose una expresión en función del coeficiente de correlación muestral de primer orden para los residuos $\hat{\rho}$,

$$d = \frac{\Sigma e_t^2 + \Sigma e_{t-1}^2 - 2\Sigma e_t e_{t-1}}{\Sigma e_t^2} \qquad 4.9$$

Los términos Σe_t^2 y Σe_{t-1}^2 difieren únicamente en una sola observación, por lo que se consideran iguales. De esta manera, la 4.9 se puede expresar de la siguiente manera como una aproximación ($\approx$).

$$d \approx 2\left(1 - \frac{\Sigma e_t e_{t-1}}{\Sigma e_t^2}\right) \qquad 4.10$$

En tanto el coeficiente de correlación empírico de primer orden se calcula

$$\hat{\rho} = \frac{\Sigma e_t e_{t-1}}{\Sigma e_t^2} \qquad 4.10$$

Por tanto, el estadístico experimental se puede expresar como:

$$d \approx 2(1 - \hat{\rho}) \qquad 4.10$$

Puesto que $-1 \leq \hat{\rho} \leq 1$, implica que

$$0 \leq d \leq 4 \qquad 4.11$$

De tal modo, se puede deducir el rango de variación del estadístico de Durbin-Watson y el signo de la autocorrelación.

- Si $\hat{\rho} = 0$ y $d \approx 2$ implica que no hay una correlación serial.
- Si $\hat{\rho} = 1$ y $d \approx 0$ implica una correlación serial positiva.
- Si $\hat{\rho} = -1$ y $d \approx 4$ implica una correlación serial negativa.

Por tanto, se asume que el estadístico experimental (d) tomará valores entre *0* y *4*, como se observa en la figura 4.2, de tal modo que cuánto más próximo a cuatro (a cero) sea el valor estadístico *d* mayor es la evidencia de autocorrelación negativa (positiva). Si el valor estadístico es dos, entonces la correlación muestral será nula y por tanto no se detectará un problema de autocorrelación entre las perturbaciones.

La prueba Durbin-Watson requiere que se cumpla con los siguientes pasos:

1. Estimar los MCO y obtener los residuos e_i.
2. Calcular d a partir de 4.8

3. Hallar los valores críticos de d_L y d_u, para un tamaño de muestra dado y un número determinado de variables explicativas.
4. Seguir las reglas de decisión de la tabla 4.1, o bien las que se ilustran en la figura 4.2.

El contraste Durbin-Watson tiene el inconveniente de no ser determinante, es por ello que se deben considerar otros criterios que sean decisivos al considerar si hay, o no, autocorrelación. Como se puede apreciar en la figura 4.1, si el estadístico de prueba, *d*, cae en la zona de indeterminación, no se puede concluir nada y menos aún si aparecen regresores estocáticos en el modelo, por lo que el estadístico Durbin-Watson presenta sesgo hacia el 2.

ACTIVIDAD DE APRENDIZAJE

Investigar otras pruebas de corrección de la autocorrelación y explicar cada una de ellas. Entregar en hojas blancas en la siguiente sesión.

AUTOEVALUACIÓN

I. Relacionar las siguientes columnas e indicar en el paréntesis la respuesta que corresponde a la afirmación.

1. Su efecto es que invalida uno de los supuestos que fundamentan el procedimiento de mínimos cuadrados ordinarios.	( ) Autocorrelación negativa con un esquema *AR*
2. Cuando $\hat{\rho} = 0$ y $d \square 2$ indica …	( ) Causa de correlación
3. Si $4 - d_L < d < 4$ se está en presencia de…	( ) Autocorrelación
4. El ordenamiento de los datos debe de tener cierto orden lógico o económico.	( ) Durbin-Watson
5. Contraste de autocorrelación que se construye usando los residuos del modelo original.	( ) $d = \frac{\sum_{t=2}^{t=N} (e_t - e_{t-1})^2}{\sum_{t=1}^{t=N} e_t^2}$
6. El coeficiente de correlación empírico de primer orden se calcula con…	( ) Correlación nula
7. Cómo se define el estadístico *d* de Durbin-Watson.	( ) Correlación espacial
8. Cómo es considerada la especificación incorrecta de la forma funcional del modelo.	( ) Autocorrelación negativa
9. Se presenta cuando los residuos son prácticamente iguales, pero de signo contrario y el estadístico será más próximo al límite superior.	( ) $\hat{\rho} = \frac{\Sigma e_t e_{t-1}}{\Sigma e_t^2}$
10. Fórmula tradicional en la cual no hay presencia de autocorrelación.	( ) $var(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma x_t^2}$

Respuestas

I. Relacionar las siguientes columnas e indicar en el paréntesis la respuesta que corresponde a la afirmación.

1. Su efecto es que invalida uno de los supuestos que fundamentan el procedimiento de mínimos cuadrados ordinarios.	(3) Autocorrelación negativa con un esquema *AR*
2. Cuando $\hat{\rho} = 0$ y $d \square 2$ indica …	(8) Causa de correlación
3. Si 4 - $d_L < d < 4$ se está en presencia de…	(1) Autocorrelación
4. El ordenamiento de los datos debe de tener cierto orden lógico o económico.	(5) Durbin-Watson
5. Contraste de autocorrelación que se construye usando los residuos del modelo original.	(7) $d = \frac{\sum_{t=2}^{t=N} (e_t - e_{t-1})^2}{\sum_{t=1}^{t=N} e_t^2}$
6. El coeficiente de correlación empírico de primer orden se calcula con…	(2) Correlación nula
7. Cómo se define el estadístico *d* de Durbin-Watson.	(4) Correlación espacial
8. Cómo es considerada la especificación incorrecta de la forma funcional del modelo.	(9) Autocorrelación negativa
9. Se presenta cuando los residuos son prácticamente iguales, pero de signo contrario y el estadístico será más próximo al límite superior.	(6) $\hat{\rho} = \frac{\Sigma e_t e_{t-1}}{\Sigma e_t^2}$
10. Fórmula tradicional en la cual no hay presencia de autocorrelación.	(10) $var(\hat{\beta}_1) = \frac{\sigma^2}{\Sigma x_t^2}$

UNIDAD 5

VARIABLES ARTIFICIALES O CUALITATIVAS

OBJETIVO

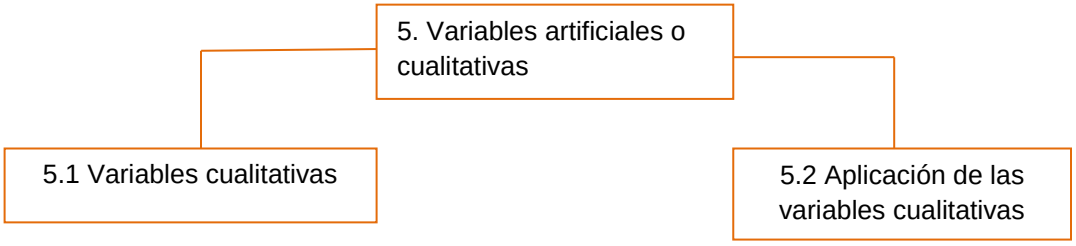
El estudiante será capaz de asignar variables cualitativas en un modelo de regresión lineal, y convertirlas en valores cuantificables. Asimismo, identificará cuándo un modelo es ANOVA o ANCOVA.

TEMARIO

5.1 VARIABLES CUALITATIVAS

5.2 APLICACIÓN DE LAS VARIABLES CUALITATIVAS

MAPA CONCEPTUAL



INTRODUCCIÓN

En un modelo de regresión lineal se introducen variables cuantitativas, de las cuales se puede obtener información de alguna base de datos, de esta manera, se tienen datos acerca de la variable dependiente e independiente.

En la presente Unidad se puede ver cómo la variable dependiente no sólo se encuentra influenciada por una variable cuantitativa, sino que también está determinada por variables cualitativas. Esta variable cualitativa, llamada también variable artificial, indica la presencia o ausencia de una cualidad o atributo, por lo que adoptan valores cuantitativos, que pueden ser 0 o 1.

En la segunda sesión se analiza la aplicación de un modelo en el que sólo se incluyen variables artificiales, en las variables explicativas, y que recibe el nombre de ANOVA. De igual modo, se verifica un modelo en que las variables explicativas se conforman tanto de variables cualitativas y cuantitativas, llamado ANCOVA.

Por último, se puede verificar que al dividir un modelo en diferentes subgrupos se puede obtener una serie de regresiones que pueden mostrar diferentes tendencias, mismas que se reflejan en la intersección y en sus pendientes.

5.1 VARIABLES CUALITATIVAS

En un análisis de regresión, normalmente, la variable dependiente se encuentra influenciada por variables que se pueden cuantificar fácilmente mediante una escala bien definida, como es el salario, el nivel educativo, el promedio de calificaciones, las ventas, etcétera. Cabe mencionar que en un trabajo empírico se deben incluir variables de tipo cualitativo.

Normalmente, se utilizan variables cuantitativas, es decir, aquéllas cuyos valores vienen expresados de manera numérica. En la Unidad 1 se mencionó que tal información puede obtenerse de datos ya existente. Estos valores cuantitativos se expresan en un modelo de regresión en el que se incluyen tanto la variable dependiente como las variables independientes. Es preciso disponer de magnitudes cuantitativas asociadas a las variables para llevar a cabo el proceso de estimación, que implica realizar cálculos numéricos.

Es importante mencionar que también existe la posibilidad de incluir en el modelo econométrico información cualitativa, siempre que la información cualitativa pueda expresarse de manera cuantitativa.

