nwxtregress: Network regressions in Stata

Stata Conference 2021

Jan Ditzen[1], William Grieser[2], Morad Zekhnini[3]

[1]Free University of Bozen-Bolzano, Italy
jan.ditzen@unibz.it

[2]Texas Christian University
w.grieser@tcu.edu

[3]Michigan State University
zekhnini@msu.edu

August 5, 2021

Economic and social agents are not independent

- Empirical analysis in social sciences (nearly) invariably relies on the assumption of cross-sectional independence.
  - E.g., the Gauss-Markov theorem assumes independence of disturbances.
- Most real-world applications involve interactions between units of observation.
  - E.g., companies buy and sell from one another, individuals share information with family and friends, etc.

Many applications of interactions are best represented using networks

- The inherent interactions between social entities has spawned a wide literature on social networks (Borgatti et al., 2009).
- Networks (or graphs) parsimoniously capture many economic settings:
  - Economic entities are the vertices or nodes of the network.
  - E.g., the US economy is made up of a collection of industries.
  - Relationships between entities are the edges, links, or ties of the network.
  - E.g., the dollar value of transactions between two industries.
- A key question remains: how do we analyze outcomes in a regression framework in the context of networks?
  - cross-sectional independence cannot be assumed!

Literature

- Network description and mathematical models (Newman, 2010; Jackson, 2010)
- Spatial econometrics
  - Initially used in regional science to model neighboring regions
  - Empirical models and estimation techniques with a priori knowledge of relationship between units (LeSage and Pace, 2009; Kelejian and Piras, 2017)
- Using spatial models to identify peer effects in social networks (Bramoullé et al., 2009; Grieser et al., 2021)

Interactions pose identification challenges

- Consider a traditional panel model with 2 units:

$$y_{1t} = X_{1t}\beta + \epsilon_{1t}$$

$$y_{2t} = X_{2t}\beta + \epsilon_{2t}$$

- The independence assumption implies: $E[\epsilon_1\epsilon_2] = E[\epsilon_1]E[\epsilon_2]$
- This rules out the possibility that units 1 and 2 interact.
- Thus, for many applications, a more appropriate model is:

$$y_{1t} = \rho y_{2t} + X_{1t}\beta + \epsilon_{1t}$$

$$y_{2t} = \rho y_{1t} + X_{2t}\beta + \epsilon_{2t}$$

- This clearly violates independence (endogenous outcome y on RHS)
- Simultaneity invalidates inferences based on direct estimation

A parsimonious model of interactions

- Generalizing the panel model gives:

$$y_{it} = \sum_{j \neq i} \rho_{ij} y_{jt} + X_{it}\beta + \epsilon_{it}$$

- Considering all interactions ($\approx N^2$) is impractical
- Ord (1975) proposed the parsimonious parameterization:

$$y_{it} = \rho \sum_{j \neq i} w_{ij,t} y_{jt} + X_{it}\beta + \epsilon_{it}$$

- w_{ij} represents a priori link between i and j

We must invert the model to solve it

- It is more convenient to use matrix notation
- If we stack all elements in conforming vectors/matrices:

$$y = \rho W y + X\beta + \epsilon$$

- This is know as the Spatial Autoregressive (SAR) model
- Estimating the model "as is" poses various challenges (Manski, 1993; Angrist, 2014)
- Solving for a reduced-form data generating process is more useful:

$$y = (I - \rho W)^{-1}(X\beta + \epsilon)$$

- Note ys only appear on LHS, but model is nonlinear in parameters

The Model implies geometrically-decaying propagation

- Given mathematical restrictions on ρ and W:

$$(I - \rho W)^{-1} = I + \rho W + \rho^2 W^2 + \dots$$

- Interpret outcome as geometric sum of:
  - Own effect (I term)
  - Immediate peers' effect (W term)
  - Peers of peers effect (W^2 term)
  - etc

