Spatial autoregressive logit and probit using Stata: The spatbinary package

Daniele Spinelli[a]
daniele.spinelli@unimib.it

Anna Gloria Billè[b]   Alessio Tomelleri [c]

[a] *Dep. of Statistics and Quantitative Methods, University of Milano-Bicocca*
[b] *Dep. of Statistical Sciences, Alma Mater Studiorum - University of Bologna*
[c] *Research Institute for the Evaluation of Public Policies - Fondazione Bruno Kessler*

Italian Stata Conference. May 2024, Florence

Official Stata

Stata 15 introduced [SP]:

- manipulation of spatial matrices (spmatrix)
- official commands for spatial regression models (spregress,spxtregress and spivregress) estimate models with continuous dependent variables.

Community contributed

- spatwmat and spmat for matrix manipulation
- spmap and geoplot draw detailed maps (Pisati 2018; Jann 2023)
- spatial regression models in terms of
  - cross-sectional data (Pisati 2001),
  - spatial panel regressions (Belotti, Hughes and Mortari 2017)
  - endogenous regressors (Drukker, Prucha and Raciborski 2013).
- calculate travel time (Huber and Rust 2016; Weber and Péclat 2017)
- patial correlation tests (spatcorr),
- geocode data (Ozimek and Miles 2011)

Aim

- Commands to estimate **spatial regressions with binary dependent variables** are not available.
- Introducing `spatbinary` (Spinelli 2022), a command to estimate spatial autoregressive probit and logit models
  - compatible with `SP`
  - estimation
  - marginal effects
  - prediction
- an empirical example with real data provided by Tomelleri and Billé 2024

The binary SAR - 1

The spatial autoregressive model with binary response (BSAR) (Pinkse and Slade 1998; Klier and McMillen 2008; Billé and Leorato 2020)

- Binary response $y_i = I(U_i > 0)$, where $y_i \in \boldsymbol{y}, u_i \in \boldsymbol{U}$
- row-standardized continguity matrix $\boldsymbol{W}$
- $\boldsymbol{\beta}$ and ρ to be estimated

$$\boldsymbol{U} = \rho \boldsymbol{W}\boldsymbol{U} + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon}$$

- U_i is the unobserved *propensity* to observe $y_i = 1$

The binary SAR - 2

- the spatial autocorrelation parameter ρ implies clustering ($\rho > 0$) or dispersion ($\rho < 0$) in space
- distributional assumptions on the residuals lead to the **probit BSAR** (normal) or to the **logit BSAR** (logistic)
- residuals are correlated and heteroscedastic
- the error term variance is proportional to

$$\boldsymbol{V} = E(\epsilon'\epsilon) = [(\boldsymbol{I} - \rho\boldsymbol{W})'(\boldsymbol{I} - \rho\boldsymbol{W})]^{-1}$$

Estimation

GMM estimator (Hansen 1982; Pinkse and Slade 1998), estimates are chosen to minimize the quantity:

$$Q = n^{-1}[\epsilon(\boldsymbol{\beta}, \rho)' \boldsymbol{ZMZ}' \epsilon(\boldsymbol{\beta}, \rho)] \tag{1}$$

- $\boldsymbol{Z}$ is a set of instruments which may include the covariates and their spatial lags (Kelejian and Prucha 1998)
- Klier and McMillen 2008: the model is a non-linear two stage least squares (N2SLS) if $\boldsymbol{M} = (\boldsymbol{Z'Z})^{-1}$ and proposed a *linearized* version
- The `spatbinary` estimates the linearized and the full N2SLS

Linearized vs N2SLS

Advantages of the linearized model

- computational: no inversion of the matrix $\boldsymbol{I} - \rho\boldsymbol{W}$ is required
- the advantage is less pronounced if $\boldsymbol{W}$ is small or sparse
- good approximation if ρ is small

Disadvantages

- less efficient the N2SLS (Billé 2013)
- upwardly biased if $|\rho| > 0.5$

The coefficients from the linearized model can be used as starting values for the N2SLS model. This is the default setting in `spatbinary`

Syntax

Data should be spset before using spatbinary. The main options are [1]:

spatbinary *depvar* [*indepvars*] [*if*] [*in*] [*weight*],wmat(*matname*) [logit probit linearized n2sls instr(*varlist*) winstr(*varlist*) impower(#)]

- wmat(*matname*), the spatial weight matrix created using spmatrix.
- probit or logit: estimate a logit or probit model
- linearized or n2sls: fits the linearized or N2SLS model . The default is linearized, if n2sls is chosen estimates from linearized are used as starting values.
- instr a *varlist* of instruments
- winstr a *varlist* of instruments to be premultiplied by the spatial weight matrix up to degree chosen by impower(#). Default is 1.

