ESTIMACION DE MODELOS ECONOMETRICOS INHERENTEMENTE LINEALES Y NO LINEALES

MARIO ALBERTO NAJERA HERNANDEZ

T E S I S

PRESENTADA COMO REQUISITO PARCIAL
PARA OBTENER EL GRADO DE
MAESTRO EN CIENCIAS
EN ESTADISTICA EXPERIMENTAL

Universidad Autónoma Agraria
"Antonio Narro"
PROGRAMA DE GRADUADOS
Buenavista, Saltillo, Coah.

MARZO DE 2000

UNIVERSIDAD AUTONOMA AGRARIA
ANTONIO NARRO

SUBDIRECCION DE POSTGRADO

ESTIMACION DE MODELOS ECONOMETRICOS INHERENTEMENTE LINEALES Y NO LINEALES

TESIS

POR

MARIO ALBERTO NAJERA HERNANDEZ

Tesis elaborada bajo la supervisión del Comité Particular de Asesoría y aprobada como requisito parcial, para obtener el grado de:

MAESTRO EN CIENCIAS
EN ESTADISTICA EXPERIMENTAL

COMITE PARTICULAR

Asesor principal: _____
M. C. Félix de Jesús Sánchez Pérez

Asesor: _____
M. C. Emilio Padrón Corral

Asesor: _____
M. C. Roberto Coronado Niño

_____
Dr. Ramiro López Trujillo
Subdirector de Postgrado

Buenavista, Saltillo, Coahuila, Marzo 2000.

DEDICATORIA

A YOLANDA ROSALIA HERNANDEZ LOPEZ

COMPENDIO

Estimación de modelos econométricos inherentemente lineales y no lineales.

POR

MARIO ALBERTO NAJERA HERNANDEZ

MAESTRIA
ESTADISTICA EXPERIMENTAL

UNIVERSIDAD AUTONOMA AGRARIA ANTONIO NARRO
BUENAVISTA, SALTILLO, COAHUILA, MARZO 2000

M. C. Félix de Jésus Sánchez Pérez. - Asesor -

Palabras clave: Modelo lineal general, estimación mínimo cuadrática, estimación máximo verosímil, algorítmo de Gauss-Newton, algorítmo de Newton-Raphson, transformación de Box y Cox.

Se analizan las propiedades de los estimadores de máxima verosimilitud y de mínimos cuadrados cuando el tamaño de muestra es suficientemente grande. Se aborda el problema de estimación no lineal a través de dos métodos iterativos, además se incluye un ejercicio numérico para ilustrar las técnicas y procedimientos estudiados.

ABSTRACT

Estimation of econometrics models inherently lineal and not lineal.

BY

MARIO ALBERTO NAJERA HERNANDEZ

MASTER OF SCIENCE
EXPERIMENTAL STATISTICS

UNIVERSIDAD AUTONOMA AGRARIA ANTONIO NARRO
BUENAVISTA, SALTILLO, COAHUILA, MARZO 2000

M. C. Félix de Jésus Sánchez Pérez. - Advisor -

Key words: Model of the general form, least square estimation, maximum likelihood estimation, algorithm of Gauss Newton, algorithm of Newton Raphson, Box and Cox transformation.

The properties are analyzed of the maximum likelihood estimators and of least squares when the size of sample is sufficiently large. The not lineal problem of estimation through two methods iteratives is undertaken, besides a numerical exercise is included to illustrate the techniques and procedures studied.

INDICE DE CONTENIDO

Página

INDICE DE CUADROS

INDICE DE FIGURAS

INTRODUCCION

Un modelo dentro de la ciencia económica puede ser considerado como la representación de un fenómeno real, el cual se encuentra representado por este para explicarlo, predecirlo y controlarlo. La modelística es una parte integral en la mayoría de las ciencias debido a que los sistemas del mundo, por lo común, son enormemente complejos; es decir, los fenómenos son tan complicados que únicamente pueden ser tratados en términos de una representación simplificada, vía un modelo.

Para propósitos económicos el tipo de modelo más importante es el algebraico, ya que permite representar el sistema del mundo real a través de un conjunto de ecuaciones. El balance adecuado entre maleabilidad y realismo son la esencia de un buen modelo, especificar las interrelaciones entre las partes de un sistema en una forma suficientemente detallada y explícita asegura que el estudio del modelo conduzca a introspecciones respecto al sistema del mundo real, al mismo tiempo las especificaciones en una forma suficientemente simplificada y maleable aseguran que el modelo pueda ser fácilmente analizado y puedan extraerse conclusiones relacionadas con el sistema del mundo real. En la medida que es posible establecer en forma precisa cómo construir un buen modelo, la modelística es en parte arte y en parte ciencia. Seguir ciertos preceptos económicos y estadísticos, conocer intentos previos para modelar un fenómeno, así como la experiencia son reglas útiles.

Un estudio econométrico empieza con un conjunto de proposiciones sobre algún aspecto de la economía. La teoría especifica un conjunto exacto de relaciones determinísticas entre variables. La investigación empírica proporciona estimaciones

de los parámetros desconocidos del modelo, y normalmente, intenta medir la validez de la teoría mediante el comportamiento de los datos observables. De esta forma, el proceso del análisis econométrico parte de la especificación de una relación teórica la cual a veces sólo es capaz de efectuar sugerencias vagamente o ambiguamente, como en el caso de la forma funcional. Esto implica que nos vemos forzados a elegir entre un largo y complicado catálogo de posibilidades. Aunque el modelo lineal es lo suficientemente flexible para permitir una gran variedad de formas en la regresión, excluye muchas formas funcionales útiles. A pesar de que los modelos intrínsecamente no lineales son a menudo difíciles de estimar, con el desarrollo de los paquetes estadísticos de fácil utilización para el usuario, se han vuelto bastante comunes.

El objetivo de este trabajo responde a la necesidad de reunir los temas implícitos en los modelos de regresión no lineales univariados. Al ser este un tema calificado con cierta complejidad debido a que la estimación mínimo cuadrática no lineal es en general una complicada función no lineal de la variable dependiente resulta imposible obtener una expresión explícita para el estimador mínimo cuadrático, se hace preciso resolver el sistema de ecuaciones normales por métodos numéricos, así como también establecer sus propiedades bajo muestras finitas para una amplia gama de especificaciones de modelos no lineales, en particular las propiedades de mejor estimador lineal insesgado del modelo lineal no pueden mantenerse.

Una de las tareas fundamentales del trabajo econométrico es la de aportar un conocimiento descriptivo de algún fenómeno económico. Uno de los principales métodos por los que dicho conocimiento se consigue es mediante el contraste de determinados supuestos alternativos que la teoría económica hace en cada aplicación empírica. De esta manera, los primeros dos capítulos de los seis en que esta organizado este trabajo, pretenden ser un simple recordatorio de estimación, haciendo

referencia a los dos métodos mas populares: el método de mínimos cuadrados y el método de máxima verosimilitud los cuales satisfacen cierto criterio funcional, es decir, en lo que atañe a técnicas numéricas estos se concentran en optimizar una cierta función. También, en estos primeros capítulos, se hará alusión a la estimación restringida y al procedimiento de contrastes de restricciones lineales sobre los parámetros, dentro del marco del modelo lineal general, para después en capítulos subsecuentes hacer únicamente referencia a las técnicas descritas en estos capítulos y posteriormente discutir especificaciones más elaboradas.

El tercer capítulo versa sobre las propiedades de los estimadores de máxima verosimilitud y de mínimos cuadrados cuando el tamaño de muestra es suficientemente grande. Se hará uso de algunas de las herramientas proporcionadas por la teoría asintótica para mostrar las propiedades asintóticas de los estimadores máximo verosímil y mínimo cuadrático. Como usualmente se hacen en el caso lineal procedimientos convencionales de estimación por intervalo y pruebas de hipótesis pueden ser llevadas a cabo bajo una colección de condiciones suficientes que son requeridas por la teoría asintótica; es decir, podrá verse que las expresiones conocidas del modelo lineal general, mostradas en el primer capítulo, permitirán los habituales contrastes de hipótesis mediante los estadísticos t y F en el modelo no lineal.

El capítulo cuarto aborda el problema de la estimación en modelos de regresión no lineales. La disponibilidad de paquetes estadísticos para el tratamiento de datos facilita con una relativa comodidad el trabajo empírico con cierto nivel de sofisticación. No se pretende llegar a grandes conclusiones sobre los distintos métodos iterativos, sino familiarizarnos solamente con dos métodos comúnmente empleados en otras disciplinas. Esto es, la idea que subyace en el procedimiento de mínimos cuadrados no depende de modo alguno de la linealidad del modelo, por lo que es aplicable también en condiciones más generales, la única variación es que la

resolución analítica del problema de estimación es bastante mas compleja que en el modelo lineal. Asimismo, al igual que en el modelo lineal, en el modelo no lineal si el término de error sigue una distribución normal y su varianza es independiente de los componentes del vector de los parámetros estimados, los estimadores de máxima verosimilitud y de mínimos cuadrados si existen coincidirán.

En el capítulo quinto, teniendo en mente la existencia de diferentes tipos de funciones, se hará uso de un mecanismo particular de transformación conocido como la transformación de Box y Cox, como un método de generalizar el modelo lineal. Esto debido a que la teoría económica deja generalmente indeterminada la forma funcional de las relaciones entre variables económicas, lo que sugiere que estas pueden ser en ocasiones intrínsecamente lineales o no lineales. Al hablar de esta dificultad, se mostraran como ejemplos de modelos no lineales en los parámetros a la función de producción de Cobb-Douglas y la función de producción de elasticidad de substitución constante (ESC).

En el sexto y último capítulo se ilustra a través de un ejemplo los procedimientos y técnicas plasmadas principalmente en el cuarto capítulo; es decir, se pretende ilustrar el funcionamiento de las técnicas, más que la evidencia empírica en si misma. Se emplea además el paquete Prostat, como herramienta de cálculo y del cual se presentarán cada uno de los pasos empleados para obtener el resultado final.

CAPITULO 1

CONCEPTOS GENERALES DENTRO DEL MODELO LINEAL GENERAL

Estimación de Máxima Verosimilitud

En el análisis y especificación del modelo estadístico lineal

$$\mathbf{y} = \mathbf{X}\beta + \mathbf{e} \tag{1.1}$$

donde $\mathbf{y}$ es un vector $(T \times 1)$ observado de valores muestrales, $\mathbf{X}$ es una matriz $(T \times K)$ de valores conocidos de las variables explicatorias, β es un vector K-dimensional de coeficientes desconocidos, y $\mathbf{e}$ es un vector inobservable $(T \times 1)$ de variables aleatorias incorrelacionadas e independientes distribuidas con media $\mathbf{0}$ y varianza σ^2, esto es, $\mathbf{e} \sim (\mathbf{0}, \sigma^2\mathbf{I})$, los supuestos estocásticos sobre el vector aleatorio pueden ser cambiados de modo que el vector se distribuya $N(\mathbf{0}, \sigma^2 I)$. Si de verdad existe independencia e idéntica distribución de los errores aleatorios $\mathbf{e}$, estos representan el efecto combinado de un gran número de variables explicatorias que han sido excluidas de la matriz $\mathbf{X}$. Este supuesto adicional cambia el modelo estadístico cambiando la cantidad de información usada en la especificación del modelo estadístico.

Se asume que los errores inobservables $\mathbf{e}$ son un vector de variables aleatorias distribuidas normalmente con vector de medias cero y matriz de covarianzas $\sigma^2\mathbf{I}$. Esto implica que el vector $\mathbf{y}$ es un vector multivariado distribuido normalmente con media $\mathbf{X}\beta$ y matriz de covarianza $\sigma^2\mathbf{I}_t$, esto es $\mathbf{y} \sim \mathbf{N}(\mathbf{X}\beta, \sigma^2\mathbf{I}_t)$. Ya que se sabe que el vector aleatorio observado $\mathbf{y}$ está distribuido normal multivariado con

vector de media $\mathbf{X}\beta$ y covarianza $\sigma^2\mathbf{I}_t$, se puede analíticamente expresar la función de densidad para una particular observación muestral como

$$f(y_t \mid \mathbf{x}_t, \beta, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} exp\Big[- \frac{(y_t - \mathbf{x}'_t\beta)^2}{2\sigma^2} \Big] \tag{1.2}$$

ya que se asume que las observaciones son independientes, se puede expresar la función de densidad conjunta de la muestra como

$$f(y_1, y_2, ..., y_T) = f(y_1)f(y_1), ..., f(y_T) \tag{1.3}$$

$$f(\mathbf{y} \mid \mathbf{X}\beta, \sigma^2) = \frac{1}{\sqrt{(2\pi\sigma^2)}} exp\Big[\frac{-(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \Big] \tag{1.4}$$

y esta puede ser usada para hacer expresiones probabilísticas sobre el vector $\mathbf{y}$. Un problema en esta formulación parametrizada, es que β y σ^2 son desconocidos e inobservables. Una vez que el modelo estadístico ha sido especificado y después que las observaciones muestrales han sido coleccionadas, se deben elegir aquellos valores de los parámetros desconocidos, en este caso β y σ^2, que bajo la especificación normal multivariada, maximizan la probabilidad de obtener la muestra actualmente observada. Este criterio hace explícita la idea de que las $y's$ observadas serán relevantes para conclusiones o evidencias sobre β y σ^2. De esta manera el concepto clave en el principio de verosimilitud es la función de verosimilitud.

A causa de que la probabilidad es una propiedad de la muestra, es costumbre usar la verosimilitud como una propiedad de los parámetros desconocidos. Una vez la muestra es obtenida, se puede expresar la función de densidad conjunta normal, que envuelve los parámetros desconocidos β y σ^2, como la función de verosimilitud.

$$\ell(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) = (2\pi\sigma^2)^{-T/2} - \exp\left[\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2}\right]. \tag{1.5}$$

Esta función, que depende del resultado de las variables aleatorias, proporciona la base para seguir el criterio de seleccionar los valores de β y σ^2 que maximizan la función de verosimilitud. El razonamiento intuitivo para recurrir a la función de verosimilitud es que, dado $\mathbf{y}$, un β para el que la función de verosimilitud es grande es más probable sea el verdadero β que un β para el que la función de verosimilitud es reducida; esto es, $\mathbf{y}$ es más plausible si la función de verosimilitud es grande. En el principio de verosimilitud se asume que toda la información experimental relevante esta contenida en la función de verosimilitud basada en la densidad de $\mathbf{y}$.

El primer paso para encontrar los estimadores de máxima verosimilitud para β y σ^2 es escribiendo la función de verosimilitud en forma logarítmica

$$\ln \ell(\beta, \sigma^2|\mathbf{y}, \mathbf{X}) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln \sigma^2 - \frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \tag{1.6}$$

debido a que es mas conveniente maximizar $\ln \ell$. Ya que el último término de $\ln \ell$ es el único que contiene a β, maximizar $\ln \ell$ con respecto a β es idéntico a maximizar

$$-\frac{(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)}{2\sigma^2} \tag{1.7}$$

con respecto a β. Además, dado el signo negativo y la constante $2\sigma^2$, es claro que maximizar $-(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)/2\sigma^2$ con respecto a β es equivalente a minimizar

$$S = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{1.8}$$

con respecto a β.

Debido a que el estimador de máxima verosimilitud es idéntico al estimador de mínimos cuadrados, esto es,

$$\hat{\beta} = \tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \tag{1.9}$$

las propiedades muestrales son muy similares. En particular, su media y matriz de covarianzas son

$$E[\tilde{\beta}] = \beta \qquad \text{y} \qquad E[(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'] = \sigma^2(\mathbf{X}'\mathbf{X})^{-1} \tag{1.10}$$

donde esta última expresión es el estimador de mínima varianza dentro de la clase de los estimadores linealmente insesgados que son funciones lineales de $\mathbf{y}$. A causa del supuesto adicional de distribución normal, no es necesario considerar solamente funciones lineales de $\mathbf{y}$; es posible mostrar que $\tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ es el estimador de mínima varianza dentro de la clase de todos los estimadores insesgados.

Debido a que $\tilde{\beta}$ es una función lineal de $\mathbf{y}$, otra consecuencia del hecho que $\mathbf{y}$ es un vector aleatorio normal es que $\tilde{\beta}$, también, será un vector normal aleatorio. Así para el estimador de máxima verosimilitud se tiene el siguiente resultado

$$\tilde{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} \sim \mathbf{N}[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}] \tag{1.11}$$

que será usado para probar hipótesis y establecer estimaciones por intervalo.

Para derivar el estimador de máxima verosimilitud de σ^2 es necesario tomar la derivada parcial del logaritmo de la función de verosimilitud con respecto a σ^2. Este procedimiento produce

$$\frac{\partial \ln \ell(\beta, \sigma^2|\mathbf{y}, \mathbf{X})}{\partial \sigma^2} = -\frac{T}{2\sigma^2} + \frac{1}{2\sigma^4}(\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta). \tag{1.12}$$

Para obtener el valor máximo se iguala esta expresión a cero; esto es,

$$-\frac{T}{2\tilde{\sigma}^2} + \frac{1}{2\tilde{\sigma}^4}(\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta}) = 0 \tag{1.13}$$

donde $\tilde{\sigma}^2$ y $\tilde{\beta}$ son los estimadores de máxima verosimilitud para σ^2 y β, respectivamente. Resolviendo para $\tilde{\sigma}^2$ se tiene

$$\tilde{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta})}{T} = \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{T} \tag{1.14}$$

donde $\tilde{\mathbf{e}} = \mathbf{y} - \mathbf{X}\tilde{\beta}$ es idéntico al vector de residuales mínimo cuadrático $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{X}\hat{\beta}$.

Así, la regla de máxima verosimilitud produce un estimador de $\tilde{\sigma}^2$ que es una función cuadrática de $\mathbf{y}$. Como otros estimadores, $\tilde{\sigma}^2$ será una variable aleatoria, variando de muestra a muestra, y sus propiedades muestrales serán, por lo tanto, de interés. Para encontrar la media de $\tilde{\sigma}^2$ es preciso recordar que

$$E[\tilde{\mathbf{e}}'\tilde{\mathbf{e}}] = E[\mathbf{e}'(\mathbf{I} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{e}] = \sigma^2(T - K) \tag{1.15}$$

Así, la media o valor esperado de $\tilde{\sigma}^2$ es

$$E[\tilde{\sigma}^2] = E\left[\frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{T}\right] = \sigma^2\frac{(T - K)}{T}. \tag{1.16}$$

Ya que se requiere $E[\hat{\sigma}^2] = \sigma^2$ para que $\hat{\sigma}^2$ sea insesgado, es claro que el estimador de máxima verosimilitud $\tilde{\sigma}^2$ es sesgado. Sin embargo, el sesgo desaparece cuando T se incrementa y K permanece fijo.

A causa de que $\tilde{\sigma}^2$ es un estimador sesgado, el estimador mas común para σ^2 es el estimador insesgado dado a continuación

$$\hat{\sigma}^2 = \frac{(\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta})}{T - K}. \tag{1.17}$$

Note que $\hat{\sigma}^2 = [T/(T-K)]\tilde{\sigma}^2$, es un simple ajuste requerido para moverse de un estimador a otro.

No sólo la media de $\tilde{\sigma}^2$ y de $\hat{\sigma}^2$ son de interés, también lo es la distribución de probabilidad. La distribución de probabilidad de $\hat{\sigma}^2$ es importante para probar hipótesis y para la estimación por intervalo. Para examinar esta distribución de probabilidad se analiza primero la distribución de probabilidad para

$$\frac{(\mathbf{y} - \mathbf{X}\tilde{\beta})'(\mathbf{y} - \mathbf{X}\tilde{\beta})}{\sigma^2} = \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{\sigma^2} = \frac{\mathbf{e}'(\mathbf{I}_T - X(X'X)^{-1}\mathbf{X}')\mathbf{e}}{\sigma^2} = \frac{\mathbf{e}'\mathbf{M}\mathbf{e}}{\sigma^2} \tag{1.18}$$

donde $\mathbf{M} = \mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'$ es una matriz idempotente. El numerador es una forma cuadrática implicando al vector aleatorio normal $\mathbf{e}$. En particular, si $\mathbf{e} \sim \mathbf{N}(\mathbf{0}, \sigma^2\mathbf{I}_T)$, y $\mathbf{M}$ es idempotente, entonces $\mathbf{e}'\mathbf{M}\mathbf{e}/\sigma^2$ tiene una distribución χ^2 con grados de libertad iguales al rango de $\mathbf{M}$. El rango de una matriz idempotente es igual a su traza, esto es,

$$\begin{aligned} tr(\mathbf{M}) &= tr[\mathbf{I}_T - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'] = tr(\mathbf{I}_T) - tr[\mathbf{X}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= T - tr(\mathbf{I}_K) = T - K \end{aligned} \tag{1.19}$$

Así, el rango de $\mathbf{M}$ es $T - K$. Coleccionando todos estos resultados se tiene

$$\frac{(T-K)\hat{\sigma}}{\sigma^2} = \frac{\tilde{\mathbf{e}}'\tilde{\mathbf{e}}}{\sigma^2} = \frac{\mathbf{e}'\mathbf{M}\mathbf{e}}{\sigma^2} \sim \chi^2_{(T-K)}. \tag{1.20}$$

Cuando se habla de la distribución de probabilidad de $\hat{\sigma}^2$, usualmente se dice $\hat{\sigma}^2 \sim [\sigma^2/(T-K)]\chi^2_{(T-K)}$. Esto es, $\hat{\sigma}^2$ se distribuye como una constante multiplicada por una distribución $\chi^2_{(T-K)}$.

Se puede usar el resultado anterior para encontrar la media y la varianza de $\hat{\sigma}^2$. Se sabe que la media de una variable aleatoria $\chi^2_{(T-K)}$ es igual a sus grados de libertad. Así,

$$E\left[\frac{(T-K)\hat{\sigma}^2}{\sigma^2}\right] = T-K \tag{1.21}$$

y multiplicando ambos lados por $\sigma^2/(T-K)$,

$$E[\hat{\sigma}^2] = \sigma^2. \tag{1.22}$$

Para la varianza, es necesario notar que la varianza de una variable aleatoria χ^2 es igual a dos veces sus grados de libertad. Así,

$$var\left[\frac{(T-K)\hat{\sigma}^2}{\sigma^2}\right] = 2(T-K) \tag{1.23}$$

o

$$\frac{(T-K)^2}{\sigma^4} var(\hat{\sigma}^2) = 2(T-K) \tag{1.24}$$

y

$$var(\hat{\sigma}^2) = \frac{2\sigma^4}{T-K} \tag{1.25}$$

El supuesto de normalidad ha proporcionado información adicional, suficiente para derivar expresiones para la varianza de $\hat{\sigma}^2$.

Otro importante resultado es que el vector aleatorio $\tilde{\beta}$ es independiente de la variable aleatoria $\hat{\sigma}^2$. Ya que $\hat{\sigma}^2 = \tilde{\mathbf{e}}'\tilde{\mathbf{e}}/(T-K)$, será verdadero si $\tilde{\mathbf{e}}$ y $\tilde{\beta}$ son independientes. Ya que $\tilde{\mathbf{e}}$ y $\tilde{\beta}$ son vectores aleatorios normales, para mostrar que ellos son independientes es suficiente mostrar que la matriz que contiene las covarianzas entre los elementos de $\tilde{\mathbf{e}}$ y los elementos de $\tilde{\beta}$ es $\mathbf{0}$. Esta matriz es

$$\begin{aligned} E[\tilde{\mathbf{e}}(\tilde{\beta}-\beta)'] &= E[(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] \\ &= (\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')E[\mathbf{e}\mathbf{e}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1} \\ &= \sigma^2[(\mathbf{I}-\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}')\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}] = \mathbf{0} \end{aligned} \tag{1.26}$$

y así $\tilde{\beta}$ y $\hat{\sigma}^2$ son independientes.

Hasta ahora ha sido posible resumir la información en el vector aleatorio $\mathbf{y}$ de T-ésimo orden en términos de las $K+1$ estadísticas $\tilde{\beta}, \hat{\sigma}^2$. Sería bueno mostrar que $\tilde{\beta}$ y $\hat{\sigma}^2$ contienen toda la información sobre β y σ^2 que la información muestral $\mathbf{y}$ contiene. Si es así, entonces se puede decir que $\tilde{\beta}$ y $\hat{\sigma}^2$ son estadísticas suficientes. Para hacer esto es necesario mostrar que $f(\mathbf{y}|\mathbf{X},\beta,\sigma^2) = g(\tilde{\beta},\hat{\sigma}^2|\beta,\sigma^2)$ donde $g(\cdot)$ contiene las observaciones $\mathbf{y}$ solamente en la forma de $\tilde{\beta}$, $\hat{\sigma}^2$. Ya que $(\mathbf{y}-\mathbf{X}\beta) = (\mathbf{y}-\mathbf{X}\tilde{\beta}+\mathbf{X}\tilde{\beta}-\mathbf{X}\beta)$, esto significa

$$\begin{aligned} (\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta) &= (\mathbf{y}-\mathbf{X}\tilde{\beta}+\mathbf{X}\tilde{\beta}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\tilde{\beta}+\mathbf{X}\tilde{\beta}-\mathbf{X}\beta) \\ &= (\mathbf{y}-\mathbf{X}\tilde{\beta})'(\mathbf{y}-\mathbf{X}\tilde{\beta}) + (\beta-\tilde{\beta})'\mathbf{X}'\mathbf{X}(\beta-\tilde{\beta}) \\ &= (T-K)\hat{\sigma}^2 + (\beta-\hat{\beta})'\mathbf{X}'\mathbf{X}(\beta-\tilde{\beta}). \end{aligned} \tag{1.27}$$

Por lo tanto,

$$f(\mathbf{y}|\mathbf{X},\beta,\sigma^2) = \frac{1}{(2\pi\sigma^2)^{T/2}}\, exp\Big[-(\mathbf{y}-\mathbf{X}\beta)'(\mathbf{y}-\mathbf{X}\beta)\frac{1}{2\sigma^2}\Big]$$

$$= \frac{1}{(2\pi\sigma^2)^{T/2}}\, exp\Big\{-\frac{1}{2\sigma^2}[(T-K)\hat{\sigma}^2+(\beta-\tilde{\beta})'X'X(\beta-\tilde{\beta})]\Big\} = g(\tilde{\beta},\hat{\sigma}^2|\beta,\sigma^2) \quad (1.28)$$

Consecuentemente, se concluye que $\tilde{\beta}$, $\hat{\sigma}^2$ y son una colección de estadísticos suficientes.

Una ventaja de encontrar una colección de estadísticos suficientes es que, bajo condiciones usuales, si los estadísticos suficientes son estimadores insesgados de algunos parámetros de interés, entonces los estimadores insesgados son los mejores estimadores insesgados. En este caso, $\tilde{\beta}$ y $\hat{\sigma}^2$ son estadísticos suficientes de la distribución normal multivariada, y a causa de que ellos son estimadores insesgados de β y σ^2, respectivamente, de ahí sigue que ellos son estimadores insesgados de mínima varianza. Note que estos resultados son mucho más fuertes que los que son dados usando el teorema de Gauss-Markov para establecer que, para el modelo estadístico lineal general sin el supuesto de normalidad, el estimador $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ es el mejor fuera de la clase de los estimadores linealmente insesgados. También indica que el estimador mínimo cuadrático $\hat{\sigma}^2 = (\mathbf{y}-\mathbf{X}\mathbf{b})'(\mathbf{y}-\mathbf{X}\mathbf{b})/(T-K)$ es insesgado.

Otro método para investigar si $\tilde{\beta}$ y σ^2 son los mejores estimadores insesgados es considerando la cota inferior que puede ser derivada usando la desigualdad de Cramer-Rao y la matriz de información resultante. La desigualdad de Cramer-Rao hace uso del hecho de que el cuadrado de una covarianza es en mucho igual al producto de las varianzas correspondientes. Expresando este resultado para el modelo lineal general normal, se representan los $(K+1)$ parámetros desconocidos como $\gamma = (\beta_1,...,\beta_K,\sigma^2)'$ y se recuerda que la función de verosimilitud $\ell(\gamma|\mathbf{y},\mathbf{X})$,

como una función de la muestra aleatoria $\mathbf{y}$, es aleatoria, y por lo tanto las derivadas de la función de verosimilitud con respecto a γ son aleatorias. Si se asume que la función de verosimilitud es dos veces diferenciable, la matriz de información para γ es definida como

$$I(\gamma) = -E\left[\frac{\partial^2 ln\ell(\gamma|\mathbf{y},\mathbf{X})}{\partial\gamma\partial\gamma'}\right] \tag{1.29}$$

que es el negativo de la esperanza de la matriz de derivadas de segundo orden.

