Spatial Regression Models: Identification strategy using STATA

TATIANE MENEZES – PIMES/UFPE

Intruduction

- Spatial regression models are usually intended to estimate parameters related to the interaction of agents across space
  - Social interactions, agglomeration externalities, technological spillovers, strategic interactions between governments etc.

- In this class we will explore estimation of Social interactions models using STATA

- Methods of estimation

- Identification strategy

- As an example we will use some data on pupils' marks and look at the peer effect.

Data Set

- The paper evaluates the friendship peer effects on student academic performance. The identification comes from the unique student friendship dataset from a Brazilian public institution (FUNDAJ), the strategy considers the architecture of these social networks within classrooms, in addition to group and individual fixed effects

- The file fundaj.dta is a random sample of 1,431 students from 120 schools in Recife city.

General set up: Peer effect at school

$$y_i = x'_i\gamma + m(y,s)\beta + m(x,s)'_i\theta + m(k,s)'_i\delta + m(v,s)'_i\lambda + \varepsilon_i$$

- y is child's math marks
- x is gender, age, parents' education, etc
- m(y,s) is average child marks peer
- m(x,s) is average gender, age, parent's education at school s_i
- m(z,s) is other stuff at school e.g. principal wage
- m(v,s) are unobserved child characteristics (e.g. inteligence)

General set up

- See e.g. Le Sage and Pace Introduction to Spatial Econometrics

$$y_i = x'_i\gamma + m(y,s)\beta + m(x,s)'_i\theta + m(k,s)'_i\delta + m(v,s)'_i\lambda + \varepsilon_i$$

- **SAR (spatial autoregressive)** effects: captured by **β**
  - Spillovers from neighbouring region outcome on regional outcome e.g. patents
- **SLX (spatially lagged X)** effects, captured by **θ**
  - Influence of neighbouring regions' observable characteristics on regional outcome e.g. R&D expenditure
- **SE (spatial error)** represents unobserved similarity between neighbours or spillovers between unobservables
  - e.g. the innovative culture

General form of spatial regression

- Spatial econometrics:

$$y = X\gamma + Wy\beta + WX\theta + WZ\delta + Wv\lambda + \varepsilon$$

- Social interaction:
  - Outcome for i depends on the expected (average) outcome for the spatial group, average characteristics of the group and average unobservables of the group
  - Or some other sort of dependence (spillover) between group members and the individual

$$y = x\gamma + E[y_i|W_i]\beta + E[x_i|W_i]\theta + E[z_i|W_i]\delta + E[v_i|W_i]\lambda + \varepsilon$$

Endogenous effect/SAR specifications

- These are specifications with a spatially lagged dependendent variable

$$y_i = x'_i\gamma + m(y, s)\beta + u_i$$
$$y = x\gamma + E[y_i|W_i]\beta + \varepsilon$$

- Theory is that children mark depends on peer effect
  - **Outcome** is dependent on the observable **outcome** for peers (neighbours)
  - ρ supposed to represent reaction functions, direct spillovers from peers (neighbours) occurring through observed behaviour.

Mechanical feedback endogeneity

- Unbiased and consistent estimation by OLS requires that error term and regressors are uncorrelated. Does this assumption hold for this model?

- Consider simple i-j case

$$y_i = \rho y_j + x_i \beta + u_i$$

$$y_j = \rho y_i + x_j \beta + u_j$$

$$\Rightarrow$$

$$y_i = \rho \left\{ \rho y_i + x_j \beta + u_j \right\} + x_i \beta + u_i$$

$$= \rho \left\{ \rho \left(\rho y_j + x_i \beta + \boxed{u_i} \right) + x_j \beta + u_j \right\} + x_i \beta + \boxed{u_i}$$

- The 'spatially lagged' or 'average neighbouring' dep. var. y_j is correlated with the unobserved error term:

Instrumental variables

- Good Instrument
  - **1. Correlated with endogenous variable z, conditional on x: 'powerful first stage'**
  - **2. Uncorrelated with v: ' satisfies the exclusion restriction'**

- Instrument is variable that predicts the endogenous variable y_j but does not affect outcome y_i directly

Instrumental variables

- Gibbons, Stephen and Overman, Henry G. (2012) Mostly pointless spatial econometrics. Journal of regional science, 52 (2). pp. 172-191

- So a possible set of 'instruments' (predictors) for **Wy** are

$$[\boldsymbol{WX}, \boldsymbol{W^2X}, \boldsymbol{W^3X}, \ldots]$$

- Correlated with peers marks but not with pupils marks

Computer exercise

Data set

- Classes room best friends of each student marques V2-V1432
- The students math marks - marks
- Student characteristics – popular and boy=1
- School characteristics – principal_wage

```
. tab idpupil v5 if idpupil<=25

             |          v5
    idpupil  |         0          1 |     Total
-------------+----------------------+----------
         10  |         1          0 |         1
         14  |         1          0 |         1
         16  |         0          1 |         1
         18  |         1          0 |         1
         21  |         1          0 |         1
         22  |         0          1 |         1
         23  |         1          0 |         1
         25  |         1          0 |         1
-------------+----------------------+----------
      Total  |         6          2 |         8
```

