Econometría I

(MSc. Economía)

Damian Clarke[1]

APUNTES PRELIMINARES

Marzo, 2019

[1]Profesor Asociado, Departamento de Economía, Universidad de Santiago de Chile. Sitio web: damianclarke.net, correo electrónico damian.clarke@usach.cl. Agradezco a Kathya Tapia Schythe por muchos comentarios y la inclusión de ejercicios aplicados.

Contents

Definiciones

Símbolo/Término	Definición
x	Un valor escalar
$\mathbf{x}$	Un vector
$\mathbf{X}$	Una matriz
iid	Independiente e idénticamente distribuida
$inid$	Independiente pero no idénticamente distribuida
$\mathbb{R}$	Número Real
$\mathbb{R}^N$	Una tupla de N números reales
$\langle \cdot, \cdot \rangle$	Espacio prehilbertiano para dos vectores
sii	si y solo si
$\lvert\mathbf{A}\rvert$	Determinante (de la matriz $\mathbf{A}$)
$\cdot \perp \cdot$	Ortogonalidad para dos vectores
$A \subset B$	Contención (el conjunto B contiene A)
$\Leftrightarrow$	Equivalencia
$A \Rightarrow B$	Implicancia (A implica B)
$\emptyset$	El conjunto vacío
$\arg\min_x f(x)$	El valor de x que minimiza la función $f(x)$
$\min_x f(x)$	El valor de la función $f(x)$ en su punto mínimo evaluado sobre x
…	

Sección 1

Introducción a la Econometría

Nota de Lectura: Para una muy buena (y breve) introducción a la econometría, se recomienda leer el capítulo 1 de Hansen (2017), específicamente las secciones 1.1 a 1.7. Se puede encontrar una definición temprana de la econometría en la primera nota editorial introduciendo la revista *Econometrica*, de Frisch (1933). El capítulo 1 de Stachurski (2016) también ofrece una buena introducción y resúmen. Ambos capítulos están libremente disponibles en línea, aquí y aquí.

1.1 Introducción al Curso de Teoria Econométrica

Las detalles completas de este curso estarán disponibles en el siguiente sitio web:

http://damianclarke.net/teaching/Teoria-Econometrica

Se sugiere revisar este sitio como la fuente oficial de la información principal del curso incluyendo el programa del mismo, el calendario, ejercicios computacionales, trabajos, y pruebas pasadas. El curso de este año difiere algo en cursos de años previos, incluyendo una sección nueva de repaso de herramientas algebraícas. Así, aunque las pruebas y exámenes anteriores pueden ser fuentes para repasar material, la estructura este año va a cambiar levamente.

Este curso está diseñado como el primer curso de un ciclo de hasta cuatro cursos de econometría en el Magíster en Ciencias Económicas. En el segundo semestre del magíster será seguido por el curso de Teoría econométrica II que introduce otros modelos y teécnicas, incluyendo modelos no-lineales. En el segundo año del magíster hay dos cursos electivos: uno "Topicos de Microeconometría" enfocado netamente de aplicaciones empíricas de modelos nuevos en microeconometría como modelos de regresión discontinua y modelos de diferencias-en-diferencias, y otro enfocado en modelos de series de tiempo, frecuentamente encontrado en aplicaciones *macro*-econométricas.

1.1.1 Estos Apuntes

Estos apuntes son un trabajo en progreso. Reunen resultados importantes para formar una base sólida en algunos modelos fundamentales de econometría, incluyendo el modelo de regresión lineal. Vienen de diversas fuentes, con referencias útiles y/o comprensivos indicados al inicio de cada sección. Además de los textos citados, incluyen resultados expuestos en los cursos de Steve Bond, Bent Nielsen, Nicolas van de Sijpe, Simon Quinn, Sonia Bhalotra, Debopam Bhattacharya y Andrew Zeitlen (algunos de lo/as profesores/as que me enseñaron econometría). Esto no es un texto original, sino apuntes de clases.

Estos apuntes fueron escritos para acompañar el curso "Teoría Econométrica I" del Magíster en Ciencias Economícas en la Universidad de Santiago de Chile. Son apuntes nuevos (en 2019), y por lo tanto deben ser considerados como un trabajo en progreso. Cualquier comentario, sugerencia, y/o corrección es muy bienvenido.

La idea de los apuntes es complementar nuestra discusión en clases, su lectura de otros libros y papers, y los problemas aplicados que revisamos en clases y en ayudantías. En varias secciones de los apuntes hay un apartado de "Nota de Lectura". Estas notas describen fuentes comprensivas para revisar el material de la sección. Las lecturas recomendadas aquí no son obligatorias, sin embargo, pueden ayudar a fortalecer su aprendizaje cuando son leídos para complementar estos apuntes. Si hay temas de los apuntes que no quedan claros o que le gustaría revisar en más detalle, estas lecturas son la mejor fuente para resolver dudas (además de las preguntas en clases). Si los libros indicados no están disponibles en la biblioteca, se los puede pedir del profesor.

1.2 Econometría

La econometría—más que ser una simple aplicación de los métodos de estadística a problemas economícos—tiene sus propios fundamentos y metas. La econometría une la teoría estadística formal con datos empíricos del mundo real y con la teoría y modelos económicos. A menudo cuando planetamos modelos econométricos nos interesa la relación *causal* entre variables. Una fortaleza de la econometría es que nos proporciona las herramientas necesarias para hablar en términos causales (o de qué pasaría *ceteris paribus* al variar una variable de interés), **si nuestros supuestos de identificación son correctos**.

En la econometría es común trabajar con datos observacionales, en vez de datos experimentales (aunque también hay aplicaciones experimentales en la econometría). Esto implica llegar con una pregunta de interés y datos que ya existen, cuando nos gustaría saber qué pasaría al variar alguna variable sin cambiar ninguna otra. Como los datos observacionales generalmente no proporcionan la variación de *una sola variable*, nuestros modelos econométricos tienen que intentar encontrar una manera de aislar estos cambios únicos utilizando la variación encontrada por sistemas naturales y económicos. Será una preocupación constante asegurar que nuestros estimadores capturan sólo el

cambio de una variable de interés, y no cambios simultáneos de otras variables. Cuando logramos hacer esto, nos permitirá hablar en términos de causalidad en vez de correlación.

La econometría es un campo bastante amplio, desde modelos estáticos con supuestos muy paramétricos, a modelos que siguen a sus observaciones durante muchos periodos, o ponen muy poca estructura en sus supuestos (eg modelos no paramétricos). En este curso nos enfocaremos en las herramientas básicas que nos sirvirían en cualquier área de econometría (probabilidad y algébra lineal), y después introducimos una serie de modelos paramétricos. En cursos futuros del magíster, se examinará otros tipos de modelos y supuestos econométricos.

Aunque la econometría se basa en supuestos económicos y de probabilidad, los estudios aplicados de econometría que abarcan áreas *muy* amplias, incluyendo salud, educación, organización industrial, economía política, y cualquier otra área de estudio donde se interesa aislar el impacto causal de una(s) variable(s) independientes sobre otras variables de interés…. A continuación discutiremos un poco acerca de las aplicaciones y estudios empíricos en econometría.

1.3 Recursos

1.3.1 Estudios Teóricos y Aplicados

El canal de comunicación principal para compartir nuevos resultados en econometría (y en economía de forma más genérica) es a través de journals (o revistas) científicas. Cuando se concentra en el desarrollo de la econometría teórica, una proporción importante de los avances en este campo se publica en journals como: Econometrica, Journal of Econometrics, The Econometrics Journal, Journal of the American Statistical Association, y Econometric Theory. Estos journals publican ediciones varias veces por año con estudios que han sido juzgados por sus pares como teóricamente correctos y como contribuciones importantes al campo de la econometría teórica. Los journals son una muy buena fuente para seguir el desarrollo de los temas activos en econometría moderna.

Además de artículos demostrando los avances en el campo de econometría teórica, hay una multiplicidad de journals que publican articulos que utilizan herramientas econométricas de forma aplicada, incluyendo: The Quarterly Journal of Economics, The American Economic Review, The Review of Economic Studies, The Journal of Political Economy, The Review of Economics and Statistics, The Journal of the European Economic Association, The American Economic Journals, y The Economic Journal. Además, papers aplicados a ciertos sub-especialidades de economía/econometría se publican en journals especializados en cada campo. Algunos ejemplos incluyen: The Journal of Labor Economics (economía laboral), Journal of Health Economics (salud), Journal of Monetary Economics (economía monetaria), Journal of Development Economics (desarrollo económico), …. Es una buena idea seguir los papers que salen en estos journals, especialmente los journals en los campos que más le interesan, para ver el estado del arte de econometría aplicada. Aunque este

curso se enfoca casi exclusivamente en la teoría econométrica en vez de aplicaciones empíricas, la lectura de estudios aplicados es una manera entretenida de entender como la teoría que revisamos en estos apuntes se traduce en aplicaciones reales.

Por último, los estudios más nuevos, antes de salir publicados en algún journal salen como un *working paper*. Los *working papers* son utilizados para comunicar resultados entre investigadores de economía/econometría y para recolectar comentarios, y además sirven como una manera para compartir resultados antes de que salgan definitivamente en el journal. Como el proceso de publicación en economía puede demorar bastante (no es atípico que un paper sale publicado dos o tres años después de salir por primera vez como un *working paper*), los *working papers* sirven como una manera más inmediata para hacer conocido resultados importantes. Si le interesa mantenerse al tanto de los desarrollos más nuevos en economía y econometría, es una buena idea inscribirse para recibir un resúmen temporal de *working papers*. Algunas buenas fuentes de estos papers son la serie del National Bureau of Economic Research, la serie de IZA Institute of Labor Economics, o inscribiéndose en los listados de NEP: New Economics Papers, que existen en casi 100 sub-especialidades de economía.

1.3.2 Herramientas

Al momento de estimar modelos econométricos con datos reales, es (casi siempre) necesario contar con un computador, e idealmente algún idioma computacional con implementaciones de modelos comunes de econometría. Existen muchas opciones muy buenas de idiomas con fortalezas en econometría. Esto incluye idiomas como Python, Julia, R, Octave y Fortran (todos libres), y Stata, SAS, y MATLAB (pagados). Como la econometría se basa en mucha álgebra lineal, es particularmente conveniente contar con un idioma basado en matrices para simulaciones y aplicaciones para explorar resultados teóricos importantes. Idiomas que son especialmente fácil con matrices incluyen Julia, MATLAB, y Mata. Este úlitm (Mata) es un tipo de sub-idioma que existe adentro de Stata con una sintaxis enfocada en algébra lineal y manipulación de matrices.

En este curso, generalmente utilizamos Stata y Mata. Estos idiomas tienen muchas herraimentas muy desarrolladas enfocadas en econometría, y una serie de documentación muy comprensiva. Pero un costo de Stata y Mata es, justamente, su costo! A diferencia de Python, Julia y otras, no es un idioma libre ni gratis. En la universidad tendrán acceso libre a Stata y Mata. Sin embargo, si le interesa trabajar con otros idiomas de los mencionados anteriormente (u otros), no hay problema! Existen muchos libros y materiales muy buenos con un enfoque de econometría computacional incluyendo Cameron and Trivedi (2009) (Stata), Adams, Clarke and Quinn (2015) (MATLAB), y el excelente curso de Stachurski y Sargent: https://lectures.quantecon.org/ (Python y Julia) con un enfoque más amplio que solo econometría, a modelos economicos más generalmente.

Sección 2

Un Repaso de Herramientas Algebraicas

Nota de Lectura: Para un repaso de álgebra lineal existen muchas fuentes interesantes, tanto de matemática como de econometría. Un análisis de bastante alto nivel está disponible en Stachurski (2016), capítulos 2 y 3. Rao (1973) es una referencia clásica, con demostraciones muy elegantes. Generalmente, los libros de texto de econometría tienen un capítulo o apéndice de revisión, por ejemplo apéndice A de Hansen (2017).

2.1 Operaciones y Elementos Básicos

2.1.1 Elementos Básicos

Un valor escalar, escrito x es un único número. Un vector, genéricamente de $\mathbb{R}^N$, contiene N elementos, o escalares y se escribe como:

$$\mathbf{x} = \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{pmatrix},$$

con cada $x_n \in \mathbb{R}$. En esta notación $\mathbb{R}$ refiere a los números reales, y N a la cantidad de números naturales contenidos en el vector, o dimensiones. En el caso de que $N = 1$, $\mathbb{R} = \mathbb{R}^1$ es la línea de números reales, que es la unión de los números racionales e irracionales. Un único elemento de un vector $\boldsymbol{x}$ es un escalar.

Definimos una matriz de $N \times K$ dimensiones, llamado $\mathbf{X}$. La matriz $\mathbf{X}$ contiene números reales en N filas y K columnas. Las matrices son particularlmente útiles para agrupar datos u operaciones algebraicas. En estos apuntes siempre utilizaremos letras minúsculas x para denotar valores escalares, letras minúsculas con negrita $\mathbf{x}$ para denotar vectores, y letras mayúsculas en negrita $\mathbf{X}$

para denotar matrices. Entonces, una matríz de tamaño $N \times K$ refiere a una colección de número reales, y se escribe como:

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix}.$$

Cuando escribimos x_{nk}, referimos a la entrada en la fila n-ésima y la columna k-ésima. En el caso de que $N = K$, la matriz $\mathbf{X}$ de $N \times N$ es conocido como una matriz cuadrada, y los elementos x_{nn} para $n = 1, 2, \ldots, N$ son conocidos como el diagonal principal de la matriz.

2.1.2 Operaciones Básicas

Las operaciones algebraicas básicas con matrices refieren a *adición*, y *multiplicación*. La adición, o sumatorias con matrices, es un proceso elemento por elemento. Para $\mathbf{W}, \mathbf{X} \in \mathbb{R}^{N \times K}$ su sumatoria es:

$$\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & \cdots & w_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NK} \end{pmatrix} + \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} = \begin{pmatrix} w_{11} + x_{11} & w_{12} + x_{12} & \cdots & w_{1K} + x_{1K} \\ w_{21} + x_{21} & w_{22} + x_{22} & \cdots & w_{2K} + x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} + x_{N1} & w_{N2} + x_{N2} & \cdots & w_{NK} + x_{NK} \end{pmatrix}.$$

De esta sumatoria, es aparente que para sumar dos (o más) matrices es necesario que sean del mismo tamaño, y en este caso se dice que las matrices son conformables para adición. La sumatoria de vectores sigue exactamente la misma lógica, ya que una matriz de $N \times 1$ o de $1 \times K$ es un vector (llamado un vector de columna o vector de fila, respectivamente), y la definición anterior cumple siempre cuando ambos vectores en la sumatoria sean del mismo tamaño.

A continuación se resumen algunas de las propiedades de la adición de matrices. En este listado simplemente resumimos las propiedades, y dejamos la demostración de estas propiedades como un ejercicio.

Propiedades de Sumatorias de Matrices

1. Asociatividad: $(\mathbf{X} + \mathbf{Y}) + \mathbf{Z} = \mathbf{X} + (\mathbf{Y} + \mathbf{Z})$
2. Conmutatividad: $(\mathbf{X} + \mathbf{Y}) = (\mathbf{Y} + \mathbf{X})$
3. Existencia de un Elemento Nulo: $\mathbf{X} + \mathbf{0} = \mathbf{X}$

El elemento nulo en el ítem 3 se refiere a un vector conformable para la adición con $\mathbf{X}$ cuyos elementos x_{ij} son todos iguales a 0. Volveremos a una serie de matrices importantes, incluyendo la matriz nula, en la sección 2.2.

Multiplicación La segunda operación básica de matrices es la multiplicación. La múltiplicación escalar—donde una matriz se multiplica con un valor escalar α—se define de la misma forma que múltiplicación en $\mathbb{R}$. Así, cuando una matriz se múltiplica con un único valor, se multiplica elemento por elemento para llegar a la solución:

$$\alpha \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} = \begin{pmatrix} \alpha x_{11} & \alpha x_{12} & \cdots & \alpha x_{1K} \\ \alpha x_{21} & \alpha x_{22} & \cdots & \alpha x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \alpha x_{N1} & \alpha x_{N2} & \cdots & \alpha x_{NK} \end{pmatrix}.$$

En la multiplicación de dos matrices para formar el producto matricial $\mathbf{Z} = \mathbf{XY}$, cada elemento z_{ij} se calcula de la siguiente forma:

$$z_{ij} = \sum_{k=1}^{K} w_{ik} x_{kj} = \langle \text{fila}_i \, \boldsymbol{W}, \text{columna}_j \, \boldsymbol{X} \rangle. \tag{2.1}$$

La segunda notación $\langle \cdot, \cdot \rangle$ refiere al espacio prehilbertiano, que es la suma del producto de los elementos de dos vectores. Este cálculo con dos matrices en forma extendida está presentado de forma esquemática en la ecuación 2.2.

$$\begin{pmatrix} w_{11} & w_{12} & \cdots & w_{1K} \\ w_{21} & w_{22} & \cdots & w_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ w_{N1} & w_{N2} & \cdots & w_{NK} \end{pmatrix} \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix} = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1K} \\ z_{21} & z_{22} & \cdots & z_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ z_{N1} & z_{N2} & \cdots & z_{NK} \end{pmatrix} \tag{2.2}$$

El elemento z_{11} se calcula a partir de la primera fila y la primera columna (los vectores azules):

$$z_{11} = w_{11}x_{11} + w_{12}x_{21} + \ldots + w_{1K}x_{N1},$$

que es el espacio prehilbertiano para fila 1 y columa 1. Todos los otros elementos se calculan de la misma forma, por ejemplo z_{2K} a partir de fila 2 y columna K. El espacio prehilbertiano es sólo definido para vectores con la misma cantidad de elementos, que implica que las matrices sólo son *conformables* (o pueden ser multiplicadas) si la cantidad de elementos en las filas de $\mathbf{W}$ es igual a la cantidad de elementos en las columnas de $\mathbf{X}$. Si $\mathbf{W}$ es $N \times K$ y $\mathbf{X}$ es $J \times M$, esto implica que una matriz es conformable solo si $K = J$. En este caso, el producto $\mathbf{Z}$ será de $N \times M$.

Por lo general, la multiplicación matricial no es conmutativa: $\mathbf{XY} \neq \mathbf{YX}$. En algunos casos, una matriz que puede ser pre-multiplicada con otra matriz ni siquera es conformable cuando se post-multiplica con la misma matriz. Por ejemplo, multiplicando una matriz de 5×2 con otra de 2×4 resulta en una matriz de 5×4, pero al revés no es conformable, dado que hay 4 columnas en la primera matriz, y 5 filas en la segunda. Y en otros casos, aunque dos matrices sean conformables,

el resultado no es lo mismo multiplicando de ambas formas. Por ejemplo, consideremos:

$$\mathbf{A} = \begin{pmatrix} 2 & 3 \\ 1 & 0 \end{pmatrix} \qquad \mathbf{B} = \begin{pmatrix} 4 & 1 \\ 2 & 1 \end{pmatrix}$$

Es fácil confirmar que:

$$\mathbf{AB} = \begin{pmatrix} 14 & 5 \\ 4 & 1 \end{pmatrix} \qquad \mathbf{BA} = \begin{pmatrix} 9 & 12 \\ 5 & 6 \end{pmatrix},$$

y en este caso $\mathbf{AB} \neq \mathbf{BA}$.

Las otras propiedades de multiplicación de matrices son parecidas a las propiedades de multiplicación de números reales. Específicamente, para matrices conformables $\mathbf{X}$, $\mathbf{Y}$, y $\mathbf{Z}$, y un escalar α:

Propiedades de Multiplicación de Matrices

1. Asociatividad: $(\mathbf{XY})\mathbf{Z} = \mathbf{X}(\mathbf{YZ})$
2. Distribuitividad por la izquierda: $\mathbf{X}(\mathbf{Y} + \mathbf{Z}) = \mathbf{XY} + \mathbf{XZ}$
3. Distribuitividad por la derecha: $(\mathbf{X} + \mathbf{Y})\mathbf{Z} = \mathbf{XZ} + \mathbf{YZ}$
4. Multiplicación Escalar: $\mathbf{X}\alpha\mathbf{Y} = \alpha\mathbf{XY}$
5. Existencia de un Elemento Neutro: si $\mathbf{X}$ es una matriz cuadrada $\mathbf{XI} = \mathbf{X}$ y $\mathbf{IX} = \mathbf{X}$

El elemento neutro en el ítem 5 se refiere a la matriz de identidad: una matriz cuadrada con valores 1 en la diagonal principal, y valores de 0 en cada otra posición. Describimos esta matriz con más detalle en sección 2.2.

Trasposición La traspuesta de una matriz $\mathbf{X}$ de $K \times N$, escrito $\mathbf{X}^\top$ o $\mathbf{X}'$ (en estos apuntos preferimos $\mathbf{X}'$), es una matriz de $N \times K$ donde $x_{nk} = x_{kn}$ para cada k y n. Las columnas de $\mathbf{X}$ se convierten en las filas de $\mathbf{X}'$, y visualmente se ve de la siguiente forma:

$$\mathbf{X} = \begin{pmatrix} 0 & 1 & 2 \\ 3 & 4 & 5 \end{pmatrix} \qquad \mathbf{X}' = \begin{pmatrix} 0 & 3 \\ 1 & 4 \\ 2 & 5 \end{pmatrix}.$$

Si la matriz $\mathbf{X}$ es cuadrada, su traspuesta se produce rotatndo la matriz alrededor de la diagonal principal.

Existen varias propiedades de traspuestas, resumidas en el siguiente listado, donde supongamos que $\mathbf{X}$ y $\mathbf{Y}$ son conformables, y α es un escalar.

Propiedades de Trasposición

1. $(\mathbf{X}')' = \mathbf{X}$
2. Traspuesta de un producto: $(\mathbf{XY})' = \mathbf{Y}'\mathbf{X}'$

3. Traspuesta de un producto extendido: $(\mathbf{XYZ})' = \mathbf{Z}'\mathbf{Y}'\mathbf{X}'$
4. $(\mathbf{X} + \mathbf{Y})' = \mathbf{X}' + \mathbf{Y}'$
5. $(\alpha\mathbf{X})' = \alpha\mathbf{X}'$

De nuevo, estas propiedades están presentadas sin demostración, y la demostración se deja como un ejercicio.

2.2 Matrices Importantes

Cuando revisamos las propiededes de multiplicación y sumatoria de matrices, conocimos dos matrices importantes. Estos son la matriz de identidad, $\mathbf{I_k}$, y la matriz nula, $\mathbf{0}$:

$$\mathbf{I_k} = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix} \qquad \mathbf{0} = \begin{pmatrix} 0 & 0 & \cdots & 0 \\ 0 & 0 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 0 \end{pmatrix}.$$

La matriz identidad es una matriz cuadrada de $k \times k$ con unos en la diagonal principal, y ceros en todas las demás posiciones. La matriz identidad es un tipo de matriz diagonal (una matriz cuadrada cuyas únicas elementos no-nulos aparecen en la diagonal principal), y un tipo de matriz escalar (una matriz diagonal con una única constante en cada posición del diagonal principal). La matriz nula es una matriz con ceros en cada posición. A diferencia de la matriz identidad, no es necesariamente una matriz cuadrada.

Una matriz cuadrada puede ser *triangular* si sólo tiene elemento nulos (i) arriba o (ii) abajo de la diagonal principal. En el primer caso la matriz se conoce como una matriz triangular inferior, y en el segundo caso la matriz se conoce como una matriz triangular superior. A continuación, se presenta la versión inferior ($\mathbf{L}$), y la versión superior ($\mathbf{U}$).

$$\mathbf{L} = \begin{pmatrix} l_{11} & 0 & \cdots & 0 \\ l_{21} & l_{22} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ l_{k1} & l_{k2} & \cdots & l_{kk} \end{pmatrix} \qquad \mathbf{U} = \begin{pmatrix} u_{11} & u_{12} & \cdots & u_{1k} \\ 0 & u_{22} & \cdots & u_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & u_{kk} \end{pmatrix} \tag{2.3}$$

Como veremos cuando hablamos de Invertibilidad y descomposiciones de Cholesky y QR en la sección 2.3, una propiedad muy conveniente de las matrices triangulares es que permiten soluciones muy simples a sistemas de ecuaciones. Por ejemplo, consideremos un sistema de ecuaciones del estilo $\mathbf{Lx} = \mathbf{b}$, donde $\mathbf{L}$ es una matriz triangular inferior (conocida), $\mathbf{b}$ es un vector de constantes

conocidos, y $\mathbf{x}$ es un vector de incógnitas. Entonces:

$$\begin{pmatrix} L_{11} & 0 & 0 \\ L_{21} & L_{22} & 0 \\ L_{31} & L_{32} & L_{33} \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix} = \begin{pmatrix} L_{11}x_1 \\ L_{21}x_1 + L_{22}x_2 \\ L_{31}x_1 + L_{32}x_2 + L_{33}x_3 \end{pmatrix} = \begin{pmatrix} b_1 \\ b_2 \\ b_3 \end{pmatrix}, \tag{2.4}$$

y existe una manera simple de resolver las incógnitas. La primera ecuación sólo incluye una incógnita x_1, y se puede encontrar su valor directamente. Después se sustituya el valor de x_1 en la segunda línea, para resolver la segunda incógnita x_2. Se sigue este proceso recursivo hasta resolver para todos los elementos del vector $\mathbf{x}$.

2.2.1 Formas Cuadráticas, y Formas Definidas

Cuando se habla de la forma cuadrática de una matriz $\mathbf{A}$ o su función cuadrática, se refiere a la forma:

$$Q = \mathbf{x}'\mathbf{A}\mathbf{x} = \sum_{i=1}^{N}\sum_{j=1}^{N} x_i x_j a_{ij}. \tag{2.5}$$

Aquí $\mathbf{A}$ es una matriz cuadrada (de $N \times N$), y $\mathbf{x} \in \mathbb{R}^N$ es un vector (de $N \times 1$). Formas cuadráticas de este estilo se encuentran frecuentemente en problemas de optimización, y en algunas clases, existen resultados muy elegantes acerca de la forma cuadrática. Específicamente, nos interesa saber si las formas definidas de estas matrices son definidas en alguna forma. Existen cuatro tipos de matrices definidas, que se definen a continuación:

1. Si $\mathbf{x}'\mathbf{A}\mathbf{x} > 0$ para cualquier candidato $\mathbf{x} \neq \mathbf{0}$, $\mathbf{A}$ se conoce como una matriz definida positiva
2. Si $\mathbf{x}'\mathbf{A}\mathbf{x} \geq 0$ para cualquier candidato $\mathbf{x}$, $\mathbf{A}$ se conoce como una matriz semi-definida positiva
3. Si $\mathbf{x}'\mathbf{A}\mathbf{x} < 0$ para cualquier candidato $\mathbf{x} \neq \mathbf{0}$, $\mathbf{A}$ se conoce como una matriz definida negativa
4. Si $\mathbf{x}'\mathbf{A}\mathbf{x} \leq 0$ para cualquier candidato $\mathbf{x}$, $\mathbf{A}$ se conoce como una matriz semi-definida negativa

Si $\mathbf{A}$ no cumple con ninguno de los puntos 1-4, la matriz se conoce como una matriz indefinida. La matriz definida positiva (y negativa) se parecen, de alguna forma, a los números reales positivos (negativos). Si se suman dos matrices definidas positivas (negativas), la matriz resultante tiene que ser positiva (negativa). Existen muchos resultados muy útiles si se sabe que una matriz es definida de cierta forma. Por ejemplo, una matriz definida positiva siempre es convexa, que tiene implicancias importantes para la optimización. Conoceremos otro resultado importante relacionado con matrices definidas positivas en la sección 2.3.4 cuando revisamos la inversión de matrices y la descomposición de Cholesky. En la econometría, existen muchas matrices importantes que son definidas de cierta forma, por ejemplo, la matriz de covarianza, que es una matriz definida positiva.

Notemos que aquí, aunque la matriz $\mathbf{A}$ es conocida en ecucación 2.5, los vectores $\mathbf{x}$ refieren a *cualquier* vector posible. A primera vista, podría parecer bastante difícil determinar el signo

de Q para cualquier vector $\mathbf{x}$, pero dada la forma cuadrática en que $\mathbf{x}$ entra la ecuación, no es tan restrictivo que parece. Como ejemplo, consideramos la matriz identidad $\mathbf{I_3}$, y cualquier vector no cero $\mathbf{x} = [x_1\ x_2\ x_3]'$. Se puede mostrar fácilmente que la matriz de identidad $\mathbf{I_3}$ (y cualquier matriz de identidad) es definida positiva, dado que:

$$Q = \mathbf{x'I_3x} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \end{bmatrix} = x_1^2 + x_2^2 + x_3^2,$$

y cada elemento $x_i^2 > 0\ \forall\ i \in 1, 2, 3 \Rightarrow Q > 0$.

2.3 El Inverso de Una Matriz

2.3.1 Definición y Uso de Inversión

Se dice que una matriz cuadrada de $N \times N$, $\mathbf{A}$ es invertible (o no singular, o no degenerada) si existe otra matriz $\mathbf{B}$ de tal forma que:

$$\mathbf{AB} = \mathbf{BA} = \mathbf{I_N}. \tag{2.6}$$

Si la ecucación 2.6 cumple, entonces $\mathbf{B}$ es única, y se conoce como el inverso de $\mathbf{A}$. El inverso se escribe como $\mathbf{A}^{-1}$. De la definición en 2.6, tenemos que $\mathbf{A^{-1}A} = \mathbf{AA^{-1}} = \mathbf{I}$, donde el tamaño de $\mathbf{I}$ es igual al tamaño de la matriz original $\mathbf{A}$ (y su inverso $\mathbf{A}^{-1}$).

Consideramos un sistema de ecuaciones lineales de la forma:

$$\mathbf{Ax} = \mathbf{b}. \tag{2.7}$$

Aquí, $\mathbf{A}$ es un matriz de $N \times N$ y $\mathbf{b}$ es un vector de $N \times 1$, ambas conocidas. El vector $\mathbf{x}$ (de $N \times 1$) contiene los valores desconocidos, que estamos buscando resolver en la ecuación. Si $\mathbf{A}$, $\mathbf{x}$, y $\mathbf{b}$ fuesen valores escalares, sería fácil encontrar la solución para $\mathbf{x}$, simplemente dividiendo ambos lados de 2.7 por $\mathbf{A}$. Sin embargo, para encontrar la solución a un sistema de ecuaciones matricial, necesitamos utilizar la idea del inverso de una matriz. Supongamos que existe una matríz $\mathbf{B}$ de la forma descrita en la ecuación 2.6. Entonces, podemos encontrar la solución $\mathbf{x}$ de la siguiente manera:

$$\mathbf{BAx} = \mathbf{Bb} \tag{2.8}$$

$$\mathbf{Ix} = \mathbf{Bb} \tag{2.9}$$

$$\mathbf{x} = \mathbf{Bb} \tag{2.10}$$

$$\mathbf{x} = \mathbf{A^{-1}b} \tag{2.11}$$

En la ecuación 2.8 pre-multiplicamos ambos lados de 2.6 por **B**. Por la naturaleza del inverso, $\mathbf{AB} = \mathbf{BA} = \mathbf{I}$, y $\mathbf{Ix} = \mathbf{x}$ ya que la matriz identidad es un elemento neutro. Por último, en 2.11 reemplazamos **B** por su notación común, llegando a la solución única para ecuación 2.7. Demostraciones de la unicidad del inverso (si una matriz es invertible) están disponibles en varios libros de econometría. Por ejemplo, una demostración de Greene (2002, §A.4.2) refiere a la matriz **A** y su inverso **B**. Ahora, supongamos que existe otro inverso **C**. En este caso:

$$(\mathbf{CA})\mathbf{B} = \mathbf{IB} = \mathbf{B} \tag{2.12}$$

$$\mathbf{C}(\mathbf{AB}) = \mathbf{CI} = \mathbf{C}, \tag{2.13}$$

y 2.12-2.13 son inconsistentes si **C** no es igual a **B**.

Algunas propiedades de los inversos de matrices simétricas e invertibles **A** y **B**, ambos del mismo tamaño de $N \times N$ y un escalar no igual a cero α son:

Propiedades de Inversión de Matrices

1. $(\mathbf{A}^{-1})^{-1} = \mathbf{A}$
2. $(\alpha\mathbf{A})^{-1} = \alpha^{-1}\mathbf{A}^{-1}$ para un escalar $\alpha \neq 0$
3. $(\mathbf{A}')^{-1} = (\mathbf{A}^{-1})'$
4. $(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}$
5. $(\mathbf{A} + \mathbf{B})^{-1} = \mathbf{A}^{-1}(\mathbf{A}^{-1} + \mathbf{B}^{-1})^{-1}\mathbf{B}^{-1}$

2.3.2 El Determinante de una Matriz

Para calcular algebraicamente el inverso de una matriz, se requiere calcular el determinante. El determinante de una matriz cuadrada **A** (el determinante solo existe para matrices cuadradas) se escribe $|\mathbf{A}|$, o $\det \mathbf{A}$. El determinante es un valor único para cada matriz, y captura el volumen de la matriz.Como veremos más adelante, el valor del determinante interviene en muchos resultados algebraicos. Por ejemplo, una matriz es invertible si y sólo si (*ssi*) su determinante no es igual a cero.

Computacionalmente, para encontrar el determinante se sigue un proceso recursivo para considerar sub-bloques en cada matriz. Sin embargo, existe una fórmula analítica para el determinante de una matriz. Siguiendo la definición de Hansen (2017, p. 460), el determinante de una matriz de $k \times k$ $|\mathbf{A}| = a_{ij}$ se puede calcular utilizando las permutaciones de $...\pi = (j_1, \ldots, j_k)$, que son todas las formas posibles para reorganizar los valores $1 \ldots, k$ (existen $k!$ posibles permutaciones). Y definimos como $\varepsilon_\pi = 1$ si la cantidad de inversiones en el orden de $1 \ldots, k$ para llegar a $\pi = (j_1, \ldots, j_k)$ es un número par, y como $\varepsilon_\pi = -1$ si la cantidad de inversiones es un número impar. Entonces, definimos a la determinante de una matriz **A** como:

$$|\mathbf{A}| = \sum_{\pi} \varepsilon_\pi a_{1j_1} a_{2j_2} \cdots a_{kj_k}. \tag{2.14}$$

Para la matriz $\mathbf{A}$, decimos que el menor para cada elemento a_{ij}, denotado M_{ij}, es el determinante de la matriz, una vez que hemos eliminido la fila i y la columna j de $\mathbf{A}$. Y definimos como al cofactor del mismo elemento $C_{ij} = (-1)^{i+j} M_{ij}$. Estos cofactores tienen un vínculo importante con el determinante de la matriz entero resumido en el Teorema de Laplace. Este teorema relaciona $|\mathbf{A}|$ con sus cofactores mediante la fórmula:

$$|\mathbf{A}| = \sum_{j=1}^{k} a_{ij} C_{ij},$$

y esta fórmula es conveniente para computar el determinante en pasos sucesivos para llegar a una matriz de 3×3 o 2×2. En el caso de una matriz de 2×2, se puede demostrar que (utilizando ecuación 2.14), que el determinante es:

$$\left| \begin{pmatrix} a & b \\ c & d \end{pmatrix} \right| = ad - bc.$$

2.3.3 Encontrando el Inverso de Una Matriz

Por lo general, al momento de encontrar el inverso de una matriz, se utiliza un algoritmo conocido (por ejemplo la eliminación de Gauss-Jordan) y un computador, aunque también es un proceso que se puede calcular 'a mano'. No revisaremos estos algoritmos aquí (si le interesa, una descripción está disponible en Simon and Blume (1994, §7.1).

Pero en casos de matrices pequeñas, es sencillo expresar la fórmula para el inverso. Por ejemplo, en el caso de una matriz $\mathbf{A}$ de 2×2, tenemos que:

$$\mathbf{A}^{-1} = \begin{pmatrix} a & b \\ c & d \end{pmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix} = \frac{1}{ad - bc} \begin{pmatrix} d & -b \\ -c & a \end{pmatrix}$$

y para un $\mathbf{A}$ de 3×3 el inverso se calcula como:

$$\mathbf{A}^{-1} = \begin{pmatrix} a & b & c \\ d & e & f \\ g & h & i \end{pmatrix}^{-1} = \frac{1}{|\mathbf{A}|} \begin{pmatrix} ei - fh & -(bi - ch) & bf - ce \\ -(di - fg) & ai - cg & -(af - cd) \\ dh - eg & -(ah - bg) & ae - bd \end{pmatrix}.$$

Adicionalmente, el determinante en el denominador se puede calcular utilizando la regla de Sarrus, que da el determinante de una matriz de 3×3 como $|\mathbf{A}| = aei + dhc + gbf - ceg - fha - ibd$. Por el momento no discutiremos los requisitos para saber si una matriz es invertible o no. Volveremos a examinar los requisitos de invertibilidad en la sección 2.4.2, después de introducir la idea de independencia en la sección 2.4.

2.3.4 ⋆La Descomposición de Cholesky

La descomposición de Cholesky es una descomposición de una matriz simétrica definida positiva ($\mathbf{A}$). La descomposición consiste en encontrar una matriz $\mathbf{L}$ (y su traspuesta $\mathbf{L}'$) de forma que:

$$\mathbf{A} = \mathbf{L}\mathbf{L}'.$$

Aquí $\mathbf{L}$ es una matriz triangular inferior como en la ecuación 2.3, y cada elemento de la diagonal principal de $\mathbf{L}$ es estrictamente positivo. Esta descomposición es particularmente útil[1] por su uso como una manera más (computacionalmente) eficiente de resolver sistemas de ecuaciones sin la necesidad de invertir una matriz. Para ver esto, partimos con una ecuación lineal de la misma forma que en 2.7. Calculamos la descomposición de Cholesky de la matriz definida positiva $\mathbf{A}$, dando $\mathbf{A} = \mathbf{L}\mathbf{L}'$. Con esto, se puede re-escribir ecuación 2.7 como:

$$\mathbf{L}\mathbf{L}'\mathbf{x} = \mathbf{b}. \tag{2.15}$$

Ahora, si definimos $\mathbf{z} = \mathbf{L}'\mathbf{x}$, finalmente se puede escribir 2.15 como:

$$\mathbf{L}\mathbf{z} = \mathbf{b}. \tag{2.16}$$

Es fácil resolver 2.16 para el desconocido $\mathbf{z}$ ya que $\mathbf{L}$ es una matriz triangular inferior, y una vez que se sabe $\mathbf{z}$, también podemos volver a $\mathbf{z} = \mathbf{L}'\mathbf{x}$ para encontrar el $\mathbf{x}$ desconocido (de interés) fácilmente, dado que $\mathbf{L}'$ es una matriz triangular superior. En ambos casos, el hecho de que las matrices en las ecuaciones son triangulares implica que se puede resolver la ecuación utilizando sustitución recursiva, un proceso simple y rápido para resolver un sistema de ecuaciones como revisamos en la ecucación 2.4.[2]

Para una matriz $\mathbf{A}$ definida positiva (y de $k \times k$), la descomposición de Cholesky es única. Las demostraciones de unicidad son inductivas. Para el caso de $k = 1$, simplemente se toma la raíz cuadrada (que es única). Para un k arbitrario, la demostración formal está disponible en Golub and Van Loan (1983, p. 88). En el texto de Hansen (2017, §A.14), se presenta una derivación muy intuitiva de la descomposición única cuando $k = 3$.

[1] En realidad, la descomposición de Cholesky es muy útil para varias razones, pero en este curso referimos principalmente a la descomposición cuando pensamos en una manera para resolver sistemas de ecuaciones en mínimos cuadrados ordinarios. Adicionalmente, la descomposición de Cholesky es de utilidad en simulaciones de números pseudo-aleatorios cuando se quiere simular múltiples variables correlacionadas. Con una base de una matriz de variables no correlacionadas $\mathbf{U}$, se puede simular una matriz de variables con una correlación deseada, $\mathbf{Z}$, utilizando la descomposición Cholesky de la matriz de covarianza $\square = \mathbf{L}\mathbf{L}'$, mediante $\mathbf{Z} = \mathbf{L}\mathbf{U}$.

[2] Refiere al programa `cholesky.do` para un ejemplo de la descomposición de Cholesky con datos simulados en Mata.

2.3.5 ⋆La Descomposición QR

La descomposición QR es otra descomposición en el estilo de Cholesky, donde se descompone la matriz $\boldsymbol{A}$ (simétrica, semi-definida positiva) como:

$$\mathbf{A} = \mathbf{QR}.$$

Este proceso descompone cualquier $\boldsymbol{A}$ invertible en dos matrices. La primera, $\boldsymbol{Q}$, es una matriz ortogonal, que implica que sus columnas y filas son vectores ortogonales unitarios, y por ende $\boldsymbol{Q}'\boldsymbol{Q} = \boldsymbol{I}$. La segunda matriz, $\boldsymbol{R}$, es una matriz triangular superior. Existen una serie de algoritmos comunes para realizar esta descomposición.

De nuevo, esta descomposición proporciona una manera eficiente para resolver sistemas de ecuaciones lineales como la ecuación 2.7. Para ver porqué, observamos:

$$\begin{aligned} \mathbf{Ax} &= \mathbf{b} \\ \mathbf{QRx} &= \mathbf{b} \\ \mathbf{Q'QRx} &= \mathbf{Q'b} \\ \mathbf{Rx} &= \mathbf{Q'b}, \end{aligned} \tag{2.17}$$

y notamos que la ecuación 2.17 se puede resolver simplemente para $\mathbf{x}$ utilizando sustitución recursive dado que $\mathbf{R}$ es una matriz triangular superior.

2.4 Independencia y Ortogonalidad (de vectores)

2.4.1 Independencia

Consideremos una serie de vectores $X = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_K\}$. Siempre se puede formar la matriz nula como una combinación lineal de los vectores de la forma:

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 + \ldots + \alpha_K\mathbf{x}_K = \mathbf{0}$$

si definimos a cada valor escalar α como 0. Decimos que X es linealmente independente si ésta es la *única* combinación de valores de α posible para formar el vector nulo. Es decir, la independencia lineal implica:

$$\sum_{k=1}^{K} \alpha_k\mathbf{x}_k = \mathbf{0} \qquad \Rightarrow \qquad \alpha_1 = \alpha_2 = \cdots = \alpha_K = 0, \tag{2.18}$$

donde el símbolo $\Rightarrow$ significa que si la primera ecuación cumple, entonces la segunda ecuación tiene que cumplir. De otra forma—si hay otras soluciones potenciales para el vector de valores α—se dice que X es linealmente dependiente.

La dependencia lineal implica que para un set de $k \geq 2$ vectores, por lo menos uno de los vectores se puede escribir como una combinación lineal de los otros vectores. Para un caso muy simple, consideremos los vectores $\mathbf{x}_1 = (2\ 1\ 3)'$ y $\mathbf{x}_2 = (-6\ -3\ -9)'$. Es fácil comprobar que los vectores $\mathbf{x}_1$ y $\mathbf{x}_2$ son linealmente dependientes, ya que $\mathbf{x}_2 = -3 \times \mathbf{x}_1$. En este caso, se puede llegar al vector $\mathbf{0} = (0\ 0\ 0)'$ a partir de la ecuación $\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2$ de varias formas:

$$\alpha_1\mathbf{x}_1 + \alpha_2\mathbf{x}_2 = 0\begin{pmatrix}2\\1\\3\end{pmatrix} + 0\begin{pmatrix}-6\\-3\\-9\end{pmatrix} = 3\begin{pmatrix}2\\1\\3\end{pmatrix} + 1\begin{pmatrix}-6\\-3\\-9\end{pmatrix} = \begin{pmatrix}0\\0\\0\end{pmatrix},$$

y por ende la matriz $\mathbf{X} = [\mathbf{x_1} \quad \mathbf{x_2}]$ es linealmente dependiente.

Independencia de Los Vectores de la Base Canónica Consideremos los vectores $\{\mathbf{e_1}, \ldots, \mathbf{e_K}\}$ donde cada $\mathbf{e_k}$ tiene cada elemento igual a 0, con la excepción de un valor de 1 en el elemento k. Estos vectores se conocen como la "base canónica" de $\mathbb{R}^K$, ya que con una combinación lineal de estos vectores de la forma:

$$\alpha_1\mathbf{e}_1 + \alpha_2\mathbf{e}_2 + \ldots + \alpha_K\mathbf{e}_K,$$

y valores elegidos para cada α_k, se puede producir cualquier vector posible $\in \mathbb{R}^K$.

Se puede demostrar que los K vectores de la base canónica son linealmente independientes en $\mathbb{R}^K$. Replicando Stachurski (2016, p. 18), definimos a los vectores canónicas $\{\mathbf{e_1}, \ldots, \mathbf{e_K}\}$, y una serie de coeficientes $\alpha_1, \ldots, \alpha_K$, de tal forma que $\sum_{k=i}^{K} \alpha_k\mathbf{e_k} = \mathbf{0}$. Ahora, esto implica:

$$\alpha_1\begin{pmatrix}1\\0\\\vdots\\0\end{pmatrix} + \alpha_2\begin{pmatrix}0\\1\\\vdots\\0\end{pmatrix} + \alpha_K\begin{pmatrix}0\\0\\\vdots\\1\end{pmatrix} = \begin{pmatrix}\alpha_1\\\alpha_2\\\vdots\\\alpha_K\end{pmatrix} = \begin{pmatrix}0\\0\\\vdots\\0\end{pmatrix}, \tag{2.19}$$

y $\alpha_k = 0$ para cada k, que es la única solución, cumpliendo con la definición de independencia lineal en la ecuación 2.18.

Independencia y Unicidad Anteriormente, vimos que la independencia implica por definición una sola solución para una ecuación lineal que forma el vector nulo (ecuación 2.18). En realidad, esta condición es mucho más generalizable y la independencia y unicidad están estrechemente vinculados. La independencia implica que la solución para *cualquier* variable $\mathbf{y}$ que es la suma de una ecuación de la forma $\sum_{k=1}^{K} \alpha_k\mathbf{x}_k$ es única (si la solución existe).

Específicamente, consideramos una colección de vectores en $\mathbb{R}^N$, $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_K\}$. Se puede demostrar que es equivalente decir: (a) $\mathbf{X}$ es linealmente independiente, y (b) que para cada $\mathbf{y} \in \mathbb{R}^N$,

existe como máximo un grupo de escalares $\alpha_1, \ldots, \alpha_K$ de tal forma que:

$$\mathbf{y} = \alpha_1 \mathbf{x}_1 + \cdots + \mathbf{x}_K. \tag{2.20}$$

La demostración formal de esta equivalencia se deja como un ejercicio.

Ejercicios:

1. Demuestra formalmente que las propiedades de sumatorias de matrices (asociatividad, conmutatividad, existencia de un elemento nulo) cumplen. Utiliza la notación definida en estos apuntes, y asegura definir cualquier otro tipo de notación necesario para hacer las demostraciones. También demuestra formalmente el cumplimiento de la propiedades de trasposición de matrices.
2. Demuestra formalmente la equivalencia entre los dos resultados descritos en la sección 2.4.1 (con ecuación 2.20). Nota que para demostrar equivalencia, es suficiente mostrar que (a) implica (b), y que (b) implica (a).

El Rango de una Matriz El rango de una matriz $\mathbf{A}$ de $N \times K$ (con $K \leq N$) es la cantidad de columnas linealmente independientes, y lo escribimos aquí como $rango(\mathbf{A})$. Se dice que $\mathbf{A}$ tiene rango completo si $rango(\mathbf{A}) = K$. Una matriz cuadrada de $K \times K$ es conocida como una matriz no-singular si tiene rango completo. De otra forma, se conoce como una matriz singular, que implicar que por lo menos dos de las columnas son linealmente dependientes.

2.4.2 La Relación entre Independencia e Invertibilidad

Cuando definimos el inverso de las matrices en la sección 2.3, nunca definimos los requisitos precisos para saber si una matriz es invertible. Ahora con la definición de independencia y el rango de una matriz, tenemos todos las detalles necesarios para introducir los requisitos de invertibilidad para cualquier matriz cuadrada $\mathbf{A}$.

Una matriz $\mathbf{A}$ de $K \times K$ es invertible si la matriz es linealmente independiente. Como hemos viston en las secciones anteriores, independencia lineal implica varias cosas:

1. La ecuación $\mathbf{Ax} = \mathbf{0}$ tiene una sola solución, de $\mathbf{x} = \mathbf{0}$
2. La ecuación $\mathbf{Ax} = \mathbf{b}$ tiene una sola solución para cada $\mathbf{b} \in \mathbb{R}^K$
3. La matriz tiene rango completo, $rango(\mathbf{A}) = k$
4. $|\mathbf{A}| \neq 0$
5. $\mathbf{A}$ es invertible (no-singular)
6. Existe una matriz $\mathbf{B}$ de $K \times K$ de tal forma que $\mathbf{AB} = \mathbf{I_K} = \mathbf{BA}$.
7. La traspuesta $\mathbf{A}'$ es invertible

Estos hechos son parte del teorema de inversión matricial. Y cada hecho es equivalente—si uno cumple, todos cumplen, y si uno no cumple, ninguno cumple—. Por lo tanto, mostrar que uno de estos hechos cumple basta para mostrar que la matriz es invertible.

2.4.3 Ortogonalidad de Vectores

La ortogonalidad es un concepto clave al momento de considerar relaciones entre variables, vectores, y matrices. Si $\mathbf{x}$ y $\mathbf{u}$ son dos vectores en $\mathbb{R}^N$, decimos que son vectores ortogonales si:

$$\langle \mathbf{x}, \mathbf{u} \rangle = 0 \tag{2.21}$$

donde la notación aquí sigue la definición en la ecuación 2.1. Esto también se puede escribir de la forma $\mathbf{x} \perp \mathbf{u}$. En dos dimensiones, la ortogonalidad implica que dos vectores son perpendiculares, o que cruzan para formar una intersección con angúlos internos de 90 grados. Volveremos a la ortogonalidad de vectores en mucho detalle cuando nos encontremos con la regresión lineal más tarde en el curso.

Clase Computacional: Una Introducción a Álgebra Lineal en Mata El programa computacional Stata tiene un sub-idioma llamado Mata. A diferencia de Stata, Mata está optimizado para trabajar con matrices y vectores. Adentro de Stata, se puede entrar a Stata escribiendo simplemente `mata`. En esta actividad, introducimos algunas funciones de Mata. Para una breve introducción a Mata, revise aquí. El manual de referencia de Mata está disponible en esta página y es la fuente comprensiva de información sobre Mata.

1. Simula una matriz d de 3×3 (por ejemplo utilizando el comando de Mata `uniform()`. Calcula el inverso utilizando un comando relevante de Mata. Calcula el inverso "a mano" (en Mata) utilizando la fórmula más extensiva de la sección 2.3.3.
2. Genere una matriz e compuesta por unos y una matriz columna f cuyo primer elemento sea el 11 y que cada elemento sea mayor en una unidad respecto al anterior, de tal forma que pueda calcular $(d + e)'$ y $(e \times f)-1$.
3. Calcule $(d + e)'$ y $(e \times f)-1$.
4. Utilice la base de datos precargada en Stata "'auto.dta"' e importe los datos del precio de los vehículos y su kilometraje a una matriz p. ¿Cuál es dimensión de p?

Ejercicios de Ayudantía:
Utilizando el archivo "'rendimiento2015.csv"', el cual contiene información acerca de las notas de todos los alumnos del país durante el año 2015, desarrolle los siguientes ejercicios:

1. Conociendo Stata:
   (a) ¿Cuántas observaciones tiene la muestra? ¿Cuántas variables han sido consideradas?
   (b) Para trabajar con mayor facilidad considere una submuestra que contenga las primeras 350.000 observaciones.
   (c) ¿El promedio general de los alumnos se encuentra en formato texto o numérico? Conviértala al formato deseado. Tabule y corrija los valores de ser necesario.
   (d) Obtenga la estadística descriptiva del promedio general de los alumnos e interprete. Genere un histograma de los promedios.
   (e) Tabule los valores que toma la variable de género de los alumnos y genere una nueva variable *hombre* que tome el valor 1 si el alumno es hombre y 0 si es mujer. ¿Qué tipo de variable es?
   (f) Obtenga la estadística descriptiva de la variable que indica el género de los alumnos e interprete.
2. Conociendo Mata:
   (a) Active "'Mata"' y luego desactívelo para volver a Stata. Active nuevamente, pero ahora ingrese el comando de Stata "'sum hombre"' antes de desactivar.
   (b) Genere una matriz de a de 3×3 cuyos elementos sean los números del 1 al 9, una matriz m de 3×3 cuyos elementos sean los números del 5 al 13 y una matriz z de unos de 3×2.
   (c) Genere una varible $constante$ que contenga sólo unos y una matriz x que contenga en cada columna las variables $constante$, $rural\_rbd$, la cual toma el valor 1 cuando el establecimiento es rural y 0 en caso que sea urbano, y $hombre$ ¿Cuáles serán las dimensiones de esta matriz x?
   (d) Sume y reste las matrices a y m y multiplique las matrices a y b.
   (e) Calcule la traspuesta y la inversa de la matriz a.
   (f) Genere una matriz y que contenga la variable $promedio\_gral$.
   (g) Genere una matriz b:
   $$b = (x'x)^{-1}(x'y) \tag{2.22}$$
   ¿Cuáles serán sus dimensiones? ¿Qué representan los elementos de la matriz b?
3. Repaso Herramientas Algebráicas:

(a) Consideremos una colección de vectores en $\mathbb{R}^N$, $X = (x_1, ..., x_K)$. Demuestre que:

X linealmente independiente $\Leftrightarrow$ para cada $y \in \mathbb{R}^N$, existe como máximo un grupo de escalares $a_1, ..., a_K$ tal que $y = \alpha_1 x_1 + ... + \alpha_K x_K$

(b) Calcule la inversa de la siguiente matriz $V = \left(\begin{smallmatrix} 8 & 2 \\ 2 & 4 \end{smallmatrix}\right)$. Utilice la fórmula:

$$V^{-1} = \frac{1}{|\ V\ |} \times adj(V^t)$$

Sección 3

Un Repaso de Herramientas Probabilísticas

Nota de Lectura: Se sugiere leer capítulo 1 de Goldberger (1991) para una introducción interesante. Una buena cobertura de todos los materiales presentados en esta sección está disponible en Casella and Berger (2002) capítulos 1-3. Sin embargo existen otros libros de probabilidad que contienen presentaciones apropiadas, como DeGroot and Schervish (2012). Una alternativa avanzada es Stachurski (2016), capítulos 8–10.

3.1 Elementos Básicos de Probabilidad

3.1.1 Una Introducción a la Probabilidad

La probabilidad es el estudio de la certidumbre que se puede asociar con eventos futuros inciertos. Hay varias concepciones de probabilidad, incluyendo una interpretación frecuentista (que considera la frecuencia relativa de distintos eventos para definir sus probabilidades), la interpretación clásica que parte de la base de igualdad de probabilidad de eventos, para así definir probabilidades iguales, y una interpretación subjetiva, que considera la probabilidad que una persona (en particular) asigna a los eventos. En estos apuntes no examinaremos la historia o filosofía detrás de la probabilidad, pero existen muchos libros de interés si le gustaría profundizar más en este tema. Una opción para partir es capítulo 1 de DeGroot and Schervish (2012) y sus referencias.

La teoría de probabilidad es la base de toda estadística, y por ende fundamental para nuestro estudio de econometría. Y la base de la probabilidad es la teoría de conjuntos. Partimos en esta sección introduciendo la noción de la teoría de conjuntos y otros aspectos fundamentales de probabilidad, antes de desarrollar algunas herramientas que serán fundamentales en nuestros modelos econométricos, como los estimadores, intervalos de confianza, y contrastes de hipótesis.

Algunas Definiciones Preliminares

Para poder introducir la notación básica de la teoría de conjuntos, primero definimos la idea de un **experimento** y un **evento**. Un experimento se define como cualquier proceso, verdadero o hipotético, en que se sabe con anterioridad todos los resultados potenciales (definición 1.3.1 de DeGroot and Schervish (2012)). Un experimento es aleatorio (o *estocástico*) si hay varios posibles resultados, y *deterministico* si sólo hay un resultado potencial. Un evento refiere a un conjunto bien definido de los resultados potenciales del experimento.

Esta definición de un 'experimento' y un 'evento' es muy amplia, y nos permite examinar muchos procesos de interés (con incertidumbre) en la econometría. Aunque un experimento puede ser tan simple como tirar un dado, la definición también incluye procesos como observar una persona para ver la educación total acumulada durante su vida, o su eventual salario laboral.

La Teoría de Conjuntos

La teoría de conjuntos sirve para una base de la teoría de probabilidad. La teoría de conjuntos es una herramienta de lógica que clasifica a elementos como perteneciente o no perteneciente a espacios—o conjuntos—particulares.

Volviendo a la idea de un experimento definido anteriormente, referimos al conjunto de todos los resultados potenciales de un experimento como el espacio muestral, o S. Por ejemplo, si el experimento consiste simplemente en lanzar un dado, definimos al espacio muestral $S \in \{1, 2, \ldots, 6\}$, y si el experimento consiste en observar el nivel de educación alcanzada en la vida de una persona elegido al azar de una población específica, el espacio muestral consistiría de $S \in \{\text{sin educación, básica, secundaria, terciaria}\}$.

Un evento, como lo definimos anteriormente, es cualquier resultado potencial del experimento de interés, y por ende es un sub-conjunto de S. Si definimos un evento A, decimos que A ocurre cuando el resultado aleatorio del experimento es contenido en el sub-conjunto A. Definimos contención de la siguiente forma:

$$A \subset B \Leftrightarrow x \in A \Rightarrow x \in B, \tag{3.1}$$

y definimos igualdad como:

$$A = B \Leftrightarrow A \subset B \text{ y } B \subset A. \tag{3.2}$$

Por último, definimos el conjunto vacío $\emptyset$ como el conjunto que consiste de ningún elemento. Trivialmente, $\emptyset \in A$ donde A es cualquier evento. Dado que el conjunto vacío no contiene ningún elemento, es cierto que todos los elementos que pertenecen a $\emptyset$ también pertenecen a A.

Los operaciones básicas de conjuntos se resumen a continuación (para dos conjuntos arbitrarios A y B), donde la nomenclatura sigue a Casella and Berger (2002, §1.1).

Operaciones Básicas de Conjuntos

1. Unión: Los elementos que pertenecen a A o a B. Se define $A \cup B = \{x : x \in A \text{ o } x \in B\}$
2. Intersección: Los elementos que pertenecen a A y a B. Se define: $A \cap B = \{x : x \in A \text{ y } x \in B\}$
3. Complemento: El complemento de A son todos los elementos que no pertenecen a A. Se define $A^c = \{x : x \notin A\}$.

Se dice que dos eventos son eventos disjuntos (o mutuamente excluyentes) si $A \cap B = \emptyset$.

La Teoría de Probabilidad

La teoría de probabilidad intenta asignar a cada evento en un experimento un valor para capturar la frecuencia del evento si el experimento fuese repetido muchas veces. La teoría de probabilidad define una serie de axiomas que caracterizan al número que captura esta frecuencia. En lo que sigue consideramos un evento A en el espacio S, y definimos a $P(A)$ como la probabilidad de que el evento A ocurre. Hay varias consideraciones técnicas acerca de exactamente cuáles son los subconjuntos de S sobre cuales se define una probabilidad, pero no las examinamos en este curso. Detalles más comprensivos están disponibles en Stachurski (2016) o Casella and Berger (2002, §1.2).

Los axiomas de probabilidad definen las propiedades para una función de probabilidad. Estos axiomas son:

Los Axiomas de Probabilidad

1. Para cada evento A, $P(A) \geq 0$.
2. $P(S) = 1$
3. Para cada secuencia infinita de eventos disjuntos $A_1, A_2, \ldots$, $P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i)$.

Los tres axiomas se refieren al evento arbitrario A, y al espacio muestral S. Entonces, el segundo axioma declara que si un evento es cierto (el resultado de un experimento por definición siempre cae en el espacio muestral), su probabilidad es igual a 1. El último axioma sigue la definición de unión (y eventos disjuntos) descrita anteriormente, pero aquí la unión es sobre una cantidad infinita de eventos.

A partir de los axiomas de probabilidad, se puede derivar una serie de otras propiedades de funciones de probabilidad. Resumimos algunas de estas propiedades aquí para una función de probabilidad P, y dos conjuntos A y B. Dejamos como un ejercicio la demostración de estas propiedades.

Otras Propiedades de Probabilidad

1. $P(\emptyset) = 0$
2. $P(A) \leq 1$
3. $P(A^c) = 1 - P(A)$
4. Si $A \subset B$, entonces $P(A) \leq P(B)$.

Notemos que con estos axiomas y propiedades adicionales, no hay una restricción acerca de exáctamente qué función de probabilidad P se debe elegir, sino que se define una serie de condiciones básicas para tener un P válido. Para cualquier espacio muestral existen muchas posibles funciones de probabilidad que cumplen con los axiomas 1-3. La definición de una función de probabilidad específica depende de la naturaleza del experimento bajo estudio, y una serie de otras propiedades que definimos a continuación.

Sin embargo, podemos definir una serie de condiciones que—si cumplen—aseguran que la función de probabilidad siempre satisface los axiomas 1-3. Consideramos un experimento con una cantidad finita de resultados potenciales. Esto implica que el espacio muestral S contiene una cantidad finita de puntos $s_1, \ldots, s_n$. En este tipo de experimento, para definir una función de probabilidad necesitamos asignar una probabilidad p_i a cada punto $s_i \in S$. Para asegurar que los axiomas de probabildad se satisfagan, los valores para $p_1, \ldots, p_n$ deben cumplir con las sugientes dos condiciones:

$$p_i \geq 0 \quad \text{para } i = 1, \ldots, n, \text{ y} \tag{3.3}$$

$$\sum_{i=1}^{n} p_i = 1. \tag{3.4}$$

Una demostración formal de este resultado está disponible en Casella and Berger (2002, Teorema 1.2.6).

