Fundamentos de Econometría

Juan León Jara Almonte & Marcelo Manuel Gallardo Burga

Prefacio

A lo largo del presente documento, se dan alcances de los fundamentos matemáticos y propiedades estadísticas relevantes para la estimación de los parámetros del modelo de regresión multivariado. Para ello, se empieza presentando los conceptos básicos del álgebra matricial y los fundamentos de optimización estática. Enseguida, se presenta el modelo de regresión multivariado y sus supuestos subyacentes, los cuales son la base para la estimación robusta de relaciones entre variables, así como para la identificación de efectos causales entre variables.

Se hace un especial énfasis en los factores que pueden llevar a una mala estimación de los parámetros y sus errores estándar. Motivo por el cual, se cuentan con capítulos que abordan los temas de cambio estructural, multicolinealidad, heterocedasticidad y autocorrelación serial. En dichos capítulos no solo se exploran los conceptos teóricos sino que también se presentan las pruebas estadísticas para detectar dichos problemas y estrategias estadísticas para poder solucionarlos. Asimismo, el documento cuenta con capítulos que abordan temas como el uso de variables cualitativas en una regresión lineal y fundamentos básicos de muestreo. En el caso de uso de variables cualitativas como explicativas, se busca que quede claro el manejo de este tipo de variables en los modelos de regresión, la interpretación del efecto de estas y se explora su uso para analizar posibles relaciones no lineales mediante los modelos de regresión con interacciones. En cuanto al tema de muestreo, se brinda un alcance sobre aspectos básicos como el tamaño de la

muestra, el margen de error, la selección de la muestra y el muestreo en el caso de diseños experimentales.

Por último, en el anexo, se incorporan los fundamentos de la teoría de la probabilidad e inferencia estadística para profundizar en la lectura y tener un mejor entendimiento de algunas propiedades y definiciones matemáticas.

En resumen, el presente documento busca ser un material de apoyo para los estudiantes de cursos básicos de econometría para entender los pasos que se deben seguir al momento de realizar un análisis de regresión lineal multivariado, no solo presentando los conceptos y pruebas estadísticas, sino también mediante ejemplos aplicados a lo largo de cada capítulo.

Juan León,
Profesor Auxiliar
del Departamento de Economía de la
Pontificia Universidad Católica del Perú

Marcelo Gallardo,
Asistente de investigación y docencia
de la Facultad de Ciencia e Ingeniería de la
Pontificia Universidad Católica del Perú

LISTA DE SÍMBOLOS

$\mathbb{N}$: conjunto de números naturales, $\mathbb{N} \triangleq \{1, 2, \dots\}$.

$\mathbb{Z}$: conjunto de números enteros, $\mathbb{Z} \triangleq \{\dots, -3, -2, -2, 0, 1, 2, 3 \dots\}$.

$\mathbb{Z}_+$: conjunto de números enteros positivos incluido el cero, $\mathbb{Z}_+ \triangleq \{0, 1, 2, 3 \dots\}$.

$\mathbb{Q}$: conjunto de números racionales.

$\mathbb{R}$: conjunto de números reales.

$\mathbb{R}_+$: conjunto de números reales mayores o iguales a cero.

$\mathbb{R}_{++}$: conjunto de números reales estrictamente mayores a cero.

$\overline{\mathbb{R}}+$: conjunto de números reales mayores o iguales a cero unidos con $\{\infty\}$.

A^c: si A es un conjunto, A^c denota el complemento de dicho conjunto.

$A \subset B$: el conjunto A está incluido en el conjunto B.

$\mathbb{R}^n$: espacio euclidiano de dimensión $n \in \mathbb{N}$.

$||x||$: si x es un vector del espacio vectorial V, $||x||$ denota la norma Euclidiana de dicho vector.

$\mathscr{B}_{\mathbb{R}^k}$ $\sigma-$álgebra de Borel en $\mathbb{R}^k$.

sup$\{A\}$: supremo del conjunto A.

ínf$\{A\}$: ínfimo del conjunto A.

$t \downarrow c$: t tiende a c y $t > c$. Lo mismo aplica para sucesiones de variables aleatorias $X_n \downarrow X$. Se define de manera análoga $t \uparrow c$ y $X_n \uparrow X$.

$\emptyset$: el conjunto vacío.

$\uplus$: unión disjunta.

$\mathbb{P}\lim$: probabilidad límite.

$\xrightarrow{d}$: converge en distribución.

$\xrightarrow{\mathbb{P}}$: converge en probabilidad.

$\mathbb{E}[\cdot]$: valor esperado.

$\text{Var}(\cdot)$: varianza.

$\text{Cov}(\cdot, \cdot)$: covarianza.

Avar: varianza asintótica.

$\mathbf{1}_A$: función indicatriz, $\mathbf{1}_A(x) = 1$ si $x \in A$ y 0 caso contrario.

Índice general

Capítulo 1

Álgebra Matricial

El álgebra matricial es una herramienta de gran utilidad en econometría y, en general, en economía. Este capítulo tiene como objetivo presentar de manera resumida dicha herramienta, manteniendo el rigor matemático y proveyendo ejemplos de aplicación en economía. El enfoque es más práctico y se invita al lector a ahondar en temas subyacentes, como lo son el Álgebra Lineal o el Análisis en Espacios Vectoriales Normados. Véase por ejemplo Axler (2015), Simon and Blume (1994) o Chavez and Gallardo (2023).

La organización de este capítulo es la siguiente: en una primera instancia, se definirán las diversas propiedades sobre el álgebra de matrices. Enseguida, se presentarán propiedades de las matrices de gran interés práctico: matriz transpuesta y rango de una matriz. Luego, se abordarán las nociones de matriz inversa y determinante. Finalmente, se estudiarán algunas aplicaciones de los conceptos teóricos abordados en este capítulo.

1.1. Matrices y operaciones

Definición 1.1.1. Una matriz, es un arreglo rectangular con entradas reales[1]:

$$A \triangleq [a_{ij}] = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mn} \end{pmatrix}.$$

El elemento (a_{ij}) denota al elemento de la $i-$ésima fila y $j-$ésima columna del arreglo.

Ejemplo 1. Considere la siguiente matriz

$$A = \begin{pmatrix} 1 & 0 & 2 \\ 3\pi & \ln(2{,}5) & -1{,}8 \\ e^2 & 2\sqrt{\pi} & 100 \end{pmatrix}.$$

Entonces, por ejemplo, $a_{11} = 1$, $a_{21} = 3\pi$ y $a_{33} = 100$.

Definición 1.1.2. Una matriz tiene dimensión $m \times n$ cuando el número de filas de dicha matriz es $m \in \mathbb{N}$, y su número de columnas es $n \in \mathbb{N}$. El espacio de matrices (o conjunto de matrices) de dimensión $m \times n$, con entradas reales, se denota $\mathcal{M}_{m\times n}$[2]. Eventualmente, si $A \in \mathcal{M}_{m\times n}$, se escribe $A \triangleq A_{m\times n}$.

[1] Si bien las entradas podrían tener elementos de $\mathbb{C}$, funciones etc., nos limitamos al estudio de matrices con entradas $a_{ij} \in \mathbb{R}$.

[2] Este conjunto es un espacio vectorial. Este concepto no se aborda en este capítulo, pero puede ser de gran interés para el lector y se le invita a consultar bibliografía relacionada al Álgebra Lineal. Véase Axler (2015) o Chavez and Gallardo (2023).

Ejemplo 2. Sean las matrices:

$$A = \begin{pmatrix} 1 & 5 \\ 3{,}5 & -6 \\ 1{,}2 & 7 \end{pmatrix}, \; B = \begin{pmatrix} 8 & 9{,}4 & 0{,}2 & 3{,}2 \\ 10550 & 103 & 97 & 1{,}2 \end{pmatrix}.$$

La matriz A tiene dimensión 3×2 mientras que la matriz B tiene dimensión 2×4.

Enseguida, presentaremos algunas propiedades de las operaciones usuales[3] con matrices: la suma, multiplicación por escalar, y finalmente, la multiplicación entre matrices.

Definición 1.1.3. Sean dos matrices A y B. Si estas matrices tienen misma dimensión, es decir, si tienen el mismo número de filas y el mismo número de columnas, definimos de la siguiente manera la suma $A + B$:

$$\begin{aligned} A + B &= \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} b_{11} & b_{12} & \cdots & b_{1n} \\ b_{21} & b_{22} & \cdots & b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ b_{m1} & \cdots & \cdots & b_{mn} \end{pmatrix} \\ &= \begin{pmatrix} a_{11} + b_{11} & a_{12} + b_{12} & \cdots & a_{1n} + b_{1n} \\ a_{21} + b_{21} & a_{22} + b_{22} & \cdots & a_{2n} + b_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} + b_{m1} & \cdots & \cdots & a_{mn} + b_{mn} \end{pmatrix}. \end{aligned}$$

Por como se ha definido la suma entre matrices, claramente $A + B = B + A$. Es decir, la suma entre matrices es una operación conmutativa.

[3] Existen otras como el producto Kronecker. Sin embargo, no lo estudiamos en este texto.

Ejemplo 3. Considere las siguientes matrices

$$A = \begin{pmatrix} 4 & 0 & 5{,}6 \\ 0{,}5 & 1 & -2 \\ 0 & 4{,}7 & 3 \end{pmatrix}, \; B = \begin{pmatrix} 4 & 4{,}2 & 0 \\ 0 & 0{,}8 & -2{,}4 \\ 0{,}3 & 2 & 11 \end{pmatrix}$$

$$\text{y } C = \begin{pmatrix} 2{,}3 & 2 & 1{,}3 \\ 54 & 88 & 1 \end{pmatrix}.$$

Las matrices A y B tienen la misma dimensión (3×3), pero la matriz C tiene dimensión 2×3. Por ende, pueden sumarse las matrices A y B, pero no pueden ser sumadas con la matriz C. En este caso

$$A + B = \begin{pmatrix} 8 & 4{,}2 & 5{,}6 \\ 0{,}5 & 1{,}8 & -4{,}4 \\ 0{,}3 & 6{,}7 & 14 \end{pmatrix}.$$

La suma entre matrices es una operación asociativa. Esto es, si se tienen 3 matrices A, B y C cuya dimensión es la misma, entonces $(A + B) + C = A + (B + C)$.

La matriz nula 0 de dimensión $m\times n$, que corresponde a la matriz cuyas entradas son iguales a cero, cumple la siguiente propiedad: $A + 0 = 0 + A = A$.

$$\begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mn} \end{pmatrix} + \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 0 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mn} \end{pmatrix}.$$

Definamos ahora el producto de una matriz por un escalar.

Definición 1.1.4. Sea $A \in \mathcal{M}_{m\times n}$ y $\alpha \in \mathbb{R}$. Definimos $\alpha \cdot A = A \cdot \alpha$ de la siguiente manera:

$$\alpha \cdot A = \alpha \cdot \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & \cdots & \cdots & a_{mn} \end{pmatrix} = \begin{pmatrix} \alpha a_{11} & \alpha a_{12} & \cdots & \alpha a_{1n} \\ \alpha a_{21} & \alpha a_{22} & \cdots & \alpha a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha a_{m1} & \cdots & \cdots & \alpha a_{mn} \end{pmatrix}.$$

Definiendo esta operación, podemos definir la resta de dos matrices A y B como la suma de A con $(-1) \cdot B$.

Ejemplo 4. Sea $\alpha = 5$ y

$$A = \begin{pmatrix} 3 & -2 & 4 & 5{,}4 \\ -7432 & 88 & 2{,}1 & 0{,}3 \end{pmatrix}.$$

Entonces,

$$\alpha A = 5A = \begin{pmatrix} 15 & -10 & 20 & 27 \\ -37160 & 440 & 10{,}5 & 1{,}5 \end{pmatrix}.$$

La multiplicación entre matrices es una operación más delicada pues, a diferencia de los números, no es conmutativa y no siempre puede efectuarse. Veamos.

Sean $A \in \mathcal{M}_{m\times n}$ y $B \in \mathcal{M}_{\ell\times p}$ dos matrices . Las matrices A y B pueden multiplicarse en el sentido $A \times B$, solo si $n = \ell$.

Ejemplo 5. Sean las matrices,

$$A = \begin{pmatrix} 5 & 4 \\ 1 & -3 \\ -18 & 11 \end{pmatrix},\ B = \begin{pmatrix} 5 & 12 & -14 & 3 \\ -2 & 1 & 9 & 8 \end{pmatrix},\ C = \begin{pmatrix} 4 & -1 \end{pmatrix}.$$

Obsérvese que $A \in \mathcal{M}_{3\times 2}$, $B \in \mathcal{M}_{2\times 4}$ y $C \in \mathcal{M}_{1\times 2}$. Las únicas multiplicaciones entre matrices posibles son $A_{3\times 2} \times B_{2\times 4}$ y $C_{1\times 2} \times B_{2\times 4}$.

Ya hemos visto cual es la condición necesaria para que puedan multiplicarse dos matrices, esto es, que el número de columnas de A sea igual al número de filas de B. Veamos ahora como se ejecuta el producto matricial.

Definición 1.1.5. Formalmente, la multiplicación de dos matrices se define de la siguiente forma:

- Sea $A_{m\times n} = (a_{ij})$ con $1 \leq i \leq m$ y $1 \leq j \leq n$.
- Sea $B_{n\times p} = (b_{ij})$ con $1 \leq i \leq n$ y $1 \leq j \leq p$.
- Entonces, $C = A \times B$ corresponde a la matriz cuyas entradas están dadas por

$$c_{ij} = \sum_{k=1}^{m} a_{ik} b_{kj}.$$

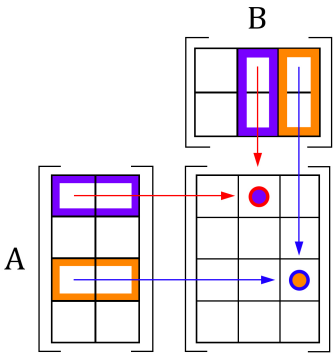

Figura 1.1 Multiplicación de matrices (1).

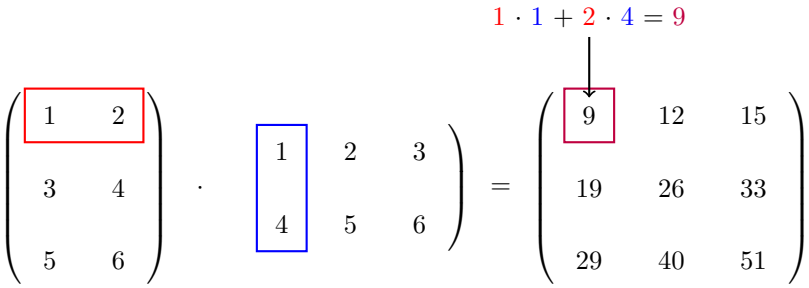

Figura 1.2 Multiplicación de matrices (2).

Note que si tenemos dos matrices $A \in \mathcal{M}_{m\times n}$ y $B \in \mathcal{M}_{n\times p}$, el producto C, es decir, la matriz generada por la multiplicación de A con B, pertenece a $\mathcal{M}_{m\times p}$.

Ejemplo 6. Sean

$$A = \begin{pmatrix} 1 & 10 \\ -2 & 15 \\ 3 & 8 \end{pmatrix} \text{ y } B = \begin{pmatrix} 5 & 6 & 1 \\ 3 & 4 & 2 \end{pmatrix}.$$

Entonces:

$$A \times B = \begin{pmatrix} 1 & 10 \\ -2 & 15 \\ 3 & 8 \end{pmatrix} \times \begin{pmatrix} 5 & 6 & 1 \\ 3 & 4 & 2 \end{pmatrix}$$

$$= \begin{pmatrix} 1\cdot 5 + 10\cdot 3 & 1\cdot 6 + 10\cdot 4 & 1\cdot 1 + 10\cdot 2 \\ -2\cdot 5 + 15\cdot 3 & -2\cdot 6 + 25\cdot 4 & -2\cdot 1 + 15\cdot 2 \\ 3\cdot 5 + 8\cdot 3 & 3\cdot 6 + 8\cdot 4 & 3\cdot 1 + 8\cdot 2 \end{pmatrix}.$$

Finalmente, operando, obtenemos

$$C = A \times B = \begin{pmatrix} 35 & 46 & 21 \\ 35 & 48 & 28 \\ 39 & 50 & 19 \end{pmatrix}.$$

El caso particular de matrices $A \in \mathcal{M}_{1\times n}$ y $B \in \mathcal{M}_{n\times 1}$ son de interés pues cuando se multiplica $A \times B$, se obtiene la operación producto interno entre dos vectores, la cual corresponde en $\mathbb{R}^n$ a

$$A \times B = \begin{pmatrix} x_1 & x_2 & \cdots x_n \end{pmatrix} \times \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \triangleq x \cdot y = \sum_{i=1}^{n} x_i y_i.$$

Ya hemos mencionado que el producto matricial es una operación delicada y más compleja que las operaciones usuales en $\mathbb{R}$. Sin embargo, también mencionamos que no es lo mismo, dadas dos matrices A y B, efectuar $A \times B$ que $B \times A$. Más aún, puede ocurrir que no sea posible efectuar $B \times A$, siendo sin embargo posible efectuar $A \times B$. No obstante, en caso sea posible, puede darse que $A \times B \neq B \times A$. Veamos.

Ejemplo 7. Considere las siguientes matrices

$$A = \begin{pmatrix} 1 & 1 \\ 3 & 2 \end{pmatrix}, \; B = \begin{pmatrix} 1 & 0 \\ -1 & 2 \end{pmatrix}.$$

Si bien

$$A \times B = \begin{pmatrix} 0 & 2 \\ 1 & 4 \end{pmatrix},$$

tenemos que

$$B \times A = \begin{pmatrix} 1 & 1 \\ 5 & 3 \end{pmatrix}.$$

Terminamos esta sección presentando 3 tipos de matrices que serán de utilidad a continuación:

1. La matriz identidad I: una matriz cuyas entradas son cero salvo en la diagonal, donde valen uno:

$$I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & 1 & 0 \\ 0 & \cdots & 0 & 1 \end{pmatrix}.$$

   Obsérvese que cuando se multiplica esta matriz por una matriz A, se obtiene la misma matriz A (siempre y cuando la multiplicación sea posible).

2. Las matrices simétricas: matrices cuyas entradas son las mismas en las coordenadas a_{ij} y a_{ji}. Por ejemplo:

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{12} & a_{22} & a_{23} & a_{24} \\ a_{13} & a_{23} & 2_{33} & a_{34} \\ a_{14} & a_{24} & a_{34} & a_{44} \end{pmatrix}.$$

3. Las matrices triangulares superiores[4]: las matrices cuyas entradas son iguales a cero por debajo de la diagonal:

$$\mathbb{T} = \begin{pmatrix} t_{11} & t_{12} & \cdots & t_{1n} \\ 0 & \ddots & \ddots & t_{2n} \\ \vdots & \ddots & t_{n-1,n-1} & t_{n-1,n} \\ 0 & \cdots & 0 & t_{nn} \end{pmatrix}.$$

En los casos anteriores, las matrices han sido cuadradas, i.e., el número de filas es igual al número de columnas. De no haber sido

[4]La definición es análoga para las matrices triangulares inferiores.

el caso, las definiciones pierden sentido. Note también que dadas 2 matrices cuadradas A y B de misma dimensión, siempre es posible multiplicar A con B así como B con A.

Ejemplo 8. La matriz Q

$$Q = \begin{pmatrix} a & b \\ c & d \end{pmatrix}$$

es una matriz cuadrada de orden 2×2.

1.2. Matriz transpuesta y rango

En esta sección presentaremos las nociones de matriz transpuesta y rango, centrales para el desarrollo de la teoría de las matrices no singulares (invertibles).

Definición 1.2.1. La transpuesta de una matriz $A_{m\times n}$, es la matriz que se obtiene al intercambiar las filas y columnas de la matriz A. Esta matriz se denota usualmente como A^T. De manera más analítica, la entrada a_{ij} toma el valor a_{ji} y la entrada a_{ji} toma el valor a_{ij}. Más aún, si la matriz era de dimensión $m \times n$, su transpuesta será de dimensión $n \times m$.

Ejemplo 9. Sean las matrices

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{pmatrix} \text{ y } B = \begin{pmatrix} b_{11} & b_{12} & b_{13} & b_{14} \\ b_{21} & b_{22} & b_{23} & b_{24} \end{pmatrix}.$$

Entonces sus transpuestas son, respectivamente

$$A^T = \begin{pmatrix} a_{11} & a_{21} & a_{31} & a_{41} \\ a_{12} & a_{22} & a_{32} & a_{42} \\ a_{13} & a_{23} & a_{33} & a_{43} \\ a_{14} & a_{24} & a_{34} & a_{44} \end{pmatrix} \text{ y } B^T = \begin{pmatrix} b_{11} & b_{21} \\ b_{12} & b_{22} \\ b_{13} & b_{23} \\ b_{14} & b_{24} \end{pmatrix}.$$

Note que si tenemos una matriz $A \in \mathcal{M}_{m\times n}$, entonces, multiplicarla por su transpuesta, siempre genera una matriz cuadrada. En este caso

$$A \times A^T = Q_{m\times m}$$
$$A^T \times A = P_{n\times n}.$$

Teorema 1. Sean $A, B \in \mathcal{M}_{m\times n}$ y $\alpha \in \mathbb{R}$. Entonces

- $(A+B)^T = A^T + B^T$.
- $(A-B)^T = A^T - B^T$.
- $(A^T)^T = A$.
- $(\alpha A) = \alpha A^T$.

Estas propiedades pueden ser probadas directamente y las dejamos como ejercicios para el lector interesado. El siguiente resultado sin embargo, es menos directo y por ello su prueba es desarrollada.

Teorema 2. Sean $A \in \mathcal{M}_{m\times n}$ y $B \in \mathcal{M}_{n\times p}$. Entonces

$$(AB)^T = B^T A^T.$$

Demostración. Permítanos denotar $(C)_{ij} = c_{ij}$. Tenemos entonces

$$\begin{aligned}((AB)^T)_{ij} &= (AB)_{ji} \\ &= \sum_k a_{jk} b_{ki} \\ &= \sum_k (A^T)_{kj} (B^T)_{ik} \\ &= \sum_k (B^T)_{ik} (A^T)_{kj} \\ &= (B^T A^T)_{ij}.\end{aligned}$$

□

Ejemplo 10. Sean A y B las siguientes matrices

$$A = \begin{pmatrix} a_{11} & a_{12} & a_{13} \\ a_{21} & a_{22} & a_{23} \end{pmatrix}, \; B = \begin{pmatrix} b_{11} & b_{12} \\ b_{21} & b_{22} \\ b_{31} & b_{32} \end{pmatrix}.$$

Mediante un cálculo directo de $A \times B$ obtenemos

$$AB = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} \\ a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}.$$

Transponiendo, obtenemos

$$(AB)^T = \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}.$$

Ahora, por otro lado,

$$B^T A^T = \begin{pmatrix} b_{11} & b_{21} & b_{31} \\ b_{12} & b_{22} & b_{32} \end{pmatrix} \times \begin{pmatrix} a_{11} & a_{21} \\ a_{12} & a_{22} \\ a_{13} & a_{23} \end{pmatrix}$$
$$= \begin{pmatrix} a_{11}b_{11} + a_{12}b_{21} + a_{13}b_{31} & a_{21}b_{11} + a_{22}b_{21} + a_{23}b_{31} \\ a_{11}b_{12} + a_{12}b_{22} + a_{13}b_{32} & a_{21}b_{12} + a_{22}b_{22} + a_{23}b_{32} \end{pmatrix}.$$

Ya habiendo expuesto la noción y propiedades de la matriz transpuesta, seguimos con el concepto de rango de una matriz.

Definición 1.2.2. Decimos que los vectores $\{v_\ell\}_{\ell=1}^n$ son linealmente independientes (l.i.) si no existen escalares $\gamma_1, ..., \gamma_n$ diferentes a cero tales que

$$\sum_{\ell=1}^{n} \gamma_\ell v_\ell = 0.$$

Contrariamente, si los vectores $\{v_\ell\}_{\ell=1}^n$ son linealmente dependientes (l.d.), entonces existen $\gamma_1, ..., \gamma_n$ no todos iguales a cero, tales que

$$\sum_{\ell=1}^{n} \gamma_\ell v_\ell = 0.$$

Alternativamente, existe al menos un vector v_j en el conjunto de vectores $\{v_\ell\}_{\ell=1}^n$ que es combinación lineal del resto:

$$v_j = \sum_{\ell \neq j} \theta_\ell v_\ell.$$

Note que si la colección de vectores $\{v_\ell\}_{\ell=1}^n$ contiene al vector nulo, siempre es l.d.

Definición 1.2.3. El rango de una matriz $A_{m\times n}$ corresponde al número de columnas (o filas) linealmente independientes. Esto es, si identificamos cada columna de la matriz A como vector en $\mathbb{R}^m$:

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1j} & \cdots & a_{1n} \\ a_{21} & \cdots & a_{2j} & \cdots & a_{2n} \\ \vdots & \vdots & \vdots & \vdots & \\ a_{m1} & \cdots & a_{mj} & \cdots & a_{mn} \end{pmatrix} \implies v_j = \begin{pmatrix} a_{1j} \\ a_{2j} \\ \vdots \\ a_{mj} \end{pmatrix}$$

entonces el rango de A es el número de vectores v_j linealmente independientes. Note que, en cualquier caso, el rango será menor o igual a $N = \text{mín}\{m, n\}$.

Ejemplo 11. Sea

$$A = \begin{pmatrix} 2 & 1 & 3 & 0 \\ 5 & 2 & 7 & 0 \\ 4 & 1 & 5 & 0 \end{pmatrix}.$$

Vemos que la tercera columna es combinación lineal de las dos primeras:

$$\begin{pmatrix} 2 \\ 5 \\ 4 \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} 3 \\ 7 \\ 5 \end{pmatrix}.$$

Por otro lado, la cuarta columna es combinación lineal del resto también pues,

$$\begin{pmatrix} 2 \\ 5 \\ 4 \end{pmatrix} + \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} - \begin{pmatrix} 3 \\ 7 \\ 5 \end{pmatrix} = 0_{3\times 1}.$$

Finalmente, como la columna 1 no es múltiplo de la segunda, el rango de la matriz A es 2.

1.3. Matriz inversa y determinante

En el caso de los números reales, si tenemos $x \in \mathbb{R}$, con $x \neq 0$, podemos encontrar un número y tal que $xy = yx = 1$. Este número $y = x^{-1} = \frac{1}{x}$ es conocido como la inversa de x. En el caso de las matrices, existe una idea muy similar. Dada una matriz $A \in \mathcal{M}_{m\times n}$ nos preguntamos cuando es posible encontrar una matriz B tal que $A \times B = B \times A = I$. Recordemos que una propiedad de interés de la matriz identidad $I \in \mathcal{M}_{n\times n}$ es que, para cualquier matriz $A \in \mathcal{M}_{n\times n}$:

$$I \times A = A \times I = A.$$

Puesto que deseamos poder realizar la multiplicación $A \times B$ tanto como la multiplicación $B \times A$, la matrices A y B deben ser cuadradas y de misma dimensión $n \times n$. La matriz identidad resultante tendrá entonces orden $n \times n$: $I \triangleq I_{n\times n}$.

Definición 1.3.1. Dada una matriz $A \in \mathcal{M}_{n\times n}$, si $B \in \mathcal{M}_{n\times n}$ es tal que

$$A \times B = B \times A = \mathbb{I},$$

entonces la matriz B es la matriz inversa de A, y se denota A^{-1}.

Ciertamente, puede que dicha matriz A^{-1} no exista. Nos preguntamos entonces ¿cuándo existe la matriz A^{-1}? Para responder esta pregunta, es necesario definir el concepto de determinante de una matriz.

Definición 1.3.2. Sea $A \triangleq (a_{ij}) \in \mathcal{M}_{n\times n}(\mathbb{R})$, definimos la aplicación

$$\det : \mathcal{M}_{n\times n}(\mathbb{R}) \to \mathbb{R},$$

conocida como determinante, de manera recursiva tal y como sigue:

- Si $n = 1$, $\det(A) = |A| = a_{11}$.
- Si $n > 1$, definimos A_{ij} como la matriz que corresponde a la matriz A eliminando la fila i y la columna j[5]:

$$|A| = a_{11}|A_{11}| + ... + (-1)^{k+1}a_{1k}|A_{1k}| + ... + (-1)^{n+1}a_{1n}|A_{1n}|.$$

Dada una matriz A,

$$A = \begin{pmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_n \end{pmatrix}$$

usualmente se denota su determinante de la manera siguiente:

$$|A| = \begin{vmatrix} a_{11} & \cdots & a_{1n} \\ \vdots & \ddots & \vdots \\ a_{n1} & \cdots & a_n \end{vmatrix}.$$

Es decir, $\det(A) \triangleq |A|$.

Ejemplo 12. Sea

$$A = \begin{pmatrix} a & b \\ c & d \end{pmatrix}.$$

Entonces, su determinante estará dado por

$$|A| = ad - bc.$$

[5]Note que $A_{ij} \in \mathcal{M}_{n-1\times n-1}(\mathbb{R})$.

Ejemplo 13. Sea

$$B = \begin{pmatrix} 1 & -2 & 3 \\ 2 & 1 & -1 \\ 1 & 5 & 2 \end{pmatrix}.$$

Entonces, su determinante estará dado por

$$|B| = 1\begin{vmatrix} 1 & -1 \\ 5 & 2 \end{vmatrix} - (-2)\begin{vmatrix} 2 & -1 \\ 1 & 2 \end{vmatrix} + 3\begin{vmatrix} 2 & 1 \\ 1 & 5 \end{vmatrix} = 44.$$

La teoría del determinante de una matriz se extiende mucho más allá de los ejemplo prácticos y la definición brindada. Se puede ahondar desde la perspectiva de las formas multilineales alternadas, introduciendo el concepto de permutación. Sin embargo, este no es el objetivo ni será necesario para los siguiente capítulos. El lector interesado puede consultar Girfone (2018) o Roman (2008). Enseguida, vamos a presentar las principales propiedades del determinante. Luego, usando esta nueva herramienta, podremos caracterizar de manera sistemática el concepto de matriz invertible y de matriz inversa.

Teorema 3. Sean $A, B \in \mathcal{M}(\mathbb{R})$, $\alpha \in \mathbb{R}$. Entonces

- $\det(AB) = \det(A) \cdot \det(B)$.
- $|A^{-1}| = \frac{1}{|A|}$.
- $|A| = (-1)^p |\sigma_p(A)|$ donde $\sigma_p(A)$ corresponde a la matriz que se genera vía la permutación (intercambio) de columnas de la

matriz inicial A, siendo p el número de cambios. [6]

- $|\alpha A| = \alpha^n A$.
- $|A| = |A_{\Sigma_i}|$, donde A_{Σ_i} es la matriz A cuyas columnas corresponden a las columnas iniciales sumándoles combinaciones lineales de las otras.[7] El mismo razonamiento se aplica al caso de las filas.

La prueba de estas propiedades son consecuencia directa de la Definición 1.3.2. Dado el enfoque de este texto, se deja como ejercicio para el lector interesado demostrar las propiedades. Como sugerencia, aplique inducción.

Definición 1.3.3. Una matriz es invertible o no singular si su determinante es diferente de cero.

Definición 1.3.4. La matriz de cofactores asociada a una matriz

[6] Por ejemplo

$$A = \begin{pmatrix} 1 & 4 \\ 3 & -2 \end{pmatrix}, \ \sigma_{p=1}(A) = \begin{pmatrix} 4 & 1 \\ -2 & 3 \end{pmatrix}.$$

[7] Por ejemplo,

$$A = \begin{pmatrix} 2 & 4 & 7 \\ 3 & -2 & 5 \\ 1 & 0 & 2 \end{pmatrix}, \ A_{\Sigma_i} = \begin{pmatrix} 2 & 13 & 7 \\ 3 & 6 & 5 \\ 1 & 3 & 2 \end{pmatrix}.$$

La matriz A_{Σ_i} es prácticamente idéntica a la matriz A salvo que a la columna 2 se le sumó (vectorialmente) las columnas 1 y 3. Obsérvese que en ambos casos el determinante es igual a 2.

A está definida de la siguiente manera[8]:

$$[\text{cof}(A)]_{ij} = (-1)^{i+j}|A_{ij}|.$$

Ejemplo 14. Sea

$$A = \begin{pmatrix} 1 & 2 & 4 \\ 3 & 5 & 1 \\ -1 & 2 & 3 \end{pmatrix}.$$

Entonces,

$$[\text{cof}(A)]_{11} = (-1)^{1+1}\begin{vmatrix} 5 & 1 \\ 2 & 3 \end{vmatrix} = 13.$$

Definición 1.3.5. Si $A \in \mathcal{M}_{n\times n}$ es invertible, entonces su inversa A^{-1} se define de la manera siguiente

$$A^{-1} = \frac{1}{|A|}\text{cof}(A)^T.$$

Ejemplo 15. Sea por ejemplo la matriz

$$A = \begin{pmatrix} 1 & 0 & 1 \\ 4 & 3 & 1 \\ 1 & 2 & 4 \end{pmatrix}.$$

Primero, calculamos el determinante y verificamos que no sea igual a cero

$$|A| = 1\begin{vmatrix} 3 & 1 \\ 2 & 4 \end{vmatrix} - 0\begin{vmatrix} 4 & 1 \\ 1 & 4 \end{vmatrix} + 1\begin{vmatrix} 4 & 3 \\ 1 & 2 \end{vmatrix} = 15.$$

[8]Recordar que A_{ij} es la matriz que resulta de eliminar la fila i y columna j de A.

Enseguida, obtenemos los cofactores

$$
\begin{aligned}
[\text{cof}(A)]_{11} &= 10 \\
[\text{cof}(A)]_{12} &= -15 \\
[\text{cof}(A)]_{13} &= 5 \\
[\text{cof}(A)]_{21} &= -10 \\
[\text{cof}(A)]_{22} &= 3 \\
[\text{cof}(A)]_{23} &= -2 \\
[\text{cof}(A)]_{31} &= -3 \\
[\text{cof}(A)]_{32} &= 3 \\
[\text{cof}(A)]_{33} &= 3.
\end{aligned}
$$

Luego, la matriz de cofactores sería:

$$
\text{cof}(A) = \begin{pmatrix} 10 & -15 & 5 \\ -10 & 3 & -2 \\ -3 & 3 & 3 \end{pmatrix}.
$$

Transponemos:

$$
[\text{cof}(A)]^T = \begin{pmatrix} 10 & -10 & -3 \\ -15 & 3 & 3 \\ 5 & -2 & 3 \end{pmatrix}.
$$

Finalmente, multiplicando por $1/|A|$, se obtiene la inversa de A:

$$
A^{-1} = \begin{pmatrix} 2/3 & 2/15 & -1/5 \\ -1 & 1/5 & 1/5 \\ 1/3 & -2/15 & 1/5 \end{pmatrix}.
$$

Terminamos esta sección con una definición que es de gran utilidad en la práctica. En la última sección, a través de tres ejemplos de interés, exhibiremos 3 como las herramientas presentadas hasta el momento son empleadas en el contexto económico.

Definición 1.3.6. Se dice que dos matrices $A, B \in \mathcal{M}_{n\times n}$ son semejantes si existe una matriz P no singular tal que

$$P^{-1}AP = B.$$

Note que dadas dos matrices semejantes,

$$\det(A) = \det(B).$$

En efecto, existe P invertible tal que

$$\begin{aligned}\det(A) &= \det(P^{-1}BP) = \det(P^{-1})\det(B)\det(P)\\ &= \frac{1}{\det(P)}\det(B)\det(P) = \det(B).\end{aligned}$$

Con esto, hemos concluido nuestro breve repaso sobre el determinante de una matriz y la inversa de una matriz. A continuación, pasamos a las aplicaciones de estas herramientas, y con ello, concluimos este breve capítulo.

1.4. Aplicaciones

Si bien existen numerosas aplicaciones de la teoría desarrollada en las secciones previas, presentamos tres que aparecen frecuentemente en la práctica. Empezamos con la aplicación en la determinación de conjuntos de vectores l.i.

Dado un conjunto finito de vectores $\{v_\ell\}_{\ell=1}^n \in \mathbb{R}^n$, podemos verificar que este conjunto es linealmente independiente definiendo la matriz A como la matriz cuyas columnas corresponden a estos vectores

$$A = \begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}$$

y verificando que su determinante es diferente de 0.

Ejemplo 16. Sean los vectores $v_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$ y $v_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$. Entonces,

$$A = \begin{pmatrix} 1 & 2 \\ 2 & 4 \end{pmatrix}.$$

Calculando, se observa que $|A| = 0$. Uno puede verificar que, en efecto, los vectores no son linealmente independientes pues $v_2 = 2v_1$.

Ejemplo 17. Sean ahora los vectores $v_1 = \begin{pmatrix} 1 \\ 3 \end{pmatrix}$ y $v_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$. Tenemos que

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 4 \end{pmatrix}.$$

Luego, $|A| = -1$. Uno puede verificar que, en efecto, los vectores son linealmente independientes pues no existe $\alpha \in \mathbb{R}$ tal que $v_2 = \alpha v_1$.

Veamos ahora el tema de la resolución de sistemas lineales, es decir, ecuaciones del tipo:

$$\begin{aligned} a_{11}x_1 + ... + a_{1n}x_n &= y_1 \\ a_{21}x_1 + ... + a_{2n}x_n &= y_2 \\ &\vdots \\ a_{n1}x_1 + ... + a_{nn}x_n &= y_n. \end{aligned}$$

Lo primero que se observa es que este tipo sistemas pueden expresarse de la manera siguiente,

$$Ax = y$$

siendo $x = (x_1, ..., x_n)^T$ (desconocido), $y = (y_1, ..., y_n)^T$ (dato) y[9]

$$A = \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}.$$

Si $|A| \neq 0$, entonces el sistema puede resolverse, y su solución será:

$$x = A^{-1}y.$$

Ejemplo 18. Considere el siguiente sistema de ecuaciones lineales

$$2x_1 + 3x_2 - x_3 = 1$$
$$4x_1 + 2x_3 = 2$$
$$x_1 + x_2 = 1.$$

Este sistema se expresa se la siguiente manera, de forma matricial:

$$\begin{pmatrix} 2 & 3 & -1 \\ 4 & 0 & 2 \\ 1 & 1 & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}.$$

Entonces,

$$x = A^{-1}y = \begin{pmatrix} 1 & 1/2 & -3 \\ -1 & -1/2 & 4 \\ -2 & -1/2 & 6 \end{pmatrix} \begin{pmatrix} 1 \\ 2 \\ 1 \end{pmatrix} = \begin{pmatrix} -1 \\ 2 \\ 3 \end{pmatrix}.$$

[9] Todos los parámetros siendo conocidos.

Pasamos a nuestra última aplicación. Si bien el enfoque a continuación es informal[10], nos permite apreciar la importancia de los conceptos presentados en este capítulo en el contexto de la teoría microeconómica.

Dado el sistema de ecuaciones

$$\underbrace{\begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}}_{=A} \begin{pmatrix} x \\ y \\ z \end{pmatrix} = \begin{pmatrix} j \\ k \\ \ell \end{pmatrix},$$

se tiene que, siempre y cuando A sea invertible,

$$x = \frac{1}{|A|} \begin{vmatrix} j & b & c \\ k & e & f \\ \ell & h & i \end{vmatrix}$$

$$y = \frac{1}{|A|} \begin{vmatrix} a & j & c \\ d & k & f \\ g & \ell & i \end{vmatrix}$$

$$z = \frac{1}{|A|} \begin{vmatrix} a & b & j \\ d & e & k \\ g & h & \ell \end{vmatrix}.$$

A esto se le conoce como la regla de Cramer.

Ejemplo 19. Dado el problema de maximización de la utilidad[11], asumiendo que se satisfacen las condiciones de Inada para la función

[10] Una justificación más rigurosa hace uso del teorema de la función implícita, véase Chavez and Gallardo (2023).

[11] En el siguiente capítulo ahondamos en temas de optimización.

de utilidad $U(x,y)$ Chavez and Gallardo (2023), el problema se escribe

$$\begin{aligned} &\text{máx } U(x,y) \\ &\text{s.a. : } p_x x + p_y y = I. \end{aligned}$$

Deseamos conocer $\dfrac{dx}{dI}$ y $\dfrac{dx}{dp_x}$. Para simplificar usamos la notación

$$\frac{\partial U}{\partial x} = U_x, \text{ y } \frac{\partial U}{\partial y} = U_y.$$

De las condiciones de primer orden, obtenemos el siguiente sistema

$$\begin{aligned} \frac{\partial \mathcal{L}}{\partial x} &= U_x - \lambda p_x = 0 \\ \frac{\partial \mathcal{L}}{\partial y} &= U_y - \lambda p_y = 0 \\ \frac{\partial \mathcal{L}}{\partial \lambda} &= I - p_x x - p_y y = 0. \end{aligned}$$

Sacando diferenciales en estas tres ecuaciones, se obtiene

$$\begin{aligned} d(U_x - \lambda p_x) &= U_{xx}dx + U_{xy}dy - d\lambda p_x - \lambda dp_x = 0 \\ d(U_y - \lambda p_y) &= U_{yy}dy + U_{yx}dx - d\lambda p_y - \lambda dp_y = 0 \\ d(I - p_x x - p_y y) &= dI - dxp_x - xdp_x - dyp_y - ydp_y = 0. \end{aligned}$$

Las condiciones de Inada aseguran que los bienes son normales y las preferencias convexas Mas-Colell et al. (1995). Por ende:

$$U_{xy} > 0, U_{xx}, U_{yy} < 0.$$

Luego, el sistema el sistema de ecuaciones se convierte en

$$\begin{pmatrix} U_{xx} & U_{xy} & -p_x \\ U_{yx} & U_{yy} & -p_y \\ -p_x & -p_y & 0 \end{pmatrix} \begin{pmatrix} dx \\ dy \\ d\lambda \end{pmatrix} = \begin{pmatrix} \lambda dp_x \\ \lambda dp_y \\ xdp_x + ydp_y - dI \end{pmatrix}.$$

Aplicando la regla de Cramer para obtener

$$\frac{dx}{dI},$$

considerando $dp_x = dp_y = 0$ (mantener precios constantes), se calcula

$$dx = \frac{1}{|A|}\begin{vmatrix} 0 & U_{xy} & -p_x \\ 0 & U_{yy} & -p_y \\ -dI & -p_y & 0 \end{vmatrix},$$

donde

$$\begin{aligned} |A| &= \det \begin{pmatrix} U_{xx} & U_{xy} & -p_x \\ U_{yx} & U_{yy} & -p_y \\ -p_x & -p_y & 0 \end{pmatrix} \\ &= \underbrace{-U_{xx}p_y^2 + U_{xy}p_xp_y + U_{yx}p_xp_y - U_{yy}p_x^2 > 0.}_{\text{preferencias convexas, i..e, } u \text{ cuasi cóncava}} \end{aligned}$$

Luego,

$$\frac{dx}{dI} = \frac{U_{xy}p_y - p_xU_{yy}}{|A|} > 0.$$

La desigualdad se obtiene teniendo en cuenta que la utilidad marginal es decreciente en cada bien $U_{xx}, U_{yy} < 0$, pero las derivadas cruzadas, puesto que las preferencias se suponen convexas, son positivas $U_{xy}, U_{yx} > 0$ Mas-Colell et al. (1995).

En relación a $\frac{dx}{dp_x}$, considerando $dp_y = dI = 0$

$$dx = \frac{1}{|A|}\begin{vmatrix} \lambda dp_x & U_{xy} & -p_x \\ 0 & U_{yy} & -p_y \\ xdp_x & -p_y & 0 \end{vmatrix}.$$

Expandiendo el determinante del numerador,

$$\begin{aligned}\frac{dx}{dp_x} &= \frac{-\lambda p_y^2 - x p_y U_{xy} + x p_x U_{xx}}{|A|} \\ &= \frac{-\lambda p_y^2 + x(U_{yy} p_x - U_{xy} p_y)}{|A|} \\ &= -\frac{\lambda p_y^2}{|A|} - x\frac{dx}{dI} < 0.\end{aligned}$$

Esta ecuación indica que cuando el precio de un bien aumenta, la demanda por este bien se ve reducida por un efecto sustitución y un efecto ingreso[12].

Si bien ya hemos concluido con los conceptos de álgebra matricial que serán requeridos a continuación, existe aún una vasta cantidad de tópicos por explorar. Estos, fundamentan algunos de los resultados que serán obtenidos en capítulos posteriores, pero van más allá del alcance de este libro. Por ejemplo, la diagonalización de matrices, la ortogonalización de vectores, la descomposición polar, la forma canónica de Jordan, etc. Estos temas pueden ser explorados en Chavez and Gallardo (2023), Axler (2015) o Roman (2008). En el siguiente capítulo, se abordarán otras herramientas que serán de gran utilidad en los capítulos destinados propiamente a la teoría econométrica. En concreto, en el siguiente capítulo se discutirán los fundamentos de la optimización estática. Para una discusión más detallada, consultar Sundaram (1996) o Chavez and Gallardo (2023).

[12] Este resultado gana en interpretación cuando se deriva la ecuación de Slutsky.

Capítulo 2

Fundamentos de Optimización Estática

La optimización es una de las ramas más activas e importantes de la matemática aplicada pues, sus aplicaciones en diversas disciplinas, como la economía o la física, son de gran amplitud. En este capítulo, haremos un repaso bastante breve de las principales técnicas de optimización que serán muy útiles en el desarrollo de futuros capítulos.

El famoso matemático Leonhard Euler (1707 - 1783) mencionó que la noción de mínimo y máximo aparece naturalmente en la gran mayoría de acontecimientos en el universo. Ciertamente Euler hacía referencia a las aplicaciones en la física, pero, como veremos más adelante, Euler acertó también en el cuadro de la economía. Desde el problema de maximización de la utilidad hasta la recta de mínimos cuadrados, los mínimos y máximos son de gran interés en las diversas ramas de esta ciencia social.

En una primera instancia, analizamos el caso de funciones a variable real. Luego, extendemos el análisis a las funciones de varias variables. Ciertamente, este capítulo juega el rol de un breve repaso y no tiene intenciones de realizar una presentación exhaustiva o integral de la teoría de la optimización[1]. Por ejemplo, no presentamos[2] en este capítulo los problemas de Lagrange o Kuhn-Tucker. Para estudiar estos tópicos, invitamos al lector interesado a revisar Chavez and Gallardo (2023), Simon and Blume (1994) o Sundaram (1996).

2.1. Funciones de variable real

En esta sección consideramos funciones de variable real: $f : \mathbb{R} \to \mathbb{R}$, que son de clase C^2, es decir, con segunda derivada $\frac{d^2 f}{dx^2} = f''(x)$ continua. Nuestro objetivo será resolver

$$\begin{aligned} &\text{opt } f(x) \\ &\text{s.a} : x \in X \subset \mathbb{R}. \end{aligned}$$

Acá opt puede significar max o min. La pregunta de interés a continuación es, dada una función $y = f(x)$ con $x \in [a, b] = I \triangleq X \subset \mathbb{R}$, ¿cómo encontrar $x^* \in I$ tal que $f(x^*) \geq f(x)$ para cualquier otro x? Es decir:

$$f(x^*) = \max_{a \leq x \leq b} f(x).$$

[1] El lector interesado puede profundizar consultando Boyd and Vandenberghe (2004), Lenberger and Ye (2021) o Chavez and Gallardo (2023).

[2] A modo de nota brindamos el enunciado del teorema pero no ejemplo.

Además, es muy pertinente preguntarse si dicho x^* existe[3].

Teorema 4. Bajo los supuestos hechos sobre la función f y suponiendo que el máximo existe y es tal que $x^* \in (a, b)$ (es decir que x^* pertenezca al interior del intervalo), entonces

$$f'(x^*) = 0.$$

Demostración. Por un lado, sabemos que, para $\epsilon > 0$ suficientemente chico,

$$f(x^* + \varepsilon) \leq f(x^*)$$
$$f(x^* - \varepsilon) \leq f(x^*).$$

Dividiendo entre ϵ y haciendo $\epsilon \to 0$, concluimos que

$$f'(x^*) \geq 0$$
$$f'(x^*) \leq 0.$$

Es decir, $f'(x^*) = 0$. □

El caso de un mínimo es análogo. Al Teorema 4 se le conoce como Condición de Primer Orden.

Por el Teorema de Weierstrass Chavez and Gallardo (2023), dado que $[a, b]$ es compacto y $f(\cdot)$ continua, siempre existe un mínimo y un máximo. Entonces, para encontrar dichos puntos, se obtiene x^* tal que $f'(x^*) = 0$ (pueden ser ciertamente varios puntos

[3] Al ser I un intervalo cerrado, es compacto en la topología usual de $\mathbb{R}$. Como f es continua (pues es C^2) el problema de optimización tiene solución por el teorema de Weierstrass. Para mayores desarrollos de este teorema, ver Abbott (2015) o Chavez and Gallardo (2023).

que cumplan con esta condición), y se compara $f(x^*)$ con $f(a)$ y $f(b)$.

Ejemplo 20. Sea $f(x) = 1 - x^2$. Ciertamente, como $x^2 \geq 0$ para todo $x \in \mathbb{R}$, $\max_{x\in\mathbb{R}} f(x) = 1$ y esto se obtiene en $x^* = 0$. Por otro lado, $f'(x^*) = 0$:

$$f'(x^*) = \left.\frac{df}{dx}\right|_{x=0} = -2x|_{x=0} = 0,$$

lo cual, de cierta forma, verifica el Teorema 4.

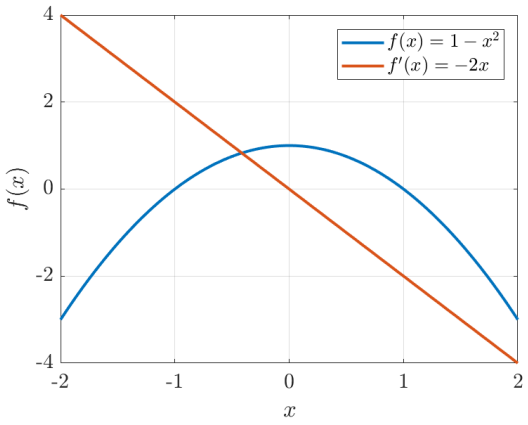

Figura 2.1 Función cuadrática.

Ejemplo 21. La función de costo total de un fabricante está dada por:

$$C(q) = \frac{q^2}{4} + 3q + 400.$$

Acá C es el costo total de producir q unidades. ¿Para qué nivel de producción q^* será el costo promedio por unidad mínimo? Como el costo promedio por unidad es

$$\frac{C}{q} = \frac{q}{4} + 3 + \frac{400}{q},$$

para obtener el candidato a mínimo derivamos e igualamos a cero

$$\begin{aligned}\frac{d}{dq}\left(\frac{C}{q}\right) &= \frac{d}{dq}\left(\frac{q}{4}+3+\frac{400}{q}\right)\\ &= \frac{1}{4}-\frac{400}{q^2}\\ &= \frac{q^2-1600}{4q^2}=0.\end{aligned}$$

Obtenemos $q^* = \pm 40$. Como $q \geq 0$, nos quedamos con $q^* = 40$.

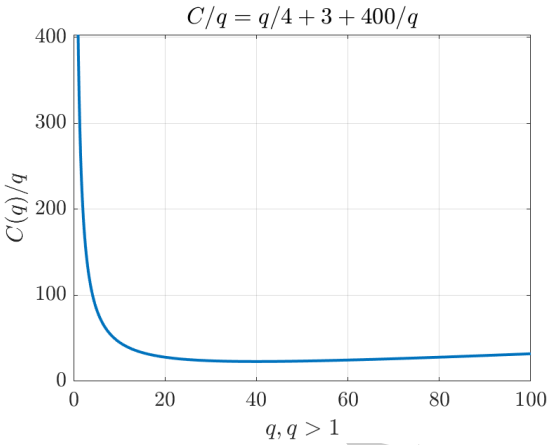

Figura 2.2 Función de costo medio.

2.1.1. Condición de segundo orden

La condición de primer orden es una condición necesaria más no suficiente. Es decir, si x^* es un punto interior de X (véase la definición de punto interior en Chavez and Gallardo (2023)) y es un máximo (o mínimo), entonces $f'(x^*) = 0$. Sin embargo, no necesariamente $f'(x^*) = 0$ implica que x^* sea un óptimo. Más aún, el máximo o mínimo puede encontrarse en el borde de X (extremos del intervalo en caso $X = I$).

Ejemplo 22. Sea $f(x) = x^3$ e $X = I = [-1, 1] \subset \mathbb{R}$. Ciertamente, el máximo de la función es alcanzado en $x = 1$. Sin embargo, $f'(0) = 0$.

El ejemplo anterior nos muestra que necesitamos condiciones adicionales que aseguren que x^*, tal que $f'(x^*) = 0$, sea un óptimo.

Teorema 5. Sea $f : I \subset \mathbb{R} \to \mathbb{R}$ una función dos veces diferenciable en su interior. Entonces:

- Si $f'(x^*) = 0$ y $f''(x^*) > 0$, con $x^* \in I$, entonces x^* es un mínimo local[4].
- Si $f'(x^*) = 0$ y $f''(x^*) < 0$, con $x^* \in I$, entonces x^* es un máximo local.

Un criterio que permite analizar si se trata de un óptimo global (sobre todo X) pasa por analiza la convexidad o concavidad de la función. Al final de este capítulo se hará un breve comentario al respecto.

Ejemplo 23. Sea nuevamente $f(x) = 1 - x^2$. Verifiquemos mediante el último criterio que $x^* = 0$ es un candidato a máximo local. De acuerdo con la Teorema 5, como

$$\frac{d^2 f}{dx^2} = -2 < 0,$$

$x^* = 0$ es un máximo local.

[4]La palabra local hace alusión a que nos movemos en una vecindad, i.e., un intervalo abierto $I = (x^* - \delta, x^* + \delta)$, del punto x^*.

Ejemplo 24. En estadística, una de las distribuciones más importantes es la distribución normal. La función de densidad[5] asociada a esta distribución es

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}},$$

donde μ es la media y σ^2 la varianza. Podemos verificar fácilmente que $x^* = \mu$ es un candidato a óptimo (máximo) local para $f(x)$. En efecto,

$$f'(x) = -\frac{(x-\mu)}{\sigma^2\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} = -\frac{(x-\mu)f(x)}{\sigma^2}.$$

Luego, evaluando en $x^* = \mu$, se obtiene $f'(x^*) = 0$. Ahora bien, por medio del Teorema 5 podemos determinar que se trata efectivamente de un máximo. Para esto, calculamos

$$f''(x) = -\frac{f(x)}{\sigma^2} - \frac{(x-\mu)^2 f(x)}{\sigma^4}.$$

Evaluando nuevamente en $x^* = \mu$, puesto que[6]

$$f''(\mu) = -\frac{f(\mu)}{\sigma^2} < 0,$$

concluimos entonces que f alcanza un máximo local cuando x es igual a μ[7].

En la práctica, los fenómenos usualmente dependen de más de una variable. Por ejemplo, la utilidad del consumo de una canasta de bienes depende de n bienes: $x_1, ..., x_n$, o la producción de una empresa depende de diferentes factores de producción como el stock de capital K o el trabajo L. Por ende, es de gran interés abordar el caso más general donde $f : \Omega \subset \mathbb{R}^n \to \mathbb{R}$.

[5] Ver apéndice sobre probabilidad e inferencia estadística.
[6] Recordemos que $f(x) > 0$ para todo $x \in \mathbb{R}$.
[7] La media: $X \sim N(\mu, \sigma^2)$, $\mathbb{E}[X] = \mu$.

2.2. Funciones de variable vectorial

El caso de funciones en varias variables es análogo al caso de una variable con la excepción que, la variable de entrada ya no es un número real x, si no, un vector $x = (x_1, ..., x_n)^T$. En este caso, buscamos resolver

$$\begin{aligned} &\text{opt } f(x) \\ &\text{s. a : } x \in \Omega \subset \mathbb{R}^n. \end{aligned}$$

A continuación haremos uso de la notación $\partial\Omega$ (donde Ω es un subconjunto de $\mathbb{R}^n$ y $\partial\Omega$ su borde) y argmax (y argmin). El lector que no conozca esta notación puede consultar Boyd and Vandenberghe (2004) o Chavez and Gallardo (2023).

Teorema 6. Si

$$x^* \in \text{argmax}_{x \in \Omega \subset \mathbb{R}^n} \ f(x), \ x^* \in \Omega / \partial\Omega$$

y f es diferenciable, entonces $\nabla f(x^*) = \left(\frac{\partial f(x^*)}{\partial x_1}, ..., \frac{\partial f(x^*)}{\partial x_n}\right)^T = 0$.

El resultado es análogo para el $x^* \in \text{argmin}_{x \in \Omega \subset \mathbb{R}^n} \ f(x)$.

Demostración. La prueba de este resultado, que generaliza el caso de funciones de variable real, hace uso de las aproximaciones de funciones de varias variables por su polinomio de Taylor. Veamos. Si x^* es un máximo local, existe una vecindad $\mathcal{V}$ de x^* tal que $\forall\ x \in \mathcal{V} \cap \Omega$,

$$f(x) \leq f(x^*).$$

La aproximación lineal de $f(x)$ en una vecindad del punto x^* es

$$f(x^*) + \nabla f(x^*)^T (x - x^*).$$

Pero entonces, como $\forall\ x \in V \cap \Omega \subset V$,

$$f(x^*) \geq f(x^*) + \nabla f(x^*)^T (x - x^*).$$

$$\nabla f(x^*)^T (x - x^*) \leq 0.$$

Luego, como $\nabla f(x^*)^T$ es la derivada en la dirección $x - x^*$, si existe x tal que

$$\nabla f(x^*)^T (x - x^*) < 0,$$

entonces $\nabla f(x^*)^T$ en la dirección opuesta es positiva, contradiciendo el hecho que x^* es un máximo local. Por ende,

$$\nabla f(x^*)^T = 0.$$

□

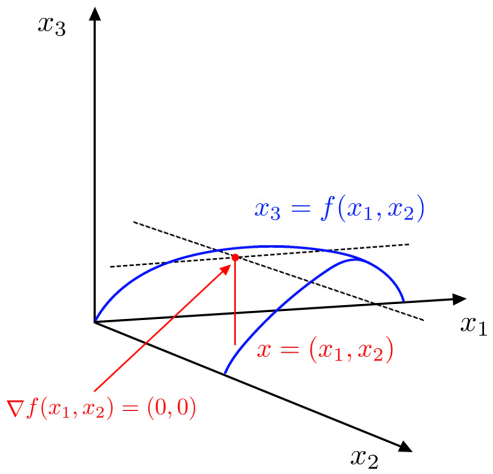

Figura 2.3 Punto rojo estacionario.

El razonamiento es análogo para un mínimo local y en la literatura, x^* tal que $\nabla f(x^*)^T = 0$, es conocido como punto

estacionario. Finalmente, permítanos enfatizar que, así como el Teorema 4, el Teorema 6 provee únicamente una condición necesaria más no suficiente. Es decir, bien podría tenerse $\nabla f(\tilde{x}) = 0$ sin que $\tilde{x}$ sea un óptimo[8].

Ejemplo 25. Sea $f(x_1, x_2) = x_1^2 + x_2^2$.

$$\nabla f(x_1, x_2) = (0, 0) \implies (x_1, x_2) = (0, 0).$$

Ahora bien, el punto $(0, 0)$, es en efecto un mínimo local (incluso global) como puede apreciarse en la siguiente figura. Esto puede deducirse también analíticamente de la expresión de $f(x_1, x_2)$. En efecto, $f(x_1, x_2) \geq 0$ sobre $\mathbb{R}^2$ y $f(0, 0) = 0$. Así, se verifica el Teorema 6.

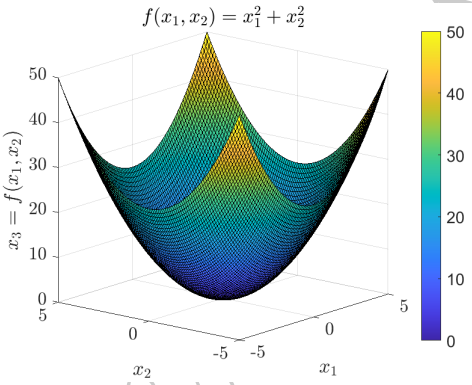

Figura 2.4 Paraboloide.

Ejemplo 26. Sea $f(x_1, x_2) = x_1^2 - x_2^2$. Por un lado, $\nabla f(0, 0) = (0, 0)$ implica que $x_1 = x_2 = 0$. Sin embargo, tal y como se aprecia en la siguiente figura, el punto $(0, 0)$ no es ni un mínimo ni un máximo local (en la literatura se le conoce como punto silla de hecho).

[8] Véase el Ejemplo 26.

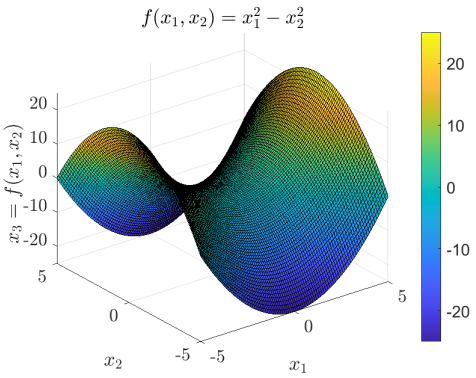

Figura 2.5 Punto silla.

Note que, para cualquier $\varepsilon > 0$, $f(0, \varepsilon) < 0 < f(\varepsilon, 0)$.

El Ejemplo 26 enfatiza que, como en el caso de funciones real valuadas, la condición $\nabla f(x) = 0$ no es suficiente para asegurar que x^* sea un óptimo. Es una condición necesaria. Más aún, es imposible determinar si se trata de un máximo o un mínimo. No obstante, el siguiente resultado permite discernir entre ambos casos y verificar que un punto estacionario es en efecto un máximo o mínimo local.

2.2.1. Condiciones de segundo orden

Definición 2.2.1. Sea $f : S \to \mathbb{R}$, $S \subset \mathbb{R}^n$ una función clase $C^2(S)$[9]. Definimos la matriz hessiana de f como

$$Hf = \begin{pmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial x_n x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{pmatrix}.$$

[9]Sus segundas derivadas parciales son continuas.

Teorema 7. Sea $A = Hf(x^*)$, la matriz Hessiana de una función f evaluada en un punto estacionario x^*, y sean $\lambda_1, ..., \lambda_n$ sus valores propios.[10] Entonces:

- Si $\lambda_i > 0$ para todo i, x^* es un mínimo local estricto.
- Si $\lambda_i < 0$ para todo i, x^* es un máximo local estricto.
- Si $\lambda_i \geq 0$ para todo i, $x*$ es un mínimo local.
- Si $\lambda_i \leq 0$ para todo i, x^* es un máximo local.
- Si existe $\lambda_i > 0$ y $\lambda_j < 0$, x^* es un punto silla.

Si $\lambda_i > 0$ para todo i, por el Teorema Espectral Axler (2015), para cualquier $x \in \mathbb{R}^n$, $x^T(Hf(x^*)x > 0$. El resultado es análogo en los otros casos. Diremos, respectivamente, que la matriz hessiana es positiva definida, negativa definida, positiva semidefinida, negativa semidefinida o indefinida, si $x^T Hf(x^*)x > 0$, $x^T Hf(x^*)x < 0$, $x^T H(x^*)x \geq 0$, $x^T Hf(x^*)x \leq 0$ o ninguno de los casos anteriores.

Usualmente, en la práctica, muchas de las funciones tienen como dominio $\Omega \subset \mathbb{R}^2$. En dicho caso,

$$Hf(x^*) = \begin{pmatrix} f_{11}(x^*) & f_{12}(x^*) \\ f_{21}(x^*) & f_{22}(x^*) \end{pmatrix}.$$

[10] Como la función es clase C^2, por el Teorema de Clairaut Tao (2016), la segundas derivadas parciales cruzadas son iguales. Es decir, $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$. Por ello, la matriz es simétrica. Aplicando el Teorema Espectral Axler (2015), se concluye que la matriz posee n valores propios diferentes. Para la definición de valores propios, consultar Chavez and Gallardo (2023) o Simon and Blume (1994).

Aplicando el caso general (Teorema 7) a esta situación, dónde $Hf(x^*) \in \mathcal{M}_{2\times 2}$:

- Si $f_{11}(x^*) \geq 0$, y $|Hf(x^*)| \geq 0$, x^* es un mínimo local.
- Si $f_{11}(x^*) > 0$, y $|Hf(x^*)| > 0$, x^* es un mínimo local estricto.
- Si $f_{11}(x) \leq 0$, y $|Hf(x^*)| \geq 0$, x^* es un máximo local.
- Si $f_{11}(x^*) < 0$, y $|Hf(x^*)| > 0$, x^* es un máximo local estricto.
- Si $|Hf(x^*)| < 0$ x^* es un punto silla.
- Si el determinante es cero, debemos efectuar una análisis ad-hoc.

