Estimación de modelos no lineales

Alfonso Novales
Departamento de Economía Cuantitativa
Universidad Complutense

Enero 2016
Versión preliminar

Contents

1 Introducción

Es bien conocido que el estimador de Mínimos Cuadrados Ordinarios de un modelo de relación lineal,

$$y_t = x_t'\beta + u_t, t = 1, 2, ..., T$$

viene dado por la expresión matricial,

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

siendo X la matriz Txk que tiene por columnas las T observaciones de cada una de las k variables explicativas contenidas en el vector x_t, e Y el vector columna, de dimensión T, formado por las observaciones de y_t. Este estimador, que es lineal (función lineal del vector Y), es insesgado. Es el de menor varianza

entre los estimadores lineales si la matriz de covarianzas de los términos de error tiene una estructura escalar,

$$Var(u) = \sigma_u^2 I_T$$

Si, además de tener dicha estructura de covarianzas, el término de error tiene una distribución Normal, entonces el estimador de Mínimos Cuadrados coincide con el estimador de Máxima Verosimilitud, siendo entonces eficiente: estimador de menor varianza, entre todos los estimadores insesgados, sea cual sea su dependencia respecto del vector de Y.

Supongamos que se pretende estimar la relación,

$$y_t = f(x_t, \beta) + u_t, \tag{1}$$

donde $f(x_t, \beta)$ es una función no lineal de los componentes del vector $kx1$, β.

El interés de un modelo no lineal es que rompe con una limitación del modelo lineal, que es que el efecto de un cambio unitario en una variable explicativa x_t sobre la variable dependiente, es constante: $dy_t/dx_t = \beta$.

Si $f(x_t, \beta)$ es no lineal únicamente en las variables explicativas x_t, un cambio de variable permite transformar el modelo anterior en un modelo lineal. Excluimos, sin embargo, inicialmente, la estimación de relaciones implícitas, representables a partir de un modelo general del tipo,

$$g(y_t, x_t, \beta) + u_t,$$

aunque pueden en muchos casos estimarse siguiendo los mismos procedimientos que explicamos en este capítulo.

Conviene observar que las posibles dificultades en estimación, y la necesidad de utilizar procedimientos adecuados para el tratamiento de modelos no lineales surge cuando el modelo es no lineal en los parámetros. Es decir, nolinealidades en las variables del modelo, por sí solas, no generan ninguna dificultad, y no requieren procedimientos especiales de estimación. Son situaciones que se reducen a modelos lineales mediante un cambio de variable apropiado. Por ejemplo, para estimar el modelo:

$$y_t = \alpha + \beta \frac{1}{1 - e^{x_t}} + u_t$$

si hacemos el cambio de variable: $z_t = \frac{1}{1-e^{x_t}}$, tenemos un modelo lineal: $y_t = \alpha + \beta z_t + u_t$, que se estima por minimos cuadrados ordinarios, sin ninguna dificultad.

Otros ejemplos de tales modelos:

$$\begin{aligned} y_t &= \alpha + \beta \frac{1}{1 - \ln x_t} + u_t \\ \sqrt{y_t} &= \alpha + \beta_1 e^{x_t} + \beta_2 z_t + u_t \end{aligned}$$

En el primer modelo, el cambio de variable: $\tilde{x}_t = \frac{1}{1-\ln x_t}$ transforma el modelo en lineal, al igual que sucede en el segundo modelo si hacemos: $\tilde{y}_t = \sqrt{y_t},\ \tilde{x}_t = e^{x_t}$. Dentro de este grupo, un modelo interesante es:

$$y_t = \alpha + \beta_1 x_t + \beta_2 x_t^2 + u_t$$

que puede generar distintos tipos de relación entre y_t y x_t, en función de los signos y magnitudes de β_1 y β_2. Si $\beta_2 > 0$, la dependencia será de acuerdo con una función convexa, como en el gráfico, siendo cóncava si $\beta_2 < 0$.

La relación podria ser estrictamente creciente o decreciente en el rango de valores admisibles de x_t (por ejemplo, si $x_t > 0$). Cuando se trabaja con funciones no lineales es siempre importante analizar las características de la dependencia entre ambas variables, tomando derivadas en la función $y_t = f(x_t, \theta)$. Por ejemplo, en esta última función, tenemos:

$$\frac{\partial y_t}{\partial x_t} = \beta_1 + 2\beta_2 x_t$$

que depende del valor numérico de x_t, al contrario de lo que sucede en un modelo de regresión lineal. Esto significa que el modelo implica que el impacto que sobre y_t tiene una determinada variación en el nivel de x_t dependen del nivel de esta última variable. El impacto que sobre y_t tiene una elevación de 1 unidad en x_t no es el mismo si $x_t = 10$ que si $x_t = 100$.

Será frecuente encontrar situaciones en que $\beta_2 < 0$, que significan que y_t crece o decrece cuando x_t varia, pero lo hace *menos que proporcionalmente* a dicha variación. En ese caso, la función será creciente para valores de $x_t < \frac{\beta_1}{2\beta_2}$.

Comenzamos analizando la estimacion de algunos modelos no lineales que no precisan para su estimación de métodos especificos. Son modelos en los que puede diseñarse una estrategia de estimación que utilice unicamente técnicas de estimación de modelos lineales, como es el método de Minimos Cuadrados Ordinarios.

2 Algunos modelos no lineales típicos

2.1 Modelo potencial

Una especificación muy natural acerca de la relación no lineal entre variables es:

$$y_t = \alpha + \beta x_t^{\gamma} + u_t, \tag{2}$$

que se reduce a una relación lineal: $y_t = \alpha + \beta x_t + u_t$, bajo la restricción $\gamma = 1$. Es decir, el modelo de relación lineal entre y_t y x_t : $y_t = \alpha + \beta x_t + u_t$, es una versión restringida del modelo (2) .

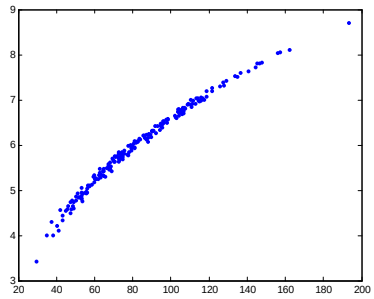
Siendo una extension natural del caso lineal, este modelo es muy apropiado para analizar posibles no linealidades en la relación entre ambas variables, una vez que se ha estimado un modelo lineal. Puede utilizarse asimismo para analizar el caracter no lineal del efecto de una variable explicativa x_t sobre y_t

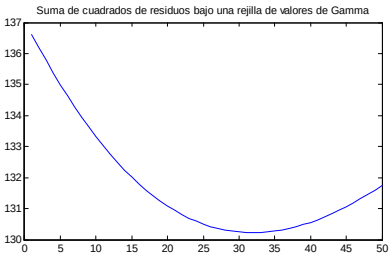
en una regresión múltiple. Una vez estimado el modelo, el contraste de linealidad equivale a contrastar la hipótesis nula: $H_0 : \gamma = 1$ frente a la hipótesis alternativa $H_1 : \gamma \neq 1$.

En este modelo, tenemos:

$$\frac{dy_t}{dx_t} = \beta x_t^{\gamma} (\ln x_t)$$

El programa *Simul_ estim.m* genera datos simulados y estima luego el modelo potencial de esta sección: a) utilizando las condiciones de optimalidad del estimador de minimos cuadrados, b) mediante una rejilla de valores de γ, c) utilizando un algoritmo numérico de optimización. Para ello, se genera primero una variable explicativa con estructura: $x_t = \bar{x} + u_x$, para luego generar datos de y_t. Con valores paramétricos: $\bar{x} = 6, u_x \sim N(0,1), \alpha = 10, \beta = 0.8, \gamma = 2.5, u_t \sim N(0, 2^2)$, en una determinada simulación se tiene la nube de puntos (x, y) :

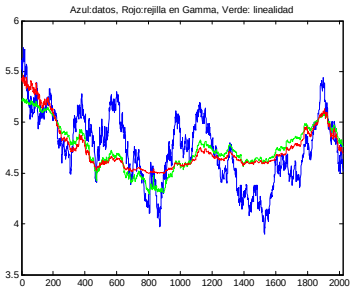




Mediante la evaluación de las condiciones de primer orden en una rejilla de valores de γ, tenemos: $\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}\right) = 12.5211; 0.6548; 2.59)$, con Suma de Cuadrados de Residuos: SCR=652.6104. Cuando utilizamos una rejilla de valores de γ entre 0.1 y 5.0, alcanzamos el menor valor numérido de la Suma de Cuadrados de Residuos en $\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}\right) = 12.7913; 0.6404; 2.60)$, con Suma de Cuadrados de Residuos: SCR=652.8019. El estadístico F para el contraste de la hipótesis de linealidad arroja un valor numérico: $F = 1191,5$, conduciendo a un rechazo claro de dicha hipótesis nula. Por último el uso de la rutina "fminunc.m" de Matlab, conduce a $\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}\right) = 12.4825; 0.6569; 2.5886)$. con SCR = 652.6074.

Que todos los procedimientos conduzcan a una estimación tan similar se debe, en parte, a que la nube de puntos generada es bastante suave, con poco error muestral (poca dispersión), y una clara curvatura. Si aumentamos la dispersión, los resultados pueden cambiar.

Realizamos asimismo la estimación de un modelo potencial para recoger la relación entre tipos a corto plazo y a largo plazo, utilizando tipos de interés diarios a 1 y 10 años, para UK, contenidos en el archivo: PCA_Spot_Curve.xls. Utilizando una rejilla de valores de γ para evaluar el valor numérico de las condiciones de primer orden del problema de minimización de la Suma de Cuadrados de Residuos obtuvimos: $\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}\right) = 4.3070; 0.0029; 3.19)$, con Suma de Cuadrados de Residuos: SCR=130.2204, mientras que utilizando una rejilla de valores de γ y estimando el modelo lineal que se obtiene al condicionar en cada valor numérico de γ, obtenemos: $\left(\hat{\alpha}, \hat{\beta}, \hat{\gamma}\right) = 4.3084; 0.0028; 3.20)$, con Suma de Cuadrados de Residuos: SCR=130.2205. El estadístico F para el contraste de la hipótesis de linealidad es 47.9544, rechazando con claridad dicha hipótesis.



2.2 Regresión por umbrales

2.2.1 Contraste de Chow

Es bien conocido el contraste de Chow para analizar la posible existencia de cambio estructural en un modelo de regresión lineal. Cuando se sospecha que el modelo ha podido variar a partir de un determinado momento, conviene estimar el modelo dos veces, con la submuestra previa a dicho instante, y con la submuestra posterior al mismo. El test de Chow consiste en evaluar si hay suficiente evidencia acerca de que las estimaciones paramétricas con ambas submuestras son diferentes entre sí. Para ello, compararemos las estimaciones obtenidas con las submuestras, con la que obtendríamos con la muestra completa. Si concluimos que no existe dicha evidencia empírica, pensaremos que no ha habido cambio estructural. El modelo restringido es el que estima con toda la muestra, mientras que le modelo sin restringir es el que considera una ecuación distinta para cada submuestra. La Suma de Cuadrados de residuos de este modelo es el agregado de la Suma de Cuadrados de residuos con ambas submuestras, anterior y posterior al posible momento de cambio estructural. El test de Chow tiene la

forma del test F clásico:

$$\frac{SCRR - SCRS}{SCRS}\frac{T-k}{q}$$

donde como siempre, el número de restricciones q será igual al número de variables explicativas (contando la constante del modelo), mientras que k es el número de parámetros estimados en el modelo Sin Restringir, es decir, el doble de q.

Para este contraste hay que fijar un posible instante en el que se hubiera producido el cambio estructural, y el resultado del contraste depende de la elección de dicho instante de tiempo, que se utiliza para dividir la muestra en las dos submuestras mencionadas.

2.2.2 Switching regressions con probabilidades exógenas

Supongamos que queremos estimar la relación:

$$y_t = x_t'\beta + u_t, t = 1, 2, ..., T$$

en la que suponemos que los parámetros β no han permanecido constantes alo largo dela muestra. Evidentemente, cuando ese es el caso, hay muchas maneras en que los β han podido variar, y no podriamos estimar el modelo salvo si establecemos un determinado supuesto acerca del modo en que los parámetros β han variado a lo largo de la muestra.

La regresión por umbrales, o modelo *switching regressions con probabilidades exógenas* surge si estamos dispuestos a suponer que el vector β solo ha tomados dos valores posibles a lo largo de la muestra, y que ello depende de los valores que ha tomado una determinada variable z. Así, suponemos que:

$$\begin{aligned} \beta &= \beta_1 \text{ si } z_t < z^* \\ \beta &= \beta_2 \text{ si } z_t > z^* \end{aligned}$$

La variable z puede ser una de las variables que integran el vector x_t, o no formar parte del mismo.

Los parámetros a estimar son $2k + 1 : (\beta_1, \beta_2, z^*)$, y la estimación es condicional en nuestra elección de la variable z_t que determina el cambio de régimen. Para estimar el modelo, condicional en un determinado valor numérico z^*, dividimos la muestra en dos submuestras, según que $z_t < z^*$ o $z_t > z^*$, y estimamos dos regresiones:

$$\begin{aligned} y_t &= x_t'\beta_1 + u_t, \text{ con la submuestra de observaciones en que } z_t < z^* \\ y_t &= x_t'\beta_2 + v_t, \text{ con la submuestra de observaciones en que } z_t > z^* \end{aligned}$$

Si agregamos las Sumas de Cuadrados de Residuos obtenidas en ambas regresiones: $SCR = SCR_1 + SCR_2$, tendremos la suma de cuadrados de residuos

de este modelo de dos regímenes. Indudablemente, la calidad del ajuste y, con ello, el valor numérico de SCR dependerá de la partición que hayamos hecho en la muestra, es decir, del valor numérico z que hayamos fijado inicialmente.

Lógicamente, dicho valor numérico no debería estar fijado. Lo que hacemos es repetir el procedimiento para distintos valores numéricos de z^* comprendidos entre $\min(z_t)$ y $\max(z_t)$ y observar para qué valor numérico de z^* se obtienen un valor menor de SCR. Esa será la estimación de z^*. No es preciso hacer nada más, pues las estimaciones de β_1 y β_2 serán las que hayamos obtenido para el valor numérico z^* que minimiza el valor numérico de SCR.

Por último, el procedimiento descrito es condicional en la elección de una determinada variable z_t que condiciona el cambio de régimen. Pero puede haber distintas elecciones alternativas para dicha variable. Podemos tomar otra variable w_t y repetir el procedimiento. Al final, comparamos $SCR(z^*)$ y $SCR(w^*)$,donde z^* y w^* son los valores numéricos de z_t y w_t que minimizan la función SCR en cada caso.

Otra cuestion abierta en la especificación del modelo es si la condición de cambio de régimen depende de un valor observable en t o en el pasado, por ejemplo, en $t-1$. La diferencia es importante cuando se trata de predecir el proceso, por cuanto que si la condición es del tipo $z_{t-1} < c$, entonces en el instante T sabemos cuál será el régimen vigente en $T+1$. Por el contrario, si la condición es del tipo $z_t < c$, entonces tendríamos que predecir en T el valor numérico de z_{T+1} y, en base, al mismo, optar por un régimen u otro. En este caso, los errores de predicción de z se añadirían al resto de errores de especificación para determinar el error de prediccion total.

Cuando la regresion tiene la forma de un modelo autoregresivo, se conoce como modelo TAR (Threshold Autoregression):

$$\begin{aligned} y_t &= \phi_{0,1} + \phi_{1,1} y_{t-1} + \varepsilon_t,\ Var(\varepsilon_t) = \sigma_1^2,\ si\ z_t < c \\ y_t &= \phi_{0,2} + \phi_{1,2} y_{t-1} + \varepsilon_t,\ Var(\varepsilon_t) = \sigma_2^2,\ si\ z_t \geq c \end{aligned}$$

aunque tambien podria tener la innovación ε_t una estructura GARCH:

$$\begin{aligned} y_t &= \phi_{0,1} + \phi_{1,1} y_{t-1} + \varepsilon_t,\ Var(\varepsilon_t) = \sigma_t^2, \\ \sigma_t^2 &= \delta_{0,1} + \delta_{1,1}\sigma_{t-1}^2 + \delta_{2,1}\varepsilon_{t-1}^2\ si\ z_t < c \\ y_t &= \phi_{0,2} + \phi_{1,2} y_{t-1} + \varepsilon_t,\ Var(\varepsilon_t) = \sigma_t^2, \\ \sigma_t^2 &= \delta_{0,2} + \delta_{1,2}\sigma_{t-1}^2 + \delta_{2,2}\varepsilon_{t-1}^2\ si\ z_t \geq c \end{aligned}$$

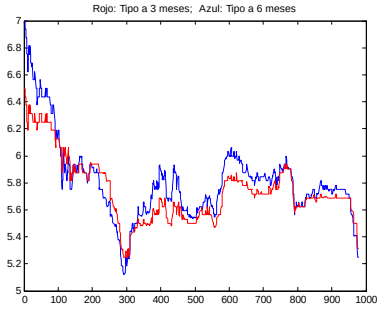
Un caso particular interesante surge cuando la variable que determina los cambios de régimen es un retardo de la propia variable dependiente, y_{t-d}, para algun valor $d > 0$, modelo que se conoce como SETAR (Self-Exciting Threshold Autoregression). Por ejemplo, con $d = 1$:

$$\begin{aligned} y_t &= \phi_{0,1} + \phi_{1,1} y_{t-1} + \varepsilon_t,\ Var(\varepsilon_t) = \sigma_1^2,\ si\ y_{t-1} < c \\ y_t &= \phi_{0,2} + \phi_{1,2} y_{t-1} + \varepsilon_t,\ Var(\varepsilon_t) = \sigma_2^2,\ si\ y_{t-1} \geq c \end{aligned}$$

o, con una estructura GARCH para la innovación ε_t :

$$\begin{aligned} y_t &= \phi_{0,1} + \phi_{1,1} y_{t-1} + \varepsilon_t,\ Var(\varepsilon_t) = \sigma_t^2, \\ \sigma_t^2 &= \delta_{0,1} + \delta_{1,1}\sigma_{t-1}^2 + \delta_{2,1}\varepsilon_{t-1}^2 \ si\ y_{t-1} < c \\ y_t &= \phi_{0,2} + \phi_{1,2} y_{t-1} + \varepsilon_t,\ Var(\varepsilon_t) = \sigma_t^2, \\ \sigma_t^2 &= \delta_{0,2} + \delta_{1,2}\sigma_{t-1}^2 + \delta_{2,2}\varepsilon_{t-1}^2 \ si\ y_{t-1} \geq c \end{aligned}$$

Ejemplo: Utilizando datos diarios de tipos de interés con vencimientos a 3 y 6 meses (Tsay), llevamos a cabo la estimación del modelo de capacidad predictiva del tipo forward.



Para ello, comenzamos estimando el tipo forward mediante: $1 + F_{t,3,6}^3/4 = \frac{1+r_{t,6}/200}{1+r_{t,3}/400}$, donde estamos trasladando el tipo a 6 meses a un período semestral, y el tipo a 3 meses a un periodo trimestral. El tipo forward resultante sería aplicable a un periodo trimestral, por lo que habrá que multiplicarlo por 4 para obtenerlo en términos anuales. Si utilizamos composición multiplicativa para los tipos, estimaríamos el tipo Forward: $1 + F_{t,3,6}^3 = \frac{(1+r_{t,6}/100)^2}{1+r_{t,3}/100}$, que es lo que se hace en el programa *Regresion_ umbrales.m* .

La hipótesis que queremos analizar es si este tipo forward adelanta al tipo de contado a 3 meses que estará vigente en el mercado dentro de 3 meses, mediante la relación:

$$r_{t,3} = \alpha + \beta F_{t-3,3,6}^3 + u_t$$

donde debe notarse el retraso de 3 meses introducido en el tipo forward. De hecho, como contamos con datos diarios y consideramos 21 dias por mes, se trata de retrasar el tipo forward 63 observaciones en dicha relación.

Estimamos primero el modelo suponiendo que la relación tiene dos regímenes diferentes, dependiendo del valor que toma la pendiente de la curva, es decir, el diferencial entre ambos tipos a 6 y 3 meses. El resultado es:

$$
\begin{aligned}
r_{t,3} &= \underset{(12,17)}{2,6443} + \underset{(14,11)}{0,5106} F^3_{t-3,3,6},\ si\ r_{6,t} - r_{3,7} < -0,025 \\
r_{t,3} &= \underset{(56,55)}{3,6777} + \underset{(31,26)}{0,3450} F^3_{t-3,3,6},\ si\ r_{6,t} - r_{3,7} > -0,025
\end{aligned}
$$

habiendo 159 observaciones en el primer régimen, que corresponde a una curva de tipos invertida, con el tipo a 6 meses por debajo del tipo a 3 meses, y 755 observaciones en el régimen de curva normal. Como se aprecia en el gráfico anterior, que muestra ambos tipos de interés, el régimen de curva de tipos invertida se produjo durante un intervalo de tiempo en la primera parte de la muestra. La Suma de Cuadrados de Residuos agregada de ambos regímenes es de 14,9073, y el estadístico F para el contraste de linealidad, es decir, de existencia de un solo régimen, toma el valor de 21,50, rechazando claramente dicha hipótesis.

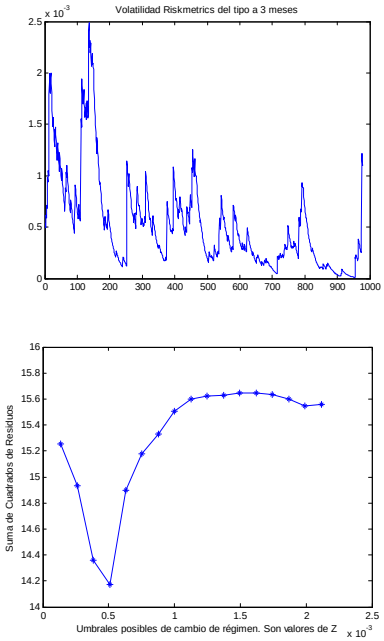
Volvemos a estimar el modelo bajo el supuesto de que es la volatilidad de los tipos de interés a corto plazo quien determina las características de la relación. Para ello hemos de comenzar estimando una serie temporal de volatilidad, pues dicha variable no es observable, y ha de ser la volatilidad de la innovación en el tipo a 3 meses. El resultado será condicional en dicha estimación, y distintos modelos de volatildad arrojarán resultados diferentes. Nosotros utilizamos el modelo de Riskmetrics, de modo que estimamos:

$$
\begin{aligned}
r_{3,t} &= \delta_0 + \delta_1 r_{3,t-1} + \varepsilon_t \\
\sigma^2_t &= 0,94\sigma^2_{t-1} + 0,06\varepsilon^2_t,\ con\ \sigma^2_0 = var(\varepsilon^2_t)
\end{aligned}
$$

El resultado es:

$$
\begin{aligned}
r_{t,3} &= \underset{(36,72)}{4,0908} + \underset{(14,44)}{0,2763} F^3_{t-3,3,6},\ si\ \sigma^2_t < 18,42\% \\
r_{t,3} &= \underset{(35,48)}{3,1062} + \underset{(29,93)}{0,4328} F^3_{t-3,3,6},\ si\ \ \sigma^2_t > 18,42\%
\end{aligned}
$$

habiendo 564 en el régimen de baja volatilidad y 350 observaciones en el régimen de alta volatilidad. La Suma de Cuadrados de Residuos agregada de ambos regímenes es de 14,1742, y el estadístico F para el contraste de linealidad, es decir, de existencia de un solo régimen, toma el valor de 42,81, rechazando claramente dicha hipótesis. El ajuste es algo mejor que bajo el supuesto de que los regímenes vienen determinados por el diferencial entre los tipos a 6 y 3 meses, pero la evidencia en contra de un solo régimen es ahora aún más clara.



2.2.3 Switching Markov regression

El modelo tiene la forma:

$$\begin{aligned} y_t &= \alpha_1 + \beta_1 x_t + u_{1t},\ u_{1t} \sim N(0, \sigma_1^2) \text{ en el estado 1} \\ y_t &= \alpha_2 + \beta_2 x_t + u_{2t},\ u_{2t} \sim N(0, \sigma_2^2) \text{ en el estado 2} \end{aligned}$$

que podemos representar en función de una variable latente s_t que toma el valor 1 si estamos en el estado 1, y toma el valor 2 si estamos en el estado 2:

$$y_t = \alpha_{s_t} + \beta_{s_t} x_t + u_{s_t t},\ u_{s_t t} \sim N(0, \sigma_{s_t}^2)$$

con matriz de probabilidades de transición:

$$\Pi = \begin{pmatrix} \pi_{11} & \pi_{21} \\ \pi_{12} & \pi_{22} \end{pmatrix} = \begin{pmatrix} \pi_{11} & 1 - \pi_{22} \\ 1 - \pi_{11} & \pi_{22} \end{pmatrix}$$

siendo π_{ij} la probabilidad de estar en el estado j en el instante $t+1$, habiendo estado en el estado i en el instante t.