Probabilidad Condicional

En nuestro análisis econométrico, a menudo cuando enfrentemos incertidumbre, contaremos con algo de información preliminar. Por ejemplo, nos podría interesar la probabilidad de que una persona tenga empleo condicional en el hecho de que la persona tiene una educación secundaria. En este caso, más que pensar en probabilidades incondicionales, vamos a querer hacer un análisis condicional.

Al contar con más información acerca de un experimento, tenemos que considerar un espacio muestral alterado. Para fijar ideas, imaginemos que nos interesa un evento A en un experimento con espacio muestral S. Ahora, imaginemos que aprendimos que ocurrió otro evento B, que reduce el espacio muestral en alguna forma. Ahora, dado que sabemos que B ocurrió, en vez de estar tratando de encontrar $P(A)$—la probabilidad incondicional de A—queremos encontrar $P(A|B)$, que es la probabildad condicioanl del evento A, dado el evento B. Computamos esta probabilidad condicional de la siguiente forma:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}. \tag{3.5}$$

Notemos dos cosas de esta ecuación. Primero, si la probabilidad que $P(B) = 0$, la probabilidad condicional no existe. Intuitivamente, esto no es muy problemático, dado que estamos condicio-

nando en el hecho de que B *ocurrió*, que implica que *ex-ante* hubo una probabilidad no nula que esto podría pasar. Y segundo, notamos que estamos buscando la unión de A y B. Esta probabilidad condicional pregunta cuál es la probabilidad de que A y B ocurrieran juntos, y lo divide por la probabilidad total de B, que de alguna forma es el espacio muestral reducido una vez que se sabe que B ha ocurrido.

Podemos invertir la ecuación 3.5 para calcular probabilidades de intersecciones entre eventos:

$$P(A \cap B) = P(B)P(A|B). \tag{3.6}$$

Este cálculo es particularmente útil en casos cuando la probabilidad condicional es fácil de calcular o asignar. Y, por simetria, si $P(A) > 0$, tenemos:

$$P(A \cap B) = P(A)P(B|A). \tag{3.7}$$

Esto se conoce como la ley de múltiplicación para probabilidades condicionales.

Otra ley importante que se basa en la probabilidad condicional es la ley de probabilidad total. Para introducir la ley de probabilidad total, necesitamos definir la idea de una partición. Una partición es una separación de un estado muestral en una serie de áreas mutuamente excluyentes, que cubren todo el espacio. Formalmente (DeGroot and Schervish, 2012, Definición 2.1.2), consideramos k eventos de un experimento (con espacio muestral S) llamados $B_1, \ldots, B_k$, de tal forma que $B_1, \ldots, B_k$ son disjuntos, y $\bigcup_{i=1}^{k} B_i = S$. Entonces, se dice que los eventos $B_1, \ldots, B_k$ forman una partición de S.

La *ley de probabilidad total* parte con una partición de eventos $B_1, \ldots, B_k$ del espacio S. Asumimos que $P(B_j) > 0$ para cada j. Entonces, para cada evento A en S:

$$P(A) = \sum_{j=1}^{k} P(A|B_j)P(B_j). \tag{3.8}$$

Examinamos algunos ejemplos como un ejercicios en clase.

Si combinamos los resultados anteriores (ecuaciones 3.6-3.8), llegamos al famoso Teorema de Bayes. Notemos que el lado izquierdo de ecucación 3.6 y 3.7 son idénticos, de tal forma que:

$$\begin{aligned} P(A)P(B|A) &= P(B)P(A|B) \\ P(B|A) &= \frac{P(B)P(A|B)}{P(A)} \end{aligned} \tag{3.9}$$

La ecuación 3.9 es la versión más simple del teorema de Bayes, que vincula la probabilidad condicional de un evento a su probabilidad incondicional, e información preliminar acerca de la probabilidad de la condición (revisamos algunos ejercicios aplicados). En el análisis Bayesiano, la

probabilidad de $B|A$ se conoce como la probabilidad posterior, ya que es la probabilidad de ocurrencia de B una vez que sabemos que A ocurrió. La probabilidad incondicional, o $P(B)$, se conoce como la probabilidad *a priori* en la ausencia de saber algo de A.

Sin embargo, también hay una versión del Teorema de Bayes para todos los eventos en una partición, y se llega a esta versión del teorema reemplazando el denominador de 3.9 por la ley de probabilidad total. Consideramos la partición $B_1, \ldots, B_k$ con $P(B_j) > 0$ para cada $j \in \{1, \ldots, k\}$, y el evento A con $P(A) > 0$. Entonces, para cada $i = 1, \ldots, k$, el Teorema de Bayes dice:

$$P(B_i|A) = \frac{P(B_i)P(A|B_i)}{\sum_{j=1}^{k} P(A|B_j)P(B_j)}. \tag{3.10}$$

Independencia

A veces, el haber aprendido que un evento A ha ocurrido no nos hace cambiar nuestra creencia acerca de la probabilidad de otro evento B. En estos casos decimos que A y B son eventos independientes. Trivialmente, consideramos un caso cuando lanzamos dos dados tradicionales, uno tras otro. Al observar un cierto resultado tras lanzar el primer dado, esto no nos entrega más información relevante acerca del resultados probable del segundo lanzamiento. En este caso, utilizando la notación anterior, tenemos que $P(B|A) = P(B)$.

Formalmente, dos eventos A y B son independientes si:

$$P(A \cap B) = P(A)P(B). \tag{3.11}$$

Para ver porqué, simplemente tenemos que volver a la ecuación 3.6 (o 3.7) y reemplazar la probabilidad condicional $P(B|A)$ con su probabilidad incondicional $P(B)$, dado que condicionar en A no cambia la probabilidad de ocurrencia. De la definición aquí, y las ecuaciones 3.6-3.7 se ve que dos eventos A y B son independientes si y sólo si (*ssi*) $P(A|B) = P(A)$, y $P(B|A) = P(B)$. Esta fórmula de independencia también se extiende para la independencia múltiple. En particular, para tres eventos A, B y C, decimos que son mutuamente independientes si:

$$P(A \cap B \cap C) = P(A)P(B)P(C). \tag{3.12}$$

3.1.2 Variables Aleatorias

Una variable aleatoria toma valores que son determinandos por un proceso aleatorio, que muchas veces es un proceso natural estocástico.[1] Una variable aleatoria es una representación númerica de

[1] Un proceso estocástico simplemente refiere a un proceso cuyo resultado no es conocido con certeza *ex-ante*. Es el opuesto a un proceso deterministico, que es un proceso que siempre produce el mismo resultado dado una condición inicial específica. Por ejemplo, lanzar una moneda es un proceso estocástico, y sumar dos números específicos es un proceso deterministo.

los resultados potenciales de un experimento, y formalmente es una función que mapea los resultados potenciales en el espacio muestral S a un número real $X \in \mathbb{R}$. La variable aleatoria captura la información contenida en todo el espacio muestral del experimento en una cantidad numérica—una manera conveniente para seguir con un análisis posterior.

El valor de un experimento aleatorio es, por definición, no conocido antes de realizar un experimento, pero dado que S es conocido, todos los valores potenciales de la variable aleatoria son conocidos. Posterior al experimento se observa el valor que la variable tomó en esta realización particular. A menudo, se refiere a una variable aleatoria por una letra mayúscula: X, y a las realizaciones específicas de la variable por una letra minúscula; x. Entonces, cuando se escribe $X = x$, o $X = 500$ (o la probabilidad $\mathrm{P}(X = x)$, o $\mathrm{P}(X = 500)$) es una combinación de la variable aleatoria con una realización específica de ella (x, o 500). Aquí, hemos pasado de hablar de la probabildad de observar una realización particular de una *variable aleatoria*, pero hasta ahora sólo sabemos que las funciones de probabilidad cumplen con los axiomas de probabilidad cuando consideramos los eventos en el espacio muestral original. Explícitamente, cuando hablamos de la probabilidad de observar un resultado particular, x_i, de una variable aleatoria X (con rango $\mathcal{X} = \{x_1, \ldots, x_k\}$), nos referimos a la función P_X:

$$P_X(X = x_i) = P(\{s_j \in S : X(s_j) = x_i\}).$$

Aquí, observamos el resultado x_i, *ssi* el resultado del experimento inicial era $s_j \in S$. Dado que P_X satisface los tres axiomas de probabilidad[2], P_X es una función de probabilidad, y podemos simplemente escribir $P()$ en vez de $P_X()$.

La definición específica de una variable aleatoria depende del experimento de interés, y el resultado particular de interés. Por ejemplo, si un experimento consiste en lanzar una moneda 25 veces, una variable aleatoria podría ser la cantidad total de 'caras' que salen, o podría ser la cantidad de lanzamientos hasta que salga una cara, o una variable binaria que toma el valor 1 si la cantidad de caras es mayor a 12, etc. Aquí se puede ver como una variable aleatoria puede resumir mucha información del espacio muestral subyacente. Consideremos el experimento de lanzar una moneda 25 veces y observar la cantidad de caras. El espacio muestral S consiste de 2^{25} elementos: cada uno un vector ordenado de 'caras' y 'sellos' de tamaño 25. Sin embargo, al definir una variable aleatoria $X =$ Cantidad total de caras, hemos reducido el espacio muestral a una serie de números enteros con rango $\mathcal{X} = \{0, 1, \ldots, 25\}$.

Las variables aleatorias pueden ser discretas, cuando toman una cantidad finita de posibles valores, o continuas, cuando pueden tomar infinitos posibles valores. En el caso de variables discretas, un caso especial es cuando toman solo dos valores (variables binarias) como el sexo, o si una per-

[2]Para ver esto, consideramos los tres axiomas. (1) Para cualquier evento A, $P_X(A) = P(\bigcup_{x_i \in A}\{s_j \in S : X(s_j) = x_i\}) \geq 0$ dado que $P()$ es una función de probabilidad; (2) $P_X(\mathcal{X}) = P(\bigcup_{i=1}^{k}\{s_j \in S : X(s_j) = x_i\}) = P(S) = 1$, y (3) si $A_1, A_2, \ldots$ son eventos disjuntos, $P_X(\bigcup_{m=1}^{\infty} A_m) = P(\bigcup_{m=1}^{\infty}\{\bigcup_{x_i \in A_m}\{s_j \in S : X(s_j) = x_i\}\}) = \sum_{m=1}^{\infty} P\{\bigcup_{x_i \in A_m}\{s_j \in S : X(s_j) = x_i\} = \sum_{m=1}^{\infty} P_X(A_m)$. Dado que los tres axiomas cumplen con P_X, se concluye que P_X es una función de probabildad válida.

sona está empleado o no, o pueden tomar más valores, por ejemplo edad en años. A pesar de tomar infinitos valores posibles, las variables continuas pueden ser limitadas en algún sentido, como por ejemplo el peso de un objeto, el cual no puede tomar valores negativos.

3.1.3 Esperanza, Momentos y Medidas Simples de Asociación

Existen una serie de valores particularmente interesante para caracterizar a una variable aleatoria. Una de ellos es la esperanza, o expectativa, de la variable aleatoria X. La expectativa es una medida del valor promedio de la variable aleatoria, o el valor esperado de un valor típico de la variable. Este promedio considera todos los valores posibles de X, ponderando por la frecuencia de ocurrencia del valor. La esperanza de una variable aleatoria X se denota $E(X)$, donde:

$$E(X) = \sum_{j=1}^{J} x_j P(X = x_j).$$

Aquí X tiene una cantidad finita de valores potenciales, y por lo tanto se puede sumar sobre todos los valores x_j. En el caso de una variable continua, definimos a la esperanza como:

$$E(X) = \int_{-\infty}^{\infty} x f(x) dx \tag{3.13}$$

donde aquí $f(x)$ refiere a la masa de probabilidad concentrada en un rango de x. Como veremos con más detalle en la sección 3.1.4, esta función $f()$ es conocida como la función de densidad de probabilidad.

Algunas Propiedades de la Esperanza

1. Si $Y = aX + b$ donde a y b son constantes, entonces $E[Y] = aE(X) + b$
2. Si $X = c$ con probabilidad 1, entonces $E(X) = c$
3. Si $X_1, \ldots, X_n$ son n variables aleatorias, entonces, $E(X_1 + \ldots + X_k) = E(X_1) + \ldots + E(X_k)$
4. Si $X_1, \ldots, X_n$ son n variables aleatorias independientes, entonces, $E\left(\prod_{i=1}^{n} X_i\right) = \prod_{i=1}^{n} E(X_i)$
5. Para funciones de variables aleatorias: $E[g(X)] = \sum_{j=1}^{k} g(x_j)P(X = x_j)$. Por lo general, $E[g(X)] \neq g(E(X))$. Una excepción son la clase de funciones lineales.

Las demostraciones de éstas propiedades se encuentran en DeGroot and Schervish (2012, §4.2).

Varianza

La esperanza proporciona *un* valor para caracterizar a una variable aleatoria. Sin embargo, esconde mucha información acerca de la variable aleatoria. Si queremos saber algo acerca del nivel de variablidad de una variable aleatoria alrededor de su expectativa, un otro valor característico importante es la varianza. La varianza de una variable aleatoria X se define como:

$$Var(X) = \sigma^2 = E[X - E(X)]^2,$$

o la distancia (en promedio) de todas las realizaciones de la variable aleatoria X de su valor esperado, al cuadardo. La desviación estandar también es una medida de dispersión, pero medida en la misma unidad que la variable original:

$$sd(X) = \sigma = \sqrt{(Var(X))}$$

Otra manera de representar la fórmula de varianza que utilizaremos a veces en estos apuntes es $Var(X) = E[X^2] - [E(X)]^2$. Para ver por qué notamos:

$$\begin{aligned} Var(X) &= E[X - E(X)]^2 \\ &= E(X^2 - 2E(X)X + E[X]^2) \\ &= E(X^2) - 2E(X)E(X) + E(X)^2 \\ &= E(X^2) - E(X)^2. \end{aligned} \tag{3.14}$$

Resumimos algunas propiedades de la varianza a continuación, para una variable aleatoria X, y dos constantes a y b.
Propiedades de la Varianza

1. Para cada X, Var$(X) \geq 0$
2. Var$(X) = 0$ ssi existe una constante c tal que $Pr(X = c) = 1$
3. Si $Y = aX + b$ donde a y b son constantes, entonces $Var[Y] = a^2 Var(X)$
4. Si $X_1, \ldots, X_n$ son n variables aleatorias *independientes*, entonces, $Var(X_1 + \ldots + X_k) = Var(X_1) + \ldots + Var(X_k)$

Nuevamente, las demostraciones formales de estas propiedades se encuentran en DeGroot and Schervish (2012), sección 4.3.

Los Momentos de una Variable Aleatoria

Para cada variable aleatoria X y cada número entero k, la esperanza $E(X^k)$ se conoce como el momento k-ésimo de X. La esperanza, como lo definimos anteriormente, es el primer momento de la distribución de X. El momento existe si $E(|X^k|) < \infty$. Ahora, supongamos que X es una variable aleatoria, y definimos a $E(X) = \mu$. Para cada número entero positivo k la esperanza $E[(X - \mu)^k]$ es conocido como el k-ésimo momento central de X. Siguendo esta notación, se observa que la varianza es el segundo momento central de X.

Los momentos sirven como medidas de distintas características de una variable aleatoria. Varios momentos o momentos centrales son cantidades muy conocidas, como por ejemplo la expecta-

tiva y la varianza. El tercer y cuarto momento central son conocidos como la asimetría estadística y la kurtosis, respectivamente. Como veremos más adelante, a veces vamos a definir nuestros estimadores en base a los momentos de la distribución observada en los datos. Lo llamaremos "método de los momentos" o "método de los momentos generalizados" (Sección 3.3.3 de los apuntes.).

Asociación: Covarianza y Correlación

Cuando trabajamos con más de una variable aleatoria, con frecuencia vamos a estar interesados en saber cómo se mueven en conjunto. La covarianza y la correlación son dos medidas simples de la asociación entre variables. Entre otras cosas, estas estadísticas son útiles para saber qué hace una variable aleatoria Y, si otra variable aleatoria X aumenta. Definimos la covarianza y la correlación ente dos variables X y Y a continuación:

$$\text{Cov}(X,Y) = E[(X - E(X))(Y - E(Y))] = E[XY] - E[X]E[Y]$$

$$\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{sd(X) \cdot sd(Y)}.$$

A veces la covarianza entre X y Y se denota σ_{xy}, y la correlación se denota ρ_{xy}.

La covarianza mide cómo las dos variables se mueven en conjunto, y está medida en las unidades de las dos variables de interés. Por ejemplo, si una variable X es medida en tiempo, y otra Y en peso, la covarianza estará expresada en tiempo×peso. A veces esto será una medida útil, pero muchas veces sería más útil tener una medida estandarizada. La versión estandarizada es la correlación, que tiene un rango de $[-1, 1]$, con -1 implicando una correlación negativa perfecta y 1 una correlación positiva perfecta. Si X e Y son independientes, $\text{Corr}(X,Y) = 0$ (pero el revés no es necesariamente cierto)[3].

Esperanzas Condicionales y Esperanzas Iteradas

Cuando consideramos múltiples variables aleatorias, también podemos considerar algunas propiedades de una variable *condicional* en ciertas carecterísticas de la otra variable. Un ejemplo de este tipo de proceso sería considerar el salario promedio de personas con educación secundaria. En términos generales, si dos variables aleatorias X e Y no son independientes, conocer algo acerca de X puede aportar información acerca de Y.

Sean X e Y dos v.a. La esperanza condicional de Y dado que $X = x$ se denota $E(Y|X = x)$, o simplemente como $E(Y|x)$. Para una variable Y continua, se define la esperanza condicional de la

[3]Para ver un caso simple en que dos variables no son independientes, pero sí tienen una correlación igual a cero, consideremos las variables Y y X, donde $Y = X^2$. En esta función cuadrática, Y claramente depende de X, pero la correlación entre las dos variables es igual a cero.

siguiente forma:

$$E[Y|X = x] = \sum_{j=1}^{m} y_j P_{Y|X}(Y = y_j|X = x).$$

Para una variable continua, definimos,

$$E[Y|x] = \int_{-\infty}^{\infty} y g_2(y|x) dy$$

donde utilizamos la misma idea de la ecuación 3.13 de una función que define la masa de probabilidad localizada en cada punto de y (para un valor dado de $X = x$), que se conoce como la función de densidad condicional. Definimos la función de densidad condicional de forma completa en la sección 3.1.4.

La Ley de las Esperanzas Iteradas (Ley de Probabilidad Total) Existe una clara relación entre la esperanza global y la esperanza condicional mediante la ley de las esperanzas iteradas, o la ley de probabildad total. La ley de esperanzas iteradas dice que la esperanza global se puede calcular tomando el promedio de todos los promedios condicionales! En notación, tenemos que:

$$E[Y] = E_X[E[Y|X]],$$

donde destacamos que la primera expectativa en el lado derecho es sobre X. Las otras dos expectativas son sobre Y, y la segunda expectativa en el lado derecho es una expectativa condicional. Con un X discreto tenemos, entonces:

$$E[Y] = \sum_{x_i} E[Y|X = x_i] \cdot Pr(X = x_i).$$

Esto implica que si ponderamos a las expectivas condicionales de X sobre Y para *todas* las valores posibles de X, volvemos a la expectativa global de Y. Para un ejemplo sencillo, si estamos interesados en el salario (Y) y la educación de las personas (X), una manera de calcular el salario promedio sería calcular el promedio de salario condicional en no tener educación; el salario promedio condicional en tener educación básica; en tener educación secundaria; y en tener educación terciaria, y después ponderamos todos estos promedios por la probabilidad de tener cada nivel de educación para llegar al salario promedio simple. Por supuesto, sería más directo simplemente calcular el salario promedio, pero la ley de esperanzas iteradas es un resultado fundamental que utilizaremos al momento de estar trabajando con análisis multivariado en econometría.[4] Revisaremos un ejemplo (sencillo) trabajado en clases.

[4]Una demostración simple de la ley de esperanzas iteradas está disponible en DeGroot and Schervish (2012, Theorem 2.1.4). Otra demostración está disponible en Casella and Berger (2002, Theorem 4.4.2).

3.1.4 Distribuciones

Una distribución para una variable X asigna probabilidades al evento en el que X cae en subespacios en $\mathbb{R}$. Cuando definimos probabilidades de realizaciones de variables aleatorias anteriormente, definimos la probabilidad de observar un resultado particular. Las distribuciones consideran probabilidades para un conjunto de valores de X.

La Funcion de Densidad Acumulada

Una clase importante de distribuciones de probabilidad son las funciones de densidad acumulada. La función de densidad acumulada (fda) de una variable aleatoria X, denotada F, es la función:

$$F(x) = Pr(X \leq x) \quad \text{para toda } x$$

y esta función está definida así tanto para variables continuas como para variables discretas. Una fda satisface las siguientes propiedades:

Propiedades de la Función de Densidad Acumulada

1. No decreciente: si $x_1 < x_2$, entonces $F(x_1) \leq F(x_2)$
2. Limites a $\pm\infty$: $\lim_{x\to-\infty} F(x) = 0$ y $\lim_{x\to\infty} F(x) = 1$
3. Continuidad desde la derecha: $\lim_{x\downarrow x_0} F(x) = F(x_0)$ en cada x_0

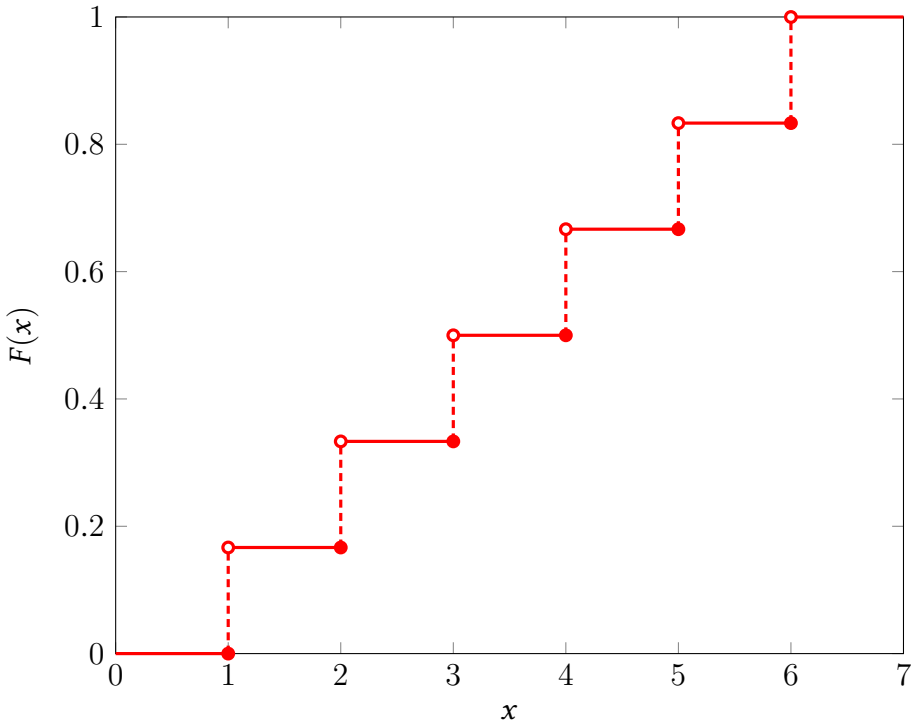
Consideremos la variable aleatoria que registra el valor que sale del lanzamiento de un dado. El ejemplo de la fda asociada con esta variable está presentada en la Figura 3.1. La probabildad de observar un valor de $x < 1$ es igual a 0. En la Figura 3.1 el eje horizontal está acotada entre 0 y 7, pero en realidad extiende entre $-\infty$ y ∞ (con $F(x) = 0$ cuando $x < 0$, y $F(x) = 1$ cuando $x > 6$). La probabildad acumulada de observar un número inferior a x salta a cada número entero, ya que existe una probabildad positiva de que salga este número al lanzar el dado.

Para considerar el caso de una fda de una variable continua, consideramos el caso de la distribución normal, o distribución Gaussiana. La distribución normal es potencialmente la distribución más frecuentamente encontrada en la econometría, reflejando su frecuencia en variables observadas en el mundo natural. La fda de una variable normal o Gaussiana tiene la siguiente distribución:

$$F(x) = \frac{1}{\sqrt{2\pi\sigma}} \int_{-\infty}^{x} \exp\left(-\frac{(t-\mu)}{2\sigma^2}\right) \mathrm{d}t \tag{3.15}$$

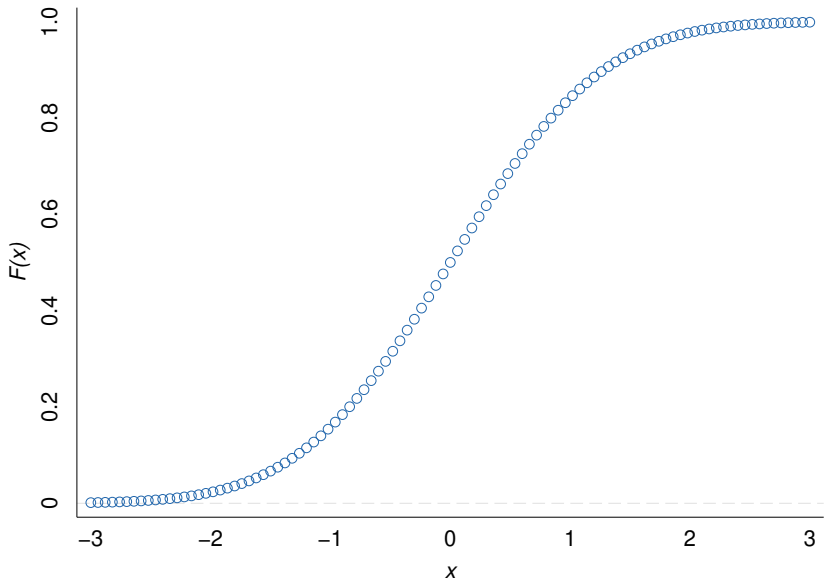
donde μ indica el promedio de la distribución, y σ su desviación estándar. Más adelante, vamos a llamar esta distribución como $\mathcal{N}(\mu, \sigma^2)$, y en el caso especial cuando $\mathcal{N}(0, 1)$, la distribución es conocida como la distribución normal estandarizada, y se denota Φ. Esta fda se grafica en la Figura 3.2. Nuevamente, aunque el eje horizontal está acotado (entre -3 y 3), su dominio es entre $-\infty$ y ∞, como fue estipulado en la propiedad 2 de la fda. Volvemos a examinar más distribuciones

Figure 3.1: Función de Densidad Acumulada de Una Variable Discreta



como un ejercicio computacional, y resumimos algunas propiedades básicas de las distribuciones importantes en la Tabla 3.1.

Figure 3.2: La Distribución Normal Estandarizada – fda



Funciones de Densidad

Las funciones de densidad presentan las probabilidades puntuales de observar un resultado x en el rango de X. La manera en que se describen estas funciones depende de si la variable subyacente X es discreta o continua. En el caso de una variable discreta, son conocidas como una función de probabilidad y en el caso de una variable continua, son conocidas como funciones de densidad de probabilidad. Examinamos ambos tipos de distribuciones (y variables) a continuación

Distribuciones Discretas Una variable aleatoria X tiene una *distribución discreta* si X sólo puede tomar un número finito de valores distintos $x_1, \ldots, x_k$, o una secuencia infinitia de valores distintos $x_1, x_2, \ldots$. Si una variable aleatoria X tiene una distribución discreta, se define la función de probabilidad, f, como la función que para cada número real x:

$$f(x) = P(X = x)$$

con el soporte $\{x : f(x) > 0\}$. Esta función $f(x)$ presenta la masa de probabildad asociada con cada punto x. Las propiedades de la función de probabilidad (y los axiomas de probabilidad) implican (i) que si x no es uno de los valores posibles de X entonces $f(x) = 0$. Y (ii) si la secuencia $x_1, x_2, \ldots$ contiene todos los valores posibles de X entonces $\sum_{i=1}^{\infty} f(x_i) = 1$.

Un ejemplo simple de una variable discreta y su función de probabilidad asociada es una variable aleatoria Z que sólo toma dos valores: 0 y 1, con $P(Z = 1) = p$. Esta variable tiene una distribución Bernoulli con parámetro p, que se caracteriza de la siguiente forma:

$$f(z; p) = \begin{cases} p & \text{if } z = 1 \\ 1 - p & \text{if } z = 0. \end{cases}$$

En esta función se escribe $f(z; p)$ para denotar que la función depende de la realización de la variable aleatoria z, pero también del parámetro p. Con frecuencia omitimos los parámetros de la definición de la función cuando es claro que refieren a parámetros que dependen del contexto del experimento. Otro ejemplo de una función de probabilidad corresponde a la Distribución Uniforme en Números Enteros. Sean $a \leq b$ dos números enteros. Supongamos que es igualmente probable que el valor de la variable aleatoria X toma el número entero $a, \ldots, b$. Entonces decimos que X tiene una distribución uniforme en los número enteros $a, \ldots, b$, y se escribe la densidad:

$$f(x) = \begin{cases} \frac{1}{b-a+1} & \text{si } x = a, \ldots, b \\ 0 & \text{si no.} \end{cases}$$

En ambos casos, es fácil confirmar que las dos propiedades descritas anteriormente en (i) y (ii) cumplen con la función de probabilidad definida.

Distribuciones Continuas En el caso de una variable continua, la probabilidad de observar cualquier realización *particular* de X es igual a cero (ver Casella and Berger (2002, p. 35)), y por lo tanto tenemos que definir la distribución utilizando integrales en vez de probabilidades puntuales. En este caso, definimos a la función de densidad de probabilidad de una variable aleatoria X de tal forma que:

$$Pr(a \leq X \leq b) = \int_a^b f(x)dx$$

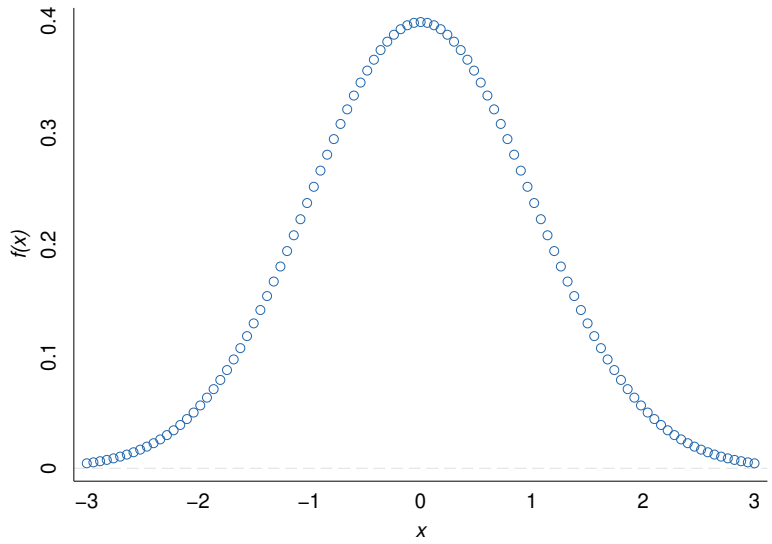
para cada intervalo cerrado $[a, b]$. Si una variable aleatoria X tiene una distribución continua, la función f se llama la función de densidad de probabilidad (fdp) de X, y el conjunto tiene el soporte $\{x : f(x) > 0\}$. En el caso de una variable continua, la fdp tiene que cumplir con las siguientes dos condiciones:
Propiedades de una Función de Densidad de Probabilidad

1. $f(x) \geq 0$ para toda x,
2. $\int_{-\infty}^{\infty} f(x)dx = 1$.

La función de densidad de probabilidad de una variable normal $\mathcal{N}(\mu, \sigma^2)$ está presentada en la Figura 3.3. En este caso, $\mu = 0$ y $\sigma = 1$, que es la densidad de la normal estandarizada, denotada ϕ. Para cualquier valor μ y σ, esta fdp se escribe:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} \exp\left(-\frac{(x-\mu)}{2\sigma^2}\right). \tag{3.16}$$

Figure 3.3: La Función de Densidad de Probabilidad de la Normal Estandarizada



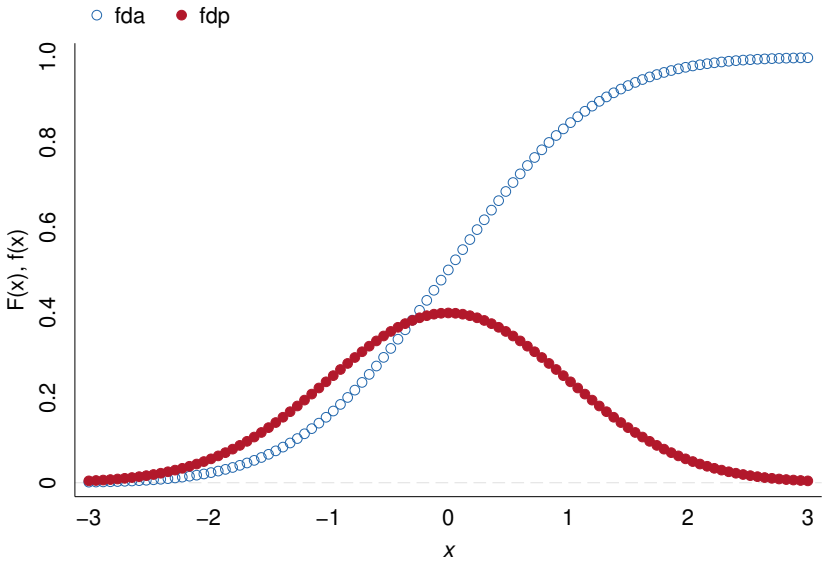
La Relación Entre la fda y la fdp Hay un vínculo muy estrecho y aparente entre las funciones de densidad acumulada, y las funciones de densidad de probabilidad (o funciones de densidad). Dado

que las fdp demuestran la masa de probabilidad en puntos de x, y las fda la probabilidad acumulada *hasta* cada punto x, se llega a la fda integrando sobre la fdp, y se llega la fdp diferenciando la fda:

$$\frac{dF(x)}{dx} = f(x) \tag{3.17}$$

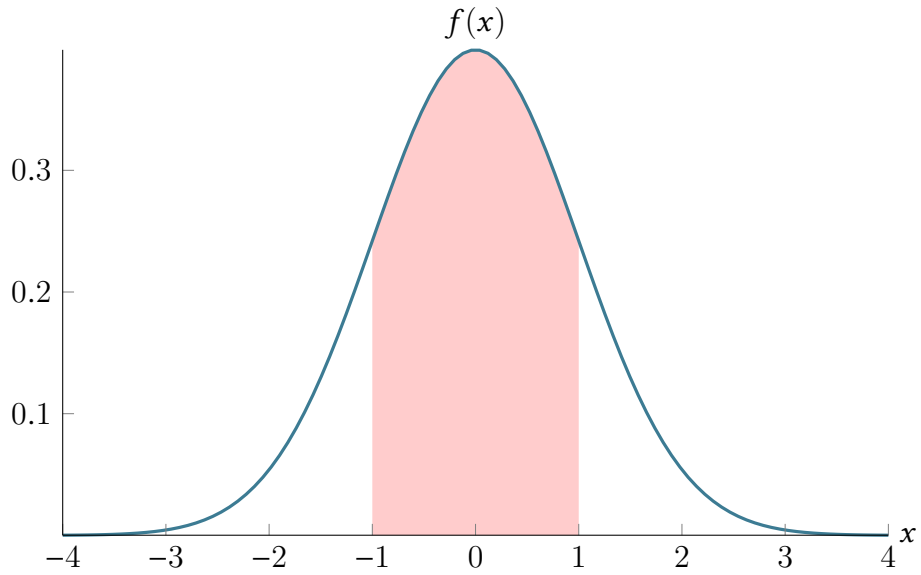
Para ver un caso particular, consideramos la fda y la fdp de la distribución normal estandardizada. Se presentan las dos funciones en la Figura 3.4. La utilidad de esta relación es aparente cuando se quiere responder a preguntas como ¿cuál es la probabildad de observar un valor de $X < 1$?; ¿cuál es la probabildiad de observar un valor de $X \geq 2$?; o ¿cuál es la probabilidad de observar un valor $-1 < X \leq 1$? Estas preguntas refieren a la masa de probabilidad baja la función de densidad de probabilidad (refiere a la Figura 3.5 para la visualización gráfica de la tercera pregunta). Sin embargo, la respuesta se obtiene sencillamente a partir de función de densidad *acumulada*. En el primer caso, o más generalmente casos cuando se quiere calcular $P(X < a)$ para algún escalar a, simplemente buscamos $F(a)$, que es justamente toda la masa de probabilidad inferior a a. En el segundo caso, o de nuevo, más generalmente casos cuando se quiere calcular $P(X > b)$ para un valor escalar b, se requiere calcular $(1 - F(b))$, donde utilizamos el hecho que $\lim_{x\to\infty} F(x) = 1$. Por último, consideramos un caso de querer calcular $P(a < X \leq b)$ para cualquier dos valores escalares a y b con $a < b$. Aquí, calculamos la probabilidad como: $F(b) - F(a)$. Notamos de la figura 3.5 (con $b = 1$ y $a = -1$), que aquí estamos simplemente calculando toda la masa acumulada hasta el valor $b = 1$, y después restando la masa de probabilidad entre $-\infty$ y a.

Figure 3.4: La fda y la fdp de la Distribución Normal Estandarizada



Para los cálculos anteriores, hemos considerado valores como $F(a)$, que son valores específicos de la función de densidad acumulada. Aunque sabemos algunas condiciones básicas de las fda (como por ejemplo que son limitado por 0 y 1), no es necesariamente trivial calcular un valor como

Figure 3.5: Area Bajo la Curva en una fdp



$F(a)$ para alguna función de densidad acumulada específica. Por ejemplo, en el caso de la fda de la distribución normal, no hay una solución de forma cerrada para calcular el valor de la función. Una manera de calcular el valor de la fda sería algún tipo de integración númerica sobre la fdp, por ejemplo utilizando la Cuadratura de Gauss. Típicamente, los lenguajes de computación estadística proveen fórmulas fáciles para calcular valores específicos de funciones de densidad acumulada. Y en algunos casos muy comunes, por ejemplo para la normal estandarizada, a menudo se ven tablas estadísticas que resumen los valores de la fda en muchos puntos de la distribución. Por ejemplo, en la Tabla 6.1 de estos apuntes, resumimos los valores de la fda de la normal estandarizada en muchos puntos (positivos). Dado que la fdp es una distribución simétrica, se puede inferir los valores de $F(x)$ para un $x < 0$. Esta tabla demuestra (por ejemplo) que 50% de la masa de probabilidad cae abajo del valor de 0.00, y 97.5% de la masa cae abajo de 1.96. Dejamos como ejercicios la revisión del cálculo de varios valores específicos, y una extensión a normales con media y desviación estándar distintas a 0 y 1.

La Función Cuantil Se utiliza la función de densidad acumulada de esta forma para responder a preguntas de la probabilidad de observar un valor de X superior, o inferior a cierto punto de interés. Pero también hay casos en que nos interesa invertir la pregunta, y saber el valor exacto de x donde la probabilidad de ocurrencia es igual a algún valor p. Por esto, utilizamos la función cuantil. Sea X una variable aleatoria con una fda F. Para cada p estrictamente entre 0 y 1, definamos $F^{-1}(p)$ como el menor valor de x tal que $F(x) \geq p$. Entonces, $F^{-1}(p)$ es el cuantil p de X, y la función está definido sobre el intervalo (0,1). Esta función F^{-1} es la función cuantil, y como la fda, generalmente no tiene una solución de forma cerrada. De igual modo que la fda, la manera más simple para calcular un valor específico (para una distribución particular) de la función cuantil es utilizando un rutina estadística de un programa computaciónal. Y también se puede encontar el valor de

una función cuantil para la distirbución normal estandarizada a partir de tablas como la Tabla 6.1. Por ejemplo, imaginamos que estamos interesados en saber el valor mínimo de la distribución normal abajo de donde cae 97.5% de la masa de probabilidad. En la Tabla 6.1 observamos que $F^{-}1(0.975) = 1.96$ para una variable normal estandarizada.

Distribuciones Bivariadas

En situaciones cuando trabajamos con dos variables, además de considerar medidas de asociación común entre variables (como la correlación y la covarianza), podemos considerar toda la distribución conjunta de ambas variables. La distribución bivariada considera el soporte de dos variables en conjunto. Para definir la distribución bivariada, consideramos dos variables aleatorias, X e Y. La distribución bivariada es la colección de probabilidades de la forma $\mathrm{P}[(X, Y) \in C]$ para todos los pares de números reales tal que $\{(X, Y) \in C\}$ es un evento. También como las funciones de distribución univariables, las distribuciones bivariadas pueden ser discretas o continuas (o mezcladas cuando una variable es discreta y otra es continua).

La función de probabilidad bivariada de X e Y, está definida como la función f tal como para cada punto (x, y) en el plano xy,:

$$f(x, y) = Pr(X = x \text{ y } Y = y),$$

que cumple las siguientes condiciones:

1. Si (x, y) no es uno de los posibles valores de los pares (X, Y), entonces $f(x, y) = 0$
2. $\sum_{Cada(x,y)} f(x, y) = 1$

Y las dos variables aleatorias X e Y tienen una distribución bivariada continua si existe una función no negativa f definida en todo el plano xy tal como para cada subconjunto C en el plano:

$$Pr[(X, Y) \in C] = \int_C \int f(x, y) dx\, dy$$

Aquí la función f se llama la función de densidad de probabilidad bivariada cuyas condiciones son:

1. $f(x, y) \geq 0$ para $-\infty < x < \infty$ y $-\infty < y < \infty$
2. $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f(x, y) dx\, dy = 1$

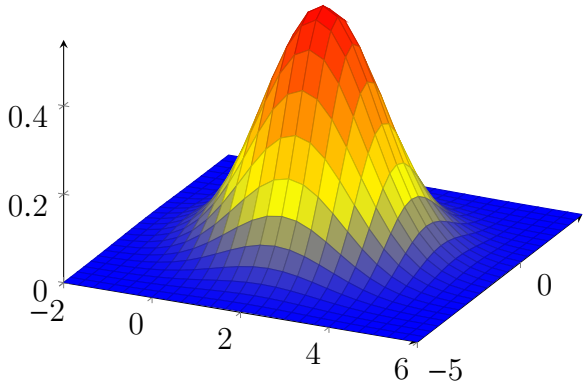
Un ejemplo sencillo de una fdp bivariada contínua es la distribución normal bivariada, representada en la Figura 3.6. En forma general, la fdp de la Normal bivariada tiene la siguente fórmula, caracterizada por el promedio de cada variable X e Y, sus desviaciones estándares respectivas, y la

correlación entre X e Y:

$$f(x,y) = \frac{1}{2\pi\sigma_x\sigma_y\sqrt{1-\rho^2}} \exp\left(-\frac{1}{2(1-\rho^2)}\left[\frac{(x-\mu_x)^2}{\sigma_x^2} + \frac{(y-\mu_y)^2}{\sigma_y^2} - \frac{2\rho(x-\mu_x)(y-\mu_y)}{\sigma_x\sigma_y}\right]\right).$$

El ejemplo a continuación (Figura 3.6) representa dos variables normales, uno con promedio 2 y otro con promedio 1, ambos con desviación estándar igual a uno, y con una correlación $\rho = 0$.

Figure 3.6: Función de Densidad de Probabilidad Normal Bivariada



Distribuciones Condicionales

Las distribuciones condicionales describen las probabilidades asociadas con una variable aleatoria, condicionando en eventos determinados por otras variables aleatorias. Después de observar los resultados de una (o varias) variables aleatorias, nos gustaría poder actualizar las probabildades asociadas con variables que aún no hemos observado. Por ejemplo, sabiendo que alguien estudio una carrera universitaria nos entrega información relevante para la distribución de su salario laboral. O saber que una empresa tiene 10 trabajadores probablemente impacte la distribución posible de su producción total.

Supongamos que X e Y son dos variables aleatorias con una distribución bivariada cuya función de probabilidad es f. Ahora, f_1 y f_2 son las funciones de probabilidad marginales (individuales) de x e y. Si observamos que $Y = y$, la probabilidad de que una variable aleatoria X tome un valor específico x está dado por la probabildad condicional:

$$\begin{aligned} Pr(X = x|Y = y) &= \frac{Pr(X = x \text{ y } Y = y)}{Pr(Y = y)} \\ &= \frac{f(x,y)}{f_2(y)} \end{aligned}$$

Ahora, en vez de una probabilidad para un $X = x$, podemos escribir la función de probabilidad

condicional entera:

$$g_1(x|y) = \frac{f(x,y)}{f_2(y)}$$

y la llamamos "la distribución condicional de X dado que $Y = y$."

Sean X e Y dos variables aleatorias con una distribución bivariada cuya función de probabilidad es f y con funciones de probabilidad marginal f_1. Sea y un valor para el cual $f_2(y) > 0$. Entonces, definimos la fdp condicional de X dado que $Y = y$ como:

$$g_1(x|y) = \frac{f(x,y)}{f_2(y)} \text{ para } -\infty < x < \infty$$

De la misma forma cuando trabajamos con probabilidades condicionales e independencia de variables en la ecuación 3.11, podemos hablar de independencia entre distribuciones enteras. Específicamente, tenemos que dos variables aleatorias X e Y con una fdp bivariada $f(x,y)$ son independientes *ssi* para cada valor de y con $f_2(y) > 0$ y cada valor de x:

$$g_1(x|y) = f_1(x).$$

Algunas Distribuciones de Interés Relacionadas con la Distribución Normal

Para cerrar nuestra discusión de distribuciones de probabilidad, introducimos una serie de distribuciones conocidas que nos serán de utilidad durante este curso. Estas incluyen la distribución log-normal, la distribucíon chi-cuadrada (o χ^2), la distribución t de Student, y la distribución F.

La Distribución log-Normal El uso de logaritmos para modelar variables es común (crecimiento, escala Richter, ...). Nos permite hablar en términos de cambios constantes (eg un log(PIB per capita) de 4.77 (Sweden) es 10 veces mayor que el de Vietnam (3.77) que es 10 veces mayor que el de la República Centroafricana). Los logaritmos tienen un soporte sobre $\mathbb{R}^+$. Decimos que si $\log(X)$ tiene una distribución normal, entonces la variable no transformada X es log-normal. En la Figura 3.7, illustramos una distribución empírica que parece ser log-Normal.

Si $\log(X)$ tiene una distribución normal con media μ ($-\infty < \mu < \infty$) y varianza σ^2 ($\sigma > 0$) decimos que X tiene una distribución log-normal.

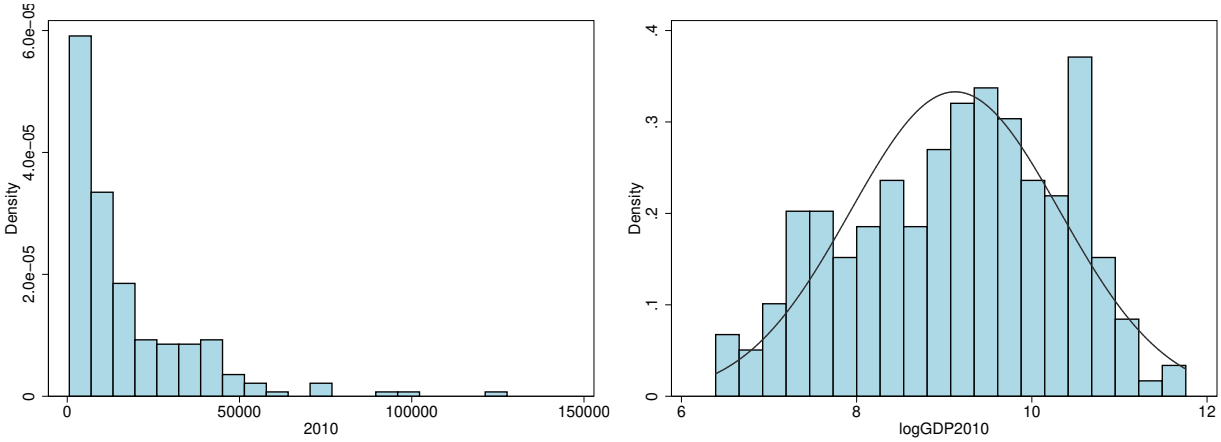
$$f(\log(x); \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp\left[-\frac{1}{2}\left(\frac{\log(x) - \mu}{\sigma}\right)^2\right] \tag{3.18}$$

Esta distribución tiene una FDA de la siguiente forma:

$$F(x) = \Phi([log(x) - \mu]/\sigma)$$

Figure 3.7: PIB per Capita (ajustado por poder de paridad de compra)

(a) PIB per Capita (2010) (b) log(PIB per Capita)



y una función cuantil:

$$F^{-1}(p) = exp(\mu + \sigma\Phi^{-1}(p))$$

La Distribución Chi-cuadrado (χ^2) Sumando N variables normales (iid) al cuadrado resulta en una distribución χ^2 con N grados de libertad. Decimos que si $\{Y_i\}_{i=1}^N$ son $\mathcal{N}(0,1)$, iid. Entonces, $X = \sum_{i=1}^N Y_i^2 \sim \chi_N^2$. Esta distribución tiene un soporte sobre $\mathbb{R}^+$.

La Distribución t de Student La distribución t de "Student" (ver Figura 3.8) es la distribución que describe una variable aleatoria formada por la ratio de una variable aleatoria normal, y la raíz cuadrada de una variable Chi-cuadrado. Si $Y \sim \mathcal{N}(0,1)$, $X \sim \chi_N^2$, y Y, X son independientes, entonces

$$T = \frac{Y}{\sqrt{(X/N)}} \sim t_N$$

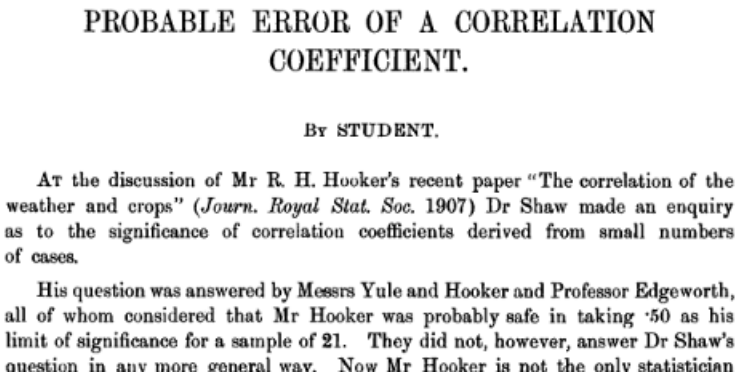
Esta distribución nos sirve para inferencia estadística en muestras pequeñas. Cuando $N \to \infty$, $t \to \mathcal{N}(0,1)$.

La Distribución F de Snedcor La distribución de Fisher con N_1 grados de libertad en el numerador y N_2 grados de libertad en el denominador es el ratio de dos variables distribuidas como un χ^2, divididas por sus respectivas grados de libertad. si $X_1 \sim \chi_{N_1}^2$ y $X_2 \sim \chi_{N_2}^2$, y X_1 y X_2 son independientes. entonces,

$$F = \frac{X_1/N_1}{X_2/N_2} \sim F(N_1, N_2)$$

Nos encontraremos con esta distribución cuando implementamos contrastes de hipótesis múltimples en la segunda mitad del curso.

Figure 3.8: "Student" Biometrika, 1908.



PROBABLE ERROR OF A CORRELATION COEFFICIENT.

By STUDENT.

At the discussion of Mr R. H. Hooker's recent paper "The correlation of the weather and crops" (*Journ. Royal Stat. Soc.* 1907) Dr Shaw made an enquiry as to the significance of correlation coefficients derived from small numbers of cases.

His question was answered by Messrs Yule and Hooker and Professor Edgeworth, all of whom considered that Mr Hooker was probably safe in taking ·50 as his limit of significance for a sample of 21. They did not, however, answer Dr Shaw's question in any more general way. Now Mr Hooker is not the only statistician

La Normal Mulitvariada : La normal multivariada es una generalización de la distribución normal y la distribución normal bivariada ya introducidas en este capítulo. La normal multivariada (de N variables) se escribe como $y \sim \mathcal{N}(\mu, \Sigma)$ donde y y μ son vectores de $N \times 1$ y Σ es una matriz de $N \times N$. Definimos la distribución marginal y condicional de la distribución normal multivariada a partir de la partición:

$$y = \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \mu = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix} \text{ y } \Sigma = \begin{pmatrix} \Sigma_{11}\Sigma_{12} \\ \Sigma_{21}\Sigma_{22} \end{pmatrix}$$

donde y_1 y μ_1 son vectores de $N_1 \times 1$, y_2 y μ_2 son vectores de $N_2 \times 1$, $N_1 + N_2 = N$, Σ_{11} es una matriz de $N_1 \times N_1$, Σ_{22} es una matriz de $N_2 \times N_2$, Σ_{12} es una matriz de $N_1 \times N_2$ y $\Sigma_{21} = \Sigma'_{12}$ es una matriz de $N_2 \times N_1$. En este caso, la distribución marginal de y_1 es:

$$y_1 \sim \mathcal{N}(\mu_1, \Sigma_{11})$$

y la distribución condicional de $y_2|y_1$ es:

$$y_2|y_1 \sim \mathcal{N}(a + B'y_1, \Sigma_{22} - B'\Sigma_{11}B)$$

con $B = (\Sigma_{11})^{-1}\Sigma_{12}$ y $a = \mu_2 - B'\mu_1$.

Una correlación de cero implica que hay independencia: si $\Sigma_{12} = 0$, entonces y_1 y y_2 son independientes. Por lo general, esto *no* es cierto para dos variables aleatorias y_1 y y_2, pero *sí* cumple si las variables son normalmente distribuidas.

Propiedades de Funciones de un Vector Normal Estándar El vector aleatorio $N \times 1$ $z \sim \mathcal{N}(0, I)$ donde I es una matriz de identidad de $N \times N$ se conoce como un vector Normal Estándar, cuyos elementos $z_i \sim \mathcal{N}(0, 1)$ para $i = 1, 2, \ldots, N$ son variables Normales estandares **independientes**.

La densidad de z

$$f(z) = (2\pi)^{\frac{-N}{2}} \exp\left(\frac{-z'z}{2}\right) = \prod_{i=1}^{N}\left[\left(\frac{1}{\sqrt{2\pi}}\right)\exp\left(\frac{-z_i^2}{2}\right)\right]$$

es el producto de N densidades Normales estandares. La forma cuadrática de un vector aleatoria normal refiere a la forma $z\prime Wz$, donde W es una matriz conformable. Algunas propiedades de la forma cuadrática de vectores aleatorias normales son:

1. Si el vector $N \times 1$ $y \sim \mathcal{N}(\mu, \Sigma)$ y el escalar $w = (y - \mu)'\Sigma^{-1}(y - \mu)$, entonces $w \sim \chi^2(N)$
2. Si el vector de $N \times 1$ $z \sim \mathcal{N}(0, I)$ y la matriz de $N \times N$ idempotente y no estocástico G tiene rango$(G) = r \leq N$, entonces el escalar $w = z'Gz \sim \chi^2(r)$

El resultado que $z'z \sim \chi^2(N)$ sigue como un caso especial de cualquiera de estas dos propiedades.

Table 3.1: Distribuciones Comúnes con su Esperanza y Varianza

Distribución	FDP	$E(X)$	$V(X)$
Distribuciones Discretas			
Bernoulli	$f(x;p) = \begin{cases} p^x(1-p)^{1-x} & \text{si } x = 0,1 \\ 0 & \text{si no} \end{cases}$	p	$p(1-p)$
Uniforme Discreta	$f(x;n) = 1/n \quad \text{si } x = 0,1,2\ldots,n$	$(N+1)/2$	$(N^2-1)/12$
Binomial	$f(x;n,p) = \begin{cases} \binom{n}{x} p^x(1-p)^{n-x} & \text{si } x = 0,1,2\ldots,n \\ 0 & \text{si no} \end{cases}$	np	$np(1-p)$
Poisson	$f(x;\lambda) = \begin{cases} \frac{\exp(-\lambda)\lambda^x}{x!} & \text{si } x = 1,2,\ldots \\ 0 & \text{si no} \end{cases}$	λ	λ
Distribuciones Continuas			
Rectangular en intervalo $[a,b]$	$f(x) = \begin{cases} \frac{1}{b-a} & \text{si } a \leq x \leq b \\ 0 & \text{si no} \end{cases}$	$(a+b)/2$	$(b-a)^2/12$
Exponencial	$f(x) = \lambda \exp(-\lambda x)$	$1/\lambda$	$1/\lambda^2$
Normal Estandarizada	$f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x}{2}\right).$	0	1
Logística Estandarizada	$f(x) = \frac{\exp(x)}{(1+\exp(x))^2}$	0	$\pi^2/3$

Refiere a Tabla 3.1 de Goldberger (1991, p. 28)

3.2 Comportamiento Asintótico

Nota de Lectura: Cameron and Trivedi (2005) tienen un apéndice que ofrece un tratamiento de los elementos más centrales de teoría asintótica aplicada a estimadores econométricos. Los capítulos de Stachurski (2016) son muy buenos, y ofrecen más detalle de los resultados que revisamos en estos apuntes. Demostraciones formales para los resultados de esta sección están disponibles en Rao (1973) (las demostraciones formales no son examinables, pero en cada caso se indica la referencia exacta de su ubicación en Rao (1973)). Para un tratamiento muy extensivo, existe el libro de texto (avanzado) de Davidson (1994). En esta sección, vamos a seguir la notación de Cameron and Trivedi (2005).

Consideremos el comportamiento de una secuencia de variables aleatorias b_N en la medida que $N \to \infty$. Utilizamos el subíndice N para indicar que esta estadística depende de la cantidad de observaciones N a partir de la cual se calcula. Por ejemplo, si nos interesa hablar de la cantidad promedio de educación en la población de Chile, el valor que estimamos con $N =\sim 15.000.000$ observaciones en el Censo es, probablemente, distinto al valor que estimamos con $N = 30.000$ observaciones aleatorias en la encuesta CASEN. Vamos a tener dos preocupaciones principales cuando hablamos del comportamiento asintótica. La primera es la **convergencia en probabilidad** de la cantidad b_N a algun límite b, que podría ser un valor escalar, o una variable aleatoria. La segunda, si el límite b es una variable aleatoria, es la **distribución límite**.

Convergencia en Probabilidad Decimos que la secuencia de variables aleatorias b_N converge en probabilidad a b si para toda $\delta > 0$:

$$Pr\left(|b_N - b| > \delta\right) \to 0 \qquad \text{a la medida que } N \to \infty. \tag{3.19}$$

Cameron and Trivedi (2005, p. 945) dan una definición más formal como su Definición A.1. Ésta nos permite elegir cualquier valor arbitrariamente pequeño para δ, y sabemos que si el N de la muestra es suficientemente grande, la probabilidad de que b_N difiere de b incluso en sólo δ unidades, es nula. Eso nos permite escribir $\text{plim}_{N\to\infty} b_N = b$, donde plim refiere a la **probabilidad límite** (o a veces por simplicidad plim $b_N = b$) o $b_N \xrightarrow{p} b$. Esta definición sirve para variables aleatorias escalares. También existe una definición básicamente idéntica para variables aleatorias vectoriales, donde se reemplaza $|b_N - b|$ de la ecuación 3.19 con $||\boldsymbol{b}_N - \boldsymbol{b}|| = \sqrt{(b_{1N} - b_1)^2 + \ldots + (b_{KN} - b_K)^2}$. Volveremos a la Convergencia en Probabilidad pronto cuando hablemos de la consistencia de estimadores.

En la econometría, generalmente nos va a interesar trabajar con promedios o sumatorias sobre variables aleatorias en una muestra de interés. Por lo tanto, nos va a interesar considerar resultados límites considerando el comportamiento de promedios. Por suerte, existen dos clases de teoremas que son de importancia fundamental en probabilidad que hablan del comportamiento de promedios

en el límite. Éstos son la **Ley de los Grandes Números** (LGN), y el **Teorema del Límite Central** (TLC). El LGN describe el comportamiento del promedio de N variables cuando N crece sin límites, el y TLC describe el comportamiento de la distribución del promedio cuando N crece sin límites. Vamos a considerar estos dos resultados con más detalle en las sub-secciones que vienen.