En los cursos de estadística ya se había definido a una variable cualitativa como aquella que describe cualidades o atributos del objeto de estudio. Por ejemplo: sexo, estado civil, raza, religión, corriente ideológica, zona geográfica, nacionalidad, etcétera. Estos son considerados factores cualitativos.

Esos factores cualitativos recogen efectos diferenciales como es el caso del sexo de una persona (si es hombre o mujer), la raza (blanco o negro), religión (católico o no católico), es decir, adoptan la forma de datos binarios. En econometría a estos datos se les conocen como: variables binarias, variable dicotómica, variables indicadoras, variables dummy, variables ficticias y variables cualitativas.

Se llaman variables dummy a las variables que tomando valores cuantitativos, tratan de representar las diferentes situaciones o casos que se producen en los factores cualitativos de interés.

Un ejemplo de lo anterior puede ser comprobar si el sexo de un trabajador influye, o tiene importancia, en el salario que percibe. En este caso,

la variable sexo incluye dos posibilidades, hombre y mujer, por lo que se tiene que atribuir un valor cuantitativo a cada uno de estos dos casos, de manera que cuando se trate de un trabajador hombre la variable ficticia tome un valor y cuando se determine un trabajador mujer se le asigne un valor diferente.

Otro ejemplo puede ser si el nivel educativo alcanzado por un trabajador, si tiene algún efecto sobre el salario, distinguiéndose los casos desde primaria o menos, secundaria, preparatoria, universidad o postgrado. Siendo cada uno de estos casos identificados de manera cuantitativa.

Los valores numéricos que se pueden atribuir son completamente arbitrarios y no tienen más efecto que establecer un código que permita distinguir numéricamente cada caso de los demás. Así, la variable dummy se define sin más que atribuir un número diferente a cada uno de los casos posibles en el factor que se considere.

Las variables cualitativas son construidas artificialmente y, generalmente, indican la presencia o ausencia de una cualidad o atributo, una manera de cuantificar tales atributos consiste en asignarle valores de 1 o 0, donde 0 indica la ausencia de un atributo y 1 la presencia de ese atributo. Por ejemplo, el sexo de una persona, mujer puede ser 1, y del hombre 0; qué persona tiene estudios profesionales se indica con 1, o que no cuenta con estudios profesionales, con 0, y así sucesivamente.

El hecho de que se utilice 0 y 1 es porque son valores arbitrarios, igual puede ser cualquier otro valor, sólo que en la captura de información cualitativa, de un modelo de regresión, el 0 y el 1 lleva a que los parámetros tengan interpretaciones naturales.

Las variables ficticias pueden incluirse tanto en modelos temporales como en modelos de corte transversal. Los paquetes computacionales realizan la transformación de las variables categóricas en las variables dummy necesarias automáticamente, y no se requiere efectuar todo el proceso manualmente, únicamente debe identificarse, en el programa computacional que se utilice, cuál es el nombre de las variables que requieren este tipo de transformación.

Al construir un modelo de regresión, es importante interpretar los resultados obtenidos, pues al aplicar variables cualitativas se obtendrá un resultado global, es decir, del modelo en general donde confluyen tanto variables cuantitativas como dicotómicas, y además se arrojarán resultados respecto a cada una de las variables *dummy* que intervinieron en el modelo. De esta manera se puede hacer una comparación de cada uno de los resultados obtenidos por las variables cualitativas, y así explicar los resultados de modo adecuado.

ACTIVIDAD DE APRENDIZAJE

Plantear de forma teórica un modelo econométrico y determinar en él las variables cualitativas y de qué manera afectan al modelo inicial, en el cual determine variables cualitativas, especificando por qué es 0 y por qué 1, e indicar de qué manera pueden afectar al modelo. Hacer un análisis de por lo menos una cuartilla. Entregar en hojas blancas en la siguiente sesión.

Se debe plantear un modelo, el cual contiene variables que son medibles, por ejemplo, con ingreso y consumo a partir de estas variables, se determinan las cualitativas que pueden ser profesionista o no profesionista.

5.2 APLICACIÓN DE LAS VARIABLES CUALITATIVAS

Al igual que las variables cuantitativas, las variables dummy se pueden utilizar con facilidad en los modelos de regresión. Incluso los modelos de regresión pueden incluir como variables explicativas sólo variables cualitativas.

Es conveniente mencionar que cuando un modelo sólo incluye variables dummy como variables explicativas, se le llama modelo de análisis de la varianza (ANOVA) y cuando un modelo incluye variables cuantitativas y cualitativas se le conoce como modelo de análisis de la covarianza (ANCOVA).

Un ejemplo de modelo ANOVA que sólo incluye variables explicativas de tipo cualitativo es:

$$Y_i = \delta + \beta D_i + u_i \qquad 5.1$$

donde Y_i = salario anual de un ingeniero

D_i = 1 si el ingeniero es hombre

= 0 si el ingeniero es mujer

La expresión 5.1 es similar a los modelos de regresión en dos variables, excepto porque en lugar de tener *X*, ahora tiene a *D* que es la variable dicotómica (la literal *D* identificará en adelante a una variable dicotómica).

La expresión 5.1 permite averiguar si el sexo tiene alguna incidencia sobre el salario de los ingenieros, manteniendo contante otras variables como edad, años de experiencia o grados universitarios alcanzados. ¿Cómo se interpreta esta expresión?, δ mide el valor medio de la variable dependiente de la categoría base o de referencia, es decir, para la que la variable dummy asume el valor 0; β mide la diferencia del punto de corte entre las dos categorías y se le llama coeficiente del punto de corte diferencial. Suponiendo que las perturbaciones satisfacen los supuestos del modelo clásico de regresión lineal, a partir de la 5.1 se obtiene:

Salario promedio de un ingeniero mujer	$E(Y_i / D_i = 0) = \delta$	5.2
Salario promedio de un ingeniero hombre	$E(Y_i / D_i = 1) = \delta + \beta$	

De esta manera, el término de intersección δ proporciona el salario promedio de los ingenieros mujeres, y el coeficiente de la pendiente β dice en cuánto difiere el salario promedio de un ingeniero hombre del salario promedio de su contraparte femenina, mientras que $\delta + \beta$ refleja el salario promedio de un ingeniero hombre.

A partir de la prueba de hipótesis se puede contrastar el hecho de que no existe discriminación sexual, esto es por medio de la hipótesis nula ($Ho: \beta = 0$), es decir, se puede llevar a cabo la corrida de la regresión establecida en la expresión 5.1 y con base en la prueba *t*, se puede averiguar si el $\hat{\beta}$ estimado es estadísticamente significativo.

Por otra parte, el modelo ANCOVA es una ampliación de los modelos ANOVA, el cual incluye variables explicativas cuantitativas que controlan estadísticamente los efectos de las variables dummy. Un ejemplo de esto, son los siguientes:

a) Regresión con una variable cuantitativa y una cualitativa con dos clases o categorías. Se tiene entonces, que la expresión 5.1 se modifica quedando como sigue:

$$Y_i = \delta_1 + \delta_2 D_i + \beta X_i + u_i \qquad 5.3$$

donde Y_i = salario anual de un ingeniero

X_i = años de experiencia

D_i = 1 si es hombre

= 0 si no lo es

Este modelo contiene una variable cuantitativa que son los años de experiencia, y una variable cualitativa que es el sexo, la cual posee dos niveles, hombre o mujer.

Por tanto, el salario promedio de un ingeniero mujer es

$$E(Y_i / X_i, D_i = 0) = \delta_1 + \beta X_i \qquad 5.4$$

y el salario promedio de un ingeniero hombre es

$$E(Y_i / X_i, D_i = 1) = (\delta_1 + \delta_2) + \beta X_i \qquad 5.5$$

El modelo 5.3 postula que el salario de los ingenieros hombres y las mujeres en relación con los años de experiencia tienen la misma pendiente β, pero diferentes intersecciones. En otras palabras, se supone que el nivel del salario promedio del ingeniero hombre es

diferente del salario promedio del ingeniero mujer (en δ_2), pero la tasa de cambio en el salario anual promedio por años de experiencia es la misma para ambos sexos. Esto se puede verificar en la siguiente figura 5.1.