La probabilidad incondicional de estar en el régimen 1 es:

$$P(s_t = 1) = \frac{1 - \pi_{22}}{2 - \pi_{11} - \pi_{22}}$$

mientras que la probabilidad incondicional de estar en el régimen 1 es:

$$P(s_t = 2) = \frac{1 - \pi_{11}}{2 - \pi_{11} - \pi_{22}}$$

y el vector de parámetros a estimar es: $\theta = \left(\alpha_1, \alpha_2, \beta_1, \beta_2 \sigma_1^2, \sigma_2^2, \pi_{11}, \pi_{22}\right)$.

La cadena de Markov se representa mediante un vector aleatorio de indicadores de estado, ξ_t, cuyo elemento i-ésimo es igual a 1 si se produce el estado i en dicho período, y es igual a 0 en caso contrario.

En una cadena con dos estados tendremos:

$$\xi_t = \begin{pmatrix} \xi_t^1 \\ \xi_t^2 \end{pmatrix} = \begin{cases} \begin{pmatrix} 1 \\ 0 \end{pmatrix} & \text{si se produce el estado 1 en el periodo } t \\ \begin{pmatrix} 0 \\ 1 \end{pmatrix} & \text{si se produce el estado 2 en el periodo } t \end{cases}$$

Pero los estados no son observables, por lo que unicamente podemos asignar probabilidades de estar en uno u otro régimen, condicionales en la información disponible hasta ese instante.

La esperanza del vector ξ_t que indica el estado en t, condicional en la información disponible hasta $t-1$ se denota por $\xi_{t|t-1}$, y por la definición de la matriz de transición, se puede probar que:

$$\xi_{t|t-1} = E_{t-1}(\xi_t) = \Pi \xi_{t-1}$$

El modelo se estima por Máxima Verosimilitud, procedimiento que se simplifica en gran medida si se supone Normalidad de los errores del modelo en cada estado. Denotamos en lo sucesivo la función de densidad Normal por $\varphi(x; \mu, \sigma^2)$.

Para el algoritmo iterativo, fijamos como condiciones iniciales: $\hat{\xi}_{1|0} = \begin{pmatrix} \hat{\xi}_{1|0}^1 \\ \hat{\xi}_{1|0}^2 \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}$ ó $\hat{\xi}_{1|0} = \begin{pmatrix} \hat{\xi}_{1|0}^1 \\ \hat{\xi}_{1|0}^2 \end{pmatrix} = \begin{pmatrix} 0 \\ 1 \end{pmatrix}$, y los valores numéricos de los parámetros los tomamos inicialmente iguales a las estimaciones de mínimos cuadrados de un modelo con un solo régimen $\hat{\alpha}_1 = \hat{\alpha}_2, \hat{\beta}_1 = \hat{\beta}_2, \hat{\sigma}_1^2 = \hat{\sigma}_2^2$. Habitualmente, se toma asimismo: $\hat{\pi}_{11} = \hat{\pi}_{22} = 0,50$.A partir de estas condiciones iniciales, iteramos:

- construimos la función de densidad: $f_t(y_t/x_t; \hat{\theta}) = \hat{\xi}_{t|t-1}^1 . \varphi(y_t; \hat{\alpha}_1 + \hat{\beta}_1 x_t, \hat{\sigma}_1^2) + \hat{\xi}_{t|t-1}^2 . \varphi(y_t; \hat{\alpha}_2 + \hat{\beta}_2 x_t, \hat{\sigma}_2^2)$
- $\hat{\xi}_{t|t} = \begin{pmatrix} \hat{\xi}_{t|t}^1 \\ \hat{\xi}_{t|t}^2 \end{pmatrix} = \begin{pmatrix} \frac{\hat{\xi}_{t|t-1}^1 . \varphi(y_t; \hat{\alpha}_1 + \hat{\beta}_1 x_t, \hat{\sigma}_1^2)}{f_t(y_t/x_t; \hat{\theta})} \\ \frac{\hat{\xi}_{t|t-1}^2 . \varphi(y_t; \hat{\alpha}_2 + \hat{\beta}_2 x_t, \hat{\sigma}_2^2)}{f_t(y_t/x_t; \hat{\theta})} \end{pmatrix}$
- $\xi_{t+1|t} = \Pi \xi_{t|t}$
- Repetimos hasta $t = T$

Estas iteraciones nos proporcionan el conjunto de densidades condicionales $\left\{f_t(y_t/x_t;\hat{\theta})\right\}_{t=1}^{T}$, así como un conjunto de probabilidades condicionales de los estados: $\left\{\hat{\xi}_{t|t}\right\}_{t=1}^{T}$.

A continuación, los parámetros del modelo se estiman resolviendo:

$$\underset{\theta}{Max} \ln L(\theta) = \sum_{t=1}^{T} \ln f_t(y_t/x_t;\theta)$$

Una página Web muy interesante sobre software para estos modelos es: https://sites.google.com/site/marceloperlin/matlab-code/classical-pairs-trading-using-matlab

2.3 Simulación de modelos

2.3.1 Simulación de un modelo de regresión por umbrales

Para simular una regresión por umbrales, debemos utilizar una variable aleatoria que nos indique en qué regimen estamos en cada período. Consideremos que hemos estimado un modelo autoregresivo:

$$r_t = \alpha + \beta r_{t-1} + u_t, t = 1, 2, ..., T$$

con

$$\begin{aligned} \alpha &= \alpha_1, \beta = \beta_1 \text{si } z_t < z^* \\ \alpha &= \alpha_2, \beta = \beta_2 \text{si } z_t > z^* \end{aligned}$$

para el que hemos estimado: $\hat{\alpha}_1, \hat{\alpha}_2, \hat{\beta}_1, \hat{\beta}_2, \hat{\sigma}_1^2, \hat{\sigma}_2^2, \hat{z}^*$.Ahora queremos simular dicho proceso a partir del final de la muestra. Esto es lo que queremos hacer para la estimación del Valor en Riesgo, por ejemplo.

Para ello, necesitamos un supuesto acerca de la evolución temporal de la variable Z_t así como un supuesto acerca de la distribución de probabilidad seguida por su innovación. Supongamos que Z_t es independiente y que sigue una distribución Normal. Estimariamos la esperanza matemática y la varianza de dicha distribución a partir de la serie temporal de Z_t.

A continuación, para cada período $T+1, ..., T+h$,simularíamos un valor numérico de dicha distribución Normal, por ejemplo, mediante una realización $\{\varepsilon_t\}_{t=T+1}^{T+h}$ de una distribución $N(0,1)$ y haciendo: $Z_t = \hat{\sigma}_Z \varepsilon_t + \hat{\mu}_Z, t = T+1, T+2, ..., T+h$. Lógicamente, si supusiesemos una distribución diferente, muestrearíamos de dicha distribución de probabilidad. Alternativamente, podemos evitar hacer dicho supuesto y muestrar mediante *bootstrapping* a partir de las realizaciones muestrales observadas para $Z_t, t = 1, 2, ..., T$.

Si Z_t tuviese dependencia temporal, tendriamos que modelizar dicha estructura, por ejemplo, mediante: $Z_t = \delta_0 + \delta_1 Z_{t-1} + \xi_t, t = 1, 2, ..., T$, hacer un supuesto acerca de la distribución de probabilidad de la innovación ξ_t, y

seguir los pasos mencionados en el párrafo anterior para obtener una senda para $\xi_t, t = T+1, ..., T+h$,y por consiguiente, para $Z_t, t = T+1, ..., T+h$.

Si $Z_{T+1} < \hat{z}^*$, eso significa que en $T+1$ estamos en el régimen 1, y hariamos:

$$\hat{r}_{T+1} = \hat{\alpha}_1 + \hat{\beta}_1 r_T + \hat{\sigma}_1 \varepsilon_{T+1}$$

A continuación, vamos a $T+2$. Si $Z_{T+2} < \hat{z}^*$, eso significa que en $T+2$ estamos de nuevo en el régimen 1, y hariamos:

$$\hat{r}_{T+2} = \hat{\alpha}_1 + \hat{\beta}_1 \hat{r}_{T+1} + \hat{\sigma}_1 \varepsilon_{T+2}$$

Si, por el contrario, hubiese sido $Z_{T+2} > \hat{z}^*$, eso significaría que en $T+2$ estamos en el régimen 2, y haríamos:

$$\hat{r}_{T+2} = \hat{\alpha}_2 + \hat{\beta}_2 \hat{r}_{T+1} + \hat{\sigma}_2 \varepsilon_{T+2}$$

Como vemos, para estimar $\hat{r}_{T+j}$ en cada periodo, utilizamos el valor numérico $\hat{r}_{T+j-1}$ estimado el periodo anterior, con independencia de que estemos en el mismo régimen o que hayamos cambiado de régimen.

Se produciría una situación especiamente interesante cuando la variable Z_t que determina el cambio de régimen estuviera relacionada con alguna de las variables explicativas del modelo. Podríamos entonces modelizar o no dicha relación. Por ejemplo, si la modelizamos:

$$\begin{aligned} y_t &= \alpha_1 + \beta_{1,1} x_t + \beta_{2,1} w_t + \varepsilon_t,\ z_t < c \\ y_t &= \alpha_2 + \beta_{1,2} x_t + \beta_{2,2} w_t + \varepsilon_t,\ z_t \geq c \\ z_t &= \delta_0 + \delta_1 w_t + \delta_2 w_t^{\gamma} + u_t \end{aligned}$$

o, si no la modelizamos, pero condicionamos en una correlación estimada: $\rho(z_t, w_t) = \rho^*$, entonces bajo supuestos de Normalidad, si son aceptables, podríamos en un ejercicio de simulación obtener sendas (z_t, w_t) con ese nivel de correlación, ρ^*.

2.3.2 Simulación de un modelo GARCH

Supongamos que hemos estimado un modelo GARCH:

$$\begin{aligned} r_t &= \alpha + \beta r_{t-1} + u_t, u_t \sim N(0, \sigma_t^2), t = 1, 2, ..., T \\ \sigma_t^2 &= \delta_0 + \delta_1 u_{t-1}^2 + \delta_2 \sigma_{t-1}^2 \end{aligned}$$

y queremos simular una realización de dicha variable para $T+1, T+2, ..., T+h$. Al estimar el modelo, habemos generado residuos: $\{\hat{u}_1, \hat{u}_2, ..., \hat{u}_T\}$,así como una serie temporal de varianzas de la innovación: $\left\{\hat{\sigma}_1^2, \hat{\sigma}_2^2, ..., \hat{\sigma}_T^2\right\}$.

El modelo nos da la predicción de la varianza para $T+1$:

$$\hat{\sigma}^2_{T+1} = \delta_0 + \delta_1 \hat{u}^2_T + \delta_2 \hat{\sigma}^2_T \tag{3}$$

en función del último residuo muestral y de la última varianza muestral.

Tomamos ahora una extracción aletoria de una distribución N(0,1), que denotamos por ε_{T+1}, y hacemos:

$$\hat{u}_{T+1} = \hat{\sigma}_{T+1} \varepsilon_{T+1} \tag{4}$$

con lo que $\hat{u}_{T+1}$ se distribuye $N(0, \hat{\sigma}^2_{T+1})$. A continuación:

$$\hat{r}_{T+1} = \hat{\alpha} + \hat{\beta} r_T + \hat{u}_{T+1}$$

y ya tenemos la realización numérica para $\hat{r}_{T+1}$.

Ahora, volvemos a repetir el proceso para $T+2$:

$$\hat{\sigma}^2_{T+2} = \delta_0 + \delta_1 \hat{u}^2_{T+1} + \delta_2 \hat{\sigma}^2_{T+1}$$

donde utilizamos para $\hat{u}_{T+1}, \hat{\sigma}^2_{T+1}$ los valores numéricos obtenidos en (4) y (3) y el proceso se repite hasta $T+h$. Posteriormente, podríamos generar cuantas realizaciones quisiésemos para ese mismo período: $T+1, T+2, ..., T+h$.

2.3.3 Simulando un modelo GARCH con cambio de régimen (probabilidades exógenas)

Supongamos que hemos estimado un modelo GARCH con cambio de régimen con probabilidades exógenas:

$$\begin{aligned} r_t &= \alpha_1 + \beta_1 r_{t-1} + u_t, u_t \sim N(0, \sigma^2_{1t}), t = 1, 2, ..., T \\ \sigma^2_{1t} &= \delta_0 + \delta_1 u^2_{t-1} + \delta_2 \sigma^2_{1-1} \end{aligned}$$

si $z_t < \hat{z}^*$, y:

$$\begin{aligned} r_t &= \alpha_2 + \beta_2 r_{t-1} + v_t, v_t \sim N(0, \sigma^2_{2t}), t = 1, 2, ..., T \\ \sigma^2_{2t} &= w_0 + w_1 u^2_{t-1} + w_2 \sigma^2_{2t-1} \end{aligned}$$

si $z_t > \hat{z}^*$.

Una vez estimado el modelo con datos: $t = 1, 2, ..., T$, tendremos que generar una senda simulada para la variable z_t como hemos comentado más arriba, para $\{T+1, ..., T+h\}$. Extraeremos asimismo una realización $N(0,1)$ para cada uno de esos períodos.

Supongamos que $Z_{T+1} < \hat{z}^*$, de modo que en $T+1$ estamos en el régimen 1. Calcularíamos:

$$\hat{\sigma}^2_{1,T+1} = \delta_0 + \delta_1 \hat{u}^2_T + \delta_2 \hat{\sigma}^2_T$$

pues tanto $\hat{u}_T$ y $\hat{\sigma}_T^2$ han sido obtenidos al estimar el modelo con la muestra de datos $t = 1, 2, ..., T$. Como en $T + 1$ estamos en el régimen 1, tomariamos la realización $N(0, 1)$ muestreada para ese periodo, ε_{T+1}, y calcularíamos: $\hat{u}_{T+1} = \sigma_{1,T+1}\varepsilon_{T+1}$, y:

$$r_{T+1} = \hat{\alpha}_1 + \hat{\beta}_1 r_T + \hat{u}_{T+1}$$

Si en $T + 2$ tenemos que $Z_{T+1} > \hat{z}^*$, estando en el régimen 2, haríamos:

$$\hat{\sigma}_{2,T+2}^2 = w_0 + w_1\hat{u}_{T+1}^2 + w_2\hat{\sigma}_{T+1}^2$$

utilizando para $\hat{\sigma}_{T+1}^2$ el valor numérico que obtuvimos en la ecuación anterior para $T + 1 : \hat{\sigma}_{T+1}^2 = \hat{\sigma}_{1,T+1}^2$.

A continuación, calcularíamos: $\hat{u}_{T+2} = \hat{\sigma}_{2,T+2}\varepsilon_{T+2}$, y:

$$r_{T+2} = \hat{\alpha}_2 + \hat{\beta}_2 r_{T+1} + \hat{u}_{T+2}$$

Es decir, como vemos, tomamos en cada periodo los parámetros correspondientes al régimen en que nos encontramos, lo que viene determinado por la realización de la senda temporal de Z_t. Sin embargo, en cada periodo, los valores retardados de la varianza y de la innovación son los calculados para el período anterior, con independencia del régimen en el que estuviésemos en dicho período.

En distintas sendas, al diferir los valores simulados para $Z_t, t = T+1, ..., T+h$, tambien diferirian los regimenes en que nos hallamos en cada periodo. En todo caso, incluso sin cambio de régimen, los valores numéricos de las innovaciones serían diferentes para las distintas sendas, obteniendo así sendas diferentes.

Si lo que queremos es generar una serie temporal para $t = 1, 2, ..., T$ a partir de un modelo GARCH teorico, nos encontramos con que no podemos simular el primer dato, por la estructura autoregresiva que tiene la ecuación de la varianza, y la ecuación de la media, además, si es un proceso autoregresivo. Así, comenzaremos en $T = 2$. Primero extraemos una realización de una $N(0, 1)$, con $T - 1$ observaciones: $t = 2, 3, ..., T$.

Supongamos que $Z_2 < \hat{z}^*$, de modo que en $T = 2$ estamos en el régimen 1. Querríamos hacer:

$$\sigma_{1,2}^2 = \delta_0 + \delta_1 u_1^2 + \delta_2\sigma_{1,1}^2$$

pero desconocemos u_1 y $\sigma_{1,1}^2$. Lo natural seria sustituirlos por sus medias muestrales. Por tanto, haríamos: $u_1 = 0$,mientras que para $\sigma_{1,1}^2$ tomaríamos la varianza a largo plazo de dicho régimen: $\sigma_1^2 = \delta_0/(1 - \delta_1 - \delta_2)$. Por tanto:

$$\hat{\sigma}_{1,2}^2 = \hat{\delta}_0 + \hat{\delta}_2\hat{\delta}_0/(1 - \hat{\delta}_1 - \hat{\delta}_2)$$

Para calcular la rentabilidad $\hat{r}_2$ nos encontramos con que desconocemos $\hat{r}_1$. Por analogía con lo anterior, sustituimos $\hat{r}_1$ por su media de largo plazo, $\alpha_1/(1-\beta_1)$. A continuación:

$$\begin{aligned}\hat{u}_2 &= \hat{\sigma}_{1,2}\varepsilon_2, \text{ con } \varepsilon_2 \sim N(0,1)\\ \hat{r}_2 &= \hat{\alpha}_1 + \hat{\beta}_1\hat{\alpha}_1/(1-\hat{\beta}_1) + \hat{u}_2\end{aligned}$$

Nótese que ya, ni $\hat{u}_2$ es igual a cero, ni $\hat{r}_2$ es igual a su media de largo plazo. Para $t = 3$, si $Z_3 < \hat{z}^*$, hariamos:

$$\begin{aligned}\sigma^2_{1,3} &= \hat{\delta}_0 + \hat{\delta}_1\hat{u}_2^2 + \hat{\delta}_2\hat{\sigma}^2_{1,2}\\ \hat{u}_3 &= \hat{\sigma}_{1,3}\varepsilon_3\\ \hat{r}_3 &= \hat{\alpha}_1 + \hat{\beta}_1\hat{r}_2 + \hat{u}_3\end{aligned}$$

Si en $t = 3$ hubiesemos estado en el régimen 2, habriamos utilizado los parámetros de dicho régimen sin mayor complicación. Unicamente, el lector debe darse cuenta de que es preciso construir las series temporales de volatilidades, innovaciones y rentabilidades para todos los períodos, si bien en cada periodo seleccionaremos unas u otras en funcion del regimen que venga indicado por la realización numérica de la variable Uniforme para ese periodo.

2.4 Regresión cuantílica

Dado un cuantil q de la variable Y, la regresión cuantílica resuelve el problema:

$$\min_{(\alpha,\beta)} \sum_{t=1}^{T} \left(q - 1_{y_t \leq \alpha+\beta x_t}\right)\left(y_t - (\alpha + \beta x_t)\right)$$

o, equivalentemente:

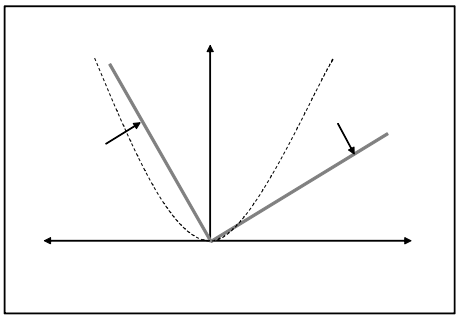
$$\min_{(\alpha,\beta)} \left[q \sum_{y_t \geq \alpha+\beta x_t} (y_t - (\alpha + \beta x_t)) - (1-q) \sum_{y_t \leq \alpha+\beta x_t} (y_t - (\alpha + \beta x_t)) \right]$$

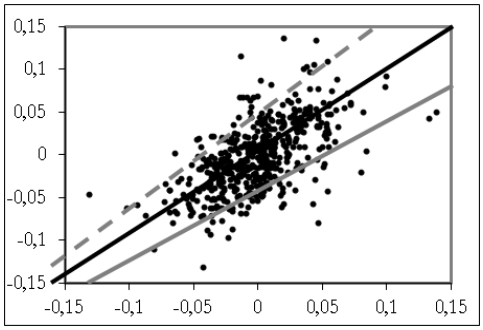
Nótese que la primera suma recoge residuos positivos, mientras que la segunda suma recoge los residuos negativos, de modo que ambas sumas entran positivamente en la función objetivo. Esta función generaliza el problema conocido como Mean Absolute Regression:

$$\min_{(\alpha,\beta)} \sum_{t=1}^{T} |y_t - (\alpha + \beta x_t)|$$

que utiliza como función de pérdida las dos bisectrices en los cuadrantes que aparecen en el gráfico izquierdo. Puede apreciarse en el gráfico izquierdo la función de pérdida cuadrática, del estimador MCO, así como la correspondiente a la regresión cuantílica. En ella se ve cómo para valores $q < 0.5$ se asigna

una mayor ponderación[1] a los residuos negativos que a los residuos positivos, sucediendo lo contrario cuando $q > 0,5$. En Finanzas, dado que habitualmente queremos cubrir el riesgo a la baja, este modelo suele utilizarse con valores de q reducidos.



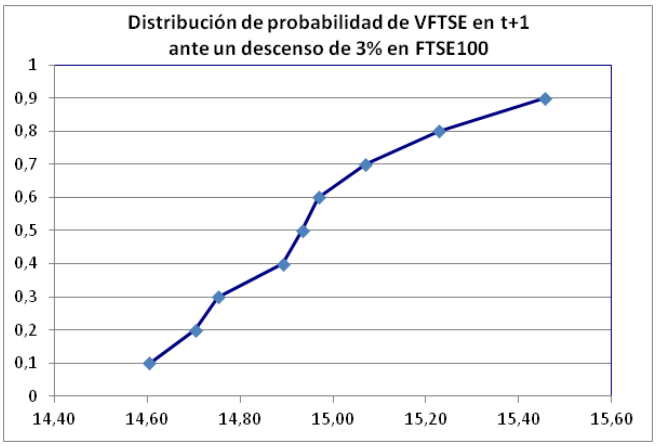


En el gráfico de la derecha se muestra el resultado de la regresión cuantílica para varios valores del cuantil q de Y. La línea punteada corresponde a $q = 0.1$, la línea gris corresponde a $q = 0.9$, y la linea central se corresponde con la mediana, es decir, con la Mean Absolute Regression. Es conocido que la recta de regresión minimocuadrática pasa por el punto $(\bar{y}, \bar{x})$. En cambio, la regresión cuantilica pasa por un cuantil de la nube de puntos. Si q es pequeño, por ejemplo, $q = 0.1$,entonces la mayoría de los puntos de la muestra quedará por debajo de la recta de regresión correspondiente al q-cuantil. Esto se debe a que los residuos negativos tienen asociado un peso muy importante en la función objetivo, por lo que los coeficientes estimados tenderán a generar pocos residuos negativos (puntos con un valor de Y inferior al esperado de acuerdo con la recta de regresión) y muchos puntos positivos (puntos con un valor de Y superior al esperado de acuerdo con la recta de regresión).

En el Case Study II.7.3.1 se analiza la relación entre la rentabilidad del indice FTSE100 y la rentabilidad del indice de volatilidad asociado, VFTSE. La regresión cuantílica arroja el resultado que se muestra en las primeras filas de la tabla. En la tabla se muestra asimismo el impacto estimado sobre la volatilidad de una caida del 3% en el índice FTSE100:

[1] Por ejemplo, si $q = 0,25$, los residuos negativos reciben una ponderación 3 veces superior a la que reciben los residuos positivos.

	Linear Quantile Regressions								
q	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9
alpha (Solver)	-0,039	-0,025	-0,016	-0,006	0,001	0,008	0,016	0,027	0,045
T-stat alpha	-21,815	-15,661	-9,716	-4,278	0,681	5,920	10,366	14,952	16,957
beta (Solver)	-6,010	-5,784	-5,609	-5,653	-5,525	-5,381	-5,388	-5,434	-5,403
T-stat beta	-22,465	-24,525	-23,263	-26,729	-27,687	-26,475	-23,975	-20,592	-13,622
VFTSE Rtn	14,10%	14,88%	15,26%	16,34%	16,67%	16,95%	17,74%	18,97%	20,76%
New VFTSE	14,60	14,70	14,75	14,89	14,93	14,97	15,07	15,23	15,46



La interpretación es que con probabilidad del 60%, el ascenso porcentual en el indice de volatilidad VFTSE sería superior al 16,34%, elevándose la volatilidad (a partir de un nivel de 12,8) en un nivel por encima del 14,89. El gráfico muestra la distribución de probabilidad de $VFTSE(t+1)$, a partir de su nivel actual $VFTSE(t) = 12,8$ en el supuesto de que FTSE100 cayera un 3% en $t+1$.