3.2.1 La Ley de los Grandes Números

La ley de los Grandes Números describe un caso especial de convergencia en probabilidad, cuando la variable b_N refiere a una media muestral, o:

$$b_N = \bar{X}_N = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{3.20}$$

Aquí X_i refiere a una única realización de una variable aleatoria. Si cada X_i comparte el mismo promedio μ definimos $E(X) = \mu$ como el promedio poblacional, y la Ley Débil de los Grandes Números dice:

$$\bar{X}_N \to \mu \qquad \text{a la medida que } N \to \infty.$$

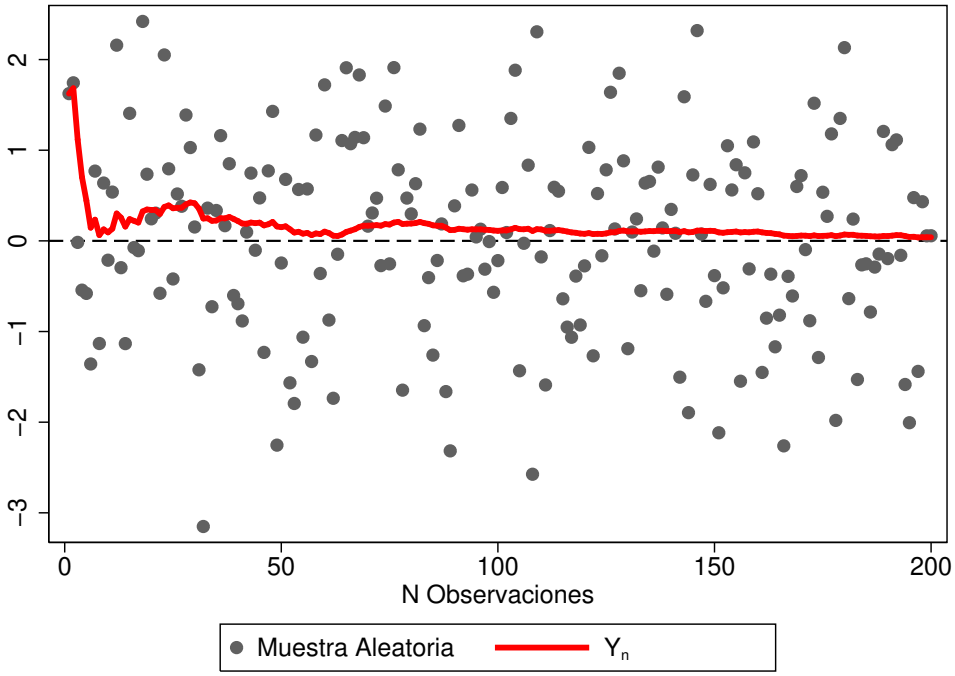
También se puede escribir de la misma forma como $\text{plim}\, \bar{X}_N = \mu$. Existe varias Leyes de Grandes Números, dependiendo de la distribución de variable que entra en el promedio, y el tipo de convergencia deseada. Detalles adicionales y definiciones formales están disponibles en Cameron and Trivedi (2005, pp. 947-948), pero brevamente, consideramos dos LGN. La LGN de Kolmogorov es la Ley relevante cuando cada X_i es iid. Si asumimos que X_i es iid, entonces la LGN de Kolmogorov demuestra convergencia al promedio μ con el único otro supuesto necesario siendo que $E(X) = \mu$. La LGN de Markov es una LGN que no requiere un μ común. Esta LGN sólo asume que cada X_i es independiente, pero no idénticamente distribuido ($inid$). A diferencia de la LGN de Kolmogorov, la LGN de Markov da como resultado que $\bar{X}_N$ converge en probabilidad a $E[\bar{X}_N]$. En el caso de la LGN de Markov, aunque se elimina el supuesto de una distribución común, es necesario agregar un supuesto adicional, de la existencia de un momento más alto que el primer momento.

La figura 3.9 demuestra una versión simulada de la LGN con variables iid. Consideramos una serie de observaciones X_i para toda $i \in \{1, \ldots, 200\}$, (que vienen de una distribución $\mathcal{N}(0,1)$). La línea roja presenta el promedio $\bar{X}_N$ de todas las observaciones hasta N en el eje horizontal. En la medida que N aumenta, observamos que $\bar{X}_N$ se acerca cada vez más a la línea punteada igual a $\mu = 0$. De la LGN de Kolmogorov, sabemos que cuando $N \to \infty$ la probabilidad de que $\bar{X}_N$ difiere de μ para cualquier valor $\delta > 0$ se acerca a 0.

3.2.2 El Teorema del Límite Central

Antes de describir el Teorema del Límite Central (TLC) y sus implicancias, vamos a introducir otra definición fundamental cuando consideremos el comportamiento asintótico de estadísticas de

Figure 3.9: La Ley de los Grandes Números



interés. Esta es la **Convergencia en Distribución**. Decimos que una secuencia de variables aleatorias b_N converge en distribución a la *variable aleatoria* b si:

$$\lim_{N\to\infty} F_N = F \tag{3.21}$$

en cada punto de continuidad de F. Aquí F_N es la distribución de b_N y F la distribución de b. Notemos que esta definición es la contraparte distribucional de la convergencia en probabilidad que definimos antes. Ahora, escribimos $b_N \xrightarrow{d} b$, y la distribución F se conoce como la distribución límite de b_N.

Cómo veremos más adelante en estos apuntes, un resultado conveniente de la convergencia en distribución es que también funciona con transformaciones de las variables aleatorias subyacentes. Un ejemplo de esto se resume en el Teorma de Slutsky. El Teorema de Slutsky implica, si $a_N \xrightarrow{d} a$ y $b_N \xrightarrow{p} b$, donde a es una variable aleatoria, y b un constante, entonces:

(i) $a_n + b_n \xrightarrow{d} a + b$
(ii) $a_n b_n \xrightarrow{d} ab$
(iii) $a_n/b_n \xrightarrow{d} a/b$, asumiendo que b es invertible.

Este teorema permite que consideremos la distribución límite y probabilidad límite de a_N y b_N por separado, en vez de tener que intentar trabajar con la distribución límite del conjunto.

La ley de los grandes número sólo nos entrega información acerca de cómo cambia la esperanza con un aumento del N. El teorema central del límite entrega información acerca de la *distribución*

entera de la esperanza estimada, y cómo esta distribución se comporta en el límite. La TLC demuestra la existencia de una relación muy regular (y sorprendente) *sin importar* la distribución original de donde se extrae la muestra (y por ende la esperanza).

El Teorema del Límite Central (Lindeberg y Lévy) Si las variables aleatorias $X_1, \ldots, X_N$ son una muestra aleatoria de tamaño N de una distribución dada con media μ y varianza finita σ^2, entonces el Teorema de Límite Central implica:

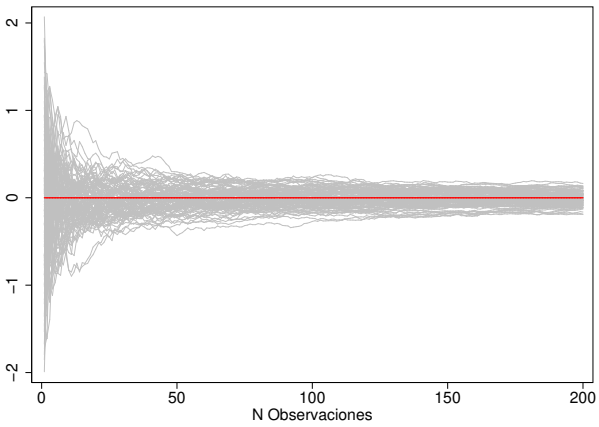
$$Z_N = \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}} \xrightarrow{d} \mathcal{N}(0,1). \tag{3.22}$$

Este resultado dice que si obtenemos una muestra aleatoria (*iid*) grande de *cualquier* distribución con media μ y varianza σ^2 (discreta o continua), la distribución de la variable aleatoria $N^{1/2}(\bar{X}_n - \mu)/\sigma$ será aproximadamente una normal estandarizada. O también, expresada en otra forma, implica que la distribución de $\bar{X}_N$ será una normal con media μ y varianza σ^2/N. Notamos que aquí la variable $(\bar{X}_N - \mu)/(\sigma/\sqrt{N})$ tendrá por definición en el límite un promedio igual a cero (ya que a $\bar{X}_N$ restamos μ), y una desviación estándar de 1, ya que se divide esta cantidad por su desviación estándar, pero la parte más importante es que sin importar la distribución base de la variable X, la distribución límite será normal. La demostración de este teorema se encuentra en Rao (1973, §2c.5)

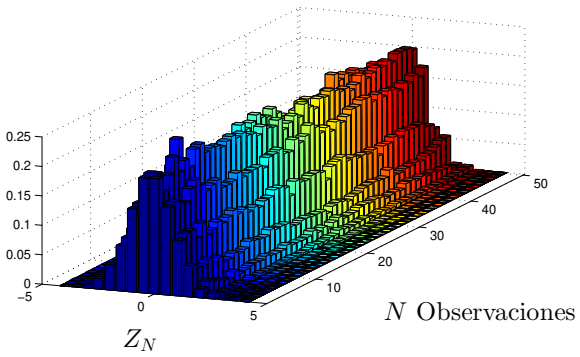
En el mundo real, muchas variables siguen una distribución normal (altura, peso, …). El TLC proporciona una posible explicación de este fenómeno. Si estas variables vienen de la influencia de muchos otros factores, entonces la distribución de la variable debería ser normal.

Figure 3.10: El Vínculo Entre la LGN y el TLC

(a) Repeticiones de la LGN



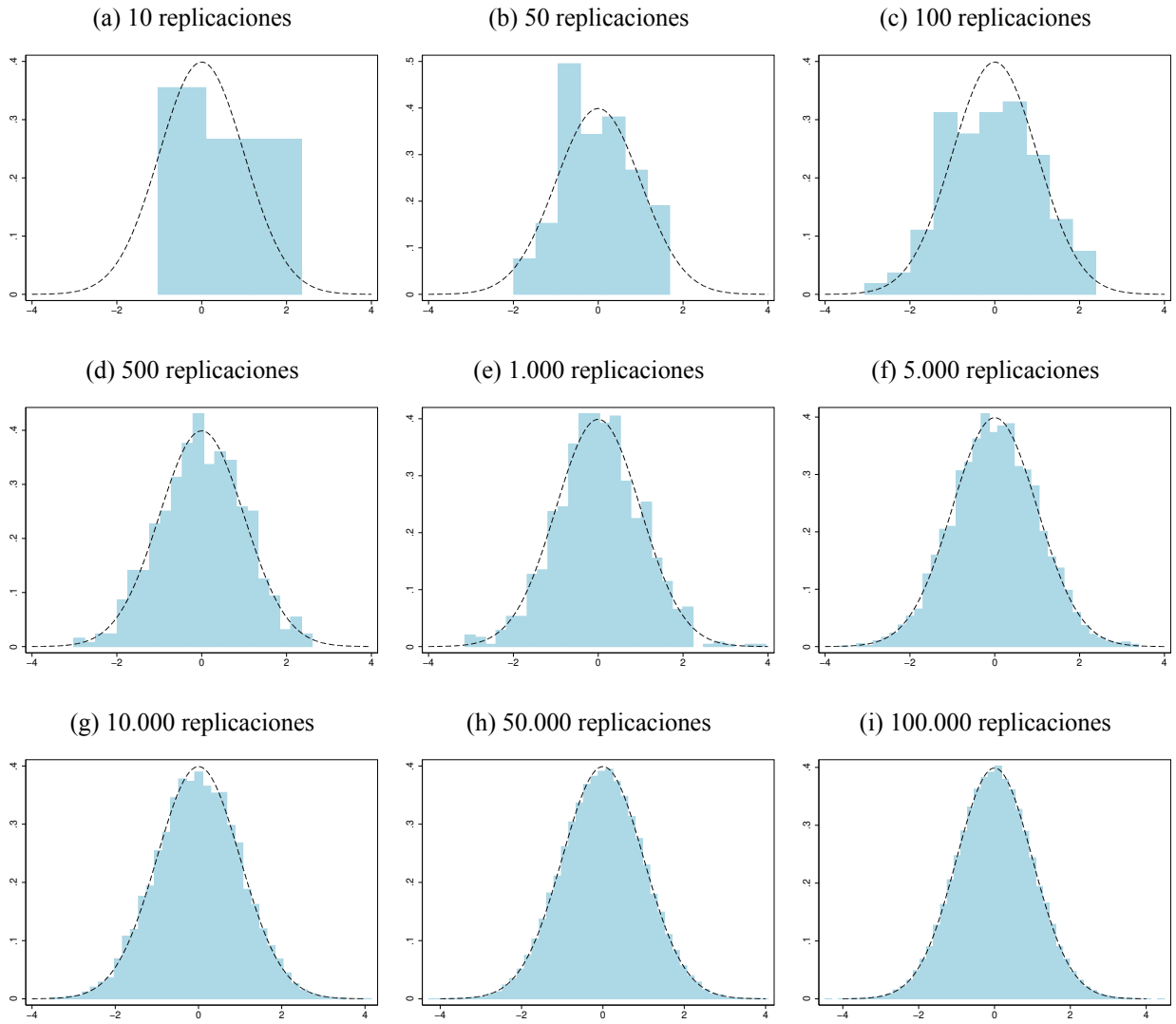
(b) fdp de Repeticiones de la LGN



En la Figura 3.10 vemos un ejemplo de este idea. En la figura (a), observamos procesos repetetidos del proceso observado en la Figura 3.9. Aquí sabemos (por la ley de los grandes números) que en el límite, este promedio se debería acercar al valor verdadero de μ. Pero cuando repetimos este

proceso muchas veces (cada línea gris es un promedio distinto), también observamos que aparece una distribución alrededor de $\bar{X}_N$. En la medida que el tamaño de la muestra (N) aumenta, esta distribución vuelve cada vez más acotada alrededor de μ. En la figura (b) observamos primero, que la distribución vuelve cada vez más precisa, y segundo, que sigue una distribución normal. Ésta distribución en la figura (b) es justamente descrita por el TLC. Cuando el tamaño de la muestra crece, la varianza cae en el orden de magnitud N^{-1}.[5] Este resultado se traduce en el teorema del límite central. Si consideramos la transformación descrita en ecuación 3.22, observamos como la distribución de esta variable se va regularizando (hacía la distribución $\Phi(x)$) en la medida que se observan más replicaciones del proceso subyacente.

Figure 3.11: El Teorema del Límite Central



Nota: Consideramos la sumatoria de $N = 100$ muestras de una variable normal. Las distintas figuras demuestran la variación en $Z_N = \frac{\bar{X}_N - \mu}{\sigma/\sqrt{N}}$ cuando Z_N se calcula una distinta cantidad de veces.

La Figura 3.11 presenta una otra ilustración del TLC. En esta figura observamos justo la aparen-

[5] Para ver esto, notemos que la ecuación 3.22 implica que $\bar{X}_N \xrightarrow{d} \mathcal{N}(\mu, \sigma^2/N)$, donde el denominador de la varianza es N.

cia de la distribución normal prometida en la medida que nos acercamos a una distribución límite. Los paneles iniciales desmuestran la distribución de Z_N de 3.22 cuando consideramos una cantidad pequeña de replicaciones del proceso generador de Z_N (con un N fijo). En la primera fila observamos que esta distribución se asemeja bastante poco a una distribución normal estandarizada con 10, 50 y 100 replicaciones. A partir de 500 replicaciones, la distribución de Z_N empieza a se bastante parecida a una distribución normal estandarizada analítica, y una vez que se observa 100.000 replicaciones del proceso, la distribución empírica y la distribución analítica casi no se diferencian. En cada caso, la distribución subyacente que genera $\bar{X}_N$ en este proceso está basada en una serie de simulaciones de números pseudo-aleatorios que son normales. En la clase computacional al final de esta sección, dejamos como ejercicio demostrar que este resultado se obtiene a partir de *otras* distribuciones también.

TLC de Liapounov (independencia) El TLC de Lindeberg Lévy asume que cada observación es independiente e idénticamente distribuida. Típicamente en econometría no tenemos observaciones con esta regularidad. Pero también existe un teorema del límite central para una secuencia de variables aleatorias que son independientes, pero no necesariamente identicamente distribuidas. Asumimos que $E(X_i) = \mu_i$ y $Var(X_i) = \sigma_i^2$. Definimos:

$$Z_N = \frac{\sum_{i=1}^N X_i - \sum_{i=1}^N \mu_i}{\left(\sum_{i=1}^N \sigma_i^2\right)^{1/2}},$$

y entonces $E(Y_n) = 0$, y $Var(Y_n) = 1$. Suponemos que las variables aleatorias $X_1, X_2, \ldots$ son independientes y $E(|X_i - \mu_i|^3) < \infty$ for $i = 1, 2, \ldots$. Entonces, $Z_N \xrightarrow{d} \mathcal{N}(0, 1)$. Notemos que aquí el resultado final es igual que el caso *iid*, pero en este caso *inid* (independiente y no idénticamente distribuida), requerimos el supuesto adicional $E(|X_i - \mu_i|^3) < \infty$. Esta versión del TLC está basada en Rao (1973). Cameron and Trivedi (2005, p. 950) presentan una versión generalizada, en base a White (2001).

Clase Computacional: Escribiendo Programas Computacionales

Cuando se trabaja con datos empíricos en un proyecto econométrico, es de importancia fundamental guardar los comandos computacionales utilizados, para poder replicar tu trabajo en una fecha futura. Esto será necesario si quieres extender tu análisis, o si simplemente quieres recordar exactamente qué has hecho, replicar tu trabajo, o pasar tu código a otro/a investigador/a para permitir que ello/as replican tu trabajo. En Stata, típicamente basta utilizar un "archivo do", que es un archivo que termina en `.do`, (por ejemplo `analisisTLC.do`) y que se puede correr en Stata escribiendo: `do analisisTLC.do`. Estos archivos do funcionan básicamente igual que si escribes cada comando en el interpretador de Stata línea por línea.

Pero a veces, más de simplemente guardar una serie de comandos para replicar un análsis específico, nos interesa escribir un programa computacional para permitirnos seguir un proceso específico con cualquier conjunto de datos. Por ejemplo, consideramos el comando `sum` en Stata. Este es un programa que permite resumir las características de cualquier variable disponible en una base de datos. Es un comando que se puede utilizar muchas veces, y siempre devolverá los mismos estadísticos (el promedio, la desviación estándar, etc.), cuyos valores dependen de los *argumentos* utilizados. Los argumentos refieren a las variables o la información pasada al programa computacional. Por ejemplo, si escribimos `sum var1 var2`, los argumentos son `var1` y `var2`, mientras si escribimos `sum ingreso`, el argumento es la variable `ingreso`. Todos los comandos de Stata son programas, en el sentido de que se puede replicar el mismo procedimiento con distintos argumentos.

En esta clase, revisaremos brevemente cómo nosotros podemos escribir programas en Stata, para nuestro uso permanente. Una ventaja de escribir un programa es que nos permite estandarizar nuestro análisis. Por ejemplo, si en muchos distintos proyectos sabemos que vamos a estar realizando un determinado tipo de operación, en vez de escribir una serie de archivos `do` que lo hace de forma *ad hoc*, podríamos escribir un único programa para utilizar en cada proyecto, simplemente cambiando los argumentos que damos al programa. Una introducción *mucho* más comprensivo a programación en Stata está disponible en el manual de Stata: https://www.stata.com/manuals13/u18.pdf#u18ProgrammingStata.

En Stata, se escriben los programas utilizando un "archivo ado", o un archivo que termina en `.ado`. Hay algunas ingredientes básicas en los programas computacionales. Uno es el sintaxis, que define los tipos de argumentos que el programa aceptará, y otro es una serie resultados que el programa devuelve al usuario. La meta del programa es tomar los argumentos que son pasados al programa por el usuario, y devolver al usuario el resultado deseado.

Para ver un ejemplo simple, podemos pensar en un programa que suma una serie de valores escalares. En realidad, si nos interesa sumar una serie de valores, podemos utilizar el comando en Stata: `dis 1+2+3` (o cualquier sumatoria de interés). Pero imaginamos por el momento que nos gustaría escribir un comando que se llama "sumatoria", y que puede tomar como argumentos una serie de valores escalares, y que devolverá su sumatoria. A continuación daremos un ejemplo de un programa de este estilo. Este programa debería ser contenido en un archivo llamado `sumatoria.ado`.

sumatoria.ado

```
cap program drop sumatoria
program sumatoria, rclass
version 11.0

syntax anything
```

```
local SUM = 0
foreach num of numlist `anything' {
    local SUM = `SUM'+`num'
}
dis "La sumatoria de estos números es: `SUM'."
return scalar sumatoria = `SUM'

end
```

Los programas en Stata siempre deben partir con las tres primeras líneas dadas arriba. Primero se borra el programa de la memoria de Stata por si ya está definida (nota, aquí es importante asegurar que Stata no tiene otro programa con el mismo nombre!), despúes se define el nombre del programa, y, opcionalmente la clase del programa (en nuestro caso utilizamos un `rclass`, que tiene implicancias para donde los resultados serán guardados, y por último se define una versión de Stata *mínima* que funcionaría con este programa. Por ejemplo, si escribimos `version 11.0`, esto implica que nuestro programa funcionará en Stata versión 11, 12, 13, 14, 15 o cualquier versión futura. En algunos casos, necesitaremos funciones de Stata que solamente aparacieron en versiones más modernas de Stata, que podría limitar las versiones que pueden correr el programa.

En Stata, la sintaxis del programa tiene una forma estandarizada (ver aquí: https://www.stata.com/manuals13/psyntax.pdf). En nuestro caso, estamos diciendo que este programa puede aceptar "cualquier" argumento utilizando `syntax anything`. Otras veces podríamos querer solamente aceptar variables (`syntax varlist`), o agregar opciones adicionales. Por el momento, utilizamos un sintaxis simple que basta para nuestra sumatoria. Las siguientes 5 líneas realizan la sumatoria. Nota que aquí, si los argumentos del programa *no* son númericos, el programa no funcionaría (¿por qué?).

Al final del programa, devolveremos el resultado a Stata como el valor escalar `sumatoria`. Dado que hemos definido que este programa es del "rclass", este valor será guardado como `r(sumatoria)`. Por último, el comando `end` dice a Stata que el programa puede terminar de ejecutar.

Si queremos correr este programa en Stata, ahora será fácil observar que el programa realmente hace lo que debería hacer. Si el programa está ubicado en el directorio donde estás trabajando (o algunos de los directorios listados cuando escribes `sysdir`), puedes simplemente escribir el nombre del programa, más los argumentos de interés. Por ejemplo:

```
. sumatoria 1 2 3
La sumatoria de estos números es: 6.
```

```
. return list
scalars:
r(sumatoria) = 6

. sumatoria 1.0001 2 3 68213213
La sumatoria de estos números es: 68213219.0001.

. dis r(sumatoria)
68213219
```

Preguntas/Actividades Refiere al código TLC.ado. Este código proporciona un programa para examinar el Teorema del Límit Central variando el tamaño de la muestra (N observaciones) la cantidad de repeticiones, y la distribución de la variable subyacente.
Su sintaxis es:
`tlc distribucion, observaciones(#) reps(#) [mu(#) sigma(#) a(#) b(#)].`
donde `distribucion` puede ser `rnormal` (para una distribución subyacente normal, o `runiform` para una distribución subyacente uniforme. El argumento `observaciones` refiere a la cantidad de observaciones en la muestra, y `reps` refiere a la cantidad de replicaciones de la sumatoria. `mu()` y `sigma()` son argumentos opcionales para variar los parámetros de la variable normal ($\mathcal{N}(mu, sigma)$) si se especifica `rnormal`, y `a()` y `b()` son argumentos opcionales para variar los parámetros de la variable uniforme ($\mathcal{U}(a,b)$) si se especifica `runiform`. Si no se especifica estos parámetros, son tomados como 0 y 1.

1. Examine el funcionamiento del programa cambiando el tipo de variable aleatoria, la cantidad de observaciones, y la cantidad de replicaciones. Hay varias cosas nuevas en `tlc.ado`, y puede ser informativo revisar el manual para el sintaxis de Stata.

2. Replica los distintos paneles de la figura 3.11.

3. (Difícil) Este código funciona para el TLC de Lindeberg–Lévy con variables iid. Escribe una versión para examinar el TLC de Liapunov para variables inid.

4. Practica haciendo programas con algunas otras funciones que podrían ser de uso para ti.

3.3 Estimadores y Estimación

Nota de Lectura: En esta sección, examinamos la estimación paramétrica, es decir, asumiendo que nuestros datos vienen de una distribución con una cantidad finita de parámetros. Detalles de estimadores en este contexto están presentados en Casella and Berger (2002, Capítulo 7) y DeGroot and Schervish (2012, Capítulo 7). Ambos textos son una buena referencia. La presentación en Stachurski (2016, Capítulo 8) es más general, presentando una teoría de estimación para situaciones paramétricas y no-paramétricas. Este libro presenta una sintésis muy comprensiva, y más detallada que la presentación aquí.

3.3.1 Una Introducción y Descripción Generalizado

Consideremos un experimento, y asumamos que podemos repetir este experimento N veces. Como resultado, observamos N variables aleatorias: $(Y_1, \ldots, Y_N) = \mathbf{Y}$. Esta muestra de N realizaciones viene de una población descrita por alguna función de densidad o función de densidad de probabilidad: $f(\mathbf{y}|\boldsymbol{\theta})$. Aquí no suponemos nada acerca de los parámetros que describen esta distribución $\boldsymbol{\theta}$, pero asumimos que la forma general de la distribución es conocida.

Cuando hablamos de estimación paramétrica, nuestra meta es "encontrar" los parámetros $(\theta_1, \ldots, \theta_K) = \boldsymbol{\theta}$ que describen esta distribución.[6] Partimos sabiendo que estos parámetros vienen del espacio Ω, que con frecuencia es un espacio poco restringido por ejemplo, podría ser $\mathbb{R}^K$. A veces vamos a querer poner restricciones en este espacio. Un ejemplo es si vamos a estimar una varianza, probablemente queramos limitar a un valor positivo. El desafío de la inferencia estadística es en cómo utilizar los valores de la muestra Y para estimar valores plausibles de $\theta_1 \ldots, \theta_K$.

Para estimar los parámetros de interés, entonces necesitamos contar con (a) una muestra de datos extraídos de la población, (b) un modelo estadístico, y (c) una manera de llevar a cabo nuestra estimación puntual. Cuando hablamos de un modelo estadístico (o económico), nos referimos a la manera en que llegamos a la clase general de distribuciones que suponemos que describe nuestra población, $f(\mathbf{y}|\boldsymbol{\theta})$. Y cuando hablamos de una manera de estimar, nos referimos al proceso de encontrar los parámetros $\boldsymbol{\theta}$ *una vez* que tengamos restringidas las funciones de densidad de población a una clase paramétrica. En esta sección nos vamos a enfocar principalmente en el paso (c), y más adelante veremos algunos ejemplos de cómo formamos un modelo estadístico para un problema específico.

Formalmente, definimos a un **estimador puntual** como cualquier función

$$\hat{\theta} = \hat{\theta}(Y_1, \ldots, Y_N),$$

[6] O a veces nos interesa una función de estos parámetros, $\tau(\boldsymbol{\theta})$. Veremos que el desarrollo aquí también para encontrar funciones de los parámetros estimados.

es decir, un estadístico de la muestra de datos. Notemos que esta definición es muy general, y de hecho, ni siquera implica alguna correspondencia entre el estimador y el parámetro que se intenta estimar. Veremos más adelante (en la sección 3.3.5), que esta correspondencia viene más bien por condiciones que vamos a exigir a nuestro estimador para considerarlo un buen estimador. De esta definición podemos desprender varios hechos. Primero, notamos que el estimador es una función de los datos $Y_1, \ldots, Y_N$, que son variables aleatorias, y por ende el estimador es una variable aleatoria. Segundo, el estimador es una regla determinśtica de cómo convertir la muestra de datos disponibles cualquiera sean los valores de esta muestra al vector o escalar $\hat{\theta}$. Una vez que sustituimos los valores de las variables aleatorias específicas en la ecuación, tenemos una **estimación**. Y por último, notamos el uso de un 'gorro' para indicar que estamos trabajando con un estimador (o estimación). Utilizaremos esta notación de forma consistente en estos apuntes.

En algunos casos puede parecer bastante obvio determinar un estimador para un estadístico de interés. Por ejemplo, si nos interesa estimar el medio o algún cuantil de una población de interés, la contraparte *muestral* es probablemente un buen candidato. Pero en otras situaciones puede no ser tan obvio cómo definir cuál es un buen candidato. En la sección 3.3.5 de estos apuntes volvemos a definir una series de características que nos permiten juzgar a los estimadores, y comparar varios estimadores potenciales para un mismo parámetro.

En esta sección de los apuntes vamos a examinar a dos grandes clases de estimadores. Estas clases nacen de distintos supuestos, y permiten estimar parámetros en un rango amplio de situaciones. Además, vamos a utilizar estos estimadores a lo largo del curso. Específicamente, vamos a introducir el estimador de Máxima Verosimilitud, y el estimador del Método de Momentos. Esto no es para sugerir que estos estimadores sean las únicas posibilidades, o incluso los más utilizados en aplicaciones econométricas, pero los introducimos aquí dada su flexibilidad y utilidad en un importante grupo de problemas.

Por último, antes de seguir, destacamos la importancia de nuestros supuestos acerca del estado del mundo cuando definimos nuestros estimadores. La base de cada estimador es un modelo de probabilidad que permite definir las clases de distribuciones que consideramos al momento de estimar nuestros parámetros de interés. Cuando partimos con supuestos muy fuertes, esto nos puede proporcionar una solución muy fácil al estimador, o producir un resultado matemático muy elegante. Sin embargo, al estar interesados en fenómenos del mundo natural, y especificamente de comportamiento humano, estos supuestos también deben ser capaces de capturar la complejidad de los fenómenos de interés. Es importante que siempre recordemos esta dualidad al momento de plantear nuestros estimadores, y más generalmente cuando pensemos en inferencia estadística. A continuación, reproducimos un párrafo de Manski (2003) que captura esta idea de forma muy simple:

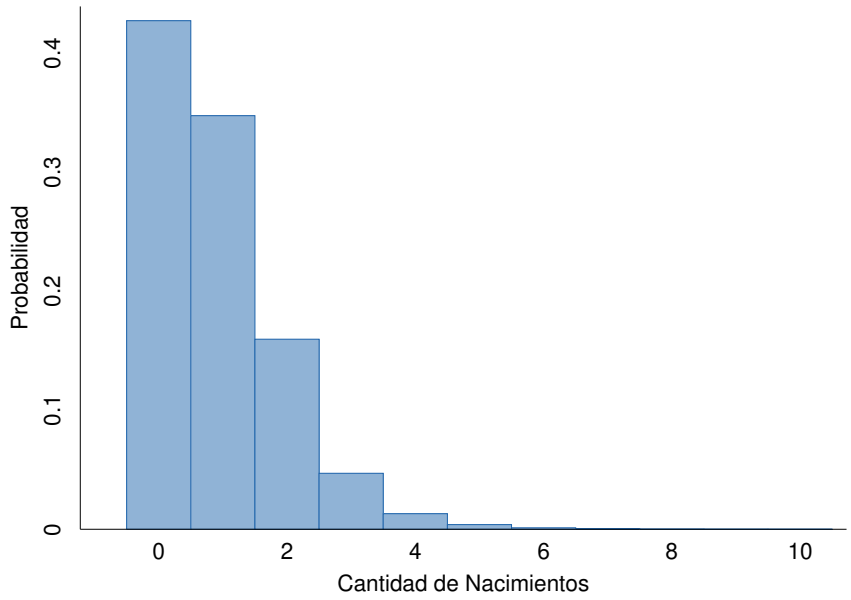
> "*The Law of Decreasing Credibility*: The credibility of inference decreases with the strength of the assumptions maintained.

> This principle implies that empirical researchers face a dilemma as the decide what assumptions to maintain: Stronger assumptions yield inferences that may be more powerful but less credible. Statistical theory cannot resolve the dilemma but can clarify its nature."
>
> Manski (2003, p. 1)

3.3.2 Una Aclaración: La Población

Cuando nos referimos a "la población", nos referimos al universo de objetos (personas, unidades, etc.) en que estamos interesado/as hacer inferencia estadística. Es una construcción teórica, y es infinita. Por ejemplo, si queremos saber algo acerca de las Pequeñas y Medianas empresas (PYMES) en Chile, la población es todas las PYMES que podrían hipotéticamente existir. Podemos observar todas las PYMES ahora en Chile, pero esto es sólo parte de la población infinita. Por lo tanto, incluso cuando tenemos un censo, hablamos de una muestra de la población. Aunque podría parecer curioso no tratar el censo como el universo de la población estadística cuando efectivamente *es* toda la población del país, esto se debe a la diferencia en la terminología de poblacón estadística, y población como típicamente se utiliza la palabra en la vida cotidiana. Nuestro universo o población estadística refiere a toda la población que podría existir a la raíz del proceso generador de datos.

Figure 3.12: Función de Probabilidad: Número de Nacimientos Anteriores de Madres, Chile 2015



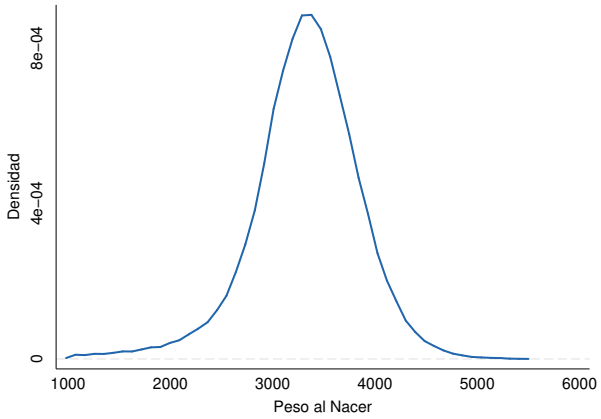
Cuando trabajamos con una muestra de la población, nos referimos a funciones de distribución *empíricas*. En las figuras 3.12-3.13 consideramos una serie de distribuciones empíricas. La figura 3.12 presenta una representación gráfica de una función de probabilidad en base a la representación empírica de la cantidad de hermano/as de todos los bebés que nacieron en Chile en 2015. Cada

resultado potencial (tener una cantidad x de hermana/os mayor o igual a cero) está asociado con una probabilidad correspondiente.

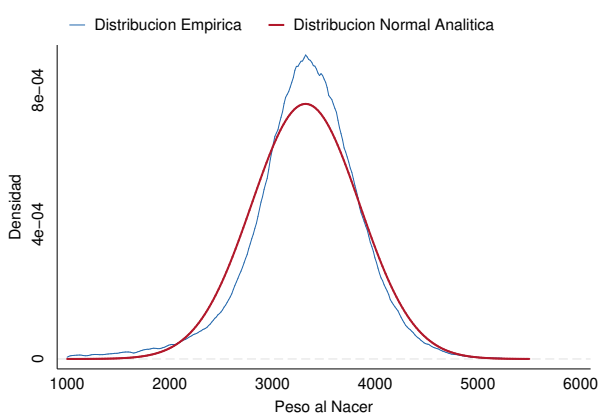
Y en el segundo caso, observamos una función de densidad de probabilidad empírica: la distribución de peso al nacer de todos los nacimientos ocurridos en Chile en el año 2015. Aquí, aunque son todos los nacimientos, sigue siendo una muestra, dada la definición de población a la que nos referimos antes. Y aquí se ve el vínculo con los modelos estadísticos que discutimos en la sección anterior. Aunque no sabemos los parámetros de la distribución poblacional que produjeron la densidad en La Figura 3.13, al observar la distribución empírica podría ser razonable suponer que esta distribución empírica fue extraída de una distribución poblacional normal. En la figura (b) se compara la distribución empírica con una distribución analítica con la misma media y desviación estándar. En las secciones que vienen, veremos con más detalle los supuestos y pasos necesarios para estimar parámetros de una distribución poblacional, a partir de una muestra empírica.

Figure 3.13: Función de Densidad: Peso al Nacer, Chile 2015

(a) Función de Densidad Empírica



(b) Distribución Empírica y Analítica



Nota: Distribuciones se basan en los 244.670 nacimientos que ocurrieron en Chile en 2015. En el segundo panel, también se presenta una distribución normal analítica con la misma media y desviación estándar que la distribución empírica.

3.3.3 Método de Estimación: Método de Momentos

En las dos sub-secciones que siguen, vamos a introducir una serie de métodos de estimación: primero el método de momentos, y luego el de máxima verosimilitud. Para introducir estas técnicas, partiremos con un caso simple: imaginemos que existe alguna variable Y que es una variable aleatoria con:

$$Y \sim \mathcal{N}(\mu, \sigma^2) \tag{3.23}$$

y vamos a suponer que tenemos una muestra de realizaciones Y_i extraída de la población. Nuestro interés es estimar $\widehat{\mu}$ y $\widehat{\sigma}$. Esta descripción es un tipo de "modelo estadístico" (aunque uno bastante sencillo). Estamos asumiendo que tenemos un proceso generador de datos (poblacional) donde Y

sigue la distribución normal. Así, estamos limitando la clase de distribuciones posibles para Y, pero no estamos restringiendo los parámetros que describen Y dentro de esta clase de distribuciones.[7]

Ahora, nosotros contamos con una muestra $\boldsymbol{Y}$ de tamaño N ($\boldsymbol{Y}$ es un vector de $N \times 1$). Nuestro interés es inferir algo acerca de los parámetros $\boldsymbol{\theta} = (\mu, \sigma^2)$ que describen el modelo poblacional 3.23. Sin embargo, sólo tenemos nuestra muestra finita (de tamaño N) para hacer deducciones. Entonces, necesitamos una técnica de estimación que vincule los datos de nuestra muestra, con nuestro modelo poblacional para estimar $\boldsymbol{\theta}$.

La técnica del Método de Momentos parte con el principio de analogía. Este principio sugiere que debemos estimar nuestros parámetros poblacionales utilizando estadísticas muestrales que tienen las mismas características en la muestra, que los parámetros poblacionales tienen en la población. Entonces, por ejemplo, si estamos interesados en estimar la media poblacional (algo que no observamos), el principio de analogia sugiere utilizar la media muestral, algo que *sí* observamos. Este es una técnica simple, y muy poderosa y generalizable. Específicamente, un estimador del método de momentos busca definir los *momentos poblacionales* que caracterizan el proceso generador de datos, y después estimarlos utilizando **los momentos análogos de la muestra**.

En el caso de interés aquí (el proceso generador de datos 3.23), tenemos dos momentos de interés (refierese a la sección 3.1.3 para una discusión acerca de los momentos de una distribución). Estos primeros dos momentos son la esperanza y la varianza, los que en la población se escriben:

$$E(Y) = \mu \tag{3.24}$$

$$Var(Y) = E[Y^2] - (E[Y]^2) = \sigma^2. \tag{3.25}$$

Referimos a las ecuaciones 3.24 y 3.25 como los momentos poblacionales.

La idea del método de momentos es simplemente utilizar las contrapartes muestrales para estimar la respectiva cantidad poblacional. En este caso, podemos escribir los momentos muestrales como:

$$\widehat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i \tag{3.26}$$

$$\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 \tag{3.27}$$

donde ahora los momentos sí son observables. En términos mecánicos, siempre escribimos a nuestros momentos como alguna cantidad igual a cero, resultando en las siguientes condiciones de

[7] En realidad, tenemos *una* restricción sobre uno de los parámetros. Sabemos que la varianza σ^2 es igual o superior a 0.

momentos poblacionales:

$$E[Y] - \mu = 0 \tag{3.28}$$
$$E[Y^2] - (E[Y]^2) - \sigma^2 = 0. \tag{3.29}$$

Hasta ahora nos hemos referido a estos momentos como "momentos centrales". Y entonces, los vamos a estimar usando las condiciones de momentos centrales análogas de la muestra que observamos:

$$\frac{1}{n}\sum_{i=1}^{n} Y_i - \widehat{\mu} = 0 \tag{3.30}$$
$$\frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \left(\frac{1}{n}\sum_{i=1}^{n} Y_i\right)^2 - \widehat{\sigma}^2 = 0 \tag{3.31}$$

donde este proceso de estimación consiste en resolver dos ecuaciones (ecuación 3.30-3.31) con dos incógnitas, resultando un proceso simple de eliminación. Más adelante en el curso, vamos a considerar estimadores de métodos de momentos bastante más complejos, con más momentos que parámetros a estimar, y en este caso el problema de estimación es un problema de minimización, esta técnica de estimación es conocida como el **método de momentos generalizados**.

El Estimador de Forma General Una distribución normal y $\boldsymbol{\theta} = (\mu, \sigma)$ es sólo uno de *muchos* posibles ejemplos del método de momentos. En forma general, se puede representar el método de momentos con: (a) Los momentos poblacionales (b) Los momentos muestrales correspondientes, y (c) La solución de los momentos muestrales. Los momentos poblacionales se escriben de forma genérica como:

$$E[\boldsymbol{h}(\boldsymbol{w}_i, \boldsymbol{\theta}_0)] = \mathbf{0}$$

donde:

- $\boldsymbol{\theta}$ es un vector de $K \times 1$ de los parámetros a estimar
- $\boldsymbol{\theta}_0$ es el valor de $\boldsymbol{\theta}$ en la población
- $\boldsymbol{h}(\cdot)$ en una función vectoral de $r \times 1$ de los momentos ($r \geq K$)
- $\boldsymbol{w}$ es un vector que incluye todos los datos observables

Y de forma parecida, escribimos los momentos muestrales como:

$$\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{h}(\boldsymbol{w}_i, \hat{\boldsymbol{\theta}}) = \mathbf{0} \tag{3.32}$$

donde aquí la única diferencia es que reemplazamos los momentos poblacionales por su contraparte muestral, y denotamos al vector de parámetros que resuelve 3.32 como $\widehat{\theta}$. Este vector es la solución a esta ecuación, y el estimador del método de momentos. A veces denotamos al estimador como

$\hat{\boldsymbol{\theta}}_{MM}$.

Cuando la cantidad de momentos es igual a la cantidad de parámetros que se busca estimar ($r = K$), se puede simplemente resolver el sistema de ecuaciones en 3.32. Sin embargo, muchas veces es más facil (computacionalmente) minimizar una función de la siguiente forma:

$$Q_N(\boldsymbol{\theta}) = \left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}(\boldsymbol{w}_i, \boldsymbol{\theta})\right]' \left[\frac{1}{N}\sum_{i=1}^{N}\boldsymbol{h}(\boldsymbol{w}_i, \boldsymbol{\theta})\right]$$

El estimador del método de momentos $\hat{\boldsymbol{\theta}}_{MM}$ es (también) la cantidad que minimiza Q_N:

$$\hat{\boldsymbol{\theta}}_{MM} = \arg\min_{\boldsymbol{\theta}} Q_N(\boldsymbol{\theta}).$$

3.3.4 Método de Estimación: Máxima Verosimilitud

El estimador del método de momentos tiene una lógica bastante razonable: nuestro valor estimado de θ es el valor de θ que resuelve la contraparte de los momentos poblacionales en la muestra. Sin embargo, no es el único estimador posible, y no utiliza toda la información disponible en nuestro modelo estadístico. En particular, el método de momentos extrae toda la información para estimar de algunos pocos puntos en la distribución de probabilidad (los momentos). En algunos casos esto puede ser conveniente, dado que es más fácil plantear momentos que toda una distribución de probabilidad (y los supuestos son más flexibles), pero en otras casos (como en el caso de interés aquí), contamos con más información que sólo los momentos. Esto nos lleva a otra manera de estimar, que es la técnica de máxima verosimilitud, con la que utilizamos *más información* para estimar que sólo los momentos.

La idea de la Máxima Verosimilitud (MV) viene de Fisher (1922) y el principio de verosimilitud:

> Eligimos como estimador para el vector de parámetros $\boldsymbol{\theta}_0$ al valor de $\boldsymbol{\theta}$ que maximiza la probabilidad de observar la muestra actual.

Notamos que no estamos diciendo nada acerca de hacer cumplir los momentos de la distribución (y de hecho, pueden haber buenos estimadores de máxima verosimilitud que no hacen cumplir los momentos), pero ahora tenemos que tener una manera de calcular la probabilidad de haber observado una serie de parámetros específicos, dado un modelo de probabilidad.

De nuevo en esta sección vamos a asumir que tenemos una muestra aleatoria (iid) extraída de una distribución normal:

$$Y_i \sim \mathcal{N}(\mu, \sigma^2)$$

La idea de MV es que podemos utilizar toda la información de la fdp (y no sólo los momentos cen-

trales). Utilizando la misma notación anterior, escribimos la función de densidad de probabilidad como $f(\boldsymbol{y}|\boldsymbol{\theta})$, o también como $f(\boldsymbol{y}|\mu,\sigma)$ si queremos destacar que los parámetros que describen la fdp en este caso son μ, su promedio, y σ, su desviación estándar.

Como tenemos una muestra de tamaño N extraída de una distribución normal, por la naturaleza de un proceso estocástico, habrán algunas observaciones que son bastante cercanas a la media μ, y otras observaciones que son más alejadas. Pero nosotros observamos todas las observaciones en la muestra, y queremos inferir a partir de estas observaciones cuáles fueron los parámetros poblacionales que las generaron. Para poder considerar la muestra entera, tenemos que partir con la fdp conjunta de todas las variables aleatorias. Esta fdp conjunta nos dice la probabilidad de haber observado toda la muestra en conjunto. Dado que contamos con una muestra de variables independientes, tenemos que:

$$\begin{aligned} f(y_1,...,y_N|\mu,\sigma) &= f(y_1|\mu,\sigma)\times\cdots\times f(y_N|\mu,\sigma) \\ &= \prod_{i=1}^{N} f(y_i|\mu,\sigma), \end{aligned} \tag{3.33}$$

donde el lado derecho de la ecuación 3.33 viene de la independencia de observaciones, y la ecuación 3.12 discutida anteriormente.

Aquí, en palabras, tenemos la probabilidad de observar una combinación $y_1,\ldots,y_N$, dado los parámetros verdaderos μ y σ. Sin embargo, lo que queremos es algo un poco distinto. Queremos saber cuál es la probabilidad de tener parámetros verdaderos μ y σ dado el conjunto de datos que hemos observado. Es decir, en vez de considerar a la función con parámetros dados y con observaciones de y_i que cambian, queremos imaginar que los y_i son dados, y que vamos variando los parámetros μ y σ.

Por esto, escribimos la **función de verosimilitud**:

$$\mathcal{L}(\mu,\sigma|y_1,\ldots,y_n) = \prod_{i=1}^{N} f(y_i|\mu,\sigma) \tag{3.34}$$

Ahora, aunque la función en el lado izquierdo igual, la interpretación es distinta. Con la función de verosimilitud, calculamos la probabilidad conjunta de haber observado el vector Y variando los parámetros μ y σ. Por ende, la función de verosimilitud cuantifica cuán probable es que hubiésemos observado los datos que observamos, para un μ y σ determinado. A partir de esta función, se estima $\hat{\mu}$ y $\hat{\sigma}$ ($\hat{\boldsymbol{\theta}}$), que son los valores que maximizan la función de verosimilitud, o los valores que hacen lo más probable el haber observado la muestra que observamos:

$$\hat{\theta}_{MV} = \arg\max_{\theta\in\Omega} \mathcal{L}(\theta|y_1,\ldots,y_N).$$

En la práctica, maximizar la función de verosimilitud puede ser computacionalmente (o alge-

braicamente) difícil. En términos computacionales, dado que la función de verosimilitud se forma multiplicando N funciones de densidad (con un rango de entre 0 y 1), cuando N es grande, este producto se vuelve muy pequeño. Por eso, definimos la función de log verosimilitud, que tiene varias propiedades muy convenientes:

$$\begin{aligned} \ell(\mu, \sigma | y_1, \ldots, y_n) &\equiv \ln \mathcal{L}(\mu, \sigma | y_1, \ldots, y_n) \\ &= \sum_{i=1}^{N} \ln f(y_i | \mu, \sigma) \end{aligned} \tag{3.35}$$

Lo más conveniente es que dado que ℓ es una función monótona creciente de $\mathcal{L}$, la función ℓ y la función $\mathcal{L}$ alcanzan sus máximos en el mismo valor de $\boldsymbol{\theta}$. Así, estimamos MV de la misma manera:

$$\hat{\theta}_{MV} = \arg\max_{\theta \in \Omega} \ell(\theta | y_1, \ldots, y_N).$$

Generalmente (pero con algunas excepciones), las funciones de verosimilitud no tienen una solución de forma cerrada. En este caso, se puede estimar usando métodos computacionales. En su forma más simple, estos métodos prueban todas las combinaciones posibles de $\hat{\theta}$ y $\hat{\sigma}$ hasta encontrar el máximo. Y en programas como Stata, existen librerías que pueden maximizar una función de manera mucho más rápida, por ejemplo, utilizando algoritmos como Newton-Raphson. Para nuestros requisitos, basta saber que una vez que escribimos una función de máximia verosimilitud, tenemos que utilizar el computador para maximizarla! Si le interesa leer más de métodos computacionales para maximizar una función, dirjase al capítulo 10 de Cameron and Trivedi (2005).

Las derivaciones anteriores han presentado de forma general el proceso de MV para cualquier modelo de probabilidad, y cualquier fdp. Pero para el caso de una distribución normal, podemos ir más lejos. En este caso, sabemos cuál es la fdp para una sola realización de la variable aleatoria (de la ecuación 3.16). Así, escribimos:

$$\begin{aligned} f(y_1, ..., y_N | \mu, \sigma) &= f(y_1 | \mu, \sigma) \times \cdots \times f(y_N | \mu, \sigma) \\ &= \prod_{i=1}^{N} f(y_i | \mu, \sigma) \\ &= \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(y_i - \mu)^2}{2\sigma^2} \end{aligned} \tag{3.36}$$

donde el último paso viene de la distribución normal 3.16. Dado esto, nuestra función de verosimilitud *para una muestra de variables extraídas de una distribución normal* es:

$$\mathcal{L}(\mu, \sigma | y_1, \ldots, y_N) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp - \frac{(y_i - \mu)^2}{2\sigma^2}$$

y el logaritmo de la función de verosimilitud es:

$$\begin{aligned}\ell(\mu,\sigma|y_1,\ldots,y_N) &= N\ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right)-\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{2\sigma^2}.\\ &= -\frac{N}{2}\ln(2\pi\sigma^2)-\sum_{i=1}^{N}\frac{(y_i-\mu)^2}{2\sigma^2}. \qquad (3.37)\end{aligned}$$

Esta última ecuación 3.37 es una función que podríamos entregar a un programa computacional para maximar, y encontrar los parámetros estimados $\hat{\boldsymbol{\theta}} = (\widehat{\mu},\widehat{\sigma})$:

$$\hat{\boldsymbol{\theta}}_{MV} = \arg\max_{\boldsymbol{\theta}} \ell(\mu,\sigma|y_1,\ldots,y_N).$$

Examinaremos un ejemplo de este proceso como un código computacional, y el comando `mlexp` de Stata.

3.3.5 Propiedades de Estimadores

Como hemos definido, la estimación consiste en utilizar una muestra de datos y otra información que tenemos *a priori* para producir un valor que es en algún sentido nuestra "mejor estimación" de un parámetro desconocido. Pero un estimador es simplemente una regla determinística para definir una estadística a partir de datos observados. En realidad, podemos imaginar muchos estimadores. Una posibilidad (bastante absurda) sería simplemente elegir nuestro número favorito para todos los parámetros estimados. Aunque hemos conocido un par de técnicas basadas en principios de estimación que parecen razonables (MV y MM), no hemos definido ninguna manera para definir cuáles son los mejores estimadores, ni comparar técnicas de estimación. Aunque lo que es "mejor estimación" depende de nuestra definición y metas al momento de estimar, podríamos esperar que nuestros estimadores cumplen con ciertas condiciones o criteros para ser considerados buenos.

En las muestras finitas (o muestras pequeñas), hay dos propiedades que son particularmente relevantes. Éstas son **Sesgo** y **Eficiencia**. Veremos ambas propiedades con un poco más de detalle.

Sesgo Una propiedad intuitiva, y una manera clásica de evaluar un estimador es el sesgo. Un estimador insesgado cumple con:

$$E[\hat{\theta}] = \theta.$$

Si $E[\hat{\theta}] \neq \theta$, decimos que el estimador $\hat{\theta}$ está sesgado. Definimos el sesgo como la diferencia entre la expectativa del estimador, y el valor verdadero del parámetro poblacional que estamos intentando estimar: $E[\hat{\theta}] - \theta = \delta$. Aquí cuando hablamos de la expectativa del estimador, nos estamos refiriendo al valor promedio si pudiésemos repetir el experimento subyacente infinitas veces, cada vez sacando una muestra de tamaño N, y estimando nuestro estimador $\hat{\theta}$. Si el estimador es ins-

esgado, encontraremos que en promedio, este estimador es igual a θ. Pero esto es un constructo teórico. En la realidad, no tenemos infinitos experimentos, sino uno. Por esto, también pensamos en la eficiencia de los estimadores.

Eficiencia Dado que generalmente observamos una sola serie de datos, la insesgadez de un estimador no es nuestra única preocupación. Si tenemos un estimador insesgado pero muy impreciso, una realización en particular de $\hat{\theta}$ podría estar bastante lejos de θ. La eficiencia considera la precisión de un estimador, y el criterio de eficiencia consiste en considerar sólo los estimadores insesgados, y elegir el que tiene menor varianza. Decimos que un estimador $\hat{\theta}$ es eficiente si:

$$\text{Var}(\hat{\theta}) \leq \text{Var}(\tilde{\theta}) \tag{3.38}$$

donde $\tilde{\theta}$ es cualquier otro estimador insesgado de θ. Una manera de comprobar que un estimador sea el estimador más eficiente de todos los estimadores posibles, es utilizando la cota inferior de Cramér-Rao. La cota inferior de Cramér-Rao, demuestra que:

$$\text{Var}(\hat{\theta}) \leq \frac{1}{-E\left[\frac{d^2\ell(\theta)}{d\theta^2}\right]}$$

donde ℓ es el logaritmo de la función de verosimilitud. Esto implica que si podemos mostrar que la varianza de un estimador es igual o menor que el término de la derecha en la ecuación 3.38, sabemos que es el mejor estimador insesgado posible.

Así aunque la eficiencia también considera la varianza del estimador, se limita sólo a estimadores insesgados. En la práctica, si estamos frente a un estimador insesgado e impreciso, y otro sesgado, pero preciso, ¿cuál es mejor? Una solución potencial viene del error cuadrático medio, que es una función que define la 'pérdida' asociada con cualquier estimador $\hat{\theta}$. El ECM se define de la siguiente forma:

$$ECM(\hat{\theta}) = E[\hat{\theta} - \theta]^2 \tag{3.39}$$

que incorpora un castigo por sesgo, y además por precisión. Para ver esto, notamos que podemos re-escribir la función de pérdida 3.39 como:

$$\begin{aligned}
ECM(\hat{\theta}) &= E[\hat{\theta} - E(\hat{\theta}) + E(\hat{\theta}) - \theta]^2 \\
&= E\{[\hat{\theta} - E(\hat{\theta})] + [E(\hat{\theta}) - \theta]\}^2 \\
&= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 + 2E[\hat{\theta} - E(\hat{\theta})][E(\hat{\theta}) - \theta] \\
&= E[\hat{\theta} - E(\hat{\theta})]^2 + [E(\hat{\theta}) - \theta]^2 \\
&= \text{var}(\hat{\theta}) + [\text{sesgo}(\hat{\theta})]^2.
\end{aligned}$$

Propiedades Asintóticas En varias estimadores que encontraremos durante el curso, no va a ser posible derivar propiedades en mustras finitas. En este caso, es necesario considerar cómo el estimador se comporta en muestras asintóticas (o muestras grandes). Hablaremos de las siguientes propiedades, que juegan el mismo papel del sesgo y eficiencia en muestras pequeñas:

1. Consistencia
2. Eficiencia asintótica

La consistencia—como el sesgo en muestras finitas—asegura que un estimador $\hat{\theta}$ estará "cerca" del valor θ cuando la muestra sea lo suficientemente grande. Formalmente, decimos que θ_N es consistente si:

$$\lim_{N\to\infty} P(|\hat{\theta}_N - \theta| < \varepsilon) = 1$$

donde ε es un número positivo y pequeño. Generalmente, escribimos lo anterior como:

$$\text{p lim}\, \hat{\theta}_N = \theta.$$

Exigir que un estimador sea insesgado es más exigente que exigir que un estimador sea consistente. Al tener un estimador insesgado, sabemos que $E(\hat{\theta}) = \theta$, incluso con un N pequeño. En el caso de un estimador consistente, sólo sabemos que $\lim_{N\to\infty} E[\hat{\theta}] = \theta$ (es decir, cuando el N va hacia infinito). Sin embargo, muchas veces vamos a tener estimadores que no son insesgados, pero sí consistentes. Un ejemplo de ello es el $\hat{\sigma}$ que estimamos esta clase con MV (más detalles en clase).

Convergencia en Distribución Como vimos en la sección 3.2, cuando el tamaño de una muestra aumenta, aunque la distribución subyacente no es necesariamente regular, la distribución límite sí lo es! Este teorema (el teorema del límite central), nos sirve bastante con estimadores en muestras grandes.

$$\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma^2)$$

Este hecho es muy conveniente cuando tenemos un estimador con una distribución desconocida en muestras pequeñas. Sin embargo, con una muestra grande sabemos:

$$\begin{aligned} \sqrt{N}(\hat{\theta} - \theta) &\dot{\sim} \mathcal{N}(0, \sigma^2) \\ \Rightarrow \hat{\theta} &\dot{\sim} \mathcal{N}(\theta, \sigma^2/N) \end{aligned}$$

donde $\dot{\sim}$ es "aproximadamante distribuida"

Eficiencia Asintótica Como último, tenemos la eficiencia asintótica, que cumple el mismo rol que la eficiencia en muestras pequeñas. Decimos que un estimador $\hat{\theta}$ es relativamente más eficiente

que otro, $\tilde{\theta}$, si se cumplen las siguientes tres propiedades:

1. $\sqrt{N}(\hat{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma_1^2)$
2. $\sqrt{N}(\tilde{\theta} - \theta) \xrightarrow{d} \mathcal{N}(0, \sigma_2^2)$
3. y: $\sigma_2^2 \geq \sigma_1^2$.

3.4 Inferencia

Cuando hablamos de sesgo y eficiencia en la sección anterior, era aparente que un estimador viene con su propia distribución. Cuando tenemos una muestra dada, una determinada estimación siempre será la misma. Pero si tuviésemos otra mustra (representativa) de la misma población, la estimación probablemente será un poco distinta, simplemente por el hecho de contar con una muestra distinta extraída de la población. En la práctica, generalmente contaremos con una sola muestra para estimar parámetros de interés. Pero para reconocer que nuestra estimación viene con su propia distribución, también nos interesaría estimar su varianza. Una vez que hemos estimado un estimador puntual, y además su varianza, podemos hacer enunciados acerca de la probabilidad que el parámetro poblacional (verdadero) caiga en un cierto rango alrededor de nuestro estimador puntual. Nos referimos a este proceso de formar enunciados de probabilidad como Inferencia estadística. La inferencia puede consistir en la formación de intervalos de confianza alrededor de un estimador puntual, o a testear formalmente hipótesis estadísticas. Estudiamos ambos casos en más detalle a continuación.

3.4.1 Estimación de Intervalos

Estimación con Varianza Conocida

Partimos considerando un caso simple. Supongamos que tenemos una muestra aleatoria $Y_1, \ldots, Y_N$ de una población con $\mathcal{N}(\mu, \sigma^2)$ con σ^2 conocida. Y supongamos que nos interesa estimar el parámetro μ. Por ahora no nos preocupamos del por qué σ^2 es conocida. Simplemente notaremos que probablemente no es tan realista pensar que podríamos saber σ^2, así que puede ser un supuesto que queremos eliminar en el futuro! Podemos mostrar que el estimador de máxima verosimilitud para μ es $\hat{\mu} = \sum_{i=1}^{N} Y_i/N$. Dejamos como un ejercicio esta demostración.

Ahora, para considerar la distribución de este estimador, partimos considerando las propiedades del estimador y su varianza. Tenemos que la expectativa del estimador es:

$$E[\hat{\mu}] = \sum_{i=1}^{N} E[Y_i/N] = \sum_{i=1}^{N} \mu/N = N\mu/N = \mu,$$

(que implica que $\hat{\mu}$ es un estimador insesgado), y que la varianza es:

$$Var[\hat{\mu}] = \sum_{i=1}^{N} Var[Y_i/N] = \frac{1}{N^2} \sum_{i=1}^{N} Var[Y_i] = N\sigma^2/N^2 = \sigma^2/N.$$

Dado que cada Y_i es *iid* de una distribución normal, podemos definir la distribución *exacta* del estimador $\hat{\mu}$ como:

$$\mathcal{N}(\mu, \sigma^2/N) \tag{3.40}$$

La ecuación 3.40 es un resultado muy conveniente, ya que víncula por primera vez la distribución del estimador con el parámetro poblacional en sí. Y esto nos permite hacer enunciados de probabilidad acerca del parámetro desconocido μ utilizando el estimador $\hat{\mu}$ (conocido) y la varianza σ^2 (conocida por nuestro supuesto anterior).

Para ver cómo podemos seguir con estos enunciados de probabilidad, partimos con una fórmula conocida de la normal estandarizada:

$$z = \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \sim \mathcal{N}(0, 1) \tag{3.41}$$

Esta variable z es una transformación de cualquier variable aleatoria normal, restando el promedio y dividiendo por su desviación estándar para expresar la variable como una normal con medio 0 y desviación estándar 1. Hacemos esta transformación dada la facilidad de trabajar con la normal estandarizada. La probabilidad de masa de probabilidad bajo la curva de la normal en cada punto se resume en tablas estadísticas como la Tabla 6.1.

