spivregress — Spatial autoregressive models with endogenous covariates

Description

`spivregress` is the equivalent of `ivregress` for spatial data. `spivregress` fits spatial autoregressive (SAR) models, also known as simultaneous autoregressive models, where the models may contain additional endogenous variables as well as exogenous variables. These models can be used to account for possible dependence between the outcome variable and the unobserved errors.

For models without endogenous regressors, see [SP] **spregress**.

If you have not read [SP] **Intro 1**–[SP] **Intro 8**, you should do so before using `spivregress`. Your data must be Sp data to use `spivregress`. See [SP] **Intro 3** for instructions on how to prepare your data.

To specify spatial lags, you will need to have one or more spatial weighting matrices. See [SP] **Intro 2** and [SP] **spmatrix** for an explanation of the types of weighting matrices and how to create them.

Quick start

Spatial autoregressive model of `y1` regressed on `x1`, `x2`, endogenous regressor `y2`, which uses `z1` as an instrument, and a spatial lag for `y1` specified by the weighting matrix `W`

```
spivregress y1 x1 x2 (y2 = z1), dvarlag(W)
```

Add an autoregressive error term with the lag given by `M`

```
spivregress y1 x1 x2 (y2 = z1), dvarlag(W) errorlag(M)
```

Add a spatial lag for the exogenous variable `x1` based on `W`

```
spivregress y1 x1 x2 (y2 = z1), dvarlag(W) errorlag(M) ivarlag(W: x1)
```

Add a second spatial lag for the outcome variable based on the weighting matrix `M`

```
spivregress y1 x1 x2 (y2 = z1), dvarlag(W) errorlag(M)        ///
   dvarlag(M) ivarlag(W: x1)
```

Add interaction between `x1` and `x2` and add categorical instrument `z2` using factor-variable notation

```
spivregress y1 x1 x2 c.x1#c.x2 (y2 = z1 i.z2), dvarlag(W)      ///
   errorlag(M) dvarlag(M) ivarlag(W: x1 x2 c.x1#c.x2)
```

Menu

Statistics > Spatial autoregressive models

Syntax

`spivregress` *depvar* [*varlist*$_1$] (*varlist*$_2$ = *varlist*$_{\text{iv}}$) [*if*] [*in*] [, *options*]

varlist$_1$ is the list of included exogenous regressors.

varlist$_2$ is the list of endogenous regressors.

varlist$_{\text{iv}}$ is the list of excluded exogenous regressors used with *varlist*$_1$ as instruments for *varlist*$_2$.

options	Description
Model	
`dvarlag(`*spmatname*`)`	spatially lagged dependent variable; repeatable
`errorlag(`*spmatname*`)`	spatially lagged errors; repeatable
`ivarlag(`*spmatname* : *varlist*`)`	spatially lagged exogenous variables from *varlist*$_1$; repeatable
`noconstant`	suppress constant term
`heteroskedastic`	treat errors as heteroskedastic
`force`	allow estimation when estimation sample is a subset of the sample used to create the spatial weighting matrix
`impower(#)`	order of instrumental-variable approximation
Reporting	
`level(#)`	set confidence level; default is `level(95)`
display_options	control columns and column formats, row spacing, line width, display of omitted variables and base and empty cells, and factor-variable labeling
Optimization	
optimization_options	control the optimization process; seldom used
`coeflegend`	display legend instead of statistics

varlist$_1$, *varlist*$_2$, *varlist*$_{\text{iv}}$, and *varlist* specified in `ivarlag()` may contain factor variables; see **[U] 11.4.3 Factor variables**.

`collect` is allowed; see **[U] 11.1.10 Prefix commands**.

`coeflegend` does not appear in the dialog box.

See **[U] 20 Estimation and postestimation commands** for more capabilities of estimation commands.

