UNIDAD II

¿Qué es Análisis de regresión lineal?

UNIDAD II

¿Qué es Análisis de regresión lineal?

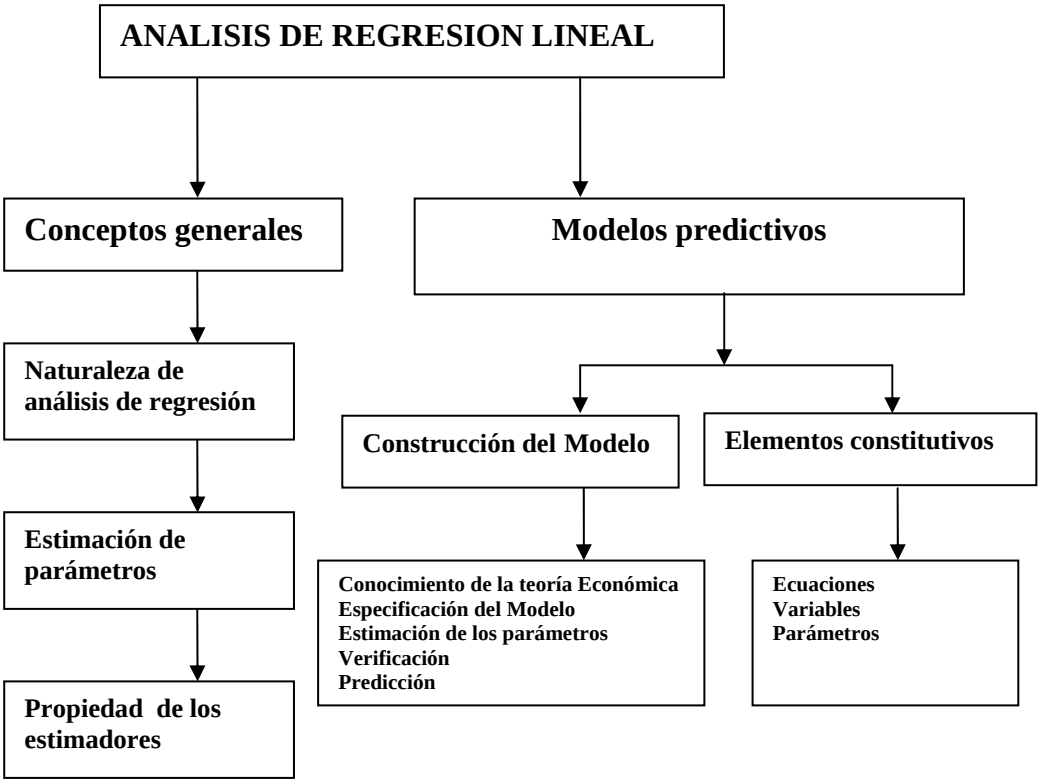
"…el pensamiento estadístico algún día será tan necesario para la ciudadanía como la capacidad de leer y escribir…"

H.G. Wells (hace 100 años)

- **¿Cuáles son los conceptos generales del análisis de regresión lineal?**
- **¿Qué es un modelo, cuáles son sus elementos constitutivos?**
- **¿Cuál es el proceso de construcción de un modelo?**
- **¿Cuáles son las características y supuestos del Modelo Lineal General?**
- **¿Qué es estimación de parámetros?**
- **¿Cuáles son las propiedades de los estimadores?**
- **¿Cómo se halla la estimación de la varianza de término de perturbación?**

ANALISIS DE REGRESION LINEAL

ESQUEMA CONCEPTUAL



COMPETENCIAS A LOGRAR

CONCEPTUAL	PROCEDIMENTAL	ACTITUDINAL
Explica qué es el análisis de regresión lineal, el modelo predictivo, sus procesos de construcción, clases y supuestos.	**Aplica las técnicas apropiadas del análisis de regresión lineal.**	**Estima o predice la media o valor promedio de la variable dependiente con base en los valores conocidos.**

CONCEPTOS –CLAVE

Variable, regresión, modelos, correlación, parámetro, muestra, población, estimador, varianza.

LECCIÓN 1

CONCEPTOS GENERALES

1. VARIABLES

Es la representación de un fenómeno (característica), el cual puede tomar diferentes valores. También se define como la característica de la muestra o de la población nque se observa.

Ejemplo:
El precio de un bien, cantidad producida, gastos en publicidad, temperatura, regiones, educación, tipo de gobierno, prioridades, etc.

2. POBLACIÓN

Conjunto total de unidades definidas en un tiempo y espacio determinadas por el investigador para realizar un análisis. Por ejemplo, si el primer ejecutivo de una gran empresa textil desea estudiar la producción de todas las fábricas en el año 2003 en el Perú; la población estaría formada por todas las plantas textiles ubicadas en el Perú en el 2003.

Ejemplo:
Todos los asalariados del Perú en el 2003.

3. PARÁMETRO

Es una medida descriptiva de la población, la cual es de interés para el investigador.

Ejemplo:
La producción total de las plantas textiles en el 2003 o los ingresos medios de todos los asalariados del Perú.

4. MUESTRA

Es la porción representativa de la población, la cual es obtenida cuando la población es demasiado grande para analizarla en su totalidad.

Ejemplo:
Relación de 400 Hogares de Lima Metropolitana que son usados para medir el ingreso de los asalariados en el 2003, esto representa una pequeña parte del total de la población (Total de hogares de Lima Metropolitana en el 2003).

5. ESTADÍSTICO (O ESTIMADOR)

Es cualquier medida descriptiva de una muestra y es usado para la estimación del parámetro correspondiente de la población.

Ejemplo:
El Ingreso medio de una muestra de 500 trabajadores calculados por el Ministerio de Trabajo es un estadístico.

6. REGRESIÓN

Es una expresión cuantitativa de la naturaleza básica de la relación entre las variables dependientes con la independiente.

Ejemplo:
Dado un modelo de regresión simple con una variable independiente, el modelo determinará si las dos variables (independiente y dependiente) tienden a desplazarse en la misma dirección - (las dos crecen o decrecen al mismo tiempo) o en sentidos opuestos (una aumenta cuando la otra disminuye). También indicará la cantidad en que Y cambiará cuando la variable independiente varíe en una unidad.

7. REGRESIÓN SIMPLE

Se presenta cuando la variable dependiente *Y* esta en función de una sola variable independiente.

La notación que lo expresa es:

$$Y = f(X_1)$$

8. REGRESIÓN MÚLTIPLE

Abarca dos o más variables independientes. Si se dice que *Y* depende de más de dos variables independientes, podemos escribir:

$$Y = f(X_1, \cdots, X_n)$$

9. CORRELACIÓN

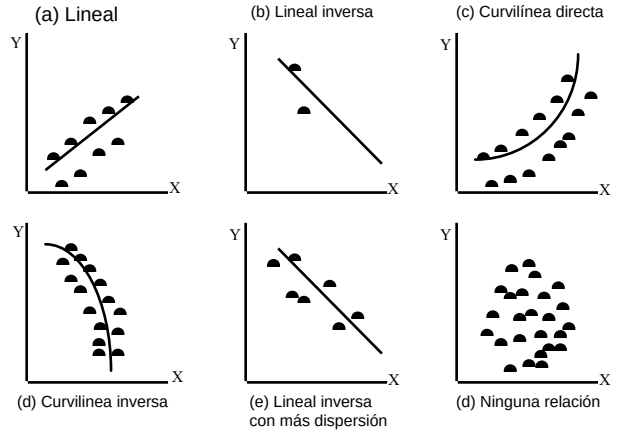
Determina la fuerza de la relación; es decir, mientras que la regresión describe la naturaleza básica de la relación entre las dos variables, la correlación mide la solidez de dicha relación.

Ejemplo:
Se puede estar interesado en conocer la correlación entre el gasto en publicidad y las ventas de una empresa; entre el nivel de producción y la inflación mensual; etc.

10. DIAGRAMAS DE DISPERSIÓN

Representan de forma gráfica las relaciones entre 2 variables (la dependiente con una independiente), lo habitual es colocar la variable dependiente en el eje vertical y la independiente en el eje horizontal.

Relaciones posibles entre X y Y vistos en diagramas de dispersión



11. ANÁLISIS DE REGRESION

Estudia la relación existente entre una variable endógena o dependiente (Y) y una o más variables exógenas o independientes (X), con el objeto de estimar la media o valor promedio poblacional de la variable dependiente en términos de los valores conocidos o fijos de las independientes.

Ejemplo:

Cuando se analiza una regresión, se trata de predecir el valor promedio de la variable; Por ejemplo: predecir el saldo de la cuenta de capitales teniendo información muestral de la tasa de interés; predecir el monto total de créditos conociendo la tasa de encaje bancario; etc.

12. DEFINICIÓN DE MODELOS

Es un conjunto de relaciones funcionales, generalmente interrelacionadas., el cual puede ser definido como una representación simplificada e idealizada de la realidad. De esta definición se desprenden dos comentarios que reflejan en parte las ventajas y limitaciones del uso de modelos:

a. Un modelo no es la realidad, pero tampoco es completamente independiente de ella.
b. Como representación idealizada, parte de la interpretación que el constructor del modelo posee de la realidad.

Los modelos están conformados por afirmaciones o hipótesis de comportamiento de unas variables en base al comportamiento de otras, que se manifiestan en relaciones funcionales expresadas matemáticamente.

LECCIÓN 2

ELEMENTOS CONSTITUTIVOS DE UN MODELO

Los elementos que integran un modelo son: las ecuaciones, las variables y los parámetros.

1. ECUACIONES

Una ecuación es una relación ponderada entre variables, que se verifica para determinados valores numéricos.

Un modelo se especifica mediante una ecuación o varias ecuaciones, en donde cada una de éstas pretende explicar un sector (agrícola, minero, manufacturero, transporte, etc.) o una categoría (consumidores, productores, intermediarios, inversionistas, etc.) de la actividad objeto de investigación.

2. VARIABLES

Las variables son magnitudes susceptibles de modificarse cuantitativamente, dentro de un cierto margen o campo de variabilidad.

Clasificación de las Variables

a. Variables endógenas o dependientes
b. Variables predeterminadas o independientes
   b.1 Exógenas
   b.2 Endógenas con retardo
c. Variables aleatorias o estocásticas

a. Variables endógenas o dependientes.- Se caracterizan porque sus valores se determinan como soluciones particulares del modelo.

b. Variables predeterminadas o independientes: Son aquellas cuyos valores no se obtienen por la solución del modelo sino que provienen fuera del mismo. Ellas contribuyen a explicar el comportamiento de las variables endógenas de un modelo sin ser explicadas por el modelo mismo. Se clasifican en:

b.1 Las **Variables Exógenas** incluyen variables económicas y no económicas que explican el modelo, pero no son explicadas por éste.
b.2 Las **Variables Endógenas con Retardo.-** Son aquellas que actúan como variables explicativas o como datos del pasado que contribuyen a explicar el presente.

c. Variables aleatorias o estocásticas.- Se incluyen para justificar la omisión de variables explicativas, los errores en la especificación de las ecuaciones y errores en la medición de variables endógenas.

Ejemplo ilustrativo de clasificación de variables:

Dado el siguiente modelo:

$$C_t = a_0 + a_1(Y_{t-1} - T_{t-1}) + a_2 C_{t-1} + \mu_{1t}, \qquad 0 < a_1 < 1$$

$$I_t = b_0 + b_1(Y_{t-1} - Y_{t-2}) + \mu_{2t}, \qquad b > 0$$

$$T_t = g_0 + g_1 Y_t + \mu_{3t}, \qquad 0 < g_1 < 1$$

$$Y_t = C_t + I_t + G_t$$

siendo :

C_t = consumo nacional **T_t** = impuestos **G_t** = gasto de gobierno

Y_t = ingreso nacional **I_t** = inversión neta

Clasificando sus variables, tenemos lo siguiente:

a. Variables Endógenas o Dependientes :

C_t = Consumo del periodo de estudio.
I_t = Inversión en el periodo de estudio.
T_t = Impuestos del periodo de estudio.
Y_t =Ingreso Nacional del periodo de estudio.

Estas variables son la razón de ser del modelo, al resolver el modelo debemos de hallarlas.

b. Variables Predeterminadas :

b.1 Variables Endógenas con Retardo:

Y_{t-1} = Ingreso Nacional del periodo anterior.
T_{t-1} = Impuestos del periodo anterior.
C_{t-1} =Consumo Nacional del periodo anterior.
Y_{t-2} = Ingreso Nacional de 2 periodos anteriores.

b.2 Variable Exógena:

G_t = Gasto de gobierno.

Todas estas son variables Predeterminadas que ayudan a explicar a las variables endógenas.

c. Variables Estocásticas o Aleatorias :

μ_{1t} μ_{2t} μ_{3t}

Son importantes porque con ellas se diferencia un modelo económico de uno econométrico; su ubicación en el modelo justifica:

- Las omisiones de otras variables explicativas.
- Los errores por hacer las ecuaciones.
- Los errores al evaluar las variables dependientes.

3. PARÁMETROS

Los parámetros son los factores de ponderación entre las variables incluidas en las ecuaciones de un modelo.