Aunque hemos extendido la regresión habitual en el sentido de permitir que los parametros estimados cambien entre distintas submuestras, todavia estamos imponiendo una relación lineal entre ambos índices. Sin embargo, cuando las variables X e Y se relacionan a través de una distribución distinta de la Normal, la relación entre ambas será no lineal.

Para ello, nos apoyamos en la teoría de cópulas. Si ambas variables tienen distribución Normal y se relacionan mediante una cópula Normal, la curva cuantil es:

$$Y = \rho X + \sqrt{1-\rho^2}\Phi^{-1}(q)$$

que es, efectivamente, lineal.

Sin embargo, si suponemos que las variables X, Y siguen distribuciones marginales F_1, F_2, que previamente hemos especificado y estimado por Máxima Verosimilitud, la curva cuantil de una cópula Normal es:

$$Y = F_2^{-1}\left[\Phi\left(\rho\Phi^{-1}(F_1(X)) + \sqrt{1-\rho^2}\Phi^{-1}(q)\right)\right]$$

mientras que la de una cópula t-Student es:

$$Y = F_2^{-1}\left[T_v\left(\rho.T_v^{-1}\left(F_1(X)\right) + \sqrt{\left(1-\rho^2\right)\frac{v + T_v^{-1}\left(F_1(\xi_1)\right)^2}{v+1}}T_{v+1}^{-1}(q)\right)\right]$$

mientras que si se relacionan a través de una cópula Clayon, la curva cuantil es:

$$Y = F_2^{-1}\left[\left(1 + F_1(X)^{-\alpha}\left(q^{-\alpha/(1+\alpha)} - 1\right)\right)^{-1/\alpha}\right]$$

y hay muchas otras cópulas para las que existe una expresión analítica como las anteriores para la curva cuantilica.

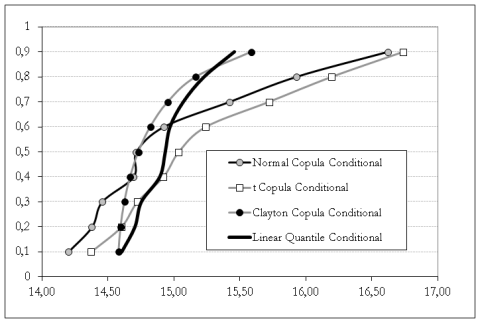
La *regresión cuantílica de cópula* es la solución al problema:

$$\min_{\theta}\sum_{t=1}^{T}\left(q - 1_{y_t \le Q_q(x_t,q;\theta)}\right)\left(y_t - \left(Q_q(x_t,q;\theta)\right)\right)$$

Estadísticamente, es más eficiente si en vez de calibrar las distribuciones marginales por separado de la cópula, estimamos todo simultáneamente, resolviendo el problema:

$$\min_{\alpha,\theta}\sum_{t=1}^{T}\left(q - 1_{y_t \le Q_q(x_t,q;\alpha,\theta)}\right)\left(y_t - \left(Q_q(x_t,q;\alpha,\theta)\right)\right)$$

Para estimar una cópula para representar la relación entre VFTSE y FTSE100 suponiendo que siguen distribuciones marginales tipo t-Student, comenzaríamos estandarizando los datos y estimando los grados de libertad de sus distribuciones marginales estandarizadas, loq ue se hace en EII.6.4. A continuación, en Case Study II.7.1 se estima una regresión cuantílica para una cópula Normal, asi como para una copula t-Student y una cópula de Clayton. Las distribuciones condicionales de VFTSE(t+1) bajo el supuesto de que FTSE caiga un 1% en t+1 se muestran en el gráfico:



donde podemos ver que tendríamos una confianza del 90% de que la volatilidad de FT100 (VFTSE) no excedería de 16,74 bajo la regresión de cópula t-Student, o de 16,62 bajo la cópula Normal o de 15,59 bajo la cópula Clayton, o de 15,46 bajo la regresión cuantilíca lineal.

2.4.1 Cobertura bajo regresiones cuantílicas de cópula

La cobertura de carteras es precisamente una de las situaciones en que un gestor de riesgos puede estar interesado en minimizar el riesgo a la baja específicamente, lo que sugiere el uso de regresiones cuantílicas en preferencia a la cobertura de Minimos Cuadrados. El Case Study II.7.3.2 analiza la cobertura de una cartera equiponderada de Vodafone, British Petroleum y HSBC. Después de construir una serie temporal de cotizaciones de dicha cartera, generamos la rentabilidad de la cartera y del indice FTSE100 y estimamos el ratio de cobertura de Mínimos Cuadrados es 0,547. Si utilizamos una regresión cuantílica con $q = 0,5$ obtenemos un ratio de cobertura de 0,496, inferior al de minimos cuadrados. Pero ninguno de estos dos modelos está diseñado para dar una consideración especial al riesgo a la baja. Para ello, hacemos $q = 2$, obteniendo un ratio de cobertura todavía inferior: 0,482. Una menor posición corta en el activo de cobertura proporciona en este caso una mejor protección frente a riesgo a la baja.

Para permitir relaciones no lineales entre las rentabilidades de la cartera y el índice, estimamos: a) una regresión de cópula Normal con $q = 0,20$, con marginales t-Student, b) una regresión de cópula t-Student para $q = 0,20$, con marginales t-Student. Para ello, comenzamos estandarizando ambas rentabilidades. Estimamos el número de grados de libertad en 9,99 para FTSE y 10,16 para la cartera equiponderada. Obtenemos un ratio de 0,557 para la cópula Normal y de 0,555 para la cópula t-Student, ambos significativamente superiores al ratio de la regresión cuantílica lineal, que era 0,482.

Por último, podríamos estimar un ratio permitiendo variación temporal. Hay varios procedimientos que podemos aplicar. El método EWMA (exponential moving average model) genera un ratio de cobertura:

$$\beta_t = \frac{Cov_\lambda(R_{ct}, I_t)}{Var_\lambda(I_t)}$$

siendo R_{ct}, I_t las rentabilidades de la cartera y el indice que utilizamos como cobertura, FTSE100, en el instante t.

Ratio cobertura EWMA

Una cobertura cambiante en el tiempo puede estimarse mediante modelos GARCH bivariantes sobre las rentabilidades del contado y del activo de cober-

tura. De este modo, el ratio de cobertura se va adaptando a las situaciones de mercado en funcion de cambios en las volatildades relativas de ambos activos y de su correlación, que pueden variar muy significativamente. [ver Lafuente y Novales (2003), Andani, Lafuente y Novales (2009), y Novales y Urtubia (2014) para coberturas cruzadas].

Este ejercicio muestra que hay un riesgo de modelo bastante significativo al decidir coberturas óptimas. En este caso, el ratio de cobertura debe ser algo mas reducido que el de minimos cuadrados si queremos que trate adecuadamente el riesgo a la baja, pero ligeramente más elevado si va a tener en cuenta la dependencia no lineal entre rentabilidades que no siguen una distribución conjunta Normal. Además, los ratios de cobertura que reflejan mejor las condiciones de mercado son considerablemente mas altos que los que se estiman utilizando promedios muestrales. Sin embargo, no hay una regla general a este respecto.

3 Las dificultades del método de Minimos Cuadrados en modelos no lineales

El procedimiento de Mínimos Cuadrados no Lineales en este modelo consiste en resolver el problema de optimización:

$$\min_{\theta} SR(\hat{\theta}) = \min_{\theta} \sum_{t=1}^{T} \hat{u}_t \left(\hat{\theta}\right) = \min_{\theta} \sum_{t=1}^{T} \left[y_t - f(x_t, \beta)\right]^2$$

lo que implica resolver el sistema de ecuaciones,

$$\left(\frac{\partial f(x_t, \beta)}{\partial \beta}\right)' y = \left(\frac{\partial f(x_t, \beta)}{\partial \beta}\right)' f(X, \beta)$$

donde el vector gradiente es Txk, y $f(X, \beta)$ es $Tx1$. Este sistema puede no tener solución, o tener múltiples soluciones. A diferencia del estimador de Mínimos Cuadrados aplicado a un modelo lineal, el estimador *no* es insesgado. La matriz de covarianzas del estimador resultante es:

$$Var(\hat{\theta}) = \sigma_u^2 \left[\left(\frac{\partial f(x_t, \beta)}{\partial \beta}\right)' \left(\frac{\partial f(x_t, \beta)}{\partial \beta}\right)\right]^{-1}$$

que se reduce a la matriz de covarianzas $\sigma_u^2 (X'X)^{-1}$ en el caso de un modelo lineal.

Si quisiéramos aplicar Mínimos Cuadrados directamente, en el modelo exponencial,

$$y_t = f(x_t, \theta) + u_t = \alpha + \beta_1 e^{\beta_2 x_t} + u_t$$

con $\theta = (\alpha, \beta_1, \beta_2)$, tendríamos que resolver el problema,

$$\min_{\theta} SR(\hat{\theta}) = \min_{\theta} \sum_{t=1}^{T} \left[\hat{u}_t\left(\hat{\theta}\right)\right]^2 = \min_{\theta} \sum_{t=1}^{T} \left[y_t - (\alpha + \beta_1 e^{\beta_2 x_t})\right]^2$$

que conduce a las condiciones de optimalidad,

$$\begin{aligned} \sum y_t &= \alpha T + \beta_1 \sum e^{\beta_2 x_t} \\ \sum y_t e^{\beta_2 x_t} &= \alpha \sum e^{\beta_2 x_t} + \beta_1 \sum e^{2\beta_2 x_t} \\ \sum y_t x_t e^{\beta_2 x_t} &= \alpha \sum x_t e^{2\beta_2 x_t} + \beta_1 \sum x_t e^{2\beta_2 x_t} \end{aligned}$$

que carece de solución explícita, por lo que debe resolverse por procedimientos numéricos.

3.1 Aproximación lineal del modelo no lineal

Para evitar recurrir a los métodos numéricos, en los que siempre es complicado saber si hemos encontrado el tipo de solución que buscábamos, un primer enfoque consiste en estimar la aproximación lineal del modelo (1) , alrededor de una estimación inicial,

$$y_t = f(x_t, \hat{\beta}) + \left(\frac{\partial f(x_t, \beta)}{\partial \beta}\right)_{\beta=\hat{\beta}} \left(\beta - \hat{\beta}\right) + u_t,$$

Haciendo el cambio de variable: $y_t^* = y_t - f(x_t, \hat{\beta}) + \left(\frac{\partial f(x_t, \beta)}{\partial \beta}\right)_{\beta=\hat{\beta}} \hat{\beta}$, y generando asimismo "datos" para cada una de las k variables definidas por el gradiente $\left(\frac{\partial f(x_t, \beta)}{\partial \beta}\right)_{\beta=\hat{\beta}}$,podemos estimar el modelo lineal

$$y_t^* \simeq \left(\frac{\partial f(x_t, \beta)}{\partial \beta}\right)_{\beta=\hat{\beta}} \beta + u_t,$$

por el procedimiento habitual de Mínimos Cuadrados.

Podemos pensar que en realidad estamos estimando un modelo distinto del que pretendíamos, y que de poco nos servirá, si el modelo que estimamos tiene una variable dependiente y unas variables explicativas diferentes de las que aparecían en el modelo original. Lo que sucede es que una vez más (como también sucede al estimar por MCG un modelo de regresión inicial en el que el término de error tiene heterocedasticidad o autotocorrelación), lo que hacemos es transformar las variables del modelo para obtener otro modelo diferente, que comparte con el primero los mismos coeficientes, y en el que la estimación de mínimos cuadrados tiene buenas propiedades. Además, veremos pronto que esta estrategia de estimación se puede interpretar como el resultado de un verdadero problema de minimización de la suma de cuadrados de residuos (ver algoritmo de Gauss Newton, más adelante).

La estimación resultante es,

$$\tilde{\beta} = \left[\left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)'_{\beta=\hat{\beta}} \left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)_{\beta=\hat{\beta}} \right]^{-1} \left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)'_{\beta=\hat{\beta}} y^*$$

donde el vector gradiente es una matriz de pseudo-datos, de dimensión Txk, e y^* es un vector $Tx1$.

Sustituyendo y^* por la expresión que utilizamos para definir a esta variable, podemos escribir el estimador como,

$$\tilde{\beta} = \hat{\beta} + \left[\left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)'_{\beta=\hat{\beta}} \left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)_{\beta=\hat{\beta}} \right]^{-1} \left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)'_{\beta=\hat{\beta}} \hat{u}.$$

Este resultado es muy interesante, pues permite poner en práctica un procedimiento iterativo, del siguiente modo: en cada etapa, partimos de unos determinados valores numéricos para los parámetros β del modelo, que utilizamos para generar los errores de ajuste, $\hat{u}$, y estimamos una regresión de dichos errores sobre las variables que configuran el vector gradiente $\frac{\partial f(x_t, \beta)}{\partial \beta}$.

Los coeficientes estimados en dicha regresión son las correcciones que hay que introducir sobre el estimador disponible en en dicha etapa para obtener un nuevo vector de estimaciones. Para comenzar este proceso, hemos de empezar con unas estimaciones iniciales, que se seleccionan bien utilizando información muestral, o bien escogiendo valores numéricos que simplifiquen el modelo.

El estimador resultante tras la convergencia del procedimiento tiene una distribución asintótica Normal, con esperanza matemática igual al verdadero vector de parámetros β, y su matriz de covarianzas puede estimarse por,

$$\hat{\sigma}_u^2 \left[\left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)'_{\beta=\tilde{\beta}} \left(\frac{\partial f(x_t, \beta)}{\partial \beta} \right)_{\beta=\tilde{\beta}} \right]^{-1} \tag{5}$$

con $\hat{\sigma}_u^2 = \frac{1}{T-k} \sum_{t=1}^{T} \tilde{u}_t^2$,siendo el residuo $\tilde{u}_t = y_t - f(x_t, \tilde{\beta})$.

Más adelante veremos una interpretación alternativa de este enfoque, que resulta si aplicamos un algoritmo numérico a la función de Suma de Cuadrados de los Residuos, directamente, no a una aproximación de la misma.

3.1.1 Ejemplo 1: Modelo exponencial con constante

Consideremos la estimación del modelo exponencial:

$$y_t = \alpha + \beta_1 e^{\beta_2 x_t} + u_t = f(x_t, \theta) + u_t$$

con $\theta = (\alpha, \beta_1, \beta_2)$. El gradiente de la función f que define la relación entre variable dependiente e independiente es,

$$\frac{\partial f(x_t,\theta)}{\partial \theta} = \left(1, e^{\beta_2 x_t}, \beta_1 x_t e^{\beta_2 x_t}\right)'$$

por lo que la aproximación lineal al modelo original es,

$$y_t \simeq f(x_t,\hat{\theta}) + \left(\frac{\partial f(x_t,\theta)}{\partial \theta}\right)'_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right) + u_t,\ t = 1,2,...,T,$$

que definiendo variables:

$$\begin{aligned} y_t^* &= y_t - f(x_t,\hat{\theta}) + \left(\frac{\partial f(x_t,\theta)}{\partial \theta}\right)'_{\theta=\hat{\theta}} \hat{\theta} = y_t + \hat{\beta}_1 \hat{\beta}_2 e^{\hat{\beta}_2 x_t} \\ z_{1t} &= e^{\hat{\beta}_2 x_t} \\ z_{2t} &= \hat{\beta}_1 x_t e^{\hat{\beta}_2 x_t} \end{aligned}$$

conduce a estimar el modelo,

$$y_t^* = \alpha + \beta_1 z_{1t} + \beta_2 z_{2t} + u_t,\ t = 1,2,...,T \tag{6}$$

A partir de unas estimaciones iniciales denotadas por el vector $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}_1, \hat{\beta}_2\right)$, generamos observaciones numéficas para la variable y_t^*, así como para las varaibles z_{1t}, z_{2t}, y procedemos a estimar el modelo (6) , obteniendo las nuevas estimaciones numéricas de los tres parámetros. Con ellos, podríamos volver a obtener observaciones numéricas de y_t^*, z_{1t}, z_{2t}, e iterar el procedimiento.

Como hemos visto antes, este procedimiento puede también ponerse en práctica estimando la regresión de los residuos sobre el vector gradiente:

$$\hat{u}_t = \delta_0 + \delta_1 z_{1t} + \delta_2 z_{2t}$$

Tanto el cálculo del vectror de residuos como la generación de datos para el vector gradiente dependerán de la estimación concreta disponible en ese momento, y procederemos a la actualización de valores numéricos de los parámetros, mediante:

$$\hat{\alpha}_n = \hat{\alpha}_{n-1} + \hat{\delta}_0;\ \hat{\beta}_{1,n} = \hat{\beta}_{1,n-1} + \hat{\delta}_1;\ \hat{\beta}_{2,n} = \hat{\beta}_{2,n-1} + \hat{\delta}_2$$

siendo $\hat{u}_t = y_t - f(x_t, \hat{\theta}_{n-1})$.

3.1.2 Ejemplo 2: Modelo potencial

Supongamos que queremos estimar el modelo potencial:

$$y_t = \alpha + \beta x_t^{\gamma} + u_t, t = 1,2,...,T$$

la función $f(x_t,\beta)$ es: $f(x_t,\beta) = \alpha + x_t^{\gamma}$, de modo que el vector gradiente es:

$$\frac{\partial f(x_t,\beta)}{\partial \beta} = \left(\frac{\partial f(x_t,\beta)}{\partial \alpha}, \frac{\partial f(x_t,\beta)}{\partial \beta}, \frac{\partial f(x_t,\beta)}{\partial \gamma} \right) = (1, x_t^{\gamma}, \beta x_t^{\gamma} \ln x_t)$$

[Recordemos que la derivada de la función x^{γ} *con respecto a* γ es igual a $x^{\gamma} \ln x$].

Nótese que para cada observación t tenemos un vector de tres valores numéricos para el vector $\frac{\partial f\ (x_t,\beta)}{\partial \beta}$, que siempre tiene como primer elemento en este caso el número 1.

A partir de unas estimaciones $\hat{\beta}$, calculamos los errores de ajuste:

$$\hat{u}_t = y_t - \hat{\alpha} - \hat{\beta} x_t^{\hat{\gamma}}, t = 1, 2, ..., T$$

y estimamos una regresión con $\hat{u}_t$ como variable dependiente, y las tres variables del vector $\frac{\partial f\ (x_t,\beta)}{\partial \beta}$ como variables explicativas. El vector de estimaciones se añade, con el signo que haya tenido (es decir, se suma si es positivo, y se resta si es negativo), de las estimaciones iniciales, para tener una nueva estimación. el algoritmo continua hasta que alcance la convergencia, y el punto al que converge se toma como estimación del vector β.

En este modelo, una estimación inicial razonable consistiría en partir de $\gamma =$ 1,que simplifica el modelo haciéndolo lineal. Si estimamos una regresión lineal por mínimos cuadrados: $y_t = \alpha + \beta x_t + u_t, t = 1, 2, ..., T$, el vector $\left(\hat{\alpha}, \hat{\beta}, 1 \right)$, donde $\hat{\alpha}$ y $\hat{\beta}$ denotan las estimaciones de mínimos cuadrados del modelo lineal, servirían como estimaciones iniciale para comenzar el procedimiento iterativo.

4 Minimización de una función

Teóricamente, para estimar por máxima verosimilitud deberíamos derivar la función de verosimilitud o su logaritmo (lo que suele ser más sencillo, al menos bajo Normalidad), respecto a cada uno de los parámetros del modelo, y al igualar a cero cada una de dichas derivadas, tendríamos tantas condiciones de optimalidad como parámetros a estimar. Resolveríamos dicho sistema encontrando valores numéricos para cada parámetro del modelo. Si se cunplen las condiciones de segundo orden (hessiano del logaritmo de la función de verosimilitud definido negativo en el vector de valores paramétricos que hemos obtenido como solución al sistema anterior, si estamos buscando un mínimo, o definido positivo, si estamos buscando un máximo), entonces podríamos decir que hemos hallado un mínimo o un máximo local, respectivamente. Nótese nuestra insistencia en que no habremos obtenido la solución al problema de optimización salvo si la función de verosimilitud es globalmente cóncava, en caso de buscar un máximo, o convexa, en caso de buscar un mínimo.

El problema básico es que, excepto en .casos muy específicos, el sistema de condiciones de primer orden no tiene solución analítica, es decir, no pueden despejarse en él los parámetros desconocidos. Ello hace necesaria la utilización de un algoritmo numérico de optimización.

Consideremos una función $F(\theta)$ cuyo mínimo estamos buscando. Supongamos que disponemos de una estimación inicial de los parámetros desconocidos, $\hat{\theta}_0$, y queremos obtener otra estimación más próxima al verdadero vector. A partir de una estimación inicial del valor de dicho vector, $\hat{\theta}_{n-1}$, aproximamos la función $F(.)$.

$$F(\theta) \simeq F\left(\hat{\theta}_n\right) + \left[\nabla F\left(\hat{\theta}_n\right)\right]' \left(\theta - \hat{\theta}_n\right) + \frac{1}{2}\left(\theta - \hat{\theta}_n\right)' \left[\nabla^2 F\left(\hat{\theta}_n\right)\right] \left(\theta - \hat{\theta}_n\right) \equiv M(\theta)$$

donde $\nabla F\left(\hat{\theta}_n\right), \nabla^2 F\left(\hat{\theta}_n\right)$ denotan, respectivamente, el vector gradiente y la matriz hessiana de la función F, evaluados en el punto $\hat{\theta}_n$. Para encontrar una estimación numérica que mejore la que teníamos hasta ahora, $\hat{\theta}_n$, podemos minimizar el valor numérico del miembro derecho de la expresión anterior, tomado como función del vector de parámetros θ, $M(\theta)$. Al igualar a cero la derivada de dicha función respecto de θ tenemos,

$$M'(\theta) = \left[\nabla F\left(\hat{\theta}_n\right)\right] + \left[\nabla^2 F\left(\hat{\theta}_n\right)\right] \left(\theta - \hat{\theta}_n\right) = 0$$

que conduce a,

$$\theta = \hat{\theta}_n - \left[\nabla^2 F\left(\hat{\theta}_n\right)\right]^{-1} \left[\nabla F\left(\hat{\theta}_n\right)\right] \tag{7}$$

valor numérico que puede tomarse como la nueva estimación, $\hat{\theta}_{n+1}$. Por supuesto, convendrá comprobar que el Hessiano $\nabla^2 F\left(\hat{\theta}_n\right)$ es definido positivo.

Este es un algoritmo iterativo, conocido como *algoritmo de Newton-Raphson*. Converge en una sóla etapa al mínimo local cuando la función $F(\theta)$ es cuadrática. En los demás casos, no hay ninguna seguridad de que el algoritmo vaya a converger. Incluso si lo hace, no hay seguridad de que converja al mínimo global, frente a hacerlo a un mínimo local. Además, no es posible saber si el límite alcanzado es o no un mínimo de naturaleza local. Por eso, conviene repetir el ejercicio partiendo de condiciones iniciales muy distintas para, si converge, certificar que lo hace a un mínimo local *peor* que el alcanzado previamente.

El algoritmo se basa en condiciones de primer orden por lo que, cuando el algoritmo converja, no sabremos si hemos alcanzado un máximo o un mínimo, y necesitaremos hacer alguna exploración adicional. Si aplicamos la expresión anterior a la minimización de una función cuadrática: $F(\theta) = a\theta^2 + b\theta + c$, obtenemos: $\hat{\theta}_n = -b/2a$, llegando a este punto crítico de la función sin necesidad de hacer ninguna iteración.

La derivada segunda de $M(\theta)$ es igual a $\left[\nabla^2 F\left(\hat{\theta}_0\right)\right]$, por lo que si este hessiano es definido positivo, estaremos aproximándonos al mínimo de la función $F(\theta)$. Una vez calculado el valor numérico de θ en (7) lo tomamos como la próxima estimación, $\hat{\theta}_1$. El procedimiento puede volver a repetirse, hasta que se consiga la convergencia a un punto mínimo. Cuando esto ocurra, sin embargo, no sabremos si el mínimo alcanzado es de naturaleza local o global, lo que habremos de explorar siguiendo las pautas que daremos más adelante.

En este tipo de algoritmos puede utilizarse un parámetro λ de longitud de paso, para tratar de controlar la velocidad de convergencia y, con ello, posibilitar que nos aproximemos al mínimo global, o que no abandonemos demasiado pronto una determinada región del espacio paramétrico:

$$\theta = \hat{\theta}_n - \lambda \left[\nabla^2 F\left(\hat{\theta}_n\right)\right]^{-1} \left[\nabla F\left(\hat{\theta}_n\right)\right]$$

Hay que tener en cuenta que posiblemente esté incorporado en el programa informático que se utilice para estimar modelos no lineales una determinada magnitud para λ, que el investigador puede alterar cuando observe cambios bruscos en el vector de parámetros.