[1] estimation options are also allowed

Postestimation

- spatbinary allows predict
- allows spatbinary_impact: a wrapper of margins that estimates measures of impact such as **direct, indirect and total marginal effects** (Billé and Leorato 2020)
- spatbinary_impact corresponds to official Stata's estat impact for spregress postestimation.

spatbinary_impact *varlist*, eyex dydx eydx dyex total direct indirect

- dydx. marginal effect of *varlist* on the predicted probability.
- eyex, eydx and dyex. Calculates the elasticities and semielasticities of the predicted probability wrt *varlist*.
- total,direct and indirect. Calculates the total, direct (own-effect) and indirect (other unit's effect) measure of impact of *varlist*,

General workflow - 1

Installation

```
. net install st0672
```

Setup using spmatrix and spset

```
. webuse homicide1990, clear
(S.Messner et al.(2000), U.S southern county homicide rates in 1990)
. copy https://www.stata-press.com/data/r17/homicide1990_shp.dta .
. spmatrix clear
. spmatrix create contiguity W2, normalize(row)
. spset
      Sp dataset: homicide1990.dta
Linked shapefile: homicide1990_shp.dta
            Data: Cross sectional
 Spatial-unit ID: _ID
     Coordinates: _CX, _CY (planar)
. quietly sum hrate, det
. gen hrate_gt_p95=hrate>r(p95)
```

General workflow - 2

Estimation (using n2sls)

```
. spatbinary hrate_gt_p95 ln_population gini, wmat(W2) n2sls
instruments set as (X,WX...W^n X)
(output omitted)
N2SLS LOGIT
```

hrate_gt_p95	Coefficient	Robust std. err.	z	P>|z|	[95% conf.	interval]
hrate_gt_p95						
ln_population	.2088806	.176295	1.18	0.236	-.1366513	.5544124
gini	41.17571	6.693724	6.15	0.000	28.05625	54.29517
_cons	-23.58003	4.516926	-5.22	0.000	-32.43304	-14.72702
rho						
_cons	-.4242538	.2173661	-1.95	0.051	-.8502837	.001776

```
Test of overidentifying restriction:
Hansen´s J chi2(1) = .0588395, p = .8083396
```

General workflow - 3

Measures of impact

```
. spatbinary_impact gini, dydx
Impact measures for gini
```

	dydx	Delta-M~d std. err.	z	p>|z|	[95 conf.	interval]
gini						
total	1.198613	.1923828	6.230356	4.65e-10	.8215498	1.575676
direct	1.698983	.2554343	6.65135	2.90e-11	1.198341	2.199625
indirect	-.5003701	.2555695	-1.957863	.0502461	-1.001277	.000537

Overview - 1

Tomelleri and Billé 2024:

Do Micro-Enterprises Ask for Local Support Measures? Evidence After the COVID-19 Pandemic using a Spatial hurdle model

- Investigate the impact of spatial dependence as a measure of interaction effects on the take-up rate of local government subsidies in 2020 in Trentino.
- Specific sub-population of firms hit particularly hard by the pandemic: micro-enterprises (MEs).
- Link with administrative data on structure and performance.
- Lack of information about the coordinates of MEs due to privacy reasons (economic metric for the weighting matrix).
- observations grouped into three areas (East, West and Central): we present results only from the Eestern Area

Overview - 2

- Covariates:
  1. *ln(turnover)* is the logarithm of the average turnover between 2017 and 2019,
  2. *imp lockdown* reports whether the firm was forced to close by the government in 2020,
  3. *employees* = 1 if the firm have more than one employee, = 0 otherwise
  4. *firm age* is the number of years since the firm was registered,
  5. *national aid* identifies firms who also resorted for national support,
  6. four dummy variables, the strategies adopted by the firm
     - resorting to self-financing;
     - resorting to borrowing from friend/family members;
     - changing payment terms with customers;
     - changing payment terms with suppliers