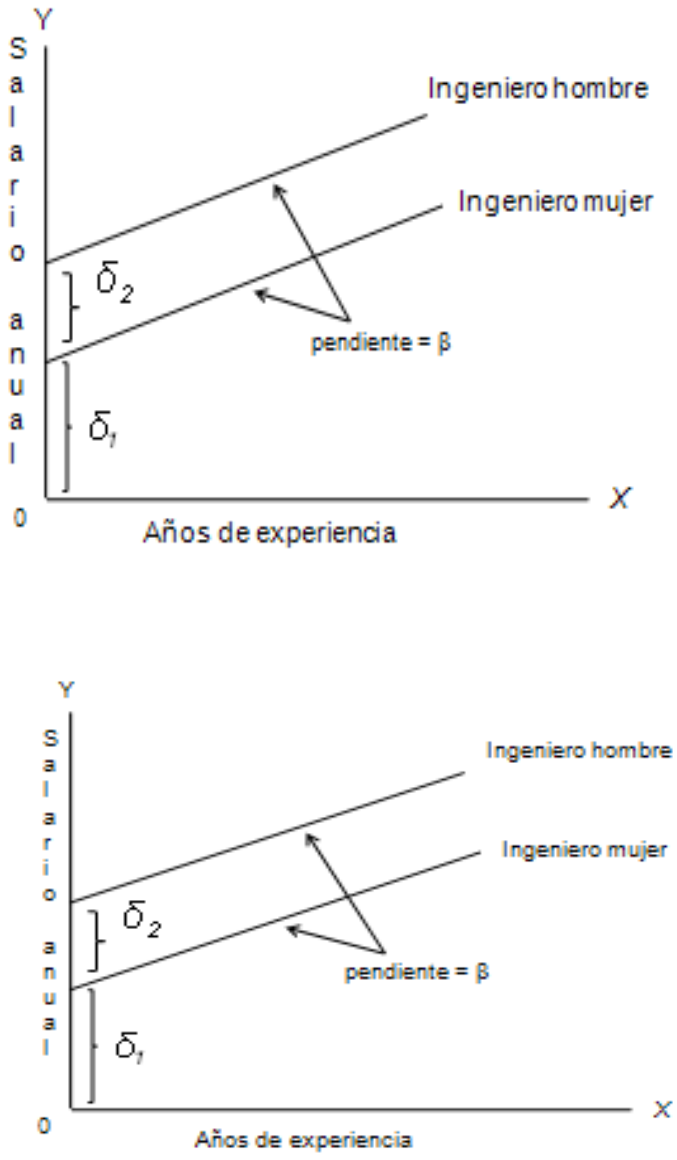


Figura 5.1

Salario anual y años de experiencia de los ingenieros.

Si el supuesto de la pendiente común es válido, una prueba de hipótesis de que las dos regresiones (5.4 y 5.5) tienen la misma intersección, es decir que no hay discriminación sexual, se puede efectuar mediante la corrida de la regresión de la expresión 5.3 y observando la significancia estadística de δ_2 estimado, con base en la prueba *t* tradicional. Si la prueba *t* muestra que δ_2 es estadísticamente significativo, se rechaza la hipótesis nula de que los niveles salariales de los ingenieros hombres y las mujeres son los mismos.

Antes de continuar con el siguiente ejemplo, es importante considerar la existencia de una regla general, la cual indica que si una variable cualitativa tiene *m* categorías, se introduzca únicamente $m-1$ variables dicotómicas. De lo contrario, se cae en la denominada trampa de la variable dicotómica, es decir, aquella donde existe multicolinealidad perfecta.

b) Regresión en una variable cuantitativa y una variable cualitativa con más de dos clases. Supóngase que con base en datos de corte transversal, por ejemplo, se tienen gastos anuales que efectuar como es en la educación por individuo; de esta manera, se tiene una variable cuantitativa que es ingreso y una variable cualitativa que es el nivel de educación alcanzado por el individuo. En este caso la variable educación es cualitativa, misma que sigue tres niveles (secundaria, preparatoria y superior), por lo que se tienen más de dos categorías. Siguiendo la regla de que el número de variables dicotómicas debe ser igual al número de categorías de las variables menos uno, se deben introducir dos variables. Esto es bajo el supuesto de que los tres niveles educativos poseen una pendiente común, pero intersecciones diferentes, por lo que el modelo de regresión del gasto anual en educación se expresa de la siguiente manera:

$$Y_i = \delta_1 + \delta_2 D_{2i} + \delta_3 D_{3i} + \beta X_i + u_i \qquad 5.6$$

donde Y_i = gastos anuales en educación

X_i = ingreso anual

D_2 = 1 si se ha culminado la educación preparatoria

= 0 los demás casos

D_3 = 1 si se ha alcanzado educación universitaria

= 0 los demás casos

En la asignación de las variables dicotómicas se trata de manera arbitraria la categoría del nivel de educación inferior a la preparatoria, como la categoría base (0). Suponiendo que $E(u_i)=0$, se obtiene de 5.6.

$$E(Y_i/D_2=0, D_3=0, X_i)=\delta_1+\beta X_i \quad 5.7$$

$$E(Y_i/D_2=1, D_3=0, X_i)=(\delta_1+\delta_2)+\beta X_i \quad 5.8$$

$$E(Y_i/D_2=0, D_3=1, X_i)=(\delta_1+\delta_3)+\beta X_i \quad 5.9$$

Cada una de estas funciones, corresponden a los gastos promedio en educación para los tres niveles de educación, nivel inferior a preparatoria, preparatoria y universidad, esto se puede verificar en la siguiente figura.

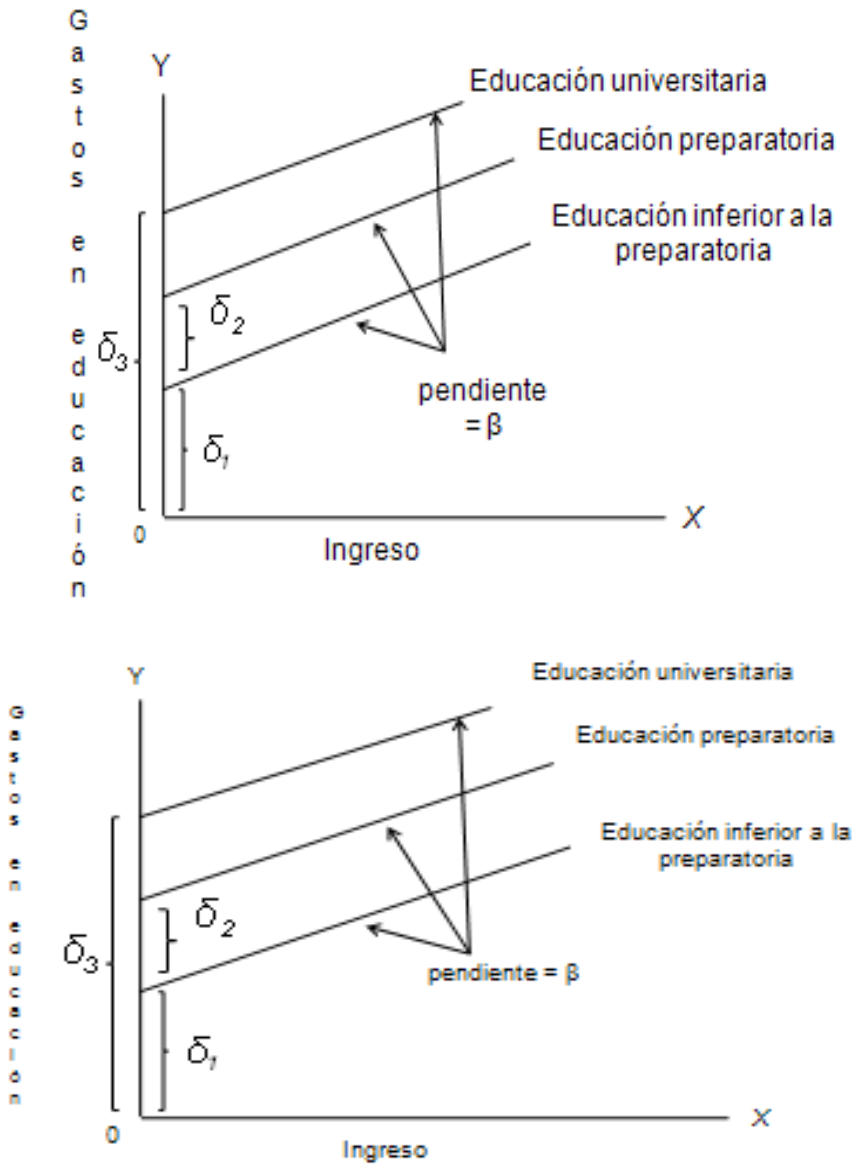


Figura 5.2

Gastos en educación en relación a los ingresos para tres niveles de educación.

Después de correr la regresión 5.6, se puede verificar si las intersecciones δ_2 y δ_3 son estadísticamente significativas en términos individuales, es decir, diferentes del grupo base. Una prueba de la hipótesis de que $\delta_2 = \delta_3 = 0$, también se puede realizar simultáneamente por medio de la técnica ANOVA y la prueba *F*.

La interpretación de la expresión 5.6 cambiaría si se hubiera adoptado un esquema diferente para asignar las variables dummy. Por tanto, si se hubiera asignado D_2 = 1 a la categoría menos que educación preparatoria y D_3 = 1 a la categoría educación preparatoria, la categoría base entonces es la educación universitaria y todas las comparaciones serán en relación con esta categoría.

c) Regresión en una variable cuantitativa y dos cualitativas. Volviendo a la regresión de los salarios de los ingenieros 5.3, ahora se supone que además del ingreso y el sexo, se considera el color de la raza (blanco o negro) del ingeniero, el cual determina de modo importante el salario. Así se tiene que 5.3 como:

$$Y_i = \delta_1 + \delta_2 D_{2i} + \delta_3 D_{3i} + \beta X_i + u_i \qquad 5.10$$

donde Y_i = salario anual

X_i = años de experiencia

D_2 = 1 si es hombre

= 0 los demás casos

D_3 = 1 si es blanco

= 0 los demás casos

Las variables cualitativas, sexo y color, tienen dos categorías, por lo que se requiere entonces de una variable dicotómica para cada una. Suponiendo nuevamente que $E(u_i) = 0$, se obtiene entonces

Salario promedio para un ingeniero negra

$$Y_i = \delta_1 + \delta_2 D_{2i} + \delta_3 D_{3i} + \beta X_i + u_i \qquad 5.11$$

Salario promedio para un ingeniero negro

$$E(Y_i / D_2 = 1, D_3 = 0, X_i) = (\delta_1 + \delta_2) + \beta X_i \qquad 5.12$$