Por la naturaleza de la normal estandarizada, podemos calcular la probabildad que z caiga entre cualquier par de valores simétricos $-z_{\alpha/2}$ y $z_{\alpha/2}$:

$$Pr[-z_{\alpha/2} \leq z \leq z_{\alpha/2}] = 1 - \alpha \tag{3.42}$$

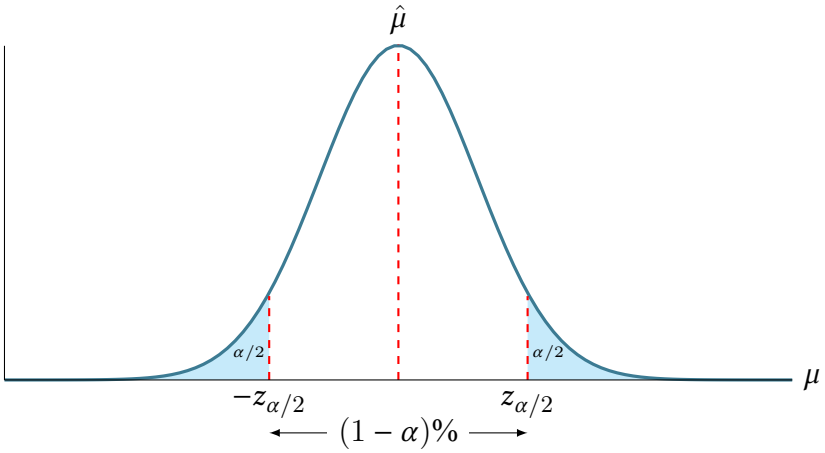
Por ejemplo, utilizando la Tabla 6.1, la probabilidad que z caiga entre -1.96 y 1.96 es de 95%. Este cálculo se realiza sumando la masa de probabilidad que cae *afuera* del rango de interés, como se resume en la Figura 3.14.

Ahora, con un poco de álgebra en las ecuaciones 3.41 y 3.42, tenemos:

$$\begin{aligned} Pr[-z_{\alpha/2} \leq z \leq z_{\alpha/2}] &= 1 - \alpha \\ Pr\left[-z_{\alpha/2} \leq \frac{\hat{\mu} - \mu}{\sigma/\sqrt{N}} \leq z_{\alpha/2}\right] &= 1 - \alpha \\ Pr\left[\hat{\mu} - z_{\alpha/2}(\sigma/\sqrt{N}) \leq \mu \leq \hat{\mu} + z_{\alpha/2}(\sigma/\sqrt{N})\right] &= 1 - \alpha \end{aligned} \tag{3.43}$$

Los puntos finales de la ecuación 3.43 son variables aleatorias ya que $\hat{\mu}$ es una variable aleatoria.

Figure 3.14: Intervalos de la Normal Estándarizada



Entonces el intervalo aleatorio:

$$[\hat{\mu} - z_{\alpha/2}(\sigma/\sqrt{N}), \hat{\mu} + z_{\alpha/2}(\sigma/\sqrt{N})]$$

es un estimador de intervalo, y contiene el parámetro poblacional μ con probabilidad $1 - \alpha$.

Una Aclaración Importante Cuando hablamos de un estimador de intervalo, **No** podemos decir que el intervalo calculado en un caso particular contiene al parámetro verdadero con una probabilidad de $1 - \alpha$. Dependiendo del parámetro poblacional μ, el intervalo de confianza o lo contiene, o no lo contiene. Cuando calculamos intervalos de confianza, estamos haciendo enunciados de probabilidad acerca del estimador del intervalo, y por lo tanto es correcto decir que tenemos un *intervalo de confianza* de $1 - \alpha\%$ para el parámetro poblacional μ.

Estimación con Varianza Desconocida

En la derivación del intervalo de confianza anterior, asumimos que σ era una cantidad conocida. En la práctica, generalmente tenemos que estimar también la cantidad σ. Por ejemplo, si nos interesa estimar la altura promedio de una población, es curioso imaginar que no sabemos el promedio, pero sí sabemos su desviación estándar. En lo que queda de esta sección, consideramos el caso mucho más realista en el que se determina un intervalo de confianza cuando σ es una cantidad desconocida.

En este caso, además de estimar $\hat{\mu}$, necesitamos estimar $\hat{\sigma}$. Cuando estimamos un parámetro adicional, ocupamos un "grado de libertad" adicional. También, si formamos una variable z (parecida a la ecuación 3.42), va a depender de dos variables aleatorias ($\hat{\mu}$ y $\hat{\sigma}$). Una variable $z(\hat{\mu}, \hat{\sigma})$ no va a tener las propiedades convenientes de la $z(\hat{\mu})$ que estamos acustombrado/as a utilizar.

Para partir, notamos una serie de resultados importantes:

1. El estimador de la varianza: $\tilde{\sigma}^2 = \sum_{i=1}^{N}(Y_i - \hat{\mu})^2/N$ será sesgado (más detalles en clases)
2. Sin embargo, $\hat{\sigma}^2 = \sum_{i=1}^{N}(Y_i - \hat{\mu})^2/(N-1)$ es un estimador insesgado para σ^2
3. Una variable que es la suma de k variables aleatorias al cuadrado sigue una distribución chi cuadrado con k grados de libertad (χ^2_k)
4. Y una variable aleatoria normal estandarizada divida por la raíz cuadrada de una variable χ^2_k divida por sus grados de libertad (es decir $\sqrt{(}\chi^2_k/k)$) sigue una distribución t con k grados de libertad

Ahora, sea $Y_1, \ldots, Y_N$ una muestra aleatoria de una población con $\mathcal{N}(\mu, \sigma^2)$. Esta vez necesitmos estimar tanto μ como σ. Tenemos las siguientes estimadores insesgados para μ y σ^2, y la distribución entera para $\hat{\mu}$:

$$\begin{aligned} \hat{\mu} &= \sum_{i=1}^{N} Y_i/N \sim \mathcal{N}(\mu, \sigma^2/N) \\ \hat{\sigma}^2 &= \sum_{i=1}^{N}(Y_i - \hat{\mu})^2/(N-1). \end{aligned}$$

Ahora, si definimos la siguiente variable aleatoria, por el hecho 3 del listado anterior tendrá una distribución chi-cuadrado:

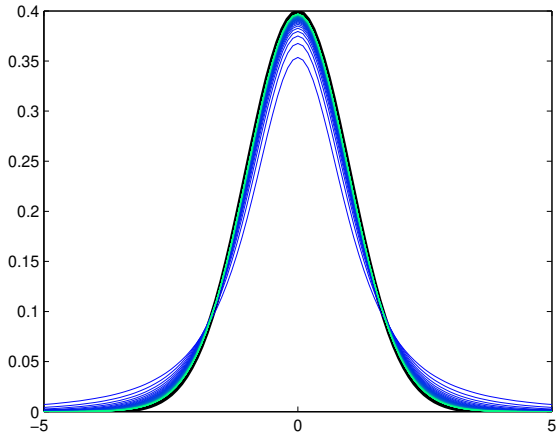
$$\frac{(N-1)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(N-1)}.$$

Con todas esos detalles, podemos derivar una variable aleatoria dependiendo sólo de cantidades observables (o estimables) y el parámetro poblacional de interés que sigue una distribución t de Student:

$$\begin{aligned} t &= \frac{\frac{\hat{\mu}-\mu}{\sigma/\sqrt{N}}}{\left\{\left[\frac{(N-1)\hat{\sigma}^2}{\sigma^2}\right]/(N-1)\right\}^{1/2}} \\ &= \frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{N}} \end{aligned} \tag{3.44}$$

Por el hecho 4, sabemos que t tiene una distribución t con $N-1$ grados de libertad. La distribución t, como la distribución normal estandarizada, tiene una función de densidad de probabilidad estándar que frecuentemente se tabula (ver por ejemplo la Tabla 6.2), y que permite cuantificar de forma muy fácil la masa de probabilidad contenida dentro de cualquier par de valores. Y de hecho, la distribución t se asemaja bastante a la distribución normal estandarizada. En la Figura 3.15, comparamos la función de densidad de la normal estandarizada (la linea negra), y la distribución t de Student variando la cantidad de grados de libertad. Como se observa, la distribución t concentra más masa de probabilidad en las colas de la función de densidad, reconociendo el aumento de la varianza por la estimación de σ^2. La distribución límite de la distribución t (cuando $N \to \infty$), es la distribución normal estandarizada.

Figure 3.15: Distribución Normal versus Distribución t con k grados de libertad



Nota: Curvas corresponden a funciones de densidad para la normal estándar (linea negra) y distribuciones t con distintas cantidad de libertad (lineas azules/verdes). Lineas máz azules tienen una cantidad menor de grados de libertad.

A partir de la ecuación 3.44 ahora por fin podemos derivar un estimador de intervalo con μ y σ desconocida:

$$\begin{aligned} Pr[-t_{(N-1,\alpha/2)} \leq t \leq t_{(N-1,\alpha/2)}] &= 1-\alpha \\ Pr\left[-t_{(N-1,\alpha/2)} \leq \frac{\hat{\mu}-\mu}{\hat{\sigma}/\sqrt{N}} \leq t_{(N-1,\alpha/2)}\right] &= 1-\alpha \\ Pr\left[\hat{\mu} - t_{(N-1,\alpha/2)}(\hat{\sigma}/\sqrt{N}) \leq \mu \leq \hat{\mu} + t_{(N-1,\alpha/2)}(\hat{\sigma}/\sqrt{N})\right] &= 1-\alpha. \end{aligned}$$

Y esto nos da nuestro estimador de intervalo:

$$\hat{\mu} \pm t_{(N-1,\alpha/2)}(\hat{\sigma}/\sqrt{N})$$

de $(1 - alpha)\%$ para parámetro desconocido μ.

3.4.2 Contrastes de Hipótesis

La lógica del intervalo de confianza es encontrar un rango donde, si se replica el experimento subyacente infinitas veces, en $1-\alpha\%$ de las replicaciones el parámetro verdadero caerá en el intervalo. Sin embargo, existen otras maneras de considerar un parámetro estimado con incertidumbre. Uno que encontraremos en varios momentos de nuestro curso son los contrastes de hipótesis. Un contraste estadístico es un problema de decisión acerca de un parámetro θ que está contenido en el conjunto Ω. Este espacio Ω puede estar particionado en dos distintos (y mutuamente excluyentes) sub-espacios Ω_0 y Ω_1. Utilizando nuestra muestra de datos, tenemos que decidir si θ pertenece a

Ω_0 o Ω_1. Dado que son espacios mutuamente exluyentes, el parámetro tiene que pertenecer a uno, y sólo uno de los espacios.

Los contrastes de hipótesis parten de la base de dos hipótesis:

1. **La Hipotesis Nula**: Denotamos a H_0 como la hipótesis nula que $\theta \in \Omega_0$
2. **La Hipotesis Alternativa**: Denotamos a H_1 como la hipótesis alternativa que $\theta \in \Omega_1$.

Y el test estadístico (o contraste de hipótesis) consiste entonces en rechazar una hipótesis (y "'aceptar"' la otra) tomando en cuenta (a) Los costos de una decisión incorrecta, y (b) Todos los datos disponibles.

Suponemos que tenemos una muestra aleatoria de datos $\boldsymbol{Y}$ que tiene fdp conjunta $f(\boldsymbol{y}|\boldsymbol{\theta})$. El conjunto de todos los valores posibles de $\boldsymbol{Y}$ es el espacio muestral del experimento. Definimos un procedimiento de prueba que divide el espacio en dos partes: uno conteniendo todos los valores de $\boldsymbol{Y}$ que hacen que no se rechace H_0, y el otro donde H_1 será nuestra conclusión (y H_0 rechazada). El segundo conjunto se llama la **región de rechazo** del test, ya que nos hará rechazar la hipótesis nula. Este espacio de $\boldsymbol{Y}$ es de dimensión N, y por lo general será muy complejo abarcar un contraste de hipótesis considerando todas las observaciones por separado. Por lo tanto, cuando realizamos un contraste de hipótesis, buscamos formar un **estadístico de prueba**, que reduce este espacio de N dimensiones en un valor escalar en $\mathbb{R}$. Por definición un estadístico de prueba tiene una distribución conocida bajo la hipótesis nula, y la región de rechazo es el conjunto de valores del estadístico de prueba para las cuales se rechzará la hipótesis nula.

Los Elementos Básicos de un Test de Hipótesis consisten de:

1. Una hipótesis nula, que será tratada como cierta hasta que haya evidencia de lo contrario
2. Una hipótesis alternativa que se adoptará si la nula es rechazada
3. Un estadístico de prueba
4. Una región de rechazo (o "valor crítico")

La hipótesis nula es, de cierta forma, nuestra posición *ex ante*. Es la posición que mantenemos si no encontramos evidencia que sugiere lo contrario. La idea de tener una hipótesis que se supone que se cumple hasta encontrar evidencia en contra es análga al principio jurídico de la presunción de inocencia. En econometría, generalmente la nula considera si un parámetro (o parámetros) toma un valor (o valores) específico(s), que con frecuencia es 0. En este caso, la alternativa es simplemente que el (los) parámetro(s) no es (son) igual al valor. Ejemplos comunes incluyen contrastes de un solo parámetro: $H_0 : \theta = 0$, $H_1 : \theta \neq 0$; contrastes de una combinación lineal de parámetros: $H_0 : \theta_1 + \theta_2 = 1$, $H_1 : \theta_1 + \theta_2 \neq 1$; contrastes acerca de varios parámetros $H_0 : \theta_1 = 0$ y $\theta_2 = 0$, H_1 : por lo menos uno de los parámetros no es igual a cero, o contrastes en base a desigualdades: $H_0 : \theta \geq 0$, $H_1 : \theta < 0$.

Para ilustrar la idea de un test de hipótesis, examinamos un caso particular. Más adelante en el curso, veremos otros ejemplos cuando lleguemos a analizar modelos de regresión. Imaginemos que queremos evaluar una hipótesis acerca de un promedio desconocido μ de una población con distribución normal y varianza conocida de $\sigma^2 = 10$. Como vimos anteriormente, aunque el supuesto de varianza conocida es bastante poco creíble, se puede eliminar el supuesto sin tantos problemas. En este caso, formamos nuestra hipótesis nula y alternativa:

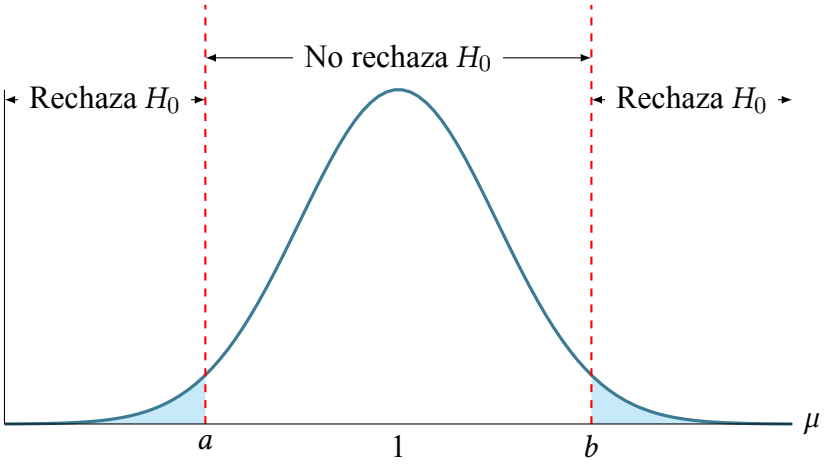
$$H_0 : \mu = 1$$
$$H_1 : \mu \neq 1$$

Aquí elegimos el valor 1 de forma arbitraria, pero por lo general, cuando realizamos un test de hipótesis la nula tiene un vínculo a alguna base teórica. E imaginemos que además tenemos una muestra de tamaño $N = 10$: $y_1, \ldots, y_{10}$. Dada esta muestra, sabemos que:

$$\hat{\mu} = \sum_{i=1}^{N} Y_i/N \sim \mathcal{N}(\mu, \sigma^2/N) = \mathcal{N}(\mu, 1)$$

Bajo nuestra hipótesis nula, es decir *imponiendo* que la nula es cierta, tenemos una distribución para $\hat{\mu}$ de la forma presentada en la Figura 3.16. Con la distribución (bajo la nula) en mano, tenemos que

Figure 3.16: La Distribución de $\widehat{\mu}$ bajo H_0



elegir valores para a o b que sugieren que debemos rechazar la nula. Estos valores son elegidos con base en que si la nula fuese cierta, parece bastante poco probable observar un valor tan extremo. Por ejemplo, si observamos un valor de $\hat{\mu} = 100$, parece muy poco probable que este valor haya salido de la distribución en la Figura 3.16, aunque teóricamente *es* probable. Además, los valores a y b dependen de nuestro nivel de significancia deseado, α. Por ejemplo, si pensamos que debemos exigir mucha evidencia para poder rechazar la nula, vamos a fijar a y b de tal modo que haya poca masa de probabilidad en las colas de la distribución. Aquí, elegimos a y b de tal modo que:

$$Pr[\hat{\mu} \leq a] = Pr[\hat{\mu} \geq b] = \alpha/2$$

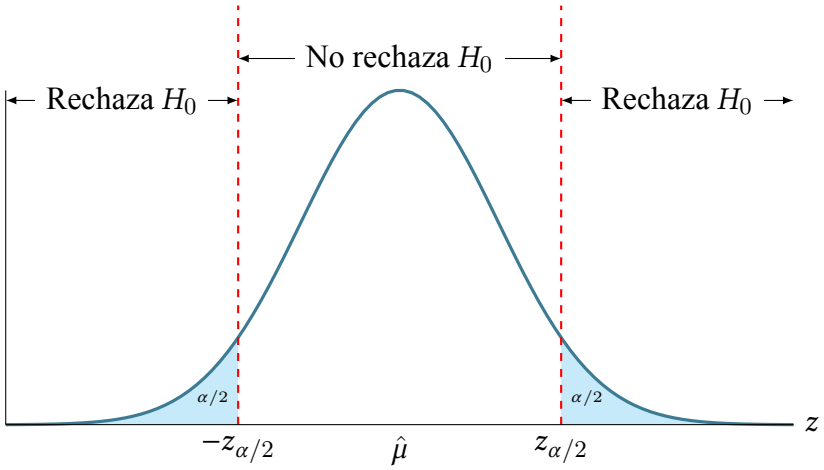
Y, definimos nuestra regla de rechazo:

- Rechaza H_0 si $\hat{\mu} \leq a$ o $\hat{\mu} \geq b$
- No se rechaza (o se"'acepta"') H_0, si $a < \hat{\mu} < b$

Por último, definimos nuestra estadístico de prueba. En el caso de este ejemplo, trabajamos con la distribución normal estandarizada que conocemos bastante bien:

$$z = \frac{\hat{\mu} - 1}{\sigma/\sqrt{N}}$$

y: $z \sim \mathcal{N}(0, 1)$ si H_0 es cierto. Ahora, utilizando nuestro valor para α, calculamos los valores que corresponden a z...



Por ejemplo, si fijamos $\alpha = 0.05$, podemos buscar los valores críticos en la Tabla estadística 6.1, encontrando valores críticos de ± 1.96.

¿Por qué rechazamos si $|z| > z_{\alpha/2}$? Dada la naturaleza de una distribución normal, nunca vamos a poder asegurar con certeza que un parámetro no es igual a algún valor. Pero aquí, si rechazamos la nula, implica uno de dos casos. Si la nula es cierta, la probabilidad de obtener un valor de z en la región de rechazo es sólo α. Y dado que este evento es poco probable (elegimos un α pequeño), y por lo tanto, concluimos que el estadístico de prueba probablemente no tiene una distribución $\mathcal{N}(0, 1)$.

Tipos de Errores en un Contraste de Hipótesis Cuando hacemos un test de hipótesis, idealmente rechazamos la hipótesis nula cuando la nula es falsa, y no rechazamos la nula cuando la alternativa sea falsa. Definimos una funcion $\Pi(\theta)$ donde:

$$\Pi(\theta) = Pr[\text{rechazar } H_0 | \theta]$$

La función $\Pi(\theta)$ es conocido como la función de potencia. Idealmente, tendremos que $\Pi(\theta) = 0 \; \forall \; \theta \in \Omega_0$ y $\Pi(\theta) = 1 \; \forall \; \theta \in \Omega_1$. Sin embargo, generalmente no existen test de hipótesis ideales. En particular, hay dos tipos de errores que podríamos cometer:

Error Tipo I: La probabilidad de rechazar H_0 cuando H_0 es verdadera.

$$P[\text{Error Tipo I}] = P[\text{rechazar } H_0 | H_0 \text{ es cierto}] \leq \alpha$$

Aquí α se conoce como el nivel de significancia del test, y α es el valor más grande de $\Pi(\theta)$ para cualquier valor de θ.

Error Tipo II: La probabilidad de no rechazar una hipotesis falsa.

$$\begin{aligned} \beta &= P[\text{Error Tipo II}] = P[\text{no rechazar} H_0 | H_1 \text{ es cierto}] \\ &= 1 - \Pi(\theta) \quad \text{para} \quad \theta \in \Omega_1 \end{aligned}$$

No existen test de hipótesis para eliminar (o hacer arbitrariamente pequeño) ambos tipos de errores. Generalmente α está definido como un valor fijo (y pequeño) ya que un error de tipo I es más grave que un error de tipo II.

3.4.3 Test de Razón de Verosimilitudes

Por último, antes de terminar esta sección, consideramos el test de Razón de Verosimilitudes. Éste es una clase generalizada de test de hipótesis, que viene de la estimación por máxima verosimilitud. El proceso general de este test es:

1. Estimar un parámetro (o parámetros) utilizando máxima verosimilitud
2. Estimar el modelo utilizando máxima verosimilitud, pero con la restricción definida por la test de hipótesis
3. Comparar los valores de las dos cantidades de máxima verosimilitud
4. Si el valor de 1 es muy distinto al valor de 2, es probable que el parámetro restringido no es el valor correcto para el parámetro poblacional

Imaginemos que estamos intersado/as en estimar un modelo con un solo parámetro μ. Queremos testear la hipótesis:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

donde μ_0 es simplemente un valor particular. Entonces, la idea del test de razón de verosimilitudes es que debemos maximizar dos funciones de verosimilitud: una función no restringida, que nos da el estimador $\hat{\mu}_{ML}$, y otra función restringida, que restringimos para dar el valor μ_0. Estos valores

estimados traen consigo un valor de la función de verosimilitud, $\ell(\hat{\mu}_{ML})$ y $\ell(\mu_0)$ respectivamente. ¿Qué podemos decir acerca de los dos valores? $\ell(\hat{\mu}_{ML}) \overset{?}{\lesseqgtr} \ell(\mu_0)$

La lógica del test es que si las funciones $\ell(\hat{\mu}_{ML})$ y $\ell(\mu_0)$ son muy distintas, haber impuesto que $\mu = \mu_0$ fue una restricción muy fuerte, y probablemente no tan realista dado los datos observados. Sin embargo, si $\ell(\hat{\mu}_{ML})$ y $\ell(\mu_0)$ son muy parecidas, es razonable pensar que el valor del parámetro podría ser μ_0. La única parte que queda es saber cómo formar el estadístico de prueba. El test de razón de verosimilitudes sugiere utilizar:

$$2\left[\ell(\hat{\mu}_{ML}) - \ell(\mu_0)\right] \sim \chi^2_{(1)}$$

Y ahora, con la distribución para el estadístico de prueba (una distribución $\chi^2_{(k)}$ donde k es la cantidad de restricciones impuestas por la hipótesis nula), podemos calcular la región de rechazo para cualquier test dado los valores calculados para cada ℓ, una tabla estadística de la distribución χ^2 y el nivel de significancia deseado (α).

Ejercicios de Ayudantía:

1. Números Aleatorios y Probabilidad:

   (a) Encuentre un número aleatorio de una distribución normal $(0, 1)$ y $(150, 30)$.

   (b) Encuentre un número aleatorio de una distribución uniforme $(0, 1)$ y $(0, 5)$.

   (c) Encuentre el mismo número aleatorio que sus compañeros en una distribución normal $(0, 1)$.

   (d) Genere una variable aleatoria z que distibuya $normal(80, 16)$ y tenga 150.000 observaciones. Grafique con un histograma que compare con la distribución normal y exporte.

   Asuma varianza desconocida:

   (e) Contraste si la media de la variable z es 50 con un 95% de confianza. ¿Le sorprenden los resultados? ¿Por qué?

   (f) Compruebe que efectivamente la media de la variable z es 80 con un 99% de confianza.

   (g) Construya un intervalo de confianza para la media variable z con un 95% de confianza.

   (h) Contraste si la varianza de la variable z es mayor de 10^2 con un 90% de confianza.

2. Repaso Herramientas Probabilísticas:
   Demuestre:

   (a) $P(\emptyset) = 0$

   (b) $P(A) \leq 1$

   (c) $P(A^c) = 1 - P(A)$

   (d) Si $A \subset B \Rightarrow P(A) \leq P(B)$

3. Trabajando con la normal estandarizada $\mathcal{N}(0, 1)$ y la tabla 6.1:

   (a) Encuentra la masa de probabilidad de la normal estandarizada que cae *abajo de 1.0*.

   (b) Encuentra la masa de probabildad de la normal estandarizada que se ubica entre -1.0 y 1.0.

   (c) ¿Abajo de qué valor se ubica 5% de la masa de probabilidad?

   (d) ¿Entre cuáles valores $-a, a$ se encuentra 99% de la masa de probabilidad?

   (e) Responde a los ejercicios (a) y (d), pero ahora en vez de referir a los valores de la normal estandarizada, encuentran los valores de una variable normal $\mathcal{N}(400, 50)$.

4. Responda las siguientes preguntas:

   (a) Queremos analizar la efectividad de una evaluación para aprobar un curso y saber si es común aprobar sin haber estudiado. Sabemos que el 90% de los estudiantes que estudian y el 5% de los que no lo hacen aprueban el examen y adicionalmente, sabemos que 30 de los 45 alumnos de la clase estudian para este. En base a esta información, si un estudiante aprueba, ¿Cuál es la probabilidad que no se haya esforzado y estudiado para el examen?

   (b) Con la siguiente información compruebe que la "'Ley de las esperanzas iteradas se cumple"':

   Promedio General

Variable	Obs	Mean	Std. Dev.	Min	Max
prom_gral	3238586	5.14507	1.774596	0	7

   Promedio General Establecimientos Municipales

Variable	Obs	Mean	Std. Dev.	Min	Max
prom_gral	1264581	4.943402	1.916149	0	7

   Promedio General Establecimientos Subvencionados

Variable	Obs	Mean	Std. Dev.	Min	Max
prom_gral	1679366	5.207575	1.684732	0	7

   Promedio General Establecimientos Pagados

Variable	Obs	Mean	Std. Dev.	Min	Max
prom_gral	245935	5.832907	1.375029	0	7

   Promedio General Corporaciones

Variable	Obs	Mean	Std. Dev.	Min	Max
prom_gral	48704	4.752792	1.710224	0	7

   (c) Construya un intervalo de confianza del 95% para la esperanza del promedio general por dependencia.

   (d) Contraste si la esperanza del promedio general es mayor a 5.0 con un 1% de significancia

   (e) Contraste si la varianza del promedio general es igual a 3 con un nivel de significancia del 5%.

Sección 4

Introducción al Modelo de Regresión Lineal

Nota de Lectura: Todos los textos de econometría que utilizamos en este curso dan una presentación del Modelo de Regresión Lineal en alguna forma. Algunos textos lo hacen de manera bastante concisa (por ejemplo Cameron and Trivedi (2005)). La presentación en Rao (1973) ofrece una serie de demostraciones extremadamente elegantes de los resultados principales discutidos aquí. Además de estos dos libros, se puede encontrar una versión de los resultados discutidos en esta sección en Greene (2002), Wooldridge (2002), Hansen (2017) y Goldberger (1991) (entre otros).

Un elemento central de la literatura en econometría empírica es el modelo lineal. Además, este modelo proporciona una muy buena base para entender otros modelos más complejos o con supuestos menos exigentes. El modelo lineal tiene varias caractertísticas muy convenientes que explican su frecuencia de uso, pero como veremos más adelante, también hay situaciones en las que plantear un modelo lineal *no* es lo más apropiado. El modelo de regresión lineal se define según:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i \tag{4.1}$$

Donde tenemos una variable de interés y_i para la observación i, una serie de K variables explicativas $x_{1i}, x_{2i}, \ldots, x_{Ki}$, y una muestra de N observaciones con indice $i = 1, 2, \ldots, N$. El término u_i en este modelo se conoce como el término de error y captura el efecto combinado de las influencias adicionales de otras variables aparte de $x_{1i}, x_{2i}, \ldots, x_{Ki}$, sobre la variable y_i, y además, cualquier impacto no-lineal de las K variables incluidas. Los supuestos acerca de este componente de error serán de importancia fundamental en nuestra consideración del modelo de regresión lineal.

Por ejemplo, si nos interesa determinar cuáles son algunos de los factores que explican la gran variación en el salario que las personas reciben en el mercado laboral, podríamos plantear un modelo donde nuestra variable de interés (y_i) es el salario mensual de persona i, y donde nuestras variables explicativas son la cantidad de años de escolaridad de la persona y su experiencia laboral

en años. Dado que hay mucha variación en el salario de las personas incluso manteniendo fijo su escolaridad y experiencia laboral, el término u_i captura esta heterogeneidad individual.

En la ecuación 4.1 hemos planteado un efecto lineal de cada x_{ki} con $k \in 1, \ldots, K$ sobre y_i (y por esto el modelo se conoce como el "modelo lineal"). Esto se observa con la definición de la serie de parámetros $\beta_1, \ldots, \beta_K$. Estos son valores escalares, e implican que—todo lo demás constante– para un aumento en una unidad de alguna x_k, la variable y_i cambia en β_k unidades en promedio. Notemos que la linealidad aquí refiere al parámetro β_k a lo largo de la distribución de la variable x_{ki}. Esta estructura impone que un aumento de x_{ki} entre 0 y 1 unidad tiene el mismo impacto sobre y_i que un aumento entre 1.000 y 1.001 unidades.

Aunque esta definición lineal de los parámetros parece ser muy restrictiva para muchas circunstancias, la especificación lineal no es tan restrictiva como inicialmente puede parecer. La parte restrictiva es que los parámetros $\beta_1, \beta_2, \ldots, \beta_K$ entran de manera lineal, y el término de error u_i entra de manera aditiva. Sin embargo, las variables (dependiente e independientes) en sí pueden ser transformaciones no-lineales. Por ejemplo, $y_i = \ln(y_i)$ o $x_{3i} = x_{2i}^2$. En este sentido, aunque el modelo es "lineal", la relación capturada entre las variables originales x_{ki} y y_i no tiene que serlo.

Por lo general, el modelo lineal incluye un término de intercepto, además de la relación entre las variables x_{ki} y y_i. En el modelo 4.1 si la variable $x_{1i} = 1$ para cada i, el modelo se puede re-escribir como:

$$y_i = \beta_1 + \beta_2 x_{2i} + \cdots + \beta_K x_{Ki} + u_i, \tag{4.2}$$

donde ahora β_1 es el término de constante, o el intercepto de un hiperplano de K dimensiones. Cuando volvemos a hablar de la interpretación de los términos β en la sección ABC, veremos que a menudo este intercepto tiene una implicancia directa del fenónemo de interés en la regresión. Antes de seguir, es importante señalar que esta notación para el intercepto (β_1) no es la única utilizada en libros de texto de econometría. Algunos autores (por ejemplo Wooldridge (2002)) denotan el intercepto como β_0, y los demás parámetros parten en β_1. Más allá que en decisión de nomenclatura, este no tiene ninguna implicancia para la interpretación o presentación del modelo de regresión.

Notación Denotamos a x_i y β como vectores de columna de $K \times 1$. Aquí el x_i refiere al conjunto de valores para las distintas *variables independientes* x_k para cada observación i, mientras que β son comúnes entre observaciones. Estos vectores son:

$$x_i = \begin{pmatrix} x_{1i} \\ x_{2i} \\ \vdots \\ x_{Ki} \end{pmatrix} \text{ y } \beta = \begin{pmatrix} \beta_1 \\ \beta_2 \\ \vdots \\ \beta_K \end{pmatrix}.$$

En este formato vectorial, podemos escribir el modelo como:

$$y_i = x_i'\beta + u_i \text{ para } i = 1, \ldots, N, \tag{4.3}$$

donde observamos que cada y_i es ahora el resultado de sumar $K + 1$ elementos. Estos elementos son las K variables independientes multiplicados por los parámetros β, más el término de error u_i. En algnunos casos, se escribe el modelo sin el subíndice i, y entonces:

$$y = \beta_1 x_1 + \beta_2 x_2 + \cdots + \beta_K x_K + u$$

o, para una observación típica:

$$y = x'\beta + u.$$

En estos apuntes, generalmente escribiremos el subíndice i cuando referimos al modelo al nivel de cada observación.

Por último, podemos juntar cada una de las $i = 1, \ldots, N$, observaciones para escribir el modelo en conjunto. En formato matricial para las N observaciones escribimos:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad X = \begin{pmatrix} x_1' \\ x_2' \\ \vdots \\ x_N' \end{pmatrix} = \begin{pmatrix} x_{11} & x_{12} & \cdots & x_{1K} \\ x_{21} & x_{22} & \cdots & x_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ x_{N1} & x_{N2} & \cdots & x_{NK} \end{pmatrix}, \quad u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}.$$

Aquí, y y u son vectores de columna de $N \times 1$ y X es una matriz de $N \times K$. En este formato es claro que cada una de las N filas de las 3 matrices corresponde a los casos descritos en ecuación 4.3. Con esto, podemos escribir el modelo de regresión lineal en forma matricial como:

$$y = X\beta + u, \tag{4.4}$$

donde seguimos la notación típica de tener matrices con letras mayúsculas (X), y vectores en letras minúsculas (y, u).

4.1 Mínimos Cuadrados Ordinarios

4.1.1 Planteando el Estimador de MCO en un Modelo Lineal

La ecuación 4.1 es parte de un modelo estadístico. Asumimos que alguna variable de interés y se explica por una serie de k variables x_k, y un error no observado. Pero nuestro interés en econometría es encontrar los coeficientes asociados con nuestros modelos estadísticos.

Asumiendo que tenemos una muestra que representa a la población de interés, cuando contamos con un modelo de la forma de 4.1, es probable que queramos estimar el vector de parámetros β

(entre otras cosas) que define la relación entre las variables x_k y la variable y. El estimador de Mínimos Cuadrados Ordinarios (MCO) es la metodología más común que se utiliza para estimar el vector de parámetros desconodidos β desde los datos de y_i y $x_{1i}, x_{2i}, \ldots, x_{Ki}$ para una muestra de $i = 1, \ldots, N$ observaciones.

MCO es un estimador, y por ende enfrenta las mismas consideraciones que revisamos en sección 3.3. En términos particulares, el objetivo del estimador MCO es encontrar el vector de parámetros β que minimiza la suma de los errores al cuadrado: $u_i = y_i - x_i'\beta = u_i(\beta)$. Formalmente[1]:

$$\begin{aligned} \widehat{\beta} &= \arg\min_{\beta} \sum_{i=1}^{N} u_i^2 = \sum_{i=1}^{N} (y_i - x_i'\beta)^2 \\ &= \arg\min_{\beta} \; u'u = (y - X\beta)'(y - X\beta). \end{aligned}$$

Definimos como μ a la función objetivo $\mu = \sum_{i=1}^{N} u_i(\beta)^2$. La función μ es una *función de pérdida*, ya que en la medida que esta función aumenta, nos alejamos del estimador elegido, y si aumenta, consideramos al β probado como una peor opción para el eventual $\widehat{\beta}$. Ésta función de pérdida de error al cuadrado castiga valores grandes del término de error, y tolera valores pequeños. Por ende, el estimador MCO encuentra un valor del vector de parámetros β que evita los errores grandes. Un costo de este estimador es que el parámetro estimado puede ser muy sensible a la inclusión o exclusión de 'outliers' (observaciones con valores muy alejados de la media en por lo menos una variable).

El estimador de MCO no es el único estimador que se podría considerar para encontrar el vector de parámetros β en el modelo 4.1. Como cualquier estimador, es simplemente una regla para calcular una cantidad en base a datos observados. Pero como veremos más adelante en este capítulo (sección 4.2.2), este estimador tiene buenas propiedades que explican su frecuencia de uso.

Podemos resolver este estimador de forma algebraica. Para obtener el estimador MCO, minimizamos $\mu = \sum_{i=1}^{N} u_i^2$ con respecto a β donde $u_i(\beta) = y_i - \beta_1 x_{1i} - \cdots - \beta_K x_{Ki}$. Derivando μ con respecto a alguna β_k tenemos:

$$\frac{\partial \mu}{\partial \beta_k} = \sum_{i=1}^{N} 2u_i \left(\frac{\partial u_i}{\partial \beta_k} \right) = \sum_{i=1}^{N} -2u_i x_{ki} \text{ para } k = 1, \ldots, K. \tag{4.5}$$

Entonces, para encontrar el punto mínimo:

$$\frac{\partial \mu}{\partial \beta_k} = 0 \leftrightarrow \sum_{i=1}^{N} x_{ki} u_i = 0 \text{ para } k = 1, \ldots, K. \tag{4.6}$$

[1] Y recordamos que $\arg\min_x f(x)$ refiere al valor de x que minimiza $f(x)$, a diferencia de $\min_x f(x)$ que refiere al valor de $f(x)$ en su punto mínimo.

Más adelante demostraremos que este punto es, en realidad, un mínimo global, y no un máximo.

Las Condiciones de Primer Orden La ecucación 4.6 da las condiciones de primer orden (CPO) del problema de minimización. Estos CPO son: $\sum_{i=1}^{N} x_{ki}u_i = 0$ para $k = 1, \ldots, K$, y se las puede escribir de manera resumida como $X'u = 0$:

$$X'u = \begin{pmatrix} x_{11} & x_{21} & \cdots & x_{N1} \\ \vdots & \vdots & \ddots & \vdots \\ x_{1K} & x_{2K} & \cdots & x_{NK} \end{pmatrix} \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_K \end{pmatrix} = \begin{pmatrix} \sum_{i=1}^{N} x_{1i}u_i \\ \vdots \\ \sum_{i=1}^{N} x_{Ki}u_i \end{pmatrix} = \begin{pmatrix} 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Aquí, X' es una matriz de $K \times N$, u es un vector de $N \times 1$, y entonces, $X'u$ es un vector de $K \times 1$. Utilizando las CPO que dan $X'u = 0$, y el hecho de que $u = y - X\beta$, obtenemos:

$$\begin{aligned} X'(y - X\beta) &= 0 \\ \leftrightarrow X'y - X'X\beta &= 0. \end{aligned} \tag{4.7}$$

Esto es un sistema de K ecuaciones para los K elementos desconocidos del vector β. Si $(X'X)^{-1}$ existe, el vector de parámetros β que minimiza $\mu = \sum_{i=1}^{N} u_i^2$ satisface:

$$\beta = (X'X)^{-1}X'y,$$

dando el estimador MCO para el vector de parámetros β como:

$$\begin{aligned} \widehat{\beta}_{MCO} &= (X'X)^{-1}X'y \qquad (4.8) \\ &= \left(\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\sum_{i=1}^{N} x_i y_i\right) \\ &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i'\right)^{-1} \left(\frac{1}{N}\sum_{i=1}^{N} x_i y_i\right). \qquad (4.9) \end{aligned}$$

La penúltima línea simplemente reemplaza $X'X$ y $X'y$ por sus respectivas fórmulas sobre las N observaciones de la muestra, y la última línea agrega el factor común $\frac{1}{N}$, que no tiene ningún impacto sobre $\widehat{\beta}_{MCO}$.

Este cálculo de minimización de los errores al cuadrado (de la ecuación 4.6) también puede proceder en forma matricial. Consideremos el problema: $\frac{\partial u'u}{\partial \beta}$ donde $u = (y - X\beta)$ y β refieren a

los vectores de $N \times 1$ y $K \times 1$ respectivamente. Entonces:

$$\begin{aligned}\frac{\partial u'u}{\partial \beta} &= \frac{\partial}{\partial \beta}(y - X\beta)'(y - X\beta) \\ &= \frac{\partial}{\partial \beta}(y'y - \beta'X'y - y'X\beta + \beta'X'X\beta) \\ &= \frac{\partial}{\partial \beta}(y'y - 2\beta'X'y + \beta'X'X\beta) \\ &= -2X'y + 2X'X\beta, \end{aligned} \tag{4.10}$$

donde el penúltimo paso viene del hecho de que la traspuesta de un escalar es el mismo escalar ($y'X\beta = (y'X\beta)' = \beta'X'y$), y el último paso viene de las reglas de derivación de matrices que implican que $\partial\beta'X'X\beta/\partial\beta = \partial\beta'A\beta/\partial\beta = 2A\beta = 2X'X\beta$ (refiere a Greene (2002, p. 839)). Para encontrar el mínimo de $u'u$, simplemente igualamos a la derivada a 0:

$$\begin{aligned} -2X'y + 2X'X\beta &= 0 \\ \Rightarrow X'X\beta &= X'y \end{aligned} \tag{4.11}$$

y siempre cuando $X'X$ es invertible, tenemos que[2]:

$$\beta = (X'X)^{-1}X'y.$$

Esta solución es la solución a un proceso de minimización no constreñida, ya que el vector de parámetros β puede tomar cualquier valor.[3]

Una motivación para el estimador MCO en los modelos lineales es que si la expectativa $E(y|X)$ es lineal, los valores predichos:

$$\hat{y} = X\hat{\beta}_{MCO}$$

son las predicciones óptimas de y que se puede construir de las observaciones de X en el sentido de minimizar la pérdida del residuo al cuadrado, es decir $\sum_{i=1}^{N}(y_i - \hat{y})^2$.

Clase Computacional: Encontrando el Vector de parámetros Las secciones anteriores sugieren que existen (por lo menos) dos maneras de encontrar el vector de parámetros $\hat{\beta}$. Uno es minimizando la función $\mu = \sum_{i=1}^{N} u_i(\beta)^2 = u'u$, y el otro es mediante la expresión matricial $\hat{\beta} = (X'X)^{-1}X'y$.

1. Con una base de datos, encuentra, en Mata (o algún otra idioma matricial), el vector de parámetros $\hat{\beta}$ utilizando cada metodo.
2. Estima los parámetros utilizando el comando `regress` de Stata.

[2] Para un repaso del proceso de inversión, refiere a la ecuación 2.11, notando que $A = X'X$ es una matriz de $K \times K$, $b = X'y$ es un vector de $K \times 1$, y $x = \beta$ es una matrix de $K \times 1$.

[3] Para una discusión de MCO con minimización constreñida (eg limitando los posibles valores para β), refiere a Rao (1973, pp. 231-232).

El Caso Simple Bivariado Un caso especial es el caso simple de un modelo lineal con un intercepto y una única variable explicativa:

$$y_i = \beta_1 + \beta_2 x_{2i} + u_i.$$

Por ejemplo, consideramos la relación del peso al nacer de los bebés y la edad de su madre. Si nos interesaba estudiar como el peso de los bebés varía con la edad de su madre, una manera de considerarlo sería mediante la regresión linea bivariada:

$$peso_i = \beta_1 + \beta_2 edad_i + u_i. \tag{4.12}$$

En este caso los coeficientes nos describen cómo el peso de un bebé (en gramos) varía con la edad de su madre (en años). Aunque volveremos a revisar la interpretación de los coeficientes en los modelos de regresión en más produndidad en la sección 4.3 de este capítulo, este modelo nos indica que al aumentar 1 año la edad de la madre, el peso del bebé en promedio cambia en β_2 gramos. Y cuando el edad de la madre es igual a cero (una situación sin una interprteación real), el peso promedio del bebé sería de β_1 gramos.

Para ver cómo son los parámetros β_1 y β_2 cuando se estima mediante MCO, volvemos a la ecuación 4.6. Para el modelo bivariado, hay dos CPO: una para $x_{1i} = 1$ (el constante) y otra para x_{2i}. Éstas CPO son:

$$\sum_{i=1}^{N}(y_i - \beta_1 - \beta_2 x_{2i}) = 0 \tag{4.13}$$

$$\sum_{i=1}^{N} x_{2i}(y_i - \beta_1 - \beta_2 x_{2i}) = 0 \tag{4.14}$$

Con un poco de álgebra en la ecuación 4.13, tenemos que

$$\sum_{i=1} y_i - N\beta_1 - N\beta_2 \sum_{i=1} x_{2i} = 0 \leftrightarrow \beta_1 = \frac{1}{N}\sum_{i=1} y_i - \beta_2 \frac{1}{N}\sum_{i=1} x_{i2} = \bar{y} - \beta_2 \bar{x}_2.$$

Y sustituyendo esta solución en 4.14 produce:

$$\sum_{i=1}^{N} x_{2i}(y_i - \bar{y} + \beta_2 \bar{x}_2 - \beta_2 x_{2i}) = 0 \leftrightarrow \sum_{i=1}^{N} x_{2i}(y_i - \bar{y}) = N\beta_2 \sum_{i=1}^{N} x_{2i}(x_{2i} - \bar{x}_2)$$

$$\leftrightarrow N\beta_2 = \frac{\sum_{i=1}^{N} x_{2i}(y_i - \bar{y})}{\sum_{i=1}^{N} x_{2i}(x_{2i} - \bar{x}_2)} \leftrightarrow \beta_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_{2i} - \bar{x}_2)(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_{2i} - \bar{x}_2)^2}.$$

Donde aquí el último paso viene del hecho general que $\sum_{i=1}^{N}(a_i - \bar{a})(b_i - \bar{b}) = \sum_{i=1}^{N} a_i(b_i - \bar{b}) - \bar{a}\sum_{i=1}^{N}(b_i - \bar{b}) = \sum_{i=1}^{N} a_i(b_i - \bar{b})$ dado que $\sum_{i=1}^{N}(b_i - \bar{b}) = 0$.

Resumiendo, las soluciones al estimador MCO para los parámetros β_1 y β_2 son, respectiva-

mente:

$$\hat{\beta}_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_{2i} - \bar{x}_2)(y_i - \bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_{2i} - \bar{x}_2)^2} \tag{4.15}$$

$$\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}_2 \tag{4.16}$$

donde $\hat{\beta}_1$ y $\hat{\beta}_2$ son las estimaciones de MCO del intercepto y la pendiente, y $\bar{y}$ y $\bar{x}_2$ son las medias muestrales de y_i y x_{2i}. Aquí agregamos el gorro a los parámetros β para indicar que son los estimadores para los parámetros MCO aplicados a la muestra de datos particular.

Estas soluciones para los estimadores de β_1 y β_2 tienen varias implicancias. Entre ellas:

1. $\hat{\beta}_1 = \bar{y} - \hat{\beta}_2\bar{x}_2 \leftrightarrow \bar{y} = \hat{\beta}_1 + \hat{\beta}_2\bar{x}_2$: La relación estimada utilizando MCO intersecta a las medias muestrales (y esto generaliza a un caso con más de una variable independiente).
2. $\hat{\beta}_2 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_{2i}-\bar{x}_2)(y_i-\bar{y})}{\frac{1}{N}\sum_{i=1}^{N}(x_{2i}-\bar{x}_2)^2}$: El parámetro estimado de la pendiente es la razón de la covarianza simple entre x_{2i} y y_i y la varianza muestral de x_{2i} (esto NO generaliza a un caso con más de una variable independiente).

En la Figure 4.1, presentamos una regresión bivariada simple siguiendo a la ecuación 4.12. Esta figura muestra todos los nacimientos de la comuna de Corral (Región de los Ríos) en Chile en el año 2016. La línea discontinua (gris) muestra la línea de regresión lineal, estimada por mínimos cuadrados ordinarios. Para comparación, la línea punteada (azul) muestra una relación de regresión no-paramétrica (que no asume linealidad). Aquí observamos que la asociación entre edad de la madre y peso al nacer de su bebé es positivo: es decir $\hat{\beta}_2 > 0$. También podemos observar que el intercepto del peso al nacer (cuando Edad de Madre $= 0$), o $\hat{\beta}_1$ será alrededor de 2900-3000 gramos.

4.1.2 Minimización

Anteriormente, llegamos a la solución del estimador de MCO derivando la función objetivo, y fijando la derivada en cero. En estricto rigor, esta solución podría representar un mínimo, un máximo, o un punto de inflexión. Debemos asegurarnos que nuestra solución a las condiciones de primer orden produce un mínimo, y no un máximo de $\mu(\beta)$.

Para hacer esto, notamos que $\hat{\beta}_{MCO}$ resuelve las CPO de tal modo que $X'\hat{u} = 0$, donde $\hat{u} = y - X\hat{\beta}_{MCO}$. Volveremos a estudiar $\hat{u}$ en más detalle en la sección 4.2.3, pero por el momento notaremos que $\hat{u}$ es una cantidad estimada a partir de los parámetros $\hat{\beta}$, y por lo tanto, *no* es lo mismo que la cantidad u de la ecuación 4.4, que es no-observable, y parte de un modelo estadístico. Graficamos la cantidad $\hat{u}$ de la regresión 4.12 en la Figura 4.2 como la diferencia entre cada observación, y la recta de regresión.

Ahora, consideramos *cualquier* otro candidato del vector de parámetros $\tilde{\beta}$, definiendo $d = \tilde{\beta} - \hat{\beta}_{MCO}$ como el vector de diferencias entre los dos vectores de parámetros considerados. Y

Figure 4.1: Una Relación Bivariada: Paso al Nacer y Edad Materna

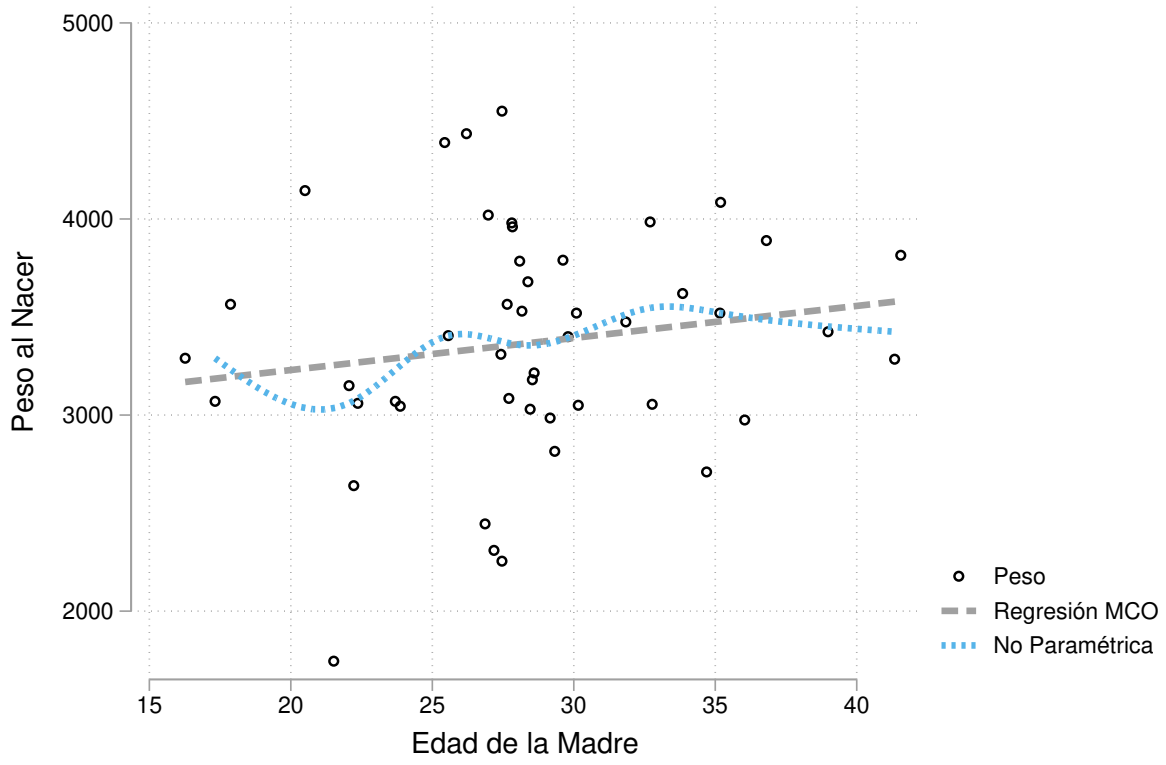
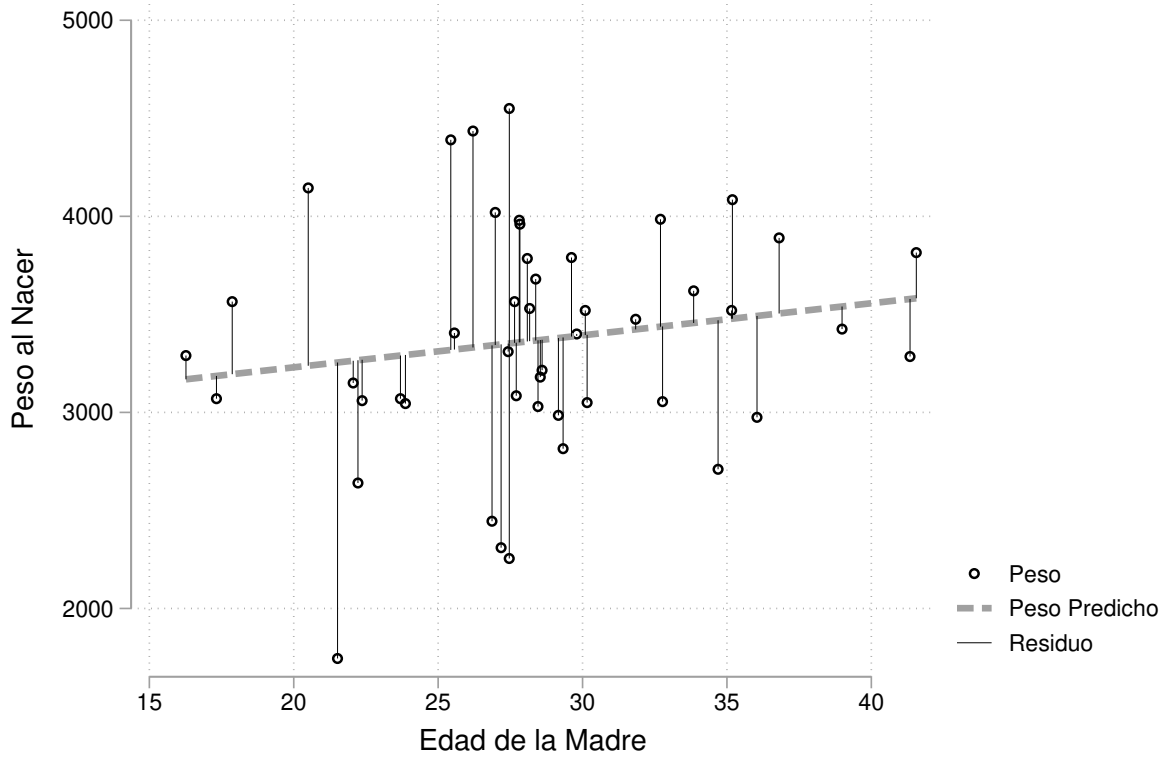


Figure 4.2: Una Relación Bivariada: Paso al Nacer y Edad Materna



definimos al vector $\tilde{u}$ como el análogo a $\hat{u}$, es decir:

$$\tilde{u} = y - X\tilde{\beta} = y - X(\hat{\beta}_{MCO} + d) = y - X\hat{\beta}_{MCO} - Xd = \hat{u} - Xd.$$

En base a la definición anterior, $\tilde{u}'\tilde{u} = (\hat{u} - Xd)'(\hat{u} - Xd)$. Dado que[4] $X'\hat{u} = 0$ (y entonces, $\hat{u}'X = 0$), tenemos:

$$\begin{aligned} \mu(\tilde{\beta}) = \tilde{u}'\tilde{u} &= (\hat{u} - Xd)'(\hat{u} - Xd) \\ &= \hat{u}'\hat{u} - \hat{u}'Xd - d'X'\hat{u} + d'X'Xd \\ &= \hat{u}'\hat{u} + d'X'Xd \\ &= \hat{u}'\hat{u} + v'v \end{aligned} \tag{4.17}$$

donde $v = Xd$ es un vector de columna de $N \times 1$. Esta cantidad $v'v$ es una escalar suma de cuadrados y por lo tanto $v'v \geq 0$, con $v'v = 0$ si y sólo si $v = 0$.

De lo anterior, demostramos que $\mu(\tilde{\beta}) \geq \mu(\hat{\beta}_{MCO}) = \hat{u}'\hat{u}$, con $\mu(\tilde{\beta}) = \mu(\hat{\beta}_{MCO})$ ssi $v = Xd = 0$. Esto implica que, el estimador MCO $\hat{\beta}_{MCO}$ minimiza $\mu(\beta) = u'u$, asegurando que nuestra solución de *mínimos* cuadrados ordinarios realmente es un *mínimo* (en vez de un punto máximo o punto de inflexión). Además, si $rango(X) = K$, el única vector de $K \times 1$ que satisface $Xd = 0$ es $d = 0$ y $\mu(\tilde{\beta}) = \mu(\hat{\beta}_{MCO})$ sii $\tilde{\beta} = \hat{\beta}_{MCO}$. En este caso, la solución a las CPO nos da el mínimo único. Es importante notar que si $rango(X) < K$, entonces $(X'X)^{-1}$ no existe (refiere a la sección 2.4.2 para un repaso), y no hay una solución única para las ecuaciones $X'u = 0$ para todos los K elementos del vector de parámetros β

Multicolinealidad Perfecta La situación en que $rango(X) < K$ se conoce como "multicolinealidad perfecta". Si estamos frente a un caso de multicolinalidad perfecta, no podemos resolver la ecuación 4.8, y por ende no existe una solución al estimador de MCO. Pero en la práctica, no es difícil evitar la situación de multicolinealidad perfecta. Simplemente no podemos incluir variables que son combinaciones lineales perfectas de otras variables explicativas en el modelo. Por ejemplo, si un modelo incluye x_{2i} y x_{3i}, no podemos incluir otra variable $x_{4i} = x_{2i} + x_{3i}$. Y por lo mismo no podemos incluir las dos variables género = feminino y género = masculino en el mismo modelo si el modelo también incluye un término constante, ya que feminino+masculino=1. El hecho de que no se puede estimar un término constante, más una serie de variables dicotómicas que cubren todas los niveles de una variable, es conocida como "La trampa de las variables dummies" (donde variable dummy refiere a una variable binaria para representar una característica). Si se quiere incluir una serie de variables dummies para capturar todos los niveles de una variable discreta, siempre es necesario omitir un nivel de la variable como el caso base. Por ejemplo, si se quisiera escribir

[4] El hecho que $X'\hat{u} = 0$ viene directamente del proceso de estimación. Recuerde que elegimos nuestro estimador de MCO para hacer cumplir la condición de primer orden (ecuación 4.7) $X'(y - X\beta) = 0$. En otras palabras, elegimos $\hat{\beta}$ para asegurar que $X'(y - X\hat{\beta}) = 0$, y sustituyendo $\hat{u} = (y - X\hat{\beta})$ muestra que por la definición del estimador MCO, $X'\hat{u} = 0$.

un modelo con variables binarias para todos los niveles de educación, se podrá incluir una variable para educación básica, media y terciaria (utilizando como caso base las observaciones sin nada de educación formal), pero no se puede tener las cuatro variables dummies en el mismo modelo.

4.2 La Regresión Lineal

4.2.1 El Modelo Clásico de Regresión Lineal

Hasta el momento, hemos planteado un modelo de regresión y sugerido una manera de estimar los parámetros de este modelo, pero no hemos dicho nada acerca de las propiedades del estimador de los parámetros de MCO. Como revisamos en bastante detalle en la sección 3.3.5, hay varias propiedades de un estimador que son favorables. Para poder establecer las propiedades del estimador MCO en versiones particulares del modelo, necesitamos hacer unos supuestos específicos acerca del modelo. Existe una versión clásica de la regresión lineal que incluye supuestos paramétricos convenientes (acerca de la forma funcional de y, X y u en el modelo). Por supuesto, como discutimos con la cita de Manski (2003) en la sección 3.3.1, siempre es importante cuestionar si los supuestos en un modelo estadístico son razonables. Por el momento, partiremos introduciendo los supuestos del modelo clásico de regresión lineal para derivar las propiedas del estimador MCO bajo condiciones básicas. Más adelante, nos preguntaremos qué podemos hacer si pensamos que estos supuestos *no* se cumplen.

Empezamos considerando algunos supuestos bajo los cuales podemos derivar propiedades que cumplen en muestras pequeñas. La versión del modelo que nos permite derivar estas propiedades de muestras finitas se conoce como el "modelo clásico de regresión lineal" Y más adelante veremos que el estimador MCO tiene algunas propiedades buenas en muestras grandes bajo supuestos bastante menos exigentes (las propiedades asintóticas del estimador).

El modelo clásico de regresión se plantea como:

$$y = X\beta + u$$

$$E(u|X) = 0,\ V(u|X) = \sigma^2 I$$

X es estocástico y de rango completo

o, de manera equivalente:

$$E(y) = X\beta$$

$$V(y) = \sigma^2 I$$

X es estocástico y de rango completo.

Aquí $V(\cdot)$ refiere a la varianza del vector no observable u, σ^2 es un constante, y I refiere a la matriz de identidad de tamaño $N \times N$.