Partial derivatives are no longer βs

- In traditional model:

$$\frac{\partial y_i}{\partial x_i} = \beta, \text{ and } \frac{\partial y_i}{\partial x_j} = 0, \; i \neq j$$

- In the model with interactions:

$$\frac{\partial y_i}{\partial x_j} = (I - \rho W)^{-1}_{ij} \beta, \; \forall i, j$$

- Listing all partial derivatives is impractical.
- LeSage and Pace (2009) propose summarizing partial derivative estimates into direct and indirect effect averages:
  - Direct: $\frac{1}{N} \sum_i \frac{\partial y_i}{\partial x_i}$
  - Indirect: $\frac{1}{N} \sum_i \sum_{j \neq i} \frac{\partial y_i}{\partial x_j}$

The SDM adds contextual effects to SAR

- The Spatial Durbin Model (SDM) is given by:

$$y = \rho Wy + X\beta + WX\theta + \epsilon$$

- The values in WX represent the covariates of peers
- The effect of these covariates is often referred to a contextual effect
- These values are assumed exogenous and do not materially change the estimation

A short primer on estimation

- Focusing on one cross-section (for notational convenience), the likelihood function of the model is:

$$
\begin{aligned}
f(Y, X; \rho, \beta, \sigma^2) &= |I_N - \rho W|(2\pi\sigma^2)^{-N/2}\exp(-\frac{e'e}{2\sigma^2}) \\
e &= (I - \rho W)Y - X\beta
\end{aligned}
$$

- If ρ is known (say ρ_0), then β (and σ^2) can be integrated out in a maximum likelihood estimation (MLE)
- The problem becomes an optimization w.r.t. ρ only
- The estimation proceeds with an MCMC sampler using the above likelihood

nwxtregress[1]

Syntax

Spatial Autocorrelation Model (SAR)

`nwxtregress` *depvar indepvars* [*if*] , *dvarlag(W1[,options1)])*
[`mcmc_options nosparse`]

Spatial Durbin Model (SDM)

`nwxtregress` *depvar indepvars* [*if*] , *dvarlag(W1[,options1)])*
ivarlag(W2[,options1) [`mcmc_options nosparse`]

- `W1` and `W2` define spatial weight matrices, default is Sp object.
- Note: nwxtregress allows for unbalanced panels and time varying W1 and W2 (unlike spxtreg)

[1] This command is work in progress. Options, functions and results might change.

nwxtregress

Spatial Weight Options

`nwxtregress` *depvar indepvars* [*if*] , *dvarlag(W1[,options1)])*

[`ivarlag(W2[,options1) mcmc_options nosparse`]

- `options1` controls the spatial weight matrices:
  - `mata` declares weight matrix is `mata` matrix.
  - `sparse` if weight matrix is sparse.
  - `timesparse` weight matrix is sparse and varying over time.
  - `id(string)` vector of IDs if W is a non sparse mata matrix.

nwxtregress

Further Options

`nwxtregress` *depvar indepvars* [*if*] , *dvarlag(W1[,options1)])*
[`ivarlag(W2[,options1) mcmc_options nosparse`]

- `nosparse` do not convert weight matrix internally to a sparse matrix.
- `mcmc_options` control the Markov Chain Monte Carlo:
  - `draws(integer 2000)` number of griddy gibs draws.
  - `gridlength(integer 1000)` grid length
  - `nomit(integer 500)` number of omitted draws
  - `barrypace(numlist)` settings for BarryPace Trick, iterations, maxorder default: 50 100
  - `usebp` use BarryPace trick instead of LUD for inverse of $I - \rho W$.
  - `seed(#)` sets the seed.