La inversa de la matriz de información proporciona la cota inferior para el muestreo preciso de estimadores insesgados de γ. Esto es, $\Sigma_{\hat{\gamma}} \geq I(\gamma)^{-1}$ en el sentido que $\Sigma_{\hat{\gamma}} - I(\gamma)^{-1}$ es positiva semidefinida. Para el modelo estadístico considerado los elementos de la matriz de información son las derivadas de segundo orden de la función de verosimilitud. Llevándose a cabo esta diferenciación, esto produce:

$$\begin{aligned}\mathbf{I}(\gamma) &= -E\begin{pmatrix} -\mathbf{X}'\mathbf{X}/\sigma^2 & -(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta)/\sigma^4 \\ -(\mathbf{X}'\mathbf{y} - \mathbf{X}'\mathbf{X}\beta)'/\sigma^4 & T/2\sigma^4 - (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta)/\sigma^6 \end{pmatrix} \\ &= \begin{pmatrix} -\mathbf{X}'\mathbf{X}/\sigma^2 & 0 \\ 0' & T/2\sigma^4 \end{pmatrix} \end{aligned} \tag{1.30}$$

y

$$\mathbf{I}(\gamma)^{-1} = \begin{pmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & 2\sigma^4/T \end{pmatrix}. \tag{1.31}$$

La covarianza para los estimadores insesgados $\tilde{\beta}$ y $\hat{\sigma}^2$ es

$$\Sigma_{(\hat{\beta},\hat{\sigma}^2)} = \begin{pmatrix} \sigma^2(\mathbf{X}'\mathbf{X})^{-1} & 0 \\ 0 & 2\sigma^4/(T-K) \end{pmatrix} \tag{1.32}$$

esto significa que la cota inferior de Cramer-Rao es alcanzada por la covarianza de $\tilde{\beta} = b$ pero no por el estimador insesgado $\hat{\sigma}^2$. Como se mencionó cuando se discutió la suficiencia, esto ofrece un resultado mas fuerte que el teorema de Gauss-Markov porque esto significa que el estimador de máxima verosimilitud de β es el mejor en una gran clase de estimadores que no incluyan la restricción lineal. También, aunque la varianza de $\hat{\sigma}^2$ no alcanza la cota inferior, puede ser mostrado por otros métodos que un estimador insesgado de σ^2 con varianza mas pequeña que $2\sigma^4/(T-K)$ no existe. Este resultado implica que $\hat{\sigma}^2$ es el mejor estimador insesgado.

Estimación Restringida de Máxima Verosimilitud

En muchos estudios econométricos, información *A priori* o de tipo no muestral puede existir y puede estar disponible en una variedad de formas. En este punto, se asume que se tiene información exacta relativa a un parámetro particular o a una combinación lineal de parámetros. Por ejemplo, en la estimación de una función de producción $y_t = \beta_1 + \beta_2 x_{t2} + \beta_3 x_{t3}$, donde x_{t2} es el logaritmo del insumo capital y x_{t3} es el logaritmo del insumo trabajo, la información que puede estar disponible es que la empresa está trabajando bajo cualquiera de tres situaciones posibles, según haya rendimientos a escala crecientes, constantes, o decrecientes. Estas tres situaciones vienen determinadas por la condición de que la suma de las tres elasticidades sea mayor, igual o menor que uno. Alternativamente, estimando una relación de demanda la información que puede estar disponible de la teoría del consumidor es la condición de homogeneidad, o el coeficiente del ingreso respuesta puede estar disponible de un trabajo empírico previo.

En cualquier evento, si información de este tipo está disponible, ésta puede ser expresada en la forma de las siguientes relaciones lineales o restricciones lineales de igualdad

$$\mathbf{R}\beta = \mathbf{r} \tag{1.33}$$

donde $\mathbf{r}$ es un vector $(J \times 1)$ de elementos desconocidos y $\mathbf{R}$ es una matriz $(J \times K)$ *A priori* de rango $J \leq K$ que expresa la estructura de la información en el parámetro individual β_i o alguna combinación lineal de los elementos del vector β. La información concerniente a los parámetros, tal como β_1 igual a un escalar k, la suma de los coeficientes igual a la unidad, y β_2 igual a β_3, puede ser especificada en el formato $\mathbf{R}\beta = \mathbf{r}$ como

$$\begin{pmatrix} 1 & 0 & 0 & 0 & ... & 0 \\ 1 & 1 & 1 & 1 & ... & 1 \\ 0 & 1 & -1 & 0 & ... & 0 \end{pmatrix} \begin{pmatrix} \beta_1 \\ \beta_2 \\ . \\ . \\ . \\ \beta_k \end{pmatrix} = \begin{pmatrix} k \\ 1 \\ 0 \end{pmatrix} \tag{1.34}$$

donde $J = 3$. Si los primeros elementos del vector de coeficientes J fueran especificados igual a un vector particular $\mathbf{r}$ de dimensión J, esta información podría ser especificada como

$$(\mathbf{I}_J \quad \mathbf{0}_{K-J}) \begin{pmatrix} \beta_j \\ \beta_{K-J} \end{pmatrix} = \mathbf{r} \tag{1.35}$$

donde $\mathbf{I}_J$ es una matriz identidad de $J -$ ésimo orden y $\mathbf{r}$ es un vector $(J \times 1)$ desconocido.

Dada la información contenida en la forma $\mathbf{R}\beta = \mathbf{r}$, la pregunta es como combinar ésta con la información contenida en las observaciones muestrales $\mathbf{y}$. A causa de que la información contenida en los parámetros individuales y las com-

binaciones están especificadas como conocidas con certeza, no existe variabilidad en el muestreo de muestra a muestra, y las relaciones de igualdad $\mathbf{R}\beta = \mathbf{r}$ pueden ser tomadas como dadas o restricciones en cualquier muestreo o proceso de estimación. En general, el modelo estadístico lineal tiene como estimadores al criterio de mínimos cuadrados o al principio de máxima verosimilitud $\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$ cuya media es el vector β y covarianza $\mathbf{\Sigma}_b = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}$. Ya que la matriz $\mathbf{X}'\mathbf{X}$ no se asume como diagonal, restricciones en coeficientes particulares o sus combinaciones lineales reflejan por la condición $\mathbf{R}\beta = \mathbf{r}$ los valores que otros coeficientes estimados pueden asumir o adoptar. Si se asume cualquier criterio, máximo verosímil o mínimo cuadrático, aplicado a la información muestral $\mathbf{y}$ y a la información no muestral $\mathbf{R}\beta = \mathbf{r}$, se esta de hecho con el problema de encontrar el vector $\mathbf{b}^*$ que minimiza la forma cuadrática

$$\mathbf{S} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) \tag{1.36}$$

sujeta a

$$\mathbf{R}\beta = \mathbf{r} \qquad \text{o} \qquad \mathbf{R}\beta - \mathbf{r} = \mathbf{0} \tag{1.37}$$

Ya que la información no muestral aparece como una restricción de igualdad lineal, procedimientos clásicos Lagrangianos pueden ser aplicados para producir la función Lagrangiana

$$\mathcal{L} = \mathbf{e}'\mathbf{e} + 2(\mathbf{r}' - \beta'\mathbf{R}')\lambda \tag{1.38}$$

o

$$\mathcal{L} = (\mathbf{y} - \mathbf{X}\beta)'(\mathbf{y} - \mathbf{X}\beta) + 2(\mathbf{r}' - \beta'\mathbf{R}')\lambda$$

$$= \mathbf{y}'\mathbf{y} - 2\beta'\mathbf{X}'\mathbf{y} + \beta'\mathbf{X}'\mathbf{X}\beta + 2(\mathbf{r}' - \beta'\mathbf{R}')\lambda \tag{1.39}$$

donde λ es el vector $(J \times 1)$ de multiplicadores de Lagrange. El dos enfrente del último término aparece para hacer las cosas mas fáciles mas tarde y este no afecta el resultado, ya que $\mathbf{r}' - \beta'\mathbf{R}' = \mathbf{0}$ es bajo el supuesto. Para determinar el valor óptimo, se colocan las derivadas parciales de $\mathcal{L}$ con respeto a β y a λ igual a cero para encontrar el punto estacionario de la función Lagrangiana.

$$(i) \qquad \frac{\partial \mathcal{L}}{\partial \beta} = -2\mathbf{X}'\mathbf{y} + 2\mathbf{X}'\mathbf{X}\mathbf{b}^* - 2\mathbf{R}'\lambda^* = \mathbf{0}$$

$$(ii) \qquad \frac{\partial \mathcal{L}}{\partial \lambda} = 2(\mathbf{r} - \mathbf{R}\mathbf{b}^*) = \mathbf{0}$$

de (i) se obtiene

$$\mathbf{X}'\mathbf{X}\mathbf{b}^* = \mathbf{X}'\mathbf{y} + \mathbf{R}'\lambda^* \tag{1.40}$$

de aquí

$$(\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{X})\mathbf{b}^* = (\mathbf{X}'\mathbf{X})^{-1}(\mathbf{X}'\mathbf{y} + \mathbf{R}'\lambda^*) \tag{1.41}$$

y así

$$\mathbf{b}^* = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\lambda^* \tag{1.42}$$

donde $\mathbf{b}$ es el estimador mínimo cuadrático irrestringido. Esta última ecuación multiplicada por $\mathbf{R}$ produce,

$$\mathbf{R}\mathbf{b}^* = \mathbf{R}\mathbf{b} + \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'\lambda^* = \mathbf{r}. \tag{1.43}$$

Ya que $(\mathbf{X}'\mathbf{X})^{-1}$ es positiva definida, $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ es una matriz positiva definida con rango J, que es menor o igual a K, el rango de $(\mathbf{X}'\mathbf{X})^{-1}$. Ya que $\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$ es no singular, se puede expresar $\mathbf{R}\mathbf{b}^*$ como

$$\lambda^* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\mathbf{b}^* - \mathbf{R}\mathbf{b}) \quad \text{o} \quad \lambda^* = [\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}) \tag{1.44}$$

a causa de las derivadas de la función Lagrangiana el problema de minimización restringido debe satisfacer la condición $\mathbf{R}\mathbf{b}^* = \mathbf{r}$. Usando este valor para el vector λ^*, se obtiene, $\mathbf{b}^*$, el estimador

$$\mathbf{b}^* = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}). \tag{1.45}$$

Esta regla usada para ambos datos muestrales y no muestrales, es llamada el estimador mínimo cuadrático restringido o estimador máximo verosímil restringido y difiere del estimador mínimo cuadrático irrestringido $\mathbf{b}$ por una función lineal del vector $(\mathbf{r} - \mathbf{R}\mathbf{b})$.

Ahora bien, ya que $\mathbf{b}$ es un vector aleatorio, la regla implicada por $\mathbf{b}^*$ significa que este es también un vector aleatorio. El vector aleatorio mínimo cuadrático restringido tiene media

$$\begin{aligned}
\mathbf{E}[\mathbf{b}^*] &= \mathbf{E}\{\mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b})\} \\
&= \mathbf{E}[\mathbf{b}] + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{E}[\mathbf{b}]) \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\beta)
\end{aligned}$$

$$= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\delta$$

$$= \beta \tag{1.46}$$

a causa de que la condición del lado derecho $\delta = \mathbf{r} - \mathbf{R}\beta = \mathbf{0}$, es verdadera bajo el supuesto. Por lo tanto, $\mathbf{b}^*$ es insesgado si $\mathbf{r} - \mathbf{R}\beta = \mathbf{0} = \delta$. Esto es, $\mathbf{b}^*$ es insesgado si la restricción es correcta. Antes de calcular la matriz de varianzas y covarianzas de $\mathbf{b}^*$ es necesario recordar que

$$\mathbf{b} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'(\mathbf{X}\beta + \mathbf{e}) = \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}. \tag{1.47}$$

Corresponde esto, a lo obtenido de $\mathbf{b}^*$, así

$$\mathbf{b}^* - \beta = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\beta - \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e})$$

$$= (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \tag{1.48}$$

que puede ser escrita como

$$\mathbf{b}^* - \beta = \mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} \tag{1.49}$$

ya que $\mathbf{r} - \mathbf{R}\beta = \mathbf{0}$ y donde $\mathbf{M}^* = \mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}$.

Es posible expresar la matriz de varianzas y covarianzas para $\mathbf{b}^*$ como

$$\mathbf{var}(\mathbf{b}^*) = \mathbf{\Sigma}_{\mathbf{b}^*} = \mathbf{E}[(\mathbf{b}^* - \mathbf{E}[\mathbf{b}^*])(\mathbf{b}^* - \mathbf{E}[\mathbf{b}^*])'] = \mathbf{E}[(\mathbf{b}^* - \beta)(\mathbf{b}^* - \beta)']$$

$$= \mathbf{E}[\mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}\mathbf{e}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}^{*'}] = \mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{E}[\mathbf{e}\mathbf{e}']\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}^{*'}$$

$$= \sigma^2\mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}^{*'} \tag{1.50}$$

ya que $\mathbf{E}[\mathbf{e}\mathbf{e}'] = \sigma^2\mathbf{I}$. Además,

$$\mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}^{*'} = [\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}](\mathbf{X}'\mathbf{X})^{-1}$$

$$\times[\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}]'$$

$$= (\mathbf{X}'\mathbf{X})^{-1} - 2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

$$+(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'$$

$$\times[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= (\mathbf{X}'\mathbf{X})^{-1} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \{\mathbf{I} - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}\}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1} \tag{1.51}$$

donde el término entre corchetes es la matriz idempotente $\mathbf{M}^*$, se puede escribir la matriz de varianzas y covarianzas Σ_{b^*}, como

$$\Sigma_{b^*} = \sigma^2\mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}$$

$$= \Sigma_b - \mathbf{C} \tag{1.52}$$

Así

$$= \Sigma_b - \Sigma_{b^*} = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1} = \mathbf{C} \tag{1.53}$$

donde $\mathbf{C}$ es una matriz positiva semidefinida. Consecuentemente la matriz de varianzas y covarianzas para el estimador mínimo cuadrático restringido Σ_{b^*} tiene elementos en la diagonal que son menores o iguales que los correspondientes elementos de estimador mínimo cuadrático Σ_b. El estimador $\mathbf{b}^*$ es el mejor estimador en la clase de los estimadores lineales que son insesgados cuando los verdaderos valores de los parámetros satisfacen la restricción. Si se asume que el vector de información muestral $\mathbf{y}$ es normal multivariado, entonces $\mathbf{b}^* \sim \mathbf{N}(\beta, \sigma^2\mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{M}^{*'})$

Sin embargo, en el trabajo aplicado si no se está seguro que la información no muestral es correcta, si las restricciones son inconsistentes con los parámetros del modelo muestral que genera los datos, esto es, si las restricciones son incorrectas y $\mathbf{r} - \mathbf{R}\beta = \delta \neq \mathbf{0}$, entonces el estimador mínimo cuadrático restringido tiene media

$$\begin{aligned}
\mathbf{E}[\mathbf{b}^*] &= \mathbf{E}[\mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b})] \\
&= \mathbf{E}[\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1} \\
&\quad \times[\mathbf{r} - \mathbf{R}\beta - \mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e}] \\
&= \beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\delta
\end{aligned} \tag{1.54}$$

que significa que el estimador mínimo cuadrático b es sesgado. Sin embargo, su matriz de covarianza es en el caso de $\mathbf{X}$ fijo

$$\begin{aligned}
\mathbf{\Sigma}_{\mathbf{b}^*} &= \mathbf{E}[(\mathbf{b}^* - \mathbf{E}[\mathbf{b}^*])(\mathbf{b}^* - \mathbf{E}[\mathbf{b}^*])'] \\
&= \mathbf{E}[\{\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\delta \\
&+\mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} - \beta - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\delta\}]
\end{aligned}$$

$$\times[\{\beta + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\delta$$

$$+\mathbf{M}^*(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{e} - \beta - (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\delta\}']$$

$$= \sigma^2(\mathbf{X}'\mathbf{X})^{-1} - \sigma^2(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}] \qquad (1.55)$$

que es igual a la matriz de covarianzas del estimador insesgado restringido mínimo cuadrático por lo tanto si las restricciones son correctas o incorrectas o el estimador restringido es sesgado o insesgado, el estimador mínimo cuadrático restringido tiene una matriz precisión que es superior al estimador mínimo cuadrático que usa solamente información muestral. Así el estimador mínimo cuadrático restringido $\mathbf{b}^*$ tiene una excelente precisión relativa a su contraparte mínimo cuadrático o máximo verosímil $\mathbf{b}$ y $\tilde{\beta}$, pero esto puede posiblemente resultar en una regla sesgada.

CAPITULO 2

INFERENCIA DENTRO DEL MODELO LINEAL GENERAL

Estimación por Intervalo

Proporcionar un estimador puntual de un parámetro desconocido no es una información de excesiva utilidad a no ser que se acompañe de un intervalo de confianza para dicha estimación. Para un nivel de significancia dado, es importante conocer el rango de valores admisibles del parámetro que se estima, y no tan solo que valor consideramos como más probable. Ya que el estimador de máxima verosimilitud $\tilde{\beta}$ está distribuido como un vector aleatorio normal con media β y covarianza $\sigma^2(\mathbf{X}'\mathbf{X})^{-1}$, el problema es ahora obtener regiones confidenciales individuales y conjuntas cuando los elementos del vector aleatorio $\tilde{\beta}$ no son independientes.

Dado que $\beta \sim N[\beta, \sigma^2(\mathbf{X}'\mathbf{X})^{-1}]$, se sabe que bajo una transformación lineal general $\mathbf{R}\tilde{\beta} \sim N[\mathbf{R}\beta, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']$ donde $\mathbf{R}$ es una matriz conocida $(J \times K)$. Si $\mathbf{R}\tilde{\beta}_1$ representa una combinación lineal individual de $\tilde{\beta}$, donde $\mathbf{R}_1$ es un vector $(1 \times K)$ de valores conocidos. Consecuentemente,

$$\mathbf{R}_1\tilde{\beta} \sim N[\mathbf{R}_1\beta, \sigma^2\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'] \tag{2.1}$$

o, alternativamente

$$\mathbf{R}_1\tilde{\beta} - \mathbf{R}_1\beta \sim N[0, \sigma^2\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'] \tag{2.2}$$

Ya que $\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'$ es un escalar, se sabe que

$$\frac{\mathbf{R}_1(\tilde{\beta}-\beta)}{\sigma\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}} = z \tag{2.3}$$

está distribuida como una variable normal estándar, con media cero y varianza unitaria. Como un resultado, si se conoce σ^2, y así σ, se puede escoger un intervalo $[-z_{(\alpha/2)}, z_{(\alpha/2)}]$ y, ya que la variable aleatoria normal estándar está tabulada, se pueden hacer expresiones probabilísticas sobre el intervalo que contiene el verdadero parámetro. En otras palabras, si se representa la variable aleatoria normal estándar por z y su función de densidad por f(z), entonces

$$\int_{z_{\alpha/2}}^{z_{\alpha/2}} f(z)dz = 1-\alpha \tag{2.4}$$

donde $-z_{(\alpha/2)}$ y $z_{(\alpha/2)}$ son los valores críticos asociados con una cierta probabilidad o nivel de significancia estadística α para la variable aleatoria normal estándar. Esto es, para los valores críticos de la variable aleatoria normal estándar $z_{\alpha/2}$,

$$Pr[-z_{\alpha/2} \leq z \leq z_{\alpha/2}] = 1-\alpha. \tag{2.5}$$

De aquí que se pueda escribir el intervalo estimado para $\mathbf{R}_1\beta$ como

$$Pr\left[-z_{\alpha/2} \leq \frac{\mathbf{R}_1(\tilde{\beta}-\beta)}{\sigma\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}} \leq z_{\alpha/2}\right] = 1-\alpha. \tag{2.6}$$

Si colocara el nivel de significancia en α=.05, se debería alcanzar, en muestras repetidas, el resultado que, en promedio, el intervalo $[-z_{(\alpha/2)}, z_{(\alpha/2)}]$ pueda contener el estadístico $\mathbf{R}_1(\tilde{\beta}-\beta)/\sigma\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}$, el 95% de las veces. Esto es, si se toman muestras repetidas de los datos y se estima $\mathbf{R}_1\beta$ por $\mathbf{R}\tilde{\beta}$, en promedio 95 de

100 de tales intervalos contendrán $\mathbf{R}_1(\tilde{\beta} - \beta)/\sigma\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}$. Si σ es conocida, entonces existen bases para desarrollar un intervalo estimado para $\mathbf{R}_1\beta$ que puede ser reescrito como

$$Pr[\mathbf{R}_1\tilde{\beta} - z_{\alpha/2}\sigma\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'} \leq \mathbf{R}_1\beta \leq \mathbf{R}_1\tilde{\beta} + \sigma z_{\alpha/2}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}] = 1 - \alpha. \tag{2.7}$$

Un problema en usar este último intervalo es que, en muchos experimentos o situaciones aplicadas, σ^2, y así σ, es desconocido. Si se usa un estimador insesgado de σ^2 se tendrá que considerar la distribución de la variable

$$\frac{\mathbf{R}_1(\tilde{\beta} - \beta)}{\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}}. \tag{2.8}$$

Afortunadamente, esta distribución fue considerada hace tiempo por Gossett considerando el problema de la inferencia que surge cuando se trabaja con la razón de una desviación normal estándar a la raíz de una variable independiente χ^2 dividida por sus grados de libertad. Si $\tilde{\beta}$ y σ^2 son independientes, un ejemplo de tal razón es una variable aleatoria distribuida como una variable aleatoria t de student con $(T - K)$ grados de libertad.

$$\frac{\mathbf{R}_1(\tilde{\beta} - \beta)}{\sigma\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}} \div \left[\frac{(T-K)\hat{\sigma}^2}{\sigma^2(T-K)}\right]^{1/2} = \frac{\mathbf{R}_1(\tilde{\beta} - \beta)}{\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}} = t_{(T-K)}. \tag{2.9}$$

Esta variable aleatoria ha sido tabulada para varios grados de libertad y niveles de significancia (α). Por lo tanto, la distribución t de Student puede ser usada para reexpresar un intervalo aleatorio como

$$Pr[-t_{(T-K,\alpha/2)} \leq t_{(T-K)} \leq t_{(T-K,\alpha/2)}]. \tag{2.10}$$

En otras palabras,

$$Pr\left[- t_{(T-K,\alpha/2)} \leq \frac{\mathbf{R}_1(\tilde{\beta} - \beta)}{\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}} \leq t_{(T-K,\alpha/2)}\right] = 1 - \alpha \tag{2.11}$$

o

$$\begin{aligned} Pr[\mathbf{R}_1\tilde{\beta} - t_{(T-K,\alpha/2)}\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'} \leq \mathbf{R}_1\beta \leq \mathbf{R}_1\tilde{\beta} \\ + \hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}\, t_{(T-K,\alpha/2)}] = 1 - \alpha \end{aligned} \tag{2.12}$$

que significa que si se computan intervalos estimados para $\mathbf{R}_1\beta$ sobre muestras repetidas, $(1-\alpha)$ 100 por ciento de los intervalos estimados contendrán la verdadera combinación lineal $\mathbf{R}_1\beta$.

Si se piensa en $\mathbf{R}_1$ como un vector renglón con todos sus elementos cero excepto uno de valor unitario, esto es, $\mathbf{R} = [0\,0...1...0]$, entonces se puede definir $t_{(T-K)}$ para un coeficiente individual y usar el resultado para especificar el intervalo estimado para un coeficiente particular. Por lo tanto, $t_{(T-K)}$ llega a ser para el coeficiente K − ésimo

$$\frac{\tilde{\beta}_K - \beta_K}{\hat{\sigma}\sqrt{a^{KK}}} = t_{(T-K)} \tag{2.13}$$

donde a^{KK} es el K-ésimo elemento de la matriz $(\mathbf{X}'\mathbf{X})^{-1}$. Consecuentemente, para un particular coeficiente el intervalo estimado es

$$Pr[\tilde{\beta}_K - t_{(T-K,\alpha/2)}\hat{\sigma}\sqrt{a^{KK}} \leq \beta_K \leq \tilde{\beta}_K + t_{(T-K,\alpha/2)}\hat{\sigma}\sqrt{a^{KK}}] = 1 - \alpha. \qquad (2.14)$$

A causa de que los puntos finales de la desigualdad son funciones de variables aleatorias observables, los intervalos son aleatorios. La expresión probabilística enfoca la proporción de veces, $(1 - \alpha)$, que la desigualdades son satisfechas.

Dados los intervalos estimados para un parámetro individual o una combinación de ellos, ahora se tratará la cuestión de estimaciones por intervalos conjuntos o simultáneos. Esta situación ocurre en el trabajo aplicado si un intervalo confidencial conjunto para el coeficiente de ingreso y precio en una relación de demanda o coeficientes de trabajo y capital en una función de producción fueran deseados. Se considerará el problema de usar estimaciones máximo verosímiles para obtener intervalos confidenciales multilples para el caso de dos combinaciones lineales de parámetros.

Suponiendo que existe interés en dos combinaciones lineales $\mathbf{R}_1\beta$ y $\mathbf{R}_2\beta$ donde $\mathbf{R}_1$ y $\mathbf{R}_2$ son vectores $(1 \times K)$. Estas dos combinaciones lineales pueden ser escritas en notación matricial como $\mathbf{R}\beta$, donde

$$\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{pmatrix} \qquad (2.15)$$

es una matriz $(2 \times K)$. A partir de los resultados de la distribución normal multivariada se puede decir que

$$\mathbf{R}\tilde{\beta} \sim N[\mathbf{R}\beta, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']. \qquad (2.16)$$

Esto es, $\mathbf{R}\tilde{\beta}$ es un vector normal bidimensional con media $\mathbf{R}\beta$ y matriz de covarianza

$$\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'.$$

Además,

$$(\mathbf{R}\tilde{\beta} - \mathbf{R}\beta)'[\sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\tilde{\beta} - \mathbf{R}\beta)$$

$$= \frac{(\tilde{\beta} - \beta)'\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\tilde{\beta} - \beta)}{\sigma^2} \sim \chi^2_{(2)} \tag{2.17}$$

Este resultado puede ser usado para obtener una región confidencial conjunta para las dos combinaciones lineales contenidas en $\mathbf{R}\beta$, si σ^2 es conocido. Dado que σ^2 es desconocido se usa el resultado de la razón de dos variables aleatorias independientes χ^2 para formar una variable aleatoria F que no depende de σ^2. Se sabe que $(T - K)\hat{\sigma}^2/\sigma^2$ está distribuida como una variable aleatoria $\chi^2_{(T-K)}$.

Ya que una variable aleatoria F esta definida como la razón de dos variables aleatorias χ^2 dividida por sus grados de libertad se puede formar la siguiente variable aleatoria

$$\lambda = [(\tilde{\beta} - \beta)'\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\tilde{\beta} - \beta)/2\sigma^2] \div \frac{(T - K)\hat{\sigma}^2}{\sigma^2}/(T - K)$$

$$= \frac{(\tilde{\beta} - \beta)'\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\tilde{\beta} - \beta)}{2\hat{\sigma}^2} \tag{2.18}$$

que sigue una distribución $F_{(2,T-K)}$. Ya que λ no contiene a σ^2, esta puede ser usada para construir una región confidencial de $(1 - \alpha)$ por ciento para las dos combinaciones lineales β contenidos en $\mathbf{R}\beta$. Trabajando en esta dirección se tiene

$$Pr[\lambda \leq F_{(2,T-K,\alpha)}] = 1 - \alpha \tag{2.19}$$

o

$$Pr\Big\{\frac{1}{2\hat{\sigma}^2}[(\tilde{\beta}-\beta)'\mathbf{R}_1'\mathbf{R}_1(\tilde{\beta}-\beta)a_1^{11} + 2(\tilde{\beta}-\beta)'\mathbf{R}_1'\mathbf{R}_2(\tilde{\beta}-\beta)a_1^{12} + (\tilde{\beta}-\beta)'\mathbf{R}_2'\mathbf{R}_2(\tilde{\beta}-\beta)a_1^{22}] \leq F_{(2,T-K)}\Big\} = 1 - \alpha \tag{2.20}$$

donde a^{ij} son los elementos de $[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}$. (Esta última expresión probabilística denota un intervalo confidencial conjunto en la forma de una elipse con $\mathbf{R}\tilde{\beta}$ como su centro).

Si $\mathbf{R}_1$ y $\mathbf{R}_2$ son vectores unitarios $\mathbf{R}_1$=(1 0 ... 0) y $\mathbf{R}_2$=(0 1 ...0) y se están considerando intervalos confidenciales conjuntos para dos coeficientes individuales, por ejemplo, β_1 y β_2 entonces la siguiente expresión

$$Pr\Big\{\frac{1}{2\hat{\sigma}^2}[(\tilde{\beta}_1-\beta_1)^2a_1^{11} + 2(\tilde{\beta}_1-\beta_1)(\tilde{\beta}_2-\beta_2)a_1^{12} + (\tilde{\beta}_2-\beta_2)^2a_1^{22}] \leq F_{2,T-K}\Big\} = 1 - \alpha \tag{2.21}$$

que es una elipse con β_1 y β_2 como su centro.

Habiendo desarrollado estimaciones individuales y conjuntas para los parámetros contenidos en el vector β, es necesario tratar ahora con σ^2, otro parámetro desconocido . Se sabe que la variable aleatoria $(T-K)\hat{\sigma}^2/\sigma^2$ esta distribuida como una variable aleatoria χ^2 con $(T-K)$ grados de libertad. Por lo tanto, dados los más bajos y más altos valores críticos $\chi^2_{(\cdot)}$, c_1=$\chi^2_{(T-K,\alpha/2)}$ y c_2=$\chi^2_{((T-K,1-\alpha/2))}$, donde α es el nivel de significancia y

$$\int_{c_1}^{c_2} f(\chi^2_{(\cdot)})d\chi^2_{(\cdot)} = 1 - \alpha \tag{2.22}$$

donde $f(\chi^2)$ es la función de densidad, se puede hacer la expresión

$$Pr[\chi^2_{(T-K,\alpha/2)} \leq \chi_{(T-K)} \leq \chi^2_{(T-K,1-\alpha/2)}] = 1 - \alpha \tag{2.23}$$

o

$$Pr\left[c_1 \leq \frac{(T-K)\hat{\sigma}^2}{\sigma^2} \leq c_2\right] = 1 - \alpha. \tag{2.24}$$

Esto implica que

$$Pr\left[\frac{(T-K)\hat{\sigma}^2}{\chi^2_{(T-K,1-\alpha/2)}} \leq \sigma^2 \leq \frac{(T-K)\hat{\sigma}^2}{\chi^2_{(T-K,\alpha/2)}}\right] = 1 - \alpha \tag{2.25}$$

ya que los puntos finales dentro de los corchetes son variables aleatorias, éstas variarán de muestra a muestra.