Salario promedio para un ingeniero blanca

$$E(Y_i/D_2=0, D_3=1, X_i)=(\delta_1+\delta_3)+\beta X_i \qquad 5.13$$

Salario promedio para un ingeniero blanco

$$E(Y_i/D_2=0, D_3=1, X_i)=(\delta_1+\delta_2+\delta_3)+\beta X_i \qquad 5.14$$

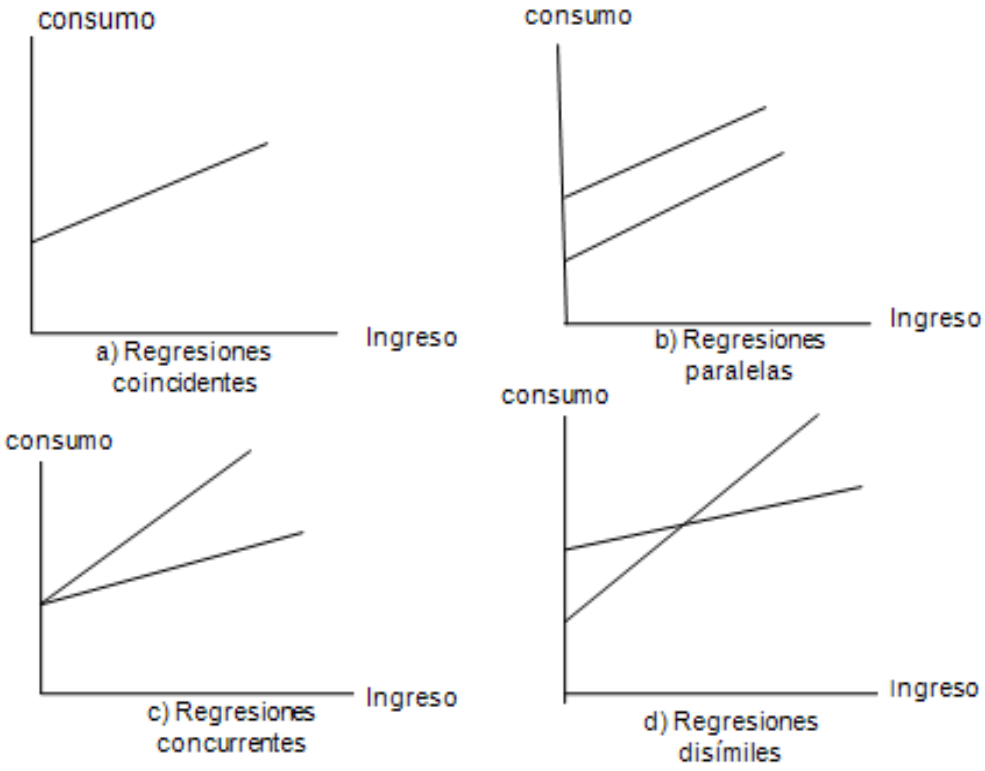
De nuevo se deduce que las regresiones anteriores difieren únicamente en el coeficiente de la intersección pero no en el coeficiente de la pendiente β .

La estimación de mínimos cuadrados ordinarios de 5.10 permite evaluar una variedad de hipótesis. De esta manera, si δ_3 es estadísticamente significativa implicará que el color afecta el salario de un ingeniero. De igual forma, si δ_2 es estadísticamente significativa implicará que el sexo también afectará el salario de un ingeniero. Así, si ambas intersecciones son estadísticamente significativas, se tiene que tanto el sexo como el color determinan en forma importante el ingreso de un ingeniero.

Hasta el momento se ha asumido en los modelos considerados en la presente sesión, que las variables cualitativas afectan la intersección, pero no el coeficiente de la pendiente de las diferentes regresiones de los subgrupos. Si se evaluará la diferencia en las intersecciones, tiene poca significancia en la práctica. Lo que hay que hacer es verificar si dos, o más, regresiones son diferentes, por lo que la diferencia puede estar en las intersecciones o en la pendiente, o en ambas. A partir de esto se puede determinar cuatro posibilidades, a saber:

1. Regresiones coincidentes: cuando no hay diferencias ni en los coeficientes del punto de corte ni de las pendientes.
2. Regresiones paralelas: las pendientes son iguales, pero los puntos de corte son distintos.

3. Regresiones concurrentes: los puntos de corte son iguales, pero las pendientes son distintas.
4. Regresiones disímiles: tanto los puntos de corte como las pendientes son distintos.



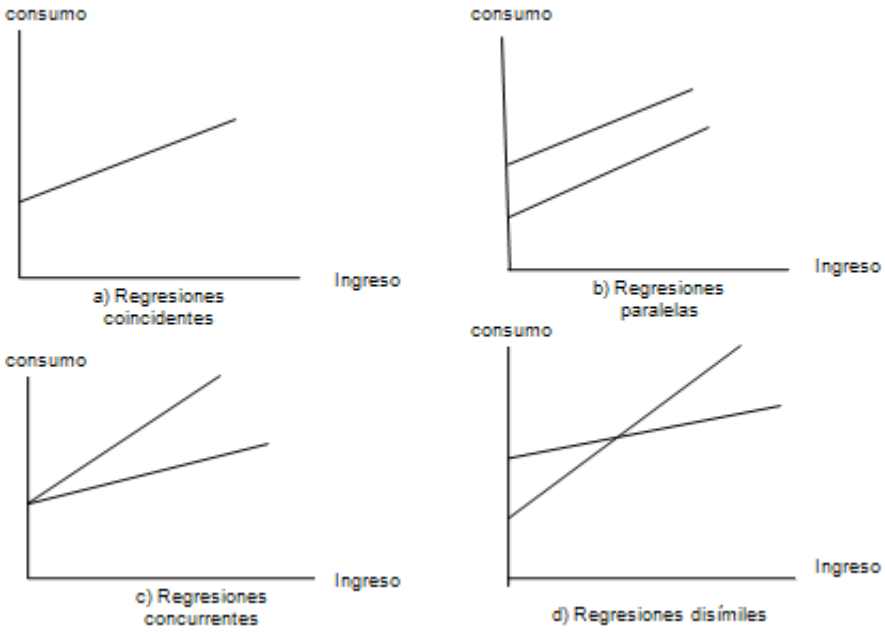


Figura 5.3

Posibles regresiones entre consumo-ingreso

Por último, las variables dicotómicas son esencialmente trucos para la clasificación de datos, ya que dividen una muestra en diferentes subgrupos con base en cualidades o atributos y se corren diferentes regresiones para cada uno de estos subgrupos. En caso de existir diferencias en la respuesta de la variable dependiente ante un cambio en las variables cualitativas en los diferentes subgrupos, esto se verá reflejado en las diferencias en los coeficientes de intersección, o de las pendientes o en ambos simultáneamente.

ACTIVIDAD DE APRENDIZAJE

Plantear dos modelos, el primer modelo que sea de tipo ANOVA, y el segundo de forma ANCOVA. Dar una explicación teórica de cada uno de los modelos establecidos. Entregar en hojas blancas en la siguiente sesión.

AUTOEVALUACIÓN

I. Relacionar las siguientes columnas e indicar en el paréntesis la respuesta que corresponde a la afirmación.

1. Los factores cualitativos recogen efectos diferenciales, qué tipo de forma adoptan.	( ) ANOVA
2. Para llevar a cabo la estimación con qué tipo de magnitudes es necesario contar para asociar a las variables.	( ) Datos binarios
	( ) ANCOVA
3. Son aquéllas que tratan de representar diferentes situaciones o casos que se producen en los factores cualitativos de interés.	( ) Variable dummy
	( ) Coincidente
4. En qué tipo de modelos se pueden incluir las variables ficticias.	( ) Cuantitativas
5. Qué nombre reciben los modelos cuando se incluyen sólo variables dummy como variables explicativas.	( ) Multicolinialidad perfecta
6. Si en un modelo no se introduce $m-1$ variables dicotómicas se puede darse la existencia de…	( ) Presencia o ausencia
7. Qué tipo de regresión se tiene cuando no hay diferencias ni en los coeficientes del punto de corte ni de las pendientes.	( ) Corte transversal y temporal.
	( ) Disímil
8. Qué nombre recibe un modelo cuando incluye variables cuantitativas y cualitativas.	
9. Qué tipo de regresión se tiene cuando tanto los puntos de corte como las pendientes son distintos.	

10. Una forma de cuantificar los atributos de las variables cualitativas es con 1 o 0, qué indican estos valores.	

Respuestas

I. Relacionar las siguientes columnas e indicar en el paréntesis la respuesta que corresponde a la afirmación.

1. Los factores cualitativos recogen efectos diferenciales, qué tipo de forma adoptan.	(5) ANOVA
2. Para llevar a cabo la estimación con qué tipo de magnitudes es necesario contar para asociar a las variables.	(1) Datos binarios
	(8) ANCOVA
3. Son aquéllas que tratan de representar diferentes situaciones o casos que se producen en los factores cualitativos de interés.	(3) Variable dummy
	(7) Coincidente
4. En qué tipo de modelos se pueden incluir las variables ficticias.	(2) Cuantitativas
5. Qué nombre reciben los modelos cuando se incluyen sólo variables dummy como variables explicativas.	(6) Multicolinialidad perfecta
	(10) Presencia o ausencia
6. Si en un modelo no se introduce $m-1$ variables dicotómicas se puede darse la existencia de…	(4) Corte transversal y temporal.
7. Qué tipo de regresión se tiene cuando no hay diferencias ni en los coeficientes del punto de corte ni de las pendientes.	(9) Dísimil

8. Qué nombre recibe un modelo cuando incluye variables cuantitativas y cualitativas. 9. Qué tipo de regresión se tiene cuando tanto los puntos de corte como las pendientes son distintos. 10. Una forma de cuantificar los atributos de las variables cualitativas es con 1 o 0, qué indican estos valores.	

