CAPÍTULO 1
INTRODUCCIÓN Y MOTIVACIÓN: ANTEPROYECTO DE INVESTIGACIÓN

God does not play dice in Universe
Albert Einstein (1879-1955)

God effectively plays dice in Universe, but he throws them where nobody is able to see
Niels Bohr (1885-1962)

1.1 RESUMEN CAPITULAR

En este capítulo se expone el soporte metodológico y conceptual del presente trabajo de tesis. Debe ser considerado como un marco de referencia para el lector, en el cual se vierten consideraciones de interés tales como los antecedentes y definición del problema, justificación, objetivos, presentación de los datos y puntos afines.

1.2 ANTECEDENTES Y BOSQUEJO HISTÓRICO

El problema de estimación es uno de los ejes centrales de la estadística moderna. De hecho, una gran porción de la literatura estadística existente se refiere a la solución de este tipo de problemas. Esto se debe en parte a que existe un gran interés práctico y teórico de la determinación de los valores de ciertos parámetros en algún fenómeno bajo estudio, así como sus propiedades de dispersión (puesto que esto permite determinar potenciales estructuras de causalidad entre las variables del fenómeno, con un entendimiento de mayor profundidad de su relación potencial que brinda un mayor poder operacional que permite, por ejemplo, pronosticar el comportamiento sobre la variable a explicarse).

Así, el problema de estimación emerge como un problema clave de la estadística. Usualmente, y sobre todo en problemas relativamente simples, lo que se busca es estimar un valor individual de un parámetro de interés. Sin embargo, y sobre todo en estudios más sofisticados, lo que se busca es estimar toda una serie de valores, o más

propiamente dicho, estimar la distribución de datos de algún cierto fenómeno. Así, por ejemplo, en un problema típico en la estadística oficial pudiera ser de interés el estimar la distribución de nacimientos en una cierta población, o estimar la distribución de opiniones acerca del comportamiento de alguna variable económica, por región. Esto es, en varios casos, más que estimar solo algunos estadísticos puntuales (como la media o la varianza), es de interés fundamental el estimar la función de distribución de toda una población.

Los ingresos de una persona, familia, entidad o un país en general no son la excepción. De hecho, este es un problema económico de vital importancia y todos los países con oficinas consolidadas de estadística llevan a cabo estudios periódicos que involucran la recopilación y análisis de variables de ingresos y gastos en los hogares. En el caso de México también se lleva a cabo este tipo de análisis mediante la llamada *"Encuesta Nacional de Ingresos y Gastos de los Hogares" (ENIGH)* cuya operación se encuentra a cargo del *Instituto Nacional de Estadística, Geografía e Informática (INEGI).*

Más aún, el tema es tan extremadamente relevante que se han llevado a cabo profundas investigaciones en diversos países y con diversas metodologías acerca de la distribución de los ingresos. Muchos de estos esfuerzos se han centrado en características descriptivas de la población de datos, aunque también se han realizado esfuerzos importantes en la parte inferencial del problema (véase el capítulo 2). Este gran esfuerzo se comprende fácilmente al considerar que una gran cantidad de políticas públicas basan sus acciones en función de dicha información para determinar, por ejemplo: las tasas tributarias y fiscales aplicables a diversos segmentos de la población según su poder adquisitivo; la adopción de políticas de abatimiento a la pobreza (la información de los ingresos es fundamental para la determinación de la llamada "línea de pobreza"); políticas de re-distribución de los ingresos; ciertos planes de proyección de egresos para ejercicios fiscales posteriores. En este orden ideas, en el capítulo 2 de este trabajo se expondrán las líneas principales de pensamiento e investigación que han merecido el interés de los principales investigadores en dicho campo a nivel mundial.

Las teorías de las distribuciones de los ingresos siempre han tenido fuertes motivaciones empíricas. Así, los primeros estudios acerca de la riqueza y el ingreso descubrieron una regularidad notable que se encuentra en todas las distribuciones conocidas de ingresos en grandes poblaciones:

- Las distribuciones de ingresos son siempre *sesgadas a la derecha*. Esto implica en particular que existe una gran cantidad de individuos con relativamente pocos ingresos, y un número reducido de individuos con ingresos sumamente altos (aunque de una gran importancia en cuanto a su aportación conjunta y relativa al ingreso nacional total).
- Sus funciones de densidad son *asimétricas* y tienen una *gran cola derecha* (lo cual se manifiesta diciendo que tienen "colas pesadas").
- Las distribuciones de los ingresos tienen medidas *positivas de sesgo* (tercer momento alrededor de la media).
- Las funciones de densidad también son *leptocúrticas* (cuarto momento positivo alrededor de la media). Esto significa en general que el grado en que se manifiesta el pico de este tipo de distribuciones es más fuerte que el de una normal asociada.

- Como ya se había comentado, las funciones de densidad tienen también una "cola pesada". Expuesto en términos ligeramente diferentes, esto significa que la media de los ingreso excede a su mediana, y los cuantiles superiores tienen una participación desproporcionadamente grande del total de los ingresos. La siguiente tabla proporciona la gráfica de deciles para los ingresos corrientes provenientes de la ENIGH 2004, así como sus porcentajes de participación en dicha variable.

Decil	**1**	**2**	**3**	**4**	**5**
Valor	6,531.89	9,882.59	12,924.67	16,115.50	19,800.01
Porcentaje	1.4	2.7	3.7	4.8	5.9
Decil	**6**	**7**	**8**	**9**	**10**
Valor	24,446.02	30,875.72	40,665.69	60,833.91	357,2218.50
Porcentaje	7.2	9.1	11.6	16.1	37.4

Cabe señalar que históricamente Pareto (1897) y Bowley (1915) fueron los pioneros en estudiar de la distribución de los ingresos personales desde un punto de vista empírico.

Se han desarrollado extensos análisis tanto macro como microecónomicos en base a diversas fuentes de datos desde la Segunda Guerra Mundial que han derivado en un entendimiento más completo del problema (y en el cual también han surgido, por supuesto, líneas de investigación muy interesantes). De hecho, una gran cantidad del trabajo se ha llevado a cabo en el sentido de tratar de formular una base estadística, económica y econométrica para ajustar formas funcionales específicas a las distribuciones empíricas de los ingresos. Así, es común el escuchar que la distribución de Pareto, la log-normal y otras son "las distribuciones" que mejor ajustan los datos de ingresos en ciertas oficinas nacionales de estadística.

A pesar de tales esfuerzos, el problema de la identificación de una estructura distribucional para los ingresos no es en absoluto trivial. Esto se debe a que existen una gran cantidad de factores que lo afectan. Por ejemplo, en México, los ingresos difieren fuertemente en las diversas categorías de agrupación, como lo son el tipo de trabajo, experiencia laboral, la edad, el sexo, las zonas de residencia (urbana, rural, semi-rural, etc.). A su vez, debido a que los datos asociados casi siempre son obtenidos mediante un proceso de muestreo, también debe de considerarse su error asociado. Una tercera consideración es el hecho de empatar los datos de los ingresos de la ENIGH con respecto los correspondientes de las oficinas de Cuentas Nacionales, situación que en sí misma merece un estudio a profundidad.

Las metodologías aplicadas al análisis de este problema cubren un amplio espectro, y para mencionar solo algunas de ellas se pueden considerar las siguientes: el ajuste por máxima verosimilitud a ciertas distribuciones clásicas, como la Pareto; la imputación de datos para aquellos registros que carecen de respuesta; el uso de registros administrativos y otras fuentes de información para obtener la información asociada; el uso de herramientas de simulación (tales como las estrategias de bootstrap y jackniffe) y empleo de métodos intensivos por computadora para obtener información acerca de los estimadores asociados y sus errores. En este sentido, existe una rama de la estadística conocida como la *"Teoría de Valores Extremos"* la cual es de gran interés tanto teórica como prácticamente puesto que cuenta con poderosos mecanismos de modelación y validación para analizar aquellos hechos que podrían catalogarse más

bien como "inusuales" más que los "usuales". De hecho, en este trabajo se empleará esta estrategia, además de algunas otras, para la analizar estadísticamente la cola derecha de la distribución de los ingresos en México (que por poseer ciertas características que serán analizadas posteriormente, sus observaciones pueden tipificarse precisamente como extremas).

Como complemento a lo anterior, y asociado con la considerable complejidad del problema, con cierta frecuencia las investigaciones centran su atención en algunos segmentos de la distribución de ingresos, más que en la estructura completa, empleado el principio básico "dividir para vencer". En este trabajo, de hecho, se toma en consideración esta óptica, y por ello se analizará básicamente la parte de la cola derecha de la distribución de ingresos, en cuanto sus máximos.

1.3 DEFINICIÓN DEL PROBLEMA

El problema a estudiarse es el de conformar un análisis comparativo entre las distribuciones "clásicas" para el ajuste de datos de ingresos y la modelación de valores extremos, con la finalidad unificada de determinar las propiedades estructurales de la cola derecha de distribución de los máximos de los ingresos corrientes en México. Es decir, se aplicará la teoría clásica de ajustes a modelos que han sido ampliamente utilizados en otros países para modelar los ingresos extremos superiores en México, así como la teoría de valores extremos mediante el análisis de los datos provenientes de la ENIGH 2004 (en la parte de la "alcances del problema" de este capítulo se expone el motivo de por que fueron seleccionados para su análisis ciertos datos de esta encuesta). El producto final será un análisis comparativo entre las distribuciones clásicas y la modelación de valores extremos para plantear las bondades y deficiencias de cada modelo, y que ayude así a proporcionar un esquema que mejore la comprensión del fenómeno de la distribución de los ingresos en México.

1.4 JUSTIFICACIÓN

Ya se han comentó en la parte de los antecedentes de este trabajo que el problema de la estimación de distribuciones es de gran interés en una gran cantidad de estudios. Esto es especialmente cierto para el caso concreto de la determinación general de la distribución de ingresos en una economía nacional, y en particular a sus valores extremos. Esto es así, en parte, por los siguientes motivos:

- Existe siempre un interés teórico profundo en cuanto a la determinación de la distribución de variables de interés en algún fenómeno, puesto que al estimar la distribución de los ingresos podría conocerse, de manera detallada, el comportamiento de la misma y conocer, por ejemplo, características poblacionales de importancia como la media, la varianza, momentos de órdenes mayores, propiedades concretas de asimetría y curtosis de la distribución, entre otras.

- La estimación de los valores extremos de la distribución de ingresos es de gran importancia, puesto que es un hecho observado que esta cola es la que concentra la

mayor cantidad de los ingresos nacionales, por lo que el hecho de conocerla implica el poder conformar de mejor manera las políticas fiscales, monetarias y sociales (por ejemplo, en la determinación del nivel de tasas tributarias, un pronostico más adecuado para proyectar el nivel de ingresos y gastos de la nación, así como en un mejor diseño de políticas de re-distribución de la riqueza hacia los grupos menos favorecidos).

- Debido a que es un hecho empírico el que la cola derecha en la distribución de los ingresos es pesada, una correcta determinación de la distribución de los datos de esta cola implica un re-ajuste de la distribución completa de ingresos.

- El contar con una estimación fidedigna de la cola derecha permitiría contar potencialmente con herramientas predictivas de mayor precisión.

- El tener una correcta distribución de los ingresos permite realizar comparaciones con respecto a otros países y así determinar avances o retrasos en ciertas áreas de la economía nacional.

- Hasta el momento de escribir este trabajo no parece existir evidencia de que la teoría de valores extremos hubiese sido aplicada para analizar la cola derecha de la distribución de ingresos en México. Por este motivo, es de interés el contar con un análisis de la calidad de ajuste entre las distribuciones clásicas y la teoría de valores extremos con la finalidad de contar con un marco de comparación de las bondades y defectos relativos de cada estrategia.

1.5 HIPÓTESIS

Sea F la distribución de ingresos corrientes de la cola derecha de México para el año de 2004. Sea $\widehat{F}$ la correspondiente función estimada mediante la Teoría de Valores Extremos (véase capítulo 3). A un cierto nivel de confianza, previamente establecido, el juego de hipótesis que se analizará en este trabajo es el siguiente:

- $H_0 : F = \widehat{F},$
- $H_1 : F \neq \widehat{F}.$

Expuesto en términos equivalentes, en este trabajo se busca estudiar el nivel de bondad ajuste de la distribución de ingresos mediante la aplicación de la teoría de valores extremos con respecto a los modelos "clásicos" de ajuste. A su vez, también será llevado a cabo el análisis del ajuste del modelo beta generalizado de segundo tipo y del modelo lognormal, con la finalidad de contar con los ya comentados elementos de comparación entre las diversas distribuciones.

1.6 OBJETIVOS

El propósito fundamental que tiene el presente estudio es llevar a cabo un análisis comparativo del nivel de ajuste de la cola derecha de la distribución de ingresos, mediante el uso de la distribución beta generaliza del segundo tipo, la lognormal y los modelos de la teoría de valores extremos.

A su vez, se busca alcanzar los siguientes objetivos secundarios:

- Realizar un análisis exhaustivo de la literatura asociada con el problema para vislumbrar el nivel del estado del arte del problema.

- Exponer de manera breve y completa la teoría de valores extremos para una audiencia cuyos intereses se centran en el análisis de datos de estadística oficial, y que pudieran ser de utilidad para la aplicación en sus problemas particulares.

- Presentar las virtudes y limitaciones de la metodología de los valores extremos en cuanto al análisis de datos de la distribución de los máximos de los ingresos.

- Describir potenciales líneas de investigación que pudieran derivarse de una continuación del presente trabajo.

1.7 ALCANCES DEL PROBLEMA

Asociado con la definición y los objetivos del problema planteados, es conveniente ahora el establecer los alcances del problema.

- Los datos a analizarse son provenientes de la ENIGH 2004. Esta selección se debió a que dicha información no presenta un elapse demasiado aletargado con respecto al momento actual, de manera tal que las conclusiones sean aún válidas. A su vez, como la información de las variables económicas de la ENIGH se presenta a precios corrientes, para no utilizar un deflactor que los lleve a precios constantes, y asociado con la relativamente baja tasa de inflación en el periodo 2004-2007, se consideró pertinente su adopción. Más aún, la ENIGH 2004 presentó un incremento sustantivo de su tamaño de muestra con respecto a su diseño inicial (para contar con un mayor grado de cobertura que solicitaron varios estados y algunas entidades gubernamentales). Esto último hace que se cuente con una mayor cantidad de datos, con los cuales puede tenerse un mejor análisis. Finalmente, la ENIGH 2004 ha sido ampliamente revisada, y algunos posibles errores de captura han sido corregidos mediante procesos de validación adecuados. Cabe señalar, sin embargo, que los procedimientos y métodos expuestos en este trabajo podrían aplicarse, con poco o ningún cambio, a los datos de las distintas ENIGH que se han realizado.

- En este trabajo se analiza únicamente la parte derecha de la distribución de los ingresos. No se realiza el análisis para los datos de la cola izquierda debido a que el interés en este trabajo, como se expuso en la sección de la justificación, reside en el otro extremo. Nuevamente, sin embargo, casi sin cambios de fondo se podría

aplicar la metodología de la parte derecha a la estimación de la distribución de los mínimos de la cola izquierda (aunque en ese caso los objetivos planteados deberían de re-estructurarse en función de determinar con mayor precisión, por ejemplo, la línea de pobreza).

1.8 PRESENTACIÓN DE LOS DATOS

En esta última sección se exponen algunos hechos relevantes con respecto a los datos de la ENIGH 2004 que son de interés para contar con un mayor entendimiento del problema por resolver (algunos de los resultados más técnicos y de índole metodológica se exponen en los dos primeros apéndices al final de este documento). Para ubicar con mayor simplicidad el significado de algunos términos del ingreso, el lector puede referirse al Anexo "A" que constituye un extracto del glosario de la ENIGH. Además, si busca ampliar un poco más su conocimiento sobre cuestiones metodológicas acerca de tal encuesta, el lector se puede referir al Anexo "B".

La Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) correspondiente al tercer trimestre de 2004, consideró un diseño con cobertura a nivel nacional que además permitiera el desglose de la información por localidades de 2,500 y más habitantes y de menos de 2,500 habitantes, así como para cinco estratos de acuerdo con su nivel de marginación. El tamaño final de muestra fue de 25,115 viviendas.

Los datos que muestra la ENIGH-2004 son comparables con las realizadas por el INEGI en el tercer trimestre de 1984, 1989, 1992, 1994, 1996, 1998, 2000 y 2002. En los ocho proyectos se utilizó el mismo marco conceptual e instrumentos de captación; además, el levantamiento se realizó en las mismas fechas y con igual procedimiento de recolección.

Las encuestas de ingresos y gastos de los hogares están basadas en la consideración de que el monto del ingreso, su procedencia y su forma de distribución condiciona, en gran medida, el nivel de bienestar de la población, puesto que es el ingreso el que determina la capacidad económica de los hogares para adquirir los bienes y servicios necesarios.

Para abordar el estudio del monto, la procedencia y la distribución del ingreso y el gasto de los hogares, se seleccionaron viviendas particulares como unidad de muestreo y la misma vivienda, así como los hogares y sus integrantes como unidades de observación y análisis.

Por otra parte, es necesario observar las características sociodemográficas de los integrantes del hogar como causa-efecto de los patrones de distribución del ingreso; así como la relación que guardan las condiciones en las que los individuos se incorporan al proceso productivo, mediante la captación de las características de la ocupación que genera al menos parte de sus ingresos y, por otro lado, las características de la vivienda y el equipamiento del hogar para conocer las condiciones de vida de los integrantes del hogar.

El marco conceptual de las ENIGH está basado en las recomendaciones internacionales de las Naciones Unidas y la Organización Internacional del Trabajo y está articulado al Sistema de Cuentas Nacionales y a otras encuestas de hogares que levanta el INEGI.
De acuerdo con su propósito de comparabilidad con ENIGH'S anteriores (1984, 1989, 1992, 1994, 1996, 1998, 2000 y 2002), la encuesta es diseñada para presentar información a nivel nacional, para localidades de 2 500 y más habitantes, para las de menos de 2 500 habitantes, así como para cinco estratos de acuerdo con su nivel de marginación.

El levantamiento de la información se realizó del 10 de agosto al 17 de noviembre, en un periodo que se hizo coincidir para las ENIGH anteriores, con el propósito de poder ejecutar la comparación de los resultados sin que éstos se vean afectados por variaciones estacionales del ingreso o del gasto de los hogares.

La recolección de la información se llevó a cabo por medio de visitas a las viviendas seleccionadas, se utilizaron instrumentos de captación especializados que permitieron la operacionalización del marco de conceptos de la encuesta; un equipo de entrevistadores y supervisores capacitados de manera exhaustiva, durante un mes, sobre los procedimientos, lineamientos y criterios establecidos en la misma.

Las unidades que conforman la muestra fueron seleccionadas con criterios probabilísticos, con el propósito de asegurar que a partir de los resultados, pudieran estimarse los indicadores correspondientes para todos los hogares.

El tamaño de la muestra nacional fue de 25,115 viviendas, lo cual garantiza que las estimaciones de cada uno de los indicadores de interés tengan una calidad aceptable.

En general, el esquema de muestreo es estratificado y bietápico, porque las unidades de análisis que se incluyen en la muestra fueron seleccionadas mediante dos etapas sucesivas. En la primera eligen grupos de viviendas y en la segunda, selecciona directamente la vivienda.

El procedimiento de selección de las unidades de muestreo, el procedimiento de estimación y el tamaño de la muestra de la ENIGH-2004, permiten estimar distribuciones porcentuales con precisión y confianza.

En el caso de estimar totales, deberá tomarse en cuenta la longitud del intervalo de confianza correspondiente. Posteriormente, durante el proceso de captura-validación, actividad que consiste en verificar y garantizar que la información captada esté completa, se realizó en dos etapas:

- **Primera etapa:** se procede a la captura de los datos que provienen de campo, cuidando su integridad y confiabilidad.
- **Segunda etapa:** los datos se validan electrónicamente para depurar las incongruencias y omisiones que pudiera contener el archivo base de datos.

La conformación de los archivos se inició con la captura de los cuestionarios, donde se desarrolló un procedimiento de captura con filtros de congruencia e integridad para sólo registrar los cuestionarios correctamente requisitados.

La validación del archivo se realizó en dos etapas; en la primera se llevó a cabo la emisión de distribución de frecuencias por código, y en la segunda, la validación de estas frecuencias haciendo las correcciones necesarias con base en la información original de los cuestionarios.

Con la información capturada y validada fue posible construir la base de datos, verificar la integridad del archivo, corregir los factores de expansión por la no respuesta, generar los tabulados básicos y los resultados de la construcción de los estratos.

Dentro de las múltiples bases de datos que se generan dentro de la ENIGH, la que será analizada en el presente trabajo es la de "concentrados", en la cual se resumen por cada folio (hogar) las características de 117 variables.

Con esta exposición se da por terminado el primer capítulo, pasando ahora a un análisis de la literatura existente con respecto al problema.

CAPÍTULO 2

REVISIÓN DE LA LITERATURA ACERCA DE LA DETERMINACIÓN DE LA DISTRIBUCIÓN DE INGRESOS

There is nothing in the Universe that lacks any maximization or minimization rule

Leonard Euler (1707-1783)

2.1 RESUMEN CAPITULAR

En el presente capítulo se realiza un análisis exhaustivo de la literatura asociada con la determinación de la estructura de la distribución de ingresos. El propósito básico es doble: por una parte, se pretende presentar al lector las estrategias y metodologías enmarcadas en la frontera del estado del arte del problema, y por la otra, justificar la aplicación de la teoría de valores extremos en la determinación de la distribución de los máximos de los ingresos (en el sentido de que el enfoque de la aplicación es original).

2.2 INTRODUCCIÓN

De manera general, los artículos, libros y literatura estadística asociada con el problema de la determinación de la estructura de los ingresos, se encuentra inmerso en alguna de las siguientes grandes categorías:

- **Modelos estocásticos.** Este tipo de modelos comienzan usualmente su análisis mediante la adopción de ciertos supuestos distribucionales acerca del comportamiento de los datos, y mediante un proceso de observación se trata de proponer una estructura estocástica que pudiera ser seguida por ellos. Este tipo de modelos se puede considerar como "a posteriori" en el sentido de que se supone que las observaciones ya se encuentran dadas y lo que se busca es estimar la mejor distribución que las ajuste.

- **Modelos de selección.** Estos también comienzan suponiendo ciertos hechos asociados con la distribución de los ingresos, pero a diferencia de los estocásticos, se supone que las decisiones de los trabajadores en cuanto a la asignación de sus habilidades en ciertos tipos de trabajos es la que genera una distribución de los ingresos. Se podría considerar a este tipo de modelos como "a priori", en el sentido de que las observaciones de los ingresos se suponen condicionadas a las decisiones de los trabajadores.

- **Teorías de capital humano.** Estos modelos contribuyen al entendimiento de como es que los trabajadores adquieren sus habilidades. Esto demuestra que la desigualdad de los ingresos es necesaria en una economía donde algunas actividades requieren mayor inversión de habilidades que otras. Estos modelos ilustran como los recursos familiares y los talentos naturales afectan las inversiones de las habilidades que generan la desigualdad de los ingresos antes mencionada.

- **Teoría de las agencias.** Esta teoría trata de establecer modelos que describan también las situaciones de la desigualdad de los ingresos, pero desde una óptica enteramente distinta. En lugar de describir como las decisiones de los trabajadores generan una distribución de los ingresos, la teoría de las agencias intentan describir como las compañías, industrias, firmas, gobiernos y demás empleadores llegan a seleccionar una distribución de ingresos para llegar a contar con los niveles deseados de productividad individual que desean. Así, las distribuciones de los sueldos altamente sesgadas podría ser un instrumento de incentivo para captar al personal más calificado.

Como cabe observar dentro de esta tipología, las tres últimas clasificaciones, a pesar de ser altamente atractivas para un análisis detallado, se centran más bien en cuestiones de análisis social, económico, demográfico o incluso psicológico. Debido al giro que tiene el presente trabajo, será de interés únicamente la primera categoría, la cual a su vez será disgregada en los siguientes apartados.

2.3 MODELOS ESTOCÁSTICOS DE DISTRIBUCIÓN DE LOS INGRESOS

Por lo ya comentado en el primer capítulo de este documento, el hecho de estimar la distribución de los ingresos ha sido un problema de gran atractivo tanto a los estadísticos como a los investigadores sociales. No es de sorprender que la literatura asociada con este tema sea abundante. Dicha situación presenta por supuesto beneficios y perjuicios: un beneficio claro es que muchas aristas del problema ya han sido investigadas, y de hecho, como se mostrará más adelante, han surgido una gran cantidad de propuestas metodológicas para resolver el problema de modelar los ingresos; un punto desfavorable de importancia es que para lograr estar en el estado del arte del problema de la distribución de ingresos, es necesario llevar a cabo estudios muy voluminosos de los trabajos desarrollados, que involucran necesariamente la inversión de grandes cantidades de recursos.

Los modelos estocásticos que se han aplicado en el análisis de la distribución de los ingresos pueden catalogarse a su vez en alguna de las siguientes categorías:

- Modelos previamente diseñados para algún otro problema, y que se adecuan al problema de la distribución de los ingresos.
- Modelos generados "ex profeso" para la distribución de los ingresos.
- Modelos estocásticos que buscan más que explicar la distribución de los ingresos, demostrar ciertos hechos de economía normativa en situaciones de pobreza, desigualdad e inequidad. Esto es, dichos modelos no tienen la finalidad de determinar la distribución de los ingresos, sino coadyuvar a la afirmación o negación de enunciados de economía normativa tales como: "La desigualdad en México se ha incrementado".

En los siguientes tres apartados se discutirán a profundidad cada una de estas tres grandes vertientes de los modelos estocásticos.

Es importante señalar que las referencias que serán dadas en estos apartados son una muestra que trató de realizarse lo más selectamente posible para no sobrecargar el número de artículos que el potencial lector de este trabajo pudiera consultar, y que se muestran en las fuentes de información del capítulo 2. Esto es, las referencias que se presentan no son por supuesto exhaustivas aunque se pretende que sean representativas del tema.

2.4 MODELOS ESTOCÁSTICOS DE DISEÑO PREVIO

Como ya se ha comentado en el apartado anterior, diversos autores han propuesto el uso de distribuciones conocidas para su ajuste al problema de la distribución de los ingresos. La estrategia básica que se sigue en dichos modelos es:

- Contar con la base de datos otorgada por alguna agencia de estadística oficial nacional.

- Proponer el modelo estocástico que pudiera ser razonable en cuanto al ajuste de dichas observaciones.
- Estimar los parámetros del modelo propuesto.
- Realizar pruebas de bondad de ajuste de los parámetros estimados, que usualmente son de tipo gráficas.
- Si es de interés, pronosticar y contrastar la validez del modelo contra de estudios posteriores de las oficinas nacionales de estadística.

En la tabla 2.1 se muestran algunas formas funcionales explícitas que se han aplicado al problema de la estimación de los ingresos o afines, así como su función de densidad de probabilidad y el dominio de sus parámetros.

Tabla 2.1 *Tabla de modelos de diseño previo, sus funciones de densidad de probabilidad y dominios de los parámetros*

		Función de densidad de probabilidad (pdf)	**Datos y restricciones de los parámetros**
Distribuciones biparamétricas	Lognormal	$f(x;\mu,\sigma)=\frac{1}{(x-\min)\sqrt{2\pi\sigma^2}}\exp\left[-\frac{[\ln(x-\min)-\mu]^2}{2\sigma^2}\right]$	$x > 0$
	Gamma	$f(x;\alpha,\beta)=\frac{(x-\min)^{\alpha-1}}{\beta^{\alpha}\Gamma(\alpha)}\exp\left(-\frac{x-\min)}{\beta}\right)$	$x>0;\ \alpha>0;\ \beta>0$ $\Gamma(\alpha)=$ función gamma
	Beta	$f(x;p,q)=\frac{1}{\mathrm{B}(p,q)}\frac{(x-\min)^{p-1}(\max-x)^{q-1}}{(\max-\min)^{p+q-1}}$	$x\min < x < x\max;\ p>0$ $B(p,q)=\int_{x\min}^{x\max}\frac{(x-x\min)^{p-1}(x-x\max)^{q-1}}{(x\max-x\min)^{p+q-1}}dx$
		o $f(y;p,q)=\frac{1}{B(p,q)}y^{p-1}(1-y)I_{(0,1)}(y)$	$0<y<1;\ p>0;$ $q>0$ $B(p,q)=\int_0^1 x^{p-1}(1-x)^{q-1}dx$
	Champer-nowne	$f(x;\mu,\theta)=\theta\frac{\mu^{\theta}x^{\theta-1}}{(\mu^{\theta}+x^{\theta})^2}$	$x>0;\ \mu>0;\ \theta>0$
Distribuciones triparamétricas	Lognormal desplazada	$f(x;\lambda,\mu,\sigma)=\frac{1}{(x-\lambda)\sqrt{2\pi\sigma^2}}\exp\left[-\frac{[\ln(x-\lambda)-\mu]^2}{2\sigma^2}\right]$	$x>\lambda$
	Singh-Maddala	$f(x;a,b,q)=\frac{qax^{a-1}}{b^a\left[1+\left(\frac{x}{b}\right)^a\right]^{1+q}}$	$x>0;\ a>0;\ \beta>0;$ $q>1/a$
	Dagum	$f(x;\beta,\lambda,\delta)=\beta\lambda\delta x^{-\delta-1}(1+\lambda x^{-\delta})^{-\beta-1}$	$x>0;\ \beta>0;\ \lambda>0;\ \delta>0$

2.5 MODELOS ESTOCÁSTICOS EX PROFESO PARA LA DISTRIBUCIÓN DE INGRESOS

Como se había comentado con anterioridad, y debido a la importancia del problema de la estimación de la distribución de los ingresos, se han propuesto una gran diversidad de enfoques que pueden conceptualizarse como métodos ex profeso para este problema, aunque algunos de ellos surgieron simultáneamente a otras aplicaciones físicas de estos modelos, como el Weibull, que también tiene áreas de aplicación muy importantes en ingeniería. A continuación se comienza ha realizar un recuento cronológico de las distribuciones que han sido propuestos para analizar la problemática.

De manera histórica, la primera propuesta de una distribución que modelara los ingresos se debió a Wilfredo Pareto en 1895. El análisis de Pareto comenzó mediante un análisis de la desigualdad de los ingresos (que se basaba en un modelo previamente diseñado por él y con datos que había recopilado para tales fines) iniciando así el debate del efecto del crecimiento económico sobre la desigualdad de los ingresos. En términos actuales, la pregunta que él se realizó podría estructurarse como: ¿son los ricos más ricos y los pobres más pobres, o las fluctuaciones económicas afectan por igual a todos? Corrado Gini no estuvo de acuerdo con la opinión de Pareto acerca de que el desarrollo económico implica mayor desigualdad. Para estudiar de mejor manera el fenómeno, él propuso una unidad de medida adimensional, conocida precisamente como el "índice de Gini", el cual sigue utilizando de manera extensiva.

Estudios posteriores mostraron que la distribución de Pareto modela de manera adecuada los niveles altos del ingreso, pero proporcionan un nivel pobre de precisión en el extremo izquierdo de la distribución. A medida que la investigación continuo, se propusieron nuevas distribuciones que se ajustaban mejor a los datos observados. Por ejemplo, Gibrat (1931) escribió un artículo famoso en el cual se sugería la distribución lognormal con dos parámetros, que fue nuevamente examinada por Aitchison y Brown (véase ref. [3]). Ammon (1895) propuso la distribución gamma, la cual se reintrodujo recientemente para ajustar los datos de los ingresos de Estados Unidos por Salem & Mount (1974). Barlets y Van Metelel (1975) sugirieron la utilización de otra función de densidad de probabilidad, la Weibull.

Aún mejores ajustes pueden realizarse utilizando distribuciones de tres parámetros. Estas incluyen la función gamma generalizada (Amoroso, 1924-25 y Taille, 1981) y beta (Thurow, 1970) así como dos modelos estrechamente relacionados los cuales son miembros de la familia Burr de distribuciones: la Singh-Maddala (1976), conocida en la literatura como la Burr 12, y la Dagum (1977), conocida como la Burr 3. Estas distribuciones permiten la intersección de las curvas de Lorenz, un fenómeno observado en ciertos datos que no pueden ser modelados por ninguna distribución biparamétrica.

McDonald (1984) introdujo la función beta generalizada del primer y segundo tipo (GB1 y GB2), dos distribuciones tetra-paramétricas las cuales no fueron solamente muy exitosas al ajustar los datos, sino que también incluían todas las mencionadas previamente como casos límite o especiales. McDonald y Mantrala (1996) encontraron que la distribución GB2 proporciona un ajuste significativamente mejor que su

distribución anidada cuando se ajustan los datos de ingresos de los Estados Unidos (por este motivo se seleccionó a esta distribución como un modelo de ajuste razonable a los datos de los ingresos a llevarse a cabo en el capítulo 4). El éxito empírico de la distribución GB2 fue complementada por el modelo teórico de la generación de ingresos de Parker (1999), mostrando ingresos que siguen la distribución GB2.

Se ha prestado una gran cantidad de tiempo y esfuerzo en el desarrollo de hipótesis económicas y econométricas que expliquen el comportamiento de los ingresos. Por ejemplo, Bresciani-Turroni (1937) llegaron a proporcionar una explicación exhaustiva acerca de la famosa ley de Pareto de los ingresos.

A su vez, también se ha tratado de vislumbrar la solución al problema mediante la aplicación de funciones acumuladas de frecuencias. Por ejemplo, Irving W. Burr (1942) escribió un artículo famoso en el cual se expone precisamente la función Burr y en la cual se hace un uso extenso de las propiedades de la función de probabilidad acumulada. En las siguientes subsecciones se mencionarán los principales avances de lo expuesto hasta aquí, señalando los posibles beneficios y desventajas de algunas de ellas.

2.5.1 Modelos paramétricos de tipo Beta Generalizados de distribución de ingresos

McDonald y Xu (1995) mostraron que las distribuciones mencionadas en el apartado anterior pueden ser consideradas como casos especiales o límites de una función penta-paramétrica, la cual es muy flexible, y que se ha denominado **modelo beta generalizado**, cuya función de densidad de probabilidad puede escribirse de la siguiente forma:

$$\mathrm{GB}(y;a,b,c,p,q)=\frac{|a|\,y^{ap-1}[1-(1-c)(y/b)^a]^{q-1}}{b^{ap}B(p,q)[1+c(y/b)^a]^{p+q}} \quad \text{para} \quad 0<y^a<b^a/(1-c)$$

y cero en otro caso, donde $0\le c\le 1$, y $b,p,q>0$. Utilizando esta parametrización, es fácil mostrar varias relaciones entre la familia de distribuciones beta generalizada. La GB2 puede ser derivada fijando el parámetro c igual a 1.

$$\mathrm{GB2}(y;a,b,p,q)=\frac{|a|\,y^{ap-1}[1-(y/b)^a]^{q-1}}{b^{ap}B(p,q)}=\mathrm{GB}(y;a,b,c=1,p,q).$$

Las distribuciones triparamétricas Dagum y Singh-Maddala corresponden a los casos siguientes:

$$\mathrm{DAGUM}(y;a,b,p)=\mathrm{GB2}(y;a,b,p,q=1),$$

$$\mathrm{SM}(y;a,b,q)=\mathrm{GB2}(y;a,b,p=1,q).$$

La distribución gamma generalizada (GG) puede verse como un caso límite definido como:

$$\mathrm{GG}(y;a,\beta,p)=\lim_{q\to\infty}\mathrm{GB}(y;a,b=q^{1/a}\beta,c=1,p,q).$$

Una forma conveniente de visualizar estas relaciones se muestra en el árbol de la figura 2.1.