Hasta aquí el armazón ya ha sido desarrollado para usar $x_0'\tilde{\beta}$ como una base para predecir el resultado promedio para $x_0'\beta$, de un nivel particular de las variables tratadas x_0'. Desafortunadamente, desde un punto de vista de la predicción o para propósitos de decisión es deseado el resultado promedio x_0, más que el resultado para y_0 de un individual o particular x_0. Si dentro del contexto del modelo estadístico lineal se utiliza este resultado para y_0 como una variable aleatoria, entonces puede ser útil determinar por adelantado un intervalo para el resultado aleatorio y_0.

Ya que la variable aleatoria $y_0 \sim N(x_0'\beta, \sigma^2)$, si se tiene la buena fortuna de conocer β y σ^2, podría ser posible desarrollar un intervalo que incluyera a y_0 con probabilidad $(1 - \alpha)$. Sin embargo, ya que β y σ^2 son desconocidos, se debe considerar cómo construir un intervalo. Si se reemplazan los parámetros desconocidos por estimaciones, entonces, en contraste a la estimación por intervalo para los

parámetros desconocidos tales como β_k o σ^2, este problema es considerado usando el predictor $\hat{y} = x_0'\tilde{\beta}$ para construir un intervalo de predicción para la variable aleatoria $y = x_0'\beta + e_0$. Tal como un intervalo predictor con probabilidad $(1-\alpha)$ en el sentido que, del muestreo repetido de la distribución conjunta para $(y'\, y_0)'$, $(1-\alpha)$ por ciento de los intervalos confidenciales contendrán los $y_0's$ realizados .

En el desarrollo del intervalo de predicción se está concentrado en el error de predicción

$$\hat{y}_0 - y_0 = x_0'\tilde{\beta} - x_0'\beta - e_0 = x_0'(\tilde{\beta} - \beta) - e_0 \tag{2.26}$$

que implica la ecuación de error e_0 y el error de estimación $\tilde{\beta} - \beta$. Esta variable aleatoria normal tiene media

$$E[x_0'(\tilde{\beta} - \beta) - e_0] = 0 \tag{2.27}$$

y varianza

$$\begin{aligned}\{E[x_0'(\tilde{\beta} - \beta) - e_0]\}^2 &= E[x_0'(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'x_0 - 2x_0'(\tilde{\beta} - \beta)e_0 + e_0^2] \\ &= E[x_0'(\tilde{\beta} - \beta)(\tilde{\beta} - \beta)'x_0] - 2E[x_0'(\tilde{\beta} - \beta)e_0] + E[e_0^2] \\ &= \sigma^2 x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + \sigma^2 = \sigma^2[x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + 1] \end{aligned} \tag{2.28}$$

donde se usa el hecho de que $\tilde{\beta}$ y e_0 son independientes y así

$$E[x_0'(\tilde{\beta} - \beta) - e_0] = E[x_0'(\tilde{\beta} - \beta)]E[e_0^2] = 0. \tag{2.29}$$

Consecuentemente, la variable aleatoria

$$\frac{x_0'\tilde{\beta} - y_0}{\sigma\sqrt{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + 1}} \tag{2.30}$$

está distribuida como una variable aleatoria normal estándar con media 0 y varianza 1. Esto significa que la variable aleatoria

$$\frac{x_0'\tilde{\beta} - y_0}{\sigma\sqrt{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + 1}} \div \sqrt{\frac{(T-K)\hat{\sigma}^2/\sigma^2}{T-K}} = \frac{x_0'\tilde{\beta} - y_0}{\hat{\sigma}\sqrt{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + 1}} \tag{2.31}$$

está distribuida como una variable t con $(T-K)$ grados de libertad. Como antes, se puede escribir

$$Pr[-t_{(T-K,\alpha/2)} \leq t_{(T-K)} \leq t_{(T-K,\alpha/2)}] = 1 - \alpha \tag{2.32}$$

o

$$Pr\left[-t_{(T-K,\alpha/2)} \leq \frac{x_0'\tilde{\beta} - y_0}{\hat{\sigma}\sqrt{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + 1}} \leq t_{(T-K,\alpha/2)}\right] = 1 - \alpha \tag{2.33}$$

que puede ser reescrita como

$$\begin{aligned} &Pr[x_0'\tilde{\beta} - t_{(T-K,\alpha/2)}\hat{\sigma}\sqrt{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + 1} \leq y_0 \\ &\leq x_0'\tilde{\beta} + t_{(T-K,\alpha/2)}\hat{\sigma}\sqrt{x_0'(\mathbf{X}'\mathbf{X})^{-1}x_0 + 1}] = 1 - \alpha. \end{aligned} \tag{2.34}$$

Se observa de nuevo que este intervalo es similar al discutido previamente en los intervalos de confianza implicando un parámetro conocido donde solamente los límites son aleatorios. La diferencia aquí es que se está concentrado con valores desconocidos de una variable aleatoria, y todo el lado derecho (expresado como variable t) es aleatorio. En términos de la interpretación, el intervalo predice con

probabilidad $(1-\alpha)$ que el valor de la variable aleatoria está contenido dentro de éste.

Prueba de Hipótesis

Muchos problemas de decisión requieren una base para decidir si un parámetro o vector de parámetros está en un subespacio específico ω del espacio de parámetros Ω. Por ejemplo, en el desarrollo de la estimación puntual máximo verosímil para β y σ^2, hay que concentrarse en el espacio paramétrico

$$\Omega = \{\beta, \sigma^2; \beta \in E_K, \sigma^2 > 0\} \tag{2.35}$$

donde E_k es el espacio Euclidiano K-dimensional, implicando el espacio de vectores K-dimensional. Consecuentemente, los elementos de β están irrestringidos y σ^2 esta restringido a ser positivo.

Alternativamente, se puede tener la conjetura o hipótesis lineal general que los parámetros del vector β están contenidos en el subespacio para el que $\mathbf{R}\beta = \mathbf{r}$, donde $\mathbf{R}$ es una hipótesis conocida expresada a través de una matriz $(J \times K)$ de rango $J \leq K$ y $\mathbf{r}$ es un vector $(J \times 1)$ conocido. $\mathbf{R}$ tiene las mismas características que la matriz usada en la forma de sistema de restricciones lineales presentada anteriormente. En el contexto de prueba de hipótesis, $H_0 : \mathbf{R}\beta = \mathbf{r}$ es la hipótesis nula. Por ejemplo, en el espacio bidimesional, si la hipótesis es $\beta_1 = \beta_2$, o $\beta_1 - \beta_2 = 0$, entonces en la notación general,

$$\mathbf{R}\beta = \mathbf{r} \tag{2.36}$$

se tendría

$$(1 \quad -1)\begin{pmatrix}\beta_1 \\ \beta_2\end{pmatrix} = 0. \tag{2.37}$$

El interés puede centrarse en hipótesis lineales compuestas tal como $\beta_1 = \beta_2$ y $\beta_1 + \beta_2 + ... + \beta_k = 1$. Dentro del esquema hasta ahora descrito, para expresar la hipótesis lineal $\mathbf{R}\beta = \mathbf{r}$ puede ser escrita como

$$R\beta = \begin{pmatrix}1 & -1 & 0 & ... & 0 \\ 1 & 1 & 1 & ... & 1\end{pmatrix}\beta = \begin{pmatrix}0 \\ 1\end{pmatrix} = 0. \tag{2.38}$$

Otro ejemplo puede ser la hipótesis

$$\begin{aligned}\beta_1 - 3\beta_3 &= 6\beta_4 \\ \frac{\beta_1}{\beta_2} &= 4\end{aligned} \tag{2.39}$$

La hipótesis puede ser expresada en el esquema lineal general como

$$\begin{pmatrix}1 & 0 & -3 & -6 & 0 & ... & 0 \\ 1 & -4 & 0 & 0 & 0 & ... & 0\end{pmatrix}\beta = \begin{pmatrix}0 \\ 0\end{pmatrix} = 0. \tag{2.40}$$

Una hipótesis lineal general a menudo expresada es $\mathbf{R}\beta = \mathbf{I}_k\beta = \mathbf{r} = \beta_0$. Bajo esta especificación, el vector de parámetros β es contrastado como igual a un vector K-dimensional β_0. En este caso, la hipótesis nula es $H_0 : \beta = \beta_0$ y la hipótesis alternativa es $H_0 : \beta \neq \beta_0$.

Dada la representación alternativa del espacio de parámetros, es necesaria una base para probar si los datos son consistentes con H_0 y una regla de decisión para determinar si se acepta o se rechaza H_0. Así es necesaria una prueba estadística y una base para particionar el espacio muestral en una región de aceptación y una de rechazo. En el desarrollo de un mecanismo de esta forma dos tipos de errores son posibles. Se puede obtener un valor muestral de esta prueba estadística que yace en

la región de rechazo y rechazar a H_0 cuando de hecho ésta es verdadera -un error de tipo I. Alternativamente, es posible obtener un valor de la prueba estadística que cae en la región de aceptación y aceptar H_0 cuando ésta de hecho es falsa - un error de tipo II.

Dado este escenario, usualmente no es posible elegir una región de aceptación y una de rechazo para una muestra dada T y minimizar ambos errores. Usualmente, un máximo aceptable de error de tipo I denotado por α es elegido y entonces la región crítica de rechazo que minimiza la probabilidad de un error de tipo II es encontrada.

Una forma de abordar el problema de encontrar una adecuada prueba es estableciendo una prueba estadística que tenga una distribución conocida cuando la hipótesis nula $R\beta = r$ es verdad, y que tiene otra distribución cuando H_0 no es verdad. El conocimiento de la distribución de la estadística bajo H_0 permite colocar la región de aceptación y la región de rechazo de tal forma que la probabilidad de un error de tipo I es fijado en algún nivel preespecificado. Es también deseable para estas pruebas que esten construidas en tal forma que tengan una alta potencia. Es desable que la probabilidad de rechazar H_0 cuando H_0 es falsa sea alta. Un método de obtener pruebas con potencias características deseables es el principio de la razón de verosimilitud.

Para probar $H_0 : \mathbf{R}\beta = \mathbf{r}$ contra la alternativa $H_1 : \mathbf{R}\beta \neq \mathbf{r}$ se procede como sigue: se comienza sugiriendo un estadístico que tiene una distribución conocida bajo H_0. Entonces, se resume como esta misma estadística puede ser obtenida usando el principio de razón de verosimilitud, probando así que ésta tiene deseables potencias características. En el proceso de ver la prueba estadística como una prueba estadística de razón de verosimilitud es posible descubrir que ésta tiene dos representaciones adicionales que proporcionan interpretaciones útiles y significativas.

Para sugerir una estadística que tiene distribución conocida bajo H_0 se sigue la misma línea de razonamiento como fue usada par establecer una región confidencial conjunta para dos combinaciones lineales de los elementos en β. Reiterando este argumento para el caso mas general de una matriz $\mathbf{R}$ que es de dimensión $(J \times K)$ mas que $(2 \times K)$. Se tiene

$$\mathbf{R}\tilde{\beta} \sim N[\mathbf{R}\beta, \sigma^2\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']. \tag{2.41}$$

Entonces,

$$Q_1 = \frac{(\mathbf{R}\tilde{\beta} - \mathbf{R}\beta)'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\tilde{\beta} - \mathbf{R}\beta)}{\sigma^2} \sim \chi^2_{(J)} \tag{2.42}$$

donde los grados de libertad J están dados por el número de elementos en el vector $\mathbf{Rb}$.

Para eliminar la σ^2 desconocida se usa el hecho que

$$Q_2 = \frac{(T-K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(T-K)} \tag{2.43}$$

y que estas dos últimas expresiones son independientes. La razón de Q_1 a Q_2, cada una dividida por sus grados de libertad, forman un estadístico F. Esto es,

$$\lambda_1 = \frac{Q_1/J}{Q_2/(T-K)}$$

$$= \frac{(\mathbf{R}\tilde{\beta} - \mathbf{R}\beta)'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\tilde{\beta} - \mathbf{R}\beta)}{J\hat{\sigma}^2} \sim F_{(J,T-K)}. \tag{2.44}$$

Ahora, cuando $H_0 : \mathbf{R}\beta = \mathbf{r}$ es verdadera, λ_1 llega a ser

$$\lambda = \frac{(\mathbf{R}\tilde{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\tilde{\beta} - \mathbf{r})}{J\hat{\sigma}^2} \sim F_{(J,T-K)}. \tag{2.45}$$

Esta expresión no depende de algún parámetro desconocido y tiene una distribución conocida cuando H_0 es verdadera. Así ésta puede ser usada como un estadístico, donde se rechaza H_0 cuando λ es más grande que el apropiado valor crítico tomado de los valores tabulados de la distribución F con J y $(T - K)$ grados de libertad. Note que $\mathbf{R}\tilde{\beta}$ puede ser considerada como el estimador irrestringido para $\mathbf{R}\beta$, y que, otras cosas siendo iguales, los demás estimadores irrestringidos $\mathbf{R}\tilde{\beta}$ son de la colección de valores restringidos $\mathbf{r}$, la más grande λ es probablemente ésta, y lo más probable es que H_0 será rechazada.

Es posible mostrar que λ puede ser vista como una estadística de prueba de razón de verosimilitud, y por lo tanto, ésta tiene una apelación intuitiva y alguna potencia característica deseable.

La estadística de prueba de razón de verosimilitud refleja la compatibilidad entre una muestra de datos y la hipótesis nula por una comparación de funciones de verosimilitud restringidas e irrestringidas. Si se presenta la hipótesis lineal general sobre los parámetros desconocidos como $\mathbf{R}\beta = \mathbf{r}$, donde $\mathbf{R}$ es una hipótesis conocida expresada como una matriz $(J \times K)$ y $\mathbf{r}$ es $(J \times 1)$, la razón de verosimilitud es

$$\frac{\max \ell(\beta, \sigma^2|\mathbf{y}, \mathbf{X})}{\max \ell(\beta, \sigma^2|\mathbf{y}, \mathbf{X}, \mathbf{R}\beta = \mathbf{r})} = \frac{\hat{\ell}(\Omega)}{\hat{\ell}(\omega)} = \lambda_0. \tag{2.46}$$

El numerador de ésta última expresión es el máximo de la función irrestringida, y el denominador es el máximo de la función restringida. Si ambas verosimilitudes son maximizadas, una restringida y la otra irrestringida, el valor de la irrestringida no será más pequeña que el valor de la restringida y de aquí la razón $\lambda_0 \geq 1$. Para ver que ésta es una plausible prueba estadística, recuerde que se piensa en la función

de verosimilitud $\ell(\cdot)$ como una medida de que también β explica la información muestral dada $\mathbf{y}$. Así $\hat{\ell}(\Omega)$ es grande comparada con $\hat{\ell}(\omega)$, entonces la muestra observada es la mejor explicada por algún β en un Ω y viceversa. La pregunta crítica, y una que depende de la distribución de probabilidad de λ_0, es por cuanto debe exceder λ_0 la unidad antes de que alguna duda surja en la validez de la hipótesis nula. En esta parte, se rechaza H_0 cuando λ_0 es mayor que algún valor crítico.

Regresando a la razón de verosimilitud, $\ell(\beta, \sigma^2|\mathbf{y}, \mathbf{X})$ ésta alcanza su máximo cuando $\beta = \tilde{\beta} = \mathbf{b}$, el estimador máximo verosímil irrestringido. Alternativamente, $\ell(\beta, \sigma^2|\mathbf{y}, \mathbf{X}, \mathbf{R}\beta = \mathbf{r})$ alcanza su máximo cuando $\beta = \mathbf{b}^*$, el estimador máximo verosímil restringido que es consistente con la hipótesis lineal general $\mathbf{R}\beta = \mathbf{r}$. Maximizar la función de verosimilitud resulta encontrando estimadores $\mathbf{b}$ y $\mathbf{b}^*$ que minimicen el respectivo error de la suma de cuadrados. Después de algo de álgebra es posible mostrar el rechazo de H_0 cuando λ_0 es más grande que una constante, es equivalente a rechazar H_0 cuando

$$\lambda_0^* = \frac{(\mathbf{y} - \mathbf{X}\mathbf{b}^*)'(\mathbf{y} - \mathbf{X}\mathbf{b}^*)}{(\mathbf{y} - \mathbf{X}\mathbf{b})'(\mathbf{y} - \mathbf{X}\mathbf{b})} = \frac{SCE_R}{SCE_I} \tag{2.47}$$

es más grande que una constante. Así, la pregunta concerniente a la compatibilidad de los datos con la hipótesis nula puede ser encuadrada en términos de la compatibilidad del estimador mínimo cuadrático restringido e irrestringido $\mathbf{b}$ y $\mathbf{b}^*$, y su respectiva suma de error al cuadrado (SCE_I y SCE_R).

Como ésta yace, la estadística λ_0^* no es muy útil porque su distribución no es fácilmente reconocida. Sin embargo, se puede aplicar una simple transformación para obtener una estadística ligeramente modificada con una distribución conocida. Esta estadísitica es,

$$\lambda = \frac{(\lambda_0^* - 1)(T-K)}{J} = \frac{[(y - Xb^*)'(y - Xb^*) - (y - Xb)'(y - Xb)]/J}{(y - Xb)'(y - Xb)/(T-K)}$$

$$= \frac{SCE_R - SCE_I}{J\hat{\sigma}^2} \tag{2.48}$$

Puede ser mostrado que esta estadística es idéntica a un λ como el expresado en (2.45), y de aquí que ésta siga una distribución F con $(J, T-K)$ grados de libertad. Claramente, si se rechaza H_0 cuando λ_0^* es más grande que alguna constante es equivalente a rechazar H_0 cuando λ es mas grande que otra constante seleccionada. La interpretación que está colocada en λ es que cuando la suma del error restringida (SCE_R) es significativamente más grande que la suma de cuadrados del error irrestringida (SCE_I), hay alguna duda sobre la validez de las restricciones así ellas (la hipotesis nula) son rechazadas.

Si se elige un nivel de significancia α para definir la región de rechazo, y así el valor crítico de $F_{(J,T-K,\alpha)}$, nuestra decisión de rechazo o aceptación depende de si el valor muestral de $F_{(J,T-K)}$ cae sobre (rechaza) o debajo (acepta) del valor crítico de la tabla de la distribución F. Por ejemplo, la decisión de rechazo sigue esta línea de razonamiento: si $\mathbf{R}\beta = \mathbf{r}$ y λ siguen una distribución F, el valor muestral F que excede el valor crítico relevante de F aparecerá por casualidad, en el sentido de un muestreo repetido, un α por ciento de las veces. Esto señala que la probabilidad de obtener un valor F, que es más grande que el valor crítico, es bastante pequeño, y por lo tanto, se tienen fundamentos para rechazar la hipótesis. La decisión de aceptar sigue de si justamente lo contrario es verdadero. Esto significa que se rechaza la hipótesis $H_0 : \mathbf{R}\beta = \mathbf{r}$ sí y sólo sí $\lambda \geq F_{(J,T-K,\alpha)}$, donde $F_{(J,T-K,\alpha)}$, es la probabilidad superior (crítica) para la variable aleatoria F central con J y $(T-K)$ grados de libertad.

Se derivará una tercera expresión que apela a la intuición. La tarea es probar que

$$SCE_R - SCE_I = \hat{\mathbf{e}}^{*\prime}\mathbf{e}^* - \hat{\mathbf{e}}'\hat{\mathbf{e}} = (\mathbf{Rb} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r}) \qquad (2.49)$$

donde $\hat{\mathbf{e}} = \mathbf{y} - \mathbf{Xb}$ es el vector de residuales mínimo cuadráticos irrestringidos y $\hat{\mathbf{e}}^* = \mathbf{y} - \mathbf{Xb}^*$ es el vector de residuales mínimo cuadráticos restringidos. Se tiene que

$$\begin{aligned}\hat{\mathbf{e}}^* &= \mathbf{y} - \mathbf{Xb}^* \\ &= \mathbf{y} - \mathbf{X}[\mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb}) \\ &= \hat{\mathbf{e}} - \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb}) \qquad (2.50)\end{aligned}$$

Así,

$$\begin{aligned}\hat{\mathbf{e}}^{*'}\hat{\mathbf{e}}^* = \hat{\mathbf{e}}'\hat{\mathbf{e}} &+ (\mathbf{r} - \mathbf{Rb})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{X} \\ &\times(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb}) \\ &-\hat{\mathbf{e}}'\mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{Rb}) \\ &-(\mathbf{r} - \mathbf{Rb})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\hat{\mathbf{e}} \\ &= \hat{\mathbf{e}}'\hat{\mathbf{e}} + (\mathbf{Rb} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{Rb} - \mathbf{r}) \qquad (2.51)\end{aligned}$$

porque $\mathbf{X}'\mathbf{e} = \mathbf{0}$. Despejando $\mathbf{e}'\mathbf{e}$ al lado izquierdo de esta última expresión proporciona el resultado.

Para derivar la tercer expresión útil para la prueba estadística λ se comienza con el estimador mínimo cuadrático restringido

$$\mathbf{b}^* = \mathbf{b} + (\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}) \tag{2.52}$$

de esto sigue que

$$\mathbf{X}(\mathbf{b}^* - \mathbf{b}) = \mathbf{X}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}). \tag{2.53}$$

Si ambos lados de esta última son premultiplicados por sus respectivas transpuestas se tiene

$$(\mathbf{b}^* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}^* - \mathbf{b}) = (\mathbf{r} - \mathbf{R}\mathbf{b})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{r} - \mathbf{R}\mathbf{b}). \tag{2.54}$$

El lado derecho de esta expresión es el numerador para λ dado con anterioridad. Así, la tercera expresión para λ es

$$\lambda = \frac{(\mathbf{b}^* - \mathbf{b})'\mathbf{X}'\mathbf{X}(\mathbf{b}^* - \mathbf{b})}{J\hat{\sigma}^2}. \tag{2.55}$$

Esta expresión es atrayente porque muestra que la más grande divergencia entre los estimadores restringidos e irrestringidos ($\mathbf{b}$ y $\mathbf{b}^*$), lo más probable es que la hipótesis nula sea rechazada.

Finalmente, es preciso notar que el caso especial $\mathbf{R}\beta = \mathbf{I}_k\beta = \mathbf{r}$ es usado en muchos libros y programas de computo. En otras palabras, cada elemento del vector K-dimensional β es establecido en la hipótesis como igual a una constante desconocida que a menudo es cero. Por lo tanto, la hipótesis nula es $H_0 : \beta = \mathbf{r}$ y la alternativa es $H_1\beta \neq r$. Esta formulación conduce, usando los procedimientos precedentes, a la prueba estadística

$$\lambda = \frac{(\mathbf{b} - \mathbf{r})'\mathbf{X}'\mathbf{X}(\mathbf{b} - \mathbf{r})}{K\hat{\sigma}^2} \tag{2.56}$$

que está distribuida como una variable aleatoria F con K y $(T - K)$ grados de libertad.

En muchos casos se esta interesado en probar una hipótesis individual implicando un coeficiente particular o algúna combinación lineal de los coeficientes. Por ejemplo, se puede tener la hipótesis $\beta_1 = r_1$ o la hipótesis $\beta_1 + \beta_2 + \beta_3 + ... + \beta_k = r_1$ donde r_1 es cualquier escalar, incluyendo el 0. Dentro del armazón de la hipótesis lineal general bajo discusión esta hipótesis individual puede ser representada como

$$\mathbf{R}_1\beta = (1 \quad 0 \quad 0 \quad ... \quad 0)\,\beta = \mathbf{r}_1 \tag{2.57}$$

y

$$\mathbf{R}_1\beta = (1 \quad 1 \quad ... \quad 1)\,\beta = \mathbf{r}_1. \tag{2.58}$$

donde $\mathbf{R}_1$ es un vector renglón de dimensión K. En este caso la hipótesis a través de la matriz $\mathbf{R}$ llega a ser un vector, y la prueba estadística

$$\lambda = \frac{(\mathbf{R}\tilde{\beta} - \mathbf{r})'[\mathbf{R}(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}']^{-1}(\mathbf{R}\tilde{\beta} - \mathbf{r})}{J\hat{\sigma}^2} \sim F_{(J,T-K)}$$

ya que $J = 1$, llega a ser,

$$\lambda = \frac{(\mathbf{R}_1\mathbf{b} - \mathbf{r}_1)[\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1']^{-1}(\mathbf{R}_1\mathbf{b} - \mathbf{r}_1)}{\hat{\sigma}^2} = \frac{(\mathbf{R}_1\mathbf{b} - \mathbf{r}_1)^2}{\hat{\sigma}^2[\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1']} \sim F_{(1,T-K)} \tag{2.59}$$

donde $\hat{\sigma}[\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1']$ es un escalar y es una estimación de la varianza de $\mathbf{R}_1\mathbf{b} - \mathbf{r}_1$.

Consecuentemente, se rechaza la hipótesis individual si $\lambda \geq F(1, T-K, \alpha)$. En este caso especial de la hipótesis lineal, implica que el coeficiente individual β_1 y $\mathbf{r}_1 = 0$, la prueba estadística llega a ser

$$\frac{\mathbf{b}_1^2}{\hat{\sigma}^2[\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1']} \sim F_{(1,T-K)}.$$

De la teoría de las distribuciones, se sabe que, para el caso especial de una hipótesis individual,

$$F_{(1,T-K)} = t_{(T-K)}^2 \tag{2.60}$$

donde $t_{(T-K)}$ es la variable aleatoria t de student con $(T-K)$ grados de libertad. Por lo tanto, llega a ser

$$\lambda = \frac{(\mathbf{R}_1\mathbf{b} - \mathbf{r}_1)^2}{\hat{\sigma}^2[\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1']} \sim t_{(T-K)}^2 \tag{2.61}$$

donde

$$\frac{(\mathbf{R}_1\mathbf{b} - \mathbf{r}_1)}{\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}} \sim t_{(T-K)}.$$

En este caso se rechaza la hipótesis si el valor absoluto de esta última es más grande que el valor crítico $t_{(T-K,\alpha/2)}$.

Para ver la relación entre la prueba de hipótesis y la estimación por intervalo, note que se acepta la hipótesis sí y sólo sí $\mathbf{r}_1 = \mathbf{R}\beta$ está en el intervalo

$$\begin{aligned}&[\mathbf{R}_1\mathbf{b} - t_{(T-K,\alpha/2)}\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'} \\ &\leq \mathbf{r}_1 \leq \mathbf{R}_1\mathbf{b} + t_{(T-K,\alpha/2)}\hat{\sigma}\sqrt{\mathbf{R}_1(\mathbf{X}'\mathbf{X})^{-1}\mathbf{R}_1'}]\end{aligned} \tag{2.62}$$

Considerando ahora la hipótesis $H_0 : \sigma^2 = \sigma_0^2$ contra la hipótesis $H_1 : \sigma^2 \neq \sigma_0^2$. El esquema para el mecanismo de esta prueba sigue del resultado

$$\frac{(T-K)\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{(T-K)}. \tag{2.63}$$

Si se supone que el escalar $\sigma^2 = \sigma_0^2$ y se asume que la hipótesis es correcta, la forma cuadrática es

$$(T-K)\frac{\hat{\sigma}^2}{\sigma_0^2} \sim \chi^2_{(T-K)}. \tag{2.64}$$

Se rechazará la hipótesis si el valor muestral de la prueba es menor que $\chi^2_{(T-K,\alpha/2)}$ o más grande que $\chi^2_{(T-K,1-\alpha/2)}$ o, dentro del contexto de un intervalo estimado, si σ_0^2 cae fuera del intervalo

$$\left[\frac{(T-K)\hat{\sigma}^2}{\chi^2_{(T-K,1-\alpha/2)}}, \frac{(T-K)\hat{\sigma}^2}{\chi^2_{(T-K,\alpha/2)}}\right]. \tag{2.65}$$

CAPITULO 3

EL COMPORTAMIENTO ASINTOTICO DE LOS ESTIMADORES EN EL MODELO DE REGRESION

Cuando se hace referencia a las propiedades del estimador mínimo cuadrático ha sido en el contexto de muestras finitas. Se deduce la distribución exacta del estimador y de algunos contrastes estadísticos bajo los supuestos de regresores no estocásticos y de perturbaciones normalmente distribuidas, siendo estos resultados independientes del tamaño de la muestra. Pero el modelo de regresión clásico con regresores no estocásticos y perturbaciones normalmente distribuidas es un caso especial que no incluye las aplicaciones mas comunes.

En muchos casos, las únicas propiedades conocidas de un estimador son aquellas que se aplican a grandes muestras. De forma que, solo se puede aproximar el comportamiento muestral finito mediante la utilización de lo que se conoce sobre las propiedades de muestras grandes, por lo que es útil conocer el comportamiento asintótico de los estimadores de los parámetros en el modelo clásico de regresión.