Data

	Mean	SD	Min	Max	N
			East		
turnover 17-19	152,412	229,468	4838	2.3e+06	367
added value 17-19	57,696	67,071	-2.8e+04	5.4e+05	367
ln(turnover 17-19)	11.26	1.11	8.48	14.64	367
ln(added value 17-19)	10.53	0.98	3.96	13.20	360
imp_lockdown	0.62	0.49	0.00	1.00	367
employees	0.30	0.46	0.00	1.00	367
firm age	20.05	11.95	3.00	60.00	367
self-financing	0.27	0.44	0.00	1.00	367
loans from family/friends	0.11	0.31	0.00	1.00	367
payment cond. customers	0.07	0.26	0.00	1.00	367
payment cond. suppliers	0.14	0.35	0.00	1.00	367
national aids	0.75	0.43	0.00	1.00	367

Model specification - 1

- In the full sample, 364 MEs were not eligible (they received a rejection). Take-up is conditional of eligibility.
- empirical strategy considers a **spatial hurdle model**
  1. **eligibility equation**. Measures the participation decisions,
  2. **main equation**. Measures the MEs decisions, among the active ones, of asking for local support measures conditional on participation.
- The second equation is estimated using `spatbinary`

Depending on eligibility ($d_i = 1$), and on covariates x_i the probability that ME i applies for local support ($y_i = 1$) is

$$P(y_i = 1|x_i) = \begin{cases} P\left(d_i = 0|x_i\right) & \textit{if} \quad y_i = 0 \\ P\left(d_i = 1|x_i\right) P(y_i = 1|d_i = 1, x_i) & \textit{if} \quad y_i = \{0, 1\} \end{cases}$$

Model specification - 2

The second equation then specify a spatial **autoregressive probit model**

$$y^* = \rho W y^* + X_2 \beta_2 + \varepsilon_2 \quad \varepsilon_2 \sim \mathcal{N}(0, I)$$

where W is an n by n matrix of weights connecting the spatial latent variable[2] y^* and ρ is the corresponding spatial autoregressive coefficient. Asking for local support be ME is observed only if

$$y = \mathrm{I}(y^* > 0)$$

- spatial spillovers can be interpreted as peer effects among MEs.
- **direct, indirect and total marginal effects** are estimated also taking into account the first equation. [3]

[2] propensity to ask for local support

[3] this requires assumptions, please see details in Tomelleri and Billé 2024

Weighting matrix

- Coordinates of MEs are unknown due to statistical confidentiality
- weighting matrix $W = \{w_{ij}\}$ is built by using an economic variable[4], i.e. the mean 2017-2019 of the micro-firms' added values ($\bar{av}$)

$$\begin{cases} w_{ij} = \frac{1}{|\bar{av}_i - \bar{av}_j|} & if \quad i \neq j \\ w_{ij} = 0 & otherwise \end{cases}$$

- takes into account similarities in terms of added value.
- W is row-normalized (i.e., $\sum_j w_{ij} = 1$)

[4] see, for instance, Case, Rosen and Hines Jr 1993 who rely on a similar economic definition of the weighting matrix.

Setup

Setup using a matrix stored in an external file

```
. clear all

. import delimited "distEst3.csv"
(encoding automatically selected: ISO-8859-2)
(368 vars, 367 obs)

. drop v1

. mkmat v*, matrix(spatmat)

. use data.dta, clear

. spset ID

Sp dataset: data.dta
Linked shapefile: <none>
Data: Cross sectional
Spatial-unit ID: _ID (equal to ID)
Coordinates: <none>

. mata: W=st_matrix("spatmat")
. mata: ID=1::rows(W)
. spmatrix spfrommata W = W ID
```

Coefficient estimates

```
 . spatbinary local_aid $X, wmat(W) probit n2sls  noc
instruments set as (X,WX...W^n X) where X= ln_ricven1719 imp_lockdown i.dip_cat i.frm_g
firm_age ib2.settore liquid_C03_3 liquid_C03_4 liquid_C03_8 liquid_C03_9 i.treatment1
and W=W where n=1
  (367 observations)
  (367 observations (places) used)
  (weighting matrix defines 367 places)
Iteration        1:   GMM criterion Q(b) =         0.020022488708
(output omitted)
Iteration        7:   GMM criterion Q(b) =         0.019416323129
N2SLS PROBIT
------------------------------------------------------------------------------------------
                          |             Robust
                local_aid | Coefficient  std. err.        z    P>|z|      [95% conf. interval]
--------------------------+---------------------------------------------------------------
local_aid                 |
             ln_ricven1719 |  -.1427016   .0284251    -5.02   0.000    -.1984137   -.0869895
              imp_lockdown |   .2735895    .165672     1.65   0.099    -.0511217    .5983006
(output omitted)
--------------------------+---------------------------------------------------------------
rho                       |
                         1 |   .3718604   .2090455     1.78   0.075    -.0378612    .7815819
------------------------------------------------------------------------------------------
Test of overidentifying restriction:
Hansen´s J chi2(16) = 7.125791, p = .9707596
```