- First we describe situation in which we have the spatial-weighting matrix precomputed and simply want to put it in an spmat object

  spmat dta peer v2-v1432, id(idchild) replace

```
. spmat summarize peer, links

Summary of spatial-weighting object peer
-----------------------------------------------
        Matrix |              Description
---------------+-------------------------------
    Dimensions |              1431 x 1431
     Stored as |              1431 x 1431
 Links         |
         total |                       3558
           min |                          1
          mean |                   2.486373
           max |                         10
-----------------------------------------------
```

- Estimate a regression to look at effect of popular, boy and principal wage on child marks using classical special econometrics model: SAR

$$y = \rho Wy + X\beta + u$$

```
spreg ml mark  popular boy principal_wage,id(idpupil) dlmat(peer) nolog

Spatial autoregressive model                          Number of obs   =      1431
(Maximum likelihood estimates)                        Wald chi2(3)    =  19.8763
                                                      Prob > chi2     =   0.0002


--------------------------------------------------------------------------------
       mark    |     Coef.    Std. Err.      z     P>|z|     [95% Conf. Interval]
---------------+----------------------------------------------------------------
mark           |
      popular  |   1.809878    .5849103    3.09    0.002     .6634747     2.95628
         boy   |    .249489    .7963501    0.31    0.754    -1.311329     1.81030
principal_wage |    -.00121    .0003863   -3.13    0.002    -.0019671    -.000452
        _cons  |   39.17803    1.745047   22.45    0.000      35.7578     42.5982
---------------+----------------------------------------------------------------
lambda         |
        _cons  |   .0315783    .0055472    5.69    0.000      .020706     .042450
---------------+----------------------------------------------------------------
sigma2         |
        _cons  |   214.8521    8.03456    26.74    0.000     199.1047     230.599
--------------------------------------------------------------------------------
```

- The estimated ρ coefficient is positive and significant, indicating SAR dependence. In other words, an exogenous shock to one pupil will cause changes in the marks in the class peers.
- The estimated θ and δ vector does not have the same interpretation as in a simple linear model, because including a spatial lag of the dependent variable implies that the outcomes are determined simultaneously.

```
. spreg gs2sls mark popular boy principal_wage, id(idpupil) dlmat(peer)

Spatial autoregressive model                          Number of obs   =      1431
(GS2SLS estimates)


-----------------------------------------------------------------------------
        mark    |      Coef.    Std. Err.       z     P>|z|     [95% Conf. Interval]
----------------+------------------------------------------------------------
mark            |
       popular  |   1.783077    .5856426      3.04    0.002     .6352389     2.93091
           boy  |   .1485575    .8012924      0.19    0.853    -1.421947     1.71906
principal_wage  |  -.0012348     .000387     -3.19    0.001    -.0019934    -.000476
         _cons  |   39.76611    1.815093     21.91    0.000     36.20859     43.3236
----------------+------------------------------------------------------------
lambda          |
         _cons  |   .0274628    .0065471      4.19    0.000     .0146307     .040294
-----------------------------------------------------------------------------
```

There are no apparent differences between the two sets of parameter estimates.

- classical special econometrics model: SARAR

$$\mathbf{y} = \boldsymbol{\rho}\mathbf{W}\mathbf{y} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{u}$$
$$\mathbf{u} = \boldsymbol{\rho}\mathbf{W}\mathbf{u} + \mathbf{e}$$

```
. spreg ml mark popular boy principal_wage,id(idpupil) dlmat(peer) elmat(peer) nolog

Spatial autoregressive model                    Number of obs   =      1431
(Maximum likelihood estimates)                  Wald chi2(3)    =   18.3844
                                                Prob > chi2     =    0.0004


--------------------------------------------------------------------------------
        mark2 |      Coef.    Std. Err.      z     P>|z|     [95% Conf. Interval]
--------------+-----------------------------------------------------------------
mark          |
      popular |   1.725857    .5877355     2.94    0.003     .573917     2.87779
          boy |    .204751    .8264258     0.25    0.804    -1.415014    1.82451
principal_wage |  -.0012425    .0004083    -3.04    0.002    -.0020429    -.000442
        _cons |   39.97857    1.864488    21.44    0.000     36.32424     43.632
--------------+-----------------------------------------------------------------
lambda        |
        _cons |   .0261876    .0066554     3.93    0.000     .0131433      .03923
--------------+-----------------------------------------------------------------
rho           |
        _cons |   .0234643    .0129945     1.81    0.071    -.0020045      .04893
--------------+-----------------------------------------------------------------
sigma2        |
        _cons |   214.2306    8.014287    26.73    0.000     198.5229     229.938
--------------------------------------------------------------------------------
```