En el caso de la estimación por máxima verosimilitud, la función que queremos minimizar es $-\ln L(\theta)$, donde $L(\theta)$ denota la función de verosimilitud. Así, tenemos el algoritmo numérico,

$$\theta = \hat{\theta}_0 - \left[\nabla^2 \ln L\left(\hat{\theta}_0\right)\right]^{-1} \nabla \ln L\left(\hat{\theta}_0\right) \tag{8}$$

La matriz de covarianzas, una vez lograda la convergencia, es

$$Cov\left(\hat{\theta}_n\right) = -\left[\nabla^2 \ln L\left(\hat{\theta}_0\right)\right]^{-1}$$

que será definida positiva en el caso de una distribución de probabilidad Normal para la innovación del modelo, puesto que la densidad Normal es estrictamente cóncava.

El estimador de máxima verosimilitud es eficiente, pero nos encontramos a dos dificultades: una, la referida acercad e nuestro desconcimietno sobre si hemos alcanzado un máximo local o global; otro, que las buenas propiedades del estimador de máxima verosimilitud descansan en que el supuesto acerca de la distribución de probabilidad que sigue la innovación del modelo sea correcto. En muchas ocasiones se calcula el estimador bajo supuestos de Normalidad porque es más sencillo, aun a sabiendas de que la distribución de probabilidad de la innovación dista de ser Normal. El estimador resultante se conoce como estimador de *quasi-máxima verosimilitud*.

4.0.3 Algunas simplificaciones

La puesta en práctica del algoritmo anterior requiere obtener las expresiones analíticas de las derivadas primeras y segundas de la función F. Ello significa calcular $k\left(\frac{k+3}{2}\right)$ derivadas, que hay que evaluar para cada dato, utilizando los valores numéricos de los parámetros que en ese momento se tienen como estimación, lo que puede ser un gran trabajo. Para evitar esta tarea pueden adoptarse algunas posibles soluciones:

- sustituir el hessiano $\nabla^2 F\left(\hat{\theta}_0\right)$ por el producto del vector gradiente por sí mismo, $\nabla F\left(\hat{\theta}_0\right) \nabla F\left(\hat{\theta}_0\right)'$, lo que genera una matriz cuadrada, simétrica, definida positiva,

- sustituir las derivadas analíticas por derivadas numéricas. Para ello, cuando disponemos de un vector de estimaciones $\hat{\theta}_{n-1}$, variamos ligeramente uno de los parámetros, y evaluamos numéricamente la función objetivo en el vector resultante. El cambio en el valor numérico de F, dividido por la variación introducida en el parámetro considerado, nos da una aproximación numérica a la derivada parcial con respecto a dicho parámetro, evaluada en el vector de estimaciones disponibles en ese momento,

- las derivadas analíticas se simplifican mucho, generalmente, si utilizamos su esperanza matemática. Ello nos llevaría al algoritmo iterativo,

$$\theta = \hat{\theta}_0 + \left[I\left(\hat{\theta}_0\right) \right]^{-1} \nabla \ln L\left(\hat{\theta}_0\right)$$

donde $I\left(\hat{\theta}_0\right)$ denota la matriz de información correspondiente a la distribución de probabilidad que se ha supuesto para la innovación del modelo: $I\left(\hat{\theta}_0\right) = E\left[-\nabla^2 \ln L\left(\hat{\theta}_0\right)\right]$. Este procedimiento se conoce como *algoritmo de scoring*, y es muy utilizado, por su simplicidad. En tal caso, la matriz de covarianzas del estimador resultante es,

$$Var\left(\hat{\theta}_n\right) = \left[I\left(\hat{\theta}_0\right) \right]^{-1}$$

4.1 Criterios de convergencia

Antes de ello, vamos a establecer criterios de convergencia: decimos que el algoritmo iterativo anterior ha convergido, y detenemos el procedimeitno numérico de estimación, cuando se cumple alguna de las siguientes condiciones:

- el valor numérico de la función objetivo varía menos que un cierto umbral previamente establecido al pasar de una estimación $\hat{\theta}_{n-1}$, a la siguiente, $\hat{\theta}_n$,

$$F\left(\hat{\theta}_n\right) - F\left(\hat{\theta}_{n-1}\right) < \varepsilon_3$$

- el gradiente de la función objetivo, evaluado en la nueva estimación, $\nabla F\left(\hat{\theta}_n\right)$, es pequeño, en el sentido de tener una norma reducida. Para comprobar el cumplimiento de esta condición, puede utilizarse la norma euclídea: raiz cuadrada de la suma de los cuadrados de los valores numéricos de cada componente del gradiente, o puede utilizarse el valor numérico de cualquier forma cuadrática calculada con el vector gradiente y una matriz definida positiva.

$$\left[\nabla F\left(\hat{\theta}_n\right)\right]' \left[\nabla F\left(\hat{\theta}_n\right)\right] < \varepsilon_2$$

- la variación en el vector de estimaciones es inferior a un umbral previamente establecido. Para comprobar esta condición utilizaríamos una norma del vector diferencia $\hat{\theta}_n - \hat{\theta}_{n-1}$,

$$\left(\hat{\theta}_n - \hat{\theta}_{n-1}\right)' \left(\hat{\theta}_n - \hat{\theta}_{n-1}\right) < \varepsilon_1$$

- se ha alcanzado el máximo número de iteraciones establecido en el programa de cálculo numérico que lleva a cabo la actualización de estimaciones descrita en (7) . Esto se hace con el objeto de que el programa de estimación no continúe iterando durante un largo período de tiempo, especialmente, si no está mejorando significativamente la situación de estimación.

El programa de estimación puede diseñarse para que se detenga cuando se cumple uno cualquiera de estos criterios, o todos ellos. Es importante puntualizar, por tanto, que al estimar mediante un algoritmo numérico, el investigador puede controlar: *i*) las estimaciones iniciales, *ii*) el máximo número de iteraciones a efectuar, y *iii*) el tamaño del gradiente, *iv*) la variación en el vector de parámetros y *v*) el cambio en el valor numérico de la función objetivo por debajo de los cuales se detiene la estimación. Cuando se utiliza una rutina proporcionada por una librería en un determinado lenguaje, dicha rutina incorpora valores numéricos para todos los criterios señalados, que pueden no ser los que el investigador preferiría, por lo que es muy conveniente poder variar dichos parámetros en la rutina utilizada. Alternativamente, lo que es mucho más conveniente, el investigador puede optar por escribir su propio programa de estimación numérica.

Estos aspectos afectan asimismo a la presentación de los resultados obtenidos a partir de un esquema de estimación numérica: como generalmente no sabemos si hemos alcanzado un óptimo local o global, esto debe examinarse volviendo a repetir el ejercicio de estimación a partir de condiciones inniciales sustancialmente diferentes de las utilizadas en primer lugar, con objeto de ver si se produce la convergencia, y cual es el valor de la función objetivo en dicho punto. Conviene repetir esta prueba varias veces. Asimismo, cuando se presentan estimaciones, deberían acompañarse de la norma del graidnet en dicho punto, así como de los umbrales utilizados para detener el proceso de estimación, tanto en términos del vector gradiente, como de los cambios en el vector de estimaciones, o en el valor numérico de la función objetivo, como hemos explicado en el párrafo anterior.

4.2 Dificultades prácticas en el algoritmo iterativo de estimación

- Cuando se utilizan algoritmos numéricos para la maximización de la función de verosimilitud es frecuente encontrar situaciones en las que el algoritmo numérico encuentra dificultades para encontrar una solución al

problema de optimización. Es muy importante que, en todos los casos en que la rutina de estimación o de optimización se detenga, examinemos cuál es el criterio de parada que ha actuado. Cuando el programa se ha escrito de modo que se detenga cuando se cumple alguno de los criterios antes señalados, conviene incluir en el programa un mensjae que haga explícito cuál de los criterios ha conducido a su parada, de modo que reduzcamos el umbral asociado a dicho criterio.

- Si la razón es que se ha excedido el máximo número de iteraciones propuesto en el programa, siempre se debe volver a ejecutar dicho programa. En la mayoría de los casos, es razonable elevar el número máximo de iteraciones y, posiblemente, comenzar a partir del vector de parámetros en el que se haya detenido.

- En ocasiones la rutina numérica itera un número reducido de veces y, sin exceder del máximo número de iteraciones, se detiene en un punto muy próximo al que hemos utilizado como condiciones iniciales. Esto puede deberse a que los umbrales de parada que hemos seleccionado, o que están escritos como valores por defecto en la rutina que implemente el algoritmo numérico son demasiado grandes. Así, en los primeros cálculos, los cambios en las estimaciones o en el valor de la función objetivo son inferiores a dichos umbrales, y el algoritmo se detiene. Deben reducirse dichos umbrales y volver a estimar.

- Si el programa se detiene sin exceder el máximo número de iteraciones, es importante comparar los valores paramétricos en los que se detiene, con los que se utilizaron como condiciones iniciales. Esta comparación que, lamentablemente, no suele efectuarse, muestra frecuentemente que en alguno de los parámetros el algoritmo no se ha movido de la condición inicial. Salvo que tengamos razones sólidas para creer que dicha condición inicial era ya buena, esto significa que, o bien el algoritmo está teniendo dificultades para encontrar en que sentido mover en la dirección de dicho parámetro para mejorar el valor numérico de la función objetivo, o no ha tenido suficiente posibilidad de iterar en esa dirección, dadas las dificultades que encuentra en otras direcciones (o parámetros). En estos casos quizá conviene ampliar el número máximo de iteraciones, y quizá también reducir la tolerancia del algoritmo (la variación en θ o en F que se ha programado como criterio de parada), para evitar que el algoritmo se detenga demasiado pronto.

- Todo esto no es sino reflejo, en general, de un exceso de parametrización, que conduce a que la superficie que representa la función objetivo, como función de los parámetros, sea plana en algunas direcciones (o parámetros). Esto hace que sea dificil *identificar* los valores numéricos de cada uno de los parámetros del modelo por separado de los demás, por lo que el algoritmo encuentra dificultades en hallar una dirección de búsqueda en la que mejore el valor numérico de la función objetivo. Una variación, incluso

si es de magnitud apreciable, en la dirección de casi cualquier parametro, apenas varía el valor numérico de la función objetivo. Por eso, el algoritmo no encuentra un modo de variar los valores paramétricos de modo que la función objetivo cambie por encima de la tolerancia que hemos fijado, y se detiene. En estos casos, el gradiente va a ser también muy pequeño, que puede ser otro motivo por el que el algoritmo se detenga. De hecho, la función objetivo varía de modo similar (poco, en todo caso) tanto si el algoritmo varía uno como si cambia varios parámetros, que es lo que genera el problema de identificación, similar al que se obtiene en el modelo lineal general cuando existe colinealidad entre alguna de las variables explicativas. Las dificultades en la convergencia del algoritmo producidas por una excesiva sobreparametrización del modelo se reflejan en unas elevadas correlaciones de los parámetros estimados. Como en cualquier otro problema de estimación, conviene examinar no sólo las varianzas de los parámetros estimados, sino también las correlaciones entre ellos.

4.3 Estimación condicionada y precisión en la estimación

Para tratar estas situaciones, cuando se identifican uno o dos parámetros altamente correlacionados con los demás, puede llevarse a cabo una estimación condicionada, fijando valores alternativos de dichos parámetros a lo largo de una red, maximizando la verosimilitud respecto de los demás, y comparando resultados para alcanzar el máximo absoluto. En otras ocasiones, sin necesidad de incurrir en dificultades numéricas, se aprecia que imponer un valor numérico para uno o dos parámetros simplifica enormemente la estructura del modelo a estimar, por ejemplo, haciéndola linear. Si este es el caso, puede establecerse una red de búsqueda en dichos parámetros y, para cada uno de ellos, estimar el modelo lineal resultante. Se resuelve así un conjunto de muchos problemas simples, frente a la alternativa de resolver un único problema complicado que es, en ocasiones, mucho más difícil.

Una limitación de esta estrategia de estimación, que tantas veces simplifica el problema computacional, es que no nos proporciona una estimación de la varianza para el parámetro o los parámetros sobre los que se ha hecho la estimación condicional. Según cuál sea el grado de simplificación alcanzado, podríamos no tener varianzas para ninguno de los parámetros. Esto sugiere una cuestión aún más profunda, acerca del significado real de las varianzas proporcionadas por el problema de estimación. En realidad, lo que el investigador quiere tener es una medida del grado de precisión obtenido en su estimación, y ello bien puede depender del objetivo final de la estimación del modelo. Por ejemplo, consideremos el habitual problema de calcular la volatilidad implícita de una opción. Obtener las sensibilidades de la respuesta a dicha pregunta a variaciones en el valor de alguno de los parámetros que se fija equivale a determinar un rango de confianza para el parámetro que se estima.

Consideremos que el subyacente de una opción call cotiza a 100, que el precio de ejercicio de la misma es 95, el tipo de interés, supuesto constante hasta el vencimiento, es 7,5%, el plazo residual es 3 meses, y el precio de la opción es de

10. La inversión de la fórmula de Black Scholes (BS) proporciona una volatilidad de 31,3%. Este no es un problema estadístico, y no se ha llevado a cabo ningún proceso de muestreo. Sin embargo, el usuario que conoce la limitación del modelo BS por los supuestos que incorpora, puede estar dispuesto a aceptar un rango de valores de volatilidad que no generen un precio teórico que se separe en más de 0,25 del precio observado en el mercado. Ello le llevará a considerar un rango de volatilidades entre 29,8% y 32,7%.

La misma idea puede aplicarse en un problema de estimación para evaluar la precisión con que se ha estimado un determinado parámetro. En función de la utilidad que se vaya a dar al modelo, el usuario puede determinar que está dispuesto a aceptar variaciones de hasta un 1% alrededor del valor de la función objetivo que ha obtenido en su estimación. Se trata entonces de perturbar el valor numérico del parámetro cuya precisión se quiere medir, y estimar condicionando en dicho valor mientras que el valor resultante para la función objetivo satisfaga la condición prefijada. Se obtiene así numericamente, un intervalo de confianza alrededor de la estimación inicialmente obtenida. En principio, esta región no tiene por qué coincidir con la tradicional región de confianza. Puede resultar extraño hablar de regiones de confianza paramétricas en el caso del cálculo de la volatilidad implícita pues, como hemos dicho, no es realmente un problema estadístico. Existe un razonamiento distinto del anterior, con más base estadística que conduce asimismo a una región de confianza paramétrica. Para ello, consideremos que el usuario de la expresión BS, consciente de que el tipo de interés relevante no va a permanecer constante hasta vencimiento, y desconociendo su evolución establece un conjunto de posibles escenarios de evolución de los tipos, cada uno acompañado de una probabilidad que recoge la mayor o menor verosimilitud asignada a dicho escenario, e identifica cada escenario con distintos niveles constantes del tipo de interés. Calculando la volatilidad implícita para cada nivel de tipos de interés considerado, mientras se mantienen constantes los restantes parámetros, generaríamos una distribución de probabilidad para la volatilidad implícita. Por supuesto, este argumento se puede generalizar el caso en que la incertidumbre a priori se recoge en la forma de una distribución de probabilidad multivariante para el vector de parámetros sobre los que se condiciona en el proceso de estimación.

5 Estimación por Mínimos Cuadrados

Si queremos obtener el estimador de Mínimos Cuadrados del modelo no lineal, querremos minimizar la función,

$$F(\theta) = \sum_{t=1}^{T} (y_t - f(x_t, \beta))^2 = SR(\beta)$$

y la regla iterativa anterior se convierte en,

$$\hat{\beta}_n = \hat{\beta}_{n-1} - \left[\nabla^2 F\left(\hat{\beta}_{n-1}\right)\right]^{-1} \left[\nabla F\left(\hat{\beta}_{n-1}\right)\right]$$

en la que es fácil ver que,

$$\begin{aligned}\nabla F\left(\hat{\beta}_{n-1}\right) &= \frac{\partial SR(\beta)}{\partial \beta} = -2\sum_{t=1}^{T}\frac{\partial f(x_t,\beta)}{\partial \beta}u_t \\ \nabla^2 F\left(\hat{\beta}_{n-1}\right) &= \frac{\partial^2 SR(\beta)}{\partial \beta \partial \beta'} = 2\sum_{t=1}^{T}\left(\frac{\partial f(x_t,\beta)}{\partial \beta}\right)\left(\frac{\partial f(x_t,\beta)}{\partial \beta}\right)' - 2\sum_{t=1}^{T}\frac{\partial^2 f(x_t,\beta)}{\partial \beta \partial \beta'}u_t\end{aligned}$$

en este caso, el algoritmo de Newton-Raphson consiste en:

$$\hat{\beta}_n = \hat{\beta}_{n-1} + \left[\sum_{t=1}^{T}\left(\frac{\partial f(x_t,\beta)}{\partial \beta}\right)\left(\frac{\partial f(x_t,\beta)}{\partial \beta}\right)' - \frac{\partial^2 f(x_t,\beta)}{\partial \beta \partial \beta}u_t\right]^{-1}\left[\sum_{t=1}^{T}\frac{\partial f(x_t,\beta)}{\partial \beta}u_t\right]$$

El estimador resultante es asintóticamente insesgado, con matriz de covarianzas,

$$\sigma_u^2\left[\nabla^2 F\left(\hat{\theta}_n\right)\right]^{-1}$$

estimándose el parámetro σ_u^2 del modo antes referido, mediante el cociente de la Suma de Cuadrados de los errores de ajuste y el número de grados de libertad del modelo.

El *algoritmo de Gauss-Newton* consiste en ignorar la presencia de la segunda derivada en la matriz inversa anterior, y considerar el esquema iterativo,

$$\hat{\beta}_n = \hat{\beta}_{n-1} + \left[\sum_{t=1}^{T}\left(\frac{\partial f(x_t,\beta)}{\partial \beta}\right)\left(\frac{\partial f(x_t,\beta)}{\partial \beta}\right)'\right]^{-1}\left[\sum_{t=1}^{T}\frac{\partial f(x_t,\beta)}{\partial \beta}u_t\right]$$

Al despreciar la segunda derivada, este algoritmo entra en dificultades cuando la superficie a optimizar no tiene suficiente curvatura que, como veremos más adelante, son las situaciones que en términos estadísticos, corresponden a identificación imperfecta de los parámetros del modelo.

El interés de este segundo algoritmo estriba en que la expresión matricial que aparece en el segundo sumando corresponde con las estimaciones de mínimos cuadrados del vector de *errores,* calculado con las estimaciones actuales, sobre las k variables definidas por el vector gradiente $\frac{\partial f(x_t,\beta)}{\partial \beta}$. Son k variables, tantas como parámetros hay que estimar, porque el vector gradiente consta de una derivada parcial con respecto a cada uno de los k parámetros del modelo. Las estimaciones resultantes son las correcciones a introducir sobre las actuales estimaciones del vector β para tener un nuevo vector de estimaciones numéricas.

Como podemos ver, este es el mismo estimador que resulta de aplicar Minimos Cuadrados a la aproximación lineal del modelo no lineal.

5.1 Ilustración: El modelo exponencial con constante

Consideremos de nuevo la estimación del modelo exponencial,

$$y_t = \alpha + \beta_1 e^{\beta_2 x_t} + u_t = f(x_t, \theta) + u_t$$

Si denotamos por $F(\theta)$ la función Suma de Cuadrados de Residuos, tenemos el gradiente y matriz hessiana,

$$\nabla F(\theta) = -2\sum \frac{\partial f(x_t,\theta)}{\partial \theta}\hat{u}_t = -2\sum \frac{\partial f_t}{\partial \theta}\hat{u}_t = -2\sum \left(1, e^{\beta_2 x_t}, \beta_1 x_t e^{\beta_2 x_t}\right)\hat{u}_t$$

$$\begin{aligned}
\nabla^2 F(\theta) &= 2\sum \left(\frac{\partial f_t}{\partial \theta}\right)\left(\frac{\partial f_t}{\partial \theta}\right)' - 2\sum \frac{\partial^2 f_t}{\partial \theta^2}\hat{u}_t = \\
&= 2\sum_{t=1}^{T} \begin{pmatrix} 1 & e^{\beta_2 x_t} & \beta_1 x_t e^{\beta_2 x_t} \\ e^{\beta_2 x_t} & e^{2\beta_2 x_t} & \beta_1 x_t e^{2\beta_2 x_t} \\ \beta_1 x_t e^{\beta_2 x_t} & \beta_1 x_t e^{2\beta_2 x_t} & \beta_1^2 x_t^2 e^{2\beta_2 x_t} \end{pmatrix} - 2\sum_{t=1}^{T} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & x_t e^{\beta_2 x_t} \\ 0 & x_t e^{\beta_2 x_t} & \beta_1 x_t^2 e^{\beta_2 x_t} \end{pmatrix}\hat{u}_t = \\
&= 2\sum_{t=1}^{T} \begin{pmatrix} 1 & e^{\beta_2 x_t} & \beta_1 x_t e^{\beta_2 x_t} \\ e^{\beta_2 x_t} & e^{2\beta_2 x_t} & -x_t e^{\beta_2 x_t}\hat{u}_t + \beta_1 x_t e^{2\beta_2 x_t} \\ \beta_1 x_t e^{\beta_2 x_t} & x_t e^{\beta_2 x_t}\left(\beta_1 e^{\beta_2 x_t} - \hat{u}_t\right) & \beta_1 x_t^2 e^{\beta_2 x_t}\left(\beta_1 e^{\beta_2 x_t} - \hat{u}_t\right) \end{pmatrix}
\end{aligned}$$

y el algoritmo de Newton-Raphson consiste en actualizar los valores numéricos de los parámetros mediante el esquema,

$$\hat{\theta}_n = \hat{\theta}_{n-1} - \left[\nabla^2 F\left(\hat{\theta}_{n-1}\right)\right]^{-1} \nabla F\left(\hat{\theta}_{n-1}\right)$$

El algoritmo de Gauss-Newton es una versión simplificada del anterior, sustituyendo la matriz hessiana por el producto,

$$\sum_{t=1}^{T} \left(\frac{\partial f_t}{\partial \theta}\right)_{\theta=\hat{\theta}} \left(\frac{\partial f_t}{\partial \theta}\right)'_{\theta=\hat{\theta}}$$

lo que equivale a despreciar las derivadas de segundo orden. La aproximación será apropiada por tanto cuando la función a optimizar sea aproximadamente cuadrática. En ese caso, el hessiano sería constante. Como en la expresión del algoritmo Newton-Raphson aparece la suma de productos del hessiano por el residuo, si el hessiano es aproximadamente constante, la suma sería proporcional a la suma de residuos, que debería ser pequeña (sería cero si el modelo fuese lineal).

Bajo esta aproximación, tenemos el esquema iterativo,

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \left[\sum_{t=1}^{T} \left(\frac{\partial f_t}{\partial \theta}\right)_{\theta=\hat{\theta}_{n-1}} \left(\frac{\partial f_t}{\partial \theta}\right)'_{\theta=\hat{\theta}_{n-1}}\right]^{-1} \left[\sum_{t=1}^{T} \frac{\partial f(x_t,\beta)}{\partial \beta}\hat{u}_t\right]$$

que, como puede verse, coincide con la estimación de la aproximación lineal al modelo no lineal que antes analizamos.

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \left[\sum_{t=1}^{T} \begin{pmatrix} 1 & e^{\beta_2 x_t} & \beta_1 x_t e^{\beta_2 x_t} \\ e^{\beta_2 x_t} & e^{2\beta_2 x_t} & \beta_1 x_t e^{2\beta_2 x_t} \\ \beta_1 x_t e^{\beta_2 x_t} & \beta_1 x_t e^{2\beta_2 x_t} & \beta_1^2 x_t^2 e^{2\beta_2 x_t} \end{pmatrix} \right]^{-1} \left[\sum_{t=1}^{T} \begin{pmatrix} \hat{u}_t \\ e^{\beta_2 x_t} \hat{u}_t \\ \beta_1 x_t e^{\beta_2 x_t} \hat{u}_t \end{pmatrix} \right]$$

Pero lo verdaderamente interesante del algoritmo de Gauss-Newton es que la actualización en el estimador puede llevarse a cabo mediante una regresión de los errores de ajuste, calculados con el estimador actualmente disponible,

$$\hat{u}_t = y_t - f(x_t, \hat{\beta})$$

sobre el vector gradiente de la función f ,$\frac{\partial f_t}{\partial \theta}$.En el modelo exponencial se trataría de una regresión de $\hat{u}_t$ sobre las tres variables explicativas:

$$\nabla f_t \equiv \frac{\partial f_t}{\partial \theta} = \begin{pmatrix} 1 & e^{\beta_2 x_t} & \beta_1 x_t e^{\beta_2 x_t} \end{pmatrix}$$

Los coeficientes estimados en esta regresión auxiliar se añaden a los actuales valores numéricos de los parámetros para obtener el nuevo estimador, y se continúa de modo iterativo hasta lograr a convergencia del algoritmo.

5.1.1 Condiciones iniciales

En algunos casos, puede comenzarse de estimaciones iniciales sencillas. En el modelo potencial:

$$y_t = \alpha + \beta x_t^{\gamma} + u_t$$

es razonable comenzar con $\gamma_0 = 1$, lo que reduciría el modelo a una regresión lineal simple. Por tanto, estimando dicha regresión, si obtenemos estimaciones α_0, β_0, el vector de estimaciones iniciales seria: $(\alpha_0, \beta_0, 1)$.