La consistencia del estimador mínimo cuadrático

$$b = (X'X)^{-1}X'y = \beta + (X'X)^{-1}X'e \tag{3.1}$$

puede ser deducida de la forma siguiente. Se empieza suponiendo que

$$\lim_{T\to\infty} \frac{1}{T} X'X = Q \tag{3.2}$$

donde Q es una matriz positiva definida. El estimador de mínimos cuadrados podría escribirse

$$b = \beta + \left(\frac{1}{T}X'X\right)^{-1}\left(\frac{1}{T}X'e\right). \tag{3.3}$$

Si Q^{-1} es finita y no singular

$$\text{plim}\, b = \beta + Q^{-1}\text{plim}\, \frac{1}{T}X'e. \tag{3.4}$$

Ahora se necesita el límite probabilístico del último término. Sea

$$\frac{1}{T}\sum x_i e_i = \frac{1}{T}\sum w_i = \bar{w} \tag{3.5}$$

entonces

$$\text{plim}\, b = \beta + Q^{-1}\text{plim}\, \bar{w}. \tag{3.6}$$

Ya que se supone que X es una matriz no estocástica

$$E(\bar{w}) = \frac{1}{T}\sum E(w_i) = \frac{1}{T}\sum x_i E(e_i) = \frac{1}{T}X'E(e) = 0 \tag{3.7}$$

y como $E(ee') = \sigma^2 I$,

$$var(\bar{w}) = E(\bar{w}\bar{w}') = E\left(\frac{1}{T}X'ee'X\frac{1}{T}\right) = \frac{1}{T}X'E(ee')X\frac{1}{T}$$

$$= \frac{1}{T} X' \sigma^2 X \frac{1}{T} = \frac{1}{T} \sigma^2 \frac{X'X}{T} = \frac{1}{T} \sigma^2 Q \tag{3.8}$$

entonces se deduce que

$$\lim_{T \to \infty} var(\bar{w}) = 0\, Q = 0. \tag{3.9}$$

Como la media de $\bar{w}$ es cero, y su varianza converge a cero, $\bar{w}$ converge en media cuadrática a cero, luego $\operatorname{plim} \bar{w} = 0$. De esta manera,

$$\operatorname{plim} \frac{1}{T} X'e = 0 \tag{3.10}$$

de modo que

$$\operatorname{plim} b = \beta + Q^{-1} 0 = \beta \tag{3.11}$$

lo que establece que b es un estimador consistente de β en el modelo clásico de regresión.

Para deducir la normalidad asintótica del estimador mínimo cuadrático, se considera a éste estimador como se expresó en la ecuación (3.3)

$$b = \beta + \left(\frac{1}{T} X'X\right)^{-1} \left(\frac{1}{T} X'e\right)$$

de lo cual se puede deducir que

$$b - \beta = \left(\frac{1}{T} X'X\right)^{-1} \left(\frac{1}{T} X'e\right) \tag{3.12}$$

$$\sqrt{T}(b-\beta) = \left(\frac{1}{T}X'X\right)^{-1}\left(\frac{1}{\sqrt{T}}X'e\right). \tag{3.13}$$

Por lo tanto si la distribución límite de esta última expresión existe, será igual a la distribución límite de la siguiente expresión,

$$\left[\lim_{T\to\infty}\left(\frac{X'X}{T}\right)^{-1}\right]\left(\frac{1}{\sqrt{T}}\right)X'e = Q^{-1}\left(\frac{1}{\sqrt{T}}\right)X'e. \tag{3.14}$$

Así se establece la distribución límite

$$\frac{1}{\sqrt{T}}X'e = \frac{1}{\sqrt{T}}\sum w_i = \sqrt{T}[\bar{w} - E(\bar{w})] \tag{3.15}$$

donde $E(\bar{w}) = 0$.

Puede emplearse la versión de Lindberg-Feller del teorema central del límite para obtener la distribución límite de $\sqrt{T}\bar{w}$. Utilizando esta formulación, $\bar{w} = (1/T)\Sigma x_i e_i$ y $var(x_i e_i) = \sigma^2 Q_i$. La varianza de $\sqrt{T}\bar{w}$ es

$$var(\sqrt{T}\bar{w}) = var(\sum w_i/\sqrt{T}) = \frac{1}{\sqrt{T}}X'E(ee')X\frac{1}{\sqrt{T}} = \frac{\sigma^2}{T}X'X. \tag{3.16}$$

Entonces se tiene que se cumple lo siguiente,

$$\lim_{T\to\infty}\sigma^2\bar{Q}_T = \sigma^2 Q. \tag{3.17}$$

Aplicando el teorema de límite central de Lindberg-Feller al vector $\sqrt{T}\bar{w}$ se tiene que

$$\left(\frac{1}{\sqrt{T}}\right)X'e \xrightarrow{d} N[0, \sigma^2 Q]. \tag{3.18}$$

Entonces se sigue que,

$$Q^{-1}\left(\frac{1}{\sqrt{T}}\right)X'e \xrightarrow{d} N[Q^{-1}0, Q^{-1}(\sigma^2 Q)Q^{-1}] \tag{3.19}$$

o reuniendo términos

$$\sqrt{T}(b-\beta) \xrightarrow{d} N[0, \sigma^2 Q^{-1}]. \tag{3.20}$$

De manera que la distribución asintótica de b es

$$b \xrightarrow{a} N\left[\beta, \frac{\sigma^2}{T}Q^{-1}\right]. \tag{3.21}$$

En la práctica se estima $(1/T)\,Q^{-1}$ por medio de $(X'X)^{-1}$ y σ^2 con $e'e/T-k$. Esta distribución en el límite está degenerada porque

$$\lim_{T\to\infty} V(\hat{\beta}) = \lim_{T\to\infty} V(b) = \lim_{T\to\infty} \frac{\sigma^2}{T}\left(\frac{X'X}{T}\right)^{-1} = 0 \tag{3.22}$$

y entonces $\hat{\beta} \xrightarrow{d} \beta$ donde β es una constante, es decir, tiene toda la masa de probabilidad concentrada en β. Para solucionar este problema, hay que realizar una transformación del estimador, de forma que su varianza no se anule cuando $T \to \infty$

$$Z_T = \sqrt{T}(\hat{\beta}-\beta). \tag{3.23}$$

Esta transformación Z_T sigue una distribución normal con

$$E(Z_T) = E[\sqrt{T}(\hat{\beta} - \beta)] = 0 \tag{3.24}$$

y

$$V(Z_T) = V[\sqrt{T}(\hat{\beta} - \beta)] = T\frac{\sigma^2}{T}\left(\frac{X'X}{T}\right)^{-1} = \sigma^2\left(\frac{X'X}{T}\right)^{-1}. \tag{3.25}$$

De forma que

$$\sqrt{T}(\hat{\beta} - \beta) \sim N\left[0, \sigma^2\left(\frac{X'X}{T}\right)^{-1}\right] \quad \forall \quad T. \tag{3.26}$$

Conforme el tamaño muestral tiende a infinito, si se satisface que Q existe, la distribución asintótica de Z_T no es degenerada

$$\sqrt{T}(\hat{\beta} - \beta) \overset{a}{\sim} N(0, \sigma^2 Q^{-1}). \tag{3.27}$$

Si las perturbaciones no siguen una distribución normal, entonces el estimador mínimo cuadrático $\hat{\beta}$ no sigue una distribución normal y por lo tanto los estadísticos t y F no siguen una distribución t de Student y una F de Snedecor, respectivamente. Esto implica que no se pueden considerar estas distribuciones para realizar inferencia sobre los parámetros del modelo. Si se pudiera derivar una distribución asintótica, se podría utilizar como aproximación a la verdadera distribución de $\hat{\beta}$ en muestras finitas para construir intervalos de confianza, hacer contrastes de hipótesis, etc., siempre que contemos con una muestra suficientemente grande. El estimador mínimo cuadrático se puede escribir como

$$\hat{\beta} = \beta + \left(\frac{X'X}{T}\right)^{-1}\left(\frac{X'e}{T}\right) \tag{3.28}$$

de forma que la transformación del mismo $\sqrt{T}(\hat{\beta} - \beta)$ quede como sigue

$$\sqrt{T}(\hat{\beta} - \beta) = \sqrt{T}\left(\frac{X'X}{T}\right)^{-1}\frac{X'e}{T} = \left(\frac{X'X}{T}\right)^{-1}\frac{X'e}{\sqrt{T}}. \tag{3.29}$$

Si las perturbaciones estan idéntica e independientemente distribuidas con $E(e) = 0$, $E(e'e) = \sigma^2 I$, los regresores son no estocásticos y se cumple la existencia de una matriz Q, se tiene por el teorema de Mann-Wald

$$\frac{X'e}{\sqrt{T}} \xrightarrow{d} N(0, \sigma^2 Q) \tag{3.30}$$

y por el teorema de Cramer

$$\left(\frac{X'X}{T}\right)^{-1}\frac{X'e}{\sqrt{T}} \xrightarrow{d} N(0, \sigma^2 Q^{-1}QQ^{-1}). \tag{3.31}$$

Por lo tanto, la distribución asintótica de $\sqrt{T}(\hat{\beta}-\beta)$ es la misma independientemente de si las perturbaciones siguen una distribución normal. Por otra parte, es preciso valorar también la consistencia de $\hat{\sigma}^2$ como un estimador de σ^2. Desarrollando

$$\hat{\sigma}^2 = \frac{1}{T-k}e'Me \tag{3.32}$$

da lugar a

$$\hat{\sigma}^2 = \frac{1}{T-k}[e'e - e'X(X'X)^{-1}X'e]$$

$$= \frac{T}{T-k}\left[\frac{e'e}{T} - \left(\frac{e'X}{T}\right)\left(\frac{X'X}{T}\right)^{-1}\left(\frac{X'e}{T}\right)\right]. \tag{3.33}$$

La constante de la primera parte converge a 1. Se puede aplicar

$$\lim_{T\to\infty} \frac{1}{T}X'X = Q \qquad y \qquad \lim_{T\to\infty} \frac{1}{T}X'e = 0 \tag{3.34}$$

dos veces, y la regla del producto para límites de probabilidad para afirmar que el segundo término en el paréntesis converge a cero. Esto da lugar a

$$\bar{e}^2 = \frac{1}{T}\sum_{t=1}^{T} e_t^2. \tag{3.35}$$

Suponiendo que las perturbaciones son independientes, $\bar{e}^2$ es la media de la muestra aleatoria. Si los términos en la suma tienen varianza finita, entonces podemos aplicar el teorema de desigualdad de Chebychev. Por tanto, se supone que

$$E[e_t^4] = \phi_e < \infty. \tag{3.36}$$

Entonces se tiene que los términos de la suma son independientes, con media σ^2 y varianza $\phi_e - \sigma^4$. Así por el teorema de convergencia en media cuadrática el primer término entre paréntesis converge en probabilidad a σ^2, lo que da lugar al resultado

$$\operatorname{plim} \hat{\sigma}^2 = \sigma^2 \tag{3.37}$$

y por la regla del producto

$$\operatorname{plim} \hat{\sigma}^2 \frac{(X'X)^{-1}}{T} = \sigma^2 Q^{-1}. \tag{3.38}$$

De manera que el estimador apropiado de la matriz de covarianza asintótica es $\sigma^2(X'X)^{-1}$.

Otra forma de probar la consistencia de la varianza del estimador mínimo cuadrático necesita un resultado adicional conocido como el teorema de Khintchine. Este expresa que la media muestral computada de una muestra aleatoria de observaciones distribuidas idéntica e independientemente será un estimador consistente de la media poblacional. Para nuestros propósitos consideramos los errores al cuadrado $e_1^2, e_2^2, ..., e_T^2$ como una muestra aleatoria de observaciones distribuidas independiente e idénticamente con media $E[e_t^2] = \sigma^2$. La media muestral es dada por

$$\frac{1}{T}\sum_{t=1}^{T} e_t^2 = \frac{e'e}{T} \tag{3.39}$$

y del teorema de Khintchine

$$\operatorname{plim} \frac{e'e}{T} = \sigma^2. \tag{3.40}$$

De esta manera, se está en condiciones de probar la consistencia del estimador de la varianza mínimo cuadrática

$$\hat{\sigma}^2 = \frac{1}{T-K} e'(I - X(X'X)^{-1}X')e = \frac{T}{T-k}\left[\frac{e'e}{T} - \left(\frac{e'X}{T}\right)\left(\frac{X'X}{T}\right)^{-1}\left(\frac{X'e}{T}\right)\right] \tag{3.41}$$

usando los resultados establecidos hasta ahora

$$\operatorname{plim} \hat{\sigma}^2 = \operatorname{plim}\left(\frac{T}{T-k}\right)\left[\operatorname{plim}\frac{e'e}{T} - \operatorname{plim}\frac{e'X}{T}\left(\operatorname{plim}\frac{X'X}{T}\right)^{-1}\operatorname{plim}\frac{X'e}{T}\right]$$

$$= 1(\sigma^2 - 0Q^{-1}0) = \sigma^2. \tag{3.42}$$

Así $\hat{\sigma}^2$ es un estimador consistente de σ^2. Si los errores estan normalmente distribuidos, es posible mostrar que $\hat{\sigma}^2$ es un estimador consistente mostrando que su sesgo y varianza se aproximan a cero cuando $T \to \infty$.

Si la función de regresión es no lineal, entonces el análisis de esta sección debe aplicarse a los pseudo regresores mas que a las variables independientes. A parte de esta consideración, no se necesita ningún resultado adicional. Por lo que podemos aplicar esta discusión al modelo linearizado.

En general, los parámetros de un modelo intrínsecamente no lineal pueden ser estimados estableciendo una función de verosimilitud y encontrando las estimaciones de máximo verosimilitud. Bajo los supuestos clásicos concernientes a las perturbaciones estocásticas y las variables explicatorias, las estimaciones resultantes tendrán todas las propiedades asintóticas deseables. Por supuesto la aplicación del método de máxima verosimilitud esta condicionado bajo el supuesto de que los errores están distribuidos normalmente. Si no queremos hacer este supuesto, obviamente tenemos la opción de poder hacer nuestras estimaciones minimizando la suma del cuadrado de las desviaciones de los valores observados de y de los valores ajustados de $\hat{y}$, esto es, el método de mínimos cuadrados cuyas propiedades asintóticas han sido ya mencionadas. Ya que los parámetros a ser estimados entran en una forma no lineal, este método es llamado método de mínimos cuadrados no lineales. La principal diferencia entre este método y el método de mínimos cuadrados ordinarios (lineal) es que en el caso lineal las estimaciones pueden ser expresadas como una función lineal de las perturbaciones. Esto generalmente no es posible en el caso no lineal. Las estimaciones obtenidas por el método de mínimos cuadrados no lineales es exactamente el mismo que las estimaciones de máxima verosimilitud

si la maximización de la función de verosimilitud es alcanzada por la minimización de la suma de cuadrados de las desviaciones de los valores observados de y a los ajustados. Puede ser mostrado que aún sin el supuesto de normalidad la distribución asintótica de las estimaciones mínimo cuadráticas no lineales son normales y tienen la misma media y varianza que las estimaciones de máxima verosimilitud para el caso de perturbaciones normales. Así el supuesto de normalidad de las perturbaciones no siempre es tan crucial. De esta manera, ahora se observaran las propiedades asintóticas de los estimadores de máxima verosimilitud. Principalmente la de consistencia y la de normalidad asintótica.

El estimador de máxima verosimilitud es en general un método para encontar estimadores de parámetros. Mientras el método de máxima verosimilitud tiene características deseables, quizá las característica más deseable es su apelación a la intuición. El estimador de máxima verosimiltud encuentra los parámetros que muy probablemente hayan generado los datos observados dada la función de distribución.

Formalmente la función de verosimilitud es entendida como la densidad conjunta evaluada en una colección de observaciones. Sean $x_1, ..., x_n$ n observaciones independientes e idénticamente distribuidas de una función $f(x;\theta)$. La función de verosimilitud será

$$L(\theta; x_1, ..., x_n) = \prod_{i=1}^{n} f(x_i;\theta) \tag{3.43}$$

donde θ es un parámetro.

A menudo es más fácil trabajar con el logaritmo de la función de verosimilitud, una función creciente de L. Ya que la función de verosimilitud es monótona y tiene una primer derivada diferente de cero, el valor de θ que maximiza la función de verosimilitud también maximiza el logaritmo de la función de verosimilitud.

Una condición para que Log L sea maximizada en $\theta = \hat{\theta}$ es

$$\frac{\partial LogL}{\partial \theta}\Big|_{\theta=\hat{\theta}} = 0. \tag{3.44}$$

Cabe hacer mención que la solución de la ecuación de verosimilitud no necesariamente maximiza la función de verosimilitud. La solución puede resultar ser un mínimo o punto de silla, o puede el máximo ocurrir en algún límite admisible de valores de θ. A veces se resuelve fácilmente pero por lo común ha de ser resuelta por métodos iterativos. En muchos resultados estadísticos concernientes a los estimadores de máxima verosimilitud se necesita poner condiciones matemáticas en la función de densidad de variables aleatorias, las cuales frecuentemente son referidas como condiciones de regularidad. Para un función de densidad son las siguientes:(1) Para casi todo x, $\partial logf/\partial\theta$, $\partial^2 logf/\partial\theta^2$, y $\partial^3 logf/\partial\theta^3$ existen para todo $\theta \in \Theta$. (2) Para todo $\theta \in \Theta$,$\mid \partial f/\partial\theta \mid < F_1(x)$, $\mid \partial^2 logf/\partial\theta^2 \mid < F_2(x)$, y $\mid \partial^3 logf/\partial\theta^3 \mid < H(x)$ existen para todo $\theta \in \Theta$, las funciones F_1 y F_2 son integrables en el intervalo $(-\infty, \infty)$ y $\int_{-\infty}^{\infty} H(x)f(x,\theta)dx < M$ donde M no depende de θ. (3) Para todo $\theta \in \Theta$, la integral $\int_{-\infty}^{\infty}(\partial logf/\partial\theta)^2 f dx$ es finita y positiva.

La primera condición asegura la expansión de una serie de Taylor, la segunda permite la diferenciación bajo una integral, y la tercera expresa que la variable aleatoria $(\partial logf/\partial\theta)$ tiene una varianza finita.

Suponga que $\{X_i\}$ están independiente e idénticamente distribuidas con densidad $f(X,\theta)$. Se define

$$Q_n(\theta) \equiv \frac{1}{n} logL_n(\theta) = \frac{1}{n}\sum_{i=1}^{n} logf(X_i,\theta) \tag{3.45}$$

donde una variable X_i aparece en el argumento de f porque se necesita considerar la

propiedad de la función de verosimilitud como una variable aleatoria. Para probar la consistencia del estimador máximo verosímil, se necesita esencialmente mostrar que $Q_n(\theta)$ converge en probabilidad a una función no estocástica de θ, denotada $Q(\theta)$, que alcanza el máximo global en el verdadero valor de θ, denotada como θ_0.

Debe mostrarse porque se puede esperar $Q_n(\theta)$ converja a $Q(\theta)$ y porque se puede esperar que $Q(\theta)$ sea maximizada en θ_0. $Q_n(\theta)$ es $1/n$ veces la suma de las variables aleatorias independientes e idénticamente distribuidas por lo que se puede aplicar la ley de grandes números de Khintchine, probando que $E \log f(X_i, \theta) < \infty$. Por lo tanto,

$$\text{plim}_{n\to\infty} Q_n(\theta) = Q(\theta) \equiv E \log f(X_i, \theta). \tag{3.46}$$

De acuerdo al teorema de Jensen, sea X una variable aleatoria propia (esto es, no una constante) y $g(\cdot)$ sea una función estrictamente concava, esto es, $g[\lambda a + (1-\lambda)b] > \lambda g(a) + (1-\lambda)g(b)$ para cualquier $a < b$ y $0 < \lambda < 1$. Entonces,

$$Eg(X) < g(EX) \tag{3.47}$$

haciendo que g sea log y X sea $f(X, \theta)/f(X, \theta_0)$ obtenemos,

$$E \log \frac{f(X,\theta)}{f(X,\theta_0)} < \log E \frac{f(X,\theta)}{f(X,\theta_0)} \qquad \text{si} \qquad \theta \neq \theta_0. \tag{3.48}$$

Pero el lado derecho de la desigualdad de arriba es igual a uno, porque

$$E \frac{f(X,\theta)}{f(X,\theta_0)} = \int_{-\infty}^{\infty} \frac{f(X,\theta)}{f(X,\theta_0)} f(X,\theta_0) dx = \int_{-\infty}^{\infty} f(X,\theta) dx = 1. \tag{3.49}$$

Por lo tanto se obtiene

$$E \log f(X, \theta) < E \log f(X, \theta_0) \qquad \text{donde} \qquad \theta \neq \theta_0. \tag{3.50}$$

Se ha probado la consistencia de un estimador máximo verosímil global. Para probar la consistencia de un estimador máximo verosímil local se debe mostrar que

$$\frac{\partial}{\partial \theta} E \log L = 0. \tag{3.51}$$

Permitiendo, de momento, que el rango de X_i dependa de los parámetros; para cada elemento, X_{ik}, $L(\theta) \leq X_{ik} \leq U(\theta)$. En donde la integral simple $\int ...dx$ se usa aquí para indicar la integral múltiple respecto a todos los elementos de X_i. Por definición,

$$\int_{L(\theta)}^{U(\theta)} f(X|\theta)dx = 1. \tag{3.52}$$

Ahora, diferenciando esta expresión con respecto a θ, y aplicando el teorema de Leibnitz, se tiene,

$$\frac{\partial \int_{L(\theta)}^{U(\theta)} f(X|\theta)dx}{\partial \theta} = \int_{L(\theta)}^{U(\theta)} \frac{\partial f(X|\theta)}{\partial \theta} dx$$

$$+ f(U(\theta)|\theta)\frac{\partial U(\theta)}{\partial \theta} - f(L(\theta)|\theta)\frac{\partial L(\theta)}{\partial \theta} = 0. \tag{3.53}$$

Si el segundo y tercer término de esta última expresión tienden a cero, entonces se pueden intercambiar las operaciones de integración, diferenciación e integración. La condición necesaria es que, en el límite, la función de densidad se anule. Una

condición suficiente es que el rango de la variable aleatoria observada, x_i, no dependa de los parámetros, lo que significa que,

$$\frac{\partial L(\theta)}{\partial \theta} = \frac{\partial U(\theta)}{\partial \theta} = 0. \tag{3.54}$$

Puesto que $L(x, \theta)$ es la función de densidad conjunta de las observaciones, se tiene $\int L(X, \theta)dx = 1$ donde $x = x_1, x_2, ..., x_n$; $dx = dx_1, dx_2, ..., dx_n$ y $\int$=integral múltiple.

Bajo el supuesto que se puede diferenciar bajo el signo de la integral.

$$\int \frac{\partial L}{\partial \theta} dy = \int \frac{1}{L} \frac{\partial L}{\partial \theta} L \, dy = E\left(\frac{\partial log L}{\partial \theta}\right) = 0, \tag{3.55}$$

si se considera ahora

$$\int \frac{\partial log L}{\partial \theta} L \, dy = 0, \tag{3.56}$$

diferenciando una vez mas con respecto a θ, se obtiene una expresión conocida como información de θ en la muestra. Esta expresión juega un papel central en la teoría de la estimación máximo verosímil, ya que comunmente, bajo las condiciones de regularidad el método sigue una distribución normal con varianza igual a la inversa de la matriz de información. De esta manera, el procedimiento que lleva a la matriz de información es el siguiente

$$\frac{\partial}{\partial \theta} \int \frac{\partial log L}{\partial \theta} L dy = 0,$$

$$\int \frac{\partial}{\partial\theta}\frac{\partial logL}{\partial\theta}Ldy = 0,$$

$$\int \left(\frac{\partial^2 logL}{\partial\theta^2}Ldy + \frac{\partial L}{\partial\theta}\frac{\partial logL}{\partial\theta}dy\right) = 0,$$

$$\int \left(\frac{\partial^2 logL}{\partial\theta^2}Ldy + \frac{\partial logL}{\partial\theta}L\frac{\partial logL}{\partial\theta}dy\right) = 0,$$

$$\int \left(\frac{\partial^2 logL}{\partial\theta^2}Ldy + \left(\frac{\partial logL}{\partial\theta}\right)^2 Ldy\right) = 0,$$

$$\int \left[\frac{\partial^2 logL}{\partial\theta^2} + \left(\frac{\partial logL}{\partial\theta}\right)^2\right]Ldy = 0,$$

$$-\int \frac{\partial^2 logL}{\partial\theta^2}Ldy = \int \left(\frac{\partial logL}{\partial\theta}\right)^2 Ldy,$$

$$-E\left(\frac{\partial^2 logL}{\partial\theta^2}\right) = E\left(\frac{\partial logL}{\partial\theta}\right)^2 = I(\theta)$$

$$E\left(-\frac{\partial^2 logL}{\partial\theta^2}\right) = E\left(\frac{\partial logL}{\partial\theta}\right)^2 = I(\theta). \tag{3.57}$$

Ahora bien, bajo las condiciones de regularidad, la estimación de la varianza asintótica del estimador de máxima verosimilitud dado por $[I(\theta)]^{-1}$ se le conoce como la información límite de la varianza o alternativamente como la cota inferior de Cramer-Rao para la varianza del estimador $\hat{\theta}$. De esta forma, se tiene la siguiente expresión:

$$v(\hat{\theta}) \geq -\frac{1}{E(\partial^2 logL/\partial\theta^2)}. \tag{3.58}$$

Suponga $X = \hat{\theta}$ y $Y = \partial logL/\partial\theta$. Entonces,

$$E(Y) = E\left(\frac{\partial logL}{\partial\theta}\right) = E\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right) = \int\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right)Ldx$$

$$= \frac{\partial}{\partial\theta}\int Ldx = \frac{\partial 1}{\partial\theta} = 0 \tag{3.59}$$

donde la integral es una n-tupla integral con respecto a $X_1, X_2, ..., X_n$. También se tiene

$$E\frac{\partial^2 logL}{\partial\theta^2} = E\frac{\partial}{\partial\theta}\frac{\partial logL}{\partial\theta} = E\frac{\partial}{\partial\theta}\left(\frac{1}{L}\frac{\partial L}{\partial\theta}\right) = -E\frac{1}{L^2}\left(\frac{\partial L}{\partial\theta}\right)^2 + E\frac{1}{L}\left(\frac{\partial^2 L}{\partial\theta^2}\right)$$

$$= -E\frac{1}{L^2}\left(\frac{\partial L}{\partial\theta}\right)^2 = -E\left(\frac{\partial logL}{\partial\theta}\right)^2 \tag{3.60}$$

donde,

$$E\left(\frac{1}{L}\right)\left(\frac{\partial^2 L}{\partial\theta^2}\right) = \int\left(\frac{\partial^2 L}{\partial\theta^2}\right)dx = \frac{\partial^2}{\partial\theta^2}(\int Ldx) = 0. \tag{3.61}$$

Por lo tanto,

$$V(y) = E(Y^2) = \frac{\partial^2 logL}{\partial\theta^2} \tag{3.62}$$

debido a que

$$E(Y) = E\left(\frac{\partial logL}{\partial \theta}\right) = 0. \tag{3.63}$$

También se tiene

$$cov(X,Y) = E\hat{\theta}\frac{\partial logL}{\partial \theta} = \int \hat{\theta}\frac{1}{L}\frac{\partial L}{\partial \theta}Ldx = \frac{\partial}{\partial \theta}\int \hat{\theta}Ldx = \frac{\partial}{\partial \theta}E\hat{\theta} = \frac{\partial \theta}{\partial \theta} = 1 \tag{3.64}$$

y por lo tanto $V(\hat{\theta})$ sigue la desigualdad de Cauchy-Schwartz.

Para las pruebas de significación y establecimiento de intervalos de confianza es preciso conocer la distribución asintótica del estimador de máxima verosimilitud. De esta manera, asumiendo condiciones de regularidad se tiene que

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N[0, M(\theta)^{-1})] \tag{3.65}$$

donde

$$M(\theta) = -E\left(\frac{\partial^2 LogL}{\partial \theta^2}\right). \tag{3.66}$$

Para poder entender esta expresión se comienza tomando una expansión de la primera derivada de la función log maximoverosímil, $l = LogL$,

$$\frac{\partial l(\hat{\theta}; x)}{\partial \theta} = \frac{\partial l(\theta; x)}{\partial \theta} + \frac{\partial^2 l(\theta; x)}{\partial \theta^2}(\hat{\theta} - \theta) + \frac{1}{2}\frac{\partial^3 l(\theta^*; x)}{\partial \theta^3}(\hat{\theta} - \theta)^2, \tag{3.67}$$

donde θ^* esta entre θ y $\hat{\theta}$. Note que es una primer derivada que está expandida en términos de $\hat{\theta}$ sobre θ y que el tercer término es un término de resto.