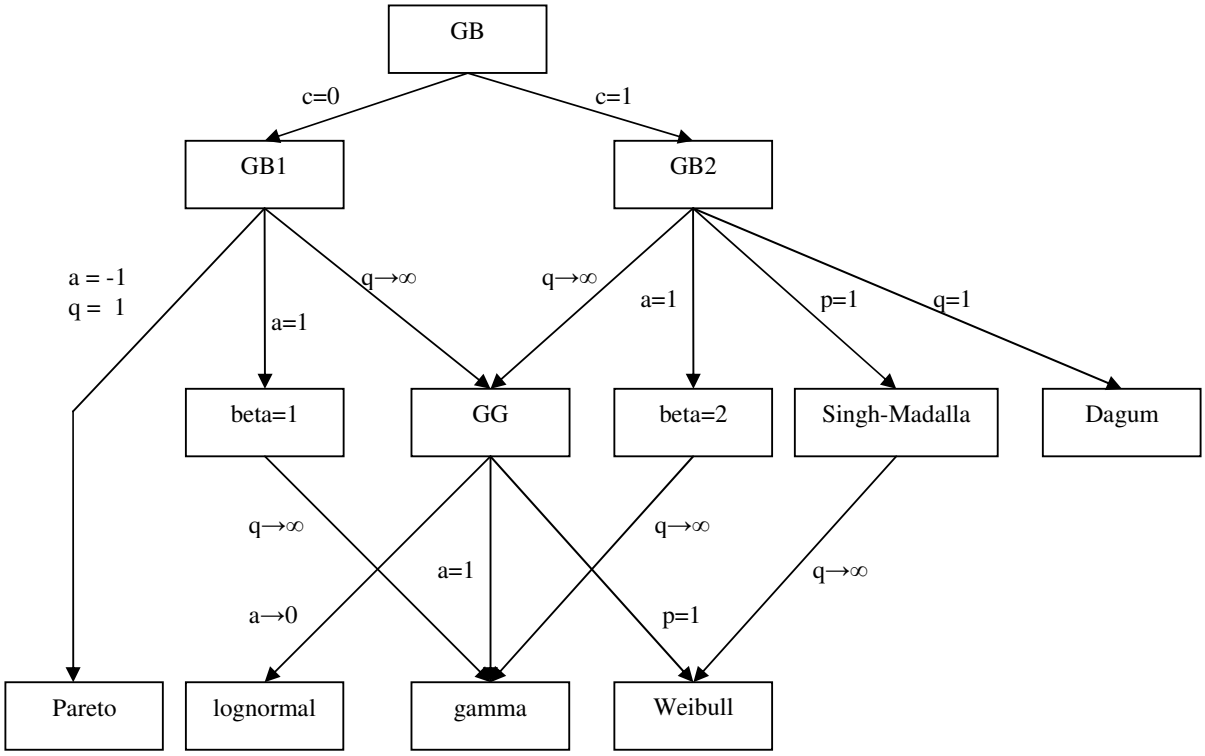

Figura 2.1 *Árbol de relaciones de distribuciones de ingresos*

En la tabla 2.2 se proporciona los desempeños empíricos relativos de algunos de estos modelos, para el ajuste a datos de ingresos de los Estados Unidos de América.

Tabla 2.2 *Tabla de desempeños relativos de algunos modelos de ingresos*

Modelo (distribución)	No. de parámetros	Desempeño a los datos de Estados Unidos de América
Pareto	2	Proporciona un ajuste excelente a la cola superior de la distribución de los ingresos, pero el ajuste al conjunto completo de datos del ingreso es pobre (véase Mandelbrot, 1960)
Lognormal	2	Ajuste bien sobre un gran parte del rango medio de los ingresos, pero da un ajuste pobre en las colas (Aitchison & Brown, 1957). Sin embargo, en el rango medio del medio se exagera la curtosis.

Gamma	2	Proporciona un mejor ajuste que la lognormal en las dos colas. En el rango medio, tanto la lognormal como la gama exageran la curtosis, pero esta tendencia es más marcada en el caso de la lognormal. Evaluando la bondad de ajuste general, la evidencia empírica favorece a la gamma sobre la lognormal (Salem & Mount, 1974).
Singh-Maddala (incluye las distribuciones de Pareto –tipo II–, Weibull y la de Fisk, como casos especiales)	3	En términos de la bondad de ajuste, este modelo es superior tanto al lognormal como a la gamma (Singh-Maddala, 1976).
Dagum	4	Tiene un mejor desempeño que la lognormal, gamma y Singh-Maddala (Dagum, 1977).
Beta generalizada II (incluye la distribuciones gamma generalizada, beta II, Singh-Maddala y la F como casos especiales)	4	Tiene una mejor desempeño que todos sus casos especiales, así como la beta generalizada I (McDonal, 1984).

Ahora bien, en cuanto a la estimación de parámetros en la utilización de estas técnicas, es común agrupar los datos en g clases, de forma tal que se busca maximizar alguna función de verosimilitud. A su vez, para evaluar la eficiencia relativa de las distribuciones, se acostumbra evaluar la suma de los errores cuadráticos, la suma de los errores absolutos y encontrar las medidas de bondad de ajuste tipo χ^2, así como el valor de la log-verosimilitud del modelo estimado y del conjunto de datos. De manera específica, se busca maximizar la log-verosimilitud

$$\ell(\theta) = \sum_{i=1}^{N} \ln[f_d(y_i;\theta)]$$

$$\ell(\theta) = \ln(N!) + \sum_{i=1}^{g} \{n_i \ln[p_i(\theta)] - \ln(n!)\}$$

donde:

- $p_i(\theta) = F_d(Y_i:\theta) - F_d(Y_{i-1}:\theta)$, $f_d(\,)$ y $F_d(\,)$, denotan la función de densidad de probabilidad y la función de distribución acumulada para la distribución de tipo d;
- θ es un vector que contiene los parámetros distribucionales;
- Y_i y Y_{i-1} son las cotas superiores del $i-$ésimo de g grupos de datos;
- n_i es el número de observaciones en el $i-$ésimo grupo;
- N es el total de observaciones.

Los métodos de optimización numérica, los cuales son necesarios para estimar los parámetros desconocidos, utilizan aplicaciones repetidas de los algoritmos de búsqueda.

Ahora, el estadístico de prueba para la razón de verosimilitud se define como:

$$LR = 2\left[\hat{\ell} - \hat{\ell}^{*}\right] \sim \chi^2(r),$$

y puede ser utilizado para comparar distribuciones anidadas donde $\hat{\ell}$ y $\hat{\ell}^{*}$ representan, respectivamente, los valores de la log-verosimilitud correspondiente a los modelos sin restricciones y anidado y r (los grados de libertad de la chi-cuadrada asintótica) es la diferencia en el número de parámetros estimados en las dos especificaciones de modelos. Así, el mejor estadístico de GB2 con respecto a la distribución de Dagum puede ser probada utilizando una distribución chi-cuadrada con un grado de libertad. Los modelos anidados sobre la frontera del espacio paramétrico podría comprometer la adecuabilidad de $\chi^2(r)$. La suma de los errores cuadráticos (SSE), la suma de los errores absolutos (SAE) y la medida de la bondad de ajuste $\chi^2(r)$ tienen las siguientes expresiones:

$$\text{SSE} = \sum_{i=1}^{g}\left(\frac{n_i}{N} - p_i(\hat{\theta})\right)^2,$$

$$\text{SAE} = \sum_{i=1}^{g}\left|\frac{n_i}{N} - p_i(\hat{\theta})\right|, \quad \text{y}$$

$$\chi^2 = N\sum_{i=1}^{g}\left[\left(\frac{n_i}{N} - p_i(\hat{\theta})\right) \div p_i(\hat{\theta})\right].$$

donde $\hat{\theta}$ denota el vector de parámetros estimados. La χ^2 está distribuida asintóticamente como una chi-cuadrada con grados de libertad igual a uno menos que la diferencia entre el número de grupos y el número de parámetros estimados. Como es conocido, esta prueba puede aplicarse de manera razonable sólo cuando cada una de las celdas tiene un conteo de datos mayor o igual a cinco observaciones. Cuando no se cumple esta condición, puede emplearse alguna otra prueba de bondad de ajuste, como la de Kolmogorov-Smirnov.

2.5.2 Modelo de Champernowne de la distribución del ingreso

En 1953, Champernowne escribió un artículo en el cual desarrollo un modelo ex profeso para la distribución de los ingresos, el cual tenía la característica de considerar que la distribución del ingreso se comportaba como un proceso estocástico sobre un conjunto enumerablemente infinito de rangos de ingresos. Antes de él, ya se había considerado esta situación, pero se asumía que la matriz estocástica asociada al modelo permanecía constante a través del tiempo. Bajo ciertas circunstancias, y suponiendo que se daban ciertas condiciones, la distribución tendía hacia una distribución única de equilibrio dependiente de cierta matriz estocástica pero no de la distribución inicial. Él

encontró, sin embargo, que bajo cierto conjunto más o menos amplio de condiciones, la distribución de equilibrio tenía una curva Pareto en el límite, y sería asintótica a una cierta línea recta.

2.5.3 La Salamandra: Un modelo de la cola derecha de la distribución de ingresos truncados por codificación

Este modelo propone una estrategia para estimar la cola derecha truncada de ingresos anuales y distribuciones de sueldos utilizando el resto de la distribución. La situación de regenerar la parte derecha de la distribución utilizando el resto de la información es lo que sugiere precisamente el nombre de este método: la Salamandra.

En esencia, este método recurre a una mezcla de funciones de densidad de probabilidad gamma en las cuales los parámetros se restringen en cierto modo particular.

Una desventaja importante de esta metodología reside en el hecho de que las codificaciones de los datos implican el hecho de que el truncamiento se da de manera arbitraria. Por ejemplo, debe de decidirse de antemano cuales serían las percentilas superiores que buscan reproducirse, e.g., la percentila del 1% superior, utilizando el 99% de los datos restantes. Esto implica una seria restricción puesto que de antemano no es claro cual es el punto de corte que debería de imponerse para la restitución de las percentilas superiores, i.e., surgen preguntas tales como: ¿sería más conveniente reproducir el 2%? ¿qué implica esto con respecto a las estimaciones asociadas?

A su vez, el análisis de la Salamandra tiene como un fin subyacente la conformación de estadísticos de desigualdad y pobreza, tales como el de Gini. Esto es, a pesar de que en principio la finalidad es el reproducir la cola derecha de la distribución de los ingresos, esto se hace con el objetivo último de obtener una mayor precisión en el coeficiente de Gini e índices afines.

Para un desarrollo más extenso, véase las referencias [4] y [5].

2.5.4 Modelos de comparación de intercambio de individuos en poblaciones con diferentes ingresos

Un problema asociado, tal y como se encuentra analizado en Dagum (1987) es el hecho de establecer diferencias métricas de la afluencia económica entre poblaciones de personas que reciben ingresos. Esto es, cuando se tienen poblaciones con medias diferentes de los ingresos, surge una propensión a emigrar, cambiando por supuesto de manera dinámica los ingresos de ambas poblaciones, y para lo cual se propone una nueva métrica conocida como la *afluencia económica relativa* (REA, por sus siglas en inglés). Este problema, el cual es sumamente interesante, llevaría a la consideración de elementos dinámicos de transferencias de ingresos.

2.5.5 La graduación de la distribución de los ingresos

En el artículo discutido por Fisk (1961), se sugiere un modelo muy interesante de los ingresos que surge como solución a cierta ecuación diferencial. En específico, con anterioridad a dicho documento se habían sugerido una cantidad importante de formas funcionales como distribuciones razonables de los ingresos. Algunas de ellas habían sido derivadas como modelos "explicativos" de la generación de una distribución de los ingresos, mientras que otras habían surgido como solamente como técnicas de ajuste de observaciones. Una que no había sido lo suficientemente estudiada era la distribución cuadrática sech (o Fisk). Esta distribución cuenta con ciertas propiedades útiles, tal como una medida simple de Lorenz de la desigualdad y un método de análisis gráfico, lo que lo hace una herramienta útil en el análisis y comparación de distribuciones de ingresos. La ecuación diferencial de la cual la distribución cuadrática sech se deriva puede ser perturbada para permitir que un amplio rango de distribuciones diferentes puedan ser ajustadas. En este sentido, existe una cierta similaridad entre esta función de distribución y las funciones de distribución de Pareto y de Champernowne.

2.5.6 Aplicación de leyes económicas a la distribución de ingresos de países particulares

Han surgido un gran número de estudios referentes a la aplicación de leyes económicas aplicadas a la distribución de ingresos. Probablemente la más importante de estas sea la ley de Pareto. Por ejemplo, tal y como se analiza en Hayakawa (1951), se ha aplicado la ley de Pareto al ajuste de datos en Japón. La ley de Pareto afirma que, de manera aproximada, se valida la siguiente expresión:

$$N = AX^{-\alpha}$$

donde X es la magnitud del ingreso, N es el número de individuos que tienen dicho ingreso o superior, y A y α son constantes a determinarse de manera empírica por las observaciones. En un papel doble logarítmico se debería de tener una línea recta con pendiente igual a $-\alpha$. Debido a ello, la expresión anterior puede re-escribirse como:

$$\log N = \log A - \alpha \log X.$$

donde α, conocida como la "constante de Pareto", ha mostrado una gran estabilidad para ingresos suficiente altos, y adquiere usualmente valores alrededor de 1.5.

Sin embargo, el mismo Pareto reconoció ciertos defectos en esta expresión, y la principal es que la fórmula $N = AX^{-\alpha}$ representa solamente una simplificación de la ecuación más general

$$\log N = \log A - \alpha \log(a + X) - \beta X,$$

donde β debería determinarse para cada región y tiempo de estudios de estadística empírica, aunque debido a que su valor es frecuentemente pequeño, Pareto mismo la desdeño.

A su vez, debido a que con frecuencia se ha utilizado $N = AX^{-\alpha}$ no solo para los valores extremos sino a toda la distribución, ha sufrido de una gran falta de ajuste asociada con estudios posteriores.

2.5.7 Modelo del ingreso basado en la elasticidad compartida de los ingresos

Un modelo adicional, desarrollado por Majumder (1990), utilizó un modelo tetra-paramétrico de la distribución del ingreso utilizando la "elasticidad compartida de los ingresos". Este modelo en particular incluye las distribuciones de Singh-Maddala (1976) y Dagum (1977) como casos especiales. La distribución beta generalizada II (GB2) de McDonald también es una variante de este modelo. Por el análisis de este modelo, se probó que proporcionaba un excelente ajuste a los datos de ingresos de Estados Unidos y su desempeño empírico resultó ser superior al de Singh-Maddala (1976), Dagum (1977) y la distribución penta-paramétrica de Champernowne (1953), así como la distribución beta generalizada II de McDonald (1984) para algunos conjuntos de datos.

2.5.8 Modelos de series de tiempo para la distribución de los ingresos

Existe también una serie de estudios que basan fundamentalmente su enfoque en la teoría de series de tiempo. Por ejemplo, en Nirei (2004), se analizan de manera empírica las distribuciones de ingresos y se propone un modelo estocástico para explicar la distribución estacionaria de los datos y sus desviaciones relativas.

2.5.9 Métodos no-paramétricos que se basan en la técnica del Kernel

El histograma es una forma gráfica que representa la frecuencia relativa de ocurrencias del valor de la variable del ingreso. Esta función tiene la característica de tener saltos en sus puntos de corte, aún cuando los datos que representan sean variables continuas. Como Resenblatt (1956) propuso, es posible transformar este histograma para obtener una función de densidad donde el paso más crítico es el de tomar la decisión de cuanto suavizar los bordes utilizando una función ponderada de kernel. El *método de Kernel* es la técnica más ampliamente estudiada desde una perspectiva matemática y se utiliza con frecuencia para realizar estimaciones de densidad no-paramétricas. En la estimación del Kernel de $f(x)$ del histograma suavizado se tiene que:

$$\hat{f}(x) = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{h}K\left(\frac{x - x_i}{h}\right)$$

donde la función Kernel $K(\cdot)$ es generalmente unimodal, simétrica y acotada, y h es llamado el *parámetro de suavizamiento*. Intuitivamente se puede imaginar esto como una protuberancia que se coloca en cada punto muestral, de forma tal que la suma de todas ellas genera la distribución de todos los puntos. La función Kernel determina la forma de cada protuberancia mientras que el parámetro de suavizamiento determina la amplitud de cada una de ellas. La función tiene las siguientes propiedades:

- No existe necesidad de conocer el rango de los datos de antemano;
- $\widehat{f}(x)$ por sí misma forma una función de densidad la cual hereda todas las propiedades de continuidad, diferenciabilidad e integrabilidad de la función Kernel;
- K y h son dos factores que afectan su precisión pero de manera esencial está afectada por el parámetro de suavizamiento.

La estimación consiste en medir y minimizar el error global entre la función de densidad y la función de densidad subyacente real del

$$\text{Error cuadrático medio integrado}(\widehat{f}, f) = E\int [\widehat{f}(x) - f(x)]dx \to 0$$

donde

$$\widehat{f}(x) = \frac{1}{N\sqrt{2\pi}} \sum_{i=1}^{N} \exp\left\{-\frac{1}{2}\left(\frac{x - x_i}{h}\right)^2\right\}$$

y

$$h = 1.06\sigma N^{-1/5}$$

si se utiliza el Kernel normal.

2.5.10 Modelos de mezclas de distribuciones

Finalmente, dentro de la revisión de la literatura, existen también modelos de mezclas de distribuciones, en los cuales en esencia se trata de modelar por partes la distribución de los ingresos utilizando una variedad de funciones de distribución (Grazia-Pittau & Zelli, 2004). En concreto, con un modelo basado en una mezcla de distribuciones, los datos observados pueden ser considerados como generados por una mezcla de g distribuciones componentes en proporciones desconocidas $\pi_1,...,\pi_g$, donde las proporciones de las mezclas π_i son no negativas y suman uno. Los modelos de mezclas pueden ser considerados como una alternativa semi-paramétrica a las densidades no paramétricas, especialmente cuando la densidad no-paramétrica presenta más de una moda, y proporciona, en general, mayor flexibilidad y precisión en la modelación de las distribuciones subyacentes de los datos muestrales.

Formalmente, la función de densidad de probabilidad del vector aleatorio X_j bajo un modelo de mezcla de $g-$componentes, se define como:

$$f(x_j, \boldsymbol{\Psi}) = \sum_{i=1}^{g} \pi_i f_i(x_j, \theta_i),$$

donde el vector $\boldsymbol{\Psi} = (\pi_1,...,\pi_{g-1}, \xi')'$ contiene todos los parámetros del modelo de mezcla; π_i, $i = 1,...,g$ representa las proporciones de la mezcla y el vector ξ contiene todos los parámetros $(\theta_1,...,\theta_g)$ conocidos a priori que sean distintos; $f_i(x_j, \theta_i)$ denota los valores de la densidad univariada especificada por el vector paramétrico θ_i; en el

caso de normalidad de los componentes, $\theta_i = (\mu, \sigma_i^2)$ denota la media y la varianza de cada $i-$ésima componente normal.

Las proporciones de la mezcla, $\pi_1,...,\pi_g$ proporcionan la probabilidad a priori de que una unidad económica pertenezca a la $i-$ésima componente de la mezcla, representando un parámetro el cual determina la importancia relativa de cada componente en la mezcla. Una de las principales ventajas en la utilización de los modelos de mezclas es que, una vez que el modelo se ha generado, las probabilidades condicionales de que un hogar con un cierto ingreso provenga de un componente de la mezcla pueden ser calculadas para cada hogar. De manera formal, la probabilidad a posteriori o condicional τ_{ij} es:

$$\tau_{ij} = \tau_i(x_j, \boldsymbol{\Psi}) = \frac{\pi_i f_i(x_j, \theta_i)}{\sum_{i=1}^{g} \pi_i f_i(x_j, \theta_i)}$$

la cual representa la probabilidad de que el $j-$ésimo hogar con ingreso x_j provenga del $i-$ésimo componente de la mezcla.

El hecho de ajustar una densidad desconocida mediante la mezcla de g componentes en proporciones iguales $1/n$, donde n es el tamaño de la muestra observada, es equivalente a una estimación de kernel no-paramétrica. Por lo tanto, los modelos de mezclas pueden ser considerados como un balance entre los modelos paramétricos, representados por una familia paramétrica individual $(g = 1)$ y un modelo no-paramétrico, representado (en el caso de que $g = n$) por el estimador de la densidad kernel.

Suponiendo que la distribución de cada componente es normal, una mezcla finita es una aproximación razonable cuando los grupos están bien separados. La suposición de normalidad podría ser muy restrictiva, puesto que en principio cualquier forma funcional puede ser considerada. Se han combinado, por ejemplo, distribuciones gamma (o lognormal) con exponenciales para aproximar la distribución de ingresos de hombres en Nueva Zelanda. La primera forma funcional se selecciona para modelar el componente medio de la mezcla, mientras que el segundo espera capturar la cola inferior, ayudando de cualquier forma a representar también la cola derecha. Sin embargo, la familia de mezclas de densidades normales de g componentes puede aproximar de manera razonable una gran cantidad de formas distribucionales, y por dicha razón es un enfoque que con frecuencia se toma al realizar análisis de mezclas de distribuciones (asociado, por supuesto, con el hecho de las bondadosas propiedades de la distribución normal). Más aún, mientras que en algunos análisis el número de componentes se selecciona a priori y por lo tanto se centra el objetivo en la selección de la forma funcional, también existen estudios en los cuales no se realiza ninguna conjetura del número de grupos y por ende se determinan a posteriori. En este caso la consideración acerca del número de componentes que debería de incluir el modelo así como la selección del menor número de componentes debe de empatarse con un análisis empírico de los datos.

La forma preferida de estimar los parámetros del modelo de mezclas es la utilización del *Algoritmo EM de Maximización de la Esperanza (Expectation-Maximization Algorithm),* sobre el cual se conjetura que sus propiedades son superiores a las de otros procedimientos que tratan de encontrar un máximo local de la función de verosimilitud.

La dificultad principal que se tiene en los modelos de mezclas, y que hasta el momento no se ha resuelto completamente de manera satisfactoria, es el hecho de determinar el número g de componentes. Por supuesto, una manera evidente de tratar de resolver esta problemática consistiría en probar el juego de hipótesis $H_0 : g = g_0$ en contra de $H_a : g = g_a$ para algún $g_a > g_0$, utilizando para ello el *estadístico de prueba de razón de verosimilitud (LTRS, likelihood ratio test statistics).* Desafortunadamente, en modelos de mezclas, no se cumplen las condiciones de regularidad usuales y la distribución asintótica del LTRS no es necesariamente chi-cuadrada. Así, se han propuesto métodos alternativos como procedimientos tipo *bootstrap* para estimar la distribución bajo la hipótesis nula. Para muestras muy grandes, sin embargo, el esfuerzo computacional puede no permitir la aplicación de esta técnica. Por ello, se han propuesto criterios basados en la verosimilitud integrada, tal como el *Criterio de Información Bayesiana (BIC, Bayesian Information Criterion) de Schwarz.* El BIC proporciona, bajo ciertas condiciones de regularidad, una aproximación fidedigna de la verosimilitud integrada. En el BIC se adiciona un término a la verosimilitud para penalizar la complejidad del modelo, por lo que podría ser maximizado mediante una parametrización parsimoniosa. Sin embargo las condiciones de regularidad para el BIC no se sostienen con cierta frecuencia.

Finalmente, es importante observar que los componentes reflejan grupos distintos en la población, los cuales pueden no necesariamente corresponder al número de modas detectadas en la distribución. Por ejemplo, aún cuando el número de modas pueda sugerir el número de distribuciones subyacentes del ingreso, un modelo de distribución puede ser también unimodal cuando los componentes no están suficientemente separados. Por otra parte, la bimodalidad no necesariamente implica que los datos se generen de una distribución de mezclas de dos componentes.

2.6 MODELOS ESTOCÁSTICOS QUE COADYUVAN A AFIRMACIONES DE INEQUIDAD DEL INGRESO

Aunque la finalidad de este trabajo no es el analizar índices de desigualdad que provengan de los datos de los ingresos, este es un tema que merece, al menos, algunos comentarios, sobre todo para visualizar la relación que tiene con la distribución de los ingresos.

La idea general cuando se realizan estudios de inequidad y pobreza es conformar índices derivados de los datos de los ingresos que coadyuven a probar o refutar, según sea el caso, hipótesis que tengan los investigadores con respecto a la distribución de los ingresos. De hecho, y en términos generales, los estudios desarrollados al respecto se refieren a tres tipos de análisis:

a) Desigualdad debida a la riqueza relativa extrema.
b) Desigualdad asociada a bajos niveles de ingresos extremos.
c) Desigualdad asociada a pobreza extrema.

Por supuesto, buena parte de la literatura asociada con este punto se centra en la problemática de tratar de definir, con la menor ambigüedad posible, el significado de términos tales como "riqueza", "pobreza", "altos niveles de ingreso" y "bajos niveles de ingreso". A su vez, una parte fundamental es el tratar de establecer "buenas" propiedades a los índices desarrollados, tales como el hecho de que sean libres de la escala de medida, eficientes, de máxima verosimilitud, etc.

Se pueden proporcionar una lista muy grande de las propiedades que se desearía que tuviesen los índices de desigualdad. A continuación se exponen los que probablemente sean los más importantes:

- *Familiaridad* con el índice y *conveniencia* de su cálculo o su estimación.
- *Imparcialidad* entre personas en el sentido que los índices sólo dependen de la distribución de frecuencias de los ingresos y no del orden en el cual los individuos se encuentran dentro de la distribución de forma tal que no exista asociación con características tales como riqueza, poder, ventajas políticas, raza o salud. Desde cierto punto de vista, tal imparcialidad podría ser considerada como una desventaja, pero para propósitos de un estudio estadísticos permite una gran simplificación.
- *Invarianza* con respecto al número de personas que reciben ingresos. De manera más precisa el índice debería, de acuerdo a este criterio, resultar imperturbado si se conserva inalterada la distribución proporcional de personas a través de la escala de los ingresos, aún si se incrementa o decrementa el número total de personas. Un ejemplo de un índice que no satisface esta propiedad es el índice de entropía de Theil, el cual consideraría la combinación de dos poblaciones, cada una con una distribución idéntica de ingresos, como menos inequitativa que la que se tendría si se estudiaran de forma separada. Esta propiedad derivaría en el hecho de que el índice tomaría el valor de "uno" en distribuciones en las cuales existiera un individuo que tuviera todo el ingreso.
- *Invarianza* con respecto a un *incremento* (o decremento) *uniforme* del tamaño de los ingresos. De manera más precisa, el índice debería de permanecer inalterado si cada ingreso se incrementa (o decrementa) en la misma proporción.
- *Eficiencia Pigout-Dalton*. Este criterio requiere que si una distribución se modifica por la alteración de *dos* ingresos que dejen inalterado el gran total, entonces el índice correspondiente debe de incrementarse, permanecer igual o decrementarse de acuerdo a si la diferencia absoluta entre los dos ingresos se incrementa, permanece sin cambios o se decrementa, respectivamente. Cualquier índice que sea la media aritmética de una función estrictamente convexa del ingreso satisface este criterio.
- *Tener un rango de cero a uno*. Se cumple esta condición cuando el índice tiene un valor de "cero" en aquella distribución donde los ingresos son iguales y un valor de "uno" cuando un solo individuo posee todo el ingreso. Esta es una condición ligeramente menos estricta que el índice no modificado de Theil y que puede ser satisfecha mediante una modificación del mismo.
- *Adecuabilidad* para que un especialista esté en condiciones de medir un aspecto particular de la desigualdad en distinción a otros rubros.

Como ejemplos clásicos de índices que se han estudiado para estos fines, están los siguientes:

- Coeficiente de variación del ingreso.
- La desviación estándar de la potencia del ingreso.
- La proporción mediante la cual el ingreso geométrico medio cae debajo del ingreso medio aritmético.
- La proporción mediante la cual la el ingreso armónico medio cae debajo del ingreso medio aritmético. Tanto este como el anterior son casos particulares del índice de desigualdad de Atkinson.
- El coeficiente de desigualdad de Gini.
- El coeficiente de desigualdad de entropía de Theil.

A su vez, como se menciona en Cowell & Victoria-Feser (1996) y Dagum (1980), se han realizado estudios extensos acerca de las propiedades de estos y varios índices más, tales como propiedades de robustez, comparación de índices a través del tiempo en diversos países y casos de aplicación asociados.

Probablemente, y de manera actual, las dos cantidades más ampliamente utilizadas sean la llamada curva de Lorenz y el índice de Gini. De hecho, se han diseñado métodos robustos y eficientes para su estimación, tal y como se menciona en Gastwirth (1972).

Con este recuento de tipo histórico acerca de la problemática de la modelación de los ingresos, se concluye el presente capítulo 2. En el siguiente se sigue una exposición de la teoría de valores extremos, del estadístico de mayor orden r y de los modelos de umbral, mismos que forman parte clave de este trabajo.

CAPÍTULO 3

MODELACIÓN ESTADÍSTICA PARA EVENTOS EXTREMOS EN LA DISTRIBUCIÓN DE INGRESOS

Mathematics possesses not only truth, but also supreme beauty
a beauty cold and austere, like that of sculpture
without appeal to any part of our weaker nature...
Sublimely pure, and capable of stern perfection
such as only the greatest art can show

Bertrand Russell (1872-1970)

3.1 RESUMEN CAPITULAR

Es fundamental contar con un conocimiento preciso de la teoría de eventos extremos para estar en condiciones de conocer las aristas que surgen en sus aplicaciones y modelado a la distribución de ingresos.

En este capítulo se exponen los fundamentos de la teoría de valores extremos que serán de importancia para contar con un marco conceptual apropiado de las aplicaciones que serán llevadas a cabo posteriormente.

3.2 INTRODUCCIÓN A LA TEORÍA DE VALORES EXTREMOS

La teoría estadística de valores extremos ha cobrado una gran importancia desde hace varios años, debido sobre todo al amplio rango de aplicaciones que tiene en la ciencia y la tecnología. Así, cabe señalar que existen aplicaciones de gran importancia en campos tan diversos como:

- Ajuste de portafolios financieros en la industria del aseguramiento.
- Análisis de riesgo en mercados financieros.
- Problemas de predicción de tráfico en líneas de telecomunicación.
- Modelación oceanográfica.
- Problemas de fallas de memoria en dispositivos electrónicos.
- Ingeniería holística.
- Planeación estratégica.
- Procesamiento de datos biomédicos.
- Termodinámica de temblores.
- Análisis de cambios meteorológicos.
- Vibraciones no lineales de cuerdas.
- Ciencia alimentaria.
- Ingeniería de estructuras.
- Ingeniería hidráulica.
- Meteorología.
- Problema de fatiga de materiales.
- Resistencia a la corrosión.
- Estudios de contaminación.

La lista podría continuar, pero con esta se cuenta con una buena percepción de la enorme importancia que tiene esta teoría así como sus aplicaciones.

La gran aceptación que ha tenido la teoría de valores extremos se asocia con su versatilidad para modelar sucesos y fenómenos en los cuales resulta de interés analizar sus comportamientos extremos. Por ejemplo, en una presa, los niveles extremos de sequedad y desbordamiento debido a épocas de sequías o de lluvias, respectivamente, son factores importantes para la proyección de dichos niveles, y su modelación es fundamental para una correcta planificación del diseño de la presa.

La característica distintiva del análisis de valores extremos reside en el hecho de que el objetivo del análisis es el cuantificar el comportamiento estocástico de un proceso en los niveles inusitadamente altos o bajos, más que los niveles promedios. De hecho, casi cualquier estudio de valores extremos requiere la estimación de la probabilidad de eventos que son más extremos que cualquier otro valor que hubiese sido observado. Esta consideración implica que los datos que son motivos de análisis son necesariamente escasos, puesto que por definición, los datos extremos así lo son, y esto tiene importantes repercusiones en cuanto a la forma de enfocar la modelación por valores extremos.

Ahora, en ausencia de guías físicas o empíricas con las cuales se pueda formular una regla de extrapolación, los modelos estándar de la teoría de valores extremos se derivan del siguiente argumento. Supóngase que $X_1, X_2, ..., X_n$ es una sucesión de valores de alguna variable. Entonces se define

$$M_n = \max\{X_1, X_2, ..., X_n\}$$

como el máximo valor sobre un "$n-$periodo" de observación. Si el comportamiento estadístico exacto de X_i fuese conocido, entonces la correspondiente naturaleza estadística de M_n también sería conocido. En la práctica, el comportamiento de X_i es desconocido, haciendo que los cálculos asociados para determinar a M_n sean imposibles. Sin embargo, bajo un cierto conjunto razonable de suposiciones, el comportamiento aproximado de M_n para valores grandes de n se puede derivar siguiendo ciertos argumentos detallados al estudiar el límite cuando $n \to \infty$, lo cual lleva a la consideración de una familia de modelos que pueden ser entonces calibrados por los valores observados de M_n. Este enfoque podría ser llamado el *paradigma de valores extremos*, puesto que lleva a la consideración de un principio para la extrapolación de modelos basado en la implementación de límites matemáticos como aproximaciones de nivel finito. Es fácil objetar en contra de este procedimiento sobre la base que, aún con el soporte de un argumento asintótico, existe una suposición implícita de que el mecanismo estocástico subyacente del proceso que está siendo modelado es lo suficientemente suave para permitir la extrapolación de niveles no observados. Sin embargo, no existe ninguna otra alternativa creíble diseñada hasta el momento.

Desde un inicio es importante percatarse de las limitaciones implicadas por la adopción del paradigma de valores extremos. Primeramente, los argumentos se desarrollan utilizando caracterizaciones asintóticas, y por ello se necesita tener cuidado para aplicarlos a muestras finitas. En segundo lugar, los modelos por sí mismos son derivados bajo circunstancias idealizadas, que podrían ser no completamente exactas para un proceso bajo estudio. En tercer lugar, los modelos podrían implicar pérdidas de información cuando se implementan en la práctica. Para clarificar este punto, una forma común de registrar los datos extremos consiste en almacenar solamente los valores máximos sobre un periodo específico (quizá los máximos anuales). Suponiendo que este valor de n es lo suficientemente grande, el argumento asintótico tiene como consecuencia otro modelo que describe las variaciones en los máximos anuales de un año a otro y el cual puede ser ajustado a los máximos anuales observados. Pero en un año particular, eventos extremos adicionales podrían haber ocurrido en forma tal que hiciera posible tener sucesos aún más extremos que el máximo sobre los otros años. Debido a que estos datos no son los máximos anuales en los años bajo consideración, podrían ser excluidos del análisis.

Todos los puntos anteriores enfatizan la importancia de implementaciones estadísticas como un complemento para el desarrollo de modelos apropiados para los valores extremos. Cuatro hechos de importancia, en particular, necesitan ser considerados:

1. **Métodos de estimación.** Estos son los medios mediante los cuales los parámetros desconocidos de un modelo se infieren sobre la base de los datos históricos. A pesar de que han sido propuestos varios métodos diferentes para la estimación de modelos de valores extremos, en el presente se adopta el de maximizar la función de verosimilitud. Todos los métodos de estimación tienen sus pros y sus contras, pero las técnicas basadas en la función de verosimilitud son únicas en su poder de adaptación al cambio de modelos. Esto es, aunque las ecuaciones de estimación cambien si un modelo es modificado, la metodología subyacente permanece esencialmente invariable. Más aún, se adopta la estrategia de la máxima verosimilitud puesto que presenta un conjunto conveniente de propiedades inferenciales en grandes muestras.

2. **Cuantificación de la incertidumbre.** En cualquier análisis estadístico, las estimaciones son las "mejores conjeturas" sobre la veracidad de la información de los datos históricos. Es un hecho implícito que otro conjunto de datos que fuese igualmente representativo del proceso real que está siendo estudiado, podría proporcionar otras estimaciones diferentes. Por lo tanto, es importante complementar la estimación de un modelo con una medida de la incertidumbre asociado a la variabilidad muestral. Esto es especialmente importante en la modelación de los valores extremos, en los cuales cambios pequeños en el modelo pueden ser magnificados fuertemente en el proceso de extrapolación. A pesar de ello, la medición de la incertidumbre ha sido frecuentemente ignorada en las aplicaciones de valores extremos. Existe una buena dosis de ironía en esto, puesto que un análisis de valores extremos tienen mayores probabilidades de contar con más fuentes de incertidumbre que la mayoría de los análisis estadísticos. Por ende, la estimación de la incertidumbre de valores extremos en un proceso puede ser un parámetro de diseño tan importante como la estimación del parámetro mismo. Se puede mostrar que al basar la inferencia de la incertidumbre sobre la función de verosimilitud, sus estimaciones se obtienen de manera relativamente fácil.