Options

Model

`dvarlag(`*spmatname*`)` specifies a spatial weighting matrix that defines a spatial lag of the dependent variable. This option is repeatable to allow higher-order models. By default, no spatial lags of the dependent variable are included.

`errorlag(`*spmatname*`)` specifies a spatial weighting matrix that defines a spatially lagged error. This option is repeatable to allow higher-order models. By default, no spatially lagged errors are included.

`ivarlag(`*spmatname* : *varlist*`)` specifies a spatial weighting matrix and a list of exogenous variables that define spatial lags of the variables. The variables in *varlist* must be a subset of the exogenous variables in *varlist*$_1$. This option is repeatable to allow spatial lags created from different matrices. By default, no spatial lags of the exogenous variables are included.

`noconstant`; see **[R] Estimation options**.

`heteroskedastic` specifies that the estimator treat the errors as heteroskedastic instead of homoskedastic, which is the default; see *Methods and formulas* in [SP] **spregress**.

`force` requests that estimation be done when the estimation sample is a proper subset of the sample used to create the spatial weighting matrices. The default is to refuse to fit the model. Weighting matrices potentially connect all the spatial units. When the estimation sample is a subset of this space, the spatial connections differ and spillover effects can be altered. In addition, the normalization of the weighting matrix differs from what it would have been had the matrix been normalized over the estimation sample. The better alternative to `force` is first to understand the spatial space of the estimation sample and, if it is sensible, then create new weighting matrices for it. See [SP] **spmatrix** and *Missing values, dropped observations, and the W matrix* in [SP] **Intro 2**.

`impower(#)` specifies the order of an instrumental-variable approximation used in fitting the model. The derivation of the estimator involves a product of # matrices. Increasing # may improve the precision of the estimation and will not cause harm, but will require more computer time. The default is `impower(2)`. See *Methods and formulas* for additional details on `impower(#)`.

Reporting

`level(#)`; see [R] **Estimation options**.

display_options: `noci`, `nopvalues`, `noomitted`, `vsquish`, `noemptycells`, `baselevels`, `allbaselevels`, `nofvlabel`, `fvwrap(#)`, `fvwrapon(`*style*`)`, `cformat(%`*fmt*`)`, `pformat(%`*fmt*`)`, `sformat(%`*fmt*`)`, and `nolstretch`; see [R] **Estimation options**.

Optimization

optimization_options: `iterate(#)`, [`no`]`log`, `trace`, `gradient`, `showstep`, `hessian`, `showtolerance`, `tolerance(#)`, `ltolerance(#)`, `nrtolerance(#)`, and `nonrtolerance`; see [M-5] **optimize()**.

The following option is available with `spivregress` but is not shown in the dialog box:

`coeflegend`; see [R] **Estimation options**.

Remarks and examples

See [SP] **Intro** for an overview of SAR models.

`spivregress` fits spatial autoregressive models that include endogenous regressors. The `spivregress` command is for use with cross-sectional data. It requires each observation to represent one unique spatial unit. See [SP] **Intro 3** and the introductory sections that follow for instructions with examples on how to prepare your data for analysis with `spivregress`.

`spivregress` fits models like the following:

```
spivregress y1 x1 x2 (y2 y3 = z1 z2 z3), dvarlag(W) errorlag(M) ///
    ivarlag(W: x1)
```

`dvarlag(W)` specifies a spatial lag of the dependent variable `y1`, with the formulation of the lag given by the spatial weighting matrix `W`. You can include multiple `dvarlag()` options, each with different weighting matrices, to model higher-order spatial lags of the dependent variable.

`errorlag(M)` specifies an autoregressive error term based on the weighting matrix `M`. You can include multiple `errorlag()` options.

ivarlag(W: x1) specifies a spatial lag of the exogenous variable x1. You cannot include in the model spatial lags of the endogenous regressors y2 and y3 or spatial lags of the excluded exogenous regressors z1, z2, and z3.

spivregress uses a generalized method of moments estimator known as generalized spatial two-stage least squares (GS2SLS), the same estimator used by spregress, gs2sls. See *Methods and formulas*. Also see *Choosing weighting matrices and their normalization* in [SP] **spregress** for details about the GS2SLS estimator.

