Tema 1. Tipos de datos

Gustavo A. García

ggarci24@eafit.edu.co

Econometría para la Toma de Decisiones

Maestría en Economía Aplicada

Escuela de Finanzas, Economía y Gobierno

Universidad EAFIT

Link slides en formato html

Link slides en formato PDF

En este tema

- La estructura de los datos económicos
- Datos de sección cruzada (*Cross-sectional data*)
- Datos de series de tiempo (*Time series data*)
- Datos panel o longitudinales (*Panel data*)

La estructura de los datos económicos

- El conjunto de datos en economía son de una variedad de tipos
- En general los métodos econométricos pueden ser aplicados con pocas modificaciones para diferentes tipos de datos
- Sin embargo, las características especiales de algunos conjuntos de datos deben ser tenidos en cuenta para ser explotados
- Entre las más importantes estructuras de datos se tienen:
  - Datos de sección cruzada (*Cross-sectional data*)
  - Datos de series de tiempo (*Time series data*)
  - Datos panel o longitudinales (*Panel data*)

Datos de sección cruzada

- **Definición**: un conjunto de datos de sección cruzada consiste en una muestra de individuos, hogares, firmas, ciudades, estados, países, o variedad de otras unidades tomadas en un dado punto en el tiempo

- Una importante característica de los datos de sección cruzada es que se asume que estos datos han sido obtenidos de una muestra aleatoria de una población objetivo. Por ejemplo: salarios, educación, experiencia y otras características de la población

- Los datos de sección cruzada son ampliamente utilizadas en economía y otras ciencias sociales

- En economía el análisis de estos tipos de datos está cercanamente alineado con el campo de la microeconomía aplicada: economía laboral, organización industrial, economía urbana, economía de la educación, economía de la salud, etc

- Datos sobre individuos, hogares, firmas, y ciudades en un punto en el tiempo son importantes para probar hipótesis en la microeconomía y evaluar políticas económicas

Datos de sección cruzada

```
library(haven) # Leyendo el paquete haven para leer datos de Stata
data1 <- read_dta("http://fmwww.bc.edu/ec-p/data/wooldridge/wage1.dta")
names(data1) # Muestra las variables que contiene la base de datos
```

```
 [1] "wage"     "educ"     "exper"    "tenure"   "nonwhite" "female"
 [7] "married"  "numdep"   "smsa"     "northcen" "south"    "west"
[13] "construc" "ndurman"  "trcommpu" "trade"    "services" "profserv"
[19] "profocc"  "clerocc"  "servocc"  "lwage"    "expersq"  "tenursq"
```

En este link se puede ver la descripción de las variables (es lo que llamamos el diccionario de variables)

```
data1[1:15,c("wage", "educ", "exper", "female", "married")]
```

```
# A tibble: 15 x 5
    wage  educ exper female married
   <dbl> <dbl> <dbl>  <dbl>   <dbl>
 1  3.10    11     2      1       0
 2  3.24    12    22      1       1
 3  3       11     2      0       0
 4  6        8    44      0       1
 5  5.30    12     7      0       1
 6  8.75    16     9      0       1
 7 11.2     18    15      0       0
 8  5       12     5      1       0
 9  3.60    12    26      1       0
10 18.2     17    22      0       1
11  6.25    16     8      1       0
12  8.13    13     3      1       0
13  8.77    12    15      0       1
14  5.5     12    18      0       0
15 22.2     12    31      0       1
```

```
data1[512:526,c("wage", "educ", "exper", "female", "married")]
```

```
# A tibble: 15 x 5
    wage  educ exper female married
   <dbl> <dbl> <dbl>  <dbl>   <dbl>
 1  4.38    13     7      0       1
 2 10       12    15      0       0
 3  4.95     7    25      0       1
 4  9       17     7      1       1
 5  1.43    12    17      1       1
 6  3.08    12     3      0       0
 7  9.33    14    12      0       1
 8  7.5     12    18      0       1
 9  4.75    13    47      0       1
10  5.65    12     2      0       0
11 15       16    14      1       1
12  2.27    10     2      1       0
13  4.67    15    13      0       1
14 11.6     16     5      0       1
15  3.5     14     5      1       0
```

Datos de sección cruzada

- Diferentes variables algunas veces corresponden a diferentes períodos de tiempo en datos de sección cruzada

- Por ejemplo, para determinar los efectos de las políticas sobre el crecimiento de largo plazo, se ha estudiado la relación entre el crecimiento en el PIB per capita real sobre un cierto período (1960-1985) y variables determinadas en parte por políticas en 1960, como consumo del gobierno como porcentaje del PIB y las tasas de adultos con secundaria

TABLE 1.2 A Data Set on Economic Growth Rates and Country Characteristics

obsno	country	gpcrgdp	govcons60	second60
1	Argentina	0.89	9	32
2	Austria	3.32	16	50
3	Belgium	2.56	13	69
4	Bolivia	1.24	18	12
.	.	.	.	.
.	.	.	.	.
.	.	.	.	.
61	Zimbabwe	2.30	17	6

