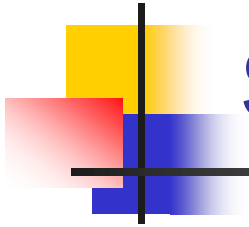
Implementing procedures for spatial panel econometrics in Stata

Gordon Hughes
University of Edinburgh

Stata User Group Meeting
15th September 2011



Spatial analysis in Stata

- Variety of special purpose routines written by users and available through SSC
  - Manipulation of spatial data
  - Cross-section spatial regressions
- StataCorp-related routines – also through SSC
  - shp2dta converts ESRI shapefiles to dta files – similar to programs converting to csv or xls files
  - spmat, spreg, spivreg, etc for construction & manipulation of spatial weights and for cross-section spatial regressions

Nature of panel data

- Large N and/or large T?
- Balanced or unbalanced panels
- Spatial weights – interactions with missing data
- Examples:
  - State tax and fiscal policies
  - Cross-country models of economic development

Econometric specification

- Fixed or random effects – can we talk about random effects with complete sample of states or countries?
- Lagged dependent variable or within panel serial correlation
- Why are data missing – missing at random assumption

Key models

Spatial auto-regression model (SAR)

$$y_{it} = \rho W y_t + X_{it}\beta + \mu_i + \varepsilon_{it}$$

Spatial Durbin model (SDM)

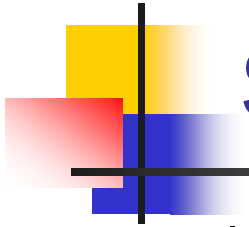
$$y_{it} = \rho W y_t + X_{it}\beta + WX_t\varphi + \mu_i + \varepsilon_{it}$$

Spatial error model (SEM)

$$y_{it} = X_{it}\beta + \mu_i + \nu_{it} \; with \; \nu_{it} = \lambda W \nu_t + \varepsilon_{it}$$

Estimation methods

- Instrumental variables
  - Use –spmat- & -ivreg2-
- GMM
  - Can be implemented using –gmm-
- Maximum likelihood
  - Routines based on Matlab procedures written by Elhorst, Pfaffermayr and others



Syntax

panel_spat_mle varlist [if] [in] , WMATrix(weight_matrix)
[FE RE TYPE (integer 1) NOConstant CLUSter
ROBust (integer -1) DLAG]

"varlist" = depvar indvars [required].

weight_matrix" should refer to an N x N Stata matrix of spatial weights [required].

“type(#)” - type of model to be estimated.

"noconstant" specifies that the model should be estimated without adding a constant term.

"re" - random effects model (default).

"fe" - fixed effects model.

"robust(#)" for Driscoll-Kraay robust standard errors (# = no of lags).

"cluster" - cluster robust standard errors should be used.

"dlag" - include the lagged dependent variable in the model.

Table of options

	Random effects		Fixed effects	
	No lagged dependent variable	With lagged dependent variable	No lagged dependent variable	With lagged dependent variable
Spatial autoregressive model (SAR)	TYPE(1) RE	TYPE(1) RE DLAG	TYPE(1) FE	TYPE(1) FE DLAG
Spatial Durbin model (SDM)	TYPE(2) RE	TYPE(2) RE DLAG	TYPE(2) FE	TYPE(2) FE DLAG
Spatial error model (SEM)	TYPE(3) RE	N/A	TYPE(3) FE	N/A

Illustration – US electricity demand

- State data – continental US, 1990-2009
  - Electricity demand by sector
  - Regressors - prices, weather (heating & cooling days)
- Focus on price elasticities and weather impacts
- Likely to be spatial interactions due to
  - Common factors in unobserved variables
  - Competition between states for industry and/or movement of households

Table 2 – Fixed effect models for residential electricity demand

	Dependent variable – ln(residential electricity consumption per person)					
Variables	Non-spatial panel	SAR	SAR + Dlag	SDM	SDM + Dlag	SEM
	(1)	(2)	(3)	(4)	(5)	(6)
Y[t-1]			0.314***		0.365***	
			(0.025)		(0.031)	
W*Y		0.471***	0.548***	0.462***	0.460***	
		(0.027)	(0.020)	(0.033)	(0.021)	
ln(GDP per person)	0.279***	0.088	0.018	0.063	0.046	0.471
	(0.045)	(0.087)	(0.064)	(0.037)	(0.037)	(11.380)
ln(Average residential price)	-0.284***	-0.269***	-0.170***	-0.311***	-0.189***	0.0884
	(0.048)	(0.041)	(0.028)	(0.039)	(0.032)	(3.677)
ln(Housing units per person)	1.101***	0.739***	0.227	0.560***	0.198**	-0.269
	(0.115)	(0.158)	(0.134)	(0.037)	(0.073)	(0.940)
ln(Cooling degree days)	0.072***	0.048***	0.065***	0.053***	0.059***	0.739
	(0.012)	(0.013)	(0.014)	(0.013)	(0.010)	(63.920)
ln(Year mean temperature)	-0.991***	-0.695***	-0.642***	-0.641***	-0.540***	0.048
	(0.099)	(0.166)	(0.130)	(0.163)	(0.139)	(0.507)
W*ln(GDP per person)				0.069	-0.02	
				(0.044)	(0.038)	
W*ln(Average residential price)				0.187***	0.083	
				(0.030)	(0.044)	
W*ln(Housing units per person)				0.256	-0.019	
				(0.204)	(0.201)	
W*ln(Cooling degree days)				0.001	0.018	
				(0.017)	(0.015)	
W*ln(Year mean temperature)				-0.091	-0.171	
				(0.248)	(0.235)	
Log-likelihood	1865	1986	2139	2030	2151	1986