Para la prueba, véase Chavez and Gallardo (2023) o Simon and Blume (1994).

Ejemplo 27. Considere la función $f(x_1, x_2) = x_1^2 + x_2^2 - 2x_1 - x_2 + 1$. Resolviendo

$$\nabla f(x_1, x_2) = \begin{pmatrix} 2x_1 - 2 \\ 2x_2 - 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix},$$

obtenemos los puntos estacionarios. En este caso, hallamos $(x_1^*, x_2^*) = (1, 1/2)$. Luego, la matriz hessiana de la función, en cualquier punto, es

$$Hf(x^*) = \begin{pmatrix} 2 & 0 \\ 0 & 2 \end{pmatrix}.$$

Como $f_{11}, f_{22}, |H| > 0$, $(1, 1/2)$ es un mínimo local.

Ejemplo 28. Sea ahora $f(x_1, x_2, x_3) = -2x_1^2 - 5x_2^2 - x_3^2 + 4x_1x_2 + 2x_2x_3 + 4$. Tenemos que,

$$\nabla f(x) = \nabla f(x_1, x_2, x_3) = \begin{pmatrix} -4x_1 + 4x_2 \\ -10x_2 + 4x_1 + 2x_3 \\ -2x_3 + 2x_2 \end{pmatrix}.$$

Resolviendo $\nabla f(x) = 0$, se encuentra que el único punto estacionario $x^* = 0$. La matriz hessiana en dicho punto es

$$Hf(0) = \begin{pmatrix} -4 & 4 & 0 \\ 4 & -10 & 2 \\ 0 & 2 & -2 \end{pmatrix}.$$

Como

$$p(\lambda) = -\lambda^3 - 16\lambda^2 - 48\lambda - 32$$

tiene tres raíces negativas, concluimos que x^* es un máximo local.

2.3. Lagrange y Karush-Kuhn-Tucker

Previamente, se ha enfatizado que, en un problema de optimización, la variable de optimización puede estar sujeta a una serie de restricciones[11]. En esta breve sección, presentamos resultados relacionados con el caso en el cual $\Omega = \{x \in \mathbb{R}^n : h(x) = a \in \mathbb{R}^m\}$ o $\Omega = \{x \in \mathbb{R}^n : g(x) \leq b \in \mathbb{R}^m\}$.

[11] Esto es lo que implícitamente se da a entender cuando se escribe $x \in \Omega$.

Teorema 8. Sean $f, h_1, \cdots, h_m \in C^1$ funciones de $\mathbb{R}^n$ en $\mathbb{R}$. Considere el siguiente problema de optimización:

$$\begin{aligned} &\text{máx } f(x) \\ &\text{s. a. } x \in C_h \end{aligned}$$

con

$$C_h = \{x \in \mathbb{R}^n : h_1(x) = a_1, \cdots, h_m(x) = a_m\}.$$

Suponga que $x^* \in C_h$ es un maximizador (o minimizador) local para el problema de optimización. Si $Dh(x^*) = J_x h(x^*) \in \mathcal{M}_{m\times n}$ es de rango completo[12], entonces existen $\nu_1^*, \cdots, \mu_m^*$ tales que (x^*, μ^*) son un punto crítico de

$$L(x, \mu) = f(x) - \sum_{i=1}^{m} \mu_i (h_i(x) - a_i).$$

A un problema como el del Teorema 8 se le denomina problema de Lagrange.

Teorema 9. Suponga que $f, g_1, \cdots, g_k \in C^1$ son funciones de $\mathbb{R}^n$ a $\mathbb{R}$. Suponga que $x^* \in \mathbb{R}^n$ es un maximizador local de f sujeto a

$$g_1(x) \leq b_1, \cdots, g_k(x) \leq b_k. \tag{2.1}$$

Por simplicidad suponga que las primeras k_0 restricciones en (2.1) se dan con igualdad. Entonces, si

$$\begin{bmatrix} \frac{\partial g_1}{\partial x_1}(x^*) & \cdots & \frac{\partial g_1}{\partial x_n}(x^*) \\ \vdots & \ddots & \vdots \\ \frac{\partial g_{k_0}}{\partial x_1}(x^*) & \cdots & \frac{\partial g_{k_0}}{\partial x_n}(x^*) \end{bmatrix}$$

[12] A esto se le conoce como condición de regularidad.

tiene rango completo, y definimos

$$L(x,\lambda) = f(x) - \sum_{j=1}^{k} \lambda_j (g_j(x) - b_j)$$

existen $\lambda_1^*, \cdots, \lambda_k^*$ tale que

$$\begin{aligned} \frac{\partial L}{\partial x_i}(x^*, \lambda^*) &= 0, \ \forall \ i = 1, ..., n \\ \lambda_j^*[g_j(x^*) - b_j] &= 0, \ \forall \ j = 1, ..., k \\ \lambda_j^* &\geq 0, \ \forall \ j = 1, ..., k \\ g_j(x^*) &\leq \lambda_j^*, \ \forall \ j = 1, ..., k. \end{aligned}$$

A un problema como el del Teorema 9 se le denomina problema de Karush-Kuhn-Tucker.

Ejemplo 29. En el problema de maximización de la utilidad, cuando $u : \mathbb{R}^n \to \mathbb{R}$ satisface las condiciones de Inada, el problema pasa de tener la forma de un problema de KKT

$$\begin{aligned} &\text{máx } u(x) \\ &\text{s. a. } p \cdot x \leq I \\ &\qquad x \geq 0 \end{aligned}$$

a tener la forma de un problema de Lagrange Chavez and Gallardo (2023)

$$\begin{aligned} &\text{máx } u(x) \\ &\text{s. a. } p \cdot x = I. \end{aligned}$$

Entonces, según el Teorema 8, si x^* resuelve el problema de maximización de la utilidad[13]

$$\frac{\partial u}{\partial x_i}(x^*) = \mu p_i$$
$$\sum_{i=1}^{n} p_i x_i^* = I.$$

Ejemplo 30. Si en el problema de maximización de la utilidad

$$u(x_1, ..., x_n; \theta) = \prod_{i=1}^{n} (x_i - a_i)^{\alpha_i}, \ \theta = (\alpha, a)^T \in \mathbb{R}^n \times \mathbb{R}^n, \quad (2.2)$$

con $a_i, \alpha_i > 0$ para todo i y $\sum_{i=1}^{n} \alpha_i = 1$, aplicando el Teorema 8, se deduce que

$$x_i^*(\theta, I) = a_i + \frac{\alpha_i}{p_i}\left[I - \sum_{i=1}^{n} p_i a_i\right].$$

A la función de utilidad de la Ecuación (2.2) se le conoce como Stone-Geary en honor a Richard Stone (Stone (1954)) y Roy Geary (Geary (1950)).

Una extensión de los dos tipos de problemas presentados en esta sección (Lagrange y Kuhn-Tucker) es el problema mixto. De manera similar a las condiciones de segundo orden abordadas previamente, existen condiciones de segundo orden para este tipo de problemas. Estas condiciones involucran lo que se conoce como Hessiano orlado. Véase Chavez and Gallardo (2023) o Simon and Blume (1994).

[13] Es fácil verificar la condición de regularidad.

2.4. Breve nota sobre la convexidad

La convexidad es una propiedad matemática ampliamente estudiante en matemáticas. Se estudia tanto la convexidad de los conjuntos como la convexidad de las funciones. En esta sección, brindamos algunas definiciones y teoremas centrales en la teoría de la optimización. Es importante mencionar que la teoría del análisis convexo se extiende al estudio de las funciones cuasi-convexas, cuasi-cóncavas, así como al estudio de los teoremas de separación, al Lema de Farkas etc. resultados ampliamente usados en teoría económica[14].

Definición 2.4.1. Decimos que un conjunto $X \subset \mathbb{R}^n$ es convexo si $\forall\ x, y \in X$ y $\theta \in [0,1]$,

$$\theta x + (1-\theta)y \in X.$$

Definición 2.4.2. Decimos que $f : X \subset \mathbb{R}$, con X convexo, es convexa si $\forall\ x, y \in X$ y $\theta \in [0,1]$

$$f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y).$$

Definición 2.4.3. Decimos que $f : X \subset \mathbb{R}$, con X convexo, es cóncava si $\forall\ x, y \in X$ y $\theta \in [0,1]$

$$\theta f(x) + (1-\theta)f(y) \leq f(\theta x + (1-\theta)y).$$

Ejemplo 31. El conjunto presupuestario del problema del consumidor

$$B(p, I) = \{x \in \mathbb{R}^n :\ p \cdot x \leq I\}$$

[14] Como por ejemplo teoría del consumidor, equilibrio general, teoría de contratos etc. Véase Mas-Colell et al. (1995).

con $p \in \mathbb{R}^n_{++}$ e $I > 0$, es convexo[15].

Ejemplo 32. La función norma Euclidiana $||\cdot||_2 : \mathbb{R}^n \to \mathbb{R}$,

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2}$$

es convexa.

Ejemplo 33. Las funciones $f(x) = x^{2k}$, $k \in \mathbb{N}$ y exponencial $g(x) = e^x$ son convexas sobre $\mathbb{R}$. Por otro lado, las funciones $h(x) = x^a$ con $a \in (0, 1)$ y $\ell(x) = \ln x$ son cóncavas sobre su dominio de definición.

Teorema 10. Sea $f : X \to \mathbb{R}$, con $X \subset \mathbb{R}^n$ convexo. Entonces, dados $x_1, ..., x_k \in X$ y $\theta_1, ..., \theta_k \geq 0$ tales que $\sum_{i=1}^{k} \theta_i = 1$, f es convexa si y solo si

$$f\left(\sum_{i=1}^{k} \theta_i x_i\right) \leq \sum_{i=1}^{k} \theta_i f(x_i). \tag{2.3}$$

Se tiene un resultado análogo para el caso de funciones cóncavas. A (2.3) se le conoce como desigualdad de Jensen[16].

Ejemplo 34. Usando la desigualdad de Jensen, es posible probar la desigualdad media-aritmética:

$$\prod_{i=1}^{n} x_i^{1/n} \leq \frac{1}{n} \sum_{i=1}^{n} x_i.$$

[15] Además, es compacto. Esto es, es acotado y cerrado. Puede considerar la bola con la norma $||x||_{\text{máx}} = \text{máx}_{1 \leq i \leq n} |x_i|$ y $B_{||\cdot||_{\text{máx}}}(0, 2I/p_{\text{mín}})$.

[16] Esta desigualdad será estudiada en el apéndice de teoría de la probabilidad en un contexto diferente.

En efecto, dada la concavidad de $\ln(\cdot)$,

$$\sum_{i=1}^{n} \frac{\ln x_i}{n} \leq \ln\left(\sum_{i=1}^{n} \frac{x_i}{n}\right).$$

Aplicando la función exponencial que es creciente y usando el hecho que $\ln\left(\prod_{i=1}^{n} x_n^{1/n}\right) = \frac{1}{n}\sum_{i=1}^{n} \ln x_i$, concluimos.

Usualmente, es bastante complicado determinar la convexidad o concavidad de una función a partir de su definición. Por ello, existen maneras alternativas de identificar cuando una función es convexa o cóncava, siempre y cuando la función en cuestión cumpla ciertas condiciones (continuidad, diferenciabilidad).

Teorema 11. Sea $f \in C^1(X)$ con $X \subset \mathbb{R}^n$ convexo y abierto. Entonces, es cóncava sobre X si y solo si para todo $x, y \in X$

$$f(y) - f(x) \leq \nabla f(x)(y - x).$$

Teorema 12. Sea $f \in C^2(X)$ con $X \subset \mathbb{R}^n$ convexo y abierto. Entonces f es cóncava si y solo si $Hf \leq 0$. Análogamente, f es convexa si y solo si $Hf \geq 0$.

Note que por medio del Teorema 12 es posible determinar la optimalidad de un punto estacionario por medio de la convexidad o concavidad de la función objetivo.

Esto concluye el breve repaso acerca de la teoría de la optimización. Lo que se ha presentado en este capítulo es una mera introducción. La teoría de la optimización se extiende a los problemas de optimización en otros espacios (superficies, variedades) y al estudio del análisis convexo (tanto en dimensión

finita como infinita). Por otro lado, una vez obtenida una solución $x^* = x^*(\alpha)$, donde α es un vector de parámetros[17] a un problema de optimización[18]. Para una presentación completa de estos temas, invitamos al lector consultar libros como Boyd and Vandenberghe (2004), Sundaram (1996), de la Fuente (2000), Lenberger and Ye (2021) o Chavez and Gallardo (2023). Aplicaciones sólidas de la teoría económica se encuentran en, por ejemplo, Mas-Colell et al. (1995).

[17]Por ejemplo, en el clásico problema de maximización de la utilidad, el vector de precios y el ingreso. En el caso de la minimización del costo, el nivel de producción requerido y el vector de precios de los insumos. Un análisis similar pero informal ya se hizo en el Capítulo 1 al introducir la regla de Cramer.

[18]En ese sentido, consultar por ejemplo el Teorema de la Envolvente.

Capítulo 3

Modelo multivariado

Muchas de las relaciones en economía plantean modelos determinísticos entre las variables, como por ejemplo

$$Y(K, L) = K^{\alpha} L^{\beta},$$

donde α y β son dos parámetros positivos, K el stock de capital, L el trabajo y Y la producción que se obtiene al emplear estos factores de producción en una economía. Si bien esta relación es bastante intuitiva, pues ciertamente la producción crece con el capital y con el trabajo ($Y_K, Y_L > 0$), ¿cómo saber que valores deben tomar los parámetros α y β? ¿Cuál es el rango de valores para estos parámetros? Esta pregunta es de gran de interés pues, determina propiedades como la concavidad o convexidad de la función

$$Y = Y(K, L).$$

A continuación, presentamos los fundamentos de una teoría establecida que permite responder preguntas complejas en el ámbito

de la econometría. En primera instancia, se definirá el modelo principal de este capítulo, junto con los supuestos necesarios para establecer los resultados fundamentales subyacentes. Este análisis se realizará mediante técnicas de álgebra matricial y optimización, justificando así la introducción previa de estos temas en los Capítulos 1 y 2. Finalmente, se ofrecerán ejemplos prácticos que consolidarán estos conceptos.

En este capítulo, se abordarán específicamente el modelo lineal k, la esperanza condicional, la interpretación geométrica, las regresiones particionadas y los momentos del estimador de Mínimos Cuadrados Ordinarios (MCO). Además, se analizará la bondad de ajuste, se discutirán aspectos relacionados con los intervalos de confianza y se mencionarán los supuestos del modelo, los cuales serán tratados en profundidad en capítulos posteriores.

3.1. El modelo $k-$lineal

Dada una variable aleatoria, esta puede descomponerse de la siguiente forma:

$$Y = \mathbb{E}[Y|X] + \varepsilon,$$

donde, recordemos $\mathbb{E}[Y|X] = \mathbb{E}[Y|\sigma(X)]$ y, debido a la ley de esperanzas iteradas, $\mathbb{E}[\varepsilon|X] = 0$. El modelo k-lineal plantea que

$$\mathbb{E}[Y|X] = X\beta = \beta_0 + \sum_{i=1}^{k} \beta_i X_i. \tag{3.1}$$

En la práctica, lo que se tiene es un conjunto de observaciones $\{Y_i\}_{1\leq i\leq n}$ y un conjunto de datos $\{X_{ji}\}_{1\leq i\leq n, 1\leq j\leq k}$. La variable

Y es una variable que buscamos predecir, mientras que X_j para $j = 1, ..., k$, es una variable «explicativa». En el modelo $k-$lineal, tal y como especifica la Ecuación 3.1, se adopta la siguiente forma funcional[1]

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \varepsilon_i, \ i = 1, ..., n. \quad (3.2)$$

Esta ecuación puede expresarse matricialmente como

$$\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{11} & X_{21} & \cdots & X_{k1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{1n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

$$Y_{n\times 1} = X_{n\times(k+1)} \ \beta_{(k+1)\times 1} + \varepsilon_{n\times 1}.$$

En este modelo, se asume que $\varepsilon \sim \mathcal{N}(0_n, \sigma^2 I_n)$.

Antes de continuar, permítanos hacer énfasis en algunos puntos. Matemáticamente, lo que tenemos es un conjunto de puntos en un espacio, usualmente $\mathbb{R}^p$, donde $X \in \mathcal{M}_{n\times k}$ representa el conjunto de observaciones de las variables predictoras, Y es el conjunto de observaciones de la variable dependiente y β es un vector de parámetros. El objetivo es determinar los parámetros.

Ahora bien, $Y_1, ..., Y_n$ es una muestra aleatoria. Es decir, cada Y_i es una variable aleatoria y observamos una realización $Y_i(\omega)$. Lo mismo aplica para los X_{ki}. Sin embargo, asumiremos que los regresores son determinísticos, por lo que $\mathbb{E}[Y|X] = \mathbb{E}[Y] = X\beta$, y todo lo estocástico queda almacenado en ε.

[1] La forma funcional es el término lineal, el error estocástico ε refleja la incertidumbre en el modelo.

Ejemplo 35. Supongamos que se busca estimar la producción en función de las variables clásicas del modelo de la demanda agregada

$$Y_i = \beta_0 + \beta_1 C_i + \beta_2 G_i + \beta_3 I_i + \beta_4 (X_i - M_i) + \varepsilon_i \ , \ i = 1, ..., n.$$

En esta ecuación, Y_i es la producción, C_i el consumo, G_i el gasto público, I_i la inversión y $X_i - M_i$ la balanza comercial. Estas últimas, son las variables explicativas en el modelo pues, en función de estas uno predice la variable dependiente. Luego, el índice i indica el elemento de la muestra considerado. Por ejemplo, cada i determina un país, un instante de tiempo (en dicho caso sería más apropiado denotar t en vez de i). El objetivo, es estimar los parámetros $\beta_0, \beta_1, \beta_2, \beta_3$ y β_4.

El Ejemplo 35 nos invita a reflexionar sobre el tipo de datos que se consideran en una regresión lineal. La siguiente definición es clave para distinguir los tipos de datos que aparecen en la práctica.

Definición 3.1.1. Identificamos tres tipos de datos:

- Datos transversales: se cuenta con solo una observación en el tiempo para diferentes variables. Por ejemplo, los datos recolectados en una encuesta $\{X_{1i}, X_{2i}, ..., X_{ki}\}_{1 \leq i \leq n}$.
- Series de tiempo: más de 2 observaciones en el tiempo: $X_0, X_1, ..., X_t, ..., X_T$. En estos casos tendríamos un modelo del tipo
$$Y_t = f(X_t, Z_t, \cdots, W_t) + \varepsilon_t.$$
- Datos panel o longitudinales: dos o más observaciones en el tiempo del mismo individuo, país $\{X_{it}\}_{1 \leq i \leq n, 1 \leq t \leq T}$.

3.1.1. Supuestos del modelo $k-$lineal

Enseguida, presentamos los supuestos del modelo (3.2). La hoja de ruta en los siguientes capítulos consiste justamente en levantar estos supuestos.

Teorema 13. En el modelo $k-$lineal se efectúan los siguientes supuestos:

- El modelo de regresión es lineal en los parámetros. Es decir, no se puede tener algo de la siguiente forma

$$Y_i = \beta_1^2 X_{1i} + \ln(\beta_2) X_{2i} + \varepsilon_i.$$

- La muestra es aleatoria. Es decir, la selección de los datos se lleva a cabo siguiendo metodologías específicas que buscan reducir el sesgo de selección y/o adaptarse a los objetivos del estudio.

- El valor esperado de los errores es igual a cero: $\mathbb{E}[\varepsilon_i] = 0, \forall\ i$. Más aún, como consecuencia

$$\mathbb{E}\left[\sum_{i=1}^{n} \varepsilon_i\right] = 0.$$

- Los errores tienen varianza constante: $\text{Var}(\varepsilon_i) = \sigma^2, \forall\ i$. Usualmente, el término de error está normalmente distribuido $\varepsilon_i \sim N(0, \sigma^2), \forall\ i$. Esto contempla los dos supuestos previos.

- No existe correlación entre las explicativas y los errores medidos: $X^T \varepsilon = 0$.

- No existe colinealidad perfecta entre las variables explicativas incluidas en el modelo:

$$\sum_{\ell=1}^{k} \gamma_\ell X_\ell = 0 \implies \gamma_\ell = 0, \ \forall \ \ell.$$

  Este supuesto es por ejemplo muy útil para asegurar más adelante la invertibilidad de $X^T X$.

- En el caso de series de tiempo, los errores no tienen correlación serial, es decir, no existe correlación entre los errores de diferentes periodos de tiempo $\text{Cov}(\varepsilon_t, \varepsilon_{t+k}) = \sigma_k = 0$. Para corte transversal, $\text{Cov}(\varepsilon_i, \varepsilon_j) = \sigma_{ij} = 0, \ \forall \ i \neq j$.

- El modelo está perfectamente identificado, esto es, el modelo incluye todas las variables explicativas relevantes. Así,

$$\mathbb{E}[\hat{\beta}] = \beta$$

  donde $\hat{\beta}$ es el vector de parámetros que se estima. Para asegurar esto, se realiza una revisión de literatura exhaustiva.

- El número de observaciones debe ser mayor al número de parámetros a estimar $n > k$.

Note que el supuesto de linealidad en los parámetros cubre casos como el siguiente

$$e^{Y_i} = \prod_{j=1}^{k} e^{\beta_j} e^{X_j} e^{\varepsilon_i},$$

pues, sacando logaritmos en ambos lados, se obtiene

$$\ln[e^{Y_i}] = \ln\left[\prod_{j=1}^{k} e^{\beta_j} e^{X_j} e^{\varepsilon_i}\right]$$
$$= \sum_{j=1}^{k} \beta_j \ln[X_j] + \varepsilon_i.$$

Por otro lado, los supuestos de no colinealidad perfecta y el hecho que $n > k$, aseguran que el rango de $X_{n\times k}$ sea k.

Ejemplo 36. Otro ejemplo de modelo de regresión lineal es el semi-logarítmico

$$\ln Y_t = X_t\beta + \delta t + \varepsilon_t.$$

Este modelo cumple la linealidad en los parámetros, que son β y δ. Note que el tiempo es un regresor. Por otro lado, el error es denotado ε_t en vez de ε_i pues las observaciones son en el tiempo. Finalmente, debe verificarse que $\mathbb{E}[\varepsilon_t] = 0$ y $\text{Var}(\varepsilon_t) = \sigma^2$, $\forall\ t$.

3.2. El problema de optimización

El método más frecuente usado para estimar los coeficientes de una regresión lineal ($k-$lineal) es el Método de Mínimos Cuadrados Ordinarios (MCO). ¿En qué consiste este método? Lo que buscamos es, dadas las observaciones, determinar un vector de parámetros que nos permita predecir la variable dependiente. Denotemos por $\hat{\beta}$ los parámetros que se obtienen luego de la estimación[2]. Entonces, si $X \in \mathbb{R}^k$ es un conjunto de variables explicativas (usadas en

[2] A continuación detallamos el origen del vector $\hat{\beta}$.

la estimación), la predicción para Y, dado dicho conjunto de información, es

$$\hat{Y} = X\hat{\beta}.$$

Cuando incorporamos una constante β_0, la primera componente de X es un 1.

Luego, la perturbación asociada a la $i-$ésima observación es igual a

$$\hat{\varepsilon}_i = Y_i - X_i^T\hat{\beta}.$$

Lo que se busca es que este error sea el más pequeño posible, para todo i. Por ello, el MCO propone minimizar la siguiente suma Greene (2015)

$$\sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - X_i^T\beta)^2 = (Y - X\beta)^T(Y - X\beta)$$

respecto al vector de parámetros. Este programa de optimización tiene una interpretación geométrica que discutiremos más adelante. Ahora bien, notemos que

$$\sum_{i=1}^{n} \varepsilon_i^2 = ||\varepsilon||_2^2.$$

Cabe la pregunta, ¿por qué no escoger $||\varepsilon||_p^p$ con $p \geq 1$ diferente de 2. Sucede que al escoger, por ejemplo $p = 1$[3], no se obtiene una solución analítica exacta al problema de optimización, dada la no diferenciabilidad del valor absoluto aparecen sub-diferenciales Boyd and Vandenberghe (2004) etc.[4]

[3] $||\varepsilon||_1 = \sum_{i=1}^{n} |\varepsilon_i|$

[4] Sin embargo, considerar $|\cdot|$ resulta ser un método insensible a outliers y se le conoce como Regresión Robusta Rau (2016).

De este modo, el problema de optimización que buscamos resolver, con la finalidad de encontrar los parámetros, es el siguiente:

$$\begin{aligned} \text{mín } Q(\beta) &= \sum_{i=1}^{n}(Y_i - X_i^T\beta) \\ \text{s.a } \beta &\in \mathbb{R}^{k+1}. \end{aligned}$$

Alternativamente, podemos escribir

$$\begin{aligned} \text{mín } Q(\beta) &= (Y - X\beta)^T(Y - X\beta) \\ \text{s.a } \beta &\in \mathbb{R}^{k+1}. \end{aligned}$$

Note que $\beta \in \mathbb{R}^{k+1}$ pues se incorpora la constante.

Teorema 14. La solución al problema de minimización es

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

Demostración. Expandiendo la función $Q(\beta)$ se tiene

$$\begin{aligned} Q(\beta) &= Y^TY - \beta^TX^TY - Y^TX\beta + \beta^TX^TX\beta \\ &= Y^TY - 2\beta^TX^TY + \beta^TX^TX\beta \\ &= Y^TY - 2Y^TX\beta + \beta^TX^TX\beta. \end{aligned}$$

Aquí hemos usado que $\beta^TX^TY = Y^TX\beta$, pues, ambos términos son escalares y la transpuesta de un escalar es el mismo escalar. Ahora bien, por las condiciones de primer orden

$$\frac{\partial Q(\beta)}{\partial \beta} = -2Y^TX + 2X^TX\beta = 0.$$

Luego,

$$X^TX\beta = X^TY.$$

Como X tiene rango completo, $\det(X^TX) \neq 0$. Así, puede invertirse y por ende[5]

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^TY \\ &= \left(\frac{1}{N}\sum_{i=1}^{N} X_i^TX_i\right)^{-1}\left(\sum_{i=1}^{N} X_i^TY_i\right)\end{aligned}$$

□

Definición 3.2.1. Una regresión lineal simple es una relación de la forma

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i,$$

siendo ε_i es un error aleatorio tal que $\varepsilon_i \sim N(0, \sigma^2)$. Este es un caso particular de (3.2), donde $k = 1$.

Ejemplo 37. En el caso de una regresión lineal simple, los estimadores obtenidos vía Mínimos Cuadrados Ordinarios se obtienen resolviendo el siguiente problema de optimización:

$$\text{mín} \sum_{i=1}^{n} \varepsilon_i^2 = Q(\beta_0, \beta_1) = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2,$$

la solución es

- $\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}$.
- $\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$.

En efecto, las condiciones de primer orden son

$$\begin{aligned}\frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_0} &= -2\sum_{i=1}^{n}(Y_i - \beta_0 - \beta_1 X_i) = 0 \\ \frac{\partial Q(\beta_0, \beta_1)}{\partial \beta_1} &= -2\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0.\end{aligned}$$

[5] El candidato a óptimo es justamente $\hat{\beta}$.

Despejando β_0 en la primera ecuación se tiene $\beta_0 = \overline{Y} - \beta_1 \overline{X}$. Ahora, de la segunda ecuación, se obtiene

$$\sum_{i=1}^{n} X_i(Y_i - \beta_0 - \beta_1 X_i) = 0.$$

Luego,

$$\begin{aligned} 0 &= \sum_{i=1}^{n} X_i Y_i - \beta_0 n\overline{X} - \beta_1 \sum_{i=1}^{n} X_i^2 \\ &= \sum_{i=1}^{n} X_i Y_i - (\overline{Y} - \beta_1 \overline{X}) n\overline{X} - \beta_1 \sum_{i=1}^{n} X_i^2 \\ &= \left(\sum_{i=1}^{n} X_i Y_i \right) - n\overline{YX} + \beta_1 n\overline{X}^2 - \beta_1 \sum_{i=1}^{n} X_i^2. \end{aligned}$$

Despejando, se obtienen

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^{n} (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^{n} (X_i - \overline{X})^2}$$

y

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Antes de pasar a las condiciones de segundo orden, proveemos una derivación alternativa del estimador de MCO desde la estadística y la ya anticipada interpretación geométrica.

Dado que $\mathbb{E}[\varepsilon|X] = 0$, se sigue a partir de

$$Y = X\beta + \varepsilon$$

y multiplicando por X^T en ambos lados

$$\begin{aligned} X^TY &= X^TX\beta + X^T\varepsilon \\ \mathbb{E}[X^TY] &= \mathbb{E}[X^TX\beta + X^T\varepsilon] \\ &= \mathbb{E}[X^TX]\beta + \underbrace{\mathbb{E}[X^T\varepsilon]}_{=0} \\ \beta &= \mathbb{E}[X^TX]^{-1}\mathbb{E}[X^TY]. \end{aligned}$$

Así pues, debido al principio de analogía Manski (1988)

$$\hat{\beta}_{MCO} = \left(\frac{1}{N}\sum_{i=1}^{N} X_i^TX_i\right)^{-1}\left(\sum_{i=1}^{N} X_i^TY_i\right).$$

Note que hemos usado las propiedades de la esperanza condicional (véase Apéndice A)

$$\begin{aligned} \int \mathbb{E}[\varepsilon|X] &= \int \varepsilon \\ \mathbb{E}[\mathbb{E}[\varepsilon|X]] &= \mathbb{E}[\varepsilon] \\ \mathbb{E}[X^T\varepsilon] &= \mathbb{E}[X^T\varepsilon|X] \\ &= X^T\mathbb{E}[\varepsilon|X] \\ &= 0. \end{aligned}$$

Respecto a la interpretación geométrica, notemos que

$$\begin{aligned} \hat{Y} &= X\hat{\beta} \\ &= X(X^TX)^{-1}X^TY \\ &= \mathrm{Proj}_X(Y) \end{aligned}$$

donde Proj_X es la matriz proyección del espacio generado vectorial

por las columnas de X. Por otro lado,

$$\begin{aligned}\hat{\varepsilon} &= Y - X\hat{\beta} \\ &= (I - X(X^TX)^{-1}X^T)Y \\ &= N_X Y\end{aligned}$$

donde N_X es la matriz proyección en el espacio nulo de las columnas de X. Note que las matrices proyección son idempotentes[6] y simétricas Axler (2015).

3.2.1. Condiciones de segundo orden

Recordemos que para asegurarnos que $\hat{\beta}$ se trata de un mínimo, es necesario verificar la condición de segundo orden. Esto es, verificar que, la matriz hessiana evaluada en el punto estacionario $\hat{\beta}$, es definida positiva, i.e., que dado cualquier vector $v \in \mathbb{R}^{k+1}$, $v^T(2X^TX)v \geq 0$.

Demostración.

$$\frac{\partial^2 Q(\beta)}{\partial\beta\partial\beta^T} = 2X^TX.$$

Luego, definiendo $z = Xv \in \mathbb{R}^n$,

$$\begin{aligned}v^T(2X^TX)v &= 2z^Tz \\ &= 2\sum_{i=1}^{n} z_i^2 \geq 0.\end{aligned}$$

□

[6] $A^2 = A$.

Ejemplo 38. En el caso de la regresión lineal simple, podemos calcular directamente la matriz hessiana de $Q(\beta_0, \beta_1)$ evaluada en el óptimo $(\hat{\beta}_0, \hat{\beta}_1)$:

$$HQ(\hat{\beta}_0, \hat{\beta}_1) = \begin{bmatrix} \dfrac{\partial^2 Q(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0^2} & \dfrac{\partial^2 Q(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0 \partial \beta_1} \\ \\ \dfrac{\partial^2 Q(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0 \partial \beta_1} & \dfrac{\partial^2 Q(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_1^2} \end{bmatrix} = \begin{pmatrix} 2n & 2n\overline{X} \\ 2n\overline{X} & 2\sum_{i=1}^n X_i^2 \end{pmatrix},$$

siendo

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Tenemos que verificar que $\dfrac{\partial^2 Q(\hat{\beta}_0, \hat{\beta}_1)}{\partial \beta_0^2}$ y $|HQ(\hat{\beta}_0, \hat{\beta}_1)|$ son positivos. Ciertamente

$$\frac{\partial^2 Q}{\partial \beta_0^2} = 2n > 0.$$

Finalmente,

$$|HQ| = 4n \left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2 \right] = 4n \left[\sum_{i=1}^{n} (X_i - \overline{X})^2 \right] \geq 0.$$

Con ello, concluimos que se cumple la condición de mínimo.

Ejemplo 39. Sean $\hat{\beta}_1$ y $\hat{\beta}_2$ el intercepto y la pendiente estimados, respectivamente, de la regresión de Y_i contra X_i para una muestra de n observaciones. Sean c_1 y c_2 dos constantes $(c_1, c_2 \neq 0)$, $\bar{\beta}_1$ y $\bar{\beta}_2$ el intercepto y la pendiente estimados, respectivamente, de la regresión $c_1 Y_i$ contra $c_2 X_i$. Buscamos una expresión de $\bar{\beta}_1$ y $\bar{\beta}_2$ en función de $\hat{\beta}_1$, $\hat{\beta}_2$ y las constantes c_1, c_2. Al plantearse el modelo

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

se estiman los parámetros $\hat{\beta}_1$ y $\hat{\beta}_2$. Ahora, nos interesamos en la regresión

$$c_1 Y_i = \beta_1 + \beta_2 c_2 X_i + \varepsilon_i.$$

Para obtener los parámetros $\overline{\beta}_1, \overline{\beta}_2$ via MCO, resolvemos el problema de minimización

$$\text{mín } Q(\overline{\beta}_1, \overline{\beta}_2) = \sum_{i=1}^{n} (c_1 Y_i - \overline{\beta}_1 - \overline{\beta}_2 c_2 X_i)^2.$$

Aplicando condiciones de primer orden, obtenemos

$$\frac{\partial Q}{\partial \overline{\beta}_1} = -2 \sum_{i=1}^{n} (c_1 Y_i - \overline{\beta}_1 - c_2 \overline{\beta}_2 X_i) = 0$$
$$\frac{\partial Q}{\partial \overline{\beta}_2} = -2 \sum_{i=1}^{n} c_2 X_i (c_1 Y_i - \overline{\beta}_1 - c_2 \overline{\beta}_2 X_i) = 0.$$

De la primera ecuación se obtiene la relación

$$c_1 \sum_{i=1}^{n} Y_i - n\overline{\beta}_1 - c_2 \overline{\beta}_2 \sum_{i=1}^{n} X_i = 0$$

y por ende

$$\overline{\beta}_1 = c_1 \overline{Y} - c_2 \overline{\beta}_2 \overline{X}.$$

Ahora, de la segunda condición de primer orden, y reemplazando con la expresión de $\overline{\beta}_1$

$$c_2 c_1 \sum_{i=1}^{n} X_i Y_i - (c_1 \overline{Y} - c_2 \overline{\beta}_2 \overline{X}) \sum_{i=1}^{n} c_2 X_i - c_2^2 \overline{\beta}_2 \sum_{i=1}^{n} X_i^2 = 0.$$

Desarrollando se tiene

$$c_1 c_2 \sum_{i=1}^{n} X_i Y_i - c_1 c_2 n \overline{YX} + c_2^2 \overline{\beta}_2 n \overline{X}^2 - c_2^2 \overline{\beta}_2 \sum_{i=1}^{n} X_i^2 = 0.$$

Luego, despejando $\overline{\beta}_2$

$$\begin{aligned}\overline{\beta}_2 &= \frac{c_1c_2n\overline{YX} - c_1c_2\sum_{i=1}^n X_iY_i}{c^2n\overline{X}^2 - c_2^2\sum_{i=1}^n X_i^2} \\ &= \frac{c_1c_2\left(\sum_{i=1}^n X_iY_i - n\overline{X}\cdot\overline{Y}\right)}{c_2^2\left(\sum_{i=1}^n X_i^2 - n\overline{X}^2\right)} \\ &= \frac{c_1}{c_2}\hat{\beta}_2,\end{aligned}$$

donde

$$\hat{\beta}_2 = \frac{\sum_{i=1}^n Y_iX_i - n\overline{YX}}{\sum_{i=1}^n X_i - n\overline{X}^2}.$$

Así pues,

$$\begin{aligned}\overline{\beta}_1 &= c_1\overline{Y} - c_2\overline{\beta}_2\overline{X} \\ &= c_1\overline{Y} - c_2\frac{c_1}{c_2}\hat{\beta}_2\overline{X} \\ &= c_1(\overline{Y} - \hat{\beta}_2\overline{X}) \\ &= c_1\hat{\beta}_1.\end{aligned}$$

Concluimos entonces que $\overline{\beta}_1 = c_1\hat{\beta}_1$ y $\overline{\beta}_2 = \frac{c_1}{c_2}\hat{\beta}_2$. Note que la convexidad del paraboloide (función objetivo) asegura que se trata de un mínimo.

Ejemplo 40. De manera similar al ejemplo anterior, sean ahora $\bar{\beta}_1$ y $\bar{\beta}_2$ el intercepto y la pendiente estimados, respectivamente, de la regresión $Y_i + c_1$ contra $X_i + c_2$. Obtengamos una expresión de $\bar{\beta}_1$ y $\bar{\beta}_2$ en función de $\hat{\beta}_1, \hat{\beta}_2$ y las constantes c_i. En este caso, la especificación del modelo es la siguiente

$$Y_i + c_1 = \beta_1 + \beta_2(X_i + c_2) + \varepsilon_i.$$

Procedemos análogamente. Buscamos minimizar

$$Q(\overline{\beta}_1, \overline{\beta}_2) = \sum_{i=1}^{n} (Y_i + c_1 - \overline{\beta}_1 - \overline{\beta}_2 X_i - \overline{\beta}_2 c_2)^2.$$

Las condiciones de primer orden nos dan

$$\frac{\partial Q}{\partial \overline{\beta}_1} = -2 \sum_{i=1}^{n} (Y_i + c_1 - \overline{\beta}_1 - \overline{\beta}_2 X_i - \overline{\beta}_2 c_2) = 0$$
$$\frac{\partial Q}{\partial \overline{\beta}_2} = -2 \sum_{i=1}^{n} (X_i + c_2)(Y_i + c_1 - \overline{\beta}_1 - \overline{\beta}_2 X_i - \overline{\beta}_2 c_2) = 0.$$

De la primera ecuación, despejando para $\overline{\beta}_1$ se obtiene

$$\overline{\beta}_1 = \overline{Y} + c_1 - \overline{\beta}_2 \overline{X} - \overline{\beta}_2 c_2.$$

Ahora, de la segunda ecuación, reemplazando con la expresión de $\overline{\beta}_1$, se tiene

$$\sum_{i=1}^{n} (X_i + c_2)(Y_i + c_1 - (\overline{Y} + c_1 - \overline{\beta}_2 \overline{X} - \overline{\beta}_2 c_2) - \overline{\beta}_2 X_i - \overline{\beta}_2 c_2) = 0.$$

Simplificando

$$\sum_{i=1}^{n} (X_i + c_2)(Y_i + c_1 - \overline{Y} - c_1 + \overline{\beta}_2 c_2 - \overline{\beta}_2 X_i - \overline{\beta}_2 c_2) = 0,$$

se llega a la siguiente expresión

$$\sum_{i=1}^{n} (X_i + c_2)(Y_i - \overline{Y} + \overline{\beta}_2 \overline{X} - \overline{\beta}_2 X_i) = 0.$$

Desarrollando obtenemos

$$\sum_{i=1}^{n} (X_i Y_i - X_i \overline{Y} + \overline{\beta}_2 \overline{X} X_i - \overline{\beta}_2 X_i^2 + c_2 Y_i - c_2 \overline{Y} + \overline{\beta}_2 c_2 \overline{X} - \overline{\beta}_2 c_2 X_i) = 0.$$

Aplicando la suma a cada término

$$\sum_{i=1}^{n} X_i Y_i - n\overline{X}\cdot\overline{Y} + \overline{\beta}_2 n\overline{X}^2 - \overline{\beta}_2 \sum_{i=1}^{n} X_i^2 + c_2 n\overline{Y} - n\overline{Y}c_2 + \overline{\beta}_2 c_2 n\overline{X} - \overline{\beta}_2 c_2 n\overline{X}.$$

Simplificando y despejando para $\overline{\beta}_2$ se llega a

$$\overline{\beta}_2 = \frac{\sum_{i=1}^{n} X_i Y_i - n\overline{XY}}{\sum_{i=1}^{n} X_i^2 - n\overline{X}^2} = \hat{\beta}_2.$$

Así,

$$\begin{aligned} \overline{\beta}_1 &= \overline{Y} + c_1 - \overline{\beta}_2\overline{X} - \overline{\beta}_2 c_2 \\ &= \overline{Y} + c_1 - \hat{\beta}_2\overline{X} - \hat{\beta}_2 c_2 \\ &= c_1 + \hat{\beta}_1 - \hat{\beta}_2 c_2. \end{aligned}$$

3.3. Análisis de los parámetros

Previamente ya se ha abordado el problema de la estimación de parámetros desde un enfoque puramente algebraico y siguiendo el método propuesto por MCO. Sin embargo, no nos hemos preguntado si este método es el más adecuado, o si existen otros métodos. Más aún, es de interés conocer las diferentes propiedades de los parámetros estimados $\hat{\beta}$ (estimadores). En ese sentido, vamos a estudiar en la presente sección, la varianza y el valor esperado de los parámetros estimados. Esto va a permitirnos introducir el Teorema de Gauss Markov, el cual explica el interés de la estimación de los parámetros usando este método.

3.3.1. Insesgadez de los parámetros

Recordemos que si $A \in \mathcal{M}_{m\times n}$ y $x : \Omega \to \mathbb{R}^n$,

$$\mathbb{E}[Ax] = A\mathbb{E}[x]$$
$$\mathrm{Var}(Ax) = A\mathrm{Var}(x)A^T.$$

Recordemos que los parámetros estimados tienen la siguiente forma

$$\hat{\beta} = (X^TX)^{-1}X^TY.$$

Luego, usando que $Y = (X\beta + \varepsilon)$,

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^TY \\ &= (X^TX)^{-1}X^T(X\beta+\varepsilon) \\ &= (X^TX)^{-1}X^T(X\beta+\varepsilon) \\ &= (X^TX)^{-1}X^TX\beta + (X^TX)^{-1}X^T\varepsilon \\ &= \beta + (X^TX)^{-1}X^T\varepsilon.\end{aligned}$$

Teorema 15. Si $\hat{\beta}$ es el vector de parámetros estimados vía MCO, entonces

$$\mathbb{E}[\hat{\beta}] = \beta.$$

Demostración. Usando el resultado anterior, aplicando las propiedades del valor esperado y usando que $\mathbb{E}[\varepsilon] = 0$, calculamos

$$\begin{aligned}\mathbb{E}[\hat{\beta}] &= \mathbb{E}[\beta + (X^TX)^{-1}X^T\varepsilon] \\ &= \mathbb{E}[\beta] + \mathbb{E}[(X^TX)^{-1}X^T\varepsilon] \\ &= \beta + (X^TX)^{-1}X^T\mathbb{E}[\varepsilon] \\ &= \beta + (X^TX)^{-1}X^T0 \\ &= \beta.\end{aligned}$$

□

Ejemplo 41. Recordemos cuales son los parámetros estimados en el caso del modelo bivariado

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n X_i Y_i - n\overline{XY}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2} = \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2}$$
$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}.$$

Veamos que $\mathbb{E}[\hat{\beta}_0] = \beta_0$ y $\mathbb{E}[\hat{\beta}_1] = \beta_1$

$$\begin{aligned}
\hat{\beta}_1 &= \frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n [(X_i - \overline{X})Y_i - (X_i - \overline{X})\overline{Y}]}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})Y_i}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \frac{\sum_{i=1}^n (X_i - \overline{X})(\beta_0 + \beta_1 X_i + u_i)}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \beta_0 \frac{\sum_{i=1}^n (X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} + \beta_1 \frac{\sum_{i=1}^n (X_i - \overline{X})X_i}{\sum_{i=1}^n (X_i - \overline{X})^2} + \frac{\sum_{i=1}^n (X_i - \overline{X})u_i}{\sum_{i=1}^n (X_i - \overline{X})^2}.
\end{aligned}$$

Luego, aplicando la linealidad del valor esperado,

$$\begin{aligned}
\mathbb{E}[\hat{\beta}_1] &= \mathbb{E}\left[\beta_1 \frac{\sum_{i=1}^n (X_i - \overline{X})X_i}{\sum_{i=1}^n (X_i - \overline{X})^2} + \frac{\sum_{i=1}^n (X_i - \overline{X})u_i}{\sum_{i=1}^n (X_i - \overline{X})^2}\right] \\
&= \beta_1 \frac{\sum_{i=1}^n \mathbb{E}[(X_i - \overline{X})X_i]}{\sum_{i=1}^n (X_i - \overline{X})^2} + \frac{\sum_{i=1}^n \mathbb{E}[X_i - \overline{X}]\cancel{\mathbb{E}[u_i]}}{\sum_{i=1}^n (X_i - \overline{X})^2} \\
&= \beta_1.
\end{aligned}$$

Finalmente,

$$\mathbb{E}[\hat{\beta}_0] = \mathbb{E}[\overline{Y} - \hat{\beta}_1 \overline{X}] = \overline{Y} - \beta_1 \overline{X} = \beta_0.$$

A esta propiedad, i.e., $\mathbb{E}[\hat{\beta}] = \boldsymbol{\beta}$ se conoce como insesgadez de los parámetros.

Ejemplo 42. Considere dos estimadores, $\hat{\beta}$ y $\bar{\beta}$, construidos para estimar el parámetro poblacional β. El primer estimador $\hat{\beta}$ es el de Mínimos Cuadrados Ordinarios (que cumple con todos los supuestos) y el segundo es otro estimador lineal e insesgado. Luego, se construye un tercer estimador β^* que es una combinación convexa de $\hat{\beta}$ y $\bar{\beta}$; es decir, $\beta^* = \delta\hat{\beta} + (1-\delta)\,\bar{\beta}$, con $\delta \in [0,1]$. Este estimador sigue siendo insesgado. En efecto

$$\begin{aligned}\mathbb{E}[\beta^*] &= \mathbb{E}[\delta\hat{\beta} + (1-\delta)\,\bar{\beta}] \\ &= \mathbb{E}[\delta\hat{\beta}] + \mathbb{E}[(1-\delta)\,\bar{\beta}] \\ &= \delta\mathbb{E}[\hat{\beta}] + (1-\delta)\mathbb{E}[\bar{\beta}].\end{aligned}$$

Como el estimador MCO es insesgado, así como el estimador lineal $\bar{\beta}$, $\mathbb{E}[\hat{\beta}] = \beta$ y $\mathbb{E}[\bar{\beta}] = \beta$. Así

$$\delta\mathbb{E}[\hat{\beta}] + (1-\delta)\mathbb{E}[\bar{\beta}] = \delta\beta + (1-\delta)\beta = \beta.$$

Concluimos que β^* es insesgado. Más aún, esto nos permite concluir que la combinación convexa de estimadores insesgados siempre provee un estimador insesgado[7].

3.3.2. Varianza de los parámetros estimados

Ya habiendo estudiado el valor esperado de los parámetros estimados por MCO, nos interesamos en la varianza de dichos parámetros.

[7] A nivel de conjunto, el conjunto de estimadores lineales insesgados es convexo.

Teorema 16. Si $\hat{\beta}$ es el vector de parámetros estimados por MCO, entonces

$$\mathrm{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1}.$$

Demostración.

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= \mathbb{E}\left[(\hat{\beta}-\beta)(\hat{\beta}-\beta)^T\right] \\
&= \mathbb{E}\left[(\beta + (X^T X)^{-1} X^T \varepsilon - \beta)(\beta + (X^T X)^{-1} X^T \varepsilon - \beta)^T\right] \\
&= \mathbb{E}\left[((X^T X)^{-1} X^T \varepsilon)((X^T X)^{-1} X^T \varepsilon)^T\right] \\
&= \mathbb{E}[(X^T X)^{-1} X^T \varepsilon \varepsilon^T X (X^T X)^{-1}] \\
&= (X^T X)^{-1} X^T \mathbb{E}[\varepsilon \varepsilon^T] X (X^T X)^{-1}.
\end{aligned}$$

Como

$$\mathbb{E}[\varepsilon \varepsilon^T] = \mathrm{Var}(\varepsilon) - \mathbb{E}[\varepsilon]\mathbb{E}[\varepsilon^T] = \mathrm{Var}(\varepsilon),$$

$$\begin{aligned}
\mathrm{Var}(\hat{\beta}) &= (X^T X)^{-1} X^T \sigma^2 I X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} X^T X (X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} (X^T X)(X^T X)^{-1} \\
&= \sigma^2 (X^T X)^{-1} I \\
&= \sigma^2 (X^T X)^{-1}.
\end{aligned}$$

□

Ejemplo 43. En el caso de la regresión lineal simple,

$$\mathrm{Var}(\hat{\beta}_0) = \sigma^2 \frac{\sum_{i=1}^n X_i^2}{n \sum_{i=1}^n (X_i - \overline{X})^2}\ ,\ \mathrm{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \overline{X})^2}.$$

Note que el parámetro σ^2 sigue siendo a priori desconocido.

Teorema 17. Se cumple que

$$\hat{\sigma}^2 = \frac{1}{n-k}\sum_{i=1}^{n}\hat{\varepsilon}_i^2.$$

Demostración. A partir de

$$\hat{\varepsilon}^T\hat{\varepsilon} = (n-k)\sigma^2,$$

y ajustando por el número de grados de libertad,

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k}$$

$$\begin{aligned}\mathbb{E}[\hat{\sigma}^2] &= \mathbb{E}\left[\frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k}\right] \\ &= \frac{1}{n-k}\mathbb{E}[\hat{\varepsilon}^T\hat{\varepsilon}] \\ &= \sigma^2.\end{aligned}$$

□

Teorema 18. Para la estimación por MCO de $Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$ para muestras grandes[8] o v.a. normales, se cumple que

- $\mathbb{E}[Y_i] = \beta_0 + \beta_1 X_i$, $\text{Var}(Y_i) = \sigma^2 = \frac{\hat{\sigma}^2}{n-2}$. Así,

$$Y_i \sim \mathcal{N}\left(\beta_0 + \beta_1 X_i, \frac{\hat{\sigma}^2}{n-2}\right).$$

- $\mathbb{E}[\hat{\beta}_0] = \beta_0$, $\text{Var}(\hat{\beta}_0) = \frac{\hat{\sigma}^2\sum_{i=1}^n X_i}{n\sum_{i=1}^n (X_i-\overline{X})^2}$. Así,

$$\hat{\beta}_0 \sim \mathcal{N}\left(\beta_0, \frac{\hat{\sigma}^2\sum_{i=1}^n X_i^2}{n\sum_{i=1}^n (X_i-\overline{X})^2}\right).$$

[8] Gracias al Teorema del Límite Central Casella and Berger (2002).

- $\mathbb{E}[\hat{\beta}_1] = \beta_1$, $\text{Var}(\hat{\beta}_1) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}$. Así,

$$\hat{\beta}_1 \sim \mathcal{N}\left(\beta_1, \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \overline{X})^2}\right).$$

Demostración. Se sigue de los teoremas 15 y 16. □

Ejemplo 44. Considere el siguiente modelo

$$Y_i = a_1 X_{1i} + a_2 X_{2i} + a_3 X_{3i} + \varepsilon_i$$

donde

$$X = \begin{pmatrix} 2 & 4 & 1 \\ 0 & -1 & 0 \\ 0 & 0 & 0 \\ -1 & 0 & 0 \\ -1 & 3 & 1 \\ 1 & -2 & 0 \\ 3 & -4 & 1 \\ -2 & 2 & 1 \\ 6 & 1 & 0 \end{pmatrix}, \quad Y = \begin{pmatrix} 6 \\ 0 \\ -1 \\ -4 \\ -1 \\ 3 \\ 1 \\ -3 \\ 2 \end{pmatrix}.$$

Con esta información, podemos obtener los parámetros estimados y analizar las diferentes propiedades exhibidas previamente. Para calcular los coeficientes estimados, usamos la fórmula

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

Usando un software de cálculo, obtenemos

$$X^T_{3\times 9} X_{9\times 3} = \begin{pmatrix} 56 & -7 & 2 \\ -7 & 51 & 5 \\ 2 & 5 & 4 \end{pmatrix}.$$

Luego

$$X_{3\times 9}^T Y_{9\times 1} = \begin{pmatrix} 41 \\ 7 \\ 3 \end{pmatrix}.$$

Así, finalmente,

$$\hat{\beta} = \begin{pmatrix} 56 & -7 & 2 \\ -7 & 51 & 5 \\ 2 & 5 & 4 \end{pmatrix}^{-1} \begin{pmatrix} 41 \\ 7 \\ 3 \end{pmatrix} = \begin{pmatrix} 0{,}7585 \\ 0{,}2337 \\ 0{,}0787 \end{pmatrix}.$$

La varianza de los errores está dada por

$$\hat{\sigma}^2 = \frac{\hat{\varepsilon}^T \cdot \hat{\varepsilon}}{n-k}$$

con n el número de observaciones y k el número de restricciones. Acá $n = 9$ y $k = 3$. Por ende,

$$\begin{aligned} \hat{\varepsilon}^T \cdot \hat{\varepsilon} &= (Y - \hat{Y})^T (Y - \hat{Y}) \\ &= (Y - X\hat{\beta})^T (Y - X\hat{\beta}) \\ &= \sum_{i=1}^{9} \hat{\varepsilon}_i^2 = 44{,}0283. \end{aligned}$$

Así,

$$\hat{\sigma}^2 = \frac{44{,}0283}{9-3} = 7{,}3380.$$

Finalmente, con todos estos datos, podemos obtener la matriz de varianzas y covarianzas del vector de parámetros :

$$\begin{aligned} \mathrm{Var}(\hat{\beta}) &= \hat{\sigma}^2 (X^T X)^{-1} \\ &= \begin{pmatrix} 0{,}1385 & 0{,}0294 & -0{,}1060 \\ 0{,}0294 & 0{,}1702 & -0{,}2275 \\ -0{,}1060 & -0{,}2275 & 2{,}1719 \end{pmatrix}. \end{aligned}$$

3.3.3. Teorema de Gauss-Markov

El siguiente resultado es de gran importancia desde un punto de vista teórico pues, explica las ventajas del uso de la estimación vía MCO. Primero requerimos de la siguiente definición.

Definición 3.3.1. Decimos que $\hat{\theta}$ es el MEL (Mejor Estimador Lineal Insesgado) de θ si

- $\hat{\theta}$ es insesgado, i.e., $\mathbb{E}[\hat{\theta}] = \theta$.
- $\hat{\theta} = \sum_{i=1}^{n} c_i X_i$ (lineal).
- $\hat{\theta}$ es el estimador más eficiente entre todos los estimadores lineales e insesgados que existen de θ (minimiza la varianza).

El MELI $\hat{\theta}_{MELI}$ resuelve entonces

$$\mathcal{P} : \begin{cases} \text{mín} & \sum_{i=1}^{n} c_i^2 \text{Var}[X_i] \\ \text{s.a :} & \sum_{i=1}^{n} c_i \mathbb{E}[X_i] = \theta. \end{cases}$$

Teorema 19. Teorema de Gauss Markov. En el modelo de regresión lineal y bajo el cumplimiento de todos los supuestos; entre ellos recordemos

$$\mathbb{E}[u_i] = 0$$
$$\text{Var}(u_i) = \sigma^2 < \infty$$
$$\text{Cov}(u_i, u_j) = 0, \ \forall \ i \neq j,$$

el estimador $\hat{\beta}$ obtenido a través de MCO es el Mejor Estimador Lineal Insesgado (MELI).

Demostración. Asumamos por contradicción que existe otro estimador lineal, que denotaremos $\tilde{\beta}$, lineal e insesgado, cuya varianza es mínima. Este es entonces de la siguiente forma

$$\tilde{\beta} = AY,\ A = (X^T X)^{-1} X^T + C^T,$$

siendo $C \in \mathcal{M}_{n\times k}$, con al menos una entrada diferente de cero. De este modo,

$$\tilde{\beta} = (X^T X)^{-1} X^T Y + C^T Y.$$

Luego, usando la expresión para $Y = X\beta + \varepsilon$, se tiene

$$\begin{aligned}\tilde{\beta} &= (X^T X)^{-1} X^T [X\beta + \varepsilon] + C^T [X\beta + \varepsilon] \\ &= (X^T X)^{-1} X^T X\beta + (X^T X)^{-1} X\varepsilon + C^T X\varepsilon + C^T \varepsilon \\ &= \beta + (X^T X)^{-1} X\varepsilon + C^T X\varepsilon + C^T \varepsilon.\end{aligned}$$

Luego, tomando el valor esperado, como $\tilde{\beta}$ es insesgado,

$$\begin{aligned}\mathbb{E}[\tilde{\beta}] &= \beta + (X^T X)^{-1} X^T \mathbb{E}[\varepsilon] + C^T X\beta + C^T \mathbb{E}[\varepsilon] \\ &= \beta + C^T X\beta = \beta.\end{aligned}$$

Esto implica que $C^T X = 0_{k\times k}$. Luego, calculamos la varianza

$$\mathrm{Var}(\tilde{\beta}) = \mathbb{E}[(\tilde{\beta} - \mathbb{E}[\tilde{\beta}])(\tilde{\beta} - \mathbb{E}[\tilde{\beta}])^T].$$

Usando que

$$(X^T X)^{-1} X^T (X\beta + \varepsilon) + C^T (X\beta + \varepsilon) - \beta = (X^T X)^{-1} X^T \varepsilon + C^T \varepsilon$$

se tiene

$$\begin{aligned}\mathrm{Var}(\tilde{\beta}) &= \mathbb{E}[((X^TX)^{-1}X^T\varepsilon + C^T\varepsilon)((X^TX)^{-1}X^T\varepsilon + C^T\varepsilon)^T] \\ &= \mathbb{E}[((X^TX)^{-1}X^T\varepsilon + C^T\varepsilon)(\varepsilon^TX(X^TX)^{-1} + \varepsilon^TC)] \\ &= (X^TX)^{-1}X^T\mathbb{E}[\varepsilon\varepsilon^T]X(X^TX)^{-1} + (X^TX)^{-1}X^T\mathbb{E}[\varepsilon\varepsilon^T]C \\ &+ C^T\mathbb{E}[\varepsilon\varepsilon^T]X(X^TX)^{-1} + C^T\mathbb{E}[\varepsilon\varepsilon^T]C.\end{aligned}$$

Usando la igualdad $\mathbb{E}[\varepsilon\varepsilon^T] = \sigma^2 I$,

$$\begin{aligned}\mathrm{Var}(\tilde{\beta}) &= (X^TX)^{-1}X^T\sigma^2 IX(X^TX)^{-1} \\ &+ (X^TX)^{-1}X^T\sigma^2\mathbb{I}C + C^T\sigma^2\mathbb{I}X(X^TX)^{-1} + C^T\sigma^2\mathbb{I}C. \\ &= \sigma^2(X^TX)^{-1}X^TX(X^TX)^{-1} + \sigma^2(X^TX)^{-1}(C^TX)^T \\ &+ \sigma^2C^TC(X^TX)^{-1} + \sigma^2C^TC.\end{aligned}$$

Luego, nuevamente, como $C^TX = 0_{k\times k}$, la expresión previa se simplifica, quedando

$$\mathrm{Var}(\tilde{\beta}) = \sigma^2[(X^TX)^{-1} + C^TC].$$

Como $\mathrm{Var}(\hat{\beta}) = \sigma^2(X^TX)^{-1}$, para que $\tilde{\beta}$ tenga menor varianza, el término σ^2C^TC tiene que ser negativo. Sin embargo, esto no es posible. Por ende,

$$\mathrm{Var}(\hat{\beta}) < \mathrm{Var}(\tilde{\beta}).$$

□

Teniendo ya el grueso de los fundamentos teóricos de la estimación vía MCO, podemos proceder al análisis de los resultados subyacentes a este método de estimación. En particular, vamos a analizar los residuos observados y la interpretación de los parámetros estimados.

3.4. Interpretaciones

3.4.1. Indicadores de ajuste global

Definición 3.4.1. Dadas las observaciones de la variable que quiere ser predecida Y_i, definimos la Suma de Cuadrados Totales como

$$SCT = \sum_{i=1}^{n} (Y_i - \overline{Y})^2.$$

Definición 3.4.2. Análogamente dados los valores de la variable de interés predecida $\hat{Y}_i$, definimos la Suma de Cuadrados Explicativos como

$$SCE = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2.$$

Definición 3.4.3. Finalmente, dados los valores de la variable de interés predecida, $\hat{Y}_i$, y los valores originales de la variable de interés Y_i, definimos la Suma de Cuadrados Residuales como

$$SCR = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Teorema 20. Se cumple

$$SCT = SCE + SCR \tag{3.3}$$
$$\sum_{i=1}^{n} (Y_i - \overline{Y})^2 = \sum_{i=1}^{n} (\hat{Y}_i - \overline{Y})^2 + \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2.$$

Demostración. Partiendo de las definiciones,

$$\begin{aligned}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 &= \sum_{i=1}^{n}(Y_i-\hat{Y}_i+\hat{Y}_i-\overline{Y})^2\\ &= \sum_{i=1}^{n}(\hat{\varepsilon}_i+(\hat{Y}_i-\overline{Y}))^2\\ &= \sum_{i=1}^{n}\hat{\varepsilon}_i^2+2\underbrace{\sum_{i=1}^{n}(Y_i-\hat{Y}_i)(\hat{Y}_i-\overline{Y})}_{=0}+\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2\\ &= \sum_{i=1}^{n}\hat{\varepsilon}_i^2+\sum_{i=1}^{n}(\hat{Y}_i-\overline{Y})^2\\ &= SCR+SCE.\end{aligned}$$

□

Definición 3.4.4. Usando las definiciones (3.4.1), (3.4.2) y (3.4.3), el R^2 se establece mediante la siguiente ecuación,

$$R^2=\frac{SCE}{SCT}=1-\frac{SCR}{SCT}. \tag{3.4}$$

En otras palabras, el coeficiente de determinación R^2 es el ratio entre la variación explicada y la variación total, es decir, qué proporción de la variación de la dependiente es explicada por la(s) variable(s) explicativa(s).

Note que la ecuación (3.4) se deduce de la igualdad (3.3). En efecto

$$\begin{aligned}SCT &= SCE+SCR\\ \frac{SCT}{SCT} &= \frac{SCE}{SCT}+\frac{SCR}{SCT}\\ 1-\frac{SCR}{SCT} &= \frac{SCE}{SCT}=R^2.\end{aligned}$$

De este modo, siguiendo la definición, si por ejemplo $R^2 = 0{,}48$, entonces, se está explicando el 48 % de la variabilidad de la variable dependiente en el conjunto de datos. Ciertamente, mientras mayor sea el R^2, menor es la fracción SCE/SCT, por lo que el ajuste es mejor. En efecto, si $SCE/SCT \to 0$, $SCT \gg SCE$, i.e., la variabilidad de la predicción es considerablemente inferior a la variabilidad original de los datos, respecto a la media muestral. Por otro lado, si $R^2 \to 0$, $SCE \simeq SCT$. O sea, la predicción tiene tanta variabilidad como los datos inicial, el ajuste es por ello bastante pobre.

El coeficiente de determinación, R^2, pertenece en general al intervalo $[0, 1]$. No obstante, puede ocurrir que $R^2 < 0$, por ejemplo, cuando la especificación del modelo es incorrecta. En general, de aquí en adelante, tendremos $R^2 \in [0, 1]$. Por ende, el análisis se verá limitado a si $R^2 \sim 0$ (mal ajuste), o si $R^2 \sim 1$ (buen ajuste). Si bien no es claro que sea un buen predictor de ajuste, es muy frecuente que se incorpore en las investigaciones Rau (2016).

En general, el R^2 es bajo (menor a 0.5) en datos de corte transversal en comparación con los datos que provienen de series de tiempo. Por ello, no se le debe dar tanto peso al tamaño del R^2 cuando se realiza análisis con este tipo de datos. Por el contrario, cuando se trabaja con bases de datos temporales, se le suele dar énfasis al tamaño del R^2.

Definición 3.4.5. El coeficiente de correlación de Pearson es una medida de la dependencia lineal entre dos variables aleatorias

cuantitativas. Se define matemáticamente de la siguiente manera

$$\rho_{XY} = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}}.$$

En el caso de una muestra aleatoria de dos variables, $\{X_1, ..., X_n\}$ y $\{Y_1, ..., Y_n\}$

$$r_{XY} = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2} \cdot \sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \in [-1, 1].$$

El coeficiente de correlación de Pearson permite distinguir 5 casos de dependencia entre dos variables aleatorias.