Marginal effects - 1

- marginal effects for the second equation
- they are to be interpreted as the change in probability of asking for local support associated to a 1% variation in turnover conditional on eligibility
- direct refers to own-effects, indirect refers to spillover effects, total aggregates them

```
. spatbinary_impact ln_ricven1719, dydx
Impact measures for ln_ricven1719
```

	dydx	Delta-M~d std. err.	z	p>|z|	[95 conf.	interval]
ln_ricv~1719						
total	-.0560494	.0229092	-2.446587	.0144216	-.1009506	-.0111481
direct	-.0358173	.0062999	-5.685387	1.31e-08	-.0481648	-.0234697
indirect	-.0202321	.0190645	-1.061243	.2885794	-.0575978	.0171337

Marginal effects - 2

- **direct marginal effects** for the second equation at the individual level

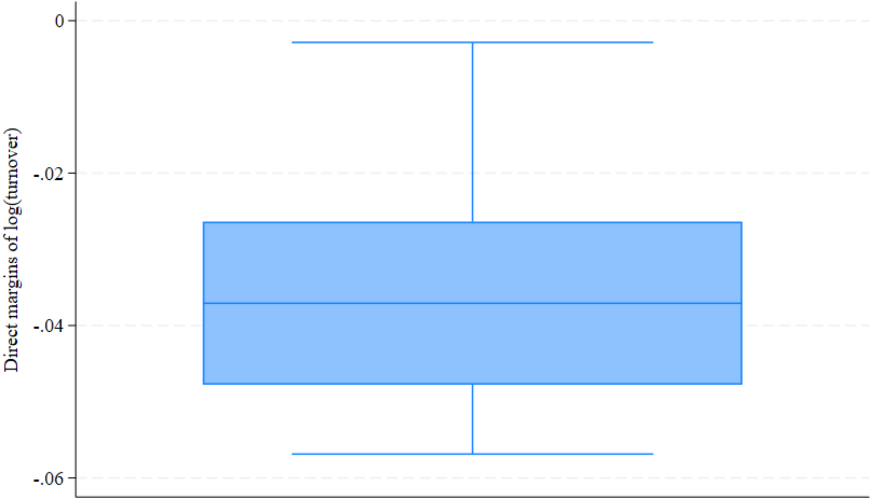
```
. predict dirmar_ln_ricven1719 , directmargin
Marginal effect

. replace dirmar_ln_ricven1719=dirmar_ln_ricven1719*_b[local_aid: ln_ricven1719]
(367 real changes made)

. summarize dirmar_ln_ricven1719
```

Variable	Obs	Mean	Std. dev.	Min	Max
dirmar_~1719	367	-.0358173	.0141344	-.0568698	-.0028516

```
. tabstat dirmar_ln_ricven1719, stat(p5 p25 p50 p75 p95)
```

Variable	p5	p25	p50	p75	p95
dirmar_~1719	-.0555428	-.0477662	-.0370793	-.0263814	-.010658

Marginal effects - 2



Marginal effects - 3

- marginal effects for the **hurdle model**, they take into account participation
- they use the `phat_1` variable: the participation probability from the first equation
- they are to be interpreted as the change in probability of asking for local support associated to a 1% variation in turnover

Marginal effects - 3

```
. margins, expression(phat_1*predict(totalmargin)*_b[local_aid: ln_ricven1719])
warning: cannot perform check for estimable functions.

Predictive margins                                    Number of obs = 367
Model VCE: Robust

Expression: phat_1*predict(totalmargin)*_b[local_aid: ln_ricven1719]
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_cons	-.032345	.0131934	-2.45	0.014	-.0582036	-.0064864

Marginal effects - 3

```
. margins, expression(phat_1*predict(directmargin)*_b[local_aid: ln_ricven1719])
warning: cannot perform check for estimable functions.

Predictive margins                                  Number of obs = 367
Model VCE: Robust

Expression: phat_1*predict(directmargin)*_b[local_aid: ln_ricven1719]
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_cons	-.0206689	.003668	-5.63	0.000	-.027858	-.0134799

```
.                  margins, expression(phat_1*predict(indirectmargin)*_b[local_aid: ln_ricven171
warning: cannot perform check for estimable functions.

Predictive margins                                  Number of obs = 367
Model VCE: Robust

Expression: phat_1*predict(indirectmargin)*_b[local_aid: ln_ricven1719]
```

	Margin	Delta-method std. err.	z	P>|z|	[95% conf.	interval]
_cons	-.0116761	.010984	-1.06	0.288	-.0332044	.0098522

Conclusion

- spatial probit and logit models using Stata
- possibile extensions:
  - the partial maximum likelihood modeling framework of Billé and Leorato 2020
  - spatial error models

Thanks

THANKS FOR YOUR ATTENTION!