Example 1: SAR models with endogenous regressors

Suppose we want to know whether prohibiting alcohol sales in a county decreases the rate of arrests for driving under the influence (DUI). We use the artificial dataset dui_southern.dta, containing DUI rates in counties in southern states of the United States.

Because the analysis dataset and the Stata-formatted shapefile must be in our working directory to spset the data, we first save both dui_southern.dta and dui_southern_shp.dta to our working directory by using the copy command. We then load the data and type spset to see the Sp settings.

```
. copy https://www.stata-press.com/data/r19/dui_southern.dta .
. copy https://www.stata-press.com/data/r19/dui_southern_shp.dta .
. use dui_southern
. spset
      Sp dataset: dui_southern.dta
Linked shapefile: dui_southern_shp.dta
            Data: Cross sectional
 Spatial-unit ID: _ID
     Coordinates: _CX, _CY (planar)
```

The outcome of interest is dui, which is the alcohol-related arrest rate per 100,000 daily vehicle miles traveled (DVMT). Explanatory variables include police, the number of sworn officers per 100,000 DVMT; nondui, the nonalcohol-related arrest rate per 100,000 DVMT; vehicles, the number of registered vehicles per 1,000 residents; and dry, a variable that indicates whether a county prohibits the sale of alcohol within its borders.

Because the size of the police force may be a function of dui and nondui arrest rates, we treat police as endogenous. We assume the variable election is a valid instrument, where election is 1 if the county government faces an election and is 0 otherwise.

We believe the DUI arrest rate to be spatially correlated, with the rate in a county affecting the rates in neighboring counties. Formally, the model we want to fit is

$$\begin{aligned}\texttt{dui} &= \beta_0 + \beta_1 \times \texttt{nondui} + \beta_2 \times \texttt{dry} + \beta_3 \times \texttt{vehicles} + \pi_1 \times \texttt{police} + \lambda \mathbf{W} \times \texttt{dui} + \mathbf{u} \\ \mathbf{u} &= \rho \mathbf{W}\,\mathbf{u} + \epsilon\end{aligned}$$

The term $\mathbf{W} \times$ dui defines a spatial lag of dui. See [SP] **Intro 2** for an explanation of how spatial lags are defined by weighting matrices, and see *Choosing weighting matrices and their normalization* in [SP] **spregress**. The equation for **u** gives the error an autoregressive form also specified by the weighting matrix **W**. The variable police is endogenous and may be correlated with the error **u**. We instrument it with the variable election. See *Methods and formulas* for how the endogeneity of police is handled by the estimator.

Before we can fit the model, we must create the weighting matrix **W**. We will create one that puts the same positive weight on contiguous counties and a 0 weight on all other counties—a matrix known as a contiguity matrix. We will use the default spectral normalization for the matrix. See [SP] **Intro 2** and [SP] **spmatrix create** for details. We type

```
. spmatrix create contiguity W
```

We fit the model by typing

```
. spivregress dui nondui vehicles i.dry (police = elect), dvarlag(W) errorlag(W)
  (1422 observations)
  (1422 observations (places) used)
  (weighting matrix defines 1422 places)
Estimating rho using 2SLS residuals:
Initial:      GMM criterion =  .00254902
Alternative:  GMM criterion =  .00377532
Rescale:      GMM criterion =  .00009468
Iteration 0:  GMM criterion =  .00009468
Iteration 1:  GMM criterion =  .00001513
Iteration 2:  GMM criterion =  .00001512
Estimating rho using GS2SLS residuals:
Iteration 0:  GMM criterion =  .00086665
Iteration 1:  GMM criterion =  .00085487
Iteration 2:  GMM criterion =  .00085486
Spatial autoregressive model                         Number of obs =   1,422
GS2SLS estimates                                     Wald chi2(5)  = 4393.21
                                                     Prob > chi2   =  0.0000
                                                     Pseudo R2     =  0.7378
```

