Estimación de modelos multiecuacionales mediante el entorno de programación R

1. Resumen

En el presente documento vamos a abordar la estimación de modelos de ecuaciones simultáneas mediante el entorno de programación R. Con tal objetivo usaremos el paquete SEM, disponible en el *Comprehensive R Achive Network* (http://cran.r-project.org/). Dicho paquete proporciona un soporte general para los modelos de ecuaciones estructurales con variables latentes (ecuaciones simultáneas) mediante el ajuste por máxima verosimilitud (asumiendo normalidad multivariante), y la estimación de una sola ecuación mínimos cuadrdos en dos etapas (MC2E). También se abordará el método de mínimos cuadrados ordinarios (MCO), ya que en la práctica la econometría se centra en la aplicación directa de MCO y MC2E. Esto se debe a que los métodos más perfectos teóricamente son más complejos y tienen un mayor el coste computacional.

2. Función *tsls*

Para la estimación de cada ecuación usaremos el método de MC2E, ya que es adecuado tanto para ecuaciones exactamente identificadas como sobreidentificadas. Dicho método se ejecuta mediante la función *tsls* de la librería sem:

tsls(formula, instruments, ...)

Las principales argumentos de dicha función son:

- Fórmula: especificación de la ecuación estructural a estimar.
- Instrumentos: especificación de las variables instrumentales.
- La función devuelve un objeto de la clase *tsls*, sobre el que se pueden ejecutar distintas órdenes: *print*, *summary*, *fitted*, *residuals* y *anova*.

Para más información ejecutar *help(tsls)* en la ventana de comandos de R.

3. Función *lm*

Aquellas ecuaciones estructurales no identificadas se pueden estimar mediante el método de MCO. Dicho método se ejecuta en R mediante la función *lm* y sus principales argumentos coinciden con los de la función *tsls* (para más información ejecutar *help(lm)*).

4. Ejemplo

Consideremos el modelo de ecuaciones simultáneas dado por las dos ecuaciones siguientes:

$$\begin{aligned} Y_{t1} &= \alpha_0 + \alpha_1 Y_{t2} + \alpha_2 X_{t1} + \alpha_3 X_{t2} + \alpha_4 X_{t3} + u_{t1}, \\ Y_{t2} &= \beta_0 + \beta_1 X_{t3} + \beta_2 X_{t4} + u_{t2}, \end{aligned}$$

donde:

- Y1 es el consumo familiar mensual (medido en miles de euros).
- Y2 es la renta familiar mensual (medida en miles de euros).
- X1 es una variable ficticia que toma el valor 1 si la familia en cuestión tiene algún tipo de deuda (hipoteca, coche, etc.) y 0 en otro caso.
- X2 es el número de hijos de cada familia.
- X3 es el número de individuos de la familian que trabajan.
- X4 es el nivel de estudios de los trabajadores de cada familia.

Puesto que la renta familiar aparece al mismo tiempo como explicada (segundo ecuación) y explicativa (primera ecuación), es evidente que nos encontramos ante un modelo de ecuaciones simultáneas.

Teniendo en cuenta que en dicho modelo podemos distinguir dos variables endógenas (consumo y renta) y cinco exógenas (constante, deuda, hijos, trabajadores y estudios), tras realizar la identificación de las ecuaciones se obtiene que la primera es exactamente identificada y que la segunda es sobreidentificada. Por tanto, ambas ecuaciones pueden ser estimadas mediante MC2E.

Para abordar la estimación de dichas ecuaciones disponemos de la información muestral obtenida a partir de 16 familias recogida en la tabla 1.

Tabla 1: Información muestral de 16 familias. Fuente: elaboración propia.

Consumo	Renta	Deuda	Hijos	Trabajadores	Estudios
1.3	1.7	1	1	1	1
2.3	2.6	1	1	2	2
2.5	3.2	1	2	2	3
1.2	2	1	0	1	1
2	3	1	3	2	2
3.2	3.8	1	2	2	3
3	5	0	1	2	1
1.8	3	1	1	1	1
2	3	0	3	1	2
2.7	4.3	1	2	2	1
1.8	2	1	1	1	1
1.5	3.3	1	0	1	3
1.5	3.1	0	0	2	2
1.1	2.8	0	0	1	3
1	1.3	1	1	1	1
1	2.1	0	0	1	1

Dichas observaciones han sido guardados en un archivo *.csv* delimitado por *;* (adviértase que el separador decimal ha de ser el punto). Para cargar los datos en R hay que especificar que el directorio de trabajo es aquel en el que se encuentra dicho archivo (esto se hace mediante *Cambiar dir...* del menú *Archivo*) y ejecutar en la ventana de comandos de R las órdenes:

```
> datos <- read.csv("consumo_renta.csv", sep=";")
> datos <- as.matrix(datos)
> consumo = datos[,1]
> renta = datos[,2]
> deuda = datos[,3]
> hijos = datos[,4]
> edad = datos[,5]
> trabajadores = datos[,6]
> estudios = datos[,7]
```

El siguiente paso es instalar el paquete sem (lo que se hace mediante *Instalar paquete(s)...* del menú *Paquetes*) y cargarlo mediante el comando:

```
> library(sem)
```

Para estimar la primera ecuación simplemente hay que intoducir la siguiente orden:

```
> eqn.1 <- tsls(consumo ~ renta + deuda + hijos + trabajadores, instruments = ~ deuda + hijos + trabajadores)
```

donde hemos especificado la ecuación mediante la fórmula *consumo ~ renta + deuda + hijos + trabajadores* y las variables instrumentales.