Ejemplo ilustrativo de interpretación de los parámetros:

Del modelo anterior, en la Ecuación (1): $C_t = a_0 + a_1(Y_{t-1} - T_{t-1}) + a_2 C_{t-1} + \mu_{1t}$

a_0: Vendría a ser el consumo autónomo. Este parámetro no está afectado por ninguna variable explicativa. También significa que así no hubiera consumo en el periodo precedente y también sino hubiera ingreso disponible el periodo anterior, siempre va a existir un consumo autónomo dado. Estadísticamente a_0 es el intercepto.

a_1: Esta es la Propensión Marginal a Consumir., y nos muestra en cuánto varía el consumo cuando varía en una unidad monetaria el Ingreso disponible (del periodo anterior, para este caso). En efecto esta variable no puede ser cero ya que eso implicaría que el consumo actual no estaría afectado por el ingreso disponible del periodo anterior (lo cual es falso porque siempre existe alguna participación). Esta variable tampoco puede ser 1, porque quiere decir que se gasta todo el ingreso disponible en el consumo, lo cual también es falso porque siempre se ahorra algo. Por tanto $0<a_0<1$.

a_2: Nos muestra en cuánto varía el consumo actual cuando varía en una unidad el consumo del periodo anterior. Acá entra a tallar lo que es los patrones de consumo de la gente.

En la Ecuación (2): $I_t = b_0 + b_1(Y_{t-1} - Y_{t-2}) + \mu_{2t}$

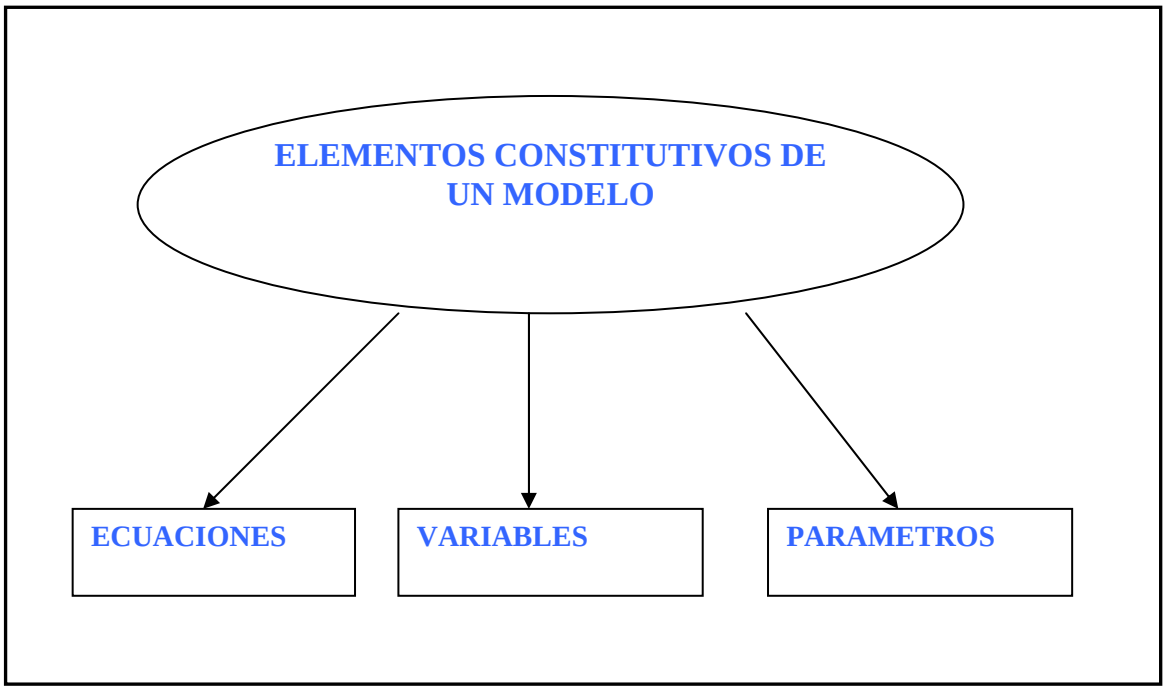
b_0: Viene a ser la Inversión Autónoma, la inversión que no está afectada por ninguna variable explicativa (ni aleatoria). Es decir, implica que así no halla existido ingresos en periodos anteriores, siempre se va a invertir algo (b_0).

b_1: Nos dice en cuánto varía la inversión ante cambios en periodos anteriores del Ingreso. Es la sensibilidad de la inversión ante cambios en ingresos anteriores.

En la Ecuación (3): $T_t = g_0 + g_1 Y_t + \mu_{3t}$

g_0: Representa el Impuesto Autónomo, esto quiere decir, que así no existan ingresos, siempre el gobierno debe recaudar impuestos (g_0).

g_1: Nos dice en cuánto varían los impuestos cuando varía en una unidad monetaria los ingresos. Mide la sensibilidad de los impuestos ante cambios en los ingresos.



LECCIÓN 3

PROCESO DE CONSTRUCCIÓN DE UN MODELO

El proceso de construcción de un modelo se puede presentar como una secuencia de etapas que a continuación se consideran:

1. **Conocimiento de la Teoría Económica:** esta etapa se refiere al conocimiento de aspectos de la teoría económica requeridos para la especificación del modelo econométrico a aplicarse. Esta etapa es la más importante en la construcción de un modelo.

2. **Especificación del Modelo Econométrico**: son todos aquellos pasos previos a la estimación, es decir, precisar las variables y sus relaciones, la estructura lógica del modelo, el período de tiempo a estudiarse, la identificación, etc.

3. **Estimación**: una vez especificado el modelo, la siguiente tarea consiste en la obtención de estimaciones (valores numéricos) de los parámetros del modelo a partir de los datos disponibles, generalmente proporcionados por el estadístico económico. Estas estimaciones le dan contenido empírico a la teoría económica.

4. **Verificación**: (Inferencia Estadística) habiendo obtenido estimaciones de los parámetros, la siguiente labor es la verificación de las hipótesis estadísticas y económicas; es decir, desarrollar los criterios apropiados para lograr establecer si las estimaciones obtenidas están de acuerdo con lo que se espera de la teoría que se esta verificando.

   Dentro de las hipótesis estadísticas nos referiremos a las pruebas de hipótesis acerca de la significación de los parámetros (Pruebas "t" y "F"), de las relaciones del modelo (coeficientes de correlación), de la validez de los supuestos de las perturbaciones (Pruebas de heterocedasticidad, autocorrelación, multiconlinealidad, etc.).

5. **Predicción o Pronóstico**: los modelos econométricos ya estimados se utilizan frecuentemente para predecir el (los) valor (es) futuro (s) de la variable dependiente con base a valores conocidos o esperados en el futuro para la (s) variable (s) explicativa (s). Supongamos por ejemplo, que el Gobierno contempla la posibilidad de una reducción de los impuestos personales con el fin de estimular su quebrantada economía, ¿Cuál será el efecto de esta política sobre el consumo y por consiguiente sobre el empleo y el ingreso?

Ejemplo Ilustrativo del Proceso de Construcción de un Modelo

1. Conocimiento de la Teoría Económica.

Para ilustrar, consideramos la teoría Keynesiana del consumo, Keynes dice: "La Ley psicológica fundamental consiste en que los hombres están dispuestos, por regla general y en promedio, a aumentar su consumo a medida que su ingreso crece, aunque no tanto como el crecimiento de dicho ingreso"[1].

[1] John Maynard Keynes: "La Teoría General del Empleo, el Interés y el Dinero". New York, 1936, p. 96.

Keynes afirma que la propensión marginal a consumir (PMC), la razón de cambio del consumo por un cambio unitario en el ingreso, es mayor que cero pero menor que uno.

2. Especificación del Modelo Econométrico

En nuestra ilustración, aunque Keynes postula una relación positiva entre consumo e ingreso, no especifica la forma precisa de la relación funcional entre las dos variables.

Para simplificar un economista matemático puede sugerir la siguiente forma para la función consumo de Keynes:

$$(1) \quad C = \alpha + \beta * Y$$

donde:

C = gastos de consumo
Y = ingreso
α , β = constantes o parámetros

el coeficiente de Y representa la pendiente o PMC.

La ecuación (1), que afirma que el consumo está relacionado linealmente con el ingreso, es de interés limitado para el analista, por cuanto supone una relación exacta o determinista entre consumo e ingreso. Sin embargo, las relaciones entre las variables económicas son inexactas.

De este modo, si fuéramos a obtener las cifras de gastos de consumo y de ingreso disponible (descontados los impuestos) de una muestra de 5,000 familias, por ejemplo, y luego dispusiéramos los datos en una gráfica, colocando los gastos de consumo en el eje vertical y el ingreso disponible en el eje horizontal, con toda seguridad, no esperaría que las 5,000 observaciones ocurrieran exactamente sobre la línea de la ecuación (1), pues *además del ingreso, existen otras variables que afectan los gastos de consumo; por ejemplo, el tamaño de la familia, la edad de sus miembros, el tiempo de constitución de la familia, la religión y otros factores que ejercen influencia en el consumo.*

Para tener en cuenta la relación inexacta entre las variables económicas, el econometrista debe modificar la función de consumo determinista de (1), de la siguiente manera:

$$(2) \quad C = \alpha + \beta * Y + \mu$$

en la que "μ" es variable aleatoria o estocástica con propiedades probabilísticas bien definidas.

La ecuación (2), plantea la hipótesis de que la variable dependiente C (consumo) está relacionada linealmente con la variable explicativa Y (ingreso), aunque no de manera exacta, puesto que está sujeta a variaciones individuales.

3. Estimación:
Así por ejemplo, si en el estudio de la función consumo Keynesiana, se encuentra que $\beta = 0.8$ este valor no solo proporciona una estimación numérica de la PMC, sino que corrobora la hipótesis Keynesiana según la cual, la PMC es menor que uno.

¿Cómo se estiman los parámetros α y β? Eso lo trataremos posteriormente en el Modelo Lineal General.

4. Verificación: (Inferencia Estadística)
Dentro de las hipótesis estadísticas nos referiremos a las pruebas de hipótesis acerca de la significación de los parámetros (Pruebas "t" y "F"), de las relaciones del modelo (coeficientes de correlación), de la validez de los supuestos de las perturbaciones (Pruebas de heterocedasticidad, autocorrelación, multicolinealidad, etc.). En cuanto a las hipótesis económicas nos referimos a la interpretación del modelo, su validez en cuanto a reproducir el período de análisis y a su poder predictivo.

Como se vio anteriormente, Keynes pretendía que la PMC fuese positiva pero menor que uno. Supongamos, por otro lado, que en un estudio de la función consumo se encuentra que la PMC = 0.9; si bien es cierto que este resultado es menor que 1, nos podemos preguntar si es suficientemente menor que 1 como para que logremos convencernos de que no es el resultado accidental de un proceso de muestreo. En otras palabras, ¿Es esta estimación estadísticamente menor que 1? Si es así, adquiere respaldo la afirmación Keynesiana, de lo contrario queda refutada.

5. Predicción o Pronóstico
Si la PMC = 0.8; quiere decir que si el ingreso aumenta en una unidad monetaria, se producirá finalmente un aumento en el consumo igual a 0.8 u.m.

Otro Ejemplo del Proceso de Construcción de un Modelo: Estimación de un Modelo Importaciones: 1998 –2004

1. Especificación del Modelo

El objetivo es el de encontrar una relación de largo plazo que explique la evolución de las importaciones dentro del periodo de estudio el mismo que se caracterizo por estar en una plena apertura comercial desde principio de los noventa.

Se sugiere el siguiente modelo para las importaciones:

$$\text{Limport} = \beta_1 - \beta_2 \text{LTCR} + \beta_3 \text{LPBI}$$

Donde:

Limport : Logaritmo de las Importaciones
LTCR : Logaritmo del Tipo de Cambio Real

LPBI : Logaritmo del Producto Bruto Interno

$\beta_1, \beta_2, \beta_3$: Constantes o parámetros

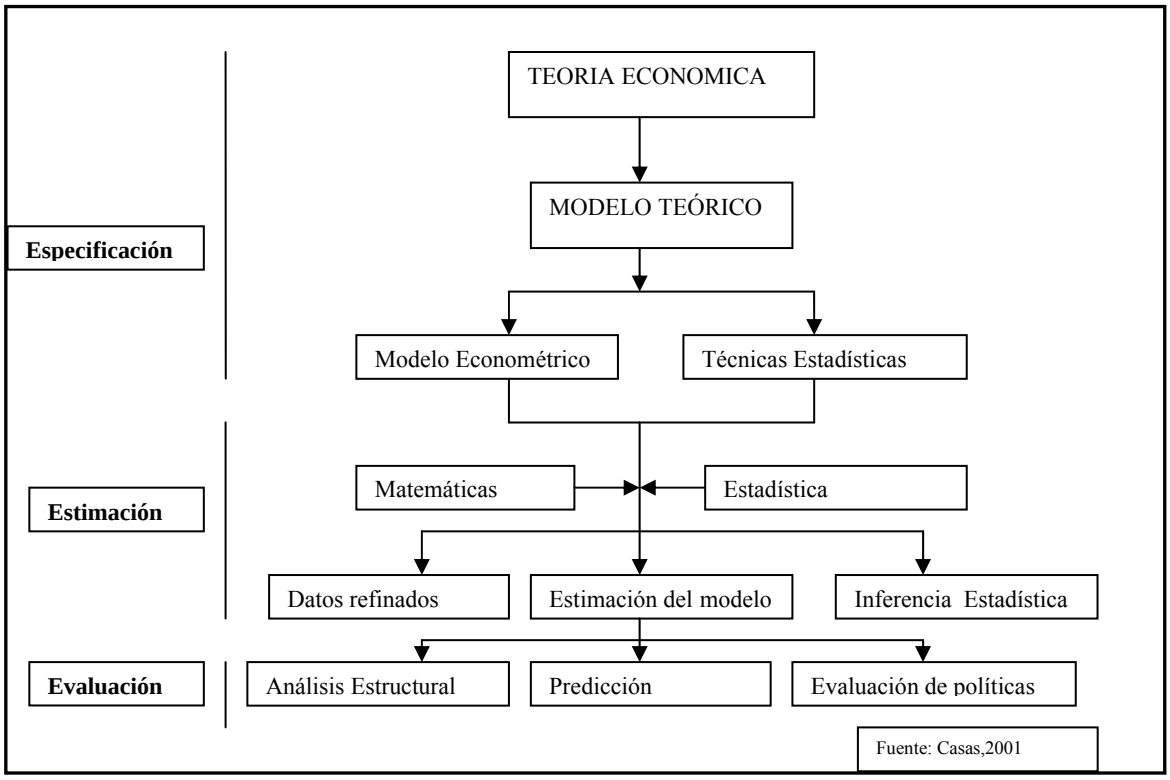
A continuación se deben realizar las siguientes etapas:

1. Estimación, Inferencia e Interpretación.
2. Test de Estabilidad de los Parámetros
   a. Test de Residuos Recursivos
   b. Test del Cusum y Cusum Cuadrado
   c. Test de Coeficientes Recursivos
   d. Test de Punto de Quiebre de Chow
   e. Test de Predicción de Chow

3. Normalidad de las Perturbaciones
4. Análisis de Multicolinealidad
5. Análisis de Heteroscedasticidad
   a. Test de White Sin Términos Cruzados
   b. Test de White Con Términos Cruzados
   c. Test ARCH LM

6. Análisis de Autocorrelación
   Para analizar la existencia de autocorrelación, podemos verificarlo utilizando los siguientes test:
   a. Test Durbin-Watson
   b. Test LM de Correlación Serial
   c. Test Box-Pierce Q

7. Especificación del Modelo
8. Evaluación de la Predicción


TEORIA ECONOMICA
MODELO TEÓRICO
Especificación
Modelo Econométrico
Técnicas Estadísticas
Matemáticas
Estadística
Estimación
Datos refinados
Estimación del modelo
Inferencia Estadística
Evaluación
Análisis Estructural
Predicción
Evaluación de políticas
Fuente: Casas,2001

LECCIÓN 4

NATURALEZA DEL ANÁLISIS DE REGRESIÓN

El análisis de regresión está relacionado con el estudio de la dependencia de una variable, la variable dependiente, que está en función de una o más variables explicativas con la perspectiva de estimar y/o predecir el valor (poblacional) medio o promedio de la primera en términos de valores conocidos o fijos (en muestreos repetidos) de las segundas.

El objetivo es determinar una ecuación de regresión que permita pronosticar el valor de una variable (denotado por Y; denominado variable dependiente) en base a otra variable (denotada por X; llamada variable independiente).

1. REGRESIÓN LINEAL SIMPLE

Establece que la variable dependiente Y es función de una sola variable independiente (X).
La notación que lo expresa es:

$$Y = \beta_1 + \beta_2 X + \mu$$

Donde:

β_1 : Es el parámetro constante del modelo.
β_2 : Es la pendiente de la ecuación poblacional
μ : Es el término de perturbación o error del modelo.

La ecuación de regresión estimada queda definida por:

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X$$

2. REGRESIÓN LINEAL MÚLTIPLE

Es la ampliación de la regresión lineal simple a dos o más variables explicativas, es decir con la regresión lineal múltiple se pueden predecir valores de la variable dependiente (Y) a través de varias variables explicativas (X_2, X_3,…,X_k)

El Modelo Estadístico es:

$$Y = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \ldots + \beta_k X_k + \mu$$

Donde
Y : Es la variable independiente
X_i: Son las variables explicativas (i = 2,…, k)

β_i : Son los parámetros correspondientes a cada variable (i = 2,…, k)

β_1 : Es el parámetro constante del modelo.

μ : Es el término de perturbación o error del modelo.

Al estimar los parámetros se determina la ecuación de regresión estimada:

$$\hat{Y} = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 + \ldots + \hat{\beta}_k X_k$$

Una vez verificado las condiciones del modelo se pueden realizar las predicciones a través de la ecuación anterior.

3. CURVA DE REGRESIÓN POBLACIONAL (Regresión de Y sobre X)

Es aquella que muestra el lugar geométrico de las medias condicionales o esperanzas de la variable endógena para los valores fijos de las variables exógenas.

Ejemplo Ilustrativo

- Se reúnen datos (X Y) cuya relación se desea estudiar y se organiza la información en una tabla que represente la **población.**

  Por ejemplo se ha efectuado una encuesta de ingresos y gastos a una población de 60 familias, que viven en un centro poblado. La información se presenta en el cuadro adjunto.

  Se desea estudiar la relación entre:
  Y = Gasto de consumo de la familia.
  X = Ingreso de la familia disponible
  Se desea predecir el nivel de la media poblacional del Y (Gasto de consumo de la familia)

- Se organiza y representa la distribución de los valores que toma Y, condicionada a los valores dados de X.

  Supuestamente, se han formado las 60 familias en 10 grupos cada uno de ellos tiene los ingresos iguales y se examinan los gastos de consumo de las familias (X). Por ejemplo con ingresos de 800 soles, existen 5 familias cuyos gasto de consumo de la familia se encuentran en el rango de S/. 520 a S/. 680. Cuando X es S/.1850 hay 5 familias cuyos gastos de consumo se encuentran entre S/.1080 y S/. 1440

	Ingreso de las Familias (X)									
Y	**650**	**800**	**950**	**1100**	**1250**	**1400**	**1550**	**1700**	**1850**	**2000**
Gasto de consumo Familiar por mes (S/.)	440	520	640	640	800	880	960	1080	1080	1240
	480	560	680	680	920	920	1080	1120	1200	1280
	520	560	720	760	840	960	1120	1080	1320	1320
	560	640	680	720	960	1040	1200	1200	1400	1400
	600	680	760	800	880	1080	1160	1240	1440	1440
		640		920	1000	1120		1320		1480
				960				1360		1560
				880						1480
E(y/x)	520	600	696	795	900	1000	1104	1200	1288	1400

- Se calcula las **probabilidades condicionales** p(Y/X) que se lee: probabilidad que Y tome un valor, dado que X ha tomado un determinado valor.

  Por ejemplo: la probabilidad de que Y tome un valor de S/.520 cuando X es igual a S/.650 es igual a 1/5, por que el número de familias que tienen este nivel de ingreso es 5, y sólo una gasta S/.520.

  $p(Y = 520/ X = 650) = 1/5$

  La probabilidad de que Y sea igual a S/.1400 cuando X es igual a S/.2000 es igual a 1/8, por que existen 8 familias que tienen un ingreso de S/.2000 aunque cada uno tiene diferentes niveles de consumo. El valor de S/.1500 sólo lo muestra una de las familias.

  $p(Y = 1400/ X = 2000) = 1/8$

  Para cada una de las distribuciones de probabilidad condicional de Y podemos calcular su media o valor promedio, conocido como la **media condicional o esperanza condicional,** que se denota: $E(Y/X = X_i) = \sum Y * p(Y / X)$

  En el ejemplo:

  Esperanza promedio: $E(Y/X = 800) = 520\left(\frac{1}{6}\right) + 560\left(\frac{2}{6}\right) + 640\left(\frac{2}{6}\right) + 680\left(\frac{1}{6}\right) = 600$

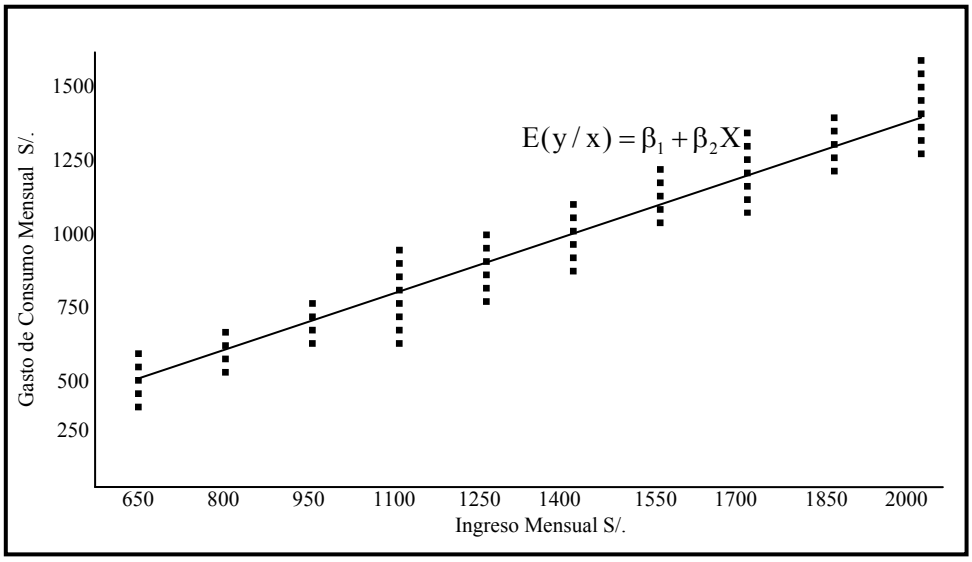
  Entonces, el consumo promedio de las familias que ganan S/.800 es S/.600.

  Esperanza promedio:

  $$E(Y/X = 2000) = 1240\left(\frac{1}{8}\right) + 1280\left(\frac{1}{8}\right) + 1320\left(\frac{1}{8}\right) + 1400\left(\frac{1}{8}\right) + 1440\left(\frac{1}{8}\right) + 1480\left(\frac{2}{8}\right) + 1560\left(\frac{1}{8}\right) = 1400$$

  El consumo promedio de las familias que ganan S/.2000 es S/.1400.

  Luego, observamos las cifras en el siguiente diagrama de dispersión



Se observa que el valor promedio del gasto de consumo tiende a aumentar a medida que el ingreso aumenta.

4. FUNCIÓN DE REGRESIÓN POBLACIONAL (FRP)

Para la construcción de la función de regresión poblacional la curva de regresión debe expresar todos los valores promedios de la variable dependiente para todos los valores fijos de la variable explicativa.

La regresión poblacional nos muestra cómo el valor promedio de Y varía en relación a los valores de la variable X.

$$E(Y/X_i) = f(X_i) = \beta_1 + \beta_2 X_i$$

En el análisis de regresión, se quiere estimar la FRP, es decir estimar los valores de β_1 y β_2 no conocidos con base en las observaciones de Y y X.

Ejemplo Ilustrativo
En el ejemplo anterior, se trata de los valores promedios de consumo en cada valor fijo del ingreso.

$$E(Y/X) = f(X_i) = \beta_1 + \beta_2 X_i$$

Donde:
β_1, β_2 son parámetros desconocidos pero fijos que se denominan **coeficiente de regresión**, también llamados **intercepto** y **coeficiente de la pendiente** de la recta formada respectivamente:

E(Y/X = 800) = 600 Valor promedio de "Y" para "X" = 800

La diferencia entre el valor promedio obtenido y cada valor observado se debe al término de perturbación (μ_i).

$$Y_i = E(Y/X) + \mu_i$$

$Yi = \beta_1 + \beta_2 X + \mu_i$, reemplazando para c/u de los valores del consumo cuando el ingreso es S/.800, nos da las siguientes expresiones:

$Y_1 = 440 = \beta_1 + \beta_2 X + \mu_1$
$Y_1 = 440 = \beta_1 + \beta_2(650) + \mu_1$
$Y_2 = 480 = \beta_1 + \beta_2(650) + \mu_2$
$Y_3 = 520 = \beta_1 + \beta_2(650) + \mu_3$
$Y_4 = 560 = \beta_1 + \beta_2(650) + \mu_4$
$Y_5 = 600 = \beta_1 + \beta_2(650) + \mu_5$

Finalmente la regresión poblacional para un valor particular de la variable dependiente es:

$$FRP \rightarrow Y_i = \beta_1 + \beta_2 X_i + \mu_i$$

En el análisis de regresión interesa es estimar la FRP, es decir, estimar los valores de β_1 y β_2 no conocidos en base a las observaciones de Y y X.

5. FUNCIÓN DE REGRESIÓN MUESTRAL (FRM)

Es la que se obtiene a partir de una muestra de observaciones y nos permite estimar los parámetros de una función de la regresión poblacional, a partir de la información proporcionada por la muestra. Su forma es la siguiente:

$$FRM \rightarrow Y_i = \hat{\beta}_1 + \hat{\beta}_2 X_i + \hat{\mu}_i$$

La diferencia con la FRP está dada en que en este último caso los valores de los parámetros son de los datos poblacionales (β_i). Asimismo el término de perturbación (μ_i) está referido a la diferencia de los valores promedios poblacionales respecto a cada uno de los valores mencionados.

Podemos afirmar lo siguiente:
$\hat{\beta}_1$ es un estimador de β_1
$\hat{\beta}_2$ es un estimador de β_2
$\hat{\mu}_i$ es un estimador de μ_i

En conclusión, lo que se trata con los modelos de regresión es estimar la función de regresión poblacional (FRP) con base en la función de regresión muestral (FRM) en la forma mas precisa posible.

6. SIGNIFICADO DEL TERMINO DE PERTURBACION (μ_i)

Se tiene un modelo general, de la siguiente forma:

$$Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + ... + \mu_i$$

Donde los valores de los parámetros (β) son referidos a la población. Suponiendo que alguien nos diera los valores de los β's, entonces nos faltaría encontrar el valor del término de perturbación (μ_i).

El μ_i se simboliza como una bolsa donde están las otras variables respectivas del modelo y que no están incluidas en el mismo. Asimismo representa efectos aleatorios de la misma naturaleza de las μ_i .

Por ejemplo:
En el caso del consumo μ_i estaría representando el efecto de otras variables como la riqueza, el tamaño de la familia, etc.