dui	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
dui						
police	-1.283189	.1138994	-11.27	0.000	-1.506428	-1.059951
nondui	-.001833	.0025467	-0.72	0.472	-.0068245	.0031585
vehicles	.0906069	.0045059	20.11	0.000	.0817755	.0994384
dry						
Yes	.4631025	.076754	6.03	0.000	.3126674	.6135377
_cons	8.714745	1.060428	8.22	0.000	6.636345	10.79315
W						
dui	.3859225	.0194397	19.85	0.000	.3478214	.4240235
e.dui	.2169234	.0496595	4.37	0.000	.1195926	.3142541

```
Wald test of spatial terms: chi2(2) = 408.78           Prob > chi2 = 0.0000
Endogenous: police (W*dui)
Exogenous:  nondui vehicles 1.dry election dui:_cons
```

When a spatial lag of the dependent variable is included in the model, covariates have both direct and indirect effects. See example 1 of [SP] **spregress** for a discussion. To obtain the direct, indirect, and total effects of the covariates, we must use `estat impact`:

```
. estat impact
progress    : 20%  40%  60%  80% 100%
Average impacts                                  Number of obs     =      1,422
```

	dy/dx	Delta-Method std. err.	z	P>|z|	[95% conf.	interval]
direct						
police	-1.313426	.1198948	-10.95	0.000	-1.548416	-1.078437
nondui	-.0018762	.0026073	-0.72	0.472	-.0069864	.003234
vehicles	.092742	.0048427	19.15	0.000	.0832504	.1022336
dry						
Yes	.4740151	.0788695	6.01	0.000	.3194336	.6285966
indirect						
police	-.6465736	.1063216	-6.08	0.000	-.8549601	-.4381871
nondui	-.0009236	.0012928	-0.71	0.475	-.0034576	.0016103
vehicles	.045655	.0057216	7.98	0.000	.0344409	.0568692
dry						
Yes	.2333482	.0464145	5.03	0.000	.1423774	.3243189
total						
police	-1.96	.2258604	-8.68	0.000	-2.402678	-1.517322
nondui	-.0027998	.0038989	-0.72	0.473	-.0104416	.0048419
vehicles	.138397	.0105248	13.15	0.000	.1177688	.1590253
dry						
Yes	.7073633	.123289	5.74	0.000	.4657213	.9490052

While it is running, `estat impact` prints percentages at the top of the output to indicate progress. Calculation of the standard errors of the effects can be intensive and take time, so it reports its progress as it does the computations.

The average direct, or own-county, effect of going from a wet county to a dry county on alcohol-related arrest rates is positive. The average indirect, or spillover, effect of going from a wet county to a dry county on alcohol-related arrest rates is also positive. The total effects are the sum of the direct and indirect effects, so these are also positive.

◁

▷ Example 2: SAR models with endogenous regressors and covariate lags

Continuing with example 1, we found that dry, we now add a spatial lag of the covariate dry.

```
. spivregress dui nondui vehicles i.dry (police = elect), dvarlag(W)
> errorlag(W) ivarlag(W: i.dry)
  (1422 observations)
  (1422 observations (places) used)
  (weighting matrix defines 1422 places)
note: exog*W:0b.dry omitted because of collinearity.
```