1. Si $r = 1$, existe una correlación perfecta entre las 2 variables, cuando una de ellas aumenta, la otra también lo hace en proporción constante.

2. Si $0 < r < 1$, existe un correlación positiva. Mientras menor sea r ($r \to 0$), más débil será esta correlación. Si una aumenta, la otra puede que también, pero la intensidad y certitud de esto es cada vez menor conforme $r \to 0$.

3. Si $r = 0$, no existe correlación alguna, si X aumenta, Y puede aumentar, como decrecer o mantenerse constante. No se puede realmente sacar conclusión alguna. Esto no significa que no exista relación alguna entre las v.a. y por ende que estas sean independiente. Puede ser simplemente que la relación sea no lineal.

4. Si $-1 < r < 0$, existe un correlación negativa. Mientras menor sea r en valor absoluto ($r \to 0$), más débil será esta

correlación. Si una aumenta, la otra puede disminuir, pero la intensidad y certitud de esto es cada vez menor conforme $r \to 0$.

5. Si $r = -1$, existe una correlación negativa perfecta. Esto señala una dependencia total entre las dos variables llamada relación inversa. Cuando una de ellas aumenta, la otra disminuye en proporción constante.

Teorema 21. El coeficiente de determinación R^2 puede definirse en el caso más simple de regresión lineal vía el coeficiente de correlación de Pearson de la siguiente manera en el caso bivariado

$$R^2 = \rho_{XY}^2 = \frac{\sigma_{XY}^2}{\sigma_X^2 \sigma_Y^2} = \frac{\mathrm{Cov}(X,Y)^2}{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)}.$$

Demostración. Recordemos que $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$

$$\begin{aligned}
R^2 &= \frac{SCE}{SCT} \\
&= \frac{\mathrm{Var}(\hat{Y})}{\mathrm{Var}(Y)} \\
&= \frac{\mathrm{Var}(\hat{\beta}_0 + \hat{\beta}_1 X)}{\mathrm{Var}(Y)} \\
&= \frac{\hat{\beta}_1^2 \mathrm{Var}(X)}{\mathrm{Var}(Y)} \\
&= \left(\frac{\sum_{i=1}^n (Y_i - \overline{Y})(X_i - \overline{X})}{\sum_{i=1}^n (X_i - \overline{X})^2} \right)^2 \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)} \\
&= \left(\frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X)} \right)^2 \frac{\mathrm{Var}(X)}{\mathrm{Var}(Y)} \\
&= \frac{\mathrm{Cov}(X,Y)}{\mathrm{Var}(X) \cdot \mathrm{Var}(Y)} \\
&= \rho_{XY}^2.
\end{aligned}$$

□

El coeficiente de determinación R^2 puede obtenerse directamente haciendo uso de los parámetros estimados. En efecto,

$$R^2 = 1 - \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{Y^TY - n\overline{Y}^2} = 1 - \frac{(Y^T - \hat{\beta}^TX^T)(Y - X\hat{\beta})}{Y^TY - n\overline{Y}^2}.$$

Desarrollando el producto,

$$R^2 = 1 - \frac{Y^TY - Y^TX\hat{\beta} - \hat{\beta}^TX^TY + \hat{\beta}^TX^TX\hat{\beta}}{Y^TY - n\overline{Y}^2}.$$

Usando la expresión para $\hat{\beta}$,

$$R^2 = 1 - \frac{Y^TY - Y^TX\hat{\beta} - \hat{\beta}^TX^TY + \hat{\beta}^TX^TX(X^TX)^{-1}X^TY}{Y^TY - n\overline{Y}^2}.$$

Simplificando, se obtiene finalmente

$$R^2 = 1 - \frac{Y^TY - Y^TX\hat{\beta}}{Y^TY - n\overline{Y}^2} = \frac{Y^TX\hat{\beta} - n\overline{Y}^2}{Y^TY - n\overline{Y}^2}. \tag{3.5}$$

En el caso de la regresión lineal múltiple, es decir, el caso general, el coeficiente de determinación R^2 puede verse comprometido por el uso excesivo de variables explicativas. En efecto, al incluir cada vez más variables, ciertamente se va a reducir la variabilidad de los datos predichos, pero sacrificando el principio de parsimonia. Es por esto que se introduce una penalidad y se define el R^2 ajustado.

Definición 3.4.6. Sea k el número de variables explicativas (sin incluir la constante), n el tamaño de la muestra y R^2 el valor que se obtiene calculando (3.5). Entonces

$$R^2_{\text{ajustado}} = 1 - \left[\frac{n-1}{n-k-1}\right](1 - R^2). \tag{3.6}$$

Teniendo en cuenta la constante,

$$R^2_{\text{ajustado con constante}} = 1 - \left[\frac{n-1}{n-k}\right](1-R^2). \tag{3.7}$$

De este modo, analizando la expresión (3.6), se deduce que si k aumenta, el ratio $(n-1)/(n-k-1)$ también y por ende el R^2_{ajustado} disminuye. De esta manera, el R^2 ajustado puede tomar valores menores o iguales al R^2. Una diferencia principal entre el R^2 ajustado y el R^2 está en que este último solo toma valores entre 0 y 1, mientras el R^2 ajustado puede tomar valores negativos debido a que:

1. el número de variables explicativas se acerque al número de observaciones; es decir, no se cuenta con grados de libertad suficientes para la estimación de los parámetros. *Usualmente se recomienda un ratio de 10 observaciones por parámetro a estimar.*

2. el coeficiente de determinación es bajo: lo que indica que se esta incluyendo simplemente variables irrelevantes en el modelo de regresión.

3.4.2. Parámetros estimados

Ya hemos atacado el problema del ajuste global, evaluando el performance del modelo de manera general. Sin embargo, no nos hemos interesado aún sobre el significado de los parámetros estimados. En el caso más simple, en el que se tiene únicamente dos parámetros,

- $\hat{\beta}_0$ corresponde al valor predicho para Y_i cuando $X_i = 0$. Por ejemplo, si X_i mide los años de educación y Y_i el salario, $\hat{\beta}_0$ será el salario promedio para personas sin educación.

- Por otro lado, $\hat{\beta}_1$ es la pendiente de la recta de regresión muestral, nos indica en cuanto varia el valor predicho $\hat{Y}_i$ ante el cambio de una unidad de X_i. Usando el ejemplo anterior, $\hat{\beta}_1$ miden en cuanto se incrementa el salario por un año adicional de estudios.

Cuando se tiene más de un regresor, cada $\hat{\beta}_k$, $k > 0$, mide en cuanto varia el valor predicho para Y ante el incremento de una unidad de X_k.

En caso la relación funcional no sea estrictamente lineal, como por ejemplo[9],

$$\ln(w_i) = \sum_{j=0}^{m} \beta_j X_j + \varepsilon_i,$$

el significado de los $\hat{\beta}_j$ ya no es el mismo debido al la presencia del $\ln(\cdot)$.

El análisis debe entonces tener en cuenta si los regresores son argumento de un logaritmo, o si la variable por predecir es argumento de un logaritmo. Se tiene entonces los siguientes 4 casos.

- Nivel - Nivel: X vs Y, $\Delta Y = \hat{\beta}_1 \Delta X$. O sea, el incremento de una unidad en X hace incrementar Y en $\hat{\beta}_1$ unidades.

- log-Nivel: $\log X$ vs Y, $\Delta Y = \left(\frac{\hat{\beta}_1}{100}\right) \%\Delta X$. O sea, el incremento en 1 % en X Incrementa Y en $\hat{\beta}_1$ unidades.

[9] Esta ecuación es conocida como la ecuación de Mincer.

- Nivel-log: X vs $\log Y$, $\%\Delta Y = 100\hat{\beta}_1 \Delta X$. O sea, $\hat{\beta}_1$ mide en que $\%$ incrementa Y ante el incremento de una unidad de X.

- $\log - \log$, $\log X$ vs $\log Y$, $\%\Delta Y = \hat{\beta}_1 \, \%\Delta X$. El coeficiente mide en que porcentaje incrementa Y ante incremento en $1\,\%$ en X. En este sentido, los β_j representan elasticidades.

Note que el uso de logaritmos permite eliminar el efecto de las unidades en los coeficientes y se emplea cuando la variable es asimétrica. A menudo resulta útil calcular elasticidades utilizando derivación logarítmica. Un resultado importante en este contexto es presentado a continuación.

Teorema 22. Si $y = \varphi(x)$, entonces

$$\varepsilon = \frac{dy/y}{dx/x} = \frac{dy}{dx}\frac{x}{y}.$$

Siempre que $x, y > 0$,

$$\varepsilon = \frac{d\ln y}{d\ln x}.$$

Demostración. Aplicando la regla de la cadena,

$$\frac{d\ln y}{dx} = \frac{d\ln y}{d\ln x}\frac{d\ln x}{dx}.$$

Luego,

$$\frac{d\ln y}{d\ln x}\frac{1}{x} = \frac{1}{y}\frac{dy}{dx}.$$

O sea,

$$\frac{d\ln y}{d\ln x} = \frac{x}{y}\frac{dy}{dx}.$$

□

En este punto, una pregunta de interés es, ¿qué sucede si $\hat{\beta}_j = 0$? Ciertamente si $\hat{\beta}_0 = 0$, significa que el promedio de los datos predichos será cero cuando $X = 0$. Sin embargo, si $\hat{\beta}_j = 0$ para $j \neq 0$, ¿acaso la contribución marginal es de 0 %? El siguiente estadístico permite estudiar este tipo de escenarios y sus consecuencias. Pero antes, recordemos ciertos conceptos de inferencia estadística.

Definición 3.4.7. Hipótesis estadística. Una hipótesis estadística es cualquier enunciado que hagamos respecto a la distribución de una o más variables aleatorias.

Definición 3.4.8. Todo contraste de hipótesis sobre un parámetro unidimensional θ posee la forma siguiente

$$H_0 : \ \theta = \theta_0 \ , \ \text{vs} \ H_1 : \theta = \begin{cases} \theta_1, & \text{simple} \\ > \theta_0, & \text{a cola derecha} \\ < \theta_0, & \text{a cola izquierda} \\ \neq \theta_0, & \text{a dos colas.} \end{cases}$$

Definición 3.4.9.

$$\alpha = \mathbb{P}(\text{Error tipo 1}) = \mathbb{P}(\text{Rechazar } H_0 | H_0 \text{ es verdadera})$$
$$\beta = \mathbb{P}(\text{Error tipo 2}) = \mathbb{P}(\text{Aceptar } H_0 | H_0 \text{ es falsa}).$$

Observe que estamos asumiendo por simplicidad que H_0 toma la forma de una igualdad ($\theta = \theta_0$). Con mayor generalidad podemos tener una hipótesis del tipo $H_0 : \theta \in \Xi \subset \Theta$. En tal situación

$$\alpha = \sup_{\theta \in \Xi} \ \mathbb{P}(\text{Rechazar} \ \ H_0).$$

Un buen contraste debería ser aquel por el cual α y β son mínimos. Sin embargo, por lo general, ambas probabilidades están en relación inversamente proporcional. Es por ello que la convención es fijar una de estas medidas, específicamente α, con el fin de encontrar el mejor contraste, definido con aquel que para este α fijo, posea el menor β, o equivalentemente, la mayor potencia

$$\phi = \mathbb{P}(\text{Rechazar } H_0 | H_0 \text{ falsa}) = 1 - \beta.$$

Ejemplo 45. Ahora, veamos un caso concreto en el cual se hace uso de los tests de hipótesis. Más adelante, se estudiará el concepto de *heterocedasticidad de los errores*. Este concepto, hace referencia a un problema que puede presentarse en la naturaleza de los errores estimados $\hat{\varepsilon}_i$ vía MCO. La hipótesis nula H_0 en este caso, es que los errores no presentan heterocedasticidad. Una forma de evaluar esta hipótesis es aplicando un test estadístico conocido como el test de White[10]. En este último, se usa como estadístico la variable nR^2, la cual se distribuye según una χ^2 de parámetro q, correspondiente al número de regresores. Si $nR^2 > \chi^2_{1-\alpha}(q)$ (siendo α un nivel de significancia fijado previamente), se rechaza la hipótesis nula. Gráficamente, la Figura (3.1) nos muestra en qué consiste el test de hipótesis, teniendo en cuenta una regresión con 2 variables base y 3 variables adicionales (interacciones); de ahí $q = 5$.

[10]Este será estudiado con detenimiento más adelante en el Capítulo 8.

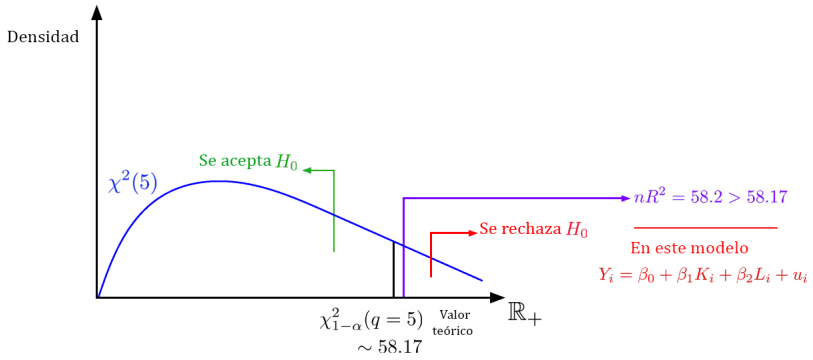

Figura 3.1 Gráfica del examen estadístico.

Note que el valor del estadístico (v.a. de la muestra) $nR^2 \in \mathbb{R}$ se obtiene previamente y simplemente se compara con el valor de tablar de $\chi^2_{1-\alpha}(5)$, siendo $\alpha = 0{,}05$ (o sea, significancia al 95 %)[11].

Definición 3.4.10. Definimos el F estadístico

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

dónde n es el número de observaciones y k el número de parámetros sin incluir la constante, como el ratio entre cuanto explican las variables incluidas en el modelo y lo que explica los errores.

La hipótesis nula H_0 de este estadístico es

$$\hat{\beta}_2 = \hat{\beta}_3 = ... = \hat{\beta}_k = 0.$$

Si $F > F_{\alpha,n,n-k-1}$, con α el nivel de significancia, rechazamos la hipótesis nula H_0. Si $F \leq F_{\alpha,n,n-k-1}$ se acepta la hipótesis nula. En

[11] Si fuese un examen de dos colas, $\alpha = 0{,}025$.

otras palabras, lo que se busca establecer mediante el estadístico F es cuanto explican los regresores en conjunto. También es posible evaluar la contribución marginal de cada variable individualmente usando el estadístico t (test t de Student)[12]. El siguiente ejemplo permitirá afianzar estas nociones.

Ejemplo 46. Mediante STATA, en el siguiente cuadro presentamos los resultados de una regresión lineal simple en la que se incluyen dos regresores, «*El índice del tipo de cambio real bilateral*» y el «*PBI de USA*», para estimar «*Las exportaciones no tradicionales de Perú en millones de dólares*».

```
. reg xnt itcrb07 pbiusa
```

Source	SS	df	MS	Number of obs	=	76
				F(2, 73)	=	900.79
Model	66739802.7	2	33369901.3	Prob > F	=	0.0000
Residual	2704289.24	73	37045.0581	R-squared	=	0.9611
				Adj R-squared	=	0.9600
Total	69444091.9	75	925921.226	Root MSE	=	192.47

xnt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
itcrb07	-32.63971	3.290984	-9.92	0.000	-39.19863	-26.08079
pbiusa	.2228377	.0109732	20.31	0.000	.2009681	.2447073
_cons	1607.755	447.0523	3.60	0.001	716.7809	2498.729

Figura 3.2 Regresión nivel-nivel, exportaciones (xnt) vs tipo de cambio bilateral (itcrb07) y PBI de USA.

La Figura (3.2) nos indica un coeficiente de determinación R^2 de 0.9611 y un R^2 ajustado de 0.9600. Debido a que los datos corresponden a una serie de tiempo, es importante analizar si

[12]Para mayor información sobre los tests estadístico ver Casella and Berger (2002).

$R^2 \sim 1$. Al ser el caso, podemos afirmar que globalmente el modelo cumple su función explicativa. Más aún, se logra explicar aproximadamente el 96 % de la variabilidad de los datos. Luego, la probabilidad que los $\hat{\beta}_j$ sean iguales a cero es prácticamente nula. Además, uno puede verificar con el valor de tablas que $F > F_{\alpha,n,n-k-1}$, siendo $n = 76$ y $k = 2$. Con esto en mente, procedemos a analizar individualmente los coeficientes β_j. Tal y como puede verse en la tabla.

1. Si el tipo de cambio bilateral (itcrb07) sube en una unidad, las exportaciones caen en 32 millones de dólares.

2. Por otro lado, si el PBI de los estados unidos se incrementa en 1 billón de dólares, las exportaciones se incrementarán en 0.22 millones de dólares.

Ambas interpretaciones son estadísticamente significativas pues el estadístico t es mayor en valor absoluto al valor de tablas para un nivel $\alpha = 0{,}05$. Recordemos que la H_0 es en este caso $\beta_1 = 0$ y $\beta_2 = 0$ (independientemente). Si $t \leq t_\alpha$, se acepta la hipótesis, lo cual implica que el parámetro puede ser igual a cero, i.e., no explicativo. Otra forma de percatarse que los coeficientes son significativos es analizando el intervalo de confianza. Este intervalo, establece con un $1 - \alpha$ de probabilidad, un rango de valores para β_j. En este caso, al 95 %, observamos que ninguno toma el valor de cero pues $0 \notin IC_{\beta_j}$. En conclusión, para establecer la significancia del modelo a nivel global, analizamos el R^2, R^2 ajustado y el F de Fisher. Enseguida, pasamos al análisis regresor por regresor. En dicho análisis, se evalúa si el t es menor al valor de tablas, si la

probabilidad de la hipótesis nula ($\beta_j = 0$) es mayor a 0.05, o si el 0 pertenece al intervalo de confianza. Finalmente, luego de descartar regresores no explicativos, se procede con la interpretación de los parámetros según la especificación del modelo (nivel o logaritmos). Por ejemplo, en (3.2), las escalas son nivel-nivel. En ese sentido, una unidad adicional en X_j representa un incremento de β_j unidades en Y_j. Sin embargo, si se tiene una especificación en logaritmos, como por ejemplo

$$\ln(\text{xnt})_t = \beta_0 + \beta_1 \ln(\text{itrb07})_t + \ln(\text{pbiusa})_t + u_t, \tag{3.8}$$

la interpretación cambia. La siguiente tabla , corresponde a dicha regresión, (3.8).

```
. reg lnxnt lnitcrm07 lnpbiusa
```

Source	SS	df	MS		
Model	28.9368316	2	14.4684158	Number of obs	= 76
Residual	2.13482357	73	.029244158	F(2, 73)	= 494.75
				Prob > F	= 0.0000
Total	31.0716552	75	.414288736	R-squared	= 0.9313
				Adj R-squared	= 0.9294
				Root MSE	= .17101

lnxnt	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
lnitcrm07	-.8589906	.5951755	-1.44	0.153	-2.045174	.3271924
lnpbiusa	2.934381	.1078397	27.21	0.000	2.719457	3.149305
_cons	-16.89463	3.310409	-5.10	0.000	-23.49227	-10.29699

Figura 3.3 Regresión log-log, exportaciones vs tipo de cambio bilateral y PBI de USA.

En este caso, el R^2 y R^2 ajustado siguen indicando una bondad de ajuste globalmente positiva. Al rededor del 90 % de la variabilidad de los datos es explicada. Tanto globalmente como individualmente el modelo es también estadísticamente significativo

pues $\mathbb{P} > F$ (probabilidad de la hipótesis nula en relación al estadístico de Fisher) y $\mathbb{P} > t$ (probabilidad de la hipótesis nula en relación al estadístico t de Student) son prácticamente nulas. El cambio más notorio es el de los valores en los parámetros estimados. Esto es sin embargo coherente con el hecho que se han tomado logaritmos en (3.8), y por ende, la contribución ya no es marginal pero más bien porcentual. Leemos que,

1. Si el tipo de cambio bilateral (itcrb07) sube en 1 %, las exportaciones caen en 0.85 % .

2. Si el PBI de USA se incrementa en 1 %, las exportaciones se incrementarán en 2,9 % por ciento.

3.5. Restricciones lineales

En esta última sección, nuestro objetivo será analizar relaciones funcionales lineales entre los parámetros. Recordemos que los tests vistos previamente han sido $H_0 : \beta_j = 0$ (para cada j independientemente) y $H_0 : \beta_j = 0$ (para cada j en conjunto). Sin embargo, considere por ejemplo la siguiente regresión especificación

$$Y_i = AK_i^{\alpha} L_i^{\beta} e^{\varepsilon_i}, \tag{3.9}$$

que es de hecho equivalente a

$$\ln Y_i = \beta_0 + \beta_1 \ln K_i + \beta_2 \ln L_i + \varepsilon_i. \tag{3.10}$$

Queremos verificar $\hat{\beta}_1 + \hat{\beta}_2 = 1$. Note que (3.9) corresponde a una forma funcional *Cobb-Douglas*, i.e., se asume que la producción

es igual a $f(K,L) = AK^{\alpha}L^{\beta}$, y se multiplica por un término estocástico e^{ε}, $\varepsilon \sim N(0,\sigma^2)$. Sacando logaritmos a la Ecuación 3.9 se obtiene (3.10); una forma lineal en las variables y se entiende que, un incremento de 1 % en la cantidad de uno de los factores de producción genera un incremento de β % en el nivel de producción Y (donde β es el parámetro asociado al regresor). La hipótesis que se busca contrastar tiene como objetivo analizar los rendimientos a escala que presenta la función de producción implícitamente definida. En este caso, analizar si f tiene rendimientos a escala constantes, es decir

$$f(\lambda K, \lambda L) = \lambda f(K,L),\ \forall\ \lambda > 0.$$

A continuación analizamos en detalle la técnica econométrica que permite incorporar restricciones.

Definición 3.5.1. Definimos la matriz de restricciones R como la matriz del sistema $R_{q\times k}\beta_{k\times 1} = r_{q\times 1}$, siendo r un vector fijo y q siendo el número de restricciones. Esta matriz es la que define las restricciones lineales.

Ejemplo 47. En el caso presentado previamente (3.9), teníamos

$$\hat{\beta}_1 + \hat{\beta}_2 = 1.$$

Por ende, $R = (1,1)^T$, $r = 1$ y $q = 1$.

Recordemos que $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$. A partir de esto se deduce que $R\hat{\beta} \sim N(R\beta = r, R\sigma^2(X^TX)^{-1}R^T)$. Más aún

$$(R\hat{\beta} - r)^T \left[\sigma^2 R(X^TX)^{-1}R^T\right]^{-1} (R\hat{\beta} - r) \sim \chi^2(q).$$

Usando la relación[13]

$$\hat{\sigma}^2 = \frac{q}{n-k-1}\hat{\varepsilon}^T\hat{\varepsilon}$$

tenemos

$$(R\hat{\beta} - r)^T \left[\frac{q}{n-k-1}\hat{\varepsilon}^T\hat{\varepsilon}R(X^TX)^{-1}R^T\right]^{-1}(R\hat{\beta} - r) \sim F_{q,n-k-1}.$$

Reemplazando con

$$\hat{\varepsilon}^T\hat{\varepsilon} = (Y - X\hat{\beta})^T(Y - X\hat{\beta}) = Y^TY - Y^TX\hat{\beta},$$

finalmente

$$(R\hat{\beta}-r)^T \left[\frac{q}{n-k-1}(Y^TY - Y^TX\hat{\beta})R(X^TX)^{-1}R^T\right]^{-1}(R\hat{\beta}-r) \sim F_{q,n-k-1}. \tag{3.11}$$

La Ecuación 3.11 nos permite testear hipótesis reemplazando r.

- Si el F calculado es mayor que el valor crítico $F_{\alpha,q,n-k-1}$ rechazamos H_0. Esto significa que hay suficiente evidencia estadística para concluir que las restricciones $R\beta = r$ no son verdaderas. En otras palabras, las restricciones impuestas por H_0 no se ajustan bien a los datos.

- Si F calculado es menor o igual al valor crítico $F_{\alpha,q,n-k-1}$, no rechazamos H_0. Esto significa que no hay suficiente evidencia estadística para rechazar las restricciones $R\beta = r$. En otras palabras, las restricciones impuestas por H_0 son consistentes con los datos.

[13] Si incluimos la constante en el conteo $n - k$.

Ejemplo 48. Se tiene el siguiente modelo de regresión

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i.$$

A partir de este último, se obtienen las matrices

$$X = \begin{pmatrix} 1 & 2 & 4 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \\ 1 & -1 & 3 \end{pmatrix}, \; Y = \begin{pmatrix} 6 \\ 0 \\ -1 \\ -4 \\ -1 \end{pmatrix}.$$

El objetivo es contrastar las siguientes hipótesis:

$$H_0 : \beta_1 = 0 \; , \; \beta_2 + \beta_3 = 1.$$

Para esto, necesitamos calcular

$$(R\hat{\beta} - r)^T \left[\frac{q}{n-k-1} \hat{\varepsilon}^T \hat{\varepsilon} R (X^T X)^{-1} R^T \right]^{-1} (R\hat{\beta} - r) \sim F_{q,n-k-1}.$$

Primero, obtenemos $\hat{\beta}$,

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= \left[\begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & -1 & -1 \\ 4 & -1 & 0 & 1 & 3 \end{pmatrix} \cdot \begin{pmatrix} 1 & 2 & 4 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \\ 1 & -1 & 3 \end{pmatrix} \right]^{-1} \begin{pmatrix} 1 & 1 & 1 & 1 & 1 \\ 2 & 0 & 0 & -1 & -1 \\ 4 & -1 & 0 & 1 & 3 \end{pmatrix} \begin{pmatrix} 6 \\ 0 \\ -1 \\ -4 \\ -1 \end{pmatrix} \\ &= \begin{pmatrix} -\frac{287}{428} & \frac{269}{107} & \frac{205}{428} \end{pmatrix}^T. \end{aligned}$$

Luego,

$$\begin{aligned}\mathrm{Var}(\hat{\varepsilon}) &= \frac{\hat{\varepsilon}^T\hat{\varepsilon}}{n-k} \\ &= \frac{(Y-X\hat{\beta})^T(Y-X\hat{\beta})}{n-k} \\ &= \frac{(Y-X\hat{\beta})^T(Y-X\hat{\beta})}{5-3}.\end{aligned}$$

Reemplazando con

$$Y - X\hat{\alpha} = \begin{pmatrix} 6 \\ 0 \\ -1 \\ -4 \\ -1 \end{pmatrix} - \begin{pmatrix} 1 & 2 & 4 \\ 1 & 0 & -1 \\ 1 & 0 & 0 \\ 1 & -1 & 1 \\ 1 & -1 & 3 \end{pmatrix} \begin{pmatrix} -287/428 \\ 269/107 \\ 205/428 \end{pmatrix}$$

y despejando $\hat{\varepsilon}^T\hat{\varepsilon}$, se obtiene

$$\hat{\varepsilon}^T\hat{\varepsilon} \simeq 2{,}64.$$

Finalmente,

$$R\hat{\beta} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix} = r.$$

Así, reemplazando con los valores numéricos ya obtenidos, se tiene que $F \simeq 8{,}40$.

3.5.1. Intervalos de confianza y $t-$Student

Cuando $q = 1$, denotando $\theta = R\beta = r$ y $\hat{\theta} = R\hat{\beta}$ [14]

$$\frac{\hat{\theta} - \theta}{\sqrt{s^2 R(X^T X)^{-1} R^T}} \sim t_{n-k}.$$

Luego, si deseamos un intervalo del $(1-\alpha)\,\%$ de confianza para el verdadero valor del parámetro θ, es suficiente con obtener las tablas de $t_{n-1}^{\frac{1-\alpha}{2}}$ e invertir el test. Esto es

$$\begin{aligned} 1-\alpha &= \mathbb{P}\left\{ t_{n-k}^{\alpha/2} \leq \frac{\hat{\theta} - \theta}{\sqrt{s^2 R(X^T X)^{-1} R^T}} \leq t_{n-k}^{\frac{1-\alpha}{2}} \right\} \\ &= \mathbb{P}\left\{ -t_{n-k}^{\frac{1-\alpha}{2}} \leq \frac{\hat{\theta} - \theta}{\sqrt{s^2 R(X^T X)^{-1} R^T}} \leq t_{n-k}^{\frac{1-\alpha}{2}} \right\} \\ &= \mathbb{P}\left\{ \theta \in \left[\hat{\theta} \pm t_{n-k}^{\frac{1-\alpha}{2}} \sqrt{s^2 R(X^T X)^{-1} R^T} \right] \right\}. \end{aligned}$$

Recordemos que la distribución t de Student se utiliza cuando se estima la media de una población normal con una muestra pequeña y se desconoce la varianza poblacional. La distribución t de Student con ν grados de libertad, denotada t_ν, es la distribución de la variable aleatoria $T = \frac{Z}{\sqrt{U/\nu}}$ donde $Z \sim N(0,1)$ es una variable aleatoria que sigue una distribución normal estándar, $U \sim \chi^2(\nu)$ es una variable aleatoria que sigue una distribución chi-cuadrado con ν grados de libertad, y Z y U son independientes. La función de densidad de probabilidad de la distribución t de Student con ν

[14] Si estamos considerando el modelo con k parámetros estimados incluyendo la constante, los grados de libertad son $n-k$. Cuando no se considera la constante, usamos $n-k-1$.

grados de libertad es

$$f(t;\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\,\Gamma\left(\frac{\nu}{2}\right)}\left(1+\frac{t^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

donde Γ es la función gamma Casella and Berger (2002). La distribución t de Student es simétrica respecto a $t = 0$. Para $\nu > 1$, $\mathbb{E}[T] = 0$. Para $\nu > 2$, $\text{Var}(T) = \frac{\nu}{\nu-2}$. Para encontrar el valor crítico $t_{n-k}^{\frac{1-\alpha}{2}}$, buscamos el punto t en el cual la integral acumulada de la densidad de probabilidad desde $-\infty$ hasta t es igual a $\frac{1-\alpha}{2}$:

$$\int_{-\infty}^{t_{n-k}^{\frac{1-\alpha}{2}}} f(t;n-k)\,dt = \frac{1-\alpha}{2}.$$

Aquí, $f(t;n-k)$ es la función de densidad de probabilidad de la distribución t de Student con $n-k$ grados de libertad.

3.5.2. Método de los residuos

Una alternativa para contrastar una serie de restricciones lineales donde no se utilizan los coeficientes estimados $\hat{\beta}$ es usar la suma de residuos al cuadrado del modelo estimado dos veces. En primer lugar, se estima el modelo sin las restricciones lineales y luego el modelo con las restricciones lineales

$$F_{q,n-k-1} = \frac{\frac{SRC_{CR}-SRC_{SR}}{q}}{\frac{SRC_{SR}}{n-k-1}}$$

con q el número de restricciones lineales, n el número de observaciones, k el número de variables explicativas sin incluir la constante, SRC_{CR} la suma de cuadrados con restricciones y SRC_{SR} la suma de cuadrados sin restricciones.

Cuando se asume que todos los coeficientes estimados de las explicativas son iguales a cero $[q = k]$, tenemos en dicho caso el siguiente resultado

$$F_{q,n-k} = \frac{\frac{SRC_{CR}-SRC_{SR}}{k}}{\frac{SRC_{SR}}{n-k-1}} = \frac{\frac{SCE}{k}}{\frac{SRC_{SR}}{n-k-1}} = \frac{R^2/k}{(1-R^2)/(n-k-1)}. \quad (3.12)$$

Retomando el caso del Ejemplo 44, usando (3.12), y teniendo en cuenta la ausencia de la constante en la especificación, se obtiene

$$F_{k,n-k} = \frac{\frac{SRC_{CR}-SRC_{SR}}{k}}{\frac{SRC_{SR}}{n-k}} \simeq 1{,}5.$$

3.5.3. Propiedades asintóticas

Para abordar las propiedades asintóticas del estimador MCO, debemos abordar algunos resultados preliminares. Algunos son detallados en el apéndice de teoría de la probabilidad.

Teorema 23. Cramer-Wald. Si $\sum_{i=1}^{k} \lambda_i X_n^i \to \sum_{i=1}^{k} \lambda_i X^i$ en distribución, entonces $X_n \to X$ en distribución.

Teorema 24. Teorema del Límite Central de Linderberg-Levy Multivariado. Sea $\overline{X}_n$ el promedio muestral de $\{X_i\}_{i=1,\dots,n}$ con $\mathbb{E}[X_i] - \mu$ y $\text{Var}(X_i) = \Sigma$. Entonces,

$$\sqrt{n}(\overline{X}_n - \mu) \to N(0, \Sigma)$$

en distribución.

Teorema 25. Si $X_n \to x_0$ en probabilidad y $g : \mathbb{R}^k \to \mathbb{R}$ es continua en x_0, entonces

$$g(X_n) \to g(x_0)$$

en probabilidad.

Teorema 26. Slutsky. Si $X_n \sim Y_n$, con $X_n, Y_n : \Omega \to \mathbb{R}^k$ y $X_n \to x_0$ en probabilidad, y $Y_n \to Y$ en distribución, entonces

1. $X_n + Y_n \to x_0 + Y$ en distribución
2. $X_n^T Y_n \to x_0^T Y$ en distribución.

Teorema 27. Mann-Wald. Si $X_n \to X$ en distribución y $g(x)$ es continua para todo x, entonces

$$g(X_n) \to g(X)$$

en distribución.

Teorema 28. Método Delta. Sea θ_n un vector aleatorio asintóticamente normal (convergencia en probabilidad) con $\sqrt{n}(\theta_n - \theta_0) \to N(0, \Sigma)$ en distribución y $g(\theta) \in C^1(V_{\theta_0})$ con Jacobiano

$$G_0 = \left.\frac{\partial g}{\partial \theta}\right|_{\theta=\theta_0}.$$

Entonces,

$$\sqrt{n}(g(\theta_n) - g(\theta_0)) \to N(0, G_0 \Sigma G_0^T)$$

en distribución.

Demostración. De acuerdo con el Teorema del Valor medio, existe $\tilde{\theta}_n \in [\theta_0, \theta_n]$ tal que

$$(g(\theta_n) - g(\theta_0)) = \left.\frac{\partial g}{\partial \theta}\right|_{\theta=\tilde{\theta}_n} (\theta_n - \theta_0)$$
$$\sqrt{n}(g(\theta_n) - g(\theta_0)) = \left.\frac{\partial g}{\partial \theta}\right|_{\theta=\tilde{\theta}_n} \sqrt{n}(\theta_n - \theta_0).$$

Dado que $\theta_n \to \theta_0$ en probabilidad, $\tilde{\theta}_n \to \theta_0$ en probabilidad. Luego, por el Teorema 25,

$$\left.\frac{\partial g}{\partial \theta}\right|_{\theta=\tilde{\theta}_n} \to G_0$$

en probabilidad. Por otro lado, $\sqrt{n}(\theta_n - \theta_0) \to N(0, \Sigma)$ en distribución. Luego, por el Teorema 26,

$$\begin{aligned}\sqrt{n}(g(\theta_n) - g(\theta_0)) &= \left.\frac{\partial g}{\partial \theta}\right|_{\theta=\tilde{\theta}_n} \sqrt{n}(\tilde{\theta}_n - \theta_0) \\ &\to G_0 N(0, \Sigma) \\ &= N(0, G_0 \Sigma G_0^T).\end{aligned}$$

□

Recordemos que

$$\hat{\beta} = (X^T X)^{-1} X^T Y = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)\left(\frac{1}{n}\sum_{i=1}^{n} X_i^T Y_i\right).$$

Ahora bien, $Y_i = X_i^T \beta + \epsilon_i$,

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^T (X_i\beta + \varepsilon_i)\right)$$

$$\hat{\beta} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^T \epsilon_i\right)$$

$$\sqrt{n}(\hat{\beta} - \beta) = \left(\frac{1}{n} X_i^T X_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i \epsilon_i\right).$$

Por la ley débil de los grandes números,

$$\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i \to \mathbb{E}[X_i^T X_i] = D$$

en probabilidad. Luego, por el teorema de la continuidad, tenemos que

$$\left(\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)^{-1} \to D^{-1}$$

en probabilidad. Como

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^T \epsilon_i = \frac{\sqrt{n}}{n}\sum_{i=1}^{n} X_i^T \epsilon_i,$$

podemos aplicar el TLC y

$$\begin{aligned}
\mathbb{E}[X_i\epsilon_i] &= \mathbb{E}[X_i(Y_i - X_i\beta)] \\
&= \mathbb{E}[X_i^T Y_i] - \mathbb{E}[X_i^T X_i]\beta \\
&= \mathbb{E}[X_i^T Y_i] - \mathbb{E}[X_i^T X_i](\mathbb{E}[X_i^T X_i])^{-1}\mathbb{E}[X_i^T Y_i] \\
&= \mathbb{E}[X_i^T Y_i] - \mathbb{E}[X_i^T Y_i] \\
&= 0 \\
\mathrm{Var}(X_i^T\epsilon_i) &= \mathbb{E}[X_i^T \epsilon_i \epsilon_i^T X_i] \\
&= \mathbb{E}[\epsilon_i^2 X_i^T X_i] \\
&= C.
\end{aligned}$$

Luego, por el TLC,

$$\frac{\sqrt{n}}{n}\sum_{i=1}^{n} X_i^T \epsilon_i \to N(0, C)$$

en distribución. Así, aplicando el Teorema 26,

$$\begin{aligned}
\sqrt{n}(\hat{\beta} - \beta) &= \left(\frac{1}{n}\sum_{i=1}^{n} X_i^T X_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} X_i^T \epsilon_i\right) \\
&\to D^{-1}N(0, C) = N(0, D^{-1}C(D^{-1})^T)
\end{aligned}$$

en distribución.

En capítulos posteriores[15], se levantarán los supuestos hechos en este capítulo. Antes, veamos una regresión lineal con un proceso iid:

$$Y_i = \beta X_i + \epsilon_i,$$

con $\mathbb{E}[\epsilon_i] = 0$ y $\text{Var}(\epsilon_i) = \sigma^2$. Luego, asumiendo que

$$\mathbb{E}[\epsilon_i^2 | X_i] = \mathbb{E}[\epsilon_i^2],$$

$$\begin{aligned} C &= \mathbb{E}[\epsilon_i^2 X_i^T X_i] \\ &= \mathbb{E}[\epsilon_i^2]\mathbb{E}[X_i^T X_i] \\ &= \sigma^2 \underbrace{\mathbb{E}[X_i^T X_i]}_{=D} \end{aligned}$$

y

$$\begin{aligned} \text{Var}(\hat{\beta}) &= D^{-1} C D^{-1} \\ &= [\mathbb{E}[X_i^T X_i]]^{-1} \sigma^2 \mathbb{E}[X_i^T X_i] \mathbb{E}[X_i^T X_i]^{-1} \\ &= \mathbb{E}[X_i^T X_i]^{-1} \sigma^2. \end{aligned}$$

Ahora bien,

$$\mathbb{P}\lim \frac{1}{N} X^T X = \mathbb{E}[X_i^T X_i]$$

por lo que

$$\sqrt{N}(\hat{\beta} - \beta) \to N(0, \sigma^2 D^{-1}) = N\left(0, \left(\mathbb{P}\lim \frac{1}{N} X^T X\right)^{-1}\right).$$

[15] En el siguiente capítulo se aborda un tipo de regresión en particular en el cual los regresores son variables discretas, categóricas. En el subsiguiente, el tema del muestreo. Ya en los que le siguen a este último, abordamos la multicolinealidad, heterocedasticidad, autocorrelación y endogeneidad.

La única diferencia con el modelo clásico yace en el hecho que ahora escribimos $\left(\mathbb{P}\lim \frac{1}{N} X^T X\right)^{-1}$ en vez de $\left(\frac{1}{N} X^T X\right)^{-1}$.

3.5.4. Estimador con restricciones

El problema de optimización en presencia de restricciones lineales corresponde analíticamente a

$$\mathcal{P}_R : \begin{cases} \text{mín}_\beta & (Y - X\beta)^T (Y - X\beta) \\ \text{s.a:} & R\beta = r. \end{cases}$$

Luego, para resolver $\mathcal{P}_R$, se plantea el Lagrangiano del problema

$$L(\beta, \lambda) = (Y - X\beta)^T (Y - X\beta) + \lambda(R\beta - r).$$

Mediante las condiciones de primer orden, se obtiene el siguiente par de ecuaciones

$$\begin{aligned} \frac{\partial L(\beta, \lambda)}{\partial \beta} &= -2X^T Y + 2X^T X\beta + 2\lambda^T R = 0 \\ \frac{\partial L(\beta, \lambda)}{\partial \lambda} &= R\beta - r = 0. \end{aligned}$$

Finalmente, luego de ciertas manipulaciones algebraicas,

$$\begin{aligned} \hat{\beta}^{LRS} &= (X^T X)^{-1} X^T Y - (X^T X)^{-1} R^T (R(X^T X) R^T)(R(X^T X)^{-1} X^T Y - r) \\ &= \hat{\beta}^{MCO} - (X^T X)^{-1} R^T (R(X^T X) R^T)(R\hat{\beta}^{MCO} - r). \end{aligned}$$

Capítulo 4

Variables cualitativas

Imagine que se busca estimar el logaritmo del salario de un trabajador. Una opción para ello es usar la ecuación de Mincer Borjas (2000), cuya especificación tradicional es[1]

$$\ln w = \ln w_0 + \rho x_1 + \beta_1 x_2 + \beta_2 x_2^2. \tag{4.1}$$

En (4.1), w es el salario, w_0 el salario promedio, x_1 los años de escolaridad, y x_2 los años de experiencia laboral. Sin embargo, en caso del Perú, podemos incluir un regresor adicional, x_3 que corresponde al sexo del trabajador. Es decir, $x_3 = \{\text{Hombre}, \text{Mujer}\}$ y

$$\ln w = \ln w_0 + \rho x_1 + \beta_1 x_2 + \beta_2 x_2^2 + \beta_3 x_3. \tag{4.2}$$

Si bien w, x_1 y x_2 son variables que pueden tomar valores en $\mathbb{R}$, x_3 no, es una variable cualitativa. Ciertamente no podemos decir que *un incremento en una unidad de* x_3 *genera un incremento en*

[1] Incorporar $\beta_2 x_2^2$ se explica en la literatura. Véase por ejemplo Polachek (2007).

1 % *del salario*. ¿Cómo interpretar entonces $\hat{\beta}_3$? Este será uno de los objetivos principales de este capítulo.

4.1. Conceptos básicos

Definición 4.1.1. Variable cualitativa. Una variable cualitativa indica la presencia o ausencia de un atributo o cualidad. Por ejemplo, *sexo, raza, religión, región, nacionalidad, afiliación política*, entre otros.

Ejemplo 49. Las siguientes ecuaciones representan especificaciones en las cuales los regresores son variables dicotómicas $x \in \{0, 1\}$, i.e., 1 indica la ausencia (o presencia) de un atributo

$$\begin{aligned} \text{Salario}_i &= \beta_1 + \beta_2 \text{Sexo}_i + \epsilon_i \\ \text{Salario}_i &= \beta_1 + \beta_2 \text{Urbano}_i + \epsilon_i \\ \ln(\text{Salario}_i) &= \beta_1 + \beta_2 \text{Sexo}_i + \beta_3 \text{Indígena}_i + \epsilon_i \end{aligned}$$

Al igual que con variables independientes cuantitativas continuas, con variable cualitativas también se usa el método de Mínimos Cuadrados Ordinarios para estimar.

Los supuestos del modelo en caso se incluyan variables no continuas son:

- Linealidad en X.
- Homocedasticidad: $\mathbb{E}[\epsilon_i] = 0$, $\text{Var}[\epsilon_i] = \sigma^2$ para todo i.
- Normalidad: las muestras de cada grupo deben provenir de poblaciones con distribución normal.

- Independencia de errores: no hay autocorrelación entre los errores de cada una de las observaciones en la muestra.

Note que deben incluirse $m-1$ categorías (si son m en total) para evitar la colinealidad, y así el modelo podrá ser estimable.

Ejemplo 50. Imaginemos que buscamos estimar el logaritmo del salario en función de la región r en la cual el trabajador habita, es decir,

$$r \in \{\text{Costa}, \text{Sierra}, \text{Selva}\}.$$

Si planteamos

$$\ln w_i = \beta_1 + \beta_2 \text{Costa}_i + \beta_3 \text{Sierra}_i + \beta_4 \text{Selva}_i + \epsilon_i,$$

como un individuo pertenece a una de las 3 regiones, se tiene (sin pérdida de generalidad) que

$$X = \begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 1 \\ \vdots & \vdots & \vdots & \vdots \\ 1 & 1 & 0 & 0 \end{pmatrix}.$$

O sea,

$$X_2 + X_3 + X_4 = X_1.$$

Sin embargo, esto viola el supuesto de multicolinealidad. Es por ello que deben solo incluirse dos categorías y no las tres. A esto se conoce como la *trampa de las dummies*.

Definición 4.1.2. La categoría a la cual no se asigna variable dicotómica se conoce como categoría base, de comparación, de

control, de referencia u omitida. Además, todas las comparaciones se hacen respecto a dicha categoría de comparación.

En concreto,

$$Y_i = \beta_1 + \sum_{j=2}^{m-1} \beta_j X_{ji}.$$

El valor de $\hat{\beta}_1$ es el valor promedio de la categoría base. Luego, $\hat{\beta}_j - \hat{\beta}_1 = \overline{Y}_j - \overline{Y}_1$. En efecto, $\hat{\beta}_1$ es el valor promedio de la variable dependiente cuando la variable explicativa toma el valor de 0.

Ejemplo 51. Retomando el modelo que busca estimar el salario de un individuo en función de la región en la cual habita, tendremos

$$\text{Salario}_i = \beta_1 + \beta_2 \text{Costa}_i + \beta_3 \text{Sierra}_i + \epsilon_i.$$

En este caso, la categoría base es la *región Selva*. Si el individuo i habita en la Costa, $\text{Costa}_i = 1$ y $\text{Sierra}_i = 0$. Análogamente, si el individuo pertenece a la Sierra, $\text{Costa}_i = 0$ y $\text{Sierra}_i = 1$. Luego,

$$\begin{aligned}\mathbb{E}[\text{Salario}_i|\text{Costa}_i] &= \mathbb{E}[\beta_1 + \beta_2 \text{Costa}_i + \beta_3 \text{Sierra}_i + \epsilon_i | \text{Costa}_i = 1, \text{Sierra}_i = 0] \\ &= \beta_1 + \beta_2\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\text{Salario}_i|\text{Sierra}_i] &= \mathbb{E}[\beta_1 + \beta_2 \text{Costa}_i + \beta_3 \text{Sierra}_i + \epsilon_i | \text{Costa}_i = 0, \text{Sierra}_i = 1] \\ &= \beta_1 + \beta_3\end{aligned}$$

$$\begin{aligned}\mathbb{E}[\text{Salario}_i|\text{Selva}_i] &= \mathbb{E}[\beta_1 + \beta_2 \text{Costa}_i + \beta_3 \text{Sierra}_i + \epsilon_i | \text{Costa}_i = 0, \text{Sierra}_i = 0] \\ &= \beta_1.\end{aligned}$$

De este modo, si por ejemplo, tenemos $\beta_1 = 1000$, $\beta_2 = 950$ y $\beta_3 = 300$, los trabajadores de la Selva ganan en promedio 1000

soles, mientras que los de la Costa ganan 1950 soles y los de la Sierra 1300 soles.

Ejemplo 52. Consideremos nuevamente el caso en el que se busca estimar el salario en soles de los trabajadores peruanos. Esta vez, se utiliza información de la educación (medida en años de estudio), el sexo del trabajador y la región natural donde vive (Costa, Sierra y Selva). La regresión lineal descrita es entonces

$$\text{Sal}_i = \beta_0 + \beta_1 \text{E}_i + \beta_2 \text{S}_i + \beta_3 \text{Si}_i + \beta_4 \text{Se}_i + \epsilon_i, \tag{4.3}$$

donde *Sal* es el salario, *E* los años de estudio, *S* el sexo (1 si es hombre, 0 si es mujer), *Si* región sierra, *Se* región selva y ϵ_i es el término de error aleatorio. Usando esta especificación (4.3), podemos calcular el promedio de los salarios en función de la región y la brecha salarial por sexo en cada región.

- $$\begin{aligned}\mathbb{E}[\text{Sal}_i|\text{S}_i = 0, \text{Costa}] &= \beta_0 + \beta_1 \text{E}_i + \beta_2(0) + \beta_3(0) + \beta_4(0) \\ &= \beta_0 + \beta_1 \text{E}_i.\end{aligned}$$

- $$\begin{aligned}\mathbb{E}[\text{Sal}_i|\text{S}_i = 1, \text{Costa}] &= \beta_0 + \beta_1 \text{E}_i + \beta_2(1) + \beta_3(0) + \beta_4(0) \\ &= \beta_0 + \beta_1 \text{E}_i + \beta_2.\end{aligned}$$

- $$\begin{aligned}\mathbb{E}[\text{Sal}_i|\text{S}_i = 0, \text{Sierra}] &= \beta_0 + \beta_1 \text{E}_i + \beta_2(0) + \beta_3(1) + \beta_4(0) \\ &= \beta_0 + \beta_1 \text{E}_i + \beta_3.\end{aligned}$$

- $$\begin{aligned}\mathbb{E}[\text{Sal}_i|\text{S}_i = 1, \text{Sierra}] &= \beta_0 + \beta_1\text{E}_i + \beta_2(1) + \beta_3(1) + \beta_4(0) \\ &= \beta_0 + \beta_1\text{E}_i + \beta_2 + \beta_3.\end{aligned}$$

- $$\begin{aligned}\mathbb{E}[\text{Sal}_i|\text{S}_i = 0, \text{Selva}] &= \beta_0 + \beta_1\text{E}_i + \beta_2(0) + \beta_3(0) + \beta_4(1) \\ &= \beta_0 + \beta_1\text{E}_i + \beta_4.\end{aligned}$$

- $$\begin{aligned}\mathbb{E}[\text{Sal}_i|\text{S}_i = 1, \text{Selva}] &= \beta_0 + \beta_1\text{E}_i + \beta_2(1) + \beta_3(0) + \beta_4(1) \\ &= \beta_0 + \beta_1\text{E}_i + \beta_2 + \beta_4.\end{aligned}$$

Luego, definiendo *SalH* el salario de los hombres y *SalM* el salario de las mujeres,

- Costa:
$$\mathbb{E}[\text{SalH}] = \beta_0 + \beta_1\text{E}_i + \beta_2$$
$$\mathbb{E}[\text{SalM}] = \beta_0 + \beta_1\text{E}_i$$
$$\mathbb{E}[\text{SalH}] - \mathbb{E}[\text{SalM}] = \beta_2.$$

- Sierra:
$$\mathbb{E}[\text{SalH}] = \beta_0 + \beta_1\text{E}_i + \beta_2 + \beta_3$$
$$\mathbb{E}[\text{SalM}] = \beta_0 + \beta_1\text{E}_i + \beta_3$$
$$\mathbb{E}[\text{SalH}] - \mathbb{E}[\text{SalM}] = \beta_2.$$

- Selva:

$$\mathbb{E}[\text{SalH}] = \beta_0 + \beta_1 \text{E}_i + \beta_2 + \beta_4$$

$$\mathbb{E}[\text{SalM}] = \beta_0 + \beta_1 \text{E}_i + \beta_4$$

$$\mathbb{E}[\text{SalH}] - \mathbb{E}[\text{SalM}] = \beta_2.$$

  En general, $\overline{\text{SalH}} - \overline{\text{SalM}} = \hat{\beta}_2$. Es este parámetro el que mide la brecha del salario por región. Note que, si se comparase entre regiones, el resultado cambiaría.

Ejemplo 53. Con el objetivo de determinar si existen o no diferencias en las calificaciones obtenidas por hombres y mujeres en una determinada asignatura, a partir de 20 observaciones se estimó el siguiente modelo

$$\text{Nota}_i = \beta_0 + \beta_1 \text{Nota media Micro}_i + \beta_2 \text{Género}_i + \epsilon_i,$$

donde la variable género toma el valor 1 si se trata de una mujer y 0 para un varón. Los resultados de la estimación fueron los siguientes

$$\widehat{\text{Nota}}_i = 25 + 0{,}75 \text{Nota media Micro}_i + 20{,}5 \text{Género}_i + \epsilon_i.$$

Luego,

$$\mathbb{E}[\text{Nota}|\text{Mujer}] - \mathbb{E}[\text{Nota}|\text{Hombre}] = 20{,}5.$$

Así, existe una diferencia en el esperado de la nota en función del género: las mujeres obtienen 20.5 puntos (en promedio) por encima que los hombres.

4.2. Interacciones

De momento, se han estudiado especificaciones de la forma

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + \epsilon_i$$

donde $X_{ji} \in \mathbb{R}$ o $X_{ji} \in \{0, 1\}$. Sin embargo, regresando al modelo (4.3), supongamos que se presume que la brecha salarial por sexo no es homogénea en cada región, i.e., que Sal_i en función de S_i depende de W_i, con

$$W = \text{Si o Se.}$$

¿Cómo contrastar dicha hipótesis? Se implementan lo que se conoce como *interacciones*.

Definición 4.2.1. Si en el modelo simplificado

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 W_i + \epsilon_i,$$

se presume que el valor de Y_i en función de X_i depende de W_i, el modelo se convertiría en

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 W_i + \beta_4 X_i W_i + \epsilon_i.$$

Se dice entonces que X interactúa con W.

Volviendo a (4.3), el modelo queda de la siguiente manera

$$\text{Sal}_i = \beta_0 + \beta_1 \text{E}_i + \beta_2 \text{S}_i + \beta_3 \text{Si}_i + \beta_4 \text{Se}_i + \beta_5 \text{S}_i \text{Si}_i + \beta_6 \text{S}_i \text{Se}_i + \epsilon_i,$$

si se presume una interacción entre el sexo del individuo y la región en la que habita.

Cuando X_i y W_i son variables binarias, el impacto se mide directamente en la suma de los coeficientes, en particular, la pendiente para X_i sería igual a $\hat{\beta}_2 + \hat{\beta}_4 W_i$.

Ejemplo 54. Suponga que se busca analizar el efecto de los años de escolaridad de la madre sobre el estado nutricional de las niñas y niños. No obstante, se presume que dicha relación puede ser afectada por la condición de pobreza de la madre. En otras palabras, lo que se plantea es que existe una diferencia en la pendiente (4.1), o relación entre el estado nutricional y años de escolaridad, en caso la madre sea pobre o no.

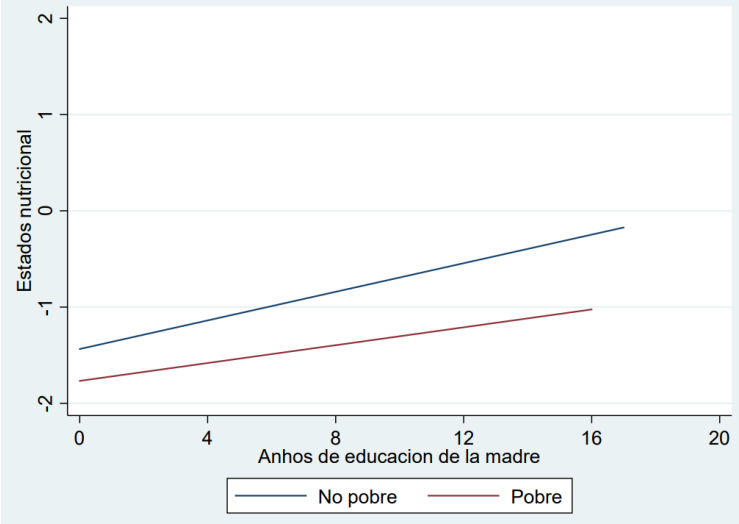

Figura 4.1 Diferencia en las pendientes.

El modelo en cuestión, teniendo en cuenta la interacción, es

$$\begin{aligned}\text{Nutrición}_i &= \beta_0 + \beta_1 \text{Educación Madre}_i + \beta_2 \text{Pobre}_i \\ &\quad + \beta_3 \text{Educación Madre}_i \cdot \text{Pobre}_i + \epsilon_i.\end{aligned}$$

De este modo, si el niño tiene una madre que es considerada *Pobre*,

$$\begin{aligned}\text{Nutrición}_i &= \beta_0 + \beta_1 \text{Educación Madre}_i + \beta_2 + \beta_3 \text{Educación Madre}_i + \epsilon_i \\ &= (\beta_0 + \beta_2) + (\beta_1 + \beta_3) \text{Educación Madre}_i + \epsilon_i.\end{aligned}$$

Así, tanto el promedio como la pendiente se ven afectados. Note que, en particular, el sentido de la relación entre el estado nutricional del niño y la educación de la madre puede cambiar.

Ejemplo 55. En la Figura (4.2) se presentan los resultados de la regresión

$$\begin{aligned} Y_i = \text{Talla para la edad}_i &= \beta_0 + \beta_1 \text{Madre trabaja}_i \\ &+ \beta_2 \text{Urbano}_i + \beta_3 \text{Riqueza}_i + \epsilon_i. \end{aligned} \tag{4.4}$$

En esta última, las variables tanto la variable *Madre trabaja* como *Urbano* son binarias. Note primero que la variable *Madre trabaja* es no significativa pues $0 \in IC = [\hat{\beta}_1 - \delta, \hat{\beta}_1 + \delta]$. Luego, si i vive en una zona urbana, en promedio, i tendrá un valor para Y_i superior por 24 unidades[2] al individuo ℓ que reside en una zona rural. Finalmente, la riqueza es una variable que influye positivamente sobre la talla para la edad. Concretamente, un incremento de una unidad en la riqueza genera un incremento de 35 unidades en la talla para la edad. Sin embargo, globalmente, para ser un modelo de corte transversal la significancia es positiva ($\mathbb{P} > F \sim 0$). No obstante, al estudiar el R^2 y R^2_{adj}, nos percatamos que el modelo explica únicamente el 1,3 % de la variabilidad de los datos. Esto hace reflexionar sobre la especificación lineal utilizada pues es un valor considerablemente bajo.

[2] Según la medida tomada para Y.

```
. regress HW70 trabaja_m urbano V190
```

Source	SS	df	MS	Number of obs	=	14,847
				F(3, 14843)	=	66.72
Model	42599065.7	3	14199688.6	Prob > F	=	0.0000
Residual	3.1588e+09	14,843	212813.028	R-squared	=	0.0133
				Adj R-squared	=	0.0131
Total	3.2014e+09	14,846	215639.421	Root MSE	=	461.32

HW70	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
trabaja_m	3.935325	7.667624	0.51	0.608	-11.09417	18.96482
urbano	24.89572	10.35635	2.40	0.016	4.595986	45.19546
V190	35.12138	3.597586	9.76	0.000	28.06967	42.1731
_cons	-171.419	9.318188	-18.40	0.000	-189.6838	-153.1542

Figura 4.2 Regresión talla para la edad, ENDES 2019.

El Ejemplo 54 se basa en los datos de la Encuesta Demográfica y de Salud Familiar - ENDES (2019). La cantidad de datos observados asciende a 14 847 individuos, número considerable considerando el número de regresores en el modelo. Justamente, es en el siguiente capítulo en el cual se abordarán los temas relacionados al tamaño de la muestra, criterios de selección de muestra, entre otros.

Capítulo 5

Muestreo

El concepto de muestra surge por la necesidad de recolectar información, datos, pero muchas veces, dada la gran cantidad de elementos (personas por ejemplo) a las cuales se les extrae la información, solamente es posible acceder a una parte del total. En este capítulo se van a estudiar los conceptos básicos del muestro y presentar algunos ejemplos en los cuales se puede apreciar la importancia de esta técnica. En una primera instancia, empezaremos con las definiciones elementales. Luego, analizaremos el problema del *tamaño de muestra*. Enseguida, presentaremos una de las dos formas de seleccionar una muestra: el *muestreo probabilístico*. Se sigue con los diseños experimentales, esenciales para el desarrollo, por ejemplo, de políticas públicas o programas sociales. Finalmente, concluimos con el estudio del método conocido de remuestreo conocido como Bootstrap.

5.1. Introducción y conceptos básicos

Definición 5.1.1. Muestra. Una muestra es un grupo de individuos u objetos de la población usadas para hacer inferencia de la misma. Esta se realiza por la falta de recursos o tiempo que demora encuestar a toda la población.

Para la elaboración de una muestra es necesario un *marco muestral*, definido a continuación.

Definición 5.1.2. Marco muestral. Es el listado de la población objetivo. Por ejemplo: listado de la clase, registro de alumnos en la universidad, listado de escuelas, entre otras.

La ventaja de las muestras aleatorias es que permiten realizar generalizaciones sobre la población. La pregunta central que debe ser formulada en este punto, es *¿qué factores influyen en la representatividad de una muestra?* En efecto, si se conocen dichos factores, se pueden elaborar estudios de forma que se pueda, con mayor certeza, generalizar sobre toda la población a partir de un subconjunto de esta (la muestra). *Grosso modo*, son 3:

1. El tamaño de la muestra.
2. El método del muestreo.
3. La tasa de respuesta.[1]

[1]La tasa de respuesta es una medida que indica el porcentaje de personas que respondieron a una encuesta o estudio en comparación con el número total de personas a las que se les solicitó participar.

En una primera instancia, se aborda el tema del *tamaño de la muestra*. Veamos.

5.2. Tamaño de muestra

A continuación, enunciamos una serie de resultados sin proveer las demostraciones. Un análisis más detallado y especializado puede encontrarse en Valdivieso (2020).

Teorema 29. El número de elementos en la muestra para una variable aleatoria binaria, i.e., $X \in \{0,1\}$ es[2]

$$n = \frac{z_{1-\alpha/2}^2 p(1-p)N}{N\epsilon^2 + z_{1-\alpha/2}^2 p(1-p)} \ , \ p = 1/2 \text{ (poblacional)}.$$

[2] $z_{1-\alpha/2}$ es el valor crítico de la distribución normal estándar correspondiente a un nivel de confianza $1-\alpha$. Se define como:

$$\mathbb{P}(Z \leq z_{1-\alpha/2}) = 1 - \alpha/2$$

donde $Z \sim N(0,1)$. Matemáticamente,

$$z_{1-\alpha/2} = \text{ínf}\left\{z \in \mathbb{R} \mid \Phi(z) \geq 1 - \alpha/2\right\}$$

con $\Phi(z)$ siendo la función de distribución acumulada (CDF) de la distribución normal estándar:

$$\Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, dt$$

Para un nivel de confianza del 95

$$1 - \alpha/2 = 0{,}975 \quad \text{y} \quad z_{0{,}975} \approx 1{,}96.$$

Más aún, si $N \to \infty$ (N el tamaño de la población)

$$n = \frac{z^2_{1-\alpha/2} p(1-p)}{\epsilon^2}. \tag{5.1}$$

Teorema 30. El número de elementos en la muestra para una variable aleatoria continua ($X \in \mathbb{R}$) es

$$n = \frac{z^2_{1-\alpha/2}\sigma^2 N}{N\epsilon^2 + z^2_{1-\alpha/2}\sigma^2} \to \frac{z^2_{1-\alpha/2}\sigma^2}{\epsilon^2}. \tag{5.2}$$

Usualmente se toma

$$\sigma^2 = \frac{R}{6} \tag{5.3}$$

con $R = X_{(n)} - X_{(1)}$.[3]

Estos resultados se deducen a la hora de trabajar con intervalos de confianza y variables pivote. Brindamos a continuación un breve resumen sobre este tópico.

5.2.1. Intervalos de confianza

Dada una muestra aleatoria $\{X_1, .., X_n\}$ de una variable aleatoria $X \sim \theta \in \Theta$[4], nos interesa estimar θ no solo por su valor (punctual), sino por un rango de valores que contengan a θ.

Definición 5.2.1. Diremos que las estadísticas L_1 y L_2 conforman un intervalo de confianza $IC = [L_1, L_2]$ al $100(1-\alpha)\,\%$ para θ si

$$\mathbb{P}(L_1 \leq \theta \leq L_2) = 1 - \alpha.$$

[3] Aquí, $X_{(i)}$ denota el i-ésimo estadístico de orden de la muestra. Es decir, $X_{(1)}$ es el valor más pequeño (mínimo) de la muestra, y $X_{(n)}$ es el valor más grande (máximo) de la muestra.

[4] La notación $X \sim \theta$ indica que X se relaciona con una distribución a priori no conocida vía el parámetro o vector de parámetros θ. Por ejemplo, media μ o varianza σ^2.

El procedimiento para estimar θ vía un parámetro es el siguiente.

1. Definir una variable pivote $W = W(X_1, ..., X_n; \theta)$, adecuada, esto es, que W solo dependa de la m.a. y de θ como único valor desconocido, y que tenga distribución conocida.