Belotti, Federico, Gordon Hughes and Andrea Piano Mortari (2017). 'Spatial panel-data models using Stata'. In: *The Stata Journal* 17.1, pp. 139–180.

Billé, Anna Gloria (2013). 'Computational Issues in the Estima tion of the Spatial Probit Model: A Comparison of Various Estimators'. In: *Review of Regional Studies* 43.2, 3, pp. 131–154.

Billé, Anna Gloria and Samantha Leorato (2020). 'Partial ML estimation for spatial autoregressive nonlinear probit models with autoregressive disturbances'. In: *Econometric Reviews* 39.5, pp. 437–475.

Case, Anne C, Harvey S Rosen and James R Hines Jr (1993). 'Budget spillovers and fiscal policy interdependence: Evidence from the states'. In: *Journal of public economics* 52.3, pp. 285–307.

Drukker, David M, Ingmar R Prucha and Rafal Raciborski (2013). 'Maximum likelihood and generalized spatial two-stage least-squares estimators for a spatial-autoregressive model with spatial-autoregressive disturbances'. In: *The Stata Journal* 13.2, pp. 221–241.

Hansen, Lars Peter (1982). 'Large sample properties of generalized method of moments estimators'. In: *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

- Huber, Stephan and Christoph Rust (2016). 'Calculate travel time and distance with OpenStreetMap data using the Open Source Routing Machine (OSRM)'. In: *The Stata Journal* 16.2, pp. 416–423.
- Jann, Ben (2023). 'geoplot: A new command to draw maps'. In.
- Kelejian, Harry H and Ingmar R Prucha (1998). 'A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances'. In: *The Journal of Real Estate Finance and Economics* 17.1, pp. 99–121.
- Klier, Thomas and Daniel P McMillen (2008). 'Clustering of auto supplier plants in the United States: generalized method of moments spatial logit for large samples'. In: *Journal of Business & Economic Statistics* 26.4, pp. 460–471.
- Ozimek, Adam and Daniel Miles (2011). 'Stata utilities for geocoding and generating travel time and travel distance information'. In: *The Stata Journal* 11.1, pp. 106–119.
- Pinkse, Joris and Margaret E Slade (1998). 'Contracting in space: An application of spatial statistics to discrete-choice models'. In: *Journal of Econometrics* 85.1, pp. 125–154.

- Pisati, Maurizio (2001). 'sg162: tools for spatial data analysis'. In: *Stata Technical Bulletin* 60, pp. 21–37.
- — (2018). 'SPMAP: Stata module to visualize spatial data'. In.
- Spinelli, Daniele (2022). 'Fitting spatial autoregressive logit and probit models using Stata: The spatbinary command'. In: *The Stata Journal* 22.2, pp. 293–318.
- Tomelleri, Alessio and Anna Gloria Billé (2024). 'Evidence after the COVID-19 pandemic using a Spatial hurdle model'. In: *SSRN*.
- Weber, Sylvain and Martin Péclat (2017). 'A simple command to calculate travel distance and travel time'. In: *The Stata Journal* 17.4, pp. 962–971.

Marginal effects for the hurdle Model

(Tomelleri and Billé 2024) The marginal effects were calculated considering the spatial hurdle model in reduced form.

The marginal effects with respect to a continuous variable x_h are calculated as follows

$$\frac{\partial P\left(y_{i2}=1|x_{i2}\right)}{\partial x_{ih}}\,|_x = \Phi(x'_{i1}\beta_1)\phi\left(\{\Sigma_{\varepsilon_2^*}\}_{ii}^{-1/2}\left\{A^{-1}X_2\right\}_i\beta_2\right)\{\Sigma_{\varepsilon_2^*}^{-1/2}\}_{ii}\{A^{-1}\}_i\beta_{2h}$$

where x_h is the n–dimensional vector of units referred to the h–th continuous regressor included *only* in the set X_2, $\{.\}_i$ is the i–th row of the matrix inside, and $\{.\}_{ii}$ is the i–th diagonal element of a square matrix.

Please, see details in Tomelleri and Billè (2024) at SSRN.