(output omitted)

```
Spatial autoregressive model                          Number of obs =    1,422
GS2SLS estimates                                      Wald chi2(6)  = 4300.29
                                                      Prob > chi2   =  0.0000
                                                      Pseudo R2     =  0.7337

------------------------------------------------------------------------------
         dui | Coefficient  Std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
dui          |
      police |  -1.301634   .1155866   -11.26   0.000     -1.52818   -1.075089
      nondui |  -.0018725   .0025746    -0.73   0.467    -.0069187    .0031737
    vehicles |    .091364   .0045754    19.97   0.000     .0823965    .1003316
             |
         dry |
        Yes  |   .4754855    .078153     6.08   0.000     .3223085    .6286626
       _cons |   8.853401    1.07409     8.24   0.000     6.748223    10.95858
-------------+----------------------------------------------------------------
W            |
         dry |
        Yes  |   .2868458   .2209814     1.30   0.194    -.1462697    .7199613
         dui |     .38758   .0196366    19.74   0.000      .349093    .4260669
       e.dui |   .2196418   .0497708     4.41   0.000     .1220929    .3171908
------------------------------------------------------------------------------
Wald test of spatial terms: chi2(3) = 405.90              Prob > chi2 = 0.0000
Endogenous: police (W*dui)
Exogenous:  nondui vehicles 1.dry election (W*0b.dry) (W*1.dry) dui:_cons
```

We use estat impact to see the effects:

```
. estat impact
progress   : 20%   40%   60%   80% 100%
Average impacts                                 Number of obs     =      1,422

------------------------------------------------------------------------------
             |            Delta-Method
             |      dy/dx   std. err.      z    P>|z|     [95% conf. interval]
-------------+----------------------------------------------------------------
direct       |
      police |  -1.332603   .1217453   -10.95   0.000    -1.571219   -1.093986
      nondui |   -.001917   .0026364    -0.73   0.467    -.0070844    .0032503
    vehicles |   .0935378   .0049201    19.01   0.000     .0838945    .1031811
             |
         dry |
        Yes  |   .5044067   .0833742     6.05   0.000     .3409963     .667817
-------------+----------------------------------------------------------------
indirect     |
      police |  -.6601862   .1089584    -6.06   0.000    -.8737408   -.4466316
      nondui |  -.0009497   .0013158    -0.72   0.470    -.0035287    .0016293
    vehicles |   .0463396   .0058501     7.92   0.000     .0348737    .0578055
             |
         dry |
        Yes  |   .6165397   .3004056     2.05   0.040     .0277555    1.205324
-------------+----------------------------------------------------------------
total        |
      police |  -1.992789   .2303197    -8.65   0.000    -2.444207   -1.541371
      nondui |  -.0028668    .003951    -0.73   0.468    -.0106106    .0048771
    vehicles |   .1398774   .0107284    13.04   0.000     .1188501    .1609047
             |
         dry |
        Yes  |   1.120946   .3442805     3.26   0.001      .446169    1.795724
------------------------------------------------------------------------------
```

The direct effect of dry is little changed when we added a lag of dry, going from 0.47 to 0.50. But the indirect effects of dry go from 0.23 to 0.62. In these fictional data, the indirect effects of dry become larger than the direct effects when there is a lag of dry in the model.

Note that spivregress does not allow the fitting of spatial lags for police, our endogenous regressor, nor for election, its instrument.

◁

▷ Example 3: SAR models with endogenous regressors and higher-order lags

In the previous models, we specified all the spatial lags with a single weighting matrix W, a contiguity weighting matrix with the default spectral normalization. Many researchers use a spatial weighting matrix whose (i, j)th element is the inverse of the distance between units i and j. With the GS2SLS estimator used by spivregress, we can include spatial lags using two spatial weighting matrices. This can be done to model a "higher-order" approximation to the true spatial process. We will now add lags specified by an inverse-distance matrix, using again a spectral normalization of the matrix.