Manipulando y multiplicando por $\sqrt{T}$,

$$\sqrt{T}(\hat{\theta}-\theta)=$$

$$\left[\frac{1}{T}\frac{\partial^2 l(\hat{\theta};x)}{\partial\theta^2}\right]^{-1}\left[\frac{1}{\sqrt{T}}\frac{\partial l(\hat{\theta};x)}{\partial\theta}-\frac{1}{\sqrt{T}}\frac{\partial l(\theta;x)}{\partial\theta}-\frac{1}{2\sqrt{T}}\frac{\partial^3 l(\theta^*;x)}{\partial\theta^3}(\hat{\theta}-\theta)^2\right] \tag{3.68}$$

examinando cada término en la ecuación de arriba. Primero

$$\frac{1}{T}\frac{\partial^2 l(\hat{\theta};x)}{\partial\theta^2}=\left(\frac{1}{T}\right)\frac{\partial^2}{\partial\theta^2}log\prod_{i=1}^{T}f(x_i;\theta)=\left(\frac{1}{T}\right)\sum_{i=1}^{T}\frac{\partial^2}{\partial\theta^2}logf(x_i;\theta). \tag{3.69}$$

Colocando $Y_i=(\partial^2/\partial\theta^2)logf(x_i;\theta)$, puede ser visto que las Y_i son variables aleatorias independientes y que

$$E[Y_i]=\int_{-\infty}^{\infty}\left(\frac{\partial^2}{\partial\theta^2}logf(x_i;\theta)\right)f(x_i;\theta)dx_i$$

$$=\int_{-\infty}^{\infty}\left(\frac{\partial}{\partial\theta}\frac{f'(x_i;\theta)}{f(x_i;\theta)}\right)f(x_i;\theta)dx_i$$

$$=\int_{-\infty}^{\infty}\left(\frac{f''(x_i;\theta)}{f(x_i;\theta)}-\left[\frac{f'(x_i;\theta)}{f(x_i;\theta)}\right]^2\right)f(x_i;\theta)dx_i$$

$$=\int_{-\infty}^{\infty}f''(x_i;\theta)dx_i-\int_{-\infty}^{\infty}\left[\frac{f'(x_i;\theta)}{f(x_i;\theta)}\right]^2f(x_i;\theta)dx_i$$

$$=-E\left(\left[\frac{\partial}{\partial\theta}logf(x_i;\theta)\right]^2\right). \tag{3.70}$$

Así, las Y_i son independientes con idéntica media, y puede aplicarse la ley fuerte de los grandes números para obtener

$$\text{plim}\left(\frac{1}{T}\right)\frac{\partial^2 l(\theta;x)}{\partial\theta^2} = \text{plim}\,\frac{1}{T}\sum_{i=1}^{T} Y_i = -E\left[\frac{\partial}{\partial\theta} log f(x_i;\theta)\right]^2. \tag{3.71}$$

El siguiente término,

$$\frac{1}{\sqrt{T}}\frac{\partial l(\hat{\theta};x)}{\partial\theta} \tag{3.72}$$

es igual a cero porque $\hat{\theta}$ se asume es una solución de la ecuación de verosimilitud.

Para la expresión

$$\frac{1}{\sqrt{T}}\frac{\partial l(\theta;x)}{\partial\theta} \tag{3.73}$$

se hace

$$Z_i = \frac{f'(x_i;\theta)}{f(x_i;\theta)} \tag{3.74}$$

entonces se tiene

$$\frac{1}{\sqrt{T}}\frac{\partial l(\theta;x)}{\partial\theta} = \left(\frac{1}{\sqrt{T}}\right)\sum_{i=1}^{T}\frac{\partial}{\partial\theta} log f(x_i;\theta) = \frac{1}{\sqrt{T}}\sum_{i=1}^{T} Z_i. \tag{3.75}$$

Fue previamente mostrado que $E(Z_i) = 0$ y por definición de Z_i,

$$E(Z_i^2) = E\left[\left[\frac{\partial}{\partial\theta} log f(x_i;\theta)\right]^2\right]. \tag{3.76}$$

De aquí, la suma

$$\frac{1}{\sqrt{T}}\sum_{i=1}^{T} Z_i \tag{3.77}$$

es la suma de variables aleatorias independientes con media cero y varianza constante idéntica. Se puede aplicar así un teorema de límite central, para obtener

$$\frac{1}{\sqrt{T}}\frac{\partial l(\theta;x)}{\partial\theta} = \frac{1}{\sqrt{T}}\sum_{i=1}^{T} Z_i \xrightarrow{d} N\Big[0, E\Big\{\Big(\frac{\partial}{\partial\theta} log f(x_i;\theta)\Big)^2\Big\}\Big]. \tag{3.78}$$

Se puede ahora examinar el término final

$$\frac{1}{2\sqrt{T}}\frac{\partial^3 l(\theta^*;x)}{\partial\theta^3}(\hat{\theta}-\theta)^2 \tag{3.79}$$

por condiciones de regularidad se tiene que esta última expresión está acotada por una constante, y se sabe que $\hat{\theta} \xrightarrow{p} \theta$; por lo tanto, este término converge en probabilidad a cero.

En resumen,

$$\Big[\frac{1}{T}\frac{\partial^2 l(\hat{\theta};x)}{\partial\theta^2}\Big]^{-1} \xrightarrow{p} M(\theta)^{-1} \tag{3.80}$$

$$\frac{1}{\sqrt{T}}\frac{\partial l(\hat{\theta};x)}{\partial\theta} = 0 \tag{3.81}$$

$$\frac{1}{\sqrt{T}}\frac{\partial l(\theta;x)}{\partial\theta} \xrightarrow{d} N[0, M(\theta)] \tag{3.82}$$

$$\frac{1}{2\sqrt{T}}\frac{1}{2}\frac{\partial^3 l(\hat{\theta}^*;x)}{\partial\theta^3}(\hat{\theta}-\theta)^2 \xrightarrow{p} 0. \tag{3.83}$$

Finalmente, aplicando las reglas de la distribución límite, se tiene que

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} -M(\theta)^{-1} \cdot N[0, M(\theta)]. \tag{3.84}$$

o

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N[0, M(\theta)^{-1}] \tag{3.85}$$

Este resultado puede ser extendido al caso multivariado. Considerando las condiciones de regularidad bajo una versión multivariada

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N[0, -I(\theta)^{-1}] \tag{3.86}$$

donde $I(\theta)$ conocida como matriz información, que es fácilmente entendida cuando se escribe en detalle

$$I(\theta) = -E \begin{pmatrix} \partial^2 l(\theta)/\partial\theta_1^2 & \cdots & \partial^2 l(\theta)/\partial\theta_1\partial\theta_k \\ \vdots & \ddots & \vdots \\ \partial^2 l(\theta)/\partial\theta_k\partial\theta_1 & \cdots & \partial^2 l(\theta)/\partial\theta_K^2 \end{pmatrix}. \tag{3.87}$$

Así, cada elemento de $-I(\theta)$ es el valor esperado de una derivada parcial de segundo orden o derivada parcial cruzada del logaritmo de la función de verosimilitud con respecto a los parámetros.

Suponga que se tienen dos estimadores $\hat{\theta}$ y $\tilde{\theta}$ de un parámetro θ tal que $\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \sigma_1^2)$ y $\sqrt{T}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \sigma_2^2)$. Si $\sigma_2^2 \geq \sigma_1^2$, entonces $\hat{\theta}$ es asintóti-

camente eficiente relativo a $\tilde{\theta}$. Si θ es un vector de parámetros estimados por $\tilde{\theta}$ y $\hat{\theta}$ son estimadores consistentes tales que

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N(0, \Sigma) \tag{3.88}$$

y

$$\sqrt{T}(\tilde{\theta} - \theta) \xrightarrow{d} N(0, \Omega). \tag{3.89}$$

Entonces $\hat{\theta}$ es asintóticamente eficiente relativo a $\tilde{\theta}$ si $\Omega - \Sigma$ es positiva definida.

Como en el caso de eficiencia de muestras finitas, mostramos que un estimador es asintóticamente eficiente relativo a cualquier otro estimador consistente, requiere el conocimiento de la distribución de origen, de manera que la matriz de información y el límite asintótico de la cota inferior de Cramer-Rao pueden ser establecidos.

Deseamos que $\hat{\theta}$ sea un estimador consistente de θ de tal forma que $\sqrt{T}(\hat{\theta}) \xrightarrow{d} N(0, \Sigma)$. Entonces el estimador $\hat{\theta}$ es asintóticamente eficiente si

$$\Sigma = \lim_{T \to \infty} \left[\frac{1}{T} I(\theta) \right]^{-1} \tag{3.90}$$

donde $I(\theta)$ es la matriz de información. Así ambos muestras pequeñas y eficiencia asintótica son establecidos usando la desigualdad de Cramer-Rao como punto de partida.

Bajo algunas estrictas condiciones de regularidad los estimadores de máxima verosimilitud son: consistentes, asintóticamente normales, asintóticamente insesga-

dos, y asintóticamente eficientes. Esto es, si $\hat{\theta}$ es el estimador máximo verosímil del vector de parámetros θ, entonces

$$\sqrt{T}(\hat{\theta} - \theta) \xrightarrow{d} N\left(0, \lim_{T\to\infty}\left[\frac{1}{T}I(\theta)\right]^{-1}\right). \tag{3.91}$$

Aún cuando no se conoce la distribución muestral en muetras pequeñas de un estimador se puede confiar de hecho en que es asintóticamente normal para probar hipótesis o hacer expresiones de intervalos de confianza.

CAPITULO 4

EL ALGORITMO DE GAUSS-NEWTON

Y EL ALGORITMO DE NEWTON-RAPHSON

La no linealidad entra en los modelos económicos en varias formas. Si solamente las variables entran en forma no lineal, el modelo tadavía puede ser manejado dentro del proyecto de los modelos de regresión lineal. Si la no linealidad es en los parámetros o en las variables y en los parámetros, es algunas veces posible encontrar una transformación para convertir el modelo en una especificación lineal. Generalmente esto no es posible, y por lo tanto se discute un modelo de la forma general

$$y_t = f(\mathbf{x}_t, \beta) + e_t \tag{4.1}$$

siendo x_t un vector $(N \times 1)$ de variables independientes, β es un vector $(k \times 1)$ de parámetros, y_t es la variable dependiente cuya media es una función de x_t y β, y e_t es un error aleatorio. En los modelos no lineales a menos que la especificación sea lineal, el número de parámetros y el número de variables independientes no necesariamente coincidirán.

Asi como en el caso lineal, la estimación está basada en optimizar un función objetivo. Para el caso no lineal se tienen dos tipos de funciones objetivo: la suma de cuadrados del error y la función de verosimilitud. Bajo supuestos estándar la minimización o maximización (optimización) puede ser llevada a cabo resolviendo

una colección de ecuaciones normales, las cuales para el modelo no lineal, son en general no lineales en los parámetros y resolverlas puede ser una tarea un tanto difícil.

Suponga se quiere encontrar las ecuaciones normales para obtener estimaciones $\hat{\beta}$ de β para el modelo

$$y_t = \beta_1 x_{t1} + \beta_1^2 x_{t2} + e_t$$

donde, e_t son variables aleatorias distribuidas independiente e idénticamente con media cero y varianza σ^2 y donde están disponibles T pares de observaciones. Encontramos que la estimación mínimo cuadrática para β_1 es aquel valor que minimiza la suma de cuadrados de los residuales.

$$S(\beta) = \sum_{t=1}^{T} e_t^2 = \sum_{t=1}^{T} [y_t - f(\mathbf{x}_t, \beta)]^2 = \sum_{t=1}^{T} [y_t - \beta_1 x_{t1} - \beta_1^2 x_{t2}]^2$$

y de acuerdo con el criterio de la primera derivada para encontrar puntos estacionarios que sean un mínimo tenemos

$$\frac{\mathrm{dS}}{\mathrm{d}\beta} = 2\sum_{t=1}^{T} [y_t - f(\mathbf{x}_t, \beta)] \frac{df(\mathbf{x}_t, \beta)}{d\beta} = 2\sum_{t=1}^{T} [(y_t - \beta_1 x_{t1} - \beta_1^2 x_{t2})(-x_{t1} - 2\beta_1 x_{t2})]$$

$$= -\sum_{t=1}^{T} y_t x_{t1} - 2\sum_{t=1}^{T} y_t \beta_1 x_{t2} + \sum_{t=1}^{T} \beta_1 x_{t1}^2 + 2\sum_{t=1}^{T} \beta_1^2 x_{t1} x_{t2} + \sum_{t=1}^{T} \beta_1^2 x_{t2} x_{t1} + 2\sum_{t=1}^{T} \beta_1^3 x_{t2}^2$$

$$= 2\beta_1^3 \sum_{t=1}^{T} x_{t2}^2 + 3\beta_1^2 \sum_{t=1}^{T} x_{t1} x_{t2} + \beta_1 (\sum_{t=1}^{T} x_{t1}^2 - 2\sum_{t=1}^{T} x_{t2} y_t) - \sum_{t=1}^{T} x_{t1} y_t = 0.$$

Esta expresión es una ecuación cúbica en β_1 y resultarán, por lo tanto, tres puntos estacionarios o tres posibles soluciones. La estimación no lineal mínimo cuadrática se interesa en aquella solución que produce el valor mas bajo para la suma de cuadrados de los residuales $S(\beta)$, es decir, se interesa en el mínimo global ya que pudieran existir otros mínimos pero estos serían locales e incluso pudiera existir hasta una máximo. En general, cualquier función no lineal posee múltiples máximos y mínimos locales y no existe una forma fácil de saber si el límite es un mínimo local o global.

Una expresión analítica para obtener el estimador de β para el problema de regresión no lineal, como la obtenida para el modelo de regresión lineal

$$\frac{\sum(X_i - \overline{x})(Y_i - \overline{y})}{\sum(X_i - \overline{x})^2}$$

no es directa e incluso hasta es computacionalmente difícil. Por lo que deben emplearse métodos iterativos o métodos numéricos en casi todos los modelos de regresión no lineal.

Las expresiones analíticas de las derivadas de segundo o primer orden de la función objetivo, que por supuesto deben existir, son requeridas en algunos algoritmos empleados para la estimación de parámetros. En todos ellos el punto del espacio paramétrico en el que se alcanza la convergencia (si es que el algoritmo converge) depende de la forma de las ecuaciones y de las condiciones iniciales.

Aquí solamente se describirán dos de ellos el método de Gauss-Newton y el método de Newton-Raphson. Otros métodos de estimación no lineal están diponibles y pueden proveer también estimaciones convergentes. Sin embargo, no hay realmente un mejor método para cualquier problema ya que puede ser que uno converja mas fácilmente mientras que otro implica un mayor gasto computacional. A menudo métodos alternativos serán usados como una forma de revisar si un mí-

nimo global de la suma de cuadrados de los residuales es el que se ha encontrado.

El Algoritmo de Gauss-Newton

Se iniciará considerando primero el algoritmo de Gauss-Newton. Suponga un modelo con un único parámetro como el expresado en la ecuación (4.1)

$$y_t = f(\mathbf{x}_t, \beta) + e_t$$

con la correspondiente función de la suma de cuadrados de los residuales

$$S(\beta) = \sum_{t=1}^{T} [y_t - f(\mathbf{x}_t, \beta)]^2 \tag{4.2}$$

y con una condición de primer orden para un mínimo dada por la siguiente expresión

$$\frac{\mathrm{dS}}{\mathrm{d}\beta} = 2\sum_{t=1}^{T} [y_t - f(\mathbf{x}_t, \beta)]\frac{df(\mathbf{x}_t, \beta)}{d\beta} = 0 \tag{4.3}$$

el problema es encontrar un valor de β que satisfaga la ecuación y que también conduzca a un mínimo global de la suma de cuadrados de los residuales.

Una forma de abordar este problema es reemplazando $f(\mathbf{x}_t, \beta)$ por una aproximación usando una serie de Taylor de primer orden. Si se comienza en algún punto β_1, entonces la aproximación de primer orden de $f(\mathbf{x}_t, \beta)$ alrededor del punto β_1 es dado por

$$f(\mathbf{x}_t, \beta) \simeq f(\mathbf{x}_t, \beta_1) + \left.\frac{\mathrm{df}(\mathbf{x}_t, \beta)}{\mathrm{d}\beta}\right|_{\beta_1} (\beta - \beta_1) \tag{4.4}$$

donde puede representarse la pendiente de la tangente de la curva $f(\mathbf{x}_t, \beta)$ en el punto β_1 como

$$\left.\frac{\mathrm{df}(\mathbf{x}_t, \beta)}{\mathrm{d}\beta}\right|_{\beta_1} \simeq \frac{\mathrm{f}(\mathbf{x}_t, \beta) - \mathrm{f}(\mathbf{x}_t, \beta_1)}{\beta - \beta_1} \tag{4.5}$$

e introduciendo una notación menos incomoda para la derivada de $f(\mathbf{x}_t, \beta)$ con respecto a β evaluada en el punto β_1 se tendrá

$$z_t(\beta_1) = \left.\frac{\mathrm{df}(\mathbf{x}_t, \beta)}{\mathrm{d}\beta}\right|_{\beta_1}. \tag{4.6}$$

Usando esta notación y sustituyendola en la aproximación de la serie de Taylor dentro de la función de la suma de cuadrados de los residuales resulta

$$S(\beta) = \sum_{t=1}^{T} [y_t - f(\mathbf{x}_t, \beta_1) - z_t(\beta_1)(\beta - \beta_1)]^2 = \sum_{t=1}^{T} [\overline{y_t}(\beta_1) - z_t(\beta_1)\beta]^2 \tag{4.7}$$

donde

$$\overline{y_t}(\beta_1) = y_t - f(\mathbf{x}_t, \beta_1) + z_t(\beta_1)\,\beta_1.$$

Dado un valor de β_1, ambos $\overline{y_t}(\beta_1)$ y $z_t(\beta_1)$ son observables. Así, la suma de cuadrados $\mathrm{S}(\beta)$ de los residuales puede ser vista como aquella que se necesita minimizar para encontrar la estimación mínimo cuadrática de β en el modelo lineal

$$\overline{y_t}(\beta_1) = z_t(\beta_1)\beta + e_t. \tag{4.8}$$

La estimación mínimo cuadrática para este modelo está dada por

$$\beta_2 = \frac{\sum_{t=1}^{T} \overline{y_t}(\beta_1) z_t(\beta_1)}{\sum_{t=1}^{T} z_t(\beta_1)^2} = [\mathbf{z}(\beta_1)'\mathbf{z}(\beta_1)]^{-1}\mathbf{z}(\beta_1)'\overline{\mathbf{y}}(\beta_1) \tag{4.9}$$

donde

$$\mathbf{z}(\beta_1) = \begin{pmatrix} z_1(\beta_1) \\ z_2(\beta_1) \\ . \\ . \\ . \\ z_T(\beta_1) \end{pmatrix} \quad \text{y} \quad \overline{\mathbf{y}}_t(\beta_1) = \begin{pmatrix} \overline{y}_2(\beta_1) \\ \overline{y}_2(\beta_1) \\ . \\ . \\ . \\ \overline{y}_T(\beta_1) \end{pmatrix}.$$

De manera que, si se aproxima $f(\mathbf{x}_t, \beta)$ por una serie de Taylor de primer orden alrededor del valor inicial β_1 de β, un segundo valor o estimación de β, llamado β_2 puede ser encontrado aplicando mínimos cuadrados al nuevo modelo linearizado. Este proceso puede repetirse usando β_2 para construir el nuevo modelo

$$\overline{\mathbf{y}}(\beta_2) = \mathbf{z}(\beta_2)\beta + \mathbf{e} \tag{4.10}$$

que produce la estimación mínimo cuadrática

$$\beta_3 = [\mathbf{z}(\beta_2)'\mathbf{z}(\beta_2)]^{-1}\mathbf{z}(\beta_2)'\overline{\mathbf{y}}(\beta_2). \tag{4.11}$$

El proceso continua y se escribe la estimación $(n+1)-$ésima como una función de la n-ésima estimación como sigue

$$\beta_{n+1} = [\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\mathbf{z}(\beta_n)'\overline{\mathbf{y}}(\beta_n)$$

$$\beta_{n+1} = [\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\mathbf{z}(\beta_n)'[\mathbf{y} - f(\mathbf{X}, \beta_n) + \mathbf{z}(\beta_n)\beta_n]$$

$$= \beta_n + [\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\mathbf{z}(\beta_n)'[\mathbf{y} - \mathbf{f}(X, \beta_n)] \tag{4.12}$$

donde $f(\mathbf{X}, \beta) = [f(\mathbf{x}_1, \beta), f(\mathbf{x}_2, \beta), ..., f(\mathbf{x}_T, \beta)]'$.

También la condición de primer orden para un mínimo $dS/d\beta$ puede ser escrita en forma matricial como

$$\mathbf{z}(\beta_n)'[\mathbf{y} - f(\mathbf{X}, \beta_n)] = 0. \tag{4.13}$$

Si dos estimaciónes sucesivas son iguales, $\beta_{n+1} = \beta_n$, se sigue de la expresión (4.12) $\beta_{n+1} = \beta_n + [\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\mathbf{z}(\beta_n)'[\mathbf{y} - f(\mathbf{X}, \beta_n)]$ que $\mathbf{z}(\beta_n)'[\mathbf{y} - f(\mathbf{X}, \beta_n)] = 0$ se cumple y por lo tanto, β_n satisface la condición necesaria para un mínimo.

En resumen, se comienza con un valor β_1, y repetidamente se aplica la fórmula

$$\beta_{n+1} = \beta_n + [\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\mathbf{z}(\beta_n)'[\mathbf{y} - f(\mathbf{X}, \beta_n)]$$

hasta que la convergencia ocurra, es decir, hasta que se alcanza un punto que es una solución a la condición de primer orden $\mathbf{z}(\beta_n)'[\mathbf{y} - f(\mathbf{X}, \beta_n)] = 0$. Sin embargo, el algoritmo de Gauss-Newton podría conducirnos a un mínimo global, un mínimo local o un máximo. Para protegernos contra esta posibilidad se comienza todo el proceso con diferentes valores iniciales (β_1). Si estos nos llevan a diferentes puntos, la estimación mínimo cuadrática será aquel mínimo que produzca la menor suma de cuadrados del error.

Por otro lado, la derivada de la función de la suma de cuadrados de los residuales

$$\frac{dS}{d\beta} = -2\sum_{t=1}^{T}[y_t - f(\mathbf{x}_t, \beta)]\left(\frac{df(\mathbf{x}_t, \beta)}{d\beta}\right) = 0$$

también puede ser escrita como

$$\frac{\mathrm{dS}}{\mathrm{d}\beta} = -2\mathbf{z}(\beta)'[\mathbf{y} - f(\mathbf{X}, \beta)]$$

y, por tanto,

$$\beta_{n+1} = \beta_n + [\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\mathbf{z}(\beta_n)'[\mathbf{y} - f(\mathbf{X}, \beta_n)]$$

puede ser escrita como

$$\beta_{n+1} = \beta_n - \frac{1}{2}[\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_n}. \tag{4.14}$$

Es posible que el cambio en β pueda sobrepasar el valor mínimo, o al menos tomar un largo tiempo en alcanzar el mínimo. Para protección contra esta posibilidad el algoritmo en esta última expresión es usualmente hecho un poco mas sofisticado mediante la introducción de una variable conocida como longitud de salto t_n. Esto conduce a un algoritmo definido por

$$\beta_{n+1} = \beta_n - t_n[\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\frac{\mathrm{ds}}{\mathrm{d}\beta}\Big|_{\beta_n} \tag{4.15}$$

o

$$\beta_{n+1} = \beta_n + 2t_n[\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\mathbf{z}(\beta_n)'[\mathbf{y} - f(\mathbf{X}, \beta_n)]. \tag{4.16}$$

El caso de un modelo con un solo parámetro puede ser generalizado a un modelo con más de un parámetro. Supongamos un modelo $y_t = f(\mathbf{x}_t, \beta) + e_t$

con $(K \times 1)$ parámetros desconocidos incluidos en un vector β. Usando notación matricial se tiene

$$\mathbf{y} = f(\mathbf{X}, \beta) + \mathbf{e} \tag{4.17}$$

se supone que $E[\mathbf{e}] = \mathbf{0}$ y $E[\mathbf{e}'\mathbf{e}] = \sigma^2\mathbf{I}$. La estimación mínimo cuadrática del vector β es el valor que minimiza la suma de cuadrados de los residuales

$$S(\beta) = (\mathbf{e}'\mathbf{e}) = [\mathbf{y} - f(\mathbf{X}, \beta)]'[\mathbf{y} - f(\mathbf{X}, \beta)]. \tag{4.18}$$

En este caso hay K condiciones de primer orden para un mínimo, definido por la colección del vector K-dimensional de derivadas de $dS/d\beta$ igual al vector $\mathbf{0}$. Estas condiciones de primer orden están dadas por

$$\frac{\partial \mathrm{S}}{\partial \beta} = -2\frac{\partial \mathrm{f}(\mathbf{X}, \beta)'}{\partial \beta}[\mathbf{y} - f(\mathbf{X}, \beta)] = \mathbf{0} \tag{4.19}$$

donde $\partial f(\mathbf{X}, \beta)'/\partial\beta$ es una matriz de derivadas $(K \times T)$ con el (k, t)-ésimo elemento dado por $\partial f(x_t, \beta)'/\partial\beta_k$.

Si se establece que $\mathbf{Z}(\beta)$ denota la transpuesta de la matriz $\partial f(\mathbf{X}, \beta)'/\partial\beta$. Esto es

$$\mathbf{Z}(\beta) = \frac{\partial \mathrm{f}(\mathbf{X}, \beta)}{\partial \beta'} = \begin{pmatrix} \partial f(\mathbf{x}_1, \beta)/\partial\beta_1 & . & . & . & \partial f(\mathbf{x}_1, \beta)/\partial\beta_K \\ & . & & . & \\ & . & & . & \\ & . & & . & \\ \partial f(\mathbf{x}_T, \beta)/\partial\beta_1 & . & . & . & \partial f(\mathbf{x}_T, \beta)/\partial\beta_K \end{pmatrix}. \tag{4.20}$$

Cuando esta matriz de derivadas es evaluada en un valor particular para β, digamos β_1, esto será escrito como $\mathbf{Z}(\beta_1)$. La condición de primer orden puede ser escrita como

$$\mathbf{Z}(\beta)'[\mathbf{y} - f(\mathbf{X}, \beta)] = \mathbf{0}. \tag{4.21}$$

Para encontrar una solución a esta última expresión se inicia una aproximación de $f(\mathbf{X}, \beta)$ a través de una serie de Taylor de primer orden alrededor de un punto inicial β_1. La aproximación para la t-ésima observación es dada por

$$f(\mathbf{x}_t, \beta) \simeq f(\mathbf{x}_t, \beta_1) + \left(\left.\frac{df(\mathbf{x}_t, \beta)}{d\beta_1}\right|_{\beta_1} \quad \cdots \quad \left.\frac{df(\mathbf{x}_t, \beta)}{d\beta_K}\right|_{\beta_1} \right)(\beta - \beta_1) \tag{4.22}$$

e incluyendo todas las T observaciones resulta

$$\mathbf{f}(\mathbf{X}, \beta) \simeq \mathbf{f}(\mathbf{X}, \beta_1) + \mathbf{Z}(\beta_1)(\beta - \beta_1).$$

Sustituyendo esta última expresión en la ecuación (4.17), $\mathbf{y} = f(\mathbf{X}, \beta) + \mathbf{e}$, resulta

$$\mathbf{y} \simeq f(\mathbf{X}, \beta_1) + \mathbf{Z}(\beta_1)(\beta - \beta_1) + \mathbf{e} \tag{4.23}$$

que puede ser expresada como

$$\overline{\mathbf{y}}(\beta_1) \simeq \mathbf{Z}(\beta_1)\beta + \mathbf{e}$$

donde

$$\overline{\mathbf{y}}(\beta_1) \simeq \mathbf{y} - f(\mathbf{X}, \beta_1) + \mathbf{Z}(\beta_1)\beta_1.$$

La estimación mínimo cuadrática para el modelo $\overline{\mathbf{y}}(\beta_1) \simeq \mathbf{Z}(\beta_1)\beta + \mathbf{e}$ provee una segunda estimación para β, o sea,

$$\beta_2 = [\mathbf{Z}(\beta_1)'\mathbf{Z}(\beta_1)]^{-1}\mathbf{Z}(\beta_1)'\overline{\mathbf{y}}(\beta_1)]$$

$$= \beta_1 + [\mathbf{Z}(\beta_1)'\mathbf{Z}(\beta_1)]^{-1}\mathbf{Z}(\beta_1)'[\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta_1)] \tag{4.24}$$

y continuando este proceso, la n-ésima iteración del algoritmo de Gauss-Newton es dada por

$$\beta_{n+1} = \beta_n + [\mathbf{Z}(\beta_n)'\mathbf{Z}(\beta_n)]^{-1}\mathbf{Z}(\beta_n)'[\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta_n)]. \tag{4.25}$$

Cuando el proceso converge en $\beta_{n+1} = \beta_n$ la condición de primer orden para un mínimo $\mathbf{Z}(\beta)'[\mathbf{y} - \mathbf{f}(\mathbf{X}, \beta)] = \mathbf{0}$ debe ser satisfecha. En estas circunstancias el punto β_n podría corresponder a un mínimo local o a un mínimo global. Este no podría ser un máximo debido a que $[Z(\beta_n)'Z(\beta_n)]^{-1}$ es definida positiva, asegurando que el cambio $\beta_{n+1} - \beta_n$ siempre será en la dirección correcta. Aunque no existe seguridad de que ha sido encontrado un mínimo global, la oportunidad de equivocación puede ser reducida intentando un número de valores diferentes para el valor inicial β_1.