3. **Diagnóstico del modelo.** La única justificación para extrapolar un modelo de valores extremos está sobre una base asintótica sobre la cual se derive. Sin embargo, si un modelo tiene un mal desempeño en términos de la representación de los valores extremos que han sido observados, existe una pequeña esperanza de que tenga un buen funcionamiento en la extrapolación. Para cada modelo de valor extremo, se deben de introducir métodos apropiados para asegurar la bondad de ajuste.

4. **Máximo uso de la información.** A pesar de que la incertidumbre es inherente en cualquier modelo estadístico, tal incertidumbre puede ser reducida mediante una selección cuidadosa del modelo y de la técnica de inferencia, y mediante la utilización de todas las fuentes posibles de información. En un contexto de valores extremos, las posibilidades incluyen el uso de modelos alternativos que explotan más datos que solamente la variable básica de interés; el uso de la información de la covarianza; la construcción de modelos multivariados y la incorporación de fuentes adicionales de conocimiento o información al análisis.

3.3 MODELOS ASINTÓTICOS

3.3.1 Formulación del modelo

En esta sección se comienza la presentación de una de las partes angulares de la teoría de valores extremos. El modelo se centra en el análisis del comportamiento estadístico de

$$M_n = \{X_1, X_2, ..., X_n\},$$

donde $X_1, X_2, ..., X_n$, es una sucesión de variables aleatorias independientes que tienen una función de distribución común F. En las aplicaciones, las $X_i's$ representan usualmente los valores de un proceso medido sobre una escala temporal regular (quizá de hora en hora, de tipo anual, etc.), por lo que M_n representa el máximo del proceso sobre n unidades de tiempo de observación. Si n corresponde al número de observaciones en un año, entonces M_n corresponde a un máximo anual.

En teoría, la distribución de M_n puede ser derivada exactamente para todos los valores de n:

$$\begin{aligned} \Pr\{M_n \le z\} &= \Pr\{X_1 \le z, X_2 \le z, ..., X_n \le z\} \\ &= \Pr\{X_1 \le z\}\Pr\{X_2 \le z\} \times \cdots \times \Pr\{X_n \le z\} \\ &= \{F(z)\}^n. \end{aligned} \tag{3.1}$$

Sin embargo, esto no es inmediatamente útil en la práctica, debido a que la distribución de F es desconocida. Una posibilidad es el utilizar técnicas estadísticas estándar para estimar F de los datos observados, y entonces sustituir esta estimación en (3.1). Desafortunadamente, discrepancias muy pequeñas en la estimación de F puede generar grandes discrepancias en F^n.

Un enfoque alternativo es aceptar que F es desconocido, y buscar familias aproximadas para F^n, las cuales pueden ser estimadas sobre las bases de los datos extremos solamente. Esto es similar a la práctica usual de aproximar la distribución muestral de medias mediante una distribución normal, justificándolo mediante el teorema del límite central. Los argumentos que serán presentados a continuación son esencialmente los análogos en valores extremos a la teoría del límite central.

Se procede a observar el comportamiento de F^n cuando $n \to \infty$. Pero esto por sí solo no es suficiente: para cualquier $z < z_+$, donde z_+ es el punto frontera superior de F (esto es, z_+ es el menor valor de z tal que $F(z) = 1$), $F^n(z) \to 0$ cuando $n \to \infty$, por lo que la distribución de M_n degenera en una masa puntual sobre z_+. Esta dificultad se evita mediante una re-normalización lineal de la variable M_n:

$$M_n^* = \frac{M_n - b_n}{a_n},$$

para sucesiones de constantes $\{a_n > 0\}$ y $\{b_n\}$. Las selecciones apropiadas de las $\{a_n\}$ y $\{b_n\}$ estabilizan la localización y la escala de M_n^* a medida que n se incrementa, evitando las dificultades que surgen con la variable M_n. Por lo tanto, el análisis se enfoca en buscar las distribuciones límite de M_n^* con selecciones apropiadas de $\{a_n\}$ y $\{b_n\}$, más que $\{M_n\}$.

3.3.2 El Teorema de Tipos Extremos

El rango completo de distribuciones límite posibles para M_n^* está dado por el teorema 3.1, conocido como el *teorema de tipos extremos*.

Teorema 3.1. **(Teorema de tipos extremos)** Si existen sucesiones de constantes $\{a_n > 0\}$ y $\{b_n\}$ tales que

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to G(z) \text{ a medida que } n \to \infty,$$

donde G es una función de distribución no degenerada, entonces G pertenece a una de las siguientes familias:

I: $$G(z) = \exp\left\{-\exp\left[-\left(\frac{z-b}{a}\right)\right]\right\}, \quad -\infty < z < \infty;$$

II: $$G(z) = \begin{cases} 0, & z \le b, \\ \exp\left\{-\left(\frac{z-b}{a}\right)^{-\alpha}\right\}, & z > b; \end{cases}$$

III: $$G(z) = \begin{cases} \exp\left\{-\left[-\left(\frac{z-b}{a}\right)^{\alpha}\right]\right\}, & z < b, \\ 1, & z \ge b, \end{cases}$$

para parámetros $a > 0$, b y, en el caso de las familias II y III, $\alpha > 0$. ■

De manera equivalente, el teorema 3.1 establece que un máximo muestral re-escalado $(M_n - b_n)/a_n$ converge en distribución a una variable que tiene una distribución dentro de una de las familias señaladas como I, II y III. De manera colectiva, estas tres clases de distribuciones se llaman distribuciones de valores extremos, con los tipos I, II y III conocidos ampliamente como las familias **Gumbel**, **Fréchet** y **Weibull**, respectivamente. Cada familia tiene un parámetro de localización y escala, b y a, de

manera respectiva; adicionalmente, las familias Fréchet y Weibull tienen un parámetro de forma α.

El teorema 3.1 implica que, cuando M_n puede ser estabilizada con sucesiones apropiadas $\{a_n\}$ y $\{b_n\}$, la variable normalizada correspondiente M_n^* tiene una distribución limite que debe ser de uno de los tres tipos de distribuciones de valores extremos. El hecho importante de este resultado es que los tres tipos de distribuciones de valores extremos son los únicos límites posibles para la distribución de los M_n^*, sin considerar la distribución F de la población. Es en este sentido que el teorema provee un análogo de valores extremos al teorema del límite central.

3.3.3 La Distribución Generalizada de Valores Extremos (GEV)

Los tres tipos de límites que resultan del teorema 3.1 tienen formas distintas de comportamiento, correspondientes a las formas diferentes de las colas de la distribución F del X_i. Esto puede hacerse preciso mediante la consideración del comportamiento de la distribución límite G en z_+, su punto frontera superior. Para la distribución Weibull, $z_+ = \infty$. Sin embargo, la densidad de G decae exponencialmente para la distribución Gumbel y polinomialmente en la distribución de Fréchet, correspondientes a tasas relativamente diferentes de decaimiento en la cola de F. Se ha visto que en las aplicaciones las tres familias proporcionan representaciones muy diferentes en cuanto al comportamiento de valores extremos. En las primeras aplicaciones de la teoría de valores extremos, era usual adoptar una de las tres familias, y entonces estimar los parámetros relevantes de esta distribución. Pero existen dos debilidades: primero, se requiere una técnica para seleccionar cual de las tres familias es la más apropiada para los datos; segundo, una vez que la decisión se ha tomado, las inferencias subsecuentes asumen que esta selección es correcta, y no se clarifica el nivel de incertidumbre asociado con esta selección, aún cuando dicha incertidumbre pudiera ser grande.

Un mejor análisis es el realizar una reformulación del modelo en el teorema 3.1. Es directo constatar que las familias Gumbel, Fréchet y Weibull pueden ser combinadas en una sola familia de modelos que tengan funciones de distribución de la siguiente forma:

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}, \tag{3.2}$$

definida sobre el conjunto $\{z : 1+\xi(z-\mu)/\sigma > 0\}$, donde los parámetros satisfacen $-\infty < \mu < \infty, \sigma > 0$ y $-\infty < \xi < \infty$. Esta es la llamada familia **generalizada de distribuciones de valores extremos** (**GEV**, por sus siglas en inglés: **Generalizad Extreme Value**). El modelo tiene tres parámetros: un parámetro de localización: μ; un parámetro de escala: σ; un parámetro de forma: ξ. Las clases de tipo II y III de valores extremos corresponden respectivamente a los casos $\xi > 0$ y $\xi < 0$ en esta parametrización. El subconjunto de la familia GEV con $\xi = 0$ se interpreta como el

límite de (3.2) cuando $\xi \to 0$, que deriva en la **familia Gumbel** con función de distribución dada por:

$$G(z) = \exp\left[-\exp\left\{-\left(\frac{z-\mu}{\sigma}\right)\right\}\right], \qquad -\infty < z < \infty.$$

La unificación de las tres familias originales de distribuciones de valores extremos en una familia simple conlleva una simplificación importante en cuanto a su implementación estadística. A través de la inferencia sobre ξ, los datos por sí mismos determinan el tipo más apropiado del comportamiento de la cola, y no existe necesidad de realizar juicios subjetivos a priori acerca de cual familia de valores extremos particular hay que adoptar. Más aún, la incertidumbre en el valor inferido de ξ mide la incertidumbre acerca de cual de los tres tipos originales es el más apropiado para un conjunto de datos particular.

Por conveniencia, se modifica el teorema 3.1 en la forma siguiente:

Teorema 3.1.1. Si existe una sucesión de constantes $\{a_n > 0\}$ y $\{b_n\}$ tales que:

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to G(z) \text{ a medida que } n \to \infty, \tag{3.3}$$

para una función de distribución no degenerada G, entonces G es una miembro de la familia GEV:

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\},$$

definida sobre $\{z : 1+\xi(z-\mu)/\sigma\}$, donde $-\infty < \mu < \infty$, $\sigma > 0$ y $-\infty < \xi < \infty$. ■

Interpretando el límite en el teorema 3.1.1 como una aproximación para valores grandes de n, se sugiere de inmediato el uso de la familia GEV para la modelación de la distribución de máximos para grandes sucesiones de valores. La dificultad aparente de que las constantes de normalización sean desconocidas, se resuelve rápidamente en la práctica. Suponiendo la validez de (3.3), se tiene que:

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \approx G(z)$$

para n suficientemente grande. De forma equivalente:

$$\begin{aligned}\Pr\{M_n \le z\} &\approx G\{(z-b_n)/a_n\} \\ &= G^*(z),\end{aligned}$$

donde G^* es otro miembro de la familia GEV. En otras palabras, si el teorema 3.1.1 permite la aproximación de la distribución de M_n^* por un miembro de la familia GEV para n grande, la distribución de M_n puede en sí misma ser aproximada por un miembro diferente de la misma familia. Debido a que los parámetros de la distribución tiene que ser estimados de cualquier forma, es irrelevante en la práctica el que los parámetros de la distribución G sean diferentes de los de G^*.

Este argumento conduce al siguiente enfoque para la modelación de una serie de observaciones extremas e independientes $X_1, X_2, \ldots$. Los datos son agrupados en una sucesión de observaciones de longitud n, generando una serie de bloques de máximos, $M_{n,1}, \ldots, M_{n,m}$ por ejemplo, sobre los cuales se puede ajustar la función de distribución GEV. Frecuentemente los bloques se seleccionan de forma tal que correspondan a un periodo de tiempo anual, en cuyo caso n es el número de observaciones en un año y el máximo del bloque es el máximo anual. Las estimaciones de los cuantiles extremos de la distribución anual máxima son obtenidas entonces mediante el despeje de z_p de la ecuación (3.2):

$$z_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1 - \{-\log(1-p)\}^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma \log\{-\log(1-p)\}, & \xi = 0. \end{cases} \tag{3.4}$$

donde $G(z_p) = 1 - p.$ En terminología común, z_p es el nivel de retorno asociado con el periodo de retorno $1/p$, debido a que con un grado razonable de precisión, se espera que el nivel z_p sea excedido, en promedio, cada $1/p$ años. De manera más precisa, z_p es excedido por el máximo anual en cualquier año en particular con una probabilidad p.

Debido a que los cuantiles permiten que los modelos de probabilidad sean expresados en la escala de los datos, la relación entre el modelo GEV a sus parámetros es más fácilmente interpretable en términos de las expresiones de cuantiles (3.4). En particular, definiendo $y_p = -\log(1-p)$, se tiene que

$$z_p = \begin{cases} \mu - \dfrac{\sigma}{\xi}\left[1 - y_p^{-\xi}\right], & \xi \neq 0 \\ \mu - \sigma \log y_p, & \xi = 0. \end{cases}$$

de lo cual se sigue que si z_p se grafica contra y_p en escala logarítmica (o, de manera equivalente, si z_p se grafica en contra de $\log y_p$), la figura resultante es una recta en el caso de que $\xi = 0.$ Si $\xi < 0,$ la gráfica es convexa con límite asintótico $\mu - \sigma/\xi$ cuando $p \to 0;$ si $\xi > 0,$ la gráfica es cóncava y no tiene cota finita. Esta figura es la **grafica de nivel de retorno**. Debido a la simplicidad de interpretación, y puesto que la selección de la escala comprime la cola de la distribución a medida que el efecto de

extrapolación se intensifica, las gráficas de nivel de retorno son particularmente convenientes tanto para la presentación del modelo como su validación.

3.3.4 Bosquejo de la demostración del Teorema de Tipos Extremos

Una justificación formal del teorema de tipos extremos es técnica, aunque no es especialmente complicada. En esta sección se proporciona una prueba de tipo informal. Primero, es conveniente hacer la siguiente definición.

Definición 3.1. Una distribución G se dice que es **max-estable** si para todo $n = 2,3,...,$ existen constantes $a_n > 0$ y β_n tales que:

$$G^n(\alpha_n z + \beta_n) = G(z).$$

Debido a que G^n es la función de distribución de $M_n = \max\{X_1,...,X_n\},$ donde las X_i son variables independientes cada una con función de distribución G, la max-estabilidad es una propiedad satisfecha por distribuciones para las cuales la operación de tomar un máximo muestral deriva en una distribución idéntica, salvo de un cambio de escala y localización. La conexión con las leyes límite de valores extremos se realiza con la ayuda del siguiente resultado:

Teorema 3.2. Una distribución es max-estable si, y sólo sí, es una distribución generalizada de valores extremos.

Se requiere solamente la utilización de álgebra simple para mostrar que todos los miembros de la familia GEV son max-estables. Lo contrario requiere algunas ideas de análisis funcional, mismas que no serán tratadas aquí.

El teorema 3.2 es utilizado directamente en la prueba del teorema de tipos extremos. La idea consiste en considerar M_{nk}, la variable aleatoria máxima en una sucesión de $n \times k$ variables para algún valor grande de n. Esto puede ser considerado como el máximo de una sucesión individual de longitud $n \times k,$ o como el máximo de k máximos, cada uno de los cuales es el máximo de n observaciones. De manera más precisa, supóngase que la distribución límite de $(M_n - b_n)/a_n$ es G. Así, para un n lo suficientemente grande,

$$\Pr\{(M_n - b_n)/a_n \le z\} \approx G(z)$$

por el teorema 3.1.1. Por lo tanto, para cualquier entero $k,$ debido a que nk es grande,

$$\Pr\{(M_{nk} - b_{nk})/a_{nk} \le z\} \approx G(z). \qquad (3.5)$$

Pero dado que M_{nk} es el máximo de k variables que tienen la misma distribución que M_n,

$$\Pr\{(M_{nk} - b_{nk})/a_{nk} \le z\} = [\Pr\{(M_n - b_n)/a_n \le z\}]^k. \qquad (3.6)$$

Por lo tanto, por (3.5) y (3.6) de forma respectiva,

$$\Pr\{M_{nk} \le z\} \approx G\left(\frac{z - b_{nk}}{a_{nk}}\right)$$

y

$$\Pr\{M_{nk} \le z\} \approx G^k\left(\frac{z - b_n}{a_n}\right).$$

Por ende, G y G^k son idénticos salvo por los coeficientes de localización y escala. Se sigue entonces que G es max-estable y por lo tanto es un miembro de la familia GEV por el teorema 3.2.

3.3.5 Ejemplos

Un hecho que no se ha discutido hasta el momento asociado con el teorema 3.1 es si puede establecerse la convergencia de una distribución de variables normalizadas M_n, dada una función de distribución F. Si es posible, existen dos cuestiones adicionales: ¿cuáles son las selecciones necesarias de las sucesiones normalizantes $\{a_n\}$ y $\{b_n\}$? y ¿cuál miembro de la familia GEV se obtiene en el límite? Debido a que la finalidad principal que tiene este trabajo es el de proporcionar un esquema de inferencia de los datos reales sobre los cuales la distribución subyacente F es desconocida, se proporcionan solo algunos pocos ejemplos que ilustren la relación que guarda la selección de las sucesiones normalizantes con la distribución límite la familia GEV, como se implica en el teorema 3.1. Estos ejemplos también sirven para ilustrar otros resultados límites que pudieran ser de interés.

Ejemplo 3.1. Si $X_1, X_2, \ldots$ es una sucesión de variables exponenciales estándar independientes exp(1), entonces $F(x) = 1 - e^{-x}$ para $x > 0$. En este caso, sean $a_n = 1$ y $b_n = n$, con lo cual

$$\begin{aligned}\Pr\{(M_n - b_n / a_n) \le z\} &= F^n(z + \log n) \\ &= \left[1 - e^{-(z+\log n)}\right]^n \\ &= \left[1 - n^{-1}e^{-z}\right]^n \\ &\to \exp(-e^{-z})\end{aligned}$$

cuando $n \to \infty$, para cada $z \in \mathbb{R}$ fijo. Por lo tanto, con la selección de a_n y b_n, la distribución límite de M_n cuando $n \to \infty$ es la distribución Gumbel, correspondiente a $\xi = 0$ en la familia GEV.

Ejemplo 3.2. Si $X_1, X_2, \ldots$ es una sucesión de variables estándar Fréchet, entonces $F(x) = \exp(-1/x)$ para $x > 0$. Sean $a_n = n$ y $b_n = 0$. Entonces,

$$\begin{aligned} \Pr\{(M_n - b_n / a_n) \le z\} &= F^n(nz) \\ &= \left[\exp\{-1/(nz)\}\right]^n \\ &= \exp(-1/z) \end{aligned}$$

cuando $n \to \infty$, para cada $z > 0$ fijo. Por lo tanto, el límite en este caso –el cual es un resultado exacto para todo n, debido a la max-estabilidad de F – es también la distribución estándar Fréchet: $\xi = 1$ en la familia GEV.

Ejemplo 3.3. Si $X_1, X_2, \ldots$ es una sucesión de variables uniformes independientes $U(0,1)$, entonces $F(x) = x$ para $0 \le x \le 1$. Para un $z < 0$ fijo, supóngase que $n > -z$ y sean $a_n = 1/n$ y $b_n = 1$. Entonces,

$$\begin{aligned} \Pr\{(M_n - b_n / a_n) \le z\} &= F^n(n^{-1}z + 1) \\ &= \left(1 + z/n\right)^n \\ &\to e^z \end{aligned}$$

cuando $n \to \infty$. Por lo tanto, la distribución límite es de tipo Weibull, con $\xi = -1$ en la familia GEV.

Existe cierta arbitrariedad en la selección de $\{a_n\}$ y $\{b_n\}$ en los ejemplos mostrados. Sin embargo, diferentes selecciones que impliquen un límite no-degenerado siempre conlleva la consideración de una distribución límite en la familia GEV con el mismo valor de ξ, aunque con algunos parámetros posiblemente diferentes de localización y escala.

3.4 MODELOS ASINTÓTICOS PARA MÍNIMOS

Algunas aplicaciones requieren modelos para observaciones extremadamente pequeñas, más que para datos con magnitudes muy grandes, como lo podrían ser modelos de sistemas de fallas, como en el caso del lapso de vida de un sistema definido como el mínimo de tiempo de vida de n componentes individuales. En este caso, el tiempo de vida del sistema completo es entonces $\widetilde{M}_n = \min\{X_1, \ldots, X_n\}$, donde la X_i denotan los periodos de vida de los componentes individuales. Suponiendo que las X_i son independientes e idénticamente distribuidas, existen argumentos similares que se aplican a $\widetilde{M}_n$ como los aplicados a M_n, proporcionando una distribución límite para una variable convenientemente re-escalada.

Los resultados son inmediatos de los correspondientes para M_n. Sea $Y_i = -X_i$ para $i = 1, ..., n,$ donde el cambio de signo significa que valores pequeños de X_i corresponden a valores grandes de Y_i. Así, si $\widetilde{M}_n = \min\{X_1, ..., X_n\}$ y $M_n = \max\{Y_1, ..., Y_n\},$ entonces $\widetilde{M}_n = -M_n$. Por lo tanto, para una n grande

$$\begin{aligned}
\Pr\left\{\widetilde{M}_n \le z\right\} &= \Pr\{-M_n \le z\} \\
&= \Pr\{M_n \ge -z\} \\
&= 1 - \Pr\{M_n \le -z\} \\
&= 1 - \exp\left\{-\left[1 + \xi\left(\frac{-z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \\
&= 1 - \exp\left\{-\left[1 - \xi\left(\frac{z-\widetilde{\mu}}{\sigma}\right)\right]^{-1/\xi}\right\},
\end{aligned}$$

sobre $\{z : 1 - \xi(z - \widetilde{\mu})/\sigma > 0\},$ donde $\widetilde{\mu} = -\mu.$ Esta es la distribución GEV para los mínimos. Se puede utilizar este resultado formalmente como un teorema análogo al teorema 3.1 para máximos.

Teorema 3.3. Si existe una sucesión de constantes $\{a_n > 0\}$ y $\{b_n\}$ tales que

$$\Pr\left\{\frac{\widetilde{M}_n - b_n}{a_n} \le z\right\} \to \widetilde{G}(z) \quad \text{cuando} \quad n \to \infty$$

Para una distribución no-degenerada $\widetilde{G},$ entonces $\widetilde{G}$ es un miembro de la familia GEV para las distribuciones de los mínimos:

$$\widetilde{G} = 1 - \exp\left\{-\left[1 - \xi\left(\frac{z-\widetilde{\mu}}{\sigma}\right)\right]^{-1/\xi}\right\},$$

definida sobre $\{z : 1 - \xi(z - \widetilde{\mu})/\sigma > 0\},$ donde $-\infty < \widetilde{\mu} < \infty,$ $\sigma > 0$ y $-\infty < \xi < \infty.$ ■

En situaciones donde es apropiado el modelo de mínimos, la distribución GEV para mínimos puede ser directamente aplicada. Una alternativa consiste en explotar la dualidad entre las distribuciones para máximos y mínimos. Dados los datos $z_1, ..., z_m$ que son realizaciones de la distribución GEV para mínimos, con parámetros $(\widetilde{\mu}, \sigma, \xi),$ esto implica el ajustar la distribución GEV para máximos para los datos $-z_1, ..., -z_m.$ La estimación máximo verosímil de los parámetros de esta distribución corresponden

exactamente a la que es requerida para la distribución GEV de mínimos, mediante la corrección de signos $\hat{\hat{\mu}} = \hat{\mu}$.

3.5 INFERENCIA SOBRE LA DISTRIBUCIÓN GEV

3.5.1 Consideraciones generales

Motivado por el teorema 3.1.1, el GEV proporciona un modelo para la distribución de máximos por bloques. Sus aplicaciones consisten en bloques de datos de igual longitud, y el ajuste del GEV al conjunto de máximos por bloques. Pero en la implementación de este modelo a una base de datos particular, la selección del tamaño de bloques puede ser crítica. La situación delicada es el intercambio que existe entre sesgo y varianza: los bloques que son muy pequeños implican que la aproximación por el modelo límite del teorema 3.1.1 pudiera ser pobre, derivando en sesgo en la estimación y en extrapolaciones; bloques grandes generan pocos bloques de máximos, generando una gran varianza estimada. Las consideraciones pragmáticas frecuentemente llevan a la adopción de bloques de longitud anual. Por ejemplo, solamente los datos anuales máximos podrían ser registrados, por lo que el uso de bloques de menor longitud no es una opción correcta. Aún cuando este no sea el caso, un análisis de los datos anuales máximos parece ser más robusta que un análisis basado en bloques de menor longitud si las condiciones del teorema 3.1.1 son violadas. Por ejemplo, es probable que la temperatura diaria varíe de acuerdo a la temporada, violando la suposición de que las X_i tienen una distribución común. Si los datos son agrupados en bloques de longitudes aproximadamente trimestrales, el máximo del bloque de verano tiene mayor probabilidad de ser más mucho grande que el bloque de invierno, y una inferencia que no tome en consideración esta no-homogeneidad podría proporcionar resultados imprecisos. Tomando, en su lugar, bloques de longitud anual implicaría que es plausible la suposición de que los individuos máximos así agrupados tienen una distribución común, a pesar de que seguiría siendo invalida una justificación formal para la aproximación GEV.

Se simplificará ahora la notación mediante la notación de que los máximos por bloques sean $Z_1, ..., Z_m$. Estos se asumen como variables independientes de una distribución GEV cuyos parámetros deberán ser estimados. Si las X_i son independientes entonces también las Z_i lo son. Sin embargo, la independencia de las Z_i es una aproximación razonable aún si las X_i son una serie dependiente. En este caso, a pesar de que no se validan las suposiciones del teorema 3.1.1, la conclusión de que las Z_i tengan una distribución GEV puede ser aún razonable.

Muchas técnicas han sido propuestas para la estimación de parámetros en los modelos de valores extremos. Estas incluyen técnicas gráficas basadas en versiones de papeles de probabilidad; técnicas basadas en momentos en las cuales los momentos se igualan a sus equivalentes empíricos; procedimientos en los cuales los parámetros son estimados como funciones específicas de estadísticos de orden; y métodos basados en la verosimilitud (véase el apéndice C). Cada técnica tiene sus pros y sus contras, pero toda

la utilidad y adaptabilidad de los modelos construidos en técnicas basadas en verosimilitud hace que este enfoque sea particularmente atractivo.

Una dificultad potencial del uso de los métodos de verosimilitud para estimar los GEV se asocia con las condiciones de regularidad que son requeridas para que las propiedades asintóticas relacionadas con el estimador de máxima verosimilitud se sostengan. Tales condiciones no se satisfacen por los modelos GEV debido a que los puntos finales de la distribución GEV son funciones de los valores de los parámetros: $\mu - \sigma/\xi$ es el punto máximo superior de la distribución cuando $\xi < 0$, y un punto mínimo inferior cuando $\xi > 0$. Esta violación a las condiciones usuales de regularidad significa que los resultados estándares de verosimilitud asintótica no son automáticamente aplicables. Smith (1985) estudio este problema en detalle y obtuvo los siguientes resultados:

- Cuando $\xi > -0.5$, los estimadores de máxima verosimilitud son regulares, en el sentido de que tienen las propiedades asintóticas usuales;
- Cuando $-1 < \xi < -0.5$, los estimadores de máxima verosimilitud generalmente se pueden obtener, pero no tienen las propiedades asintóticas usuales;
- Cuando $\xi < -1$, es improbable que se pueden obtener los estimadores de máxima verosimilitud.

El caso en que $\xi \leq -0.5$ corresponde a distribuciones con una cola derecha muy pequeña y acotada. Esta situación se encuentra raramente en aplicaciones de modelación de valores extremos, así que las limitaciones teóricas del enfoque de máxima verosimilitud no son usualmente un obstáculo en la práctica.

3.5.2 Estimación por Máxima Verosimilitud

Bajo el supuesto de que $Z_1, ..., Z_m$ son variables independientes que tienen una distribución GEV, la log-verosimilitud para los parámetros GEV cuando $\xi \neq 0$ es

$$\ell(\mu, \sigma, \xi) = -m \log \sigma - (1 + 1/\xi) \sum_{i=1}^{m} \log \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right] - \sum_{i=1}^{m} \left[1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) \right]^{-1/\xi}, \quad (3.7)$$

bajo el supuesto de que

$$1 + \xi \left(\frac{z_i - \mu}{\sigma} \right) > 0, \quad \text{para} \quad i = 1, ..., m. \quad (3.8)$$

En una combinación de parámetros en la cual se viole la condición (3.8), que corresponda a una configuración en la cual al menos uno de los datos observados caiga más allá de un punto extremo de la distribución, implica que la verosimilitud sea igual a cero y que la log-verosimilitud sea igual a $-\infty$.

El caso en el que $\xi = 0$ requiere un tratamiento aparte mediante el uso del límite Gumbel de la distribución GEV. Esto implica la consideración de la log-verosimilitud siguiente:

$$\ell(\mu,\sigma) = -m\log\sigma - \sum_{i=1}^{m}\left(\frac{z_i-\mu}{\sigma}\right) - \sum_{i=1}^{m}\exp\left[-\left(\frac{z_i-\mu}{\sigma}\right)\right]. \tag{3.9}$$

La maximización de la pareja de ecuaciones (3.7) y (3.9) con respecto al vector de parámetros (μ,σ,ξ) proporciona la estimación de máxima verosimilitud con respecto a la familia GEV completa. Generalmente no existe una solución analítica cerrada, pero para un conjunto de datos dado la maximización es directa mediante la utilización de algoritmos estándares de optimización numérica. Sin embargo, debe prestarse especial cuidado en que tales algoritmos no muevan la combinación de parámetros que resulte en una violación a la condición (3.8), así como en las dificultades numéricas que podrían derivarse de la evaluación de (3.7) en una vecindad de $\xi = 0$. Este último problema se resuelve fácilmente utilizando (3.9) en lugar de (3.7) para valores de ξ que caen dentro de una pequeña región alrededor del cero.

Sujeto a las limitaciones de ξ discutidas con anterioridad, la distribución aproximada de $(\hat{\mu},\hat{\sigma},\hat{\xi})$ es una normal multivariada con media (μ,σ,ξ) y matriz de varianza-covarianza igual a la inversa de la matriz de información evaluada en el estimador de máxima verosimilitud (véase Anexo "C"). A pesar de que esta matriz puede ser calculada de manera analítica, es más sencillo utilizar técnicas de diferenciación numérica para evaluar las segundas derivadas y utilizar rutinas numéricas estándar para llevar a cabo la inversión. Los intervalos de confianza y otras formas de inferencia se siguen de manera inmediata de la normalidad aproximada del estimador.

3.5.3 Inferencia para Niveles de Retorno

Mediante la sustitución de los estimadores de máxima verosimilitud de los parámetros GEV en (3.4), la estimación de máxima verosimilitud de z_p para $0 < p < 1$, el nivel de retorno $1/p$ se obtiene como:

$$\hat{z}_p = \begin{cases} \hat{\mu} - \dfrac{\hat{\sigma}}{\hat{\xi}}\left(1 - y_p^{-\hat{\xi}}\right), & \text{para} \quad \hat{\xi} \neq 0, \\ \hat{\mu} - \hat{\sigma}\log y_p, & \text{para} \quad \hat{\xi} = 0. \end{cases} \tag{3.10}$$

donde $y_p = -\log(1-p)$. Además, mediante el método delta,

$$\text{var}(\hat{z}_p) \approx \nabla z_p^T V \nabla z_p, \tag{3.11}$$

donde V es la matriz de varianza-covarianza de $(\hat{\mu},\hat{\sigma},\hat{\xi})$ y

$$\nabla z_p^T = \left[\frac{\partial z_p}{\partial \mu}, \frac{\partial z_p}{\partial \sigma}, \frac{\partial z_p}{\partial \xi}\right] = \left[1, -\xi^{-1}\left(1 - y_p^{-\xi}\right), \sigma\xi^{-2}\left(1 - y_p^{-\xi}\right) - \sigma\xi^{-1} y_p^{-\xi} \log y_p\right]$$

evaluada en $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$.

Es el caso que grandes periodos de retorno correspondan a valores pequeños de p, por lo que son los de mayor interés. Si $\hat{\xi} < 0$ es posible realizar inferencias acerca de punto terminal superior de la distribución, el cual es el "periodo de retorno de una observación infinita", que corresponde a z_p con $p = 0$. El estimador de máxima verosimilitud es

$$\hat{z}_0 = \hat{\mu} - \frac{\hat{\sigma}}{\hat{\xi}},$$

y la expresión (3.11) es aún válida con

$$\nabla z_0^T = \left[1, -\xi^{-1}, \sigma\xi^{-2}\right],$$

evaluada, nuevamente, en $(\hat{\mu}, \hat{\sigma}, \hat{\xi})$. Cuando $\hat{\xi} \geq 0$ el estimador de máxima verosimilitud del punto extremo superior es infinito.

Se requiere precaución en la interpretación de las inferencias de los niveles de retorno, especialmente los niveles de retorno correspondientes a grandes periodos de retorno. Primeramente, la aproximación normal a la distribución del estimador de máxima verosimilitud podría ser pobre. Generalmente se obtienen mejores aproximaciones mediante la función de verosimilitud perfil (véase Anexo "C"). De manera más fundamental, los estimadores y sus mediciones de precisión se basan en el supuesto de que el modelo es correcto. A pesar de que el modelo GEV tiene un soporte matemático firme, su utilización en la extrapolación se basa en supuestos verificables, y las medidas de incertidumbre de los niveles de retorno deberían ser propiamente consideradas como cotas inferiores que podrían ser mucho mayores si la incertidumbre asociada a la corrección del modelo fuese tomada en consideración.

3.5.4 Verosimilitud Perfil

La evaluación numérica de la verosimilitud perfil para cualquiera de los parámetros individuales μ, σ o ξ es directa. Por ejemplo, para obtener la verosimilitud perfil de ξ, se fija $\xi = \xi_0$, y maximizar la log-verosimilitud (3.7) con respecto a los parámetros restantes, μ y σ. Esto se repite para un rango de valores de ξ_0. Los valores correspondientes a este proceso de maximización de la log-verosimilitud constituye la log-verosimilitud perfil para ξ, y mediante la cual y utilizando el siguiente teorema, se pueden obtener intervalos de confianza apropiados.

Teorema 3.4. Sean $x_1,...,x_n$ realizaciones de una distribución dentro de una familia paramétrica F, y sea $\hat{\theta}_0$ el estimador de máxima verosimilitud del parámetro $d-$dimensional del modelo $\theta_0=(\theta^{(1)},\theta^{(2)})$, donde $\theta^{(1)}$ es un subconjunto $k-$dimensional de θ_0. Entonces, bajo condiciones convenientes de regularidad, para un n grande

$$D_p\left(\theta^{(1)}\right)=2\left\{\ell\left(\hat{\theta}_0\right)-\ell_p\left(\theta^{(1)}\right)\right\}\dot{\sim}\chi^2_k.$$

Esta metodología puede ser aplicada cuando la inferencia es requerida en alguna combinación de los parámetros. En particular, se pueden obtener intervalos de confianza para cualquier nivel de retorno especificado z_p. Esto requiere una re-parametrización del modelo GEV, así que z_p es uno de los parámetros del modelo, después del cual la log-verosimilitud perfil se obtiene mediante la maximización con respecto a los parámetros restantes de la forma usual. La reparametrización es directa:

$$\mu=z_p+\frac{\sigma}{\xi}\left[1-\left\{-\log\left(1-p\right)\right\}^{-\xi}\right], \tag{3.12}$$

así que el reemplazo de μ en (3.7) con (3.12) tiene el efecto de expresar el modelo GEV en términos de los parámetros (z_p,σ,ξ).

3.5.5 Pruebas al Modelo

A pesar de que es imposible justificar sin ninguna duda la validez de una extrapolación basada en un modelo GEV, puede llevarse a cabo una valoración con referencia a los datos observados. No es suficiente para justificar la extrapolación, pero es un pre-requisito razonable. En este sentido, los papeles de probabilidad y las gráficas de cuantiles pueden ser utilizados para visualizar la validez del modelo GEV, por lo cual ahora se describen dos formas de evaluar la bondad de ajuste mediante procedimientos gráficos.