We create the inverse-distance matrix `M` and use `spmatrix dir` to list our Sp matrices.

```
. spmatrix create idistance M

. spmatrix dir
```

Weighting matrix name	N x N	Type	Normalization
M	1422 x 1422	idistance	spectral
W	1422 x 1422	contiguity	spectral

We fit the model including both weighting matrices for all the lags:

```
. spivregress dui nondui vehicles i.dry (police = elect), dvarlag(W)
> errorlag(W) ivarlag(W: i.dry) dvarlag(M) errorlag(M) ivarlag(M: i.dry)
  (1422 observations)
  (1422 observations (places) used)
  (weighting matrices define 1422 places)
note: exog*W:0b.dry omitted because of collinearity.
note: exog*M:0b.dry omitted because of collinearity.
```

(output omitted)

```
Spatial autoregressive model                         Number of obs =    1,422
GS2SLS estimates                                     Wald chi2(8)  = 6447.40
                                                     Prob > chi2   =  0.0000
                                                     Pseudo R2     =  0.8058
```

dui	Coefficient	Std. err.	z	P>|z|	[95% conf.	interval]
dui						
police	-.9762511	.0782532	-12.48	0.000	-1.129625	-.8228777
nondui	-.0010538	.002093	-0.50	0.615	-.005156	.0030484
vehicles	.0786513	.0031164	25.24	0.000	.0725432	.0847594
dry						
Yes	.4207578	.0631514	6.66	0.000	.2969833	.5445323
_cons	6.067956	.749057	8.10	0.000	4.599831	7.536081
W						
dry						
Yes	.2354006	.2272314	1.04	0.300	-.2099647	.6807659
dui	.3335356	.0134262	24.84	0.000	.3072207	.3598505
e.dui	.2206959	.0630469	3.50	0.000	.0971263	.3442655
M						
dry						
Yes	-.0924289	2.709074	-0.03	0.973	-5.402117	5.217259
dui	.000521	.0112679	0.05	0.963	-.0215636	.0226056
e.dui	-.1069377	.5910201	-0.18	0.856	-1.265316	1.05144

```
Wald test of spatial terms: chi2(6) = 649.09             Prob > chi2 = 0.0000
Endogenous: police (W*dui) (M*dui)
Exogenous:  nondui vehicles 1.dry election (W*0b.dry) (W*1.dry) (M*0b.dry)
            (M*1.dry) dui:_cons
```

All the spatial lags specified by the inverse-distance matrix `M` are nonsignificant. We conclude that there are no inverse-distance-type effects after we account for contiguity-type effects.

◁

Stored results

spivregress stores the following in e():

Scalars	
e(N)	number of observations
e(k)	number of parameters
e(df_m)	model degrees of freedom
e(df_c)	degrees of freedom for comparison test
e(iterations)	number of generalized method of moments iterations
e(iterations_2sls)	number of two-stage least-squares iterations
e(rank)	rank of e(V)
e(r2_p)	pseudo-R^2
e(chi2)	χ^2
e(chi2_c)	χ^2 for comparison test
e(p)	p-value for model test
e(p_c)	p-value for test of spatial terms
e(converged)	1 if generalized method of moments converged, 0 otherwise
e(converged_2sls)	1 if two-stage least-squares converged, 0 otherwise
Macros	
e(cmd)	spivregress
e(cmdline)	command as typed
e(depvar)	name of dependent variable
e(indeps)	names of independent variables
e(idvar)	name of ID variable
e(estimator)	gs2sls
e(title)	title in estimation output
e(constant)	hasconstant or noconstant
e(endog)	names of endogenous variables
e(exog)	names of exogenous variables
e(dlmat)	names of spatial weighting matrices applied to *depvar*
e(elmat)	names of spatial weighting matrices applied to errors
e(het)	heteroskedastic or homoskedastic
e(chi2type)	Wald; type of model χ^2 test
e(properties)	b V
e(estat_cmd)	program used to implement estat
e(predict)	program used to implement predict
e(marginsok)	predictions allowed by margins
e(marginsnotok)	predictions disallowed by margins
e(asbalanced)	factor variables fvset as asbalanced
e(asobserved)	factor variables fvset as asobserved
Matrices	
e(b)	coefficient vector
e(delta_2sls)	two-stage least-squares estimates of coefficients in spatial lag equation
e(rho_2sls)	generalized method of moments estimates of coefficients in spatial error equation
e(V)	variance–covariance matrix of the estimators
Functions	
e(sample)	marks estimation sample