UNIDAD 6

SERIES TEMPORALES

OBJETIVO

El estudiante identificará series de tiempo económicas en un modelo de regresión y comprenderá su estimación mediante un modelo de rezagos distribuidos. Distinguirá las etapas de predicción de series de tiempo mediante el modelo ARIMA.

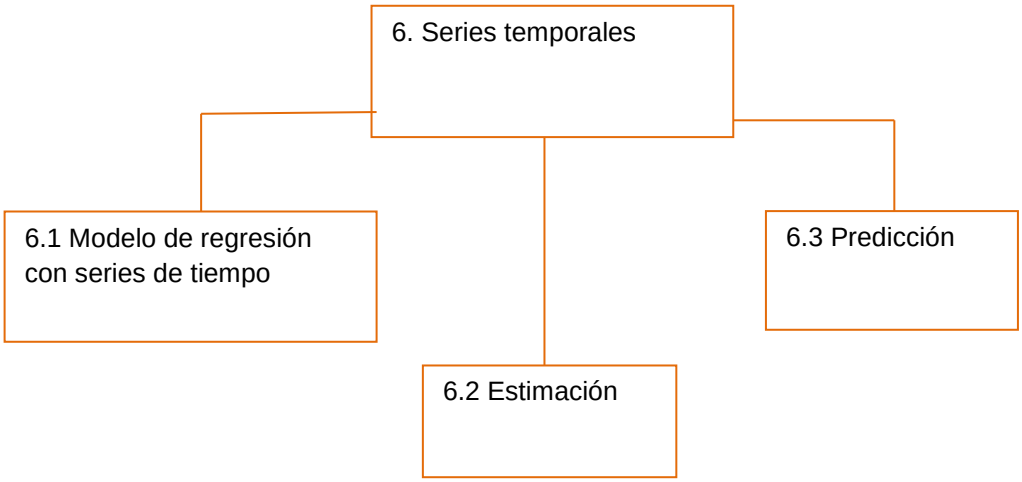
TEMARIO

6.1 MODELO DE REGRESIÓN CON SERIES DE TIEMPO

6.2 ESTIMACIÓN

6.3 PREDICCIÓN

MAPA CONCEPTUAL



INTRODUCCIÓN

Una de las categorías más importantes de los modelos de regresión lineal es la estimación de series de tiempo, las cuales son principalmente aplicables en series económicas.

En la presente Unidad se analizará qué es una serie de tiempo y cuál es su comportamiento o tendencia de forma gráfica. En el análisis de regresión que contiene series de tiempo, no sólo se incluye valores actuales sino también valores rezagados (pasados).

Se verifica la estimación de mínimos cuadrados ordinarios mediante el método de rezagos distribuidos, esto se produce con la aplicación de expresiones matemáticas.

Por último, se estudia el modelo de predicción para series de tiempo estacionarias desarrollado por Box y Jenkins, tal modelo de regresión con variables independientes es denominado ARIMA.

6.1 MODELO DE REGRESIÓN CON SERIES DE TIEMPO

Una serie de tiempo es una secuencia cronológica de observaciones de una variable o conducta particular durante un periodo determinado. Una característica de los datos de las serie de tiempo que los distingue de los de corte transversal es su orden temporal. Para analizar los datos de series de tiempo en las ciencias sociales se debe de reconocer que el pasado influye en el futuro.

El patrón de una serie temporal puede asumir diferentes formas dependiendo del factor involucrado. Un primer factor de variación es la tendencia que produce en la serie de tiempo un movimiento ascendente o descendente en un periodo determinado, el cual se refleja en un crecimiento o desvanecimiento en la serie, tal como se muestra en la figura 6.1

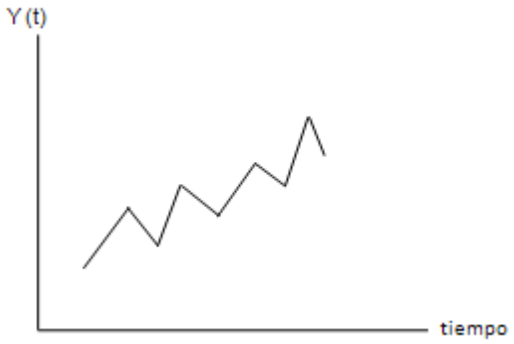


Figura 6.1
Tendencia

Un segundo patrón es la llamada variación cíclica que se refiere a un movimiento recurrente de arriba hacia abajo alrededor de un nivel de tendencia presente en un periodo fijo. La duración del periodo puede ser un año, un trimestre, un mes, un día, etcétera. Suele hacerse distinción entre cíclicas y estacionarias. La variación cíclica puede referirse a ciclos grandes, tal puede ser el comportamiento de la economía en que hay periodos de crecimiento y otros de estancamiento, es en este último en el que se puede encontrar estacionalidad, por ejemplo, en los niveles de empleo.

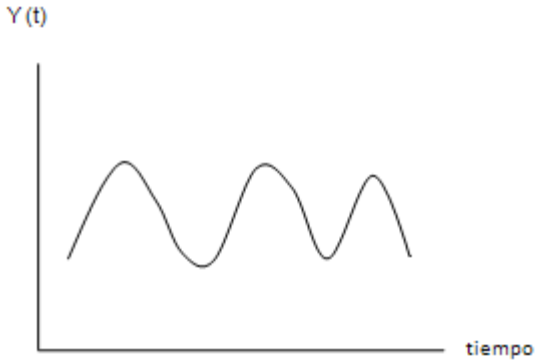


Figura 6.2
Variaciones cíclicas

Una tercera forma es la que recoge movimientos erráticos que no siguen un patrón regular, llamada fluctuaciones irregulares. La mayoría de las veces estas fluctuaciones irregulares son producto de eventos inusuales que no pueden evitarse y que obedecen a fallas en los sistemas de observación o a errores aleatorios.

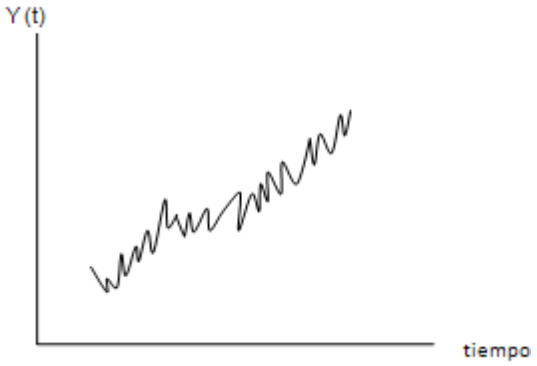


Figura 6.3
Irregulares

A continuación se da un ejemplo de una serie de tiempo correspondiente a cifras anualizadas del porcentaje de inflación registrado desde el año de 1999 hasta el año 2010

AÑO	INFLACIÓN
1999	12,32
2000	8,96
2001	4,4
2002	5,7
2003	3,98
2004	5,19
2005	3,33
2006	4,05
2007	3,76
2008	6,53
2009	3,57
2010	4,4

Fuente: INEGI. Sistema de Cuentas Nacionales de México, año base 2010=100

Tabla 6.1

Tabla de porcentaje de inflación en México 1999-2010

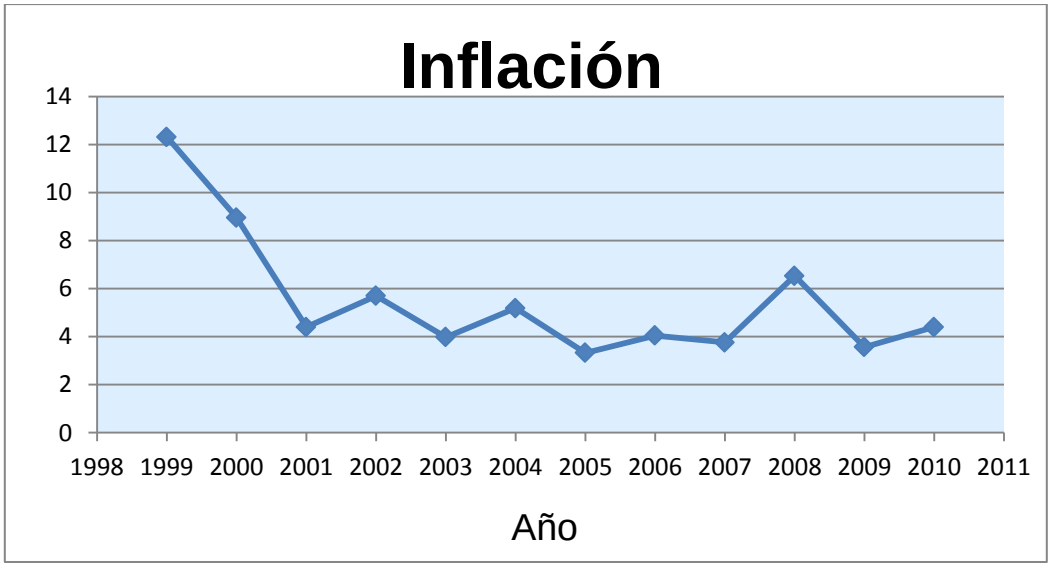


Figura 6.4

Porcentaje de inflación anual desde 1999 hasta 2010

Como se puede apreciar en la figura 6.4, se tiene una serie temporal de tendencia, al mostrar un comportamiento ascendente y descendente.