Un papel de probabilidad es una comparación de las funciones de distribución empírica y ajustada. Con los máximos ordenados por bloques $z_{(1)}\leq z_{(2)}\leq\cdots\leq z_{(m)}$, la función de distribución empírica evaluada en $z_{(i)}$ está dada por

$$\widetilde{G}\left(z_{(i)}\right)=\frac{i}{m+1}.$$

Mediante la sustitución de los estimadores de los parámetros en (3.2), los estimadores correspondientes del modelo son

$$\widehat{G}\left(z_{(i)}\right)=\exp\left\{-\left[1+\hat{\xi}\left(\frac{z_{(i)}-\hat{\mu}}{\hat{\sigma}}\right)\right]^{-1/\hat{\xi}}\right\}.$$

Si el modelo GEV es adecuado, entonces

$$\widehat{G}\left(z_{(i)}\right)=\widetilde{G}\left(z_{(i)}\right)$$

para cada $i,$ por lo cual, un papel de probabilidad que consista en los puntos

$$\left\{\left(\widehat{G}(z_{(i)}),\widetilde{G}(z_{(i)})\right),i=1,2,...,m\right\},$$

debería de estar cerca de una línea recta de 45º. Cualquier desviación importante de esta linealidad es un indicativo de la falla del modelo GEV.

Una debilidad del papel de probabilidad para los modelos de valores extremos es que tanto $\widehat{G}(z_{(i)})$ como $\widetilde{G}(z_{(i)})$ se aproximan acotadamente a 1 a medida que $z_{(i)}$ se vuelve más grande, y es usualmente la precisión del modelo para grandes valores de z lo que constituye el principal cuestionamiento. Esto es, el papel de probabilidad proporciona la menor cantidad de información donde radica el interés principal. Esta deficiencia puede evitarse mediante la gráfica de cuantiles, que consiste en los puntos

$$\left\{\left(\widehat{G}^{-1}\left(\frac{i}{m+1}\right),z_{(i)}\right),i=1,...,m\right\}, \tag{3.13}$$

donde, de la expresión (3.10),

$$\widehat{G}^{-1}\left(\frac{i}{m+1}\right)=\hat{\mu}-\frac{\hat{\sigma}}{\hat{\xi}}\left[1-\left\{-\log\left(\frac{i}{m+1}\right)\right\}^{-\hat{\xi}}\right].$$

A mayores desviaciones en la linealidad de los puntos de la gráfica de cuantiles se tienen mayores indicios de que el modelo falle.

Como se ha discutido con anterioridad, la gráfica de nivel de retorno, consiste de una gráfica de

$$z_p=\mu-\frac{\sigma}{\xi}\left[1-\left\{-\log\left(1-p\right)\right\}^{-\xi}\right]$$

en contra de $\hat{y}_p=-\log(1-p)$ sobre una escala logarítmica, y es particularmente conveniente para interpretar los modelos de valores extremos. La cola de la distribución se comprime, por lo que las estimaciones para los niveles de retorno para grandes periodos de retorno se muestran, mientras que la linealidad de la gráfica en el caso de

que $\xi = 0$ proporciona una línea base en contra de la cual se pueda juzgar el efecto del parámetro de forma estimado.

Una forma de resumir el modelo ajustado consiste en considerar la gráfica de nivel de retorno consistente en el lugar geométrico de los puntos

$$\left\{(\log y_p, \hat{z}_p) : 0 < p < 1\right\},$$

donde $\hat{z}_p$ es el estimador de máxima verosimilitud de z_p. Los intervalos de confianza pueden ser añadidos a tal gráfica para incrementar el grado de información. Los estimados empíricos de la función de nivel de retorno que se obtiene de los puntos (3.13), puede también ser adicionados, permitiendo que la gráfica de nivel de retorno pueda ser utilizada como una herramienta extra al diagnóstico del modelo. Si el modelo GEV es factible para los datos, la curva basada en el modelo y los estimadores empíricos deberían estar en concordancia razonable. Cualquier desviación sustancial o sistemática, después de admitir el error muestral, sugeriría una inadecuación del modelo GEV.

Las gráficas de probabilidad, cuantiles y niveles de retorno se basan en la comparación del los estimadores del modelo y los empíricos de la función de distribución. Por completitud, un diagnóstico equivalente basado en la función de densidad estaría asociado con la comparación de la función de densidad teórica con el histograma de los datos. Este procedimiento es generalmente menos informativo que las gráficas anteriores, debido a que la forma de un histograma puede variar sustancialmente con la selección de los intervalos de agrupamiento. Esto es, en contraste con la función de distribución empírica, no existe un estimador empírico único de la función de densidad, haciendo que la comparación con el estimador basado en el modelo sea difícil y subjetiva.

3.6 GENERALIZACIÓN DEL MODELO: EL MODELO DEL MAYOR ESTADÍSTICO DE ORDEN r

3.6.1 Formulación del modelo

Una dificultad implícita en cualquier análisis de valores extremos es la cantidad limitada de datos para la estimación del modelo. Los extremos son escasos, por definición, por lo que los estimadores del modelo, especialmente los niveles de retorno extremos, tienen una gran varianza. Este hecho ha motivado la investigación de la caracterización del comportamiento de los valores extremos que permitan la modelación de otros datos además de los extremos.

Existen dos caracterizaciones generales bien conocidas. Una se basa en el nivel de excedencia de un umbral grande y el otro se basa en el comportamiento del mayor estadístico de orden r dentro de un bloque, para valores pequeños de r. En esta sección se analiza el segundo enfoque.

Como en las secciones previas, supóngase que $X_1, X_2, \ldots$ es una sucesión de variables aleatorias independientes e idénticamente distribuidas, y se busca caracterizar el comportamiento extremo de las X_i. Con anterioridad se obtuvo que la distribución límite de M_n cuando $n \to \infty$, convenientemente re-escalado, es GEV. Primeramente se extiende este resultado a otros estadísticos de orden extremos, mediante la consideración de la siguiente notación:

$$M_n^{(k)} = \text{mayor } k - \text{ésimo de } \{X_1, \ldots, X_n\},$$

e identificando el comportamiento límite de esta variable, para un k fijo, cuando $n \to \infty$. El siguiente es un resultado que generaliza el teorema 3.1.

Teorema 3.5. Si existe una sucesión de constantes $\{a_n > 0\}$ y $\{b_n\}$ tales que

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to G(z) \text{ cuando } n \to \infty$$

para una función de distribución no-degenerada G, por lo que G es la función de distribución GEV dada por (3.2). Entonces, para un k fijo,

$$\Pr\left\{\frac{M_n^{(k)} - b_n}{a_n} \le z\right\} \to G_k(z)$$

definida sobre $\{z : 1 + \xi(z - \mu)/\sigma > 0\}$, donde

$$G_k(z) = \exp\{-r(z)\}\sum_{s=0}^{k-1}\frac{\tau(z)^s}{s!} \tag{3.14}$$

con

$$\tau(z) = \left[1 + \xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}. \qquad \blacksquare$$

El teorema 3.5 implica que si el $k-$ésimo mayor estadístico de orden en un bloque es normalizado en exactamente la misma forma como el de los máximos, entonces la distribución límite es de la forma (3.14) y los parámetros corresponden a los parámetros de la distribución límite GEV de los máximos por bloques. Nuevamente, absorbiendo las constantes de escala desconocidas en los parámetros de localización y escala del modelo se sigue que, para una n grande, la distribución aproximada de $M_n^{(k)}$ está dentro de la familia (3.14).

Existe, sin embargo, una dificultad al utilizar (3.14) como modelo. La situación radica en el hecho de que con frecuencia se tiene cada uno de los mayores estadísticos de

orden r dentro de cada uno de los diversos bloques, para algún valor de r. Esto es, usualmente se tiene un vector completo

$$\mathbf{M}_n^{(r)} = \left(M_n^{(1)},...,M_n^{(r)}\right)$$

para cada uno de los diversos bloques. Mientras que el teorema 3.4 proporciona una familia de distribuciones aproximadas de cada uno de los componentes de $\mathbf{M}_n^{(r)}$, no exhibe la distribución conjunta de $\mathbf{M}_n^{(r)}$. Más aún, los componentes no pueden ser independientes: $M_n^{(2)}$ no puede ser mayor que $M_n^{(1)}$, por ejemplo, por lo que la salida de cada componente influye en la distribución de la otra. Por lo tanto, el resultado del teorema 3.5 no proporciona un modelo para $\mathbf{M}_n^{(r)}$. En su lugar, se requiere una caracterización de la distribución conjunta límite del vector completo $\mathbf{M}_n^{(r)}$. Con un re-escalamiento apropiado esto puede ser llevado, pero la distribución conjunta límite no es sencilla. Sin embargo, el teorema siguiente proporciona la función de densidad conjunta de la distribución límite.

Teorema 3.6. Si existe una sucesión de constantes $\{a_n > 0\}$ y $\{b_n\}$ tal que

$$\Pr\left\{\frac{M_n - b_n}{a_n} \le z\right\} \to G(z) \quad \text{cuando} \quad n \to \infty$$

para alguna función de distribución no degenerada G, entonces, para un r fijo, la distribución límite cuando $n \to \infty$ de

$$\widetilde{\mathbf{M}}_n^{(r)} = \left(\frac{M_n^{(1)} - b_n}{a_n},...,\frac{M_n^{(r)} - b_n}{a_n}\right)$$

se encuentra dentro de la familia de probabilidad conjunta siguiente

$$f\left(z^{(1)},...,z^{(r)}\right) = \exp\left\{-\left[1+\xi\left(\frac{z^{(r)}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\} \times \prod_{k=1}^{r} \sigma^{-1}\left[1+\xi\left(\frac{z^{(k)}-\mu}{\sigma}\right)\right]^{-1/\xi-1}, \quad (3.15)$$

donde:

$-\infty < \mu < \infty;\ \sigma > 0;\ -\infty < \xi < \infty;\ z^{(r)} \le z^{(2)} \le \cdots \le z^{(1)};\ z^{(k)}: 1+\xi(z^{(k)}-\mu)/\sigma > 0$ para $k = 1,...,r.$ ■

En el caso en que $r = 1$, (3.15) se reduce a la familia GEV de funciones de densidad. El caso $\xi = 0$ en (3.15) se interpreta como la forma límite cuando $\xi \to 0$, que lleva a considerar la siguiente familia de funciones de densidad

$$f\left(z^{(1)},...,z^{(r)}\right)=\exp\left\{-\exp\left[-\left(\frac{z^{(r)}-\mu}{\sigma}\right)\right]\right\}\times\prod_{k=1}^{r}\sigma^{-1}\exp\left[-\left(\frac{z^{(k)}-\mu}{\sigma}\right)\right], \qquad (3.16)$$

en la cual, en el caso en que $r=1,$ se reduce a la familia Gumbel.

3.6.2 Modelación de los mayores estadísticos de orden r

Comenzando con una serie de variables independientes e idénticamente distribuidas, se supondrá que los datos están agrupados en m bloques. En el bloque i se registran las observaciones más grandes, que lleva a la consideración de las series $\mathbf{M}_i^{(r_i)}=(z_i^{(1)},...,z_i^{(r_i)}),$ para $i=1,...,m.$ Es usual asignar $r_1=\cdots=r_m=r$ a algún valor especificado $r,$ a pesar de que existan menos datos en algunos bloques.

Como en el modelo GEV, el variar los tamaños de los bloques implica un intercambio entre el sesgo y la varianza el cual usualmente se resuelve mediante una selección apropiada del tal tamaño, como asociar el tamaño de los bloques a un año. El número de estadísticos de orden utilizados en cada bloque tiene también influencia en el intercambio entre el sesgo y la varianza: valores pequeños de r generan pocos datos con gran varianza; valores grandes de r son probables que violen el soporte asintótico del modelo, que deriva en sesgo. En la práctica es usual seleccionar las r_i tan grandes como sea posible, sujeto a la adecuación del diagnóstico del modelo.

La verosimilitud para este modelo se obtiene de (3.15) y (3.16), mediante el procedimiento usual de la absorción de los coeficientes desconocidos de localización y escala, y tomando productos a través de los bloques. Así, cuando $\xi\neq 0,$

$$L(\mu,\sigma,\xi)=\prod_{i=1}^{m}\left(\exp\left\{-\left[1+\xi\left(\frac{z_i^{(r_i)}-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}\times\prod_{k=1}^{r_i}\sigma^{-1}\left[1+\xi\left(\frac{z_i^{(k)}-\mu}{\sigma}\right)\right]^{-1/\xi-1}\right), \qquad (3.17)$$

bajo el supuesto de que $1+\xi(z^{(k)}-\mu)/\sigma>0,\ \ k=1,...,r_i,\ \ i=1,...,m;$ de otra forma la verosimilitud es cero. Cuando $\xi=0,$

$$L(\mu,\sigma,\xi)=\prod_{i=1}^{m}\left(\exp\left\{-\exp\left[-\left(\frac{z_i^{(r_i)}-\mu}{\sigma}\right)\right]\right\}\times\prod_{k=1}^{r_i}\sigma^{-1}\exp\left[-\left(\frac{z_i^{(k)}-\mu}{\sigma}\right)\right]\right). \qquad (3.18)$$

La verosimilitud (3.17) y (3.18) o, más comúnmente, la correspondiente log-verosimilitud, puede ser maximizada numéricamente para obtener los estimadores de máxima verosimilitud. La teoría estándar de verosimilitud asintótica proporciona a su vez errores estándar aproximados e intervalos de confianza. En el caso especial en que $r_i=1$ para cada $i,$ la función de verosimilitud se reduce a la verosimilitud del modelo GEV para máximos por bloques. De manera más general, el modelo de los mayores estadísticos de orden r provee una verosimilitud cuyos parámetros corresponden a aquellos de la distribución GEV del máximo por bloques, pero el cual incorpora más de

los datos extremos observados. Así, relativo al análisis estándar de máximos por bloques, la interpretación de los parámetros no se altera, pero la precisión debería de mejorarse debido a la inclusión de la información extra.

3.7 MODELOS DE UMBRAL

3.7.1 Consideraciones generales

Como se ha discutido hasta el momento, la sola modelación de los máximos por bloques es un enfoque que pierde información valiosa cuando se tienen más observaciones que únicamente los extremos. A pesar de que el modelo estadístico de mayor orden r es una mejor alternativa, no es usual el contar con datos dados en esta forma. Sin embargo, aún este método puede perder información valiosa de los datos si, por ejemplo, en uno de los bloques contiene más eventos extremos que algún otro. Si una serie de tiempo de longitud de una hora o un día está disponible, entonces sería mejor utilizar los datos quitando completamente el procedimiento de conformar bloques.

Sean $X_1, X_2, \ldots$ una sucesión de variables aleatorias independientes e idénticamente distribuidas, que tienen una función de distribución marginal F. Es natural considerar los eventos extremos como aquellos X_i que exceden un cierto umbral u. Denotando un término arbitrario X_i por X, se sigue que una descripción del comportamiento aleatorio de los eventos extremos se encuentra dado por la probabilidad condicional

$$\Pr\{X > u + y \,|\, X > u\} = \frac{1 - F(u+y)}{1 - F(u)}, \quad y > 0. \tag{3.19}$$

Si la distribución padre F es conocida, la distribución de las excedencias de umbral en (4.1) también lo sería. Debido a que en las aplicaciones prácticas este no es el caso, se realizan ciertas aproximaciones que sean ampliamente aplicables para valores grandes del umbral. Esto es similar al uso del modelo GEV como una aproximación de la distribución de máximos de una sucesión grande cuando la población padre es desconocida.

3.7.2 La Distribución Generalizada de Pareto

El resultado principal se encuentra en el siguiente teorema.

Teorema 3.7. Sea $X_1, X_2, \ldots$ una sucesión de variables aleatorias independientes con función de de distribución común F, y sea

$$M_n = \max\{X_1, \ldots, X_n\}.$$

Denótese un término arbitrario en las X_i por X, y supóngase que F satisface el teorema 3.1.1, aquí que para una n grande,

$$\Pr\{M_n \leq z\} \approx G(z),$$

donde

$$G(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

para algunos μ, $\sigma > 0$ y ξ. Entonces, para una μ lo suficientemente grande, la función de distribución de $(X-\mu)$, condicionada sobre $X > u$, es aproximadamente

$$H(y) = 1-\left(1+\frac{\xi y}{\tilde{\sigma}}\right)^{-1/\xi} \tag{3.20}$$

definida sobre $\{y : y > 0 \text{ y } (1+\xi y/\tilde{\sigma}) > 0\}$, donde

$$\tilde{\sigma} = \sigma + \xi(u-\mu). \tag{3.21}$$

El teorema 3.7 puede considerarse realizado de manera más precisa justificando (3.20) como una distribución límite a medida que u se incrementa. En la siguiente sección se proporciona un bosquejo de la demostración de este teorema.

La familia de distribuciones definida por la ecuación (3.20) se llama la **familia generalizada de Pareto**. El teorema 3.7 implica que si el máximo por bloques tiene una distribución aproximada G, entonces la excedencia de umbral tiene una distribución correspondiente aproximada dentro de la familia generalizada de Pareto. Sin embargo, los parámetros de la distribución generalizada de Pareto de la excedencia de umbral se encuentran determinados de manera única por aquellos asociados con la distribución GEV de los máximos por bloques. En particular, el parámetro ξ es igual a su correspondiente en la distribución GEV. Seleccionando un tamaño de bloque n diferente, pero aún grande, afectaría los valores de los parámetros GEV, pero no los correspondientes a la distribución generalizada de Pareto de las excedencias de umbral: ξ es invariante al tamaño de bloque, mientras que los cálculos de $\tilde{\sigma}$ en (3.21) no se perturba mediante cambios en μ y σ los cuales se auto-compensan.

La dualidad entre el modelo GEV y la familia generalizada de Pareto significa que el parámetro de forma ξ es dominante en la determinación del comportamiento cualitativo de la distribución generalizada de Pareto, justamente como en la distribución GEV. Si $\xi < 0$ la distribución de excesos tiene una cota superior de $u - \tilde{\sigma}/\xi$; si $\xi > 0$ la distribución no tiene límite superior. La distribución no está acotada si $\xi = 0$, la cual debe interpretarse nuevamente como el proceso de tomar el límite $\xi \to 0$ en (3.20), lo que implica que

$$H(y) = 1-\exp\left(-\frac{y}{\tilde{\sigma}}\right), \quad y > 0, \tag{3.22}$$

correspondiente a una distribución exponencial con parámetro $1/\widetilde{\sigma}$.

3.7.3 Bosquejo de la justificación del modelo generalizado de Pareto

Esta sección proporciona un bosquejo de la prueba del teorema 3.7.

Sea X una función de distribución F. Mediante el supuesto del teorema 3.7, para una n grande,

$$F^n(z) = \exp\left\{-\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}\right\}$$

para algunos parámetros $\mu,\ \sigma>0$ y ξ. Por lo tanto,

$$n\log F(z) \approx -\left[1+\xi\left(\frac{z-\mu}{\sigma}\right)\right]^{-1/\xi}. \tag{3.23}$$

Pero para valores grandes de $z,$ una expansión de Taylor implica que

$$\log F(z) \approx -\{1-F(z)\}.$$

La sustitución en (3.23), seguida de una re-arreglo de términos, implica que

$$1-F(u) \approx \frac{1}{n}\left[1+\xi\left(\frac{u-\mu}{\sigma}\right)\right]^{-1/\xi}$$

para algún $\mu,\sigma\xi$ grande. De manera similar, para $y>0,$

$$1-F(u+y) \approx \frac{1}{n}\left[1+\xi\left(\frac{u+y-\mu}{\sigma}\right)\right]^{-1/\xi}. \tag{3.24}$$

Por ende,

$$\begin{aligned}\Pr\{X>u+y\,|\,X>u\} &\approx \frac{n^{-1}[1+\xi(u+y-\mu)/\sigma]^{-1/\xi}}{n^{-1}[1+\xi(u-\mu)/\sigma]^{-1/\xi}}\\ &= \left[1+\frac{\xi(u+y-\mu)/\sigma}{1+\xi(u-\mu)/\sigma}\right]^{-1/\xi}\\ &= \left[1+\frac{\xi y}{\widetilde{\sigma}}\right]^{-1/\xi},\end{aligned} \tag{3.25}$$

donde

$$\widetilde{\sigma} = \sigma+\xi(u-\mu),$$

como se requería.

3.7.4 Ejemplos

Se consideran ahora los tres ejemplos teóricos analizados en la sección 3.3.5 en términos de los modelos de excedencias de umbral.

Ejemplo 3.4. Para el modelo exponencial, $F(x)=1-\exp(-x)$, para $x>0$. Mediante un cálculo directo, se tiene que

$$\frac{1-F(u+y)}{1-F(u)}=\frac{e^{-(u+y)}}{e^{-u}}=e^{y}$$

para todo $y>0$. Consecuentemente, la distribución límite de las excedencias de umbral es la distribución exponencial, correspondiente al caso en que $\xi=0$ y $\tilde{\sigma}=1$ en la familia generalizada de Pareto. Además, este es un resultado exacto para todos los umbrales tales que $u>0$.

Ejemplo 3.5. Para el modelo Fréchet estándar, $F(x)=\exp(-1/x)$, para $x>0$. Por lo tanto,

$$\frac{1-F(u+y)}{1-F(u)}=\frac{1-\exp\{-(u+y)^{-1}\}}{1-\exp(-u^{-1})}\sim\left(1+\frac{y}{u}\right)^{-1}$$

cuando $u\to\infty$, para todo $y>0$. Esto corresponde a la distribución generalizada de Pareto con $\xi=1$ y $\tilde{\sigma}=u$.

Ejemplo 3.6. Para el modelo de distribución uniforme $U(0,1)$, $F(x)=x$, para $0\le x\le 1$. Por lo tanto,

$$\frac{1-F(u+y)}{1-F(u)}=\frac{1-(u+y)}{1-u}=1-\frac{y}{1-u}$$

para $0\le y\le 1-u$. Esto corresponde a la distribución generalizada de Pareto con $\xi=-1$ y $\tilde{\sigma}=1-u$.

Una comparación de las familias límite que se obtuvieron aquí para las excedencias de umbral con los límites de bloques máximos correspondientes visto con anterioridad confirma la dualidad de las dos formulaciones de modelos límite implicado en el teorema 3.7. En particular, los valores de ξ son comúnes en los dos modelos.

Además, el valor de $\tilde{\sigma}$ es dependiente del umbral, excepto en el caso donde el modelo límite tiene $\xi=0$, como se muestra en la expresión 3.21.

En este punto se ha utilizado la notación $\tilde{\sigma}$ para denotar el parámetro de escala de la distribución generalizada de Pareto, así como para distinguirlo de su parámetro

correspondiente en la distribución GEV. Por conveniencia notacional, en este punto se disuelve esta distinción, utilizando σ para denotar el parámetro de escala en cualquiera de las dos familias.

3.8 MODELACIÓN DE LAS EXCEDENCIAS DE UMBRAL

3.8.1 Selección del umbral

El teorema 3.7 sugiere el siguiente marco conceptual para la modelación de valores extremos. Los datos originales consisten en una sucesión de mediciones independientes e idénticamente distribuidas $x_1,...,x_n$. Los eventos extremos son identificados mediante la definición de un umbral grande u, para el cual las excedencias son $\{x_i : x_i > u\}$. Estas cantidades se denotan por $x_{(1)},...,x_{(k)}$, y se definen de manera precisa mediante $y_j = x_{(j)} - u$, para $j = 1,...,k$. Por el teorema 3.7, las y_j podrían ser consideradas como realizaciones independientes de una variable aleatoria cuya distribución puede ser aproximada por un miembro de la familia generalizada de Pareto. La inferencia consiste en ajustar la familia generalizada de Pareto a las excedencias de umbral observadas, seguida de una verificación del modelo así como la extrapolación.

Este enfoque contrasta con el de máximos por bloques por la caracterización de una observación como extrema si esta excede un umbral alto. Pero la consideración de la selección de un umbral es análoga a la selección del tamaño de bloque en el enfoque de máximos por bloques, implicando un balance entre el sesgo y la varianza. En este caso, un umbral muy pequeño es probable que viole el sesgo asintótico del modelo, dando como resultado un gran sesgo; un umbral muy grande generará pocas excedencias sobre las cuales pudiera estimarse el modelo, generando entonces una gran varianza. El enfoque estándar consiste en adoptar un umbral tan pequeño como sea posible, sujeto a que el modelo limite proporcione una aproximación razonable. Existen dos métodos disponibles para este propósito: uno consiste en una técnica exploratoria llevada a cabo de manera previa a la estimación del modelo; la otra consiste en el aseguramiento de la estabilidad de los estimadores del parámetro, basada en el ajuste de modelos a través de un rango de umbrales diferentes.

En un mayor detalle, el primer método se fundamenta en la media de una distribución generalizada de Pareto. Si Y tiene una distribución generalizada de Pareto con parámetros σ y ξ, entonces

$$E(Y) = \frac{\sigma}{1-\xi}, \tag{3.26}$$

suponiendo que $\xi < 1$. Cuando $\xi \geq 1$ la media es infinita. Ahora, supóngase que la distribución generalizada de Pareto es válida como un modelo para las excedencia de un umbral u_0 generado por una sucesión $X_1,...,X_n$, de las cuales un término arbitrario se denota por X. Por la expresión (3.26),

$$E(X-u_0 \mid X>u_0)=\frac{\sigma_{u_0}}{1-\xi},$$

suponiendo que $\xi<1$, donde se adopta el convenio de utilizar σ_{u_0} para denotar el parámetro de escala correspondiente a las excedencia del umbral u_0. Pero si la distribución generalizada de Pareto es válida para las excedencia del umbral u_0, sería igualmente válida para todos los umbrales $u>u_0$, sujeto a los cambios de escala apropiados de σ_u. Por lo tanto, para $u>u_0$,

$$E(X-u \mid X>u)=\frac{\sigma_u}{1-\xi}=\frac{\sigma_{u_0}+\xi u}{1-\xi} \tag{3.27}$$

por (3.21). Así, para $u>u_0$, $E(X-u \mid X>u)$ es una función lineal de u. Además, $E(X-\mu \mid X>u)$ es simplemente la media de las excedencias del umbral u, para el cual la media muestral de las excedencias del umbral de u proporcionan una estimación empírica. De acuerdo a (3.27), se espera que estas estimaciones cambien linealmente con respecto a u, en los niveles de u para los cuales el modelo generalizado de Pareto es apropiado. Esto lleva a la consideración del siguiente procedimiento. El lugar geométrico de los puntos

$$\left\{\left(u,\frac{1}{n_u}\sum_{i=1}^{n_u}(x_{(i)}-u)\right):u<x_{\max}\right\},$$

se llama la **gráfica de los residuales promedios de vida**, y donde $x_{(1)},...,x_{(n)}$ consiste de las n_u observaciones que exceden a u, y $x_{\max}$ es el más grande de los X_i. Por arriba de un umbral u_0 en el cual la distribución generalizada de Pareto proporciona una aproximación adecuada de la distribución de excedencias, la gráfica de los residuales promedios de vida debería ser aproximadamente lineal con respecto a u. Por supuesto, se pueden añadir intervalos de confianza a esta gráfica, los cuales se basen en la normalidad aproximada de las medias muestrales.

3.8.2 Estimación de los parámetros

Una vez determinado el umbral, los parámetros de la distribución generalizada de Pareto pueden ser estimados mediante máxima verosimilitud. Supóngase que los valores $y_1,...,y_k$ son las k excedencias de un umbral u. Para $\xi\neq 0$ la log-verosimilitud se deriva de (3.20) como

$$\ell(\sigma,\xi)=-k\log\sigma-(1+1/\xi)\sum_{i=1}^{k}\log(1+\xi y_i/\sigma), \tag{3.28}$$

suponiendo que $(1+\sigma^{-1}\xi y_i)>0$ para $i=1,...,k;$ de otra forma, $\ell(\sigma,\xi)=-\infty.$ En el caso en que $\xi=0$ la log-verosimilitud se obtiene de la expresión (3.22) como

$$\ell(\sigma)=-k\log\sigma-\sigma^{-1}\sum_{i=1}^{k}y_i.$$

La maximización analítica de la log-verosimilitud no es posible, así que nuevamente se requieren técnicas numéricas, debiéndose tomar en consideración el hecho de evitar inestabilidades cuando $\xi=0$ en la expresión (3.28), y asegurar que el algoritmo no falle debido a una evaluación fuera del espacio paramétrico permisible. Los errores estándar y los intervalos de confianza de la distribución generalizada de Pareto se obtienen de la forma usual de la teoría estándar de verosimilitud.

3.8.3 Niveles de Retorno

Como se ha discutido con anterioridad, es usualmente más conveniente el hecho de interpretar los modelos de valores extremos en términos de los cuantiles o niveles de retorno, más que en términos de los valores paramétricos individuales. Así, supóngase que la distribución generalizada de Pareto con parámetros σ y ξ es un modelo razonable para las excedencias de un umbral u de una cierta variable X. Esto es, para $x>u,$

$$\Pr\{X>x\,|\,X>u\}=\left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}.$$

Se sigue que

$$\Pr\{X>x\}=\varsigma_u\left[1+\xi\left(\frac{x-u}{\sigma}\right)\right]^{-1/\xi}, \tag{3.29}$$

donde $\varsigma_u=\Pr\{X>u\}.$ Por lo tanto, el nivel x_m es la excedencia de la media dado que cada una de las m observaciones es solución de

$$\varsigma_u\left[1+\xi\left(\frac{x_m-u}{\sigma}\right)\right]^{-1/\xi}=\frac{1}{m}. \tag{3.30}$$

Reordenando,

$$x_m=u+\frac{\sigma}{\xi}\left[(m\varsigma_u)^{\xi}-1\right], \tag{3.31}$$

suponiendo que m es suficientemente grande para asegurar que $x_m>u.$ Todo esto asume que $\xi\neq 0.$ Si $\xi=0,$ procediendo de la misma forma como en (3.22) lleva a la consideración de

$$x_m=u+\sigma\log(m\varsigma_u), \tag{3.32}$$

suponiendo nuevamente que m es lo suficientemente grande.

Por construcción, x_m es el **nivel de retorno de las** m **observaciones**. De (3.31) y (3.32), si se grafica x_m en contra de m sobre una una escala logarítmica, se produce el mismo hecho cualitativo como en las gráficas de niveles de retorno del modelo GEV: linealidad si $\xi = 0$; concavidad si $\xi > 0$; convexidad si $\xi < 0$. Por presentación, es frecuentemente más conveniente el proporcionar niveles de retorno sobre una escala anual, así que el nivel de retorno de tamaño N es tal que el nivel esperado sea excedido cada N años. Si existen n_y observaciones por año, esto corresponde a un nivel de retorno de m observaciones, donde $m = N \times n_y$. Por lo tanto, el nivel de retorno de N años se define por

$$z_N = u + \frac{\sigma}{\xi}\left[\left(Nn_y\varsigma_u\right)^{\xi} - 1\right],$$

donde $\xi = 0,$ en cuyo caso

$$z_N = u + \sigma \log(Nn_y\varsigma_u).$$

La estimación de los niveles de retorno requiere la sustitución de los valores de los parámetros por sus estimadores. Para σ y ξ esto corresponde a la sustitución por los correspondientes estimadores de máxima verosimilitud, pero en el caso de ς_u, también es necesario conocer la probabilidad de que una observación individual sobrepase el umbral. Existe un estimador natural:

$$\hat{\varsigma}_u = \frac{k}{n},$$

que es la proporción muestral de los puntos que exceden a u. Debido a que el número de excedencias de u sigue una distribución binomial $\text{Bin}(n, \varsigma_u)$, ς_u también es un estimador de máxima verosimilitud de ς_u.

Los errores estándar o intervalos de confianza para x_m pueden derivarse por el método delta, pero la incertidumbre del estimador de ς_u también debería ser incluida en los cálculos. De las propiedades de la distribución binomial, $\text{var}(\hat{\varsigma}_u) \approx \hat{\varsigma}_u(1-\hat{\varsigma}_u)/n,$ por lo que la matriz completa de varianza-covarianza para $(\hat{\varsigma}_u, \hat{\sigma}, \hat{\xi})$ es aproximadamente

$$V = \begin{bmatrix} \hat{\varsigma}_u(1-\hat{\varsigma}_u)/n & 0 & 0 \\ 0 & v_{1,1} & v_{1,2} \\ 0 & v_{2,1} & v_{2,2} \end{bmatrix},$$

donde $v_{i,j}$ denota el (i, j) término de la matriz de varianza-covarianza de $\hat{\sigma}$ y $\hat{\xi}$. Por lo tanto, por el método delta,

$$\text{var}(\hat{x}_m) = \nabla x_m^T V \nabla x_m,$$

donde

$$\nabla x_m^T = \left[\frac{\partial x_m}{\partial \varsigma_u}, \frac{\partial x_m}{\partial \sigma}, \frac{\partial x_m}{\partial \xi}\right] = [\sigma m^{\xi} \varsigma_u^{\xi-1}, \xi^{-1}\{(m\varsigma_u)^{\xi} - 1\},$$
$$-\sigma\xi^{-2}\{(m\varsigma_u)^{\xi} - 1\} + \sigma\xi^{-1}(m\varsigma_u)^{\xi}\log(m\varsigma_u)],$$

evaluado en $(\hat{\varsigma}_u, \hat{\sigma}, \hat{\xi})$.

Como en los modelos previos, los mejores estimadores de precisión para los parámetros y niveles de retorno se obtienen mediante la verosimilitud perfil apropiada. Para σ o ξ esto es directo. Para los niveles de retorno, es requerida una reparametrización. Se simplifican las cosas si se ignora la incertidumbre en ς_u, la cual es usualmente pequeña relativa a los otros parámetros. De (3.31) y (3.32)

$$\sigma = \begin{cases} \dfrac{(x_m - u)\xi}{(m\varsigma_u)^{\xi} - 1}, & \text{si } \xi \neq 0; \\ \dfrac{x_m - u}{\log(m\varsigma_u)}, & \text{si } \xi = 0. \end{cases}$$

Con un x_m fijo, la sustitución en (3.28) proporciona una verosimilitud uni-paramétrica que puede ser maximizada con respecto a ξ. Como una función de x_m, esta es la log-verosimilitud perfil del nivel de retorno $m-$observacional.

3.8.4 Revisión de la selección del umbral

Se ha visto que la grafica de los residuos promedios de vida puede ser difícil de interpretar como un método para la selección del umbral. Una técnica complementaria consiste en ajustar la distribución generalizada de Pareto sobre un rango de umbrales y buscar la estabilidad de los estimadores de los parámetros. El argumento es como sigue.

Por el teorema 3.7, si una distribución generalizada de Pareto es un modelo razonable para la excedencia de un umbral u_0, entonces la excedencia de un umbral más alto u debería también seguir una distribución generalizada de Pareto. Los parámetros de forma de las dos distribuciones son idénticos. Sin embargo, denotando por σ_u al valor del parámetro de escala de la distribución generalizada de Pareto para un umbral $u > u_0$, se sigue de (3.21) que

$$\sigma_u = \sigma_{u_0} + \xi(u - u_0), \tag{3.34}$$

por lo que el parámetro de escala cambia con u a menos que $\xi = 0$. Esta dificultad puede ser resuelta mediante una reparametrización de los parámetros de escala generalizados de Pareto como

$$\sigma^* = \sigma_u - \xi u,$$

el cual es constante con respecto a u por la expresión (3.34). Por lo tanto, los estimadores tanto de σ^* y ξ deberían ser constantes por encima de u_0, si u_0 es un umbral válido para las excedencias que concuerda con la distribución generalizada de Pareto. La variabilidad muestral significa que los estimadores de estas cantidades no son exactamente constantes, pero deberían ser estables después de la consideración de sus errores muestrales.