Derivamos esta especificación, partiendo con un modelo lineal para observación i

$$\begin{aligned} y_i &= \beta_1 x_{1i} + \beta_2 x_{2i} + \ldots + \beta_K x_{Ki} + u_i \\ &= x_i'\beta + u_i \text{ para } i = 1, 2, \ldots, N \end{aligned}$$

Assumption 1 *Primero, asumimos que la expectativa condicional de y_i dado x_i es lineal, dando $E(y_i|x_i) = x_i'\beta$, o de manera equivalente, $E(u_i|x_i) = 0$. Esto es el **supuesto de expectativa condicional lineal**.*

Supuesto 1: Expectativa Condicional Lineal

Cuando hablamos de un modelo de regresión *lineal*, nos referimos a la forma de la relación entre las variables independientes, y la variable dependiente. La linealidad refiere a la relación entre cada x_k y y, y asume que, en promedio, un aumento en una unidad en una variable independiente tiene el mismo impacto sobre la variable dependiente, sin importar el nivel de la variable independiente. Para ver esto, consideramos un coeficiente del modelo lineal, β_1. En este modelo, β_1 captura cuánto aumenta y_i al aumentar x_{1i} en una unidad:

$$\frac{\partial y_i}{\partial x_{1i}} = \beta_1.$$

Es importante notar que aquí, la derivada es un constante β_1, y no depende de x_{1i}. Para considerar un caso puntual, imaginemos que estamos intentando capturar la relación entre educación (x_{1i}) y el salario de una persona (y_i). El supuesto de linealidad implica que al aumentar desde 0 a 1 año de educación tiene el mismo impacto sobre salario que cambiar desde 11 a 12 años de educación. Es inmediatamente claro aquí que una especificación lineal probablemente *no* es una manera apropiada para capturar esta relación. Por ejemplo, hay varios niveles de educación que resultan en saltos cuantitivos de salario (secundaria completa, terciaria completa, etc.). Y además, simplemente asumiendo que los retornos son constantes en los años completados tampoco permite capturar si hay retornos crecientes o decrecientes a la educación en el mercado laboral.

Aunque a primera vista puede parecer que el modelo lineal no es tan razonable, resulta ser mucho más flexible de lo que parece. En realidad, aunque asume linealidad en los parámetros estimados (los β_k), hay varias maneras de capturar relaciones no-lineales entre variables. Esto incluye funciones cuadráticas, por ejemplo incluyendo dos variables: años de educación, y años de educación al cuadrado en el modelo de regresión, o incluyendo variables indicadores para cada nivel de educación, (eg educación básico, educación secundaria, y educación terciara) que permite un retorno distinto para cada nivel de educación. Volveremos a evaluar las maneras de interpretar las coeficientes en una regresión lineal—y maneras de especificar un modelo bastante flexible a pesar de la linealidad de las coeficientes—en la sección 4.3.

Assumption 2 *Segundo, suponemos que existe* ***exogeneidad estricta****:* $E(u|x_1, \ldots, x_k) = E(u|X) = 0$.

Exogeneidad Estricta

El supuesto de exogeneidad estricta es clave en el modelo clásico de regresión lineal. Este supuesto además implica:

(i) $E(u) = 0$

(ii) $E(u_i|x_{jk}) = 0$ para cualquier i, j y k

(iii) $E(u_i x_{jk}) = 0$ para cualquier i, j y k (ortogonalidad de u y x_k)

(iv) $Cov(u_i, x_{jk}) = 0$ para cualquier i, j y k.

Dejamos la demostración de estos resultados como una pregunta al final de esta sección. El supuesto de exogeneidad estricta $E(u|X) = 0$ dice que, condicional en cada observación x_i, el término estocástico del modelo tiene media 0. En otras palabras, implica que ninguna observación de x entrega información acerca del componente no observado (o que u y cada x son ortogonales). Llamamos a regresores exógenos a las variables que cumplen con esta condición de exogeneidad estricta. Si una variable por alguna razón no cumple con esta condición, es conocida como una variable endógena. Algunos casos específicos de endogenidad ocurren cuando hay variables relevantes omitidas del modelo de regresión o errores de medición en variables independientes. Consideramos las implicancias de este caso en el capítulo 5.

Notemos los puntos (ii)-(iv) que la exogeneidad *estricta* refiere no sólo a una ausencia de correlación entre u_i y las realizaciones de $x_{i1}, \ldots, x_{Ki}$ para la misma observación, sino que a una ausencia de correlación entre el componente no observada de cada individuo, y las realizaciones de $x_{j1}, \ldots, x_{Kj}$ para cada otra observación j. Es por este razón que el supuesto es un supuesto *estricto*.

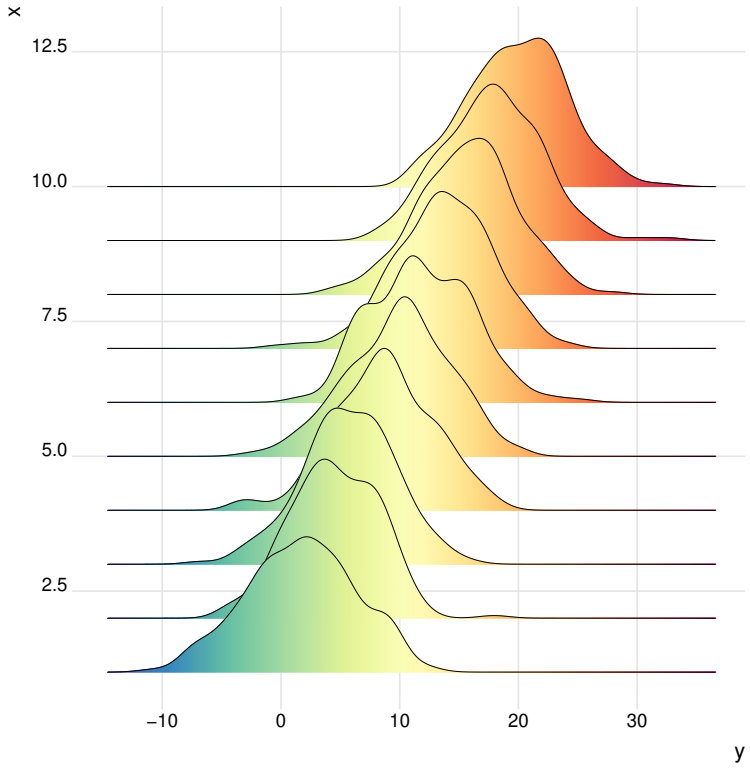
Assumption 3 *Tercero, asumimos que la varianza condicional de* y_i *dado* x_i *es un factor común para todas las observaciones* $i = 1, 2, \ldots, N$*, siendo* $V(y_i|x_i) = \sigma^2$*, o de manera equivalente,* $V(u_i|x_i) = \sigma^2$*. Esto es el* ***supuesto de homoscedasticidad condicional****.*

Supuesto 3: Homoscedasticidad Condicional y no autocorrelación

La homoscedasticidad condicional refiere a la varianza del término no observado, u. La homoscedasticidad condicional implica que la varianza de este término, σ^2, es constante para cada observación i. Esto implica que la varianza del componente no observado en el modelo de regresión no depende de la caracteristicas observadas de cada individuo en la muestra. En la Figura 4.3,

se grafica un ejemplo (simulado) donde la homoscedasticidad parece cumplir. Aquí, la variable dependiente y se gráfica en el eje horizontal, y la variable independiente se gráfica en el eje vertical (damos vuelta a las ejes para simplificar comprensión). Aunque en la medida que x aumenta, el promedio de la distribución de y aumenta, al parecer la varianza de cada distribución es idéntica, y no depende de x. Notemos que el supuesto de homoscedasticidad no es un supuesto *distribucional*, ya que solo toma supuestos acerca del término de la varianza de u y no de toda su distribución.

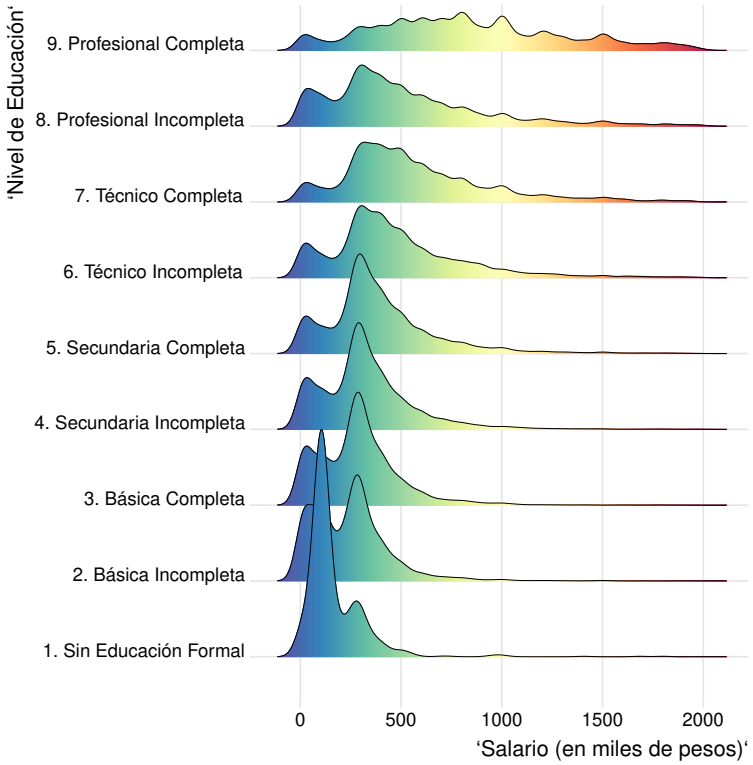
Figure 4.3: Homoscedasticidad Condicional



Pero hay muchos situaciones en que probablemente no es tan razonable suponer que estamos frente a una situación de homoscedasticidad condicional. Por ejemplo, en el caso de un modelo que relaciona el salario en el mercado laboral con el nivel más alto de educación alcanzada, es poco probable que la variación en el salario alrededor de la media para cada nivel de educación sea parecida sobre todos los niveles. En la Figura 4.4, se muestran estos patrones en Chile utilizando los datos reales del CASEN 2015. Se observa que para las observaciones con un menor nivel de educación la distribución de salario es bastante concentrada, con una marcada proporción que ganan el salario mínimo. Y mientras más alto es el nivel de educación completada, más se aumenta la varianza en el salario observado. Aquí claramente estamos en una situación en que no es razonable suponer que la varianza es constante para cada observación, sino depende de la variable "nivel de educación". En este capítulo vamos a mantener el supuesto de homoscedasticidad, pero en el siguiente capítulo vamos a ver qué debemos hacer cuando sospechamos que el modelo sufre de heteroscedasticidad.

Adicionalmente, suponemos que cada realización u_i no tiene relación con cualquier otra real-

Figure 4.4: Heteroscedasticidad Relacionada con Años de Educación



ización u_j. Esta independencia implica $Cov(u_i, u_j|x_1, x_2, \ldots, x_N) = 0$ para cada $i \neq j$, y permite que podemos escribir la matriz de varianza-covarianza de $N \times N$ para u como:

$$E[uu'|X] = V(u|X) = \begin{pmatrix} \sigma^2 & 0 & \ldots & 0 \\ 0 & \sigma^2 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & \sigma^2 \end{pmatrix} = \sigma^2 I.$$

Assumption 4 *Por último, tenemos el supuesto de que X (una matriz de $N \times K$) es estocástico y de rango completo: $rango(X) = K$.*

Supuesto 4: X es de Rango Completo

Este supuesto es un supuesto de identificación. Es necesario tener variación independente de cada variable $x_1, \ldots, x_K$ para poder estimar K parámetros $\beta_1, \ldots, \beta_K$. Este es el supuesto de **No Multicolinealidad**, y además implica que N tiene que ser por lo menos tan grande como K.

4.2.2 Propiedades de Muestra Finita de MCO

Expectativa de $\hat{\beta}$

El vector de parámetros $\hat{\beta}$ estimado por MCO es insesgado. Para ver esto, partimos con la fórmula MCO para $\hat{\beta} = (X'X)^{-1}X'y$, y definimos a $A = (X'X)^{-1}X'$. Entonces, $\hat{\beta} = Ay$, y A es conocido (no estocástico) condicional en X, mientras que y es un vector aleatorio condicional en X. La expectativa condicional de $\hat{\beta}_{MCO}$ es:

$$\begin{aligned} E(\hat{\beta}_{MCO}|X) &= E(Ay|X) \\ &= AE(y|X) \\ &= AX\beta \qquad (4.18) \\ &= (X'X)^{-1}X'X\beta \qquad (4.19) \\ &= \beta \end{aligned}$$

El paso 4.18 requiere el supuesto de expectativa condicional lineal, y el paso 4.19 requiere del supuesto de no mutlicolinealidad (para asegurar que la matriz de $K \times K$ $(X'X)$ es invertible). El resultado de $E(\hat{\beta}_{MCO}|X) = \beta$ dice que el estimador MCO $\hat{\beta}_{MCO}$ es insesgado condicional en los valores de X realizados en nuestra muestra. Esta demostración implica que $E(\hat{\beta}_{MCO} - \beta|X) = 0$, o en otras palabras, que $\hat{\beta}_{MCO}$ es condicionalmente insesgado.

En esta demostración, la expectativa de $\hat{\beta}_{MCO}$ se toma sobre la (no especificada) distribución condicional $u|X$. O de manera equivalente (dado que $y = X\beta + u$ y $X\beta$ no es estocástica condicional en X), sobre la distribución condicional $y|X$. La idea filosófica es que fijamos los valores X, y calculamos $\hat{\beta}_{MCO}$ para muchas distintas muestras de $u|X$ o $y|X$ En promedio, vamos a estimar el valor verdadero del vector de parámetros. Revisamos este idea con simulaciones Monte Carlo en los ejercicios computacionales al final de esta sección.

Este resultado condicional requiere de un vector de variables independientes X fijo. Sin embargo, tipicamente en nuestro trabajo econométrico, nos será más útil poder estipular que el estimador MCO es insesgado idependiente de la muestra de X observada en una situación determinada. Afortunadamente, es simple demostrar que $\hat{\beta}_{MCO}$ también es insesgada *no condicionando en X* utilizando la **Ley de Esperanzas Iteradas**:

$$E(\hat{\beta}_{MCO}) = E_X[E(\hat{\beta}_{MCO})|X] = E[\beta] = \beta \qquad (4.20)$$

En esta ecuación, hacemos explícito que la expectativa, y por ende la ley de expectativas iteradas, está operando sobre X. Como sabemos que cada $E(\hat{\beta}_{MCO})|X = \beta$, y adicionalmente que el vector de parámetros β no es estocástico sino una cantidad poblacional dada, llegamos al resultado de insesgadez incondicional.

La Varianza Condicional de $\hat{\beta}$

Aunque *en promedio* el estimador de MCO para β devuelve el valor poblacional de β, esto no implica nada acerca de una realización particular de $\hat{\beta}_{MCO}$. Por lo tanto, también nos interesa saber cuán variable es el estimador sobre distintas realizaciones de datos. Para cuantificar la variabilidad del estimador, podemos calcular su varianza.

Para derivar la varianza condicional del vector de parámetros $\hat{\beta}_{MCO}$, sabemos que $\hat{\beta} = (X'X)^{-1}X'y$, y hemos definido a $A = (X'X)^{-1}X'$. Entonces, simplemente utilizando las propiedades de la varianza de la sección 3.1.3:

$$\begin{aligned} V(\hat{\beta}|X) &= AV(y|X)A' = A(\sigma^2 I)A' = \sigma^2 AA' \\ &= \sigma^2(X'X)^{-1}X'X(X'X)^{-1} \\ &= \sigma^2(X'X)^{-1} \end{aligned}$$

dado que[5] $(X'X)^{-1}$ es una matriz simétrica de tal modo que $[(X'X)^{-1}]' = (X'X)^{-1}$.

Este resultado de la varianza depende del supuesto de homoscedastidad e independencia, para poder escribir $V(y|X) = \sigma^2 I$. Esta fórmula para la varianza sirve para varias cosas. Entre ellas,

1. nos permite comparar el comportamiento del estimador con otros estimadores en términos de su varianza, y
2. nos permite realizar contrastes de hipótesis acerca de los valores poblacionales para β en el modelo lineal.

En términos del primer punto, se puede demostrar que no existe otro estimador para el modelo lineal que sea insesgado, y que además tenga un varianza más pequeña del estimador MCO. Este resultado de eficiencia del estimador es muy poderoso y se conoce como **El Teorema de Gauss Markov**. Veremos el teorema de Gauss-Markov en las siguientes páginas. Y en términos del segundo punto, si se agrega un supuesto distribucional o de distribución asintótica, se tendrán todos los ingredientes necesarios para hacer contrastes de hipótesis formal acerca de los parámetros del modelo poblacional β. Volveremos a este punto en la sección 4.4 de estos apuntes.

El Teorema de Gauss Markov

El teorema de Gauss Markov dice que en el modelo clásico de regresión lineal, los coeficientes estimados por MCO son los mejores estimadores lineales insesgados (MELI). Es decir:

$$V(\hat{\beta}_{MCO}|X) \leq V(\tilde{\beta}|X)$$

donde $\tilde{\beta}$ es cualquier otro estimador lineal insesgado. El teorma de Guass Markov demuestra que el estimador lineal de MCO para β es eficiente en la clase de estimadores lineales insesgados.

[5] Esto es relevante para la segunda linea de la derivación, ya que: $AA' = (X'X)^{-1}X'[(X'X)^{-1}X']' = (X'X)^{-1}X'X[(X'X)^{-1}]'$.

La manera típica de demostrar el teorema[6] es considerar todas las estimadores lineales potenciales, y después mostrar que MCO tendrá la mínima varianza entre la clase de los estimadores insesgados. Para esto, definamos a $\tilde{\beta} = \tilde{A}y$ para una matriz $\tilde{A}$ de $K \times N$ que no es estocástico condicional en X. Para tener un $\tilde{\beta}$ insesgado condicional en X (que es un requisito para clasificar en la consideración de estimadores alternativos), necesitamos que:

$$E(\tilde{\beta}|X) = \tilde{A}E(y|X) = \tilde{A}X\beta = \beta \tag{4.21}$$

Por lo tanto, se requiere que $\tilde{A}X = I$. Y notemos que, por definición, la varianza de la clase de estimadores $\tilde{\beta}$ es:

$$V(\tilde{\beta}|X) = \tilde{A}V(y|X)\tilde{A}' = \sigma^2\tilde{A}\tilde{A}'. \tag{4.22}$$

El desafío del Teorema es demostrar que no hay ningun estimador que cumple con la condición de insesgades en 4.21, y que además tiene una varianza $\sigma^2\tilde{A}\tilde{A}' < \sigma^2AA'$ (la varianza del estimador MCO).

Partimos escribiendo a $\tilde{A}$ como $\tilde{A} = (A + D)$ donde $A = (X'X)^{-1}X'$, y la matriz $D = \tilde{A} - A$ es una matriz de $K \times N$. Ahora, $\tilde{A}X$ de la ecuación 4.21 es igual a $(A + D)X = AX + DX$, y sabemos que $AX = (X'X)^{-1}X'X = I$, dando que $\tilde{A}X = I + DX$. Para tener un $\tilde{\beta}$ insesgada condicional en X, necesitamos que $\tilde{A}X = I$, implicando que $DX = 0$ para considerar $\tilde{\beta}$ como un estimador insesgado. Notemos que $DX = 0$ a la vez implica que $DX(X'X)^{-1} = DA' = 0$, que implica que $[DA'] = AD' = 0$.

Ahora, $\tilde{A}\tilde{A}'$ de la ecuación 4.22 es igual a $(A + D)(A + D)' = AA' + AD' + DA' + DD'$. Si $\tilde{\beta}$ está insesgada condicional en X, tenemos que $AD' = DA' = 0$, y la ecuación anterior simplifica a $\tilde{A}\tilde{A}' = AA' + DD'$. Dado que $V(\tilde{\beta}|X) = \sigma^2\tilde{A}\tilde{A}'$, tenemos:

$$V(\tilde{\beta}|X) = \sigma^2AA' + \sigma^2DD'$$

Recordemos que $V(\hat{\beta}|X) = \sigma^2AA'$. Entonces, tenemos que $V(\tilde{\beta}|X) = V(\hat{\beta}|X) + \sigma^2DD'$. Para cualquier matriz D de $K \times N$, la matriz DD' es semidefinida positiva (sdp). Dado que $\sigma^2 = V(u_i|X) > 0$, también tenemos que σ^2DD' es sdp. Entonces, $V(\tilde{\beta}|X) - V(\hat{\beta}_{MCO}|X)$ es sdp, demostrando la declaración $V(\tilde{\beta}|X) \geq V(\hat{\beta}_{MCO}|X)$. Con esto, $\hat{\beta}_{MCO}$ es el MELI.

[6] Por ejemplo, ver Greene (2002, pp. 45–47) o Goldberger (1991, 165–166). Ambas utilizan notación un poco distinta. La notación aquí parece más a la notación de Goldberger (1991).

Implicancias del Teorema de Gauss Markov Notemos que la varianza condicional del estimador $\hat{\beta}_{MCO}$ es una matriz de $K \times K$, que se escribe de la siguiente forma:

$$V(\hat{\beta}|X) = \begin{pmatrix} V(\hat{\beta}_1|X) & Cov(\hat{\beta}_1, \hat{\beta}_2|X) & \cdots & Cov(\hat{\beta}_1, \hat{\beta}_K|X) \\ Cov(\hat{\beta}_2, \hat{\beta}_1|X) & V(\hat{\beta}_2|X) & \cdots & Cov(\hat{\beta}_2, \hat{\beta}_K|X) \\ \vdots & \vdots & \ddots & \vdots \\ Cov(\hat{\beta}_K, \hat{\beta}_1|X) & Cov(\hat{\beta}_K, \hat{\beta}_2|X) & \cdots & V(\hat{\beta}_K|X) \end{pmatrix}$$

donde $\hat{\beta}_k$ denota el elemento k-ésimo de $\hat{\beta}$ para $k = 1, 2, \ldots, K$. Las matrices de varianza son simétricas: $Cov(\hat{\beta}_j, \hat{\beta}_k|X) = Cov(\hat{\beta}_k, \hat{\beta}_j|X)$. Los elementos en el diagonal principal son la varianza de cada parámetro individual estimado por MCO.

Dos implicancias inmediatas del teorema Gauss Markov son:

1. Para *cada elemento* del vector de parámetros, el estimador MCO tiene la varianza más pequeña en la clase de estimadores lineales insesgados

2. Para *cualquier combinación lineal de los parámetros* β_k, el estimador MCO tiene la varianza más pequeña en la clase de estimadores lineales insesgados

Para ver porqué el primer punto es cierto, notemos que $V(\tilde{\beta}|X) - V(\hat{\beta}_{MCO}|X)$ también es una matriz simétrica sdp de $K \times K$. Una matriz simétrica que es sdp tiene números no negativos en su diagonal principal. Entonces, una $V(\tilde{\beta}|X) - V(\hat{\beta}_{MCO}|X)$ sdp implica que los elementos diagonales $V(\tilde{\beta}_k) - V(\hat{\beta}_k) \geq 0$, o: $V(\tilde{\beta}_k) \geq V(\hat{\beta}_k)\ \forall\ k$

Y para ver porqué el segundo punto es cierto, notemos que podemos fomar cualquiera combinación lineal de los parámetros β_k como $\theta = h'\beta$ para un vector no stocástico h de $K \times 1$. El vector h es un vector que se define para determinar la combinación lineal de parámetros de interés. Por ejemplo, $\beta_1 - \beta_2$ se obtiene como $h' = (1, -1, 0, \ldots, 0)$ tal que $h'\beta = \beta_1 - \beta_2$. El estimador $\hat{\theta}_{MCO}$ en base de $\hat{\beta}_{MCO}$ es $\hat{\theta}_{MCO} = h'\hat{\beta}_{MCO}$. El estimador $\tilde{\theta}$ en base de $\tilde{\beta}$ es $\tilde{\theta} = h'\tilde{\beta}$. Entonces:

$$V(\hat{\theta}_{MCO}|X) = h'V(\hat{\beta}_{MCO}|X)h, \text{ y}$$

$$V(\tilde{\theta}|X) = h'V(\tilde{\beta}|X)h$$

dando: $V(\tilde{\theta}|X) - V(\hat{\theta}_{MCO}|X) = h'[V(\tilde{\theta}|X) - V(\hat{\theta}_{MCO}|X)]h$. Para cualquier vector h de $K \times 1$ y matriz G sdp, el valor escalar $h'Gh \geq 0$. Por lo tanto:

$$\begin{aligned} V(\tilde{\theta}|X) &- V(\hat{\theta}_{MCO}|X) \geq 0 \\ V(\tilde{\theta}|X) &\geq V(\hat{\theta}_{MCO}|X). \end{aligned}$$

Actividad Computacional: Simulando $\hat{\beta}$ con β conocida

Consideremos un modelo $y = 2x_1 + u$ donde $x_1 \sim \mathcal{N}(0,1)$ y $u \sim \mathcal{N}(0,3)$, ambas independientes. En esta actividad realizaremos 500 simulaciones "Monte Carlo" para estimar el parámetro β con distintas realizaciones de u. Partimos simulando una "base de datos" con 1,000 observaciones de x_1. Para las preguntas a continuación, realizaremos $m = 500$ replicaciones Monte Carlo. En cada caso, generamos 1,000 realizaciones de u, y generamos la varible y como:

$$y^m = 2x_1 + u^m.$$

Aquí utilizamos el superíndice m para indicar que cada realización de u y y vienen de una realización distinta, mientras el x_1 lo mantenemos fijo.

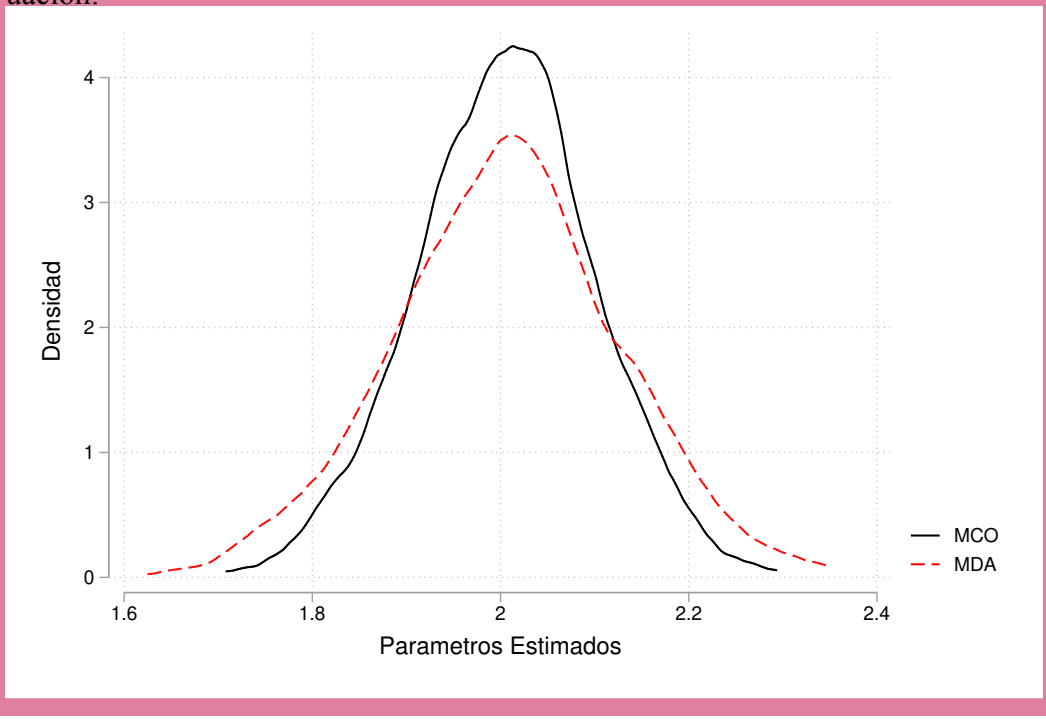
1. En cada una de las 500 replicaciones, estima $\hat{\beta}^m_{MCO}$, el estimador de mínimos cuadrados ordinarios para β, y guardamos el coeficiente estimado.

2. Además, en cada replicación, estima el mismo parámetro utilizando el estimador de mínimas desviaciones absoultas: $\widehat{\beta}^m_{MDA} = \sum_{i=1}^{N} |y_i - x_i'\beta|$. Guardamos los 500 parámetros estimados.

3. ¿Cuál es el promedio de los $\hat{\beta}^m_{MCO}$ y $\hat{\beta}^m_{MDA}$? Comenta.

4. ¿Cuál es la varianza de cada estimador en estas 500 realizaciones? Comenta. *Pista:* Aquí se refiere a la varianza $Var(\hat{\beta}^m_{MCO})$ sobre las m realizaciones, no el estimador de la varianza de $\hat{\beta}^m_{MCO}$ estimada por MCO en alguna realización específica.

5. Gráfica la distribución de las 500 estimaciones, por ejemplo como el gráfico a continuación:



6. Realiza una simulación más de u y y. Estima $V(\hat{\beta}_{MCO})$ utilizando la fórmula típica de MCO. ¿Cómo se compara con la varianza del parámetro $\hat{\beta}_{MCO}$ estimada a partir de las simulaciones Monte Carlo?

4.2.3 Predicciones y Bondad de Ajuste

Valores Predichos El valor predicho para la observación i es el valor que resulta de multiplicar el vector de coeficientes estimados $\hat{\beta}_{MCO}$ del modelo con el vector de características de cada observación en el modelo:

$$\hat{y}_i = x_i' \hat{\beta}_{MCO}$$

Es la mejor predicción del valor de y_i que el modelo de regresión lineal puede producir dadas sus variables independientes x_i. Notemos que aquí, cada observación i con el mismo vector de características observadas x_i tendrá el mismo valor predicho $\hat{y}$. El vector de los valores predichos para todas las N observaciones es:

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

donde el resultado final resulta al sustituir $\hat{\beta}_{MCO}$ por su fórmula matricial de ecuación 4.8. A veces, se escribe $X(X'X)^{-1}X'y = Py$, donde $P = X(X'X)^{-1}X'$ se conoce como la matriz de proyección, una matriz de $N \times N$, una matriz que proyecta y a sus valores predichos $\hat{y}$.

Residuos El residuo para la observación i corresponde al componente de y_i no explicado por el modelo. Se define como la diferencia entre la observación, y su valor predicho:

$$\hat{u}_i = y_i - x_i' \hat{\beta}_{MCO} = y_i - \hat{y}_i.$$

El vector de los residuous para todas las N observaciones es:

$$\begin{aligned} \hat{u} &= y - X\hat{\beta}_{MCO} = y - \hat{y} \\ &= y - X(X'X)^{-1}X'y \\ &= (I - P)y. \end{aligned}$$

La matriz $I - P$ de $N \times N$ se conoce como M, o la matriz generadora de residuos, donde $M = I - P = I - X(X'X)^{-1}X'$.

Las K ecuaciones normales utilizadas para obtener el estimador MCO implican que $X'\hat{u} = 0$:

$$\begin{pmatrix} \sum_{i=1}^{N} x_{1i}\hat{u}_i \\ \sum_{i=1}^{N} x_{2i}\hat{u}_i \\ \vdots \\ \sum_{i=1}^{N} x_{Ki}\hat{u}_i \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix},$$

o que las residuos de MCO $\hat{u}_1, \ldots, \hat{u}_N$ no están correlacionados con cada una de las variables explicativas del modelo. Esta propiedad de los residuos MCO es una consecuencia de estimación y sigue necesariamente de la definición del estimador MCO. No nos dice nada acerca de la validez del supuesto de esperanza condicional lineal:

$$E(u_i|x_i) = E(u_i|x_{1i}, x_{2i}, \ldots, x_{Ki}) = 0$$

o el supuesto (mas fuerte) de exogeneidad estrícta:

$$E(u_i|x_1, x_2, \ldots, x_K) = 0$$

que hemos tomado. Estos últimos supuestos están tomados en base al constructo teórico (no observado) u_i, no el constructo mecánico $\hat{u}_i$, que es un elemento observable que resulta del proceso de estimación.

Estimación de σ^2 Notemos que anteriormente, la varianza del estimador MCO dependía de un parámetro σ^2, que generalmente no es conocido. Para que el resultado de $V(\hat{\beta}_{MCO}|X) = \sigma^2(X'X)^{-1}$ sea útil, necesitamos un estimador para σ^2. Este parámetro es la varianza del término no observado u. El estimador MCO de σ^2 es:

$$\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K} = \frac{\sum_{i=1}^{N} \hat{u}_i^2}{N-K}$$

donde $\hat{u} = y - X\hat{\beta}_{MCO}$ son los residuos MCO. Notemos que este estimador parece a la varianza del término $\hat{u}$. Si queremos calcular la varianza del término $\hat{u}$ lo podríamos escribir como:

$$\frac{\sum_{i=1}^{N} (\hat{u}_i - \bar{\hat{u}}_i)^2}{N},$$

y dado que $\bar{\hat{u}}_i = 0$, este término simplifica a:

$$\frac{\sum_{i=1}^{N} \hat{u}_i^2}{N}.$$

La única diferencia entre este término es el denominador, donde en vez de dividir por N, se divide por $N-K$. La división por $N-K$ (en vez de N) es una corrección para los K grados de libertad que se utilizó para estimar los parámetros $\hat{\beta}_{MCO}$, y por ende, el término $\hat{u}$ que aparece en el estimador. Se puede demostrar que el estimador de MCO para σ^2 es insesgado. Dejamos esta demostración como una actividad.

Bondad de Ajuste

La bondad de ajuste se refiere al poder predictivo del modelo. Típicamente en la microeconometría nos interesa mucho más la estimación consistente o insesgadez de los coeficientes del modelo que el poder predictivo, pero hay algunos contextos en que el poder predictivo es relevante.[7] Una medida común de la bondad de ajuste del modelo lineal de regresión es el R-cuadrado, o R^2. A continuación derivamos algunos detalles acerca del R^2.

Notemos que, por definición, $y = \hat{y} + \hat{u}$, y que:

$$\hat{y}'\hat{u} = (X\hat{\beta}_{MCO})'\hat{u} = (\hat{\beta}'_{MCO}X')\hat{u} = \hat{\beta}'_{MCO}(X'\hat{u}) = 0.$$

La última igualdad viene directamente de los condiciones de primer orden de MCO que implican que $X'\hat{u} = 0$. Los residuos de MCO $\hat{u}$ son ortogonales a las variables independientes X, y además a los valores predichos $\hat{y} = X\hat{\beta}_{MCO}$ en la muestra. Entonces:

$$\begin{aligned} y'y &= (\hat{y}+\hat{u})'(\hat{y}+\hat{u}) = \hat{y}'\hat{y} + \hat{y}'\hat{u} + \hat{u}'\hat{y} + \hat{u}'\hat{u} \\ &= \hat{y}'\hat{y} + \hat{u}'\hat{u} \end{aligned}$$

o, de manera equivalente (algebraicamente),

$$\sum_{i=1}^{N} y_i^2 = \sum_{i=1}^{N} \hat{y}_i^2 + \sum_{i=1}^{N} \hat{u}_i^2 \tag{4.23}$$

Esto es simplemente una descomposición de la "suma de los cuadrados" de las observaciones y el modelo.

También tenemos que $y_i = \hat{y}_i + \hat{u}_i$ para $i = 1, \ldots, N$, implicando:

$$\begin{aligned} \sum_{i=1}^{N} y_i &= \sum_{i=1}^{N} \hat{y}_i + \sum_{i=1}^{N} \hat{u}_i \\ \text{y: } \bar{y} &= \bar{\hat{y}} + \bar{\hat{u}} \end{aligned} \tag{4.24}$$

Ahora, siempre que $\bar{\hat{u}} = 0$, obtenemos:

$$\bar{y} = \bar{\hat{y}}$$

[7] La importancia de predicción es central en *machine learning*, y herramientas de *machine learning* son cada vez más utilizadas en econometría. Ver por ejemplo Mullainathan and Spiess (2017) para discusión.

y restando $N\bar{y}^2$ de ambos lados de 4.23 produce

$$\left(\sum_{i=1}^{N} y_i^2\right) - N\bar{y}^2 = \left(\sum_{i=1}^{N} \hat{y}_i^2\right) - N\bar{y}^2 + \left(\sum_{i=1}^{N} \hat{u}_i^2\right).$$

Notando que $\left(\sum_{i=1}^{N} y_i^2\right) - N\bar{y}^2 = \sum_{i=1}^{N}(y_i - \bar{y})^2$, y de nuevo, siempre cuando $\bar{\hat{u}} = 0$, $\left(\sum_{i=1}^{N} \hat{y}_i^2\right) - N\bar{y}^2 = \sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2$, obtenemos:

$$\underbrace{\sum_{i=1}^{N}(y_i - \bar{y})^2}_{\text{Suma Total de Cuadrados}} = \underbrace{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}_{\text{Suma Explicada de Cuadrados}} + \underbrace{\sum_{i=1}^{N}\hat{u}_i^2}_{\text{Suma Residuo de Cuadrados}}$$

Esta fórmula proporciona una descomposición de la desviación al cuadrado de la media muestral (la "suma total de cuadrados"). Se descompone en un componente predicho por el modelo (la "suma explicada de cuadrados") y un componente no predicho por el modelo (la "suma residuo de cuadrados").

El R^2 cuantifica la proporción de la variación en las observaciones y_i que está explicada por el modelo. Si dividimos ambos lados de esta descomposición por la suma total de los cuadrados:

$$1 = \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} + \frac{\sum_{i=1}^{N}\hat{u}_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

y reorganizando da el R^2:

$$\begin{aligned} R^2 &= \frac{\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} = 1 - \frac{\sum_{i=1}^{N}\hat{u}_i^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2} \\ &= \frac{SEC}{STC} = 1 - \frac{SRC}{STC} \end{aligned}$$

donde *SEC* refiere a la suma explicada de cuadrados, *STC* refiere al suma total de los cuadrados, y *SRC* refiere al suma residuo de los cuadrados. Así, observamos que el R^2 es la proporción total de la variación del y_i que el modelo logra explicar.

Siempre cuando $\bar{\hat{u}} = 0$, $0 \leq R^2 \leq 1$. Para ver porque, consideramos los dos casos límites. Un $R^2 = 1$ resulta de un ajuste perfecto del modelo, cuando $y = X\hat{\beta}_{MCO}$. De este modo, $y - \hat{y} = \hat{u} = 0$, y por lo tanto $\hat{u}'\hat{u} = \sum_{i=1}^{N}\hat{u}_i^2 = 0$. Nota que dado que la suma residuo de los cuadrados nunca puede ser negativo, ya que es una cantidad cuadrática. Así, la función $1-(SRC/STC)$ nunca puede exceder 1. Y en el otro extremo de $R^2 = 0$ ocurre cuando para cada elemento $y = X\hat{\beta}_{MCO} = \bar{y}$, del modo que $\sum_{i=1}^{N}(\hat{y}_i - \bar{y})^2 = 0$. Nuevamente, la suma explicada de cuadrados es un término cuadrático, y tiene un valor mínimo de 0. Así la función SEC/STC no puede ser inferior a 0.

Para entender un poco más el funcionamiento del R^2, consideramos un modelo con $R^2 = 0$. Si X

es un vector de $N \times 1$ con cada elemento igual a 1, entonces $X'\hat{u} = \sum_{i=1}^{N} \hat{u}_i$. En este caso, $X'\hat{u} = 0$ (el resultado que sale de las condiciones de primer orden del estimador MCO), es equivalente a $\frac{1}{N}\sum_{i=1}^{N} \hat{u}_i = 0$. Definiendo el coeficiente del intercepto como β_1 produce:

$$\frac{1}{N}\sum_{i=1}^{N}(y_i - \hat{\beta}_1) = 0 \leftrightarrow \hat{\beta}_1 = \frac{1}{N}\sum_{i=1}^{N} y_i = \bar{y}$$

y entonces, $\hat{y}_i = \hat{\beta}_1 \times 1 = \bar{y}$ para $i = 1, 2, \ldots, N$. Entonces, el R^2 mide el bondad de ajuste de un modelo con variables adicionales *relativa* a un modelo que sólo consiste de un término de intercepto.

En varios puntos de esta sección, hemos requerido que el promedio de los residuos muestrales sea igual a cero. La inclusión de un término de intercepto asegura $\sum_{i=1}^{N} \hat{u}_i = 0 \leftrightarrow \bar{\hat{u}} = \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i = 0$. A la vez, esto asegura que $0 \leq R^2 \leq 1$. Para un modelo de regresión lineal que *no* incluye un término de intercepto (o que no asegura que $\bar{\hat{u}} = \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i = 0$ en alguna otra manera), no se debe utilizar el R^2 como medida de ajuste.

Volviendo al R^2 máximo de 1, notamos que en una muestra de tamaño N, siempre podemos obtener un ajuste perfecto ($R^2 = 1$) mediante una regresión de y_i sobre N variables explicativas independientes. De manera más generalizada, aumentando la cantidad de variables explicativas no puede reducir el R^2. Si se agrega una variable irrelevante, el ajuste del modelo no cambia. Y si se agrega una variable relevante (que aporta poder predictivo al modelo), el ajuste mejora. Dado este resultado mecánico, en algunos casos se utiliza en R^2 ajustado ($\bar{R}^2$), definido como:

$$(1 - \bar{R}^2) = \frac{(N-1)(1-R^2)}{(N-K)}.$$

La idea del R^2 ajustado es que permite comparar el poder predictivo de dos modelos (con la misma variable dependiente) con una cantidad distinta de variables, sin necesarimanete aumentar siempre cuando se incluya otra variable poca relevante. El R^2 ajustado descuenta mejoras en el ajuste mientras la cantidad de variables K aumenta en una muestra dado de tamaño N. Aunque claramente este ajuste "castiga" la inclusión de variables adicionales (por el término K en el denominador), no hay ninguna justificación teórica para este ajuste en particular.

Una desvantaja en particular del R^2 como una medida de bondad de ajuste es que el R^2 *no* es invariante a transformaciones lineales del modelo…Por ejemplo, comparamos:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

y

$$(y_i - x_{3i}) = y_i^* = \beta_1 + \beta_2 x_{2i} + (\beta_3 - 1)x_{3i} + u_i$$

Estos modelos son equivalentes y resultan en las mismas estimaciones de MCO ($\beta_1, \beta_2, \beta_3$), no-

tando que la coeficiente en x_{3i} en el segundo modelo es $\beta_3 - 1$. Tienen el mismo término de error u_i. Además tienen los mismos residuos $\hat{u}_i$, y por lo tanto el mismo suma de los residuos al cuadrado $SRC = \sum_{i=1}^{N} \hat{u}_i^2$. Sin embargo, tienen distintas variables dependientes, y por lo general:

$$\sum_{i=1}^{N}(y_i - \bar{y})^2 \neq \sum_{i=1}^{N}(y_i^* - \bar{y^*})^2.$$

Por consiguiente, tienen distintos valores para $R^2 = 1 - \frac{SRC}{STC}$. Por lo tanto, el R^2 no se debe utilizar para comparar la bondad de ajuste de modelos con distintas variables dependientes. Una manera más apropiada para comparar en esta situación será comparar $\sigma^2 = \frac{\hat{u}'\hat{u}}{N-K}$, que es un estimador insesgado de la varianza del término de error u_i.

Preguntas:

1. Demuestra que el estimador de MCO para el término de varianza σ^2, $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K} = \frac{\sum_{i=1}^{N} \hat{u_i}^2}{N-K}$ es un estimador insesgado.
2. Hemos definido **exogeneidad estricta** como $E(u|x_1, \ldots, x_k) = E(u|X) = 0$. Demuestra que la exogeneidad estricta implica además las siguientes condiciones:
   (i) $E(u) = 0$
   (ii) $E(u_i|x_{jk}) = 0$ para cualquier i, j y k
   (iii) $E(u_i x_{jk}) = 0$ para cualquier i, j y k (ortogonalidad de u y x_k)
   (iv) $Cov(u_i, x_{jk}) = 0$ para cualquier i, j y k.

4.3 Coeficientes

4.3.1 Entendiendo los Coeficientes en Modelos de Regresión Lineal

Las coeficientes β_k en un modelo de regresión lineal miden el cambio promedio en la variable dependiente y, dado un cambio de una unidad en la variable independiente x_k, manteniendo fijas las demás variables en el modelo. En el caso de una variable continua (o categórica con muchos niveles), el supuesto de linealidad implica que no importa el nivel de x_k cuando consideramos el cambio, β_k simplemente mide el cambio promedio en la población al aumentar x_k desde algún valor d a otro valor $d + 1$. Aunque β_k mide el cambio promedio en la población, es posible que haya mucha varianza de este término en distintos miembros de la población. Por ejemplo, si estimamos que un año más de educación aumenta el salario promedio en $\$50,000$ por mes no implica que el impacto será lo mismo para cada persona, y ni siquiera implica que el impacto será positivo

para cada persona.[8] Si todos los supuestos Gauss-Markov cumplen, los coeficientes capturan una relación causal entre y y x_k.

Algunas Formas Funcionales en Modelos de Regresión Como discutimos brevamente cuando introducimos el supuesto de linealidad en el modelo clásico, hay varias maneras de capturar relaciones no-lineales entre la variable dependiente y las variables independientes, aún respetando la linealidad inherente en el modelo. Una manera de capturar relaciones no lineales es utilizando funciones polinómicas de variables independientes. Por ejemplo, consideramos la conocida "regresión de Mincer" que plantea un modelo para el salario laboral como:

$$salario_i = \beta_0 + \beta_1 educ_i + \beta_2 exper_i + \beta_3 exper_i^2 + u_i.$$

En este modelo, $educ_i$ refiere a los años de educación del individuo i, y $exper_i$ y $exper_i^2$ refieren a la cantidad de años de experiencia laboral y su cuadrado. En este modelo, cada coeficiente por si solo representa una relación lineal. Pero si consideramos el impacto de un cambio en los años de experiencia sobre el salario laboral, notemos que:

$$\frac{\partial salario_i}{\partial exper_i} = \beta_2 + 2 \cdot \beta_3 exper_i.$$

Aquí, la relación entre salario y experiencia laboral es una relación no lineal, ya que depende del nivel de experiencia de cada persona, ya que $exper$ aparece en la derivada. Específicamente, estas 'formas funcionales' cuadráticas permiten capturar retornos crecientes y decrecientes a escala. En el caso de la experiencia laboral, se suele esperar que el retorno de la experiencia sea alto al inicio de la carrera y, por ejemplo, el salario esperado de una persona puede aumentar bastante al aumentar desde 0 a 1 años de experiencia. Sin embargo, este retorno cae sobre el tiempo y, por ejemplo, al aumentar desde 30 a 31 años de experiencia, el cambio salarial sea bastante menor. En este caso específico (de retornos decrecientes), el segundo término (β_3) es negativo. Si hay una relación con retornos crecientes, el término β_3 será positivo, y el cambio en la variable dependiente será mayor mientras más alto es la variable independiente.

Otra manera común de incorporar relaciones no-lineales en el modelo de regresión lineal es mediante transformaciones logarítmicas de algunas variables del modelo. Tomando transformaciones logarítmicas, por ejemplo reemplazando la variable $salario$ para el logaritmo natural del salario, $ln(salario)$ se tiene varias ventajas en la interpretación del modelo. Específicamente, permite interpretar algunas relaciones en términos de elasticidad.[9] Pero por supuesto también tiene algunas limitaciones. Una limitación clara es que la transformación solamente será posible con variables

[8] Hay técnicas para considerar cómo estos valores varian en la población, por ejemplo la regresión cuantílica. Conversamos de estas técnicas en el segundo año del magíster.

[9] Recordamos que la definición de elasticidad-x de y es $\frac{\partial ln(y)}{\partial ln(x)}$, interpretado como el cambio porcentual de y dado un cambio de 1% de x. A diferencia, en los modelos de regresión que hemos visto hasta ahora, las coeficientes β capturan el cambio en unidades de y dado un cambio de una unidad de x.

con un soporte estrictamente positivo, ya que el logaritmo de un valor negativo no es definido.

El interés en incorporar transformaciones logarítmicas a modelos de regresión es por su interpretación como un cambio porcentual. Por ejemplo, volviendo al ejemplo de salario y educación, posiblemente no pensamos que el salario aumenta en una cantidad constante para cada año adicional de educación, pero sí aumenta en un porcentaje constante. Un modelo de la siguiente forma:

$$ln(y_i) = x_{1i} + \beta_2 x_{2i} + u_i$$

es conocida como un modelo de log-nivel (ya que la variable dependiente está medida en logaritmos, y la variable independente en nivel), o un modelo de semi-elasticidad. En este caso, el cambio porcentual de la expectativa de y (la variable medida en unidades originales), frente a un cambio de 1 una unidad de x es aproximadamente $100 \times \hat{\beta}_2$. Es importante notar que esto es sólo una aproximación, y es una mejor aproximación cuando el cambio de x es pequeño.

Modelos con transformaciones logarítmicas de este estilo también pueden ser utilizados cuando la variable independiente está transformada en unidades logarítmicas, o cuando tanto la variable dependiente como la(s) variable(s) independiente(s) están transformadas. En el primer caso, un modelo de semi-elasticidad de y contra $ln(x)$, un cambio de un porcentaje de x aproximadamente cambiaría la expectativa de y en $\beta_2/100$ unidades. Y en el último caso, de un modelo de log-log (o elasticidad), un cambio de un porcentaje de x está asociado con un cambio de $\beta_2\%$ en la expectativa de y.

Variables Binarias y Términos de Interacción La discusión anterior está centrada en variables continuas, o aproximadamente continuas. Sin embargo, es común observar variables binarias, o dicotomicas, en modelos de regresión. Cuando hay una variable binaria (una variable que toma solamente valores de 1 o 0) como una variable independiente en una regresión, su interpretación es muy sencilla. Consideremos una versión un poco distinto del modelo de Mincer:

$$salario_i = \beta_0 + \beta_1 educmedia_i + \beta_2 exper_i + \beta_3 exper_i^2 + u_i.$$

Aquí, la variable $educmedia_i$ toma el valor de 1 si la persona ha completado educación media, y 0 si no. En este modelo el parámetro β_2 simplemente mide el retorno a educación media. En cualquier modelo que tenga una variable binaria, la coeficiente sobre una variable binaria captura la diferencia condicional entre personas que tienen esta característica y las que no. En la notación de esta ecuación, tenemos:

$$\beta_1 = E[salario_i | educmedia_i = 1, exper_i] - E[salario_i | educmedia_i = 0, exper_i].$$

Este uso de variables binarias se extiende a modelos con múltiples variables binarias. Por ejemplo, en el modelo anterior, podríamos haber tenido una variable para educación básica, una variable

educación media y una variable para educación terciaria, y en cada caso el coeficiente asociado con cada variable sería el retorno de haber completado este nivel de educación, versus el caso base cuando una persona no tiene ningún nivel de educación formal.[10] Veremos algunos ejemplos de este tipo de modelo en las preguntas al final de esta sección.

Por último, notamos que las variables binarias también pueden ser combinados para permitir diferencias en retornos a *otras variables*. Para ver porqué, consideramos el siguiente modelo:

$$salario_i = \beta_0 + \beta_1 educmedia_i + \beta_2 exper_i + \beta_3 (educmedia_i \times exper_i) + u_i.$$

Aquí, β_1 sigue siendo el retorno de educación para personas con una educación media. Pero ahora, si queremos cuantificar el retorno a *experiencia*, esto también depende de la educación de la persona. Para una persona sin educación media, el tercer término $educmedia_i \times exper_i = 0 \times exper_i = 0$, y por lo tanto, el cambio en y asociado con un cambio en experiencia sería simplemente β_2. Pero para una persona que si tiene educación media, su retorno a experiencia sería $\beta_2 + \beta_3$:

$$\frac{\partial salario_i}{\partial exper_i}\Big|_{educmedia_i=1} = \beta_2 + \beta_3 \times 1.$$

De esta manera, se puede formar modelos lineales bastante completos, que permiten tanto interceptos como pendiente distintos para distintos sub-grupos en la muestra de interés.

Preguntas:

1. Considera el modelo:

   $$salario_i = \beta_0 + \beta_1 educbasica_i + \beta_2 educmedia_i + \beta_3 educterciaria_i + u_i$$

   donde $educbasica = 1$ si el nivel más alto de educación de individuo i es básica o 0 si no, $educmedia = 1$ si el nivel más alto de educación de individuo i es la educación media o 0 si no, y $educterciaria = 1$ si el nivel más alto de educación de individuo i es terciaria, y 0 si no. Cada variable es mutuamente excluyente. Ahora, si las variables fueron redefinidas para ser 1 si la persona tiene *por lo menos* este nivel, de tal modo que no son mutuamente excluyentes, (eg una persona con educación media tendría $educbasica_i = 1$ y $educmedia_i = 1$) ¿cómo se comparían las coeficientes estimadas en cada modelo si fueran estimadas con la misma muestra de observaciones? Explica este resultado.

2. Considere el modelo (2) presentado en el trabajo "'Growth, Girls' Education, and Female Labor: A Longitudinal Analysis"' Lincove (2008) el cual relaciona el crecimiento económico y educación con la participación femenina en el mercado laboral en distintos

[10] Notemos que en este caso, siempre tenemos que tener un caso base, cuya variable binaria es omitida. En formas prácticas, esto sirve para comparar las coeficientes estimadas con algún grupo de referencia. Pero en términos más fundamentales para el modelo, si tenemos una serie de variables binarias que son perfectamente correlacionadas entre si, estaremos frente a un caso de multicolinealidad.

países, controlando por los cambios estructurales del mercado laboral y los obstáculos culturales:

$$\begin{aligned} FEP_{(45-59)} &= \alpha + \beta_1 ln(GDPpc) + \beta_2 ln(GDPpc)^2 + \beta_3 GER_{t-30} + \beta_4 GER_{t-30}^2 + \\ &\quad \beta_5 industry + \beta_6 service + \beta_7 Islamic + \beta_8 Catholic + \beta_9 CivilLiberties \\ &\quad + \epsilon \end{aligned}$$

Donde $FEP_{(45-59)}$ corresponde a la tasa de participación femenina en el mercado laboral de las mujeres entre 45 y 59 años, *GDPpc* corresponde al PIB per cápita en dólares y *GER* es la tasa bruta de matrícula femenina de las mujeres estudiadas, es decir, la matrícula de 30 años atrás. $industry$ y $service$ corresponden al porcentaje de la fuerza laboral empleada en cada sector económico, $Islamic$ y $Catholic$ toman el valor 1 si la mayoríade los ciudadanos profesan la religión islámica y católica, respectivamente, y $CivilLiberties$ corresponde a un índice de 1a 7 asignado a cada país de manera que mientras mayor es el índice, mayores libertades civiles tienen los ciudadanos.

Resultados de estimación

FEP	Coeficiente
ln(GDPpc)	-19.557**
ln(GDPpc)^2	0.989
GER$_{t-30}$	-0.500**
GER$^2_{t-30}$	0.008***
industry	-0.314*
service	-0.146
Islamic	-9.102**
Catholic	-8.635***
Civil Liberties	3.519***
Constante	147.752***
R-cuadrado	0.556
R-cuadrado Ajustado	0.519
Observaciones	119

De acuerdo a los resultados de estimación presentados, responda:

(a) Interprete el coeficiente de la variable $CivilLiberties$.

(b) La literatura plantea que la relación en forma de U entre la tasa de participación femenina y la educación. ¿Los resultados de estimación apoyan esta hipótesis? Explique.

(c) ¿Los países con mayoría islámica tienen una mayor o menor participación laboral? Interprete el parámetro de interés estimado .

(d) ¿Cuál es el impacto estimado de *GDPpc* sobre la participación laboral femenina para un país con PIB per cápita de $2,000? Interprete.

4.3.2 Los Coeficientes y el Álgebra de Regresión

Para entender qué están haciendo los coeficientes en un modelo de regresión lineal, y por qué nos referimos a los coeficientes como "manteniendo todo lo demás constante", consideremos el teorema de Frisch-Waugh-Lovell. Este teorema toma su nombre de dos articulos: Frisch and Waugh (1933) y Lovell (1963), y demuestra otra manera de llegar a los coeficientes de MCO de un modelo múltiple, utilizando un modelo bivariado, siempre y cuando se condicione cada variable en el modelo bivariado en cada variable en el modelo original.

Para ver las implicancias del teorema Frisch-Waugh-Lovell, consideremos el modelo de regresión múltiple con $K = 2$:

$$y = X_1\beta_1 + X_2\beta_2 + u = X\beta + u$$

Asumimos para simplicidad que $\bar{y} = \bar{X}_1 = \bar{X}_2$. En realidad, esta simplificación no es muy restrictiva, ya que siempre podemos estandarizar variables para ser expresadas en términos de desviaciones de las medias muestrales. ¿Cómo interpretamos las estimaciones MCO de las coeficientes individuales $\hat{\beta}_1$ y $\hat{\beta}_2$ en el modelo de regresión múltiple? Procedemos en partes para ver cuando recuperamos, y cuando no recuperamos, el valor poblacional verdadero en distintos modelos.

1. Regresión Bivariada X_1

Primero, consideraremos la regresión simple de X_1 sobre y. Escribimos este modelo como:

$$y = X_1\gamma + r_y,$$

donde γ es el coeficiente de la regresión bivariada, y r_y es el error. Considerando el estimador MCO para γ, tenemos:

$$\begin{aligned}
\hat{\gamma}_{MCO} &= (X_1'X_1)^{-1}X_1'y = A_1y \\
\tilde{y} &= X_1\hat{\gamma}_{MCO} = X_1(X_1'X_1)^{-1}X_1'y \\
\hat{r}_y &= y - \tilde{y} = (I - X_1(X_1'X_1)^{-1}X_1')y = M_1y,
\end{aligned}$$

donde hemos definido la matriz de proyección $A_1 = (X_1'X_1)^{-1}X_1'$, y la matriz generadora de residuos $M_1 = (I - X_1(X_1'X_1)^{-1}X_1')$. Por nuestras definiciones anteriores (en la sección 4.2.3) $\tilde{y}$ son los

valores predichos, y $\hat{r}_y$ son los residuos de la regresión. Ahora, podemos escribir:

$$\hat{\gamma}_{MCO} = A_1 y = A_1[X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{u}] \tag{4.25}$$

2. Regresión Bivariada de X_1 sobre X_2

Consideraremos también la regresión simple de X_1 sobre X_2:

$$X_2 = X_1\delta + r_2$$

Nuevamente, por las definiciones del estimador MCO, tenemos:

$$\begin{aligned}
\hat{\delta}_{MCO} &= (X_1'X_1)^{-1}X_1'X_2 = A_1X_2 \\
\hat{X_2} &= X_1\hat{\delta}_{MCO} = X_1(X_1'X_1)^{-1}X_1'X_2 \\
\hat{r}_2 &= X_2 - \hat{X_2} = (I - X_1(X_1'X_1)^{-1}X_1')X_2 = M_1X_2
\end{aligned}$$

con $X_1'\hat{r}_2 = 0$. Y ahora, volviendo a la ecuación 4.25, podemos simplificador el estimador de $\hat{\gamma}_{MCO}$ de la regresión bivariada de y sobre X_1:

$$\begin{aligned}
\hat{\gamma}_{MCO} &= A_1 y = A_1[X_1\hat{\beta}_1 + X_2\hat{\beta}_2 + \hat{u}] \\
&= (X_1'X_1)^{-1}(X_1'X_1)\hat{\beta}_1 + (X_1'X_1)^{-1}(X_1'X_2)\hat{\beta}_2 + (X_1'X_1)^{-1}X_1'\hat{u} \\
&= \hat{\beta}_1 + \hat{\delta}_{MCO}\hat{\beta}_2 + 0 \\
&= \hat{\beta}_1 + \hat{\delta}_{MCO}\hat{\beta}_2
\end{aligned}$$

La penúltima linea sigue de (a) la invertibilidad de $(X_1'X_1)$, (b) la definición del coeficiente de regresión $\hat{\delta}_{MCO}$, y (c) la ortogonalidad (por construcción) de $\hat{u}$ y X_1. Entonces, $\hat{\beta}_1 \neq \hat{\gamma}_{MCO}$ salvo que $\hat{\beta}_2 = 0$ o $\hat{\delta}_{MCO} = 0$ (o ambos). Volveremos a explorar estos dos casos en más profundidad cuando hablamos del "sesgo de variable omitida" en la sección 5.3. Por el momento, notamos que *por lo general, la coeficiente estimada ($\hat{\beta}_1$) en X_1 en la regresión múltiple de y en X_1 y X_2 es diferente a la coeficiente estimado ($\hat{\gamma}_{MCO}$) en X_1 en la regresión simple de X_1 sobre y.*

3. Regresión de Residuos

Ahora, por último, consideremos la regresión de $\hat{r}_y$ en $\hat{r}_2$. Ésta se conoce como una "regresión de residuos", ya que estamos utilizando los residuos de las dos regresiones anterieros. Este modelo es escribe:

$$\hat{r}_y = \hat{r}_2 b + v_2$$

donde los $\hat{r}_y$ son los residuos de la regresión simple de y sobre X_1, $\hat{r}_2$ son los residuous de la regresión simple de X_1 sobre X_2, y b es el coeficiente de la regresión bivariada.

Siguiendo la definición MCO, podemos escribir el estimador para b como:

$$\hat{b}_{MCO} = (\hat{r}_2'\hat{r}_2)^{-1}\hat{r}_2'\hat{r}_y$$

De las secciones anteriores, sabemos que $\hat{r}_2 = M_1X_2$ y $\hat{r}_y = M_1y$. Y notamos además que la M_1 es una matriz simétrica e idempotente (es decir, $M_1M_1 = M_1$).[11] Con un poco de álgebra, podemos re-escribir los términos $\hat{r}_2'\hat{r}_2$ y $\hat{r}_2'y$ como:

$$\begin{aligned}
\hat{r}_2'\hat{r}_2 &= (M_1X_2)'M_1X_2 = (X_2'M_1')M_1X_2 = X_2'M_1M_1X_2 = X_2'M_1X_2 \\
\hat{r}_2'\hat{r}_y &= (M_1X_2)'M_1y = (X_2'M_1')M_1y = X_2'M_1M_1y = X_2'M_1y,
\end{aligned}$$

y entonces, $\hat{b}_{MCO} = (\hat{r}_2'\hat{r}_2)^{-1}\hat{r}_2'\hat{r}_y = (X_2'M_1X_2)^{-1}X_2'M_1y$.