Otro concepto fundamental en el análisis de series de tiempo, es el de estacionariedad, que se refiere a la ausencia de cualquier tipo de variabilidad, ya sea de tendencia o cíclica. En este caso, las variables que no muestran tendencia a crecer a lo largo del tiempo.

El análisis de series de tiempo consiste en el examen del patrón histórico generado por el evento en observación con la esperanza. De esta manera, los datos de series de tiempo obtenidos se utilizan para estimar la totalidad del proceso de estudio.

A la colección de variables aleatorias ordenadas en el tiempo se le llama proceso estocástico o aleatorio, o bien, proceso de serie de tiempo (estocástico es sinónimo de aleatorio). Al hacer la recopilación de series de tiempo se tiene un proceso estocástico o aleatorio, el cual se realiza una vez, esto es que sólo se está manejando un rango en donde los valores no se pueden modificar, ya están dados, pues no hay retroceso en el tiempo. En el análisis de corte transversal, la recopilación de las diferentes series de tiempo adopta la forma de población.

Para el manejo de series de tiempo, el proceso estocástico debe ser estacionario, es decir, que la media y la varianza sea constante en el tiempo, siendo que el valor de la covarianza va a depender de la distancia o rezago entre dos periodos.

En el análisis de regresión que contiene series de tiempo, no sólo se incluye valores actuales sino también valores rezagados (pasados) de las variables explicativas (las X), y se le denominan modelos de rezagos distribuidos. Es importante mencionar que si el modelo incluye uno o más valores rezagados de la variable dependiente entre sus variables explicativas, a esto se le conoce con el nombre de modelo autorregresivo.

ACTIVIDAD DE APRENDIZAJE

Graficar una serie de tiempo respecto al comportamiento del Producto Nacional Bruto desde 1989 al 2010 a precios constantes, hacer un análisis de su

comportamiento e indicar qué tipo de variación es. Entregar en hojas blancas en la siguiente sesión.

6.2 ESTIMACIÓN

Los modelos de series de tiempo son útiles en el análisis empírico, y se estiman con facilidad mediante mínimos cuadrados ordinarios, son diversos los métodos que se pueden emplear para la estimación de series de tiempo, de manera particular se atiende en la siguiente sesión el de rezagos distribuidos.

Un modelo de rezagos distribuidos se representa mediante la siguiente expresión:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + u_t \qquad 6.1$$

en tanto, un modelo autorregresivo se determina del siguiente modo:

$$Y_t = \alpha + \beta X_t + \gamma Y_{t-1} + u_t \qquad 6.2$$

En economía, la dependencia de una variable *Y* con respecto a otra variable *X* (variable explicativa) suele no ser inmediata. Con frecuencia *Y* responde a *X* con un lapso; este tiempo se denomina rezago. De manera general se puede escribir un modelo de rezagos distribuidos como:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + \beta_k X_{t-k} + u_t \qquad 6.3$$

este rezago es finito en *k* periodos, donde β_0 se conoce como multiplicador de impacto o de corto plazo, por representar el cambio en el valor medio de *Y* después de un cambio unitario en *X*; por lo que $\beta_0 + \beta_1)$ corresponden al cambio en (el valor promedio de) *Y* en el siguiente periodo, $\beta_0 + \beta_1 + \beta_2)$ en el siguiente y así sucesivamente. A estas sumas parciales se les denomina multiplicadores intermedios. Después de *k* periodos se obtiene:

$$\sum_{i=0}^{k} \beta_i = \beta_0 + \beta_1 + \beta_2 + \ldots + \beta_k = \beta \qquad 6.4$$

conocido como multiplicador a largo plazo o total. Si se define

$$\beta_i^* = \frac{\beta_i}{\Sigma \beta_i} = \frac{\beta_i}{\beta} \qquad 6.5$$

se obtiene el β_i estandarizado. Las sumas parciales de los β_i estandarizados generan la proporción del impacto total o a largo plazo que se experimenta para un cierto periodo de tiempo.

Los rezagos ocupan un lugar fundamental en la economía, lo cual se refleja al tratar fenómenos de corto y largo plazo, de ahí que es importante saber cómo deben estimarse. Suponiendo que se tiene una variable explicativa, el modelo de rezagos distribuidos infinito se expresa de la siguiente manera:

$$Y_t = \alpha + \beta_0 X_t + \beta_1 X_{t-1} + \beta_2 X_{t-2} + \ldots + u_t \qquad 6.6$$

Para estimar α y β de la expresión 6.6 se pueden adoptar dos enfoques: la estimación ad hoc y las restricciones a priori sobre las β, suponiendo que estas siguen un patrón sistemático.

La estimación ad hoc supone que la variable explicativa X_t es no estocástica, así como X_{t-1}, X_{t-2} y así sucesivamente. O bien, que el término de perturbación u_t no está correlacionado. De esta manera, el método de mínimos cuadrados ordinarios se puede aplicar a 6.6. Este enfoque sugiere que para estimar 6.6 se proceda secuencialmente, es decir, que primero se regresa Y_t en X_t, luego Y_t en X_t y X_{t-1}, a continuación se regresa Y_t en X_t, X_{t-1}, y X_{t-2}, y así sucesivamente. Este procedimiento se detiene cuando los coeficientes de regresión comienzan a ser estadísticamente insignificantes, o bien cuando el

coeficiente de las variables empieza a cambiar de signo, es decir, de positivo a negativo o bien a la inversa de negativo a positivo.

Este enfoque presenta algunas complicaciones como las siguientes: primero, no hay una guía respecto a la máxima longitud del rezago; las series económicas tienden a estar altamente correlacionadas, tornándose la multicolinealidad en un factor de cuidado, pues esto ocasiona una estimación imprecisa de los coeficientes, reflejando con esto que los errores estándar sean grandes en relación a los coeficientes estimados.

Debido a las complicaciones que se presentan, el enfoque ad hoc tiende a ser poco recomendable.

Otro enfoque para la estimación de modelos de rezagos distribuidos es el de Koyck, suponiendo el modelo de rezagos de la expresión 6.6 y que todas las β tienen el mismo signo, éstas disminuyen geométricamente así:

$$\beta_k = \beta_0 \lambda^k \qquad k = 0,1,... \tag{6.7}$$

donde λ es: $0 \prec \lambda \prec 1$ y se le conoce como tasa de disminución del rezago distribuido y $1 - \lambda$ es la velocidad de ajuste.

La 6.7 indica que cada coeficiente sucesivo de β es numéricamente menor al β que le precede, de manera que cada vez que se retrocede al pasado, el efecto de ese rezago sobre Y_t se hace progresivamente más pequeño.

Por otra parte, cuanto más cerca esté λ a 1, más lenta será la velocidad de disminución de β_k, mientras que cuanto más cerca esté de 0 más rápidamente declinará β_k. Los valores más antiguos de X ejercen un impacto considerable sobre Y_t en tanto que en el último caso su influencia sobre Y_t disminuirá rápidamente.

La suma de los β proporciona un multiplicador finito a largo plazo.

$$\sum_{k=0}^{\infty} \beta_k = \beta_0 \left(\frac{1}{1 - \lambda} \right) \tag{6.8}$$

Como resultado de 6.7, el modelo de rezagos infinitos 6.6 se expresa como sigue:

$$Y_t = \alpha + \beta_0 X_t + \beta_0 \lambda X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \ldots + u_t \qquad 6.9$$

En este caso λ no es lineal, por lo que el método de regresión lineal en los parámetros no se puede aplicar al modelo de la 6.9, además de que hay que estimar una cantidad infinita de parámetros. Es con el método de Koyck que se puede solucionar este problema, el cual consiste en rezagar la expresión 6.9 en un periodo.

$$Y_{t-1} = \alpha + \beta_0 X_{t-1} + \beta_0 \lambda X_{t-2} + \beta_0 \lambda^2 X_{t-3} + \ldots + u_{t-1} \qquad 6.10$$

Multiplicando ambos lados de la expresión 6.10 por λ

$$\lambda Y_{t-1} = \lambda\alpha + \lambda\beta_0 X_{t-1} + \beta_0 \lambda^2 X_{t-2} + \beta_0 \lambda^3 X_{t-3} + \ldots + \lambda u_{t-1} \qquad 6.11$$

Restando 6.11 de 6.9, se obtiene

$$Y_t - \lambda Y_{t-1} = \alpha(1-\lambda) + \beta_0 X_t + (u_t - \lambda u_{t-1}) \qquad 6.12$$

reordenando esta expresión se tiene:

$$Y_t = \alpha(1-\lambda) + \beta_0 X_t + \lambda Y_{t-1} + v_t \qquad 6.13$$

siendo, esto corresponde al promedio móvil de u_t y u_{t-1}. De esta manera, con 6.13 se puede estimar α, β_0 y λ. Debido a que Y_{t-1} es una variable explicativa el modelo se convierte en autorregresivo.

En este caso, sólo se atiende a las aplicaciones o al método, para que se pueda dar en la realidad económica se tienen que tener series de tiempo de alguna variable macroeconómica, y los resultados que arroje se pueden obtener mediante la corrida con paquetería computacional.

ACTIVIDAD DE APRENDIZAJE

Investigar otros procesos estocásticos y realizar un cuadro en que se diferencie cada uno. Entregar en hojas blancas en la siguiente clase.