Methods and formulas

We consider a cross-sectional spatial autoregressive model with possible endogenous covariates and spatial autoregressive disturbances (SARAR), allowing for higher-order spatial dependence in the dependent variable, the exogenous variables, and the spatial errors. The model is

$$\begin{aligned} \mathbf{y} &= \sum_{j=1}^{J} \pi_j \tilde{\mathbf{y}}_j + \sum_{k=1}^{K} \beta_k \mathbf{x}_k + \sum_{p=1}^{P} \gamma_p \mathbf{W}_p \mathbf{x}_p + \sum_{r=1}^{R} \lambda_r \mathbf{W}_r \mathbf{y} + \mathbf{u} \\ \mathbf{u} &= \sum_{s=1}^{S} \rho_s \mathbf{M}_s \mathbf{u} + \boldsymbol{\epsilon} \end{aligned} \tag{1}$$

where

$\mathbf{y}$ is an $n \times 1$ vector of observations on the dependent variable;

$\tilde{\mathbf{y}}_j$ is an $n \times 1$ vector of observations on the jth endogenous variable; π_j is the corresponding scalar parameter;

$\mathbf{x}_k$ is an $n \times 1$ vector of observations on the kth exogenous variable; β_k is the corresponding scalar parameter;

$\mathbf{W}_p$, $\mathbf{W}_r$, and $\mathbf{M}_s$ are $n \times n$ spatial weighting matrices;

$\mathbf{W}_p \mathbf{x}_p$, $\mathbf{W}_r \mathbf{y}$, and $\mathbf{M}_s \mathbf{u}$ are $n \times 1$ spatial lags for the exogenous variable, dependent variable, and error terms; γ_p, λ_r, and ρ_s are scalar parameters; and

$\boldsymbol{\epsilon}$ is an $n \times 1$ vector of innovations.

The J endogenous variables $\tilde{\mathbf{y}}_j$ are correlated with the errors $\mathbf{u}$. To estimate the model parameters, we need Q instrumental variables $\mathbf{x}_1^e, \mathbf{x}_2^e, \dots, \mathbf{x}_Q^e$ with $Q \geq J$ that are correlated with the endogenous variables in $\tilde{\mathbf{y}}_j$ and uncorrelated with the errors $\mathbf{u}$.

The model in (1) is frequently referred to as a higher-order spatial autoregressive model with spatial autoregressive disturbances, or namely, a SARAR(R, S) model.

The innovations $\boldsymbol{\epsilon}$ are assumed to be independent and identically distributed or independent but heteroskedastically distributed, where the heteroskedasticity is of unknown form. The generalized spatial two-stage least-squares (GS2SLS) estimator implemented in `spivregress` produces consistent estimates in both cases when the `heteroskedastic` option is specified.

For the first-order SARAR model, `spivregress` implements the GS2SLS estimator discussed in Arraiz et al. (2010) and Drukker, Egger, and Prucha (2013). This estimation strategy builds on Kelejian and Prucha (1998, 1999, 2010) and references cited therein. For higher-order SARAR(R, S) models, `spivregress` implements an extension of GS2SLS in Badinger and Egger (2011) to allow endogenous covariates.