Sea el modelo $Y = \beta_1 + \beta_2 X_2$, en el cual se ha estimado lo siguiente:
$\beta_1 = 10$; $\beta_2 = 2$ $\mu_i \sim N(0, 25)$

X_2	Valor *Teórico* (Y_i)	$\hat{\mu}_i$	Valor *Empírico* (Y_i)
2	14	-2	12
5	20	5	25
4	18	0	18
6	22	-3	19

Valor promedio

$\beta_1 + \beta_2 X_2 = E(Y/X)$ $E(Y/X) + \hat{\mu}_i = Y_i$

$10 + 2(2) = 14$ $14 - 2 = 12$

$10 + 2(5) = 20$ $20 + 5 = 25$

$10 + 2(4) = 18$ $18 + 0 = 18$

$10 + 2(6) = 22$ $22 - 3 = 19$

LECCIÓN 5

MODELO LINEAL GENERAL

1. MODELO LINEAL SIMPLE

El modelo lineal de dos variables es denominado también modelo lineal simple. Este caso bivariante donde la variable Y es explicada por la variable X, está representada por la siguiente expresión:

$$Y_i = \beta_1 + \beta_2 X_{i2} + \mu_i \qquad (1)$$
$$i = 1,2,\ldots,n$$

La expresión (1) muestra el modelo a partir de cada una de las observaciones. Sin embargo, el modelo se puede expresar de forma alternativa, en la que utilizando la notación matricial, se recogen todas las observaciones del modelo.

$$\underset{(n\times 1)}{Y} = \underset{(n\times 2)}{X}\underset{(2\times 1)}{\beta} + \underset{(n\times 1)}{\mu} \qquad (2)$$

$$\text{en donde } \underset{(2\times 1)}{\beta} = \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} \text{ y } \underset{(n\times 2)}{X} = \begin{bmatrix} 1 & X_1 \\ 1 & X_2 \\ \vdots & \vdots \\ 1 & X_n \end{bmatrix}$$

2. HIPÓTESIS DEL MODELO DE REGRESIÓN LINEAL

Para obtener los estimadores de los parámetros desconocidos del modelo, así como para realizar contrastes de hipótesis y la verificación del modelo, se necesitan un conjunto de hipótesis que se irán desarrollando en esta sección a medida que se vayan necesitando. Asimismo se hará referencia de ellas en el momento en que se utilicen.

El conjunto de hipótesis sobre las que se basa el modelo de regresión versa sobre los siguientes aspectos:

1) Forma funcional de la relación. (supondremos que es lineal)
2) Correcta especificación del modelo (es decir, que X es la única variable explicativa)
3) La variable X es no estocástica.
4) Identificabilidad de los parámetros. (β_1 y β_2 se podrán estimar de forma única)
5) Valor esperado de la perturbación dada la información observada. ($E(\mu) = 0_{(n\times 1)}$)
6) Varianzas y covarianzas de las perturbaciones dada la información observada. $E[\mu\mu'] = \sigma_\mu^2 I$

7) Distribución de probabilidad de la parte estocástica del modelo.

A continuación, se enumerarán y comentarán las hipótesis básicas del modelo lineal simple.

Hipótesis 1: El modelo es lineal tanto en las variables como en los parámetros.
Esto es, que las variables entran en el modelo de forma lineal ya sea en sus variables originales o después de alguna transformación previa. Los parámetros asociados a dichas variables también aparecen de forma lineal. Esta hipótesis es fundamental debido a que si el modelo no cumple con este supuesto habrá que utilizar técnicas no lineales que suponen un mayor grado de complicación.

Por ejemplo, el modelo $Y_i = \beta_1 + X_i\beta_2 + \mu_i$ es lineal en sus parámetros mientras que el modelo $Y_i = \beta_1 X_i^{\beta_2} + \mu_i$ no lo es.

Hipótesis 2: El modelo está correctamente especificado.
Esta hipótesis implica:

- Que se ha incluido la variable explicativa correcta.
- Que no se han omitido variables explicativas relevantes para explicar a la variable endógena.
- Que la relación es constante en todo el período muestral lo que implica que los coeficientes del modelo son constantes.

Hipótesis 3: Regresores no estocásticos.
Las observaciones de X_i son fijas durante todo el proceso de selección de muestra. De este modo, sólo se supone que el modelo de regresión y sus supuestos se aplican al conjunto particular de las X que se han observado. Así, la matriz X definida en (2) es de constantes conocidas.

Hipótesis 4: Identificabilidad de los parámetros.
Esta hipótesis se traduce en que los coeficientes β_1 y β_2 se podrán estimar de forma única a partir de unas observaciones dadas. Esto sucede cuando la variable X_i no sea constante, es decir, que presente variabilidad.

Si la variable explicativa fuese constante, el modelo presentaría dos términos constantes: el asociado al parámetro β_1 y el asociado al parámetro β_2 y ambos coeficientes medirían el mismo efecto.

Hipótesis 5: La esperanza de las perturbaciones condicionada a la información dada es nula.

Lo que significa que el valor esperado de las perturbaciones son cero, matricialmente se denota por:

$$E[\mu_i] = 0 \Rightarrow E[\mu_i] = \begin{bmatrix} E[\mu_1] \\ E[\mu_2] \\ \vdots \\ E[\mu_n] \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

Por lo tanto este supuesto conduce a que las observaciones de X no contengan información sobre el valor esperado de μ_i.

Ejemplo: En el ejercicio anterior de las 60 familias para un ingreso dado (650), existen 5 familias que tienen diferentes niveles de consumo, siendo su promedio igual a 650. El μ_i para cada uno de estás familias se obtiene a partir de la siguiente tabla:

Y_i	$E(Y/X_i)=650$	μ_i	$P(\mu_i)$	$\mu_i P(\mu_i)$
440	520	-80	1/5	-80(1/5)
480	520	-40	1/5	-40(1/5)
520	520	0	1/5	0(1/5)
560	520	40	1/5	40(1/5)
600	520	80	1/5	80(1/5)

E $(Y/X_i=650) = 520$

$$E(\mu_i / x_j) = \sum \mu_i p(\mu_i) = \left(\frac{1}{5}\right)\sum \mu_i = 0 \qquad \frac{1}{5}(-80-40+0+40+80) = 0$$

Por ser $P(\mu_i) = \frac{1}{5}$ constante.

En lo sucesivo se utilizará E $(u_i)=0$

Hipótesis 6: Las perturbaciones son esféricas.
Este supuesto se refiere a que las perturbaciones presentan varianza constante y están incorrelacionadas entre sí. Esta hipótesis encierra dos supuestos:

- Perturbaciones homocedásticas: $\text{Var}(\mu_j / X) = \sigma_s^2$ para $i = 1,2,...n$, es decir que las varianzas de las perturbaciones son iguales.
- Perturbaciones incorreladas o ausencia de autocorrelación: $\text{Cov}(\mu_i, \mu_j / X) = 0$, lo que significa que no existe relación alguna entre las perturbaciones consideradas.

Estos dos supuestos se pueden expresar conjuntamente para un modelo más general de la forma siguiente:

$$\sum_{\mu} = Var(\mu / X) = E[\mu\mu' / X] = \begin{bmatrix} \sigma^2 & 0 & ... & 0 \\ 0 & \sigma^2 & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \sigma^2 \end{bmatrix} = \sigma_{\mu}^2 I_n$$

donde I_n es una matriz identidad de orden n.

Este supuesto describe la información sobre las varianzas y covarianzas que es proporcionada por las variables independientes. Si μ satisface esta hipótesis, se dice que las perturbaciones son esféricas.

Hipótesis 7: Las perturbaciones recogidas en "μ" se distribuyen de forma normal ó Gaussiana

Esta hipótesis se establece por conveniencia, debido a que las derivaciones de los contrastes son mucho más sencillas. Además permite la estimación del modelo lineal y gaussiano por máxima verosimilitud. Sin embargo, se puede utilizar cualquier otra función de probabilidad sobre la distribución de las perturbaciones, cambiando algunos de los resultados que se verán posteriormente.

Analíticamente, este supuesto se puede expresar: $\mu \sim N(0, \sigma_\mu^2 I_n)$

3. MODELO LINEAL GENERAL

Este modelo establece una relación lineal entre un conjunto de k-1 variables explicativas (exógenas) y una variable a explicar (variable endógena).

Hipótesis:

Supongamos que existe una relación lineal entre una variable Y_i con k-1 variables explicativas X_2, X_3,..... X_k y un término de perturbación (μ), la cual podemos escribir como:

Ecuación tradicional

$$Y_i = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + ... + \beta_k X_{ik} + \mu_i \qquad (i = 1,2,....n)$$

Y: Es la variable endógena o explicada cuyo comportamiento se quiere analizar.

X: Cada una de las variables exógenas o explicativas y que son consideradas como las causas que crean transformaciones en la variable endógena.

β_1, β_2: Son los parámetros cuyo valor se desconoce y se va a estimar. A través de la estimación de los parámetros se obtiene una cuantificación de las relaciones existentes entre la variable endógena (Y) y cada una de las variables explicativas (X)..

μ_i: Perturbación aleatoria que recoge el efecto conjunto de otras variables no directamente explicitadas en el modelo, cuyo efecto individual sobre la endógena no resulta relevante.

i: Es el subíndice que hace referencia a las diversas observaciones para las cuales se establece su validez. Según el tipo de valores con los que esté trabajando, el subíndice hará referencia a distintos momentos del tiempo (series temporales: las cotizaciones en bolsa diarias, los índices de precio al consumo mensuales, los datos anuales del PIB de un país, etc.) o a distintas unidades económicas.

El parámetro que corresponde al término constante debe ser interpretado como el valor que toma la variable endógena cuando el resto de variables explicativas valen cero. Por ejemplo, en una función de consumo, aunque éste dependa de la renta y de otras variables, cuando todas ellas valen cero el individuo realiza un consumo para sobrevivir, lo que es conocido como "autoconsumo". Ese valor queda recogido en el modelo básico de regresión lineal a través del parámetro que corresponde al término constante.

El resto de parámetros que acompañan a las variables explicativas miden la relación entre estas y la variable endógena a través de su signo y su cuantía. El signo mide si la relación entre las variables es directa o inversa (si a medida que la variable explicativa se incrementa también lo hace la endógena o viceversa). La cuantía sirve para medir que variable explicativa, de todas las explicitadas en el modelo, es más importante para explicar el comportamiento de la endógena, de tal manera que si todas las variables están medidas en las mismas unidades de medida, la variable más importante será la que tenga un mayor valor de su parámetro.

Por tanto, el análisis de los parámetros estimados permite conocer la estructura económica del fenómeno que estamos analizando, entendiendo por estructura el patrón de comportamiento de acuerdo con el cual se desarrolla una acción. Por ejemplo, en el modelo que trata de explicar la evolución del consumo en función de la renta y de los tipos de interés, la estructura económica quedará definida como incrementos de consumo a medida que incrementa la renta; y reducciones de consumo cuando se incrementan los tipos de interés.

4. CÁLCULO MATRICIAL DEL MODELO LINEAL GENERAL

Para efectos del cálculo matricial tenemos los siguientes:

$$
\begin{aligned}
Y_1 &= \beta_1 + \beta_2 X_{12} + \beta_3\ X_{13} + ... + \beta_k X_{1k} + \mu_1 \\
Y_2 &= \beta_1 + \beta_2 X_{22} + \beta_3\ X_{23} + ... + \beta_k X_{2k} + \mu_2 \\
Y_3 &= \beta_1 + \beta_2 X_{32} + \beta_3\ X_{33} + ... + \beta_k X_{3k} + \mu_3 \\
&\ . \\
&\ . \\
Y_n &= \beta_1 + \beta_2 X_{n2} + \beta_3 X_{n3} + ... + \beta_k X_{nk} + \mu_n
\end{aligned}
$$

que puede escribirse matricialmente

$$
\begin{array}{l}
i = 1 \\ i = 2 \\ i = 3 \\ \ . \\ \ . \\ i = n
\end{array}
\quad
\begin{bmatrix} Y_1 \\ Y_2 \\ Y_3 \\ . \\ . \\ Y_n \end{bmatrix}
=
\begin{bmatrix} 1 & X_{12} & X_{13} & .X_{1k} \\ 1 & X_{22} & X_{23} & .X_{2k} \\ 1 & X_{32} & X_{33} & .X_{3k} \\ . & . & . & . \ . \\ . & . & . & . \ . \\ 1 & X_{n2} & X_{n3} & .X_{nk} \end{bmatrix}
*
\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \\ . \\ . \\ \beta_k \end{bmatrix}
+
\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ . \\ . \\ \mu_n \end{bmatrix}
$$

o simplemente: $Y = X * \beta + \mu$

Entonces, la forma extendida del Modelo Lineal General (MLG) puede compactarse utilizando al análisis matricial:

$$Y_{n1} = X_{nk}\beta_{k1} + \mu_{n1}$$

Para la estimación del MLG se asume lo siguiente:

a. El modelo es lineal en los parámetros.
b. Las variables explicativas, definidas como las columnas de la matriz X son determinísticas y linealmente independientes.
c. Los parámetros del modelo son constantes a lo largo de la muestra.
d. Existe una relación de causalidad desde las variables exógenas hacia la variable endógena y no viceversa.
e. El vector (μ) es un vector de variables aleatorias que cumplen con:

   $E(\mu) = 0$; $Var(u) = E(\mu\mu') = \sigma_{\mu}^2 I$

5. SUPUESTOS DEL MODELO LINEAL GENERAL

Complementándose con los supuestos mencionados para el modelo lineal simple, el modelo lineal general, además supone:

1) Las variables $X_2, X_3 \ldots X_K$ son variables no aleatorias.