Este argumento sugiere graficar tanto $\widehat{\sigma}^*$ y $\hat{\xi}$ en contra de u, junto con los intervalos de confianza de cada una de estas cantidades, y seleccionando a u_0 como el menor valor de u para el cual los estimadores están cercanos a ser constantes. Los intervalos de confianza para $\hat{\xi}$ se obtienen inmediatamente de la matriz de varianza-covarianza V. Los intervalos de confianza para $\widehat{\sigma}^*$ requiere la utilización del método delta, utilizando

$$\text{var}(\sigma^*) \approx \nabla \sigma^{*T} V \nabla \sigma^*,$$

donde

$$\nabla \sigma^{*T} = \left[\frac{\partial \sigma^*}{\partial \sigma_u}, \frac{\partial \sigma^*}{\partial \xi} \right] = [1, -u].$$

3.8.5 Diagnóstico del modelo

Los papeles de probabilidad, las gráficas de cuantiles, las gráficas de niveles de retorno y las gráficas de densidad son útiles para establecer la calidad del modelo generalizado de Pareto. Suponiendo un umbral u, las excedencias de umbral $y_{(1)} \leq \cdots \leq y_{(k)}$ y un modelo estimado $\widehat{H}$, el papel de probabilidad consiste de los pares

$$\{(i/(k+1), \widehat{H}(y_{(i)})) : i = 1, ..., k\},$$

donde

$$\widehat{H}(y) = 1 - \left(1 + \frac{\hat{\xi} y}{\widehat{\sigma}}\right)^{-1/\xi},$$

suponiendo que $\hat{\xi} \neq 0$. Si $\hat{\xi} = 0$ es la gráfica se construye utilizando (3.22) en lugar de (3.24). Nuevamente, suponiendo $\hat{\xi} \neq 0$, la gráfica de cuantiles consiste en los pares

$$\{(\widehat{H}^{-1}(i/(k+1)), y_{(i)}) : i = 1, ..., k\},$$

donde

$$\widehat{H}^{-1}(y) = u + \frac{\widehat{\sigma}}{\widehat{\xi}}\left[y^{-\hat{\xi}} - 1 \right].$$

Si el modelo generalizado de Pareto es razonable para modelar las excedencias de u, entonces tanto las gráficas de probabilidad y de cuantiles deberían de consistir de puntos que fuesen aproximadamente lineales.

Un nivel de retorno consiste en el lugar geométrico de los puntos $\{(m, \hat{x}_m)\}$ para valores grandes de m, donde $\hat{x}_m$ es el estimador del nivel de retorno $m-$observacional

$$\hat{x}_m = u + \frac{\widehat{\sigma}}{\widehat{\xi}}\left[(m\hat{\varsigma}_u)^{\hat{\xi}} - 1 \right],$$

nuevamente modificado si $\hat{\xi} = 0$. Como en la gráfica de nivel de retorno GEV, es usual el graficar la curva de nivel de retorno sobre una escala logarítmica para enfatizar el efecto de la extrapolación, así como adicionar las cotas de confianza y los estimadores empíricos de los niveles de retorno.

Finalmente, la función de densidad del modelo generalizado ajustado de Pareto puede ser comparado con el histograma de las excedencias de umbral.

Con esto se da por finalizado el presente capítulo, con lo que se concluye la revisión documental y de marco teórico, para ahora proceder a la parte de las aplicaciones y análisis comparativo de los modelos a considerarse en el siguiente capítulo.

CAPÍTULO 4
APLICACIÓN DE LA TEORÍA CLÁSICA DE AJUSTES Y SU COMPARACIÓN CON LA TEORÍA DE VALORES EXTREMOS

Experiment!
Make it your motto day and night
Experiment,
And it will lead you to the light.

The apple on the top of the tree
Is never too high to achieve,
So take and example from Eve…
Experiment!

Be curious,
Though interfering friends may frown.
Get furious…

…At each attempt to hold you down.
If this advice you only employ,
The future can offer you infinite joy
And merriment…
Experiment
And you'll see!

Cole Porter (1891-1964)

4.1 RESUMEN CAPITULAR

En este capítulo se presenta el núcleo del trabajo, que consiste en la aplicación de diversas metodologías con la finalidad de conformar un marco comparativo de los modelos beta generalizado del segundo tipo, el lognormal y los distintos modelos de valores extremos presentados en los capítulos 2 y 3, respectivamente.

4.2 PRESENTACIÓN DE LAS BASES DE DATOS DE TRABAJO

Las bases de datos que son el insumo para los análisis posteriores se tomaron de la información proporcionada por la ENIGH 2004, la cual, como se había comentado, en uno de sus productos (conocido como el "Concentrado por hogar"), se tiene una base de datos de un poco más de 25,000 registros que corresponden a los hogares que se tomaron en muestra en dicho año. A su vez, contiene 114 variables que corresponden a los diversos conceptos captados por la encuesta. De estos, para los fines que tiene el presente trabajo, se analiza únicamente el ingreso el corriente. Esto es así puesto que es la variable que se utiliza más frecuentemente para diversos análisis, como estudios de pobreza y desigualdad (sin embargo, el análisis puede llevarse a cabo sobre casi cualquier otra variable de la encuesta).

Para una definición precisa de esta variable, así como de otros tipos de ingresos, véase el apéndice "A".

En los siguientes apartados se comenzará el análisis de las técnicas presentadas en los capítulos 2 y 3.

4.3 MODELACIÓN DE VALORES EXTREMOS

En las siguientes sub-secciones se presentan los resultados de la aplicación del ajuste de los modelos de valores extremos al ingreso corriente comentado con anterioridad (conviene aclarar que los datos de los ingresos se analizan de manera trimestral, por lo que hacen dos cálculos previos: una división entre tres para mensualizarlos, y otra división entre 1000, para presentar los datos en montos de miles de pesos).

4.3.1 Análisis de datos muestrales

En este caso, como en los que siguen, las bases de datos tienen 32 registros (que corresponden a los 32 máximos de los ingresos corrientes por entidad federativa). Los estimadores de máxima verosimilitud son:

$$\hat{\mu} = 96.7472019,\ \hat{\sigma} = 45.9867624,\ \ \hat{\xi} = 0.7335034.$$

En cuyo caso la log-verosimilitud maximizada es igual a 186.2534. La matriz aproximada de varianza-covarianza de los estimadores de los parámetros es:

$$V = \begin{bmatrix} 89.5089566 & 80.9045078 & -0.48524169 \\ 80.9045078 & 110.9888776 & 0.37863181 \\ -0.4852417 & 0.3786318 & 0.04618974 \end{bmatrix}$$

La diagonal de esta matriz corresponden a las varianzas de los parámetros individuales (μ, σ, ξ). Tomando las raíces cuadradas, los errores estándar son:

$$SE_{\hat{\mu}} = 9.4609173,\ SE_{\hat{\sigma}} = 10.5351259,\ SE_{\hat{\xi}} = 0.2149180.$$

Combinando los valores de los estimadores con estos errores estándar, se tienen los siguientes intervalos de confianza aproximados del 95% de confianza. Para $\hat{\mu}$: [78.2038,115.2906]; para $\hat{\sigma}$: [25.33792,66.63561]; para $\hat{\xi}$: [0.3122642,1.154743]. Es de notarse que el estimador para el parámetro de forma, ξ, adquiere un valor positivo, y el intervalo de confianza no toma como valor al cero, con lo cual se concluye, hasta esto momento, que los estimadores de máxima verosimilitud son regulares, entendiendo con ello que guardan las propiedades asintóticas usuales, y a su vez, se tiene una fuerte evidencia de que la distribución de los ingresos corrientes es acotada. A su vez, se observa que este modelo corresponde al caso Fréchet. Ahora, para obtener un intervalo de confianza más fino, se puede utilizar la verosimilitud perfil. La figura 4.1 muestra la forma de la verosimilitud perfil para ξ, de la cual se obtiene que un intervalo de confianza del 95% es [0.3750751,1.2328328], el cual no difiere demasiado del intervalo encontrado anteriormente, aunque vale la pena apreciar que se encuentra más separado del origen.

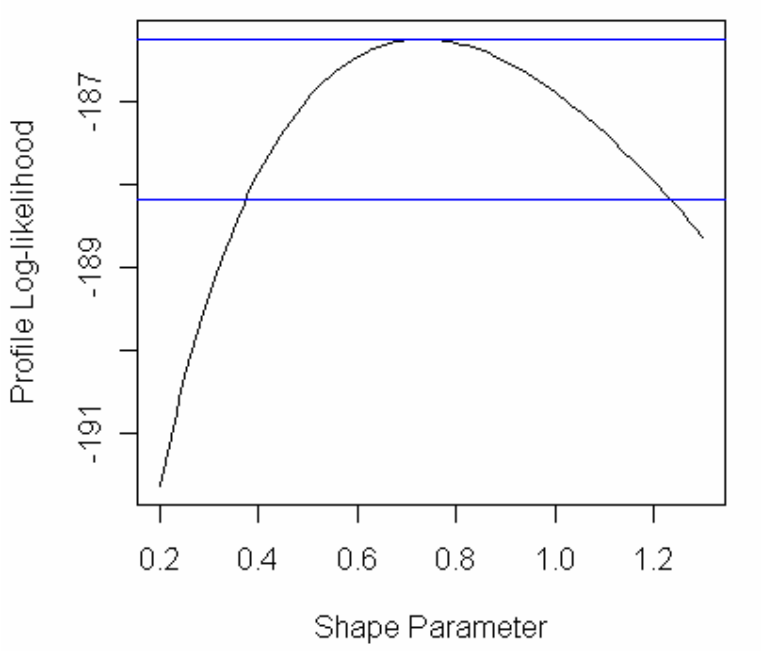

Figura 4.1 *Verosimilitud perfil para ξ del ingreso corriente muestral*

Ahora, los estimadores e intervalos de confianza de los niveles de retorno se obtienen mediante la sustitución de $(\hat{\mu},\hat{\sigma},\hat{\xi})$ en (3.10) y (3.11). A continuación se presentan los resultados en la tabla 4.1 la cual muestra una lista de valores seleccionados de p iguales a $1/10,\ 1/20,\ 1/50,\ 1/100,\ 1/1000,$ para un un nivel de confianza del 95%, donde SD es la desviación estándar, LI y LS significa límite inferior y superior, respectivamente, de los intervalos de confianza.

Tabla 4.1 *Tabla de los valores de retornos y sus intervalos de confianza para diferentes valores de* p

p	z_p	var(z_p)	SD(z_p)	LI	LS
1/10	360.7154	12445.84	111.5609	142.0600	579.3708
1/20	587.9198	65373.51	255.6824	86.79137	1089.048
1/50	1131.083	486551.6	697.5325	-236.0559	2498.221
1/100	1864.841	2008736	1417.299	-913.0137	4642.695
1/1000	9978.522	154919736	12446.68	-14416.51	34373.56

Se observa que el intervalo de confianza para el nivel de retorno que se asocia a $p = 1/50$ es negativo, lo cual no tiene sentido. De hecho, haciendo un recorrido valor por valor, se puede observar que el nivel de retorno más grande en el cual todavía es positivo corresponde a $p = 1/29,$ con un valor para el límite inferior de 4.659732, pues cuando $p = 1/30,$ el valor asociado es igual a $-5.539881.$

Se puede obtener un mejor nivel de precisión haciendo uso de la verosimilitud perfil. Para el caso en que $p = 1/10,$ se muestra en la figura 4.2 la verosimilitud perfil asociada para el nivel de retorno. En este caso, el intervalo de confianza asociado es [225.7576,854.5455].

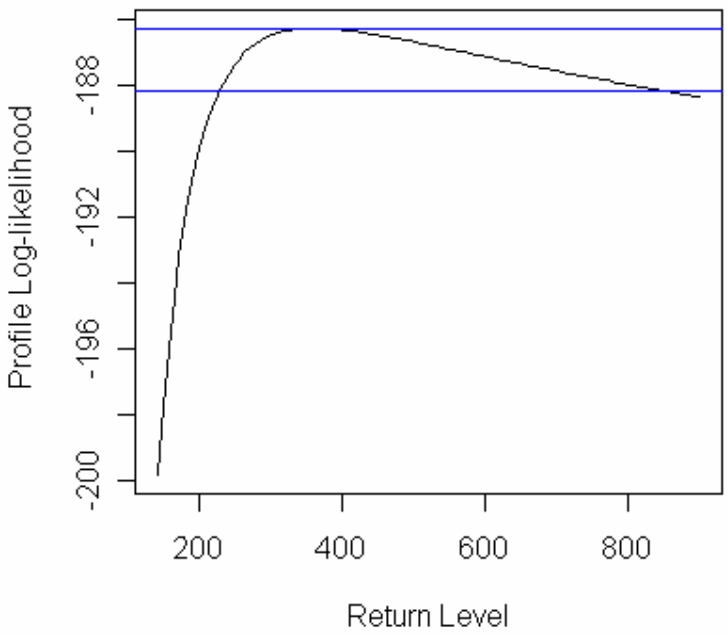

Figura 4.2 *Verosimilitud perfil para un nivel de retorno de 10 años en el ingreso corriente muestral*

Se presentan ahora diversos gráficos diagnósticos para dilucidar acerca del nivel de precisión del modelo GEV, mismos que están en la gráfica 4.3. El papel de probabilidad resulta aproximadamente razonable, aunque no de la misma forma la gráfica de cuantiles, debido a que hay tres puntos que se separan de manera importante de la recta. Sin embargo, la curva del periodo de retorno proporciona una representación razonablemente satisfactoria a los estimadores empíricos, especialmente

cuando la variabilidad muestral se toma en consideración. Finalmente, la densidad estimada correspondiente parece ser consistente con el histograma de los datos. Así, salvo el caso de los cuantiles, las otras tres gráficas parecen ser coherentes con el hecho de que se tenga un modelo GEV genuino.

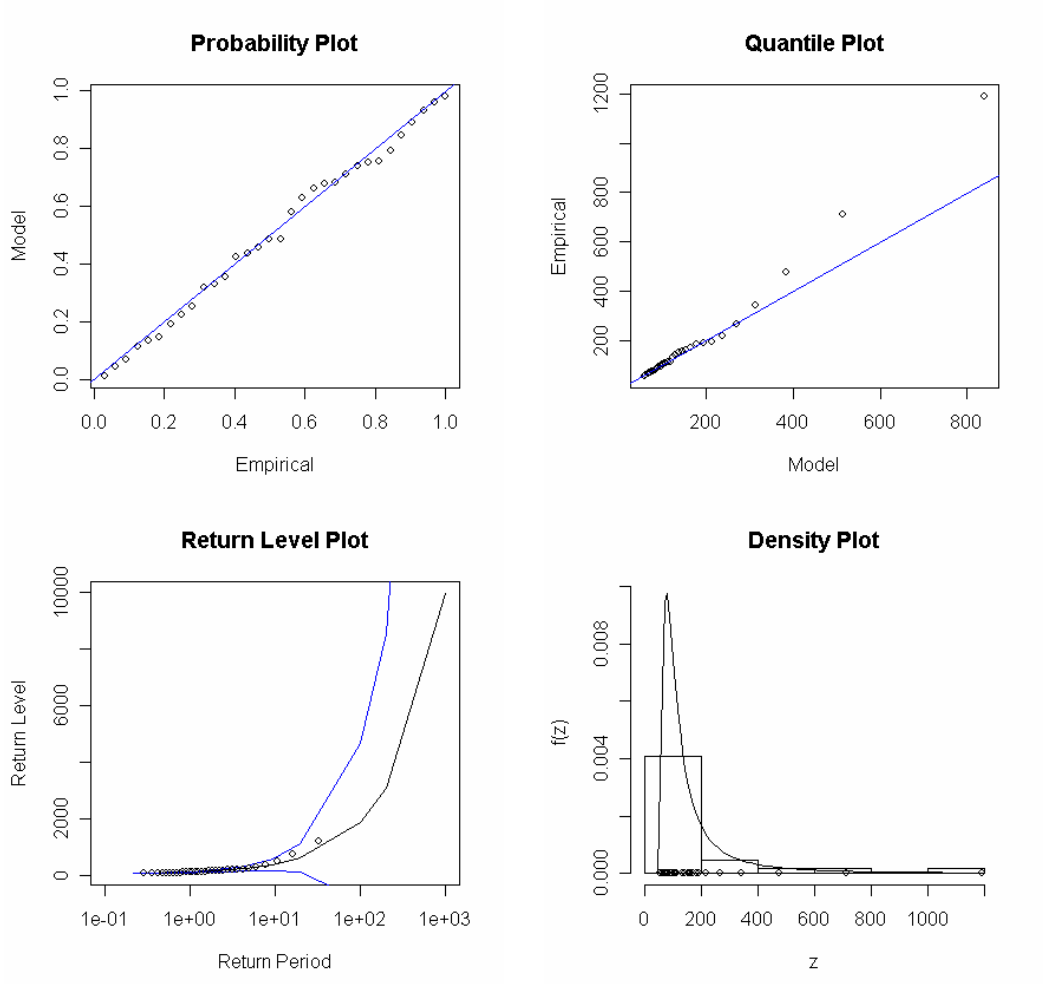

Figura 4.3 *Figuras de diagnóstico del GEV para los ingresos corrientes muestrales*

Para finalizar este apartado, se tienen algunos comentarios. La versión original del teorema de los tipos extremos, tal y como se presentó en el teorema 3.1, identifica tres posibles familias para la distribución límite de los máximos. Antes de la unificación de las tres distribuciones en la familia GEV, sería natural el realizar una selección preliminar de un modelo que anteceda al proceso de estimación de parámetros. Sin embargo, este procedimiento tiene poco mérito a la luz de la modelación de la familia GEV completa. Asociado con el valor positivo del parámetro de forma encontrado, se observa que en este primer caso analizado, tal y como ya se había comentado, la familia GEV corresponde a una distribución Fréchet. Este resultado será de utilidad cuando se realicen los subsiguientes análisis y comparaciones.

4.3.2 Análisis de datos expandidos

En esta sección la base de datos se expande, para lo cual se considera la tabla 4.2 en la cual se muestran, por entidad, los factores de expansión asociados en la ENIGH 2004.

Tabla 4.2 *Tabla de factores de expansión por entidad federativa en la ENIGH 2004*

ENTIDAD	ESTADO	FACTOR	ENTIDAD	ESTADO	FACTOR
1	Aguascalientes	1069	17	Morelos	1250
2	BCN	3790	18	Nayarit	539
3	BCS	95	19	NL	121
4	Campeche	352	20	Oaxaca	595
5	Coahuila	1424	21	Puebla	1798
6	Colima	288	22	Querétaro	211
7	Chiapas	1262	23	Qroo	1153
8	Chihuahua	1466	24	SLP	153
9	DF	701	25	Sinaloa	1484
10	Durango	845	26	Sonora	232
11	Guanajuato	950	27	Tabasco	651
12	Guerrero	229	28	Tamaulipas	1464
13	Hidalgo	590	29	Tlaxcala	6652
14	Jalisco	1289	30	Veracruz	803
15	EdoMex	2876	31	Yucatán	885
16	Michoacán	1093	32	Zacatecas	465

Así, se consideran ahora 36,775 registros en lugar de sólo 32 (se generó una pequeña función en R para conformar la base de datos expandidos). Igual que antes, los datos se dividen entre 3 para transformarlos a una base mensual y entre 1000 para manejar montos en miles de pesos.

Los estimadores de máxima verosimilitud son:

$$\widehat{\mu} = 97.8780383,\ \widehat{\sigma} = 39.8132452,\ \widehat{\xi} = 0.5717471.$$

Ahora bien, la log-verosimilitud maximizada es igual a 210446.9. La matriz aproximada de varianza-covarianza de los estimadores de los parámetros es:

$$V = \begin{bmatrix} 0.0576050180 & 0.0445119938 & -0.0004002606 \\ 0.0445119938 & 0.0577778468 & 0.0001094840 \\ -0.0004002606 & 0.0001094840 & 0.0000399032 \end{bmatrix}$$

La diagonal de esta matriz corresponden a las varianzas de los parámetros individuales (μ, σ, ξ). Tomando las raíces cuadradas, los errores estándar son:

$$\mathrm{SE}_{\widehat{\mu}} = 0.240010454,\ \mathrm{SE}_{\widehat{\sigma}} = 0.240370229,\ \mathrm{SE}_{\widehat{\xi}} = 0.005830122.$$

Combinando los valores de los estimadores con estos errores estándar, se tienen los siguientes intervalos de confianza aproximados del 95% de confianza: $\hat{\mu}$: [97.40763,98.34845]; $\hat{\sigma}$: [39.34213,40.28436]; $\hat{\xi}$: [0.5603203,0.5831739]. Nuevamente, el estimador para el parámetro de forma ξ, adquiere un valor positivo, y el intervalo de confianza no toma como valor al cero, con lo cual se sigue que los estimadores de máxima verosimilitud son regulares. Se observa también que debido a lo pequeño de los errores (puesto que una gran cantidad de observaciones repiten sus valores numéricos), los intervalos son muy estrechos.

Para obtener un intervalo de confianza más fidedigno, se puede utilizar la verosimilitud perfil. La figura 4.4 muestra la forma de la verosimilitud perfil para ξ, de la cual se obtiene que un intervalo de confianza del 95% es [0.5555556,0.5777778], el cual es más fino que el anterior, aunque está ligeramente más cerca del origen.

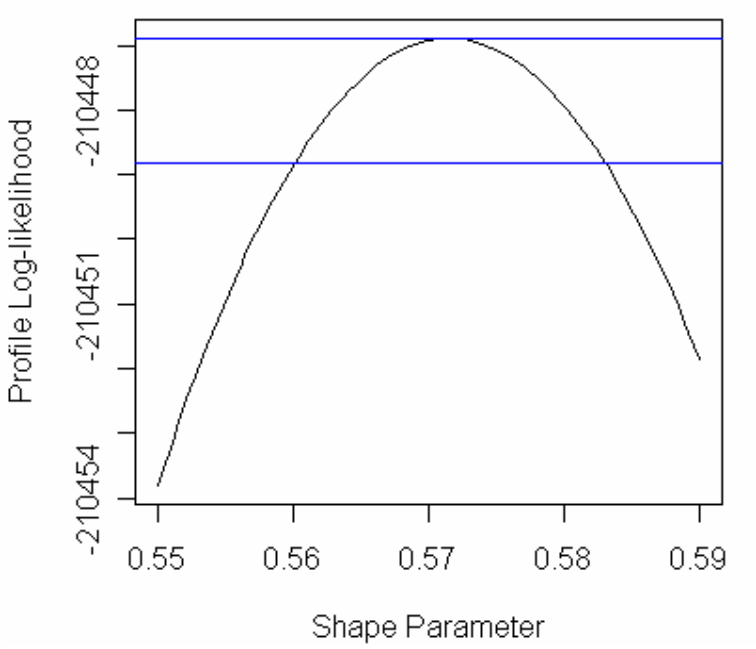

Figura 4.4 Verosimilitud *perfil para* ξ *del ingreso corriente expandido*

Los estimadores e intervalos de confianza de los niveles de retorno se obtienen sustituyendo en $(\hat{\mu},\hat{\sigma},\hat{\xi})$ en (3.10) y (3.11). En la tabla 4.3 se presentan los resultados de los diferentes valores de retorno, sus varianzas, desviaciones estándar, límites de confianza inferior y superior para valores seleccionados de p iguales a $1/10$, $1/20$, $1/50$, $1/100$, $1/1000$, todos a un nivel de confianza del 95%, donde nuevamente SD es la desviación estándar, LI y LS son los límites inferior y superior, respectivamente, de los intervalos de confianza.

Tabla 4.3 *Tabla de los valores de retornos y sus intervalos de confianza para diferentes valores de* p

p	z_p	var(z_p)	SD(z_p)	LI	LS
1/10	280.3622	3.823629	1.95541	276.5296	284.1947
1/20	408.7332	16.49758	4.061721	400.7724	416.6940
1/50	676.4314	95.41643	9.768133	657.2862	695.5765
1/100	994.4527	324.5039	18.01399	959.1459	1029.759
1/1000	3641.849	12780.98	113.0530	3420.269	3863.429

En este caso se puede observar que para ningún nivel de retorno se tienen intervalos de confianza que contengan números negativos.

Se puede obtener un mejor nivel de precisión haciendo uso de la verosimilitud perfil. Para el caso en que $p = 1/10$, se muestra en la figura 4.2 la verosimilitud perfil asociada para el nivel de retorno. En este caso, el intervalo de confianza asociado es [276.7677,283.8384], el cual prácticamente es igual al anterior.

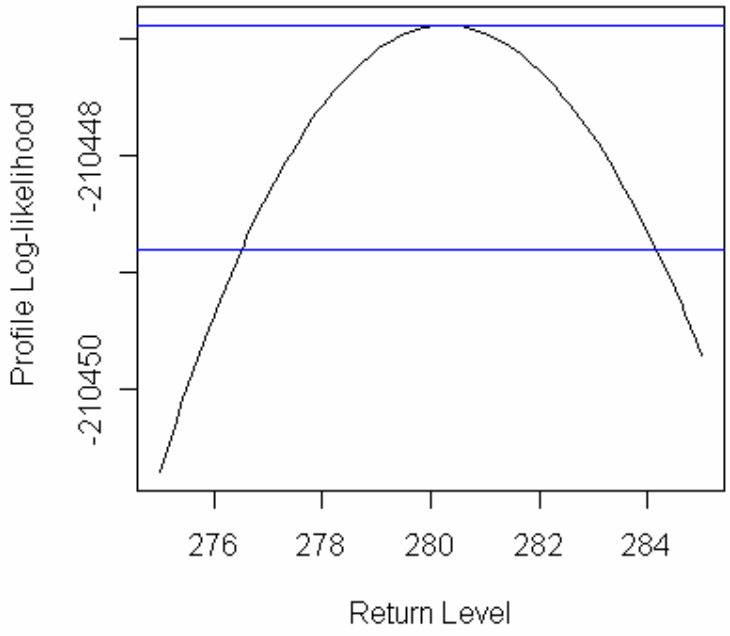

Figura 4.5 *Verosimilitud perfil para un nivel de retorno de 10 años en el ingreso corriente expandido*

Los gráficos diagnósticos se presentan en la gráfica 4.6. Como resulta claro del análisis de dichas gráficas, no se puede justificar el ajuste del modelo GEV a este conjunto de observaciones, pues sobre todo las tres primeras gráficas presentan problemas importantes. La explicación es doble: el conjunto de datos es relativamente grande, y más aún, existen grandes subconjuntos de datos repetidos, lo cual anula por supuesto el supuesto de independencia de los datos. Con esto se concluye parcialmente que la técnica GEV no puede aplicarse de manera razonable en el formato expuesto, puesto que finalmente la ENIGH al ser una muestra y no un estudio poblacional, debe de recurrir a factores de expansión, lo cual implica que forzosamente las cifras se repiten.

Es más, realizando pruebas de introducir perturbaciones aleatorias (mediante la función "jitter" de R y otras afines), no se corrige esta situación. Por este motivo, en los siguientes apartados no se continuará el análisis de los datos expandidos.

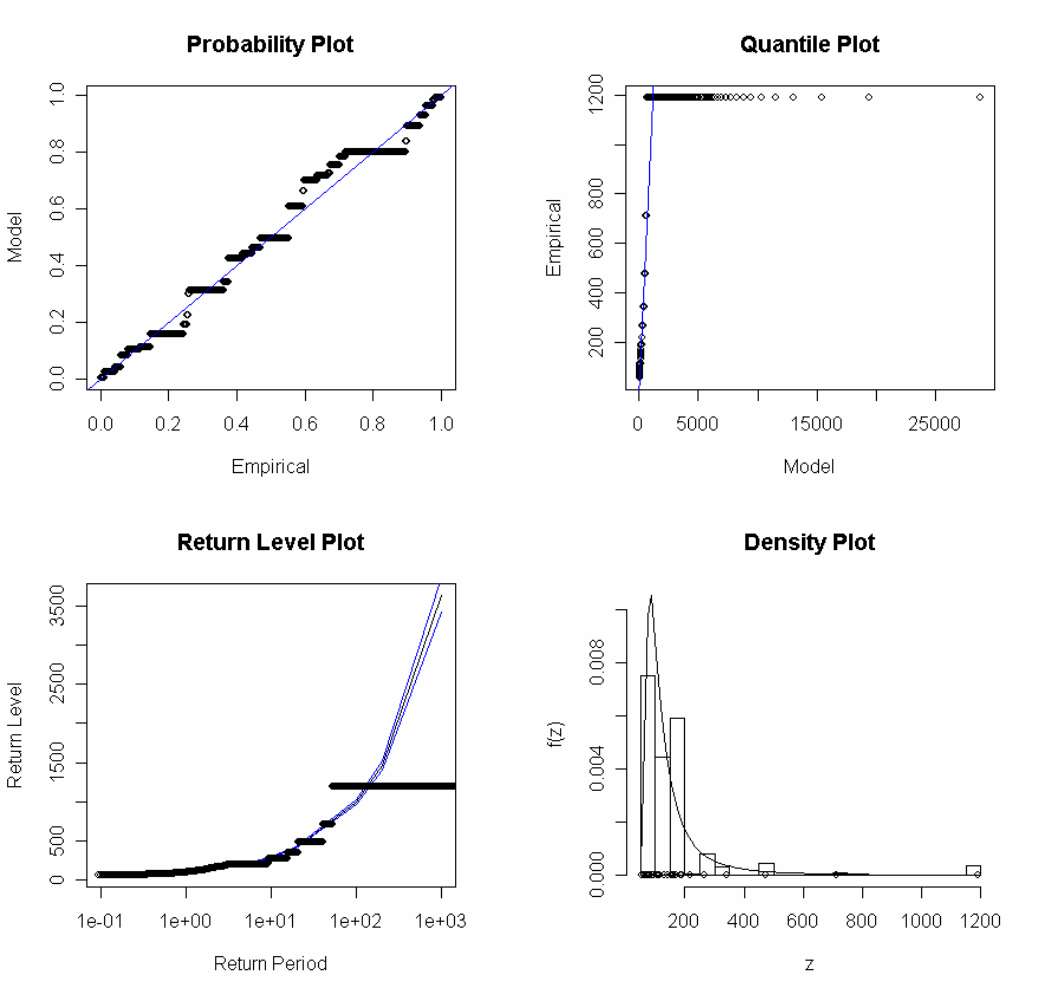

Figura 4.6 *Figuras de diagnóstico del GEV para los ingresos corrientes expandidos*

Se pasa ahora a realizar un análisis similar para el caso de los ingresos totales y monetarios. Debido a la similitud de las observaciones, en estos sub-apartados no se realizarán mayores comentarios salvo cuando los resultados difieran de los presentados con anterioridad.

4.4 MODELACIÓN DEL MAYOR ESTADÍSTICO DE ORDEN r

Los datos ahora están constituidos por 11 columnas y 32 filas, de las cuales la primera es categórica e indica la entidad federativa. La primera de las 10 restantes indica el monto máximo del ingreso corriente. La segunda es el siguiente mayor ingreso, y así sucesivamente. Entonces se aplica el modelo (3.15) para cualquiera de los valores

$r = 1,...,10$. Los estimadores de máxima verosimilitud y los errores estándar se encuentran dados en la siguiente tabla (conviene aclarar que nuevamente los datos fueron divididos entre 3 por ser trimestral la información de origen, y entre 10,000 para que los datos se den en decenas de miles de pesos –no se dividió solo entre 1,000 como antes por motivos de situación de desbordamiento en los cálculos–):

Tabla 4.4 *Tabla de los valores de máxima verosimilitud, valores de los estimadores y desviaciones estándar (en paréntesis) del mayor estadístico de orden r para los ingresos corrientes con diferentes valores de r*

r	l	$\hat{\mu}$	$\hat{\sigma}$	$\hat{\xi}$
1	112.5706	9.6726536 (0.9454867)	4.5955733 (1.0522120)	0.7333521 (0.2147981)
2	180.8059	9.8346805 (0.8520110)	5.3217175 (0.9769762)	0.6006352 (0.1281792)
3	220.2122	10.1782587 (0.8849974)	5.7471100 (1.0801218)	0.6559741 (0.1260355)
4	238.8337	10.5173094 (0.9199496)	6.0890276 (1.1875823)	0.7031854 (0.1225096)
5	251.6549	11.1019218 (0.91613978)	6.1324087 (1.07616196)	0.6060055 (0.09438516)
6	257.2880	11.4379924 (0.93522316)	6.3102925 (1.07958444)	0.5935319 (0.08619768)
7	257.6421	11.7631766 (0.93931364)	6.3856233 (1.04568401)	0.5636176 (0.07782278)
8	248.6247	11.1781777 (0.76163516)	5.7097576 (0.79823035)	0.5251102 (0.06555131)
9	228.9636	11.9789095 (0.93732075)	6.4392685 (1.02443504)	0.5601436 (0.06927137)
10	210.0924	12.1211447 (0.93654245)	6.4686667 (1.00810101)	0.5509865 (0.06593525)

Mediante el análisis de las gráficas que se muestran en la figura 4.7, la cual manifiesta el comportamiento de los estimadores y sus correspondientes errores estándar, se observa que en el caso de $\hat{\mu}$ y $\hat{\sigma}$, a medida que se incrementa el valor de r, en general crece también el valor del estimador, salvo cuando $r = 8$, situación que es hasta cierto punto inversa en el caso de $\hat{\xi}$. Se observa que las gráficas asociadas con los errores estándar son descendentes, siendo el caso más evidente la última figura correspondiente al caso de $\hat{\xi}$. Sin embargo, si la aproximación asintótica es válida para una selección particular de r, entonces los estimadores de los parámetros deberían ser estables cuando el modelo se ajusta con un número menor de estadísticos de orden. Pero de la tabla 4.4 y de la figura 4.7 se sigue que hay poca evidencia de estabilidad sobre el parámetro de forma, sobre todo cuando $r \geq 8$. Este lleva a dudar de la validez del modelo, al menos cuando $r \geq 8$.

Debido a que los parámetros de μ, σ y ξ corresponden exactamente a los parámetros GEV para la distribución del máximo anualizado, una afirmación más detallada del ajuste del modelo puede ser derivada de las curvas de niveles de retorno de la distribución anualizada de los máximos. Estas se construyen de la misma forma que en el modelo de máximo por bloques, pero esta vez utilizando los estimadores de máxima verosimilitud y las matrices de varianzas y covarianzas para el modelo del mayor estadístico de orden r.