Estimation using IV/2SLS

- Use spmat to creat spatial lag of *mark, boy* and *popular*

spmat lag double wmark peer mark

spmat lag double wpopular peer popular

spmat lag double wboy peer boy

```
. sum wmark wpop wboy mark2 pop boy

    Variable |        Obs        Mean    Std. Dev.       Min        Max
-------------+--------------------------------------------------------
       wmark |      1,431    104.5283    67.11725          0        465
    wpopular |      1,431    3.259958    2.151483          1         14
        wboy |      1,431    .9357093    1.225055          0          8
        mark |      1,431    41.16352    14.95653          0         85
     popular |      1,431    1.341719    .6647463          1          3
-------------+--------------------------------------------------------
         boy |      1,431    .4255765     .494603          0          1
```

- Including the spatial lag of *mark, sex* and *popular* in the regressions

```
. regress mark wmark popular wpopular boy wboy principal_wage, cluster(idesc)


Linear regression                              Number of obs    =       1,431
                                               F(6, 110)        =        8.87
                                               Prob > F         =      0.0000
                                               R-squared        =      0.0437
                                               Root MSE         =      14.657


                                (Std. Err. adjusted for 111 clusters in idesc)
------------------------------------------------------------------------------
               |               Robust
          mark |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
---------------+--------------------------------------------------------------
         wmark |   .0527915   .0141398     3.73   0.000     .0247697    .080813
       popular |   1.828549   .5648417     3.24   0.002     .7091654    2.94793
      wpopular |  -.3729638   .3839338    -0.97   0.333     -1.13383    .387902
           sex |   1.877473     1.0537     1.78   0.078    -.2107139     3.9656
          wsex |  -.9674357   .4718134    -2.05   0.043    -1.902459   -.032412
principal_wage |  -.0010861   .0003935    -2.76   0.007    -.0018659   -.000306
         _cons |   37.96964   1.968381    19.29   0.000     34.06877    41.8705
------------------------------------------------------------------------------
```

- Estimate the 2SLS/IV regression using wpopular and wboy as instruments for wmark – FIRST STAGE

. reg wmark wpopular wboy boy popular principal_wage ,cluster(idesc)

```
Linear regression                               Number of obs   =        1,431
                                                F(5, 110)       =       156.68
                                                Prob > F        =       0.0000
                                                R-squared       =       0.7154
                                                Root MSE        =       35.868

                              (Std. Err. adjusted for 111 clusters in idesc)
------------------------------------------------------------------------------
                |               Robust
         wmark  |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
----------------+-------------------------------------------------------------
       wpopular |   24.10899    1.15788     20.82   0.000     21.81434    26.4036
           wboy |   7.565048   2.166797      3.49   0.001     3.270965    11.8591
            boy |  -16.92019   2.544356     -6.65   0.000     -21.9625   -11.8778
        popular |   -2.92456   1.723615     -1.70   0.093    -6.340361    .491240
 principal_wage |   -.003971    .0015175    -2.62   0.010    -.0069785   -.000963
          _cons |   42.61488   7.048311      6.05   0.000     28.64678    56.5829
------------------------------------------------------------------------------
```

. testparm wpopular wboy
 (1)   wpopular = 0
 (2)   wboy = 0
       F(2,   110) = **254.35**
            Prob > F =   0.0000

- Estimate the 2SLS/IV regression using wpopular and wboy as instruments for wmark – IV

```
. ivreg mark (wmark= wpopular wboy) popular boy principal_wage ,cluster (idesc)

Instrumental variables (2SLS) regression        Number of obs   =       1,431
                                                 F(4, 1430)      =        9.31
                                                 Prob > F        =      0.0000
                                                 R-squared       =      0.0382
                                                 Root MSE        =      14.689

                              (Std. Err. adjusted for 1,431 clusters in idesc)
------------------------------------------------------------------------------
                              Robust
         mark       Coef.    Std. Err.     t     P>|t|     [95% Conf. Interval]
--------------+---------------------------------------------------------------
        wmark    .0283833    .0077867    3.65    0.000      .0129518     .04381
      popular    1.789071     .587324    3.05    0.003      .6251313    2.95301
          boy    .1711308    .8216546    0.21    0.835     -1.457196    1.79945
principal_wage  -.0012293      .00041   -2.96    0.004     -.0020536    -.00040
        _cons    39.63459    2.199265   18.02    0.000      35.27616   43.99301
------------------------------------------------------------------------------
Instrumented:  wmark
Instruments:   popular boy principal_wage
               wpopular wboy
------------------------------------------------------------------------------
```

Limitations of this approach

- Following Gibbons et. al. (2012):
  - IV/2SLS relies on instruments WX, W^2X etc. having no direct effect on y

  - In principle you can use W^2X... W^3X as instruments, for Wy in the equation assuming W^2X... W^3X don't belong in this equation:

$$y = \rho Wy + X\beta_1 + WX\beta_2 + e$$

  - Difficult to justify if W chosen arbitrarily

  - Also WX, W^2X... W^3X are all likely to be very highly correlated (remember these are all averages) so W^2X... W^3X not likely to be a good predictor of Wy, conditional on WX