Sin embargo, la sencillez puede generar dificultades numéricas. Por ejemplo, la estructura del modelo exponencial sugiere comenzar de $\beta_2 = 0$, con lo que desaparecería el término exponencial, y $\alpha = 0$, con lo que tendríamos $\beta_1 = \bar{y}$, y residuos: $\hat{u}_t = y_t - \bar{y}$. Sin embargo, en este caso, las matrices a invertir en los algoritmos de Newton- Raphson y Gauss-Newton resultan, respectivamente:

$$2\sum_{t=1}^{T} \begin{pmatrix} 1 & 1 & \bar{y}x_t \\ 1 & 1 & -x_t\hat{u}_t + \bar{y}x_t \\ \bar{y}x_t & -x_t\hat{u}_t + \bar{y}x_t & -x_t^2\bar{y}\hat{u}_t + \bar{y}^2x_t^2 \end{pmatrix} = 2\sum_{t=1}^{T} \begin{pmatrix} 1 & 1 & \bar{y}x_t \\ 1 & 1 & -x_ty_t + 2\bar{y}x_t \\ \bar{y}x_t & -x_ty_t + 2\bar{y}x_t & -x_t^2\bar{y}y_t + 2\bar{y}^2x_t^2 \end{pmatrix};$$
$$\sum_{t=1}^{T} \begin{pmatrix} 1 & 1 & \bar{y}x_t \\ 1 & 1 & \bar{y}x_t \\ \bar{y}x_t & \bar{y}x_t & \bar{y}^2x_t^2 \end{pmatrix}$$

siendo la segunda de ellas singular.

Afortunadamente, las condiciones de optimalidad del procedimiento de Mínimos Cuadrados nos sugieren cómo obtener estimaciones iniciales razonables. Notemos que la primera condición puede escribirse,

$$\alpha = m(y) - \beta_1 m(e^{\beta_2 x_t})$$

que, sustituida en la segunda, nos proporciona,

$$m(y_t e^{\beta_2 x_t}) = m(e^{\beta_2 x_t})m(y) - \beta_1 \left[m(e^{\beta_2 x_t})\right]^2 + \beta_1 m(e^{2\beta_2 x_t})$$

Dado un valor numérico de β_2, tenemos,

$$\beta_1 = \frac{m(y_t e^{\beta_2 x_t}) - m(e^{\beta_2 x_t})m(y)}{m(e^{2\beta_2 x_t}) - \left[m(e^{\beta_2 x_t})\right]^2}$$

que, como es habitual, tiene la forma de cociente entre una covarianza y una varianza muestrales.

La última condición de optimalidad nos dice,

$$m\left(y_t x_t e^{\beta_2 x_t}\right) = \alpha m\left(x_t e^{2\beta_2 x_t}\right) + \beta_1 m\left(x_t e^{2\beta_2 x_t}\right)$$

que proporcionaría otra elección de β_1,

$$\beta_1 = \frac{m(y_t x_t e^{\beta_2 x_t}) - m(x_t e^{2\beta_2 x_t})m(y)}{m(x_t e^{2\beta_2 x_t}) - \left[m(x_t e^{2\beta_2 x_t})\right]^2}$$

Podríamos optar por escoger el valor numérico de β_1 con cualquiera de ellas. También podríamos caracterizar la intersección, si existe, de las dos curvas para elegir ambos parámetros, β_1 y β_2.

Ejemplo 4: Un modelo no identificado Supongamos, por último, que pretendemos estimar el modelo,

$$y_t = \alpha + \beta_1 \beta_2 x_t + u_t$$

en el que la aplicación del algoritmo de Newton-Raphson resulta en,

$$\begin{pmatrix} \alpha^{(n)} \\ \beta_1^{(n)} \\ \beta_2^{(n)} \end{pmatrix} = \begin{pmatrix} \alpha^{(n-1)} \\ \beta_1^{(n-1)} \\ \beta_2^{(n-1)} \end{pmatrix} + \left[\sum_{t=1}^{T} \begin{pmatrix} 1 & \beta_2 x_t & \beta_1 x_t \\ \beta_2 x_t & \beta_2^2 x_t^2 & \beta_1 \beta_2 x_t^2 \\ \beta_1 x_t & \beta_1 \beta_2 x_t^2 & \beta_1^2 {x_t}^2 \end{pmatrix} - \sum_{t=1}^{T} \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & x_t \\ 0 & x_t & 0 \end{pmatrix} u_t \right]^{-1} \cdot \left[\sum_{t=1}^{T} \begin{pmatrix} 1 \\ e^{\beta_2 x_t} \\ \beta_1 x_t e^{\beta_2}$$

mientras que el algoritmo de Gauss-Newton consistiría en:

$$\begin{pmatrix} \alpha^{(n)} \\ \beta_1^{(n)} \\ \beta_2^{(n)} \end{pmatrix} = \begin{pmatrix} \alpha^{(n-1)} \\ \beta_1^{(n-1)} \\ \beta_2^{(n-1)} \end{pmatrix} + \left[\sum_{t=1}^{T} \begin{pmatrix} 1 & \beta_2 x_t & \beta_1 x_t \\ \beta_2 x_t & \beta_2^2 x_t^2 & \beta_1 \beta_2 x_t^2 \\ \beta_1 x_t & \beta_1 \beta_2 x_t^2 & \beta_1^2 {x_t}^2 \end{pmatrix}\right]^{-1} \cdot \left[\sum_{t=1}^{T} \begin{pmatrix} 1 \\ e^{\beta_2 x_t} \\ \beta_1 x_t e^{\beta_2 x_t} \end{pmatrix} \hat{u}_t\right]$$

y, como puede apreciarse facilmente, la matriz a invertir es singular, por lo que no puede aplicarse este algoritmo. Esto se debe a que l modelo presenta un obvia dificultad, comol es qeu no puede sintiguir entre los parámetros β_1 y β_2. Puede estimarse con cierta precisión su producto, pero no sus valores individuales. El modelo no está identificado, y esa es la razón de que las filas de la matriz Hessiama sean proporcionales entre sí, generando la singularidad. De hecho, la matriz a invertir en el algoritmo Newton-Raphson será asimismo aproximadamente singular, y muy probablemente, el investigador encontraría muy serios problemas numéricos.

Ejemplo 5: Modelo potencial Consideremos la utilización del modelo potencial para estimar la relación entre el tipo de interés a largo plazo R_t y el tipo de interés a corto plazo r_t,

$$R_t = \beta_1 + \beta_2 r_t^\gamma + u_t$$

son,

$$\begin{aligned}
\sum_{t=1}^{T} (R_t - \beta_1 - \beta_2 r_t^\gamma) &= 0 \\
\sum_{t=1}^{T} (R_t - \beta_1 - \beta_2 r_t^\gamma)\, r_t^\gamma &= 0 \\
\beta_2 \sum_{t=1}^{T} (R_t - \beta_1 - \beta_2 r_t^\gamma)\, r_t^\gamma \ln r_t &= 0
\end{aligned}$$

que constituyen las ecuaciones normales del problema de estimación. De las dos primeras ecuaciones, obtenemos,

$$\begin{aligned}
\sum_{t=1}^{T} R_t &= T\beta_1 + \beta_2 \sum_{t=1}^{T} r_t^\gamma \Rightarrow Tm(R) = T\beta_1 + \beta_2 Tm(r^\gamma) \Rightarrow \beta_1 = m(R) - \beta_2 m(r^\gamma) \\
\sum_{t=1}^{T} R_t r_t^\gamma &= \beta_1 \sum_{t=1}^{T} r_t^\gamma + \beta_2 \sum_{t=1}^{T} r_t^{2\gamma} \Rightarrow Tm(Rr^\gamma) = Tm(R)m(r^\gamma) - \beta_2 Tm(r^\gamma)^2 + \beta_2 Tm(r^{2\gamma}) \Rightarrow \\
&\Rightarrow \quad \beta_2 = \frac{m(Rr^\gamma) - m(R)m(r^\gamma)}{m(r^{2\gamma}) - m(r^\gamma)^2}
\end{aligned}$$

El primer resultado sugiere que la estimación del término independiente se obtenga, una vez estimados β_2 y γ, de modo similar a como se recupera el término independiente en la estimación de un modelo lineal.

Lo más interesante es observar que la segunda ecuación sugiere estimar el parámetro β_2 en función de momentos muestrales de algunas funciones de los tipos a largo y a corto plazo. Para calcular dichos momentos precisamos conocer

el parámetro γ, pero también podemos poner en marcha una búsqueda de red puesto que, por las características de la función de consumo, dicho parámetro ha de ser positivo y no muy elevado. Por tanto, una red que cubra el intervalo $(0.5, 2.0)$ puede ser suficiente. De hecho, para cada valor numérico posible de γ podemos utilizar la expresión anterior para estimar β_2,sin necesidad de optimizar, y después utilizar la primera condición de optimalidad para estimar β_1.

Ejemplo 6: Una función de consumo (Una aplicación distinta del mismo modelo anterior) Para apreciar el grado de dificultad, consideremos las condiciones de optimalidad correspondientes a la estimación por mínimos cuadrados del modelo de consumo,

$$C_t = \beta_1 + \beta_2 Y_t^\gamma + u_t$$

en el que la función $f(X, \beta)$ tiene gradiente:

$$\frac{\partial f(x_t, \beta)}{\partial \beta} = (1, Y_t^\gamma, \beta_2 Y_t^\gamma \ln Y_t)$$

que son,

$$\begin{aligned}
\sum_{t=1}^{T} (C_t - \beta_1 - \beta_2 Y_t^\gamma) &= 0 \\
\sum_{t=1}^{T} (C_t - \beta_1 - \beta_2 Y_t^\gamma)\, Y_t^\gamma &= 0 \\
\beta_2 \sum_{t=1}^{T} (C_t - \beta_1 - \beta_2 Y_t^\gamma)\, Y_t^\gamma \ln Y_t &= 0
\end{aligned}$$

que constituyen las ecuaciones normales del problema de estimación. De las dos primeras ecuaciones, obtenemos,

$$\begin{aligned}
\sum_{t=1}^{T} C_t &= T\beta_1 + \beta_2 \sum_{t=1}^{T} Y_t^\gamma \Rightarrow Tm(C) = T\beta_1 + \beta_2 Tm(Y^\gamma) \Rightarrow \beta_1 = m(C) - \beta_2 m(Y^\gamma) \\
\sum_{t=1}^{T} C_t Y_t^\gamma &= \beta_1 \sum_{t=1}^{T} Y_t^\gamma + \beta_2 \sum_{t=1}^{T} Y_t^{2\gamma} \Rightarrow Tm(CY^\gamma) = Tm(C)m(Y^\gamma) - \beta_2 Tm(Y^\gamma)^2 + \beta_2 Tm(Y^{2\gamma}) \Rightarrow \\
&\Rightarrow \quad \beta_2 = \frac{m(CY^\gamma) - m(C)m(Y^\gamma)}{m(Y^{2\gamma}) - m(Y^\gamma)^2}
\end{aligned}$$

Este procedimiento funciona muy bien desde el punto de vista numérico, como puede verse en el archivo *Ajuste_ consumo.xls*. La única limitación del método es que no proporciona la estructura de varianzas y covarianzas que permitiría llevar a cabo el análisis de inferencia estadística al modo habitual. Puede

analizarse, sin embargo, la región paramétrica consistente con un incremento en la Suma de Cuadrados de Residuos inferior a un cierto umbral de, por ejemplo, un 5%. Esto sería como construir una región de confianza del 95% para el vector de parámetros.

Ejemplo 7: Modelo exponencial sin constante. Consideremos ahora la estimación del modelo,

$$y_t = \alpha e^{\beta x_t} + u_t = f(x_t, \theta) + u_t$$

con $\theta = (\alpha, \beta)$. Entre muchas otras aplicaciones, este modelo se ha utilizado para representar una función de demanda de dinero, que relaciona la cantidad de saldos monetarios reales en la economía en función de las expectativas de inflación:

$$\left(\frac{M_t}{P_t}\right)^d = \alpha e^{\beta \pi_t^e} + u_t,\ t = 1, 2, ..., T,\ \alpha > 0,\ \beta < 0$$

El gradiente de la función f que define la relación entre variable dependiente e independiente, es,

$$\frac{\partial f(x_t, \theta)}{\partial \theta} = \left(e^{\beta x_t}, \alpha x_t e^{\beta x_t}\right)'$$

Es importante apreciar la expresión analítica de las derivadas parciales de esta función,

$$\frac{\partial y}{\partial x} = \alpha\beta e^{\beta x_t},\ \frac{\partial^2 y}{\partial x^2} = \alpha\beta^2 e^{\beta x_t},$$

Como la función exponencial es positiva con independencia del signo de β y de x_t, tenemos que la primera derivada tendrá el signo del producto $\alpha\beta$, mientras que la segunda derivada tendrá el signo del parámetro α. Esto nos puede dar pautas para la elección de condiciones iniciales. Por ejemplo, si la nube de puntos de y_t sobre x_t tiene un perfil decreciente y convexo, tendríamos un valor positivo de α, debido a la convexidad, junto con un valor negativo de β.

Aproximación lineal La aproximación lineal a este modelo es,

$$y_t \simeq f(x_t, \hat{\theta}) + \left(\frac{\partial f(x_t, \theta)}{\partial \theta}\right)'_{\theta=\hat{\theta}} \left(\theta - \hat{\theta}\right) + u_t,\ t = 1, 2, ..., T,$$

que, definiendo las variables $y_t^* = y_t - f(x_t, \hat{\theta}) + \left(\frac{\partial f(x_t, \theta)}{\partial \theta}\right)'_{\theta=\hat{\theta}} . \hat{\theta}$, $z_{1t} = e^{\hat{\beta} x_t}$, $z_{2t} = \hat{\alpha} x_t e^{\hat{\beta} x_t}$, puede escribirse:

$$y_t^* = \alpha z_{1t} + \beta z_{2t} + u_t,\ t = 1, 2, ..., T, \tag{9}$$

A partir de unas estimaciones iniciales denotadas por el vector $\hat{\theta} = \left(\hat{\alpha}, \hat{\beta}\right)$, generamos observaciones numéricas para la variable y_t^*, así como para las variables z_{1t}, z_{2t}, y procedemos a estimar el modelo (9), obteniendo las nuevas estimaciones numéricas de α y β. Con ellos, podríamos volver a obtener series temporales para las variables y_t^*, z_{1t}, z_{2t}, e iterar el procedimiento.

Como es sabido, este procedimiento puede también ponerse en práctica estimando la regresión,

$$\hat{u}_t = \delta_1 z_{1t} + \delta_2 z_{2t}$$

y procediendo a la actualización de valores numéricos de los parámetros,

$$\hat{\alpha}_n = \hat{\alpha}_{n-1} + \hat{\delta}_1;\ \hat{\beta}_n = \hat{\beta}_{n-1} + \hat{\delta}_2$$

siendo $\hat{u}_t = y_t - f(x_t, \hat{\theta}_{n-1})$.

Condiciones iniciales Si denotamos por $F(\theta)$ la función Suma de Cuadrados de Residuos,

$$\min_{\theta} SR(\hat{\theta}) = \min_{\theta} \sum_{t=1}^{T} \hat{u}_t \left(\hat{\theta}\right) = \min_{\theta} \sum_{t=1}^{T} \left(y_t - f(x_t, \theta)\right)^2 = \min_{\theta} \sum_{t=1}^{T} \left(y_t - \alpha e^{\beta x_t}\right)^2$$

que conduce a las condiciones de optimalidad,

$$\begin{aligned} \sum y_t e^{\beta x_t} &= \alpha \sum e^{2\beta x_t} \\ \sum y_t x_t e^{\beta x_t} &= \alpha \sum x_t e^{2\beta x_t} \end{aligned}$$

donde la primera condición sugiere tomar como estimación inicial,

$$\hat{\alpha} = \frac{m(ye^{\beta x})}{m(e^{2\beta x})}$$

mientras que de la segunda condición tenemos:

$$\hat{\alpha} = \frac{m(yxe^{\beta x})}{m(xe^{2\beta x})}$$

Ejercicio práctico con rutina Matlab El programa *demdir.m* analiza detalladamente este modelo $y_t = \alpha e^{\beta x_t} + u_t$.

El programa comienza generando una serie temporal de datos simulando la variable x_t a partir de un proceso i., id., $N(\mu, \sigma_x^2)$, y para el término de error del modelo a partir de un proceso $N(0, \sigma_u^2)$. Por último, generamos la serie temporal de datos para y_t utilizando la estructura del modelo y las series temporales de x_t y de u_t, una vez que hemos fijado valores numéricos para los parámetros α y β.

Con las series temporales $\{y_t, x_t\}_{t=1}^T$, podemos estimar el modelo siguiendo varios procedimientos:

- Utilizando la instrucción "fminunc" de Matlab, para minimizar la suma de cuadrados de los residuos o errores de ajuste $\underset{\alpha,\beta}{Min} \sum_{t=1}^T \left[y_t - \alpha e^{\beta x_t}\right]^2$.

- Utilizando la instrucción "fsolve" de Matlab, que encuentra las raices o soluciones de una ecuación lineal o no lineal, lo que se puede aplicar al sistema formado por las dos condiciones de optimalidad o de primer orden del problema de minimización de la suma de cuadrados de los errores,

$$\begin{aligned} -2\sum_{t=1}^T \left(y_t - \alpha e^{\beta x_t}\right) e^{\beta x} &= 0 \\ -2\sum_{t=1}^T \left(y_t - \alpha e^{\beta x_t}\right) \alpha x_t e^{\beta x} &= 0 \end{aligned}$$

- Utilizando el algoritmo de Gauss-Newton (13), con expresiones analíticas para el gradiente (10) y el hessiano (11) de la función objetivo, que es la Suma de Cuadrados de los errores de ajuste. Tenemos el gradiente y matriz hessiana,

$$\nabla F(\theta) = -2\sum \frac{\partial f(x_t,\theta)}{\partial \theta}\hat{u}_t = -2\sum \frac{\partial f_t}{\partial \theta}\hat{u}_t = -2\sum \left(e^{\beta x_t}, \alpha x_t e^{\beta x_t}\right)\hat{u}_t \quad (10)$$

$$\begin{aligned} \nabla^2 F(\theta) &= 2\sum_{t=1}^T \begin{pmatrix} e^{2\beta x_t} & \alpha x_t e^{2\beta x_t} \\ x_t \alpha e^{2\beta x_t} & \alpha^2 x_t^2 e^{2\beta x_t} \end{pmatrix} - 2\sum_{t=1}^T \begin{pmatrix} 0 & x_t e^{\beta x_t} \\ x_t e^{\beta x_t} & x_t^2 \alpha e^{\beta x_t} \end{pmatrix} \hat{u}_t \quad (11) \\ &= 2\sum_{t=1}^T \begin{pmatrix} e^{2\beta x_t} & x_t e^{\beta x_t}\left(\alpha e^{\beta x_t} - \hat{u}_t\right) \\ x_t e^{\beta x_t}\left(\alpha e^{\beta x_t} - \hat{u}_t\right) & x_t^2 \alpha e^{\beta x_t}\left(\alpha e^{\beta x_t} - \hat{u}_t\right) \end{pmatrix} \end{aligned}$$

por lo que el algoritmo de Newton-Raphson sería,

$$\hat{\theta}_n = \hat{\theta}_{n-1} - \left[\sum_{t=1}^T \begin{pmatrix} e^{2\beta x_t} & x_t e^{\beta x_t}\left(\alpha e^{\beta x_t} - \hat{u}_t\right) \\ x_t e^{\beta x_t}\left(\alpha e^{\beta x_t} - \hat{u}_t\right) & x_t^2 \alpha e^{\beta x_t}\left(\alpha e^{\beta x_t} - \hat{u}_t\right) \end{pmatrix}\right]^{-1} \left[\sum_{t=1}^T \begin{pmatrix} e^{\beta x_t} \\ \alpha x_t e^{\beta x_t} \end{pmatrix} \hat{u}_t\right] \quad (12)$$

mientras que el algoritmo de Gauss-Newton sería,

$$\hat{\theta}_n = \hat{\theta}_{n-1} - \left[\sum_{t=1}^T \begin{pmatrix} e^{2\beta x_t} & \alpha x_t e^{2\beta x_t} \\ \alpha x_t e^{2\beta x_t} & \alpha^2 x_t^2 e^{2\beta x_t} \end{pmatrix}\right]^{-1} \left[\sum_{t=1}^T \begin{pmatrix} e^{\beta x_t} \\ \alpha x_t e^{\beta x_t} \end{pmatrix} \hat{u}_t\right] \quad (13)$$

- Utilizando el algoritmo de Gauss-Newton (13), con evaluación numérica de las derivadas parciales que aparecen en el gradiente (10) y el hessiano (**??**) de la función objetivo, que es la Suma de Cuadrados de los Errores:

$$\frac{\partial f}{\partial x_i} = \lim_{\varepsilon \to 0} \frac{f(x_1, .., x_i + \varepsilon, .., x_n) - f(x_1, .., x_i - \varepsilon, .., x_n)}{2\varepsilon},\ i = 1, 2, ..., n$$

siendo las derivadas segundas: $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial g}{\partial x_j}$, donde $g = \frac{\partial f}{\partial x_i}$, de modo que:

$$\frac{\partial^2 f}{\partial x_i \partial x_j} = \lim_{\varepsilon \to 0} \frac{f(x_1, .., x_i + \varepsilon, .., x_j + \varepsilon, .., x_n) - f(x_1, .., x_i + \varepsilon, .., x_j - \varepsilon, .., x_n) - f(x_1, .., x_i - \varepsilon, .., x_j + \varepsilon, .., x_n)}{4\varepsilon^2}$$

6 Estimador de Máxima Verosimilitud

Otra estrategia de estimación consiste en utilizar un procedimiento de Máxima Verosimilitud, lo que requiere establecer un determinado supuesto acerca del tipo de distribución que sigue el término de error (innovación) del modelo. El estimador resultante es eficiente supuesto que la hipótesis acerca del tipo de distribución sea correcta. En el caso de que supongamos que $u_t \sim N(0, \sigma_u^2)$, la función de verosimilitud es,

$$L(\beta, \sigma_u^2) = \left(\frac{1}{2\pi\sigma_u^2}\right)^{T/2} \exp\left[-\frac{1}{2\sigma_u^2}\sum_{t=1}^{T}\left(y_t - f(x_t, \beta)\right)^2\right]$$

y su logaritmo,

$$\ln L(\beta, \sigma_u^2) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\sigma_u^2 - \frac{1}{2\sigma_u^2}\sum_{t=1}^{T}\left(y_t - f(x_t, \beta)\right)^2$$

cuyo gradiente, de dimensión $k+1$ hay que igualar a 0_{k+1} para obtener la estimación de Máxima Verosimilitud.

En el caso del modelo exponencial:

$$\ln L(y_t, x_t, \theta, \sigma_u^2) = -\frac{T}{2}\ln 2\pi - \frac{T}{2}\ln\sigma_u^2 - \frac{1}{2\sigma_u^2}\sum_{t=1}^{T}\left(y_t - (\alpha + \beta_1 e^{\beta_2 x_t})\right)^2$$

tendremos el conocido resultado de que, bajo el supuesto de Normalidad para el término de error, los valores numéricos para los componentes de $\theta = (\alpha, \beta_1, \beta_2, \sigma_u^2)$ que maximizan la función de verosimilitud coinciden con los valores numéricos que minimizan la suma de cuadrados de los errores de estimación.

En este procedimiento, sin embargo, a diferencia de la estimación por Mínimos Cuadrados, consideramos la estimación de la varianza del término de error, σ_u^2, simultáneamente con la de los parámetros que componen el vector

$\beta = (\alpha, \beta_1, \beta_2)$. La ecuación de optimalidad correspondiente nos dirá, como también es habitual, que la estimación de máxima verosimilitud de dicho parámetro se obtiene dividiendo por T la suma de cuadrados de los residuos que resultan al utilizar las estimaciones de máxima verosimilitud de los parámetros que entran en θ.

Si queremos maximizar el logaritmo de la función de verosimilitud, tendremos $F(\theta) = -\ln L(\beta, \sigma_u^2)$ y el algoritmo Newton-Raphson es,

$$\hat{\theta}_n = \hat{\theta}_{n-1} - \left(\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'} \right)^{-1}_{\theta=\hat{\theta}_{n-1}} \cdot \left(\frac{\partial \ln L(\theta)}{\partial\theta} \right)_{\theta=\hat{\theta}_{n-1}}$$

y el estimador resultante es asintóticamente insesgado, con distribución Normal y matriz de covarianzas,

$$Var\left(\hat{\theta}_n\right) = -\left(\frac{\partial^2 \ln L(\theta)}{\partial\theta\partial\theta'} \right)^{-1}_{\theta=\hat{\theta}_n}$$

que será definida positiva en el caso de una distribución de probabilidad Normal para la innovación del modelo, puesto que la densidad Normal es estrictamente cóncava.