Example: BEA I/O Tabels

Data

- We collect USE/MAKE table data from the BEA's website
- These data represent the goods that were used (USE) and made (MAKE) by each industry in the US
- To construct links between industries, we convert into flows between industries
- Loaded data as *Sp* matrix using `spmatrix fromdata W = sam* , replace`, but only for year 1998.
- We also collect key variables about each industry: capital consumption, compensation, and net surplus

SAR

Time constant spatial weights

```
. nwxtregress cap_cons compensation net_surplus , ///
> dvarlag(W) seed(1234)
Griddy Gibbs (2000)
----+--- 10 ---+--- 20 ---+--- 30 ---+--- 40 ---+--- 50   %
..................................................   50
..................................................  100

Spatial SAR                            Number of obs      =       1358
Panel Variable (i): ID                 Number of groups   =         62
Time Variable (t): Year                Obs. of group:               22
                                                     min =          19
                                                     avg =          22
                                                     max =          22
                                       R-squared          =       0.73
                                       Adj. R-squared     =       0.73
```

cap_cons	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
compensation	-1.303834	.0230025	-56.68	0.000	-1.386128	-1.232569
net_surplus	-1.187692	.0236269	-50.27	0.000	-1.268585	-1.100094
W						
cap_cons	.1119915	.0272865	4.10	0.000	.007	.199
\sigma_u	.2659483	.0103939			.2328073	.3016006

Example

Time varying spatial weight

- Saved network data in *timesparse* format in mata as W.
- The first column identifies the year, second and third the IDs and the last one the value of the weight.
- Non standardized timesparse W:

```
. mata W[1..10,.]
                 1             2             3             4
     +---------------------------------------------------------+
   1 |        1997             1             1    120.445105   |
   2 |        1997             1             2   2646.806067   |
   3 |        1997             1             3             0   |
   4 |        1997             1             4   1594.653373   |
   5 |        1997             1             5   93.56892452   |
   6 |        1997             1             6   67.68815816   |
   7 |        1997             1             7   668.9182689   |
   8 |        1997             1             8   607.2025953   |
   9 |        1997             1             9   444.9500985   |
  10 |        1997             1            10   1884.318874   |
     +---------------------------------------------------------+
```

SAR

```
. nwxtregress cap_cons compensation net_surplus , ///
> dvarlag(W,mata timesparse) seed(1234)
Griddy Gibbs (2000)
────┼─── 10 ──┼─── 20 ──┼─── 30 ──┼─── 40 ──┼─── 50   %
..................................................    50
..................................................   100

Spatial SAR                                       Number of obs     =       1358
Panel Variable (i): ID                            Number of groups  =         62
Time Variable (t): Year                           Obs. of group:              22
                                                                min =         19
                                                                avg =         22
                                                                max =         22
                                                  R-squared         =       0.73
                                                  Adj. R-squared    =       0.73
──────────────────────────────────────────────────────────────────────────────────
       cap_cons │      Coef.   Std. Err.      z    P>|z|     [95% Conf. Interval]
────────────────┼─────────────────────────────────────────────────────────────────
   compensation │  -1.310997   .0226358   -57.92   0.000    -1.391276    -1.24096
    net_surplus │  -1.195375   .0232523   -51.41   0.000    -1.274683   -1.109078
────────────────┼─────────────────────────────────────────────────────────────────
W               │
       cap_cons │    .094567    .025236     3.75   0.000        -.003        .175
────────────────┼─────────────────────────────────────────────────────────────────
       \sigma_u │   .2665574   .0104167                        .2333413    .3022325
──────────────────────────────────────────────────────────────────────────────────
```

SAR

Direct Indirect Effects

```
. estat impact
Average Impacts                                 Number of obs =       1358
```

cap_cons	dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
direct						
compensation	-1.316631	.0227273	-57.93	0.000	-1.393996	-1.245314
net_surplus	-1.200513	.0233336	-51.45	0.000	-1.279492	-1.111448
indirect						
compensation	-.118322	.0315852	-3.75	0.000	-.2292999	.0038375
net_surplus	-.1078809	.0287743	-3.75	0.000	-.2066906	.0034616
total						
compensation	-1.434953	.0406058	-35.34	0.000	-1.612898	-1.297665
net_surplus	-1.308393	.0386178	-33.88	0.000	-1.453865	-1.161205