El Algoritmo de Newton-Raphson

Para introducir el algoritmo de Newton-Raphson se supone un modelo con un único parámetro idéntico al expresado en 4.17,

$$\mathbf{y} = f(\mathbf{X}, \beta) + \mathbf{e} \tag{4.26}$$

donde, nuevamente se supone que $E[\mathbf{e}] = \mathbf{0}$ y $E[\mathbf{e}\mathbf{e}'] = \sigma^2\mathbf{I}$. La correspondiente suma de cuadrados de los residuales viene dada por

$$S(\beta) = [\mathbf{y} - f(\mathbf{X}, \beta)]'[\mathbf{y} - f(\mathbf{X}, \beta)] = \sum_{t=1}^{T}[y_t - f(\mathbf{x}_t, \beta)]^2. \tag{4.27}$$

Cuando se considera el algoritmo de Gauss-Newton, se inicia remplazando $f(\mathbf{x}_t, \beta)$ con una aproximación de primer orden mediante series de Taylor alrededor del punto inicial β_1. Ahora en algoritmo de Newton-Raphson se comienza reemplazando $S(\beta)$ con una aproximación de segundo orden mediante series de Taylor, o sea

$$S(\beta) \simeq S(\beta_1) + \frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_1}(\beta - \beta_1) + \frac{1}{2}\frac{\mathrm{d}^2\mathrm{S}}{\mathrm{d}\beta^2}\Big|_{\beta_1}(\beta - \beta_1)^2. \tag{4.28}$$

El objetivo de nueva cuenta, de acuerdo al problema de minimización, es encontrar el valor de β que minimice $S(\beta)$ usando la notación

$$h(\beta_1) = \frac{\mathrm{d}^2\mathrm{S}}{\mathrm{d}\beta^2}\Big|_{\beta_1} \tag{4.29}$$

se tiene que

$$\frac{\mathrm{dS}}{\mathrm{d}\beta} \simeq \frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_1} + h(\beta_1)(\beta - \beta_1) \tag{4.30}$$

e igualando esto último a cero y resolviendo para β se encuentra un segundo valor de β, es decir β_2, que esta dado por

$$\beta_2 = \beta_1 - h(\beta_1)^{-1}\frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_1}. \tag{4.31}$$

Si $S(\beta)$ es cuadrática, entonces β_2 será entonces exactamente la estimación mínimo cuadrática. Pero en el caso de que $S(\beta)$ no sea cuadrática, como es usual, β_2 entonces no será el valor mínimo debido a que

$$\frac{\mathrm{dS}}{\mathrm{d}\beta} \simeq \frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_1} + h(\beta_1)(\beta - \beta_1)$$

es solamente una aproximación. El procedimiento conduce a un valor (n+1)-ésimo para β dado por

$$\beta_{n+1} = \beta_n - h(\beta_n)^{-1} \frac{\mathrm{dS}}{\mathrm{d}\beta}\bigg|_{\beta_n}. \tag{4.32}$$

Si el proceso converge en el sentido de que $\beta_{n+1} = \beta_n$, entonces debe ser cierto que $dS/d\beta|_{\beta_n} = 0$, que es la condición necesaria para un mínimo (o un máximo).

El algoritmo conduce a la dirección correcta, es decir hacia un mínimo, del punto β_1 si la segunda derivada $h(\beta_1)$ siempre es positiva en la vecindad de un mínimo, se ira en la correcta dirección si β_1 es suficientemente cercano al valor mínimo. No obstante, es posible sobrepasarlo. Para protección contra esto, la variable t_n, conocida como longitud de salto, puede ser introducida, quedando expresado ahora el algoritmo como

$$\beta_{n+1} = \beta_n - t_n h(\beta_n)^{-1} \frac{\mathrm{dS}}{\mathrm{d}\beta}\bigg|_{\beta_n}. \tag{4.33}$$

Cada iteración t_n es encontrada de tal forma que $S(\beta_{n+1}) < S(\beta_n)$. Si se comienza en un punto (β_1) que es cercano a un máximo en el sentido de que $h(\beta_1)$ es negativa , entonces $\beta_{n+1} = \beta_n - h(\beta_n)^{-1} \frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_n}$ nos conduce a una incorrecta dirección, es decir hacia a un máximo. De nueva cuenta, el criterio para establecer si se enfrenta a un mínimo local o global deberá ser probando diferentes valores iniciales.

En el caso general de K parámetros, se supone que el modelo general es de la forma

$$\mathbf{y} = f(\mathbf{X}, \beta) + \mathbf{e} \tag{4.34}$$

donde β es un vector k-dimensional de parámetros desconocidos y donde además se supone que $E[\mathbf{e}] = \mathbf{0}$ y $E[\mathbf{ee}'] = \sigma^2\mathbf{I}$, la n-ésima iteración del algoritmo de Newton-Raphson designado para encontrar el valor de β que minimiza $S(\beta) = (\mathbf{e}'\mathbf{e})$ es dado por

$$\beta_{n+1} = \beta_1 - H_n^{-1}\frac{\partial \mathrm{S}}{\partial \beta}\Big|_{\beta_n} \tag{4.35}$$

donde

$$\frac{\partial \mathrm{S}}{\partial \beta}\Big|_{\beta_n} = \left(\frac{\partial S}{\beta_1}, \frac{\partial S}{\beta_2}, ..., \frac{\partial S}{\beta_K}\right)'\Big|_{\beta_n} \tag{4.36}$$

es el vector gradiente evaluado en β_n, y H_n es la matriz Hessiana $(K \times K)$ evaluada en β_n. Esto es

$$H_n = \frac{\partial^2 \mathrm{S}}{\partial\beta\partial\beta'}\Big|_{\beta_n} = \begin{pmatrix} \partial^2 S/\partial\beta_1^2 & \cdots & \partial^2 S/\partial\beta_1\partial\beta_K \\ \vdots & & \vdots \\ \partial^2 S/\partial\beta_K\partial\beta_1 & \cdots & \partial^2 S/\partial\beta_K^2 \end{pmatrix}\Big|_{\beta_n}. \tag{4.37}$$

La ecuación $\beta_{n+1} = \beta_1 - H_n^{-1}\frac{\partial \mathrm{S}}{\partial \beta}\Big|_{\beta_n}$ puede ser derivada encontrando el valor de β que minimiza $S(\beta)$ que ha sido aproximada por una K-aproximación de una serie de Taylor de segundo orden alrdedor del valor predeterminado β_n. Si H_n es positiva definida, entonces el cambio de $\beta_{(n+1)} - \beta_{(n)}$ será en la dirección correcta (hacia un mínimo). Si β_n no está suficientemente cercano a un mínimo, H_n no podra ser positiva definida y el algoritmo se dirigirá hacia una dirección equivocada. También, si un mínimo es alcanzado, este podrá ser un mínimo local más que un mínimo global.

La comparación de los dos algoritmos vistos aquí puede hacerse de la forma siguiente: el algoritmo de Gauss-Newton fue definido como

$$\beta_{n+1} = \beta_n - \frac{1}{2}[\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1}\frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_n} \tag{4.38}$$

mientras que el Newton-Raphson fue definido como

$$\beta_{n+1} = \beta_n - h(\beta_n)^{-1}\frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_n} \tag{4.39}$$

observándose que ambos son de la forma

$$\beta_{n+1} = \beta_n - p_n\frac{\mathrm{dS}}{\mathrm{d}\beta}\Big|_{\beta_n} \tag{4.40}$$

donde

$$p_n = \begin{cases} \frac{1}{2}[\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n)]^{-1} \\ h(\beta_n)^{-1} \end{cases} \tag{4.41}$$

para Gauss-Newton y para Newton-Raphson, respectivamente.

Recordando las definiciones $Z(\beta)$ y $h(\beta)$, tenemos

$$\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n) = \sum_{t=1}^{T}\left(\frac{df(\mathbf{x}_t,\beta)}{d\beta}\right)^2$$

y

$$h(\beta) = \frac{\mathrm{d}^2\mathrm{S}}{\mathrm{d}\beta^2} = \frac{\mathrm{d}^2}{\mathrm{d}\beta^2}\Big[\sum_{t=1}^{T}[y_t - f(\mathbf{x}_t,\beta)]^2\Big] = \frac{\mathrm{d}}{\mathrm{d}\beta}\Big[-2\sum_{t=1}^{T}[y_t - f(\mathbf{x}_t,\beta)]\frac{df(\mathbf{x}_t,\beta)}{d\beta}\Big]$$

$$= 2\mathbf{z}(\beta_n)'\mathbf{z}(\beta_n) - 2\sum_{t=1}^{T}[y_t - f(\mathbf{x}_t,\beta)]\frac{d^2f(\mathbf{x}_t,\beta)}{d\beta^2} \tag{4.42}$$

por lo que los dos algoritmos son idénticos excepto por el segundo término de la última línea. Ya que $E[y_t] = f(x_t, \beta)$, este término tiene una esperanza igual a cero. Esto es,

$$E\left[\frac{1}{2}h(\beta)\right] = \left[\frac{1}{2}\frac{\mathrm{d}^2\mathrm{S}}{\mathrm{d}\beta^2}\right] = \mathbf{z}(\beta_n)'\mathbf{z}(\beta_n). \tag{4.43}$$

Los algoritmos de Gauss-Newton y de Newton-Raphson son dos de un gran número de posibles algoritmos. Muchos tienen la forma

$$\beta_{n+1} = \beta_n - t_n P_n \gamma_n \tag{4.44}$$

donde $\gamma_n = dS/d\beta|_{\beta_n}$ es el vector gradiente, P_n es deseable sea una matriz positiva definida conocida como la matriz dirección, y t_n es un número positivo conocido como la longuitud de salto. Muchos algoritmos incluyen algún procedimiento para determinar una longuitud de salto óptima en cada iteración. La característica que distingue los algoritmos alternativos es la definición de P_n.

Un modelo intrínsecamente no lineal y que no permite la utilización de logaritmos es la función de producción de elasticidad de substitución constante (CES). El empleo del algoritmo de Gauss-Newton en un modelo de este tipo con término de error aditivo

$$Q_t = \beta_1 L_t^{\beta_2} K_t^{\beta_3} + e_t \tag{4.45}$$

distribuido independiente e idénticamente con media cero y varianza σ^2 requiere: la matriz de primeras derivadas $\mathbf{z}(\beta) = \partial f(x_t, \beta)/\partial\beta'$, cuyo t^{th} elemento es dado por

$$f(x_t, \beta) = \beta_1 L_t^{\beta_2} K_t^{\beta_3} + e_t \tag{4.46}$$

donde

$$x_t' = (L_t, K_t), \qquad \beta' = (\beta_1, \beta_2, \beta_3) \tag{4.47}$$

y el t^{th} renglón de $\mathbf{z}(\beta)$ es dado por

$$\frac{\partial f(x_t, \beta)}{\partial \beta'} = \left[\frac{\partial f(x_t, \beta)}{\partial \beta_1}, \frac{\partial f(x_t, \beta)}{\partial \beta_2}, \frac{\partial f(x_t, \beta)}{\partial \beta_3}\right]$$

$$= \left[L_t^{\beta_2} K_t^{\beta_3}, (\ln L_t)\beta_1 L_t^{\beta_2} K_t^{\beta_3}, (\ln K_t)\beta_1 L_t^{\beta_2} K_t^{\beta_3}\right]. \tag{4.48}$$

El procedimiento para poder alcanzar un mínimo global a través de la aplicación del método de Gauss-Newton utiliza la matriz $\mathbf{z}(\beta)'\mathbf{z}(\beta)$ escrita como

$$\mathbf{z}(\beta)'\mathbf{z}(\beta) =$$

$$\begin{pmatrix} \sum_{t=1}^T [LK]^2 & \beta_1 \sum_{t=1}^T [LK]^2 \ln L_t & \beta_1 \sum_{t=1}^T [LK]^2 \ln K_t \\ \beta_1 \sum_{t=1}^T [LK]^2 \ln L_t & \beta_1^2 \sum_{t=1}^T [LK]^2 [\ln L_t]^2 & \beta_1^2 \sum_{t=1}^T [LK]^2 [\ln L_t][\ln K_t] \\ \beta_1 \sum_{t=1}^T [LK]^2 \ln K_t & \beta_1^2 \sum_{t=1}^T [LK]^2 [\ln L_t][\ln K_t] & \beta_1^2 \sum_{t=1}^T [LK]^2 [\ln K_t]^2 \end{pmatrix} \tag{4.49}$$

donde, $[L_t^{\beta_2} K_t^{\beta_3}]^2 = [LK]$; y el vector $\mathbf{z}(\beta)'[y - f(x_t, \beta)]$ está dado por

$$\mathbf{z}(\beta)'[y - f(x_t, \beta)] = \begin{pmatrix} \sum_{t=1}^T [L_t^{\beta_2} K_t^{\beta_3}][y_t - \beta_1 L_t^{\beta_2} K_t^{\beta_3}] \\ \beta_1 \sum_{t=1}^T \ln L_t [L_t^{\beta_2} K_t^{\beta_3}][y_t - \beta_1 L_t^{\beta_2} K_t^{\beta_3}] \\ \beta_1 \sum_{t=1}^T \ln K_t [L_t^{\beta_2} K_t^{\beta_3}][y_t - \beta_1 L_t^{\beta_2} K_t^{\beta_3}] \end{pmatrix} \tag{4.50}$$

el mecanismo que habrá de minimizar la suma de cuadrados del error ya ha sido descrito renglones arriba.

CAPITULO 5

LA ESPECIFICACION DE LA RELACION FUNCIONAL ENTRE VARIABLES

Un problema frecuentemente es investigar la relación entre una variable dependiente y y algún conjunto de variables independientes $X_1, X_2, ..., X_k$, es decir, tratar de especificar la relación funcional entre las variables. Aunque usualmente información *A priori* sobre las variables que deben incluirse en una relación particular suele estar disponible, existe poca información sobre la forma funcional precisa. En este capítulo se referirá a la forma funcional del modelo de regresión. Teniendo en mente la existencia de diferentes tipos de funciones, se hará uso de un mecanismo particular de transformación, para generar una gran variedad de modelos. La discusión estará confinada a unos pocos casos de interés. Se analizarán las características especiales de cada modelo, los casos en los cuales su uso es apropiado y la forma como estos son estimados.

El Artificio de Transformación de Box y Cox

La transformación de Box y Cox es un artificio útil para determinar la forma funcional apropiada. Para ser más específicos, en la relación de dos variables considere la siguientes transformaciones de y por $y^{(\lambda)}$ y X por $X^{(\lambda)}$, donde λ es el único parámetro del cual depende la transformación de las variables así pues

$$y^{(\lambda)} = \begin{cases} (y^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \ln y & \lambda = 0 \end{cases} \tag{5.1}$$

y similarmente

$$X^{(\lambda)} = \begin{cases} (X^{\lambda} - 1)/\lambda & \lambda \neq 0 \\ \ln X & \lambda = 0 \end{cases} . \tag{5.2}$$

Considerando el caso de y: cuando $\lambda = 1$, $y^{(\lambda)} = y - 1$; cuando $\lambda = -1$, $y^{(\lambda)} = -y^{-1} + 1$; y cuando $\lambda = 0$, $y^{(\lambda)} = \ln y$. Originalmente la sustitución de λ por cero parece generar expresiones indeterminadas en , $(y^{\lambda} - 1)/\lambda$ y $(x^{\lambda} - 1)/\lambda$. Sin embargo, la aplicación de la regla de L'Hospital muestra que

$$\lim_{\lambda \to 0} y^{(\lambda)} = \lim_{\lambda \to 0} \frac{(d/d\lambda)(y^{\lambda} - 1)}{(d/d\lambda)(\lambda)} = \lim_{\lambda \to 0} (y^{\lambda} \ln y) = \ln y \tag{5.3}$$

lo mismo se cumpliría para X.

Hasta aquí las últimas expresiones han sido especificadas en términos de logaritmos de base e, sin embargo, se podría tomar logaritmos de base 10 en un trabajo empírico. En la práctica se puede utilizar logaritmos comunes, es decir, logaritmos con base 10. La relación entre el logaritmo natural y el logaritmo común es: $\ln_e X = 2.3026 = \log_{10} X$.

Otra forma de mostrar lo mismo, sería utilizando un número positivo, digamos Z, escrito como

$$Z = \exp(\log Z) \tag{5.4}$$

donde la base del logaritmo es e, y que $\exp(\log Z)$ puede ser expandido como

$$\exp(\log Z) = 1 + \log Z + \frac{1}{2!}(\log Z)^2 + \frac{1}{3!}(\log Z)^3 + ... \tag{5.5}$$

Por lo tanto, se tiene que

$$\frac{y_i^\lambda - 1}{\lambda} = \frac{1}{\lambda}\left[1 + \lambda \log y_i + \frac{1}{2!}(\lambda \log y_i)^2 + ... - 1\right]$$

$$= \log y_i + \frac{\lambda}{2!}(\lambda \log y_i)^2 + \frac{\lambda^2}{3!}(\lambda \log y_i)^3 + ... \tag{5.6}$$

Para $\lambda = 0$,

$$\frac{y_i^\lambda - 1}{\lambda} = \log Y_i \tag{5.7}$$

y, similarmente,

$$\frac{x_i^\lambda - 1}{\lambda} = \log X_i. \tag{5.8}$$

Ahora, considerando la transformación de variables en el siguiente modelo,

$$y^{\lambda_1} = \alpha_0 + \beta X^{\lambda_2} + e \tag{5.9}$$

que posee cinco parámetros básicos α_0, β, λ_1, λ_2, y σ^2. Este permitirá considerar algunos casos especiales correspondientes a valores particulares de λ_1 y λ_2. El primer modelo es el modelo lineal, este surge cuando $\lambda_1 = \lambda_2$ son iguales a 1, y combinando las ecuaciones arriba descritas se tiene

$$y = \alpha + \beta X + e \tag{5.10}$$

donde $\alpha = 1+\alpha_0-\beta$. Este modelo es ampliamente conocido y comúnmente estimado utilizando el procedimiento de mínimos cuadrados ordinarios, sin necesidad de que los datos hayan sido transformados.

Un segundo modelo denominado log-log, doble log, o log-lineal es originado cuando $\lambda_1 = \lambda_2$ son iguales a 0, por lo que tenemos el siguiente modelo

$$\ln y = \alpha_0 + \beta \ln X + e. \tag{5.11}$$

El método de mínimos cuadrados se aplica una vez los datos han sido transformados. La estimación de α_0 será efectado por la elección de la base del logaritmo empleado, pero β no lo será. Ignorando el término de perturbación la relación entre X y y es

$$y = \beta_1 X^{\beta_2} \tag{5.12}$$

donde, $\ln \beta_1 = \alpha_0$. Este modelo es conocido como el modelo de regresión exponencial.

Una característica del modelo log-log que lo ha hecho muy popular, es que el coeficiente de la pendiente β mide la elasticidad de y con respecto a X, es decir, el cambio porcentual en y ante un cambio porcentual en X. Además, en el modelo log-log un cambio en $\ln y$ por unidad de cambio en $\ln X$ (es decir, la elasticidad, β) permanece igual sin importar en cual $\ln X$ midamos la elasticidad, de aquí su nombre alterno modelo de elasticidad constante. El coeficiente de elasticidad, en la notación de cálculo se define como

$$\left(\frac{dy/y}{dX/X}\right) = \frac{dy}{dX}\frac{X}{y}.$$

En el modelo log-log, β es efectivamente el coeficiente de elasticidad. Esto resulta evidente del hecho que $d(\ln X)/dX = 1/X$ o $d(lnX) = dX/X$, es decir, para cambios infinitesimalmente cortos, un cambio en $\ln X$ es igual al cambio relativo o proporcional en X. Hay una diferencia notable entre el coeficiente de la pendiente y la medida de elasticidad. En el modelo log-log la pendiente (β) es el coeficiente de elasticidad, esto es la pendiente dy/dX multiplicada por la razón X/y. Sin embargo, la pendiente en el modelo lineal es solamente dy/dX.

Un tercer modelo conocido como semilogarítmico se da cuando $\lambda_1 = 0$ y $\lambda_2 = 1$

$$\ln y = \alpha + \beta X + e \tag{5.13}$$

donde $\alpha = \alpha_0 - \beta$. Modelos como este se denominan semilog porque solamente una variable aparece en forma logarítmica. Un modelo en el cual la variable independiente aparece en forma logarítmica se denomina log-lin. En este modelo el coeficiente de la pendiente mide el cambio proporcional constante o relativo en y para un cambio absoluto en X. Un caso especial ocurre cuando X denota el tiempo y la función entonces describe una variable y que despliega una tasa de crecimiento ($\beta > 0$) o de decremento ($\beta < 0$).

Un cuarto modelo se produce cuando se combinan valores de $\lambda_1 = 1$ y $\lambda_2 = -1$. Estos valores dan la relación

$$y = \alpha + \beta\left(\frac{1}{X}\right) + e. \tag{5.14}$$

Este modelo tiene las siguientes características: a medida que $X \to \infty$, el término $\beta(1/X) \to 0$ y y se aproxima a un valor límite o asintótico α. La pendiente de éste modelo es

$$\frac{dy}{dX} = -\beta\left(\frac{1}{X^2}\right) \tag{5.15}$$

e implica que si β es positivo, la pendiente siempre es negativa y si β es negativa, la pendiente es positiva. Una de las aplicaciones mas interesantes de este modelo se da cuando $\alpha < 0$ y $\beta > 0$, haciendo referencia a la curva de Phillips, que relaciona la tasa de cambio en los salarios sobre la tasa de desempleo.

Otro modelo se presenta con la elección de parámetros $\lambda_1 = 0$ y $\lambda_2 = -1$, a este modelo se le conoce como modelo recíproco logarítmico

$$\ln y = \alpha - \beta\left(\frac{1}{X}\right) + e. \tag{5.16}$$

Ignorando el término de perturbación, este modelo podríamos escribirlo como

$$y = e^{\alpha-(\beta/X)} \tag{5.17}$$

y damos cuenta que, y no esta definida para $X = 0$, $y \to 0$ cuando $X \to 0$, pero $y \to e^{\alpha}$ cuando $X \to \infty$. Podemos definir $y(0)$ como cero y entonces tendremos una función que es continua a la derecha del origen. De esta manera, la pendiente

$$\frac{dy}{dX} = \left(\frac{\beta}{X^2}\right)e^{\alpha-\beta/X} \tag{5.18}$$

es positiva para $\beta > 0$. En la segunda derivada

$$\frac{d^2y}{dX^2} = \left(\frac{\beta^2}{X^4} - \frac{2\beta}{X^3}\right)e^{\alpha-\beta/X} \tag{5.19}$$

existe un punto de inflexión cuando $X = \beta/2$. Su principal característica es que cerca del origen esta aumentá en una tasa creciente (convexa) y después del punto $X = \beta/2$ se incrementa en una tasa decreciente (cóncava).

Estos últimos modelos han dependido de la elección particular de 1, 0, o -1 para el parámetro λ. Otros valores de λ dan lugar a otras muchas formas funcionales más. El modelo de Box y Cox es una formulación útil que incluye muchos modelos que se incluyen como casos especiales. En los modelos anteriormente presentados el método de mínimos cuadrados ordinarios puede ser aplicado, probando que los supuestos del término de perturbación se cumplen. Desde luego, el modelo de Box y Cox en si mismo planeta sólo una ventaja relativa, si λ es conocido, simplemente introducimos el valor conocido y obtenemos un modelo que puede ser estimado con relativa facilidad. Excepto para los casos polares en que λ es igual a 1, 0, o -1, es difícil imaginar situaciones en las cuales un valor particular pueda especificarse *A priori.* Evitar restringir las consideraciones a justamente tres posibles valores de λ, mediante el tratamiento de λ como un parámetro adicional desconocido en la ecuación, se obtiene una gran flexibilidad, permite que las λ puedan ser estimadas y se puedan probar hipótesis sobre ellas pudiéndose así discriminar entre la forma funcional. El costo de la inclusión de un parámetro adicional consiste en que el modelo se convierte en uno no lineal implicando procedimientos de estimación iterativos.

La aplicación inmediata de la transformación de Box y Cox a todas las variables en una aplicación particular produce el modelo

$$y_t = \beta_1 + \beta_2 X_{t2}^{\lambda} + \beta_3 X_{t3}^{\lambda} + ... + \beta_k X_{tk}^{\lambda} + e_t. \tag{5.20}$$

Obviamente este modelo incluye a los modelos lineal y logarítmicos como casos especiales. Cuando $\lambda = 0$, este modelo es idéntico al log-lineal

$$\ln y_t = \beta_1 + \beta \ln X_{t2} + \beta \ln X_{t3} + ... + \beta_k \ln X_{tk} + e_t. \tag{5.21}$$

Para $\lambda = 1$

$$y_t - 1 = \beta_1 + \beta_2(X_{t2} - 1) + \beta_3(X_{t3} - 1) + ... + \beta_k(X_{tk} - 1) + e_t \tag{5.22}$$

o

$$y_t = (\beta_1 - \beta_2 - ... - \beta_1 + 1) + \beta_2 X_{t2} + \beta_3 X_{t3} + ... + \beta_k X_{tk} + e_t \tag{5.23}$$

$$y_t = \beta_1^* + \beta_2 X_{t2} + \beta_3 X_{t3} + ... + \beta_k X_{tk} + e_t. \tag{5.24}$$

Este modelo es equivalente al modelo lineal. Otros valores de λ definen otros modelos, y, desde un punto de vista econométrico, la idea es estimar λ junto con $\beta' = (\beta_1, \beta_2, ..., \beta_k)$, y por lo tanto estimar la forma funcional. Una familia más flexible de funciones es definida si especificamos una transformación de parámetros diferente para cada variable y de aquí intentamos estimar la función

$$y_t^{\lambda_1} = \beta_1 + \beta_2 X_{t2}^{\lambda_2} + \beta_3 X_{t3}^{\lambda_3} + ... + \beta_k X_{tk}^{\lambda_k} + e_t. \tag{5.25}$$

Esta familia podría incluir, por ejemplo, algunas variables enteramente lineales ($\lambda_k = 1$), algunas enteramente en términos de logaritmos ($\lambda_k = 0$), y algunas

enteramente como recíprocos ($\lambda_k = -1$). Esta flexibilidad adicionada no viene sin algún costo. Sin un número grande de observaciones puede ser difícil asegurar estimaciones de todos los parámetros de este modelo.

Para estimar el modelo

$$y_t^{\lambda} = \beta_1 + \beta_2 X_{t2}^{\lambda} + \beta_3 X_{t3}^{\lambda} + ... + \beta_k X_{tk}^{\lambda} + e_t \tag{5.26}$$

ha sido una práctica común asumir una distribución normal de los errores y aplicar el procedimiento de máxima verosimilitud. Una dificultad con este enfoque es que el supuesto de normalidad es incompatible con la transformación. La función de verosimilitud implica términos de la forma $\ln y_t$ que estan indefinidos para y_t negativos, y aún, si e_t están normalmente distribuidos, la posibilidad de y_t negativos existe. Alternativamente, la transformación implica que la distribución de los e_t este truncada, una característica no poseída por la distribución normal. Para proceder pragmáticamente con una estimación máximo verosímil es necesario asumir que el efecto de truncamiento no es importante y que los e_t son aproximadamente independientes e idénticamente distribuidos como una variable aleatoria normal con media 0 y varianza σ^2. Bajo este supuesto, la función de densidad conjunta para $\mathbf{e} = (e_1, e_2, ..., e_T)'$ es dada por

$$f(\mathbf{e}) = (2\pi\sigma^2)^{(-T/2)} \exp\left(\frac{\mathbf{e}'\mathbf{e}}{2\sigma^2}\right). \tag{5.27}$$

Se iniciará con el caso donde todas las variables son transformadas usando el mismo parámetro λ. La función de densidad conjunta para

$$\mathbf{y} = (y_1, y_2, ..., y_T)'$$

es dada por

$$f(\mathbf{y}) = (2\pi\sigma^2)^{(-T/2)}\exp\left(\frac{\mathbf{e}'\mathbf{e}}{2\sigma^2}\right)\left|\frac{\partial \mathbf{e}'}{\partial \mathbf{y}}\right| \tag{5.28}$$

y considerando a nuestro modelo con los parámetros de transformación como

$$\mathbf{y}^{(\lambda)} = \mathbf{X}^{(\lambda)}\beta + \mathbf{e} \tag{5.29}$$

el logaritmo de la función de verosimilitud para una muestra de T observaciones es

$$\ln L = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{t=1}^{T} e_t^2 \tag{5.30}$$

donde se puede deducir lo siguiente

$$\mathbf{e} = \mathbf{y}^{(\lambda)} - \beta\mathbf{X}^{(\lambda)}, \tag{5.31}$$

por tanto el Jacobiano es

$$\left|\frac{\partial \mathbf{e}'}{\partial \mathbf{y}}\right| = \prod_{t=1}^{T} y_t^{\lambda-1}. \tag{5.32}$$

Llevando a cabo la sutitución y multiplicando por el Jacobiano, se obtiene el logaritmo de la función de verosimilitud para el modelo Box-Cox:

$$\ln L(\beta, \sigma^2, \lambda|\mathbf{y}, \mathbf{X}) = -\frac{T}{2}\ln(2\pi) - \frac{T}{2}\ln\sigma^2 + (\lambda - 1)\sum_{t=1}^{T}\ln y_t$$

$$- \frac{1}{2\sigma^2}\sum_{t=1}^{T}(y^{(\lambda)} - x^{(\lambda)}\beta)^2. \tag{5.33}$$

La notación y^λ y X^λ es usada para denotar la colección de todas las T observaciones transformadas en la variable dependiente y variables explicatorias, respectivamente. Esto es, $\mathbf{y}^\lambda = (y_1^\lambda, y_2^\lambda, ..., y_T^\lambda)'$ y $\mathbf{X}^\lambda = (\mathbf{j}, X_1^\lambda, X_2^\lambda, ..., X_T^\lambda)'$ donde $\mathbf{j}$ es un vector n-dimensional de unos. Diferenciando la función (5.33) con respecto a β y σ^2, y colocando las derivadas igual a cero resulta

$$\hat{\beta}(\lambda) = (\mathbf{X}^{(\lambda)'}\mathbf{X}^{(\lambda)})^{-1}\mathbf{X}^{(\lambda)'}\mathbf{y}^{(\lambda)} \qquad \text{y} \qquad \hat{\sigma}^2(\lambda) = \frac{(\mathbf{y}^\lambda - \mathbf{X}^\lambda\hat{\beta}(\lambda))'(\mathbf{y}^\lambda - \mathbf{X}^\lambda\hat{\beta}(\lambda))}{T}. \tag{5.34}$$

Así para un λ conocido, los estimadores de máxima verosimilitud son encontrados aplicando las fórmulas convencionales a las observaciones transformadas. Para un λ desconocido, es necesario encontrar el valor maximizado $\tilde{\lambda}$, del cual los estimadores de máxima verosimilitud para β y σ^2 son dados por $\tilde{\beta} = \hat{\beta}(\tilde{\lambda})$ y $\tilde{\sigma}^2 = \hat{\sigma}^2(\tilde{\lambda})$, respectivamente.