El algoritmo conocido como quadratic hill-climbing consiste en sustituir en cada iteración la matriz hessiana por,

$$\nabla^2 F\left(\hat{\theta}_{n-1}\right) + \mu I_k$$

de modo que sea siempre definida positiva. Cuando esta corrección se introduce en el algoritmo de Gauss-Newton, se tiene el *algoritmo de Marquardt*.

El algoritmo de *scoring* consiste en sustituir la matriz hessiana del logaritmo de la verosimilitud, por su esperanza matemática, la matriz de información cambiada de signo, lo que simplifica mucho su expresión analítica y, por tanto, los cálculos a efectuar en cada etapa del algoritmo,

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \left[I(\hat{\theta}_{n-1}) \right]^{-1}_{\theta=\hat{\theta}_{n-1}} \cdot \left(\sum_{t=1}^{T} \frac{\partial \ln l_t(\theta)}{\partial\theta} \right)_{\theta=\hat{\theta}_{n-1}}$$

y la matriz de covarianzas del estimador resultante es, por supuesto, la inversa de la matriz de información.

Su matriz de covarianzas es la inversa de la matriz de información,

$$Var\left(\hat{\theta}_{MV}\right) = \left[I(\theta, \sigma_u^2) \right]^{-1} = \left[-E\frac{\partial^2 \ln L(\theta)}{\partial^2\theta} \right]^{-1} = -\left[\sum_{t=1}^{T} E\frac{\partial^2 \ln l_t(\theta)}{\partial^2\theta} \right]^{-1}$$

donde $\theta = \left(\beta, \sigma_u^2\right)$ y $\ln l_t(\theta)$ denota el logaritmo de la función de densidad correspondiente a un período de tiempo. En el caso habitual en que los parámetros de la matriz de covarianzas son diferentes de los parámetros que entran en el modelo $f(x_t, \beta)$, es fácil probar que esta matriz de covarianzas es diagonal a

bloques, en β y σ_u^2, por lo que la estimación del vector β y del parámetro σ_u^2 son independientes, siendo por tanto, estadísticamente eficiente llevarlas a cabo por separado.

El estimador de máxima verosimilitud es eficiente, pero nos encontramos a dos dificultades: una, la referida acerca de nuestro desconcimiento sobre si hemos alcanzado un máximo local o global; otro, que las buenas propiedades del estimador de máxima verosimilitud descansan en que el supuesto acerca de la distribución de probabilidad que sigue la innovación del modelo sea correcto. En muchas ocasiones se calcula el estimador bajo supuestos de Normalidad porque es más sencillo, aun a sabiendas de que la distribución de probabilidad de la innovación dista de ser Normal. El estimador resultante se conoce como estimador de *quasi-máxima verosimilitud*.

El algoritmo de *Gauss-Newton*, aplicado a la estimación por máxima verosimilitud, es,

$$\hat{\theta}_n = \hat{\theta}_{n-1} + \left[\sum_{t=1}^{T}\left(\frac{\partial \ln l_t(\theta)}{\partial\theta}\right)\left(\frac{\partial \ln l_t(\theta)}{\partial\theta}\right)'\right]^{-1}_{\theta=\hat{\theta}_{n-1}} \cdot \left(\sum_{t=1}^{T}\frac{\partial \ln l_t(\theta)}{\partial\theta}\right)_{\theta=\hat{\theta}_{n-1}}$$

En este caso, el algoritmo Gauss-Newton está justificado por la conocida propiedad teórica de la función de verosimilitud,

$$E\left[\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)\left(\frac{\partial \ln L(\theta)}{\partial\theta}\right)'\right] = -\left[E\frac{\partial^2 \ln L(\theta)}{\partial^2\theta}\right]^{-1}$$

y, como vemos, el agoritmo Gauss Newtpn consiiste en actualizar el vector de estiamciones mediante los coeficientes estimados en una regresión lineal de un vector de unos: 1_T , sobre el vector gradiente $\frac{\partial \ln l_t(\theta)}{\partial\theta}$.

En el caso del modelo exponencial, el gradiente de la función logaritmo de la verosimilitud es,

$$\nabla \ln L(y_t, x_t, \theta, \sigma_u^2) = \frac{1}{\sigma_u^2}\begin{pmatrix} \sum_{t=1}^{T}\hat{u}_t \\ \sum_{t=1}^{T} e^{\beta_2 x_t}\hat{u}_t \\ \sum_{t=1}^{T}\beta_1 x_t e^{\beta_2 x_t}\hat{u}_t \\ -\frac{T}{2\sigma_u^2} + \frac{1}{2(\sigma_u^2)^2}\sum \hat{u}_t^2 \end{pmatrix}$$

Para este modelo exponencial, la matriz hessiana es:

$$H = -\frac{1}{\sigma_u^2}\sum_{t=1}^{T}\begin{pmatrix} 1 & e^{\beta_2 x_t} & \beta_1 x_t e^{\beta_2 x_t} & -\frac{1}{\sigma_u^2}\sum_{t=1}^{T}\hat{u}_t \\ e^{\beta_2 x_t} & e^{2\beta_2 x_t} & \beta_1 x_t e^{2\beta_2 x_t} & -\frac{1}{\sigma_u^2}\sum_{t=1}^{T} e^{\beta_2 x_t}\hat{u}_t \\ \beta_1 x_t e^{\beta_2 x_t} & \beta_1 x_t e^{2\beta_2 x_t} & \beta_1^2 x_t^2 e^{2\beta_2 x_t} & -\frac{1}{\sigma_u^2}\sum_{t=1}^{T}\beta_1 x_t e^{\beta_2 x_t}\hat{u}_t \\ -\frac{1}{\sigma_u^2}\sum_{t=1}^{T}\hat{u}_t & -\frac{1}{\sigma_u^2}\sum_{t=1}^{T} e^{\beta_2 x_t}\hat{u}_t & -\frac{1}{\sigma_u^2}\sum_{t=1}^{T}\beta_1 x_t e^{\beta_2 x_t}\hat{u}_t & \frac{T}{2(\sigma_u^2)^2} - \frac{1}{(\sigma_u^2)^3}\sum \hat{u}_t^2 \end{pmatrix}$$

Al tomar la esperanza matemática de los elementos de la matriz hessiana y cambiar su signo, obtenemos la matriz de información, que tendrá ceros en

la última fila y columna, correspondientes a la estimación de σ_u^2, excepto en su elemento diagonal.

$$I\left(\theta,\sigma_u^2\right)=\frac{1}{\sigma_u^2}\sum_{t=1}^{T}\begin{pmatrix}1 & e^{\beta_2 x_t} & \beta_1 x_t e^{\beta_2 x_t} & 0\\ e^{\beta_2 x_t} & e^{2\beta_2 x_t} & \beta_1 x_t e^{2\beta_2 x_t} & 0\\ \beta_1 x_t e^{\beta_2 x_t} & \beta_1 x_t e^{2\beta_2 x_t} & \beta_1^2 x_t^2 e^{2\beta_2 x_t} & 0\\ 0 & 0 & 0 & \frac{T}{2(\sigma_u^2)^2}\end{pmatrix}$$

que demuestra que el estimador de máxima verosimilitud de dicho modelo es estadísticamente independiente de los estimadores de los restantes parámetros, lo que no sucede con los estimadores de máxima verosimilitud de estos entre sí, que tienen covarianzas no nulas.

La puesta en práctica del algoritmo anterior requiere obtener las expresiones analíticas de las derivadas primeras y segundas de la función F , si se va a seguir un algoritmo del tipo Newton-Raphson. Ello significa calcular $k\left(k+3\right)/2$ derivadas, que hay que evaluar para cada dato, utilizando los valores numéricos de los parámetros que en ese momento se tienen como estimación, lo que puede ser un gran trabajo. Para evitar esta tarea pueden adoptarse algunas posibles soluciones: a) utilizar el algoritmo de Gauss-Newton, que sustituye el hessiano $\nabla^2 F\left(\hat{\theta}_0\right)$ por el producto del vector gradiente por sí mismo, $\nabla F\left(\hat{\theta}_0\right)\nabla F\left(\hat{\theta}_0\right)'$, lo que evita trabajar con derivadas segundas, b) sustituir las derivadas analíticas por derivadas numéricas. Para ello, cuando disponemos de un vector de estimaciones $\hat{\theta}_{n-1}$, variamos ligeramente uno de los parámetros, y evaluamos numéricamente la función objetivo en el vector resultante. El cambio en el valor numérico de F, dividido por la variación introducida en el parámetro considerado, nos da una aproximación numérica a la derivada parcial con respecto a dicho parámetro, evaluada en el vector de estimaciones disponibles en ese momento, c) utilizar el *algoritmo de scoring*, que, al tomar esperanzas matemáticas, simplifica mucho las expresiones analiticas de las derivadas.

7 Zero coupon curve estimation

7.1 Modelo polinómico

Before describing the use of the Principal Component technique for risk management in fixed income markets, let us remember the main idea behind zero coupon curve estimation.

Note: Zero coupon curves are estimated using market prices for bonds that pay coupon. As illustration for those of you interested, I leave the 'polynomial zero coupon curve.xls' file, that solves the following exercise. A .zip file named 'nelson_siegel' will also be made available for those of yo interested in estimating Nelson-Siegel and Svensson models of zero coupon curves using Matlab.

Consider the following exercise. Today is November 5, 2011. The first column of file 'polynomial zero coupon curve.xls' contains the coupon of each bond traded in the secondary market for Government debt. The second column contains the maturity date, the third column the date the bond was first issued, which is assumed to be the same for all bonds, 15/08/2011. Each bond is assumed to have a nominal of 100 monetary units. This is just for simplification, and it cold be changed without any difficulty. Finally, we see the (average) market price for each bond.

We assume a polynomial discount function,

$$d(t) = a + bt + ct^2 + dt^3 + et^4$$

to be applied to each cash flow.

Hence, the price of a bond can be represented:

$$\begin{aligned} P_{it} &= \sum_{j=1}^{n_i} c_{ij} d_j(t) = \sum_{j=1}^{n_i} c_{ij} \left(a + bt_{ij} + ct_{ij}^2 + dt_{ij}^3 + et_{ij}^4\right) = \\ &= a\sum_{j=1}^{n_i} c_{ij} a + b\sum_{j=1}^{n_i} c_{ij} t_{ij} + c\sum_{j=1}^{n_i} c_{ij} t_{ij}^2 + d\sum_{j=1}^{n_i} c_{ij} t_{ij}^3 + e\sum_{j=1}^{n_i} c_{ij} t_{ij}^4 \end{aligned}$$

where n_i denotes the number of cash-flows to be paid by the i-th bond before maturity. We assume that all bonds pay coupon each semester (half of the annual amount).

For each vector of parameter values (a, b, c, d) we have a theoretical price for each bond. We want to find the parameter values so that

$$\underset{(a,b,c,d)}{Min} \sum_{i=1}^{N} (P_{it}^M - P_{it}^T)^2$$

where P_{it}^M denotes the market price for each bond, and P_{it}^T denotes the theoretical price for that parameter vector.

The market price is 'ex coupon', meaning that we need to add to it the part of the coupon which would correspond to the current holder since the last date that a coupon was paid. To calculate that amount, we multiply the size of the next coupon payment by the proportion of the 2-month interval that has already gone by. Adding that to the 'ex coupon' market price, we get the true traded price.

The polynomial function $d_j(t)$ is the discount function, giving us the price of a bond that would mature at any future date, with a single payment, to be effective at maturity. This would be a zero coupon bond maturing t periods from now.

Estimate a discount function using a polynomial of degree 2, and another one using a polynomial of degree 4, and represent both discount functions. Draw a bar diagram with the market and the theoretical prices for each bond under each specification of the discount function.

The zero coupon curve itself, that represents zero coupon interest rates as a function of maturity, is obtained from:

$$r_t = 100\left(\left(\frac{1}{d_t}\right)^{1/t} - 1\right)$$

7.2 Modelo de Nelson Siegel

El modelo de Nelson y Siegel parte de una representación del tipo instantáneo que en t se espera que sea aplicable dentro de s periodos:

$$\varphi_t(s) = \beta_0 + \beta_1 e^{-s/\tau} + \beta_2 \frac{s}{\tau} e^{-s/\tau}$$

Por lo que el tipo de interés de contado (cupón cero) a plazo t_i es:

$$r_t(t_i) = \frac{1}{t_i}\int_0^{t_i} \varphi_t(s)ds$$

Integrando:

$$\begin{aligned}
\int_0^{t_i} e^{-s/\tau} ds &= -\tau e^{-s/\tau} \mid_0^{t_i} = -\tau e^{-t_i/\tau} + \tau \\
\int_0^{t_i} \frac{s}{\tau} e^{-s/\tau} ds &= \int_0^{t_i} \frac{s}{\tau} d\left(-\tau e^{-s/\tau}\right) = s e^{-s/\tau} \mid_0^{t_i} + \int_0^{t_i} \tau \frac{1}{\tau} e^{-s/\tau} ds = t_i e^{-t_i/\tau} - \tau e^{-t_i/\tau} + \tau
\end{aligned}$$

por lo que:

$$\begin{aligned}
r_t(t_i) &= \frac{1}{t_i}\int_0^{t_i} \varphi_t(s)ds = \frac{1}{t_i}\left[\beta_0 t_i + \beta_1\left(-\tau e^{-t_i/\tau} + \tau\right) + \beta_2\left(t_i e^{-t_i/\tau} - \tau e^{-t_i/\tau} + \tau\right)\right] = \\
&= \beta_0 + (\beta_1 + \beta_2)\frac{\tau}{t_i} - (\beta_1 + \beta_2)\frac{\tau}{t_i} e^{-t_i/\tau} - \beta_2 e^{-t_{i/\tau}}
\end{aligned}$$

El precio teórico de un bono en el instante t de acuerdo con este modelo será entonces:

$$P_t^{NS} = \sum_{i=1}^{k} C_{t_i} e^{-r_t(t_i)} = \sum_{i=1}^{k} C_{t_i} \exp\left[\beta_0 + (\beta_1 + \beta_2)\frac{\tau}{t_i}\left(1 - e^{-t_i/\tau}\right) - \beta_2 e^{-t_{i/\tau}}\right]$$

Para estimar los parámetros $\theta = (\beta_0, \beta_1, \beta_2, \tau)$ del modelo a partir de los precios de mercado de n bonos, resolvemos:

$$\min_{\theta} \sum_{i=1}^{n} \left(P_{it}^m - P_{it}\right)^2$$

o podemos utilizar ponderaciones:

$$\min_{\theta} \sum_{i=1}^{n} \omega_i \left(P_{it}^m - P_{it}\right)^2$$

7.3 Modelo de Svensson (1994)

Tipo instantáneo que en t se espera que se aplicable dentro de s periodos:

$$\varphi_t(s) = \beta_0 + \beta_1 e^{-s/\tau_1} + \beta_2 \frac{s}{\tau_1} e^{-s/\tau_1} + \beta_3 \frac{s}{\tau_2} e^{-s/\tau_2}$$

Tipo de interés de contado (cupón cero) a plazo t_i:

$$r_t(t_i) = \frac{1}{t_i} \int_0^{t_i} \varphi_t(s) ds$$

Integrando:

$$\begin{aligned}
\int_0^{t_i} e^{-s/\tau_1} ds &= -\tau_1 e^{-s/\tau_1} \mid_0^{t_i} = -\tau_1 e^{-t_i/\tau_1} + \tau_1 \\
\int_0^{t_i} \frac{s}{\tau} e^{-s/\tau} ds &= \int_0^{t_i} \frac{s}{\tau} d\left(-\tau e^{-s/\tau}\right) = s e^{-s/\tau} \mid_0^{t_i} + \int_0^{t_i} \tau \frac{1}{\tau} e^{-s/\tau} ds = t_i e^{-t_i/\tau} - \tau e^{-t_i/\tau} + \tau
\end{aligned}$$

por lo que el tipo cupón cero a plazo t_i es, de acuerdo con este modelo:

$$\begin{aligned}
r_t(t_i) &= \frac{1}{t_i} \int_0^{t_i} \varphi_t(s) ds = \frac{1}{t_i} [\beta_0 t_i + \beta_1 \left(-\tau_1 e^{-t_i/\tau_1} + \tau_1\right) + \beta_2 \left(t_i e^{-t_i/\tau_1} - \tau_1 e^{-t_i/\tau_1} + \tau_1\right) + \\
&\quad + \beta_3 \left(t_i e^{-t_i/\tau_2} - \tau_2 e^{-t_i/\tau_2} + \tau_2\right)] \\
&= \beta_0 + (\beta_1 + \beta_2) \frac{\tau_1}{t_i} - (\beta_1 + \beta_2) \frac{\tau_1}{t_i} e^{-t_i/\tau_1} - \beta_2 e^{-t_{i/\tau_1}} + \beta_3 \left(-e^{-t_i/\tau_2} - \frac{\tau_2}{t_i} e^{-t_i/\tau_2} + \frac{\tau_2}{t_i}\right) = \\
&= \beta_0 + (\beta_1 + \beta_2) \frac{\tau_1}{t_i} \left(1 - e^{-t_i/\tau_1}\right) - \beta_2 e^{-t_{i/\tau_1}} + \beta_3 \frac{\tau_2}{t_i} \left(1 - e^{-t_i/\tau_2}\right) - \beta_3 e^{-t_i/\tau_2}
\end{aligned}$$

El precio teórico de un bono en el instante t de acuerdo con este modelo será entonces:

$$P_t^S = \sum_{i=1}^{k} C_{t_i} e^{-r_t(t_i)} = \sum_{i=1}^{k} C_{t_i} \exp \left[\begin{array}{c} \beta_0 + (\beta_1 + \beta_2) \frac{\tau_1}{t_i} \left(1 - e^{-t_i/\tau_1}\right) - \beta_2 e^{-t_{i/\tau_1}} + \\ \beta_3 \frac{\tau_2}{t_i} \left(1 - e^{-t_i/\tau_2}\right) - \beta_3 e^{-t_i/\tau_2} \end{array} \right]$$

8 Un modelo general de tipos de interés

Para explicar la evolución temporal de los tipos de interés, consideremos la siguiente ecuación diferencial estocástica

$$dr_t = (\alpha + \beta r_t)\, dt + \sigma r_t^{\gamma} dW_t$$

como en Chan et al. (1992a) [CKLS], donde $r_t, t > 0$, es un proceso estocástico real en tiempo continuo, y α, β, γ y σ son parámetros estructurales cuyo valor numérico es desconocido. Esta ecuación general anida como casos particulares diversos modelos que han sido propuestos en la literatura.

Discretización exacta

Bergstrom (1984) prueba que el modelo discreto correspondiente al anterior es,

$$r_t = e^{\beta} r_{t-1} + \frac{\alpha}{\beta}\left(e^{\beta} - 1\right) + \eta_t,\ t = 1, 2, ..., T \tag{14}$$

con,

$$E\left(\eta_t \eta_s\right) = 0,\ s \neq t;\ E\left(\eta_t^2\right) = \frac{\sigma^2}{2\beta}\left(e^{2\beta} - 1\right) r_{t-1}^{2\gamma} = m_t^2$$

Si denotamos por $\theta = \left(\alpha, \beta, \gamma, \sigma^2\right)$ el vector de parámetros del modelo, tenemos el logaritmo de la función de verosimilitud $L(\theta)$,

$$L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{\left[r_t - e^{\beta} r_{t-1} - \frac{\alpha}{\beta}\left(e^{\beta} - 1\right)\right]^2}{m_t^2}\right);\ m_t^2 = \frac{\sigma^2}{2\beta}\left(e^{2\beta} - 1\right) r_{t-1}^{2\gamma}$$

y tenemos,

$$L(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(2\ln m_t + \varepsilon_t^2\right)$$

donde $\varepsilon_t, t = 1, 2, ..., T$ puede calcularse utilizando,

$$m_t \varepsilon_t = \eta_t$$

ya que ε_t no es sino una versión de η_t normalizado en varianza.

Discretización aproximada

Una discretización rápida del modelo en tiempo continuo puede obtenerse como,

$$r_t - r_{t-1} = \alpha + \beta r_{t-1} + \eta_t \tag{15}$$

con:

$$\begin{aligned} E\eta_t &= 0 \\ E\left(\eta_t^2\right) &= \sigma^2 r_{t-1}^{2\gamma} \end{aligned} \tag{16}$$

La aproximación lineal de la función e^β alrededor de $\beta = 0$ es: $e^\beta = 1 + \beta$, por lo que (14) puede escribirse,

$$r_t = (1+\beta)\, r_{t-1} + \alpha + \eta_t,\ t = 1, 2, ..., T$$

que coincide con (15), lo que nos da una idea de la diferencia entre ambas expresiones, que será mayor cuanto mayor sea el valor absoluto de β.

Bajo Normalidad del término de error tendremos la función de verosimilitud,

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \gamma\sum_{t=2}^{T}\ln r_{t-1} - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(\frac{(r_t - r_{t-1}) - \alpha - \beta r_{t-1}}{r_{t-1}^{\gamma}}\right)^2$$

Veremos más adelante que cuando $\beta = 0$ las dos discretizaciones, exacta y aproximada, coinciden.

8.1 Estimación por Máxima Verosimilitud

Si queremos estimar la discretización exacta, es razonable utilizar (**??**) para obtener condiciones iniciales para la estimación por maxima verosimilitud mediante una regresión lineal de r_t sobre r_{t-1} :

$$r_t = \delta_0 + \delta_1 r_{t-1} + u_t$$

para obtener: $\hat{\delta}_o, \hat{\delta}_1, \hat{\sigma}_u^2$. A partir de aqui, puesto que: $\delta_0 = \frac{\alpha}{\beta}\left(e^\beta - 1\right)$ y $\delta_1 = e^\beta$, obtenemos estimaciones iniciales mediante:

$$\hat{\beta} = \ln(\hat{\delta}_1);\ \hat{\alpha} = \frac{\hat{\beta}}{\exp(\hat{\beta}) - 1}\hat{\delta}_0;\ \hat{\sigma}_t^2 = \frac{\hat{\sigma}_u^2}{2\hat{\beta}}\left(e^{2\hat{\beta}} - 1\right) r_{t-1}^{2\gamma}$$

Para estimar γ, estimamos una regresión auxiliar de los residuos al cuadrado, como proxy de la varianza en cada periodo:

$$\ln\hat{\eta}_t^2 = \xi_0 + \xi_1 \ln r_{t-1}$$

en la que, segun el modelo teórico: $\xi_0 = \ln\left[\frac{\hat{\sigma}^2}{2\hat{\beta}}\left(e^{2\hat{\beta}} - 1\right)\right]$; $\xi_1 = 2\gamma$, de donde estimamos:

$$\hat{\sigma}^2 = \frac{2\hat{\beta}}{e^{2\hat{\beta}} - 1}e^{\hat{\xi}_0};\ \ \hat{\gamma} = \hat{\xi}_1/2;$$

y los cuatro valores numéricos $(\hat{\alpha}, \hat{\beta}, \hat{\gamma}, \hat{\sigma}^2)$ se llevan a la función de verosimilitud para iniciar el proceso de optimización numérica:

$$\min_{(\alpha,\beta,\gamma,\sigma^2)} L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{\left[r_t - e^{\beta}r_{t-1} - \frac{\alpha}{\beta}\left(e^{\beta}-1\right)\right]^2}{m_t^2}\right);\ m_t^2 = \frac{\sigma^2}{2\beta}\left(e^{2\beta}-1\right)r_{t-1}^{2\gamma}$$

El programa *estima.m* y la función asociada *loglik.m* llevan a cabo este ejercicio de estimación.

Para estimar la discretización aproximada bajo Normalidad, maximizaremos:

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \gamma\sum_{t=2}^{T}\ln r_{t-1} - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(\frac{(r_t - r_{t-1}) - \alpha - \beta r_{t-1}}{r_{t-1}^{\gamma}}\right)^2$$

lo cual, en el caso general, debe hacerse por procedimientos numéricos. Podria hacerse condicionando en un valor de γ, lo que se llevaria a cabo estableciendo una rejilla de valores de dicho parámetro y estimando condicional en cada uno de ellos, de modo similar a como se describe en el modelo CIR que se explica más abajo.