2. Encontar a, b tal que

$$\mathbb{P}(a \leq W \leq b) = 1 - \alpha.$$

3. Despejar la inecuación para obtener

$$\mathbb{P}(L_1 = L_1(X_1, ..., X_n) \leq \theta \leq L_2 = L_2(X_1, ..., X_n)) = 1 - \alpha.$$

Es usual, sobre todo si la distribución de la variable pivote es simétrica, tomar áreas iguales en las colas de la distribución de W. Esto es, considerar a, b tal que

$$\mathbb{P}(W \leq a) = \mathbb{P}(W > b) = \frac{\alpha}{2}.$$

Usualmente la variable pivote W se forma partiendo del estimador de máxima verosimilitud $\hat{\theta}_{MV}$ de θ aprovechando que asintóticamente se tiene $\hat{\theta}_{MV} \sim N(\theta, \sigma_\theta^2)$. Así, una variable pivote podría tomarse en la construcción de un IC aproximado para θ es

$$W = \frac{\theta_{MV} - \theta}{\sigma_\theta}.$$

Teorema 31. Sea $X_1, ..., X_n$ una muestra aleatoria de una variable $X \sim \mathcal{N}(\mu, \sigma^2)$ y S^2 la varianza muestral

1. Para la media μ con varianza σ^2 conocida

$$IC = \left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}; \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\right].$$

2. Cuando la varianza σ^2 es desconocida y se desea estimar μ

$$IC = \left[\overline{X} - t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}; \overline{X} + t_{1-\alpha/2}(n-1)\frac{S}{\sqrt{n}}\right].$$

   Acá la variable pivote es $T = \frac{\overline{X}-\mu}{S/\sqrt{n}} \sim t(n-1)$.

3. Para estimar σ^2 usamos

$$IC = \left[\frac{(n-1)S^2}{\chi^2_{1-\alpha/2}(n-1)}; \frac{(n-1)S^2}{\chi^2_{\alpha/2}(n-1)}\right].$$

   Acá la variable pivote es $W = \frac{(n-1)S^2}{\sigma^2} \sim \chi^2(n-1)$.

Si p denota la proporción de una población con característica A, y $\overline{p} = \frac{X}{n}$ la proporción en una muestra, con x el número de elementos con la propiedad A en la muestra y $n \geq 30$ el tamaño de esta, entonces, X se distribuye como una variable aleatoria binomial

$$Z = \frac{X - np}{\sqrt{np(1-p)}} = \frac{\overline{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1).$$

En este contexto, $\frac{\overline{p}-p}{\sqrt{\frac{p(1-p)}{n}}}$ es la variable pivote. El intervalo de confianza para p es

$$IC = \left[\overline{p} - z_{1-\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}, \overline{p} + z_{1-\alpha/2}\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\right],$$

siendo $\overline{p}$ la proporción observada.

Teorema 32. Consideremos

- N población total.
- n población muestral.

- p proporción en la población.
- $\overline{p}$ proporción en la muestra.

Se cumple que

$$Z = \frac{\overline{p} - p}{\sqrt{\frac{\overline{p}(1-\overline{p})}{n}}\sqrt{\frac{N-n}{N-1}}} \sim N(0,1).$$

Así, esta es una variable pivote. El intervalo de confianza al $100(1-\alpha)\,\%$ para p es entonces

$$IC = \left[\overline{p} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}, \overline{p} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}\right]$$

y el IC para la media poblacional μ

$$IC = \left[\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}\sqrt{1-\frac{n}{N}}\right]. \quad (5.4)$$

Así, de (5.4) y las consideraciones en relación al rango de valores y el valor de p, se deducen (5.1) y (5.2).

5.2.2. Aplicaciones

Teniendo presente las expresiones (5.1) y (5.2), veamos algunos ejemplos de aplicación directa y otros con contexto.

Ejemplo 56. En la siguiente tabla, podemos determinar usando p, $z_{1-\alpha}$ y ϵ el valor de n.

Confianza al 95 %[5]	p	Error muestral ϵ	Muestra n
1.96	0.5	0.01	9604
1.96	0.5	0.05	384
1.96	0.5	0.07	196
1.96	0.5	0.10	96

Entonces, si queremos representar a una población con un nivel de confianza del 95 % y margen de error del 5 %, se necesita una muestra de 384 observaciones.

Ejemplo 57. Del ejemplo anterior, si queremos retratar a la población peruana bastaría con una muestra de 384 observaciones y tendríamos un margen de error del 5 %.

	N	%	Margen de error
Total	400	100	5 %
Hombre	200	50	7 %
Mujer	200	50	7 %
Costa	200	50	7 %
Sierra	150	38	8 %
Selva	50	12	14 %

De esta manera, se puede apreciar que si quisiéramos hablar de la población de la selva, el margen de error sería de 14 %[6]. Por este motivo, al momento de elaborar una muestra hay que pensar en los grupos sobre los que se quiere sacar conclusiones.

En caso la población bajo estudio sea pequeña, se debe aplicar la corrección por poblaciones pequeñas. En efecto, recuérdese que para encontrar (5.1) y (5.2) se considera $N \to \infty$. La formula es la siguiente

$$n_1 = \frac{n_0}{\left(1 + \left(\frac{n_0 - 1}{N}\right)\right)}$$

donde:

[6]Se despeja ϵ en términos de N: $\epsilon = \sqrt{1{,}96^2 \cdot 1/2(1-1/2)/N}$.

1. n_0 es el tamaño de muestra original,

2. n_1 es el tamaño de muestra corregido,

3. y N es el tamaño de la población.

Esta corrección se aplica para $N < 10000$.

Ejemplo 58. Asumamos que tenemos una población de 500 habitantes. Los tamaños de muestra para los diferentes tamaños de error son

Confianza 95 %	p	ϵ	n_0	n_1
1.96	0.5	0.01	9604	475
1.96	0.5	0.05	384	217
1.96	0.5	0.07	196	141
1.96	0.5	0.10	96	81

Tal y como se puede apreciar en los ejemplos anteriores, la fórmula del tamaño de muestra depende del tamaño de la población, pero también, de como se seleccionan a los individuos. En efecto, en función de esto, se tomará en consideración ciertas características sobre la distribución poblacional. A continuación, vamos a presentar las diferentes formas de ejecutar un muestreo (seleccionar elementos para muestra). Este proceso no es homogéneo y tiene importantes consecuencias.

5.3. Selección de la muestra

Definición 5.3.1. Tipos de muestreo.

- **Muestro probabilístico**: se le da una probabilidad diferente a cero a cada uno de los elementos o individuos que se seleccionan de la población. Solo este tipo de muestreos aseguran representatividad de la muestra que se obtiene de la población (y por ende se puede hacer inferencia).
- **Muestreo no probabilístico**: los individuos que se seleccionan de la población no tiene una probabilidad de ser elegidos. Es decir, al seleccionar se realiza siguiendo ciertos criterios, procurando que la muestra sea representativa quitando el factor aleatorio de por medio.

En este texto, nos interesamos exclusivamente por el muestro probabilístico.

Definición 5.3.2. Muestro aleatorio simple (MAS). Es el tipo de muestreo más simple. Se asigna con número o etiqueta a cada miembro de la población, y después, se usa algún medio automático para seleccionar a los individuos (generación de números aleatorios).

Las *semillas* permiten generar números aleatorios pero, manteniendo la aleatoridad, se puede repetir la selección.

Definición 5.3.3. Muestro aleatorio sistemático. Se seleccionan a los individuos de la siguiente manera. Se enumera / etiqueta y se escoge aleatoriamente a un elemento entre 1 y n. Después,

definiendo

$$k = \text{mín}\left\{\left\lfloor \frac{N}{n} \right\rfloor, 1\right\}$$

con N el tamaño de la muestra, se van seleccionando a los individuos espaciados de k. Es decir, si el individuo seleccionado aleatoriamente al inicio es i, el segundo es el de la posición $i+k$, y así sucesivamente.

Definición 5.3.4. Muestreo aleatorio estratificado. Permite reducir el error muestral para un tamaño de muestra dado. La idea es considerar categorías o grupos (estratos). Cada estrato es homogéneo de acuerdo a una determinada característica (sexo, género...). La idea, es que todos los estratos estén representados en la muestra. Finalmente, la distribución de la muestra en los diferentes estratos se puede hacer simple (MAS) o proporcional de acuerdo al tamaño de la población en cada estrato.

Definición 5.3.5. Muestreo por conglomerado. Es una técnica de muestreo utilizada cuando hay agrupamientos *naturales* relativamente homogéneos en una población estadística. En esta técnica, la población total se divide en estos grupos (o clusters) - escuelas, hospitales - y vía un MAS se selecciona a individuos de estos grupos, previamente definidos.

Definición 5.3.6. Muestreos probabilísticos complejos. Los muestreos complejos por lo general involucran dos o más etapas de selección de la muestra o individuos bajo estudio. En otras palabras, se cuenta con más de una unidad de muestreo. Por ejemplo, *escuelas* $\rightarrow$ *aulas o secciones* $\rightarrow$ *estudiantes*. Finalmente, al interior de cada sub-categoría, se aplica un MAS.

Estos son esencialmente todas las metodologías relativas al muestro probabilístico. Más adelante, veremos como estos se aplican en casos muy concretos. Antes de pasar al estudio de los diseños experimentales, brindamos una breve nota sobre el muestro no probabilístico y sobre los pesos muestrales.

Como ya se mencionó, el muestro no probabilístico es el que contempla cierta designación no aleatoria a la hora de seleccionar los elementos para construir la muestra. Esencialmente, se tienen los siguientes tipos de muestro no probabilístico.

- **Muestro por cuotas**: se eligen por características específicas (edad, género, niveles educativos). Se usa en encuestas de opinión.
- **Muestreo intencional o por conveniencia**: se selecciona individuos de acuerdo a su accesibilidad o por ciertos criterios específicos de interés para, por ejemplo, anuncios etc...
- **Bola de nieve o en cadena**: consiste en localizar algunos individuos, los cuales posteriormente referirán a otros.
- **Muestreo de casos extremos**: consiste en selecciona individuos alejados de la *normalidad*, por ejemplo, para seleccionar personas sumamente violentas podríamos selecciona una muestra de pandilleros.

Definición 5.3.7. Pesos muestrales. En ocasiones el número de individuos u objetos muestreados por un determinado grupo es mayor a la proporción que representan en la población. Entonces,

al realizar los estadísticos descriptivos de dicha muestra no van a darnos resultados similares a los de la población, motivo por el cual se requiere el uso de pesos muestrales. Entonces, los pesos muestrales sirven para poder ajustar la muestra seleccionada y esta pueda representar de manera adecuada a la población. El *peso* dado es

$$w_i = \frac{1}{p_i},$$

donde p_i es la probabilidad de selección del individuo[7]. En caso de muestreos multi-etapas,

$$w_i = \frac{1}{p_{i_1}} \frac{1}{p_{i_2}} \frac{1}{p_{i_3}} \cdots \frac{1}{p_{i_k}}.$$

Definición 5.3.8. Ajuste por tasa de no respuesta. Si uno o más individuos no fueron cubiertos, se realiza la siguiente corrección

$$a_i = \frac{n_i}{N_i}$$

siendo n_i y N_i la cantidad encuestada finalmente y la población objetivo inicial. Así, el peso final sería

$$w_f = a_i w_i.$$

Ejemplo 59. Un estudiante de economía de la PUCP decide hacer un estudio sobre el bullying a los estudiantes de secundaria en las escuelas de Lima Metropolitana. Motivo por el cuál primero decide estimar cuantos estudiantes necesita encuestar para poder hablar de los estudiantes de escuelas públicas y privadas. Sus estimados le dan que debe tener aproximadamente 400 estudiantes de escuelas

[7] Es decir, x/N con x el número de individuos del grupo y N el total.

públicas y 400 de escuelas privadas. Para hacer su marco muestral usa el padrón de instituciones educativas secundarias del Ministerio de Educación para Lima Metropolitana y lo divide en públicas y privadas. Luego, al interior de cada grupo decide seleccionar de forma aleatoria un total de 40 escuelas y al interior de cada escuela selecciona de forma aleatoria a una sección por grado y dos estudiantes por sección también de forma aleatoria. De esta forma, el estudiante lograría encuestar un total de 800 estudiantes de Lima Metropolitana. El muestro en cuestión corresponde ciertamente a un muestreo probabilístico por conglomerados pues, se identifican cadenas de grupos y se seleccionan elementos en función de estos últimos. Cabe resaltar que en cada selección, se emplea un MAS.

5.4. Diseños experimentales

En esta sección, abordamos los fundamentos y principales conceptos que aparecen en el *diseño de experimentos*.

Definición 5.4.1. Tamaño del efecto. Es la magnitud del efecto que se está estudiando, en el caso de un diseño experimental seria la diferencia entre el grupo tratado y control.

Distinguimos las diferentes medidas para dicho tamaño de efecto.

1. Diferencia simple de promedios

$$\overline{X}_T - \overline{X}_C.$$

2. Diferencia estandarizada de los promedios.

a) La d de Cohen,

$$d = \frac{\overline{X}_T - \overline{X}_C}{\text{DE}_{\text{pooled}}}, \ \text{DE}_{\text{pooled}} = \sqrt{\frac{\sigma_T^2 + \sigma_C^2}{2}}. \tag{5.5}$$

b) La g de Hedges,

$$g = \frac{\overline{X}_T - \overline{X}_C}{\text{DE}_{\text{pooled}}}, \ \text{DE}_{\text{pooled}} = \sqrt{\frac{\sigma_T^2(n_T - 1) + \sigma_C^2(n_C - 1)}{n_T + n_C - 2}}. \tag{5.6}$$

c) La Δ de Glass

$$\Delta = \frac{\overline{X}_T - \overline{X}_C}{\text{DE}_{\text{C}}}, \ \text{DE}_{\text{C}} = \sqrt{\frac{\sum(X_C - \overline{X}_C)}{n_C}}. \tag{5.7}$$

El término *pooled* hace referencia a la pertenencia a un mismo grupo (mismas característica). Por otro lado, la letra C designa el conjunto de control y T el conjunto tratado. Finalmente, note que $N = n_C + n_T > 2$, (5.6).

En Cohen (1988), se indica que un tamaño de efecto es pequeño si esta por encima de 0.2 DE pero debajo de 0.5 DE; es mediano si se ubica por encima de 0.5 DE pero por debajo de 0.8 DE; y se considera un tamaño de efecto grande si este es igual o mayor a 0.8 DE.

Así como en el caso de las encuestas, es de interés conocer el tamaño de muestra en el diseño de experimentos. Para ello, necesitamos previamente introducir los siguientes dos conceptos, el nivel de confianza y el poder de análisis (mencionados previamente a la hora de estudiar los tests de hipótesis).

Definición 5.4.2. Nivel de confianza. Es el nivel de confianza que se tiene de los resultados, es decir, nos indica la probabilidad

que tenemos de que el parámetro estimado se encuentre dentro del intervalo asumido.

Definición 5.4.3. Poder de análisis. Es la probabilidad que tiene la muestra para poder detectar el parámetro de interés y tamaño de efecto deseado.

Usualmente, debido a la normalidad, usamos $z_{\alpha/2}$ en lo que concierne el nivel de confianza y $z_{\beta/2}$ en lo que concierne el poder de análisis.

A continuación se presenta la fórmula para el tamaño de muestra por grupo asumiendo:

1. La variable de resultado sigue una distribución normal,
2. El número de observaciones por grupo es igual
3. Las varianzas de ambos grupos son iguales.

Teorema 33. Sea TE el tamaño de efecto, previamente determinado, entonces si la variable resultado es continua,

$$n = \frac{(\sigma_1^2 + \sigma_2^2)[z_{\alpha/2} + z_\beta]^2}{TE^2}.$$

Siendo σ_i la desviación estándar de $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$, i.e. la v.a. representativa del grupo i.

Teorema 34. Sea TE el tamaño de efecto, previamente determinado, entonces si la variable resultado es binaria,

$$n = \frac{[p_1(1-p_1) + p_2(1-p_2)][z_{\alpha/2} + z_\beta]^2}{TE^2}. \tag{5.8}$$

Siendo σ_i la desviación estándar de $X_i \sim N(\mu_i, \sigma_i^2)$, i.e. la v.a. representativa del grupo i, y p_i la probabilidad de que la variable resultado tome el valor del estado bajo estudio, relativa al grupo i.

Teorema 35. En caso se requiera hacer una desigual distribución de tratados y controles, se debe de incorporar un parámetro adicional que es el ratio entre grupos $r = (n_1/n_2)$

1. Variable de resultado continua,
$$n = \frac{(r+1)(\sigma_1^2 + \sigma_2^2)[z_{\alpha/2} + z_\beta]^2}{rTE^2}.$$
2. Variable de resultado binaria,
$$n = \frac{(r+1)(p_1(1-p_1)) + p_2(1-p_2)[z_{\alpha/2} + z_\beta]^2}{rTE^2}.$$

Ejemplo 60. Se quiere estimar el tamaño de muestra necesario para evaluar un programa relacionado al lavado de manos en niños y niñas menores de cinco años. Se cuenta con la información que este tipo de programas permiten reducir la incidencia de episodios de diarrea en un 10 % en promedio para los grupos intervenidos. Además, se cuenta con el dato que la prevalencia promedio de episodios de diarrea en niños y niñas menores de cinco años es de 30 % de acuerdo a la Encuesta Demográfica y de Salud Familiar del 2020. Dado lo anterior nos preguntamos ¿cuál es el tamaño de muestra necesario para poder evaluar el programa con un nivel de confianza del 95 % y un poder de análisis del 70 %? Usando (5.8), calculamos
$$n = \left\lfloor \frac{[(0{,}3)(1-0{,}3) + (0{,}2)(1-0{,}2)][1{,}96^2 + 0{,}33^2]}{0{,}1^2} \right\rfloor + 1 = 147.$$

5.5. Bootstrap

Sea F la distribución conjunta de las observaciones $\{(x_i, y_i)\}_{i=1}^n$ y

$$T_n = T_n((x_1, y_1), \cdots, (x_n, y_n))$$

un estadístico de interés[8]. La distribución de T_n es

$$G_n(z, F) = \mathbb{P}\{T_n \leq z | F\}.$$

Idealmente, uno quiere hacer inferencia a partir de $G_n(z, F)$. No obstante, esto no es posible dado que F es desconocido.

Un enfoque, ya abordado, consiste en aproximar $G_n(z, F)$ haciendo $n \to \infty$. Sin embargo, Efron (1979) propone un nuevo método. Este método consiste en reemplazar F_n por F y así obtener $\tilde{G}_n(z) = G_n(z, F_n)$.

Por un lado,

$$F(x, y) = \mathbb{P}\{x_i \leq x, y_i \leq y\} = \mathbb{E}[\mathbf{1}_{\{x_i \leq x\}} \mathbf{1}_{\{y_i \leq y\}}].$$

Por otro lado,

$$F_n(x, y) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \leq x\}} \mathbf{1}_{\{y_i \leq y\}}$$

es la función de distribución empírica (EDF). Note que F_n es un estimador no paramétrico de F. Lo importante es que

$$F_n(x, y) \to F(x, y)$$

en probabilidad. Más aún,

$$\text{Var}(\mathbf{1}_{\{x_i \leq x\}} \mathbf{1}_{\{y_i \leq y\}}) = F(x, y)(1 - F(x, y)).$$

[8] Desde un estimador hasta un $t-$test $\frac{\hat{\theta} - \theta}{\text{sd}(\theta)}$.

Así, por el TLC

$$\sqrt{n}(F_n(y,x) - F(x,y)) \to N(0, F(x,y)(1-F(x,y))).$$

Ejemplo 61. Si $(\tilde{x}_i, \tilde{y}_i)$ es un par aleatorio con distribución $F_n(x,y) = \mathbb{P}\{\tilde{x}_i \leq x, \tilde{y}_i \leq y\}$, podemos calcular

$$\begin{aligned}\mathbb{E}[h(\tilde{x}_i, \tilde{y}_i)] &= \sum_{i=1}^{n} h(x_i, y_i) \underbrace{\mathbb{P}\{\tilde{x}_i = x_i, \tilde{y}_i = y_i\}}_{=1/n} \\ &= \frac{1}{n}\sum_{i=1}^{n} h(x_i, y_i).\end{aligned}$$

Dado un conjunto de datos originales $\{(x_i, y_i)\}_{i=1}^n$, el proceso de bootstrap no paramétrico se puede describir de la siguiente manera. Primero, se construye la función de distribución empírica (EDF), definida como

$$F_n(x,y) = \frac{1}{n}\sum_{i=1}^{n} 1_{\{x_i \leq x\}} \cdot 1_{\{y_i \leq y\}},$$

En la EDF, cada par de observaciones (x_i, y_i) tiene una probabilidad de $\frac{1}{n}$ de ser seleccionado en la muestra bootstrap. Para generar una muestra bootstrap, se seleccionan n pares de observaciones (x_i, y_i) de los datos originales con reemplazo, generando n variables aleatorias $I_1, I_2, \ldots, I_n$ independientes e idénticamente distribuidas (i.i.d.) que siguen una distribución uniforme discreta en el conjunto $\{1, 2, \ldots, n\}$. La muestra bootstrap $\{(x_i^*, y_i^*)\}_{i=1}^n$ se forma seleccionando las observaciones correspondientes a los índices generados: $(x_i^*, y_i^*) = (x_{I_i}, y_{I_i})$. Este proceso se repite $B > 0$ veces para generar B muestras bootstrap. Cada muestra bootstrap es un conjunto

de datos del mismo tamaño que los datos originales y sigue la misma distribución.

Para cada una de las B muestras bootstrap generadas, calculamos el estadístico de interés, denotado como T_n^*. Si el estadístico de interés es por ejemplo la media, se calcula

$$T_n^{*(b)} = \frac{1}{n}\sum_{i=1}^{n} y_i^* \quad \text{para} \quad b = 1, 2, \ldots, B.$$

El sesgo del estimador T_n se calcula como la diferencia entre la media de los estadísticos bootstrap T_n^* y el estadístico calculado a partir de los datos originales T_n:

$$\text{Sesgo} = \frac{1}{B}\sum_{b=1}^{B} T_n^{*(b)} - T_n.$$

La varianza del estimador T_n se calcula como la varianza de los estadísticos bootstrap T_n^*:

$$\text{Var}(T_n) = \frac{1}{B-1}\sum_{b=1}^{B}\left(T_n^{*(b)} - \frac{1}{B}\sum_{b=1}^{B} T_n^{*(b)}\right)^2.$$

Los intervalos de confianza para el estadístico T_n se pueden obtener a partir de la distribución de los estadísticos bootstrap T_n^*. Uno de los métodos comunes es el método percentil, que se basa en los percentiles de la distribución de T_n^*. Para un intervalo de confianza del 95 %, ordenamos los B valores de T_n^* y seleccionamos el percentil 2.5 % y el percentil 97.5 % como los límites inferior y superior del intervalo de confianza. A continuación detallamos este aspecto.

Para una distribución $G_n(z, F)$ sea $q_n(\alpha, F)$ un cuantil, es decir, la función que satisface

$$G_n(q_n(\alpha, F), F) = \alpha.$$

Si tenemos contrapartes de bootstrap $q_n^*(\alpha, F) = q_n^*(\alpha)$, a un $(1-\alpha)\,\%$. Primero, consideremos el cuantil de la distribución original $q_n(\alpha, F)$. Dado que no conocemos la verdadera distribución G_n, utilizamos la función de distribución empírica (EDF) F_n y aplicamos el método bootstrap para generar una distribución estimada. La notación $q_n^*(\alpha, F)$ representa el cuantil α de la distribución estimada mediante bootstrap. Dado que esta estimación se basa en la EDF, podemos simplificar la notación a $q_n^*(\alpha)$, que es el cuantil α estimado por bootstrap sin necesidad de especificar F explícitamente.

A continuación, se discute el tema de los intervalos de confianza en el contexto del bootstrap. El intervalo de confianza de Efron para $T_n = \hat{\theta}$ es $I_1 = [q_n^*(\alpha/2), q_n^*(1-\alpha/2)]$. Según Efron (1979), se recomienda estimar $q_n(\alpha, F)$ mediante el método bootstrap, obteniendo $q_n^*(\alpha, F_n) = q_n^*(\alpha)$. Si $T_n = \hat{\theta}$, el intervalo de confianza se construye con la distribución empírica como

$$I_1 = [q_n^*(\alpha/2); q_n^*(1-\alpha/2)].$$

Este método es intuitivo ya que consiste en cortar las colas de la distribución, aunque carece de una justificación estadística sólida Rau (2016). Si se considera $T_n = \hat{\theta} - \theta$ y $q_n(\alpha)$ como el cuantil α de T_n, se puede estimar $q_n^*(\alpha)$ mediante bootstrap y construir el intervalo de confianza como $I_1 = [\hat{\theta}+q_n^*(\alpha/2); \hat{\theta}+q_n^*(1-\alpha/2)]$, lo cual compensa el desplazamiento inicial. I_1 es la contraparte bootstrap del intervalo de confianza teórico

$$\tilde{I}_1 = [\hat{\theta} + q_n(\alpha/2); \hat{\theta} + q_n(1-\alpha/2)].$$

La probabilidad de que θ_0 esté en $\tilde{I}_1$ es la siguiente:

$$\begin{aligned}\mathbb{P}(\theta_0 \in \tilde{I}_1) &= \mathbb{P}[\hat{\theta} + q_n(\alpha/2) \leq \theta_0 \leq \hat{\theta} + q_n(1-\alpha/2)] \\ &= \mathbb{P}[-q_n(1-\alpha/2) \leq \hat{\theta} - \theta_0 \leq -q_n(\alpha/2)] \\ &= G_n(-q_n(\alpha/2), F_0) - G_n(-q_n(1-\alpha/2), F_0)\end{aligned}$$

Esto es igual a $1-\alpha$ solo si G_n es simétrica. El método de Hall (véase Rau (2016)) aborda el problema de simetría: si $T_n(\theta) = \hat{\theta}-\theta$ y $q_n(\alpha)$ es el cuantil α de T_n, entonces $\mathbb{P}[q_n(\alpha/2) \leq T_n(\theta_0) \leq q_n(1-\alpha/2)] = 1-\alpha$. Esto se traduce a $\mathbb{P}[\hat{\theta}-q_n(1-\alpha/2) \leq \theta_0 \leq \hat{\theta}-q_n(\alpha/2)] = 1-\alpha$. Estimando F mediante bootstrap, se obtiene:

$$I_2 = [\hat{\theta} - q_n^*(1-\alpha/2); \hat{\theta} - q_n^*(\alpha/2)],$$

El cual generalmente no es igual a I_1 a menos que $G_n^*(z)$ sea simétrica respecto a $\hat{\theta}$.

El «Percentile-t Equal-tailed Interval» se aplica para probar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta < \theta_0$ con una significancia de $\alpha\,\%$. Para ello, se construye el estadístico t como $T_n(\theta) = \frac{\hat{\theta}-\theta}{\mathrm{sd}(\hat{\theta})}$. Se rechaza H_0 si $T_n(\theta) < c$, donde c es tal que

$$\mathbb{P}(T_n(\theta_0) < c) = \alpha.$$

Estimando mediante bootstrap, se construye el intervalo:

$$\mathbb{P}[q_n(\alpha/2) \leq \frac{\hat{\theta} - \theta_0}{\mathrm{sd}(\hat{\theta})} \leq q_n(1-\alpha/2)] = 1-\alpha,$$

que se traduce a:

$$I_3 = \left[\hat{\theta} - \mathrm{sd}(\hat{\theta})q_n(1-\alpha/2); \hat{\theta} - \mathrm{sd}(\hat{\theta})q_n(\alpha/2)\right].$$

Finalmente, el «Symmetric Percentile-t Interval» se utiliza para probar $H_0 : \theta = \theta_0$ frente a $H_1 : \theta \neq \theta_0$ con una significancia de $\alpha\,\%$. Se usa $T_n(\theta) = \frac{\hat{\theta}-\theta}{\mathrm{sd}(\hat{\theta})}$ y se rechaza si $|T_n(\theta)| > c$. El valor c se obtiene resolviendo $\mathbb{P}(|T_n(\theta)| > c) = \alpha$, estimado por bootstrap ordenando de manera ascendente $|T_n^*(\theta)| = \frac{|\hat{\theta}^*-\hat{\theta}|}{\mathrm{sd}(\hat{\theta}^*)}$ y tomando el cuantil $1-\alpha$. El intervalo es:

$$I_4 = [\hat{\theta} - \mathrm{sd}(\hat{\theta})q_n^*(\alpha); \hat{\theta} - \mathrm{sd}(\hat{\theta})q_n^*(\alpha)],$$

y se rechaza H_0 si $|T_n(\theta_0)| > q_n^*(\alpha)$. Este método es preferible si no se puede asumir simetría.

Ejemplo 62. El modelo de regresión lineal se puede expresar como

$$Y_i = X_i^T\beta + \epsilon_i, \quad \mathbb{E}[\epsilon_i|X_i] = 0.$$

Para hacer inferencia sobre β mediante bootstrap, se utiliza el método de bootstrap no paramétrico, que remuestrea pares (Y_i^*, X_i^*) de la función de distribución empírica (EDF). Sin embargo, esto no garantiza que $\mathbb{E}[\epsilon_i^*|X_i^*] = 0$, lo cual puede hacer que el estimador sea ineficiente si los supuestos del modelo de regresión se cumplen. Para imponer independencia entre los errores y los regresores remuestreados, se pueden seguir diferentes enfoques. Primero, se obtienen los errores bootstrap ϵ_i^* remuestreando los errores obtenidos mediante MCO $\hat{\epsilon}_i$. Alternativamente, se pueden generar errores bootstrap de una distribución paramétrica, por ejemplo, $N(0, \hat{\sigma}^2)$. Para los regresores, se obtienen X_i^* remuestreando de la EDF de los regresores $\{X_1, X_2, \ldots, X_n\}$, se generan regresores bootstrap de una distribución paramétrica, por ejemplo, $N(\bar{X}, \mathrm{Var}(X))$, o se mantienen

$X_i^* = X_i$, tratando a los regresores como fijos en muestras repetidas. Estos métodos generan errores independientes de los regresores y son válidos bajo el supuesto de homocedasticidad.

Capítulo 6

Multicolinealidad

Desde este capítulo en adelante, la hoja de ruta será sustancialmente diferente a lo abordado previamente. Hasta el momento, se han presentado los detalles de la metodología de estimación fundamental en econometría, así como sus variantes (datos no continuos) y el concepto de muestro, altamente ligado al de la estimación pues permite establecer una base sólida para el análisis empírico. Sin embargo, el presente capítulo así como los siguientes, abordan el levantamiento de los supuestos, enunciados en el Teorema (13). En particular, se analizará en este capítulo el supuesto de la *no multicolinealidad*, es decir, que no existe colinealidad perfecta entre las variables explicativas incluidas en el modelo. Recordemos que la no colinealidad perfecta se expresa matemáticamente de la siguiente manera:

$$\nexists\ \gamma_1, \gamma_2 ..., \gamma_k \neq 0 :\ \gamma_1 X_{1i} + \gamma_2 X_{2i} + ... + \gamma_k X_{ki} = 0.$$

Este supuesto evita que $\det(X^TX) = 0$. Sin embargo, es posible que $\text{Corr}(X_{\ell i}, X_{ji}) \sim 1$, lo que haría que $\det(X^TX) \to 0$. Por ende, como

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

en presencia de multicolinealidad respecto a la variable j,

$$\text{Var}(\hat{\beta}) \cdot e_j \to \infty.$$

6.1. Análisis de la varianza

Dos o más variables independientes se correlacionan fuertemente, por tanto, es difícil poder determinar cual explica la variable dependiente. Caso extremo es que una variable sea combinación lineal de otra (colinealidad perfecta). O sea,

$$X_j = \sum_{i \neq j} \gamma_i X_i.$$

Ejemplo 63. Consideremos el modelo de regresión

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 Z_i + \epsilon_i.$$

Entonces

$$\text{Var}(\hat{\beta}_2) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (X_i - \overline{X})(1 - r_{XZ}^2)}$$

$$\text{Var}(\hat{\beta}_3) = \frac{\hat{\sigma}^2}{\sum_{i=1}^n (Z_i - \overline{Z})(1 - r_{XZ}^2)}$$

con

$$r_{XZ} = \frac{\sum_{i=1}^n (X_i - \overline{X})(Z_i - \overline{Z})}{\sqrt{\sum_{i=1}^n (X_i - \overline{X})} \sqrt{\sum_{i=1}^n (Z_i - \overline{Z})}}.$$

Por ende, si $r_{XZ} \to 1$, $\text{Var}(\beta_i) \to \infty$.

Si $\text{Var}(\hat{\beta})$ aumenta, el error estándar aumenta, y por ende

$$t = \frac{\hat{\beta}}{\text{sd}(\hat{\beta})}$$

disminuye, con lo cual, la significancia individual del regresor asociado, cae.

Ejemplo 64. Se tiene el siguiente modelo:

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 S_{1i} + \beta_4 S_{2i} + u_i$$

donde Y es el salario en soles, X representa la educación medida en años de estudio, S_1 es una variable binaria que toma el valor de 1 si i es hombre y 0 si es mujer, y S_2 es una variable binaria que toma el valor de 1 si i es mujer y 0 si es hombre. ¿Existe algún problema para estimar dicho modelo? Sí. El modelo anterior viola el supuesto de colinealidad pues, dada la matriz

$$X = [\mathbf{1}, X, S_1, S_2] = \begin{bmatrix} 1 & X_1 & S_{11} & S_{21} \\ 1 & X_2 & S_{12} & S_{22} \\ \vdots & \vdots & \vdots & \vdots \\ 1 & X_n & S_{1n} & S_{2n} \end{bmatrix},$$

bajo el modelo anterior, $S_{1i} + S_{2i} = \mathbf{1}$. Por ende, tendremos que $X^T X$ no es invertible. Para arreglar el modelo, los podemos transformar de la manera siguiente

$$Y_i = \beta_1 + \beta_2 X_i + \beta_3 S_{1i} + \epsilon_i,$$

de modo que $S_{1i} = \{1, 0\}$. Más aún, en este caso, $\beta_1 + \beta_2 X_i + \beta_3 = \mathbb{E}[Y_i | S_{1i} = 1]$. Es decir, β_3 equivale a un incremento (o reducción si $\beta_3 < 0$) de los salarios de los hombres respecto a las mujeres.

6.2. Métodos de detección

Ya hemos explicado la importancia del problema de la colinealidad entre las regresores. La pregunta de interés ahora es ¿cómo detectar este problema? Siguiendo el análisis anterior, ciertamente una forma de efectuar aquello es estudiando la matriz de correlaciones de las variables independientes antes de estimar el modelo e identificar si existen variables fuertemente correlacionadas ($r > 0{,}90$). Otra forma es, al momento de realizar la regresión, identificar que los coeficientes de la regresión son no significativos pero el modelo de manera global es significativo. Es decir, t estadísticos bajos pero un F estadístico o R^2 alto. Finalmente, una prueba bastante común es el *Variance Inflation Test*, el cual describimos a continuación.

Definición 6.2.1. Test de Inflación de la Varianza. Sea

$$\text{VIF}_j = \frac{1}{1 - R_j^2}$$

con R_j^2 el coeficiente de determinación de la variable j de la regresión de la variable j en función de las demás variables explicativas. Si $VIF \to \infty$, hay fuertes indicios de multicolinealidad, respecto al regresor j.

Siguiendo el criterio de una correlación mayor a 0.9, usualmente se establece que, para $VIF > 10$ (o sea $R_j^2 > 0{,}9$), el problema de la multicolinealidad está presente en el modelo.

Ejemplo 65. Considere el siguiente conjunto de datos a partir del cual se estima el salario de los individuos en base a su edad en años y a sus años de experiencia.

Salario	Edad	Experiencia
2112	40	10
1967	38	10
1378	27	7
1842	34	9
1512	29	7

Al momento de realizar la regresión donde desea estimar los efectos asociados de la edad y la experiencia en el salario, se encuentra que ninguna de estas variables es significativa a pesar de que habría indicios de que el modelo en conjunto es bueno para estimar dicha variable.

Estadístico	**Valor**
R^2	0.98
$P > F$	0.0135
IC de la constante	[-42.18, 145.42]
IC de β_1	[-334.11, 357.93]
IC de β_2	[-744.56, 595.16]

Cuadro 6.1 Resultados del modelo de regresión

¿Qué es lo que podría estar aconteciendo? Debido a una baja significancia individual pero una alta significancia global, se sospecha de un problema de multicolinealidad (siguiendo los criterios establecidos previamente). Tal y como se sabe, existen fundamentalmente tres formas que nos permiten detectar si se trata de un problema de multicolinealidad.

1. Contrastar la significancia global con la de los parámetros uno por uno, esto es, al momento de realizar la regresión, se observa si los coeficientes de regresión no son significativos pero el modelo de manera global si es significativo, teniendo un R^2 alto.

2. Revisar la matriz de correlaciones de las variables independientes (explicativas). Si hay variables fuertemente correlacionadas ($r > 0{,}90$) se detecta colinealidad.

3. Aplicar el test de inflación de la varianza (VIF). Si el valor hallado es igual o mayor de 10, esto indica que se presenta el problema de colinealidad.

El primer indicio ya está establecido. Queda entonces por verse los otros dos. Primero, se calcula la correlación entre las explicativas (X_i =edad, Z_i =experiencia), con $N = 5$

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Z_i - \overline{Z})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})}\sqrt{\sum_{i=1}^{n}(Z_i - \overline{Z})}}.$$

Se obtiene $r = 0{,}978. > 0{,}90$. Esto implica que (en efecto) se hace frente a un problema de colinealidad. Finalmente, ya sea para reforzar la hipótesis o emplear una forma de detección alternativa, podemos calcular

$$\text{VIF}_j = \frac{1}{1 - R_j^2}.$$

En este caso, se obtiene

$$\frac{1}{1 - 0{,}9865} \sim 74 > 10.$$

Por ende, se refuerza la idea de que existe un problema de colinealidad entre la edad y la experiencia.

6.3. Soluciones ante casos de multicolinealidad

Ya hemos visto que la multicolinealidad es un problema de gran importancia a la hora de efectuar la estimación de los coeficientes en una regresión lineal, debido al efecto sobre la varianza de estos últimos. No obstante, queda la duda de cómo afrontar este problema, o si para empezar, es posible. La respuesta a esta última interrogante es afirmativa, existen diversas maneras de lidiar con la multicolinealidad, y ese será el tema de esta breve sección. De manera concisa, para solucionar la multicolinealidad, es posible:

1. Replantear el modelo a estimar, eliminando una de las variables que ocasionan el problema de colinealidad.

2. En caso de tener un reducido número de observaciones, se puede incrementar el tamaño de la muestra. De esta manera, se espera que las observaciones adicionales permitan eliminar la dependencia entre los regresore X_j.

3. Transformar las variables, ya sea diferenciándolas ($Z_{ji} = X_j - X_i$) o ponderando ($Z_j = w_j X_j$) las variables en función de alguna de las variables del modelo.

4. Generar índices sintéticos con las variables que presentan alta colinealidad (por ejemplo: sumatoria normalizada, componentes principales, percentiles, ranking).

Definición 6.3.1. Un indicador sintético es la combinación de dos o más indicadores simples o individuales (por ejemplo: años de

escolaridad o ingreso per capita). De esta forma, los indicadores sintéticos permiten condensar la información de un grupo de variables altamente correlacionadas que reflejan un aspecto latente detrás. En otras palabras son una suma ponderada de los diferentes indicadores simples empleados. Matemáticamente,

$$IS_i = \sum_{j=1}^{n}\sum_{i=1}^{k} w_i X_{ij}, \ \sum_{i=1}^{k} X_i = 1,$$

n el tamaño de muestra respecto a los regresores en cuestión.

El inconveniente más señalado respecto a los indicadores sintéticos es la determinación de los pesos w_i. Usualmente, o se le asigna el mismo valor a cada peso, i.e., $w_i = 1/k$, o se fijan pesos de manera ad-hoc, o bien en función de la correlación entre las variables[1].

Finalmente, es importante mencionar que es posible efectuar otro tipo de transformaciones afines[2] (a parte de las diferencias y ponderación) usando $X_{(1)}$ y $X_{(n)}$ (i.e., $X_{\text{mín}}$ y $X_{\text{máx}}$) o $\overline{X}$ y $\sqrt{S_X}$. Esto último consiste en normalizar cada una de las variables para que varíen entre los valores de 0 y 1. Así, al poner a las variables en una misma escala es posible sumarlas y generar un indicador sintético. Para ello, se computa

$$ZX_i = \frac{X_i - X_{\text{mín}}}{X_{\text{máx}} - X_{\text{mín}}} \triangleq \frac{X_i - X_{(1)}}{X_{(n)} - X_{(n)}}$$

o bien (normalización clásica):

$$ZX_i = \frac{X_i - \overline{X}}{\sqrt{S_X}}.$$

[1] Mayor o menor correlación, mayor o menor el valor de $- < w_i \leq 1$

[2] $x \to ax + b$.

Capítulo 7

Estabilidad de los parámetros estimados

En la estimación de los parámetros en el caso de enumeración temporal (series de tiempo), i.e., $X_i = X_t, t \leq n$, nos interesamos en si los parámetros del modelo son los mismos para todo t. En efecto, cuando estimamos una relación usando el MCO se considera que el efecto marginal es fijo o el mismo para los diferentes periodos de tiempo. Sin embargo, puede existir t^* tal que $t \leq t^*$, $\hat{\beta} = \hat{\alpha}$ y para $t > t^*$, $\hat{\beta} = \hat{\gamma}$. Este tipo de cambio estructural en los parámetros puede deberse a eventos externos significativos, cambios en la política económica o cambios en el comportamiento de los agentes económicos. Identificar y modelar adecuadamente estos puntos de cambio es crucial para obtener estimaciones precisas. Métodos como las pruebas de Chow Chow (1960) y la técnica de regresiones segmentadas Bai and Perron (1998) son comúnmente utilizados para detectar cambios estructurales en series de tiempo.

Observemos por ejemplo el siguiente gráfico.

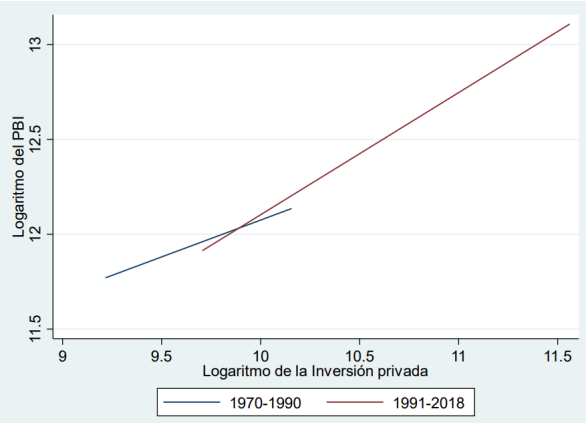

Figura 7.1 ln(PBI) vs ln(Inversión privada).

En la Figura (7.1), se observa que la pendiente, dada por β tal que

$$\ln(\text{PBI})_t = \alpha + \beta \ln(\text{Inversión privada})_t,$$

no es la misma si consideramos $t \in [1970 - 1990]$, a si consideramos $t \in [1991 - 2018]$. Ciertamente el parámetro β va a depender de la muestra que se tome, o sea, β depende de I, donde $t \in I \subset \{t_0, t_1, ..., T\}$. Sin embargo, nos preguntamos hasta que punto esto sería meramente explicado por un factor aleatorio al considerar periodos diferentes, i.e., cuando esta diferencia es estadísticamente significante.

Definición 7.0.1. El tiempo t^* es conocido como punto de quiebre, y se puede originar por fenómenos naturales, crisis financieras, implementación de Tratados de Libre Comercio, entre otros.

Hay dos formas de determinar el punto de quiebre t^*:

1. Fijarlo de manera arbitraria basado en conocimiento previo sobre el contexto del país o tema que se está analizando. Por ejemplo, teniendo en cuenta una crisis económica-sanitaria (COVID 19), vigencia de TLC (Tratados de Libre Comercio), etc.

2. El método de residuos recursivos, detallado en la siguiente sección.

7.1. Residuos Recursivos

Los residuos recursivos son un método para analizar la estabilidad de los parámetros β_j. Este método consiste en estimar el modelo MCO de un modo recursivo, es decir, aumentando la muestra paulatinamente. El error de predicción de Y_t se obtiene de la siguiente manera

$$\hat{\beta}_{t-1} = (X_{t-1}^T X_{t-1})^{-1} X_{t-1}^T Y_{t-1}.$$

Esto es, los parámetros estimados con el set

$$\{X_{1,1}, ..., X_{1,t-1}, ..., X_{k,1}, ..., X_{k,t-1}\}.$$

Entonces, el error de predicción para la observación t sería

$$v_t = Y_t - \hat{Y}_t = Y_t - X_t \hat{\beta}_{t-1}. \tag{7.1}$$

Luego, a partir de (7.1), computamos el residuo recursivo normalizado w_t:

$$w_t = \frac{Y_t - X_t\hat{\beta}_{t-1}}{\sqrt{\sigma^2(1 + X_t^T(X_{t-1}^T X_{t-1})^{-1})X_t}} \tag{7.2}$$

$$= \frac{v_t}{\sqrt{\sigma^2(1 + X_t^T(X_{t-1}^T X_{t-1})^{-1})X_t}}, \tag{7.3}$$

$t = k + 1, ..., n$.

A la hora de computar (7.2), se tiene en cuenta que

$$\begin{aligned}\mathbb{E}[v_t|X_t] &= \mathbb{E}[Y_t - X_t\hat{\beta}_{t-1}|X_t] \\ &= 0\end{aligned}$$

y, que

$$\mathrm{Var}(v_t) = \sigma^2(1 + X_t^T(X_{t-1}^T X_{t-1})^{-1}X_t).$$

Enseguida, definimos lo que será la herramienta principal para determinar la significancia estadística de desvíos sistemáticos en los β.

Definición 7.1.1. Cumulative Sum Control Chart: CUSUM. Es la suma acumulada de los residuos normalizados, y el CUSUM cuadrado consiste en emplear los cuadrados de los residuos normalizados. Ambos estadísticos permiten comprobar desviaciones no aleatorias o desvíos sistemáticos.

$$\mathrm{CUSUM} = \frac{1}{\hat{\sigma}} \sum_{j=k+1}^{n} w_j, \ \hat{\sigma} = \frac{\mathrm{SRC}}{n-k}$$

$$\mathrm{CUSUM}_t^2 = \frac{\sum_{j=k+1}^{t} w_j^2}{\sum_{j=k+1}^{n} w_j^2}, \ t = k+1, ..., n.$$

En el caso del CUSUM, el estadístico debe estar alrededor de 0 en caso no existan desvíos sistemáticos dado que el esperado de los errores es 0. En el caso del CUSUM cuadrado, el estadístico oscila entre 0 y 1. Los paquetes estadístico trabajan con estos dos estadísticos, y, en función de estos, permiten determinar la significancia estadística del cambio en los parámetros estimados. Concretamente, se proveen bandas de confianza, ilustradas en las siguientes dos figuras.

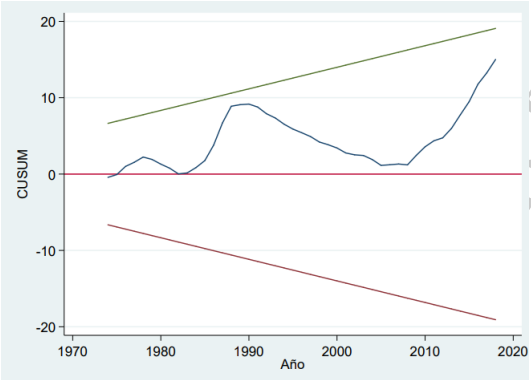

Figura 7.2 $\sum_{j=k+1}^{n} w_j$.

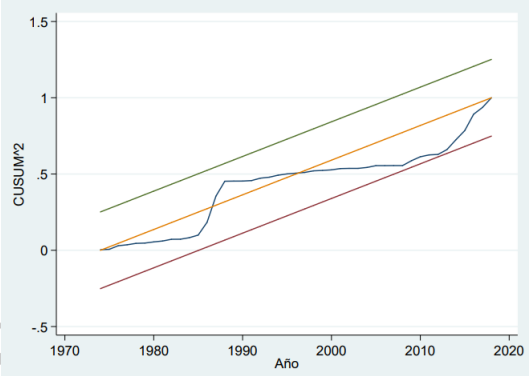

Figura 7.3 $\frac{\sum_{j=k+1}^{t} w_j^2}{\sum_{j=k+1}^{n} w_j^2}$.

En caso el CUSUM o CUSUM cuadrado se *salga* de las bandas, se confirma la significancia estadística del cambio estructural en los parámetros (o sea, que efectivamente hay un quiebre). Queda entonces claro que cuando se trabaja con datos que son series temporales, puede pasar que exista un cambio structural en la relación entre la variable dependiente y las independientes. Este cambio estructural se puede deber a causas exógenas (por ejemplo: fenómeno del niño, epidemia), o debido a cambios en la política pública de un país (por ejemplo: cambio en el sistema de conversión del tipo de cambio) u otra causa exógena. A continuación, presentamos el test estadístico por excelencia a la hora de determinar cambios estructurales en los parámetros.

7.2. Test de Chow

Primero, se estima la regresión con todos los datos y se obtiene la sumatoria de residuos al cuadrado SRC_{CR}[1]

$$\sum_t (\hat{Y}_t - Y_t)^2.$$

Los grados de libertad de este estadístico es $n_1 + n_2 - k - 1$[2], con k el número de explicativas en el modelo. En este modelo, se acepta que los coeficientes son iguales en ambos periodos, por lo que sería el modelo con restricciones. En segundo lugar, se estiman los modelos para cada uno de los periodos donde se espera que haya quiebre y se guardan la sumatoria de residuos al cuadrado

[1] También llamado suma de cuadrados residuales SCR.

[2] Siendo n_1 y n_2 el número de periodos en cada una de las 2 etapas.

de cada uno. Los grados de libertad de cada estadístico serán el número de observaciones, menos la cantidad de parámetros menos 1 (constante). Esto es

$$\text{SRC}_{SR} = \text{SRC}_{SR1} + \text{SRC}_{SR2},$$

con $n_1 + n_2 - 2k - 2$ grados de libertad. Finalmente, definimos el estadístico

$$F = \frac{\dfrac{\text{SRC}_{CR} - \text{SRC}_{SR}}{k+1}}{\dfrac{\text{SRC}_{SR}}{n_1 + n_2 - 2k - 2}}. \tag{7.4}$$

Con los instrumentos previamente calculados, ya es posible definir en qué consiste el test de Chow, detallado en la siguiente definición.

Definición 7.2.1. Test de Chow. El estadístico (7.4) se compara con F_{k+1,n_1+n_2-2k-2} (valor de tablas o teórico). Acá k es el número de parámetros (no cuenta la constante). La hipótesis nula es que los parámetros son iguales ambos periodos (no hay cambio estructural). Si $F_{\text{calculado}} > F_{\text{tablas}}$ se rechaza la hipótesis nula.

Luego de aplicar el test de Chow, en función de los resultados (si se determina el cambio estructural), el modelo original puede ser sustituido por un modelo con dummies. Esto es, si tenemos 2 periodos de tiempo, $D_i = 0 \vee 1$ (0 para el primer periodo, 1 para el periodo dos). El modelo sería entonces

$$\begin{aligned} Y_t = \beta_0 + \beta_1 X_{1t} + \beta_2 X_{2t} + ... + \beta_k X_{kt} + \gamma_0 D_t \\ + \sum_{i=1}^{k} \gamma_i D_i X_{it} + \epsilon_t. \end{aligned}$$

Se incluye $\gamma_0 D_t$ para analizar un cambio en el intercepto.

Ejemplo 66. Un investigador quiere estimar el ahorro en función del ingreso disponible, para ello cuenta con datos anuales del ahorro personal (S) y el ingreso disponible (I) para el periodo 1988-2005. Luego, estima el siguiente modelo:

$$S_t = \beta_1 + \beta_2 I_t + \epsilon_t.$$

El investigador obtiene los siguientes resultados,

Variable	Coeficiente	Error estándar
Constante	-1.082	0.145
I_t	0.118	0.009
R^2	0.92	-
SRC	0.572	-

Tomando en consideración que a mediados del año 1997 se dio una grave crisis financiera en Asia que afectó a varios países, el investigador plantea la posibilidad de un cambio estructural en dicho año. Por lo que, se desea analizar si ocurrió un cambio solo en intercepto, o solo en pendiente, o en intercepto y pendiente a la vez. Teniendo en cuenta los siguientes datos

	Periodo 1988-1996	1997-2005
$\sum(I_t - \overline{I})^2$	28.2622	89.62
$\sum(S_t - \overline{S})^2$	0.2022	2.2217
$\sum(I_t - \overline{I})(S_t - \overline{S})$	1.3291	13.4833

es posible, al 95 % de confianza determinar si ocurrió, a la vez, un cambio en el intercepto y en la pendiente. En efecto, basta con

aplicar el test de Chow.

$$\begin{aligned}
\text{SRC} &= \sum_{t=1}^{n}(S_t - \hat{S}_t)^2 \\
&= \sum_{t=1}^{n}(S_t - \overline{S})^2 - \sum_{t=1}^{n}(\hat{S}_t - \overline{S})^2 \\
&= \sum_{t=1}^{n}(S_t - \overline{S})^2 - \sum_{t=1}^{n}(\hat{\beta}_1 + \hat{\beta}_2 I_t - \hat{\beta}_2\overline{I} - \hat{\beta}_1)^2 \\
&= \sum_{t=1}^{n}(S_t - \overline{S})^2 - \hat{\beta}_2^2 \sum_{t=1}^{n}(I_t - \overline{I})^2 \\
&= \sum_{t=1}^{n}(S_t - \overline{S})^2 - \left[\frac{\sum_{t=1}^{m}(I_t - \overline{I})(S_t - \overline{S})}{\sum_{t=1}^{n}(I_t - \overline{I})^2}\right]^2 \sum_{t=1}^{n}(I_t - \overline{I})^2.
\end{aligned}$$

Note que se ha usado el hecho que $\hat{\beta}_1 = \overline{S} - \hat{\beta}_2\overline{I}$ y que $\hat{\beta}_2 = \frac{\sum_{t=1}^{m}(I_t-\overline{I})(S_t-\overline{S})}{\sum_{t=1}^{n}(I_t-\overline{I})^2}$. De ahí,

$$\begin{aligned}
\text{SRC}_{SR_1} &= 0{,}1396 \\
\text{SRC}_{SR_2} &= 0{,}1931 \\
\text{SRC}_{SR} &= 0{,}3327
\end{aligned}$$

y

$$F_{\text{estadístico}} = \frac{\frac{\text{SRC}_{CR} - \text{SRC}_{SR}}{1+1}}{\frac{\text{SRC}_{SR}}{18-2-2}} = 5{,}039 > F_{\text{tablas.}}$$

Se rechaza entonces la H_0, i.e., sí hubo un cambio estructural. Veamos ahora si ocurrió un cambio solo en la pendiente o en ambas y no solo en la pendiente. Para ello, definimos la siguiente variable dicotómica

$$D_t = \begin{cases} 1, & \text{si } t \in [1988, 1996) \\ 0, & t \in [1996, T] \end{cases}$$

y consideremos los siguientes modelos.

Modelo	Especificación	SRC
Sin cambios	$S_t = \beta_1 + \beta_2 I_t + u_t$	0.572
Cambio en β_2	$S_t = (\beta_1 + \beta_3 D_t) + (\beta_2 + \beta_4 D_t) I_t + u_t$	0.332
Cambio en β_1	$S_t = (\beta_1 + \beta_3 D_t) + \beta_2 I_t + u_t$	0.563

Para poder realizas los exámenes estadísticos, recordemos que $F(0{,}95; 1; 14) = 4{,}6$, $F(0{,}95; 1; 15) = 4{,}54$, $F(0{,}95; 2; 14) = 3$. Ahora, contrastamos el modelo 1 con el 3, es decir, $H_0 : \beta_3 = 0$. El estadístico F es

$$F = \frac{\frac{\text{SRC}_R - \text{SRC}_I}{q}}{\frac{\text{SRC}_I}{N-m}},$$

donde

1. SRC_R es la suma de cuadrados residuales en el modelo restricto.
2. SRC_I es la suma de cuadrados residuales en el modelo irrestricto.
3. q el número de restricciones.
4. N el número de observaciones y m el número de parámetros en el modelo irrestricto.

De ahí, como $q = 1$, $N = 18$ y $m = 3$,

$$F = \frac{\frac{0{,}572 - 0{,}362}{1}}{\frac{0{,}562}{15}} = 0{,}257.$$

Así, como $0{,}257 < 4{,}54$, no se rechaza la hipótesis nula, la pendiente no cambia. Finalmente, para contrastar los modelos 2

y 3, procediendo de manera análoga, pero teniendo en cuenta que $m = 4$,

$$F = \frac{\frac{0{,}562-0{,}362}{1}}{\frac{0{,}332}{14}} = 9{,}67 > 4{,}6.$$

Así, se rechaza la hipótesis nula, hay efectivamente un cambio en pendiente.

Capítulo 8

Heterocedasticidad

Una forma informal pero concisa de resumir el contenido trabajado y el que se va a trabajar en este capítulo, es la siguiente: *se estudia el modelo $k-$lineal y progresivamente, se van levantando los supuestos*. Ya hemos analizado el aspecto del muestreo y la multicolinealidad. En este capítulo, nos enfocamos en las propiedades de los errores ϵ. Concretamente, en aspectos relacionados a la normalidad de los errores y su varianza. Recordemos que

1. $\mathbb{E}\left[\sum_{i=1}^{n} \epsilon_i\right] = 0$.
2. $\text{Var}(\epsilon_i^2) = \sigma^2$, $\forall\ i = 1, ..., n$.

Nos preguntamos primero, ¿qué sucede cuando los errores no tienen varianza constante? Al asumir la normalidad del vector de parámetros

$$\hat{\beta} \sim \mathcal{N}(\beta, \sigma^2(X^T X)^{-1})$$

se pueden efectuar los tests estadísticos sobre estos últimos:

1. $t-$Student.

2. F-Fisher.

3. χ^2.

Más aún, en caso de normalidad, se tiene entonces que

$$f(Y|X\beta,\sigma^2)=\frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(Y-X\beta)^T(Y-X\beta)}{2\sigma^2}}.$$

Es decir, la variable dependiente condicionada al conjunto de variables independientes sigue una distribución normal. El hecho que los errores sigan una distribución normal permite asegurar que su varianza es constante y que la media de los mismos es cero.

La hoja de ruta es la siguiente. En primer lugar, se explorará de que manera es posible identificar la propiedad de normalidad de los errores. Luego, se estudiarán formas de corregir la heterocedasticidad.

8.1. Tests de normalidad

Una variable aleatoria X que sigue una distribución normal se caracteriza por una serie de propiedades. Primero, requerimos de algunas definiciones.

Definición 8.1.1. Sea F una ley de distribución relativa al conjunto de datos $\{y_1,...,y_n\}$. El $Q-Q$ plot es la representación gráfica de los cuantiles teóricos de F

$$F^{-1}\{1/(n+1)\},\ F^{-1}\{2/(n+1)\},...,F^{-1}\{n/(n+1)\}.$$

versus los estadístico de orden $y_{(1)}\leq y_{(2)}\leq ...\leq y_{(n)}$.

Teorema 36. En caso F sea la distribución de una normal $\mathcal{N}(\mu, \sigma^2)$. Entonces, el $Q-Q$ plot no es nada menos que una recta de intercepto μ y pendiente σ.

Demostración. Sea

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt.$$

Luego,

$$\begin{aligned} \Phi_{\mu,\sigma^2}(x) &= \int_{-\infty}^{x} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(u-\mu)^2}{2\sigma^2}} du \\ &= \frac{1}{2}\left[1 + \text{erf}^{-1}\left(\frac{x-\mu}{\sigma\sqrt{2}}\right)\right], \ x \in \mathbb{R}. \end{aligned}$$

Así,

$$\Phi^{-1} = \mu + \sigma\sqrt{2}\text{erf}^{-1}(2p-1), \ p \in (0,1).$$

□

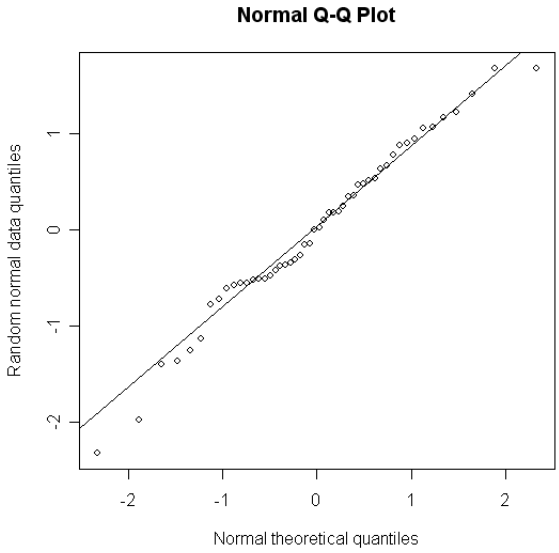

Figura 8.1 Gráfico $Q-Q$ normal de datos $N(0,1)$ generados aleatoriamente .

Si bien la representación gráfica es bastante útil, no es un método exacto, de máxima precisión e infalible[1]. Por ello, se introducen una serie de contrastes estadísticos fundamentados en las siguientes propiedades.

Teorema 37. Si $X \sim \mathcal{N}(\mu_X, \sigma_X^2)$:

1. $\mathbb{E}[X] = \mu_X$.
2. $\text{Var}[X] = \mathbb{E}[(X-\mu_X)^2] = \sigma_X^2$.
3. $\overline{\mu}_3 = \mathbb{E}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^3\right] = 0$.

[1]Después de todo, es un análisis visual.

4. $\overline{\mu}_4 = \mathbb{E}\left[\left(\frac{X-\mu_X}{\sigma_X}\right)^4\right] = 3.$

En relación al Teorema (37):

1. La notación μ_X y σ_X hace énfasis en el hecho que X puede venir de una muestra aleatoria.

2. $\overline{\mu}_3 = S$ se conoce como la asimetría de X.

3. $\overline{\mu}_4 = \kappa$ se conoce como la curtosis de X.

4. De manera general, si $X \sim F$, tal y como se precisa en el apéndice de teoría de la probabilidad,

$$\mathbb{E}[X^n] = \begin{cases} \sum x^n \mathbb{P}(x) & \text{si la variable es discreta.} \\ \int x^n dF(x), & \text{si la variable es continua.} \end{cases} \tag{8.1}$$

Usando (8.1), es posible computar S y κ.

¿De qué manera podemos analizar si $\{X_i\}_{i=1}^n$ posee la propiedades señaladas en el Teorema (37)? Aplicando alguno de los siguientes tres tests estadísticos

1. Jarque-Bera.

2. Kolmogorov-Smirnov.

3. Shapiro-Wilks.

Veamos en qué consisten dichos contrastes estadísticos.

Definición 8.1.2. Jarque-Bera. La prueba estadística de Jarque-Bera es una prueba de *bondad de ajuste* cuyo objetivo es comprobar si una muestra de datos $\{X_i\}_{i=1}^n$ tiene la asimetría y la curtosis de una distribución normal. Se definen:

- La hipótesis nula $H_0 : S = 0$ y $\kappa = 3$ (la muestra se distribuye normalmente).
- La hipótesis alternativa $H_1 : S \neq 0$ y/o $\kappa \neq 3$.
- El estadístico

$$JB = \frac{n}{6}\left[S^2 + \frac{(\kappa - 3)^2}{4}\right] \simeq \chi^2(2)$$

donde n es el número de observaciones, S la asimetría y κ la curtosis.

A un nivel de significancia del 5 % el estadístico χ^2 tiene como valor crítico o de tablas de 5.99. Entonces, si el valor del estadístico estimado es menor al de tablas, se acepta la hipótesis nula de normalidad de la variable; en caso contrario, se rechaza la nula y la variable seria no normal.

Ejemplo 67. Sea $S = 0{,}06$ y $\kappa = 3{,}39$. Entonces, para $n = 7111$

$$JB = \frac{7111}{6}\left[(0{,}06)^2 + \frac{(3{,}39 - 3)^2}{4}\right] = 49{,}30 > 5{,}99.$$

Así, se rechaza la H_0: la muestra no se distribuye normalmente.

Definición 8.1.3. Kolmogorov-Smirnov. Sean $X_1, ..., X_n$ iid, que toman valores en $\mathbb{R}$ y cuya función de distribución

$$F_X(x) = \mathbb{P}(X \leq x)$$

es F. Recordemos que la función de distribución empírica (EDF) de la muestra F_n se define como

$$F_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{\{X_i \leq x\}}.$$

Luego, definimos

$$D_n = \sup_{x \in \mathbb{R}} |F_n(x) - F(x)|.$$

De este modo, si F es la distribución de una normal, es posible analizar si la muestra se distribuye según una normal si $D_n \to 0$.