Datos de series de tiempo

- **Definición**: los datos de series de tiempo consiste de observaciones sobre una o varias variables sobre el tiempo
- Ejemplos: los precios, la oferta monetaria, IPC, PIB, tasas de homicidios, etc
- A diferencia del los datos de sección cruzada, el orden cronológico de las observaciones en los datos de series de tiempo conlleva a potencialmente importante información
- Dos características importantes en los datos de series de tiempo son la dependencia temporal y la presencia de tendencias sobre el tiempo. Estos factores deben ser tenidos en cuenta antes aplicar técnicas econométricas estándar
- Otra característica a tener en cuenta en este tipo de datos es la frecuencia de los datos: datos diarios, semanales, mensuales, trimestrales o anuales. Este tipo de datos presentan fuertes patrones estacionales, lo cual debe ser tenido en cuenta en los modelos de regresión

Datos de series de tiempo

Datos anuales de Puerto Rico entre 1950 y 1987 sobre la tasa de empleo, salario mínimo y otras variables usadas en Castillo-Freeman y Freeman (1992) para estudiar los efectos del salario mínimo sobre el empleo

```
data2 <- read_dta("http://fmwww.bc.edu/ec-p/data/wooldridge/prminwge.dta")
names(data2)
```

```
 [1] "year"     "avgmin"   "avgwage"  "kaitz"    "avgcov"   "covt"
 [7] "mfgwage"  "prdef"    "prepop"   "prepopf"  "prgnp"    "prunemp"
[13] "usgnp"    "t"        "post74"   "lprunemp" "lprgnp"   "lusgnp"
[19] "lkaitz"   "lprun_1"  "lprepop"  "lprep_1"  "mincov"   "lmincov"
[25] "lavgmin"
```

En este link se puede ver la descripción de las variables

```
data2[1:15,c("year","prepop", "usgnp", "mincov")]
```

```
# A tibble: 15 x 4
    year prepop usgnp mincov
   <dbl>  <dbl> <dbl>  <dbl>
 1  1950  0.470 1204.  0.100
 2  1951  0.449 1328.  0.106
 3  1952  0.434 1380   0.121
 4  1953  0.428 1435.  0.150
 5  1954  0.415 1416.  0.138
 6  1955  0.419 1495.  0.159
 7  1956  0.412 1526.  0.182
 8  1957  0.412 1551.  0.174
 9  1958  0.397 1539.  0.184
10  1959  0.394 1629.  0.194
11  1960  0.403 1665.  0.198
12  1961  0.397 1709.  0.187
13  1962  0.385 1799.  0.211
14  1963  0.395 1873.  0.195
15  1964  0.396 1973.  0.217
```

```
data2[24:38,c("year","prepop", "usgnp", "mincov")]
```

```
# A tibble: 15 x 4
    year prepop usgnp mincov
   <dbl>  <dbl> <dbl>  <dbl>
 1  1973  0.421 2744.  0.250
 2  1974  0.405 2729.  0.373
 3  1975  0.368 2695   0.434
 4  1976  0.364 2827.  0.446
 5  1977  0.358 2959.  0.429
 6  1978  0.362 3115.  0.450
 7  1979  0.360 3192.  0.458
 8  1980  0.359 3187.  0.455
 9  1981  0.343 3249.  0.458
10  1982  0.318 3166   0.448
11  1983  0.321 3279.  0.438
12  1984  0.334 3501.  0.436
13  1985  0.331 3608.  0.425
14  1986  0.351 3713.  0.412
15  1987  0.369 3820.  0.400
```

Datos panel o longitudinales

- **Definición**: los datos panel o longitudinales consisten en series de tiempo para unidad de sección cruzada

- Ejemplo: información sobre salarios, educación y empleo para un conjunto de individuos sobre un período de 10 años. O datos de crimen de dos años para 150 ciudades

TABLE 1.5 A Two-Year Panel Data Set on City Crime Statistics

obsno	city	year	murders	population	unem	police
1	1	1986	5	350000	8.7	440
2	1	1990	8	359200	7.2	471
3	2	1986	2	64300	5.4	75
4	2	1990	1	65100	5.5	75
.	.	.	.	.	.	.
.	.	.	.	.	.	.
.	.	.	.	.	.	.
297	149	1986	10	260700	9.6	286
298	149	1990	6	245000	9.8	334
299	150	1986	25	543000	4.3	520
300	150	1990	32	546200	5.2	493

Datos panel o longitudinales

Los datos panel presentan importantes ventajas sobre los datos de sección cruzada:

- la existencia de múltiples observaciones sobre la misma unidad permite controlar por características inobservadas de los individuos, firmas, etc
- permiten estudiar la importancia de los efectos de los rezagos. Esto es de particular importancia teniendo en cuenta que muchas medidas de política puede ser esperadas a tener un impacto después de un tiempo