6.3 PREDICCIÓN

La información presente y pasada permite hacer una estimación acerca del futuro, a esto se le llama predicción. En el campo de la economía es ampliamente utilizada, mediante series temporales, pues permite planificar o prever el comportamiento de una variable explicativa.

Una categoría de métodos de predicción en los valores previamente observados en la serie de tiempo, y que se ocupan como variables independientes en los modelos de regresión, es el modelo autorregresivo integrado de promedios móvil (ARIMA). El método más amplio para el uso de esta categoría fue desarrollado por Box y Jenkins, llamándosele método de Box-Jenkins.

Así, el modelo general de series de tiempo que describe el componente estocástico se modela de la siguiente manera:

AR que significa autorregresivo, queda definido como sigue

$$(Y_t - \delta) = \alpha_1 (Y_{t-1} - \delta) + u_t \qquad 6.14$$

donde δ es la media de *Y.*

Y_t tiene un proceso estocástico autorregresivo de primer orden AR(1), es decir, el valor de *Y* en el tiempo *t* depende de su valor en el periodo anterior y un término aleatorio (α_1), este proceso se presenta debido a que u_t es el término de perturbación no correlacionado con media cero y varianza constante.

Si se considera este modelo como:

$$(Y_t - \delta) = \alpha_1 (Y_{t-1} - \delta) + \alpha_2 (Y_{t-2} - \delta) + u_t \qquad 6.15$$

al igual que en la expresión 6.14, los valores de *Y* de la 6.15 se encuentran expresados alrededor del valor de su media δ. En este caso, la expresión 6.15 muestra que el valor *Y* en el tiempo *t* depende de sus valores en dos periodos anteriores, se dice entonces que Y_t sigue un proceso autorregresivo de segundo orden AR(2).

Así de forma general la expresión 6.15 queda de la siguiente manera:

$$(Y_t - \delta) = \alpha_1 (Y_{t-1} - \delta) + \alpha_2 (Y_{t-2} - \delta) + \ldots + \alpha_p (Y_{t-p} - \delta) + u_t \qquad 6.16$$

ahora Y_t sigue un proceso autorregresivo de orden *p*, es decir AR(p).

En los tres modelos anteriores se ha considerado valores actuales y anteriores de *Y*, esto es un modelo de forma reducida.

Para el proceso de media móvil (MA) se considera un modelo de *Y* de la siguiente manera:

$$Y_t = \mu + \beta_0 u_t + \beta_1 u_{t-1} \qquad 6.17$$

μ es una constante y *u* es el termino de perturbación estocástico. En tanto, *Y* en el periodo *t* es una constante más un promedio móvil de los errores presentes y pasados. Se dice entonces que Y sigue un proceso de promedio móvil de primer orden, MA(1).

Un proceso MA(2) queda como sigue:

$$Y_t = \mu + \beta_0 u_t + \beta_1 u_{t-1} + \beta_2 u_{t-2} \qquad 6.18$$

y de forma general se expresa a continuación como:

$$Y_t = \mu + \beta_0 u_t + \beta_1 u_{t-1} + \beta_2 u_{t-2} + \ldots + \beta_q u_{t-q} \qquad 6.19$$

Este es un proceso MA(q). El proceso de media móvil es una combinación lineal de los términos de perturbación estocásticos.

Si se hace una combinación de los procesos AR y MA, Y_t sigue un proceso ARMA(1,1) y se determina con la siguiente expresión:

$$Y_t = \theta + \alpha_1 Y_{t-1} + \beta_0 u_t + \beta_1 u_{t-1} \qquad 6.20$$

en este caso se tiene un término autorregresivo y otro de media móvil, siendo θ un término constante. Así, en el proceso ARMA(p, q) hay *p* términos autorregresivos y *q* términos de media móvil.

Para el caso de la *I* significa modelo integrado, en este tema es necesario que la serie de tiempo muestre estacionariedad, es decir, que su media, su varianza y su covarianza, en los diferentes rezagos, sea la misma no importando el momento en que se mida, significa que no varían en el tiempo. Para que una serie de tiempo sea estacionaria se debe diferenciar *d* veces y luego aplicar el modelo ARMA(p, q). De esta manera la serie de tiempo original es ARIMA(p, d, q) y se dice que es una serie de tiempo autorregresiva integrada de media móvil.

P representa el número de términos autorregresivos, *d* es el número de veces que se efectuó la diferenciación ($Y_t - Y_{t-1}$) para hacerla estacionaria y el término q es el número de parámetros de media móvil. Cuando se tiene una ARIMA (3, 2, 3) significa que se realizaron dos diferenciaciones (d=2) antes de ser estacionaria, mientras que ha sido modelada con un proceso ARMA (2, 2), es decir, tiene dos términos autorregresivos y dos medias móvil.

Para hacer uso de la metodología de Box-Jenkins es necesario tener una serie de tiempo estacionaria, una vez que se ha realizado la(s) diferenciación(es).

El modelo estimado se utiliza para predicción, el cual supone características constantes a lo largo del tiempo. De esta manera, para la elaboración del modelo ARIMA se deben seguir cuatro etapas:

1. Identificación. Para identificar el modelo de serie de tiempo que describe la serie temporal en consideración, se emplea la función de autocorrelación (FAC) y la función de autocorrelación parcial (FACP), así como de los correlogramas que resulten, esto es los gráficos de FAC y FACP. Cada modelo está determinado por el comportamiento de la autocorrelación teórica y autocorrelación parcial teórica, se dice que es teórica cuando ARIMA(0, 0, q). Con esto, el modelo de series de tiempo a elegir es aquel cuya función de autocorrelación teórica y autocorrelación teórica parcial se asemeja a la FAC y a la FACP de la serie de tiempo observada.

   Para la identificación de un proceso de promedio móviles depende del número de coeficientes de la función de autocorrelación teórica estadísticamente significativos y de la forma de la función de autocorrelación parcial teórica. Para identificar cuántos coeficientes son estadísticamente significativos, se debe identificar el momento en que la función teórica se corta o desaparece después de un determinado retroceso, *q*.

   La persona que realiza la investigación sólo necesita comparar la FAC y la FACP del modelo autorregresivo con la FAC y la FACP de los promedios móviles.

   Para que la identificación se dé con alto grado de certeza, se deben tener mínimo 30 observaciones.
2. Estimación. Para estimar los parámetros de los términos autorregresivos y de media móvil incluidos en el modelo, es necesario haber identificado los valores apropiados de *p* y *q*. Para realizar esta estimación se pueden efectuar cálculos de mínimos cuadrados. Es importante mencionar que para realizar esta estimación se requiere de paquetería estadística, por lo que no es necesario llevar a cabo los desarrollos matemáticos.

3. Verificación. Una vez que se ha identificado el modelo y se ha realizado la estimación, es necesario observar si los parámetros del modelo caen dentro de los intervalos de estacionariedad. Otra prueba consiste en examinar el comportamiento diferencia o del residuo entre el dato observado y el que predice el modelo. El análisis de estos residuos se efectúa mediante la función de autocorrelación residual, los cuales si adquieren la forma de un proceso de ruido blanco (media cero, varianza constante y autocorrelaciones nulas), puede aceptarse el ajuste. Eso es que el modelo de series de tiempo ARIMA es estocástico.
4. Pronóstico. Una vez que se ha realizado la verificación del modelo ARIMA, éste se utiliza para pronósticos con un mínimo de error en la predicción.

Se tiene entonces que el modelo general de series de tiempo que describe el componente estocástico se denomina autorregresivo integrado de promedios móviles, ARIMA (p, d, q), puede ser descrito por un modelo autorregresivo o por uno de promedios móviles. Al darse la estimación de este modelo se puede efectuar la predicción, al suponer que se mantiene constante en el tiempo.

ACTIVIDAD DE APRENDIZAJE

Consultar y describir otros enfoques de predicción económica basados en las series de tiempo, determinar las ventajas y desventajas que presentan.
Entregar en hojas blancas la siguiente sesión.

AUTOEVALUACIÓN

I. Relacionar las siguientes columnas e indicar en el paréntesis la respuesta que corresponde a la afirmación.

1. Los modelos de series de tiempo han sido de utilidad en el análisis empírico y se estiman con facilidad por medio de …	( ) Modelo de series de tiempo
2. La función de autocorrelación (FAC) y la función de autocorrelación parcial (FACP), así como los correlogramas ayudan a identificar…	( ) Fluctuaciones irregulares ( ) p, q
3. Representan el número de términos autorregresivos y el número de parámetros de media móvil.	( ) Mínimos cuadrados ordinarios
4. Es aquella que recoge movimientos erráticos.	( ) Estocástico estacionario
5. Qué tipo de proceso se tiene cuando la media y la varianza es constante en el tiempo	

II. Completar las siguientes expresiones:

1. La _____ es una secuencia cronológica de observaciones de una variable o conducta particular durante un periodo determinado.
2. Al movimiento ascendente o descendente en un periodo determinado se le llama _____, el cual se refleja en un crecimiento o desvanecimiento en la serie.
3. Cuando en una serie de tiempo hay ausencia de cualquier tipo de variabilidad hay presencia de _____
4. El _____consiste en el examen del patrón histórico generado por el evento en observación con la esperanza.

5. Se le llama estimación _____a aquella que supone que la variable explicativa X_t es no estocástica, así como X_{t-1}, X_{t-2} y así sucesivamente.