Definición 8.1.4. Shapiro-Wilks. El estadístico de contraste del test de Shapiro-Wilks es

$$W = \frac{\left(\sum_{i=1}^{n} a_i X_{(i)}\right)^2}{\sum_{i=1}^{n}(X_i - \overline{X})^2}$$

donde $\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$, $X_{(i)}$ el i-ésimo estadístico de orden y los coeficientes a_i se calculan de la siguiente manera:

$$(a_1, ..., a_n) = \frac{m^T \Sigma^{-1}}{C},\ C = ||\Sigma^{-1} m|| = \sqrt{m^T \Sigma^{-1} \Sigma^{-1} m}$$

donde $m = (m_1, ..., m_n)^T$ los valores esperados de los estadísticos de orden de variables aleatorias independientes e idénticamente distribuidas según una ley normal, y Σ es la matriz de varianzas y covarianzas de dichos estadísticos de orden.

La H_0 se rechaza si $W \to 0$. El valor de W puede oscilar entre 0 y 1. Para efectuar el contraste, se usan los valores de tabla, que se calculan vía métodos más avanzados que escapan de los temas abordados en este texto.

Los contrastes estadísticos Jarque-Bera, Shapiro-Wilks y Kolmogorov-Smirnov nos permiten identificar la normalidad de los errores aplicando dichos tests a los errores $\hat{\epsilon}_i = Y_i - \hat{Y}_i$. En caso no exista normalidad en los residuos, uno de los causantes es la presencia

de *heterocedasticidad*. ¿Qué origina la heterocedasticidad? ¿qué es? y ¿cómo detectarla? Estas interrogantes serán respondidas en la siguiente sección.

8.2. Métodos de detección de heterocedasticidad

En los modelos de regresión lineales se dice que hay heterocedasticidad cuando la varianza de los errores no es igual en todas las observaciones realizadas. Matemáticamente, es la negación del siguiente enunciado

$$\text{Var}(\epsilon_i|X) = \sigma^2, \ \forall \ i = 1, ..., n.$$

Las causas más comunes de la heterocedasticidad son

1. Una mala especificación del modelo: regresores omitidos.
2. Forma funcional incorrecta en las variables usadas en el modelo: una de las variables tiene una relación no lineal con la dependiente.
3. Un cambio estructural puede provocar una estimación errónea de los parámetros. Esto se produce en algunas secciones de la muestra y genera diversos problemas en el modelo.

Note que el último ítem ya fue analizado de forma extensiva previamente. Queda entonces únicamente una interrogante: ¿cómo se detecta la heterocedasticidad? Existen (por fortuna) diversos métodos:

1. Un análisis gráfico de los errores estimados al cuadrado en función de la muestra.

2. La prueba de Park.

3. La prueba de Glesjer.

4. La prueba de Breuch-Pagan-Godfrey.

5. La prueba de White.

El análisis gráfico consiste en estudiar si existe alguna relación directa o patrón de los errores estimados al cuadrado $\hat{\epsilon}^2$ en función de X.

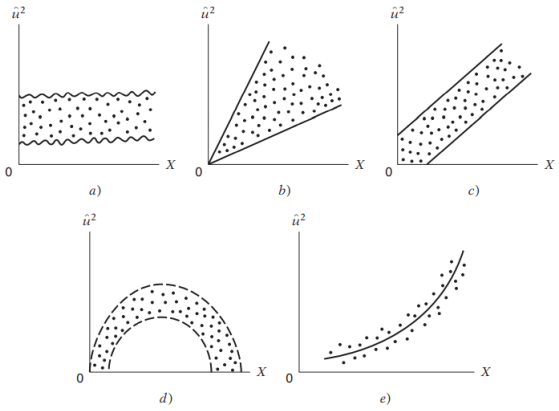

Figura 8.2 Heterocedasticidad: correlación errores. La figura ha sido extraída de Gujarati and Porter (2010).

La presencia de un patrón son indicios de heterocedasticidad. Si bien el análisis gráfico provee una buena primera forma de detectar el problema, no constituye un método formal. Por ello, se recurre a las pruebas estadísticas que detallaremos a continuación.

Definición 8.2.1. Prueba de Park. Primero, se estima vía MCO

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i.$$

Enseguida, se calcula

$$\hat{\epsilon} = Y - \hat{Y}.$$

Esto es

$$\hat{\epsilon}_i = Y_i - \hat{Y}_i, \ \forall \ i = 1, ..., n.$$

Enseguida, se efectúa la siguiente regresión

$$\ln \hat{\epsilon}_i^2 = \alpha_0 + \ln X_{1i} + ... + \alpha_p \ln X_{pi} + \epsilon_i$$

donde u_i un error aleatorio "well behaved"(normalmente distribuido), y $p \leq k$. Si los coeficiente son significativos, se confirma la presencia de heterocedasticidad.

Ejemplo 68. Supongamos que se tiene el siguiente modelo Y =salario promedio en miles de dólares y X =productividad promedio en miles de dólares

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i.$$

Vía MCO se obtiene

$$\hat{Y}_i = 1992{,}3452 + 0{,}2329 X_i.$$

Los parámetros $\hat{\beta}_1 = 1992{,}3452$ y $\hat{\beta}_2 = 0{,}2329$ tienen respectivamente error estándar

$$ee = (936{,}4791)(0{,}0998).$$

Luego, $t = \hat{\beta}/ee(\hat{\beta})$. Así,

$$t = (2{,}1275)(2{,}333),\ R^2 = 0{,}4375.$$

Los resultados revelan que el coeficiente de la pendiente estimado es significativo en el nivel de 5 % con base en una prueba t de una cola (umbral en 1.96). La ecuación muestra que, a medida que aumenta la productividad laboral, por ejemplo, en un dólar, el salario aumenta, en promedio, alrededor de 23 centavos de dólar. Ahora, en la regresión de los residuos sobre la explicativa, se obtiene

$$\ln \hat{\epsilon}_i^2 = 35{,}817 - 2{,}8099 \ln X_i$$

$$ee = (38{,}319)(4{,}216)$$

$$t = (0{,}934)(-0{,}667),\ R^2 = 0{,}0595.$$

No hay una relación estadísticamente significativa entre ambas variables. Según la prueba de Park, se puede concluir que no hay heterocedasticidad en la varianza del error.

Definición 8.2.2. Prueba de Glesjer. La prueba de Glesjer consiste en realizar un análisis de significancia a los parámetros estimados del siguiente modelo

$$|\hat{\epsilon}_i| = \alpha_0 + \alpha_1 X_{1i} + ... + \alpha_p X_{pi} + u_i.$$

Acá, $u_i \sim N(0, \sigma^2)$.

A veces se efectúa la regresión

$$|\hat{\epsilon}_i| = \alpha_0 + \alpha_1 X_{1i}^{\gamma_1} + ... + \alpha_p X_{pi}^{\gamma_p} + v_i,\ \gamma_i = \pm 1/2.$$

Definición 8.2.3. Prueba de Breusch-Pagan-Godfrey. La prueba de prueba de Breusch-Pagan-Godfrey consiste en lo siguiente. Dado el modelo de regresión lineal

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i,\ i = 1, ..., n,$$

suponemos que

$$\sigma_i^2 = f(\alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_p Z_{pi})$$

con $Z_{ij} \in \{X_{ij}\}_{1 \leq j \leq n}$. Es decir, una función no estocástica de un subconjunto de los regresores. En particular, tomamos $f(\cdot)$ lineal

$$\sigma_i^2 = \alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_p Z_{pi}.$$

Si $\alpha_j = 0, \forall\ 1 \leq j \leq n$, los errores son homocedásticos. Ahora bien, queda la interrogante ¿de qué manera se implementa la prueba usando lo datos (dado que los σ_i son parámetros)? Se procede de la siguiente manera

- Se estima vía MCO el modelo y se obtienen los $\hat{\epsilon}_i$.
- Se calcula $\hat{\sigma}_2 = \frac{\sum_{i=1}^n \hat{u}_i^2}{n-k}$.
- Se define $\gamma_i = \frac{\hat{\epsilon}_i^2}{\hat{\sigma}_2}$.
- Efectuamos la regresión

  $$\gamma_i = \alpha_0 + \alpha_1 Z_{1i} + ... + \alpha_p Z_{pi} + u_i$$

  con el objetivo de calcular la $SCE = \sum_{i=1}^n (\hat{\gamma}_i - \overline{\gamma})^2$.
- Enseguida, definimos

  $$\Theta = \frac{SCE}{2} \sim \chi^2(p-1).$$

- Si $\Theta = \frac{SCE}{2} > \chi^2_{1-\alpha}(p-1)$ se rechaza la hipótesis nula (homocedasticidad).

Definición 8.2.4. Test de White. Dado el modelo

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \epsilon_i,\ i = 1, ..., n,$$

se estiman vía MCO los errores $\hat{\epsilon}_i^2 = (Y_i - \hat{Y}_i)^2$. Enseguida, se plantea la regresión

$$\hat{\epsilon}_i^2 = \alpha_0 + \alpha_1 X_{1i} + \alpha_2 X_{2i} + \alpha_3 X_{1i}^2 + \alpha_4 X_{2i}^2 + \alpha_5 X_{1i} X_{2i} + u_i.$$

De este modelo se obtiene el R^2 y se define el estadístico $nR^2 \sim \chi^2(q)$ (en este caso $q = 5$). Si $nR^2 > \chi^2(q)$, se rechaza $H_0 : \alpha_j = 0$, $\forall\ j$.

Las pruebas de Park, Glesjer y Breusch-Pagan-Godfrey son metodologías que permiten analizar la heterocedasticidad de los errores. Sin embargo, en la práctica, es el test de White, el que se usa con mayor frecuencia. Esto se explica debido a lo siguiente:

- En los tests de Park y Glesjer se requiere saber qué variables causan la heterocedasticidad.
- La prueba de Glesjer requiere normalidad en los residuos.
- Irónicamente, los errores v_i en el caso de Park y Glesjer pueden ser heterocedásticos o presentar autocorrelación serial.
- Las pruebas de Breusch-Pagan-Godfrey y de White no requieren que se conozca la fuente de heterocedasticidad y tampoco requieren normalidad en los residuos.

- La prueba de White detecta interacción entre las explicativas en relación al término de error. Revelando un posible problema de multicolinealidad.

Ya se poseen las herramientas para detectar el problema de heterocedasticidad. Una vez detectada, ¿cómo corregirla? Esta interrogante será abordada a continuación.

8.3. Métodos para corregir la heterocedasticidad

Recordemos que el problema de heterocedasticidad se debe a que las varianzas no son constantes

$$\text{Var}(\epsilon_i) = \sigma_i^2 \ , \ 1 \leq i \leq n.$$

Sin embargo, de momento, $\text{Cov}(\epsilon_i, \epsilon_j) = 0$. De forma matricial, se puede entonces expresar el problema de heterocedasticidad de la siguiente manera:

$$\mathbb{E}[\epsilon\epsilon^T] = \Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & \sigma_n^2 \end{pmatrix}.$$

Efectuando la siguiente transformación, la matriz de varianzas y covarianzas se escribe como

$$\Sigma = \sigma^2 \begin{pmatrix} \omega_1 & 0 & \cdots & 0 \\ 0 & \omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & \omega_n \end{pmatrix} = \sigma^2 \Omega. \tag{8.2}$$

El término ω_i es el causante de la heterocedasticidad en los errores estimados. ¿Qué implica esta situación para los parámetros estimados mediante MCO? Recordemos una vez más que en el modelo de $k-$ variables, se tenía

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

A pesar del problema de heterocedasticidad, $\mathbb{E}[\hat{\beta}] = \beta$. Sin embargo,

$$\text{Var}(\hat{\beta}) = \sigma^2 (X^T X)^{-1} X^T \Omega X (X^T X)^{-1} \neq \sigma^2 (X^T X)^{-1}. \tag{8.3}$$

En resumen, si bien la presencia de heterocedasticidad no introduce sesgo en los parámetros estimados, si origina problemas en la validez de las inferencias estadísticas dado que el estimador deja de ser eficiente.

A partir de dicha información, expresada por las ecuaciones (8.2) y (8.3), y siguiendo White (1980), se ejecutan los siguientes pasos para corregir el problema de heterocedasticidad. La estrategia consiste en transformar el modelo original de tal manera que los coeficientes estimados no cambien y solo sea la matriz de varianzas y covarianzas del modelo la que cambie, de tal forma que los nuevos errores estimados tengan varianza media 0 y varianza constante.

Para lograr este fin, se pre-multiplica a todas las variables del modelo de regresión por una matriz P de dimensión $n \times n$.

$$\begin{aligned} Y &= X\beta + \epsilon \text{ modelo original} \\ PY &= P(X\beta) + P\epsilon \text{ modelo transformado} \\ Y^* &= X^*\beta + \epsilon^*. \end{aligned}$$

Con la transformación realizada, se puede apreciar que $Y^*(= PY)$ y $\epsilon^*(= P\epsilon)$ siguen siendo vectores de dimensión $n \times 1$ con la diferencia que ahora cada observación de la variable Y^* es una combinación lineal de las n observaciones del vector Y, encontrándose los coeficientes de dichas combinaciones lineales en la matriz P. Lo anterior también es aplicable para la matriz de variables explicativas y los errores del modelo original. Por otro lado, ninguna de estas nuevas variables tienen un significado económico claro; sin embargo, dado el supuesto de linealidad, se tiene que los coeficientes β_j siguen siendo los mismos del modelo original.

En relación a la matriz de varianzas y covarianzas de los errores, ahora se tiene

$$\text{Var}(\epsilon) = \text{Var}(P\epsilon) = P\text{Var}(\epsilon)P^T = \sigma^2 P\Omega P^T.$$

Así, el objetivo es que $P\Omega P^T = I_n$.

Ahora bien, como Ω es una matriz simétrica, por la descomposición de Cholesky, existe una matriz cuadrada V triangular superior, tal que $\Omega = VV^T$. Así, debemos tener $PVV^TP^T = I_n$, de donde, $P = V^{-1}$.

Definición 8.3.1. Estimador de Mínimos Cuadrados Generalizados. El estimador de Mínimos Cuadrados Generalizados es

$$\hat{\beta}^* = (X^{*T}X^*)^{-1}X^{*T}Y^*. \tag{8.4}$$

Así, usando que $X^* = V^{-1}X$ y $Y^* = V^{-1}Y$

$$\begin{aligned}\hat{\beta}^* &= (X^T(V^{-1})^TV^{-1}X)^{-1}X^T(V^{-1})^TY \\ &= (X^T\Omega^{-1}X)^{-1}X^T\Omega^{-1}Y.\end{aligned}$$

Entonces conociendo la fuente de la heterocedasticidad Ω, se puede corregir el modelo.

Teorema 38. El estimador de Mínimos Cuadrados Generalizados es

- Es insesgado: $\mathbb{E}[\hat{\beta}^*] = \beta$.
- Tiene matriz de varianzas y covarianzas : $\text{Var}(\hat{\beta}^*) = \sigma^2(X^T\Omega^{-1}X)^{-1}$.

Hay que encontrar entonces la matriz Ω que será igual a VV^T. Una vez encontrada V, se pre-multiplica el modelo de regresión por la inversa de esta matriz y luego se estima el modelo MCO con las variables transformadas. Si se conoce la matriz Ω simplemente se reemplaza en las formulas halladas para la estimación de los coeficientes y la varianza.

Ejemplo 69. Considérese el modelo lineal

$$Y_i = \beta_0 + \sum_{j=1}^{k} \beta_j X_{ji} + \epsilon_i.$$

La matriz de varianzas y covarianzas de los errores es

$$\mathrm{Var}(\epsilon_i) = \begin{pmatrix} \sigma_1^2 & & & 0 \\ & \sigma_2^2 & & \\ & & \ddots & \\ 0 & & & \sigma_n^2 \end{pmatrix}.$$

Luego,

$$\Sigma = \begin{pmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_n \end{pmatrix} \begin{pmatrix} \sigma_1 & & & 0 \\ & \sigma_2 & & \\ & & \ddots & \\ 0 & & & \sigma_n \end{pmatrix} = VV^T.$$

$$V^{-1} = \begin{pmatrix} 1/\sigma_1 & & & 0 \\ & 1/\sigma_2 & & \\ & & \ddots & \\ 0 & & & 1/\sigma_n \end{pmatrix}.$$

Así

$$X^* = V^{-1}X = \begin{pmatrix} 1/\sigma_1 & X_{11}/\sigma_1 & \cdots & X_{k1}/\sigma_1 \\ 1/\sigma_2 & X_{12}/\sigma_2 & \cdots & X_{k2}/\sigma_n \\ \vdots & & & \vdots \\ 1/\sigma_n & X_{1n} & \cdots & X_{kn}/\sigma_n \end{pmatrix}$$

y

$$Y^* = V^{-1}Y = \begin{pmatrix} Y_1/\sigma_1 \\ Y_2/\sigma_2 \\ \vdots \\ Y_n/\sigma_n \end{pmatrix}.$$

Con lo cual, a partir de (8.4)

$$\hat{\beta}_{MCG} = \underbrace{\begin{pmatrix} \sum_{i=1}^n \frac{1}{\sigma_i} & \sum_{i=1}^n \frac{X_{1i}}{\sigma_i} & \cdots & \sum_{i=1}^n \frac{X_{ki}}{\sigma_i} \\ \vdots & \sum_{i=1}^n \frac{X_{1i}^2}{\sigma_i^2} & \cdots & \sum_{i=1}^n \frac{X_{1i}X_{ki}}{\sigma_i^2} \\ \vdots & & \ddots & \vdots \\ \sum_{i=1}^n \frac{1}{\sigma_i} & & & \sum_{i=1}^n \frac{X_{ki}^2}{\sigma_i^2} \end{pmatrix}^{-1}}_{(X^{*T}X^*)^{-1}} \underbrace{\begin{pmatrix} \sum_{i=1}^n \frac{Y_i}{\sigma_i^2} \\ \sum_{i=1}^n \frac{X_{1i}Y_i}{\sigma_i^2} \\ \vdots \\ \sum_{i=1}^n \frac{X_{ki}Y_i}{\sigma_i^2} \end{pmatrix}}_{=X^{*T}Y}.$$

Ejemplo 70. Considere el siguiente modelo heterocedástico

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,\ n = 5,\ \mathrm{Var}(\epsilon_i) = \sigma^2 X_i.$$

Y es el gasto en salud anual y X la renta anual de las familias. Lo que nos estaría indicando las forma de los errores es que las familias de rentas altas son las que tienen mayor variación en el gasto por salud, a diferencia de las familias de bajos niveles de ingresos. Supóngase que obtenemos los siguientes datos vía estimación MCO:

Familia (i)	Gasto Y_i	Ingreso X_i
1	7.0	10
2	12.8	20
3	18.3	35
4	25.3	50
5	33.4	60

Nuestro objetivo es obtener $\hat{\beta}_{MCG}$. Primero, identificamos las matrices Σ, Ω y V.

$$\Sigma = \begin{pmatrix} \sigma X_1 & & & 0 \\ & \sigma X_2 & & \\ & & \ddots & \\ 0 & & & \sigma X_5 \end{pmatrix},\ \Omega = \begin{pmatrix} X_1 & & & 0 \\ & X_2 & & \\ & & \ddots & \\ 0 & & & X_5 \end{pmatrix}$$

$$V = \begin{pmatrix} \sqrt{X_1} & & & 0 \\ & \sqrt{X_2} & & \\ & & \ddots & \\ 0 & & & \sqrt{X_5} \end{pmatrix}, \; V^{-1} = \begin{pmatrix} 1/\sqrt{X_1} & & & 0 \\ & 1/\sqrt{X_2} & & \\ & & \ddots & \\ 0 & & & 1/\sqrt{X_5} \end{pmatrix}.$$

De ahí, es posible computar directamente $\hat{\beta}_{MCG}$ haciendo

$$\hat{\beta}_{MCG} = ((V^{-1}X)^T(V^{-1}X))^{-1}((V^{-1}X)V^{-1}Y).$$

Sin embargo, haciendo el cambio de variable $Y^* = \frac{Y}{\sqrt{X}}$, $X = \frac{1}{\sqrt{X}}$, $\epsilon^* = \frac{\epsilon}{\sqrt{X}}$ y $\beta_0^* = \frac{1}{\sqrt{X}}$:

$$\begin{aligned} \text{Var}(\epsilon^*) &= \text{Var}\left(\frac{\epsilon}{\sqrt{X}}\right) \\ &= \frac{1}{X}\text{Var}(\epsilon) \\ &= \frac{\sigma^2 X}{X} \\ &= \sigma^2. \end{aligned}$$

Aplicando este cambio de variable, se aplica directamente el la fórmula usual de los estimadores $\beta = (X^TX)^{-1}X^TY$ (pero post transformación). Los datos transformados proveen la siguiente tabla.

Familia (i)	Gasto Y_i^*	Ingreso X_i^*
1	2.214	3.162
2	2.862	4.472
3	3.093	5.916
4	3.678	7.071
5	4.312	7.746

Finalmente,

$$\hat{\beta}_{MCG} = (X^{*T}X^*)^{-1}X^{*T}Y^* = \begin{pmatrix} 2{,}125 \\ 0{,}496 \end{pmatrix}.$$

Puede ocurrir que no se conozca la posible fuente de heterocedasticidad en nuestro modelo; i.e., no se conoce Σ. En estos casos, se puede usar el estimador de White:

$$\text{Var}(\hat{\beta}) = (X^TX)^{-1}X^T\mathbb{E}[\epsilon\epsilon^T]X(X^TX)^{-1}.$$

Usamos entonces el siguiente estimador White (1980), para $\frac{X^T\Omega X}{n}$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{\epsilon}_i^2 X_i X_i^T.$$

White demostró que esta estimación puede realizarse de forma que las inferencias estadísticas sean asintóticamente válidas.

Ejemplo 71. Dado el modelo

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i, \ , \ \text{Var}(\epsilon_i) = \sigma_i^2$$

como los σ_i^2 no son directamente observados, para calcular

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2\sigma_i^2}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2}$$

se usa $\hat{\epsilon}_i$ en vez de σ_i^2

$$\text{Var}(\hat{\beta}_1) = \frac{\sum_{i=1}^{n}X_i^2\hat{\epsilon}_i^2}{\left(\sum_{i=1}^{n}(X_i - \overline{X})^2\right)^2}.$$

En el caso general

$$Y_i = \beta_0 + \beta_1 X_{1i} + ... + \beta_k X_{ki} + \epsilon_i, \ , \ \text{Var}(\epsilon_i) = \sigma_i^2.$$

La varianza para cada coeficiente es

$$\text{Var}(\hat{\beta}_j) = \frac{\sum_{i=1}^{n} \hat{w}_{ij}\hat{\epsilon}_i^2}{\left(\sum_{i=1}^{n} \hat{w}_{ij}^2\right)^2}$$

con $\hat{w}_j$ los errores de la regresión de X_j en función de los demás regresores.

A continuación, brindamos un ejemplo más completo que permite sintetizar lo abordado en esta sección.

Ejemplo 72. Consideremos la siguiente especificación de función de producción que depende únicamente del capital y el empleo:

$$Y_i = \beta_0 + \beta_1 K_i + \beta_2 L_i + \epsilon_i. \tag{8.5}$$

Acá Y_i representa el valor agregado de la empresa medido en millones de soles, K_i a los activos fijos de la empresa en millones de soles y L_i al número de trabajadores en millones de personas. Recordemos que el test de White tiene como objetivo analizar si se presenta el problema de heterocedasticidad, i.e., la varianza de los errores no es constante en las observaciones:

$$\text{Var}(\epsilon_i) \neq \sigma^2 \ , \ \forall \ 1 \leq i \leq n.$$

Usando la regresión de la ecuación (8.5), el test de White consiste en efectuar los siguientes pasos:

1. Se estima la regresión

$$Y_i = \beta_0 + \beta_1 K_i + \beta_2 L_i + \epsilon_i,$$

vía MCO, y se obtienen los errores estimados $\hat{\epsilon}_i = (Y_i - \hat{Y}_i)$.

2. Luego, se elevan al cuadrado: $\hat{\epsilon}_i^2 = (Y_i - \hat{Y}_i)^2$.

3. Se efectúa la regresión de estos errores al cuadrado versus las variables explicativas iniciales, sus interacciones, y sus cuadrados. Para este modelo, dicha regresión sería:

$$\hat{\epsilon}_i^2 = \alpha_0 + \alpha_1 K_i + \alpha_2 L_i + \alpha_3 K_i L_i + \alpha_4 K_i^2 + \alpha_5 L_i^2 + u_i, \quad (8.6)$$

donde u_i es un término de error, aleatorio y *bien comportado*. Se estima (8.6) vía MCO y se obtiene el R^2.

4. Se define el estadístico nR^2, con n el número de observaciones. Este, sigue una distribución $\chi^2(q)$, con q el número de términos en la regresión (sin contar la constante). En este caso, $q = 5$.

5. Planteamos la hipótesis nula H_0 : los errores no presentan el problema de heterocedasticidad.

6. Si $nR^2 > \chi^2_{1-\alpha}(q)$, se rechaza la H_0. Caso contrario, se acepta. Acá α es la significancia.

La siguiente tabla resume el examen estadístico necesario para obtener la información requerida:

```
White's test for Ho: homoskedasticity
         against Ha: unrestricted heteroskedasticity

         chi2(5)      =     58.17
         Prob > chi2  =    0.0000

Cameron & Trivedi's decomposition of IM-test
```

Source	chi2	df	p
Heteroskedasticity	58.17	5	0.0000
Skewness	16.02	2	0.0003
Kurtosis	7.23	1	0.0072
Total	81.42	8	0.0000

Figura 8.3 Test de White vía Stata.

Con estos datos, es ahora posible graficar el examen estadístico:

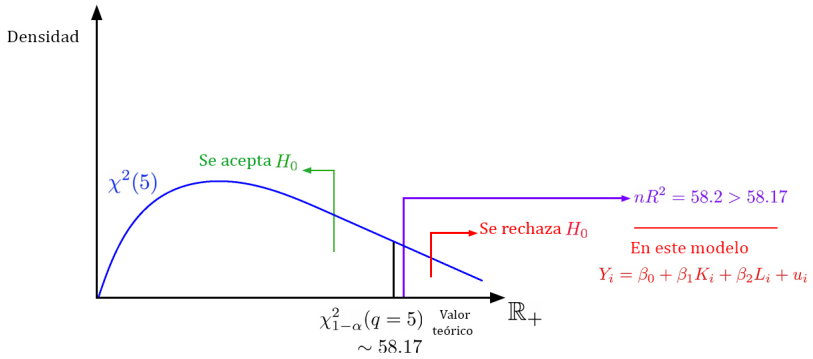

Figura 8.4 Gráfica del examen estadístico.

A partir de esta información y los siguientes resultados, es posible concluir sobre la heterocedasticidad.

```
. reg resid2 K L K2 L2 KL
```

Source	SS	df	MS		
Model	43.779011	5	8.7558022	Number of obs	= 733
Residual	507.834927	727	.698534976	F(5, 727)	= 12.53
				Prob > F	= 0.0000
Total	551.613938	732	.753570954	R-squared	= 0.0794
				Adj R-squared	= 0.0730
				Root MSE	= .83578

resid2	Coef.	Std. Err.	t	P>|t|	[95% Conf.	Interval]
K	-1.829969	.3402904	-5.38	0.000	-2.498038	-1.161899
L	.6620759	.3920022	1.69	0.092	-.1075156	1.431667
K2	.0677918	.013448	5.04	0.000	.0413901	.0941934
L2	.0652812	.0249826	2.61	0.009	.0162346	.1143279
KL	-.0840858	.0320042	-2.63	0.009	-.1469174	-.0212541
_cons	13.9769	2.266904	6.17	0.000	9.526437	18.42735

Figura 8.5 Regresión para los $\hat{u}_i^2$.

Supongamos ahora que, hay presencia de heterocedasticidad en el modelo (no necesariamente en el conjunto de datos anterior). Concretamente, $\text{Var}(\epsilon_i) = cL_i^2$, $c > 0$. Luego,

$$\Sigma = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & \sigma_N^2 \end{pmatrix} = \begin{pmatrix} cL_1^2 & 0 & \cdots & 0 \\ 0 & cL_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & cL_N^2 \end{pmatrix}.$$

Factorizando la constante c

$$\Sigma = c \begin{pmatrix} L_1^2 & 0 & \cdots & 0 \\ 0 & L_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & L_N^2 \end{pmatrix} = c\Omega.$$

Recordemos ahora que, debido a que Ω es simétrica (por la descomposición de Cholesky),

$$\Omega = VV^T.$$

Dada la presencia de términos nulos en todas las entradas de la matriz a excepción de la diagonal, es posible identificar que dicha matriz V estará definida de la siguiente manera

$$V = \begin{pmatrix} L_1 & 0 & \cdots & 0 \\ 0 & L_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & L_N \end{pmatrix}.$$

Asimismo[2],

$$P = V^{-1} = \begin{pmatrix} 1/L_1 & 0 & \cdots & 0 \\ 0 & 1/L_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 1/L_N \end{pmatrix}.$$

En efecto,

$$PP^T = P^2 = \begin{pmatrix} 1/L_1^2 & 0 & \cdots & 0 \\ 0 & 1/L_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 1/L_N^2 \end{pmatrix}$$

y

$$\Omega \cdot \begin{pmatrix} 1/L_1^2 & 0 & \cdots & 0 \\ 0 & 1/L_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 1/L_N^2 \end{pmatrix} = I_N.$$

[2]De manera general, si $D \in \mathcal{M}_{N\times N}$ es una matriz diagonal $D = (a_{11}, a_{22}, ..., a_{NN})$, si $\forall\ i$: $a_{ii} \neq 0$,

$$D^{-1} = (a_{11}^{-1}, a_{22}^{-1}, ..., a_{NN}^{-1}).$$

De este modo, la matriz P cumple la propiedad deseada, i.e. : $PP^T = \Omega^{-1}$. Ejecutemos ahora la transformación del modelo

$$\begin{aligned} Y^* &= PY \\ X^* &= PX = P[\mathbf{1}, K, L] \\ \epsilon^* &= P\epsilon. \end{aligned}$$

Calculamos

$$Y^* = \begin{pmatrix} 1/L_1 & 0 & \cdots & 0 \\ 0 & 1/L_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 1/L_N \end{pmatrix} \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_N \end{pmatrix} = \begin{pmatrix} Y_1/L_1 \\ Y_2/L_2 \\ \vdots \\ Y_N/L_N \end{pmatrix}$$

$$X^* = \begin{pmatrix} 1/L_1 & 0 & \cdots & 0 \\ 0 & 1/L_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \cdots & 1/L_N \end{pmatrix}_{N\times N} \begin{pmatrix} 1 & K_1 & L_1 \\ 1 & K_2 & L_2 \\ \vdots & \vdots & \vdots \\ 1 & K_N & L_N \end{pmatrix}_{N\times 3}$$

$$= \begin{pmatrix} 1/L_1 & K_1/L_1 & 1 \\ 1/L_2 & K_2/L_2 & 1 \\ \vdots & \vdots & \vdots \\ 1/L_N & K_N/L_N & 1 \end{pmatrix}_{N\times 3}$$

$$\epsilon^* = \begin{pmatrix} 1/L_1 & 0 & \cdots & 0 \\ 0 & 1/L_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & 0 \\ 0 & \cdots & \cdots & 1/L_N \end{pmatrix} \begin{pmatrix} \epsilon_1 \\ \epsilon_2 \\ \vdots \\ \epsilon_N \end{pmatrix} = \begin{pmatrix} \epsilon_1/L_1 \\ \epsilon_2/L_2 \\ \vdots \\ \epsilon_N/L_N \end{pmatrix}.$$

El nuevo modelo tendría entonces la siguiente formulación

$$Y_i^* = \beta_0 L_i^{-1} + \beta_1 (K_i/L_i) + \beta_2 + \epsilon_i^*$$
$$y_i = \beta_0 L_i^{-1} + \beta_1 k_i + \beta_2 + \epsilon_i^*.$$

La transformación del modelo permite corregir el problema de heterocedasticidad. En efecto,

$$\text{Var}(\epsilon_i^*) = \text{Var}\left(\frac{\epsilon_i}{L_i}\right) = \frac{1}{L_i^2}\text{Var}(\epsilon_i) = \frac{cL_i^2}{L_i^2} = c \ , \ \forall \ 1 \leq i \leq N.$$

Ahora, el nuevo modelo, que consiste en una transformación de escala a una función de producción lineal $Y = \beta_0 + \beta_1 K + \beta_2 L + u$, expresa la producto medio en función del ratio capital-trabajo. Este ratio es de sumo interés pues define la proporción del trabajo destinada al uso del stock de capital (maquinaría etc.). En otras palabras, define la asignación entre los factores de producción. Como los parámetros iniciales β's deben ser positivos (las funciones de producción son cóncavas crecientes), el factor $1/L$ indica como el producto medio se incrementa cada vez menos conforme el factor trabajo aumenta[3]. Esto confirma de cierta forma las condiciones de Inada[4], supuesto clave sobre la función de producción en los modelos de crecimiento.

[3] La función $f(L) = 1/L$, $L > 0$ $f : \mathbb{R}_{++} \to \mathbb{R}_{++}$ es decreciente en L.
[4] Véase la definición en Barro and Martin (2003)

Capítulo 9

Autocorrelación serial

El problema de una varianza no constante en los errores ya fue abordado en el capítulo anterior. Sin embargo, se precisó que en dicho caso, la matriz de varianza y covarianzas toma la forma $\Sigma = \text{diag}[\sigma_1^2, ..., \sigma_n^2]$. Esto a su vez implica que $\text{Cov}(\epsilon_i, \epsilon_j) = 0,\ i \neq j$. En este capítulo, nos interesemos en el caso en el que los datos provienen de una serie de tiempo. Esto es, una sucesión de datos medidos en ciertos momentos y ordenados cronológicamente. Usualmente, una serie de tiempo se denota de la siguiente forma

$$\{X_t :\ t \in [t_0, T]\}.$$

Recordemos que uno de los supuestos del modelo de regresión lineal es que los errores no tienen autocorrelación serial, es decir, no existe correlación entre los errores de diferentes periodos de tiempo.

Definición 9.0.1. La autocorrelación serial se da cuando los errores en el tiempo t no tienen covarianza nula con los errores

de tiempos pasados. Esto es

$$\text{Cov}(\epsilon_t, \epsilon_{t-1}) = \mathbb{E}[\epsilon_t \epsilon_{t-1}] - \underbrace{\mathbb{E}[\epsilon_t]}_{=0}\mathbb{E}[\epsilon_{t-1}] \neq 0.$$

Como es costumbre, la primera pregunta que surge en estos casos (a la hora de levantar un supuesto del modelo $k-$lineal) es qué es lo que origina el problema. En el caso de la auto-correlación serial, esta puede tener como origen lo siguiente:

- La omisión de variables relevantes en el modelo.
- La existencia de ciclos o tendencias.
- Presencia de relaciones no lineales.
- Uso de modelos autoregresivos (la variable dependiente, depende de sus rezagos).

Ahora, la segunda interrogante natural es ¿qué implica la autocorrelación? El estimador $\hat{\boldsymbol{\beta}}$ sigue siendo insesgado, pero deja de ser eficiente:

$$\text{Var}(\epsilon) = \mathbb{E}[\epsilon\epsilon^T] = \begin{bmatrix} \sigma^2 & \sigma_{21} & \cdots & \sigma_{T1} \\ \sigma_{21} & \sigma^2 & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{1T} & \cdots & \cdots & \sigma^2 \end{bmatrix}. \tag{9.1}$$

Note que en (9.1) se ha asumido homocedasticidad.

La hoja de ruta es la siguiente. En este capítulo se estudiarán esencialmente dos tipos de modelos en los cuales la matriz de varianzas y covarianzas toma la forma de (9.1). Luego, se presentarán los contrastes de autocorrelación más usados. Finalmente, se brindarán ejemplos, así como métodos correctivos.

9.1. Modelo autorregresivo AR

En general, la correlación entre momentos diferentes del tiempo no se limita a dos periodos sucesivos, sino que se mantiene para cualquier distancia entre esos dos momentos del tiempo. Esto se conoce como **Modelo Autorregresivo** (de orden p) o **AR**(p).

$$\begin{aligned}\epsilon_t &= \phi_1\epsilon_{t-1} + \phi_2\epsilon_{t-2} + \cdots + \phi_p\epsilon_{t-p} + u_t \\ &= u_t + \sum_{j=1}^{p} \phi_j\epsilon_{t-j}\end{aligned}$$

con $u_t \sim N(0, \sigma_u^2)$. El modelo $AR(1)$ es un caso particular que tiene la forma

$$\epsilon_t = \rho\epsilon_{t-1} + u_t \ , \ u_t \sim N(0, \sigma_u^2) \ , \ |\rho| < 1.$$

En particular, cuando $t \to \infty$,

$$\epsilon_t = \sum_{i=0}^{t} \rho^i u_{t-i} \implies \lim_{t\to\infty} \epsilon_t = \sum_{i=0}^{\infty} \rho^i u_{t-i}.$$

Ahora, el valor esperado del error ϵ_t es igual a cero:

$$\mathbb{E}[\epsilon_t] = \mathbb{E}\left[\sum_{i=0}^{t} \rho^i u_{t-i}\right] = \sum_{i=0}^{t} \rho^i \mathbb{E}[u_{t-i}] = 0.$$

Por otro lado, la varianza converge a $\frac{\sigma^2}{1-\rho^2}$:

$$\begin{aligned}\text{Var}(\epsilon_t) &= \lim_{t\to\infty} \text{Var}\left(\sum_{i=0}^{t} \rho^i u_{t-i}\right) \\ &= \sum_{i=0}^{\infty} \text{Var}\left(\rho^i u_{t-i}\right) \\ &= \sum_{i=0}^{\infty} \rho^{2i}\text{Var}(u_{t-i}) \\ &= \sum_{i=0}^{\infty} \rho^{2i}\sigma^2 = \frac{\sigma^2}{1-\rho^2}.\end{aligned}$$

Finalmente, la covarianza de ϵ_t con ϵ_{t-s}, $1 \leq s \leq t$, es $\frac{\rho^s\sigma^2}{1-\rho^2}$. En efecto:

$$\begin{aligned}\mathbb{E}[\epsilon_t\epsilon_{t-1}] &= \mathbb{E}[(\rho\epsilon_{t-1} + u_t)\epsilon_{t-1}] \\ &= \mathbb{E}[\rho\epsilon_{t-1}^2 + u_t\epsilon_{t-1}] \\ &= \rho\mathbb{E}[\epsilon_{t-1}^2] + \mathbb{E}[u_t\epsilon_{t-1}] \\ &= \rho\text{Var}[\epsilon_{t-1}] = \rho\frac{\sigma^2}{1-\rho^2}.\end{aligned}$$

Usando que $\text{Cov}(\epsilon_t, \epsilon_{t-s}) = \mathbb{E}[\epsilon_t\epsilon_{t-s}]$ y

$$\epsilon_t = \rho^s\epsilon_{t-s} + \sum_{i=0}^{s-1} \rho^i u_{t-i},$$

se tiene:

$$\begin{aligned}\text{Cov}(\epsilon_t, \epsilon_{t-s}) &= \mathbb{E}\left[\left(\rho^s\epsilon_{t-s}^2 + \sum_{i=0}^{s-1} \rho^i u_{t-i}\right)\epsilon_{t-s}\right] \\ &= \mathbb{E}\left[\rho^s\epsilon_{t-s}^2\right] + \mathbb{E}\left[\epsilon_{t-s}\sum_{i=0}^{s-1} \rho^i u_{t-i}\right] \\ &= \rho^s\mathbb{E}[\epsilon_{t-s}^2] = \frac{\rho^s\sigma^2}{1-\rho^2}.\end{aligned}$$

Se ha considerado que $t - s \to \infty$. En efecto,

$$\mathbb{E}\left[\epsilon_{t-s}\sum_{i=0}^{s-1}\rho^i u_{t-i}\right] = \sum_{i=0}^{s-1}\rho^i \mathbb{E}[u_{t-i}\epsilon_{t-s}] = 0.$$

Así,

$$\text{Var}(\epsilon) = \sigma^2\Omega = \frac{\sigma^2}{1-\rho^2}\begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \cdots & & \vdots \\ \rho^2 & & \ddots & & \vdots \\ \vdots & & & \ddots & \rho \\ \rho^{T-1} & \cdots & \cdots & \rho & 1 \end{bmatrix} \tag{9.2}$$

Las expresiones obtenidas previamente son aproximaciones. En efecto, la dimensión de la matriz (9.2) es $T \times T$. Sin embargo, en los cálculos anteriores, se ha asumido que $T \to \infty$.

9.2. Modelo de medias móviles MA

El modelo de medias móviles de orden q tiene la siguiente estructura

$$\epsilon_t = u_t + \sum_{i=1}^{q}\theta_i u_{t-i}$$

donde los θ_i son los parámetros del modelo, y u_{t-i} son términos de error. Los modelos de media móvil o de memoria finita solo mantienen la correlación entre períodos de tiempo determinados. En un modelo de medias móviles de orden 1,

$$\epsilon_t = u_t + \theta u_{t-1}\ ,\ u_t \sim N(0, \sigma^2).$$

A diferencia de un proceso $AR(1)$, no se necesita imponer supuesto alguno sobre el coeficiente asociado a los errores rezagados[1]. Por un lado,

$$\mathbb{E}[\epsilon_t] = \mathbb{E}[u_t + \theta u_{t-1}] = \mathbb{E}[u_t] + \theta\mathbb{E}[u_{t-1}] = 0.$$

Por otro lado, la varianza es igual a

$$\begin{aligned}\mathbb{E}[\epsilon_t^2] &= \mathbb{E}[(u_t + \theta u_{t-1})^2] \\ &= \mathbb{E}[u_t^2] + 2\theta \underbrace{\mathbb{E}[u_t u_{t-1}]}_{=0} + \theta^2 \underbrace{\mathbb{E}[u_{t-1}^2]}_{=\mathrm{Var}(u_t)} \\ &= \sigma^2(1+\theta^2).\end{aligned}$$

Por inducción, es posible probar que $\mathbb{E}[\epsilon_t\epsilon_{t-1}] = \theta\sigma^2$ y $\mathbb{E}[\epsilon_t\epsilon_{t-j}] = 0$, $\forall\ j \geq 2$. Así,

$$\mathrm{Var}(\epsilon) = \mathbb{E}[\epsilon\epsilon^T] = \sigma^2 \begin{bmatrix} 1+\theta^2 & \theta & 0 & \cdots & 0 \\ \theta & 1+\theta^2 & \theta & \cdots & \vdots \\ 0 & \theta & 1+\theta^2 & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \theta \\ 0 & \cdots & \cdots & \theta & 1+\theta^2 \end{bmatrix}$$

9.3. Contrastes estadísticos de detección

Ya habiendo trabajado los dos modelos representativos del problema de la autocorrelación serial, presentamos los contrastes estadísticos, enfocándonos en los modelos autorregresivos. En efecto, los contrastes que realizaremos, nos permitirán identificar la

[1] Recordemos que $|\rho| < 1$.

presencia de un comportamiento autorregresivo $AR(p)$ o de medias móviles $MA(q)$[2]. Esencialmente, se cuentan con las siguientes pruebas:

- Durbin-Watson.
- Ljung-Bonx.
- Breusch-Godfrey.

A continuación, usaremos la notación $T = n$.

Definición 9.3.1. Durbin Watson. El test de Durbin Watson Durbin and Watson (1950) pone a prueba la existencia de un comportamiento autorregresivo de los errores de orden 1. Es decir, evidencia un $AR(1)$.

$$\epsilon_t = \rho\epsilon_{t-1} + u_t \ , \ u_t \sim N(0, \sigma^2).$$

En este test $H_0 : \rho = 0$ y $H_1 : \rho \neq 0$, $|\rho| < 1$. En caso $\rho > 0$, se dice que la autocorrelación es positiva. Caso contrario, es negativa.

Luego, el estadístico de Durbin-Watson (DW) es

$$\begin{aligned} DW &= \frac{\sum_{t=2}^{n}(\epsilon_t - \epsilon_{t-1})^2}{\sum_{t=1}^{n}\epsilon_t^2} \\ &= \frac{\sum_{t=2}^{n}(\epsilon_t^2 - 2\epsilon_t\epsilon_{t-1} + \epsilon_{t-1}^2)}{\sum_{t=1}^{n}\epsilon_t^2} \\ &= 1 - 2r + \frac{\epsilon_n^2 - \epsilon_n^2 + \sum_{t=2}^{n}\epsilon_{t-1}^2}{\sum_{t=1}^{n}\epsilon_t^2} \\ &= 1 - 2\xi + 1 - \frac{\epsilon_n^2}{\sum_{t=1}^{n}\epsilon_t^2} \sim 2(1-r) = d. \end{aligned}$$

[2]En general estos modelos son más complejos dado que los errores son no observables.

Acá $\xi = \frac{\sum_{t=2}^{n} \epsilon_t \epsilon_{t-1}}{\sum_{t=2}^{n} \epsilon_t^2}$. A partir del valor del estadístico d, se contrasta la hipótesis nula usando la siguiente tabla.

Hipótesis nula	Si	Decisión
No hay autocorrelación positiva	$0 < d < d_L$	Rechazar
No hay autocorrelación positiva	$d_L \leq d \leq d_U$	Sin decisión
No hay correlación negativa	$4 - d_L < d < 4$	Rechazar
No hay correlación negativa	$4 - d_U \leq d \leq 4 - d_L$	Sin decisión
No hay autocorrelación	$d_U < d < 4 - d_U$	No rechazar

Los valores d_L y d_U dependen de n y el número de regresores. Ahora bien, las principales limitaciones del contraste Durbin-Watson son:

- Sólo es válido para la autocorrelación de la perturbación autorregresiva de orden 1.
- Requiere $n > 15$.
- Presenta zonas (rango de valores para d) de indeterminación.

Definición 9.3.2. Ljung-Box. Este test Ljung and Box (1978) utiliza el coeficiente de correlación simple y sólo puede ser aplicado cuando el conjunto de variables explicativas son todas exógenas. La hipótesis nula es que no existe autocorrelación serial. El estadístico Ljung-Box es

$$Q = n(n+2) \sum_{i=1}^{r} \frac{\rho_i^2}{n-1} \sim \chi^2(r)$$

con

$$\rho_i = \frac{\sum_{t=2}^{i} \epsilon_t \epsilon_{t-1}}{\sum_{t=2}^{i} \epsilon_t^2}.$$

Usualmente el número de rezagos $r \sim n/4$ y se rechaza la nula si $Q > \chi^2_{1-\alpha}(r)$.

Definición 9.3.3. Breusch-Godfrey. A diferencia del contraste anterior (Ljun-Box), el test de Breusch-Godfrey Godfrey (1978) permite contrastar que los errores sigue un comportamiento autorregresivo de orden p o de media móvil de orden q.

- Primero, se estima el modelo[3] $Y_t = \beta_0 + \beta_1 X_t + \beta_2 Z_t + \epsilon_t$.
- Se plantea $\epsilon_t = \alpha_1 X_t + \alpha_2 Z_t + \sum_{i=1}^{p} \rho_i \epsilon_{t-i} + u_t$.
- Al estimar el modelo para Y_t, se obtienen los $\hat{\epsilon}_t$ y se efectúa la regresión

$$\hat{\epsilon}_t = \alpha X_t + \alpha Z_t + \sum_{i=1}^{p} \rho_i \hat{\epsilon}_{t-i} + u_t.$$

- Se obtiene el R^2. Luego, $(n-p)R^2 \sim \chi^2(p)$, y se efectúa el test de hipótesis nula $H_0 : \rho_i = 0$ usando dicho estadístico.

Antes de concluir con los aspectos teóricos y proceder con algunos ejemplos, se presentan una serie de metodologías que permiten corregir los problemas de autocorrelación serial. En concreto:

- Tomar primeras diferencias.
- Usar el método iterativo de Cochrane-Orcutt.
- Aplicar Mínimos Cuadrados Generalizados.

[3] Usamos uno con dos regresores por simplicidad.

9.4. Métodos correctivos

Definición 9.4.1. Primeras diferencias. Se plantea

$$Y_t - Y_{t-1} = \beta_1(X_t - X_{t-1}) + (\epsilon_t - \epsilon_{t-1}),$$

i.e.,

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

donde Δ es el operador de diferencias.

El método de primeras diferencias se aplica cuando la correlación excede 0.8.

Definición 9.4.2. Método iterativo de Cochrane-Orcutt. De manera algorítmica, el método iterativo de Cochrane-Orcutt consiste en lo siguiente.

- Se estima el modelo original vía MCO.
- Se guardan los residuos $\hat{\epsilon}_t$ y se corre la regresión $\hat{\epsilon}_t = \rho\hat{\epsilon}_{t-1} + u_t$. (Esto se puede generalizar al caso $AR(p)$.
- Se obtiene el estimado de ρ.
- Usar el parámetro ρ estimado para transformar las variables y estimar el nuevo modelo por MCO.
- Estas iteraciones se deben repetir hasta un nivel de convergencia considerado de antemano. Inicialmente $\rho = 0$

La transformación de las variables es la siguiente

$$\begin{aligned} y_t^* &= y_t - \rho y_{t-1} \\ x_t^* &= x_t - \rho x_{t-1} \\ y_t^* &= \beta_0(1-\rho) + \beta_1 x_t^* + u_t. \end{aligned}$$

El procedimiento se repite hasta que (comúnmente)

- $|\rho^i - \rho^{i-1}| < 10^{-5}$.
- $|\beta^i - \beta^{i-1}| < 10^{-5}$.
- $\left|(\sum \hat{\epsilon}_t^2)^i - (\sum \hat{\epsilon}_t^2)^{i-1}\right| < 10^{-5}$.

En este contexto, i denota la iteración.

A continuación, veamos cuál es el estimador de MCOG para los casos estudiados en este capítulo.

Definición 9.4.3. MCOG.

$$\hat{\beta}_{MCG} = (X^T \Omega^{-1} X)^{-1} X^T \Omega^{-1} Y$$

donde (para un $AR(1)$)

$$\Omega = \begin{bmatrix} 1 & \rho & \rho^2 & \cdots & \rho^{T-1} \\ \rho & 1 & \cdots & \vdots & \vdots \\ \rho^2 & \vdots & \ddots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \ddots & \rho \\ \rho^{T-1} & \cdots & \cdots & \rho & 1 \end{bmatrix}.$$

Si se conoce el valor de ρ, es posible calcular Ω^{-1}. En dicho caso,

$$\text{Var}(\hat{\beta}_{MCG}) = \sigma^2(X^T\Omega^{-1}X)^{-1}.$$

Si ρ es desconocido, usualmente se usa el estimador

$$r = \frac{\sum_{t=2}^{T}\hat{\epsilon}_t\hat{\epsilon}_{t-1}}{\sum_{t=2}^{T}\hat{\epsilon}_t^2}.$$

A continuación, brindamos un ejemplo que integra los conceptos abordados a lo largo de este capítulo.

Ejemplo 73. Se trata de estimar el efecto traspaso del tipo de cambio al nivel de precios con el siguiente modelo:

$$\ln P_t = \alpha + \beta \ln E_t + \varepsilon_t, \tag{9.3}$$

donde $\varepsilon_t \sim N(0, \sigma^2)$ y $\text{Cov}(\varepsilon_t, \varepsilon_{t-k}) = 0$ para $k \neq 0$. Se decide tomar la cuarta diferencia a esta ecuación para expresar el modelo en diferencias porcentuales anuales. La variable endógena se transforma de la siguiente manera:

$$\Delta \ln P_t = \ln P_t - \ln P_{t-4}$$

con lo que el modelo se simplifica a:

$$\Delta \ln P_t = \Delta \ln E_t + u_t$$
$$u_t = \varepsilon_t - \varepsilon_{t-4}.$$

Se asume una muestra de 11 observaciones trimestrales para las variables. El modelo en diferencias presenta autocorrelación serial, veamos esto a continuación. Primero,

$$\begin{aligned}\text{Cov}(u_t, u_{t-k}) &= \mathbb{E}[(u_t - \mathbb{E}[u_t])(u_{t-k} - \mathbb{E}[u_{t-k}])] \\ &= \mathbb{E}[u_t u_{t-k}].\end{aligned}$$

$$\mathbb{E}[u_t] = \mathbb{E}[\varepsilon_t - \varepsilon_{t-4}] = \mathbb{E}[\varepsilon_t] - \mathbb{E}[\varepsilon_{t-4}] = 0 - 0 = 0.$$

Analicemos caso por caso, para $k = 0, 1, 2, 3, 4...$

- $k = 0$, $t \geq 4$:

$$\begin{aligned}\text{Cov}(u_t, u_{t-k}) &= \text{Cov}(u_t, u_t) \\ &= \mathbb{E}[u_t^2] \\ &= \text{Var}(u_t) - \mathbb{E}[u_t]^2 \\ &= \text{Var}(\varepsilon_t - \varepsilon_{t-4}) - 0^2.\end{aligned}$$

  Ahora, recordemos que $\text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)$, X, Y dos variables aleatorias. Entonces,

  $\text{Var}(\varepsilon_t - \varepsilon_{t-4}) = \text{Var}(\varepsilon_t) + \text{Var}(\varepsilon_{t-4}) - 2\text{Cov}(\varepsilon_t, \varepsilon_{t-4}) = \sigma^2 + \sigma^2 - 0.$

  Así, $\text{Cov}(u_t, u_{t-k}) = 2\sigma^2$.

- $k = 1$, $t \geq 5$:

$$\text{Cov}(u_t, u_{t-1}) = \mathbb{E}[u_t u_{t-1}] = \mathbb{E}[(\varepsilon_t - \varepsilon_{t-4})(\varepsilon_{t-1} - \varepsilon_{t-5})].$$

  Expandiendo el producto y aplicando la linealidad del valor esperado, se obtiene

  $\text{Cov}(u_t, u_{t-1}) = \mathbb{E}[\varepsilon_t \varepsilon_{t-1}] - \mathbb{E}[\varepsilon_{t-4}\varepsilon_{t-1}] - \mathbb{E}[\varepsilon_t \varepsilon_{t-5}] + \mathbb{E}[\varepsilon_{t-4}\varepsilon_{t-5}].$

  Esto es igual a 0 pues $\text{Cov}(\varepsilon_t, \varepsilon_{t-k}) = 0$ para $k \neq 0$.

- $k = 2$, $t \geq 6$, análogamente

$$\begin{aligned}\text{Cov}(u_t, u_{t-2}) &= \mathbb{E}[u_t u_{t-2}] \\ &= \mathbb{E}[(\varepsilon_t - \varepsilon_{t-4})(\varepsilon_{t-2} - \varepsilon_{t-6})] \\ &= \mathbb{E}[\varepsilon_t \varepsilon_{t-2}] - \mathbb{E}[\varepsilon_{t-4}\varepsilon_{t-2}] - \mathbb{E}[\varepsilon_t \varepsilon_{t-6}] + \mathbb{E}[\varepsilon_{t-4}\varepsilon_{t-6}] \\ &= 0.\end{aligned}$$

- $k = 3$, $t \geq 7$:

$$\begin{aligned}\text{Cov}(u_t, u_{t-3}) &= \mathbb{E}[u_t u_{t-3}] \\ &= \mathbb{E}[(\varepsilon_t - \varepsilon_{t-4})(\varepsilon_{t-3} - \varepsilon_{t-7})] \\ &= \mathbb{E}[\varepsilon_t \varepsilon_{t-3}] - \mathbb{E}[\varepsilon_{t-4}\varepsilon_{t-3}] - \mathbb{E}[\varepsilon_t \varepsilon_{t-7}] + \mathbb{E}[\varepsilon_{t-4}\varepsilon_{t-7}] \\ &= 0.\end{aligned}$$

- $k = 4$, $t \geq 8$:

$$\begin{aligned}\text{Cov}u_t, u_{t-4}) &= \mathbb{E}[u_t u_{t-4}] \\ &= \mathbb{E}[(\varepsilon_t - \varepsilon_{t-4})(\varepsilon_{t-4} - \varepsilon_{t-8})] \\ &= \mathbb{E}[\varepsilon_t \varepsilon_{t-4}] - \mathbb{E}[\varepsilon_{t-4}\varepsilon_{t-4}] - \mathbb{E}[\varepsilon_t \varepsilon_{t-8}] + \mathbb{E}[\varepsilon_{t-4}\varepsilon_{t-8}] \\ &= -\mathbb{E}[\varepsilon_{t-4}^2] \\ &= -\sigma^2.\end{aligned}$$

Para un trimestre, con $11 - (r-1)$ observaciones ($r = 4$ en este modelo), tendremos:

$$V = \begin{pmatrix} 2\sigma^2 & 0 & 0 & 0 & -\sigma^2 & 0 & 0 & 0 \\ 0 & 2\sigma^2 & 0 & 0 & 0 & -\sigma^2 & 0 & 0 \\ 0 & 0 & 2\sigma^2 & 0 & 0 & 0 & \sigma^2 & 0 \\ 0 & 0 & 0 & 2\sigma^2 & 0 & 0 & 0 & -\sigma^2 \\ -\sigma^2 & 0 & 0 & 0 & 2\sigma^2 & 0 & 0 & 0 \\ 0 & -\sigma^2 & 0 & 0 & 0 & 2\sigma^2 & 0 & 0 \\ 0 & 0 & -\sigma^2 & 0 & 0 & 0 & 2\sigma^2 & 0 \\ 0 & 0 & 0 & -\sigma^2 & 0 & 0 & 0 & 2\sigma^2 \end{pmatrix}_{8\times 8}.$$

Si deseamos extender el análisis a un año (4 trimestres), la matriz

preserva la misma estructura:

$$V = \begin{pmatrix} 2\sigma^2 & 0 & 0 & 0 & -\sigma^2 & \cdots & 0 & 0 \\ 0 & 2\sigma^2 & 0 & 0 & 0 & \ddots & \vdots & 0 \\ 0 & 0 & \ddots & 0 & 0 & 0 & \ddots & \vdots \\ 0 & 0 & 0 & \ddots & 0 & 0 & 0 & -\sigma^2 \\ -\sigma^2 & 0 & 0 & 0 & 2\sigma^2 & 0 & 0 & 0 \\ \vdots & \ddots & 0 & 0 & 0 & 2\sigma^2 & 0 & 0 \\ 0 & \vdots & -\sigma^2 & 0 & 0 & 0 & 2\sigma^2 & 0 \\ 0 & 0 & \cdots & -\sigma^2 & 0 & 0 & 0 & 2\sigma^2 \end{pmatrix}_{T-3\times T-3}.$$

Ahora, ordenemos en función de su pertinencia los exámenes estadísticos presentados previamente, en función del caso concreto que se está analizando. El test de Durbin-Watson, dada su sencillez, es de gran utilidad a la hora de analizar procesos $AR(1)$. Esto es,

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

siendo ε_t un error idiosincrásico. No obstante, en este tipo de modelos,

$$\begin{aligned} \text{Cov}(u_t u_{t-1}) &= \mathbb{E}[u_t u_{t-1}] \\ &= \mathbb{E}[(\rho u_{t-1} + \varepsilon_t) u_{t-1}] \\ &= \mathbb{E}[\rho u_{t-1}^2 + \varepsilon_t u_{t-1}] \\ &= \rho \mathbb{E}[u_{t-1}^2] + \mathbb{E}[\varepsilon_t u_{t-1}] \\ &= \rho \mathbb{E}[u_{t-1}^2] \\ &= \rho \text{Var}(u_{t-1}^2) \neq 0. \end{aligned}$$

En este modelo $\text{Cov}(u_t u_{t-1}) = 0$. Por ende, no es oportuno aplicar

el test de Durbin-Watson. Queda por analizar el test Llung-Box y el test Breusch-Godfrey. Recordemos las siguientes características:

- Llung-Box: el número de rezagos a testear debe ser aproximadamente $N/4$ (no más). Proporciona la existencia de un $AR(p)$ o $MA(q)$, pero, no da necesariamente el orden. Sin embargo, nos permite ver la autocorrelación (normal y parcial), analizando entonces el grado de correlación.

- Breusch-Godfrey: analizar el estadístico $(N-p)R^2$, el R^2 proviniendo de la regresión donde $\hat{u}_t$ es la variable dependiente. Este examen estadístico sí nos permite encontrar el orden de los rezagos.

Volviendo a nuestro modelo, para un trimestre, si se cuenta únicamente con 8 observaciones, el test de Llung-Box no sería relevante pues, solo podríamos testear 2 rezagos y ciertamente, la covarianza de los errores no es igual a cero en errores con cuatro unidades de diferencia temporal. Es por ello que el test de Breusch-Godfrey sería el más apropiado. Le sigue sin embargo el Llung-Box, puesto que el Durbin-Watson, no solo requiere más de 15 observaciones, pero además, restringe el estudio al modelo $AR(1)$.

Ejemplo 74. A partir de la especificación Cobb-Douglas $Y = F(K, L) = K^\alpha L^\beta$, se plantea el modelo

$$Y_t = AK_t^\alpha L_t^\beta e^{a_t},$$

donde a_t es un error aleatorio. Luego, sacando logaritmos y usando al notación $x_t = \ln X_t$:

$$y_t = c + \beta_1 \ell_t + \beta_2 k_t + a_t.$$

Primero, estimamos el modelo vía MCO. Se resuelve

$$\begin{pmatrix} \hat{c} \\ \hat{\beta}_1 \\ \hat{\beta}_2 \end{pmatrix} = \left(\begin{pmatrix} 1 & \cdots & 1 \\ \ell_1 & \cdots & \ell_n \\ k_1 & \cdots & k_n \end{pmatrix} \begin{pmatrix} 1 & \ell_1 & k_1 \\ \vdots & \vdots & \vdots \\ 1 & \ell_n & k_n \end{pmatrix} \right)^{-1} \begin{pmatrix} 1 & \cdots & 1 \\ \ell_1 & \cdots & \ell_n \\ k_1 & \cdots & k_n \end{pmatrix} \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}.$$

Luego, a partir de esta estimación, se computan

$$\hat{a}_t = y_t - \hat{y}_t = y_t - (\hat{c} + \hat{\beta}_1 \ell_t + \hat{\beta}_2 k_t).$$

Enseguida, se estima ρ en la regresión

$$\hat{a}_t = \rho \hat{a}_{t-1} + \varepsilon_t.$$

A partir de $\hat{\rho}$, se efectúan los cambios de variable

$$\begin{aligned} y_t^* &= y_t - \hat{\rho} y_{t-1} \\ \ell_t^* &= \ell_t - \hat{\rho} \ell_{t-1} \\ k_t^* &= k_t - \hat{\rho} k_{t-1}. \end{aligned}$$

Se repiten los pasos (j veces) hasta que, eventualmente,[4]

$$|\hat{\rho}_j - \hat{\rho}_{j-1}| < 10^{-5}$$
$$||\hat{\beta}_j - \hat{\beta}_{j-1}|| < 10^{-5}$$
$$\left| \sum_{t=1}^{n} (\hat{u}_t^2)^j - \sum_{t=1}^{n} (\hat{u}_t^2)^{j-1} \right| < 10^{-5}.$$

9.5. Mínimos Cuadrados No Lineales

Antes de terminar con el capítulo, vamos a comentar brevemente el modelo de mínimos cuadrados no lineales. En este modelo, relajamos el supuesto de que $\mathbb{E}(y|x,\beta)$ es lineal en parámetros y asumimos que la forma funcional es conocida, $\mathbb{E}(y|x,\theta) = g(x,\theta)$, donde $g(x,\theta)$ es conocida y diferenciable[5]. La estimación se realiza mediante la resolución del siguiente problema de optimización:

$$\hat{\theta}_{\text{MCNL}} = \operatorname{argmin} S_N(\theta)$$

donde

$$S_N(\theta) = \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i,\theta))^2.$$

[4]Usualmente, se utiliza una de las siguientes normas $||\cdot||$: dado $x = (x_1, ..., x_d) \in \mathbb{R}^d$

$$||x||_2 = \sqrt{\sum_{i=1}^{d} x_i^2}$$
$$||x||_1 = \sum_{i=1}^{d} |x_i|$$
$$||x||_{\text{máx}} = \max_{1 \le i \le d} \{|x_i|\}.$$

[5]Por ejemplo, $g(x,\theta) = \theta_1 + \theta_2 e^{\theta_3 x}$.