Para encontar $\tilde{\lambda}$, sustituimos $\hat{\beta}(\lambda)$ y $\hat{\sigma}^2(\lambda)$ en $L(\beta, \sigma^2, \lambda|\mathbf{y}, \mathbf{X})$, produciendo la función de verosimilitud concentrada

$$L^*(\lambda) = \text{constante} - \frac{T}{2}\ln\hat{\sigma}^2(\lambda) + (\lambda - 1)\sum_{t=1}^{T}\ln y_t. \tag{5.35}$$

De manera que, la estimación procede primero numéricamente encontrando el valor de λ que maximiza $L^*(\lambda)$, con un algoritmo iterativo, y después usando $\tilde{\lambda}$ en la expresión $\hat{\beta}(\lambda)$ y $\hat{\sigma}^2(\lambda)$. Sea $\theta' = (\beta', \sigma^2, \lambda)$, una estimación de la matriz de covarianza asintótica para $\tilde{\theta}' = (\tilde{\beta}', \tilde{\sigma}^2, \tilde{\lambda})$ es dado por la inversa de la matriz de información $\mathbf{I}(\theta)$ evaluada en $\tilde{\theta}$, o, alternativamente, por la inversa del negativo del Hessiano del logaritmo de la función de verosimilitud, evaluada en $\tilde{\theta}$. Este estimador es dado por

$$\left[-\frac{\partial^2 L}{\partial\theta\partial\theta'} \right]^{-1} \Big|_{\theta=\tilde{\theta}}. \tag{5.36}$$

Si λ es conocida, entonces una estimación de la matriz de covarianza $\hat{\beta}(\lambda)$ es dada por

$$\hat{\sigma}^2(\lambda)(\mathbf{X}^{(\lambda)'}\mathbf{X}^{(\lambda)})^{-1}. \tag{5.37}$$

Los errores estándar computados de esta última expresión con λ reemplazado por $\hat{\lambda}$ puede ser considerado como un error estándar condicional, en $\lambda = \hat{\lambda}$. Estos errores estándar serán mas bajos que los incondicionales proporcionados por la inversa del negativo del Hessiano del logaritmo de la función de verosimilitud y por lo tanto sobreexpresará la seguridad del estimador para β. Si solamente σ^2 está concentrado fuera del logaritmo de la función de verosimilitud, es directamente mostrado que esta función es

$$L^*(\beta,\lambda) = C_1 - \frac{T}{2}\ln\left[\frac{(\mathbf{y}^\lambda - \mathbf{X}^\lambda\beta)'(\mathbf{y}^\lambda - \mathbf{X}^\lambda\beta)}{T}\right] + (\lambda-1)\sum_{t=1}^{T}\ln y_t \tag{5.38}$$

donde C_1 es una constante. Si C_2 es una nueva constante, puede ser mostrado que esta última expresión es igual a

$$L^*(\beta,\lambda) = C_2 - \frac{T}{2}\ln\left[\frac{(\mathbf{y}^{(\lambda)} - \mathbf{X}^{(\lambda)}\beta)'(\mathbf{y}^{(\lambda)} - \mathbf{X}^{(\lambda)}\beta)}{\bar{\mathbf{y}}_G^{2(\lambda)}}\right] \tag{5.39}$$

donde $\bar{\mathbf{y}}_G = (y_1, y_2, ..., y_n)^{1/T}$ es la media geométrica de y_t. Encontrar aquellos valores para (β,λ) que maximizan $L^*(\beta,\lambda)$ es equivalente a encontar aquellos valores que minimizan $\mathbf{e}^{*'}\mathbf{e}^*$ donde

$$\mathbf{e}^* = \frac{(\mathbf{y}^\lambda - \mathbf{X}^\lambda \beta)}{\bar{\mathbf{y}}_G^\lambda}. \tag{5.40}$$

Así, es posible escribir el problema de maximización dentro un cuerpo general de mínimos cuadrados no lineales, un cuerpo que es particularmente útil si algún tipo de software esta disponible.

Regresando de nuevo a la especificación

$$y_t^{\lambda_1} = \beta_1 + \beta_2 x_{t2}^{\lambda_2} + \beta_3 x_{t3}^{\lambda_3} + ... + \beta_k x_{tk}^{\lambda_k} + e_t, \tag{5.41}$$

donde cada variable es sujeta a una transformación diferente, el logaritmo de la función de verosimilitud puede ser escrita como

$$L(\beta, \sigma^2, \lambda, \lambda_1) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln \sigma^2 - \frac{1}{2\sigma^2}(y^{\lambda_1} - X^\lambda \beta)'(y^{\lambda_1} - X^\lambda \beta)$$

$$+ (\lambda_1 - 1)\sum_{t=1}^{T} \ln y_t \tag{5.42}$$

donde $\lambda' = (\lambda_2, \lambda_3, ..., \lambda_k)$ y $\mathbf{X}^{(\lambda)} = (\mathbf{j}, x_2^{(\lambda_2)}, ..., x_k^{(\lambda_k)})$. Los mismos pasos trazados pueden ser aplicados a esta función, la diferencia será que en este caso, tendremos que tratar con $(k-1)$ parámetros adicionales que no pueden ser concentrados fuera.

La Prueba de Linealidad

La estimación de parámetros de un modelo no lineal presupone que la forma funcional de la ecuación de regresión poblacional es conocida, o asumida, *A priori.*

Si este no es el caso, se considerará la especificación de la forma funcional una hipótesis comprobable más que una hipótesis mantenida. Esto es, frecuentemente se deseará probar la hipótesis que la ecuación de regresión poblacional es lineal con respecto a las variables contra alguna hipótesis alternativa. La más simple prueba de linealidad es en el caso en que la hipótesis alternativa es una ecuación de regresión que envuelve una función de potencia de un cierto grado. Su desventaja es que se debe expresar una función de potencia de un grado específico como la alternativa del modelo lineal. La idea básica de esta prueba descansa en el hecho de que una función lineal es un caso especial de una función de potencia, o sea una función de potencia de grado uno. Si los coeficientes conectados a las potencias más altas de las variables explicatorias son todos cero, la función de potencia dada se reduce a una regresión lineal simple. Esta idea puede ser explotada especificando otras formas funcionales, que incluyen linealidad como un caso especial. Particularmente adecuado para este propósito es la transformación de Box-Cox especificándola ahora como

$$\frac{y_i^{\lambda}-1}{\lambda}=\alpha+\beta\left(\frac{X_i^{\lambda}-1}{\lambda}\right). \tag{5.43}$$

En general, debido a que diferentes valores de λ en la ecuación de regresión especificada a través de la transformación de Box y Cox conducen a diferentes especificaciones funcionales de la ecuación de regresión. Esto permite probar la hipótesis de linealidad contra la hipótesis alternativa que la ecuación de regresión es alguna función de regresión no lineal dentro de la familia de funciones definida por la ecuación de regresión especificada a través de la transformación de Box y Cox. Formalmente,

$$H_O : \lambda = 1$$

$$H_1 : \lambda \neq 1$$

para llevar a cabo la prueba, necesitamos una estimación de λ y su error estándar. Podemos usar la prueba de razón de verosimilitud para probar la hipótesis de linealidad en que estamos interesados. La prueba estadística (para un muestra grande) es dada como

$$LR = -2[L(\lambda = 1) - L(\hat{\lambda})] \sim \chi_1^2 \tag{5.44}$$

donde $L(\lambda = 1)$ es igual al máximo valor de L para el modelo de regresión lineal, y $L(\hat{\lambda})$ representa el máximo valor de L cuando λ es igual a su valor de máxima verosimilitud.

La Función de Producción Cobb-Douglas y Función de Elasticidad de Sustitución Constante

Hay por supuesto una gran variedad de modelos empleados dentro del contexto de la teoría económica. La atención se concentrará en algunos casos de interés, en particular, sobre modelos no lineales asociados con las funciones de producción Cobb-Douglas y la función de elasticidad de sustitución constante. Suponga que se esta interesado en estimar la relación entre el producto en una industria y dos insumos, trabajo y capital. Si se utiliza una función de producción Cobb-Douglas para escribir esta relación podemos expresar esta como

$$y = \alpha L^{\beta_2} K^{\beta_3}. \tag{5.45}$$

Varias propiedades de esta función pueden ser descritas. Pero de particular interés

son los productos marginales, la tasa marginal de sustitución, y la elasticidad de sustitución de esta función. Los productos marginales están dados por

$$Pmg_l = \frac{\partial y}{\partial L} = \frac{\beta_2 y}{L} \tag{5.46}$$

$$Pmg_k = \frac{\partial y}{\partial K} = \frac{\beta_3 y}{K}. \tag{5.47}$$

Una isocuanta es una curva que describe estas combinaciones de trabajo y capital que producen un nivel dado de producto. La pendiente de la isocuanta mide la tasa en que capital puede sustituir trabajo (o viceversa) manteniéndose un nivel constante de producción. Esta pendiente es llamada tasa marginal de sustitución (TMS) y es dada por la razón de productos marginales. Esto es

$$TMS = \frac{Pmg_K}{Pmg_L} = \frac{\partial y/\partial K}{\partial y/\partial L} = \frac{\beta_3 y/K}{\beta_2 y/L} = \frac{\beta_3 L}{\beta_2 K}. \tag{5.48}$$

Una dificultad con la tasa marginal de sustitución como una medida de cuanto capital puede ser sustituido por trabajo es que depende de las unidades de medida de capital y trabajo. No es significativo comparar tasas marginales de sustitución de dos industrias que usan diferentes unidades de medida. Para superar este problema, el concepto de elasticidad de sustitución se hace preciso. Esta elasticidad es definida como

$$ES = \frac{d\ln(L/K)}{d\ln(TMS)}. \tag{5.49}$$

Cuando un ligero cambio en la pendiente (la TMS) es asociado con un gran cambio en la razón capital-trabajo (K/L), la isocuanta es plana, y un alto grado

de sustitución es posible; el valor de ES es grande. Cuando un gran cambio en la pendiente es asociado con un ligero cambio en la razón capital-trabajo, la isocuanta esta bien redondeada, y una ligera sustitución es posible; el valor de ES es ligero. En general, ES esta definida de 0 a ∞, y esta no depende de las unidades de medida de capital y trabajo. Para encontrar la elasticidad de sustitución de la función de producción de Cobb-Douglas, tomamos logaritmos en ambos lados de la TMS y reagrupando ésta obtenemos

$$\ln(L/K) = -\ln(\beta_3/\beta_2) + ln(TMS). \tag{5.50}$$

Encontrar la elasticidad de sustitución de esta ecuación es como encontrar la derivada dy/dx en la ecuación y=a+x. Esto es dado por

$$ES = \frac{d\ln(L/K)}{d\ln(TMS)} = 1 \tag{5.51}$$

es decir, la elasticidad sustitución de una función de producción de Cobb-Douglas es siempre igual a la unidad. Este resultado implica que la función de producción de Cobb-Douglas es bastante restrictiva; especificando una función de producción de Cobb-Douglas automáticamente se especifica la tasa en que dos insumos, trabajo y capital, son sustituidos uno con otro para alcanzar un nivel de producto. Bajo ciertas circunstancias sería más apropiado si se pudiera especificar una función donde se estimará la elasticidad de sustitución a partir de los datos. La función de producción ESC fue desarrollada con esta idea en mente.

La elasticidad de sustitución constante es una función de producción más general que la Cobb-Douglas. Esta es especificada como

$$y = \alpha[\delta K^{-\rho} + (1-\delta)L^{-\rho}]^{-\eta/\rho}. \tag{5.52}$$

Los parámetros desconocidos en esta función son el parámetro de eficiencia ($\alpha > 0$), el parámetro de retorno a la escala ($\eta > 0$), el parámetro de sustitución ($\rho > -1$), y el parámetro de distribución ($0 < \delta < 1$) que relaciona la porción de producto a los dos insumos. Dados los productos marginales, la tasa marginal de sustitución, y la elasticidad de sustitución, para esta función se puede ver como la elasticidad de sustitución esta relacionada al parámetro de sustitución ρ e indicar como la Cobb-Douglas puede ser vista como un caso especial de la función de producción ESC.

Los productos marginales de la función de producción ESC están dados por

$$Pmg_K = \frac{\partial y}{\partial K} = \eta\alpha^{-\rho/\eta}\delta K^{(-1+\rho)}y^{(1+\rho/\eta)} \tag{5.53}$$

$$Pmg_L = \frac{\partial y}{\partial L} = \eta\alpha^{-\rho/\eta}(1-\delta)L^{(-1+\rho)}y^{(1+\rho/\eta)}. \tag{5.54}$$

Tomando la razón de los productos marginales para obtener la tasa marginal de sustitución se tiene

$$TMS = \frac{Pmg_L}{Pmg_K} = \frac{\partial y/\partial K}{\partial y/\partial L} = \frac{\delta}{1-\delta}\left(\frac{L}{K}\right)^{1+\rho} \tag{5.55}$$

aplicando logaritmos en ambos lados de esta ecuación, reagrupando y diferenciando, se obtiene la elasticidad sustitución,

$$ES = \frac{d\ln(L/K)}{d\ln(TMS)} = \frac{1}{1+\rho}. \tag{5.56}$$

Así, en la función de producción CES, la elasticidad sustitución depende del parámetro desconocido ρ. Estimando ρ a partir de los datos se está sugiriendo una elasticidad sustitución. Este resultado se contrasta con la función de producción Cobb-Douglas donde la elasticidad de sustitución es siempre igual a la unidad. Es preciso notar que $ES \to 0$ cuando $\rho \to \infty$; $ES = 1$ cuando $\rho = 0$ y $ES \to \infty$ cuando $\rho \to -1$. De manera que el valor más grande de ρ, es el menor valor de elasticidad de sustitución.

También, usando límites, puede mostrarse que cuando $\rho \to 0$, la función de producción ESC se aproxima a una función de producción Cobb-Douglas con elasticidad de sustitución igual a uno. En este sentido la función de producción Cobb-Douglas puede ser vista como un caso especial de la función de producción ESC, y se puede probar lo apropiado de una función de producción de Cobb-Douglas probando si ρ es significativamente diferente de cero.

El modelo económico

$$y = \alpha[\delta K^{-\rho} + (1-\delta)L^{-\rho}]^{-\eta/\rho} \tag{5.57}$$

puede ser expresado como un modelo estadístico

$$y_t = \alpha[\delta K_t^{-\rho} + (1-\delta)L_t^{-\rho}]^{-\eta/\rho} \exp\{e_t\} \tag{5.58}$$

aplicando logaritmos se tiene

$$\ln(y_t) = \beta - \frac{\eta}{\rho}\ln[\delta K_t^{-\rho} + (1-\delta)L_t^{-\rho}] \tag{5.59}$$

donde $\ln\alpha = \beta$. Además, se asume que los errores están distribuidos independiente

e idénticamente, con media cero y varianza constante. Para estimar cada uno de los parámetros $(\beta, \eta, \rho, \delta)$ establecemos la función de la suma de cuadrados del error

$$S(\beta, \eta, \rho, \delta) = \sum_{t=1}^{T} e_t^2 = \sum_{t=1}^{T} \{\ln Y_t - \beta + \frac{\eta}{\rho} \ln[\delta L^{-\rho} + (1-\delta)K^{-\rho}]\}^2 \qquad (5.60)$$

y encontramos las estimaciones que minimicen esta suma. A causa de que hay cuatro parámetros implicados, es difícil encontrar el mínimo de esta función. Sin embargo, a través de mínimos cuadrados no lineales y con la ayuda de algún software puede encontrarse. Si se obtienen las derivadas parciales de la suma de cuadrados del error con respecto a cada uno de los parámetros desconocidos, y se igualan las derivadas parciales a cero, se obtiene una colección de cuatro ecuaciones en los cuatro parámetros desconocidos. Las estimaciones mínimo cuadráticas serían una solución a este conjunto de ecuaciones. Desafortunadamente, las cuatro ecuaciones serán no lineales.

Una alternativa mas simple de estimación de la función de producción ESC es posible si se reemplaza por su aproximación lineal con respecto a ρ. Usando series de Taylor expandiendo $\log Q$ alrededor de $\rho = 0$, y eliminando los términos que implican potencias de ρ más altas que uno, se obtiene

$$\log Q_i = \log \gamma - \nu\delta \log K_i + \nu(1-\delta) \log L_i - \frac{1}{2}\rho\nu\delta(1-\delta)[\log K_i - \log L_i]^2 + e_i. \quad (5.61)$$

Note que el lado derecho puede ser convenientemente separado en dos partes, una corresponde a la función de producción de Cobb-Douglas y una representa una corrección debido a la desviación de ρ de cero. La parte dada por

$$-\frac{1}{2}\rho\nu\delta(1-\delta)[\log K_i - \log L_i]^2, \tag{5.62}$$

desaparecerá si $\rho = 0$. De esta manera, este modelo puede representarse como uno intrínsecamente lineal,

$$\log Q_i = \beta_1 - \beta_2 \log K_i + \beta_3 \log L_i + \beta_4[\log K_i - \log L_i]^2 + e_i \tag{5.63}$$

donde:

$$\gamma = \exp(\beta_1), \quad \delta = \frac{\beta_2}{\beta_2 + \beta_3}, \quad \nu = \beta_2 + \beta_3, \quad \text{y} \quad \rho = \frac{-2\beta_4(\beta_2 + \beta_3)}{\beta_2\beta_3} \tag{5.64}$$

así se podrán utilizar estimaciones mínimo cuadráticas para obtener estimaciones de los parámetros $\beta's$.

La elección de la forma funcional debe ser hecha con fundamentos teóricos y empíricos. A causa de que tales fundamentos son frecuentemente difíciles de establecer, se han hecho esfuerzos en varias formas para resolver el problema de la elección. Un enfoque ha sido el uso de formas funcionales lo suficientemente flexibles que permitan la aproximación a una variedad de formas especiales. En el contexto de las funciones de producción la forma más flexible es la llamada función de producción transcendental (o translog). Esta es obtenida de la aproximación a la función de producción ESC permitiendo que los coeficientes $(\log K_i)^2$, $(\log L_i)^2$, y $-2(\log K_i)(\log L_i)$ difieran. Así la función de producción translog para dos insumos es

$$\log Q_i = \alpha - \beta_K \log K_i + \beta_L \log L_i + \beta_{KK}(\log K_i)^2$$

$$+ \beta_{LL}(\log L_i)^2 + \beta_{KL}(\log K_i)(\log L_i) + e_i. \tag{5.65}$$

Otra forma flexible comúnmente utilizada y ya enunciada anteriormente es la transformación de Box y Cox. Si aun tales formas flexibles son consideradas demasiado restrictivas, pueden ser usada una variable dummy en la regresión. Tal formulación solamente hace posible probar la relevancia de una variable explicatoria pero no estima su efecto en la variable dependiente.

La Función Logística

Ya se ha mencionado que los modelos lineales no son los únicos que pueden utilizarse para representar el mundo real, la ciencia económica hace uso también de otros modelos empleados en otras disciplinas como es el caso de la función logística.

$$y = \frac{a}{1 + be^{-ct}} \tag{5.66}$$

donde a, b, y c son parámetros a ser determinados y donde se ha escrito a y como una función del tiempo t, sin embargo en algunas aplicaciones es bastante posible reemplazar t por alguna variable independiente X. La curva logística tiene una altura asintótica en algun nivel finito y una parte baja asintótica en cero cuya expresión, por ejemplo, bien puede usarse en ajustar tendencias de crecimiento de cualquier población, ya sea que sea bacterial, animal, humana, o económica, donde el crecimiento es pensado positivamente relacionado al tamaño de población existente y negativamente relacionado a la distancia de un nivel de saturación; o bien puede usarse para representar el proceso de introducción de un nuevo producto donde la tasa de adopción es lenta en las etapas iniciales, luego sube y finalmente

se estabiliza.

En la función logística es claro que

$$y \to a \qquad \text{cuando} \qquad t \to \infty \qquad \text{y} \qquad y \to 0 \qquad \text{cuando} \qquad t \to -\infty$$

así que a es la parte asintótica de arriba y cero en la parte asintótica de abajo. La primera derivada de la función logística es

$$\frac{dy}{dt} = \frac{abce^{-ct}}{(1+be^{-ct})^2} = \left(\frac{y}{a}\right)abce^{-ct} = \frac{c}{a}y(ybe^{-ct}) = \frac{c}{a}y(y-a). \tag{5.67}$$

Así la tasa de cambio de y con respecto a t es proporcional al nivel corriente y y también a la distancia para alcanzar el nivel de saturación. La primera derivada es positiva para todos los valores de t. A partir de la primera derivada

$$\frac{dy}{dt} = cy - \frac{c}{a}y^2 \tag{5.68}$$

se obtiene la segunda derivada que puede ser escrita como

$$\frac{d^2y}{dt^2} = c\frac{dy}{dt} - 2\frac{c}{a}y\frac{dy}{dt} = \frac{c}{a}(a-2y)\frac{dy}{dt}. \tag{5.69}$$

Igualando la segunda derivada a cero se encuentra un punto de inflexión en $y = a/2$ y $t = (1/c)\ln b$, esto es,

$$a = 2y, \qquad y = \frac{a}{2}, \qquad \frac{a}{2} = \frac{a}{1+be^{-ct}}, \tag{5.70}$$

y sustituyendo $a/2$ en y tenemos

$$2 = 1 + be^{-ct}, \qquad ct = \ln b, \qquad t = \frac{1}{c} \ln b. \tag{5.71}$$

Así cuando $y < a/2$, el valor más grande de $a-y$ domina al valor mas pequeño de y en la ecuación $dy/dt = \frac{c}{a}y(y-a)$ y causa que dY/dt se incrementa en forma más que proporcional con el tiempo hasta el instante que $a - 2y = 0$ o $(y = a/2)$. Cuando y crece hacia $a/2$, el balance relativo de las dos fuerzas cambia así que dy/dt alcanza un valor máximo cuando $y = a/2$ y después de ello declina constantemente cuando y crece hacia el nivel de saturación a, obteniéndose un crecimiento menos que proporcional si $y > a/2$, puesto que entonces $y'' < 0$. La representación gráfica de la curva logística es la de una función exponencial, creciente en todo el campo de variación en el tiempo con dos asíntotas paralelas al eje de las abscisas (Y=h, Y= a +h) y un punto de inflexión en $Y = (1/c) \ln b$ y $y = a/2$, considerándose aquí el caso de h =0.

La estimación de todos los parámetros a, b, y c no pueden ser alcanzada por los métodos ordinarios. De la ecuación dy/dt se puede escribir

$$\frac{1}{y}\frac{dy}{dt} = c - \left(\frac{c}{a}\right)y. \tag{5.72}$$

Si el tiempo es medido en unidades constantes, el lado izquierdo de esta ecuación es aproximadamente, la tasa proporcional de crecimiento de y. De manera que podría ser ajustada una regresión lineal

$$\frac{y_{t+1} - y_t}{y_t} = \hat{c} - \left(\frac{\hat{c}}{a}\right)y_t + e_t \tag{5.73}$$

que resulta de la estimación de a y c. Para obtener una estimación de b, la función

logística puede ser arreglada para dar

$$b = \frac{a - y}{y} e^{ct}. \tag{5.74}$$

De forma que un valor de b puede ser computado para cada y_t , dando la estimación de c y la estimación de a . Una estimación podría ser entonces promediando algunos o todos los valores computados de b, o alternativamente sustituyendo la media de y y la media de t en la ecuación $b = (a - y)e^{ct}/y$.

Hay dos dificultades principales con este simple procedimiento para estimar la función logística. Primero se tiene una estimación puntual de los parámetros, pero procedimientos de inferencia son difíciles especialmente para b y a. Segundo, hay evidencia que el procedimiento es insatisfactorio comparado con una estimación directa con métodos no lineales.

La velocidad (ritmo) de crecimiento, determinada por la función logística en cualquier momento, es decir, la primera derivada de la función y con respecto al tiempo t, (dY/dt), es proporcional al valor existente y, y al factor de retardo $(a-y)$, o sea, a la distancia del valor existente y con relación al nivel de saturación a o la ordenada de la asíntota a la curva logística. Pudiéndose representar la propiedad anterior, conocida tambien como ley de crecimiento o crecimiento de acuerdo a la Ley de Robertson, por la siguiente relación

$$\frac{dy}{dt} = ky(a - y) \tag{5.75}$$

en la cual $k > 0$ denota el coeficiente de proporcionalidad entre la velocidad de crecimiento de la función y el producto del valor de la función y que se incrementa con el tiempo t y el factor $(a - y)$ cuya magnitud decrece con el tiempo. Cuando

el proceso de expansión se acerca a la saturación, el factor de retardo se acerca a cero y la velocidad de crecimiento dY/dt tiende a cero, esto significa que la función comienza a estabilizarse cerca del nivel $y = a$.

La ley de crecimiento también puede escribirse de la siguiente forma

$$\frac{1}{y}\frac{dy}{dt} = k(a - y) \qquad \text{o} \qquad \frac{d\log y}{dt} = k(a - y) \tag{5.76}$$

en donde puede verse que el logaritmo de la función y se incrementa a una velocidad igual a una constante k multiplicada por el factor de retardo. Si el factor de retardo no existiera en el proceso dado, entonces se tendría

$$\frac{1}{y}\frac{dy}{dt} = \frac{d\log y}{dt} = k \tag{5.77}$$

lo que significa que la tasa de crecimiento (incremento relativo de la funcion) sería constante y el crecimiento tendría lugar de acuerdo con una curva exponencial del tipo

$$y = ab^t. \tag{5.78}$$

La fórmula de la función logística puede deducirse de la ley de crecimiento expresada como una ecuación diferencial. Analizando el crecimiento en el tiempo t de una población y, que se desarrolla en un espacio finito la hipótesis de trabajo sería la siguiente, si $0 < y_t < a$ que limita a a al posible número de habitantes que pueden habitar aquel espacio finito, entonces

$$\frac{dy}{dt} = ky(a - y), \qquad k > 0 \tag{5.79}$$

que puede ser expresada como

$$\frac{dy}{y(a-y)} = kdt \tag{5.80}$$

y dado que

$$\frac{1}{y(a-y)} = \frac{1}{a}\left(\frac{1}{y} + \frac{1}{a-y}\right) \tag{5.81}$$

integrando la ecuación diferencial $dy/dt = ky(a-y)$ resulta que

$$\int \left(\frac{1}{y} + \frac{1}{a-y}\right) dy = a \int kdt$$

$$\log \frac{y}{a-y} + c_1 = akt + c_2$$

$$\log(y) - \log(a-y) + c_1 = akt + c_2 \tag{5.82}$$

donde c_1 y c_2 son constantes de integracion,

$$\log \frac{y}{a-y} = akt + c_2 - c_1 \tag{5.83}$$

o bien,

$$\frac{y}{a-y} = e^{akt+c_2-c_1}, \qquad ye^{-akt}e^{c_1-c_2} = a-y \qquad , a-y = ye^{-ct}b \tag{5.84}$$

donde

$$b = e^{c_2 - c_1} \qquad \text{y} \qquad c = ak$$

se obtiene

$$a = ye^{-ct}b + y \tag{5.85}$$

encontrando y en esta ecuación tenemos una primera expresión de la curva logística para cualquier clase de población.

CAPITULO 6

EJEMPLO DE ESTIMACION E INFERENCIA DENTRO DEL CONTEXTO DE UN MODELO DE REGRESION NO LINEAL

Estimación de los Parámetros de Regresión No Lineales

En este capítulo se presentan las técnicas básicas de estimación e inferencia en modelos de regresión no lineal a través de un ejemplo. Se describe el proceso de estimación de parámetros e inferencia empleando las técnicas plasmadas en el capítulo 4. Se ilustran los procedimientos en una forma sencilla, mediante un ejemplo donde el modelo de regresión incluye solamente dos parámetros y se utiliza una muestra de tamaño moderada; es decir, el ejemplo permitirá ilustrar el desarrollo de las técnicas, mas que la evidencia empírica en si misma. Puesto que la mayoría de los paquetes estadísticos consideran la regresión no lineal, se emplea como herramienta auxiliar de cálculo en el ejercicio el paquete Prostat, del cual se presentará cada uno de los pasos para obtener el resultado final.