Notemos que:

$$\begin{aligned}
M_1y &= M_1[X_1\hat{\beta}_1 + X_2\beta_2 + \hat{u}] \\
&= M_1X_1\hat{\beta}_1 + M_1X_2\hat{\beta}_2 + M_1\hat{u} \\
&= 0 + M_1X_2\hat{\beta}_2 + M_1\hat{u}
\end{aligned}$$

dado que $M_1X_1 = 0$. Y, dado que $X_1'\hat{u} = 0$:

$$\begin{aligned}
M_1\hat{u} &= (I - X_1(X_1'X_1)^{-1}X_1')\hat{u} \\
&= \hat{u} - X_1(X_1'X_1)^{-1}X_1'\hat{u} \\
&= \hat{u}
\end{aligned}$$

Con estos resultados, tenemos:

$$\begin{aligned}
\hat{b}_{MCO} &= (X_2'M_1X_2)^{-1}X_2'M_1y \\
&= (X_2'M_1X_2)^{-1}X_2'[M_1X_2\hat{\beta}_2 + \hat{u}] \\
&= (X_2'M_1X_2)^{-1}(X_2'M_1X_2\hat{\beta}_2) + (X_2'M_1X_2)^{-1}(X_2'\hat{u}) \\
&= \hat{\beta}_2.
\end{aligned}$$

Esto demuestra que el coeficiente estimado ($\hat{\beta}_2$) en X_2 en una regresión múltiple de y en X_1 y X_2 **es igual** al coeficiente estimado ($\hat{b}_{MCO}$) en $\hat{r}_2$ en la regresión simple de $\hat{r}_y$ en $\hat{r}_2$. Ésta es una regresión de los residuos MCO ($\hat{r}_y$) de la regresión de y en X_1 en los residuos MCO ($\hat{r}_2$) de la regresión de X_2 en X_1. Este resultado es un caso específico del teorema Frisch-Waugh-Lovell. Con este procedimiento de generación de residuos, estamos "sacando" (o concentrando) el efecto de X_1 de ambas variables: y y X_2. Una vez que hemos sacado la variación de y y X_2 que viene de X_1, sólo consideramos la relación que queda entre y y X_2 controlando por el efecto externo de X_1.

[11] Dejamos como un ejercicio la demostración que $M_1M_1 = M_1$.

Regresión Múltiple con $K > 2$ Este resultado generaliza a los modelos con más de dos variables explicativas. En el modelo con K variables explicativas:

$$y_i = \beta_1 x_{1i} + \beta_2 x_{2i} + \cdots + \beta_{K-1} x_{(K-1)i} + \beta_K x_{Ki} + u_i,$$

podemos "concentrar" variables $x_1, \ldots, x_{K-1}$ de y, y de X_K estimando las siguientes estimaciones:

$$\begin{aligned} y_i &= \gamma_1 x_{1i} + \gamma_2 x_{2i} + \cdots + \gamma_{K-1} x_{(K-1)i} + r_{yi} \\ x_{Ki} &= \delta_1 x_{1i} + \delta_2 x_{2i} + \cdots + \delta_{K-1} x_{(K-1)i} + r_{Ki}, \end{aligned}$$

y después calculando los residuos $\hat{r}_{yi}$ y $\hat{r}_{Ki}$. Por último, si después estimamos la regresión de residuos de la siguiente forma:

$$\hat{r}_{yi} = b\hat{r}_{Ki} + v_i$$

nuevamente tendremos que:

$$\hat{\beta}_K = \hat{\beta}_{MCO} = (\hat{r}_K' \hat{r}_K)^{-1} \hat{r}_K' \hat{r}_y = \frac{\sum_{i=1}^N \hat{r}_{yi} \hat{r}_{Ki}}{\sum_{i=1}^N \hat{r}_{Ki}^2} = \frac{\frac{1}{N}\sum_{i=1}^N \hat{r}_{yi} \hat{r}_{Ki}}{\frac{1}{N}\sum_{i=1}^N \hat{r}_{Ki}^2}.$$

Aquí hemos 'sacado' los efectos de *todas* las otras $K-1$ variables explicativas en el modelo de regresión múltiple para construir los residuos MCO de $\hat{r}_y$ y $\hat{r}_K$. Por construcción, estos residuos MCO son ortogonales a $(X_1, X_2, \ldots, X_{K-1})$. Es decir, tenemos $X_j' \hat{r}_y = X_j' \hat{r}_K = 0$ para $j = 1, 2, \ldots, K-1$.

Clase Computacional: Aplicación Simulada Simula un modelo:

$$y_i = 2 + 3x_{2i} + 4x_{3i} + u_i$$

donde $u_i \sim \mathcal{N}(0, 3)$, $x_{2i} \sim U[0, 1]$, $x_{3i} = x_{2i} + \eta_i$, y $\eta_i \sim U[0, 1]$.

1. Estima los coeficientes $\hat{\beta}_1, \hat{\beta}_2, \hat{\beta}_3$ del modelo simulado.
2. Estandariza cada variable (y_i, x_{2i}, x_{3i}) de la siguiente forma: $(y_i - \bar{y}_i)/\sigma_{y_i}$ (replicando la misma fórmula para x_{2i} y x_{3i}. ¿Cuál es el promedio y desviación estándar de cada variable reparametrizada?
3. Estime el modelo anterior, pero ahora utilizando las variables reparametrizadas. ¿Cómo se comparan los coeficientes? ¿Por qué?
4. Aplica el teorema Frisch-Waugh-Lovell con la variable x_{3i} (es decir, estima el modelo sin x_{3i}, concentrando x_{3i} de cada otra variable). ¿Son idénticos los coeficientes?
5. Ahora, estima el modelo original, simplemente omitiendo x_{3i}. ¿Son idénticas los coeficientes?

6. Si la respuesta a las preguntas anteriores no es afirmitiva, demuestra (computacionalmente) a qué se debe la diferencia.

4.4 Inferencia

4.4.1 Intervalos y Tests de Hipótesis Acerca de Un parámetro

El estimador MCO nos produce **estimaciones puntuales** para los β_K parámetros en nuestro modelo. Por ejemplo, si estimamos $\hat{\beta}_1 = 1.081$, nuestra mejor estimación es que el valor para el parámetro β_1 es 1.081. Pero por supuesto, β_1 es un valor poblacional, y algo que nunca podemos saber con certeza. Y aún más, sabemos que esta mejor aproximación está casi seguramente equivocada. El valor poblacional se refiere al valor en una población teórica, que nunca podemos observar completamente. Aunque nuestro estimador está formado a partir de una muestra representativa de esta población, la varianza muestral, o la aleatoridad de la muestra, implicará que el estimador puntual típicamente será distinto al valor poblacional.

Dada esta inseguridad, generalmente nos interesa especificar un intervalo de valores que contienen al parámetro verdadero con una probabilidad especificada. Esta es la idea de la **estimación de intervalos**. De manera relacionada, si estimamos que $\hat{\beta}_1 = 1.081$, podríamos querer comprobar si el valor verdadero de β_1 podría posiblemente ser 1. Esta es la idea de los **contrastes de hipótesis**. En esta sección, revisaremos los detalles formales detrás de la construcción de intervalos de confianza, y la realización de contrastes de hipótesis en modelos lineales de regresión.

Distribución del Estimador Nuestro resultado de que la varianza condicional $V(\hat{\beta}_{MCO}|X) = \sigma^2(X'X)^{-1}$ en el modelo clásico de regresión lineal con regresores estocásticos nos será útil para la estimación de intervalos y los test de hipótesis. Sin embargo, sabiendo la expectativa condicional y la varianza del estimador MCO no será suficiente. Hay infinitas posibles distribuciones que pueden cumplir con una media y varianza específica, pero para contruir intervalos de confianza y contrastes de hipótesis, vamos a necesitar información acerca de distintos cuantiles de la distribución: algo que no sabemos con solamente un término para la media y varianza. Por ende, necesitamos tener información acerca de la **distribución** condicional del estimador MCO para proceder con la estimación de intervalos de confianza y realizar test de hipótesis.

Exploramos dos maneras potenciales de avanzar. La primera es agregando un supuesto acerca de la distribución de $u|X$ (o de manera equivalente, de $y|X$). En la práctica, *se supone* que $u|X$ está distribuida según una Normal en cualquier muestra, incluyendo muestras finitas. Este supuesto nos permite derivar la distribución exacta de $\hat{\beta}_{MCO}|X$. La segunda manera potencial de avanzar hacía informacional distribucional es no tomar ningún supuesto acerca de la distribución de $u|X$, pero apelar a un argumento asintótico, y esperar que nuestra muestra sea suficientemente grande

para ser aproximada por supuestos asintóticos. Sin especificar la forma de $u|X$, aún así podemos caracterizar la distribución de $\hat{\beta}_{MCO}$ en la medida que $N \to \infty$.

Un motivo particular para suponer que los términos del error estocástico vienen de una distribución Normal en el límite (sin un supuesto explícito que sea así) es que el término de error captura el efecto combinado de muchos factores diferentes que no están incluidos en nuestro modelo. Entonces, podemos citar el TLC que sugiere que la suma de muchas variables independientes tendrá una distribución Normal

Cuando agregamos el supuesto que $u|X$ tiene una distribución Normal a los supuestos anteriores para especificar el modelo clásico de regresión lineal con regresores estocásticos nos da **el modelo clásico de regresión lineal con errores normales**. Este modelo, se especifica de la siguiente forma:

$$y = X\beta + u$$

$$u|X \sim \mathcal{N}(0, \sigma^2 I)$$

X es estocástico y de rango completo

o, de manera equivalente:

$$y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$$

X es estocástico y de rango completo.

Este modelo clásico con errores normales es una extensión del modelo introducido en la sección 4.2.1. Ahora, hemos agregado un supuesto distribucional completo, en vez de solamente un supuesto acerca de la media y varianza. Desde luego, los supuestos detrás del modelo actual son más fuertes, y es necesario cuestionar si los supuestos son razonables. Volveremos a este punto en el capítulo 5.

El resultado clave que surge de este modelo es:

$$\hat{\beta}_{MCO}|X \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1}). \tag{4.26}$$

Ya habíamos visto que $E(\hat{\beta}_{MCO}|X) = \beta$ y $V(\hat{\beta}_{MCO}|X) = \sigma^2(X'X)^{-1}$, y estos los resultados los pudimos comprobar en el modelo clásico de regresión lineal *sin* tener que utilizar el supuesto de errores normales. La parte que hemos agregado es que el vector aleatorio $\hat{\beta}_{MCO}|X$ también tiene una distribución Normal. Este resultado viene de la propiedad que, condicional en X, $\hat{\beta}_{MCO}|X$ es una función lineal del vector aleatorio $y|X$. Y hemos tomado el supuesto que $y|X$ tiene una

distribución Normal.[12]

Una implicancia inmediata de la propiedad de Distribución Marginal de la normal multivariada (revisa la sección 3.1.4) es que dado $\hat{\beta}_{MCO}|X \sim \mathcal{N}(\beta, \sigma^2(X'X)^{-1})$, tenemos:

$$\hat{\beta}_k|X \sim \mathcal{N}(\beta_k, v_{kk}) \text{ para } k = 1, \ldots, K$$

donde $\hat{\beta}_k$ es el elemento k-ésimo de $\hat{\beta}_{MCO}$, β_k es el elemento k-ésimo del vector de parámetros verdaderos β, y $v_{kk} = V(\hat{\beta}_k|X)$ es el elemento en la fila k y columna k de $V(\hat{\beta}_{MCO}|X) = \sigma^2(X'X)^{-1}$. El supuesto de normalidad de $u|X$ implica que cada elemento de $\hat{\beta}_{MCO}$ sigue una distribución normal.

Para avanzar con la generación de intervalos de confianza y contrastes de hipótesis, notemos que la función lineal de $\hat{\beta}_k$:

$$z_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) = \left(\frac{-\beta_k}{\sqrt{v_{kk}}}\right) + \left(\frac{1}{\sqrt{v_{kk}}}\right)\hat{\beta}_k$$

tiene una distribución Normal estandarizada:

$$z_k|X \sim \mathcal{N}(0, 1) \text{ para } k = 1, \ldots, K.$$

Además, como la distribución condicional de $z_k|X$ es idéntica para cualquiera realización de X, también tenemos que la distribución no condicional de z_k es una Normal estandarizada:

$$z_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}\right) \sim \mathcal{N}(0, 1) \text{ para } k = 1, \ldots, K. \tag{4.27}$$

En contraste, la distribución condicional de $\hat{\beta}_k \sim \mathcal{N}(\beta_k, v_{kk})$ sí depende de la realización de X, y por lo tanto no tenemos una distribución Normal para el parámetro estimado $\hat{\beta}_k$ (en el modelo con regresores estocásticos). Por lo mismo, siempre vamos a utilizar la variable estandarizada en la contrucción de intervalos y estadísticos de prueba.

Si el parámetro de la varianza fuese conocida, podríamos utilizar el resultado de 4.27 directamente. Por el momento, imaginaremos que σ^2 es conocida, y por lo tanto conocemos $V(\hat{\beta}_{MCO}|X) = \sigma^2(X'X)^{-1}$ y v_{kk}. Después consideramos el caso (mucho más razonable) en que v_{kk} no es conocida.

[12] Utilizamos el hecho de que si el vector y de $N \times 1$ es estocástico condicional en X con $y|X \sim \mathcal{N}(\mu, \Sigma)$, y si el vector a es no-estocástico condicional en X y la matriz B de $K \times N$ es no-estocástico condicional en X y con rango completo de fila (rango$(B) = K$), entonces el vector de $K \times 1$ $z = a + By$ es estocástico condicional en X, con $z|X \sim \mathcal{N}(a + B\mu, B\Sigma B')$. Dado que $\hat{\beta}_{MCO} = (X'X)^{-1}X'y = Ay$ con un A no-estocástico condicional en X y rango$(A) = K$, podemos utilizar este hecho para obtener la distribución condicional de $\hat{\beta}_{MCO}|X$. Dejamos como una actividad el paso entre esta nota de pie, y el resultado dado en ecuación 4.26.

Inferencia con Varianza Conocida

Intervalos de Confianza con Varianza Conocida Para crear un intervalo de confianza de x% (donde x es cualquier porcentaje entre 0 y 100), necesitamos considerar la masa de distribución de una variable normal estandarizada. Por ejemplo, si nos interesa construir un intervalo de confianza de 95%, partimos con el resultados que 95% de la masa de probabilidad en una variable normal estandarizada cae entre -1.96 y 1.96 (ver tabla 6.1). Partiendo con esto, y la ecuación 4.27, podemos derivar el siguiente resultado:

$$\begin{aligned} P(-1.96 < z_k < 1.96) &= 0.95 \\ P\left(-1.96 < \frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} < 1.96\right) &= 0.95 \\ P(-1.96\sqrt{v_{kk}} < \hat{\beta}_k - \beta_k < 1.96\sqrt{v_{kk}}) &= 0.95 \\ P(-1.96\sqrt{v_{kk}} < \beta_k - \hat{\beta}_k < 1.96\sqrt{v_{kk}}) &= 0.95 \\ P(\hat{\beta}_k - 1.96\sqrt{v_{kk}} < \beta_k < \hat{\beta}_k + 1.96\sqrt{v_{kk}}) &= 0.95 \end{aligned} \tag{4.28}$$

Los primeros pasos utilizan álgebra simple. El penúltimo paso utiliza la simetría del intervalo alrededor de cero. La ecuación 4.28 nos da el **intervalo de confianza de 95%** para el valor verdadero de β_k como:

$$\hat{\beta}_k \pm 1.96\sqrt{v_{kk}}.$$

Contrastes de Hipótesis con Varianza Conocida Para realizar un contraste de hipíotesis acerca del valor verdadero de β_k, utilizamos el mismo z_k como estadístico de prueba. Supongamos que queremos comprobar si β_k podría ser algún candidato β_k^0. Esto equivale a un contraste de hipótesis de la siguiente forma:

$$\begin{aligned} H_0 : \beta_k &= \beta_k^0 \\ H_1 : \beta_k &\neq \beta_k^0 \end{aligned} \tag{4.29}$$

En este caso, sabemos que bajo la hipótesis nula, la variable:

$$z_k = \frac{\hat{\beta}_k - \beta_k^0}{\sqrt{v_{kk}}}$$

tendrá una distribución Normal estandarizada. Esta z_k es nuestro estadístico de prueba. Si con nuestro estimador $\hat{\beta}_k$, observamos un valor de z_k que parece ser poco probable para una variable Normal estandarizada, concluiremos que es poco probable que la hipótesis sea cierta, y optaremos por rechazar la hipótesis nula. En el caso opuesto, no rechazamos la hipótesis nula. Los demás detalles del contraste de hipótesis son idénticos a los detalles revisados en la sección 3.4.2 de este documento. Dejamos algunos ejemplos aplicados al final de esta sección.

Inferencia con Varianza Desconocida

Utilizamos una metodología parecida en el caso más realista en que el parámetro de varianza σ^2 no es conocido. La diferencia es que además de estimar el parámetro $\hat{\beta}_k$, tendremos que estimar el parámetro de varianza σ^2 utilizando nuestros datos. Así obtenemos una estimación para $V(\hat{\beta}_{MCO}|X)$ y v_{kk}. Para poder estimar σ^2, podemos ocupar el estimador MCO: $\hat{\sigma}^2 = \frac{\hat{u}'\hat{u}}{N-K}$ y estimar $V(\hat{\beta}_{MCO})$ utilizando $\hat{V}(\hat{\beta}_{MCO}|X) = \hat{\sigma}^2(X'X)^{-1}$. En la práctica, el hecho que ahora estamos trabajando con una $\hat{v}_{kk}$ estimada implica algunas complicaciones al formar un estadístico de prueba con una distribución conocida.

Estadísticos de Prueba con Varianza Desconocida Llamando al elemento k-ésimo de la diagonal principal de $\hat{V}(\hat{\beta}_{MCO}|X)$ como $\hat{v}_{kk}$, utilizaremos el estadístico estandarizado:

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}} = \frac{\hat{\beta}_k - \beta_k}{ee_k} \tag{4.30}$$

donde $ee_k = \sqrt{\hat{v}_{kk}}$ se conoce como el error estándar del parámetro estimado $\hat{\beta}_k$. Sin embargo, reemplazando el desconocido $\sqrt{v_{kk}}$ con una estimación $ee_k = \sqrt{\hat{v}_{kk}}$ cambia la distribución de muestra finita del estadístico, y entonces es necesario derivar la distribución de t_k.

Recordamos de la sección 3.1.4 que una variable que es la razón entre una variable normal estandarizada y la raíz cuadarda de una variable chi-cuadrado dividida en sus grados de libertad (ambas independientes entre si), sigue una distribución t. En lo que sigue, vamos a demostrar que la variable t_k de ecuación 4.30 se define así.

Primero, notamos que siempre podemos reparametrizar el componente estocástico del modelo de regresión lineal clásico como una función de una variable aleatoria normal estandarizada, escribiendo el modelo de la siguiente forma:

$$y = X\beta + \sigma z, \quad \text{donde } u = \sigma z \quad \text{con } z|X \sim \mathcal{N}(0,1).$$

Y recordamos que $\hat{u} = My$, y $MX = 0$. Entonces, utilizando esta misma reparametrización, podemos escribir $\hat{u}$ y $\hat{u}'\hat{u}$ como:

$$\hat{u} = My = M[X\beta + \sigma z] = MX\beta + M\sigma z = \sigma Mz \tag{4.31}$$

$$\hat{u}'\hat{u} = (\sigma Mz)'(\sigma Mz) = \sigma^2(z'M')Mz = \sigma^2 z'MMz = \sigma^2 z'Mz \tag{4.32}$$

Ahora, definimos una variable:

$$w_0 = \frac{\hat{u}'\hat{u}}{\sigma^2} = \frac{\sigma^2 z'Mz}{\sigma^2} = z'Mz \tag{4.33}$$

La matriz M tiene rango de $N-K$, y $z|X$ es una variable normal estándar. Entonces, por definición

(por la propiedad 2 de la forma cuadrática de vectores Normales introducido en sección 3.1.4), $w_0|X \sim \chi^2(N-K)$. Esto es un resultado clave para establecer la distribución t_k.

Si volvemos a t_k, un poco de álgebra nos da:

$$t_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}} = \frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}} \frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}} = \frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}} z_k \tag{4.34}$$

donde $z_k|X \sim \mathcal{N}(0,1)$. Y notamos que $V(\hat{\beta}_{MCO}|X)$ y su estimador $\hat{V}(\hat{\beta}_{MCO}|X)$ solo difieren en el término σ:

$$V(\hat{\beta}_{MCO}|X) = \sigma^2(X'X)^{-1} = \sigma^2 Q^{-1} \tag{4.35}$$

$$\hat{V}(\hat{\beta}_{MCO}|X) = \hat{\sigma}^2(X'X)^{-1} = \hat{\sigma}^2 Q^{-1}. \tag{4.36}$$

Lo mismo ocurre con v_{kk} y $\hat{v}_{kk}$, los elementos k-ésimos del diagonal principal de la matriz de varianza covarianza:

$$v_{kk} = \sigma^2 q_{kk} \tag{4.37}$$

$$\hat{v}_{kk} = \hat{\sigma}^2 q_{kk}. \tag{4.38}$$

Con todo esto,

$$\frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}} = \frac{\sigma\sqrt{q_{kk}}}{\hat{\sigma}\sqrt{q_{kk}}} = \frac{\sigma}{\hat{\sigma}},$$

implicando que podemos escribir la ecuación 4.34 como:

$$t_k = \frac{z_k}{\hat{\sigma}/\sigma}.$$

La distribución del numerador de 4.34 es conocida (una variable normal estandarizada). Utilizando nuestro estimador MCO de $\hat{\sigma}^2$, el cuadrado del denominador se escribe:

$$\frac{\hat{\sigma}^2}{\sigma^2} = \left(\frac{1}{\sigma^2}\right)\left(\frac{\hat{u}'\hat{u}}{N-K}\right) = \left(\frac{1}{N-K}\right)\left(\frac{\hat{u}'\hat{u}}{\sigma^2}\right) = \frac{w_0}{N-K}, \tag{4.39}$$

donde anteriormente hemos demostrado que $w_0|X \sim \chi^2(N-K)$. Entonces, juntando todos los resultados anteriores, podemos escribir t_k de la siguiente forma:

$$t_k = \frac{\sqrt{v_{kk}}}{\sqrt{\hat{v}_{kk}}} z_k = \frac{\sigma}{\hat{\sigma}} z_k = \frac{z_k}{(\hat{\sigma}/\sigma)} = \frac{z_k}{\sqrt{(\hat{\sigma}^2/\sigma^2)}} = \frac{z_k}{\sqrt{w_0/(N-K)}},$$

donde ahora es claro que t_k es el razon de una variable normal estandarizada, y la raíz cuadrada de una variable χ^2 dividido en sus grados de libertad. Además, se puede demostrar[13] que el numerador

[13] La independencia viene del hecho que cualquier función de $\hat{u}|X$ y de $\hat{\beta}_{MCO}|X$ son independientes (no demostramos este resultado aquí, pero está disponible, por ejemplo, en Goldberger (1991, p. 224)). La variable $z_k = \frac{\hat{\beta}_k - \beta_k}{\sqrt{v_{kk}}}$ es una

y denominador son independientes entre sí, resultando que el escalar

$$t_k = \left(\frac{\hat{\beta}_k - \beta_k}{\sqrt{\hat{v}_{kk}}} \right) = \frac{z_k}{\sqrt{w_0/(N-K)}} \sim t_{N-K},$$

se distribuye como una variable t de Student.

Utilizamos este resultado para formar **intervalos de confianza** y para construir **test de hipótesis** acerca de los parámetros en el modelo de regresión lineal con σ desconocida. En ambos casos, no hay ninguna diferencia mecánica entre los procesos discutidos en inferencia en el modelo con varianza conocida; solamente en este caso los valores críticos se observan de la distribución t. Entonces, por ejemplo, el intervalo de confianza se construye de la siguiente forma, donde c refiere al valor crítico de interés:

$$P(\hat{\beta}_k - c_{0.025}(N-K)ee_k < \beta_k < \hat{\beta}_k + c_{0.025}(N-K)ee_k) = 0.95.$$

En el caso de una regresión con $N - K = 50$ grados de libertad, este valor es 2.004 (versus 1.96 para el caso con varianza conocidad). Y para realizar un contraste de hipótesis, el estadístico de prueba

$$t_k = \frac{\hat{\beta}_k - \beta_k^0}{ee_k}$$

tiene una distribución conocida bajo la nula $H_0 : \beta_k = \beta_k^0$. Nuevamente, el valor crítico de rechazo viene de la distribución t de Student (ver Tabla 6.2 para la tabulación de algunos de estos valores).

4.4.2 Combinaciones Lineales de parámetros

Hasta ahora hemos derivado resultados para poder realizar un test de hipótesis acerca de un **elemento individual** del vector de parámetros β. Sin embargo, a menudo estamos interesados en realizar un test acerca de una combinación lineal de parámetros, o una hipotesis conjunta. Por ejemplo, si estimamos un modelo de la siguiente forma:

$$ln(salario)_i = \beta_1 + \beta_2 educBasica_i + \beta_3 educMedia_i + \beta_4 educUniversitario_i + u_i$$

si queremos comprobar la **hipótesis conjunta** de un retorno de educación igual a cero, esta hipótesis consiste en la nula: $H_0 : \beta_2 = \beta_3 = \beta_4 = 0$. También en otros contextos se podría querer comprobar una **hipotesis de una combinación lineal de parámetros**, del estilo $\beta_1 + \beta_2 = 1$. Para realizar esta clase más general de restricciones lineales en el modelo, necesitamos derivar un estadístico de prueba más general.

Para este nueva clase de contrastes de hipótesis, vamos a introducir una nueva herramienta. Esta herramienta es una matriz simple que nos permite seleccionar combinaciones lineales o conjuntas

función de $\hat{\beta}_{MCO}$, y la variable $w_0 = \frac{\hat{u}'\hat{u}}{\sigma^2}$ es una función de $\hat{u}$, dando independencia.

de parámetros del vector de parámetros $\hat{\beta}_{MCO}$. Esta matriz lo llamaremos H (en la notación de Cameron and Trivedi (2005) esta matriz se llama R), cuya dimensión es de $p \times K$. Como siempre, K refiere a la cantidad de parámetros en $\hat{\beta}_{MCO}$, y p depende del estilo de contraste de hipótesis que queremos realizar. La idea de H es que nos permitirá formar un vector aleatorio $\hat{\theta}_{MCO}$ en base a los parámetros de interés de $\hat{\beta}_{MCO}$. Especificamente,

$$\hat{\theta}_{MCO} = H\hat{\beta}_{MCO}$$

donde $\hat{\theta}_{MCO}$ es un vector aleatorio de $p \times 1$.

Para ver un ejemplo, consideramos un modelo con $K = 3$, $y = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$. Y consideramos dos distintas hipótesis de interés:

Hipótesis A:

$$\begin{aligned} H_0 &: \beta_2 + \beta_3 = 0 \\ H_1 &: \beta_2 + \beta_3 \neq 0 \end{aligned} \tag{4.40}$$

Hipótesis B:

$$\begin{aligned} H_0 &: \quad \beta_2 = \beta_3 = 0 \\ H_1 &: \quad H_0 \text{ no es cierta} \end{aligned} \tag{4.41}$$

El primer caso es una hipótesis acerca de una combinación lineal de parámetros (con una sola restricción), y el segundo caso es una hipótesis conjunta acerca de dos parámetros (con dos restricciones). Para alguna matriz no estocástica H de $p \times K$ con rango$(H) = p$ (rango completo de fila). Nuestra meta con la matriz H es formar un vector de parámetros que nos permite crear la combinación lineal para la hipótesis A, y dos distintos parámetros para la hipótesis B.

En el primer caso, consideramos la matriz de 1×3:

$$H = \begin{pmatrix} 0 & 1 & 1 \end{pmatrix}.$$

Al multiplicar H por $\hat{\beta}_{MCO}$ tenemos:

$$\begin{aligned} \hat{\theta}_{MCO} &= H\hat{\beta}_{MCO} \\ \hat{\theta}_{MCO} &= \begin{pmatrix} 0 & 1 & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} \\ \hat{\theta}_{MCO} &= \hat{\beta}_2 + \hat{\beta}_3, \end{aligned}$$

que es la combinación lineal de interés. En el segundo caso, podemos formar la matriz de 2×3:

$$H = \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

y siguiendo el mismo proceso da

$$\begin{aligned} \hat{\theta}_{MCO} &= H\hat{\beta}_{MCO} \\ \hat{\theta}_{MCO} &= \begin{pmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \hat{\beta}_1 \\ \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} \\ \hat{\theta}_{MCO} &= \begin{pmatrix} \hat{\beta}_2 \\ \hat{\beta}_3 \end{pmatrix} \end{aligned}$$

que son los dos parámetros relevantes para las dos restricciones impuestas en la hipótesis B. Siguendo este mismo procedimiento, se puede formar cualquier combinación lineal de parámetros y/o hipótesis conjunta, siempre cuando la matriz H tenga K filas. Volveremos a una serie de ejemplos en las preguntas al final de esta sección. Por último, es importante destacar que en cada caso, la cantidad de filas p se refiere a la cantidad de restricciones impuestas en la hipótesis nula, (la cantidad de signos de igualdad). En cada caso, $rango(H) = p$.

$\hat{\theta}_{MCO}$ es un estimador que proviene del modelo de regresión lineal. Como se basa en el estimador insesgado $\hat{\beta}_{MCO}$, también es un estimador insesgado bajo los mismos supuestos clásicos:

$$E(\hat{\theta}_{MCO}|X) = HE(\hat{\beta}_{MCO}|X) = H\beta = \theta.$$

La varianza se escribe de la siguiente forma:

$$V(\hat{\theta}_{MCO}) = HV(\hat{\beta}_{MCO}|X)H' = H[\sigma^2(X'X)^{-1}]H' = \sigma^2 H(X'X)^{-1}H' = \sigma^2 D^{-1}$$

donde $D = H(X'X)^{-1}H'$ es no estocástico condicional en X. Por linealidad del operador H, y el hecho de que una transformación lineal de un vector aleatorio normal también tiene una distribución normal tenemos:

$$\hat{\theta}_{MCO}|X \sim \mathcal{N}(\theta, \sigma^2 D^{-1}).$$

Por último, definamos al estimador MCO de la varianza de $\hat{\theta}_{MCO}$ como:

$$\hat{V}(\hat{\theta}_{MCO}) = H\hat{V}(\hat{\beta}_{MCO}|X)H' = H[\hat{\sigma}^2(X'X)^{-1}]H' = \hat{\sigma}^2 H(X'X)^{-1}H' = \hat{\sigma}^2 D^{-1}.$$

Utilizando la propiedad 1 de la forma cuadrática de los vectores Normales introducida en sec-

ción 3.1.4, definimos una variable $w|X \sim \chi^2(p)$:

$$\begin{aligned} w &= (\hat{\theta}_{MCO} - \theta)'[\sigma^2 D^{-1}]^{-1}(\hat{\theta}_{MCO} - \theta) \\ &= \left(\frac{1}{\sigma^2}\right)(\hat{\theta}_{MCO} - \theta)'D(\hat{\theta}_{MCO} - \theta). \end{aligned} \tag{4.42}$$

Y consideremos su versión estimada $\hat{w}$:

$$\hat{w} = \frac{1}{\hat{\sigma}^2}(\hat{\theta}_{MCO} - \theta)'D(\hat{\theta}_{MCO} - \theta) \tag{4.43}$$

y

$$v = \frac{\hat{w}}{p} = \left(\frac{1}{p\hat{\sigma}^2}\right)(\hat{\theta}_{MCO} - \theta)'D(\hat{\theta}_{MCO} - \theta).$$

A diferencia de la ecuación 4.42, la ecuación 4.43 se basa en un estimador para el término σ^2, y por lo tanto ya no sigue una distribución χ^2. Dado que la diferencia entre $\hat{w}$ y w viene del estimador de σ^2, podemos re-escribir este estadístico como:

$$v = \frac{\hat{w}}{p} = \left(\frac{\sigma^2}{\hat{\sigma}^2}\right)\left(\frac{w}{p}\right).$$

Nuevamente, volvemos a tener $\sigma^2/\hat{\sigma}^2$. En la ecucación 4.39 demostramos que su inverso es igual a $\hat{\sigma}^2/\sigma^2 = w_0/(N-K)$ donde $w_0 = (\hat{u}'\hat{u})/\sigma^2$, y donde w_0 es una variable distribuida según una χ^2 con $N-k$ grados de libertad. Entonces,

$$v = \left(\frac{\sigma^2}{\hat{\sigma}^2}\right)\left(\frac{w}{p}\right) = \frac{(w/p)}{(\hat{\sigma}^2/\sigma^2)} = \frac{(w/p)}{(w_0/(N-K))}$$

es la razón de dos variables aleatorias chi-cuadrado, cada uno divida por sus grados de libertad. Además, $w|X \sim \chi^2(p)$ es una función de $\hat{\beta}_{MCO}$, y $w_0|X \sim \chi^2(N-K)$ es una función de $\hat{u}$, implicando que las dos variables son independientes.

Junto, esto implica que

$$v|X \sim F_{p,N-K},$$

donde p son la cantidad de grados de libertad del numerador (la cantidad de restricciones impuestas por el contraste múltiple), y $N-K$ son los grados de libertado del denominador. Y adema, dado que esta distribución condicional es la misma (F con p y $N-K$ grados de libertad) para cualquier realización de X, de nuevo tenemos que la distribución no condicional es la misma:

$$v \sim F_{p,N-K}.$$

Hemos demostrado que la variable aleatoria escalar:

$$\begin{aligned} v &= \left(\frac{1}{p\hat{\sigma}^2}\right)(\hat{\theta}_{MCO}-\theta)'D(\hat{\theta}_{MCO}-\theta) \\ &= \left(\frac{1}{p}\right)(\hat{\theta}_{MCO}-\theta)'[\hat{V}(\hat{\theta}_{MCO}|X)]^{-1}(\hat{\theta}_{MCO}-\theta) \sim F_{p,N-K}, \end{aligned} \tag{4.44}$$

donde $\theta = H\beta$ y $\hat{\theta}_{MCO} = H\hat{\beta}_{MCO}$. Podemos utilizar este resultado para formar el estadístico de prueba para comprobar restricciones lineales involucrando uno o más de los elementos del vector de parámetros β. Exploramos unos ejemplos más tarde en esta sección.

Una restricción como caso especial Esta derivación funciona para una cantidad arbitraria de p (siempre y cuando $p < (N-K)$). Sin embargo, notamos que el caso de una sola restricción es un caso especial. Consideramos la matriz H de $1 \times K$:

$$H = \begin{pmatrix} 0 & \cdots & 0 & 1 & \cdots & 0 \end{pmatrix}$$

con el escalar 1 como su elemento k-ésimo y ceros en las otras posiciones. En esta situación, tenemos $\theta = H\beta = \beta_k$, y $\hat{\theta}_{MCO} = \hat{\beta}_k$. Y nuestra estadística de prueba v se convierte en:

$$\begin{aligned} v &= \left(\frac{1}{p}\right)(\hat{\theta}_{MCO}-\theta)'[\hat{V}(\hat{\theta}_{MCO}|X)]^{-1}(\hat{\theta}_{MCO}-\theta) \\ &= \frac{(\hat{\beta}_k-\beta_k)^2}{\hat{v}_{kk}} \\ &= t_k^2 \sim F_{1,N-K}. \end{aligned}$$

Esto en realidad es un resultado más general: una variable aleatoria con distribución t_{N-K} al cuadrado tiene una distribución $F_{1,N-K}$.[14]

Realizando Contrastes de Hipótesis Múltiples Para hacer un contraste de hipótesis acerca de p combinaciones lineales de los parámetros β_k, formulamos la hipótesis nula de la forma:

$$H_0 : H\beta = \theta^0$$

[14]COULD LEAVE THIS AS A QUESTION? Si las variables aleatorias $z \sim \mathcal{N}(0,1)$ y $w \sim \chi^2(n)$ son independientes, entonces el escalar:

$$u = \frac{z}{\sqrt{(w/n)}} \sim t(n).$$

La razón de una variable aleatoria Normal estandarizada y la raíz cuadrada de una variable chi-cuadrado dividida por sus grados de libertad *independientes* tiene una distribución **t de Student** con la misma cantidad de grados de libertad. Y dado que $z^2 \sim \chi^2(1)$, la distribución de esta variable u al cuadrado es:

$$u^2 = \frac{z^2}{(w/n)} = \frac{(z^2/1)}{(w/n)} \sim F_{1,n}$$

demostrando que una variable t_n al cuadrado tiene una distribución $F_{1,n}$

donde H es una matriz no estocástica de $p \times K$ con rango$(H) = p$. La hipótesis alternativa es simplemente:

$$H_1 : H\beta \neq \theta^0.$$

También se puede escribir la hipótesis nula y alternativa en forma extensiva como (por ejemplo) $H_0 : \beta_1 = 0$ y $\beta_2 = 0$ versus $H_1 : H_0$ no es verdadera.

Para realizar el contraste de hipótesis, utilizamos el estimador $\hat{\theta}_{MCO} = H\hat{\beta}_{MCO}$ y construimos la estadística de prueba escalar:

$$v = \left(\frac{1}{p}\right)(\hat{\theta}_{MCO} - \theta^0)'[\hat{V}(\hat{\theta}_{MCO}|X)]^{-1}(\hat{\theta}_{MCO} - \theta^0)$$

donde $\hat{V}(\hat{\theta}_{MCO}|X) = H\hat{V}(\hat{\beta}_{MCO}|X)H' = \hat{\sigma}^2 H(X'X)^{-1}H'$ en el modelo clásico de regresión lineal con errores distribuidos normalmente. Si la nula es cierta, sabemos que:

$$v \sim F_{p,N-K}$$

y el contraste de hipótesis consiste en preguntar si el valor de v que obtenemos es "poco probable" dada esta distribución. Rechazamos $H_0 : H\beta = \theta^0$ a favor de $H_1 : H\beta \neq \theta^0$, si la probabilidad de observar el valor v de la distribución F es menor que nuestro nivel de significancia elegido. Elegimos el nivel de significancia F-crítico en base a la cantidad de grados de libertad p y $N - K$, y además un α elegido por el/la investigador/a, que mide la tasa de error tipo-I aceptada. En la Tabla 6.3 tabulamos valores críticos para un $\alpha = 0.05$ con una distinta cantidad de p y $N - K$. Notemos que la forma cuadrática implica que el escalar v es positivo, así que solamente tenemos que considerar la cola de arriba de la distribución F.

Actividad Computacional: Similando contrastes conjuntos

Suponga un proceso generador de datos:

$$y = 5 + 2x_1 + 1x_2 + 0x_3 + \varepsilon_i$$

donde $E[\varepsilon_i x_k] = 0$ para cada k. Simulemos 2000 observaciones de las variables x_k con la siugiente forma:

$$x_1 \sim U[0,1] \quad x_2 \sim U[0,2] \quad x_3 \sim U[0,3].$$

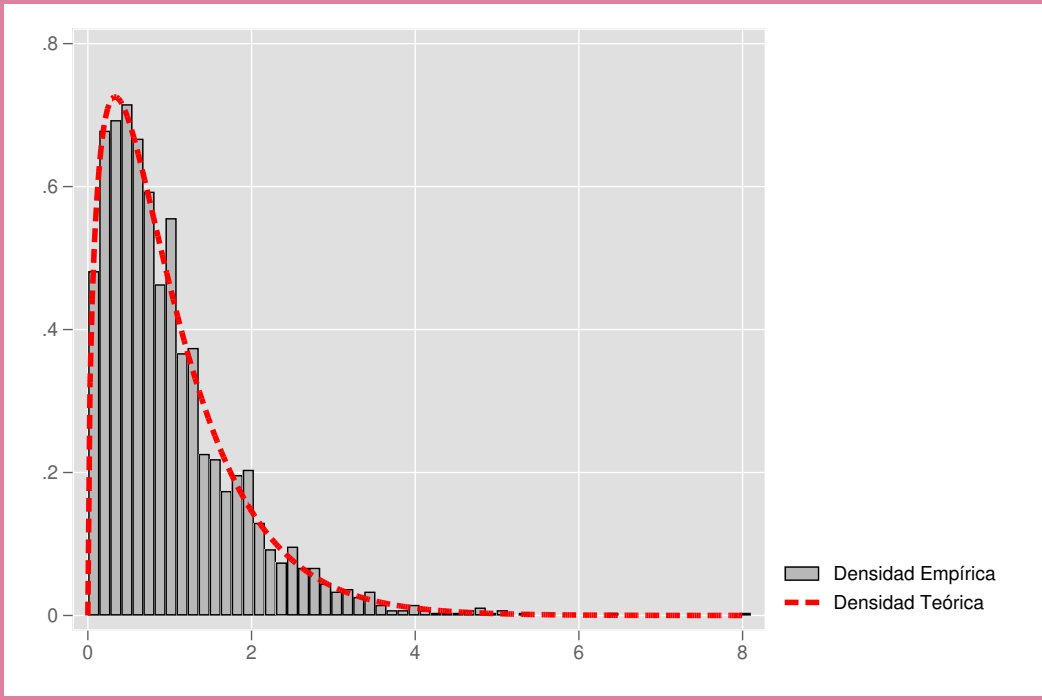
En estos ejercicios, nos interesa realizar el contraste conjunto:

$$\begin{aligned} H_0 &: \beta_1 = 2 \text{ y } \beta_2 + \beta_3 = 1 \text{ y } \beta_0 = 5 \\ H_1 &: H_0 \text{ no es cierta .} \end{aligned}$$

Aquí, β_0 refiere al término de constante, y β_1, β_2, y β_3 refieren a las coeficientes sobre x_1, x_2, y x_3 respectivamente.

1. Repetimos 1500 veces el proceso generador de datos. En cada repetición, genera $\varepsilon \sim \mathcal{N}(0, 3)$ y y siguiendo el modelo poblacional descrito arriba. En cada caso, estima el vector de parámetros $\hat{\beta}_{MCO}$ y la matriz de varianza-covarianza $\hat{V}(\hat{\beta}_{MCO})$. Por último realiza el test de hipótesis conjunta, y calcula el valor del estadístico de prueba v utilizando el método matricial y la matriz H. Guarda las 1500 valores de v calculadas.

2. Grafica la distribución empírica de las realizaciones v, y además una distribución teórica $F(df_1, df_2)$, donde df_1 y df_2 refieren a los grados de libertad relevantes (ver ejemplo a continuación).



3. Encuentra el valor $F^{-1}(df_1, df_2)_{0.05}$ (es decir, el valor crítico de rechazo a un 5%). ¿Cuántas observaciones de $v > F^{-1}(df_1, df_2)_{0.05}$? ¿Qué porcentaje de realizaciones? Comenta.

4. (Difícil) Compare formalmente la distribución empírica graficada con la distribución teórica. Realiza un contraste formal de igualdad de las distribuciones. *Pista:* Una manera de realizar un constrate formal de igualdad de dos distribuciones es mediante la prueba de Kolmogorov-Smirnov.

Utilizando Trasformaciones Lineales Para Test de Hipótesis Simples

En algunos casos, por ejemplo el contraste de hipótesis indicada en la pregunta 2 de los ejercicios anteriores, existe otra manera posibile de realizar el test: re-parametrizar el modelo para hacer el test con un sólo coeficiente de la regresión. Estas re-parametrizaciones sirven en el caso de hipótesis conjuntas acerca de una sola restricción en el modelo.

Por ejemplo, consideramos el modelo:

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i \tag{4.45}$$

y un caso en que queremos realizar el contraste $H_0 : \beta_2 + \beta_3 = 1$ contra $H_1 : \beta_2 + \beta_3 \neq 1$. Notamos, además, que la hipótesis nula también se puede escribir como $\beta_2 + \beta_3 - 1 = 0$. Hay una manera para re-parametrizar (o reorganizar) el modelo para tener *un coeficiente* que es $\beta_2 + \beta_3 - 1$, y así podemos simplemente hacer un test-t acerca de este parámetro, para comprobar la nula.

Para ver esto, restamos x_{3i} de ambos lados del modelo 4.45, para tener:

$$y_i - x_{3i} = \beta_1 + \beta_2 x_{2i} + (\beta_3 - 1)x_{3i} + u_i.$$

Ahora, restando $\beta_2 x_{3i}$ del segundo término y sumando al tercero término da:

$$y_i - x_{3i} = \beta_1 + \beta_2(x_{2i} - x_{3i}) + (\beta_2 + \beta_3 - 1)x_{3i} + u_i,$$

donde el término sobre x_{3i} ahora es el parámetro de interés para la hipótesis nula. Entonces, una manera realizar el contraste de hipótesis de interés en este caso consiste en:

1. Formar las variables nuevas $(y_i - x_{3i})$ y $(x_{2i} - x_{3i})$
2. Estimar la regresión de $(y_i - x_{3i})$ sobre $(x_{2i} - x_{3i})$ y x_{3i}
3. Realizar un test-t simple, con la nula que el coeficiente sobre x_{3i} es igual a 0.

Utilizando el R^2 Para Comprobar Restricciones de Exclusión El contraste matricial en base a parámetros y la matriz de varianza-covarianza no es la única manera de realizar un contraste de hipótesis múltiple. Una manera alternativa utiliza el poder predictivo del modelo, medido por el R^2. La idea detrás de este contraste es parecido a un contraste de razón de verosimilitudes: se estima dos modelos; uno sin restricciones, y el otro con la nula impuesta, y se compara el R^2 asociado a cada modelo. Si el R^2 cae mucho al imponer la nula, sugiere que el modelo no restringido es mejor, y se rechaza la nula. De otra forma, si el R^2 es parecido en los modelos con y sin las restricciones, sugiere que la hipótesis nula es razonable.

Para ver esto en más detalle, consideremos el modelo lineal de regresión con $K = 4$ y $x_{1i} = 1$ para $i = 1, \ldots, N$.

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i \tag{4.46}$$

Suponemos que queremos hacer un test de hipótesis que $\beta_3 = \beta_4 = 0$. Podríamos formar una estadística de prueba de $F(2, N - K)$ en la manera introducida en la sección 4.4.2. Pero un test equivalente consiste en comparar el R^2 del modelo no-restringido (4.46) y el modelo restringido: $y_i = \beta_1 + \beta_2 x_{2i}$ que tiene las restricciones impuestas.

Denotamos el R^2 del modelo no restringido como R^2_U y el R^2 del modelo restringido como R^2_R, notando que $R^2_U \geq R^2_R$. Podemos escribir el estadístico de prueba como:

$$v = \left(\frac{N-K}{2}\right)\left(\frac{R^2_U - R^2_R}{1-R^2_U}\right) \sim F(2, N-K)$$

y esto extiende de manera natural a p restricciones de exclusión con:

$$v = \left(\frac{N-K}{p}\right)\left(\frac{R^2_U - R^2_R}{1-R^2_U}\right) \sim F(p, N-K) \tag{4.47}$$

Se rechaza la nula, si la exclusión de estas p variables explicativas produce una caída suficientemente grande en el R^2, o si el modelo pierde mucho poder explicativo.

Se puede utilizar el mismo test para comprobar la restricción de que todos los $K-1$ coeficientes de pendientes en el modelo lineal son iguales a cero – es decir si se puede excluir todas las variables explicativas del modelo. En este caso, el modelo restringido es: $y_i = \beta_1 + u_i$ con $R^2_R = 0$. Y el estadístico de prueba se simplifica a:

$$v = \left(\frac{N-K}{K-1}\right)\left(\frac{R^2}{1-R^2}\right) \sim F(K-1, N-K)$$

donde $R^2 = R^2_U$ refiere al R^2 del modelo no restringido. A veces esto se conoce como el test-F del modelo.

Preguntas: Consideremos un modelo de K=3, y $x_{1i} = 1$ para cada $i = 1, \ldots, N$

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

1. ¿Cuáles serían las matrices H para cada una de las siguientes hipótesis? En cada caso indica la matriz H, y la variana de θ_{MCO} que será relevante en el contraste.
   (a) $H_0: \ \beta_1 = 0$
   (b) $H_0: \ \beta_1 = \beta_2 = 0$
   (c) $H_0: \ \beta_1 = 0$ y $\beta_2 + \beta_3 = 0$
   (d) $H_0: \ \beta_2 - \beta_3 = 0$
   (e) $H_0: \ 2\beta_1 = 1$ y $\beta_2 - \beta_3 = 0$
2. Para el contraste de hipótesis que $\beta_2 = 0$ y $\beta_3 = 0$ contra la alternativa que $\beta_2 \neq 0$ o $\beta_3 \neq 0$, (a) Plantea la hipótesis nula, y (b) deriva el estadístico de prueba v en forma matricial.
3. Para el contraste de hipótesis que $\beta_2 + \beta_3 = 1$ contra la alternativa que $\beta_2 + \beta_3 \neq 1$, (a) plantea la hipótesis nula, y (b) deriva el estadístico de prueba v en forma matricial.

4. Para el contraste de hipótesis indicado en la pregunta (2), deriva el estadístico de prueba para un test-t, en vez de un test F. *Pista*: El error estándar asociado a la suma de dos parámetros β_j y β_k es igual a $\sqrt{\hat{v}_{jj} + 2\hat{v}_{jk} + \hat{v}_{kk}}$, donde $\hat{v}$ refiere a elementos de la matriz de varianza-covarianza. ¿Cómo se comparan los estadísticas de prueba para los dos contrastes?

5. (difícil) Demuestra que la variable v en la ecuación 4.47 sigue una distribución F con p y $N - K$ grados de libertad. *Pista*: Parte escribiendo los R^2 en términos de la suma total de los cuadrados del modelo.

4.5 Máxima Verosimilitud y Método de Momentos

4.5.1 Estimación Por Máxima Verosilimitud

La estimación que hemos hecho hasta ahora siempre ha requerido minimizar la suma de los cuadrados de los residuos del modelo de regresión. Es lo que llamamos mínimos cuadrados ordinarios. Sin embargo, utilizando el **mismo modelo** y los **mismos supuestos** (o supuestos más débiles), podemos *estimar* de manera diferente. En este sentido, la interpretación de los coeficientes será la misma y la teoría atrás del modelo será la misma. Simplemente estamos cambiando la mecánica en el punto de estimar.

Explorar la estimación por máxima verosimilitud con un modelo de regresión lineal es, en algún sentido redundante. Dado lo que vimos en el Teorema de Gauss Markov, ya tenemos una manera eficiente para estimar este modelo: utilizando MCO! Pero hay una ventaja importante de aprender máxima verosimilitud con el modelo lineal: en este caso los distintos métodos de estimación (MCO y MV) dan el **mismo resultado**. Este es un resultado que veremos más tarde en esta sección. Así, cuando aprendemos a implementar el estimador de MV en la práctica, podemos estar seguros que está funcionando correctamente si devuelve el mismo vector de parámetros que el estimador MCO.

Sin embargo, en algunos casos, MCO no es una manera fáctible de estimar. Por ejemplo, hay varios modelos no-lineales cuando, por definición, MCO no es un estimador aplicable. En estos casos será *necesario* estimar con alguna otra metodología, y un estimador muy práctico y común en estos casos es el estimador de máxima verosimilitud. Así, al entender el proceso de estimar con MV en un caso "simple" (el modelo lineal), será más fácil extender la aplicación a un caso más complejo. En la práctica, casi siempre estimamos los modelos clásicos con MCO.

Mantenemos los mismos supuestos que en el modelo clásico de regresión lineal con errores normales (definidos en la sección 4.4.1). A diferencia de MCO, el ingrediente fundamental en MV es el supuesto distribucional que hemos tomado acerca de $u|X$. Hemos supuesto:

$$u|X \sim \mathcal{N}(0, \sigma^2 I),$$

que implica que $y|X \sim \mathcal{N}(X\beta, \sigma^2 I)$. Este es nuestro modelo de probabilidad, y como revisamos anteriormente en la sección 3.3.4, si podemos escribir la función de densidad de probabilidad asociada al modelo, podemos maximizarlo para encontrar los parámetros más probables para haber generado los datos observados. La fdp de una variable normal y con esperanza μ y desviación estándar σ es:

$$f(y|\mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(y-\mu)^2}{2\sigma^2}.$$

La gran diferencia entre el modelo lineal y el caso general de máxima verosimilitud con una variable normal que vimos en la sección 3.3.4 es que ahora, en vez de estar parametrizado por un solo término μ, el promedio es parametrizado por todo el vector de parámetros β. Tenemos para cada y_i que $y_i|X_i \sim \mathcal{N}(X\beta, \sigma^2)$, y podemos escribir la fdp de arriba como:

$$f(y_i|X_i; \beta, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(y-X\beta)^2}{2\sigma^2}.$$

Y con observaciones independientes, la fdp conjunta que necesitamos maximizar es el producto de cada fdp individual de *todas* las observaciones $i = 1, \ldots, N$

$$f(y|X; \beta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(y-X\beta)^2}{2\sigma^2}. \tag{4.48}$$

La función 4.48 está entendida como la densidad del argumento y, con X, β y σ^2 dado. Como revisamos con más detalle en la sección 3.3.4 de estos apuntes, para formar la función de verosimilitud, cambiamos el enfoque para ver todo como una función de los parámetros β y σ^2, pero con y y X tomados como dados:

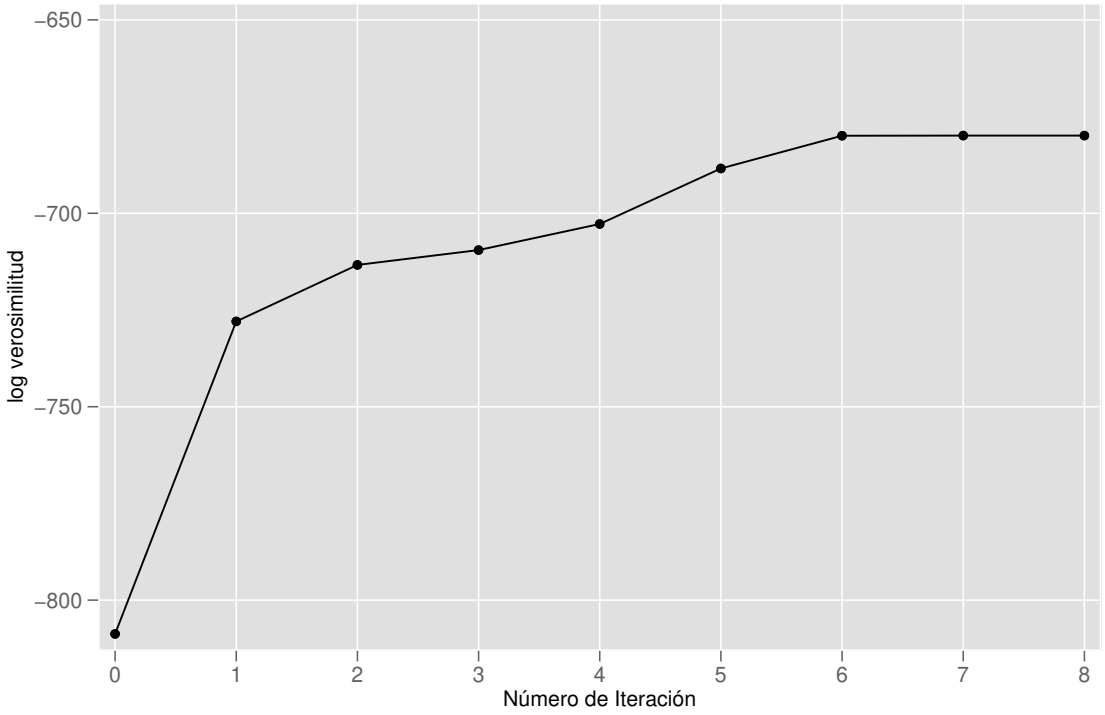
$$\mathcal{L}(\beta, \sigma^2|y, X) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp -\frac{(y-X\beta)^2}{2\sigma^2}. \tag{4.49}$$

Esta ecuación 4.49 es la función de verosimilitud para el modelo lineal, y hace explícito que ahora depende de *todo* el vector de parámetros β, y además la varianza σ^2. Notemos también, que dado que $u = y - X\beta$, a veces se escribe la función de forma más resumida como:

$$\mathcal{L}(\beta, \sigma^2) = \prod_{i=1}^{N} \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{u'u}{2\sigma^2}\right).$$

Por último, tomando logaritmos para la conveniencia computacional de la maximización, podemos escribir la función de verosimilitud $\mathcal{L}(\beta, \sigma^2)$ como una función de log verosimilitud $\ell = ln\mathcal{L}(\beta, \sigma^2)$, que tendrá su máximo en el mismo punto para (β, σ^2). Esto da la función de log

Figure 4.5: El Proceso de Iteración de la función de log verosimilitud



verosimilitud para el modelo lineal:

$$\begin{aligned} .\ell(\beta,\sigma^2) &= ln\mathcal{L}(\beta,\sigma^2) \\ &= ln\prod_{i=1}^{N}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{u'u}{2\sigma^2}\right) \\ &= \sum_{i=1}^{N} ln\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{u'u}{2\sigma^2}\right) \\ \ell &= -\frac{N}{2}ln2\pi - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}u'u \end{aligned} \tag{4.50}$$

Si, por ejemplo, nuestro modelo de regresión es $y = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$, la función de arriba es:

$$\ell(\beta_1,\beta_2,\beta_3,\sigma^2) = -\frac{N}{2}ln2\pi - \frac{N}{2}ln(\sigma^2) - \frac{1}{2\sigma^2}(y-\beta_1-\beta_2x_2-\beta_3x_3)'(y-\beta_1-\beta_2x_2-\beta_3x_3)$$

Maximizando la función de log verosimilitud La función de log verosimilitud depende de los argumentos β y σ^2. Al igual que en la sección 3.3.4, el proceso de estimación consiste en encontrar los valores $\hat{\beta}$ y $\hat{\sigma}^2$ que maximizan esta función, tomando como dado la muestra de y y X. Estos valores $\hat{\beta}_{MV}$ y $\hat{\sigma}_{MV}$ son los estimadores de máxima verosimilitud. En la práctica, los procesos de maximización que se utilizan son procesos iterativos. En la mayoría de las implementaciones computacionales de librerias de optimización, el usuario puede (o debe) especificar un punto de

partida para la estimación para el vector de parámetros (β, σ^2). Es común observar en estos procesos que el proceso de convergencia del proceso de optimización parte bastante lejos del máximo final, y después converge con cambios pequeños hasta llegar al punto máximo encontrado. Este tipo de comportamiento (ver la figura 4.5 para un ejemplo), es inherente a la naturaleza de los algoritmos de optimización, que están descritos en Cameron and Trivedi (2005, capítulo 10).

Maxima Verosimilitud versus MCO Tipicamente, un estimador de máxima verosimilitud no tendrá una solución de forma cerrada, y por lo tanto, utilizamos herramientas númericas para encontrar el máximo, y los valores de los parámetros estimados. Pero en este caso específico, podemos encontrar una solución de forma cerrada para el vector de parámetros β a partir de la función de log verosimilitud. Para encontrar el máximo, derivamos la función ℓ, y lo igualamos a cero:

$$\begin{aligned} \frac{\partial \ell(\beta, \sigma^2)}{\partial \beta} &= 0 \\ \frac{\partial(-\frac{1}{2\sigma^2}(y - X\beta)'(y - X\beta))}{\partial \beta} &= 0 \\ \Rightarrow \frac{1}{\sigma^2}X'(y - X\beta) &= 0 \\ \beta &= (X'X)^{-1}X'y \end{aligned} \tag{4.51}$$

donde la segunda linea se escribe a partir de la ecucación 4.50, dado que β solamente aparece en el tercer término de la función de log verosimilitud.[15]

La solución 4.51 es el máximo de la función ℓ, y por ende, $\hat{\beta}_{MV}$. Es inmediamente claro de 4.51 que $\widehat{\beta}_{MV} = \widehat{\beta}_{MCO}$. Esto también es intuitiva al observar la función de log verosimilitud. Como solamente el término final de la función de verosimilitud depende de β, maximizar $\ell(\beta, \sigma^2)$ con respecto a β es equivalente a minimizar la suma de los errores cuadrados $u'u = \sum_{i=1}^{N} u_i^2$ con respecto a β. Entonces, *en el caso del modelo clásico de regresión lineal con errores Normales*, el estimador $\hat{\beta}_{MV}$ es idéntica al estimador $\hat{\beta}_{MCO}$.

Test de la Razón de Verosimilitudes El test de la razón de verosimilitudes introducido en la sección 3.4.3 de este documento se extiende fácilmente a modelos de regresión. Para implementar este test, es necesario estimar dos modelos mediante máximia verosimilitud: un modelo no restringida que es el modelo propuesto, y un modelo restringido que es el modelo con la hipótesis nula impuesta.

Este test funciona tanto para hipótesis simples como para hipótesis múltiples. Por ejemplo,

[15] Notemos que puede ser más simple hacer este cálculo de maximización si escribimos la función de log verosimilitud como:

$$\ell = -\frac{N}{2}\ln 2\pi\sigma^2 - \frac{\sum_{i=1}^{N}(y_i - \beta_1 x_{1i} - \ldots - \beta_K x_{Ki})^2}{2\sigma^2},$$

y procedemos a maximizar ℓ respecto a cada β_k uno por uno.

si nuestra modelo a estimar tiene un intercepto y tres otras variables independientes, y queremos comprobar la hipótesis:

$$H_0 : \beta_2 = 0 \text{ y } \beta_3 = 0 \qquad H_1 : H_0 \text{ no es cierto,}$$

Esto implica que, bajo la nula, el modelo no incluye las variables x_2 y x_3. Por lo tanto, nuestro modelo no restringido es

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

y llamamos la función de log verosimilitud que corresponde a este modelo a $\ell(\hat{\beta}_{ML})$. Y el modelo restringido (bajo la nula) es:

$$y = \beta_1 + \beta_4 x_4 + u$$

y la función de log verosimilitud correspondiente se llama $\ell(\hat{\beta}_0)$.