Dado que no siempre existe una solución analítica, se usan métodos numéricos como el algoritmo iterativo de Gauss-Newton. Bajo la continuidad y diferenciabilidad de $g(x,\theta)$, se aplica el teorema de Taylor de primer orden en torno de θ_0, obteniendo:

$$\hat{\theta} = \theta_0 + \left[\sum_{i=1}^{n} \frac{\partial g(x_i,\theta)}{\partial \theta}\bigg|_{\theta_0} \frac{\partial g(x_i,\theta)}{\partial \theta^T}\bigg|_{\theta_0}\right]^{-1} \sum_{i=1}^{n} \frac{\partial g(x_i,\theta)}{\partial \theta}\bigg|_{\theta_0} (y_i - g(x_i,\theta_0))$$

Así vemos que Gauss-Newton es un algoritmo iterativo, donde la fórmula iterativa es

$$\hat{\theta}_j = \hat{\theta}_{j-1} + \left[\sum_{i=1}^{n} \frac{\partial g(x_i,\theta)}{\partial \theta}\bigg|_{\hat{\theta}_{j-1}} \frac{\partial g(x_i,\theta)}{\partial \theta^T}\bigg|_{\hat{\theta}_{j-1}}\right]^{-1} \sum_{i=1}^{n} \frac{\partial g(x_i,\theta)}{\partial \theta}\bigg|_{\theta_{j-1}} (y_i - g(x_i,\theta_{j-1}))$$

Los pasos del algoritmo incluyen la selección de valores iniciales, que pueden basarse en teoría, alteraciones de la función para obtener una solución analítica, valores obtenidos por mínimos cuadrados ordinarios, o gráficos de la función. Luego, se procede con la iteración hasta que se cumpla una regla de parada, que puede ser absoluta ($||\theta_j - \theta_{j-1}|| <$ tolerancia, generalmente 10^{-6}) o relativa ($\frac{||\theta_j - \theta_{j-1}||}{||\theta_{j-1}||} <$ tolerancia).

Respecto a la distribución asintótica, tenemos que

$$\sqrt{n}(\hat{\theta} - \theta) \approx \left[\frac{1}{n}\sum_{i=1}^{n} \frac{\partial g(x_i,\theta)}{\partial \theta}\bigg|_{\theta} \frac{\partial g(x_i,\theta)}{\partial \theta^T}\bigg|_{\theta}\right]^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{\partial g(x_i,\theta)}{\partial \theta}\bigg|_{\theta} (y_i - g(x_i,\theta)).$$

Asumiendo simplicidad en $g_{\theta i} = \frac{\partial g(x_i,\theta)}{\partial \theta^T}\Big|_{\hat{\theta}}$, tenemos:

$$\mathbb{E}(g_{\theta i} g_{\theta i}^T) = \mathbb{E}\left(\frac{\partial g(x_i,\theta)}{\partial \theta} \frac{\partial g(x_i,\theta)}{\partial \theta^T}\right)$$

$$\mathbb{E}(g_{\theta i} g_{\theta i}^T \epsilon_i^2) = \mathbb{E}\left(\frac{\partial g(x_i,\theta)}{\partial \theta} \frac{\partial g(x_i,\theta)}{\partial \theta^T} \epsilon_i^2\right)$$

La varianza asintótica es:

$$V_\theta = \left(\mathbb{E}(g_{\theta i} g_{\theta i}^T)\right)^{-1} \left(\mathbb{E}(g_{\theta i} g_{\theta i}^T \epsilon_i^2)\right) \left(\mathbb{E}(g_{\theta i} g_{\theta i}^T)\right)^{-1}$$

Definiendo,

$$\frac{1}{n}\sum_{i=1}^{n} \hat{g}_{\theta i}\hat{g}_{\theta i}^T = \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial g(x_i,\theta)}{\partial \theta}\right|_{\hat{\theta}} \left.\frac{\partial g(x_i,\theta)}{\partial \theta^T}\right|_{\hat{\theta}}$$

$$\frac{1}{n}\sum_{i=1}^{n} \hat{g}_{\theta i}\hat{g}_{\theta i}^T \epsilon_i^2 = \frac{1}{n}\sum_{i=1}^{n} \left.\frac{\partial g(x_i,\theta)}{\partial \theta}\right|_{\hat{\theta}} \left.\frac{\partial g(x_i,\theta)}{\partial \theta^T}\right|_{\hat{\theta}} \epsilon_i^2,$$

la varianza asintótica estimada es:

$$\hat{V}_\theta = \left[\frac{1}{n}\sum_{i=1}^{n} \hat{g}_{\theta i}\hat{g}_{\theta i}^T\right]^{-1} \left[\frac{1}{n}\sum_{i=1}^{n} \hat{g}_{\theta i}\hat{g}_{\theta i}^T \epsilon_i^2\right] \left[\frac{1}{n}\sum_{i=1}^{n} \hat{g}_{\theta i}\hat{g}_{\theta i}^T\right]^{-1}.$$

Capítulo 10

Endogeneidad

El análisis de datos con regresores endógenos (variables explicativas observables correlacionadas con términos de error no observables) es, probablemente un de las contribuciones fundamentales de la econometría a la estadística. Si bien la endogeneidad puede surgir de distintas fuentes como regresores con error de medida, selección muestral, efecto tratamiento heterogéneo, etc. el término apareció inicialmente en el contexto de ecuaciones simultáneas, por ejemplo, ecuaciones de oferta y demanda. En este capítulo nos concentraremos en el caso en que existe una ecuación lineal de interés, llamada la ecuación estructural, y alguno de los regresores está correlacionado con el término de error. Una referencia clásica y completa para este tema es Angrist and Pischke (2009).

10.1. Variables Instrumentales

Considere el siguiente modelo lineal: $Y = X\beta + \epsilon$ donde (X, Y) representa una observación de dimensión $(1 \times (k+1))$, β es un vector de parámetros y ϵ es un término de error no observable. El supuesto de identificación fundamental de Mínimos Cuadrados Ordinarios es que las variables explicativas no estén correlacionadas con el término de error, esto es: $\mathbb{E}(X^T\epsilon) = 0$. Note que el parámetro poblacional β puede ser expresado en momentos de las variables observables explotando el supuesto recién presentado: $X^TT = X^TX\beta + X^T\epsilon$. Tomando valor esperado tenemos que: $\beta = \mathbb{E}[X^TX]^{-1}\mathbb{E}[X^TY]$. Dado que (X, Y) es observable, β es identificado. Recordemos que el principio de la analogía para escoger un estimador dice que transformemos los momentos poblacionales en momentos muestrales. Haciendo eso, obtenemos el estimador MCO:

$$\hat{\beta}_{\text{MCO}} = \left(\frac{1}{n}\sum_{i=1}^{n} X_i^TX_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} X_i^TY_i\right) \equiv (X^TX)^{-1}X^TY.$$

Un ejemplo clásico de endogeneidad es el caso de la relación entre educación y salario, donde la habilidad no observada está correlacionada con la educación, sesgando así los estimadores de MCO. Veamos esto a detalle en el siguiente ejemplo.

Ejemplo 75. En la literatura económica, es común analizar la relación entre educación y salario. Sin embargo, uno de los desafíos principales en esta estimación es la endogeneidad. La

habilidad innata de los individuos, que no es observable, puede estar correlacionada tanto con la educación como con el salario, generando un sesgo en las estimaciones. Consideremos el siguiente modelo de salario:

$$\ln(w_i) = \beta_0 + \beta_1 x_i + \epsilon_i \tag{10.1}$$

donde $\ln(w_i)$ es el logaritmo del salario del individuo i, x_i es el nivel de educación del individuo i y ϵ_i es el término de error. Si la habilidad (A_i) está correlacionada con la educación y afecta directamente el salario, podemos expresar el salario como:

$$\ln(w_i) = \beta_0 + \beta_1 x_i + \gamma A_i + u_i \tag{10.2}$$

donde u_i es el nuevo término de error. Si A_i no se incluye en el modelo y está correlacionada con x_i, entonces x_i está endógena y la estimación de β_1 estará sesgada. Para demostrar este sesgo, partimos del modelo estimado:

$$\ln(w_i) = \beta_0 + \beta_1 x_i + \epsilon_i \tag{10.3}$$

donde $\epsilon_i = \gamma A_i + u_i$. El estimador de mínimos cuadrados ordinarios (MCO) de β_1 es:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(\ln(w_i) - \overline{\ln(w)})}{\sum_i (x_i - \bar{x})^2} \tag{10.4}$$

Sustituyendo $\ln(w_i)$ en el numerador:

$$\hat{\beta}_1 = \frac{\sum_i (x_i - \bar{x})(\beta_0 + \beta_1 x_i + \gamma A_i + u_i - \overline{\ln(w)})}{\sum_i (x_i - \overline{x})^2} \tag{10.5}$$

Separando los términos,

$$\hat{\beta}_1 = \beta_1 + \frac{\sum_i (x_i - \bar{x})(\gamma A_i + u_i)}{\sum_i (x_i - \bar{x})^2} \tag{10.6}$$

El término $\frac{\sum_i (x_i - \overline{x}) u_i}{\sum_i (x_i - \overline{x})^2}$ es cero en promedio, ya que u_i es ruido blanco no correlacionado con x_i. Sin embargo, el término $\frac{\sum_i (x_i - \overline{x}) \gamma A_i}{\sum_i (x_i - \overline{x})^2}$ no es cero si x_i y A_i están correlacionados. Por lo tanto, podemos escribir

$$\hat{\beta}_1 \approx \beta_1 + \gamma \cdot \frac{\sum_i (x_i - \overline{x}) A_i}{\sum_i (x_i - \overline{x})^2} \tag{10.7}$$

El término $\frac{\sum_i (x_i - \overline{x}) A_i}{\sum_i (x_i - \overline{x})^2}$ representa la correlación entre la educación y la habilidad. Si esta correlación es positiva (lo cual es común, ya que individuos con mayor habilidad innata tienden a obtener más educación), entonces el estimador $\hat{\beta}_1$ estará sesgado hacia arriba.

La estimación de β_1 en el modelo (10.3) estará sesgada si la habilidad innata A_i está correlacionada con la educación x_i. Este sesgo surge porque la habilidad no observada, que afecta tanto a la educación como al salario, no se incluye en el modelo, lo que lleva a una correlación entre el término de error ϵ_i y la variable explicativa x_i.

Para corregir este sesgo, es necesario utilizar métodos econométricos como las variables instrumentales, que permiten aislar la variación exógena en la educación que no está correlacionada con la habilidad innata. Veremos esto a continuación.

Cuando se viola el supuesto de exogeneidad de las variables explicativas con respecto al término de error, las variables X incluyen un subconjunto de variables que son endógenas, lo que significa que: $\mathbb{E}(X^T \epsilon) \neq 0$. Esto genera un problema de identificación. No es posible encontrar una expresión del parámetro poblacional β en función de momentos poblacionales de variables observables a no ser que contemos con otro set de variables Z que

cumpla las siguientes condiciones: $\mathbb{E}[Z^T\epsilon] = 0$ y $\mathbb{E}[Z^TX] \neq 0$. Note que implícitamente estamos asumiendo que el producto Z^TX es realizable, esto implica que el orden de las matrices es el mismo. Note que con esta nueva variable podemos proceder de la misma manera que lo hicimos para MCO. Podemos pre-multiplicar la ecuación estructural por Z^T y obtener un sistema de ecuaciones: $Z^TY = Z^TX\beta + Z^T\epsilon$ y por lo tanto obtenemos el siguiente sistema de ecuaciones: $\mathbb{E}[Z^TX]\beta = \mathbb{E}[Z^TY]$ donde $\mathbb{E}[Z^TX]$ es de orden $K \times K$ y $\mathbb{E}[Z^TY]$ es de orden $K \times 1$. Por lo tanto, la ecuación representa un sistema de K ecuaciones con K incógnitas dadas por $\beta_1, \beta_2, ..., \beta_K$. Este sistema tiene solución única si la matriz $\mathbb{E}[Z^TX]$ es invertible, lo cual sucede si el rango de esta es completo e igual a K. Luego, si tenemos una muestra aleatoria (Y_i, X_i, Z_i) y siguiendo el principio de la analogía tenemos que el estimador de variables instrumentales está dado por:

$$\hat{\beta}_{\text{IV}} = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i^T X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i^T Y_i\right) \equiv (Z^TX)^{-1}Z^TY.$$

Cuando buscamos instrumentos para una variable endógena, los supuestos $\mathbb{E}[Z^T\epsilon] = 0$ (exogeneidad[1]) y $\mathbb{E}[Z^TX] \neq 0$ (relevancia[2].) son igualmente importantes para identificar β. Sin embargo, hay una diferencia, el supuesto $\mathbb{E}[Z^T\epsilon] = 0$ no puede ser testeado. La razón de esto es simple: no observamos ϵ como para realizar un test. Por otra parte, el supuesto $\mathbb{E}[Z^TX] \neq 0$ puede y debe ser testeado. Más adelante veremos que es relativamente sencillo hacerlo y no

[1] $\text{Cov}(Z, \epsilon) = 0$.

[2] $\text{Cov}(X, Z) \neq 0$.

requiere más instrumental que un test-t o F. Cuando la correlación de las variables instrumentales con las endógenas es pequeña se dice que estamos en presencia de instrumentos débiles.

El método de variables instrumentales se implementa en dos etapas. Primero, se realiza una regresión de X_i sobre Z_i y se guardan los valores predichos $\hat{X}_i$. Luego, se usa $\hat{X}_i$ en vez de X_i en $Y_i = \beta_0 + \beta_1 \hat{X}_i + \epsilon_i$.

A continuación un ejemplo de variable instrumental.

Ejemplo 76. Se cuenta con una muestra de niños y niñas entre dos y cinco años a nivel nacional y se quiere predecir en qué medida el estado nutricional de ellos y ellas influye en sus habilidades cognitivas. Así, se plantea la siguiente ecuación:

$$\begin{aligned} \text{vocabulario}_i = \delta_0 + \delta_1 \text{nutrición}_i + \delta_2 \text{edad}_i + \delta_3 \text{mujer}_i \\ + \delta_4 \text{nse}_i + \delta_5 \text{urbano}_i + \epsilon_i \end{aligned} \tag{10.8}$$

1. vocabulario_i representa las habilidades cognitivas medidas a través del vocabulario del niño i.

2. nutrición_i se refiere al estado nutricional del niño i, medido a través de la talla para la edad.

3. edad_i es la edad del niño i.

4. mujer_i es una variable indicadora que toma el valor de 1 si el niño i es una niña, y 0 si es un niño.

5. nse_i representa el nivel socioeconómico del hogar del niño i.

6. urbano$_i$ es una variable indicadora que toma el valor de 1 si el niño i vive en una zona urbana, y 0 si vive en una zona rural.

7. ϵ_i es el término de error.

En este caso, tenemos que la variable *nutrición*, que hace referencia a la talla para la edad, sería una variable endógena que estaría correlacionada con el término de error. Una variable que podríamos usar como instrumento para este modelo es el estado nutricional de la madre, medido a partir de la talla de la misma. Sería una variable que está fuertemente correlacionada con el estado nutricional de los niños y niñas pero no con el término de error asociado a las habilidades cognitivas de los niños y niñas.

Las variables instrumentales han surgido como una técnica importante en trabajos de investigación para corregir problemas de endogeneidad. Por ejemplo, Angrist and Krueger (1991b) utilizan el «quarter of birth» como instrumento de la educación para estudiar el retorno de la educación. Angrist (1990b) emplea el «draft number» de la lotería para servir en la guerra de Vietnam como instrumento para la participación en la guerra, en el estudio del impacto de servir en la guerra sobre el ingreso. Card (1995) usan la proximidad a una universidad (college) como instrumento de la educación de la persona para estimar el retorno a la educación. Frankel and Romer (1999) utilizan la proximidad a otros países y el tamaño como instrumentos del comercio internacional para estudiar el impacto del comercio sobre el PBI.

Antes de continuar con la siguiente sección, discutamos acerca de la consistencia del estimado IV. La consistencia de este estimador sigue principalmente de la ley de los grandes números. Note que podemos escribir el estimador de variables instrumentales como sigue,

$$\hat{\beta}_{\text{IV}} = \beta + \left(\frac{1}{n}\sum_{i=1}^{n} Z_i^T X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n} Z_i^T \epsilon_i\right)$$

Luego, se deduce claramente que $\mathbb{P}\lim \hat{\beta}_{\text{IV}} = \beta$. Ahora bien, podemos generar la expresión clásica ajustada por $\sqrt{n}$:

$$\sqrt{n}(\hat{\beta}_{\text{IV}} - \beta) = \left(\frac{1}{n}\sum_{i=1}^{n} Z_i^T X_i\right)^{-1}\left(\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i^T \epsilon_i\right)$$

donde el primer término del lado derecho de la ecuación convergerá a $\mathbb{E}[Z^T X] = M_{ZX}$ por la ley débil de los grandes números y el segundo término converge en distribución a una normal por el Teorema Central del Límite

$$\frac{1}{\sqrt{n}}\sum_{i=1}^{n} Z_i^T \epsilon_i \xrightarrow{d} N(0, V_0)$$

donde $V_0 = \mathbb{E}(\epsilon^2 Z^T Z)$. Por lo tanto,

$$\sqrt{n}(\hat{\beta}_{\text{IV}} - \beta) \xrightarrow{d} N(0, M_{ZX}^{-1} V_0 (M_{ZX}^{-1})^T).$$

10.2. Múltiples instrumentos 2SLS

Sea Z la matriz de instrumentos de orden $n \times L$ y X la matriz de variables independientes de orden $n \times K$. Cuando hay más instrumentos que variables endógenas (o más de un instrumento

para una variable endógena), tenemos el caso sobre identificado. Esto significa que tenemos más ecuaciones que incógnitas en nuestro sistema. Es prácticamente imposible encontrar una solución que satisfaga todas las ecuaciones, excepto en casos muy particulares.

Una manera ineficiente de resolver este problema es eliminar instrumentos, igualando así el orden de las matrices Z y X. Sin embargo, esto resulta en la pérdida de información valiosa. Otra manera de resolver este problema es post-multiplicando la matriz de instrumentos Z por otra matriz Λ de orden $L \times K$. Luego, la matriz $Z\Lambda$ es de dimensión $n \times K$.

Explotando la condición de identificación $\mathbb{E}[Z^T\epsilon] = 0$, tenemos:

$$\Lambda^T Z^T Y = \Lambda^T Z^T X\beta + \Lambda^T Z^T \epsilon.$$

Podemos identificar β tomando el valor esperado:

$$\beta = [\mathbb{E}(\Lambda^T Z^T X)]^{-1}\mathbb{E}[\Lambda^T Z^T Y]$$

Siguiendo el principio de analogía, el estimador está dado por:

$$\hat{\beta} = \left(\frac{1}{n}\sum_{i=1}^{n}\Lambda^T Z_i^T X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\Lambda^T Z_i^T Y_i\right) \equiv (\Lambda^T Z^T X)^{-1}\Lambda^T Z^T Y.$$

El asunto por discutir es la matriz Λ. Esta puede ser desconocida, por lo que necesitaremos un estimador de Λ. Suponiendo que tenemos un estimador de Λ dado por $\hat{\Lambda}$, el estimador generalizado de variables instrumentales (GIV) o de método de momentos generalizado (GMM) está dado por:

$$\hat{\beta}_{\text{GIV}} = \left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}^T Z_i^T X_i\right)^{-1}\left(\frac{1}{n}\sum_{i=1}^{n}\hat{\Lambda}^T Z_i^T y_i\right) \equiv (\hat{\Lambda}^T Z^T X)^{-1}\hat{\Lambda}^T Z^T y$$

Usualmente, $\hat{\Lambda} = (Z^T Z)^{-1} Z^T X$, que corresponde al estimador MCO de la regresión de X sobre Z. Así, el estimador de mínimos cuadrados en dos etapas (2SLS) es:

$$\begin{aligned}\hat{\beta}_{2SLS} &= (X^T Z (Z^T Z)^{-1} Z^T X)^{-1} X^T Z (Z^T Z)^{-1} Z^T Y \\ &= (\hat{X}^T \hat{X})^{-1} \hat{X}^T Y.\end{aligned}$$

donde $\hat{X} = Z(Z^T Z)^{-1} Z^T X$. El nombre de mínimos cuadrados en dos etapas (2SLS) proviene de la interpretación de Theil que muestra que el estimador se puede obtener de la siguiente manera:

$$X = Z\Lambda + u$$

$$Y = X\beta + \epsilon$$

así, se estima la primera etapa y se utilizan los valores estimados $\hat{X} = Z\hat{\Lambda} = Z(Z^T Z)^{-1} Z^T X$ en la segunda etapa.

Luego, si

$$\sqrt{\frac{1}{N}} Z^T \epsilon \xrightarrow{d} N(0, V_0)$$

donde, nuevamente, $V_0 = \mathbb{E}(\epsilon^2 Z^T Z)$ y $\hat{\Lambda} \xrightarrow{\mathbb{P}} \Lambda$, se puede demostrar que:

$$\sqrt{n}(\hat{\beta}_{\text{GIV}} - \beta) \xrightarrow{d} N(0, [\Lambda^T M_{ZX}]^{-1} \Lambda^T V_0 \Lambda [\Lambda^T M_{ZX}^{-1}]^T)$$

Esta expresión depende de Λ, V_0 y M_{ZX}. Aunque no se conozca Λ, podemos estimarla como la proyección ortogonal de X sobre Z y, en consecuencia, obtener la distribución asintótica del estimador de 2SLS. Por la Ley de los Grandes Números, $\hat{\Lambda} \xrightarrow{\mathbb{P}} \Lambda \equiv [\mathbb{E}(Z^T Z)]^{-1} \mathbb{E}(Z^T X) = M_{ZZ}^{-1} M_{ZX}$.

Reemplazando $\hat{\Lambda}$ en la expresión anterior, obtenemos la varianza asintótica del estimador de 2SLS:

$$\sqrt{n}(\hat{\beta}_{2SLS} - \beta) \xrightarrow{d} N(0, \sigma^2[M_{XZ}^T M_{ZZ}^{-1} M_{ZX}]^{-1}).$$

Para estimar σ^2, definamos el residuo de 2SLS como $\hat{\epsilon} = Y - X\hat{\beta}_{2SLS}$. El estimador de σ^2 se define de la forma tradicional:

$$\hat{\sigma}^2 = \frac{1}{n-k} \sum_{i=1}^{n} \hat{\epsilon}_i^2$$

Si ϵ y Z no son independientes, pero $\mathbb{E}[Z^T \epsilon] = 0$, se puede estimar consistentemente la matriz de varianzas y covarianzas usando Eicker-Huber-White o Newey-West dependiendo si los errores son autocorrelacionados, véase Wooldridge (2001).

El estimador de Eicker-Huber-White de la varianza asintótica de $\hat{\beta}_{2SLS}$ está dado por

$$\text{Avar}(\hat{\beta}_{2SLS}) = (\hat{X}^T \hat{X})^{-1} \left(\sum_{i=1}^{n} \hat{\epsilon}_i^2 Z_i^T Z_i \right) (\hat{X}^T \hat{X})^{-1}$$

Desde el enfoque numérico, en Stata, usando el comando `ivregress 2sls` con la opción `vce(robust)` nos proporcionará los errores estándar de la matriz descrita anteriormente. También se puede usar el comando `ivreg2` con la opción `robust`, véase el Manual de Stata.

10.3. Método Generalizado de Momentos

Un estimador alternativo a 2SLS en presencia de endogeneidad y variables instrumentales es el estimador de GMM. Definamos las

condiciones de momento como

$$m(Z, X, \beta) = Z^T(Y - X\beta)$$

donde Z es una realización del vector de instrumentos de $L \times 1$ y X es una realización del vector de variables endógenas de $K \times 1$. Suponiendo que $L > K$, tenemos un sistema sobre identificado; véase Rau (2016). Dado el supuesto de identificación $\mathbb{E}[Z\epsilon] = 0$, tenemos que:

$$\mathbb{E}[m(Z, X, \beta)] = 0.$$

Luego, el valor esperado de cada condición de momento es cero. Además, recordemos que cada condición de momento poblacional tiene su contraparte muestral dada por

$$m(Z, X, \beta) = \frac{1}{n}\sum_{i=1}^{n} Z^T(Y - X\beta) = \frac{1}{n}Z^T\epsilon.$$

El problema que tenemos es el usual: «la probabilidad de encontrar una solución que satisfaga las L (siendo que tenemos K incógnitas) es casi 0» a menos que reduzcamos el orden multiplicando por alguna matriz como se hizo en 2SLS. En el caso que $L = K$ se tiene un sistema exactamente identificado y la solución está dada por $m(Z, X, \beta) = 0$, con lo cual se obtiene la misma solución que $\hat{\beta}_{IV}$.

Para el caso sobre-identificado, el estimador de GMM es aquel que minimiza la siguiente forma cuadrática:

$$\min_{\beta} m(Z, X, \beta)^T W^{-1} m(Z, X, \beta)$$

donde W^{-1} es una matriz de $L \times L$ con lo cual el sistema es de $K \times K$.

Se define el estimador de GMM eficiente (EGMM) como aquel que utiliza como matriz de ponderación:

$$W = \text{Var}(m(Z, X, \beta)) = \mathbb{E}[\epsilon^2 Z^T Z] = V_0.$$

Por lo tanto, en el caso que $W = V_0$, tenemos que $\hat{\beta}_{\text{GMM}}$ minimiza la siguiente expresión:

$$\min_{\beta} \frac{1}{n}(Y - X\beta)^T Z V_0^{-1} Z^T (Y - X\beta)$$

luego,

$$\hat{\beta}_{\text{EGMM}} = [X^T Z V_0^{-1} Z^T X]^{-1} X^T Z V_0^{-1} Z^T Y.$$

Lo único que falta es un estimador consistente de V_0. Bajo el supuesto de heterocedasticidad, podemos usar el estimador de Eicker-White (véase Wooldridge (2001)), con lo cual la varianza asintótica estaría dada por

$$\text{Avar}(\hat{\beta}_{\text{EGMM}}) = (M_{ZX}^T V_0^{-1} M_{ZX})^{-1}$$

Luego, se puede implementar el estimador EGMM en tres etapas:

1. Se estima el modelo por 2SLS y obtenga los residuos de la manera antes descrita $\hat{\epsilon} = Y - X\hat{\beta}_{2SLS}$.
2. Construya la matriz $\hat{V}_0 = \frac{1}{n}\sum \hat{\epsilon}_i^2 Z_i^T Z_i$.
3. Estime mediante EGMM usando $\hat{V}_0$ como matriz de ponderación.

En Stata, esto se puede implementar con el comando `ivreg2` con la opción `gmm`.

En el caso general, cuando los errores son heterocedásticos y/o autocorrelacionados y $V_0 \neq \sigma^2 M_{ZZ}$, el estimador 2SLS (caso particular del GIV) no tendrá la menor varianza asintótica. Para obtener un estimador eficiente necesitamos escoger una matriz Λ que minimice la varianza asintótica.

Luego queremos minimizar con respecto a Λ la siguiente expresión:

$$\text{Avar}(\hat{\beta}_{\text{GIV}}) = [\Lambda^T M_{ZX}]^{-1} \Lambda^T V_0 \Lambda ([\Lambda M_{ZX}]^{-1})^T$$

Se puede demostrar que

$$\Lambda^* = V_0^{-1} M_{ZX} = \arg \min_{\Lambda} \text{Avar}(\hat{\beta}_{\text{GIV}}(\Lambda)).$$

Pero en la práctica no podemos disponer de Λ^*, incluso si suponemos V_0 conocida. Necesitamos un estimador consistente de M_{ZX}. La Ley Débil de los Grandes Números nos garantiza que si $\{X_i, Z_i\}$ son i.i.d con primer y segundo momento acotados,

$$\hat{\Lambda}^* = V_0^{-1} \frac{1}{n} \sum Z_i^T X_i \xrightarrow{p} V_0^{-1} M_{ZX} = \Lambda^*$$

Por lo tanto, el estimador generalizado de variables instrumentales eficiente corresponde al estimador eficiente de método de momentos y es igual a:

$$\hat{\beta}_{\text{EGIV}} = \hat{\beta}_{\text{EGMM}} = [X^T Z V_0^{-1} Z^T X]^{-1} X^T Z V_0^{-1} Z^T Y$$

con distribución asintótica:

$$\sqrt{n}(\hat{\beta}_{\text{EGMM}} - \beta) \xrightarrow{d} N(0, (M_{ZX}^T V_0^{-1} M_{ZX})^{-1})$$

Siempre que se estima usando variables instrumentales en el caso de sobre-identificación, es posible testear si los instrumentos no están correlacionados con el término de error. En el caso de GMM, esto se hace testeando que las condiciones de momento muestrales, en conjunto, no sean diferentes de cero (en términos estadísticos). Esto solo se puede realizar cuando se tienen más instrumentos excluidos que variables endógenas, es decir, cuando $L > K$. Este test de alguna manera testea la especificación del modelo y las condiciones de ortogonalidad. Si rechazamos la hipótesis nula (que el modelo está correctamente especificado y que las condiciones de momento son válidas) uno debe preocuparse porque los instrumentos no son limpios (están correlacionados con el término de error); o el modelo está incorrectamente especificado en el sentido de las restricciones de exclusión (qué instrumentos se excluyen de la ecuación estructural).

En el contexto de GMM, este test se realiza usando el estadístico J de Hansen (1982). Este estadístico no es más que la función objetivo evaluada en $\hat{\beta}_{\text{EGMM}}$ y se distribuye como χ^2_{L-K}, así:

$$J(\hat{\beta}_{\text{EGMM}}) = nm(Z, X, \hat{\beta}_{\text{EGMM}})^T \hat{V}_0^{-1} m(Z, X, \hat{\beta}_{\text{EGMM}}) \xrightarrow{d} \chi^2_{L-K}.$$

Claramente, un valor grande de $J(\hat{\beta}_{\text{EGMM}})$, es decir, mayor al valor crítico obtenido de la tabla, nos da indicios de que el modelo está mal especificado o que los instrumentos no son limpios.

En Stata, el comando `ivreg2` con la opción `robust` estima por EGMM y además entrega el estadístico J. En el caso de 2SLS, existe el test de Sargan (1958) para restricciones de sobre-identificación.

Es muy simple:

$$\text{Sargan} = \frac{\hat{\epsilon}^T \text{Proy}_Z \hat{\epsilon}}{\hat{\epsilon}^T \hat{\epsilon}/n} \xrightarrow{d} \chi^2_{L-K}.$$

Una manera sencilla de obtener el estadístico de Sargan es correr el modelo por 2SLS, obtener $\hat{\epsilon}$, correr la regresión auxiliar de $\hat{\epsilon}$ sobre todas las variables exógenas (X y Z) y obtener el R^2[3].

En Stata, después del comando `ivregress 2sls` ejecute `estat overid` para que el programa nos entregue el estadístico de Sargan. Para versiones anteriores, la secuencia es primero `ivreg2` y luego `overid`.

10.4. Instrumentos débiles

Cuando los instrumentos están débilmente correlacionados[4] con las variables endógenas, el uso de variables instrumentales (en muestra finita) puede ser perjudicial John Bound (1995). En presencia de instrumentos débiles, el estimador por variables instrumentales puede estar sesgado en la misma dirección que el estimador por MCO y puede no ser consistente John Chao (2005).

[3] De hecho, es posible demostrar que Sargan $= n \times R^2$.

[4] Las complicaciones en general son por la no linealidad en la primera etapa (si la relación entre los instrumentos y la variable endógena es curiosamente no lineal, los instrumentos pueden ser débiles), outliers: dada la falta de resistencia estadística de OLS, la presencia de outliers puede llevar a que los instrumentos sean débiles y/o correlación con la variable endógena en una subpoblación: si los instrumentos están correlacionados con la(s) variable(s) endógena(s) solo en una subpoblación y esta correlación se diluye en el total, los instrumentos pueden ser débiles.

Además, los tests tienen una medida incorrecta y los intervalos de confianza presentan problemas.

La medida de fortaleza de los instrumentos está dada por el parámetro de concentración

$$\mu^2 = \frac{\Lambda^T Z^T Z \Lambda}{\sigma_\epsilon^2},$$

el cual está relacionado con el estadístico F de la primera etapa para testear la hipótesis de relevancia $\Lambda = 0$.

Considere el siguiente modelo para un regresor endógeno:

$$Y = X\beta + \epsilon$$
$$X = Z\Lambda + u$$

donde Y es un vector de $n \times 1$, X es una matriz de $n \times 1$, Z es una matriz de $n \times l$, y ϵ es un vector de $n \times L$ con varianza σ_ϵ^2.

El estimador 2SLS minimiza $(Y - X\beta)^T P_Z (Y - X\beta)$ y se define como

$$\hat{\beta}_{2SLS} = (X^T P_Z X)^{-1} (X^T P_Z Y),$$

donde P_Z es la matriz de proyección a las columnas del espacio vectorial generado por las Z. Rothenberg (1984) muestra que a medida que μ^2 crece, el estimador 2SLS converge en probabilidad y su distribución es estándar[5]:

$$\mu(\hat{\beta}_{2SLS} - \beta) = \left(\frac{\sigma_e}{\sigma_\epsilon}\right)\left(\frac{z_e + S_{eu}/\mu}{1 + 2z_u/\mu + S_u u/\mu^2}\right)$$

donde

$$z_\epsilon = \frac{\Lambda^T Z^T \epsilon}{\sigma_\epsilon \sqrt{\Lambda^T Z^T Z \Lambda}}; \quad z_u = \frac{\Lambda^T Z^T u}{\sigma_u \sqrt{\Lambda^T Z^T Z \Lambda}}$$

[5]Siguiendo la notación en Rau (2016).

$$S_{\epsilon u} = \frac{u^T P_Z \epsilon}{\sigma_\epsilon \sigma_u}; \quad S_{uu} = \frac{u^T P_Z u}{\sigma_u^2}$$

Es posible demostrar que bajo los supuestos de instrumentos fijos y errores normales, z_ϵ y z_u son variables aleatorias normales con coeficientes de correlación ρ, y $S_{\epsilon u}$ y S_{uu} son formas cuadráticas de variables aleatorias normales con respecto a la matriz de proyección P_Z. Como las distribuciones de z_ϵ, z_u, $S_{\epsilon u}$ y S_{uu} no dependen del tamaño de la muestra n, el tamaño de la muestra entra sólo a través del parámetro de concentración μ^2. Note que μ^2 juega el rol del tamaño muestral, es decir, si μ^2 es suficientemente grande entonces se tiene la aproximación normal usual. Por otro lado, si μ^2 es pequeño, la distribución asintótica no es estándar.

Cuando $\mu^2 = 0$, entonces

$$\mathbb{P}\lim(\hat{\beta}_{2SLS}) = \beta + \left(\frac{\sigma_\epsilon^2}{\sigma_\epsilon^2}\right)\rho \neq \beta.$$

Este caso extremo deja en evidencia cuán sensible puede ser el estimador 2SLS a la fuerza de los instrumentos.

Para determinar la validez de un instrumento, es crucial argumentar que no existe una relación directa entre el instrumento y la variable dependiente. En el ejemplo anterior, es necesario argumentar que la talla de la madre no influye en las habilidades cognitivas de los niños y niñas más allá de su efecto a través del estado nutricional de los mismos.

La validez de un instrumento puede ser evaluada mediante el test de Stock and Yogo (2005), que examina el porcentaje de sesgo aceptable. Se recomienda utilizar un 10 % de sesgo, con la hipótesis nula de que los instrumentos son débiles. Si $F_{\text{1etapa}} >$

Valor crítico al 10 %, entonces el instrumento es robusto; de lo contrario, es débil. Este test evalúa el ajuste conjunto de las variables en la primera etapa de la regresión.

Otra prueba utilizada es el test de Sargan (1958), que verifica si los instrumentos no están correlacionados con el término de error del modelo en la segunda etapa de la estimación por variables instrumentales (MC2E). Los pasos son:

1. estimar el modelo de MC2E y guardar los residuos,
2. realizar una regresión de los residuos con los instrumentos y
3. construir la prueba χ^2 donde los grados de libertad son $q-1$ (número de instrumentos menos uno). La hipótesis nula es que los instrumentos son exógenos.

El test de Hansen (1982) es similar al de Sargan, pero utiliza errores estándar robustos tanto en la estimación de los residuos como en la relación de los instrumentos con los errores del modelo de MC2E.

En general, la selección de instrumentos adecuados es crucial y puede ser evaluada mediante las pruebas mencionadas para asegurar la validez de los resultados obtenidos John Bound (1995); John Chao (2005); Douglas Staiger (1997); Hausman (1978); Rothenberg (1984); John G. Cragg (1993); Stock and Yogo (2005).

Ejemplo 77. Para evaluar la validez de los instrumentos en un modelo de variables instrumentales (IV), como vimos previamente, se pueden utilizar varios tests. A continuación, se presenta un ejemplo que muestra los resultados de varias pruebas de identificación y sobreidentificación.

lehr	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
yrsqual	.3276378	.0227545	14.40	0.000	.2830399	.3722357
sexo	-.5912689	.0419384	-14.10	0.000	-.6734666	-.5090711
urbano	.1793616	.0954352	1.88	0.060	-.007688	.3664112
exper	.0304977	.0057491	5.30	0.000	.0192297	.0417657
exper2	-.0006119	.0001496	-4.09	0.000	-.0009052	-.0003186
_cons	.5407223	.4104245	1.32	0.188	-.263695	1.34514

Underidentification test (Anderson canon. corr. LM statistic): 362.564
Chi-sq(2) P-val = 0.0000

Weak identification test (Cragg-Donald Wald F statistic): 202.316
Stock-Yogo weak ID test critical values: 10% maximal IV size 19.93
15% maximal IV size 11.59
20% maximal IV size 8.75
25% maximal IV size 7.25
Source: Stock-Yogo (2005). Reproduced by permission.

Sargan statistic (overidentification test of all instruments): 0.196
Chi-sq(1) P-val = 0.6576

Instrumented: yrsqual
Included instruments: sexo urbano exper exper2
Excluded instruments: pvnuml secomas_i

Figura 10.1 Stock y Yogo, Sargan.

El Underidentification test (Anderson canon. corr. LM statistic) verifica si el modelo está subidentificado. En el ejemplo, el estadístico de la prueba es 362.564 con un p-value de 0.0000, lo que indica que el modelo no está subidentificado. El Weak identification test John G. Cragg (1993) evalúa la fortaleza de los instrumentos. En el ejemplo, el estadístico de la prueba es 202.316. Según los valores críticos de Stock y Yogo Stock and Yogo (2005), que dependen del tamaño del sesgo que se quiera aceptar (10 %, 15 %, 20 %, 25 %), el valor crítico para un sesgo máximo del 10 % es 19.93. Dado que el estadístico es mayor que este valor crítico, se concluye que los instrumentos no son débiles. El Sargan statistic (overidentification test) descrito por Sargan (1958), verifica si los instrumentos están correlacionados con el término de error del modelo. En el ejemplo, el estadístico de la prueba es 0.196 con un $p-$value de 0.6576. La hipótesis nula es que los instrumentos son exógenos, y dado que el $p-$value es mayor que 0.05, no se rechaza la

hipótesis nula, sugiriendo que los instrumentos son válidos. El test de Hansen (1982) es similar al test de Sargan, pero utiliza errores estándar robustos. Aunque no se muestra en la figura, es otro test comúnmente utilizado para evaluar la sobreidentificación de los instrumentos. En resumen, estos tests proporcionan una manera robusta de evaluar la validez y fortaleza de los instrumentos en un modelo de variables instrumentales, asegurando que los resultados obtenidos sean fiables.

Previamente en este capítulo, hemos abordado temas cruciales como las variables instrumentales, el método de estimación en dos etapas (2SLS), el método de momentos generalizados (GMM), los instrumentos débiles y los tests estadísticos relevantes. Si bien no hemos tocado el estadístico de John G. Cragg (1993) ni el método de máxima verosimilitud de información limitada (LIML), permítanos comentar un poco al respecto. El estadístico de Cragg-Donald es fundamental para evaluar la fuerza de los instrumentos en modelos de regresión, proporcionando un criterio importante para determinar la validez de los instrumentos utilizados. Por otro lado, el LIML[6] es una técnica alternativa que, aunque menos común, ofrece ventajas específicas en ciertos contextos econométricos,

[6] Angrist and Krueger (1991a) en adelante AK, estiman el retorno a la educación (ecuación de Mincer) utilizando como instrumento el trimestre de nacimiento de las personas, basándose en los datos del Censo de 1980. La justificación es la ley estadounidense que estipula que los niños pueden comenzar el primer grado con seis años cumplidos al 30 de junio (empiezan en agosto). Aquellos nacidos en el primer y segundo trimestre comienzan la escuela en agosto con la edad ya cumplida. Dado que se puede abandonar la escuela a los 16 o 17 años, los nacidos en el primer o segundo trimestre lo

especialmente cuando los instrumentos son débiles o el número de instrumentos es grande en relación al tamaño de la muestra. Con estos fundamentos en mente, procederemos a explorar en detalle el estimador de Wald en la sección siguiente.

10.5. Estimador de Wald

Un caso particular de endogeneidad en el modelo de regresión lineal ocurre cuando la variable endógena es binaria. Este escenario es común en la literatura de evaluación de programas, ya que frecuentemente hay un programa o tratamiento que no se asignó aleatoriamente.

Supóngase que se tiene el siguiente modelo lineal:

$$Y_i = \alpha + \beta X_i + \epsilon_i$$

donde X_i es una variable binaria. Se puede observar que el estimador OLS es un estimador de diferencias. Este estimador representa la diferencia de medias condicionales de Y en X para $X = 1$ y $X = 0$.

harán antes de completar el año escolar, teniendo así menos escolaridad que los nacidos en el tercer y cuarto trimestre. Las variables dependientes son el logaritmo del salario por hora, escolaridad y dummies por año de nacimiento, y los instrumentos son dummies por trimestre de nacimiento, dummies por año de nacimiento e interacciones entre ambas. AK demuestran que, aunque los instrumentos pueden parecer débiles según algunos criterios, son válidos para identificar el modelo. Además, comparan los resultados de 2SLS con LIML y encuentran que mientras 2SLS converge a OLS cuando se utilizan muchos instrumentos, LIML resulta ser más consistente.

Para ilustrarlo un poco más, se debe notar que para lograr la identificación de β, OLS asume que $\mathbb{E}[\epsilon_i|X_i] = 0$. Esto implica que la variable binaria X_i de tratamiento es independiente del nivel medio condicional de los no observables. Es decir, en promedio, no hay factores no observables que se relacionen con el hecho de recibir o no el tratamiento. Esto se logra mediante aleatorización. Bajo ese supuesto, el estimador OLS se puede expresar de la siguiente manera:

$$\hat{\beta}_{OLS} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(X_i - \bar{X})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \tag{10.9}$$

$$= \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i} - \frac{\sum_{i=1}^{n} Y_i(1 - X_i)}{\sum_{i=1}^{n}(1 - X_i)}$$

$$= \hat{\mathbb{E}}[Y|X = 1] - \hat{\mathbb{E}}[Y|X = 0], \tag{10.10}$$

donde[7]

$$\hat{\mathbb{E}}[Y|X = 1] = \frac{\sum_{i=1}^{n} Y_i X_i}{\sum_{i=1}^{n} X_i}$$

y

$$\hat{\mathbb{E}}[Y|X = 0] = \frac{\sum_{i=1}^{n} Y_i(1 - X_i)}{\sum_{i=1}^{n}(1 - X_i)}.$$

Así, el estimador OLS en este caso es un estimador de diferencias,

$$\hat{\beta}_{OLS} = \bar{Y}_T - \bar{Y}_C = \hat{\mathbb{E}}[Y|X = 1] - \hat{\mathbb{E}}[Y|X = 0].$$

En consecuencia, si X_i fuera un tratamiento producto de un experimento aleatorio, el estimador OLS estima el efecto causal del tratamiento sobre la variable de resultados Y.

[7] Tenemos. $\mathbb{E}[Y_i|X_i = 0] = \alpha$ Por lo tanto, la diferencia en el valor esperado de Y_i entre aquellos con $X_i = 1$ y $X_i = 0$ es $\mathbb{E}[Y_i|X_i = 1] - \mathbb{E}[Y_i|X_i = 0] = (\alpha + \beta) - \alpha = \beta$.

Sin embargo, si $\mathbb{E}[\epsilon_i|X_i] \neq 0$, el tratamiento no proviene de un experimento aleatorio y está correlacionado con alguna característica no observable del individuo i. Esto ocurre, por ejemplo, si los individuos se autoseleccionan en un determinado programa. En este caso, el parámetro β no está identificado y el estimador de OLS es inconsistente y sesgado.

Por otro lado, si se dispone de una variable instrumental binaria Z_i que cumple con los supuestos fundamentales: $\mathbb{E}[\epsilon_i|Z_i] = 0$ y $\mathbb{E}[Z_i X_i] \neq 0$, se puede identificar β y obtener un efecto causal del tratamiento sobre Y mediante el estimador de variables instrumentales,

$$\hat{\beta}_{\text{Wald}} = \frac{\sum_{i=1}^{n}(Y_i - \bar{Y})(Z_i - \bar{Z})}{\sum_{i=1}^{n}(X_i - \bar{X})(Z_i - \bar{Z})}.$$

Este estimador también puede expresarse como un estimador de diferencias pero con un ajuste en el denominador. Note que con un poco de álgebra se obtiene:

$$\hat{\beta}_{IV} = \frac{\hat{\mathbb{E}}[Y|Z=1] - \hat{\mathbb{E}}[Y|Z=0]}{\hat{\mathbb{E}}[X|Z=1] - \hat{\mathbb{E}}[X|Z=0]}$$

donde

$$\hat{\mathbb{E}}[Y|Z=1] = \frac{\sum_{i=1}^{n} Y_i Z_i}{\sum_{i=1}^{n} Z_i}$$

$$\hat{\mathbb{E}}[Y|Z=0] = \frac{\sum_{i=1}^{n} Y_i(1-Z_i)}{\sum_{i=1}^{n}(1-Z_i)}$$

$$\hat{\mathbb{E}}[X|Z=1] = \frac{\sum_{i=1}^{n} X_i Z_i}{\sum_{i=1}^{n} Z_i}$$

y

$$\hat{\mathbb{E}}[X|Z=0] = \frac{\sum_{i=1}^{n} X_i(1-Z_i)}{\sum_{i=1}^{n}(1-Z_i)}.$$

Este estimador ($\hat{\beta}_{\text{Wald}}$) se conoce como Estimador de Wald.

Ejemplo 78. La lotería de Vietnam. Angrist (1990a) estudia el impacto de servir en la guerra de Vietnam en el salario de los veteranos, años después. Esta pregunta es interesante porque existe la hipótesis de que haber servido en Vietnam podría compensar la pérdida de experiencia en el mercado laboral. Sin embargo, existe un problema de endogeneidad en la estimación de un modelo de regresión lineal del efecto de haber servido en la guerra en el salario. El problema de endogeneidad es claro, ya que las personas que sirven voluntariamente en una guerra tienen otras características no observables que pueden estar correlacionadas con habilidad, capital social, etc. Durante la Guerra de Vietnam se realizaron cinco loterías en Estados Unidos para enviar jóvenes a la guerra. Por ejemplo, la lotería de 1970 cubrió a jóvenes entre 19 y 26 años. Se sortearon números (sin reemplazo) del 1 al 365, asignando cada número a una fecha de nacimiento (día, mes). Las personas eran llamadas según una secuencia de números (del 1 al 365) hasta cumplir la cuota requerida por el Departamento de Defensa. Posteriormente, se realizaban exámenes médicos y se seleccionaba a quienes irían a la guerra. Aquí se puede utilizar un estimador de Wald. Se puede crear un instrumento binario (1 si se tiene un número de sorteo bajo, 0 si es alto). Este instrumento está correlacionado con servir en la guerra, pero no con otras características. Así, el estimador de variables instrumentales es:

$$\hat{\beta}_{IV} = \frac{\hat{\mathbb{E}}[Y|Z=1] - \hat{\mathbb{E}}[Y|Z=0]}{\hat{\mathbb{E}}[X|Z=1] - \hat{\mathbb{E}}[X|Z=0]}$$

Note que el numerador es un estimador de diferencias, pero no condicional en el tratamiento, sino en el instrumento. El

denominador proporciona la diferencia de las probabilidades de ser tratado condicional al instrumento. En este caso, condicional a tener un número bajo o alto de lotería.

Capítulo 11

Máxima Verosimilitud

A lo largo de este texto, la herramienta de estimación principal ha sido la estimación vía Mínimos Cuadrados Ordinarios. Sin embargo, en este capítulo, presentamos un segundo método de estimación que resulta de gran interés en la práctica y que, si bien no ha sido usado previamente, aparece con frecuencia en el análisis estadístico y econométrico a la hora de, por ejemplo, trabajar con modelos donde la variable de regresión es dicotómica (modelo de probabilidad lineal, Logit, Probit).

11.1. Estimación

Definición 11.1.1. Sea $f(\cdot|\theta)$ con $\theta \in \Theta \subset \mathbb{R}^k$ una familia paramétrica de distribuciones. Sea $X = (X_1, ..., X_n)$ una muestra aleatoria iid de una distribución $g(\cdot|\theta_0)$ con $\theta_0 \in \Theta$. Entonces, la

densidad conjunta es

$$f(x|\theta) = \prod_{i=1}^{n} g(x_i|\theta),$$

donde x_i es la observación de X_i. Definimos la función de verosimilitud de esta muestra aleatoria de la siguiente manera

$$L(\theta|x) = \begin{cases} \prod_{i=1}^{n} \mathbb{P}_{X_i}(x_i|\theta) & \text{variable discreta} \\ \prod_{i=1}^{n} f_{X_i}(x_i|\theta) & \text{variable continua.} \end{cases}$$

Note que las distribuciones dependen desde ahora de un (vector) de parámetros $\theta \in \Theta \subset \mathbb{R}^k$. En ese sentido, podemos escribir, teniendo en cuenta que $X_1, ..., X_n$ son iid

$$f_{(X_1,\cdots,X_n)}(x, \theta) = \prod_{i=1}^{n} f_X(x_i, \theta).$$

Eventualmente, para abreviar la notación, escribiremos $L(\theta)$.

Ejemplo 79. Sea

$$f_X(x; \theta) = \begin{cases} \theta^2 x e^{-\theta x}, & \text{si } x > 0 \\ 0, & \text{en otro caso.} \end{cases}$$

La función de densidad de una variable aleatoria continua, $\theta > 0$. Entonces,

$$\begin{aligned} L(\theta) &= \prod_{i=1}^{n} f_{X_i}(x_i|\theta) \\ &= \prod_{i=1}^{n} \theta^2 x_i e^{-\theta x_i} \\ &= \theta^{2n} \left(\prod_{i=1}^{n} x_i \right) e^{-\theta \sum_{i=1}^{n} x_i}. \end{aligned}$$

Definición 11.1.2. Estimador de máxima verosimilitud. El estimador de máxima verosimilitud

$$\hat{\theta}_{MV} = \phi(X_1, ..., X_n)$$

es el estimador que resuelve

$$\mathcal{P} : \begin{cases} \text{máx} & L(\theta) \\ \text{s. a} & \theta \in \Theta. \end{cases}$$

En relación al estimador de máxima verosimilitud, es posible que la m.a. $X_1, ..., X_n$ no sea iid. Por otro lado, usualmente es más sencillo maximizar $\ell(\theta) = \ln(L(\theta))$.

Luego, la CPO provee

$$\frac{1}{n}\sum_{i=1}^{n} \frac{\partial \ell(\hat{\theta}_{MV}; x_i)}{\partial \theta} = 0.$$

Ejemplo 80. Una variable aleatoria X posee la siguiente función de densidad

$$f(x) = \begin{cases} \frac{1}{\theta} e^{-\frac{x}{\theta}}, & \text{si } x \geq 0 \\ 0, & \text{caso contrario.} \end{cases}$$

Veamos que el estimador de máxima verosimilitud $\hat{\theta}$ es igual a la media muestral. Primero, notemos que, efectivamente $f(\cdot)$ es un densidad para $\theta > 0$:

$$\int_{\mathbb{R}} f(x)dx = \int_0^{\infty} \frac{1}{\theta} e^{-\theta x} dx = \lim_{t \to \infty} (e^{-t/\theta} + 1) = 1.$$

Ahora, siguiendo la definición dada,

$$\hat{\theta} \in \text{argmax}_{\theta \in \Theta} \prod_{i=1}^{n} f_{X_i}(x_i|\theta).$$

Dado que $\ln(\cdot)$ es una función creciente,

$$\hat{\theta} \in \text{argmax}_{\theta \in \Theta} \underbrace{\ln\left(\prod_{i=1}^{n} f_{X_i}(x_i|\theta)\right)}_{=K(\theta)}.$$

Primero, calculamos $K(\theta)$

$$\begin{aligned} K(\theta) &= \ln\left(\prod_{i=1}^{n} f(x_i)\right) \\ &= \ln\left(\prod_{i=1}^{n} \frac{e^{-x_i/\theta}}{\theta}\right) \\ &= \ln\left(\frac{1}{\theta^n} e^{-\frac{\sum_{i=1}^{n} x_i}{\theta}}\right) \\ &= -n\ln\theta - \frac{1}{\theta}\sum_{i=1}^{n} x_i. \end{aligned}$$

Aplicando condiciones de primer orden,

$$-\frac{n}{\hat{\theta}} + \frac{1}{\hat{\theta}^2}\sum_{i=1}^{n} x_i.$$

Así,

$$\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Ejemplo 81. Sea $(x_1, x_2, ..., x_n)$ una muestra aleatoria correspondiente a una distribución normal $NS(\mu, \sigma^2)$. La función de verosimilitud es

$$\begin{aligned} L(x_1, x_2, ..., x_n; \mu, \sigma^2) &= \prod_{i=1}^{n} f(x_i; \mu, \sigma^2) \\ &= \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}. \end{aligned} \tag{11.1}$$

Computemos los estimadores de máxima verosimilitud de μ y σ^2 (que se denotan por $\hat{\mu}$ y $\hat{\sigma}^2$). Nuestro objetivo es resolver

$$\mathcal{P}: \begin{cases} \text{máx}_{\mu,\sigma^2} & L(x_1,...,x_n;\mu,\sigma^2) \\ s.a.: & (\mu,\sigma) \in \Theta = \mathbb{R}\times\mathbb{R}_+. \end{cases}$$

Un primer enfoque, para poder encontrar los candidatos a óptimos locales, es aplicar directamente las condiciones de primer orden a la función (11.1):

$$\frac{\partial L}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} = 0 \tag{11.2}$$

$$\frac{\partial L}{\partial \sigma^2} = \frac{-n}{(\sigma^2)^{\frac{n+1}{2}}\sqrt{2\pi}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} + \frac{\sum_{i=1}^{n}(x_i-\mu)^2}{(\sigma^2)^{n/2}\sqrt{2\pi}\cdot(\sigma^2)^{3/2}}e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} = 0. \tag{11.3}$$

De (11.2), como la exponencial es siempre positiva, se tiene que

$$\sum_{i=1}^{n}(x_i-\mu) = 0$$
$$\frac{\sum_{i=1}^{n}x_i}{n} = \overline{x} = \hat{\mu}.$$

Luego, teniendo en cuenta nuevamente que $e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2} > 0$ y simplificando (11.3), obtenemos

$$\frac{1}{\sigma^n\sqrt{2\pi}}\cdot\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^3} = \frac{n}{\sigma^{n+1}\sqrt{2\pi}^n}$$
$$\frac{\sum_{i=1}^{n}(x_i-\mu)^2}{\sigma^{n+3}} = \frac{n}{\sigma^{n+1}}$$
$$m\sigma^2 = \sum_{i=1}^{n}(x_i-\mu)^2.$$

Evaluando en el óptimo,

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Ahora bien, dada la estructura de la función de máxima verosimilitud, considerando sobre todo las condiciones de segundo orden, lo más acertado es maximizar la función de log-verosimilitud, definida de la siguiente manera

$$\begin{aligned} K(\theta) &= K(\mu, \sigma^2|x) \\ &= \ln(L(\mu, \sigma^2; x_1, ..., x_n)) \\ &= \ln\left[\left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2}\right] \\ &= -n\ln\sqrt{2\pi} - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2}\ln\sigma^2. \end{aligned}$$

Debido a las propiedades de la función logaritmo neperiano, $K(\theta)$ posee los mismos óptimos, y de misma naturaleza, que $L(x;\theta)$ Casella and Berger (2002).

Por ende, bastaba con aplicar las CPO a $K(\mu, \sigma^2)$

$$\begin{aligned} \frac{\partial K}{\partial \mu} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 = 0 \\ \frac{\partial K}{\partial \sigma^2} &= \frac{1}{2(\sigma^2)^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{n}{2\sigma^2}. \end{aligned}$$

Despejando, se vuelven a obtener los candidatos a máximo local

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n}x_i = \overline{x}$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\hat{\mu})^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \frac{1}{n}\sum_{j=1}^{n}x_j\right)^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i-\overline{x})^2.$$

Queda entonces por analizar las condiciones de segundo orden. Esto es[1], verificar que $D_1 \leq 0$ y $D_2 \geq 0$, siendo D_i los menores principales de la matriz $HK(\mu, \sigma^2) \in \mathcal{M}_{2\times 2}$ evaluada en $\hat{\theta} = (\hat{\mu}, \hat{\sigma}^2)$.

$$\begin{aligned}
\frac{\partial^2 K}{\partial \mu^2} &= -\frac{n}{\sigma^2} \\
\frac{\partial^2 K}{\partial \mu \partial \sigma^2} &= \frac{\partial^2 K}{\partial \sigma^2 \partial \mu} = -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \\
\frac{\partial^2 K}{\partial (\sigma^2)^2} &= \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2.
\end{aligned}$$

Luego,

$$HK(\mu, \sigma^2) = \begin{bmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) \\ -\frac{1}{\sigma^4}\sum_{i=1}^{n}(x_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\sum_{i=1}^{n}(x_i - \mu)^2 \end{bmatrix}.$$

Por ende,

$$D_1 = H_{11}(K(\hat{\mu}, \hat{\sigma}^2)) = -\frac{n}{\hat{\sigma}^2} < 0$$

[1] Ver Casella and Berger (2002).

y

$$\begin{aligned}
D_2 &= \det(H(K(\hat{\mu}, \hat{\sigma}^2))) \\
&= \frac{1}{\sigma^6}\left[-\frac{n^2}{\sigma^2} + \frac{n}{\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2 - \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)\right)^2\right]\Bigg|_{\mu=\overline{x},\sigma^2=\hat{\sigma}^2} \\
&= \frac{1}{\sigma^6}\left[-\frac{n^2}{\sigma^2} + \frac{n}{\sigma^2}\sigma^2 - \frac{1}{\sigma^2}\left(\sum_{i=1}^{n}(x_i-\mu)\right)^2\right]\Bigg|_{\mu=\overline{x},\sigma^2=\hat{\sigma}^2} \\
&= \frac{1}{\hat{\sigma}^6}\left[-\frac{n^2}{\hat{\sigma}^2} + \frac{n}{\hat{\sigma}^2}\hat{\sigma}^2 - \frac{1}{\hat{\sigma}^2}\left(\sum_{i=1}^{n}(x_i-\hat{\mu})\right)^2\right] \\
&= -\frac{n}{\hat{\sigma}^2} - \frac{n}{2\hat{\sigma}^6} \\
&= \frac{n^2}{2\hat{\sigma}^6} > 0.
\end{aligned}$$

Concluimos entonces, a través de las condiciones de segundo orden, que $(\hat{\mu}, \hat{\sigma}^2)$ es en efecto un local.

Ejemplo 82. Supongamos que

$$Y_i = \beta_1 + \beta_2 X_i + \epsilon_i,\ \epsilon_i \sim N(0, \sigma^2).$$

En este caso, $\mathbb{E}[Y_i|X_i] = \beta_1 + \beta_2 X_i$. Así,

$$\begin{aligned}
L(\theta) &= L(\beta_1, \beta_2, \sigma) \\
&= \prod_{i=1}^{n} f_{Y_i}(y_i|\beta, \sigma) \\
&= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(Y_i - \beta_1 - \beta_2 X_i)^2}{2\sigma^2}\right) \\
&= \frac{1}{(\sqrt{2\pi\sigma^2})^n} \exp\left(-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \beta_1 - \beta_2 X_i)^2\right).
\end{aligned}$$

Luego de resolver el problema de maximización para obtener los estimadores, se llega a

$$\begin{aligned}\hat{\beta}_1 &= \overline{Y} - \hat{\beta}_2\overline{X}\\ \hat{\beta}_2 &= \frac{\sum_{i=1}^n X_iY_i - \overline{XY}}{\sum_{i=1}^n X_i^2 - n\overline{X}^2}\\ \hat{\sigma}^2 &= \frac{1}{n}\sum_{i=1}^n \hat{\epsilon}_i^2\end{aligned}$$

con $\hat{\epsilon}_i = Y_i - \hat{Y}_i = Y_i - \hat{\beta}_1 - \hat{\beta}_2 X_i$.

Ejemplo 83. En el modelo de k-variables,

$$Y_i = \beta_1 + \beta_2 X_{2i} + \beta_3 X_{23} + ... + \beta_k X_{2k} + \epsilon_i,$$

que matricialmente se expresa como

$$\underbrace{\begin{bmatrix} Y_1 \\ \vdots \\ Y_n \end{bmatrix} = \begin{bmatrix} 1 & X_{21} & X_{31} & \cdots & X_{k1} \\ \vdots & \vdots & \vdots & \cdots & \vdots \\ 1 & X_{2n} & X_{2n} & \cdots & X_{kn} \end{bmatrix} \begin{bmatrix} \beta_1 \\ \vdots \\ \beta_k \end{bmatrix} + \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix}}_{=Y_{n\times 1}=X_{n\times k}\beta_{k\times 1}+\epsilon_{n\times 1}}$$

teníamos que $\hat{Y} = X\hat{\beta}$ y $\epsilon \sim N(0_n, \sigma^2 I_n)$. Así,

$$L(\theta) = \frac{1}{(2n\sigma^2)^{\frac{n}{2}}} e^{-\frac{(Y-X\beta)^T(Y-X\beta)}{2\sigma^2}}.$$

Las CPO proveen

$$\begin{aligned}\hat{\beta} &= (X^TX)^{-1}X^TY\\ \hat{\sigma}^2 &= \frac{\hat{\epsilon}^T\hat{\epsilon}}{n}.\end{aligned}$$

11.2. La cota inferior de Cramer-Rao

Teorema 39. Supongamos que θ es un parámetro determinístico no conocido que debe ser estimado a partir de una muestra de n observaciones iid con densidad $f(x;\theta)$. La varianza de cualquier estimador insesgado $\hat{\theta}$ de θ es acotado inferiormente por la recíproca de la información de Fisher:

$$\mathrm{Var}(\hat{\theta}) \geq \frac{1}{nI(\theta)} \tag{11.4}$$

con

$$I(\theta) = \mathbb{E}_{X;\theta}\left[\left(\frac{\partial \ell(X;\theta)}{\partial \theta}\right)^2\right]$$

siendo $\ell(x;\theta) = \ln L(\theta;x)$ el logaritmo neperiano de la función de verosimilitud y $\mathbb{E}_{X;\theta}$ el valor esperado con respecto a la densidad $f(x;\theta)$ de X.

En el caso más general, si $\delta(X)$ es un estimador insesgado de $g(\theta)$, (11.4) se convierte en

$$\mathrm{Var}(\delta) \geq \frac{[g'(\theta)]^2}{nI(\theta)}.$$

Teorema 40. Asumiendo que se cumple la regla de Leibniz, se tiene que

$$I(\theta) = \mathrm{Var}\left[\frac{\partial}{\partial \theta}\ln f(X|\theta)\right]^2 = -\mathbb{E}\left[\frac{\partial^2}{\partial \theta^2}\ln f(X|\theta)\right].$$

Demostración. Para simplificar la notación, escribimos $\ell(\theta) =$

$\ln f(x|\theta)$. Luego,

$$\begin{aligned}
\mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(\theta)\right] &= \int \left[\frac{\partial}{\partial\theta}\ell(\theta)\right] f(x|\theta)dx \\
&= \int \left[\frac{\partial}{\partial\theta}f(x|\theta)\right] \frac{1}{f(x|\theta)} f(x|\theta)dx \\
&= \int \left[\frac{\partial}{\partial\theta}f(x|\theta)\right] dx \\
&= \frac{\partial}{\partial\theta}\int f(x|\theta)dx \\
&= \frac{\partial}{\partial\theta}(1) = 0.
\end{aligned}$$

Ahora bien, como $\mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(\theta)\right] = 0$

$$\begin{aligned}
0 &= \frac{\partial}{\partial\theta}\int \left[\frac{\partial}{\partial\theta}\ell(\theta)\right] f(x|\theta)dx \\
&= \int \left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] f(x|\theta)dx + \int \left[\frac{\partial}{\partial\theta}\ell(\theta)\right]\left[\frac{\partial}{\partial\theta}f(x|\theta)\right] dx \\
&= \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] + \int \left[\frac{\partial}{\partial\theta}\ell(\theta)\right]\left[\frac{\partial}{\partial\theta}\ell(\theta)\right] f(x|\theta)dx \\
&= \mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right] + \mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(\theta)\right]^2.
\end{aligned}$$

Así,

$$\mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(\theta)\right]^2 = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right].$$

Finalmente, concluimos notando que

$$\text{Var}\left[\frac{\partial}{\partial\theta}\ell(\theta)\right] = \mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(\theta)\right]^2 - \underbrace{\left(\mathbb{E}\left[\frac{\partial}{\partial\theta}\ell(\theta)\right]\right)^2}_{=0}.$$

□

A continuación, probamos el Teorema de la cota inferior de Cramer-Rao en el contexto más general, es decir, considerando $g(\theta)$.