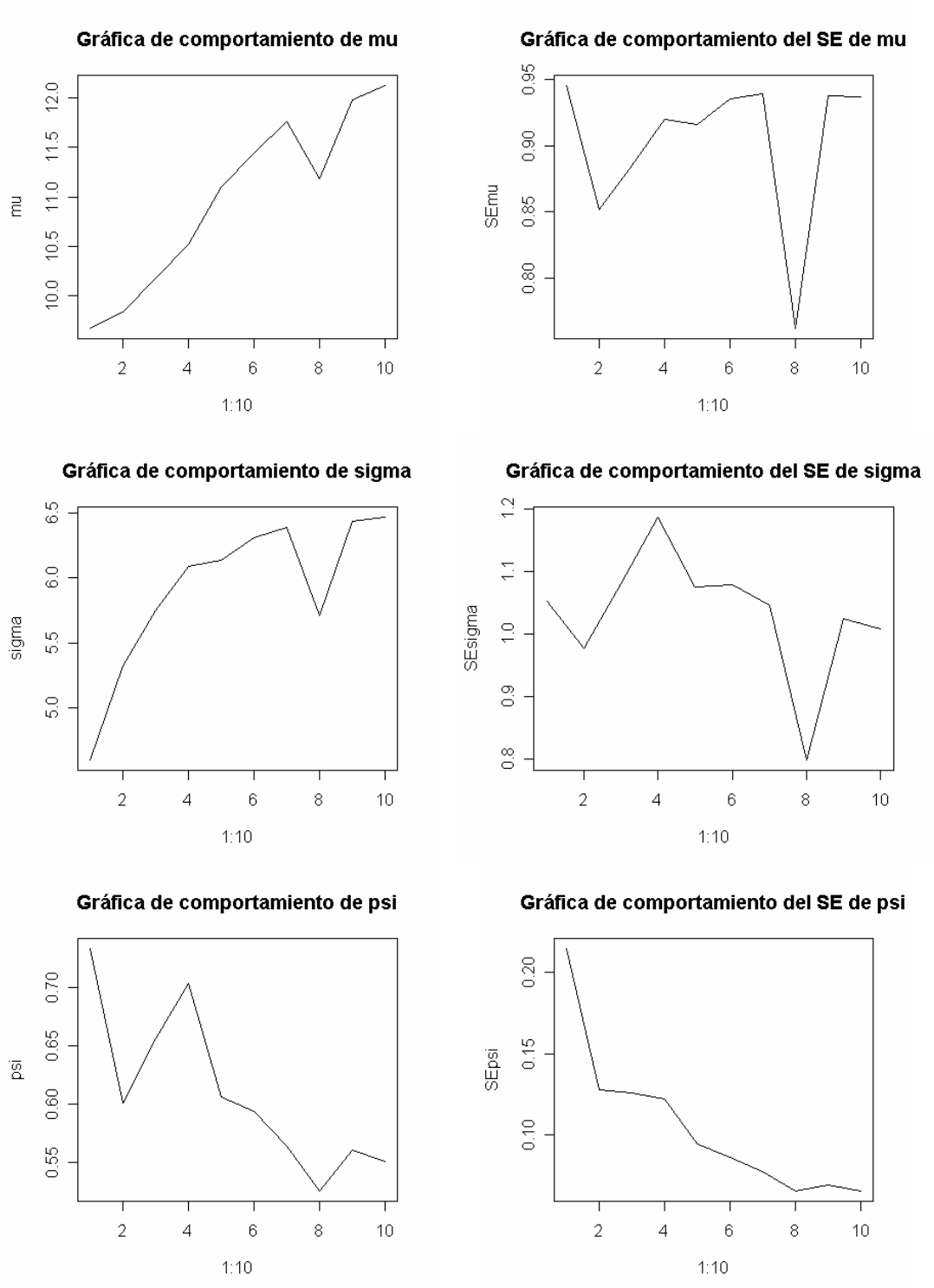

Figura 4.7 *Gráficas del comportamiento de los estimadores y sus desviaciones estándar del modelo del estadístico de mayor orden r para los ingresos corrientes*

En la figura 4.8 se muestran las gráficas de los niveles de retorno estimados con intervalos de confianza del 95%, para valores de r en el rango de 2 a 10. Se observa que en todos los casos el ajuste parecería ser razonable. Un hecho interesante es que estas gráficas muestran la correspondencia entre el modelo y los datos decrece a medida que r incrementa su valor, a pesar de que los intervalos de confianza se vuelven menos grandes. Esta es una ilustración gráfica del intercambio entre el sesgo y la varianza determinado por la selección de r.

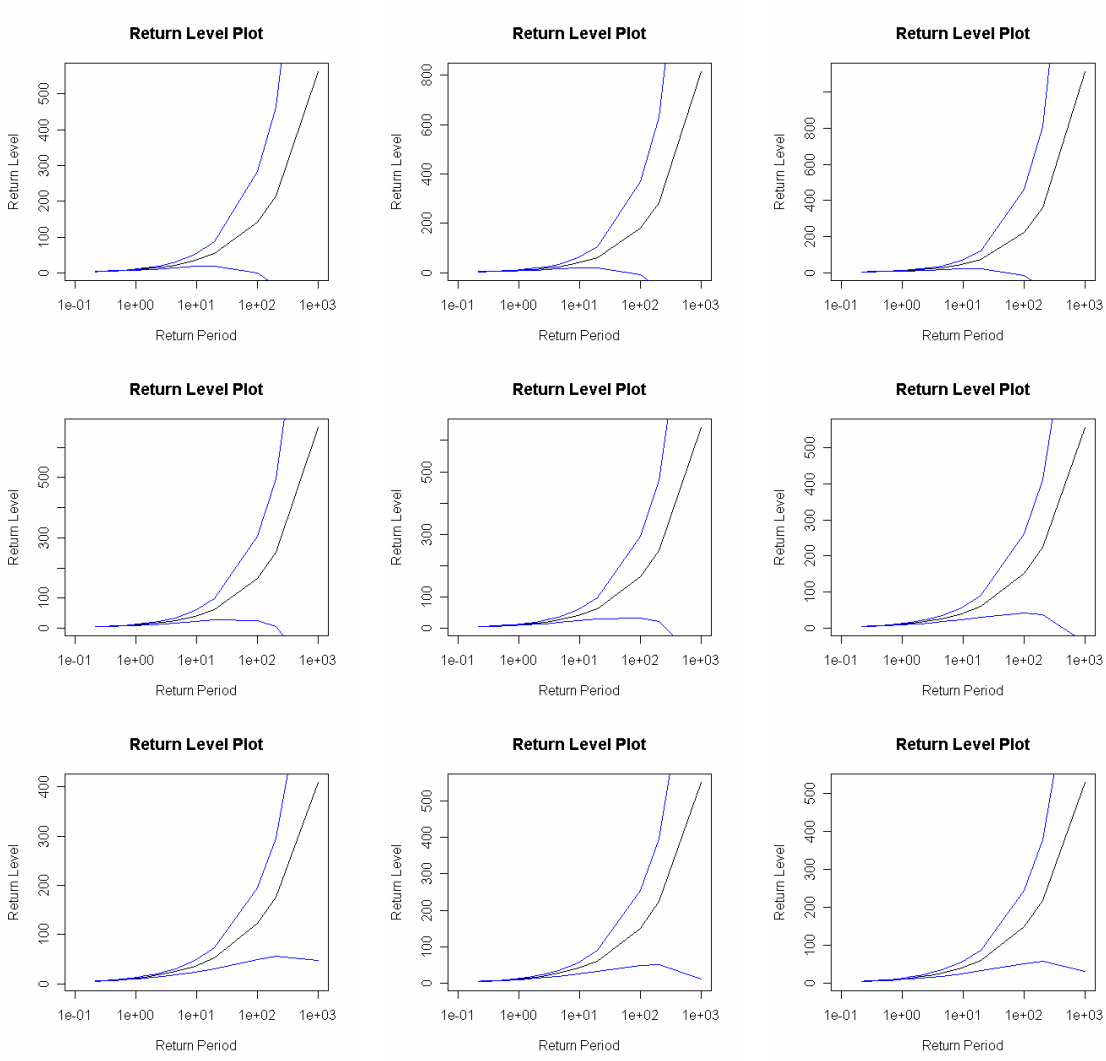

Figura 4.8 *Estimaciones de niveles de retorno con intervalos de confianza del 95% para las distribuciones máximas basadas en el modelo del mayor estadístico de orden r para los ingresos corrientes*

Para cualquier valor particular de r la precisión del ajuste puede ser examinada en un mayor detalle. En la figura 4.9 se muestran las gráficas de diagnóstico tipo GEV para el mayor estadístico de orden r cuando $r = 5$. En la primera de estas gráficas los puntos tienen a seguir un comportamiento lineal, aunque es curioso que todas las observaciones están por debajo de la recta. En la segunda gráfica se tienen dos puntos

que se alejan de manera considerable de la línea recta. Sin embargo, para el periodo de retorno se observa, tal y como ya se había comentado, un comportamiento razonable. Finalmente, el histograma muestra una concordancia razonable con la densidad propuesta. Tal y como ocurre con las gráficas de niveles de retorno, estas se obtienen en exactamente la misma forma que en el caso del modelo por bloques, sustituyendo la estimación de los parámetros y la matriz de varianzas y covarianzas con aquellas que se obtienen de la maximización de (3.17). En este caso, por los comentarios anteriores, el modelo parece ajustar de manera razonable a los ingresos corrientes. Sin embargo, al analizar los papeles de probabilidad y de cuantiles para cada uno de los cuatro mayores estadísticos de orden, pueden surgir sospechas importantes de cierta falta de ajuste, sobre todo para el caso en que $k = 2$ (como puede observarse en la figura 4.10).

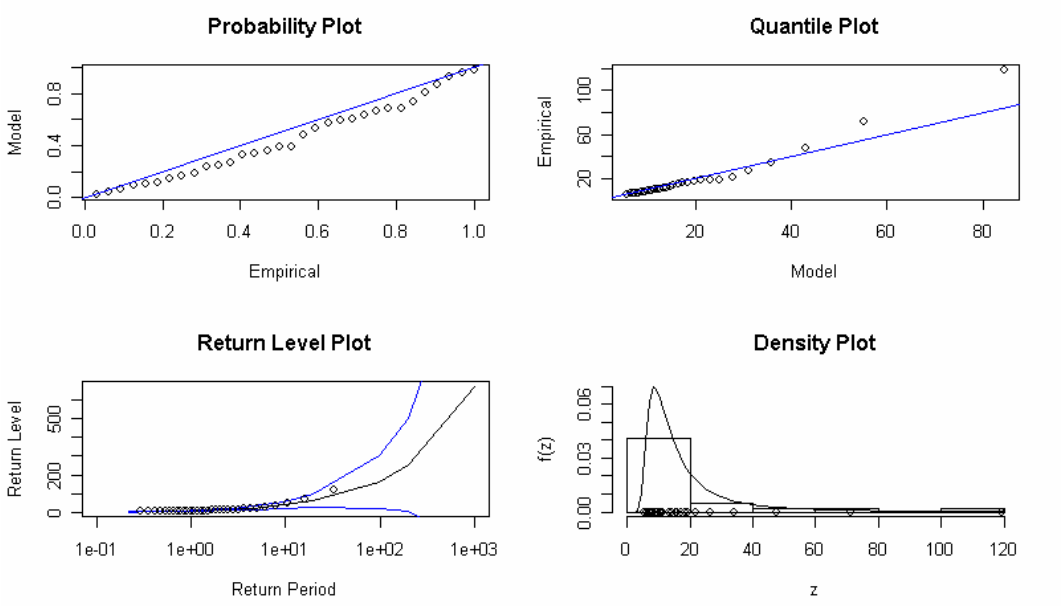

Figura 4.9 *Gráficas de diagnóstico tipo GEV para los máximos de los ingresos corrientes sobre la base del mayor estadístico de orden r cuando $r = 5$*

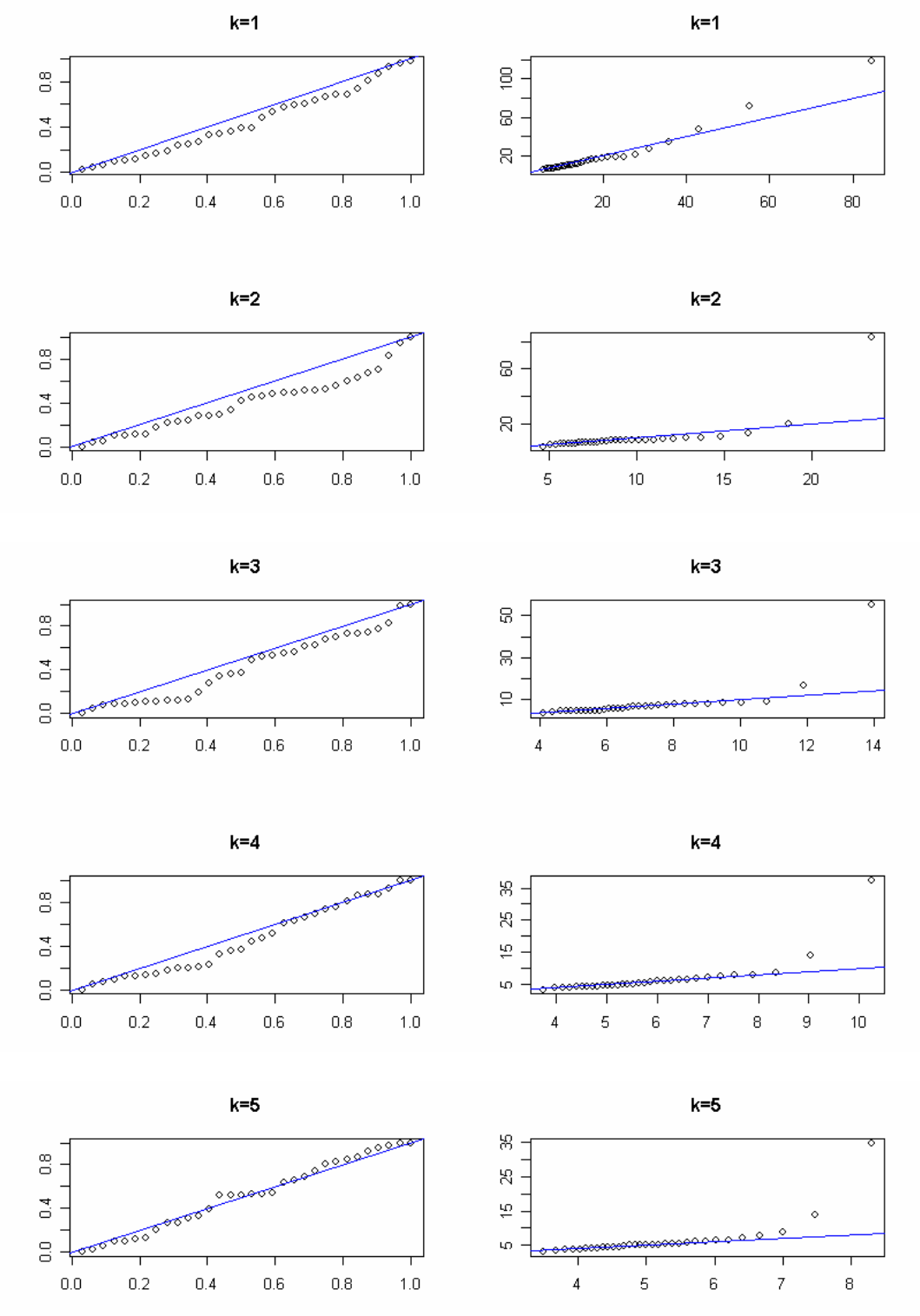

Figura 4.10 *Diagnósticos del modelo de los ingresos corrientes sobre la base del modelo del mayor estadístico de orden* r *cuando* $r=5$. *Las gráficas que se muestran son los papeles de probabilidad (columna de la izquierda) y los diagramas de cuantiles (columna de la derecha) para el* $k-$*ésimo estadístico de orden cuando* $k=1,...,5$.

4.5 MODELACIÓN DE UMBRALES

En este caso la base de datos se conforma de la totalidad de los ingresos corrientes en la ENIGH 2004. La figura 4.11 muestra el gráfica de los residuales promedios de vida estos datos. En esta se puede observar que de manera aproximada la gráfica es línea hasta un nivel de 200, por lo cual este valor se sugiere como el umbral para comenzar el análisis.

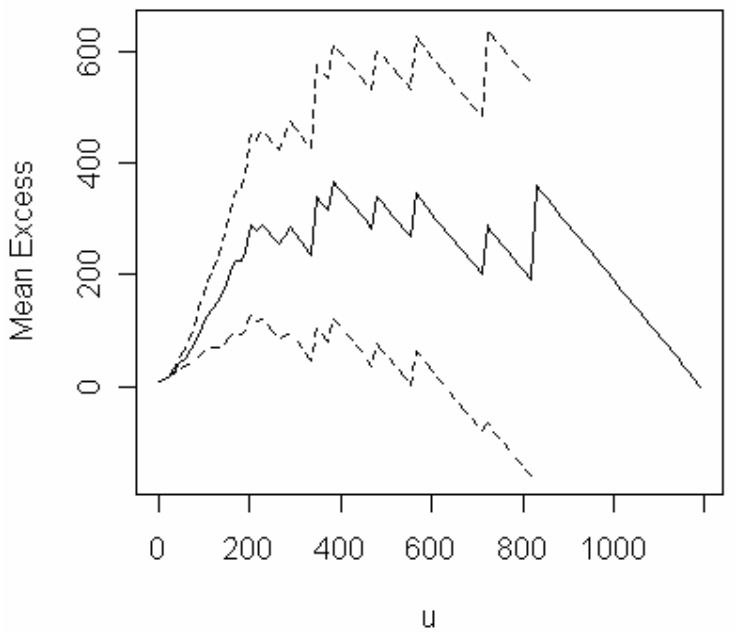

Figura 4.11 *Gráfica de los residuales promedios de vida de los ingresos corrientes*

Adicionalmente, en las figura 4.12 se muestran los estimados de los parámetros en contra del umbral para los ingresos corrientes. Con estas figuras se re-estructura la apreciación de esta afirmación y por el análisis de dichas figuras ahora se propone un valor de umbral de 100. Con esto cota, existen 58 valores que satisfacen este requisito.

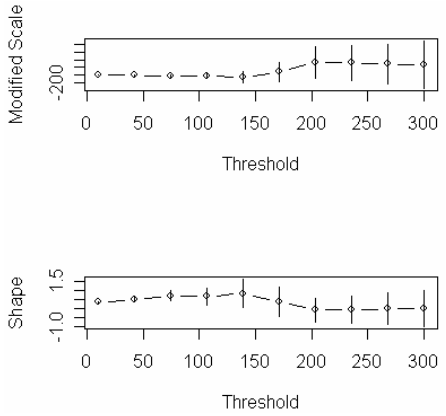

Figura 4.12 *Estimados de los parámetros en contra del umbral para los ingresos corrientes*

Los estimadores de máxima verosimilitud en este caso son:

$$(\hat{\sigma},\hat{\xi}) = (38.2290945, 0.7883592)$$

con una log-verosimilitud correspondiente de 315.0304. La matriz de varianza y covarianzas es:

$$\begin{bmatrix} 99.392459 & -1.4215219 \\ -1.421522 & 0.0615261 \end{bmatrix},$$

con errores estándar iguales a 9.9695767 y 0.2480446 para $\hat{\sigma}$ y $\hat{\xi}$, respectivamente. Así, los intervalos de confianza del 95% son (18.68908,57.7691) y (0.3022007,1.274518) para $\hat{\sigma}$ y $\hat{\xi}$, de manera respectiva. El estimador de máxima verosimilitud corresponde, por lo tanto a una distribución no acotada (puesto que $\hat{\xi} > 0$), siendo fuerte la evidencia a esta afirmación, puesto que el intervalo del 95% para ξ está exclusivamente dentro del dominio positivo.

Debido a que existen 58 excedencias en el conjunto completo de 22,595, el estimador de máxima verosimilitud de la probabilidad de excedencias es

$$\hat{\varsigma}_u = 58/22{,}595 = 0.002566940,$$

con varianza aproximada igual a

$$\text{var}(\hat{\varsigma}_u) = \hat{\varsigma}_u(1-\hat{\varsigma}_u)/22{,}595 = 1.133149\times10^{-7}.$$

Por lo tanto, la matriz completa de varianzas y covarianzas para $(\hat{\varsigma},\hat{\sigma},\hat{\xi})$ es

$$V = \begin{bmatrix} 1.133149\times10^{-7} & 0 & 0 \\ 0 & 99.392459 & -1.4215219 \\ 0 & -1.421522 & 0.0615261 \end{bmatrix}.$$

Debido a que $\hat{\xi} > 0$, no es muy útil el llevar a cabo un análisis inferencial detallado del límite superior.

Finalmente, los gráficos de diagnóstico del modelo generalizado de Pareto se muestran en la figura 4.13. Estas gráficas indican un ajuste razonable a dicho modelo, puesto que no se presentan desviaciones importantes en ninguna de ellas.

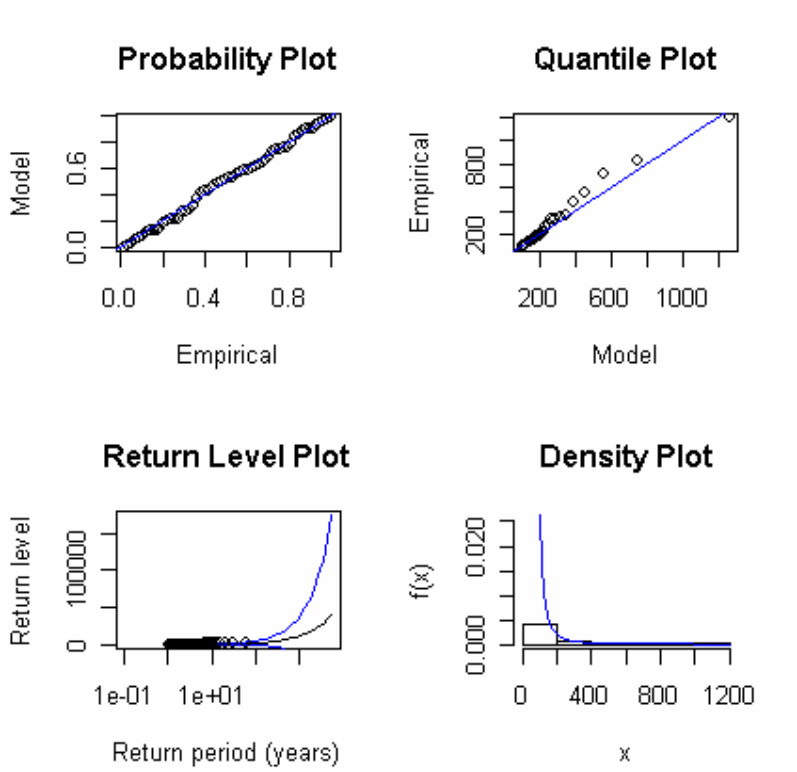

Figura 4.13 *Gráficas de diagnóstico para el modelo de umbral de excedencias para los ingresos corrientes*

4.6 MEDIDAS DE BONDAD DE AJUSTE DE DIVERSOS MODELOS

En esta sección, la última de este trabajo, se realizarán diversos ajustes a formas distribucionales que han sido ampliamente utilizadas en la modelación de los ingresos en diversos contextos. El primero de ellos se suscita al considerar únicamente los extremos de los ingresos corrientes ya comentados. Finalmente, como los ingresos extremos en sí mismos tienen un comportamiento altamente sesgado a la derecha, tal y como se analizó en el análisis de los datos extremos, las distribuciones del capítulo 2 también pueden ser aplicadas en cuanto a un ajuste de los datos. Se recordará que el análisis fundamental consiste en la cola derecha de la distribución del ingreso, y por este motivo se selecciona así el ajuste (con dicha información).

4.6.1 Modelo beta generalizado II

Esta distribución, como fue analizado en el capítulo 2, es ampliamente utilizada para la modelación de los ingresos (más que la misma beta generalizada), debido sobre todo a que incluye una gran cantidad de distribuciones que ha su vez también se utilizan ampliamente para dichos fines.

La forma distribucional ajustada es la siguiente:

$$f(y)=\frac{ay^{ap-1}}{b^{ap}B(p,q)\{1+(y/b)^{a}\}^{p+q}},\quad a>0,b>0,p>0,q>0,y>0.$$

Aquí B es la función beta, y b es un parámetro de escala, mientras que los otros son parámetros de forma. El valor esperado es

$$E(Y)=\frac{b\Gamma(p+1/a)\Gamma(q-1/a)}{\Gamma(p)\Gamma(q)}$$

Siempre que $-ap<1<aq$.

En este caso, el ajuste asociado proporciona los siguientes estimadores:

$$a=1.248196\times10^{12},\ b=2.909646\times10^{1},\ p=5.031926\times10^{-12},\ q=7.173998\times10^{-13}.$$

Se observa que el valor del primer parámetro es muy grande, mientras que los valores para p y q son muy pequeños.

El histograma de los datos sobre la curva ajustada se muestra en la figura 4.14a. En esta figura se observa que el ajuste es relativamente razonable, sin embargo se observa en la figura 4.14b que su gráfica qq es sumamente irregular, con lo cual se concluye que el ajuste no es del todo bueno (es conveniente señalar que se acostumbra en la práctica estadística el hecho de tomar como indicador de ajuste las gráficas qq, más que estadísticos de bondad de ajuste, pues tienen problemas en determinarse de manera precisa sobre todo en las colas de la distribución, que es justamente el interés que se tiene).

Llevando a cabo la prueba de Kolmogorov-Smirnov , se tienen los siguientes hallazgos: El estadístico de Kolmogorov-Smirnov tiene un valor de 0.75, p-value = 0.002342.

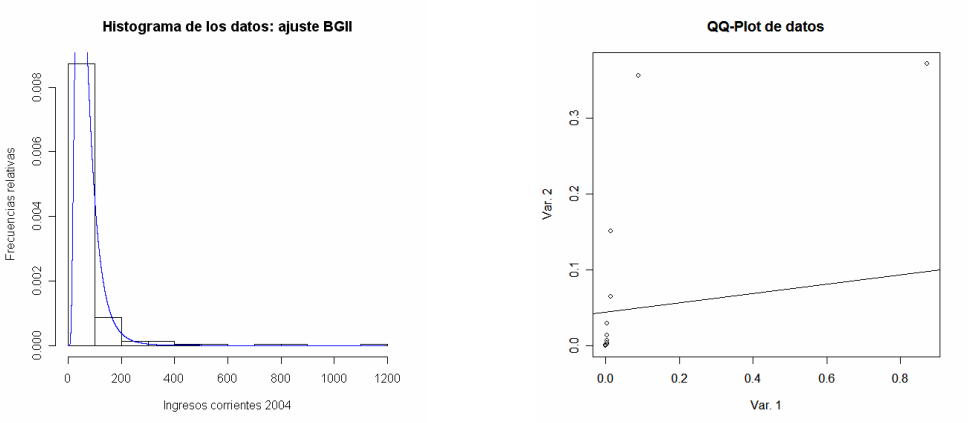

Figura 4.14 *Histograma de observaciones extremas y su ajuste al modelo beta generalizado II*

4.6.2 Modelo lognormal

Una distribución que ha sido importante en el estudio de ajustes a datos de los ingresos es la lognormal. Una variable aleatoria lognormal X lo es si su logaritmo natural es una variable normal. Más específicamente, puede mostrarse que la función de densidad de una variable lognormal con parámetros μ y σ adquiere la siguiente forma:

$$f(x;\mu,\sigma) = \frac{1}{x\sigma\sqrt{2\pi}} \exp\left(-\frac{(\ln x - \mu)^2}{2\sigma^2}\right)$$

En la figura 4.15 se muestran algunas formas gráficas que adquiere la distribución lognormal, cuando $\mu = 0$. Se puede apreciar que efectivamente la distribución es muy dinámica y con el cambio de los valores de sus parámetros se tienen formas parecidas a las distribuciones de los ingresos. Más aún, el modelo lognormal también es de colas pesadas, que describen de manera razonable el fenómeno de los ingresos bajo estudio.

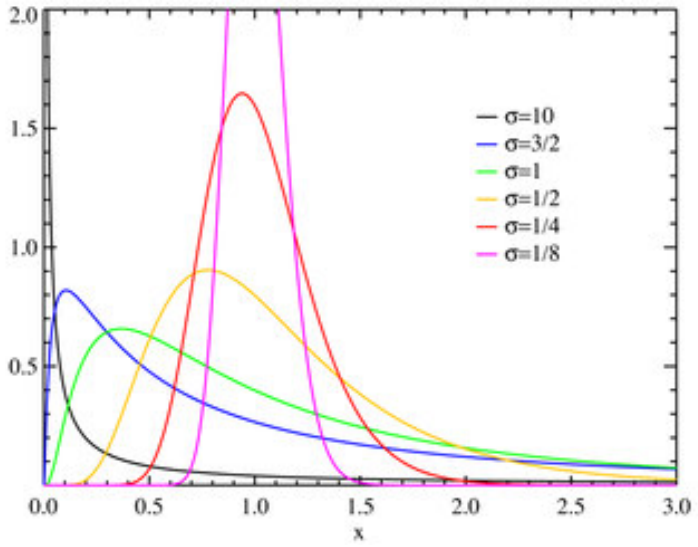

Figura 4.15 *Formas de la densidad de una variable lognormal cuando* $\mu = 0$

Los parámetros ajustados resultan ser

$$\log \mu = 4.9336979;\ \log \sigma = 0.7011166.$$

En la figura 4.16a se muestra el histograma de los datos junto con el ajuste a la distribución lognormal. Se observa un fuerte parecido con respecto a la gráfica correpondiente del modelo beta generalizado II. A pesar de que efectivamente el modelo lognormal es un caso particular del primero, al no considerar parámetros que no aportan una descripción significativa a los datos, puede obtenerse mejores valores en cuanto a los ajustes. De hecho, así ocurre, puesto que el estadístico de Kolmogorov-Smirnov es 0.3333, y el $p-$valor asociado tiene un valor de 0.5176. Se observa, sin embargo, que el gráfico qq-plot es muy deficiente, lo cual se señala en la figura 4.16b

(se observa que existe una observación excesivamente extrema que no permite una pendiente más pronunciada de la recta).

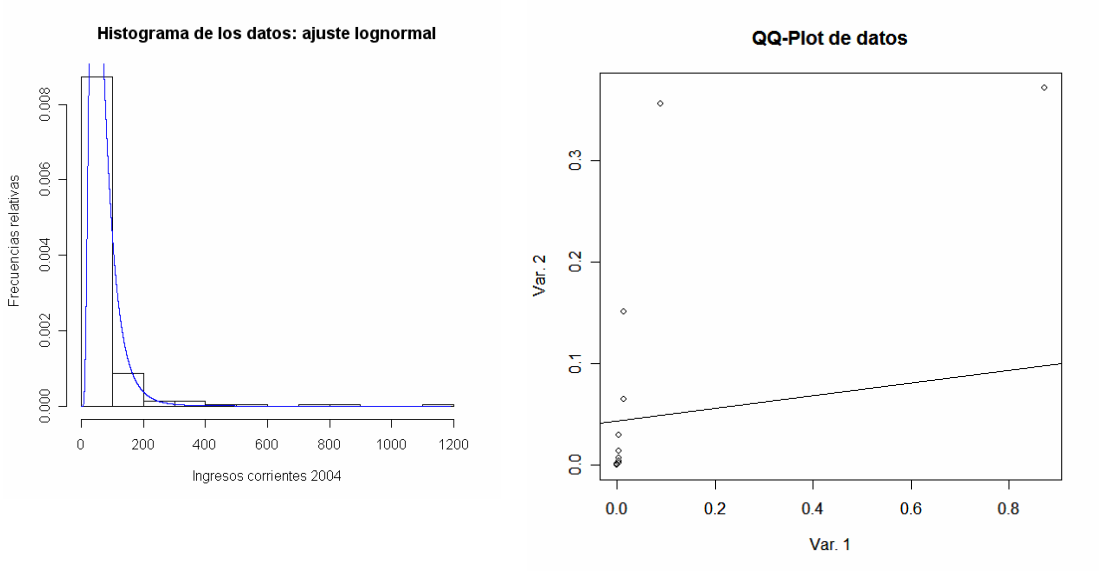

Figura 4.16 *Ajuste de los datos a una distribución lognormal y qq-plot de datos a la distribución lognormal*

4.6.3 Modelo Fréchet

En el capítulo 3 se proporcionó la función de distribución del modelo Fréchet, considerada dentro del contexto de la familia unificada GEV. La función de densidad correspondiente es:

$$f(x;\mu,\sigma,\xi)=\frac{1}{\sigma}\left[1+\xi\left(\frac{x-\mu}{\sigma}\right)\right]^{-1/\xi-1}.$$

Al ajustar los estimadores de este modelo se tiene que son iguales a:

$$(\hat{\mu},\hat{\sigma},\hat{\xi})=(12.1211447,6.46686667,0.5509865).$$

En la figura 4.17a se muestra el histograma de los datos del ingreso corriente según la ENIGH-2004 con respecto al modelo ajustado. Se observa un ajuste razonable. A su vez, el qq plot asociado tiene un comportamiento relativamente adecuado. En este caso, el estadístico de Kolmogorov-Smirnov es igual a $D=0.4167$, con $p-$valor igual a 0.2485.

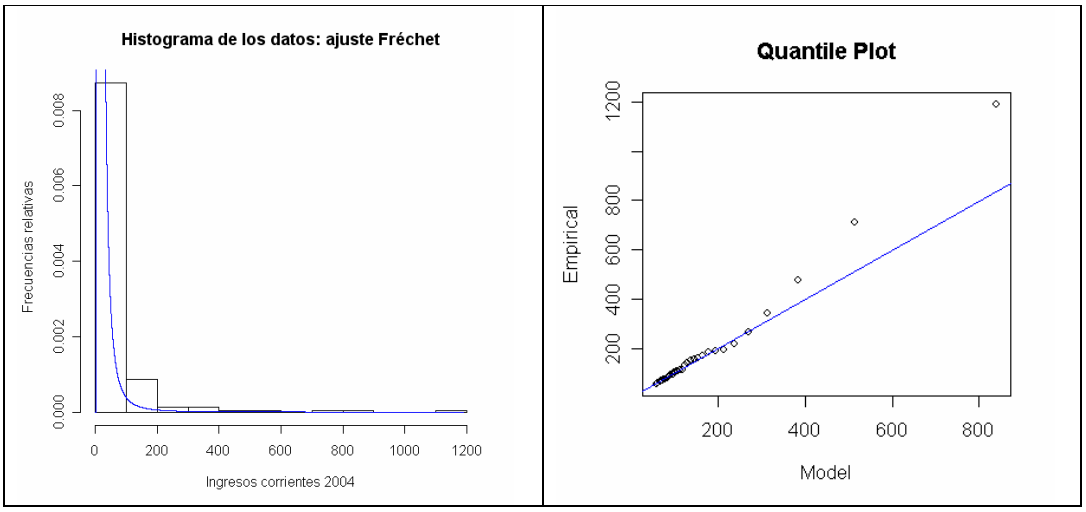

Figura 4.17 *Histograma de observaciones extremas y la gráfica qq asociada*

4.6.4 Modelo Fréchet sin considerarlo como proveniente de la familia GEV

En la sección anterior se considero el modelo Fréchet como proveniente de la familia GEV. Sin embargo, es de interés realizar el ajuste de este modelo no tomando en cuenta que esta unificación, puesto que dicha caracterización se otorga probabilidades no negativas a los datos negativos. Sin embargo, los ingresos nunca son negativos, por lo cual se pierde potencia el la prueba unificada.

Para realizar el ajuste, se adoptó la siguiente forma funcional para la densidad:

$$f(y)=\frac{sb}{(y-a)^2}\exp\left[-\left(\frac{b}{y-a}\right)^s\right]\left[\left(\frac{b}{y-a}\right)^{s-1}\right]$$

para $y>a$ y con parámetro de escala $b>0$. El parámetro de forma positivo es s. La función de distribución acumulada es:

$$F(y)=\exp\left[-\left(\frac{b}{y-a}\right)^s\right].$$

La media de Y es $a+b\Gamma(1-1/s)$ para $s>1$. La varianza de Y es

$$b^2[\Gamma(1-2/s)-\Gamma(1-1/s)^2] \text{ para } s>2.$$

Los valores estimados de los parámetros son:

$$\hat{a}=21.672545,\quad \hat{b}=63.510575,\quad \hat{s}=1.388958.$$

La forma distribucional sobre el histograma de los datos se muestra en la figura 4.18a y en la 4.18b se muestra el qq-plot asociado. Se observa que la figura del qq plot resulta relativamente razonable.

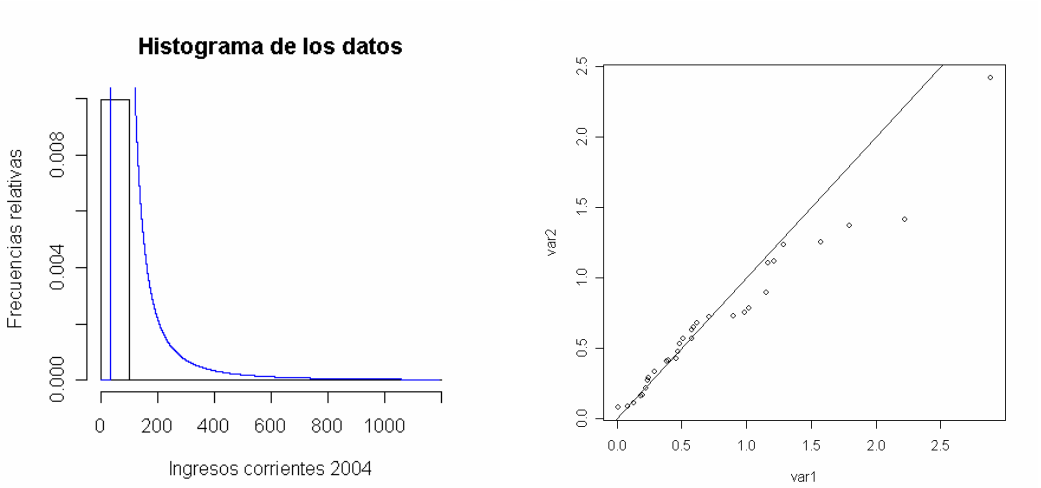

Figura 4.18 *Histograma de los ingresos corrientes y el ajuste del modelo Fréchet sin unificación GEV y qq-plot*

4.6.5 Base de datos completa

Por completitud, en este punto se llevará a cabo el ajuste de las tres distribuciones consideradas con respecto a la base de datos completa.