De esta manera, se inicia con un ejemplo teniendo en cuenta los argumentos proporcionados por la la teoría económica, la cual indica que la demanda de cualquier bien depende generalmente del precio del bien, de los precios de otros bienes que compiten con él o que son complementarios al bien primario, y del ingreso del consumidor. Para llevar a cabo dicha tarea se estructuró un modelo de demanda no lineal con término de error aditivo en el que solamente se relaciona

la variable cantidad de demanda con la variable precio y en el cual se utilizan los datos que se presentan en la cuadro 5.1 y que fueron estudiados por Gujarati (1997) considerando el período de 1970 a 1980 utilizando para ello un modelo exponencial con un término de error multiplicativo. Se decidió conocer lo apropiado del modelo de regresión con dos parámetros

$$y = \beta_1 X^{\beta_2} + e$$

es decir, se quizó saber si el modelo era apto en la estimación de la función de demanda.

Como se mencionó en el capítulo 4, el método de mínimos cuadrados para modelos de regresión no lineal hace uso del mismo criterio empleado en modelos de regresión lineal, el cual consiste en la minimización de la suma de cuadrados del error (SCE) que en nuestro modelo esta expresado por

$$SCE = \sum_{t=1}^{T} (y - \beta_1 X^{\beta_2})$$

Aquellos valores de β_1 y β_2 que minimizan la suma de cuadrados del error (SCE) para la muestra dada de observaciones y_t y X_t son las estimaciones mínimo cuadráticas y son denotadas por b_1 y b_2.

Un método para encontrar las estimaciones mínimo cuadráticas es por medio de procedimientos de búsqueda numéricos. En esta parte, se ilustrará el método de Gauss Newton, también llamado método de linearización. Este método usa una expansión en serie de Taylor para aproximar un modelo de regresión no lineal con términos lineales y después emplea mínimos cuadrados ordinarios para estimar los parámetros. La iteración de estos pasos generalmente conduce a una solución al problema de regresión no lineal.

Ambos métodos descritos en este trabajo, el algoritmo de Gauss Newton y el de Newton Raphson comienzan con valores iniciales para los parámetros de re-

gresión β_1 y β_2, los cuales pueden ser obtenidos de estudios previos o de estudios relacionados, expectativas teóricas, o una búsqueda preliminar de valores de los parámetros que conducen a un valor comparativamente bajo de la suma de cuadrados del error (SCE). Es decir, se evalúa la suma de cuadrados del error (SCE) para diferentes valores de β_1 y β_2, variando estos sistemáticamente hasta que el valor mínimo es encontrado.

Cuadro 5.1 Consumo de Café en los Estados Unidos 1970-1980.

AÑO	Y (TAZAS DIARIAS POR PERSONA)	X (U.S. DLLS POR LIBRA)
1970	2.57	0.77
1971	2.50	0.74
1972	2.35	0.72
1973	2.30	0.73
1974	2.25	0.76
1975	2.20	0.75
1976	2.11	1.08
1977	1.94	1.81
1978	1.97	1.39
1979	2.06	1.20
1980	2.02	1.17

Fuente: D.N. Gujarati, Econometría, McGraw Hill, Colombia, 1997, pp. 81.

Cabe mencionar la existencia de un segundo método para encontrar estimaciones mínimo cuadráticas por medio de las ecuaciones normales. Con este enfoque se encuentran analíticamente las ecuaciones normales mínimo cuadráticas derivando

la suma de cuadrados del error (SCE) con respecto a β_1 y β_2 e igualando las ecuaciones a cero. La solución de las ecuaciones normales resultan ser las estimaciones mínimo cuadráticas. Sin embargo, es más práctico en muchos problemas de regresión no lineal, encontrar las estimaciones mínimo cuadráticas por procedimientos de búsqueda numéricos directos en lugar de obtener las ecuaciones normales y luego usar métodos numéricos para encontrar iterativamente la solución para éstas ecuaciones. La mayoría de los paquetes computacionales estadísticos emplean uno o más procedimientos de búsqueda numérica directa para resolver problemas de regresión no lineal.

En nuestro modelo de regresión no lineal, el criterio mínimo cuadrático en ésta etapa requiere evaluar la función $y = \beta_1 X^{\beta_2}$ para cada nivel de precio, utilizando los valores de los parámetros iniciales $b_1 = 2.1$ y $b_2 - 0.1$. Por ejemplo, para el nivel de precio, $X_1 = 0.77$, la cantidad de café demandada es 2.15561016. Ya que la cantidad demandada en el año de 1970 es igual a 2.57, la desviación de la respuesta media será entonces igual a 0.414338984. Vemos que ésta desviación es el residuo para el precio del año de 1970 para ésta primera estimación inicial. De manera que podemos obtener un vector de respuesta media:

$$\mathbf{y} - f(\mathbf{X}, \beta_1) = \begin{pmatrix} Y_1 - b_1 X_1^{b_2} \\ . \\ . \\ . \\ Y_{11} - b_1 X_{11}^{b_2} \end{pmatrix} = \begin{pmatrix} 0.414338984 \\ . \\ . \\ . \\ -0.4728669 \end{pmatrix}$$

siendo el criterio mínimo cuadrático en ésta etapa inicial la suma de cuadrados de los residuales en la etapa 1,

$$SCE(1) = e'e = [\mathbf{y} - f(\mathbf{X}, \beta_1)]'[\mathbf{y} - f(\mathbf{X}, \beta_1)] = 0.3526781$$

Para revisar los valores iniciales de los parámetros se requiere de la matriz $\mathbf{z}(\beta)$ y del conjunto de valores iniciales b_1 y b_2, esto es,

$$\mathbf{z}(\beta_1) = \begin{pmatrix} X_1^{b_2} & b_1 X_1^{b_2} \ln(X_1) \\ . & . \\ . & . \\ . & . \\ X_{11}^{b_2} & b_1 X_{11}^{b_2} \ln(X_{11}) \end{pmatrix}$$

La cuál permite obtener la matriz

$$\mathbf{z}(\beta_1)'\mathbf{z}(\beta_1) = \begin{pmatrix} 11.106623872 & -1.337310167 \\ -1.337310167 & 4.551839763 \end{pmatrix}$$

y el vector

$$\mathbf{z}(\beta_1)'[\mathbf{y} - f(\mathbf{X}, \beta_1)] = \begin{pmatrix} 1.10895974 \\ -0.84867256 \end{pmatrix}$$

De manera que se puede obtener una estimación revisada mínimo cuadrática, β_2, al final de la primera iteración

$$\beta_2 = \beta_1 + [\mathbf{z}(\beta_1)'\mathbf{z}(\beta_1)']^{-1}\mathbf{z}(\beta_1)'[\mathbf{y} - f(\mathbf{X}, \beta_1)]$$

$$= \begin{pmatrix} 2.1 \\ -0.1 \end{pmatrix} + \begin{pmatrix} 0.08023563 \\ -0.16287378 \end{pmatrix} = \begin{pmatrix} 2.18023563 \\ -0.26287378 \end{pmatrix}$$

En éste punto se puede examinar si los coeficientes de regresión representan arreglos en la dirección apropiada. Cabe hacer mención que, en el cálculo de la suma de cuadrados del error se utiliza la función de regresión no lineal, y no la aproximación lineal en la expansión de serie de Taylor. De esta forma, al término de la primera iteración el criterio mínimo cuadrático produce una segunda suma de cuadrados del error $SCE(2)$ igual a 0.12445532. Lo cual indica que el método de Gauss-Newton trabaja efectivamente en la dirección correcta ya que $SCE(2) < SCE(1)$, de manera

que los coeficientes de regresión revisados son mejores estimaciones que los valores utilizados inicialmente.

Los coeficientes de regresión han sido moderadamente revisados y la condición de primer orden

$$\mathbf{z}(\beta)'[\mathbf{y} - f(\mathbf{X}, \beta)] = 0$$

aun no se cumple, ya que este vector hasta la actual iteración tiene elementos diferentes a los que incluye el vector cero, por lo que se continúa revisando la estimación de los parámetros para la cual se requiere nuevamente la matriz $\mathbf{z}(\beta)$ y el vector $\mathbf{z}(\beta)'[\mathbf{y} - f(\mathbf{X}, \beta)] = 0$ pero ahora empleando el conjunto de valores revisados β_2. A partir del cual se generaran nuevas estimaciones de los coeficientes de regresión a partir de la expresión ya enunciada en el capítulo 4.

$$\beta_3 = \beta_2 + [\mathbf{z}(\beta_2)'\mathbf{z}(\beta_2)']^{-1}\mathbf{z}(\beta_2)'[\mathbf{y} - f(\mathbf{X}, \beta_2)]$$

Obteniéndose como resultado el vector

$$\beta_3 = \begin{pmatrix} 2.17725719 \\ -0.26006446 \end{pmatrix}$$

y una suma de cuadrados del error $SCE(3) = 0.1242813$, inferior a la suma de cuadrados del error originada al término de la suma de cuadrados de la primera iteración. El proceso iterativo continúa hasta que la diferencia entre coeficientes estimados sucesivos y/o la diferencia entre criterios mínimos cuadráticos sucesivos $SCE(n+1) - SCE(n)$ se vuelve insignificante. En nuestro ejemplo, el procedimiento continúa hasta la cuarta iteración en la cual no se registra cambio en los coeficientes estimados, ni se registra un mejor ajuste de acuerdo al criterio mínimo cuadrático. De aquí que la estimación final de los coeficientes de regresión sea

$$\beta_5 = \beta_4 + [\mathbf{z}(\beta_4)'\mathbf{z}(\beta_4)']^{-1}\mathbf{z}(\beta_4)'[\mathbf{y} - f(\mathbf{X}, \beta_4)]$$

$$= \begin{pmatrix} 2.177225724 \\ -0.26006418 \end{pmatrix} + \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 2.177225724 \\ -0.26006418 \end{pmatrix}$$

con una suma de cuadrados del error igual a 0.1242812998.

Las estimaciones mínimo cuadráticas no lineales se presentan en el cuadro 5.2. Los procedimientos de prueba de hipótesis e intervalos de confianza descritos en la siguientes sección utilizan la última iteración.

Cuadro 5.2. Estimación de la función de demanda.

Iteración	S.C.E.	β_1	β_2
0	0.3526578148	2.10000000	-0.10000000
1	0.1244553190	2.18023563	-0.26287379
2	0.1242813038	2.17725312	-0.26009434
3	0.1242812998	2.17725719	-0.26006446
4	0.1242812998	2.17725724	-0.26006419

Inferencia sobre los Parámetros de Regresión No Lineales

Para efectuar las pruebas de hipótesis y establecer las estimaciones por intervalo, en nuestro ejemplo, se puede utilizar los procedimientos habituales, teniendo en consideración los resultados presentados en secciones anteriores referentes a las propiedades asintóticas de los estimadores. De esta manera, por lo que se refiere a la inferencia sobre los parámetros de regresión no lineales se requiere una estimación del término varianza del error $\hat{\sigma}^2$. Esta estimación es similar a la de la regresión lineal:

$$\hat{\sigma}^2 = \frac{SCE}{T-k} = \frac{[\sum_{t=1}^{T} \mathbf{y} - f(\mathbf{X}, \beta)]^2}{T-k}$$

donde β es el vector de las estimaciones finales de los parámetros. Para la regresión no lineal, $\hat{\sigma}^2$ no es un estimador insesgado de σ^2, pero el sesgo es mas pequeño cuando el tamaño de la muestra es grande. Cuando los términos del error son independientes y normalmente distribuidos y el tamaño de la muestra es razonablemente grande los estimadores mínimo cuadráticos β en el caso de la regresión no lineal se distribuyen aproximadamente normal y son insesgados.

En la estimación de la matriz de varianzas y covarianzas de los coeficientes de regresión se utiliza la expresión:

$$\widehat{var}(b) = \hat{\sigma}^2[\mathbf{z}(b)'\mathbf{z}(b)]^{-1}$$

ya descrita en el capítulo 4, donde $\mathbf{z}(b)$ es la matriz de derivadas parciales evaluada en la estimación final mínimo cuadrática b. Es preciso mencionar que la estimación aproximada de la matriz de varianzas y covarianzas $\widehat{var}(b)$ es exactamente de la misma forma como en la regresión lineal, con $\mathbf{z}(b)$ jugando de nuevo el papel de la matriz X. En el ejemplo de la demanda de café, la varianza estimada es

$$\hat{\sigma}^2 = \frac{SCE}{T-k} = \frac{0.1242812998}{11-2} = 0.013809033$$

y una estimación aproximada de la matriz de varianzas y covarianzas de los coeficientes de regresión es

$$\widehat{var}(b) = \hat{\sigma}^2[\mathbf{z}(b)'\mathbf{z}(b)]^{-1} = \begin{pmatrix} 0.001321343 & 0.000573814 \\ 0.000573814 & 0.0031303587 \end{pmatrix}$$

Por lo que respecta a la estimación por intervalo de un único parámetro, cuando los términos de error en el modelo de regresión no lineal son independientes y normalmente distribuidos, el siguiente resultado aproximado se mantiene cuando el tamaño de muestra es grande

$$\frac{b-\beta}{\sqrt{\hat{var}(b)}} \sim z$$

donde z es una variable normal estándar. De aquí que, aproximadamente $1-\alpha$ límites de confianza para cualquier único parámetro esta formado en la forma usual por medio de la utilización de percentiles de la distribución normal estándar

$$b + z(1-\frac{\alpha}{2})\widehat{var}(b)$$

donde $z(1-\alpha/2)$ es el $(1-\alpha/2)100$ percentil de la distribución normal estándar. En nuestro ejemplo, si se desea efectuar estimaciones por intervalo para cada uno de los dos parámetros del modelo, con un 95% de confianza, estos serán:

$$2.10601069 \leq \beta_0 \leq 2.24850379$$

y para β_1

$$-0.36925549 \leq \beta_1 \leq -0.15087289.$$

Así se concluye con aproximadamente 95% de confianza que β_1 esta entre 2.10601069 y 2.24850379, mientras que β_2 esta entre -0.36925549 y -0.15087289.

Para muestras de tamaño pequeño, algunos estadísticos recomiendan que las inferencias sobre los parámetros de los modelos de regresión no lineal se hagan usando la distribución t basados en $T-k$ grados de libertad en lugar de la distribución normal estándar. Ya que el valor t será siempre más grande que el valor Z, este argumento es un poco mas conservador. Para ilustrar esto, en nuestro ejemplo, una estimación por intervalo del 95% de confianza, se requiere un valor t(0.975,9)=2.262774. El límite de confianza entonces será para β_0

$$2.09500477 \leq \beta_0 \leq 2.25950971$$

y para β_1

$$-0.38612298 \leq \beta_1 \leq -0.1340054$$

estos intervalos son ligeramente mas amplios que los estructurados con la distribución normal.

En cuanto a las pruebas concernientes a un único parámetro, estas son ejecutadas en la en la forma usual. Esto es para probar

$$H_0 : \beta_i = \beta_{io}$$

$$H_a : \beta_i \neq \beta_{io}$$

donde β_{io} es el valor especificado de β_i, podemos usar el estadístico de prueba z cuando n es razonablemente grande

$$z = \frac{b - \beta}{\sqrt{\widehat{var}(b)}}$$

La regla de decisión para controlar el riesgo de cometer un error de tipo I en un α aproximado es entonces:

$$Si \quad |z| \leq z(1 - \alpha/2), \quad \text{concluimos} H_0$$

$$Si \quad |z| > z(1 - \alpha/2), \quad \text{concluimos} H_a$$

En nuestro ejemplo deseamos probar que $\beta_2 = 1$ una prueba t asintótica basado en una distribución normal estandarizada, se lleva a cabo utilizando

$$z = \frac{-0.26006419 - 1}{0.055709847} = -22.618338729$$

este valor es menor que el valor crítico de -1.96 para el nivel de significación del 5% y, por tanto, se rechaza el modelo lineal en favor de la regresión no lineal.

Cuando se desea una prueba concerniente a varios parámetros simultáneamente, se usa el mismo enfoque como en la prueba lineal general, primero se ajusta el modelo completo y se obtiene $SCE(C)$, después se ajusta el modelo reducido y se obtiene la $SCE(R)$, y finalmente se calcula la misma prueba estadística como en la regresión lineal

$$F = \frac{SCE(R) - SCE(F)}{(g.l._R - g.l._c)} \div CME(c).$$

Para un tamño de muestra grande, esta prueba estadística esta distribuida aproximadamente como $F(g.l._R - g.l._c, g.l._c)$ cuando H_0 se mantiene.

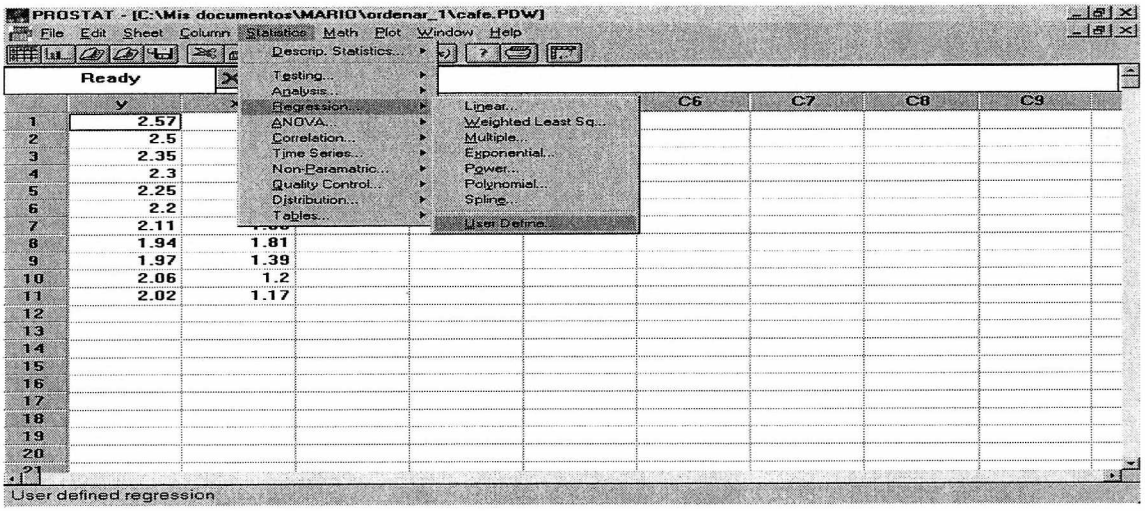
Procedimientos para el Ajuste de un Modelo No Lineal

Prostat for Windows habilita el procedimiento user-defined non-linear fitting para obtener estimaciones de los parámetros de un modelo no lineal (respecto a sus parámetros). Si un modelo es lineal en los parámetros y no lineal en las variables se llama modelo intrínsecamente lineal, por ser precisamente linealizable mediante la aplicación de transformaciones adecuadas. Pero si un modelo es no lineal en los parámetros, no puede transformarse a lineal y se denomina intrínsecamente no lineal. Este tipo de modelos son los que trata el procedimiento user-defined non-linear fitting. Este procedimiento usa un algoritmo iterativo de búsqueda de soluciones aproximadas para las estimaciones de los parámetros partiendo de ciertos valores iniciales lógicos definidos por el usuario. Los modelos intrínsecamente no lineales no pueden ser estimados directamente mediante el procedimiento de mínimos cuadrados ordinarios.

La pantalla de entrada del procedimiento user-defined non-linear fitting es la presentada en la figura 5.1. Para acceder a esta pantalla se tienen las siguientes

instrucciones: Statistics / Regression / User Defined command.

Figura 6.1. Pantalla de entrada al procedimiento User Defined command.



Un editor sencillo es proporcionado para redactar las ecuaciones del modelo. Este editor usa combinaciones sencillas de teclado y ratón para obtener la información. Todos los caracteres uniformes de ASCII se aceptan. Pero se debe tener las ecuaciones disponibles para describir los datos antes de que haga el ajuste de la curva. El método dará el mejor ajuste de los parámetros, no las ecuaciones.

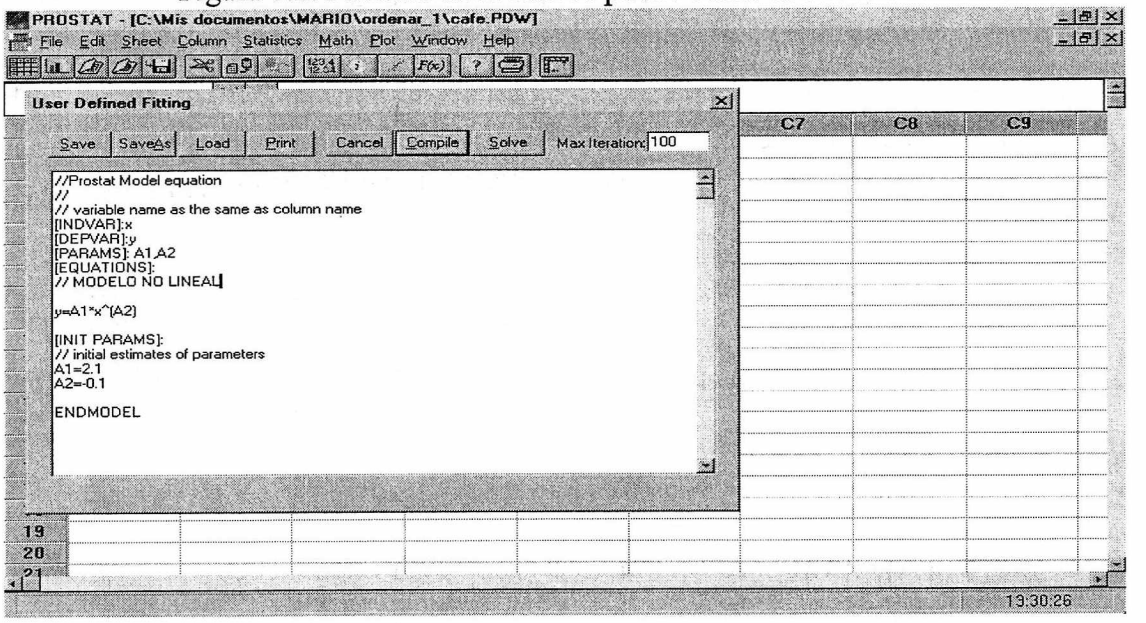
Los archivos de las ecuaciones del modelo usado para el ajuste no lineal queda en Prostat en un archivo de texto de ASCII con ciertas convenciones, los cuales pueden ser guardados en disco. La extensión sugerida es "*.eqn", pero puede usarse el nombre que se quiera. El archivo de la ecuación del modelo puede ser editado en el editor de ecuaciones en Prostat o en cualquier editor de texto.

Para compilar el modelo se elige el comando compile. Este comando revisa errores de sintaxis en el editor de ecuaciones, los cuales a menudo surgen por no usar el nombre de una columna como nombre de las variables o porque algunas palabras claves en los paréntesis cuadrados son modificados.

Después de que el modelo es cargado en la ventana que contiene el editor, se procede a hacer click en compile o presionar $< E >$ para compilar el modelo

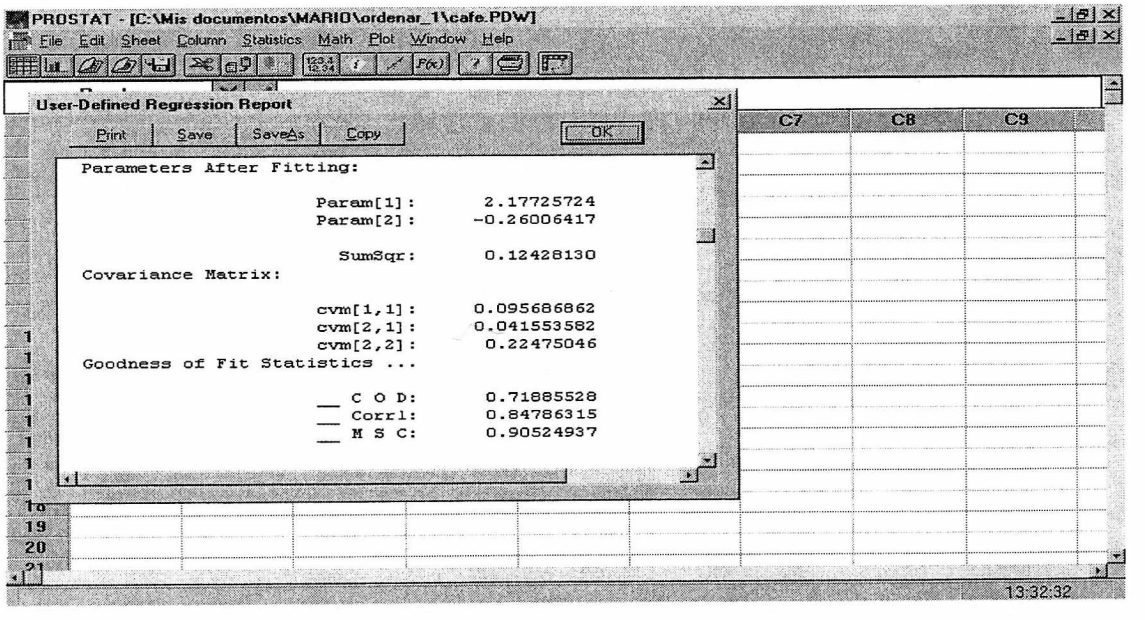
(checando errores de sintaxis). Si la compilación es exitosa, se procede a hacer click en el botón Solve para comenzar el proceso de ajuste.

Figura 6.2. Pantalla en la cual se puede editar la ecuación del modelo



Después de haber efectuado todos estos pasos. El programa proporciona un reporte con la principal información estadística para evaluar el ajuste del modelo.

Figura 6.3. Pantalla en la cual se presenta el reporte con información referente al ajuste del modelo.



CONCLUSIONES

Dentro del contexto del modelo estadístico lineal, los planteamientos sobre estimación e inferencia tanto bajo el criterio de máxima verosimilitud como de mínimos cuadrados tienen interesantes implicaciones cuando se considera el supuesto de normalidad del vector aleatorio **e**.

Por otra parte, por no poder obtener una expresión explícita para el estimador de mínimos cuadrados en el contexto de modelos no lineales, hay que resolver el sistema de ecuaciones normales o el problema de optimización de que éstas proceden por métodos numéricos (algoritmos). Además, ya que el estimador mínimo cuadrático no lineal **b** es una complicada función no lineal de **y**, es imposible establecer sus propiedades muestrales finitas para la amplia variedad de modelos no lineales que pueden existir. En particular, las propiedades de mejor estimador linealmente insesgado no pueden llevarse a cabo del o desde el modelo lineal. Sin embargo, es posible considerar sus propiedades asintóticas. Fue mostrado que bajo apropiadas condiciones, el estimador mínimo cuadrático no lineal **b** es un estimador consistente, y que $\sqrt{T}(b-\beta)$ tiene una distribución normal en el límite con media cero y varianza $\sigma^2[z(\beta)'z(\beta)/T]^{-1}$. Así para propósitos de inferencia b puede ser tratado posee una distribución aproximadamente normal con media β y varianza $\sigma^2[z(\beta)'z(\beta)]^{-1}$. La varianza es consistentemente estimada por

$$\widehat{var}(b) = \sigma^2[z(\beta)'z(\beta)]^{-1}$$

donde

$$\sigma^2 = \frac{S(b)}{T-1}.$$

Asimismo los procedimientos convencionales de estimación por intervalo y de pruebas de hipótesis pueden ser llevados a cabo de la forma usual (lineal), excepto que $X'X$ es reemplazada por $z(\beta)'z(\beta)$. Es decir, cuando el tamaño de muestra

crece, el estimador de mínimos cuadrados obtenido por alguno de los algoritmos numéricos mencionados, tiene una distribución normal, con expresiones para la media y para la varianza que se reducen a las ya conocidas en el caso del modelo lineal salvo que $X'X$ es reemplazada por $z(\beta)'z(\beta)$. De esta manera, como consecuencia de la distribución normal, los habituales contrastes de hipótesis mediante estadísticos t y F son válidos, si incluyen la existencia de un mínimo global para la función de suma de cuadrados del error, y la singularidad de la matriz $[lim\ z(\beta)'z(\beta)/T]$.

LITERATURA CITADA

Box , G. E. P., and D. R. Cox. 1964. An analysis of transformations, Journal of the Royal Statistical Society, Series B, 211-264.

Chiang, A. C. 1987. Métodos fundamentales de economía matemática, México, D.F., McGraw-Hill.

Christ, C. F. 1974. Modelos y métodos econométricos, México, D.F., Limusa.

Davidson, R. and J. G. MacKinnon. 1993. Estimation and inference in econometrics, New York, Oxford University Press, Inc.

Draper, N. and Smith. 1980. Applied Regresion Analysis, New York, John Wiley & Sons, Inc.

Eisenpress, H. and J. Greenstadt. 1966. The estimation of nonlinear econometric systems, Econometrica, 34, 851-61.

Greene, W. H. 1998. Econometric Analysis, Printice Hall, Inc.

Gujarati, D. N. 1997. Econometría, Santa Fé de Bogotá, 3th. Ed., McGraw-Hill.

Hartley, H. O. 1961. The modified Gauss-Newton method for the fitting of the nonlinear regression functions by least squares, Technometrics, 3, 269-80.

Henderson, J. M. and R. E. Quandt. 1980. Microeconomic theory, Singapore, McGraw-Hill.

Johnston, J. And J. Dinardo. 1997. Econometric methods, Singapore, 4th. Ed., McGraw-Hill.

Kmenta, J. 1971. Elements of econometrics, New York, Macmillan Publishing Co. Inc.

Maddala, G. S. 1988. Econometría, México, D.F., McGraw-Hill.

Nakamura, S. 1997. Análisis numérico y visualización gráfica, México, D.F., Printice Hall, Inc.

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