Demostración. Sea

$$g(\theta) = \iint \cdots \iint \delta(x) f(x|\theta) dx. \tag{11.5}$$

Tomando la derivada respecto a θ en ambos lados de la Ecuación 11.5,

$$\begin{aligned} g'(\theta) &= \iint \cdots \iint \delta(x) \frac{\partial}{\partial \theta} f(x|\theta) dx. \\ &= \iint \cdots \iint \delta(x) \frac{\partial}{\partial \theta} \ln f(x|\theta) dx. \\ &= \mathbb{E}\left[\delta(X) \frac{\partial}{\partial \theta} \ln f(X|\theta)\right]. \end{aligned}$$

Ahora bien, como

$$\mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(X|\theta)\right] = n\mathbb{E}\left[\frac{\partial}{\partial \theta} \ln f(X|\theta)\right] = 0,$$

concluimos que

$$g'(\theta) = \text{Cov}\left(\delta(X), \frac{\partial}{\partial \theta} \ln f(X|\theta)\right).$$

Debido a la desigualdad de Cauchy-Schwarz,

$$\text{Cov}(X, Y)^2 \leq \text{Var}(X)\text{Var}(Y),$$

con igualdad solo si $X = aY + b$,

$$\begin{aligned} [g'(\theta)]^2 &\leq \text{Var}(\delta(X))\text{Var}\left(\frac{\partial}{\partial \theta} \ln f(x|\theta)\right) \\ &= \text{Var}(\delta(X)) \cdot n \cdot \text{Var}\left(\frac{\partial}{\partial \theta} \ln f(X|\theta)\right) \\ \text{Var}(\delta(X)) &\geq \frac{[g'(\theta)]^2}{nI(\theta)}. \end{aligned}$$

□

11.3. Propiedades asintóticas

Recordemos que $\hat{\theta}_{MV}$ maximiza

$$\ell_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} \ln f(x_i|\theta).$$

Definamos

$$\ell(\theta) = \mathbb{E}_{\theta_0}[\ell(x|\theta)],$$

donde θ_0 es el parámetro de la muestra $X_1, ..., X_n$. En caso la distribución fuese continua,

$$\ell(\theta) = \int \ln f(x|\theta) f(x|\theta_0) dx.$$

Por la ley de los grandes números,

$$\ell_n(\theta) \to \ell(\theta).$$

Teorema 41. Para todo $\theta \in \Theta$,

$$L(\theta) \leq L(\theta_0).$$

Más aún, la desigualdad es estricta salvo que

$$\mathbb{P}_{\theta_0}(f(x|\theta) = f(x|\theta_0)) = 1.$$

Demostración. Consideremos

$$\begin{aligned} L(\theta) - L(\theta_0) &= \mathbb{E}_{\theta_0}[\ln f(x|\theta) - \ln f(x|\theta_0)] \\ &= \mathbb{E}_{\theta_0} \ln \frac{f(x|\theta)}{f(x|\theta_0)}. \end{aligned}$$

Dado que $\ln t \leq t - 1$,

$$\begin{aligned}
\mathbb{E}_{\theta_0} \ln \frac{f(x|\theta)}{f(x|\theta_0)} &\leq \mathbb{E}_{\theta_0} \left[\frac{f(x|\theta)}{f(x|\theta_0)} - 1 \right] \\
&= \int \left(\frac{f(x|\theta)}{f(x|\theta_0)} - 1 \right) f(x|\theta_0) dx \\
&= \int f(x|\theta) dx - \int f(x|\theta_0) dx \\
&= 1 - 1 = 0.
\end{aligned}$$

Esto nos permite concluir. □

Teorema 42. En el contexto descrito en este capítulo, si θ_0 es tal que para todo $\theta \neq \theta_0$ existe x con $f(x|\theta) \neq f(x|\theta_0)$[2], el soporte[3] de $f(\cdot|\theta)$ no depende de θ y $\theta_0 \in \Theta^\circ$, entonces $\hat{\theta}_{ML} \to \theta_0$ en probabilidad.

Teorema 43. Supongamos que se satisfacen las condiciones del Teorema 42. Asumamos además que $g(x_i|\theta)$ es clase $C^3(\Theta)$, que se cumplen las hipótesis del teorema de Leibiniz[4] en el par (x, θ) y que[5]

$$\left| \frac{\partial^3 \ln g(x_i|\theta)}{\partial \theta^3} \right| \leq M(x), \ \mathbb{E}[M(X_i)] < \infty, \ \forall \ i.$$

[2] La generalización es $\mathbb{P}\{f(x|\theta) \neq f(x|\theta_0) > 0$.

[3] S tal que $\int_S dF = 1$.

[4] $f(x,t)$ función tal que $f_x(x,t)$ es continua en t y x para alguna región del plano que incluye $a_1(x) \leq t \leq a_2(x)$, $x_0 \leq x \leq x_1$. Supóngase además que $a, b \in U$ con $[x_0, x_1] \subset U$. Entonces,

$$\frac{d}{dx} \left(\int_{a(x)}^{b(x)} f(x,t) dt \right) = f(x, b(x)) b'(x) - f(x, a(x)) a'(x) + \int_{a(x)}^{b(x)} f_x(x,t) dt.$$

[5] $M_X(t) = e^{tX}$.

Entonces,[6]

$$\sqrt{n}(\hat{\theta}_{MV} - \theta_0) \to N(0, I_g^{-1}(\theta_0))$$

con

$$I_g(\theta) = \mathbb{E}_\theta \left[\left(\frac{\partial \ell(\theta|X)}{\partial \theta} \right)^2 \right].$$

Las pruebas de los Teoremas 42 y 43 se encuentran en las siguientes notas de clase.

Los resultados de esta sección se resumen en el siguiente teorema.

Teorema 44. Sea $X_1, ..., X_n$ una muestra aleatoria de una distribución con parámetro θ. Sea $\hat{\theta}_{MV}$ el estimador de máxima verosimilitud de θ. Entonces, bajo ciertas condiciones de regularidad estándares[7],

1. $\hat{\theta}_{MV}$ es asintóticamente consistente, es decir, para todo $\epsilon > 0$

$$\lim_{n\to\infty} \mathbb{P}\{|\hat{\theta}_{MV} - \theta| > \epsilon\} = 0.$$

2. $\hat{\theta}_{MV}$ es asintóticamente insesgado, es decir,

$$\lim_{n\to\infty} \mathbb{E}[\hat{\theta}_{MV}] = \theta$$

3. Cuando $n \to \infty$,

$$\frac{\hat{\theta}_{MV} - \theta}{\sqrt{\text{Var}(\hat{\theta}_{MV})}} \to N(0, 1)$$

en distribución.

[6] La convergencia es en distribución.

[7] Compacidad de Θ, existencia de $\mathbb{E}[\sup_{\theta\in\Theta|\ell_{\theta\theta}|}]$, $\theta_0 \in \Theta^\circ$, diferenciabilidad y validez de la regla de Leibniz, entre otros.

11.4. Computación

A continuación algunas técnicas para computar el estimado de máxima verosimilitud. Seguimos fundamentalmente a Wooldridge (2001) y Rau (2016). Esto además cierra el capítulo e invitamos al lector consultar, por ejemplo, Weiss (1971), Self and Liang (1987) o Wooldridge (2001) para una discusión más profunda y extensa sobre los temas abordados en este capítulo.

Lo que buscamos es $\hat{\theta}_N$ tal que

$$\mathbb{E}_N[L_\theta(\hat{\theta}_N)] = \frac{1}{N}\sum_{i=1}^{N} \ell(\theta, x) = 0.$$

1. Búsqueda de grilla: se busca resolver $\max_{\theta\in[a,b]} R(\theta)$. Para ellos se divide $[a, b]$ en sub-intervalos $\{[a, \theta_1], \cdots [\theta_n, b]\}$ y se evalúa R en θ_i. Luego, se escoge donde R toma el valor más grande (digamos θ_i) y se escogen los intervalos $[\theta_{i-1}, \theta_i]$ y $[\theta_i, \theta_{i+1}]$. Se itera (se suele iterar) hasta que $|R(\theta_i) - R(\theta_{i+1})| < 10^{-5}$ o $|\theta_i - \theta_{i+1}| < 10^{-5}$.

2. Aproximación por polinomios: $R(\theta) = a + b(\theta - \theta_0) + \frac{1}{2}c(\theta - \theta_0)^2$. La CPO provee $\theta^* = \theta_0 - \frac{b}{c}$. Los coeficientes a, b y c se obtiene aplicando una expansión de Taylor de orden 2. Esto se repite para diferentes θ_0 y se elige aquel que maximice la función objetivo.

3. Búsqueda de línea: dado un valor inicial θ_1 y una dirección de búsqueda δ, resolvemos

$$\lambda^* = \text{argmax}_\lambda \; R(\theta_1 + \lambda\delta).$$

Luego, se toma $\theta_2 = \theta_1 + \lambda^* \delta$.

4. Forma cuadrática: planteamos

$$R(\theta) = a + b^T \theta + \frac{1}{2} \theta^T C \theta,$$

con C simétrica. Luego,

$$\frac{\partial R}{\partial \theta} = b + C\theta$$
$$\frac{\partial^2 R}{\partial \theta^2} = C^T.$$

R alcanza su máximo en $\theta^* = -C^{-1}b$. Notemos que

$$\theta^* = -C^{-1}b$$
$$\theta^* = \theta_1 - C^{-1}(b + C\theta_1)$$
$$\theta^* = \theta_1 - R_{\theta\theta}(\theta_1)^{-1} R_\theta(\theta_1).$$

Haciendo $\delta = -R_{\theta\theta}(\theta_1)^{-1} R_\theta(\theta_1)$ y $\lambda = 1$, se aplica una búsqueda lineal.

5. Newton-Raphson: sea

$$\delta_{NR} = \left\{ -\mathbb{E}_N \left[\frac{\partial^2 \ell}{\partial \theta^2}(\theta_1) \right] \right\}^{-1} \mathbb{E}_N \left[\frac{\partial \ell}{\partial \theta}(\theta_1) \right]$$

y λ_1. Así,

$$\theta_{k+1} = \theta_k + \delta_{NR}.$$

La derivación de Newton-Raphson se basa en una expansión de Taylor de segundo orden de $\mathbb{E}_N[\ell(\theta)]$ en θ_1.

6. Algoritmo BHHH (Brend-Hall-Hall-Hausman): se imputa en la búsqueda lineal

$$\delta_{BHHH} = \left\{ \mathbb{E}_N \left[\frac{\partial \ell}{\partial \theta}(\theta_1) \frac{\partial \ell}{\partial \theta}(\theta_1)^T \right] \right\} \mathbb{E}_N \left[\frac{\partial \ell}{\partial \theta}(\theta_1) \right].$$

Finalmente, un criterio muy común de convergencia[8], es

$$||\theta_i - \theta_{i-1}|| < 10^{-5}.$$

Cuando la función de log verosimilitud es estrictamente cóncava, entonces los métodos descritos funcionan por lo general bien[9] pues solo existe un único máximo global.

[8]Tanto en econometría como en otras disciplinas: física computacional, química computacional etc.

[9]Asumiendo diferenciabilidad del orden adecuado.

Apéndices

Apéndice A

Elementos de teoría de la probabilidad

En este apéndice brindamos los fundamentos de la teoría de la probabilidad que son de gran utilidad para tener un entendimiento más profundo y adecuado de los temas desarrollados en este texto. Asimismo, constituye una base sólida para el estudio de temas más avanzados en econometría. Este apéndice está basado en las notas de clase de los cursos dictados en la PUCP, Análisis Real 2 (dictado por el profesor Johel Beltrán) y Probabilidad y Estadística (dictado por el profesor Jonathan Farfán).

Definición A.0.1. Dado un conjunto Ω, un $\sigma-$álgebra sobre Ω es una colección de conjuntos de Ω, $\mathscr{F} \subset \mathcal{P}(\Omega)$[1] tales que

1. $\Omega \in \mathscr{F}$

[1] $\mathcal{P}(\Omega)$ denota el conjunto potencia de Ω: todos los posibles sub-conjuntos de Ω

2. $A \in \mathscr{F} \implies A^c \in \mathscr{F}$

3. $A_n \in \mathscr{F},\ \forall\ n \in \mathbb{N} \implies \bigcup_{n\in\mathbb{N}} A_n \in \mathscr{F}$.

Definición A.0.2. Sea $(\Omega, \mathscr{F})$ un espacio medible[2] Una medida sobre $(\Omega, \mathscr{F})$ es una aplicación $\mu : \mathscr{F} \to \mathbb{R}_+ \cup \{\infty\}$[3] tal que

1. $\mu(\emptyset) = 0$

2. si $A_n \in \mathscr{F},\ \forall\ n \in \mathbb{N}$ y $A_n \cap A_m = \emptyset$, $n \neq m$, entonces

$$\mu\left(\biguplus_{n\in\mathbb{N}} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

Definición A.0.3. Una medida de probabilidad $\mathbb{P}$ sobre un espacio de medida $(\Omega, \mathscr{F})$ es una medida tal que

$$\mathbb{P} : \mathscr{F} \to [0, 1]$$

y se cumple que $\mathbb{P}(\Omega) = 1$. A partir de esto se deduce que

1. $\mathbb{P}(A) + \mathbb{P}(A^c) = 1,\ \forall\ A \in \mathscr{F}$.

2. $\{A_1, ..., A_n, ...\}$ disjuntos dos a dos: $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

3. $\mathbb{P}(A \cup B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B)$.

4. $\mathbb{P}(A) + \mathbb{P}(B - A) = \mathbb{P}(A \cup B)$.

5. $\mathbb{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathbb{P}(A_n)$.

A continuación, consideramos en todo momento un espacio de probabilidad implícito $(\Omega, \mathscr{F}, \mathbb{P})$.

[2] Ω es cualquier conjunto y $\mathscr{F}$ un $\sigma-$álgebra.

[3] Véase recta real extendida en Folland (1984).

Definición A.0.4. Probabilidad condicional. Sea $\mathbb{P}$ una probabilidad. Entonces, la probabilidad de A dado B, con $A, B \in \mathscr{F}$ y $\mathbb{P}(A), \mathbb{P}(B) > 0$ es

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

Teorema 45. Regla de Bayes. Se cumple que

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

Demostración. Por definición,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}$$
$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)}.$$

Entonces,

$$\mathbb{P}(A|B)\mathbb{P}(B) = \mathbb{P}(B|A)\mathbb{P}(A) = \mathbb{P}(A \cap B).$$

Así,

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B)}.$$

□

Si tenemos $A_1, ..., A_n \in \mathscr{F}$, y $\Omega = \biguplus_{k=1}^{n} A_k$, entonces

$$\mathbb{P}(A) = \sum_{k=1}^{n} \mathbb{P}(A \cap A_k).$$

Por otro lado, si $\mathbb{P}(A_k) > 0$

$$\mathbb{P}(A) = \sum_{k=1}^{n} \mathbb{P}(A|A_k)\mathbb{P}(A_k).$$

Definición A.0.5. Sean A_λ con $\lambda \in \Lambda$ eventos. Decimos que los eventos son independientes cuando

$$\mathbb{P}\left(\bigcap_{k=1}^{n} A_{\lambda_k}\right) = \prod_{k=1}^{n} \mathbb{P}(A_k),\ \forall\ n \in \mathbb{Z},\ \forall\ \lambda_1, \cdots, \lambda_k \in \Lambda.$$

Definición A.0.6. Sean $A_1, A_2, \cdots$ una sucesión de conjuntos. Definimos

$$\lim\sup A_n = \bigcap_{n=1}^{\infty} \bigcup_{k=n}^{\infty} A_k$$
$$\lim\inf A_n = \bigcup_{n=1}^{\infty} \bigcap_{k=n}^{\infty} A_k.$$

Note que si $\omega \in \lim\sup A_n$, entonces $\omega \in A_k$ para infinitos A_k. Por otro lado, si $\omega \in \lim\inf A_n$, $\omega \notin A_k$ para un número finito de k's.

Teorema 46. Se cumple que

1. $\lim\inf A_n \subset \lim\sup A_n$
2. $B_n = \bigcup_{k\geq n} A_k$, entonces $B_n \downarrow \lim\sup A_n$
3. $C_n = \bigcap_{k\geq n} A_k$, entonces $C_n \uparrow \lim\inf A_n$

Teorema 47. Primer Lema de Borel Cantelli. Sean $A_1, A_2, \cdots, A_n, ...$ una sucesión de eventos tales que $\sum_n \mathbb{P}(A_n) < \infty$. Entonces,

$$\mathbb{P}(\lim\sup A_n) = 0.$$

Teorema 48. Segundo Lema de Borel Cantelli. Sean $A_1, A_2, \cdots, A_n, ...$ una sucesión de eventos independientes tales que $\sum_n \mathbb{P}(A_n) = \infty$. Entonces,

$$\mathbb{P}(\lim\sup A_n) = 1.$$

Definición A.0.7. Un vector aleatorio es una función $X : \Omega \to \mathbb{R}^k$. Cuando $k = 1$, diremos que se trata de una variable aleatoria (v.a.).

Cuando trabajamos con $\mathbb{R}^k$, consideramos, salvo que se diga lo contrario, el $\sigma-$álgebra de Borel $\mathscr{B}_{\mathbb{R}^k}$ Folland (1984).

Definición A.0.8. La ley o distribución de un vector aleatorio X es una medida de probabilidad en $(\mathbb{R}^k, \mathscr{B}_{\mathbb{R}^k})$:

$$\begin{aligned} \mathbb{P}_X \ &: \mathscr{B}_{\mathbb{R}^k} \to \mathbb{R} \\ &A \to \mathbb{P}(X^{-1}(A)). \end{aligned}$$

Ejemplo 84. Sea $X \sim B(n, p)$. Entonces,

$$\mathbb{P}_X\{k\} = \mathbb{P}(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \ k = 0, ..., n.$$

De hecho, dado $A \in \mathscr{B}_{\mathbb{R}}$,

$$\begin{aligned} \mathbb{P}_X : \mathscr{B}_{\mathbb{R}} &\to \mathbb{R} \\ A &\to \sum_{k \in A} \binom{n}{k} p^k (1-p)^{n-k}. \end{aligned}$$

Ejemplo 85. Sea $X \sim N(0, 1)$

$$\mathbb{P}_X([a, b]) = \mathbb{P}(a \leq X \leq b) = \int_a^b \varphi(x) dx, \ \forall \ a < b,$$

donde

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}.$$

Para poder integrar sobre todo los Borelianos, en relación al Ejemplo 85.

A continuación, nos enfocamos en los vectores aleatorios discretos. Esto es, aquellos que poseen la siguiente propiedad:

$$\exists \ A \subset \mathbb{R}^k \text{ enumerable} : \ \mathbb{P}(X \in A) = 1.$$

Definición A.0.9. La función de probabilidad de un vector aleatorio discreto es

$$\begin{aligned}\mathbb{P}_X : A &\to \mathbb{R} \\ x &\to \mathbb{P}\{X = x\}.\end{aligned}$$

Definición A.0.10. Sea A tal que $\mathbb{P}\{X \in A\} = 1$, con A enunerable,. La esperanza de una variable aleatoria discreta es

$$\mathbb{E}[X] = \sum_{k=1}^{n} x_k \mathbb{P}_X(x_k) \quad \text{para el caso finito}$$

$$\mathbb{E}[X] = \sum_{k=1}^{\infty} x_k \mathbb{P}_X(x_k) \quad \text{para el caso enumerable no finito.}$$

Teorema 49. Sean X, Y dos variables aleatorias discretas y $\alpha \in \mathbb{R}$:

1. αX es discreta y $\mathbb{E}[\alpha X] = \alpha \mathbb{E}[X]$
2. $\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y]$
3. Si $X \leq Y$ $(X(\omega) \leq Y(\omega),\ \forall\, \omega \in \Omega)$, entonces $\mathbb{E}[X] \leq \mathbb{E}[Y]$
4. Si $g : \mathbb{R} \to \mathbb{R}$ es Borel medible, entonces $\mathbb{E}[g(X)] = \sum_{k=1}^{n} g(x_k)\mathbb{P}_X(x_k)$ (en el caso finito), $\mathbb{E}[X] = \sum_{k=1}^{\infty} g(x_k)\mathbb{P}_X(x_k)$ (en el caso enumerable no finito).

Definición A.0.11. Sea X una v.a. tal que $\mathbb{E}[X] < \infty$,

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

A $\sqrt{\mathrm{Var}[X]}$ se le conoce como desviación estándar y se le denota σ_X.

Teorema 50. Sea X una variable aleatoria discreta con media finita. Entonces:

1. Si $c \in \mathbb{R}$, $\mathrm{Var}(cX) = c^2\mathrm{Var}(X)$.
2. Si $c \in \mathbb{R}$, $\mathrm{Var}(X + c) = \mathrm{Var}(X)$.
3. $\mathrm{Var}(X) = \mathbb{E}[X^2] - (\mathbb{E}[X])^2$.

Abordamos a continuación la cuestión de la independencia de vectores aleatorios, siempre en el caso discreto.

Sean $X_\lambda : \Omega \to \mathbb{R}^{n_\lambda}$, $\lambda \in \Lambda$ una familia de vectores aleatorios. Decimos que son independientes cuando

$$\mathbb{P}\{X_{\lambda_1} = x_1, \cdots, X_{\lambda_k} = x_k\} = \prod_{i=1}^{k} \mathbb{P}\{X_{\lambda_i} = x_i\},$$

$\forall\ k \geq 2,\ \forall\ \lambda_1, ..., \lambda_k \in \Lambda,\ \forall\ x_j \in \mathbb{R}^{n_{\lambda_j}}$.

A continuación, abordamos el caso de vectores aleatorios que ya no son necesariamente discretos.

Definición A.0.12. Sea $X : \Omega \to \mathbb{R}^k$ un vector aleatorio continuo. Decimos que X es un vector aleatorio absolutamente continuo cuando existe una función $f : \mathbb{R}^k \to \mathbb{R}_+$ tal que

$$\mathbb{P}\{X \leq a\} = \int_{-\infty}^{a_1} dx_1 \int_{-\infty}^{a_2} \cdots \int_{-\infty}^{a_k} dx_k\ f(x_1, \cdots, x_k),\ \forall\ a \in \mathbb{R}^k.$$

La función f es conocida como la densidad de X.

Note que[4]

$$\mathbb{P}_X(A) = \underbrace{\int_A f(x)dx}_{\text{integral de Lebesgue}},\ \forall\ A \in \mathscr{B}_{\mathbb{R}^k}.$$

[4]Más adelante definimos con rigor la integral de Lebesgue.

Por otro lado, denotaremos

$$f(x) = \frac{d\mathbb{P}_X}{dx}.$$

Cuando X es una variable aleatoria absolutamente continua con función de densidad f y $g : \mathbb{R}^k \to \mathbb{R}$ Borel medible,

$$\mathbb{E}[X] = \int_{\mathbb{R}} x f(x) dx$$
$$\mathbb{E}[g(X)] = \int_{\mathbb{R}} g(x) f(x) dx$$

Definición A.0.13. Sea $X : \Omega \to \mathbb{R}^k$ un vector aleatorio. La función de distribución de X es la función $F : \mathbb{R}^k \to \mathbb{R}$ dada por

$$F_X(a) = \mathbb{P}\{X \leq a\} = \mathbb{P}_X \left\{ \prod_{i=1}^{k} (-\infty, a_i] \right\}.$$

Teorema 51. Sea X una v.a.

1. $0 \leq F_X(t) \leq 1$, $\forall\ t \in \mathbb{R}$.
2. F_X es una función no decreciente. En particular, existen los límites por la izquierda.
3. F_X es continua por la derecha.
4. $\lim_{t \to -\infty} F_X(t) = 0$ y $\lim_{t \to \infty} F_X(t) = 1$.

Definición A.0.14. El $\sigma-$álgebra generado por un vector aleatorio $X : \Omega \to \mathbb{R}^k$ es

$$\sigma(X) = \{X^{-1}(B) :\ B \in \mathscr{B}_{\mathbb{R}^k}\}.$$

Si $X_\lambda : \Omega \to \mathbb{R}^{n_\lambda}$, $\lambda \in \Lambda$ es una familia de vectores aleatorios, el $\sigma-$álgebra generado por la familia es

$$\sigma(\{X_\lambda : \lambda \in \Lambda\}) = \sigma(\{X_\lambda^{-1}(B) : \ \lambda \in \Lambda, B \in \mathscr{B}_{\mathbb{R}^k}\}).$$

Definición A.0.15. Diremos que los vectores aleatorios $\{X_\lambda\}_{\lambda\in\Lambda}$ son independientes si $\sigma(X_\lambda)$ son independientes.[5]

Para poder definir correctamente la noción de esperanza cuando la variable aleatoria ya no es discreta ni absolutamente continua, requerimos una definición formal de la integral de Lebesgue. Para esto, recordemos algunos aspectos claves de la teoría de la integración.

Sea $(\Omega, \mathscr{F}, \mu)$ un espacio de medida.

$$\mu : \mathscr{F} \to [0, \infty]$$
$$\mu(\emptyset) = 0$$
$$A_1, A_2, ..., \in \mathscr{F}$$
$$A_i \cap A_j = \emptyset, \ \forall \ i \neq j : \mu\left(\bigcup_{k\in\mathbb{N}} A_k\right) = \sum_{k=1}^{\infty} \mu(A_k).$$

Definición A.0.16. Una función $f : \Omega \to \mathbb{R}$ es medible si $f^{-1}(B) \in \mathscr{F}$ para todo $A \in \mathscr{B}_\mathbb{R}$.

Definición A.0.17. Decimos que una función medible f es simple cuando f toma una cantidad finita de valores. Es decir,

$$f(\Omega) = \{f(\omega) : \ \omega \in \Omega\}$$

es finito .

[5]Para la noción de independencia de $\sigma-$álgebras, consultar Gall (2022).

Definición A.0.18. Decimos que f es una función positiva cuando $f(\omega) \geq 0,\ \forall\ \omega \in \Omega$.

Sea $f : \Omega \to \mathbb{R}$ medible, simple y positiva. Entonces,

$$f(\Omega) = \{x_1, ..., x_n\},$$

y, si $A_k = f^{-1}(\{x_k\})$:

$$\Omega = \sum_{k=1}^{n} A_k = \begin{cases} A_1, A_2, ... & \text{disjuntos 2 a 2} \\ \biguplus_{k=1}^{n} A_k = \Omega \end{cases}$$

$$f(\omega) = \begin{cases} x_1, & \text{si } \omega \in A_1 \\ x_2, & \text{si } \omega \in A_2 \\ & \vdots \\ x_n, & \text{si } \omega \in A_n. \end{cases}$$

Así pues,

$$f = \sum_{k=1}^{n} x_k \mathbf{1}_{A_k}.$$

Definimos[6]

$$\int_\Omega f d\mu = \sum_{k=1}^{n} x_k \mu(A_k).$$

[6]

$$\frac{1}{b-a}\int_a^b f(x)dx \simeq \underbrace{\frac{1}{b-a}\sum_{k=1}^{n} f(x_k^*)(t_k - t_{k-1})}_{\text{promedio ponderado}}$$

$$\frac{1}{b-a}\sum_{k=1}^{n}(t_k - t_{k-1}) = 1 = \sum_{k=1}^{n}\lambda_k.$$

Por convención, consideramos que

$$a \cdot \infty = \infty, \ a > 0$$

$$a \cdot \infty = 0, \ a = 0.$$

Si $y_1, ..., y_m \geq 0$ y $B_1, ..., B_m \in \mathscr{F}$ y

$$g = \sum_{j=1}^{m} y_j \mathbf{1}_{B_j}.$$

Entonces, g es medible (combinación lineal de medibles), simple (toma a lo mucho 2^m valores), positiva ($y_i \geq 0$) y

$$\int_\Omega g d\mu = \sum_{j=1}^{m} y_j \mathbf{1}_{B_j}.$$

Sea $f : \Omega \to \overline{\mathbb{R}}$ una función medible positiva.

$$\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty, \infty\}$$

$$\mathscr{B}(\overline{\mathbb{R}}) = \{A = B \cup C : \ B \in \mathscr{B}(\mathbb{R}) \ \wedge \ C \subset \{-\infty, \infty\}\}.$$

Definición A.0.19. Definimos

$$\int_\Omega f d\mu = \sup \left\{ \int_\Omega g d\mu : \ g \text{ es medible simple positiva y } g \leq f \right\}.$$

Definición A.0.20. Dado $A \subset \mathbb{R}$

$$\sup \ A = \begin{cases} \sup \ A & \text{si } A \neq \emptyset \text{ y } A \text{ acotado superiormente} \\ \infty & \text{si } A \neq \emptyset \text{ y } A \text{ no es acotado superiormente} \\ -\infty, & \text{si } \ A = \emptyset. \end{cases}$$

Teorema 52. Sea $f : \Omega \to \overline{\mathbb{R}}$ una función medible positiva. Para cada $n \geq 1$, sea

$$f_n(\omega) = \sum_{k=0}^{n2^n - 1} \frac{k}{2^n} \mathbf{1}_{f^{-1}\left(\left[\frac{k}{2^n}, \frac{k+1}{2^n}\right)\right)} + n \mathbf{1}_{f^{-1}(\infty)}.$$

Entonces, $f_n \uparrow f$. Esto es

$$\begin{cases} f_1(\omega) \leq f_2(\omega) \leq \cdots & \forall\ \omega \in \Omega \\ \lim_{n\to\infty} f_n(\omega) = f(\omega) & ,\ \forall\ \omega \in \Omega. \end{cases}$$

Teorema 53. Sean $f, g : \Omega \to \overline{\mathbb{R}}$ funciones medibles positivas.

1. Si $\alpha \geq 0$ entonces αf es medible positiva y $\int \alpha f = \alpha \int f$
2. $f + g$ es medible positiva y $\int f + g = \int f + \int g$. Lo mismo vale para simples.
3. Si $f \leq g$, entonces $\int f \leq \int g$.

Teorema 54. Convergencia monótona. Sean $f, f_1, f_2, ...$ funciones medibles positivas. Si $f_n \uparrow f$ entonces $\int f_n \to \int f$.

Teorema 55. Fatou. Sea $(\Omega, \mathscr{F}, \mu)$ un espacio de medida y $f_1, f_2, ... : \Omega \to \overline{\mathbb{R}}$ funciones medibles positivas

$$\int \lim\text{ínf}\, f_n \leq \lim\text{ínf} \int f_n.$$

Note que si definimos $x_n = \int_\Omega f_n d\mu \in [0, \infty]$,

$$\lim\text{ínf}\, x_n = \sup_{n\geq 1} \text{ínf}\{x_k : k \geq n\} = \lim_{n\to\infty} \text{ínf}\{x_k : k \geq n\}.$$

Así, el lím ínf de las integrales está bien definido. Por otro lado,

$$h(\omega) = \lim\text{ínf}\, f_n(\omega) \in [0, \infty]$$

es una función medible bien definida.

Teorema 56. Fijamos $(\Omega, \mathscr{F}, \mu)$, espacio de medida. Sea $f : \Omega \to \overline{\mathbb{R}}$ una función medible positiva. Entonces

a) La función $\nu : \mathscr{F} \to [0, \infty]$ tal que

$$\nu(A) = \int_\Omega f \cdot \mathbf{1}_A d\mu$$

es una medida en $(\Omega, \mathscr{F})$.

b) Si $g : \Omega \to \overline{\mathbb{R}}$ es una función medible positiva, entonces

$$\int_\Omega g d\nu = \int_\Omega g \cdot f d\mu.$$

Teorema 57. Cambio de Variable. Sean $(\Omega_1, \mathscr{F}_1)$ y $(\Omega_2, \mathscr{F}_2)$ espacios medibles y $g : \Omega_2 \to \overline{\mathbb{R}}$ una función medible positiva.

a) Si μ es una medida en $(\Omega_1, \mathscr{F}_1)$ entonces la función $\nu = \mu f^{-1} : \mathscr{F}_2 \to [0, \infty]$, tal que $A \to \mu(f^{-1}(A))$ es una medida en $(\Omega_2, \mathscr{F}_2)$.

b) $\int_{\Omega_2} g d\nu = \int_{\Omega_1} g \circ f d\mu$.

Sea $f : \Omega \to \overline{\mathbb{R}}$ una función medible. Definimos la parte positiva de f y la parte negativa de f por

$$f^+ : \Omega \to \overline{\mathbb{R}}, \quad f^- : \Omega \to \overline{\mathbb{R}}$$

$$f^+(\omega) = \begin{cases} f(\omega), & \text{si } f(\omega) \geq 0 \\ 0, & \text{si } f(\omega) < 0 \end{cases}$$

y

$$f^-(\omega) = \begin{cases} 0, & \text{si } f(\omega) > 0 \\ f(\omega), & \text{si } f(\omega) \leq 0 \end{cases}$$

Cuando al menos una de las integrales $\int f^+$, $\int f^-$ es finita, decimos que la integral de f con respecto a μ está bien definida y su valor es

$$\int_\Omega f d\mu = \int_\Omega f^+ d\mu - \int_\Omega f^- d\mu.$$

Luego, cuando ambas $\int f^+$, $\int f^-$ son finitas, decimos que f es $\mu-$integrable.

Teorema 58. Sea $f : \Omega \to \overline{\mathbb{R}}$ una función medible. Entonces,

$$f \text{ integrable} \Leftrightarrow \int_\Omega |f| d\mu < \infty.$$

Teorema 59. Sean f y g funciones medibles integrables.

1. Si $c \in \mathbb{R}$ entonces cf es integrable y
$$\int cf = c \int f.$$
2. $f + g$ es integrable y $\int f + g = \int f + \int g$
3. Si $f \leq g$ entonces $\int f \leq \int g$. En particular, $|\int f| \leq \int |f|$.
4. Si $A_1, ..., A_n \in \mathscr{F}$ y $A_1, ..., A_n$ son disjuntos dos a dos,
$$\int_{\bigcup_{k=1}^n A_k} f d\mu = \sum_{k=1}^n \int_{A_k} f d\mu.$$
5. Lo anterior vale para una colección infinita numerable
$$\int_{\bigcup_{k=1}^\infty A_k} f d\mu = \sum_{k=1}^\infty \int_{A_k} f d\mu.$$
6. Si $A_1, A_2, ... \in \mathscr{F}$ y $A_n \downarrow A$ o bien $A_n \uparrow A$, entonces
$$\lim_{n\to\infty} \int_{A_n} f d\mu = \int_A f d\mu.$$

Teorema 60. Sean $f, g, f_1, f_2, ... : \Omega \to \overline{\mathbb{R}}$ funciones medibles tales que

1. $\lim_{n\to\infty} f_n = f$.
2. $|f_n| \leq g$, para todo $n \geq 1$
3. g es integrable.

Entonces $\lim_{n\to\infty} \int f_n = \int f$.

Los resultados enunciados previamente valen si se agrega «casi seguramente». Esto es, que la propiedad vale salvo eventualmente en un conjunto $\mathcal{N}$ tal que $\mu(\mathcal{N}) = 0$.

Los siguientes dos teoremas requieren de ciertos preliminares que no vamos a presentar (teorema de Carathéodory, medida producto, sigma álgebra producto). El lector puede consultar Folland (1984).

Teorema 61. Tonelli. Sean $(\Omega_1, \mathscr{F}_1, \mu_1)$, $(\Omega_2, \mathscr{F}_2, \mu_2)$ espacios de medida $\sigma-$finitos[7] y $f : \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$ una función $\mathscr{F}_1 \otimes \mathscr{F}_2-$medible positiva. Entonces:

1. Para cada $x \in \Omega_1$, la función $f_x : \Omega_2 \to \overline{\mathbb{R}}$, $y \to f(x, y)$ es $\mathscr{F}_2$ medible y positiva.
2. La función $\varphi : \Omega_1 \to \overline{\mathbb{R}}$ tal que $x \to \int_{\Omega_2} f(x, y) d\mu_2(y)$ es $\mathscr{F}_1$ medible.

[7] μ_i es medida sobre Ω_i y existen $\{A_k^i\}_{k\geq 1}$ tales que $\mu(A_k^i) < \infty$ y $\bigcup_{k=1}^{\infty} A_k^i = \Omega_i$.

3. Para cada $y \in \Omega_2$, la función $f^y : \Omega_1 \to \overline{\mathbb{R}}$ es $\mathscr{F}_1-$medible positiva.

4. La función $\psi : \Omega_2 \to \overline{\mathbb{R}}$, $y \to \int_{\Omega_1} f^y(x) d\mu_1(x)$ es $\mathscr{F}_2$ medible.

5. Y

$$\int_{\Omega_1 \times \Omega_2} f(x,y) d(\mu_1 \times \mu_2)(x,y) = \int_{\Omega_1} \varphi(x) d\mu_1(x) = \int_{\Omega_2} \psi(y) d\mu_2(y).$$

Teorema 62. Fubini. Sean $(\Omega_1, \mathscr{F}_1, \mu_1)$, $(\Omega_2, \mathscr{F}_2, \mu_2)$ espacios de medida $\sigma-$finitos y $f : \Omega_1 \times \Omega_2 \to \overline{\mathbb{R}}$ una función $\mathscr{F}_1 \otimes \mathscr{F}_2$ medible integrable. Entonces,

1. Para cada $x \in \Omega_1$ la función $f_x : \Omega \to \overline{\mathbb{R}}$, $y \to f(x,y)$ es $\mathscr{F}_2$ medible.

2. f_x es μ_2 integrable μ_1-c.s.

3. La función $\varphi : \Omega_1 \to \overline{\mathbb{R}}$ tal que $x \to \int_{\Omega_2} f_x(y) d\mu_2(y)$ definida μ_1-c.s. es $\mathscr{F}_1$ medible y μ_1 integrable.

4. Para cada $y \in \Omega_2$, la función $f^y : \Omega_1 \to \overline{\mathbb{R}}$ es $\mathscr{F}_1-$medible.

5. f^y es μ_1-integrable μ_2-c.s.

6. La función $\psi : \Omega_2 \to \overline{\mathbb{R}}$ tal que $y \to \int_{\Omega_1} f(x,y) d\mu_1(x)$ definida μ_2 c.s. es $\mathscr{F}_2$ medible y μ_2 integrable.

7. $\int_{\Omega_1 \times \Omega_2} f(x,y) d(\mu_1 \times \mu_2)(x,y) = \int_{\Omega_1} \varphi(x) d\mu_1(x) = \int_{\Omega_2} \psi(y) d\mu_2(y)$.

Antes de terminar con el breve repaso de teoría de la medida, enunciamos el teorema de Radon-Nikodym. Enseguida, pasamos a la definición de esperanza condicional en el caso general.

Sean μ y ν medidas en el espacio medible $(\Omega, \mathscr{F})$. Decimos que ν es absolutamente continua con respecto a μ cuando

$$\nu(A) = 0, \ \forall \ A \in \mathscr{F}, \ \text{con } \mu(A) = 0.$$

Denotamos esta situación por $\nu \ll \mu$.

Teorema 63. Radon-Nikodym. Sean μ y ν medidas $\sigma-$finitas en el espacio medible $(\Omega, \mathscr{F})$ tales que $\nu < \mu$.

1. Existe $f : \Omega \to \overline{\mathbb{R}}$ medible positiva tal que $d\nu = f d\mu$: $f = \frac{d\nu}{d\mu}$.
2. f es μ-única (si g es otra función, $g = f$ c.s.)

Regresamos al mundo de la probabilidad. Sea $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad y $X : \Omega \to \overline{\mathbb{R}}$ una variable aleatoria. La media o esperanza de X es definida por

$$\mathbb{E}[X] = \int_{\Omega} X d\mathbb{P}.$$

Si $X : \Omega \to \mathbb{R}^k$ es un vector aleatorio y $g : \mathbb{R}^k \to \mathbb{R}$ es una función Borel medible, entonces

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^n} g(x) \underbrace{d\mathbb{P}_X(x)}_{\text{medida que induce } X}.$$

Definición A.0.21. Sean $(\Omega_1, \mathscr{F}_1, \mathbb{P}^1)$, $(\Omega, \mathscr{F}_2, \mathbb{P}^2)$ espacios de probabilidad y $X : \Omega_1 \to \mathbb{R}^k$ y $Y : \Omega_2 \to \mathbb{R}^k$ vectores aleatorios. Diremos que X e Y están idénticamente distribuidos si $\mathbb{P}^1_X = \mathbb{P}^2_Y$. Denotamos esto por $X \sim Y$.

Teorema 64. Si $X \sim Y$, entonces:

1. Si $g : \mathbb{R}^k \to \mathbb{R}$ es función Borel medible, entonces

$$g(X) = g(Y).$$

2. Si $k = 1$, y las v.a son integrables, $\mathbb{E}[X] = \mathbb{E}[Y]$.

Teorema 65. Sean $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad y $X : \Omega \to \overline{\mathbb{R}}$ una v.a. positiva. Entonces

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)dx = \int_0^\infty \mathbb{P}(X \geq x)dx.$$

Demostración. Tenemos que

$$\int_0^\infty \mathbb{P}(X > x)dx = \int_0^\infty \left[\int_\Omega \mathbf{1}_{X>x}(\omega)d\mathbb{P}(\omega)\right] dx.$$

Ahora bien

$$\mathbf{1}_{X>x}(\omega) = \begin{cases} 1, & \text{si } \underbrace{X(\omega) > x}_{x\in[0,X(\omega)} \\ 0, & \text{caso contrario.} \end{cases}$$

Luego, por Tonelli

$$\begin{aligned} \int_0^\infty \left[\int_\Omega \mathbf{1}_{X>x}(\omega)d\mathbb{P}(\omega)\right] dx &= \int_\Omega \left[\int_0^\infty \mathbf{1}_{[0,X(\omega))}(x)dx\right] d\mathbb{P}(\omega) \\ &= \int_\Omega X(\omega)d\mathbb{P}(\omega) \\ &= \mathbb{E}[X]. \end{aligned}$$

□

Teorema 66. Sea $X : \Omega \to \mathbb{R}$ una variable aleatoria positiva. Entonces

$$\sum_{n=1}^\infty \mathbb{P}(X \geq n) \leq \mathbb{E}[X] \leq 1 + \sum_{n=1}^\infty \mathbb{P}(X \geq n).$$

Demostración. Para $n \geq 1$, sea $A_n = \{n-1 \leq X < n\}$. Entonces, definiendo $B_n = \{X \geq n\}$[8]

$$\begin{aligned}
\mathbb{E}[X] &= \int_\Omega X d\mathbb{P} \\
&= \sum_{n=1}^{\infty} \int_{A_n} X d\mathbb{P} \\
&\leq \sum_{n=1}^{\infty} \int_{A_n} n d\mathbb{P} \\
&= \sum_{n=1}^{\infty} \mathbb{P}(A_n) \\
&= \sum_{n=1}^{\infty} n[\mathbb{P}(B_{n-1}) - \mathbb{P}(B_n)] \\
&= \lim_{n\to\infty} \sum_{k=1}^{n+1} k[\mathbb{P}(B_{k-1}) - \mathbb{P}(B_k)] \\
&= \lim_{n\to\infty} \left[\mathbb{P}(B_0) + \sum_{k=1}^{n} \mathbb{P}(B_k) - (n+1)\mathbb{P}(B_n)\right] \\
&\leq 1 + \lim_{n\to\infty} \sum_{k=1}^{n} \mathbb{P}(B_k) \\
&= 1 + \sum_{k=1}^{\infty} \mathbb{P}(B_k).
\end{aligned}$$

Respecto a la primera desigualdad,

$$\mathbb{E}[X] \geq \left[\lim_{m\to\infty} \sum_{n=1}^{m} \mathbb{P}(B_n) - m\mathbb{P}(B_m)\right].$$

Si $\mathbb{E}[X] = \infty$ ya está. En caso $\mathbb{E}[X] < \infty$, como $B_m \downarrow \emptyset$, ν tal que

[8] $B_{n-1} = A_n \cup B_n$.

$d\nu = X d\mathbb{P}$ es finita[9]

$$\nu(B_m) = \int_{B_m} X d\mathbb{P} \geq \int_{B_m} m d\mathbb{P} = m\mathbb{P}(B_m) \geq 0$$
$$\implies \lim_m \; m \cdot \mathbb{P}(B_m) = 0.$$

Así,

$$\mathbb{E}[X] \geq \sum_{n=1}^{\infty} \mathbb{P}(B_n).$$

□

Sean $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad y $X_1, X_2, ..., X_n$ variables aleatorias. Sea $X = (X_1, X_2, \cdots, X_n) : \Omega \to \mathbb{R}^n$ vector aleatorio.

- P_X es una medida de probabilidad en $(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n})$.
- $\mathbb{P}_{X_1}, ..., \mathbb{P}_{X_n}$ son medidas de probabilidad en $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$ y $\mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}$ es una medida de probabilidad en

$$(\mathbb{R}^n, \mathscr{B}_{\mathbb{R}} \otimes \cdots \otimes \mathscr{B}_{\mathbb{R}}) = (\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n}).$$

Teorema 67. Son equivalentes

1. $X_1, ..., X_n$ son v.a. independientes
2. $\mathbb{P}_{(X_1,...,X_n)} = \mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n}$.

Teorema 68. Si $X_1, ..., X_n$ son variables aleatorias independientes e integrables, entonces

$$\mathbb{E}\left[\prod_{i=1}^{n} X_i\right] = \prod_{i=1}^{n} \mathbb{E}[X_i].$$

[9] Pues $\nu(\Omega) = \int_\Omega X d\mathbb{P} < \infty$.

Demostración. Consideramos $X : (\Omega, \mathscr{F}) \to (\mathbb{R}^n, \mathscr{B}_{\mathbb{R}^n})$ y $g : \mathbb{R}^n \to \mathbb{R}$, $g(x_1, \cdots, x_n) = x_1 \cdots x_n$. Entonces, $X_1 \cdots X_n = g \circ X$. Ahora bien, $\tilde{g}(x_1, ..., x_n) = |x_1 \cdots x_n|$.

$$\begin{aligned}
\mathbb{E}[|X_1 \cdots X_n|] &= \int_\Omega \tilde{g}(X) d\mathbb{P} \\
&= \int_{\mathbb{R}^n} \tilde{g}(x) d\mathbb{P}_X(x) \\
&= \int_{\mathbb{R}^n} |x_1 \cdots x_n| d(\mathbb{P}_{X_1} \times \cdots \times \mathbb{P}_{X_n})(x) \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \cdots \left[\int_{\mathbb{R}} |x_1| \cdots |x_n| d\mathbb{P}_{X_n}(x_n) \right] \cdots d\mathbb{P}_{X_2}(x_2) \right] d\mathbb{P}_{X_1}(x_1) \\
&= \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \cdots |x_1| \cdots |x_{n-1}| \left[\int_{\mathbb{R}} |x_n| d\mathbb{P}_{X_n}(x_n) \right] \cdots d\mathbb{P}_{X_2}(x_2) \right] d\mathbb{P}_{X_1}(x_1) \\
&= \mathbb{E}[X_n] \cdots \mathbb{E}[X_{n-1}] \cdots \mathbb{E}[|X_1|] < \infty.
\end{aligned}$$

Así, $X_1 \cdots X_n$ es integrable. Luego,

$$\begin{aligned}
\mathbb{E}[X_1 \cdots X_n] = \int_\Omega h(X) d\mathbb{P} &= \int_{\mathbb{R}^n} g(x) d\mathbb{P}_X(x) \\
&= \cdots = \mathbb{E}[X_n] \cdots \mathbb{E}[X_1].
\end{aligned}$$

□

Sean $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad y $X : \Omega \to \mathbb{R}$ una variable aleatoria integrable. Recordemos que la varianza de X se define como

$$\mathrm{Var}(X) = \mathbb{E}[(X - \mathbb{E}[X])^2].$$

Teorema 69. Sea X v.a.

a) $\mathrm{Var}(X) = \mathbb{E}[X^2] - \mathbb{E}[X]^2$.

b) $\mathrm{Var}(X) \geq 0$. En particular, $\mathbb{E}[X^2] \geq \mathbb{E}[X]^2$.

c) $\text{Var}(X) = 0$, entonces $X = \mathbb{E}[X]$ c.s.

d) $\text{Var}(cX) = c^2\text{Var}(X)$ para todo $c \in \mathbb{R}$ y $\text{Var}(X+c) = \text{Var}(X)$.

e) Si $X_1, ..., X_n$ son v.a. independientes e integrables tales que $\mathbb{E}[X_j^2] < \infty$ para $j = 1, ..., n$. Entonces

$$\text{Var}(X_1 + \cdots + X_n) = \sum_{j=1}^{n} \text{Var}(X_j).$$

A continuación, algunas desigualdades clásicas en teoría de la probabilidad que aparecen con frecuencia a la hora de hacer inferencia estadística desde la perspectiva teórica.

Teorema 70. Desigualdad de Markov. Sea X una v.a. y $t > 0$. Entonces

$$\mathbb{P}(\{|X| \geq t\}) \leq \frac{\mathbb{E}[|X|]}{t}.$$

Demostración. Tenemos

$$\begin{aligned}
\mathbb{E}[|X|] &= \int_{\Omega} |X| d\mathbb{P} \\
&= \int_{\{|X| \geq t\}} |X| d\mathbb{P} + \int_{\{|X| < t\}} |X| d\mathbb{P} \\
&\geq \int_{\{|X| \geq t\}} t d\mathbb{P} = t\mathbb{P}\{|X| \geq t\}.
\end{aligned}$$

□

Teorema 71. Desigualdad de Chebyshev. Sea X una variable aleatoria integrable y $t > 0$, entonces

$$\mathbb{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\text{Var}(X)}{t^2}.$$

Demostración. Escribimos $Y = X - \mathbb{E}[X]$

$$\begin{aligned}
\mathrm{Var}(X) &= \mathbb{E}[Y^2] \\
&= \int_\Omega Y^2 d\mathbb{P} \\
&= \int_{\{|Y|\geq t\}} Y^2 d\mathbb{P} + \int_{\{|Y|<t\}} Y^2 d\mathbb{P} \\
&\geq \int_{\{|Y|\geq t\}} t^2 d\mathbb{P} + \int_{\{|Y|<t\}} Y^2 d\mathbb{P} \\
&\geq \int_{\{|Y|\geq t\}} t^2 d\mathbb{P} \\
&= t^2 \mathbb{P}(\{|Y| \geq t\}).
\end{aligned}$$

□

Teorema 72. Sean Z una variable aleatoria positiva, $\varphi : [0, \infty) \to [0, \infty)$ estrictamente creciente y $t > 0$. Entonces,

$$\mathbb{P}(Z \geq t) \leq \frac{\mathbb{E}[\varphi(Z)]}{\varphi(t)}.$$

Demostración. Tenemos

$$\begin{aligned}
\mathbb{E}[\varphi(Z)] &= \int_{\{Z\geq t\}} \varphi(Z) d\mathbb{P} + \int_{\{Z<t\}} \varphi(Z) d\mathbb{P} \\
&\geq \int_{\{Z\geq t\}} \varphi(t) d\mathbb{P} \\
&= \varphi(t)\mathbb{P}(Z \geq t).
\end{aligned}$$

□

Teorema 73. Desigualdad de Jensen. Sean X una v.a. y $\varphi : \mathbb{R} \to \mathbb{R}$ una función convexa tales que X y $\varphi(X)$ son integrables. Entonces,

$$\mathbb{E}[\varphi(X)] \geq \varphi(\mathbb{E}[X]).$$

A continuación definimos la función generadores de momentos y la función característica asociada a una variable o vector aleatorio.

Definición A.0.22. Sea X una v.a. La función generadora de momentos de X es definida por

$$\psi_X(t) = \mathbb{E}[e^{tX}],\ \forall\ t \in \mathbb{R}.$$

Definición A.0.23. La función característica de X es definida por

$$\varphi_X(t) = \mathbb{E}[e^{itX}] = \mathbb{E}[\cos(tX) + i\sin(tX)].$$

Teorema 74. Si $\mathbb{E}[e^{\delta|X|} < \infty]$ para algún $\delta > 0$ entonces $X^{(k)}$ es integrable y $\mathbb{E}[X^k] = \psi_X^{(k)}(0)$, $\forall\ k \in \mathbb{Z}_+$.

Teorema 75. Si $\mathbb{E}[|X|^r] < \infty$ para algún $r \in \mathbb{Z}_+$ entonces φ_X es de clase C^r y

$$\varphi_X(t) = \mathbb{E}[(iX)^r e^{itX}],\ \forall\, t \in \mathbb{R}.$$

Demostración. Supongamos primero que $r = 1$.

$$\begin{aligned}
\frac{\varphi(t+h) - \varphi(t)}{t} &= \frac{\mathbb{E}[e^{i(t+h)X}] - \mathbb{E}[e^{iXt}]}{h} \\
&= \frac{1}{h}\left[\int_{\mathbb{R}} e^{i(t+h)x} d\mathbb{P}_X(x) - \int_{\mathbb{R}} e^{itx} d\mathbb{P}_X(x)\right] \\
&= \int_{\mathbb{R}} \frac{e^{i(t+h)x} - e^{itx}}{h} d\mathbb{P}_X(x) \\
&= \int_{\mathbb{R}} \frac{e^{itx}(e^{ihx} - 1)}{h} d\mathbb{P}_X(x) \\
&= \int_{\mathbb{R}} ix \frac{e^{itx}(e^{ihx} - 1)}{ixh} d\mathbb{P}_X(x).
\end{aligned}$$

Haciendo $h \to 0$, y usando el TCD

$$\lim_{h\to 0} \frac{\varphi(t+h) - \varphi(t)}{t} = \int_{\mathbb{R}} ixe^{itx} d\mathbb{P}_X(x).$$

En efecto, para $h \in (-1,1)$, $x \neq 0$

$$\begin{aligned}\left|ixe^{itx}\frac{e^{ihx}-1}{ihx}\right| &\leq \left|ixe^{itx}\frac{-1+\cos(hx)+i\sin(hx)}{hx}\right| \\ &\leq |ixe^{itx}| \cdot \left(\underbrace{\left|\frac{1-\cos(hx)}{hx}\right|}_{\leq \theta_0} + \underbrace{\left|\frac{\sin(hx)}{hx}\right|}_{\leq 1}\right) \\ &\leq \theta_1|x|.\end{aligned}$$

De este modo, como

$$\int_{\mathbb{R}} \theta_1|x|dP_X(x) = \theta_1\mathbb{E}[|X|] < \infty,$$

se sigue que

$$\lim_{h\to 0}\frac{\varphi(t+h)-\varphi(t)}{h} = \int_{\mathbb{R}} ixe^{itx}d\mathbb{P}_X(x) = \mathbb{E}[iXe^{itX}].$$

Ahora bien, para probar que es C^1, tenemos

$$\varphi(t+h)-\varphi(t) = \int_{\mathbb{R}} ixe^{itx}[e^{ihx}-1]d\mathbb{P}_X(x).$$

Como

$$|ixe^{itx}[e^{ihx}-1]| \leq 2|x|$$

por el Teorema de la Convergencia Dominada

$$\lim_{h\to 0}\varphi(t+h)-\varphi(t) = \int_{\mathbb{R}} 0d\mathbb{P}_X(x) = 0.$$

Sea ahora $r \in \mathbb{Z}_+$. Primero, se cumple que $\mathbb{E}[|X|^{r+1}] < \infty$ implica $\mathbb{E}[|X|^r] < \infty$. En efecto, en general, si $0 < \alpha < \beta$, $\mathbb{E}[|X|^\beta] < \infty \implies \mathbb{E}[|X|^\alpha] < \infty$. Basta que probemos para $\alpha = 1$

pues siempre podemos tomar $Y = |X|^\alpha$ y considerar Y^γ, donde $\gamma = \beta/\alpha > 1$. Entonces,

$$\begin{aligned}\mathbb{E}[Y] &= \int_{\{Y\geq 1\}} Y + \int_{\{Y<1\}} Y \\ &\leq \int_{\{Y\geq 1\}} Y^\gamma + \int_{\{Y<1\}} Y \\ &\leq \mathbb{E}[Y^\gamma] + 1 < \infty.\end{aligned}$$

Otra forma es usando la desigualdad de Jensen: t^γ para $\gamma > 1$ es convexa. Así,

$$\mathbb{E}[Y_N^\gamma] \geq (\mathbb{E}[Y_N])^\gamma.$$

Ahora bien, haciendo $N \to \infty$, por el Teorema de la Convergencia Monótona

$$\mathbb{E}[Y^\gamma] \geq (\mathbb{E}[Y])^\gamma.$$

Ahora bien,

$$\frac{\varphi^{(r)}(t+h) - \varphi^{(r)}(t)}{h} = \int_{\mathbb{R}} (ix)^r e^{itx} \left[\frac{e^{ihx}-1}{h}\right] d\mathbb{P}_X(x).$$

Luego,

$$\begin{aligned}(ix)^r e^{itx}\left[\frac{e^{ihx}-1}{h}\right] &\to (ix)^{r+1}e^{itx} \\ \left|(ix)^r e^{itx}\left[\frac{e^{ihx}-1}{h}\right]\right| &\leq \theta_1 |x|^{r+1} \\ \int_{\mathbb{R}} \theta_1 |x|^{r+1} d\mathbb{P}_X(x) &= \theta_1 \mathbb{E}[|X|^{r+1}] < \infty.\end{aligned}$$

Por el Teorema de la Convergencia Dominada

$$\begin{aligned}\lim_{h\to 0} \frac{\varphi^{(r)}(t+h) - \varphi^{(r)}(t)}{h} &= \int_{\mathbb{R}} (ix)^{r+1} e^{itx} d\mathbb{P}_X(x) \\ &= \mathbb{E}[(iX)^{r+1} e^{itX}].\end{aligned}$$

La continuidad de $\varphi^{(r+1)}$ se prueba de forma análoga. □

Note que

1. $|\varphi(t)| \leq \varphi(0) = 1, \forall\ t \in \mathbb{R}$
2. φ es uniformemente continua.
3. Si $a, b \in \mathbb{R}$, $\varphi_{aX+b}(t) = \varphi(at)e^{itb}$.
4. $\varphi(-t) = \overline{\varphi}(t)$.
5. φ toma valores reales si y solo si X es una v.a. simétrica. Esto es, $\mathbb{P}_X(B) = \mathbb{P}_X(-B)$.

Demostración. Inciso por inciso:

1. $|\varphi(t)| \leq |\mathbb{E}[e^{itX}]| \leq \mathbb{E}[|e^{itX}|] = 1 = \varphi(0), \forall\ t \in \mathbb{R}$.
2. $|\varphi(t+h) - \varphi(t)| = \left|\int_{\mathbb{R}} e^{itx}(e^{ih}-1)d\mathbb{P}_X(x)\right| \leq \int_{\mathbb{R}} |e^{ihx} - 1| d\mathbb{P}_X(x)$.
3. $\varphi_{aX+b}(t) = \mathbb{E}[e^{it(aX+b)}] = e^{itb}\mathbb{E}[e^{itaX}] = e^{itb}\varphi(ta)$.
4. $\varphi(-t) = \mathbb{E}[e^{-itX}] = \mathbb{E}[\overline{e^{itX}}] = \overline{\mathbb{E}[e^{itX}]} = \overline{\varphi(t)}$.
5. $\varphi_{-X}(t) = \mathbb{E}[e^{it(-X)}] = \varphi_X(-t) = \overline{\varphi}_X(t) = \varphi_X(t), \forall\ t \in \mathbb{R}$. Recíprocamente,

$$\varphi_X(t) = \varphi_{-X}(t) = \overline{\varphi_X(t)} \implies \varphi_X(t) \in \mathbb{R}.$$

□

Teorema 76. Sea φ la función característica del vector aleatorio $X : \Omega \to \mathbb{R}^n$. Entonces,

1. $|\varphi(t)| \leq \varphi(0) = 1$.

2. φ es uniformemente continua.

3. Si $a \in \mathbb{R}$ y $b \in \mathbb{R}^n$, entonces

$$\varphi_{aX+b}(t) = \varphi(at)e^{it\cdot b},\ \forall\ t \in \mathbb{R}^n.$$

4. Las v.a. $X_1, ..., X_n$ son independientes si y solo si $\varphi(t) = \prod_{k=1}^{n} \varphi_{X_k}(t_k)$, $\forall\ t \in \mathbb{R}^n$.

5. Si $Y : \tilde{\Omega} \to \mathbb{R}$ es un vector aleatorio y $\varphi_Y = \varphi$, entonces $Y \sim X$.

Las funciones características son de mucha utilidad a la hora de probar cuestiones relacionadas a la convergencia de variables aleatorias. Esto se ilustra a continuación.

Teorema 77. Sean $X_1, X_2, ...$ v.a. independientes e idénticamente distribuidas (iid). Si

$$\mathbb{E}[X_1] = m < \infty$$

entonces

$$\underbrace{\frac{1}{n}\sum_{k=1}^{n} X_k}_{=S_n/n} \to m = Y$$

en distribución.

Demostración.

$$\begin{aligned}
\varphi_{S_n/n}(t) &= \mathbb{E}[e^{it\frac{S_n}{n}}] \\
&= \mathbb{E}\left[e^{\sum_{k=1}^{n}\frac{itX_k}{n}}\right] \\
&= \mathbb{E}\left[\prod_{k=1}^{n} e^{\frac{itX_k}{n}}\right] \\
&= \prod_{k=1}^{n}\mathbb{E}[e^{\frac{itX_k}{n}}] \\
&= \prod_{k=1}^{n}\mathbb{E}[e^{\frac{itX_k}{n}}] \\
&= \left(\mathbb{E}[e^{\frac{itX_1}{n}}]\right)^n \\
&= \varphi_{X_1}\left(\frac{t}{n}\right)^n.
\end{aligned}$$

Usando una aproximación lineal de primer orden,

$$\varphi_{X_1}\left(\frac{t}{n}\right)^n = \left[1 + im\frac{t}{n} + o\left(\left|\frac{t}{n}\right|\right)\right]^n.$$

Haciendo $n \to \infty$,

$$\lim_{n\to\infty} \varphi_{S_n/n}(t) = e^{itm} = \varphi_Y(t).$$

Así, concluimos, pues la convergencia de las funciones características implica la convergencia en distribución. □

Teorema 78. Límite central. Sean $X_1, X_2, ...$ v.a. independientes e idénticamente distribuidas (iid). Si $\mathbb{E}[X_1^2] < \infty$, $\mathbb{E}[X_1] = m$ y $\mathrm{Var}(X_1) = \sigma^2$, entonces

$$\frac{S_n - mn}{\sigma\sqrt{n}} \to N(0,1)$$

en distribución.

Demostración. Sea $Y_k = X_k - \mathbb{E}[X_k] = X_k - m$ y $T_n = \sum_{k=1}^{n} Y_k = S_n - mn$.

$$\begin{aligned}
\varphi_{\frac{T_n}{\sigma\sqrt{n}}} &= \prod_{k=1}^{n} \mathbb{E}\left[e^{\frac{itY_k}{\sigma\sqrt{n}}}\right] \\
&= \left(\mathbb{E}\left[e^{\frac{itY_1}{\sigma\sqrt{n}}}\right]\right)^n \\
&= \left(\varphi_{Y_1}\left(\frac{t}{\sigma\sqrt{n}}\right)\right)^n \\
&= \left[1 + i \cdot 0 \frac{t}{\sigma\sqrt{n}} + \frac{1}{2} i^2 \sigma^2 \left(\frac{t}{\sigma\sqrt{n}}\right)^2 + o\left(\frac{t^2}{\sigma^2 n}\right)\right]^n \\
&= \left[1 + \frac{t^2}{n}\left(-\frac{1}{2} + \frac{b_n}{\sigma^2}\right)\right]^n.
\end{aligned}$$

Como $\frac{b_n t^2}{\sigma^2 n} \to 0$,

$$\varphi_{\frac{T_n}{\sigma\sqrt{n}}} \to e^{-\frac{t^2}{2}} = \varphi_Z(t).$$

□

Seguimos con una breve discusión acerca de los espacios L^p. Esta última nos permitirá abordar el tema de los modos de convergencia de las variables aleatorias.

Sean $(\Omega, \mathscr{F}, \mu)$ un espacio de medida y $p > 0$. Definimos

$$L^p(\Omega, \mathscr{F}, \mu) = \left\{ f : \Omega \to \mathbb{R} : \int_\Omega |f|^p d\mu < \infty \right\}$$

para f medible.

Teorema 79. El espacio $L^p(\Omega, \mathscr{F}, \mu)$ es un espacio vectorial[10].

[10] Véase la definición en Axler (2015).