Let's first rewrite (1) in a compact form.

$$\begin{aligned} \mathbf{y} &= \mathbf{Z}\delta + \mathbf{u} \\ \mathbf{u} &= \overline{\mathbf{U}}\boldsymbol{\rho} + \boldsymbol{\epsilon} \end{aligned} \tag{2}$$

where

$\mathbf{Z}$ is the matrix of observations on all the variables in the equation for $\mathbf{y}$; $\mathbf{Z}$ contains the endogenous covariates $\tilde{\mathbf{y}}_1, \dots, \tilde{\mathbf{y}}_J$, the exogenous covariates $\mathbf{x}_1, \dots, \mathbf{x}_K$, the spatially lagged exogenous covariates $\mathbf{W}\mathbf{x}_1, \dots, \mathbf{W}\mathbf{x}_P$, and the spatially lagged dependent variables $\mathbf{W}\mathbf{y}_1, \dots, \mathbf{W}\mathbf{y}_R$;

$\overline{\mathbf{U}}$ contains all the spatial lags of the errors $\mathbf{u}$ that appear in (1); $\overline{\mathbf{U}}$ contains $\mathbf{M}_1\mathbf{u}, \ldots, \mathbf{M}_S\mathbf{u}$;

$\boldsymbol{\delta} = (\pi_1, \ldots \pi_J, \beta_1, \ldots, \beta_K, \gamma_1, \ldots, \gamma_P, \lambda_1, \ldots, \lambda_R)'$ is a vector of all the coefficients on the variables in the equation for $\mathbf{y}$; and

$\boldsymbol{\rho} = (\rho_1, \ldots, \rho_S)$ is the vector of coefficients on the spatially lagged errors.

Given these definitions, the estimator implemented in `spivregress` is a simple extension to the GS2SLS estimator documented in the *Methods and formulas* of `spregress`.

Specifically, after adding the instrumental variables $\mathbf{x}_1^e, \mathbf{x}_2^e, \ldots, \mathbf{x}_Q^e$ to the list of exogenous variables $\mathbf{X}_f$ used to create the matrix of instruments $\mathbf{H}_1$ in `spregress`, the other formulas in `spregress` specify how the estimator implemented in `spivregress` works. See *Methods and formulas* in [SP] **spregress** for further details.

References

Arraiz, I., D. M. Drukker, H. H. Kelejian, and I. R. Prucha. 2010. A spatial Cliff–Ord-type model with heteroskedastic innovations: Small and large sample results. *Journal of Regional Science* 50: 592–614. https://doi.org/10.1111/j.1467-9787.2009.00618.x.

Badinger, H., and P. H. Egger. 2011. Estimation of higher-order spatial autoregressive cross-section models with heteroscedastic disturbances. *Papers in Regional Science* 90: 213–235. https://doi.org/10.1111/j.1435-5957.2010.00323.x.

Drukker, D. M., P. H. Egger, and I. R. Prucha. 2013. On two-step estimation of a spatial autoregressive model with autoregressive disturbances and endogenous regressors. *Econometric Reviews* 32: 686–733. https://doi.org/10.1080/07474938.2013.741020.

Kelejian, H. H., and I. R. Prucha. 1998. A generalized spatial two-stage least squares procedure for estimating a spatial autoregressive model with autoregressive disturbances. *Journal of Real Estate Finance and Economics* 17: 99–121. https://doi.org/10.1023/A:1007707430416.

———. 1999. A generalized moments estimator for the autoregressive parameter in a spatial model. *International Economic Review* 40: 509–533. https://doi.org/10.1111/1468-2354.00027.

———. 2010. Specification and estimation of spatial autoregressive models with autoregressive and heteroskedastic disturbances. *Journal of Econometrics* 157: 53–67. https://doi.org/10.1016/j.jeconom.2009.10.025.

Also see

[SP] **spivregress postestimation** — Postestimation tools for spivregress

[SP] **estat moran** — Moran's test of residual correlation with nearby residuals

[SP] **Intro** — Introduction to spatial data and SAR models

[SP] **spmatrix** — Categorical guide to the spmatrix command

[SP] **spregress** — Spatial autoregressive models

[SP] **spxtregress** — Spatial autoregressive models for panel data

[R] **ivregress** — Single-equation instrumental-variables regression

[U] 20 Estimation and postestimation commands

For suggested citations, see the FAQ on citing Stata documentation.