2) La variable explicada Y_i es aleatoria con *media*:

$$E(Y_i) = \beta_1 + \beta_2 X_{i2} + \beta_3 X_{i3} + \ldots\ldots + \beta_k X_{ik}$$

O también: $E(Y)=XB$

Varianza: $E\left[(Y_i - E(Y_i))\right]^2 =$

$$E\left[(\beta_1 + \beta_2 X_{i2} + \ldots + \beta_{ik} X_k + \mu_i) - \beta_1 - \beta_2 X_{i2} - \ldots - \beta_k X_{ik}\right]^2 = E(\mu_i)^2 = \sigma_u^2$$

3) La variable Y_i (explicada) y $X_2, X_3 \ldots X_k$ (explicativas) no tienen errores de observación.

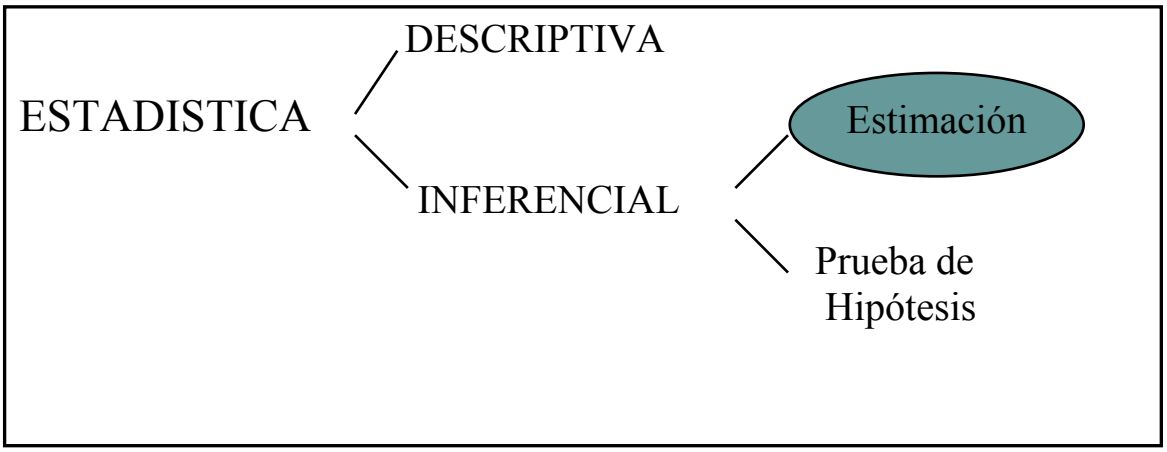
4) Entre las variables: $X_2, X_3 \ldots X_k$ no debe haber relación lineal (no están correlacionados entre ellas.) es decir, $Cov(X_i X_j)=0$ cuando $i \neq j$

5) Lo anterior significa que el rango de la matriz X debe ser k; por consiguiente ninguna columna debe ser linealmente dependiente de otra columna. ´

6) Para poder estimar el modelo se requiere tomar una muestra de n elementos, tal que $n>k$.

LECCIÓN 6

ESTIMACIÓN DE LOS PARÁMETROS



1. CRITERIOS PARA SELECCIONAR UN ESTIMADOR

- **Coherencia**: si al aumentar *n*, el estimador se aproxima al parámetro.
- **Eficiencia**: Cuando proporciona menor error estándar que otros estimadores.
- **Suficiente**: Si utiliza mayor cantidad de la información contenida en la muestra que otro estimador.
- **Insesgado (o imparcial)**: Si el estimador tiende a tomar valores por encima y por debajo del parámetro que estima, con la misma frecuencia.

2. MÉTODO DE ESTIMACIÓN DE LOS PARÁMETROS - MÍNIMOS CUADRADOS ORDINARIOS (MCO)

Es el método más usado, eficaz y conocido del análisis de regresión debido al contenido de las propiedades estadísticas que posee. El principio sobre el cual descansa esta metodología consiste en la minimización de la raíz cuadrada de la sumatoria de cada uno de los errores o perturbaciones.

Principio básico

El principio básico para estimar los parámetros es que la suma de los residuales de cada valor observado respecto al estimado sea lo más pequeña. Pero, $\sum \mu_i = \sum (Y_i - \hat{Y}_i) = 0$ porque la recta estimada corta a los residuales por encima y debajo de manera que se compensa. En consecuencia se debe de minimizar la suma de los cuadrados de cada uno de los residuales, obteniéndose los estimadores de los parámetros (β) que posean la menor varianza en comparación con cualquier otro método, sus valores en forma matricial serán:

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

Donde su varianza será :

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

Si se considera que se repite el proceso de muestreo, entonces las variables (X) permanecen fijas de muestra a muestra, pero cada muestra dará un conjunto diferente de μ, y por lo tanto un vector $\hat{\beta}$ diferente, en donde se expresa a $\hat{\beta}$ como una función lineal del verdadero β y de las perturbaciones μ.

Ejercicio Ilustrativo de Estimación de Parámetros en un Modelo Lineal Simple (MCO)

Se dispone de información de los ingresos totales y gastos en alimentación de 12 familias

Familia	**Gasto alimentación (nuevos soles)**	**Ingreso Total (nuevos soles)**
1	830	2100
2	510	1100
3	420	900
4	560	1600
5	1250	3200
6	840	2300
7	720	1800
8	490	700
9	690	1300
10	850	2400
11	550	1200
12	780	1700

Se planteará un modelo de regresión lineal y se especificará el papel que desempeña cada una de las variables en función al estudio.
Variable explicativa (X) es el ingreso familiar
Variable explicada (Y) es el gasto en alimentos de la familia

$$Y_i = \beta_1 + \beta_2 X$$

Familia	Y_i	X_i	$X_i Y_i$	X^2	$\hat{Y}_i$	$\mu_i = Y_i - \hat{Y}_i$
1	830	2,100	1,743,000	4,410,000	830.22	-0.22
2	510	1,100	561,000	1,210,000	529.69	-19.69
3	420	900	378,000	810,000	469.58	-49.58
4	560	1,600	896,000	2,560,000	679.95	-119.95
5	1,250	3,200	4,000,000	10,240,000	1160.80	89.20
6	840	2,300	1,932,000	5,290,000	890.32	-50.32
7	720	1,800	1,296,000	3,240,000	740.06	-20.06
8	490	700	343,000	490,000	409.48	80.52
9	690	1,300	897,000	1,690,000	589.79	100.21
10	850	2,400	2,040,000	5,760,000	920.37	-70.37
11	550	1,200	660,000	1,440,000	559.74	-9.74
12	780	1,700	1,326,000	2,890,000	710.00	70.00
Totales	8,490	20,300	16,072,000	40,030,000	8,490	-7.96E-13

Solución
Como los parámetros a estimar son β_1 y β_2 se establece las ecuaciones normales siguientes:

$$\sum Y = n\beta_1 + \beta_2 \sum X_i \quad (1)$$

$$\sum YX = \beta_1 \sum X_i + \beta_2 \sum X_i^2 \quad (2)$$

Y reemplazando, se tiene:

En (1) $8490 = 12\hat{\beta}_1 + 20300\hat{\beta}_2$

En (2) $16072000 = 20300\hat{\beta}_1 + 40030000\hat{\beta}_2$

Si se despeja de la primera ecuación el intercepto y se reemplaza dicho valor en le segunda se obtienen los siguientes estimadores:

$\hat{\beta}_1 = 199.108 \qquad \hat{\beta}_2 = 0.301$

La función de regresión muestral, es decir la regresión de Y con respecto a X:

$$\hat{Y}_i = 199.108 + 0.301X_i$$

Sustituyendo las observaciones muestrales de X en la ecuación anterior se obtiene la columna 6 de la tabla.

Comparando estos valores con aquellos observados para la variable dependiente hallamos los errores correspondientes a cada observación de la muestra. Se verifica que la suma de errores estimados es 0. (Columna μ_i)

Método Matricial

Familia	Y_i	X_i	X_iY_i	X^2	$\hat{Y}_i$	$\mu_i = Y_i - \hat{Y}_i$
1	830	2,100	1,743,000	4,410,000	830.22	-0.22
2	510	1,100	561,000	1,210,000	529.69	-19.69
3	420	900	378,000	810,000	469.58	-49.58
4	560	1,600	896,000	2,560,000	679.95	-119.95
5	1,250	3,200	4,000,000	10,240,000	1160.80	89.20
6	840	2,300	1,932,000	5,290,000	890.32	-50.32
7	720	1,800	1,296,000	3,240,000	740.06	-20.06
8	490	700	343,000	490,000	409.48	80.52
9	690	1,300	897,000	1,690,000	589.79	100.21
10	850	2,400	2,040,000	5,760,000	920.37	-70.37
11	550	1,200	660,000	1,440,000	559.74	-9.74
12	780	1,700	1,326,000	2,890,000	710.00	70.00
Totales	8,490	20,300	16,072,000	40,030,000	8,490	-7.96E-13

La ecuación matricial se escribe de la siguiente forma:

$$\begin{bmatrix} Y_1 \\ Y_2 \\ \\ . \\ . \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & . \\ 1 & X_{22} & . \\ & & . \\ . & . & . \; . \\ . & . & . \; . \\ 1 & X_{2k} & . \end{bmatrix} \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \\ . \\ . \\ \end{bmatrix} + \begin{bmatrix} \hat{\mu}_1 \\ \hat{\mu}_2 \\ \\ . \\ . \\ \hat{\mu}_n \end{bmatrix}$$

O simplemente: $Y = X\beta + \mu$

Para el caso de 2 variables: $(X'X)\hat{\beta} = (X'Y)$

$$(X'X) = \begin{pmatrix} n & \sum X_i \\ \sum X_i & \sum X_i^2 \end{pmatrix} \quad y \quad X'Y = \begin{pmatrix} \sum Y_i \\ \sum X_i Y_i \end{pmatrix}$$

$$(X'X) = \begin{bmatrix} 12 & 20300 \\ 20300 & 40030000 \end{bmatrix} \quad y \quad X'Y = \begin{pmatrix} 8490 \\ 16072000 \end{pmatrix}$$

$$(X'X)^{-1} = \begin{bmatrix} 0.586348323 & -0.000297349 \\ -0.000297349 & 1.75773E-07 \end{bmatrix}$$

$$\hat{\beta} = \begin{bmatrix} 0.586348323 & -0.000297349 \\ -0.000297349 & 1.75773E-07 \end{bmatrix} \begin{bmatrix} 8490 \\ 16072000 \end{bmatrix} = \begin{bmatrix} 199.10795 \\ 0.3005273 \end{bmatrix}$$

Los β son los mismos obtenidos que el método anterior.

Ejercicio Ilustrativo de Estimación de Parámetros en un Modelo Lineal General (MCO)

El director de una agencia de viajes quiere estudiar el sector turístico en Perú. Para ello dispone de información relativa al grado de ocupación hotelera (**Y**), número medio de turistas ($\mathbf{X_2}$), medido en miles de turistas, y estancia media ($\mathbf{X_3}$), medida en días.

OBSERVACIÓN I	**Nº DE OCUPACIÓN HOTELERA**	**TURISTAS (MILES)**	**DÍAS DE ESTANCIA**
	Y_i	X_2	X_3
1	5	2	3
2	8	3	4
3	8	5	6
4	9	4	5
5	9	6	7
6	13	2	6
7	6	3	4
8	9	4	5
9	4	5	4
10	3	6	3

Solución

En este caso se tienen 2 variables independientes, por lo que será conveniente hacer uso de la forma matricial, por lo tanto:

Modelo Lineal General: $Y_i = \beta_1 + \beta_2 X_2 + \beta_3 X_3 + \mu_i$**,** donde n =10; k=3

$$(X'X) = \begin{bmatrix} n & \sum X_{i2} & \sum X_{i3} \\ \sum X_{i2} & \sum X_{i2}^2 & \sum X_{i2}X_{i3} \\ \sum X_{i2} & \sum X_{i2}x_{i3} & \sum X_{i3}^2 \end{bmatrix} \qquad X'Y = \begin{bmatrix} \sum Y_1 \\ \sum X_{i2}Y_i \\ \sum X_{i3}Y_i \end{bmatrix}$$

Los coeficientes del modelo serán:

$$\hat{\beta} = \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{bmatrix} = (X'X)^{-1}X'Y = \begin{bmatrix} 2.5529 \\ -1.0821 \\ 1.9608 \end{bmatrix}$$

Luego, el modelo estimado es:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 = 2.5529 - 1.0821X_2 + 1.9608X_3$$

LECCIÓN 7

PROPIEDADES DE LOS ESTIMADORES

1. INSESGABILIDAD

Los β estimados son insesgados, es decir, si se obtuvieran los $\hat{\beta}$ de las muestras posibles, en promedio daría el verdadero valor del β poblacional.

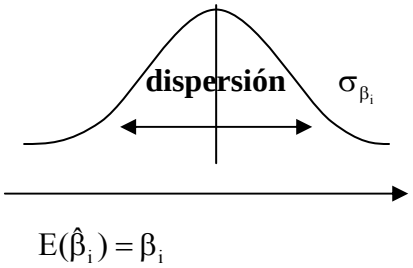
$$E(\hat{\beta}) = E\left[\beta + (X'X)^{-1}X'\mu\right] = \beta + (X'X)^{-1}X'E(\mu) = \beta$$

Entonces:

$E(\hat{\beta}_1) = \beta_1$
$E(\hat{\beta}_2) = \beta_2$
.
.
$E(\hat{\beta}_k) = \beta_k$

Es decir, que el valor esperado del estimador (y no el valor del estimador) coincide con el valor poblacional (desconocido) del parámetro.

Gráfico de Distribución de los Estimadores de β a partir de todas las Muestras Posibles



Cada uno de los $\hat{\beta}$ es igual a los verdaderos β más algo. Como los X's son valores fijos, entonces este algo va a depender de la perturbación (u).

En consecuencia los β estimados a partir de muestras diferirán entre ellos a partir de las diferencias en su término de perturbación (u). En efecto, diferentes muestras de valores produciran diferentes β (parámetros).