Demostración. Dado que $L^p(\Omega, \mathscr{F}, \mu)$ es subconjunto de $\{f : \Omega \to \mathbb{R}\}$, solo debemos probar que $f + \lambda g \in L^p$ cuando $f, g \in L^p$ y $\lambda \in \mathbb{R}$.

$$\begin{aligned} |f + \lambda g|^p &\leq (|f| + |\lambda g|)^p \\ &=\leq (2\,\text{máx}\{|f|, |\lambda g|\})^p \\ &= 2^p(\text{máx}\{|f|^p, |\lambda g|^p\}) \\ &\leq 2^p(|f|^p + |\lambda|^p \cdot |g|^p) \\ \int |f + \lambda g|^p &\leq 2^p \left[\int |f|^p + |\lambda|^p \int |g|^p\right] < \infty \\ f + \lambda g &\in L^p. \end{aligned}$$

□

Teorema 80. Desigualdad de Young. Sean $p, q > 1$ tales que $\frac{1}{p} + \frac{1}{q} = 1$. Si $x, y \geq 0$, entonces

$$xy \leq \frac{x^p}{p} + \frac{y^q}{q}.$$

Demostración. Si $x = 0$ o $y = 0$ es directo. Si $x, y > 0$, sean $a = \ln x$ y $b = \ln y$. Entonces,

$$\frac{x^p}{p} + \frac{y^q}{q} = \frac{1}{p}e^{ap} + \frac{1}{q}e^{bq} \geq e^{a+b} = xy.$$

Estamos usando que $x \to e^x$ es convexa. □

Note, en relación a la desigualdad de Young, que también vale que si $p_1, \cdots, p_k > 0$ y $\sum_{i=1}^k \frac{1}{p_i} = 1$ y $x_1, \cdots, x_k \geq 0$, entonces

$$\prod_{i=1}^k x_i \leq \sum_{i=1}^k \frac{x_i^{p_i}}{p_i}.$$

Teorema 81. Desigualdad de Holder. Sean $p, q > 1$ tales que $\frac{1}{p} + \frac{1}{q} = 1$ y $f, g : \Omega \to \mathbb{R}$ funciones medibles. Entonces

$$\int |fg| \leq \left(\int |f|^p \right)^{1/p} \left(\int |g|^q \right)^{1/q}.$$

Además, cuando $f \in L^p$, $g \in L^q$, vale la desigualdad si y solo si existen constantes $a, b \geq 0$ tales que $a^2 + b^2 \neq 0$ y $b|f|^p = a|g|^q$ c.s.

Demostración. Primero, si $\int |f|^p = 0$, $f = 0$ c.s., por lo que $|fg| = 0$ c.s. Análogo si $\int |g|^q = 0$. En caso $\int |f|^p = \infty$ o $\int |g|^q = \infty$, también tenemos la desigualdad.
En caso $\int |f|^p = \int |g|^p = 1$, por la desigualdad de Young

$$1 = \frac{1}{p} \int |f|^p + \frac{1}{q} \int |g|^q \geq \int |fg|.$$

Si $\int |f|^p, \int |g|^q \in (0, \infty)$, sean

$$\tilde{f} = \frac{f}{\left(\int |f|^p \right)^{1/p}}, \ \tilde{g} = \frac{g}{\left(\int |g|^q \right)^{1/q}}$$

Entonces, $\int |\tilde{f}|^p = \int |\tilde{g}|^q = 1$ y concluimos. □

También vale que, dados $p_1, \cdots, p_k > 0$ con $\frac{1}{p_1} + \cdots + \frac{1}{p_k} = 1$ y $f_1, \cdots, f_k : \Omega \to \mathbb{R}$ funciones medibles, entonces

$$\prod_{i=1}^{k} \left(\int |f_i|^p \right)^{\frac{1}{p_i}} \geq \int |f_1 \cdots f_k|.$$

Teorema 82. Desigualdad de Minkowski. Sean $p \geq 1$ y $f, g \in L^p$. Entonces,

$$||f + g||_p \leq ||f||_p + ||g||_p.$$

Demostración. Para $p = 1$ es consecuencia de la desigualdad triangular. Para $p > 1$, sea $q = \frac{p}{p-1}$ y $\theta = \frac{p}{q} = p - 1$. Entonces,

$$\begin{aligned} ||f||_p|| \cdot || \ |(f+g)|^\theta \ ||_q &\geq \int |f| \cdot |f+g|^\theta \\ ||g||_p \cdot || \ |f+g|^\theta \ ||_q &\geq \int |g| \cdot |f+g|^\theta \\ (||f||_p + ||g||_p)||f+g||_p^\theta &\geq \int (|f| + |g|)|f+g|^\theta \\ &\geq \int |f+g|^p. \end{aligned}$$

□

¿Es $|| \cdot ||_p$ una norma en L^p? Tenemos la desigualdad triangular (Minkowski), homogeneidad y $f = 0 \implies ||f||_p = 0$. Sin embargo, no tenemos que $||f||_p = 0$ implique $f = 0$, podemos concluir solo que $f = 0$ .c.s. Es por ello que frecuentemente se considera $L^p/\sim$: el espacio L^p cocientado por la relación de equivalencia «$f = g$ c.s. $\Leftrightarrow f \sim g$».

Teorema 83. El conjunto de funciones simples es denso en L^p.

Demostración. Para la prueba véase Folland (1984). □

Sea $(\Omega, \mathscr{F}, \mu)$ un espacio de medida. Para cada función $f : \Omega \to \mathbb{R}$ definimos el supremo esencial de $|f|$ como

$$||f||_\infty = \text{ínf}\{a \in \mathbb{R} : \mu(\{|f| > a\}) = 0\}.$$

El espacio $L^\infty = L^\infty(\Omega, \mathscr{F}, \mu)$ es definido como el conjunto de funciones medibles $f : \Omega \to \mathbb{R}$ tales que $||f||_\infty < \infty$.

Teorema 84. Si $0 < p < q < r \leq \infty$, entonces

1. $L^q \subset L^p + L^r$.

2. $L^p \cap L^r \subset L^q$ y $||f||_q \leq ||f||_p^{\lambda} ||f||_r^{1-\lambda}$, con $\lambda = \frac{q^{-1}-r^{-1}}{p^{-1}-r^{-1}}$.

Demostración. Para (1), escribimos $f = f\mathbf{1}_{\{|f|\leq 1\}} + f\mathbf{1}_{\{|f|>1\}} = g + h$. Entonces,

$$\begin{aligned}
&\int |g|^r = \int |f|^r \mathbf{1}_{\{|f|\leq 1\}} \leq \int |f|^q \mathbf{1}_{\{|f|\leq 1\}} < \infty \\
&\implies g \in L^r \\
&\int |h|^p = \int |f|^p \mathbf{1}_{\{|f|>1\}} \\
&\qquad \leq \int |f|^q \mathbf{1}_{\{|f|>1\}} < \infty \\
&\implies h \in L^p.
\end{aligned}$$

Luego, para (2), sean $\alpha = \frac{r-p}{r-q}$, $\beta = \frac{r-p}{q-p}$, $a = \frac{p(r-q)}{r-p}$, $b = \frac{r(q-p)}{r-p}$. Por Holder (Teorema 81),

$$|| \, |f|^q \, ||_1 \leq || \, |f|^a \, ||_\alpha \, || \, |f|^b \, ||_\beta.$$

Entonces,

$$\begin{aligned}
\int |f|^q &\leq \left(\int |f|^p \right)^{\frac{r-q}{r-p}} \left(\int |f|^r \right)^{\frac{q-p}{r-p}} \\
&= ||f||_p^{\frac{p(r-q)}{r-p}} \, ||f||_r^{\frac{r(q-p)}{r-p}} \\
||f||_q &\leq ||f||_p^{\frac{p(r-q)}{q(r-p)}} ||f||_r^{\frac{r(q-p)}{q(r-p)}}.
\end{aligned}$$

□

Teorema 85. Si μ es una medida finita y $0 < p < q \leq \infty$, entonces $L^q \subset L^p$ y $||f||_p \leq ||f||_1 q\mu(\Omega)^{\frac{1}{q}-\frac{1}{p}}$, para toda $f : \Omega \to \mathbb{R}$ medible.

Demostración. Sean $\alpha = \frac{q}{p}$, $\beta = \frac{q}{q-p}$. Se tiene que $\frac{1}{\alpha} + \frac{1}{\beta} = 1$. Luego,

$$|| \ |f|^p \ ||_1 \leq || \ |f|^p \ ||_\alpha || \ 1 \ ||_\beta$$
$$\int |f|^p \left(\int |f|^q \right)^{\frac{p}{q}} \left(\int 1 \right)^{\frac{q-p}{q}}$$
$$||f||_p \leq ||f||_q (\mu(\Omega))^{\frac{q-p}{qp}}.$$

□

Continuamos este apéndice con el estudio de los modos de convergencia.

Definición A.0.24. Sean (X_n) una sucesión de variables aleatorias y X una variable aleatoria. Decimos que X_n converge casi seguramente (c.s) a X cuando

$$\mathbb{P}\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\} = 1.$$

Se denota $X_n \to X$ c.s.

Definición A.0.25. Decimos que X_n converge a X en probabilidad cuando

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0, \ \forall \ \varepsilon > 0.$$

Esto es,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1, \ \forall \ \varepsilon > 0.$$

Se denota $X_n \to X$ en $\mathbb{P}$.

Definición A.0.26. Decimos que X_n converge a X en L^p con $p > 0$ cuando

1. $X, X_1, X_2, X_3, ..., \in L^p$

2. $\lim_{n\to\infty} \mathbb{E}(|X_n - X|^p) = 0$.

Denotamos esta situación $X_n \to X$ en L^p.

Un caso interesante de la convergencia en L^p es el caso de la convergencia en media cuadrática véase Rau (2016) o Casella and Berger (2002).

Definición A.0.27. Decimos que X_n converge a X en distribución (o en ley) cuando se cumple una (y por lo tanto todas) de las siguientes condiciones

1. $\lim_{n\to\infty} F_{X_n}(t) = F_X(t)$, para todo t que es punto de continuidad de F_X.

2. $\lim_{n\to\infty} \varphi_{X_n}(t) = \varphi_X(t)$, $\forall\ t \in \mathbb{R}$.

3. $\lim_{n\to\infty} \mathbb{E}[g(X_n)] = \mathbb{E}[g(X)]$ para todo $g : \mathbb{R} \to \mathbb{R}$ Borel medible y acotada. A esto se le denomina convergencia débil.

El siguiente teorema relaciona los modos de convergencia (casi segura, en probabilidad, en L^p y en distribución).

Teorema 86. Sean (X_n) una sucesión de variables aleatorias y X una v.a. Entonces,

1. Si $X_n \to X$ c.s., entonces $X_n \to X$ en probabilidad.

2. Si $X_n \to X$ en L^p, entonces $X_n \to X$ en probabilidad.

3. Si $X_n \to X$ en probabilidad, entonces $X_n \to X$ en distribución.

Teorema 87. Si $X_n \to X$ en probabilidad, entonces existe una subsucesión $(X_{n_k})_{k\in\mathbb{N}}$ de (X_n) tal que $X_{n_k} \to X$ c.s.

Para la prueba de los Teoremas 86 y 87, consultar por ejemplo Gall (2022). Por otro lado, como consecuencia del Teorema 87, dada una sucesión de v.a., podemos concluir que son equivalentes

1. $X_n \to X$ en probabilidad.
2. Toda sub-sucesión de X_n posee una sub-sucesión que converge a X c.s.

A continuación una serie de propiedades que se cumplen en función del modo de convergencia:

- Sean $X, X_1, ..., Y, Y_1, ... : \Omega \to \mathbb{R}$ variables aleatorias tales que $X_n \to X$ c.s. y $Y_n \to Y$ c.s. Entonces:
  - $\forall\ c \in \mathbb{R}$, $cX_n \to cX$ c.s.
  - $X_n + Y_n \to X + Y$ c.s.
  - $X_n Y_n \to XY$ c.s.
  - $\varphi : \mathbb{R} \to \mathbb{R}$ continua, entonces $\varphi \circ X_n \to \varphi \circ X$ c.s.
- Sean $X, X_1, ..., Y, Y_1, ... : \Omega \to \mathbb{R}$ variables aleatorias tales que $X_n \to X$ c.s. y $Y_n \to Y$ en $\mathbb{P}$. Entonces:
  - $\forall\ c \in \mathbb{R}$, $cX_n \to cX$ en $\mathbb{P}$
  - $X_n + Y_n \to X + Y$ en $\mathbb{P}$
  - $X_n Y_n \to XY$ en $\mathbb{P}$
  - $\varphi : \mathbb{R} \to \mathbb{R}$ continua, entonces $\varphi \circ X_n \to \varphi \circ X$ en $\mathbb{P}$.

- Sean $p \geq 1$ y $X, X_1, ..., Y, Y_1, ...$ v.a. en L^p. Entonces:
  - Si $c \in \mathbb{R}$ y $X_n \to X$ en L^p, entonces $cX_n \to cX$ en L^p.
  - Si $X_n \to X$ en L^p y $Y_n \to Y$ en L^p entonces $X_n + Y_n \to X + Y$ en L^p.
  - Si $X_n \to X$ en L^p y $Y_n \to Y$ en L^q donde $\frac{1}{p} + \frac{1}{q} = 1$, entonces $X_n Y_n \to XY$ en L^1.
- Sea $c \neq 0$ y suponga que $X_n \to X$ en distribución y $Y_n \to y_0$ en distribución[11]. Entonces:
  - $cX_n \to cX$ en distribución
  - $X_n + Y_n \to X + y_0$ en distribución
  - $X_n Y_n \to 0$ en distribución (para $y_0 = 0$).
  - Si g es continua, $g(X_n) \to g(X)$ en distribución.

  A los ítems 1 y 2 se les conoce como Teorema de Slutsky, mientras que al ítem 4 se le conoce como teorema de Mann-Wald.

Teorema 88. Sea $(X_n)_{n\in\mathbb{N}}$ una sucesión tal que $X_n \to X$ en probabilidad y tal que existe $r \in (1, \infty)$ tal que $\{\mathbb{E}[|X_n|^r]\}_{n\in\mathbb{N}}$ es acotada. Entonces, $\mathbb{E}[|X|^r] < \infty$ y para todo $p \in [1, r)$ la sucesión $(X_n)_{n\in\mathbb{N}}$ converge casi seguramente a X en L^p.

Llegamos finalmente al último tópico de teoría de la probabilidad que se expone en este apéndice: el concepto de esperanza condicional.

[11] Como y_0 es una constante, entonces la convergencia es en probabilidad.

Dada una variable aleatoria X integrable y $\mathscr{G}$ un sub $\sigma-$álgebra de $\mathscr{F}$, queremos encontrar una v.a. Y tal que

1. Y es $\mathscr{G}-$ medible.
2. $\int_A Y d\mathbb{P} = \int_A X d\mathbb{P}$, $\forall\ A \in \mathscr{G}$.

$\mathbb{P} : \mathscr{F} \to \mathbb{R}$, $\mathbb{P}|_{\mathscr{G}} : \mathscr{G} \to \mathbb{R}$ es una medida de probabilidad en $(\Omega, \mathscr{G})$. Sean $\nu_1 : \mathscr{G} \to \mathbb{R}$ que hace $A \to \int_A X^+ d\mathbb{P}$ y $\nu_2 : \mathscr{G} \to \mathbb{R}$ que hace $A \to \int_A X^- d\mathbb{P}$. Entonces, $\nu_1, \nu_2 \ll \mathbb{P}_{\mathscr{G}}$. Por Radon-Nikodym, $d\nu_1 = Y_1 d\mathbb{P}|_{\mathscr{G}}$ y $d\nu_2 = Y_2 d\mathbb{P}|_{\mathscr{G}}$. Definimos $Y = Y_1 - Y_2$. Por un lado, es $\mathscr{G}$ medible. Por otro lado, si $A \in \mathscr{G}$

$$\nu_1(A) = \int_A Y_1 d\mathbb{P}|_{\mathscr{G}} = \int_A Y_1 d\mathbb{P}.$$

Luego,

$$\int_A X d\mathbb{P} = \int_A Y d\mathbb{P},\ \forall\ A \in \mathscr{G}.$$

Teorema 89. Sean X una v.a. integrable y $\mathscr{G}$ un sub-σ-álgebra de $\mathscr{F}$.

a) Existe una v.a. Y tal que
   - Y es $\mathscr{G}-$medible.
   - $\int_A Y d\mathbb{P} = \int_A X d\mathbb{P}, \forall\ A \in \mathscr{G}$.

b) Si Z es otra v.a. que cumple con las 2 condiciones, entonces $Z = Y$ c.s.

A Y se le conoce como la esperanza condicional de X dado $\mathscr{G}$ y es denotada por $\mathbb{E}[X|\mathscr{G}]$.

Ejemplo 86. Si $X = c$ constante, entonces $\mathbb{E}[c|\mathscr{G}] = c$.

Ejemplo 87. Si $\mathscr{G} = \{\Omega, \emptyset\}$, entonces $\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X]$.

Ejemplo 88. Si $\Omega = \sum_{k=1}^{n} B_k$ con $B_1, ..., B_n \in \mathscr{F}$, $\mathscr{G} = \sigma(\{B_1, B_2, ..., B_n\})$, entonces

$$\mathbb{E}[X|\mathscr{G}] = \sum_{1\leq k\leq n} \left(\frac{1}{\mathbb{P}(B_n)} \int_{B_k} X d\mathbb{P}\right) \mathbf{1}_{B_k}.$$

Ejemplo 89. Si $\Omega = \sum_{k=1}^{\infty} B_k$ con $B_1, ..., B_n, ... \in \mathscr{F}$, y definimos $\mathscr{G} = \sigma(\{B_k : \ k \geq 1\})$, entonces

$$\mathbb{E}[X|\mathscr{G}] = \sum_{k\geq 1} \left(\frac{1}{\mathbb{P}(B_k)} \int_{B_k} X d\mathbb{P}\right) \mathbf{1}_{B_k}.$$

Ejemplo 90. Sea $A \in \mathscr{F}$ y $\mathscr{G} \subset \mathscr{F}$. La probabilidad condicional de A dado $\mathscr{G}$ es definida como

$$\mathbb{P}[A|\mathscr{G}] = \mathbb{E}[\mathbf{1}_A|\mathscr{G}].$$

Teorema 90. Sean X, Z variables aleatorias integrables y $\mathscr{G}$ un sub $\sigma-$álgebra de $\mathscr{F}$.

1. Si $\alpha \in \mathbb{R}$, entonces $\mathbb{E}[\alpha X|\mathscr{G}] = \alpha\mathbb{E}[X|\mathscr{G}]$.
2. $\mathbb{E}[X + Z|\mathscr{G}] = \mathbb{E}[X|\mathscr{G}] + \mathbb{E}[Z|\mathscr{G}]$.
3. Si $X \leq Z$, entonces $\mathbb{E}[X|\mathscr{G}] \leq \mathbb{E}[Z|\mathscr{G}]$.
4. $\mathbb{E}(\mathbb{E}[X|\mathscr{G}]) = \mathbb{E}[X]$.
5. Si $\mathscr{H} \subset \mathscr{G}$, entonces

$$\mathbb{E}[\mathbb{E}[X|\mathscr{G}]|\mathscr{H}] = \mathbb{E}[X|\mathscr{H}].$$

6. Si $Z \in \mathscr{G}$ y ZX es integrable, entonces

$$\mathbb{E}[ZX|\mathscr{G}] = Z\mathbb{E}[X|\mathscr{G}].$$

7. Si X es independiente de $\mathscr{G}$, entonces

$$\mathbb{E}[X|\mathscr{G}] = \mathbb{E}[X].$$

Teorema 91. Desigualdad de Jensen para esperanza condicional. Supongamos que ϕ es una función convexa y X una variable aleatoria sobre la cual se define la esperanza condicional respecto a una σ-álgebra $\mathcal{G}$. Entonces,

$$\phi(\mathbb{E}[X|\mathcal{G}]) \leq \mathbb{E}[\phi(X)|\mathcal{G}].$$

En todo momento, se asume integrabilidad de.

Teorema 92. Sea (X_n) una sucesión de v.a. integrables:

1. Si $X_n \uparrow X$ entonces $\mathbb{E}[X_n|\mathscr{G}] \uparrow \mathbb{E}[X|\mathscr{G}]$.
2. Si $X_n \downarrow X$ entonces $\mathbb{E}[X_n|\mathscr{G}] \downarrow \mathbb{E}[X|\mathscr{G}]$.
3. Si $X_n \geq 0$, $\forall\ n \geq 1$ entonces

$$\mathbb{E}[\liminf_n X_n|\mathscr{G}] \leq \liminf \mathbb{E}[X_n|\mathscr{G}].$$

4. Si $|X_n| \leq Z$ para todo $n \geq 1$ y $X_n \to X$, entonces $\mathbb{E}[X_n|\mathscr{G}] \to \mathbb{E}[X|\mathscr{G}]$.

La prueba de los Teoremas 90, 91 y 92 se encuentran en Gall (2022).

Definición A.0.28. Sea $(\Omega, \mathscr{F}, \mathbb{P})$ un espacio de probabilidad y X una v.a. integrable. Si Y es una v.a. definimos

$$\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)].$$

Si Y_i, $i \in I$ son v.a., definimos

$$\mathbb{E}[X|Y_i,\ i \in I] = \mathbb{E}[X|\sigma(\{Y_i :\ i \in I\})].$$

Teorema 93. Sean Y, Z variables aleatorias. Las siguientes condiciones son equivalentes:

1. $\sigma(Z) \subset \sigma(Y)$.
2. Z es $\sigma(Y)-$medible.
3. Existe una función Borel medible $g : \mathbb{R} \to \mathbb{R}$ tal que $Z = g(Y)$.

Para la prueba de este teorema, sugerimos consultar Gall (2022). Como consecuencia tenemos que, si X es una v.a. integrable y Y es una v.a. entonces

$$\mathbb{E}[X|Y] = g(Y), \tag{A.1}$$

donde $g : \mathbb{R} \to \mathbb{R}$ es una función Borel medible. Luego, si $y \in \mathbb{R}$ y g cumple (A.1), definimos

$$\mathbb{E}[X|Y = y] = g(y).$$

Teorema 94. Sean X una v.a. positiva integrable y Y una v.a. Sean $\mathbb{P}_Y$ la ley de Y y $\nu : \mathscr{B}_{\mathbb{R}} \to \mathbb{R}$,

$$\nu :\ A \to \int_{Y^{-1}(A)} X d\mathbb{P}.$$

Entonces,

1. ν es una medida en $(\mathbb{R}, \mathscr{B}_{\mathbb{R}})$.
2. $\mathbb{E}[X|Y=y] = \frac{d\nu}{d\mathbb{P}_Y}(y)$.

Note que si $y_0 \in \mathbb{R}$ es tal que $\mathbb{P}(Y=y_0) > 0$, entonces

$$\begin{aligned}\mathbb{E}[X|Y=y_0] &= \mathbb{E}[X|\{Y=y_0\}] \\ &= \frac{1}{\mathbb{P}\{Y=y_0\}} \int_{\{Y=y_0\}} X d\mathbb{P}.\end{aligned}$$

Teorema 95. Sean $X, Y \in L^2$. Entonces:

1. Las variables aleatorias $X, Y, \mathbb{E}[X|Y], X\mathbb{E}[X|Y], \mathbb{E}[Y|X]^2$ son integrables.
2. $\mathrm{Cov}(X,Y) = \mathrm{Cov}(X, \mathbb{E}[Y|X])$.
3. $\mathrm{Var}[\mathbb{E}[Y|X]] \leq \mathrm{Var}[Y]$.

Demostración. Dado que $X, Y \in L^2$, sabemos que $\mathbb{E}[X^2] < \infty$ y $\mathbb{E}[Y^2] < \infty$. Esto implica que tanto X como Y son integrables.

- $\mathbb{E}[X|Y]$ es integrable porque $\mathbb{E}[\mathbb{E}[X|Y]^2] \leq \mathbb{E}[X^2] < \infty$ (por la desigualdad de Jensen para esperanzas condicionales).
- $X\mathbb{E}[X|Y]$ es integrable, dado que:

$$\begin{aligned}\mathbb{E}[(X\mathbb{E}[X|Y])^2] &\leq \mathbb{E}[X^2\mathbb{E}[X|Y]^2] \\ &\leq \mathbb{E}[X^2]\mathbb{E}[\mathbb{E}[X|Y]^2] \\ &\leq \mathbb{E}[X^2]\mathbb{E}[X^2] < \infty.\end{aligned}$$

- $\mathbb{E}[Y|X]^2$ es integrable ya que:

$$\int \mathbb{E}[Y|X]^2 \leq \int \mathbb{E}[Y^2|X] = \int Y^2 < \infty.$$

Para demostrar que $\mathrm{Cov}(X, Y) = \mathrm{Cov}(X, \mathbb{E}[Y|X])$, observamos que:

$$\mathrm{Cov}(X, Y) = \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y],$$
$$\mathrm{Cov}(X, \mathbb{E}[Y|X]) = \mathbb{E}[X\mathbb{E}[Y|X]] - \mathbb{E}[X]\mathbb{E}[\mathbb{E}[Y|X]].$$

Dado que $\mathbb{E}[\mathbb{E}[Y|X]] = \mathbb{E}[Y]$ y

$$\mathbb{E}[XE[Y|X]] = \mathbb{E}[\mathbb{E}[XY|X]] = \mathbb{E}[XY]$$

concluimos que ambas covarianzas son iguales.

Finalmente, la varianza condicional satisface la desigualdad:

$$\begin{aligned}\mathrm{Var}(\mathbb{E}[Y|X]) &= \mathbb{E}[\mathbb{E}[Y|X]^2] - (\mathbb{E}[\mathbb{E}[Y|X]])^2 \\ &= \mathbb{E}[\mathbb{E}[Y|X]^2] - \mathbb{E}[Y]^2 \\ &\leq \mathbb{E}[Y^2] - \mathbb{E}[Y]^2\end{aligned}$$

que proviene directamente de la ley de la varianza total y la desigualdad de Jensen. □

Definición A.0.29. Definimos la varianza condiciones de una variable aleatoria Y dada una variables aleatoria X como

$$\mathrm{Var}[Y|X] = \mathbb{E}[(Y - \mathbb{E}[Y|X])^2|X].$$

La varianza es la desviación cuadrática esperada entre una variable aleatoria (digamos, Y) y su valor esperado. El valor esperado puede considerarse una predicción razonable de los resultados del experimento aleatorio. De hecho, el valor esperado es la mejor predicción constante cuando las predicciones se evalúan por el error cuadrático medio esperado. Así, una interpretación de

la varianza es que proporciona el menor error cuadrático medio posible.

Si tenemos conocimiento de otra variable aleatoria X que podemos usar para predecir Y, podemos potencialmente usar este conocimiento para reducir el error cuadrático medio esperado. Resulta que la mejor predicción de Y dado X es la esperanza condicional.

En particular, para cualquier función medible $f : \mathbb{R} \to \mathbb{R}$:

$$\begin{aligned}\mathbb{E}[(Y - f(X))^2] &= \mathbb{E}[(Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - f(X))^2] \\ &= \mathbb{E}[\mathbb{E}\{(Y - \mathbb{E}(Y|X) + \mathbb{E}(Y|X) - f(X))^2|X\}] \\ &= \mathbb{E}[\mathrm{Var}(Y|X)] + \mathbb{E}[(\mathbb{E}(Y|X) - f(X))^2].\end{aligned}$$

Apéndice B

Elementos de estadística

En este apéndice nos concentramos en ciertos aspectos fundamentales de la inferencia estadística. En particular, estudiamos las distribuciones bivariadas, las distribuciones multivariadas y aspectos vinculados a las muestras aleatorias. Los temas de estimación puntual y por intervalos son abordados a lo largo del cuerpo principal de este texto. Fundamentalmente seguimos a Casella and Berger (2002) y para la estructura[1], las notas de clase del profesor Tomás Rau de la Pontificia Universidad Católica de Chile (Rau (2016)).

En el Apéndice A, ya hemos definido formalmente lo que es una variable aleatoria, su función de distribución y, cuando existe, la función de densidad. Suponga que X es una v.a. con función de distribución F_X, y considere una variable aleatoria $Y = aX + b$ donde $a > 0$ y $b \in \mathbb{R}$. Entonces,

$$F_Y(y) = F_X\left(\frac{y-b}{a}\right).$$

[1] Secuencia de temas presentados.

Además, si X es continua y posee función de densidad,

$$f_Y(y) = \frac{1}{a} f_X\left(\frac{y-b}{a}\right).$$

Nos preguntamos a continuación si hay una forma sistemática de analizar, dada una v.a. X, la distribución y densidad de $Y = g(X)$, con g una función Borel medible.

Teorema 96. Sea X una v.a. con distribución $F_X(x)$. Sea $Y = g(X)$ y $\mathcal{X} = \{x : \ f_X(x) > 0\}$ y $\mathcal{Y} = \{y : y = g(x), \ x \in \mathcal{X}\}$. Entonces:

1. Si g es creciente sobre $\mathcal{X}$[2],

$$F_Y(y) = F_X(g^{-1}(y)), \ y \in \mathcal{Y}.$$

2. Si g es decreciente en $\mathcal{X}$ y X es continua,

$$F_Y(y) = 1 - F_X(g^{-1}(y)), \ \forall \ y \in \mathcal{Y}.$$

La demostración y enunciado original del Teorema 96 se encuentran en Casella and Berger (2002).

Ejemplo 91. Sea $X \sim U[0,1]$ (véase el Apéndice C). Sea $Y = g(X) = -\ln X$. Entonces, como g es estrictamente decreciente sobre $\mathbb{R}_{++}$ y $g^{-1}(y) = e^{-y}$:

$$F_Y(y) = 1 - F_X(g^{-1}(y)) = 1 - F_X(e^{-y}) = 1 - e^{-y}.$$

[2]El soporte de f

Teorema 97. Sea X con función de densidad $f_X(x)$ y $Y = g(X)$ con g una función monótona. Sean $\mathcal{X} = \{x : \ f_X(x) > 0\}$ y $\mathcal{Y} = \{y : y = g(x), \ x \in \mathcal{X}\}$. Supongamos que f_X es continua sobre $\mathcal{X}$ y que $g^{-1} \in C^1(\mathcal{Y})$. Entonces,

$$f_Y(y) = \begin{cases} f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|, & \text{si } y \in \mathcal{Y} \\ 0, & \text{caso contrario.} \end{cases}$$

La prueba es simplemente derivar aplicando regla de la cadena.

Ejemplo 92. Sea $f_X(x)$ la densidad de $X \sim \Gamma(n, \beta)$:

$$f_X(x) = \frac{1}{(n-1)!\beta^n} x^{n-1} e^{-x/\beta}, \ 0 < x < \infty.$$

Sea $g(x) = \frac{1}{x}$. En este caso, $\mathcal{X} = \mathcal{Y} = \mathbb{R}_{++}$. Si hacemos $y = g(x)$, entonces $g^{-1}(y) = 1/y$ y

$$\frac{d}{dy} g^{-1}(y) = -\frac{1}{y^2}.$$

Así, de acuerdo con el Teorema 97

$$f_Y(y) = \frac{1}{(n-1)!\beta^n} \left(\frac{1}{y}\right)^{n-1} e^{-\frac{1}{\beta y}} \frac{1}{y^2} = \frac{1}{(n-1)!\beta^n} \left(\frac{1}{y}\right)^{n+1} e^{-\frac{1}{\beta y}}.$$

A continuación, abordamos el caso de las distribuciones bivariadas y multivariadas. Dicho análisis nos conduce al estudio de los resultados presentados anteriormente en el caso más general.

Definición B.0.1. Un vector aleatoria bivariado es un vector (X, Y) es un vector (X, Y) donde X e Y son variables aleatorias (definidas en un espacio de probabilidad implícito $(\Omega, \mathscr{F}, \mathbb{P})$).

En este caso, $X : \Omega \to \mathbb{R}^2$ induce un espacio de probabilidad $(\mathbb{R}^2, \mathscr{B}_{\mathbb{R}^2}, \mathbb{P}_{X,Y})$, donde

$$\mathbb{P}_{X,Y}(B) = \mathbb{P}\{\omega \in \Omega : (X(\omega), Y(\omega)) \in B\}.$$

Definición B.0.2. Una función de distribución conjunta de (X, Y) es la función $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ definida por

$$\begin{aligned} F_{X,Y}(x, y) &= \mathbb{P}_{X,Y}((-\infty, x], (-\infty, y]) \\ &= \mathbb{P}\{\omega : X(\omega) \leq x, Y(\omega) \leq y\}, \ (x, y) \in \mathbb{R}^2. \end{aligned}$$

Note que si conocemos $F_{X,Y}$ conocemos $\mathbb{P}_{X,Y}$.

Teorema 98. Una función $F : \mathbb{R}^2 \to \mathbb{R}$ es una función de distribución si y solo si

1. $\lim_{x\to-\infty} F(x, y) = 0$ para todo y, $\lim_{y\to\infty} F(x, y) = 0$ para cualquier x y donde $\lim_{x\to\infty, y\to\infty} F(x, y) = 1$.
2. F no es decreciente, esto es, $F(x', y') \geq F(x, y)$ cuando $x' \geq x$, $y' \geq y$.
3. F es continua por la derecha.

Cuando (X, Y) admite una densidad,

- cuando (X, Y) es discreto

$$F_{X,Y}(x, y) = \sum_{s\leq x} \sum_{t\leq y} f_{X,Y}(s, t), \ \forall \ (x, y) \in \mathbb{R}^2$$

- cuando (X, Y) es continuo

$$F_{X,Y}(x, y) = \int_{(-\infty,x]} \int_{(-\infty,y]} f_{X,Y}(t, s) ds dt, \ \forall \ (x, y) \in \mathbb{R}^2.$$

Note que $f_{X,Y}$ es densidad si y solo si Casella and Berger (2002)

- para el caso discreto

$$\sum_{(x,y)\in\mathbb{R}^2} f(x,y) = 1$$

- para el caso continuo

$$\int_{\mathbb{R}}\int_{\mathbb{R}} f(x,y)dxdy = 1.$$

Ejemplo 93. Sea (X,Y) con densidad conjunta

$$f(x,y) = \begin{cases} 6xy^2, & \text{si } 0<x<1,\ 0<y<1 \\ 0, & \text{caso contrario.} \end{cases}$$

Se cumple que

$$\begin{aligned} \iint_{\mathbb{R}^2} f(x,y)dxdy &= \int_0^1\int_0^1 6xy^2dxdy \\ &= \int_0^1 3xY2y^2|_0^1 fy \\ &= \int_0^1 3y^2dy = 1. \end{aligned}$$

A partir de la densidad conjunta, es posible recuperar las densidades marginales $f_X(x)$ y $f_Y(y)$. En efecto,

$$f_X(x) = \sum_{y\in\mathbb{R}} f_{X,Y}(x,y)$$

en el caso discreto, y[3]

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x,y)dy.$$

[3]Cuando $\int_{\mathbb{R}} f_{X,Y}(x,y)dy < \infty$.

Definición B.0.3. Sea (X, Y) un vector aleatorio con densidad conjunta $f_{X,Y}$ y f_X densidad marginal de X. Para cualquier $x \in \mathcal{X}$[4], la densidad condicional de Y dado X es

$$f_{Y|X}(y|x) = \frac{f_{X,Y}(x,y)}{f_X(x)}, \ \forall\ y \in \mathbb{R}.$$

El resultado es análogo cuando intercambiamos X por Y:

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, \ \forall\ x \in \mathbb{R}.$$

Esto aplica para el caso discreto como continuo.

Definición B.0.4. Sea (X, Y) un vector aleatorio con densidad $f_{X,Y}$ y marginales f_X y f_Y. Si X, Y son independientes (véase el Apéndice A), entonces

$$f_{X,Y}(X,Y) = f_X(x) f_Y(y).$$

Esto aplica para el caso discreto como continuo.

Definición B.0.5. Sea (X, Y) un vector aleatorio discreto y sea $g : \mathbb{R} \to \mathbb{R}$ una función Borel medible. Para cualquier y tal que $f_Y(y) > 0$, el valor esperado condicional de $g(X)$ dado $Y = y$, denotado $\mathbb{E}[g(X)|Y = y]$ está dado

$$\mathbb{E}[g(X)|Y = y] \sum_{x \in \mathbb{R}} g(x) f_{X|Y}(x,y),$$

siempre y cuando $g(X)$ sea integrable. Para el caso continuo esto es análogo,

$$\mathbb{E}[g(X)|Y = y] = \int_{\mathbb{R}} g(x) f_{X|Y}(x|y) dx.$$

[4]El soporte de f_X.

En particular, la media condicional de X dado $Y = y$ es $\mathbb{E}[X|Y = y]$ y la varianza condicional de X dado $Y = y$ es

$$\mathbb{E}[X^2|Y = y] - \mathbb{E}[X|y]^2.$$

Definición B.0.6. Sea (X, Y) un vector aleatorio bivariado con densidad conjunta $f_{X,Y}$. Sea $g : \mathbb{R}^2 \to \mathbb{R}$ Borel medible. Entonces,

$$\mathbb{E}[g(X, Y)] = \sum_{(x,y)\in\mathbb{R}^2} g(x, y) f_{X,Y}(x, y).$$

Definición B.0.7. La covarianza de X e Y, con $X, Y, XY \in L^1(\Omega)$ se define como

$$\begin{aligned}\mathrm{Cov}(X, Y) &= \mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])] \\ &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y] \\ &= \mathrm{Cov}(Y, X).\end{aligned}$$

Definición B.0.8. La correlación de X y Y ρ_{XY} es definido por

$$\rho_{XY} = \frac{\mathrm{Cov}(X, Y)}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}}.$$

Definición B.0.9. Sea (X, Y) un vector aleatorio bivariado. La media de (X, Y) es

$$\mathbb{E}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \mathbb{E}[X] \\ \mathbb{E}[Y] \end{bmatrix}.$$

Definición B.0.10. La matriz de covarianza de (X, Y) es

$$\mathrm{Var}\begin{bmatrix} X \\ Y \end{bmatrix} = \begin{bmatrix} \mathrm{Var}(X) & \mathrm{Cov}(X, Y) \\ \mathrm{Cov}(X, Y) & \mathrm{Var}(Y) \end{bmatrix}.$$

Teorema 99. Sea (X, Y) un vector aleatorio bivariado. Si X e Y son independientes, entonces $\text{Cov}(X, Y) = \rho_{XY} = 0$.

Teorema 100. Si (X, Y) es un vector aleatorio bivariado,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y).$$

La prueba es por definición.

Teorema 101. Desigualdad de Cauchy-Schwarz. Si $X, Y \in L^2$ es un vector aleatorio bivariado, entonces

$$|\mathbb{E}[XY]| \leq \mathbb{E}[|XY|] \leq \sqrt{\mathbb{E}[X^2]}\sqrt{\mathbb{E}[Y^2]}.$$

Sea ahora (X, Y) un vector aleatorio con distribución conjunta conocida (distribución de probabilidad conocida) y definamos $U = g_1(X)$ y $V = g_2(Y)$, donde g_1, g_2 son Borel medibles. Si $B \subset \mathbb{R}^2$, entonces

$$(U, V) \in B \Leftrightarrow (X, Y) \in A = \{(x, y) : (g_1(x, y), g_2(x, y)) \in B\}.$$

Por ende,

$$\mathbb{P}\{(U, V) \in B\} = \mathbb{P}\{(X, Y) \in A\}.$$

A continuación vemos como computar $\mathbb{P}\{(U, V) \in B\}$. Para esto, asumiremos que conocemos $f_{X,Y}(x, y)$. Sea

$$J(u, v) = \begin{bmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{bmatrix}.$$

Entonces,

$$f_{U,V}(u, v) = f_{X,Y}(h_1(u, v), h_2(u, v))|J|$$

con $x = h_1(u, v)$ y $x_2 = h_2(u, v)$ y

$$J(u, v) = |J(u, v)| = \frac{\partial x}{\partial u}\frac{\partial y}{\partial v} - \frac{\partial x}{\partial v}\frac{\partial y}{\partial u}.$$

Ejemplo 94. Sea $X \sim \text{Beta}(\alpha, \beta)$ y $Y \sim \text{Beta}(\alpha+\beta, \gamma)$ variables aleatorias independientes. Entonces,

$$f_{X,Y}(x,y) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1} \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha+\beta)\Gamma(\gamma)} y^{\alpha+\beta-1}(1-y)^{\gamma-1}.$$

Consideremos $U = XY$ y $V = X$. Entonces, $B = \{(u,v) : 0 < u < v < 1\}$ y $x = h_1(u,v) = v$, $y = h_2(u,v) = \frac{u}{v}$. Luego,

$$J = \begin{vmatrix} 0 & 1 \\ \frac{1}{v} & -\frac{u}{v^2} \end{vmatrix} = -\frac{1}{v}.$$

Así,

$$f_{U,V}(u,v) = \frac{\Gamma(\alpha+\beta+\gamma)}{\Gamma(\alpha)\Gamma(\beta)\Gamma(\gamma)} v^{\alpha-1}(1-v)^{\beta-1} \left(\frac{u}{v}\right)^{\alpha+\beta-1} \left(1-\frac{u}{v}\right) \gamma - 1 \frac{1}{v}.$$

A continuación pasamos al análisis de las distribuciones multivariadas.

Definición B.0.11. Un vector aleatorio $n-$dimensional es un vector $X = (X_1, \cdots, X_k)^T$ donde $X_1, \cdots, X_k$ son variables aleatorias definidas en un mismo espacio de probabilidad. La medida de probabilidad inducida en $\mathbb{R}^k$ por X es

$$\mathbb{P}_X(B) = \mathbb{P}\{\omega \in \Omega : X(\omega) \in B\},\ B \in \mathscr{B}_{\mathbb{R}^k}.$$

Definición B.0.12. La distribución (conjunta) de un vector aleatorio $k-$dimensional X es la función $F_X : \mathbb{R}^k \to [0,1]$ definida por

$$F_X(x) = \mathbb{P}_X \left\{ \prod_{i=1}^{k} (-\infty, x_i] \right\},\ x = (x_1, \cdots, x_k)^T \in \mathbb{R}^k.$$

Definición B.0.13. Sea X un vector aleatorio $k-$dimensional con distribución conjunta F_X.

1. En caso X sea discreto con densidad f_X[5]

$$F_X(x) = \sum_{t \leq x} f_X(t),\ x \in \mathbb{R}^k.$$

2. En caso X sea continua con densidad f_X

$$F_X(x) = \int\int \cdots \int\int_{t \leq x} f_X(t)dt,\ x \in \mathbb{R}^k.$$

Note que la notación $t \leq x$ simboliza $t_1 \leq x_1, \cdots, t_k \leq x_k$ con $t = (t_1, \cdots, t_k)^T$ y $x = (x_1, \cdots, x_k)^T$.

Definición B.0.14. Sea X un vector aleatorio discreto $k-$dimensional con densidad conjunta f_X. Sea $g : \mathbb{R}^k \to \mathbb{R}$ Borel medible. Entonces, en caso $g(X)$ sea integrable,

$$\mathbb{E}[g(X)] = \sum_{t \in \mathbb{R}^k} g(t) f_X(t).$$

En caso X sea continuo, la situación es análoga y

$$\mathbb{E}[g(X)] = \int_{\mathbb{R}^k} g(x) f_X(x) dx.$$

En el caso más general, en el que $g : \mathbb{R}^k \to \mathbb{R}^{p \times m}$,

$$\mathbb{E}[g(X)] = \begin{bmatrix} E[g_{11}(X)] & \cdots & \mathbb{E}[g_{1m}(X)] \\ \vdots & \ddots & \vdots \\ \mathbb{E}[g_{p1}(X)] & \cdots & \mathbb{E}[g_{pm}(X)] \end{bmatrix},$$

donde

$$\mathbb{E}[g_{ij}(X)] = \int_{\mathbb{R}^k} g_{ij}(x) f_X(x) dx < \infty,$$

provisto que la integral esté bien definida.

[5] $f_X(t) = \mathbb{P}\{X = t\}$.

Definición B.0.15. Sea $X : \Omega \to \mathbb{R}^k$. Entonces,

$$\mathbb{E}[X] = \mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix}$$

y

$$\mathrm{Var}(X) = \mathbb{E}[(X-\mu)(X-\mu)^T] = \Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{bmatrix}.$$

En este contexto,

$$\sigma_{ij} = \mathrm{Cov}(X_i, X_j),\ 1 \leq i, j \leq k.$$

Definición B.0.16. Sea $X = (X_1, \cdots, X_k)^T$ un vector aleatorio $k-$dimensional. Si $a_1, \cdots, a_k, b_1, \cdots, b_k$ con constantes, entonces

$$\mathbb{E}\left[\sum_{i=1}^{k} a_i X_i\right] = \sum_{i=1}^{k} a_i \mathbb{E}[X_i]$$

$$\mathrm{Cov}\left(\sum_{i=1}^{k} a_i X_i, \sum_{j=1}^{k} b_j X_j\right) = \sum_{i=1}^{k}\sum_{j=1}^{k} a_i b_j \mathrm{Cov}(X_i, X_j).$$

En el caso especial en el que las X_i son independientes,

$$\mathrm{Cov}\left(\sum_{i=1}^{k} a_i X_i, \sum_{j=1}^{k} b_j X_j\right) = \sum_{i=1}^{k} a_i^2 \mathrm{Var}(X_i).$$

Definición B.0.17. Sea

$$X = \begin{bmatrix} Y \\ Z \end{bmatrix} \in \mathbb{R}^k$$

con $Y \in \mathbb{R}^n$ y $Z \in \mathbb{R}^{k-n}$. Esto es $Y = (X_1, \cdots, X_n^T)$ y $Z = (X_{n+1}, \cdots, X_k)^T$. Entonces, sobre el soporte[6] de Z

$$f_{Y|Z}(y|z) = \frac{f_{Y,Z}(y,z)}{f_Z(z)}, \ \forall\ y \in \mathbb{R}^n, \ \forall\ z \in \mathbb{R}^{k-n} \cap \text{supp}(f_Z(\cdot))$$

Esto tanto para el caso discreto como continuo.

A continuación, una de las caracterizaciones más usuales de la independencia.

Definición B.0.18. Sean $X_1, ..., X_n$ vectores aleatorios discretos o continuos, no necesariamente de misma dimensión; $X_i : \Omega \to \mathbb{R}^{k_i}$ con densidad f_{X_i}. Entonces,

$$f_{X_1 \cdots X_n}(x_1, \cdots, x_n) = \prod_{i=1}^{n} f_{X_i}(x_i), \ \forall\ x_1, \cdots, x_n.$$

A continuación, una de las distribuciones multivariadas más importantes y frecuentes en la práctica.

Definición B.0.19. Un vector aleatorio $X = (X_1, \cdots, X_k)^T : \Omega \to \mathbb{R}^k$ está normalmente distribuido, con media $\mu = \begin{bmatrix} \mu_1 \\ \vdots \\ \mu_k \end{bmatrix}$ y varianza $\Sigma = \begin{bmatrix} \sigma_{11} & \cdots & \sigma_{1k} \\ \vdots & \ddots & \vdots \\ \sigma_{k1} & \cdots & \sigma_{kk} \end{bmatrix}$, denotado $X \sim N(\mu, \Sigma)$, si X es continuo y con densidad conjunta dada por

$$f_X(x) = \frac{1}{(2\pi)^{k/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right), \ x \in \mathbb{R}^k.$$

[6] $\text{supp}(f_Z(\cdot))$.

Teorema 102. Sea $X \sim N(\mu, \Sigma)$ un vector aleatorio k dimensional. Si $A \in \mathcal{M}_{m\times k}$ y $b \in \mathbb{R}^m$. Entonces,

$$AX + b \sim N(A\mu + b, A\Sigma A^T).$$

Demostración. Por definición. □

Sea

$$X = [X_1, X_2]^T \sim N\left(\mu = \begin{bmatrix}\mu_1 \\ \mu_2\end{bmatrix}, \Sigma = \begin{bmatrix}\Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22}\end{bmatrix}\right),$$

donde $X_1 : \Omega \to \mathbb{R}^n$ y $X_2 : \Omega \to \mathbb{R}^{k-n}$. Entonces,

$$\begin{aligned}
X_1 &\sim N(\mu_1, \Sigma_{11}) \\
X_2 &\sim N(\mu_2, \Sigma_{22}) \\
X_1|X_2 = x_2 &\sim N\left(\mu_1 + \Sigma_{12}\Sigma_{22}^{-1}(x_2 - \mu_2), \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{12}^T\right) \\
X_2|X_1 = x_1 &\sim N\left(\mu_2 + \Sigma_{12}^T\Sigma_{11}^{-1}(x_1 - \mu_1), \Sigma_{22} - \Sigma_{12}^T\Sigma_{11}^{-1}\Sigma_{12}\right).
\end{aligned}$$

Concluimos este apéndice con el tema de las muestras aleatorias.

Definición B.0.20. Sea $X = (X_1, \cdots, X_n)$ un vector aleatorio n dimensional. Las variables aleatorias $X_1, ..., X_n$ se llaman muestra aleatoria si es que son mutuamente independientes y además, tienen la misma distribución (marginal). Esto se denota X_i iid. En dicho caso,

$$\begin{aligned}
F_{X_1\cdots X_n}(x_1, \cdots, x_n) &= \prod_{i=1}^n F_{X_i}(x_i) = \prod_{i=1}^n F(x_i) \\
f_{X_1\cdots X_n}(x_1, \cdots, x_n) &= \prod_{i=1}^n f_{X_i}(x_i) = \prod_{i=1}^n f(x_i).
\end{aligned}$$

Definición B.0.21. Dada una muestra aleatoria $X_1, \cdots, X_n$ y una función medible $T : \mathbb{R}^n \to \mathbb{R}^m$, el vector aleatorio

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

se llama estadístico, y su distribución se llama distribución muestral de Y.

Ejemplo 95. Un ejemplo de estadístico es la media muestral:

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Otro estadístico ampliamente usado es la varianza muestral

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left[X_i - \frac{1}{n}\sum_{i=1}^{n} X_i\right]^2.$$

Quizás llame la atención el factor $\frac{1}{n-1}$ en la definición de la varianza muestral: uno podría anticipar un factor $\frac{1}{n}$. El motivo se explica por el siguiente teorema.

Teorema 103. Sea $X_1, \cdots, X_n$ una muestra aleatoria tal que $\mathbb{E}[X_1] = \mu$ y $\mathrm{Var}(X_1) = \sigma^2$. Entonces,

1. $\mathbb{E}[\overline{X}] = \mu$.
2. $\mathrm{Var}(\overline{X}) = \frac{\sigma^2}{n}$.
3. $\mathbb{E}[S^2] = \sigma^2$.

Demostración. Inciso, por inciso:

$$\begin{aligned}
\mathbb{E}[\overline{X} &= \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\mathbb{E}\left[\sum_{i=1}^{n} X_i\right] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}[X_i] \\
&= \frac{1}{n}\sum_{i=1}^{n} \mu \\
&= \mu.
\end{aligned}$$

Luego,

$$\begin{aligned}
\mathrm{Var}(\overline{X}) &= \mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\mathrm{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n} \mathrm{Var}(X_i) \\
&= \frac{\sigma^2}{n}.
\end{aligned}$$

Finalmente,

$$\begin{aligned}
\mathbb{E}[S^2] &= \mathbb{E}\left[\frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right)\right] \\
&= \frac{1}{n-1}\mathbb{E}\left[\sum_{i=1}^{n} X_i^2 - n\overline{X}^2\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} \mathbb{E}[X_i^2] - n\mathbb{E}[\overline{X}^2]\right] \\
&= \frac{1}{n-1}\left[\sum_{i=1}^{n} (\sigma^2 + \mu^2) - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right] \\
&= \sigma^2.
\end{aligned}$$

Note que se ha hecho uso de la siguiente relación,

$$\mathbb{E}[X_i^2] = \mathrm{Var}(X_i) + \mathbb{E}[X_i]^2 = \sigma^2 + \mu^2.$$

□

Esto concluye el apéndice sobre elementos de estadística. A lo largo del cuerpo principal de este documento se abordan otros temas de la inferencia estadística: los estimadores, los intervalos de confianza etc.

Apéndice C

Distribuciones usuales

1. Binomial $B(n,p)$

   a) $\mathbb{P}\{X=k\} = \binom{n}{k} p^k (1-p)^{n-k}$

   b) $M_X(t) = (1-p+pe^t)^n$

   c) $\varphi_X(t) = (1-p+pe^{it})^n$

   d) $\mathbb{E}[X] = np$

   e) $\mathrm{Var}[X] = np(1-p)$.

2. Geométrica $G(p)$

   a) $\mathbb{P}\{X=k\} = p(1-p)^{k-1}$

   b) $M_X(t) = \frac{pe^t}{1-(1-p)e^t}$

   c) $\varphi_X(t) = \frac{pe^{it}}{1-(1-p)e^{it}}$

   d) $\mathbb{E}[X] = \frac{1}{p}$

   e) $\mathrm{Var}[X] = \frac{1-p}{p^2}$

3. Binomial negativa $BN(r,p)$

   a) $\mathbb{P}\{X=k\} = p^r(1-p)^k\binom{r+k-1}{k}$

   b) $M_X(t) = \left(\frac{p}{1-(1-p)e^t}\right)^r$

   c) $\varphi_X(t) = \left(\frac{p}{1-(1-p)e^{it}}\right)^r$

   d) $\mathbb{E}[X] = \frac{r(1-p)}{p}$

   e) $\mathrm{Var}[X] = \frac{r(1-p)}{p^2}$

4. Multinomial $B(n,p_1,...,p_k)$

   a) $\mathbb{P}\{X=(X_1,\cdots X_k)=(n_1,...,n_k)\} = \frac{n!}{n_1!\cdots n_k!}p_1^{n_1}\cdots p_k^{n_k}$

   b) $M_X(t) = \left(\sum_{i=1}^k p_i e^{t_i}\right)^n$

   c) $\varphi_X(t) = \left(\sum_{j=1}^k p_j e^{it_j}\right)^n$

   d) $\mathbb{E}[X_i] = np_i$

   e) $\mathrm{Var}[X_i] = np_i(1-p_i)$

5. Poisson $\mathcal{P}(\lambda)$

   a) $\mathbb{P}\{X=k\} = \frac{\lambda^k e^{-\lambda}}{k!}$

   b) $M_X(t) = e^{\lambda(e^t-1)}$

   c) $\varphi_X(t) = e^{\lambda(e^{it}-1)}$

   d) $\mathbb{E}[X] = \lambda$

   e) $\mathrm{Var}[X] = \lambda$

6. Normal $N(\mu,\sigma)$

   a) $\mathbb{P}\{X\in A\} = \int_A \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(x-\mu)^2}{2\sigma^2}}dx$

b) $M_X(t) = e^{\mu t + \frac{\sigma^2 t^2}{2}}$

c) $\varphi_X(t) = e^{\mu i t - \frac{\sigma^2 t^2}{2}}$

d) $\mathbb{E}[X] = \mu$

e) $\text{Var}[X] = \sigma^2$

7. Uniforme $U([a, b])$

a) $\mathbb{P}\{X \in A\} = \int_A \mathbf{1}_{[a,b]} \frac{1}{b-a} dx$

b) $M_X(t) = \frac{e^{tb} - e^{ta}}{t(b-a)}$

c) $\varphi_X(t) = \frac{e^{itb} - e^{ita}}{it(b-a)}$

d) $\mathbb{E}[X] = \frac{a+b}{2}$

e) $\text{Var}[X] = \frac{(b-a)^2}{12}$

8. Exponencial $Exp(\lambda)$

a) $\mathbb{P}\{X \in A\} = \int_A \mathbf{1}_{[0,\infty)} \lambda e^{-\lambda t} dt$[1]

b) $M_X(t) = \frac{\lambda}{\lambda - t}$

c) $\varphi_X(t) = \frac{\lambda}{\lambda - it}$

d) $\mathbb{E}[X] = \frac{1}{\lambda}$

e) $\text{Var}[X] = \frac{1}{\lambda^2}$

9. Gamma $\Gamma(\alpha, \lambda)$

a) $\mathbb{P}\{X \in A\} = \int_A \mathbf{1}_{\mathbb{R}_{++}} \frac{\lambda(\lambda x)^{\alpha-1} e^{-\lambda x}}{\Gamma(\alpha)} dx$

b) $M_X(t) = \left(\frac{\lambda}{\lambda - t}\right)^{\alpha}$

c) $\varphi_X(t) = \left(\frac{\lambda}{\lambda - it}\right)^{\alpha}$

[1] $F_X(t) = 1 - e^{-\lambda t}$

d) $\mathbb{E}[X] = \frac{\alpha}{\lambda}$

e) $\text{Var}[X] = \frac{\alpha}{\lambda^2}$

10. Weibull $W(\alpha, \lambda)$

a) $\mathbb{P}\{X \in A\} = \int_A \mathbf{1}_{\mathbb{R}_{++}} \lambda\alpha(\lambda x)^{\alpha-1} e^{-(\lambda x)^\alpha} dx$

b) $M_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!\lambda^n} \Gamma\left(1 + \frac{n}{\alpha}\right)$

c) $\varphi_X(t) = \sum_{n=0}^{\infty} \frac{(it)^n}{n!\lambda^n} \Gamma\left(1 + \frac{n}{\alpha}\right)$

11. Cauchy $\mathcal{C}(x_0, \gamma)$

a) $\mathbb{P}\{X \in A\} = \int_A \frac{1}{\pi\gamma\left[1+\left(\frac{x-x_0}{\gamma}\right)^2\right]}$

b) $\varphi_X(t) = e^{x_0 it - \gamma|t|}$

c) $\mathbb{E}[X] = x_0$

12. Lognormal[2]

a) $\mathbb{P}\{X \in A\} = \int_A \mathbf{1}_{\mathbb{R}_{++}} \frac{1}{x\sigma\sqrt{2\pi}} e^{-\frac{(\ln x - \mu)^2}{2\sigma^2}} dx$

b) $\mathbb{E}[X] = e^{\mu + \frac{\sigma^2}{2}}$

c) $\text{Var}[X] = (e^{\sigma^2} - 1)(e^{2\mu+\sigma^2})$.

Note que algunas distribuciones, como la Cauchy, Weibull o Lognormal, no tienen definidas una función generadora de momento o su varianza.

Por otro lado, recordemos los siguiente. Cuando tenemos

$$\mathbb{P}\{X \in A\} = \int_A f_X(x)dx,$$

[2]Si $X \sim Ln(\mu, \sigma^2)$, entonces $Y = \ln X \sim N(\mu, \sigma^2)$

entonces decimos que $f_X(x)$ es la densidad de X. Además, para una v.a. continua, podemos definir

$$F(x) = \mathbb{P}\{X \leq x\} = \int_{-\infty}^{x} f_X(x)dx,$$

que resultar la función de distribución o densidad acumulada de X. En el caso discreto,

$$F_X(x) = \sum_{t \leq x} p_X(t),\ p_X(t) = \mathbb{P}\{X = t\}.$$

Notar que, en ambas situaciones $f_X(x) \geq 0$ ($p_X(t) \geq 0$) y

$$\sum_{x \in \mathbb{R}} f(x) = 1$$

$$\sum_{t \in I \subset \mathbb{R}} p_X(t) = 1,\ I \text{ enumerable infinito o finito.}$$

A I se le conoce como el soporte de X.

Bibliografía

Abbott, S. (2015). *Understanding Analysis*. Springer, 2 edition.

Angrist, J. (1990a). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *The American Economic Review*, 80(3):313–336.

Angrist, J. and Krueger, A. (1991a). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

Angrist, J. and Pischke, J.-S. (2009). *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, 1 edition.

Angrist, J. D. (1990b). Lifetime earnings and the vietnam era draft lottery: Evidence from social security administrative records. *American Economic Review*, 80(3):313–336.

Angrist, J. D. and Krueger, A. B. (1991b). Does compulsory school attendance affect schooling and earnings? *Quarterly Journal of Economics*, 106(4):979–1014.

Axler, S. (2015). *Linear Algebra Done Right*. Springer, 3 edition.

Bai, J. and Perron, P. (1998). Estimating and testing linear models with multiple structural changes. *Econometrica*, 66(1):47–78.

Barro, R. and Martin, X. S. I. (2003). *Economic Growth*. MIT Press, 1 edition.

Borjas, G. (2000). *Labor Economics*. McGraw-Hill, 2 edition.

Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, 1 edition.

Card, D. (1995). Using geographic variation in college proximity to estimate the return to schooling. *National Bureau of Economic Research Working Paper*, (4483).

Casella, G. and Berger, R. (2002). *Statistical Inference*. Thomson Learning, 2 edition.

Chavez, J. and Gallardo, M. (2023). *Álgebra Lineal y Optimización para el Análisis Económico*. Pre-published, 1 edition.

Chow, G. (1960). Tests of equality between sets of coefficients in two linear regressions. *Econometrica*, 28(3):591–605.

Cohen, J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 2 edition.

de la Fuente, A. (2000). *Mathematical Methods and Models for Economists*. Cambridge University Press.

Douglas Staiger, J. H. S. (1997). Instrumental variables regression with weak instruments. *Econometrica*, 65(3):557–586.

Durbin, J. and Watson, G. (1950). Testing for serial correlation in least squares regression: I. *Biometrika*, 37(3):409–428.

Efron, B. (1979). Bootstrap methods another look at the jackknife. *The Annals of Statistics*, 7(1):1–26.

Folland, G. (1984). *Real Analysis: Modern Techniques and Their Applications*. Wiley, 1 edition.

Frankel, J. A. and Romer, D. (1999). Does trade cause growth? *American Economic Review*, 89(3):379–399.

Gall, J.-F. L. (2022). *Measure Theory, Probability, and Stochastic Processes*. Springer Verlag, 1 edition.

Geary, R. (1950). A note on a constant-utility index of the cost of living. *The Review of Economic Studies*, 18(1):65–66.

Girfone, J. (2018). *Algebre Linéaire*. Cepadues, 6 edition.

Godfrey, L. (1978). Testing against general autoregressive and moving average error models when the regressors include lagged dependent variables. *Econometrica*, 46(6):1293–1301.

Greene, W. (2015). *Econometric Analysis*. Prentice Hall, 5 edition.

Gujarati, D. and Porter, D. (2010). *Econometría*. McGraw Hill, 5 edition.

Hansen, L. P. (1982). Large sample properties of generalized method of moments estimators. *Econometrica*, 50(4):1029–1054.

Hausman, J. A. (1978). Specification tests in econometrics. *Econometrica*, 46(6):1251–1271.

John Bound, David A. Jaeger, R. M. B. (1995). Problems with instrumental variables estimation when the correlation between the instruments and the endogenous explanatory variable is weak. *Journal of the American Statistical Association*, 90(430):443–450.

John Chao, N. S. (2005). Consistent estimation with a large number of weak instruments. *Econometrica*, 73(5):1673–1692.

John G. Cragg, S. G. D. (1993). Testing identifiability and specification in instrumental variable models. *Econometric Theory*, 9(2):222–240.

Lenberger, D. and Ye, Y. (2021). *Linear and Nonlinear Programming*. Springer Verlag, 5 edition.

Ljung, G. and Box, G. P. (1978). On a measure of lack of fit in time series models. *Biometrika*, 65(2):297–303.

Manski, C. (1988). *Analog Estimation Methods in Economics*. Chapman and Hall, 1 edition.

Mas-Colell, A., Whinston, M., and Green, J. (1995). *Microeconomic Theory*. Oxford University Press, 1 edition.

Polachek, S. (2007). Earning over the lifecycle the mincer earning function and its applications. *IZA*, (3181).

Rau, T. (2016). *Teoría Econométrica I*. Notas de clase, 1 edition.

Roman, S. (2008). *Advanced Linear Algebra*. Springer Verlag.

Rothenberg, T. J. (1984). Approximating the distributions of econometric estimators and test statistics. *Handbook of Econometrics*, 2:881–935.

Sargan, J. D. (1958). The estimation of economic relationships using instrumental variables. *Econometrica*, 26(3):393–415.

Self, S. and Liang, K.-Y. (1987). Asymptotic properties of maximum likelihood estimators and likelihood ratio tests under nonstandard conditions. *Journal of the American Statistical Association*, 82(398).

Simon, C. and Blume, L. (1994). *Mathematics for Economists*. W.W Norton and Company, 1 edition.

Stock, J. H. and Yogo, M. (2005). Testing for weak instruments in linear iv regression. *Cambridge University Press*.

Stone, R. (1954). Linear expenditure systems and demand analysis an application to the pattern of british demand. *The Economic Journal,*, 64(255):511–527.

Sundaram, R. (1996). *A First Course in Optimization Theory*. Camrbdige University Press, 1 edition.

Tao, T. (2016). *Analysis 2*. Springer, 3 edition.

Valdivieso, L. (2020). *Notas de Técnicas de Muestreo*. Fondo Editorial PUCP, 1 edition.

Weiss, L. (1971). Asymptotic properties of maximum likelihood estimators in some nonstandard cases. *Journal of the American Statistical Association*, 66(334):345–350.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica*, 48.

Wooldridge, J. (2001). *Econometric Analysis of Cross Section and Panel Data*. MIT University Press, 1 edition.