Como definimos anteriormente, y como mostró formalmente Wilks (1938), la estadística de prueba para este test es:

$$2\left[\ell(\hat{\mu}_{ML}) - \ell(\mu_0)\right] \overset{a}{\sim} \chi^2_{(2)}.$$

Notemos aquí que ahora el valor del crítico del test depende de una distribución χ^2 con 2 grados de libertad. La cantidad de grados de libertad simplemente viene de la cantidad de restricciones impuestas en la nula. En el modelo anterior, impongamos dos restricciones ($\beta_2 = 0$ y $\beta_3 = 0$), pero el test de razón de verosimilitudes funciona de igual forma con 1 o más de 2 restricciones.

4.5.2 Estimación Por Método de Momentos

De la introducción a métodos de momentos en la sección 3.3.3, sabemos que la estimación por método de momentos requiere plantear *momentos poblacionales* y formar momentos análagos en la muestra. La idea de método de momentos es encontrar *momentos* que cumplen en la población dado el modelo a estimar. Entonces, estimamos reemplazando estos momentos poblacionales con sus contrapartes muestrales, y encontrando la solución a los momentos muestrales.

Del modelo lineal, sabemos:

$$y_i = x_i'\beta + u_i$$

con $E[u_i|x_i] = 0$. El supuesto $E[u_i|x_i] = 0$ implica (por las condiciones de la covarianza):

$$E[x_i u_i] = 0$$

Este supuesto acerca de los valores poblacionales $E[x_i u_i]$ nos da los momentos poblacionales para el modelo de regresión lineal. También lo podemos escribir en forma vectorial:

$$E[xu] = 0$$

o, con $u = y - X\beta$, como:

$$E[x(y - x\beta)] = 0.$$

Notemos que en esta situación, tenemos K momentos—uno para cada variable:

$$\begin{aligned} E(x_1 u) &= 0 \\ E(x_2 u) &= 0 \\ &\vdots \\ E(x_K u) &= 0 \end{aligned}$$

Y tenemos un sistema de ecucaciones "identificada". Tenemos K incógnitas (las $\beta_1, \ldots, \beta_K$ que necesitamos estimar), y K ecuaciones. La estimación por MCO consiste en encontrar los valores para β, que vamos a llamar $\hat{\beta}_{MM}$, que resuelvan las contrapartes muestrales.

$$\underbrace{E[x(y - x\beta)] = 0}_{\text{Momentos Poblacionales}} \qquad \underbrace{\frac{1}{N}\sum_{i=1}^{N} x_i(y_i - x_i\beta) = 0}_{\text{Momentos Muestrales}}$$

Nuestra solución consiste en resolver los momentos muestrales:

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} x_i(y_i - x_i\hat{\beta}_{MM}) &= 0 \\ \frac{1}{N}\sum_{i=1}^{N} x_i y_i - \frac{1}{N}\sum_{i=1}^{N} x_i x_i \hat{\beta}_{MM} &= 0 \\ \hat{\beta}_{MM} &= \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} x_i y_i \end{aligned}$$

$$\hat{\beta}_{MM} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} x_i y_i$$

Siempre cuando $\frac{1}{N}\sum_{i=1}^{N} x_i x_i$ es invertible. Sin embargo, notemos que en forma matricial

$$\begin{aligned} \frac{1}{N}\sum_{i=1}^{N} x_i x_i &= X'X \\ \frac{1}{N}\sum_{i=1}^{N} x_i y_i &= X'y \end{aligned}$$

y por lo tanto:

$$\hat{\beta}_{MM} = \left(\frac{1}{N}\sum_{i=1}^{N} x_i x_i\right)^{-1} \frac{1}{N}\sum_{i=1}^{N} x_i y_i = (X'X)^{-1}X'y = \hat{\beta}_{MCO}.$$

Nuevamente, como vimos con el estimador $\hat{\beta}_{MV}$, el estimador de mínimos cuadrados ordinarios y el estimador de método de momentos para el vector de parámetros β es idéntica (en este caso del modelo lineal de regresión).

El método de momentos consiste en resolver un sistema de j ecuaciones para j incógnitas. Sin embargo, en algunos casos tenemos más momentos válidos que parámetros a estimar (sobre-identificación). En estos casos, *minimizamos* una cantidad que depende del sistema de ecucaciones. Esto se conoce como el Método de Momentos Generalizados (o GMM). Una estimación así permite agregar más información al sistema para resolver, abriendo la posibilidad de otros tipos de contrastes de hipótesis y analisis de los supuestos atrás del modelo. Volverán a este punto en mucho más profundidad en el segundo semestre del magíster.

Clase Computacional: Explorando Máxima Verosimilitud En este clase considerarmos un modelo de la siguiente forma:

$$y = 3 + 6x_1 + 4x_2 + 2x_3 + \varepsilon,$$

donde $\varepsilon \sim \mathcal{N}(0, 3)$. Partiremos en Stata simulando este modelo. Para poder simularlo, supongamos que x_1, x_2 y x_3 se distribuyen según una variable uniforme [0,1], con cada variable independiente entre sí. Simula 1000 observaciones para $(x_1, x_2, x_3, \varepsilon)$, y por último genera y como:

```
gen y = 3 + 6*x1 + 4*x2 + 2*x3 + epsilon
```

Ejercicios:

1. Genera una función (un 'programa') en Stata que es la función de log verosimilitud del modelo lineal. Utilizando esta función y los comandos `ml model` y `ml maximize` encuentra el vector de parámetros que maximiza ℓ.
2. Escriba una función en Mata, y utilizando los mismos datos generados arriba, encuentra el vector de parámetros que maximiza ℓ con el comando `optimize`
3. Calcula el valor de la función de log verosimilitud para los parámetros $\beta = (3, 6, 4, 2)$ y $\sigma^2 = 3$. ¿Por qué no alcanza un máximo absoulto en este punto a pesar del proceso generador de datos?
4. Utilizando un test de razón de verosimilitudes, comprueba la hipótesis nula: $H_0 : x_3 = 0$ contra la alternativa $H_1 : X_3 \neq 0$. Explora este proceso en Mata y en Stata.
5. Gráfica el proceso de maximización de la función de log verosimilitud con un gráfico parecido a la figura 4.5.

4.6 Comportamiento Asintótico

4.6.1 La Teoría Asintótica

Las propiedades de los estimadores MCO derivadas en la sección 4.2.2 cumplen en muestras finitas (y muestras de cualquier tamaño). pero dependen del supuesto de normalidad del término estacástico u. En esta sección, vamos a examinar las propiedades del estimador MCO en **muestras grandes**, o, en términos precisos, cuando el tamaño de la muestra N se acerca a infinito manteniendo constante la cantidad de variables explicativas K. Cuando estimamos modelos con una muestra grande de datos, los resultados asintóticos nos proporcionarán un acercamiento útil al comportamiento de los estimadores y estadísticas de prueba. La utilidad de esto es que podemos establecer resultados útiles con supuestos **menos exigentes** a los que hemos utilizado para establecer las distribuciones exactas en muestras finitas. Específicamente, podemos seguir sin la necesidad de tomar supuestos distribucionales acerca del término no observado u, simplemente utilizando los resultados de un TLC apropiado.

Primero, consideramos una serie de supuestos suficientes para tener un estimador consistente de MCO en el modelo lineal.

(i) $y_i = x_i'\beta + u_i$ para $i = 1, \ldots, N$, o $y = X\beta + u$.

(ii) Los datos en (y_i, x_i) son independientes sobre $i = 1, \ldots, N$, con $E(u_i) = 0$ y $E(x_i u_i) = 0$ para cada $i = 1, \ldots, N$.

(iii) X es estocástico y de rango completo

(iv) La matriz de $K \times K$ $M_{XX} = p\lim_{N\to\infty}\left(\frac{X'X}{N}\right) = p\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} x_i x_i'$ existe y no es singular.

Bajo las 4 condiciones anteriores, $\text{plim}_{\text{N}\to\infty}\hat{\beta}_{MCO} = \beta$, o, $\hat{\beta}_{MCO} \xrightarrow{P} \beta$. Para demostrar esto, notemos que podemos escribir nuestro estimador MCO de la siguiente forma:

$$\begin{aligned}\hat{\beta}_{MCO} &= (X'X)^{-1}X'y = (X'X)^{-1}X'(X\beta + u) \\ &= (X'X)^{-1}X'X\beta + (X'X)^{-1}X'u = \beta + (X'X)^{-1}X'u.\end{aligned}$$

Y multiplicando y dividiendo el segundo término por N (que mantiene constante el estimador) da:

$$\hat{\beta}_{MCO} = \beta + N(X'X)^{-1}\left(\frac{1}{N}\right)X'u. \tag{4.52}$$

Para demostrar que el estimador $\hat{\beta}_{MCO}$ es consistente, necesitamos demostrar que $\text{plim}_{\text{N}\to\infty}\hat{\beta}_{MCO} = \beta$, o de manera equivalente, que el limite de probabilidad del segundo término es 0. Tomando la

probabilidad límite de 4.52 da:

$$\begin{aligned}\text{plim}_{N\to\infty}\hat{\beta}_{MCO} &= \text{plim}_{N\to\infty}\beta + \left(\text{plim}_{N\to\infty}\left(\frac{X'X}{N}\right)\right)^{-1}\left(\text{plim}_{N\to\infty}\left(\frac{X'u}{N}\right)\right)\\ &= \beta + M_{XX}^{-1}\,\text{plim}_{N\to\infty}\left(\frac{X'u}{N}\right),\end{aligned}$$

y la consistencia del estimador requiere:

$$\text{plim}_{N\to\infty}\left(\frac{X'u}{N}\right) = \text{plim}_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N}x_iu_i\right) = 0.$$

Esta última igualdad es cierto, y viene del supuesto $E(x_iu_i) = 0$. Utilizando una ley de grandes números apropiada para las observaciones $i = 1,\ldots,N$, tenemos $\text{plim}_{N\to\infty}\left(\frac{1}{N}\sum_{i=1}^{N}x_iu_i\right) = E(x_iu_i) = 0$, los grandes números de Kolmogorov y para observaciones independientes pero no idénticamente distribuidos, la de Markov.

Notemos que aquí, para establecer que el estimador MCO es consistente, *no* tuvimos que suponer que haya homoscedasticidad condicional ($V(u_i|x_i) = \sigma^2$), ni tampoco normalidad. El supuesto clave para la consistencia es que el término de error u_i no esté correlacionado con cada variable explicativa x_i, como lo señalado por el supuesto $E(x_iu_i) = 0$.

4.6.2 Normalidad Asintótica

La consistencia es una propiedad muy útil para un estimador, especificamente si estamos trabajando con bases de datos grandes. Pero como sabemos, la consistencia no basta para proceder con inferencia. Para obtener intervalos de confianza y para realizar contrastes de hipótesis necesitamos saber algo acerca de la distribución asintótica de los estimadores. Eso es, algo que víncula la distribución del estimador en muestras grandes con una distribución conocida (y ojalá conveniente).

Planteamos aquí una serie de supuestos para tener un estimador MCO resultante cuya distribución límite sea una función de una distribución Normal. Para partir, vamos a suponer que la varianza condicional del término de error ($E(u_i^2|x_i)$) satisface el supuesto de homoscedasticidad condicional. En el siguiente capítulo, estudiaremos cómo se puede relajar este supuesto. El supuesto de homoscedasticidad va a ser necesario para demostrar que el estimador MCO tiene la varianza asintótica *específica* que derivamos aquí, pero *no* es necesario para establecer una distribución asintótica normal. Estos supuestos son:

(i) $y_i = x_i'\beta + u_i$ para $i = 1,\ldots,N$, o $y = X\beta + u$.

(ii) Los datos en (y_i, x_i) son independientes sobre $i = 1,\ldots,N$, con $E(u_i) = 0$ y $E(x_iu_i) = 0$ y $E(u_i^2|x_i) = \sigma^2$ para cada $i = 1,\ldots,N$.

(iii) X es estocástico y de rango completo

(iv) La matriz de $K \times K$ $M_{XX} = p\lim_{N\to\infty}\left(\frac{X'X}{N}\right) = p\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} x_i x_i'$ existe y no es singular

(v) El vector de $K \times 1$ $\left(\frac{X'u}{\sqrt{N}}\right) = \frac{1}{N}\sum_{i=1}^{N} x_i u_i \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{XX})$

Si estos supuestos cumplen, Entonces $\sqrt{N}(\hat{\beta}_{MCO} - \beta) \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{XX}^{-1})$. Notemos que aquí hemos agregado dos supuestos a la serie de supuestos planteado en la sección 4.6.1. El primero es el supuesto de homoscedasiticad condicional (planteado en rojo), y el segundo es un supuesto acerca de la distribución límite de una función de u. El supuesto (v) parece ser un supuesto bastante fuerte, al suponer una distribución para un componente no observado. Pero se puede derivar el supuesto (v): $\left(\frac{X'u}{\sqrt{N}}\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{XX})$ desde supuestos mucho más primitivos. Demostramos esto en la sigueinte sub-sección (de lectura no obligatoria). Por el momento, notemos que basta utilizar el supuesto (ii) y un TLC. Con el supuesto de homoscedasticidad, es el TLC de Lindeberg Lévy para vectores aleatorios iid.

Para demostrar normalidad asintótica, partimos con $\hat{\beta}_{MCO}$,

$$\hat{\beta}_{MCO} = \beta + \left(\frac{X'X}{N}\right)^{-1}\left(\frac{X'u}{N}\right)$$

Y reorganizamos como:

$$\begin{aligned}\hat{\beta}_{MCO} - \beta &= \left(\frac{X'X}{N}\right)^{-1}\left(\frac{X'u}{N}\right)\\ \sqrt{N}(\hat{\beta}_{MCO} - \beta) &= \left(\frac{X'X}{N}\right)^{-1}\left(\frac{X'u}{\sqrt{N}}\right).\end{aligned}$$

Del supuesto (iv) indicado anteriormente, tenemos que la matrix de $K \times K$: $\left(\frac{X'X}{N}\right)^{-1} \xrightarrow{P} M_{XX}^{-1}$. Y del supuesto (v), el vector de $K \times 1$: $\left(\frac{X'u}{\sqrt{N}}\right) \xrightarrow{D} N(0, \sigma^2 M_{XX})$. Se puede demostrar (por el ley de producto) que $\left(\frac{X'X}{N}\right)^{-1}\left(\frac{X'u}{\sqrt{N}}\right)$ tiene la misma distribución límite que $M_{XX}^{-1}\left(\frac{X'u}{\sqrt{N}}\right)$. De lo anterior[16], tenemos

$$\left(\frac{X'X}{N}\right)^{-1}\left(\frac{X'u}{\sqrt{N}}\right) \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{XX}^{-1} M_{XX} M_{XX}^{-1})$$

Y por lo tanto:

$$\sqrt{N}(\hat{\beta}_{MCO} - \beta) \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{XX}^{-1}). \tag{4.53}$$

En 4.53, hemos derivado un resultado de **Normalidad Asintótica**. Pero en estricto rigor, esta distribución es para una función del estimador $\hat{\beta}_{MCO}$, mientras nos interesa solo en el estimador

[16] El ley de producto normal límite dice que si un vector $a_N \xrightarrow{D} \mathcal{N}(\mu, A)$ y una matríz $H_N \xrightarrow{P} H$ donde H es positiva definida, entonces $H_N a_N \xrightarrow{D} \mathcal{N}(H\mu, HAH')$.

$\hat{\beta}_{MCO}$. Por suerte esto no es problemático: podemos reorganizar la ecuación. Simplemente multiplicamos ambos lados por el escalar $\frac{1}{\sqrt{N}}$, y sumamos el vector no estocástico β a ambos lados:

$$\hat{\beta}_{MCO} \overset{a}{\sim} \mathcal{N}\left(\beta, \left(\frac{1}{N}\right)\sigma^2 M_{XX}^{-1}\right)$$

Ahora, tenemos la varianza aproximada del estimador MCO como la matriz $\left(\frac{1}{N}\right)\sigma^2 M_{XX}^{-1}$. Esta distribución asintótica se denota como $avar(\hat{\beta}_{MCO})$. Y con un estimador para σ^2 y M_{XX}, ya podemos seguir con inferencia. La inferencia asintótica consiste en formar los mismos estadísticos de prueba que hemos visto hasta ahora (t, F).

Los estimadores más obvios para M_{XX} y σ^2 son, respectivamentes, $\widehat{M}_{XX} = \left(\frac{X'X}{N}\right)$ ya que:

$$\text{plim}_{\text{N}\to\infty}\, \widehat{M}_{XX} = \text{plim}_{\text{N}\to\infty}\left(\frac{X'X}{N}\right) = M_{XX},$$

y $\hat{\sigma}^2 = \frac{1}{N-K}\sum_{i=1}^{N}\hat{u}_i^2 = \frac{\hat{u}'\hat{u}}{N-K}$. Los dos son estimadores consistentes. Por último, podemos escribir el estimador de la varianza asintótica como:

$$\begin{aligned}
\widehat{avar}(\hat{\beta}_{MCO}) &= \left(\frac{1}{N}\right)\widehat{\sigma}^2\widehat{M}_{XX}^{-1} \\
&= \left(\frac{1}{N}\right)\widehat{\sigma}^2\left(\frac{X'X}{N}\right)^{-1} \\
&= \left(\frac{1}{N}\right)\widehat{\sigma}^2 N(X'X)^{-1} \\
&= \widehat{\sigma}^2(X'X)^{-1}
\end{aligned}$$

y, dado que $\widehat{avar}(\hat{\beta}_{MCO}) \xrightarrow{P} avar(\hat{\beta}_{MCO})$, tenemos:

$$\hat{\beta}_{MCO} \overset{a}{\sim} \mathcal{N}\left(\beta, \widehat{avar}(\hat{\beta}_{MCO})\right)$$

$$\text{o: } \hat{\beta}_{MCO} \overset{a}{\sim} \mathcal{N}\left(\beta, \widehat{\sigma}^2(X'X)^{-1}\right).$$

4.6.3 ⋆Derivando la Distribución Límite con un TLC

El modelo lineal de regresión clásico se basa en los siguientes supuestos (son los supuestos asintóticos, no los supuestos de muestra finita).

(i) $y_i = x_i'\beta + u_i$ para $i = 1, \ldots, N$, o $y = X\beta + u$.

(ii) Los datos en (y_i, x_i) son independientes sobre $i = 1, \ldots, N$, con $E(u_i) = 0$ y $E(x_i u_i) = 0$ y $E(u_i^2|x_i) = \sigma^2$ para cada $i = 1, \ldots, N$.

(iii) X es estocástico y de rango completo

(iv) La matriz de $K \times K$ $M_{XX} = p\lim_{N\to\infty}\left(\frac{X'X}{N}\right) = p\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} x_i x_i'$ existe y no es singular

(v) El vector de $K \times 1$ $\left(\frac{X'u}{\sqrt{N}}\right) = \frac{1}{N}\sum_{i=1}^{N} x_i u_i \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{XX})$

En base a estos 5 supuestos, hemos demostrado que:

$$\sqrt{N}(\hat{\beta}_{MCO} - \beta) \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{XX}^{-1}),$$

y ésta es una fórmula útil para derivar una distribución límite conocida para el estimador $\hat{\beta}_{MCO}$:

$$\hat{\beta}_{MCO} \overset{a}{\sim} \mathcal{N}\left(\beta, \left(\frac{1}{N}\right)\sigma^2 M_{XX}^{-1}\right). \tag{4.54}$$

El supuesto (v) fue central en esta derivación, ya que, utilizando el Teorema de Slutsky, nos da el resultado de normalidad asintótica para $\hat{\beta}_{MCO}$. Sin embargo, el supuesto (v) se puede derivar desde supuestos más primitivos, sin tener que suponer normalidad de $\left(\frac{X'u}{\sqrt{N}}\right)$. Sólo se necesita el supuesto (ii), y un Teorema Central de Límite apropiado, que en el caso de homoscedasticidad sería el de Lindeberg-Lévy. Aquí demostraremos esto.

Si $\{z_i : i = 1, 2, \ldots\}$ es una sequencia iid de $K \times 1$ vectores aleatorias con $E(z_i) = 0$, y $V(z_i) = E(z_i z_i') = \Sigma$ finita, entonces[17]:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} z_i \xrightarrow{D} \mathcal{N}(0, \Sigma).$$

Definiendo $z_i = x_i u_i$, tenemos $K \times 1$ vectores aleatorias iid con $E(x_i u_i) = 0$. Suponiendo una varianza finita obtenemos:

$$\frac{1}{\sqrt{N}}\sum_{i=1}^{N} x_i u_i = \left(\frac{X'u}{\sqrt{N}}\right) \xrightarrow{D} \mathcal{N}(0, V(x_i u_i))$$

Además,

$$V(x_i u_i) = E[(x_i u_i)(x_i u_i)'] = E(x_i u_i u_i' x_i') = E(u_i^2 x_i x_i')$$

ya que u_i es un escalar. Pero,

$$\begin{aligned} E(u_i^2 x_i x_i') &= E_x[E_{u|x}(u_i^2 x_i x_i'|x_i)] = E_x[x_i x_i' E_{u|x}(u_i^2|x_i)] \\ &= E_x[x_i x_i' \sigma^2] = \sigma^2 E_x(x_i x_i') = \sigma^2 M_{XX} \end{aligned}$$

utilizando el supuesto de homoscedasticidad que dice que $E(u_i^2|x_i) = \sigma^2$.

Entonces, $V(x_i u_i) = \sigma^2 M_{XX}$ y $\left(\frac{X'u}{\sqrt{N}}\right) = \frac{1}{N}\sum_{i=1}^{N} x_i u_i \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{XX})$ como se dijo en el supuesto (v).

[17] Por el teorema del límite central multivariada. Para detalles, ver, por ejemplo, Rao (1973, p. 128).

El supuesto de homoscedasticidad es solo necesario para obtener esta expresión en particular para la distribución límite de $\left(\frac{X'u}{\sqrt{N}}\right)$, y por lo tanto, la distribución límite de $\hat{\beta}_{MCO}$ en (4.54). Con independencia y heteroscedasticidad (es decir $E(u_i^2|x_i) \neq \sigma^2$ para cada $i = 1, \ldots, N$), podemos utilizar el Teorema del Límite Central de Liapounov para vectores aleatorios independientes, pero no idénticamente distribuidos. Con esto, se obtiene que la distribución límite de $\left(\frac{X'u}{\sqrt{N}}\right)$ es Normal, pero con una expresión diferente para la varianza.

Ejercicios de Ayudantía:

1. Considere el siguiente modelo del rendimiento escolar de los estudiantes chilenos durante el año 2015:

$$prom\_gral_i = \beta_1 + \beta_2 asistencia_i + \beta_3 mujer_i + \beta_4 subvenc_i + \beta_5 pagado_i + \beta_6 corp_i + u_i$$

Donde *prom_gral* corresponde al promedio general del alumno, *asistencia* mide su porcentaje de asistencia, *mujer* toma el valor 1 si el estudiante es mujer y 0 si es hombre y *subvenc*, *pagado* y *corp*, toman el valor 1 si la dependencia es subvencionada, pagada o corporación, respectivamente, y 0 en caso contrario. Utilizando la base de datos $rendimiento2015\_3.dta$ resuelva los siguientes ejercicios:

   (a) Genere una matriz y que contenga la variable dependiente y una matriz x que contenga las variables independientes.

   (b) Calcule el estimador MCO de manera matricial. Presente los resultados en forma de ecuación y exporte los resultados.

   (c) Interprete el impacto de la asistencia. ¿Es estadísticamente significativo al 5%? Decida con p-valor.

   (d) Interprete el parámetro $\widehat{\beta_3}$. Contraste si a las mujeres les va mejor que a los hombres con un nivel de significancia del 10%. Decida con valor crítico.

   (e) De acuerdo a los resultados de estimación ¿Les va mejor a los estudiantes de los establecimientos municipales o pagados? ¿Cuánto mejor/peor?

   (f) ¿Existen diferencias en el rendimiento derivadas de la dependencia de los establecimientos? Contraste con un nivel de significancia del 1%. Utilice el test-F y el test con matrices. Presente los resultados con p-valor y valor crítico.

2. Formas Funcionales:
   Considere las siguientes estimaciones:

$$\widehat{salario} = -2.4195 + 0.5622 educacion + 0.3299 permanencia - 0.0055 permanencia^2 \tag{4.55}$$

$$\widehat{\log(salario)} = 1.5502 + 0.1077 \log(experiencia) - 0.3856 mujer \tag{4.56}$$

Donde la variable *salario* mide el salario por hora en dólares, *educacion* corresponde a los años de educación, *permanencia* corresponde a los años que se ha permanecido en el

trabajo actual, *experiencia* mide la experiencia laboral en años y *mujer* toma el valos 1 cuando el individuo es mujer.

(a) Interprete los resultados de la ecuación (1)

(b) Encuentre el nivel de permanencia en el trabajo actual que maximiza el salario.

(c) Interprete el impacto de la experiencia laboral y género en la ecuación (2)

(d) Estime el salario de una mujer con 20 años de experiencia.

3. Considere el siguiente modelo que explica el salario:

$$\begin{aligned}\widehat{\log(salario)} &= \underset{(0.101)}{0.490} + \underset{(0.007)}{0.084}educacion + \underset{(0.002)}{0.003}experiencia + \underset{(0.003)}{0.017}permanencia - \\ &\quad \underset{(0.037)}{0.286}mujer + \underset{(0.040)}{0.126}casado \\ n &= 526, \quad R^2 = 0.4036\end{aligned}$$

$$\begin{aligned}\widehat{\log(salario)} &= \underset{(0.104)}{0.284} + \underset{(0.007)}{0.092}educacion + \underset{(0.002)}{0.004}experiencia + \underset{(0.003)}{0.022}permanencia \\ n &= 526, \quad R^2 = 0.3160\end{aligned}$$

Donde las variables *educacion*, *experiencia*, *permanencia* y *mujer* son las mismas descritas en el ejercicio anterior, y *casado* es una variable que toma el valor 1 si el individuo está casado y 0 en caso contrario.

(a) Interprete el impacto de la educación, ¿Es estadísticamente significativo al 5%?

(b) Contraste la significancia conjunta del género y estado civil con un nivel del significancia del 10% en el modelo (3). Calcule el test utilizando R^2 y la forma matricial vista en clases.

(c) Interprete el impacto del género y contraste si existe discriminación salarial contra la mujer con un nivel de significancia del 1% en el modelo (3).

(d) Contraste la significancia global del modelo (3). Calcule el test utilizando R^2.

Sección 5

El Modelo de Regresión Lineal II – Relajando Algunos Supuestos

5.1 Heteroscedasticidad

Nota de Lectura: Una discusión de heteroscedasticidad y soluciones para hacer inferencia válida están disponibles en basicamente todos los textos de economía. Un ejemplo es la sección 4.2.3 de Wooldridge (2002). El capítulo 11 de Greene (2002) es una fuente muy comprensiva para esta material. Hay una discusión fantástica de ponderadores, encuestas y regresiones en Deaton (1997). Las secciones 1.1 y 2.1 de Deaton son muy recomendables. Para una discusión aplicada de ponderadores en regresión, refiere a Solon, Haider and Wooldridge (2015). Angrist and Pischke (2009, p. 91) también ofrecen algunas reflexiones.

Partimos planteando el modelo clásico de regresión lineal en la sección anterior, aún cuando reconocimos que todos los supuestos no iban a ser razonables en cada contexto. Ahora vamos a considerar cómo relajar algunos de estos supuestos en el modelo de regresión lineal.

Hemos mantenido el supuesto de **homoscedasticidad** en la especificación del modelo, y además durante el proceso de inferencia. Este supuesto requiere que la varianza condicional $V(u_i|x_i) = \sigma^2$ es igual para todas las observaciones $i = 1, \ldots, N$. En muchas situaciones (incluyendo el caso de la Figura 4.4) es un supuesto muy cuestionable. Y con frecuencia, es fácil observar en los datos que la homoscedasticidad probablemente *no* cumple.

En muchas situaciones, vamos a estar preocupados de que la varianza del término de error podría variar entre las observaciones i en una manera que es difícil de predecir y/o modelear. Por ejemplo, la varianza en los cambios al PIB en un país que exporta cobre (o recursos naturales generalmente) probablemente es muy diferente a la varianza del PIB en lugares que exportan servicios, o un país que exporta una variedad de bienes. Y la varianza de los ingresos en un hogar rural tra-

bajando en agricultura es muy diferente a la varianza de un hogar urbano con ingresos proveniente de salarios. Hay muchos posibles otros ejemplos en nuestros modelos econométricos.

¿Cómo proceder con inferencia bajo heteroscedasticidad? Cuando introducimos el modelo clásico de regresión lineal, vimos que la propieded de insesgadez (o la consistencia en sección 4.6) no requiere de homoscedasticidad. Cuando se sospecha que hay heteroscedasticidad en los errores estocásticos, la inferencia tiene que proceder utilizando herramientas asintóticas. En esta sección vamos a demostrar que la propiedad de Normalidad Asintótica (refiere a la sección 4.6.2 para la derivación en el caso clásico) del estimador tampoco requiere homoscedasticidad. La varianza asintótica del estimador MCO tiene una forma distinta en el caso más general de heteroscedasticidad condicional, pero se puede estimar esta varianza de manera consistente. La clave es en encontrar un teorema del límite central apropiado para un modelo con un término de error que *no* es idénticamente distribuido.

5.1.1 Estimadores Robustos a la Heteroscedasticidad

White (1980), introdujó un estimador de varianza asintótica robusto a heteroscedasiticidad. También existen ideas parecidas desde más temprano en la literatura estadística (Huber, 1967; Eicker, 1967), y por lo tanto, los errores estándar robustos a heteroscedasticidad a veces son conocidos como "errores estándar de White", o "errores estándar de Eicker-Huber-White". Como veremos aquí, la solución es muy fácil de implementar en computadores modernos, y por lo tanto, este tipo de error estándar ya es un paquete u opción básica en la gran mayoría de los idiomas computacionales. Dejamos como una actividad al final de la sección 5.2 su estimación de forma "manual" en Stata o Mata. Es importante notar que como el estimador de MCO es consistente sin tener que suponer homoscedasticidad, al corregir por heteroscedasticidad, los estimadores puntuales (los $\hat{\beta}_{MCO}$) nunca deben cambiar. Un estimador robusto a heteroscedasticidad se refleja sólo en un cambio en la varianza estimada.

Planteamos los supuestos del modelo básico potencilamente *con* heteroscedasticidad de la siguiente forma:

(i) $y_i = x_i'\beta + u_i$ para $i = 1, \ldots, N$, o $y = X\beta + u$.

(ii) Los datos en (y_i, x_i) son independientes sobre $i = 1, \ldots, N$, con $E(u_i) = 0$ y $E(x_i u_i) = 0$ y $E(u_i^2 | x_i) = \sigma_i^2$ para cada $i = 1, \ldots, N$.

(iii) X es estocástico y de rango completo

(iv) La matriz de $K \times K$ $M_{XX} = p\lim_{N\to\infty}\left(\frac{X'X}{N}\right) = p\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N} x_i x_i'$ existe y no es singular

(v) El vector de $K \times 1$ $\left(\frac{X'u}{\sqrt{N}}\right) = \frac{1}{N}\sum_{i=1}^{N} x_i u_i \xrightarrow{D} \mathcal{N}(0, \sigma^2 M_{X\Omega X})$ donde $M_{X\Omega X} = \text{plim}_{\text{N}\to\infty}\left(\frac{X'uu'X}{N}\right) = \text{plim}_{\text{N}\to\infty}\sum_{i=1}^{N} u_i^2 x_i x_i'$.

Notamos que estos supuestos son idénticos a los supuestos indicados en la sección 4.6.2, pero ahora permitimos heteroscedasticidad. El cambio en los supuestos está indicado en rojo. Si los supuestos (i) a (v) son correctos, entonces:

$$\sqrt{N}(\hat{\beta}_{MCO} - \beta) \xrightarrow{D} \mathcal{N}(0, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}).$$

Este resultado nos permitirá hacer inferencia acerca de los parámetros $\hat{\beta}_{MCO}$ incluso si los errores son heteroscedásticos.

Desafios Del supuesto (ii), que $E(u_i^2|x_i) = \sigma_i^2$ y el hecho de que tenemos observaciones independientes sobre $i = 1, \ldots, N$, definamos la matriz de varianza condicional $E(uu'|X) = \Omega$ como una matriz de $N \times N$ con elementos σ_i^2 en su diagonal, y ceros en las demás posiciones:

$$E(uu'|X) = \Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix}.$$

En la matriz Ω, hay N parámetros (no ceros), $\sigma_1^2, \sigma_2^2, \ldots, \sigma_N^2$. Consecuentemente, no podemos estimar Ω consistentemente con una muestra de N observaciones. El número de parámetros a estimar aumenta paralelamente al tamaño de la muestra. Pero, por suerte no requerimos un estimador consistente para Ω para obtener un estimador consistente de $avar(\hat{\beta}_{MCO}) = \left(\frac{1}{N}\right) M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}$. Volveremos a este procedimiento más adelante en esta sección.

Normalidad Asintótica La demostración de normalidad asintótica sigue los mismos pasos que en el caso que derivamos con homoscedasticidad en la sección 4.6.2. Escribimos:

$$\sqrt{N}(\hat{\beta}_{MCO} - \beta) = \left(\frac{X'X}{N}\right)^{-1} \left(\frac{X'u}{\sqrt{N}}\right).$$

Del supuesto (iv), la matrix de $K \times K$: $\left(\frac{X'X}{N}\right)^{-1} \xrightarrow{P} M_{XX}^{-1}$. Del supuesto (v), el vector de $K \times 1$: $\left(\frac{X'u}{\sqrt{N}}\right) \xrightarrow{D} \mathcal{N}(0, M_{X\Omega X})$. Y por la ley de producto, $\left(\frac{X'X}{N}\right)^{-1} \left(\frac{X'u}{\sqrt{N}}\right)$ tiene la misma distribución límite que $M_{XX}^{-1} \left(\frac{X'u}{\sqrt{N}}\right)$, y entonces:

$$\sqrt{N}(\hat{\beta}_{MCO} - \beta) \xrightarrow{D} \mathcal{N}(0, M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}) \tag{5.1}$$

Aunque este resultado de normalidad asintótica permite hacer inferencia acerca de los parámetros $\hat{\beta}_{MCO}$, el término de varianza es un poco más complejo que la varianza en el caso de homoscedasticidad condicional dada en la ecuación 4.53.

Nuevamente, como vimos en la sección (no obligatoria) 4.6.3, se puede derivar el supuesto (v) utilizando supuestos más primitivos en base a un teorema del límite central. En este caso, se requeriría el teorema del límite central de Liapounov. Es un teorema para vectores aleatorios independientes pero *no* idénticamente distribuidos. Y aplicando este teorema, se puede demostrar que $\frac{1}{N}\sum_{i=1}^{N} x_i u_i \xrightarrow{D} \mathcal{N}(0, M_{X\Omega X})$ donde $M_{X\Omega X} = \text{plim}_{\text{N}\to\infty} \frac{1}{N}\sum_{i=1}^{N} u_i^2 x_i x_i'$. Los pasos son idénticos a los pasos descrito en la sección 4.6.3.

Para ser útil para hacer inferencia respecto a β, el resultado asintótico en 5.1 tiene que ser presentado en términos de $\hat{\beta}_{MCO}$, el parámetro estimado. Actualmente, tenemos la distribución límite para $\sqrt{N}(\hat{\beta}_{MCO} - \beta)$. En la práctica, es simple reorganizar 5.1 para tener una distribución límite para $\hat{\beta}_{MCO}$: solo necesitamos multiplicar ambos lados por el escalar $\frac{1}{\sqrt{N}}$, y sumar el vector no estocástico β a ambos lados. Con esto, tenemos la "distribución asintótica" para $\hat{\beta}_{MCO}$.

La Distribución Asintótica de $\hat{\beta}_{MCO}$

$$\hat{\beta}_{MCO} \overset{a}{\sim} \mathcal{N}\left(\beta, \left(\frac{1}{N}\right) M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}\right)$$

donde,

$$avar(\hat{\beta}_{MCO}) = \left(\frac{1}{N}\right) M_{XX}^{-1} M_{X\Omega X} M_{XX}^{-1}$$

Para ser útil, esta varianza tiene que ser estimable. Para poder estimar la varianza, necesitamos estimadores consistentes de las matrices de $K \times K$ M_{XX} y $M_{X\Omega X}$. Consideramos estas dos matrices una por una.

1. $\boldsymbol{M_{XX}}$: La elección más razonable para estimar $M_{XX} = \text{plim}_{\text{N}\to\infty}\left(\frac{X'X}{N}\right)$ es:

$$\widehat{M}_{XX} = \left(\frac{X'X}{N}\right).$$

   Dado que $\text{plim}_{\text{N}\to\infty} \widehat{M}_{XX} = \text{plim}_{\text{N}\to\infty}\left(\frac{X'X}{N}\right) = M_{XX}$, nos da un estimador consistente de M_{XX}.
2. $\boldsymbol{M_{X\Omega X}}$: De manera parecida, $\widehat{M}_{X\Omega X} = \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i^2 x_i x_i'$ proporciona un estimador consistente para $M_{X\Omega X}$.

$$\text{plim}_{\text{N}\to\infty} \widehat{M}_{X\Omega X} = \text{plim}_{\text{N}\to\infty} \frac{1}{N}\sum_{i=1}^{N} \hat{u}_i^2 x_i x_i' = \text{plim}_{\text{N}\to\infty} \frac{1}{N}\sum_{i=1}^{N} u_i^2 x_i x_i' = M_{X\Omega X}$$

   dado que $y_i - x_i'\hat{\beta}_{MCO} = \hat{u}_i \xrightarrow{P} u_i$

Un punto clave en 2 es que no obtenemos en ningún momento un estimador consistente de la matriz de $N \times N$ de varianza condicional $\Omega = E(uu'|X)$. Sólo necesitamos un estimador consistente de la matriz de $K \times K$ $M_{X\Omega X} = \text{plim}_{\text{N}\to\infty} \frac{1}{N}\sum_{i=1}^{N} u_i^2 x_i x_i'$, y esto lo logramos a partir de los residuos de la

regresión original, $\hat{u}_i$. Aquí, incluso no tenemos que consumir un grado de libertad adicional, dado que los residuos salen de forma "gratuita" una vez que hemos estimados los coeficientes $\hat{\beta}_{MCO}$.

Entonces, ahora sustituyendo los estimadores consistentes de M_{XX} y $M_{X\Omega X}$:

$$\begin{aligned}\widehat{M}_{XX}^{-1}\widehat{M}_{X\Omega X}\widehat{M}_{XX}^{-1} &= \left(\frac{X'X}{N}\right)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)\left(\frac{X'X}{N}\right)^{-1} \\ &= N(X'X)^{-1}\left(\frac{1}{N}\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)(X'X)^{-1}\end{aligned}$$

es un estimador consistente para $M_{XX}^{-1}M_{X\Omega X}M_{XX}^{-1}$, y

$$\begin{aligned}\widehat{avar}(\hat{\beta}_{MCO}) &= \left(\frac{1}{N}\right)\widehat{M}_{XX}^{-1}\widehat{M}_{X\Omega X}\widehat{M}_{XX}^{-1} \\ &= (X'X)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)(X'X)^{-1}\end{aligned}$$

es un estimador consistente de $avar(\hat{\beta}_{MCO}) = \left(\frac{1}{N}\right)M_{XX}^{-1}M_{X\Omega X}M_{XX}^{-1}$.

Ahora, por fin tenemos:

$$\hat{\beta}_{MCO} \overset{a}{\sim} \mathcal{N}\left(\beta, \widehat{avar}(\hat{\beta}_{MCO})\right)$$

donde

$$\widehat{avar}(\hat{\beta}_{MCO}) = (X'X)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)(X'X)^{-1}$$

y podemos calcular $\widehat{avar}(\hat{\beta}_{MCO})$ utilizando los datos sobre X y los residuos MCO, $\hat{u}$. Este estimador de $\widehat{avar}(\hat{\beta}_{MCO})$ es consistente en caso de heteroscedasticidad. Y con esto, podemos proceder exactamente como indicada en la sección 4.4. Podemos realizar test-t asintóticos, y también test-F asintóticos (conocido como test de Wald).

Decisiones Prácticas La inferencia basada en un estimador para la varianza $\widehat{V}(\hat{\beta}_{MCO}) = \hat{\sigma}^2(X'X)^{-1}$ que derivamos anteriormente es válido *sólo* bajo el supuesto restrictivo de homoscedasticidad condicional. Pero dado que $E(u_i^2|x_i) = \sigma^2$ es sólo un caso especial del supuesto general $E(u_i^2|x_i) = \sigma_i^2$, la inferencia asintótica en base al estimador robusto $\widehat{avar}(\hat{\beta}_{MCO}) = (X'X)^{-1}\left(\sum_{i=1}^{N}\hat{u}_i^2 x_i x_i'\right)$ también es válido si el modelo satisface el supuesto de homoscedasticidad. Cuando contamos con una muestra grande de datos, una respuesta común si se sospecha que podría existir heteroscedasticidad es utilizar el estimador MCO, pero con los errores estándar robustos a heteroscedasticidad en lugar de los errores estándar tradicionales. El estimador MCO sigue siendo consistente, y la inferencia es válida en muestras grandes. Esta es una respuesta 'pasiva' a la heteroscedasticidad. El estimador MCO *no* es asintóticamente eficiente en el caso de heteroscedasticidad, pero se pueden realizar contrastes de hipótesis de manera válida. A continuación, veremos algunas maneras de

comprobar si los datos podrían ser heteroscedásticos.

5.1.2 Tests para la presencia de heteroscedastdad

Existen varios tests con el poder de detectar la presencia de heteroscedasticidad condicional en los residuos MCO (o rechazar la nula de homoscedasticidad condicional). Entre ellas:

1. Test de Breusch and Pagan (1979)

2. Test de White (1980).

En ambas contrastes, la idea básica es que especificamos a σ_i^2 como alguna función desconocida $f(z_i)$ de un vector de las variables observadas z_i e incluidas en el modelo de regresión. Si alguna función de σ_i^2 está correlacionada con las variables incluidas en el modelo, es evidencia a favor de la heteroscedasticidad. Utilizamos los residuos al cuadrado $\widehat{u}_i^2$ como una proxy para σ_i^2. Es necesario especificar alguna función para las x_k como una aproximación a la función $f(z_i)$. La diferencia entre los dos tests es en la manera en que especifican esta función $f(z_i)$.

El Test de Breusch-Pagan El Test de Breusch y Pagan está diseñado para detectar cualquier forma lineal de homoscedasticidad. Breusch and Pagan (1979) sugieren predecir los residuos de la regresión, y comprobar si estos residuos al cuadrado están relacionados en forma multiplicativa con cualquiera de las variables explicativas del modelo de regresión. Se hace simplemente estimando una regresión de la forma

$$\widehat{u}_i^2 = \gamma_1 + \gamma_2 x_{2i} + \gamma_3 x_{3i} + \cdots + \gamma_K x_{Ki} + v_i$$

y si se rechaza la nula $\gamma_2 = \gamma_3 = \cdots = \gamma_K = 0$, sugiere que por lo menos una de las variables está relacionada de manera mulitplicativa con la varianza de los errores (que es evidencia a favor de la heteroscedasticidad).

El Test de White El Test de White (1980) sigue una lógica parecida, pero sugiere regresar los residuos al cuadrado de MCO sobre un término constante, todas las variables explicativas, sus cuadrados, y sus productos cruzados. Esto permite capturar formas más complejas de heteroscedasticidad.

Por ejemplo, en un modelo con $K = 3$

$$y_i = \beta_1 + \beta_2 x_{2i} + \beta_3 x_{3i} + u_i$$

el Test de White consiste en la regresión:

$$\widehat{u}_i^2 = \gamma_1 + \gamma_2 x_{2i} + \gamma_3 x_{3i} + \gamma_4 x_{2i}^2 + \gamma_5 x_{3i}^2 + \gamma_6 x_{2i} x_{3i} + v_i$$

y un test de la nula $H_0 : \gamma_2 = \gamma_3 = \cdots = \gamma_6 = 0$. Si hay homoscedasticidad condicional $E(u_i^2|x_i) = \sigma^2$, la nula es cierta. El test de White también puede capturar formas de heteroscedasticidad no lineales que serían difíciles de detectar con el test de Breusch-Pagan.

5.1.3 Un Estimador de Mínimos Cuadrados Generalizados Factibles (FGLS)

Una respuesta activa a la heteroscedasticidad consiste en intentar modelearla directamente. En este caso, podemos formar un estimador consistente y más eficiente que el estimador de MCO con errores estándar robustos a la heteroscedasticidad. El desafío desde luego es poder modelar la forma de heteroscedasticidad. Esta respuesta 'activa' a la heteroscedasticidad puede ser apropiada cuando se considera que la eficiencia del estimador es muy importante. Para ver la idea básica, partimos con un modelo de regresión lineal clásico con heteroscedasticidad cuya forma es conocida

Asumimos:

$$\begin{aligned} E(y|X) &= X\beta \\ V(y|X) &= \Omega, \end{aligned}$$

donde $\Omega \neq \sigma^2 I$ es una matriz definida positiva $N \times N$ **conocida**, y X es estocástico y de rango completo. Dado que Ω es definida positiva y conocida, podemos encontrar una matrix $N \times N$ no estocástica H tal que $H'H = \Omega^{-1}$ y $H\Omega H' = I$. La idea de esta matriz H es que nos permitirá transformar (o re-ponderar) las variables para eliminar la heteroscedasticidad.

Ahora, sea $y^* = Hy$ y $X^* = HX$. Entonces,

$$\begin{aligned} E(y^*|X) &= HE(y|X) = HX\beta = X^*\beta \\ V(y^*|X) &= HV(y|X)H' = H\Omega H' = I (= \sigma^2 I \text{ para } \sigma^2 = 1) \end{aligned}$$

$X^* = HX$ es estocástico y de rango completo. El punto clave aquí es que el modelo transformado:

$$y^* = X^*\beta + u^* \text{ con } V(u^*|X^*) = I$$

es un modelo clásico de regresión lineal con **homoscedasticidad** condicional, que satisface los supuestos del teorema Gauss-Markov. Además, los coeficientes β son idénticos a los coeficientes en el proceso generador de datos.

El Teorema de Aitken El estimador MCO de β en este modelo transformado (conocido como el estimador de Minimos Cuadrados Generalizados) es eficiente en la clase de estimadores lineales insesgados.

$$\begin{aligned}\hat{\beta}_{MCG} &= (X^{*\prime}X^{*})^{-1}X^{*\prime}y^{*} \\ &= (X'H'HX)^{-1}X'H'Hy \\ &= (X'\Omega^{-1}X)^{-1}X'\Omega^{-1}y\end{aligned}$$

Notemos que aquí el estimador de mínimos cuadrados generalizados *es* un estimador ponderado. Si comparamos el estimador β_{MCG} con el estimador de β_{MCP} (ecuación 5.2), los pesos del estimador son simplemente $W = \Omega^{-1}$. Por lo tanto, este estimador de mínimos cuadrados generalizados da menos peso a las observaciones con una variazna más alta, y más peso a las observaciones con una varianza más baja, así logrando ser un estimador más eficiente que MCO con errores estándar robustos.

Mínimos Cuadrados Generalizados Factibles En la práctica, no podemos computar el estimador MCG dado que no conocemos la varianza condicional Ω. Esto es el desafío de estimar MCG. El estimador Mínimos Cuadrados Generalizados Factibles (MCGF) reemplaza el desconocido Ω por un estimador $\widehat{\Omega}$. Esto da:

$$\hat{\beta}_{MCGF} = (X'\widehat{\Omega}^{-1}X')^{-1}X'\widehat{\Omega}^{-1}y,$$

Y ahora sí se puede computar el estimador, utilizando $y^{*} = \widehat{H}y$ y $X^{*} = \widehat{H}X$ donde ahora requeremos que $\widehat{H}'\widehat{H} = \widehat{\Omega}^{-1}$ y $\widehat{H}\widehat{\Omega}\widehat{H}' = I$.

Las propiedades del estimador MCGF dependen de las propiedades de $\widehat{\Omega}$ como estimador de Ω. Si $\widehat{\Omega}$ es un estimador consistente de Ω, entonces bajo condiciones bastante generales encontramos que $\hat{\beta}_{MCGF}$ tiene la misma distribución asintótica que el infactible $\hat{\beta}_{MCG}$, dando:

$$\hat{\beta}_{MCGF} \overset{a}{\sim} \mathcal{N}\left(\beta, (X'\hat{\Omega}^{-1}X)^{-1}\right).$$

En este caso, $\hat{\beta}_{MCGF}$ es asintóticamente eficiente. Sin embargo, la parte difícil es cómo encontrar un estimador consistente para $\widehat{\Omega}$…

La matrix Ω que es simétrica y $N \times N$ tiene $N(N+1)/2$ elementos distintos. Incluso si restringimos que todos los elementos no-diagonales sean ceros (que es natural cuando las observaciones son independentes) para tener:

$$\Omega = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_N^2 \end{pmatrix},$$

aun así tenemos N distintos elementos. No podemos estimar los parámetros consistentemente utilizando nuestra muestra de tamaño N. La estimación consistente requiere que especifiquemos un modelo para Ω, de la forma $\Omega = \Omega(\phi)$, donde $\Omega(\phi)$ es una función del vector ϕ que contiene un número finito de parámetros adicionales que no crece con el tamaño de la muestra que se puede estimar utilizando los datos. Si la especificación de heteroscedasticidad condicional $V(y|X) = \Omega(\phi)$ es correcto y podemos encontrar un estimador consistente $\hat{\phi}$ para ϕ, podemos utilizar el estimador consistente $\widehat{\Omega} = \Omega(\hat{\phi})$ para obtener el estimador MCGF.

Como un ejemplo muy simple, podríamos especificar la varianza condicional $V(y_i|X) = \sigma_i^2$ como proporcional a los valores cuadrados de uno de los regresores, eg x_{Ki}, dando $\sigma_i^2 = \sigma^2 x_{Ki}^2$. Un ejemplo de este estilo sería si pensamos que la heteroscedasticidad en un modelo de salario laboral solamente dependía de forma positiva y cuadrático de los años de educación de cada persona. Y en este caso, tenemos que $y_i^* = \frac{y_i}{x_{Ki}}$ y $x_i^* = \frac{x_i}{x_{Ki}}$ $\forall k = 1, \ldots, K$. En el modelo transformado $y_i^* = x_i^{*\prime}\beta + u_i^*$, tenemos $V(y_i^*|X) = \frac{\sigma_i^2}{x_{Ki}^2} = \frac{\sigma^2 x_{Ki}^2}{x_{Ki}^2} = \sigma^2$ para cada $i = 1, \ldots, N$. Y ahora, el modelo transformado cumple con homoscedasticidad, y lo estimamos utilizando MCO. Notemos que, nuevamente, este es además un estimador de MCP, y la transformación da más peso a observaciones cuya varianza es menor.

5.1.4 Ponderadores

Mínimos cuadrados generalizados (factibles) es un tipo de **Mínimos Cuadrados Ponderados**. En mínimos caudrados ponderados, en vez de minimizar la suma de los cuadrados de los residuos, se minimiza la suma de los cuadrados de los residuos dando un peso, o énfasis, específica a cada observación. Al realizar una regresión, nuestra interés en econometría es hacer inferencia acerca de una población de interés. A veces, por el diseño o el levantamiento de los datos, la base de datos no es una muestra aleatoria de la población, sino sobre-representa algunas poblaciones de interés. Una manera común de diseñar encuestas es utilzando un diseño estratificado. En este caso, primero se define distintos estratos de la población (por ejemplo cada comuna podría ser un estrato), y después dentro de cada estrato, se toma una muestra aleatoria. Pero como algunas comunas son más pequeñas que otras, si se toma una muestra de igual tamaño en cada comuna, esto efectivamante sobrerepresentará estas comunas al momento de realizar el análisis si no se corrigiera por el diseño de la encuesta.

Un ejemplo de este tipo de encuesta es la encuesta CASEN. En la CASEN cada hogar y cada observación vienen con su propio peso, y si queremos hacer inferencia acerca de la población chilena, es necesario ponderar utilizando estos pesos al momento de realizar una regresión. En este caso, el estimador de interés es el estimador de mínimos cuadrados ponderados:

$$\hat{\beta}_{MCP} = (X'WX)^{-1}X'Wy \tag{5.2}$$

donde W es una matriz que define los pesos de cada observación en el diagonal principal. Por

ejemplo, si cada observación tuvo una probabilidad ex-ante de ω_i de haber sido incluido en la encuesta y queremos ponderar para este probabilidad de inclusión, definimos a nuestra matriz $W = diag(1/\omega_1, 1/\omega_2, \ldots, 1/\omega_N)$. Así, damos más peso a las observaciones cuya probabilidad de inclusión es más baja, haciendo que la muestra re-ponderada en la regresión refleje la composición de la población de interés. En la práctica, es simple implementar este estimador en la mayoría de los idiomas estadísticos. Por ejemplo, en Stata se indica pesos en una regresión utilizando: `reg y x [pweight=pesos]`, donde `pesos` es la variable que define los pesos de inclusión en la muestra.

5.2 Clusterización

Nota de Lectura: Probablemente la mejor fuente de información acerca de errores estándares clusterizados (con un enfoque aplicado) es Cameron and Miller (2015). Una introducción en un libro de texto está disponbiel en Cameron and Trivedi (2005).

Un supuesto que todavía no hemos relajado es el supuesto de independencia entre observaciones. Aún cuando permitimos que la varianza del término de error sea distinta para cada observación, no hemos permitido ninguna forma de dependencia *entre* observaciones. Es decir $\rho_{u_i,u_j} = 0 \forall i \neq j$.. Este supuesto se puede relajar en varias maneras límitadas. Un ejemplo es cuando trabajamos con datos de sección cruzada y las observaciones están agrupadas de manera lógica. Podemos permitir que haya dependencia dentro de un grupo, pero no entre grupos.

En este caso, consideramos a distintos grupos de observaciones como "*clusters*". Por ejemplo, podríamos pensar que hay dependencia entre los componentes no observados de: todos los estudiantes en un colegio o clase; todos los miembros de un hogar/aldea/región; toda/os la/os trabajadora/es de una empresa; etc. La idea tras de estos modelos es que estos grupos podrían compartir un evento no observado común, por ejemplo todos los estudiantes en una sala de clases están expuestos a las mismas factores idiosincráticos de su profesor/a.

Supongamos que cada una de las $i = 1, 2, \ldots, N$ individuos pertenece a uno de $c = 1, 2, \ldots, C$ *clusters*. Podemos re-escribir nuestro modelo lineal como:

$$y_{ic} = x'_{ic}\beta + u_{ic} \qquad \text{para } i = 1, \ldots, N \quad \text{y} \quad c = 1, \ldots, C,$$

donde cada una de las $i = 1 \ldots, N$ observaciones está asociado con uno (y sólo uno) de los $c = 1, \ldots, C$ grupos (o *clusters*). Por ende, y_{ic} es la variable dependiente para observación i en el cluster c, y lo mismo para x_{ic} y u_{ic}. Si denotamos como N_c la cantidad de observaciones en cluster c, tenemos que $N_1 + N_2 + \ldots + N_C = \sum_{c=1}^{C} N_c = N$. Notemos que por definición, $C \leq N$.

Agrupamos a las observaciones para tener las primeras n_1 observaciones las en cluster 1, las próximas n_2 observaciones las de cluster 2, y así sucesivamente. Entonces escribimos el modelo

como:

$$y_c = X_c\beta + u_c \qquad \text{para} \qquad c = 1, 2, \ldots, C$$

donde y_c es el vector $n_c \times 1$ que contiene y_{ic} para las n_c observaciones de cluster c. Definimos u_c ($n_c \times 1$) y X_c ($n_c \times K$) del mismo modo. Cada fila de X_c contiene el vector de fila x'_{ic} para una observación en el cluster c. Ahora, definamos los vectores de $n \times 1$ y y u, y la matríz de $n \times K$: X como:

$$y = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_C \end{pmatrix} X = \begin{pmatrix} X_1 \\ X_2 \\ \vdots \\ X_C \end{pmatrix} u = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_C \end{pmatrix}$$

y podemos escribir el modelo como:

$$y = X\beta + u.$$

Con esto, el estimador MCO de β es:

$$\widehat{\beta}_{MCO} = (X'X)^{-1}X'y = \left(\sum_{c=1}^{C} X'_c X_c\right)^{-1} \left(\sum_{c=1}^{C} X'_c y_c\right).$$

En este modelo, el supuesto de independencia es reemplazado por el supuesto de que las observaciones (y_{ic}, x'_{ic}) ahora son independientes entre clusters, pero no dentro de los clusters. Esto permite que los términos de error u_{ic} y u_{jc} estén correlacionados para observaciones distintas en el mismo cluster c y permite patrones de heteroscedasticidad y correlación dentro del cluster. Se derivan los resultados asintóticos bajo la lógica de que la cantidad de clusters $C \to \infty$. Y en el nuevo modelo el supuesto clave que mantenemos es:

$$E(u_{ic}|x_{ic}) = 0$$

$$E(u_{ic}u_{jd}|x_{ic}, x_{jd}) = 0 \qquad \text{para } i \neq j, \qquad \text{salvo que } c = d.$$

Estos supuestos implican:

$$E(uu'|X) = \Omega = \begin{pmatrix} \Omega_1 & 0 & \cdots & 0 \\ 0 & \Omega_2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \Omega_C \end{pmatrix}$$

donde cada Ω_c es simétrica de $n_c \times n_c$ sin ninguna otra restricción! Por ejemplo, consideramos un modelo (muy simple) con 6 observaciones en 3 clusters, con $N_1 = 3, N_2 = 2, yN_3 = 1$. En este

caso, la matriz Ω se escribiría de la siguiente forma general:

$$E(uu'|X) = \Omega = \begin{pmatrix} \Omega_1 & 0 & 0 \\ 0 & \Omega_2 & 0 \\ 0 & 0 & \Omega_3 \end{pmatrix} = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \sigma_{13} & 0 & 0 & 0 \\ \sigma_{12} & \sigma_2^2 & \sigma_{23} & 0 & 0 & 0 \\ \sigma_{13} & \sigma_{23} & \sigma_3^2 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sigma_4^2 & \sigma_{45} & 0 \\ 0 & 0 & 0 & \sigma_{45} & \sigma_5^2 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sigma_6^2 \end{pmatrix}$$

Con una cantidad grande de clusters tenemos el resultado de distribución límite:

$$\sqrt{C}(\hat{\beta}_{MCO} - \beta) \xrightarrow{D} \mathcal{N}(0, V)$$

que implica que el estimador MCO es consistente y asintóticamente normal. Un estimador consistente de la varianza V es:

$$\widehat{V} = C(X'X)^{-1} \left(\sum_{c=1}^{C} X_c' \widehat{u}_c \widehat{u}_c' X_c \right) (X'X)^{-1}$$

donde $\widehat{u}_c = y_c - X_c \widehat{\beta}_{MCO}$ es el vector de residuos de MCO para las n_c observaciones en el cluster c. Y considerando la aproximación asintótica al estimador MCO, tenemos que:

$$\widehat{\beta}_{MCO} \overset{a}{\sim} \mathcal{N}\left(\beta, \frac{\widehat{V}}{C} \right)$$

que sugiere el siguiente estimador robusto a clusterización para la matriz de varianza de $\widehat{\beta}_{MCO}$:

$$\widehat{V}/C = (X'X)^{-1} \left(\sum_{c=1}^{C} X_c' \widehat{u}_c \widehat{u}_c' X_c \right) (X'X)^{-1}$$

Esta varianza permite construir errores estándar y estdaísticos de prueba robustos a clusterización. Y es "fácil" calcular, y ya viene programado en muchos idiomas computacionales (eg `vce(cluster nombre)` en Stata). La matriz de varianza robusta a clusterización se simplifica al estimador de White (1980) (robusto a heteroscedasticidad) en el caso especial de cada observación es un cluster. Las derivaciones hasta aquí asumen $C \to \infty$. Hay varias consideraciones para muestras pequeñas (de clusters). Es un tema que volveremos a ver el año siguiente en el ramo de microeconometría aplicada. Este tema es muy relevante para muchas aplicaciones empíricas (datos de panel, diferencias en diferencias, …), y hay mucho trabajo relevante de aplicación y extensión bastante nuevo. El mejor lugar para partir es Cameron and Miller (2015): "*A Practitioner's Guide to Cluster-Robust Inference*".

Clase Computacional: Programando errores estándar en Stata o Mata
En esta clase exploramos en más detalle el proceso de estimación de los errores estándar bajo distintos supuestos. Pueden utilizar Mata o Stata para esta actividad (o alguna otra idioma matricial si prefieres), y cualquier base de datos (o datos simulados) para estimar las regresiones.

1. Escribe una función para computar errores estándar robustos a heteroscedasticidad en Mata o Stata. *Pista: si utilizas Stata para hacer este cálculo, la función matricial* `opaccum` *puede ser muy útil.*
2. Escribe una función para computar errores estándar clusterizados en Mata o Stata. *Pista: si utilizas Stata para hacer este cálculo, la función matricial* `opaccum` *puede ser muy útil.*
3. En ambos casos, utiliza estas funcionas con datos reales desde Stata y compara los errores estándar con las versiones programadas en Stata. *Pista: Las versiones de los errores estándar calculado en Stata vienen con una pequeña corrección por grados de libertad de* $N/(N-K)$ *en el caso de errores robustos, y* $\frac{N-1}{N-k}\frac{C}{C-1}$ *en el caso de errores clusterizados. Ver aquí:* https://www.stata.com/manuals14/rregress.pdf
4. ⋆ Crea tu propio comando de Stata que permite hacer regresiones con errores estándar robustos y clusterizados sin tener que utilizar las opciones correspondientes (`robust` y `cluster()`) de Stata.