En general las diferencias entre los β estimados por cada muestra serán parecidas, por lo tanto cercano a los verdaderos β's, ello en la medida que el componente aleatorio (u) sea lo más reducido posible.

2. EFICIENCIA

Decir que un estimador es eficiente se refiere a que posee la menor varianza, es decir, un estimador es más eficiente que otro, cuando obtenidos de la misma muestra, la varianza del primero es menor que la del segundo.

$$\text{Var}(\beta_i) \leq \text{Var}(\beta_i^*)$$

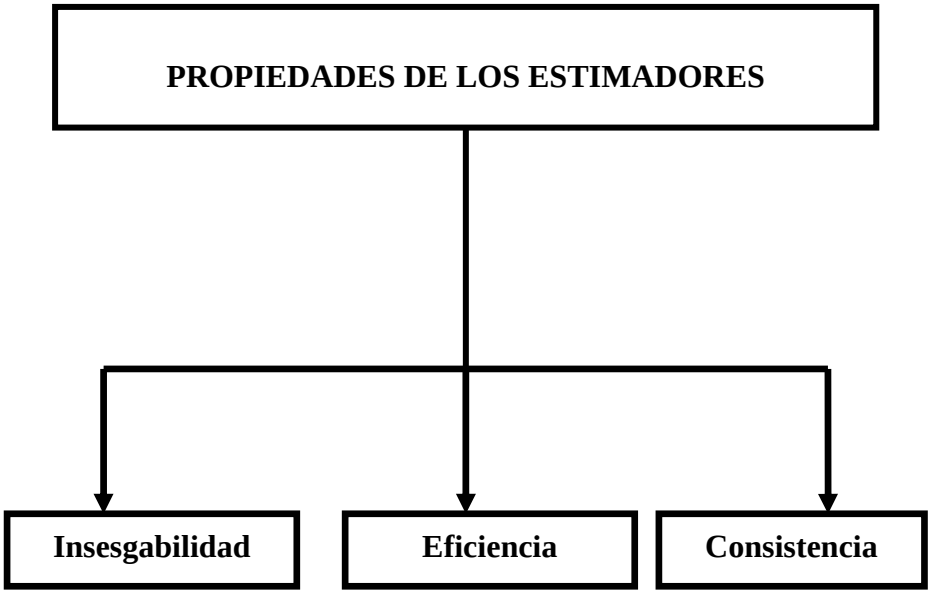
Donde:

β_i^* es obtenido por otro procedimiento

3. CONSISTENCIA

Un estimador $\hat{\beta}_i$ es consistente cuando a medida que aumente el tamaño de la muestra la media de la distribución de éste será más próxima al valor verdadero del parámetro β_i ,es decir, que $\hat{\beta}_i$ al cumplir con esta propiedad será la media β_i de tal distribución.

La importancia de esta propiedad radica en que si un estimador resulta sesgado cuando se posee una muestra de tamaño reducido (menor de 25), el investigador puede eliminar el sesgo se aumenta el tamaño de la muestra, por tanto, para poder garantizar que el estimador sea insesgado se deberá utilizar muestras grandes.


PROPIEDADES DE LOS ESTIMADORES
Insesgabilidad
Eficiencia
Consistencia

LECCIÓN 8

ESTIMACIÓN DE LA VARIANZA DEL TÉRMINO DE PERTURBACION

Un estimador del término de perturbación sería el residual (e). En consecuencia la varianza del residual podría utilizarse como estimador de la varianza del término de perturbación (μ).

Fórmulas

La fórmula usada para el Modelo Lineal Simple, que como se sabe tiene dos parámetros (β_1 β_2) es:

$$S^2 = \hat{\sigma}_\mu^2 = \frac{\sum (Y - \hat{Y})^2}{n-2} = \frac{\sum \mu_i^2}{n-2}$$

Entonces la fórmula usada para el Modelo Lineal General que tiene k parámetros (β_1 ; β_2 ; …; β_k) será:

Forma Matricial:

$$\sigma_\mu^2 = \frac{\mu'\mu}{n-k} = \frac{Y'Y - \hat{\beta}'X'Y}{n-k}$$

$$S^2 = \frac{\sum (Y - \hat{Y})^2}{n-k} = \frac{\sum \mu_i^2}{n-k}$$

Donde:

n = número de observaciones
k = número de parámetros

Ejercicio 1

El director de una empresa piensa que el nivel de ventas de un producto que él comercializa depende únicamente los gastos realizados en publicidad de este producto. Para estudiar las ventas de este producto pretende estimar el siguiente modelo:

$$Y_t = \beta_0 + \beta_1 X_t + \mu$$

Donde Y_t es la cantidad vendida anualmente del bien Y en el año t y X_t es gasto en publicidad durante el año t, por lo tanto

$$Y = X\beta + \mu$$

Se dispone de los siguientes datos muestrales:

Año	X_t (Publicidad)	Y_t (Ventas)
1994	80	70
1995	100	65
1996	120	90
1997	140	95
1998	160	110
1999	180	115
2000	200	120
2001	220	140
2002	240	155
2003	260	150

Para este caso se observa que sólo se tiene una variable independiente que es gasto en publicidad (X) , mientras que la dependiente son las ventas (Y)
Por lo que el modelo será:

$$\begin{bmatrix} 70 \\ 65 \\ 90 \\ . \\ . \\ . \\ 150 \end{bmatrix} = \begin{bmatrix} 1 & 80 \\ 1 & 100 \\ 1 & 120 \\ . & . \\ . & . \\ . & . \\ 1 & 260 \end{bmatrix} \begin{bmatrix} \beta_1 \\ \beta_2 \end{bmatrix} + \begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \\ . \\ . \\ . \\ \mu_{10} \end{bmatrix}$$

Para estimar a los 2 parámetros se hará uso de $\hat{\beta} = (X'X)^{-1}X'Y$, luego:

$$(X'X)^{-1} = \left[\begin{pmatrix} 1 & 1 & 1 & \ldots\ldots\ldots\ldots & 1 \\ 80 & 100 & 120 & \ldots\ldots\ldots & 260 \end{pmatrix} \begin{pmatrix} 1 & 80 \\ 1 & 100 \\ 1 & 120 \\ . & . \\ . & . \\ . & . \\ 1 & 260 \end{pmatrix} \right]^{-1} = \begin{pmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{pmatrix}^{-1} = \begin{pmatrix} 10 & 1700 \\ 1700 & 322000 \end{pmatrix}^{-1} =$$

$$\frac{1}{330000}\begin{pmatrix}322000 & -1700\\ -1700 & 10\end{pmatrix}=\begin{pmatrix}0.975757 & -0.005152\\ -0.005152 & 0.0000303\end{pmatrix}$$

$$X'Y=\begin{pmatrix}1.....1\\ 80....260\end{pmatrix}\begin{pmatrix}70\\ 65\\ .\\ .\\ .\\ 150\end{pmatrix}=\begin{pmatrix}1110\\ 205500\end{pmatrix}$$

Los coeficientes estimados del modelo serán:

$$\hat{\beta}=\begin{pmatrix}0.975757 & -0.005152\\ -0.005152 & 0.0000303\end{pmatrix}\begin{pmatrix}1110\\ 205500\end{pmatrix}=\begin{pmatrix}\hat{\beta}_1\\ \hat{\beta}_2\end{pmatrix}=\begin{pmatrix}24.4545\\ 0.50909\end{pmatrix}$$

La Varianza del Término de Perturbación

Se sabe que:

$$\boxed{\sigma_\mu^2=\frac{\mu'\mu}{(n-k)}=\frac{Y'Y-\beta'X'Y}{n-k}}$$

Haciendo las operaciones respectivas:

$$Y'Y=\begin{pmatrix}70 & 65.....150\end{pmatrix}\begin{pmatrix}70\\ 65\\ .\\ .\\ .\\ 150\end{pmatrix}=132100$$

$$\hat{\beta}'X'Y=\begin{pmatrix}24.4545 & 0.50909\end{pmatrix}\begin{pmatrix}1110\\ 205500\end{pmatrix}=131762.49$$

Reemplazando en la fórmula de la varianza de la perturbación, tenemos:

$$\sigma_\mu^2=\frac{132100-131762.49}{10-2}=\frac{337.51}{8}=42.18875$$

Calculo de la varianza estimada de los parámetros de regresión:

$$\text{Var}(\hat{\beta}) = (42.18875)\begin{pmatrix} 0.975757 & -0.005152 \\ -0.005152 & 0.0000303 \end{pmatrix}$$

$$\sigma^2_{\beta 1} = 42.18875(0.975757) = 41.16596813$$
$$\sigma^2_{\beta 2} = 42.18875(0.0000303) = 0.001278319125$$

La desviación estandar estimada será :

$$\hat{\sigma}_{\beta 1} = 6.416071082$$
$$\hat{\sigma}_{\beta 2} = 0.0357533$$

Ejercicio

En el ejercicio anterior del modelo lineal general: Grado de ocupación hotelera (**Y**) en función del número medio de turistas (**X_2**), medido en miles de turistas, y estancia media (**X_3**), medida en días:

$$\hat{Y}_i = \hat{\beta}_1 + \hat{\beta}_2 X_2 + \hat{\beta}_3 X_3 = 2.5529 - 1.0821 X_2 + 1.9608 X_3$$

La varianza del término de perturbación es:

$$\sigma^2_{\mu} == \frac{Y'Y - \beta' X'Y}{n - k} = 0.9914$$

ANEXO

OPERACIONES CON MATRICES

En este anexo se presentarán las nociones básicas del álgebra matricial, la cual es necesaria para poder entender los capítulos subsiguientes.
Dado los siguientes datos hipotéticos (Periodo 1991-1995)

AÑO	Y	X1	X2
1991	3	3	5
1992	1	1	4
1993	8	5	6
1994	3	2	4
1995	5	4	6

Se desea estimar el siguiente modelo de regresión lineal:

$Y_t = \beta_1 + \beta_2 X_{1t} + \beta_3 X_{2t} + \mu_t$
Donde:

- Y_t es la variable dependiente o endógena.
- X_1, X_2 son variables independientes o exógenas.
- β_1, β_2 y β_3 son parámetros desconocidos. A β_1 se le conoce con el nombre de intercepto, a los β_2 y β_3 se les llaman coeficientes de regresión.
- μ_t es una variable aleatoria no correlacionada y no observable.

A partir de los datos se crean las siguientes matrices:

$$Y = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} \qquad X = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix}$$

En este caso: n = 5 (numero de observaciones)
k = 3 (numero de parámetros del modelo)

Primeramente se tiene que tener en claro que una matriz es un arreglo de números o elementos en filas y en columnas. Cuando se habla del orden de una matriz se refiere a la cantidad de elementos ordenados en filas y columnas, por ejemplo las matrices X es una matriz de orden (3x5), mientras que la matriz Y es de (5x1).

Para estimar el modelo se hará uso de:

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

Por lo que para encontrar esos valores será necesario realizar ciertos cálculos matriciales previos tales como:

TRANSPUESTA DE UNA MATRIZ

La transpuesta de una matriz X de orden (5x3) la cual se denota por X', es una matriz de orden (3x5), la cual es obtenida a partir de cambiar las filas por las columnas, es decir que por ejemplo la primera fila de X se convierte la primera columna de X'.

Por lo tanto se tendrá que las transpuestas de X e Y serán:

$$X' = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 1 & 5 & 2 & 4 \\ 5 & 4 & 6 & 4 & 6 \end{bmatrix}$$

$$Y' = \begin{bmatrix} 3 & 1 & 8 & 3 & 5 \end{bmatrix}$$

MULTIPLICACIÓN DE MATRICES

Cada elemento de esta nueva matriz se obtiene sumando los valores que resultan de multiplicar los elementos de una fila de la matriz (por ejemplo de X') por su columna correspondiente de la otra matriz (por ejemplo Y), lo que originará que se forme una matriz de orden (3x1) la cual proviene de que la primera matiz tenga 3 filas y la segunda 1 columna. Por ejemplo:

$$X'Y = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 1 & 5 & 2 & 4 \\ 5 & 4 & 6 & 4 & 6 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 1\times3+1\times1+1\times8+1\times3+1\times5 \\ 3\times3+1\times1+5\times8+2\times3+4\times5 \\ 5\times3+4\times1+6\times8+4\times3+6\times5 \end{bmatrix} = \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix}$$

$$Y'Y = \begin{bmatrix} 3 & 1 & 8 & 3 & 5 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix} = \begin{bmatrix} 3\times3+1\times1+8\times8+3\times3+5\times5 \end{bmatrix} = 108$$

De manera similar se calcula (X'X)

$$X'X = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 \\ 3 & 1 & 5 & 2 & 4 \\ 5 & 4 & 6 & 4 & 6 \end{bmatrix} \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix} = \begin{bmatrix} 5 & 15 & 25 \\ 15 & 55 & 81 \\ 25 & 81 & 129 \end{bmatrix}$$

DETERMINANTE

El determinante es un valor que obtiene a partir de una matriz cuadrada (posee igual número de filas y columnas), el valor del determinante de una matriz es presentado por: la matriz encerrada por unas llaves: $|A|$.

Por simplicidad se mostrará a continuación como obtener una matriz de orden 2 y 3, para los otros casos es más conveniente hacer uso del computador ya que son operaciones que requieren de una considerable cantidad de operaciones.

Hallar un determinante de una matriz de orden 2:

Sea la matriz $A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$

su determinante estará dado por: $|A| = a \times d - b \times c$

Hallar un determinante de una matriz de orden 3:

Sea la matriz $A = \begin{bmatrix} a & b & c \\ d & e & f \\ m & n & p \end{bmatrix}$

su determinante estará dado por:

$$|A| = a \times e \times p - a \times f \times n + b \times f \times m - b \times d \times p + c \times e \times n - c \times e \times m$$

INVERSA[2]

Se debe tener presente que la inversa de una matriz origina otra matriz la cual se podrá calcular solamente cuando tenga la misma cantidad de filas y columnas, además su determinante debe ser diferente de cero.

