(#365). BÁSICOS DE ECUACIONES ESTRUCTURALES (V): MATRIZ DE DATOS BRUTOS

[MONOTEMA] En esta quinta entrega, vamos a explicar en qué consiste la matriz de datos brutos que debemos emplear como entrada para realizar los análisis, que no es más que la matriz de covarianzas entre todos los observables de la muestra. Esa matriz se suele denominar como *S*.

El profesor Leslie A. Hayduk, explica perfectamente en su libro de 1987 cómo se construye esa matriz (pp. 62-63). Lo que quizá llame la atención a los estudiantes es que esa matriz de entrada no es una matriz de *n* x *m*, es decir *n* casos en filas y *m* variables en columnas, que es el tipo de disposición habitual para realizar un análisis de regresión por mínimos cuadrados, por ejemplo, donde debemos especificar las puntuaciones de cada caso en cada variable.

Aquí no es así. Y no lo es porque, como ya hemos explicado en capítulos anteriores, las relaciones entre los coeficientes de los modelos se pueden obtener a partir de las covarianzas directamente. Obviamente, las covarianzas se construyen desde los datos individuales, pero no necesitamos especificarlos en SEM.

Si partimos de una matriz *X* de *n* filas y *m* columnas (casos x variables), podemos construir la matríz *X'X*, donde *X'* es la transpuesta de *X*. La matriz *X'* es de dimensión *m* x *n*. De este modo, la matriz resultante es una matriz de *m* x *m*, es decir, una matriz cuadrada donde sólo hay relaciones entre las variables observables.

Esas relaciones son relaciones de covarianza cuando se divide esa matriz resultante por *n*, es decir, *S=Cov(X'X)=(X'X)/n*. Eso es así porque recordemos que los datos estaban tomados en

desviaciones sobre la media, y que por tanto la multiplicación de las dos matrices da una suma de cuadrados. La matriz *S* es simétrica, y en la diagonal están las varianzas de los observables.

Ejemplo con Stata

Vamos a realizar una entrada manual de datos en Stata a través de una matriz de 3 casos x 2 variables, muy sencillo por tanto.

```
/*Generamos la matriz, primero metiendo las filas y después las columnas*/
matrix input X = (3,2\1,0\0,1)
/*Le pedimos un listado para asegurarnos que los datos están como queremos*/
matrix list X
/*Calculamos la matriz traspuesta*/
mat Xtraspuesta=X'
/*Le pedimos un listado para asegurarnos que los datos están como queremos*/
matrix list Xtraspuesta
/*Multiplicamos ambas matrices, y nos da una suma de cuadrados*/
mat sumcuad=X'*X
/*Le pedimos un listado para asegurarnos que los datos están como queremos*/
matrix list sumcuad
/*Dividimos la suma de cuadrados por el tamaño de la muestra (3 casos)*/
mat covar=sumcuad/3
/*Y le pedimos un listado para ver la matriz de covarianzas, que es la matriz S de datos brutos*/
matrix list covar
```

De este modo, por muy grande que sea el número de casos, nuestra matriz de datos brutos siempre tendrá el tamaño de *m* x *m* variables observables.

Dos cosas importantes a considerar son: (1) Los datos de entrada están en desviaciones con respecto a la media. Es decir, los vectores de datos el ejemplo con Stata [3,1,0] y [2,0,1] son datos en desviaciones sobre la media. Si no lo están, basta con hacer ese cálculo previo para seguir con el procedimiento indicado; (2) De momento no vamos a considerar ni la posibilidad de incluir en la matriz de entrada los valores medios, ni la ocurrencia de casos perdidos (ambos temas de índole más avanzado).

El efecto del tamaño de muestra

Aunque usemos una matriz de covarianzas, el tamaño de la muestra, es decir, el número de casos, es fundamental. Quizá se pueda pensar que el tamaño muestral no importa si dos matrices de covarianzas son iguales cuando provienen de muestras de tamaño diferente (lo que puede suceder perfectamente).