8.1.1 Merton (1973): $\beta = 0, \gamma = 0$

Con $\gamma = 0$, la varianza es constante en este modelo:

$$dr_t = \alpha dt + \sigma dW_t$$

Notemos que $\lim_{\beta\to 0}\frac{e^{\beta}-1}{\beta} = 1$, $\lim_{\beta\to 0}\frac{\sigma^2}{2\beta}\left(e^{2\beta}-1\right) = \sigma^2$, por lo que tenemos:

Discretización exacta:

$$r_t = r_{t-1} + \alpha + \eta_t,\ t = 1, 2, ..., T$$

con función de verosimilitud (bajo Normalidad),

$$L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{[r_t - r_{t-1} - \alpha]^2}{m_t^2}\right);\ m_t^2 = \sigma^2,\ \text{constante}$$

La estructura de dicha función de verosimilitud revela que la estimación de α ha de ser la media muestral de las variaciones en el nivel del tipo de interés, $\hat{\alpha} = \frac{1}{T}\sum_{t=1}^{T}(r_t - r_{t-1}) = \overline{\Delta r_t}$, con $\Delta r_t = r_t - r_{t-1}$,mientras que la estimación de σ^2 es la suma de cuadrados de los errores de ajuste, dividida por el tamaño muestral: $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(r_t - r_{t-1} - \hat{\alpha}\right)^2 = var(\Delta r_t)$.

Discretización aproximada:

$$r_t - r_{t-1} = \alpha + \eta_t;\ E\eta_t = 0;\ E\left(\eta_t^2\right) = \sigma^2$$

con funciones de verosimilitud,

$$L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{[r_t - r_{t-1} - \alpha]^2}{m_t^2}\right);\ m_t^2 = \sigma^2,\ \text{constante}$$

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left((r_t - r_{t-1}) - \alpha\right)^2$$

En este caso, las funciones de verosimilitud de ambas discretizaciones coinciden.

8.1.2 Vasicek (1977): $\gamma = 0$

Modelo en tiempo continuo,

$$dr_t = (\alpha + \beta r_t)\,dt + \sigma dW_t$$

Discretización exacta,

$$r_t = e^{\beta}r_{t-1} + \frac{\alpha}{\beta}\left(e^{\beta} - 1\right) + \eta_t,\ Var(\eta_t) = \frac{\sigma^2}{2\beta}\left(e^{2\beta} - 1\right), t = 1, 2, ..., T$$

con función de verosimilitud,

$$L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{\left[r_t - e^{\beta}r_{t-1} - \frac{\alpha}{\beta}\left(e^{\beta} - 1\right)\right]^2}{m_t^2}\right);\ m_t^2 = \frac{\sigma^2}{2\beta}\left(e^{2\beta} - 1\right)$$

Las condiciones iniciales para α y β se obtienen como en el caso del modelo general, a partir de la regresión lineal de r_t sobre r_{t-1} : $r_t = \delta_0 + \delta_1 r_{t-1} + \eta_t$. En este caso m_t es constante, y una estimación inicial es: $\sigma^2 = \frac{2\beta}{e^{2\beta}-1}Var(\hat{\eta}_t)$. Los valores numéricos $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ se llevan a la función de verosimilitud para iniciar el proceso de optimización numérica. El programa *estima_ vasicek.m* y la función asociada *vasicek.m* realizan esta estimación.

Discretización aproximada

$$r_t - r_{t-1} = \alpha + \beta r_{t-1} + \eta_t;\ E\eta_t = 0;\ E\left(\eta_t^2\right) = \sigma^2$$

con función de verosimilitud:

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left((r_t - r_{t-1}) - \alpha - \beta r_{t-1}\right)^2$$

que se maximiza mediante:

$$\hat{\beta} = \frac{\sum_{t=1}^{T}(r_t - r_{t-1})\,r_{t-1}}{\sum_{t=1}^{T}(r_{t-1} - \bar{r})^2},\ \hat{\alpha} = \sum_{t=1}^{T}\frac{r_t - r_{t-1}}{T} - \hat{\beta}\sum_{t=1}^{T}\frac{r_{t-1}}{T},\ \hat{\sigma}^2 = \sum_{t=1}^{T}\frac{\left(r_t - r_{t-1} - \hat{\alpha} - \hat{\beta}r_{t-1}\right)^2}{T}$$

8.1.3 Cox, Ingersoll, Ross (1985): $\gamma = 1/2$.

Modelo en tiempo continuo:

$$dr_t = (\alpha + \beta r_t)\, dt + \sigma\sqrt{r_t} dW_t$$

Discretización exacta,

$$r_t = e^{\beta} r_{t-1} + \frac{\alpha}{\beta}\left(e^{\beta} - 1\right) + \eta_t,\ Var(\eta_t) = \frac{\sigma^2}{2\beta}\left(e^{2\beta} - 1\right) r_{t-1},\ t = 1, 2, ..., T$$

con función de verosimilitud,

$$L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{\left[r_t - e^{\beta} r_{t-1} - \frac{\alpha}{\beta}\left(e^{\beta} - 1\right)\right]^2}{m_t^2}\right);\ m_t^2 = \frac{\sigma^2}{2\beta}\left(e^{2\beta} - 1\right) r_{t-1}$$

Las condiciones iniciales para α y β se obtienen como en el caso del modelo general, a partir de la regresión lineal de r_t sobre r_{t-1} : $r_t = \delta_0 + \delta_1 r_{t-1} + \eta_t$. En la regresión auxiliar: $\ln\hat{\eta}_t^2 = \xi_0 + \xi_1 \ln r_{t-1}$, el coeficiente ξ_1 es ahora: $\xi_1 = 1$, por lo que tenemos:[2] $\xi_0 = T^{-1}\sum_{t=1}^{T}(\ln\hat{u}_t^2 - \ln r_{t-1})$ que nos permite estimar $\hat{\xi}_0$ y a continuación recuperar la estimación de σ^2:

$$\hat{\sigma}^2 = \frac{2\hat{\beta}}{e^{2\hat{\beta}} - 1} e^{\hat{\xi}_0},$$

y las estimaciones $(\hat{\alpha}, \hat{\beta}, \hat{\sigma}^2)$ se llevan a la función de verosimilitud para iniciar el proceso de optimización numérica.

Discretización aproximada

$$r_t - r_{t-1} = \alpha + \beta r_{t-1} + \eta_t;\ E\eta_t = 0;\ E\left(\eta_t^2\right) = \sigma^2 r_{t-1}$$

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \frac{1}{2}\sum_{t=2}^{T}\ln r_{t-1} - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(\frac{(r_t - r_{t-1}) - \alpha - \beta r_{t-1}}{\sqrt{r_{t-1}}}\right)^2$$

La verosimilitud aproximada se maximiza aplicando mínimos cuadrados generalizados, tras imponer la estructura de heterocedasticidad teórica de este modelo, es decir, estimando por mínimos cuadrados ordinarios el modelo,

$$\frac{r_t - r_{t-1}}{\sqrt{r_{t-1}}} = \alpha\frac{1}{\sqrt{r_{t-1}}} + \beta\sqrt{r_{t-1}} + \frac{\eta_t}{\sqrt{r_{t-1}}};\ Var\left(\frac{\eta_t}{\sqrt{r_{t-1}}}\right) = \frac{\sigma^2 r_{t-1}}{r_{t-1}} = \sigma^2$$

obteniendo así las estimaciones de α y β y, posteriormente, $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(r_t - r_{t-1} - \hat{\alpha} - \hat{\beta} r_{t-1}\right)^2$.

[2] También podriamos mantener la estimacion del modelo general: $\xi_0 = \ln\left[\frac{\hat{\sigma}^2}{2\hat{\beta}}\left(e^{2\hat{\beta}} - 1\right)\right] \Rightarrow \hat{\sigma}^2 = \frac{2\hat{\beta}}{e^{2\hat{\beta}} - 1} e^{\hat{\xi}_0}$

8.1.4 Dothan: $\alpha = 0, \beta = 0, \gamma = 1$

Modelo en tiempo continuo

$$dr_t = \sigma r_t dW_t$$

Discretización exacta:

$$r_t = r_{t-1} + \eta_t,\ t = 1, 2, ..., T$$

Discretización aproximada:

$$r_t - r_{t-1} = \eta_t;\ E\eta_t = 0;\ E\left(\eta_t^2\right) = \sigma^2 r_{t-1}^2$$

con funciones de verosimilitud,

$$L_e\left(\theta\right) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{\left[r_t - r_{t-1}\right]^2}{m_t^2}\right);\ m_t^2 = \sigma^2 r_{t-1}^2$$

$$L_a\left(\theta\right) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \sum_{t=2}^{T}\ln r_{t-1} - \frac{1}{2\sigma^2}\sum_{t=2}^{T}\left(\frac{r_t - r_{t-1}}{r_{t-1}}\right)^2$$

Ambas funciones de verosimilitud coinciden, y se maximizan mediante $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\frac{(r_t - r_{t-1})^2}{r_{t-1}^2}$.

8.1.5 Movimiento browniano geométrico: $\alpha = 0, \gamma = 1$

Modelo en tiempo continuo

$$dr_t = \beta r_t dt + \sigma r_t dW_t$$

Discretización exacta:

$$r_t = e^{\beta} r_{t-1} + \eta_t,\ t = 1, 2, ..., T$$

con funciones de verosimilitud (bajo una distribución Normal),

$$L_e\left(\theta\right) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{\left[r_t - e^{\beta} r_{t-1}\right]^2}{m_t^2}\right);\ m_t^2 = \frac{\sigma^2}{2\beta}\left(e^{2\beta} - 1\right) r_{t-1}^2$$

Discretización aproximada:

$$r_t - r_{t-1} = \beta r_{t-1} + \eta_t;\ E\eta_t = 0;\ E\left(\eta_t^2\right) = \sigma^2 r_{t-1}^2$$

con verosimililtud:

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \sum_{t=2}^{T}\ln r_{t-1} - \frac{1}{\sigma^2}\sum_{t=2}^{T}\left(\frac{(r_t - r_{t-1}) - \beta r_{t-1}}{r_{t-1}}\right)^2$$

La verosimilitud aproximada se maximiza aplicando mínimos cuadrados generalizados, tras imponer la estructura de heterocedasticidad de este modelo, es decir, estimando por mínimos cuadrados ordinarios el modelo,

$$\frac{r_t - r_{t-1}}{r_{t-1}} = \beta + \frac{\eta_t}{r_{t-1}};\ Var\left(\frac{\eta_t}{r_{t-1}}\right) = \frac{\sigma^2 r_{t-1}^2}{r_{t-1}^2} = \sigma^2$$

obteniendo así la estimación de β, $\hat{\beta} = \frac{1}{T}\sum_{t=1}^{T}\frac{r_t - r_{t-1}}{r_{t-1}}$ y, posteriormente, $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{r_t - r_{t-1}}{r_{t-1}} - \hat{\beta}\right)^2$.

8.1.6 Brennan y Schwartz (1980): $\gamma = 1$

Modelo en tiempo continuo

$$dr_t = (\alpha + \beta r_t)\,dt + \sigma r_t dW_t$$

Discretización exacta:

$$r_t = e^{\beta} r_{t-1} + \frac{\alpha}{\beta}\left(e^{\beta} - 1\right) + \eta_t,\ t = 1, 2, ..., T$$

con función de verosimilitud (bajo innovaciones Normales),

$$L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{\left[r_t - e^{\beta} r_{t-1} - \frac{\alpha}{\beta}\left(e^{\beta} - 1\right)\right]^2}{m_t^2}\right);\ m_t^2 = \frac{\sigma^2}{2\beta}\left(e^{2\beta} - 1\right) r_{t-1}^2$$

Discretización aproximada:

$$r_t - r_{t-1} = \alpha + \beta r_{t-1} + \eta_t;\ E\eta_t = 0;\ E\left(\eta_t^2\right) = \sigma^2 r_{t-1}^2$$

con verosimilitud:

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \sum_{t=2}^{T}\ln r_{t-1} - \frac{1}{\sigma^2}\sum_{t=2}^{T}\left(\frac{(r_t - r_{t-1}) - \alpha - \beta r_{t-1}}{r_{t-1}}\right)^2$$

La verosimilitud aproximada se maximiza aplicando mínimos cuadrados generalizados, tras imponer la estructura de heterocedasticidad de este modelo, es decir, estimando por mínimos cuadrados ordinarios el modelo,

$$\frac{r_t - r_{t-1}}{r_{t-1}} = \alpha\frac{1}{r_{t-1}} + \beta + \frac{\eta_t}{r_{t-1}};\ Var\left(\frac{\eta_t}{r_{t-1}}\right) = \frac{\sigma^2 r_{t-1}^2}{r_{t-1}^2} = \sigma^2$$

obteniendo así las estimaciones de α y β y, posteriormente, $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{r_t - r_{t-1}}{r_{t-1}} - \left(\hat{\alpha}\frac{1}{r_{t-1}} + \hat{\beta}\right)\right]^2$.

8.1.7 Cox, Ingersoll, Ross (180): $\alpha = 0, \beta = 0, \gamma = 3/2$.

Modelo en tiempo continuo

$$dr_t = \sigma r_t^{3/2} dW_t$$

Discretización exacta:

$$r_t = r_{t-1} + \eta_t,\ t = 1, 2, ..., T$$

con función de verosimilitud bajo Normalidad,

$$L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{[r_t - r_{t-1}]^2}{m_t^2}\right);\ m_t^2 = \sigma^2 r_{t-1}^3$$

Discretización aproximada:

$$r_t - r_{t-1} = \eta_t;\ E\eta_t = 0;\ E\left(\eta_t^2\right) = \sigma^2 r_{t-1}^3$$

con verosimililtud:

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \frac{3}{2}\sum_{t=2}^{T}\ln r_{t-1} - \frac{1}{\sigma^2}\sum_{t=2}^{T}\left(\frac{r_t - r_{t-1}}{r_{t-1}^3}\right)^2$$

Ambas funciones de verosimilitud coinciden, y se maximizan aplicando mínimos cuadrados generalizados, tras imponer la estructura de heterocedasticidad de este modelo, es decir, estimando por mínimos cuadrados ordinarios el modelo,

$$\frac{r_t - r_{t-1}}{\sqrt{r_{t-1}^3}} = \alpha\frac{1}{\sqrt{r_{t-1}^3}} + \beta\frac{1}{\sqrt{r_{t-1}}} + \frac{\eta_t}{\sqrt{r_{t-1}^3}};\ Var\left(\frac{\eta_t}{\sqrt{r_{t-1}^3}}\right) = \frac{\sigma^2 r_{t-1}^3}{r_{t-1}^3} = \sigma^2$$

obteniendo así las estimaciones de α y β y, posteriormente, $\hat{\sigma}^2 = \frac{1}{T}\sum_{t=1}^{T}\left[\frac{r_t - r_{t-1}}{\sqrt{r_{t-1}^3}} - \left(\hat{\alpha}\frac{1}{\sqrt{r_{t-1}^3}} + \hat{\beta}\frac{1}{\sqrt{r_{t-1}}}\right)\right]^2$.

8.1.8 Elasticidad de la varianza constante: $\alpha = 0$.

Modelo en tiempo continuo

$$dr_t = \beta r_t dt + \sigma r_t^{\gamma} dW_t$$

Discretización exacta,

$$r_t = e^{\beta} r_{t-1} + \eta_t,\ t = 1, 2, ..., T$$

con función de verosimilitud bajo Normalidad,

$$L_e(\theta) = -\frac{T}{2}\ln 2\pi - \frac{1}{2}\sum_{t=2}^{T}\left(\ln m_t^2 + \frac{\left[r_t - e^{\beta}r_{t-1}\right]^2}{m_t^2}\right);\ m_t^2 = \frac{\sigma^2}{2\beta}\left(e^{2\beta}-1\right)r_{t-1}^{2\gamma}$$

Discretización aproximada

$$r_t - r_{t-1} = \beta r_{t-1} + \eta_t;\ E\eta_t = 0;\ E\left(\eta_t^2\right) = \sigma^2 r_{t-1}^{2\gamma}$$

con verosimilitud:

$$L_a(\theta) = -\frac{T}{2}\ln 2\pi - T\ln\sigma - \gamma\sum_{t=2}^{T}\ln r_{t-1} - \frac{1}{\sigma^2}\sum_{t=2}^{T}\left(\frac{(r_t - r_{t-1}) - \beta r_{t-1}}{r_{t-1}^{\gamma}}\right)^2$$

La maximización de la función de verosimilitud aproximada puede llevarse a cabo condicionando en un valor numérico de γ, para aplicar mínimos cuadrados generalizados, estimando el modelo

$$\frac{r_t - r_{t-1}}{r_{t-1}^{\gamma}} = \beta r_{t-1}^{1-\gamma} + \frac{\eta_t}{r_{t-1}^{\gamma}};\ E\eta_t = 0;\ E\left(\frac{\eta_t}{r_{t-1}^{\gamma}}\right)^2 = \frac{\sigma^2 r_{t-1}^{2\gamma}}{r_{t-1}^{2\gamma}} = \sigma^2$$

para obtener $\hat{\beta}(\gamma)$ y, posteriormente, $\hat{\sigma}^2(\gamma) = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{r_t - r_{t-1}}{r_{t-1}^{\gamma}} - \hat{\beta}r_{t-1}^{1-\gamma}\right)^2$. Una vez realizado este ejercicio para una red de valores de γ, seleccionaríamos aquél que proporciona la menor estimación de $\hat{\sigma}^2(\gamma)$, junto con la estimación asociada de β.

9 Método Generalizado de Momentos

El Método Generalizado de Momentos se basa en condiciones de ortogonalidad del tipo:

$$E\left[h(\theta_0, w_t)\right] = 0 \tag{17}$$

donde w_t es un $m-$vector de variables observables en el período t, θ_0 es un vector de k parámetros, y h es un vector de r funciones reales: $h : R^m x R^k \longrightarrow R^r$. Es muy importante tener en cuenta que, en este método de estimación, juega un papel fundamental el período en que las variables pasan a ser conocidas, con independencia del subíndice que tengan.

Algunos modelos teóricos implican este tipo de condiciones de ortogonalidad. De hecho, algunos modelos teóricos generan condiciones en términos de esperanzas condicionales en un determinado conjunto de información:

$$E\left[f(\theta_0, w_t) \mid \Omega_{t-1}\right] = E_{t-1}\left[f(\theta_0, w)\right] = 0$$

donde Ω_{t-1} denota el conjunto de información disponible al agente económico cuando toma sus decisiones en el período $t-1$. Pero tal condicion implica que si Z_{t-1} es una variable contenida en Ω_{t-1}, tenemos:

$$E_{t-1}\left[Z_{t-1}.f(\theta_0, w_t)\right] = Z_{t-1}E_{t-1}\left[f(\theta_0, w_t)\right] = 0$$

y entonces:

$$E\left(E_{t-1}\left[Z_{t-1}.f(\theta_0, w_t)\right]\right) = 0$$

que es de la forma: $E\left[h(\theta_0, w_t)\right] = 0$ con $h(\theta_0, w_t) = f(\theta_0, w_t).Z_{t-1}$.

Por ejemplo, consideremos la maximización de la utilidad agregada intertemporal de un agente económico, sujeto a una sucesión de restricciones presupuestarias, del tipo:

$$\underset{\{c_t, b_t\}}{Max}\ E_t\left(\sum_{s=1}^{\infty}\beta^s U(c_{t+s})\right)$$
$$sujeto\ a \quad : \quad c_{t+s} + b_{t+s} = (1 + r_{t+s})b_{t+s-1} + y_{t+s}$$

siendo c_t el nivel de consumo, b_t su nivel de ahorro (ambas en términos, reales), r_t el tipo de interés real, exógeno para el decisor individual, que suponemos conocido en t, e y_t denota una renta exógena, aleatoria, recibida cada período. El consumidor solo puede maximizar la esperanza matematica, basada en la informacion de que dispone cuando toma decisiones en t , ya que desconoce su renta futura y, por tanto, tambien desconoce sus posibilidades de consumo futuras. Una función de utilidad más cóncava hace que el consumidor/inversor prefiera una senda temporal de consumo mas suave.

La resolución de este problema, que requiere utilizar multiplicadores de Lagrange estocásticos, conduce a las condiciones:

$$\frac{U'(c_t)}{\beta E_t U'(c_{t+1})} = 1 + r_t,\ t = 1, 2, 3, ...$$

que pueden escribirse:

$$E_t\left[\beta(1 + r_t)U'(c_{t+1}) - U'(c_t)\right] = 0 \tag{18}$$

denotemos por Y_t el $Tm-$vector que contiene las observaciones sobre las variables w_t en una muestra de tamaño T. El método generalizado de momentos consiste en encontrar los valores numéricos θ_0 de los parámetros que satisfagan en la muestra las condiciones análogas a las condiciones de ortogonalidad (17):

$$g(\theta, Y_t) = \frac{1}{T}\sum_{t=1}^{T} h(\theta, w_t) = 0 \tag{19}$$

Si el vector $w_t = w$ es estacionario y las funciones h son continuas, entonces podemos esperar que se cumpla la ley de los grandes números:

$$g(\theta, Y_T) \underset{T\to\infty}{\longrightarrow} E\left[h\left(\theta, w_t\right)\right]$$

por lo que los valores paramétricos que resuelven aproximadamente el sistema de m ecuaciones (19) sea muy similar al vector de parámetros que resolveria (17) [ver Hansen (1982), Econometrica].

9.1 El estimador GMM

El problema que vamos a resolver es:

$$\underset{\theta}{Min}\ J_T = \underset{\theta}{Min}\left(\parallel g(\theta, Y_t) \parallel\right) = \underset{\theta}{Min}\left(\parallel \frac{1}{T}\sum_{t=1}^{T} h(\theta, w_t) \parallel\right)$$

Nótese que no tomamos la suma de las normas de cada $h(\theta, w_t)$ sino la norma de la suma (o del promedio) de las $h(\theta, w_t)$.

Para definir una norma del vector $h = (h_1, h_2, ..., h_T)$, escogemos una matriz de ponderaciones A_T definida positiva, y consideramos el problema,

$$\underset{\theta}{Min}\ Q(\theta, Y_T) = \underset{\theta}{Min}\ \left[\left(\frac{1}{T}\sum_{t=1}^{T} h(\theta, w_t)\right)' A_T \left(\frac{1}{T}\sum_{t=1}^{T} h(\theta, w_t)\right)\right] \qquad (20)$$

Hay dos razones para considerar la minimización de esta forma cuadrática, en vez de resolver directamente el conjunto de ecuaciones (19) . Una, que las condiciones de ortogonalidad pueden no satisfacerse exactamente en la muestra; otras, que podemos tener mas condiciones de ortogonalidad (ecuaciones en (19) que parámetros, en cuyo caso la forma cuadrática nos permite encontrar el vector de parámetros que con mayor aproximación permite el cumplimiento en la muestra de las condiciones de ortogonalidad. Además, la matriz S_T nos permite ponderar de distinta manera unas condiciones de ortogonalidad de otras.

La distribución de probabilidad asintótica del estimador resultante depende de la elección de la matriz A. Hansen y Singleton (1982) probaron que la elección óptima de matriz de ponderaciones A, en el sentido de minimizar la matriz de covarianzas del estimador MGM resultante se consigue utilizando una aproximación muestral a la inversa de la varianza asintótica de la media muestral de $h\left(\theta, w_t\right)$:

$$S = \lim_{T\longrightarrow\infty} T.E\left(\left[g(\theta_0, Y_T)\right]\left[g(\theta_0, Y_T)\right]'\right)$$

por lo que el estimador generalizado de momentos eficiente se obtiene minimizando:

$$\min_{\theta} Q(\theta, Y_T) = g(\theta, Y_T)' S\ g(\theta, Y_T)$$

solo que desconocemos la matriz S. Si las funciones $h(\theta, w_t)$ carecen de autocorrelación, un estimador consistente de S sería la matriz:

$$S_T^* = \frac{1}{T}\sum_{t=1}^{T}\left[h(\theta_0, w_t)\right]\left[h(\theta_0, w_t)\right]'$$

lo que tampoco podemos calcular, pues desconocemos los verdaderos valores de los parámetros. Bajo ciertas condiciones, si $\hat{\theta}_T$ es un estimador consistente de θ, y si las funciones $h(\theta, w_t)$ carecen de autocorrelación, se tiene:

$$\hat{S}_T = \frac{1}{T}\sum_{t=1}^{T}\left[h(\hat{\theta}_T, w_t)\right]\left[h(\hat{\theta}_T, w_t)\right]' \longrightarrow S \tag{21}$$

Por lo tanto el procedimiento se pone en práctica del siguiente modo:

1. Se obtiene un estimador inicial $\hat{\theta}_T^{(0)}$ minimizando (20) para una matriz de pesos arbitraria, que habtuamente es I_r
2. Este estimador de θ_0, que resulta ser consistente, se utiliza en (21) para tener un estimador inicial $\hat{S}_T^{(0)}$ de la matriz de ponderaciones
3. Se minimiza (20) con $A_T = \left[\hat{S}_T^{(0)}\right]^{-1}$ para obtener un nuevo estimador GMM, $\hat{\theta}_T^{(1)}$,
4. el procedimiento se itera hasta que se cumplan los criterios de convergencia que se impongan. En todo caso, las propiedades teóricas del estimador obtenido en la primera etapa son idénticas a las del estimador resultante tras alcanzar convergencia.