5.3 Endogeneidad

Nota de Lectura: Para el enfoque de error de medición, pueden refereir a la sección 4.4 de Wooldridge (2002). Para el análisis de sesgo de variables omitidas, Cameron and Trivedi (2005) sección 4.7.4 presenta una derivación concisa. Greene (2002) tiene una presentación un poco distinta en su sección 8.1.

Un supuesto clave necesario para establecer la consistencia del estimador MCO es la de exogeneidad. Si el supuesto de que $Cov(x_i, u_i) = 0$ no cumple, esto trae consigo cosecuencias graves para el estimador MCO. Cuando una o más de las variables explicativas está correlacionadas con el término de error u_i, tendremos que $E(u_i|x_i) \neq 0$ y $E(u_i x_i) \neq 0$. En este caso, el estimador MCO será sesgado e inconsistente.

En la práctica, el problema es que el término de error contiene variables que deberían estar en el modelo, y los coeficientes sobre las variables que sí están en el modelo capturan parcialmente el efecto de las variables omitidas. Es un problema que con una regresión MCO siempre está potencialmente presente. Con estos modelos *la única manera* de estimar coeficientes de forma insesgada es especificar bien el modelo y medir las variables correctamente. Al final de esta clase (y en todos los ramos de econometría que quedan) veremos opciones de cómo seguir en la presencia

de endogeneidad. En esta sección, introducimos las consecuencias que la endogeneidad puede tener sobre los coeficientes estimados.

Consideramos específicamente dos situaciones en que la **endogeneidad** o **simultaneadad** ocurre:

1. Un modelo lineal en que $Cov(x_i, u_i) = 0$ es la especificación correcta, pero con una (o más) de las variables explicativas medidas con error

2. Un modelo lineal en que $Cov(x_i, u_i) = 0$ es la especificación correcta, pero una (o más) de las variables explicativas no es observada, y por ende omitida del modelo

5.3.1 Error de Medición/Errores en Variables

Una consideración importante en la econometría aplicada es que las variables explicativas relevantes podrían estar medidas incorrectamente. Algunos ejemplos en datos de encuestas que pueden causar problemas de medición son los "sesgos de recuerdo" (*recall bias*). Por ejemplo, muchas veces hay preguntas relevantes que se realizan posterior al momento del hecho, y algunas veces las personas que responden simplemente no recuerdan correctamente. También es común sospechar que podría haber un "sesgo de redondeo". Observando datos de encuesta, es común ver masas de probabilidad grandes en valores que terminan en 0 o 5, y menos masa en varlores que no son redondeados.

Bajo ciertos supuestos, el error de medición en una variable explicativa resulta en un **sesgo de atenuación**. El sesgo de atenuación es un sesgo siempre hacia cero. Si el efecto verdadero es positivo, un parámetro estimado con sesgo de atenuación es menos positivo. Y si el efecto verdadero es negativo, un parámetro estimado con sesgo de atenuación es menos negativo. El sesgo no desaparece incluso en muestras grandes (MCO es inconsistente). Un error de medición en la variable dependiente *no* produce el mismo sesgo.

Consideremos el modelo con una sola variable explicativa y sin un término de intercepto

$$y_i^* = x_i^* \beta + u_i$$

donde y_i^* y x_i^* denotan a los valores verdaderos de estas variables, que no necesarimente observamos. Para simplificar, suponemos que $E(u_i) = E(x_i^*) = E(y_i^*) = 0$ (simplemente una estandarización de las variables como desviaciones de sus medias muestrales). Nos enfocaremos en las propiedades en muestras grandes y suponemos que $E(x_i^* u_i) = 0$ con observaciones independientes. Bajo estas condiciones, $\hat{\beta}_{MCO}$ sería un estimador consistente para β si observamos los valores verdaderos de y_i^* y x_i^*.

1. Error de Medición Aditiva en la Variable Dependiente

Primero, consideramos error de medición aditiva con media cero en la variable dependiente solamente:

$$y_i = y_i^* + v_i \leftrightarrow y_i^* = y_i - v_i$$

donde y_i es el valor observado, y_i^* es el valor verdadero, v_i es el error de medición, con $E(v_i) = 0$, y los valores verdaderos x_i^* son observados. Sustituyemos la expresión para y_i^* en el modelo verdadero:

$$\begin{aligned} (y_i - v_i) &= x_i^*\beta + u_i \\ \text{o} \quad y_i &= x_i^*\beta + (u_i + v_i) \end{aligned}$$

La consistencia requiere que x_i^* no esté correlacionado con el término de error $(u_i + v_i)$. Dado que $E(x_i^* u_i) = 0$, el requisito adicional es que $E(x_i^* v_i) = 0$ para $i = 1, \ldots, N$. Es decir, el error de medición en la variable dependiente no puede estar correlacionado con la variable explicativa. Si esto es cierto, el error de medición en la variable *dependiente* no provoca sesgo en el parámetro estimado $\hat{\beta}_{MCO}$.

2. Error de Medición Aditiva en la Variable Explicativa Ahora consideramos un error aditivo con medio cero en la variable explicativa (solamente). Nuevamente, en vez de observar la variable verdadera de interés x_i^*, observaremos:

$$x_i = x_i^* + e_i \leftrightarrow x_i^* = x_i - e_i$$

Para ver qué problemas ocurre si sustituyemos la variable de interés x_i^* para su valor observado x_i, sustituyemos la expresión para x_i^* en el modelo verdadero:

$$\begin{aligned} y_i^* &= (x_i - e_i)\beta + u_i \\ \text{o} \quad y_i^* &= x_i\beta + (u_i - e_i\beta) \end{aligned}$$

Aquí el estimador MCO de β es sesgado e inconsistente. Para un valor dado de x_i^*, la x_i observada y el error de medición están correlacionadas positivamente, y esto implica que existe una correlación distinto a cero entre x_i y el término de error en el modelo de arriba $(u_i - e_i\beta)$. Específicamente:

- Para $\beta > 0$, implica una correlación negativa entre x_i y $(u_i - e_i\beta)$
- Para $\beta < 0$, implica una correlación positiva entre x_i y $(u_i - e_i\beta)$

- Para $\beta > 0$, el estimador MCO de β será sesgado hacía abajo
- Para $\beta < 0$, el estimador MCO de β será sesgado hacía arriba

En ambos casos, el estimador MCO de β será sesgado hacia cero: **sesgo de atenuación**. Veremos la derivación formal de esto a continuación, invocando los "supuestos clásicos de errores en variables".

Los supuestos clásicos de errores en variables son los siguientes:

$E(x_i^* e_i) = 0$	Error de medición no correlacionado con el verdadero x_i^*
$E(u_i e_i) = 0$	Error de medición no correlacionado con error del modelo
$V(e_i) = \sigma_e^2$	Error de medición es homoscedástico
$V(x_i^*) = \sigma_{x^*}^2$	Varianza poblacional del verdadero x_i^* existe y es finita

Para ver las implicancias formales del error de medición con estos supuestos tiene sobre el coeficiente estimado, partimos con la fórmula MCO para $\hat{\beta}_{MCO}$:

$$\hat{\beta}_{MCO} = (X'X)^{-1}X'y^* = \frac{\sum_{i=1}^N x_i y_i^*}{\sum_{i=1}^N x_i^2} = \frac{\frac{1}{N}\sum_{i=1}^N x_i y_i^*}{\frac{1}{N}\sum_{i=1}^N x_i^2}$$

Ahora, utilizando $x_i = x_i^* + e_i$ y $y_i^* = x_i^*\beta + u_i$ con los supuestos de arriba, obtenemos:

$$\text{plim}_{N\to\infty}\hat{\beta}_{MCO} = \frac{p\lim \frac{1}{N}\sum_{i=1}^N (x_i^* + e_i)(x_i^*\beta + u_i)}{p\lim \frac{1}{N}\sum_{i=1}^N (x_i^* + e_i)^2}$$

$$= \frac{\left(p\lim \frac{1}{N}\sum_{i=1}^N x_i^{*2}\right)\beta + p\lim \frac{1}{N}\sum_{i=1}^N x_i^* u_i + \left(p\lim \frac{1}{N}\sum_{i=1}^N x_i^* e_i\right)\beta + p\lim \frac{1}{N}\sum_{i=1}^N u_i e_i}{p\lim \frac{1}{N}\sum_{i=1}^N x_i^{*2} + 2p\lim \frac{1}{N}\sum_{i=1}^N x_i^* e_i + p\lim \frac{1}{N}\sum_{i=1}^N e_i^2}$$

$$= \frac{E(x_i^{*2})\beta + E(x_i^* u_i) + E(x_i^* e_i)\beta + E(u_i e_i)}{E(x_i^{*2}) + 2E(x_i^* e_i) + E(e_i^2)} = \frac{E(x_i^{*2})\beta + +0 + 0 + 0}{E(x_i^{*2}) + 0 + E(e_i^2)}$$

$$= \left(\frac{\sigma_{x^*}^2}{\sigma_{x^*}^2 + \sigma_e^2}\right)\beta = \frac{\beta}{1 + (\sigma_e^2/\sigma_{x^*}^2)} \neq \beta \qquad \text{si } \sigma_e^2 > 0 \tag{5.3}$$

La ecuación 5.3 implica que:

$$\text{plim}_{N\to\infty}\hat{\beta}_{MCO} = \frac{\beta}{1 + (\sigma_e^2/\sigma_{x^*}^2)} < \beta \quad \text{para } \beta > 0 \text{ y } \sigma_e^2 > 0$$

$$\text{plim}_{N\to\infty}\hat{\beta}_{MCO} = \frac{\beta}{1 + (\sigma_e^2/\sigma_{x^*}^2)} > \beta \quad \text{para } \beta < 0 \text{ y } \sigma_e^2 > 0$$

El estimador MCO para β es inconsistente, con un sesgo hacia cero que no disminuye cuando la muestra crece. Para un $\sigma_{x^*}^2$ dado, la severidad del sesgo de atenuación aumenta con la varianza

del error de medición (σ_e^2). El tamaño de la inconsistencia depende inversamente de la razón entre "señal" y "ruido" ($\sigma_{x^*}^2/\sigma_e^2$).

La Varianza del Error de Medición Bajo los supuestos clásicos de errores en variables con error de medición homoscedástico, la presencia del error de medición afecta la pendiente estimada, pero no la linealidad de la relación entre y_i^* y la variable observada x_i. Con error de medición heteroscedástico el error de medición también puede introducir de manera equivocada una relación no-lineal. Por ejemplo si $\beta > 0$ y $V(e_i)$ suele ser más grande para observaciones con x_i^* más grande, existiría una relación no lineal entre y_i^* y la variable observada x_i.

Regresión Múltiple con Errores en Variables Volvemos a considerar el modelo:

$$y_i^* = x_i^{*\prime}\beta + u_i$$
$$x_i' = x_i^{*\prime} + e_i'$$

donde x_i', $x_i^{*\prime}$ y e_i son vectores de $1 \times K$. Y como antes:

$$y_i^* = x_i'\beta + (u_i - e_i'\beta)$$

Por lo general, el estimador MCO del vector de parámetros β de $K \times 1$ será sesgado e inconsistente, dado que $E[x_i(u_i - e_i'\beta)] \neq 0$. En el caso especial cuando sólo una variable explicativa x_i está medida con error se puede demostrar que:

- El estimador MCO del coeficiente en esta variable está sesgado hacia cero (atenuación)
- El estimador MCO de los otros coeficientes también están sesgados en una dirección no determinada

Si varias variables explicativas están medidas con error es muy difícil inferir la dirección del sesgo para cualquier variable.

5.3.2 Variables Omitidas

El proceso generador de datos (PGD) de un determinado modelo explica el proceso que produce alguna variable de interés y_i en el mundo real. Es un desafío (no menor) saber qué PGD es el proceso verdadero para una variable de interés. Pero incluso si supiesemos perfectamente un PGD, existe una posibilidad muy real de que habrían variables que son relevantes, pero que simplemente no las observamos para incluirlas en nuestro modelo de regresión. Esto puede deberse a que simplemente no contamos con las variables en los datos, o porque son inherentemente inobservables, por ejemplo características personales importantes como habilidad o motivación u otras

características psicológicas. En un modelo de salario laboral, estas caraceteristicas probablemente son muy relevantes, pero es difícil argumentar que existen datos que tienen observaciones creíbles de todas estas características.

Con una única variable omitida y una única variable incluida en un modelo, el estimador MCO será sesgado si la variable omitida es relevante en el modelo y correlacionada con la variable incluida. El sesgo no desaparece en muestras grandes (inconsistente), y el signo del sesgo depende de la correlación entre la variable incluida y la variable omitida.

Por ende, las variables omitidas o 'heterogeneidad no-observada' es un desafio muy importante al momento de llegar a inferencia causal en regresiones de sección cruzada. Existe un peligro grave de que las variables observadas e inluidas están simplemente sirviendo como un proxy para factores no observados (y excluidos) – y no tengan un efecto directo y causal sobre el outcome de interés. Esto no es solo un problema en la economía empírica. Cualquier paper que sugiere que ha estimado un efecto causal sólo es confiable si se ha controlado por todos los otros factores relevantes, o utilizado otro tipo de diseño para llegar a una estimación creíblemente causal. Es importante ejercer precaución con resultados llamativos y sorprendentes de la prensa popular, o hasta en revistas académicas, cuando vienen de una sección cruzada y MCO!

Primero, consideremos un modelo con una variable incluida (x_{1i}) y una variable omitida (x_{2i})

$$y_i = x_{1i}\beta_1 + x_{2i}\beta_2 + u_i \quad \text{para } i = 1, \ldots, N,$$

con $E(u_i) = E(x_{1i}) = E(x_{2i}) = 0$, y $E(x_{1i}u_i) = E(x_{2i}u_i) = 0$. Bajo estos supuestos, si pudiésemos observar x_{1i} y x_{2i}, $\hat{\beta}_1$ y $\hat{\beta}_2$ serían insesgados. Pero aquí, el modelo que estimamos no incluye x_{2i}, relegando esta variable al término de error no observado:

$$y_i = x_{1i}\beta_1 + (u_i + x_{2i}\beta_2) \quad \text{para } i = 1, \ldots, N.$$

En forma matricial, escribimos este modelo como:

$$y = X_1\beta_1 + (u + X_2\beta_2),$$

y el estimador MCO de β_1 es:

$$\hat{\beta}_1 = (X_1'X_1)^{-1}X_1'y$$

Substityendo y del modelo verdero $y = X_1\beta_1 + X_2\beta_2 + u$, tenemos:

$$\begin{aligned}
\hat{\beta}_1 &= (X_1'X_1)^{-1}X_1'(X_1\beta_1 + X_2\beta_2 + u) \\
&= \beta_1 + [(X_1'X_1)^{-1}X_1'X_2]\beta_2 + (X_1'X_1)^{-1}X_1'u \\
&= \beta_1 + \widehat{\delta}\beta_2 + (X_1'X_1)^{-1}X_1'u.
\end{aligned}$$

Aquí, $\widehat{\delta} = (X_1'X_1)^{-1}X_1'X_2$ es el estimador MCO de la coeficiente de una regresión de la variable omitida sobre la variable x_{1i}:

$$x_{2i} = x_{1i}\delta + e_i$$

Y, entonces, tomando límites de probabilidad y utilizando el hecho de que $E(x_{1i}u_i) = 0$ obtenemos:

$$\text{plim}_{N\to\infty}\,\widehat{\beta}_1 = \beta_1 + (\text{plim}_{N\to\infty}\,\widehat{\delta})\beta_2 \tag{5.4}$$

De la ecuación 5.4, resulta evidence que el estimador MCO en el modelo con x_{2i} omitida es inconsistente salvo que una (o ambas) de los siguientes condiciones es cierta: (a) $\text{plim}_{N\to\infty}\,\widehat{\delta} = 0$, que implica que la variable omitida es ortogonal a x_{1i}, o (b) $\beta_2 = 0$, que implica que la variable omitida no es una variable relevante en el modelo verdadero. Entonces si omitimos una variable explicativa relevante ($\beta_2 \neq 0$), el único caso en que $\widehat{\beta}_1$ sigue siendo un estimador consistente para β_1 es el caso en que $\text{plim}_{N\to\infty}\,\widehat{\delta} = 0$, es decir, cuando x_{1i} y x_{2i} no están correlacionadas.

De la ecuación 5.4, también se nota que si x_{1i} y x_{2i} están positivamente correlacionadas, tenemos $\text{plim}_{N\to\infty}\,\widehat{\delta} > 0$. Si β_2 también es positiva, esperamos tener un sesgo hacía arriba del estimador MCO $\widehat{\beta}_1$. E inversamente, si x_{1i} y x_{2i} están negativamente correlacionadas ($\text{plim}_{N\to\infty}\,\widehat{\delta} < 0$) y $\beta_2 > 0$, tendremos un sesgo hacia abajo del estimador MCO $\hat{\beta}_1$. Intuitivamente, el estimador MCO estará incluyendo (de manera errónea) una relación indirecta entre x_{1i} y y_i dado que x_{1i} sirve parcialmente como un proxy para la variable x_{2i} omitida, además del efecto directo (causal) de x_{1i} sobre y_i, que es, al final, lo que nos interesa. Entonces, si un modelo de regresión omite factores relevantes (pero tal vez no observados/medibles) que están correlacionados con las variables incluidas, **no podemos deducir una inferencia causal** de las correlaciones parciales entre las variables observadas.

Algunos Ejemplos Ejemplo 1: La Hipótesis de Cantidad–Calidad

Existe mucha evidencia—al nivel micro y a nivel macro—que lo/as niño/as de familias más grandes tienen peores indicadores de capital humano acumulado. Esto ha resultado en la hipótesis de "cantidad–calidad" de fertilidad (Becker and Lewis, 1973; Becker and Tomes, 1976). Pero, la relación entre número de hermano/as y rendimiento educativa es causal? Una regresión de fertilidad sobre cantidad de hermano/as también capturará todos los aspectos de decisiones familiares relacionados con inversión en educación y decisiones de fertilidad. Por ejemplo: acceso a información puede afectar a la fertilidad y además, a la inversión en educación de la próximo generación. Es una pregunta aún abierta si las correlaciones observadas de sección cruzada representan una relación causal.

Ejemplo 2: Retornos a Educación

La relación entre educación y el salario laboral es una de las más antiguas preguntas en economía laboral. ¿Por qué ganan más en promedio los individuos con más educación? ¿Es una relación directa (causal): la educación aumenta productividad y por ende salarios? ¿O una correlación

con una variable omitida: las personas más productivas también tienden en promedio a tener más educación?

Regresión Múltiple con Variables Omitidas Este análisis sigue en una manera parecida al análisis con una sola variable incluida y una variable excluida. Escribimos un modelo que consiste de variables incluidas (X_1) y variables excluidas (X_2):

$$y = X_1\beta_1 + (X_2\beta_2 + u) \tag{5.5}$$

donde ahora X_1 es $N \times K_1$, β_1 es $K_1 \times 1$, X_2 es $N \times K_2$ y β_2 es $K_2 \times 1$ (es decir, hay K_1 regresores incluidas y K_2 variables omitidas). Y como en la versión con una sola variable podemos obtener:

$$\text{plim}_{N\to\infty}\,\widehat{\beta_1} = \beta_1 + (\text{plim}_{N\to\infty}(X_1'X_1)^{-1}X_1'X_2)\beta_2. \tag{5.6}$$

Cada columna de la matriz de $K_1 \times K_2$ $(X_1'X_1)^{-1}(X_1'X_2)$ es un vector de $K_1 \times 1$ de los estimadores MCO de los coeficientes en una regresión múltiple de la columna correspondente a X_2 en todas las variables incluidas X_1.

Por lo general, es mucho más difícil[1] estar seguro de la dirección del sesgo esperado en los coeficientes estimados de β_1. Si hay **una sola variable omitida** ($K_2 = 1$), que está correlacionada con varias de las variables incluidas, se puede demostrar que el sesgo en cada coeficiente dependerá de su correlación *parcial* con la variable omitida, y no simplemente la correlación *simple* entre la variable incluida y la variable excluida. Es decir, la dirección del sesgo depende del signo de los coeficientes en una regresión múltiple de la variable omitida con todas las variables incluidas juntas. Si hay **mútliples variables omitidas** es difícil decir algo acerca del signo del sesgo, pero el estimador MCO $\widehat{\beta_1}$ es un estimador sesgado e inconsistente para β_1 en el modelo verdadero, salvo en el caso especial en que *todas* las variables omitidas sean ortogonales a *todas* las variables incluidas (ver Basu (2018) para una discusión).

Para ver el caso con una sola variable omitida, escribimos:

$$y_i = \beta_1 + \beta_2 x_{2i} + \ldots + \beta_{K-1}x_{K-1,i} + (\beta_K x_{Ki} + u_i) \tag{5.7}$$

donde $E(x_{ki}u_i) = 0$ para $k = 1, \ldots, K$ (y $x_{1i} = 1$ para cada $i = 1, \ldots, N$). Ahora, si realizamos la siguiente regresión (o "proyección lineal") de la variable omitida sobre las variables incluidas:

$$x_{Ki} = \delta_1 + \delta_2 x_{2i} + \ldots + \delta_{K-1}x_{K-1,i} + v_i,$$

[1] Una descripción de este problema está disponible en Clarke (2005). Hay algunos trabajos recientes que describen este sesgo (Basu, 2018), o documenten una solución simple para el sesgo con varias variables omitidas e incluidas en situaciones específicas (Clarke, 2019).

podemos reescribir la ecuación 5.7 como:

$$y_i = (\beta_1 + \beta_K\delta_1) + (\beta_2 + \beta_K\delta_2)x_{2i} + \ldots + (\beta_{K-1} + \beta_K\delta_{K-1})x_{K-1,i} + (u_i + \beta_K v_i).$$

Sabemos que $E(x_{ki}u_i) = 0$ por las propiedades de una regresión, y $E(x_{ki}v_i) = 0$ por las propiedades de una proyección lineal. Por lo tanto: $E(x_{ki}(u_i + \beta_K v_i)) = 0$ para $k = 1, \ldots, K-1$. Entonces:

$$\text{plim}_{N\to\infty}\,\widehat{\beta}_k = \beta_k + \beta_K\delta_k \quad \text{para } k = 1, \ldots, K-1.$$

La inconsistencia depende del signo de las correlaciones parciales reflejadas por los δ_k, y no de las correlaciones simples entre cada x_{ki} y x_{Ki}.

Ejercicios de Ayudantía:

1. Considere el estimador de MCO para un modelo con k variables independientes que cumple los supuestos de Gauss-Markov:

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

   (a) Demuestre la insesgadez del estimador.

   (b) Derive la varianza de los estimadores.

   (c) Explique cómo contrastaría la presencia de hetrocedasticidad con el contraste de Breush-Pagan.

2. Considere el siguiente modelo e hipótesis nula:

$$y = \beta_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4 + u$$

$$H_0 : \beta_3 = -\beta_2$$

   (a) Reparametrice el modelo de tal forma que pueda contrastar directamente la hipótesis nula con un sólo coeficiente del modelo reparametrizado.

   (b) ¿Cómo contrastaría la hipótesis planteada empleando el estadístico F con matrices?

   (c) Describa cómo se puede comprobar esta hipótesis utilizando el R^2 de dos modelos.

3. * Considere el modelo:

$$y_i = \beta_0 + \beta_1 x_{1i} + \beta_2 x_{2i} + \beta_3 x_{3i} + \beta_4 x_{4i} + u_i$$

   donde cada $x_{ki} \forall k \in \{1, \ldots, 5\}$ es una variable binaria, y cada variable es mutuamente exclusivo en el sentido que si una $x_{ki} = 1$, ninguna otra variable x_{-ki} es igual a 1. Ahora, suponga que se estima el modelo omitiendo las variables x_{3i} y x_{4i}. Derive el sesgo por variables omitidas en el parámetro estimado β_2. *Pista*: Podría ser útil definir un escalar N_{xk} para cada x que es la cantidad de veces que esta variable tome valor 1. Además, podría ser relevante referir a la fórmula para el inverso de una matriz, tipo "Arrowhead".

Sección 6

Una Breve Introducción a las Variables Instrumentales

Nota de Lectura: Hay muchas buenas exposiciones de variables instrumentales en libros de texto. El capítulo de 10 de Hansen (2017) es muy bueno. El Capítulo 5 de Wooldridge (2002) también es muy comprensivo. Cameron and Trivedi (2005) tiene una exposición concisa y estrechamente vinculada con el modelo lineal clásico.

6.1 Introducción a las Variables Instrumentales

Hemos visto que el supuesto de exogeneidad, $E(x_i u_i) = 0$ es fundamental para la consistencia (e insesgadez) del estimador $\widehat{\beta}_{MCO}$ del modelo de regresión lineal. Pero también sabemos que existen muchos casos bajo los cuales NO tendremos una nula correlación entre el error estocástico u y variables independientes X. Entre ellos, la endogeneidad puede ser provocada por errores de medición, variables omitidas, o simultaneidad en las relaciones de interés.

A diferencia de supuestos como el de heteroscedasticad, no multicolinealidad o independencia entre observaciones, la endogeneidad no tiene una solución "simple" que permita seguir utilizando el modelo de regresión lineal clásico.[1] Una solución potentical a la endogeneidad es utilizar **variables instrumentales**. Brevemente, consideremos un modelo lineal simple:

$$y_i = x_i'\beta + u_i \qquad \text{para } i = 1, 2, \ldots, N. \tag{6.1}$$

Aquí x_i y u_i están correlacionadas. El uso de variables instrumentales busca reemplazar esta variable endógena x_i por valores predichos de x_i que cumplen con dos condiciones fundamentales: (1)

[1] Por supuesto, si el problema es por una variable omitida y se puede medir e incluir la variable en el modelo lineal, entonces no tenemos un problema, ya que de esta forma recuperamos el supuesto de exogeneidad. En estos apuntes estamos suponiendo que no podemos simplemente incluír variables observables y disponibles a nuestro modelo y así rescatar exogeneidad. Por lo tanto, tenemos que buscar una otra solución.

Deben estar relacionados con la variable actual x_i, pero, a la vez, (2) No deben estar correlacionados con u_i.

Estos valores predichos están formados mediante una proyección lineal de la variable endógena x_i sobre una serie de variables instrumentales, que requieren dos propiedades, análogas a (1) y (2) anteriormente:

1. Las variables instrumentales deben estar correlacionadas con la(s) variable(s) endógena(s) x_i
2. Las variables instrumentales no deben estar correlacionadas con los errores u_i

Estas dos propiedades se conocen como **Relevancia** y **Validez**. Bajo estas dos propiedades, demostraremos que el estimador de variables instrumentales explicado más adelante es un estimador consistente para β cuya varianza asintótica es estimable.

El desfío principal con variables instrumentales es encontrar una variable que cumpla con estas dos propiedades! Por lo general no es difícil encontrar variables que estén correlacionadas con alguna variable x_i, pero poder argumentar que no están correlacionadas con u_i resulta difícil. Y viceversa, sería fácil encontrar una variable no correlacionada con u_i si la variable es completamente aleatoria, pero en este caso, no va a estar correlacionada con x_i. Como veremos más adelante, estos supuestos son fundamentales en modelos de variables instrumentales. Incluso si los supuestos sólo son "un poco" inválidos, el estimador de variables instrumentales puede ser peor aún que el estimador (sesgado e inconsistente) de $\widehat{\beta}_{MCO}$.

6.2 Estimación Utilizando Variables Instrumentales

Existen varias maneras de estimar parámetros de interés utilizando variables instrumentales. En este curso vamos a considerar varias maneras de estimar parámetros utilizando instrumentos, las que consisten en minimizar los residuos al cuadrado. Pero también se pueden estimar los parámetros con variables instrumentales utilizando el **Método de Momentos** o el **Método de Momentos Generalizados** (GMM). Van a encontrar estos métodos en el segundo curso de econometría del magíster, y además, varias pruebas de interés relacionados con la estimación por GMM.

6.2.1 El Estimador de Variables Instrumentales

Para partir, imaginemos que tenemos un modelo con una variable instrumental para cada variable endógena (una situación conocida como identificación exacta). Por ahora no nos preguntaremos por qué creemos que los instrumentos son válidos, sólo supongamos que cada instrumento (llamado z_i) cumple con $E(z_i u_i) = 0$.

Partiendo con este supuesto de validez $E(z_i u_i) = 0$, podemos reeemplazar u_i por $u_i = y_i - x_i'\beta$ (implicado en 6.1):

$$\begin{aligned} E(z_i u_i) &= 0 \\ E(z_i(y_i - x_i'\beta)) &= 0 \\ E(z_i y_i) - E(z_i x_i')\beta &= 0 \end{aligned} \tag{6.2}$$

Ahora, utilizando la igualdad en 6.2, podemos resolver para β siempre y cuando $E(z_i x_i')$ sea invertible:

$$\beta = E(z_i x_i')^{-1} E(z_i y_i).$$

Por suerte, la invertibilidad de $E(z_i x_i')$ ya viene por el supuesto de relevancia discutido anteriormente. Notemos que en el caso en que tenemos múltiples variables endógenas y múltiples variables instrumentales, esto implica que las variables instrumentales z_i tienen que tener poder explicativo para cada variable endógena después de condicionar en cualquier otra variable incluida en el modelo (incluyendo los otros instrumentos). Más adelante vamos a definir esto con un poco más de formalidad.

El estimador de variables instrumentales $\widehat{\beta}_{IV}$ reemplaza los momentos poblacionales en 6.2 con sus momentos muestrales. Encontramos:

$$\begin{aligned} \widehat{\beta}_{IV} &= \left(\frac{1}{n}\sum_{i=1}^{n} z_i x_i\right)^{-1} \left(\frac{1}{n}\sum_{i=1}^{n} z_i y_i\right) \\ &= \left(\sum_{i=1}^{n} z_i x_i\right)^{-1} \left(\sum_{i=1}^{n} z_i y_i\right) \\ \widehat{\beta}_{IV} &= (Z'X)^{-1}(Z'y) \end{aligned} \tag{6.3}$$

El estimador en 6.3 es la versión matricial del estimador de variables instrumentales.

6.2.2 Mínimos Cuadrados en Dos Etapas

En el caso de $\widehat{\beta}_{IV}$, hemos asumido que tenemos tantas variables instrumentales como variables endógenas. Ahora, ¿qué pasaría si tenemos más variables instrumentales que variables endógenas? Para casos de este tipo, podemos utilizar un estimador alternativo de variables instrumentales: el estimador de Mínimos Cuadrados en Dos Etapas (MC2E). Este estimador sirve en la clase general de circunstancias cuando $L \geq K$, donde L se refiere a la cantidad de instrumentos, y K a la cantidad de variables endógenas. Cuando $L > K$ esta situación se conoce como "sobre-identificación", y puede ser bastante útil para considerar la plausibilidad de los supuestos de variables instrumentales. Si estamos en un caso cuando $L < K$, se conoce como sub-identificación, y no tenemos la información suficiente para estimar los parámetros β.

La necesidad de tener por lo menos tantos instrumentos como variables endógenos se resume en la condición de rango. Consideremos un caso con L variables instrumentales para K variables endógenas. Entonces la condición de rango dice que la matriz de $L \times K$, $E(z_i x_i')$ tiene rango completo de columna:

$$rango(E(z_i x_i')) = K.$$

En otras palabras, los instrumentos tienen que tener poder suficiente para explicar por lo menos algo de cada variable endógena. Este supuesto es necesario y suficiente para poder identificar β.

El estimador de mínimos cuadrados en dos etapas consiste en – lógicamente – realizar dos etapas. Primero, proyectamos la(s) variable(s) endógena(s) en función de los instrumentos, y segundo estimamos el modelo de interés donde cada variable endógena es reemplazada por su valor predicho en la primera etapa. En formato matricial, escribimos:

$$y = X\beta + u \tag{6.4}$$
$$X = Z\pi + v \tag{6.5}$$

Aquí, la primera etapa es 6.5, y los parámetros de interés son los β de 6.4.

Para partir, con la estimación de β, estimamos la(s) coeficiente(s) en la primera etapa, π, utilizando MCO:

$$\widehat{\pi} = (Z'Z)^{-1}Z'X,$$

y formamos los valores predichos de X como:

$$\widehat{X} = Z\widehat{\pi} = Z(Z'Z)^{-1}Z'X. \tag{6.6}$$

Ahora, podemos formar la segunda etapa de la regresión utilizando este X predicho, lo que escribimos como:

$$y = \widehat{X}\beta + (u + (X - \widehat{X})\beta). \tag{6.7}$$

Notemos que aquí por el supuesto de validez de los instrumentos, $\widehat{X}$ *no* está correlacionado con u, aunque la variable original X es una variable endógena. Como veremos más adelante, este segundo término del error $(X - \widehat{X})\beta$ será una fuente de sesgo, pero *no* de inconsistencia.

La estimación del parámetro β se procede utilizando MCO en la regresión de segunda etapa descrita en ecuación 6.7. Esto es simplemente una aplicación de la fórmula ya conocida de MCO:

$$\begin{aligned}
\widehat{\beta}_{MC2E} &= (\widehat{X}'\widehat{X})^{-1}\widehat{X}'y \\
&= [(Z(Z'Z)^{-1}Z'X)'(Z(Z'Z)^{-1}Z'X)]^{-1}(Z(Z'Z)^{-1}Z'X)'y \\
&= [(X'Z(Z'Z)^{-1}Z')(Z(Z'Z)^{-1}Z'X)]^{-1}(X'Z(Z'Z)^{-1}Z')y \\
&= [X'Z(Z'Z)^{-1}Z'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y \\
&= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y.
\end{aligned} \tag{6.8}$$

Arriba, el primer paso viene de sustituir $\widehat{X} = Z(Z'Z)^{-1}Z'X$ de la ecuación 6.6 y el segundo paso viene de las reglas de transposición de matriz y utilizando la simetría de $(Z'Z)^{-1}$ que implica que $[(Z'Z)^{-1}]' = (Z'Z)^{-1}$.[2]

Notemos que si volvemos a sustituir la expresión de la ecuación 6.6, podemos re-escribir el estimador como:

$$\begin{aligned}\widehat{\beta}_{MC2E} &= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= [(Z(Z'Z)^{-1}Z'X)'X]^{-1}(Z(Z'Z)^{-1}Z'X)'y \\ &= (\widehat{X}'X)^{-1}\widehat{X}'y \end{aligned} \tag{6.9}$$

Este estimador se ve muy parecido al estimador típico de MCO, y, de hecho, si $z_i = x_i$, tenemos que $\widehat{\beta}_{MC2E} = \widehat{\beta}_{MCO}$. En otras palabras, si utilizamos x_i como un instrumento para sí mismo, volvemos a tener el estimador MCO. En realidad, esto es un resultado bastante intuitivo: si proyectamos x_i sobre sí mismo, las predicciones son perfectas, y la segunda etapa de MC2E es simplemente la regresión MCO. Además, es claro que en este caso, el supuesto clave de variables instrumentales: $E(z_i u_i) = 0$ vuelve a ser el mismo supuesto de exogeneidad de MCO: $E(x_i u_i) = 0$.

El estimador de mínimos cuadrados en dos etapas funciona incluso cuando tenemos múltiples instrumentos en la matriz Z (o también con múltiples variables endógenas X, siempre y cuando haya por lo menos tantos instrumentos como variables endógenas). En el caso especial en que la cantidad de instrumentos es igual a la cantidad de variables endógenas, el estimador de MC2E se convierte en $\widehat{\beta}_{MC2E} = (Z'X)^{-1}Z'y$ (es decir, se convierte en el estimador $\widehat{\beta}_{IV}$). Para ver porqué es así, consideramos que en este caso, $X'Z$, $Z'Z$ and $Z'X$ son todos matrices cuadradas. De las propiedades de inversión de matrices, sabemos que para matrices invertibles de $n \times n$ A y B, es cierto que $(AB)^{-1} = B^{-1}A^{-1}$. Entonces,

$$\begin{aligned}\widehat{\beta}_{MC2E} &= [X'Z(Z'Z)^{-1}Z'X]^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}(Z'Z)(X'Z)^{-1}X'Z(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}(Z'Z)(Z'Z)^{-1}Z'y \\ &= (Z'X)^{-1}Z'y \\ &= \widehat{\beta}_{IV}.\end{aligned}$$

Es particularmente fácil ver esto en el caso de 1 instrumento y una variable endógena, utilizando

[2] La transpuesta de una matriz de la forma $(A_1A_2\cdots A_{k-1}A_k)$ es igual a $(A_1A_2\cdots A_{k-1}A_k)' = (A_k'A_{k-1}'\cdots A_2'A_1')$. El cálculo en 6.8 es simplemente una aplicación de ésta, donde k es igual a 4, y $A_1 = Z$, $A_2 = (Z'Z)^{-1}$, $A_3 = Z'$, y $A_4 = X$.

la ecuación 6.10 y $\widehat{X} = Z\widehat{\pi}$, donde $\widehat{\pi}$ es un escalar:

$$\begin{aligned}
\widehat{\beta}_{MC2E} &= (\widehat{X}'X)^{-1}\widehat{X}'y \\
&= [(Z\widehat{\pi})'X]^{-1}(Z\widehat{\pi})'y \\
&= [\widehat{\pi}(Z'X)]^{-1}\widehat{\pi}(Z'y) \\
&= \frac{1}{\widehat{\pi}}(Z'X)^{-1}\widehat{\pi}(Z'y) \\
&= (Z'X)^{-1}Z'y.
\end{aligned}$$

6.3 Consistencia del Estimador y Teoría Asintótica

6.3.1 Consistencia

La estimación por MC2E *nunca* produce un estimador insesgado de β. Con variables instrumentales, siempre hay un sesgo en muestras finitas.Sin embargo, este sesgo desaparece en la medida que la muestra crece sin límites: por lo menos el estimador de variables instrumentales es un estimador consistente. Para ver esto, partimos con el estimador MC2E $\widehat{\beta}_{MC2E} = (\widehat{X}'X)^{-1}\widehat{X}'y$. Sustituyendo $y = X\beta + u$ nos da:

$$\begin{aligned}
\widehat{\beta}_{MC2E} &= (\widehat{X}'X)^{-1}\widehat{X}'(X\beta + u) \\
&= (\widehat{X}'X)^{-1}\widehat{X}'X\beta + (\widehat{X}'X)^{-1}\widehat{X}'u \\
&= \beta + \left(\frac{\widehat{X}'X}{N}\right)^{-1}\left(\frac{\widehat{X}'u}{N}\right). \qquad (6.10)
\end{aligned}$$

Tomamos límites en probabilidad de la ecuación 6.10 para examinar el comportamiento del estimador en muestras grandes:

$$\text{plim}_{\text{N}\to\infty}\,\widehat{\beta}_{MC2E} = \beta + \text{plim}_{\text{N}\to\infty}\left(\frac{\widehat{X}'X}{N}\right)^{-1}\text{plim}_{\text{N}\to\infty}\left(\frac{\widehat{X}'u}{N}\right). \qquad (6.11)$$

Para mostrar la consistencia del estimador de mínimos cuadrados en dos etapas tenemos que demostrar que $\text{plim}_{\text{N}\to\infty}\,\widehat{\beta}_{MC2E} = \beta$. De la ecuación 6.11, esto requiere que: (a) $\text{plim}_{\text{N}\to\infty}[(\widehat{X}'u)/N] = 0$ y (b) $\text{plim}_{\text{N}\to\infty}[(\widehat{X}'X)/N]^{-1}$ existe y es finito.

Considerando (a), sabemos de nuestro supuesto de validez instrumental que $E(z_i u_i) = 0$. Y de la Ley de los Grandes Números, el vector de medias muestrales converge en probabilidad, con

$E(z_i u_i) = 0$:

$$\frac{1}{n}\sum_{i=1}^{n} z_i u_i \xrightarrow{P} E(z_i u_i) = 0$$
$$\left(\frac{Z'u}{n}\right) \xrightarrow{P} 0 \tag{6.12}$$

donde el lado izquierdo de (6.12) es simplemente la forma matricial de $\frac{1}{n}\sum_{i=1}^{n} z_i u_i$. Dado que $\widehat{X} = Z\widehat{\pi}$, se puede expresar:

$$\left(\frac{\widehat{X}'u}{n}\right) = \left(\frac{(Z\widehat{\pi})'u}{n}\right) = \left(\frac{(\widehat{\pi}'Z')u}{n}\right) = \widehat{\pi}'\left(\frac{Z'u}{n}\right). \tag{6.13}$$

Finalmente, $\widehat{\pi}$ es un estimador consistente de π en la primera etapa, y por lo tanto, podemos escribir $\widehat{\pi} \xrightarrow{P} \pi \neq 0$, donde la última desigualdad viene del supuesto de relevancia. Ahora, combinando (6.12), (6.13), la consistencia de $\widehat{\pi}$, e invocando el Teorema de Slutsky[3], tenemos:

$$\text{plim}_{\text{N}\to\infty}\left(\frac{\widehat{X}'u}{n}\right) = \text{plim}_{\text{N}\to\infty}\left[\widehat{\pi}'\left(\frac{Z'u}{n}\right)\right] = \pi'0 = 0, \tag{6.14}$$

que es la parte estipulada de (a).

Y para la parte (b), partimos desde el otro supuesto—el supuesto de relevancia: $E(z_i x_i) \neq 0$. De nuevo, de la LGN, afirmamos que el vector de medias muestrales converge a $E(z_i x_i) \neq 0$:

$$\frac{1}{n}\sum_{i=1}^{n} z_i x_i \xrightarrow{P} E(z_i x_i) = M_{ZX} \neq 0$$
$$\left(\frac{Z'X}{n}\right) \xrightarrow{P} M_{ZX} \neq 0. \tag{6.15}$$

Ahora, de forma análoga a 6.13, podemos escribir $\left(\frac{\widehat{X}'X}{n}\right)$ como:

$$\left(\frac{\widehat{X}'X}{n}\right) = \left(\frac{(Z\widehat{\pi})'X}{n}\right) = \left(\frac{(\widehat{\pi}'Z')X}{n}\right) = \widehat{\pi}'\left(\frac{Z'X}{n}\right). \tag{6.16}$$

De nuevo aplicamos el Teorema de Slutsky con (6.16) que da: $\text{plim}_{\text{N}\to\infty}\left(\frac{\widehat{X}'X}{n}\right) = \pi' M_{ZX} \neq 0$, y por lo tanto

$$\text{plim}_{\text{N}\to\infty}\left(\frac{\widehat{X}'X}{n}\right)^{-1} = (\pi' M_{ZX})^{-1} \tag{6.17}$$

es finita.

[3] El teorema de Slutsky declara que para dos vectores o matrices aleatorias X_n y Y_n, si X_n converge en probabilidad a un vector/matriz aleatorio X e Y_n converge en probabilidad a un constante c, entonces, $X_n Y_n \xrightarrow{d} cX$, donde $\xrightarrow{d}$ denota convergencia en distribución.

Combinando los dos resultados de (6.14) y (6.17) con la ecuación (6.11) obtenemos el resultado de consistencia del estimador de MC2E:

$$\text{plim}_{N\to\infty}\widehat{\beta}_{MC2E} = \beta \qquad \text{o} \qquad \widehat{\beta}_{MC2E} \xrightarrow{P} \beta.$$

Así, concluimos que con los supuestos de validez instrumental y relevancia instrumental, el método de variables instrumentales produce un estimador consistente de β.

6.3.2 Resultados de Distribución Límite

Los resultados asintóticos fundamentales de MC2E vienen de White (1982) y Hansen (1982). El resultado más importante para nosotros es que con instrumentos relevantes y válidos, y observaciones que son independientes e idénticamente distribuidas, tenemos un resultado de normalidad asintótica:

$$\sqrt{n}(\widehat{\beta}_{MC2E} - \beta) \xrightarrow{D} \mathcal{N}(0, V),$$

o, de la misma forma:

$$\widehat{\beta}_{MC2E} \overset{a}{\sim} \mathcal{N}(\beta, V/n).$$

Con esto, en muestras grandes, podemos construir intervalos de confianza y comprobar pruebas de hipótesis acerca del vector de parámetros β, si encontramos un estimador consistente de V. Bajo el supuesto de homoscedasticidad condicional: $E(u_i^2|x_i) = \sigma^2$ para cada i, un estimador consistente de V es:

$$\begin{aligned}\widehat{V} &= n\widehat{\sigma}^2(X'Z(Z'Z)^{-1}Z'X)^{-1} \\ &= n\widehat{\sigma}^2(\widehat{X}'\widehat{X})^{-1}.\end{aligned}$$

Se puede comprobar que el segundo paso sigue del primer paso sustituyendo $\widehat{X} = Z(Z'Z)^{-1}Z'X$ en el segundo paso, y simplificando (utilizando el hecho que $(Z'Z)^{-1} = (Z'Z)^{-1\prime}$. La cantidad $\widehat{\sigma}^2$ se puede estimar utilizando $\widehat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n \widehat{u}_i^2$ donde $\widehat{u}_i = y - x_i'\widehat{\beta}_{MC2E}$. Esto sugiere un estimador para la varianza de $\widehat{\beta}_{MC2E}$ de:

$$\widehat{V}(\widehat{\beta}_{MC2E}) = \frac{\widehat{V}}{n} = \widehat{\sigma}^2(\widehat{X}'\widehat{X})^{-1}.$$

Pero, notemos que esta varianza no es actualmente correcta! Esto se debe a que los residuos en el calculo de $\widehat{\sigma}^2$ son:

$$\widehat{u}_i = y - x_i'\widehat{\beta}_{MC2E} \neq y - \hat{x}_i'\widehat{\beta}_{MC2E}.$$

De esta forma, $\widehat{\sigma}^2$ no es consistente utilizar los residuos de la segunda etapa, dado que existe el segundo término $(x_i' - \hat{x}_i')\beta$ en el término de error en la segunda etapa. La versión "correcta" del estimador de varianza está disponible en la mayoría de las implementaciones computacionales de variables instrumentales. Es importante recordar que si se quiere estimar una regresión de vari-

ables instrumentales utilizando MC2E "a mano", aunque los parámetros de la segunda etapa serán correctos, los errores estándar no lo serán a menos que se haga el ajuste para calcular $\widehat{\sigma}^2$ de forma consistente.

Ejercicios de Ayudantía:

1. El modelo que describe la salud de los individuos toma la siguiente forma:

$$salud_i = \beta_1 + \beta_2 fumador_i + \beta_3 educacion_i + \beta_4 tasa\_descuento_i + u_i \tag{6.18}$$

   Donde *salud* es un índice de buena salud, $fumador$ toma el valor 1 si el individuo fuma, *educacion* corresponde a los años de educación y $tasa\_descuento$ es un indicador del peso que se da al bienestar presente con respecto al futuro. Sin embargo, el investigador omite la tasa de descuento y considera el siguiente modelo mal especificado:

$$salud_i = \beta_1 + \beta_2 fumador_i + \beta_3 educacion_i + v_i \tag{6.19}$$

   (a) Derive la dirección probable del sesgo de los parámetros estimados $\tilde{\beta}_2$ y $\tilde{\beta}_3$ del modelo (2).

   (b) ¿Cuáles son los requisitos formales de un instrumento?

   (c) ¿Qué podría ser un buen instrumento en cada caso?

   (d) ¿Se pueden comprobar los requisitos de los instrumentos? ¿Cómo lo haría?

2. Considere el siguiente modelo con problemas de endogeneidad:

$$Y = X\beta + V \tag{6.20}$$

   Demuestre que cuando la cantidad de instrumentos es igual a la cantidad de variables endógenas, es decir, hay identificación exacta, el estimador MC2E es igual al estimador de VI.

Bibliography

Adams, Abi, Damian Clarke, and Simon Quinn. 2015. *Microeconometrics and MATLAB: An Introduction.* Oxford University Press.

Angrist, Joshua D., and Jörn-Steffen Pischke. 2009. *Mostly Harmless Econometrics: An Empiricist's Companion.* Princeton University Press.

Basu, Deepankar. 2018. "Bias of OLS Estimators due to Exclusion of Relevant Variables and Inclusion of Irrelevant Variables." University of Massachusetts Amherst, Department of Economics UMASS Amherst Economics Working Papers 2018-19.

Becker, Gary S, and H Gregg Lewis. 1973. "On the Interaction between the Quantity and Quality of Children." *Journal of Political Economy*, 81(2): S279–88.

Becker, Gary S, and Nigel Tomes. 1976. "Child Endowments and the Quantity and Quality of Children." *Journal of Political Economy*, 84(4): S143–62.

Breusch, T. S., and A. R. Pagan. 1979. "A Simple Test for Heteroscedasticity and Random Coefficient Variation." *Econometrica*, 47(5): 1287–1294.

Cameron, A. Colin, and Douglas L. Miller. 2015. "A Practitioner's Guide to Cluster-Robust Inference." *The Journal of Human Resources*, 50(2): 317–72.

Cameron, A. Colin, and Pravin K. Trivedi. 2005. *Microeconometrics: Methods and Applications.* Cambridge University Press.

Cameron, A. Colin, and Pravin K. Trivedi. 2009. *Microeconometrics Using Stata.* Stata Press.

Casella, George, and Roger L Berger. 2002. *Statistical Inference.* . 2 ed., Duxberry Thomson.

Clarke, Damian. 2019. "A Convenient Omitted Variable Bias Formula for Treatment Effect Models." *Economics Letters*, 174: 84–88.

Clarke, Kevin A. 2005. "The Phantom Menace: Omitted Variable Bias in Econometric Research." *Conflict Management and Peace Science*, 22(4): 341–352.

Davidson, James. 1994. *Stochastic Limit Theory.* Oxford University Press.

Deaton, Angus. 1997. *The Analysis of Household Surveys – A Microeconometric Approach to Development Policy.* The Johns Hopkins University Press.

DeGroot, Morris H., and Mark J. Schervish. 2012. *Probability and Statistics.* . 4 ed., Addison-Wesley.

Eicker, Friedhelm. 1967. "Limit theorems for regressions with unequal and dependent errors." *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1: 59–82.

Fisher, R. A. 1922. "On the Mathematical Foundations of Theoretical Statistics." *Philosophical Transactions of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 222(594-604): 309–368.

Frisch, Ragnar. 1933. "Editor's Note." *Econometrica*, 1(1): 1–4.

Frisch, Ragnar, and Frederick V. Waugh. 1933. "Partial Time Regressions as Compared with Individual Trends." *Econometrica*, 1(4): 387–401.

Goldberger, Arthur S. 1991. *A Course in Econometrics.* Harvard University Press.

Golub, G.H., and C.F. Van Loan. 1983. *Matrix Computations.* Johns Hopkins University Press.

Greene, William H. 2002. *Econometric Analysis.* . 5 ed., Pearson.

Hansen, Bruce. 2017. *Econometrics.* Online Manuscript, http://www.ssc.wisc.edu/~bhansen/econometrics/, descargado 26/12/2017.

Hansen, Lars. 1982. "Large Sample Properties of Generalized Method of Moments Estimators." *Econometrica*, 50(4): 1029–54.

Huber, P J. 1967. "The behavior of maximum likelihood estimates under nonstandard conditions." *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, 1: 221–233.

Lincove, Jane Arold. 2008. "Growth, Girls' Education, and Female Labor: A Longitudinal Analysis." *The Journal of Developing Areas*, 41(2): 45–68.

Lovell, Michael C. 1963. "Seasonal Adjustment of Economic Time Series and Multiple Regression Analysis." *Journal of the American Statistical Association*, 58(304): 993–1010.

Manski, Charles F. 2003. *Partial Identification of Probability Distributions.* Springer.

Mullainathan, Sendhil, and Jann Spiess. 2017. "Machine Learning: An Applied Econometric Approach." *Journal of Economic Perspective*, 31(2): 87–106.

Rao, C. Radhakrishna. 1973. *Linear Statistical Inference and its Applications.* John Wiley and Sons.

Simon, Carl P., and Lawrence Blume. 1994. *Mathematics for Economists.* New York, N.Y.:W. W. Norton & Company, Inc.

Solon, Gary, Steven J. Haider, and Jeffrey M. Wooldridge. 2015. "What Are We Weighting For?" *Journal of Human Resources*, 50(2): 301–316.

Stachurski, John. 2016. *A Primer in Econometric Theory.* The MIT Press.

White, Halbert. 1980. "A Heteroskedasticity-Consistent Covariance Matrix Estimator and a Direct Test for Heteroskedasticity." *Econometrica*, 48(4): 817–838.

White, Halbert. 1982. "Maximum Likelihood Estimation of Misspecified Models." *Econometrica*, 50(1): 1–25.

White, Halpert. 2001. *Asymptotic Theory for Econometricians.* San Diego, Academic Press.

Wilks, S. S. 1938. "The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses." *The Annals of Mathematical Statistics*, 9(1): 60–62.

Wooldridge, J. M. 2002. *Econometric Analysis of Cross Section and Panel Data.* Cambridge, Massachusetts:The MIT Press.

Table 6.1: fda Normal Estandarizada

x	0.00	0.01	0.02	0.03	0.04	0.05	0.06	0.07	0.08	0.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7703	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998
3.5	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998	0.9998
3.6	0.9998	0.9998	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.7	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.8	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999	0.9999
3.9	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000

Table 6.2: Valores Críticos de la Distribución t de Student

k	60.0%	66.7%	75.0%	80.0%	87.5%	90.0%	95.0%	97.5%	99.0%	99.5%	99.9%
1	0.325	0.577	1.000	1.376	2.414	3.078	6.314	12.706	31.821	63.657	318.31
2	0.289	0.500	0.816	1.061	1.604	1.886	2.920	4.303	6.965	9.925	22.327
3	0.277	0.476	0.765	0.978	1.423	1.638	2.353	3.182	4.541	5.841	10.215
4	0.271	0.464	0.741	0.941	1.344	1.533	2.132	2.776	3.747	4.604	7.173
5	0.267	0.457	0.727	0.920	1.301	1.476	2.015	2.571	3.365	4.032	5.893
6	0.265	0.453	0.718	0.906	1.273	1.440	1.943	2.447	3.143	3.707	5.208
7	0.263	0.449	0.711	0.896	1.254	1.415	1.895	2.365	2.998	3.499	4.785
8	0.262	0.447	0.706	0.889	1.240	1.397	1.860	2.306	2.896	3.355	4.501
9	0.261	0.445	0.703	0.883	1.230	1.383	1.833	2.262	2.821	3.250	4.297
10	0.260	0.444	0.700	0.879	1.221	1.372	1.812	2.228	2.764	3.169	4.144
11	0.260	0.443	0.697	0.876	1.214	1.363	1.796	2.201	2.718	3.106	4.025
12	0.259	0.442	0.695	0.873	1.209	1.356	1.782	2.179	2.681	3.055	3.930
13	0.259	0.441	0.694	0.870	1.204	1.350	1.771	2.160	2.650	3.012	3.852
14	0.258	0.440	0.692	0.868	1.200	1.345	1.761	2.145	2.624	2.977	3.787
15	0.258	0.439	0.691	0.866	1.197	1.341	1.753	2.131	2.602	2.947	3.733
16	0.258	0.439	0.690	0.865	1.194	1.337	1.746	2.120	2.583	2.921	3.686
17	0.257	0.438	0.689	0.863	1.191	1.333	1.740	2.110	2.567	2.898	3.646
18	0.257	0.438	0.688	0.862	1.189	1.330	1.734	2.101	2.552	2.878	3.610
19	0.257	0.438	0.688	0.861	1.187	1.328	1.729	2.093	2.539	2.861	3.579
20	0.257	0.437	0.687	0.860	1.185	1.325	1.725	2.086	2.528	2.845	3.552
21	0.257	0.437	0.686	0.859	1.183	1.323	1.721	2.080	2.518	2.831	3.527
22	0.256	0.437	0.686	0.858	1.182	1.321	1.717	2.074	2.508	2.819	3.505
23	0.256	0.436	0.685	0.858	1.180	1.319	1.714	2.069	2.500	2.807	3.485
24	0.256	0.436	0.685	0.857	1.179	1.318	1.711	2.064	2.492	2.797	3.467
25	0.256	0.436	0.684	0.856	1.178	1.316	1.708	2.060	2.485	2.787	3.450
26	0.256	0.436	0.684	0.856	1.177	1.315	1.706	2.056	2.479	2.779	3.435
27	0.256	0.435	0.684	0.855	1.176	1.314	1.703	2.052	2.473	2.771	3.421
28	0.256	0.435	0.683	0.855	1.175	1.313	1.701	2.048	2.467	2.763	3.408
29	0.256	0.435	0.683	0.854	1.174	1.311	1.699	2.045	2.462	2.756	3.396
30	0.256	0.435	0.683	0.854	1.173	1.310	1.697	2.042	2.457	2.750	3.385
35	0.255	0.434	0.682	0.852	1.170	1.306	1.690	2.030	2.438	2.724	3.340
40	0.255	0.434	0.681	0.851	1.167	1.303	1.684	2.021	2.423	2.704	3.307
45	0.255	0.434	0.680	0.850	1.165	1.301	1.679	2.014	2.412	2.690	3.281
50	0.255	0.433	0.679	0.849	1.164	1.299	1.676	2.009	2.403	2.678	3.261
55	0.255	0.433	0.679	0.848	1.163	1.297	1.673	2.004	2.396	2.668	3.245
60	0.254	0.433	0.679	0.848	1.162	1.296	1.671	2.000	2.390	2.660	3.232
∞	0.253	0.431	0.674	0.842	1.150	1.282	1.645	1.960	2.326	2.576	3.090

Table 6.3: Valores Críticos de la Distribución F con $\alpha = 0.05$

p N-K	1	2	3	4	5	6	7	8	9	10
10	4.965	4.103	3.708	3.478	3.326	3.217	3.135	3.072	3.020	2.978
20	4.351	3.493	3.098	2.866	2.711	2.599	2.514	2.447	2.393	2.348
30	4.171	3.316	2.922	2.690	2.534	2.421	2.334	2.266	2.211	2.165
40	4.085	3.232	2.839	2.606	2.449	2.336	2.249	2.180	2.124	2.077
50	4.034	3.183	2.790	2.557	2.400	2.286	2.199	2.130	2.073	2.026
60	4.001	3.150	2.758	2.525	2.368	2.254	2.167	2.097	2.040	1.993
70	3.978	3.128	2.736	2.503	2.346	2.231	2.143	2.074	2.017	1.969
80	3.960	3.111	2.719	2.486	2.329	2.214	2.126	2.056	1.999	1.951
90	3.947	3.098	2.706	2.473	2.316	2.201	2.113	2.043	1.986	1.938
100	3.936	3.087	2.696	2.463	2.305	2.191	2.103	2.032	1.975	1.927
125	3.917	3.069	2.677	2.444	2.287	2.172	2.084	2.013	1.956	1.907
150	3.904	3.056	2.665	2.432	2.274	2.160	2.071	2.001	1.943	1.894
175	3.895	3.048	2.656	2.423	2.266	2.151	2.062	1.992	1.934	1.885
200	3.888	3.041	2.650	2.417	2.259	2.144	2.056	1.985	1.927	1.878
225	3.883	3.036	2.645	2.412	2.254	2.139	2.050	1.980	1.922	1.873
250	3.879	3.032	2.641	2.408	2.250	2.135	2.046	1.976	1.917	1.869
275	3.875	3.029	2.637	2.404	2.247	2.132	2.043	1.972	1.914	1.865
300	3.873	3.026	2.635	2.402	2.244	2.129	2.040	1.969	1.911	1.862
325	3.870	3.024	2.632	2.399	2.242	2.127	2.038	1.967	1.909	1.860
350	3.868	3.022	2.630	2.397	2.240	2.125	2.036	1.965	1.907	1.858
375	3.866	3.020	2.629	2.396	2.238	2.123	2.034	1.963	1.905	1.856
400	3.865	3.018	2.627	2.394	2.237	2.121	2.032	1.962	1.903	1.854
425	3.863	3.017	2.626	2.393	2.235	2.120	2.031	1.960	1.902	1.853
450	3.862	3.016	2.625	2.392	2.234	2.119	2.030	1.959	1.901	1.852
475	3.861	3.015	2.624	2.391	2.233	2.118	2.029	1.958	1.900	1.851
500	3.860	3.014	2.623	2.390	2.232	2.117	2.028	1.957	1.899	1.850
550	3.858	3.012	2.621	2.388	2.230	2.115	2.026	1.955	1.897	1.848
600	3.857	3.011	2.620	2.387	2.229	2.114	2.025	1.954	1.895	1.846
650	3.856	3.010	2.619	2.386	2.228	2.113	2.024	1.953	1.894	1.845
700	3.855	3.009	2.618	2.385	2.227	2.112	2.023	1.952	1.893	1.844
750	3.854	3.008	2.617	2.384	2.226	2.111	2.022	1.951	1.892	1.843
800	3.853	3.007	2.616	2.383	2.225	2.110	2.021	1.950	1.892	1.843
850	3.852	3.006	2.615	2.382	2.225	2.109	2.020	1.949	1.891	1.842
900	3.852	3.006	2.615	2.382	2.224	2.109	2.020	1.949	1.890	1.841
950	3.851	3.005	2.614	2.381	2.224	2.108	2.019	1.948	1.890	1.841
1000	3.851	3.005	2.614	2.381	2.223	2.108	2.019	1.948	1.889	1.840
∞	3.841	2.996	2.605	2.372	2.214	2.099	2.010	1.938	1.880	1.831