Pero sí que importa, porque esas covarianzas estarán mejor estimadas si provienen de muestras más grandes. Para comprobar empíricamente esta cuestión, vamos a ejecutar el siguiente código:

```
/* Borrramos lo anterior */
clear
/* Generamos una muestra aleatoria Normal (500 casos) con media 0 y varianza=1*/
drawnorm x1, n(500)
/* Ahora dividimos la muestra en 10 grupos diferentes*/
gen g1=1 in 1/50
gen g2=2 in 51/100
gen g3=3 in 101/150
gen g4=4 in 151/200
gen g5=5 in 201/250
gen g6=6 in 251/300
gen g7=7 in 301/350
gen g8=8 in 351/400
gen g9=9 in 401/450
gen g10=10 in 451/500
/* Generamos la desviación típica para cada uno de los grupos*/
egen SD = sd(x1), by(g1 g2 g3 g4 g5 g6 g7 g8 g9 g10)
/* Convertimos esas desviaciones típicas en varianzas*/
gen varianza=SD*SD
/* Listamos el primer valor de cada grupo, generando las 10 varianzas diferentes*/
list var in 1
list var in 51
list var in 101
list var in 151
list var in 201
list var in 251
list var in 301
list var in 351
list var in 401
list var in 451
/* Obtenemos la varianza de la muestra de 500 casos y así podemos compararla con las obtenidas para las submuestras*/
sum x1, detail
```

Lo que hemos hecho es un simple ejercicio de simulación donde podemos ver que, siempre que corremos el código, la varianza de la muestra de 500 casos es muy cercana a 1. Sin embargo, en los 10 subgrupos esa varianza puede oscilar mucho, pudiendo obtener varianzas muy cercanas pero también muy lejanas a 1.

Por tanto, aunque la varianza de la población sea realmente 1, escoger muestras pequeñas puede hacer que algunas varianzas (y también covarianzas) de la matriz S disten mucho de su valor real, lo que a su vez puede distorsionar la comparación que ha de hacerse con la matriz implicada por el modelo (que explicaremos en capítulos posteriores).

Conclusión

La matriz de entrada en SEM es una matriz cuadrada de covarianzas de las variables observables, donde en la diagonal están las varianzas,y que además es simétrica. Esa matriz se denomina *S*, y está sujeta a la inherente variabilidad muestral.

Tamaños de muestra pequeños pueden distorsionar de forma importante esta matriz, por lo que aunque el número de casos no esté explícitamente expresado en *S*, es fundamental para que *S* sea válida. Es cierto que, en algunas ocasiones SEM puede funcionar bien con muestras relativamente pequeñas, pero probablemente sea arriesgarse demasiado.

Todos los posts relacionados

[(#384). MEJOR UN SÓLO ITEM QUE VARIOS PARA MEDIR ACTITUDES]

[(#380). ÍNDICES APROXIMADOS FLEXIBLES EN ECUACIONES ESTRUCTURALES]

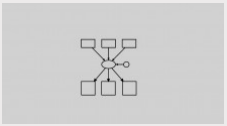
[(#365). BÁSICOS DE ECUACIONES ESTRUCTURALES (V): MATRIZ DE DATOS BRUTOS]

[(#358). BÁSICOS DE ECUACIONES ESTRUCTURALES (IV): REGRESIÓN LINEAL SIMPLE CON FIABILIDADES DIVERSAS]

[(#354). BÁSICOS DE ECUACIONES ESTRUCTURALES (III): COVARIANZA Y CORRELACIÓN ENTRE VARIABLES LATENTES]

[(#353). BÁSICOS DE ECUACIONES ESTRUCTURALES (II): VARIABLES LATENTES Y FIABILIDAD]

[(#350). BÁSICOS DE ECUACIONES ESTRUCTURALES (I): COVARIANZAS Y DESVIACIONES SOBRE LA MEDIA]



[(#331). LOS CONSTRUCTOS FORMATIVOS NO REFLEJAN ESTADOS PSICOLÓGICOS]

[(#287). MEDIAS VERDADES Y OCULTACIÓN; ACEPTACIÓN DEL ENGAÑO DE CORPORACIONES]

[(#163). CRECEN LAS DUDAS SOBRE PLS (PARTIAL LEAST SQUARES)]

[(#128). INDICADORES FORMATIVOS EN ECUACIONES ESTRUCTURALES]

[(#127). AJUSTE EXACTO EN ECUACIONES ESTRUCTURALES]