Los pasos para obtener la inversa de un matriz son:
Hallar el determinante de la matriz, si es diferente de cero será posible calcularlo.
Si se cumple con el punto anterior el paso siguiente consiste en reemplazar cada elemento de la matriz por su cofactor correspondiente, obteniéndose así la matriz de cofactores.

El cofactor de una matriz A de orden (nxn) es denotado por c_{ij}, el cual esta definido por:

$$c_{ij} = (-1)^{i+j} |M_{ij}|$$

[2] La inversa de una matriz puede ser halla por medio de calculadoras matriciales, esto resulta útil para el ahorro de tiempo en los cálculos.

donde $\left|M_{ij}\right|$ es el determinante que resulta de eliminar la iésima fila y la jésima columna de la matriz considerada inicialmente

Luego de obtener la matriz de cofactores se halla su transpuesta, la cual es conocida como la matriz adjunta (Adj).

Como último paso se procede a calcular la inversa de la forma siguiente:

$$A^{-1}\frac{1}{|A|}\times \text{Adj}(A)$$

Su determinante se obtiene por la fórmula mostrada anteriormente: $|X'X| = 20$

El paso que sigue es la obtención de la matriz de cofactores, a al cual la llamaremos C:

$$C=\begin{bmatrix} \begin{vmatrix}55 & 81\\ 81 & 129\end{vmatrix} & -\begin{vmatrix}-7 & 11\\ -9 & -3\end{vmatrix} & \begin{vmatrix}15 & 55\\ 25 & 81\end{vmatrix} \\ -\begin{vmatrix}15 & 25\\ 81 & 129\end{vmatrix} & \begin{vmatrix}17 & -13\\ -9 & -3\end{vmatrix} & -\begin{vmatrix}17 & -3\\ -9 & 3\end{vmatrix} \\ \begin{vmatrix}-3 & -13\\ -3 & 11\end{vmatrix} & -\begin{vmatrix}17 & -13\\ -7 & 11\end{vmatrix} & \begin{vmatrix}17 & -3\\ -7 & -3\end{vmatrix} \end{bmatrix} = \begin{vmatrix}534 & -120 & -160\\ -90 & -168 & -24\\ -72 & 168 & -72\end{vmatrix}$$

$$\text{Adj}=\begin{vmatrix}534 & -120 & -160\\ -90 & -168 & -24\\ -72 & 168 & -72\end{vmatrix}$$

Luego:

$$(X'X)^{-1}=\begin{bmatrix}5 & 15 & 25\\ 15 & 55 & 81\\ 25 & 81 & 129\end{bmatrix}^{-1}=\frac{1}{20}\times\text{Adj}=\begin{bmatrix}26.7 & 4.5 & -8\\ 4.5 & 1 & -1.5\\ -8 & -1.5 & 2.5\end{bmatrix}$$

Si se utiliza la fórmula anteriormente dada se obtendrán los estimadores de parámetro $\hat{\beta}$

$$\hat{\beta} = \begin{bmatrix} 26.7 & 4.5 & -8 \\ 4.5 & 1 & -1.5 \\ -8 & -1.5 & 2.5 \end{bmatrix} \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix} = \begin{bmatrix} 4 \\ 2.5 \\ -1.5 \end{bmatrix}$$

Valor estimando de la varianza de los términos de perturbación: $\hat{\sigma}_{\mu}^{2}$

En el modelo de regresión lineal se obtiene a partir de:

$$\hat{\sigma}_{\mu}^{2} = (Y'Y - \hat{\beta}'X'Y)/(n-k)$$

$$\hat{\beta}' = \begin{bmatrix} 4 & 2.5 & -1.5 \end{bmatrix}$$

$$\hat{\beta}'X'Y = \begin{bmatrix} 4 & 2.5 & -1.5 \end{bmatrix} \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix} = 106.5$$

$$\hat{\sigma}_{\mu}^{2} = \frac{(Y'Y - \hat{\beta}'X'Y)}{n-k} = \frac{108 - 106.5}{5-3} = 0.75$$

Estimación de la matriz de varianzas y covarianzas:

$$\text{var}(\hat{\beta}) = \hat{\sigma}_{\mu}^{2}(X'X)^{-1}$$

$$\text{var}(\hat{\beta}) = 0.75 \begin{bmatrix} 26.7 & 4.5 & -8 \\ 4.5 & 1 & -1.5 \\ -8 & -1.5 & 2.5 \end{bmatrix} = \begin{bmatrix} 20.025 & 3.375 & -6 \\ 3.375 & 0.75 & -1.125 \\ -6 & -1.125 & 1.875 \end{bmatrix}$$

Estimando el R^2

$$R^2 = (\hat{\beta}'X'Y - n\overline{Y}^2)/(Y'Y - n\overline{Y}^2)$$

$$\overline{Y} = \frac{3+1+8+3+5}{5} = 4$$

$$\overline{Y}^2 = 16$$

$$\hat{\beta}'X'Y = \begin{bmatrix} 4 & 2.5 & -1.5 \end{bmatrix} \begin{bmatrix} 20 \\ 76 \\ 109 \end{bmatrix} = 106.5 \qquad R^2 = \frac{106.5 - 5\times 16}{108 - 5\times 16} = 0.9464$$

Estimando el R^2 ajustado

$$R^2_{ajustado} = \frac{(n-1)}{(n-k)}\left(\frac{1-(Y'Y-\hat{\beta}'X'Y)}{(Y'Y-n\overline{Y}^2)}\right) = \frac{(n-1)}{(n-k)}R^2$$

$$R^2_{ajustado} = \left(\frac{5-1}{5-3}\right)\left(\frac{106.5-5\times 16}{108-5\times 16}\right) = \left(\frac{4}{3}\right)\times 0.9464 = 0.8929$$

Para el cálculo del F estadístico, utilizaremos la siguiente formula:

$$F_c = \frac{R^2/(k-1)}{(1-R^2)/(n-k)}$$

Reemplazando en la formula obtuvimos el siguiente valor:

$$F_c = \frac{0.9464 \ /(3-1)}{(1-0.9464)/(5-3)} = 17.6667$$

PREDICCION[3]

Predicción en Media: $\hat{Y}_i = X'_i\hat{\beta}$

Predicción de un valor puntual: $\hat{Y}_i/(x_0) = X'_0\hat{\beta}$

[3] Este tema se vera detalladamente en la Unidad IV

Sea $X_0 = \begin{bmatrix} 1 \\ 6 \\ 8 \end{bmatrix}$ Entonces $X'_0 = \begin{bmatrix} 1 & 6 & 8 \end{bmatrix}$ $\qquad \hat{\beta}_0 = \begin{bmatrix} 4 \\ 2.5 \\ -1.5 \end{bmatrix}$

Finalmente: $\hat{Y}_i /(x_0) = X'_0 \hat{\beta} = 7$

VARIANZA DE LA PREDICCION

$$(X'X)^{-1} = \begin{bmatrix} 26.7 & 4.5 & -8 \\ 4.5 & 1 & -1.5 \\ -8 & -1.5 & 2.5 \end{bmatrix} \qquad \hat{\sigma}^2_{\mu} = 0.75$$

$$X_0'(X'X)^{-1} = \begin{bmatrix} -10.3 & -1.5 & 3 \end{bmatrix}$$

$$Xo'(X'X)^{-1}Xo = 4.7$$

Remplazando en las fórmulas para obtener las varianzas, tenemos:

Varianza de la predicción promedio

$$var(Y / xo) = \hat{\sigma}^2_{\mu}\, x'_o\, (X'X)^{-1} x_o$$

$$var(Y / xo) = 0.75(4.7) = 3.525$$

DS: Desviación Estándar: $DS = \sqrt{\hat{\sigma}^2_{\mu}\, x'_o\, (X'X)^{-1} x_o}$ $\qquad$ Entonces $DS = \sqrt{3.525} = 1.8755$

El valor promedio de Y se encuentra en el intervalo comprendido entre:

$$Yo - t_{\alpha/2}\sqrt{\sigma^2_{\mu} X'o(X'X)^{-1}Xo} \le E(Y / Xo) \le Y_o + t_{\alpha/2}\sqrt{\sigma^2_{\mu} X'o(X'X)^{-1}Xo}$$

$Y_0 = 7$ Donde: $t_{\alpha/2} = 3.182$ Con 3 grados de libertad y un nivel de significancia del 5%

Reemplazando los datos, tenemos que el valor promedio de Y se encuentra comprendido en el intervalo: $[1.0258, 12.974]$

VARIANZA DE LA PREDICCION INDIVIDUAL:

$$var(Y / xo) = \hat{\sigma}^2_{\mu}\left[1 + x'_o\, (X'X)^{-1} x_o\right]$$

$$var(Y / xo) = 0.75[1 + 4.7] = 4.275$$

$$DS = \sqrt{\hat{\sigma}_{\mu}^{2}\left[1 + x'_{o}(X'X)^{-1}x_{o}\right]} \quad \text{Entonces } DS = \sqrt{4.275} = 2.0676$$

El intervalo de confianza al 95% para la predicción puntual se calcula mediante la siguiente fórmula:

$$Yo - t_{\alpha/2}\sqrt{\hat{\sigma}_{\mu}^{2}\left[1 + x'_{o}(X'X)^{-1}x_{o}\right]} \le E(Y/Xo) \le Yo - t_{\alpha/2}\sqrt{\hat{\sigma}_{\mu}^{2}\left[1 + x'_{o}(X'X)^{-1}x_{o}\right]}$$

$Y_0 = 7$

Donde: $t_{\alpha/2} = 3.182$

Entonces el intervalo de confianza para la predicción individual es: $[0.42087, 13.579]$

EVALUACION DEL MODELO ESTIMADO (PARA PREDECIR)

$$X = \begin{bmatrix} 1 & 3 & 5 \\ 1 & 1 & 4 \\ 1 & 5 & 6 \\ 1 & 2 & 4 \\ 1 & 4 & 6 \end{bmatrix} \qquad \hat{\beta} = \begin{bmatrix} 4 \\ 2.5 \\ -1.5 \end{bmatrix} \qquad \hat{Y}_t = \begin{bmatrix} 4 \\ 0.5 \\ 7.5 \\ 3 \\ 5 \end{bmatrix} \qquad Y_t = \begin{bmatrix} 3 \\ 1 \\ 8 \\ 3 \\ 5 \end{bmatrix}$$

$$\hat{Y}_t - Y_t = \begin{bmatrix} 1 \\ -0.5 \\ -0.5 \\ 0 \\ 0 \end{bmatrix} \qquad (\hat{Y}_t - Y_t)^2 = \begin{bmatrix} 1 \\ 0.25 \\ 0.25 \\ 0 \\ 0 \end{bmatrix} \qquad n = 5$$

$$\sum_{t=1}^{n}(\hat{Y}_t - Y_t)^2 = (1 + 0.25 + 0.25 + 0 + 0) = 1.5$$

Raíz Cuadrática Media (rms):

$$rms = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{Y}_t - Y_t\right)^2}$$

Donde: $\hat{Y}_t$ es el valor estimado de Y_t

Y_t es el valor observado de Y_t

Reemplazando los datos: $\text{rms} = \sqrt{\frac{1}{5}(1.5)} = 0.5477$

La rms, debe ser lo más pequeño posible para que el modelo sea bueno para predecir.

Coeficiente de Theil (U):

$$U = \frac{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{Y}_t - Y_t\right)^2}}{\sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{Y}_t\right)^2} + \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(Y_t\right)^2}}$$

Donde: $\hat{Y}_t$ Es el valor estimado de Y_t

Y_t Es el valor observado de Y_t

$$(\hat{Y}_t)^2 = \begin{bmatrix} 16 \\ 0.25 \\ 56.25 \\ 9 \\ 25 \end{bmatrix} \qquad (Y_t)^2 = \begin{bmatrix} 9 \\ 1 \\ 64 \\ 9 \\ 25 \end{bmatrix} \qquad \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(\hat{Y}_t - Y_t\right)^2} = \sqrt{\frac{1}{5}(1.5)} = 0.5477$$

$n = 5$

$\sum(\hat{Y}_t)^2 = (16 + 0.25 + 56.25 + 9 + 25) = 106.5$

$\sum(Y_t)^2 = (9 + 1 + 64 + 9 + 25) = 108$

$\sqrt{\frac{1}{n}\sum(\hat{Y}_t)^2} = \sqrt{\frac{1}{5}(106.5)} = 4.6152$ $\qquad$ $\sqrt{\frac{1}{n}\sum(\hat{Y}_t)^2} = \sqrt{\frac{1}{5}(108)} = 4.6476$

Reemplazando datos se tiene, que U es igual a:

$$U = \frac{0.5477}{4.6152 + 4.6476} = 0.0591$$

El índice de Theil nos dice que cuanto más cercano a cero, el modelo será bueno para predecir. Este coeficiente mide la rms en términos relativos.

Ejercicio de autoconocimiento

¿Porqué hacer un análisis de regresión lineal?

	SI	NO	NO SÉ
1. Porque considero que es una técnica estadística importante para una buena toma de decisiones empresariales.			
2. Porque permite recomendar un tratamiento para los problemas en el comportamiento de los agentes.			
3. Para analizar el pasado y predecir el futuro de la empresa.			
4. Especifica la relación entre variables.			
5. Para utilizar el modelo correcto y adecuado para un pronóstico			
6. Para establecer la importancia del estudio de las variables.			
7. Porque realiza la distinción entre variable dependiente y la independiente.			
8. Para interpretar los elementos constitutivos del modelo de predicción.			
9. Para desarrollar la posibilidad de utilizar el análisis de regresión para estimar intervalos y contrastar hipótesis.			
10. Para predecir sucesos futuros.			