Note: Robust standard errors in brackets under the coefficients - *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Table 3 – Fixed effect models for industrial electricity demand

	Dependent variable - ln(industrial electricity demand per person)					
	Non-spatial panel	SAR	SAR + Dlag	SDM	SDM + Dlag	SEM
Variables	(1)	(2)	(3)	(4)	(5)	(6)
Y[t-1]			0.044		0.044	
			(0.034)		(0.034)	
W*Y		0.116*	0.780***	0.133**	0.760***	
		(0.046)	(0.018)	(0.045)	(0.019)	
ln(GDP per person)	-0.380**	-0.334***	-0.115*	0.940***	0.214	0.116**
	(0.128)	(0.042)	(0.055)	(0.195)	(0.459)	(0.037)
ln(Average industrial price)	-0.392*	-0.376***	-0.174	-0.482***	-0.187	-0.334**
	(0.168)	(0.106)	(0.092)	(0.098)	(0.155)	(0.109)
ln(Year mean temperature)	0.656	0.645	0.260	1.426*	0.713	-0.376***
	(0.364)	(0.478)	(0.620)	(0.644)	(1.191)	(0.025)
W*ln(GDP per person)				-1.357***	-0.367	
				(0.214)	(0.447)	
W*ln(Average industrial price)				0.174*	-0.011	
				(0.075)	(0.194)	
W*ln(Year mean temperature)				-0.957	-0.623	
				(0.728)	(1.046)	
Log-likelihood	195	198	733	249	743	198

Note: Robust standard errors in brackets under the coefficients - *** $p < 0.001$, ** $p < 0.01$, * $p < 0.05$.

Calculating elasticities

- Direct effect (spatial Durbin model)

$$M_{dir}(k) = trace([I - \rho W]^{-1}[I_N \beta_k + W\varphi_k])(1/N)$$

  - impact of a unit change in variable X_k in state i on demand in state i averaged over all states i = 1...N

- Total effect

$$M_{tot}(k) = i_N'([I - \rho W]^{-1}[I_N \beta_k + W\varphi_k])i_N(1/N)$$

  - the impact of the same unit change in variable X_k in all states on demand in state i, again averaged over all states

Direct and total price elasticities

	Residential		Industrial	
	Direct	Total	Direct	Total
Non-spatial panel	-0.28		-0.39	
SAR	-0.29	-0.51	-0.38	-0.43
SAR + Dlag - short run	-0.19	-0.38	-0.23	-0.79
SAR + Dlag - long run	-0.27	-0.55	-0.24	-0.83
SDM	-0.32	-0.45	-0.53	-2.12
SDM + Dlag - short run	-0.20	-0.39	-0.36	-2.30
SDM + Dlag - long run	-0.32	-0.61	-0.38	-2.41

Source: Derived from coefficient estimates in Tables 2 & 3.

Unbalanced panels - options

- Listwise deletion
  - Can mean loss of all or most of sample
- Single imputation
  - Particularly useful for spatial lags
  - See Cameron & Trivedi, Chap 27
- Multiple imputation
  - Computationally expensive
  - Will consider adding to -panel_spat_mle-

Spatial random effects model 1

- Pfaffermayr examines generalisation of RE-SEM by Baltagi et al

$$y_t = X_t\beta + m + \eta_t$$

$$where\ m = \rho Wm + \mu, \eta_t = \lambda W\eta_t + \varepsilon_t \ \ for\ t = 1...T$$

- Balanced panel model can be estimated by ML using methods similar to the RE-SEM model
  - manageable for models with N ~ 400

Spatial random effects model 2

- Unbalanced panel model can estimated by ML if we assume data is missing at random (MAR)
  - Likelihood derived from marginal distribution of the true model with respect to the observed model
  - Computation involves matrices of order min(M, NT-M) where M=the total number of missing observations
  - ML estimates are consistent estimates of the full assuming $\omega = M/NT$ converges to a constant as N increases

Spatial random effects model 3

- Stata version of Pfaffermayr's routine
  - Performs satisfactorily for datasets with N ~ 400 and N*T ~ 5000 with ω ~ 18% and ω ~ 30%
  - At present implemented as a separate procedure but the balanced data version could be added to panel_spat_mle as an additional model type

Random coefficient models

- Spatial mean groups
  - Development of Eberhardt's –xtmg- for spatial error model
  - Spatially lagged dependent variable – modify –xtmg- to use –ivreg2-
- FGLS
  - Methods for both SEM & SAR outlined by Elhorst but can be expensive since manipulation of large (N*T or N*K) matrices can't be avoided
  - Subject to Beck & Katz's observations on poor properties of FGLS for random coefficient models