4.6.5.1 Modelo beta generalizado II

Los valores estimados de los parámetros ajustados son:

$$\hat{a} = 1.292145, \hat{b} = 6.633437, \hat{p} = 1.952230, \hat{q} = 1.951993.$$

La forma distribucional sobre el histograma de los datos se muestra en la figura 4.19a y en la 4.19b se muestra el qq-plot asociado (nuevamente, muy deficiente)

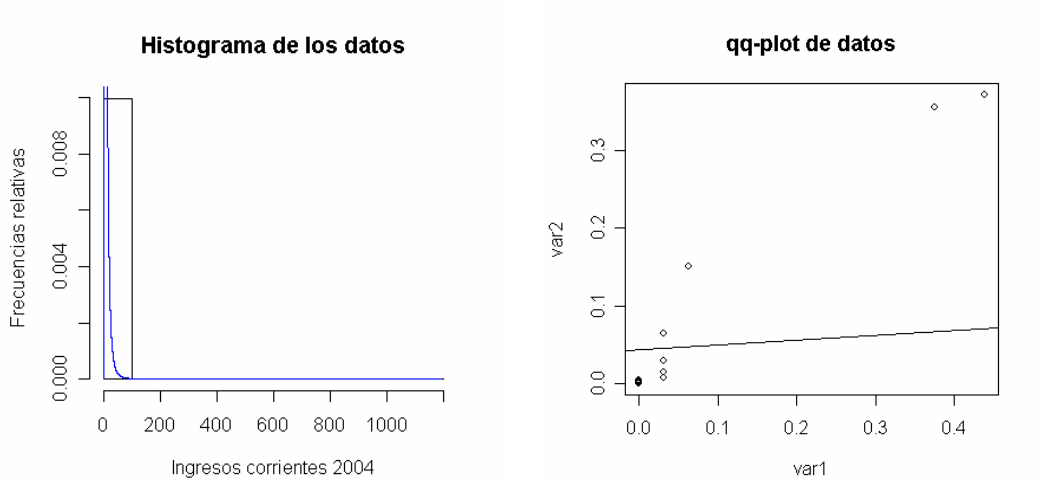

Figura 4.19 *Histograma de los ingresos corrientes y el ajuste del modelo beta generalizado II y qq-plot*

La prueba de Kolmogorov-Smirnov tuvo como resultados un estadístico asociado $D = 0.25$, con un $p-$valor igual a 0.8475. Se observa que la evidencia empírica *no* lleva a rechazar el hecho de que la distribución beta generalizada proporcione un ajuste razonable a los datos.

4.6.5.2 Modelo lognormal

Al realizar el ajuste, los valores de los parámetros son los siguientes:

$$\hat{\mu} = 1.8922448, \hat{\sigma} = 0.8920993.$$

El histograma con la función ajustada se muestran en la figura 4.20a junto con el modelo lognormal, y en la 4.20b se muestra la gráfica qq.

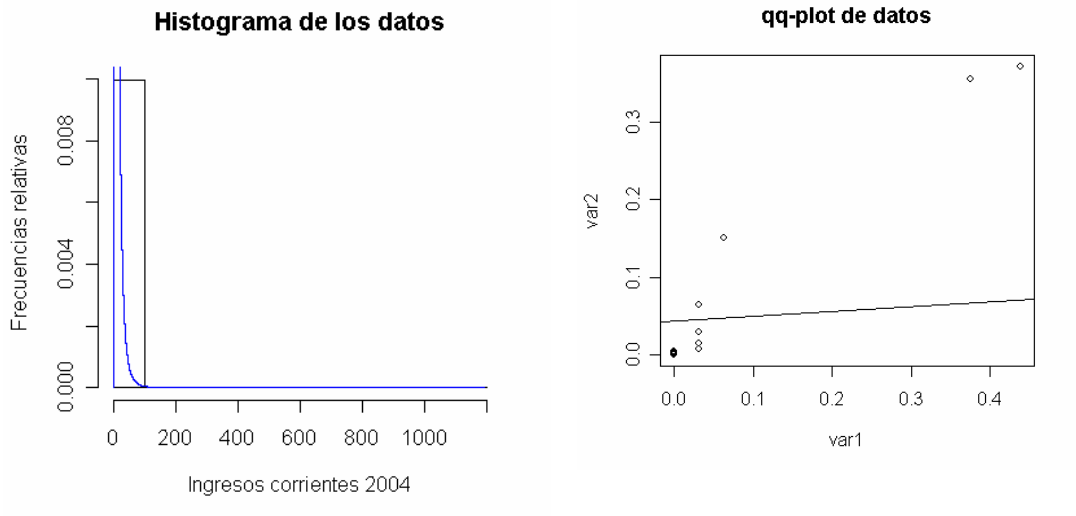

Figura 4.20 *Histograma de los ingresos corrientes y el ajuste del modelo lognormal*

Se observa que la figura tiene un gran parecido con la función beta generalizada II, lo cual no es de extrañar, puesto que ésta es una generalización a la lognormal.

Ahora, en cuanto a la bondad de ajuste de la distribución, se tiene que el estadístico de Kolgomorov-Smirnov es $D = 0.4167$ con un $p-$valor igual a 0.2485. En este caso el ajuste ya no es tan bueno como en el caso de la beta generalizada II, sin embargo sigue siendo razonable, desde la percepción de este estadístico.

4.6.5.3 Modelo Fréchet

Se recordará que en la modelación GEV (capítulo 3), los valores estimados de los parámetros fueron:

$$\hat{\mu} = 96.7472019,\ \hat{\sigma} = 45.9867624,\ \ \hat{\xi} = 0.7335034.$$

El histograma, junto con la función ajustada de Fréchet ajustada se muestra en la figura 4.21a y el qq plot se muestra la 4.21b. Se recordará, a punto de comparación, las gráficas de cuantiles que fueron presentas en el capítulo 3.

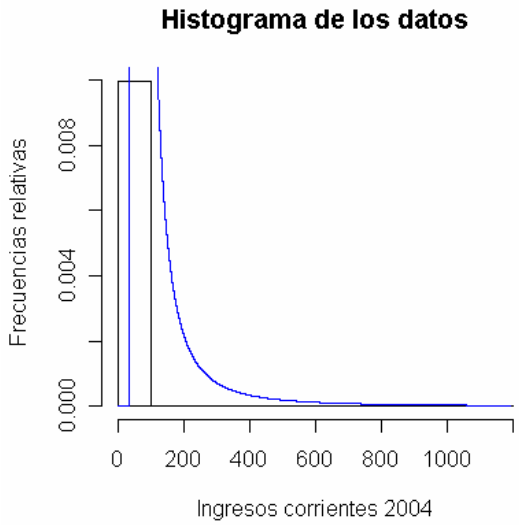

Figura 4.21 *Histograma de los ingresos corrientes y el ajuste del modelo Fréchet*

En este caso, el estadístico de Kolmogorov-Smirnov fue igual a 0.3889, con un $p-$valor asociado igual a 0.1314. Se observa que el p valor no lleva rechazar el ajuste (aún cuando la modelación de valores extremos no tomó explícitamente a toda la base de datos para sus análisis).

Conclusiones y recomendaciones

1. El análisis de modelación de valores extremos proporciona un ajuste razonable para los datos de los ingresos corrientes de México provenientes de la ENIGH 2004, en el sentido de que las medidas de bondad de ajuste adquieren valores razonables.

2. La modelación de valores extremos comparada con la beta generalizada II y la log-normal parece ser superior en cuanto al ajuste, sólo basados en el análisis de los p-valores del estadístico de Kolmogorov-Smirnov. Sin embargo, y debido a la gran gran dificultad que tienen este y virtualmente todos los estadísticos de bondad de ajuste en cuanto a su precisión en las colas de la distribuciones, los investigadores han optado usualmente por utilizar más bien herramientas gráficas de diagnóstico para el análisis, siendo probablemente las más utilizados los qq-plots. En todos los casos, estas gráficas señalaron que los modelos beta generalizado II y lognormal no son superiores (e incluso muy posiblemente son inferiores) al modelo de Fréchet proveniente de la teoría de valores extremos.

3. Un punto conveniente de la modelación de valores extremos consiste en que simplifica sustancialmente los ajustes al considerar la función generalizada de valores extremos, puesto que engloba la distribución Gumbel, Fréchet y Weibull.

4. Asociado con el punto anterior, cabe destacar que debido al hecho de que el parámetro ξ resultó positivo en todos los casos analizados, y su intervalo de confianza asociado (al 95%) no contenía valores negativos ni al cero, el modelo que mejor ajusta a los datos extremos del ingreso corriente es el Fréchet, desde la óptica de la modelación de valores extremos, del estadístico de mayor orden r y de la modelación de umbral.

5. La modelación de valores extremos además de proporcionar un ajuste razonable a los datos, proporciona toda una serie de estrategias y procedimientos que permiten al usuario establecer el grado de certidumbre con el cual se adopta alguna de las tres modelaciones sugeridas en el punto 3, y a su vez, permite establecer mecanismos analíticos y gráficos intuitivos para verificar la validez del modelo, lo cual hace que este enfoque sea sumamente atractivo para modelar datos extremos.

6. Este trabajo centró su interés únicamente en la variable del ingreso corriente. Existen diferentes variables dentro de la ENIGH que mide segmentos particulares del ingreso (como ingreso por remuneraciones y otras) las cuales por supuesto podrían estudiarse con la modelación de valores extremos y las funciones de ajuste analizadas. De hecho, de las más del centenar de variables que contiene la ENIGH, virtualmente en casi todas podría aplicarse esta modelación, por lo cual se apertura una gran cantidad de investigación asociado con ello.

7. A su vez, este trabajo centró solo su atención en la ENIGH 2004. Podría analizarse de manera separada cada una de las ENIGH a lo largo de los diversos años en que se ha llevado a cabo y/o analizar los datos de manera conjunta. Esto también apertura una gran brecha de investigación al respecto.

8. Más aún, la modelación podría darse de manera multivariada, en el sentido de que no se analice de manera aislada sólo una de las variables, sino con variables que pudieran ser auxiliares a ella. Esto es, si por ejemplo la variable del ingreso corriente es la fundamental en el análisis, podría utilizarse la información de otras variables auxiliares de manera conjunta multivariada. O bien, podría darse el caso de tratar de encontrar la distribución conjunta de las variables de ingreso y gasto de manera simultánea. Nuevamente, por la gran cantidad de combinaciones que pueden darse de subconjuntos de variables, también aquí se tiene una gran cantidad de investigación potencial.

9. En la práctica existe un gran problema referente a la poca disponibilidad que tienen algunos de los informantes a proporcionar información acerca de su ingreso, y esto es especialmente cierto en los informantes de los estratos superiores del ingreso (que puede darse por cuestiones de seguridad o por suspicacias de filtrado de información no deseado). Esto lleva a plantear la necesidad de realizar un ajuste en la distribución del ingreso tomando en cuenta, por ejemplo, la información que proporcionan la oficina de Cuentas Nacionales en México. Este mecanismo a pesar de que ha sido explorado por algunos investigadores, todavía no está lo suficientemente desarrollado tanto en sus cuestiones teóricas como en las operativas. Así, a pesar de que no es un problema que se derive de manera directa del estudiado en este trabajo, parece conveniente explicitar esta necesidad con el objetivo de que este y muchos otros análisis más se vean beneficiados al contar con información más precisa.

10. Un problema adicional es el estudiar los datos expandidos (poblacionales) asociados a la ENIGH. Como se vio en el capítulo 4, la modelación clásica de valores extremos exige la independencia de las observaciones. Al ser los datos expandidos claramente dependientes, no es posible estudiar de esta manera el problema. Sin embargo, puede plantearse esquemas de análisis que involucren sucesiones dependientes en el tiempo del tipo de cadenas de Markov en las cuales se permitan estructuras de dependencia como las ya señaladas. Nuevamente, esto apertura líneas interesantes de investigación.

11. Algunos de los análisis que fueron llevados a cabo se basaron en la consideración de los 32 máximos por entidad federativa, o en los 10 máximos por cada entidad federativa. Esto fue realizado así puesto que no se ve claro que esto podría disolver el supuesto de independencia de las observaciones, y a su vez, proporciona un conjunto de observaciones lo suficiente basto para conformar algún cierto estudio de interés. Por supuesto, existen otras formas de conformar la muestra, por ejemplo, por estratos geográficos, lo cual, nuevamente abre líneas adicionales de investigación de interés.

ANEXO A: GLOSARIO DE TÉRMINOS DE LA ENIGH

El propósito de este anexo es que el lector pueda ubicar rápidamente algunos términos utilizados en la ENIGH, referente sobre todo a las variables y hechos más importantes de los ingresos.

INGRESO TOTAL. Percepciones en efectivo y/o en especie recibidas durante el periodo de referencia a cambio del trabajo asalariado a una empresa, institución a las órdenes de un patrón, incluye el ingreso en efectivo y/o en especie de un negocio agropecuario o no agropecuario, los rendimientos derivados de cooperativas de producción, así como los ingresos derivados de la posesión de activos físicos y no físicos, las transferencias recibidas y otros ingresos corrientes. Comprende las percepciones por retiro de ahorro, la venta de bienes inmuebles, muebles o activos físicos o no físicos, la disposición de capital invertido, las transferencias y financiamientos recibidos, la recuperación de préstamos otorgados a otras unidades ajenas al hogar. Comprende el valor estimado a precios de menudeo, de los productos y servicios recibidos por otros hogares, instituciones sin fines de lucro o por parte del empleo asalariado del autoconsumo o auto suministro. Se consideró la estimación del alquiler de la vivienda que se hubiera tenido que pagar por la vivienda propia.

INGRESO CORRIENTE MONETARIO. Percepciones en efectivo provenientes del trabajo asalariado en una empresa, institución o a las órdenes de un patrón, incluye el ingreso en efectivo y/o en especie de un negocio agropecuario o no agropecuario, los rendimientos derivados de cooperativas de producción, así como los ingresos derivados de la posesión de activos físicos y no físicos, las transferencias recibidas y otros ingresos corrientes.

INGRESO POR REMUNERACIONES AL TRABAJO. Percepciones totales en dinero que recibieron los asalariados determinadas por su participación en actividades de empresas y negocios establecidos en un contrato verbal o escrito con sus empleadores.

• **Ingresos por sueldos, salarios o jornal.** Percepciones en efectivo regulares pagadas a los trabajadores asalariados como retribución al trabajo realizado por éste durante un periodo de tiempo determinado y establecido en un contrato verbal o escrito.

• **Ingresos por destajo.** Percepciones en efectivo en forma regular recibidas por los salariados determinadas por la cantidad de trabajo o servicio que realice o la venta de productos.

• **Ingresos por comisiones y propinas.** Percepciones en efectivo recibidas por los asalariados ya sean pagadas por los empleadores o terceros a cambio de la producción de cierto número de mercancías a la venta o de cierto número de productos y/o servicios realizados.

• **Ingresos por horas extras.** Percepciones en efectivo recibidas por los asalariados como compensación por el tiempo dedicado al trabajo fuera del horario normal por el que fueron contratados, es adicional al sueldo o salario establecido en un contrato verbal o escrito.

• **Ingresos por aguinaldo.** Percepciones extraordinarias en efectivo otorgadas a los trabajadores por parte del patrón, empresa o institución una vez al año.

• **Ingresos por incentivos, gratificaciones o premios.** Pago en efectivo otorgados a los trabajadores asalariados que cumplen con los lineamientos establecidos por la empresa en convenios o programas de trabajo, establecidas en un contrato verbal o escrito.

• **Ingresos por bono, percepción adicional o sobresueldo.** Percepciones en efectivo recibidas por los asalariados ya sea obligatorias establecidas en un contrato verbal o escrito, o bien regulares recibidas como compensación por la responsabilidad del trabajo realizado.

• **Ingresos por primas vacacionales y otras prestaciones.** Percepciones en efectivo recibidas por los asalariados como aporte a su(s) periodo(s) vacacional(es), ayuda de despensa, transporte, útiles escolares, etcétera, por parte de la empresa donde trabaja.

• **Ingresos por reparto de utilidades.** Percepciones en efectivo que reciben los 490 trabajadores asalariados de los beneficios o utilidades que genera la empresa donde trabajan.

INGRESOS POR NEGOCIOS PROPIOS. Percepciones en efectivo o en especie, obtenidas de unidades de producción que no están constituidas como entidades separadas de sus propietarios y no llevan una contabilidad completa incluida el balance entre ingresos y gastos.

• **Ingresos por negocios industriales.** Percepciones en efectivo provenientes de cualquier actividad relacionada con la transformación mecánica, física o química de materiales o substancias con el fin de obtener productos nuevos. Incluyen las actividades de maquila, el ensamble de partes y componentes o productos fabricados, la instalación en construcciones de equipo y materiales prefabricados, la construcción de obras en combinación con actividades de servicios, incluye la extracción de petróleo y gas de minerales metálicos y no metálicos. La minería incluye la explotación de canteras, operaciones en pozos, operaciones de beneficio, la explotación y a las actividades de preparación y acondicionamiento de las minas, así como a los servicios de apoyo a la producción y explotación de minas y la instalación o desmantelamiento de torres de perforación. Se deberá considerar también la generación, transmisión y suministro de energía eléctrica para su venta; la captación, potabilización y suministro

de agua potable y residual así como, el suministro de gas por ductos al consumidor final.

• **Ingresos por negocios comerciales.** Percepciones en efectivo provenientes de cualquier actividad con la compra-venta (sin transformación) de bienes de consumo intermedio (como bienes de capital, materias primas, suministros utilizados en la producción, y bienes de consumo final) para ser vendidos a otros comerciantes, distribuidores, fabricantes y productores de bienes y servicios. Incluidas las actividades que sólo realizan una parte de este proceso (la compra o la venta) agentes importadores y exportadores, así como los servicios integrados a la venta de los bienes, como clasificación inventariado, embalaje, empacado y etiquetado.

Los comerciantes minoritarios que venden o promueven la compra-venta a cambio de una comisión o pago.

• **Ingresos por prestación de servicios.** Percepciones en efectivo provenientes de diversos servicios.

• **Ingresos por negocio agrícola.** Percepciones en efectivo provenientes de las actividades (en terrenos, predios o parcelas, patios, azoteas, huertas, invernaderos y viveros) relacionados con la explotación de especies vegetales cultivadas con el fin de obtener alimentos para consumo humano y animal, así como las materias primas para la industria y servicios.

• **Ingresos por negocios de cría, explotación y productos derivados de animales.** Percepciones en efectivo provenientes de actividades relacionadas a la explotación de animales en todas sus fases, incluyendo la cría, explotación, engorda y uso de ganado bovino, porcino, ovino, caprino, equino, etcétera, la explotación en ambientes controlados de avicultura (aves), cunicultura (conejos), animales de pelaje fino (chinchillas, zorros, llamas), abejas, cría de animales con otros fines como perros y gatos, gallos de pelea, toros de lidia y ratones para experimentos. La administración de empresas ganaderas, alquiler de maquinaria y equipo con operador, trasquila, inseminación artificial, inspección sanitaria, albergue, castración, limpieza, recolección de estiércol, baños, parasiticidas, clasificación de huevo, registro de pedigrí, herraje de caballos, cruza y marcado de animal y nubulización de ganado y la acuicultura animal, en ambiente controlado.

• **Ingresos por negocios de reproducción, recolección de productos forestales y tala de árboles.** Percepciones en efectivo provenientes de actividades de plantación, reforestación y conservación con el propósito de realizar la venta en pie de árboles maderables, la recolección de productos forestales como goma, resinas, fibras y heno y a la tala de árboles en superficies forestales y sus actividades en el mismo lugar como descortezado, producción de trozos, astillas y rajas de madera.

• **Ingresos por negocios de pesca, caza y captura de animales.** Percepciones en efectivo provenientes de la pesca, caza y captura de animales en su hábitat natural y a la operación y administración de reservas para caza.

INGRESOS POR COOPERATIVAS. Percepciones en efectivo provenientes de los rendimientos o ganancias generadas en un periodo de tiempo determinado por la administración, gestión y distribución de los beneficios de una empresa constituida como cooperativa.

• **Ingresos por sueldos o salarios.** Percepciones recibidas por pertenecer a una cooperativa, además de trabajar activamente en el desarrollo de la misma.

• **Ingresos por ganancias o utilidades.**
Percepciones en efectivo que se tienen por derecho como resultado de poner su capital (dinero, bienes, maquinaria, equipo, tierras, etcétera) o trabajo a disposición de la cooperativa.

INGRESOS POR SOCIEDADES. Percepciones en efectivo que recibieron por ser propietarios de manera colectiva de una sociedad.

• **Sociedad.** Negocio o empresa creada con el fin de producir bienes y/o servicios para el mercado que pueden ser fuente de beneficios monetarios y de otras ganancias para sus propietarios; es propiedad de varias personas que aportaron capital (dinero, bienes, maquinaria, equipo, tierras, etcétera) o trabajo.

• **Ingresos por ganancias o utilidades.** Percepciones en efectivo a que se tiene derecho como resultado de poner su capital (dinero, bienes, maquinaria, equipo, tierras, etcétera) o trabajo a disposición de las sociedades.

INGRESOS POR EMPRESAS QUE FUNCIONAN COMO SOCIEDADES. Percepciones en efectivo obtenidas de unidades de producción que están constituidas como entidades separadas de sus propietarios y que llevan una contabilidad completa, incluido el balance entre ingresos y gastos, pero ante la ley no están registradas como sociedades.

• **Empresas que funcionan como sociedad.** Unidades de producción que están constituidas como entidades separadas de sus propietarios que llevan una contabilidad completa, incluido el balance entre ingresos y gastos, ante la ley no están registradas como sociedades.

• **Ingresos por ganancias o utilidades.** Percepciones en efectivo a que se tiene derecho como resultado de poner su capital (dinero, bienes, maquinaria, equipo, tierras, etcétera) o trabajo a disposición de este tipo de empresas.

INGRESOS POR RENTA DE LA PROPIEDAD. Percepciones en efectivo recibidas a cambio de poner a disposición de otros (hogares, empresas, etcétera) su dinero, valores, bienes o propiedades.

• **Ingresos por intereses provenientes de inversiones a plazo fijo.** Percepciones en efectivo que recibieron instituciones financieras por ser poseedores o titulares de una cuenta de inversión o plazo fijo en un periodo de tiempo determinado sin reducir el total de dicha inversión.

• **Ingresos por intereses provenientes de cuentas de ahorro.** Percepciones en efectivo que recibieron de las instituciones por ser poseedor o el titular de una cuenta de ahorro en un periodo determinado sin reducir el total del ahorro de dicha cuenta.

• **Ingresos por intereses provenientes de préstamos a terceros.** Percepciones en efectivo que se recibieron de personas ajenas al hogar los cuales se comprometieron a pagar como resultado de préstamos realizados.

• **Ingresos por rendimientos provenientes de acciones o dividendos, bonos y cédulas.** Percepciones en efectivo que las instituciones financieras, empresas o sociedades se comprometen a pagar al poseedor de un documento de valor a largo plazo o un documento donde se reconoce una deuda u obligación. Sin que este monto sea parte del valor del capital invertido en dichos documentos.

• **Ingresos por el alquiler de marcas, patentes y derechos de autor.** Regalías recibidas por poner a disposición de terceros para su explotación, distribución y comercialización, los artículos o servicios patentados por los propietarios.

INGRESOS POR TRANSFERENCIAS. Percepciones regulares o frecuentes recibidas en efectivo o a la entrega de depósitos transferibles, provenientes de instituciones, empresas en otros hogares, sin proporcionar a cambio contrapartida alguna.

• **Ingresos por jubilaciones y/o pensiones originadas dentro del país.** Transferencias en dinero o en depósitos transferibles que se reciben como consecuencia de una jubilación y/o pensión de seguridad social provenientes de otros hogares, instituciones o empresas que están dentro del país sin contrapartida alguna.

• **Ingresos por jubilaciones y/o pensiones provenientes de otros países.** Transferencias en dinero o en depósitos transferibles que se reciben como consecuencia de una jubilación y/o pensión de seguridad social provenientes de otros hogares, instituciones o empresas que están fuera del país sin contrapartida alguna.

• **Ingresos por indemnizaciones recibidas de seguros contra riesgos a terceros.** Percepciones en efectivo recibidas de seguros contra riesgos, por accidentes, enfermedades, incendios, inundaciones o siniestros similares, ocasionados por personas ajenas al hogar, a sus bienes o incluso a su propia persona.

• **Ingresos por ganancias o utilidades.** Percepciones en efectivo a que se tiene derecho como resultado de poner su capital (dinero, bienes, maquinaria, equipo, tierras, etcétera) o trabajo a disposición de la cooperativa.

• **Ingresos por indemnizaciones por accidente de trabajo.**
• **Ingresos por indemnizaciones por despido y retiro voluntario.**
• **Ingresos por becas provenientes de organizaciones no gubernamentales.** Transferencias regulares o frecuentes sin contrapartida, que se reciben como apoyo al desempeño académico de organizaciones gubernamentales.

• **Ingresos por becas provenientes del gobierno.** Transferencias regulares o frecuentes sin contrapartida, que se reciben como apoyo al desempeño académico provenientes del gobierno.

• **Ingresos por donativos provenientes de organizaciones no gubernamentales.** Transferencias regulares frecuentes sin contrapartida, provenientes de organizaciones no gubernamentales.

• **Ingresos por regalos provenientes del gobierno y/o donativos en dinero provenientes de otros hogares.** Transferencias en dinero o en depósitos transferibles de otros hogares sin contrapartida alguna.

• **Ingresos provenientes de otros países.** Transferencias en efectivo recibidas de personas o instituciones que residen fuera del país.

• **Ingresos por beneficio de PROGRESA U OPORTUNIDADES.** Percepciones en efectivo derivadas del beneficio directo para la salud, alimentación y educación proporcionados por la Secretaría de Desarrollo Social (SEDESOL), a través del programa PROGRESA U OPORTUNIDADES.

• **Ingresos por beneficio de PROCAMPO.** Percepciones en efectivo derivadas como beneficio directo al campo para las actividades agrícolas.

OTROS INGRESOS CORRIENTES NO CONSIDERADOS EN LOS ANTERIORES. Percepciones en efectivo, provenientes de fuentes ajenas al trabajo, renta de la propiedad o transferencia y que no provienen de la venta de bienes propiedad del hogar.

INGRESOS POR PERCEPCIONES FINANCIERAS Y DE CAPITAL MONETARIAS. Percepciones en efectivo recibidas por retiro de ahorro, la venta de inmuebles, muebles o activos físicos o no físicos por la disposición de capital invertido, las transferencias y los financiamientos recibidos, la recuperación de préstamos otorgados a otras unidades ajenas al hogar.

• **Ingresos por retiro de inversiones, ahorros, tandas, cajas de ahorro, etcétera.** Percepciones en efectivo obtenidas del retiro de dinero de cuentas de inversión en instituciones financieras o de particulares.

• **Ingresos por préstamos recibidos de personas que no pertenecen al hogar o instituciones.** Percepciones en efectivo derivadas de financiamientos provenientes de instituciones financieras privadas, personas ajenas al hogar o de otra índole.

• **Ingresos por venta de acciones, bonos y cédulas.** Percepciones en efectivo que obtienen a cambio de vender a personas ajenas al hogar los documentos que les hacen participar en la propiedad de una empresa constituida en sociedad y que les da derecho a recibir parte de los beneficios o ganancias de la empresa, o bien donde se hace constar la propiedad de un valor o se reconoce una deuda u obligación.

• **Ingresos por venta de marcas, patentes y derechos de autor.** Ingresos obtenidos por el hecho de otorgar a personas ajenas al hogar o instituciones, los derechos de vender, reproducir o representar una obra, o bien por la venta del registro o los derechos de patentes o marcas de algún invento.

INGRESOS POR PERCEPCIONES FINANCIERAS Y DE CAPITAL NO MONETARIAS. Valor estimado a precios de menudeo de los bienes de capital recibidos por otros hogares o por parte del empleo asalariado, del autoconsumo o auto suministro. Se clasifican en:

• **Autoconsumo:** Estimación con base al valor en el mercado a precio de menudeo, de los bienes de capital que los hogares tomaron de su propia producción o de su negocio comercial.

• **Pago en especie:** Estimación con base al valor en el mercado a precio de menudeo, de los bienes de capital que recibieron los trabajadores asalariados, a cambio de su trabajo. Se incluye el valor de los productos que recibieron los trabajadores por cuenta propia o los patrones por un trabajo realizado.

• **Regalos:** Estimación con base al valor en el mercado a precio de menudeo, de los bienes de capital que fueron recibidos como regalo de personas que no eran miembros del hogar o de prestaciones sin fines de lucro.

HOGAR. Conjunto formado por una o más personas que residen habitualmente en la misma vivienda y se sostienen de un gasto común principalmente para alimentarse y pueden ser parientes o no.

• **Hogar principal.** Es aquel en que alguno de sus residentes de la vivienda se declara como dueño de la vivienda; si es rentada, con el que se haya hecho contrato escrito o verbal del arrendamiento; al que le prestan la vivienda, o que la recibió como prestación por parte de su trabajo.

CLASE DE HOGAR. Diferenciación de los hogares a partir del tipo de relación consanguínea, legal, de afinidad o de costumbre entre el jefe(a) y los otros integrantes del hogar, sin considerar a los trabajadores domésticos y a los familiares de éstos ni a los huéspedes.

Se clasifican en:
• **Unipersonal:** Hogar formado por una sola persona que es el jefe(a).
• **Nuclear:** Hogar constituido por un solo grupo familiar primario.
• **Ampliado:** Hogar formado por el jefe(a) y su grupo familiar primario más otros grupos familiares u otros parientes.
• **Compuesto:** Hogar formado por un hogar nuclear o ampliado con personas sin parentesco con el jefe(a).
• **De co-residentes:** Hogar formado por dos o más personas que no tienen parentesco con el jefe(a).

VIVIENDA. Espacio delimitado por paredes y techos de cualquier material de construcción donde viven, duermen, preparan alimentos, comen y se protegen de las inclemencias del tiempo una o más personas. La entrada debe ser independiente, es decir, que sus ocupantes puedan entrar o salir de ella sin pasar por el interior de otra vivienda.

ANEXO B: NOTA METODOLÓGICA DE LA ENIGH

B.1 OBJETIVO

La Encuesta Nacional de Ingresos y Gastos de los Hogares (ENIGH) es un proyecto de generación de estadísticas en el INEGI, que tiene como objetivo proporcionar información sobre la distribución, monto y estructura del ingreso y gasto de los hogares.

Así mismo, permite generar información de la estructura del ingreso corriente de los hogares, según la fuente de donde provenga; la estructura del gasto corriente en la adquisición de bienes de consumo final (duraderos y no duraderos); el valor de los bienes y servicios que auto consumen los hogares, el pago en especie y los regalos recibidos, así como la estructura de las erogaciones y las percepciones financieras y de capital.

También es el campo de estudio para conocer las características sociodemográficas, la condición de actividad y las características ocupaciones de los integrantes del hogar de 12 y más años a la vez que se estudian las características de infraestructura de la vivienda y equipamiento del hogar.

B.2 ANTECEDENTES

La Encuesta Nacional de Ingresos y Gastos de los Hogares surge en el año de 1984; a partir de 1992 se realiza con una periodicidad de levantamiento de cada dos años, con excepción de 2005, ya que fue un levantamiento extraordinario para tener cifras actualizadas sobre las condiciones de vida de los hogares. Al tiempo que se conservó la comparabilidad del marco conceptual, periodos de referencia, unidades de análisis, cobertura geográfica, instrumentos de captación, diseño muestral y procedimientos operativos utilizados en la generación de datos, también se actualizaron otros aspectos relevantes como la metodología, y se incorporaron nuevos productos con el objetivo de adecuarse a los cambios económicos del país y obtener resultados que reflejan la realidad.

B.3 IMPORTANCIA

Disponer de la información estadística que genera la ENIGH, permite conocer el nivel de bienestar de la población, bajo la consideración de que el monto del ingreso, su procedencia y forma de distribución lo condiciona en gran medida.

Los resultados de la ENIGH son utilizados para distintos fines, entre los cuales se pueden mencionar los siguientes:

- Generación de ponderadores para la realización del Índice Nacional de Precios al Consumidor.
- Construcción de indicadores para el estudio de la pobreza.
- Cálculo de estadísticas sobre los niveles de vida.
- Estudios del comportamiento de la economía nacional en el ámbito de la economía de los hogares y comparativos con otros países.

B.4 UNIDAD DE OBSERVACIÓN

La unidad de observación que la ENIGH considera para su estudio es el "Hogar", el cual se define como el conjunto formado por una o más personas que residen habitualmente en la misma vivienda y se sostienen de un gasto común, principalmente para alimentarse y pueden ser parientes o no.

B.5 MÉTODO DE CAPTACIÓN

La generación de estadísticas de la ENIGH se basa en la aplicación de un esquema de muestreo probabilístico, a su vez el diseño es bietápico, estratificado y por conglomerados, donde la unidad última de selección es la vivienda y la unidad de observación es el hogar y en consecuencia los resultados obtenidos de la encuesta se generalizan a toda la población.

B.5.1 MARCO DE MUESTREO

El marco muestral utilizado es el de propósitos múltiples del INEGI, constituido por la información demográfica y cartográfica obtenida a partir del levantamiento del Censo de Población y Vivienda del 2000.

En este marco de muestreo se excluyen a todas las viviendas colectivas y las de diplomáticos extranjeros, ya que para fines de la encuesta no son objeto de estudio.

B.5.2 SELECCIÓN Y TAMAÑO DE LA MUESTRA

El procedimiento para conformar la muestra se realiza de la siguiente manera:

- Se seleccionan las viviendas, distinguiendo entre dos clases; la vivienda particular y la colectiva, siendo objeto de la encuesta sólo la primera.

- Se realiza en forma independiente para cada entidad y estrato; el procedimiento varía dependiendo la zona. La selección de la muestra, está calculada para dar estimaciones a los siguientes niveles de desagregación:
- Nivel nacional.
- Localidades de 2 500 y más habitantes.
- Localidades de menos de 2 500 habitantes

MUESTRAS ENIGH 2000 – 2005			
AÑO	**REFERENCIA**	**MUESTRA**	**PORCENTAJE**
2000	NACIONAL	11,657	100.00
	1	7,784	66.78
	2	3,873	33.22
2002	NACIONAL	19,856	100.00
	1	14,539	73.22
	2	5,317	26.78
2004	NACIONAL	25,115	100.00
	1	19,190	76.41
	2	5,925	23.59
2005	NACIONAL	25,443	100.00
	1	18,490	72.67
	2	6,953	27.33

1 Localidades de 2 500 y más viviendas
2 Localidades de menos de 2 500 viviendas

B.6 DIAGRAMA GENERAL DEL PROCESO

Para llevar a cabo la generación de las estadísticas se realizó una serie de tareas que van desde la planeación hasta la presentación de resultados, para lo cual se toma como base la estructura del INEGI que cuenta con oficinas centrales en Aguascalientes, donde se realiza la planeación general, el diseño técnico y metodológico; diez oficinas regionales, que realizan labores de planeación regional y seguimiento del trabajo de campo y las oficinas del Instituto en las 32 entidades federativas del país, las cuales realizan el operativo de campo con las fuentes informantes y la captura-validación de los cuestionarios.