CALIFICACION

Puntuar con un punto cada respuesta "SI".
Si obtienes de de 1 - 3 puntos tienes pocas expectativas de hacer un buen análisis de regresión lineal.
Si tienes entre 4 – 7, tienes buenas expectativas de hacer un buen análisis de regresión lineal.
Y si tienes entre 8 – 10, denotas excelentes expectativas de hacer un buen análisis de regresión lineal.

RESUMEN

Los elementos que integran un modelo son: las ecuaciones, las variables y los parámetros.

El proceso de construcción de un modelo se puede presentar como una secuencia de etapas que a continuación vamos a presentar:

-Conocimiento de la Teoría Económica
-Especificación del Modelo Econométrico
-Estimación
-Verificación
-Predicción

La idea clave del análisis de regresión es la dependencia estadística de una variable, la variable dependiente, sobre una o más variables, las variables explicativas.

El objetivo de este análisis es estimar y/o predecir la media o el valor promedio de la variable dependiente con base en los valores conocidos o determinados de las variables explicativas.

Resumen de fórmulas

- La función de regresión lineal: $Y_i = \beta_1 + \beta_2 X_2 + \mu_i$
- Intercepto: β_1
- Coeficiente de la pendiente: β_2
- Perturbación estructural o estocástica de la población: $\mu_i = Y_1 - E(Y/X)$
- Función de regresión poblacional: $FRP \rightarrow Y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + \mu_i$
- Función de regresión muestral: $FRM \rightarrow Y_i = \hat{\beta}_1 + \hat{\beta}_2 x_i + e_i$
- La forma extendida del Modelo Lineal General (MLG) puede compactarse así: $Y_{n1} = X_{nk}\beta_{k1} + \mu_{n1}$
- Métodos de estimación de los parámetros: Mínimos Cuadrados ordinarios (MCO), Método de momentos y Máxima verosimilitud
- Modelo Lineal Simple**:** $y_i = \beta_1 + \beta_2 x_{i2} + \mu_I$
- Cálculo de estimadores:

$$b_2 = \frac{\sum (x_2 - \bar{x})(y_i - \bar{y})}{\sum (x_2 - \bar{x})}$$

$b_1 = \bar{y} + b_2\bar{x}_2$

- Modelo Lineal General**:** $y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_3 + \mu_1$

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

EXPLORACION ON LINE

1. Análisis de regresión lineal

   http://www.udc.es/dep/mate/estadística 2

2. Regresión lineal entre dos variables.

   http://bayes.escet.urjc.es/~jmmarin/libroelec,tema8.

3. Fundamentos del análisis de regresión lineal

   http://www.termodinamica.us.es/tecnicas/como/node

4. Introducción al análisis de regresión lineal.

   http://www.cuspide.com/detalle_libro.php?isbn=9702403278

5. Optimización y análisis de regresión lineal.

   http: //members.lycos.co.uk/k59/artha/artha1.

LECTURA

IMPORTANCIA DE LA REGRESIÓN.

Uno de los usos más frecuentes de la regresión no es el de pronosticar, como en este ejemplo, sino que usamos la regresión para el propósito de hablar de una relación más o menos general entre las variables involucradas. Regresando al ejemplo, el argumento sería el siguiente:

si esta compañía de bolígrafos encontró una relación entre las ventas y estas variables independientes, ese mismo tipo de relación se debe presentar en otra compañía comercializadora o productora de bolígrafos

Una extensión de resultados como ésta se sale ya del ámbito de la estadística y se adentra más en el del sentido común.

En parecidas circunstancias, se encontraría una afirmación que extendiera los resultados de un análisis de regresión en el tiempo. Diciendo algo como

si en los últimos siete años ha habido una relación entre las ventas y estas variables independientes, ese mismo tipo de relación será cierta en el próximo año

A veces llegamos, incluso, al extremo de decir

si en esta compañía del ramo mercantil se ha presentado esta relación, en otras compañías del mismo ramo se presentará también; nuestra compañía debe poner más cuidado al elaborar su política de ventas en los factores siguientes:

- *número de agentes de ventas*
- *número de spots en la televisión local y*
- *eficiencia de los mayoristas*

Claro que esta afirmación se halla mucho más retirada de la frialdad de los números y de los mínimos cuadrados.

Para poder calificar la validez de afirmaciones como las anteriores debemos fijarnos en la cuestión de los *sesgos*.

En este caso, habría que ver qué tanto se parecen nuestros datos a una muestra al azar de observaciones de la situación mayor. Regresando a nuestro ejemplo habría que considerar qué tan parecidas son las circunstancias de mercado de las compañías a las que queremos extender los resultados con las de la compañía de donde se sacaron los datos analizados; habría que hacer el mismo tipo de consideraciones en caso de querer extender los resultados en el tiempo.

Fuente: Mendoza Durán, 2003

ACTIVIDADES

1. Dado el modelo**:**

(1) $PBI_t = I_t + CP_t + CG_t + BC_t + VE_t$

(2) $I_t = \alpha_1 + \alpha_2 DEI_t + \alpha_3 YT_t + \alpha_4 LSF_t + \alpha_5 IDE_t + \alpha_6 DEI_{t-1} + \alpha_7 YT_{t-1} + \alpha_8 LSF_{t-1} + \alpha_9 IDE_{t-1} + \alpha_{10} I_{t-1} + \mu_{2t}$

(3) $LSF_t = \beta_1 + \beta_2 PBI_t + \beta_3 TIPR_t + \beta_4 PBI_{t-1} + \beta_5 TIPR_{t-1} + \beta_6 LSF_{t-1} + \mu_{3t}$

(4) $IDE_t = \gamma_1 + \gamma_2 PBI_t + \gamma_3 REM_t + \gamma_4 PBI_{t-1} + \gamma_5 REM_{t-1} + \gamma_6 IDE_{t-1} + \mu_{4t}$

Donde:

PBI	:	Producto Bruto Interno
I	:	Inversión Total
LSF	:	Liquidez Total del Sistema Financiero
IDE	:	Inversión Directa Extranjera
CP	:	Consumo Privado Total
CG	:	Consumo del Gobierno Total
BC	:	Balanza Comercial
VE	:	Variación de Existencias
DEI	:	Deuda Externa para la Inversión
YT	:	Ingresos Tributarios
TIPR	:	Tasa de Interés Pasiva Real del Sistema Financiero
REM	:	Remesas de utilidades al exterior

a. Clasificar las variables.
b. Interpretar los parámetros.

2. Una agente desea invertir 1 millón de soles en acciones que se coticen en Bolsa. Después de evaluar las distintas alternativas se plantea la decisión de invertir entre dos opciones: acciones de la empresa A o acciones de la empresa B. En principio, su criterio de elección se basa en preferir la compra de acciones de aquella empresa en la que espere obtener un rendimiento por sol invertido más elevado y a la vez que presente mayor seguridad. Para ayudarse en la toma de decisión plantea un modelo econométrico donde establece que la rentabilidad por cada 1000 soles invertidos en acciones de cada empresa (REN) depende de dos variables:

- Volumen de beneficios reales obtenidos por la empresa en millones de soles (BEN)
- Volumen de activos medio mantenido en millones de soles (ACT).

Dado que se evalúan dos empresas, se plantea dos modelos independientes. Para ello dispone de la siguiente información obtenida con los datos de los últimos 20 años:

Empresa "A"

$$X'X = \begin{bmatrix} 20 & 10 & 25 \\ & 35 & 30 \\ & & 50 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 0.17 & 0.05 & -0.11 \\ & 0.07 & -0.07 \\ & & 0.12 \end{bmatrix} \qquad X'Y = \begin{bmatrix} 40 \\ 105 \\ 115 \end{bmatrix}$$

Empresa "B"

$$X'X = \begin{bmatrix} 20 & 8 & 20 \\ & 22 & 25 \\ & & 40 \end{bmatrix} \qquad (X'X)^{-1} = \begin{bmatrix} 0.15 & 0.10 & -0.14 \\ & 0.23 & -0.20 \\ & & 0.22 \end{bmatrix} \qquad X'Y = \begin{bmatrix} 30 \\ 90 \\ 110 \end{bmatrix}$$

Teniendo en cuenta la información proporcionada, se pide:

a) Efectuar la especificación de cada uno de los modelos
b) Calcular los estimadores mínimos cuadráticos de los parámetros.
c) Estimar las varianzas de los términos de perturbación
d) Si el agente conoce los siguientes datos en el periodo n+1 :

	Empresa "A"	**Empresa "B"**
BEN	3.2	3.5
ACT	3.8	3.5

donde se estima que los valores de las variables explicativas son similares a sus medias. ¿En cuál de las dos empresas decidirá invertir teniendo en cuenta sus criterios de inversión?

3. Sea Y: demanda de trabajo y X: tasa de desempleo.
Con los siguientes datos:

ΣX_i^2 =432.970 ΣY_i^2 = 262467.06
ΣX_i = 82.7000 ΣY_i = 1995.2000
X = 5.16875 Y = 1224.7000
$\Sigma X_i Y_i$ = 10042.3

Estime los parámetros del modelo, presente la función de regresión muestral ¿Qué comentarios le merece?

4. Un comerciante al menudeo llevó a cabo un estudio para determinar la relación entre los gastos de publicidad semanal y las ventas. Se obtuvieron los siguientes datos:

X	: Costos de publicidad ($)	Y	:	Ventas ($)
ΣX_i^2	= 15650	ΣY_i^2	=	2512925
ΣX_i	= 410	ΣY_i	=	5445
$\Sigma X_i Y_i$	= 191325	n	=	12

Estime los parámetros del modelo, presente la función de regresión e interprete.

5. La Empresa 'The Home" produce y comercializa muebles para el hogar. Esta empresa tiene cierto poder en el mercado, en el sentido que puede manejar el precio de sus productos, además el constante gasto en publicidad diferencia sus productos de los de la competencia.

Sin embargo, últimamente la participación de la empresa en el mercado de muebles para el hogar ha disminuido. El Gerente General atribuye esto al hecho que no ha existido una política clara en cuanto la fijación de precios y publicidad. Por tanto, se pide un estudio que determine el efecto que tiene en las ventas las variables precio y publicidad.
Además, se pide proponer una política en precio y publicidad si se desea aumentar las ventas para el próximo periodo en 10%. Los datos son los siguientes:

VENTAS (miles de soles)	**PRECIO soles/u**	**PUBLICIDAD (miles de soles)**
180.6	2.1	30
213.3	4.5	55
174.6	2.9	25
189.3	3.6	36
209.1	15	60
248.1	7.7	82
253.9	5.8	73
215.8	3.2	58
218.1	5	58
206.6	12.3	49

AUTOEVALUACIÓN

Encierra en un círculo la letra que contenga la alternativa correcta.

1. ¿Cuál de las siguientes etapas no pertenecen al proceso de construcción de un modelo:

   a) Conocimiento de la Teoría Económica
   b) Especificación del Modelo Econométrico
   c) Estimación
   d) Verificación
   e) Propiedad de los estimadores

2. La verdadera expresión de la ecuación lineal con dos variables independientes es:

   a) $\hat{Y} = b_1 + b_2X_2 + b_3X_3$
   b) $\hat{Y} = f(X_1, X_2, X_3, ...)$
   c) $\mu_i = Y_1 - E(Y/X)$
   d) $T_t = \alpha + \beta Y_t + u_t \;;\; o < \beta < 1$
   e) N.A.

3. La ecuación tradicional del Modelo Lineal General es

   a) $\mu_i = Y_1 - E(Y/X)$
   b) $\hat{Y} = f(X_1, X_2, X_3, ...)$
   c) $Y_i = \beta_1 + \beta_2 x_{i2} + \beta_3 x_{i3} + ... + \beta_k x_{ik} + \mu_i$
   d) Dependiente
   e) N.A.

4. El grado de dureza de un mineral depende del contenido de otros dos minerales. Los siguientes datos corresponden a dichas barras de minerales estudio.

Dureza	6	13	13	29	33	32	46	117.5
c_1	1	2	3	4	5	6	8	20
c_2	10	10	12	11	14	15	18	30

Considerar el modelo $d_i = \beta_1 + \beta_2 c_{2i} + \beta_3 c_{2i}$
con los supuestos habituales:

4.1 Prueba la hipótesis: $H_{0:}\ \beta_1 = 0$
con $\alpha = 5\%$, $H_{1:}\ \beta_1 \neq 0$

   a. Se acepta la H_0 con un t calculado igual a 0.68.
   b. Se rechaza la H_0 con un t calculado igual a 5.66.
   c. Se acepta la H_0 con un t calculado igual a 1.98

4.2 Calcule un intervalo de confianza del 99 % para el valor esperado de la dureza de una barra tal que $c_{1=}5$ $c_{2=}10$

$(X'X)^{-1} =$

15.258919	1.9536424	-1.806665
1.9536424	0.2649007	-0.238411
-1.806665	-0.238411	0.2177953

$$X'Y = \begin{pmatrix} 289.5 \\ 3262 \\ 5960 \end{pmatrix} \qquad Y'Y = 19250.25 \qquad \hat{\beta} = (22.51, 8.76, -2.66)$$

$S^2 = 8.94$

a. [17.61,61.71]
b. [20.35,56.34]
c. [24.45,70.94]

4.3 Calcule un intervalo de confianza de 95 % para $\sigma 2$

a. [6.76,34.45]
b. [7.35,46.49]
c. [3.48,53.79]

RESPUESTAS DE CONTROL

1e, 2.a, 3.c, 4.1c, 4.2a, 4.3c