Si las funciones $h(\theta, w_t)$ presentan autocorrelación, entonces un estimador consistente de la matriz S, propuesto por Newey y West (1987) es:

$$\hat{S}_T = \hat{\Gamma}_{0,T} + \sum_{i=1}^{i=L}\left(1 - \frac{i}{L}\right)\left(\hat{\Gamma}_i + \hat{\Gamma}_i'\right), \; \textit{donde } \Gamma_i = E\left[h(\theta, w_t)h(\theta, w_{t-i})\right]$$

matrices que estimamos mediante:

$$\Gamma_j = \frac{1}{T}\sum_{t=j+1}^{T} h(\hat{\theta}, w_t)h(\hat{\theta}, w_{t-j})'$$

donde L debe escogerse igual al orden de la autocorrelación que se estima para el vector h_t. Finalmente, en (20) tomamos:

$$A_T = \hat{S}_T^{-1}$$

9.2 Distribución asintótica del estimador GMM

El estimador que minimiza la forma cuadrática anterior se distribuye, asintóticamente,

$$\sqrt{T}(\hat{\theta}_T - \theta) \to N(0, \Sigma)$$

siendo $\Sigma = \left(DS^{-1}D'\right)^{-1}$, donde S es la matriz de varianzas y covarianzas de las condiciones de ortogonalidad antes definida, que se estima mediante (**??**) y D es el límite en probabilidad del Jacobiano de las condiciones de ortogonalidad respecto a los parámetros del modelo,

$$D = p\lim E\left(\frac{\partial g(\theta, Y_T)}{\partial \theta}\right)_{\theta=\theta_0}$$

Por tanto, podemos aproximar:

$$\hat{\theta}_T \to N\left(\theta_0, \frac{1}{T}\hat{\Sigma}_T\right)$$

siendo la matriz $\hat{\Sigma}_T$ una aproximación a Σ, definida mediante $\hat{\Sigma}_T = (\hat{D}_T \hat{S}_T^{-1} \hat{D}_T)^{-1}$, con:

$$D_T = \frac{1}{T}\sum_{t=1}^{T}\left(\frac{\partial h(\theta, Y_t)}{\partial \theta}\right)_{\theta=\hat{\theta}_T}$$

Puesto que $g(\theta, Y_T)$ es la media muestral de un proceso cuya esperanza matemática es cero, cabe esperar que bajo determinadas condiciones (entre otras: w_t estacionarias, h continuas) $g(\theta_0, Y_T)$ satisfaga el teorema central del limite:

$$\sqrt{T}g(\theta_0, Y_T) \longrightarrow N(0, S)$$

lo cual implicaría:

$$\left[\sqrt{T}g(\theta_0, Y_T)\right]' S^{-1}\left[\sqrt{T}g(\theta_0, Y_T)\right] \longrightarrow \chi^2_r$$

Por otra parte, si $\hat{\theta}_T$ es un óptimo interior el problema de optimización (20) con $A_T = \hat{S}_T^{-1}$, $\hat{\theta}_T$ sería una solución del sistema de ecuaciones:

$$\left[\left(\frac{\partial g(\theta, Y_T)}{\partial \theta}\right)_{\theta=\hat{\theta}_T}\right]' \hat{S}_T^{-1}\left[g(\theta, Y_T)\right] = 0_k$$

(la primera matriz es kxr, la segunda es rxr, y el tercer factor es $rx1$), por lo que tendremos:

$$\left(\frac{\partial g(\hat{\theta}_T, Y_T)}{\partial \theta}\right)' \hat{S}_T^{-1}\left[g(\hat{\theta}_T, Y_T)\right] = 0_k$$

Esto significa que hay k combinaciones lineales del vector $g(\hat{\theta}_T, Y_T)$ que son exactamente iguales a cero, por lo que el vector $g(\hat{\theta}_T, Y_T)$ solo contiene $r-k$ variables no degeneradas y el contraste de cumplimiento de las cndiciones de ortogonalidad, que se conoce como contraste de sobreidentificación, se basa en la distribución:

$$T\left[g(\hat{\theta}_T, Y_T)\right]' \hat{S}_T^{-1}\left[g(\hat{\theta}_T, Y_T)\right] \longrightarrow \chi^2_{r-k}$$

9.3 Estimación por método generalizado de los momentos

9.3.1 El modelo CCAPM

Si retomamos el problema de maximización de la utilidad intertemporal agregada en el tiempo, tendremos una condicion como (18) para cada activo, es decir:

$$\frac{U'(c_t)}{\beta E_t U'(c_{t+1})} = 1 + r_t,\ i = 1, 2, ..., m,\ t = 1, 2, 3, ...$$

Si suponemos que el inversor tiene una funcion de utilidad: $U(c_t) = \frac{c_t^{1-\gamma}}{1-\gamma}$, tendremos las condiciones de optimalidad:

$$E_t\left[\beta\left(1 + r_{it}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma}\right] = 1,\ i = 1, 2, ..., m$$

que significa que, en equilibrio (si todos los agentes son idénticos, todos optimizan, y los mercados se vacían (la oferta de bienes y de activos es igual a su demanda), entonces, ponderados por la relación marginal de sustitución, todos los activos ofrecen la misma rentabilidad esperada.

Para cualquier variable Z_t conocida en el momento de tomar las decisioens del período t, tendremos:

$$E_t\left[1 - \beta\left(1 + r_{it}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} Z_t\right] = 0,\ i = 1, 2, ..., m$$

Si denotamos $\theta = (\gamma, \beta)$, y por w_t al vector de variables: $w_t = (r_{1t}, r_{2t}, ..., r_{mt}, c_{t+1}; c_t, Z_t)$, tenemos r condiciones de ortogonalidad:

$$h(\theta, w_t) = \begin{pmatrix} 1 - \beta\left(1 + r_{1t}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} Z_t \\ 1 - \beta\left(1 + r_{2t}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} Z_t \\ ... \\ 1 - \beta\left(1 + r_{mt}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} Z_t \end{pmatrix}$$

al que aplicaríamos el método de estimación del modo antes descrito. En esta aplicación, la propia teoría sugiere que $1-\beta\left(1+r_{it}\right)\left(\frac{c_{t+1}}{c_t}\right)^{-\gamma} Z_t$ está incorrelacionado con variables en Ω_t, por lo que podemos utillizar la expresión más sencilla del estimador $\hat{S}_T$. En la estimación del modelo, Hansen y Singleton (1982) utilizan como instrumentos: $Z_t = (1, \frac{c_t}{c_{t-1}}, \frac{c_{t-1}}{c_{t-2}}, ..., \frac{c_{t-l+1}}{c_{t-l}}, r_{1t}, r_{1t-1}, ..., r_{1,t-l+1}, r_{2t}, r_{2t-1}, ..., r_{2,t-l+1})$, sieno r_{1t} la rentabilidad, ajustada por inflación, de un US$ invertido en cada acción que cotiza en NYSE, mientars que r_{2t} es la rentablidad, ajustada por inflación, de la cartera completa del NYSE, ponderada por valor (capitalización) ajustada por inflación.

9.3.2 El estimador MCO en una regresión lineal

Como es sabido, el estimador de minimos cuadrados del modelo de regresión lineal: $y_t = x_t'\beta + u_t$ es el conjunto de valores numéricos $\hat{\beta}$ de los coeficientes β tales que los residuos generados $\hat{u}_t = y_t - x_t'\hat{\beta}$ satisfagan:

$$E(x_t'\hat{u}_t) = E\left[x_t'(y_t - x_t'\hat{\beta})\right] = 0$$

Por lo tanto, en términos de la notación GMM, $h(\theta, w_t) = x_t'(y_t - x_t'\hat{\beta})$ por lo que tenemos:

$$g(\hat{\theta}, Y_T) = \frac{1}{T}\sum_{t=1}^{T} x_t'(y_t - x_t'\hat{\beta}) = 0$$

un sistema de ecuaciones que nos prporciona el estimador MCO. La matriz Jacobiana es en este caso:

$$\begin{aligned} \hat{D}_T' &= \left(\frac{\partial g(\theta, Y_T)}{\partial \theta'}\right)_{\theta=\hat{\theta}_T} = \frac{1}{T}\sum_{t=1}^{T} x_t x_t' \\ S &= \lim \frac{1}{T}\sum_{t=1}^{T}\sum_{l=-\infty}^{l=\infty} E\left(u_t u_{t-l} x_t x_{t-l}'\right) \end{aligned}$$

si u_t carece de autocorrelación y de heterocedasticidad, con $Var(u_t) = \sigma^2 \forall t$, entonces:

$$\begin{aligned} E\left(u_t u_{t-l} x_t x_{t-l}'\right) &= \sigma^2 E(x_t x_t') \; si \; l = 0 \\ &= 0 \; si \; l \neq 0 \end{aligned}$$

de modo que:

$$\hat{S}_T = \hat{\sigma}_T^2 \frac{1}{T}\sum_{t=1}^{T} x_t x_t', \; con \; \hat{\sigma}_T^2 = T^{-1}\sum_{t=1}^{T} \hat{u}_t^2$$

es decir:

$$\frac{1}{T}\hat{\Sigma}_T = \frac{1}{T}\left[\left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)\left(\hat{\sigma}_T^2 \frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)^{-1}\left(\frac{1}{T}\sum_{t=1}^{T} x_t x_t'\right)\right]^{-1} = \hat{\sigma}_T^2\left(\sum_{t=1}^{T} x_t x_t'\right)^{-1}$$

9.3.3 Proceso de difusión de tipos de interés

Si consideramos nuevamente la discretización aproximada del modelo de tipos de interés,

$$r_t - r_{t-1} = \alpha + \beta r_{t-1} + \eta_t$$

con,

$$\begin{aligned} E_{t-1}\eta_t &= 0 \\ E_{t-1}\eta_t^2 &= \sigma^2 r_{t-1}^{2\gamma} \end{aligned} \tag{22}$$

La condición sobre el momento de segundo orden puede escribirse,

$$E_{t-1}\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) = 0 \tag{23}$$

por lo que tenemos en el modelo que dos funciones del término de error tienen esperanza condicional igual a cero.

Como vimos antes, utilizamos en la estimación condiciones algo más débiles, como son,

$$\begin{aligned} E\left[z_{t-1}\eta_t\right] &= 0 \\ E\left[z_{t-1}\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right)\right] &= 0 \end{aligned} \tag{24}$$

donde z_{t-1} es cualquier variable contenida en el conjunto de información disponible en $t-1$. Las variables z_{t-1} utilizadas en la estimación del modelo reciben el nombre de instrumentos, en línea con la denominación habitual en econometría, puesto que (24) muestra que son variables incorrelacionadas con el término de error del modelo.

Para cada conjunto de instrumentos tenemos un estimador MGM. Además, hemos de tener presente que este estimador utiliza un conjunto de condiciones más débiles que las que realmente tenemos disponibles. Si escribimos las condiciones anteriores como,

$$\begin{aligned} Eh_{1t}\left(z_{t-1}, r_t, r_{t-1}; \theta\right) &= 0,\ h_{1t} \equiv z_{t-1}\eta_t \\ Eh_{2t}\left(z_{t-1}, r_t, r_{t-1}; \theta\right) &= 0,\ h_{2t} \equiv z_{t-1}\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) \end{aligned}$$

formamos un vector de funciones de dimensión $2k$ (en general, qk), siendo k el número de variables instrumentales seleccionadas, y buscar en el espacio paramétrico el valor numérico del vector θ que minimiza una norma (forma cuadrática con matriz definida positiva) de dicho vector de funciones, evaluadas en la muestra disponible,

$$\underset{\theta}{Min}\ J_T = \underset{\theta}{Min}\left(\parallel \frac{1}{T}\sum_{t=1}^{T} h_t \parallel \right) \tag{25}$$

donde $h_t' = (h_{1t}^1, h_{2t}^1, h_{1t}^2, h_{2t}^2, ..., h_{1t}^k, h_{2t}^k,)$, es un vector fila de dimensión $2k$, y la diferencia entre $h_{1t}^i, h_{2t}^j, i, j = 1, 2, ..., k$, estriba en que utilizamos en su cálculo instrumentos distintos z_{t-1}^i, z_{t-1}^j :

$$\begin{aligned} h_1^1 &= \frac{1}{T}\sum z_{t-1}^1 \eta_t,\ h_1^2 = \frac{1}{T}\sum z_{t-1}^2 \eta_t, ...,\ h_1^k = \frac{1}{T}\sum z_{t-1}^k \eta_t \\ h_2^1 &= \frac{1}{T}\sum z_{t-1}^1 \left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right),\ h_2^2 = \frac{1}{T}\sum z_{t-1}^2 \left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right), ...,\ h_2^k = \frac{1}{T}\sum z_{t-1}^k \left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) \end{aligned}$$

donde las variables z_{t-1}^i, z_{t-1}^j pueden ser: 1, r_{t-1}, r_{t-2}, etc.. Como puede apreciarse, el número de condiciones de ortogonalidad muestrales de que disponemos para la estimación es igual al producto del número de condiciones de ortogonalidad poblacionales (funciones h) multiplicado por el número de instrumentos (z) que utilicemos en cada una de ellas, que supondremos el mismo.

En este caso, tendremos:

$$D_T = \frac{1}{T}\sum_t \begin{pmatrix} X_{t-1}\frac{\partial \eta_t}{\partial \theta} \\ X_{t-1}\frac{\partial\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right)}{\partial \theta} \end{pmatrix}$$

En consecuencia, puede apreciarse que la expresión analítica para la obtención del estimador MGM puede escribirse, tomando derivadas en (25) ,

$$\left(\frac{1}{T}\sum_{t=1}^{T} \begin{matrix} X_{t-1}\frac{\partial \eta_t}{\partial \theta} \\ X_{t-1}\frac{\partial\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right)}{\partial \theta} \end{matrix}\right)' A \left(\frac{1}{T}\sum_{t=1}^{T} \begin{matrix} X_{t-1}\eta_t \\ X_{t-1}\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) \end{matrix}\right) = 0$$

donde los órdenes de los factores son $qxnk, nkxnk$ y $nkx1$, siendo n el número de condiciones de ortogonalidad poblacionales, 2 en nuestro caso, y k el número de instrumentos. Estas ecuaciones serán lineales si el gradiente $\begin{pmatrix} \frac{\partial \eta_t}{\partial \theta} \\ \frac{\partial\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right)}{\partial \theta} \end{pmatrix}$ lo es, como ocurre en un modelo lineal y sin heterocedasticidad.

Para iniciar el proceso iterativo de estimación, en el que la matriz A_T se va actualizando en cada etapa, se comienza tomando $A_T = I_{nk}$, para obtener en la primera etapa el estimador que minimiza:

$$\left[\frac{1}{T}\sum_{t=1}^{T} \begin{pmatrix} X_{t-1}\eta_t & X_{t-1}\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) \end{pmatrix}\right]\left(\frac{1}{T}\sum_{t=1}^{T} \begin{matrix} X_{t-1}\eta_t \\ X_{t-1}\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) \end{matrix}\right).$$

A partir de las estimaciones obtenidas, se calculan las matrices arriba indicadas y se itera el procedimiento.

Como el número de condiciones de ortogonalidad utilizado en la estimación debe ser mayor que el número de parámetros a estimar, existe un número de grados de libertad, y podemos contrastar la medida en que las condiciones de ortogonalidad no utilizadas para obtener las estimaciones de los parámetros, se satisfacen. Para ello, conviene saber que el valor mínimo alcanzado por la forma cuadrática (25) , multiplicado por el tamaño de la muestra, T, se distribuye como una χ^2_{gdl}, siendo gdl el número de grados de libertad la diferencia entre el número de condiciones de ortogonalidad utilizadas, y el número de parámetros estimados.

Si tomamos como instrumentos una constante y r_{t-1}, tenemos las condiciones,

$$\begin{aligned}
E\eta_t &= E\left(r_t - r_{t-1} - \alpha - \beta r_{t-1}\right) = 0 \\
E\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) &= E\left[\left(r_t - r_{t-1} - \alpha - \beta r_{t-1}\right)^2 - \sigma^2 r_{t-1}^{2\gamma}\right] = 0 \\
E\left(r_{t-1}\eta_t\right) &= E\left[\left(r_t - r_{t-1} - \alpha - \beta r_{t-1}\right) r_{t-1}\right] = 0 \\
E\left[\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) r_{t-1}\right] &= E\left(\left[\left(r_t - r_{t-1} - \alpha - \beta r_{t-1}\right)^2 - \sigma^2 r_{t-1}^{2\gamma}\right] r_{t-1}\right) = 0
\end{aligned}$$

que en la muestra se corresponden con:

$$\begin{aligned}
\frac{1}{T}\sum_{t=1}^{T} h_{11,t} &= \frac{1}{T}\sum_{t=1}^{T}\left[\left(r_t - r_{t-1}\right) - \beta r_{t-1} - \alpha\right] = 0 \\
\frac{1}{T}\sum_{t=1}^{T} h_{12,t} &= \frac{1}{T}\sum_{t=1}^{T}\left[\left(\left(r_t - r_{t-1}\right) - \beta r_{t-1} - \alpha\right)^2 - \sigma^2 r_{t-1}^{2\gamma}\right] = 0 \\
\frac{1}{T}\sum_{t=1}^{T} h_{21,t} &= \frac{1}{T}\left(\sum_{t=1}^{T}\left[\left(r_t - r_{t-1}\right) r_{t-1} - \beta r_{t-1}^2 - \alpha r_{t-1}\right]\right) = 0 \\
\frac{1}{T}\sum_{t=1}^{T} h_{22,t} &= \frac{1}{T}\sum_{t=1}^{T}\left(r_{t-1}\left[\left(r_t - r_{t-1}\right) - \beta r_{t-1} - \alpha^2 - \sigma^2 r_{t-1}^{2\gamma}\right]\right) = 0
\end{aligned}$$

que son las 4 condiciones de momentos $(h_{11,t}, h_{12,t}, h_{21,t}h_{22,t})$ que vamos a utilizar en la estimación. Son cuatro ecuaciones que dependen de momentos muestrales de distintas funciones de los tipos de interés, todas ellas calculables a partir de la información muestral, y de los cuatro parámetros desconocidos. En este caso, tenemos un sistema exactamente identificado. El problema es que, como fácilmente se aprecia, el sistema de ecuaciones no puede resolverse analíticamente, fundamentalmente porque, salvo en casos muy simples, es un sistema de ecuaciones no lineales en las incógnitas, que son los parámetros del modelo.

En este caso, si tomamos $L = 0$, la matriz D_T tiene por columnas las derivadas parciales de cada condicion de ortogonalidad con respecto a los parámetros: $(\alpha, \beta, \gamma, \sigma)$:

$$D_T = \frac{1}{T-1}\sum_{t=2}^{T}\begin{pmatrix} -1 & -2\eta_t & -r_{t-1} & -2\eta_t r_{t-1} \\ -r_{t-1} & -2\eta_t r_{t-1} & -r_{t-1}^2 & -2\eta_t r_{t-1}^2 \\ 0 & -2\sigma^2 r_{t-1}^{\gamma} & 0 & -2\sigma^2 r_{t-1}^{2\gamma+1}\ln r_{t-1} \\ 0 & -2\sigma r_{t-1}^{2\gamma} & 0 & -2\sigma r_{t-1}^{2\gamma+1} \end{pmatrix} =$$

tomando esperanzas :

$$= \frac{1}{T-1}\sum_{t=2}^{T}\begin{pmatrix} -1 & -2\eta_t & -r_{t-1} & -2\eta_t r_{t-1} \\ -r_{t-1} & -2\eta_t r_{t-1} & -r_{t-1}^2 & -2\eta_t r_{t-1}^2 \\ 0 & -2\sigma^2 r_{t-1}^{\gamma} & 0 & -2\sigma^2 r_{t-1}^{2\gamma+1}\ln r_{t-1} \\ 0 & -2\sigma r_{t-1}^{2\gamma} & 0 & -2\sigma r_{t-1}^{2\gamma+1} \end{pmatrix}$$

mientras que A_T tiene una estructura:

$$A_T = \left[\frac{1}{T-1}\sum_{t=2}^{T}\begin{pmatrix} h_{11,t} \\ h_{12,t} \\ h_{21,t} \\ h_{22,t} \end{pmatrix}\begin{pmatrix} h_{11,t} & h_{12,t} & h_{21,t} & h_{22,t} \end{pmatrix}\right]^{-1} =$$

$$= \left[\frac{1}{T-1}\sum_{t=2}^{T}\begin{pmatrix} h_{11,t}^2 & h_{11,t}h_{12,t} & h_{11,t}h_{21,t} & h_{11,t}h_{22,t} \\ h_{12,t}h_{11,t} & h_{12,t}^2 & h_{12,t}h_{21,t} & h_{12,t}h_{22,t} \\ h_{21,t}h_{11,t} & h_{21,t}h_{12,t} & h_{21,t}^2 & h_{21,t}h_{22,t} \\ h_{22,t}h_{11,t} & h_{22,t}h_{12,t} & h_{22,t}h_{21,t} & h_{22,t}^2 \end{pmatrix}\right]^{-1}$$

Habitualmente, en el cálculo del estimador del método generalizado de momentos se utilizan más condiciones de ortogonalidad que parámetros se pretenden estimar, lo que permite contrastar la sobreidentificación del modelo. Por ejemplo, en la estimación de la discretizacion anterior podriamos utilizar asimismo r_{t-2} como instrumento, añadiendo entonces dos condiciones de ortogonalidad:

$$\begin{aligned} E\left(r_{t-2}\eta_t\right) &= E\left[\left(r_t - r_{t-1} - \alpha - \beta r_{t-1}\right) r_{t-2}\right] = 0 \\ E\left[\left(\eta_t^2 - \sigma^2 r_{t-1}^{2\gamma}\right) r_{t-2}\right] &= E\left(\left[\left(r_t - r_{t-1} - \alpha - \beta r_{t-1}\right)^2 - \sigma^2 r_{t-1}^{2\gamma}\right] r_{t-2}\right) = 0 \end{aligned}$$

y sus correspondientes momentos análogos muestrales:

$$\begin{aligned} \frac{1}{T}\sum_{t=1}^{T} h_{31,t} &= \frac{1}{T}\sum_{t=1}^{T}\left[\left(r_t - r_{t-1}\right) r_{t-2} - \beta r_{t-1} r_{t-2} - \alpha r_{t-2}\right] = 0 \\ \frac{1}{T}\sum_{t=1}^{T} h_{32,t} &= \frac{1}{T}\sum_{t=1}^{T}\left(r_{t-2}\left[\left(\left(r_t - r_{t-1}\right) - \beta r_{t-1} - \alpha\right)^2 - \sigma^2 r_{t-1}^{2\gamma}\right]\right) = 0 \end{aligned}$$

se añadirían a las anteriores, configurando un vector de 6 condiciones de ortogonalidad para estimar los 4 parámetros: $(\alpha, \beta, \gamma, \sigma)$. La matriz D_T tendría dimensión $4x6$, pues en cada columna tendriamos las derivadas parciales de cada función h con respecto a los 4 parámetros estructurales. La matriz A_T tendría dimensión $6x6$, pues se forma a partir de los productos cruzados de las funciones h.

9.3.4 Ejercicio

1. Obtener la estimaciones, por el Método Generalizado de Momentos, de los parámetros $\alpha, \rho, \sigma_\varepsilon^2$ del modelo de regresión constante con errores AR(1).

Solución: Utilizaríamos el hecho de que, bajo el supuesto de que el modelo esté correctamente especificado, se tienen las propiedades: $Ey_t = \alpha, Var(y_t) = \sigma_u^2, \rho = \frac{Cov(y_t, y_{t-1})}{Var(y_t)}, \sigma_\varepsilon^2 = \sigma_u^2 \left(1 - \rho^2\right)$, por lo que, sustituyendo momentos poblacionales por muestrales en las igualdades anteriores, tendríamos,

$$\hat{\alpha} = \frac{1}{T} \sum_1^T y_t;\ \hat{\rho} = \frac{\sum_1^T (y_t - \bar{y})(y_{t-1} - \bar{y})}{\sum_1^T (y_t - \bar{y})^2}$$

$$\hat{\sigma}_u^2 = \frac{1}{T} \sum_1^T (y_t - \bar{y})^2;\ \hat{\sigma}_\varepsilon^2 = \hat{\sigma}_u^2 \left(1 - \hat{\rho}^2\right) = \left(\frac{1}{T} \sum_1^T (y_t - \bar{y})^2\right) \left(1 - \left[\frac{\sum_1^T (y_t - \bar{y})(y_{t-1} - \bar{y})}{\sum_1^T (y_t - \bar{y})^2}\right]^2\right)$$

La estimación de ρ coincide con la estimación de mínimos cuadrados que hemos propuesto más arriba. No así la de σ_ε^2 ni la de σ_u^2. Tampoco será exactamente coincidente la estimación del término independiente α si bien, el argumeo efectuado al presentar el estimador de Máxima Verosimilitud garantiza que la diferencia entre los valores numéricos de ambos estimadores no será muy elevada en muestras grandes.