Respuestas

I. Relacionar las siguientes columnas e indicar en el paréntesis la respuesta que corresponde a la afirmación.

1. Los modelos de series de tiempo han sido de utilidad en el análisis empírico y se estiman con facilidad por medio de …	(2) Modelo de series de tiempo
2. La función de autocorrelación (FAC) y la función de autocorrelación parcial (FACP), así como los correlogramas ayudan a identificar…	(4) Fluctuaciones irregulares
3. Representan el número de términos autorregresivos y el número de parámetros de media móvil.	(3) p, q
4. Es aquella que recoge movimientos erráticos.	(1) Mínimos cuadrados ordinarios
5. Qué tipo de proceso se tiene cuando la media y la varianza es constante en el tiempo	(5) Estocástico estacionario

II. Completar las siguientes expresiones:

1. La _____**serie de tiempo**_____ es una secuencia cronológica de observaciones de una variable o conducta particular durante un periodo determinado.
2. Al movimiento ascendente o descendente en un periodo determinado se le llama _____**variación de tendencia**____, el cual se refleja en un crecimiento o desvanecimiento en la serie.

3. Cuando en una serie de tiempo hay ausencia de cualquier tipo de variabilidad hay presencia de ____ **Estacionariedad**_____

4. El _____**análisis de serie de tiempo**_____consiste en el examen del patrón histórico generado por el evento en observación con la esperanza.

5. Se le llama estimación ___ **ad hoc**_____a aquella que supone que la variable explicativa X_t es no estocástica, así como X_{t-1}, X_{t-2} y así sucesivamente.

BIBLIOGRAFÍA

Gallastegui, Fernández, Alonso, *Econometría*, México, Pearson, 2005.

Goldberger, Arthur S., *Econometric Theory*, John Wiley & Sons, Inc., New York, 1964.

Maddala, G.S., *Econometría*, México, Mc Graw Hill, 1985.

Malinvaud, E., *Statistical Methods of Econometrics*, Rand Mc Nally & Co., Chicago, 1966.

Martínez, Garza Ángel, *Métodos econométricos. Proyecto de Investigación*. Colegio de Postgraduados, 1982.

Ramírez, Arellano Gerardo, *Introducción a la econometría*, Universidad Autónoma de Ciudad Juárez, 2005.

Samuelson, P. A., T.C. Koopmans, and J. R. N. Stone, “Report of the Evaluative Committee for Econometrica”, Econometrica, Vol. 22, No. 2, abril, 1954.

Theil, H., *Principles of Econometrics*, John Wiley & Sons, Inc., New York, 1971.

Tintner, Gerhard, *Econometrics*, John Wiley & Sons, Inc., New York, 1965.

Tintner, Gerhard, *Methodology of Mathematical Economics and Econometrics*, The University of Chicago, 1968.

Wooldridge, Jeffrey M., *Introducción a la econometría. Un enfoque moderno*, Thomson Learning, 2001.

GLOSARIO

Análisis de regresión: Tipo de análisis utilizado para describir la estimación y la inferencia en el modelo de regresión.

Análisis empírico: Estudio que utiliza datos en un análisis econométrico formal para probar una teoría, estimar una relación o determinar la eficiencia de un proyecto establecido.

Análisis residual: Análisis que estudia el signo y magnitud de los residuos de determinadas observaciones después de estimar el modelo de regresión.

Aleatorio: Se da en un experimento repetido indefinidamente presenta siempre resultados totalmente impredecibles.

Asimetría: Es cuando los datos pierden su simetría respecto a la media.

Autorregresivo: Una variable o conjunto de variables se explican al menos en parte, en función del pasado de la misma variable.

Ceteris paribus: Todos los demás factores relevantes se mantienen fijos.

Coeficiente de correlación: Es el cociente de dividir la covarianza de una distribución bidimensional entre las desviaciones típicas de X e Y respectivamente.

Coeficiente de determinación: Es el cociente entre la varianza explicada y la total en un ajuste a la recta de regresión.

Covarianza: Es la varianza conjunta en una distribución en la que se encuentran dos variables *X, Y*. Es el cociente del producto de la diferencia de la media de *X* con los X_i, con la media de *Y* y *Yi*, entre el número de observaciones *X*-*Y.*

Correlación: Es la relación que existe entre dos variables *X*, *Y*. Su valor está entre -1 y 1. Al observase un valor negativo significa que mientras una variable crece, la otra tiende a decrecer, representa que hay una relación inversa. Si el valor es positivo hay una relación directa entre las variable, es decir, ambas (*X*, *Y*) van en la misma dirección.

Dato: Es el valor cuantitativo o cualitativo que representa un atributo o medida en la población.

Desviación típica: Es la raíz cuadrada de la varianza.

Distribución bidimensional: Dadas las variables *X*, *Y* y a partir de la realización de una prueba se obtienen sus medidas.

Distribución muestral: Distribución de probabilidad de un estimador en todos los resultados posibles de la muestra.

Error de predicción: Diferencia entre el resultado actual y una predicción de tal resultado.

Esperanza condicional: Valor esperado o promedio de una variable aleatoria, llamada dependiente o explicada, que depende de los valores de otra u otras variables, llamadas independientes o explicativas.

Estacionariedad en covarianza: Proceso de series de tiempo con media y varianzas constantes, y en donde la covarianza entre cualquiera de dos variables aleatorias de la serie depende de la distancia que las separa.

Estadística de prueba: Regla para probar hipótesis donde cada resultado muestral produce un valor numérico.

Estimación: A partir de un parámetro de la población y con la aplicación de una serie de cálculos, su valor será igual al estadístico de prueba que se calculo con la muestra.

Estimador: Regla para combinar los datos para producir un valor numérico para un parámetro poblacional; la forma de la regla no depende de la muestra obtenida.

Estimador insesgado: Es cuando su media muestral coincide con el parámetro.

Grados de libertad (gl): En el análisis de regresión es el número de observaciones menos número de parámetros estimados.

Heterocedasticidad: Dadas las variables explicativas, la varianza del término de error no es constante.

Homocedasticidad: En un modelo de regresión los errores tienen una varianza constante que depende de las variables explicativas.

Intervalo de confianza: Regla para establecer un intervalo aleatorio tal que el porcentaje de todos los datos, determinado por el nivel de confianza, proporcione intervalo que comprenda el valor proporcional.

Mejor estimador lineal insesgado (MELI): Entre todos los estimadores lineales insesgados, el que tenga la mínima varianza.

Mínimos cuadrados ordinarios (MCO): Método para estimar los parámetros de un modelo de regresión lineal. Los estimadores de mínimos cuadrados ordinarios se obtienen minimizando la suma residual de cuadrados.

Modelo de regresión lineal simple: Modelo en que la variable dependiente es una función lineal de una sola variable independiente, más un término de error.

Modelo de rezagos distribuidos: Modelo de series de tiempo que relaciona la variable dependiente con los valores actuales y pasados de una variable explicativa.

Modelo econométrico: Ecuación que relaciona la variable dependiente con un conjunto de variables explicativas y perturbaciones inobservables, en el que los parámetros desconocidos de la población determinan el efecto ceteris paribus de cada explicativa.

Modelo económico: Relación derivada de la teoría económica o de un razonamiento económico menos formal.

Muestra: Elementos extraídos aleatoriamente de una población en estudio. A partir de la aplicación de inferencias estadísticas de la muestra se puede deducir los resultados de la población.

Multicolinialidad: Se refiere a la correlación entre las variables independientes de un modelo de regresión.

Nivel de confianza: En una inferencia estadística se tiene que la probabilidad de que un valor se encuentre dentro del intervalo de confianza, el de mayor uso es de 95%, pero también son usuales el de 90% y 99%.

Nivel de significancia: Es la probabilidad de que los valores caigan en la región de rechazo, el nivel de significancia es α.

Parámetro: Valor desconocido que describe una relación poblacional.

Parámetro de intercepción: Parámetro en un modelo de regresión que da el valor esperado de la variable dependiente cuando todas las independientes son igual a cero.

Población: Conjunto de elementos que forman parte de una misma especie y de los cuales se puede efectuar un estudio o investigación. Cuando no se cuenta con los recursos necesarios para obtener la población, se recurre a tomar una muestra, que es aquélla que forma parte de la población.

Predicción: Estimación de un resultado que se obtiene introduciendo valores específicos de las variables explicativas en un modelo estimado.

Proceso con tendencia: Proceso de series de tiempo cuyo valor esperado es una función creciente o decreciente del tiempo.

Proceso estocástico: Sucesión de variables aleatorias indexadas en el tiempo.

Regresión: A partir de un modelo econométrico en el que interviene una variable dependiente, y una o varias variables independientes, se aplica este método estadístico, y de esta manera determinar la relación que existe entre las variables.

Residuo: Diferencia entre el valor actual y el ajustado, hay un residuo para cada observación de la muestra usada para obtener una línea de regresión de mínimos cuadrados ordinarios.

Teorema de Gauss-Markov: Teorema que afirma que en contexto de las suposiciones de Gauss-Markov, el estimador de MCO es MELI, dependiendo de los valores en la muestra de las variables explicativas.

Variable: Del conjunto de datos que conforman una muestra, cada uno de ellos toma un valor distinto.

Variable cualitativa: Característica que recoge una cualidad de los individuos de la muestra.

Variable cuantitativa: Es aquélla que se puede cuantificar o medir dado un conjunto de elementos.

Varianza: Es una medida de dispersión, en la que a partir de una muestra de mediciones, se tiene que es la suma del cuadrado de la diferencia entre el valor X_i y la media de X_1, X_2, X_3,…. X_n dividida entre el número de observaciones y

su valor mayor o igual a cero, al sacar la raíz cuadrada de la varianza se obtiene la desviación estándar.