SDM

```
. nwxtregress cap_cons compensation net_surplus , ///
> dvarlag(W,mata timesparse) ///
> ivarlag(W: compensation,mata timesparse )  seed(1234)
Griddy Gibbs (2000)
────┼─── 10 ───┼─── 20 ───┼─── 30 ───┼─── 40 ───┼─── 50   %
..................................................     50
..................................................    100

Spatial SDM                                       Number of obs     =       1358
Panel Variable (i): ID                            Number of groups  =         62
Time Variable (t): Year                           Obs. of group:              22
                                                              min =         19
                                                              avg =         22
                                                              max =         22
                                                  R-squared         =       0.73
                                                  Adj. R-squared    =       0.73
```

cap_cons	Coef.	Std. Err.	z	P>|z|	[95% Conf.	Interval]
compensation	-1.311007	.023423	-55.97	0.000	-1.399435	-1.235371
net_surplus	-1.195046	.0239872	-49.82	0.000	-1.267235	-1.108135
W						
cap_cons	.1033565	.0270195	3.83	0.000	.013	.19
compensation	.0177968	.0267131	0.67	0.505	-.0679343	.107463
\sigma_u	.2669513	.0101299			.2355723	.3033771

SDM

Direct Indirect Effects

```
. estat impact
Average Impacts                                 Number of obs =       1358
```

cap_cons	dy/dx	Std. Err.	z	P>|z|	[95% Conf.	Interval]
direct						
compensation	-1.311734	.0234539	-55.93	0.000	-1.40014	-1.23592
net_surplus	-1.195803	.0240079	-49.81	0.000	-1.26789	-1.108757
indirect						
compensation	-.1318438	.0523719	-2.52	0.012	-.3261685	.0389637
net_surplus	-.1382094	.0400208	-3.45	0.001	-.2875287	-.0158213
total						
compensation	-1.443578	.059526	-24.25	0.000	-1.656334	-1.250907
net_surplus	-1.334012	.0484298	-27.55	0.000	-1.526842	-1.179319

Future steps

- Most social networks contain very few edges relative to possible edges
  - Such networks are best represented via sparse matrices
  - Current support for sparse matrices in Mata is limited
  - Future improvements will rely on additional sparse matrix functions
- Streamlined support for fixed effects/intercept
- Implementation of convex combination of multiple networks (Debarsy and LeSage, 2020)
- Available on GitHub (https://janditzen.github.io/nwxtregress/) or directly in Stata:

  net install nwxtregress ,
  from(https://janditzen.github.io/nwxtregress/)
- Please, help us by providing feedback

THANK YOU!

References I

Borgatti, S. P., A. Mehra, D. J. Brass, and G. Labianca. 2009. Network analysis in the social sciences. science 323(5916): 892–895.

Bramoullé, Y., H. Djebbari, and B. Fortin. 2009. Identification of peer effects through social networks. Journal of econometrics 150(1): 41–55.

Debarsy, N., and J. P. LeSage. 2020. Bayesian model averaging for spatial autoregressive models based on convex combinations of different types of connectivity matrices. Journal of Business & Economic Statistics 1–33.

Grieser, W., C. Hadlock, J. LeSage, and M. Zekhnini. 2021. Network Effects in Corporate Financial Policies. Journal of Financial Economics (Forthcoming) .

Jackson, M. O. 2010. Social and economic networks. Princeton university press.

Kelejian, H., and G. Piras. 2017. Spatial Econometrics. Academic Press.

References II

LeSage, J. P., and R. K. Pace. 2009. Introduction to Spatial Econometrics. Florida CRC Press.

Newman, M. 2010. Networks: An Introduction. Oxford University Press.