En la figura B.1 de la página siguiente se describe el proceso para la generación de las estadísticas de la ENIGH, con el detalle de las actividades realizadas, de acuerdo a la estructura del INEGI y su realización con las fuentes informantes.

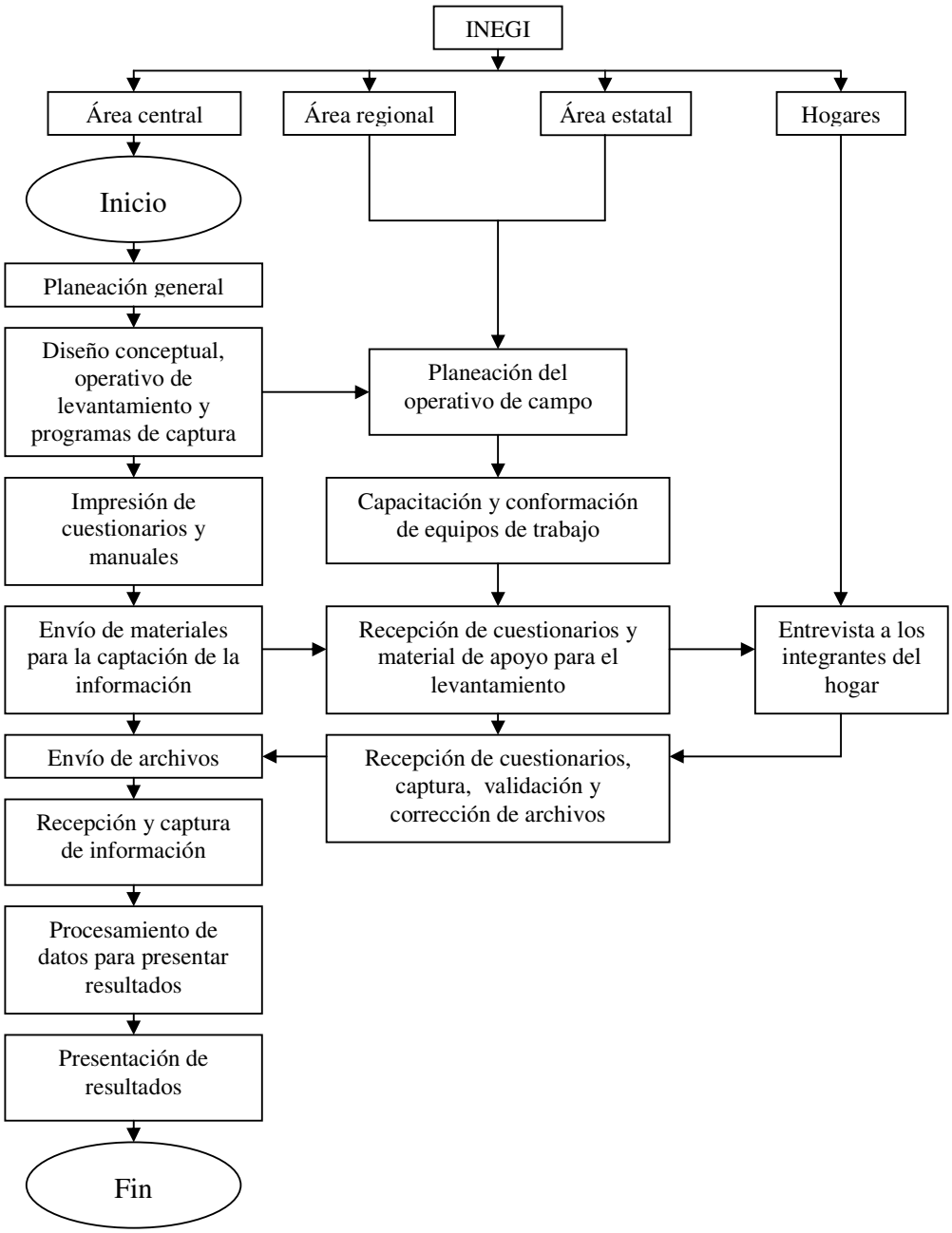

Figura B.1 *El proceso para la generación de las estadísticas de la ENIGH*

ANEXO C: VEROSIMILITUD PERFIL Y TEMAS AFÍNES

Por conveniencia del lector, se presenta en este anexo algunas consideraciones sobre la verosimilitud perfil y algunos temas afines que puedan ayudar a aclarar algunas cuestiones al respecto.

C.1 INTRODUCCIÓN

Considérese un conjunto de observaciones $x_1,...,x_n$ que pueden plantearse de manera vectorial simplemente como $\mathbf{x}$. Si $\mathbf{f}$ representa la función de densidad conjunta de estos datos, y Θ es el vector de parámetros (usualmente desconocido), se define entonces la **verosimilitud** como

$$L(\Theta;\mathbf{x}) = L(\Theta) = \mathbf{f}(\mathbf{x};\Theta). \qquad \text{(C.1)}$$

Se enfatiza el hecho de que la verosimilitud se considera como una función de los parámetros desconocidos, y no tanto función de las observaciones. Este es el motivo del porque la escritura de la primera igualdad, en la cual se enfatiza la dependencia funcional sobre el parámetro y no sobre los datos.

Usualmente se considera, sobre por simplicidad operativa, que las $x_i's$ son variables independientes entre sí, y además, si es el caso que son idénticamente distribuidas, entonces (C.1) adquiere la forma siguiente:

$$L(\Theta) = \prod_{i=1}^{n} f(x_i;\Theta).$$

Esta factorización se debe al hecho de la independencia, y la escritura de la función sin texto negro enfatiza el hecho de la igualdad de la distribución de las observaciones.

Ahora, la filosofía básica de la estimación por máxima verosimilitud es la siguiente. Si se ha seleccionado un modelo f del cual se piensa que provienen los datos, y dado que $x_1,...,x_n$ es una muestra, la pregunta sería: ¿Cuáles son los valores de Θ que hacen que la probabilidad de haber obtenida tales observaciones sea máxima? Pensado así, se tiene el esquema básico de la estimación por verosimilitud. En este sentido, lo que ahora se trataría de hacer es maximizar la función de verosimilitud para cada uno de los componentes del vector Θ considerando como dadas las observaciones.

De manera operativa, es frecuentemente más sencillo encontrar el máximo del logaritmo de la función de verosimilitud, puesto que por una parte es más fácil derivar sumas que productos, y por la otra con frecuencia se tienen potencias, las cuales se reducen a productos cuando se toman logaritmos. Así, se considera ahora la función de **log-verosimilitud** definida como:

$$\ell(\Theta) = \log L(\Theta) = \sum_{i=1}^{n} \log f(x_i; \Theta).$$

Esto no implica ningún problema en cuanto a la localización de los puntos críticos de la función de verosimilitud original, puesto que el logaritmo es una función monótona creciente.

Un beneficio importante de adoptar la máxima verosimilitud como un principio para la estimación de parámetros es que están disponibles un gran número de aproximaciones estándar y ampliamente aplicables de distribuciones de muestreo. Esto conlleva a aproximaciones razonables para los errores estándar e intervalos de confianza. A su vez, se tienen varios resultados útiles acerca de los estimadores de máxima verosimilitud. Supóngase que se tiene una muestra de realizaciones independientes $x_1, ..., x_n$ de una variable aleatoria X que tiene una distribución $F \in \mathcal{F}$. La familia $\mathcal{F}$ está indexada por un parámetro $d-$dimensional θ y la verdadera distribución F tiene $\theta = \theta_0$. El estimador de máxima verosimilitud de θ_0 se denota por $\hat{\theta}_0$.

Estrictamente hablando, cada uno de los resultados es una ley límite asintótica que se obtiene al considerar que el tamaño de muestra n tiene al infinito. Estas son sólo válidas sobre ciertas condiciones de regularidad. Se supone que estas condiciones son válidas y proporcionan sus resultados como aproximaciones cuya exactitud se incrementa a medida que n lo hace. Se presenta ahora un teorema fundamental que refiere hechos acerca del comportamiento aproximadamente normal de los estimadores de máxima verosimilitud.

Teorema C.1. Sea $x_1, ..., x_n$ realizaciones independientes de una distribución dentro de una familia paramétrica $\mathcal{F}$, y sean $\ell(\cdot)$ y θ_0 la función de log-verosimilitud y el estimador de máxima verosimilitud del parámetro θ_0 de un modelo $d-$dimensional. Entonces, bajo ciertas condiciones apropiadas de regularidad, y para una n grande, se tiene que:

$$\hat{\theta}_0 \dot{\sim} \text{MVN}_d(\theta_0, I_E(\theta_0)^{-1}),$$

donde el símbolo $\dot{\sim}$ indica que se distribuye de manera aproximadamente igual, MVN_d indica una función multivariada normal $d-$dimensional con media en el primer parámetro y matriz de varianzas y covarianzas en el segundo parámetro, la cual está dada por:

$$I_E(\theta)=\begin{bmatrix} e_{1,1}(\theta) & \cdots & e_{1,d}(\theta) \\ \vdots & \ddots & \vdots \\ e_{d,1}(\theta) & \cdots & e_{d,d}(\theta) \end{bmatrix},$$

Con

$$e_{i,j}(\theta)=E\left\{-\frac{\partial^2}{\partial\theta_i\partial\theta_j}\ell(\theta)\right\}.$$

La matriz $I_E(\theta)$, la cual mide el valor esperado de la curvatura de la superficie de log-verosimilitud, se conoce usualmente como la **matriz esperada de información**.

El teorema anterior puede ser utilizado para obtener de manera aproximada intervalos de confianza para componentes individuales de $\theta_0=(\theta_1,...,\theta_d)$. Denotando un término arbitrario en la inversa de $I_E(\theta_0)$ por $\psi_{i,j}$, se sigue de las propiedades de la distribución normal multivariada que, para una n grande

$$\hat{\theta}_i \dot{\sim} N(\theta,\psi_{i,j}).$$

Por lo tanto, si $\psi_{i,j}$ son conocidos, un intervalo de confianza aproximado del $(1-\alpha)100\%$ de confianza para θ_i se encuentra dado por

$$\hat{\theta}_i \pm z_{\alpha/2}\sqrt{\psi_{i,j}}, \qquad \text{(C.2)}$$

donde $z_{\alpha/2}$ es el $(1-\alpha/2)$ cuantil de la distribución normal estándar. Debido a que el verdadero valor de $\psi_{i,j}$ es frecuentemente desconocido, es usual el aproximar los términos de I_E mediante la matriz de información observada, definida por:

$$I_O(\theta)=\begin{bmatrix} -\frac{\partial^2}{\partial\theta_1^2}l(\theta) & \cdots & -\frac{\partial^2}{\partial\theta_1\partial\theta_d}l(\theta) \\ \vdots & \ddots & \vdots \\ -\frac{\partial^2}{\partial\theta_d\partial\theta_1}l(\theta) & \cdots & -\frac{\partial^2}{\partial\theta_1^2}l(\theta) \end{bmatrix}$$

y evaluada en $\theta=\hat{\theta}$. Denotando los términos de la inversa de esta matriz por $\tilde{\psi}_{i,j}$, se sigue que una intervalo de confianza aproximado del $(1-\alpha)100\%$ para θ_i es

$$\tilde{\theta}_i \pm z_{\alpha/2}\sqrt{\tilde{\psi}_{i,j}}.$$

A pesar de la aproximación adicional, estos intervalos son frecuentemente más precisos que los que se obtienen por (C.2).

A pesar de que la familia parámetrica $\mathscr{F}$ podría estar indicada por un parámetro θ, del cual θ_0 representa su verdadero valor, podría ser el caso que θ_0 no fuera de particular interés sino alguna cierta función de él $\phi_0 = g(\theta_0)$ que se busca estimar, donde ϕ_0 podría tener una dimensión diferente a θ_0. Se restringe la atención a la situación donde ϕ_0 es una función escalar de θ_0. Esto es frecuentemente útil en la modelación de valores extremos, donde θ_0 es el vector de parámetros de una distribución representativa del comportamiento de valores extremos, pero se necesita la probabilidad de algún evento extremo –el cual es una función de θ_0– . Los siguientes dos resultados permiten inferencias de máxima verosimilitud sobre θ_0 el cual se transforma para proporcionar inferencias sobre θ_0.

Teorema C.2. Si $\hat{\theta}_0$ es el estimador de máxima verosimilitud para θ_0, y $\phi = g(\theta)$ es una función escalar, entonces el estimador de máxima verosimilitud para ϕ_0 está dado por $\hat{\phi}_0 = g(\hat{\theta}_0)$.

Este resultado significa que, una vez que el estimador de máxima verosimilitud para θ_0 ha sido calculado, el estimador de máxima verosimilitud para cualquier función de θ_0 se obtiene por simple sustitución.

Teorema C.3. Si $\hat{\theta}_0$ es un estimador de máxima verosimilitud proveniente de una muestra grande de un parámetro $d-$dimensional θ_0 con una matriz de varianzas y covarianzas aproximada V_θ. Entonces si $\phi = g(\theta)$ es una función escalar, el estimador de máxima verosimilitud de $\phi_0 = g(\theta_0)$ satisface

$$\hat{\phi}_0 \dot{\sim} N(\phi_0, V_\phi),$$

donde

$$V_\phi = \nabla\phi^T V_\theta \nabla\phi,$$

con

$$\nabla\phi = \left[\frac{\partial\phi}{\partial\theta_1},...,\frac{\partial\phi}{\partial\theta_1}\right]^T$$

evaluado en $\hat{\theta}_0$.

El teorema C.3 es usualmente referido como el **método delta**. En la misma forma en que la normalidad aproximada de $\hat{\theta}_0$ puede ser utilizada para obtener intervalos de confianza para los componentes individuales de θ_0, el método delta permite que la

normalidad aproximada de $\hat{\phi}_0$ pueda ser utilizado para obtener intervalos de confianza para ϕ_0.

C.2 INFERENCIA APROXIMADA UTILIZANDO LA FUNCIÓN DE DEVIANCIA

Un método alternativo para cuantificar la incertidumbre en el estimador de máxima verosimilitud se basa en la función de deviancia, definida como

$$D(\theta) = 2\{\ell(\hat{\theta}_0) - \ell(\theta)\}. \tag{C.3}$$

Valores de θ con una deviancia pequeña corresponden a modelos con una alta verosimilitud, y por lo tanto un criterio natural para regiones de confianza es especificarlas como una región de confianza

$$C = \{\theta : D(\theta) \leq c\}$$

Para alguna selección de c. Idealmente, se intentaría seleccionar c de manera tal que la región correspondiente C tenga alguna probabilidad pre-establecida, por ejemplo $(1-\alpha)100\%$, de que contenga el verdadero parámetro poblacional θ_0. En general esto no es posible debido a que se requiere el conocimiento de la distribución poblacional. Aun si esta distribución fuese conocida, los cálculos exactos necesitar determinar la distribución de $D(\theta)$, los cuales son usualmente intratables algebraicamente. Estas dificultades se resuelven usualmente mediante la utilización de la distribución de muestreo que es válida para grandes tamaños de muestra.

Teorema C.4. Sea $x_1,...,x_n$ realizaciones independientes de una distribución dentro de una familia paramétrica $\mathcal{F}$, y sea $\hat{\theta}_0$ el estimador de máxima verosimilitud del parámetro del modelo $d-$dimensional θ_0. Entonces, para una n grande, bajo condiciones de regularidad apropiadas, la función de deviancia (C.3) satisface

$$D(\theta_0) \dot{\sim} X_d^2.$$

Se sigue del teorema C.4 que una región de confianza del $(1-\alpha)100\%$ para θ_0 está dado por

$$C_\alpha = \{\theta : D(\theta) \leq c_\alpha\},$$

Donde c_α es el $(1-\alpha)$ cuantil para la distribución χ_d^2. Esto aproximación es usualmente más exacta que aquella que se basa en la normalidad asintótica del estimador de máxima verosimilitud, aunque la carga computacional también es mayor.

C.3 INFERENCIAS UTILIZANDO LA FUNCIÓN DE VEROSIMILITUD PÉRFIL

Hasta el momento se ha descrito un método para realizar inferencias de un componente particular θ_i de un vector de parámetros θ. Una alternativa, y usualmente más precisa, es el método de la verosimilitud perfil. La log-verosimilitud para θ puede escribirse formalmente como $\ell(\theta_i, \theta_{-i})$, donde θ_{-i} denotan todos los componentes de θ excluyendo θ_i. La **log-verosimilitud** para θ_i está definida como

$$\ell_p(\theta_i) = \max_{\theta_{-i}} \ell(\theta_i, \theta_{-i}).$$

Esto es, para cada valor de θ_i, la log-verosimilitud perfil es la log-verosimilitud maximizada con respecto a todos los otros componentes de θ. En otras palabras, $\ell_p(\theta_i)$ es el perfil de la superficie de log-verosimilitud vista desde el eje θ_i.

Esta definición generaliza la situación en la cual θ puede ser particionado en dos componentes, $(\theta^{(1)}, \theta^{(2)})$, del cual $\theta^{(1)}$ es un vector $k-$dimensional de interés y $\theta^{(2)}$ corresponde a las restantes $(k-d)$ componentes. La log-verosimilitud perfil para $\theta^{(1)}$ se define ahora como

$$\ell_p(\theta^{(1)}) = \max_{\theta^{(2)}} \ell(\theta^{(1)}, \theta^{(2)}).$$

Si $k=1$ esto se reduce a la definición previa.

El siguiente resultado, el cual es una generalización del teorema C.4, proporciona un procedimiento de inferencias aproximadas sobre el estimador de máxima verosimilitud $\theta^{(1)}$.

Teorema C.5. Sea $x_1, ..., x_n$ realizaciones independientes de una distribución dentro de una familia paramétrica $\mathcal{F}$, y sea $\hat{\theta}_0$ el estimador de máxima verosimilitud del parámetro del modelo $d-$dimensional $\theta_0 = (\theta^{(1)}, \theta^{(2)})$, donde $\theta^{(1)}$ es un subconjunto $k-$dimensional de θ_0. Entonces, para una n grande y bajo condiciones de regularidad apropiadas, se tiene que

$$D_p(\theta^{(1)}) = 2\{\ell(\hat{\theta}_0) - \ell_p(\theta^{(1)})\} \dot{\sim} X_k^2.$$

El teorema C.5 es frecuentemente utilizado en dos situaciones diferentes. Primero, para una componente simple de θ_i, $C_\alpha = \{\theta_i : D_p(\theta_i) \le c_\alpha\}$ es un intervalo del $(1-\alpha)100\%$ de confianza, donde c_α es el cuantil $(1-\alpha)$ de una distribución χ_k^2. Para verificar si M_0 es un modelo reducido plausible de M_1, es suficiente con ver si el 0 está en C_α, el

cual es equivalente a ver si $D < c_\alpha$. Esto se llama la **prueba de la razón de verosimilitud**, como se resume en el siguiente resultado.

Teorema C.6. Supóngase que M_0 con parámetro $\theta^{(2)}$ es el sub-modelo de M_1 con parámetro $\theta_0 = (\theta^{(1)}, \theta^{(2)})$ bajo la restricción de que el sub-vector $k-$dimensional $\theta^{(1)} = 0$. Sean $\ell_0(M_0)$ y $\ell_1(M_1)$ los valores maximizados de la log-verosimilitud para los modelos M_0 y M_1 respectivamente. Una prueba de la validez del modelo M_0 relativa a M_1 a un nivel de significancia α es el rechazar M_0 a favor de M_1 si $D = 2\{\ell_1(M_1) - \ell_0(M_0)\} > c_\alpha$, donde c_α es el cuantil $(1-\alpha)$ de la distribución χ^2_k.

Finalmente, bajo condiciones de regularidad, cada una de las aproximaciones de muestras grandes descritas es valida cuando $x_1, ..., x_n$ son independientes aunque no necesariamente distribuidas de manera idéntica por una familia indexada por un parámetro θ. Por ejemplo, en un modelo clásico de regresión, $X_i \sim D(\alpha + \beta w_i)$ para $i = 1, ..., n,$ donde $D(\theta)$ denota una distribución con parámetro θ y $w_1, ..., w_n$ son constantes independientes. Aunque cada una de las X_i tiene una distribución diferente, el estimador de máxima verosimilitud de (α, β) satisface las propiedades de grandes muestras establecidas en los teoremas anteriores.

FUENTES DE INFORMACIÓN

Capítulo 1

[1] **Sampieri,** et. al, *Metodología de la Investigación*, McGraw-Hill, México, D.F., 2002.

[2] Aspectos metodológicos de la Encuesta Nacional de Ingreso y Gasto de los Hogares. http://www.inegi.gob.mx/est/default.aspx?c=2604

Capítulo 2

[3] **Aitchison**, J. & **Brown**, J.A.C. (1957), *The Lognormal Distribution,* Cambridge Univeristy Press, Cambridge.

[4] **Ammon**, O. (1895), *Die Gesellschaftsordnung und ihre Naturlichen Grundlagen,* Jena.

[5] **Amoroso**, L. (1924-1925), *Richerche Intorno alla Curva dei Redditi,* Annali di Matematica Pura ed Applicata, Series 4-21, vol. II, 123-157.

[6] **Angle**, J. (2002), *The Salamander: A Model of the Right Tail of the Wage Distribution Truncated by Top coding., The Pareto, Zipf and other power laws,* Liga de Internet: http://www.fcsm.gov/03papers/Angle_Final.pdf

[7] **Angle**, J. (2006), *The Inequality Process as a Wealth Maximizing Process, the Pareto, Zipf and other power laws,* Liga de Internet: http://adsabs.harvard.edu/abs/2006APS..MARA33005A

[8] **Arnold**, B. (1980), *Pareto Distributions: Pareto and Related Heavy-tailed Distributions,* University of California at Riverside.

[9] **Atoda**, N., **Suruga,** T. & **Tchibanaki**, T. (1980), *Statistical Inference of Functional Forms for Income Distribution,* Kyoto University.

[10] **Bandourian**, R., **McDonald**, J.B. & **Turley**, R. (2002), *A Comparison of Parametric Models of Income Distribution Across Countries and Over Time,* Department of Economics, Brigham Young University. Liga de Internet: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=324900

[11] **Bartels**, C.P. & **Metelen,** van (1975), *Alternative Probability Density Function of Income,* Vrije University Amerstadam: Research memorandum 29, 30 pages.

[12] **Boccanfuso**, D., **Decaluwé**, B. & **Savard**, L. (2003), *Poverty, Income Distribution and CGE Modeling: Does the Functional Form of Distribution Matter?,* Centre Interuniversitaire sur le risqué, les politiques économiques et l'emploi, Working Paper 03-32. Liga de Internet: http://ideas.repec.org/p/lvl/lacicr/0332.html

[13] **Bresciani-Turroni,** C. (1937), *On Pareto's Law,* Journal of the Royal Statistical Society, Vol. 100, No. 3, pp. 421-432.

[14] **Burr,** I.W. (1942), *Cumulative Frequency Functions,* Annals of Mathematical Statistics, Vol. 13, No. 2, pp. 215-235.

[15] **Champernowne**, D.G. (1952), *The Graduation of Income Distributions,* Econometrica, Vol. 20, No. 4, pp. 591-615.

[16] **Champernowne**, D. G. (1953), *A model for income distribution,* The Economic Journal, Vol. 63, No. 250, pp. 318-351.

[17] **Champernowne**, D.G. (1974), *A Comparison of Measures of Inequality of Income Distribution,* The Economic Journal, Vol. 84, No. 336, pp. 787-816.

[18] **Cowell**, F.A. & **Victoria-Feser**, M.P. (1996), *Robustness Properties of Inequality Measures,* Econometrica, Vol. 64, No. 1, pp. 77-101.

[19] **Cronin**, D.C. (1979), *A Function for Size Distribution of Incomes: A Further Comment,* Econometrica, Vol. 47, No. 3, pp. 773-774.

[20] **Dagum**, C. (1977), *A new model for personal income distribution: specification and estimation,* Economie Applique'e, 30, pp. 413-437.

[21] **Dagum**, C. (1980), *Inequality Measures between Income Distributions with Applications,* Econometrica, Vol. 48, No. 7, pp. 1791-1803.

[22] **Dagum**, C. (1987), *Measuring the Economic Affluence between Population of Income Recievers,* Journal of Business & Economic Statistics, Vol. 5, No. 1, pp. 5-12.

[23] **Di Guilmi**, C. (2003), *Power Law Scaling in the World Income Distribution,* Economics Bulletin, Vol. 15, No. 6, pp. 1-7.

[24] **Fisk**, P.R. (1961), *The Graduation of Income Distributions,* Econometrica, Vol. 29, No. 2, pp. 171-185.

[25] **Francois**, J.F. & **Kaplan**, S. (1996), *Aggregate Demand Shifts, Income Distribution and the Linder Hypothesis,* The Review of Economics and Statistics, Vol. 78, No. 2, pp. 244-250.

[26] **Gastwirth,** J.L. (1972), *The Estimation of the Lorenz Curve and Gini Index,* The Review of Economics and Statistics, Vol. 54, No. 3, pp. 306-316.

[27] **Gibrat, R**. (1931), *Les Egalites Economiques,* Sirely, Paris.

[28] **Glejberman**, D. (2006), *Distribución de Pareto con dos parámetros,* Liga de Internet: http://www.ccee.edu.uy/ensenian/catest2/pareto.PDF

[29] **Hayakawa**, M. (1951), *The Application of Pareto's Law of Income to Japanese Data,* Econometrica, Vol. 19, No. 2, pp. 174-183.

[30] **Kloek**, T. & **Van Dijk**, H.K. (1980), *Inferential Procedures in Stable Distributions for Class Frequency Data on Incomes,* Econometrica, Vol. 48, No. 5, pp. 1139-1148.

[31] **Maclachlan**, F.C. & **Reith**, J.E. (2002), *The Beaman Distribution: A New Descreptive Model for the Size Distribution of Incomes,* Liga de Internet: http://home.manhattan.edu/~fiona.maclachlan/beaman/

[32] **Majumbder**, A. & **Chakravarty**, S.R. (1990), *Distribution of Personal Income: Development of a new Model and its Applications to U.S. Income Data,* Journal of Applied Econometrics, Vol. 5, No. 2, pp. 189-196.

[33] **Mandelbrot**, B. (1960), *The Pareto-Levy Law and the Distribution of Income,* International Economic Review, Vol. 1, No. 2, pp. 79-106.

[34] **McDonald,** J.B., **Dastrup**, S. & **Hartshorn**, R. (2006), *The Impact of Taxes and Transfer Payments on the Distribution of Income: A Parametric Comparison,* Luxembourg Income Study Working Paper Series, Working Paper No. 401. Liga de Internet: http://www.lisproject.org/publications/liswps/401.pdf

[35] **McDonald,** J.B. & **Jensen**, B.C. (1979), *An Analysis of Some Properties of Alternative Measures of Income Inequality Based on the Gamma Distribution Function,* Journal of American Statistical Association, Vol. 74, No. 368, pp. 856-860.

[36] **McDonald**, J.B., **Ransom,** M.R. (1979), *Functional Forms, Estimation Techniques and the Distribution of Income,* Econometrica, Vol. 47, No. 6, pp. 1513-1526.

[37] **McDonald**, J.B. (1984), *Some Generalized Functions for the Size Distribution of Income,* Econometrica, Vol. 52, No. 3, pp. 647-664.

[38] **McDonald,** J.B. & **Mantrala**, A. (1996), *The Distribution of Personal Income: Revisited,* Journal of Applied Econometrics, Vol. 10, No. 2, pp. 201-204.

[39] **Metcalf,** E.C. (1972), *An Econometric Model of the Income Distribution,* Markham Publishing Co., Chicago.

[40] **Morgan**, J. (1962), *The Anatomy of Income Distribution,* The Review of Economics and Statistics, Vol. 44, No. 3, pp. 270-283.

[41] **Neal**, D. & **Rosen**, S. (1995), *Theories of the Distribution of Labour Earnings,* Working Paper 6378, National Bureau of Economic Research, Massachusetts (1997). Liga de Internet: http://www.nber.org/papers/W6378

[42] **Nirei**, M. & **Souma**, W. (2004), *Two factor model of income distribution dynamics,* Liga de Internet: http://www.comdig.org/index.php?id_issue=2004.48

[43] **Mitnik**, O.A. (1998), *Notas Docentes sobre Distribución del Ingreso y Pobreza,* Liga de Internet: http://economia.uahurtado.cl/pdf/publicaciones/docente-8.pdf

[44] **Ojha**, P.D. & **Bhatt**, V.V. (1964), *Pattern of Income Distribution in an Underveloped Economy: A case of Study of India,* The American Economic Review, Vol. 54, No. 5, pp. 711-720.

[45] **Parker**, S.C. (1999), *The generalized beta as a model for the distribution of earnings,* Economics Letters 62, pp. 197-200.

[46] **Pena Trapero**, J.B, **Callealta Barroso**, F.J. & **Núñez Velázquez**, J.J. (2002), *Encuestas de presupuestos familiares, renta de las familias y estudio de la distribución personal de la renta: una experiencia española,* Universidad de Alcalá. Liga de internet: http://www.uah.es/otrosweb/inves/DocsComun/MemoInves/2001_02/Dpto.%20Est,%20Estr%20y%20OEI%2001-02.pdf

[47] **Pittau**, M.G. & **Zelli**, R. (2005), *Trends in Income Distribution in Italy: A Non Parametric and Semi-Parametric Analysis,* Paper prepared for the 28th General Conference of The International Association for Research in Income and Wealth, Cork, Ireland, August 22-28.

[48] **Reed**, W.J. (2000), *The Pareto, Zipf and other power laws,* Liga de Internet: http://linkage.rockefeller.edu/wli/zipf/reed01_el.pdf

[49] **Rhodes,** E.C. (1944), *The Pareto Distribution of Incomes,* Economica, New Series, Vol. 11, No. 41, pp. 1-11.

[50] **Roy,** A.D. (1950), *The Distribution of Earnings and of Individual Output,* The Economic Journal, Vol. 60, No. 239, pp. 489-505.

[51] **Roy,** A.D. (1951), *Some Thoughts on the Distribution of Earnings,* Oxford Economic Papers, New Series, Vol. 3, No. 2, 135-146.

[52] **Rutherford,** R.S.G. (1955), *Income Distributions: A New Model,* Econometrica, Vol. 23, No. 3, pp. 277-294.

[53] **Sain,** S.R. & **Scout**, D.W. (1996), *On Locally Adaptive Density Estimation,* Journal of the American Statistical Association, 91, No. 436, pp. 1525-1534.

[54] **Salem,** A.B.Z. & **Mount**, T.D. (1974), *A Convenient Descriptive Model of Income Distribution: The Gamma Density,* Econometrica, Vol. 42, No. 6, pp. 1115-1127.

[55] **Singh,** S.K. & **Maddala**, G.S. (1976), *A function for Size distribution of Incomes,* Econometrica, Vol. 44, No. 5, pp. 963-970.

[56] **Stone,** R., **Champernowne**, D.G. & **Meade**, J.E. (1942), *The Precision of National Income Estimates,* The Review of Economic Studies, Vol. 9, No. 2, pp. 111-125.

[57] **Thurow,** L.C. (1970), *Analyzing the American Income Distribution,* The American Economic Review, Vol. 60, No. 2, pp. 261-269. Papers and Proceedings of the Eighty-second Annual Meeting of the Amercian Economic Association.

[58] **Tadikamalla,** P.R. (1980), *A look at the Burr and Related Distributions,* International Statistical Review, Vol. 48, No. 3, pp. 337-344.

[59] **Taille,** C. (1981), *Lorenz Ordering within the Generalized Gamma Family of Income Distributions,* in Statistical Distributions in Scientific Work, Vol. 6, ed. By C. Taille, G.P. Patil y B. Balderssari. Boston: Reidel, 1981, pp. 181-192.

[60] **Thurow,** L.C. (1970), *Analyzing the American Income Distribution,* Papers and Proceedings, American Economics Association 60, 261-269.

[61] **Victoria-Feser**, M.P. (2000), *Robust Income Distribution Estimation with Missing Data,* Distributional Analysis Research Programme, Discussion Paper No. DARP 57, London, UK.

[62] **Victoria-Feser**, M.P. (2000), *Robust Methods for the Analysis of Income Distribution, Inequality and Poverty,* International Statistical Institute, Vol. 68, No. 3, pp. 227-293.

[63] **Wu**, X. & **Perloff**, J.M. (2004), *China's Income Distribution over Time: Reasons for Rising Inequality,* Liga de Internet: http://siteresources.worldbank.org/INTDECABC2006/Resources/Pradeep.PDF

[64] **Wu**, X. & **Perloff**, J.M. (2005), *China's Income Distribution, 1985-2001,* Liga de Internet: http://are.berkeley.edu/~perloff/PDF/china.pdf

[65] **Yeh**, H.C., **Arnold**, B.C. & **Robertson**, C.A. (1988), *Pareto Process,* Journal of Applied Probability, Vol. 25, No. 2, pp. 291-301.

[66] **Zenga**, M. & **Zini**, A. (2000), *A modification of the right tail for heavy-tailed income distributions,* Liga de Internet: http://econpapers.repec.org/article/mtnancoec/2001_3A3_3A02.htm

Capítulo 3

[67] **Bensalah,** Y. (2000), *Steps in Applying Extreme Value Theory to Finance: A Review,* Liga de Internet: http://www.bankofcanada.ca/en/res/wp/2000/wp00-20.pdf

[68] **Castillo,** E., Hadi, A.S., Balakrishnan (2005), N. .S., *Extreme Value and Related Models with Applications in Engineering and Science.* Wiley Series in Probability and Statistics, First Edition.

[69] **Coles**, S. (2004), *An Introduction to Statistical Modeling of Extreme Values,* Springer, London, UK.

[70] **Chávez-Demoulin**, V. & **Roehrl**, A. (2004), *Extreme Value Theory can save your neck,* Liga de Internet: http://www.approximity.com/papers/evt_wp.pdf

[71] **Dowd,** K. (2006), *The Extreme Value Approach to VaR - An Introduction,* Liga de Internet: http://www.fenews.com/fen11/extreme.html

[72] **Gumbel**, E.J. (1935), *Les valeurs extrêmes des distributions statistiques,* Annales de l'I.H.P., Tome 5, No. 2, pp. 115-158.

[73] **Meeker,** W. & **Escobar,** L. (1998), *Statistical Methods for Reliability Data,* Wiley Series in Probability and Statistics.

[74] *Extreme Value Theory,* Engineering Statistics Handbook. Liga de Internet: http://www.itl.nist.gov/div898/handbook/apr/section1/apr163.htm

Capítulo 4

[75] **Becker**, R.A. & **Chambers**, J.M., (1988), *The New S Lenguage,* Chapman & Hall, New York.

[76] **Carmona**, F. (2004), *Curso básico de R,* Liga de Internet: http://cran.r-project.org/

[77] **Chambers**, J.M. & **Hastie**, T.J., (1992), *Statistical Models in S,* Chapman & Hall, New York.

[78] **Díaz-Uriarte**, R. (2003), *Introducción al uso y programación del sistema estadístico R,* Liga de Internet: http://cran.r-project.org/

[79] **Heiberger**, R.M. & **Holland**, B. (2003), *Statistical Analysis And Data Display: an intermediate course with examples in S-PLUS, R, and SAS,* Springer-Verlag.

[80] **Paradis**, E. (2003), *R para principiantes,* Liga de Internet: http://cran.r-project.org/

[81] **Stephenson**, A. (2006), *The evd package,* Liga de Internet: http://cran.r-project.org/

[82] **Venables**, B. & Ihaka, R. (1997), *Notas sobre R: Un entorno de programación para Análisis de Datos y Gráficos,* Liga de Internet: http://cran.r-project.org/

[83] **Venables**, W.N. & **Smith**, D.M. (2004), *An Introduction to R,* Liga de Internet: http://cran.r-project.org/

[84] **Verzani**, J. (2002), *Using R for Introductory Statistics,* Liga de Internet: http://cran.r-project.org/