Mediante el comando:

```
> summary(eqn.1)
```

se obtiene la información:

```
 2SLS Estimates

Model Formula: consumo ~ renta + deuda + hijos + trabajadores

Instruments: ~deuda + hijos + trabajadores

Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-2.590000 -1.320000  0.000376  0.445000  2.230000  4.350000

             Estimate Std. Error    t value Pr(>|t|)
(Intercept)    3.9804   20819757  1.912e-07        1
renta         -2.6669   13975776 -1.908e-07        1
deuda         -1.1497    7868223 -1.461e-07        1
hijos          0.4452    1323320  3.364e-07        1
trabajadores   3.7800   16317192  2.317e-07        1

Residual standard error: 2.5651 on 11 degrees of freedom
```

Podemos observar que la información resultante es la fórmula de la ecuación, las variables instrumentales, información sobre los residuos, la estimación de cada variable, junto a su correspondiente desviación típica y p-valor asociado (adviértase que unas desviaciones típicas estimadas de los coeficientes estimados tan altos nos hacen pensar que pueda haber un problema de multicolinealidad en la ecuación, por lo que las estimaciones obtenidas quedan en cuarentena).

También se pueden ejecutar los comandos print, fitted o residual sobre el objeto eqn.1 obteniéndose, respectivamente, las estimaciones de los coeficientes, los valores estimados de la variable endógena de la ecuación y los residuos del modelo ajustado.

Para la segunda ecuación, sin más que ejecutar:

```
> eqn.2 <- tsls(renta ~ trabajadores + estudios, instruments = ~ trabajadores + estudios)
> summary(eqn.2)
```

se obtienen los siguientes resultados:

```
 2SLS Estimates

Model Formula: renta ~ trabajadores + estudios

Instruments: ~trabajadores + estudios

Residuals:
     Min.   1st Qu.    Median      Mean   3rd Qu.      Max.
-9.95e-01 -5.04e-01 -2.45e-01 -8.33e-17  6.23e-01  1.54e+00

             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.0177     0.6565  1.5502  0.14510
trabajadores   1.1671     0.4079  2.8614  0.01336
estudios       0.1098     0.2440  0.4499  0.66021

Residual standard error: 0.7802 on 13 degrees of freedom
```

En este caso se aprecia cómo la única variable significativa en la renta familiar es el número de trabajadores, de manera que a mayor número de trabajadores mayor renta familiar).

Por tanto, la estimación por MC2E de las dos ecuaciones consideradas será:

$$\begin{aligned}\widehat{Y}_{t1} &= 3.9804 - 2.6669Y_{t2} - 1.1497X_{t1} + 0.4452X_{t2} + 3.78X_{t3},\\ \widehat{Y}_{t2} &= 1.0177 + 1.1671X_{t3} + 0.1098X_{t4}.\end{aligned}$$

Además, como en la segunda ecuación no hay variables endógenas que aparezcan como explicativas, el método de MC2E no es más que MCO. Luego vamos a estimar a continuación la segunda ecuación también por éste segundo método:

```
> eqn.2 <- lm(renta ~ trabajadores + estudios)
> summary(eqn.2)
```

obteniéndose:

```
Call:
lm(formula = renta ~ trabajadores + estudios)

Residuals:
    Min      1Q  Median      3Q     Max
-0.9946 -0.5038 -0.2446  0.6231  1.5384

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)    1.0177     0.6565   1.550   0.1451
trabajadores   1.1671     0.4079   2.861   0.0134 *
estudios       0.1098     0.2440   0.450   0.6602
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7802 on 13 degrees of freedom
Multiple R-squared: 0.429,      Adjusted R-squared: 0.3411
F-statistic: 4.883 on 2 and 13 DF,  p-value: 0.02620
```

Como se observa se obtienen exactamente las mismas estimaciones (como no podía ser de otra forma). En este caso se nos ofrece más información, por ejemplo, tenemos un coeficiente de determinación de 0.3411, que si bien parece bajo, el p-valor asociado al estadístico F (del ANOVA) valida el modelo ya que se rechaza la hipótesis nula de que todos los coeficientes son cero de forma simultánea.

Referencias

J. Fox (2006), "Structural Equation Modeling With the sem Package in R." *Structural Equation Modeling*, 13:465-486 (http://socserv.mcmaster.ca/jfox/Misc/sem/SEM-paper.pdf).